Paolo Spirito

Elettronica digitale

McGraw-Hill Libri Italia srl

Milano • New York • St. Louis • San Francisco • Auckland • Bogotá
Caracas • Lisboa • London • Madrid • Mexico City • Montreal
New Delhi • San Juan • Singapore • Sydney • Tokyo • Toronto

Indice

Prefazione

Un dato ormai ampiamente riconosciuto è l'importanza che oggi riveste l'area dei circuiti digitali nel progetto dei sistemi elettronici. Occorre d'altra parte rilevare che, nel loro complesso, gli insegnamenti fondamentali di Elettronica nei corsi di Ingegneria trattano prevalentemente dell'elettronica analogica; a riprova di ciò si può citare l'organizzazione dei testi classici di Elettronica, che riservano all'analisi dei circuiti digitali solo alcuni capitoli, con un dimensionamento degli argomenti di solito inadeguato per coprire un corso base di Elettronica Digitale. Ciò può essere giustificato sia da ragioni storiche (ricordiamo che l'elettronica si è sviluppata in tempi relativamente recenti a partire dai sistemi di telecomunicazione, nei quali inizialmente l'elaborazione dell'informazione era analogica), sia dall'articolazione usuale delle materie presentate nel corso di laurea in Ingegneria Elettronica, che prevede insegnamenti sui dispositivi a semiconduttore e sui circuiti elettronici, materie che possono essere considerate propedeutiche ad un corso di Elettronica Digitale.

In base alla attuale articolazione dei corsi di laurea nel Settore dell'Ingegneria dell'Informazione, è oggi possibile pensare ad una organizzazione degli insegnamenti base dell'area Elettronica in "parallelo" anziché in serie, in modo da poter prevedere una utilizzazione anche di una sola di queste materie nei curricula di Ingegneria delle Telecomunicazioni e Ingegneria Informatica; in tal caso per questi corsi la scelta dovrebbe essere quella di un corso di Elettronica Digitale (o Elettronica dei sistemi digitali), che però non debba richiedere necessariamente il supporto propedeutico di contenuti forniti in altri insegnamenti come quelli sopra indicati. Ciò è a maggior ragione valido per l'articolazione dei corsi di Diploma Universitario del settore dell'Informazione.

Questo libro si propone di presentare gli elementi fondamentali dell'elettronica dei sistemi digitali, per un supporto didattico all'insegnamento dell'Elettronica Digitale dei corsi del settore dell'Informazione, senza richiedere necessariamente a priori la conoscenza dei principi di funzionamento dei circuiti analogici e dei dispositivi a semiconduttore, ma presentando ed utilizzando le metodologie, proprie dell'analisi dei circuiti non lineari, direttamente per la comprensione e lo studio dei circuiti digitali.

L'esposizione può essere suddivisa in tre parti.

Nella *prima parte* si introducono e si analizzano in dettaglio i blocchi funzionali di base – le porte logiche elementari – che costituiscono i "tasselli" con cui vengono realizzati i sottosistemi digitali più complessi. La descrizione e l'analisi dei blocchi funzionali elementari è essenzialmente quella circuitale, cioè in termini di tensioni e correnti invece che in termini di funzioni logiche (algebra di Boole, tabelle della verità). Si vuole infatti sottolineare che nei circuiti digitali effettivamente realizzabili, sia l'ampiezza dei segnali che la loro evoluzione temporale sono legate ad un comportamento "analogico" delle reti, che il progettista elettronico deve ben conoscere per poter garantire le specifiche di livelli logici e di ritardi temporali richieste dal sistema; questo approccio d'altra parte permette di sottolineare l'interazione dispositivo-circuito, e fornire tecniche di analisi nonlineare, tipicamente introdotte nei corsi di Elettronica, per la comprensione in generale dei circuiti elettronici nonlineari. In questa parte si è cercato di fornire gli elementi necessari alla comprensione del funzionamento dei dispositivi MOS e bipolari, in regime sia statico che dinamico; la fisica del funzionamento è stata introdotta su base essenzialmente descrittiva, ma cercando di fornire gli strumenti necessari sia per una valutazione quantitativa del loro funzionamento nei circuiti, che per una giustificazione dei modelli ad ampi segnali utilizzati dai simulatori circuitali, per i quali occorre fornire i parametri elettrici e geometrici necessari. In tutto il libro viene mantenuto un forte aggancio tra le caratteristiche elettriche dei circuiti e la loro realizzazione tecnologica come circuiti integrati, e nella prima parte vengono fornite le informazioni base sulle diverse tecnologie utilizzabili, essenzialmente al fine di poter considerare le limitazioni tecnologiche come elemento fondamentale nei criteri di scelta delle differenti soluzioni, e poter introdurre le regole di progetto dei tracciati (lay-out) dei circuiti analizzati come componente fondamentale per la definizione delle loro prestazioni.

La *seconda parte* introduce gradualmente il lettore ai problemi di interconnessione dei blocchi elementari precedentemente studiati, al fine di realizzare funzioni logiche più complesse: verranno quindi introdotte le modifiche alle porte elementari al fine di realizzare efficacemente interconnessioni e interfacciamenti tra queste, le soluzioni per la realizzazione dei circuiti di ingresso/uscita. Infine verranno presentate le soluzioni alternative per la realizzazione di circuiti a larga scala di integrazione (VLSI) con le attuali tecnologie MOS (logiche a porte di trasmissione, logiche dinamiche), e le possibilità che da queste ulteriori strutture logiche ne derivano.

Nella *terza parte* vengono utilizzate le conoscenze e gli strumenti introdotti precedentemente per analizzare e progettare i sottosistemi digitali più significativi, sia nel campo dei circuiti combinatori sia di quelli sequenziali e delle memorie, discutendone gli aspetti progettuali con le diverse tecnologie utilizzabili. Questo approccio è favorito dal fatto che nei sistemi digitali, ben più diffusamente che in quelli analogici, la progettazione di sistemi complessi si basa sull'utilizzo e sull'iterazione di un numero relativamente contenuto di strutture elementari, che combinate opportunamente danno luogo a sistemi altamente complessi.

La comprensione del funzionamento dei circuiti logici elementari è stata essenzialmente affidata a una descrizione analitica del loro comportamento, sia statico

che dinamico, che rendesse possibile una valutazione quantitativa, se pur approssimata, dei principali parametri elettrici di interesse applicativo, sia statici (livelli logici del segnale, margini di rumore, caratteristiche di trasferimento, di ingresso e di uscita) che dinamici (tempi di transizione, ritardi di propagazione, prodotto ritardo-potenza). La trattazione analitica permette infatti di comprendere meglio il funzionamento dei circuiti elementari, e di evidenziare le connessioni tra le caratteristiche elettriche presentate e i parametri progettuali a disposizione, nonché di paragonare le prestazioni e le limitazioni delle diverse famiglie tecnologiche oggi utilizzabili. Inoltre l'analisi "manuale" dei circuiti è la base necessaria per un efficiente utilizzo degli strumenti CAD (di fondamentale importanza per il progettista di circuiti digitali).

L'analisi sperimentale è stata sempre seguita da quella numerica, sia per sviluppare analisi più approfondite, sia per una migliore valutazione delle approssimazioni contenute nelle formulazioni analitiche. Nel testo è stato utilizzato intensivamente il simulatore SPICE nell'analisi sia delle porte elementari che dei circuiti combinatori e sequenziali, per una migliore comprensione dei fenomeni che determinano le loro prestazioni sia statiche che dinamiche, e come ausilio alla loro progettazione.

L'organizzazione del materiale presentato è la seguente.

Nel Capitolo 1 sono introdotti in via generale i principali parametri elettrici che definiscono le porte logiche elementari, con enfasi sulla funzione di discretizzazione dei segnali.

Nel Capitolo 2 vengono brevemente descritte le operazioni tecnologiche necessarie per la realizzazione di circuiti integrati, e si richiama il ruolo fondamentale delle "regole di progetto" nelle prestazioni dei circuiti.

I Capitoli 3 e 6 presentano i concetti fondamentali alla base del funzionamento rispettivamente del transistore MOS e di quello bipolare, e definiscono i parametri dinamici da introdurre nei modelli ad ampi segnali per una descrizione quantitativa, anche se approssimata, del funzionamento in commutazione. Vengono infine introdotti i modelli dei dispositivi utilizzati nel simulatore SPICE e definite la grandezze che occorre specificare per descrivere questi dispositivi.

I Capitoli 4 e 5 sono dedicati all'analisi delle porte elementari in tecnologia MOS (NMOS e CMOS), mentre i Capitoli 7, 8, 9 sono dedicati alle porte logiche bipolari, a partire da quelle "storiche" (RTL e DTL), per passare a quelle TTL e ECL.

Il Capitolo 10 è dedicato ai problemi di interconnessione, interfacciamento, e ai circuiti di ingresso/uscita. Vengono trattate le porte A-O-I, gli stadi buffer, quelli tri-state, le logiche BiCMOS. Inoltre vengono presentati i circuiti combinatori di maggior impiego, come gli operatori numerici, i circuiti di indirizzamento e di codifica, i PLA.

Nel Capitolo 11 sono presentate le configurazioni basate su tecnologia MOS, che trovano larga applicazione negli attuali circuiti integrati ad alta densità di integrazione (VLSI e ULSI), come le logiche a porte di trasmissione e le logiche dinamiche, e vengono discusse le implementazioni dei circuiti combinatori con queste strutture.

Il Capitolo 12 tratta i circuiti sequenziali, a partire dalle configurazioni elementari bistabili, passando per le famiglie di flip-flop, fino alle funzioni logiche realizzabili con questi ultimi, come i registri e i contatori. L'analisi è sviluppata sia per versioni MOS che bipolari, con attenzione ai problemi legati alla dinamica della commutazione e alla temporizzazione dei segnali.

Infine il Capitolo 13 è dedicato alla descrizione dei circuiti di memoria, comprendendo sia le memorie ROM che quelle W/R. Per queste ultime si è descritto il funzionamento dinamico delle celle di memoria elementari, bipolari e MOS, descrivendo il funzionamento sia delle memorie RAM statiche che di quelle dinamiche.

Alla fine di ogni capitolo sono riportati numerosi esercizi di riepilogo, allo scopo di sollecitare lo studente ad una utilizzazione delle formulazioni analitiche presentate per la risoluzione di problemi anche non specificamente indicati nel testo, e di familiarizzarlo alle verifiche, mediante uso del simulatore SPICE, delle analisi sviluppate. Nelle Appendici sono riportati i richiami all'uso del simulatore SPICE, le schede .MODEL dei dispositivi attivi utilizzati, e i listati dei file necessari alle simulazioni SPICE per la maggior parte delle porte logiche e dei circuiti presentati nel testo.

Il testo è dimensionato per un corso base di Elettronica indirizzato ai circuiti digitali; per contenere il materiale presentato nelle dimensioni utilizzabili da un corso universitario, si è preferito escludere alcuni argomenti che avrebbero richiesto un maggiore approfondimento degli aspetti tecnologici connessi, come ad esempio le logiche ad iniezione di corrente (I^2L), e le logiche all'arseniuro di Gallio.

Il libro è basato su di un'ampia rielaborazione ed aggiornamento del materiale presentato nel testo *Elettronica dei Sistemi Digitali* edito dall'Ateneo, ed utilizzato nel corso di Elettronica II tenuto presso l'Università di Napoli. In particolare, sulla base dell'esperienza didattica, sono stati completamente rielaborati e sviluppati i capitoli sulle logiche CMOS, quelli sugli stadi buffer e sui circuiti combinatori, quelli sulle logiche FCMOS e BiCMOS, quelli sulle logiche dinamiche e a porte di trasmissione, e il capitolo sulle memorie.

Si ringraziano gli studenti del corso per i loro contributi alla stesura di questo testo, e per le segnalazioni dei numerosi errori tipografici della prima edizione. Si desidera inoltre ringraziare sentitamente i colleghi del Dipartimento di Elettronica per i loro numerosi consigli e le osservazioni critiche. Infine un ringraziamento particolare va a mia figlia Francesca per la cura e l'impegno da lei dedicati alla correzione della bozze.

Paolo Spirito

1

Circuiti digitali

1.1 Introduzione

Nella normale accezione del termine, mutuato dal termine *digit* (*cifra*) del linguaggio anglosassone, l'*elettronica digitale* riguarda l'analisi e il progetto dei circuiti che trattano dati ed informazioni espresse attraverso due soli stati possibili, indicati rispettivamente con le cifre binarie 1 e 0. Questi due stati vengono rappresentati nei circuiti elettrici da valori discreti di una grandezza elettrica (di solito la grandezza utilizzata è la tensione), che può assumere quindi due valori differenti e sufficientemente diversificati, in modo da poter essere facilmente riconosciuti come rappresentativi dell'uno o dell'altro stato; si comprende quindi come gli stati del circuito possano anche essere identificati come stato alto o basso, oppure come conduzione o interdizione.

Come è noto, un dato, o il valore di una grandezza, può essere codificato in un sistema binario, cioè a due sole cifre, dette *bit* (da *binary digit*, *cifra binaria*), secondo il codice binario: ad esempio il numero 13 (in notazione decimale) può essere rappresentato in codice binario dalla sequenza 01101. Ciò permette di utilizzare i circuiti digitali, in cui il segnale può assumere solo due valori discreti a cui vengono rispettivamente associati i bit 1 e 0, per elaborare i dati, purché questi siano codificati opportunamente in codice binario.

Questa possibilità ha portato all'utilizzo dei circuiti digitali per l'elaborazione elettronica dei dati, e ha creato un settore applicativo, quello dei calcolatori, che ha avuto uno sviluppo eccezionale nell'ambito dell'industria elettronica.

Lo sviluppo delle tecnologie microelettroniche e l'impatto dei circuiti integrati sulle prestazioni e sul costo degli apparati elettronici hanno spinto la diffusione dei circuiti digitali anche in campi come quello delle comunicazioni, dei controlli e dell'elettronica di consumo, nei quali inizialmente venivano utilizzati circuiti analogici; anche in questi campi l'impiego dei circuiti digitali ha largamente sostituito in molte applicazioni l'elaborazione analogica dei segnali, principalmente a causa delle più elevate prestazioni, in termini di immunità ai rumori,

e ai ridotti costi di progettazione e di realizzazione presentati dai circuiti digitali. In questo caso l'informazione, ossia il segnale che deve essere elaborato, viene prima campionato e codificato in un numero binario mediante un circuito che effettua una conversione da analogico a digitale (*Convertitore analogico-digitale*, ADC), e successivamente elaborato mediante tecniche numeriche, ed infine riconvertito in segnale analogico mediante una ulteriore conversione da digitale ad analogico (*Convertitore digitale-analogico*, DAC). Un esempio generico di elaborazione di segnali rispettivamente mediante circuiti analogici o circuiti digitali è indicato in Figura 1.1.

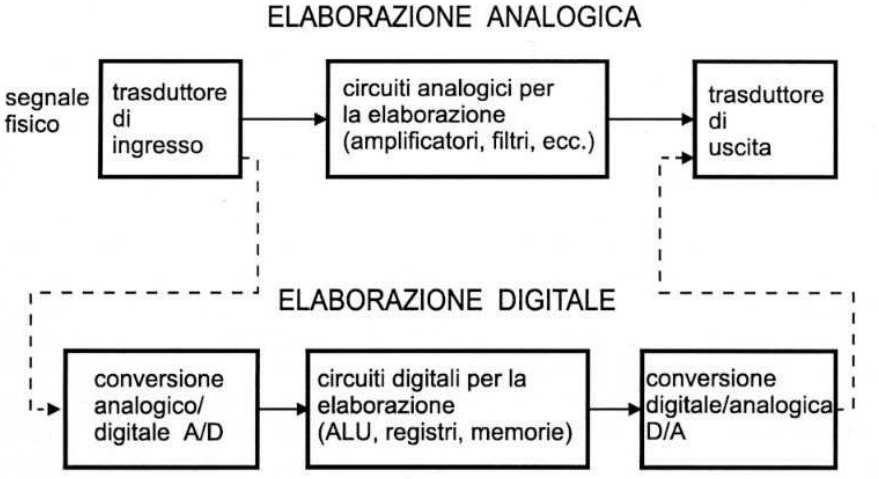


Figura 1.1 Elaborazione dei segnali mediante circuiti analogici o digitali

Una ulteriore ed importante possibilità dei circuiti digitali è quella di poter realizzare non solo delle operazioni *matematiche* tra le grandezze codificate, ma anche delle operazioni *logiche* tra le grandezze digitali, considerate ora come variabili logiche, in base alla logica Booleana che definisce le relazioni possibili fra variabili che possono assumere solo due stati. In altre parole i circuiti digitali possono effettuare delle relazioni logiche dell'algebra Booleana tra le variabili fornendo in uscita una grandezza (binaria) Y che è il risultato della relazione logica voluta tra più grandezze (binarie) A, B, ... N in ingresso ai circuiti stessi.

Questa fondamentale possibilità dei circuiti digitali (che più correttamente in tal caso sono detti circuiti *logici*) permette di realizzare sistemi complessi che possono effettuare scelte, valutare eventi, identificare situazioni, e che sono prepotentemente entrati in tutti i campi delle scienze applicate grazie alle loro sempre crescenti possibilità di elaborazione.

Lo studio e la progettazione di questi sistemi elettronici richiede, come in molti altri sistemi complessi, un'insieme di metodologie e di competenze che spaziano

nei diversi ambiti del settore dell'informazione e che debbono confluire nella fase di definizione e di progettazione del sistema stesso. Nell'ambito dello studio dei sistemi elettronici basati sui circuiti digitali, il campo dell'elettronica digitale riguarda essenzialmente la comprensione del funzionamento dei circuiti elettronici elementari, che costituiscono i blocchi base, ossia i tasselli elementari di cui sono costituiti i sistemi più complessi, e la loro progettazione, sia a livello di circuito elettronico, che come struttura integrata nel chip di silicio.

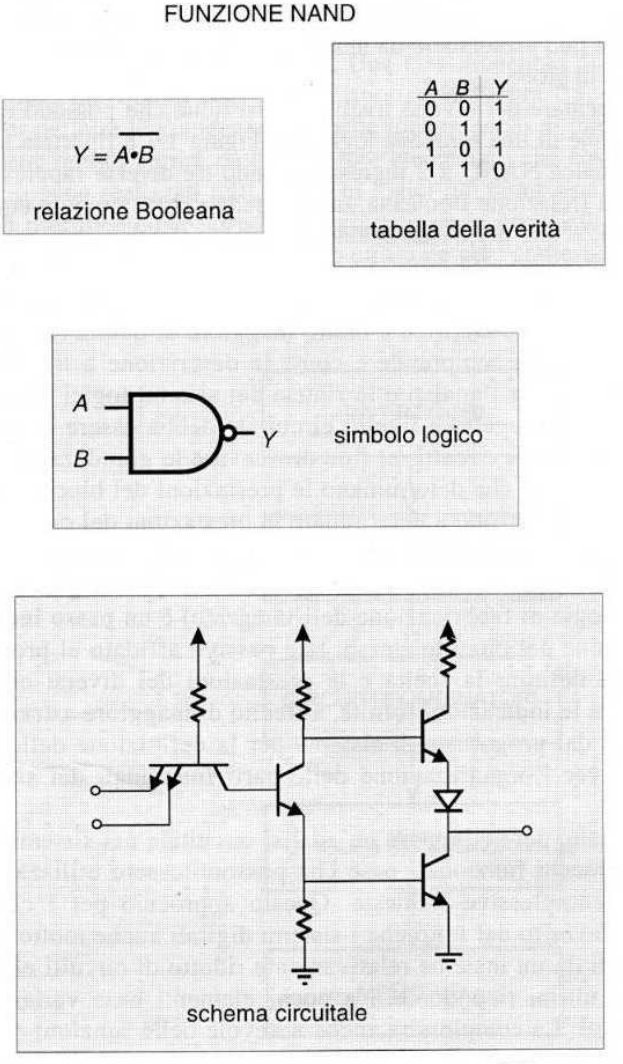


Figura 1.2 Rappresentazioni logiche e circuitali di una porta NAND

Da un punto di vista funzionale, cioè facendo riferimento alla funzione logica che tali circuiti debbono svolgere, e ricordando che questi circuiti trattano segnali con due soli livelli logici, la descrizione di un circuito logico può essere data ad un livello di astrazione più elevato facendo riferimento ai legami tra le variabili logiche applicate; tale descrizione è quella usualmente utilizzata nell'analisi delle reti logiche secondo metodologie tipiche delle discipline informatiche. Tuttavia nessun circuito digitale, per quanto ben realizzato, si comporta in maniera ideale: tutti i circuiti *digitali* sono in realtà circuiti in cui le grandezze variano in maniera *analogica* da un valore all'altro in un tempo finito, con ritardi rispetto ai segnali in ingresso e con livelli variabili rispetto a quelli ideali prefissati, per cui il loro studio e la loro progettazione non può prescindere da una descrizione e da un'analisi di tipo elettrico delle grandezze in gioco.

Come esempio elementare dei diversi livelli di astrazione che possono essere utilizzati per la descrizione di una funzione logica, in Figura 1.2 è riportata la descrizione di una porta logica NAND a 2 ingressi secondo tre diverse rappresentazioni logiche equivalenti (relazione Booleana, tabella della verità, simbolo equivalente di una rete logica), e una rappresentazione circuitale (circuito elettronico in tecnologia TTL).

Come si può vedere da questo pur semplice esempio, la complessità di descrizione di un circuito a livello elettrico è molto maggiore di quella richiesta a livello logico, per cui è facile comprendere come la descrizione a livello più astratto sia quella utilizzata per l'analisi e la sintesi dei sistemi logici più complessi, mentre l'analisi e il progetto a livello circuitale debba essere di norma applicato ai blocchi basilari dei circuiti, al fine di ricavare le grandezze elettriche sia statiche che dinamiche che determinano le prestazioni del blocco stesso e che, attraverso questo, concorrono a determinare le prestazioni del circuito da esso costituito. D'altra parte come si è detto prima, l'approccio circuitale (come quello di descrizione del tracciato topologico sul silicio e più in generale la definizione della tecnologia di fabbricazione dell'integrato) è un passo ineliminabile per la progettazione del circuito stesso; tale passo è affidato al progettista elettronico che, nel definire la scelta e le prestazioni dei diversi circuiti elettronici, deve recepire le indicazioni fornite, a livello di maggiore astrazione, dal progettista logico e dal progettista di sistema per la definizione delle funzioni logiche volute e per l'organizzazione delle parti funzionali del sistema finale.

È quindi fondamentale, per sviluppare un'analisi circuitale nei sistemi elettronici, identificare i blocchi funzionali base che possono essere utilizzati per realizzare le funzioni complessive richieste. Questo approccio per i circuiti logici è in certo modo favorito dal fatto che i sistemi digitali anche molto complessi risultano costituiti da un insieme relativamente ridotto di circuiti elettronici elementari, questi ultimi riconducibili a pochi elementi base variamente interconnessi ed utilizzati. La complessità anche notevole delle funzioni realizzate dai sistemi digitali è ottenuta attraverso l'iterazione di strutture regolari formate dai blocchi funzionali elementari che permettono di realizzare le diverse funzioni elettroniche e logiche.

Le funzioni fondamentali di un sistema logico complesso possono essere ricondotte a quelle di

a. elaborazioni numeriche dei dati (somma, sottrazione, moltiplicazione, ecc.);
b. operazioni logiche sulle variabili binarie (NAND, NOR, NOT, ecc.);
c. memorizzazione dei dati e delle variabili;
d. trasferimento alle interfacce ingresso/uscita.

Si verificherà nel corso dell'esposizione dei diversi circuiti logici, come tutte queste operazioni siano possibili in base a blocchi elettronici elementari chiamati *porte logiche*; queste ultime possono a loro volta considerarsi costituite a partire da circuiti elementari detti *invertitori*. Una classificazione di tipo gerarchico dei circuiti logici può quindi essere quella riportata in Figura 1.3.

Per definire le grandezze elettriche fondamentali di interesse nei circuiti logici, conviene quindi partire dall'invertitore elementare che costituisce l'elemento base con cui sono realizzati tutti i blocchi logici più complessi. A quest'ultimo faremo perciò riferimento per introdurre e giustificare le principali grandezze elettriche sia statiche che dinamiche che vengono utilizzate per caratterizzare le reti logiche.

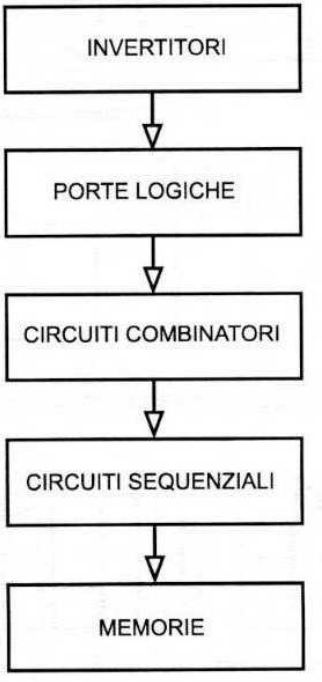


Figura 1.3 Suddivisione gerarchica dei circuiti logici

1.2 Discretizzazione dei segnali

La caratteristica fondamentale dei circuiti logici (digitali) è quella di operare con segnali che idealmente dovrebbero assumere solo due possibili valori. Ciò in pratica non può essere garantito, sia a causa delle inevitabili variazioni delle grandezze elettriche in gioco (tensioni di alimentazione dei circuiti, variazioni delle funzioni di trasferimento delle reti elettriche), che a causa di disturbi elettrici che si sovrappongono ai segnali stessi in maniera aleatoria, facendo variare questi ultimi rispetto ai loro valori nominali.

Una caratteristica molto importante dei circuiti digitali è quella di ripristinare i livelli logici dei segnali che attraversano i circuiti stessi; questa caratteristica permette di tollerare anche disturbi relativamente elevati, perché questi ultimi, come vedremo in seguito, vengono per così dire filtrati ad ogni successiva elaborazione se sono inferiori a determinati valori, in modo che il rapporto segnale/rumore rimane inalterato o addirittura migliorato alla fine della elaborazione.

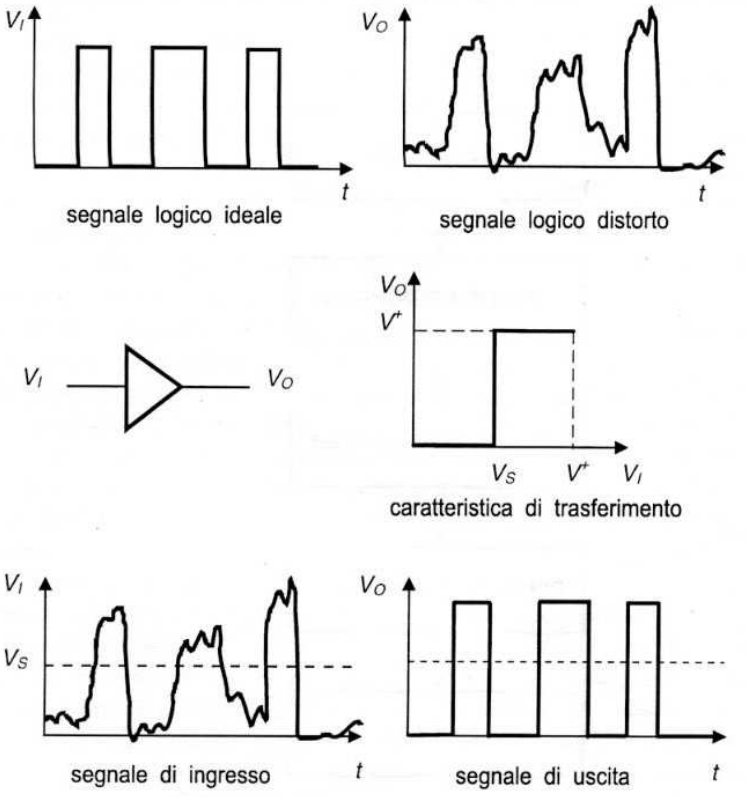


Figura 1.4 Ripristino dei livelli logici di un segnale con un circuito a soglia ideale

Questa caratteristica di relativa immunità ai disturbi per i segnali che attraversano circuiti digitali ha portato a scegliere quando possibile la conversione analogico/digitale dell'informazione e la sua successiva elaborazione in forma digitale; d'altra parte una elevata immunità ai disturbi è essenziale nel campo delle applicazioni logiche in cui anche pochi segnali spuri potrebbero alterare in maniera inaccettabile le operazioni previste.

L'operazione di ripristino dei livelli logici 1 e 0 in principio viene effettuata utilizzando un circuito con funzione di trasferimento ideale a soglia, come quella riportata in Figura 1.4. In questo circuito un segnale di ingresso distorto V_I viene ripristinato ai livelli logici primitivi, in quanto i livelli inferiori alla soglia V_S vengono riportati al valore basso (0 nell'esempio) e quelli superiori a V_S al valore alto (V^+).

1.3 L'invertitore ideale

L'invertitore ideale è un circuito che abbina alla caratteristica di trasferimento a soglia, necessaria per la rigenerazione dei segnali logici, la funzione di inversione dei livelli logici dei segnali. In Figura 1.5 è indicato lo schema di principio dell'invertitore, con la sua funzione di trasferimento ideale. L'invertitore è realizzato in via di principio mediante un interruttore ideale che viene pilotato dalla variabile di ingresso V_I, in maniera tale che esso è chiuso se V_I è superiore al livello di soglia $V^+/2$, mentre è aperto se V_I è inferiore a $V^+/2$.

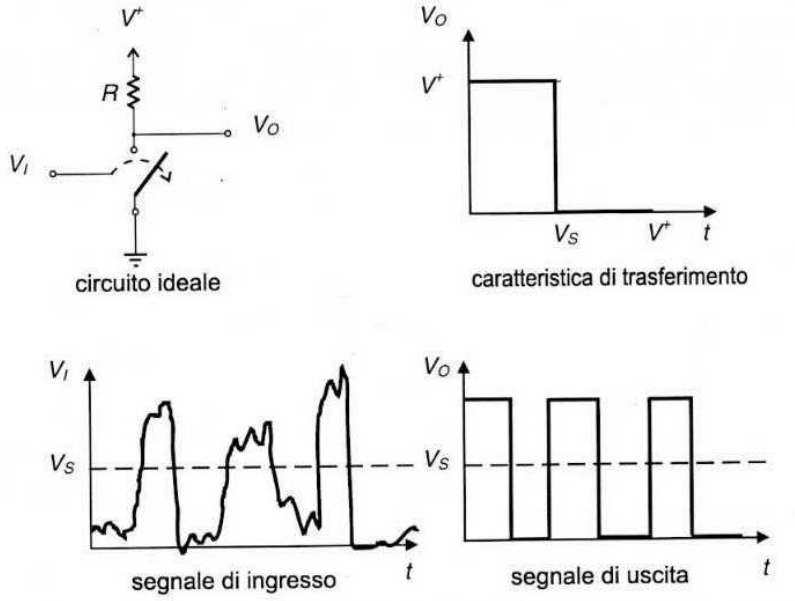


Figura 1.5 Ripristino dei livelli logici con l'invertitore ideale

Quando l'interruttore è aperto ($V_I < V^+/2$) la tensione di uscita V_O è quella dell'alimentazione perché non circola corrente nella resistenza R (si assume che i circuiti connessi eventualmente al terminale di uscita non assorbano corrente da quest'ultimo); quando l'interruttore è chiuso ($V_I > V^+/2$) la tensione V_O è nulla (assumendo che l'interruttore sia un cortocircuito ideale quando è chiuso) e tutta la tensione V^+ cade sulla resistenza R. I segnali in ingresso con valori inferiori alla soglia $V^+/2$ vengono quindi riportati in uscita al livello logico alto (V^+ nel nostro caso), mentre quelli superiori alla soglia vengono riportati al livello logico basso (0). La funzione logica realizzata dall'invertitore è quindi la funzione Booleana NOT descritta dalla relazione $Y = \overline{A}$, oppure $Y = \text{NOT } A$.

Nei circuiti elettronici la funzione dell'*interruttore comandato* è realizzata dai *dispositivi attivi* a tre terminali, quali il transistore MOS o il transistore bipolare, che verranno richiamati nei Capitoli 3 e 6. Rimandando ad uno studio successivo il comportamento degli invertitori realizzati con tali dispositivi, si farà ora riferimento ad invertitori generici con una caratteristica di trasferimento più realistica di quella idealizzata di Figura 1.5, in modo da poter definire in via generale i principali parametri che caratterizzano il comportamento di tali circuiti, comunque realizzati.

1.4 L'invertitore reale

1.4.1 Caratteristica di trasferimento e margini di rumore

La caratteristica di trasferimento di un invertitore reale (intesa come il legame grafico tra la tensione di uscita V_O e quella di ingresso V_I) non presenta una discontinuità per il valore della soglia logica V_S, ma assume in generale un andamento caratterizzato da due tratti a debole pendenza interconnessi da un tratto ad elevata pendenza, corrispondente al passaggio del circuito dallo stato di uscita alta (bassa) a quello di uscita bassa (alta), come è genericamente riportato in Figura 1.6.

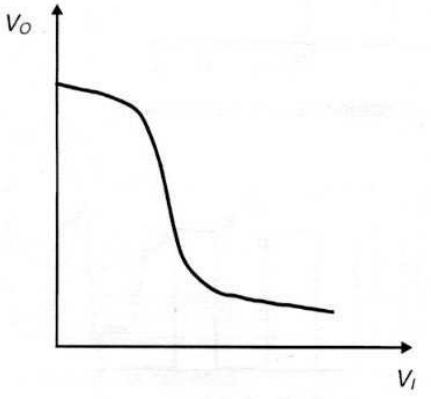


Figura 1.6 Caratteristica di trasferimento di un invertitore reale

Una caratteristica di questo tipo presenta naturalmente una minore capacità di ripristinare i livelli logici del segnale nel passaggio attraverso l'invertitore stesso, e quindi è fondamentale definire tale capacità in termini quantitativi in base alla caratteristica di trasferimento, in modo da poter confrontare le prestazioni di invertitori reali costruiti con differenti soluzioni circuitali.

Il primo problema che si pone in un invertitore generico descritto sinteticamente dalla sua funzione di trasferimento è quello della identificazione dei livelli logici alto e basso, in quanto per tale funzione ad ogni valore della grandezza di ingresso corrisponde un diverso valore di quella di uscita. Il problema viene risolto considerando che nei circuiti digitali i segnali di ingresso non sono forniti da generatori indipendenti, ma vengono di norma forniti dalle uscite di altri circuiti digitali (dello stesso tipo e tecnologia) connessi a monte del circuito in esame.

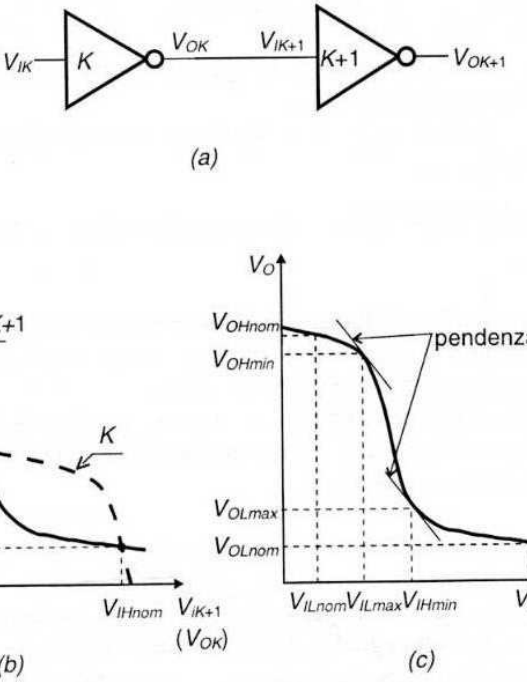


Figura 1.7 a) Catena di invertitori; b) costruzione grafica per la determinazione dei valori nominali di ingresso e di uscita; c) grandezze caratteristiche della funzione di trasferimento

Facendo quindi riferimento ad una serie di invertitori uguali connessi in cascata come in Figura 1.7a, in cui il precedente fornisce il segnale di ingresso al successivo, e ad una generica funzione di trasferimento uguale per ciascun invertitore, si

possono definire i valori *nominali* dei livelli logici basso ed alto, rispettivamente in ingresso e in uscita, in assenza di disturbi eventualmente sovrapposti ai segnali stessi.

Dallo schema di Figura 1.7a si può notare come la grandezza logica V_{OK} dell'invertitore K coincida con quella d'ingresso V_{IK+1} dello stadio $K+1$; a sua volta il livello logico in uscita dallo stadio $K+1$ corrisponde a quello di ingresso dello stadio K in quanto $V_{K+1} = \overline{V_{OK}} = \overline{\overline{V_{IK}}} = V_{IK}$. Questi valori possono essere ottenuti dalla costruzione grafica di Figura 1.7b, ribaltando e ruotando di 90° la caratteristica di trasferimento dello stadio K in modo che l'asse delle tensioni in ingresso allo stadio $K+1$ (asse delle ascisse) coincida con quello delle tensioni in uscita dello stadio K, e quello delle tensioni di uscita dello stadio $K+1$ (che è anche l'uscita dello stadio $K-1$) corrisponda all'asse delle tensioni di ingresso dello stadio K; questa costruzione corrisponde alla condizione imposta dalla connessione in cascata degli invertitori $K-1$, K e $K+1$.

I *valori nominali* di V_I e V_O sono quindi quelli corrispondenti alle due possibili coppie di valori determinate dalle due intersezioni:

V_{OHnom} il valore della tensione di uscita alta dell'invertitore corrispondente all'ingresso basso nominale V_{ILnom};

V_{OLnom} il valore della tensione di uscita bassa, corrispondente all'ingresso alto nominale V_{IHnom}.

Oltre a questi due valori, si definiscono altri due valori caratteristici (vedi Figura 1.7c), definiti come:

V_{ILmax} il massimo valore della tensione di ingresso per il quale l'uscita è ancora al valore alto V_{OHmin}, corrispondente al punto in cui la caratteristica presenta una tangente con inclinazione pari a -1;

V_{IHmin} il minimo valore della tensione di ingresso per il quale l'uscita è ancora al valore basso V_{OHmax}, corrispondente ancora al punto in cui la tangente alla curva ha inclinazione -1.

Per questi ultimi valori occorre giustificare il perché si sceglie la condizione di tangente pari a -1 per la definizione dei valori di ingresso V_{ILmax} e V_{IHmin}. Ciò si spiega se si ricorda che la tangente alla curva di trasferimento rappresenta l'amplificazione (a piccoli segnali) che un disturbo di piccola entità subisce nel passare attraverso l'invertitore stesso. I punti corrispondenti ad una pendenza -1 della curva separano quindi per così dire le regioni della caratteristica con guadagno minore di uno (attenuazione) da quelle con guadagno maggiore di uno (amplificazione). Quindi un eventuale disturbo sovrapposto alla tensione di ingresso verrà attenuato (e quindi soppresso da una catena di invertitori di uguali caratteristiche) se il segnale di ingresso è inferiore (superiore) a V_{ILmax} (V_{IHmin}), mentre verrà al contrario amplificato per segnali compresi tra V_{ILmax} e V_{IHmin}, portando gli invertitori successivi in stati logici diversi da quelli previsti, in dipendenza

del numero degli invertitori (pari o dispari) e del segno del disturbo (positivo o negativo).

In Figura 1.8 è riportato il comportamento di una cascata di tre invertitori in presenza di un segnale logico V_I su cui è sovrapposto un disturbo D simulato da un'oscillazione di ampiezza decrescente; se la tensione di ingresso (segnale + disturbo) supera il valore V_{ILmax} l'uscita a valle dei tre invertitori cambia di stato rispetto a quella prevista per il segnale logico, mentre per livelli inferiori del disturbo tali che $V_I + D < V_{ILmax}$ ($>V_{ILmin}$), il disturbo viene eliminato dall'uscita e il livello logico è completamente ripristinato.

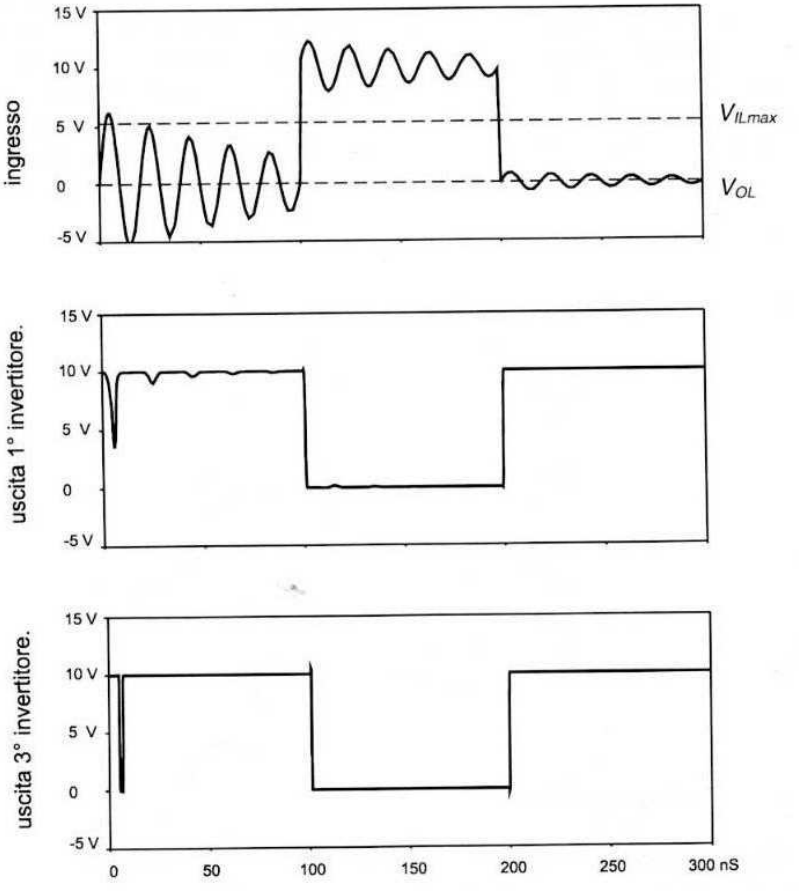


Figura 1.8 Effetto di un disturbo sul segnale logico in uscita da una cascata di invertitori

Da questo esempio si comprende come le differenze tra i valori massimi (minimi) di ingresso ammissibili per un'uscita alta (bassa), e quelli nominali bassi (alti) vengano indicate come *margini di rumore* dell'invertitore, in quanto queste

differenze indicano il massimo livello del disturbo ammissibile che, sommandosi in modo algebrico al segnale, fornisce un'uscita ancora riconducibile allo stato logico previsto. In particolare si definisce (vedi Figura 1.9):

margine di rumore per ingresso alto: $NM_H = V_{OHnom} - V_{IHmin}$;

margine di rumore per ingresso basso: $NM_L = V_{ILmax} - V_{OLnom}$.

In realtà, come si vede dal diagramma delle tensioni di ingresso e di uscita della serie di invertitori di Figura 1.9, sarebbe più corretto definire i margini di rumore nel *caso peggiore*, con riferimento ai valori *minimi* di V_{OH} e V_{IH} e a quelli *massimi* di V_{IL} e V_{OL}, ma per semplicità di definizione e di valutazione ci si riferisce alle grandezze su indicate, che nel seguito saranno indicate più sinteticamente come: V_{OH} (corrispondente a V_{OHnom}), V_{IH} (corrispondente a V_{IHmin}), V_{IL} (corrispondente a V_{ILmax}) e V_{OL} (corrispondente a V_{OLnom}). Nel caso di margini di rumore diversi per i due stati logici, le prestazioni dell'invertitore sono determinate dal più piccolo dei due.

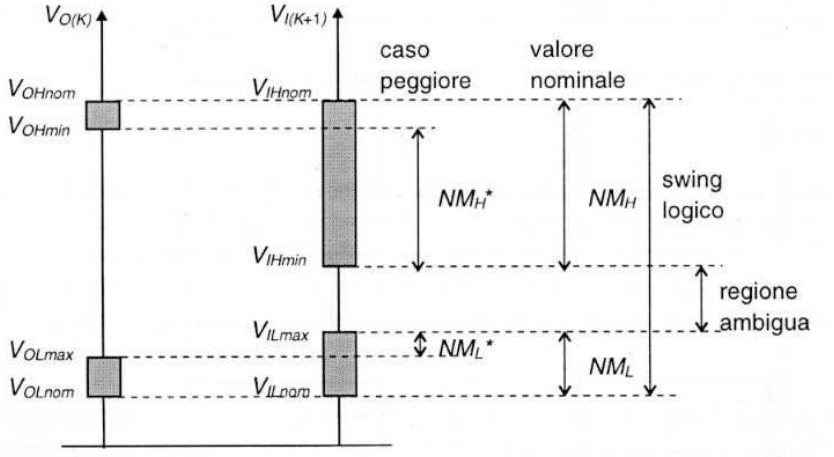


Figura 1.9 Definizione dei margini di rumore

Dal diagramma di Figura 1.9 risulta immediato che, per un buon funzionamento dell'invertitore, occorre che sia $V_{OH} > V_{IH}$ e $V_{OL} < V_{IL}$; quanto più grande è il campo delle variazioni $V_{OH} - V_{OL}$ rispetto a quello delle variazioni $V_{IH} - V_{IL}$, tanto maggiore sarà la capacità a tollerare disturbi anche relativamente ampi. Il caso di massima tolleranza ai disturbi è quello dell'invertitore ideale che, in base alla funzione di trasferimento di Figura 1.5, presenta il livello V_{OH} pari alla tensione di alimentazione V^+, il livello V_{OL} pari a 0, $V_{IH} = V_{IL} = V^+/2$, e quindi margini di rumore uguali e pari anche essi a $V^+/2$. Dalla Figura 1.9 si ricava anche la definizione di *escursione*

(*swing*) *logica*, data da $V_{OHnom} - V_{OLnom}$ e coincidente sia in ingresso che in uscita, essendo le grandezze nominali di ingresso e di uscita semplicemente invertite, e la *regione ambigua*, definita come il campo di segnali di ingresso che non forniscono lo stato logico previsto in uscita. Infine viene definita *soglia logica* V_{SL} la tensione che si ottiene identificando sulla caratteristica di trasferimento il valore della tensione di ingresso V_I uguale a quello V_O di uscita; quest'ultima è in effetti il livello di tensione ideale che separa i valori bassi da quelli alti, ed è definita dalla intercetta della curva di trasferimento con una retta con pendenza 45° passante per l'origine.

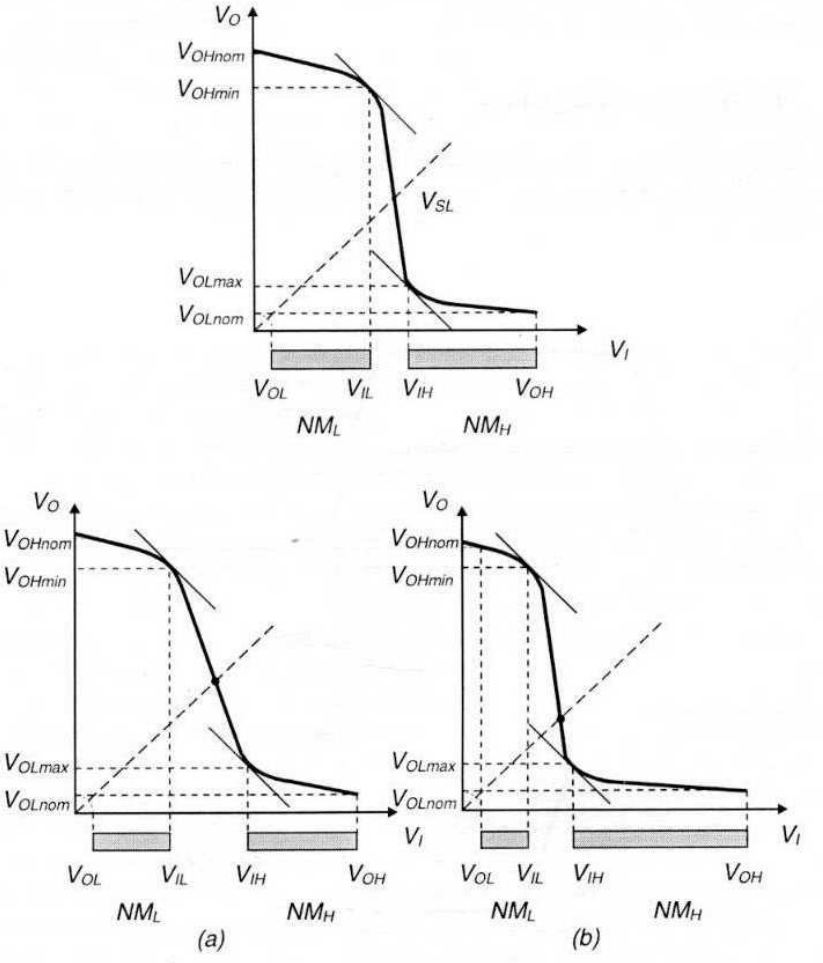


Figura 1.10 Influenza della forma della caratteristica di trasferimento sui margini di rumore: a) variazione della pendenza; b) traslazione della caratteristica

La forma della caratteristica di trasferimento gioca, come è prevedibile, in maniera significativa sui valori dei margini di rumore. In particolare una riduzione della pendenza della caratteristica tra i punti con tangente unitaria aumenta la regione ambigua e quindi riduce i margini di rumore sia nello stato logico basso che in quello alto (vedi Figura 1.10a). Anche una traslazione della soglia della caratteristica di trasferimento rispetto al valore ideale $V_{SL} = V^+/2$ comporta un aumento di uno dei margini di rumore a scapito dell'altro, quindi ancora una degradazione delle prestazioni logiche dell'invertitore (Figura 1.10b); da quest'ultima osservazione discende che un parametro di merito dell'invertitore è quello di avere una caratteristica di trasferimento con una soglia logica quanto più possibile vicina alla metà dell'escursione logica; in questo caso i margini di rumore NM_H e NM_L risultano uguali e quindi la reiezione ai disturbi è la più elevata, a parità di regione ambigua $V_{IH} - V_{IL}$.

1.4.2 Tempo di propagazione

Una caratteristica importante per l'invertitore e, più in generale, per i circuiti digitali è quella della rapidità della risposta ai segnali logici che si presentano all'ingresso.

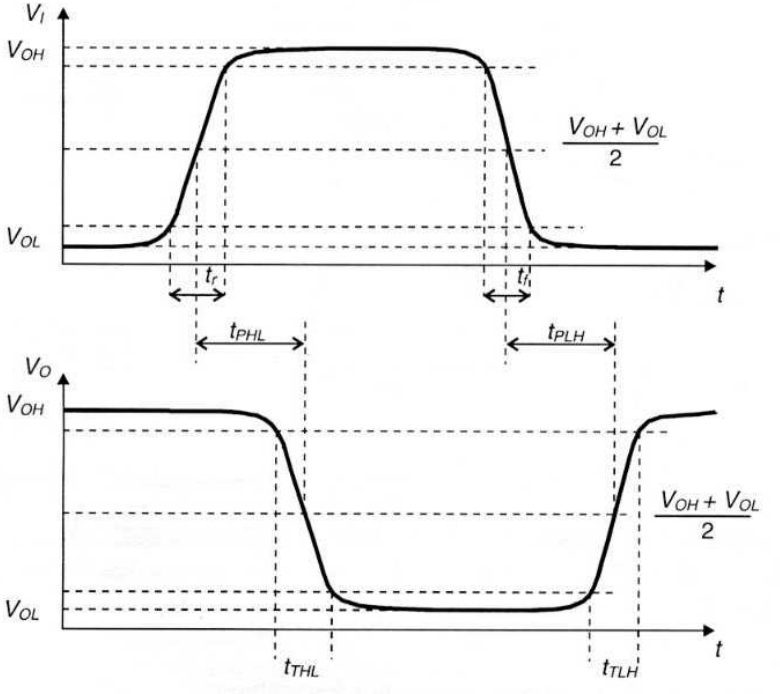


Figura 1.11 Definizione delle grandezze dinamiche di un invertitore

In Figura 1.11 sono rappresentate le forme d'onda di un segnale in ingresso che passa dallo zero logico all'uno logico (arrivo di un bit 1), e quella corrispondente all'uscita dell'invertitore. Il segnale di ingresso è di solito fornito da un circuito collegato all'ingresso dell'invertitore e quindi sarà anch'esso rallentato nel fronte di salita e di discesa; si definiscono tempi di salita t_r e di discesa t_f gli intervalli di tempo corrispondenti rispettivamente al passaggio del segnale dal 10% al 90% della escursione logica $V_{OH} - V_{OL}$ e al passaggio dal 90% al 10% della stessa escursione.

Nel segnale di uscita si identificano come parametri dinamici i *tempi di transizione* t_{THL} e t_{TLH} corrispondenti al passaggio dal 90% al 10% dell'uscita (transizione alto-basso) e al passaggio dal 10% al 90% dell'uscita (transizione basso-alto), mentre si definiscono *tempi di propagazione* t_{PHL} e t_{PLH} gli intervalli di tempo corrispondenti al ritardo tra il segnale di ingresso e l'uscita corrispondente, nel passaggio per il valore del 50% della escursione logica: $(V_{OH} + V_{OL})/2$.

Questi ultimi due parametri identificano un *ritardo* tra la presenza del segnale logico in ingresso e la sua elaborazione (in questo caso la sua negazione logica) in uscita, ed infatti viene chiamato *ritardo di propagazione* t_P il valore medio tra questi due ritardi, definito come:

$$t_P = \frac{t_{PHL} + t_{PLH}}{2} \tag{1.1}$$

Nel caso in cui i due tempi di propagazione t_{PHL} e t_{PHL} siano uguali, il ritardo $t_P = t_{PHL} = t_{PLH}$ corrisponde proprio alla traslazione temporale con cui si presenta in uscita il segnale rispetto all'ingresso.

Il ritardo di propagazione t_P è la grandezza dinamica più importante per i circuiti logici, non solo perché in effetti definisce la minima durata che deve avere il segnale di ingresso per provocare una variazione logica corrispondente in uscita, e quindi implicitamente vincola la massima frequenza con cui i bit si possono susseguire in ingresso, ma anche perché il ritardo del segnale logico elaborato in uscita può portare a delle operazioni logiche non desiderate nei circuiti a valle, dove la simultaneità dei segnali provenienti da percorsi differenti per le operazioni logiche successive può essere essenziale per una corretta esecuzione della funzione logica; si verificherà questo aspetto in particolare nell'esame dei circuiti sequenziali.

Il ritardo di propagazione ha la proprietà di essere una grandezza additiva per i circuiti connessi in cascata, nel senso che nel passaggio attraverso una serie di n invertitori (o più in generale di porte logiche) il ritardo di propagazione complessivo è dato dalla somma dei singoli ritardi di propagazione di ogni porta. Questa proprietà è dimostrata schematicamente nella Figura 1.12 in cui si considera una serie di invertitori connessi in cascata, ognuno definito da un ritardo t_{Pi}, e si riportano i segnali di ingresso ad ogni invertitore; si vede come, all'ingresso dell'invertitore $n + 1$, il ritardo di propagazione complessivo sia pari alla somma dei ritardi di propagazione degli stadi a monte di quello considerato, e cioè: $t_{PT} = \sum_n t_{Pi}$.

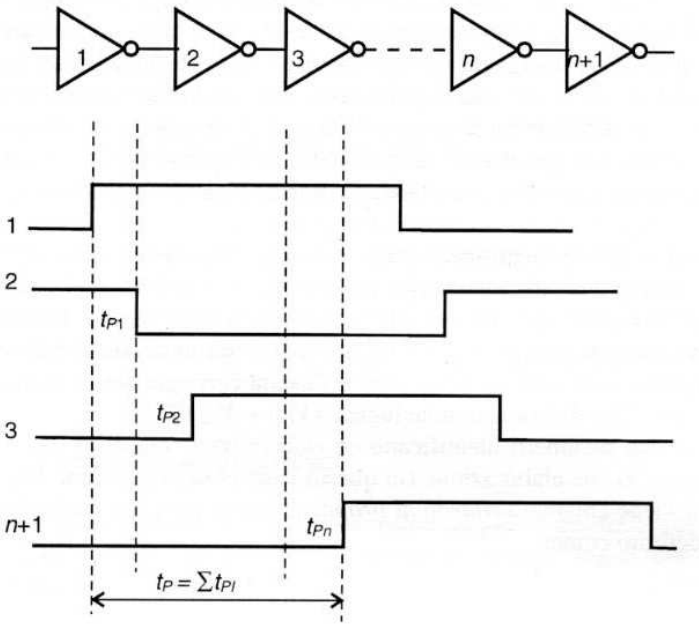


Figura 1.12 Ritardo di propagazione totale di una cascata di invertitori (per semplicità si sono assunti tempi di transizione nulli per le forme d'onda)

La proprietà additiva del ritardo di propagazione viene usualmente sfruttata per una misura sperimentale del ritardo di propagazione stesso. Infatti, poiché con le tecnologie attuali le porte logiche elementari nei circuiti integrati hanno ritardi di propagazione ben inferiori al ns, risulta molto difficile una misura diretta di questa grandezza sulla singola porta, e in ogni caso la misura è fortemente approssimata sia a causa della limitazione della strumentazione utilizzata, sia del disturbo non trascurabile introdotto dalla misura stessa sulla dinamica delle porte.

In questi casi si ricorre ad una connessione in cascata di n invertitori uguali (con n numero dispari dell'ordine di qualche decina), realizzati nello stesso circuito integrato che si vuole caratterizzare, riportando il segnale dell'ultimo stadio direttamente all'ingresso del primo. Tale configurazione si chiama *oscillatore ad anello*, in quanto è facile verificare (vedi Figura 1.13) che l'uscita di questo circuito oscilla periodicamente tra il valore logico alto e quello basso; in base al ritardo di propagazione complessivo che è nt_P (dove t_P è il ritardo di propagazione di ogni invertitore assunto uguale per ipotesi per tutti gli invertitori), si ricava una semplice relazione tra il periodo T della forma d'onda (facilmente misurabile con un frequenzimetro) e il ritardo di propagazione incognito del singolo invertitore data da: $t_P = T/2n$; in tal caso si trasferisce la misura di t_P in quella, più agevole, di T.

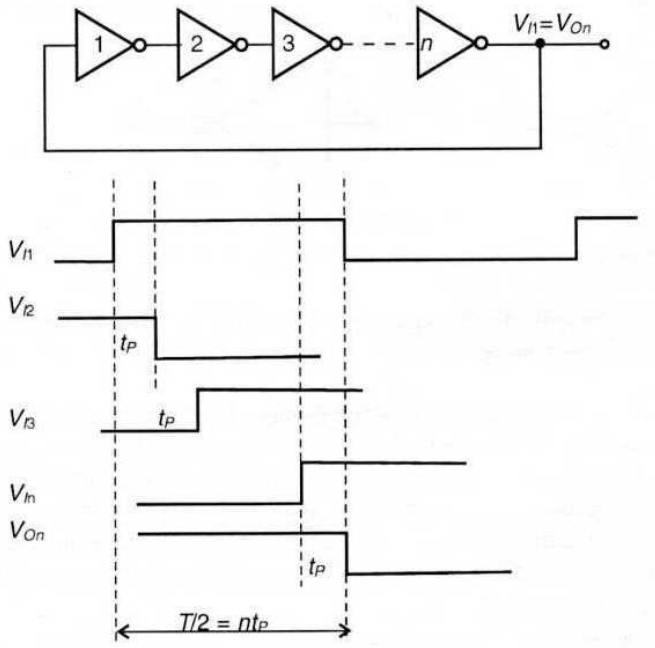


Figura 1.13 Oscillatore ad anello per la misura del tempo di propagazione t_P

1.4.3 Potenza dissipata

Per *potenza dissipata* di un invertitore (e più in generale di una porta logica) si intende la potenza fornita dall'alimentazione che viene assorbita dalla porta logica nel suo funzionamento ed è quindi definita come $P_D = V^+I$, dove V^+ è la tensione di alimentazione e I è la corrente assorbita dalla porta durante il suo funzionamento. La dissipazione di potenza ha due componenti: una statica ed una dinamica.

Il *consumo di potenza statico* P_S corrisponde alla potenza assorbita dal circuito quando questo non cambia stato logico; in generale la potenza V^+I_H assorbita quando l'uscita è nello stato logico alto sarà differente da quella V^+I_L assorbita nello stato logico basso, e si definisce potenza media dissipata P_{Sav} il valore medio tra le due:

$$P_{Sav} = V^+ \frac{I_H + I_L}{2} \tag{1.2}$$

assumendo che il circuito si trovi mediamente per metà del tempo nello stato alto e per metà in quello basso.

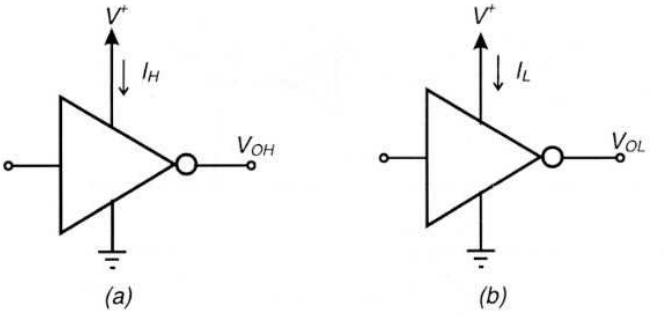


Figura 1.14 Potenza dissipata da un invertitore in condizioni statiche: a) uscita logica alta; b) uscita logica bassa

Il *consumo di potenza dinamico* avviene durante le transizioni da uno stato logico all'altro, e dipende dalla corrente assorbita dal circuito durante i tempi t_{TLH} e t_{THL}; la corrente assorbita, a sua volta, è costituita da una prima componente assorbita nell'invertitore stesso per cambiare stato, e da una seconda componente necessaria in ogni caso per caricare e scaricare la capacità C_L che costituisce per ogni invertitore il carico associato in uscita. Queste due componenti sono schematicamente indicate nella Figura 1.15.

La seconda componente della dissipazione di potenza dinamica è usualmente la più rilevante ed è facilmente valutabile assumendo per semplicità che la tensione di uscita dell'invertitore nello stato logico basso sia nulla (capacità scarica).

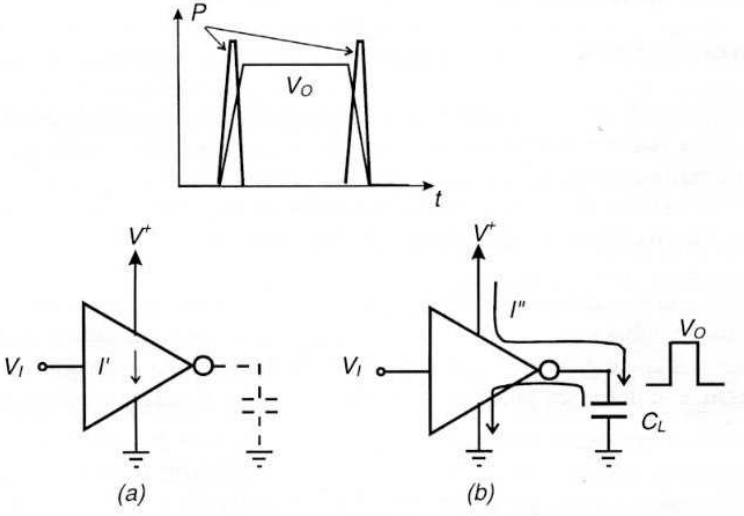


Figura 1.15 Dissipazione di potenza dinamica: a) contributo legato alla corrente I' assorbita dall'invertitore per cambiare stato; b) contributo relativo alla corrente I'' di carica della capacità C_L

Indicando con V^+ la tensione di alimentazione, l'energia assorbita nel passaggio dallo stato basso a quello alto vale:

$$E_{AH} = V^+ \int i_C dt = V^+ Q_C = V^{+2} C_L \tag{1.3}$$

Questa energia viene per metà dissipata nell'invertitore e per metà immagazzinata nella capacità, essendo quest'ultima, con la capacità carica a V^+, pari a $1/2V^{+2}C_L$. Nel passaggio dallo stato alto a quello basso la capacità si scarica (verso massa) e perde l'energia immagazzinata, per cui l'energia totale persa nelle due transizioni sarà proprio E_{AH}, e la potenza dissipata, assumendo che l'invertitore effettui f passaggi dallo stato basso all'alto e viceversa per secondo (ossia che la frequenza di ripetizione dei bit in ingresso sia f), vale:

$$P_D = f \cdot E_{AH} = f \cdot C_L \cdot V^{+2} \tag{1.4}$$

La dissipazione di potenza di una porta logica assume un ruolo rilevante non solo perché definisce la richiesta di potenza che l'alimentatore deve fornire all'insieme (di solito rilevante) delle porte logiche, ma anche perché, essendo questi circuiti realizzati nella loro totalità, come vedremo in seguito, in forma *integrata*, cioè realizzando l'insieme dei circuiti in un piccolo tassello (*chip*) di silicio con opportuna tecnologia, la dissipazione di potenza delle singole porte elementari determina in ultima analisi il numero massimo di circuiti che possono essere realizzati in uno stesso chip. Quest'ultimo infatti può dissipare una potenza dell'ordine di qualche watt, in dipendenza del tipo di contenitore utilizzato, per cui il numero di porte logiche elementari che il chip può contenere può essere limitato dalla potenza dissipata dalla porta invece che dall'area occupata dalla porta stessa; ad esempio in un circuito integrato che può dissipare una potenza massima di 1 watt, il numero di porte elementari può superare il migliaio (Integrazione a Larga Scala, LSI) solo se le porte elementari dissipano meno di un mW ciascuna.

1.4.4 Prodotto ritardo-potenza dissipata

Le due grandezze caratteristiche precedentemente introdotte sono tra le più rilevanti per le prestazioni dei circuiti logici, perché permettono di valutare il livello di integrazione realizzabile e la massima velocità di operazione. In generale il miglioramento di una di queste due grandezze va a scapito dell'altra, per cui una riduzione della potenza dissipata comporta un aumento del tempo di propagazione e viceversa, ponendo così al progettista un problema di ottimizzazione. Vedremo meglio questo aspetto nell'analisi dei circuiti logici realizzati con differenti tecnologie, ma è possibile una verifica in via semplificata, con riferimento al circuito idealizzato di Figura 1.5 per l'invertitore elementare, basato su un interruttore controllato dal segnale di ingresso V_I (interruttore che idealizza il dispositivo attivo), caricato in

uscita da una capacità C_L che simula l'effetto dei circuiti a valle dell'invertitore stesso (vedi Figura 1.16); si assume inoltre per l'interruttore un comportamento ideale con tempi infinitesimi di apertura e chiusura e una caduta di tensione nulla in condizione di cortocircuito.

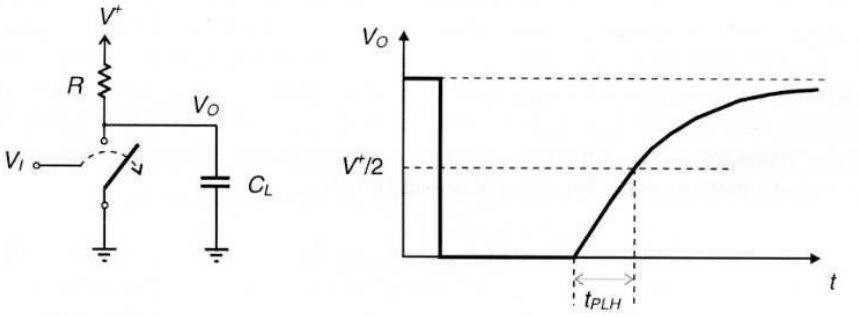


Figura 1.16 Ritardo di propagazione per l'invertitore idealizzato

La potenza dissipata statica in tal caso è solo quella assorbita nello stato di uscita bassa (nello stato di uscita alta non circola corrente nell'invertitore) e vale, in base all'Equazione (1.2):

$$P_{Sav} = \frac{V^{+2}}{2R} \tag{1.5}$$

Il ritardo di propagazione dell'invertitore, considerando un segnale di ingresso idealizzato con tempi di salita e discesa nulli, dipenderà solo dal tempo che impiega la capacità C_L a caricarsi attraverso la resistenza R, perché la scarica di C_L avviene in tempo nullo attraverso l'interruttore chiuso in cortocircuito. Il tempo di propagazione sarà quindi proporzionale alla costante di tempo di carica della capacità RC_L, e sarà dato da:

$$t_P = \frac{t_{PLH} + t_{PHL}}{2} = \frac{t_{PLH}}{2} = \frac{0.69 \cdot RC_L}{2} = 0.34 \cdot RC_L \tag{1.6}$$

e quindi il prodotto $t_P \cdot P_D$, indicato come $D \cdot P$, è dato da:

$$D \cdot P = 0.17 \cdot V^{+2} C_L \tag{1.7}$$

e risulta dipendente solo dalla tensione di alimentazione e dalla capacità di carico, ma non dalle caratteristiche dell'invertitore (in questo caso identificate dal valore della resistenza di carico R). Il prodotto $D \cdot P$ identifica quindi un *fattore di merito*

dell'invertitore, nel senso che più piccolo tale valore è, maggiore sarà la velocità di operazione a parità di potenza dissipata, o minore sarà il consumo a parità di velocità operativa.

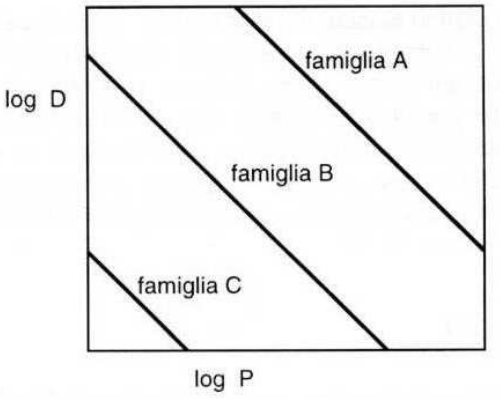


Figura 1.17 Curve ritardo-potenza dissipata per diverse famiglie logiche

È d'uso confrontare le diverse famiglie logiche realizzate con diverse tecnologie in termini del prodotto $D \cdot P$ come in Figura 1.17, riportando in scala log-log i valori del ritardo su un asse e quelli della potenza dissipata sull'altro asse. In tal caso le curve corrispondenti ad uguale valore del prodotto $D \cdot P$ sono delle rette inclinate di 45° rispetto agli assi; i valori del prodotto (che ha le dimensioni di un'energia) vengono riportati in picojoule, se le potenze vengono espresse in milliwatt e i ritardi in nanosecondi.

1.4.5 Fan-in e fan-out

Ogni invertitore (e più in generale ogni porta logica) può pilotare più circuiti logici, ed occorre definire la capacità dell'invertitore di pilotare correttamente questi circuiti. Viene definito *fan-out* il massimo numero di porte logiche (uguali a quella considerata), che possono essere connesse in uscita ad una data porta (in questo caso l'invertitore), mantenendo la degradazione del segnale di uscita in limiti accettabili. La definizione di limite accettabile fa riferimento alle specifiche fornite per la porta in esame; per quanto riguarda le specifiche statiche, occorre che il segnale di uscita sia ancora riconosciuto come livello logico non ambiguo e che quindi se basso non sia superiore a V_{OLmax} e se alto non sia inferiore a V_{OHmin}; per quanto riguarda le caratteristiche dinamiche, è necessario che il ritardo di propagazione non superi un valore prefissato dal costruttore come massimo. Si vedrà come l'inserzione di qualsiasi porta logica in uscita introduce una capacità di carico C_L aggiuntiva e quindi degrada il ritardo di propagazione; questo è il caso in particolare per le porte logiche realizzate in tecnologia MOS, come vedremo più avanti. Per alcune famiglie logiche realizzate in tecnologia bipolare, l'inserzione di porte in

uscita della porta in esame introduce sul terminale di uscita di questa anche un carico ohmico che modifica la caratteristica di trasferimento e i margini di rumore; in tal caso il fan-out è essenzialmente determinato dalle considerazioni su parametri statici.

Il *fan-in* definisce dualmente il massimo numero di porte logiche che il circuito in esame può accettare in ingresso prima che il segnale in uscita si degradi oltre le specifiche ammesse. Questo parametro non si applica per gli invertitori, che hanno un solo ingresso, ma ha significato per porte logiche a più ingressi, come ad esempio porte NAND o NOR a più ingressi, e va tenuto in conto in fase di progettazione delle porte e non in fase della loro utilizzazione, in quanto in questo caso il numero di ingressi è già definito. Vedremo che questo parametro usualmente pone un limite al numero di ingressi realizzabili per le porte NAND.

1.5 Porte logiche elementari

Le caratteristiche su esposte si applicano non solo agli invertitori, ma anche alle *porte logiche* che, come si è detto, possono essere realizzate a partire dall'invertitore elementare. Si definiscono porte logiche i circuiti digitali che implementano una particolare funzione di una, due o più variabili dell'algebra Booleana.

Una prima funzione di una variabile è la funzione $Y = NOT\ A$, e il circuito che la realizza è appunto l'invertitore, che è quindi una porta logica ad un ingresso in quanto opera su una sola variabile. Le altre funzioni possono legare due o più variabili e quindi il numero di ingressi della porta logica definisce anche il numero di variabili considerato nella relazione Booleana implementata.

Nella Tabella 1.1 sono riportate le funzioni Booleane fondamentali, con riferimento al caso di due variabili A e B, e in Figura 1.18 i simboli logici e le tabelle della verità delle porte logiche corrispondenti.

Tabella 1.1 Funzioni Booleane fondamentali

Funzione	*Espressione*
NOT	$Y = \overline{A}$
OR	$Y = A + B$
AND	$Y = A \cdot B$
NOR	$Y = \overline{A + B}$
NAND	$Y = \overline{A \cdot B}$
EX - OR	$Y = A \oplus B$

La realizzazione di una funzione NOR a partire dall'invertitore elementare si ottiene in linea di principio ponendo in parallelo le uscite di N invertitori elementari (se N è il numero di variabili di ingresso della porta NOR), con un solo carico R in

comune agli N interruttori e un'unica alimentazione V^+, come è indicato in Figura 1.19.

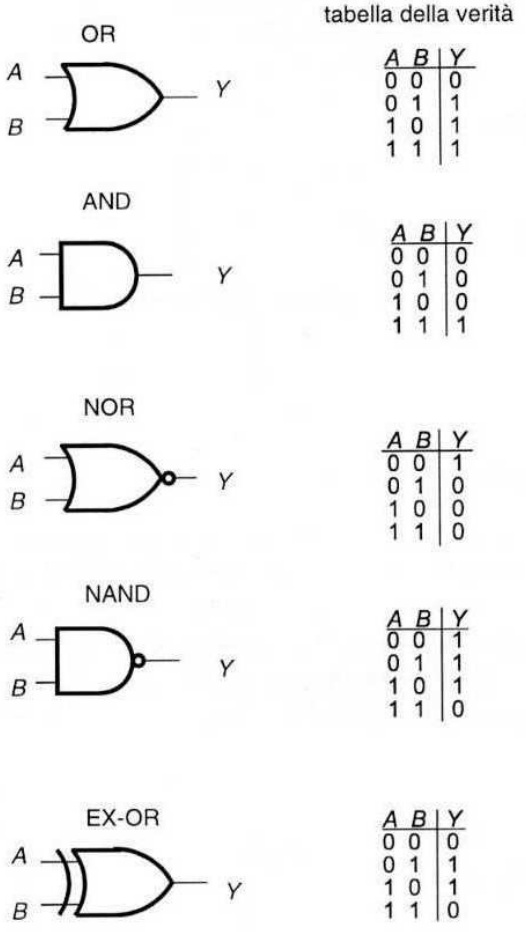


Figura 1.18 Simboli logici e tabelle della verità delle porte logiche elementari

Per comprendere la funzione implementata, basandosi sulla situazione di interruttore chiuso o aperto, si vede che:

- l'uscita è bassa (0) se l'interruttore *A* o l'interruttore *B* è chiuso (10, 01);
- l'uscita è bassa (0) se sia l'interruttore *A* che l'interruttore *B* sono chiusi (11);
- l'uscita è alta (1) se sia l'interruttore *A* che quello *B* sono aperti (00);

il che corrisponde alla tabella della verità della funzione NOR (si può facilmente vedere che, se gli interruttori si trovano tra l'alimentazione e il carico, e quest'ultimo è collegato a massa, il circuito realizza la funzione OR).

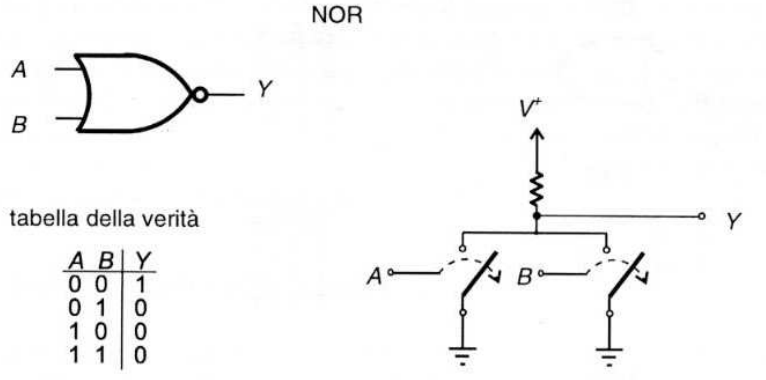


Figura 1.19 Porta NOR realizzata con invertitori idealizzati

Dualmente la funzione NAND può essere realizzata ponendo in serie N invertitori elementari con un unico carico per la serie (vedi Figura 1.20). In questo caso si ha che:

- l'uscita è alta (1) se solo l'interruttore A o l'interruttore B è chiuso (10, 01);
- l'uscita è alta (1) se sia l'interruttore A che quello B sono aperti (00);
- l'uscita è bassa (0) se sia l'interruttore A che l'interruttore B sono chiusi (11);

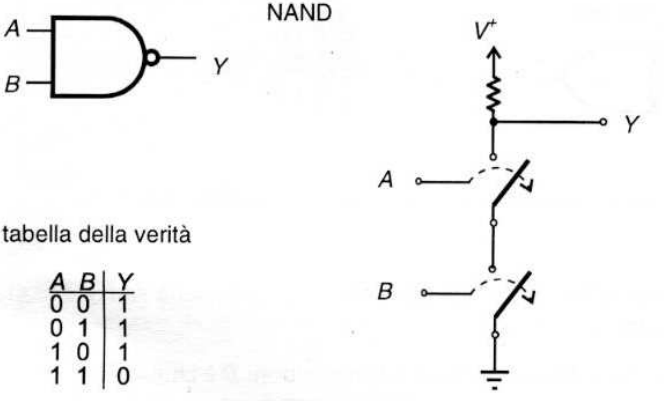


Figura 1.20 Porta NAND realizzata con invertitori idealizzati

il che corrisponde alla tabella della verità della funzione NAND (anche in questo caso, se si scambiano tra loro gli interruttori con il carico si realizza la funzione AND).

Ovviamente le funzioni OR e AND possono essere realizzate rispettivamente dalle porte logiche NOR e NAND ponendo in uscita un invertitore (funzione NOT). Infine la funzione EX-OR (OR ESCLUSIVO) può essere realizzata con porte AND e OR, in base alla relazione Booleana valida per la funzione EX-OR:

$$A \oplus B = A \cdot \overline{B} + \overline{A} \cdot B$$

Tabella 1.2 Alcuni teoremi dell'algebra Booleana

T1)	$A + 0 = A$	T2)	$A \cdot 0 = 0$
T3)	$A + \overline{A} = 1$	T4)	$A \cdot \overline{A} = 0$
T5)	$A + A = A$	T6)	$A \cdot A = A$
T7)	$A + 1 = 1$	T8)	$A \cdot 1 = A$
T9)	$A + A \cdot B = A$	T10)	$A \cdot (A + B) = A$
T11)	$\overline{A + B} = \overline{A} \cdot \overline{B}$	T12)	$\overline{A \cdot B} = \overline{A} + \overline{B}$
T13)	$A + B \cdot C = (A + B) \cdot (A + C)$	T14)	$A \cdot (B + C) = (A \cdot B) + (A \cdot C)$

L'analisi e la sintesi delle reti logiche basate sulla logica Booleana trascendono lo scopo di questo libro, e sono esaurientemente trattate in molti testi dedicati a questo argomento. Ci limiteremo a riportare in Tabella 1.2 alcuni dei teoremi dell'algebra Booleana che verranno richiamati in seguito nello studio di circuiti logici realizzati a partire da porte logiche elementari. Questi teoremi possono facilmente essere estesi al campo di n variabili e possono essere utilizzati per elaborare e modificare espressioni logiche.

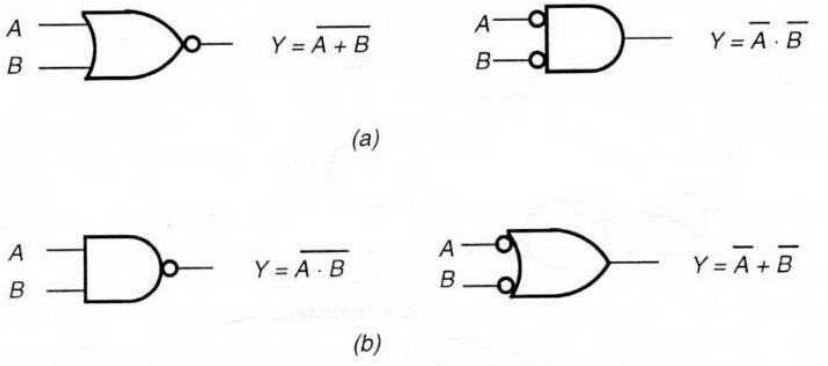


Figura 1.21 Rappresentazioni grafiche equivalenti di porte a) NOR, b) NAND

Utilizzando i teoremi dell'algebra Booleana è possibile ridurre qualsiasi espressione logica ad un insieme di somme di prodotti (o di prodotti di somme) e trasformare le reti logiche in circuiti basati unicamente su porte NOR o porte NAND. Può essere utile, per le analisi di alcuni semplici circuiti combinatori che saranno presentati nel seguito, richiamare una interpretazione grafica dei teoremi T11 e T12, conosciuti come *teoremi di De Morgan*, riportata in Figura 1.21. In base al teorema T11, infatti, si ricava che l'uscita di una porta NOR pilotata dalle variabili A e B, è uguale a quella di una porta AND pilotata dalle stesse variabili negate, e dualmente, per il teorema T12 l'uscita di una porta NAND è equivalente a quella di una porta OR pilotata dalle variabili negate. La negazione della variabile può essere rappresentata più sinteticamente dalla presenza di un cerchietto sul terminale rispettivo, e questo sia in ingresso che in uscita.

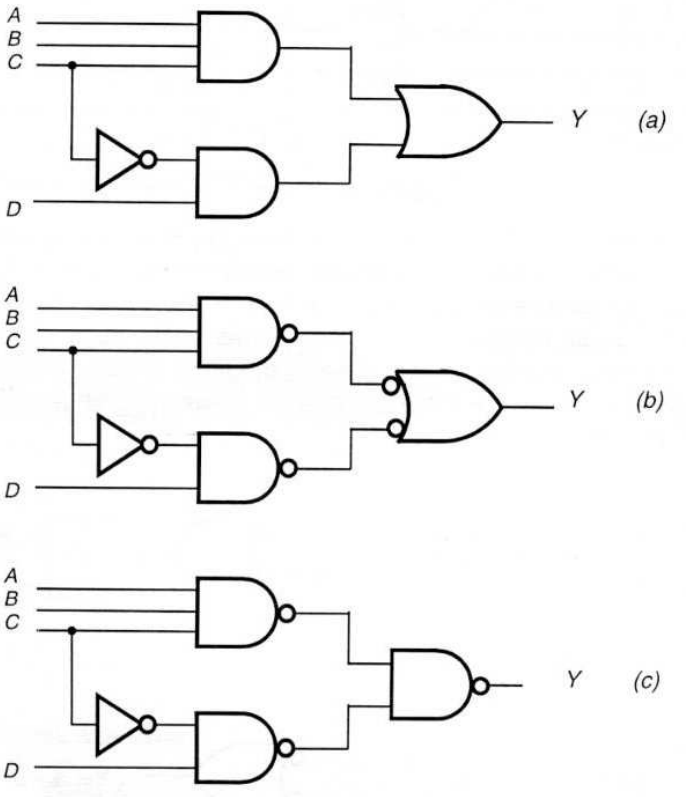


Figura 1.22 a) Rete logica che implementa l'espressione $Y = A \cdot B \cdot C + \overline{C} \cdot D$; b) sostituzione delle porte AND con porte NAND; c) versione del circuito che utilizza solo porte NAND

Questa rappresentazione grafica permette di semplificare rapidamente reti logiche semplici, annullando ad esempio negazioni doppie nelle trasformazioni di porte NOR in porte NAND e viceversa, o identificando i punti in cui occorre inserire delle negazioni aggiuntive per la trasformazione della rete.

Un esempio di queste trasformazioni è riportato in Figura 1.22, in cui è riportato il circuito logico che realizza l'espressione $Y = A \cdot B \cdot C + C \cdot D$ e una versione con sole porte NAND. Il circuito logico di Figura 1.22a è la realizzazione diretta dell'espressione logica data, e utilizza porte AND, OR e un invertitore per la negazione della variabile C. Il circuito di Figura 1.22b è ottenuto dal primo aggiungendo due negazioni sia all'ingresso che all'uscita per ognuno dei due collegamenti tra la porta AND e quella OR; con il primo si è in effetti sostituita la porta AND con una NAND, mentre con il secondo si è ottenuta una porta NOT-OR, che, in base alla Figura 1.21, è equivalente ad una porta NAND. Il circuito finale di Figura 1.22c riporta quindi la versione del circuito realizzata con sole porte NAND (l'invertitore presente nella rete può ovviamente essere realizzato sia con porte NAND che NOR, in base ai teoremi T5 e T6, ed in questo caso si intende realizzato con una porta NAND a due ingressi entrambi pilotati dalla variabile C).

1.6 Progetto dei sistemi digitali

Come si è detto precedentemente, la realizzazione dei sistemi digitali si basa su una organizzazione di gerarchie di circuiti via via più complessi, partendo dalle porte logiche che ne costituiscono gli elementi base; si comprende quindi come il progetto dei circuiti integrati si debba sviluppare necessariamente a diversi livelli di astrazione, partendo da una descrizione comportamentale del sistema e giungendo fino ad una descrizione fisica del circuito che dovrà essere implementato nel silicio, livelli sinteticamente indicati nello schema di Figura 1.22. Gli ultimi due passi di progetto, la rappresentazione circuitale delle diverse parti del sistema, e la definizione del tracciato per il circuito da realizzare, sono necessari quando non ci si debba limitare ad utilizzare componenti standard a basso livello di integrazione, da dovere assemblare su piastre per realizzare le funzioni volute, ma si desideri sfruttare la possibilità, sempre più sentita dai costruttori di sistemi, di realizzare i sistemi per quanto possibile per via integrata, approfittando delle capacità di realizzazione di circuiti specifici con elevata densità di componenti a costi contenuti e continuamente decrescenti.

Molti di questi livelli di progettazione possono essere notevolmente semplificati dall'aiuto del calcolatore (*Computer Aided Design*, progettazione assistita dal calcolatore, CAD). Gli strumenti CAD permettono, ad esempio, di verificare gli schemi logici, di analizzare i comportamenti dei circuiti elettrici, di effettuare automaticamente il posizionamento e le interconnesssioni (*placement and routing*) dei tracciati, e di estrarre le descrizioni logiche da descrizioni comportamentali. In ogni caso il progetto procede usualmente dall'alto verso il basso (approccio *top-down*) nello schema indicato, ma poiché la definizione temporale e più in generale la verifica delle specifiche di progetto dipende essenzialmente dalle scelte a livello cir-

cuitale e a livello di definizione del tracciato, è necessario ritornare con un approccio dal basso verso l'alto (*bottom-up*) su questi passi di progetto, fino alla definizione finale.

In questo testo, finalizzato allo studio del comportamento elettronico dei circuiti digitali, ci si limiterà a sviluppare gli elementi necessari per poter comprendere, analizzare e progettare, con l'aiuto di simulatori circuitali come lo SPICE, i circuiti che effettuano le funzioni logiche essenziali per i sistemi digitali, cercando di evidenziare i legami esistenti tra le prestazioni elettriche dei circuiti e le scelte di dimensionamento dei componenti e delle porte logiche elementari che li costituiscono; queste ultime valutazioni verranno effettuate utilizzando semplici versioni di strumenti CAD per l'elaborazione dei tracciati.

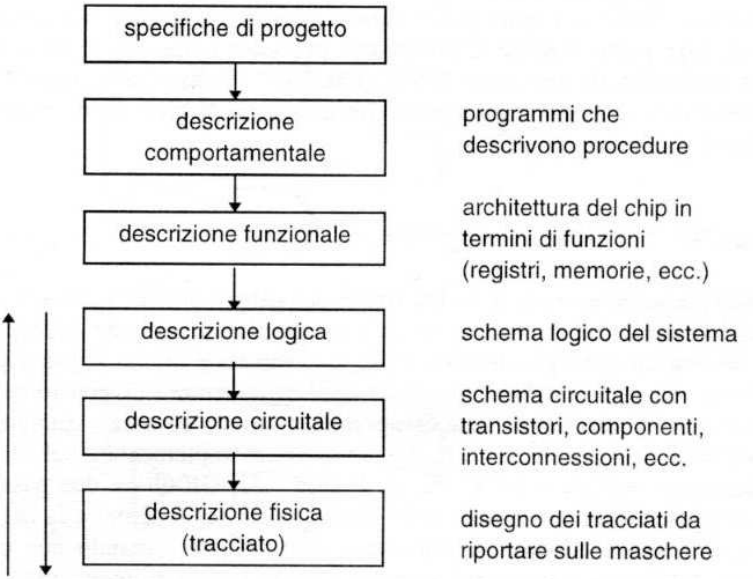


Figura 1.23 Livelli di astrazione nel progetto di sistemi digitali

Esercizi di riepilogo

1.1 Determinare i livelli logici ed i margini di rumore dell'invertitore descritto dalla funzione di trasferimento di Figura E1.1.

1.2 Spiegare perché l'invertitore che presenta la caratteristica di trasferimento di Figura E1.2 non può essere utilizzato per pilotare altri invertitori dello stesso tipo.

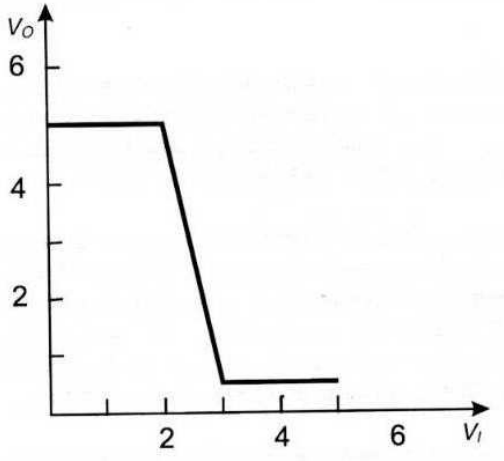


Figura E1.1

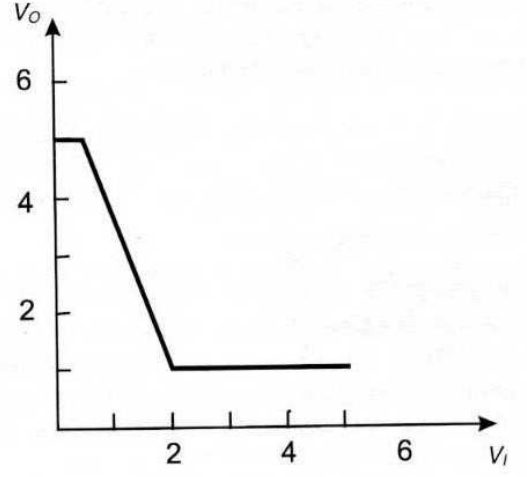


Figura E1.2

1.3 Identificare il percorso del segnale logico nel circuito di Figura E1.3 che presenta il maggiore ritardo di propagazione complessivo.

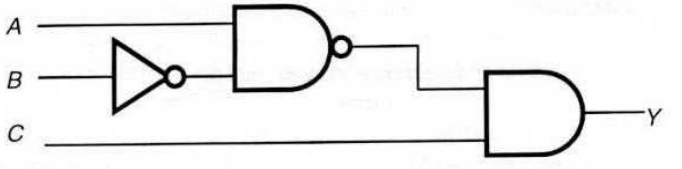


Figura E1.3

1.4 Con riferimento al circuito di Figura E1.3, assumendo un ritardo di propagazione $t_P = 5$ ns uguale per tutte le porte e per l'invertitore, e assumendo $C = 0$, mentre A e B passano contemporaneamente da 0 a 1, quali

sono le minime durate dei segnali A e B perché l'uscita Y rimanga alta per una durata di 10 ns?

1.5 Un invertitore che presenta la caratteristica di trasferimento di Figura E1.1 ed è alimentato a 5 V, assorbe una corrente $I_L = 0.4$ mA quando l'uscita è al valore logico basso, e $I_H = 0$ mA quando l'uscita è alta. Calcolare la dissipazione di potenza statica media.

1.6 Per l'invertitore dell'Esercizio 1.5, assumendo una capacità di carico $C_L = 20$ pF, e considerando solo la dissipazione dinamica di potenza dovuta alla capacità C_L, valutare la frequenza del segnale di ingresso per cui la dissipazione totale di potenza dell'invertitore è il doppio di quella statica. Qual è l'aumento possibile della frequenza del segnale, a parità di dissipazione, se la capacità C_L si riduce a 1 pF e la tensione di alimentazione si riduce a 3.3 V?

1.7 Un circuito integrato montato in un contenitore plastico che può dissipare 1 W, deve contenere 10.000 porte logiche elementari. Qual è la potenza massima che può dissipare ciascuna porta? Qual è il prodotto $P \cdot D$ delle porte se il massimo ritardo di propagazione ammissibile è di 5 ns?

1.8 Un circuito logico utilizza porte NOR e NAND a N ingressi ma ha solo $N-1$ variabili per ogni porta. A quale livello logico debbono essere rispettivamente connessi gli ingressi non utilizzati delle porte NOR e di quelle NAND?

1.9 a) Identificare il circuito logico che implementa in termini di porte logiche elementari la funzione $Y = A \cdot B + C \cdot (\overline{D + E})$; b) realizzare il circuito precedente con sole porte NAND; c) assumendo uguali ritardi di propagazione per tutte le porte dei circuiti dei casi a) e b), valutare quale delle due soluzioni presenta il maggior ritardo di propagazione per il percorso peggiore.

Riferimenti bibliografici

H. Taub, D. Schilling, *Elettronica integrata digitale*, Jackson, Milano, 1981.

J. Millman, *Circuiti e sistemi microelettronici*, Bollati Boringhieri, Torino, 1985.

A.S. Sedra, K.C. Smith, *Circuiti per la microelettronica*, Ingegneria 2000, Roma, 1993.

J. Millman, A. Grabel, *Microelettronica*, McGraw-Hill Libri Italia, Milano, 1994.

2

Tecnologie dei circuiti integrati

2.1 Introduzione

La diffusione dei sistemi elettronici in generale e dei sistemi digitali in particolare è dovuta essenzialmente alla possibilità di realizzare i circuiti che li costituiscono in forma *integrata*, cioè realizzando tutti i loro componenti e le relative interconnessioni direttamente nel silicio in forma miniaturizzata; la realizzazione integrata di circuiti elettronici nel materiale semiconduttore va sotto il nome di *microelettronica*. Il passaggio dalla realizzazione dei circuiti e sistemi elettronici in forma discreta (ossia assiemando i singoli componenti – resistori, condensatori, dispositivi attivi – su piastre sulle quali sono realizzati i collegamenti tra i diversi componenti mediante circuiti stampati), a quella integrata, che permette di realizzare interi circuiti nella singola tessera elementare (*chip*) di silicio, con operazioni tecnologiche sostanzialmente equivalenti a quelle necessarie per realizzare i singoli dispositivi, ha mostrato una forza dirompente attraverso la capacità di integrare circuiti sempre più complessi in dimensioni sempre più piccole. Infatti ciò non ha portato solo ad una *riduzione delle dimensioni* dei sistemi elettronici, con conseguente riduzione sia del numero di piastre elettroniche necessarie che del numero di "componenti" (intesi ora come microcircuiti in singolo contenitore) da assemblare per ogni piastra, ma ha portato ulteriori e significativi vantaggi nella *velocità di operazione*, essendo tutti i componenti ridotti dalle dimensioni di millimetri a quelle di micron e quindi anche le capacità parassite ridotte in maniera equivalente, e nel *consumo di potenza*, poiché di nuovo i dispositivi attivi hanno dimensioni sempre più ridotte e assorbono quindi minore corrente a parità di alimentazione. Tutto ciò si riflette evidentemente in notevoli benefici in termini di *costo* del sistema stesso a parità di prestazioni e contribuisce alla diffusione sempre più estesa di tali sistemi.

L'ingegnere progettista dei circuiti elettronici deve quindi utilizzare al meglio questa risorsa tenendo conto, nella progettazione del circuito, delle limitazioni e delle possibilità offerte dalla integrazione nel silicio. Lo studio delle tecnologie microelettroniche esorbita dalle tematiche di questo libro; al lettore che intendesse

approfondire l'argomento si suggerisce la lettura dei testi indicati in bibliografia. In questo capitolo si vogliono fornire solo dei richiami (per chi abbia già le basi di microelettronica) o delle informazioni (per chi sia a digiuno di questi aspetti) per poter meglio capire le considerazioni, legate alla integrazione dei circuiti, che saranno effettuate nel corso dell'esposizione della materia.

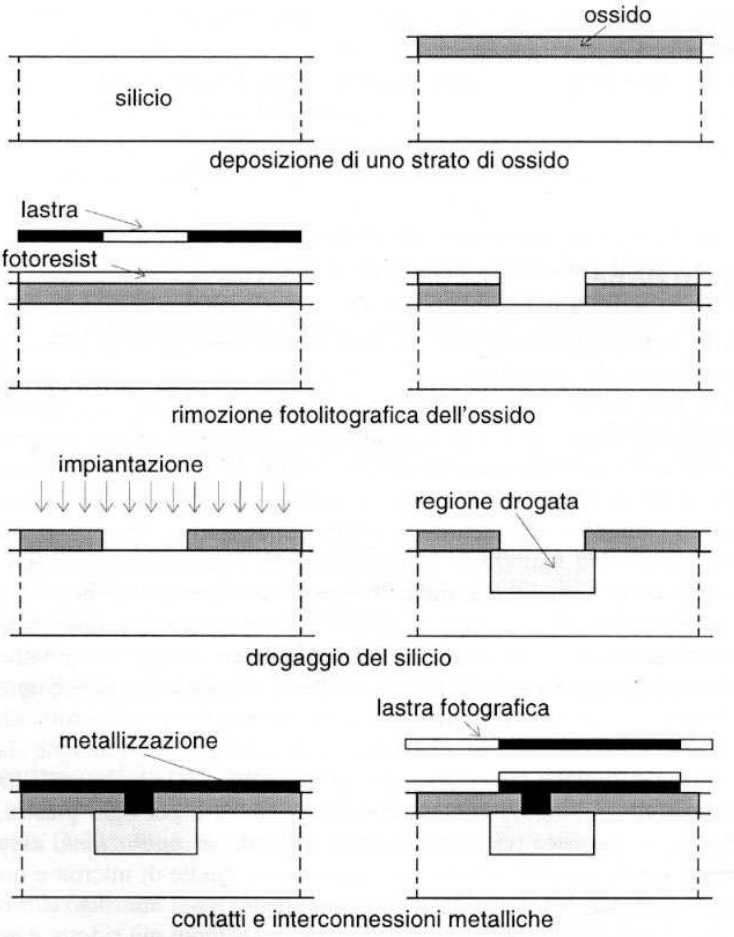


Figura 2.1 Processi tecnologici fondamentali per i circuiti integrati

La possibilità, introdotta per la prima volta nel 1962, di realizzare direttamente nel silicio, con opportune tecnologie, non solo dispositivi attivi, ma anche quant'altro serve per realizzare un circuito elettronico (resistori, condensatori, diodi e interconnessioni) è basata sulla iterazione di poche operazioni tecnologiche base, ancorché di sofisticazione sempre più spinta, che vengono condotte "in parallelo"

sulle fette (*wafers*) di silicio, in modo da fornire alla fine del processo (che, come si è detto, è poco più complesso di quello necessario per realizzare i singoli dispositivi), un elevato numero di microcircuiti uguali, pari al numero di chips che sono contenuti in una singola fetta.

Le operazioni base per realizzare i dispositivi e i componenti sono quelle schematicamente indicate in Figura 2.1, ed elencate di seguito:

a) deposizione di uno strato di ossido (SiO_2) o di nitruro (Si_3N_4)
L'ossido di silicio viene realizzato sul silicio sia direttamente per via termica in atmosfera di ossigeno, portando il silicio a temperature tra i 900 °C e i 1200 °C, che per deposizione chimica da fase vapore (CVD); il nitruro viene depositato per via CVD. Entrambi questi strati vengono utilizzati per mascherare le aree della fetta di silicio che non debbono essere esposte ai processi successivi (come drogaggio, metallizzazione, ecc.), mentre possono essere facilmente rimossi per i processi successivi, mediante attacco chimico. Questa rimozione può avvenire anche selettivamente mediante il processo fotolitografico, come indicato nel processo seguente, in modo da lasciare scoperte dall'ossido le sole aree che debbono essere esposte. Il nitruro inoltre, se depositato direttamente sul silicio, impedisce (nell'area su cui questo è rimasto dopo una rimozione selettiva) la crescita dell'ossido termico, e quindi agisce come maschera anche per l'ossidazione.

b) rimozione dell'ossido o del nitruro
La rimozione selettiva dello strato depositato avviene ricoprendo lo strato con una pellicola fotosensibile (*fotoresist*) ed impressionando questa attraverso una lastra fotografica che contiene (in positivo o in negativo) il disegno della o delle zone che si intende rimuovere. Una operazione di sviluppo (simile a quella di un normale processo fotografico) dello strato di fotoresist elimina la pellicola nelle aree dove essa è stata impressionata, lasciandola in quelle non illuminate dalla lastra. Una successiva esposizione a particolari attacchi chimici che rimuovono l'ossido (o il nitruro), ma non attaccano il fotoresist che è di materiale organico, permette l'eliminazione dello strato protettivo dal silicio nelle aree volute. Questa sequenza di operazioni prende il nome di *fotolitografia*, ed è il meccanismo base con cui è possibile realizzare i circuiti in via integrata.

c) drogaggio di tipo P o N
L'operazione fondamentale per la realizzazione dei dispositivi attivi (e dei componenti come resistori, diodi) nel silicio è quella del drogaggio della fetta in modo da realizzare opportune regioni di tipo P e/o N in aree definite. Questa operazione viene effettuata o per *diffusione termica*, introducendo le fette in forni a temperature tra 900 °C e 1200 °C in presenza di fosforo o arsenico (per drogaggio di tipo N) o boro (per drogaggio di tipo P), o, secondo la tecnologia oggi prevalentemente utilizzata, per *impiantazione ionica*, cioè accelerando sotto vuoto mediante un opportuno campo elettrico gli ioni del drogante (P, As, B) contro la superficie del silicio, in modo da farli penetrare ad una profondità dipendente dal potenziale di accelerazione degli ioni e creare così un sottile strato superficiale drogato; succes-

sivamente, mediante riscaldamento del wafer, il drogante impiantato viene fatto diffondere nel silicio (*drive-in*) fino alla profondità desiderata, realizzando così le regioni drogate. In entrambe le tecniche, lo strato di ossido (o di nitruro) depositato sul silicio agisce da maschera, impedendo l'introduzione del drogante nel silicio nelle aree su cui questi strati sono presenti, per cui l'operazione di drogaggio può avvenire solo nelle aree libere dall'ossido e quindi esposte al processo di impiantazione.

d) deposizione di strati conduttori
La contattazione delle regioni drogate dei singoli dispositivi e la creazione delle linee di interconnessione metalliche tra i dispositivi, vengono realizzate depositando per evaporazione sotto vuoto un sottile strato di alluminio su tutta la fetta, e utilizzando ancora lo strato di ossido che, opportunamente rimosso nelle aree da contattare con operazione fotolitografica, si comporta da maschera per la contattazione. Le piste di interconnessione vengono in seguito realizzate rimuovendo il metallo in eccesso mediante una ulteriore operazione di fotolitografia, utilizzando il fotoresist come maschera per l'attacco dell'alluminio.

Per le contattazioni di aree critiche dei dispositivi attivi, come la gate dei transistori MOS e l'emettitore e la base dei transistori bipolari avanzati, si usa anche uno strato di *polisilicio*, ossia film sottile di silicio depositato da fase vapore che assume una struttura policristallina (da qui il nome per distinguerlo dal silicio *monocristallino*, cioè realizzato come unico cristallo) che, come sarà spiegato nel Paragrafo 2.2, permette di ridurre le aree attive dei dispositivi e di ridurre le capacità parassite.

La sequenza di queste operazioni permette di realizzare in aree specifiche del chip sia i dispositivi attivi (transistori MOS o bipolari) che i componenti passivi (resistori, condensatori, diodi) nonché le regioni di isolamento tra i vari componenti e le linee di interconnessione del circuito elettrico, fino alle piazzole di saldatura (*pad*) a cui vanno saldati i terminali esterni del contenitore. Nei successivi paragrafi verranno richiamate le sequenze di processi essenziali per la realizzazione delle due famiglie tecnologiche, MOS e bipolari, e per la realizzazione degli altri componenti elettrici nelle stesse tecnologie.

2.2 Processi di fabbricazione per i transistori MOS

La sequenza dei processi per la realizzazione di un transistore MOS a canale N (o P) prevede essenzialmente le operazioni indicate schematicamente in Figura 2.2, con riferimento a un MOS a canale N realizzato in un wafer di silicio di tipo P detto *substrato*:

1. Apertura di una finestra nell'ossido di campo e crescita dell'ossido sottile di gate sull'area aperta, che costituirà il MOS.

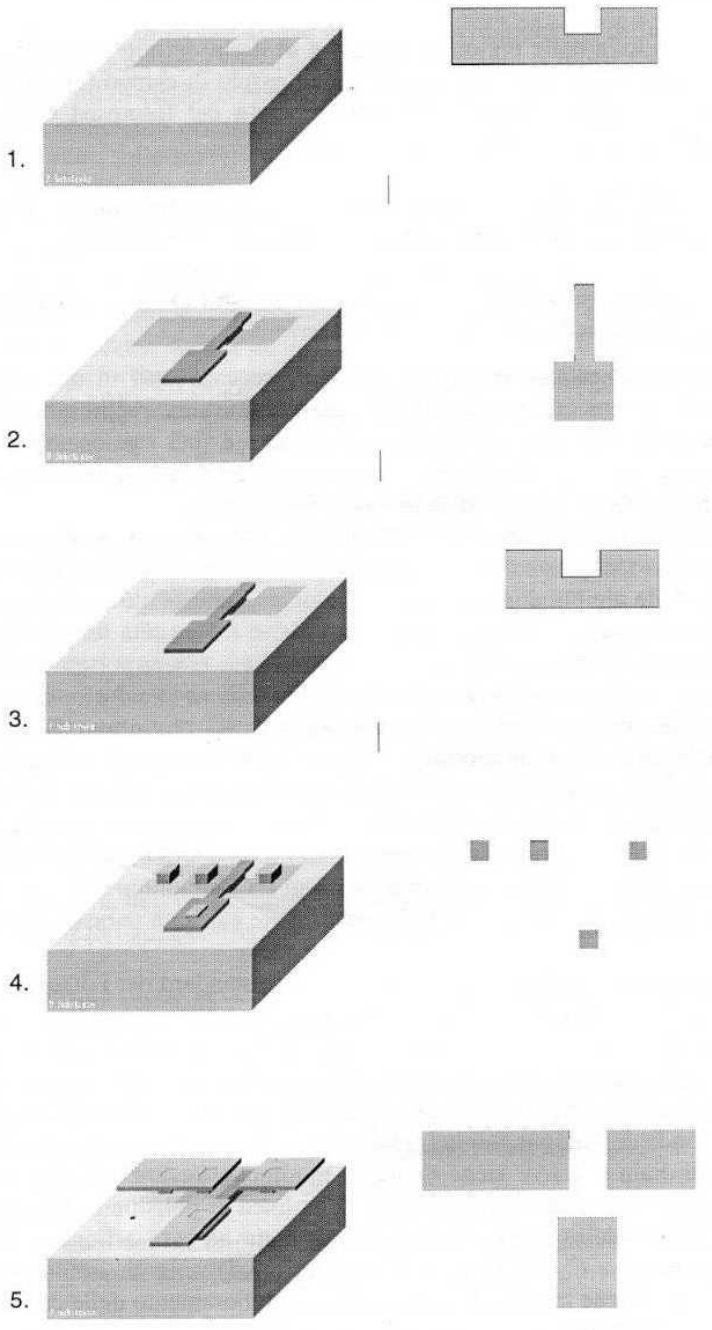


Figura 2.2 Processi di realizzazione di un transistore MOS

2. Deposizione di uno strato di polisilicio che realizza la gate e definizione fotolitografica della geometria della gate mediante rimozione del polisilicio.

3. Impiantazione selettiva delle regioni di source e di drain definite automaticamente dalla posizione del polisilicio di gate e dall'ossido di campo (la regione più chiara che costituirà il contatto N^+ è mascherata dal fotoresist durante l'impiantazione P e viene impiantata successivamente).

4. Copertura dell'area mediante ossido depositato CVD, e apertura dei buchi (*vias*) per la contattazione delle regioni di gate, source, drain e substrato.

5. Metallizzazione della fetta e determinazione fotolitografica della geometria delle metallizzazioni e delle eventuali interconnessioni.

Nella figura sono indicati gli effetti dei passi di processo visti in un'immaginaria prospettiva tridimensionale e in pianta; in quest'ultima sono rappresentate le diverse geometrie delle maschere fotolitografiche utilizzate (per ogni passo si è omesso di riportare lo strato di ossido spesso che ricopre tutto il wafer, detto ossido di campo, che impedirebbe la visione della successione dei passi).

Nelle prime generazioni di transistori MOS la gate era realizzata in alluminio invece che in polisilicio; in questo caso, non potendosi depositare l'alluminio prima della realizzazione delle regioni di drain e source, in quanto queste ultime richiedono dopo l'impiantazione un processo di diffusione (drive-in) ad alta temperatura (900-1000 °C) non compatibile con l'alluminio (che non deve essere portato a temperature superiori ai 450 °C), occorreva depositare l'alluminio successivamente alla realizzazione delle due regioni e utilizzare un ulteriore processo fotolitografico per la definizione della gate. Ciò comportava, a causa delle inevitabili tolleranze nell'allineamento delle diverse operazioni fotolitografiche di cui si parlerà nel Paragrafo 2.6, che la lunghezza L del canale tra source e drain non potesse scendere a valori di qualche micron, il che riduceva la corrente di uscita del transistore; inoltre si dovevano tollerare delle significative sovrapposizioni tra l'alluminio della gate e parte delle regioni di source e drain, con incremento delle capacità parassite equivalenti.

Nel processo con gate in polisilicio, che è diventato lo standard per i dispositivi utilizzati nei circuiti integrati LSI e VLSI, il polisilicio può essere depositato e definito prima della realizzazione delle regioni di source e di drain, in modo da poter essere utilizzato come maschera nella fase successiva di impiantazione di queste regioni. In questo modo la lunghezza di gate può essere portata al minimo valore compatibilmente con la definizione fotolitografica disponibile, e la sovrapposizione tra gate e regioni di drain e source viene ridotta al minimo, attraverso un processo definito "autoallineato", in quanto la definizione della linea di polisilicio definisce automaticamente i bordi delle regioni di source e di drain ad esso affiancate. Le capacità di sovrapposizione C_{GDO} e C_{GSO} (vedi Paragrafo 3.5) vengono quindi drasticamente ridotte e si ottiene in definitiva un notevole miglioramento delle caratteristiche elettriche dei transistori.

2.3 Processi per i transistori bipolari

La tecnologia di fabbricazione dei transistori bipolari si differenzia in partenza da quella utilizzata per i transistori MOS, in quanto il transistore bipolare è intrinsecamente un dispositivo realizzato in *verticale* rispetto alla superficie della fetta di silicio, anziché in *orizzontale* come il MOS, e la corrente di uscita (del collettore) va riportata sulla superficie superiore per realizzare un circuito integrato. Inoltre i singoli dispositivi realizzati non sono intrinsecamente isolati dal resto della fetta che li contiene, e quindi occorre creare delle regioni di isolamento per i singoli dispositivi.

Per tali ragioni il materiale di partenza utilizzato per i circuiti integrati bipolari è una fetta di silicio (*substrato*) usualmente di tipo P, sulla quale vengono inizialmente realizzate delle impiantazioni selettive di uno strato N^+ molto drogato detto *strato sepolto* (*buried layer*), che costituisce il collettore equivalente del transistore verticale, nelle aree in cui dovranno essere realizzati i singoli transistori. Su questo substrato di tipo P (contenente lo strato sepolto) si deposita un ulteriore strato di silicio cristallino di tipo N attraverso un processo che viene chiamato *crescita epitassiale* (in pratica si espone la superficie del wafer ad un'atmosfera di silicio in fase gassosa, a temperature superiori a 1200 °C, in modo da accrescere sulla superficie del wafer uno strato cristallino con la stessa orientazione del substrato, ma con drogaggio determinato dalle impurità introdotte durante la crescita), nel quale sarà realizzato il transistore bipolare. Poiché quest'ultimo nei circuiti integrati deve avere tutti i terminali accessibili sulla faccia superiore del wafer, lo strato sepolto ha anche la funzione di elettrodo di raccolta della corrente di collettore, riportandola attraverso una via di bassa resistenza al terminale di collettore, che è posto sulla superficie superiore della fetta, agendo quindi di fatto come uno strato conduttore, integrato nello spessore di silicio.

Le fasi di processo essenziali per la realizzazione di transistori bipolari integrati sono elencate in Figura 2.3, per iseguenti passi:

1. apertura, nell'ossido cresciuto sul substrato di tipo P, delle aree in cui verranno realizzati i dispositivi per l'impiantazione dello strato sepolto;

2. apertura (dopo la crescita epitassiale) delle regioni di isolamento nell'ossido e impiantazione di drogante P^+ per le regioni di isolamento;

3. apertura dell'area per l'impiantazione della regione P di base;

4. apertura delle aree per la impiantazione della regione N^+ di emettitore e per la realizzazione del contatto di collettore;

5. apertura dei buchi (vias) per la contattazione delle regioni di emettitore, base, collettore (e substrato) ;

6. metallizzazione della fetta e definizione fotolitografica della geometria delle metallizzazioni e delle eventuali interconnessioni.

indicando, nella figura, sia le modifiche introdotte nella sezione verticale del wafer che le corrispondenti mascherature delle aree superficiali, ossia la successione dei tracciati (lay-out) del dispositivo.

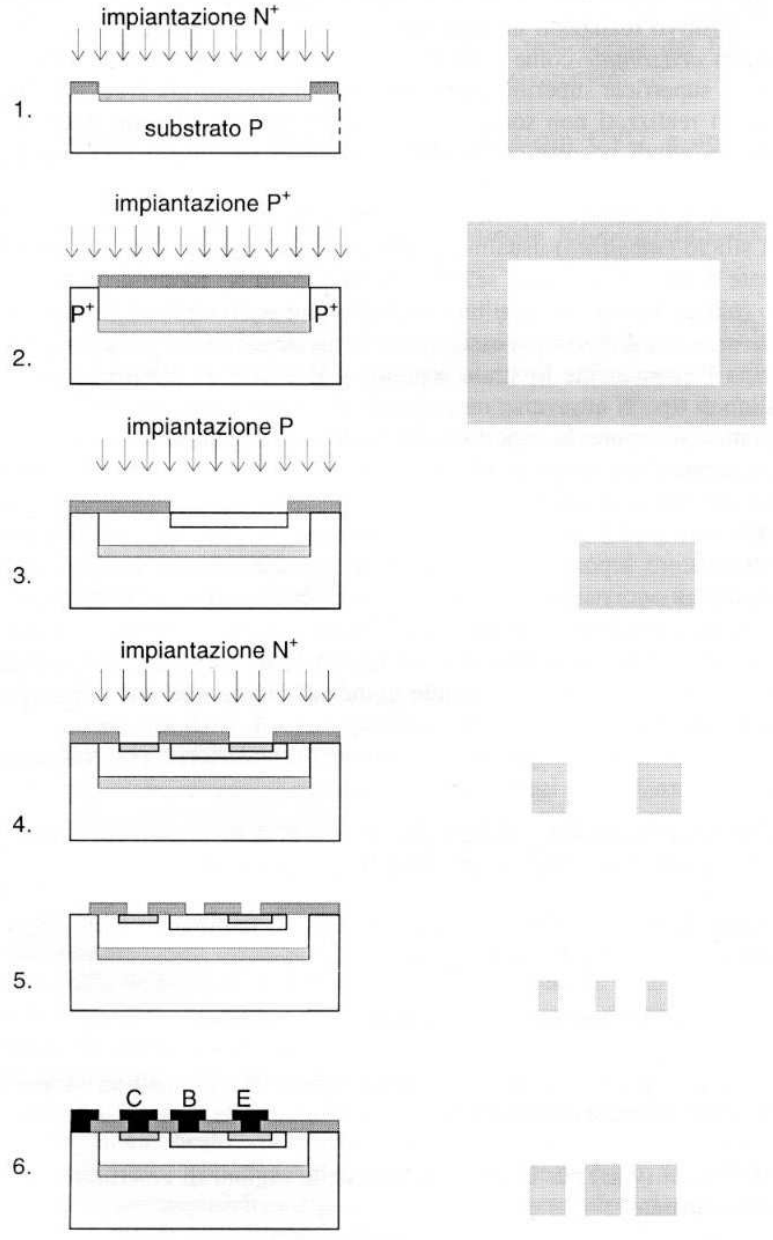


Figura 2.3 Processi di realizzazione di transistori bipolari integrati

Anche per i transistori bipolari si è avuta una notevole evoluzione nei processi di fabbricazione nell'ultimo decennio, per cui il processo qui descritto è stato notevolmente migliorato con l'introduzione dell'isolamento laterale mediante ossido, e la riduzione delle dimensioni dell'emettitore e della base mediante tecniche "autoallineanti" che fanno uso del polisilicio. Si ritornerà su questi aspetti discutendo dei miglioramenti dei transistori bipolari nel Capitolo 6.

2.4 Altri componenti

Nelle tecnologie integrate i processi fondamentali sono quelli utilizzati per realizzare i dispositivi attivi, e proprio per questo i circuiti integrati vengono divisi in due grandi famiglie: integrati con tecnologia MOS o con tecnologia bipolare, a seconda che i circuiti vengano realizzati con l'uno o l'altro tipo di dispositivi. Occorre quindi che gli altri componenti delle reti elettriche che debbono essere realizzati, possano essere fabbricati utilizzando alcuni (o tutti) i processi utilizzati per i dispositivi attivi, in modo da poter essere facilmente integrati nel processo di realizzazione del circuito complessivo, utilizzando la tecnologia MOS o bipolare.

Resistenze

Per i resistori si utilizzeranno le diffusioni che invertono il segno del drogante per realizzare delle regioni isolate, in cui vengono impiantate le regioni di opportuna resistività che, una volta contattate, permettono di realizzare le resistenze di valore voluto, secondo lo schema di Figura 2.4. Nel processo MOS andranno utilizzate le tasche (*well*) come regione di isolamento, e le diffusioni utilizzate per creare le regioni di source/drain per i resistori veri e propri, mentre nel processo bipolare si userà una regione epitassiale N circondata dalle regioni di isolamento P e si realizze-

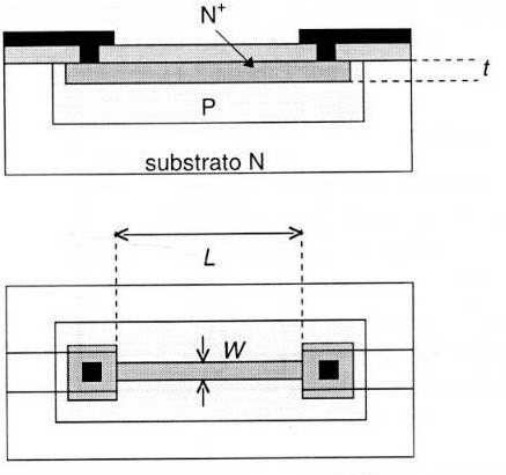


Figura 2.4 Realizzazione di resistori integrati

rà il resistore utilizzando l'impiantazione della regione di base. Facendo riferimento a una geometria di tipo rettangolare, detta t la profondità della regione drogata, W la larghezza e L la lunghezza della regione, la resistenza si ricava secondo la formula:

$$R = \rho \frac{L}{S} = \rho \frac{L}{tW} \tag{2.1}$$

La resistività ρ è data dal numero di atomi droganti impiantati per unità di volume N secondo l'espressione $\rho = 1/(qN\mu)$. La grandezza $(N \cdot t)$ rappresenta il numero di atomi impiantati nel silicio per unità di area, e quindi è da considerarsi una costante del processo di impiantazione, per cui la grandezza $\rho/t = 1/(qN\mu\, t)$ viene definita come *resistenza di strato* R_S. In base alla definizione della resistenza di strato R_S si può scrivere l'Equazione (2.1) come:

$$R = R_S \frac{L}{W}, \quad \text{con } R_S \text{ espressa in ohm per quadro } (\Omega/\square) \tag{2.2}$$

(la definizione di R_S deriva dal fatto che $R = R_S$ se ci si riferisce a una regione quadrata, con $L = W$); la resistenza complessiva viene calcolata dalla resistenza di strato R_S moltiplicata per il numero di quadrati di lato W che entrano nella lunghezza L.

Nel processo MOS, poiché l'impiantazione viene utilizzata per realizzare le regioni di source/drain che debbono avere un elevato drogaggio, la resistenza di strato è relativamente bassa, tra 20 e 50 $\Omega/\square$, e quindi risulta difficile integrare resistenze superiori a 5 KΩ perché risulterebbero troppo lunghe occupando uno spazio eccessivo. Nel processo bipolare invece, si può utilizzare il processo di impiantazione della base (che richiede drogaggi più bassi), per realizzare resistori a più elevato valore, in quanto con il processo di base si ottengono resistenze di strato tra 200 e 500 $\Omega/\square$. È possibile anche utilizzare, come vedremo nel caso delle celle RAM, la resistenza offerta da uno strato di polisilicio, nel caso che si debba realizzare una resistenza di valore molto elevato (vedi Tabella 2.1).

Tabella 2.1 Valori tipici di resistenza di strato R_S

materiale	$\Omega/\square$
alluminio	0.05
regioni source/drain	25
regione di base	200
polisilicio (non drogato)	5M
polisilicio (drogato)	50

Capacità

Nella tecnologia MOS i condensatori vengono realizzati usualmente con strutture metallo-ossido-semiconduttore, e cioè ricoprendo con il metallo un sottile ossido

cresciuto sulla superficie del semiconduttore, come per la realizzazione della gate dei MOS. La capacità offerta da una struttura metallo-ossido-semiconduttore, come quella riportata in Figura 2.5, si può esprimere in via semplificata secondo la relazione (valida, come verrà ricordato nel Capitolo 3, per tensioni relativamente basse applicate all'elettrodo metallico):

$$C_{MOS} = A\frac{\varepsilon_{OX}}{t_{OX}} \tag{2.3}$$

dove A è l'area della regione MOS, ε_{OX} è la costante dielettrica dell'ossido di silicio (pari a 3.9 ε_O), t_{OX} lo spessore dell'ossido.

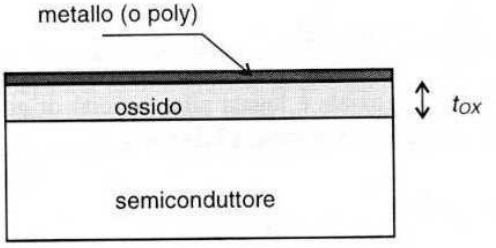


Figura 2.5 Capacità della struttura MOS

Una via alternativa è data dalla realizzazione di una giunzione P/N (o N/P) contropolarizzata, che presenta come è noto una capacità dipendente dalla zona di svuotamento; questa via è quella utilizzata in tecnologia bipolare. Le capacità di giunzione sono in ogni caso presenti in tutti i componenti realizzati in tecnologia integrata, poiché questi vengono usualmente realizzati in regioni isolate dal resto del substrato mediante giunzioni P/N contropolarizzate; si richiamerà quindi la loro espressione, con riferimento ad una giunzione N^+P contropolarizzata, come quella riportata in Figura 2.6.

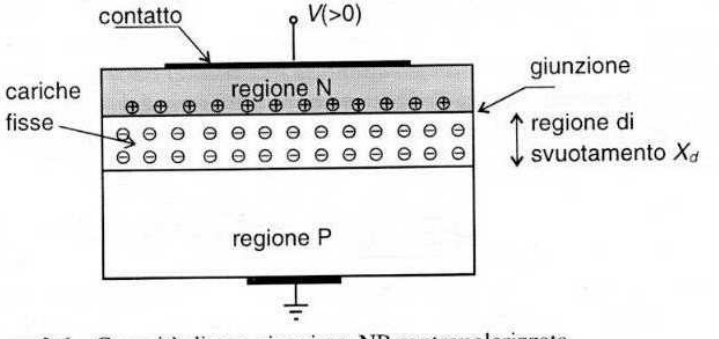


Figura 2.6 Capacità di una giunzione NP contropolarizzata

Si ricorda che in condizione di contropolarizzazione, intorno alla giunzione si crea una regione di svuotamento, priva di cariche mobili e contenente solo le cariche fisse dovute agli ioni degli atomi di drogante N_D e N_A (rispettivamente per le regioni N e P); questa regione si estende essenzialmente nella regione a basso drogaggio, poiché le cariche fisse (cioè gli ioni degli atomi droganti) nelle due parti della regione di svuotamento debbono equilibrarsi, e $N_D^+ >> N_A$. La carica per unità di area nella regione di svuotamento Q_{SC} sarà data da:

$$Q_{SC} = -qN_A X_d \tag{2.4}$$

Se il potenziale ai capi della giunzione aumenta di dV, la regione di svuotamento aumenta di dX_d e la carica Q_{SC} varierà della quantità:

$$dQ_{SC} = -qN_A dX_d \tag{2.5}$$

Ricordando che la variazione di potenziale è legata alla capacità di giunzione C_J secondo la relazione $dV = d\,Q_{SC}/C_J$ e utilizzando la (2.5) si ha:

$$dV = \frac{dQ_{SC}}{C_J} = \frac{dQ_{SC} X_d}{\varepsilon_{SI}} = -q\frac{N_A X_d dX_D}{\varepsilon_{SI}} \tag{2.6}$$

Integrando la (2.6) sulla regione di svuotamento X_J si ha:

$$\int_{V_N}^{V_P} dV = -\int_0^{X_d} q\frac{N_A X_d}{\varepsilon_{SI}} dX_d; \quad V + \phi_0 = q\frac{N_A X_d^2}{2\varepsilon_{SI}} \tag{2.7}$$

dove ϕ_0 è il *potenziale di barriera (built-in voltage)* presente ai capi della regione di svuotamento nel caso di assenza di contropolarizzazione, dipendente dal drogaggio delle regioni P e N secondo la relazione:

$$\phi_0 = V_T\, ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{2.8}$$

dove V_T è la tensione termica, e n_i la concentrazione intrinseca del silicio. Dalla (2.7), specializzata per il caso di $V = 0$, si ricava l'espressione dello svuotamento X_{dO} in assenza di contropolarizzazione:

$$X_{d0} = \sqrt{\frac{2\varepsilon_{SI} \cdot \phi_0}{qN_A}} \tag{2.9}$$

Si definisce quindi una capacità di giunzione in assenza di contropolarizzazione C_{J0} (sempre per unità di area) data da:

$$C_{J0} = \frac{\varepsilon_{SI}}{X_{d0}} = \sqrt{q \frac{\varepsilon_{SI} \cdot N_A}{2\phi_0}} \tag{2.10}$$

In presenza di polarizzazione V della giunzione, la capacità C_J, data dalla. (2.7), si può esprimere in funzione di C_{J0} attraverso la (2.10) come:

$$C_J(V) = A \cdot C_{J0} \left(\frac{1}{1 + V / \phi_0} \right)^{1/2} \tag{2.11}$$

dove A è l'area della giunzione della regione in esame. Dalla (2.11) si può notare come la capacità di una giunzione decresca all'aumentare (in modulo) della tensione di contropolarizzazione della giunzione stessa. In Tabella 2.2 sono riportati i valori delle costanti che intervengono nelle espressioni delle capacità di giunzione e dell'ossido.

Tabella 2.2

grandezza	*valore*
ε_{SI}	0.1 fF/μm
ε_{OX}	0.034 fF/μm
ϕ_0	0.86 V

Altri componenti

I diodi sono parte del processo bipolare e quindi vengono realizzati sia utilizzando le impiantazioni di emettitore nelle regioni di base che quelle delle basi nello strato epitassiale; per il processo MOS si possono utilizzare le giunzioni ottenute dalle regioni di source/drain nelle tasche di isolamento.

Per quanto riguarda gli induttori, va detto che questi ultimi, a differenza degli altri componenti, non sono facilmente integrabili. Tuttavia va ricordato che gli induttori non sono considerati componenti utili per i circuiti digitali; al contrario le inevitabili induttanze parassite create dalle linee di interconnessione possono creare problemi per la propagazione dei segnali.

2.5 Interconnessioni

Le interconnessioni tra i diversi dispositivi e componenti che costituiscono il circuito integrato vengono, come si è detto, realizzate con il processo di metallizzazione della fetta e la successiva delineazione delle piste e dei contatti mediante un

processo fotolitografico apposito. Con l'aumentare della densità dei componenti integrati in un singolo chip (le massime densità attuali di integrazione superano i 10^8 componenti per chip) il problema dell'allocazione dei componenti e dei collegamenti tra le diverse parti del chip, definito con termine inglese come *placement and routing*, è divenuto uno degli aspetti cruciali del progetto del chip.

Come si può facilmente capire, se si utilizza un solo strato di metallo per la definizione delle piste si creano non indifferenti problemi topologici a causa della necessità di evitare gli attraversamenti tra le piste stesse, problemi che diventano rapidamente insormontabili con l'aumentare della complessità del circuito.

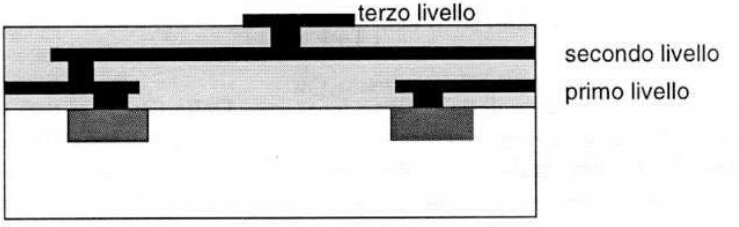


Figura 2.7 Metallizzazioni e interconnessioni a più livelli

È quindi diventata la norma nei circuiti a larga scala di integrazione (LSI e VLSI) realizzare le interconnessioni a due o più livelli di metallo, e cioè depositando e delineando in sequenza più strati di metallo separati da strati di ossido, in cui linee di interconnessione si attraversano senza toccarsi, perché separate in verticale dallo strato di ossido, secondo lo schema indicato in Figura 2.7; i collegamenti in verticale tra un piano e l'altro della metallizzazione sono realizzati con opportuni fori aperti nell'ossido, chiamati *vias*, nei quali vengono utilizzati metalli diversi dall'alluminio, come il molibdeno, per migliorare il riempimento del canale e ridurre la resistenza di contatto. Attualmente sono utilizzati processi a partire da due livelli di metallo fino a 5-6 livelli per i microprocessori più avanzati.

Le linee di interconnessione possono anche usare il polisilicio, che in effetti sostituisce l'alluminio per la contattazione di alcune regioni dei dispositivi attivi; tuttavia, poiché il polisilicio ha una resistività ben maggiore di quella del metallo, si utilizza il primo solo per brevi collegamenti (ad esempio per connettere due gate di dispositivi MOS) lasciando all'alluminio le interconnessioni di maggiore lunghezza e il trasporto di correnti elevate.

Un aspetto indesiderabile che presentano le linee di interconnessione è quello della inevitabile capacità parassita associata alla struttura metallo-ossido-semiconduttore che costituisce la linea stessa. Questa capacità C_l, descritta dalla relazione:

$$C_l = \frac{\varepsilon_{OX}}{t_{OX}} L \cdot W \tag{2.12}$$

dove L è la lunghezza della linea e W la sua larghezza, pur se costruita sull'ossido di campo, cioè un ossido ben più spesso (10-20 volte) di quello di gate, può diventare rilevante rispetto alla capacità di ingresso di un MOS se la lunghezza della linea raggiunge le frazioni di millimetro.

I percorsi più lunghi vengono di solito realizzati mediante le interconnessioni di secondo livello, che presentano una capacità più bassa per unità di area rispetto a quelle di primo livello (o di polisilicio), dovuta al maggiore spessore di isolante tra la linea e il silicio delle prime.

2.6 Tracciati e regole di progetto

Da quanto su esposto, si comprende come la realizzazione dello specifico circuito integrato sia ampiamente affidata alla realizzazione degli opportuni tracciati (*layout*) da riportare sulle differenti maschere usate per i processi fotolitografici, che nella loro successione definiscono le dimensioni e le posizioni dei dispositivi, componenti ed interconnessioni che realizzano il circuito stesso.

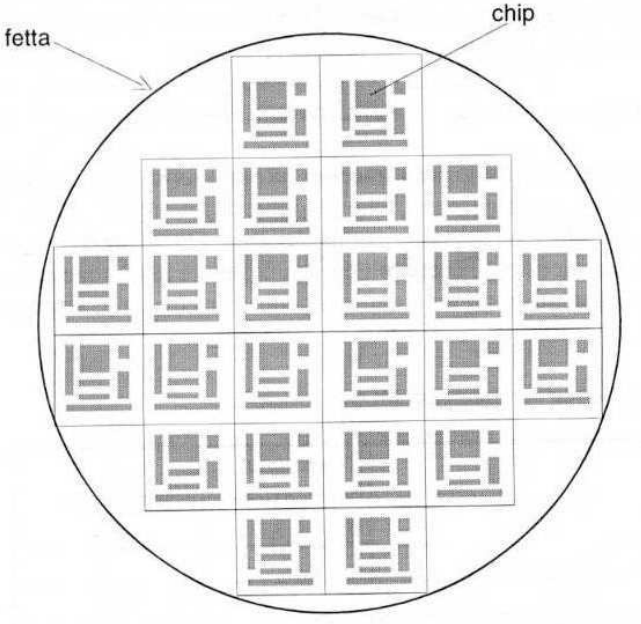


Figura 2.8 Iterazione dei chips sulla fetta di silicio

La fabbricazione dei differenti circuiti integrati, o chip, in una fabbrica che possegga le tecnologie necessarie, è quindi realizzata disegnando il set di maschere opportuno per i diversi circuiti, e definendo la sequenza di passi tecnologici che vanno applicati alle fette con quel set di maschere. La produzione dei

circuiti integrati ha avuto un continuo calo dei costi con l'aumentare della densità di integrazione, la standardizzazione dei processi e l'automazione di alcuni dei passi di progettazione dei tracciati delle maschere, nonché con l'aumentare del volume di produzione legato alla capacità di operare con fette di diametro crescente.

In effetti la rivoluzione che si è verificata nella realizzazione dei sistemi elettronici con l'integrazione su larga scala è per molti versi simile a quella avutasi nella scrittura dei libri nel quindicesimo secolo con l'invenzione della stampa. Anche in quel caso si è passati da una realizzazione "in serie" di caratteri in modo manuale per la realizzazione del testo, che andava ripetuta per ogni esemplare da realizzare, ad una preparazione "in parallelo" delle pagine mediante "matrici" che poi venivano utilizzate per un numero illimitato di stampe, e quindi di esemplari.

Nel caso delle maschere per fotolitografia il disegno del singolo tracciato, per ogni operazione fotolitografica, viene realizzato con tecniche CAD e trasferito sulla lastra (ricoperta da una pellicola metallica). Questo disegno viene quindi trasferito sul fotoresist con opportune macchine per l'esposizione, dette *allineatori*, che permettono di centrare le maschere successive rispetto a segni di allineamento lasciati sulla fetta dalle precedenti operazioni, in modo da assicurare la compatibilità delle operazioni successive con le precedenti. La maschera usualmente contiene il tracciato di un singolo chip, e questo tracciato viene iterato sulla fetta in modo da dar luogo al numero di chip (uguali) che si otterranno alla fine del processo, come in Figura 2.8.

Un aspetto importante nel disegnare i tracciati fotolitografici è quello delle cosiddette *regole di progetto* (*design rules*) che bisogna rispettare per ottenere una resa elevata alla fine dei processi di realizzazione dei chip. Queste regole in pratica determinano le minime dimensioni accettabili per le aperture e le minime spaziature tra le diverse parti che garantiscono un comportamento corretto delle strutture realizzate in ogni posizione sulla fetta.

Tabella 2.3 Regole di progetto per transistori MOS

regione	*minima dimensione*	*minima distanza*
polisilicio	2 λ	3 λ
sovrapp. polisilicio / regioni diffus.		2 λ
regioni P o N	3 λ	3 λ
tasche di isolamento	10 λ	6 λ
bordo regioni diffus. / isolamento		6 λ
metallo	3 λ	3 λ
apertura contatti	2 λ	3 λ
contatti / regioni diffus., poly		2 λ
sovrapp. metallizzazione / contatto		2 λ
metallizzazione	3 λ	3 λ

La maggior parte di queste regole è legata alla presenza di inevitabili tolleranze nei processi fotolitografici. Infatti le operazioni di allineamento tra una maschera e l'altra possono essere realizzate fino ad una tolleranza minima sotto la quale non si può garantire il risultato; per ogni sistema fotolitografico viene quindi definita una dimensione minima (*feature size*) λ, che nel tempo si è andata riducendo da valori di parecchi micron a meno di 0.5 micron.

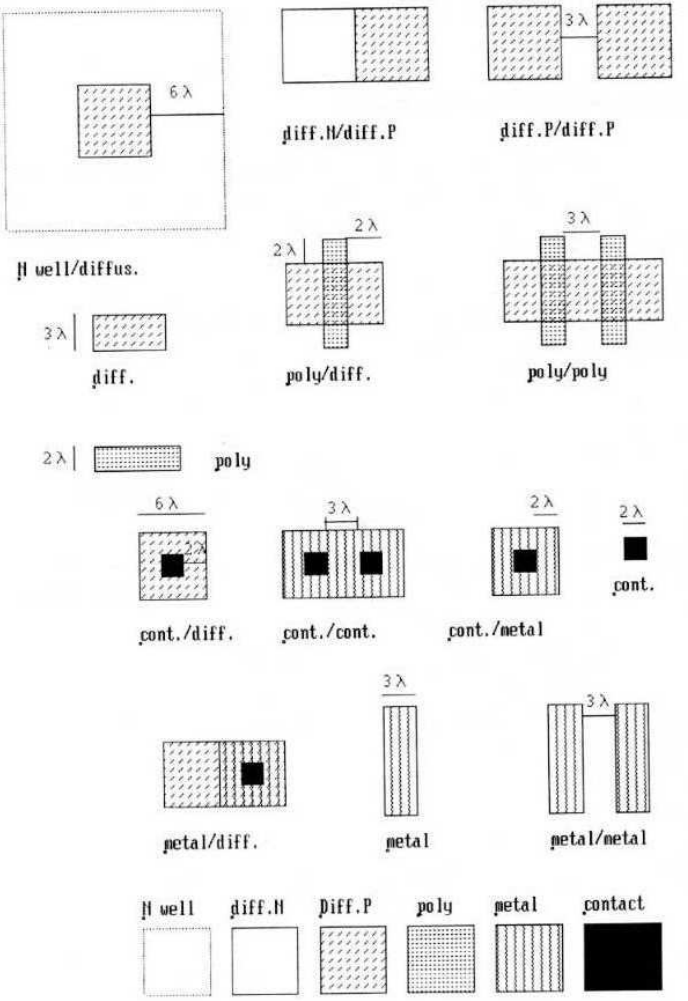


Figura 2.9 Regole di progetto per un processo N-well. Le dimensioni sono in multipli di λ

Oltre a questa grandezza legata al processo fotolitografico occorre tenere conto di altre variabili legate ai processi tecnologici coinvolti, come l'attacco laterale

degli strati di ossido o di nitruro, la diffusione laterale del drogante nei processi termici, ecc. Da questi effetti derivano le regole di progetto, legate alla tecnologia utilizzata, che definiscono le minime dimensioni delle diverse regioni (polisilicio, diffusioni P e N, regioni di isolamento, linee di metallo, vias) e che di solito sono indicate in multipli interi della dimensione minima λ. A titolo di esempio in Tabella 2.3 sono riportate alcune regole di progetto per la realizzazione di circuiti CMOS con tasca di isolamento N, e in Figura 2.9 sono esemplificate alcune delle regole riportate; nella stessa figura sono anche riportate le codifiche di tratteggio delle diverse regioni, che verranno adottate nel seguito per i lay-out dei circuiti esaminati.

2.7 Scala di integrazione dei circuiti

Come si è visto nel paragrafo precedente, la capacità di integrare un numero elevato di dispositivi in un chip è legata alla riduzione della minima dimensione (feature size) del processo microelettronico utilizzato. Quest'ultima grandezza si è continuamente ridotta nell'evoluzione delle tecnologie, passando dalla decina di micron dei primi circuiti integrati a valori sensibilmente inferiori al micron dei processi attuali, avvicinandosi a quello che è ritenuto un limite ultimo della fotolitografia ottica, e cioè la lunghezza d'onda della luce (ultravioletto) utilizzata per l'esposizione del fotoresist.

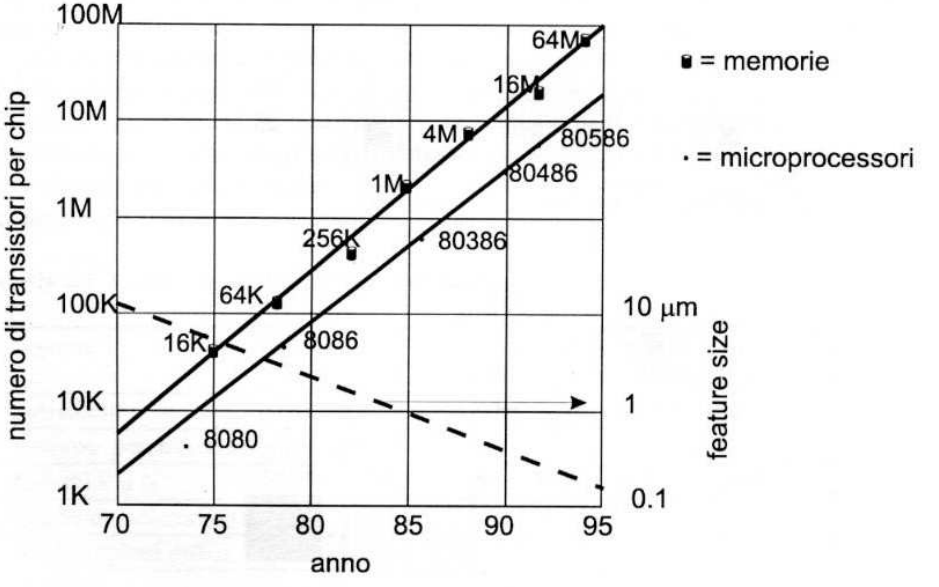


Figura 2.10 Evoluzione del numero dei dispositivi integrabili in un chip

In dipendenza della diminuzione della dimensione minima, si è avuta la crescita del numero di dispositivi integrabili in un chip, riportata in Figura 2.10 per i circuiti

a più alta densità di integrazione come le memorie e i circuiti più complessi come i microprocessori, con un ritmo che presenta un raddoppio del numero di dispositivi ad ogni anno. La densità di componenti integrabili nel singolo chip viene usualmente indicata come:

da 2 a 200	dispositivi:	SSI	(Small Scale Integration)
da 200 a 2.10^3	dispositivi:	MSI	(Medium Scale Integration)
da 2.10^4 a 2.10^4	dispositivi:	LSI	(Large Scale Integration)
da 2.10^4 a 2.10^5	dispositivi:	VLSI	(Very Large Scale Integration)
oltre 2.10^5	dispositivi:	ULSI	(Ultra Large Scale Integration)

Nei primi anni (65-75) dell'evoluzione dei circuiti integrati il livello di integrazione raggiungibile era quello SSI e MSI, che permetteva l'integrazione di porte logiche e semplici circuiti sequenziali, come flip-flop e registri. Questi circuiti erano prodotti come componenti standard, con caratteristiche prefissate ed in grande quantità. Il progetto di un sistema digitale consisteva quindi essenzialmente nella definizione dello schema logico per la realizzazione delle specifiche, e non occorreva sviluppare il progetto circuitale né tanto meno il suo tracciato, ma piuttosto la disposizione dei componenti logici elementari sulla piastra. Oggi l'evoluzione delle capacità di integrazione su chip, associate allo sviluppo di *tools* di progettazione automatica con tecniche CAD, hanno portato un numero sempre più elevato di industrie di sistemi a progettare e a far realizzare circuiti integrati per applicazioni specifiche (ASICS), con una notevole diminuzione dei costi di produzione, e con la possibilità di introdurre innovazioni circuitali e di funzioni nei sistemi prodotti, in maniera non facilmente riproducibile. Le tecniche di progettazione e di descrizione dei circuiti integrati sono quindi uscite dall'ambito ristretto dei laboratori di ricerca e sviluppo delle industrie produttrici di chip, per giungere a quelle produttrici di apparati e sistemi. Le prime hanno aperto le loro capacità tecnologiche alla produzione dei chip commissionati dai progettisti delle seconde e, per indicare questa funzione di esecuzione tecnologica di integrati progettati da altri, gli stabilimenti di produzione dei chip vengono indicati come Fonderie di silicio (Silicon Foundries).

Esercizi di riepilogo

2.1 Disegnare le sezioni verticali della struttura MOS per ognuno dei passi di processo riportati in Figura 2.2.

2.2 Disegnare le maschere necessarie per realizzare un transistore MOS in base alle regole di progetto di Figura 2.9.

2.3 Disegnare le maschere necessarie per realizzare un transistore bipolare integrato in base alle regole di progetto di Figura 2.9.

2.4 Calcolare il valore della resistenza integrata di Figura E2.1, assumendo una resistenza di strato di 50 Ω/□. Valutare l'area occupata dalla diffusione assumendo $\lambda = 1\mu m$. Quale sarà l'occupazione di area se il valore della resistenza è di 10 KΩ?

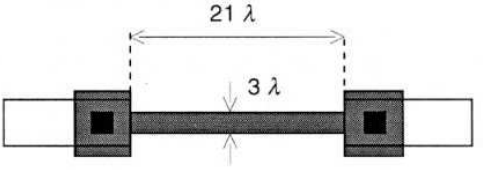


Figura E2.1

2.5 Determinare la capacità unitaria presentata da una giunzione N^+/P con drogaggi $N^+=10^{19}$ cm^{-3}, $P = 10^{15}$ cm^{-3} quando la tensione ai suoi capi è nulla. Qual è il valore della capacità con una tensione di contropolarizzazione V = 5 V?

2.6 Calcolare la capacità offerta da una linea di interconnessione di lunghezza 5 mm e larghezza 3 μm, assumendo uno spessore dell'ossido di campo su cui è depositata la linea di 1 μm.

2.7 Calcolare la resistenza introdotta da una linea di interconnessione in alluminio di larghezza 3 μm e lunghezza 600 μm; valutare l'incremento di resistenza se la stessa linea di interconnessione è realizzata con polisilicio drogato.

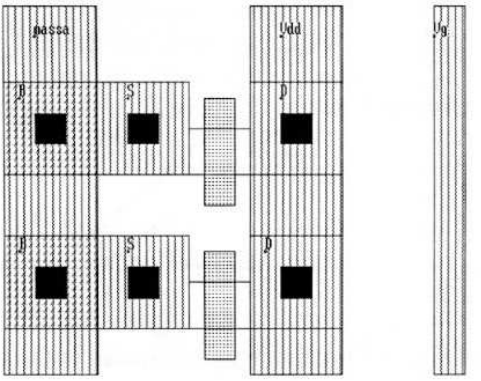


Figura E2.2

2.8 Completare il lay-out di Figura E2.2, collegando le gate dei due transistori MOS, connessi in parallelo, alla linea di metallo indicata con *Vg*, soddisfacendo alle regole di progetto di Tabella 2.2 e riducendo al massimo la lunghezza delle interconnessioni in polisilicio.

Riferimenti bibliografici

R.S. Muller, T.I. Kamins, *Dispositivi elettronici nei circuiti integrati*, Bollati Boringhieri, Torino, 1982.

G. Soncini, *Tecnologie Microelettroniche*, Bollati Boringhieri, Torino, 1986.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

3

Il transistore MOS

3.1 Introduzione

La classificazione delle tecnologie di realizzazione dei circuiti integrati fa essenzialmente riferimento ai dispositivi attivi che sono la parte fondamentale di qualunque circuito e che richiedono la maggior parte dei passi di processo delle tecnologie impiegate; per tale ragione i circuiti integrati vengono classificati secondo la loro realizzazione come integrati in tecnologia MOS o in tecnologia bipolare.

Nei prossimi capitoli si presenteranno i circuiti logici elementari realizzati in tecnologia MOS. Per questi (come per i circuiti realizzati con tecnologia bipolare, introdotti successivamente) è fondamentale la conoscenza del funzionamento dei dispositivi che li compongono, e in particolare la comprensione dei diversi modelli utilizzati per una descrizione del loro comportamento elettrico, sia in regime statico che dinamico. Tale conoscenza è necessaria per diversi aspetti; infatti da un lato occorre disporre di modelli analitici, anche se approssimati, per poter sviluppare analisi semplificate del loro comportamento, fondamentali per una comprensione del funzionamento e delle prestazioni dei circuiti, e necessarie per un primo livello di progettazione dei circuiti elettronici in genere; dall'altro occorre conoscere quali siano i modelli dei dispositivi (di solito più sofisticati di quelli utilizzati nelle analisi manuali) che vengono impiegati nei simulatori circuitali, e quali siano i (numerosi) parametri che definiscono il modello, per i quali bisogna fornire al simulatore i valori specifici che dipendono dalla realizzazione dei dispositivi stessi.

Si presume che i lettori abbiano già adeguate conoscenze della fisica dei dispositivi a semiconduttore e dei loro modelli, di solito fornite in corsi di fisica dei semiconduttori, tuttavia si ritiene utile sintetizzare in forma descrittiva i concetti fondamentali alla base del funzionamento dei dispositivi, e richiamare i modelli ad ampi segnali di prima approssimazione dei dispositivi stessi, che saranno frequentemente utilizzati per studiare il comportamento delle porte logiche elementari e valutare le loro prestazioni. In questo capitolo verranno richiamati i concetti fon-

damentali alla base del funzionamento dei dispositivi MOS, in particolare per quanto riguarda i modelli analitici ad ampi segnali del funzionamento statico, e gli effetti delle capacità, che intervengono significativamente nel funzionamento in transitorio e quindi determinano, nel caso dei circuiti digitali, la risposta temporale ai segnali logici applicati.

3.2 Struttura del transistore MOS

Nel Paragrafo 2.2 si è sinteticamente descritta la sequenza dei processi di fabbricazione di un transistore MOS per circuiti integrati, con riferimento ad una tecnologia con gate in polisilicio (*polysilicon gate*) per strutture CMOS. Va detto che, per applicazioni in cui sono tollerate prestazioni dinamiche meno spinte e dimensioni maggiori dei singoli dispositivi a vantaggio di un costo più basso di fabbricazione, si realizzano ancora MOS con gate in metallo (alluminio), il che giustifica la dizione MOS (*Metallo-Ossido-Semiconduttore*) data alla struttura. Sia nel caso di dispositivi con gate in polisilicio che in quelli con gate in alluminio il principio di funzionamento è lo stesso, cambiando solo in via quantitativa le caratteristiche elettriche, in particolare quelle dinamiche; faremo quindi riferimento ad una struttura schematizzata come in Figura 3.1, con riferimento ad un transistore NMOS, cioè ad un transistore realizzato a partire da un substrato di tipo P e in cui, come vedremo, si induce un canale di tipo N tra le regioni di source e drain. È naturalmente possibile realizzare una struttura duale partendo da un substrato di tipo N e realizzando le regioni di source e drain di tipo P; in tal caso il transistore sarà a canale P e viene indicato come PMOS.

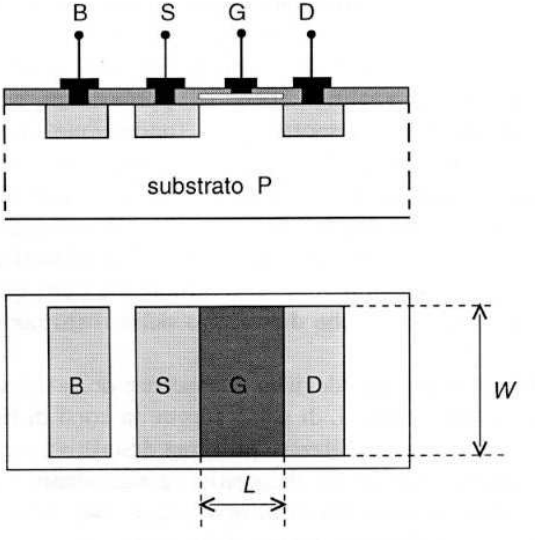


Figura 3.1 Struttura semplificata di un transistore NMOS

Nella Figura 3.1 è schematicamente indicato uno spessore relativamente elevato (circa 1 μm) dell'ossido nel quale vengono realizzate le regioni di source e drain (detto ossido di campo nella sequenza di operazioni riportate nel Paragrafo 2.2), mentre lo spessore dell'ossido di gate (che è lo strato di ossido sottostante il polisilicio o l'alluminio) è molto più sottile (inferiore a 0.1 μm). I terminali elettrici della struttura sono quelli di source e drain, che contattano le rispettive regioni diffuse, e quello di gate, che contatta l'elettrodo di gate (alluminio o polisilicio). Si definisce inoltre substrato (*body*) il substrato di silicio nel quale viene realizzato il transistore MOS e terminale di substrato l'elettrodo che è a contatto con il substrato di silicio non diffuso; in questo caso il terminale di substrato è collegato ad una regione P^+ che effettua un buon contatto ohmico con la fetta (assunta di tipo P) di silicio.

3.3 La tensione di soglia

Si supponga in un primo momento di portare il terminale di substrato allo stesso potenziale di quello di source (si noti che la struttura è simmetrica, per cui i terminali di drain e di source possono essere scambiati), e di applicare una tensione positiva tra drain e source. In assenza di polarizzazione applicata al terminale di gate, tra questi due terminali non può circolare corrente perché la struttura presenta tra questi due terminali due giunzioni N/P e P/N, formate rispettivamente dalle regioni source/substrato e substrato/drain, di cui una è sempre contropolarizzata, qualunque sia il segno della tensione tra drain e source; in questa situazione si definisce il dispositivo come interdetto, in altre parole non può circolare corrente tra i terminali di source e drain ed il circuito elettrico equivalente può essere assimilato a quello di un interruttore aperto (se si trascura la debolissima corrente inversa della giunzione contropolarizzata).

Se ora si applica una tensione sufficientemente elevata (e di segno opportuno, come vedremo) al terminale di gate, si viene a creare un *canale* conduttore tra source e drain, che permette la circolazione di una significativa corrente tra i due terminali. La tensione minima necessaria a formare questo canale di conduzione viene detta *tensione di soglia* V_T (*threshold voltage*) del MOS, ed è la grandezza più rilevante nel funzionamento del dispositivo.

La comprensione degli effetti che sono alla base della creazione del canale indotto nel semiconduttore per effetto della tensione di gate, necessaria per una valutazione quantitativa del valore della tensione di soglia, richiede approfondite conoscenze di fisica dei semiconduttori; in questo contesto ci si limiterà a dare una spiegazione intuitiva del fenomeno, basata sul comportamento della capacità della struttura MOS. In effetti considerando la regione di gate (riportata in Figura 3.2) come un condensatore di cui il polisilicio (o il metallo) di gate è una armatura, l'ossido è l'isolante e il substrato (drogato P in questo caso) è l'altra armatura, se si applica una tensione positiva ma minore di V_T tra gate e substrato (Figura 3.2b), le cariche positive mobili (lacune) vengono allontanate dalla superficie, e si crea una regione di svuotamento, in tutto simile a quella creata dal campo elettrico in una

giunzione P/N (discussa nel Paragrafo 2.4); in questa regione vi sono le cariche fisse negative (dovute agli ioni di drogante P) che vengono bilanciate da un'eguale quantità di cariche positive sull'elettrodo di gate.

Il potenziale positivo applicato alla superficie del silicio, mentre respinge le lacune (portatori maggioritari), attrae i portatori minoritari che sono sempre presenti nel semiconduttore, per cui nelle immediate prossimità della superficie la densità di elettroni (portatori minoritari in questo caso) cresce all'aumentare della tensione V_{GS} fino al punto in cui la concentrazione degli elettroni alla superficie del silicio (limitatamente alla regione al di sotto dell'elettrodo di gate) diviene uguale a quella delle lacune nel substrato in assenza di potenziale alla gate (Figura 3.2c).

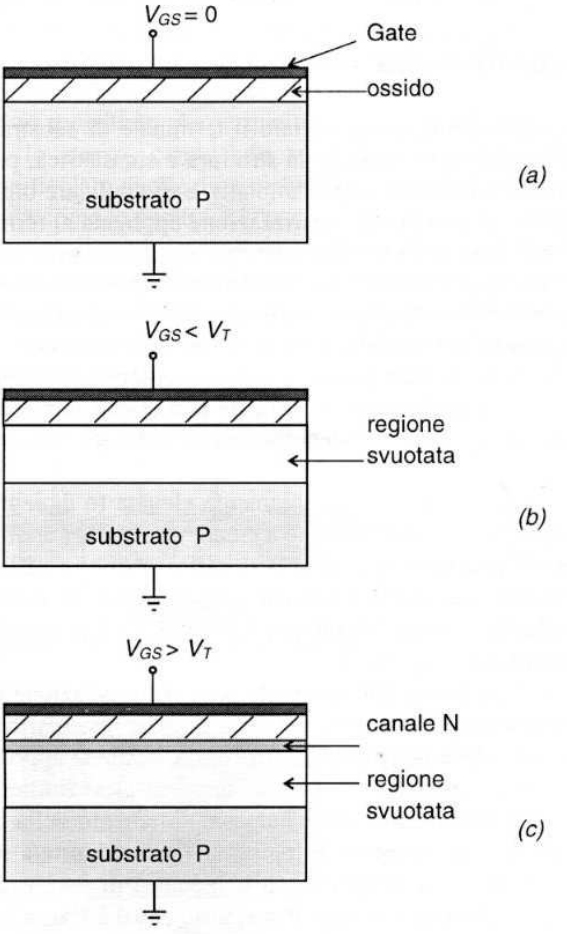


Figura 3.2 Comportamento della struttura MOS: a) in assenza di polarizzazione; b) con polarizzazione sotto la soglia; c) con polarizzazione sopra la soglia

In queste condizioni si dice che si è raggiunta la condizione di *inversione* del semiconduttore alla superficie; in altre parole si venuto a creare alla superficie un sottile strato di materiale di tipo N (dovuto agli elettroni che in questa regione hanno una concentrazione uguale a quella delle lacune nel resto del materiale, mentre, sempre nella stessa regione, le lacune sono ridotte ad un livello trascurabile rispetto a questi). Questa regione, dello stesso segno di quello delle regioni di source e di drain, in effetti costituisce il canale di cui si è detto, e permette il passaggio della corrente di tipo ohmico tra source e drain, in quanto a causa della presenza del canale N non vi sono più le barriere di potenziale delle giunzioni N/P e P/N che ostacolano il cammino delle cariche.

La tensione di soglia V_T che permette la creazione di questo canale può essere definita in via approssimata come la tensione da applicare alla gate affinché sulla superficie del silicio si raggiunga il potenziale critico necessario a dar luogo alla inversione del semiconduttore, valore indicato con ϕ^*. Quindi il valore della tensione di soglia V_T è dato dalla somma (algebrica) delle seguenti componenti:

a. la differenza di potenziale di contatto $\Phi_{GS} = \phi_G - \phi_S$ tra il materiale di gate e il substrato di silicio;
b. il potenziale critico alla superficie ϕ^*;
c. la tensione necessaria ad equilibrare la carica degli atomi droganti Q_{SI} nella regione di silicio svuotata delle cariche maggioritarie, tensione che è data da $-Q_{SI}/C_{OX}$ (assumendo la carica Q_{SI} positiva, nel caso di un substrato di tipo N).

In pratica, la tensione di soglia dipende anche dalla carica presente all'interfaccia ossido/silicio, dove termina il reticolo cristallino del silicio e si creano delle cariche localizzate, e all'eventuale carica intrappolata nell'ossido, dipendente dai processi di realizzazione dell'ossido di gate; quindi è da considerare un'ulteriore componente necessaria ad equilibrare la carica Q_{OX} eventualmente presente nell'ossido, data da $-Q_{OX}/C_{OX}$. Si ha in definitiva per la tensione di soglia l'espressione approssimata:

$$V_T = \Phi_{GS} + \phi^* - \left(\frac{Q_{SI}}{C_{OX}} + \frac{Q_{OX}}{C_{OX}} \right) \tag{3.1}$$

Queste componenti dipendono (in valore e in segno) dal tipo di materiale della gate, dal tipo di substrato e dal drogaggio di substrato. Nella struttura con substrato P, considerata nella Figura 3.2, la carica Q_{SI} è negativa, mentre la carica Q_{OX} è positiva e di valore assoluto molto minore della prima, per cui nella Equazione (3.1) il termine di carica più rilevante è quello relativo alla carica della regione di svuotamento Q_{SI}, che dipende dal drogaggio della regione della superficie. È quindi possibile modificare la tensione di soglia rispetto al valore definito dal drogaggio di substrato realizzando una debole impiantazione di drogante nel canale; dalla Equazione (3.1) si deduce che, in un substrato di tipo P, impiantando una dose ridotta di drogante P si aumenta il valore della V_T, mentre impiantando un drogante di tipo N in un substrato N si diminuisce il valore (negativo) della V_T.

La tensione di soglia dipende in ogni caso dallo spessore dell'ossido (attraverso la relazione $C_{OX} = \varepsilon_{OX}/t_{OX}$ nella Equazione (3.1)), per cui al diminuire dello spessore dell'ossido diminuisce anche la V_T, anche se in maniera più debole di quanto appaia dalla Equazione (3.1), a causa degli effetti delle cariche nell'ossido e all'interfaccia.

Nel valutare la tensione di soglia determinata dalla Equazione (3.1) si è assunto che il terminale di substrato sia allo stesso potenziale di quello di source. Nei circuiti integrati tuttavia non è possibile connettere i source di tutti i transistori al substrato, perché in tal caso i transistori avrebbero tutti i terminali di source connessi in parallelo. D'altra parte nei chip NMOS il substrato deve essere posto al più negativo tra i potenziali presenti nel circuito, per garantire che le giunzioni tra le regioni di isolamento dei vari componenti e il substrato stesso siano sempre contropolarizzate (per i circuiti in tecnologia PMOS valgono le considerazioni duali per cui il substrato deve essere al più alto potenziale tra quelli presenti nel circuito).

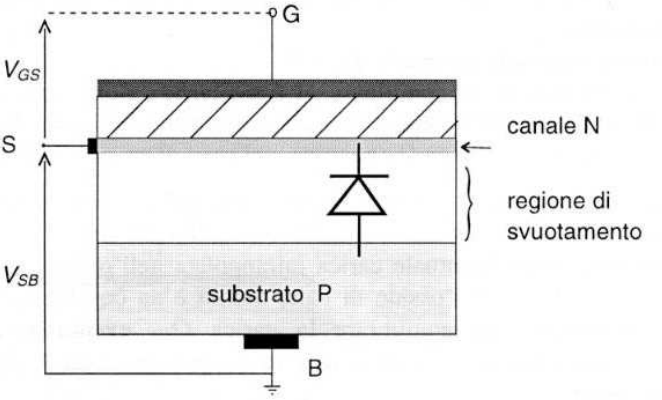


Figura 3.3 Effetto della tensione tra source e substrato sulla tensione di soglia

Questo vincolo porta ad avere un potenziale positivo (negativo) tra source e substrato per qualcuno dei transistori nei chip NMOS (PMOS), e ciò comporta un aumento della tensione di soglia corrispondente (per il PMOS l'aumento è del modulo della tensione essendo questa negativa). Una spiegazione intuitiva di questo aumento può essere data identificando la regione tra il canale di inversione (collegato elettricamente al source) e il substrato come quella di una giunzione P-N che si forma tra il substrato non svuotato (di tipo P) e il canale N (con cariche mobili negative), separate dalla regione di svuotamento indotta dalla tensione di gate, come è indicato in Figura 3.3. Per effetto del potenziale positivo applicato tra source e substrato la regione di svuotamento si allarga, e quindi la carica Q_{SI} aumenta, incrementando la tensione di soglia. Per tener conto di questo effetto, indicato come *effetto substrato* (*body Effect*), si può correggere l'espressione della tensione di soglia data dall'Equazione (3.1), secondo l'espressione seguente, che definisce il valore di V_T in presenza di una tensione V_{SB}:

$$V_T = V_{T0} + \gamma \cdot \left(\sqrt{|\phi^* + V_{SB}|} - \sqrt{|\phi^*|} \right) \tag{3.2}$$

dove V_{T0} è il valore dato dall'Equazione (3.1), V_{SB} la tensione tra source e substrato, e γ un fattore, detto coefficiente dell'effetto di substrato, variabile tra 0.2 e 1 $V^{1/2}$ (l'espressione è valida sia per transistori NMOS che PMOS tenendo conto dei segni dei parametri per i due casi). In Tabella 3.1 sono riportati i segni e i valori tipici delle diverse grandezze che intervengono nella tensione di soglia, per transistori NMOS (substrato di tipo P) e PMOS (substrato di tipo N).

Tabella 3.1

parametro	*valore*	*segno*	
substrato		P	N
Φ_{GS} (alluminio)		– (0.9)	– (0.3)
Φ_{GS} (polisilicio)		– (0.9)	– (0.2)
ϕ^*	0.6-0.7	+	–
Q_{SI}	$2\text{-}5 \cdot 10^{-8}$	–	+
Q_{OX}	10^{-9}	+	+
V_{SB}	-	+	–
γ	0.3-0.7	+	–
V_T	0.6-1	+	–

La tensione di soglia, sia per i transistori NMOS che PMOS, dipende dalla temperatura; in pratica il suo valore assoluto diminuisce con la temperatura con un coefficiente di circa –3 mV/°C.

3.4 Caratteristiche corrente-tensione

Una volta creatosi il canale tra source e drain, la corrente può fluire tra queste due regioni se è applicata una tensione V_{DS} tra gli elettrodi corrispondenti. Questa corrente dipende non solo dalla V_{DS}, ma anche dalla tensione V_{GS} tra gate e source, che determina il numero di cariche mobili disponibili nel canale.

Una espressione semplificata della relazione tra la corrente di drain I_D e le tensioni V_{DS}, V_{GS} può essere ottenuta dallo schema di Figura 3.4, facendo riferimento alle cariche indotte nel canale dalla capacità della struttura MOS. La corrente che fluisce nel canale, di resistenza finita, provoca una caduta di potenziale variabile tra source e drain; ne consegue che il potenziale sulla superficie del silicio (e cioè sull'armatura inferiore del condensatore MOS) varia con l'ascissa y, aumentando verso il drain. La carica $Q_C(y)$ nel canale (carica per unità di larghezza W del canale stesso) può essere scritta, ricordando che questa viene creata solo per tensioni in

eccesso della tensione di soglia V_T, come il prodotto della capacità dell'ossido per la differenza di tensione (in eccesso di V_T) tra gate e canale:

$$Q_C(y) = C_{OX}\left[(V_{GS} - V_T - V(y)\right] \tag{3.3}$$

La resistenza dR dell'elemento di canale dy all'ascissa y è data da:

$$dR = \frac{dy}{W\mu_n Q_C(y)} \tag{3.4}$$

e la caduta di potenziale $dV(y)$ sull'elemento dR, dalla Equazione (3.4) è data da:

$$dV(y) \equiv I_D dR = \frac{I_D dy}{W\mu_n Q_C(y)} \tag{3.5}$$

L'Equazione (3.5) è un'equazione differenziale a variabili separabili, che integrata fornisce il legame voluto:

$$\int_0^L I_D dy = \int_0^{V_{DS}} W\mu_n C_{OX}(V_{GS} - V_T - V(y))dV(y) \tag{3.6}$$

$$I_D = \frac{1}{2}\mu_n C_{OX}\frac{W}{L}\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] = K\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] \tag{3.7}$$

dove il termine K, definito come *trasconduttanza del dispositivo*, va considerato, a parità di tensioni applicate, come un fattore di scala per la corrente, che dipende dal *fattore di forma* (*aspect ratio*) W/L della regione di gate del MOS, e dallo spessore dell'ossido di gate, attraverso C_{OX}.

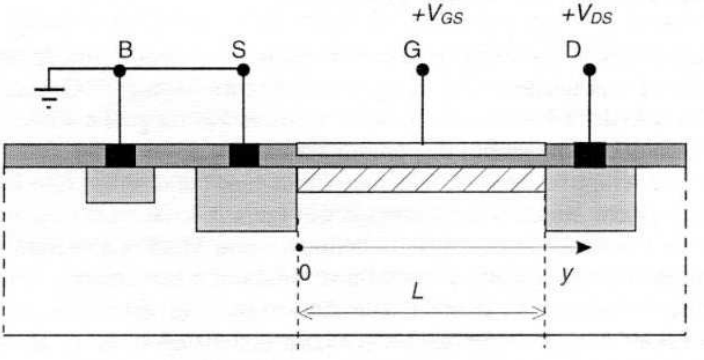


Figura 3.4 Caratteristiche *I-V* del transistore NMOS

Spesso conviene separare la dipendenza dalla geometria del dispositivo da quella legata al processo tecnologico, considerando una trasconduttanza di processo k' funzione della mobilità dei portatori nel canale e dello spessore dell'ossido:

$$k' = \frac{1}{2}\mu_n C_{OX} = \frac{1}{2}\mu_n \frac{\varepsilon_{OX}}{t_{OX}}; \quad K = k'\frac{W}{L} \tag{3.8}$$

Il valore del parametro k' dipende essenzialmente dallo spessore dell'ossido di gate e dal valore della mobilità μ dei portatori, e quindi k' sarà diverso per un dispositivo NMOS o per uno PMOS in dipendenza della differente mobilità degli elettroni (μ_n) o lacune (μ_p); per uno spessore di ossido di gate t_{OX} = 200 nm, k' vale circa 50 $\mu A/V^2$ per un NMOS e 20 $\mu A/V^2$ per un PMOS.

L'espressione di I_D fornita dall'Equazione (3.7) vale solo se in ogni punto del canale l'eccesso di tensione $V_{GS} - V(y) > V_T$; ciò comporta che il massimo valore di $V(y) = V_{DS}$ sia ancora tale da soddisfare la disequazione precedente.

Per valori di $V_{DS} > V_{GS} - V_T$ una parte del canale è in condizioni di *strozzamento* (*pinch-off*) ossia il canale scompare a partire dall'ascissa L' per la quale $V(L') = V_{GS} - V_T$, come è riportato in Figura 3.5. In questo caso l'Equazione (3.6) va integrata solo fino alla lunghezza L' del canale effettivamente formato sul quale cade una tensione fissa $V_{GS} - V_T$, mentre l'eccesso di tensione di drain rispetto a questo valore cade sulla estensione più o meno grande della regione L–L' di pinch-off. La corrente I_D in tal caso non risulta più dipendente dalla V_{DS}, in quanto la tensione ai capi del canale ohmico risulta costante; le cariche mobili che si muovono nel canale, raggiunta l'ascissa L', vengono iniettate nella regione svuotata L–L' e mosse dal campo elettrico esistente in questa regione fino a raggiungere il drain.

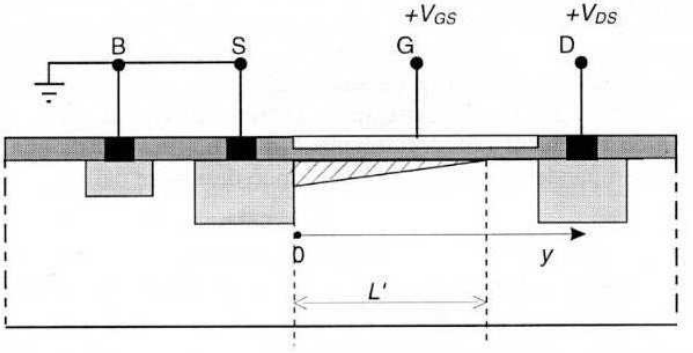


Figura 3.5 Condizione di pinch-off del transistore NMOS

In effetti, poiché il canale si è ridotto ad una lunghezza effettiva L_{eq}, la corrente I_D dopo il pinch-off aumenta leggermente a causa del ridursi del termine L_{eq} nell'espressione del fattore di scala K della Equazione (3.8). In queste condizioni la

corrente I_D è quella al limite del pinch-off, ottenuta ponendo: $V_{DS} = V_{GS} - V_T$ nella Equazione (3.7), ma corretta da un termine proporzionale alla tensione V_{DS} che tiene conto dell'effetto di modulazione del canale L':

$$I_D = K(V_{GS} - V_T)^2 \cdot (1 + \lambda V_{DS}) \tag{3.9}$$

Nel seguito, per le analisi dei circuiti digitali, che prevedono tensioni di drain relativamente basse, si trascurerà questa debole dipendenza di I_D da V_{DS} oltre il pinch-off, poiché il coefficiente λ ha valori contenuti (0.01 - 0.1 V^{-1}). La corrente di drain del MOS risulta quindi espressa da due relazioni differenti, a seconda del valore delle tensioni V_{DS} e V_{GS}:

$$I_D = K\left[2(V_{GS} - V_T)V_{DS} - V_{DS}^2\right] \quad \left| \text{ per } V_{DS} \leq V_{GS} - V_T \right. \tag{3.10a}$$

$$I_D = K(V_{GS} - V_T)^2 \quad \left| \text{ per } V_{DS} \geq V_{GS} - V_T \right. \tag{3.10b}$$

I due campi di validità sono separati dalla condizione limite del pinch-off $V_{DS} = V_{GS} - V_T$, che, sostituita nella Equazione (3.11), definisce una curva limite nel piano I_D - V_{DS} tra le due parti della caratteristica complessiva:

$$V_{DS} = V_{GS} - V_T \qquad \text{(condizione di pinch-off)}$$

$$I_{D\text{lim}} = K(V_{DS\text{lim}})^2 \qquad \text{(curva limite di pinch-off)}$$

In Figura 3.6 è riportata una famiglia di caratteristiche I-V di un dispositivo NMOS in base alle Equazioni (3.10). In essa si può identificare una regione, detta *nonlineare* o *di triodo*, in cui vale la relazione (3.10a), e una regione detta di *pinch-off* o *di saturazione*, in cui le caratteristiche non dipendono più dalla V_{DS} (Equazione (3.10b)). La prima parte della regione di triodo viene anche detta *regione ohmica* perché il comportamento è quello (lineare) di una resistenza R_{MOS} di valore dipendente dalla V_{GS}, come si può desumere dalla (3.10a) per $V_{DS} << (V_{GS} - V_T)$:

$$R_{MOS} = \frac{V_{DS}}{I_D} = \frac{1}{K \cdot 2(V_{GS} - V_T)} \tag{3.11}$$

Il funzionamento di un transistore PMOS (MOS a canale P) è del tutto simile a quello visto per il caso del transistore NMOS, purché si scambi, in tutto quanto detto precedentemente, il drogaggio delle regioni P con N e viceversa. Infatti nel PMOS, sempre a causa del segno delle cariche nella regione di svuotamento e delle cariche nel canale, la tensione di soglia sarà negativa, come anche sarà negativa la corretta polarizzazione di drain; la corrente in un

PMOS fluisce quindi dal source al drain; le caratteristiche I_D-V_{DS} giacciono quindi nel terzo quadrante, ma usualmente vengono riportate ancora nel primo quadrante facendo riferimento ai valori assoluti sia delle tensioni che delle correnti.

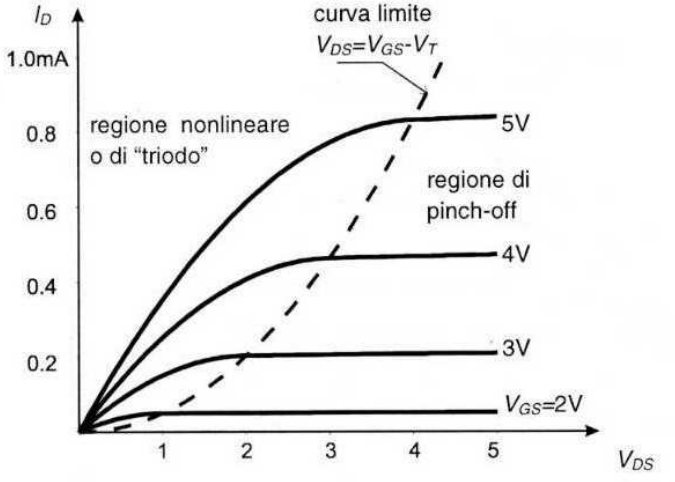


Figura 3.6 Caratteristiche *I-V* per un transistore NMOS con $k' = 10\text{mA/V}^2$ e rapporto di forma $W/L = 2$

I dispositivi precedentemente descritti vengono chiamati MOS ad *arricchimento* (*enhancement*), poiché il canale di conduzione viene creato ed arricchito di cariche per effetto della tensione di gate quando questa supera il valore di soglia. È tuttavia possibile, con un opportuno step tecnologico nel processo di fabbricazione (impiantazione prima della formazione dell'ossido di gate), creare una sottile regione superficiale tra source e drain, con drogaggio di segno opposto a quello del substrato; in tal caso il canale di conduzione è presente anche a tensione di gate nulla, e il dispositivo è normalmente in conduzione, a meno che non si applichi una tensione di gate di segno tale da respingere le cariche mobili del canale, e quindi svuotare il canale stesso e impedire la conduzione.

I dispositivi di questo tipo vengono detti MOS a *svuotamento* (*depletion*) e la tensione necessaria per il completo strozzamento del canale viene detta *tensione di punch-through* V_p. Questa tensione è di segno opposto a quella di soglia per un transistore con lo stesso tipo di canale ma del tipo ad arricchimento (ad esempio negativa per un transistore NMOS a svuotamento); a partire da questa tensione tuttavia la dipendenza della corrente di drain dalle tensioni di gate e di drain è la stessa di quella discussa per il MOS ad arricchimento, in quanto il meccanismo di controllo è lo stesso per entrambi i dispositivi. Si utilizzeranno quindi espressioni analoghe a quelle (3.10-3.11) del MOS ad arricchimento, in questo caso riferite ad una tensione di soglia equivalente V_{TD} pari

alla tensione di punch-through V_P; quest'ultima è, come abbiamo detto, negativa, e di valore maggiore, in modulo, della tensione di soglia di transistori MOS ad arricchimento, assumendo valori dell'ordine di – 3 V.

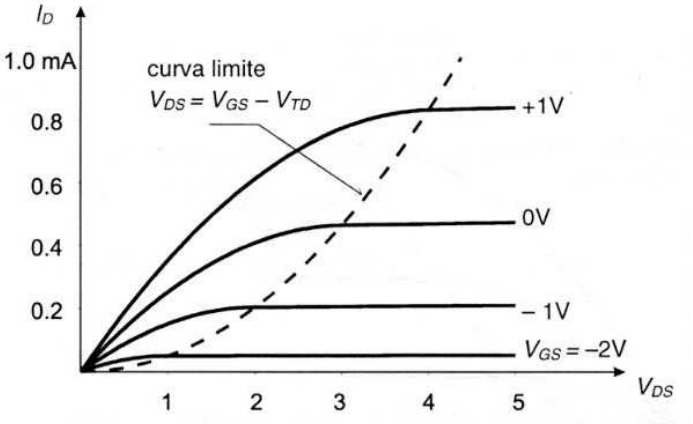


Figura 3.7 Caratteristiche *I-V* di un NMOS a svuotamento con $k' = 10\text{mA/V}^2$ e rapporto di forma $W/L = 2$

Nelle (3.12) e (3.13) sono riportate le relazioni valide per un NMOS a svuotamento, dove il parametro K assume lo stesso significato della Equazione (3.8):

$$I_D = K\left[2(V_{GS} + |V_{TD}|)V_{DS} - V_{DS}^2\right] \quad \text{per } V_{DS} \leq V_{GS} + |V_{TD}| \tag{3.12}$$

$$I_D = K(V_{GS} + |V_{TD}|)^2 \quad \text{per } V_{DS} \geq V_{GS} + |V_{TD}| \tag{3.13}$$

In Figura 3.7 sono riportate le caratteristiche *I-V* di uscita di un NMOS a svuotamento, con una tensione di soglia pari a $V_{TD} = -3$ V.

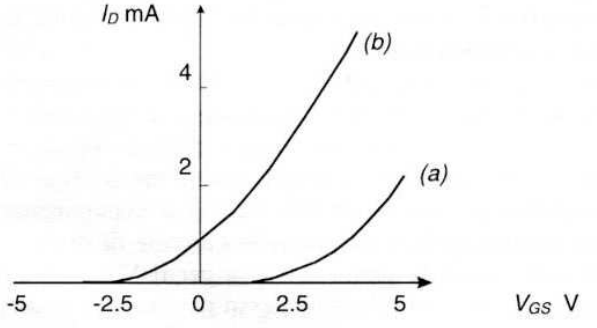


Figura 3.8 Caratteristiche di trasferimento di NMOS: a) ad arricchimento; b) a svuotamento

Per i dispositivi attivi si definiscono oltre alle caratteristiche I_D-V_{DS}, dette *caratteristiche di uscita*, anche le *caratteristiche di trasferimento* I_D-V_{GS}, che legano un parametro di uscita con uno di ingresso. Queste caratteristiche, di solito considerate nella sola regione di pinch-off, sono espresse dalle Equazioni (3.10b) o (3.13) rispettivamente per i MOS ad arricchimento o a svuotamento, e sono riportate in Figura 3.8 per il caso di MOS a canale N.

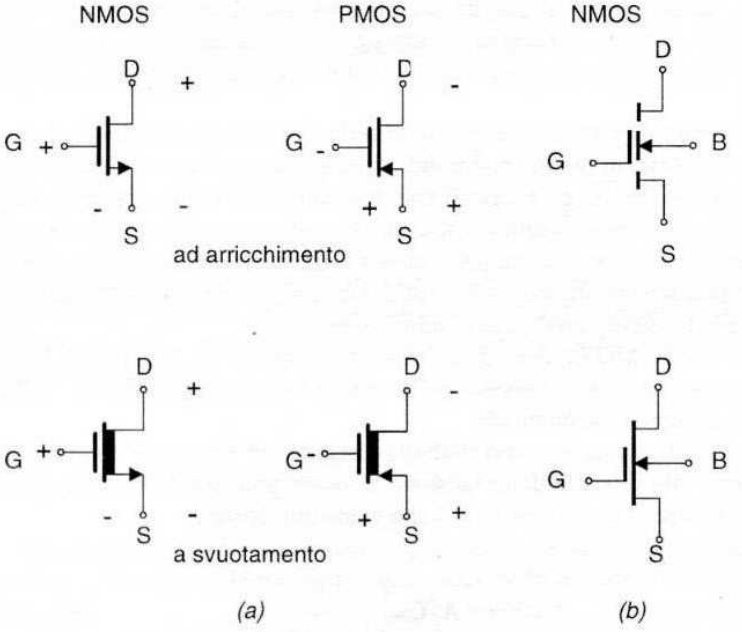


Figura 3.9 a) Simboli elettrici semplificati e polarità delle tensioni per i diversi transistori MOS; b) esempio di simboli (per i transistori NMOS) che indicano esplicitamente il terminale di substrato

In definitiva sono possibili quattro tipi differenti di dispositivi MOS in dipendenza del segno delle cariche del canale e della modalità di formazione del canale stesso, e cioè MOS ad arricchimento (a canale N o P) e MOS a svuotamento (a canale N o P). I simboli elettrici e le polarità delle tensioni ai terminali sono indicate in Figura 3.9a, per transistori in cui non occorre esplicitare il terminale di substrato (o questo è direttamente connesso al terminale di source), mentre in Figura 3.9b sono riportati i simboli, per transistori NMOS sia ad arricchimento che a svuotamento, in cui è indicato anche il terminale di substrato. Si noti che il verso della freccia sul terminale di source indica anche il verso della corrente circolante tra il source e il drain; la freccia sul terminale di substrato indica il verso della (ipotetica) corrente circolante nella giunzione P/N (per transistori NMOS) o N/P (per transistori PMOS) che si crea tra substrato e canale.

3.5 Capacità del dispositivo

I modelli che descrivono i dispositivi attivi debbono includere anche i componenti che ne influenzano il comportamento in regime dinamico, e che limitano quindi le loro prestazioni nei transitori di commutazione. Nei modelli semplificati che vengono utilizzati per una descrizione analitica del comportamento dinamico dei dispositivi viene usualmente assunta l'ipotesi di comportamento *quasi-stazionario* del dispositivo; in altre parole si assume che lo spostamento da un punto di funzionamento all'altro avvenga attraverso il passaggio per una serie di stati equivalenti a quelli che si assumerebbero in regime stazionario con opportune condizioni di polarizzazione.

Questo comporta che la ridistribuzione delle cariche mobili nei vari stati di funzionamento avvenga in tempi trascurabili rispetto a quelli necessari per la commutazione. In queste ipotesi gli elementi che determinano il comportamento dinamico dei dispositivi sono assimilabili a capacità che vengono poste tra gli elettrodi del dispositivo stesso, e che determinano, insieme agli elementi nonlineari che descrivono il comportamento statico, delle costanti di tempo equivalenti che agiscono sui transitori delle tensioni variabili ai terminali stessi.

Per i transistori MOS queste capacità sono di due tipi: a) le capacità delle giunzioni P/N contropolarizzate presenti nella struttura e b) le capacità legate alla struttura metallo-ossido-semiconduttore.

Le capacità di giunzione sono state già discusse nel Paragrafo 2.4, dove si è visto che la capacità dipende dalla tensione di contropolarizzazione della giunzione stessa, espressa dalla Equazione (2.11) che viene qui ripetuta:

$$C_J(V) = A \cdot C_{J0}\left(\frac{1}{1+V/\phi_0}\right)^{1/2} \tag{3.14}$$

dove C_{J0} è la capacità per unità di area in assenza di contropolarizzazione, A è l'area della giunzione e V è la tensione di contropolarizzazione della giunzione stessa (si ricorda che il valore di V da utilizzare è quello in modulo).

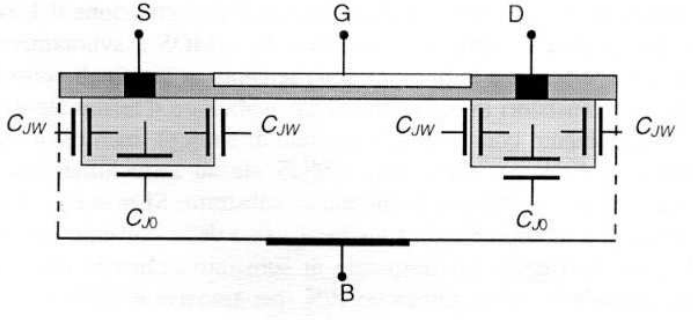


Figura 3.10 Capacità delle regioni di source e drain

Nel caso del transistore MOS le giunzioni contropolarizzate sono quelle di drain/substrato e di source/substrato, per cui nel modello elettrico del dispositivo compariranno le capacità C_{SB} e C_{DB}, dipendenti dal valore della capacità unitaria C_{J0} di giunzione, moltiplicata rispettivamente per l'area della regione di source e di drain, e funzioni rispettivamente delle tensioni V_{SB} e V_{DB}. Oltre a queste componenti si definisce una capacità laterale corrispondente alla parte laterale della giunzione, di area pari al perimetro della regione P_S (o P_D) per la profondità X_J della giunzione stessa (vedi Figura 3.10). Definendo una capacità laterale unitaria C_{JW} per unità di lunghezza perimetrica (espressa in fF/μm), le capacità totali di source e di drain saranno date da:

$$C_{SB(DB)} = (C_{J0} \cdot A_{S(D)} + C_{JW} \cdot P_{S(D)}) \left(\frac{1}{1 + V_{SB(DB)} / \phi_0} \right)^{1/2} \tag{3.15}$$

Oltre alle capacità delle regioni di source e drain si debbono considerare le capacità dipendenti dalla struttura M-O-S del gate. Queste in generale, in assenza di polarizzazione dell'elettrodo di gate, si definiscono secondo la semplice relazione già vista nel Paragrafo 2.2:

$$C_{OX} = A \frac{\varepsilon_{OX}}{t_{OX}}$$

dove A è l'area dell'ossido interessata e t_{OX} lo spessore dell'ossido di gate (i valori delle costanti che intervengono nelle espressioni delle capacità sono riportate nella Tabella 2.2).

Nel caso del transistore MOS, la determinazione delle capacità legate alla gate è complicata sia dalla presenza di una tensione di gate V_G che dalle tensioni applicate agli altri elettrodi di source, drain e substrato della struttura in esame.

L'effetto della polarizzazione dell'elettrodo di gate sulle cariche mobili del substrato, presentato nel Paragrafo 3.3, e legato alla presenza di un ossido molto sottile sotto l'elettrodo di gate, induce una forte dipendenza della capacità tra gate e substrato dalla tensione V_G. Possiamo fare di nuovo riferimento all'analisi qualitativa riportata in Figura 3.2 che mostra le diverse situazioni indotte nel substrato all'aumentare della tensione V_G, per valutarne gli effetti sulla capacità di gate C_{GB}.

Per una tensione V_G molto minore di quella di soglia, la superficie del substrato di silicio non è svuotata delle cariche mobili e quindi la regione isolante del condensatore è confinata allo spessore dell'ossido di gate (molto sottile), per cui la capacità di gate è quella C_{OX} definita dalla Equazione (3.15) (relativamente grande). Se V_G aumenta (restando sempre sotto soglia), il potenziale alla superficie induce una regione di svuotamento nel substrato, e quindi lo spessore della regione isolante aumenta, riducendo la capacità complessiva, che può essere vista come la

serie della capacità dell'ossido C_{OX} e della capacità della sola regione di svuotamento C_{DEPL}, ossia: $C_{GB} = \dfrac{C_{OX} C_{DEPL}}{C_{OX} + C_{DEPL}}$.

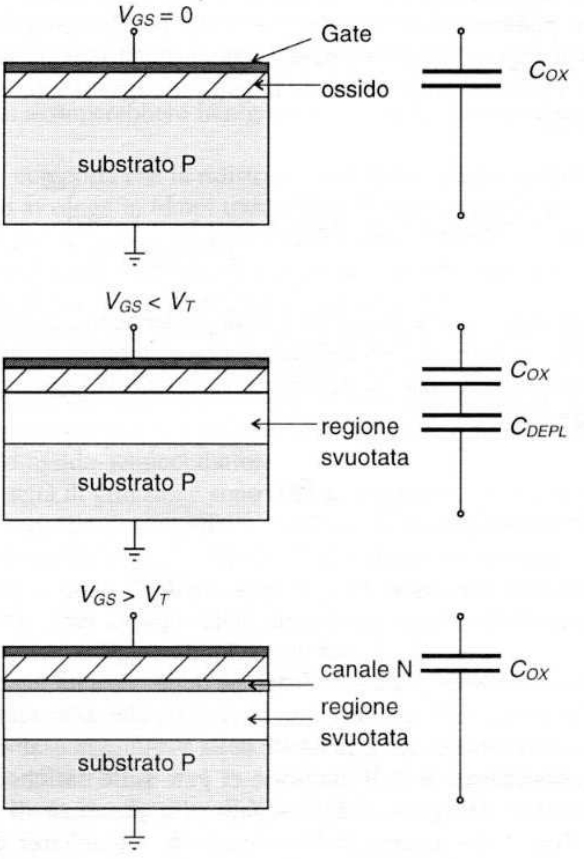


Figura 3.11 Dipendenza della capacità C_{GB} dalla tensione di gate

Poiché la regione di svuotamento in prossimità della soglia ha uno spessore molto maggiore di quello dell'ossido, il valore della capacità complessiva di gate si riduce rispetto al valore C_{OX}. Se V_G aumenta oltre il valore di soglia, si induce alla superficie uno strato di cariche N di concentrazione circa uguale a quella (di tipo P) del substrato, che agisce da strato conduttore; se questo è in collegamento elettrico con il terminale di substrato, la capacità C_{GS} (per basse frequenze) ritorna al valore C_{OX}. Se tuttavia si valuta la capacità a frequenza elevata, i portatori minoritari non hanno il tempo di ridistribuirsi in funzione della elevata velocità di variazione della tensione, e la capacità C_{GB} rimane al valore basso.

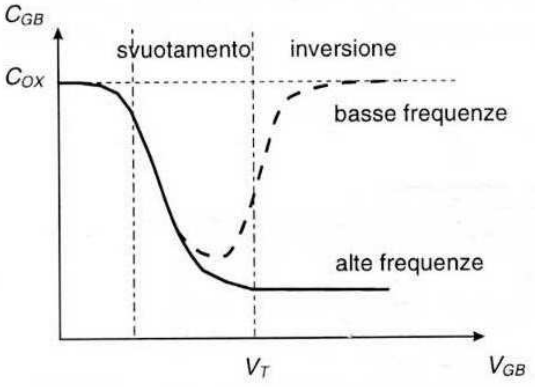


Figura 3.12 Dipendenza della capacità di una struttura MOS al variare della tensione di gate

Se la struttura MOS è quella della regione di gate di un transistore MOS, la capacità complessiva relativa al terminale di gate si divide in tre componenti, relative ai terminali di source, drain e substrato, e le tre componenti capacitive sono in questo caso funzione non solo della tensione di gate V_G, ma anche delle tensioni ai terminali di source e drain, come è indicato in Figura 3.13.

In assenza di canale, cioè con polarizzazione della gate sotto la tensione di soglia V_T, la capacità predominante è quella tra gate e substrato (body), poiché quest'ultimo non è schermato dal canale, e le regioni di source e drain sono isolate dal substrato.

In condizione di canale formato e polarizzazione in regime lineare, il canale si estende con continuità tra source e drain, e l'elevata carica di inversione presente nel canale corrisponde all'armatura metallica inferiore del condensatore; quest'armatura equivalente è collegata con i due elettrodi di source e drain, e scherma completamente il terminale di substrato, quindi la capacità C_G si divide in parti uguali tra source e drain. Infine in condizione di pinch-off il canale è collegato solo al source, per cui la capacità tra gate e drain è trascurabile; il substrato è ancora praticamente schermato dalla presenza del canale.

Ne risulta, come è indicato in Figura 3.13, che la capacità totale di gate C_G rimane essenzialmente costante, pur dividendosi fra i tre terminali in funzione del modo di funzionamento del MOS.

Oltre alle componenti discusse precedentemente, vi è un'altra componente per le capacità tra gate e source, gate e drain, e gate e substrato, dovuta alla inevitabile sovrapposizione (*overlap*) di una regione dL tra il polisilicio della gate e le regioni corrispondenti di source, di drain o di substrato (come è indicato in Figura 3.14 per le prime due), per cui si definiscono tre ulteriori componenti C_{GSO}, C_{GDO} e C_{GBO}, pari rispettivamente a:

$$C_{GSO} = C_{GDO} = W \cdot dL \cdot \frac{\varepsilon_{OX}}{t_{OX}}; C_{GBO} = L \cdot dW \cdot \frac{\varepsilon_{OX}}{t'_{OX}} \tag{3.16}$$

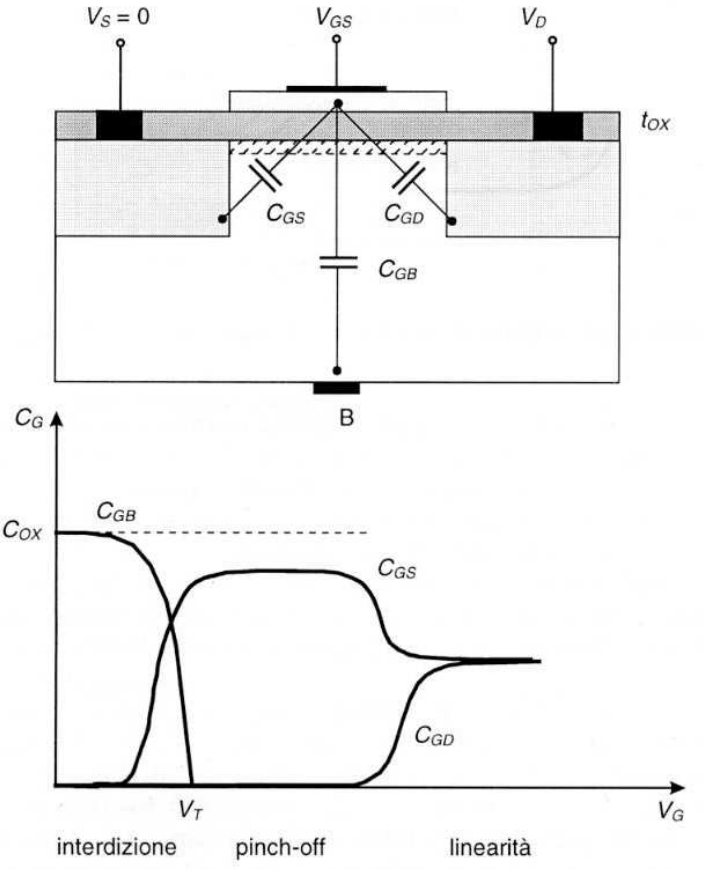


Figura 3.13 Componenti della capacità di gate di un transistore MOS

Si noti che le prime due capacità possono essere valutate moltiplicando le capacità unitarie (definite in fF/μm) per la larghezza W della regione di sovrapposizione, mentre la capacità C_{GBO}, che dipende dalla sovrapposizione dell'estensione del film di polisilicio oltre la larghezza W del canale sull'ossido di campo, che ha spessore maggiore t'_{OX}, viene valutata moltiplicando la capacità unitaria per la lunghezza L del canale di gate.

In definitiva, il modello dinamico del MOS a cui si farà riferimento prevede una rete di capacità presenti tra i diversi elettrodi, secondo lo schema di Figura 3.15, di cui quelle C_{SB}, C_{DB}, espresse dalle (3.14) e (3.15), sono dipendenti dalla tensione applicata alla giunzione, mentre quelle C_{GS}, C_{GB}, C_{GD}, (non considerando

le componenti di sovrapposizione definite dalla (3.16)), sono dipendenti in modo complesso dalla condizione di funzionamento secondo quanto riportato in Figura 3.13, e tali che la capacità totale C_G è, in buona approssimazione, data da:

$$C_G = C_{GS} + C_{GB} + C_{GD} = W \cdot L \cdot \frac{\varepsilon_{OX}}{t_{OX}} \tag{3.17}$$

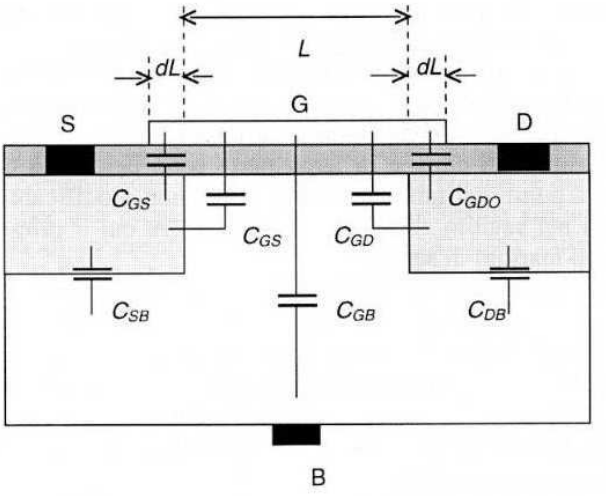


Figura 3.14 Capacità presenti nella struttura del transistore MOS

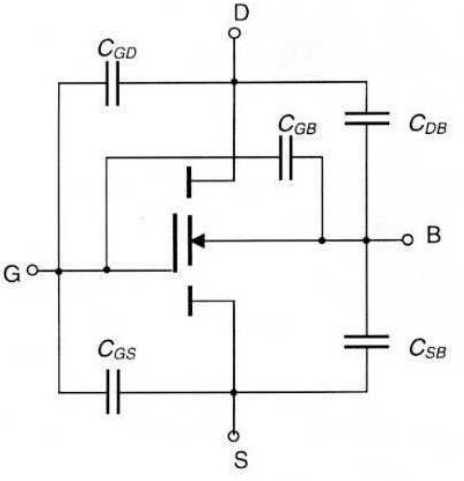


Figura 3.15 Capacità tra i terminali di un transistore NMOS

3.6 Tracciato del dispositivo MOS

I modelli del MOS, a partire da quelli analitici elementari presentati nei paragrafi precedenti, fino a quelli più sofisticati utilizzati nei simulatori circuitali, richiedono la conoscenza della geometria del dispositivo, ed in particolare i tracciati (*lay-out*) delle diverse regioni che lo compongono e che vengono trasferiti sul silicio attraverso le maschere di progettazione. Le geometrie di questi tracciati infatti determinano sia i valori statici (ad esempio il fattore di scala K delle caratteristiche I-V attraverso i valori di L e W), che le varie capacità. Nei modelli intervengono anche parametri non determinabili sulla base della conoscenza dei tracciati delle maschere, come ad esempio lo spessore dell'ossido, i drogaggi delle regioni impiantate e del substrato, parametri che influenzano il valore della tensione di soglia e delle capacità per unità di area delle giunzioni di isolamento. Questi parametri sono tuttavia caratteristici del processo impiegato per la realizzazione dei chip, e in larga parte sono dettati da condizioni tecnologiche messe a punto dal fabbricante del chip e non modificabili, per cui verranno considerati nel seguito come dati di ingresso su cui il progettista del circuito non ha grossi margini di scelta.

Le geometrie delle diverse regioni del dispositivo sono invece grandezze su cui il progettista può operare, e che in larga parte determineranno le prestazioni sia statiche che dinamiche del dispositivo stesso, e quindi del circuito elettrico da questi composto. Le uniche limitazioni che debbono essere tenute in conto nel definire il tracciato del dispositivo sono le *regole di progetto* richiamate nel Paragrafo 2.6, che definiscono le geometrie minime e le tolleranze tra le varie regioni che sono permesse dalla tecnologia a disposizione.

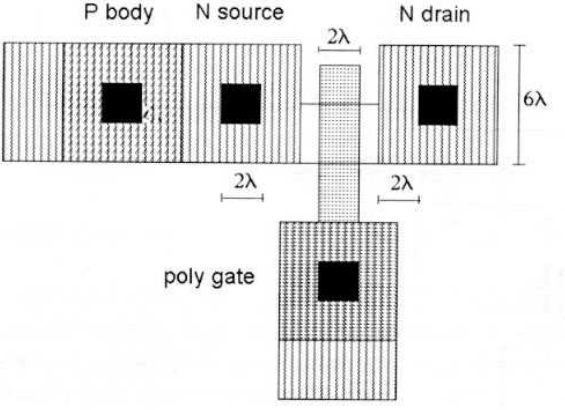


Figura 3.16 Tracciato di un NMOS ad area minima secondo le regole di progetto di Tabella 2.1 (le dimensioni sono in multipli di λ)

Applicando ad esempio le regole di Tabella 2.1 per un transistore MOS con gate di polisilicio ed area minima, si ottiene il tracciato riportato in Figura 3.16. Da

questo tracciato si ricavano i valori di W ed L (in questo caso rispettivamente 3λ e 2λ), le aree delle regioni di source A_S e drain A_D, nonché i valori delle lunghezze dL di sovrapposizione (legate alla minima dimensione geometrica ammissibile λ); una volta conosciuti i parametri del processo utilizzato, come la tensione di soglia, lo spessore dell'ossido t_{OX}, la profondità delle giunzioni delle regioni impiantate X_J,si possono determinare quindi i valori dei parametri dei modelli, come ad esempio il valore di K attraverso la (3.8), e le capacità tra i vari terminali secondo le (3.14, 3.15, 3.16, 3.17).

3.7 Modelli CAD di dispositivi MOS

Come si è detto nell'Introduzione, la presentazione dei concetti base dei circuiti digitali (ciò vale anche per i circuiti analogici) richiede, come elemento fondamentale per la comprensione del funzionamento dei circuiti, e come primo passo indispensabile per una progettazione critica di qualunque circuito, una utilizzazione significativa di strumenti e modelli analitici. Tuttavia l'uso di strumenti CAD (*Computer Aided Design*) per l'affinamento della progettazione in ambiente più realistico è divenuto un complemento essenziale allo studio dei circuiti elettronici, grazie anche alla sempre maggiore diffusione di simulatori efficienti e a basso costo, disponibili anche su Personal Computer di prestazioni limitate. Nel seguito quindi, molte delle analisi sviluppate con semplici strumenti analitici verranno confrontate con i risultati ottenuti con simulatori circuitali; in particolare si farà essenzialmente riferimento al programma di simulazione circuitale SPICE (*Simulation Program with Integrated Circuits Emphasis*), che è uno degli strumenti CAD più utilizzati sia per circuiti analogici che digitali, e che viene usualmente impiegato per l'analisi ed il progetto dei circuiti elettronici.

I modelli dei dispositivi utilizzati nei simulatori numerici dei circuiti sono di norma ben più complessi di quelli adottati per una descrizione analitica semplificata, essenzialmente perché tengono in conto una serie di dipendenze nonlineari dei parametri, e di effetti "del secondo ordine", che sarebbe troppo oneroso considerare in analisi "manuali". In questo paragrafo si elencheranno sinteticamente i parametri richiesti dal simulatore SPICE per descrivere i dispositivi MOS, rimandando ai riferimenti bibliografici per una descrizione più completa dei modelli di questo, come degli altri dispositivi, che saranno presentati in questo testo.

Il modello SPICE del MOS a cui si farà riferimento è quello definito LEVEL 1, che fa uso delle equazioni I-V in regime stazionario del dispositivo presentate nei Paragrafo 3.3 e 3.4, e delle capacità tra i vari elettrodi presentate nel Paragrafo 3.5. In particolare la suddivisione della capacità totale di gate nelle componenti C_{GS}, C_{GB}, C_{GD}, viene effettuata dal simulatore in funzione del punto di funzionamento, e le dipendenze delle capacità di giunzione dalla tensione di contropolarizzazione ammettono dipendenze funzionali analoghe a quelle riportate nella (3.21) ma con esponente anche diverso da 1/2 per tener conto del comportamento di giunzioni non necessariamente brusche. Come si è detto precedentemente, per le capacità sour-

ce/substrato e drain/substrato il modello prevede una divisione delle capacità di giunzione in due componenti, legate rispettivamente alla capacità unitaria C_{J0} che va moltiplicata per l'area corrispondente alla parte inferiore della giunzione (area equivalente a quella del tracciato della regione considerata) e a quella C_{JW} che va moltiplicata per il perimetro della giunzione.

Tabella 3.2

capacità	$t_{OX} = 100$ nm $\lambda = 4$ µm	$t_{OX} = 20$ nm $\lambda = 0.6$ µm
C_{OX}	0.34 fF/µm^2	1.7 fF/µm^2
C_{J0}	0.07 fF/µm^2	0.3 fF/µm^2
C_{JW}	0.2 fF/µm	0.4 fF/µm
C_{GDO}, C_{GSO}	0.34 fF/µm	0.2 fF/µm
C_{GBO}	0.4 fF/µm	0.2 fF/µm

Occorre anche fornire le componenti di sovrapposizione (overlap) delle capacità gate/drain, gate/source e gate/substrato, indicate con C_{GDO}, C_{GSO}, C_{GBO}. Le componenti suddette (come quelle laterali di giunzione C_{JW}) vengono fornite al simulatore per unità di perimetro, in quanto l'altra dimensione dell'area (dL per le prime e X_J per le seconde) è una grandezza tipica del processo (non del tracciato) e va inglobata nel valore della capacità.

Come esempio, in Tabella 3.2 sono riportati i valori delle diverse capacità, espresse rispettivamente per unità di area o di perimetro, ed in assenza di polarizzazione, rispettivamente per dispositivi con gate di alluminio di vecchia generazione (con spessore dell'ossido $t_{OX} = 100$ nm, $\lambda = 4$ µm, profondità di giunzione $X_J = 1$ µm, e sovrapposizione $dL = 1$ µm) e dispositivi più attuali con gate in polisilicio (con $t_{OX} = 20$ nm, $\lambda = 0.6$ µm).

Il circuito equivalente utilizzato nel simulatore è in definitiva quello riportato in Figura 3.14, dove vengono indicati i valori complessivi delle capacità (variabili) tra i vari terminali e non le singole componenti, che vengono in effetti specificate nella dichiarazione .MODEL del file. Nel modello compaiono anche i diodi corrispondenti alle giunzioni source/substrato e drain/substrato, che possono entrare in conduzione per particolari condizioni di funzionamento, e le resistenze associate ai terminali esterni, che tengono conto sia della resistenza delle regioni drogate che della resistenza delle interconnessioni (ad esempio per simulare la resistenza di contatto del polisilicio o della metallizzazione con queste regioni). In Appendice A sono riportate le schede .MODEL dei transistori MOS (NMOS, PMOS, NMOS a svuotamento) utilizzate per le simulazioni dei circuiti analizzati nel seguito.

È possibile utilizzare anche altri modelli dei dispositivi MOS nel simulatore SPICE, in particolare il LEVEL 3 che tiene conto degli effetti di canale corto dei MOS di ultima generazione, e il LEVEL 4 che è utilizzato per una descrizione del dispositivo a partire dai parametri di processo se non sono disponibili i valori dei parametri elettrici (quali tensioni di soglia, resistenze, capacità, ecc.).

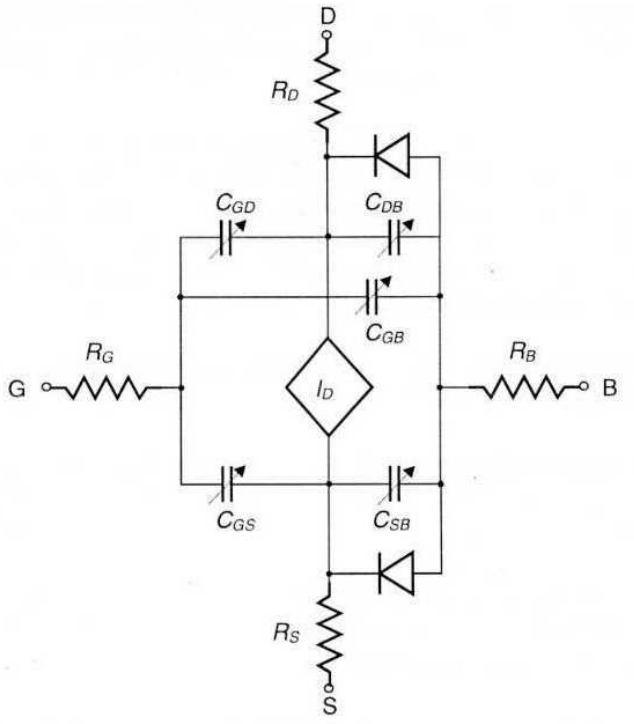


Figura 3.17 Circuito equivalente ad ampi segnali SPICE di un dispositivo MOS

In ogni caso l'accuratezza della descrizione di questi modelli è tutt'altro che soddisfacente se si richiede una simulazione accurata dei dispositivi realizzati con uno specifico processo tecnologico; usualmente per una migliore accuratezza dei modelli in fase di progettazione, i parametri elettrici da introdurre vengono estratti dai dispositivi realizzati, attraverso misure preventive su strutture di test (*test patterns*) di preproduzione, in modo da definire i parametri più corretti per una adeguata simulazione dei circuiti realizzati con un determinato processo tecnologico.

Esercizi di riepilogo

3.1 Per un transistore NMOS con $\gamma = 0.5\ V^{1/2}$, $\phi^* = 0.6$ V e con tensione di soglia $V_{TO} = 0.7$ V per $V_{SB} = 0$, valutare come si modifica la tensione V_T se è presente una tensione V_{SB} pari a 1 V o 5 V.

3.2 Calcolare la corrente di drain di un NMOS con i seguenti parametri: $k' = 40\ \mu A/V^2$, $V_{TO} = 0.7$ V, $W = 3\ \mu m$, $L = 1\ \mu m$, $\phi^* = 0.7V$, $\gamma = 0.5\ V^{1/2}$, per tensioni $V_{GS} = V_{DS} = 5$ V, $V_{SB} = 0$ V. Ripetere il calcolo per una tensione $V_{SB} = 4$ V.

3.3 Per un transistore NMOS con i valori dei parametri e delle tensioni riportate nell'Esercizio 3.2, determinare la tensione V_{DS} corrispondente al passaggio tra la regione lineare e quella di pinch-off.

3.4 Per il transistore NMOS dell'Esercizio 3.1 operante in regime di pinch-off, assumendo una dipendenza della tensione V_{TO} con la temperatura di -3 mV/°C, e una dipendenza lineare della mobilità μ_n con la temperatura con coefficiente pari a -3 cm^2/Vs°C, valutare il valore della tensione V_{GS} per la quale la corrente di drain passa da un coefficiente di temperatura positivo ad uno negativo (si consideri la caratteristica di trasferimento $I_D = f(V_{GS})$).

3.5 Determinare la resistenza offerta, in regime di funzionamento lineare, da un NMOS con $k' = 40$ μA/V^2, $V_{TO} = 0.7$V, $W = 4$ μm, $L = 2$ μm, con $V_{GS} = 5$V.

3.6 Determinare la resistenza offerta da un PMOS, in regime di funzionamento lineare, con i seguenti valori: $k' = 10$ μA/V^2, $V_{TO} = 0.7$V, $W = 4$ μm, $L = 2$ μm, con $V_{GS} = 5$V.

3.7 Determinare le correnti di drain in pinch-off, rispettivamente per un NMOS e un PMOS ad area minima, assumendo i seguenti valori: $k'_N = 50$ μA/V^2, $k'_P = 20$ μA/V^2, $V_{TNO} = |V_{TPO}| = 0.7$V, $V_{GS} = V_{DS} = 5$ V.

3.8 Calcolare la corrente I_D per un NMOS a svuotamento con i seguenti parametri: $k' = 30$ μA/V^2, $W = 3$ μm, $L = 9$ μm, $V_{TD} = -3$V, $V_{GS} = 0$V, $V_{DS} = 3$V.

3.9 Calcolare per quale valore della tensione di contropolarizzazione la capacità di una giunzione si dimezza rispetto al valore in condizione di polarizzazione nulla.

3.10 Per un MOS con area di gate di 8×2 μm^2, aree di source e drain di 8×6 μm^2, determinare il rapporto tra le capacità C_G e C_{DB0} per i due processi riportati in Tabella 3.2.

3.11 Disegnare il tracciato di un NMOS ad area minima utilizzando le regole di progetto della Tabella 2.2 e valutare le capacità C_G, C_{SB0}, C_{DB0}, utilizzando i valori unitari riportati in Tabella 3.2 per un processo con $\lambda = 0.6$ μm.

3.12 Ripetere il problema dell'Esercizio 3.10 per un nuovo tracciato con $W = 9\ \lambda$, $L = 6\ \lambda$; quanto vale in questo caso il rapporto tra la capacità C_{DB0} e quella C_G, e come si è modificato rispetto a quello relativo al caso dell'Esercizio 3.10?

Riferimenti bibliografici

R.S. Muller, T.I. Kamins, *Dispositivi elettronici nei circuiti integrati*, Bollati Boringhieri, Torino, 1982.

G. Soncini, *Tecnologie Microelettroniche*, Bollati Boringhieri, Torino, 1986.

P. Antognetti, G. Massobrio, *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York, 1987.

4

Porte elementari NMOS

4.1 Introduzione

In ordine cronologico di apparizione sul mercato, le famiglie logiche bipolari hanno preceduto quelle basate su componenti e tecnologie MOS; tuttavia si partirà da queste ultime nello studio delle porte logiche elementari, per una più graduale presentazione della materia. Ciò essenzialmente perché il funzionamento dei dispositivi MOS, che sono dispositivi unipolari e possono essere considerati come dei resistori nonlineari controllati, è più facile da descrivere rispetto al funzionamento dei dispositivi bipolari; inoltre i circuiti delle porte logiche MOS sono più direttamente riconducibili agli invertitori elementari che le compongono, e permettono quindi una più semplice analisi basata sul funzionamento di questi ultimi. D'altra parte questo aspetto di più semplice progettazione dei circuiti MOS è tra le principali ragioni che, dopo una prima fase di sviluppo di una tecnologia affidabile per la realizzazione dei dispositivi, negli anni '70, ha portato ad un così ampio sviluppo dei sistemi digitali basati su tecnologie MOS, in particolare per quanto riguarda i sistemi ad elevata densità di integrazione, come i circuiti integrati VLSI e ULSI.

Nello studio delle porte logiche con tecnologia MOS faremo riferimento, per le caratteristiche elettriche, agli invertitori, che sono le porte logiche più elementari, differenziando questi in base alla tecnologia utilizzata per la realizzazione dei dispositivi stessi. In questo capitolo verranno quindi trattati gli invertitori realizzati a partire da dispositivi ad arricchimento con canale di tipo N (NMOS ad arricchimento), e quelli con dispositivi NMOS sia ad arricchimento che a svuotamento, e successivamente le porte logiche basate su questi invertitori. Nel Capitolo 5 verranno invece discussi gli invertitori che fanno uso della tecnologia CMOS, una tecnologia che permette di realizzare sia dispositivi NMOS che PMOS realizzati su uno stesso substrato, e che oggi è la più diffusa, sia per le logiche standard che in generale per i circuiti digitali ad alta scala di integrazione; nel seguito dello stesso capitolo verranno introdotte le porte logiche basate sugli invertitori elementari CMOS.

4.2 Invertitore NMOS con carico resistivo

Il più semplice circuito invertitore che utilizza dispositivi MOS è quello che discende direttamente dallo schema di principio dell'invertitore riportato in Figura 1.10, sostituendo l'interruttore controllato con un transistore NMOS, come indicato in Figura 4.1; nello schema la tensione di alimentazione è indicata con V_{DD}, la tensione di ingresso V_I coincide con la V_{GS} e quella di uscita V_O con la V_{DS}.

Un'analisi del comportamento ad ampi segnali in regime stazionario di questo circuito (ossia non tenendo in conto i transitori di commutazione che coinvolgono le capacità del circuito), è fattibile sia per via analitica, che per via grafica. Nel primo caso occorre risolvere, assegnato un valore della grandezza di ingresso V_{GS}, il sistema formato dalla equazione delle caratteristiche I-V di uscita del MOS e da quella che descrive il vincolo posto dalla resistenza di carico R:

$$\begin{cases} I_D = f(V_{GS}, V_{DS}) \\ RI_D = V_{DD} - V_{DS} \end{cases} \tag{4.1}$$

Per effettuare l'analisi grafica, occorre riportare sul grafico della famiglia di caratteristiche di uscita (vedi Figura 4.1) il vincolo della seconda delle (4.1) che nel piano I_D, V_{DS} è una retta, detta *retta di carico* dell'invertitore.

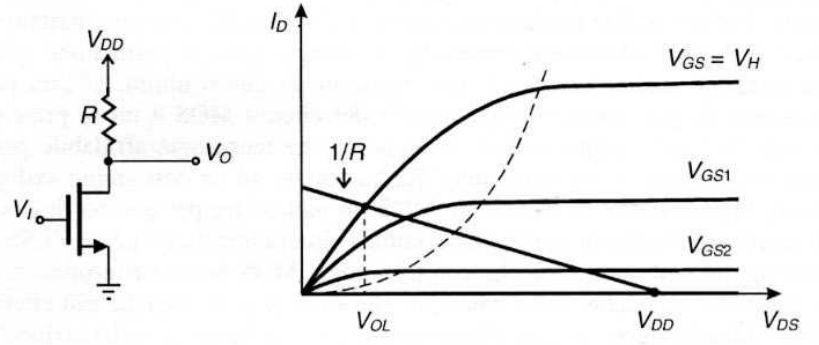


Figura 4.1 Invertitore NMOS con carico resistivo

I singoli punti di funzionamento (I_D, V_{DS}), al variare della tensione di ingresso V_{GS}, vengono immediatamente determinati dall'intersezione della retta di carico con la caratteristica corrispondente a quel valore di V_{GS}; è quindi immediato ottenere per punti la caratteristica di trasferimento $V_O - V_I$, da cui vengono ricavati i valori delle tensioni di funzionamento e i margini di rumore, assegnando alla grandezza V_I i valori di V_{GS} corrispondenti alle singole caratteristiche di uscita. Ad esempio il valore della minima tensione di uscita V_{OL} della Figura 4.1 corrisponde all'intersezione della retta di carico con la caratteri-

stica di uscita corrispondente alla tensione V_{GS} pari alla massima tensione (V_H) di ingresso.

L'analisi grafica, anche se più approssimata di quella analitica, viene frequentemente utilizzata perché permette una comprensione più diretta e globale del funzionamento dei circuiti elementari; ad esempio, nel caso dell'invertitore di Figura 4.1 è immediato rendersi conto, dall'analisi grafica, che lo swing logico è tanto più ampio quanto più è inclinata verso il basso la retta di carico, cioè quanto più è grande il valore della resistenza R, a parità di tensione di alimentazione V_{DD}, perché l'aumento del valore della R riduce il valore della tensione logica bassa V_{OL}.

In Figura 4.2 sono riportate le caratteristiche di trasferimento corrispondenti a tre valori diversi della resistenza di carico R, per un invertitore a NMOS con rapporto $W/L = 5$, $k' = 10\mu A/V^2$ e tensione di soglia $V_T = 1.5$ V, ottenute con il simulatore circuitale SPICE. Come era intuibile in base alla costruzione grafica di Figura 4.1, una riduzione della resistenza di carico comporta un aumento della tensione di uscita nello stato basso V_{OL}; anche i valori delle altre grandezze significative come V_{IH} e V_{IL} si degradano se R diminuisce, e quindi in definitiva i margini di rumore vengono ad essere ridotti sensibilmente. Da un esame di queste grandezze per i tre casi esaminati si verifica che i margini di rumore sono accettabili solo per valori della resistenza R superiori a 50 kΩ, mentre valori dell'ordine di 10 kΩ portano a margini di rumore troppo bassi, e a valori di V_{OL} troppo elevati.

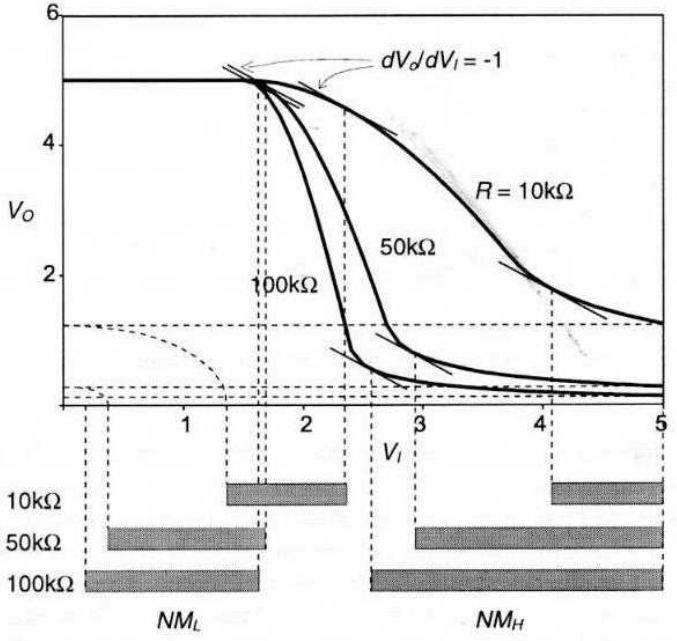


Figura 4.2 Caratteristiche di trasferimento e margini di rumore di un invertitore NMOS con $W/L = 5$, per tre valori della resistenza di carico

I risultati presentati sono relativi alle caratteristiche di un NMOS con fattore di forma $W/L = 5$, e quindi la situazione peggiora se ci si riferisce a dispositivi con minore area e quindi con rapporto W/L vicino all'unità, perché in tal caso la scala di correnti (legata al fattore K) delle caratteristiche si riduce, ed eguali intersezioni (in termini di grafico) con la retta di carico corrispondono a resistenze di valore corrispondentemente più elevato.

L'osservazione sui valori di resistenza necessari per un buon funzionamento dell'invertitore è legata alla necessità di dovere realizzare queste resistenze *per via integrata*, secondo le tecniche presentate nel Paragrafo 2.4; si è visto che con i processi di drogaggio utilizzati per i MOS non è agevole realizzare resistenze integrate con valori superiori a qualche kΩ.

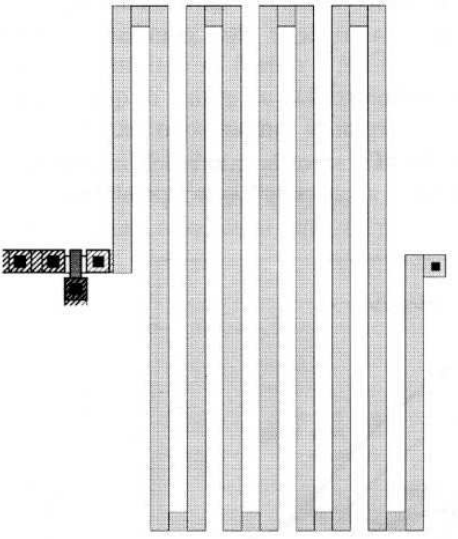


Figura 4.3 Area occupata da una resistenza integrata da 10 kΩ in un invertitore NMOS con carico resistivo

In Figura 4.3 è rappresentato il tracciato che assumerebbe un invertitore NMOS con una resistenza di carico integrata da 10 kΩ utilizzando regioni drogate con resistenza di strato di 50 Ω/□ per cui occorre un rettangolo di lunghezza 200 volte la larghezza; l'area occupata dal resistore, già per questo valore di R, è inaccettabilmente grande rispetto a quella del transistore MOS. Poiché, come si è detto, la minimizzazione dell'area occupata assume un ruolo fondamentale nella scelta del circuito, la realizzazione di efficienti invertitori con tecnologia MOS richiede l'utilizzazione di carichi non ohmici, detti *carichi attivi*, che verranno descritti nei paragrafi seguenti.

4.3 Il dispositivo MOS come carico attivo

Le considerazioni del paragrafo precedente suggeriscono l'utilizzo di un dispositivo MOS anche come resistenza di carico, vista la sua ridotta area di occupazione rispetto a quella di una resistenza integrata. In questo caso si utilizzerà il dispositivo come resistore nonlineare, in una configurazione in cui il terminale di controllo (gate) è connesso con uno degli altri due terminali, in modo da trasformare il dispositivo attivo in un bipolo. Le possibili scelte in principio sono due, e cioè con la gate connessa o al source o al drain; se si utilizzano MOS ad arricchimento, la gate non può essere collegata al source perché in tal caso il dispositivo è permanentemente interdetto, e quindi la resistenza del bipolo equivalente tra drain e source risulta infinita. Per questi dispositivi l'unica configurazione possibile è quella di Figura 4.4, con la gate connessa al drain.

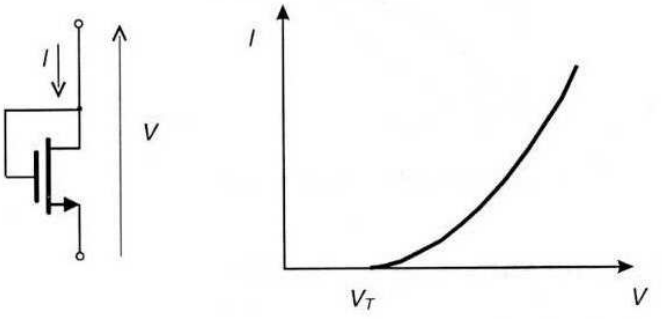


Figura 4.4 Transistore NMOS ad arricchimento come carico attivo

La caratteristica *I-V* di questo bipolo (vista tra i terminali di drain e source) si ottiene direttamente dalla (3.11) (in quanto, per $V_{DG} = 0$, la condizione di pinch-off è sempre verificata), ponendo $V_{GS} = V_{DS} = V$:

$$I = K \cdot (V - V_T)^2 \tag{4.2}$$

L'andamento di questa curva è di tipo superlineare, e la scala delle correnti in gioco viene semplicemente definita dal fattore K, e quindi, in termini di tracciato, dal rapporto W/L del dispositivo; ciò permette di realizzare elevati valori di resistenza equivalente in maniera semplice, riducendo il rapporto W/L, senza per questo richiedere elevate aree di ingombro.

Un secondo modo di realizzazione di carichi attivi è basato sull'utilizzazione di dispositivi MOS a svuotamento; in questo caso è possibile realizzare un resistore nonlineare connettendo la gate al source, poiché il canale è già formato anche per $V_{GS} = 0$, come è indicato in Figura 4.5. Il bipolo presenterà una curva *I-V* corrispondente alla caratteristica di uscita per $V_{GS} = 0$, descritta dal set di Equazioni (3.12-3.13):

$$I = K \cdot (-2V_{TD} \cdot V - V^2) \qquad \text{per} \quad V < -V_{TD}$$
$$I = K \cdot V_{TD}^2 \qquad \text{per} \quad V \geq -V_{TD} \tag{4.3}$$

In questo caso, il dispositivo potrà lavorare sia in regione nonlineare che di pinch-off a seconda del potenziale V (V_{DS}) applicato ai terminali. La curva I-V è in questo caso sublineare, e quindi il comportamento come carico attivo del MOS a svuotamento è in certo modo complementare a quello del MOS ad arricchimento.

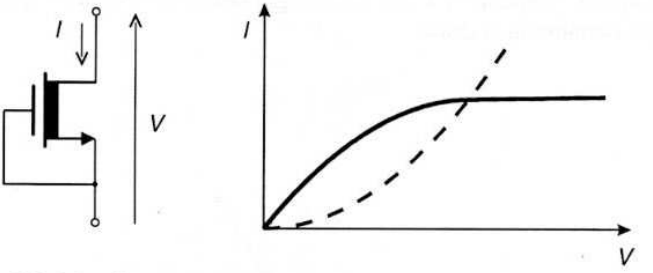


Figura 4.5 Transistore NMOS a svuotamento come carico attivo

4.4 Invertitori con carico attivo NMOS

Gli schemi elettrici di invertitori elementari che impiegano rispettivamente dispositivi NMOS ad arricchimento o a svuotamento come carico attivo sono riportati in Figura 4.6.

Per questi circuiti l'analisi può essere svolta analogamente a quanto indicato nel Paragrafo 4.2 per l'invertitore con carico resistivo; in particolare per l'analisi grafica si può considerare al posto della retta di carico di Figura 4.1 una *curva di carico* costituita dal bipolo nonlineare costituito dal NMOS di carico. Questa curva di carico è costituita dalla curva $V(I)$ del bipolo, riportata in Figura 4.4, ribaltata rispetto all'asse delle ordinate e traslata di una quantità pari a V_{DD} sull'asse delle ascisse, in accordo alla relazione:

$$V_O = V_{DD} - V(I) \tag{4.4}$$

valida per qualsiasi curva di carico.

È bene sottolineare che la curva I-V del transistore NM_2 di carico in questi casi non è data dalle semplici espressioni (4.2) o (4.3), perché bisogna tener conto, per questo transistore, dell'effetto di substrato (*body effect*) sulla tensione di soglia. Infatti (vedi Figura 4.6) i terminali di substrato dei due transistori sono connessi allo stesso potenziale, in quanto il substrato (body) è comune ai due transistori e corrisponde al silicio del wafer in cui essi sono realizzati. Que-

sto substrato va portato al potenziale più basso, in questo caso la massa, per assicurare la contropolarizzazione di tutte le regioni di drain e source; quindi per tensioni di uscita $V_O > 0$, ricordando che V_O corrisponde alla tensione V_{S2} del source del MOS di carico, si svilupperà una tensione positiva tra source e substrato, il che comporta un aumento della tensione di soglia V_T secondo l'Equazione (3.2).

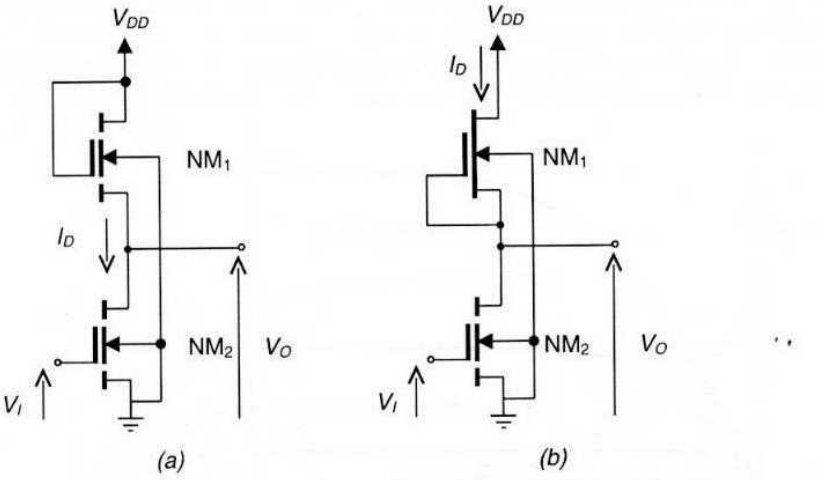


Figura 4.6 Schema elettrico di: a) invertitore con carico ad arricchimento; b) invertitore con carico a svuotamento

L'effetto body sulla tensione di soglia è tanto maggiore quanto più alta è la tensione di uscita (e quindi V_{S2}), mentre è nullo per tensione $V_O = 0$. Questo comporta una modifica delle curve di carico, come si può vedere nelle costruzioni grafiche di Figura 4.7, in cui la curva di carico corrispondente al caso ideale senza effetto body (curva tratteggiata) si discosta da quella reale (curva a tratto pieno) per la presenza dell'effetto body nel transistore NM_2, effetto che diventa sempre più rilevante per valori crescenti di V_O.

Le costruzioni di Figura 4.7 sono relative ad un invertitore con valore di K (rapporto W/L) di NM_1 maggiore di quello di NM_2, in quanto dall'analisi del Paragrafo 4.2 si è visto come sia utile un aumento della resistenza equivalente del carico per un miglioramento dei margini di rumore dell'invertitore ed in particolare per una riduzione di V_{OL}. La realizzazione del carico attivo con un MOS permette di ridurre la corrente circolante nel carico (e quindi di aumentare la resistenza equivalente) senza aumentare di molto l'area impegnata, poiché è possibile ridurre il fattore di scala K scegliendo un rapporto W/L minore dell'unità (ricordiamo che il termine k' per un NMOS è dell'ordine di qualche decina di $\mu A/V^2$).

Dalla costruzione grafica si evince che un rapporto elevato tra i valori K_1 del dispositivo NM_1 e K_2 del carico attivo NM_2 permette di ottenere bassi valori di V_{OL};

definiremo quindi questo rapporto come $K_R = K_1/K_2$. Il valore di K_R è un parametro rilevante per le prestazioni statiche dell'invertitore e in genere delle porte elementari; si vedrà che questo rapporto interviene anche nella determinazione dei parametri dinamici dell'invertitore, e per tale ragione questo tipo di logica basata su dispositivi NMOS viene detta *logica a rapporto*, in quanto le sue principali proprietà dipendono dalla scelta di questo rapporto, che è una delle principali variabili a disposizione del progettista del circuito.

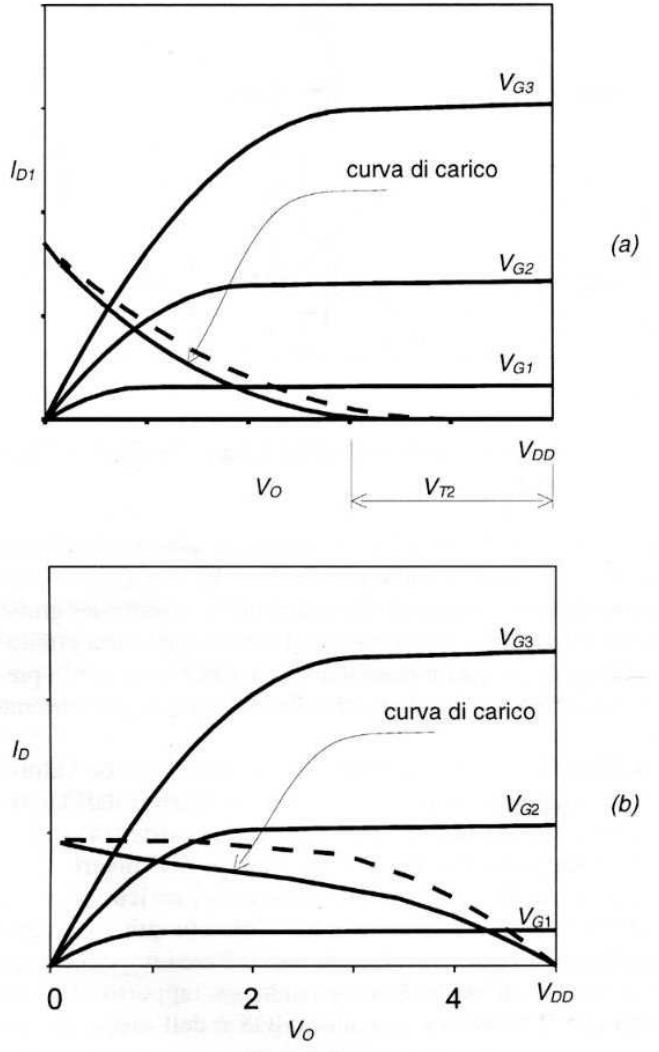


Figura 4.7 Analisi grafica di un invertitore NMOS con carico attivo; a) carico con MOS ad arricchimento; b) carico con MOS a svuotamento

4.5 Caratteristica di trasferimento e margini di rumore

4.5.1 Carico ad arricchimento

Utilizzando l'analisi grafica nel piano delle caratteristiche di uscita per un invertitore con carico attivo ad arricchimento (o invertitore E-E), riportata in Figura 4.8a, si può ricavare per punti la caratteristica di trasferimento dell'invertitore, come è mostrato in Figura 4.8b. In questa si identificano tre regioni, definite dai rispettivi punti limite riportati nella stessa curva, corrispondenti alle diverse condizioni di funzionamento dell'invertitore, che possono essere facilmente desunte dall'analisi grafica di Figura 4.8a.

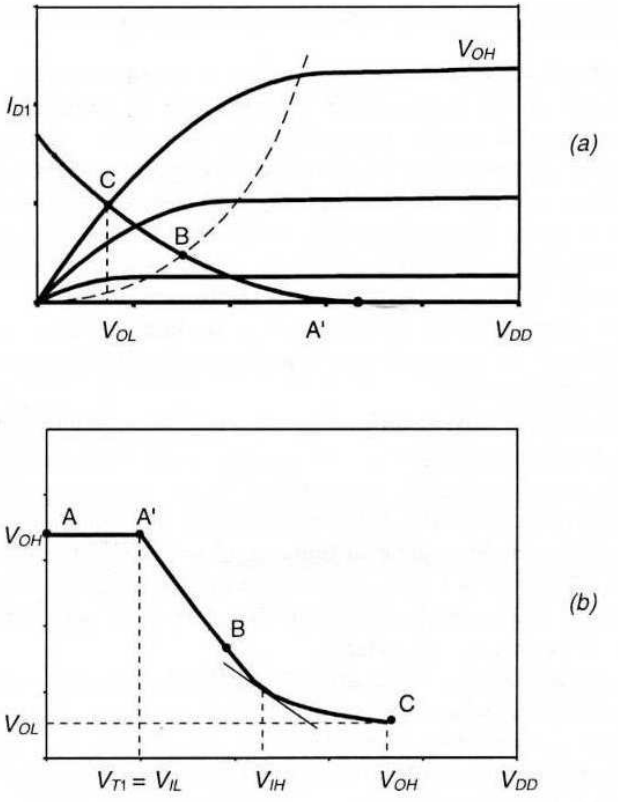


Figura 4.8 a) Caratteristiche di uscita e b) funzione di trasferimento di un invertitore NMOS con carico ad arricchimento

Una prima regione di funzionamento, tra i punti A e A', corrisponde a valori della tensione V_I in ingresso minori della tensione di soglia V_{T1} del transistore NM_1.

Il punto di funzionamento nel piano *I-V* di uscita è quello A' corrispondente ad una tensione $V_{DS1} = V_O = V_{DD} - V_{DS2}$ (si ricorda che il transistore NM_2 presenta una caduta fissa ai suoi capi pari a V_{T2} anche per corrente nulla). Quando la tensione di ingresso V_I supera la tensione di soglia V_{T1} il transistore NM_1 comincia a condurre, e i punti di funzionamento corrispondono alle intersezioni delle caratteristiche *I-V* con la curva di carico. Nella seconda regione (tra i punti A' e B) queste intersezioni avvengono nella regione di pinch-off di NM_1, mentre nella terza regione (tra i punti B e C) i punti di funzionamento corrispondono ad un funzionamento di NM_1 in regione nonlineare. Si può facilmente dimostrare che nella seconda regione la curva di trasferimento ha un andamento lineare, in quanto la dipendenza quadratica della corrente di NM_1 dalla tensione di ingresso è compensata dalla dipendenza, anch'essa quadratica, della corrente di carico dalla tensione di uscita V_O; nella terza regione l'andamento della curva di trasferimento diventa sublineare e tende asintoticamente ad una tensione nulla, valore raggiungibile solo per tensioni di ingresso idealmente infinite.

Il punto C corrisponde all'intersezione della curva di carico con la caratteristica di NM_1 corrispondente ad una $V_{GS} = V_I = V_{OH}$, cioè ad un pilotaggio dell'invertitore con la tensione di uscita nominale dello stato alto (quella fornita dall'invertitore a monte in una catena di invertitori o di porte logiche); quindi il valore di V_O al punto C corrisponde al valore V_{OL} cioè alla tensione nominale di uscita nello stato basso. Quest'ultima determina con la V_{OH} l'escursione logica della porta e quindi, per un buon progetto dell'invertitore, la prima deve essere di valore sufficientemente piccolo (ad es. minore del 10% della tensione nello stato alto V_{OH}); in ogni caso è necessario che il valore di V_{OL} sia inferiore alla tensione di soglia V_{T1} del MOS NM_1, in quanto in corrispondenza del livello logico basso quest'ultimo deve essere interdetto.

Gli altri due punti significativi della caratteristica di trasferimento sono i punti V_{IL} e V_{IH} per i quali la curva presenta una pendenza pari a -1 (vedi Paragrafo 1.4). Il primo corrisponde al punto A' nella caratteristica di trasferimento, in quanto a sinistra di questo la pendenza è nulla e a destra la pendenza è costante e in modulo maggiore di 1 (si ricorda che la regione di transizione in un invertitore deve presentare una pendenza maggiore di 1, altrimenti i margini di rumore sarebbero nulli); il secondo si trova nella terza regione (tra i punti B e C) perché per quanto detto, dal punto A' al punto B la pendenza è costante.

Una valutazione quantitativa delle dipendenze funzionali dei valori V_{OL}, V_{IL}, V_{IH}, V_{OH} dalle caratteristiche costruttive dell'invertitore può essere sinteticamente ottenuta per via analitica, eguagliando le correnti I_{D1} e I_{D2} dei due dispositivi, rispettivamente date dalle (3.10) e (4.2), e approssimando opportunamente le espressioni delle caratteristiche dei dispositivi nelle diverse regioni di funzionamento.

Determinazione di V_{OL} (NM_1 in regione triodo, NM_2 in pinch-off)

Sostituendo nelle (3.10a) e (4.2) alle rispettive variabili V_{GS} e V i valori corrispondenti del circuito V_I e $V_{DD} - V_O$ si ha:

$$I_{D1} \equiv K_1\left[2(V_I - V_{T1})V_O - V_O^2\right] = I_{D2} \equiv K_2\left[V_{DD} - V_O - V_{T2}\right]^2 \qquad (4.5)$$

Il valore di V_{OL} si ricava dalla (4.5) specificando in questo caso $V_I = V_{OH}$, la tensione alta di uscita del circuito a monte, e $V_O = V_{OL}$:

$$I_{D1} \equiv K_1\left[2(V_{OH} - V_{T1})V_{OL} - V_{OL}^2\right] = I_{D2} \equiv K_2\left[V_{DD} - V_{OL} - V_{T2}\right]^2 \qquad (4.6)$$

L'approssimazione che si farà in questo caso è di considerare NM_1 in regione lineare, trascurando cioè nel primo membro della (4.6) il termine quadratico in V_{OL} rispetto a quello lineare, ossia assumendo $V_{OL} << V_{OH} - V_{T1}$; questa approssimazione è tanto più valida quanto più piccola è V_{OL} rispetto a V_{OH}, condizione che si cerca di realizzare in un buon progetto dell'invertitore. In queste ipotesi, è possibile trascurare anche il termine V_{OL} nella parentesi a destra dell'uguaglianza, e si ha:

$$2K_R(V_{OH} - V_{T1})V_{OL} \equiv (V_{DD} - V_{T2})^2 \qquad (4.7)$$

dove si è posto $K_R = K_1/K_2$, rapporto tra i fattori di scala dei due MOS dell'invertitore. Dalla (4.7) si ottiene V_{OL}:

$$V_{OL} \cong \frac{(V_{DD} - V_{T2})^2}{2K_R(V_{OH} - V_{T1})} \qquad (4.8)$$

(Si ricorda che alla tensione V_{OL}, il dispositivo NM_2 ha un effetto body trascurabile per cui è lecito sostituire la tensione di soglia V_{T2} con quella V_{T20} in assenza di effetto body).

Determinazione di V_{IH} (NM_1 in regione triodo, NM_2 in pinch-off)

Questo punto è definito dalla condizione che la pendenza della curva sia negativa e unitaria, cioè:

$$\frac{dV_O}{dV_I} = -1 \qquad (4.9)$$

Derivando entrambi i membri della (4.5) rispetto a V_O si ha:

$$K_1\left[2\frac{dV_1}{dV_O}V_O + 2(V_1 - V_{T1}) - 2V_O\right] = -2K_2\left[V_{DD} - V_O - V_{T2}\right] \qquad (4.10)$$

Imponendo la condizione (4.9) nella (4.10), specializzata per $V_I = V_{IH}$ si ha:

$$2V_O - (V_{IH} - V_{T1}) = \frac{K_2}{K_1}(V_{DD} - V_O - V_{T2})$$

da cui, assumendo trascurabile il secondo termine, in quanto il rapporto $K_2/K_1 = 1/K_R$ deve essere piccolo per garantire un basso valore di V_{OL} (Equazione (4.8)), si ottiene una relazione tra V_{IH} e V_O:

$$V_{IH} = 2V_O + V_{T1} \tag{4.11}$$

Sostituendo questo valore di V_{IH} al posto di V_I nella (4.5) si ha:

$$K_R(3V_O^2) = (V_{DD} - V_O - V_{T2})^2 \tag{4.12}$$

da cui si ricava il termine V_O:

$$V_O = \frac{V_{DD} - V_{T2}}{\sqrt{3K_R} + 1} \tag{4.13}$$

Sostituendo la (4.13) nella (4.11) si ottiene per V_{IH}:

$$V_{IH} \equiv V_{T1} + 2\frac{V_{DD} - V_{T2}}{\sqrt{3K_R} + 1} \tag{4.14}$$

Determinazione di V_{IL} (NM_1, NM_2 in pinch-off)

La condizione di pendenza unitaria per il punto corrispondente alla tensione V_{IL} identifica il punto di discontinuità A' della curva di trasferimento perché come si è detto, a destra di questo la pendenza in modulo è maggiore di 1. La tensione corrispondente è quindi quella di soglia del dispositivo NM_1:

$$V_{IL} = V_{T1} \tag{4.15}$$

Determinazione di V_{OH} (NM_1, NM_2 in pinch-off)

Il valore di V_{OH} è facilmente ricavabile se si considera che, in un invertitore correttamente progettato, il valore della tensione di ingresso nello stato basso è certamente inferiore alla tensione di soglia V_{T1}; quindi il punto di funzionamento dell'invertitore si trova all'intersezione della curva di carico con l'asse delle ascisse (vedi Figura 4.8) e:

$$V_{OH} = V_{DD} - V_{T2} \tag{4.16}$$

In questo caso, la tensione di soglia di NM_2 è modificata significativamente dall'effetto body che, come si è detto, aumenta con il crescere della tensione di source di NM_2 e cioè con la tensione di uscita dell'invertitore; il valore di V_{T2} viene determinato dalla (3.2) come:

$$V_{T2} = V_{T20} + \gamma\left[\sqrt{\phi^* + V_{OH}} - \sqrt{\phi^*}\right] \tag{4.17}$$

dove V_{T20} è la tensione di soglia in assenza di effetto body, e le altre grandezze sono state definite nel Paragrafo 3.3. Come esempio, per dispositivi con $\gamma = 0.5$, $V_{T20} = 1$ V, per una $V_{OH} = 3.4$ V (caso di Figura 4.7) si ha una tensione di soglia $V_{T2} = 1.6$ V.

Le espressioni approssimate ricavate per le tensioni caratteristiche di ingresso e di uscita dell'invertitore NMOS dipendono dalle tensioni di soglia dei due dispositivi, che sono parametri legati alla tecnologia di realizzazione dell'integrato e non facilmente modificabili, e dalla tensione di alimentazione, anch'essa una grandezza standardizzata nelle famiglie logiche commerciali; tuttavia le espressioni di V_{OL} e V_{IH} mostrano anche una dipendenza dal rapporto K_R, che a sua volta dipende dalle dimensioni geometriche delle gate dei due MOS:

$$K_R = \frac{W_1/L_1}{W_2/L_2} \tag{4.18}$$

ed è quindi una grandezza a disposizione del progettista del circuito per definirne le prestazioni elettriche.

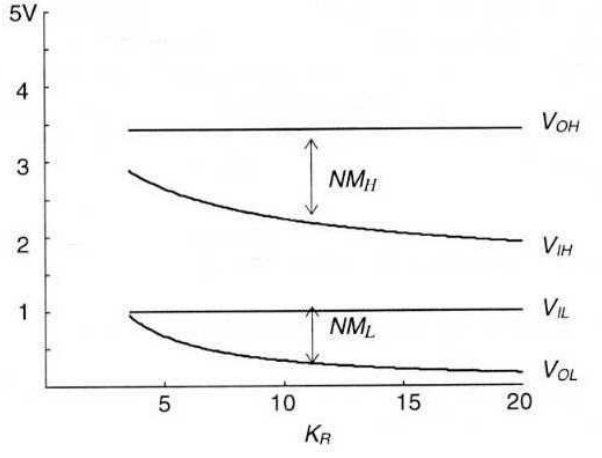


Figura 4.9 Dipendenza dei margini di rumore forniti dalle (4.8), (4.14), (4.15), (4.16) dal rapporto K_R, per invertitori NMOS con carico ad arricchimento

Dalle Espressioni approssimate (4.8), (4.14), (4.15), (4.16), riportate in Figura 4.9 in funzione di questo rapporto per un invertitore con tensioni di soglia $V_{T1} = V_{T20} = 1$ V, $V_{DD} = 5$ V, $\gamma = 0.5$, si vede che sia V_{IH} che V_{OL} decrescono all'aumentare di K_R, con effetti benefici per entrambi i margini di rumore NM_H e NM_L. Il margine di rumore più basso è in questo caso quello NM_L; per avere valori significativi di quest'ultimo occorre scegliere valori di K_R superiori a 10.

Una ulteriore osservazione è la penalizzazione sull'escursione logica disponibile indotta dall'effetto body di NM_2, che limita il valore di V_{OH} a valori significativamente inferiori alla tensione di alimentazione disponibile V_{DD}; si vedrà che questo inconveniente viene eliminato negli invertitori con carico a svuotamento.

4.5.2 Carico a svuotamento

L'analisi per l'invertitore con carico attivo a svuotamento (detto invertitore E-D) viene svolta allo stesso modo del caso precedente. L'analisi grafica nel piano delle caratteristiche di uscita (Figura 4.10a) permette di ricavare la caratteristica di trasferimento dell'invertitore, come è mostrato in Figura 4.10b. In questa si identificano quattro regioni, corrispondenti alle diverse condizioni di funzionamento dei due dispositivi dell'invertitore, facilmente desunte dall'analisi grafica di Figura 4.10a.

La prima regione di funzionamento (A-A') corrisponde a valori della tensione V_I in ingresso minori della tensione di soglia V_{T1} del transistore NM_1, e il punto di funzionamento nel piano I-V di uscita è quello A' corrispondente ad una tensione $V_{DS1} = V_O = V_{DD}$. La seconda regione (A'-B), per tensioni di ingresso V_I maggiori di V_{T1}, corrisponde alle intersezioni con il tratto della caratteristica di NM_2 in regime nonlineare, mentre la terza regione (B-C) si percorre quando entrambi i dispositivi sono in regime di pinch-off. Infine la quarta regione (C-D) corrisponde a punti di funzionamento in cui il transistore NM_1 lavora in regime nonlineare.

In questo tipo di invertitore si può subito rilevare che la tensione di uscita nello stato alto V_{OH} coincide con la tensione di alimentazione V_{DD}; questo è un notevole vantaggio dovuto al carico attivo a svuotamento, e per questa regione questo invertitore è preferito a quello con il carico ad arricchimento.

Anche in questo caso i valori V_{OL}, V_{IL}, V_{IH}, V_{OH} possono essere ottenuti per via analitica, eguagliando le correnti I_{D1} e I_{D2} dei due dispositivi, rispettivamente date dalle (3.10) e (4.3), con opportune approssimazioni.

Determinazione di V_{OL} (NM_1 in regione triodo, NM_2 in pinch-off)

In questo caso l'eguaglianza delle correnti in NM_1 e NM_2, specificando per V_O il valore V_{OL} e per V_I quello V_{OH} fornisce:

$$I_{D1} \equiv K_1\left[2(V_{OH} - V_{T1})V_{OL} - V_{OL}^2\right] = I_{D2} \equiv K_2 \cdot |V_{TD}|^2 \qquad (4.19)$$

Trascurando anche in questo caso il termine quadratico rispetto a quello lineare nella (4.19), e ricordando la definizione di K_R come rapporto tra i due K, si ottiene:

$$V_{OL} \cong \frac{|V_{TD}|^2}{2K_R(V_{DD} - V_{T1})} \tag{4.20}$$

(Ricordiamo che nella (4.20) che definisce la tensione V_{OL}, la tensione di soglia V_{TD} può essere approssimata con quella V_{TDO} in assenza di effetto body, in quanto in questo punto il source di NM_2 è a un potenziale molto vicino a quello di massa.)

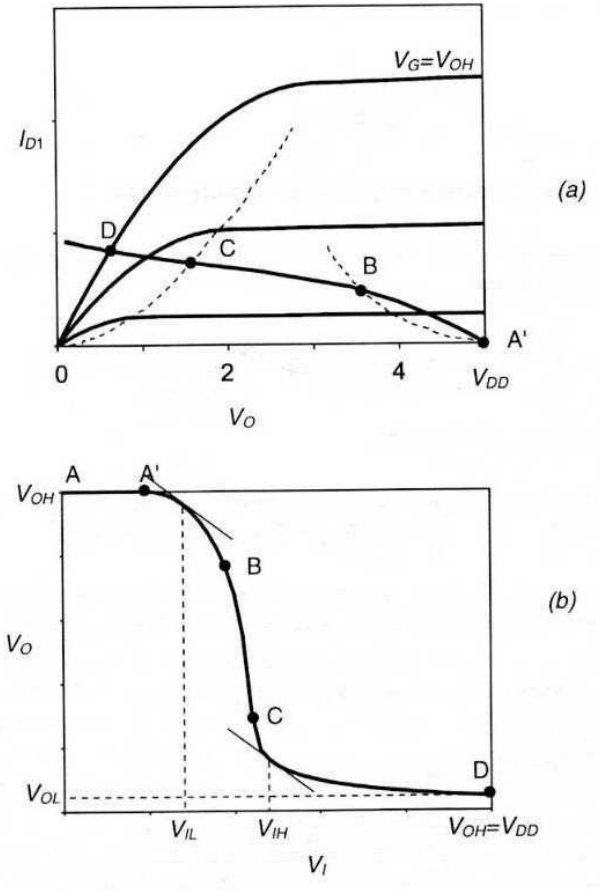


Figura 4.10 a) Caratteristiche di uscita e b) funzione di trasferimento di un invertitore NMOS con carico a svuotamento

Determinazione di V_{IH} (NM_1 in regione triodo, NM_2 in pinch-off)

La condizione sulla pendenza della curva $dV_O/dV_I = -1$ fornisce, sviluppando i calcoli analogamente al caso con carico ad arricchimento (Equazione (4.10)) lo stesso legame tra V_{IH} e V_O della (4.11):

$$V_{IH} = 2V_O + V_{T1} \tag{4.21}$$

Sostituendo il valore di V_O ricavato dalla (4.21) nella (4.19) si ha:

$$K_1\left[2(V_{IH} - V_{T1})\frac{V_{IH} - V_{T1}}{2} - \left(\frac{V_{IH} - V_{T1}}{2}\right)^2\right] = K_2 \cdot |V_{TD}|^2 \tag{4.22}$$

da cui:

$$V_{T1} + 2\frac{|V_{TD}|^2}{\sqrt{3K_R}} \tag{4.23}$$

Determinazione di V_{IL} (NM_1 in pinch-off, NM_2 in regione triodo)

In questo caso l'eguaglianza delle correnti, specializzata per i campi di funzionamento di MN_1 e MN_2, fornisce:

$$I_{D1} \equiv K_1(V_I - V_{T1})^2 = I_{D2} \equiv K_2\left[2|V_{TD}|(V_{DD} - V_O) - (V_{DD} - V_O)^2\right] \tag{4.24}$$

Derivando ambo i membri per V_O si ha:

$$K_1\left[2(V_I - V_{T1})\frac{dV_I}{dV_O}\right] = K_2\left[-2|V_{TD}| + 2(V_{DD} - V_O)\right] \tag{4.25}$$

da cui, sostituendo la condizione sulla pendenza della (4.9) si ha:

$$K_1(V_I - V_{T1}) = K_2\left[|V_{TD}| - (V_{DD} - V_O)\right] \tag{4.26}$$

Si ottiene quindi il valore di V_{IL} dalla (4.26) come:

$$V_{IL} = V_{T1} + \frac{|V_{TD}| - (V_{DD} - V_O)}{K_R} \tag{4.27}$$

e, trascurando al numeratore della frazione il termine in parentesi rispetto al modulo di V_{TD} (in quanto V_O è alto per $V_I = V_{IH}$ e prossimo a V_{DD}), si ottiene:

$$V_{IL} \cong V_{T1} + \frac{|V_{TD}|}{K_R} \tag{4.28}$$

Determinazione di V_{OH} (NM_1, NM_2 in pinch-off)

Dall'analisi grafica di Figura 4.10 si ricava facilmente che, con $V_I = V_{OL}$ (ingresso logico basso), NM_1 è in interdizione e quindi si ha:

$$V_{OH} = V_{DD} \tag{4.29}$$

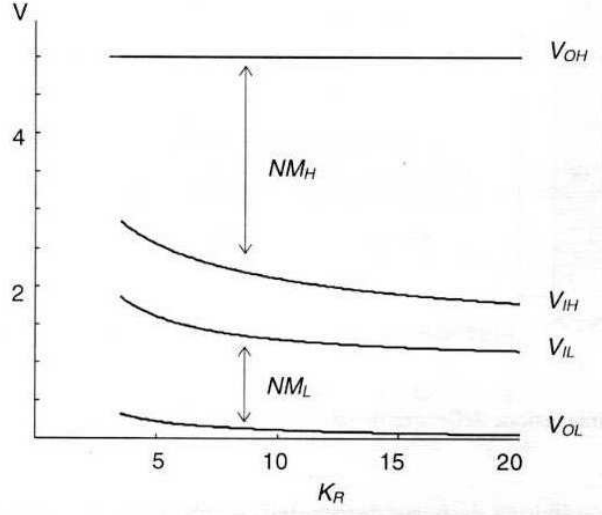


Figura 4.11 Dipendenza dei margini di rumore da K_R per un invertitore con carico a svuotamento

Anche nell'analisi dell'invertitore con carico a svuotamento si nota che il parametro progettuale che entra nelle espressioni di alcuni dei valori caratteristici di V_I e V_O è il rapporto K_R; in Figura 4.11 sono riportati i valori ottenuti dalle (4.20), (4.23), (4.28), (4.29) in funzione di K_R per un invertitore con V_{T1} = 1V, V_{TD} = −3V, V_{DD} = 5V. Anche in questo caso il margine di rumore NM_L è quello più piccolo e quindi è quello che limita le prestazioni dell'invertitore; si può vedere tuttavia che i margini di rumore per l'invertitore E-D sono meno sensibili ai valori di K_R che nel caso dell'invertitore E-E e sono ancora accettabili per valori di K_R pari a 4, mentre per l'invertitore E-E occorre arrivare a valori superiori a 10. Questa è una caratteristica importante a favore dell'invertitore E-D, perché, come vedremo, una diminuzione del K_R permette una minore occupazione di area (costo del chip) e migliori prestazioni dinamiche (velocità di operazione).

4.6 Ottimizzazione dell'area dell'invertitore NMOS

Per entrambi i casi esaminati le prestazioni statiche dell'invertitore dipendono dal rapporto K_R, che a sua volta dipende dal rapporto $(W_1/L_1)/(W_2/L_2)$ delle dimensioni geometriche delle aree di gate dei dispositivi NM_1 e NM_2. D'altra parte occorre minimizzare l'ingombro in area dell'invertitore, perché questo permette una più elevata densità di integrazione nel chip di silicio e quindi riduce il costo per porta elementare del circuito stesso.

Sorge quindi un problema di ottimizzazione delle dimensioni geometriche delle aree dei due dispositivi, una volta assegnato un determinato valore del rapporto K_R, in modo da minimizzare l'area totale dell'invertitore stesso.

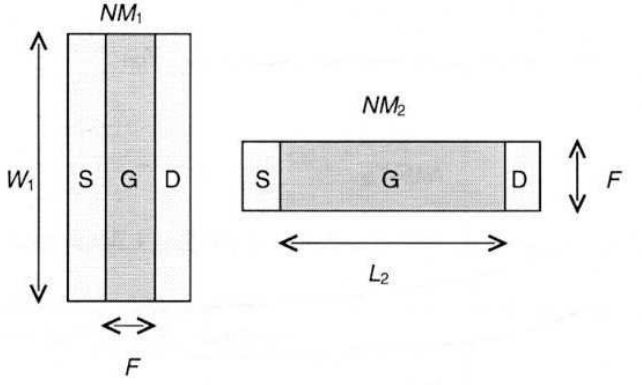


Figura 4.12 Minimizzazione delle aree di gate dell'invertitore NMOS

Una prima condizione da porre deriva dall'osservazione, fatta nel Capitolo 3, che la capacità di gate di un MOS dipende dall'area della regione di gate; poiché, come si vedrà nei paragrafi seguenti, la capacità di gate limita le prestazioni dinamiche dei circuiti logici, è importante minimizzare quest'area.

Ciò comporta che nel termine W/L una delle due grandezze geometriche venga ridotta alla minima dimensione compatibile con le regole di progetto assegnate, in modo da minimizzare l'area di gate data da $W \cdot L$ (questa dimensione minima nelle regole di progetto di Tabella 2.1 è 2λ per L mentre è 3λ per W, ma per semplicità in questa analisi verrà supposta eguale per i due casi, il che non comporta una perdità di validità dei risultati).

Questa dimensione minima, indicata con F in Figura 4.12, sarà quella L_1 per NM_1 e quella W_2 per NM_2, poiché si desidera che K_R sia elevato.

L'area dell'invertitore sarà proporzionale alla somma delle aree di gate dei due dispositivi (in quanto, per ogni dispositivo, le aree di drain e di source sono correlate all'area di gate corrispondente in base alle regole di progetto); quindi una minimizzazione dell'area complessiva delle regioni di gate, indicata con A_{GT}, mini-

mizza in definitiva l'area dell'invertitore a parità di K_R. Applicando le condizioni precedenti all'area A_{GT} si ha:

$$A_{GT} = W_1 \cdot F + F \cdot L_2$$
$$\text{con} \quad K_R = \frac{W_1}{F} \cdot \frac{L_2}{F} \tag{4.30}$$

Sostituendo nella prima delle (4.30) il valore di L_2 ricavato dalla seconda si ha:

$$A_{GT} = W_1 \cdot F + \frac{F^3 K_R}{W_1} \tag{4.31}$$

Derivando la (4.31) rispetto a W_1 ed eguagliando a zero la derivata, si ottiene il valore di W_1 che minimizza il termine A_{GT}:

$$W_{1\min} = F \cdot \sqrt{K_R} \tag{4.32}$$

e quindi sostituendo la (4.32) nella (4.31) si ottiene il valore minimo A_{GTMIN}:

$$A_{GTMIN} = F^2 \sqrt{K_R} + F^2 \sqrt{K_R} = 2F^2 \sqrt{K_R} \tag{4.33}$$

Infine per i due rapporti W_1/L_1 e W_2/L_2, ricordando la (4.32), si ha:

$$\frac{W_1}{L_1} = \sqrt{K_R} \quad ; \quad \frac{W_2}{L_2} = \frac{F \cdot W_1}{F^2 K_R} = \frac{1}{\sqrt{K_R}} \tag{4.34}$$

quindi la minimizzazione dell'area richiede, per un dato K_R, che $(W/L)_1 = (L/W)_2 = K_R^{1/2}$. In questo caso anche le aree di gate dei due MOS saranno uguali e pari alla metà dell'area minima A_{GTMIN}.

Il risultato ottenuto può essere facilmente esteso al caso in cui le dimensioni minime siano 2λ per L e 3λ per W, ottenendo al posto della (4.33):

$$A_{GTMIN} = 6\lambda^2 \sqrt{K_R} + 6\lambda^2 \sqrt{K_R} \tag{4.33b}$$

e le seguenti espressioni per i rapporti W/L dei due transistori:

$$\frac{W_1}{L_1} = \frac{3}{2}\sqrt{K_R} \quad ; \quad \frac{W_2}{L_2} = \frac{3}{2}\frac{1}{\sqrt{K_R}} \tag{4.34b}$$

quindi anche nel caso più generale di dimensioni minime diverse per L_1 e W_2, le aree delle due regioni di gate sono uguali, mentre i valori dei due rapporti W/L vanno scelti in funzione del valore di $(K_R)^{1/2}$ con un fattore correttivo pari a 3/2.

4.7 Tracciato dell'invertitore e capacità del circuito

Un esempio di tracciato per un invertitore NMOS con carico ad arricchimento è riportato in Figura 4.13 (si è adottata per semplicità una uguale dimensione minima $F = 3\lambda$ per L_1 e W_2). Le linee superiori ed inferiori corrispondono alle metallizzazioni che connettono l'invertitore rispettivamente alla tensione di alimentazione (V_{DD}) e a quella di massa (V_{SS}).

Il segnale di ingresso è applicato alla linea di polisilicio che definirà anche la regione di gate di M1, mentre quello di uscita è prelevato sulla metallizzazione che contatta anche la regione N; quest'ultima corrisponde sia alla regione di drain di M1 che a quella di source di M2, che coincidono in modo da ridurre l'area dell'invertitore.

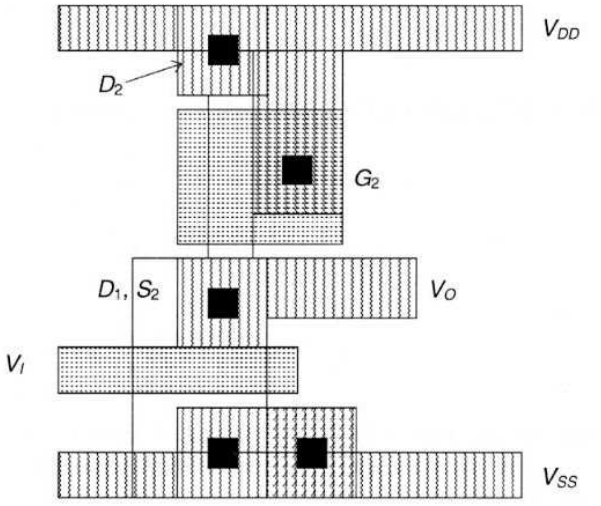


Figura 4.13 Tracciato di un invertitore NMOS con carico ad arricchimento, per $K_R = 9$, $W/L|_1 = L/W|_2 = 3$

Si può notare come l'area complessiva dell'invertitore (intesa come il rettangolo che contiene il tracciato dell'invertitore) sia ben maggiore della somma delle due aree di gate dei due MOS, a causa dell'area necessaria per le contattazioni di gate, source e drain, e di quella imposta dai vincoli delle regole di progetto. Dal tracciato è possibile ottenere le aree delle differenti regioni dei due dispositivi e quindi determinare le capacità presenti tra i terminali, in base alle (3.21), (3.24), (3.25).

Ad esempio, per l'invertitore di Figura 4.13, con $K_R = 9$, si possono determinare le dimensioni delle diverse regioni in termini di multipli di λ, e si ottiene, per le aree e i perimetri delle regioni dei due MOS:

$$A_{S1} = 9\lambda \times 7\lambda;\ A_{G1,2} = 9\lambda \times 3\lambda;\ A_{D1} + A_{S2} = 9\lambda \times 6\lambda + 3\lambda^2;\ A_{D2} = 6\lambda \times 6\lambda + 3\lambda^2$$

$$P_{S1} = 32\lambda;\quad P_{G1,2} = 24\lambda;\quad P_{D1} + P_{S2} = 32\lambda;\quad P_{D2} = 26\lambda$$

(si considera un'area complessiva per le regioni di drain del dispositivo 1 e di source del dispositivo 2, in quanto queste regioni non hanno terminali separati e quindi danno luogo ad un'unica area somma delle due, collegata tra il terminale di drain e la massa). Con questi valori delle regioni, e con i valori delle capacità unitarie (ad esempio considerando un processo con $\lambda = 1$ μm, e assumendo per le capacità unitarie i valori riportati in Tabella 3.2 per il processo a 0.6 μm), si possono determinare le capacità ai terminali dei due dispositivi NMOS dell'invertitore come:

$$C_{SB1} = C_{J0} \cdot A_{S1} + C_{JW} \cdot P_{S1} = 0.3 \cdot 63 + 0.4 \cdot 32 = 31.7 \text{ fF}$$
$$C_{G1,2} = C_{OX} \cdot A_{G1,2} + C_{GDO} + C_{GSO} + C_{GBO} \cdot P_{G1,2} = 1.7 \cdot 27 + 0.2 \cdot 24 = 50.7 \text{ fF}$$
$$C_{DB1} + C_{SB2} = C_{J0} \cdot A_{D1+S2} + C_{JW} \cdot P_{D1+S2} = 0.3 \cdot 57 + 0.4 \cdot 32 = 31.7 \text{ fF}$$
$$C_{DB2} = C_{J0} \cdot A_{D2} + C_{JW} \cdot P_{D2} = 0.3 \cdot 39 + 0.4 \cdot 26 = 22.1 \text{ fF}$$

dove i valori per le capacità di giunzione sono valutati nel caso di tensione nulla ai capi della giunzione.

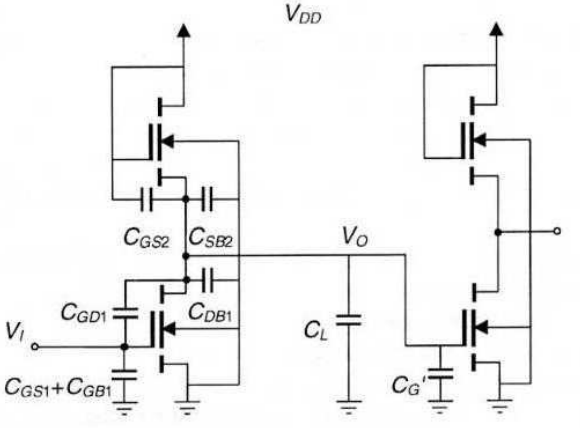


Figura 4.14 Capacità di un invertitore NMOS con carico ad arricchimento

Le capacità dei dispositivi intervengono nel determinare il funzionamento in regime dinamico dell'invertitore, in quanto richiedono del tempo per la carica o la scarica attraverso percorsi resistivi. Non tutte le capacità intervengono nel passaggio dell'invertitore da una all'altra condizione di funzionamento ($1 \rightarrow 0$ o $0 \rightarrow 1$), in quanto vanno considerate solo quelle capacità ai capi delle quali la tensione varia durante i transitori, ossia quelle connesse a nodi in cui le tensioni variano. Ad esempio nell'invertitore E-E di Figura 4.14 la capacità C_{DB2} è connessa tra due punti del circuito a tensione fissa (alimentazione e massa) e quindi non è inserita nello schema; la capacità C_{GD2} è invece cortocircuitata dalla linea di metallo. Al contrario per un invertitore E-D, essendo la gate del MOS di carico in cortocircuito con il source, comparirà la capacità C_{GD2}, mentre non vi sarà la capacità C_{GS2}.

In entrambi i casi queste capacità possono essere conglobate in un'unica capacità posta rispettivamente in ingresso (C_G) e in uscita (C_T) dello stadio invertitore, date da:

$$C_G = C_{GS1} + C_{GB1} + C_{GD1} \cdot (1 - \frac{\Delta V_O}{\Delta V_I})$$
$$C_T = C_{DB1} + C_{SB2} + C_{GS2} + C_{GD1} \cdot (1 - \frac{\Delta V_I}{\Delta V_O}) + C_L + C_G' \qquad \text{(invertitore E-E)} \quad (4.35)$$

per l'invertitore con carico ad arricchimento e

$$C_G = C_{GS1} + C_{GB1} + C_{GD1} \cdot (1 - \frac{\Delta V_O}{\Delta V_I})$$
$$C_T = C_{DB1} + C_{SB2} + C_{GD2} + C_{GD1} \cdot (1 - \frac{\Delta V_I}{\Delta V_O}) + C_L + C_G' \qquad \text{(invertitore E-D)} \quad (4.36)$$

per l'invertitore con il carico a svuotamento (in quest'ultimo caso non compare la capacità C_{GB2} perché essa è trascurabile in quanto il MOS di carico lavora in pinch-off o in regione lineare). Il termine correttivo sulla capacità C_{GD1} detto "effetto Miller" dipende dal fatto che questa capacità connette un terminale di ingresso e uno di uscita di un quadripolo, entrambi con tensioni variabili; la capacità C_G' è l'equivalente della capacità totale C_G di ingresso, riferita allo stadio invertitore collegato in uscita.

L'analisi corretta di questi pur semplici circuiti è complicata dalla dipendenza delle capacità C_{SB2}, C_{DB1} dalla tensione V_O, oltre che dall'effetto Miller su C_{GD1}, e per questo è usualmente svolta con l'ausilio di un simulatore circuitale; una trattazione analitica approssimata può farsi utilizzando il circuito semplificato di Figura 4.15a, con un'unica capacità C_T in uscita per ognuno degli stadi invertitori, data dalla (4.35) o (4.36), ma assumendo queste capacità costanti (il trascurare la dipendenza dalla tensione sulle capacità di giunzione approssima per eccesso il calcolo dei tempi di commutazione e porta ad una analisi "per il caso peggiore" – *worst-case*).

Dai valori delle diverse capacità precedentemente calcolate risulta che una componente rilevante della capacità totale C_T è fornita dalla capacità totale di gate C_G; questo è tanto più vero se il fan-out dell'invertitore è $N>>1$ (Figura 4.15b), perché in tal caso la capacità totale approssimata è:

$$C_T = C_{DB1} + C_{SB2} + C_{GS2(GD2)} + C_L + N \cdot C_G \tag{4.37}$$

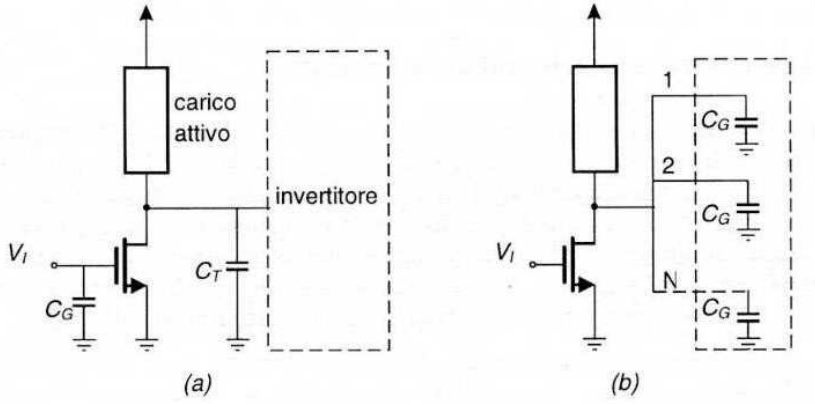


Figura 4.15 a) Circuito semplificato utilizzato per l'analisi dei transitori; b) invertitore con fan-out = N

e il termine NC_G diviene rapidamente preponderante rispetto agli altri.

Anche il valore della capacità della linea di connessione non va sottovalutato; ad esempio per una linea di larghezza 4 μm e lunghezza 200 μm con uno spessore di ossido di campo di 1 μm, dalla (3.22) si ha: C_L = 27.2 fF. La C_T viene detta anche *capacità di carico* dell'invertitore, perché l'unico carico da considerare nei circuiti MOS è l'assorbimento di corrente capacitivo nelle commutazioni, non essendovi una corrente assorbita dallo stadio a valle negli stati stazionari.

4.8 Analisi dinamica e tempi di propagazione

È possibile svolgere un'analisi approssimata della dinamica di commutazione di un invertitore elementare NMOS, considerando, come si è visto nel paragrafo precedente, una singola capacità C_T di carico, e assumendo un segnale ideale in ingresso, con tempi di salita e discesa nulli. Una ulteriore ipotesi semplificativa è quella di considerare ancora valido nelle transizioni il modello statico dei dispositivi MOS, per cui si può immaginare che il dispositivo NM_1 commuti istantaneamente dalla condizione stazionaria corrispondente al segnale basso (alto) a quella corrispondente al segnale alto (basso), e l'evoluzione della tensione V_O in uscita dall'invertitore dipenda unicamente dall'evoluzione della carica o scarica della capacità C_T.

Si deve tuttavia sottolineare che le analisi svolte con i simulatori circuitali come SPICE, mentre tengono in conto le diverse capacità con le loro dipendenze nonlineari dalla tensione, e possono considerare segnali realistici in ingresso all'invertitore, sono soggette anch'esse all'approssimazione relativa al modello quasi-statico dei dispositivi; ciò limita la validità dei risultati per segnali molto veloci, ossia con tempi di salita inferiori ai tempi di propagazione tra ingresso e uscita. In ogni caso, nelle applicazioni pratiche, il segnale di ingresso ad un circuito è fornito dall'uscita di un altro circuito ed inoltre è rallentato dalla presenza delle capacità parassite della linea, per cui l'ipotesi di quasi-stazionarietà è ragionevole.

4.8.1 Invertitore con carico ad arricchimento

Facciamo riferimento al segnale di ingresso idealizzato di Figura 4.16 applicato all'invertitore E-E; quando questo è al valore basso (V_{OL}) il MOS NM_1 è interdetto, e la capacità C_T è carica alla tensione $V_{DD} - V_{T2} = V_{OH}$. Quando l'ingresso passa al valore alto (V_{OH}), NM_1 passa dall'interdizione alla conduzione, e la capacità C_T inizia a scaricarsi attraverso NM_1 fino a raggiungere la tensione V_{OL}; successivamente, al passaggio del segnale V_I dal valore basso a quello alto, la capacità si caricherà (in un tempo maggiore, come si nota dalla Figura 4.16) fino al valore V_{OH}.

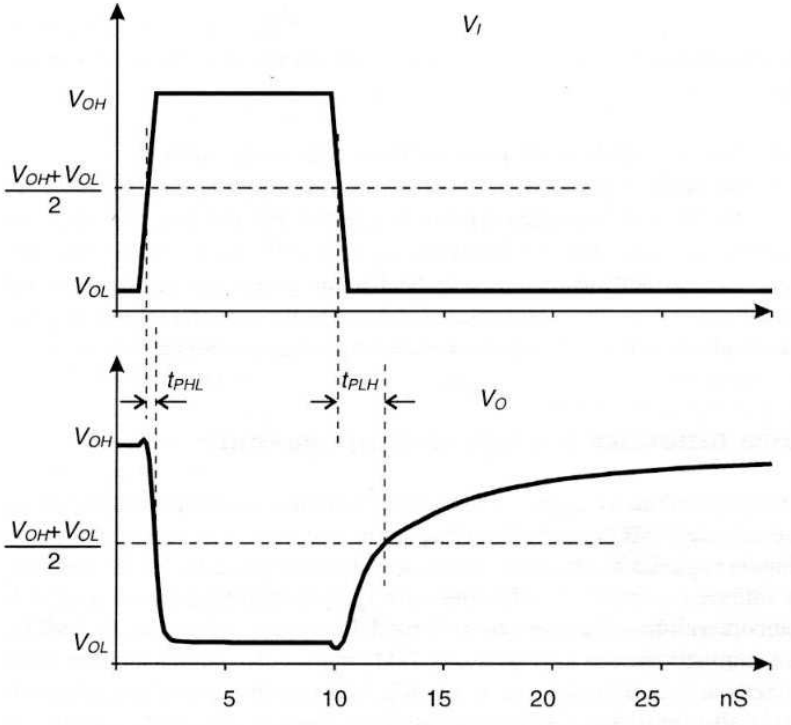


Figura 4.16 Tempi di propagazione dell'invertitore NMOS

Queste fasi sono descrivibili in base alla costruzione grafica di Figura 4.17: immediatamente prima della commutazione il punto di funzionamento è il punto A ($v = V_{OH}$, $i = 0$) in quanto il condensatore è carico e in NM_2 non circola corrente; un istante dopo la commutazione NM_1 presenta una resistenza nonlineare pari alla caratteristica corrispondente a $V_G = V_{OH}$, e il punto di funzionamento salta in B ($v = V_{OH}$, $i = I_H$), perché la capacità non può variare istantaneamente la tensione ai suoi capi. Successivamente la capacità inizia a scaricarsi attraverso NM_1; man mano che C_T si scarica la tensione ai suoi capi si riduce e il punto di funzionamento si sposta lungo la curva *I-V*. Per ogni valore di tensione, la corrente di scarica del condensatore è la differenza tra quella della curva *I-V* di NM_1 e quella della curva di carico di NM_2. Nel punto D questa differenza è nulla e l'invertitore si trova nel secondo stato stazionario, con C_T carica a V_{OL}.

Nel secondo transitorio dovuto al passaggio di V_I dal valore V_{OH} a V_{OL}, NM_1 si interdice e la capacità C_T si carica attraverso la corrente fluente attraverso NM_2. La costruzione grafica permette ancora di descrivere la fase di carica di C_T; il punto di funzionamento si sposta da D verso A percorrendo la curva di carico.

Per ogni valore della tensione di uscita la corrente di carica di C_T coincide con quella della curva di carico, e dalla figura risulta evidente che questa corrente è mediamente molto più bassa di quella di scarica. Si giustifica così l'andamento temporale della tensione di uscita di Figura 4.16 (ottenuta da una simulazione SPICE per un invertitore E-E con $K_R = 9$), che mostra un transitorio di scarica di C_T ben più rapido di quello di carica. Infatti poiché, in base alla relazione tra tensione e corrente ai capi di una capacità, i tempi dei transitori capacitivi sono inversamente proporzionali alle correnti medie:

$$\Delta T \equiv \int_{t_1}^{t_2} dt = C_T \int_{V_1}^{V_2} \frac{dV}{i} \equiv \frac{C_T}{< I >} (V_2 - V_1) \tag{4.38}$$

l'intervallo di tempo ΔT relativo alla carica risulta più elevato di quello di scarica essendo la carica effettuata ad una corrente media $<I>$ più bassa.

I tempi di propagazione t_{PHL} e t_{PLH} si definiscono come i tempi necessari per il passaggio rispettivamente da B a C o da D a E; si può approssimare per eccesso la corrente nell'integrale a secondo membro della (4.38) con un valore costante rispettivamente I_H (trascurando la corrente in NM_2 rispetto a quella in NM_1) o I_L (vedi Figura 4.17), ottenendo per i due tempi di propagazione le relazioni approssimate:

$$t_{PHL} = \frac{C_T \Delta V}{< I_{Q1} - I_{Q2} >} > \frac{C_T}{I_H} \left(V_{OH} - \frac{V_{OH} + V_{OL}}{2} \right) = \frac{C_T}{2 I_H} (V_{OH} - V_{OL})$$

$$t_{PLH} = \frac{C_T \Delta V}{< I_{Q2} >} > \frac{C_T}{I_L} \left(\frac{V_{OH} + V_{OL}}{2} - V_{OL} \right) = \frac{C_T}{2 I_L} (V_{OH} - V_{OL}) \tag{4.39}$$

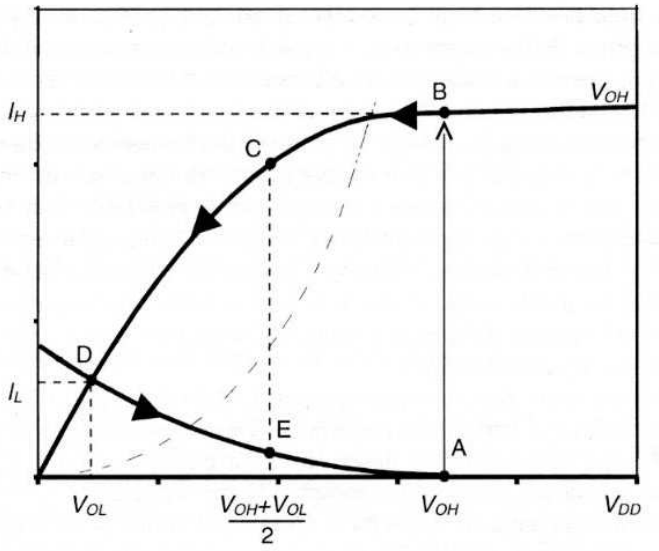


Figura 4.17 Analisi grafica dei transitori di commutazione in un invertitore E-E

da cui, ricavando I_H e I_L dalle espressioni delle correnti di NM_1 e NM_2, si ha:

$$\frac{t_{PLH}}{t_{PHL}} \cong \frac{I_H}{I_L} \cong K_R \frac{(V_{OH} - V_{T1})^2}{(V_{OH} - V_{OL})^2} \tag{4.40}$$

Dalla (4.40) ne consegue che, per valori di K_R superiori a 10, il tempo di propagazione totale t_P si può approssimare a:

$$t_P = \frac{1}{2}(t_{PLH} + t_{PHL}) \cong \frac{t_{PLH}}{2} = \frac{C_T}{4K_2(V_{OH} - V_{OL})} \tag{4.41}$$

Si può avere una espressione più corretta di t_{PLH} risolvendo l'integrale nella (4.38) tra i limiti V_{OL} e $V_O^* = 1/2(V_{OH} + V_{OL})$:

$$t_{PLH} = C_T \int_{V_{OL}}^{V_O^*} \frac{dV_O}{K_2(V_{DD} - V_O - V_{T2})^2} \tag{4.42}$$

Assumendo V_{T2} costante e ponendo $V_{DD} - V_{T2} = V_{OH}$ l'integrale vale:

$$t_{PLH} = \frac{C_T}{K_2}\left[\frac{1}{V_{OH} - V_O}\right]_{V_{OL}}^{V_O^*} = \frac{C_T}{K_2}\frac{1}{(V_{OH} - V_{OL})} \tag{4.43}$$

4.8.2 Invertitore con carico a svuotamento

Anche in questo caso l'analisi grafica (riportata in Figura 4.18) è di aiuto nella comprensione dei transitori in uscita dall'invertitore E-D. Le principali differenze rispetto al caso dell'invertitore E-E sono dovute a) all'aumento della tensione V_{OH} che nell'invertitore E-D coincide con quella di alimentazione e b) ad un percorso di carica (tratto D-A) con corrente mediamente più elevata (a parità di valore I_L), dovuta alla differente forma della curva di carico. Entrambe queste cause fanno sì che l'approssimazione di carica e scarica della capacità a corrente costante negli intervalli B-C e D-E (utilizzati per la definizione dei tempi di propagazione), sia più vicina al comportamento reale del circuito.

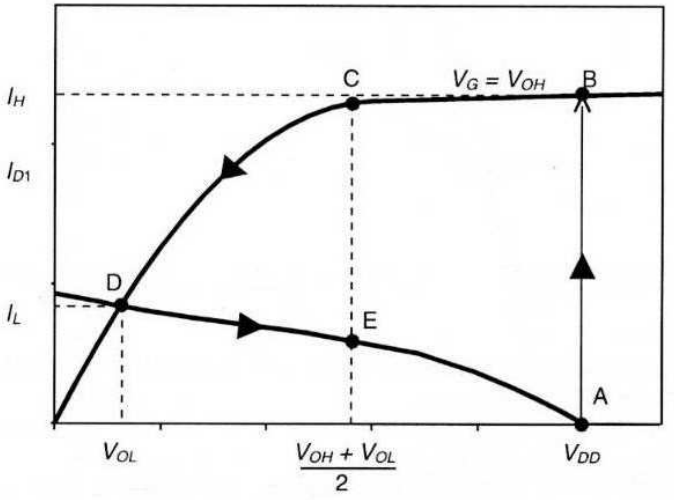


Figura 4.18 Analisi grafica dei transitori in un invertitore E-D

Il confronto tra le forme d'onda in uscita (Figura 4.19) di un invertitore E-E (curva tratteggiata) ed E-D (curva continua), entrambi con $K_R = 4$, mostra che le differenze di comportamento sono più grandi di quanto emerga dal confronto dei tempi di propagazione, in particolare per l'andamento asintotico della transizione $0 \rightarrow 1$ nel caso E-E (dovuta al fatto che approssimandosi alla tensione alta, la corrente della curva di carico tende progressivamente a zero), laddove l'invertitore E-D presenta una crescita più regolare della tensione (dovuta ad una corrente di carica relativamente poco variabile fin nei pressi della tensione V_{DD}).

I tempi di propagazione possono essere calcolati allo stesso modo che per il caso precedente, in base alla (4.39), sostituendo le espressioni corrispondenti per i valori I_H e I_L (quest'ultima è ora dipendente dalla caratteristica I-V in regione di pinch-off del NMOS a svuotamento utilizzato come carico):

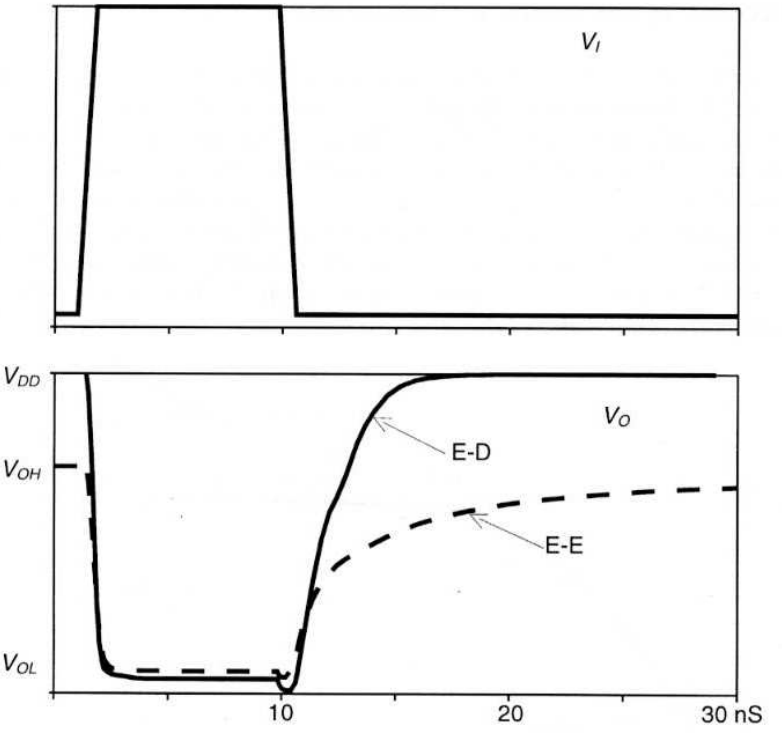


Figura 4.19 Confronto tra il comportamento dinamico di un invertitore E-E ed E-D, entrambi con $K_R = 4$

$$t_{PHL} = \frac{C_T (V_{OH} - V_{OL})}{2 < i_1 - i_2 >} \cong \frac{C_T}{2 I_H} (V_{OH} - V_{OL}) = C_T \frac{(V_{DD} - V_{OL})}{2 K_1 (V_{DD} - V_{T1})^2}$$

$$t_{PLH} \cong \frac{C_T (V_{OH} - V_{OL})}{2 I_L} = C_T \frac{(V_{DD} - V_{OL})}{2 K_2 |V_{TD}|^2} \tag{4.44a}$$

$$\frac{t_{PLH}}{t_{PHL}} \cong \frac{I_H}{I_L} = K_R \frac{(V_{DD} - V_{T1})^2}{|V_{TD}|^2} > K_R \tag{4.44b}$$

Quindi, anche per questo invertitore, il tempo di propagazione t_{PLH} è più elevato, per cui in prima approssimazione si può definire il ritardo di propagazione t_P come:

$$t_P \cong \frac{t_{PLH}}{2} = \frac{C_T}{4 K_2} \frac{(V_{DD} - V_{OL})}{|V_{TD}|^2} \tag{4.45}$$

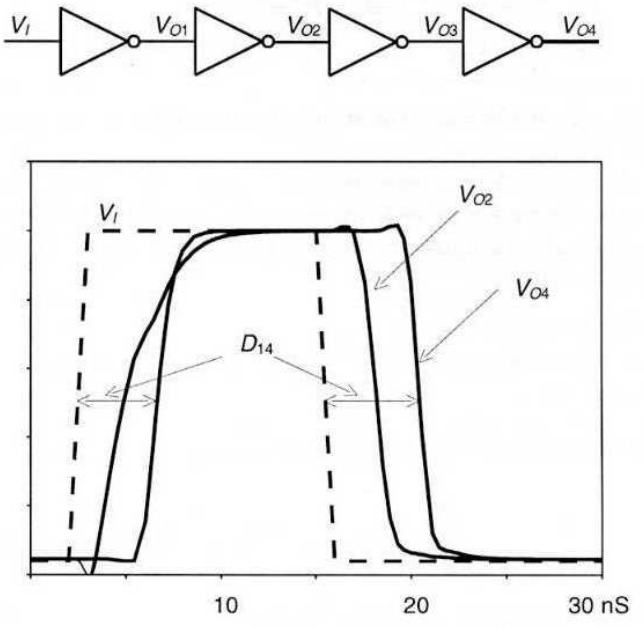


Figura 4.20 Simulazioni SPICE delle uscite da una catena di invertitori E-D con $K_R = 4$

Il significato del *ritardo di propagazione* diviene più evidente con riferimento alle relazioni temporali delle uscite di una catena di più invertitori, come è ad esempio riportato in Figura 4.20. L'uscita del primo invertitore pilota il secondo e così via, per cui le forme d'onda in uscita dagli invertitori successivi tendono a simmetrizzarsi, e vengono ritardate nel passaggio in ogni stadio rispetto all'ingresso di un ritardo pari a t_P (nel caso di Figura 4.20 il ritardo di propagazione t_P per ogni stadio è di 1.25 ns e il ritardo D_{14} dopo quattro stadi è di 4.9 ns).

4.9 Potenza dissipata dall'invertitore

Una caratteristica rilevante ai fini dell'integrazione introdotta nel Paragrafo 1.4 è la dissipazione di potenza per singola porta logica; questo dato determina in prima approssimazione il massimo numero di porte integrabili in un singolo chip.

La potenza dissipata statica dipende dallo stato logico in cui si trova l'invertitore. Per un ingresso basso ($V_I = V_{OL}$), NM_1 è interdetto e quindi non circola corrente nel carico per cui $P_D = 0$. L'invertitore consuma potenza solo quando l'ingresso è nello stato alto ($V_I = V_{OH}$), e questa potenza è data dal prodotto della corrente che circola in NM_2 per la tensione di alimentazione:

$$\begin{aligned} &\text{per } V_I = V_{OL} \quad \Rightarrow \quad P_D = 0 \\ &\text{per } V_I = V_{OH} \quad \Rightarrow \quad P_D = I_{D2}(V_{OL}) \cdot V_{DD} \end{aligned} \tag{4.46}$$

dove la corrente I_{D2} va calcolata in base al tipo di dispositivo di carico e per una tensione di uscita pari a V_{OL}.

In mancanza di informazioni dettagliate sulla sequenza di stati logici applicati all'ingresso si assume in base a considerazioni statistiche che il tempo medio di permanenza nello stato alto sia uguale a quello nello stato basso per cui la potenza media dissipata nell'invertitore è:

$$P_{Dmedia} = \frac{P_D}{2} = \frac{I_{D2} \cdot V_{DD}}{2} \tag{4.47}$$

Per l'invertitore ad arricchimento la (4.47) fornisce:

$$P_{D(E-E)} = \frac{K_2}{2}(V_{DD} - V_{OL} - V_{T2})^2 \cdot V_{DD} \tag{4.48}$$

Per l'invertitore a svuotamento dalla (4.47) si ha:

$$P_{D(E-D)} = \frac{K_2}{2}|V_{TD}|^2 \cdot V_{DD} \tag{4.49}$$

4.10 Prodotto ritardo-potenza dissipata

Confrontando le espressioni (4.41) o (4.45) relative al ritardo di propagazione dell'invertitore E-E o E-D con quelle (4.48) o (4.49) per la potenza dissipata, si vede come il parametro K_2 del dispositivo di carico giochi in maniera duale nei due casi; in altre parole un aumento di K_2 riduce il ritardo di propagazione ma aumenta la dissipazione di potenza. Si comprende così come il prodotto *ritardo-potenza dissipata* costituisca un parametro globale di merito delle porte logiche; questo prodotto vale, sia per l'invertitore E-E che per quello E-D:

$$P_{Dmedia} \cdot t_P \equiv P \cdot D = \frac{C_T}{8}(V_{OH} - V_{OL}) \cdot V_{DD} \tag{4.50}$$

e dipende solo dalla tensione di alimentazione e dalla capacità di carico, ma non da K_R. Ciò vuol dire che è possibile utilizzare come scelta progettuale il valore di K_R per ridurre il ritardo di propagazione a spese della potenza dissipata (logiche veloci) o viceversa (logiche a basso consumo). L'apparente indipendenza del prodotto PD dal parametro K_R va tuttavia corretta se si tiene conto che la capacità C_T è anch'essa funzione dell'area di gate $W \cdot L$ del MOS invertitore e quindi, attraverso la (4.33), dipende da $(K_R)^{1/2}$; ad esempio, per un invertito-

re E-D con $L_1 = W_2 = F = 5$ μm, e per $V_{DD} = 5$ V, i risultati delle simulazioni SPICE per due diversi valori di K_R forniscono:

$K_R = 4$	$t_P = 1.7$ ns	P_D media = 120 mW	$P.D = 0.2$ pJ
$K_R = 16$	$t_P = 6.6$ ns	P_D media = 62 mW	$P.D = 0.41$ pJ

da cui si vede che il prodotto $P{\cdot}D$ varia secondo il fattore $(K_R)^{1/2}$ come previsto dalle relazioni approssimate precedentemente introdotte.

La (4.50) mostra anche che una riduzione della tensione di alimentazione comporta un miglioramento del prodotto $P{\cdot}D$ all'incirca secondo il quadrato di tale diminuzione; questo spiega la tendenza, nell'evoluzione dei circuiti logici, ad una riduzione di quest'ultima, che è passata da 10 V agli attuali 5 V, e sono già presenti sul mercato le nuove famiglie di circuiti da 3.3 V.

4.11 Porte logiche elementari NMOS

Sulla base dell'invertitore elementare studiato nei paragrafi precedenti è immediato realizzare le porte elementari NOR e NAND, utilizzando i principi presentati nel Paragrafo 1.5 per la realizzazione di porte logiche sulla base dell'invertitore ideale con interruttore controllato.

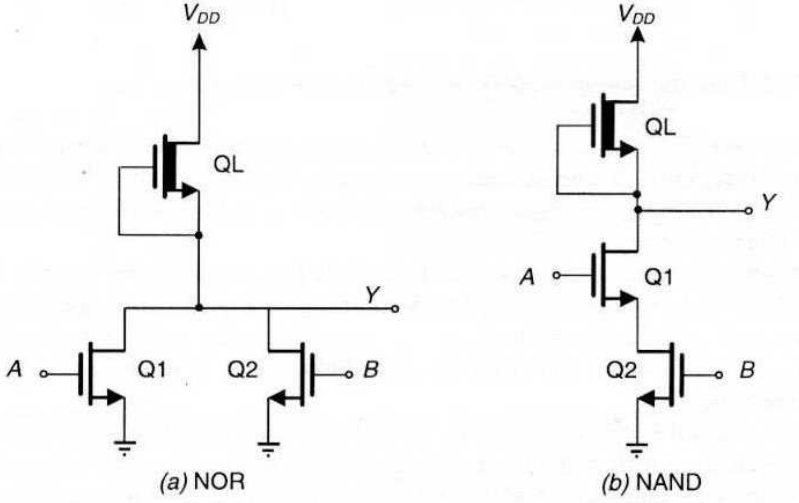


Figura 4.21 a) Porta NOR NMOS a due ingressi; b) porta NAND a due ingressi

Facendo quindi riferimento all'invertitore con carico a svuotamento (ma nulla cambia se ci riferisce all'invertitore E-E), la funzione (e quindi la porta) NOR viene realizzata ponendo più invertitori *in parallelo* per quanto riguarda le uscite, con un unico MOS di carico per tutti gli invertitori, come riportato in Figura 4.21a per il caso di una porta NOR a due ingressi. Da tale schema è immediato vedere come l'uscita Y sia bassa (stato 0) se Q1 o Q2 (o entrambi) sono in conduzione, cioè se gli ingressi A o B (o entrambi) sono alti (stato 1), mentre occorre che sia A che B siano bassi perché l'uscita sia alta. Le condizioni su A, B e Y corrispondono alla tabella della verità di una porta NOR (vedi Figura 1.12), e possono essere estese direttamente al caso di N ingressi ponendo in parallelo nello schema di Figura 4.21 N transistori invece di due.

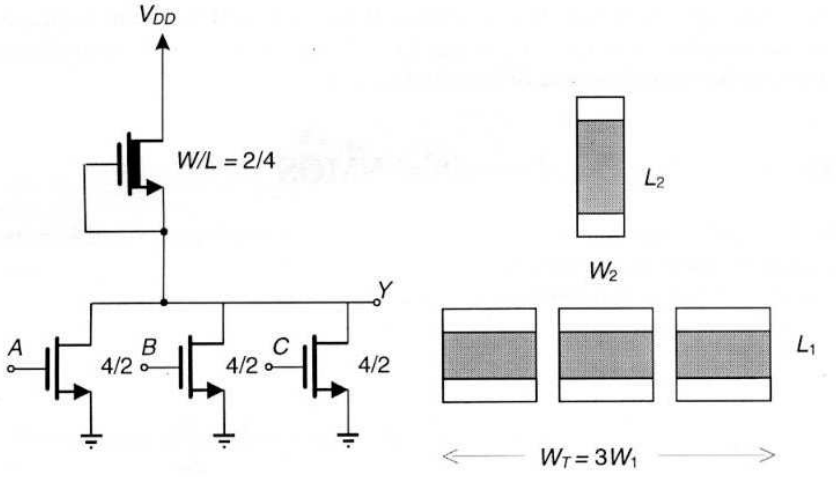


Figura 4.22 Dimensionamento dei MOS in una porta NOR a tre ingressi

La versione circuitale di una porta NAND segue ancora il criterio generale presentato nel Paragrafo 1.5, che richiede in questo caso di porre *in serie* gli interruttori, con un unico carico, ed è implementata nello schema di Figura 4.21b per il caso di porta a due ingressi.

Il dimensionamento di una porta NOR richiede la definizione dei rapporti W/L dei singoli transistori in base a considerazioni analoghe a quelle svolte per l'invertitore, ma nelle porte a più ingressi occorre tener conto che nello stato di uscita bassa (V_{OL}) i transistori in conduzione nel parallelo variano a seconda dei segnali applicati agli ingressi.

Il caso peggiore per cui va dimensionata la porta corrisponde ad un solo segnale alto ad uno degli N ingressi, con gli altri a livello basso; in tal caso la porta si comporta come un invertitore e le caratteristiche statiche e dinamiche sono definite dal rapporto K_R tra i due transistori attivi (nel caso di Figura 4.22 $K_R = 4$). Se più di un ingresso è al valore alto, il valore della tensione V_{OL} in uscita decresce perché aumenta il numero di transistori in parallelo che sono in

conduzione; se ad esempio nel caso di Figura 4.22 tutti e tre gli ingressi sono alti, la porta equivale ad un invertitore ottenuto sostituendo al parallelo dei tre MOS inferiori un unico MOS che presenta una $W_T = 3W_1$ e $L_T = L_1$; il rapporto K_R sarà quindi:

$$K_{Req} = \frac{N \cdot W_1}{L_1} \frac{L_2}{W_2} = N \cdot K_R \tag{4.51}$$

In generale per una porta NOR a N ingressi, con J ingressi alti, i margini di rumore sono quelli corrispondenti a un invertitore equivalente con $K_{Req} = JK_R$, dove K_R corrisponde a quello con due soli MOS basso e alto (lo stesso vale per i tempi di propagazione); tuttavia il dimensionamento della porta va fatto rispetto alla condizione peggiore, che è quella, come si è visto, con un solo ingresso alto, e quindi in questo caso si può dire che la porta NOR si comporta come un invertitore con uno degli N NMOS come dispositivo attivo.

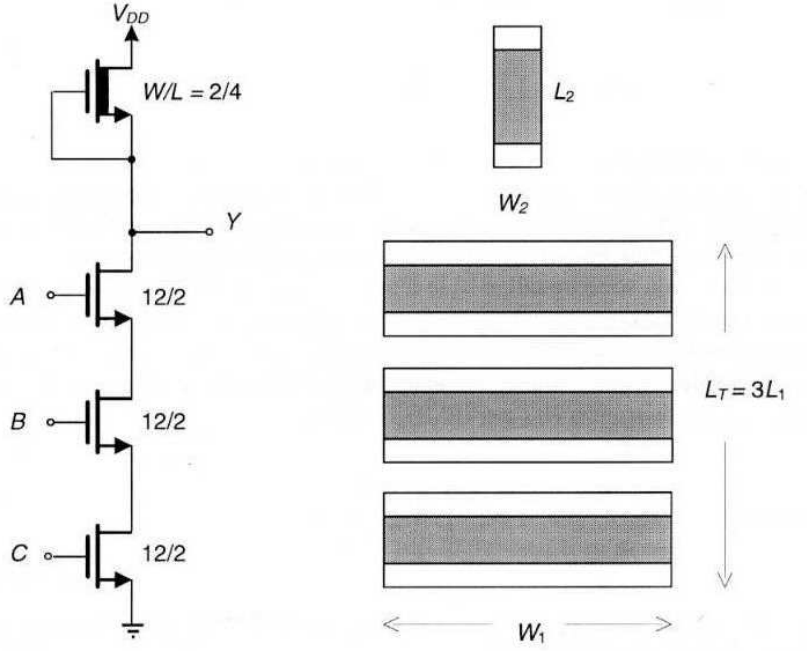


Figura 4.23 Dimensionamento dei MOS in una porta NAND a tre ingressi

Nel caso delle porte NAND si possono applicare considerazioni analoghe, che però portano a risultati differenti. Infatti in questo caso la (unica) condizione di uscita bassa è quella con tutti gli ingressi alti; la porta si può quindi considerare equivalente ad un unico invertitore che come dispositivo invertitore ha

un MOS equivalente alla serie degli N MOS. In tal caso (vedi Figura 4.23) si può considerare che il MOS equivalente (con tutti e tre gli ingressi allo stesso potenziale) abbia una regione di gate complessiva con una $L_T = NL_1$ e con $W_T = W_1$. Per avere una tensione V_{OL} uguale a quella della porta NOR (nel caso peggiore per quest'ultima), la porta NAND deve quindi avere un uguale K_R e cioè:

$$K_{R(NAND)} \equiv \frac{W_1}{N \cdot L_1}\frac{L_2}{W_2} = K_{R(NOR)}\Big|_{N=1} \equiv \frac{W_1}{L_1}\frac{L_2}{W_2} \Rightarrow W_{1\,NAND} = N \cdot W_{1\,NOR} \quad (4.52)$$

La condizione imposta dalla (4.52) comporta che, a parità di prestazioni elettriche delle due porte (con uguale numero di ingressi) una porta NAND occuperà un'area maggiore di quella di una porta NOR, come è sinteticamente indicato nelle Figure 4.22 e 4.23; infatti, ricordando che nel dimensionamento ad area minima l'area di gate del MOS di carico è uguale a quella del MOS pilotato per qualunque K_R, si ha:

$$\begin{aligned} A_{NOR} &= (W_2 \cdot L_2) + N \cdot (W_1 \cdot L_1) = A_{\min}(1+N) \\ A_{NAND} &= (W_2 \cdot L_2) + N \cdot (NW_1 \cdot L_1) = A_{\min}(1+N^2) \end{aligned} \quad (4.53)$$

e la differenza in area diventa rapidamente crescente per N>2.

L'ipotesi di considerare una serie di N MOS (con le gate allo stesso potenziale) come un unico MOS con lunghezza di gate N volte maggiore di quella del singolo MOS è una approssimazione, in quanto non si tiene conto che a) le tensioni V_{GS} dei diversi dispositivi non sono uguali, e b) le tensioni V_{SB} (per i MOS che non hanno il source connesso a massa) sono diverse da zero, e quindi vi è un effetto substrato che aumenta per i MOS nella serie più lontani dalla massa. Nell'ipotesi di funzionamento dei MOS nel tratto lineare (ottenuto dalla Equazione (3.10a) con $V_{DS} \to 0$) si può definire una resistenza R_{ON} per il MOS data da:

$$R_{ON} = \frac{V_{DS}}{I_D} = \frac{1}{k'\dfrac{W}{L}2(V_{GS} - V_T)} \quad (4.54)$$

In tal caso, per la serie di N MOS la resistenza totale tra uscita e massa è data dalla somma delle resistenze e quindi:

$$R_{TON} = NR_{ON} = N\frac{L}{k'W2(V_{GS} - V_T)} \Rightarrow L_T = NL \quad (4.55)$$

In effetti l'approssimazione citata è ragionevole anche per tensioni di drain non trascurabili, e per un funzionamento dei MOS in regime nonlineare, come si può vedere dai risultati di Figura 4.24, dove si è confrontata la caratteristica I-V di un

transistore NMOS con rapporto $W/L = 1/3$ con quella complessiva della serie di 3 NMOS ognuno dei quali con rapporto $W/L = 1$; si può vedere che delle due cause citate, l'effetto di substrato è quella più rilevante nel modificare la caratteristica complessiva della serie, tuttavia l'errore commesso, per quanto riguarda la corrente, non è eccessivo anche nella regione di pinch-off. Questa osservazione sarà utile quando si analizzeranno le porte logiche CMOS a più ingressi.

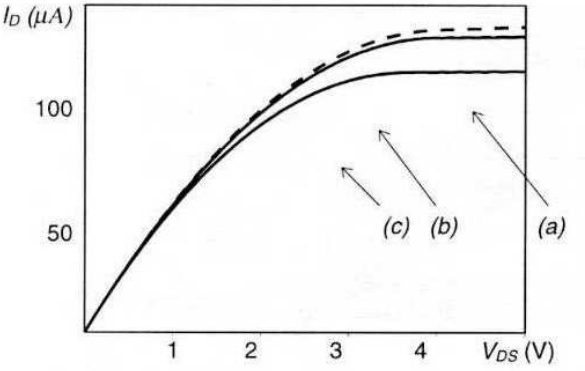


Figura 4.24 Caratteristiche *I-V* di a) un NMOS con W/L = 1/3; b) una serie di 3 NMOS con W/L = 1 senza effetto substrato; c) come per a) ma tenendo conto dell'effetto substrato

Anche il confronto delle prestazioni dinamiche è a favore della porta NOR. Infatti la capacità di ognuno degli ingressi A, B, C della porta NAND di Figura 4.23 è maggiore di un fattore N di quella del rispettivo ingresso della porta NOR di Figura 4.22, in quanto l'area della regione di gate è cresciuta di questo fattore; ne consegue un ritardo di propagazione (a parità del valore K_2 del MOS di carico) N volte maggiore per la porta NAND che per quella NOR.

Queste considerazioni spiegano perché per i circuiti con tecnologia NMOS venga preferita la realizzazione di circuiti logici *basati su porte NOR*.

4.12 Tracciati delle porte logiche NMOS

Il tracciato di una porta NOR in tecnologia NMOS può essere sviluppato sulla base del circuito di Figura 4.21a e delle considerazioni svolte per le dimensioni di gate dei transistori MOS coinvolti.

In Figura 4.25 è riportato il tracciato di una porta NOR a due ingressi, utilizzando le regole di progetto del Capitolo 2 con $W_{MIN} = 3\lambda$ e $L_{MIN} = 2\lambda$; in tal caso quindi, in base ai valori di W/L estratti dalle Espressioni (4.34b) per il progetto ad

area minima, si ottiene, per un valore di $K_R = 4$, un rapporto $W/L = 6\lambda/2\lambda$ per i MOS Q1 e Q2, e $W/L = 3\lambda/4\lambda$ per il MOS QL.

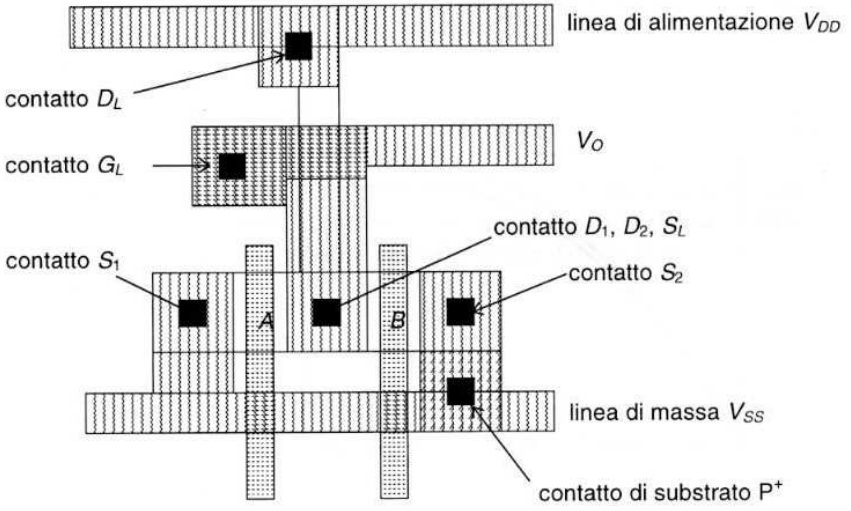


Figura 4.25 Tracciato della porta NOR NMOS con carico a svuotamento relativo al circuito di Figura 4.21(a) con $K_R = 4$: $W/L|_{QL} = 3/4$, $W/L|_{Q1,Q2} = 2/6$

La scelta di fondere insieme le aree dei drain dei MOS pilotati con quella di source del MOS di carico (MOS a svuotamento) è dettata dalla convenienza di ridurre la capacità delle regioni di drain e di source presente al nodo di uscita, in modo da diminuire la capacità totale in uscita C_T e migliorare le prestazioni dinamiche. Il polisilicio della gate di QL è contattato mediante metallo e collegato al source S_L; la stessa linea di metallo è utilizzata per prelevare il segnale di uscita V_O. I due source di Q1 e Q2 sono collegati tra loro e alla massa (V_{SS}) da una linea di metallo, che passa sopra le linee in polisilicio a cui vengono applicati i segnali di ingresso, in quanto il polisilicio e il metallo sono tra loro separati da uno strato intermedio di ossido (vedi il Capitolo 2) e quindi possono intersecarsi tra loro.

Il tracciato di una porta NAND a due ingressi è riportato in Figura 4.26 per il circuito di Figura 4.21b, utilizzando anche in questo caso le Espressioni (4.34b) per i valori di W/L dei MOS secondo il progetto ad area minima. In questo caso, assumendo ancora per K_{Req} un valore di 4 (nel caso in cui tutti e due i MOS conducono), il MOS QL avrà ancora un valore $W/L = 3\lambda/4\lambda$, mentre per ognuno dei MOS in serie occorre scegliere un valore $W/L = 12\lambda/2\lambda$, tale che con i due MOS in serie si abbia un $W/L_{eq} = 6\lambda/2\lambda$.

Anche in questo caso, nel tracciato l'area del source di QL è conglobata con quella del drain di Q1 in modo da ridurre la capacità parassita al nodo di uscita e migliorare le prestazioni dinamiche. Un'ulteriore compattazione è possibile nella realizzazione della serie di MOS, in quanto non occorre prevedere una regione di contattazione tra il drain di Q2 e il source di Q1, e le linee di polisilicio per le regioni di gate possono essere spaziate tra loro con una distanza minima di 3λ; quindi l'area totale occupata da N transistori MOS in serie è certamente minore di quella occupata da N transistori MOS in parallelo, e questo compensa in parte l'aumento di area di gate per ciascuno dei MOS di una porta NAND rispetto al caso NOR.

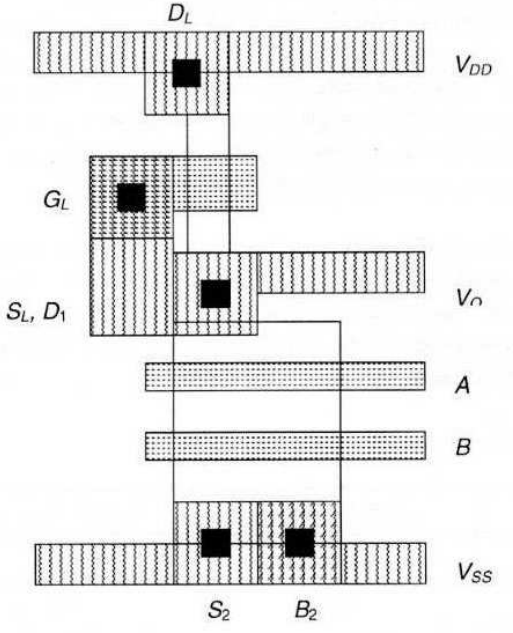


Figura 4.26 Tracciato della porta NAND NMOS con carico a svuotamento relativo al circuito di Figura 4.21b con $K_R = 4$: $W/L|_{QL} = 3/4$, $W/L|_{Q1,Q2} = 2/12$

Esercizi di riepilogo

4.1 Determinare il valore della resistenza di carico di un invertitore NMOS con carico resistivo che impiega un NMOS con: $k' = 40\ \mu A/V^2$; $V_{TO} = 0.8V$ e $W/L = 4$, per ottenere un valore di $V_{OL} = 0.1$ V.

4.2 Per l'invertitore NMOS con carico nonlineare definito da una caratteristica *I-V* del tipo riportato in Figura E4.1, e con il transistore NMOS caratterizzato da $k' = 40\ \mu A/V^2$; $V_{TO} = 0.8V$ e $W/L = 6$, determinare il valore del livello logico basso V_{OL} e della dissipazione di potenza statica P_D.

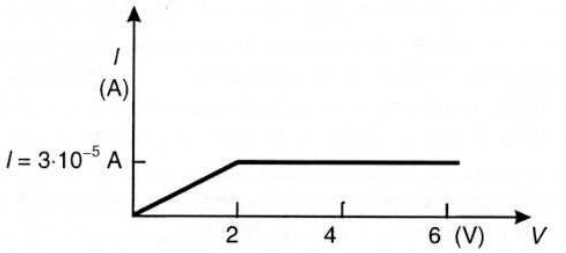


Figura E4.1

4.3 Considerando un invertitore con carico ad arricchimento, con k'= 40 μA/V^2, V_{TO} = 0.8V per entrambi i MOS, V_{DD} = 5 V, e con un rapporto W/L = 6/2 per il MOS attivo e 2/6 per quello di carico: a) valutare i livelli logici e i margini di rumore mediante le espressioni approssimate da (4.8) a (4.16) in assenza di effetto substrato; b) valutare gli stessi con le espressioni precedenti in presenza di effetto substrato, assumendo una variazione di della tensione di soglia data dalla Equazione (4.17), con γ= 0.5 e ϕ^* = 0.6V.

4.4 Per un invertitore NMOS con carico a svuotamento con i valori riportati in Tabella E4.1 e per V_{DD} = 5 V, determinare il valore del rapporto K_R affinché si abbia un valore V_{OL} = 0.2 V (si trascuri l'effetto di substrato). Utilizzando le relazioni per un progetto ad area minima, ed assumendo un valore λ = 1 μm, determinare il valore della potenza statica P_D.

Tabella E4.1

	NMOS attivo	*NMOS carico*
k'	40 μA/V^2	40 μA/V^2
V_{TO}	0.8 V	–3 V
C_{OX}	1 fF/μm^2	1 fF/μm^2

4.5 Utilizzando il tracciato dell'invertitore NMOS riportato in Figura 4.13, assumendo per i parametri dei MOS k' = 40 μA/V^2, V_{TO} = 0.8V, per la tensione di alimentazione V_{DD} = 5 V, e per le capacità unitarie i valori riportati in Tabella 3.2 relativi al processo con λ = 0.6 μm, valutare il tempo di propagazione per via analitica utilizzando come capacità complessiva di carico a) la sola capacità di gate dell'invertitore a valle, e b) quella determinata considerando tutte le capacità che insistono sul nodo di uscita senza considerare il fattore correttivo della tensione per le capacità di giunzione. Paragonare i risultati ottenuti nei due casi con il tempo di propagazione determinato mediante una simulazione SPICE del circuito in esame, utilizzando le schede .MODEL dei dispositivi con i parametri utilizzati per l'analisi manuale.

4.6 Considerando un invertitore analogo a quello dell'Esercizio 4.3, con un rapporto $W/L = 4/2$ per il MOS attivo e 2/4 per quello di carico, valutare per via analitica il ritardo di propagazione t_P quando l'invertitore è caricato da una capacità di carico $C_T = 1$ pF (si assumono trascurabili i contributi delle capacità dei MOS dell'invertitore).

4.7 Valutare per via analitica il tempo di propagazione t_P dell'invertitore con carico a svuotamento con i parametri dei transistori riportati in Tabella E4.1, con $W/L|_1 = L/W|_2 = 2$, e con CDD = 5 V, quando questo è caricato da un uguale invertitore (assumendo trascurabile il contributo delle capacità di source e di drain dei MOS). Si determini inoltre il fan-out ammissibile per un ritardo di propagazione massimo di 1 nS.

4.8 Dimensionare i transistori MOS per realizzare una porta NOR a 3 ingressi con carico a svuotamento che presenti un $K_R = 9$ nel caso peggiore. Determinare per via analitica i valori dei livelli logici nominali e i margini di rumore, per un ingresso A che passa da 0 a 1, con $B = C = 0$, e per tutti gli ingressi che passano da 0 a 1. Determinare infine, mediante il simulatore SPICE, le caratteristiche di trasferimento e i livelli logici per i due casi precedenti.

4.9 Dimensionare i transistori MOS per una porta NAND a 3 ingressi con carico a svuotamento, che presenti un $K_R = 9$ nel caso peggiore. Determinare per via analitica i valori dei livelli logici nominali e dei margini di rumore, per un ingresso A che passa da 0 a 1, con $B = C = 1$, e per tutti gli ingressi che passano da 0 a 1. Confrontare i risultati con quelli ottenuti con il simulatore SPICE.

4.10 Determinare i tempi di propagazione t_{PHL} e t_{PLH} della porta NOR dell'Esercizio 4.8 mediante le formule analitiche approssimate, nei seguenti casi: 1) A 0→1, $B = C = 0$; 2) A, B 0→1, $C = 0$; 3) A, B, C 0→1; confrontare i risultati con quelli ottenuti mediante simulazioni SPICE.

Riferimenti bibliografici

A.S. Sedra, K.C. Smith, *Microelectronic Circuits*, Saunders College publ., 1991.

B. Riccò, F. Fantini, P. Brambilla, *Introduzione ai circuiti integrati digitali*, Zanichelli, Bologna, 1991.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

5

Porte elementari CMOS

5.1 Il processo CMOS

Lo sviluppo della tecnologia MOS, ed in particolare la capacità di controllare il valore della tensione di soglia attraverso una migliore qualità dell'ossido di gate e una modifica del drogaggio sottostante l'ossido attraverso l'impiantazione, ha portato negli anni '70 al processo CMOS (*Complementary-MOS*) che rende possibile l'integrazione contemporanea nello stesso wafer di dispositivi sia PMOS che NMOS.

Per poter consentire la realizzazione di dispositivi sia a canale N che P nello stesso substrato, occorre modificare in via preliminare il drogaggio di una regione del wafer, nella quale si deve realizzare il dispositivo con canale dello stesso segno del substrato di partenza. La principale modifica introdotta quindi da questo processo, rispetto al processo NMOS schematizzato in Figura 2.2, è legata alla realizzazione di queste regioni, dette tasche (*well*), nelle aree dove dovranno essere realizzati i transistori di canale duale rispetto a quello permesso dal substrato utilizzato. In Figura 5.1 è riportata sinteticamente la sequenza di passi di processo per la realizzazione di un processo CMOS a partire da un substrato di tipo P, utilizzando una tasca di tipo N per la realizzazione dei dispositivi PMOS; questo processo è indicato come processo a tasca N (*N-well*), e realizza i dispositivi NMOS direttamente nel substrato di partenza e quelli PMOS nelle tasche N. La sequenza delle operazioni (con riferimento alla Figura 5.1) è:

1. apertura delle aree nell'ossido di campo per la realizzazione dei dispositivi NMOS e della tasca N;
2. impiantazione di drogante N per la tasca N (l'area del NMOS è protetta dal fotoresist);
3. crescita dell'ossido sottile di gate;
4. deposizione e definizione del polisilicio (che agisce da maschera per le successive impiantazioni);

5. impiantazione di drogante N per la realizzazione delle regioni di source e drain del NMOS, e del contatto di substrato del PMOS;
6. impiantazione di drogante P per la realizzazione delle regioni di source e drain del PMOS, e del contatto di substrato del NMOS;
7. deposizione dell'ossido su tutto il wafer e apertura dei contatti per le diverse regioni dei MOS e per i contatti di gate;
8. deposizione del film di metallo e delineazione delle interconnessioni e delle piste di metallizzazione (si è supposto un collegamento tra PMOS e NMOS corrispondente alla realizzazione dell'invertitore, come si vedrà nel seguito).

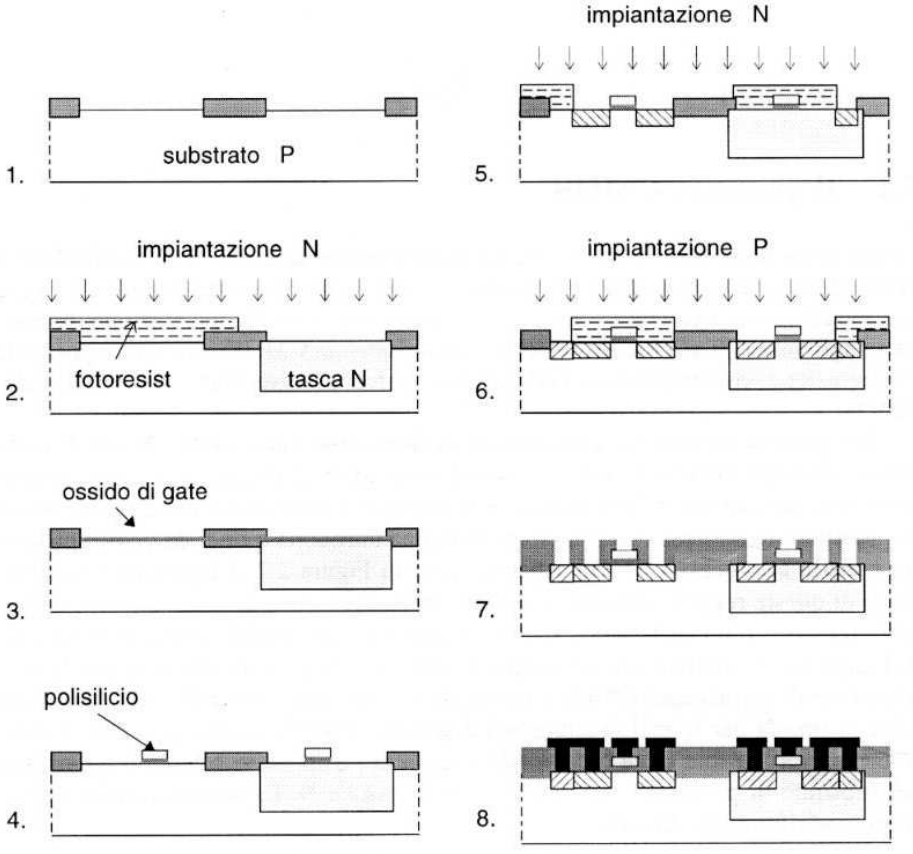


Figura 5.1 Processo di fabbricazione per strutture CMOS

Il processo può anche utilizzare tasche P, se si utilizzano substrati di tipo N, e anche entrambi i tipi di tasche (*twin-well*); quest'ultimo processo è più flessibile e permette un controllo migliore sia delle tensioni di soglia dei PMOS che degli NMOS, ma d'altra parte richiede un'area maggiore per l'invertitore.

Il primo ed immediato vantaggio della tecnologia CMOS è quello di creare substrati equivalenti (le tasche N o P), fisicamente separati dal substrato effettivo dalle giunzioni di isolamento tra tasca e substrato, per cui è possibile porre, per entrambi i transistori NMOS e PMOS, i contatti di substrato con quelli di source sia per il PMOS che per il NMOS; in questo modo l'effetto body, che degrada le prestazioni elettriche di un MOS, viene ad essere eliminato (vedremo che ciò è vero solo se tra alimentazione e massa vi è una sola coppia di CMOS; se più NMOS o PMOS sono in serie tra V_{DD} e massa vi sarà ancora un effetto body per qualcuno di questi). Il vantaggio tuttavia dell'uso di una tecnologia CMOS nelle porte logiche elementari è ben più rilevante e riveste diversi aspetti sia per quanto riguarda le prestazioni statiche che dinamiche; questi saranno considerati nell'analisi dell'invertitore CMOS sviluppata nel seguito.

5.2 L'invertitore CMOS

Lo schema dell'invertitore elementare CMOS, riportato in Figura 5.2, discende da quello dell'invertitore con carico attivo; in questo caso il dispositivo NMOS di carico di Figura 4.6 viene sostituito dal PMOS. In realtà la differenza tra i due schemi è più profonda, perché nel caso dell'invertitore CMOS *entrambi i dispositivi sono pilotati*, e quindi il carico (se così lo si può ancora chiamare) rappresentato dal PMOS è variabile non solo in dipendenza della tensione ai suoi capi (come per ogni carico attivo) ma anche e principalmente a causa della tensione al suo ingresso. Dallo schema di Figura 5.2 si vede che, mentre la tensione di gate V_{GSN} di QN è data dall'ingresso V_I, quella V_{GSP} di QP è data da $V_I - V_{DD}$ (ricordiamo che per un PMOS V_{GS} e V_{DS} hanno valori negativi); quindi un aumento dell'ingresso V_I porta contemporaneamente

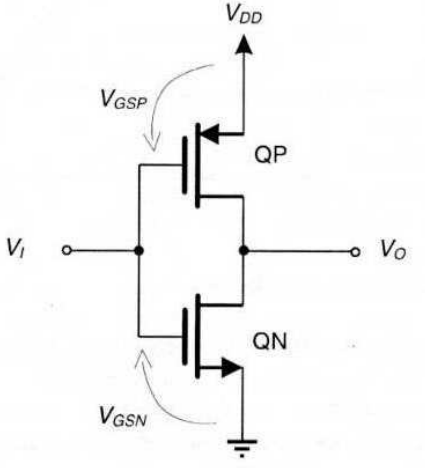


Figura 5.2 Schema di invertitore CMOS

in maggiore conduzione QN e in minore conduzione QP, con un effetto cumulativo per la variazione della tensione di uscita. I due casi limite corrispondono ad una tensione di ingresso $V_I > V_{DD} - |V_{TP}|$, per cui il transistore QP è interdetto e QN conduce, e ad una tensione di ingresso $V_I < V_{TN}$, per cui il transistore QP è in conduzione e QN è ora interdetto; in entrambi i casi la corrente I circolante nell'invertitore è nulla, e non vi è caduta di tensione sul transistore in conduzione.

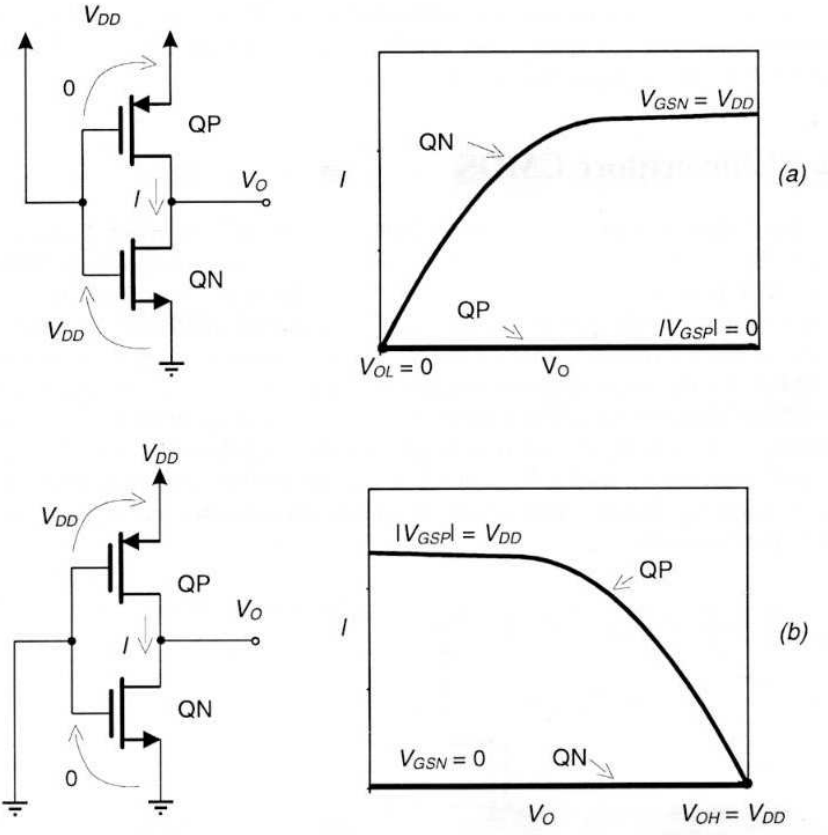


Figura 5.3 Analisi grafica per a) $V_I = V_{DD}$ e b) $V_I = 0$

In particolare, con riferimento ai due casi estremi di ingresso pari a V_{DD} o a 0 riportati in Figura 5.3, i rispettivi punti di funzionamento sono riportati in Tabella 5.1. I risultati dell'analisi grafica e i dati sui livelli logici riportati nella tabella mostrano che in questo tipo di tecnologia non occorre giocare sul valore del rapporto K_R tra i due MOS per ottenere un valore basso di V_{OL}, in quanto quest'ultimo è zero indipendentemente dal valore di K_2; la logica CMOS quindi

non è del tipo a rapporto (*logica ratioless*) come nel caso NMOS, e questo, come vedremo, porta a notevoli miglioramenti nelle prestazioni delle porte elementari.

Tabella 5.1 Punti di funzionamento dell'invertitore per i valori limite di ingresso

$V_I = V_{DD}$		$V_I = 0$	
$V_{GSN} = V_{DD}$	QN conduce	$V_{GSN} = 0$	QN interdetto
$V_{GSP} = 0$	QP interdetto	$\|V_{GSP}\| = V_{DD}$	QP conduce
$I_{DP} = I_{DN} = 0$	corrente nulla	$I_{DP} = I_{DN} = 0$	corrente nulla
$V_{OL} = 0$	uscita bassa	$V_{OH} = V_{DD}$	uscita alta

5.3 Caratteristica di trasferimento e margini di rumore

Per la caratteristica di trasferimento faremo quindi riferimento a un invertitore con uguali valori di K per il transistore NMOS e per quello PMOS. L'uguaglianza di K_N e K_P comporta per il PMOS un rapporto $W/L|_P = 2.5\ W/L|_N$. Infatti per entrambi i MOS il fattore K dato dalla (3.8) è proporzionale alla mobilità dei portatori di canale, che per il NMOS sono gli elettroni, mentre per il PMOS sono le lacune, per cui le rispettive mobilità sono legate approssimativamente dalla relazione: $\mu_N = 2.5\ \mu_P$. Inoltre assumeremo per i due MOS una tensione di soglia $V_{TN} = |\ V_{TP}\ | = V_T$; queste due condizioni rendono le caratteristiche dei due MOS simmetriche rispetto alla tensione di ingresso V_I e, come vedremo, ottimizzano la caratteristica di trasferimento dell'invertitore.

Si può utilizzare ancora la costruzione grafica riportata in Figura 5.4 per ricavare la caratteristica di trasferimento dell'invertitore. Il primo tratto della curva (A-B) (vedi Figura 5.4b) corrisponde a tensioni V_I inferiori alla soglia V_T; QN è quindi interdetto, QP conduce (ma in esso non passa corrente per l'interdizione di QN) e la tensione V_{OH} è quindi pari a V_{DD}. Nella regione B-C il transistore QN comincia a condurre, e l'uscita corrisponde ad intersezioni delle caratteristiche dei due MOS con QP nella regione di triodo e QN in pinch-off; in questa regione giace il punto V_{IL} perché nella regione C-D la pendenza è (approssimativamente) verticale. La regione C-D infatti corrisponde al caso in cui tutti e due i MOS lavorano in pinch-off (come è indicato in Figura 5.4a) e i due punti limite C e D corrispondono alle intersezioni rispettivamente con QP sulla curva limite di pinch-off o con QN sulla curva limite.

La tensione V_I che porta entrambe le curve a coincidere nella zona di pinch-off è $V_I = V_{DD}/2$, e corrisponde alla soglia logica dell'invertitore; questo valore è facilmente determinabile eguagliando le due correnti di QP e QN in regime di pinch-off. I punti C e D sono quelli corrispondenti alle intersezioni delle caratteri-

stiche sulle due curve di pinch-off rispettivamente di QP e QN, per il valore di $V_I = V_{DD}/2$:

$$\begin{aligned} C &\Rightarrow V_{DD} - V_O(QP) = V_{DD} - (V_I - V_T);\ V_O(C) = \frac{V_{DD}}{2} + V_T \\ D &\Rightarrow V_O(D) = V_I - V_T = \frac{V_{DD}}{2} - V_T \end{aligned} \tag{5.1}$$

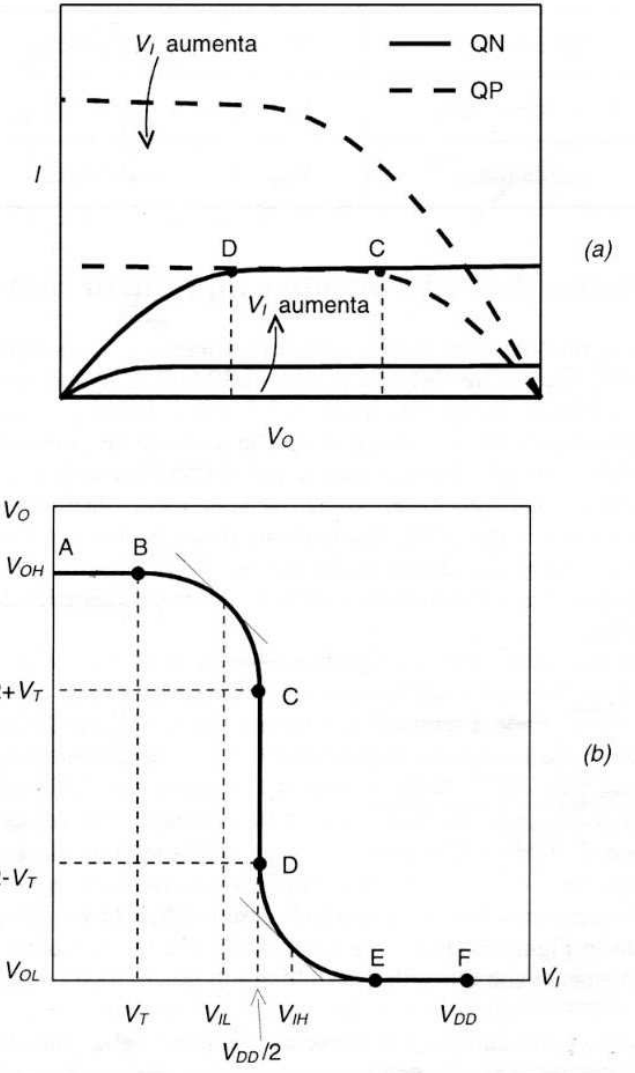


Figura 5.4 a) Costruzione grafica per un invertitore CMOS; b) caratteristica di trasferimento con $K_N = K_P$

La regione D-E è la corrispondente di quella B-C ma con QN nella regione di triodo e QP in pinch-off, quella E-F corrisponde a tensioni di ingresso V_I tali che $|V_I - V_{DD}| < V_T$, per cui QP è interdetto; il punto V_{IH} si troverà nella regione D-E analogamente a quanto detto per V_{IL}.

Per il calcolo dei valori caratteristici che determinano i margini di rumore, occorre calcolare solo il termine V_{IH}, perché V_{IL} può essere ottenuto da questo sfruttando la condizione di simmetria della caratteristica di trasferimento rispetto al valore $V_{DD}/2$, e gli altri termini sono, per quanto detto sopra: $V_{OH} = V_{DD}$; $V_{OL} = 0$.

Calcolo di V_{IH} (QN in regime triodo, QP in pinch-off)

Dall'uguaglianza delle correnti nei due dispositivi, nell'ipotesi $K_N = K_P$, $V_{TN} = |V_{TP}| = V_T$ si ha:

$$I_{DN} = I_{DP} \Rightarrow 2(V_I - V_T)V_O - V_O^2 = (V_{DD} - V_I - V_T)^2 \tag{5.2}$$

Derivando ambo i membri della (5.2) rispetto a V_I:

$$2(V_I - V_T)\frac{dV_O}{dV_I} + 2V_O - 2V_O\frac{dV_O}{dV_I} = -2(V_{DD} - V_I - V_T) \tag{5.3}$$

Ricordando che la condizione sul punto $V_I = V_{IH}$ è data da:

$$\frac{dV_O}{dV_I} = -1 \tag{5.4}$$

e sostituendo la (5.4) nella (5.3) si ha:

$$-2(V_{IH} - V_T) + 4V_O = -2(V_{DD} - V_{IH} - V_T) \tag{5.5}$$

da cui:

$$V_{IH} = V_O + \frac{V_{DD}}{2} \tag{5.6}$$

Sostituendo il valore di V_{IH} della (5.6) nella (5.2), si ottiene il valore di V_O:

$$V_O = \frac{1}{4}\left(\frac{V_{DD}}{2} - V_T\right) \tag{5.7}$$

che sostituito nella (5.6) fornisce:

$$V_{IH} = \frac{1}{4}\left(\frac{5V_{DD}}{2} - V_T\right) \tag{5.8}$$

Il valore di V_{IL} si ottiene direttamente con considerazioni di simmetria con V_{IH} rispetto alla soglia logica $V_{DD}/2$:

$$V_{IH} - \frac{V_{DD}}{2} = \frac{V_{DD}}{2} - V_{IL} \quad \Rightarrow \quad V_{IL} = \frac{1}{4}\left(\frac{3V_{DD}}{2} + V_T\right) \tag{5.9}$$

In definitiva i valori caratteristici per l'invertitore CMOS sono:

$$V_{OH} = V_{DD}$$
$$V_{OL} = 0$$
$$V_{IH} = \frac{1}{4}\left(\frac{5V_{DD}}{2} - V_T\right)$$
$$V_{IL} = \frac{1}{4}\left(\frac{3V_{DD}}{2} + V_T\right)$$

Dai risultati ottenuti per i valori V_{IH} e V_{IL} si vede che l'assunzione $K_N = K_P$ è perfettamente adeguata per l'invertitore CMOS. In effetti con questa ipotesi la caratteristica di trasferimento diventa simmetrica e i margini di rumore NM_H e NM_L sono uguali (situazione ottimale per un invertitore) :

$$NM_H = NM_L = \frac{3V_{DD}}{8} + \frac{V_T}{4} \tag{5.10}$$

Vedremo che questa condizione determina anche tempi di transizione uguali nella commutazione dell'invertitore, ed è quindi una scelta conveniente per il progetto del circuito.

5.4 Tracciato di un invertitore CMOS

Il tracciato di un invertitore CMOS deve tener conto: a) dello spazio richiesto dalle tasche (N e/o P) e della distanza minima delle aree di diffusione dai bordi della tasca, e b) della differenza delle aree di gate in quanto quella di gate del PMOS deve essere maggiore di quella del NMOS se si vuole avere $K_P = K_N$ (quest'ultima non è una condizione obbligatoria, si può accettare, come vedremo, di avere un peggioramento delle prestazioni elettriche ponendo le aree di gate uguali, perché ciò viene ripagato da una riduzione dell'area dell'invertitore). La prima condizione comporta un aumento dell'area totale dell'invertitore non indifferente, come si può vedere dall'esempio di tracciato riportato in Figura 5.5, relativo al caso di una tec-

nologia CMOS a tasca *N* (*N-well*), in quanto le regole di progetto richiedono che le regioni impiantate di drain e di source siano ad una distanza minima di 6λ dai bordi della tasca di isolamento, e che quest'ultima abbia una dimensione minima di 10λ.

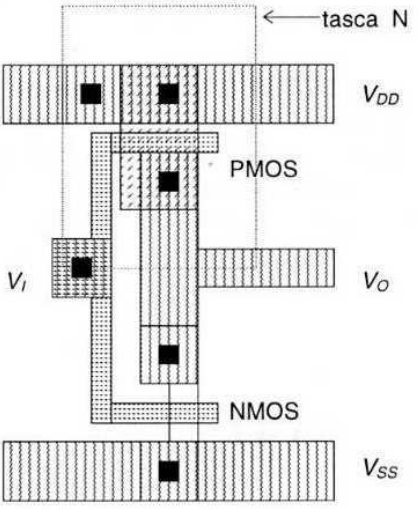


Figura 5.5 Tracciato di un invertitore CMOS a tasca *N*, con $W/L|_N = 3\lambda/2\lambda$, $W/L|_P = 8\lambda/2\lambda$

La seconda condizione implica che il rapporto *W/L* del PMOS sia circa 2.5 volte quello *W/L* del NMOS, e quindi, nell'ipotesi di area minima per ciascuno dei due dispositivi (L_N, $L_P = F$), anche l'area di gate $W{\cdot}L|_P$ sarà 2.5 volte quella $W{\cdot}L|_N$. Nel caso del tracciato della Figura 5.5, si è adottato un rapporto $W_P/W_N = 8/3 = 2.66$, in quanto il vincolo sulle dimensioni pari a multipli interi di λ non permette di assumere questo rapporto pari a 2.5.

Le due condizioni suddette fanno sì che l'area dell'invertitore CMOS sia maggiore di quella dell'invertitore NMOS a parità di valori di *K*, e questo spiega perché la massima densità di integrazione si raggiunge con tecnologia NMOS; tuttavia i miglioramenti nelle prestazioni sia statiche che (come vedremo) dinamiche dovuti alla tecnologia CMOS sono tali da far preferire questa ultima nella maggiore parte delle applicazioni, anche VLSI e ULSI.

5.5 Comportamento dinamico e tempi di propagazione

Anche per l'invertitore CMOS il comportamento dinamico dipende dalla capacità di carico C_T in uscita. Questa capacità è in effetti una capacità equivalente, come nel caso dell'invertitore NMOS, che tiene conto delle diverse capacità connesse al terminale di uscita. L'analisi approssimata che svilupperemo tiene conto essenzial-

mente la capacità di drain dei due MOS e quella di gate C_G dello stadio a valle, eventualmente moltiplicata per il numero N degli stadi in uscita (fan-out). Quest'ultima è più rilevante che nel caso NMOS, perché l'ingresso è connesso alle gate dei due dispositivi, e l'area della gate del PMOS è (per il caso $K_P = K_N$) 2.5 volte quella del NMOS; quindi la capacità totale di gate sarà:

$$C_{GT} = C_{G_N} + C_{G_P} = C_{OX}(W_N L_N + 2.5 \cdot W_N L_N) = 3.5 \cdot C_{G_N} \tag{5.11}$$

Nonostante questo aumento della capacità di carico, l'invertitore CMOS presenta una dinamica migliore di quella di un invertitore NMOS, ciò principalmente a causa del fatto che i due tempi di propagazione t_{PHL} e t_{PLH} sono uguali in base all'uguaglianza delle due correnti di scarica e di carica della capacità e, come si vedrà, relativamente più brevi che nel caso dell'invertitore NMOS.

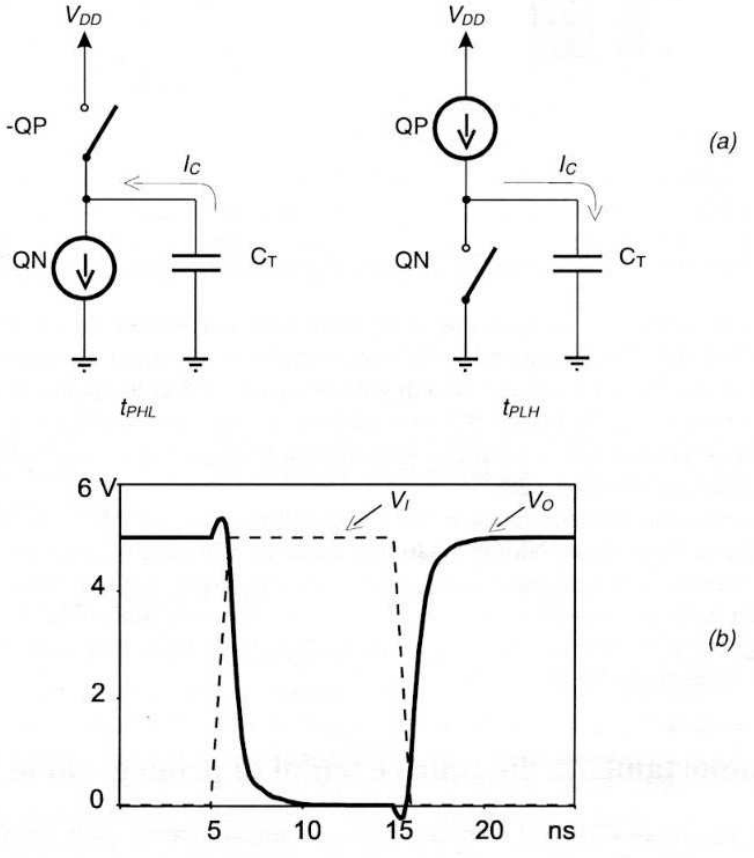


Figura 5.6 a) Circuiti semplificati per il calcolo dei tempi di propagazione; b) simulazione SPICE per l'uscita di un invertitore CMOS con $K_P = K_N$ e $W/L_N = 2$

Il circuito semplificato a cui fare riferimento per l'analisi dei tempi di propagazione è quello di Figura 5.6a; si può far ricorso anche in questo caso all'analisi grafica sviluppata per l'invertitore NMOS, per determinare la dinamica del punto di funzionamento dei due MOS (Figura 5.7), assumendo un segnale di ingresso idealizzato con tempi di salita e discesa nulli.

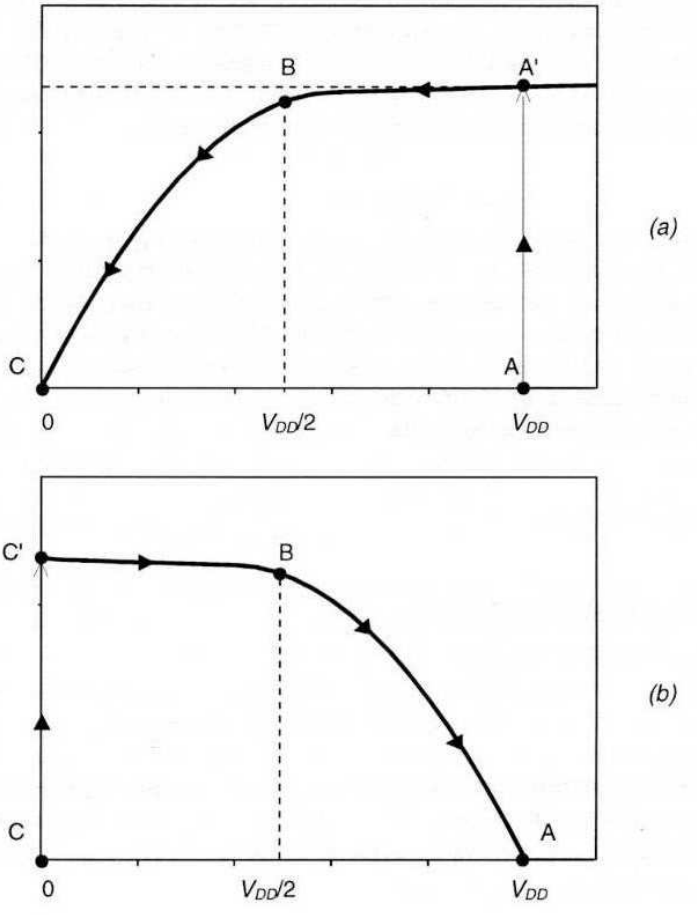


Figura 5.7 Analisi grafica dei transitori di commutazione in un invertitore CMOS: a) transitorio di scarica di C_T; b) transitorio di carica di C_T

Nel passaggio del segnale di ingresso dal livello basso a quello alto ($0 \rightarrow 1$) il PMOS si interdice e la capacità C_T si scarica attraverso il NMOS (Figura 5.7a); quest'ultimo può essere approssimato nel funzionamento in regione di pinch-off con un generatore a corrente costante, e con una resistenza nonlineare nella regione di triodo. Per la determinazione del tempo di propagazione occorre valutare il tempo di scarica dal punto A' al punto B, per il quale si è già nella regione di triodo, in

quanto il valore di soglia $V_{DD}/2$ è minore di $V_{DD} - V_T$; tuttavia si può approssimare in questo intervallo la corrente di scarica al valore massimo I_N pari a:

$$I_N = K_N(V_{DD} - V_{TN})^2 \tag{5.12}$$

Nel passaggio del segnale di ingresso dal livello alto a quello basso (1→ 0) la situazione si inverte e la capacità si carica attraverso il PMOS (Figura 5.7b). Anche in questo caso occorre valutare il tempo necessario a passare dal punto C' al punto B; in questo intervallo si può assumere, con la stessa approssimazione precedente, che la corrente di carica sia approssimativamente costante e pari a:

$$I_P = K_P(V_{DD} - |V_{TP}|)^2 \tag{5.13}$$

Nell'ipotesi $K_N = K_P = K$, $V_{TN} = |V_{TP}| = V_T$, i due transitori saranno uguali; in particolare la carica di C_T sarà resa più rapida dalla corrente relativamente elevata fornita dal PMOS, rispetto al caso dell'invertitore NMOS con carico attivo. Approssimando quindi, per il calcolo di t_{PHL} (t_{PLH}), la corrente di scarica (carica) approssimativamente uguale a quella massima del tratto di pinch-off data dalla (5.12) (o (5.13)), (che indicheremo con I_H), si ha:

$$t_{PLH} = t_{PHL} \cong \frac{C_T(V_{OH} - V_{OL})}{2I_H} = \frac{C_T V_{DD}}{2K(V_{DD} - V_T)^2} \tag{5.14}$$

e quindi il ritardo di propagazione sarà:

$$t_P = \frac{2t_{PHL}}{2} = \frac{C_T V_{DD}}{2K(V_{DD} - V_T)^2} \tag{5.15}$$

Quest'espressione mostra che il ritardo di propagazione è proporzionale al rapporto C_T/K; nel caso di un invertitore CMOS caricato da un ulteriore invertitore di uguale dimensionamento, poiché la capacità di carico $C_T \propto 3.5\ C_{GN}$, il ritardo di propagazione sarà dato da:

$$t_P = \frac{3.5C_{GN}V_{DD}}{2K(V_{DD} - V_T)^2} \tag{5.16}$$

È possibile valutare, in base a queste espressioni del tempo di propagazione, il comportamento dinamico di un invertitore CMOS in cui sia il PMOS che il NMOS hanno uguale area minima, ossia $W/L|_N = W/L|_P$. In questo caso i tempi di propagazione t_{PHL} e t_{PLH} non sono uguali, e il ritardo di propagazione t_P sarà pari al valor medio tra i due. Per il caso in esame:

$$t_{PHL} = \frac{2C_{GN}V_{DD}}{2K_N(V_{DD} - V_T)^2};\ t_{PLH} = \frac{2C_{GN}V_{DD}}{2K_P(V_{DD} - V_T)^2}$$

Poiché $K_P = K_N/2.5$, si può scrivere il ritardo di propagazione come:

$$t_P = \frac{t_{PHL} + t_{PLH}}{2} = \frac{C_{GN}V_{DD}}{K_N(V_{DD} - V_T)^2}\left(\frac{1+2.5}{2}\right) = \frac{3.5C_{GN}V_{DD}}{2K_N(V_{DD} - V_T)^2} \qquad (5.17)$$

Il ritardo di propagazione t_P in questo caso è uguale a quello dato dalla (5.16), pur essendo diversi i valori di t_{PHL} e t_{PLH}.

5.6 Potenza dissipata e prodotto ritardo-potenza

Come si è visto nell'analisi statica del Paragrafo 5.2, sia per ingresso alto (V_{DD}) che basso (0) la corrente nell'invertitore è nulla, per cui non vi è dissipazione *statica* di potenza in nessuno dei due stati possibili; questo è uno degli aspetti più interessanti della tecnologia CMOS. La potenza dissipata dalle porte corrisponde quindi solo a quella *dinamica*, cioè relativa alle *transizioni* da uno stato all'altro. Questa potenza viene a dipendere da due cause: a) dalla corrente che circola tra un MOS e l'altro nella fase di transizione in cui tutti e due i MOS sono momentaneamente in conduzione e b) dalla potenza spesa per la carica della capacità. Consideriamo separatamente i due contributi.

Per il caso a), facendo riferimento alla curva di trasferimento di Figura 5.7, nel passaggio della tensione di ingresso V_I da 0 a V_{DD} ($0 \rightarrow 1$) la corrente tra i due invertitori comincia a circolare per $V_I \geq V_T$ e raggiunge un massimo per $V_I = V_{DD}/2$; infine la corrente si annulla per $V_I \geq V_{DD} - V_T$. L'andamento della corrente nel piano I_D-V_I è facilmente ricavabile se si considera che la corrente è limitata dal MOS che conduce di meno, e che l'andamento è simmetrico rispetto a $V_I = V_{DD}/2$; per cui nel tratto da V_T a $V_{DD}/2$ la corrente dipende quadraticamente da V_I secondo la relazione:

$$I_D = K(V_I - V_T)^2 \qquad (5.18)$$

ed assume il valore massimo I_{MAX} per $V_I = V_{DD}/2$.

Per calcolare la potenza dissipata nella transizione si può approssimare un andamento lineare nel tempo della tensione di ingresso: $V_I = \alpha t$, nei tratti di salita e di discesa del segnale, per cui l'espressione della potenza media dissipata, calcolata sommando i due contributi uguali, relativi alle due transizioni Δt nel periodo T del segnale di comando, può essere scritta valutando l'energia totale assorbita nelle due transizioni e dividendola per il tempo T. Operando il cambio di variabile $dV_I = \alpha\, dt$, e moltiplicando per 2 il contributo relativo a una transizione l'integrale relativo può essere scritto come:

$$P_D' = \frac{2}{T}\int_0^{\Delta t} V_{DD} \cdot I_D(t)dt = \frac{4}{\alpha \cdot T}\int_{V_T}^{V_{DD}/2} V_{DD} \cdot I_D(V_I)dV_I \tag{5.19}$$

e, ricordando la (5.18), si ottiene:

$$P_D' = \frac{4}{3\alpha \cdot T} V_{DD} \cdot K\left(\frac{V_{DD}}{2} - V_T\right)^3 = \frac{4V_{DD} I_{MAX}}{3\alpha \cdot T}\left(\frac{V_{DD}}{2} - V_T\right) \tag{5.20}$$

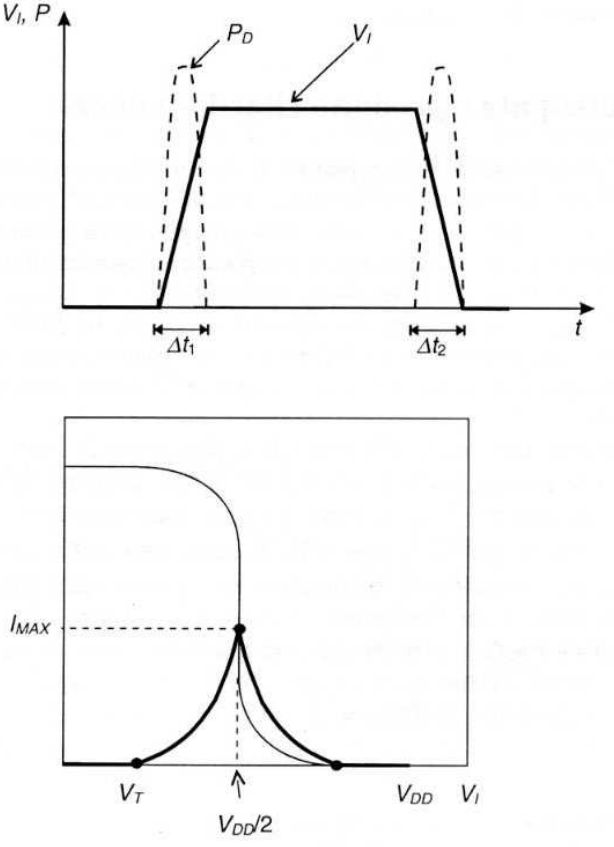


Figura 5.8 Potenza dissipata nei MOS e corrente di conduzione nei due MOS durante le transizioni di segnale

Il secondo contributo (caso b) è dovuto all'energia spesa per la carica della capacità di uscita C_T, e non dipende dalla realizzazione dell'invertitore. Nel Paragrafo 1.4 si è valutata questa dissipazione di potenza dinamica per l'invertitore idealizzato, che per il nostro caso si scrive:

$$P_D'' = f \cdot C_T \cdot V_{DD}^2 \tag{5.21}$$

dove f è il numero delle transizioni ($0 \rightarrow 1 + 1 \rightarrow 0$) nel secondo, cioè la frequenza (di clock) del segnale di ingresso.

Un paragone tra i due contributi può effettuarsi ricordando che, dalla definizione di α si può approssimare il tempo di transizione $t_T = V_{DD}/\alpha$ e che $f = 1/T$, per cui:

$$\frac{P_D'}{P_D''} = \frac{t_T \cdot K}{C_T} \cdot \frac{4}{3 \cdot V_{DD}^2} \left(\frac{V_{DD}}{2} - V_T \right) \tag{5.22}$$

Se si suppone che la capacità di carico sia quella di un uguale invertitore CMOS, e cioè $C_T = 3.5 C_{GN}$, si nota che il rapporto della (5.22) è abbastanza invariante rispetto alle scelte progettuali, in quanto, a parità di valori di tensione, nel primo termine si è evidenziato il termine K/C che è inversamente proporzionale a t_P, e quindi anche in buona approssimazione a t_T, per cui questo rapporto non può essere modificato variando i valori W/L dei MOS. Nei casi pratici il rapporto P'_D/P''_D è notevolmente minore di 1, il che permette di trascurare il primo termine rispetto al secondo; come esempio, per un invertitore CMOS con $k' = 10\ \mu A/V^2$, $W/L = 2$, $V_{DD} = 5$ V, il rapporto della (5.22) vale circa 0.04.

Il prodotto $P \cdot D$ per l'invertitore CMOS vale quindi, considerando il contributo del solo secondo termine per la potenza dinamica:

$$P \cdot D \equiv P_D'' \cdot t_P = \frac{f \cdot C_T^2 \cdot V_{DD}^3}{2K(V_{DD} - V_T)^2} \tag{5.23}$$

5.7 Confronto tra le prestazioni di invertitori NMOS e CMOS

Può essere utile a questo punto fare un confronto tra le caratteristiche più significative degli invertitori esaminati, basati sulla tecnologia MOS, basandosi sulle analisi approssimate svolte per i vari casi.

Margini di rumore

NMOS	basati su logica a rapporto; per l'invertitore NMOS E-E per avere un margine di rumore NM_L prossimo ad 1 V occorre un K_R più elevato (circa 15) che per il NMOS E-D (circa 4)
CMOS	basati su logica "ratioless"; i margini di rumore sono uguali se $K_N = K_P$, e con valori superiori a 2 V per $V_{DD} = 5$ V

Tempi di propagazione

Si possono confrontare le espressioni del tempo di propagazione t_{PLH} ricavate per i tre casi (4.39, 4.44a, 5.14) esprimendo tutte le relazioni in base al valore K_1 del MOS verso massa dell'invertitore:

NMOS E-E $$t_{PLH} = \frac{K_R C_T}{2K_1} \frac{1}{(V_{OH} - V_{OL})}$$

NMOS E-D $$t_{PLH} = \frac{K_R C_T}{2K_1} \frac{(V_{DD} - V_{OL})}{|V_{TD}|^2}$$

CMOS $$t_{PLH} = t_P = \frac{C_T}{2K_1} \frac{V_{DD}}{(V_{DD} - V_T)^2}$$

Dalle espressioni sopra elencate si vede che, a parte l'influenza del secondo termine correttivo legato ai termini di tensione, non molto diverso nei tre casi, il primo termine per l'invertitore NMOS è aumentato del fattore K_R. Si può quindi dedurre che l'invertitore CMOS permette di pilotare, a parità di tempo di propagazione t_P, capacità di carico K_R volte più elevate di quelle pilotabili dagli invertitori NMOS. Ricordando che il fan-out di un invertitore MOS è essenzialmente legato alla massima capacità vista dall'uscita come carico, questo risultato si trasferisce in un più elevato fan-out per l'invertitore CMOS.

Prodotto ritardo-potenza

NMOS E-E $$P \cdot D = \frac{C_T \cdot V_{DD}(V_{OH} - V_{OL})}{8}$$

NMOS E-D $$P \cdot D = \frac{C_T \cdot V_{DD}(V_{DD} - V_{OL})}{8} \cong \frac{C_T \cdot V_{DD}^2}{8}$$

CMOS $$P \cdot D = \frac{C_T V_{DD}^2 \cdot t_P}{T} \qquad \left(\frac{t_P}{T} \leq 0.1 \div 0.01\right)$$

Per l'invertitore CMOS il prodotto $P{\cdot}D$ è una funzione della frequenza di funzionamento. È tuttavia da sottolineare che l'aumento della frequenza di funzionamento va di pari passo con i miglioramenti tecnologici che portano ad una continua riduzione del "*feature size*"; quindi all'aumentare della frequenza (e cioè al diminuire del periodo T) corrisponde un analogo miglioramento (diminuzione) di t_P, che ha portato a conservare la competitività della tecnologia CMOS nel tempo.

Occupazione di area e dissipazione di potenza
Dalle Figure 4.13 e 5.5 che riportano i tracciati di due invertitori, rispettivamente di un NMOS (con $K_R = 9$) e di un CMOS, si può vedere come l'area utilizzata per il CMOS sia superiore a quella del NMOS, essenzialmente a causa dell'utilizzo di un'area in eccesso per la formazione della tasca N, e delle regole di progetto che prevedono una distanza minima tra le diffusioni e i bordi delle tasche e tra queste ultime.

Tuttavia la maggiore densità di integrazione per le porte NMOS non è utilizzabile se si vogliono raggiungere livelli di integrazione ULSI, cioè con densità dell'ordine di 10^6 componenti, a causa della (relativamente) elevata dissipazione di potenza per la singola porta (ricordiamo che nelle porte NMOS vi è una dissipazione di potenza statica, che per un NMOS E-D con $K_R = 4$ è circa 120 μW, come si ricava dalla (4.49). Quindi, assumendo una dissipazione di potenza massima di circa 10 W per un chip di 10×10 mm^2, pur potendo integrare (idealmente) $10^8/55 = 2 \cdot 10^6$ invertitori dal punto di vista dell'occupazione di area, si è limitati per problemi di dissipazione di potenza a integrare solo $10/(120 \cdot 10^{-6}) = 8 \cdot 10^4$ invertitori. Nel caso CMOS invece, essendo la dissipazione di potenza limitata a quella dinamica e molto più bassa, si può raggiungere un livello di integrazione determinato dall'occupazione di spazio disponibile, e cioè nel caso in esempio, $10^8/105 = 10^6$ invertitori.

5.8 Porte logiche elementari CMOS

Anche per la logica CMOS le porte elementari vengono realizzate a partire dall'invertitore; tuttavia in questo caso bisogna estendere gli algoritmi riportati nel Paragrafo 1.5 per la realizzazione di porte NAND o NOR con invertitori basati su dispositivi controllati e carico, in quanto nel caso degli invertitori CMOS sia il PMOS che il NMOS sono controllati dal segnale di ingresso. Nel Paragrafo 1.5 si è visto che i dispositivi NMOS sono assimilabili ad interruttori che si aprono se in ingresso vi è un livello logico basso, e si chiudono se l'ingresso è a livello logico alto; in base a questa rappresentazione si è estratta la regola di porre in parallelo i dispositivi controllati per realizzare la funzione NOR e in serie per quella NAND.

Questa regola può essere ancora applicata per i dispositivi NMOS dell'invertitore, assumendo che i PMOS agiscano come carico, ma occorre dualmente verificare che la funzione logica voluta venga anche realizzata considerando i PMOS come interruttori e gli NMOS come carico. Per estendere l'analogia di interruttori pilotati ai dispositivi PMOS, si deve considerare che questi sono *interdetti* quando l'ingresso è alto (mentre gli NMOS sono *in conduzione* quando l'ingresso è alto), per cui i dispositivi PMOS sono assimilabili a interruttori pilotati da variabili invertite, o negate, e con il carico connesso verso massa (i dispostivi NMOS), il che elimina la negazione in uscita della funzione logica realizzata, al contrario di ciò che accade nel caso di interruttori inseriti tra carico e massa.

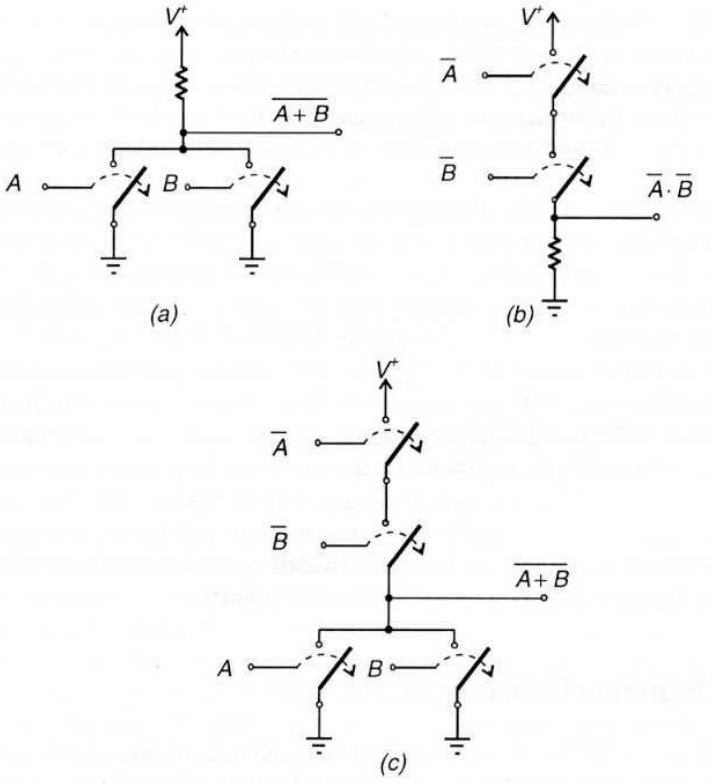


Figura 5.9 Realizzazione di una funzione NOR a partire a) da interruttori pilotati da variabili dirette, b) da variabili negate, c) da entrambi i tipi di interruttori

In Figura 5.9 è indicata la combinazione dei circuiti ad interruttori che realizzano una funzione NOR. Il primo caso è quello visto per gli invertitori NMOS (Figura 5.9a). Nel secondo caso (Figura 5.9b) si ipotizzano interruttori pilotati da variabili negate, per cui la funzione realizzata è una funzione AND tra le variabili negate (e non una NAND), in quanto gli interruttori sono connessi tra l'alimentazione e il carico, e non tra carico e massa. Ricordando il teorema T10 di Tabella 1.2 si vede che la combinazione dei due circuiti (escludendo per ognuno di questi la resistenza di carico) costituisce ancora una rete di interruttori pilotati che realizza la funzione NOR. In generale, ricordando che le due funzioni NOR e NAND possono essere espresse in due modi diversi, sfruttando i teoremi di De Morgan, basandosi su variabili non negate e realizzando la negazione all'esterno, oppure basandosi su variabili negate, come è indicato in Tabella 5.2, si deve usare la regola che scaturisce dalla seconda colonna per la connessione dei dispositivi PMOS, e cioè una connessione *serie* per la funzione NOR e *parallelo* per quella NAND.

Tabella 5.2 Funzioni NOR e NAND con variabili dirette e negate

Funzione	*negazione all'esterno (dispositivi NMOS)*	*variabili negate (dispositivi PMOS)*
NOR	$\overline{A+B}$	$\overline{A}\cdot\overline{B}$
NAND	$\overline{A\cdot B}$	$\overline{A}+\overline{B}$

In definitiva le connessioni dei due invertitori elementari necessarie per realizzare una porta NOR o NAND a 2 ingressi sono quelle riportate in Figura 5.10, dove si vede che sono applicate le due regole di Tabella 5.2 rispettivamente per i dispositivi NMOS e PMOS.

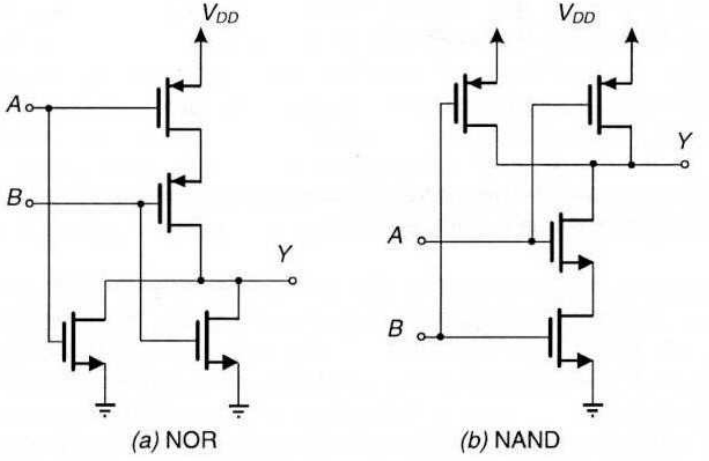


Figura 5.10 a) Porta NOR CMOS a due ingressi; b) porta NAND a due ingressi

Le porte CMOS (come anche quelle NMOS) vanno dimensionate facendo riferimento a un "invertitore equivalente" a cui possono essere riportate le porte per ogni data combinazione delle variabili logiche in ingresso. Ricordando quanto detto nel Paragrafo 4.11 per la valutazione del K equivalente di una connessione di transistori MOS in serie o in parallelo (estendendo il concetto di lunghezza equivalente di canale data dalla (4.54) anche per tensioni di drain non trascurabili) si ha che una connessione di N MOS uguali in parallelo è equivalente a un singolo MOS con un $K_{eq} = NK$, mentre una connessione di N MOS uguali in serie è equivalente ad un unico MOS con un $K_{eq} = K/N$, indicando in entrambi i casi con K il valore del fattore di scala del singolo MOS. Ne consegue che, in qualunque condizione di funzionamento il comportamento sia statico che dinamico della porta CMOS in esame è riconducibile a quello di un "invertitore CMOS equivalente" con

opportuni valori di K_{Peq} e K_{Neq}. Ad esempio in Figura 5.11 è riportato il dimensionamento dell'"invertitore equivalente" di una porta NOR a 3 ingressi, per il caso: A = 1→ 0, B = 0, C = 0.

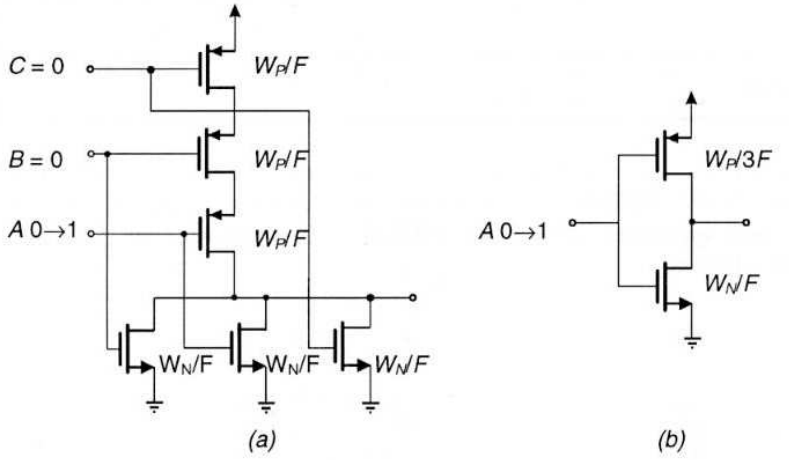


Figura 5.11 a) Dimensionamento di una porta NOR a tre ingressi; b) invertitore equivalente per la combinazione di variabili assegnata

Si può comprendere da questo esempio che, per la maggior parte delle combinazioni di segnali, le porte a più ingressi non sono riconducibili ad un invertitore con $K_{Peq} = K_{Neq}$; ciò comporta una dissimmetria nella caratteristica di trasferimento e quindi nei margini di rumore per le diverse combinazioni di ingressi, e comporta anche tempi di propagazione (e tempi di transizione) diversi nella carica e scarica della capacità di carico. In questi casi si fa usualmente riferimento per il dimensionamento della porta alla condizione peggiore (*worst case design*), in modo da garantirsi un comportamento migliore di quello valutato, per tutte le altre situazioni, rispetto a quella peggiore.

Anche per la versione CMOS si può valutare l'occupazione totale di area per le due porte logiche supposte a N ingressi. L'analisi va fatta in questo caso sulla base dell'uguaglianza dei tempi di propagazione t_{PHL} e t_{PLH}, in quanto i margini di rumore sono sufficientemente elevati anche al variare delle aree di gate dei dispositivi.

L'analisi va fatta sempre nel caso peggiore, che è quello della conduzione di uno solo dei MOS in parallelo, perché dà luogo al transitorio più lento sulla capacità, mentre il transitorio sui MOS in serie prevede in ogni caso la conduzione di tutti i MOS. Per la porta NOR questo corrisponde alla conduzione di un solo NMOS (un solo ingresso alto) per t_{PLH}, perché in tal caso la corrente di scarica di C_T (attraverso gli NMOS) è la minima, essendo quella di carica limitata dalla serie di N PMOS. In questo caso si possono considerare, nella carica della capacità di carico C_T, gli N PMOS in serie come un unico MOS con una $L_{eq} = NL_P$, mentre

nella scarica di C_T nel caso peggiore compare un unico NMOS. Si ha quindi per l'uguaglianza dei due tempi di propagazione t_{PLH} e t_{PLH} nella condizione peggiore:

$$t_{PLH} = t_{PHL} \Rightarrow K_{Peq} = K_{Neq} \Rightarrow \frac{W_P}{NL_P} = 2.5\frac{W_N}{L_N} \tag{5.24}$$

Nel caso di una porta NAND a N ingressi il caso peggiore è quello del passaggio di uno solo degli ingressi da alto a basso, rimanendo tutti gli altri ingressi alti. In questo caso l'eguaglianza dei tempi di propagazione t_{PHL} e t_{PLH} porta a:

$$K_{Neq} = K_{Peq} \Rightarrow 2.5\frac{W_N}{NL_N} = \frac{W_P}{L_P} \tag{5.25}$$

In questo caso saranno gli NMOS a dover presentare le aree maggiori, in quanto nel caso di più di due ingressi la corrente portata dal singolo PMOS è superiore a quella portata dalla serie degli NMOS a pari area. Se si assume di progettare ad area minima, cioè con $L_N = L_P = F$, e si impone l'ulteriore condizione che i tempi di propagazione nel caso peggiore siano uguali per le porte NAND e NOR a parità di numero di ingressi, ossia che le porte abbiano uguali prestazioni dinamiche, si ricava, eguagliando i secondi membri delle (5.24) e (5.25), la nota relazione: $W_P = 2.5W_N$, e dalle (5.24) e (5.25) si ottengono le seguenti relazioni per le aree (di gate) delle due porte, in funzione dell'area A_{MIN} del singolo NMOS della porta NOR:

$$\begin{aligned} \text{Area minima NOR} &= NW_N F + 2.5N^2 W_N F = NA_{MIN}(1 + 2.5N) \\ \text{Area minima NAND} &= NW_P F + \frac{N^2}{2.5} W_P F = NA_{MIN}(2.5 + N) \end{aligned} \tag{5.26}$$

dove si è assunto: $W_P F = W_N F \cdot 2.5 = A_{MIN} \cdot 2.5$.

Da queste relazioni consegue che, con l'aumentare del numero di ingressi, l'ingombro in area della porta NOR diventa maggiore di quello di una porta NAND con lo stesso numero di ingressi, per cui nella tecnologia CMOS la logica sviluppata è basata su porte NAND.

Spesso il progetto delle porte elementari CMOS prevede l'utilizzo di dispositivi elementari di dimensioni assegnate, come ad esempio NMOS con rapporto $W/L|_N = 3\lambda/2\lambda$ e PMOS con rapporto $W/L|_P = 2.5\ W/L|_N$; in tal caso le porte logiche realizzate con questi dispositivi presenteranno (in linea di principio) uguale occupazione di area, in quanto per entrambe le porte si utilizzano N NMOS e N PMOS di area determinata, mentre non risulteranno uguali i tempi di propagazione t_{PHL} e t_{PLH}. Il tempo di propagazione maggiore tra i due sarà dato dall'attivazione del ramo in cui i dispositivi sono in serie, siano essi NMOS o PMOS. In questo caso, assumendo tutti gli NMOS con rapporto $W/L|_N$ minimo, e tutti i PMOS con $W/L = 2.5\ W/L|_N$, si

avrà che i due tempi di propagazione peggiori (t_{PHL} per la porta NAND e t_{PLH} per quella NOR) saranno uguali, in quanto:

$$t_{PHL\,NOR} \propto \frac{C_T}{K_{Neq}} = \frac{NFC_T}{W_N}; t_{PLH\,NAND} \propto \frac{C_T}{K_{Peq}} = \frac{2.5NFC_T}{W_P} = \frac{2.5NFC_T}{2.5W_N} \quad (5.27)$$

5.9 Fan-in e fan-out delle porte CMOS

Si è definito in generale come fan-in di una porta logica il numero N di ingressi che la porta presenta (e quindi il numero di variabili logiche che possono presentarsi all'ingresso per effettuare la funzione logica assegnata). Dall'analisi della dinamica delle porte CMOS effettuata nel Paragrafo 5.8 si è visto come il numero di ingressi condizioni il numero di transistori in serie (PMOS per le porte NOR e NMOS per quelle NAND), e quindi in ultima analisi degradi i tempi di propagazione delle porte stesse. Il fan-in di una porta CMOS sarà quindi essenzialmente determinato dalla massima degradazione del tempo di propagazione dovuta all'aumentare degli ingressi logici della porta stessa. La degradazione del tempo di propagazione è immediatamente determinabile, in prima approssimazione, se si considera un progetto delle porte basato su transistori di dimensioni W/L assegnate, come si è visto nell'analisi precedente. Se si assume come carico di riferimento per le porte quello minimo di un invertitore CMOS con area minima ($W/L|_N = 3\lambda/2\lambda$; $W/L|_P = 2.5\ W/L|_N$), detto $C_L = 3.5C_{OX}W_NF$, si ha, in base alla (5.27), per la porta NOR:

$$t_{PLH} = N\frac{FC_LV_{DD}}{2k_P'W_P(V_{DD}-V_T)^2} = N \cdot t_{PO} \quad (5.28)$$

dove si è indicato con t_{PO} il tempo di propagazione di un invertitore ad area minima, dato dalla (5.17), caricato da un analogo invertitore. Per la porta NAND si ha analogamente:

$$t_{PHL} = N\frac{FC_LV_{DD}}{2k_N'W_N(V_{DD}-V_T)^2} = N \cdot t_{PO} \quad (5.29)$$

Da queste espressioni si vede che il tempo di propagazione cresce linearmente con il numero di ingressi, e ciò pone un limite superiore al valore di N. Va sottolineato che in questa analisi non si è considerato l'effetto body presentato rispettivamente dai dispositivi NMOS in serie a quello con il source connesso a massa (e quindi collegato al contatto di substrato), e dai dispositivi PMOS in serie a quello con il source connesso all'alimentazione (e quindi collegato al contatto della tasca N); questo effetto riduce ulteriormente la corrente che può circolare nella serie di N dispositivi MOS e quindi degrada ulteriormente il tempo di propagazione legato alla conduzione del ramo serie. In pratica non si supera un fan-in di 5÷6 sia per le porte NOR che NAND; se è necessario un numero maggiore di ingressi,

può risultare più conveniente realizzare la funzione logica in più passi, piuttosto che ricorrere ad una sola porta logica a molti ingressi, per migliorare le prestazioni dinamiche. Ad esempio in Figura 5.12 si riportano due versioni di porta NAND a 8 ingressi, realizzate con singola porta o con una combinazione di porte a 4 ingressi.

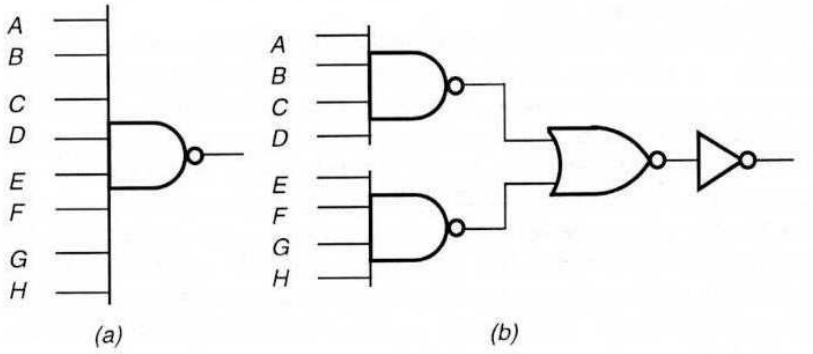


Figura 5.12 Realizzazione di una funzione NAND a 8 ingressi: a) con singola porta; b) con porte NAND, NOR e INVERT

Il fan-out delle porte CMOS è ancora determinato dalla massima degradazione ammissibile del tempo di propagazione, che aumenta in dipendenza del carico capacitivo visto dalla porta in esame, e quindi aumenta linearmente con il numero di porte collegate all'uscita. Ricordando che il carico capacitivo che ogni ingresso delle porte a valle presenta alla porta a monte è circa 3.5 C_{GN}, l'aumento del ritardo di propagazione con N porte in uscita è dato da:

$$t_P = \frac{N 3.5 C_{GN} V_{DD}}{2K(V_{DD} - V_T)^2} = N \cdot t_{PO} \tag{5.30}$$

dove in questo caso si è definito con t_{PO} il ritardo di propagazione per un carico di una sola porta in uscita.

Da questa espressione si vede come le porte CMOS, analogamente a quelle NMOS, non permettano un fan-out elevato, in quanto già con un carico di una decina di porte il ritardo di propagazione aumenta di un ordine di grandezza rispetto alle prestazioni ottenibili dalle porte a carico contenuto. Per pilotare un numero elevato di porte in uscita, e più in generale per pilotare carichi capacitivi elevati rispetto a quelli corrispondenti all'ingresso di una sola porta a valle con aumenti contenuti del ritardo di propagazione, è necessario inserire tra l'uscita e il carico un adeguato numero di stadi di separazione, che permettono un adattamento del carico all'uscita della porta in esame. Studieremo il dimensionamento degli stadi separatori di uscita (detti anche stadi *buffer*) nel paragrafo successivo.

5.10 Stadi separatori di uscita

Per le porte logiche CMOS occorre prevedere la necessità di pilotare capacità molto più grandi di quella di una porta logica; questo è ad esempio il caso dell'uscita di un segnale di clock che deve controllare molte porte, o ancora il caso di uscite che debbano alimentare circuiti fuori del chip, passando quindi per una piastra stampata (il supporto di assemblaggio dei diversi componenti integrati), il che incrementa notevolmente la capacità di carico.

In questi casi occorre prevedere degli stadi di uscita, che permettono di adattare gradualmente i circuiti logici alla elevata capacità di carico, senza degradare eccessivamente le prestazioni dinamiche delle porte. Questi sono ancora degli stadi invertitori, capaci di fornire correnti più elevate di quella disponibile dalla porta elementare, attraverso un aumento del fattore K dei transistori dell'invertitore. Non è conveniente utilizzare un solo stadio di uscita perché, posto che la capacità di carico C_L sia M volte più grande della capacità C_T pilotabile dalla porta tipica del circuito, occorrerebbe un aumento di K dello stesso valore M per ottenere un uguale tempo di propagazione su C_T, ma questo richiede un aumento di area di gate dello stadio di uscita ancora pari a M; in questo caso la porta da adattare vedrebbe ancora una capacità $M \cdot C_T$ in uscita, vanificando l'effetto dello stadio di separazione.

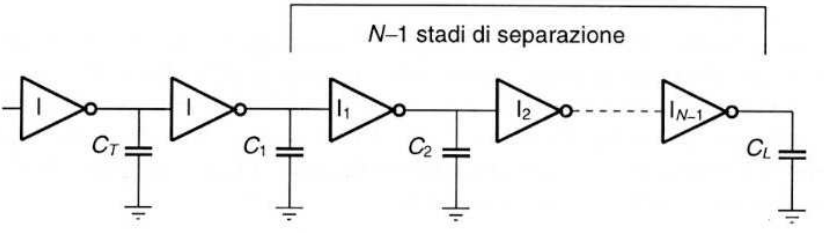


Figura 5.13 Stadi di separazione per alimentare una capacità C_L elevata

Facciamo riferimento per semplicità ad una porta CMOS schematizzata come invertitore generico I e consideriamo una cascata di $N-1$ stadi invertitori di separazione, come indicato in Figura 5.13, necessari per adattare l'uscita della porta I alla capacità di carico C_L. Assumiamo di ripartire in maniera uguale fra gli stadi l'incremento dei valori dei K_J degli $N-1$ stadi necessari per passare dalla capacità C_T (in uscita al generico invertitore I) a quella C_L, e definiamo $G = K_{J+1}/K_J$ l'incremento tra i valori K degli $N-1$ stadi. Le capacità di ingresso di questi stadi saranno a loro volta incrementate dello stesso fattore $G = C_{J+1}/C_J$ perché $C_J \propto K_J$, per cui il rapporto tra C_L e C_T sarà dato da:

$$\frac{C_L}{C_T} = \frac{C_L}{C_{N-1}} \cdots \frac{C_j}{C_{j-1}} \cdots \frac{C_1}{C_T} = G^N \Rightarrow N = \frac{1}{ln\, G} ln\left(\frac{C_L}{C_T}\right) \tag{5.31}$$

Occorre quindi trovare i valori di N e G che minimizzano l'aumento del tempo di propagazione attraverso gli N–1 stadi. Ricordando la (5.15) per il tempo di propagazione dell'invertitore CMOS (ma le considerazioni svolte valgono anche per invertitori NMOS), da cui discende che t_P è proporzionale al rapporto C_T/K, si ha per i tempi di propagazione dei singoli stadi (comprendendo anche l'ultimo stadio I che è caricato da una capacità diversa da C_T):

$$t_{P1} \propto \frac{C_1}{K} = G\frac{C_T}{K} = G \cdot t_P$$

$$t_{P2} \propto \frac{C_2}{K_1} = \frac{G^2 C_T}{GK} = G \cdot t_P = t_{P3} = \ldots = t_{PN-1} \tag{5.32}$$

$$t_{PN} \propto \frac{C_L}{K_{N-1}} = \frac{G^N C_T}{G^{N-1} K} = G \cdot t_P$$

Dalle (5.32) si ottiene per il tempo di propagazione totale degli N stadi:

$$t_{PTOT} = \sum_1^N t_{Pj} = N \cdot G \cdot t_P \tag{5.33}$$

Sostituendo l'espressione di N ricavata dalla (5.31) nella (5.33) si ha:

$$t_{PTOT} = \frac{1}{\ln G} \cdot G \cdot \ln\left(\frac{C_L}{C_T}\right) \cdot t_P \tag{5.34}$$

e minimizzando t_{PTOT} rispetto a G si ha:

$$\frac{dt_{PTOT}}{dG} = 0 \quad \Rightarrow \quad \ln G = 1 \quad \Rightarrow \quad G = e \cong 2.7 \tag{5.35}$$

Si ottiene quindi un valore di 2.7 per l'incremento G di K tra i diversi stadi; utilizzando questo valore nella (5.31) si ottiene N come:

$$N = \ln\frac{C_L}{C_T} \tag{5.36}$$

Con gli stadi di separazione si avrà quindi un ritardo di propagazione totale dato dalla (5.34), mentre se si fosse collegata direttamente la capacità C_L alla porta I il ritardo di propagazione sarebbe risultato:

$$t'_{PTOT} = \frac{C_L}{K} = G^N \cdot \frac{C_T}{K} = G^N \cdot t_P \tag{5.37}$$

Un esempio ci permetterà di valutare il vantaggio sul ritardo di propagazione come espresso dalla (5.33) invece che dalla (5.37). Supponiamo di dovere alimentare una capacità C_L =10 pF in uscita da una catena di invertitori che presentano una capacità di ingresso C_T =188 fF. Con una catena di 4 stadi di separazione (in quanto dalla (5.37) $\ln(C_L/C_T) = 4$), con $G = 2.7$ si ha per t_{PTOT}:

$$t_{PTOT} = 4 \cdot 2.7 \cdot t_P = 11 \cdot t_P$$

mentre, connettendo C_L direttamente all'uscita dell'invertitore non adattato si ha:

$$t'_{PTOT} = \frac{C_l}{C_T}\frac{C_T}{K} = 53 \cdot t_P$$

con significativo peggioramento del ritardo di propagazione.

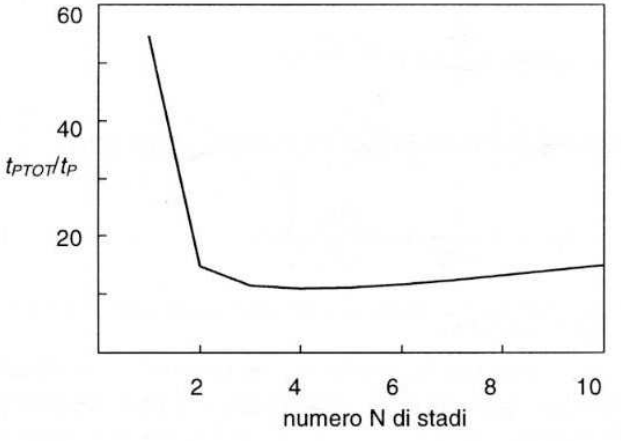


Figura 5.14 Dipendenza del ritardo di propagazione totale dal numero di stadi buffer, per il caso C_T = 188 fF, C_L = 10 pF

La riduzione del tempo di propagazione non dipende criticamente dal numero di stadi buffer utilizzati, anche se il minimo del rapporto t_{PTOT}/t_P si ha per il valore $G = e$, e quindi per un numero N definito dalla (5.36), come si può vedere dal diagramma di Figura 5.10 per il caso riportato in esempio; si vede infatti che l'incremento del ritardo di propagazione per un numero di stadi maggiore di quello ottimale di 4 è molto contenuto rispetto al valore del minimo ideale.

5.11 Tracciati delle porte NAND e NOR

In Figura 5.15 è riportato un possibile tracciato di una porta NAND CMOS a 2 ingressi, realizzata a partire da transistori con dimensioni fissate e cioè NMOS con rapporto $W/L = 6\lambda/2\lambda$, e PMOS con rapporto $W/L = 15\lambda/2\lambda$. Si può notare, confrontando questo tracciato con quello dell'analoga porta in tecnologia NMOS riportata in Figura 4.25, che in questo caso l'occupazione di area è maggiore, principalmente a causa della presenza della tasca N e della necessaria distanza che deve essere rispettata tra questa e le diffusioni N dei transistori NMOS.

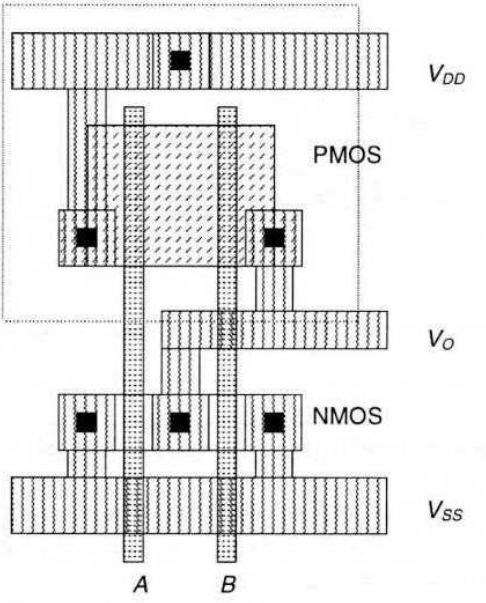


Figura 5.15 Tracciato di una porta NAND CMOS a 2 ingressi con $W/L|_N = 6\lambda/2\lambda$, $W/L|_P = 15\lambda/2\lambda$

Per completare il confronto si è riportato in Figura 5.16 il tracciato di una porta NOR CMOS a due ingressi, anche questa realizzata con transistori di dimensioni fissate, e precisamente con un rapporto $W/L|_P = 2.5\ W/L|_N$. Il confronto dei tracciati delle due porte mostra come le aree totali occupate non siano molto diverse tra loro, come si desume dall'analisi sviluppata per questo caso nel Paragrafo 5.8. Con questo dimensionamento dei MOS le porte presenteranno tempi di propagazione t_{PHL} e t_{PHL} diversi tra loro, e in entrambi i casi il tempo di propagazione maggiore è quello che coinvolge il ramo in cui i MOS sono in serie (PMOS per la porta NAND e NMOS per la porta NOR). Il dimensionamento del ramo serie per avere uguali tempi di propagazione richiede un aumento delle dimensioni di questi ultimi, e quindi un aumento dell'area totale delle porte stesse.

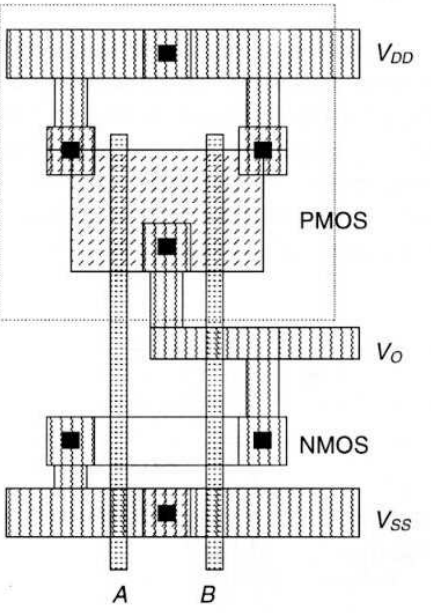


Figura 5.16 Tracciato di una porta NOR CMOS a 2 ingressi con $W/L|_N = 6\lambda/2\lambda$, $W/L|_P = 15\lambda/2\lambda$

5.12 Riduzione di scala dei circuiti CMOS

La tecnologia MOS, ed in particolare quella CMOS, permette di sfruttare al meglio i continui miglioramenti introdotti nei processi tecnologici, specialmente per quanto riguarda la minima dimensione (*feature size*) del processo fotolitografico, che, come si vede dal grafico di Figura 2.7 è diminuita costantemente negli ultimi due decenni. Come conseguenza di questa diminuzione delle dimensioni delle strutture integrabili si è avuto un continuo miglioramento delle prestazioni dinamiche dei circuiti stessi.

Il miglioramento è immediatamente comprensibile se si considera che, riducendo il valore della minima dimensione F, si riducono sia W che L (e le aree associate di drain e source) riducendo quindi la capacità equivalente C_T, mentre può rimanere costante il rapporto W/L e quindi la corrente di drain, per cui il tempo di propagazione t_P, dato dalla (5.15), diminuisce. Si può avere un effetto più rilevante dalla riduzione di scala se si ipotizza di scalare anche grandezze non legate alla fotolitografia, come ad esempio lo spessore dell'ossido di gate t_{OX}, i drogaggi e le tensioni di alimentazione. Questo porta alle regole di scala per i circuiti MOS che hanno determinato la forza di spinta della riduzione continua delle dimensioni dei dispositivi MOS negli ultimi due decenni. Le regole di scala si ricavano direttamente dalle relazioni riportate per il ritardo di propagazione t_P, della potenza dissipata P_D

e del prodotto DP, ricordando le relazioni tra i parametri fisici e geometrici scalabili e le grandezze K, C_G:

$$C_G = W \cdot L \cdot \frac{\varepsilon_{OX}}{t_{OX}}; \qquad K = \frac{1}{2}\mu \frac{\varepsilon_{OX}}{t_{OX}} \frac{W}{L}$$

$$t_P = \frac{C_G V_{DD}}{2K(V_{DD} - V_T)^2} \qquad (C_T \equiv C_G)$$

$$P_D = f \cdot C_G \cdot V_{DD}^2$$

$$P \cdot D = \frac{t_P}{T} C_G V_{DD}^2$$

Assumendo quindi di poter scalare sia le grandezze geometriche W, L, t_{OX} che le tensioni V_{DD}, V_T del fattore $1/x$, C_G si riduce di $(1/x)^2$ $x = 1/x$, K aumenta del fattore x. Lo scalamento di questi termini provoca uno scalamento delle grandezze elettriche riportate nelle relazioni precedenti.

Tabella 5.3 Regole di scala per circuiti CMOS

grandezza	*scalamento completo*
W, L, t_{OX}	$1/x$
V_{DD}, V_T, N_{SI}	$1/x$
C_G	$1/x$
K	x
ritardo di propagazione t_P	$1/x$
potenza dissipata P_D	$(1/x)^2$
prodotto ritardo-potenza $P \cdot D$	$(1/x)^3$

Si riportano in Tabella 5.3 le regole di scala per uno scalamento completo di tutte le grandezze. Nella valutazione del prodotto $P \cdot D$ si è considerato che il rapporto t_P/T rimanga costante in quanto una riduzione del ritardo di propagazione permette un aumento della frequenza di operazione a parità di prestazioni; per la stessa ragione si considera nell'espressione di P_D che la frequenza aumenti secondo il fattore x.

L'ipotesi di scalare anche le tensioni non è tuttavia così praticabile come quella di scalare le dimensioni geometriche. Per la tensione di alimentazione sorgono dei vincoli di compatibilità tra i diversi sottosistemi per cui le tensioni di alimentazione sono standardizzate (si passa dagli attuali 5 V a 3.3V e in previsione a 1.5 V); per

quanto riguarda la tensione di soglia, questa è la grandezza meno facilmente modificabile perché pur dipendendo dallo spessore dell'ossido non è semplicemente scalabile con questo a causa della dipendenza dagli stati di interfaccia alla superficie del silicio e dalle cariche nell'ossido.

Anche la riduzione delle dimensioni geometriche dei dispositivi non può essere portata avanti indefinitamente, sia pure supponendo di poter disporre di processi microelettronici che possano operare con dimensioni delle centinaia di nanometri (1 nm = 0.001 μm). La riduzione della lunghezza del canale L del MOS a valori inferiori al micron introduce una serie di effetti del secondo ordine nelle caratteristiche del transistore definiti come effetti di canale corto (*short channel effects*), che riducono sia la corrente di drain che la resistenza differenziale nel regime di pinch-off; al di sotto di 0.25 μm il trasporto dei portatori nel canale avviene con leggi differenti da quelle valide per canali lunghi e le analisi svolte non sono più valide.

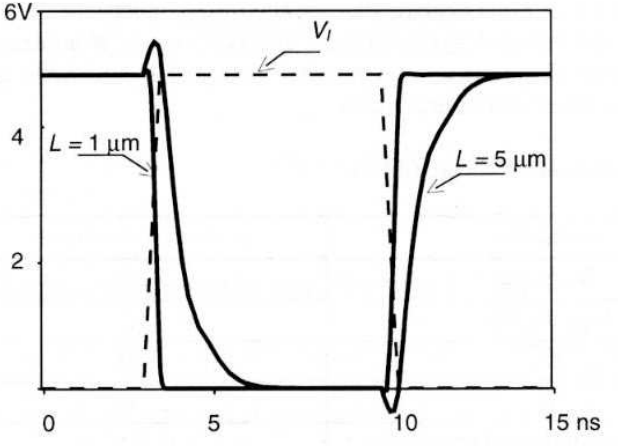


Figura 5.17 Caratteristiche dinamiche di due invertitori CMOS con $K_P = K_N$, $W/L|_N = 2$, e con MOS aventi spessore dell'ossido di gate di 100 nm o 30 nm e L pari a 5 o 1 μm.

La riduzione dello spessore t_{OX} dell'ossido di gate al di sotto di 20 nm, oltre a rendere più sensibile il circuito integrato alla presenza di microscopici vuoti nell'ossido, che portano al cortocircuito della gate e quindi creano scarti di produzione, provoca in ogni caso un aumento della corrente di tunnellaggio tra gate e body dovuta all'effetto Fowler-Nordheim in ossidi sottili (è noto che la probabilità di tunnellaggio di un elettrone con energia inferiore alla barriera di energia del sistema metallo-ossido silicio è inversamente proporzionale allo spessore della barriera, in questo caso lo spessore dell'ossido). Infine la resistenza degli ossidi sottili agli elevatissimi campi applicati (5 V applicati alla gate corrispondono ad un campo di 5 MV/cm su un ossido di 10 nm) si riduce in presenza di elettroni ener-

getici che passando nel canale possono essere intrappolati nell'ossido e ridurne il valore del campo di rottura (circa $2 \cdot 10^7$ V/cm nel caso ideale).

Tuttavia la riduzione delle dimensioni dei dispositivi è ancora il principale strumento per il miglioramento delle prestazioni dei circuiti integrati digitali; basti dire che oggi la minima dimensione del canale di MOS, anche per circuiti integrati VLSI, garantita da parecchie Silicon Foundries per circuiti progettati dall'utente (*Full custom circuits*) è inferiore a 0.6 μm, che per i microprocessori più avanzati in produzione si raggiungono dimensioni minime di canale di 0.35 μm e che sono in fase di realizzazione prototipale circuiti con dimensioni di 0.25 e 0.15 μm, si capisce come vi sia ancora spazio per significativi miglioramenti per i prossimi anni. Come esempio degli effetti della riduzione di scala sulle prestazioni dei dispositivi CMOS nell'ultimo decennio, si possono confrontare in Figura 5.17 le caratteristiche dinamiche di un invertitore CMOS realizzato con dispositivi con dimensione minima F di 2.5 μm e spessore dell'ossido di gate di 100 nm (0.1μm) con quelle ottenute da dispositivi prodotti attualmente, con dimensione minima F di 0.5 μm e spessore dell'ossido di gate di 30 nm; si può notare come nel primo caso, oltre ad un aumento dei tempi di commutazione, si presentino anche delle significative sovraelengazioni (sottoelongazioni) precedenti il fronte di discesa (di salita), dovute all'effetto di bypass della capacità C_{GD}, quest'ultima essenzialmente dovuta alla componente C_{GDO} legata alla parziale sovrapposizione tra l'elettrodo di gate e il drain.

5.13 Il fenomeno dell'aggancio (latch-up) nei CMOS

La presenza, nel processo CMOS, di una tasca di drogaggio opposto a quello del substrato, e delle regioni di source e drain dei transistori NMOS e PMOS, introduce, nella struttura esaminata, dei transistori bipolari parassiti di tipo NPN e PNP che possono dar luogo, in condizioni particolari di funzionamento, al fenomeno dell'aggancio (*latch-up*); questo fenomeno è da evitare perché, se innescato, può portare alla distruzione del circuito integrato.

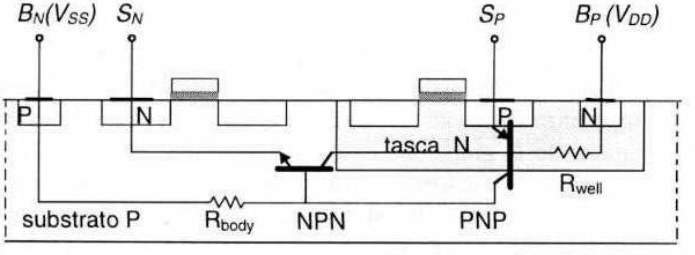


Figura 5.18 Struttura CMOS con i transistori parassiti PNP e NPN che danno origine al fenomeno del latch-up

In Figura 5.18 è riportata la struttura di un tipico processo CMOS con l'indicazione dei transistori NPN e PNP che vengono a formarsi a causa dell'alternanza di regioni P e N nella struttura realizzata. Come si può vedere dalla figura, la regione P del source S_P del PMOS, quella N della tasca, e quella P del substrato, danno luogo ad un transistore bipolare parassita PNP verticale, mentre l'alternanza della regione N della tasca, di quella P del substrato e di quella N del source S_N del NMOS, danno luogo ad un transistore bipolare parassita NPN laterale. Questi transistori sono normalmente interdetti, in quanto in regime di normale funzionamento le giunzioni tra S_P e tasca N, o tra substrato e S_N, sono polarizzate a zero volt, ma possono entrare in conduzione in particolari conduzioni, innescando così il fenomeno di "aggancio" o di latch-up, che porta ad una forte circolazione di corrente tra il terminale V_{DD} e quello di massa V_{SS}.

Per comprendere meglio il meccanismo che porta ad una elevata circolazione di corrente in questa struttura, facciamo riferimento allo schema di connessione dei due transistori riportato in Figura 5.19; in questa figura sono riportati i collegamenti tra i due transistori NPN e PNP evidenziati nella struttura CMOS di Figura 5.18, e le resistenze R_{WELL}, dovuta alla resistenza distribuita della tasca N, e R_{BODY}, dovuta alla resistenza del substrato, che giocano un ruolo rilevante nel fenomeno.

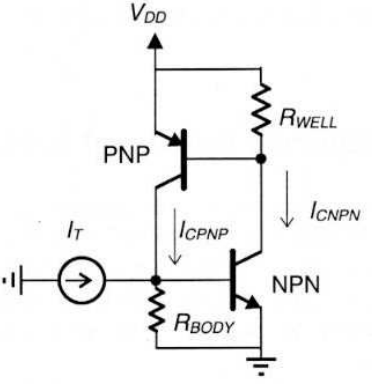


Figura 5.19 Connessione dei transistori parassiti PNP e NPN nella struttura CMOS

Supponiamo che per qualche motivo (ritorneremo su questo punto più avanti) vi sia un'iniezione di corrente I_T nel nodo di base del transistore NPN, tale da portare la caduta sulla resistenza R_{BODY} ad un valore di circa 0.7 V. Questo valore è tale da polarizzare direttamente la giunzione base-emettitore del transistore stesso, per cui quest'ultimo andrà in conduzione e farà circolare una corrente di collettore I_{CNPN}; a sua volta questa corrente, circolando nella resistenza R_{WELL} polarizzerà direttamente il transistore PNP che potrà andare in conduzione. La corrente di collettore I_{CPNP} a sua volta circolerà nella resistenza R_{BODY}, dando luogo ad un aumento della conduzione del NPN e quindi ad un aumento della sua corrente di collettore I_{CNPN}. Si innesca così un fenomeno rigenerativo che conduce alla piena conduzione dei due transistori, e che si autosostiene anche se si elimina la causa

che ha prodotto l'iniziale aumento di corrente nella base di uno dei transistori. La caduta totale di tensione tra alimentazione e massa (che prima dell'innesco del fenomeno è pari alla tensione di alimentazione V_{DD} in quanto le due giunzioni base-emettitore sono polarizzate a 0 V) crolla a valori di qualche volt, in quanto i due transistori si portano vicino alla saturazione, e vale $V_{BE}|_{NPN} + V_{EB}|_{PNP} + V_{CB}|_{NPN}$; la corrente circolante (entrante dal source del PMOS ed uscente dal source del NMOS) può essere relativamente elevata (alcuni mA) e tale da danneggiare il circuito integrato nella maggior parte dei casi.

In condizioni normali di funzionamento, come si è detto, le due giunzioni base-emettitore dei due transistori corrispondono alle giunzioni tra source S_P e tasca N o fra source S_N e substrato P, per cui il transistore parassita è certamente interdetto, e inoltre non circolano correnti significative nel substrato o nella tasca (le uniche correnti che consideriamo sono quelle di canale che circolano tra source e drain). Tuttavia la situazione idealizzata riportata in Figura 5.1, in cui i source del PMOS e del NMOS sono realizzati in contiguità con i rispettivi contatti di substrato, e a questi ultimi connessi direttamente con una metallizzazione, non è sempre verificata, potendosi avere più di un transistore PMOS in ogni tasca, non tutti fisicamente connessi con il contatto di tasca, come anche più transistori NMOS, non tutti connessi fisicamente al terminale di substrato. In questi casi è possibile che, a causa di transitori dovuti ai circuiti connessi con i terminali in questione, i potenziali dei source siano diversi da quelli dei contatti di substrato o di tasca, e si inneschi quindi una corrente transitoria che, circolando nelle resistenze distribuite di tasca o di substrato, instauri il fenomeno del latch-up della struttura.

Le tecniche per prevenire il fenomeno del latch-up sono diverse e comportano un innalzamento della soglia del fenomeno in modo che questo non possa aver luogo in nessuna condizione ragionevole di funzionamento. La più diretta è quella di una progettazione più accurata del tracciato (lay-out) del circuito in modo da ridurre la distanza e la resistenza tra i contatti di substrato o di tasca e quelli di source dei rispettivi MOS, utilizzando per quanto possibile un contatto di substrato per ogni terminale connesso all'alimentazione o a massa, e riducendo la distanza tra i contatti di substrato e di source. Una ulteriore contromisura è quella di ridurre i valori delle resistenze distribuite R_{WELL} e R_{BODY} aumentando i drogaggi rispettivi; inoltre una efficace azione è quella di ridurre il guadagno di corrente dei due transistori parassiti, e specialmente quello verticale PNP che presenta un guadagno di corrente significativo a causa dello spessore ridotto della tasca (la base del PNP).

Un processo che riduce significativamente il fenomeno del latch-up è quello riportato in Figura 5.20, che fa uso come materiale di partenza di un substrato molto drogato con un sottile strato di silicio epitassiale poco drogato sulla superficie superiore (ritorneremo su questo processo epitassiale nel Capitolo 6, parlando dei transistori bipolari). In questo caso si può utilizzare un processo a doppia tasca per realizzare sia i transistori PMOS che NMOS, mediante una doppia impiantazione sia N che P, in modo da realizzare entrambi i substrati equivalenti con un drogaggio relativamente alto. Inoltre la corrente di collettore del PNP verticale viene raccolta dal

In Figura 5.18 è riportata la struttura di un tipico processo CMOS con l'indicazione dei transistori NPN e PNP che vengono a formarsi a causa dell'alternanza di regioni P e N nella struttura realizzata. Come si può vedere dalla figura, la regione P del source S_P del PMOS, quella N della tasca, e quella P del substrato, danno luogo ad un transistore bipolare parassita PNP verticale, mentre l'alternanza della regione N della tasca, di quella P del substrato e di quella N del source S_N del NMOS, danno luogo ad un transistore bipolare parassita NPN laterale. Questi transistori sono normalmente interdetti, in quanto in regime di normale funzionamento le giunzioni tra S_P e tasca N, o tra substrato e S_N, sono polarizzate a zero volt, ma possono entrare in conduzione in particolari conduzioni, innescando così il fenomeno di "aggancio" o di latch-up, che porta ad una forte circolazione di corrente tra il terminale V_{DD} e quello di massa V_{SS}.

Per comprendere meglio il meccanismo che porta ad una elevata circolazione di corrente in questa struttura, facciamo riferimento allo schema di connessione dei due transistori riportato in Figura 5.19; in questa figura sono riportati i collegamenti tra i due transistori NPN e PNP evidenziati nella struttura CMOS di Figura 5.18, e le resistenze R_{WELL}, dovuta alla resistenza distribuita della tasca N, e R_{BODY}, dovuta alla resistenza del substrato, che giocano un ruolo rilevante nel fenomeno.

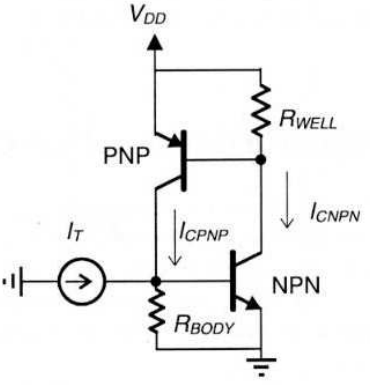


Figura 5.19 Connessione dei transistori parassiti PNP e NPN nella struttura CMOS

Supponiamo che per qualche motivo (ritorneremo su questo punto più avanti) vi sia un'iniezione di corrente I_T nel nodo di base del transistore NPN, tale da portare la caduta sulla resistenza R_{BODY} ad un valore di circa 0.7 V. Questo valore è tale da polarizzare direttamente la giunzione base-emettitore del transistore stesso, per cui quest'ultimo andrà in conduzione e farà circolare una corrente di collettore I_{CNPN}; a sua volta questa corrente, circolando nella resistenza R_{WELL} polarizzerà direttamente il transistore PNP che potrà andare in conduzione. La corrente di collettore I_{CPNP} a sua volta circolerà nella resistenza R_{BODY}, dando luogo ad un aumento della conduzione del NPN e quindi ad un aumento della sua corrente di collettore I_{CNPN}. Si innesca così un fenomeno rigenerativo che conduce alla piena conduzione dei due transistori, e che si autosostiene anche se si elimina la causa

che ha prodotto l'iniziale aumento di corrente nella base di uno dei transistori. La caduta totale di tensione tra alimentazione e massa (che prima dell'innesco del fenomeno è pari alla tensione di alimentazione V_{DD} in quanto le due giunzioni base-emettitore sono polarizzate a 0 V) crolla a valori di qualche volt, in quanto i due transistori si portano vicino alla saturazione, e vale $V_{BE}|_{NPN} + V_{EB}|_{PNP} + V_{CB}|_{NPN}$; la corrente circolante (entrante dal source del PMOS ed uscente dal source del NMOS) può essere relativamente elevata (alcuni mA) e tale da danneggiare il circuito integrato nella maggior parte dei casi.

In condizioni normali di funzionamento, come si è detto, le due giunzioni base-emettitore dei due transistori corrispondono alle giunzioni tra source S_P e tasca N o fra source S_N e substrato P, per cui il transistore parassita è certamente interdetto, e inoltre non circolano correnti significative nel substrato o nella tasca (le uniche correnti che consideriamo sono quelle di canale che circolano tra source e drain). Tuttavia la situazione idealizzata riportata in Figura 5.1, in cui i source del PMOS e del NMOS sono realizzati in contiguità con i rispettivi contatti di substrato, e a questi ultimi connessi direttamente con una metallizzazione, non è sempre verificata, potendosi avere più di un transistore PMOS in ogni tasca, non tutti fisicamente connessi con il contatto di tasca, come anche più transistori NMOS, non tutti connessi fisicamente al terminale di substrato. In questi casi è possibile che, a causa di transitori dovuti ai circuiti connessi con i terminali in questione, i potenziali dei source siano diversi da quelli dei contatti di substrato o di tasca, e si inneschi quindi una corrente transitoria che, circolando nelle resistenze distribuite di tasca o di substrato, instauri il fenomeno del latch-up della struttura.

Le tecniche per prevenire il fenomeno del latch-up sono diverse e comportano un innalzamento della soglia del fenomeno in modo che questo non possa aver luogo in nessuna condizione ragionevole di funzionamento. La più diretta è quella di una progettazione più accurata del tracciato (lay-out) del circuito in modo da ridurre la distanza e la resistenza tra i contatti di substrato o di tasca e quelli di source dei rispettivi MOS, utilizzando per quanto possibile un contatto di substrato per ogni terminale connesso all'alimentazione o a massa, e riducendo la distanza tra i contatti di substrato e di source. Una ulteriore contromisura è quella di ridurre i valori delle resistenze distribuite R_{WELL} e R_{BODY} aumentando i drogaggi rispettivi; inoltre una efficace azione è quella di ridurre il guadagno di corrente dei due transistori parassiti, e specialmente quello verticale PNP che presenta un guadagno di corrente significativo a causa dello spessore ridotto della tasca (la base del PNP).

Un processo che riduce significativamente il fenomeno del latch-up è quello riportato in Figura 5.20, che fa uso come materiale di partenza di un substrato molto drogato con un sottile strato di silicio epitassiale poco drogato sulla superficie superiore (ritorneremo su questo processo epitassiale nel Capitolo 6, parlando dei transistori bipolari). In questo caso si può utilizzare un processo a doppia tasca per realizzare sia i transistori PMOS che NMOS, mediante una doppia impiantazione sia N che P, in modo da realizzare entrambi i substrati equivalenti con un drogaggio relativamente alto. Inoltre la corrente di collettore del PNP verticale viene raccolta dal

substrato molto drogato e quindi non è efficace nel far condurre il transistore NPN laterale.

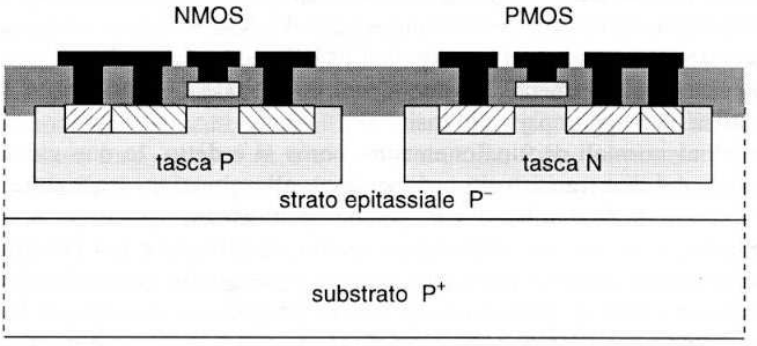


Figura 5.20 Processo a doppia tasca su substrato epitassiale per la riduzione del fenomeno del latch-up

Esercizi di riepilogo

5.1 Dato un invertitore CMOS con i dispositivi caratterizzati dai seguenti parametri: $k'_N = 50\ \mu A/V^2$, $k'_P = 20\ \mu A/V^2$, $V_{TN} = |V_{TP}| = 0.8$ V, $W_N = W_P = 2$ µm, $L_N = L_P = 1$ µm, e con una tensione di alimentazione $V_{DD} = 5$ V, determinare il valore della tensione di soglia della caratteristica di trasferimento.

5.2 Disegnare il tracciato di un invertitore CMOS che impiega un NMOS ad area minima e un PMOS con $W_P = 2.5W_N$, realizzati con un processo con $\lambda = 0.6$ µm, riferendosi alle regole di progetto riportate in Tabella 2.2. Valutare il ritardo di propagazione per $V_{DD} = 5$ V, assumendo l'invertitore caricato da un uguale invertitore, utilizzando le formule analitiche approssimate. Si consideri come capacità equivalente di carico la sola capacità totale di gate e si utilizzi il valore riportato in Tabella 3.2 per la sua capacità unitaria.

5.3 Per l'invertitore dell'Esercizio 5.2, si valuti l'influenza delle capacità di drain dei MOS dell'invertitore calcolando il ritardo di propagazione con le formule analitiche, considerando il contributo di queste ultime capacità nel valore della capacità totale di carico (si utilizzi ancora la Tabella 3.2 per i valori delle capacità unitarie C_{JO} C_{JW} e di quelle di overlap). Confrontare i risultati ottenuti sia nel caso della sola capacità di gate, sia considerando la somma delle capacità di gate e di drain, con quelli ottenuti da simulazioni SPICE per l'invertitore assegnato (si utilizzino nella scheda .MODEL dei MOS gli stessi valori delle capacità unitarie assunti per l'analisi manuale).

5.4 Per l'invertitore dell'Esercizio 5.2, valutare l'influenza del tempo di salita sul tempo di propagazione mediante un'analisi SPICE, variando il tempo di salita e di discesa del segnale di ingresso da 0.1 ns a 2 ns. Si estrapoli dalla curva ottenuta il valore del ritardo di propagazione per un tempo di salita nullo e lo si confronti con i risultati dell'analisi approssimata svolti nell'Esercizio 5.3.

5.5 Per un invertitore con i seguenti parametri dei dispositivi: $k'_N = 50\ \mu A/V^2$, $k'_P = 20\ \mu A/V^2$, $V_{TN} = |V_{TP}| = 0.8$ V, $W_N = W_P = 2\ \mu m$, $L_N = L_P = 1\ \mu m$, caricato da un uguale invertitore, valutare mediante le formule approssimate i due contributi P'_D e P''_D di potenza dissipata, per una frequenza del segnale di ingresso di 50 MHz (si assumano i tempi di transizione in salita e in discesa del segnale di ingresso, nella valutazione del contributo P'_D, pari al doppio del tempo di propagazione dell'invertitore).

5.6 Disegnare i tracciati rispettivamente per una porta NAND e per una porta NOR, entrambi a tre ingressi, dimensionando gli NMOS e i PMOS in modo da ottenere uguali valori per i tempi di propagazione t_{PLH} e t_{PHL} in base alle formule teoriche approssimate, identificando per ogni caso la condizione peggiore nella transizione dell'uscita da un valore logico all'altro. Confrontare quindi le aree di gate e quelle complessive delle due porte.

5.7 Per l'invertitore dell'Esercizio 5.2 determinare il fan-out assumendo come massimo valore accettabile per il ritardo di propagazione il valore di 1 ns.

5.8 Si desidera realizzare un segnale di fase ϕ e il suo negato $\bar{\phi}$ a partire da un unico segnale ad onda quadra, realizzato da una catena di invertitori che hanno una capacità di ingresso di 25 fF e tempo di propagazione $t_P = 0.1$ ns. Progettare due serie di stadi di separazione per pilotare, con i segnali ϕ e $\bar{\phi}$, due circuiti che presentano capacità di ingresso equivalente pari a 10 pF, mantenendo contenuto il ritardo di propagazione complessivo ed inalterata la condizione di complementarietà all'ingresso dei due circuiti. Valutare quindi il ritardo di propagazione così ottenuto rispetto a quello ottenibile da una sola catena di stadi di separazione ottimizzata.

5.9 Con riferimento alla porta NAND a tre ingressi di Figura E5.1, si disegni il tracciato in base ad un processo con $\lambda = 0.6\ \mu m$, riferendosi alle regole di progetto riportate in Tabella 2.2, e si estragga il listato SPICE del circuito, dimensionando i MOS con i seguenti parametri: $k'_N = 50\ \mu A/V^2$, $k'_P = 20\ \mu A/V^2$, $V_{TN} = |V_{TP}| = 0.8$ V, $W_N = 2\ \mu m$, $W_P = 2.5\ W_N$, $L_N = L_P = 1\ \mu m$, utilizzando i valori riportati in Tabella 3.3 per le capacità unitarie, e assumendo la porta caricata da un invertitore con uguali valori dei parametri dei MOS. Si valutino infine, mediante simulazioni SPICE, i tempi di propagazione t_{PHL} nel nodo Y nei due casi: a) $A = 1$, $B = C = 0 \rightarrow 1$; b) $A = B = C = 0 \rightarrow 1$.

Spiegare perchè il t_{PHL} nel caso a) è maggiore che nel caso b) (suggerimento: si valuti mediante analisi SPICE il comportamento nel nodo 1).

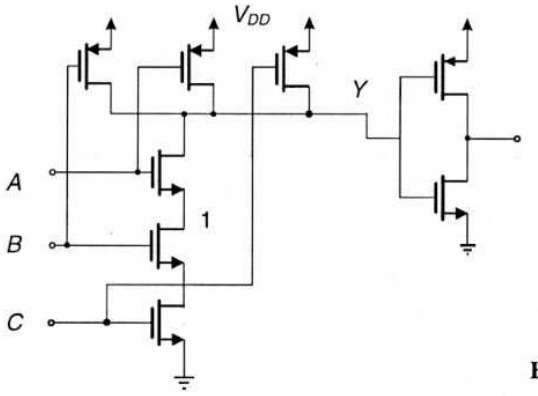


Figura E5.1

5.10 Progettare il tracciato di una porta NAND a 6 ingressi con dispositivi ad area minima ($W_N = 3\lambda/2\lambda$, $W_P = 2.5W_N$, $L_N = L_P = 2\lambda$,) realizzati con un processo con $\lambda = 0.6\mu m$, riferendosi alle regole di progetto riportate in Tabella 2.2. Valutare mediante analisi teoriche il tempo di propagazione nella condizione peggiore, considerando come carico un invertitore con gli stessi valori di W e L utilizzati per la porta, e confrontare questi valori con quelli corrispondenti alla soluzione circuitale di Figura 5.12b (per il caso di 6 ingressi), assumendo lo stesso invertitore di carico per le due soluzioni. Paragonare infine questi risultati con quelli ottenuti mediante analisi SPICE dei due circuiti, confrontando inoltre gli andamenti dei transitori di commutazione per le due soluzioni.

Riferimenti bibliografici

G.M. Glansford, *Digital Electronic Circuits*, Prentice Hall Int., 1988.

A.S. Sedra, K.C. Smith, *Microelectronic Circuits*, Saunders College publ., 1991.

B. Riccò, F. Fantini, P. Brambilla, *Introduzione ai circuiti integrati digitali*, Zanichelli, Bologna, 1991.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

6

Il transistore bipolare

6.1 Struttura del transistore bipolare

I transistori bipolari costituiscono i dispositivi attivi di una famiglia tecnologica che da questi prende il nome di tecnologia bipolare, e che comprende, oltre ai transistori, anche diodi, resistori e condensatori con i quali vengono realizzati i circuiti integrati delle famiglie logiche bipolari.

Nel Paragrafo 2.3 si è sinteticamente riportata la sequenza di operazioni tecnologiche necessarie per la realizzazione di un transistore bipolare a partire da un substrato di tipo P. Come si è già detto, i transistori bipolari sono dispositivi intrinsecamente verticali (in altre parole la corrente controllata fluisce in direzione perpendicolare alla superficie del wafer, al contrario dei MOS che sono dispositivi in cui la corrente fluisce nel canale parallelamente alla superficie del wafer); questo crea problemi sia per la tecnologia di realizzazione che per l'area occupata, in quanto occorre che la corrente assorbita dalla regione di collettore (lo strato sepolto N^+) sia riportata in superficie mediante un'opportuna regione di collegamento con lo strato sepolto. Inoltre, non essendo i dispositivi intrinsecamente isolati dallo strato epitassiale in cui questi sono creati (diversamente dai MOS che sono intrinsecamente isolati dal substrato), occorre creare delle regioni di isolamento intorno alle aree attive, con ulteriore occupazione di spazio.

Il tracciato di un transistore bipolare ad area minima discende abbastanza direttamente dalle regole di progetto utilizzate, come nel caso di un transistore MOS. Queste sono riportate nella Tabella 6.1 per transistori bipolari con tecnologia standard (cioè con isolamento a giunzione, senza processi autoallineanti e utilizzo di polisilicio: di questi processi innovativi daremo qualche cenno nel Paragrafo 6.7); rispetto alle regole di progetto presentate per la tecnologia MOS le differenze principali riguardano la regione di strato sepolto, le distanze delle diffusioni dalla regione di isolamento, nonché la distanza tra la diffusione di emettitore e quella di base che la deve contenere.

Tabella 6.1 Regole di progetto per il tracciato di transistori bipolari

regione	*minima apertura*	*minima separazione*
apertura contatti (vias)	2λ	
larghezza metallo	3λ	3λ
metallizzazione contatti	4λ	
diffusione emettitore	3λ	
diffusione base-diffusione emettitore		2λ
diffusione base-diffusione isolamento		4λ
diffusione strato sepolto-isolamento		4λ
diffusione isolamento	3λ	

Sulla base delle regole di progetto indicate, in Figura 6.1 sono sinteticamente riportate la struttura verticale e il tracciato orizzontale di un transistore bipolare convenzionale, con isolamento a giunzione.

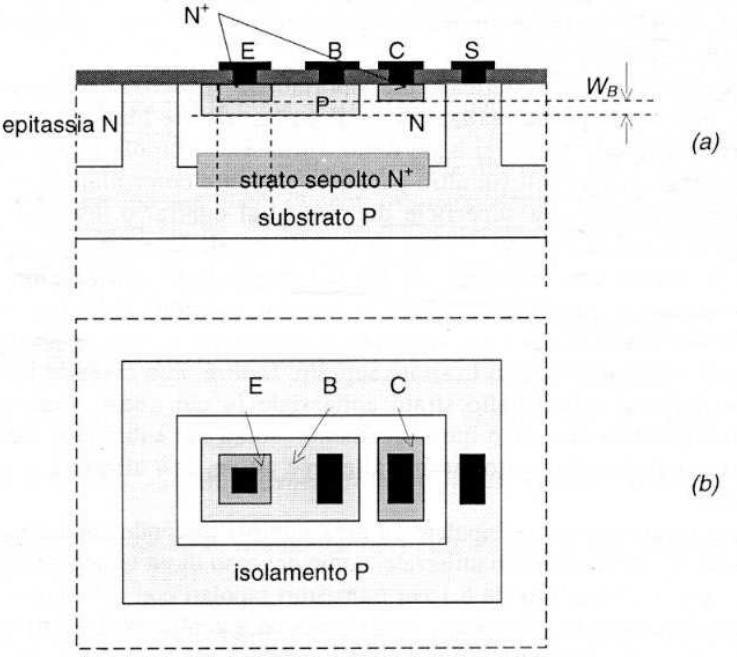


Figura 6.1 a) Sezione verticale e b) tracciato orizzontale di un transistore bipolare

Come è noto, il transistore bipolare è basato sulla stretta interazione di due giunzioni *N/P* e *P/N* poste ad una distanza W_B; queste due giunzioni sono realizzate

creando una regione di drogaggio P (base) in uno strato epitassiale N (collettore), e successivamente una seconda regione N (emettitore) nella regione di base così formata. Dalla Figura 6.1 si può vedere come la regione di base si estenda lateralmente ben oltre l'area di emettitore per permettere la connessione al contatto corrispondente; anche lo strato sepolto, che agisce essenzialmente da strato conduttore in cui la corrente circola con bassa caduta per essere poi estratta dal contatto di collettore in superficie, si estende oltre la regione di base per minimizzare la resistenza tra lo strato sepolto e il contatto di collettore. Infine la regione di isolamento occupa un'area relativamente grande rispetto alle regioni precedenti, e questo perché la diffusione successiva all'impiantazione deve raggiungere il substrato attraverso lo strato epitassiale e il processo di diffusione necessario tende ad allargare la regione anche in direzione orizzontale.

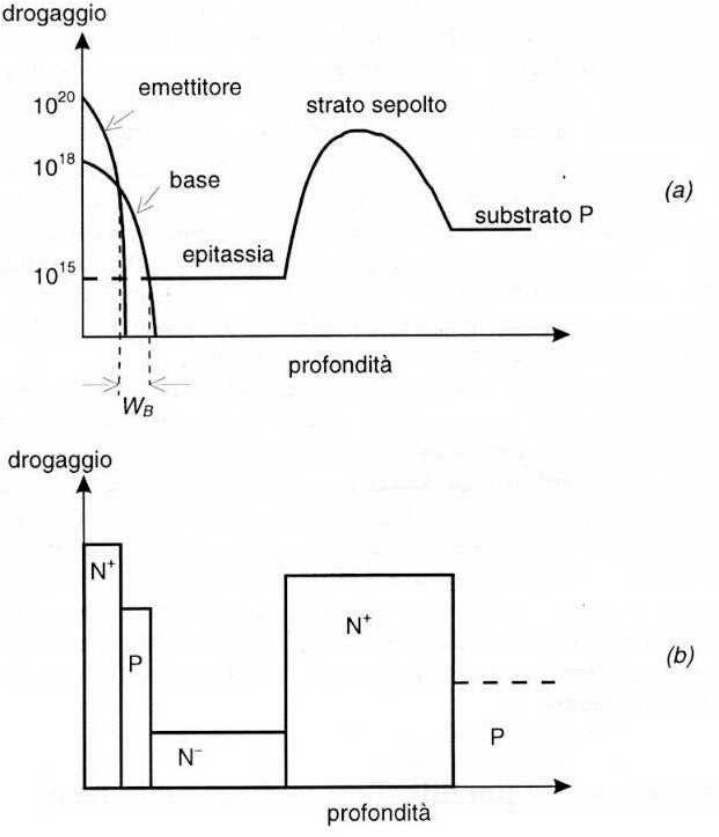


Figura 6.2 a) Profilo del drogaggio lungo la verticale sotto l'emettitore del transistore intrinseco; b) approssimazione a drogaggio costante per le regioni di emettitore, base e collettore

Tutto ciò fa sì che la parte della struttura che è effettivamente attiva e costituisce il *transistore intrinseco* (sezione tra le linee tratteggiate nella Figura 6.1a) sia una parte molto ridotta della struttura totale (questo non è più vero nei dispositivi di nuova generazione che fanno uso di tecnologie più avanzate e di processi autoallineanti, che verranno discussi con riferimento alle famiglie logiche avanzate). La parte attiva del transistore è in effetti la regione corrispondente alla proiezione verticale dell'emettitore nella base e poi nello strato epitassiale fino allo strato sepolto (regione di collettore), ed è costituita dall'alternanza di tre regioni rispettivamente drogate N (emettitore), P (base), ed ancora N (collettore).

Perché il transistore bipolare abbia un efficace controllo della corrente principale (corrente di collettore) tramite il terminale di controllo (base) è necessario che:

a. il drogaggio dell'emettitore sia più elevato di quello della base;
b. lo spessore della regione di base sia quanto più piccolo possibile.

Entrambe le condizioni vengono realizzate mediante il processo tecnologico di diffusione dell'emettitore nella base. In Figura 6.2 è riportato un tipico profilo della distribuzione di drogaggio delle tre regioni lungo una linea verticale che attraversa la regione di emettitore; la distanza W_B tra la giunzione di emettitore e quella di base è inferiore al micron e può essere controllata fino a valori dell'ordine di 0.1 μm. Nonostante lo spessore infinitesimo della regione di base, le caratteristiche elettriche del transistore sono essenzialmente determinate dalla distribuzione dei portatori all'interno di questa regione, che quindi assume importanza rilevante nella struttura globale (in analogia con questo aspetto, nel transistore MOS troviamo una corrispondenza con il canale sotto la gate, che nonostante il suo ridottissimo spessore e la sua lunghezza contenuta rispetto alle altre dimensioni della struttura determina in larga misura le caratteristiche elettriche del MOS).

Le distribuzioni pseudo-gaussiane del drogaggio che derivano dai processi di impiantazione e diffusione del drogante sono di solito approssimate, nelle analisi del primo ordine, con delle distribuzioni uniformi di drogaggio nelle tre regioni, in modo da semplificare la descrizione del comportamento delle due giunzioni di emettitore e di collettore, e pervenire a semplici relazioni analitiche per le caratteristiche del dispositivo. Richiameremo qui le principali relazioni che determinano il comportamento sia statico che dinamico del transistore bipolare, procedendo in maniera euristica ed intuitiva, analogamente al caso del transistore MOS, in modo da ricavare in via semplificata i modelli analitici che vengono utilizzati nei simulatori circuitali come SPICE, e definire i parametri fondamentali che li reggono.

6.2 Distribuzione dei portatori minoritari nella base

Con riferimento all'approssimazione di drogaggio uniforme nelle tre regioni del transistore (Figura 6.2b), in assenza di polarizzazione su ciascuna delle regioni i portatori maggioritari e minoritari di ciascuna regione si trovano in equilibrio termico, e le loro concentrazioni obbediscono alla legge:

$$p \cdot n = n_i^2 = 2 \cdot 10^{20} \tag{6.1}$$

con n_i concentrazione dei portatori (maggioritari e minoritari) nel volume unitario per un materiale intrinseco (cioè non drogato), pari a $\cong 1.4 \cdot 10^{14}$ cm^{-3} a temperatura ambiente.

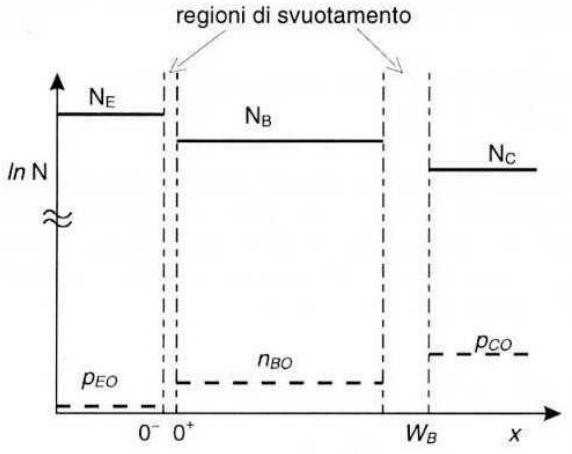


Figura 6.3 Distribuzioni dei portatori all'equilibrio

Questo valore di concentrazione intrinseca è ben inferiore a qualsiasi valore di drogaggio realizzabile nel silicio, per cui nel materiale drogato la concentrazione dei portatori introdotti dal drogante (detti portatori maggioritari) sarà pari a quella del drogante, e quella dei portatori di segno opposto (detti portatori minoritari) sarà corrispondentemente ridotta, in accordo con la (6.1). Quindi nelle tre regioni di emettitore, base e collettore, con diverso valore (e tipo) di drogaggio, i portatori maggioritari saranno pari al valore del drogaggio della regione stessa (vedi Figura 6.3), e quelli minoritari rispettivamente pari a:

$$p_{E0} = \frac{n_i^2}{N_E} \cong \frac{2 \cdot 10^{20}}{10^{19}} = 20 \tag{6.2}$$

$$n_{B0} = \frac{n_i^2}{N_B} \cong \frac{2 \cdot 10^{20}}{10^{17}} = 2 \cdot 10^3 \tag{6.3}$$

$$p_{C0} = \frac{n_i^2}{N_C} \cong \frac{2 \cdot 10^{20}}{10^{15}} = 2 \cdot 10^5 \tag{6.4}$$

In presenza di polarizzazione diretta sulla giunzione emettitore-base, la barriera di potenziale tra le due regioni dovuta alla differenza di drogaggio si riduce, e una parte delle cariche maggioritarie di ciascuna regione diffonde nell'altra regione, dove diventa (a causa del differente segno del drogaggio di questa) portatore minoritario. Per ciascuna delle due regioni, le concentrazioni di cariche minoritarie iniettate rispettivamente al confine della regione di base o di emettitore sono legate a quelle di equilibrio n_{B0} o p_{E0} (in assenza di polarizzazione) dalle relazioni:

$$n_B(0^+) = n_{B0} \cdot \exp\left(\frac{V_{BE}}{V_T}\right); \quad \left[p_E(0^-) = p_{E0} \cdot \exp\left(\frac{V_{BE}}{V_T}\right)\right] \tag{6.5}$$

dove V_T è la tensione termica KT/q pari a circa 25 mV a temperatura ambiente, e V_{BE} la tensione sulla giunzione base-emettitore (positiva sulla base). Queste relazioni valgono anche nel caso di polarizzazione inversa (negativa sulla base), per la quale si ha che la concentrazione di portatori minoritari si riduce rispetto al valore di equilibrio, fino a tendere a zero per $|-V_{BE}| >> V_T$.

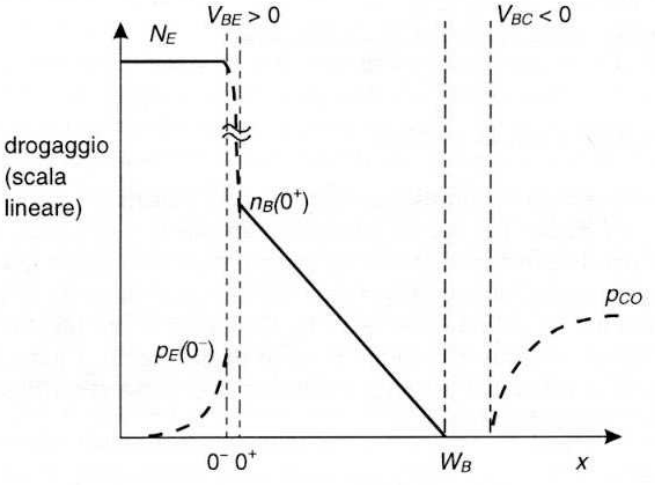


Figura 6.4 Distribuzione dei portatori in regime attivo diretto

In base alla (6.5), che costituisce l'equazione fondamentale per il comportamento di giunzioni P/N, si può comprendere la distribuzione dei portatori minoritari nelle tre regioni del transistore elementare di Figura 6.4, nell'ipotesi di polarizzazione diretta della giunzione base-emettitore e inversa di quella base-collettore (base negativa rispetto al collettore), attraverso i rispettivi terminali esterni. Ai capi della giunzione direttamente polarizzata (B-E) vi è una iniezione rispettivamente di lacune dalla base nell'emettitore, e di elettroni dall'emettitore nella base, in eccesso sui valori di equilibrio, mentre ai capi della giunzione contropolarizzata (B-C) vi è

una diminuzione delle rispettive concentrazioni di minoritari rispetto ai valori di equilibrio. Si comprende anche perché il collettore debba avere drogaggio minore di quello della base: in questo modo in condizione di contropolarizzazione anche relativamente elevata della giunzione di collettore, la regione di svuotamento si estende essenzialmente nel collettore (a più basso drogaggio) con possibilità di sostenere una maggiore tensione massima (tensione di *breakdown*), evitando il fenomeno di perforazione (*punch-through*) della base, che si avrebbe se la regione di svuotamento si sviluppasse nella regione di base e raggiungesse la giunzione di emettitore.

La condizione di polarizzazione diretta per la giunzione di emettitore e inversa per quella di collettore è la condizione di normale funzionamento del transistore bipolare, detta *regime attivo diretto*. In questo caso la concentrazione di elettroni nella base al limite della giunzione di emettitore è data dalla (6.5), e quella al limite della giunzione di collettore vale:

$$n_B(W_B) = n_{B0} \exp\frac{V_{BC}}{V_T} \cong 0 \qquad \left[V_{BC} < 0, |V_{BC}| > V_T\right] \tag{6.6}$$

Se la distanza W_B tra le due giunzioni è piccola, tale cioè che gli elettroni la possano percorrere in un tempo minore di quello necessario per ricombinarsi (detto *vita media* nella base τ_B), la distribuzione sarà lineare, a partire dal valore $n_B(0^+)$ iniettato dall'emettitore fino a zero sulla giunzione di collettore.

Questa distribuzione dà luogo ad una corrente di diffusione di elettroni I_n, (negativa per le convenzioni adottate su I_C e I_E), legata al gradiente dell'eccesso n'_B della concentrazione di portatori rispetto al valore di equilibrio dalla relazione:

$$I_n = -qAD_n \frac{d(n_B - n_{B0})}{dx} \cong qAD_n \frac{n_B(0^+) - n_{B0}}{W_B} \equiv qAD_n \frac{n'_B(0)}{W_B} \tag{6.7}$$

dove D_n è il coefficiente di diffusione degli elettroni, A l'area in cui vengono iniettati i portatori, e l'approssimazione è legata all'ipotesi di andamento lineare delle cariche. Sostituendo la (6.5) nella (6.7) si ha per la corrente di diffusione dall'emettitore:

$$I_{nE} = q\frac{AD_n n_{B0}}{W_B}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \tag{6.8}$$

6.3 Regimi di funzionamento del transistore bipolare

Regione attiva

L'espressione della corrente I_{nE} è formalmente identica a quella della corrente di un diodo P/N direttamente polarizzato, ed infatti rappresenta la componente I_{CF} della

corrente dell'emettitore che viene iniettata nel collettore per una data polarizzazione diretta V_{BE}; il pedice F sta ad indicare il *funzionamento attivo diretto (forward)* del transistore.

La corrente totale I_{EF} al terminale di emettitore è formata oltre che da questa componente (che è la parte rilevante) anche dalla corrente di lacune iniettata dalla base (che però è molto più piccola perché anch'essa espressa da una relazione simile alla (6.8) ma riferita alla concentrazione delle lacune nell'emettitore p_{EO} che per la (6.2) è molto minore di n_{BO}), per cui si può esprimere la corrente di diffusione di elettroni I_{CF} come aliquota della corrente I_{EF}:

$$I_{nE} = \alpha_F I_{EF} = \alpha_F I_{ES}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \tag{6.9}$$

dove I_{ES} rappresenta, come nella equazione di un diodo, la corrente inversa di saturazione della giunzione di emettitore, e α_F rappresenta l'aliquota I_{nE} della corrente I_{EF} ($\alpha_F < 1$ ma molto vicino a 1) che viene iniettata nel collettore.

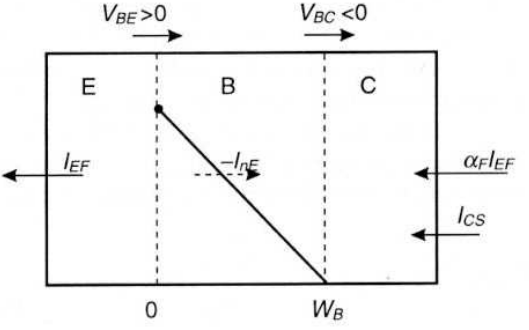


Figura 6.5 Componenti delle correnti ai terminali in regime attivo diretto

Si comprende da queste considerazioni perché è necessario che il drogaggio dell'emettitore debba essere ben maggiore di quello della base per un buon funzionamento del transistore: solo in tal caso $I_{nE} \equiv I_{EF} >> I_{nB}$, e quindi $\alpha_F \equiv 1$, il che, come vedremo, comporta un significativo valore del guadagno in corrente β_F.

In Figura 6.5 sono sinteticamente indicate le componenti delle correnti di emettitore e collettore legate nel funzionamento attivo diretto (la freccia tratteggiata indica la componente di diffusione degli elettroni nella base che, per la carica negativa dell'elettrone, corrisponde ad una corrente di segno opposto).

Da questa figura si vede che la corrente complessiva di collettore I_{CF} può scriversi come la somma della componente I_{nE} e quella (ridottissima) della corrente inversa della giunzione base-collettore I_{CS}:

$$I_{CF} = \alpha_F I_{EF} + I_{CS} = \alpha_F I_{ES}\left(\exp\frac{V_{BE}}{V_T} - 1\right) + I_{CS} \tag{6.10}$$

Data la struttura del transistore, è possibile scambiare tra loro i morsetti di collettore e di emettitore, in altre parole si può polarizzare direttamente la giunzione base-collettore, che diventa la giunzione iniettante, ed inversamente quella base-emettitore, che diventa quella di raccolta. In questo modo di funzionamento il transistore opera in *regime attivo inverso*, indicato con il pedice R (*reverse*), dove le distribuzioni di corrente e le correnti sono quelle schematizzate nella Figura 6.6.

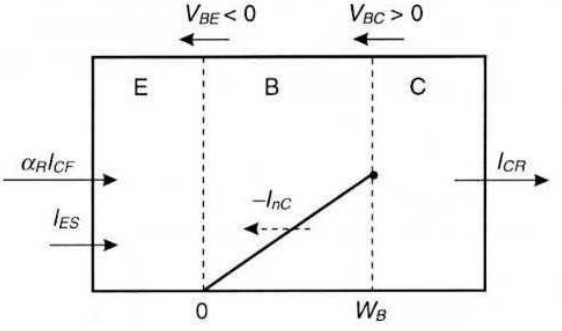


Figura 6.6 Componenti delle correnti ai terminali in regime attivo inverso

Le relazioni tra le correnti sono formalmente uguali a quelle (6.10) valide per il regime diretto, pur di scambiare tra loro i pedici di emettitore e collettore; una differenza fondamentale però consiste nel valore del coefficiente α_R. Per quanto detto precedentemente sull'effetto che ha la differenza di drogaggio tra emettitore e base sul valore di α_F, in questo caso, essendo l'emettitore equivalente (ora collettore) molto meno drogato della base, la componente della corrente di collettore che viene iniettata nella base è minore di quella iniettata dalla base nel collettore, per cui α_R sarà molto minore di 1. La corrente di emettitore (considerata sempre uscente dal terminale di emettitore), che ora corrisponde alla corrente di collettore equivalente nel regime inverso, sarà espressa da:

$$I_{ER} = -\alpha_R I_{CR} - I_{ES} = -\alpha_R I_{CS}\left(\exp\frac{V_{BC}}{V_T} - 1\right) - I_{ES} \tag{6.11}$$

dove, in analogia con la (6.10), la corrente di collettore in regime inverso I_{CR} è espressa dall'equazione del diodo, con $V_{BC} > 0$; I_{CS}, I_{ES} sono le correnti di saturazione inverse delle due giunzioni (già definite nella (6.10)).

Regione di saturazione

È anche possibile il funzionamento del transistore con entrambe le giunzioni direttamente polarizzate; questo modo di funzionamento è definito *regime di saturazione*, perché il transistore in questo caso ha una corrente di uscita che non è più dipendente da quella di controllo, e una tensione di uscita pari alla minima possibile. La distribuzione di corrente in questo regime può essere ricavata con il principio di sovrapposizione degli effetti, combinando insieme i regimi attivo diretto (nel quale la giunzione di emettitore è direttamente polarizzata) con quello inverso (nel quale la giunzione di collettore è direttamente polarizzata). In Figura 6.7 è riportata la combinazione delle due distribuzioni di portatori nella base per il regime di saturazione; si vede che la distribuzione risultante ha un gradiente minore di quello di ognuna delle due distribuzioni precedenti, indicando quindi una corrente di uscita minore di quella del regime attivo (a parità di polarizzazione diretta del terminale di controllo), e tanto minore quanto più la tensione sulla giunzione di collettore è prossima a quella di emettitore (e quindi $n_B(W_B)$ è circa uguale a $n_B(0)$).

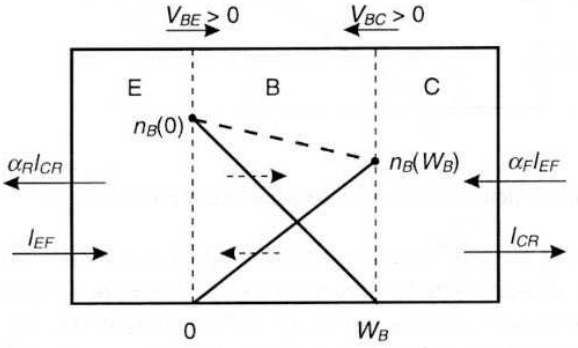


Figura 6.7 Componenti delle correnti ai terminali in regime di saturazione

6.4 Modello di Ebers Moll e caratteristiche *I-V*

Nel caso più generale di polarizzazioni qualsiasi per ognuna delle due giunzioni, le correnti ai tre terminali possono essere descritte da un sistema di equazioni, che definiscono il modello statico ad ampi segnali del primo ordine, detto *modello di Ebers e Moll*, e che scaturiscono da una sovrapposizione di effetti dei modi di funzionamento diretto e inverso, in funzione delle polarizzazioni V_{BE} e V_{BC} alle due giunzioni base-emettitore e base-collettore, che determinano con il loro segno il regime di funzionamento del transistore.

$$I_C = \alpha_F I_{EF} - I_{CR} = \alpha_F I_{ES}\left(exp\frac{V_{BE}}{V_T} - 1\right) - I_{CS}\left(exp\frac{V_{BC}}{V_T} - 1\right) \tag{6.12a}$$

$$I_E = I_{EF} - \alpha_R I_{CR} = I_{ES}\left(exp\frac{V_{BE}}{V_T} - 1\right) - \alpha_R I_{CS}\left(exp\frac{V_{BC}}{V_T} - 1\right) \tag{6.12b}$$

$$I_B = I_E - I_C \tag{6.12c}$$

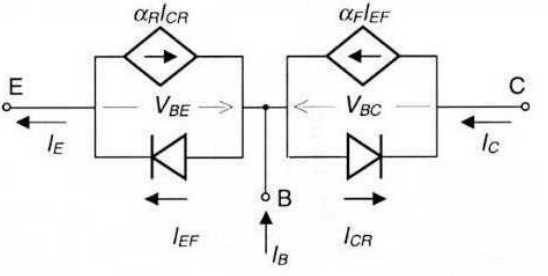


Figura 6.8 Modello circuitale di Ebers e Moll in regime stazionario

Questo modello è rappresentabile con un semplice circuito equivalente elettrico formato da diodi e generatori controllati di corrente, e per tale ragione è immediatamente implementabile in un simulatore circuitale come SPICE. Il circuito è riportato nella Figura 6.8, in cui per sinteticità si sono indicate con I_{CR}, I_{EF} le componenti di diodo delle correnti nelle giunzioni definite dai termini esponenziali delle (6.12). Per ogni terminale la corrente totale è somma di quella del diodo corrispondente e di quella di un generatore di corrente controllato dalla corrente che circola nel diodo dell'altra giunzione; a sua volta la corrente di ognuno dei diodi dipende dalla polarizzazione della giunzione corrispondente. Il modello è determinato da quattro parametri: I_{ES}, I_{CS}, α_F, α_R, che non sono però completamente indipendenti tra di loro; infatti si può dimostrare, anche in condizioni meno restrittive di quelle di drogaggio uniforme assunte per l'analisi, che vale un teorema di reciprocità per cui :

$$\alpha_F I_{ES} = \alpha_R I_{CS} \tag{6.13}$$

In base a questa relazione si giustifica l'assunzione di un valore piccolo per α_R; infatti, ricordando che α_F è prossima all'unità si ha per α_R:

$$\alpha_R = \alpha_F \frac{I_{ES}}{I_{CS}} \cong \frac{I_{ES}}{I_{CS}} \tag{6.14}$$

dove la corrente di saturazione inversa della giunzione di emettitore è inferiore a quella della giunzione di collettore, sia per la minore area della prima, che per i drogaggi più elevati che comportano una riduzione della densità di corrente (ricordiamo che queste sono legate ai portatori minoritari all'equilibrio per ciascuna regione, date dalle (6.2-6.4)).

Le caratteristiche *I-V* del transistore bipolare, in base a questo modello, sono quelle riportate (per il modo diretto di funzionamento) rispettivamente in Figura 6.9a per la caratteristica di ingresso (I_B-V_{BE}) e in Figura 6.9b per quella di uscita (I_C -V_{CE}).

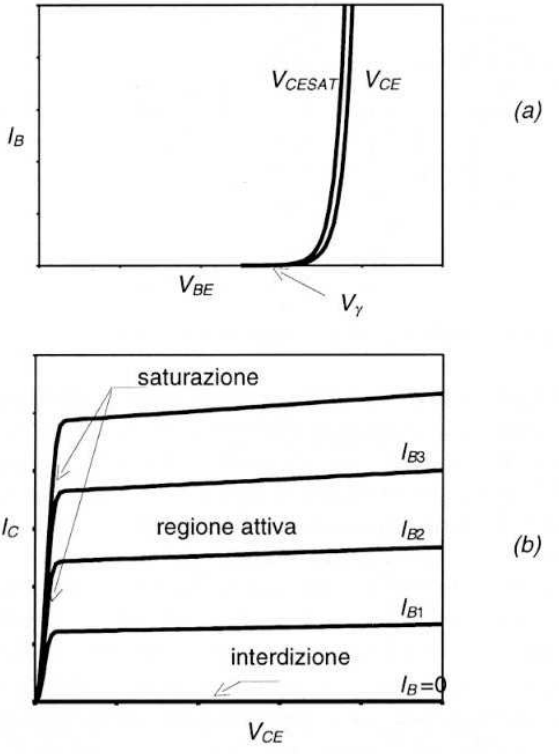


Figura 6.9 Caratteristiche di ingresso (a) e di uscita (b) di un transistore bipolare secondo il modello di Ebers e Moll

Nel piano delle caratteristiche di uscita si distinguono tre regioni:

a. la regione attiva, in cui la corrente di collettore è controllata dalla corrente di base;
b. la regione di saturazione, in cui tutte le caratteristiche di uscita convergono in una, indipendentemente dalla corrente di controllo, con una tensione di uscita molto bassa;

c. la regione di interdizione, in cui non circola corrente di uscita perché la corrente di base è nulla.

Regione attiva ***($V_{BE} > 0$, $V_{BC} < 0$)***

In questa regione, assumendo che V_{BE}, $|V_{BC}| >> V_T$, le correnti di saturazione inversa (dell'ordine di 10^{-15} A) sono trascurabili rispetto alla componente di diodo per la giunzione di emettitore, e quindi dalle (6.12) si ha:

$$I_C = \alpha_F I_{EF} + I_{CS} \equiv \alpha_F I_{EF} \equiv \alpha_F I_E \qquad [I_{CS}, \alpha_R I_{CS} << I_{EF}] \tag{6.15}$$

e ricordando la (6.12c) si ha:

$$I_C = \frac{\alpha_F}{1 - \alpha_F} I_B = \beta_F I_B \tag{6.16}$$

Dalla (6.16) si vede che nella regione attiva la corrente di collettore è proporzionale a quella di base secondo il fattore (amplificativo) β_F. Questo parametro, detto *guadagno di corrente*, è la grandezza più importante del transistore bipolare, analogamente alla tensione di soglia per i dispositivi MOS. La sua dipendenza complessa dai parametri fisici e tecnologici legati al processo impiegato per la sua realizzazione rende poco significativa una sua espressione analitica semplificata; per i dispositivi utilizzati per i circuiti digitali i valori tipici di β_F variano da 20 a 100. La relazione (6.16) ci permette subito di vedere la principale differenza tra il transistore bipolare e quello MOS: nel primo la corrente di uscita è controllata dalla *corrente* di ingresso, mentre per il MOS questa è controllata dalla *tensione* di ingresso.

Regione di saturazione ($V_{BE} > 0$, $V_{BC} > 0$)

In questo caso nelle (6.12) si trascura l'unità rispetto agli esponenziali in entrambi i termini di I_C e I_E, ottenendo:

$$I_C = \alpha_F I_{ES}\left(\exp\frac{V_{BE}}{V_T}\right) - I_{CS}\left(\exp\frac{V_{BC}}{V_T}\right) \tag{6.17}$$

$$I_E = I_{ES}\left(\exp\frac{V_{BE}}{V_T}\right) - \alpha_R I_{CS}\left(\exp\frac{V_{BC}}{V_T}\right) \tag{6.18}$$

La tensione V_{CE} in saturazione è la differenza tra le tensioni ai capi delle due giunzioni di emettitore e di collettore, per cui è inferiore ad ognuna delle due:

$$V_{CESAT} = V_{BESAT} - V_{BCSAT} \tag{6.19}$$

La tensione V_{BESAT} si ottiene dalle (6.17), (6.18) eliminando il secondo termine $I_{CS}\exp(V_{BC}/V_T)$ dalle due relazioni, ottenendo:

$$V_{BESAT} = V_T \ln \frac{I_B + I_C(1-\alpha_R)}{I_{ES}(1-\alpha_F\alpha_R)} \tag{6.20}$$

Analogamente, la V_{BCSAT} si ottiene dalle (6.17) e (6.18) eliminando il primo termine della (6.18) dalle due relazioni:

$$V_{BCSAT} = V_T \ln \frac{\alpha_F I_B - I_C(1-\alpha_F)}{I_{CS}(1-\alpha_F\alpha_R)} \tag{6.21}$$

Dalle (6.20) e (6.21) si ottiene la tensione V_{CE} come differenza tra le due:

$$V_{CESAT} = V_T \ln\left(\frac{I_B + I_C(1-\alpha_R)}{\alpha_F I_B - I_C(1-\alpha_F)} \frac{I_{CS}}{I_{ES}}\right) \tag{6.22}$$

da cui, ricordando la (6.13) e la definizione di β_F (β_R) si ha:

$$V_{CESAT} = V_T \ln \frac{\dfrac{1}{\alpha_R} + \dfrac{I_C}{I_B}\dfrac{1}{\beta_R}}{1 - \dfrac{I_C}{I_B}\dfrac{1}{\beta_F}} \tag{6.23}$$

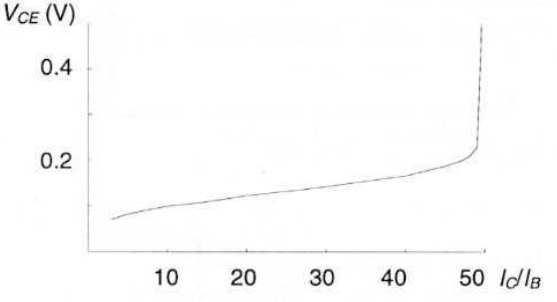


Figura 6.10 Dipendenza della tensione V_{CESAT} dal rapporto I_C/I_B, per $\beta_F = 50$, $\beta_R = 0.2$

Si nota dalla (6.23) che per $I_C = \beta_F I_B$ la tensione V_{CESAT} tende all'infinito; infatti questa è la condizione di uscita dalla saturazione e di ingresso nella regione attiva, che, nell'ipotesi di trascurare la pendenza finita delle caratteristiche *I-V* fornisce

una tensione indefinitamente grande a parità di corrente di collettore. In Figura 6.10 si riporta la variazione di V_{CESAT} con il rapporto I_C / I_B per un transistore con $\beta_F = 50$.

Regione di interdizione ($V_{BE} < 0$, $V_{BC} < 0$)

Nella regione di interdizione entrambe le giunzioni sono contropolarizzate per cui circolano solo le correnti inverse di saturazione. Dalle (6.12) si ha:

$$I_C = I_{CS} - \alpha_F I_{ES} \; ; \qquad I_E = \alpha_R I_{CS} - I_{ES} \tag{6.24}$$

quindi le correnti in gioco sono trascurabili rispetto a quelle del funzionamento normale.

Approssimazioni per i diversi regimi di funzionamento:

Le relazioni su esposte per le diverse regioni di funzionamento ottenute dal modello di Ebers e Moll, portano a significative semplificazioni per le grandezze elettriche in gioco, approssimazioni che possono essere utilmente impiegate per analisi approssimate del comportamento in regime stazionario di circuiti con transistori bipolari.

a) regione di interdizione

La verifica del funzionamento in questa regione può essere effettuata in base alle tensioni sulle due giunzioni: in pratica la condizione matematica (V_{BE}, $V_{BC} < -4V_T$) per trascurare nelle (6.12) l'esponenziale rispetto a 1, può essere notevolmente rilassata se si considera che il transistore è "interdetto" se le correnti nelle giunzioni sono trascurabili rispetto alle correnti nel funzionamento normale ($10^{-3} \div 10^{-2}$ A). Dall'equazione del diodo per ciascuna giunzione, si ricava che la tensione V_γ detta *tensione di soglia*, necessaria per fare circolare nelle giunzioni correnti 1000 volte più piccole delle correnti in regione attiva (dell'ordine della decina di mA), ossia pari a 10 μA, assumendo le correnti di saturazione dell'ordine di 10^{-15} A, vale:

$$V_{BE\gamma(BC\gamma)} \cong V_T \ln \frac{I_{E(C)}}{I_{ES(CS)}} = 0.026 \cdot \ln(10^{10}) = 0.598 \; V \tag{6.25}$$

Si considereranno quindi le giunzioni interdette nelle applicazioni pratiche se ai loro capi vi è una tensione minore della tensione di soglia pari a 0.6 V.

b) regione attiva

Nella regione attiva, come si è visto, vale la relazione (6.16) di proporzionalità tra corrente di base e di collettore; una ulteriore condizione può ricavarsi sulla tensione della giunzione di emettitore ricordando che in base alla (6.15) la corrente di collettore è espressa in funzione della tensione V_{BE} dall'espressione

del diodo attraverso la I_{EF}. Al variare quindi della I_C dall'interdizione ai valori massimi (decine di mA), utilizzando ancora la (6.25) si trova che la V_{BE} varierà da 0.6 V a 0.8 V. Un'approssimazione valida (con errore di ± 0.1 V) è quindi quella di assumere nella regione attiva una tensione intermedia $V_{BE} = 0.7$ V, all'incirca costante al variare della corrente di collettore. Questa approssimazione, che può sembrare a prima vista troppo drastica, permette notevolissime semplificazioni nelle analisi manuali dei circuiti pur conservando validità quantitativa nei risultati.

c) regione di saturazione

Il funzionamento in questa regione può essere individuato in più modi: 1) verificando che la tensione V_{CE} sia quella di saturazione (dai risultati della (6.23) si assume un valore massimo di $V_{CESAT} = 0.2$ V); 2) verificando la disequazione:

$$I_{CSAT} < \beta_F I_B \tag{6.26}$$

3) verificando che V_{BC} sia maggiore o uguale alla tensione di soglia V_γ. Quest'ultima condizione giustifica perché, in congruenza con l'assunzione $V_{CESAT} \cong 0.2$ V, si assume anche $V_{BESAT} = V_{CESAT} + V_{BC\gamma} \cong 0.8$ V.

Tabella 6.2 Approssimazioni utilizzate per il funzionamento statico di transistori bipolari

grandezza	*interdizione*	*regione attiva*	*saturazione*
V_{BE}	< 0.6 V	$\cong 0.7$ V	$\cong 0.8$ V
V_{BC}	< 0.6 V	< 0.6 V	≥ 0.6 V
V_{CE}	-	> 0.2 V	$\cong 0.2$ V
I_C	$\cong 0$	$\beta_F I_B$	$< \beta_F I_B$
I_B	$\cong 0$	I_C/β_F	$> I_C/\beta_F$

Dualmente, se si conosce il regime di funzionamento del transistore bipolare per altre vie, si possono assumere in via approssimata le relazioni precedentemente indicate per le grandezze elettriche più importanti. Queste approssimazioni semplificano notevolmente l'analisi del regime statico di funzionamento dei circuiti con transistori bipolari, in quanto riconducono l'analisi ad una rete composta da resistenze e generatori di tensione e/o corrente, questi ultimi controllati da un'altra corrente che è quella che fluisce nel terminale di ingresso del transistore in esame. Le relazioni suddette sono riportate nella Tabella 6.2; i parametri indicati fanno riferimento al modo di funzionamento diretto, in quanto nei pedici com-

pare il simbolo F, ma le relazioni sono ugualmente valide per il modo inverso, scambiando il pedice F con R.

6.5 Capacità della struttura e comportamento dinamico

Come in tutti i dispositivi, il comportamento dinamico è determinato in larga parte dai componenti reattivi equivalenti della struttura che debbono essere inseriti nei modelli utilizzati per le analisi in regime dinamico, ossia in transitorio. Questi, nel limite di approssimazione quasi-stazionaria per il funzionamento (analogamente a quanto detto per il transistore MOS, Paragrafo 3.5) sono assimilabili a capacità presenti tra i terminali del dispositivo.

Le capacità direttamente estraibili dalla struttura indicata in Figura 6.1 sono quelle di svuotamento delle giunzioni p/n, in particolare delle giunzioni di emettitore-base, base-collettore, e collettore-substrato; queste sono esprimibili dalle stesse relazioni (2.10), (2.11) presentate per le giunzioni P/N contropolarizzate, che qui si specificano:

$$C_{BE} = \frac{C_{JE0}}{\left(1 - V_{BE}/\phi_E\right)^{me}} \quad \text{con} \quad C_{JE0} = \sqrt{q\frac{\varepsilon_{SI} N_B}{2|\phi_E|}} \tag{6.27}$$

$$C_{BC} = \frac{C_{JC0}}{\left(1 - V_{BC}/\phi_C\right)^{mc}} \quad \text{con} \quad C_{JC0} = \sqrt{q\frac{\varepsilon_{SI} N_C}{2|\phi_C|}} \tag{6.28}$$

$$C_{CS} = \frac{C_{JS0}}{\left(1 - V_{CS}/\phi_S\right)^{ms}} \quad \text{con} \quad C_{JS0} = \sqrt{q\frac{\varepsilon_{SI} N_C}{2|\phi_S|}} \tag{6.29}$$

In queste relazioni, le capacità si intendono estese alle relative aree delle giunzioni; i coefficienti *me*, *mc*, *ms*, tengono conto del profilo del drogaggio nelle varie regioni (m = 1/2 per drogaggio a variazione brusca, m = 1/3 per variazione graduale), il parametro ϕ è ancora espresso dalla (2.8), specificando i valori dei drogaggi N_A (della regione P) e N_D (della regione N) per le singole giunzioni.

Tuttavia, in analogia con lo studio effettuato sul MOS, dove si è visto che queste capacità sono meno rilevanti di quella di gate legata alla struttura Metallo-Ossido-Semiconduttore, così anche per i transistori bipolari le capacità più importanti ai fini del comportamento dinamico sono quelle legate alle cariche nella regione di base. In questo caso però il problema della identificazione della capacità equivalente richiede qualche considerazione aggiuntiva sul ruolo delle cariche minoritarie iniettate nella regione di base (come anche nelle altre due regioni), in funzione delle polarizzazioni delle due giunzioni. Queste considerazioni portano ad interpretare il funzionamento del transistore nei due modi, attivo diretto ed attivo inverso, presentato nel Paragrafo 6.3, attraverso la *carica immagazzinata* nella

regione di base (e più in generale nelle diverse regioni); per tale ragione il modello che ne deriva, e che è quello maggiormente utilizzato per la descrizione del transistore in regime dinamico per grandi segnali, è detto *Modello a Controllo di Carica*. Si farà riferimento, per semplificare l'analisi, alla distribuzione delle cariche nella sola regione di base, analogamente a quanto esposto nella descrizione del modello di Ebers e Moll, correggendo in seguito i risultati per tenere conto delle cariche minoritarie nelle altre due regioni.

a) Carica iniettata dall'emettitore

La distribuzione di portatori minoritari iniettati dall'emettitore nella base, riportata in Figura 6.5, permette di definire l'eccesso di carica minoritaria Q_F accumulata nella regione di base in questo regime (Figura 6.11), data da:

$$Q_F = q\frac{A_E W_B n'_B(0)}{2} = q\frac{A_E W_B n_{B0}}{2}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \equiv Q_{F0}\left(\exp\frac{V_{BE}}{V_T} - 1\right) \quad (6.30)$$

Ricordando la espressione (6.8) per la corrente $I_{nE} \equiv \alpha_F I_{EF}$, e sostituendo in questa la (6.30) si ha:

$$I_{nE} = \frac{2D_B}{W_B^2} Q_F = \frac{Q_F}{\tau_F} \qquad \text{con} \qquad \tau_F = \frac{W_B^2}{2D_B} \quad (6.31)$$

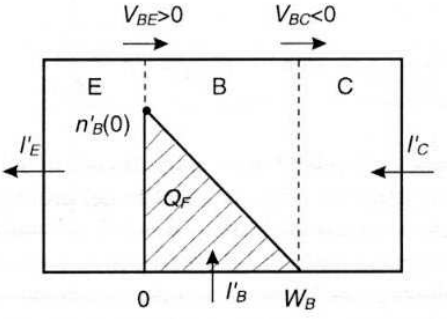


Figura 6.11 Carica iniettata nella base nel funzionamento in modo attivo diretto

La (6.31) mostra che è possibile esprimere la corrente I_{nE} come rapporto tra la carica Q_F iniettata nella base ed un tempo caratteristico τ_F detto *tempo di transito* nella base, in quanto esso corrisponde al tempo che mediamente impiegano gli elettroni ad attraversare la regione di base. Questa corrente, indicata in questo modello come I_C', rappresenta la componente di corrente di collettore dovuta all'iniezione di cariche dall'emettitore e corrisponde formalmente al primo termine della (6.12a), mostrando quindi, come si vedrà meglio in seguito,

l'equivalenza dei due modelli in regime statico. La carica Q_F nella base (costituita da portatori minoritari che tendono a ricombinarsi nel tempo τ_{BF}) non può conservarsi se non vengono fornite dalla corrente di base I_B le cariche (maggioritarie) ricombinatesi con le cariche minoritarie Q_F nel tempo τ_{BF}, per cui si può descrivere anche la corrente di base con una equazione legata alla carica Q_F e cioè:

$$I_B' = \frac{Q_F}{\tau_{BF}} \tag{6.32}$$

Questa descrizione delle correnti in funzione della carica accumulata nella base può facilmente estendersi al caso dinamico, considerando che in un transitorio in cui, al variare della tensione sulla giunzione di emettitore, varia la carica accumulata $Q_F(t)$, occorrerà fornire una corrente di base (positiva o negativa a seconda che Q_F si incrementa o diminuisce) che bilanci questa variazione nel tempo. Quindi le relazioni (6.31) e (6.32) possono essere estese al caso di grandezze variabili nel tempo secondo le relazioni seguenti:

$$i_C'(t) = \frac{Q_F(t)}{\tau_F} \tag{6.33a}$$

$$i_B'(t) = \frac{Q_F(t)}{\tau_{BF}} + \frac{dQ_F(t)}{dt} \tag{6.33b}$$

$$i_E'(t) = i_C'(t) + i_B'(t) \tag{6.33c}$$

dove l'apice indica le componenti delle correnti dovute all'iniezione dall'emettitore (modo attivo diretto). Nell'estendere la (6.31) al regime dinamico si è assunta implicitamente l'ipotesi di regime quasi-stazionario introdotto nel Paragrafo 3.5 per il MOS, cioè si assume che la ridistribuzione delle cariche mobili nell'evoluzione dinamica avvenga in tempi trascurabili rispetto a quello di variazione del segnale, (in questo caso la tensione sulla giunzione di emettitore); questo significa che la tensione V_{BE} deve variare in tempi più lunghi del tempo di transito nella base τ_F.

b) Carica iniettata dal collettore

Dalla Figura 6.12 si possono ottenere le relazioni corrispondenti per le componenti dovute all'iniezione di cariche dal collettore (modo attivo inverso), ricordando che ora la carica Q_R è data da:

$$Q_R = q\frac{A_C W_B n_B'(W_B)}{2} = q\frac{A_C W_B n_{B0}}{2}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \equiv Q_{R0}\left(\exp\frac{V_{BC}}{V_T} - 1\right) \tag{6.34}$$

e che questa carica può essere utilizzata per esprimere la corrente I_E'', formalmente corrispondente al secondo termine della (6.12b), secondo la relazione:

$$I_E'' = -q\frac{A_C D_n n_B'(W_B)}{W_B} = -\frac{Q_R}{\tau_R} \tag{6.35}$$

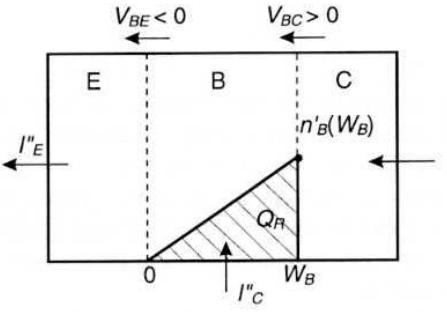


Figura 6.12 Carica iniettata nella base nel funzionamento in modo attivo inverso

Va osservato che in questa relazione valida per il regime inverso, se ci si limitasse alla sola carica di base e ad una struttura unidimensionale, si avrebbe $\tau_R = \tau_F$ (e anche $\tau_{BR} = \tau_{BF}$); questo porterebbe ad avere, nell'equivalente stazionario di questo modello, $\alpha_R = \alpha_F$. La diversificazione dei parametri τ nei due modi di funzionamento tiene quindi conto di un'analisi più dettagliata per il Modello a Controllo di Carica, che considera anche le cariche accumulate nelle altre due regioni di emettitore e di collettore, cariche che nella trattazione semplificata su esposta sono state trascurate. Queste cariche ulteriori, dovute ai contributi di iniezione rispettivamente nell'emettitore e nel collettore (diversi per le due giunzioni a causa dei differenti drogaggi e aree), possono essere aggiunte rispettivamente alle cariche Q_F e Q_R e costituiscono ulteriori componenti di corrente che vanno ad aggiungersi alle correnti nei tre terminali; si può dimostrare che il loro effetto può essere tenuto in conto modificando opportunamente i coefficienti τ corrispondenti, che quindi perdono il significato fisico introdotto precedentemente.

Con riferimento ai versi delle componenti di corrente indicate in Figura 6.12 per il modo attivo inverso, le relazioni corrispondenti in regime dinamico sono equivalenti a quelle per il modo attivo diretto, pur di scambiare il pedice F con quello R e il terminale di emettitore con quello di collettore:

$$i_E''(t) = -\frac{Q_R(t)}{\tau_R} \tag{6.36a}$$

$$i''_B(t) = \frac{Q_R(t)}{\tau_{BR}} + \frac{dQ_R(t)}{dt} \tag{6.36b}$$

$$i''_C(t) = i''_E(t) - i''_B(t) \tag{6.36c}$$

6.6 Modello a Controllo di Carica per il comportamento dinamico

Sommando termine a termine le (6.33) e (6.36) per i due regimi, si ottengono le relazioni per le tre correnti, valide per entrambi i modi di funzionamento, e quindi anche per il regime di saturazione, che si è visto essere la combinazione dei due modi di funzionamento:

$$i_C(t) = \frac{Q_F(t)}{\tau_F} - Q_R(t)\left(\frac{1}{\tau_R} + \frac{1}{\tau_{BR}}\right) - \frac{dQ_R(t)}{dt} \tag{6.37a}$$

$$i_B(t) = \frac{Q_F(t)}{\tau_{BF}} + \frac{Q_R(t)}{\tau_{BR}} + \frac{dQ_F(t)}{dt} + \frac{dQ_R(t)}{dt} \tag{6.37b}$$

$$i_E(t) = -\frac{Q_R(t)}{\tau_R} + Q_F(t)\left(\frac{1}{\tau_F} + \frac{1}{\tau_{BF}}\right) + \frac{dQ_F(t)}{dt} \tag{6.37c}$$

Queste equazioni costituiscono la base del Modello a Controllo di Carica del transistore bipolare, detto anche modello di Gummel e Poon, che permette di descrivere il comportamento del transistore bipolare oltre che in regime stazionario, anche in regime dinamico e nel campo degli ampi segnali (modello nonlineare dinamico).

Questo modello è utilizzato nella maggior parte dei simulatori circuitali, come lo SPICE, perché, analogamente al modello di Ebers e Moll, permette di sviluppare un'interpretazione circuitale delle equazioni costituenti, in termini di componenti come diodi, capacità e generatori pilotati. Per comprendere l'interpretazione circuitale dei diversi termini, riscriviamo le espressioni (6.37a) e (6.37c) delle correnti di collettore e di emettitore (quella di base può essere ottenuta dalle altre due), esprimendo le dipendenze funzionali dei termini Q dalle tensioni sulle giunzioni:

$$i_C(t) = \frac{Q_F(V_{BE})}{\tau_F} - \frac{Q_R(V_{BC})}{\tau_R^*} - \frac{dQ_R(V_{BC})}{dt} \tag{6.38a}$$

$$i_E(t) = -\frac{Q_R(V_{BC})}{\tau_R} + \frac{Q_F(V_{BE})}{\tau_F^*} + \frac{dQ_F(V_{BE})}{dt} \tag{6.38b}$$

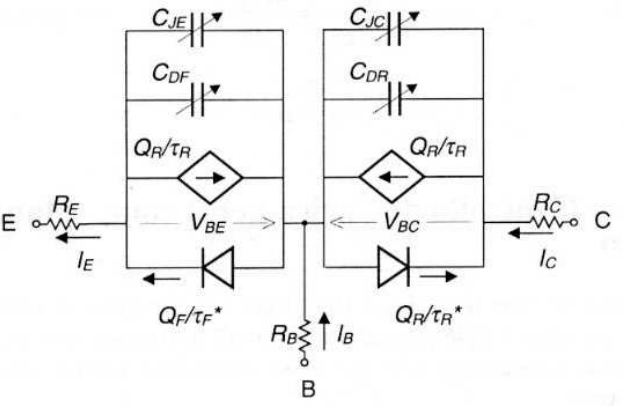


Figura 6.13 Circuito equivalente del Modello a Controllo di Carica

Il termine Q_R/τ_R^* dipende dalla tensione ai capi della giunzione di collettore in modo esponenziale secondo la (6.34) quindi è rappresentato da un diodo posto tra collettore e base; analogamente Q_F/τ_F^* in base alla (6.30) è rappresentato da un diodo posto tra emettitore e base. I termini Q_F/τ_F e Q_R/τ_R sono stati già identificati come formalmente uguali rispettivamente al primo termine della (6.12a) e al secondo della (6.12b) e sono quindi rappresentati anche in questo circuito come generatori di corrente controllati dalle correnti dei diodi corrispondenti. Rimangono i termini legati alle variazioni temporali di $Q_F(Q_R)$; questi sono descritti da capacità nonlineari C_{DF} (C_{DR}) in base a:

$$\frac{dQ_F}{dt} = \frac{dQ_F}{dV_{BE}}\frac{dV_{BE}}{dt} = \frac{Q_{F0}}{V_T}\exp\frac{V_{BE}}{V_T}\cdot\frac{dV_{BE}}{dt} = C_{DF}(V_{BE})\frac{dV_{BE}}{dt} \qquad (6.39)$$

Le capacità C_{DF} e C_{DR} vengono definite *capacità di diffusione*, per distinguerle dalle capacità di svuotamento delle giunzioni, definite dalle (6.27-6.29), e dipendono dalle cariche minoritarie iniettate in condizione di polarizzazione diretta della giunzione; per la loro dipendenza esponenziale dalla tensione di giunzione, nel regime di funzionamento attivo esse assumono valori ben maggiori delle capacità di svuotamento. In definitiva, aggiungendo anche le capacità di svuotamento delle giunzioni al modello presentato, si può descrivere quest'ultimo mediante il circuito equivalente riportato in Figura 6.13 (in questo sono state anche inserite le resistenze parassite R_C, R_E, R_B dei rispettivi terminali; in particolare la R_B, che tiene conto della resistenza distribuita dello strato di base, non è trascurabile nei dispositivi effettivi, e la R_E, che, anche se piccola, esercita una reazione non trascurabile sulla tensione d'ingresso).

I parametri che determinano il circuito equivalente sono:

$$Q_{F0} \quad Q_{R0} \quad \tau_F \quad \tau_R \quad \tau_{BF} \quad \tau_{BR} \quad C_{JC0} \quad C_{JE0}$$

Dalle Equazioni (6.33) per il regime diretto e (6.36) per quello inverso, specializzate per il caso stazionario ($d/dt = 0$), si ricava un legame tra i coefficienti β e i τ:

$$\frac{I_C'}{I_B'} \equiv \beta_F = \frac{\tau_{BF}}{\tau_F}; \qquad -\frac{I_E''}{I_B''} \equiv \beta_R = \frac{\tau_{BR}}{\tau_R} \tag{6.40}$$

inoltre, in base al teorema di reciprocità che porta alla (6.13) si trova un analogo legame tra le correnti inverse delle due giunzioni:

$$I_S = \frac{Q_{F0}}{\tau_F} = \frac{Q_{R0}}{\tau_R} \tag{6.41}$$

Quindi, in definitiva i parametri indipendenti che occorre specificare nel simulatore SPICE (aggiungendo anche la capacità di substrato non indicata in Figura 6.12) sono:

$$I_S = \frac{Q_{F0}}{\tau_F} = \frac{Q_{R0}}{\tau_R} \quad \tau_F \quad \tau_R \quad \beta_F = \frac{\tau_{BF}}{\tau_F} \quad \beta_R = \frac{\tau_{BR}}{\tau_R} \quad C_{JC0} \quad C_{JE0} \quad (C_{JS})$$

Questo modello verrà impiegato nelle simulazioni SPICE per l'analisi dei circuiti con dispositivi bipolari; verranno tuttavia utilizzate direttamente le equazioni del Modello a Controllo di Carica nell'analisi di porte elementari con transistori bipolari, per valutare in via analitica approssimata il comportamento dinamico delle porte stesse.

6.7 Miglioramenti tecnologici dei transistori bipolari

Anche i transistori bipolari, come i dispositivi MOS, hanno potuto beneficiare della evoluzione dei processi di fabbricazione e di nuove tecnologie microelettroniche, che hanno portato ad una riduzione di scala sulle dimensioni, e reso possibili significativi miglioramenti nelle prestazioni, in particolare per quanto riguarda quelle dinamiche.

Ricordiamo che il transistore bipolare è intrinsecamente un dispositivo verticale, per cui le principali caratteristiche dipendono dalle dimensioni verticali della struttura. Lo scalamento delle dimensioni verticali coinvolge quindi gli spessori delle tre regioni, e quello dello strato epitassiale di partenza. Per quest'ultimo, utilizzando nuove tecniche di crescita epitassiale a bassa temperatura, si è potuto ridurre lo spessore da circa 6 μm a meno di 2 μm, con un corrispondente aumento del drogaggio dello strato da circa 10^{15} cm^{-3} a più di 10^{16} cm^{-3}; ciò comporta una diminuzione della resistenza estrinseca di collettore, e la riduzione nella massima tensione V_{BR} che ne consegue è compatibile con la riduzione della tensione di ali-

mentazione. Lo spessore della regione di base si è ridotto fino a valori di circa 0.1 μm con l'impiego dell'impiantazione e di un annealing termico rapido, il che riduce con legge quadratica il valore del parametro τ_F definito dalla (6.31) come:

$$\tau_F \equiv \frac{Q_F}{I_C} = \frac{W_B^2}{2\mu_n V_T} \tag{6.42}$$

La diminuzione delle dimensioni verticali è accompagnata da un'analoga riduzione di quelle laterali, quest'ultima resa possibile dalla riduzione dello spessore dello strato epitassiale di partenza. Una prima riduzione dell'area della struttura elementare (Figura 6.1) del transistore bipolare è legata alla possibilità di effettuare *ossidazioni selettive* della superficie del silicio. Questa tecnologia, indicata con il termine LOCOS (*Local Oxidation of Silicon*) si basa sull'impiego di uno strato di nitruro di silicio (Si_3N_4), deposto preventivamente sul silicio e delineato con una operazione fotolitografica (vedi Figura 6.14). Il nitruro impedisce la crescita dell'ossido nelle aree su cui è depositato, e può essere successivamente eliminato mediante un attacco selettivo che non agisce sull'ossido; la crescita termica dell'ossido di silicio, che utilizza parte del silicio della fetta, penetra anche nel silicio stesso, secondo un fattore che è il 45% dello spessore totale di ossido cresciuto. Questo processo può quindi essere utilizzato per isolare lateralmente i singoli transistori realizzati nello strato epitassiale, mediante un anello di ossido al posto della giunzione di isolamento. La profondità dell'isolamento può essere aumentata attraverso un attacco selettivo del silicio prima dell'ossidazione, in modo da abbassare il livello della superficie nelle aree in cui sarà cresciuto l'ossido, e quindi permettere una maggiore profondità dell'ossido di isolamento.

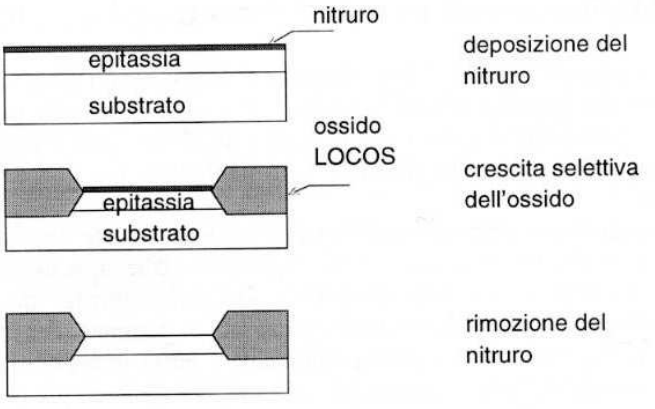


Figura 6.14 Realizzazione di ossidazione selettiva di silicio mediante mascheratura con nitruro

Questa tecnica è stata ulteriormente migliorata mediante impiego degli attacchi di silicio mediante gas ionizzati (*plasma etching*) invece che mediante le usuali soluzioni liquide; questi attacchi, detti attacchi a secco (*dry etching*), forniscono un'elevata selettività ed anisotropia (velocità di attacco verticale molto maggiore di quella orizzontale), e quindi permettono lo scavo di trincee relativamente profonde nel silicio, che vengono poi riempite di ossido e polisilicio depositati da fase vapore, in modo da isolare i singoli dispositivi con un ridottissimo consumo di area attiva.

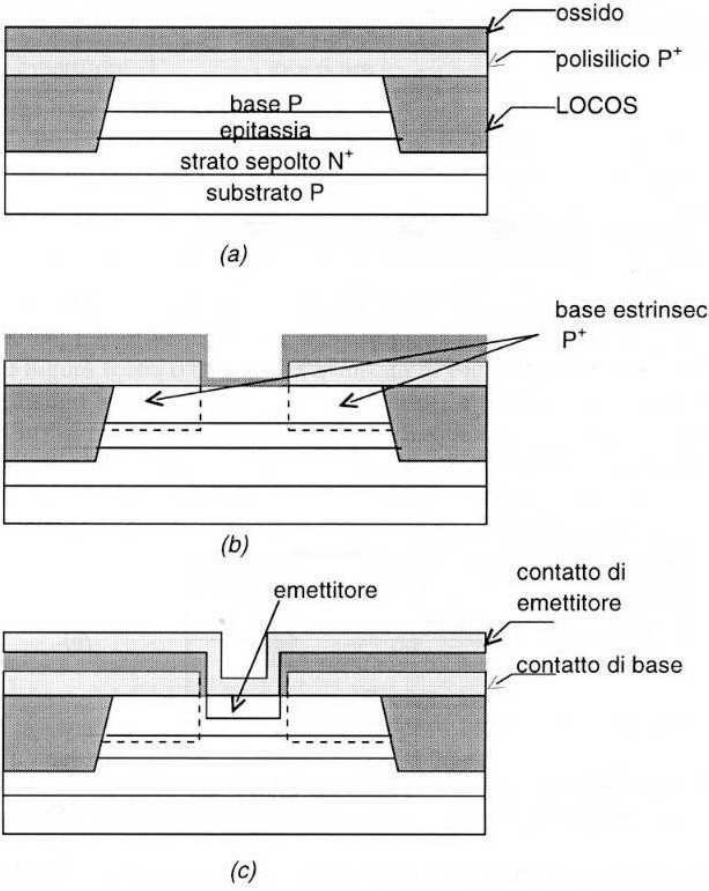


Figura 6.15 Fabbricazione di un transistore bipolare mediante polisilicio con processo autoallineato

Un'ulteriore riduzione delle dimensioni laterali della struttura, essenziale per ridurre le capacità delle giunzioni, è stata resa possibile dall'utilizzazione di *processi autoallineati* per la formazione dei contatti di base e di emettitore. Questi processi si basano sull'uso del polisilicio come strato di contatto, sia per le regioni

di base che di emettitore, oltre che sull'uso delle trincee di isolamento, e sono esemplificati in Figura 6.15. La realizzazione del contatto di base viene in questo caso effettuata attraverso la deposizione di uno strato di polisilicio drogato P^+ su tutta l'area dell'impiantazione di base, su cui viene depositato uno strato di ossido di silicio (Figura 6.15a). Una successiva mascheratura apre con un'unica operazione fotolitografica l'area in cui verrà realizzato l'emettitore e il suo contatto, e in questa viene cresciuto termicamente uno strato sottile di ossido, che viene rimosso dalla superficie del silicio con un attacco direzionale a secco (Figura 6.15b); il polisilicio P^+ crea una regione più drogata sotto il contatto di base che riduce la resistenza di base estrinseca. Infine viene depositato un secondo strato di polisilicio che viene impiantato N^+ e che agisce da sorgente di drogante per realizzare l'emettitore nel silicio (Figura 6.15c).

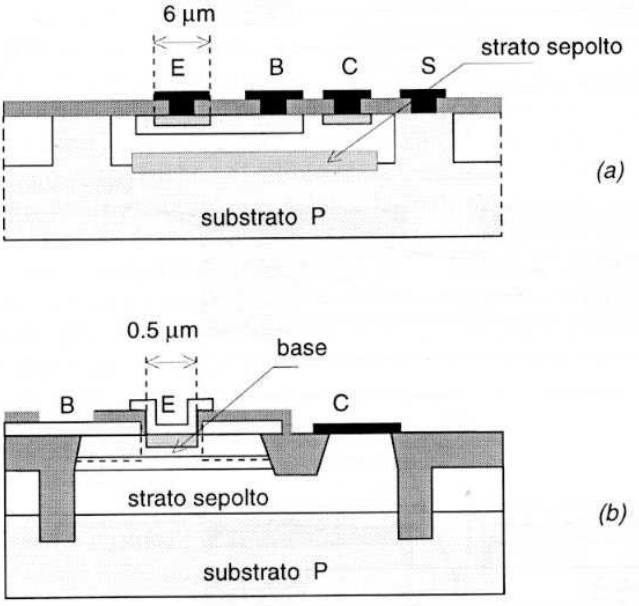


Figura 6.16 Confronto tra: a) la struttura di un transistore con isolamento a giunzione e b) un transistore autoallineato con isolamento a trincea (non in scala)

In questo modo l'emettitore è autoallineato con il suo contatto (realizzato dallo stesso strato di polisilicio utilizzato per la sua formazione) e distanziato dal contatto di base di uno spessore minimo pari all'ossido cresciuto lateralmente sul polisilicio P^+. La larghezza dell'emettitore può essere ridotta ad una dimensione pari alla minima dimensione fotolitografica λ, e la distanza tra i contatti di base e di emettitore è in pratica nulla; anche l'area della giunzione base-collettore viene notevolmente ridotta e così anche la capacità associata, e l'area di contatto del collettore, separata dall'ossido di separazione, non contribuisce ad allargare l'area attiva del dispositi-

vo, essendo in contatto solo con lo strato sepolto. In Figura 6.16 viene paragonata la struttura di un transistore tradizionale con quella di un processo autoallineato e con isolamento a trincea (*Trench isolation*), e risulta evidente la riduzione dell'area di quest'ultima.

Tabella 6.3 Parametri per transistori bipolari tradizionali e autoallineati

parametro	*transistore tradizionale*	*transistore autoallineato*
τ_F	100 ps	6 ps
τ_R	10 ns	5 ns
area emettitore	6 × 6 μm	0.5 × 1 μm
area base	10 × 14 μm	1.5 × 2 μm
C_{JE}	500 fF	4 fF
C_{JC}	500 fF	2 fF
C_{JS}	1000 fF	3 fF

In Tabella 6.3 sono riportati i valori di alcuni dei parametri del modello a controllo di carica, rispettivamente per transistori bipolari realizzati con la tecnologia riportata in Figura 6.16a o con quella avanzata di Figura 6.16b; con questi parametri si giustificano tempi di propagazione inferiori a 0.1 ns, che permettono, come vedremo, l'utilizzo di circuiti logici bipolari a frequenze ben superiori al GHz.

Può essere interessante fare un confronto degli effetti determinati dalla riduzione di scala sui dispositivi MOS e quelli bipolari, per quanto riguarda i limiti delle prestazioni dinamiche. Per entrambi i dispositivi i tempi di propagazione possono essere sinteticamente descritti da una relazione del tipo:

$$t_P = \frac{C_T}{I}\Delta V \tag{6.43}$$

dove C_T è la capacità totale di carico; questa per i MOS è somma delle capacità di giunzione C_J del dispositivo, di quelle C_L della linea, e di quella C_G di gate dello stadio successivo, mentre per i bipolari è somma delle due prima elencate più la capacità di ingresso associata alla carica immagazzinata $Q_F = \tau_F I_C$. Quindi la (6.43) si specializza per i due casi in:

$$t_P(MOS) = \frac{C_J + C_l}{I_D}\Delta V + \frac{C_G}{I_D}\Delta V \tag{6.44a}$$

$$t_P(Bipolare) = \frac{C_J + C_l}{I_C}\Delta V + \tau_F \tag{6.44b}$$

Con la progressiva riduzione delle dimensioni minime delle strutture, il primo termine delle (6.44) tende ad essere trascurabile rispetto al secondo; questo effetto è però più accentuato nei transistori bipolari nei quali la corrente dipende esponenzialmente dalla tensione applicata all'ingresso e quindi è più grande di quella fornita dai MOS a parità di area. Le espressioni dei secondi termini sono molto simili formalmente, perchè valgono:

$$\frac{C_G}{I_D}\Delta V \cong \frac{2L^2}{\mu_n(V_{DD}-V_T)} \qquad \text{(MOS)}$$

$$\tau_F \cong \frac{W_B^2}{4D_n} \qquad \text{(Bipolari)}$$

e dipendono entrambi in maniera quadratica da una dimensione minima; questa però nei transistori bipolari è lo spessore di base, che non dipende da un processo fotolitografico e può raggiungere dimensioni inferiori a 0.1 μm, mentre questi valori sono notevolmente più difficili da realizzare per la lunghezza di canale L nei MOS. Questo giustifica perché ancora oggi i transistori bipolari siano più veloci dei dispositivi MOS a parità di dimensioni minime realizzabili.

Esercizi di riepilogo

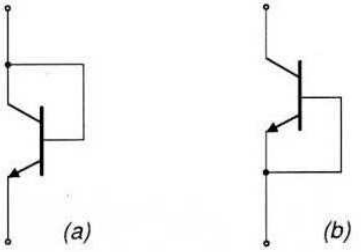
6.1 Per una giunzione base-emettitore di un transistore NPN, valutare, utilizzando l'espressione riportata nella (6.9) ed assumendo una corrente inversa della giunzione $I_{ES} = 10^{-12}$ A, le correnti I_E corrispondenti a tensioni V_{BE} di valore a) 0.6 V, e b) 0.8 V.

6.2 Determinare il valore della corrente inversa I_{CS} di un transistore che ha un valore $I_{ES} = 2{\cdot}10^{-15}$ A, e con $\alpha_F = 0.98$, $\alpha_R = 0.2$.

6.3 Determinare i regimi di funzionamento di un transistore bipolare che presenta i seguenti valori dei parametri: $I_{ES} = 10^{-15}$ A, $I_{CS} = 10^{-16}$ A, $\beta_F = 50$, $\beta_R = 0.5$, per le seguenti condizioni elettriche ai terminali: a) $I_C = 1$ mA, $I_E = 1.04$ mA; b) $V_{BE} = 0.7$ V, $V_{BC} = 0.3$ V; c) $V_{BC} = 0.8$ V, $V_{EC} = 0.8$ V.

6.4 Tracciare le caratteristiche I-V a temperatura ambiente, in regime sia diretto che inverso, per un transistore che presenta i seguenti parametri: $I_{ES} = 10^{-15}$ A, $I_{CS} = 10^{-16}$ A, $\alpha_F = 98$, $\alpha_R = 0.1$. Valutare il valore del guadagno di corrente nei due regimi, diretto e inverso.

6.5 Determinare le espressioni analitiche delle caratteristiche I-V del transistore connesso come diodo, nei due casi indicati in Figura E6.1, e valutarne i valori per i parametri: $I_{ES} = 10^{-15}$ A, $I_{CS} = 10^{-16}$ A, $\alpha_F = 98$, $\alpha_R = 0.1$.



(a) (b) **Figura E6.1**

6.6 Utilizzando il simulatore SPICE, ricavare l'andamento delle caratteristiche ideali in regime diretto, nel campo di valori V_{CE} 0÷5 V, e I_C 0÷20 mA, per un transistore NPN che presenta i seguenti parametri: $I_S = 10^{-15}$ A, $\beta_F = 50$, $\beta_R = 1$, $Vaf = 50$, $Var = 50$, $C_{JE} = 0.1$pF, $C_{JC} = 0.05$pF, $\tau_F = 0.06$ ns, $\tau_R = 5$ ns. Valutare inoltre le modifiche sulle caratteristiche introdotte dalle resistenze parassite $R_B = 50\ \Omega$, $R_C = 5\ \Omega$, $R_E = 1\ \Omega$. Spiegare perché le modifiche maggiori si ritrovano nella regione di saturazione.

6.7 Utilizzando i due circuiti equivalenti dei modelli di Ebers e Moll e di Gummel e Poon, riportati rispettivamente in Figura 6.8 e 6.13, e le rispettive rappresentazioni analitiche, ricavare un'equivalenza tra i parametri dei due modelli, limitatamente al comportamento statico (si considerino a tal fine trascurabili i componenti capacitivi).

Riferimenti bibliografici

R.S. Muller, T.I. Kamins, *Dispositivi elettronici nei circuiti integrati*, Bollati Boringhieri, Torino, 1982.

G. Soncini, *Tecnologie Microelettroniche*, Bollati Boringhieri, Torino, 1986.

P. Antognetti, G. Massobrio, *Semiconductor Device Modeling with SPICE*, McGraw-Hill, New York, 1987.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

7

Invertitori elementari con BJT

7.1 Introduzione

Le logiche bipolari comprendono tutti i circuiti in cui l'elemento di controllo dell'invertitore elementare è realizzato con transistori bipolari. In questo capitolo verranno introdotti i più semplici schemi di invertitori bipolari, che hanno dato origine alle prime famiglie logiche, quali la Logica Resistore-Transistore (RTL) e la Logica Diodo-Transistore (DTL). Queste logiche, ormai superate nelle logiche standard da altre famiglie logiche, verranno presentate essenzialmente per introdurre le differenze fondamentali tra le logiche bipolari e quelle MOS, differenze che sono visibili già nell'esame di queste prime realizzazioni circuitali.

L'analisi di questi circuiti ci permetterà da una parte di comprendere come i problemi riscontrati in queste prime famiglie logiche siano stati risolti in quelle successive; lo studio di questi semplici invertitori, come quello RTL, ci permetterà inoltre di introdurre modalità di analisi estremamente semplificate per i circuiti con componenti bipolari, in base alle considerazioni effettuate nel capitolo precedente, analisi che verranno utilizzate anche per le logiche bipolari successive.

In particolare si farà riferimento nelle analisi statiche alle approssimazioni indicate in Tabella 6.2 per le grandezze dei transistori in base al modello di Ebers-Moll (in particolare nella determinazione della funzione di trasferimento, potenza dissipata, margini di rumore), che permetteranno un'agevole valutazione di prima approssimazione delle condizioni di funzionamento dei transistori e una soluzione molto rapida delle analisi statiche. Per le analisi dinamiche si utilizzeranno le relazioni fornite dal modello a Controllo di Carica in forma semplificata, che ci permetteranno di valutare, seppur in via approssimata, i principali parametri del comportamento dinamico, come i tempi di propagazione e i ritardi. L'utilizzo di queste equazioni ci permetterà inoltre di valutare il ruolo essenziale giocato dalla carica minoritaria accumulata nella base, e di definire i tempi caratteristici di commutazione in funzione dei parametri del dispositivo.

7.2 L'invertitore RTL

Il più semplice schema utilizzabile per realizzare un invertitore elementare in cui l'interruttore pilotato è costituito da un transistore bipolare è quello di Figura 7.1. Esso consiste in un transistore caricato da una resistenza R_C, e pilotato dal segnale logico di ingresso tramite una ulteriore resistenza R_B in serie alla base. Questa resistenza di base è necessaria per limitare la corrente di ingresso nello stato alto; infatti se si applicasse, nello stato logico alto, la tensione V_{OH} (circa 5 V) direttamente alla base, circolerebbe una corrente così elevata, essendo la tensione di saturazione V_{BEsat} circa 0.8 V, da distruggere il transistore stesso. La resistenza R_B in effetti trasforma il generatore di tensione in ingresso in uno equivalente di corrente visto dal transistore, che permette il controllo del transistore bipolare attraverso la corrente di base invece che attraverso la tensione.

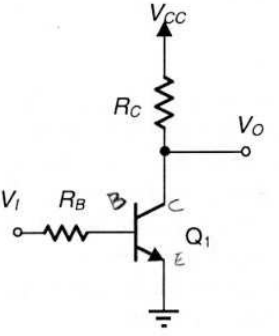


Figura 7.1 Schema circuitale dell'invertitore RTL

La presenza di una corrente nella maglia di ingresso per il transistore bipolare complica notevolmente l'analisi rispetto al caso del transistore MOS, per il quale la corrente di ingresso è nulla. Le condizioni statiche di funzionamento in presenza di segnale all'ingresso possono essere ottenute in via generale utilizzando le equazioni del modello di Ebers-Moll nelle reti della maglia di ingresso e di uscita; questo però non è agevole per un'analisi manuale del circuito, a causa delle descrizioni delle correnti in funzione dell'esponenziale delle tensioni sulle giunzioni. È invece possibile (e porta a risultati sufficientemente accurati) un'analisi semplificata che utilizzi le approssimazioni riportate in Tabella 6.2, *purché si conosca il regime di funzionamento* del dispositivo. Quest'ultimo è tuttavia facilmente desumibile nella maggior parte dei casi, e può essere sempre individuato con semplici dimostrazioni per assurdo, ossia negando la tesi e verificando l'impossibilità dell'ipotesi.

Nel caso dell'invertitore RTL, un'analisi grafica simile a quella utilizzata nel Paragrafo 4.2 per l'invertitore MOS, utilizzando la retta di carico definita dalla resistenza R_C nel piano delle caratteristiche di uscita (vedi Figura 7.2a), porta a

definire due possibili punti di funzionamento: A (interdizione) e B (saturazione), rispettivamente per ingresso basso (uscita alta) e ingresso alto (uscita bassa).

La condizione A si realizza quando la tensione di ingresso soddisfa all'ovvia relazione:

$$V_I \leq V_{BE\gamma} \tag{7.1}$$

Questo comporta che la tensione di uscita sarà $V_{OH} = V_{CC}$, in quanto il transistore è interdetto e $I_C = 0$; quindi la tensione di ingresso nello stato alto (che è anche la tensione di uscita di uno stadio precedente) sarà $V_H = V_{CC}$.

La condizione B (saturazione) è identificata da una disequazione (Equazione 6.26) tra le correnti I_C e I_B; queste ultime possono però essere relazionate alle tensioni di ingresso e di uscita attraverso l'analisi grafica con la retta di carico determinata dalla R_B nella maglia di ingresso (Figura 7.2b). Infatti, assegnata la tensione alta di ingresso V_H, dalla Figura 7.2b si estrae la corrente I_{BMAX} che circola, e quindi dalla Figura 7.2a si identifica il valore I_{CMAX} con cui verificare la (6.26).

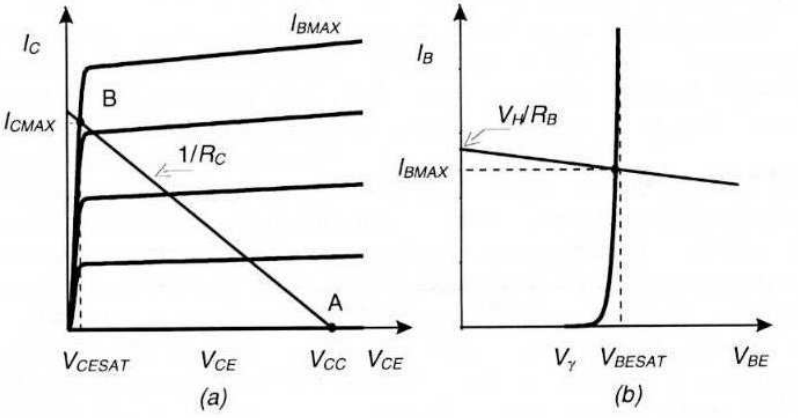


Figura 7.2 Analisi grafica per l'invertitore; a) maglia di uscita; b) maglia di ingresso

Questa procedura può essere semplificata ulteriormente utilizzando, come si è detto, le approssimazioni legate ai diversi modi di funzionamento. Supponendo che nel punto B il transistore si trovi in saturazione, si può scrivere per la corrente di collettore I_{CMAX}:

$$I_{CMAX} \equiv I_{CSAT} = \frac{V_{CC} - V_{CESAT}}{R_C} \cong \frac{V_{CC}}{R_C} \tag{7.2}$$

dove si è trascurato $V_{CESAT} \cong 0.2V$ rispetto a V_{CC}. Analogamente si ha per la corrente di base:

$$I_{BMAX} = \frac{V_{OH} - V_{BESAT}}{R_B} = \frac{V_{CC} - V_{BESAT}}{R_B} \cong \frac{V_{CC}}{R_B} \tag{7.3}$$

dove si può trascurare $V_{BESAT} \cong 0.8V$ rispetto a V_{CC}. Quindi la disequazione (6.26) valida per la regione di saturazione comporta in questo caso:

$$I_{CMAX} < \beta_F I_{BMAX} \quad \Rightarrow \quad \frac{R_B}{\beta_F R_C} < 1 \tag{7.4}$$

Questa relazione pone un vincolo tra le resistenze di base e di collettore ed è fondamentale per la valutazione del grado di saturazione del dispositivo; quanto più piccolo dell'unità è il primo termine della disequazione, tanto maggiore è il forzamento in saturazione del transistore. Nel caso dell'analisi approssimata la verifica di questa disequazione per l'invertitore di Figura 7.1 permette anche di verificare l'ipotesi di saturazione alla base delle (7.2) e (7.3), e quindi di confermare le approssimazioni assunte in queste ultime.

A questo punto può essere utile sottolineare una significativa diversità tra le porte con tecnologia MOS e quelle con dispositivi bipolari, che si può riscontrare già nell'invertitore RTL e che si ritroverà nelle porte bipolari sviluppate successivamente. Negli invertitori MOS si è riscontrata la maggior convenienza ad utilizzare, per i dispositivi, dei carichi attivi al posto delle resistenze, perché queste ultime avrebbero dovuto presentare valori molto elevati e quindi non facilmente integrabili per ridurre i valori della tensione di uscita nello stato basso V_{OL} a valori accettabilmente piccoli (vedi Paragrafo 4.2). Questa limitazione viene a cadere per gli invertitori con transistori bipolari, essenzialmente a causa della caratteristica di saturazione di questi ultimi, che non presenta un andamento lineare con la corrente come nei MOS, ma una tensione V_{CESAT} all'incirca costante al variare della corrente di collettore, fino al valore I_{CMAX} definito dalla (7.2). Ciò implica che non è necessario utilizzare come carico una resistenza di valore molto elevato, in quanto basta soddisfare alla disequazione (7.4) tra R_B e R_C, con R_C moltiplicato per il fattore $\beta_F >> 1$. In effetti per i circuiti a transistore sono sufficienti resistenze di carico dell'ordine dei kΩ, che possono essere integrate senza eccessivo consumo di area (ricordiamo che, come si è detto nel Paragrafo 2.4, si può utilizzare, nel processo di realizzazione dei transistori bipolari, la diffusione di base che fornisce resistenze di strato tra 200 e 500 Ω/□, il che comporta per una resistenza da 1kΩ, un'area di $3\lambda \times 6 \div 15\lambda$).

La possibilità di integrazione di resistenze di valore opportuno nella tecnologia bipolare giustifica il fatto che nelle porte bipolari, sia quelle RTL e DTL che, come vedremo, quelle TTL ed ECL discusse nei capitoli seguenti, si utilizzeranno essenzialmente carichi resistivi.

7.3 Caratteristica di trasferimento e margini di rumore

La caratteristica di trasferimento dell'invertitore, riportata in Figura 7.3, è facilmente interpretabile in base alle relazioni precedentemente richiamate.

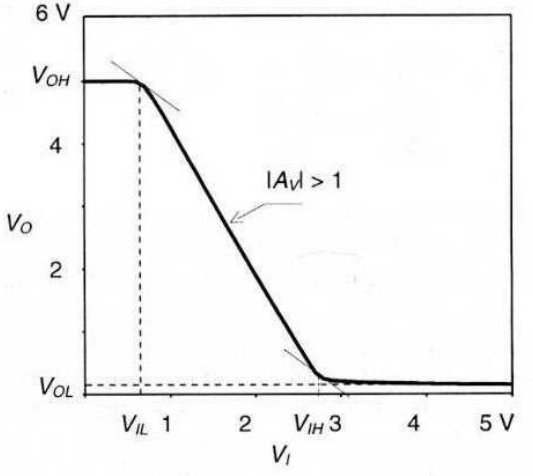


Figura 7.3 Caratteristica di trasferimento dell'invertitore RTL

Per tensioni minori della tensione di soglia $V_{BE\gamma}$ la corrente di base è nulla, e la tensione di uscita V_{OH} è pari a V_{CC}. Per ingressi $V_I > V_{BE\gamma}$ il transistore entra in zona attiva di funzionamento; la corrente di base può essere ancora descritta da una relazione analoga alla (7.3):

$$I_B = \frac{V_I - V_{BE}}{R_B} \cong \frac{V_I - 0.7\ \text{V}}{R_B} \tag{7.5}$$

e la corrente di collettore da:

$$I_C = \frac{V_{CC} - V_{CE}}{R_C} \tag{7.6}$$

In zona attiva vale la relazione (6.16) tra le correnti I_C e I_B:

$$I_C = \beta_F I_B \tag{7.7}$$

che, sostituendo in questa le (7.5) e (7.6), fornisce il legame tra tensione di ingresso e di uscita:

$$\frac{V_{CC} - V_{CE}}{R_C} \equiv \frac{V_{CC} - V_O}{R_C} = \beta_F \frac{V_I - 0.7\ \text{V}}{R_B} \tag{7.8}$$

La (7.8) fornisce un legame lineare tra ingresso e uscita, il che corrisponde ad una pendenza costante della caratteristica di trasferimento tra interdizione e saturazione (pendenza che corrisponde all'amplificazione nel regime di piccoli segnali), data da:

$$\frac{\Delta V_O}{\Delta V_I} = -\frac{\beta_F R_C}{R_B} \tag{7.9}$$

In base alla condizione (7.4), che abbiamo visto essere necessaria per portare il transistore in saturazione con l'ingresso logico alto, la (7.9) ci dice che la pendenza della caratteristica di trasferimento nella regione attiva è maggiore di 1. Questo risultato ha come importante corollario che la tensione V_{IL} (corrispondente al punto della curva con pendenza pari a -1) coincide con la tensione di soglia, e cioè con la discontinuità della caratteristica di trasferimento. Analoghe considerazioni valgono per il valore V_{IH}, che per la stessa ragione coincide con la tensione di ingresso che porta il transistore in saturazione. Quest'ultima si ricava ancora dalla (7.8), ma specificando per il valore V_{CE} quello V_{CESAT} al limite della saturazione:

$$\frac{V_{CC} - V_{CESAT}}{R_C} = \beta_F \frac{V_{IH} - V_{BESAT}}{R_B} \tag{7.10}$$

da cui:

$$V_{IH} = \frac{R_B}{\beta_F R_C}(V_{CC} - V_{CESAT}) + V_{BESAT} \tag{7.11}$$

Il valore di V_{OH} identificato tramite l'analisi grafica precedente è valido per un invertitore non connesso ad altre porte logiche; per determinare i livelli logici nominali e i margini di rumore occorre invece considerare l'effetto di carico che deriva dalla connessione in uscita di un successivo invertitore, che assorbe corrente nello stato di uscita alta per l'invertitore a monte, come è indicato nel circuito riportato in Figura 7.4; l'assorbimento di corrente dello stadio a valle modifica il valore V_{OH} fornito dallo stadio a monte. Con riferimento allo schema elettrico di Figura 7.4, assumendo Q_1 interdetto e Q_2 in saturazione, la tensione V_{OH} in uscita da Q_1 si può determinare in questo caso come:

$$V_{OH} = (V_{CC} - V_{BESAT}) \frac{R_B}{R_B + R_C} + V_{BESAT} \tag{7.12}$$

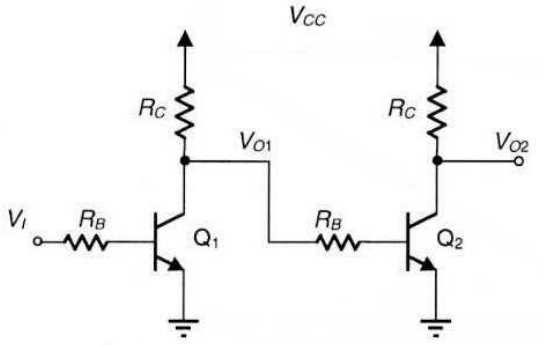


Figura 7.4 Connessione in serie di due invertitori per il calcolo di V_{OH}

In definitiva i valori delle quattro grandezze caratteristiche che determinano i margini di rumore NM_L e NM_H sono:

$$V_{OH} = (V_{CC} - V_{BESAT}) \frac{R_B}{R_B + R_C} + V_{BESAT} \tag{7.13a}$$

$$V_{IH} = \frac{R_B}{\beta_F R_C} (V_{CC} - V_{CESAT}) + V_{BESAT} \tag{7.13b}$$

$$V_{IL} = V_{BE\gamma} \cong 0.6 \text{ V} \tag{7.13c}$$

$$V_{OL} = V_{CESAT} \cong 0.2 \text{ V} \tag{7.13d}$$

da cui si ottengono i margini di rumore:

$$NM_H = V_{OH} - V_{IH}; \qquad NM_L = V_{IL} - V_{OL} \tag{7.14}$$

Per l'invertitore bipolare si vede dalle (7.13) che il parametro a disposizione del progettista per modificare i margini di rumore è il rapporto R_B/R_C, che oltre a dover soddisfare la (7.4), influisce sul valore di V_{OH} e V_{IH}. In Figura 7.5 sono riportate le dipendenze dei termini delle (7.13) in funzione del rapporto R_B/R_C.

Dalla Figura 7.5 si vede che, per avere valori di NM_H sufficientemente elevati, occorre scegliere valori del rapporto R_B/R_C ben inferiori al valore di β_F; d'altra parte il valore di questo rapporto non può essere troppo basso per non ridurre significativamente il valore di V_{OH} rispetto a V_{CC}. In ogni caso il margine di rumore NM_L è relativamente basso, e questo pone un limite significativo all'impiego di questo tipo di invertitore.

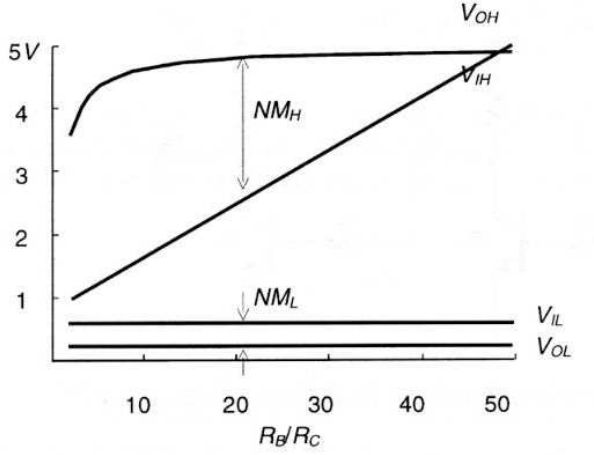


Figura 7.5 Dipendenza dei margini di rumore di un invertitore RTL dal rapporto R_B/R_C, per un valore $\beta_F = 50$

7.4 Fan-out e dissipazione di potenza

Il fan-out per gli invertitori bipolari è determinato dall'assorbimento massimo di corrente che l'invertitore può sopportare in uscita con una degradazione accettabile della tensione V_{OH}; ogni ulteriore stadio connesso in uscita assorbe infatti una corrente dall'ingresso per essere pilotato nello stato alto, e questa deve essere fornita dall'uscita dell'invertitore di pilotaggio. In base allo schema di Figura 7.6, il carico di n invertitori uguali connessi all'uscita di quello di pilotaggio modifica la tensione in uscita secondo la relazione:

$$V_{OH} = (V_{CC} - V_{BESAT}) \frac{R_B / n}{R_B / n + R_C} + V_{BESAT} \tag{7.15}$$

La (7.15) mostra come il fan-out per questo invertitore sia relativamente basso; infatti, con un rapporto $R_B/R_C = 10$ per il singolo invertitore, già con un numero $n = 5$ di invertitori in uscita la tensione di uscita V_{OH} si riduce da 5 V a 3.6 V. Da questo esempio si deduce che il fan-out massimo per l'invertitore è limitato da una massima degradazione ammissibile del margine di rumore alto fino ad un limite minimo NM_{Hmin} definito dalle specifiche della porta stessa:

$$V_{OH}(n) - V_{IH} \equiv NM_{H\min} \tag{7.16}$$

La dissipazione di potenza statica è nulla se l'ingresso è nello stato logico basso, perché il transistore è interdetto in questo caso; si ha invece consumo di

potenza nello stato logico alto, dovuto alla corrente I_{CSAT} che circola nel transistore.

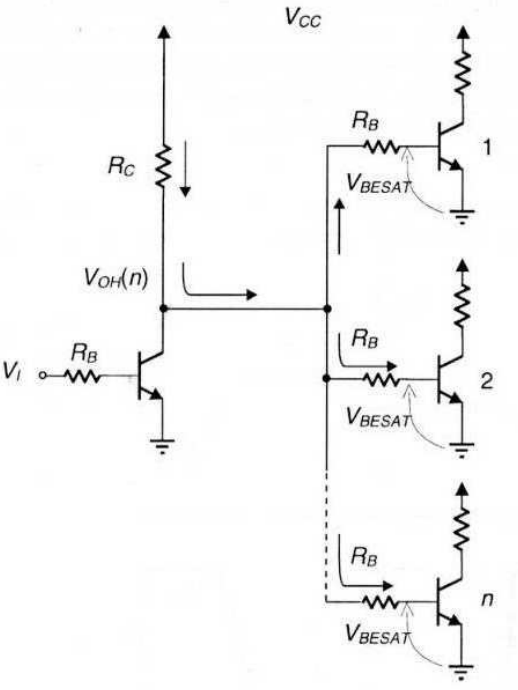


Figura 7.6 Calcolo della V_{OH} con fan-out di n

Poiché la dissipazione di potenza viene valutata in base alla media delle potenze dissipate nei due possibili stati dell'invertitore, la potenza dissipata P_D è data da:

$$P_D = \frac{V_{CC} I_{CSAT}}{2} = \frac{V_{CC}}{2} \frac{V_{CC} - V_{OL}}{R_C} \cong \frac{V_{CC}^2}{2R_C} \tag{7.17}$$

Confrontando la (7.17) con la (7.15) relativa al fan-out si può già vedere come la scelta del valore della resistenza di carico R_C sia legata ad un compromesso tra diverse esigenze: infatti la riduzione della potenza dissipata dall'invertitore, che comporta un aumento della resistenza di carico, si scontra con la necessità di non degradare troppo la tensione alta in presenza di fan-out maggiore di 1, il che comporta una riduzione di R_C; ulteriori e più significativi vincoli sulle scelte delle resistenze vengono dall'analisi dinamica che verrà sviluppata nel prossimo paragrafo.

7.5 Comportamento dinamico dell'invertitore

L'esame di una tipica forma d'onda della tensione in uscita di un invertitore RTL pilotato da un segnale impulsivo di tipo logico all'ingresso è simile a quella presentata per invertitori MOS, con una significativa differenza: in questo caso la tensione di uscita passa dallo stato basso a quello alto dopo un tempo finito, durante il quale il transistore rimane ancora in saturazione, per un tempo finito dopo che il segnale di ingresso è ritornato al livello basso.

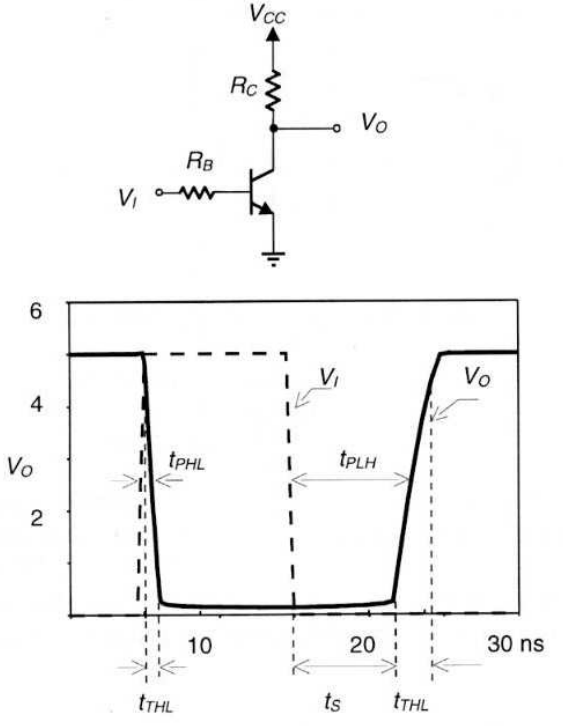


Figura 7.7 Forme d'onda di uscita di un invertitore RTL con $\beta_F = 50$, $\tau_F = 0.06$ ns, $\tau_R = 10$ ns, per un rapporto $R_B/R_C = 10$

Questo tempo caratteristico, indicato con t_S nel grafico di Figura 7.7, è detto *tempo di accumulo* (*storage time*) del transistore, e viene a sommarsi ai tempi di transizione t_{THL} e t_{TLH}, indicati nel Capitolo 1. Il tempo di accumulo è dovuto al tempo richiesto per smaltire la carica dei portatori minoritari accumulata nella base nella fase di saturazione del transistore (si dovrebbe più correttamente chiamare tempo di smaltimento), e contribuisce significativamente al ritardo di propagazione delle porte bipolari che operano in regime di saturazione.

I tempi caratteristici del comportamento dinamico dell'invertitore possono essere ricavati direttamente dalle equazioni del modello a Controllo di Carica, sotto

opportune semplificazioni. Si supporrà, in accordo con l'analisi sviluppata per gli invertitori NMOS e CMOS, che il segnale di ingresso vari tra 0 e V_{CC} con tempi di salita e di discesa nulli; inoltre si trascureranno gli effetti delle capacità di svuotamento delle giunzioni, considerando solo gli effetti sulla dinamica dovuti alla carica nella base.

Dalla Figura 7.8 si vede che l'analisi dei tempi di propagazione definiti sulla forma d'onda della tensione (in particolare è evidenziato il tempo t_{PHL}) può essere riferita alla forma d'onda della corrente, essendo la variazione di quest'ultima proporzionale a quella della tensione, a parte il segno ($\Delta I_C = -\Delta V_{CE}/R_C$). Ricaveremo quindi le espressioni di t_{PLH}, t_{PHL} con riferimento alla corrente I_C che è una delle variabili esplicite delle equazioni del modello a Controllo di Carica.

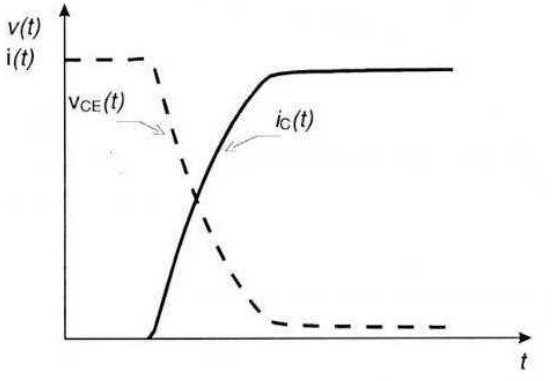


Figura 7.8 Forme d'onda di corrente e tensione di collettore

a) Commutazione dal livello alto al basso

Con il segnale di ingresso allo 0 logico, il transistore si trova in interdizione, e la carica immagazzinata nella base è nulla ($Q_F = Q_R = 0$). Al passaggio dell'ingresso dallo 0 all'1 logico ($0 \rightarrow V_{CC}$) il transistore si porta a funzionare in *modo attivo diretto*, e le equazioni del modello che reggono questo comportamento sono le (6.33) relative a tale modo:

$$i_B(t) = \frac{Q_F}{\tau_{BF}} + \frac{dQ_F}{dt} \tag{7.18a}$$

$$i_C(t) = \frac{Q_F}{\tau_F} \tag{7.18b}$$

Nell'invertitore RTL (vedi Figura 7.7) la corrente di base durante la commutazione è in pratica costante, essendo data da:

$$i_B(t) = \frac{V_{CC} - V_{BE}(t)}{R_B} \cong \frac{V_{CC}}{R_B} \equiv I_{B1} \tag{7.19}$$

Quindi la soluzione dell'equazione differenziale (7.18a), con la condizione iniziale $Q_F(0) = 0$, e con il termine $i_B(t) = I_{B1}$, vale:

$$Q_F(t) = \tau_{BF} I_{B1} \left[1 - \exp\left(-\frac{t}{\tau_{BF}} \right) \right] \tag{7.20}$$

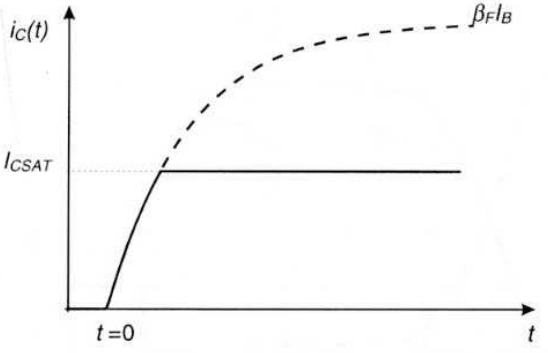


Figura 7.9 Soluzione dell'Equazione 7.21 per la corrente i_C

Dalla (7.18b), ricordando la (6.40) che lega β_F a τ_F, τ_{BF}, si ha:

$$i_C(t) = \beta_F I_{B1} \left[1 - \exp\left(-\frac{t}{\tau_{BF}} \right) \right] \tag{7.21}$$

La (7.21) mostra che la corrente cresce tendendo asintoticamente al valore $I_{CMAX} = \beta_F I_B$; tuttavia la crescita si arresta quando la corrente raggiunge il valore di saturazione I_{CSAT}, dato da:

$$I_{CSAT} = \frac{V_{CC} - V_{CESAT}}{R_C} \cong \frac{V_{CC}}{R_C} \tag{7.22}$$

Il tempo di propagazione t_{PHL} (riferito alla tensione) è definito dall'istante in cui la corrente I_C raggiunge il valore $I_{CSAT}/2$, ed in base alla (7.21) vale:

$$t_{PHL} = \tau_{BF} \ln\left[\frac{1}{1 - \dfrac{I_{CSAT}}{2\beta_F I_{B1}}}\right] \cong \tau_{BF} \ln\left[\frac{1}{1 - \dfrac{R_B}{2\beta_F R_C}}\right] \tag{7.23}$$

dove si è posto $I_{CSAT}/I_B \cong R_B/R_C$.

Ricordiamo che per forzare il transistore in saturazione occorre soddisfare la (7.4); se la disequazione è molto minore di 1 (forte saturazione), lo sarà anche il secondo termine al denominatore nella parentesi quadra della (7.23), e si può arrestare al primo termine lo sviluppo del termine logaritmico, da cui:

$$t_{PHL} \cong \tau_{BF} \frac{R_B}{2\beta_F R_C} = \tau_F \frac{R_B}{2R_C} \tag{7.24}$$

b) Tempo di accumulo

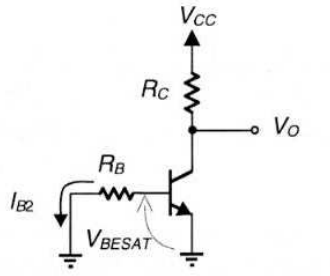
In questa fase (Figura 7.10) il segnale di ingresso passa da V_{CC} a 0, e quindi la corrente di base circola *dalla base verso il generatore di segnale* perché la tensione sulla base è ancora V_{BESAT} mentre quella di ingresso è nulla; il transistore è infatti ancora in saturazione (finché la carica in eccesso nella base non viene estratta). La distribuzione della carica è determinata dalla sovrapposizione dei modi attivo diretto ed attivo inverso (Equazione (6.37)), ma per l'analisi approssimata è più conveniente considerare questa distribuzione come dovuta alla somma di due cariche Q_A e Q_S, con quest'ultima che rappresenta l'eccesso di carica (a distribuzione uniforme) che non contribuisce alla corrente di collettore (perché $dn/dx = 0$) (Figura 7.10b).

L'uscita dalla saturazione, che comporta una corrente costante durante il tempo di accumulo t_S, può quindi definirsi in base al tempo necessario per completare l'estrazione della carica Q_S, tempo durante il quale la carica Q_A rimane inalterata. Poiché Q_A rimane costante durante il tempo t_S, la si può esprimere in base alla componente di corrente di base al limite della saturazione I_{BSAT}:

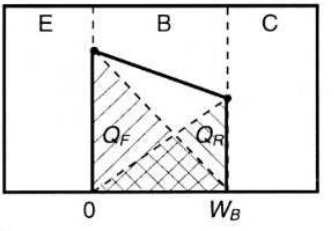
$$I_{BSAT} = \frac{Q_A}{\tau_{BF}} \qquad \left[\frac{dQ_A}{dt} = 0\right] \tag{7.25}$$

dove la corrente I_{BSAT} è definita come quella corrente di base che al limite della regione attiva fornisce la corrente di collettore I_{CSAT}:

$$I_{BSAT} = \frac{I_{CSAT}}{\beta_F} \cong \frac{V_{CC}}{\beta_F R_C} \tag{7.26}$$



(a)



(b)

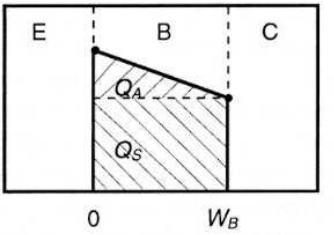


Figura 7.10 a) Estrazione di corrente durante il tempo di accumulo; b) distribuzione delle cariche nella base per la valutazione di t_S

In queste ipotesi, con riferimento alla sola componente di carica Q_S ed alla corrente di base $i_B(t) - I_{BSAT}$ si può scrivere un nuovo sistema di equazioni di controllo di carica, valido nella regione di saturazione:

$$i_B(t) - I_{BSAT} = \frac{Q_S}{\tau_S} + \frac{dQ_S}{dt} \tag{7.27a}$$

$$i_C(t) = I_{CSAT} \tag{7.27b}$$

dove si è assunto come tempo caratteristico della ricombinazione nella base la grandezza τ_S (ricordiamo che in questo regime coesistono i due modi di funzionamento diretto ed inverso), che si può dimostrare essere legata ai tempi τ_{BF} e τ_{BR} secondo la relazione:

$$\tau_S = \frac{\tau_{BF}(\beta_R + 1) + \tau_{BR}\beta_F}{1 + \beta_F + \beta_R} \cong \tau_{BR} \qquad [\beta_F >> \beta_R \cong 1\ ;\ \tau_{BR} \geq \tau_{BF}] \quad (7.28)$$

La (7.27a) può essere integrata con la condizione iniziale:

$$Q_S(0) = \tau_S(I_{B1} - I_{BSAT})$$

assumendo anche in questo caso una corrente di estrazione di base costante:

$$i_B(t) = -\frac{v_{BE}(t)}{R_B} \cong -\frac{V_{BESAT}}{R_B} \equiv I_{B2} \quad (7.29)$$

e ottenendo:

$$Q_S(t) = \tau_S\left[I_{B2} - I_{BSAT} + (I_{B1} - I_{B2})\exp\left(-\frac{t}{\tau_S}\right)\right] \quad (7.30)$$

Per definizione del tempo di accumulo, $Q_S(t_S) = 0$; sostituendo questa condizione nella (7.30) si ha:

$$t_S = \tau_S \ln\left(\frac{I_{B1} - I_{B2}}{I_{BSAT} - I_{B2}}\right) \quad (7.31)$$

Ricordando le espressioni di I_{B1}, I_{BSAT}, I_{B2} date dalle (7.19), (7.26), (7.29) si può scrivere t_S come:

$$t_S = \tau_S \ln\left(\frac{V_{CC} + V_{BESAT}}{\frac{R_B}{\beta_F R_C} V_{CC} + V_{BESAT}}\right) \quad (7.32)$$

Dalla (7.32) si vede (come era prevedibile) che $t_S \to 0$ per $R_B/\beta_F R_C \to 1$; questa condizione equivale a porre $I_{B1} = I_{BSAT}$, cioè corrisponde alla condizione per cui il punto di lavoro è al limite tra saturazione e zona attiva. Il tempo di accumulo cresce quindi all'aumentare del forzamento in saturazione, e questo pone un limite alla possibilità di ridurre il rapporto R_B/R_C, condizione che invece migliorava i margini di rumore; occorre quindi trovare una condizione di compromesso tra questi due aspetti nel progetto dell'invertitore RTL.

c) Passaggio dalla saturazione all'interdizione

In questo caso l'analisi procede analogamente al caso a), in quanto il transistore opera in modo attivo diretto dopo il tempo t_S (la carica Q_S si è annullata e rimane quella Q_A che va diminuendo via via che si riduce il valore $n_B(0)$ sulla giunzione di emettitore). Ridefinendo questa carica come $Q_F(t)$, il sistema di equazioni è ancora quello delle (7.18) che ora va risolto con la condizione iniziale:

$$Q_F(0) \equiv Q_A = I_{BSAT} \cdot \tau_{BF} \tag{7.33}$$

e con corrente di base $i_B(t) = I_{B2}$ (in questo caso l'approssimazione è più drastica, perché la giunzione di base si va portando verso l'interdizione e quindi verso la fine del transitorio la corrente di base tenderà a 0), ottenendo:

$$Q_F(t) = I_{B2}\tau_{BF} + (I_{BSAT} - I_{B2})\tau_{BF} \exp\left(-\frac{t}{\tau_{BF}}\right) \tag{7.34}$$

da cui, sostituendo nella (7.18b):

$$i_C(t) = \beta_F I_{B2} + \beta_F (I_{BSAT} - I_{B2}) \exp\left(-\frac{t}{\tau_{BF}}\right) \tag{7.35}$$

Anche in questo caso la corrente segue una legge esponenziale tendendo ad un valore negativo dato da $\beta_F I_{B2}$ (si ricorda che I_{B2} è negativa), ma la legge si arresta quando la corrente di collettore si annulla, in quanto a questo punto si è annullata la carica Q_F e anche la corrente di base. Definiamo il tempo caratteristico t_{LH}, tale che $t_S + t_{LH} = t_{PLH}$, come quello per cui la corrente di collettore si riduce a $I_{CSAT}/2$; dalla (7.35) si ha, ricordando la (7.26):

$$t_{LH} = \tau_{BF} \ln\left[\frac{I_{CSAT} - \beta_F I_{B2}}{I_{CSAT}/2 - \beta_F I_{B2}}\right] \tag{7.36}$$

Sostituendo anche in questa le espressioni di I_{CSAT}, I_{B2} date dalle (7.22), (7.29) si ha:

$$t_{LH} = \tau_{BF} \ln\left[\frac{1 + \dfrac{R_B}{\beta_F R_C}\dfrac{V_{CC}}{V_{BESAT}}}{1 + \dfrac{R_B}{2\beta_F R_C}\dfrac{V_{CC}}{V_{BESAT}}}\right] \tag{7.37}$$

Dalle espressioni (7.32) e (7.39) di t_S e t_{LH} si può infine ottenere il valore di t_{PLH} come: $t_{PLH} = t_S + t_{LH}$.

7.6 Ritardo di propagazione e prodotto potenza-ritardo

In base all'analisi precedente si vede che i tre tempi caratteristici t_{PHL}, t_S, t_{LH}, legati alla dinamica di commutazione dell'invertitore, sono funzione del rapporto:

$$R^* \equiv \frac{R_B}{\beta_F R_C} < 1 \tag{7.38}$$

che assume una rilevanza pari a quella del rapporto K_R per gli invertitori NMOS. Indicando con A il rapporto V_{CC}/V_{BESAT} si hanno le seguenti espressioni per i tempi di propagazione:

$$t_{PHL} = \tau_{BF} \ln\left(\frac{1}{1 - R^*/2}\right) \tag{7.39}$$

$$t_{PLH} = t_S + t_{LH} = \tau_S \ln\left(\frac{1 + A}{1 + A \cdot R^*}\right) + \tau_{BF} \ln\left(\frac{1 + A \cdot R^*}{1 + A \cdot R^*/2}\right) \tag{7.40}$$

$$t_P = \frac{t_{PHL} + t_{PLH}}{2} = \frac{1}{2}\left[\tau_S \ln\left(\frac{1 + A}{1 + AR^*}\right) + \tau_{BF} \ln\left(\frac{1 + AR^*}{(1 + AR^*/2)(1 - R^*/2)}\right)\right] \tag{7.41}$$

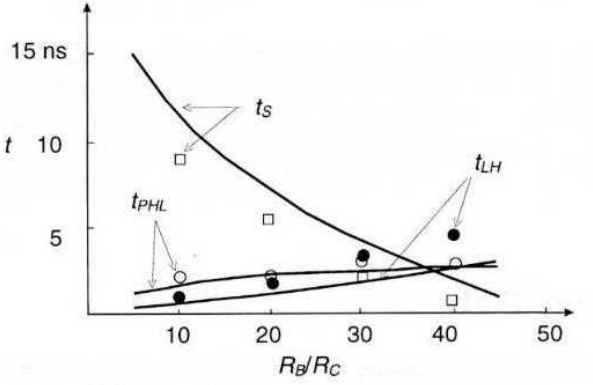


Figura 7.11 Dipendenza dei tempi di commutazione dal rapporto R_B/R_C per un transistore con $\beta_F = 50$, $\beta_R = 1$, $\tau_F = 0.1$ ns, $\tau_R = 10$ ns. In linea continua sono riportati i valori delle (7.23), (7.32), (7.37), mentre i valori discreti sono ottenuti da simulazioni con SPICE

In Figura 7.11 sono riportati i valori di t_{PHL}, t_S, t_{LH} forniti dalle (7.23), (7.32), (7.37), al variare del rapporto R_B/R_C, per un transistore con $\beta_F = 50$. Il confronto con i risultati ottenuti da simulazioni SPICE è in buon accordo sia con le dipendenze funzionali che con i valori, considerando le approssimazioni alla base delle relazioni analitiche. Da questi grafici si vede come la riduzione del rapporto R_B/R_C e cioè del forzamento in saturazione del transistore, se da una parte accelera le transizioni alto-basso e basso-alto, dall'altra aumenta molto il tempo di accumulo, che diviene la parte più rilevante del ritardo di propagazione a forzamenti elevati.

Il prodotto ritardo-potenza $P \cdot D$ è dato dal prodotto della (7.17) e (7.41); nell'ipotesi di forzamento significativo in saturazione, il maggior contributo al ritardo di propagazione t_P è dato dal tempo di accumulo t_S, ed un'espressione approssimata del prodotto può essere scritta come:

$$P \cdot D = \frac{V_{CC}^2 \tau_S}{4R_C} \ln\left(\frac{V_{CC} + V_{BESAT}}{\frac{R_B}{\beta_F R_C} V_{CC} + V_{BESAT}}\right) \tag{7.42}$$

da cui si vede che, anche per l'invertitore bipolare, se il carico aumenta la dissipazione di potenza diminuisce ma il ritardo di propagazione aumenta. In questo caso però, diversamente da quanto visto per gli invertitori MOS, il prodotto ritardo-potenza non è indipendente dal valore del carico R_C in quanto questo compare nel ritardo di propagazione sotto logaritmo.

Come ordine di grandezza di questo prodotto per un invertitore RTL, assumendo $R_C = 1$ kΩ, $R_B/R_C = 20$, $\beta_F = 50$, con i valori della Figura 7.11 si ottiene:

$$t_P = 6.3 \text{ ns}; \quad P_D = 12.5 \text{ mW}; \quad P \cdot D = 78.7 \text{ pJ}$$

Ricordiamo che dal punto di vista statico il forzamento in saturazione migliora i margini di rumore e la potenza dissipata; tuttavia questo miglioramento viene pagato con un aumento del tempo di accumulo e quindi del ritardo di propagazione. In definitiva le condizioni contrastanti legati alla scelta del rapporto R_B/R_C rendono questo tipo di invertitore (e la logica che ne deriva) poco flessibile e di limitata applicazione.

7.7 L'invertitore DTL

Una logica sviluppata successivamente alla logica RTL è la logica a diodi e transistori detta *logica DTL* (*Diode-Transistor Logic*), che presenta miglioramenti rispetto a quella RTL essenzialmente nei margini di rumore; nell'invertitore elementare DTL la rete resistiva di ingresso viene sostituita da una rete a diodi, da cui il nome della logica. I diodi possono essere utilizzati, come elementi nonlineari, per realizzare operazioni logiche, per cui la logica DTL può

anche essere considerata come una logica a diodi posta in ingresso ad un invertitore a transistore che ripristina i livelli logici. Sebbene anche questa logica sia attualmente superata, lo studio dell'invertitore elementare può aiutare a comprendere meglio la logica TTL, che ne è la naturale evoluzione, e che è la logica bipolare standard più utilizzata.

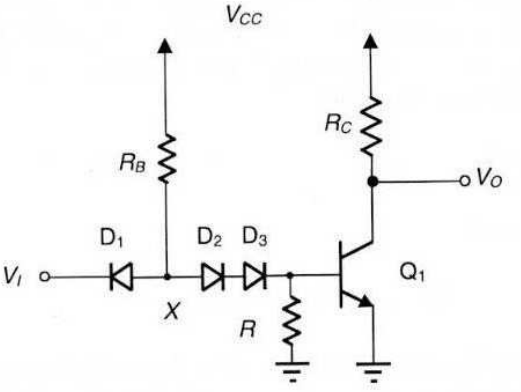


Figura 7.12 Schema dell'invertitore DTL

Il circuito base dell'invertitore DTL è quello di Figura 7.12. Confrontando questo schema con quello RTL di Figura 7.1 si nota l'inserimento dei diodi D_1, D_2, D_3, e delle resistenze di polarizzazione R_B e R nella maglia d'ingresso dell'invertitore. Analizziamo il circuito per tensioni di ingresso V_I crescenti da 0 a V_{CC} in modo da trovare i valori caratteristici della funzione di trasferimento necessari per la valutazione dei margini di rumore, utilizzando le approssimazioni già applicate nell'analisi dell'invertitore RTL.

Per $V_I = 0$, il diodo D_1 conduce (perché se non circolasse corrente in R_B, V_X sarebbe pari a V_{CC} e quindi non è possibile che D_1 sia interdetto; se invece conducessero solo D_2 e D_3 la tensione V_X sarebbe maggiore della tensione di soglia di D_1 e quindi anche in questo caso l'assunto di D_1 interdetto è errato). La tensione V_X sarà quindi data da: $V_X = V_D \cong 0.7$ V, ed essendo inferiore alla somma delle tensioni di soglia dei diodi D_2, D_3, se ne deduce che D_2, D_3 sono interdetti. Quindi anche Q_1 è interdetto e:

$$V_O(0) \equiv V_{OH} = V_{CC} \tag{7.43}$$

Al crescere di V_I la tensione $V_X = V_I + V_D$ aumenta, fino a che i diodi D_2 e D_3 conducono e circola corrente in R. Il valore della tensione di ingresso V_I tale che la caduta su R raggiunge il valore $V_{BE\gamma}$ e il transistore comincia a condurre, è per definizione la tensione V_{IL}, perché per tensioni superiori la pendenza (amplificazione) della caratteristica di trasferimento è in modulo maggiore di 1 (il circuito corrisponde nel comportamento per le componenti variabili ad un amplificatore ad

emettitore comune, in quanto $\Delta V_I = \Delta V_{BE}$ perché le cadute V_D sui diodi si suppongono costanti); si ha quindi per la tensione di ingresso V_{IL}:

$$V_{IL} = -V_{D1} + V_{D2} + V_{D3} + V_{BE\gamma} \cong -0.7 + 1.4 + 0.6 = 1.3 \text{ V} \tag{7.44}$$

All'aumentare della tensione di ingresso oltre questo valore, aumenta anche la tensione sulla base di Q_1 perché aumenta la caduta su R, fino a portare Q_1 in saturazione. A questo punto la pendenza della curva di trasferimento si annulla, e la tensione corrispondente di ingresso, V_{IH}, vale:

$$V_{IH} = -V_{D1} + V_{D2} + V_{D3} + V_{BESAT} \cong -0.7 + 1.4 + 0.8 = 1.5 \text{ V} \tag{7.45}$$

e la tensione di uscita V_O vale:

$$V_O(V_{IH}) \equiv V_{OL} = V_{CESAT} \tag{7.46}$$

Al crescere ulteriore di V_I, essendo il valore massimo di V_X vincolato dal lato della base del transistore ad una tensione di 1.4 + 0.8 = 2.2 V, il diodo D_1 si interdice per $V_I = 2.2 - 0.6 = 1.6$ V, mentre i diodi D_2 e D_3 continuano a condurre; la tensione in uscita rimane al valore V_{OL}. I margini di rumore dell'invertitore DTL sono:

$$NM_H = V_{OH} - V_{IH} \cong 5 - 1.5 = 3.5 \text{ V} \tag{7.47a}$$

$$NM_L = V_{IL} - V_{OL} \cong 1.3 - 0.2 = 1.1 \text{ V} \tag{7.47b}$$

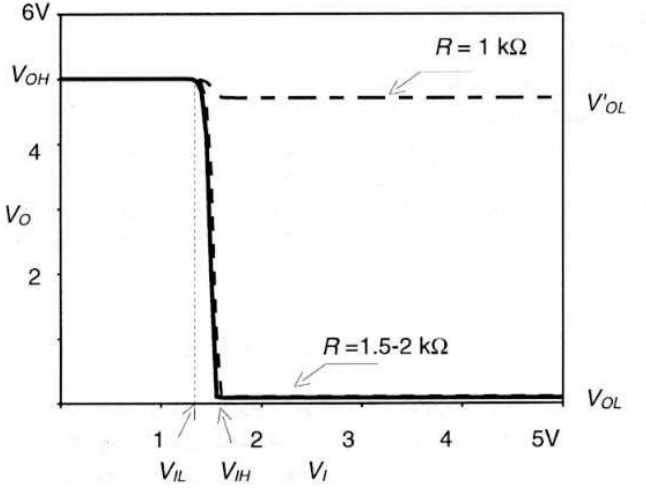


Figura 7.13 Caratteristiche di trasferimento dell'invertitore DTL per tre diversi valori di R, e con: $R_B = 4$ kΩ, $R_C = 1$ kΩ

con NM_L significativamente più elevato di quello dell'invertitore RTL, e NM_H che risulta indipendente dalla scelta della resistenza di carico R_C. Come si può vedere dalle curve di trasferimento di Figura 7.13, vi è un valore minimo di resistenza R al di sotto del quale il transistore non va in saturazione; questo valore può essere ricavato ricordando che per portare in saturazione il transistore con $V_I = V_{IH}$ deve essere $I_B \geq I_C/\beta_F$. Poiché per tensioni di ingresso appena superiori a V_{IH} il diodo D_1 va in interdizione, dal circuito di Figura 7.12 si ha che:

$$I_B \geq \frac{I_C}{\beta_F} \Rightarrow \frac{V_{CC} - 2V_D - V_{BESAT}}{R_B} - \frac{V_{BESAT}}{R} \geq \frac{V_{CC}}{\beta_F R_C} \tag{7.48}$$

da cui:

$$V_{CC}\left(1 - \frac{R_B}{\beta_F R_C}\right) - 2V_D - V_{BESAT}\left(1 + \frac{R_B}{R}\right) \geq 0 \tag{7.49}$$

Nell'ipotesi di poter trascurare nella parentesi relativa al primo termine della disequazione il termine $R_B/\beta_F R_C$ rispetto all'unità (assumendo cioè $\beta_F R_C >> R_B$), si ottiene la seguente condizione per R:

$$R \geq \frac{V_{BESAT}}{V_{CC} - V_{BESAT} - 2V_D} R_B \cong 0.3 \cdot R_B \tag{7.50}$$

che giustifica i risultati delle simulazioni SPICE di Figura 7.13, in quanto per R_B = 4 kΩ, deve essere $R > 1.2$ kΩ.

La resistenza R_B d'altra parte non deve essere troppo bassa, in quanto il suo valore limita il fan-out dell'invertitore. Infatti, con riferimento al circuito di Figura 7.12, con n invertitori in uscita, quando Q_1 è in saturazione ($V_O = V_{OL} = V_{CESAT}$), i diodi D_1 di ognuno degli invertitori in uscita sono in conduzione e la corrente che circola nel collettore di Q_1 sarà data da:

$$I_{C1} = I_{CSAT} + n \cdot I_I = \frac{V_{CC} - V_{CESAT}}{R_C} + n\left[\frac{V_{CC} - V_D - V_{CESAT}}{R_B}\right] \leq \beta_F I_{B1} \tag{7.51}$$

e deve essere minore della massima corrente in saturazione $\beta_F I_B$; con un R_B = 4 kΩ la corrente I_I iniettata da ciascuno degli invertitori in uscita è di circa 1 mA.

Confrontando i tempi di commutazione della Figura 7.14 dell'invertitore DTL con quelli della Figura 7.7 per un RTL, si può notare la maggiore velocità del primo, essenzialmente a causa del minor valore della resistenza di base R, che riduce la I_{B1} di forzamento in saturazione, e, in fase di spegnimento del transistore, permette una estrazione maggiore di corrente e quindi un minor tempo di accumulo e di transizione. In definitiva la logica DTL permette significativi miglioramenti rispetto alla RTL, principalmente a causa dell'indi-

pendenza dei margini di rumore dalle resistenze, il che permette una più ampia scelta di valori che contribuiscono a migliorare le caratteristiche dinamiche.

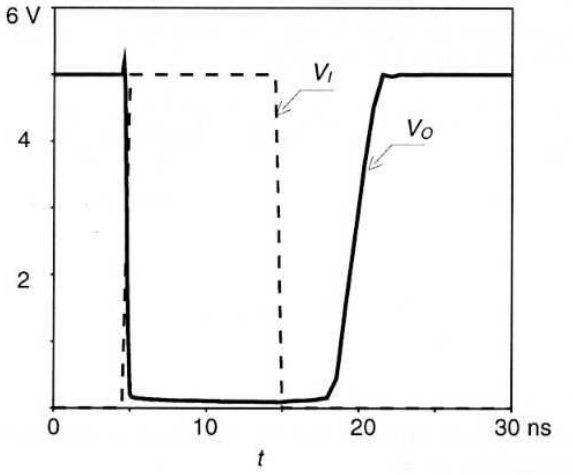


Figura 7.14 Forme d'onda di ingresso e di uscita di un invertitore DTL con $\beta_F = 50$, $\tau_F = 0.06$ ns, $\tau_R = 10$ ns; $R_C = 1$ kΩ, $R_B = 4$ kΩ, $R = 3$ kΩ

Occorre però fare attenzione nel valutare le prestazioni dinamiche degli invertitori bipolari con carico resistivo, per quanto riguarda il pilotaggio di capacità di carico relativamente elevate. Ad esempio, con una capacità di carico $C_L = 10$ pF le prestazioni dinamiche di un invertitore DTL vengono significativamente peggiorate, come si può vedere dalla Figura 7.15, in particolare per quanto riguarda la transizione in uscita dal valore basso a quello alto. L'aumento del tempo di transizione t_{TLH} è dovuto al fatto che, anche supponendo una commutazione praticamente istantanea del transistore dopo il tempo di accumulo, occorre un tempo relativamente lungo per caricare la capacità C_L attraverso la resistenza di carico R_C, dipendente dalla costante di tempo $C_L R_C$. Anche il tempo di transizione t_{THL} risulta aumentato, perché la corrente di scarica della capacità attraverso il transistore determina un aumento della corrente di collettore, e quindi un minor forzamento in saturazione (a parità di corrente iniettata in base) il che comporta un aumento del tempo di commutazione del transistore stesso.

Questa degradazione dell'uscita con carichi capacitivi relativamente elevati è in certo modo confrontabile con la situazione già esaminata per le porte NMOS e CMOS, sebbene la resistenza di carico R_C per le porte bipolari possa avere un valore relativamente più basso rispetto alla resistenza di un transistore MOS ad area minima; si determina quindi un'analoga necessità di avere degli stadi di disaccoppiamento (buffer) per le uscite delle logiche DTL collegabili a circuiti esterni al chip. Vedremo invece che per le porte TTL questo problema verrà superato per merito di un opportuno stadio di uscita.

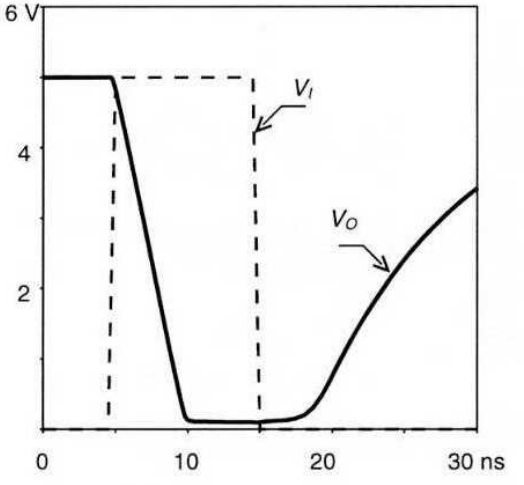


Figura 7.15 Forme d'onda in uscita dell'invertitore di Figura 7.12 con una capacità di carico $C_L = 10$ pF

7.8 Porte logiche DTL

La rete di diodi inseriti all'ingresso dell'invertitore DTL può essere anche impiegata per effettuare una funzione logica a più variabili, compattando quindi il circuito della porta logica corrispondente rispetto alla versione che utilizza opportune combinazioni di invertitori elementari come nel caso MOS.

La Figura 7.16 presenta il circuito per una porta NAND a due ingressi in logica DTL. In questo caso la funzione NAND è ottenuta aggiungendo un solo diodo all'invertitore DTL di Figura 7.12, con evidenti vantaggi di compattazione del circuito. Osservando il funzionamento della rete di diodi contenuta nel blocco tratteggiato, si vede che la tensione V_X (che deve essere inferiore a 2 V per mantenere l'uscita nello stato alto) si riduce al valore V_D della caduta sul diodo di ingresso se almeno uno dei diodi D_A o D_B conduce, ossia se almeno uno degli ingressi è basso; invece V_X rimane al valore alto solo se entrambi i diodi D_A e D_B sono interdetti, ossia se tutti gli ingressi sono alti. Quella descritta è un'operazione AND tra le tensioni in ingresso e la grandezza V_X, ed è effettuata tramite la rete dei diodi D_A e D_B e la resistenza R_B; l'invertitore con il transistore Q_1 effettua solo un'inversione della variabile, per cui il risultato complessivo è quello di un'operazione NAND tra gli ingressi e l'uscita (la porta NAND può essere estesa a n ingressi semplicemente ponendo n diodi con gli anodi tutti connessi al punto X). Da questo esempio si comprende perché la porta più conveniente per la logica DTL è la porta NAND, da cui vengono costruite le altre funzioni logiche.

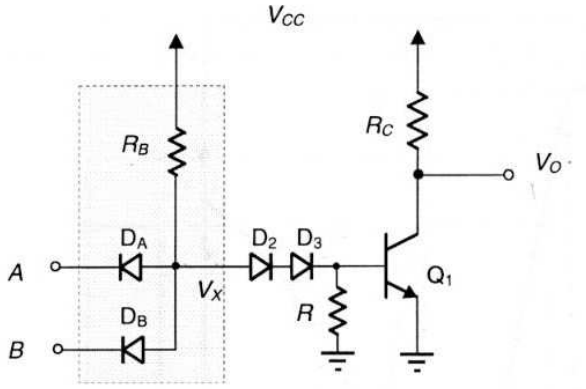


Figura 7.16 Circuito per una porta NAND DTL a due ingressi

7.9 Tracciato di una porta NAND DTL

L'esecuzione di un tracciato di una porta logica bipolare non si differenzia molto da quella per una porta logica in tecnologia MOS; anche in questo caso occorre rispettare le relative regole di progetto, che, oltre alle condizioni indicate per la tecnologia MOS, debbono tener conto anche di condizioni aggiuntive, quali quella relativa allo strato sepolto, e alle diffusioni di base e di emettitore. Una condizione molto più gravosa per i circuiti integrati bipolari nasce dalla considerazione che i dispositivi attivi non sono intrinsecamente isolati l'uno dall'altro, come avviene invece per i transistori MOS, per cui occorre contornare con regioni di isolamento tutti i transistori che presentino i collettori non connessi tra loro, per cui l'ingombro di area non è trascurabile.

Anche i componenti resistivi vanno isolati dal substrato da una regione di isolamento che li contorna, e questo aumenta sia l'occupazione di area che la complessità topologica dei collegamenti tra i diversi componenti del circuito. Per ridurre l'ingombro delle regioni di isolamento, si tende a raggruppare i diversi componenti in un ridotto numero di aree che vanno contornate dalla regione di isolamento; ad esempio si può considerare una regione che contiene i componenti resistivi, una per i diodi (in quanto questi vengono realizzati utilizzando le regioni di base e di emettitore relative al processo di realizzazione del transistore bipolare, e utilizzano la regione epitassiale N come regione di isolamento tra i diodi realizzati nella stessa area), e una per ognuno dei transistori bipolari del circuito.

In Figura 7.17 è riportato come esempio un tracciato di una porta logica NAND DTL a due ingressi, il cui schema elettrico è quello della Figura 7.16. Il circuito prevede l'utilizzo di un substrato di tipo P, con uno strato epitassiale N nel quale vengono realizzati i singoli componenti, e con regioni di isolamento ottenute per impiantazione di drogante di tipo P, che per diffusione raggiunge il substrato nelle aree in cui viene realizzata l'impiantazione, creando il cosiddetto isolamento per diffusione.

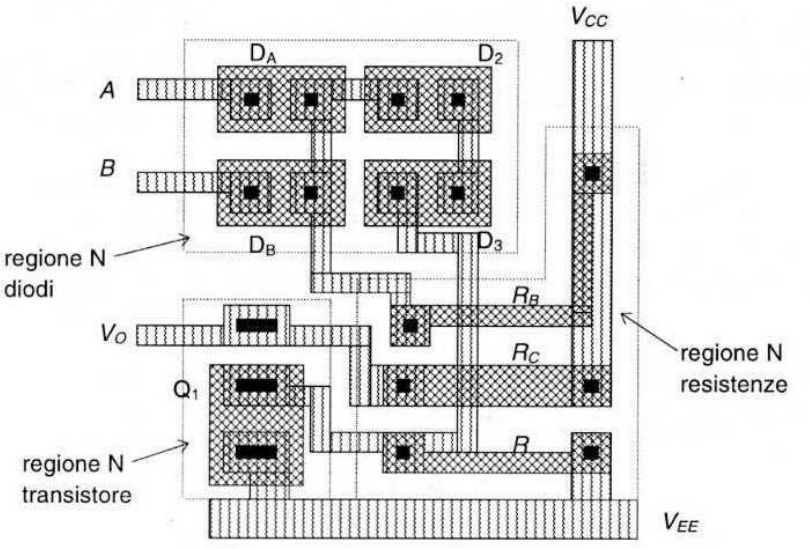


Figura 7.17 Tracciato di una porta NAND DTL a due ingressi

Le regioni di isolamento P nel tracciato della figura circondano le tre regioni N nelle quali vengono realizzati rispettivamente i diodi, le resistenze e il transistore. La tensione di alimentazione è indicata con V_{CC}, mentre la tensione di massa, in analogia con la convenzione utilizzata nei tracciati MOS, è indicata con il simbolo V_{EE}.

7.10 Porte logiche HTL

Una variante della logica DTL è la logica ad alto livello di soglia, o logica HTL (*High-Threshold Logic*), che si basa sulla sostituzione dei due diodi D_2 e D_3 dell'invertitore DTL con un diodo Zener (ricordiamo che il diodo Zener viene utilizzato in polarizzazione inversa fino al breakdown, per cui la tensione che si stabilisce ai suoi capi è la tensione di Zener, che porta ad un passaggio di corrente relativamente elevata, per effetto del campo che si crea nella regione di svuotamento in contropolarizzazione e che permette la liberazione di alcuni degli elettroni degli atomi di silicio), in modo da avere ai capi di questo elemento una tensione V_Z superiore a quella $2V_D = 1.4$ V del circuito DTL. Il circuito dell'invertitore elementare è riportato in Figura 7.18; in questo la tensione V_Z sostituisce la caduta 2 V_D, e la (7.44) che definisce la tensione V_{IL} si scrive ora:

$$V_{IL} = -V_{D1} + V_Z + V_{BE\gamma} \cong -0.7 + 6 + 0.6 = 5.9 \text{ V} \tag{7.52}$$

e la stessa correzione si applica per V_{IH} che ora vale:

$$V_{IH} = -V_{D1} + V_Z + V_{BESAT} \cong -0.7 + 6 + 0.8 = 6.1 \text{ V} \tag{7.53}$$

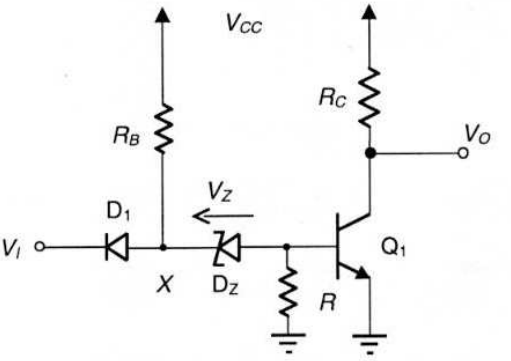


Figura 7.18 Invertitore HTL

Questi valori sono compatibili solo con una tensione di alimentazione V_{CC} maggiore di 5 V, usualmente da 12 a 15 V, che viene utilizzata per questa logica. Ne risultano margini di rumore simmetrici e notevolmente elevati (NM_H = 5.9 V; NM_L = 5.7 V per V_{CC} = 12 V) che rendono interessante l'uso di questa logica in ambienti con forti disturbi elettrici.

Esercizi di riepilogo

7.1 Per un invertitore RTL con un transistore NPN con $\beta_F = 20$, una resistenza di carico R_C = 1 kΩ, e una tensione V_{CC} = 5 V, determinare il valore della resistenza R_B che fornisce un valore di V_{OL} = 0.2 V e un valore V_{IH} = 1.8 V.

7.2 Determinare il fan-out di un invertitore RTL caratterizzato dai seguenti valori: R_C = 1 kΩ, R_B = 10 kΩ, V_{CC} = 5 V, e con un transistore con β_F = 20, assumendo un valore minimo ammissibile di V_{OH} = 4 V.

7.3 Disegnare lo schema elettrico di una porta NOR RTL a tre ingressi.

7.4 Disegnare lo schema elettrico di una porta NAND RTL a due ingressi.

7.5 Per l'invertitore RTL definito dai parametri dell'Esercizio 7.2, con un transistore definito da $\tau_F = 0.06$ ns, $\tau_R = 10$ ns, $\beta_F = 20$, determinare il tempo di propagazione t_{PLH} mediante le formule analitiche approssimate, e paragonare i risultati con quelli ottenuti con una simulazione SPICE, adottando per gli altri parametri del transistore quelli della scheda .MODEL riportata nell'Appendice.

7.6 Valutare per via analitica il ritardo di propagazione, e la potenza dissipata per un invertitore RTL con i seguenti parametri: $R_B = 10$ kΩ, $V_{CC} = 5$ V, $\beta_F = 20$, $\tau_F = 0.06$ ns, $\tau_R = 5$ ns, e per valori di R_C compresi tra 0.5 e 5 kΩ. Determinare per punti il valore di R_C che corrisponde al minimo del prodotto ritardo-potenza.

7.7 Per l'invertitore RTL con i parametri dell'Esercizio 7.2, determinare le variazioni dei livelli logici e del fan-out se il β_F varia di ± 20% rispetto al valore nominale di 20.

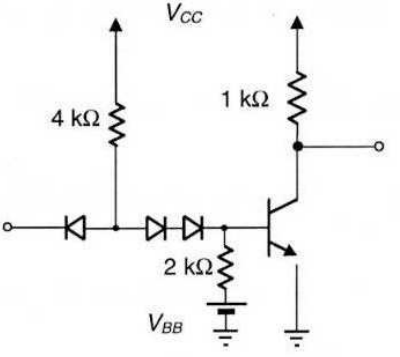


Figura E7.1

7.8 Per l'invertitore DTL riportato in Figura E7.1, con i seguenti parametri per il transistore: $\beta_F = 50$, $\tau_F = 0.06$ ns, $\tau_R = 5$ ns, e con $V_{CC} = 5$ V, $V_{BB} = 0$, determinare mediante simulazione SPICE il ritardo di propagazione. Ripetere la simulazione per un valore $V_{BB} = -2$ V. Giustificare la diminuzione del ritardo di propagazione nel secondo caso.

7.9 Determinare mediante analisi SPICE l'andamento della caratteristica di ingresso dell'invertitore DTL di Figura E7.1. Determinare per via analitica approssimata il valore e il verso della corrente di ingresso dell'invertitore quando questo è pilotato rispettivamente dal livello logico nominale alto o basso.

7.10 Per l'invertitore DTL di Figura E7.1 con i parametri del transistore dell'Esercizio 7.8, determinare il massimo numero di invertitori che possono essere

connessi in uscita mantenendo l'invertitore in saturazione quando l'uscita è al livello logico basso.

7.11 Valutare, mediante analisi SPICE, l'effetto di un carico capacitivo C_L = 0.1, 0.5, 1 pF sul ritardo di propagazione dell'invertitore DTL dell'Esercizio 7.8.

7.12 Disegnare lo schema elettrico di una porta NOR DTL a due ingressi.

Riferimenti bibliografici

H. Taub, D. Schilling, *Elettronica Integrata Digitale*, Jackson, Milano, 1981.

J. Millman, *Circuiti e sistemi microelettronici*, Bollati Boringhieri, Torino, 1985.

G.M. Glansford, *Digital Electronic Circuits*, Prentice Hall, Englewood Cliffs, 1988.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

A.S. Sedra, K.C. Smith, *Microelectronic Circuits*, Saunders College, Philadelphia, 1991.

B. Riccò, F. Fantini, P. Brambilla, *Introduzione ai circuiti integrati digitali*, Zanichelli, Bologna, 1991.

J. Millman, A. Grabel, *Microelettronica*, McGraw-Hill Italia Libri, Milano, 1994.

8

Porte logiche TTL

8.1 Introduzione

Le porte logiche Transistore-Transistore (TTL) sono state sviluppate all'inizio degli anni '70 come evoluzione delle logiche DTL, da cui discendono. Questa famiglia logica ha avuto un grande successo e si è venuta via via trasformando negli anni, assorbendo molte delle innovazioni tecnologiche dei transistori bipolari che venivano via via alla luce, e generando in effetti una "famiglia" di famiglie logiche, dalle prime porte TTL della serie 74 a quelle 74S, 74LS, fino a quelle attuali, 74AS, 74ALS e 74F.

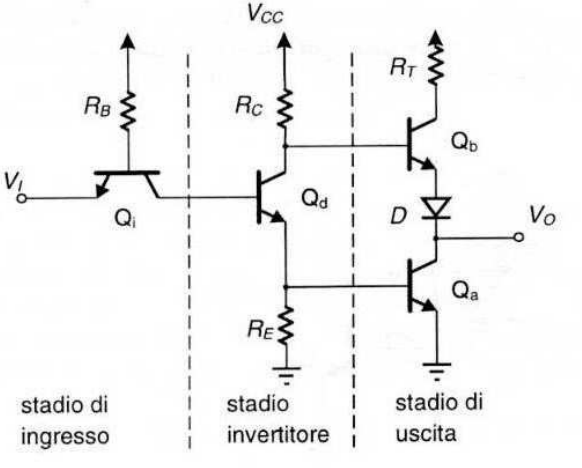


Figura 8.1 Schema circuitale dell'invertitore TTL elementare

Dal punto di vista circuitale, le porte TTL contengono una serie di ingegnose utilizzazioni delle caratteristiche dei transistori bipolari, che applicate via via nelle evoluzioni della famiglia logica ne hanno determinato il costante miglioramento delle caratteristiche sia statiche che dinamiche. Seguiremo quindi nell'esposizione e nell'analisi delle porte TTL l'evoluzione avutasi in questa famiglia, partendo dal circuito più semplice di invertitore TTL che è quello su cui è basata la prima versione di porta TTL sviluppata nel 1963, in modo da comprendere come già vi fosse un numero di modifiche di significativo rilievo concettuale rispetto alle porte DTL.

Il circuito dell'invertitore TTL è riportato in Figura 8.1. Questo è composto da tre sezioni, e cioè lo stadio di ingresso formato dal transistore Q_i, quello invertitore propriamente detto con il transistore Q_d, e lo stadio di uscita formato dai transistori Q_a e Q_b. Esamineremo separatamente le diverse parti nei paragrafi successivi.

8.2 Lo stadio di ingresso

Trascurando per un momento la resistenza R_E sull'emettitore di Q_d, le prime due sezioni sono riconducibili ad un invertitore elementare DTL in cui la rete di diodi in ingresso è sostituita dal transistore Q_i, come esemplificato nello schema di Figura 8.2. In effetti il transistore bipolare è rappresentabile con due diodi tra emettitore e collettore, connessi con gli anodi in comune sul terminale di base (ricordiamo il modello Ebers-Moll di Figura 6.8) ma con un'importante diversità (dovuta ai generatori di corrente controllati della stessa figura) e cioè la possibilità di variare sia il verso che il livello della corrente circolante tra emettitore e collettore, a seconda che il transistore funzioni in regime *diretto* o *inverso*, come si è visto nel Capitolo 6. Questa proprietà del transistore bipolare viene utilmente sfruttata mediante Q_i nella rete elementare di Figura 8.2 per migliorare l'uscita dalla saturazione del transistore Q_d nel passaggio dalla saturazione all'interdizione, e quindi per migliorare la dinamica dell'invertitore, in particolare per ridurre il tempo t_{PLH} che risulta il più gravoso per gli invertitori bipolari già visti.

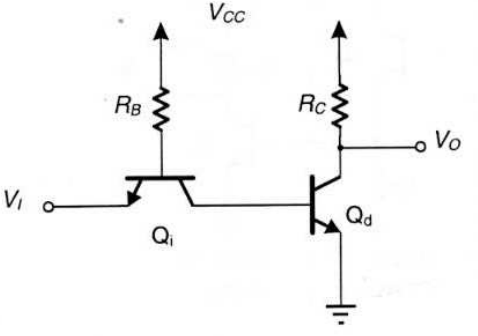


Figura 8.2 Schema dello stadio di ingresso dell'invertitore TTL elementare

Per comprendere meglio quanto detto, si può sviluppare un'analisi approssimata della rete di Figura 8.2 in condizioni dinamiche, quando il segnale logico in ingresso passa dal valore alto a quello basso (per cui in uscita dell'invertitore la tensione passerà dalla saturazione all'interdizione dopo il tempo di accumulo t_S).

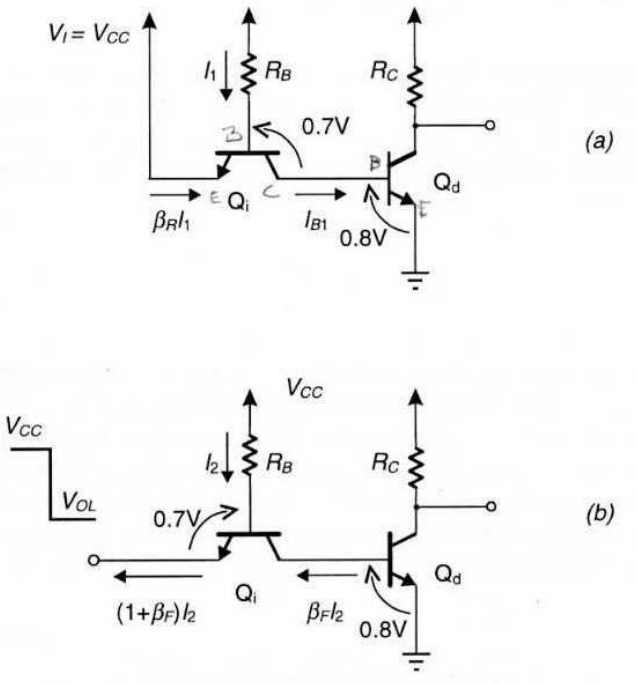


Figura 8.3 Analisi dello stadio di ingresso: a) con ingresso alto; b) durante il transitorio dovuto al passaggio dell'ingresso nello stato basso

Assumendo come valore alto la tensione di alimentazione V_{CC}, dalla Figura 8.3a si vede che, anche se non circolasse corrente in R_B, la giunzione base-emettitore di Q_i sarebbe polarizzata a 0 V e quindi sotto soglia; in realtà, poiché può circolare corrente dalla base verso il collettore, la giunzione base-emettitore è contropolarizzata e quella base-collettore direttamente polarizzata. Q_i lavora quindi in modo inverso, e la tensione al terminale di base (rispetto massa) V_{Bi} può essere facilmente determinata assumendo che Q_d sia in saturazione, per cui:

$$V_{BEd} = V_{BESAT} \cong 0.8 \text{ V}; \qquad V_{Bi} = V_{BEd} + V_{BCi} \cong 0.8 + 0.7 = 1.5 \text{ V} \tag{8.1}$$

e la corrente di base di Q_i sarà data da:

$$I_1 = \frac{V_{CC} - V_{Bi}}{R_B} \cong \frac{V_{CC} - 1.5V}{R_B} \tag{8.2}$$

La corrente di base di Q_d, ricordando che Q_i opera in regime inverso, sarà quindi:

$$I_{Bd1} = (1 + \beta_R) I_1 \cong I_1 \tag{8.3}$$

L'approssimazione nella (8.3) è giustificata in quanto nelle porte TTL si fa in modo che il transistore Q_i abbia un $\beta_R << 1$ (~ 0.02) scegliendo opportunamente le aree di emettitore e di base. La verifica che Q_d sia in saturazione può essere fatta agevolmente a questo punto applicando la disequazione (6.26) tra le correnti di base e collettore di Q_d che, in base alla (8.2) e (8.3), fornisce:

$$\frac{R_B}{\beta_{Fd} R_C} < 1 - \frac{1.5V}{V_{CC}} \tag{8.4}$$

Al passaggio del segnale di ingresso V_I dal valore alto a quello basso la situazione è quella di Figura 8.3b, in cui si assume un valore di $V_I = V_{OL} = 0.2$ V (pari alla V_{CESAT} dello stadio precedente). In questo caso la giunzione base-emettitore di Q_i viene ad essere polarizzata direttamente, e la tensione sulla base sarà $V_{Bi} = V_{OL} + V_{BEi} \cong 0.2 + 0.7 = 0.9$ V, mentre la giunzione base-collettore è polarizzata sotto V_γ. Q_i opera quindi in modo attivo diretto, con una corrente di base I_2 (il pedice 2 sta ad indicare la nuova situazione con ingresso basso) data da:

$$I_2 = \frac{V_{CC} - V_{BEi} - V_{OL}}{R_B} \cong \frac{V_{CC} - 0.9V}{R_B} \tag{8.5}$$

e questa corrente comporta una corrente entrante nel collettore di Q_i pari a $\beta_F I_2$. Quest'ultima è anche la corrente I_{Bd2} *uscente* dalla base di Q_d, ossia:

$$I_{Bd2} = \beta_{Fi} I_2 = -\beta_{Fi} I_1 \frac{V_{CC} - 0.9V}{V_{CC} - 1.5V} = -\beta_{Fi} I_1 \cdot 1.17 \tag{8.6}$$

La corrente I_{Bd2} estratta dalla base di Q_d provoca una riduzione significativa del tempo di accumulo; quando tutta la carica di base è estratta, Q_d si interdice, la tensione di base scende sotto il valore di soglia $V_{BE\gamma}$ e la corrente di base si annulla. Il tempo di accumulo t_S dato dalla (7.31), ricordando le espressioni di I_{Bd1} e I_{Bd2}, vale:

$$t_S = \tau_S \ln\left(\frac{I_{Bd1} - I_{Bd2}}{\frac{I_{CSAT}}{\beta_{Fd}} - I_{Bd2}}\right) = \tau_S \ln\left(\frac{(1 + 1.17 \cdot \beta_{Fi}) I_1}{\frac{I_{CSAT}}{\beta_{Fd}} + 1.17 \cdot \beta_{Fi} I_1}\right) \tag{8.7}$$

e analogamente il tempo t_{LH}, definito in generale dalla (7.36), vale:

$$t_{LH} = \tau_{BF} \ln\left(\frac{I_{CSAT} - \beta_{Fd} I_{Bd2}}{I_{CSAT}/2 - \beta_{Fd} I_{Bd2}}\right) = \tau_{BF} \ln\left(\frac{I_{CSAT} + 1.17 \cdot \beta_{Fd}\beta_{Fi} I_1}{I_{CSAT}/2 + 1.17 \cdot \beta_{Fd}\beta_{Fi} I_1}\right) \tag{8.8}$$

Utilizzando queste espressioni per il caso di Figura 8.2 con $R_C = 1.6$ kΩ, $R_B = 4$ kΩ, e con i seguenti parametri dei transistori Q_i e Q_d:

$$\beta_{Fi} = 20; \quad \beta_{Fd} = 50; \quad \tau_S = 10 \text{ ns}; \quad \tau_{BF} = 5 \text{ ns}$$

si ottiene dalla (8.2) $I_1 = 0.87$ mA e $I_{CSAT} = 3$ mA, e dalle (8.7) e (8.8) si ha per i tempi di commutazione: $t_S = 0.4$ ns, $t_{LH} \cong 0$.

Questo esempio, ancorché basato su un'analisi approssimata, mostra come il transistore Q_i effettui un'efficace estrazione di corrente dalla base di Q_d e quindi permetta un'interdizione molto veloce di quest'ultimo.

Ciò ovviamente non vuol dire che il tempo di propagazione t_{PLH} sia trascurabile, perché, come già indicato nel Paragrafo 7.7, questa analisi prescinde dalla presenza di un'inevitabile capacità di uscita che, dovendosi caricare attraverso la resistenza R_C, determina un tempo di salita di solito non trascurabile anche in presenza di un transistore che idealmente commutasse in un tempo nullo. Per ovviare a questo problema, comune a tutti gli invertitori con carico resistivo, si è introdotto per la porta TTL lo stadio di uscita indicato in Figura 8.1, che verrà analizzato nel seguito.

8.3 Lo stadio di uscita

Come si è visto nell'analisi degli invertitori con carico resistivo e con una capacità in uscita, il tempo di scarica della capacità attraverso il transistore in saturazione (che presenta una bassa resistenza), tramite la corrente di scarica I_{HL}, è molto minore di quello di carica, che avviene attraverso la resistenza di carico, tramite la corrente di carica I_{LH}; questi tempi sono identificati nei diagrammi della Figura 8.4a. È possibile spostare la resistenza R_C dal collettore all'emettitore, come è indicato in Figura 8.4b, nel qual caso è il transistore ad agire come carico della capacità nella fase di carica della stessa, mentre la scarica avviene attraverso la resistenza. In questo secondo caso sarà il tempo di carica ad essere più breve perché questa avviene attraverso la resistenza equivalente del transistore in saturazione e quindi con una corrente di carica I_{LH} relativamente elevata, mentre la scarica sarà relativamente più lenta perché si sviluppa attraverso la resistenza R_E, e con una corrente di carica I_{HL}

limitata. Si noti che in questo secondo caso il circuito non inverte il segnale, in altre parole un segnale di tensione alta in ingresso corrisponde ad un valore alto anche in uscita.

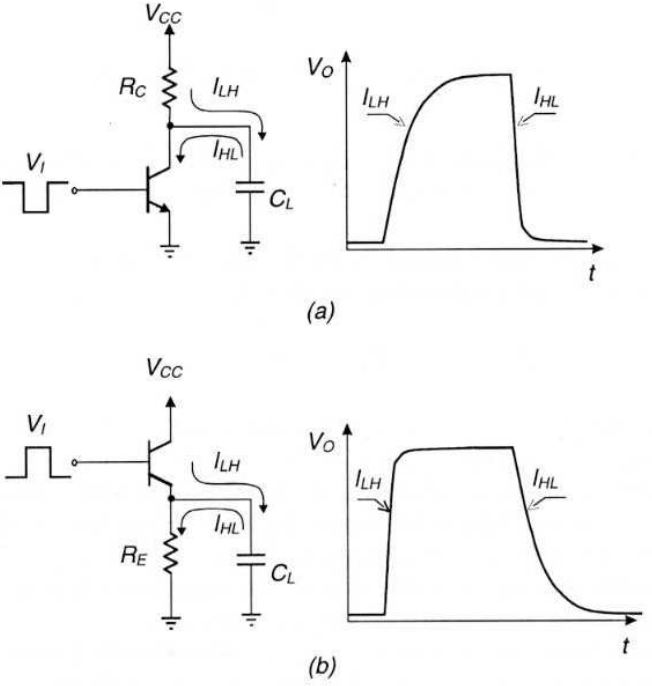


Figura 8.4 Transitorio in uscita da un invertitore con a) resistenza di carico sul collettore, b) resistenza di carico sull'emettitore

Dall'osservazione del comportamento complementare di questi due circuiti scaturisce la possibilità di utilizzarli *entrambi* per la carica e scarica della capacità in uscita, mediante il circuito di Figura 8.5. Questo circuito viene denominato *totem pole*, perché, in analogia con le figure dei pali totem dei nativi americani, presenta il transistore Q_b "seduto" su quello Q_a. Il circuito permette di ridurre sia il tempo di scarica che quello di carica, perché in entrambi i casi la capacità vede la resistenza equivalente di un transistore in saturazione che è molto bassa. In questo circuito infatti, per:

V_{Ia} = 0 logico $\Rightarrow$ Q_a interdetto
V_{Ib} = 1 logico $\Rightarrow$ Q_b in saturazione $\Rightarrow$ V_O = 1 logico

mentre per:

V_{Ia} = 1 logico $\Rightarrow$ Q_a in saturazione
V_{Ib} = 0 logico $\Rightarrow$ Q_b interdetto $\Rightarrow$ V_O = 0 logico

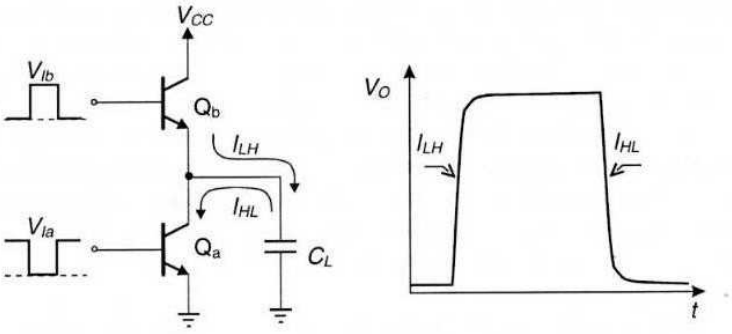


Figura 8.5 Configurazione *totem pole* per lo stadio di uscita

Un ulteriore significativo vantaggio di questo circuito è quello di avere uno dei due transistori interdetto, sia quando esso si trova nello stato alto che in quello basso, e quindi la dissipazione di potenza statica risulta nulla, in maniera simile a quanto avviene per l'invertitore CMOS. Rispetto a quest'ultimo, tuttavia, vi è la complicazione di dovere utilizzare per il comando due segnali opposti in fase, essendo i due dispositivi dello stesso tipo, e non complementari come nel caso CMOS.

I due segnali complementari di comando per Q_a e Q_b possono essere prelevati dal transistore Q_d mediante due resistenze di carico R_C e R_E, poste rispettivamente sul collettore e sull'emettitore. Poiché la tensione sul collettore di Q_d è data da $V_{CC} - R_C I_C$ e quella su R_E è $R_E I_E$, ricordando che: $I_C \cong I_E$, si comprende come i segnali su queste due resistenze siano in opposizione di fase tra loro e con i moduli dipendenti dai valori delle resistenze.

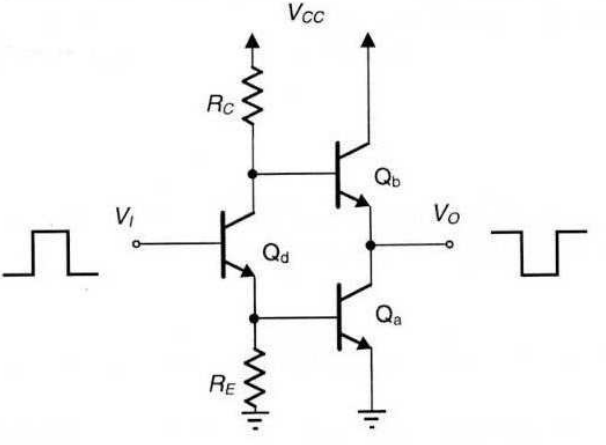


Figura 8.6 Pilotaggio dello stadio di uscita con il transistore Q_d

Il circuito totem pole può quindi essere pilotato dalle due uscite di Q_d come indicato schematicamente in Figura 8.6. Rispetto al segnale di ingresso di Q_d l'uscita del totem pole risulta invertita, per cui la cascata dello stadio Q_d e di quello di uscita si comporta come un semplice stadio invertitore.

Anche per il circuito di uscita occorre verificare se i transistori Q_a e Q_b possano commutare rapidamente nel passaggio dalla saturazione all'interdizione; ricordiamo che la velocità di commutazione dipende da un'efficace estrazione delle cariche immagazzinate, estrazione che, come si è detto precedentemente, per il transistore Q_d viene effettuata attraverso il transistore Q_i che opera in regime diretto.

Dall'osservazione del circuito di Figura 8.6 si può notare che in questo caso il ruolo svolto da Q_i nello stadio di ingresso viene ora effettuato dal transistore Q_d nei riguardi del transistore di uscita Q_b; quando Q_d passa dall'interdizione alla conduzione, nel suo collettore viene assorbita anche la corrente transitoria di base di Q_b che si trova in saturazione e che viene rapidamente portato in interdizione dall'elevata corrente I_{Bb2} uscente dalla base di Q_b. Il problema dell'estrazione delle cariche si pone invece per il transistore Q_a per il quale il passaggio dalla saturazione all'interdizione avviene quando Q_d passa all'interdizione, mediante una (relativamente debole) corrente di estrazione di base $I_{Ba2} = -V_{BESAT}/R_E$, come nel caso dell'invertitore RTL.

In questo caso, tuttavia, vi è un altro fenomeno che contribuisce ad accelerare la dinamica del passaggio dalla saturazione all'interdizione, e cioè l'aumento della corrente di collettore di Q_a durante la transizione basso-alto rispetto al valore di regime. Infatti nel passaggio del segnale di ingresso V_I dal valore alto a quello basso, il transistore Q_b passa rapidamente in conduzione, mentre Q_a (che stava in saturazione) non esce istantaneamente dalla saturazione; quindi durante il transitorio legato al tempo di accumulo t_S fluisce un'elevata corrente dall'emettitore di Q_b nel collettore di Q_a (vedi Figura 8.7), che contribuisce ad eliminare le cariche immagazzinate nella base. La corrente transitoria fornita da Q_b può essere molto elevata perché durante tutto il tempo di accumulo di Q_a il transistore Q_b opera in regione attiva ($V_{CEb} = V_{OH} - V_{OL} >> V_{CESAT}$) per cui:

$$I_{Eb} \cong I_{Ca} \cong (\beta_{Fb} + 1) I_{Bb1} \tag{8.9}$$

Il tempo di accumulo, con questo valore della corrente di collettore di Q_a, si riduce, dall'espressione (7.31) valida per un invertitore RTL, a:

$$t_S = \tau_S \ln\left(\frac{I_{Ba1} - I_{Ba2}}{I_{Ca} / \beta_F - I_{Ba2}}\right) \tag{8.10}$$

dove il termine I_{Ca}/β_F sostituisce quello I_{BSAT} della (7.31), con riduzione significativa del tempo di accumulo. Questo effetto è chiaramente visibile nelle forme d'onda delle correnti e tensioni di Q_a riportate in Figura 8.7 che mostrano come il tempo di

accumulo si riduca in presenza della corrente transitoria I_{Eb} fornita da Q_b, rispetto al caso dell'invertitore RTL.

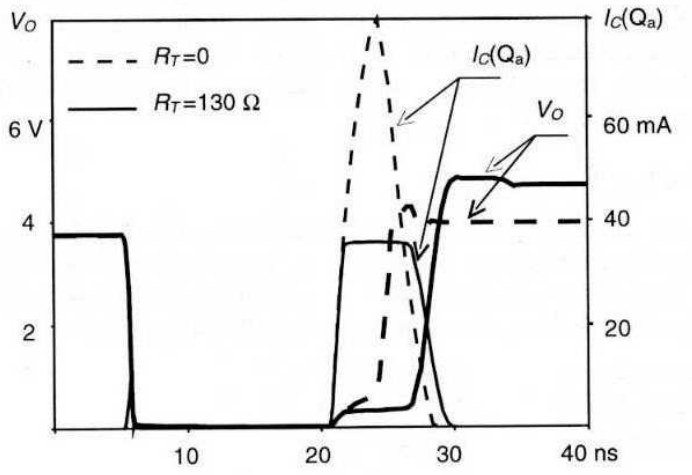


Figura 8.7 Dinamica della transizione dell'uscita dal valore basso a quello alto. Sono riportati gli andamenti di V_O e di $I_C(Q_a)$ nel caso di uscita non protetta o con una resistenza di protezione $R_T = 130\ \Omega$

Tuttavia la corrente I_{Eb} non può essere troppo elevata perché potrebbe danneggiare il transistore Q_b per eccessiva potenza di picco dissipata; ad esempio con $\beta_F = 50$ e $I_{Bb1} = 2$ mA si avrebbe $I_{Eb} = 100$ mA e la potenza istantanea sarebbe di circa 500 mW. La corrente di Q_b viene ridotta inserendo in serie al collettore di Q_b una resistenza R_T di protezione, che porta Q_b in saturazione e limita la corrente massima al valore:

$$I_{CM} = \frac{V_{CC} - 2V_{CESAT}}{R_T} \tag{8.11}$$

La scelta del valore di R_T è legata quindi ad un compromesso tra velocità di commutazione e sicurezza del transistore; usualmente si adotta il valore di 130 Ω, in modo da limitare I_{Cb} ad un valore inferiore a 50 mA (vedi Figura 8.7) .

8.4 Caratteristica di trasferimento dell'invertitore TTL

Ritornando allo schema completo dell'invertitore TTL elementare riportato in Figura 8.1, notiamo che rispetto al circuito di uscita a totem discusso precedentemente vi è l'aggiunta del diodo D tra i transistori Q_b e Q_a; vedremo che questo diodo è

necessario per garantire l'interdizione del transistore Q_b quando il segnale di ingresso è alto.

Ricaviamo ora la caratteristica di trasferimento di questo invertitore utilizzando un'analisi statica approssimata basata sulle assunzioni di Tabella 6.2 per il funzionamento dei transistori nei diversi regimi di funzionamento, assumendo per le resistenze i seguenti valori, tipici di porte TTL della famiglia 74: R_B = 4 kΩ, R_C = 1.6 kΩ, R_E = 1 kΩ, R_T = 130 Ω. Nella caratteristica si identificano 4 regioni (Figura 8.16):

Regione I: $0 < V_I < V_{IL}$

Si è già visto (Paragrafo 8.2) che con il segnale di ingresso al livello logico basso il transistore Q_i opera in modo diretto. Poiché a regime la corrente estratta dalla base di Q_d è nulla, assumiamo per l'analisi del circuito in questa condizione (verificheremo la congruenza delle assunzioni con i risultati dell'analisi) che:

Q_d sia in interdizione	$(V_{BEd} < V_{BE\gamma})$
Q_a sia in interdizione	$(V_{BEa} < V_{BE\gamma})$
Q_b sia in conduzione	$(V_{BEb} > V_{BE\gamma})$

Il circuito in queste condizioni è riportato in Figura 8.8, dove si è evidenziata in grigio l'area inattiva, essendo i transistori interdetti.

In base a queste assunzioni, dalla Figura 8.8 si ricava il valore della corrente I che fluisce nella base di Q_i:

$$I = \frac{V_{CC} - V_{BEi} - V_I}{R_B} \cong \frac{V_{CC} - 0.8 - V_I}{R_B} \tag{8.12}$$

Poiché la corrente di collettore di Q_i è trascurabile, il transistore Q_i è in forte saturazione, e si assume $V_{CESAT} \cong 0.1$ V. La tensione sulla base di Q_d vale quindi:

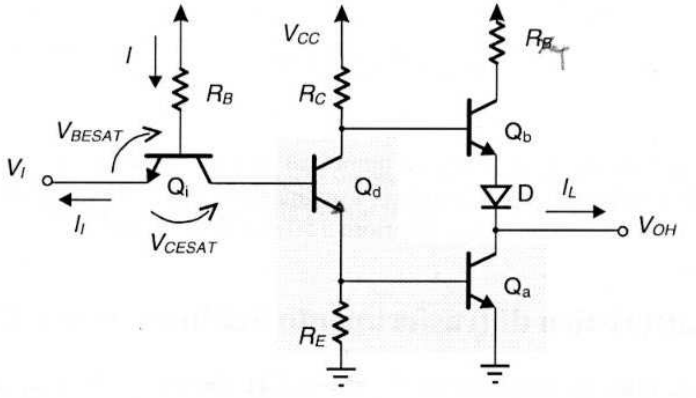


Figura 8.8 Analisi del circuito per $V_I < V_{IL}$; i valori del circuito sono: R_B = 4 kΩ, R_C = 1.6 kΩ, R_E = 1 kΩ, R_T = 130 Ω

$$V_{BEd} = V_I + V_{CESATi} < V_{BE\gamma} \tag{8.13}$$

e si verifica che Q_d è in interdizione finché $V_I < 0.5$ V. La corrente $I_{Ed} = 0$, e quindi anche la tensione V_{BEa} di ingresso a Q_a è nulla, per cui anche l'assunzione su Q_a è verificata.

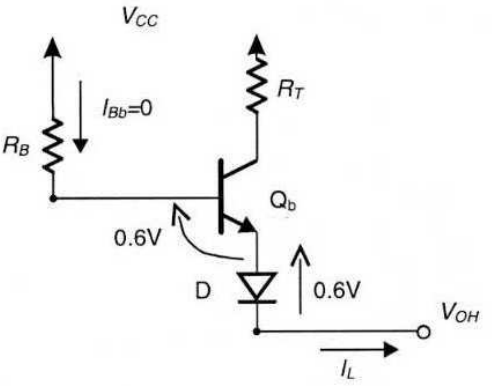


Figura 8.9 Circuito di uscita per $V_I < V_{IL}$

Il circuito si riduce quindi alla rete di Figura 8.9, che, nell'assunzione di carico nullo in uscita (e quindi di corrente I_L trascurabile), fornisce:

$$I_{Bb} = \frac{I_L}{\beta_F + 1} \cong 0; \quad V_O = V_{CC} - V_{BE\gamma} - V_D \cong V_{CC} - 1.2 \text{ V} = 3.8 \text{ V} \tag{8.14}$$

Q_b si trova al limite della conduzione; vedremo successivamente l'influenza della corrente I_L sul regime di funzionamento di Q_b, nel caso di carico non trascurabile.

Regione II: $V_{IL} < V_I < V_{L'}$

Al crescere di V_I oltre il valore V_{OL}, in base alla (8.13) la tensione di base di Q_d cresce, e per $V_I = V_{IL}$ questa raggiunge il valore $V_{BE\gamma}$, limite dell'interdizione; vedremo che oltre questo valore la caratteristica di trasferimento assume una pendenza maggiore dell'unità, per cui l'inizio della conduzione di Q_d corrisponde alla definizione di V_{IL}. Questa quindi vale:

$$V_{IL} = V_{BE\gamma(d)} - V_{CESAT(i)} \cong 0.6 - 0.1 = 0.5 \text{ V} \tag{8.15}$$

All'aumentare di V_I oltre il valore V_{IL} Q_d entra in conduzione; le condizioni di funzionamento assunte per i quattro transistori in questa regione di funzionamento sono:

Q_i in saturazione	(modo diretto)
Q_d in conduzione	$(V_{BE} > V_{BE\gamma})$
Q_b in conduzione	$(V_{BE} > V_{BE\gamma})$
Q_a in interdizione	$(V_{BE} < V_{BE\gamma})$

Il circuito cui fare riferimento in questo caso è quello di Figura 8.10, dove è indicato in grigio il transistore Q_a che è interdetto.

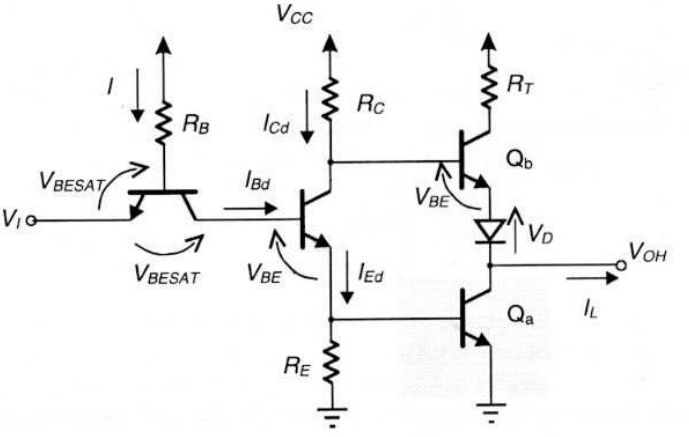


Figura 8.10 Analisi del circuito per $V_{IL} < V_I < V_L'$

Man mano che aumenta V_I, aumenta anche la corrente I_E uscente dall'emettitore di Q_d, ma finché la caduta su R_E non raggiunge 0.6 V il transistore Q_a non entra in conduzione. Definiamo V_L' il valore della tensione di ingresso per cui Q_a inizia a condurre; dal circuito di Figura 8.10 V_L' è dato da:

$$V_L' = V_{BE(d)} + V_{BE\gamma(a)} - V_{CESAT(i)} \cong 0.7 + 0.6 - 0.1 = 1.2 \text{ V} \tag{8.16}$$

In questo intervallo di valori della tensione di ingresso V_I la caratteristica di trasferimento (vedi Figura 8.13) presenta una tensione di uscita V_O che dipende dall'ingresso in maniera lineare, con una pendenza maggiore dell'unità (ciò permette di definire come V_{IL} la tensione per cui si ha l'inizio della conduzione di Q_d); la pendenza può essere determinata valutando il rapporto $\Delta V_O / \Delta V_I$ in questa regione, in altre parole la sua amplificazione, che è determinata essenzialmente dal transistore Q_d perché Q_b opera nella connessione a collettore comune (e quindi con guadagno unitario). Infatti le tensioni V_O e V_I (vedi Figura 8.10) sono date da:

$$V_O = V_{CC} - R_C I_{Cd}(V_I) - V_{BEb} - V_D \tag{8.17}$$

$$V_I = R_E I_{Ed}(V_I) + V_{BEd} - V_{CESATi} \tag{8.18}$$

dove le correnti di collettore I_{Cd} e di emettitore I_{Ed} sono gli unici termini che dipendono da V_I. L'amplificazione del circuito sarà quindi data, ricordando i valori assunti per le resistenze del circuito, da:

$$\frac{\Delta V_O}{\Delta V_I} \equiv A_V = -\frac{\Delta I_{Cd} R_C}{\Delta I_{Ed} R_E} \cong -\frac{R_C}{R_E} = -1.6 \tag{8.19}$$

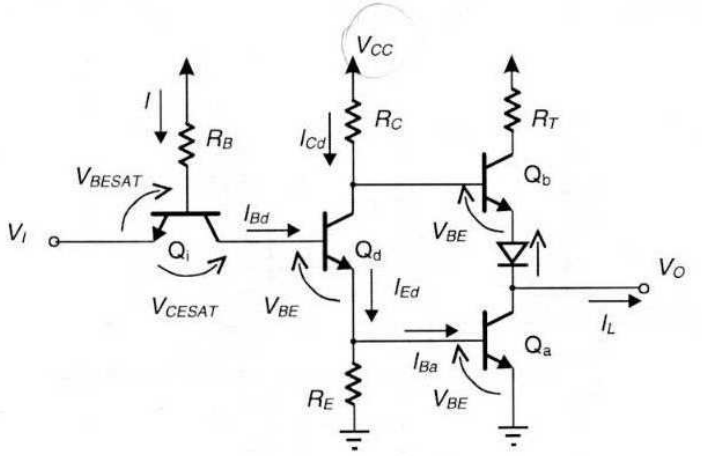


Figura 8.11 Analisi del circuito per $V_L' < V_I < V_{IH}$

Regione III: $V_L' < V_I < V_{IH}$

Nella 3ª regione comincia a condurre anche il transistore Q_a. La pendenza della caratteristica di trasferimento in questa regione è molto più elevata che nella 2ª regione, perché ora l'amplificazione è data dal prodotto delle due amplificazioni, quella relativa al transistore Q_d e quella relativa a Q_a (quest'ultimo funzionante nella connessione ad emettitore comune e quindi con amplificazione elevata). Per $V_I = V_{IH}$ si raggiunge la saturazione di Q_a; a questo punto la tensione sul collettore di Q_i diventa praticamente costante, e Q_i passa dal modo diretto a quello inverso. La tensione V_{IH} (dalla Figura 8.11) è data da:

$$V_{IH} = -V_{CESATi} + V_{BEd} + V_{BESATa} \cong 0 + 0.7 + 0.8 \cong 1.5 \text{ V} \tag{8.20}$$

dove si è assunta una tensione V_{CESATi} approssimativamente nulla in corrispondenza del passaggio di Q_i dal modo diretto a quello inverso.

Regione IV: $V_I > V_{IH}$

In questa regione Q_i passa dal funzionamento in modo diretto al modo inverso, in quanto la tensione di collettore non può crescere oltre il valore $V_{BESATd} + V_{BESATa}$ mentre quella di emettitore cresce con V_I; la giunzione base-emettitore diventa

quindi inversamente polarizzata e quella base-collettore direttamente polarizzata, come si è visto nell'analisi di Figura 8.3, portando Q_i a lavorare in regione attiva inversa. Si assume quindi:

Q_i in modo inverso	$(V_{BC} > V_\gamma)$
Q_d in saturazione	$(V_{BE} = V_{BESAT})$
Q_b in interdizione	$(V_{BE} < V_{BE\gamma})$
Q_a in saturazione	$(V_{BE} = V_{BESAT})$

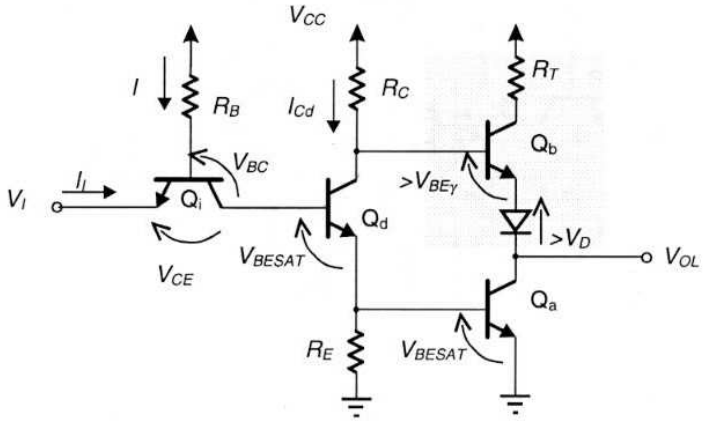


Figura 8.12 Analisi per $V_I > V_{IH}$

Occorre verificare sia l'assunzione su Q_d che quella su Q_b che potrebbe non essere verificata. Dal circuito di Figura 8.12, nell'ipotesi di saturazione di Q_d si ricavano le tensioni di base e di collettore corrispondenti:

$$V_{Bd} = V_{BESATd} + V_{BESATa} = 1.6 \text{ V}; \; V_{Cd} = V_{BESATa} + V_{CESATd} = 1 \text{ V} \Rightarrow V_{BCd} \cong 0.6 \text{ V}$$

da cui si vede che l'assunzione di Q_d in saturazione è consistente. Poiché anche Q_a è in saturazione, $V_O = V_{CESAT} = 0.2$ V; quindi la differenza di potenziale tra la base di Q_b e l'uscita è pari a: 1V – 0.2V = 0.8 V. Questa tensione è inferiore alla somma delle tensioni minime $V_{BE\gamma} + V_D = 1.3$ V per la conduzione di Q_b, per cui Q_b è interdetto, come si voleva. Quest'ultima analisi conferma la necessità dell'inserzione del diodo D per garantire l'interdizione di Q_b con l'ingresso nello stato alto.

La caratteristica di trasferimento ottenuta con l'analisi approssimata su esposta è tracciata in Figura 8.13a, mentre in Figura 8.13b si riporta per confronto il risultato ottenuto dalla simulazione SPICE del circuito; la differenza nei valori di V_{OL} si spiega considerando che il valore di 0.2 V dell'analisi approssimata è giustificato in

presenza di carico significativo, mentre nella simulazione si è considerato un carico molto basso di 100 kΩ.

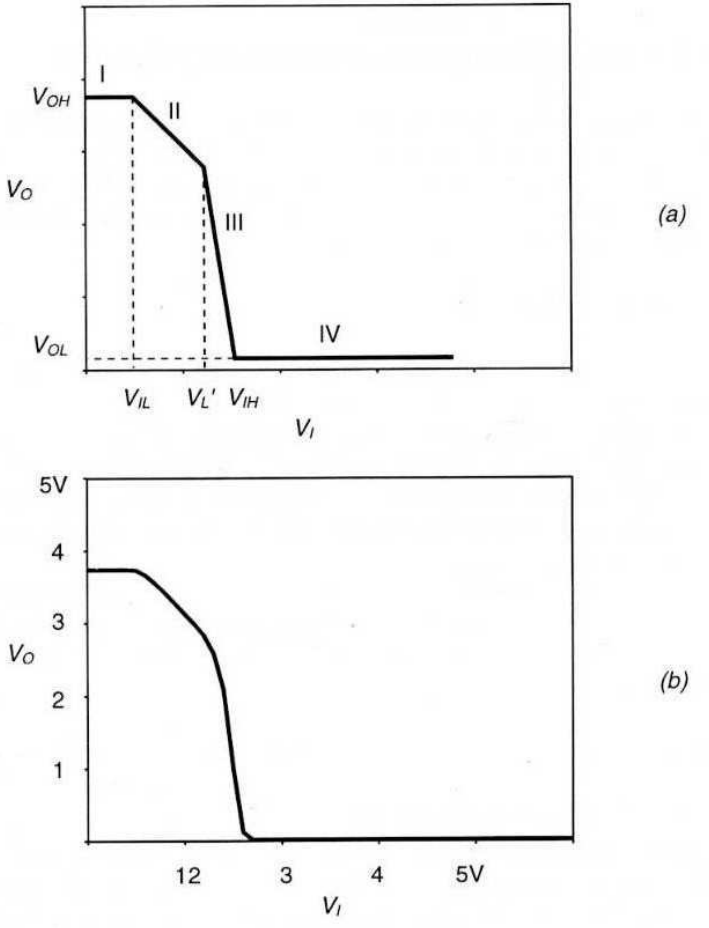


Figura 8.13 a) Caratteristica di trasferimento secondo l'analisi approssimata del circuito TTL; b) simulazione SPICE con carico di 100kΩ

I margini di rumore dell'invertitore TTL, definiti in base all'analisi sviluppata precedentemente, sono:

$$NM_L = V_{IL} - V_{OL} \cong 0.5 - 0.2 = 0.3 \text{ V} \tag{8.21}$$

$$NM_H = V_{OH} - V_{IH} \cong 3.8 - 1.5 = 2.3 \text{ V} \tag{8.22}$$

Il margine di rumore nello stato basso è molto ridotto e limita notevolmente le prestazioni della porta; vedremo che con le versioni modificate delle porte TTL è possibile migliorare significativamente questo valore.

8.5 Caratteristiche di ingresso e di uscita e fan-out

Poiché nelle porte bipolari vi è circolazione di corrente sia in ingresso che in uscita anche in condizioni stazionarie, a causa della resistenza finita di ingresso, per la caratterizzazione delle porte è necessario conoscere le caratteristiche di ingresso e di uscita, definite rispettivamente come le dipendenze tra corrente e tensione di ingresso e quelle tra corrente e tensione di uscita.

8.5.1 Caratteristica di ingresso

La caratteristica di ingresso rappresenta la dipendenza della corrente di ingresso I_I della porta dalla tensione V_I applicata. Nell'analisi del Paragrafo 8.4 si è visto che il transistore Q_i opera in modo diretto se l'ingresso è a livello logico basso, mentre è in modo inverso quando V_I è a livello logico alto. Ciò comporta che la corrente di ingresso, che coincide in modulo con quella di emettitore di Q_i ($I_I = -I_{Ei}$), sarà uscente dalla porta nel primo caso, ed entrante nel secondo.

Partendo da un valore nullo della tensione di ingresso V_I, e all'aumentare di V_I, poiché Q_i è in saturazione, $I_{Ei} = I_{Ci} + I_{Bi}$, ed essendo $I_{Ci} \cong 0$, si ha $I_{Ei} = I_{Bi}$. Ricordando la (8.12) per la corrente di base di Q_i, si può esprimere la corrente di ingresso in funzione di V_I come:

$$I_I' = -I_{Ei} \cong -I = -\frac{V_{CC} - V_{BEi} - V_I}{R_B} \tag{8.23}$$

da cui si vede che la dipendenza di I_I da V_I è lineare secondo l'inverso di R_B (vedi Figura 8.14); con il valore assunto per $R_B = 4$ kΩ, per $V_I = V_{OL}$ la corrente di ingresso *uscente* è di circa 1 mA; l'apice ' sta ad indicare che la (8.23) vale solo finché Q_i opera in regime diretto.

Quando V_I supera il valore V_{IH}, Q_i passa ad operare in modo inverso, e la corrente *entrante* nell'emettitore, con l'analisi di Figura 8.12 vale:

$$I_I'' = \beta_R I = \beta_R \frac{V_{CC} - 2V_{BESAT} - V_{BCi}}{R_B} \cong \beta_R \frac{V_{CC} - 2.3V}{R_B} \tag{8.24}$$

dove l'apice " sta ora ad indicare il campo di valori corrispondenti al funzionamento di Q_i in modo inverso. Ricordando che β_R per Q_i è molto minore dell'unità (si è assunto 0.02 nell'analisi), I_I " è dell'ordine delle decine di μA.

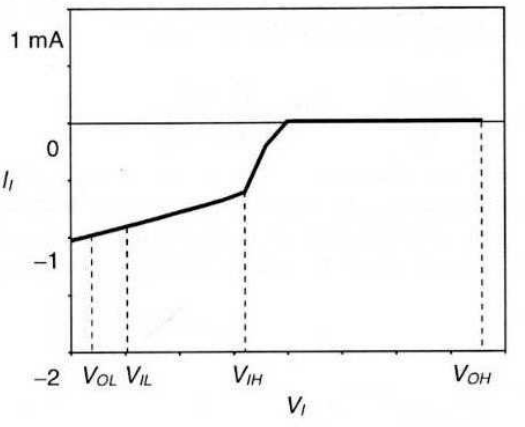


Figura 8.14 Caratteristica di ingresso della porta TTL

Dall'analisi precedente emerge che in una connessione in cascata di porte TTL, le porte connesse in uscita si comportano da carico per la porta di pilotaggio (cioè assorbono una corrente dall'uscita di questa) se l'uscita della porta di pilotaggio è alta (come indicato in Figura 8.15a), mentre agiscono come generatori di corrente (cioè iniettano una corrente nell'uscita) se l'uscita è bassa (Figura 8.15b).

8.5.2 Caratteristiche di uscita

Dall'analisi del circuito a totem di uscita della porta TTL si comprende che vi sono due differenti caratteristiche di uscita: la prima è relativa al caso di uscita al valore alto V_{OH} (Q_b in conduzione), la seconda è relativa al caso di uscita al valore basso V_{OL} (Q_a in conduzione).

Nel primo caso il circuito a cui fare riferimento è quello di Figura 8.9, già esaminato per il caso di corrente $I_L \equiv 0$. Con corrente di carico trascurabile si può verificare che Q_b lavora in regione attiva perché $V_{CEb} = V_{CC} - (V_{OH} + V_D) = 0.6$ V, per cui: $V_{OH} = V_{CC} - R_C I_B - V_{BE} - V_D$. Ricordando che $I_B = I_I/(\beta_F + 1)$, si ha per la tensione di uscita V_{OH}:

$$V_{OH} = V_{CC} - \frac{R_C}{\beta_F + 1} I_L - V_{BE} - V_D \tag{8.25}$$

Il termine $R_B/(\beta_F + 1)$ fornisce la pendenza della caratteristica di uscita nella prima regione (vedi Figura 8.16) in cui Q_b lavora in regione attiva. All'aumentare di I_L si raggiunge un valore I_L^* tale che Q_b entra in saturazione. Questo valore si ricava dall'analisi della maglia base-collettore di Q_b in cui si può scrivere la relazione:

$$R_c I_B + V_{BC} - R_T I_C = 0$$

da cui, ricordando che in regione attiva $I_C = \beta_F I_B$ e $I_L = I_B + I_C$, si ricava il valore di I_L^* al limite della saturazione, imponendo $V_{BC\gamma} = 0.6$ V:

$$I_L^* = \frac{\beta_F + 1}{\beta_F} I_C = 0.6 \frac{\beta_F + 1}{\beta_F R_T - R_C} \tag{8.26}$$

Per $I_L > I_L^*$, Q_b entra in saturazione e va sostituito nel circuito di Figura 8.9 con un generatore di tensione V_{CESAT} tra i terminali di uscita; da questo si ricava la tensione di uscita V_{OH}, che, per valori di $R_B >> R_T$, mostra una dipendenza lineare della da I_L con coefficiente $\sim R_T$:

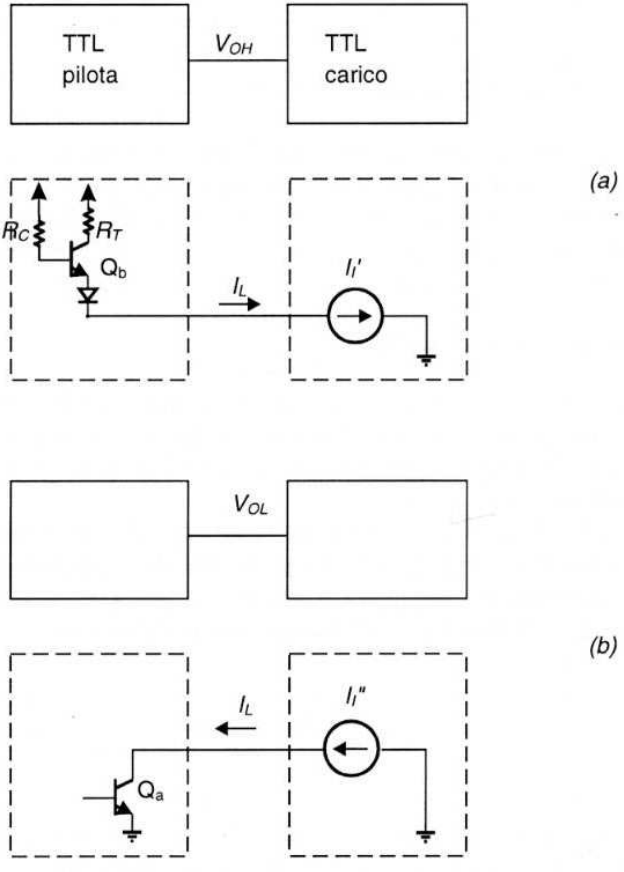


Figura 8.15 Connessione di due porte TTL in cascata: a) circuito equivalente dell'uscita per uscita alta; b) circuito equivalente per uscita bassa

$$V_{OH} = V_{CC} - \frac{R_C V_{CESAT} + R_T V_{BE}}{R_C + R_T} - V_D - \frac{R_C R_T}{R_C + R_T} I_L \cong V_{CC} - V_{CESAT} - V_D - R_T I_L \tag{8.27}$$

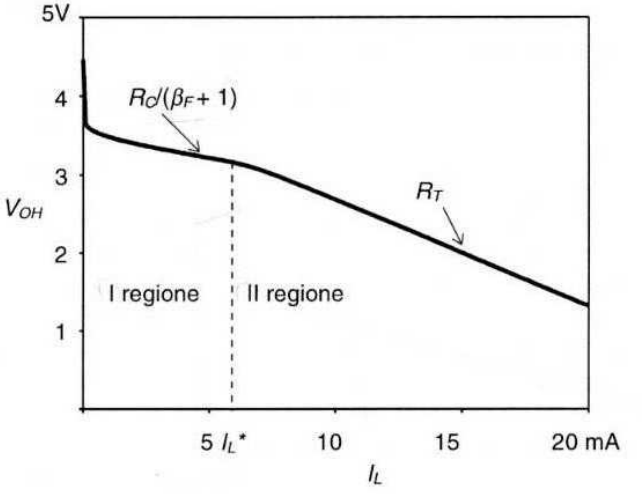


Figura 8.16 Caratteristica di uscita per $V_O = V_{OH}$

Il grafico di Figura 8.16 mostra la caratteristica di uscita per una porta TTL con i valori delle resistenze di Figura 8.8. Per corrente di carico trascurabile sia la tensione V_{BE} che V_D sono minori del valore di soglia di 0.6V e la tensione V_{OH} supera il valore nominale di 3.8V, ma già per correnti I_L delle decine di microampere la tensione scende al valore di 3.8V e dipende dalla corrente secondo la (8.25); oltre il valore I_L*, valutato mediante la (8.26), Q_b va in saturazione e la caratteristica assume una pendenza lineare con coefficiente R_T.

La seconda caratteristica di uscita si riferisce all'uscita V_{OL} dell'invertitore; in questo caso le porte connesse in uscita iniettano, come si è visto, una corrente I_I nell'uscita dell'invertitore, che contribuisce ad aumentare la corrente di collettore I_{Ca} che circola in Q_a. La tensione V_{OL}, che corrisponde alla V_{CESAT} di Q_a, aumenta con la corrente I_{Ca} secondo la (6.23), e la caratteristica di uscita ha l'andamento riportato in Figura 8.17 per il caso $\beta_F = 50$. Ne consegue che la massima corrente ammissibile in uscita con Q_a in saturazione è data da:

$$I_{CMAX} = \beta_F I_{Ba} \tag{8.28}$$

dove I_{Ba} è la corrente iniettata in base di Q_a per $V_I = V_{OH}$. Questa può essere ricavata in via approssimata, ricordando che Q_d è in saturazione, dal circuito di Figura 8.12 come:

$$I_{Ba} = I_{Cd} + I_{Bd} - I_{RE} = \frac{V_{CC} - V_{CESATd} - V_{BESATa}}{R_C} + \frac{V_{CC} - 2.3V}{R_B} - \frac{V_{BESATa}}{R_E} \quad (8.29)$$

e per i valori delle resistenze adottati vale circa 2.3 mA.

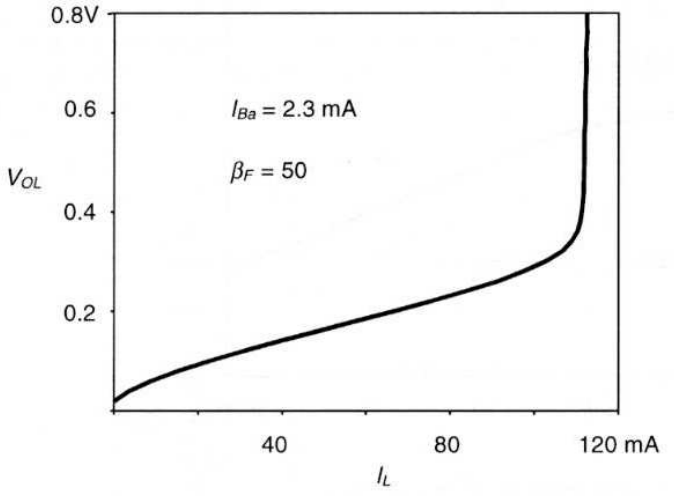


Figura 8.17 Caratteristica di uscita per $V_O = V_{OL}$

8.5.3 Fan-out

Dalle caratteristiche di ingresso e uscita si può facilmente determinare il fan-out dell'invertitore TTL, che è determinato essenzialmente da considerazioni sulle caratteristiche statiche, ed in particolare dalla corrente di ingresso delle porte collegate in uscita. Infatti nelle porte bipolari il fan-out è determinato dal massimo numero di porte che può essere collegato in uscita con una data degradazione dei margini di rumore o (il che è equivalente) con una data degradazione delle tensioni di uscita nei due stati logici.

Ad esempio, utilizzando i risultati dell'analisi svolta, nello stato basso di uscita, dalla (8.28) e (8.29) e ricordando il valore della corrente I_I iniettata dalle porte collegate in uscita si ottiene un fan-out N:

$$N = \frac{\beta_F I_{Ba}}{I_I} = \frac{50 \cdot 2.3}{1} \cong 115 \text{ porte}$$

con un valore di V_{OLMAX} di circa 0.3 V. Accettando per lo stato alto un valore V_{OHMIN} pari a 2.7 V (valore usuale per le TTL commerciali), si ricava dalla (8.27) una corrente I_{LMAX} nello stato alto pari a:

$$I_{Lmax} = \frac{V_{CC} - 0.9 - V_{OH\,min}}{R_T} \cong \frac{1.4}{130} \cong 10 \text{ mA}$$

e quindi un fan-out molto più elevato (10 mA/13 μA ≅ 770). Questo esempio mostra chiaramente che il fan-out è in pratica legato alla degradazione dell'uscita nello stato basso V_{OL}.

8.6 Dissipazione di potenza

La dissipazione di potenza di una porta TTL è essenzialmente quella statica dissipata nei due stati logici possibili. In assenza di carico significativo in uscita la potenza dissipata nello stadio a totem è trascurabile in entrambi gli stati di uscita, in quanto uno dei due transistori è in ogni caso in interdizione. Le altre due vie possibili di dissipazione sono quelle legate rispettivamente alla corrente circolante nella resistenza R_B e nella resistenza R_C. Nel caso di ingresso basso ($V_I = V_{OL}$) ricordiamo che il transistore Q_d è interdetto e quindi non circola corrente in R_C (vedi l'analisi circuitale di Figura 8.8); quindi l'unica dissipazione di potenza si ha attraverso la corrente che circola in R_B. Ricordando l'espressione della corrente I data dalla (8.12) la potenza dissipata in questo stato è data da:

$$P_{DL} = V_{CC} \frac{V_{CC} - V_{BEi} - V_{OL}}{R_B} \cong V_{CC} \frac{V_{CC} - 1V}{R_B} \tag{8.30}$$

Nello stato corrispondente ad un ingresso alto ($V_I = V_{OH}$) sia Q_i che Q_d conducono e quindi vi è assorbimento di corrente sia nella resistenza R_B che in R_C; le rispettive correnti valgono ora (vedi l'analisi del circuito in Figura 8.12):

$$I = \frac{V_{CC} - V_{BCi} - 2V_{BESAT}}{R_B}; \quad I_{Cd} = \frac{V_{CC} - V_{CESATd} - V_{BESATa}}{R_C} \tag{8.31}$$

e la dissipazione di potenza, con le consuete assunzioni, è data da:

$$P_{DH} = V_{CC}\left(\frac{V_{CC} - 2.3\text{ V}}{R_B} + \frac{V_{CC} - 1\text{ V}}{R_C}\right) \tag{8.32}$$

Dalle (8.30) e (8.32) si vede che la dissipazione di potenza con ingresso alto è maggiore di quella con ingresso basso. La potenza media dissipata è quindi:

$$< P_D >= \frac{1}{2}(P_{DL} + P_{DH}) \tag{8.33}$$

Ad esempio, con i valori delle resistenze assunti per la porta TTL di Figura 8.8, si ha:

$$P_{DL} = 5 \text{ mW}; \quad P_{DH} = 15.8 \text{ mW}; \quad < P_D >= 10.4 \text{ mW}$$

8.7 Tempo di propagazione e prodotto potenza-ritardo

Il tempo di propagazione, come si è visto nell'analisi delle prestazioni dinamiche dei vari sottoinsiemi dell'invertitore, è essenzialmente dovuto al tempo di accumulo legato al passaggio dalla saturazione all'interdizione del transistore Q_a dello stadio di uscita. Infatti il transistore Q_d dello stadio invertitore viene rapidamente interdetto dalla corrente estratta da Q_i, e l'interdizione di Q_b viene a sua volta accelerata dalla corrente assorbita da Q_d. Il tempo di accumulo di Q_a è ridotto dalla corrente di conduzione di Q_b nella fase di transizione, ma quest'ultima non può essere troppo elevata per ragioni di sicurezza dello stadio di uscita, nonché per considerazioni sulla dissipazione di potenza dinamica, e viene limitata dalla presenza della resistenza R_T ad un valore massimo I_T dato da:

$$I_T = \frac{V_{CC} - 2V_{CESAT} - V_D}{R_T} \cong \frac{V_{CC} - 1 \text{ V}}{R_T} \tag{8.34}$$

I risultati della simulazione SPICE del comportamento dinamico di una porta TTL sono riportati in Figura 8.18; si può verificare dal grafico della corrente di Q_d che il tempo di accumulo di quest'ultimo è molto più piccolo di quello del transistore Q_a, il quale contribuisce in maniera predominante al tempo di propagazione globale della porta.

Si può valutare in via analitica il tempo di propagazione della porta utilizzando i risultati dell'analisi dei tempi di propagazione del transistore bipolare, in funzione delle correnti di pilotaggio ed estrazione, presentati nel Paragrafo 7.5 per l'invertitore RTL, e cioè le espressioni di t_{PLH}, t_S, t_{LH} delle (7.23), (7.31), (7.36) riferite al transistore Q_a, sostituendo il valore I_T a quello I_{CSAT}:

$$t_{PHL} = \tau_{BF} \, ln\left(\frac{1}{1 - I_T / 2\beta_F I_{B1}}\right)$$

$$t_S = \tau_S \ln\left(\frac{I_{Ba1} - I_{Ba2}}{I_T / \beta_F - I_{Ba2}}\right) \tag{8.35}$$

$$t_{LH} = \tau_{BF} \ln\left(\frac{I_T - \beta_F I_{B2}}{I_T / 2 - \beta_F I_{B2}}\right)$$

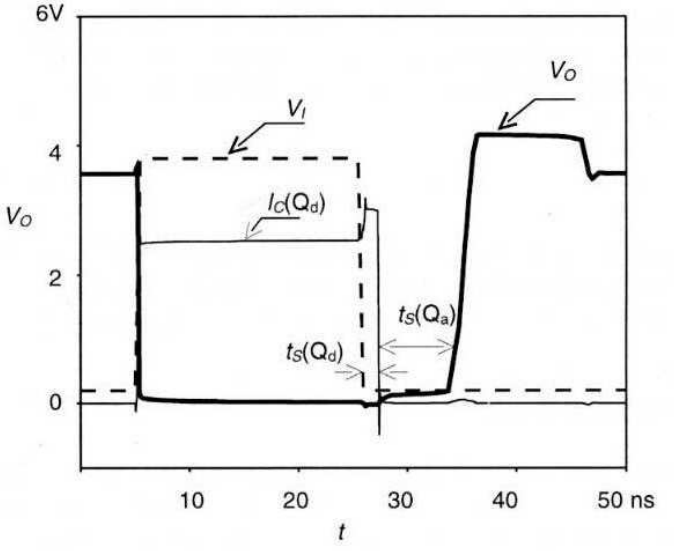


Figura 8.18 Simulazione SPICE del transitorio in uscita per una porta TTL con parametri del circuito di Figura 8.8

La corrente I_{B1} è la corrente iniettata in saturazione nella base di Q_a, espressa dalla (8.29), e $I_{B2} = -0.8\text{V}/R_E$. Utilizzando i parametri del transistore $\tau_{BF} = \beta_F \tau_F = 50 \cdot 0.06$ ns, $\tau_S \cong \tau_R = 10$ ns, si ottiene per i tempi di propagazione:

$$t_{PHL} = \tau_{BF} \cdot 0.13 = 0.4ns; \quad t_S = \tau_S \cdot 0.79 = 7.9ns; \quad t_{LH} = \tau_{BF} \cdot 0.24 = 0.7ns$$

che forniscono un ritardo di propagazione $t_P = 1/2\ (t_{PLH} + t_S + t_{LH}) = 4.5$ ns molto vicino al valore di $t_P = 4.6$ ns fornito dalla simulazione SPICE della porta senza carico. Il transitorio presentato dalla tensione V_O al termine del tempo t_{TLH}, con un innalzamento della tensione rispetto al valore V_{OH} di regime, è dovuto all'andata in saturazione del transistore Q_b con un momentaneo aumento della tensione di uscita, finchè non vengono smaltite le cariche accumulate nella base di Q_b.

Per quanto riguarda il prodotto potenza-ritardo della porta TTL, questo è relativamente elevato, principalmente per l'elevata potenza media dissipata; nell'esempio trattato esso vale:

$$P \cdot D = < P_D > t_P = 10.4 \text{ mW} \cdot 4.5 \text{ ns} = 46.8 \text{ pJ}$$

L'analisi svolta nel Paragrafo 8.6 sulla potenza dissipata non considera la componente di dissipazione di potenza legata alle transizioni della porta, che porta ad un contributo dinamico della dissipazione di potenza, proporzionale alla frequenza di funzionamento. Questa componente non è trascurabile nelle porte TTL standard,

ed è legata alla transizione t_{LH}, principalmente a causa della elevata corrente circolante nello stadio di uscita durante il tempo di accumulo t_S. Si può valutare in via approssimata questa componente di potenza P_{Dd}, assumendo costante la corrente I_T assorbita dal transistore Q_b nel tempo t_S, in base alla relazione:

$$P_{Dd} = \frac{t_S}{T}(I_T \cdot V_{CC}) \tag{8.35b}$$

dove t_S è il tempo di accumulo di Q_a, I_T la corrente che circola nello stadio di uscita nel transitorio di spegnimento di Q_a. Questo termine tuttavia è molto ridotto nelle porte TTL avanzate e in quelle Schottky che presenteremo nel Paragrafo 8.11.

8.8 Porte logiche TTL

Ricordando che le porte logiche TTL sono state sviluppate come evoluzione di quelle DTL, risulta chiaro che la funzione logica più facilmente realizzabile con porte TTL, analogamente al caso DTL, è la funzione NAND. In effetti la realizzazione di questa funzione richiede una modifica minima rispetto al circuito dell'invertitore elementare; non bisogna infatti iterare N volte tutto lo stadio invertitore come nel caso delle logiche CMOS, ma basta una piccola modifica nel transistore di ingresso Q_i.

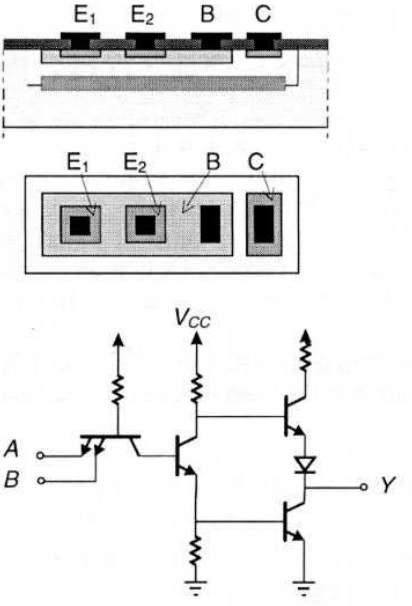


Figura 8.19 Struttura del transistore multiemettitore e porta NAND a due ingressi

Ricordiamo che per realizzare la funzione NAND a N ingressi nell'invertitore DTL occorre aggiungere $N-1$ diodi al nodo X (vedi Figura 7.16); poiché il diodo di ingresso della porta DTL è sostituito dalla giunzione emettitore-base del transistore Q_i nella TTL, la modifica consiste nell'*aggiungere* ulteriori emettitori al transistore Q_i. Questo è possibile con la struttura di transistore a multiemettitori di Figura 8.19, in cui più regioni di emettitore (tipicamente quattro) vengono realizzate nell'area di base del transistore, e contattate separatamente in modo da realizzare i diversi ingressi della porta. In questo transistore basta che una sola delle giunzioni base-emettitore sia polarizzata direttamente per portare Q_i a funzionare in modo diretto (e quindi portare all'interdizione Q_d), analogamente a quanto accadeva per i diodi della porta DTL; gli altri emettitori portati al livello alto mantengono interdette le giunzioni relative e non sono quindi efficaci per il funzionamento del transistore. Solo se tutti gli emettitori sono al potenziale alto, Q_i si porterà ad operare in modo inverso, e l'uscita sarà bassa; questo in effetti realizza la funzione NAND tra i diversi ingressi e l'uscita.

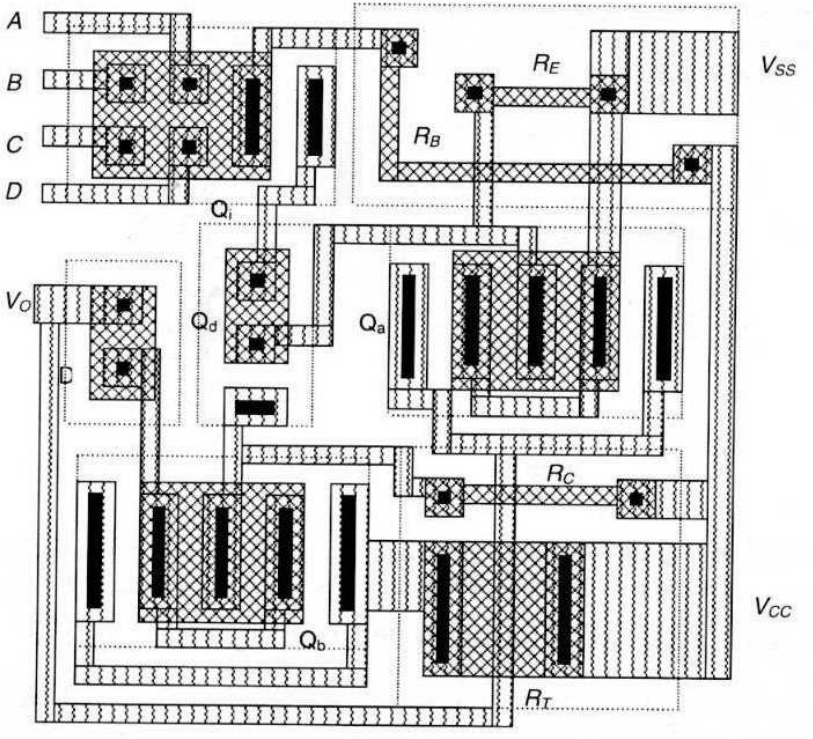


Figura 8.20 Tracciato di una porta NAND TTL a quattro ingressi

In Figura 8.20 è riportato il tracciato di una porta NAND TTL a quattro ingressi della serie 74. Nella figura si possono identificare le diverse regioni di isolamento in cui debbono essere realizzati i componenti circuitali, le resistenze integrate, di sezione e lunghezza differente a seconda del loro valore ($R_B > R_C$, $R_E >> R_T$), il transistore di ingresso Q_i a quattro emettitori, il diodo D ed i transistori di uscita Q_a e Q_b di area maggiore per aumentare la corrente di uscita (si noti che i transistori di uscita hanno un doppio contatto sia per l'emettitore che per il collettore, per ridurre le resistenze parassite relative). Sebbene l'area dell'invertitore TTL sia relativamente grande, se paragonata a quella di un invertitore CMOS, l'implementazione di porte NAND con più ingressi non richiede praticamente nessun ulteriore aumento di area, per cui in definitiva la porta NAND risulta conveniente dal punto di vista dell'occupazione di area.

La funzione NOR viene invece realizzata a spese di un maggior consumo di area, dovendosi in principio mettere in parallelo più invertitori con un unico carico R_C; in realtà si risparmia dell'area perché l'iterazione degli invertitori va realizzata solo a livello degli stadi di ingresso, in quanto il parallelo viene effettuato in uscita dallo stadio invertitore Q_d, e si utilizza un unico stadio di uscita per tutta la porta NOR, come esemplificato in Figura 8.21; tuttavia occorre iterare N volte gli stadi di ingresso se si vuole realizzare una porta NOR a N ingressi. Vedremo in seguito che con la logica TTL è però possibile realizzare facilmente porte complesse del tipo AND-OR-INVERT (A-O-I).

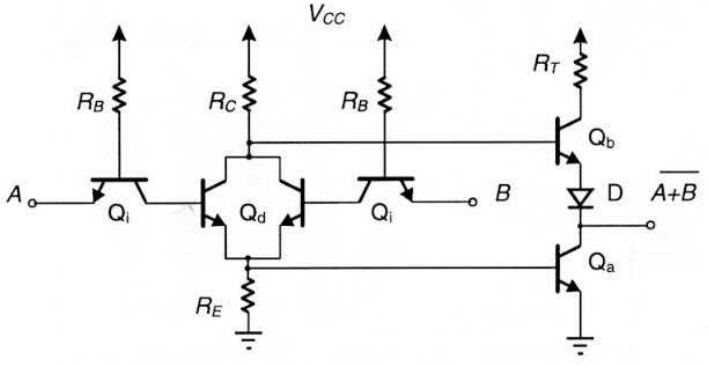


Figura 8.21 Porta NOR TTL a due ingressi

8.9 Reti attive di pilotaggio dell'uscita

Un significativo miglioramento delle caratteristiche sia statiche che dinamiche è stato ottenuto con l'impiego di transistori nelle reti di pilotaggio dei due transistori (detti di *pull-up* e *pull-down*) dello stadio di uscita. Queste reti sono state introdotte nelle versioni successive delle porte TTL (porte TTL-S, descritte nel Paragrafo 8.11), ma vengono presentate con riferimento alle porte TTL standard per poterne

più direttamente valutare gli effetti rispetto ai valori ricavati attraverso l'analisi precedente. Esaminiamo separatamente le due modifiche, che in effetti coesistono nelle porte TTL.

8.9.1 Rete di pull-up

La rete di pilotaggio del transistore Q_b (detta rete di pull-up), può essere migliorata con l'inserzione di un transistore Q_c tra il collettore di Q_d e la base di Q_b montato nella configurazione a collettore comune, come indicato in Figura 8.22. La presenza di questo transistore rende inutile il diodo D, in quanto nella maglia di base di Q_b la caduta V_D, necessaria per interdire Q_b quando l'uscita è bassa, è ora sostituita dalla caduta V_{BEc}, del tutto equivalente a tal fine. La presenza di Q_c aumenta il guadagno di corrente nel pilotaggio di Q_b e quindi riduce il tempo di commutazione di Q_b nella transizione dall'uscita bassa a quella alta, perché forza una corrente I_{Bb1} maggiore, secondo la relazione:

$$I_{Bb1} = (\beta_F + 1) I_{Bc} - \frac{V_{BEb}}{R} >> I_{Bc} \tag{8.36}$$

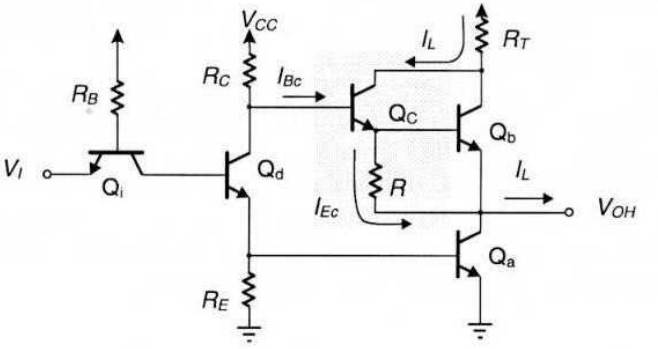


Figura 8.22 Rete di pull-up per l'invertitore TTL

dove la corrente I_{Bc} è equivalente alla corrente di base di Q_b nel circuito senza rete di pull-up. Un ulteriore vantaggio di questa rete è quello di aumentare il valore della tensione di uscita nello stato alto V_{OH}; infatti dal circuito di Figura 8.22 si vede che per $V_O = V_{OH}$ (Q_a interdetto) e per corrente di carico I_L trascurabile, la corrente di emettitore di Q_c (che contribuisce alla I_L) sarà anch'essa trascurabile e quindi la caduta su R non può portare in conduzione Q_b.

Quindi la tensione di uscita sarà data da:

$$V_{OH} \cong V_{CC} - R_C I_{Bc} - V_{BEc} - R I_{Ec} \cong V_{CC} - V_{BE\gamma c} \cong 4.4 \text{ V}$$

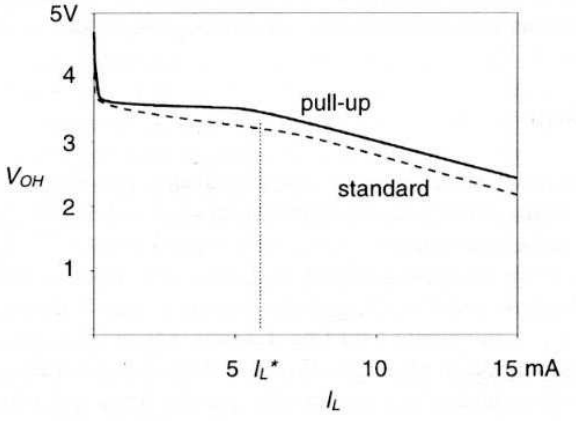


Figura 8.23 Simulazioni SPICE per la caratteristica di uscita per uscita alta: a) per un invertitore TTL standard; b) con rete di pull-up

poiché $I_{bc} << I_{Ec} = I_L \cong 0$. Tuttavia questo aumento si riduce per correnti di uscita non trascurabili, dovendosi sottrarre anche la caduta RI_{Ec}; ad esempio con una resistenza R di 3.5 kΩ, per una corrente di uscita $I_L = 0.2$ mA si ha già una caduta di 0.6 V, sufficiente a portare in conduzione il transistore Q_b, per cui già da tale valore di corrente di uscita la tensione V_{OH} scende a 3.8 V.

L'effetto principale della inserzione di Q_c è quello di ridurre la pendenza della caratteristica di uscita nella regione in cui Q_b lavora in regime attivo, come si può vedere dal grafico della Figura 8.23, dove si confronta la caratteristica di uscita di un invertitore TTL standard con quella di una porta TTL con rete di pull-up. Nel tratto della caratteristica compreso tra 0 e I_L* l'andamento della caratteristica di uscita è praticamente costante; infatti la resistenza di base vista da Q_b vale ora $R_C/(\beta_F + 1)$ per cui la (8.25) si trasforma in:

$$V_{OH} = V_{CC} - V_{BEc} - \frac{R_C}{(\beta_{Fc} + 1)(\beta_{Fb} + 1)} I_L - V_{BEb} \tag{8.37}$$

e la pendenza della curva è data da $R_C/(\beta_{Fc} + 1)(\beta_{Fb} + 1)$.

La connessione del transistore Q_c tra base e collettore di Q_b evita che quest'ultimo possa andare in saturazione, perché V_{BCb} può raggiungere un valore massimo pari a $- V_{CESATc} = - 0.2$ V < 0, tuttavia ciò non implica che la caratteristica di uscita sia definita dalla (8.37) per ogni valore di I_L, in quanto all'aumentare della corrente I_L di uscita aumenta anche in questo caso la caduta su R_T, e la tensione V_{CE} di Q_c si riduce finché quest'ultimo va in saturazione. Quando Q_c entra in saturazione la tensione V_{CE} di Q_b è fissata in ogni caso perché varrà: V_{CE}*$(Q_b) = V_{BEb} + V_{CESATc} = 0.9$ V. In definitiva il tratto a pendenza costante della caratteristica di uscita termina per un valore I_L* molto prossimo a quello definito dalla (8.26) per la TTL standard, come si può vedere dalla Figu-

ra 8.23. Al di là di questo valore, la caratteristica presenta anche in questo caso una pendenza circa uguale a R_T, in quanto la relazione approssimata che la definisce è analoga alla (8.27), se si sostituire il valore di V_{CE}*(Q_b) a quello di $V_{CESAT} + V_D$.

Il funzionamento in regime attivo del transistore Q_b migliora infine la dinamica del passaggio dell'uscita dal valore basso a quello alto anche con carichi capacitivi elevati, in quanto permette un transitorio con corrente di uscita più elevata.

8.9.2 Rete di pull-down

La rete di pilotaggio del transistore Q_a detta anche rete di pull-down, è indicata in Figura 8.24. Questa rete permette di migliorare sensibilmente la caratteristica di trasferimento ed in particolare il margine di rumore basso (che nella porta TTL standard è molto ridotto), migliorando nel contempo anche la dinamica di commutazione di Q_a.

Nella rete di pull-down la resistenza R_E viene in effetti sostituita dal bipolo non lineare di Figura 8.25a, che presenta la caratteristica corrente-tensione riportata in Figura 8.25b: finché la tensione V ai suoi capi non supera la tensione di soglia $V_{BE\gamma}$ della giunzione base-emettitore di Q_e questo non conduce, e quindi la corrente I assorbita dal bipolo è nulla. Per $V>V_{BE\gamma}$ la corrente assorbita dal bipolo è essenzialmente quella di collettore I_C perché quella di base è β volte più piccola, quindi si può trascurare la caduta su R' e la tensione V coinciderà praticamente con quella di ingresso V_{BE}. Oltre il valore V_{BESAT} la caratteristica I-V assume una pendenza approssimativamente data dal parallelo delle due resistenze R' e R'', come si può facilmente desumere da un'analisi delle variazioni della rete, assumendo le tensioni $V_{BESAT} = V_{CESAT}$ = costante.

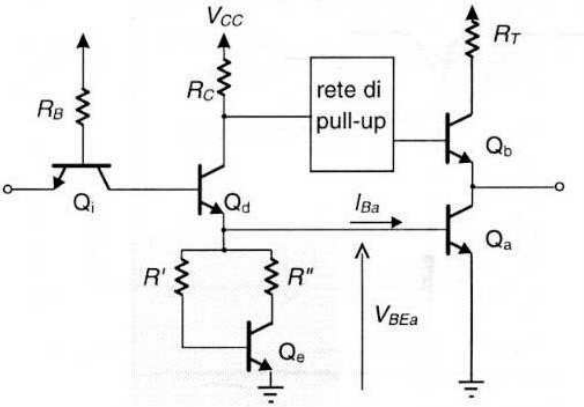


Figura 8.24 Rete di pull-down della porta TTL

Per comprendere l'effetto di questa rete sulla caratteristica di trasferimento, ricordiamo che quest'ultima nelle porte TTL standard presenta due regioni (vedi Figura 8.13): la prima (regione II) in cui conduce solo Q_d, e la seconda (regione III) in cui conducono sia Q_d che Q_a.

Poiché in questo caso la tensione ai capi del bipolo è anche la tensione V_{BE} di Q_a, ne consegue che la condizione per l'entrata in conduzione di Q_a ($V_{BE} > V_{BE\gamma}$) coincide con l'inizio della conduzione di Q_d ($I_{Ed} = I > 0$); quindi Q_d rimane interdetto finché la tensione di ingresso V_I raggiunge il valore $V_I^* = -V_{CESATi} + 2\,V_{BE\gamma} \cong 1.1$ V.

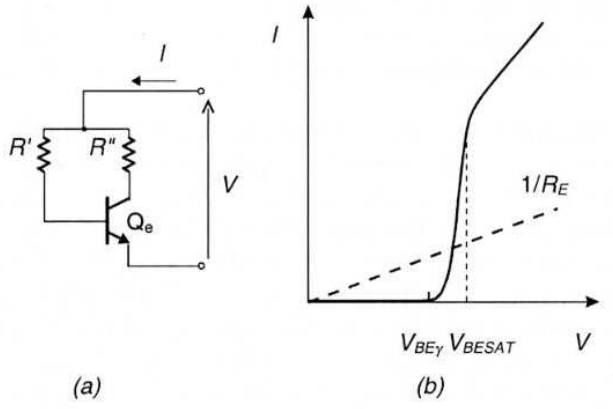


Figura 8.25 Analisi della rete di pull-down: a) rete elettrica; b) curva *I-V* del bipolo

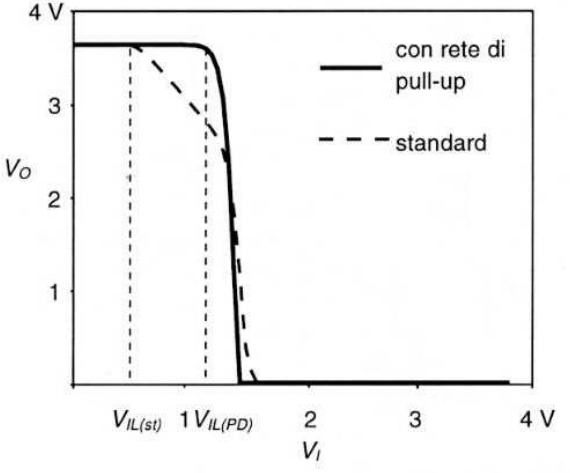


Figura 8.26 Caratteristica di trasferimento di una porta TTL con rete di pull-down, confrontata con quella di una TTL standard

Quest'ultimo valore per definizione corrisponderà al valore V_{IL}, cioè al massimo valore dell'ingresso con uscita nello stato alto, e la caratteristica di trasferimento assume l'andamento squadrato riportato in Figura 8.26. Quindi i margini di rumore della porta TTL così modificata varranno:

$$NM_L = V_{IL} - V_{OL} \cong 1.1 - 0.2 = 0.9 \text{ V} \tag{8.38}$$

$$NM_H = V_{OH} - V_{IH} \cong 3.8 - 1.5 = 2.3 \text{ V} \tag{8.39}$$

Il comportamento fortemente nonlineare della caratteristica I-V del bipolo equivalente della rete di pull-down è utile anche per migliorare la dinamica della commutazione del transistore Q_a, che è il maggiore responsabile del tempo di propagazione della porta.

Infatti, nella fase di conduzione di Q_d (ingresso alto) la corrente di pilotaggio della base di Q_a (che nella TTL standard era data dalla differenza tra la corrente di emettitore I_{Ed} e quella assorbita dalla resistenza R_E) ora viene ridotta perché una maggiore aliquota viene assorbita dalla rete di pull-down (vedi Figura 8.24) che al di sopra di $V_{BE\gamma}$ presenta una resistenza differenziale bassa. Questo riduce il forzamento in saturazione e quindi anche il tempo di accumulo t_S che è diretta conseguenza del primo. Infine, nella fase di passaggio di Q_a dalla saturazione all'interdizione, il tempo di accumulo t_S viene ulteriormente ridotto dall'aumento di corrente I_{B2} estratta dalla base, sempre a causa della minor resistenza offerta dal bipolo per $V = V_{BESAT}$.

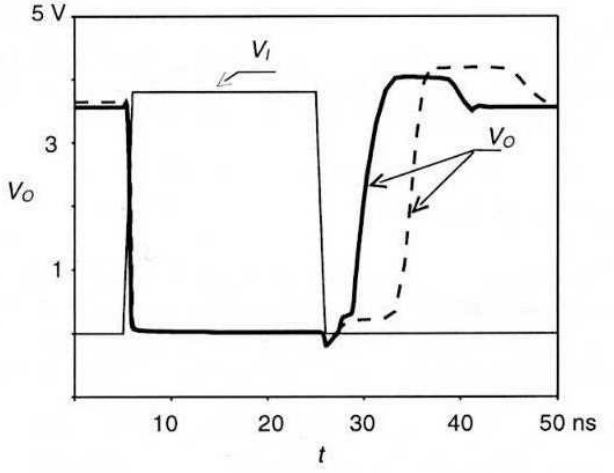


Figura 8.27 Forme d'onda di commutazione ottenute con simulazioni SPICE per una porta TTL con reti di pull-up e pull-down (linea continua), confrontate con quelle di una TTL standard (linea tratteggiata)

In definitiva l'impiego di queste due reti (in particolare quella di pull-down) migliora significativamente le prestazioni della porta. In Figura 8.27 si possono confrontare le forme d'onda dei transitori in uscita di una porta TTL con reti di pull-up e pull-down, con quelle della porta standard di Figura 8.18; si vede che il tempo di accumulo (di Q_a) si riduce sensibilmente.

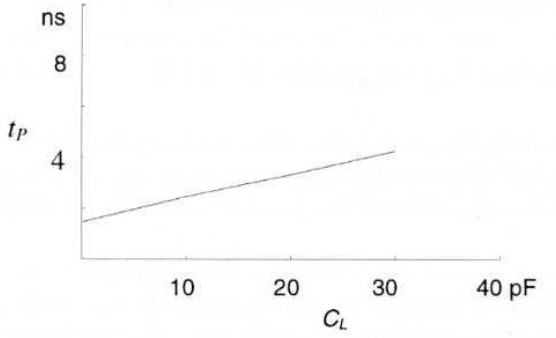


Figura 8.28 Dipendenza del tempo di propagazione dalla capacità di carico C_L

La dipendenza del tempo di propagazione dal valore della capacità di carico C_L per le porte TTL è sensibilmente ridotta sia rispetto alle porte DTL che alle porte CMOS, a causa della capacità dello stadio di uscita di fornire correnti elevate di carico nel transitorio di commutazione, in quanto i transistori di uscita operano in regione attiva durante le commutazioni. In Figura 8.27 sono riportati i risultati di simulazioni SPICE per la porta TTL con reti di pull-up e pull-down al variare della capacità di carico C_L; si vede che la porta può alimentare anche capacità relativamente grandi con un aumento contenuto del tempo di propagazione (specie se lo si confronta con la dipendenza del tempo di propagazione con C_L delle porte DTL o CMOS), il che comporta un significativo vantaggio per la logica TTL.

8.10 Il transistore Schottky

Un ulteriore miglioramento delle prestazioni delle porte logiche TTL è venuto dall'impiego di *transistori Schottky* nella realizzazione del circuito integrato. Il transistore Schottky è un transistore bipolare in cui viene integrata in fase di realizzazione una giunzione metallo-semiconduttore (detta *diodo Schottky*) connessa tra base e collettore.

La giunzione metallo-semiconduttore ha una caratteristica I-V simile a quella di un diodo P/N, ma con una ridotta tensione di soglia V_γ, a causa della ridotta barriera di potenziale nella giunzione metallo-semiconduttore rispetto a quella tra semiconduttore drogato P e drogato N. In questo diodo (vedi Figura 8.29a) il metallo (usualmente alluminio, titanio o platino) si comporta da anodo del diodo, mentre il

semiconduttore drogato N corrisponde al catodo, e la caratteristica I-V (Figura 8.29b) segue la nota legge esponenziale del diodo P/N:

$$I = I_o\left(\exp\frac{V}{V_T} - 1\right) \tag{8.40}$$

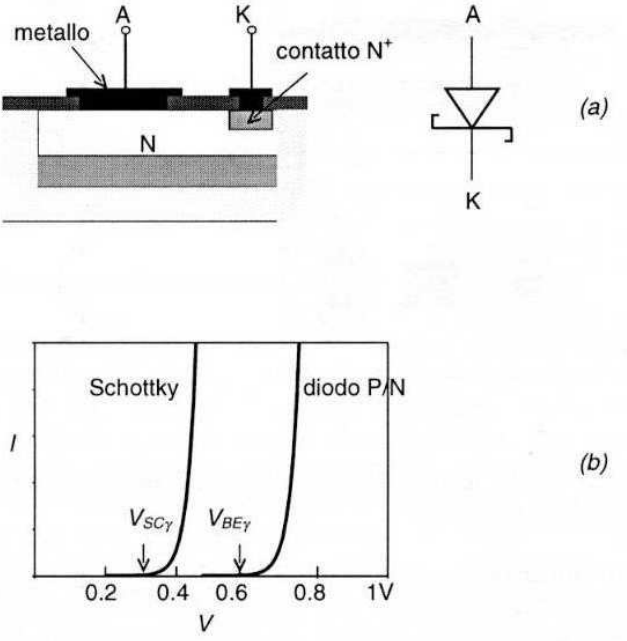


Figura 8.29 a) Struttura del diodo Schottky: b) simbolo elettrico e caratteristica *I-V*

ma con una corrente I_O molto più elevata, il che comporta una tensione di soglia $V_\gamma \cong 0.3$ V, ed una tensione di conduzione $V_{SC} \cong 0.5$ V. Un'ulteriore importante differenza è dovuta al fatto che la corrente del diodo Schottky è dovuta ai portatori maggioritari del semiconduttore, per cui non vi sono fenomeni di accumulo delle cariche minoritarie come nel diodo P/N, e quindi il comportamento dinamico è molto più rapido.

Il diodo Schottky viene impiegato in connessione con il transistore bipolare, per evitare l'entrata in saturazione del transistore a seguito di un segnale di ingresso elevato; infatti collegando il diodo tra base e collettore come in Figura 8.30b si garantisce che la tensione V_{BC} del transistore non superi mai il valore di $V_{SC} \cong 0.5$ V della piena conduzione del diodo Schottky, evitando quindi l'entrata in saturazione del transistore (che richiede $V_{BC} \geq 0.6$ V). L'eccesso di corrente eventualmente fornita dall'ingresso rispetto a quella $I_B = I_C/\beta_F$ viene quindi deviata dal diodo direttamente in uscita. L'integrazione del diodo

Schottky nel transistore è semplice (vedi Figura 8.30a), in quanto il metallo viene depositato direttamente a contatto del collettore N (dove si crea la giunzione metallo-semiconduttore) e viene poi esteso al contatto di base in modo da realizzare l'interconnessione tra base e anodo del diodo; l'aumento di area è ridotto quindi al minimo.

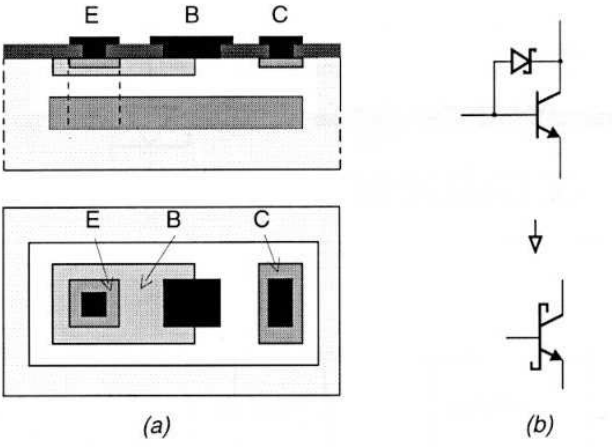


Figura 8.30 a) Struttura del dispositivo integrato; b) circuito equivalente e simbolo elettrico del transistore Schottky

Il vantaggio principale di questa modifica è quello di una drastica riduzione del tempo di accumulo del transistore, nel successivo passaggio dalla saturazione all'interdizione di quest'ultimo. Questo vantaggio viene pagato da un aumento della tensione del transistore nello stato basso, in quanto non si raggiunge la piena saturazione, e quindi la $V_{CEMIN} = V_{BESAT} - V_{BC} \cong 0.3 \div 0.4\ \text{V} > V_{CESAT}$.

L'impiego di transistori Schottky nelle porte TTL ha prodotto una serie di famiglie logiche indicate come TTL-Schottky, che verranno discusse nei paragrafi seguenti.

8.11 Logiche TTL-Schottky

8.11.1 Logiche TTL-Schottky veloci

La prima versione di porta logica TTL-Schottky (denominata TTL-S) è quella esemplificata in Figura 8.31 per il caso di una porta NAND a due ingressi. Si può notare che tutti i transistori della porta TTL sono stati sostituiti da transistori Schottky, eccetto Q_b che in ogni caso non va in saturazione per la presenza del transistore Q_c di pull-up, come già visto nel Paragrafo 8.9.

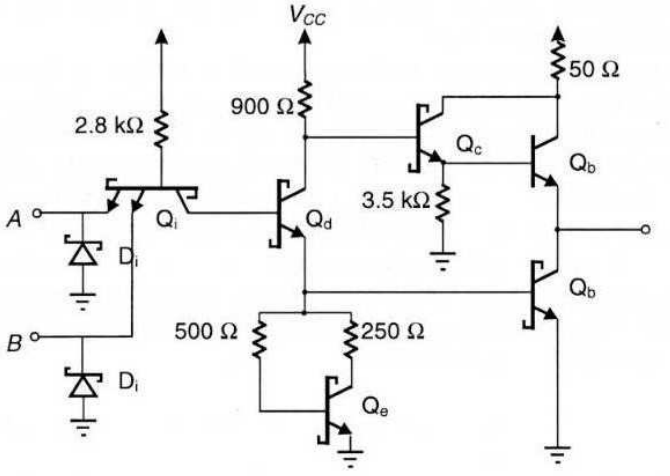


Figura 8.31 Schema elettrico di una porta TTL-Schottky veloce

Il circuito utilizza per le resistenze R_B, R_C, R_T valori più bassi di quelli impiegati nelle porte TTL standard, per aumentare la velocità di commutazione della porta, già notevolmente migliorata. Infatti nei transistori Schottky, in assenza dei tempi di accumulo dovuti alla saturazione, i tempi di transizione sono determinati essenzialmente dalle capacità di svuotamento delle giunzioni e da quelle di diffusione della base, e quindi riducendo i valori delle resistenze collegate alle giunzioni di base e di collettore si riducono le costanti di tempo associate e si velocizzano i transitori; inoltre la riduzione delle resistenze di base aumenta le correnti di pilotaggio dei transistori e questo contribuisce ad accelerare la commutazione dall'interdizione alla saturazione e viceversa. La resistenza R in questo caso non viene connessa all'uscita ma verso massa, in modo da ridurre la resistenza vista dalla base di Q_b e favorire l'estrazione della carica nello spegnimento di Q_b.

Nel caso della porta TTL di Figura 8.31, utilizzando per la simulazione SPICE gli stessi valori dei parametri dei transistori impiegati per le analisi di Figura 8.18 e 8.27, si ottengono tempi di propagazione inferiori al nanosecondo. Con commutazioni così rapide, le interconnessioni non possono essere più considerate come dei semplici collegamenti, ma piuttosto delle linee di trasmissione a costanti distribuite; è noto che se il carico (in questo caso l'ingresso della porta successiva) non ha lo stesso valore dell'impedenza caratteristica della linea, si creano delle riflessioni all'uscita che creano delle oscillazioni smorzate (*ringing*) lungo la linea, come sarà meglio visto nel Paragrafo 9.8. Per evitare che queste oscillazioni possano creare delle commutazioni indesiderate nelle porte vengono inseriti i diodi (Schottky) D_I collegati con i catodi ai singoli ingressi. I diodi sono interdetti nel normale funzionamento (tensioni positive in ingresso), mentre conducono nella semionda negativa di un'eventuale oscillazione, mantenendo la massima tensione negativa alla tensione $V_{SC} = 0.4$

V; quindi le successive semionde positive saranno inferiori a questo valore e rimarranno sotto il valore V_{IL} della porta.

Il prezzo da pagare per questo miglioramento delle prestazioni dinamiche è un aumento della potenza dissipata dalla porta. Ricordiamo che nel Paragrafo 8.6 si è valutata la dissipazione di potenza in base alle (8.30) e (8.32), con uscita rispettivamente bassa ed alta; queste relazioni mostrano una dipendenza della corrente assorbita dall'inverso delle resistenze R_B e R_C, per cui la dissipazione di potenza aumenta sia nello stato di uscita bassa che di uscita alta. Oltre a ciò, nel caso della porta TTL-S, vi è assorbimento di potenza anche da parte del transistore Q_c nella condizione di uscita bassa V_{OL}, in quanto vi è circolazione di corrente nella resistenza R quando l'ingresso V_I è alto, e il transistore Q_d è interdetto, come è evidenziato nello schema elettrico di Figura 8.32.

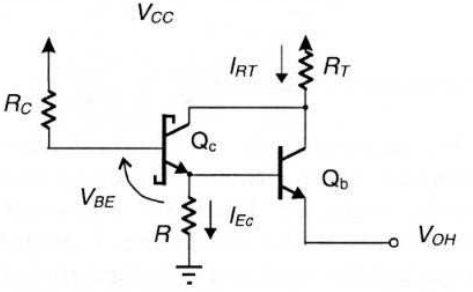


Figura 8.32 Schema elettrico per l'assorbimento di corrente del transistore Q_c per $V_O = V_{OH}$

In questo caso, anche se non circola corrente nel transistore Q_b, perché Q_a è interdetto, può circolare corrente attraverso R_T, Q_c e R. Trascurando anche in questo caso la caduta su R_C dovuta alla corrente di base di Q_c, la corrente I_{Ec} in R sarà approssimativamente uguale alla corrente I_{RT} circolante in R_T. La corrente I_{Ec} può essere definita come:

$$I_{Ec} = \frac{V_{Ec}}{R} = \frac{V_{CC} - V_{BEc}}{R} \tag{8.41}$$

e le dissipazioni di potenza P_{DL} e P_{DH} della porta possono essere scritte in questo caso, utilizzando ancora la (8.30) per l'uscita V_{OL} e aggiungendo il termine I_{Ec} alle (8.31, 8.32), come:

$$P_{DL} = V_{CC} \frac{V_{CC} - V_{BESATi} - V_{OL}}{R_B} \tag{8.42}$$

$$P_{DH} = V_{CC}\left(\frac{V_{CC} - V_{BCi} - 2V_{BESAT}}{R_B} + \frac{V_{CC} - V_{CE(sc)d} - V_{BESATa}}{R_C} + \frac{V_{CC} - V_{BEc}}{R}\right) \quad (8.43)$$

dove si è sostituito nelle (8.31, 8.32) il valore $V_{CE(SC)d} = 0.3$V al posto di $V_{CESATd} = 0.2$ V. Per il caso del circuito di Figura 8.31 si ha:

$$P_{DL} = 7.14 \text{ mW}; \quad P_{DH} = 32.5 \text{ mW}; \quad < P_D >= 19.8 \text{ mW}$$

La caratteristica di trasferimento di queste porte ha la forma squadrata già vista per la porta TTL con rete di pull-down, ma con un valore più alto di $V_{OL} \cong 0.3$ V, a causa del transistore Schottky Q_a in uscita.

8.11.2 Logiche TTL-Schottky a basso consumo

Le logiche Schottky a basso consumo (indicate con la sigla TTL-LS) sono state introdotte con lo scopo di convertire la potenziale alta velocità di queste logiche in un consumo di potenza più ridotto, essenzialmente aumentando i valori delle resistenze R_B, R_C, R', R'' rispetto alla versione standard (e a quella Schottky veloce sopra descritta). In queste porte sono state introdotte ulteriori modifiche essenzialmente miranti a recuperare velocità nelle transizioni e quindi contenere l'aumento dei tempi di propagazione, modifiche indicate nello schema circuitale di una tipica porta NAND TTL-LS a due ingressi, riportato in Figura 8.33.

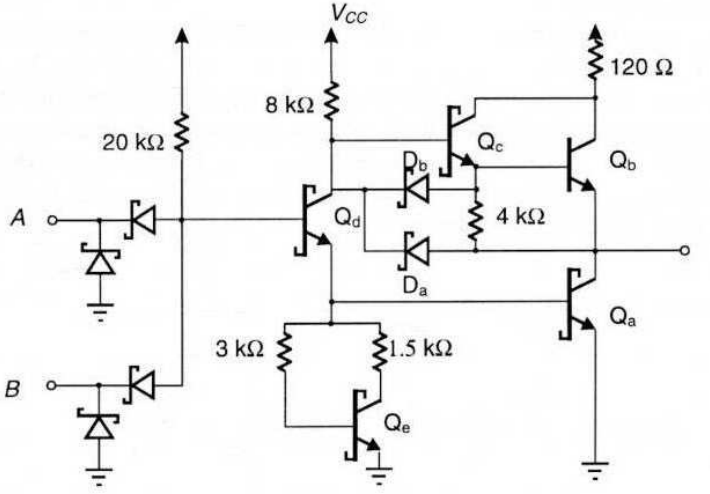


Figura 8.33 Schema di una porta NAND a due ingressi in versione Schottky a basso consumo (TTL-LS)

Oltre all'aumento delle resistenze di cui si è detto, in questo schema si può notare una modifica nello stadio di ingresso: al posto del transistore multiemettitore Q_i si è introdotta una rete a diodi (Schottky) che di fatto realizza la funzione AND come nell'ingresso delle porte DTL (si noti che con l'uso dei diodi Schottky che hanno una tensione di conduzione $V_{SC} < 0.5$ V e con la presenza della rete di pull-down non è più necessaria la presenza di diodi in serie sul terminale di base di Q_d come nella DTL). La ragione della sostituzione di Q_i con i diodi è giustificata dall'assenza di saturazione per il transistore Q_d, che quindi non richiede più l'estrazione rapida della carica accumulata attraverso Q_i. D'altra parte l'utilizzazione dei diodi al posto di Q_i riduce l'area delle giunzioni in gioco nella maglia di ingresso e quindi le capacità parassite di queste ultime, che nei transistori Schottky giocano un ruolo non trascurabile nel rallentare le commutazioni.

Un'ulteriore modifica utile per il miglioramento delle prestazioni dinamiche è l'introduzione dei diodi (Schottky) D_a e D_b nella rete di pull-up; questi diodi servono per velocizzare la transizione dell'uscita dal valore alto a quello basso. Per comprendere il loro ruolo si consideri che il transistore Q_d (ed in particolare il nodo di collettore, che è un nodo interno al circuito integrato) commuta più rapidamente dell'uscita, che nei casi reali è connessa a carichi capacitivi C_L non trascurabili. I diodi nel funzionamento stazionario della porta sono interdetti, essendo in parallelo alla giunzione base-emettitore di Q_c che presenta tensioni V_{BE} positive, mentre entrano in conduzione quando Q_d passa in conduzione mentre l'uscita (e Q_b) non hanno ancora cambiato stato. La conduzione di D_b permette un aumento della corrente I_{B2} estratta dalla base di Q_b che ne favorisce l'interdizione. La conduzione di D_a permette di estrarre una maggiore corrente dalla capacità C_L di uscita nei primi istanti in cui Q_a non è ancora in piena conduzione, e quindi favorisce anch'esso una più rapida commutazione della tensione di uscita.

La dissipazione di potenza nei due stati logici differisce da quella della porta TTL-S essenzialmente perché in questa versione non vi è il contributo dovuto alla corrente circolante nel transistore Q_c nella uscita alta V_{OH}. In questo caso infatti la resistenza R è collegata al collettore di Q_a anziché a massa, per cui nel ramo di Q_c non può circolare corrente quando l'uscita è alta, perché in questo caso Q_a è interdetto. La dissipazione di potenza può essere espressa, con riferimento al circuito di Figura 8.33, dalle relazioni seguenti:

$$P_{DL} = V_{CC} \frac{V_{CC} - V_{SC} - V_{OL}}{R_B} \tag{8.44}$$

$$P_{DH} = V_{CC} \left(\frac{V_{CC} - 2V_{BESAT}}{R_B} + \frac{V_{CC} - V_{CE(sc)d} - V_{BESATa}}{R_C} \right) \tag{8.45}$$

Con i valori delle resistenze del circuito di Figura 8.33 si ha:

$$P_{DL} = 1.05 \text{ mW}; \quad P_{DH} = 3.28 \text{ mW}; \quad < P_D >= 2.17 \text{ mW}$$

e quindi si ha una riduzione di potenza di un fattore 9 rispetto a quella delle porte TTL-S.

La riduzione della dissipazione di potenza conseguente all'aumento delle resistenze del circuito provoca un aumento del tempo di propagazione di queste porte; ad esempio una simulazione SPICE del circuito di Figura 8.33 con i parametri già utilizzati per le altre analisi SPICE, fornisce per il tempo di propagazione $t_P = 1.4$ ns.

8.12 Logiche TTL-Schottky avanzate

Le più recenti versioni delle porte logiche TTL-S includono i miglioramenti tecnologici dei transistori presentati nel Paragrafo 6.7, che permettono una maggiore velocità di commutazione ed una minore area dei dispositivi (e del circuito complessivo); queste vengono indicate con le sigle TTL-AS per le versioni ad alta velocità e TTL-ALS per quelle a bassa dissipazione; esse si diversificano anche per ulteriori modifiche degli stadi di ingresso, come l'inserzione di uno stadio preamplificatore che aumenta il valore V_{IL} e quindi il margine di rumore per ingresso basso.

In Tabella 8.1 sono riportati i valori tipici presentati dalle diverse famiglie logiche TTL per i principali parametri di caratterizzazione delle porte, desunti dai dati delle specifiche tecniche. In particolare le grandezze V_{ILMAX} e V_{IHMIN} sono i valori rispettivamente massimi e minimi garantiti dal costruttore; i valori di uscita V_{OHMIN} e V_{OLMAX} sono le tensioni di uscita nei due stati con correnti di uscita massime date rispettivamente da I_{OLMAX} e I_{OHMAX}. Dai dati riportati in tabella si vede come la famiglia TTL-ALS presenti i valori più bassi del prodotto $P \cdot D$, e per questa caratteristica è quella più utilizzata nelle applicazioni che richiedono componenti logici standard per la realizzazione di circuiti logici.

Tabella 8.1 Caratteristiche elettriche delle diverse famiglie TTL

grandezza		*TTL*	*TTL-S*	*TTL-LS*	*TTL-AS*	*TTL-ALS*
ritardo t_P	nS	10	3	9	1.7	4
consumo P_D	mW	10	20	2	8	1.2
prodotto $P \cdot D$	pJ	100	60	18	13.6	4.8
V_{ILMAX}	V	0.8	0.8	0.8	0.8	0.8
V_{OLMAX}	V	0.5	0.5	0.5	0.5	0.5
V_{IHMIN}	V	2.0	2.0	2.0	2.0	2.0
V_{OHMIN}	V	2.7	2.7	2.7	2.7	2.7
I_{OLMAX}	mA	20	20	8	20	8
I_{OHMAX}	mA	−1.0	−1.0	−0.4	−1.0	−0.4

Esercizi di riepilogo

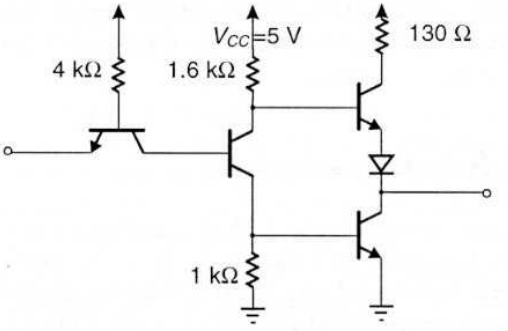


Figura E8.1

8.1 Per l'invertitore TTL di Figura E8.1 determinare i valori della corrente di ingresso nei due stati logici, assumendo per i transistori i seguenti parametri: $\beta_F = 50$, $\beta_{Ri} = 0.06$. Determinare inoltre il fan-out della porta, assumendo come valori accettabili $V_{OHMIN} = 2.7$ V, e $V_{OLMAX} = 0.3$ V. Quale delle due condizioni determina il fan-out della porta?

8.2 Per l'Esercizio 8.1 valutare l'effetto, sul valore del fan-out determinato, di una variazione del β_F di ±20% sul valore nominale.

8.3 Sempre con riferimento ai risultati dell'Esercizio 8.1, valutare l'effetto che ha sul valore del fan-out, un valore di $\beta_{Ri} = 0.1$.

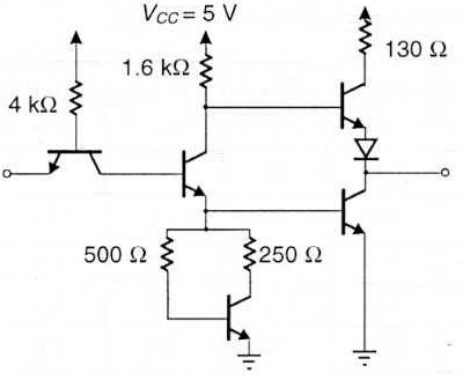


Figura E8.2

8.4 Con riferimento all'invertitore TTL di Figura E8.1, assumendo per i parametri dei transistori: $\beta_F = 50$, $\tau_F = 0.06$ ns, $\tau_S = 10$ ns, determinare i punti notevoli della caratteristica di trasferimento mediante analisi approssimata

per tre valori di R_E: 500 Ω, 1 kΩ, 2 kΩ. Calcolare inoltre, utilizzando le formule analitiche approssimate, il tempo di propagazione t_P per i tre casi e giustificare con questi valori la scelta del valore di 1 kΩ delle porte commerciali.

8.5 Per l'invertitore di Figura E8.2 determinare il valore della corrente di base I_{Ba1} per un ingresso al valore logico alto V_{OH}, e quella I_{Ba2} di estrazione quando l'ingresso passa da V_{OH} a V_{OL}. Confrontare i valori ottenuti con quelli ottenibili dall'analisi dell'invertitore TTL standard di Figura E8.1.

8.6 Si analizzi la potenza dissipata totale (statica + dinamica) dell'invertitore TTL di Figura E8.1 con i parametri: $\beta_F = 50$, $\beta_{Ri} = 0.02$, $\tau_F = 0.06$ ns, $\tau_S = 10$ ns, collegato ad un uguale invertitore, al variare della frequenza $f = 1/T$ di ripetizione del segnale di ingresso, supposto con 50% di *duty cycle*, e si determini il valore di frequenza del segnale per cui la dissipazione di potenza totale è il doppio di quella statica.

8.7 Per l'analisi dell'Esercizio 8.4, confrontare i risultati ottenuti con un'analisi approssimata con quelli di simulazioni SPICE per i tre casi indicati, utilizzando per i transistori le schede .MODEL riportate in Appendice.

8.8 Per l'invertitore TTL standard di Figura E8.1 caricato in uscita da una capacità C_L, ricavare, mediante simulazioni SPICE, la dipendenza del ritardo di propagazione t_P dalla capacità C_L al variare di quest'ultima da 0.1pF a 20 pF, utilizzando per i transistori le schede .MODEL riportate in Appendice.

8.9 Ricavare, mediante simulazioni SPICE, il grafico della caratteristica di uscita per uscita alta V_{OH}, di un invertitore TTL con rete di pull-up con i valori delle resistenze: $R_B = 4$ kΩ, $R_C = 1.6$ kΩ, $R_T = 130$ Ω, $R = 3.5$ kΩ, ed utilizzando per i transistori le schede .MODEL riportate in Appendice.

8.10 Ricavare, mediante simulazioni SPICE, il grafico della caratteristica di uscita, per uscita bassa V_{OL}, di un invertitore TTL con rete di pull-down con i valori delle resistenze: $R_B = 4$ kΩ, $R_C = 1.6$ kΩ, $R_T = 130$ Ω, $R = 3.5$ kΩ, ed utilizzando per i transistori le schede .MODEL riportate in Appendice. Si confronti il grafico con quello ottenuto con una simulazione SPICE in cui sono poste uguali a zero le resistenze parassite di base, collettore ed emettitore dei transistori dello stadio di uscita.

8.11 Valutare analiticamente il comportamento statico della rete di pull-down dell'invertitore TTL dell'Esercizio 8.10 al fine di valutare la corrente I_{Ba1} di pilotaggio del transistore Q_a per uscita al valore V_{OL}. Nell'ipotesi che il fan-out di questo invertitore sia determinato dalla degradazione del livello logico basso, e assumendo un valore massimo di $V_{OL} = 0.5$ V, determinare inoltre il valore del fan-out in questo caso; si assuma $\beta_F = 50$ per i transistori Q_e e Q_a.

8.12 Valutare analiticamente la caratteristica di ingresso per l'invertitore TTL-LS di Figura 8.31, e confrontare i valori di I_L e I_H ottenuti con quelli dell'invertitore TTL standard dell'Esercizio 8.1.

8.13 Determinare analiticamente i punti notevoli della caratteristica di trasferimento della porta NAND TTL-LS a due ingressi di Figura 8.31, per il caso in cui A = 1, B passa da 0 a 1.

8.14 Analizzare mediante simulazioni SPICE la dissipazione di potenza statica per gli invertitori TTL-S di Figura 8.30, e TTL-LS di Figura 8.31, ricavando le correnti circolanti nelle resistenze R_B, R_C, R_T al variare dell'ingresso da 0 a 5 V, e confrontare i risultati con quelli dell'analisi approssimata svolta nel testo.

Riferimenti bibliografici

H. Taub, D. Schilling, *Elettronica integrata digitale*, Jackson, Milano, 1981.

J. Millman, *Circuiti e sistemi microelettronici*, Bollati Boringhieri, Torino, 1985.

G.M. Glansford, *Digital Electronic Circuits*, Prentice Hall, Englewood Cliffs, 1988.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

B. Riccò, F. Fantini, P. Brambilla, *Introduzione ai circuiti integrati digitali*, Zanichelli, Bologna, 1991.

A.S. Sedra, K.C. Smith, *Microelectronic Circuits*, Saunders College, Philadelphia, 1991.

9

Porte logiche ECL

9.1 La configurazione differenziale

Le porte logiche bipolari che sono state presentate nei capitoli precedenti utilizzano transistori che, in dipendenza del segnale logico di ingresso, operano in una delle due possibili regioni di funzionamento: interdizione o saturazione. Ciò vale sia per le porte RTL, DTL e TTL standard, che per quelle TTL-Schottky; per queste ultime infatti i transistori vengono comunque portati a lavorare nello stato ON nella regione di saturazione, e cioè con una tensione $V_{BC} > 0$ della giunzione base-collettore, anche se la presenza del diodo Schottky evita l'accumulo di cariche nella base dovuto al forzamento in profonda saturazione.

È tuttavia possibile utilizzare per le porte logiche elementari particolari circuiti che permettono ai transistori di lavorare tra interdizione e regione attiva, evitando comunque l'ingresso nella regione di saturazione, anche in assenza di diodi Schottky sulla giunzione base-collettore. Le logiche basate su questi circuiti vengono chiamate *logiche non saturate*, e sono le logiche più veloci, sia rispetto alle altre famiglie bipolari che alle logiche NMOS e CMOS.

Il circuito elementare con cui viene realizzata questa famiglia logica non si basa più sull'interruttore elementare di Figura 7.1 (che in termini di circuito analogico viene identificato come amplificatore invertente in configurazione ad emettitore comune), ma utilizza uno schema differenziale (che nell'elettronica analogica è la base della struttura più versatile di amplificatore, e cioè l'amplificatore operazionale, o OP AMP).

Nello schema differenziale di principio, riportato in Figura 9.1, la corrente I che fluisce nel ramo comune a cui sono connessi i due emettitori viene divisa in due componenti uguali e pari a $I/2$, uscenti da ciascuno degli emettitori dei due transistori, quando i segnali applicati alle due basi sono di uguale valore; se invece questi ultimi sono diversi, il rapporto tra le correnti che fluiscono in ognuno dei due transistori (e cioè in ognuno dei rami della configurazione differenziale) dipende dalla *differenza* delle tensioni applicate alle due basi (ingressi) del circuito, che per tale

ragione viene definito come *configurazione differenziale*. Il circuito viene studiato approfonditamente nell'ambito dell'elettronica analogica, per cui si richiameranno qui solo le caratteristiche utili per comprenderne l'impiego nelle porte logiche non saturate.

Supponendo che i due transistori lavorino in regione attiva (ricordiamo che l'impiego di questa configurazione nelle porte logiche ha proprio questo scopo), le due correnti di emettitore possono essere approssimate, in base alle relazioni (6.12) del modello di Ebers-Moll, trascurando nella (6.12b) gli altri termini rispetto all'esponenziale di V_{BE}/V_T, e sostituendo a V_{BE} la differenza tra le tensioni V_1 e V_E ai due terminali, come:

$$I_{E1} \cong I_{ES} \exp\left(\frac{V_1 - V_E}{V_T}\right) ; \qquad I_{E2} \cong I_{ES} \exp\left(\frac{V_2 - V_E}{V_T}\right) \tag{9.1}$$

dove V_E è la tensione del punto comune ai due emettitori dei transistori.

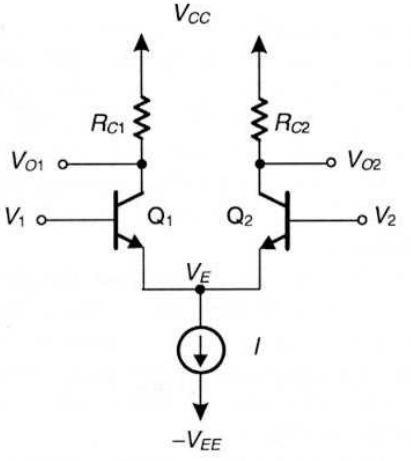


Figura 9.1 Schema di principio della configurazione differenziale

Dalla (9.1) il rapporto tra le due correnti può scriversi come:

$$\frac{I_{E1}}{I_{E2}} = \exp\left(\frac{V_1 - V_2}{V_T}\right) \tag{9.2}$$

da cui, aggiungendo 1 a primo e secondo membro della (9.2) e del suo inverso si ottiene:

$$\frac{I_{E1}+I_{E2}}{I_{E2}} = 1+\exp\left(\frac{V_1-V_2}{V_T}\right); \quad \frac{I_{E1}+I_{E2}}{I_{E1}} = 1+\exp\left(\frac{V_2-V_1}{V_T}\right) \tag{9.3}$$

Dalle (9.3), ricordando che $I_{E1} + I_{E2} = I$, e che le correnti di collettore sono legate a quelle di emettitore mediante α_F, si possono scrivere I_{C1} e I_{C2} come:

$$I_{C1} = \frac{\alpha_F I}{1+\exp\left(\frac{V_2-V_1}{V_T}\right)}; \quad I_{C2} = \frac{\alpha_F I}{1+\exp\left(\frac{V_1-V_2}{V_T}\right)} \tag{9.4}$$

Queste relazioni mostrano come le due correnti di collettore (e di emettitore) dipendano dalla *differenza* delle due tensioni $V_1 \equiv V_{B1}$ e $V_2 \equiv V_{B2}$ applicate alle basi.

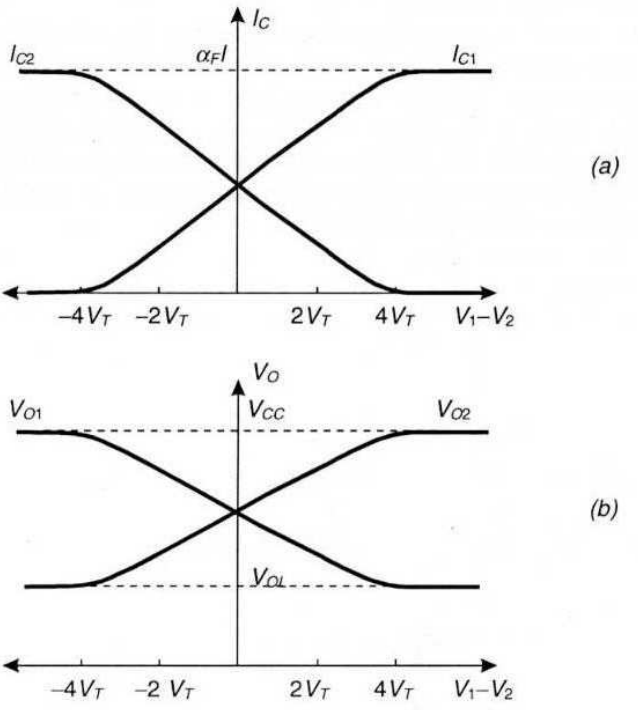


Figura 9.2 a) Andamento delle correnti di collettore nei due rami in funzione della differenza $V_1 - V_2$; b) andamento delle tensioni V_O alle due uscite

In Figura 9.2a sono riportati gli andamenti delle due correnti in funzione della differenza $V_1 - V_2$ delle tensioni ai due ingressi; da questi si vede come la

corrente I, che si divide in due parti uguali nei due rami per $V_1 = V_2$, tende a portarsi tutta nel ramo del transistore Q_1 per $V_1 - V_2 \cong 4V_T = 0.1$ V, e nel ramo di Q_2 per $V_1 - V_2 \cong -4V_T = -0.1$ V. In effetti la funzione del circuito nell'impiego come interruttore logico è proprio quella di deviare la corrente da un ramo all'altro del circuito a seconda di uno sbilanciamento rispettivamente positivo o negativo delle tensioni tra i due ingressi dell'ordine di 100 mV.

L'andamento delle tensioni di uscita $V_{O1,2}$ dei due rami segue quello delle correnti, a parte una inversione ed una traslazione, secondo la relazione: $V_O = V_{CC} - R_C I_C$ (nell'ipotesi di uguali resistenze R_C nei due rami). Se una corrente di collettore (ad esempio I_{C1}) è nulla, la corrispondente tensione (V_{O1}) sarà pari a V_{CC} ed il transistore (Q_1) è interdetto; quando la corrente I_C in uno dei due rami è la massima e pari a $\alpha_F I$, il transistore non sarà in saturazione se la resistenza R_C è scelta in modo che la differenza di tensione $V_{OMIN} - V_E >> V_{CESAT}$. Ciò comporta, come si vedrà meglio in seguito, una escursione della tensione in uscita significativamente ridotta rispetto alla tensione di alimentazione $V_{CC} + V_{EE}$, in quanto il valore minimo deve essere maggiore di V_E e quest'ultimo a sua volta deve essere superiore al valore $-V_{EE}$.

La limitazione nell'escursione logica del segnale in uscita è una caratteristica delle logiche non saturate, e comporta problemi sia per il pilotaggio delle porte a valle, che in generale per i margini di rumore, che sono inevitabilmente più bassi; tuttavia comporta anche un vantaggio in termini di prestazioni dinamiche del circuito, in quanto le capacità (sia dei dispositivi che di carico) con una data corrente di carica o di scarica impiegheranno un tempo minore per variare la tensione ai loro capi se queste variazioni sono più contenute.

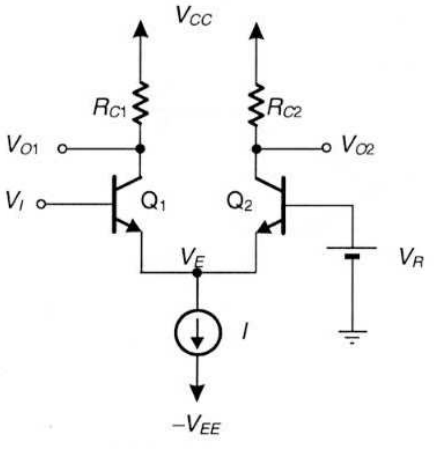


Figura 9.3 Invertitore elementare in logica non saturata

La porta più elementare di questa famiglia, ossia l'invertitore, può essere realizzata a partire dalla configurazione differenziale applicando una tensione di riferimento V_R ad uno degli ingressi ed il segnale logico V_I all'altro ingresso, come indicato in Figura 9.3; in tal caso la differenza degli ingressi corrisponde a $V_I - V_R$ dove V_R è una tensione fissa. In base ai risultati dell'analisi riportati in Figura 9.2, se il segnale V_I è inferiore a V_R di almeno 0.1 V (ingresso logico basso) il transistore Q_1 è interdetto e l'uscita V_{O1} sul collettore di Q_1 sarà alta, mentre se V_I è più grande di V_R di almeno 0.1 V, Q_1 sarà in conduzione e l'uscita V_{O1} sarà bassa.

Dallo schema di principio di Figura 9.3 si può notare come in questo circuito siano in realtà disponibili sia la variabile negata dell'ingresso (all'uscita 1) che quella non negata (all'uscita 2); questa prerogativa delle porte logiche non saturate è molto utile nella realizzazione di funzioni logiche più complesse per le quali saranno disponibili le uscite con variabile sia diretta che negata.

Da un'analisi elementare del circuito di Figura 9.3 si ricava che le tensioni di uscita nei due stati logici non possono essere utilizzate come ingressi per porte successive della stessa famiglia, in altre parole non sono compatibili con i valori logici necessari per l'ingresso; infatti, assumendo Q_2 in conduzione (ingresso V_I basso) si ha per l'uscita logica bassa V_{OL2} :

$$V_{OL2} = V_{C2\,MIN} > V_R \tag{9.5}$$

dove l'ultima disequazione si riferisce al fatto che, per garantire che Q_2 operi fuori dalla saturazione, la giunzione base-collettore deve essere polarizzata negativamente. D'altra parte, perché l'uscita V_{OL} sia utilizzabile come ingresso basso per pilotare una porta successiva dello stesso tipo, occorre verificare che $V_{OL} < V_{IL}$ (massima tensione di ingresso nello stato basso). Questa tensione può essere definita come la massima tensione di ingresso per la quale Q_1 sia ancora in interdizione; da quanto detto precedentemente ne risulta quindi:

$$V_{OL2} < V_{IL} = V_R - 0.1\ \text{V} \tag{9.6}$$

e le due disequazioni (9.5) e (9.6) non possono essere simultaneamente soddisfatte.

Per realizzare quindi l'invertitore elementare con lo schema differenziale occorre traslare le tensioni sia in ingresso che in uscita, in modo da verificare le condizioni di pilotaggio necessarie per ogni porta logica, e cioè segnale di ingresso nello stato basso $V_I\,(0) \equiv V_{OL} < V_{IL}$ e segnale di ingresso nello stato alto $V_I\,(1) \equiv V_{OH} > V_{IH}$.

Queste considerazioni portano allo schema della porta logica su cui è costruita la famiglia logica bipolare non saturata più diffusa, definita come Logica ad Accoppiamento di Emettitore o ECL (*Emitter Coupled Logic*), che sarà discussa nel seguito, ed alle sue varianti, indicate come *logiche a commutazione di corrente* (*Current Mode Logic*, CML), in quanto il passaggio da un livello logico all'altro è effettuato attraverso una deviazione della corrente da un ramo all'altro della configurazione.

9.2 Invertitore elementare in logica ECL

Lo schema dell'invertitore elementare ECL è riportato in Figura 9.4. In esso possiamo identificare tre sezioni: la configurazione differenziale, che effettua l'operazione di inversione (e il suo negato), il circuito generatore della tensione di riferimento V_R, e gli stadi disaccoppiatori (e traslatori di tensione) sulle due uscite della configurazione differenziale.

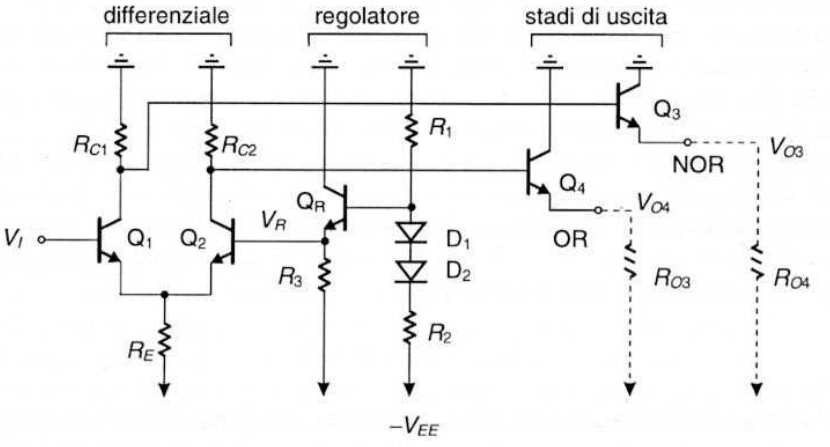


Figura 9.4 Schema della porta elementare ECL

In questo schema si può notare che la tensione di riferimento positiva V_{CC} è stata posta a massa, lasciando solo l'alimentazione negativa $-V_{EE}$. Ciò viene fatto per operare una traslazione dei livelli di tensione in uscita, ma comporta anche un effetto positivo sulla immunità ai disturbi dei livelli logici, come vedremo in seguito a valle dell'analisi del circuito ECL.

Un'ulteriore modifica nello stadio differenziale consiste nella realizzazione del generatore di corrente sugli emettitori mediante la rete formata dalla resistenza R_E e dall'alimentazione $-V_{EE}$. Questa rete in effetti non fornisce una corrente I rigorosamente costante al variare del segnale di ingresso V_I, ma come vedremo le variazioni di corrente possono essere abbastanza contenute.

L'alimentazione per sole tensioni negative ha come conseguenza che le grandezze di ingresso e di uscita sono tutte negative rispetto al riferimento di massa; quindi i livelli logici alti (1 logico) corrispondono a tensioni negative, ed in modulo minori di quelle (più negative) corrispondenti ai livelli logici bassi (0 logico). Questa inversione rispetto alla convenzione adottata nelle porte precedenti in cui si associano i livelli logici "1" alle tensioni elevate e quelli "0" alle tensioni basse viene indicata come "logica negativa", e richiede una conversione alla logica positiva nel caso di utilizzazione combinata di porte logiche ECL con quelle TTL o CMOS.

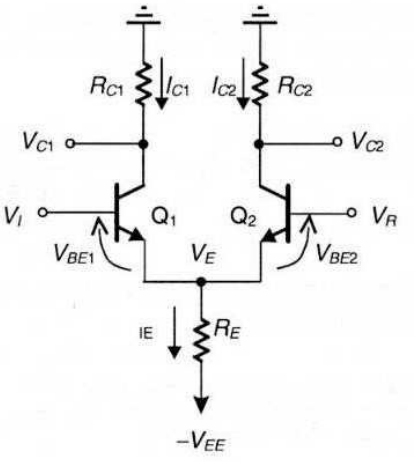


Figura 9.5 Analisi dello stadio differenziale ECL

Analizziamo per primo lo stadio differenziale (Figura 9.5), per valutare le condizioni di funzionamento nei due stati limite corrispondenti alla conduzione di un solo transistore della coppia differenziale. Per un segnale in ingresso $V_I = V_R - 0.1$ V (ricordiamo che in questo circuito sia V_R che V_I e V_O hanno valori negativi), Q_1 va all'interdizione e la corrente I del ramo comune ai due emettitori coincide con quella di emettitore di Q_2:

$$I \equiv I_{E2} = \frac{V_E' - (-V_{EE})}{R_E} = \frac{V_R - V_{BE2} + V_{EE}}{R_E} \tag{9.7}$$

dove V_E' è la tensione del nodo comune ai due emettitori quando la corrente circola solo in Q_2.

Per $V_I = V_R + 0.1$ V, la situazione si inverte: Q_1 conduce e Q_2 va in interdizione; la tensione nel nodo comune V_E'' sarà diversa da quella del caso precedente, e la corrente I vale:

$$I \equiv I_{E1} = \frac{V_E'' - (-V_{EE})}{R_E} = \frac{V_R - V_{BE\gamma 2} + V_{EE}}{R_E} \tag{9.8}$$

La variazione di corrente tra questi due ingressi è quindi:

$$\Delta I = \frac{V_{BE} - V_{BE\gamma}}{R_E} = \frac{0.1\text{V}}{R_E} \; ; \quad \frac{\Delta I}{I} \cong \frac{0.1\text{V}}{V_R - V_{BE} + V_{EE}} \tag{9.9}$$

che, con i valori tipici per V_R e V_{EE} utilizzati nelle porte, fornisce una variazione relativa di corrente dell'ordine del 2%.

9.2 Invertitore elementare in logica ECL

Lo schema dell'invertitore elementare ECL è riportato in Figura 9.4. In esso possiamo identificare tre sezioni: la configurazione differenziale, che effettua l'operazione di inversione (e il suo negato), il circuito generatore della tensione di riferimento V_R, e gli stadi disaccoppiatori (e traslatori di tensione) sulle due uscite della configurazione differenziale.

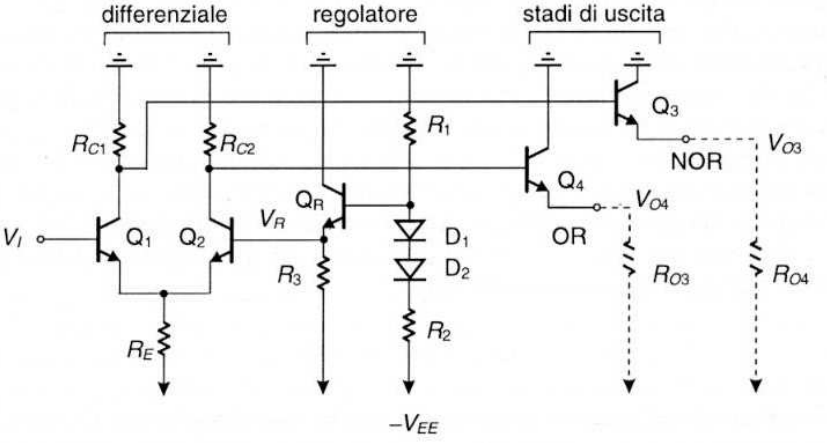


Figura 9.4 Schema della porta elementare ECL

In questo schema si può notare che la tensione di riferimento positiva V_{CC} è stata posta a massa, lasciando solo l'alimentazione negativa $-V_{EE}$. Ciò viene fatto per operare una traslazione dei livelli di tensione in uscita, ma comporta anche un effetto positivo sulla immunità ai disturbi dei livelli logici, come vedremo in seguito a valle dell'analisi del circuito ECL.

Un'ulteriore modifica nello stadio differenziale consiste nella realizzazione del generatore di corrente sugli emettitori mediante la rete formata dalla resistenza R_E e dall'alimentazione $-V_{EE}$. Questa rete in effetti non fornisce una corrente I rigorosamente costante al variare del segnale di ingresso V_I, ma come vedremo le variazioni di corrente possono essere abbastanza contenute.

L'alimentazione per sole tensioni negative ha come conseguenza che le grandezze di ingresso e di uscita sono tutte negative rispetto al riferimento di massa; quindi i livelli logici alti (1 logico) corrispondono a tensioni negative, ed in modulo minori di quelle (più negative) corrispondenti ai livelli logici bassi (0 logico). Questa inversione rispetto alla convenzione adottata nelle porte precedenti in cui si associano i livelli logici "1" alle tensioni elevate e quelli "0" alle tensioni basse viene indicata come "logica negativa", e richiede una conversione alla logica positiva nel caso di utilizzazione combinata di porte logiche ECL con quelle TTL o CMOS.

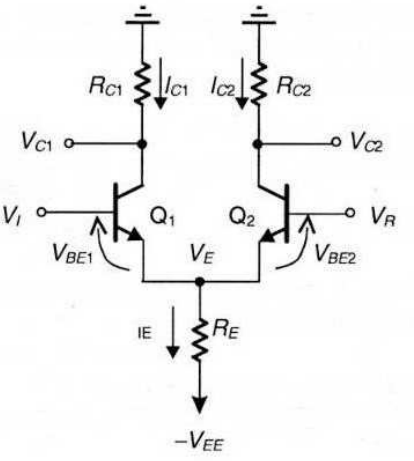


Figura 9.5 Analisi dello stadio differenziale ECL

Analizziamo per primo lo stadio differenziale (Figura 9.5), per valutare le condizioni di funzionamento nei due stati limite corrispondenti alla conduzione di un solo transistore della coppia differenziale. Per un segnale in ingresso $V_I = V_R - 0.1$ V (ricordiamo che in questo circuito sia V_R che V_I e V_O hanno valori negativi), Q_1 va all'interdizione e la corrente I del ramo comune ai due emettitori coincide con quella di emettitore di Q_2:

$$I \equiv I_{E2} = \frac{V_E' - (-V_{EE})}{R_E} = \frac{V_R - V_{BE2} + V_{EE}}{R_E} \tag{9.7}$$

dove V_E' è la tensione del nodo comune ai due emettitori quando la corrente circola solo in Q_2.

Per $V_I = V_R + 0.1$ V, la situazione si inverte: Q_1 conduce e Q_2 va in interdizione; la tensione nel nodo comune V_E'' sarà diversa da quella del caso precedente, e la corrente I vale:

$$I \equiv I_{E1} = \frac{V_E'' - (-V_{EE})}{R_E} = \frac{V_R - V_{BE\gamma 2} + V_{EE}}{R_E} \tag{9.8}$$

La variazione di corrente tra questi due ingressi è quindi:

$$\Delta I = \frac{V_{BE} - V_{BE\gamma}}{R_E} = \frac{0.1\text{V}}{R_E} \; ; \quad \frac{\Delta I}{I} \equiv \frac{0.1\text{V}}{V_R - V_{BE} + V_{EE}} \tag{9.9}$$

che, con i valori tipici per V_R e V_{EE} utilizzati nelle porte, fornisce una variazione relativa di corrente dell'ordine del 2%.

Dalle (9.7) e (9.8) si desume che le correnti di collettore quando un solo transistore è in conduzione non sono esattamente uguali per i due transistori, ed in particolare la corrente I_{C1} è leggermente maggiore di I_{C2} in quanto:

$$I_{C1} = \alpha_F I_{E1} = \alpha_F \frac{V_R - V_{BE\gamma(2)} + V_{EE}}{R_E}; I_{C2} = \alpha_F I_{E2} = \alpha_F \frac{V_R - V_{BE(2)} + V_{EE}}{R_E} \tag{9.10}$$

quindi, per ottenere eguali escursioni di tensioni alle due uscite occorre scegliere il rapporto tra le resistenze R_C dei due rami tale che:

$$\frac{R_{C2}}{R_{C1}} = \frac{I_{C1}}{I_{C2}} = \frac{V_R - V_{BE\gamma(2)} + V_{EE}}{V_R - V_{BE(2)} + V_{EE}} \tag{9.11}$$

il che comporta un valore di R_{C1} leggermente inferiore a R_{C2}. Con questa condizione le tensioni V_{OL} per ognuna delle due uscite saranno uguali e pari a:

$$V_{OL} = -R_{C2} I_{C2} = -R_{C1} I_{C1} \tag{9.12}$$

e, ricordando le (9.10), possono essere scritte come:

$$V_{OL} = -R_{C2}\alpha_F \frac{V_R - V_{BEon} + V_{EE}}{R_E} = -R_{C1}\alpha_F \frac{V_R - V_{BE\gamma} + V_{EE}}{R_E} \tag{9.13}$$

Anche le escursioni della tensione di uscita per ognuna delle due uscite saranno uguali, in quanto $V_{OH} = 0$ per ogni uscita.

Per il corretto funzionamento del circuito occorre garantire che i transistori non vadano in saturazione nella condizione di uscita bassa V_{OL}. Assumendo quindi una tensione tra base e collettore $V_{BC} \leq 0$ V per garantirsi un margine sufficiente rispetto alla condizione di saturazione, si ha che $V_{CE} \geq 0.7$ V. Dall'analisi del circuito nel caso di Q_2 in conduzione si ha:

$$V_{OL} \geq V_E + 0.7V; \quad \text{con} \quad V_{OL} = -I_{C2} R_{C2} \cong -I_{E2} R_{C2} \tag{9.14}$$

Sostituendo nella (9.14) il valore di I_{E2} dato dalla (9.7), e ricordando che $V_E = V_R - V_{BE}$, si ottiene:

$$-\frac{R_{C2}}{R_E}(V_R - V_{BE2} + V_{EE}) \geq V_R - V_{BE2} + 0.7V \tag{9.15}$$

da cui, ricordando che $V_{BE2} = 0.7$ V, si ha una relazione tra R_C e R_E in funzione delle tensioni V_R e V_{EE}:

$$\frac{R_{C2}}{R_E} \leq \frac{V_R}{V_R - V_{BE2} + V_{EE}} \tag{9.16}$$

Questo rapporto è quindi inferiore all'unità, e va determinato in funzione del valore di V_R. Quest'ultimo valore viene essenzialmente definito in base alla condizione di uguali margini di rumore per i due stati logici, condizione essenziale per un buon funzionamento della logica ECL, per bilanciare la limitazione dovuta al ridotto swing logico.

La condizione sui margini di rumore comporta che i valori V_{OL} e V_{OH} dell'uscita siano rispettivamente equidistanti dai valori degli ingressi V_{IL} e V_{IH}, come indicato in Figura 9.6; in altri termini ciò comporta (vedi Paragrafo 1.4) che la soglia logica V_{SL} del circuito nel piano della funzione di trasferimento sia centrata sia rispetto al segmento $V_{IH} - V_{IL}$ che a quello $V_{OH} - V_{OL}$.

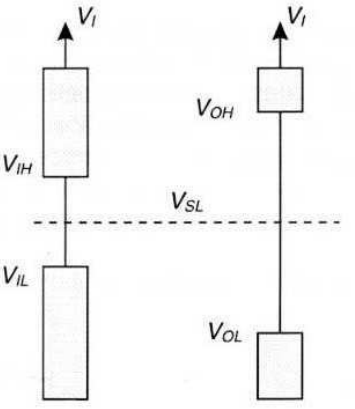


Figura 9.6 Soglia logica ottimale per un invertitore ECL

Nel circuito differenziale ECL il valore mediano rispetto alle tensioni V_{IH} e V_{IL} è dato proprio dalla tensione V_R, che è in effetti quel valore della tensione di ingresso in cui la funzione di trasferimento per ognuna delle due uscite passa per il valore intermedio tra V_{OH} e V_{OL}. I valori V_{IL} e V_{IH} corrispondono con buona approssimazione ai punti $V_R - 0.1$ V e $V_R + 0.1$ V per i quali rispettivamente Q_1 o Q_2 è al limite dell'interdizione; infatti per valori V_I maggiori (minori) di $V_R - 0.1$ V ($V_R + 0.1$ V) l'uscita dipende dall'ingresso e presenta una pendenza maggiore di -1 in quanto l'amplificazione differenziale è ben maggiore dell'unità.

Si può facilmente verificare che, col solo circuito invertitore di Figura 9.5, la tensione V_R non può essere anche centrata rispetto ai valori di V_{OH} e V_{OL}, se il generatore della tensione di riferimento è realizzato con il circuito a collettore comune

di Figura 9.4. Infatti la tensione minima V_{OL} sul collettore di Q_1 si ha in corrispondenza dell'ingresso alto $V_I = V_{OH} = 0$ V. Per mantenere Q_1 in regione attiva anche in questa condizione occorre che:

$$V_{IMAX} - V_{OL} \equiv V_{OH} - V_{OL} = -V_{OL} \leq V_{BC\gamma} \cong 0.6 \text{ V}$$

il che porta a definire come minima tensione bassa in uscita il valore $V_{OL} = -0.6$ V. Per avere una simmetria delle tensioni di uscita nei due stati rispetto al valore V_R deve essere:

$$V_R = \frac{V_{OH} + V_{OL}}{2} = \frac{V_{OL}}{2} \cong -0.3 \text{ V}$$

e questo valore di V_R non è compatibile con il circuito del generatore di tensione, perché la tensione V_R deve essere almeno pari a $-V_{BE}$ (−0.7 V) anche considerando nulla la caduta su R_1.

Gli stadi di uscita formati da Q_3 e Q_4 connessi alle due uscite del differenziale servono quindi a traslare in basso i livelli delle tensioni di uscita V_{OH} e V_{OL} in modo da realizzare la richiesta condizione di simmetria rispetto al valore di V_R. Inoltre la traslazione verso il basso della tensione V_{OH} permette di abbassare ulteriormente il valore di V_{CMIN} compatibile con il funzionamento in regione attiva, e quindi di aumentare in definitiva l'escursione logica in uscita, come vedremo meglio nel Paragrafo 9.3, analizzando la caratteristica di trasferimento complessiva del circuito.

Assumendo quindi che, in uscita dall'invertitore ECL dello schema di Figura 9.4 si abbiano i seguenti valori dei livelli logici:

$$V_{OH} = -V_{BE2,4}; \quad V_{OL} = V_{CMIN} - V_{BE2,4}$$

la condizione di uguali margini di rumore comporta:

$$\frac{V_{OH} + V_{OL}}{2} = \frac{-V_{BE} + (V_{CMIN} - V_{BE})}{2} = V_R$$

Imponendo il vincolo ulteriore che il valore minimo di collettore (sul collettore di Q_2) sia superiore al valore V_R per evitare l'entrata in saturazione di Q_2 (si assume un valore di $V_{BCMAX} = -0.2$ V per garantire un margine di sicurezza sufficiente rispetto all'entrata in saturazione di Q_2), si ottiene per il valore di V_R:

$$\frac{-V_{BE} + (V_{CMIN} - V_{BE})}{2} = \frac{-2V_{BE} + V_R + 0.2V}{2} = V_R \Rightarrow V_R = -1.2V \qquad (9.17)$$

Dalle considerazioni precedentemente esposte, risulta un numero di vincoli al progetto dello stadio differenziale che porta a valori abbastanza determinati per le resistenze del circuito di Figura 9.7, una volta scelta la tensione di alimentazione.

In particolare si può determinare un valore massimo per il rapporto tra le resistenze R_C e R_E in base alla (9.16), assumendo un valore di $V_R \cong -1.2$ V, ed un valore di $-V_{EE} = -5.2$ V, tipico delle porte ECL:

$$\frac{R_{C2}}{R_E} \leq \frac{V_R}{V_R - V_{BE2} + V_{EE}} \leq 0.36 \tag{9.18}$$

(in pratica si assume un valore del rapporto pari a 0.3).

Assumendo per il transistore in conduzione $I_{CMAX} \cong I_E$ (vedi Figura 9.7), si ha che le cadute di tensione su R_C e R_E stanno tra loro nello stesso rapporto delle resistenze, e cioè $|V_{CMIN}| \cong 0.3\ (V_E + V_{EE})$. A queste si deve aggiungere, per bilanciare la tensione di alimentazione $-V_{EE}$, la tensione $V_{BE} \cong 0.7$ V e quella $V_{CB} \geq 0.2$ V. Assumendo quest'ultima pari a $\cong 0.3$ V, per garantirsi un margine di sicurezza, rimangono $\cong 4.2$ V a disposizione per le due cadute, che, in base al vincolo di un rapporto minimo di 0.3 già ricordato, vengono divise in 1 V su R_C e 3.2 V su R_E. Si ha quindi $V_{CMIN} = -R_C I_C \cong -1$ V, $V_R \cong -1.3$ V, $V_{CEMAX} = 0$ V. Infine, le tensioni sugli emettitori degli stadi di uscita, che operano come amplificatori a collettore comune, sono traslate di $-V_{BE} = -0.7$ V rispetto agli ingressi V_{C1} e V_{C2}.

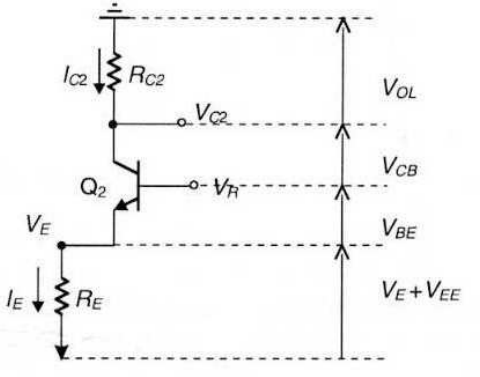


Figura 9.7 Analisi delle cadute sul ramo in conduzione del differenziale

La corrente I_E viene scelta tipicamente pari a 4 mA, per cui, in base alle cadute di tensione determinate dall'analisi precedente, i valori delle resistenze saranno: $R_C \cong 250\ \Omega$, $R_E \cong 800\ \Omega$.

9.3 Caratteristiche di trasferimento della porta ECL

In base all'analisi effettuata nel paragrafo precedente, è possibile ora determinare le caratteristiche di trasferimento dell'intero circuito di Figura 9.4; quella relativa all'uscita V_{O4} (detta uscita OR, come vedremo in seguito), e quella relativa all'uscita V_{O3} (detta uscita NOR); ricordiamo che entrambe le caratteristiche vengono riportate nel 3° quadrante del piano V_I, V_O perché sia le tensioni di ingresso che quelle di uscita hanno valori negativi.

9.3.1 Uscita OR

La caratteristica di trasferimento dell'uscita OR ottenuta da una simulazione SPICE del circuito ECL è riportata in Figura 9.8. Per $V_I = V_{OH}$ (ingresso alto) il transistore Q_2 deve trovarsi in interdizione, e la tensione di collettore di Q_2 è $V_{C2} = 0$ V; l'uscita della porta, presa sull'emettitore di Q_4 che agisce da stadio di uscita, viene traslata di $-V_{BE} \cong -0.7$ V, per cui $V_{OH} = -0.7$ V. Con questo valore è possibile verificare che l'assunto di Q_2 in interdizione è corretto: infatti se $V_R \cong -1.2$ V, la tensione di ingresso $V_I = -0.7$ V è maggiore di $V_R + 0.1$ V e quindi Q_1 è in conduzione e Q_2 è interdetto.

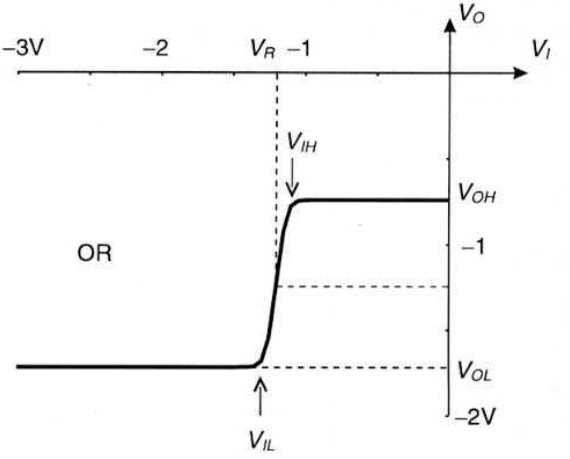


Figura 9.8 Caratteristica di trasferimento dell'uscita OR mediante simulazione SPICE; i valori delle resistenze sono $R_{C1} = 220\ \Omega$, $R_{C2} = 245\ \Omega$, $R_E = 800\ \Omega$

Al diminuire di V_I (i valori delle tensioni sono negativi), si raggiunge il valore $V_{IH} \cong V_R + 0.1$ V $\cong -1.1$ V, sotto il quale comincia a condurre Q_2.

Quando V_I scende sotto il valore $V_{IL} \cong V_R - 0.1$ V $\cong -1.3$ V, Q_1 va in interdizione e la tensione di uscita sul collettore di Q_2 raggiunge il valore minimo $V_{C2MIN} \cong -1$ V. La tensione di uscita V_{OL} della porta sarà data da $V_{OL} = V_{C2MIN} - V_{BE4} \cong -1.7$ V.

L'uscita rimane al valore basso V_{OL} anche con un'ulteriore diminuzione di V_I fino allo stesso valore.

Come si è detto nel Paragrafo precedente, questi valori di V_{OH} e V_{OL} permettono di scegliere un valore di V_R tale che $V_R = (V_{OH} + V_{OL})/2$ e che sia anche compatibile con il vincolo imposto dal circuito del regolatore di tensione, e cioè $V_R < -V_{BE}$; infatti dai valori precedenti si ha:

$$V_R = \frac{V_{OH} + V_{OL}}{2} = \frac{-0.7 + (-1.7)}{2} = -1.2 \text{ V} \tag{9.19}$$

valore compatibile sia con l'analisi del circuito differenziale svolta precedentemente, che con quella del circuito regolatore che effettueremo successivamente (in realtà poiché nelle porte ECL le aree dei transistori sono minori che per quelle TTL, questi lavorano a densità di corrente più elevate e le cadute V_{BE} sono dell'ordine di $0.75 \div 0.8$ V; questo comporta una tensione V_R più negativa del valore calcolato nella (9.19) e tipicamente pari a -1.3 V, ma l'analisi che svolgeremo rimane sostanzialmente valida).

I valori significativi della curva sono quindi:

$$V_{OH} = -V_{BE4} \cong -0.7 \text{ V} ; \quad V_{OL} = -I_{C2} R_{C2} - V_{BE4} \cong -1.7 \text{ V} \tag{9.20a}$$

$$V_{IH} = V_R + 0.1 \text{ V} \cong -1.1 \text{ V} ; \quad V_{IL} = V_R - 0.1 \text{ V} \cong -1.3 \text{ V} \tag{9.20b}$$

I margini di rumore del circuito nei due stati logici saranno quindi uguali e varranno:

$$NM_H = V_{OH} - V_{IH} = -0.7 - (-1.1) = 0.4 \text{ V} \tag{9.21a}$$

$$NM_L = V_{IL} - V_{OL} = -1.3 - (-1.7) = 0.4 \text{ V} \tag{9.21b}$$

9.3.2 Uscita NOR

La caratteristica di trasferimento riferita all'uscita su Q_3 (uscita NOR) non differisce dalla precedente solo per l'inversione dei valori logici del segnale di uscita V_{O3} rispetto all'ingresso V_I, ma anche per il comportamento del transistore Q_1 nella regione di conduzione (vedi Figura 9.9).

In questo caso Q_1 è in interdizione se $V_I = V_{OL}$ (ingresso basso), e rimane interdetto finché V_I aumenta fino al valore $V_{IL} = V_R - 0.1 \text{ V} \cong -1.3$ V; la tensione sul collettore è quindi $V_{C1MAX} = 0$ V. Per $V_I = V_{IH} = V_R + 0.1 \text{ V} \cong -1.1$ V, Q_2 va in interdizione e la tensione V_{C1} sul collettore di Q_1 vale -1 V. All'aumentare di V_I oltre V_{IH} la tensione V_{C1} diminuisce ulteriormente, in quanto la corrente I_{E1} può crescere oltre il valore indicato dalla (9.8) se V_I aumenta oltre il valore V_{IH}; il ramo del differenziale con Q_1 si comporta in queste condizioni come un amplificatore a doppio

carico (R_{C1} e R_E), analogamente al caso dell'invertitore Q_d della porta TTL. Nell'ipotesi di $I_C \cong I_E$ si ha:

$$\Delta V_{C1} = -\Delta V_I \frac{R_{C1}}{R_E}; \quad \Rightarrow \quad \frac{\Delta V_{C1}}{\Delta V_I} = -\frac{R_{C1}}{R_E} \tag{9.22}$$

e quindi la caratteristica di trasferimento presenta una pendenza negativa pari a circa 0.3, a partire dalla tensione di ingresso V_{IH}. All'aumentare di V_I si raggiunge infine un valore V_S che porta il transistore Q_1 in saturazione; infatti se V_I aumenta e V_{C1} diminuisce si raggiunge la condizione per cui:

$$V_{B1} - V_{C1} = V_S + I_{C1}R_{C1} = V_{BC\gamma}; \quad V_S = V_{BC\gamma} - I_{C1}R_{C1} \cong 0.6\text{V} - 1\text{V} \cong -0.4\text{V}$$

Da questo punto la caratteristica presenta una pendenza positiva e circa unitaria, in quanto le variazioni dell'uscita seguono quelle dell'ingresso perché in saturazione sia V_{BC1} che V_{BE3} rimangono all'incirca costanti. L'ingresso in saturazione di Q_1 oltre il valore V_S non è in contrasto con l'assunzione di funzionamento della porta in regime non saturato, perché in condizioni normali di funzionamento il segnale di ingresso massimo è – 0.7 V e non raggiunge il valore V_S necessario per la saturazione di Q_1.

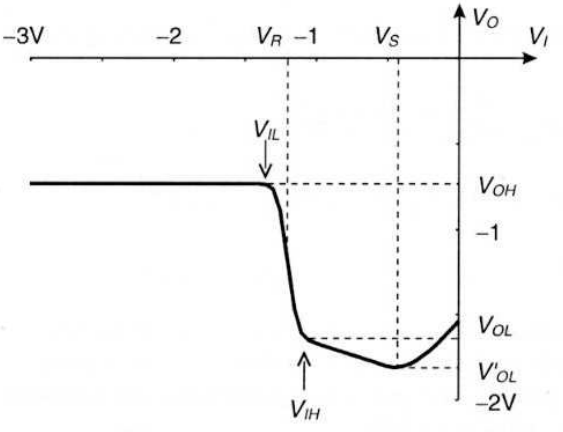


Figura 9.9 Caratteristica di trasferimento dell'uscita NOR ottenuta con SPICE per lo stesso caso di Figura 9.8

Se si considera come uscita V_{OL} quella corrispondente alla tensione di ingresso V_{IH} (condizione più conservativa nei riguardi dell'escursione logica), si ottengono anche per questa uscita margini di rumore uguali tra loro e pari a quelli definiti per l'uscita OR.

9.4 Lo stadio regolatore di tensione

Questo stadio deve fornire il valore della tensione di riferimento V_R necessaria ad un corretto funzionamento della porta logica, e cioè una tensione pari al valore medio tra V_{OH} e V_{OL}. La tensione è ottenuta come uscita di un transistore montato a collettore comune (vedi Figura 9.10), e polarizzato in ingresso dalla rete formata dalle resistenze R_1, R_2 ed i due diodi D_1, D_2 (vedremo successivamente che il ruolo dei diodi è quello di compensare gli effetti delle variazioni termiche sui livelli logici).

Trascurando la corrente di base di Q_R rispetto a quella del partitore, la tensione V_B (riferita a massa) in base al circuito di Figura 9.10 vale:

$$V_B = \frac{2V_D - V_{EE}}{R_1 + R_2} R_1$$

La tensione V_R vale quindi:

$$V_R = V_B - V_{BE} = \frac{2V_D - V_{EE}}{R_1 + R_2} R_1 - V_{BE} = -3.8 \frac{R_1}{R_1 + R_2} - 0.7\,\text{V} \qquad (9.23)$$

Con un'opportuna scelta di R_1 e R_2 si può ottenere qualunque valore di V_R compreso tra – 0.7 V e – 4.5 V. Conviene, come si è già detto, scegliere come valore di V_R il valore medio tra V_{OH} e V_{OL}, ossia V_R = – 1.2 V. Con questo valore, in base alla (9.24), si ottiene un valore del rapporto:

$$\frac{R_1}{R_1 + R_2} = 0.13 \qquad (9.24)$$

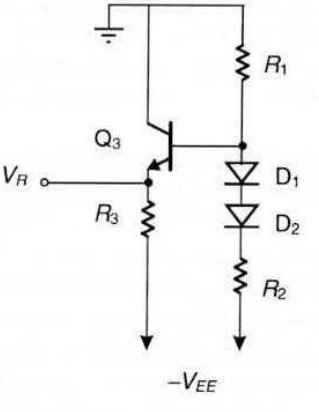


Figura 9.10 Circuito per la generazione della tensione V_R

Nelle porte ECL commerciali viene scelto un valore del rapporto $R_1/(R_1+R_2) = 0.15$, in modo da compensare le cadute ohmiche della giunzione B-E che comportano valori di V_{BE} più elevati rispetto a quelli teorici. I valori di R_1 e R_2 di queste porte vengono scelti pari a $R_1 \cong 900\ \Omega$, $R_2 \cong 5\ k\Omega$.

Il circuito regolatore presenta una bassa resistenza di uscita ed un'elevata corrente disponibile in quanto sfrutta una configurazione a collettore comune; ciò consente di utilizzare un solo circuito regolatore per fornire la tensione di riferimento V_R a più differenziali (e cioè a più porte ECL elementari) con guadagno di spazio e di potenza dissipata. Questo circuito svolge anche un ruolo essenziale sulla stabilizzazione dei livelli logici (e dei margini di rumore) nei riguardi sia delle variazioni della temperatura che di quelle della tensione di alimentazione, come si vedrà nei Paragrafi successivi.

9.4.1 Analisi del comportamento termico

Nelle porte bipolari precedentemente analizzate (RTL, DTL, TTL) non si è posta particolare attenzione agli effetti delle variazioni di temperatura sui livelli logici, nonostante che i transistori bipolari siano dispositivi molto sensibili alla temperatura. In effetti in queste porte i transistori operano come interruttori pilotati, cioè lavorano nelle condizioni limite di interdizione o di saturazione, dove gli effetti della temperatura sul punto di funzionamento sono limitati. Nelle porte a logica non saturata invece gli effetti delle variazioni di temperatura sui livelli logici assumono particolare importanza, per due ragioni: a) i dispositivi in una delle due condizioni stabili di funzionamento lavorano in regione attiva, in condizioni di funzionamento che sono molto sensibili alla temperatura; b) le escursioni del segnale logico sono relativamente limitate, per cui gli effetti della temperatura sulle condizioni di funzionamento hanno maggiore rilevanza. In queste porte è quindi necessario valutare gli effetti della temperatura, e prevedere tecniche di compensazione al fine di ridurre questi effetti sui livelli logici dei segnali. La presenza dei diodi nella rete di polarizzazione del generatore della tensione di riferimento V_R è proprio prevista per bilanciare gli effetti delle variazioni di temperatura sulle caratteristiche statiche della porta, in modo da non peggiorare i già ridotti margini di rumore al variare della temperatura di funzionamento della porta.

Le cause principali delle variazioni delle grandezze caratteristiche della porta (V_{OH}, V_{OL}, V_{IH}, V_{IL}) con la temperatura sono dovute alle variazioni delle tensioni V_{BE} delle giunzioni base-emettitore dei transistori che, come è noto, dipendono dalla temperatura con un coefficiente lineare negativo pari a circa -2 mV/°C. Non potendo eliminare le dipendenze delle tensioni V_{BE} dalla temperatura, si è provveduto all'introduzione dei diodi nella rete di polarizzazione del circuito che fornisce la tensione di riferimento, in modo da dar luogo, attraverso la stessa dipendenza della tensione V_D del diodo dalla temperatura, un'opportuna variabilità di V_R. In questo modo è possibile ottenere che quest'ultima, anche al variare della temperatura, sia sempre centrata rispetto ai valori (variabili inevitabilmente con la temperatu-

ra) di V_{OH} e V_{OL}; ricordiamo che V_{IH} e V_{IL} sono in ogni caso legati alla V_R e centrati rispetto a questo valore.

Per comprendere l'effetto di compensazione di questa rete si può effettuare un'analisi di quest'ultima alle variazioni, assumendo gli ingressi elettrici costanti e facendo variare la temperatura ambiente; in questo caso si propagheranno nel circuito solo le variazioni di tensione dovute alle variazioni termiche delle cadute sulle giunzioni (e sui diodi). Definita quindi una data variazione di temperatura ΔT, ogni giunzione subirà una variazione di tensione $\Delta V = -2\ \text{mV}\cdot\Delta T$ e si comporterà come un generatore di tensione in un circuito che considera solo le variazioni di tensione. Un'analisi corretta per questo caso richiede la trattazione della rete in termini di circuito equivalente linearizzato, utilizzando per i transistori i loro circuiti equivalenti per piccoli segnali; tuttavia per valutare gli effetti di compensazione è sufficiente un'analisi più approssimata che considera le correnti di base trascurabili rispetto a quelle di collettore e di emettitore dei transistori, e quindi permette di separare agevolmente le maglie ed eliminare l'interazione (del secondo ordine) tra uscite ed ingressi.

a) Analisi di $V_{OH}(T)$

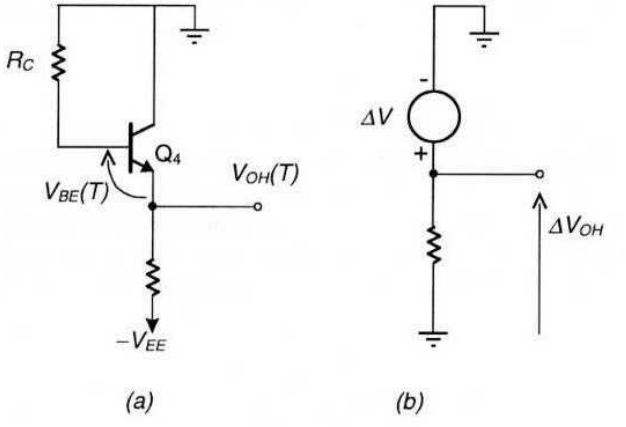


Figura 9.11 a) Analisi degli effetti della temperatura su V_{OH}; b) circuito equivalente per le variazioni termiche

Il circuito che è coinvolto per gli effetti termici nella condizione di uscita alta V_{OH} è quello riportato in Figura 9.11a, valido sia per l'uscita OR che NOR. L'unica sorgente di variazioni di tensione dovute alla temperatura è la giunzione B-E di $Q_4(Q_3)$ ed il circuito equivalente semplificato è quello riportato in Figura 9.11b (assumendo trascurabile la corrente di base e con questa la caduta su R_C). Da questo si ricava subito che la variazione di V_{OH} è:

$$\Delta V_{OH} \cong \Delta V \tag{9.25}$$

b) Analisi di $V_R(T)$

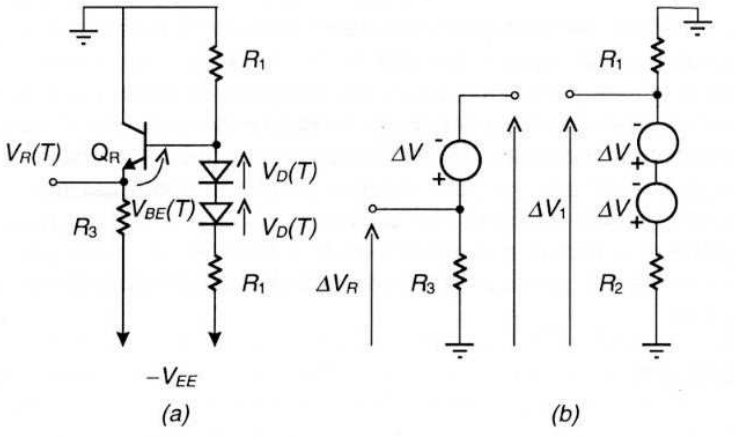


Figura 9.12 a) Analisi degli effetti della temperatura su V_R; b) circuito equivalente per le variazioni termiche

Il circuito che fornisce la tensione di riferimento V_R è indicato in Figura 9.12a, mentre in Figura 9.12b è riportato il suo circuito equivalente semplificato per le variazioni termiche, nel quale si è separata la maglia di base da quella di emettitore, in accordo con l'ipotesi di considerare trascurabile la corrente di base. In questo caso le grandezze dipendenti dalla temperatura sono le tensioni sui diodi e la V_{BE} di Q_R. La rete di polarizzazione fornisce ai capi di R_1 la tensione:

$$\Delta V_1 = -2\Delta V \frac{R_1}{R_1 + R_2}$$

(anche in questo caso si trascura la corrente di base di Q_R rispetto alla corrente che fluisce nel partitore). Questa variazione di tensione viene trasmessa alla base di Q_R, e si somma alla variazione di tensione della giunzione base-emettitore, per cui la variazione di V_R, ricordando il valore del rapporto $R_1 / (R_1 + R_2)$ dato dalla (9.24), sarà data da:

$$\Delta V_R \cong \Delta V_1 + \Delta V = \Delta V\left(1 - 2\frac{R_1}{R_1 + R_2}\right) = \Delta V(1 - 2 \cdot 0.13) = 0.74\Delta V \tag{9.26}$$

c) Analisi di $V_{OL}(T)$

In questo caso sono coinvolti sia il ramo $Q_2(Q_1)$ del differenziale che lo stadio di uscita (Figura 9.13a). Il circuito equivalente per le variazioni termiche è riportato in Figura 9.13b.

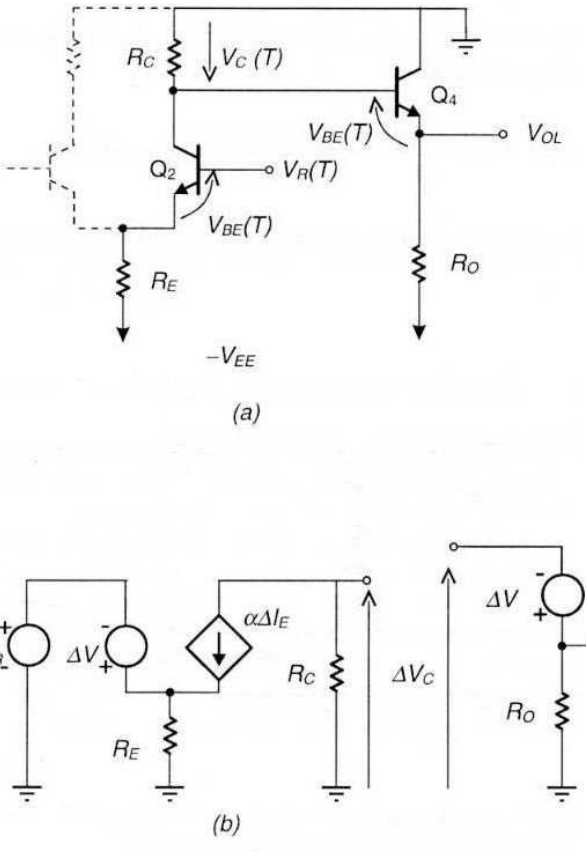


Figura 9.13 a) Analisi degli effetti della temperatura su V_{OL}; b) circuito equivalente per le variazioni termiche

La variazione di tensione ΔV_R e quella ΔV della giunzione di base di Q_2 bilanciano la caduta $\Delta V_E = R_E \Delta I_E$ sul ramo di emettitore. La variazione della corrente di collettore in base al modello equivalente del transistore è legata a quella di emettitore dalla relazione: $\Delta I_C = \alpha\, \Delta I_E$ e quindi la variazione di tensione su R_C sarà:

$$\Delta V_C = -\alpha \Delta I_E R_C \cong -\alpha(\Delta V_R + \Delta V)\frac{R_C}{R_E} \tag{9.27}$$

La variazione di tensione ΔV_{OL} si ottiene aggiungendo la variazione di tensione sulla base di Q_4 a quella ΔV_E data dalla (9.27); assumendo un valore di R_C/R_E = 0.31, valore tipico delle porte ECL, si ha:

$$\Delta V_{OL} = \Delta V - \alpha(\Delta V_R + \Delta V)\frac{R_C}{R_E} \cong \Delta V(1 - (0.7 + 1) \cdot 0.31) \cong 0.47\Delta V \qquad (9.28)$$

Dalle (9.25), (9.26) e (9.28) si ricava che la variazione di V_R è ancora la metà della variazione totale $\Delta V_{OH} + \Delta V_{OL}$:

$$\Delta V_R \cong 0.74\Delta V; \quad \frac{\Delta V_{OH} + \Delta V_{OL}}{2} \cong \frac{\Delta V + 0.47\Delta V}{2} = 0.73\Delta V \qquad (9.29)$$

e quindi la simmetria dei margini di rumore non viene degradata dalle variazioni di temperatura.

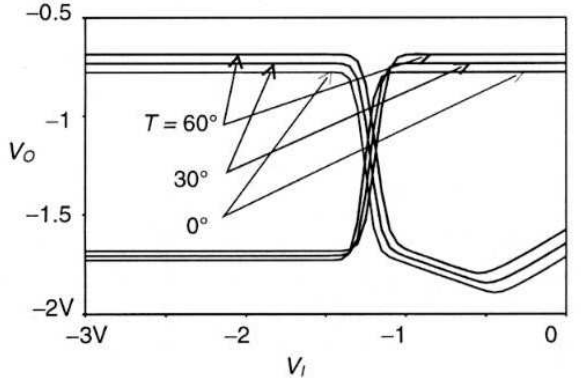


Figura 9.14 Simulazione SPICE delle caratteristiche di trasferimento OR e NOR di una porta ECL al variare della temperatura di funzionamento

L'analisi effettuata è confermata dalle simulazioni SPICE di un circuito ECL effettuate a diverse temperature di funzionamento, riportate in Figura 9.14. In particolare si può notare (Figura 9.15) che le variazioni inevitabili delle grandezze V_{OH} e V_{OL} al variare della temperatura sono seguite da quelle di V_R in modo da ottenere che la tensione di riferimento sia sempre il valore medio tra V_{OH} e V_{OL}. Poiché i valori di V_{IH} e V_{IL} sono legati al valore di V_R, la condizione imposta di mantenere quest'ultima sempre al valore medio tra V_{OH} e V_{OL} comporta il conservare i margini di rumore uguali ed approssimativamente costanti al variare della temperatura in un ampio campo di valori.

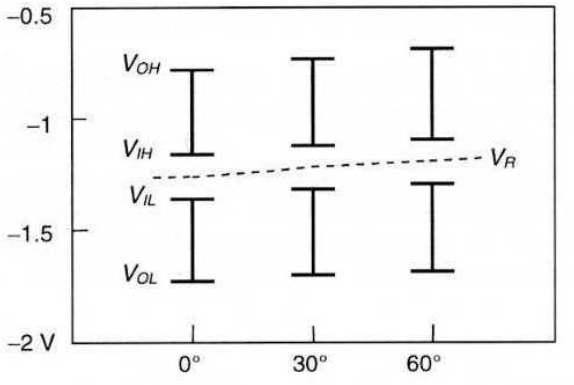


Figura 9.15 Variazioni delle grandezze caratteristiche della porta di Figura 9.14 al variare della temperatura

9.4.2 Analisi della variazione della tensione di alimentazione

Si è accennato, nella introduzione all'invertitore ECL del Paragrafo 9.2, alla convenienza di utilizzare una alimentazione negativa $-V_{EE}$ per ridurre i disturbi introdotti da variazioni della tensione di alimentazione sui livelli logici. Queste sono dovute alla presenza di una inevitabile resistenza interna dell'alimentatore, che comporta una variazione della tensione fornita ai circuiti se varia l'erogazione di corrente dell'alimentatore. Una tra le principali cause di disturbo è quindi la variazione di corrente durante le commutazioni delle altre porte, che inducono degli inevitabili disturbi ΔV sulla tensione nominale di alimentazione. Nel caso dello schema differenziale elementare di Figura 9.3, le tensioni di uscita nei due stati in presenza di un disturbo ΔV sulla tensione di alimentazione sono:

$$V_{OH} = V_{CC} + \Delta\text{V}; \quad V_{OL} = V_{CC} + \Delta V - R_C I$$

e quindi risentono direttamente dei disturbi indotti su questa tensione, che possono essere inaccettabili per questo tipo di logica che presenta swing logici molto ridotti rispetto alle altre logiche e quindi bassi margini di rumore.

Nell'invertitore ECL il problema viene sensibilmente ridotto sostituendo a V_{CC} la tensione di massa come riferimento, poiché, in presenza di disturbi nella tensione di alimentazione $-V_{EE}$, i livelli logici saranno dati in questo caso da:

$$V_{OH} = V_{BE}; \quad V_{OL} = V_{BE} - R_C I_E(\Delta V)$$

dove con $I_E(\Delta V)$ si è indicata la dipendenza della corrente I_E dalla variazione della tensione $-V_{EE}$, dipendenza più debole di quella unitaria presentata dal circuito di Figura 9.3.

Vediamo più in dettaglio gli effetti delle variazioni della tensione di alimentazione, chiamando in causa anche il circuito regolatore di tensione V_R. La rete di polarizzazione del circuito regolatore di tensione infatti contribuisce ad attenuare gli effetti delle possibili variazioni della tensione di alimentazione nei riguardi dei margini di rumore della porta. Ricordiamo che, dalle (9.20a), la tensione V_{OH} non dipende dalla tensione di alimentazione $-V_{EE}$, mentre sono funzioni della tensione di alimentazione sia la tensione V_{OL}, che per l'uscita OR è data dalla (9.20a):

$$V_{OL} = -I_{C2}R_{C2} - V_{BE4} = -(V_R - V_{BE2} + V_{EE})\frac{R_{C2}}{R_E} - V_{BE4} \qquad (9.30a)$$

che la V_R, che è data dalla (9.24):

$$V_R = V_B - V_{BE} = \frac{2V_D - V_{EE}}{R_1 + R_2}R_1 - V_{BE} \qquad (9.30b)$$

Da queste espressioni si vede che l'espressione della V_{OL} data dalla (9.30a) dipende anche dalla V_R. Effettuando la derivata di quest'ultima rispetto a V_{EE} si ottiene, in base alla (9.24):

$$\frac{dV_R}{dV_{EE}} = -\frac{R_1}{R_1 + R_2} = -0.13$$

La derivata di V_{OL} rispetto alla tensione V_{EE}, ricordando il valore del rapporto dV_R/dV_{EE} determinato precedentemente, vale quindi:

$$\frac{dV_{OL}}{dV_{EE}} = -\frac{R_{C2}}{R_{EE}}\left(1 + \frac{dV_R}{dV_{EE}}\right) = -0.26 \qquad (9.31)$$

Dalla (9.31) si vede che le variazioni del valore logico basso V_{OL} risultano significativamente ridotte rispetto a quelle della tensione di alimentazione V_{EE}. Inoltre la tensione V_R varia all'incirca della metà di quanto varia V_{OL}, garantendo quindi ancora una volta margini di rumori simmetrici per i due stadi al variare della tensione di alimentazione V_{EE}.

9.5 Lo stadio di uscita

Gli stadi di uscita per le due uscite complementari sono realizzati, come si è visto in Figura 9.4, con un transistore montato a collettore comune. Le resistenze sull'e-

mettitore sono riportate con linee tratteggiate per indicare che queste sono generalmente *esterne* alla porta, in quanto la porta stessa termina con gli emettitori (aperti) dei transistori di uscita; ciò permette sia di realizzare più semplicemente le connessioni tra le porte e le funzioni logiche cablate (come verrà discusso nel seguito), sia di lasciare libera la scelta della resistenza di uscita per un efficace adattamento alle linee di interconnessione, come si vedrà nel Paragrafo 9.8.

La presenza di questo stadio, come si è già detto, trasla in basso la tensione di uscita corrispondente e ciò permette di avere uscite compatibili con gli ingressi logici richiesti e margini di rumori accettabili. Questo però non è l'unico vantaggio, in quanto la presenza dello stadio di uscita migliora sia la caratteristica di uscita che le prestazioni dinamiche della porta e permette, come vedremo, di adattare l'uscita alle linee di trasmissione necessarie per il collegamento tra porte ECL distanti.

Caratteristica di uscita

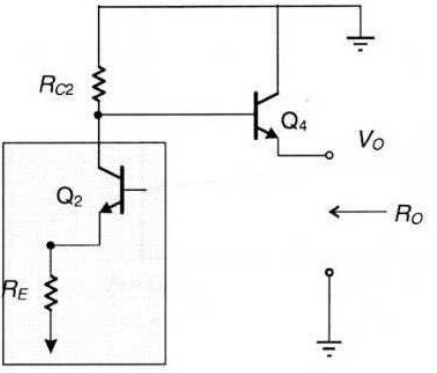


Figura 9.16 Analisi dello stadio di uscita

L'utilizzo di una configurazione a collettore comune per lo stadio di uscita riduce la resistenza di uscita dello stadio stesso; questa caratteristica viene ampiamente discussa nelle analisi degli amplificatori elementari riportate nei testi di elettronica analogica, ma può essere giustificata in maniera semplificata senza chiamare in causa i circuiti equivalenti a piccoli segnali, in maniera analoga a quanto visto per la caratteristica di uscita delle porte TTL.

Con riferimento al circuito di Figura 9.16 per lo stadio di uscita (riferito all'uscita OR per esemplificare), la tensione di uscita nello stato alto V_{OH} in funzione della corrente di uscita I_L è data, ricordando che il transistore Q_4 lavora in regime attivo, dalla relazione:

$$V_{OH}(I_L) = -R_{C2} I_{B4} - V_{BE4} = -\frac{R_{C2}}{\beta_F + 1} I_{E4} - V_{BE4} \equiv V_{OH}{}^* - R_O I_L \qquad (9.32)$$

dove V_{OH} * è la tensione di uscita per corrente di uscita nulla, e R_O è per definizione la resistenza interna dello stadio di uscita, considerato come generatore di segnale. Ad esempio questa, assumendo per il transistore $\beta_F = 50$ e ricordando il valore tipico di R_{C2}, vale:

$$R_O = \frac{R_{C2}}{\beta_F + 1} = \frac{245}{51} \cong 4.8\ \Omega \tag{9.33}$$

ed è quindi relativamente bassa, il che comporta la possibilità di erogare correnti relativamente elevate con piccole riduzioni della tensione in uscita.

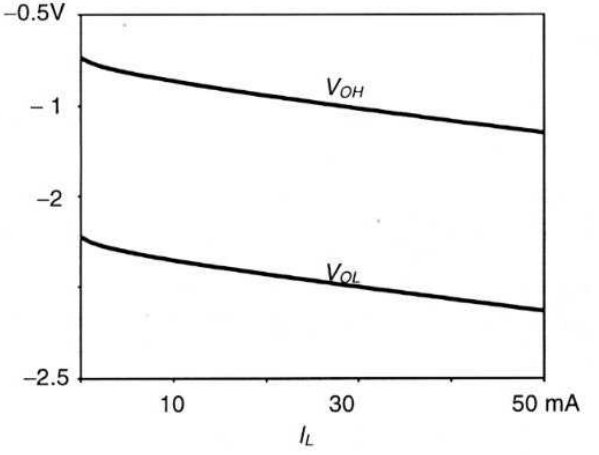


Figura 9.17 Simulazioni SPICE delle caratteristiche di uscita di una porta ECL con i valori: $R_{C2} = 245\ \Omega$, $\beta_F = 50$, $r_B = 50\ \Omega$

Ciò vale anche per l'uscita nello stato basso V_{OL}, in quanto in questo caso la rete connessa al terminale di base del transistore Q_4 di uscita può essere sempre sostituita da un generatore di tensione equivalente, secondo Thevenin, pari alla tensione V_{C2MIN}, con in serie una resistenza pari alla resistenza R_{C2} nella quale scorre la corrente I_{B4} assorbita dal transistore Q_4 (o Q_3); per questo secondo caso la (9.32) si modifica in:

$$V_{OL}(I_L) = V_{C2MIN} - \frac{R_{C2}}{\beta_F + 1} I_{E4} - V_{BE4} \equiv V_{OL}{}^* - R_O I_L \tag{9.34}$$

Le caratteristiche di uscita nei due casi sono riportate in Figura 9.17. La pendenza delle due curve è uguale sia per V_{OH} che per V_{OL} ed è molto prossima al valore definito dalle (9.32) o (9.34); nella valutazione della resistenza R_O bisogna tenere

in conto ovviamente nella maglia di base anche la resistenza di base del transistore, e la debole dipendenza della V_{BE} da I_B che non è tenuta in conto nelle relazioni approssimate precedenti.

9.6 Fan-out

Il fan-out statico di una porta può essere valutato in base alla conoscenza delle caratteristiche di uscita e di ingresso della porta stessa, che permettono di determinare rispettivamente a) quanta corrente può essere erogata dalla porta a monte con determinata degradazione della tensione dello stato logico corrispondente e b) quanta corrente è assorbita dalla porta a valle in ognuno degli stati logici.

La caratteristica di uscita della porta ECL è stata definita in base all'analisi dello stadio di uscita; quella di ingresso è dipendente dal modo di funzionamento del transistore Q_1 dello stadio differenziale.

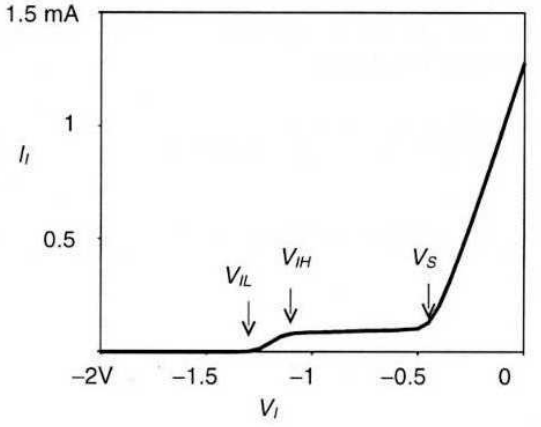


Figura 9.18 Caratteristica di ingresso della porta ECL con i valori precedentemente indicati per le resistenze del circuito e con $\beta_F(Q_1) = 50$

In particolare, ricordando che per tensioni di ingresso $V_I = V_{OL} < V_R - 0.1$ V il transistore Q_1 è interdetto, si deduce che quest'ultimo assorbirà corrente a partire da tensioni di ingresso superiori a $V_R - 0.1$ V $\equiv V_{IL}$. A partire da questo valore di V_I la corrente di ingresso, che è anche la corrente di base di Q_1, cresce con la corrente di collettore I_{C1} poiché Q_1 opera in regione attiva, finché V_I raggiunge il valore $V_R + 0.1$ V $\equiv V_{IH}$. Oltre tale valore Q_2 si interdice e nel transistore Q_1 circola tutta la corrente I_E data dalla (9.9), che risulta, per quanto visto precedentemente, poco dipendente dalla tensione di ingresso V_I. La corrente di ingresso in questa regione è quindi praticamente costante e legata alla corrente I_E secondo la relazione:

$$I_I = \frac{I_E}{\beta_F + 1} \tag{9.35}$$

Se la tensione di ingresso V_I supera il valore V_S che porta in saturazione Q_1, la corrente di ingresso cresce rispetto al valore dato dalla (9.35), con dipendenza lineare da V_I; tuttavia il valore massimo della tensione di ingresso in condizioni normali, cioè V_{OH}, è inferiore a V_S. La simulazione SPICE della caratteristica di ingresso riportata in Figura 9.18 permette di verificare l'analisi esposta.

In base alla (9.7) o (9.8) per i valori delle tensioni e delle resistenze definiti precedentemente si ottiene una corrente I_E di circa 4 mA, per cui la corrente di ingresso I_{IH} assorbita nello stato logico alto, per un valore $\beta_F = 50$, è di circa 80 μA.

Il fan-out statico può quindi calcolarsi in base alle valutazioni della corrente I_{IH} ed alla massima degradazione ammissibile della tensione di uscita con la corrente di carico I_L, definita dalla caratteristica di uscita. Facendo riferimento per quest'ultima all'uscita alta V_{OH} (perché con uscita bassa non vi è apprezzabile assorbimento di corrente dalle porte connesse in uscita), ed assumendo una degradazione ammissibile ΔV_{OH} del 10% sul valore nominale $V_{OH} = -0.7$ V si ottiene dalla (9.32) un valore massimo ammissibile di I_L pari a:

$$I_{LMAX} = \frac{\Delta V_{OH}(\beta_F + 1)}{R_C} \cong \frac{0.07 \cdot 51}{250} \cong 14 \text{ mA} \tag{9.36}$$

In base a questo valore di I_{LMAX} si determina il fan-out $N = I_{LMAX}/I_{IH}$ che, per l'esempio in questione vale:

$$N = \frac{I_{LMAX}}{I_{IH}} \cong \frac{14}{0.08} \cong 175$$

Questo valore così elevato giustifica la considerazione che il fan-out delle porte ECL è essenzialmente determinato da considerazioni sulle prestazioni dinamiche della porta, in particolare sui tempi di propagazione, analogamente al caso delle porte CMOS, e non da considerazioni statiche, come nel caso delle porte TTL.

9.7 Comportamento dinamico e tempi di propagazione

Le prestazioni dinamiche delle porte ECL sono favorite dalla condizione di funzionamento dei transistori in regione attiva e dall'assenza di fenomeni di accumulo di cariche nella base dovute alla saturazione. In particolare la dinamica dello stadio differenziale è molto rapida, essendo legata alle costanti di tempo dei collettori dei due transistori Q_1 e Q_2, dipendenti dalle resistenze di carico R_{C1} o R_{C2}, di valore

relativamente basso, e dalle capacità di ingresso dei due stadi di uscita, inferiori al centinaio di fF.

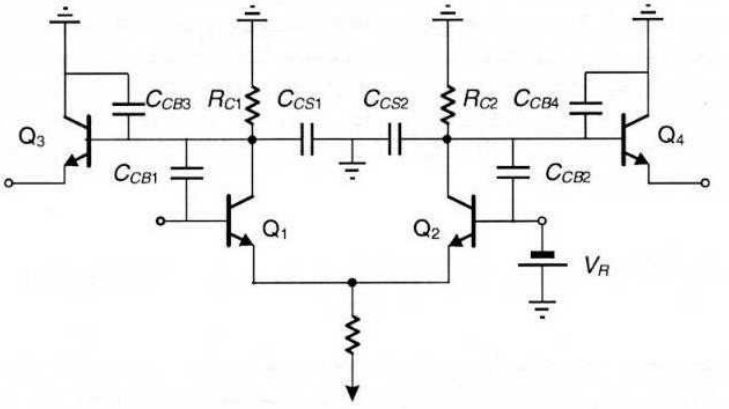


Figura 9.19 Capacità delle giunzioni coinvolte nelle transizioni dei nodi interni di un invertitore ECL

In Figura 9.19 sono indicate le capacità dei transistori del circuito differenziale e degli stadi di uscita coinvolte nella dinamica delle transizioni da un valore logico all'altro; il circuito regolatore non è riportato in quanto questo non vede tensioni variabili e quindi non è coinvolto nella dinamica delle commutazioni, e nella Figura 9.19 è stato sostituito in via approssimata da un generatore ideale di tensione V_R. Le capacità evidenziate sono quelle C_{CB} tra base e collettore dei transistori e quelle C_{CS} tra collettore e substrato, queste ultime di solito di valore più elevato (vedi Tabella 6.3); queste capacità sono connesse tra il nodo di collettore di Q_1 (o Q_2) ed una tensione di riferimento (la massa per $C_{CB3,4}$ e $C_{CS1,2}$, la tensione V_R per C_{CB2}), mentre la capacità C_{CB1} è connessa al terminale di ingresso che è connesso alla tensione variabile V_I. Trascurando in prima approssimazione l'effetto Miller su quest'ultima capacità, si può considerare l'effetto di queste capacità sulla dinamica di variazione della tensione del nodo di collettore, sostituendo a queste un'unica capacità equivalente C_T pari alla somma delle capacità, connessa tra collettore e massa. Con riferimento al terminale di collettore di Q_2 si ha quindi per la capacità equivalente C_T:

$$C_T = C_{BC2} + C_{CS2} + C_{CB4} \tag{9.37}$$

e il transitorio della tensione V_{C2} quando l'ingresso passa dal valore basso a quello alto, a cui corrisponde una tensione V_{C2} che passa dal valore $-I_{C2}R_{C2}$ a 0, è descritto dalla relazione legata alla costante di tempo totale del nodo di collettore:

$$V_{C2} = -I_{C2} R_{C2} \exp\left(-\frac{t}{R_{C2} C_T}\right) \tag{9.38}$$

L'evoluzione di V_{C2} assume quindi un andamento esponenziale se si assume in via approssimata che la capacità C_T non dipenda dalla tensione del nodo di collettore. Approssimando il segnale di ingresso con tempi di salita e di discesa nulli (come è usuale nelle espressioni analitiche dei tempi di propagazione), il tempo di propagazione t_{PLH} sarà quindi dato dal tempo in cui la tensione V_{C1} raggiunge il valore $(-I_{C2} R_{C2})/2$, e cioè:

$$\frac{-I_{C2} R_{C2}}{2} = -I_{C2} R_{C2} \exp\left(-\frac{t_{PLH}}{R_{C2} C_T}\right) \Rightarrow t_{PLH} = 0.69(R_{C2} C_T) \tag{9.39}$$

Analogamente, quando il segnale di ingresso V_I passa dal valore alto a quello basso, la tensione V_{C2} passa dal valore 0 a $-I_{C2} R_{C2}$ ed il transitorio della tensione è descritto dalla relazione:

$$V_{C2} = -I_{C2} R_{C2}\left(1 - \exp\left(-\frac{t}{R_{C2} C_T}\right)\right) \tag{9.40}$$

Anche in questo caso il tempo di propagazione t_{PHL}, definito come il tempo in cui V_{C2} passa per il valore $(-I_{C2} R_{C2})/2$, vale:

$$\frac{-I_{C2} R_{C2}}{2} = -I_{C2} R_{C2}\left(1 - \exp\left(-\frac{t_{PHL}}{R_{C2} C_T}\right)\right) \Rightarrow t_{PHL} = 0.69(R_{C2} C_T) \tag{9.41}$$

Il tempo di propagazione complessivo t_P sarà quindi dato da:

$$t_P = \frac{t_{PLH} + t_{PHL}}{2} = 0.69(R_{C2} C_T) \tag{9.42}$$

Per gli stadi di uscita, nell'ipotesi di funzionamento in regione attiva, il comportamento dinamico è anch'esso molto buono, se si suppone il carico costituito da una linea di trasmissione che presenta essenzialmente un carico ohmico pari a R_O. Anche nel caso di connessione diretta dell'uscita ad un'ulteriore porta ECL, la capacità di uscita è corrispondente alla capacità di ingresso della porta ECL e cioè alla C_{BE1}, che è relativamente bassa; ricordiamo inoltre che dall'analisi in regime lineare dello stadio a collettore comune si ricava una banda passante più ampia che per lo stadio ad emettitore comune, a causa dell'assenza dell'effetto Miller sulla capacità di ingresso.

Tuttavia il comportamento dinamico si modifica se ci si riferisce ad un comportamento ad ampi segnali e si considera in uscita una capacità di carico non trascurabile, come accade nei casi pratici quando si debbono connettere le porte a circuiti esterni al chip di silicio. In questo caso, se le capacità di uscita sono dell'ordine di $1 \div 10$ pF e quindi notevolmente superiori a quelle che agiscono sui collettori dello stadio differenziale, e con resistenze di emettitore dell'ordine di $0.5 \div 2$ kΩ, le costanti di tempo $R_O C_L$ associate alla maglia di uscita sono superiori a quelle interne $R_{C1,2} C_{C1,2}$, e quindi le variazioni di tensione del terminale di emettitore sono più lente di quelle del terminale di base.

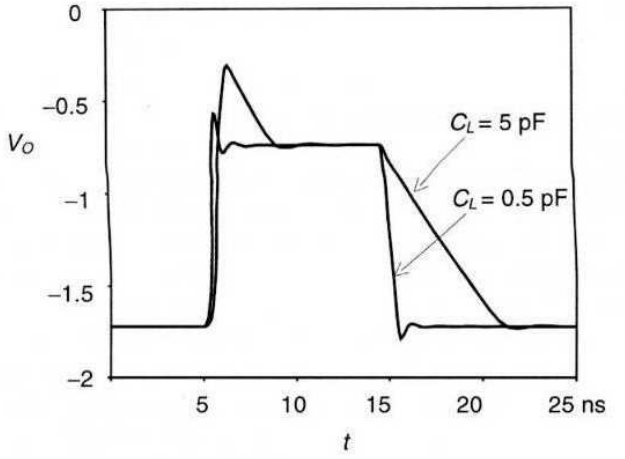


Figura 9.20 Transitorio di commutazione dello stadio di uscita con capacità di carico C_L = 0.5 pF o 5 pF

Nel passaggio dallo stato basso a quello alto (vedi Figura 9.20) il transitorio di carica di C_L è tuttavia aiutato dal funzionamento del transistore che presenta all'uscita una resistenza di uscita molto bassa, data dalla (9.33), per cui la costante di tempo dell'uscita in fase di carica di C_L non è molto più lenta di quella del nodo interno di base; ciò non accade invece nel passaggio dallo stato alto a quello basso. In questo caso infatti la più piccola costante di tempo associata alla base fa sì che la tensione di base scenda più rapidamente di quella di emettitore e quindi il transistore si interdice durante la discesa di V_O. La resistenza di uscita in questo caso sarà la resistenza fisicamente connessa all'emettitore e la capacità C_L si potrà scaricare solo attraverso quest'ultima con un tempo di transizione $t_{THL} > t_{TLH}$.

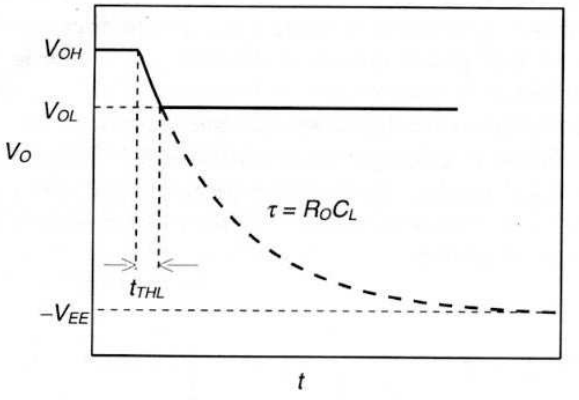


Figura 9.21 Transitorio di uscita nella transizione da V_{OH} a V_{OL}

Questo comportamento era stato già richiamato nel Paragrafo 8.3 per gli stadi di uscita delle porte TTL, dove si era visto che il passaggio da V_{OL} a V_{OH}, per uno stadio con carico sull'emettitore, era intrinsecamente più veloce di quello da V_{OH} a V_{OL}. In questo caso, tuttavia, il fenomeno è ridotto dalla limitata escursione possibile per la tensione di uscita rispetto al valore di tensione a cui tende la capacità nella scarica, pari a $-V_{EE}$. Facendo riferimento al transitorio dell'uscita riportato schematicamente in Figura 9.21, si può notare che nel caso dello stadio di uscita la tensione di pilotaggio della base non raggiunge il riferimento negativo $-V_{EE}$, ma si porta a -1 V, per cui a regime la tensione di uscita si arresta a $V_{OL} = -1.7$ V; in questo caso la capacità inizia a scaricarsi su R_O tendendo a $-V_{EE}$ con la costante di tempo $R_O C_L$, ma la scarica si arresta quando la tensione V_O scende sotto il valore $V_{OL} = -1.7$ V.

Il tempo di transizione t_{THL} è quindi corrispondentemente ridotto dalla limitata escursione della tensione in uscita rispetto al tempo necessario a scaricarla completamente al valore $-V_{EE}$, e vale:

$$t_{THL} = R_O C_L \ln\left(\frac{V_{OH} + V_{EE}}{V_{OL} + V_{EE}}\right) \cong 0.24 \cdot R_O C_L \tag{9.43}$$

Questo giustifica l'asserzione, fatta precedentemente, che la riduzione della escursione di tensione delle porte ECL contribuisce a ridurre i tempi di transizione e quindi più in generale migliora le prestazioni dinamiche della porta. Nell'esempio di Figura 9.20, la porta presenta, con $C_L = 5$ pF, un ritardo di propagazione $t_P = 1.6$ ns, essenzialmente dipendente dal tempo di propagazione t_{PHL}.

L'influenza della capacità di carico C_L sul ritardo di propagazione totale mostra come per queste porte il fan-out sia determinato essenzialmente dal massimo valore della capacità di carico (in questo caso la somma delle capacità di ingresso delle N

porte in uscita), compatibile con un ritardo di propagazione fissato; questa condizione, tenendo conto che non è desiderabile ridurre la velocità di queste porte che è la loro caratteristica più rilevante, porta ad un fan-out N molto minore di quello calcolato in base a considerazioni statiche (circa 10 porte se il ritardo di propagazione deve essere contenuto in qualche ns).

9.8 Adattamento alle linee di trasmissione

La velocità di transizione dei segnali in uscita dalle porte ECL pone dei problemi nella trasmissione di questi segnali da una porta all'altra; ciò in particolare nei casi in cui la connessione non sia limitata a porte realizzate nello stesso chip, ma avvenga attraverso il circuito stampato della piastra su cui sono assemblati i diversi componenti integrati o attraverso cavi coassiali.

È infatti noto che nei conduttori le grandezze elettriche non variano istantaneamente in ogni punto del conduttore, ma impiegano un tempo finito per propagarsi da un estremo all'altro. La velocità di propagazione dei segnali si dimostra dipendere dalla capacità C e dall'induttanza L per unità di lunghezza del conduttore, secondo la relazione:

$$v = \frac{1}{\sqrt{LC}} \tag{9.44}$$

e quindi il segnale impiega un tempo finito τ a propagarsi da un estremo all'altro del sistema di conduttori di lunghezza l dato da:

$$\tau = \frac{l}{v} \tag{9.45}$$

Se ne deduce che, per segnali elettrici variabili in tempi dell'ordine di grandezza di τ, ogni conduttore è da considerarsi una linea di trasmissione per la quale la propagazione del segnale da un estremo all'altro della linea avviene mediante due onde di tensione e di corrente, dipendenti dalla sollecitazione all'estremo di ingresso, che si propagano lungo la linea secondo le note leggi delle linee di trasmissione. Si può dimostrare che la velocità di propagazione v dipende dalla costante dielettrica del mezzo in cui sono immersi i due conduttori; se questo è l'aria la velocità di propagazione è quella della luce, e cioè 3.10^8 m/s. Quindi un sistema di conduttori in aria di lunghezza l = 30 cm avrà tempi di propagazione di 1 ns e dovrà considerarsi una linea di trasmissione se il segnale in ingresso varia in tempi comparabili al tempo di propagazione. Nel caso di conduttori in un mezzo con costante dielettrica maggiore, come ad esempio il circuito stampato di una piastra, la velocità di propagazione è minore e quindi anche piste di collegamento di una decina di cm vanno considerate come linee di trasmissione se collegate a porte ECL che presentano tempi di transizione inferiori al ns. Si richiameranno quindi sinteticamente i problemi che possono

sorgere nella propagazione di segnali rapidamente variabili lungo le linee di trasmissione.

Il rapporto tra l'ampiezza dell'onda di tensione e quella di corrente che si propagano lungo una linea di trasmissione è chiamata *impedenza caratteristica* R_O della linea, data da:

$$R_O = \sqrt{L/C} \tag{9.46}$$

ed è compresa nei casi pratici tra 50 e 200 Ω. La linea si dice *adattata* al generatore o al carico, se questi presentano una resistenza (rispettivamente R_S o R_L) pari a quella caratteristica della linea. Nei casi in cui $R_S \neq R_O$ ($R_L \neq R_O$) la linea è *disadattata* in ingresso (in uscita).

a) Caso di linea disadattata in uscita

Riferiamoci allo schema di Figura 9.22 ipotizzando un generatore di tensione che fornisca un gradino di tensione (quindi variabile in un tempo infinitesimo) all'ingresso di una linea di trasmissione, rappresentata per esemplificare da un cavo coassiale di lunghezza l, ed avente tempo di propagazione τ.

Dopo l'istante $t = 0$ di applicazione del gradino di tensione si propagano lungo la linea un fronte di tensione $V(x) = V = V_S/2$ ed uno di corrente $I(x) = V/R_O$ entrambi con velocità v (Figura 9.22a). Dopo il tempo τ i due fronti d'onda raggiungono l'ascissa l dove vi è il carico R, su cui il rapporto $V(l)/I(l)$ vale $R_L \neq R_O$. Ciò comporta che quando i due fronti d'onda raggiungono l'uscita della linea, si creano due ulteriori onde riflesse di tensione e corrente, che si propagano verso l'ingresso, proporzionali alle onde incidenti secondo un coefficiente di riflessione ρ_L per cui:

$$\Delta V(l)_{t=\tau} = V \cdot \rho_L \; ; \quad \Delta I(l)_{t=\tau} = -\frac{V}{R_O} \cdot \rho_L \tag{9.47}$$

e tali che:

$$\frac{V(l)}{I(l)} \equiv R_L = \frac{V(1+\rho_L)}{V/R_O(1-\rho_L)} = R_O \frac{1+\rho_L}{1-\rho_L} \tag{9.48}$$

Dalla (9.48) si ricava l'espressione del coefficiente di riflessione ρ_L:

$$\rho_L = \frac{R_L/R_O - 1}{R_L/R_O + 1} \qquad [-1 \leq \rho \leq 1] \tag{9.49}$$

Dopo un ulteriore tempo di propagazione τ le due onde riflesse arrivano all'ingresso dove trovano una resistenza $R_S = R_O$; non si creano quindi ulteriori

riflessioni e le tensioni e correnti lungo la linea raggiungono la condizione di regime.

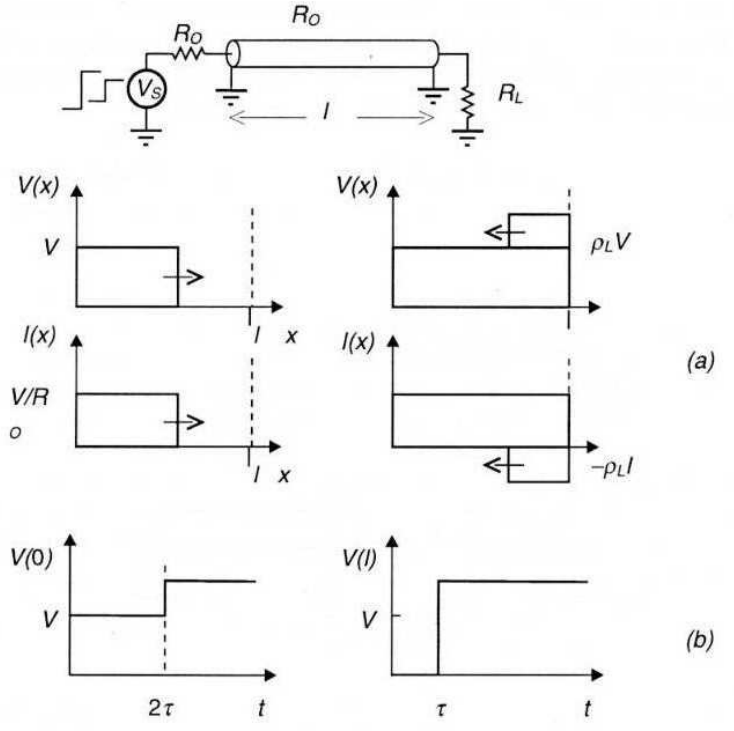


Figura 9.22 Propagazione di un gradino in una linea di trasmissione disadattata all'uscita; a) transitori lungo la linea; b) forme d'onda della tensione *V(0)* all'ingresso e di quella *V(l)* all'uscita della linea di trasmissione

Le forme d'onda della tensione *V(0)* all'ingresso della linea di trasmissione e di quella *V(l)* all'uscita della linea di lunghezza l sono riportate in Figura 9.22b; da queste forme d'onda si nota come il transitorio in ingresso si esaurisce dopo un tempo 2τ mentre quello in uscita dopo un tempo τ, dove τ è il ritardo di propagazione della linea.

b) Linea disadattata in ingresso ed in uscita

Una situazione più complessa si ha quando anche l'ingresso è disadattato. In questo caso si definisce, oltre al coefficiente di riflessione ρ_L in uscita, un coefficiente di riflessione all'ingresso ρ_S dato da:

$$\rho_S = \frac{R_S / R_O - 1}{R_S / R_O + 1} \qquad [-1 \le \rho \le 1] \tag{9.50}$$

In questo caso la situazione è rappresentata schematicamente in Figura 9.23; all'applicazione dell'impulso si ha in ingresso una tensione $V = V_S R_O/(R_S + R_O)$ ed una corrente $I = V/R_O$. Le onde di tensione e di corrente generate dalla riflessione a $x = l$ si propagano verso l'ingresso e lo raggiungono al tempo 2τ; qui si crea un'ulteriore riflessione per le onde $V\rho_L$ e $-I\,\rho_L$ definite dalla (9.41), con la creazione di due ulteriori onde che si propagano di nuovo verso l'uscita:

$$\Delta V(0)_{t>2\tau} = (V \cdot \rho_L) \cdot \rho_S \; ; \quad \Delta I(0)_{t>2\tau} = -(I \cdot \rho_L) \cdot \rho_S \tag{9.51}$$

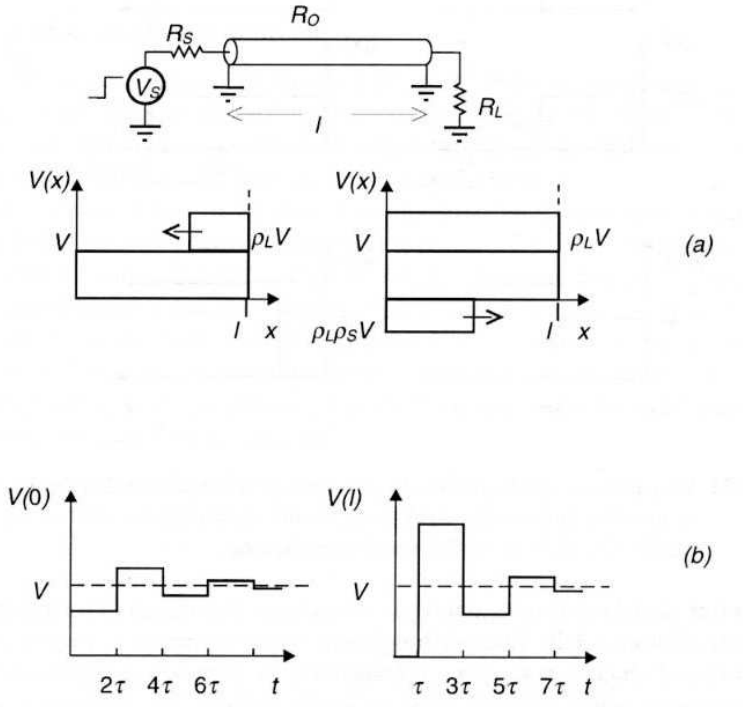


Figura 9.23 Transitori lungo una linea di trasmissione disadattata sia in ingresso che in uscita: a) tensione lungo la linea; b) forme d'onda in ingresso ed in uscita

Il processo di riflessione avviene ancora sia all'uscita che all'ingresso, con onde riflesse via via più piccole se i coefficienti di riflessione sono in modulo minori di 1, come è indicato nella figura. Il caso di interesse per l'accoppiamento di porte ECL è quello relativo ad un disadattamento all'ingresso della linea con $R_S < R_O$ e quindi con un coefficiente di riflessione $\rho_S < 0$ (ricordiamo che in questo caso la resistenza R_S è la resistenza di uscita della porta, dell'ordine di una decina di Ω,

molto minore della resistenza caratteristica della linea); in questo caso la situazione delle tensioni in ingresso e in uscita della linea è riportata in Figura 9.23. La tensione in ingresso ai diversi tempi vale:

$$V_i(0) = V;\ V_i(2\tau) = V\left[1 + \rho_L + \rho_S\rho_L\right];\ V_i(4\tau) = V\left[1 + \rho_L + \rho_S\rho_L + \rho_S\rho_L^2 + \rho_S^2\rho_L^2\right]$$

e quella in uscita vale:

$$V_L(\tau) = V\left[1 + \rho_L\right];\quad V_L(3\tau) = V\left[1 + \rho_L + \rho_S\rho_L + \rho_S\rho_L^2\right]$$

La tensione in uscita presenta quindi, nel caso di coefficiente di riflessione ρ_S negativo e vicino all'unità, una serie di impulsi progressivamente più attenuati fino a raggiungere il valore di regime dato da $V_L = V_S\, R_L/(R_S + R_L)$. Questi possono essere interpretati dalle porte a valle come livelli logici alti e bassi consecutivi, e dare luogo a commutazioni non volute. La scelta ottimale per evitare questo tipo di risposta, indicato come "*ringing*", è quella di adattare la linea sia in ingresso che in uscita. In uscita l'adattamento richiede una resistenza aggiuntiva $R_L R_O/(R_L + R_O)$ in parallelo all'ingresso dello stadio successivo, di valore circa pari a R_O in quanto la resistenza di ingresso dello stadio a valle è molto elevata. In ingresso, utilizzando la configurazione delle porte ECL che prevede l'uscita su emettitore aperto, occorre inserire una resistenza aggiuntiva pari a $R_S - R_O$ in serie tra l'emettitore dello stadio di uscita della porta e la linea. In pratica l'adattamento in ingresso non è conveniente, perché l'inserzione di una resistenza R_S circa pari a R_O attenuerebbe troppo la tensione applicata alla linea rispetto a quella di uscita dallo stadio ($V_i = V_O/2$), per cui si preferisce adattare solo l'uscita in modo da evitare su questa una riflessione che possa propagarsi di nuovo verso l'ingresso.

9.9 Potenza dissipata e prodotto potenza-ritardo

La potenza dissipata nelle porte ECL è relativamente alta, e ciò in accordo con la elevata velocità che queste porte presentano.

Il ramo differenziale della porta ECL di Figura 9.4 assorbe una corrente relativamente costante al variare del livello logico applicato in ingresso, come si è visto, per cui la potenza dissipata nei due stati logici non varia apprezzabilmente, e vale:

$$P_{D1} = V_{EE} \cdot I_E \cong V_{EE}\,\frac{V_R - V_{BE} + V_{EE}}{R_E} \tag{9.52a}$$

Per un valore di $R_E = 800\ \Omega$, la potenza dissipata in questo ramo vale:

$$P_{D1} = 5.2 \cdot 4.1 = 21\ \text{mW}$$

Anche lo stadio regolatore di tensione assorbe una corrente costante; la potenza dissipata vale:

$$P_{D2} = V_{EE}\left[\frac{V_{EE} + 2V_D}{R_1 + R_2} + \frac{V_{EE} + V_R}{R_3}\right] \tag{9.52b}$$

e, per un valore della resistenza $R_3 = 6$ kΩ (questo valore viene scelto in modo da avere nel ramo di emettitore del transistore Q_R una corrente $I = (V_R + V_{EE})/R_3$ pari a quella circolante nel partitore $I_R = (2V_D + V_{EE})/(R_1+R_2)$ e pari a circa 0.66 mA), si ha per la componente P_{D2}:

$$P_{D2} = 5.2(0.64 + 0.64) = 6.6 \text{ mW}$$

La potenza assorbita da ognuno degli stadi di uscita dipende invece dallo stato logico in uscita (V_{OH} o V_{OL}) e dalla resistenza di carico sull'uscita, e vale:

$$P_{D3(OH)} = V_{EE}\frac{V_{EE} - V_{OH}}{R_O}; \quad P_{D3(OL)} = V_{EE}\frac{V_{EE} - V_{OL}}{R_O} \tag{9.52c}$$

Nel caso di porte connesse direttamente in uscita senza adattamento, la resistenza R_O è relativamente elevata (5 kΩ) e la potenza dissipata dal singolo stadio di uscita vale:

$$P_{D3} = \frac{P_{D3(OH)} + P_{D3(OL)}}{2} \cong \frac{4.7 + 3.6}{2} \cong 4.17 \text{ mW}$$

La potenza dissipata nell'intera porta, tenendo conto dei contributi dei singoli stadi dati dalle (9.52a,b,c), vale quindi:

$$P_D = P_{D1} + P_{D2} + 2P_{D3} \cong 21 + 6.6 + 8.34 \cong 36 \text{ mW}$$

Se invece l'uscita è adattata con una resistenza R_O relativamente bassa, la potenza dissipata aumenta; in questo caso però la potenza dissipata nella resistenza di carico non va riferita alla singola porta in esame, in quanto il carico adattato è posto all'uscita della linea di accoppiamento ed usualmente va riferito a più porte di carico.

La considerazione che le uniche correnti variabili significativamente nel funzionamento della porta sono quelle degli stadi di uscita, suggerisce nelle porte ECL la scelta di soluzioni circuitali in cui sono utilizzate due tensioni di alimentazione e due riferimenti (masse) separati per la sezione differenziale + regolatore e per quella di uscita. In tal modo gli inevitabili disturbi indotti sull'alimentazione (dello stadio di uscita) dalle brusche variazioni di correnti relativamente elevate, non si

propagano nella sezione differenziale e nel regolatore di tensione che sono più sensibili ad eventuali disturbi. Con questa soluzione è anche possibile alimentare con tensioni di valore ridotto gli stadi di uscite in modo da ridurre la dissipazione di potenza in questi ultimi nel caso di resistenze di uscita relativamente basse, necessarie per adattare le uscite alle linee di trasmissione.

In ogni caso il prodotto potenza-ritardo è relativamente elevato rispetto alle altre porte, e dell'ordine di alcune decine di pJ; l'interesse di queste porte d'altra parte non è legato al basso valore di questo prodotto, ma all'elevata velocità di funzionamento ed al ridotto ritardo di propagazione che le pongono come le porte più veloci tra le famiglie logiche commerciali.

9.10 Porte logiche ECL

Partendo dall'invertitore elementare riportato in Figura 9.4 è immediato realizzare delle porte NOR ponendo in parallelo più transistori nel ramo del differenziale a cui si applica il segnale di ingresso, come è indicato in Figura 9.24 per il caso di una porta a due ingressi.

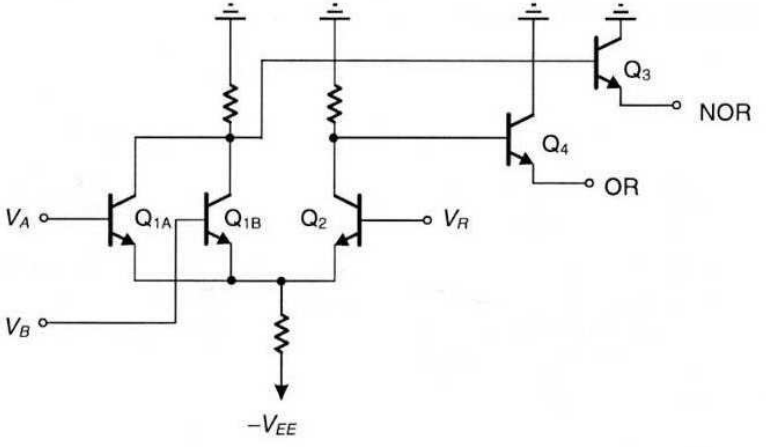


Figura 9.24 Schema di porta logica NOR-OR a due ingressi

In questo caso i transistori Q_{1A} e Q_{1B} connessi in parallelo con un unico carico realizzano la funzione NOR sull'uscita comune dei collettori; sul collettore del transistore Q_2 si ottiene la funzione OR (uscita complementare). Le due funzioni sono replicate dagli stadi di uscita Q_3 e Q_4 che non effettuano nessuna funzione aggiuntiva sul segnale logico in ingresso; si comprende quindi perché le due uscite sono state indicate rispettivamente come uscita NOR ed uscita OR.

Da quanto visto si comprende come in questa famiglia logica la struttura preferenziale sia quella basata su porte NOR (e OR), in quanto la funzione logica NAND è più difficile da implementare. Vedremo in seguito (Capitolo 10) che le due uscite con emettitore aperto permettono di realizzare funzioni più complesse che derivano dalla connessione in uscita di più porte con unica resistenza di carico; in questo caso si può trarre vantaggio dalla simultanea presenza delle funzioni complementari OR e NOR alle due uscite.

9.11 Porte ECL avanzate

Una versione migliorata delle porte ECL esaminate precedentemente è quella (indicata come serie 100K) che impiega generatori di corrente realizzati con circuiti attivi, al posto delle resistenze di polarizzazione, sia per la configurazione differenziale che per il generatore della tensione di riferimento e per gli stadi di uscita. Questa configurazione è sinteticamente riportata in Figura 9.25; i generatori di corrente vengono realizzati con versioni modificate della configurazione detta "specchio di corrente" (*current mirror*) utilizzata anche nei circuiti elettronici analogici per la polarizzazione dei dispositivi.

La polarizzazione con generatori di corrente comporta diversi vantaggi nel funzionamento delle porte ECL, esaminati nel seguito.

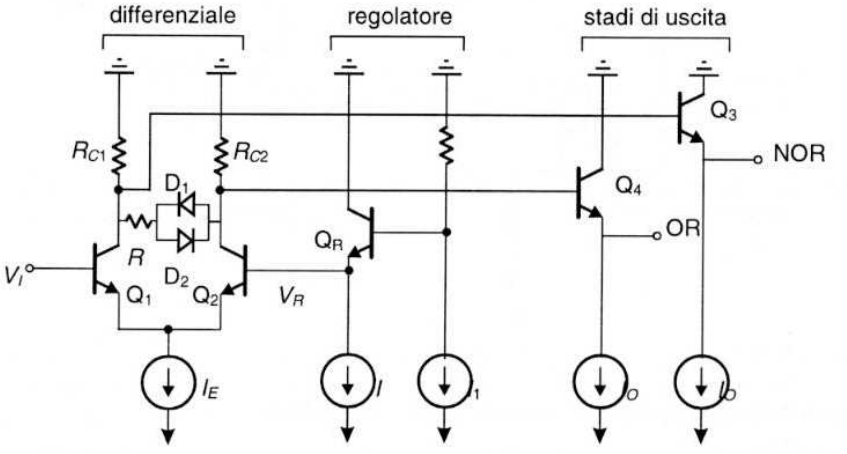


Figura 9.25 Schema della porta ECL con generatori di corrente (serie 100 K)

9.11.1 Caratteristica di trasferimento e margini di rumore

La corrente I_E è in questo caso imposta dal generatore di corrente ed è quindi costante al variare dell'ingresso V_I; ciò comporta che la corrente che circola in ognuno dei due transistori quando l'altro è interdetto (ossia per $V_I < V_R - 0.1$ V o $V_I > V_R + 0.1$ V) è uguale e costante, quindi le due caratteristiche di trasferimento OR e NOR sono simmetriche almeno fino al valore $V_I = V_S$ oltre il quale Q_1 va in saturazione, come è indicato in Figura 9.26.

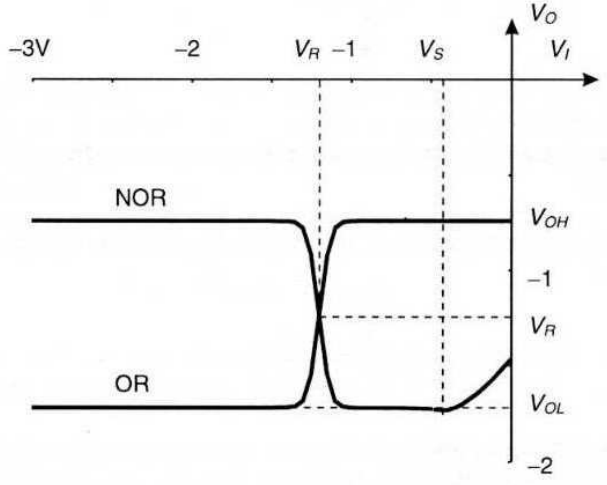


Figura 9.26 Caratteristiche di trasferimento della porta con generatori di corrente

I margini di rumore risultano migliorati in quanto la polarizzazione con generatori di corrente, e la presenza della rete di compensazione con i diodi D_1 e D_2 nella configurazione differenziale, permettono una stabilizzazione molto efficace delle grandezze caratteristiche nei riguardi sia della variazione di tensione che di quella in temperatura.

Nel primo caso la stabilizzazione è immediatamente comprensibile se si considera che le correnti di polarizzazione, nel caso di generatori di corrente ideali, non si modificano al variare di V_{EE}; quindi le caratteristiche di trasferimento (OR e NOR) non dipendono da quest'ultima ed i margini di rumore risultano inalterati rispetto a queste variazioni.

9.11.2 Comportamento alle variazioni termiche

Il comportamento nei riguardi delle variazioni termiche può essere valutato con riferimento ad una versione semplificata (riportata in Figura 9.27) del circuito specchio di corrente utilizzato per generare le correnti di polarizzazione dei diversi rami del circuito di Figura 9.25.

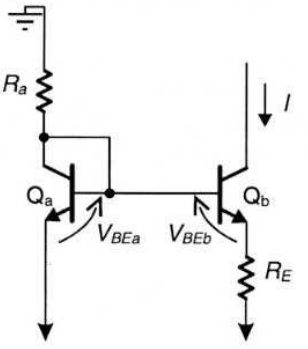


Figura 9.27 Schema semplificato del generatore di corrente realizzato con lo specchio di corrente

Assumendo un $\beta >> 1$ per i transistori, la corrente I può essere valutata in base alla relazione:

$$I = \frac{V_{BEa} - V_{BEb}}{R_E} \tag{9.53}$$

da cui si ha che la differenza tra i valori V_{BE} dei due transistori per una data corrente di polarizzazione I dipende dal valore della resistenza R_E. Ricordiamo che il coefficiente di temperatura di un diodo P/N, in base alla espressione analitica della sua caratteristica I-V, vale:

$$\frac{dV_{BE}}{dT} = \frac{1}{T}\left[V_{BE} - 3V_T - V_G\right] \cong \frac{1}{T}\left[V_{BE} - 1.2V\right] \tag{9.54}$$

dove $V_T = kT/q$ è la tensione termica, pari a circa 26 mV e $V_G = E_G/kT$ è la tensione corrispondente all'energia di banda proibita, pari a 1.12 V per il silicio. Dalla (9.54) si ricava un coefficiente di temperatura di circa −1.6 mV/°C a temperatura ambiente, che aumenta al crescere di V_{BE}.

Quindi dalla (9.53), poiché il valore di V_{BEa} è maggiore di quello V_{BEb} dovendosi sottrarre dal primo la caduta su R_E, si ottiene per I una dipendenza positiva dalla temperatura:

$$\frac{dI}{dT} = \frac{1}{R_E}\left[\frac{dV_{BEa}}{dT} - \frac{dV_{BEb}}{dT}\right] > 0 \tag{9.55}$$

che può essere modificata entro ampi limiti attraverso un'opportuna scelta del valore di R_E.

La tensione V_R generata dallo stadio regolatore di tensione (Figura 9.25) vale:

$$V_R = -R_1 I_1 - V_{BER} \tag{9.56}$$

e quindi si può scegliere la resistenza R_E in modo da annullare in pratica la dipendenza dalla temperatura di V_R, in quanto quest'ultima può essere scritta come:

$$\frac{dV_R}{dT} = -\frac{R_1}{R_E}\left[\frac{dV_{BEa}}{dT} - \frac{dV_{BEb}}{dT}\right] - \frac{dV_{BER}}{dT} \tag{9.57}$$

e quindi, con un'opportuna scelta di R_E, si può rendere il primo termine della (9.57) (che è negativo, ricordando che il termine in parentesi quadra è positivo, dalla (9.55)), uguale in modulo al secondo termine (che è positivo, in quanto la derivata dV_{BE}/dt è negativa).

La stessa considerazione vale per la tensione al livello basso V_{OL} che è data da:

$$V_{OL} = -V_{BE(3,4)} - R_{C1,2} I_E \tag{9.58}$$

e che quindi può essere compensata in temperatura analogamente alla (9.57), utilizzando il coefficiente di temperatura positivo di I_E realizzato con analogo specchio di corrente.

Infine la V_{OH} viene compensata in temperatura utilizzando ancora il coefficiente di temperatura positivo di I_E e la rete di diodi D_1, D_2 con la resistenza R inserita tra i collettori di Q_1 e Q_2. Infatti quando ad esempio Q_2 è interdetto, $V_{C1} \cong -1$ V; il diodo D_1 conduce e circola una (debole) corrente in R_{C2} e R. Al variare della temperatura, assumendo per la variazione ΔV_{C1} su R_{C1} indotta da I_E un valore $\cong \Delta V$, si ha una variazione totale ai capi delle due resistenze:

$$\Delta V_{C1} - \Delta V_{D1} \cong \Delta V - (-\Delta V) \cong 2\Delta V \tag{9.59}$$

dove ΔV è la variazione di temperatura di circa 1.6 mV/°C di una giunzione. Le resistenze R_{C2} e R vengono scelte uguali in modo che la variazione di tensione ΔV_{C2} sia pari a $(\Delta V_{C1} + \Delta V_{C2})/2 = +\Delta V$. La variazione di V_{OH} con la temperatura sarà quindi:

$$\Delta V_{OH} \cong \Delta V_{C2} + \Delta V_{BE4} \cong \Delta V + (-\Delta V) \cong 0 \tag{9.60}$$

Lo stesso discorso vale per l'uscita su Q_3 considerando ora il diodo D_2 e Q_1 interdetto. La porta è quindi compensata in temperatura.

Le porte ECL, come quelle TTL, hanno avuto una larga applicazione, in particolare negli elaboratori di grosse dimensioni dove la velocità di operazione delle unità di processo centrali (CPU) è un parametro fondamentale nelle prestazioni del sistema. Le versioni più recenti delle porte ECL impiegano le strutture bipolari

avanzate presentate nel Paragrafo 6.7 e presentano i valori più bassi dei tempi di ritardo (dell'ordine delle decine di ps) rispetto ad ogni altra porta logica realizzata in silicio.

Esercizi di riepilogo

9.1 Nel circuito differenziale di Figura 9.3, determinare: a) di quanto deve aumentare il valore di V_I rispetto al valore V_R perché la corrente I_{E1} raggiunga il 99% del valore I imposto dal generatore di corrente; b) di quanto deve diminuire il valore di V_I rispetto al valore V_R perché la corrente I_{E1} si riduca all'1% del valore I.

9.2 Per il circuito differenziale di Figura 9.5, assumendo un valore di $R_E = 800\ \Omega$ e $V_R = -1.2$ V, determinare la differenza percentuale tra i valori della corrente I_E nella resistenza RE per i valori della tensione di ingresso $V_I = V_R - 0.1$ V e $V_I = V_R + 0.1$ V.

9.3 Determinare i valori di R_{C1}, R_{C2}, R_E per il circuito differenziale di Figura 9.5 in modo da avere uguali margini di rumore NM_H e NM_L con un valore $V_E = -1.3$ V, ed una dissipazione di potenza statica di 5 mW.

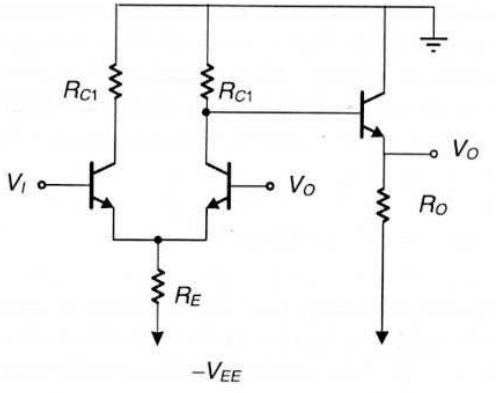


Figura E9.1

9.4 Per l'invertitore ECL di Figura 9.4, assumendo i seguenti valori dei parametri: $R_E = 800\ \Omega$, $R_{C1} = 220\ \Omega$, $R_{C2} = 250\ \Omega$, $R_1 = 900\ \Omega$, $R_2 = 5$ KΩ, $R_3 = 6$ kΩ, determinare per via analitica approssimata a) i livelli logici per l'OR; b) il valore logico basso V_{OL} corrispondente ad un ingresso $V_I = V_{IH}$; c) il valore logico basso V'_{OL} corrispondente ad un ingresso $V_I = V_S$.

9.5 Per l'invertitore ECL dell'Esercizio 9.4, determinare gli effetti sui livelli logici nominali, sulla tensione V_R e sui margini di rumore, di una variazione del ±10% della tensione di alimentazione $-V_{EE} = -5.2$ V.

9.6 Per l'invertitore ECL di Figura E9.1, con i seguenti valori: $R_E = 800\ \Omega$, $R_{C1} = 220\ \Omega$, $R_{C2} = 250\ \Omega$, $\beta_F = 50$, determinare il valore della resistenza di carico R_O compatibile con una degradazione del livello logico alto del 20%.

9.7 Con riferimento ad un invertitore ECL con i seguenti valori: $R_E = 800\ \Omega$, $R_{C1} = 220\ \Omega$, $R_{C2} = 250\ \Omega$, $\beta_F = 50$, assumendo di dover collegare l'uscita ad una linea di trasmissione con impedenza caratteristica $R_O = 200\Omega$, determinare i livelli logici presentati all'uscita della linea se questa viene adattata in uscita chiudendola su una resistenza $R_L = R_O$.

9.8 Per l'invertitore ECL di Figura E9.1, con i valori del circuito: $R_E = 800\ \Omega$, $R_{C1} = 220\ \Omega$, $R_{C2} = 250\ \Omega$, $R_O = 2\ \text{k}\ \Omega$, e con i parametri dei transistori: $\beta_F = 50$, $C_{BE} = C_{BC} = 0.3$ pF, $C_{CS} = 1$ pF, determinare a) il tempo di propagazione; b) la potenza dissipata; c) il prodotto ritardo-potenza.

9.9 Ripetere l'Esercizio 9.8 aumentando tutte le resistenze del circuito di un fattore 5. Come si è modificato il prodotto ritardo-potenza?

Riferimenti bibliografici

H. Taub, D. Schilling, *Elettronica Integrata Digitale*, Jackson, Milano, 1981.

G.M. Glansford, *Digital Electronic Circuits*, Prentice Hall, Englewood Cliffs, 1988.

A.S. Sedra, K.C. Smith, *Microelectronic Circuits*, Saunders College, Philadelphia, 1991.

B. Riccò, F. Fantini, P. Brambilla, *Introduzione ai circuiti integrati digitali*, Zanichelli, Bologna, 1991.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

10

Circuiti combinatori

10.1 Circuiti logici standard

Le porte elementari CMOS, TTL ed ECL studiate nei capitoli precedenti formano la base delle rispettive *famiglie logiche standard*, intendendo con questo termine i circuiti logici elementari basati su una data tecnologia, che utilizzano un numero limitato di porte, e che sono forniti dalle ditte costruttrici in contenitori standard con specifiche ben definite. Le logiche standard utilizzano quindi circuiti integrati con un livello di integrazione relativamente basso (SSI o MSI), in modo da poter essere utilizzate come componenti logici per un'elevata gamma di applicazioni. I circuiti SSI sono quelli che contengono un certo numero di porte elementari non interconnesse, usualmente in contenitori dual-in-line con 14 piedini (7 per lato).

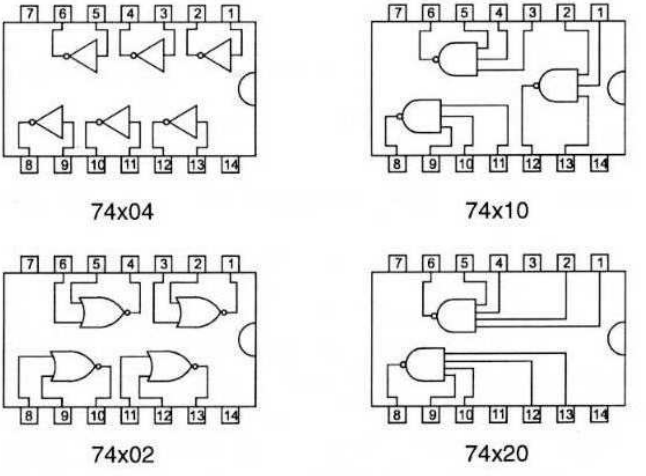


Figura 10.1 Simboli elettrici di alcuni componenti standard SSI

In Figura 10.1 sono riportati i simboli di alcuni di questi componenti standard, che possono contenere, ad esempio, invertitori, porte NAND o NOR a 2, 3, 4 o 8 ingressi, porte AND o OR fino a 4 ingressi, porte EX-OR o EX-NOR, ed altre combinazioni; il piedino 14 porta la tensione di alimentazione e quello 7 il potenziale di massa.

Le prestazioni e le caratteristiche elettriche di questi componenti sono legate alle famiglie tecnologiche utilizzate per la realizzazione dell'integrato; un singolo componente logico, definito da un determinato numero di codice (ad esempio il numero 74 × 27 corrisponde a 3 porte NOR a 3 ingressi), può essere realizzato in tecnologia CMOS, TTL o ECL (queste indicate con differenti lettere al posto della x del codice) ed avere quindi differenti specifiche pur realizzando la stessa funzione logica. Le analisi sviluppate precedentemente per le diverse famiglie logiche ci permettono di comprendere le principali differenze nelle caratteristiche delle porte standard CMOS, TTL ed ECL, che, a seconda delle differenti applicazioni, possono rendere più conveniente l'impiego dell'una o dell'altra famiglia logica. Le caratteristiche elettriche più significative di porte commerciali sono esemplificate in Tabella 10.1 con riferimento alle famiglie logiche rispettivamente CMOS-AC, TTL-ALS, ed ECL 10K e 100K.

Tabella 10.1 Caratteristiche elettriche di porte standard CMOS, TTL, ECL

famiglia logica	*CMOS-AC*	*TTL-ALS*	*ECL 10K*	*ECL 100K*
t_P (ns)	4.75	4	2	0.7
P_D/porta (mW)	statica: 0.005	1.2	(senza carico)	(senza carico)
	totale:		26	40
	f = 0.1MHz 0.08			
	f = 1 MHz 0.75			
	f = 10 MHz 7.5			
$t_P \cdot P_D$ (pJ)	f = 0.1MHz 0.4	5	52	28
	f = 1 MHz 3.6			
	f = 10 MHz 36			

Non deve sorprendere che i valori del ritardo di propagazione per le porte CMOS siano superiori a quelli valutati nel Capitolo 5 per il caso di porte CMOS con tecnologia avanzata (AC), in quanto i dati riportati si riferiscono non a singole porte integrate in circuiti VLSI, ma a componenti standard che prevedono degli stadi di uscita (stadi di buffer) in modo da poter essere collegati a carichi capacitivi relativamente elevati (5÷10 pF), in dipendenza del montaggio dell'integrato in una piastra di connessione e del suo collegamento con altri componenti integrati.

Le caratteristiche sintetizzate nella Tabella 10.1 possono essere commentate in base alle analisi svolte precedentemente.

Per le porte CMOS la potenza dissipata in condizioni di quiescenza o di bassa frequenza di pilotaggio è trascurabile; questo aspetto delle porte CMOS è di grande rilevanza e favorisce questa famiglia per molte applicazioni dove la frequenza di operazione non deve essere molto elevata. La potenza dissipata è tuttavia funzione della frequenza di pilotaggio, come riportato dalla (5.19), e quindi il bassissimo valore di P_D e del prodotto potenza-ritardo viene ad essere notevolmente ridimensionato nell'impiego a frequenze di pilotaggio relativamente elevate, e questo anche in dipendenza dell'area relativamente elevata dei MOS degli stadi di uscita, necessari per pilotare carichi capacitivi elevati.

Per le porte TTL lo stadio totem pole di uscita fornisce correnti di uscita sufficientemente elevate che permettono di tollerare carichi capacitivi anche significativi. Questo elimina la necessità di stadi di buffer e determina un prodotto potenza-ritardo più basso per queste ultime anche con frequenze di pilotaggio elevate.

Le porte ECL presentano i più bassi valori del tempo di propagazione, e quindi sono impiegate in quelle applicazioni in cui occorrono le più elevate frequenze di funzionamento, e le velocità di commutazione più rapide. Il funzionamento in logica non saturata richiede però un prezzo in termini di potenza dissipata, per cui il prodotto potenza-ritardo per queste logiche è elevato, specialmente se si utilizzano carichi adattati per gli stadi di uscita.

Le porte logiche elementari permettono di effettuare semplici operazioni logiche tra le variabili; definiremo più in generale *circuito combinatorio* un circuito logico che presenta in ogni istante alla uscita (o alle uscite) variabili logiche che dipendono solo dalla combinazione delle variabili logiche presenti nello stesso istante ai suoi ingressi (cioè non dipendono dagli stati precedenti). Tale circuito è quindi la realizzazione elettrica di una espressione logica basata su una tabella della verità tra ingressi ed uscite. I circuiti combinatori sono impiegati per realizzare funzioni logiche più complesse di quelle fornite dalle porte logiche elementari, sia per quanto riguarda le diverse funzioni presentate alle uscite (di solito per questi circuiti è prevista più di una uscita) che per le singole funzioni logiche, le quali richiedono usualmente più livelli di logica, intendendo con questi ultimi le singole operazioni di somma o di prodotto tra le variabili logiche di ingresso.

Sebbene sia possibile, come si è detto, realizzare qualsiasi espressione logica di più variabili a partire da sole porte NOR o NAND, è desiderabile ottenere funzioni logiche più complesse già nei componenti standard. Esamineremo quindi nel seguito le principali funzioni logiche disponibili nei circuiti logici combinatori standard, trattando in un successivo capitolo i circuiti digitali realizzabili con progettazione specifica (*custom design*).

Inizieremo l'argomento affrontando gli aspetti legati alle interconnessioni ed ai collegamenti tra le porte logiche, e considereremo alcune varianti delle porte logiche elementari, al fine di realizzare funzioni logiche più complesse, ottenute introducendo opportune modifiche alle configurazioni circuitali elementari viste nei capitoli precedenti; queste sono ad esempio le porte dette A-O-I, da AND-OR-INVERT e le porte per logica cablata. Tratteremo in questo contesto anche alcune

versioni di invertitori (e più in generale di porte elementari) utilizzate per le funzioni di interfacciamento tra famiglie logiche differenti. Presenteremo quindi le porte "tri-state" e gli invertitori "con isteresi", impiegate nelle funzioni ingresso/uscita, ed introdurremo in questo contesto anche gli stadi di disaccoppiamento e le porte logiche basate su tecnologia mista bipolare-CMOS, detta BiCMOS. Successivamente consideremo i circuiti combinatori più usuali sia nel campo della elaborazione dei dati, che dei circuiti di indirizzamento, codifica e decodifica.

10.2 Porte A-O-I

Queste porte realizzano una funzione logica a due livelli AND-OR-INVERT tra più variabili, ad esempio permettono di realizzare la funzione logica $Y = AB + CD$ come indicato in Figura 10.1, che è una funzione logica a due livelli in quanto contiene i prodotti tra le variabili logiche (AND) ed al secondo livello la somma negata di questi prodotti (OR-INVERT). Queste porte sono ottenibili dalle porte elementari introdotte precedentemente con modifiche contenute, permettendo quindi una maggiore compattazione del circuito combinatorio realizzato.

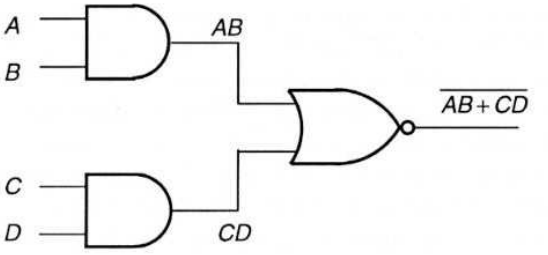


Figura 10.2 Schema logico di una porta A-O-I

Una configurazione A-O-I molto comune è quella basata su porte NAND TTL. Nella configurazione TTL, le sezioni formate dal transistore multiemettitore e dal transistore con doppio carico sono replicate m volte ed i collettori e gli emettitori di questi ultimi sono posti in parallelo con un unico carico R_C e R_E rispettivamente, in modo da pilotare un unico stadio totem pole di uscita. Viene indicata come "larghezza m" (*m-wide*) della porta A-O-I il numero di porte AND (o di ingressi OR) del circuito, mentre è indicato con n il numero degli ingressi disponibili per ogni porta AND; in Figura 10.3 è riportato lo schema elettrico di una porta A-O-I con $m = 2$ e $n = 3$.

Dal circuito si vede che la funzione AND è realizzata dal transistore multiemettitore, in quanto la tensione all'ingresso di Q_d dipende da un'operazione AND tra gli ingressi. L'operazione OR è a rigore effettuata solo sulla uscita su R_E in co-

mune agli emettitori ed è legata al parallelo dei due transistori Q_d assimilabili ad interruttori che sono collegati al livello alto (1) mentre il carico è collegato a massa (0); quest'uscita è infine applicata al transistore Q_a che funziona da invertitore elementare. Sull'uscita in comune ai collettori di Q_d si ritrova direttamente l'operazione NOR tra i rispettivi ingressi, in quanto ora il carico è connesso al livello alto (1) e gli interruttori in parallelo a quello basso (0); in questo caso però lo stadio di uscita (transistore Q_b) corrisponde ad uno stadio a collettore comune che non inverte il segnale, per cui la funzione NOR viene replicata in uscita.

Il risparmio di componenti che si ottiene è di due transistori (dello stadio totem pole) per ogni incremento di 1 unità della larghezza della porta, per cui si risparmiano $2(m-1)$ transistori per una porta A-O-I di larghezza m. Oltre a questo vantaggio sull'occupazione di area, ve ne è uno ancora più importante sul ritardo di propagazione; questo è infatti essenzialmente introdotto dal comportamento dinamico dello stadio di uscita, e quindi nel caso di una porta A-O-I di larghezza m esso si riduce a quello del singolo stadio di uscita, eliminando il ritardo aggiuntivo del secondo livello di logica (introdotto dalla porta NOR in cascata).

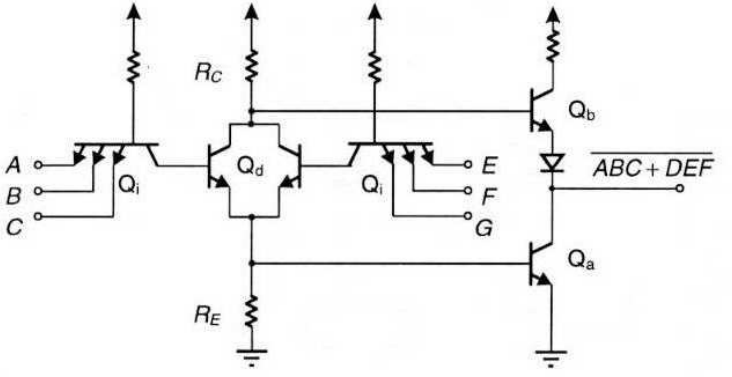
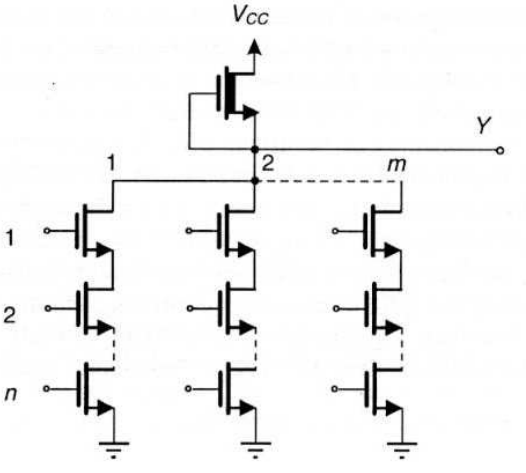


Figura 10.3 Porta TTL A-O-I con 3 ingressi e larghezza 2

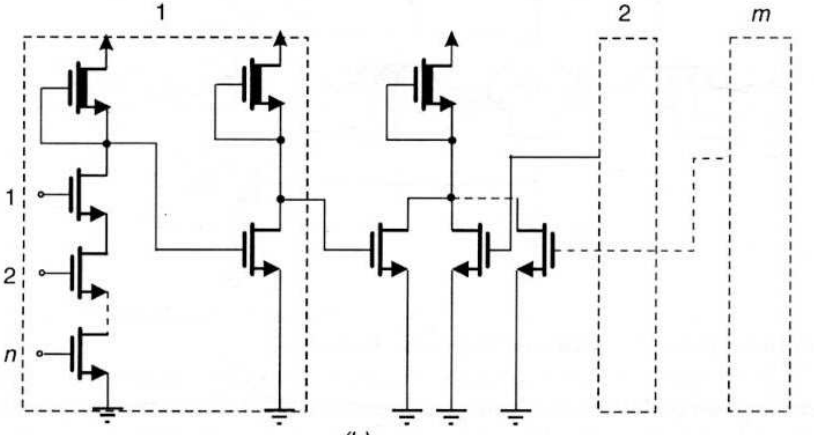
Per le porte NMOS la realizzazione di porte A-O-I è direttamente riconducibile a quella delle funzioni elementari realizzate ponendo in serie e/o parallelo gli interruttori elementari con carico comune, secondo quanto indicato in Figura 10.3.

Ad esempio, in Figura 10.4 è riportato lo schema elettrico di una porta NMOS A-O-I a n ingressi e larghezza m, confrontato con un'ipotetica realizzazione con singole porte NAND, NOR e NOT. Nella porta A-O-I gli interruttori del singolo ramo possono essere sostituiti da un unico interruttore equivalente che è *chiuso* (ingresso equivalente al livello 1) *solo se tutti gli interruttori del ramo* sono *chiusi* (tutti gli ingressi alti), mentre è aperto se anche un solo interruttore del ramo è aperto (un ingresso a livello 0); viene quindi realizzata l'operazione AND tra gli ingressi in serie per ogni ramo e l'ingresso virtuale dell'interruttore equivalente,

mentre nella connessione in parallelo dei rami (e cioè degli interruttori equivalenti) con un unico carico si effettua l'operazione NOR all'uscita rispetto agli ingressi virtuali degli interruttori equivalenti.



(a)



(b)

Figura 10.4 a) Porta NMOS A-O-I a n ingressi e larghezza m; b) realizzazione con singole porte NAND-NOT-NOR

Dal confronto di Figura 10.4 si ricava che il numero di transistori necessari per realizzare una porta A-O-I a n ingressi e larghezza m è di $n \cdot m + 1$ mentre quello necessario per una realizzazione con porte singole sarebbe di $n \cdot m + 4m + 1$; il risparmio di MOS con una porta A-O-I è quindi di $4m$. Anche in questo caso oltre al risparmio dei transistori è importante la riduzione del ritardo di propagazione che si

ottiene utilizzando per una espressione logica a due livelli una sola porta invece che tre porte elementari.

Lo schema elettrico di una porta CMOS A-O-I è riportato in Figura 10.5 per il caso di due ingressi e larghezza 2; la configurazione discende da quella utilizzata per il caso NMOS, ricordando che le configurazioni elementari NAND e NOR in versione CMOS richiedono connessioni dei PMOS rispettivamente in parallelo ed in serie (Figura 5.8). È possibile realizzare con lo stesso principio una porta OR-AND-INVERT (O-A-I) con lo stesso numero di transistori. Anche per questi casi vale quanto detto precedentemente per le porte NMOS per quanto riguarda il risparmio dei transistori ed il ritardo di propagazione.

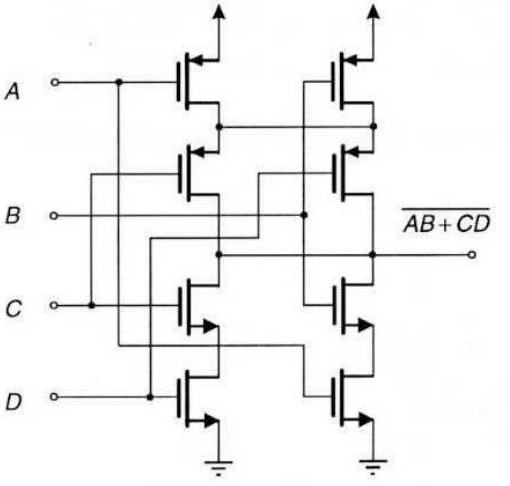


Figura 10.5 Porta CMOS AND-OR-INVERT a due ingressi e larghezza 2

10.3 Porte per logica cablata

Un'ulteriore possibilità per la realizzazione di funzioni logiche più complesse di quelle realizzate nelle porte elementari deriva dal connettere in uscita più porte con un unico carico, effettuando sul terminale di uscita una funzione logica detta "logica cablata" (*wired logic*). La connessione delle uscite è in ogni caso necessaria quando occorra convogliare su un unico collegamento (bus) le uscite di più circuiti logici; questa connessione può dar luogo ad un'ulteriore funzione logica tra le variabili delle singole uscite delle porte e la grandezza logica presente effettivamente sul collegamento comune, se si utilizzano porte che non hanno il bipolo di carico già connesso internamente.

Ad esempio collegando tra di loro le uscite di porte NAND NMOS ad un unico carico connesso alla tensione di alimentazione, come in Figura 10.6, l'uscita del punto comune sarà alta solo se tutte le uscite delle singole porte sono alte, mentre

basta che una sola di queste sia bassa perché passi corrente nel carico e la tensione del punto comune alle uscite scenda al valore basso. Il collegamento tra le uscite effettua quindi una funzione *AND cablata* (*wired AND*) tra le uscite, ed introduce un ulteriore livello di logica in quanto realizza complessivamente rispetto agli ingressi una funzione logica NAND-AND.

In realtà questa funzione è equivalente a quella presentata in Figura 10.4 per una porta A-O-I, sia in termini logici che circuitali. Infatti la connessione in parallelo delle uscite di singole porte NAND con un unico NMOS a svuotamento utilizzato come carico per tutte le porte realizza il circuito di Figura 10.4a; d'altra parte in base al teorema di De Morgan si dimostra l'uguaglianza delle due relazioni logiche fornite dai due circuiti:

$$\underbrace{\overline{A \cdot B} \cdot \overline{C \cdot D} \cdot \overline{E \cdot F}}_{\text{(NAND-AND)}} = \underbrace{\overline{A \cdot B + C \cdot D + E \cdot F}}_{\text{(AND-OR-INVERT)}} \qquad (10.1)$$

da cui si vede che le porte A-O-I possono essere considerate come casi di logica cablata.

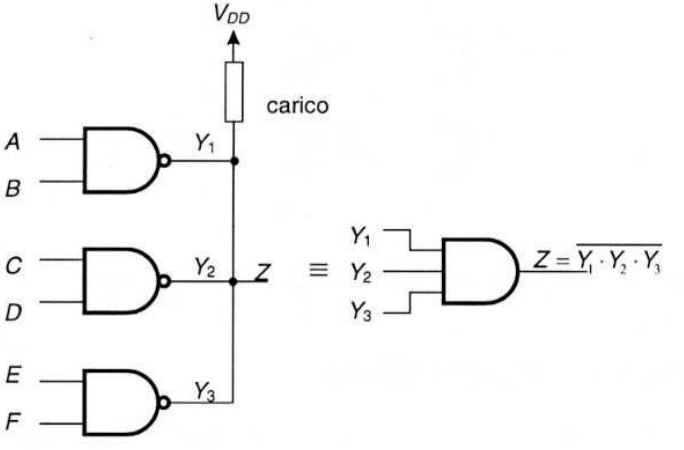


Figura 10.6 Funzione logica *wired AND* tra le uscite di più porte NAND NMOS

Con la logica cablata è possibile realizzare anche funzioni *wired OR* se si utilizzano connessioni diverse e si modifica la posizione del carico rispetto all'alimentazione. Un caso di logica cablata di questo tipo è quello che deriva dalla connessione in parallelo degli emettitori degli stadi di uscita di porte ECL in configurazione senza carico, come mostrato in Figura 10.7. La funzione logica cablata che ne consegue si può ricavare facilmente ricordando che i transistori di uscita possono essere sostituiti dai rispettivi diodi di ingresso nei ri-

guardi del loro comportamento alle variabili di ingresso (base) e di uscita (emettitore). Se uno degli ingressi è alto e quindi la rispettiva uscita è corrispondentemente alta (diodo in conduzione), gli altri diodi risulteranno interdetti perché gli ingressi sono bassi mentre l'uscita comune è alta in quanto tenuta alta dall'unico diodo in conduzione. Solo se tutti gli ingressi sono bassi l'uscita sarà bassa, in quanto in quest'ultimo caso tutti i diodi sono di nuovo in conduzione e l'uscita è ancora relazionata all'ingresso.

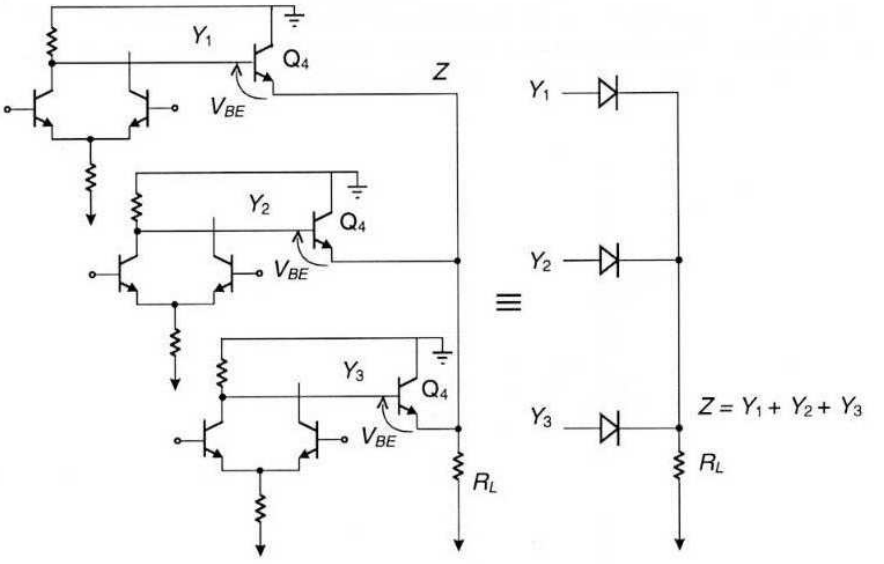


Figura 10.7 a) Logica cablata OR tra le uscite di porte ECL; b) rete equivalente a diodi per le uscite

Nel caso della Figura 10.7 sono state utilizzate le uscite NOR per effettuare la connessione in logica cablata, quindi rispetto alle variabili in ingresso alle singole porte viene effettuata una funzione NOR-OR cablata in uscita. A seconda che si utilizzino le uscite OR o NOR disponibili dalle singole porte ECL, si realizzerà una funzione OR-OR o NOR-OR. Nel primo caso tuttavia la funzione logica risultante è ancora una funzione logica ad un livello, in quanto l'uscita è ancora una funzione OR degli ingressi. Nel secondo caso, assumendo ad esempio due variabili per porta, l'uscita, in base ai teoremi di De Morgan, può essere scritta come:

$$Y = \overline{A+B} + \overline{C+D} + \overline{E+F} = \overline{(A+B)\cdot(C+D)\cdot(E+F)}$$

e quindi si ottiene una funzione logica a due livelli, e cioè prodotti di somme, in altre parole una funzione del tipo OR-AND-INVERT (O-A-I).

La connessione delle uscite in logica cablata crea problemi nel caso di porte CMOS o TTL, per le quali lo stadio di uscita è realizzato con transistori di pull-up e pull-down che agiscono alternativamente da interruttore e da carico.

Ad esempio, nel caso di porte TTL, la connessione delle uscite su un unico collegamento, o bus, pone seri problemi nel caso che una delle uscite sia alta e l'altra bassa, come indicato in Figura 10.8; non solo la grandezza sul bus risulta indefinita, in quanto sia il transistore Q_b della porta 1 che Q_a di quella 2 conducono, ma (ciò che è ben più grave) la corrente che fluisce tra i due transistori risulta inaccettabilmente alta in quanto entrambi sono in piena conduzione, e può raggiungere livelli incompatibili con l'integrità dei componenti.

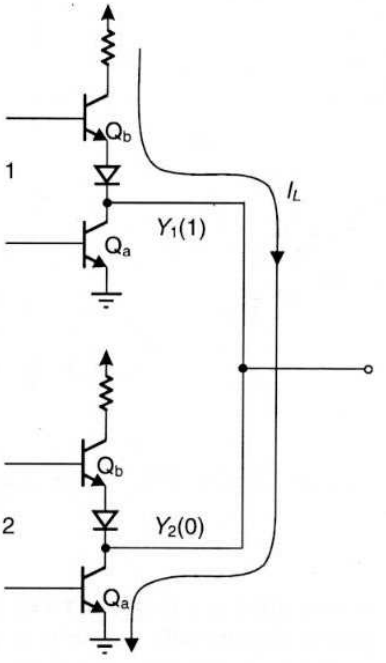


Figura 10.8 Circolazione di corrente nella connessione in uscita di porte TTL

Lo stesso problema si verifica per la connessione in uscita di porte CMOS, nel caso di contemporanea conduzione di un NMOS ed un PMOS di due porte diverse con uscite rispettivamente basse ed alte. Per le connessioni in logica cablata queste porte prevedono una versione in cui i transistori di pull-up sono sostituiti da un circuito aperto, come è indicato sinteticamente in Figura 10.9 per il caso di porte NAND.

Questa versione delle porte viene definita "a collettore aperto" (*open collector*) per le TTL e "a drain aperto" (*open drain*) per le CMOS, e viene indicata nei simboli logici delle porte con il simbolo aggiuntivo indicato in figura. Le porte richiedono l'inserzione (dall'esterno dell'integrato) di un carico R_L e di una alimentazione per l'unico transistore di uscita; il valore di R_L viene determinato in funzione del numero di porte collegate in uscita, ed è limitato, secondo le considerazioni svolte rispettivamente nei Paragrafi 4.2 e 7.2, per i valori bassi dalla massima corrente assorbibile e per quelli alti dal massimo tempo di propagazione ammissibile.

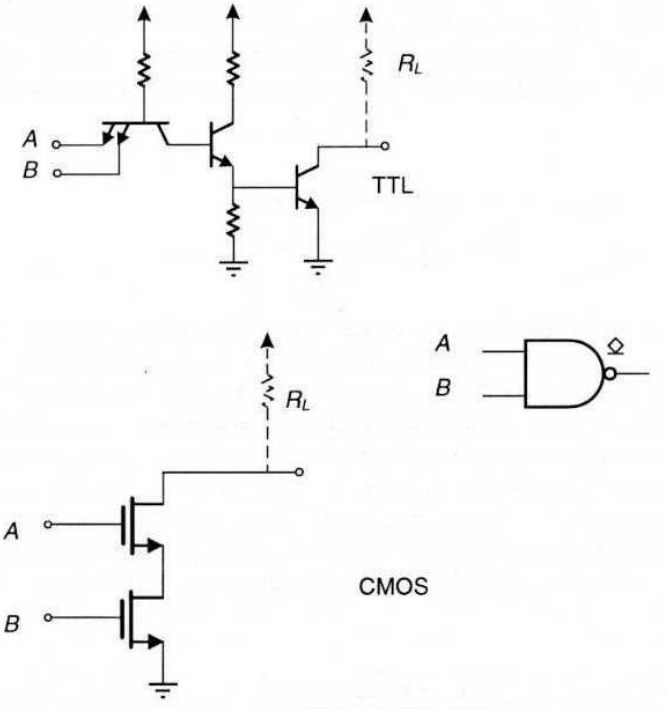


Figura 10.9 Uscite di porte NAND TTL e CMOS per logica cablata

In particolare, per la valutazione della massima corrente negli stadi di uscita, occorre considerare il caso peggiore in cui una sola delle (singole) uscite è bassa, per cui la corrente viene assorbita solo da una porta; in tal caso il carico R_L deve essere tale da mantenere nei limiti specificati il livello basso V_{OL} sul bus di uscita. Per la valutazione del tempo di propagazione occorre considerare come caso peggiore il passaggio dal livello basso a quello alto (tutte le uscite alte), considerando nella costante di tempo RC di carica oltre alla capacità del bus di uscita anche la somma di quelle di uscita di tutte le porte connesse. Queste due condizioni impongono quindi un limite superiore ed inferiore ai valori di R_L che possono essere utilizzati per una logica cablata.

Ad esempio, facendo riferimento alla porta CMOS "*open drain*" di Figura 10.9, si comprende come il valore di V_{OL} in uscita non sarà nullo, qualunque sia il valore di R_L. Si può determinare il valore di R_L eguagliando tra loro la corrente circolante nei due NMOS e quella del carico:

$$K_{NEQ}2(V_{DD} - V_T)V_{OL} = \frac{V_{DD} - V_{OL}}{R_L} \tag{10.2}$$

dove si è approssimata l'equazione della corrente nei NMOS al tratto lineare (poiché $V_{OL} << (V_{OH} - V_T)$), si è assunto $V_{OH} = V_{DD}$, e si è considerato un valore di K_{NEQ} relativo alla serie di N NMOS in serie, che, come si è visto dalla (5.25), è pari a $k'_N (W_N/NL_N)$. Se si accetta per V_{OL} un valore massimo di 0.2 V (valore compatibile con la logica TTL), dalla (10.2) si ricava un valore minimo della R_L data da:

$$R_{LMIN} = \frac{V_{DD} - 0.2}{2k_N' \dfrac{W_N}{NL_N}(V_{DD} - V_T) \cdot 0.2} \tag{10.3}$$

Questo valore della resistenza di carico, come si è detto, peggiora la dinamica della porta "*open drain*" rispetto alla versione standard della porta, che ha la rete di pull-up costituita da PMOS, in quanto nella porta "open drain" il tempo di propagazione t_{PLH} che coinvolge la carica della capacità di carico C_T attraverso la resistenza R_L sarà molto più grande di quello t_{PHL} relativo alla scarica di C_T attraverso i transistori NMOS. Si può ricavare una relazione approssimata del rapporto tra questi due tempi, ricordando che, nel caso della carica di C_T attraverso una resistenza R_L, il tempo di propagazione è dato da $0.69R_LC_L$, e che nelle porte CMOS, la rete di pull-up dei PMOS ha un $K_{PEQ} = K_{NEQ}$:

$$\frac{t_{PLH(O.D.)}}{t_{PLH}} = \frac{0.69(V_{DD} - V_{OL})C_T}{2K_{NEQ}(V_{DD} - V_T)V_{OL}} \frac{2K_{PEQ}(V_{DD} - V_T)^2}{C_T V_{DD}} \cong \frac{0.69(V_{DD} - V_T)}{V_{OL}} \tag{10.4}$$

e, nel caso considerato, con $V_{OL} = 0.2$ V, si hà un tempo di propagazione $t_{PLH\,(O.D.)} \cong 14\ t_{PLH}$; il tempo di propagazione totale t_P sarà quindi corrispondentemente più elevato di quello della porta standard che ha $t_P = t_{PHL} = t_{PLH}$.

10.4 Porte a tre stati

Le porte logiche *a tre stati* presentano in uscita tre differenti stati (elettrici) di funzionamento: lo stato logico alto, lo stato logico basso, e quello di uscita disabilitata o aperta (stato ad alta impedenza). Mentre i primi due stati di uscita corrispondono

alle specifiche combinazioni delle variabili di ingresso, secondo la funzione logica della porta, il terzo stato non dipende dalle variabili logiche in ingresso (che quindi possono assumere qualsiasi valore), ma dalla presenza di un segnale di abilitazione (*enable*) o disabilitazione (*disable*) della porta stessa, applicato ad un particolare ingresso aggiuntivo della porta.

Queste porte sono ad esempio utilizzate nella connessione di più porte ad uno stesso bus di uscita, utilizzando l'ingresso di abilitazione delle porte in modo che solo una delle porte sia di volta in volta abilitata e quindi effettivamente connessa al bus di uscita, mentre le altre sono poste nella condizione di porta disabilitata, e cioè di uscita nello stato ad alta impedenza, in modo da non disturbare il segnale logico fornito al bus dalla porta abilitata.

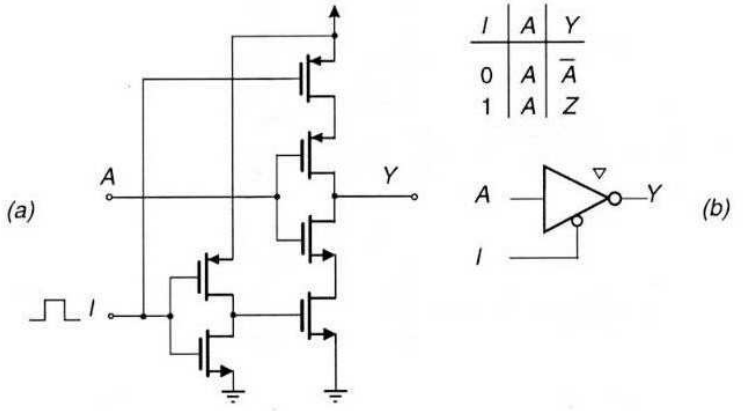


Figura 10.10 Invertitore CMOS a tre stati: a) schema elettrico; b) simbolo logico

In particolare le porte a tre stati sono utilizzate per le porte TTL e CMOS, per le quali nella connessione in uscita insorgono i problemi elettrici visti nel paragrafo precedente, a causa degli stadi di uscita con transistori di pull-up e pull-down.

Una porta CMOS a tre stati si ottiene ponendo in serie ai PMOS di pull-up ed a quelli NMOS di pull-down della porta, rispettivamente un PMOS ed un NMOS, che vengono pilotati da due segnali complementari, in modo da essere entrambi in conduzione o in interdizione; in questo secondo caso la porta è nello stato di disabilitazione e l'uscita è nello stato ad alta impedenza, indicato con Z.

In Figura 10.10 è rappresentato un invertitore CMOS a tre stati, in cui è applicato il principio esposto. In questo caso il PMOS che disconnette l'uscita verso l'alimentazione è pilotato direttamente dal segnale *I* (che in questo caso è un segnale di inibizione), mentre il NMOS che disconnette l'uscita verso massa è pilotato dal suo complementare, ottenuto dall'uscita di un secondo invertitore sulla via del segnale di inibizione: quando questo è basso entrambi i MOS sono in conduzione e l'invertitore funziona nel modo normale, cioè con un'uscita $Y = \text{NOT } A$; quan-

do il segnale di inibizione è alto i due MOS sono entrambi interdetti e l'uscita è isolata sia dall'alimentazione che dalla massa, per cui la tensione in uscita assume il valore imposto su questo terminale dalle uscite delle altre porte. La presenza di un ulteriore MOS in serie sia al PMOS che al NMOS degrada i tempi di propagazione in quanto riduce le correnti di carica e scarica della capacità di carico, se non vengono modificati i K dei MOS; per mantenere i tempi di propagazione uguali occorre quindi dimensionare opportunamente sia i MOS della porta che quelli aggiuntivi in serie.

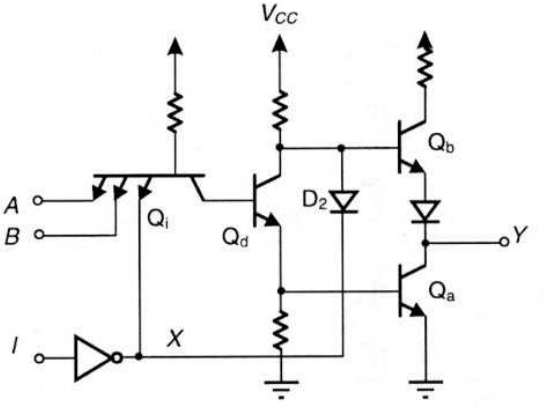


Figura 10.11 Porta TTL a tre stati

La versione TTL di una porta a tre stati è riportata in Figura 10.11: in questo caso è stato aggiunto un invertitore e un diodo allo schema standard.

Quando il segnale I (inibizione) è basso, l'uscita X dell'invertitore sarà alta e questo segnale, applicato ad un ulteriore emettitore del transistore di ingresso non altera la funzione logica effettuata in ingresso, che è una operazione AND (in altre parole, la presenza di un ulteriore segnale alto all'ingresso di una porta AND non altera la funzione logica AND tra gli altri ingressi ($A \cdot 1 = A$)). D'altra parte il segnale X alto di valore circa pari a V_{CC} (si noti che l'invertitore connesso al punto X non eroga corrente perché il diodo D_2 non permette il passaggio di corrente dal catodo all'anodo) mantiene interdetto il diodo D_2 sia quando la porta presenta un'uscita bassa V_{OL} che quando l'uscita è alta, e quindi non perturba l'uscita della porta stessa. Quando invece il segnale I è alto, l'uscita X dell'invertitore sarà bassa (V_{OL}); questo valore basso porta ad interdire il transistore Q_d come per qualunque altro ingresso basso, ed inoltre porta in conduzione il diodo D_2 vincolando così la tensione sulla base di Q_b ad essere $V_{OL} + V_D = 0.9$ V, valore inferiore al valore necessario per mantenere Q_b in conduzione. Quindi sia Q_a (per effetto dell'interdizione di Q_d) che Q_b (per effetto di D_2) sono interdetti e la porta è nello stato di disabilitazione, con l'uscita disconnessa sia da V_{CC} che da massa.

Gli invertitori ed i buffer di disaccoppiamento (cioè stadi non invertenti) a tre stati vengono di solito impiegati quando si voglia utilizzare un unico collegamento di uscita (detto *bus*) per convogliare le uscite logiche provenienti da più circuiti logici, in modo da far sì che i segnali logici vengano trasmessi al bus da un solo circuito alla volta, come è schematicamente rappresentato in Figura 10.12a.

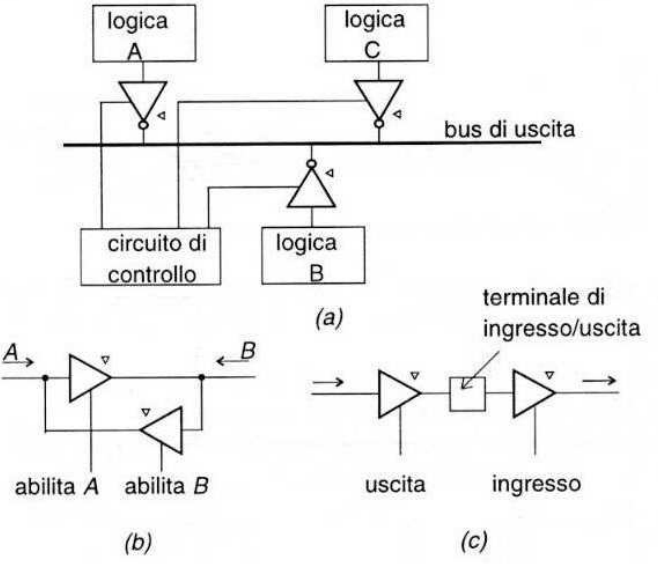


Figura 10.12 Applicazioni delle porte tri-state: a) connessioni multiple ad un singolo bus di uscita; b) buffer bidirezionale; c) terminale ingresso/uscita

In questo caso, tramite un circuito di controllo, verrà di volta in volta abilitato il buffer che è connesso al circuito logico di cui si desidera l'uscita, mentre gli altri verranno disabilitati, in modo che le uscite degli altri circuiti non interferiscano con quella voluta.

Mediante il collegamento di due buffer tri-state in opposizione (Figura 10.12b) è possibile realizzare un collegamento bidirezionale, ossia una linea che può trasmettere i segnali logici sia nell'una che nell'altra direzione. Infine è possibile utilizzare un terminale (*pin*) di un integrato, sia come terminale di ingresso che di uscita, collegandolo a due buffer tri-state nel modo indicato in Figura 10.12c; questa possibilità è utilizzata nei circuiti integrati complessi, nei quali il numero dei pin è limitato rispetto alle funzioni che il circuito deve svolgere, per cui un singolo pin può essere utilizzato sia per immettere i dati nel circuito che per prelevare i dati in uscita in tempi successivi.

10.5 Invertitori con isteresi

Un altro tipo di invertitore (o di buffer) utilizzato nelle interfacce di ingresso dei circuiti digitali è l'invertitore con isteresi, detto anche "invertitore Schmitt-trigger". Questa porta prevede una soglia logica diversa a seconda che il segnale di ingresso passi dallo stato basso a quello alto o da quello alto a quello basso; la soglia logica in particolare è più elevata nel primo caso e più bassa nel secondo. Un circuito con queste caratteristiche è particolarmente utile nel caso di segnali lentamente variabili, in quanto previene commutazioni spurie in presenza di un rumore sovrapposto al segnale.

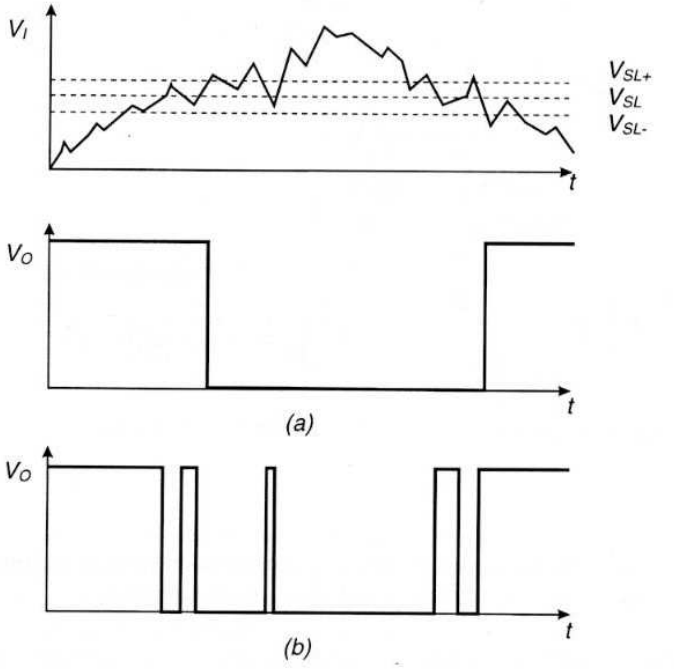


Figura 10.13 Comportamento di un invertitore con isteresi: a) uscita dall'invertitore con isteresi; b) uscita da un invertitore a soglia unica

Ad esempio, con riferimento alla Figura 10.13, assumendo una soglia $V_{SL}+ = 3$ V quando il segnale passa dal livello basso a quello alto, ed una soglia $V_{SL}- = 2$ V nel caso opposto, in presenza di un rumore sovrapposto al segnale di ampiezza minore a $V_{SL}+ - V_{SL}-$, si avrà una sola commutazione dell'uscita dall'invertitore con isteresi (Figura 10.13a), mentre nel caso di un invertitore senza isteresi, con una soglia logica $V_{SL} = 2.5$ V, si avrebbe una serie di commutazioni spurie come nel caso di Figura 10.13b, se il tempo di salita del segnale è maggiore di quello dell'invertitore.

Facendo riferimento alla logica CMOS, per realizzare un invertitore che presenti due differenti tensioni di soglia logica, è necessario realizzare due diverse

caratteristiche di trasferimento, a seconda che il segnale logico passi dal livello basso a quello alto, o viceversa, come è indicato in Figura 10.14a. Questo comportamento può essere ottenuto con una reazione interna inserita nel circuito, che modifichi la rete dell'invertitore in funzione del livello logico del segnale.

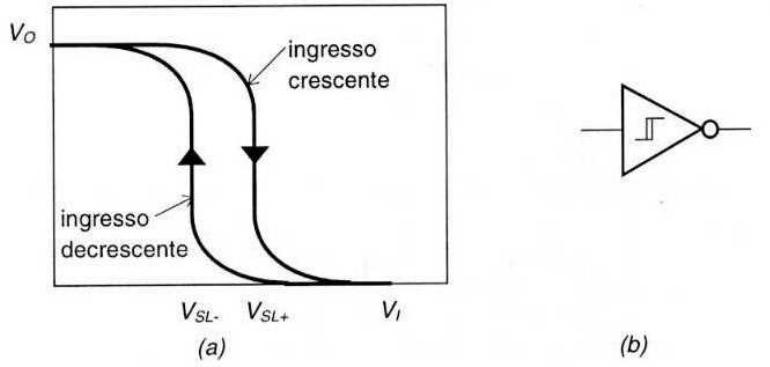


Figura 10.14 a) Caratteristiche di trasferimento per un invertitore "Schmitt-trigger" CMOS; b) simbolo logico dell'invertitore "Schmitt-trigger"

Dall'analisi della caratteristica di trasferimento dell'invertitore CMOS effettuata nel Paragrafo 5.3, ricordiamo che il valore della soglia logica (tratto verticale della caratteristica) dipende dai valori di K_N e K_P rispettivamente del NMOS e del PMOS; se i due K sono uguali $V_{SL} = V_{DD}/2$. Più in generale il valore di V_{SL} nel caso di $K_N \neq K_P$ si può ottenere eguagliando le correnti dei due MOS in regime di pinch-off:

$$K_N\left[V_I - V_{TN}\right]^2 = K_P\left[V_{DD} - V_I - |V_{TP}|\right]^2 \tag{10.5}$$

da cui si ricava per il valore di $V_I = V_{SL}$:

$$V_I \equiv V_{SL} = V_{DD}\frac{\sqrt{\frac{K_P}{K_N}}}{1+\sqrt{\frac{K_P}{K_N}}} \tag{10.6}$$

Dalla (10.6) si ricava, ad esempio, che se si dimensiona l'invertitore con $K_P = 2K_N$, si ha una tensione $V_{SL} = 0.58V_{DD}$, mentre per $K_P = 0.5K_N$, si ha una tensione $V_{SL} = 0.41V_{DD}$; occorre quindi che una o entrambe le reti di pull-up e pull-down siano modificabili in funzione di una reazione interna che abilita o meno uno dei MOS in parallelo in funzione del livello del segnale logico, in modo da variare il K_{EQ} di una rete (o di entrambi).

Una possibile realizzazione del circuito dell'invertitore con isteresi è quella riportata in Figura 10.15. In questo circuito la rete di pull-down è realizzata con due NMOS (N_A e N_B) in parallelo (non consideriamo per il momento il transistore N_C), di cui uno è pilotato dal segnale di ingresso e l'altro dall'uscita di un ulteriore invertitore, a sua volta pilotato dal segnale in uscita dell'invertitore con isteresi.

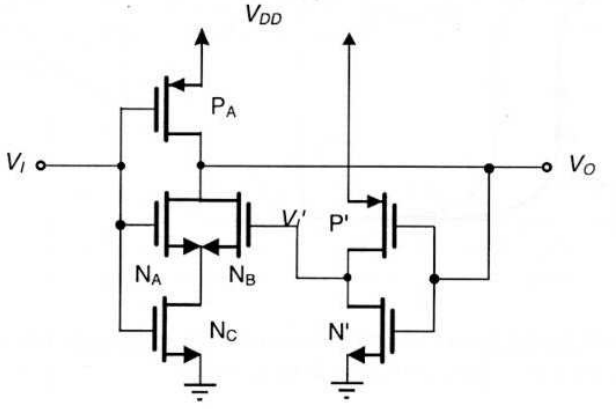


Figura 10.15 Schema circuitale di un invertitore con isteresi

Quando V_I è al valore logico basso, è bassa anche la tensione V_I' all'ingresso di N_B (che è invertita rispetto a V_O dal secondo invertitore), ed il transistore N_B è interdetto, per cui il K_{NEQ} della rete NMOS è più basso, mentre quando V_I è al livello logico alto N_B va in conduzione, ed il K_{NEQ} della rete è più alto.

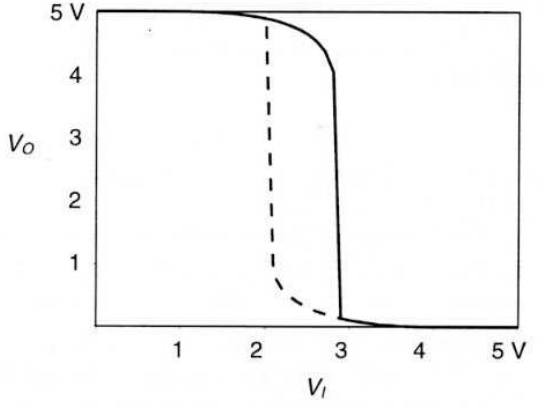


Figura 10.16 Caratteristiche di trasferimento del circuito di Figura 10.14 per valori crescenti di V_I (curva continua) e per valori decrescenti (curva tratteggiata). I valori di dimensionamento dei MOS sono: $W/L|_{PA} = 20/2$, $W/L|_{NA} = 4/2$, $W/L|_{NB} = 28/2$, $W/L|_{NC} = 32/2$, $W/L|_{P'} = 10/2$, $W/L|_{N'} = 4/2$

L'inserzione di N_C serve a impedire la conduzione del ramo degli NMOS quando V_I è basso; in assenza di N_C il valore di V_I' si posizionerebbe presso il valore della soglia logica del secondo invertitore, portando il MOS N_B in conduzione in modo da avere una $V_O \cong V_I'$ ed un funzionamento non corretto del circuito. Per i valori di *W/L* dei MOS indicati in Figura 10.16, i valori dei K_{NEQ} delle due reti di pull-down sono rispettivamente: per N_B interdetto $K_{NEQ} \cong 0.5K_P$; per N_B in conduzione $K_{NEQ} = 2K_P$, ed i valori delle due soglie logiche ricavati dall'analisi SPICE del circuito, riportata nella stessa figura, sono prossimi a quelli determinati in base alla (10.6).

10.6 Interfacciamento di famiglie logiche differenti

Nei sistemi digitali occorre in qualche caso utilizzare contemporaneamente diverse famiglie logiche per sfruttare al meglio le caratteristiche peculiari di ciascuna di esse. Questo può essere ad esempio il caso di una realizzazione in logica CMOS di un circuito ad alta integrazione che sfrutta la bassa dissipazione di potenza di questa logica, ma con periferiche TTL per migliorare il trasferimento dei segnali sui bus di interconnessione con capacità di carico relativamente elevate; un altro esempio è l'utilizzazione di circuiti TTL per realizzare funzioni logiche con relativamente bassa frequenza di operazione per sfruttare il basso prodotto potenza-ritardo di questa famiglia, interfacciandola con logiche ECL nelle unità logico-aritmetiche (ALU) per migliorare le prestazioni temporali del sistema, in quanto questa famiglia permette un elevato numero di operazioni al secondo.

Occorre quindi prevedere dei circuiti che permettano di modificare, o in ogni caso di adattare, i livelli logici forniti da una logica con quelli necessari per pilotare una logica differente: questa operazione di adattamento dei segnali logici viene detta *interfacciamento*, e circuiti di interfacciamento sono i circuiti che effettuano questo adattamento dei livelli tra logiche specifiche.

a) Interfacciamento ECL-TTL

L'interfacciamento tra l'uscita di una porta ECL e l'ingresso di una porta TTL richiede non solo una modifica dei valori dei livelli logici bassi ed alti, ma anche una traslazione di questi livelli dai valori negativi presenti nella logica ECL a quelli positivi della logica TTL.

L'operazione di traslazione può essere effettuata con una versione modificata della stessa porta ECL secondo lo schema semplificato di Figura 10.17. Il circuito utilizza la sezione destra della coppia differenziale della porta ECL, alimentata con la tensione positiva V^+ pari a quella della porta TTL, per effettuare la traslazione dei livelli dell'uscita (presa sul collettore di Q_2); lo stadio di uscita a collettore comune trasferisce (con un'ulteriore traslazione) il segnale all'ingresso della porta TTL. Sull'altro ramo della coppia differenziale si possono connettere più transistori in parallelo per effettuare operazioni OR tra i segnali di ingresso.

Con riferimento al singolo segnale A in ingresso da trasferire alla porta TTL, se questo è basso Q_2 conduce e la tensione V_I all'ingresso del circuito TTL vale:

$$V_L \cong V^+ - V_{BE4} - R_{C2} I_E = V_{OL(TTL)} \tag{10.7}$$

da cui, ricordando che la corrente I_E è circa 4 mA, si ricava il valore di R_{C2}:

$$R_{C2} = \frac{5 - 0.7 - 0.2}{4} \cong 1\ \text{k}\Omega \tag{10.8}$$

Se A è alto, Q_2 è interdetto e la tensione V_I vale:

$$V_H \cong V^+ - V_{BE4} = V_{OH(TTL)} \cong 4.3\ \text{V} \tag{10.9}$$

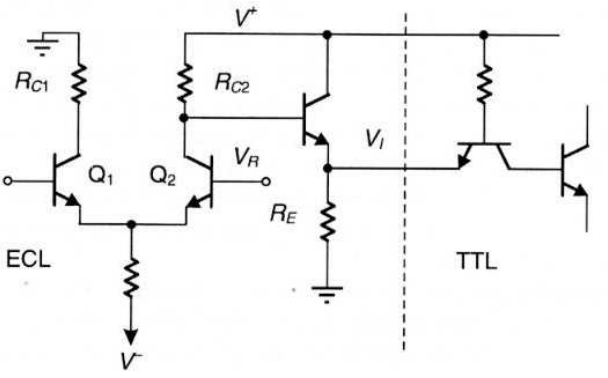


Figura 10.17 Circuito di interfacciamento ECL-TTL

Il valore della resistenza R_E d'altra parte non deve essere troppo elevato; il valore superiore di R_E è determinato dalla corrente I_L iniettata in questa dalla porta TTL in uscita, che non deve dar luogo ad una caduta $I_L R_E > V_{OL(TTL)}$.

Si può interfacciare anche l'uscita NOR della ECL con l'ingresso TTL, collegando R_{C2} a massa e R_{C1} alla tensione V^+.

b) Interfacciamento TTL-ECL

In questo caso occorre traslare a valori negativi i livelli di uscita della porta TTL perché siano utilizzabili dalle logiche ECL. Con riferimento allo schema di Figura 10.18, l'operazione viene effettuata per mezzo della rete di resistenze e diodo D.

Nel caso di uscita alta (V_{OH}) dalla porta TTL, occorre che il diodo D sia interdetto e cioè:

$$V_A = V^+ - \frac{(V^+ - V^-)R_A}{R_A + R_B + R_C} < V_{OH(TTL)} \tag{10.10}$$

Se D è interdetto, la tensione V_I in ingresso alla ECL, trascurando la (debole) corrente assorbita dall'ingresso della porta ECL, vale:

$$V_I = \frac{V^+ R_C + V^- (R_A + R_B)}{R_A + R_B + R_C} = V_{OH(ECL)} \cong -0.7 \text{ V} \tag{10.11}$$

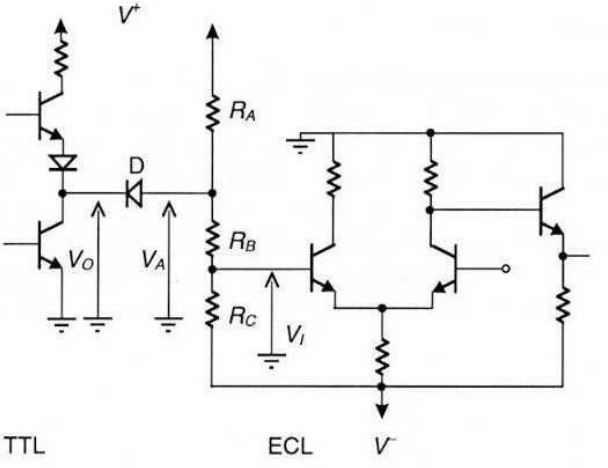


Figura 10.18 Circuito di interfacciamento TTL-ECL

Nel caso di uscita bassa (V_{OL}) dalla TTL, il diodo D conduce se $V_A - V_{OL} > V_D$ (questa condizione pone quindi un ulteriore vincolo sul valore di V_A); in questo caso la tensione V_I vale:

$$V_I = \frac{(V_{OL(TTL)} + V_D)R_C + V^- R_B}{R_B + R_C} = V_{OL(ECL)} \cong -1.7 \text{ V} \tag{10.12}$$

Le tre relazioni permettono di determinare i valori delle resistenze R_A R_B R_C della rete di interfacciamento. Con le relazioni (10.10) ÷ (10.12) si definiscono i valori dei rapporti tra le resistenze, e non i valori assoluti; per determinare questi valori occorre introdurre ulteriori elementi di valutazione, come la dissipazione di potenza e la degradazione della dinamica delle transizioni, introdotte dalla rete di interfacciamento. Infatti, per ridurre la dissipazione di potenza causata dalla rete, occorre scegliere valori elevati delle resistenze, in particolare occorre che la potenza assorbita dalla rete di interfacciamento, sia nel caso di ingresso logico alto che di ingresso basso, sia non superiore a quella assorbita

dalle porte ECL e TTL. Per non degradare le transizioni del segnale logico, invece, occorre scegliere bassi valori delle resistenze, affinché la costante di tempo τ introdotta dalla rete:

$$\tau = \frac{R_A(R_B + R_C)}{R_A + R_B + R_C} C_I$$

dove C_I è la capacità di ingresso della porta ECL, sia piccola rispetto ai tempi di transizione introdotti dalle porte stesse. La scelta andrà fatta con un criterio di compromesso tra le due esigenze contrastanti.

c) Interfacciamento TTL-CMOS

In questo caso l'uscita bassa della porta TTL è perfettamente compatibile con quella richiesta dalle porte CMOS, in quanto è minore ($\cong$ 0.2 V) della tensione di soglia dei MOS delle porte CMOS ($\cong$ 0.8 V). L'uscita alta invece può non essere sufficiente ad interdire il PMOS, se vi sono altre porte TTL connesse in uscita, in quanto la $V_{OH\,(TTL)} \cong 3.8 \div 2.7$ V, mentre occorre una tensione superiore a 4 V ($V^+ - V_T$) per interdire il PMOS; in questo caso (vedi Figura 10.19) si inserisce all'ingresso della porta CMOS una resistenza R_P di pull-up per riportare l'ingresso della porta CMOS alla tensione V^+ interdicendo il diodo in uscita della TTL nel caso di uscita alta V_{OH}.

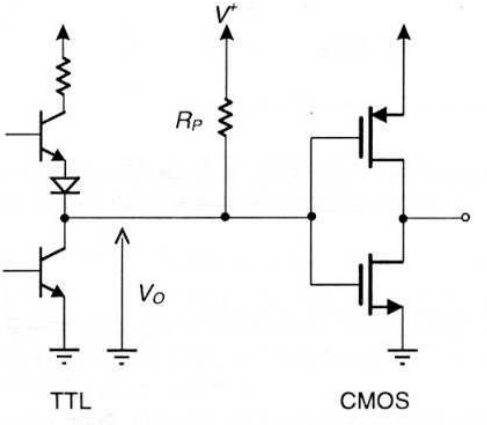


Figura 10.19 Interfacciamento TTL-CMOS

d) Interfacciamento CMOS-TTL

In questo caso i livelli di uscita della logica CMOS sono già compatibili con quelli richiesti all'ingresso delle TTL. Occorre tuttavia prevedere per lo stadio di interfacciamento CMOS un K_N del NMOS (e quindi un rapporto W/L) maggiore di

quello K_P del PMOS, in quanto il primo deve assorbire la corrente I_{IL} erogata dalla porta TTL nello stato basso, mentre il PMOS deve fornire la corrente I_{IH} (ben minore della prima) che la porta TTL assorbe nello stato alto.

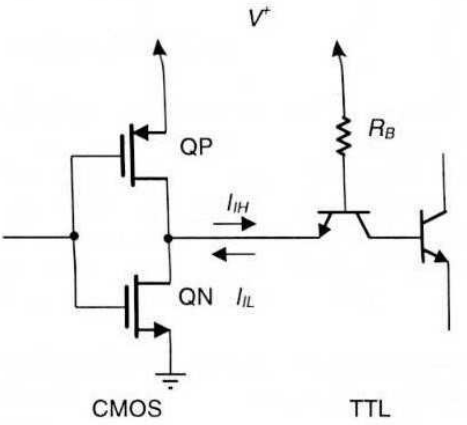


Figura 10.20 Interfacciamento CMOS-TTL

Nel caso della porta TTL standard, la corrente I_{IL}, quando l'ingresso è al livello logico basso, è data dalla (8.23), e per una resistenza $R_B = 4$ kΩ vale circa 1 mA; questa corrente circolerà nel MOS QN (vedi Figura 10.20), per cui il livello logico basso V_{OL} in uscita dalla porta CMOS sarà diverso da zero. Per avere un valore V_{OL} pari al livello logico basso della TTL, occorre che la caduta su QN in presenza della corrente I_{IL} sia pari al valore $V_{OL\,(TTL)} = 0.2$ V. Ricordando che per basse tensioni di drain il MOS ha un comportamento ohmico, con un valore di resistenza R_{MOS} dato dalla (3.11), si può determinare il valore del rapporto W/L per QN dalla relazione:

$$V_{OL(TTL)} = \frac{I_{IL}}{K_N\, 2(V_{GS} - V_T)} = \frac{I_{IL}}{k'_N \dfrac{W}{L} 2(V_{DD} - V_T)} \tag{10.13}$$

dove si è assunto per V_{GS} il valore $V_{OH\,(CMOS)} = V_{DD}$.

In Figura 10.21 sono riportate le caratteristiche di trasferimento di un invertitore CMOS caricato da una porta TTL. Nel caso (a) (curva tratteggiata) i due MOS hanno lo stesso valore di K, con un rapporto $W/L = 2$ per il NMOS; in questo caso, quando la porta TTL comincia a fornire una corrente I_{IL} la tensione di uscita dell'invertitore CMOS (tensione di drain di QN) aumenta notevolmente, e per un valore $V_I = 5$ V non si raggiunge una tensione $V_{OL} < V_{IL}$. Nel caso (b) (curva continua), si è scelto un valore $W/L = 40$ per QN, e la tensione

V_{OL} si riduce fino ad un valore di circa 0.2 V, compatibile con il valore V_{OL} di una TTL.

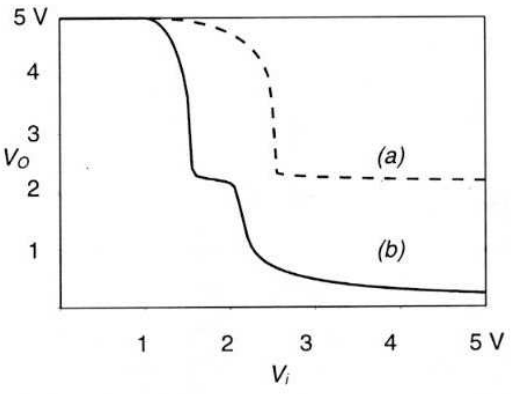


Figura 10.21 Caratteristica di trasferimento di un invertitore CMOS caricato da una porta TTL: a) con $K_N = K_P$; b) con $K_N = 20\ K_P$

Se si utilizza un interfacciamento con porte TTL-LS, la corrente I_{IL} sarà circa 5 volte più piccola (in quanto aumenta di un fattore 5 la resistenza R_B) e quindi il rapporto *W/L* per QN può essere scelto pari a 8.

10.7 Invertitori e porte logiche BiCMOS

La logica BiCMOS (da *Bipolar-CMOS*) è stata sviluppata verso la fine degli anni '80, con lo scopo di combinare la tecnologia CMOS, che presenta i vantaggi di flessibilità progettuale, elevato livello di integrazione e basso consumo di potenza, con la tecnologia bipolare, che presenta una maggiore capacità di pilotaggio di carichi capacitivi elevati, in particolare nel caso dello stadio di uscita delle logiche TTL.

Questa logica è basata sulla possibilità di inserire i passi tecnologici necessari per realizzare i transistori bipolari nel processo base della tecnologia CMOS; ciò è stato realizzato a valle dell'introduzione dei processi a doppia tasca e dei substrati con epitassia, utilizzati per migliorare le prestazioni dei circuiti CMOS (vedi Paragrafo 5.13), e delle innovazioni introdotte nei processi per transistori bipolari avanzati, quali gli emettitori in polisilicio, le strutture autoallineate, e gli isolamenti LOCOS. I processi tecnologici per realizzare strutture BiCMOS sono quindi più sofisticati di quelli utilizzati per logiche solo CMOS o solo bipolari, ma sono impiegati in tutte quelle applicazioni nelle quali un aumento della velocità di funzionamento bilanci il maggior costo della tecnologia. Le applicazioni più utili per questi circuiti sono nel pilotaggio di bus di interconnessione, nei circuiti ingresso-uscita, nell'alimentazione di lun-

ghe linee dati, ed in generale in tutti quei casi in cui occorre un pilotaggio di carichi capacitivi elevati in tempi brevi.

Considereremo gli aspetti essenziali di questa logica, discutendo dell'invertitore elementare BiCMOS, e partendo da considerazioni basate sui vincoli circuitali necessari per la connessione tra la sezione CMOS (usualmente costituente lo stadio di ingresso) e quella bipolare (con cui è realizzato lo stadio di uscita).

10.7.1 Invertitore BiCMOS

Lo schema di principio di un invertitore che combina uno stadio di ingresso CMOS con uno di uscita di tipo bipolare è quello indicato in Figura 10.22. Il circuito è formato in effetti da un invertitore CMOS, connesso ad uno stadio di uscita di tipo totem pole come quello delle logiche TTL, mediante una rete di interfaccia necessaria per pilotare correttamente i due transistori dello stadio di uscita, i quali richiedono (come si è visto in particolare nel Paragrafo 8.3) due segnali in opposizione di fase per il pilotaggio dei due transistori Q_a e Q_b.

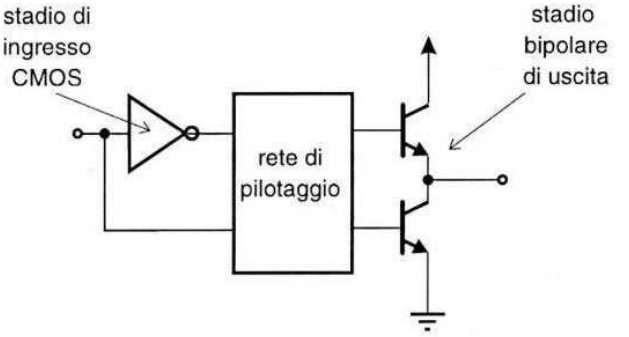


Figura 10.22 Schema di principio dell'invertitore BiCMOS

Un primo schema circuitale dell'invertitore è quello riportato in Figura 10.23. La rete di interfaccia è costituita dai due NMOS N_2 e N_3 connessi in serie, e pilotati rispettivamente dal segnale di ingresso dell'invertitore CMOS e da quello di uscita, in modo da avere un funzionamento di tipo complementare per i due NMOS, analogo a quello dell'invertitore CMOS ma ottenuto utilizzando solo dispositivi NMOS (di area più ridotta); si supporrà inoltre che il carico sia costituito da una capacità C_L che deve portarsi ai livelli logici alto e basso.

Quando il segnale V_I in ingresso all'invertitore CMOS è al livello logico basso, N_2 è interdetto, e N_3 è in conduzione perchè V_O' è al valore V_{DD}. Quindi il transistore Q_a è interdetto perché $I_{Ba} = 0$, e Q_b è in conduzione perché $V_{Bb} = V_{DD}$. La tensione di uscita V_O si porterà al livello alto, ma non può superare il valore $V_{DD} - V_{BE\gamma}$, in quanto per V_O pari a questo valore Q_b è al limite della interdizione e la tensione di uscita non può crescere ulteriormente. Per V_I al livello logico alto (V_{DD}), V_O' è al

livello basso (0), e N_3 si interdice; contemporaneamente N_2 conduce perché pilotato da $V_I = V_{DD}$, e quindi Q_b passa all'interdizione, mentre Q_a va in conduzione, essendo la base connessa (tramite N_2) al collettore. La conduzione di Q_a si arresta se V_O scende sotto il valore V_γ, perché in questo caso, essendo $V_O = V_{BEa}$ anche quest'ultima scenderebbe sotto il valore di interdizione, e quindi C_L non può scaricarsi oltre tale valore.

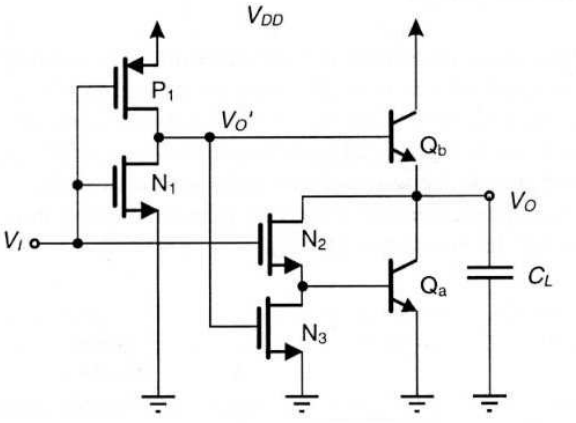


Figura 10.23 Schema elettrico dell'invertitore elementare BiCMOS

I livelli logici di questo invertitore sono quindi:

$$V_{OH} = V_{DD} - V_{BE\gamma} \; ; \qquad V_{OL} = V_{BE\gamma} \tag{10.14}$$

In ognuno dei due stati logici l'invertitore non presenta dissipazione di potenza statica. Ciò è caratteristica nota dell'invertitore CMOS, ma è anche vero per la rete NMOS di interfaccia, perché essa è pilotata da due segnali complementari, per cui se un NMOS conduce l'altro è interdetto; infine anche lo stadio bipolare di uscita, corrispondente al totem pole della logica TTL non assorbe potenza sia quando l'uscita è alta che bassa, per cui il circuito presenta solo una dissipazione di potenza dinamica e ben si presta all'integrazione in circuiti VLSI.

La caratteristica di trasferimento dell'invertitore elementare BiCMOS, ricavata mediante simulazione SPICE, è riportata in Figura 11.24. Dalla simulazione si può notare che l'escursione della tensione di uscita è inferiore a V_{DD}, anche se le differenze sia nel valore alto che in quello basso sono minori del valore $V_{BE\gamma}$; ciò è dovuto al fatto che la simulazione tiene conto della caratteristica effettiva delle giunzioni base-emettitore, che presentano tensioni anche inferiori a $V_{BE\gamma}$ per correnti trascurabili, per cui le tensioni V_{BE} senza carico, come in questo caso, sono inferiori a $V_{BE\gamma}$.

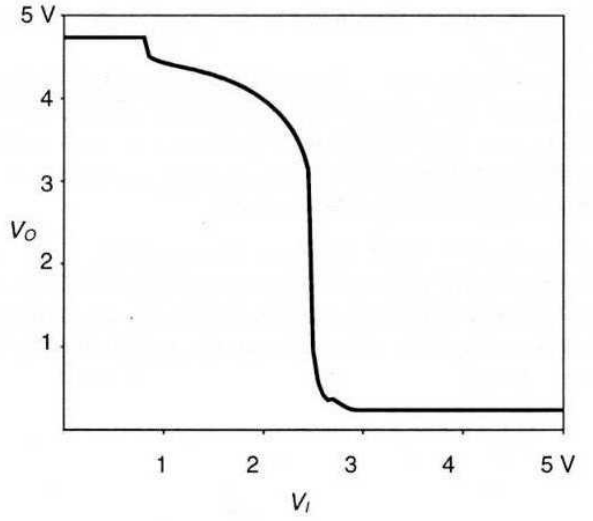


Figura 10.24 Simulazione SPICE della caratteristica di trasferimento dell'invertitore BiCMOS di Figura 10.23. I valori dei parametri sono: $W/L|_N$= 2 μm/1 μm, $W/L|_P$= 5 μm/1 μm, $V_{TN} = |V_{TP}| = 0.8$ V, $\beta_F = 50$.

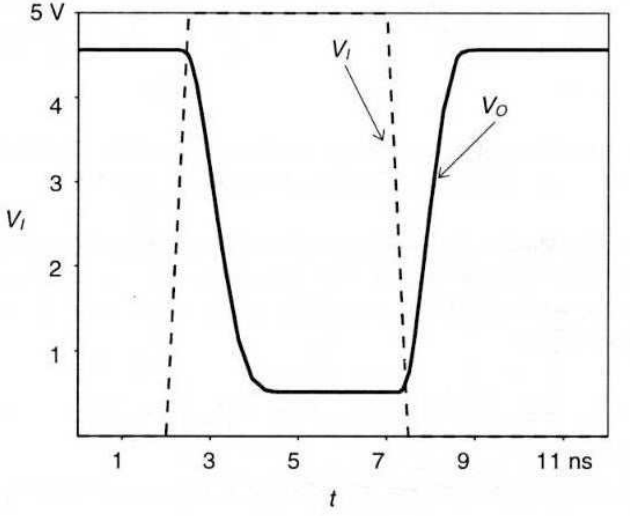


Figura 10.25 Comportamento dinamico dell'invertitore BiCMOS caricato da una capacità C_L= 5 pF

La riduzione dell'escursione logica risulta invece evidente se si considera una capacità di carico di valore non trascurabile all'uscita, e si analizza la dinamica dell'invertitore così caricato. Questa analisi dinamica è riportata in Figura 10.25; in

questo caso si vede che la capacità arresta la sua carica al valore $V_{DD} - V_{BE\gamma}$, e termina la sua scarica al valore $V_{BE\gamma}$.

Si può effettuare un'analisi dinamica semplificata del circuito di Figura 10.22, utilizzando le due reti riportate in Figura 10.26, alle quali si riduce, in opportune ipotesi semplificative, il circuito dell'invertitore, rispettivamente per la transizione della tensione di uscita dal livello logico alto a quello basso o viceversa.

Le ipotesi semplificative per l'analisi dinamica sono:

a. il segnale V_I in ingresso presenta fronti di salita e discesa nulli;
b. la tensione V_O' in uscita dall'invertitore CMOS varia più rapidamente di quella in uscita, e si assume che essa sia già al valore di regime durante la transizione;
c. la dinamica dei transistori bipolari dello stadio di uscita è molto più rapida di quella della capacità di uscita.

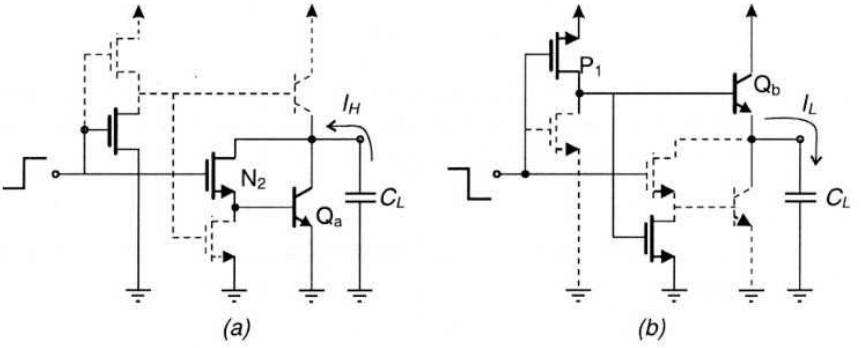


Figura 10.26 Reti equivalenti semplificate per la dinamica dell'invertitore BiCMOS: a) rete per la transizione alto-basso; b) rete per la transizione basso-alto

In tali ipotesi, e ricordando che i transistori bipolari lavorano in regime attivo diretto quando sono in conduzione, poiché per ognuno di essi $V_{BC} \cong 0$, in quanto la giunzione è cortocircuitata dal MOS in conduzione, si ha dalla rete per la transizione di discesa (Figura 10.26a):

$$I_H = I_{Qa} + I_{N2} = (\beta_F + 1) K_{N2} [V_{DD} - V_{BEa} - V_T]^2 \qquad (10.15a)$$

Quindi il tempo di propagazione t_{PHL} per il fronte di discesa sarà dato da:

$$t_{PHL} = \frac{C_L (V_{DD} - 2V_{BE\gamma})}{2(\beta_F + 1) K_{N2} [V_{DD} - V_{BEa} - V_T]^2} \qquad (10.16a)$$

Dalla rete per la transizione di salita (Figura 10.26b) si ottiene analogamente:

$$I_L = I_{Qb} + I_{P1} = (\beta_F + 1)K_{P1}[V_{DD} - V_{BEb} - V_T]^2 \qquad (10.15b)$$

e il tempo di propagazione t_{PLH} di questo fronte sarà dato da:

$$t_{PLH} = \frac{C_L(V_{DD} - 2V_{BE\gamma})}{2(\beta_F + 1)K_{P1}[V_{DD} - V_{BEb} - V_T]^2} \qquad (10.16b)$$

Dalle espressioni (10.16a) e (10.16b) si deduce che i due tempi di propagazione sono inferiori di $(\beta_F + 1)$ a quelli di un invertitore CMOS caricato dallo stesso valore di C_L. I due tempi t_{PHL} e t_{PLH} saranno inoltre uguali se si dimensionano i valori dei K dei transistori P_1 e N_2 in modo che:

$$K_{N2} = K_{P1} \Rightarrow \left.\frac{W}{L}\right|_{P1} = 2.5\left.\frac{W}{L}\right|_{N2} \qquad (10.17)$$

I valori dei livelli logici indicati nella (10.14) possono creare problemi nel pilotaggio di porte CMOS in uscita, in quanto i valori di $V_{BE\gamma}$ sono praticamente uguali a quelli delle tensioni di soglia V_{TN} e $|V_{TP}|$ dei MOS dell'attuale tecnologia, e quindi le porte CMOS pilotate da questo invertitore possono presentare una dissipazione di potenza statica. Una versione di invertitore BiCMOS che presenta un'escursione logica pari all'intera tensione di alimentazione, e quindi presenta livelli logici $V_{OH} = V_{DD}$, e $V_{OL} = 0$ è quella del circuito di Figura 10.27.

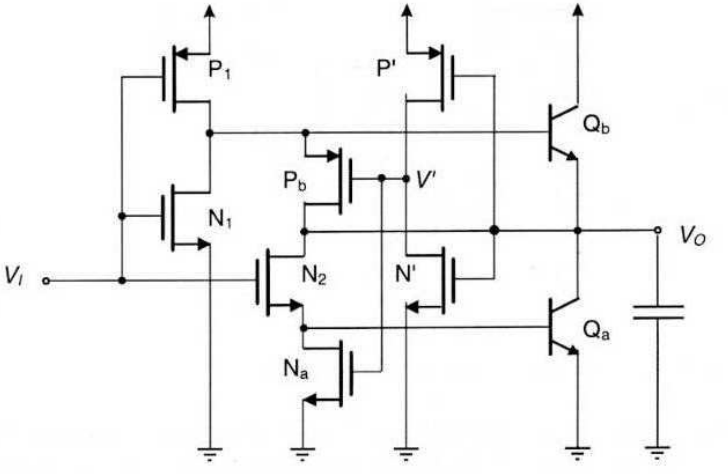


Figura 10.27 Invertitore BiCMOS con escursione logica completa

Il circuito prevede due MOS, P_b e N_a, rispettivamente in parallelo alle giunzioni base-emettitore dei transistori Q_b e Q_a; questi MOS vengono pilotati dall'uscita V' di un ulteriore invertitore CMOS formato dai transistori N' e P', a sua volta pilotato dall'uscita V_O. Quando l'uscita V_O è al valore alto, V' è basso e P_b è in conduzione; quindi in parallelo alla giunzione base-emettitore di Q_b compare la resistenza equivalente di P_b e, per correnti trascurabili (cioè al termine della carica di C_L), la tensione V_{BEb} ai capi di questa giunzione tende a zero invece che a V_γ. Analogamente, quando l'uscita passa al valore basso, V' va al valore alto e porta in conduzione il MOS N_a, e la tensione V_{BEa} della giunzione base-emettitore del transistore Q_a tende a zero per correnti trascurabili.

In tal modo l'invertitore presenterà come valore logico alto $V_{OH} = V_{DD}$, e come valore logico basso $V_{OL} = 0$, e si ottiene un'escursione logica pari all'intera tensione di alimentazione V_{DD}.

10.7.2 Porte logiche BiCMOS

La tecnologia BiCMOS può essere utilizzata non solo per realizzare invertitori e stadi di disaccoppiamento, ma anche porte logiche elementari e complesse. Lo schema di principio che permette di realizzare funzioni logiche a più variabili con strutture BiCMOS è quello indicato in Figura 10.28.

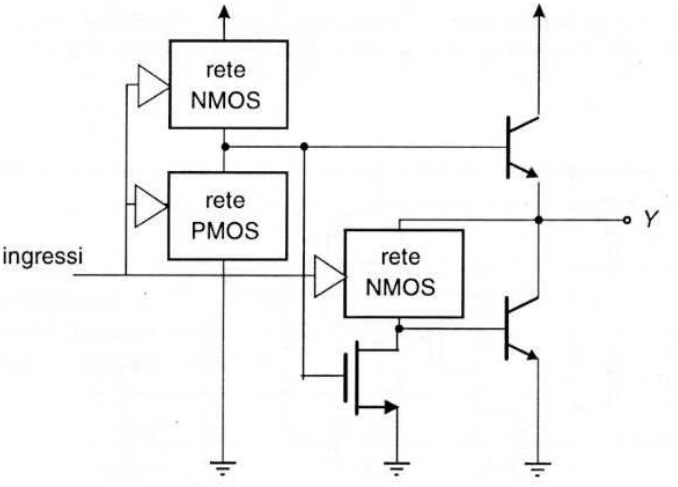


Figura 10.28 Circuito di principio di una porta logica BiCMOS

La funzione logica voluta viene realizzata sostituendo rispettivamente al PMOS ed al NMOS dell'invertitore CMOS di ingresso due opportune reti di transistori PMOS e NMOS che realizzano la funzione voluta, analogamente al caso delle porte logiche CMOS. Inoltre il NMOS N_2 è anch'esso sostituito da una rete di NMOS che realizza la stessa funzione, per cui all'ingresso di Q_a si fornisce l'uscita negata ri-

spetto a quella fornita a Q_b (si ricorda che il pilotaggio di Q_a viene fornito dall'uscita presa *tra la rete NMOS e massa*).

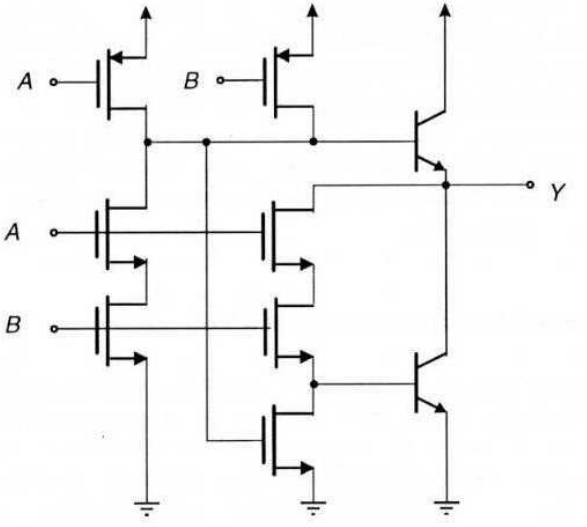


Figura 10.29 Porta logica NAND BiCMOS a due ingressi

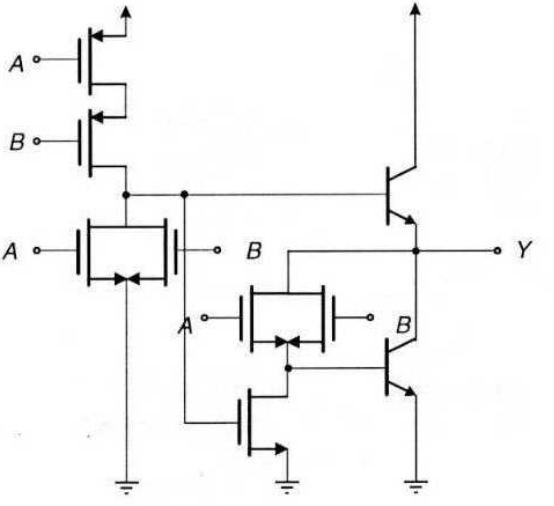


Figura 10.30 Porta logica NOR BiCMOS a due ingressi

Ad esempio, in Figura 10.29 è riportato lo schema di una porta NAND BiCMOS a due ingressi; questa viene realizzata sostituendo una porta NAND CMOS al posto dell'invertitore CMOS di ingresso, ed una rete NMOS, che realizza la funzione AND in uscita, al posto di N_2. Infine in Figura 10.30 è riportato lo schema di una porta NOR BiCMOS a due ingressi, sostituendo in questo caso una porta NOR CMOS all'invertitore CMOS, ed una rete NMOS che realizza la funzione OR in uscita al posto del transistore N_2.

Si può estendere questa modifica al fine di realizzare porte complesse BiCMOS, in analogia a quanto detto per le porte complesse CMOS; occorre inserire le porta complessa CMOS al posto dell'invertitore e una rete NMOS, che realizza la funzione negata, al posto di N_2.

10.8 Circuiti sommatori e comparatori

I circuiti logici possono essere utilizzati per effettuare operazioni aritmetiche tra numeri binari. Per l'operazione base, cioè la somma di due numeri binari, poiché questi possono assumere solo i valori 0 e 1, la somma consiste nel considerare, per ognuna delle posizioni nella sequenza di bit dei due numeri binari, la somma dei due addendi come indicato in Figura 10.31, tenendo conto di un "riporto" di 1 se i due addendi sono entrambi 1. Basicamente questo comporta un circuito che effettui l'operazione di OR esclusivo (XOR) tra due ingressi, in quanto la presenza di due 1 agli ingressi deve fornire uno 0 in uscita, con l'indicazione aggiuntiva di un riporto da sommare ai bit della posizione successiva.

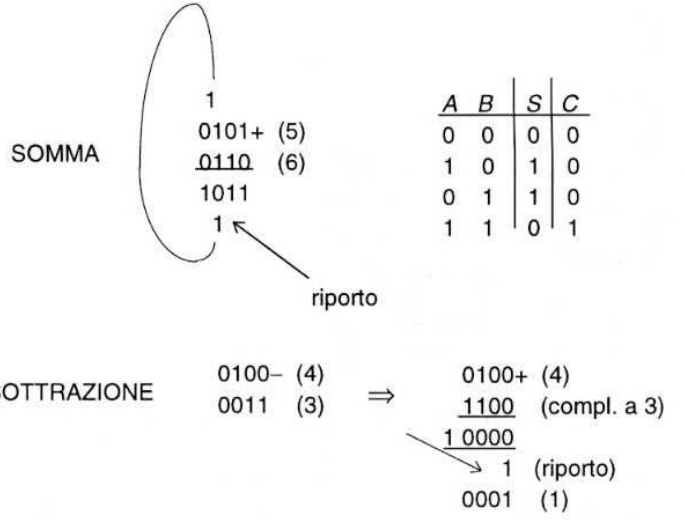


Figura 10.31 Somma e sottrazione di numeri binari

L'operazione di sottrazione può essere ricondotta ad un'operazione di somma effettuando il complemento a 1 del sottraendo e sommando ulteriormente un 1 alla cifra meno significativa, come è mostrato in Figura 10.31. Anche l'operazione di moltiplicazione si può riportare ad una sequenza di somme tra il moltiplicando ed i prodotti parziali di questo per ognuna delle cifre del moltiplicatore. Ognuno di questi prodotti va spostato di una posizione rispetto al precedente; i prodotti parziali sono poi o uguali al moltiplicando o pari a 0 secondo le relazioni $A \cdot 1 = A$, $A \cdot 0 = 0$. Infine l'operazione di divisione può essere ricondotta ad una serie di sottrazioni successive del dividendo a partire dalle cifre più significative, utilizzando il divisore se questo è inferiore al dividendo residuo o 0 se questo è superiore.

Dalla tabella della verità di Figura 10.31 si desume che il risultato della somma S tra due bit dei due numeri binari è realizzata dalla porta XOR, già richiamata nel Paragrafo 1.5, che effettua l'operazione di OR esclusivo tra due variabili A e B; a questa occorre aggiungere l'operazione che fornisce il riporto C che è effettuata da una funzione AND tra A e B. Quindi la somma completa di due numeri ad un bit è effettuata dalle due espressioni:

$$S = A \oplus B \quad \text{(somma)} \qquad C = A \cdot B \quad \text{(riporto)} \tag{10.18}$$

La porta XOR è quindi alla base dei circuiti che effettuano operazioni aritmetiche tra numeri binari. L'operazione logica alla base di una porta XOR è descritta dall'espressione logica: l'uscita è alta se è alto (*A OR B*) *AND NOT* (*A AND B*). Questa espressione è descritta dall'equazione logica:

$$Y = (A + B) \cdot \overline{A \cdot B} \equiv \overline{\overline{A + B} + A \cdot B} \equiv A \cdot \overline{B} + B \cdot \overline{A} \equiv \overline{A + \overline{B}} + \overline{B + \overline{A}} \tag{10.19}$$

e dalle sue equivalenti secondo i teoremi di De Morgan.

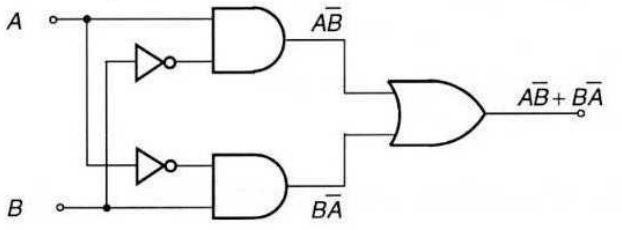


Figura 10.32 Realizzazione di una porta OR Esclusivo (XOR) mediante porte AND ed OR

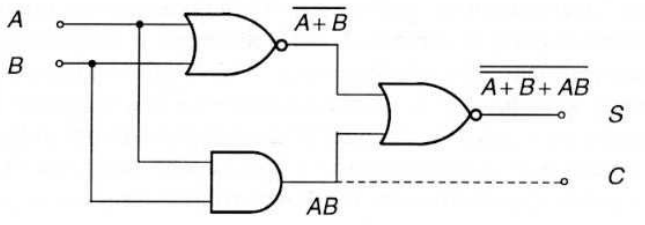


Figura 10.33 Realizzazione di una porta XOR mediante porte AND e NOR

Queste ulteriori espressioni suggeriscono alcune realizzazioni della porta XOR in base alle porte elementari introdotte precedentemente; ad esempio la seconda funzione viene implementata con due porte NOR ed una porta AND (Figura 10.33), mentre la terza espressione viene implementata con una configurazione AND-OR (Figura 10.32) che può utilizzare le porte A-O-I viste precedentemente.

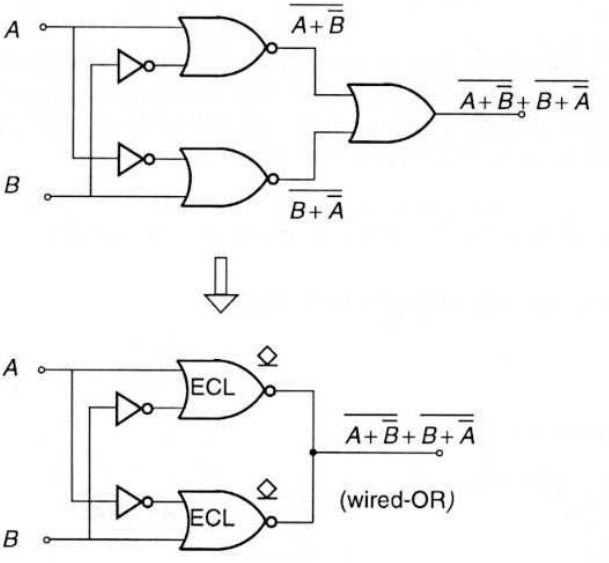
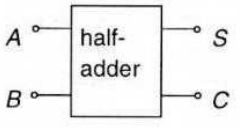


Figura 10.34 Realizzazione di una porta XOR mediante porte NOR ECL e logica cablata

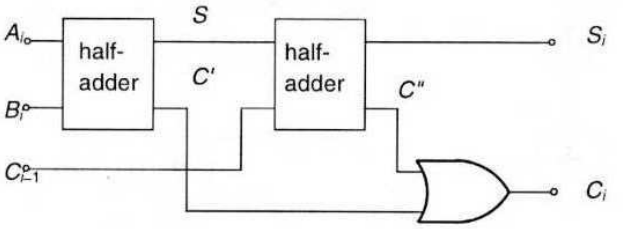
Un'ulteriore possibilità di realizzazione viene dall'ultima relazione che può essere implementata con porte NOR in versione ECL, per cui l'ulteriore porta OR richiesta in uscita può essere sostituita da una logica cablata sulle due uscite NOR connettendo come si è visto gli emettitori in comune e riducendo quindi la complessità della porta XOR (Figura 10.34).

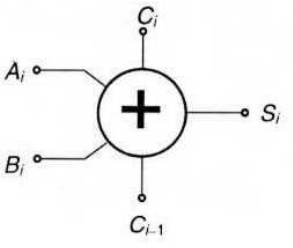
Nella realizzazione di Figura 10.33 si può utilizzare l'uscita della porta AND per ottenere contemporaneamente la funzione $A \cdot B$, e cioè il riporto di una somma di due numeri da un bit. In tal caso il circuito che ne risulta, formato da una porta XOR con un'uscita aggiuntiva viene detto *semi-addizionatore* (*half-adder*) e presenta due ingressi a cui sono inviati i due bit A_i e B_i da sommare, e due uscite, una per la somma S_i ed una per il riporto C_i, descritte dalle espressioni riportate nella (10.18).



A	B	S	C
0	0	0	0
1	0	1	0
0	1	1	0
1	1	0	1

Figura 10.35 Schema logico e tabella della verità di un semi-addizionatore





C_{i-1}	A_i	B_i	S_i	C_i
0	0	0	0	0
0	1	0	1	0
0	0	1	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	0	1

Figura 10.36 Schema logico e tabella della verità di un addizionatore completo

Un circuito *addizionatore completo* (*full-adder*) prevede di sommare ai singoli bit A_i e B_i anche il riporto C_{i-1} che può provenire dalla somma dei bit precedenti di due numeri a più bit, e si ottiene combinando due semi-addizionatori (di qui il nome di questi ultimi) come indicato nello schema logico di Figura 10.36; le uscite soddisfano alle seguenti espressioni, valide per una somma di due bit considerando anche il riporto di bit precedenti:

$$S_i = C_{i-1} \oplus A_i \oplus B_i \qquad C_i = A_i \cdot B_i + (A_i \oplus B_i) \cdot C_{i-1} \tag{10.20}$$

come si può verificare dalla tavola della verità corrispondente o dallo schema logico basato sugli half-adder.

Combinando più full-adders secondo lo schema logico di Figura 10.37a si può realizzare un addizionatore a n bit, detto "addizionatore a propagazione del riporto" (*ripple carry adder*). In questo schema, i riporti degli addizionatori dei bit meno significativi sono riportati in cascata agli addizionatori successivi, mentre i termini somma sono ottenuti in parallelo da ogni addizionatore elementare; la propagazione del riporto è quindi rallentata nel caso peggiore da un tempo di propagazione che è $n\ t_{PA}$ dove t_{PA} è il tempo di propagazione del segnale di riporto nel singolo full-adder. Vi sono molte altre architetture di addizionatori che permettono di velocizzare l'operazione di somma rispetto a quella elementare qui presentata, e si rimanda chi volesse approfondire l'argomento ai testi riportati in bibliografia.

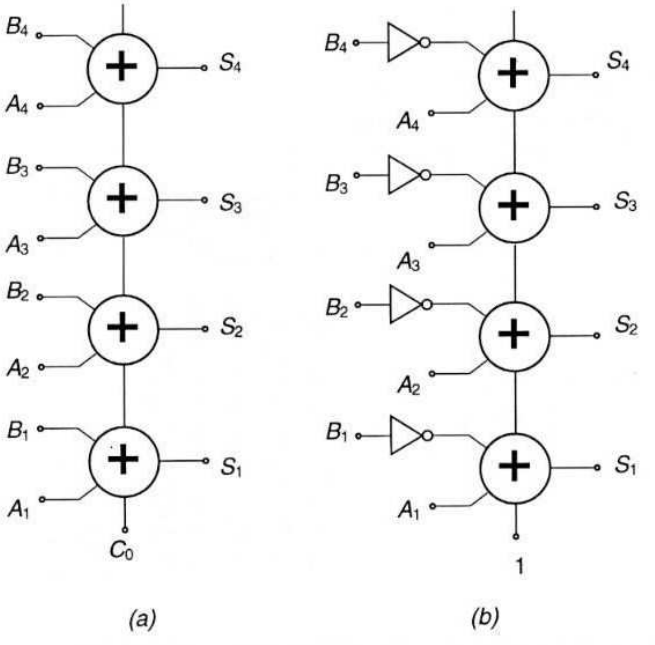


Figura 10.37 Schema logico di a) un addizionatore e b) un sottrattore a n bit

Un sottrattore a n bit può facilmente essere realizzato modificando l'addizionatore a n bit; ricordando che l'operazione di sottrazione di due numeri binari A–B si effettua come un'addizione tra il minuendo A ed il complemento a 1 del sottraendo B, ed aggiungendo 1 alla somma, come indicato in Figura 10.31, si ottiene lo schema di Figura 10.37b, in cui il complemento a 1 si realizza inserendo un inver-

titore su ogni ingresso dei bit del minuendo B, e l'aggiunta di 1 si realizza attraverso l'ingresso del riporto *C*.

Un'ulteriore applicazione delle porte XOR (XNOR) è quella nei circuiti comparatori, cioè circuiti che verificano l'uguaglianza di due numeri (o parole) binari, in altre parole verificano posizione per posizione la corrispondenza dei rispettivi bit nelle due parole da confrontare. Questa operazione viene fatta sul singolo bit dalla porta XOR (XNOR) che, in base alla sua tabella della verità, verifica la coincidenza di due 1 o di due 0 con un'uscita bassa (alta), mentre l'uscita è alta (bassa) se uno dei due bit è differente dall'altro.

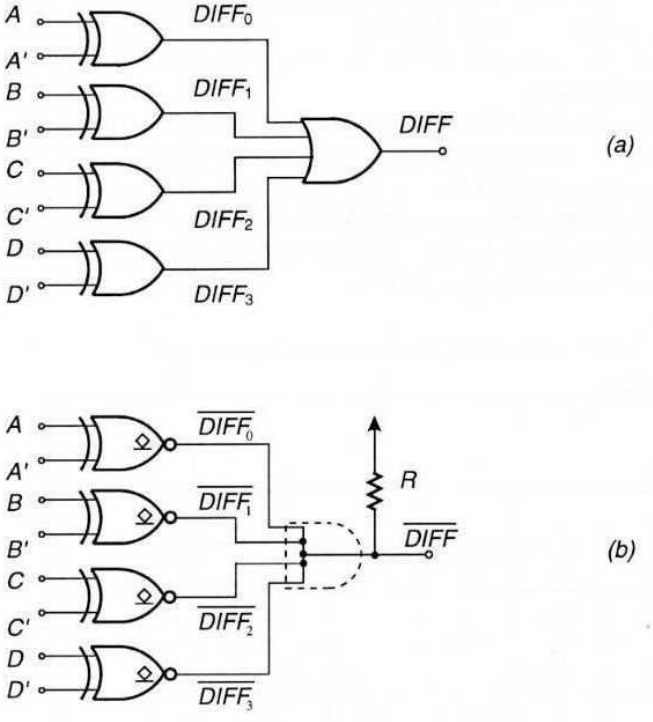


Figura 10.38 Circuito comparatore a 4 bit: a) con porte XOR ed OR; b) con porte XNOR open-collector ed AND cablato

Un circuito di principio per un comparatore di numeri o parole a 4 bit è riportato in Figura 10.38a, ed utilizza 4 porte XOR ed una NOR a 4 ingressi; in questo circuito l'uscita DIFF sarà alta se le due parole non coincidono, bassa se coincidono. Una versione più compatta prevede l'uso di porte XNOR in versione open-collector, in modo da realizzare il secondo livello di logica in AND-cablato (wired-AND) ed ottenere quindi un'uscita alta quando le due parole

sono uguali; in tal caso la degradazione nella dinamica legata all'uso di una resistenza di carico per l'uscita viene compensata da un risparmio in tempi di propagazione dovuto all'eliminazione di un livello di logica.

10.9 Circuiti codificatori e decodificatori

Un circuito *decodificatore* (*decoder*) è un circuito logico che seleziona una particolare uscita in funzione di un numero o parola binaria in ingresso, secondo una determinata legge di attribuzione, o *decodifica*. Il circuito presenta quindi tanti ingressi quanti sono i bit della parola binaria da decodificare, e tante uscite quanti sono i valori differenti assunti in base al codice scelto. Normalmente gli ingressi sono inferiori alle uscite, in quanto le combinazioni (codifiche) dei bit nella parola binaria sono più elevate del numero di bit presenti nella parola stessa.

La codifica più usuale è quella del codice binario, dove ogni parola di n bit rappresenta uno dei numeri naturali da 0 a $2^n - 1$. Un ulteriore codice molto usato è quello che codifica i numeri binari con 4 bit in modo da rappresentare le dieci cifre decimali da 0 a 9; in questo caso, poiché $2^4 = 16 > 10$, vi sono delle parole che non vengono utilizzate, e vi sono più codici per la corrispondenza tra i numeri binari e le cifre decimali, come il codice BDC.

Tabella 10.2 Tabella della verità per un decodificatore da 2 bit

ingressi			*uscite*			
En	A_0	A_1	Y_0	Y_1	Y_2	Y_3
0	x	x	0	0	0	0
1	0	0	1	0	0	0
1	1	0	0	1	0	0
1	0	1	0	0	1	0
1	1	1	0	0	0	1

I decodificatori che utilizzano il codice binario sono detti decodificatori binari; essi presentano n ingressi e 2^n uscite, ed operano in modo da portare una sola delle 2^n uscite al valore alto (o basso) in funzione della parola binaria presentata agli n ingressi. Di solito vi è un ulteriore ingresso per un segnale di abilitazione che con la sua presenza abilita l'operazione di decodifica, analogamente a quanto visto nel caso delle porte a tre stati; questo ingresso di abilitazione è previsto per la maggior parte dei circuiti combinatori. In Tabella 10.2 è riportata la tabella della verità per un decodificatore binario con 2 bit in ingresso.

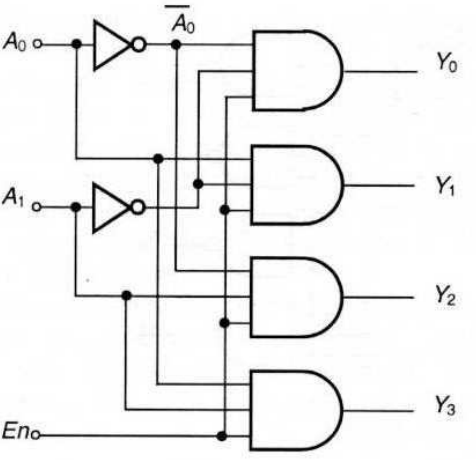


Figura 10.39 Schema logico di un decodificatore a 4 uscite (2-4)

L'operazione di decodifica viene effettuata in maniera semplice se sono disponibili sia i valori delle variabili (i bit della parola) ai singoli ingressi, che i loro valori negati (i complementi dei bit della parola). Lo schema logico per un decodificatore da 2 a 4 (parole da 2 bit in ingresso, $2^2 = 4$ linee da selezionare in uscita) è riportato in Figura 10.39; dalla tabella della verità corrispondente all'operazione voluta per il circuito, si possono scrivere le seguenti equazioni logiche per le uscite:

$$Y_0 = \overline{A_0} \cdot \overline{A_1}; \quad Y_1 = A_0 \cdot \overline{A_1}; \quad Y_2 = \overline{A_0} \cdot A_1; \quad Y_3 = A_0 \cdot A_1 \qquad (10.21)$$

Queste operazioni sono realizzate da porte AND ai cui ingressi sono inviate le variabili A_0, A_1 o i loro complementi in accordo alla (10.21). L'abilitazione è effettuata inviando il segnale *En* a tutte le porte, in quanto per le porte AND basta che uno degli ingressi sia basso per inibire l'uscita, ossia renderla bassa. In definitiva per decodificare parole da n bit occorrono 2^n porte AND con $n + 1$ ingressi ciascuna.

Gli ingressi con le variabili negate vengono ottenuti dalle singole variabili mediante invertitori, che agiscono anche da stadi separatori (buffer); anche sugli ingressi con le variabili non negate vengono aggiunti degli stadi buffer (realizzati con due invertitori) perché il carico in termini di porte per ciascun ingresso negato, o non negato, è relativamente elevato (corrisponde infatti a $2^n/2$ porte per decodificatori di parole da n bit), e può eccedere il fan-out dei circuiti che forniscono queste variabili al decodificatore; ciò vale a maggior ragione per l'ingresso di abilitazione che deve essere inviato a tutte le porte (2^n). Inoltre si preferisce utilizzare porte NAND invece che AND perché le prime, come abbiamo visto, sono più facili da realizzare e presentano minore ritardo di quelle non invertenti, in tal caso la linea identificata per l'uscita è quella al livello basso, mentre tutte le altre restano al livello alto; il circuito viene quindi realizzato secondo lo schema logico di Figura 10.40a.

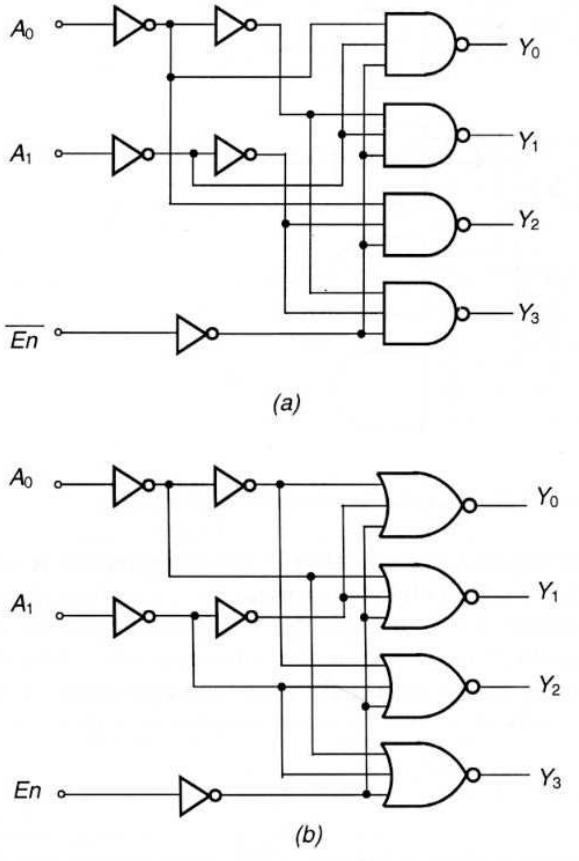


Figura 10.40 Schema logico di decodificatore 2-4: a) con porte NAND; b) con porte NOR

D'altra parte è possibile trasformare le relazioni logiche della (10.21) in modo da descrivere le uscite in termini di operazioni NOR, utilizzando ancora la possibilità, offerta dal circuito di ingresso, attraverso gli stadi invertitori che permettono di avere a disposizione sia le variabili che i loro valori negati, in base alle relazioni seguenti:

$$\begin{aligned} Y_0 &= \overline{A_0} \cdot \overline{A_1} = \overline{A_0 + A_1} \\ Y_1 &= A_0 \cdot \overline{A_1} = \overline{\overline{A_0} + A_1} \\ Y_2 &= \overline{A_0} \cdot A_1 = \overline{A_0 + \overline{A_1}} \\ Y_3 &= A_0 \cdot A_1 = \overline{\overline{A_0} + \overline{A_1}} \end{aligned} \qquad (10.22)$$

e quindi si può realizzare il decodificatore mediante porte NOR, come riportato in Figura 10.40b.

La scelta tra le due possibili configurazioni, con porte NAND o NOR, viene fatta in base alla maggior convenienza di realizzazione compatta delle porte, a seconda della tecnologia utilizzata. In ogni caso, visto l'elevato numero di porte elementari necessarie per questi circuiti (un semplice decodificatore di parole da 4 bit richiede 16 porte elementari più 9 invertitori), occorre semplificare al massimo la realizzazione della singola funzione logica NAND o NOR in modo da ridurre le dimensioni del circuito stesso.

La riduzione dei margini di rumore e di pilotaggio delle singole porte in questo caso non è vincolante, in quanto il circuito che realizza la funzione combinatoria è collegato all'esterno mediante degli stadi buffer sia in ingresso che in uscita, necessari sia per invertire le variabili di ingresso o di uscita, che per pilotare i carichi capacitivi elevati legati all'elevato numero di porte da pilotare; questi stadi buffer svolgono anche la funzione di ripristino dei livelli logici dei segnali e permettono di semplificare la parte interna del circuito combinatorio. La presenza di stadi buffer è una caratteristica comune a tutti i circuiti logici MSI e LSI, dove la parte interna del circuito (inaccessibile dall'esterno) può essere progettata con specifiche più ridotte in termini di flessibilità della singola porta, in quanto questa va utilizzata in un contesto specifico definito dal progettista, mentre le condizioni di flessibilità per la connessione con altri circuiti vengono assicurate da circuiti aggiuntivi di ingresso ed uscita realizzati mediante stadi di separazione. Utilizzando invertitori a tre stati per gli stadi buffer di ingresso e di uscita si può inoltre implementare la funzione di abilitazione per l'intero circuito.

La versione dei decodificatori in tecnologia MOS utilizza circuiti NMOS invece che CMOS per risparmiare area, e di preferenza impiega porte NOR per le ragioni indicate nel Paragrafo 4.11 (le porte NAND a molti ingressi presentano inoltre un effetto body significativo per i MOS in serie più distanti dalla massa, per cui le caratteristiche dinamiche vengono ulteriormente peggiorate).

Uno schema circuitale per un decodificatore NMOS da 3 a 8 è riportato in Figura 10.41; il carico è costituito da un NMOS a svuotamento che può essere dimensionato con un rapporto K_R di circa 2 in modo da avere tempi di propagazione non molto diversi, potendo accettare valori di V_{OL} relativamente elevati (fino a 0.5 V), visto che il pilotaggio avviene attraverso stadi buffer. Le singole celle NOT collegate ad ogni linea di uscita vengono pilotate dai singoli bit della parola di indirizzo o dai bit negati, a seconda dell'espressione logica per quella linea, analogamente a quanto indicato nella (10.22), in modo che l'operazione NOR venga effettuata dalla connessione in parallelo dei MOS sulla singola linea di uscita. In tal modo la struttura del decodificatore assume l'aspetto di una matrice rettangolare, in cui le linee dei bit di indirizzo pilotano gli ingressi dei MOS e le linee di uscita connettono i drain. In generale, per un decodificatore a n bit, il circuito utilizza $n \cdot 2^n$ transistori; il consumo di potenza è relativamente elevato, poiché in ogni condizione di funzionamento vi è dissipazione di potenza per ognuna delle 2^n-1 linee (cioè porte NOR) che sono al livello logico basso. Si può anche implementare, per ciascuna delle uscite, la

funzione negata rispetto a quella indicata nelle relazioni (10.22), rendendo bassa la linea indirizzata e mantenendo tutte le altre alte, ma occorre sempre inserire degli invertitori e quindi in questo caso saranno gli 2^n-1 invertitori pilotati dal livello alto che dissiperanno potenza.

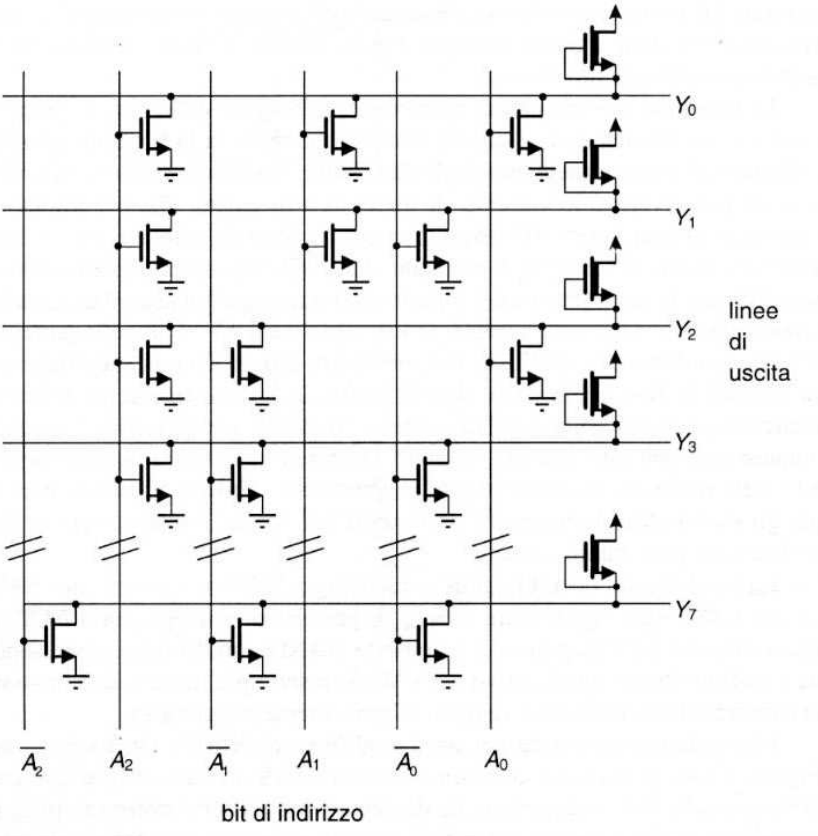


Figura 10.41 Schema circuitale di un decodificatore 3-8 con tecnologia NMOS e porte NOR

Una versione di decodificatore in tecnologia bipolare utilizza per le singole celle logiche della matrice dei semplici diodi Schottky. Questi effettuano l'operazione AND tra le variabili di ingresso (analogamente al caso degli ingressi di porte DTL, vedi Figura 7.16). Il decodificatore assume ancora una struttura a matrice rettangolare come indicato in Figura 10.42; l'uscita si porta al valore di conduzione dei diodi Schottky V_{SC} = 0.5 V se almeno una linea di ingresso è a valore basso, mentre si porta al valore V^+ se tutte le linee di ingresso sono al valore alto V^+.

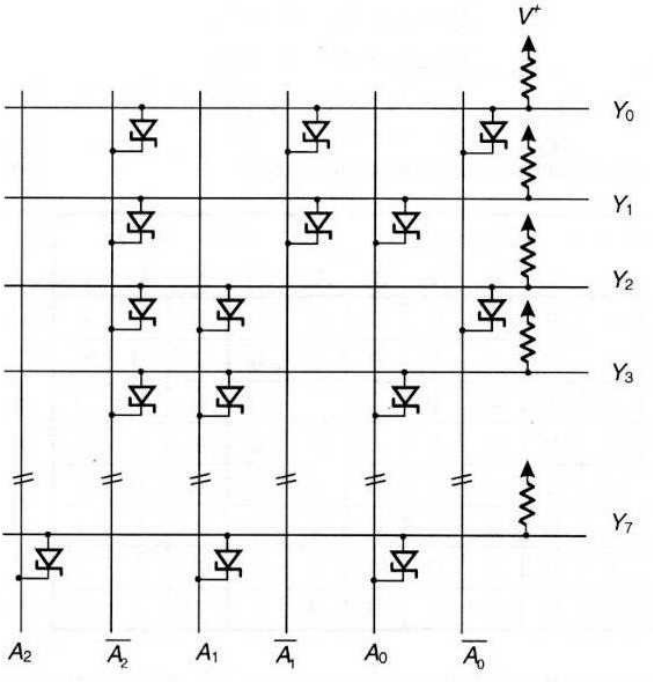


Figura 10.42 Schema circuitale di un decodificatore 3-8 con tecnologia bipolare e porte elementari AND

Il circuito *codificatore* (*encoder*) è un circuito logico che associa (o codifica) una determinata parola di n bit abilitando uno degli m ingressi disponibili. L'operazione di codifica può essere vista come l'inverso di quella di decodifica; i circuiti codificatori hanno quindi più ingressi che uscite in quanto la parola codificata ha un numero di bit inferiore al numero di combinazioni possibili. Il più usuale e semplice circuito decodificatore è quello che associa ad ognuno di 2^n ingressi considerato come numero progressivo, la codifica binaria di questo numero con n bit: tale codificatore viene definito codificatore binario o codificatore $2^n/n$; in tal caso il codificatore avrà 2^n ingressi e n uscite. Un altro usuale codificatore è quello che associa i numeri decimali da 0 a 9 con il corrispondente codice binario a 4 bit; in questo caso 4 delle combinazioni possibili in uscita non verranno utilizzate in quanto $2^4 = 16 > 10$.

In Tabella 10.3 è riportata la tabella della verità per un codificatore binario a 8 bit in ingresso e 3 bit in uscita, indicato usualmente come codificatore 8-3. In base a questa tabella le relazioni logiche per le uscite sono:

$$
\begin{aligned}
Y_0 &= A_1 + A_3 + A_5 + A_7 \\
Y_1 &= A_2 + A_3 + A_6 + A_7 \\
Y_2 &= A_4 + A_5 + A_6 + A_7
\end{aligned}
\tag{10.23}
$$

Tabella 10.3 Tabella della verità per il codificatore 8-3

ingressi								*uscite*		
A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7	Y_0	Y_1	Y_2
1	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0
0	0	1	0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0	1	1	0
0	0	0	0	1	0	0	0	0	0	1
0	0	0	0	0	1	0	0	1	0	1
0	0	0	0	0	0	1	0	0	1	1
0	0	0	0	0	0	0	1	1	1	1

Queste relazioni sono implementate con porte OR nello schema logico di Figura 10.43; non sono riportati nello schema per semplicità gli stadi separatori (buffer) che sono di norma necessari quando i singoli ingressi sono collegati a più porte, per evitare di caricare troppo in termini di fan-out i circuiti di pilotaggio.

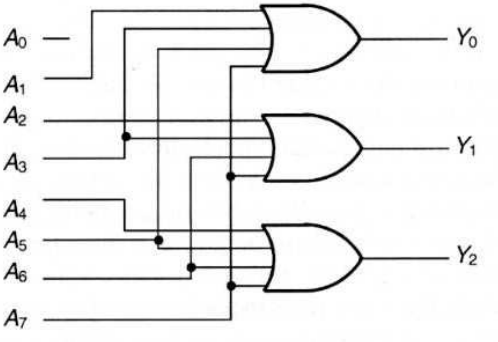


Figura 10.43 Schema logico di un codificatore 8-3

Le realizzazioni circuitali dei codificatori sono di solito in forma di matrice rettangolare, come nei casi già visti di circuiti decodificatori, di cui condividono gran parte delle soluzioni circuitali utilizzate per le differenti tecnologie di realizzazione.

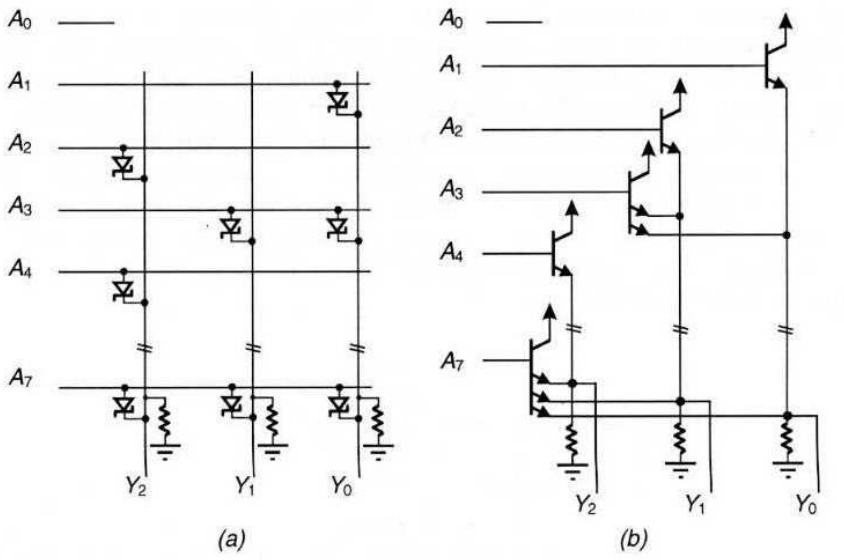


Figura 10.44 a) Matrice a diodi per un codificatore 8-3; b) versione con transistori a multiemettitore

In tecnologia MOS la struttura è equivalente a quella riportata in Figura 10.41 che è infatti realizzata con porte NOR; in tecnologia bipolare si può adottare una matrice a diodi (schema di Figura 10.44a) o quella con transistori a multiemettitore (Figura 10.44b). Quest'ultima versione è più compatta ed utilizza la funzione OR-cablato che si realizza connettendo tra loro gli emettitori di transistori connessi a collettore comune analogamente a quanto visto per le uscite delle porte ECL; i componenti sono quelli impiegati in tecnologia TTL, e la stessa tecnologia può essere utilizzata per gli stadi di disaccoppiamento di ingresso ed uscita. Inoltre l'uso di transistori nel montaggio a collettore comune permette di realizzare un guadagno di corrente in uscita ed una riduzione della resistenza interna vista dall'emettitore, favorendo quindi il pilotaggio di carichi capacitivi elevati che sono una condizione ineliminabile in questi circuiti a matrice.

10.10 Circuiti multiplexer e demultiplexer

I circuiti *demultiplexer* sono basicamente dei commutatori digitali, in altre parole essi selezionano una delle n linee in uscita ed inviano i dati (forniti in maniera seriale in ingresso, ossia inviati come sequenza temporale di bit) alla linea selezionata attraverso una parola di indirizzo.

Si comprende quindi come il nucleo base di un demultiplexer sia un circuito decodificatore, che effettua l'operazione di selezione di una tra n linee in uscita mediante una parola di indirizzo; occorre a questo aggiungere un ingresso a cui è

inviata la sequenza di bit (ossia il *dato* in forma binaria) che deve essere instradato sulla linea abilitata. In realtà quest'operazione è automaticamente effettuata dall'ingresso di abilitazione; infatti, inviando il dato come sequenza di bit binari 1 o 0 a quest'ingresso, se il bit è 1 l'abilitazione è alta e quindi la linea abilitata presenta un 1, mentre se il bit è 0 la linea selezionata (insieme a tutte le altre, ma questo è inessenziale) è bassa e presenta quindi uno 0; quindi la sequenza di 1 e 0 all'ingresso di abilitazione è replicata sull'uscita selezionata. Per mantenere anche la funzione di abilitazione sulla linea selezionata si utilizzano due ingressi di abilitazione di cui uno è utilizzato come ingresso dati e l'altro è quello di abilitazione con ingresso attivo basso (0).

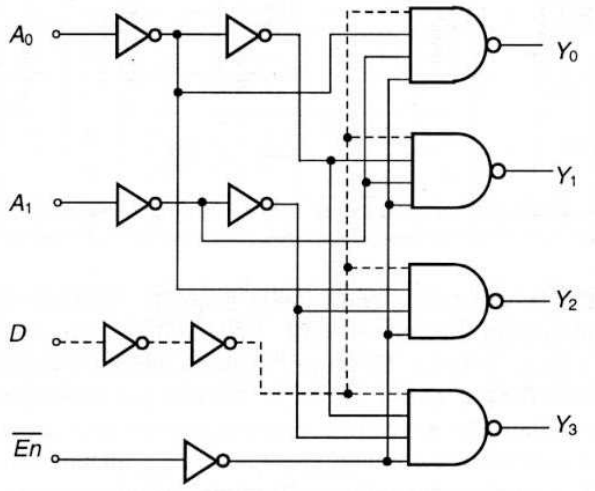


Figura 10.45 Schema logico di un demultiplexer 2-4

Lo schema logico di Figura 10.45 è quello del decodificatore di Figura 10.40, con l'ingresso dati indicato con linea tratteggiata. La realizzazione circuitale segue quindi gli schemi già visti per i circuiti decodificatori, che sono alla base di questo circuito logico. L'inserzione del dato *D* richiede l'aggiunta di un ingresso ulteriore per le porte AND (o analogamente per le versioni NOR) del decodificatore. Sono stati inseriti due invertitori per realizzare un efficiente disaccoppiamento del circuito che fornisce il segnale *D* che deve essere inviato a tutte le porte del circuito.

I circuiti *multiplexer* sono anch'essi dei commutatori digitali, che effettuano la funzione inversa a quella dei circuiti demultiplexer, ossia ricevono dati da *n* linee di ingresso e, attraverso una parola di indirizzo, selezionano tra questi un particolare ingresso e convogliano i dati di quell'ingresso sull'unica linea in uscita. La loro utilizzazione più comune è quella di convogliare su un unico bus di interconnessione i segnali provenienti da diversi reti logiche.

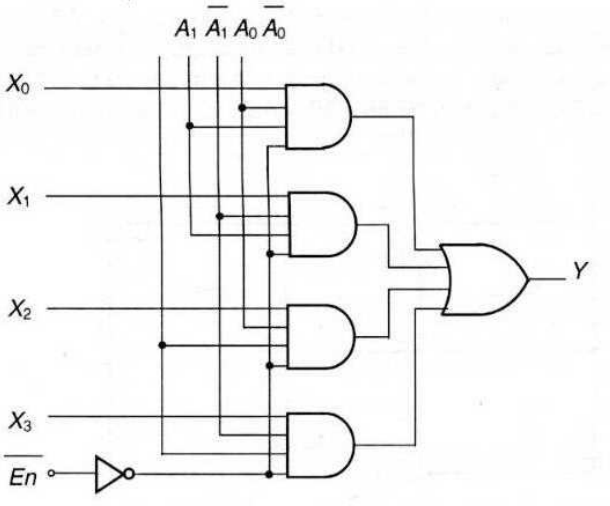


Figura 10.46 Schema logico di un multiplexer con 4 ingressi

Anche in questo caso il nucleo base del circuito è un decodificatore, che in questo caso permette di selezionare una delle porte AND in base alla parola di indirizzo. Le porte AND sono tante quante le *n* linee dati in ingresso ed ognuna delle linee va ad aggiungersi agli ingressi di una delle porte secondo lo schema di Figura 10.46 (non sono stati indicati per semplicità gli stadi invertitori necessari per formare i bit negati della parola di indirizzo). In tal modo solo l'ingresso connesso alla porta abilitata dalla parola di indirizzo viene trasmesso all'uscita corrispondente; basta poi inviare tutte le uscite ad una porta OR per ottenere sull'unica uscita del multiplexer la sequenza di dati presente sulla particolare linea selezionata.

Poiché questi circuiti utilizzano le stesse configurazioni dei circuiti decodificatori da cui discendono non verranno ripetuti gli schemi circuitali di questi ultimi; gli esempi presentati dimostrano ancora una volta come alla base dei circuiti più complessi utilizzati largamente nei sistemi digitali vi sono essenzialmente poche configurazioni base che si differenziano nelle scelte circuitali principalmente in base alla tecnologia dei componenti elementari ed alle caratteristiche elettriche delle porte base delle diverse famiglie logiche utilizzate.

10.11 Componenti Logici Programmabili (PLD)

L'applicazione più versatile e flessibile dei circuiti combinatori è quella dei *Componenti Logici Programmabili* (*Programmable Logic Device, PLD*), che sono essenzialmente dei circuiti logici a struttura regolare integrati in un unico chip, i quali possono essere modificati nelle connessioni interne dall'utente in modo da realizzare una qualsiasi espressione logica relativamente complessa basata su espressioni del tipo somme di prodotti. La versione più generale di questi circuiti è quella delle

Matrici Logiche Programmabili (*Programmable Logic Arrays, PLA*) che è basata su una struttura regolare di celle AND connesse in una matrice, detta appunto matrice (o piano) AND, le cui uscite sono connesse ad una seconda matrice di porte OR, detta matrice (o piano) OR, come è indicato schematicamente in Figura 10.47.

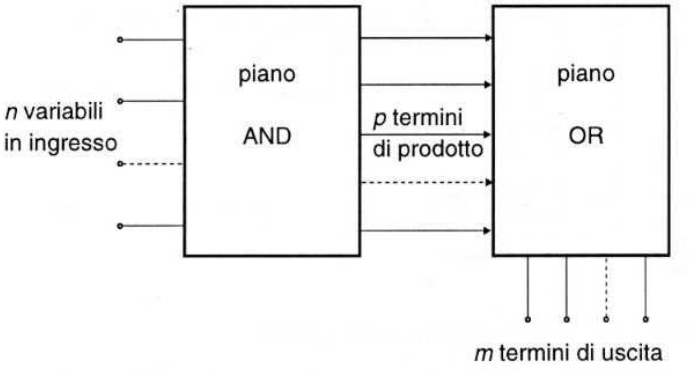


Figura 10.47 Schema a blocchi di un PLA

La matrice di p porte AND, ognuna delle quali presenta n ingressi, può fornire p termini di prodotto ognuno con n variabili; la matrice di m porte OR può fornire m termini di somma tra i p prodotti, per cui in definitiva si possono ottenere m espressioni logiche ognuna con p termini di somma e n prodotti per ogni termine. Le specifiche funzioni logiche che si desidera realizzare vengono implementate nel circuito o eliminando le connessioni non volute rispettivamente nel piano AND ed in quello OR, o inserendo connessioni elettriche tra collegamenti, mediante speciali dispositivi detti rispettivamente fusibili o antifusibili; le operazioni per la programmazione mediante questi collegamenti da inserire o eliminare vengono effettuate dall'utente direttamente sull'integrato. Ad esempio la funzione di fusibile è realizzata utilizzando metalli opportuni con una sezione ridotta nel percorso della corrente; applicando una corrente determinata superiore a quella di esercizio il collegamento si interrompe perché il metallo evapora. La funzione di antifusibile viene realizzata separando due linee di metallo con un sottile strato di dielettrico; se si applica una tensione opportuna il dielettrico viene forato e si crea un contatto con resistenza dell'ordine delle centinaia di ohm.

In Figura 10.48 è riportato lo schema di una piccola PLA di dimensioni $n = 4$, $p = 6$, $m = 3$. In questo schema le x indicano le connessioni possibili delle matrici che possono essere conservate o eliminate; ogni ingresso viene inviato a stadi invertitori che forniscono sia le variabili dirette che quelle negate ai possibili ingressi delle porte AND. Le realizzazioni circuitali delle matrici AND ed OR delle PLA sono analoghe a quelle viste nel Paragrafo 10.9 rispettivamente per i circuiti decodificatori e codificatori, e possono essere realizzate sia in tecnologia bipolare che MOS.

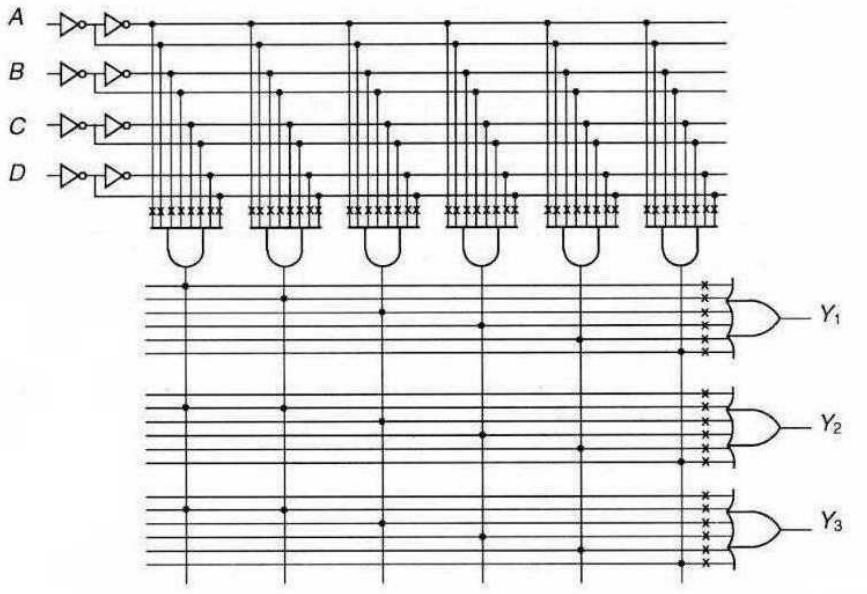


Figura 10.48 Schema logico di una PLA con 4 ingressi, 3 uscite e 6 termini di prodotto

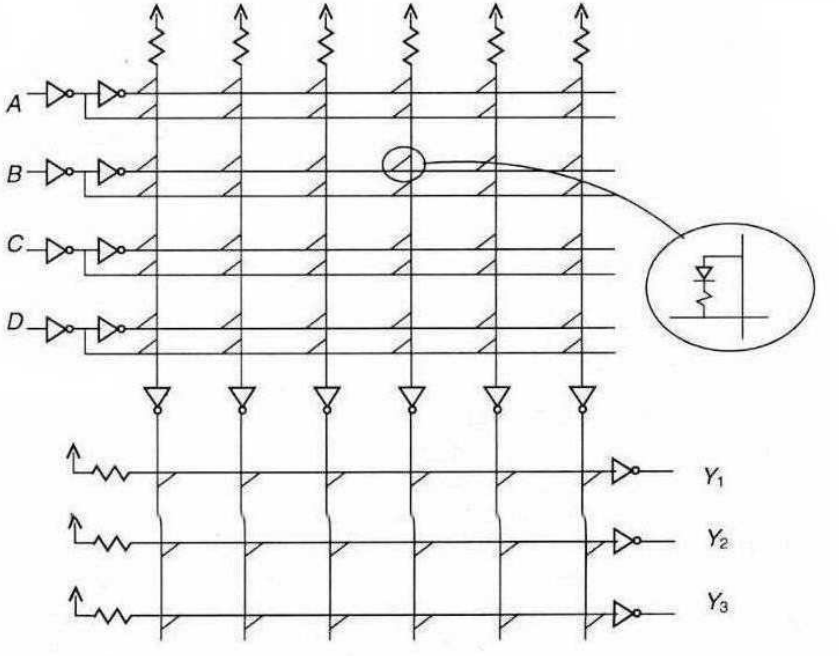


Figura 10.49 Matrici a diodi per PLA a 4 ingressi e 3 uscite

Ad esempio, una versione di una PLA in tecnologia bipolare che utilizza diodi sia nel piano AND che in quello OR è riportata in Figura 10.49; le possibili connessioni sono realizzate mediante diodi connessi in serie con fusibili che possono essere aperti applicando una tensione opportuna alle specifiche righe e colonne della matrice. Le uscite della matrice superiore sono inviate a invertitori, per cui si ottiene una funzione logica complessiva NAND; anche la seconda matrice, che realizza la funzione NAND tra righe e colonne, ha le uscite connesse a stadi invertitori (che effettuano anche la funzione di stadi buffer), per cui la funzione complessiva effettuata dalle due matrici è una funzione NAND-NAND tra le variabili, che equivale ad una funzione AND-OR.

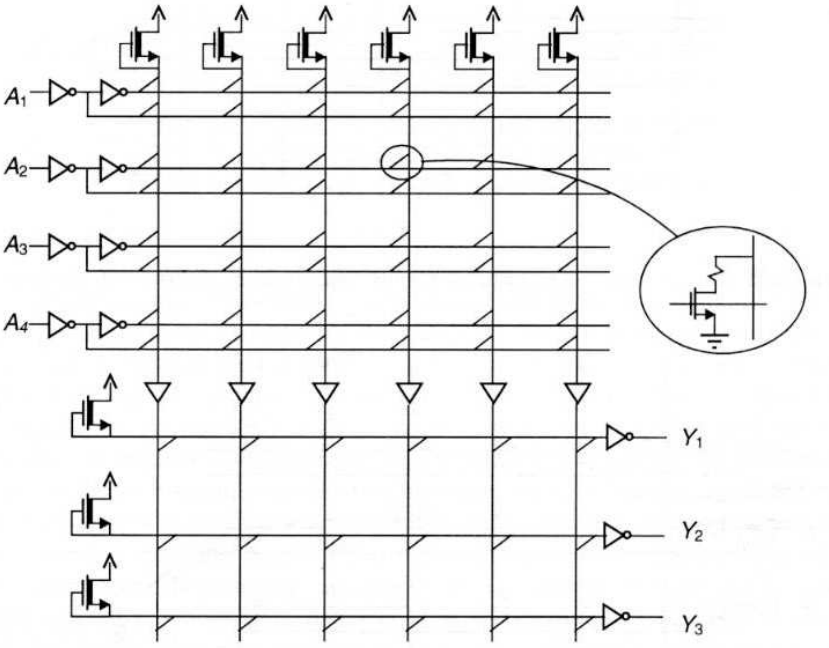


Figura 10.50 Matrici a MOS per PLA a 4 ingressi e 3 uscite

Una versione in tecnologia NMOS delle matrici dei piani AND ed OR di una PLA è riportata in Figura 10.50. In questo caso, analogamente a quanto visto per le matrici di decodifica a MOS, conviene adottare porte elementari NOR, dove le linee degli ingressi pilotano le diverse gate dei MOS e le uscite connettono i drain in parallelo. Il collegamento tra i due piani viene effettuato con stadi buffer non invertenti, per cui la prima matrice, scegliendo opportunamente le variabili nel modo già visto per il decodificatore NMOS di Figura 10.41, equivale alla funzione AND tra gli ingressi e le uscite. Il secondo piano realizza ancora una funzione NOR, ma le uscite sono applicate a stadi invertitori e realizzano la funzione OR; la

funzione logica complessiva è perciò quella AND-OR richiesta. Per queste strutture vi è inoltre la possibilità di effettuare elettricamente da parte dell'utente sia la programmazione che la cancellazione della programmazione effettuata, utilizzando particolari dispositivi su cui ritorneremo nel Capitolo 13.

Gli schemi a matrici delle figure precedenti giustificano la rappresentazione compatta usualmente utilizzata per le PLA, riportata in Figura 10.51 (che corrisponde allo schema esteso di Figura 10.48). In questo caso il simbolo della porta AND indica la funzione complessiva realizzata tra gli ingressi connessi alla singola linea di ingresso a quella porta; lo stesso vale per il simbolo della porta OR per quanto riguarda le uscite AND connesse al singolo ingresso della porta OR.

Nel caso della Figura 10.51, per esemplificare l'interpretazione dello schema logico, le connessioni nei piani AND ed OR realizzano le seguenti funzioni logiche:

$$\begin{aligned} Y_1 &= A_1 \cdot A_2 + \overline{A_1} \cdot \overline{A_2} \cdot A_3 \cdot \overline{A_4} \\ Y_2 &= A_1 \cdot \overline{A_3} \cdot A_4 + \overline{A_1} \cdot A_3 + A_2 \\ Y_3 &= A_1 \cdot A_2 + A_1 \cdot \overline{A_3} \cdot A_4 + \overline{A_1} \cdot \overline{A_2} \cdot \overline{A_4} \end{aligned} \tag{10.24}$$

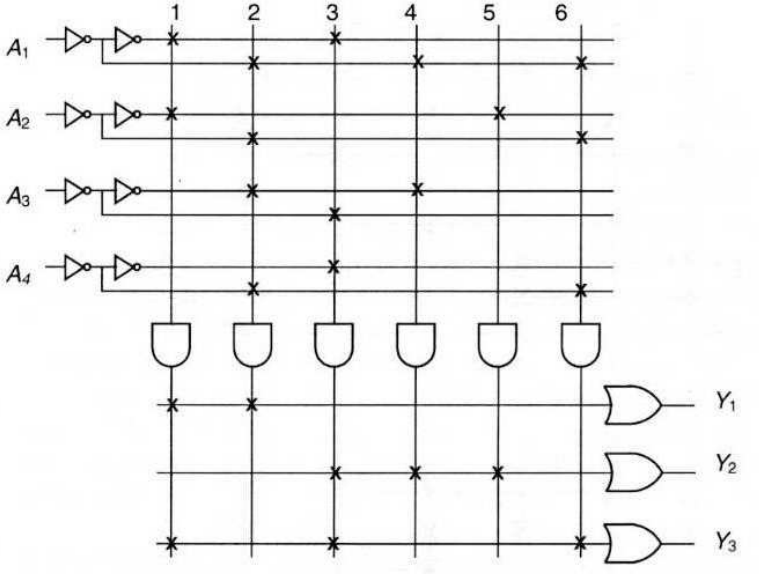


Figura 10.51 Schema logico compatto della PLA di Figura 10.48

Poiché nei PLA non si utilizzano tutte le combinazioni possibili degli ingressi nei due piani, ma solo quelle richieste per implementare le funzioni logiche deside-

rate, si può avere una forte dispersione dei dispositivi (MOS o diodi) nelle rispettive matrici, con conseguente sottoutilizzazione dell'area a disposizione. Ad esempio nella matrice AND del PLA della Figura 10.51 solo 15 delle 48 possibili intersezioni della matrice sono utilizzate (ossia prevedono un dispositivo per la funzione logica da realizzare).

È possibile compattare la matrice effettuando operazioni di taglio lungo le linee degli ingressi in modo da alimentare una parte della linea con la variabile vera ed il resto della linea con quella negata (ricordiamo che ogni porta prevede l'uso di una variabile o del suo negato e non di entrambi), e di riorganizzazione dell'ordine delle linee a valle del taglio; come esempio si è effettuata questa operazione di taglio e riorganizzazione della sequenza di linee per lo schema compatto della matrice AND del PLA di Figura 10.51, giungendo alla matrice più compatta di Figura 10.52. In quest'ultima, si arriva ad un grado di utilizzazione della matrice del 62% rispetto al 31% della matrice di partenza.

I componenti PLA hanno avuto un largo sviluppo nella realizzazione dei sistemi digitali per la loro flessibilità di impiego; ciò vale in particolare per le versioni con tecnologia MOS che permettono la riprogrammazione della funzione logica effettuata mediante operazioni direttamente realizzate dall'utente; ritorneremo su questo argomento nel Capitolo 13 parlando delle memorie ROM programmabili.

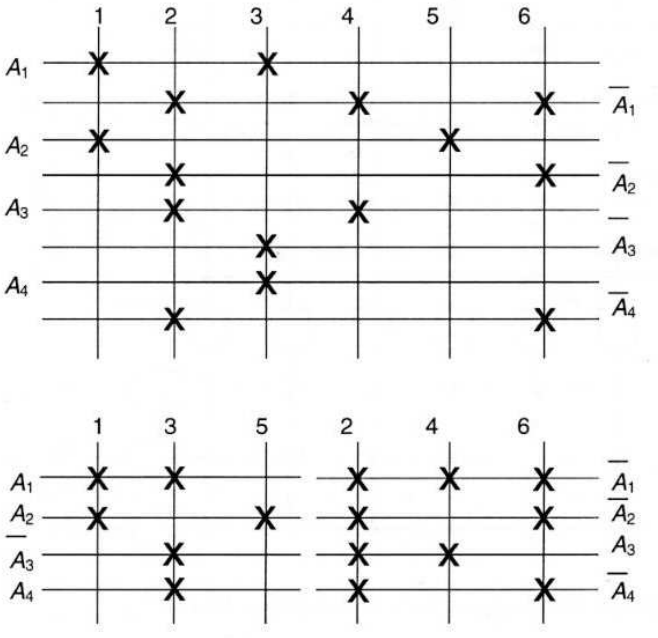


Figura 10.52 Taglio e riorganizzazione della matrice AND del PLA di Figura 10.51

Esercizi di riepilogo

10.1 Valutare il numero di transistori necessari per realizzare una porta A-O-I a 4 ingressi e larghezza 4 con tecnologia NMOS, e confrontare questo numero con quello necessario a realizzare la stessa funzione con porte logiche elementari NAND-NOT-NOR.

10.2 Ripetere il caso dell'Esercizio 10.1 per il caso di porte in tecnologia CMOS.

10.3 Valutare per il circuito di Figura E10.1, costituito da una porta NOR ed una porta NAND in tecnologia CMOS connesse (in maniera non corretta) ad un bus comune in uscita, il valore della tensione V_O al bus di uscita per i seguenti casi: a) $A = C = 1$, $B = D = 0$; b) $A = B = 0$, $C = D = 1$.

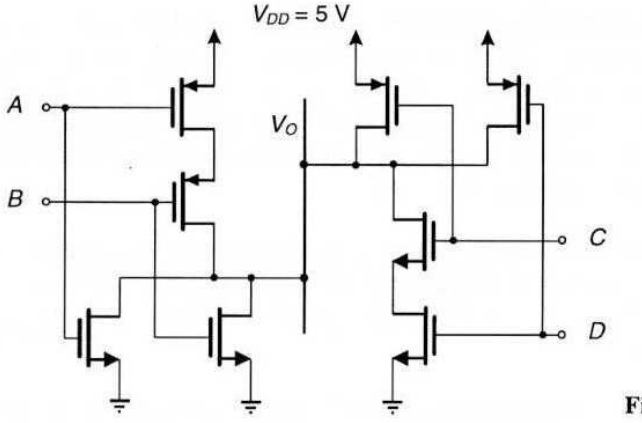


Figura E10.1

10.4 Il circuito logico di Figura E10.2 realizza una funzione AND in logica cablata tra le uscite di due porte NAND "open collector", mentre l'invertitore è in logica TTL. Determinare i valori massimi e minimi che può assumere la resistenza di carico R_L affinché i valori logici siano: $V_{OLMAX} \leq 0.4$ V, $V_{OHMIN} \geq 4.3$ V, assumendo che i transistori NMOS delle porte NAND abbiano un $k' = 4 \cdot 10^{-5}$ A/V^2, $W/L = 2/1$, $V_T = 0.8$ V, e che l'invertitore TTL abbia una corrente di ingresso $I_I = 0.1$ mA per un ingresso basso, e $I_I = -1$mA per un ingresso alto.

10.5 Con riferimento alla connessione in uscita di due porte TTL riportata in Figura 10.8, assumendo $R_T = 120\ \Omega$, $V_{CC} = 5$ V, calcolare: a) il valore della tensione di uscita nel punto comune se la porta 1 presenta un'uscita alta e la porta 2 un'uscita bassa; b) il valore della corrente di collettore del transistore Q_b della porta 1.

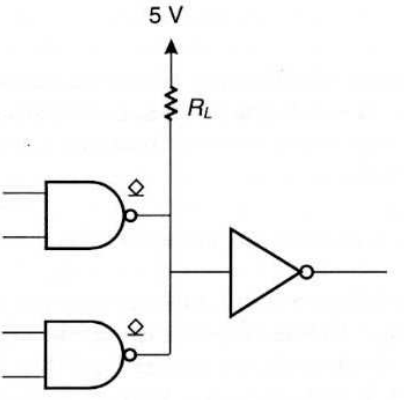


Figura E10.2

10.6 Valutare il tempo di propagazione dell'invertitore CMOS a tre stati di Figura 10.10, supposto caricato da un uguale invertitore, assumendo i seguenti valori dei transistori MOS: $k'_N = 50\ \mu A/V^2$, $k'_P = 20\ \mu A/V^2$, $W_N = 2\ \mu m$, $W_P = 5\ \mu m$, $L_N = L_P = 1\ \mu m$, $V_{TN} = |V_{TP}| = 0.8$ V, $V_{DD} = 5$ V; confrontare il tempo di propagazione determinato con quello presentato dall'invertitore CMOS standard per uguali valori dei transistori.

10.7 Determinare il valore della differenza delle tensioni di soglia $V_{SL}+ - V_{SL}-$ per l'invertitore ad isteresi di Figura 10.15, con i seguenti valori dei transistori: $W/L|_{PA} = 30/2$, $W/L|_{NA} = 4/2$, $W/L|_{NB} = 68/2$, $W/L|_{NC} = 72/2$, $W/L|_{P'} = 10/2$, $W/L|_{N'} = 4/2$.

10.8 Determinare i valori massimi e minimi delle resistenze R_A, R_B, R_C di una rete di interfacciamento tra una porta TTL ed una ECL, che comportano a) una potenza dissipata di 1 mW, e b) un tempo di propagazione aggiuntivo di 1 ns (si assuma la capacità di ingresso della porta ECL pari a 0.1 pF).

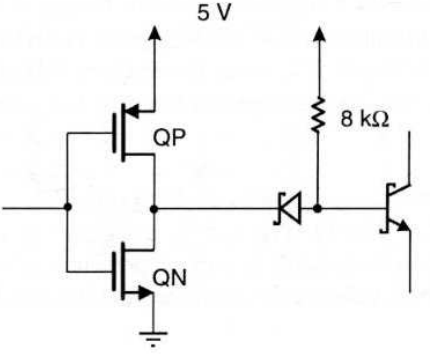


Figura E10.3

10.9 Per il circuito di interfacciamento CMOS/LS-TTL di Figura E10.3, dimensionare i transistori NMOS e PMOS in modo da avere un livello logico basso $V_{OL} = 0.3$ V. Determinare inoltre i valori dei margini di rumore NM_H e NM_L per l'invertitore CMOS così dimensionato.

10.10 Determinare il ritardo di propagazione dell'invertitore BiCMOS di Figura 10.23, caricato da una capacità di 5 pF, per i seguenti valori dei parametri del circuito: $V_{DD} = 5$ V, $\beta_F = 20$, $k_N' = 50\ \mu A/V^2$, $k_P' = 20\ \mu A/V^2$, $W_N = 2\ \mu m$, $L_N = 1\ \mu m$, $W_P = 5\ \mu m$, $L_P = 1\ \mu m$, $V_{TN} = |V_{TP}| = 0.8$ V. Verificare inoltre che il tempo di propagazione dell'uscita V_O' dell'invertitore CMOS sia ben inferiore al tempo di propagazione all'uscita dello stadio bipolare.

10.11 Determinare, mediante simulazioni SPICE, la variazione del tempo di propagazione dell'invertitore BiCMOS al variare della capacità di carico da 1 a 10 pF, assumendo per i dispositivi i valori di dimensionamento indicati nell'Esercizio 10.10 ed utilizzando le schede .MODEL riportate in Appendice.

10.12 Disegnare lo schema elettrico di una porta BiCMOS che realizzi la seguente funzione logica: $Y = \overline{A \cdot B + C}$.

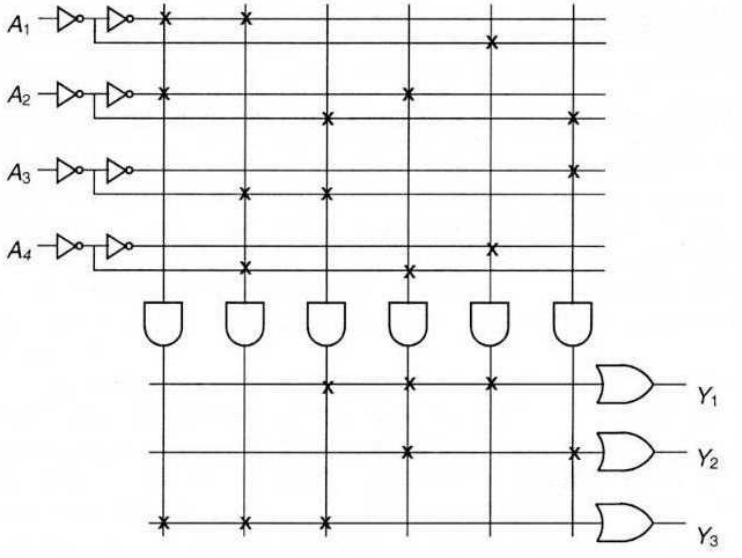


Figura E10.4

10.13 Per il decodificatore NMOS di Figura 10.41, alimentato con $V_{DD} = 5$ V, determinare: a) il tempo di propagazione t_{PLH} della linea indirizzata; b) la potenza dissipata nel decodificatore per qualsiasi combinazione dei bit di in-

gresso. Si assumano gli NMOS di carico con i seguenti parametri: $k'_N = 50$ $\mu A/V^2$, $W_N = 2$ μm, $L_N = 4$ μm, $V_{TD} = -3$ V, e quelli ad arricchimento con: $k'_N = 50$ $\mu A/V^2$, $W_N = 4$ μm, $L_N = 2$ μm, $V_T = 0.8$ V; si utilizzino i valori di Tabella 3.2 per le capacità unitarie dei MOS.

10.14 Ricavare le funzioni logiche realizzate alle uscite Y_1, Y_2, Y_3 del PLA programmato come in Figura E10.4.

10.15 Utilizzare le tecniche di taglio e riorganizzazione delle matrici sul PLA di Figura E10.4 per migliorare il grado di occupazione delle due matrici.

Riferimenti bibliografici

H. Taub, D. Schilling, *Elettronica integrata digitale*, Jackson, Milano, 1981.

G.M. Glansford, *Digital Electronic Circuits*, Prentice Hall, Englewood Cliffs, 1988.

N.E. Weste, K. Eshraghian, *Principles of CMOS VLSI Design, A Systems Perspective,* 2nd ed., Addison-Wesley, Reading, Ma., 1993.

J.F. Wakerly, *Digital Design, Principles and Practices,* 2nd ed., Prentice Hall, Englewood Cliffs, 1994.

J. Millman, A. Grabel, *Microelettronica*, McGraw-Hill Libri Italia, Milano, 1994.

11

Strutture CMOS per circuiti VLSI

11.1 Funzioni logiche complesse con CMOS

Nella realizzazione dei sistemi digitali è possibile sviluppare il progetto utilizzando componenti logici standard, che poi vengono assemblati su schede con circuiti stampati per realizzare i sottosistemi desiderati. È tuttavia più efficace e vantaggiosa la realizzazione degli interi sottosistemi digitali in circuiti a larghissima scala di integrazione (VLSI), che permettono una progettazione specifica dei vari circuiti con significativi vantaggi per quanto riguarda la compattazione dell'area utilizzata, del consumo di potenza e delle prestazioni dinamiche dei circuiti stessi. Nei circuiti VLSI la tecnologia CMOS è quella più utilizzata per le sue caratteristiche di trascurabile dissipazione di potenza statica e di facile integrazione. Con la tecnologia CMOS è possibile, oltre alle porte base NAND o NOR, realizzare qualunque funzione logica più complessa mediante opportune porte logiche che utilizzano una rete di transistori PMOS e una NMOS, costituite da un eguale numero di transistori ciascuna, e tali da dar luogo alla funzione logica richiesta, secondo lo schema generale di Figura 11.1.

Questo schema è stato già utilizzato per la realizzazione delle porte CMOS di tipo A-O-I, che di fatto realizzano già una funzione logica a due livelli. Una generalizzazione del procedimento riportato nel Paragrafo 5.8, per la realizzazione di funzioni logiche più complesse con porte CMOS, prevede di realizzare la rete di transistori NMOS considerando che una funzione OR tra le variabili richiede la connessione in parallelo degli interruttori equivalenti, mentre una funzione AND richiede la loro connessione in serie; una somma di prodotti viene poi realizzata ponendo in parallelo più rami di MOS in serie, e un prodotto di somme richiede di porre in serie i rami di MOS in parallelo. La topologia della rete di transistori PMOS sarà invece duale della prima, in quanto per i transistori PMOS pilotati dalle stesse variabili in ingresso viene scambiata la connessione in parallelo con quella in serie e viceversa.

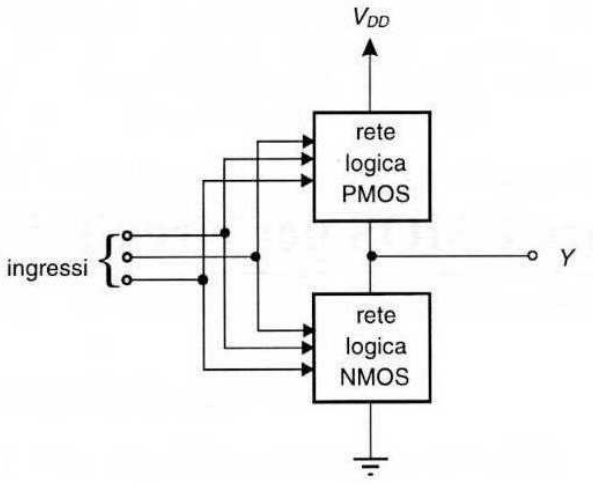


Figura 11.1 Struttura generica di una porta complessa in tecnologia CMOS

Questa topologia duale per la rete PMOS viene giustificata, analogamente al caso delle porte logiche elementari, ricordando che mentre alla rete di NMOS sono applicate le variabili logiche effettive e vi è un'operazione di inversione (NOT) tra le variabili di ingresso e di uscita, il funzionamento dei PMOS come interruttori pilotati è analogo a quello dei NMOS solo se si considerano come grandezze di pilotaggio le variabili negate (e in questo caso non vi è inversione nella funzione logica ottenuta in uscita perché gli interruttori sono connessi tra l'alimentazione e l'uscita). L'uscita deve quindi essere ottenuta sia come funzione logica negata delle variabili che come funzione logica delle variabili negate; nel secondo caso la trasformazione della funzione logica è immediatamente eseguibile in base ai teoremi di De Morgan T11 e T12 della Tabella 1.2, che scambiano il segno di operazione logica con quello di negazione.

Ad esempio l'espressione logica:

$$Y = \overline{A \cdot B + C \cdot (D + E)} = \overline{Y_1 + Y_2 \cdot Y_3} \tag{11.1}$$

che ha tre livelli di logica, e richiede una porta OR, due porte AND e una NOR come indicato in Figura 11.2a, viene realizzata in logica complessa CMOS con la rete di Figura 11.2b (dove per semplicità i transistori PMOS sono stati indicati con un cerchietto sul terminale di gate, ad indicare il comportamento invertito rispetto a quello NMOS). Per questa porta, la rete NMOS è ricavata dalla (11.1) secondo le regole suddette, e cioè ponendo in parallelo due rami che realizzano rispettivamente la funzione Y_1 e quella $Y_2 \cdot Y_3$; a sua volta Y_1 è realizzata ponendo in serie due NMOS, e $Y_2 \cdot Y_3$ ponendo un NMOS in serie al parallelo di altri due NMOS. La rete di PMOS viene definita in base alla trasformazione della (11.1):

$$Y = \overline{A \cdot B + C \cdot (D + E)} = \overline{A \cdot B} \cdot \overline{C \cdot (D + E)} = (\overline{A} + \overline{B}) \cdot (\overline{C} + (\overline{D + E})) = (\overline{A} + \overline{B}) \cdot (\overline{C} + \overline{D} \cdot \overline{E})$$

che esprime l'uscita in base alle variabili negate.

Connettendo ad un'unica uscita le due reti così definite si ottiene in definitiva la funzione desiderata; la porta complessa realizza quindi la funzione logica con un solo stadio CMOS, e cioè con un solo passaggio tra ingressi e uscita, invece che con tre stadi elementari, con vantaggi sul ritardo di propagazione.

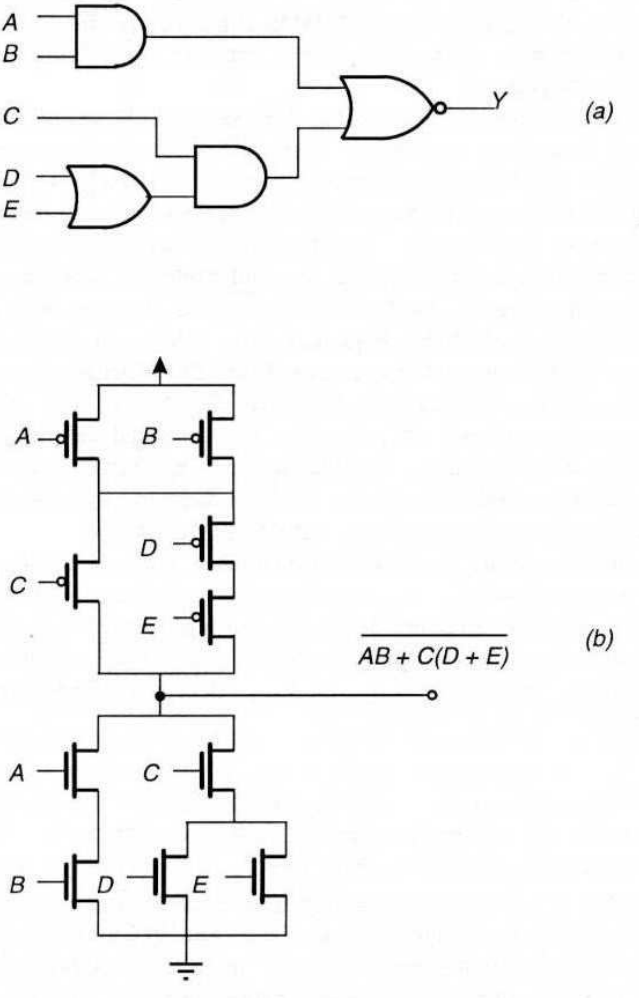


Figura 11.2 Esempio di realizzazione di una funzione logica complessa in tecnologia CMOS: a) schema logico; b) realizzazione circuitale

11.1.1 Dimensionamento dei MOS nelle porte complesse

Nelle porte complesse occorre dimensionare opportunamente i singoli transistori secondo i criteri indicati nel Paragrafo 5.8 per garantire un dato ritardo di propagazione nella condizione più sfavorevole. In generale il numero di transistori in serie deve essere limitato per evitare tempi di transizione troppo lunghi; si ricorda che, sia per i transistori NMOS che per quelli PMOS in serie al primo, le prestazioni sono degradate dall'effetto body che si verifica quando i transistori NMOS non hanno il source connesso a massa (o i PMOS il source connesso all'alimentazione). Inoltre è bene limitare il numero di drain che sono connessi al terminale di uscita, per ridurre la capacità di uscita e migliorare la dinamica. Ad esempio, un'inversione tra il parallelo degli NMOS D e E con quello C di Figura 11.2b, pur essendo perfettamente compatibile con la funzione logica voluta, fornirebbe prestazioni dinamiche peggiori.

Il dimensionamento dei transistori NMOS e PMOS va fatto con riferimento al concetto di invertitore equivalente, cioè alla struttura elementare formata da un PMOS e un NMOS rispettivamente equivalenti alle due reti in esame, struttura a cui si può riportare ogni porta comunque complessa per una determinata combinazione di ingressi. Ad esempio, con riferimento alla realizzazione della porta logica complessa riportata in Figura 11.2b, la condizione più gravosa nel passaggio dell'uscita dal valore basso a quello alto (che implica la carica della capacità di uscita attraverso la rete dei PMOS), è quella corrispondente alla conduzione di uno solo dei due rami PMOS in parallelo per ognuno dei due blocchi in serie, e in particolare nel secondo blocco la conduzione della serie dei PMOS D e E. Per quanto riguarda la condizione più gravosa nel passaggio dell'uscita dal valore alto a quello basso (che implica la scarica della capacità di uscita attraverso la rete NMOS), questa implica ancora la conduzione di uno solo dei due rami in parallelo, e, nel caso del ramo di destra, la conduzione di uno solo dei NMOS D o E.

Un criterio adottato per il dimensionamento di queste reti è quello di imporre tempi di propagazioni t_{PLH} e t_{PHL} uguali per le configurazioni di ingressi più gravose nei riguardi delle variazioni dell'uscita, o, in altre parole, di imporre che gli invertitori equivalenti, a cui si può riportare la porta complessa per queste combinazioni di ingressi, presentino un $K_{NEQ} = K_{PEQ}$. Ricordando i diversi valori di k'_N e k'_P, ciò comporta:

$$\frac{W_{PEQ}}{L_{PEQ}} = 2.5\frac{W_{NEQ}}{L_{NEQ}} \tag{11.2}$$

dove per W_{EQ} e L_{EQ} si intendono rispettivamente le dimensioni W e L dei MOS equivalenti a cui sono riconducibili le connessioni effettive delle due reti PMOS e NMOS con date combinazioni di ingressi. In base alle approssimazioni già introdotte nel Paragrafo 4.11, si possono generalizzare quei risultati assumendo, per i rapporti W/L di un MOS equivalente a J MOS connessi rispettivamente in parallelo o in serie, i valori:

$$\left.\frac{W}{L}\right|_{EQ} = \sum_J \frac{W_J}{L_J} \quad \text{per J MOS in parallelo} \tag{11.3}$$

$$\left.\frac{L}{W}\right|_{EQ} = \sum_J \frac{L_J}{W_J} \quad \text{per J MOS in serie} \tag{11.4}$$

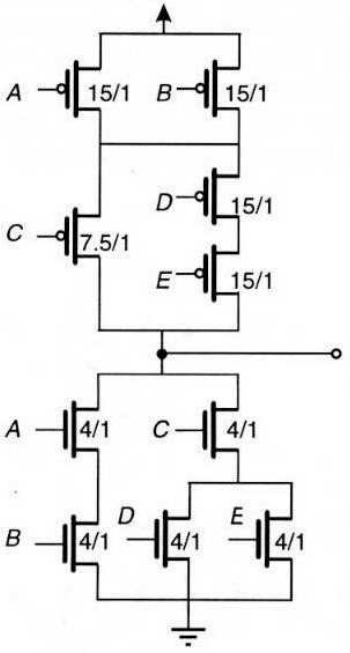


Figura 11.3 Dimensionamento della porta complessa di Figura 11.2

Ad esempio, il dimensionamento dei MOS della porta complessa di Figura 11.2b può essere effettuato come indicato in Figura 11.3, imponendo la condizione (11.2), e considerando per i PMOS una volta la configurazione: A (B) in serie con $D + E$, e una seconda volta la configurazione: A (B) in serie con C, mentre per i NMOS quella: C (A) in serie con D (E). Se si prende come riferimento un invertitore equivalente con: $W/L|_N = 2/1$ μm, $W/L|_P = 5/1$ μm, in base alle (11.3), (11.4) si determinano i valori dei rapporti W/L per i diversi MOS riportati nella Figura 11.3. La serie dei tre PMOS con $W/L = 15/1$ comporta un $W/L|_{EQ} = 15/3 = 5/1$, come anche la serie dei PMOS A e C comporta ancora un $W/L|_{EQ} = 5/1$; per ognuno dei due rami con due NMOS in serie si ha un $W/L|_{EQ} = 4/2 = 2/1$. Con questi valori di dimensionamento, le transizioni: $A,B\ 0 \rightarrow 1$; $C,D\ 0 \rightarrow 1$; $C, E\ 0 \rightarrow 1$; $A, C\ 1 \rightarrow 0$; $B, C\ 1 \rightarrow 0$; $A, D, E\ 1 \rightarrow 0$; $B, D, E\ 1 \rightarrow 0$ presenteranno uguali tempi di propagazione, mentre le combinazioni di ingressi che portano più di un

MOS tra quelli in parallelo a condurre presenteranno tempi di propagazione minori.

11.1.2 Tracciato delle porte complesse CMOS

Lo schema circuitale di una porta complessa CMOS può essere ricavato in via alternativa utilizzando un grafo ad archi per rappresentare le connessioni tra i MOS sia per la rete NMOS che PMOS.

Si parte dalla rete NMOS che viene realizzata tra il nodo di uscita Y e quello di massa V_{SS}. Con riferimento alla Figura 11.4 (che rappresenta i passi di questa costruzione per la rete di Figura 11.2), si riportano inizialmente una serie di archi che rappresentano i singoli NMOS, e i cui estremi rappresentano il source e il drain; in questo caso il nodo n corrisponde nello schema elettrico al punto intermedio della serie tra il NMOS C e i NMOS D, E in parallelo (Figura 11.4a).

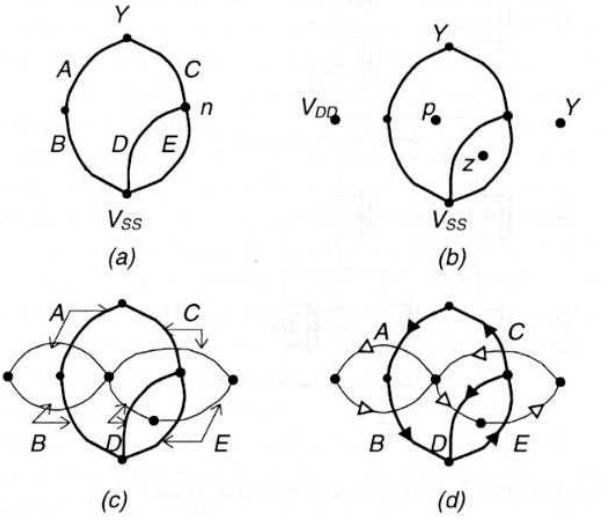


Figura 11.4 Costruzione dei grafi complementari per la porta complessa CMOS di Figura 11.2: a) grafo per la rete NMOS; b) nodi per il grafo duale; c) grafo della rete PMOS (a linea sottile) sovrapposto a quello NMOS; d) percorso di Eulero tra i nodi dei grafi complementari

Una volta realizzato il grafo ad archi per la rete NMOS (dal circuito elettrico o direttamente per ispezione della funzione logica da implementare), la rete PMOS può essere ricavata direttamente, senza ricorrere all'elaborazione dell'espressione logica con le variabili invertite, costruendo un grafo complementare per la rete PMOS, inserendo in ogni area completamente racchiusa dagli archi del primo grafo un nuovo nodo (nella Figura 11.4b i nodi p e z) e due nodi ulteriori, all'esterno delle aree racchiuse dal grafo, corrispondenti all'alimentazione V_{DD} e all'uscita Y. Tutti questi nodi vengono connessi da archi che ta-

gliano i corrispondenti archi che definiscono la rete NMOS (Figura 11.4c), e gli archi così creati assumono lo stesso nome di quelli che intersecano (in altre parole i MOS sono pilotati dallo stesso ingresso). In questo modo si viene a determinare la topologia da dare alla rete PMOS una volta determinata quella NMOS.

La rappresentazione mediante il grafo ad archi delle strutture della porta complessa CMOS è di notevole aiuto per definire il tracciato della porta stessa nel silicio. La scelta ottimale per questo genere di porte è quella di utilizzare due linee parallele di transistori, rispettivamente NMOS e PMOS, tra le linee di alimentazione e massa, ognuno dei quali scambia il terminale di source o di drain con il MOS adiacente, e con i collegamenti di gate comuni alle coppie di transistori NMOS e PMOS per ogni data variabile. È possibile realizzare una tale configurazione per la porta se nel grafo ad archi che definisce il circuito elettrico della porta si possono percorrere entrambi i grafi (corrispondenti alle reti NMOS e PMOS), partendo per ogni grafo da un nodo e toccandoli tutti una sola volta (nella stessa sequenza in entrambi i grafi), senza percorrere i singoli archi per più di una volta.

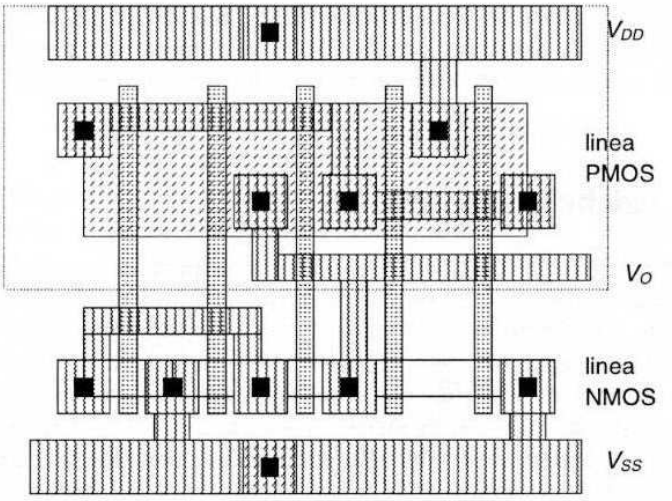


Figura 11.5 Tracciato della porta CMOS di Figura 11.2 con $W/L|_P = 15/1$ e $W/L|_N = 4/1$

Questo percorso dei nodi del grafo che soddisfa alle condizioni suddette viene detto *percorso di Eulero*; in Figura 11.4d il percorso identificato è quello *DECAB*. Se non è possibile trovare un percorso di Eulero unico per tutto il grafo, non è possibile realizzare la rete con un'unica striscia di transistori NMOS e PMOS contigui, ma occorre prevedere contatti separati per qualcuno dei source o dei drain dei MOS; in tal caso è possibile minimizzare queste interruzioni scegliendo opportu-

namente la sequenza dei nodi da percorrere e modificando la topologia delle reti. In Figura 11.5 è riportato il tracciato della porta CMOS utilizzando il percorso *DECAB*; in questo caso il dimensionamento dei transistori è stato effettuato scegliendo per tutti i PMOS il valore *W/L* maggiore, pari a 15/1, come anche per tutti gli NMOS quello pari a 4/1, in quanto la riduzione della dimensione *W* per qualcuno dei MOS nella stessa striscia non comporta nessun vantaggio di occupazione di area e complica inutilmente il tracciato della porta.

Le logiche basate su porte complesse come quelle descritte vengono dette logiche CMOS pienamente complementari, o FCMOS (*Fully Complementary MOS*), in quanto utilizzano sempre un numero di transistori PMOS pari a quello NMOS. Come si è visto, queste logiche presentano rispetto alle logiche NMOS il vantaggio di avere una dissipazione di potenza molto ridotta e margini di rumore elevati, ma comportano anche degli svantaggi, quali quello di una occupazione relativamente elevata di area di silicio, e una capacità di ingresso (per ogni terminale di pilotaggio) pari a circa 3.5 volte quella di un ingresso a singolo transistore NMOS, in quanto ogni ingresso va a pilotare sia il NMOS che il PMOS. Per i circuiti VLSI l'occupazione di spazio è un requisito determinante, mentre si possono tollerare margini di rumore anche più ridotti, in quanto le interconnessioni tra i circuiti elementari avvengono nello stesso chip di silicio, e i segnali vengono trasferiti alle uscite attraverso opportuni stadi disaccoppiatori. Si sono quindi sviluppate strutture alternative a quelle basate su logiche CMOS pienamente complementari, che permettono di ridurre l'occupazione di area, e che verranno presentate nei paragrafi seguenti.

11.2 Logiche pseudo-NMOS

I circuiti logici pseudo-NMOS sono identici a quelli delle logiche NMOS, con la differenza di utilizzare come carico attivo, al posto del transistore NMOS a svuotamento, un transistore PMOS con il gate connesso a massa. La logica realizzata è del tipo "a rapporto" come nel caso delle logiche NMOS, in quanto le prestazioni statiche e dinamiche dipendono dal rapporto dei *K* dei transistori P ed N, e presenta un consumo di potenza anche in condizioni di riposo. Rispetto alle logiche NMOS, vi è in ogni caso il vantaggio che il transistore PMOS di carico non presenta effetto body, contrariamente a quanto avveniva per quello a svuotamento della logica NMOS. Inoltre la corrente fornita dal bipolo di carico è maggiore, a parità di rapporto *W/L* (e quindi di area occupata), perché il PMOS con il gate connesso a massa ha la massima tensione V_{GS} (in modulo), pari a $|V_{DD}|$.

In Figura 11.6 è riportato lo schema elettrico di una porta che realizza la stessa funzione logica della porta CMOS di Figura 11.2b; da questo schema si può notare che la rete dei 5 PMOS della parte alta della porta è sostituita da un solo PMOS. Il risparmio in transistori per una porta a *n* ingressi è quindi di $n-1$. Un ulteriore vantaggio è quello della maggiore semplicità di interconnessione che deriva dal non dovere inviare i singoli ingressi sia agli NMOS che ai PMOS, il che nelle logiche

CMOS può creare problemi di incrocio tra le linee di collegamento e quindi richiede topologie più complicate.

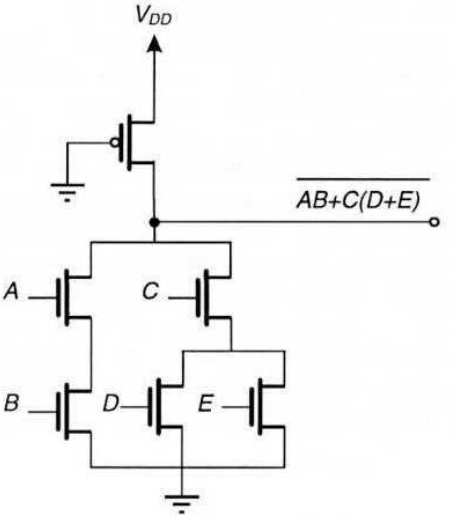


Figura 11.6 Porta logica complessa con struttura pseudo-NMOS

La capacità di ingresso per ognuno degli ingressi della porta è minore di quella dell'analoga struttura CMOS perché ogni ingresso vede solo la capacità di gate del NMOS; questo tuttavia non comporta direttamente un miglioramento del ritardo di propagazione, in quanto la corrente di carica della capacità è limitata dal valore di K_P, e quest'ultimo viene scelto più piccolo di K_{NEQ} per mantenere un valore di tensione V_{OL} non troppo elevato. Poiché la corrente di scarica della capacità di uscita è la differenza delle correnti I_N e I_P circolanti rispettivamente nella rete NMOS e nel PMOS, mentre quella di carica è la sola corrente I_P, la condizione di uguaglianza dei tempi di propagazione t_{PLH} e t_{PHL} comporta la condizione $I_N - I_P = I_P$ da cui deriva $I_P = I_N/2$. Ciò comporta che nelle strutture pseudo-NMOS sia le capacità di uscita che le correnti sono circa dimezzate rispetto a quelle delle strutture CMOS, e quindi i tempi di ritardo sono comparabili, per cui la differenza sta essenzialmente nell'occupazione di area e nella dissipazione di potenza.

11.3 Logiche con porte di trasmissione

Una notevole possibilità dei circuiti MOS è quella offerta dall'inserimento di transistori *in serie* ai terminali di ingresso e uscita, anziché *in parallelo*, come visto finora. Questi transistori inseriti in serie nel circuito e controllati dal terminale di gate agiscono come interruttori lungo la via del segnale, e pertanto vengono detti *porte di trasmissione*, in quanto, attraverso il comando sulla gate, permettono o non

il passaggio del segnale lungo il collegamento in cui sono inseriti. Il segnale di comando ϕ è detto *fase* perché in generale, come vedremo nelle logiche dinamiche, si utilizzano più segnali logici di controllo, che operano con la stessa frequenza e legami di fase (cioè di sfasamenti temporali) ben definiti tra loro.

In generale una porta di trasmissione è un circuito logico che presenta in uscita la variabile di ingresso se il segnale di controllo è alto, e non presenta il segnale se il controllo è basso, secondo la tabella della verità di Figura 11.7a. I transistori NMOS sono in effetti degli interruttori che realizzano bene lo stato di circuito aperto, e male quello di corto circuito, perché in quest'ultimo caso essi presentano una resistenza interna (tra drain e source) non trascurabile; tuttavia se l'uscita viene applicata alla gate di un successivo circuito MOS, questa non assorbe corrente, e quindi il segnale Y in uscita non viene attenuato dalla resistenza interna del MOS. Questa considerazione spiega anche perché non è possibile realizzare porte di trasmissione in tecnologia bipolare; in quest'ultimo caso, anche se i dispositivi presentano una resistenza più bassa in saturazione, la corrente richiesta dal carico (e cioè dalle basi dei transistori connessi in uscita) darebbe luogo a cadute di tensione nella resistenza R_{ON}, e quindi ad una perdita di livello logico per ogni porta posta in serie.

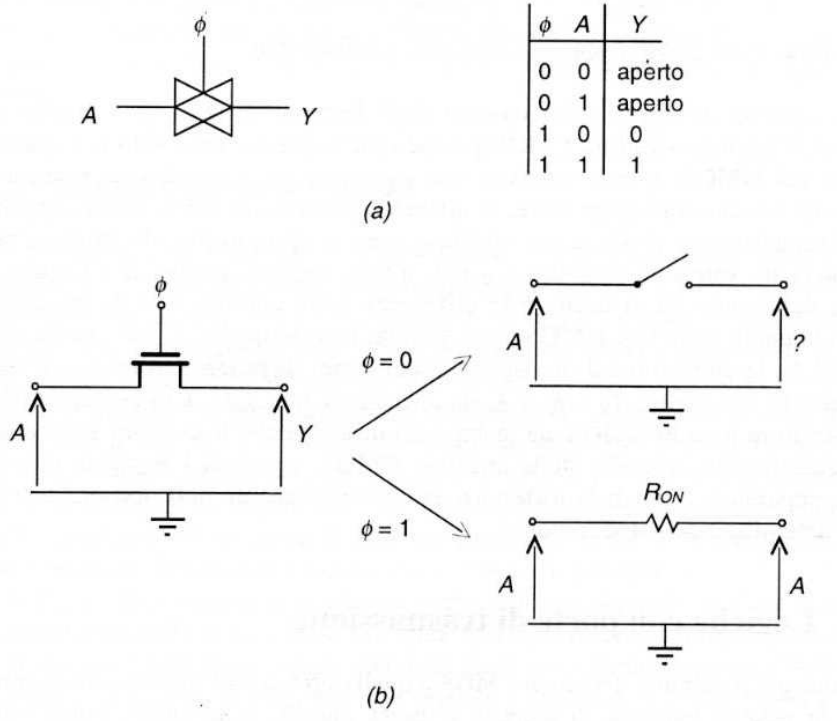


Figura 11.7 a) Simbolo logico di una porta di trasmissione; b) porta di trasmissione a NMOS

Con il segnale di fase al livello alto, il MOS agisce come porta bidirezionale, cioè trasmette il segnale in entrambe le direzioni, per cui è inessenziale definire quale è il source e quale il drain del transistore. In effetti, con riferimento alla Figura 11.7b, se il segnale di ingresso A è più grande di quello di uscita Y, la corrente passerà (nel transitorio) da A a Y e quindi il terminale connesso all'ingresso agirà in questo caso da drain e quello di uscita da source; viceversa, se A è ad un livello più basso di Y (come è il caso se $A = 0$, e Y, prima dell'apertura della porta, era al valore 1), il terminale connesso all'ingresso agirà da source e quello connesso all'uscita da drain.

Consideriamo ora il caso in cui l'ingresso sia al valore alto (V_{DD}) e l'uscita sia inizialmente al valore basso (ad esempio considerando la capacità di gate del MOS a cui è connessa la porta inizialmente scarica). Se il segnale di fase ϕ al livello alto è anch'esso al valore V_{DD}, la tensione di uscita non potrà raggiungere il valore V_{DD}, in quanto per $V_O = V_{DD} - V_T$ si avrà una tensione $V_{GS} = V_T$, e il transistore non conduce più. La funzione di trasferimento di questa porta è quindi quella di Figura 11.8 e l'uscita è inferiore all'ingresso di un valore pari alla tensione di soglia V_T. Quest'ultima è incrementata dell'effetto body dovuto alla tensione di source $V_S > 0$, per cui la perdita di tensione a regime corrisponde alla tensione di soglia:

$$V_T = V_{TO} + \gamma\left(\sqrt{\phi^* + V_{DD} - V_T} - \sqrt{\phi^*}\right) \tag{11.5}$$

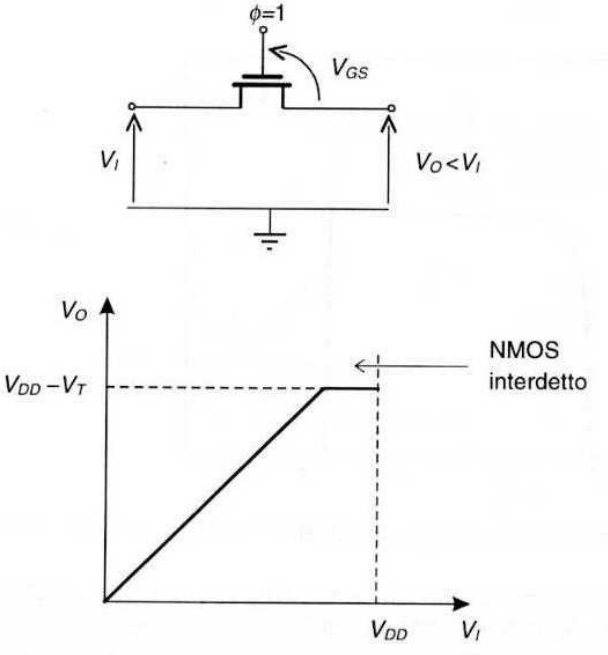


Figura 11.8 Funzione di trasferimento di una porta di trasmissione NMOS

La porta di trasmissione può essere considerata un circuito logico non rigenerativo, cioè un circuito che non ripristina i livelli logici, in quanto, come si vede dalla Figura 11.8, la funzione di trasferimento ha una caratteristica con pendenza unitaria.

Per le prestazioni dinamiche, la porta NMOS si comporta come un circuito RC con una resistenza R nonlineare in serie dovuta al NMOS, e una capacità in parallelo dovuta alla capacità di ingresso del circuito a valle e alla capacità source-substrato del NMOS (Figura 11.9). La porta presenta un tempo di propagazione maggiore nel trasmettere il livello logico alto, legato all'elevata resistenza nella trasmissione dei livelli logici alti (1 logico) in quanto il MOS va verso l'interdizione quando l'uscita cresce verso il valore $V_{DD} - V_T$, essa invece trasmette bene il livello logico basso (0 logico), in quanto per uscita V_O bassa la porta presenta la minima resistenza.

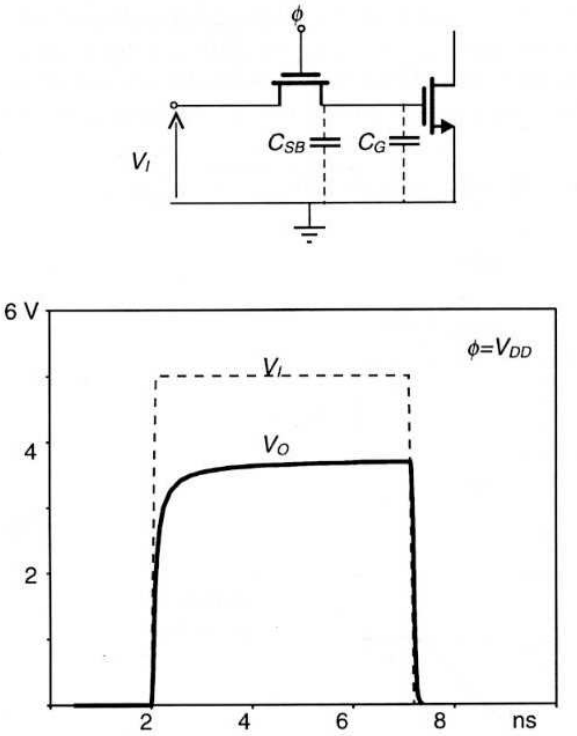


Figura 11.9 Tempi di propagazione della porta NMOS

La porta di trasmissione può essere realizzata anche con tecnologia CMOS, ponendo in antiparallelo un NMOS e un PMOS, pilotati rispettivamente dal segnale ϕ e da quello $\overline{\phi}$, secondo lo schema di Figura 11.10. In tal caso, poiché la tensione di pilotaggio del PMOS è data da $\overline{\phi} - V_I$, il PMOS trasmette bene i segnali logici alti,

in quanto in tal caso la tensione gate-source è la massima in modulo ($V_{GS} = -V_{DD}$), mentre il transistore NMOS come si è visto trasmette bene i livelli bassi. Quindi la funzione di trasferimento complessiva della porta CMOS è unitaria da 0 a V_{DD}, perché vi è almeno un MOS che conduce per qualsiasi segnale di ingresso, come si vede dalla funzione di trasferimento complessiva di Figura 11.10, ottenuta da quella di Figura 11.8 considerando che per il PMOS la limitazione si ha per tensioni $V_I < V_T$. La porta di trasmissione CMOS non ha quindi la perdita di V_T sulla tensione di uscita e si comporta come la porta ideale logica di Figura 11.7a; tuttavia richiede per il controllo sia il segnale ϕ che il suo negato.

Anche il comportamento dinamico della porta CMOS è migliore di quello della porta NMOS in quanto, essendovi sempre un transistore in conduzione nella regione di linearità, i tempi di risposta della rete *RC* equivalente sono rapidi sia per la trasmissione del livello alto che di quello basso.

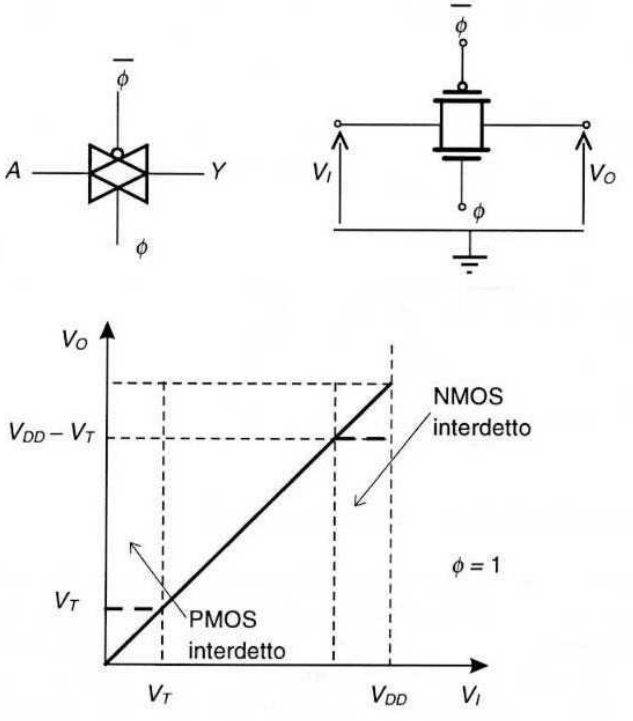


Figura 11.10 Porta di trasmissione CMOS

Le porte di trasmissione permettono una maggiore flessibilità nella realizzazione di funzioni logiche, in quanto le variabili logiche possono essere applicate, oltre che all'ingresso, anche al terminale di controllo; ciò permette in alcuni casi, come vedremo nel paragrafo seguente, di ottenere le funzioni logiche volute con reti

molto compatte. Tuttavia il progetto di circuiti logici con porte di trasmissione richiede particolare attenzione per i seguenti aspetti:

a. le porte connesse ad una stessa uscita non possono essere contemporaneamente in conduzione con livelli logici in ingresso diversi, in quanto la tensione di uscita assumerebbe un livello intermedio e non correlato con nessuno degli ingressi;
b. il ritardo di propagazione di più porte in serie sulla stessa linea corrisponde a quello di una rete a celle R_C in cascata, e quindi aumenta secondo il quadrato del numero di porte in serie e non linearmente con il numero delle porte, come nel caso di una connessione in serie di porte logiche standard.

Un ulteriore problema che complica la progettazione dei circuiti con porte di trasmissione è quello del disturbo introdotto in uscita dal segnale di fase. Per questi circuiti la porta di trasmissione è connessa in serie al percorso del segnale, e l'uscita è connessa ad un carico capacitivo, che corrisponde alla (o alle) gate dei MOS connessi in uscita.

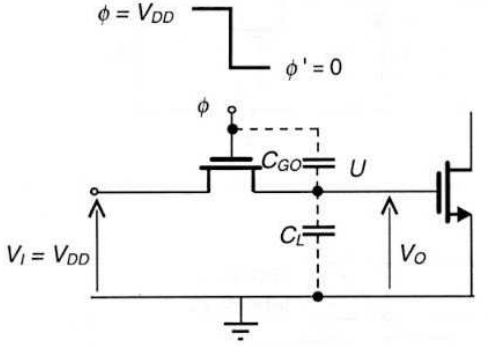


Figura 11.11 Disturbo introdotto dal segnale di fase sull'uscita

Quando la porta viene chiusa dal segnale di fase che passa dal livello logico alto a quello basso, la capacità tra il terminale di gate e quello di uscita (source o drain), che indicheremo con C_{GO}, trasmette parte della variazione del segnale di fase sul terminale di uscita, secondo il partitore capacitivo indicato in Figura 11.11. Indicando con l'apice ' i valori dopo la transizione $1 \rightarrow 0$, e considerando che il MOS è al limite dell'interdizione per $\phi = V_{DD}$ (in quanto $V_O = V_{DD} - V_T$), si può scrivere la carica sul nodo U di uscita prima della transizione come:

$$Q_U = C_L V_O + (-C_{GO} V_T) \tag{11.6}$$

Subito dopo la transizione della fase ϕ da V_{DD} a 0, la carica Q'_U non è variata rispetto a Q_U, e dall'uguaglianza della carica si ha:

$$C_L V_O + (-C_{GO} V_T) = (C_L + C_{GO}) V_O' \Rightarrow V_O' = \frac{C_L V_O - C_{GO} V_T}{C_L + C_{GO}} \qquad (11.7)$$

Ricordando che $V_O = V_{DD} - V_T$, e sostituendo nella (11.7) si ha:

$$V_O' = \frac{C_L}{C_L + C_{GO}} V_{DD} - V_T \qquad (11.8)$$

e il salto di tensione trasmesso all'uscita sul valore V_O sarà dato da:

$$\Delta V_O = V_O' - V_O = \frac{C_L}{C_L + C_{GO}} V_{DD} - V_T - (V_{DD} - V_T) = -V_{DD}\left(\frac{C_{GO}}{C_L + C_{GO}}\right) \qquad (11.9)$$

Per ridurre questo disturbo introdotto sull'uscita dal segnale di fase occorre ridurre quanto possibile il valore della capacità C_{GO} (ossia le capacità gate-drain e gate-source) rispetto a quella di uscita (tipicamente la capacità C_G di un NMOS). Questo richiede un dimensionamento del MOS utilizzato come porta ad area minima, in particolare con un valore minimo di W in modo da ridurre i valori di C_{GSO} e C_{GDO}.

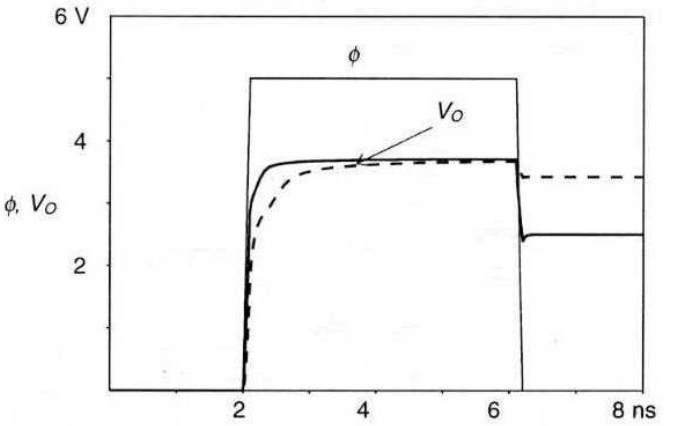


Figura 11.12 Andamento della tensione in uscita della porta di trasmissione con NMOS di Figura 11, per un rapporto W/L = 8/1 (linea continua) e W/L = 2/1 (linea tratteggiata)

Tuttavia la scelta di un dimensionamento ad area minima penalizza il tempo di carica della capacità di uscita quando la porta viene aperta e vi è un livello logico alto in ingresso; ad esempio in Figura 11.12 sono riportati gli andamenti dell'uscita V_O, in corrispondenza di un ingresso $V_I = V_{DD}$, durante e dopo l'applicazione del

segnale di fase ϕ, per un dimensionamento del NMOS con due diversi rapporti W/L. Con un valore minimo di W/L il salto di tensione ΔV_O sull'uscita è ridotto, ma il tempo di carica della capacità in uscita è relativamente grande. Se si sceglie un rapporto W/L più grande, la carica della capacità è più rapida, ma il salto di tensione ΔV_O è più grande, e può portare il MOS di uscita in interdizione se $V_{DD} - V_T + \Delta V_O < V_T$. Il dimensionamento delle porte di trasmissione va quindi effettuato tenendo conto di queste esigenze contrastanti.

11.3.1 Circuiti combinatori con porte di trasmissione

L'applicazione più efficace delle porte di trasmissione è nella realizzazione di circuiti multiplexer e demultiplexer, per i quali la funzione di interruttore serie esplicata dalle porte è direttamente implementabile nella funzione di selezione delle linee dati voluta.

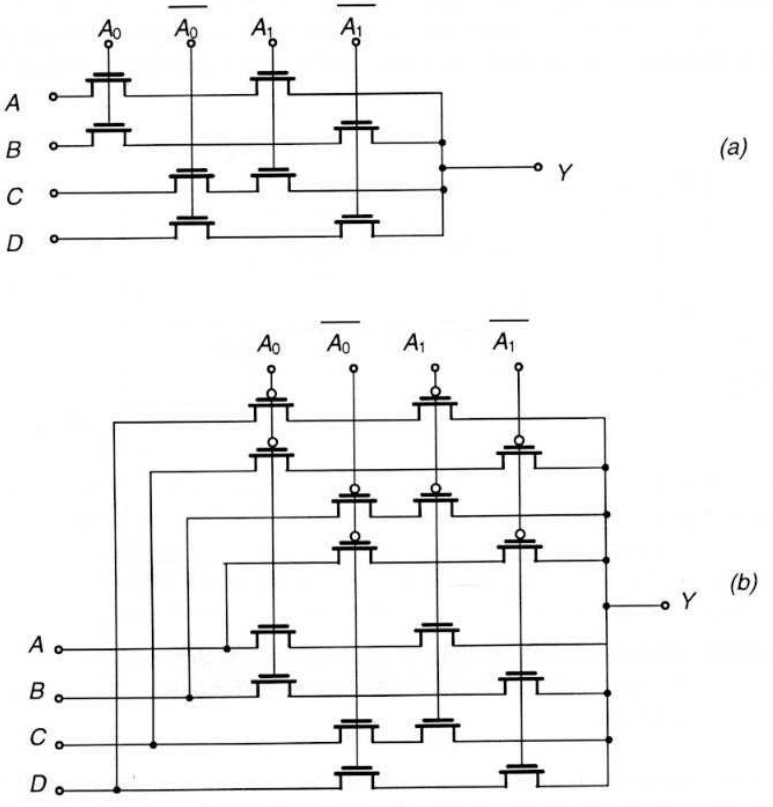


Figura 11.13 Multiplexer a 4 vie; a) con porte di trasmissione NMOS; b) con porte CMOS

In Figura 11.13 è riportato un esempio di realizzazione di multiplexer sia con porte NMOS che con porte CMOS; in quest'ultimo caso nello schema circuitale sono state omesse, perché inutili, le connessioni dirette tra i drain NMOS e i source PMOS per le singole porte, dato che la singola linea è attivata solo se tutte le porte su quella linea sono attivate. Anche nel caso di realizzazione con porte di trasmissione CMOS, per un multiplexer a quattro ingressi si utilizzano solo 16 transistori, rispetto ai 32 transistori necessari per una realizzazione con porte logiche standard (4 porte NAND a 3 ingressi più 1 porta NOR a 4 ingressi). Una versione ancora più compatta è quella che utilizza una configurazione ad *albero* per gli interruttori equivalenti, come quella riportata in Figura 11.14, che utilizza per la stessa funzione con porte CMOS solo 12 transistori.

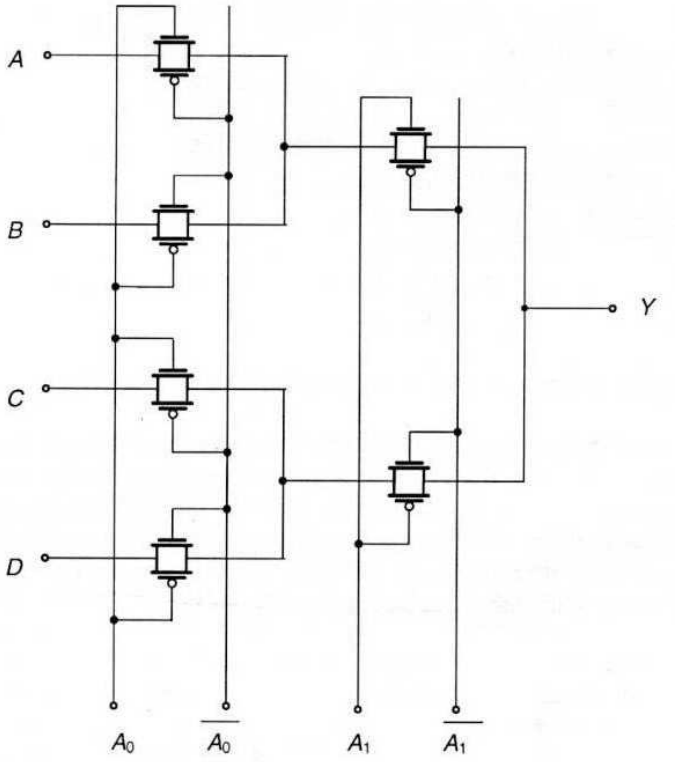


Figura 11.14 Struttura ad "albero" di un multiplexer a 4 ingressi con porte CMOS

Il tracciato della struttura ad albero del multiplexer di Figura 11.14 è riportato in Figura 11.15. Le linee che portano i bit di indirizzo e che pilotano i gate dei transistori NMOS e PMOS sono realizzate in polisilicio, in modo da poter essere intersecate dalle connessioni metalliche dei transistori che connettono le linee dati all'uscita attraverso le porte di trasmissione; in questo tracciato i transistori PMOS

sono dimensionati con un rapporto $W/L = 10\lambda/2\lambda$, mentre quelli NMOS con $W/L = 6\lambda/2\lambda$, in quanto per le porte CMOS non è necessario imporre la condizione $K_N = K_P$.

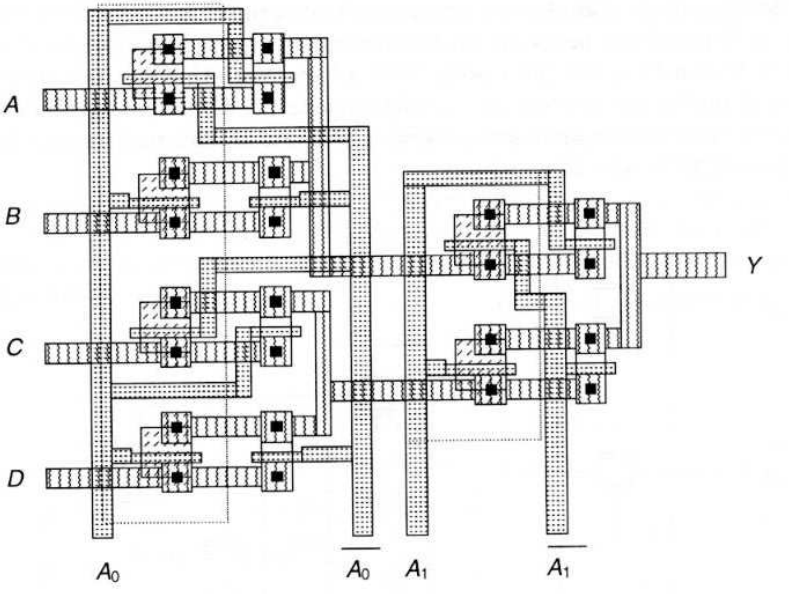
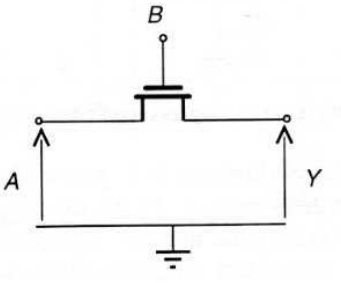


Figura 11.15 Tracciato della struttura ad albero del multiplexer di Figura 11.14

Un ulteriore vantaggio delle porte di trasmissione è legato alla possibilità di applicare la variabile logica sia alla gate che a uno dei terminali di ingresso-uscita della porta (source o drain); questo permette di effettuare una funzione logica AND tra queste due variabili nella singola porta di trasmissione (quindi con un solo transistore NMOS al limite) invece che con una porta logica con due ingressi, come indicato sinteticamente dalla tabella della verità in Figura 11.16.



A	B	Y
0	1	0
1	0	0
0	0	0
1	1	1

Figura 11.16 Funzione AND realizzata con porta di trasmissione

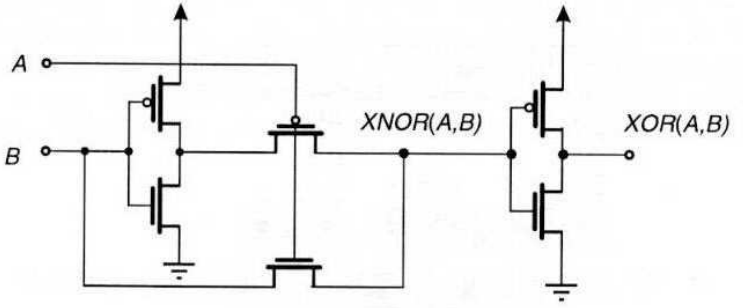


Figura 11.17 Realizzazione di una porta XOR con porte di trasmissione NMOS e PMOS

Questa possibilità di uso delle variabili permette una realizzazione di funzioni logiche in maniera molto compatta; un esempio è quello della funzione XOR che è la funzione base per la realizzazione dei sommatori binari, come si è visto nel Capitolo 10. In Figura 11.17 è mostrato il circuito di una porta XOR realizzata con due sole porte di trasmissione, rispettivamente NMOS e PMOS. In questo caso la funzione viene realizzata in maniera molto compatta, sfruttando la possibilità di pilotare la porta PMOS con la variabile B e di applicare come variabile alla porta la variabile $\overline{A}$ in uscita dall'invertitore, in modo da realizzare la funzione AND $\overline{A} \cdot \overline{B}$; la porta NMOS realizza invece la funzione AB, per cui all'ingresso del secondo invertitore si ritrova la funzione XNOR = $\overline{A} \cdot \overline{B} + AB$ e all'uscita la funzione XOR = $\overline{\overline{A} \cdot \overline{B} + AB} = \overline{\overline{A + B} + AB}$ (vedi le Equazioni (10.19)). L'impiego dell'invertitore in uscita, oltre ad essere necessario per effettuare la negazione della funzione XNOR fornita a valle delle porte, migliora anche le prestazioni statiche (in termini di livelli logici) della porta logica, ripristinando il valore della tensione di soglia V_T perso dalle porte di trasmissione con un solo transistore, e l'immunità ai disturbi.

Il tracciato della porta XOR di Figura 11.17 è riportato nella Figura 11.18. In questo tracciato i transistori NMOS sono stati dimensionati con un rapporto $W/L = 4\lambda/2\lambda$, e quelli PMOS con un rapporto $W/L = 10\lambda/2\lambda$, pari a 2.5 quello dei transistori NMOS; le variabili sono applicate mediante linee in polisilicio, mentre le linee di alimentazione e di massa sono in metallo, come anche è in metallo la linea che fornisce l'uscita.

Ricordiamo che nel Capitolo 10 si è visto come un addizionatore completo (*full adder*) di due bit può essere realizzato con due porte XOR in cascata per il termine somma S_i, e utilizzando una funzione logica complessa per il riporto C_i, secondo la (10.20) che può anche essere scritta in maniera leggermente diversa, come indicato nella (11.10):

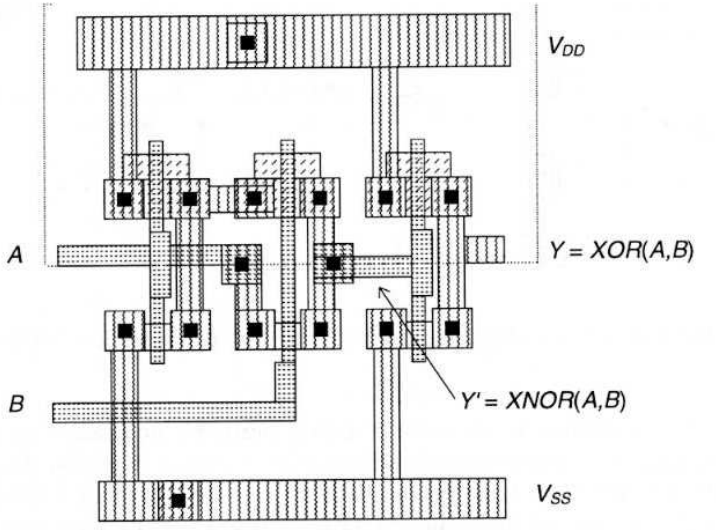


Figura 11.18 Tracciato della porta XOR di Figura 11.17

$$S_i = C_{i-1} \oplus A_i \oplus B_i \qquad C_i = A_i \cdot B_i + (A_i + B_i) \cdot C_{i-1} \tag{11.10}$$

dove il termine C_i può essere realizzato con una porta logica complessa CMOS che realizza la funzione negata $\overline{C_i}$, seguito da un invertitore per l'ulteriore negazione.

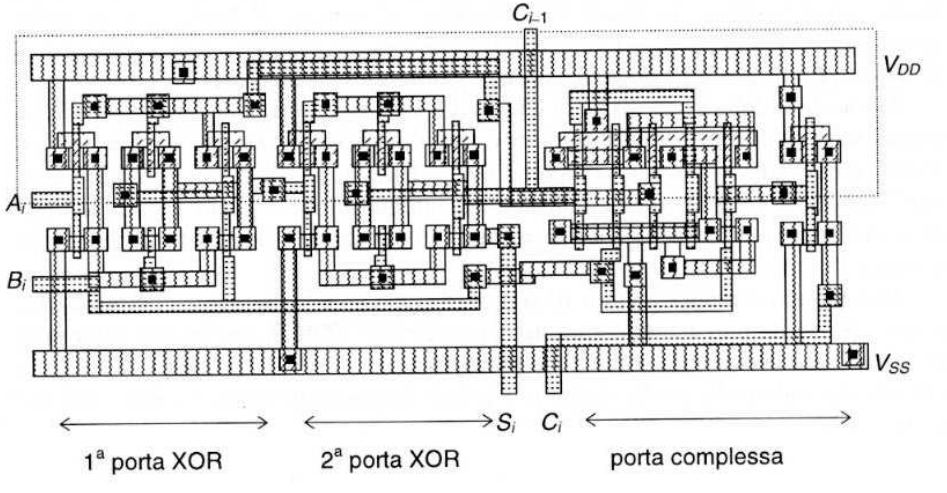


Figura 11.19 Tracciato di un *full adder* basato sulla porta XOR di Figura 11.18

Un'ulteriore applicazione delle porte di trasmissione è nella realizzazione dell'Unità Logica Booleana, ossia di un circuito che fornisce in uscita tutte le funzioni booleane delle variabili A e B, a seconda dei valori logici di opportune variabili P_i di ingresso. Questo circuito è basato sul circuito multiplexer nella versione sia a porte NMOS che a quelle CMOS; quest'ultima versione è riportata in Figura 11.20.

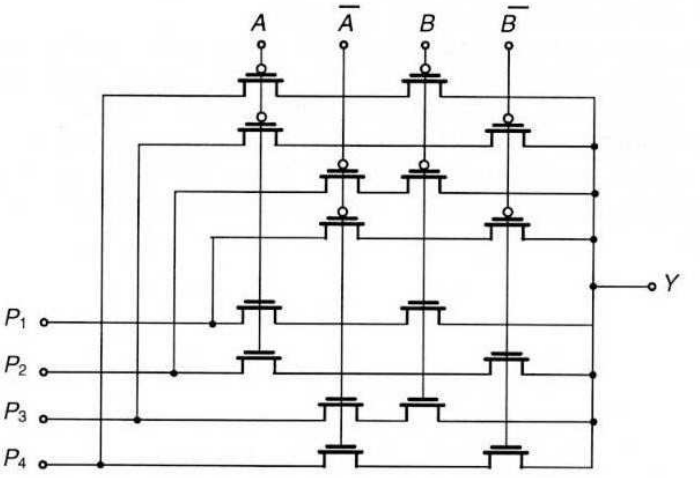


Figura 11.20 Unità Logica Booleana con porte di trasmissione CMOS

Il funzionamento del circuito si basa sul fatto che ogni diversa funzione booleana tra due variabili ha una differente tabella della verità delle quattro combinazioni possibili delle variabili, e quindi si possono associare i valori delle variabili A e B ad ognuna di queste combinazioni, fornite come ingressi alle quattro linee dati, come si può verificare dalla Tabella 11.1, che riporta le diverse funzioni tra le variabili A e B ottenibili in uscita a seconda dei valori dati agli ingressi P_i.

Tabella 11.1 Funzioni implementabili nell'Unità Logica Booleana di Figura 11.20

Y	P_1	P_2	P_3	P_4
OR(A,B)	1	1	1	0
NOR(A,B)	0	0	0	1
AND(A,B)	1	0	0	0
NAND(A,B)	0	1	1	1
XOR(A,B)	0	1	1	0
XNOR(A,B)	1	0	0	1

11.4 Logiche dinamiche MOS

Un campo di applicazione di notevole importanza e che offre significative possibilità nel progetto delle reti logiche è quello dei circuiti logici MOS dinamici. Queste logiche combinano la riduzione dell'occupazione di area e le minori capacità di ingresso delle logiche pseudo-NMOS, con i vantaggi di una logica non a rapporto (*ratioless*), come quella CMOS, per la quale non vi è consumo di potenza statico e vengono migliorate le prestazioni dinamiche a causa della maggiore corrente utilizzata per la carica delle capacità di ingresso delle porte.

Lo schema base di principio di un circuito dinamico è quello riportato in Figura 11.21, che corrisponde ad una singola cella logica dinamica in un circuito più complesso. Il blocco indicato in figura realizza la funzione logica voluta in base a soli transistori NMOS connessi in serie e/o in parallelo (corrisponde cioè alla rete di NMOS nelle versioni logiche a porte FCMOS o pseudo-NMOS); esso ha tanti ingressi quante sono le variabili logiche, e ogni ingresso corrisponde alla gate di un singolo NMOS.

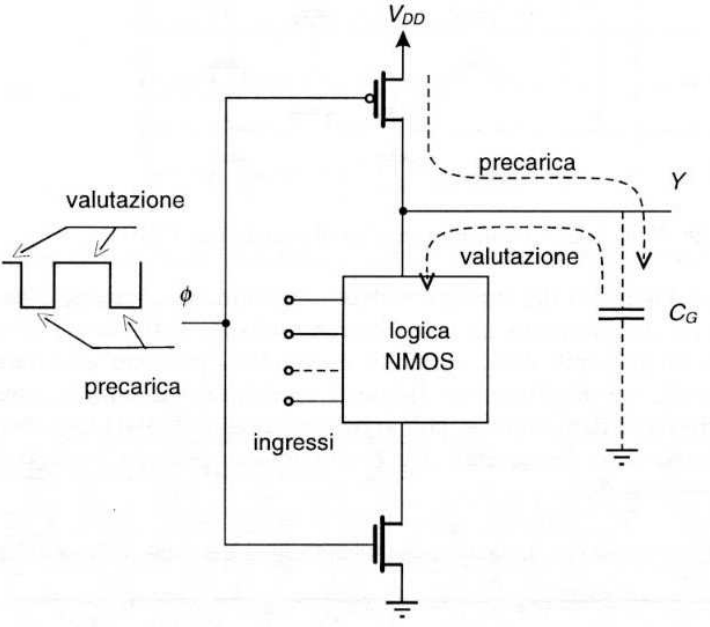


Figura 11.21 Schema di principio di un circuito logico MOS dinamico

Il transistore di carico di questa rete è un PMOS che viene però pilotato da un segnale di controllo ϕ (detto segnale di *fase*) insieme ad un ulteriore transistore NMOS connesso tra la rete logica e massa. I due transistori PMOS e NMOS comandati dalla fase ϕ, agiscono come porte di trasmissione che connettono alternativamente la rete all'alimentazione o alla massa; quindi la rete logica in qualsiasi

stato stazionario non può essere attraversata dalla corrente di alimentazione e la potenza statica dissipata è nulla, analogamente alle reti FCMOS.

Il PMOS agisce come interruttore pilotato, e quindi non è richiesta una riduzione del valore di K_P come nella logica pseudo-NMOS per ottenere bassi valori di V_{OL}; la corrente fornita dal PMOS può quindi essere aumentata in modo da velocizzare le transizioni in uscita dal livello basso a quello alto (si noti che, nelle transizioni dinamiche legate al segnale di fase, nella fase in cui il PMOS conduce, il NMOS verso massa è aperto, e quindi tutta la corrente fornita dal PMOS viene utilizzata per la carica della capacità di uscita C_G).

La caratteristica fondamentale di questi circuiti è che lo stato di uscita della porta è affidato alla carica immagazzinata nella capacità del nodo Y di uscita, capacità che è costituita di norma dalla capacità di ingresso C_G del NMOS a cui è connessa l'uscita. Questa capacità viene *precaricata* al livello elevato V_{DD} (1 logico) durante l'intervallo di conduzione del PMOS, e cioè quando il segnale di fase ϕ è basso, mentre quando ϕ è alto (e il NMOS in conduzione collega la rete a massa) essa si può eventualmente scaricare attraverso la rete logica, se gli ingressi alla rete NMOS prevedono un'uscita logica bassa, e cioè un percorso di conduzione tra i vari NMOS che costituiscono la rete.

L'intervallo di tempo in cui il segnale ϕ è al livello basso è detto fase di *precarica*, mentre l'intervallo di tempo in cui ϕ è al livello alto è detto fase di *valutazione*, in quanto è durante questo intervallo di tempo che viene valutato lo stato logico della rete NMOS. Se lo stato logico è tale che l'uscita deve presentare uno zero logico, la capacità si scaricherà attraverso la rete e conserverà questa informazione rimanendo alla tensione 0, mentre se lo stato logico è tale che l'uscita debba essere alta, non si creerà nessun percorso di conduzione nella rete, per cui la capacità non potrà scaricarsi e conserverà l'informazione mantenendo la tensione alta (V_{DD}) ai suoi capi. Questi livelli di tensione sono mantenuti durante tutta la fase di valutazione, ossia nella fase in cui occorre valutare lo stato di tutte le celle logiche connesse nel circuito; in questa fase i segnali debbono essere stati già applicati ai singoli ingressi e debbono rimanere costanti, pena la non corretta valutazione dello stato della porta.

Il nome di *logiche dinamiche* dato a questi circuiti deriva dal fatto che il funzionamento logico della porta è legato al comportamento dinamico del circuito, che alterna, attraverso il segnale ϕ, le due fasi di precarica e di valutazione. L'informazione dell'uscita della cella è conservata dinamicamente mediante lo stato di carica del condensatore di uscita; quest'ultimo tuttavia non può mantenere indefinitamente questo stato, che deve essere ripristinato in un intervallo di tempo inferiore a quello in cui la capacità si scarica attraverso le correnti inverse delle giunzioni drain-substrato e source-substrato dei MOS connessi all'uscita. Ad esempio con una corrente inversa di 100 pA, la tensione ai capi di una capacità $C_G = 0.1$ pF si riduce di 1 volt in:

$$\Delta T = \frac{C_G}{I_S}\Delta V = \frac{10^{-13}}{10^{-10}} \cdot 1 = 1 \text{ ms} \tag{11.11}$$

occorre quindi che il periodo del segnale di fase sia inferiore al valore di ΔT che comporta una riduzione ΔV superiore a quella accettabile.

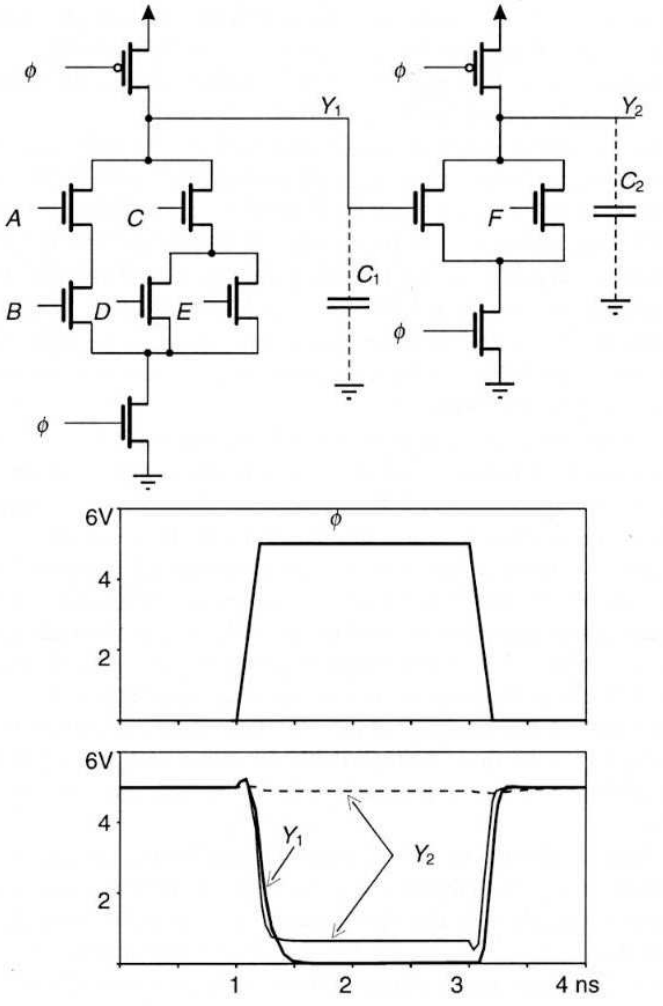


Figura 11.22 Problemi nella connessione in cascata di più celle dinamiche controllate dalla stessa fase.

L'inconveniente di questo circuito è legato al fatto che, al termine della fase di precarica e durante quella di valutazione, lo stato logico in uscita dalla singola cella logica dinamica è conservato grazie alla carica immagazzinata dalla capacità di uscita, per cui quest'ultima, se viene scaricata durante la fase di valutazione non può più ricaricarsi fino all'inizio della nuova fase di precarica. Quindi qualsiasi

transitorio sugli ingressi che comporta un'uscita bassa, anche per un breve intervallo di tempo, altera in maniera definitiva la carica della capacità connessa a questa uscita, e quindi il livello logico dell'uscita. Il problema si complica quando l'uscita di una porta dinamica 1 è utilizzata come ingresso per una successiva porta dinamica 2, in quanto, con un solo segnale di fase per tutte le porte, la fase di valutazione della porta 1 e di quella 2 coincidono; in questo caso se l'uscita della porta 1 presenta uno 0, si avrà un transitorio dovuto alla scarica condizionata della capacità del nodo di uscita attraverso la rete logica, e questo verrà visto dalla cella 2 proprio durante la fase di valutazione.

Ne consegue che, se il transitorio di valutazione della porta 2 è contemporaneo o più breve di quello della porta 1, la porta 2 può memorizzare sulla sua capacità di uscita un valore logico non corretto.

Per esemplificare questo problema si sono riportate in Figura 11.22 le grandezze in uscita dalla connessione in cascata di due porte dinamiche di cui la prima realizza la funzione complessa di Figura 11.2, e la seconda realizza una funzione NOR a due ingressi. Supponiamo che la porta 1 abbia in ingresso le variabili logiche seguenti: $A = B = 1$, $C = D = E = 0$, a cui corrisponde un'uscita $Y = 0$. Supponiamo ancora per la porta 2 una variabile logica $F = 0$, per cui se $Y_1 = 0$ debba essere $Y_2 = 1$. Il transitorio di scarica di C_1 è rallentato dalla presenza di tre transistori NMOS in serie, mentre la scarica di C_2 coinvolge solo due NMOS. Questo comporta che la capacità C_2 si scarica apprezzabilmente durante il transitorio della cella 1 (curva continua), portandosi nella fase di valutazione ad un valore logico scorretto. Solo se il valore di C_2 è molto più elevato di C_1 (curva tratteggiata), la prima si scarica di un valore inapprezzabile durante il transitorio di valutazione della cella 1, e il valore logico in uscita dalla cella 2 è quello corretto; questa tuttavia è una soluzione inaccettabile perché penalizza eccessivamente il tempo di propagazione della cella 2.

La soluzione per risolvere questo problema, nel caso di più celle dinamiche in serie, è quella di comandare in maniera sequenziale le operazioni di precarica e di valutazione per le diverse celle, in modo da evitare una fase di valutazione comune per tutte le celle in cascata. Esempi di queste logiche sono presentati nei due paragrafi seguenti.

11.5 Logica dinamica a due fasi

I circuiti logici a due fasi si basano sull'uso di due differenti segnali di controllo ϕ_1 e ϕ_2, utilizzati rispettivamente per tutte le celle di indice dispari o pari della connessione in cascata, e sull'inserzione di porte di trasmissione tra le celle consecutive, pilotate dai segnali di fase negati ϕ_1 e ϕ_2. I segnali di fase sono tali che la fase di precarica della cella precedente (ϕ_1 basso) corrisponde alla fase di valutazione della cella successiva (ϕ_2 alto). In tal modo durante la fase di valutazione di ogni singola cella (fase ϕ alta), la porta di trasmissione tra la porta precedente e quella in esame è chiusa, e questo impedisce che eventuali transitori della porta precedente possano alterare i valori in ingresso alla porta in valutazione.

Il funzionamento della logica è esemplificato in Figura 11.23, dove sono indicate le relazioni di fase tra i due segnali ϕ_1 e ϕ_2. Durante la fase di precarica della cella 1 la porta di trasmissione tra la cella 1 e 2 è chiusa, e resta chiusa anche all'inizio della valutazione della cella 1; nella fase di precarica della cella 2 la porta di trasmissione si apre, ma l'eventuale residuo di transitorio non altera lo stato della capacità di uscita della cella 2 che è in fase di carica. La fase di valutazione della cella 2 si svolge quando la porta di trasmissione 1-2 è chiusa, per cui i segnali in ingresso non possono più variare (in ogni caso il transitorio di valutazione della cella 1 dura ovviamente molto meno di tutta la fase di valutazione di ϕ_1, per cui i segnali di ingresso alla porta di trasmissione si sono in ogni caso stabilizzati alla fine del tempo di apertura della stessa).

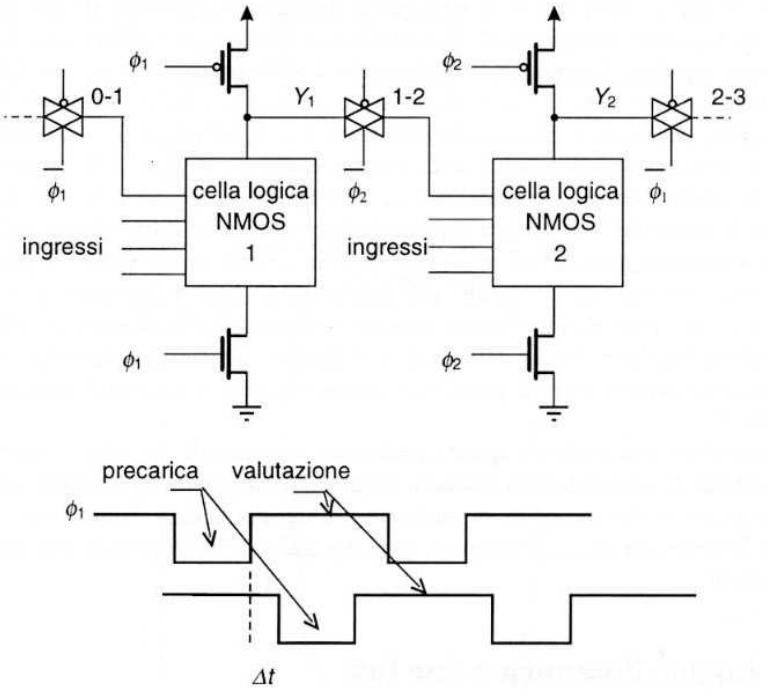


Figura 11.23 Funzionamento dei circuiti logici dinamici a 2 fasi

In linea di principio si potrebbero generare i due segnali di fase utilizzando il segnale ϕ e il suo negato; in pratica questa soluzione è inaffidabile, perché basta una piccola differenza di ritardo tra i percorsi dei due segnali di fase per portare le due fasi di precarica a sovrapporsi parzialmente, con la porta di trasmissione tra le celle contigue ancora aperta, e ciò potrebbe portare ad una scarica anomala della seconda cella all'inizio della sua fase di valutazione. In pratica quindi i due segnali vengono scelti in modo che vi sia un intervallo di tempo Δt tra i tempi di precarica

di ϕ_1 e ϕ_2, in modo da evitare sovrapposizioni tra i transitori di precarica delle due celle.

Occorre puntualizzare che in questo schema di funzionamento dinamico, la fase di precarica della porta 1 coinvolge solo la capacità C_{u1} del nodo di uscita della cella (vedi Figura 11.24) e non quella C_{i2} di ingresso della cella 2, perché durante questa fase la porta di trasmissione 1-2 è chiusa; se ora la porta 1 prevede un'uscita alta in fase di valutazione, quest'informazione viene immagazzinata nella carica di C_{u1} nell'intervallo Δt e viene trasferita alla C_{i2} durante la fase di apertura della porta di trasmissione (ϕ_2 basso). Il trasferimento dello stato logico alto (1 logico) viene quindi in pratica effettuato attraverso una ridistribuzione della tensione V_{OH1} tra le due capacità secondo la relazione:

$$Q_{C1} = V_{OH1} \cdot C_{u1} = V_{OH2}(C_{u1} + C_{i2}) \quad \Rightarrow \quad V_{OH2} = V_{OH1} \frac{C_{u1}}{C_{u1} + C_{i2}} \qquad (11.12)$$

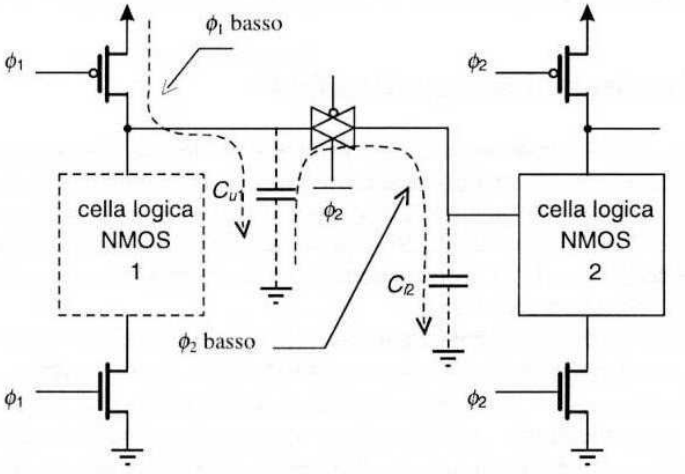


Figura 11.24 Ridistribuzione della carica nella logica a 2 fasi

Ciò implica che il livello logico alto viene degradato se la capacità di uscita C_{u1} non è elevata rispetto a quella di ingresso C_{i2}. Questo non succede quando l'uscita della cella 1 è uno 0 logico, in quanto in questo caso nella fase di valutazione vi è una via di conduzione nella rete NMOS, e questa rimane durante la fase di apertura della porta di trasmissione, per cui la capacità C_{i2} vede un circuito a bassa resistenza ohmica e si scarica anch'essa; quindi lo 0 logico viene ben trasmesso all'ingresso della porta 2. Il problema si complica quando la cella 1 debba pilotare più celle in uscita (ossia quando il fan-out della cella 1 è maggiore di 1), perché in tal caso la capacità a valle della porta di trasmissione 1-2 è la somma delle capacità di ingresso delle singole celle e la relazione di disuguaglianza $C_{u1} > \Sigma C_{i2}$ è più dif-

ficile da verificare. Occorre in tal caso verificare che la tensione al livello logico alto trasmessa alla capacità C_{i2} non scenda sotto il valore V_{IH} per mantenere i margini di rumore, ed in ogni caso questa non può scendere sotto il valore della tensione di soglia del NMOS pena il non corretto funzionamento della porta logica pilotata.

Per quanto riguarda la velocità di funzionamento di questa logica, confrontata con quella delle logiche FCMOS o pseudo-NMOS, occorre tener presente che nel cammino di scarica delle capacità di ingresso della porta successiva (e cioè del carico della cella precedente) vi sono, per una data funzione logica della cella, due transistori NMOS in più, quello della porta di trasmissione e quello verso massa; questo rallenta la transizione dal livello alto a quello basso. Inoltre il periodo dei segnali di fase, che in principio potrebbe essere poco superiore al tempo necessario per lo stabilizzarsi dei segnali agli ingressi delle celle e quindi paragonabile al ritardo di propagazione di una porta, deve tener conto in pratica del più lungo tra i ritardi di propagazione delle diverse porte logiche, e dell'inevitabile tolleranza sui legami di fase tra i due segnali; da ciò discende che la frequenza di funzionamento di queste logiche dinamiche è inferiore o circa uguale a quella delle porte statiche.

11.6 Logica dinamica a quattro fasi

L'inconveniente della degradazione dei livelli logici alti dovuta alla ridistribuzione delle cariche può essere eliminato utilizzando circuiti logici a quattro fasi. Questi richiedono una sequenza di quattro diversi segnali di fase che si sovrappongono parzialmente tra di loro durante le quattro operazioni elementari di 1) precarica del nodo di uscita della cella 1, 2) apertura della porta di trasmissione, 3) valutazione della cella 1, 4) valutazione della cella 2.

Il funzionamento di questa logica può essere esemplificato dallo schema di Figura 11.25. Con riferimento al diagramma temporale dei quattro segnali di fase, durante l'intervallo in cui ϕ_{12} è basso la capacità di uscita della cella 1 si precarica; questa rimane in connessione con l'alimentazione (precarica) anche durante la prima parte dell'intervallo ϕ_{23} in cui la porta di trasmissione si apre, per cui la capacità di ingresso della cella 2 viene anch'essa precaricata allo stato alto. Quando ϕ_{12} ritorna alto la cella 1 va in fase di valutazione, e durante la prima parte di questa fase la porta di trasmissione è ancora aperta, per cui se l'uscita della cella 1 si porta al livello basso, la capacità di ingresso della cella 2 si può scaricare. Quando ϕ_{23} ritorna alto la cella 2 va in fase di valutazione e la porta di trasmissione si chiude; quindi l'ingresso della cella 2 non viene interessato dalle eventuali variazioni che possono aversi all'uscita della cella 1.

La sovrapposizione dei segnali di precarica e di apertura della porta di trasmissione risolve il problema della ridistribuzione di cariche tra le capacità dei nodi di uscita e ingresso delle celle consecutive, anche nel caso di fan-out delle celle maggiore di uno. Tuttavia questa logica richiede una stretta sincronizzazione dei quattro segnali di fase lungo tutto il circuito e questo può non essere facile da realizzare, in presenza di cammini diversi dei segnali e di porte con differente ritardo di propaga-

zione; anche per questa logica vale quanto detto per la logica a due fasi sul periodo minimo dei segnali di fase, che deve essere significativamente più elevato del ritardo di propagazione di una singola porta logica.

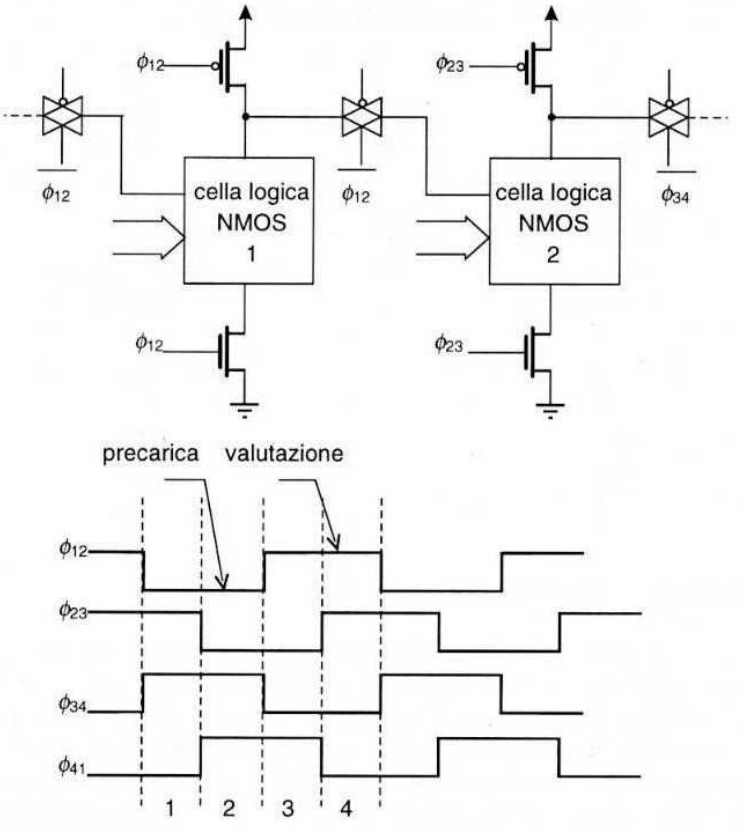


Figura 11.25 Schema di un circuito logico a quattro fasi

11.7 Logica dinamica Domino

La struttura logica dinamica detta "Domino" è una delle logiche dinamiche più utilizzate nei circuiti VLSI; essa deriva dalla logica ad una fase, in quanto utilizza una sola fase per sincronizzare tutte le celle logiche del circuito, ma con una modifica fondamentale che elimina l'inconveniente dovuto alla possibile scarica della capacità di ingresso nel transitorio di valutazione della cella a monte. Lo schema della cella elementare è riportato in Figura 11.26; in questo si ritrovano i transistori PMOS di precarica e NMOS di valutazione della singola cella logica, ma lungo la linea di connessione tra le celle è inserito un invertitore CMOS.

Il ruolo di questo invertitore essenzialmente è quello di invertire la transizione possibile all'uscita della cella in fase di valutazione, in modo da evitare i problemi visti nelle celle dinamiche ad una fase; inoltre esso separa l'uscita della cella precedente dagli ingressi delle altre celle, e fornisce un'elevata corrente di carica alle capacità di ingresso, migliorando quindi il fan-out della cella.

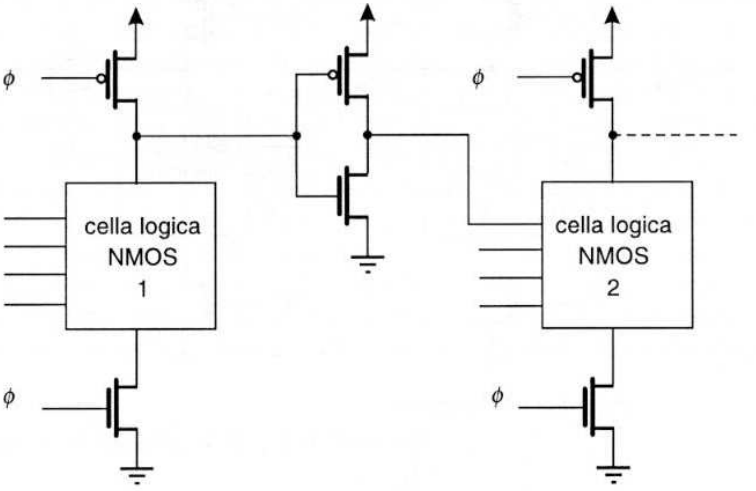


Figura 11.26 Circuito a logica dinamica Domino

Ricordiamo che nelle celle dinamiche a una fase (ma anche nelle successive), come si è visto nel Paragrafo 11.4, durante la fase di precarica l'uscita è nello stato alto, per cui nella fase di valutazione la transizione in uscita dalla cella è un passaggio condizionato dallo stato alto a quello basso, se i segnali logici agli ingressi sono tali da creare una via di conduzione nella rete di NMOS; se questa transizione non è molto veloce la fase di valutazione della cella successiva può essere alterata perché all'inizio del transitorio quest'ultima vede segnali ancora alti, mentre questi a regime dovrebbero essere bassi, e ciò può causare la scarica della capacità in uscita della cella a valle durante il transitorio di valutazione. Nella logica Domino la presenza dell'invertitore fa sì che i livelli inviati alla cella successiva siano invertiti rispetto a quelli sopra ricordati: durante lo stato di precarica, le celle successive vedono un segnale *basso*, e quindi gli NMOS sono aperti; durante la valutazione, l'uscita dell'invertitore (e quindi gli ingressi delle celle successive) non può che passare *dal livello basso a quello alto* (nel caso ovviamente che la cella precedente presenti un'uscita bassa per effetto dei segnali agli ingressi), mentre non è mai possibile la transizione dall'alto al basso.

Questo comporta che tutte le celle in un circuito Domino possono essere pilotate dallo stesso segnale di fase ϕ, in quanto finché gli ingressi non effettuano delle transizioni (da 0 a 1) le celle rimangono nello stato alto; gli eventuali transitori nelle transizioni delle uscite non comportano altro che un ritardo corrispondente nell'eventuale passaggio allo stato basso delle celle pilotate da queste uscite. In

effetti durante la fase di valutazione (comune a tutte le celle) le celle in cascata passano una dopo l'altra dallo stato alto a quello basso (condizionatamente agli ingressi) con un ritardo pari al transitorio della transizione da 0 a 1, analogamente alle tessere del gioco del domino che cadono l'una dopo l'altra, ognuna spinta dalla caduta della precedente, da cui il nome della logica.

Il vantaggio evidente della logica Domino è quello di fare uso di un solo segnale di fase, con notevole semplificazione dei circuiti di sincronizzazione e di ritardo. L'uso di un solo segnale di fase riduce ai soli tempi di transizione il ritardo delle celle, eliminando i tempi morti dovuti alla successione di più fasi dei circuiti precedenti. Inoltre la presenza dell'invertitore CMOS migliora i margini di rumore e fornisce una corrente più elevata per il pilotaggio delle capacità di ingresso delle porte a valle, permettendo quindi un più elevato fan-out.

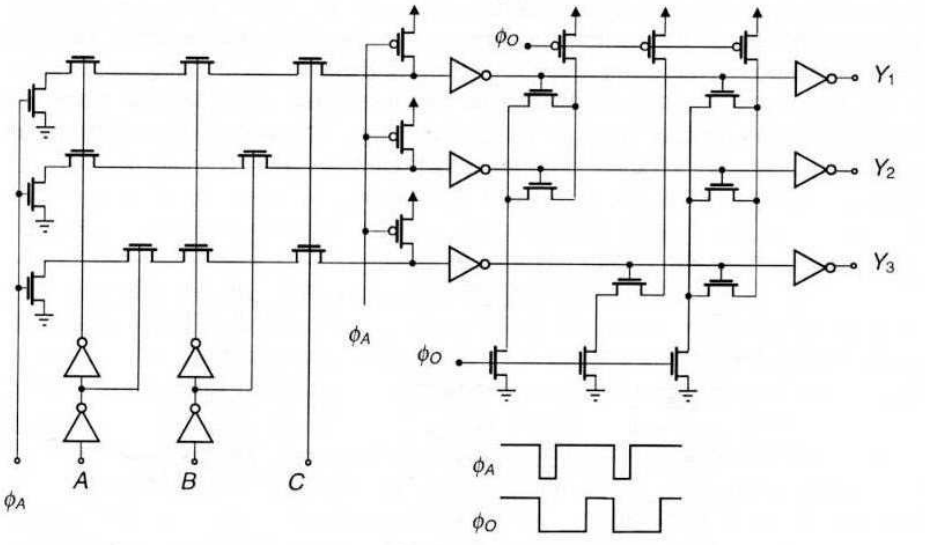


Figura 11.27 PLA a 3 ingressi, 3 termini di prodotti e tre uscite, realizzato con logica Domino

L'inconveniente di questa logica è quello di non fornire l'inversione del segnale logico all'uscita della singola cella, a causa della presenza dell'invertitore aggiuntivo; in tal modo si possono avere solo funzioni AND e OR (e le loro combinazioni), ma non funzioni NAND o NOR, il che non permette di realizzare con questo circuito qualsiasi espressione logica. L'inversione delle variabili deve essere quindi effettuata prima di entrare in una catena di celle Domino, o alla fine della stessa, ma non tra una cella e l'altra. La logica Domino è quindi particolarmente adatta a realizzare quelle funzioni logiche che non richiedono inversione in uscita: un esempio è quello delle strutture PLA viste nel Paragrafo 10.9, che richiedono i due piani AND e OR per effettuare le funzioni somma di prodotti; in questo caso

non è richiesta l'inversione in uscita dai due piani, e le due funzioni AND e OR sono direttamente implementate nelle celle Domino.

In Figura 11.27 è riportato come esempio lo schema circuitale di un PLA con 3 ingressi, 3 termini di prodotto e 3 termini di somme, realizzato con logica Domino. I segnali di fase ϕ_A e ϕ_O rispettivamente per il piano AND e quello OR, che in principio possono essere un unico segnale, sono di diversa durata per il valore basso (fase di precarica), in modo che la valutazione del piano AND preceda temporalmente quella del piano OR e permetta il completamento della fase di valutazione del primo anteriormente a quella del secondo.

11.8 Logica NORA CMOS

Una struttura logica derivata dalla logica Domino e che permette l'inversione delle variabili in uscita è quella detta NORA da "*no race*" in quanto questa logica non ha problemi di "corsa" o transitori degli ingressi come quelli dei circuiti a una fase, pur eliminando l'invertitore presente nella logica Domino.

Il circuito utilizza alternativamente per le singole celle logiche transistori NMOS e PMOS, per cui viene chiamata anche logica "Domino NP", e il funzionamento di questa logica è schematizzato nella Figura 11.28.

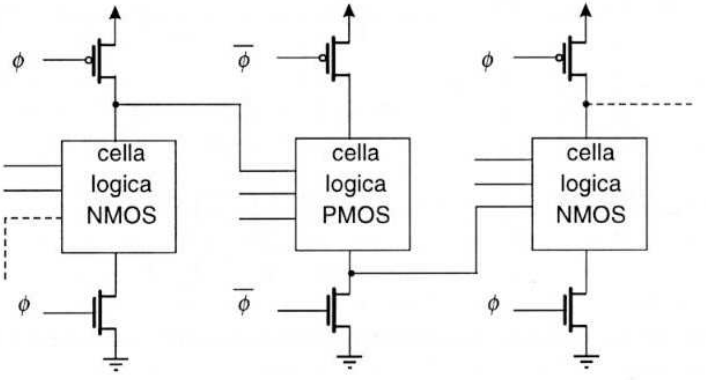


Figura 11.28 Logica Domino modificata, o NORA

La prima cella contiene una rete NMOS e quindi in fase di precarica la sua uscita è alta. La cella successiva è realizzata con transistori PMOS ed il segnale di fase è invertito rispetto a quello della cella precedente, per cui nella fase di precarica il NMOS conduce e il PMOS è aperto; in effetti per la cella P la fase di precarica è svolta dal NMOS e quella di valutazione dal PMOS. L'uscita della seconda cella in precarica è quindi bassa, e durante la fase di valutazione può portarsi condizionatamente alta se tutti gli ingressi (della rete PMOS) sono bassi. Il funzionamento è analogo a quello della logica Domino, con la variante che le celle sono alternativa-

mente a precarica bassa e alta, e al posto dell'inversione dell'uscita vengono utilizzati MOS alternativamente P e N che di fatto invertono le variabili di ingresso.

Anche in questa logica le singole uscite effettuano al più un'unica transizione che porta sempre i MOS della cella successiva ad andare dall'interdizione alla conduzione, eliminando così i problemi della scarica delle capacità della cella a valle durante i transitori. Il vantaggio di questa logica è quello di permettere espressioni logiche con inversione delle variabili in uscita e quindi di realizzare qualunque funzione logica, al contrario della logica Domino. L'inconveniente è quello di usare per le celle P reti di transistori PMOS, che sono, come si è visto, intrinsecamente più lenti a parità di area occupata. Inoltre, non essendovi lo stadio invertitore, i margini di rumore sono più bassi.

11.9 Confronto tra le logiche dinamiche

I circuiti logici dinamici hanno un significativo campo di applicazione nei circuiti integrati ad alta densità di integrazione, in quanto come si è visto richiedono un minor numero di transistori elementari rispetto alle versioni a logica pienamente complementare FCMOS.

Vi sono anche altri aspetti che differenziano le prestazione di queste logiche da quelle statiche. In particolare, per quanto riguarda i margini di rumore, le logiche dinamiche hanno un valore V_{IL} minore di quello delle porte FCMOS, in quanto la capacità di uscita della cella comincia a scaricarsi quando gli ingressi superano il valore V_T di soglia dei MOS, mentre per le porte FCMOS $V_{IL} \cong V_{DD}/2$. La riduzione del valore di soglia logica d'altra parte comporta una riduzione dei tempi di propagazione della cella dinamica, in quanto la porta inizia a commutare quando gli ingressi raggiungono il valore V_T e quindi prima dell'istante in cui raggiungono $V_{DD}/2$.

Per quanto riguarda l'occupazione di area, le logiche statiche FCMOS a m ingressi richiedono $2m$ transistori, mentre quelle dinamiche richiedono m transistori per la cella logica, più 2 transistori di precarica e di valutazione, più eventualmente altri 2 transistori per la porta di trasmissione o per l'invertitore.

Infine, per quanto riguarda i tempi di propagazione, occorre considerare diversamente le due transizioni nelle differenti logiche. Il passaggio da 0 a 1 dell'uscita per le celle dinamiche (e per le celle N della logica NORA) è effettuato durante la fase di precarica e quindi non è legato all'arrivo dei segnali in ingresso, per cui non è possibile valutare il tempo di propagazione t_{PLH}. Per le logiche pseudo-NMOS, la riduzione della capacità di uscita (corrispondente all'ingresso della cella successiva) rispetto alla logica FCMOS è bilanciato dalla riduzione della corrente di carica perché parte di questa viene assorbita dalla rete NMOS. Per quanto riguarda il passaggio da 1 a 0, nelle logiche dinamiche a 2 o 4 fasi la capacità di uscita è equivalente a quella del caso pseudo-NMOS ma la scarica della capacità dipende anche dal numero dei MOS connessi in serie nella via verso la massa, numero che è maggiorato (rispetto alla logica FCMOS e pseudo-NMOS) dal NMOS di valutazione e da quello della

porta di trasmissione; nel caso della logica Domino, la capacità di uscita della cella (ingresso dell'invertitore) è più elevata, ma si elimina un NMOS nel percorso di scarica.

Tabella 11.2 Confronto tra le diverse logiche MOS per celle NAND o NOR a m ingressi

logica	*soglia logica*	t_{PLH}	t_{PHL}	*numero MOS*
FCMOS	$\cong V_{DD}/2$			$2m$
NAND		5τ	$2m\tau$	
NOR		$5m\tau$	2τ	
pseudo-NMOS	$>V_T$			$m+1$
NAND		2.5τ	$m\tau$	
NOR		2.5τ	τ	
dinamica (2-4 fasi)	V_T			$m+4$
NAND		-	$(m+2)\tau$	
NOR		-	3τ	
domino	V_T			$m+4$
NAND		-	$(m+1)2\tau$	
NOR		-	4τ	

Queste caratteristiche sono sintetizzate nella Tabella 11.2 che confronta le prestazioni salienti delle diverse logiche esaminate, assumendo tutti i transistori, sia NMOS che PMOS, di area unitaria ($W = L$ = minima dimensione). I tempi di transizione delle diverse logiche sono stati valutati come multipli di un tempo di propagazione unitario τ, legato alla scarica o carica della capacità unitaria C_G attraverso un NMOS di area minima, e assumendo la condizione più sfavorevole per le porte NOR (uno solo dei MOS in conduzione). Quindi per le logiche FCMOS il carico capacitivo è doppio che per le logiche pseudo-NMOS e dinamiche, mentre la carica della capacità unitaria attraverso un PMOS comporta un tempo 2.5 volte maggiore che la scarica attraverso un NMOS. Vedremo ulteriori applicazioni delle logiche dinamiche nei circuiti sequenziali; inoltre le celle logiche dinamiche saranno largamente utilizzate nelle memorie RAM sia per i sottosistemi di indirizzamento che per le celle di memoria elementari.

Esercizi di riepilogo

11.1 Disegnare lo schema elettrico di una porta complessa CMOS che realizzi in uscita la funzione $Y = \overline{A \cdot B + C \cdot (A \cdot E + F)}$ per il circuito realizzato, assumendo di dimensionare tutti gli NMOS con un rapporto $W/L = 2$, e tutti i PMOS con un rapporto $W/L = 5$, definire la combinazione di variabili che determina: a) il valore più grande di t_{PLH}, b) il valore più grande di t_{PHL}.

11.2 Determinare i tempi di propagazione t_{PLH} e t_{PHL} nel caso più sfavorevole per la porta logica complessa di Figura 11.3, supponendo l'uscita caricata dall'invertitore di riferimento con $W/L|_N$ = 1μm/1μm, $W/L|_P$ = 2.5 μm/1μm.

11.3 Determinare la rete PMOS della porta complessa CMOS di Figura E11.1, assegnata la rete NMOS. Che funzione logica viene realizzata con questa porta?

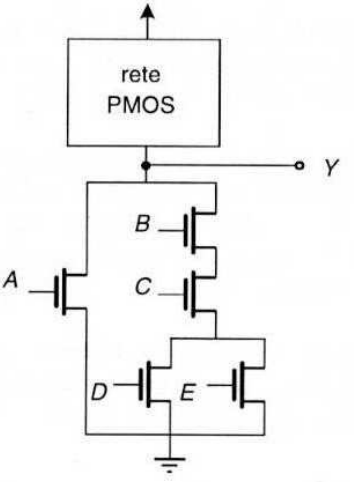


Figura E11.1

11.4 Disegnare il tracciato della porta complessa di Figura E11.1 costruendo un percorso di Eulero sul grafo ad archi della porta per la realizzazione del numero minimo di interruzioni tra le connessioni contigue dei transistori NMOS e PMOS.

11.5 Dimensionare i transistori della porta pseudo-NMOS di Figura 11.6 in modo da avere nel caso più sfavorevole (ingressi applicati alla rete NMOS tali da dar luogo alla minore corrente di scarica per la capacità di uscita) una corrente $I_N = 2I_P$. Valutare, sempre in questo caso: a) i tempi di propagazione t_{PLH} e t_{PHL} assumendo come carico un invertitore di riferimento pseudo-NMOS con $W/L|_N$ = 1μm/1μm; b) il valore della V_{OL}.

11.6 Per la porta di trasmissione NMOS di Figura E11.2, determinare il valore della tensione V_O in uscita a regime, assumendo per il NMOS i valori: V_{T0} = 0.7 V, ϕ^* = 0.6 V, γ = 0.5 $V^{1/2}$.

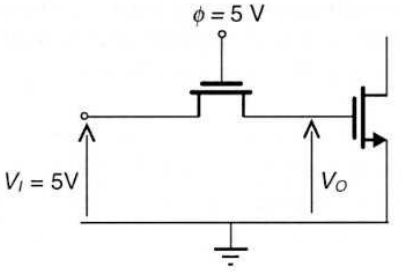


Figura E11.2

11.7 Per il multiplexer con porte NMOS di Figura 11.13a, determinare il livello logico alto del segnale di uscita Y assumendo per i MOS i valori dei parametri dell'Esercizio 11.6, e per $A = B = C = D = 5$ V, $A_0 = A_1 = 5$ V.

11.8 Dimensionare il transistore NMOS della porta di Figura E11.2, considerando come carico un NMOS ad area minima, in modo da avere, nel passaggio del segnale di fase ϕ dal valore alto (5 V) a quello basso (0 V), un salto di tensione ΔV_O in uscita inferiore a 1 V. Verificare il risultato mediante una simulazione SPICE del circuito così dimensionato.

11.9 Disegnare lo schema elettrico di un multiplexer a 8 ingressi con porte di trasmissione CMOS e struttura ad albero. Quanti transistori si risparmiano rispetto allo schema standard analogo a quello di Figura 11.13?

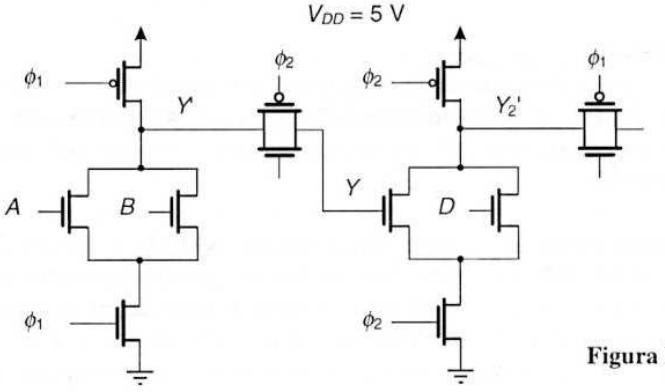


Figura E11.3

11.10 Per il circuito a porte NOR di Figura E11.3 realizzato con logica dinamica a due fasi, determinare, per il caso $A = B = 0$, il valore a cui si porta la tensione V_O all'uscita Y quando la porta di trasmissione ϕ_2 si apre, in seguito alla ridistribuzione della carica tra Y' e Y; si assuma che tutti i transistori NMOS abbiano uguali dimensioni e un rapporto $W/L = 2\lambda/1\lambda$, con $\lambda = 1$ μm, e si utilizzino le regole di progetto del Capitolo 2 per il dimensionamento dei drain. Si valuti infine l'effetto che ha la variazione della tensione V_Y sul tempo di

propagazione t_{PHL} all'uscita Y_2', nel caso $D = 1$, paragonandolo a quello che si avrebbe se V_Y rimanesse al valore ideale V_{DD}.

11.11 Con riferimento allo schema di PLA di Figura 11.27, identificare la combinazione di ingressi e il percorso che determina il maggior ritardo di propagazione complessivo nella fase di valutazione.

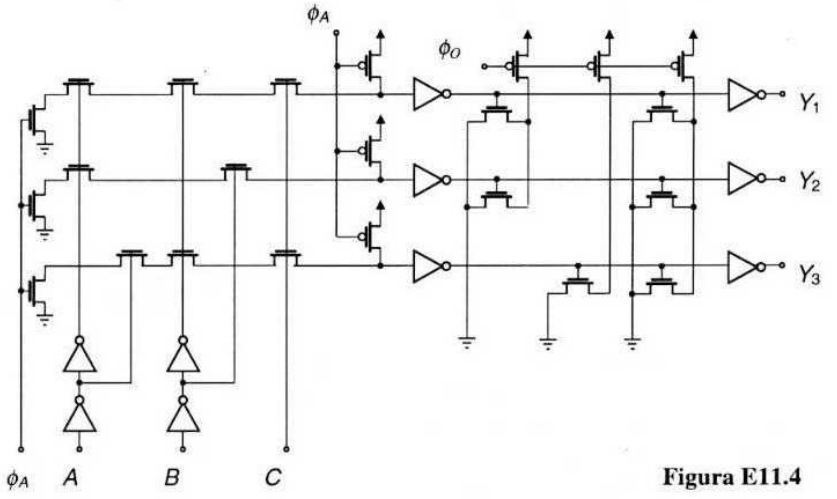


Figura E11.4

11.12 Il circuito di Figura E11.4 è una versione modificata del PLA di Figura 11.27, nella quale sono stati eliminati i transistori NMOS di valutazione nel piano OR. Cosa comporta questa modifica nel funzionamento logico del circuito? Quale è il vantaggio in termini di ritardo di propagazione complessivo nella fase di valutazione?

Riferimenti bibliografici

B. Riccò, F. Fantini, P. Brambilla, *Introduzione ai circuiti integrati digitali*, Zanichelli, Bologna, 1991.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

N.E. Weste, K. Eshraghian, *Principles of CMOS VLSI Design - A system perspective*, 2nd ed., Addison-Wesley, 1993.

12

Circuiti sequenziali

12.1 Introduzione

I circuiti esaminati nei capitoli precedenti sono detti *combinatori*, perché in ogni istante le loro uscite dipendono da un'appropriata combinazione degli ingressi presenti in quegli stessi istanti. I circuiti *sequenziali* sono invece circuiti logici le cui uscite dipendono non solo dagli ingressi presenti in quel momento, ma anche dalla storia passata degli ingressi, in altre parole da una data sequenza (corta o lunga) di eventi logici che si sono succeduti nel tempo agli ingressi.

Questi circuiti sono essenziali per creare le cosiddette *macchine a stati finiti*, che permettono di generare un insieme di stati logici possibili (e di uscite) a partire da un'opportuna sequenza di ingressi, e quindi di eseguire una sequenza di elaborazioni dipendenti dalla serie di "istruzioni" (ingressi) ricevute. Un esempio di macchina a stati finiti è quella riportata in Figura 12.1, detta di Mealy, che utilizza in uscita da una generica rete logica combinatoria una rete di memoria, la quale memorizza gli stati logici precedenti e li rimanda in ingresso in modo che l'uscita dipenda sia dagli ingressi correnti che da quelli precedenti.

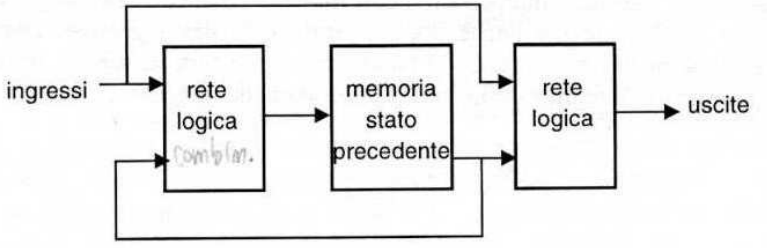


Figura 12.1 Struttura di una macchina a stati finiti

I circuiti sequenziali, per la definizione datane, e cioè di presentare degli stati logici in uscita dipendenti anche da grandezze logiche applicate in tempi precedenti a quello in esame, debbono necessariamente contenere degli elementi di *memoria*, cioè circuiti capaci di "ricordare" lo stato logico precedente e di conservarne gli effetti.

Gli elementi di memoria possono essere realizzati basicamente in due modi differenti:

a. utilizzando circuiti che riportano, mediante una reazione, i segnali in uscita di nuovo in ingresso in modo tale che questi ultimi autosostentino le uscite stesse (reazione positiva degenerativa);
b. utilizzando un elemento circuitale capace di incamerare temporaneamente l'informazione per mezzo di un'opportuna grandezza elettrica (ad esempio carica accumulata in un condensatore).

In entrambi i casi questi componenti debbono presentare due possibili stati stabili, in dipendenza della memorizzazione di uno dei due livelli logici possibili (uscita alta o bassa per i circuiti con reazione, carica immagazzinata alta o bassa per la capacità), per cui vengono detti *circuiti bistabili* (in termine inglese *latch* = chiavistello, che può assumere due posizioni, aperto o chiuso, e rimane in ognuna di queste posizioni dopo l'evento di apertura o chiusura effettuato).

Nei paragrafi seguenti verranno studiati i circuiti *bistabili* (o *latch)* che impiegano reti di reazione, e i circuiti sequenziali su questi basati. Successivamente verranno analizzati i circuiti sequenziali basati sull'utilizzo delle capacità come elementi di memoria, circuiti che utilizzano le celle logiche dinamiche MOS già introdotte nel Capitolo 11.

12.2 Circuiti bistabili

Per ottenere da un circuito reazionato un comportamento bistabile, cioè la possibilità che l'uscita possa permanere in uno o l'altro dei possibili livelli logici anche in assenza di un segnale in ingresso, occorre che: 1) il circuito sia capace di amplificare (ossia rigenerare) il segnale nel passaggio dall'ingresso all'uscita; 2) la reazione sia tale da portare il circuito nel campo dell'instabilità, che per definizione corrisponde al caso in cui l'uscita è indipendente dall'ingresso (o anche al caso in cui l'uscita permane indefinitamente anche dopo la rimozione dell'ingresso). Dalla teoria della reazione negli amplificatori, si ha che l'uscita a piccoli segnali di un amplificatore reazionato in regime di linearità è descritta dalla relazione:

$$dX_0 = dX_I \frac{A}{1 - A\beta} \tag{12.1}$$

dove A è la funzione di trasferimento a piccoli segnali del circuito amplificatore (ossia l'amplificazione) e β quella della rete di reazione, indicate nello schema a blocchi di Figura 12.2.

La condizione di instabilità si ha quando la variazione del segnale di ingresso dX_I, amplificata del guadagno di anello $A\beta$ (prodotto dell'amplificazione A nel passaggio attraverso l'amplificatore e dell'eventuale attenuazione β attraverso la rete di reazione) e riportata in ingresso con il valore $dX_F = A\beta dX_I$, è maggiore in modulo di quella di partenza dX_I; in questo caso l'uscita diverge via via dalla condizione iniziale portando il circuito in uno dei due possibili stati limite, interdizione o saturazione. La condizione matematica per l'instabilità è data dalla disequazione $A\beta>1$ e il raggiungimento di una delle due condizioni limite è tanto più veloce quanto più forte è la disuguaglianza. Ciò comporta: a) che la rete di reazione attenui il meno possibile, e b) che l'amplificatore abbia una funzione di trasferimento A positiva e ben maggiore dell'unità nel tratto lineare.

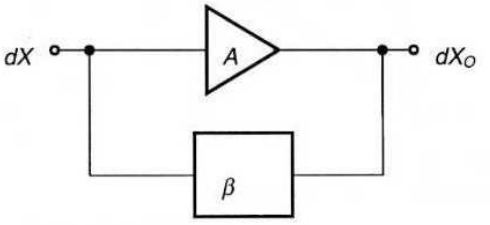


Figura 12.2 Schema a blocchi di amplificatore reazionato, per le componenti variabili di ingresso e di uscita

Per il punto a) la rete di reazione è sostituita da un collegamento diretto (che implica $\beta = 1$), mentre per il punto b) l'amplificazione positiva del segnale può essere ottenuta ponendo in cascata due stadi amplificatori invertenti. Questi possono essere realizzati mediante circuiti invertitori, come è indicato in Figura 12.3a; questi infatti, come si è visto nello studio della funzione di trasferimento, presentano nel tratto a pendenza negativa un valore di amplificazione maggiore dell'unità, per cui il prodotto delle due amplificazioni (e cioè la pendenza della funzione di trasferimento complessiva) sarà positiva ed elevata.

Il circuito di Figura 12.3a può essere ridisegnato come in Figura 12.3b, in modo da evidenziare le grandezze presenti all'ingresso e all'uscita di ognuno dei due invertitori. In Figura 12.3c è tracciata la costruzione grafica che riporta sulla caratteristica di trasferimento (X_{I1}, X_{O2}) della serie dei due invertitori (detta ad anello aperto perché ricavata nell'ipotesi di aprire la connessione di reazione tra il punto X_{O2} e quello X_{I1}), il vincolo ulteriore del collegamento tra uscita ed ingresso dato da $X_{I1} = X_{O2}$, definito da una retta a 45° nel piano X_{I1}, X_{O2}. La combinazione delle due curve definisce quindi la funzione di trasferimento a ciclo chiuso, o dell'anello di reazione; le tre intersezioni (A, B, C) dei due vincoli costituiscono i possibili punti di funzionamento stazionario del sistema.

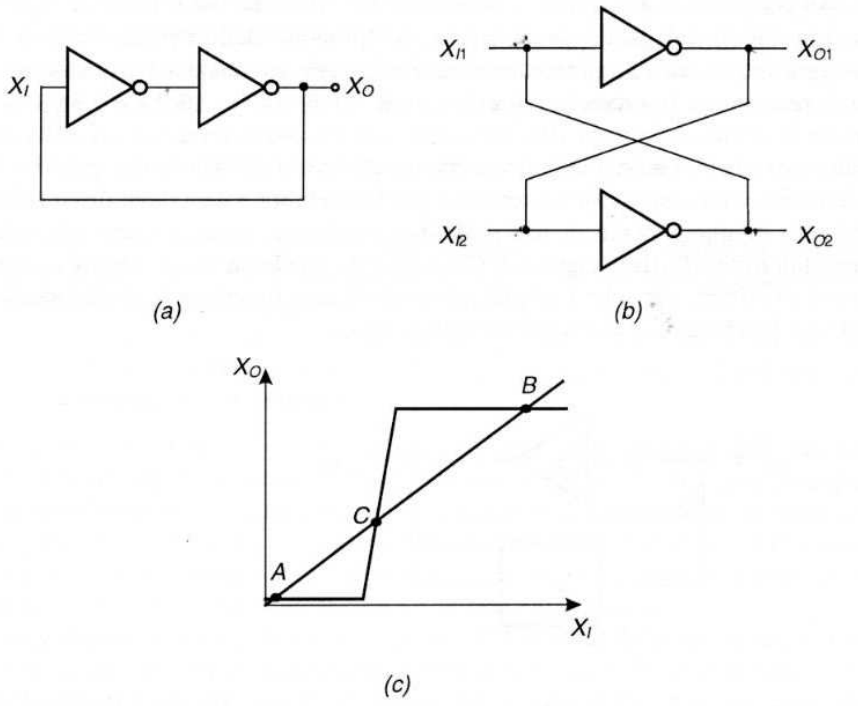


Figura 12.3 a) Schema elementare di un circuito bistabile; b) Schema alternativo del bistabile; c) Punti stabili (A, B) e instabili (C) sulla funzione di trasferimento a ciclo chiuso

Di queste intersezioni i punti A o B rappresentano due possibili situazioni *stabili* del sistema, in quanto in entrambi i punti la pendenza della funzione di trasferimento a ciclo aperto è nulla e quindi si ha un'amplificazione complessiva $A = 0$. Una eventuale perturbazione del regime di equilibrio dovuta a un disturbo di piccola entità porterà quindi, in base alla (12.1) e alla teoria della reazione, ad un'evoluzione in uscita rapidamente smorzata intorno al punto di funzionamento, in quanto, nel nostro caso, si ha per il circuito:

$$dX_O = dX_i \frac{A}{1 - A} \tag{12.1a}$$

ed in base alla (12.1a) $dX_O \to 0$ se $A = 0$ (punti A e B).

Il circuito quindi può permanere stabilmente in ognuno dei due punti di funzionamento A o B in cui è stato portato in precedenza, e per questo è detto *circuito bistabile*. Il punto di funzionamento C, che appartiene al tratto a pendenza elevata della caratteristica di trasferimento a circuito aperto, è invece un punto *instabile*,

perché in questo punto, sempre in base alla (12.1a), si ottiene per la variazione dell'uscita $dX_O \to \infty$ in quanto $A \geq 1$ (condizione di instabilità), e il punto di funzionamento si allontana indefinitamente dal punto C per raggiungere uno dei due punti stabili A o B, a seconda di quale sia stato il senso della perturbazione iniziale applicata nel punto C.

Quindi il punto C non può essere un punto di funzionamento stazionario del circuito; quest'ultimo, anche se portato inizialmente in C, dopo un transitorio si porterà a regime in uno dei due punti stabili A o B. I due punti stabili A e B corrispondono a due possibili livelli logici, che quindi costituiscono i due possibili *stati,* del circuito; in particolare, per il circuito di Figura 12.3 il punto A corrisponde all'uscita del secondo invertitore nello stato logico basso (0 logico), e quello B allo stato logico alto (1 logico). Il circuito può mantenere indefinitamente ciascuno dei due livelli, una volta portato in questo da una sollecitazione esterna, quindi esso è capace di *memorizzare* l'informazione logica applicata mediante la conservazione dello stato corrispondente: ad esempio l'uscita alta del secondo invertitore (o quella bassa del primo) corrisponderà per convenzione alla memorizzazione di un 1 logico, mentre l'uscita bassa del secondo (o quella alta del primo) corrisponderà alla memorizzazione di uno 0 logico.

12.3 Il bistabile *SR*

La rappresentazione del bistabile di Figura 12.3b rende evidenti le connessioni tra le due possibili uscite X_{O1}, X_{O2} e gli ingressi X_{I1}, X_{I2}. Da queste connessioni risulta chiaro che le due uscite non possono essere contemporaneamente alte o basse, e che il livello alto ad una delle due uscite implica automaticamente quello basso all'altra. Da questo schema risulta peraltro evidente che ognuna delle due uscite si trova in connessione diretta con un ingresso, e questo rende impossibile l'applicazione di un segnale logico ad un ingresso, perché questo, anche se tale da non alterare lo stato memorizzato, viene in ogni caso trasmesso in uscita.

Per separare i terminali delle uscite (sulle quali vanno letti gli stati logici del circuito) da quelli degli ingressi (ai quali vanno applicati i segnali logici che debbono essere memorizzati dal circuito) si può sostituire ognuno dei due invertitori (porte NOT) con porte NOR a due ingressi, come indicato in Figura 12.4.

In questo caso gli stati delle due uscite (indicati nel seguito con i simboli Q e $\overline{Q}$) dipenderanno sia dai segnali riportati in reazione su uno dei due ingressi della porta, sia da quelli applicati all'altro dei due ingressi (indicati con S da "*set*" e R da "*reset*"); a questi ultimi si possono quindi applicare gli ingressi opportuni per "inserire" l'informazione logica da memorizzare nel circuito.

Il circuito così realizzato prende il nome di *bistabile SR* (*latch SR*), ed è in generale definito nel suo funzionamento dalla tabella della verità riportata in Tabella 12.1.

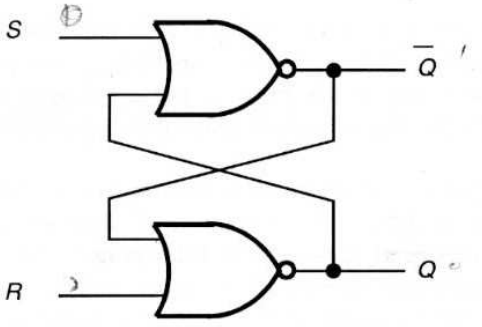


Figura 12.4 Bistabile *SR* con porte NOR

Tabella 12.1 Tabella della verità del bistabile *SR*

stato precedente		*periodo ΔT di applicazione S o R*				*stato successivo*	
Q_n	$/Q_n$	R	S	Q	$/Q$	Q_{n+1}	$/Q_{n+1}$
1	0	0	0	1	0	1	0
0	1	0	0	0	1	0	1
1	0	1	0	0	1	0	1
0	1	0	1	1	0	1	0
0	1	1	0	0	1	0	1
1	0	0	1	1	0	1	0
1	0	1	1	0	0	?	?
0	1	1	1	0	0	?	?

Nella tabella della verità del circuito (Tabella 12.1) si suppone che gli ingressi S o R assumano i valori logici indicati per un dato intervallo di tempo ΔT; lo stato delle uscite precedente all'applicazione di un segnale agli ingressi è indicato con il pedice n e quello successivo a tale applicazione con il pedice $n+1$, l'uscita negata è indicata con $/Q$. Dalla tabella si vede che, se entrambi i segnali S e R si portano (o rimangono) al livello logico basso (0), le due uscite mantengono lo stato logico che avevano precedentemente all'applicazione dei segnali; se l'ingresso S si porta al livello alto (1) si ha $Q_{n+1} = 1$ qualunque sia lo stato precedente delle uscite, e questo stato permane anche quando S ritorna a 0; se invece $R = 1$ si ha il comportamento duale, e $Q_{n+1} = 0$. Si giustifica così l'indicazione di "set" (assegnare) all'ingresso S che, portandosi alto, assegna il valore logico 1 all'uscita Q (e quindi per convenzione memorizza un 1 nella memoria binaria del bistabile) e quella di "reset" (riazzerare) all'ingresso R che, portandosi alto, riporta l'uscita Q a 0 e quindi inserisce nella memoria uno 0 (annullando eventualmente il valore 1 memorizzato). Infine

dalla tabella emerge una situazione anomala nel caso che entrambi gli ingressi S e R siano alti: in questo caso durante l'applicazione dei segnali entrambe le uscite sono basse (porte NOR), il che sembra contraddire l'osservazione fatta precedentemente di uscite complementari negli stati stabili. In effetti questa situazione non corrisponde a uno degli stati stabili ma è transitoria perché ottenuta in presenza di segnali alti ad entrambi gli ingressi, e non è destinata a durare anche in assenza degli ingressi. Tuttavia a livello di schema logico non si può prevedere quale sia la situazione finale, per cui la presenza contemporanea di segnale alto su S e R deve essere evitata e le uscite corrispondenti non vengono utilizzate (ritorneremo su questo punto nell'esaminare il comportamento temporale del circuito SR).

In effetti la tabella della verità della Tabella 12.1 è ridondante e può essere compressa nella Tabella 12.2 in quanto i valori delle uscite nello stato precedente non interessano nei casi $S = 1$ o $R = 1$, perché in ogni caso il valore dell'uscita non dipende da quegli stati, mentre per $S = R = 0$ in ogni caso le uscite rimangono quelle dello stato precedente.

Tabella 12.2 Tabella della verità sintetica del bistabile SR

R	S	Q_{n+1}	$/Q_{n+1}$
0	0	Q_n	$/Q_n$
0	1	1	0
1	0	0	1
1	1	?	?

Una differente schematizzazione del comportamento del bistabile SR è quella che fa riferimento ai diagrammi di temporizzazione dei segnali di ingresso e di uscita del circuito, riportata in Figura 12.5.

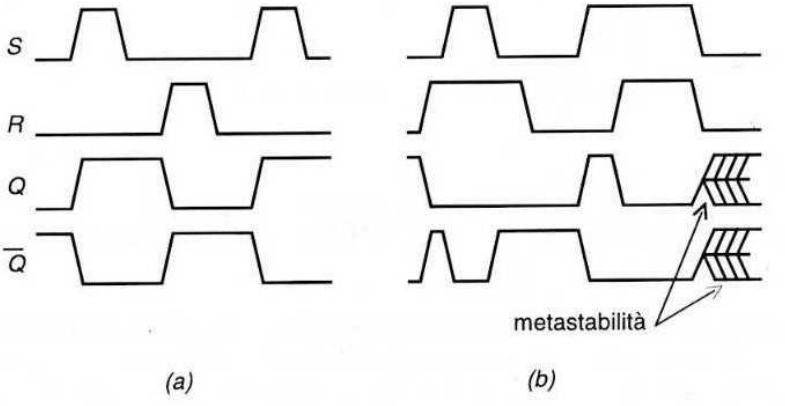


Figura 12.5 Temporizzazioni dei segnali in un bistabile SR; a) ingressi normali; b) situazioni ambigue quando $S = R = 1$

Da questa risulta chiaro il comportamento del bistabile, che si porta in uscita (Q) nello stato alto o basso, a seguito rispettivamente dell'applicazione di un segnale alto di set S o di reset R (Figura 12.5a), e rimane nella situazione precedente anche dopo l'applicazione dei segnali. Le situazioni ambigue o indesiderate sono indicate nella Figura 12.5b, e corrispondono a segnali S e R entrambi alti (e quindi ad uscite Q e $\overline{Q}$ temporaneamente basse).

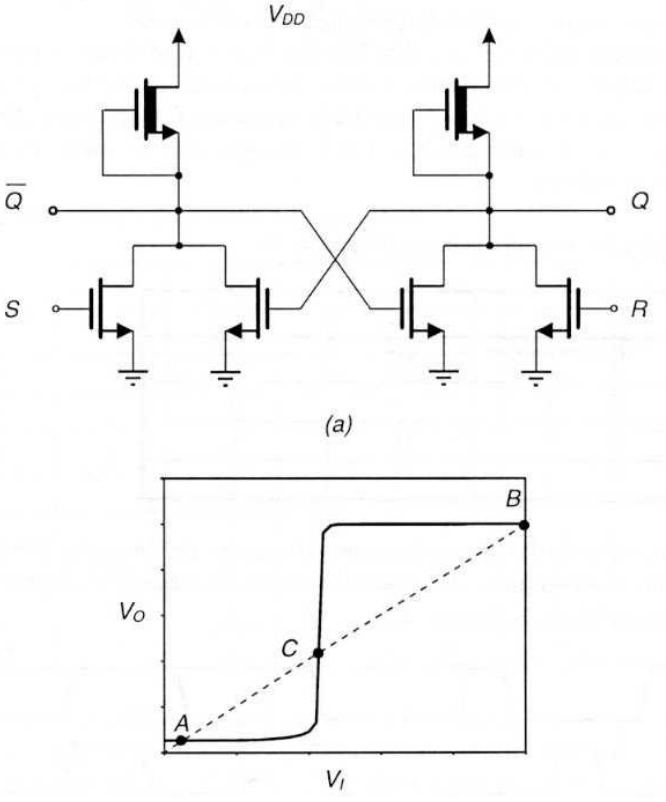


Figura 12.6 a) Bistabile con porte NOR NMOS; b) caratteristica di trasferimento del circuito

Se gli istanti in cui S e R ritornano allo stato basso sono ben separati tra loro, il circuito memorizza la situazione corrispondente all'annullamento del più lungo degli ingressi alti (ad esempio, se R passa al livello basso ben dopo S, l'uscita Q rimane bassa e quella $\overline{Q}$ passa allo stato alto, come se fosse stato applicato solo un ingresso R). Se invece sia S che R ritornano contemporaneamente allo stato basso,

sia l'uscita Q che quella $\overline{Q}$ si portano in un uno stato intermedio, dal quale escono per portarsi in due stati logici complementari non predicibili a priori. Questo comportamento è detto *comportamento metastabile*.

Per chiarire meglio questi differenti comportamenti del bistabile *SR*, faremo riferimento, come esempio specifico, allo schema circuitale di un bistabile *SR* in tecnologia NMOS con porte NOR, indicato in Figura 12.6a. La caratteristica di trasferimento del circuito è riportata in Figura 12.6b, nella quale sono indicati i due punti stabili di funzionamento A, B, e quello instabile C.

Per questo circuito è stata effettuata un'analisi dinamica, in un primo caso (Figura 12.7) al variare del livello del segnale S, e per un secondo caso (Figura 12.8) al variare della durata dello stesso segnale, mantenendo in questo caso il livello logico alto (5 V).

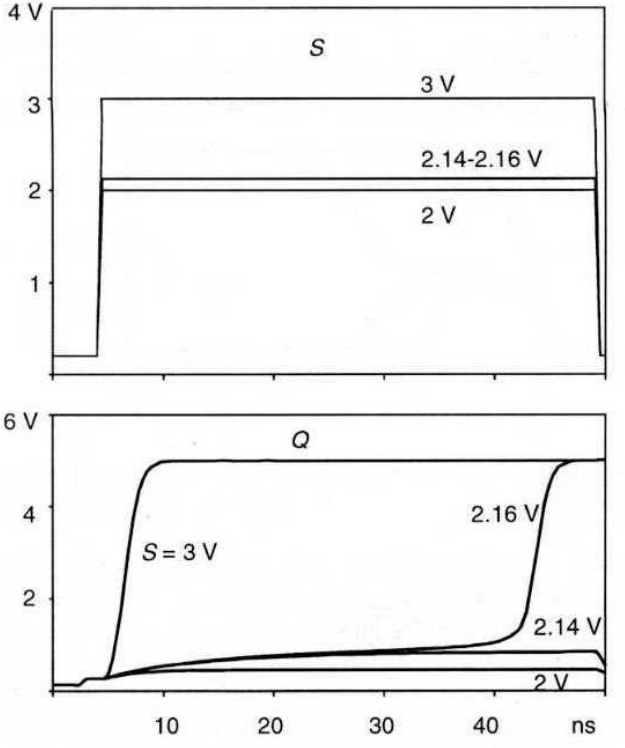


Figura 12.7 Simulazioni SPICE del comportamento dinamico del bistabile di Figura 12.6 al variare dell'ampiezza dell'ingresso S

Nel primo caso (Figura 12.7) si vede che occorre superare una certa soglia per portare l'uscita nello stato alto. In particolare per livelli molto prossimi alla soglia il circuito si porta in uno stato intermedio, detto appunto *metastabile*, di durata im-

precisata, e tanto più lunga quanto più si è vicini al valore della soglia; da questo stato il circuito esce per portarsi nello stato stabile basso o alto in maniera non prevedibile a priori. Il livello di soglia sull'ingresso è legato al valore della tensione V_t(C) dell'ingresso che corrisponde al punto instabile *C*: solo se si supera questo valore il sistema evolve verso l'altro punto stabile (*B*), altrimenti si ricade nel punto di partenza (*A*) e quindi l'uscita non cambia stato.

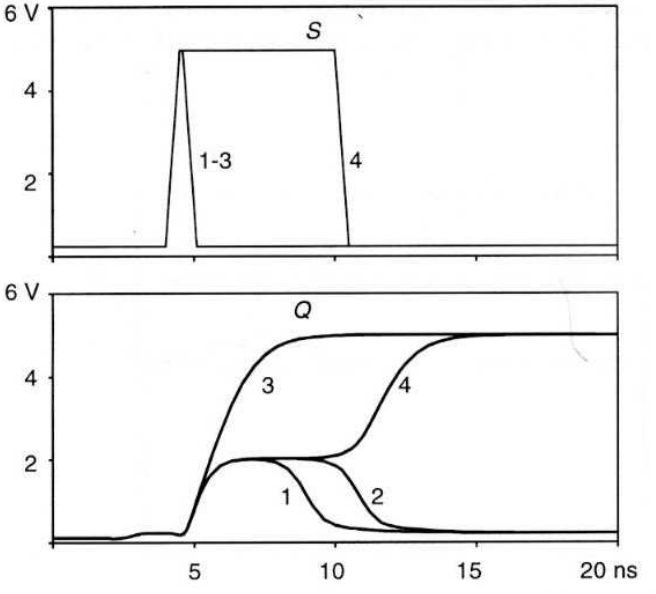


Figura 12.8 Simulazioni SPICE del comportamento dinamico del bistabile di Figura 12.6 al variare della durata dell'ingresso *S* (le durate sono crescenti da 1 a 4)

Anche nel secondo caso (Figura 12.8), se la durata dell'impulso *S* è inferiore ad un determinato valore, detto Δt_{min} (in questo caso circa 0.6 ns a $V_{DD}/2$), l'uscita non riesce a portarsi al valore alto ma entra nello stato metastabile, da cui esce dopo un certo tempo, per ricadere nello stato basso o passare a quello alto, in maniera non nota a priori. La durata minima Δt_{min} dell'impulso di comando *S* è legata al tempo di propagazione della serie delle due porte NOR, in quanto, per settare l'uscita (e cioè per portare quest'ultima nello stato stabile alto) occorre un tempo di propagazione (necessario per variare l'uscita) e un ulteriore tempo di propagazione perché questa uscita, applicata all'ingresso della seconda porta faccia variare l'uscita di quest'ultima, sostenendo quindi lo stato della prima porta anche in assenza dell'ingresso *S*. Il valore del tempo di settaggio in pratica è superiore al ritardo di propagazione t_P della singola porta

ed inferiore a $2t_P$, in quanto per innescare la reazione degenerativa basta che il segnale riportato all'ingresso della prima porta sia sufficiente a portare sopra la soglia il NMOS di questa porta.

In definitiva si vede che, per ottenere il comportamento corretto del circuito, si deve superare un livello di soglia sia per quanto riguarda l'ampiezza del segnale S (ciò vale anche per il livello dell'ingresso R nel passaggio dell'uscita Q allo stato basso) che per la durata minima del segnale S (e di quello R). Da quest'ultima osservazione si comprende come si possa avere un comportamento metastabile nel caso di Figura 12.5b quando i due segnali S e R si annullano contemporaneamente dopo essersi portati entrambi allo stato alto. Infatti, non essendo realistica una discesa *contemporanea* dei due segnali, nella realtà vi sarà sempre un intervallo di tempo, per quanto minimo, che differenzia le due discese; poiché la fase di set va valutata a partire dall'istante in cui il segnale R è diventato basso, se l'intervallo di tempo tra la discesa del segnale S e quello R è inferiore o paragonabile a questo tempo minimo, si innesca la condizione che porta alle situazioni di Figura 12.8. Lo stesso può dirsi se la discesa del segnale S precede leggermente quella di R; in tal caso il bistabile vede una fase di reset a partire dall'annullamento del segnale S, e se questo intervallo di tempo tra i due segnali è paragonabile al tempo minimo definito precedentemente, si innesca anche in questo caso un comportamento metastabile.

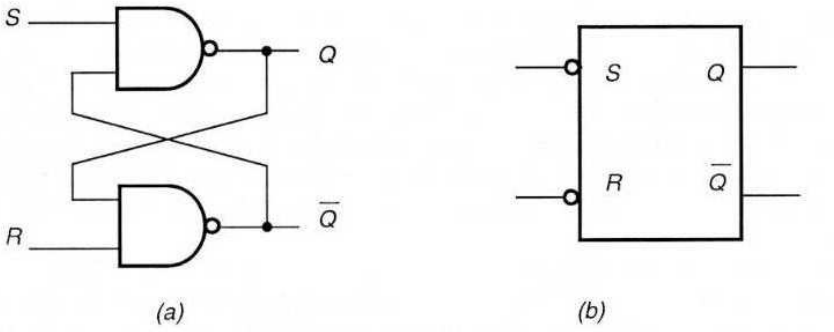


Figura 12.9 a) Schema logico di un bistabile *SR* con porte NAND; b) simbolo logico

Il bistabile *SR* può essere realizzato anche a partire da porte NAND, come è indicato in Figura 12.9a. In questo caso, come è facile verificare dalla tabella della verità corrispondente, riportata nella Tabella 12.3, per attivare l'uscita Q alta occorre che il segnale S sia basso, mentre per portarla allo stato basso occorre che R sia basso; ancora la conservazione degli stati precedenti delle due uscite si ha quando sia S che R sono alti (1), mentre lo stato ambiguo si ha per $S = R = 0$. Anche per questo circuito valgono le considerazioni fatte precedentemente circa la situazione di metastabilità che, in questo caso, si può instaurare con segnali $S = R = 0$.

Con le porte NAND quindi il bistabile funziona con ingressi attivi bassi. Il simbolo logico di questo circuito è quello riportato in Figura 12.9b e la notazione relativa agli ingressi S e R è quella relativa ad ingressi negati, per indicare questa modalità di funzionamento; si può riottenere il comportamento del bistabile SR con ingressi alti inserendo un invertitore su ognuno degli ingressi S e R relativi al bistabile a porte NAND.

Tabella 12.3 Tabella della verità sintetica del bistabile SR

S	R	Q_{n+1}	$/Q_{n+1}$
1	1	Q_n	$/Q_n$
0	1	1	0
1	0	0	1
0	0	?	?

12.4 Realizzazioni circuitali del bistabile *SR*

12.4.1 Tecnologia MOS

La realizzazione in tecnologia NMOS del bistabile SR con porte NOR è quella riportata in Figura 12.6. Per questi circuiti, come si è visto, la tensione di soglia è compresa tra i valori V_{IL} e V_{IH} della porta NOR corrispondente, in quanto il punto C che definisce il confine tra il passaggio da uno stato all'altro si trova necessariamente nel tratto a forte pendenza della caratteristica di trasferimento complessiva (per il quale l'amplificazione $A >> 1$). Il valore di soglia per il bistabile NMOS è quindi modificabile in funzione della scelta del valore $K_R = K_1/K_2$ dei MOS della porta.

La versione del bistabile SR in tecnologia CMOS è riportata in Figura 12.10, e discende direttamente dallo schema logico con due porte NOR, in base alla configurazione standard delle porte NOR in tecnologia CMOS. Anche in questo caso la soglia logica per gli impulsi S e R è definita dal punto di instabilità C, in base alla costruzione grafica di Figura 12.3. In questo caso, ricordando la simmetria della caratteristica di trasferimento dell'invertitore CMOS con $K_P = K_N$, la tensione di soglia è pari a $V_{DD}/2$.

La pendenza della caratteristica di trasferimento nell'intorno del punto C è molto elevata (la curva è verticale se si trascura la pendenza delle caratteristiche I-V in regione di pinch-off), per cui l'instabilità in quest'intorno è più accentuata, e il comportamento metastabile è più difficile da instaurarsi e di più breve durata rispetto al caso precedente. Un ulteriore vantaggio implicito nell'uso di strutture CMOS è nella maggiore capacità di fan-out di queste porte, e quindi della più elevata velocità di transizione del bistabile in presenza di carichi capacitivi in uscita non trascurabili.

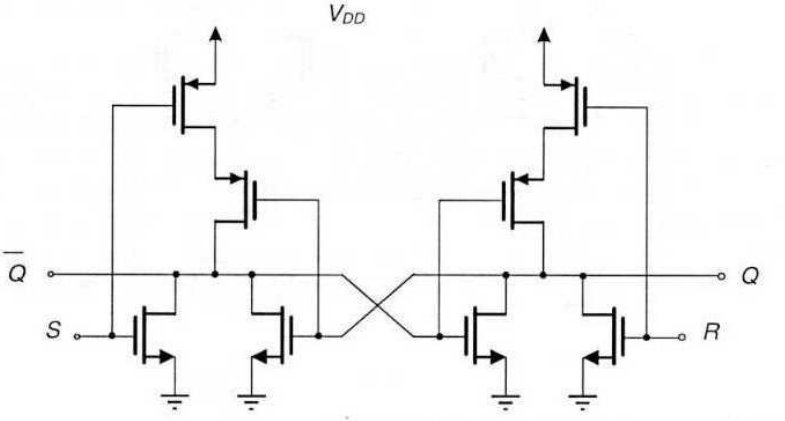


Figura 12.10 Bistabile *SR* con porte NOR CMOS

In Figura 12.11 è riportata una simulazione SPICE del funzionamento del bistabile *SR* CMOS. Dalla simulazione si ottiene un tempo di settaggio dell'uscita Q rispetto al segnale S di 0.29 ns; questo tempo, come si è detto, è leggermente minore di $2t_P$, dove t_P è tempo di propagazione t_P di ognuna delle due porte NOR che è dato dalla media dei due tempi di propagazione t_{PHL} e t_{PLH}, ed in questo caso vale $t_P = 0.166$ ns, per cui $2t_P = 0.33$ ns.

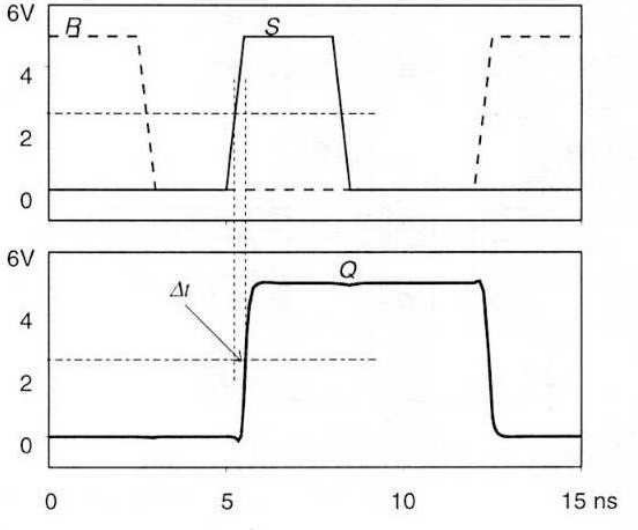


Figura 12.11 Simulazione SPICE del comportamento del bistabile *SR* CMOS. I valori di dimensionamento dei MOS sono: $W/L|_N = 2\mu m/1\mu m$, $W/L|_P = 5\mu m/1\mu m$

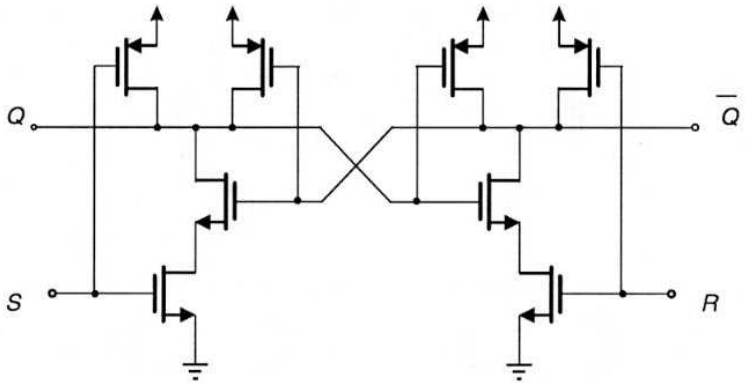


Figura 12.12 Bistabile *SR* con porte NAND CMOS

In tecnologia CMOS, come si è visto nel Capitolo 5, si preferisce impiegare porte NAND al posto di porte NOR, per utilizzare nel ramo in serie transistori NMOS che portano più corrente di quelli PMOS a parità di area. Lo schema elettrico della versione a porte NAND di un *SR* CMOS è riportato in Figura 12.12, mentre il tracciato corrispondente è riportato in Figura 12.13.

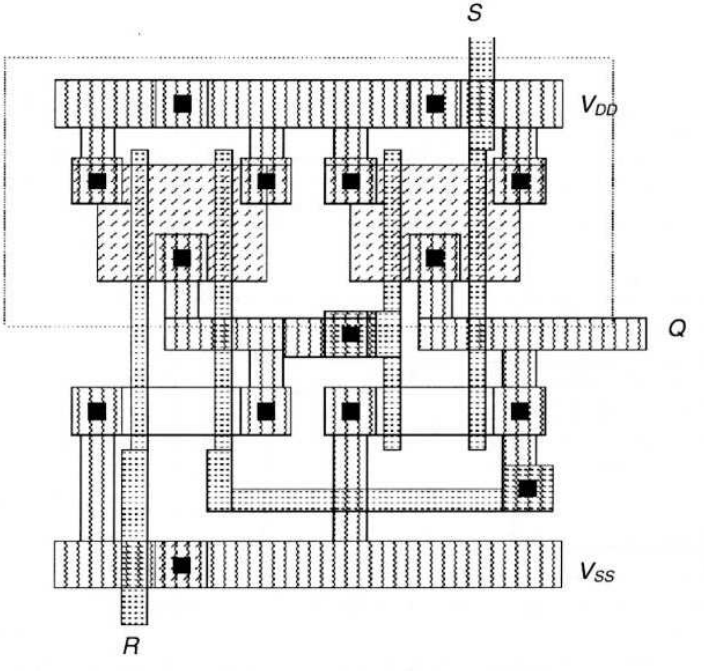


Figura 12.13 Tracciato del bistabile *SR* CMOS con porte NAND

12.4.2 Tecnologia bipolare

Le realizzazioni del bistabile *SR* con tecnologia bipolare prevedono l'uso di porte ECL o TTL, modificate rispetto alle versioni standard per compattare i circuiti risultanti e velocizzare il comportamento dinamico.

La versione del bistabile *SR* con porte NOR può essere direttamente implementata con logica ECL, le cui porte forniscono la funzione NOR ad una delle due uscite, come indicato in Figura 12.14; in questo schema i due transistori Q_{1a} e Q_{2a} (Q_{1b} e Q_{2b}) sono le coppie differenziali della porta a (porta b), e i transistori Q_{Sa} o Q_{Rb} effettuano l'operazione NOR rispetto a Q_{1a} o Q_{1b}.

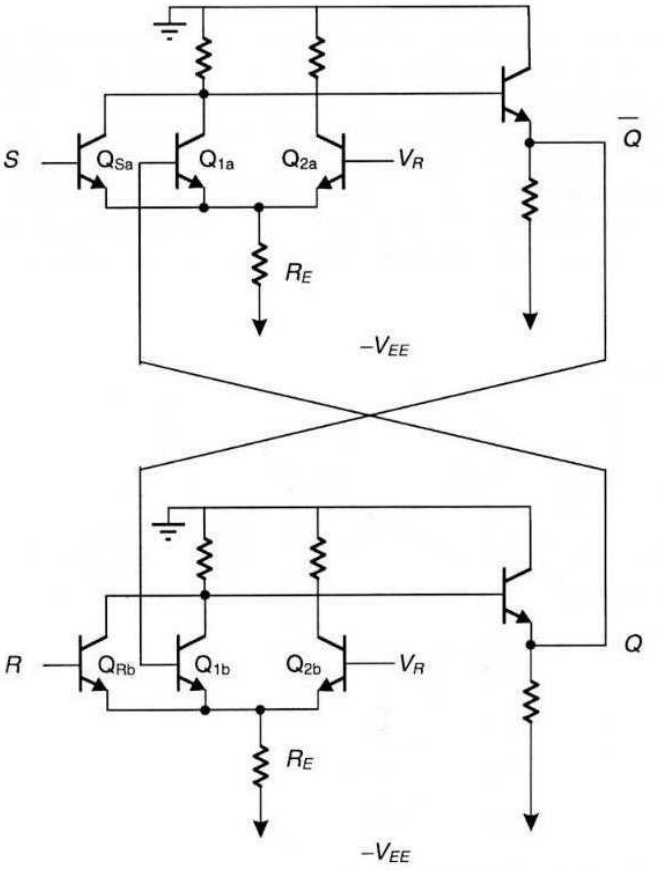


Figura 12.14 Bistabile *SR* con porte NOR ECL

Questo schema può essere semplificato in base alle seguenti considerazioni: nelle porte ECL il ramo con il transistore Q_2 polarizzato dalla tensione di riferimento V_R svolge la funzione di permettere il passaggio della corrente nella resisten-

za R_F quando il transistore Q_1 è interdetto, fissando nel contempo un riferimento di tensione sull'emettitore di quest'ultimo, che lo porta in interdizione quando il segnale di ingresso è basso ($V_I = V_{OL} = -1.7$ V).

Infatti nel caso che Q_2 conduca si ha:

$$V_E = V_R - V_{BEon} \cong -1.2 - 0.7 = -1.9 \text{ V} \tag{12.2}$$

e quindi, con un segnale di ingresso al livello logico basso V_{OL} si ha per il transistore Q_1:

$$V_{BE(Q1)} = V_{OL} - V_E \cong -1.7 - (-1.9) = 0.2 \text{ V} < V_\gamma \tag{12.3}$$

e quest'ultimo è interdetto. Nel circuito di Figura 12.14, in base al funzionamento del bistabile, si ha che quando un'uscita (ad esempio Q) è alta, l'altra è bassa, per cui se Q_{1b} è interdetto Q_{1a} è in conduzione e viceversa; si può quindi formare un'unica coppia differenziale utilizzando i transistori Q_{1a} e Q_{1b} con gli emettitori connessi ad un'unica resistenza R_E ed eliminando i transistori Q_{2a} e Q_{2b} nonché il circuito per la tensione di riferimento V_R. Si giunge così allo schema semplificato di Figura 12.15, in cui vi è una sola coppia differenziale formata da Q_{1a} e Q_{1b} e dai due transistori Q_{Sa} e Q_{Rb} per realizzare la funzione NOR sui due rami.

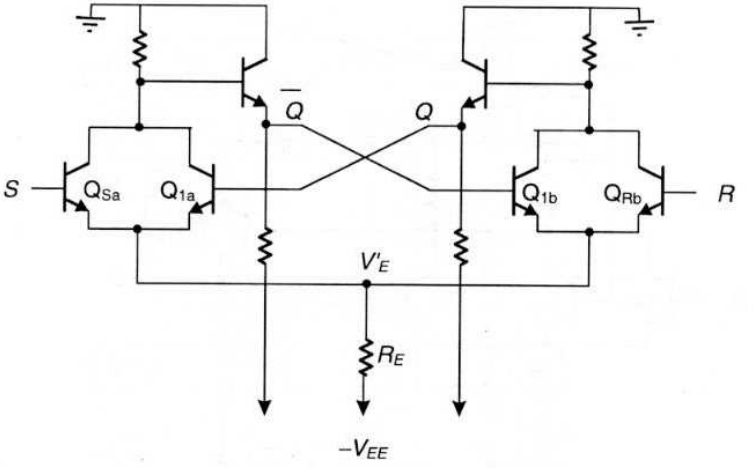


Figura 12.15 Bistabile ECL in forma compattata.

In questo circuito è facile verificare che, se ad esempio l'uscita Q è alta ($V(Q) = V_{OH}$), il transistore Q_{1a} conduce, e la tensione V_E' sul punto comune ai due emettitori, imposta dalla conduzione di Q_{1a} sarà:

$$V_E' = V_{OH} - V_{BEon} \cong -0.7 - 0.7 = -1.4 \text{ V} \tag{12.4}$$

e quindi è sufficiente per mantenere Q_{1a} interdetto, in quanto:

$$V_{BE(Q1)} = V_{OL} - V_E' \cong -1.7 - (-1.4) = -0.3 \text{ V} < V_\gamma \tag{12.5}$$

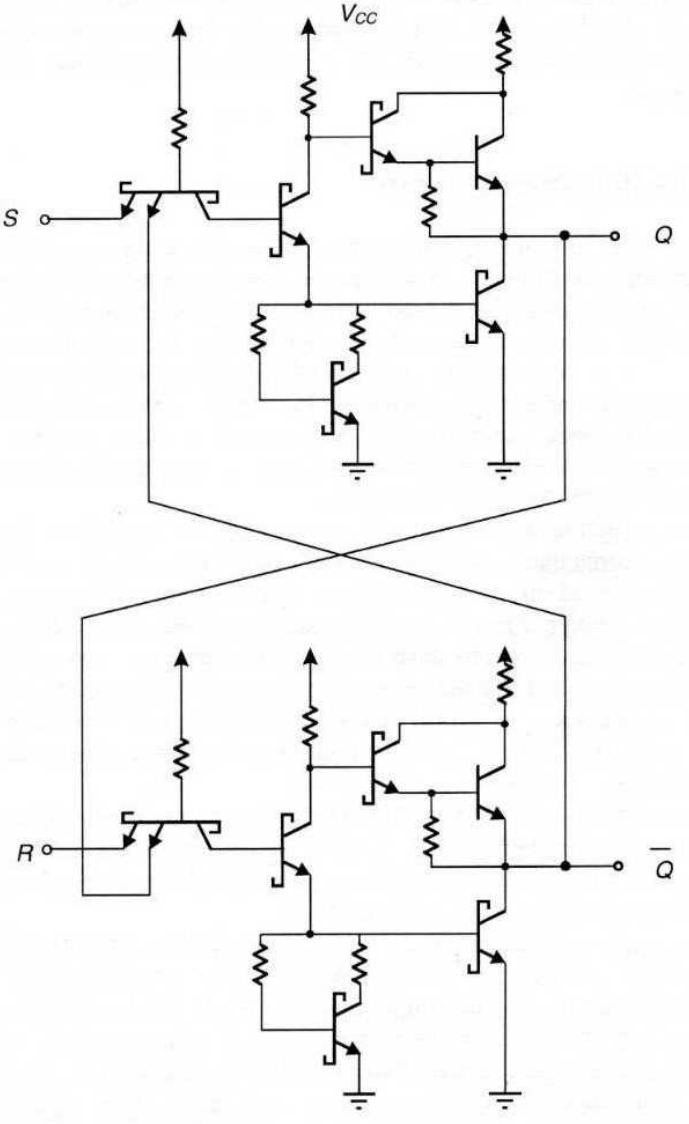


Figura 12.16 Bistabile *SR* con porte NAND TTL

Infine, in Figura 12.16 è riportata la versione TTL di un *SR* che utilizza due porte NAND TTL-LS (Schottky a basso consumo) per realizzare il latch *SR* con ingressi attivi bassi di Figura 12.9. In questo caso i segnali di set (o reset) sono alti nello stato di riposo, o di memorizzazione dello stato precedente; gli ingressi debbono abbassarsi sotto il livello $V_{IH} \cong 1.5$ V per indurre il cambiamento di stato nel latch. Anche in questo caso il circuito elettrico può essere compattato, tenendo conto che lo stadio di uscita totem-pole può essere eliminato nel collegamento tra le due porte, ed eventualmente inserito solo come stadio di interfaccia per garantire un elevato fan-out nel pilotaggio di circuiti logici a valle. Vedremo qualche esempio di compattazione nelle versioni circuitali TTL di altri circuiti sequenziali, riportati nei paragrafi seguenti.

12.5 I flip-flop sincronizzati

Si è già visto come in molti circuiti combinatori le operazioni vengano effettuate solo in coincidenza con impulsi di abilitazione, che permettono di cadenzare opportunamente le operazioni dei diversi sottoinsiemi del sistema. Ciò è a maggior ragione valido per i circuiti sequenziali, che debbono operare con una successione di eventi. In tal caso è conveniente sincronizzare queste sequenze di eventi su una cadenza prefissata di impulsi che definiscono la temporizzazione delle diverse operazioni; questa frequenza base viene detta *frequenza di orologio* (*clock*) e i sistemi di questo tipo sono detti *sincroni* (abbiamo già visto alcuni casi di sistemi sincroni nell'analisi delle strutture logiche dinamiche).

Nel caso dei latch la sincronizzazione permette di eliminare alcuni degli eventi non desiderabili, come quello della (quasi) contemporanea variazione dei segnali *S* e *R*, che può portare ad uno stato metastabile. Con l'impiego del segnale di clock *CK* il bistabile è abilitato a modificare il suo stato solo negli intervalli di tempo in cui *CK* è presente; quest'ultimo deve quindi essere applicato con un sufficiente ritardo rispetto all'applicazione dei segnali *S* o *R* in modo che durante la sua applicazione i segnali *S* e/o *R* sono stabili; inoltre il segnale di clock è di solito di durata più breve di quella di *S* o *R*, in modo che il circuito non venga influenzato dal passaggio al livello basso di *S* o *R*.

La sincronizzazione del latch può essere realizzata inserendo delle porte aggiuntive agli ingressi del latch stesso, in modo da realizzare una sezione di logica che precede quella di memoria dello stato, la quale è affidata al latch elementare già visto, come è indicato ad esempio in Figura 12.17.

I circuiti latch con sincronizzazione vengono anche definiti come "*flip-flop*". In generale, nel campo dell'elettronica digitale, si definiscono "*latch*" i circuiti che cambiano i loro stati logici in uscita negli istanti di tempo in cui si presentano le transizioni dei segnali in ingresso, mentre si definiscono "*flip-flop*" quei circuiti nei quali la modifica degli stati logici in uscita avviene solo in istanti di tempo determinati dal segnale di clock, e non dagli istanti di tempo in cui si hanno le transizioni dei segnali di ingresso. In base a questa definizione, i latch sincronizzati non dovrebbero essere definiti a rigore flip-flop, in

quanto, se il clock ha una durata finita e i segnali di ingresso variano durante l'intervallo di tempo in cui il clock è alto, gli stati logici delle uscite possono variare in corrispondenza delle transizioni dei segnali di ingresso, per cui si dice che il latch è "trasparente" agli ingressi. Tuttavia, nel seguito definiremo come flip-flop anche i latch sincronizzati, assumendo che la durata del clock sia sufficientemente breve, e che gli ingressi non varino durante l'applicazione del clock. Vedremo in seguito come le configurazioni di tipo master-slave dei circuiti sequenziali rispettino rigorosamente la definizione di flip-flop data.

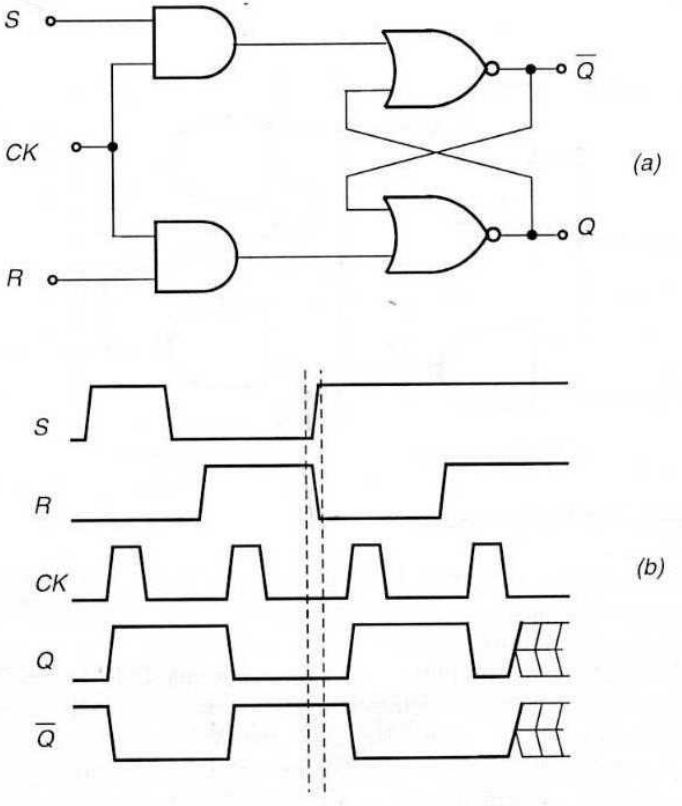


Figura 12.17 a) Flip-flop *SR* sincronizzato; b) temporizzazione dei segnali.

La sincronizzazione di un latch *SR* (che indicheremo come *flip-flop SR*), può essere realizzata mediante due porte AND aggiuntive, come indicato in Figura 12.17a. Il relativo diagramma di temporizzazione è riportato in Figura 12.17b; da questo si vede come, mediante un'opportuna scelta della durata del segnale *CK* rispetto a quelle dei segnali *S* e *R*, alcune delle situazioni critiche indicate in Figura 12.5 vengono eliminate, come ad esempio quella della variazione contemporanea di

S e R. La condizione ambigua con S e R entrambi attivi non è in ogni caso eliminata perché alla discesa del segnale di clock il bistabile si porta in una condizione non definita a priori a seguito di un comportamento metastabile. La tabella della verità è in ogni caso inalterata rispetto al caso del bistabile SR, ma si intende che le variazioni degli stati di uscita possono avvenire solo se CK è presente (alto).

Una versione del flip-flop SR, che prevede l'uso di porte NAND sia per la sezione logica che per quella di memoria, è quella riportata in Figura 12.18. In questo caso le porte NAND della sezione logica effettuano anche un'inversione dei segnali S e R, per cui i livelli logici bassi S' ed R' necessari per pilotare il latch SR con porte NAND corrispondono rispettivamente a $(S = CK = 1)$ e $(R = CK = 1)$, e quindi il flip-flop SR è nel suo complesso pilotato da ingressi alti.

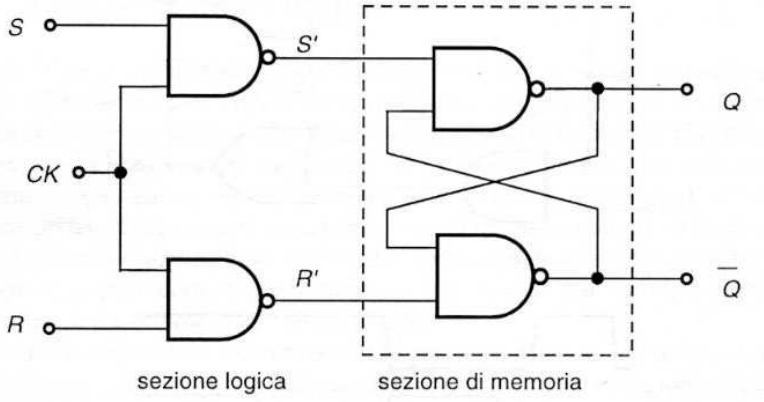


Figura 12.18 Flip-flop SR con porte NAND

Lo schema logico di Figura 12.17 che utilizza porte AND e NOR in cascata suggerisce la realizzazione di questo circuito mediante porte A-O-I. Una realizzazione circuitale di questo schema logico è quella dello schema semplificato di Figura 12.19 realizzato con porte A-O-I in tecnologia TTL. In questo circuito, il transistore Q_S (Q_R) a doppio emettitore effettua la funzione AND con il segnale di clock per la sincronizzazione del flip-flop, mentre le due coppie di transistori in parallelo realizzano le due funzioni NOR del latch. La presenza del transistore Q_i (che non effettua un'operazione logica) al secondo ingresso della NOR è necessaria per adattare il livello alto dell'uscita, $V_{OH} = V_{CC}$, a quello massimo ammissibile sulla base, pari a V_{BESAT}. Questo circuito va completato con una opportuna rete di uscita (ad esempio lo stadio totem-pole) nel caso di utilizzazione come componente logico standard, mentre può essere utilizzato in questa forma compatta in circuiti sequenziali più complessi, come ad esempio i flip-flop master-slave di tipo D o T che verranno descritti successivamente; in questo caso l'esigenza dello stadio di uscita è meno sentita in quanto il carico è prefissato in via progettuale dal circuito stesso.

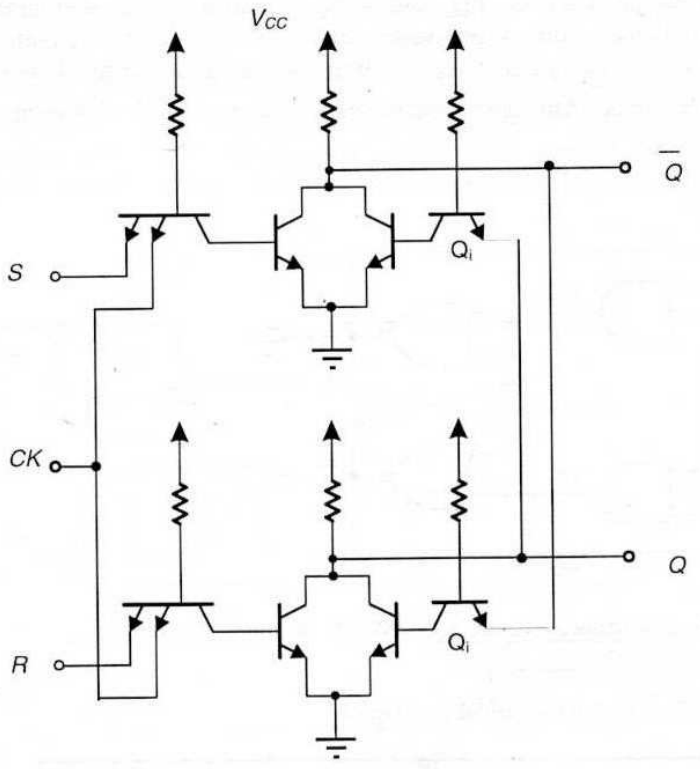


Figura 12.19 Schema circuitale di flip-flop *SR* con reti TTL

12.6 Flip-flop *JK*

Lo stato non ammesso o ambiguo dei flip-flop *SR*, che si verifica quando gli ingressi S e R sono entrambi attivi, può essere sostituito da uno stato utile mediante il riporto degli stati delle due uscite alla sezione logica in ingresso, in modo da effettuare un'operazione AND tra gli ingressi e le uscite corrispondenti. Questo circuito è detto flip-flop *JK* e il suo schema logico è riportato in Figura 12.20.

Come si vede dallo schema, questo è formato da un flip-flop *SR* sincronizzato, con l'ulteriore aggiunta di due collegamenti che riportano le uscite $\overline{Q}$ e Q in reazione agli ingressi delle due porte AND a cui sono applicati rispettivamente i segnali S o R (e il segnale di clock). Nel caso di ingressi S e R entrambi alti, quella tra le uscite $\overline{Q}$ e Q che è alta abilita la porta AND corrispondente, mentre l'uscita bassa disabilita la seconda porta AND e impedisce che il latch *SR* a valle veda entrambi gli ingressi S' e R' alti. Quindi il latch vede alto solo uno dei due segnali (S' o R'), anche nel caso di $S = R = 1$, e in questo caso opera un cambio di

stato sulle uscite rispetto alla situazione presente prima dell'applicazione dell'impulso di clock, come si può vedere dalla tabella della verità riportata in Tabella 12.4. Infatti, a causa dei collegamenti in reazione tra le uscite e le porte AND, se $\overline{Q}$ è alta viene abilitato il segnale di set S', mentre se Q è alto viene abilitato quello di reset R'.

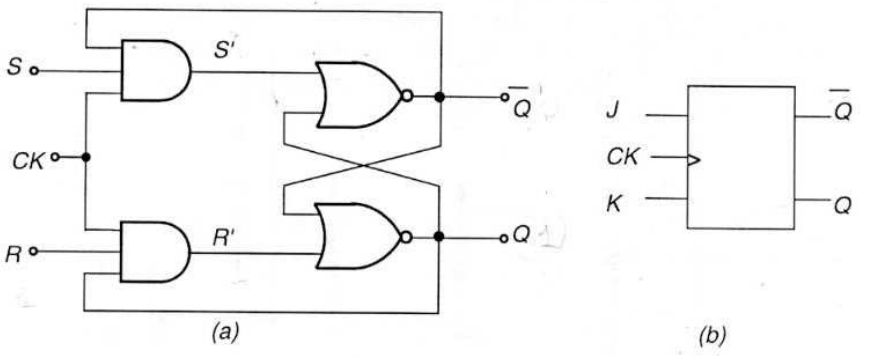


Figura 12.20 a) Schema logico di un flip-flop *JK* ; b) simbolo logico

Tabella 12.4 Tabella della verità del flip-flop *JK*

S	R	Q_n	$/Q_n$	S'	R'	Q_{n+1}	$/Q_{n+1}$
0	0	-	-	0	0	Q_n	$/Q_n$
1	0	1	0	0	0	1	0
1	0	0	1	1	0	1	0
0	1	1	0	0	1	0	1
0	1	0	1	0	0	0	1
1	1	1	0	0	1	$/Q_n$	Q_n
1	1	0	1	1	0	$/Q_n$	Q_n

Occorre però considerare gli effetti delle diverse temporizzazioni dei segnali di reazione rispetto a quelli applicati dall'esterno, in questo modo di funzionamento del flip-flop *JK*.

In Figura 12.21 sono riportati i diagrammi di temporizzazione dei segnali nel caso di un solo segnale S o R alto, e nel caso di entrambi i segnali alti. Per schematizzare il problema si è assunto un ritardo di propagazione t_P uguale per ognuna delle porte che costituiscono il circuito di Figura 12.20, e si è considerato per semplicità che il latch completa la transizione sull'uscita della porta a cui non è applicato il segnale di comando in un tempo pari a $2t_P$ (ritardo dovuto al passaggio attraverso due porte in cascata). Se i segnali S e R sono entrambi alti quando

viene applicato l'impulso di clock, nell'ipotesi di Q alto nello stato precedente, dopo un tempo t_P il segnale R' passa alto, dopo un ulteriore tempo t_P l'uscita Q passa al valore basso, e dopo un ulteriore t_P $\overline{Q}$ passa al valore alto. A questo punto, se il segnale di clock è ancora presente (caso riportato in linea tratteggiata in Figura 12.18), lo stato delle due uscite si inverte ulteriormente (dopo un tempo $3t_P$) e così via finché l'impulso di clock non ritorna basso impedendo l'ulteriore alternanza dei due stati.

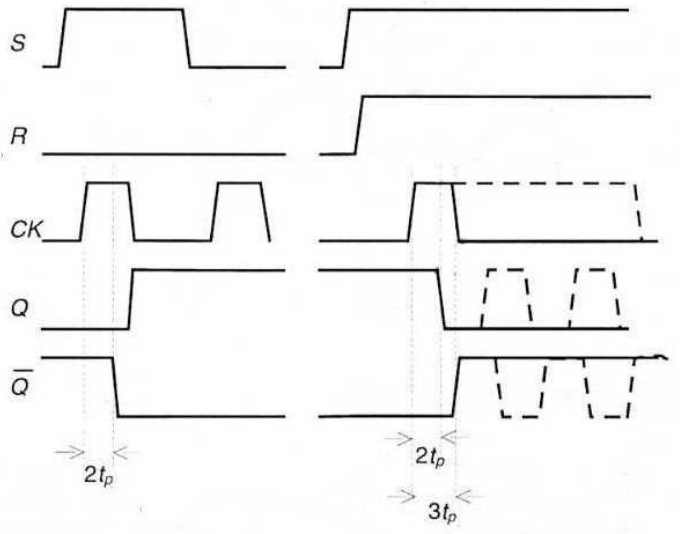


Figura 12.21 Problemi di temporizzazione per il flip-flop *JK* in funzione della durata dell'impulso di clock

Il flip-flop in questo caso alla fine del clock si porta in uno stato che non è detto sia quello indicato dalla tabella della verità per la situazione $S = R = 1$. Da quanto detto risulta che l'impulso di clock deve avere una durata inferiore a $3t_P$ per evitare le inversioni successive; d'altra parte l'impulso deve avere una durata di almeno $2t_P$ per permettere la commutazione del latch, come si è visto precedentemente. Quindi le condizioni sulla durata del clock sono molto stringenti e difficilmente realizzabili con i circuiti logici veloci per i quali i tempi di propagazione sono inferiori al ns.

È possibile ritardare i segnali di reazione inserendo nel collegamento un'ulteriore porta che presenta anch'essa un ritardo t_P. Un possibile circuito è quello di Figura 12.22 per il caso di realizzazione con porte NAND; in questo caso, per la presenza di un ulteriore ritardo nella maglia di reazione di t_P, ottenuto inserendo una porta NAND per ognuno dei due collegamenti (gli altri ingressi non riportati possono servire per l'inizializzazione del flip-flop), l'im-

pulso di clock deve essere compreso tra $2t_P$ e $4t_P$ per evitare i problemi su esposti.

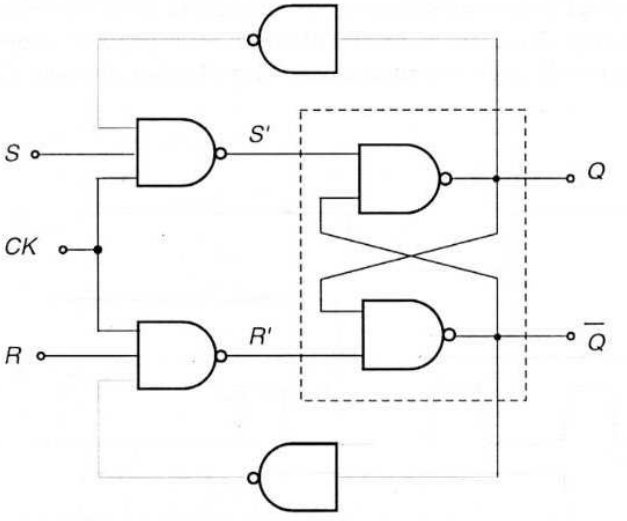


Figura 12.22 Flip-flop *JK* con reazione ritardata

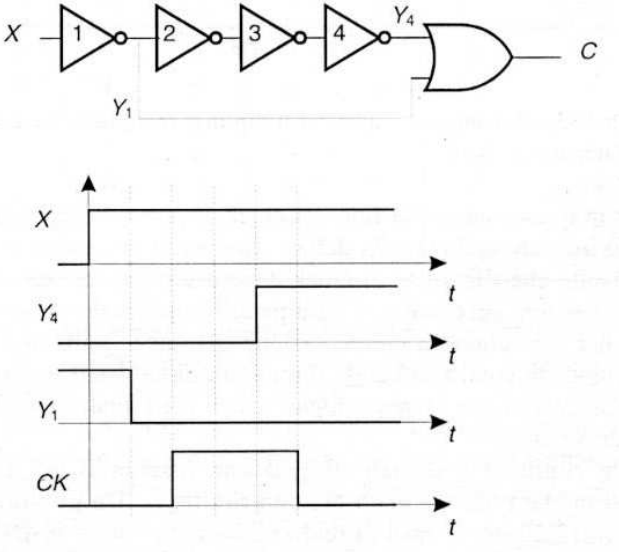


Figura 12.23 Circuito per definire la durata del clock; l'intervallo di tempo tra le linee tratteggiate è pari a t_P

Per determinare con sufficiente accuratezza la durata del clock è possibile utilizzare porte logiche con ritardi di propagazione uguali a quelli delle porte del flip-flop (e cioè porte realizzate con gli stessi componenti e con le stesse regole di progetto); ad esempio nella Figura 12.23 è indicato un modo per realizzare una durata del clock pari a $3t_P$ utilizzando quattro invertitori e una porta OR, e sollecitando il sistema con un'onda rettangolare di durata maggiore di $3t_P$.

L'interesse di questa soluzione è legato al fatto che il tempo di propagazione non è una costante, ma dipende, per una data realizzazione circuitale di porta logica, oltre che dal tempo di salita del segnale di comando, anche dalle variazioni di parametri come la tensione di alimentazione, la tensione di soglia e la temperatura ambiente.

Facendo ad esempio riferimento all'espressione approssimata del ritardo di propagazione t_P di un invertitore CMOS caricato da un altro invertitore:

$$t_P = \frac{C_G V_{DD}}{2k' \frac{W}{L} (V_{DD} - V_T)^2} \tag{12.6}$$

si vede che il valore di t_P dipende direttamente dal valore della tensione di alimentazione V_{DD}, ed implicitamente dalla temperatura, che agisce sul termine k' attraverso la dipendenza della mobilità dalla temperatura (in prima approssimazione μ diminuisce con la temperatura con un coefficiente di circa -3 cm^2/Vs°C) e sulla V_T (che diminuisce con la temperatura con un coefficiente di circa -3 mV/°C). La dipendenza più rilevante è quella della mobilità; ad esempio, per una variazione di temperatura da 25° a 50°, il tempo di propagazione aumenta di circa il 10%.

È quindi importante che i segnali di clock con specifiche di periodo legate ai valori dei tempi di propagazione delle porte, vengano realizzati con porte simili a quelle che vengono utilizzate nei circuiti sequenziali, in modo che eventuali variazioni di t_P vengano automaticamente legate alla variazione del periodo del segnale di clock.

12.7 Flip-flop master-slave

Il funzionamento dei flip-flop visti precedentemente, in particolare quello del flip-flop *JK*, dipende in maniera critica dalla durata dell'impulso di clock. Si possono invece rendere i flip-flop insensibili alla durata o al valore dell'impulso di clock, attivando le transizioni di stato solo quando il clock passa da basso ad alto o viceversa; questi flip-flop vengono detti flip-flop *comandati dalle transizioni* dell'impulso (*edge-triggered*).

Un circuito che realizza questa condizione e che è largamente impiegato nelle reti sequenziali è il flip-flop *master-slave*, che in effetti è formato dalla cascata di due flip-flop controllati dallo stesso segnale di clock in maniera opportuna. Una versione di circuito master-slave che realizza in effetti un flip-flop *JK* comandato da transizioni negative del clock è quella di Figura 12.24. Dallo schema logico si

evidenzia che il circuito è formato dalla connessione in cascata di due flip-flop *SR* sincronizzati a porte NAND, con i collegamenti di reazione per il funzionamento *JK* effettuati tra le uscite dello slave e gli ingressi del master.

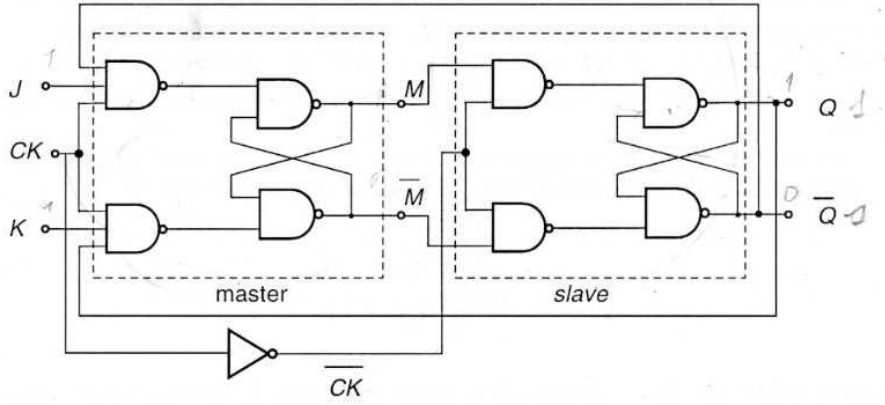


Figura 12.24 Schema di flip-flop master-slave *JK* con porte NAND

Quando il clock *CK* è alto, il flip-flop master è abilitato, mentre lo slave è disabilitato; quindi per tutto il tempo in cui il *CK* è alto le uscite Q e $\overline{Q}$ dello slave non variano e rimangono quelle dello stato precedente (indicato con n nella Tabella 12.5, che riporta la tabella della verità del flip-flop), mentre quelle del master variano in accordo con la tabella della verità di un flip-flop *JK* in cui le reazioni vengono effettuate a partire dalle uscite degli stati precedenti dello slave.

Tabella 12.5 Tabella della verità del flip-flop *JK* master-slave

J	K	CK	M	$/M$	$/CK$	Q_{n-1}	$/Q_{n-1}$
0	0	1	M_n	$/M_n$	0	Q_n	$/Q_n$
1	0	1	1	0	0	Q_n	$/Q_n$
1	0	0	1	0	1	1	0
0	1	1	0	1	0	Q_n	$/Q_n$
0	1	0	0	1	1	0	1
1	1	1	$/Q_n$	Q_n	0	Q_n	$/Q_n$
1	1	0	$/Q_n$	Q_n	1	$/Q_n$	Q_n

Se sia J che K sono alti, le reazioni delle uscite dello slave sono tali da avere rispettivamente sulle uscite M e $\overline{M}$ del master i valori (invertiti) delle uscite Q e $\overline{Q}$

dello stato precedente. Le uscite del master sono viste come segnali S e R dallo slave nella fase in cui il clock è basso, e quest'ultimo è quindi abilitato. Quindi se M è alto Q sarà alto e viceversa; nel caso di $J = K = 1$, l'inversione delle uscite dello stato precedente dello slave (conservate dalle uscite del master) verranno a modificare gli ingressi dello slave in modo da ottenere l'inversione anche delle uscite di quest'ultimo, quando CK diventa basso.

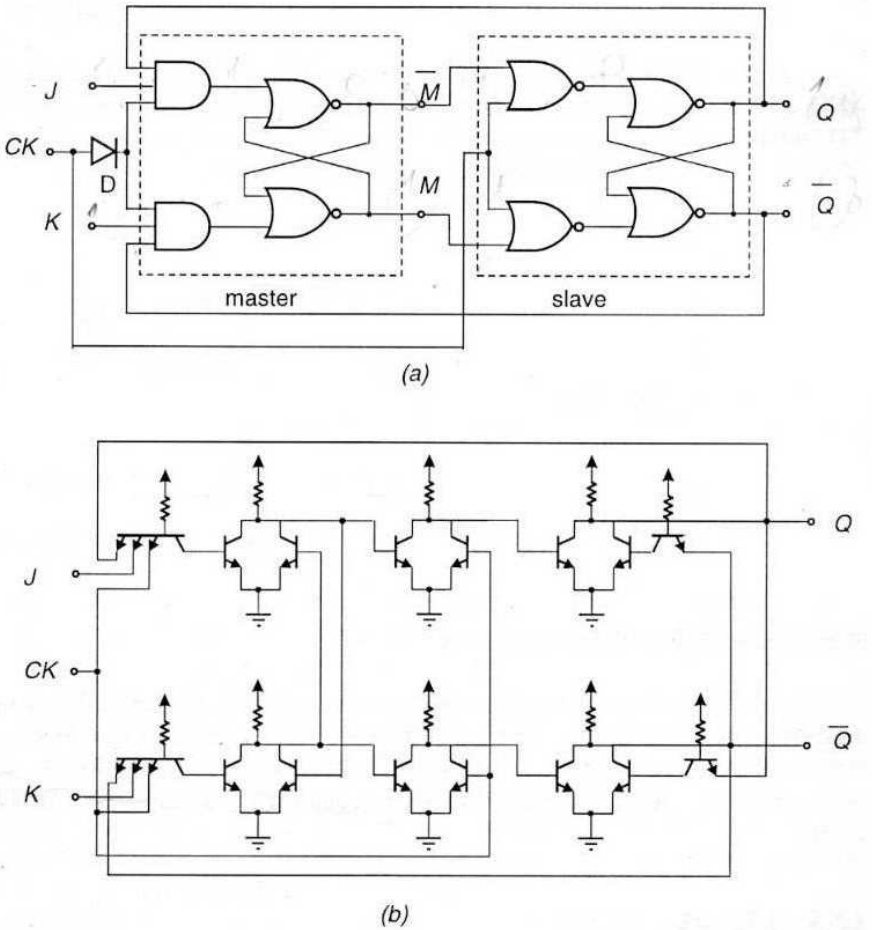


Figura 12.25 a) Flip-flop master-slave con porte AND e NOR; b) implementazione con porte TTL

Come si vede dal funzionamento descritto (e dalla tabella della verità della Tabella 12.5) il cambiamento di stato delle uscite del master-slave (uscite Q e $\overline{Q}$) avviene in corrispondenza della discesa del clock, e cioè quando lo slave è abilitato (ovviamente dopo il ritardo di propagazione della stadio slave); il cambiamento di

stato di quest'ultimo non può provocare nessuna azione indesiderata sul master, contrariamente al *JK* a singolo flip-flop, perché nel tempo in cui l'uscita dello slave varia, il master è già disabilitato in quanto il clock è basso. È importante, per assicurare la corretta temporizzazione dei due flip-flop in cascata, che la disabilitazione dello slave preceda di un Δt finito l'abilitazione del master, per evitare che lo slave possa essere ancora abilitato quando il master cambia stato; questo può essere assicurato o attraverso il ritardo t_P del flip-flop master, se questo è relativamente lento, o con un ritardo aggiuntivo del segnale *CK* rispetto a quello invertito applicato allo slave.

In Figura 12.25 è riportata una versione di flip-flop master-slave con porte AND e NOR che si presta ad essere implementata direttamente con porte A-O-I TTL modificate. Il circuito richiede 16 transistori, mentre la realizzazione diretta con porte AND e OR ne richiederebbe un numero maggiore.

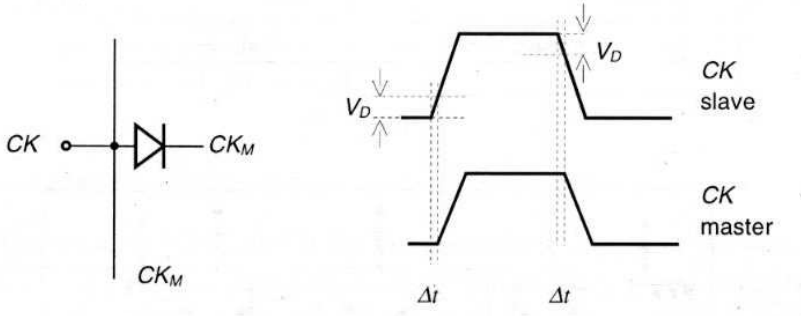


Figura 12.26 Circuito di ritardo del clock

Il diodo D sull'ingresso del clock al master serve ad anticipare il segnale di disabilitazione CK_S allo slave (in questo caso alto) rispetto a quello CK_M del master, utilizzando un impulso a tempo di salita relativamente lento e la tensione di soglia V_γ del diodo per ritardare il segnale CK_2 di un tempo Δt (Figura 12.26).

12.8 Flip-flop *D* e *T*

Le versioni di flip-flop comandate dal fronte del segnale di abilitazione sono utilizzate per realizzare due tipi di flip-flop che costituiscono la struttura base di due circuiti sequenziali molto comuni nei sistemi digitali: i registri e i contatori.

Per la realizzazione di registri viene utilizzato il *flip-flop D* (da *Delay*, ritardo), che agisce come un elemento di ritardo in quanto fornisce all'uscita Q (dopo un ritardo ΔT) la variabile logica D applicata in ingresso. Il flip-flop D è realizzato con un master-slave *JK* comandato dal fronte di discesa del clock, e con il segnale D

applicato direttamente all'ingresso J, e invertito a quello K, secondo lo schema logico di Figura 12.27.

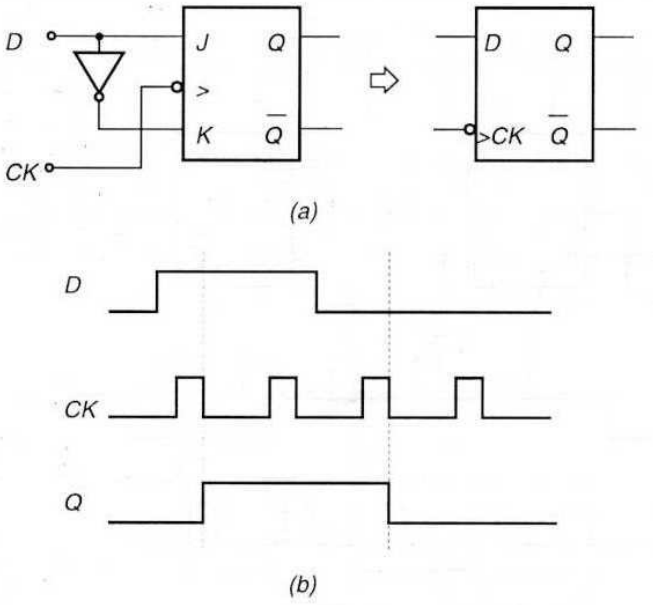


Figura 12.27 a) Schema logico del flip-flop D; b) temporizzazione dei segnali

Il segnale D, applicato direttamente all'ingresso J e invertito a quello K, pone il flip-flop in set se D è alto ($D = 1$) e in reset se D è basso ($D = 0$); tuttavia il flip-flop cambia di stato ($Q = 1$ se $D = 1$; $Q = 0$ se $D = 0$) solo quando il segnale CK scende a zero per la prima volta dopo l'applicazione del segnale D. L'uscita segue quindi il valore logico del segnale D con un ritardo ΔT dall'arrivo di quest'ultimo, che può variare da zero fino al periodo del segnale di clock, a seconda dell'istante di arrivo di D relativo alla sequenza degli impulsi CK. Questo comporta che, in una cascata di flip-flop D sincronizzati, in cui l'uscita del precedente varia in corrispondenza del fronte di discesa di CK, il ritardo tra le uscite di due flip-flop contigui è pari al periodo T del segnale di clock.

Il flip-flop T (da *Toggle*, commutatore) è invece l'elemento base per la realizzazione dei circuiti contatori, in quanto opera in modo da invertire il suo stato di uscita ad ogni impulso CK (e quindi conta in sistema binario alternando l'uscita tra 0 e 1 a seconda del numero di impulsi che sono arrivati all'ingresso CK). Il flip-flop T viene realizzato a partire da un master-slave JK comandato dalla transizione di discesa del segnale CK, ma con i due ingressi J e K connessi ad un unico segnale di ingresso T, come è indicato nello schema logico di Figura 12.28. In tal caso, ricordando il funzionamento del JK, se $T = 1$, all'applicazione del segnale CK le uscite vengono scambiate, per cui $Q_{n+1} = \overline{Q}_n$; al successivo impulso CK, se l'ingresso T è

ancora alto, l'uscita si inverte nuovamente, e così di seguito. In questo caso il segnale che di fatto opera il flip-flop T è quello CK che ad ogni periodo determina le alternanze dell'uscita; il segnale T è in realtà un segnale di abilitazione, in quanto per $T = 1$ si ha l'inversione delle uscite, mentre per $T = 0$ l'uscita rimane inalterata ($Q_{n+1} = Q_n$).

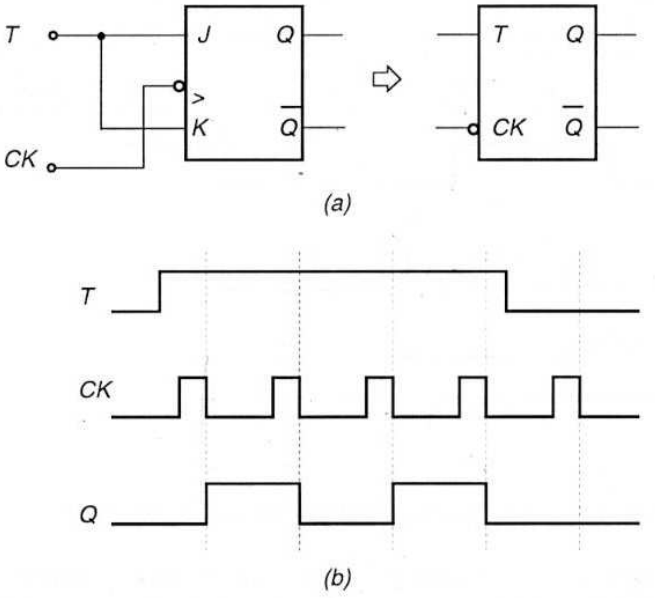


Figura 12.28 a) Schema logico di flip-flop T; b) temporizzazione dei segnali

12.9 Registri e contatori

Come si è già detto, i flip-flop D e T sono gli elementi base di circuiti sequenziali di grande impiego nei sistemi digitali, come i registri e i contatori. Si presenteranno ora, a scopo di esemplificazione del modo di funzionamento di questi circuiti, le strutture più comuni basate sull'utilizzo dei flip-flop presentati e il loro modo di funzionamento, senza voler coprire le diverse configurazioni che questi circuiti possono assumere nelle differenti applicazioni.

Il *registro* è un circuito che permette di immagazzinare in modo temporaneo una parola di n bit, memorizzando i singoli bit in ognuno degli n flip-flop di cui il registro è formato; il circuito può essere visto come un circuito di memoria, perché permette di immagazzinare una parola di n bit nei suoi n stadi, ma è usualmente utilizzato per la gestione sequenziale dei dati nei sistemi digitali, in quanto può incamerare una parola binaria fornita dai circuiti ad esso

connessi e presentarla in uscita (in maniera seriale o parallela) in un tempo successivo per l'ulteriore elaborazione.

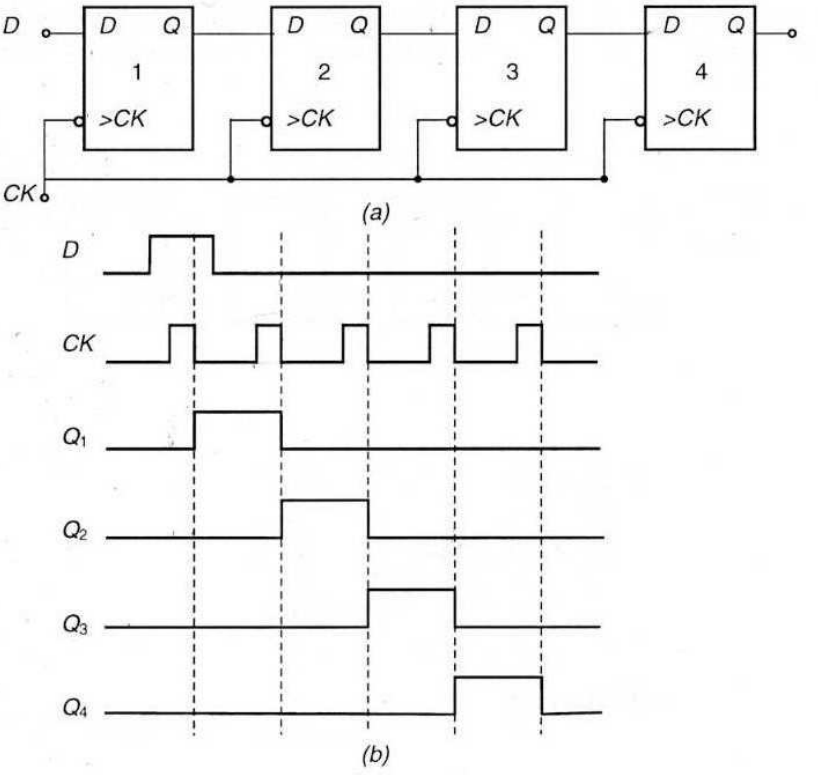
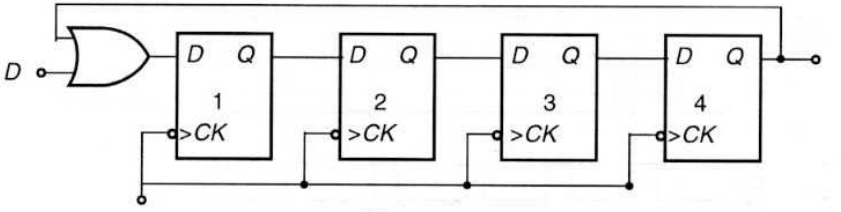


Figura 12.29 a) Schema di registro a scorrimento a quattro stadi; b) temporizzazione dei segnali

Il più semplice registro è quello detto a *scorrimento* (*Shift register*), realizzato da una connessione in cascata di n flip-flop D, che, come si è visto, introducono un ritardo tra l'ingresso D e l'uscita Q in dipendenza del periodo T del segnale di clock. Il funzionamento di un registro a scorrimento a quattro stadi è indicato in Figura 12.29. Dai diagrammi temporali delle uscite dei singoli stadi si vede che il segnale alto in uscita si propaga negli stadi successivi con un ritardo pari al periodo T di CK, e si presenta quindi all'uscita del registro (cioè all'uscita Q_4 del quarto flip-flop) dopo un tempo pari a $3T$ rispetto al passaggio alto della prima uscita Q_1; una prima applicazione di questo circuito è quindi quella di un circuito di ritardo con ritardo definito dal periodo di clock e dal numero di stadi. Nel diagramma di temporizzazione dei segnali della Figura 12.29b si è supposto un segnale D non sincronizzato con il segnale di clock, per cui il ritardo introdotto dalla prima cella è

differente da quello delle celle successive e dipende dal ritardo di fase del segnale di clock rispetto a *D*. Se (come è la norma nelle reti sincrone) il segnale *D* è anch'esso sincronizzato con il clock, allora anche la prima cella introduce un ritardo pari a *T* e il ritardo complessivo del registro a 4 celle è di 4*T*.

Se in ingresso ad un registro a *n* stadi viene applicata una sequenza di bit, il registro può incamerare fino a *n* bit nei suoi stadi, dopo di che il primo bit esce dall'ultimo stadio e via via in sequenza i successivi *n*−1 bit, secondo la tabella di Figura 12.30.



(a)

D	Q_1	Q_2	Q_3	Q_4
1	1	0	0	0
0	0	1	0	0
1	1	0	1	0
0	0	1	0	1
1	1	0	1	0
0	0	1	0	1
1	1	0	1	0
0	0	1	0	1

(b)

Figura 12.30 a) Registro a scorrimento con ricircolazione dei dati; b) posizioni dei dati nelle celle dopo ogni impulso di clock

Se i dati vengono fatti ricircolare in ingresso mediante il collegamento tra l'ultimo stadio e il primo, si ottiene il *registro a ricircolo* (Figura 12.30), che può incamerare una parola di *n* bit e presentarla in uscita in maniera seriale con la cadenza del segnale di clock. Il circuito prende anche il nome di contatore ad anello, nome che si riferisce all'osservazione che il dato *D* (supposto sincrono con il segnale di clock *CK*) si ripresenta in ingresso dopo ogni sequenza di *n* impulsi di clock, per cui il circuito è capace contenere al suo interno *n* bit, facendoli ricircolare tra ingresso e uscita ad ogni *n* impulsi. Un altro modo di definire questo circuito è quello di divisore di frequenza in quanto, con un solo bit 1 memorizzato, l'uscita cambia stato per ogni *n* alternanze del segnale di clock e quindi il periodo di *Q* è *n* volte maggiore di quello di *CK*.

I circuiti *contatori* sono circuiti basati su una catena di flip-flop che assume stati differenti ad ogni arrivo di un impulso. Se si utilizzano n stadi le uscite dei flip-flop possono assumere fino ad un massimo di 2^n configurazioni differenti; se queste cambiano all'arrivo di ogni singolo impulso si può ottenere un conteggio massimo di 2^n impulsi prima che il contatore assuma una delle configurazioni precedenti: si dice quindi che il contatore conta in modulo 2^n.

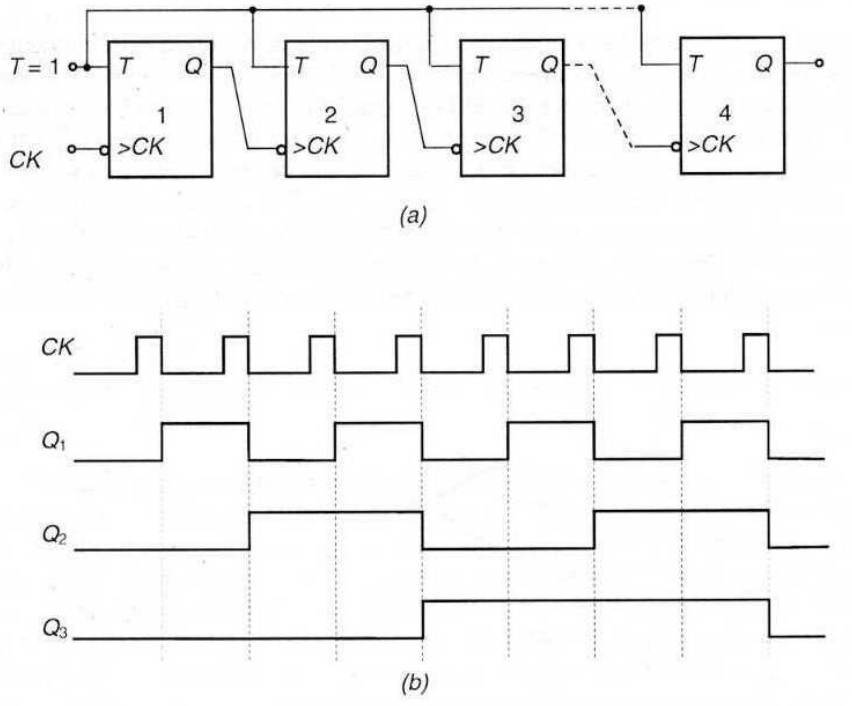


Figura 12.31 a) Schema di contatore in modulo k; b) temporizzazioni dei segnali

Il modo più semplice per realizzare un contatore in modulo 2^n è quello di connettere in cascata n flip-flop di tipo T i quali cambiano stato in uscita ad ogni arrivo di impulso successivo, come è indicato in Figura 12.31. Dalle temporizzazioni dei segnali in uscita si vede che ogni stadio divide per due il numero degli impulsi forniti dallo stadio precedente ossia conta in codice binario. In questo circuito tuttavia non solo l'uscita dell'ultimo stadio indica l'arrivo di 2^n impulsi, ma qualsiasi numero M di impulsi (tra 1 e 2^n-1) viene identificato attraverso lo stato delle uscite Q_i dei diversi stadi, a partire dal primo stadio che fornisce il bit meno significativo fino all'ultimo che fornisce quello più significativo relativo al numero stesso secondo l'espressione:

$$M = 2^{n-1} \cdot Q_n + \ldots + 2^2 \cdot Q_3 + 2^1 \cdot Q_2 + 2^0 \cdot Q_1 \qquad (12.7)$$

Naturalmente occorre azzerare il contatore prima del conteggio, ossia posizionare tutti i flip-flop in modo che ogni uscita Q_i sia bassa; questo si effettua con ingressi aggiuntivi di azzeramento (*clear*) ai singoli flip-flop, che possono ad esempio essere realizzati aggiungendo un ingresso ulteriore alle porte della sezione logica del singolo flip-flop.

12.10 Latch e flip-flop con logiche dinamiche MOS

Le strutture logiche MOS permettono di realizzare celle di memoria estremamente semplici, basate sull'accumulo di carica nella capacità di ingresso di un MOS. I due stati possibili di questo "latch" sono quelli di capacità carica (1 logico memorizzato) e di capacità scarica (0 logico memorizzato). Lo stato di carica della capacità tende in ogni caso a degradarsi per effetto delle correnti di dispersione e di contropolarizzazione delle regioni di isolamento, come si è visto in generale per le logiche dinamiche, per cui anche questo tipo di latch richiede un ripristino della carica (o un arrivo di un nuovo dato in ingresso) entro un tempo inferiore a quello di scarica della capacità stessa, e viene quindi considerato come *elemento di memoria dinamico*.

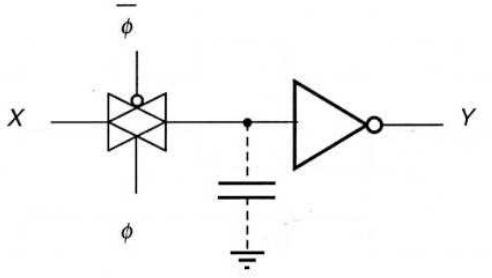


Figura 12.32 Circuito di principio di un latch in logica dinamica

Il circuito di principio di un bistabile basato su logica dinamica è quello di Figura 12.32, basato su una porta di trasmissione per la carica o scarica della capacità in accordo al segnale logico in ingresso, e di un invertitore che presenta in uscita (invertito) lo stato logico della capacità; quest'ultima è in realtà la stessa capacità di ingresso dell'invertitore, per cui non occorre una capacità aggiuntiva.

Il circuito più semplice, e con minor numero di transistori, utilizza un solo NMOS per la porta di trasmissione e un invertitore, per un totale di tre MOS, come è riportato in Figura 12.33a; questa versione come si è visto soffre della riduzione di V_T sulla tensione immagazzinata nella capacità e della dissipazione di potenza statica dovuta alla logica a rapporto utilizzata. Un circuito più efficiente da questo punto di vista è quello in tecnologia CMOS (Figura 12.33b) che prevede una coppia di transistori PMOS-NMOS sia per la porta di trasmissione che per l'invertitore, ossia quattro transistori.

Con entrambi i circuiti elementari è possibile realizzare in via molto semplice dei registri a scorrimento, basati sul ritardo che viene introdotto nella trasmissione dei segnali dalle uscite agli ingressi successivi, attraverso un comando sequenziale per l'apertura delle porte di trasmissione tra uno stadio e il successivo.

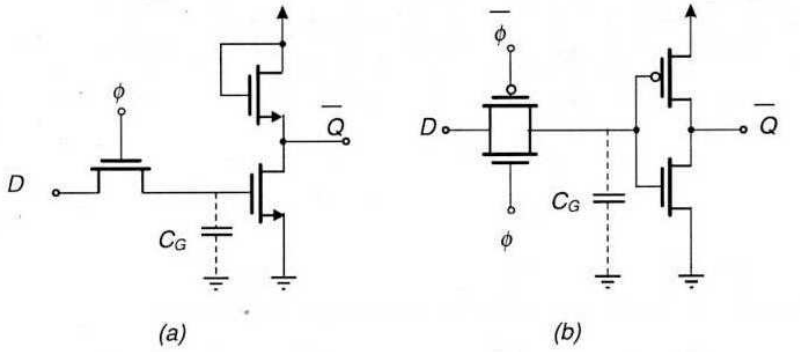


Figura 12.33 a) Circuito di memoria dinamico in tecnologia NMOS e b) in tecnologia CMOS

Per introdurre un ritardo tra l'uscita e l'ingresso e per evitare l'inversione tra l'uscita e l'ingresso del blocco elementare di memoria occorre porre in cascata due di questi stadi pilotati da due segnali di fase non sovrapposti, in modo da realizzare il funzionamento logico di un flip-flop di tipo D, ossia la cella elementare di un registro a scorrimento. In Figura 12.34 è riportato lo schema circuitale di una cella elementare del registro a scorrimento dinamico, realizzata in logica CMOS per esemplificare (il circuito basato sulle celle NMOS è di immediata deduzione).

Ognuna delle celle elementari D richiede due segnali di fase ϕ_1 e ϕ_2 per il pilotaggio delle due porte di trasmissione. La porta 1 si apre quando ϕ_1 è alto, memorizzando il segnale D in ingresso nella capacità C_A (in questo intervallo ϕ_2 è basso per cui l'uscita Q_1 non è modificata); successivamente quando ϕ_2 è alto il segnale in uscita dal primo invertitore viene trasmesso attraverso la porta 2 e immagazzinato nella capacità C_B (in questo intervallo di tempo ϕ_1 è basso e la porta 1 è chiusa per cui l'eventuale variazione del segnale D non viene a modificare il valore immagazzinato da C_A). Si ritrova quindi in uscita un segnale Q ritardato rispetto all'ingresso D.

Il ritardo della prima cella dipende dalla relazione temporale del fronte di salita dell'ingresso D rispetto al fronte di salita del clock ϕ_2; come si è già detto discutendo del funzionamento del registro a scorrimento, nell'ipotesi di un segnale D sincrono con il segnale di clock (in questo caso un segnale con il fronte di salita corrispondente a quello di ϕ_2), il ritardo dopo una generica cella del registro è pari al

periodo T del segnale di fase ϕ_2 (e quindi anche di ϕ_1), e il funzionamento del registro a n stadi è quello riportato nella Figura 12.29.

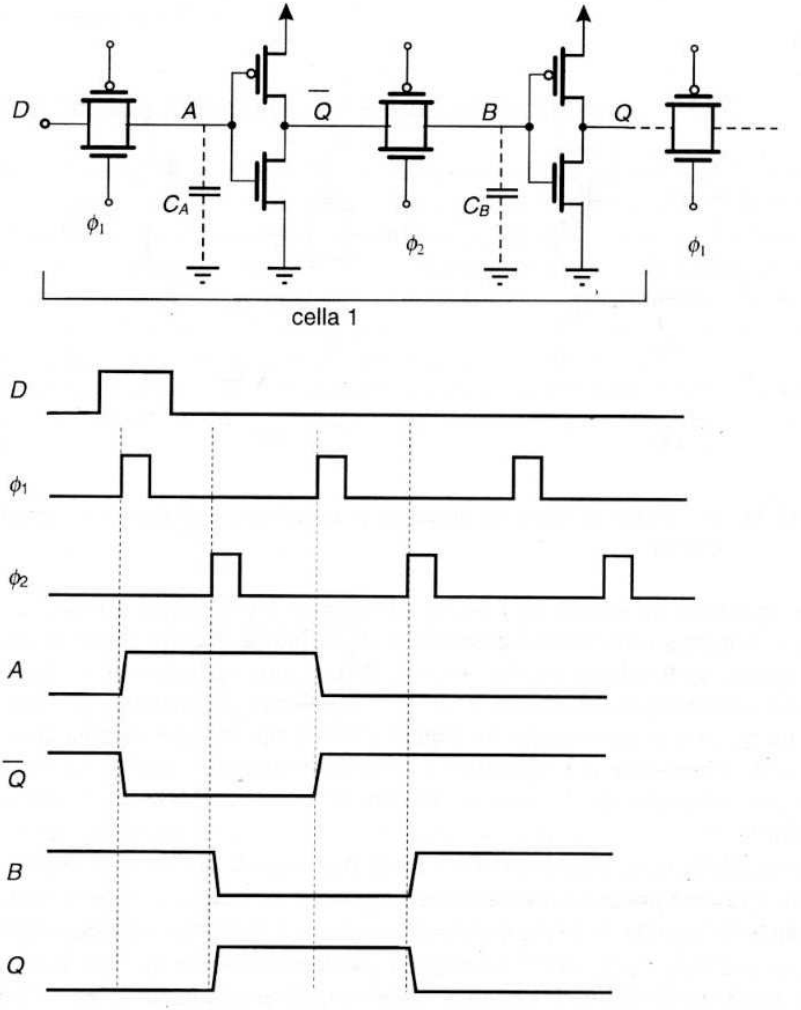


Figura 12.34 Registro a scorrimento con elementi di memoria in logica dinamica CMOS

In questo caso la cella elementare agisce come flip-flop D master-slave, con la porta 1 e invertitore 1 che realizzano la sezione master, e la porta 2 con l'invertitore 2 che realizzano quella slave, e il flip-flop è abilitato dal fronte di salita del segnale di fase ϕ_2.

In effetti il funzionamento di questo circuito può essere ricondotto a quello di una logica dinamica a due fasi, in cui le celle logiche dinamiche sono semplificate rispetto allo schema del Paragrafo 11.5, in quanto la rete logica NMOS è eliminata (l'unica operazione che deve effettuare in questo caso la rete logica è di invertire in uscita il segnale dipendente dallo stato di carica della capacità di ingresso). La fase di valutazione può essere quindi affidata alla stessa tensione ai capi della capacità: se questa è bassa l'uscita rimane alta, mentre solo se la tensione su C è alta la fase di valutazione permette la scarica del nodo di uscita e quindi l'instaurarsi di una uscita bassa. Il tracciato della cella elementare è riportato in Figura 12.35; si può notare come il circuito sia molto compatto conseguendo un risparmio di area notevole rispetto alle versioni che utilizzano flip-flop D master-slave statici.

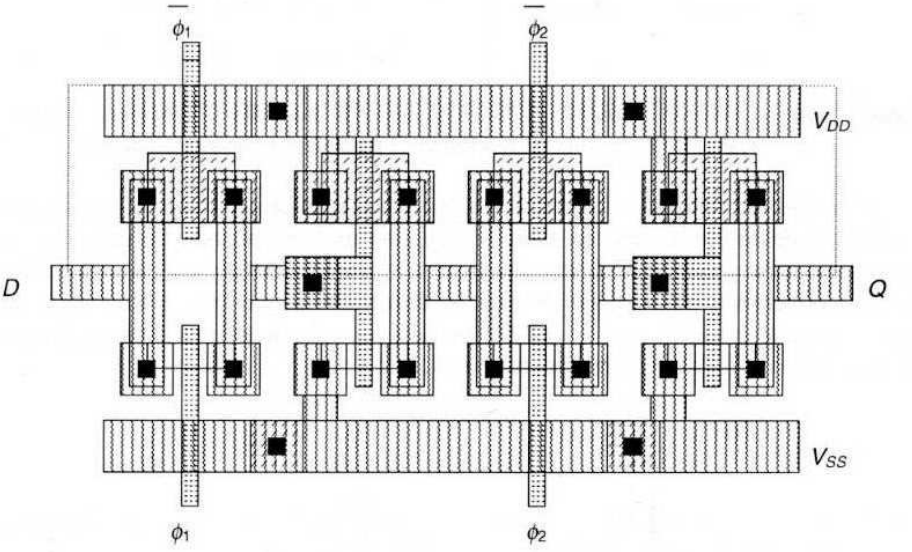


Figura 12.35 Tracciato della cella elementare di un registro a scorrimento in logica dinamica. I valori di dimensionamento sono: $W/L|_N$ = 3μm/2μm; $W/L|_P$ = 8μm/2μm

Per il corretto funzionamento del registro è necessario che le due porte di trasmissione di ognuna delle celle elementari D siano pilotate da due segnali di fase ϕ_1 e ϕ_2 non sovrapponibili, in modo da non avere mai contemporaneamente entrambe le porte aperte. Sebbene in linea di principio ciò possa essere ottenuto con segnali ricavati da un segnale di fase ϕ e dal suo negato $\overline{\phi}$ (Figura 12.36a), se i relativi percorsi comportano una differenza anche piccola dei tempi di propagazione, e in presenza di tempi di salita e discesa non nulli, si può avere una parziale sovrapposizione dei due segnali. In tal caso, nell'intervallo di tempo Δt in cui i due segnali si sovrappongono, il segnale in ingresso D si propagherebbe attraverso entrambi gli

stadi e raggiungerebbe l'uscita modificandola istantaneamente, come è schematicamente indicato in Figura 12.36b.

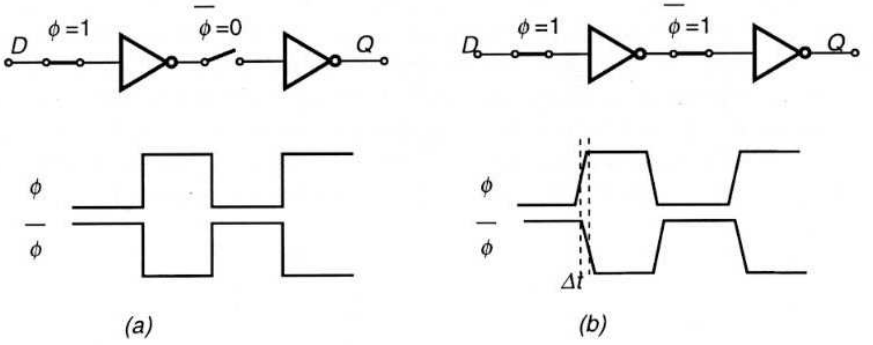


Figura 12.36 a) Schematizzazione del funzionamento di una cella *D* con segnali di fase ideali; b) effetto della sovrapposizione parziale dei due segnali di fase

Il flip-flop *D* realizzato in logica dinamica è più compatto delle versioni viste precedentemente (nella versione CMOS vengono usati 8 transistori invece che 16 transistori nella versione TTL master-slave o in quella CMOS statica), ma soffre degli inconvenienti già visti per le logiche dinamiche a due fasi, e come tutte le logiche dinamiche ha il problema della perdita del segnale memorizzato se la cadenza di arrivo dei dati è relativamente bassa.

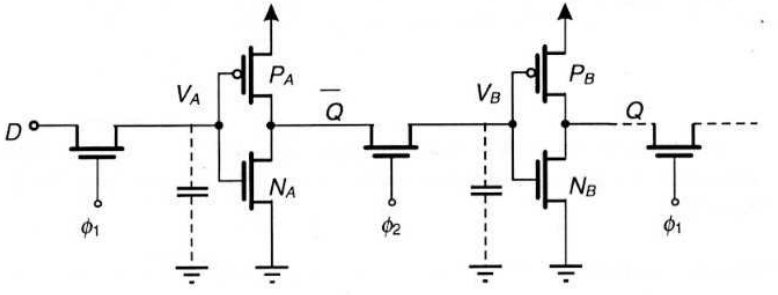


Figura 12.37 Cella di registro a scorrimento dinamico con porte NMOS

Una versione ancora più compatta di quella riportata in Figura 12.34 è quella che fa uso di porte di trasmissione a "pass transistor" NMOS invece che porte CMOS, secondo il circuito riportato in Figura 12.37, che utilizza solo 6 transistori. L'inconveniente presentato delle porte NMOS, che comportano una riduzione del livello alto della tensione in uscita del valore della tensione di soglia V_T, viene ad essere in questo caso meno risentito, in quanto l'invertitore CMOS che viene inse-

rito per invertire il segnale tra una porta e l'altra ripristina i livelli logici del segnale e riporta l'escursione logica al valore V_{DD}, come si può vedere dall'esame delle forme d'onda dei segnali all'ingresso e all'uscita del primo invertitore riportate nella Figura 12.38.

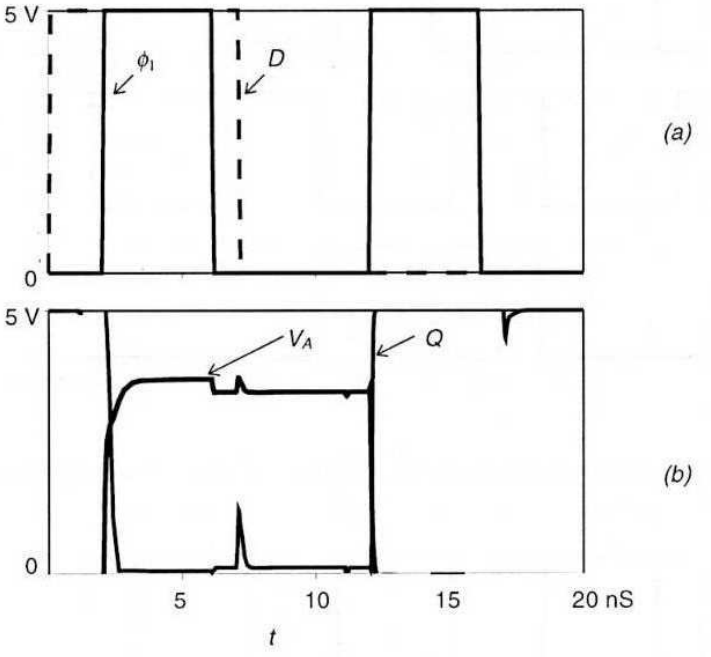


Figura 12.38 Simulazioni SPICE del circuito di Figura 12.37: a) segnale di ingresso D e segnale di fase ϕ_1; b) segnali in ingresso e in uscita dell'invertitore A

Le temporizzazioni complete relative alla cella di Figura 12.37, ottenute da un'analisi SPICE del circuito, sono riportate in Figura 12.39. Si può notare che sia l'uscita $\overline{Q}$ all'uscita del primo invertitore, che quella Q all'uscita del secondo, siano ripristinate nei valori logici, e depurate del rumore introdotto dalle transizioni dei segnali di fase e di ingresso, che passano attraverso le porte di trasmissione ma vengono filtrate dagli invertitori. Come si è detto precedentemente, il ritardo tra l'uscita Q e il segnale D è inferiore al periodo del segnale di fase, perché si è supposto un segnale D non sincrono con quest'ultimo. Per questa versione del registro con porte a un solo transistore rimane in ogni caso l'inconveniente di un maggior tempo di risposta delle porte NMOS nel presentare l'ingresso alto all'uscita.

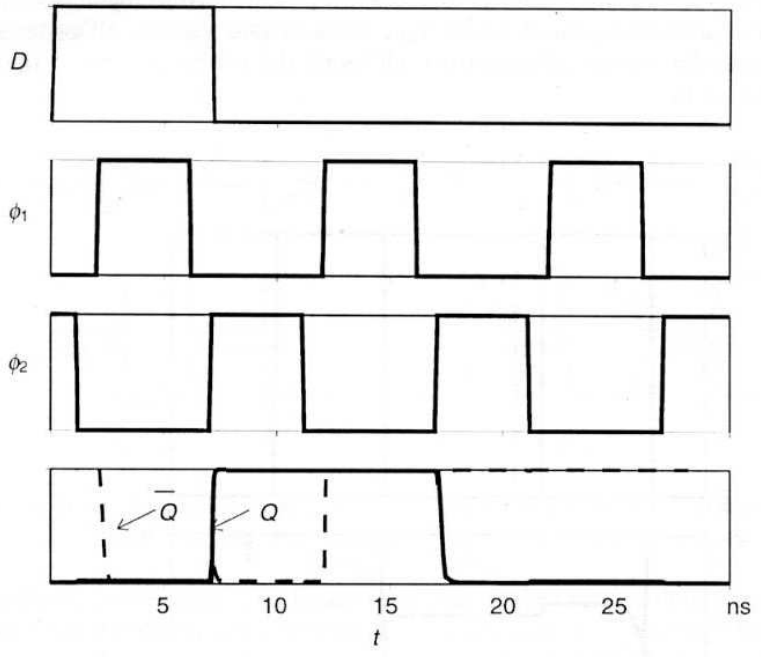


Figura 12.39 Andamenti dei segnali per il circuito di Figura 12.37; a) segnale D in ingresso; b) segnale di fase ϕ_1; c) segnale ϕ_2; d) segnali Q e $\overline{Q}$

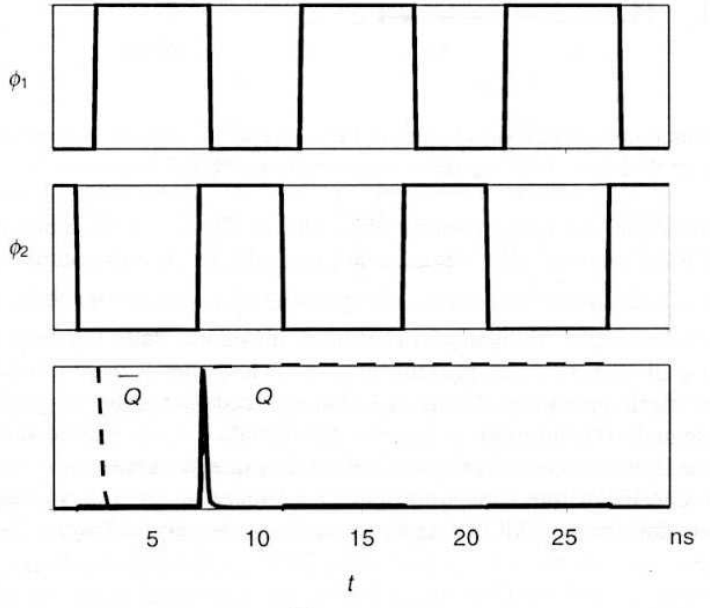


Figura 12.40 Forme d'onda Q e $\overline{Q}$ nel caso di una parziale sovrapposizione di ϕ_1 e ϕ_2

Per questo circuito si può verificare, mediante analisi SPICE, l'effetto sull'uscita di una parziale sovrapposizione dei due segnali di fase ϕ_1 e ϕ_2, effetto schematizzato nella Figura 12.36. L'analisi è riportata in Figura 12.40, in cui si è supposta una sovrapposizione dei due segnali di fase, eventualmente dovuta ad una variazione tra i due tempi di propagazione complessivi delle due linee, per una durata di 1 ns. Anche con una sovrapposizione così ridotta, si nota come il segnale Q all'uscita della cella si annulli nell'intervallo di tempo in cui ϕ_2 abilita la porta 2 e nello stesso momento ϕ_2 mantiene ancora aperta la porta 1, per cui il funzionamento del circuito non corrisponde più a quello di un flip-flop *D*.

Una versione di flip-flop *D* "statico" basato ancora su celle logiche dinamiche MOS è quella che sfrutta una seconda rete dinamica per riportare in ingresso il segnale di uscita in determinati intervalli di tempo, in modo da ripristinare il livello del segnale alto anche se l'intervallo tra due segnali di ingresso è molto lungo.

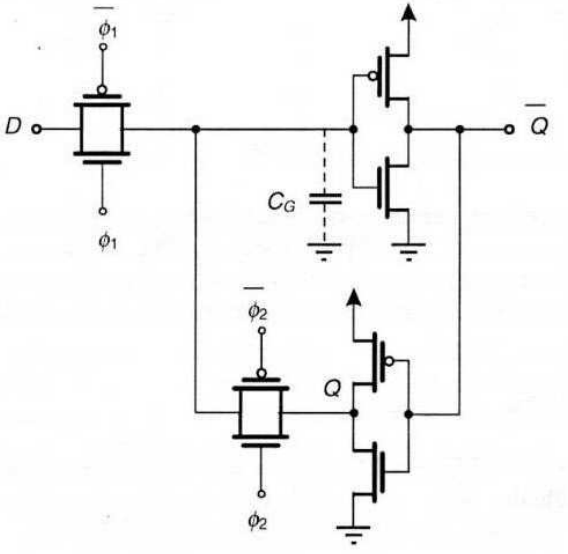


Figura 12.41 Flip-flop *D* statico con celle dinamiche CMOS

Lo schema di questo flip-flop in tecnologia CMOS è riportato in Figura 12.41; il circuito si basa su due porte di trasmissione pilotate dai due segnali di fase ϕ_1 e ϕ_2 non sovrapposti. La prima porta aprendosi trasferisce il segnale *D* all'ingresso del primo invertitore, che fornisce un'uscita $\overline{Q}$; quando si apre la seconda porta il segnale all'uscita del secondo invertitore *Q* viene riportato all'ingresso del primo, mentre nel contempo la prima porta è chiusa separando quindi il segnale *D* dall'uscita del secondo invertitore. La capacità in ingresso conserva l'informazione di *D* e quindi, se il segnale *D* era alto, in questo intervallo la tensione all'ingresso del primo invertitore viene ripristinata al livello primitivo, recuperando l'eventuale

scarica della capacità. Anche in questo caso è essenziale che i due segnali di fase non siano sovrapposti anche per tempi ridotti, altrimenti si avrebbe un conflitto in ingresso del primo invertitore.

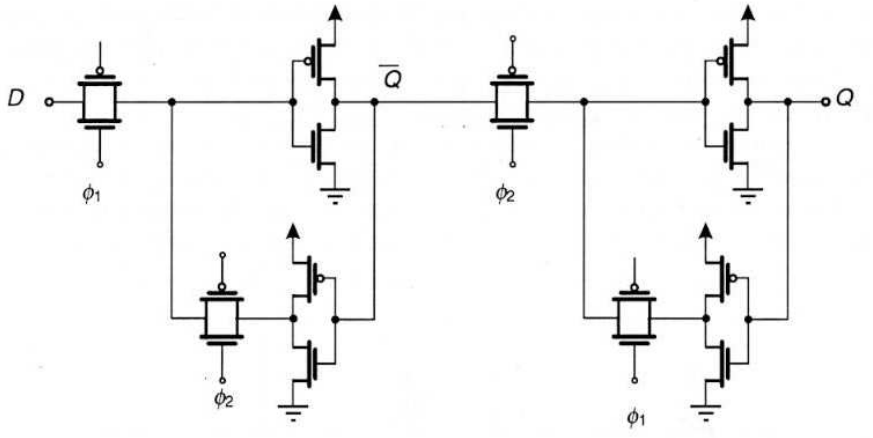


Figura 12.42 Flip-flop *D* master-slave con celle CMOS

Questo schema può essere iterato in modo da formare un flip-flop *D* di tipo master-slave ponendo in cascata due flip-flop statici come è indicato in Figura 12.42; in questo caso il circuito ha prestazioni migliori rispetto alle versioni precedenti, ma richiede 16 transistori e quindi ha un'occupazione di area circa doppia.

Esercizi di riepilogo

12.1 Per il latch *SR* NMOS di Figura 12.6, determinare tramite le caratteristiche di trasferimento delle due porte i valori della tensione Q nei due stati stabili A e B e quello dello stato metastabile C, utilizzando i seguenti valori: $V_{DD} = 5$ V; NMOS ad arricchimento: $V_T = 0.7$ V, $k' = 50$ μA/V^2, $W/L = 2$ μm/1 μm; NMOS a svuotamento: $V_T = -3$ V, $k' = 50$ μA/V^2, $W/L = 1$ μm/2 μm.

12.2 Disegnare lo schema circuitale di un latch *SR* NMOS con porte NAND, e determinare, mediante simulazioni SPICE del circuito con diverse durate del segnale S, la minima durata del segnale S nel passaggio dal valore alto a quello basso per avere una uscita Q al valore alto, assumendo per i parametri dei transistori quelli indicati nell'Esercizio 12.1, e per le capacità unitarie delle differenti regioni quelle riportate in Tabella 3.2.

12.3 Valutare con analisi approssimata i tempi di settaggio di un latch *SR* CMOS con porte NOR, per i seguenti valori dei dispositivi: $V_{TN} = |V_{TP}| = 0.7$ V, $k_N' =$

50 μA/V^2, $k_P' = 20$ μA/V^2, $W/L|_N = 2$ μm/1 μm, $W/L|_P = 5$ μm/1 μm, capacità unitarie $C_{OX} = 1.7$ fF/μm^2, $C_{J0} = 0.3$ fF/μm^2, e con $V_{DD} = 5$ V. Verificare i risultati mediante simulazioni SPICE del circuito.

12.4 Per il bistabile *SR* ECL di Figura 12.15, determinare i livelli logici delle uscite Q e $\overline{Q}$ e la potenza dissipata dal circuito, assumendo i seguenti valori dei parametri: $\beta_F = 50$, $R_{C1,2} = 250\ \Omega$, $R_{E1,2} = 800\ \Omega$, $R_O = 5$ kΩ, $V_{EE} = -5$ V.

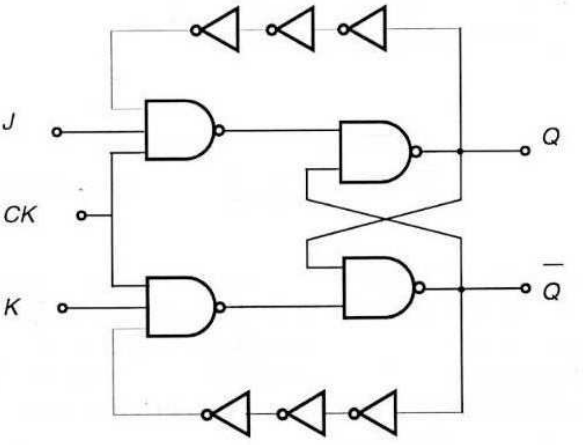


Figura E12.1

12.5 Con riferimento al flip-flop *JK* di Figura E12.1, determinare la minima e massima durata del segnale di clock per un funzionamento corretto del circuito, assumendo per gli invertitori un ritardo di propagazione $t_P = 0.5$ ns, e per le porte NAND un ritardo di propagazione $t_P = 1$ ns.

12.6 Modificare lo schema logico del circuito di Figura 12.23 per ottenere ancora una durata del clock di $3t_P$, dovendo utilizzare una porta NOR invece che OR, e assumendo anche in questo caso ritardi uguali per ogni porta e pari a t_P.

12.7 Determinare la variazione del tempo di propagazione t_P di un invertitore CMOS dimensionato con $W/L|_N = 2$ μm/1 μm, $W/L|_P = 5$ μm/1 μm, $k'_N = 50$ μm/V^2, $k'_P = 20$ μm/V^2, $V_{TN} = |V_{TP}| = 0.7$ V, per una variazione della temperatura ambiente da 0 °C a 50°C.

12.8 Il latch *SR* di Figura E12.2 è utilizzato per realizzare in uscita due segnali di fase ϕ_1 e ϕ_2 non sovrapposti. Assumendo sia gli invertitori che le porte NOR con uguale ritardo di propagazione t_P, determinare le forme d'onda di ϕ_1 e ϕ_2 e l'intervallo di tempo tra i valori alti di ϕ_1 e ϕ_2 assumendo un segnale di

clock con $T_1 = T_2 = 7t_P$. Cosa succede nei legami di fase di ϕ_1 e ϕ_2 se il clock ha i valori: $T_1 = 5t_P$, $T_2 = 7t_P$?

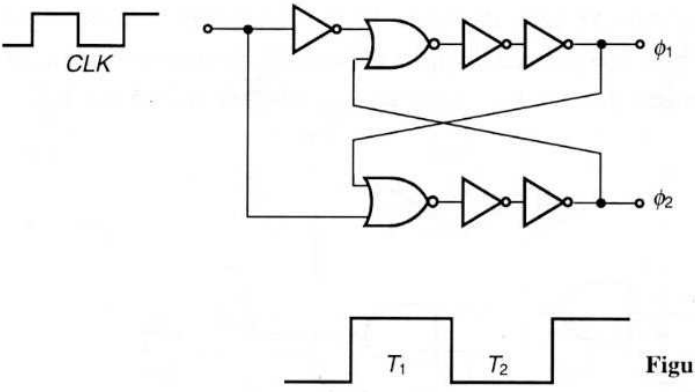


Figura E12.2

12.9 Nel registro a scorrimento di Figura 12.29a, assumendo che il segnale D abbia una durata dello stato alto di 10 ns, che il suo fronte di salita preceda di 5 ns quello del segnale di clock, e che quest'ultimo abbia una durata dello stato alto di 4 ns e un periodo di 20 ns, determinare: a) il ritardo tra l'arrivo del segnale D e quello dell'uscita Q_1 del primo stadio; b) il ritardo tra l'uscita Q_2 e quella Q_3; c) il ritardo tra D e Q_4.

12.10 Ripetere l'analisi dell'Esercizio 12.9 supponendo che il segnale D abbia il fronte di discesa sincrono con quello del segnale di clock, e che tutti gli altri dati rimangano invariati.

12.11 Analizzare il registro a scorrimento di Figura 12.29a, con i dati dell'esercizio 12.10, assumendo che ognuno dei flip-flop D del circuito sia comandato dalla transizione di salita del clock.

12.12 Per il contatore in modulo 2^6 determinare i livelli delle uscite $Q_1 \div Q_6$ dopo l'arrivo di 10 impulsi di clock. Attraverso quale indicazione si desume che sono arrivati 10 impulsi al contatore?

12.13 Disegnare il tracciato del flip-flop D statico di Figura 12.38, e paragonare l'occupazione di area con quella del flip-flop dinamico riportato in Figura 12.35, utilizzando gli stessi valori di dimensionamento dei transistori NMOS e PMOS.

Riferimenti bibliografici

G.M. Glansford, *Digital Electronic Circuits*, Prentice Hall, Englewood Cliffs, 1988.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

N. Weste, K. Eshraghian, *Principles of CMOS VLSI Design - A system perspective*, 2nd ed., Addison-Wesley, 1993.

J. Millman, A. Grabel, *Microelettronica*, McGraw-Hill Italia, Milano, 1994.

J.F. Wakerly, *Digital Design, Principles and Practices*, 2nd ed., Prentice Hall, Englewood Cliffs, 1994.

13

Memorie

13.1 Introduzione

La maggior parte dei sistemi digitali di una certa complessità contiene degli elementi di memoria, capaci di immagazzinare dati e istruzioni, da fornire in tempi successivi alle unità di elaborazione o alle unità di ingresso/uscita. Sebbene i circuiti sequenziali possano essere considerati dei circuiti di memoria, in quanto possono immagazzinare degli stati logici e fornirli all'uscita, si definiscono memorie quei circuiti che possono contenere un numero elevato di informazioni binarie in maniera organizzata e fornirle in uscita mediante una operazione detta di lettura della memoria stessa.

A seconda della modalità con cui i dati vengono immagazzinati nella memoria e quindi letti in uscita, le memorie si dividono in a) *memorie sequenziali* e b) *memorie ad accesso casuale* (*Random Access Memory* o RAM).

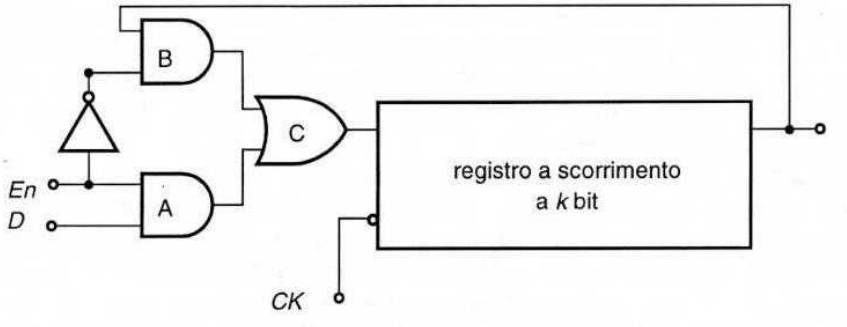


Figura 13.1 Memoria sequenziale a k bit con registro a scorrimento

Nelle prime i dati sono immagazzinati in maniera sequenziale in un supporto che permette la loro registrazione in serie, e anche la lettura avviene in maniera sequenziale, in quanto per leggere in uscita il n^{mo} bit, occorre attendere che scorrano in uscita tutti i bit incamerati precedenti a quello in esame. Un esempio di memoria sequenziale è il nastro magnetico su cui sono registrate le parole digitali nella sequenza in cui queste si presentano alla testina di scrittura, e per accedere ad una data parola in fase di lettura è necessario fare scorrere il nastro sulla testina di lettura fino alla posizione in cui questa è stata registrata. Una memoria di questo tipo realizzata a partire dai circuiti sequenziali visti nel Capitolo 12 è quella che fa uso di un registro a scorrimento con ricircolo, il cui schema di principio è riportato in Figura 13.1 (in questo caso il registro è l'equivalente di un nastro magnetico chiuso in circolo e fatto ruotare continuamente avanti alle testine di scrittura e lettura). Le porte logiche all'ingresso servono per eseguire l'operazione di scrittura della parola di k bit nel registro o per permetterne il ricircolo (e quindi la lettura in maniera sequenziale); in presenza del segnale *En* di abilitazione, la porta AND B si disabilita, e la sequenza dei k bit da memorizzare, presente all'ingresso D della porta AND A, viene inserita nel registro a scorrimento dopo k cicli del clock *CK*. Quando il segnale *En* è basso, la porta A si disabilita e non permette l'ingresso di dati, mentre quella B si abilita e permette il ricircolo dei dati presenti nel registro. La lettura può essere effettuata sia all'ingresso che all'uscita del registro nella fase di ricircolo dei dati, a partire dall'istante in cui si presenta il primo dei bit incamerati.

È evidente che le memorie sequenziali non permettono un rapido accesso all'informazione in quanto, riferendosi ad esempio al registro a scorrimento di Figura 13.1, occorrono n cicli del segnale di clock per estrarre l'informazione corrispondente al bit più lontano dall'uscita in una sequenza di n bit, e questo tempo diventa quindi inaccettabilmente lungo per le operazioni di lettura se la memoria sequenziale deve memorizzare un elevato numero di bit. Una possibilità è quella di connettere più unità di memoria di questo tipo in parallelo, ognuna con il compito di memorizzare un numero ridotto di bit (4, 8, 16 bit), ma in tal caso il numero di componenti elementari tende a crescere eccessivamente perché *per ogni parola* di k bit da memorizzare occorre un registro con k flip-flop D, più i circuiti accessori di scrittura e lettura, e ciò porta a numeri inaccettabili di porte elementari se si vogliono realizzare memorie con capacità già di decine di kbit, anche nel caso di circuiti a logica dinamica. Le memorie basate sui registri a ricircolo sono quindi utilizzate come memorie temporanee per le operazioni sequenziali che coinvolgono singole parole logiche e non come memorie di grande capacità.

Le memorie ad accesso casuale (RAM) sono invece basate su una organizzazione di tipo matriciale delle singole celle di memoria: queste non sono connesse in serie come le celle del registro a scorrimento (o per rimanere nell'esempio già fatto, come i dipoli magnetici deposti lungo il nastro magnetico) ma sono poste sulle intersezioni di una serie di righe e di colonne che costituiscono i terminali da cui si accede al contenuto di informazione della singola cella di memoria, come indicato schematicamente in Figura 13.2. In tal caso l'accesso all'informazione immagazzi-

nata nella cella C_{jk} viene effettuato attraverso l'abilitazione della riga j e della colonna k che corrispondono alla locazione jk; quindi il tempo di accesso alla memoria è uguale per tutte le locazioni e non è in principio influenzato dal numero di bit memorizzabile dalla memoria stessa.

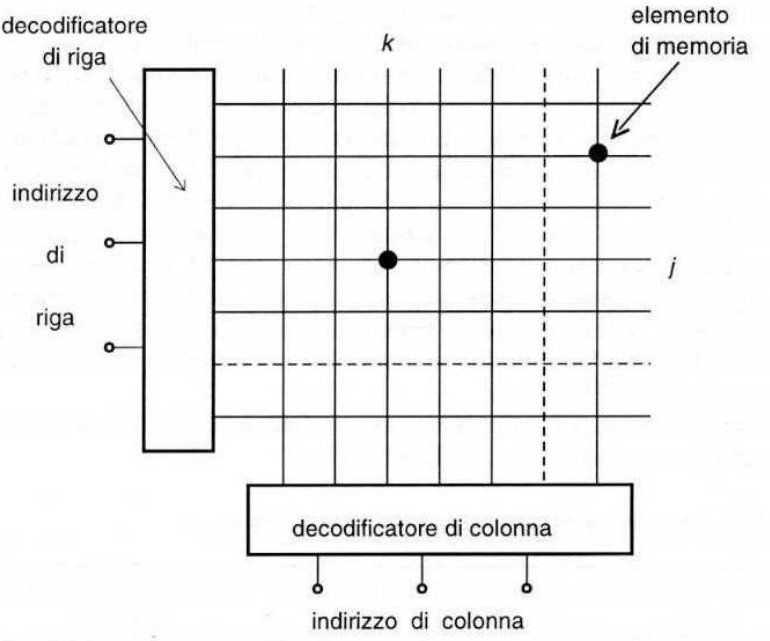


Figura 13.2 Organizzazione a matrice delle memorie ad accesso casuale

Una diversa classificazione per i circuiti di memoria divide le memorie ad accesso casuale in *memorie a sola lettura* (*Read-Only Memory,* ROM) e *memorie a lettura/scrittura* (*Read/Write Memory,* RWM). Nelle prime le informazioni sono immagazzinate nelle celle di memoria all'atto della realizzazione del circuito e possono essere solo lette, indirizzando opportunamente la memoria stessa; nelle seconde le informazioni possono essere ripetutamente scritte nelle singole celle e lette, con operazioni dette di scrittura e lettura. In realtà la dizione di memorie ad accesso casuale (RAM) è oggi riservata alle sole memorie di lettura/scrittura (RWM), pur essendo quelle a sola lettura (ROM) basate anch'esse su un accesso di tipo casuale per la lettura, in altre parole con un tempo di lettura che non dipende dalla locazione del dato nella memoria.

Una ulteriore classificazione per le memorie RAM è quella basata sul tipo di logica utilizzata per la loro realizzazione: le memorie RAM *statiche* (SRAM) sono basate su circuiti a logica statica e quindi l'informazione è conservata nelle celle di memoria per un tempo indefinito o finché non viene scritta una diversa informazione, mentre le memorie RAM *dinamiche* (DRAM) sono realizzate con logiche di-

namiche a MOS, e quindi richiedono un periodico ripristino dell'informazione immagazzinata nelle celle di memoria anche se non viene scritta una nuova informazione.

Infine una classificazione delle memorie riguarda la capacità di conservare le informazioni memorizzate anche quando viene a mancare l'alimentazione del sistema: si dicono *memorie non volatili* quelle memorie che conservano l'informazione anche in assenza di alimentazione elettrica del circuito, e *memorie volatili* quelle in cui l'informazione viene persa in assenza di alimentazione. Le memorie ROM, ad esempio, sono memorie non volatili, mentre le memorie RAM, sia statiche che dinamiche, sono memorie volatili. Tuttavia oggi vengono indicate come memorie non volatili quelle memorie che hanno capacità di scrittura delle informazioni in essa contenute, e non solo di lettura, come nel caso delle ROM, e che possono conservare l'informazione scritta anche quando l'apparato di cui esse fanno parte è disconnesso dall'alimentazione (o, come si dice comunemente, è *spento*). Queste memorie hanno un interesse crescente negli attuali apparati digitali, permettendo la sostituzione di sistemi di memoria di massa come i dischi magnetici, e l'utilizzo di memorie in sistemi sistemi digitali portatili a minimo ingombro (schede "intelligenti").

Una tabella riassuntiva della classificazione delle memorie in base alle caratteristiche suddette è quella riportata in Tabella 13.1.

Tabella 13.1 Classificazione delle memorie ad accesso casuale

memorie a sola lettura (ROM)	memorie non volatili (NVRWM)	memorie a lettura/scrittura (RWM)
ROM	EPROM	SRAM
PROM	EEPROM	DRAM
	FLASH	

Alla base delle differenti scelte tecnologiche utilizzate per la realizzazione dei circuiti di memoria vi è la necessità di memorizzare un numero elevato di bit per chip di memoria (ROM da 128 k × 8 bit sono oggi considerate un componente standard e RAM con capacità di 16 Mbit sono disponibili commercialmente): questo richiede scelte di configurazioni che riducano al minimo il numero di transistori per bit memorizzato.

13.2 Memorie a sola lettura (ROM)

Le memorie a sola lettura (ROM) sono dei circuiti in cui le informazioni consistono in determinate funzioni logiche immagazzinate nella matrice di circuiti combinatori che costituiscono la memoria stessa, e che possono essere presentate alle uscite in funzione degli indirizzi logici forniti agli ingressi. Applicazioni tipiche delle memorie ROM sono ad esempio quella di conservare (e fornire) le istruzioni di un

programma di controllo di un processore o i dati di una tabella di valori (*look-up table*) per realizzare una funzione matematica.

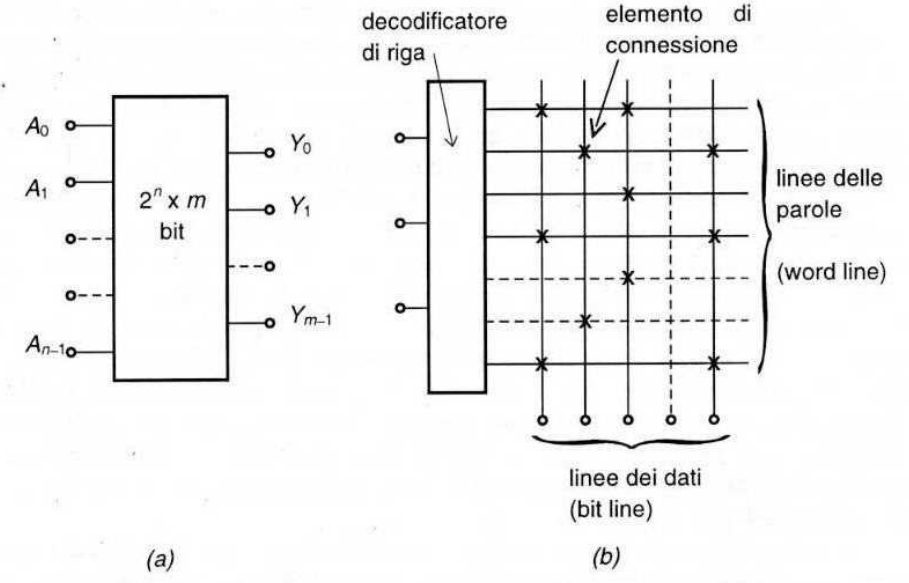


Figura 13.3 a) Simbolo logico di una memoria ROM a $2^n \times m$ bit; b) organizzazione della ROM con decodificatore e matrice di codifica

In senso stretto questi circuiti non appartengono alla classe dei circuiti sequenziali a cui fanno capo gli elementi di memoria, ma a quella dei circuiti combinatori, in quanto forniscono in uscita diverse relazioni combinatorie in funzione degli ingressi abilitati; tuttavia si definiscono memorie (ma a sola lettura) in quanto questi circuiti hanno la possibilità di "memorizzare" queste relazioni nel circuito, ossia di memorizzare una tabella della verità di relazioni tra gli ingressi e le uscite. Questi circuiti forniscono quindi in uscita le diverse parole logiche corrispondenti a queste relazioni in dipendenza dei diversi indirizzi (associati alle singole parole) inviati agli ingressi per effettuare l'operazione di lettura.

Lo schema logico generico per una memoria ROM è quindi quello di un circuito combinatorio, indicato in Figura 13.3a, che fornisce in uscita una serie di dati $Y_0 \div Y_{m-1}$ in corrispondenza di una serie di ingressi (indirizzi) $A_0 \div A_{n-1}$. In effetti con n bit in ingresso si possono avere 2^n combinazioni di parole in uscita, ognuna formata da m bit. Queste "combinazioni" memorizzate nella memoria possono essere definite sia in sede di realizzazione della stessa che in fase di programmazione del circuito, se la memoria è di tipo programmabile, come vedremo in seguito; in ogni caso le informazioni vengono conservate permanentemente nella configurazione del circuito anche se questo non è alimentato, e quindi la memoria è non volatile.

L'organizzazione in termini di blocchi combinatori di una memoria ROM è basata su un circuito di decodifica dell'indirizzo in ingresso, che abilita una delle linee di un circuito di codifica, circuito che presenta in uscita la parola definita in base alla codifica prescelta. La struttura interna è analoga a quella delle matrici logiche di tipo PLA, presentate nel Paragrafo 10.9, descritte genericamente in Figura 10.47 come formate da una matrice di decodifica dell'indirizzo di linea (piano AND) e una matrice per la codifica dei dati da presentare in uscita (piano OR); questa organizzazione è indicata schematicamente in Figura 13.3b. La differenza base della struttura ROM da quella PLA è nel fatto che nelle matrici PLA solo alcune delle combinazioni possibili delle variabili di ingresso sono implementate nelle uscite, mentre la memoria ROM prevede uscite differenti per ogni possibile combinazione degli ingressi (ossia per ogni diversa parola di indirizzo fornita all'ingresso). Il circuito viene considerato una *memoria* perché in effetti le diverse combinazioni presentate in uscita a seguito dell'abilitazione di una delle linee in ingresso al circuito di decodifica (e quindi all'arrivo di una determinata parola di indirizzo) sono registrate (ossia *memorizzate*) nella seconda matrice di codifica. Questa registrazione viene effettuata, come si è visto nei circuiti di codifica del Paragrafo 10.9, mediante l'inserzione di opportuni elementi (diodi, MOS o transistori bipolari) nei nodi nella matrice nei quali si vuole codificare una dipendenza tra bit della riga di ingresso e bit di uscita.

Le scelte delle tecnologie e dei dispositivi da utilizzare per la realizzazione di questi circuiti sono essenzialmente dettate dai requisiti richiesti sul numero di bit (o di parole) da memorizzare, ossia dalla capacità di memoria, e dai tempi di accesso alle informazioni, in questo caso dal tempo di lettura richiesto. Per le memorie ROM ad alta capacità la scelta è orientata su tecnologie MOS e CMOS, mentre le tecnologie bipolari sono riservate alle applicazioni nelle quali vi è necessità di bassi tempi di lettura. Nel seguito di questa sezione dedicata alle memorie ROM faremo riferimento a realizzazioni in tecnologia MOS, che sono come si è detto la scelta preferita per applicazioni ad elevata capacità di memoria.

Nell'ambito della tecnologia MOS, l'uso di porte elementari CMOS non è la scelta migliore a causa della necessità di dover utilizzare $2n$ dispositivi per memorizzare n bit, mentre, come si è visto, i circuiti di codifica e decodifica a NMOS utilizzano solo $n + 1$ dispositivo per bit memorizzato. La soluzione più conveniente è in questo caso l'impiego di reti pseudo-NMOS, che mantengono il numero di dispositivi ancora pari a $n + 1$ pur migliorando i tempi di propagazione t_{PLH} per le caratteristiche migliori presentate dal PMOS che agisce a carico attivo. Anche in questi casi, come si è già visto per i circuiti combinatori a larga scala di integrazione, è possibile aumentare la corrente fornita dal PMOS di carico, migliorando la dinamica e il pilotaggio degli stadi a valle. L'inevitabile peggioramento che ne consegue, sia sul valore del livello logico basso V_{OL} che dei margini di rumore, può essere accettato in quanto il ripristino dei livelli e il filtraggio dai disturbi vengono affidati ai circuiti di interfaccia previsti per la connessione con il resto del sistema, sia all'interno del chip che a maggior ragione ai terminali di ingresso e uscita.

In Figura 13.4 come esempio è riportata una ROM con matrice di codifica NOR in tecnologia pseudo-NMOS. Ricordiamo che nei circuiti di codifica tutte le linee di ingresso vengono poste usualmente al livello basso, e solo la linea indirizzata si porta al livello alto. La codifica desiderata dei bit sulle diverse uscite (*bit line*), per ogni linea di ingresso (*word line*) abilitata, può essere implementata in maniera diretta nella matrice di MOS: se l'incrocio di una word line con una bit line non prevede un MOS pilotato dalla word line, il livello alto (1) della bit line attraversata non verrà alterato; se invece all'intersezione vi è un MOS pilotato dalla word line, questo condurrà quando la word line è alta, e porterà la bit line al livello basso (0).

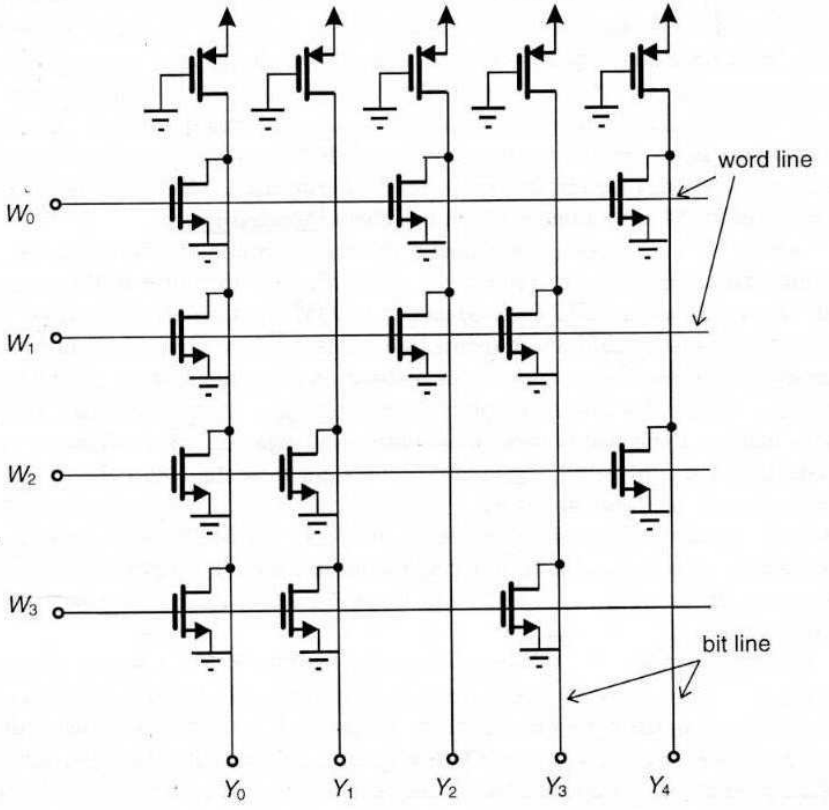


Figura 13.4 Matrice di codifica a porte NOR pseudo-NMOS

In questa matrice, l'abilitazione della linea W_0 fornisce alle uscite $Y_0 \div Y_4$ la parola 01010, quella W_1 la parola 01001, quella W_2 la parola 00110, quella W_3 la parola 00101. La rete di decodifica è analoga a quella a porte NOR NMOS riportata in Figura 10.41, a patto di sostituire i transistori NMOS a svuotamento di carico con i PMOS.

L'introduzione delle informazioni desiderate nella ROM, e cioè la codifica voluta per ogni indirizzo (ossia per ogni word line abilitata) consiste quindi nel definire le posizioni nella matrice nelle quali debbono essere inseriti i MOS, riga per riga. Questa "scrittura" delle informazioni può essere fatta a diversi livelli di realizzazione della ROM.

Una prima modalità è quella della definizione della ROM a livello di realizzazione del tracciato del circuito integrato; questa modalità è quella che permette la maggiore flessibilità di progettazione, ma viene utilizzata solo per applicazioni specifiche, nelle quali la ROM è parte di un circuito integrato più grande e va progettata in questo contesto.

Una seconda possibilità è quella che richiede la realizzazione di una struttura generale da parte dell'industria che realizza il circuito integrato, con MOS presenti in ogni nodo della matrice ma non contattati, in quanto il processo si arresta al livello di realizzazione della metallizzazione. Quest'ultima va in seguito realizzata in accordo con le specifiche relative alla codifica richiesta dall'utilizzatore; in tal modo si definisce il pattern di metallizzazione e apertura di contatti, determinando le posizioni della matrice in cui i transistori MOS vengono contattati e quindi sono attivi nel definire i legami ingresso-uscita. Questo tipo di ROM viene indicato con il nome di ROM programmabile con maschera (*Masked*-ROM).

Infine la terza possibilità è quella di una programmazione effettuata direttamente dall'utilizzatore su componenti standard, detti memorie ROM *programmabili* (*Programmable Read Only Memory,* PROM), indicando con questo nome le memorie programmabili direttamente dall'utente. Queste memorie hanno anch'esse transistori MOS ad ogni nodo della matrice, ma completamente contattati con la metallizzazione. La codifica voluta dall'utilizzatore viene realizzata eliminando dalla matrice i transistori non voluti, attraverso una rottura di elementi circuitali quali fusibili o antifusibili, secondo la tecnica già presentata nel Paragrafo 10.11. In particolare si possono utilizzare i collegamenti fusibili, restringendo opportunamente la pista metallica di connessione dei drain dei MOS; se si applica una alimentazione opportuna alla bit line e alla word line a cui è connesso il transistore da eliminare, si brucia il collegamento di drain e il MOS viene disconnesso dal nodo della matrice.

Queste memorie PROM vengono anche definite ROM a singola programmazione, in quanto possono essere programmate una sola volta e non possono più essere riprogrammate per altre funzioni, in quanto il pattern della matrice di codifica non può essere più alterato. Oltre a queste, vi sono altre famiglie di memorie ROM programmabili e cancellabili, che permettono una grande flessibilità d'uso e su cui ritorneremo successivamente.

13.2.1 Struttura interna delle ROM

La struttura della matrice di codifica di una memoria ROM a MOS è dettata, oltre che dalla necessità di minimizzare l'occupazione d'area, anche da considerazioni legate alla realizzazione di una struttura regolare, in modo da poter definire la codifica voluta con un numero contenuto di operazioni sulla maschera di metallizzazio-

ne per definire la programmazione voluta. Ad esempio, prendendo come riferimento lo schema elettrico semplificato di una matrice a 4 ingressi e 5 uscite di Figura 13.4, si può realizzare una notevole compattazione e una riduzione delle connessioni dei transistori realizzando la connessione tra i source dei MOS presenti tra due successive word line, mediante una linea di diffusione comune a tutti i source, come è indicato in Figura 13.5.

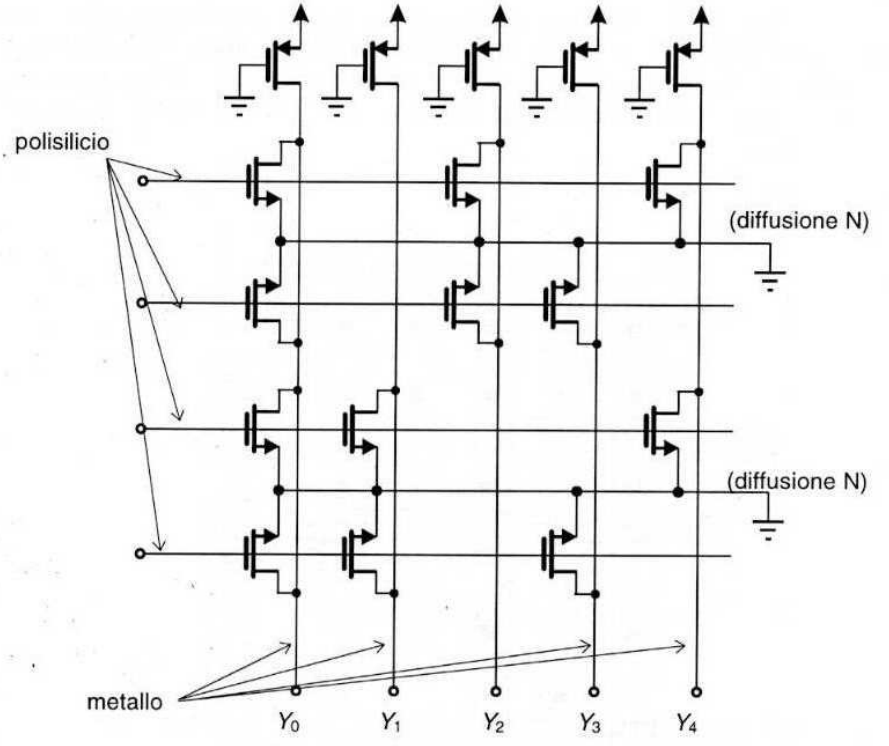


Figura 13.5 Connessioni tra i MOS della ROM

In tal modo si utilizzano tre diverse modalità di collegamento per le tre linee che debbono connettere gate, source e drain dei diversi NMOS; infatti la connessione delle gate dei MOS è realizzata con linee in polisilicio, quella dei source attraverso le linee di diffusione N comuni, e quella dei drain attraverso le linee di metallo. Queste tre modalità permettono una intersezione delle topologie di collegamento in quanto ognuna delle tre linee è separata in verticale dall'altra da uno strato di ossido, e quindi rendono possibile la realizzazione di una struttura a matrice che altrimenti risulterebbe di difficile soluzione.

Il tracciato corrispondente alla configurazione della matrice ROM di Figura 13.5 è riportato in Figura 13.6. Dal tracciato riportato si vede come la configurazio-

ne specifica della ROM, ossia l'inserzione dei NMOS nei nodi voluti della matrice di codifica è stata effettuata su una struttura regolare e periodica, semplicemente definendo la posizione dei contatti (*vias*) lungo le linee di metallo corrispondenti alle bit line della matrice. Le diffusioni di source delle coppie di transistori NMOS pilotati da due word line adiacenti si fondono in un'unica linea di diffusione che viene contattata da una linea di metallo ogni k linee in modo da ridurre la resistenza della diffusione (si noti che la linea di diffusione aumenta la resistenza dei MOS verso massa ma non la capacità, in quanto i source sono tutti connessi a massa). Tutti i transistori NMOS hanno la connessione di gate con la word line (che è in effetti la linea di polisilicio che realizza la gate stessa), ma possono rimanere sconnessi dalla bit line (e quindi non effettuare la funzione NOR dell'ingresso) se il drain non viene connesso alla linea di metallo (bit line) attraverso il contatto (quadrato nero).

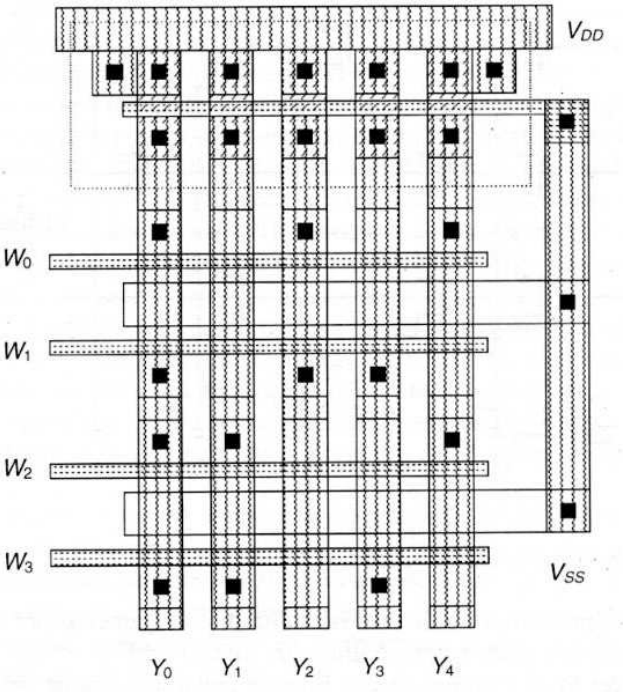


Figura 13.6 Tracciato relativo alla ROM di Figura 13.5; sia i PMOS che i NMOS sono dimensionati con $W/L = 6/2\ \lambda$; la distanza tra le colonne è $4\ \lambda$

Il dimensionamento dei transistori PMOS e NMOS è scelto di solito in modo da migliorare la velocità di risposta e l'uguaglianza dei tempi di propagazione, piuttosto che ridurre il livello logico basso V_{OL}. Come si è visto nel Paragrafo 11.2 ciò comporta una scelta di $K_P = 1/2\ K_N$; nel tracciato in esame si è scelto quindi un

valore del rapporto *W/L* per i PMOS uguale a quello degli NMOS, in modo da ottenere tempi di propagazione $t_{PLH} \cong t_{PHL}$. Il valore di V_{OL} risulta in questo caso molto elevato, circa 1 V, ma questo valore può essere accettato nelle ROM, in quanto in uscita dalle bit line vi saranno opportuni circuiti di lettura e di disaccoppiamento che possono essere resi compatibili con questi valori.

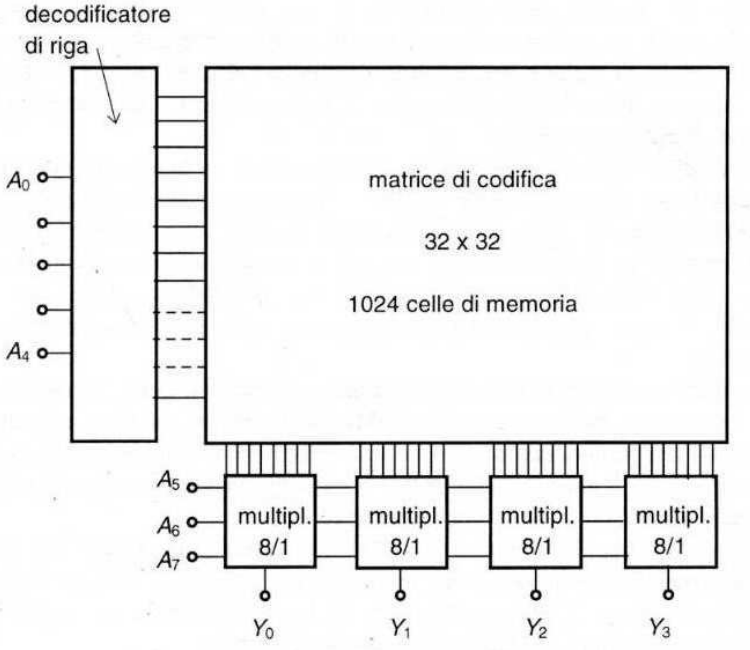


Figura 13.7 Indirizzamento bidimensionale di una ROM

La scelta della dimensione della matrice di decodifica, in termini di prodotto delle righe (word line) per le colonne (bit line), è legata sia al numero di bit delle parole di ingresso e di uscita, sia, come vedremo, alle prestazioni dinamiche della ROM stessa. Una memoria ROM da n bit usualmente prevede tutte le combinazioni possibili degli n bit in ingresso e quindi ha una matrice di codifica che fornisce tutte le 2^n combinazioni degli m bit in uscita. Il numero di bit in uscita è legato alla lunghezza (in bit) delle parole che debbono essere utilizzate dal resto del sistema, di solito 4, 8, 16, 32, 64 bit, mentre le linee di ingresso alla matrice sono 2^n dove n è il numero di bit di indirizzo in ingresso; se il numero di bit degli indirizzi di ingresso non è molto diverso da quello delle parole di uscita, la matrice di codifica (formata da: numero di righe di uscita del decodificatore × numero di colonne dei bit in uscita), risulta fortemente asimmetrica. Ad esempio per una memoria ROM indirizzabile con parole di 8 bit e con 4 bit in uscita, che fornisce $2^8 = 256$ combinazioni di uscite da 4 bit, occorre una matrice di codifica di 256 righe per 4 colonne. È possibile realizzare, per la stessa capacità di memoria (nel nostro esempio

256 × 4 = 1024 bit) una matrice meno rettangolare della prima utilizzando un indirizzamento in due fasi, secondo lo schema indicato in Figura 13.7, e cioè utilizzando un multiplexing delle uscite. Si possono utilizzare ad esempio 5 bit per gli indirizzi di riga (il che corrisponde ad avere 2^5 = 32 righe) e 3 bit di indirizzo per ognuno dei 4 multiplexer 8/1, ognuno dei quali seleziona una delle 8 linee in ingresso a seconda dei 3 bit di indirizzo, fornendo complessivamente i bit alle 4 uscite. In tal caso il numero totale di bit immagazzinati è lo stesso (32 × 32 =1024), come anche è lo stesso il numero di bit di indirizzo e quelli delle parole in uscita, ma la matrice di codifica è quadrata, con una riduzione del numero di righe e un aumento di quello delle colonne. Questo tipo di indirizzamento è detto *bidimensionale* o *X-Y* in quanto l'indirizzo è suddiviso sui due assi *X* e *Y* della matrice.

La scelta di realizzare quanto più quadrata possibile la matrice di codifica della ROM (detta anche nucleo centrale (*core*), in quanto richiede la maggior area del chip dovendo contenere tanti transistori MOS quante sono le intersezioni della matrice e quindi i bit della ROM stessa) è dettata oltre che dalla considerazione di un utilizzo migliore dell'area stessa del chip e da una riduzione del circuito di decodifica, anche (e principalmente) da considerazioni sulle prestazioni dinamiche del circuito stesso.

Come semplice esempio dell'influenza delle dimensioni della matrice di codifica sulle prestazioni dinamiche della ROM, valutiamo in maniera estremamente semplificata il tempo necessario affinché siano disponibili i dati in uscita, in seguito all'applicazione di un indirizzo alla ROM; questo tempo è detto *tempo di lettura*, ed è il parametro dinamico più importante per le ROM.

Consideriamo per semplicità una ROM con capacità di 65.536 bit (indicata usualmente come ROM da 64 Kbit), realizzata mediante una matrice NOR come nell'esempio di Figura 13.5 e nel tracciato di Figura 13.6 (che in questo caso si riferiscono solo ad una parte ridotta della matrice effettiva).

Consideriamo per questa matrice due possibili realizzazioni, la prima con matrice fortemente rettangolare, e cioè 4096 righe × 16 colonne, e la seconda con matrice quadrata, ossia 256 righe × 256 colonne. Il tempo di lettura della ROM, che corrisponde al tempo affinché la combinazione dei bit di indirizzo all'ingresso modifichino la parola in uscita, può essere valutato in maniera estremamente semplificata valutando il tempo necessario affinché ogni bit in ingresso modifichi il livello logico della linea di uscita della matrice di codifica della ROM. Questo tempo si può valutare come somma del tempo necessario per portare una riga di uscita del decodificatore al livello alto (1), assumendo che all'ingresso del decodificatore sia presente l'indirizzo che abilita quella linea, e di quello necessario a portare una colonna della matrice di codifica, pilotata da quella word line, allo stato basso (0).

Dal tracciato della matrice riportato in Figura 13.6, assumendo una *feature size* $\lambda = 0.5\ \mu m$, si ricavano i seguenti valori per le aree e i perimetri di gate e di drain dei singoli MOS:

$$\text{area gate} = 2\lambda \cdot 6\lambda = 3\ \mu m^2\ ;\ \text{perimetro gate} = 2(2\lambda + 6\lambda) = 8\ \mu m \qquad (13.1)$$

$$\text{area drain} = 6\lambda \cdot 6\lambda = 9\ \mu m^2;\ \text{perimetro drain} = 2(6\lambda + 6\lambda) = 12\ \mu m \qquad (13.2)$$

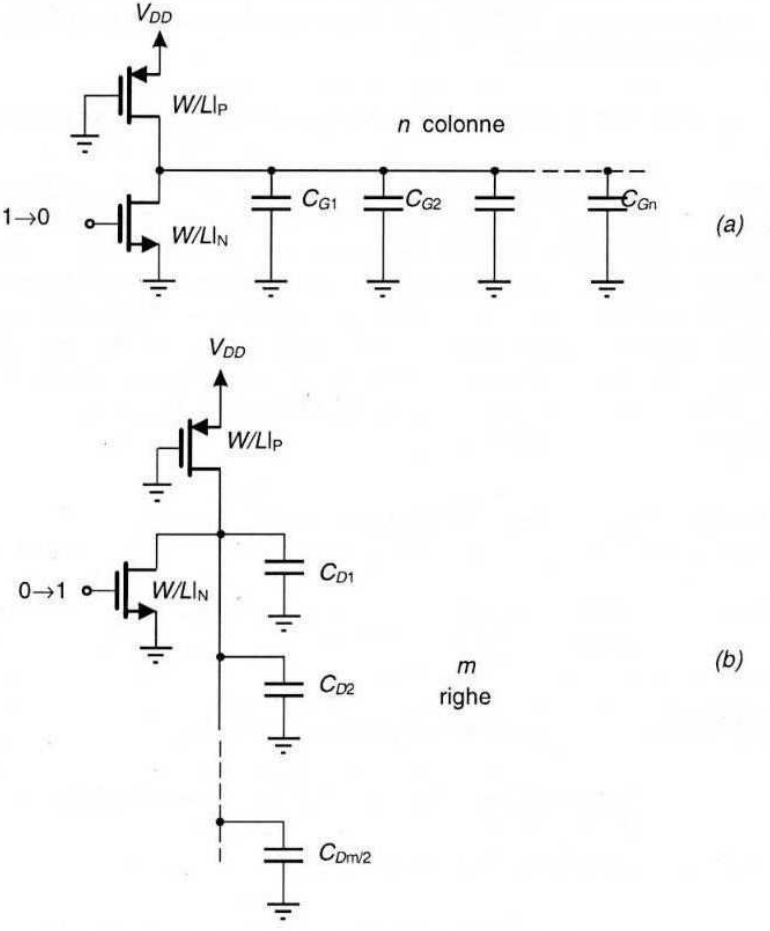


Figura 13.8 Schema elettrico semplificato per la riga (a), e per la colonna (b), della matrice di decodifica dell'esempio svolto

Assumendo come esempio per i valori unitari di capacità di gate e drain quelli riportati in Tabella 3.2, si ottengono i seguenti valori delle capacità dei transistori NMOS, sommando i valori di area e quelli di perimetro:

$$C_G = 1.7 \cdot 3 + 0.2 \cdot 8 = 6.7\ \text{fF}\ ;\ C_D = 0.3 \cdot 9 + 0.4 \cdot 12 = 7.5\ \text{fF} \qquad (13.3)$$

Il tempo di transizione $t_{LH(W.L.)}$ dal valore basso a quello alto per la riga (word line), e quello di transizione $t_{HL(B.L.)}$ dal valore alto a quello basso per la colonna (bit line) possono essere valutati in base agli schemi elettrici di Figura 13.8, assumendo sia la riga che la colonna caricate da un PMOS con $W/L|_P$ = $W/L|_N$ = 3/1 μm. Assumendo una tensione di alimentazione V_{DD} =3.3 V, per considerazioni sulla dissipazione di potenza, le correnti I_P di carica delle capacità di gate attraverso il PMOS, e I_N di scarica delle capacità di drain attraverso il NMOS saranno date da:

$$I_P = 20 \cdot 3 \cdot (V_{DD} - |V_T|)^2 = 318\ \mu A\ ;\ I_N = 50 \cdot 3 \cdot (V_{DD} - V_T)^2 = 795\ \mu A \quad (13.4)$$

e i tempi di transizione vengono calcolati, in analogia con il calcolo dei tempi di propagazione, mediante l'espressione $(C\ \Delta V)/I$, con $\Delta V = V_{DD} - V_{OL} = 3.3 - 1 = 2.3$ V. Nel calcolo di $t_{LH(W.L.)}$ si debbono considerare i contributi delle capacità di gate dei NMOS delle n colonne (in quanto tutti i transistori della matrice hanno i gate connessi alle word line); nel calcolo di $t_{HL(B.L.)}$ si debbono considerare le capacità di drain dei transistori effettivamente connessi alle bit line in funzione della codifica voluta (per le m word line solo $m/2$ posizioni conterranno un transistore e quindi alla bit line saranno connesse $m/2$ capacità di drain). Si ha quindi, per una generica matrice di m righe e n colonne:

$$t_{LH(W.L.)} \cong \frac{C_G \Delta V}{I_P} \cdot (\text{n bit}) = \frac{6.7 \cdot 2.3}{318} \cdot (\text{n bit})\ \text{ns} \quad (13.5)$$

$$t_{HL(B.L.)} \cong \frac{C_D \Delta V}{I_N} \cdot (\frac{\text{m}}{2}\ \text{bit}) = \frac{7.5 \cdot 2.3}{795} \cdot (\frac{\text{m}}{2}\text{bit})\ \text{ns} \quad (13.6)$$

Nel primo caso considerato (m = 4096, n = 16) si ha:

$$t_{LH(W.L.)} = 1\ \text{ns};\quad t_{HL(B.L.)} = 44.4\ \text{ns};\quad t_{tot} = 45.4\ \text{ns} \quad (13.7)$$

Nel secondo caso (m = 256, n = 256) si ha:

$$t_{LH(W.L.)} = 12.4\ \text{ns};\quad t_{HL(B.L.)} = 2.77\ \text{ns};\quad t_{tot} = 15.17\ \text{ns} \quad (13.8)$$

Questa semplice valutazione dei tempi di lettura nei due casi fa comprendere perché si cerchi di realizzare matrici di dimensioni quadrate invece che fortemente rettangolari.

È bene sottolineare che questa valutazione è solo una valutazione di larga massima, anche se permette di comprendere l'influenza delle dimensioni geometriche della matrice sulle prestazioni elettriche. Infatti in questo calcolo non si sono tenuti in conto i ritardi aggiuntivi inseriti dai circuiti di pilotaggio del decodificatore di ingresso e di quelli di disaccoppiamento in uscita; inoltre non si è tenuto conto del

fatto che le righe della matrice sono realizzate in polisilicio anziché in metallo: questo comporta una resistenza non più trascurabile tra le singole capacità, per cui non si può considerare la capacità globale tutta concentrata sul terminale di uscita del MOS, ma si deve considerare la linea di interconnessione come una linea distribuita a celle *RC*, per la quale il ritardo di propagazione cresce con il quadrato del numero di celle, e può essere espresso come $1/2(R \cdot C_G)^2$. Il polisilicio in questo caso è drogato (vedi Tabella 2.1), e quindi presenta una resistenza per quadro di circa 50 Ω/□, ma per un numero relativamente elevato di celle in serie questo ritardo può essere molto più elevato di quello calcolato in base alla (13.5). In questi casi si usa cortocircuitare ad intervalli regolari la linea di polisilicio con una di metallo, in modo da ridurre il numero di celle contenute tra i punti in cui è connessa la linea di metallo, e contenere quindi il ritardo globale a valori accettabili. Ad esempio, volendo valutare il ritardo della word line assumendo per il tempo di propagazione l'espressione per una linea *RC* pari a $1/2(R \cdot C_G)^2$, tenendo conto di una resistenza del polisilicio di 50 Ω/□, e valutando dal tracciato di Figura 13.6 una distanza di 10 λ tra le celle (gate), si ha una resistenza R per cella di $50 \cdot 5 = 250\ \Omega$, e si ottiene un valore del ritardo, nel secondo caso ($m = 256$):

$$t_{(W.L)} = \frac{250 \cdot 6.7 \cdot 10^{-15}}{2} (256)^2 = 54 \text{ ns} \tag{13.9}$$

che è ben maggiore di quello calcolato con la (13.5), per cui occorre limitare il numero di celle che possono essere pilotate dalla linea in polisilicio tra due successivi cortocircuiti della linea di metallo. In alternativa si può alimentare la linea di polisilicio dalle due estremità in modo da ridurre la lunghezza equivalente vista da un estremo.

Per quanto riguarda la potenza dissipata, occorre ricordare che le soluzioni presentate, basate su configurazioni pseudo-NMOS, presentano una dissipazione di potenza statica, che è aggravata dalla scelta di voler utilizzare per i PMOS di carico un rapporto *W/L* relativamente elevato in modo da migliorare le caratteristiche dinamiche. Ad esempio, nella ROM analizzata precedentemente, ammettendo che in media una metà delle colonne siano basse (e quindi che i MOS delle colonne relative siano in conduzione), si ha un assorbimento di corrente pari a:

$$P_D = V_{DD} \cdot I_P \cdot \frac{n}{2} = 3.3 \cdot 318 \cdot 10^{-6} \cdot 128 = 134.3 \text{ mW} \tag{13.10}$$

che è un valore troppo elevato, se si pensa alla capacità relativamente ridotta della ROM. Per memorie a elevata capacità la soluzione è quella di utilizzare i circuiti a logica dinamica, come la logica domino. In effetti la realizzazione di una ROM in logica domino è analoga a quella, già esaminata, di una matrice logica PLA, per cui si rimanda a queste ultime per gli aspetti realizzativi.

13.3 Memorie non volatili (NVRWM)

Le memorie viste nei paragrafi precedenti non permettono di registrare le informazioni se non in sede di realizzazione della memoria stessa (masked ROM), o in sede di prima definizione della memoria da parte dell'utente (PROM). In questo paragrafo tratteremo delle memorie ROM in cui l'informazione può essere programmata dall'utente stesso per via elettrica, e può essere in seguito cancellata per programmare una nuova serie di istruzioni. Una prima famiglia è quella delle ROM *programmabili e cancellabili*, dette EPROM (*Erasable Programmable Read Only Memory*), ossia memorie che possono essere programmate elettricamente, ma non cancellabili per via elettrica. Queste memorie utilizzano nella matrice di codifica particolari dispositivi, detti FAMOS (*Floating-gate Avalanche-injection MOS*), che sono ottenuti dalla tecnologia MOS mediante particolari modifiche della struttura base.

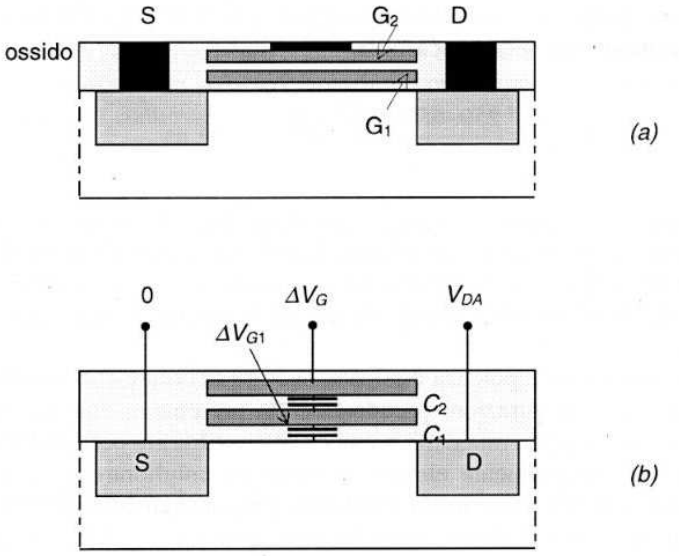


Figura 13.9 a) Dispositivo FAMOS per EPROM; b) polarizzazione del gate isolato mediante partitore capacitivo

I dispositivi FAMOS sono basati sulla presenza di due gate sovrapposte sul canale tra source e drain, come è indicato in Figura 13.9a. In questo dispositivo, la gate più vicina al canale è realizzata da uno strato di polisilicio completamente isolato perché circondato dall'ossido, mentre la gate posta superiormente è connessa all'ingresso come in un normale MOS. La programmazione viene effettuata portando la gate connessa con il terminale esterno ad una elevata tensione positiva ΔV_G, e contemporaneamente polarizzando positivamente il drain a tensioni V_{DA}

superiori a quelle normali di esercizio; attraverso il partitore capacitivo la gate interna (isolata) si porta ad un potenziale $\Delta V_{G1} \cong \Delta V_G/2$ se gli spessori t_1 e t_2 dell'ossido sono circa uguali (Figura 13.9b). Gli elettroni che circolano nel canale raggiungono energie sufficientemente elevate verso il drain (dove il campo è più elevato) fino a creare una ionizzazione a valanga, e contemporaneamente il campo elevato trasversale applicato tra il canale e la gate permette a qualcuno di questi di superare la barriera dell'ossido e raggiungere la gate isolata G_1. Queste cariche si accumulano sul gate e non possono più sfuggire da questa in quanto l'ossido è un perfetto isolante, per cui rendono negativo G_1. Il fenomeno è auto-limitante in quanto, all'aumentare della carica negativa accumulata su G_1, la gate diventa sempre più negativa, fino ad impedire il passaggio di ulteriori cariche nel canale e quindi ad interdire il MOS. Quando le tensioni di programmazione vengono rimosse, la carica negativa accumulata su G_1 mantiene il MOS interdetto per le normali tensioni di ingresso applicabili sulla gate G_2. Ciò equivale a rimuovere la presenza del MOS dal nodo in questione, o in altre parole a memorizzare permanentemente un 1 logico (livello alto) in quel nodo. Se invece il FAMOS non viene programmato, esso funziona come un normale NMOS connesso al nodo stesso, e va in conduzione se viene applicato un livello alto in ingresso alla gate G_2.

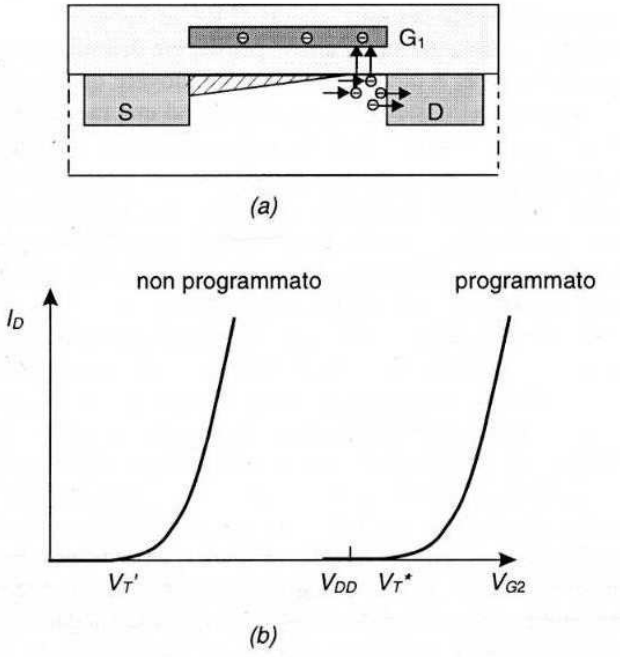


Figura 13.10 a) Iniezione delle cariche per la programmazione del FAMOS; b) caratteristica di trasferimento con e senza programmazione

La tensione di gate V_{G2} in assenza di programmazione è data da:

$$V_{G2} = \left(1 + \frac{C_1}{C_2}\right) V_{G1} \quad \Rightarrow \quad V_T' = \left(1 + \frac{C_1}{C_2}\right) V_T \tag{13.11}$$

e, assumendo che lo spessore di ossido tra G_1 e G_2 sia uguale a quello tra G_1 e silicio, e quest'ultimo sia pari a quello di un NMOS standard, si ha che $C_1 \cong C_2$, per cui quando $V_{G1} = V_T$, $V_T' \cong 2\ V_T$. In seguito alla programmazione, si accumula una carica $-Q$ sulla gate G_1, e la tensione V_{G2} si può determinare come:

$$C_1 V_{G1} + C_2 (V_{G1} - V_{G2}) = -Q \quad \Rightarrow \quad V_{G2} = \left(1 + \frac{C_1}{C_2}\right) V_{G1} + \frac{Q}{C_2} \tag{13.12}$$

Anche in questo caso, il valore che deve assumere V_{G2} affinché V_{G1} sia pari a V_T fornisce il nuovo valore di soglia V_T*, dato da:

$$V_{G2} = \left(1 + \frac{C_1}{C_2}\right) V_T + \frac{Q}{C_2} = V_T * \tag{13.13}$$

e il dispositivo resta interdetto per tutti i valori di V_G, se $V_T* > V_{DD}$.

La carica rimane accumulata su G_1 anche per decine di anni, per cui la memoria programmata è non volatile; è tuttavia possibile rimuovere questa carica (da cui il nome EPROM) esponendo l'ossido che circonda G_1 ad una radiazione ultravioletta; questa esposizione alla radiazione rende l'ossido debolmente conduttore per cui la carica immagazzinata svanisce dopo una esposizione di qualche decina di minuti. Per riprogrammare le EPROM, è necessario prevedere una finestra trasparente nel contenitore plastico del chip, in modo da permettere l'esposizione dell'ossido alla radiazione ultravioletta; naturalmente in questo modo si annullano tutte le programmazioni eseguite sui diversi FAMOS della matrice.

La memoria EEPROM (*Electrically Erasable Programmable ROM*) è simile ad una EPROM, con la possibilità ulteriore di poter cancellare singolarmente per via elettrica le singole informazioni immagazzinate. I dispositivi utilizzati sono ancora MOS a doppia gate, ma queste si estendono parzialmente sulla regione di drain (vedi Figura 13.11), e in questa sovrapposizione lo spessore dell'ossido è molto più sottile (inferiore a 10 nm). In tal caso, applicando sempre una tensione positiva elevata alla gate G_1, gli elettroni del drain possono passare attraverso il sottile strato di ossido sulla gate G_1 per effetto tunnel. In questo caso la programmazione del MOS (interdetto per la carica negativa degli elettroni che hanno raggiunto G_1) può essere annullata applicando un impulso di polarità opposta (negativa) su G_2; questo crea ancora un effetto tunnel ma con passaggio di elettroni da G_1 verso il drain, annullando la carica negativa accumulata su G_1 e ripristinando il normale modo di operazione del MOS.

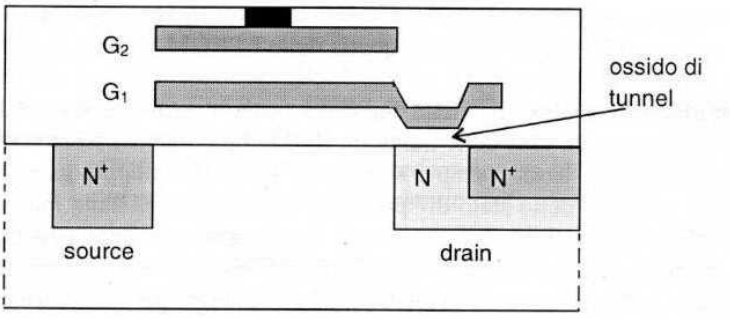


Figura 13.11 MOS a doppia gate per memorie EEPROM

In Figura 3.12 si può vedere un possibile tracciato di questa struttura. La regione del drain (con drogaggio inferiore a quello normale) in cui si può manifestare l'effetto tunnel è esposta alla sola gate isolata G_1, mentre sia G_1 che G_2 agiscono tra le regioni di source e drain del MOS definendo in effetti la condizione di programmazione (conduzione o interdizione) del MOS.

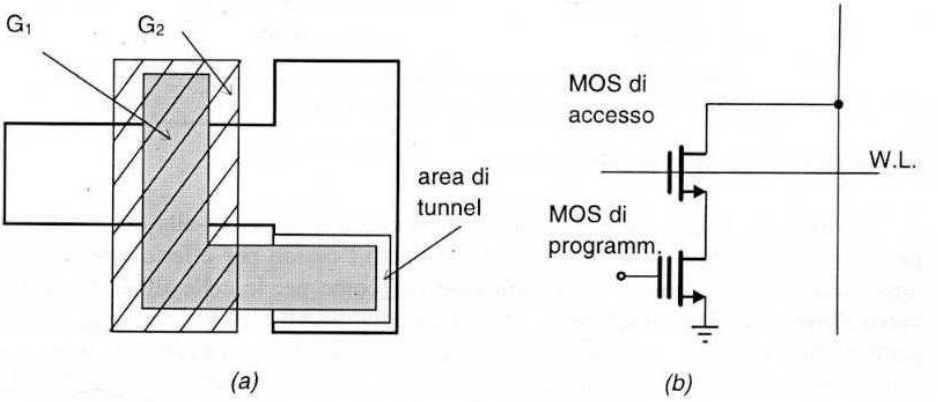


Figura 13.12 a) Tracciato di un MOS a doppia gate per EEPROM; b) connessione alla bit line mediante transistore di accesso

L'operazione di cancellazione attraverso un effetto tunnel inverso tra G_1 e drain può portare ad una "sovrascarica" della gate G_1, ossia ad una iniezione di cariche positive che portano il MOS in conduzione con un canale sempre aperto, anche quando la tensione V_{G2} è nulla. Si inserisce quindi in serie al MOS a doppia gate (che viene utilizzato per la programmazione), un ulteriore MOS detto "di accesso", che viene abilitato dalla word line quando si vuole leggere l'informazione contenuta nel MOS a doppia gate. Questa necessità di impiego di un MOS di accesso per

ogni MOS di programmazione aumenta di un fattore circa 2 l'occupazione di area delle memorie EEPROM rispetto a quella delle EPROM a parità di capacità di memoria.

Un'ulteriore evoluzione delle memorie programmabili è data dalle memorie Flash, che sono anch'esse delle memorie ROM che possono essere scritte e cancellate elettricamente, basate sempre su dispositivi MOS a doppia gate per la memorizzazione elettrica dello stato di "presenza" o "rimozione" dalla matrice di codifica, ma che utilizzano per i dispositivi a doppia gate entrambi i meccanismi visti rispettivamente per le EPROM e le EEPROM, ossia una programmazione basata sull'effetto di iniezione in G_1 di elettroni ad alta energia per ionizzazione valanga, e una cancellazione basata sull'estrazione degli elettroni da G_1 per effetto tunnel attraverso un ossido sottile al source. La struttura di un MOS per memoria Flash è riportata in Figura 13.13.

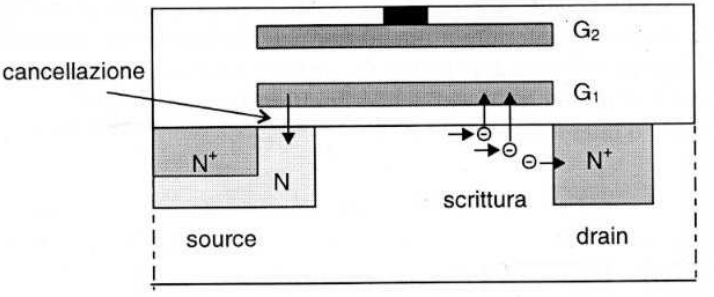


Figura 13.13 Cella MOS per memoria Flash

L'ossido tra la gate G_1 e il silicio ha uno spessore inferiore alla decina di nm per permettere il passaggio di elettroni attraverso l'ossido per effetto tunnel sotto opportuna polarizzazione. La scrittura avviene (come per le celle EPROM) attraverso l'iniezione in G_1 di elettroni ad alta energia dovuti alla ionizzazione da impatto vicino al drain, con polarizzazione elevata sia della gate che del drain. La cancellazione avviene mediante estrazione degli elettroni per effetto tunnel tra la gate G_1 e la regione poco drogata del source, polarizzando negativamente G_1 e positivamente il source. In base alla configurazione del source e alla polarizzazione adottata non si ha l'effetto di sovrascarica di G_1 per cui non è necessario l'impiego del transistore MOS di accesso, e quindi, a parità di capacità di memoria, l'area utilizzata da una memoria Flash è prossima a quella di una memoria ROM o EPROM.

Le memorie Flash hanno un'importanza crescente nei sistemi digitali, in quanto permettono di programmare e cancellare le istruzioni contenute in memorie ROM di alta capacità di memoria (4 MBit, 16 MBit), con elevato numero di cicli di cancellazione ($\cong 10^5$ cicli), tempi di programmazione di qualche μs, tempi di cancellazione di secondi, e tempi di lettura comparabili a quelli delle normali ROM.

13.4 Memorie a lettura e scrittura (RAM)

Anche le memorie a lettura e scrittura (RWM) sono organizzate secondo uno schema a matrice per permettere l'accesso diretto ad ognuna delle celle elementari di memoria, e questo sia per la lettura dei bit immagazzinati che per la loro scrittura nelle celle stesse; in effetti l'acronimo RAM (*Random Access Memory*), che individua la struttura di indirizzamento per righe e colonne della memoria, è essenzialmente utilizzato per identificare questo tipo di memorie; uno schema generale di organizzazione della memoria è riportato in Figura 13.14.

In queste memorie l'indirizzamento delle celle è sempre bidimensionale, cioè vanno selezionate con opportune parole di indirizzo sia le righe che le colonne, e la matrice è di solito organizzata in forma quadrata, cioè vi sono 2^n righe e 2^n colonne che vanno selezionate con parole di indirizzo da n bit attraverso opportuni decodificatori di riga e di colonna. Una volta selezionata la cella di memoria opportuna, occorre utilizzare opportuni circuiti aggiuntivi per scrivere o leggere le informazioni nella cella stessa; questi circuiti sono indicati come circuiti di lettura/scrittura e vengono abilitati dalla funzione richiesta per l'operazione voluta sulla memoria. Le operazioni di lettura e scrittura dei bit nelle singole celle selezionate avvengono tramite le colonne della matrice, che vengono anche per queste memorie indicate come *linee dei dati* (*bit lines*), mentre le righe della matrice che servono (insieme alle colonne) per selezionare le celle di memoria, vengono dette *linee di parola* (*word lines*).

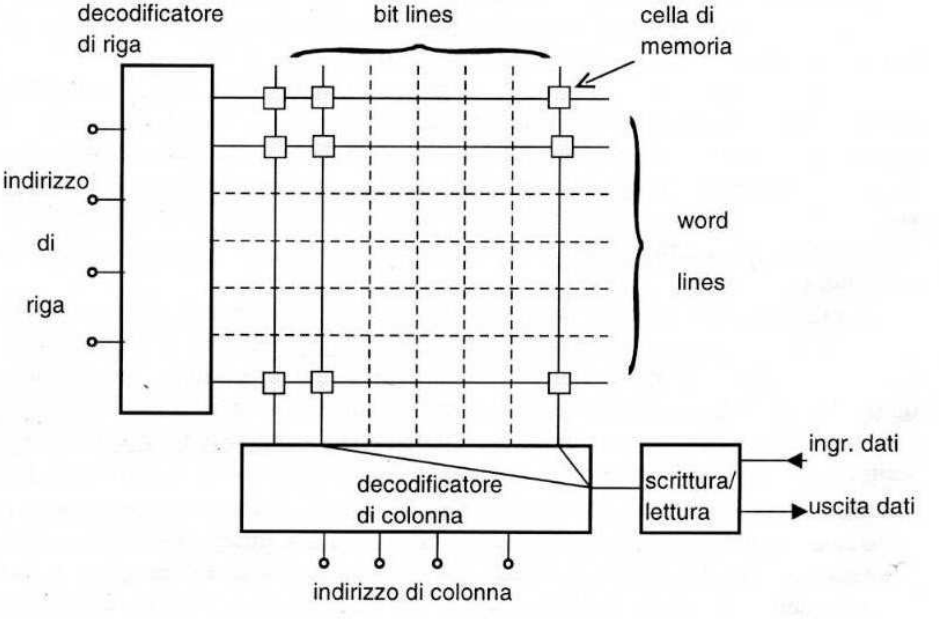


Figura 13.14 Organizzazione in matrice di una memoria RAM

Dallo schema generale di Figura 13.14 si identificano i seguenti sottosistemi fondamentali per il funzionamento di una memoria RAM:

- decodificatore di riga
- decodificatore di colonna
- celle elementari di memoria
- circuiti di lettura e scrittura

Discuteremo nel seguito questi componenti, tralasciando i decodificatori che sono stati già presentati nei circuiti combinatori e nelle memorie ROM.

13.5 Celle elementari per RAM statiche (SRAM)

Nelle memorie RAM il maggior numero di dispositivi è utilizzato per realizzare le celle elementari di memoria, in quanto occorrono tante celle elementari quanti sono i bit immagazzinabili dalla memoria stessa, mentre i circuiti decodificatori richiedono un numero molto inferiore di componenti. Ad esempio si consideri il caso di una memoria RAM indirizzabile con parole da 16 bit, ossia 8 bit per le righe e 8 bit per le colonne. I decodificatori di riga e colonna (da 8 a $2^8 = 256$) richiederanno $8 \times 256 = 2048$ dispositivi ciascuno, gli amplificatori di lettura e scrittura (uno per colonna, come vedremo) saranno 256, mentre le celle di memoria debbono essere $2^8 \times 2^8 = 65.536$ (questa memoria viene sinteticamente indicata come memoria da 64 kbit, dove $k = 2^{10}$ bit $\cong 1000$).

Questo esempio mostra come sia essenziale ridurre il numero di componenti per singola cella di memoria, la quale deve essere realizzata nella forma più elementare possibile rispetto ai circuiti bistabili visti nel Capitolo 12.

Anche in questo caso, come si è visto per i circuiti PLA e per le memorie ROM, la necessità di realizzare memorie di grande capacità porta a diminuire le richieste in termini di dinamica e di margini di rumore delle celle elementari, a favore di una elevata integrazione del circuito, e comporta quindi maggiore attenzione nella progettazione dei circuiti per l'invio e il prelievo delle informazioni, ossia i circuiti di lettura e scrittura, che sono gli elementi più critici nella progettazione del sistema.

In Figura 13.15 è rappresentata schematicamente la connessione delle celle elementari alle righe e colonne della matrice utilizzata nelle memorie di capacità inferiore al Mbit. La riga viene utilizzata come ingresso di abilitazione per le celle e quindi la riga selezionata abilita tutte le celle connesse a quella riga. La colonna selezionata definisce la cella in cui si effettueranno le operazioni di lettura o di scrittura. Si può notare come ogni colonna venga in realtà divisa in due linee, una che porta il dato D e l'altra che porta il dato negato $\overline{D}$; ciò serve essenzialmente a realizzare più efficacemente le operazioni di lettura e scrittura, attraverso una ridondanza di collegamenti che permette di semplificare le singole operazioni, come vedremo nell'analisi delle celle di memoria. Le due linee dati vengono caricate da un'opportuna rete che si differenzia, come vedremo, a seconda del tipo di celle di memoria utilizzate. Al termine di ogni coppia di linee dati che corrispondono alla

colonna nominale k vi sono gli amplificatori di lettura/scrittura, che vengono abilitati dal decodificatore di colonna. Quest'ultimo opera come un multiplexer perché permette anche l'instradamento dei dati (in ingresso o in uscita) alla linea abilitata.

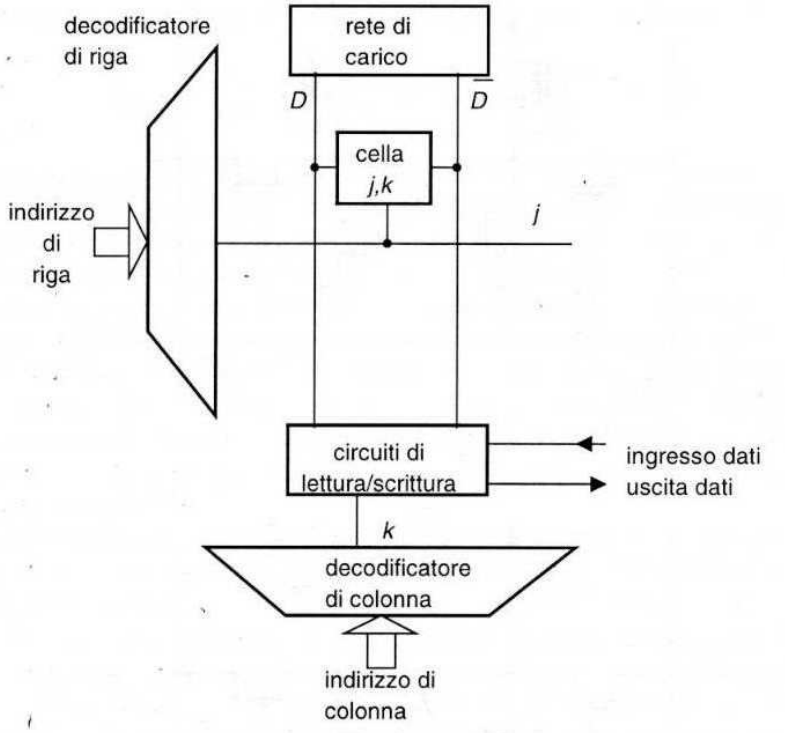


Figura 13.15 Organizzazione della lettura/scrittura dei dati nelle celle

Le celle di memoria andranno quindi realizzate utilizzando il circuito bistabile più elementare, formato da due invertitori in cascata, derivato dallo schema base di Figura 12.3; per poter accedere al bistabile in fase di lettura o scrittura dei bit nella memoria è necessario tuttavia aggiungere due porte di trasmissione agli ingressi X_{i1} e X_{i2} per la connessione alle linee dati.

13.5.1 Celle in tecnologia MOS

La versione della cella di memoria realizzata in tecnologia NMOS è quella riportata in Figura 13.16; le porte di accesso sono realizzate con porte di trasmissione a NMOS, e vengono abilitate (aperte) dalla word line, che passa al livello alto quando essa è abilitata.

Per questa tecnologia, che usa una logica "a rapporto", il problema maggiore nella realizzazione delle memorie RAM (ed in genere di circuiti a larga scala di

integrazione) è quello della potenza dissipabile, in quanto nelle singole celle si dissipa potenza anche nello stato quiescente.

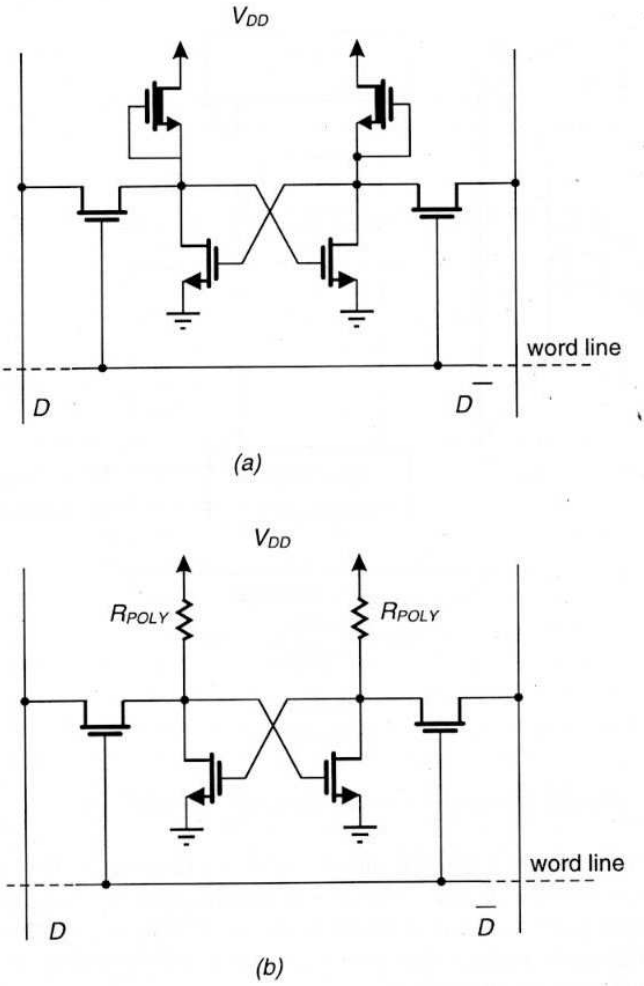


Figura 13.16 Cella di memoria in tecnologia NMOS; a) con invertitori E-D; b) con invertitori a carico resistivo

Un semplice esempio permette di vedere i vincoli che questa dissipazione comporta nella progettazione della memoria. Supponiamo di dovere realizzare una memoria da 128 k, il che comporta la presenza di ~ 128.000 celle elementari; il circuito integrato va inserito in un contenitore dual-in-line che può dissipare una potenza termica di 320 mW. La potenza dissipabile per cella vale quindi:

$$P_{Dcella} = \frac{320 \cdot 10^{-3}}{128 \cdot 10^{3}} = 2.5 \cdot 10^{-6} \text{ W / cella} \qquad (13.14)$$

da cui, assumendo una tensione di alimentazione V_{DD} = 5 V e ricordando che la potenza dissipata dal bistabile corrisponde a quella dissipata nell'invertitore che presenta uscita bassa, si ha per la corrente assorbita I_L nello stato basso:

$$P_{Dcella} = P_{DL} = V_{DD} \cdot I_L \;\Rightarrow\; I_L = \frac{2.5 \cdot 10^{-6}}{5} = 0.5\,\mu\text{A} \qquad (13.15)$$

Questa corrente deve essere fornita dall'elemento di carico del NMOS, che deve quindi limitare la corrente a valori estremamente bassi. Ciò spiega perché non viene utilizzata la versione dell'invertitore con carico realizzato da un NMOS a svuotamento: quest'ultimo dovrebbe avere un rapporto W/L minore di 10^{-2} per ridurre la corrente a valori inferiori al μA, e quindi dovrebbe presentare lunghezze di canale L inaccettabilmente grandi, con eccessiva occupazione di area. La soluzione possibile è con carico resistivo realizzato con polisilicio, ma non drogato, che può assumere valori di resistenza di strato dell'ordine di ~ $10^6 \div 10^8$ Ω/□. Supponendo di utilizzare polisilicio con resistenza di strato di 5MΩ/□, per realizzare resistenze dell'ordine di $V_{DD}/I_L \cong 10$ MΩ, bastano 2 quadrati di lato 2λ, per una lunghezza totale di 4λ.

L'operazione più complessa nel funzionamento delle memorie RAM è quella di lettura del dato immagazzinato nella cella; infatti mentre è relativamente semplice scrivere un bit nelle celle, in quanto si tratta di forzare un ingresso al valore alto e l'altro al valore basso attraverso opportune tensioni applicate alle linee D e $\overline{D}$, non è altrettanto semplice leggere il bit, cioè lo stato di uscita alta o bassa per le due uscite del bistabile, lasciando tuttavia inalterata l'informazione contenuta nella memoria. Questa operazione infatti richiede che le tensioni sulle due uscite del bistabile (V_{OH} e V_{OL}) modifichino i valori delle tensioni presenti sulle linee D e $\overline{D}$, quando queste vengono messe in contatto con le due uscite del bistabile attraverso le porte di trasmissione. Ricordiamo che il bistabile elementare è un circuito sequenziale "trasparente", in quanto le sue uscite sono direttamente connesse agli ingressi; poiché le tensioni delle linee non saranno ovviamente (almeno per una delle due) uguali a quelle degli ingressi del bistabile, vi è il rischio che l'operazione di lettura faccia "commutare" il bistabile, o, in altre parole, che durante la lettura venga cancellata l'informazione in memoria. Questo rischio è tanto più elevato quanto più piccola è la corrente in uscita del bistabile; quindi per le celle NMOS, che presentano correnti deboli per contenere la potenza dissipata, il problema si aggrava.

Per chiarire ulteriormente questo punto, facciamo riferimento al caso in cui l'informazione della cella di memoria elementare a NMOS con carico resistivo in polisilicio debba essere letta mediante la connessione alle linee dati D e $\overline{D}$, polarizzate entrambi al valore intermedio $V_{DD}/2$ attraverso una rete di carico R_L, secondo lo schema indicato in Figura 13.17a.

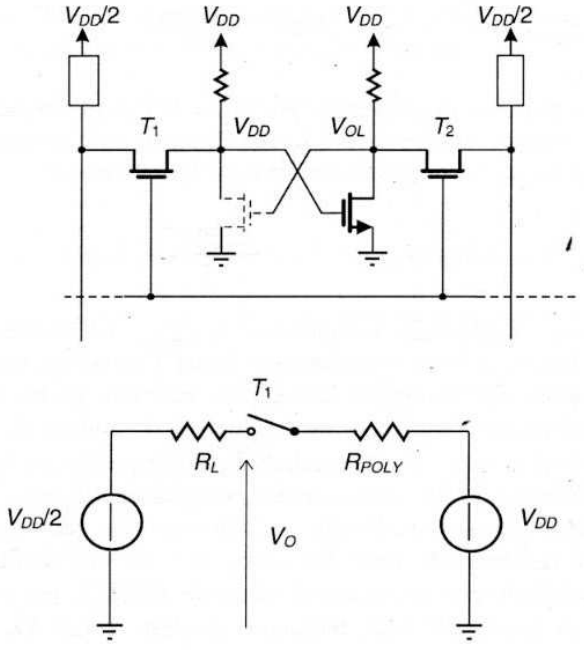


Figura 13.17 a) Cella NMOS con carico resistivo e linee con carico attivo; b) circuito equivalente per la determinazione di V_O alla chiusura della porta di trasmissione

Il circuito equivalente in regime statico della rete formata dalla linea D e l'uscita alta V_{DD} del bistabile è indicato in Figura 13.17b; sia la linea che l'uscita del bistabile sono rappresentate secondo il teorema di Thevenin da due generatori ideali di tensione che forniscono le rispettive tensioni a vuoto, con in serie le resistenze interne. Queste sono per il circuito di polarizzazione della linea, la resistenza R_L, e per l'uscita del bistabile la resistenza R_{POLY}, in quanto il NMOS è interdetto (si è trascurata rispetto a R_{POLY} la resistenza interna della porta T). Poiché $R_{POLY} >> R_L$, la tensione di uscita sulla linea rimarrà praticamente inalterata al valore $V_{DD}/2$, mentre la tensione di uscita V_O del bistabile verrà anch'essa portata al valore $V_{DD}/2$; quando la porta T si richiude il bistabile riparte dalla condizione instabile $V_{DD}/2$ ritornando ad una delle due condizioni stabili che non può essere predetta, per cui l'informazione memorizzata precedentemente viene persa.

Una soluzione che riduce il rischio di alterare l'informazione nella cella durante la lettura è quella di precaricare le linee dati prima dell'operazione di lettura ad una tensione determinata, e di lasciarle isolate, cioè come capacità caricate, durante la lettura. In tal caso, con riferimento all'analisi precedente e allo schema equivalente di Figura 13.17b, le linee vengono rappresentate da un generatore a resistenza infinita (in quanto esse risultano aperte nell'istante di lettura), e quindi non vi è il rischio di una cancellazione dell'informazione contenuta nel bistabile. Le linee sono in questo caso sostituite nel circuito dalle capacità che esse presentano, entrambe

caricate ad un valore prefissato; quando le porte di trasmissione si aprono, le capacità delle linee D e $\overline{D}$ vengono collegate alle due uscite del bistabile, e quella delle due che ha una tensione differente da quella dell'uscita corrispondente viene caricata o scaricata attraverso la corrente di uscita del bistabile.

Ricordando quanto detto precedentemente circa il valore della resistenza di carico dell'invertitore NMOS, si comprende facilmente che non conviene scegliere come valore di precarica delle linee il valore logico basso, in quanto in tal caso il tempo di lettura risulta troppo elevato. Ad esempio, considerando un chip di memoria di circa 1 cm di lato (per una memoria da SRAM da 64 k), e supponendo una lunghezza delle colonne di 8 mm, con una larghezza del metallo della linea di 4 μm, si ha una superficie totale della linea: $A = L \cdot W = 3.2 \cdot 10^2\ \mu m^2$; assumendo uno spessore dell'ossido di campo di 1 μm, si ha per la capacità della linea:

$$C_{metal} = 3 \cdot 10^{-5}\ \text{pF}/\mu m^2;\quad C_{TOT} = 3 \cdot 10^{-5} \cdot 3.2 \cdot 10^4 \cong 1\ \text{pF} \qquad (13.16)$$

La linea che è connessa all'uscita alta del bistabile si caricherà attraverso la resistenza di carico R_{POLY} di quest'ultimo. La costante di tempo della carica di questa capacità attraverso la resistenza di carico del NMOS interdetto è quindi dell'ordine di $R_{POLY} C_{TOT} \cong 10$ μs, e il tempo di lettura è inaccettabilmente lungo. Una soluzione migliore è quella di precaricare le due linee al livello logico alto, ossia alla tensione di alimentazione. In tal caso la capacità della linea connessa all'uscita bassa del bistabile si scaricherà al valore V_{OL} attraverso il NMOS che conduce, e quindi con una corrente di scarica ben più elevata ($\cong 800$ μA con $W/L = 1$) di quella di carica (0.5 μA), per cui il tempo di lettura diviene accettabile.

Lo schema di una cella di memoria NMOS con linee precaricate a V_{DD} mediante due PMOS è riportato in Figura 13.18a; la scelta di transistori PMOS per la precarica è da preferirsi perché l'impiego di transistori NMOS per la rete di precarica avrebbe portato ad una tensione di precarica sulla capacità di linea pari a $V_{DD} - V_T$, in base a quanto detto per le porte di trasmissione NMOS e PMOS (si consideri che il MOS di precarica agisce come una porta di trasmissione che ha in ingresso la tensione V_{DD} e in uscita la capacità di linea). In Figura 13.18b è indicata la necessaria temporizzazione dei segnali di abilitazione della rete di precarica, della word line, e le uscite delle linee D e $\overline{D}$. La fase di precarica deve essere terminata prima che venga abilitata la word line, e cioè prima che inizi la fase di lettura. Si suppone che la cella di memoria abbia immagazzinato un 1, cioè sia alta l'uscita collegata alla linea D. Le due linee, originariamente caricate a V_{DD}, dopo la fase di lettura si differenziano nella tensione immagazzinata sulle rispettive capacità, e quella $\overline{D}$ si porta al livello basso. La lettura del bit immagazzinato (1 o 0) viene quindi effettuata identificando la differenza di tensioni sulle due linee mediante un opportuno circuito di lettura. Per evitare che la cella possa commutare involontariamente nell'altro stato stabile a seguito dell'operazione di lettura occorre che i transistori delle porte di trasmissione T_1 e T_2 non siano a bassissima resistenza, in quanto in tal

caso tutti e due gli ingressi del bistabile vedrebbero una tensione V_{DD} e quindi il circuito potrebbe portarsi in uno qualsiasi dei due stati possibili, indipendentemente dal bit memorizzato. I MOS T_1 e T_2 delle porte di trasmissione vengono quindi scelti con un rapporto W/L tale da costituire un partitore di tensione con gli NMOS degli invertitori del bistabile, in modo da portare nel transitorio di apertura della porta la tensione, all'uscita V_{OL} che è connessa con la gate del transistore NMOS interdetto, ad un valore inferiore al valore della tensione di soglia V_T.

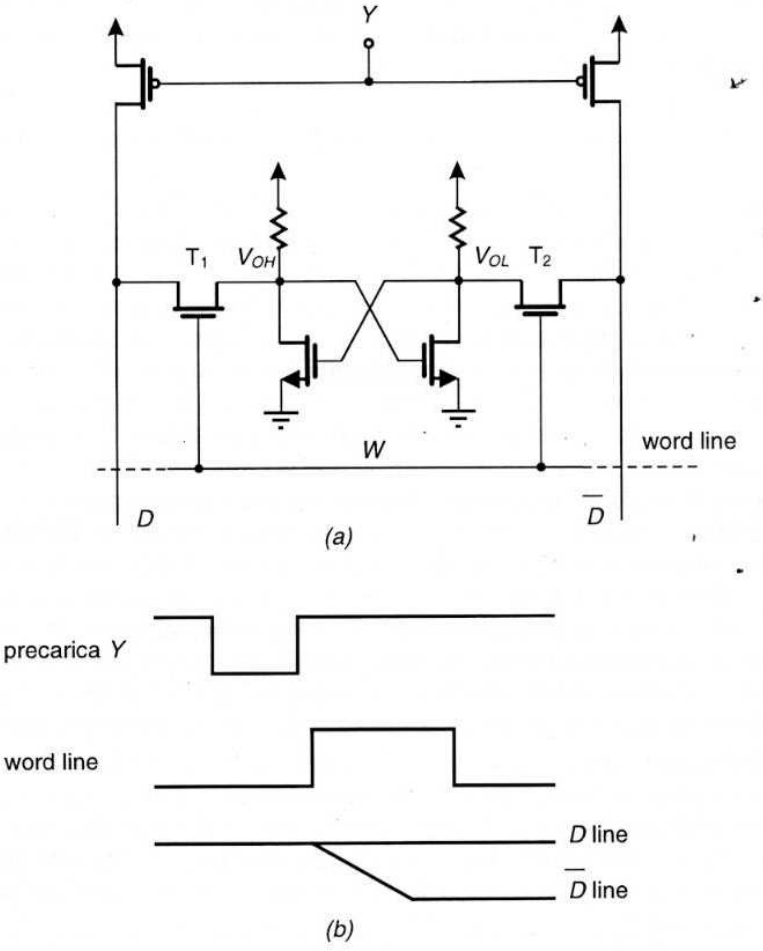


Figura 13.18 a) Cella NMOS con linee a rete di precarica; b) temporizzazione dei segnali

In Figura 13.19 sono riportati gli andamenti delle tensioni delle linee dati e degli ingressi del bistabile in seguito alla lettura di un 1 logico, ottenuti con simula-

zioni SPICE per la cella di Figura 13.18, con i transistori NMOS del bistabile dimensionati ad area minima (W/L = 2/1), e i MOS delle porte di trasmissione con W/L = 1/2; i transistori PMOS utilizzati per la precarica delle linee hanno invece un elevato W/L per caricare rapidamente le linee; la loro dimensione non è critica per le dimensioni complessive del chip in quanto vi sono solo due di questi transistori per colonna.

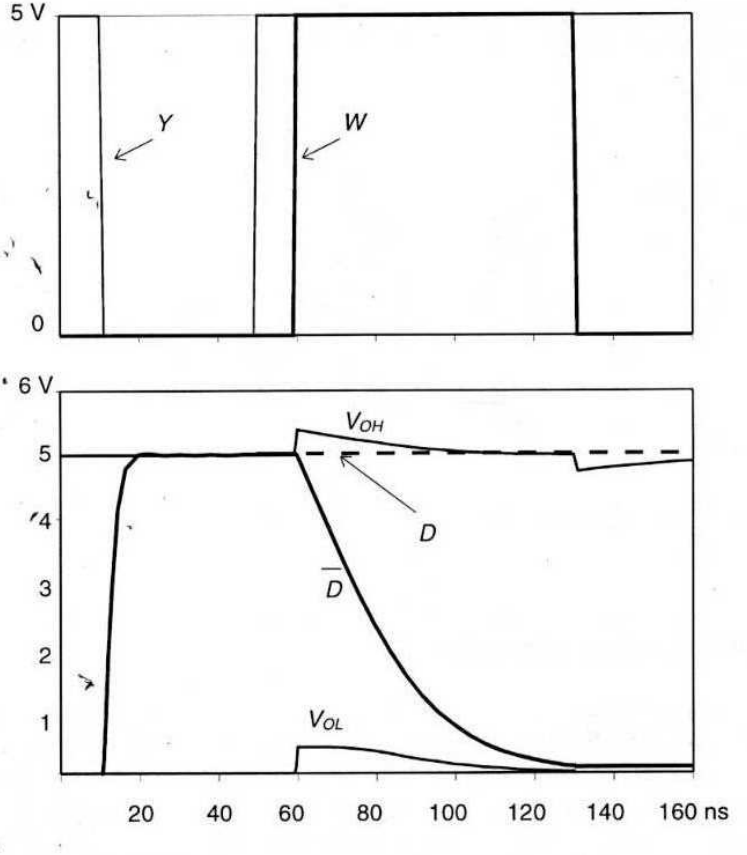


Figura 13.19 Simulazione SPICE della fase di lettura per la cella NMOS di Figura 13.18a con porte di trasmissione dimensionate con W/L = 1/2, e NMOS degli invertitori con W/L = 2/1: a) segnali Y di precarica e W di lettura; b) tensioni sulle linee dati e sui due drain del bistabile

La soluzione con cella NMOS e resistenze di carico a polisilicio richiede valori della resistenza R_{POLY} ben più elevate di quelle considerate nell'esempio precedente, se è richiesta una elevata capacità di memoria, e questo per ridurre a valori accettabili la dissipazione di potenza per cella; per capacità dell'ordine del Mbit si raggiungono valori dell'ordine del GΩ. Questi valori di resistenza, anche se sono

ottenibili con occupazione di area ancora contenuta utilizzando polisilicio con resistenza per quadro di $10^8\ \Omega/\square$, penalizzano in ogni caso il tempo di settaggio del bistabile, che è rallentato dalla elevata costante di tempo $R_{POLY}C_G$ per ognuno degli invertitori. Una soluzione migliore, che è ampiamente impiegata nelle memorie SRAM, è quella di realizzare il bistabile con invertitori CMOS, secondo lo schema riportato in Figura 13.20; questa cella elementare è detta a *6 transistori* perché impiega nella sua realizzazione 6 MOS. In queste celle il problema della dissipazione di potenza, che richiedeva una riduzione drastica della corrente assorbita, non si pone perché gli invertitori CMOS non dissipano potenza statica; per ridurre le dimensioni della cella si può scegliere un rapporto *W/L* minimo sia per gli NMOS che per i CMOS.

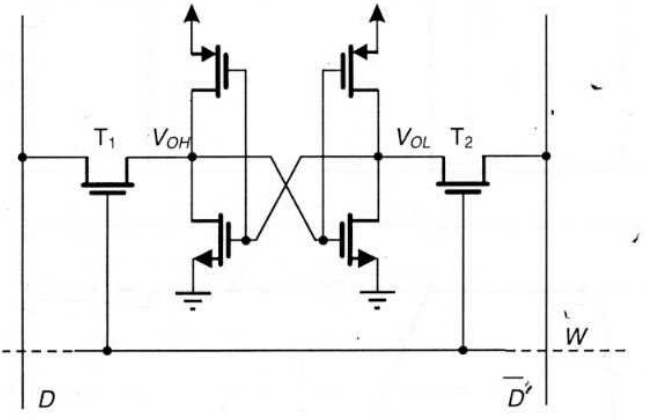


Figura 13.20 Cella di memoria statica CMOS

Nella versione CMOS sia il tempo di settaggio del bistabile che i tempi di scarica delle capacità di linea sono ridotti in quanto i PMOS di carico possono erogare una corrente relativamente elevata, e quindi mantenere il bistabile nello stato memorizzato anche durante l'operazione di lettura; i transistori T_1 e T_2 possono essere realizzati anch'essi ad area minima ($W/L = 1$), in quanto la dinamica del bistabile è più robusta che nel caso precedente, come si può vedere dall'andamento delle tensioni della simulazione SPICE della cella CMOS riportata in Figura 13.21. Da questa simulazione si vede come il tempo di lettura per il caso della cella CMOS si riduca significativamente rispetto al caso della cella NMOS. In questo caso ad esempio si può ridurre la resistenza dei MOS T_1 e T_2 in quanto si può tollerare un maggiore aumento della tensione V_{OL} nel transitorio rispetto al caso della cella con NMOS; infatti nel primo caso è necessario che questa sollecitazione su V_{OL} non raggiunga il valore di soglia V_T dei NMOS, mentre nel secondo basta che la sollecitazione su V_{OL} sia inferiore al valore di soglia dell'invertitore CMOS, che è circa $V_{DD}/2$. La riduzione della

resistenza dei transistori T_1 e T_2, d'altra parte, riduce il tempo di scarica della capacità di linea, e quindi riduce il tempo di lettura.

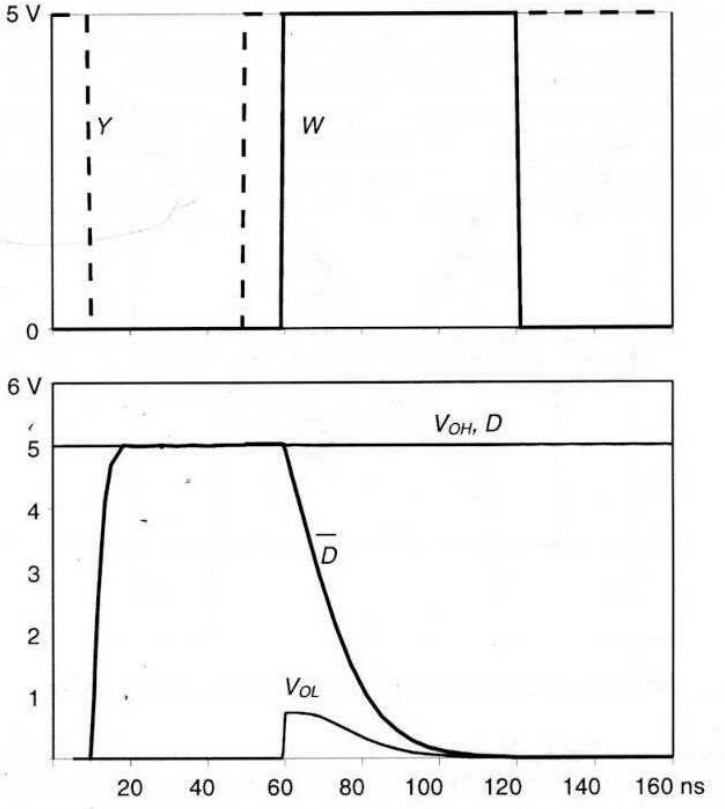


Figura 13.21 Simulazione SPICE della fase di lettura per la cella CMOS di Figura 13.20 con rete di precarica delle linee; $W/L|_{T1,2}$ = 2/2 μm, $W/L|_{N1,2}$ = $W/L|_{P1,2}$ = 2/1 μm.

Con le celle CMOS è possibile effettuare l'operazione di lettura senza ricorrere a reti di precarica delle linee; in tal caso le linee dati sono caricate dai due MOS di carico, operanti come carico attivo; in Figura 13.22 è riportato uno schema che utilizza transistori PMOS come carico. In tal caso l'operazione di lettura richiede un tempo minore, perché non bisogna attendere il tempo necessario alla precarica delle linee, ma si rende più critica la condizione di non alterare l'informazione contenuta nella cella nella fase di lettura. In questo caso infatti, assumendo, come in Figura 13.22, che l'invertitore 2 del bistabile presenti uscita bassa V_{OL}, occorre che i rapporti W/L tra i MOS P_L di carico, T_2 della porta e N_2 dell'invertitore siano scelti in modo tale che la tensione V'_{OL}, che si stabilisce all'ingresso basso

dell'invertitore quando la porta di apre, non superi il valore di soglia del bistabile, e nello stesso momento che la tensione V' sulla linea $\overline{D}$ scenda abbastanza sotto V_{DD} da poter essere rilevata dal circuito di lettura.

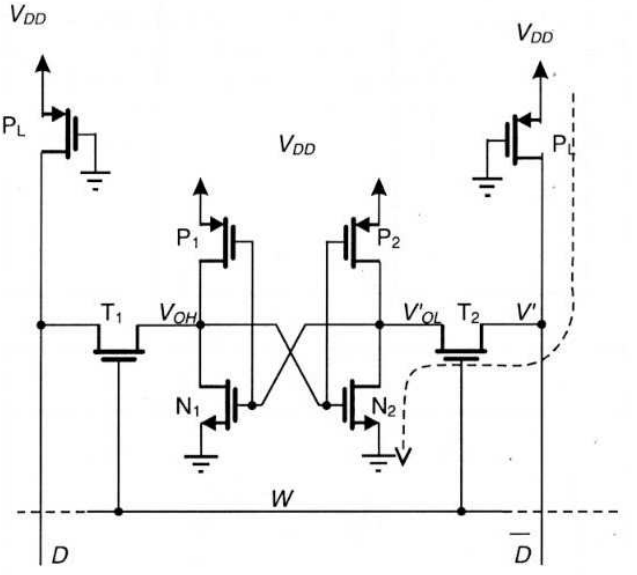


Figura 13.22 Cella di memoria CMOS con transistori di carico delle linee dati

Assumendo che P_L, T_2, N_2 lavorino nella regione lineare delle caratteristiche *I-V* (regione ohmica), e ricordando che in tale regione le resistenze dei MOS sono proporzionali all'inverso dei fattori *K* e quindi dei rapporti *W/L*, si possono scrivere le seguenti relazioni:

$$V' = V_{DD}\frac{R_{N2}+R_{T2}}{R_{N2}+R_{T2}+R_{PL}} = V_{DD}\frac{\left.\frac{L}{W}\right|_{N2}+\left.\frac{L}{W}\right|_{T2}}{\left.\frac{L}{W}\right|_{N2}+\left.\frac{L}{W}\right|_{T2}+\left.\frac{2.5L}{W}\right|_{PL}} < V_{DD}-\Delta V \quad (10.17)$$

$$V'_{OL} = V_{DD}\frac{R_{N2}}{R_{N2}+R_{T2}+R_{PL}} = V_{DD}\frac{\left.\frac{L}{W}\right|_{N2}}{\left.\frac{L}{W}\right|_{N2}+\left.\frac{L}{W}\right|_{T2}+\left.\frac{2.5L}{W}\right|_{PL}} < V_{TN} \quad (13.18)$$

che permettono di dimensionare in prima approssimazione i valori di *W/L* dei diversi MOS. In particolare la (13.18) garantisce che il MOS N_2 non sia in conduzione per tutta la fase di lettura, il che aumenterebbe la dissipazione di potenza della cella e renderebbe poco conveniente l'impiego della tecnologia CMOS; la condizione imposta dalla (13.17), tenendo conto del vincolo imposto dalla (13.18), comporta che il fattore di forma *W/L* di P_L non sia troppo elevato, ma sia paragonabile a quello di N_2, e che quello di T_2 sia significativamente minore di entrambi. Ciò porta ad una inevitabile riduzione del valore della variazione di tensione ΔV rilevabile per l'operazione di lettura, rispetto al valore ideale di V_{DD}; in tal caso occorre utilizzare circuiti di lettura che amplifichino la variazione di tensione disponibile tra le due linee dati. Ritorneremo su questo aspetto nel Paragrafo 13.7 parlando dei circuiti di lettura.

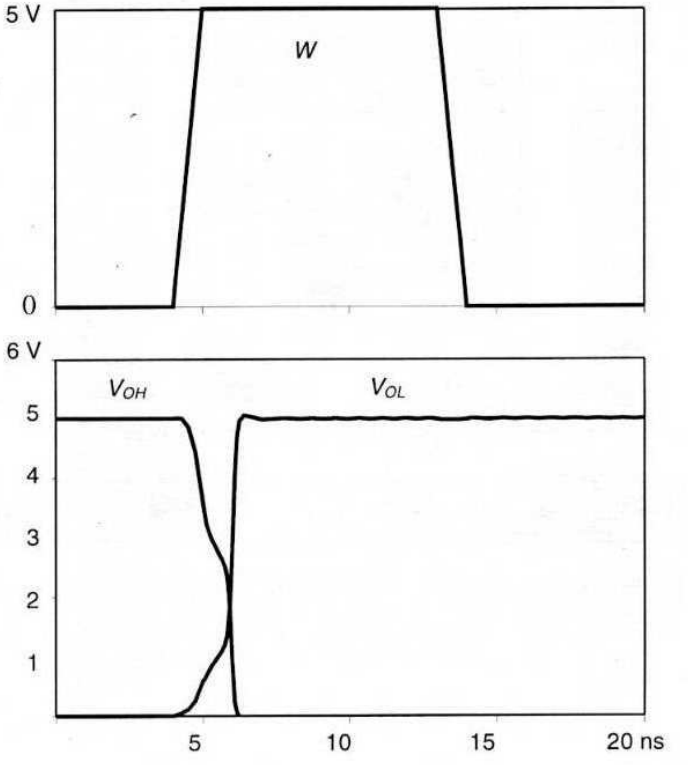


Figura 13.23 Simulazione SPICE della scrittura del bit 0 nella cella di memoria CMOS di Figura 13.20 che memorizzava un bit 1

L'operazione di scrittura, per tutte le celle esaminate, non presenta particolari problemi, in quanto è facilitata dalla possibilità di inviare sia il bit D che quello negato $\overline{D}$ ai due ingressi del bistabile attraverso le due linee dati. L'operazione è in

principio la stessa sia per linee a precarica che per quelle con carico attivo: le linee vengono pilotate dalle uscite del circuito di scrittura (vedi Paragrafo 13.7) in modo tale che, per scrivere un 1 nella cella, la linea D si porta alta e quella $\overline{D}$ bassa; poiché i MOS che pilotano le linee sono ad elevato rapporto W/L, questi valori della tensione vengono forzati nella cella di memoria che si porta nello stato stabile corrispondente. In Figura 13.23 si riporta il risultato di una simulazione SPICE per l'operazione di scrittura di uno 0 logico nella cella di memoria CMOS in cui era stato precedentemente memorizzato un 1.

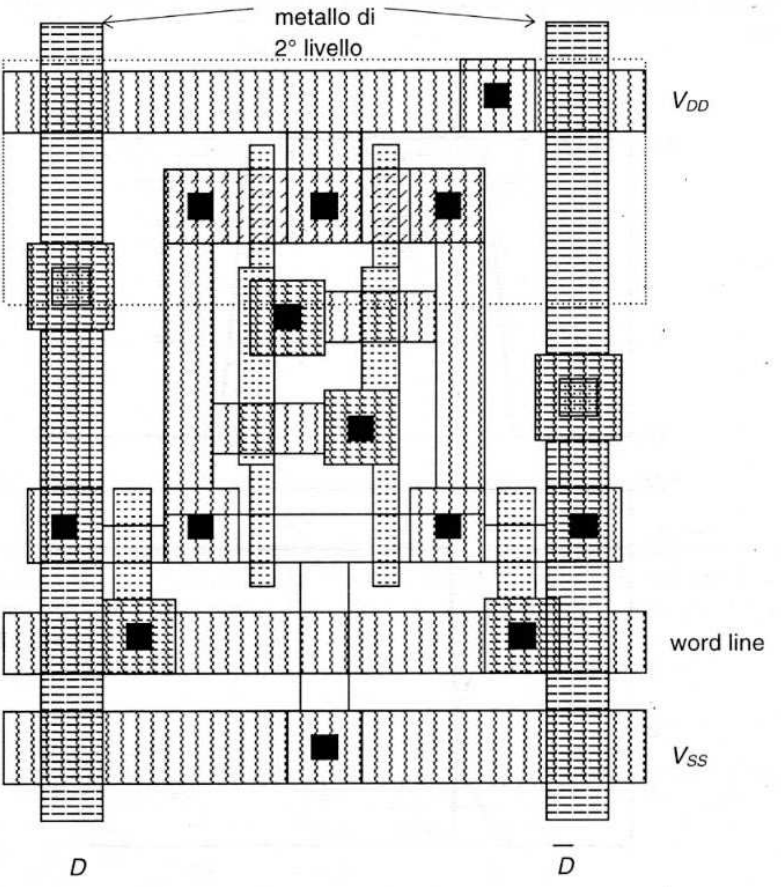


Figura 13.24 Tracciato di una cella di memoria CMOS a 6 transistori; $W/L|_{T1,2} = 3\lambda/3\lambda$, $W/L|_{N1,2}\ 4\lambda/2\lambda = W/L|_{P1,2} = 6\lambda/2\lambda$

Si è supposto che il circuito di scrittura ha portato la linea D alla tensione V_{DD} e quella $\overline{D}$ alla tensione 0. Quando il comando W della word line passa alto si aprono le porte T_1 e T_2 e le tensioni alle due uscite corrispondenti del bistabile comin-

ciano a variare portandosi verso la soglia logica; appena superata (in entrambi i sensi) la soglia logica si instaura l'azione degenerativa del bistabile che si porta rapidamente nell'altro strato stabile, memorizzando il bit 1, valore che viene mantenuto dalla cella anche quando il comando W è tornato basso. La fase di scrittura avviene quindi in un tempo molto minore di quello di lettura.

Un esempio di tracciato per una cella di memoria CMOS che utilizza un processo tecnologico a due livelli di metallo è mostrato nella Figura 13.24. In questo tracciato, le linee di alimentazione e di massa delle celle, come anche le contattazioni dei source e drain dei transistori sono realizzate con il primo livello di metallizzazione, mentre le linee dati con il secondo livello di metallizzazione, in modo da poter intersecarsi; è previsto un contatto tra il primo e il secondo livello.

13.5.2 Celle in tecnologia bipolare

Le memorie RAM bipolari hanno capacità di memoria più limitata di quelle in tecnologia CMOS a causa delle limitazioni imposte dalla dissipazione di potenza che non è trascurabile; esse presentano tuttavia un tempo di accesso più basso a causa della maggiore velocità delle strutture bipolari e della maggior capacità di pilotare carichi capacitivi elevati.

Per ridurre la dissipazione di potenza, nelle celle bipolari si alimenta la cella ad una tensione ridotta nelle fasi di attesa, e si porta l'alimentazione ai valori nominali solo durante le operazioni di lettura o scrittura, in modo da ridurre la dissipazione media (che è essenzialmente quella dello stato di riposo).

Una versione di cella bipolare ad accoppiamento di emettitore è quella riportata in Figura 13.25a, ed utilizza transistori a doppio emettitore per i due invertitori del bistabile; i due emettitori svolgono la funzione di porta di accesso alle linee dati per le operazioni di lettura o scrittura. La riga (word line) è sdoppiata in due linee, indicate come W e W^* per permettere la doppia alimentazione della cella in condizioni di quiescenza o di lettura/scrittura; anche per queste celle di memoria ogni colonna è sdoppiata in due linee dati D e $\overline{D}$.

Il funzionamento di questo circuito può essere spiegato con riferimento al circuito di Figura 13.25 e alle temporizzazioni dei segnali riportate. Le due linee di riga sono portate nella situazione di quiescenza ai valori rispettivamente di 1.3 V (W^*) e 0.3 V (W), per cui la cella è alimentata a 1 V. Si suppone che sia memorizzato un 1 nella cella se il transistore Q_2 è interdetto; per la connessione in anello dei due invertitori il transistore Q_1 conduce e la corrente passa nell'emettitore connesso alla linea W. La tensione sulla base di Q_1 vale quindi $0.3 + 0.7 \cong 1$ V; le due linee dati sono connesse ad un'alimentazione comune di 1.5 V attraverso le resistenze R_3 e R_4, e quindi l'emettitore E_2 è interdetto in quanto polarizzato ad una tensione superiore a quella della base di Q_1. Per abilitare la cella all'operazione di lettura o scrittura le linee W e W^* si portano ad un valore alto, rispettivamente di 4.3 V (W^*) e 2 V (W); quest'ultima tensione è scelta in modo tale che l'emettitore E_1 connesso a W sia interdetto, per cui la corrente del transistore che conduce non può che circolare attraverso l'emettitore E_2.

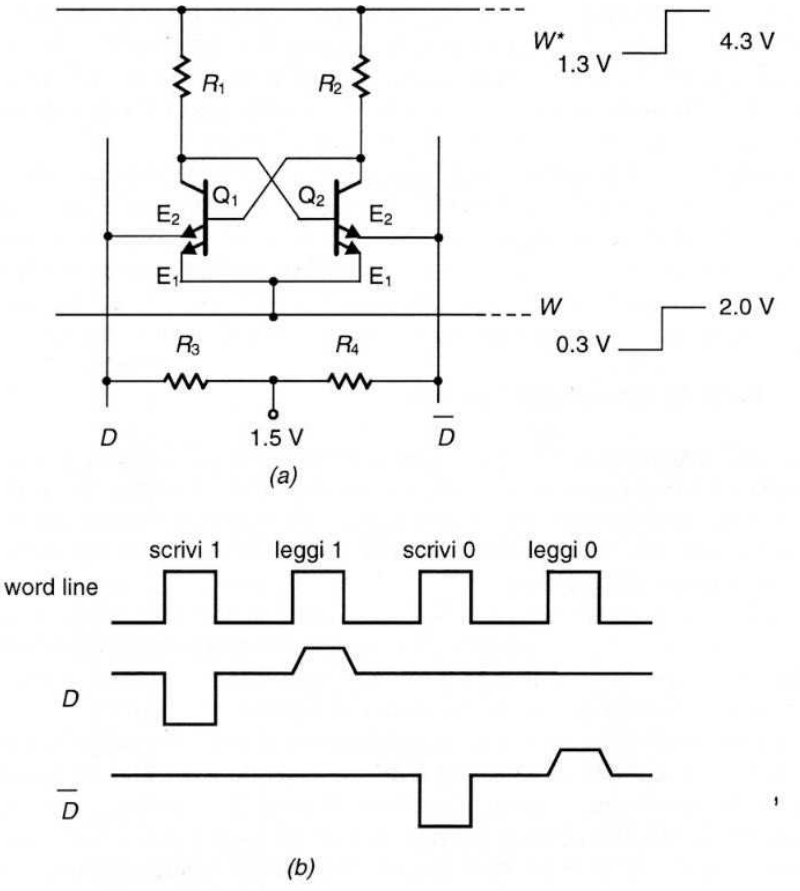


Figura 13.25 a) Cella di memoria bipolare ad accoppiamento di emettitore; b) diagrammi di temporizzazione dei segnali

Nella fase di lettura, sempre supponendo Q_1 in conduzione (1 memorizzato), questa corrente passa attraverso R_3 e si chiude attraverso la tensione di 1.5 V; ciò provoca un aumento della tensione della linea D rispetto a 1.5 V, mentre la linea $\overline{D}$ rimane a questa tensione, e questa differenza viene rilevata dal circuito di lettura. Nella fase di scrittura la tensione della linea D (per scrivere 1) o $\overline{D}$ (per scrivere 0) viene portata a zero, e questo forza il bistabile ad assumere lo stato imposto da questa condizione, in quanto viene posto in conduzione rispettivamente Q_1 o Q_2 tramite la conduzione dell'emettitore E_2 corrispondente.

La fase di lettura è quella più lenta in quanto la corrente di uscita dagli emettitori E_2 deve caricare la capacità della linea dati, che come si è visto è relativamente elevata; in ogni caso la lettura è più rapida che nel caso delle celle in tecnologia MOS, per la maggior capacità di erogazione di corrente dei transistori bipolari, e in

particolare in questo caso in quanto l'uscita sull'emettitore è a bassa resistenza interna.

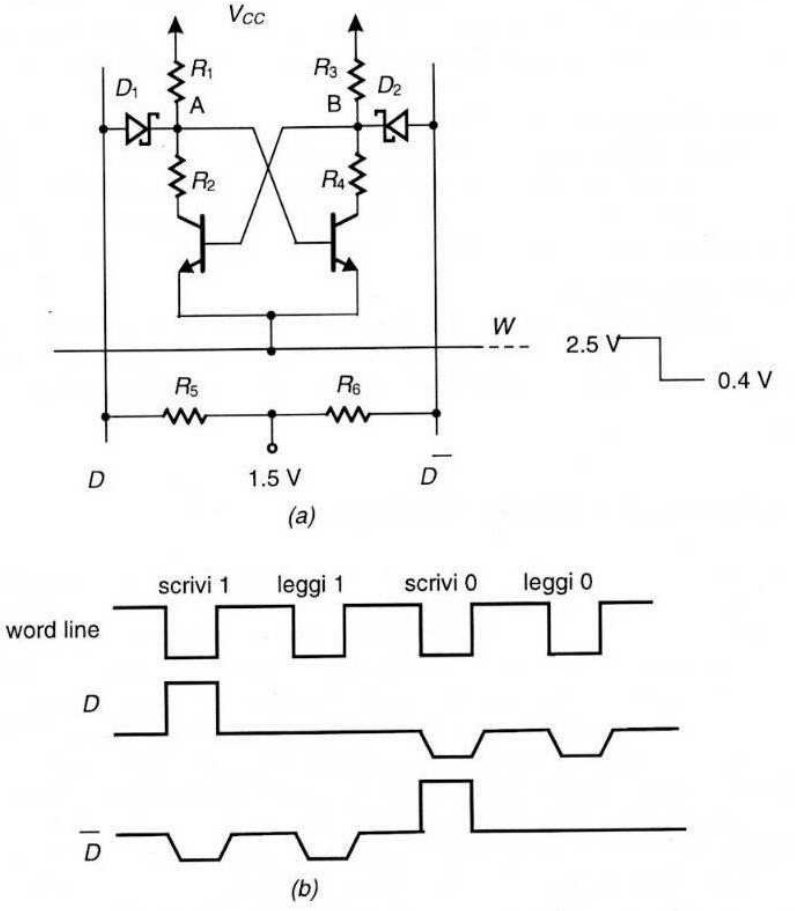


Figura 13.26 a) Cella di memoria bipolare con accoppiamento a diodi; b) temporizzazioni dei segnali.

Una seconda versione di cella bipolare utilizza un accoppiamento a diodi (in particolare diodi Schottky) per la connessione con le linee dati, utilizzando una configurazione di tipo logica RTL per i due invertitori del bistabile, secondo lo schema di Figura 13.26a. Questa cella permette di utilizzare una sola linea per riga anziché le due della cella precedente, e quindi tale configurazione è preferita perché permette una minore occupazione di spazio. Anche in questa cella le due linee dati D e $\overline{D}$ sono collegate ad un'alimentazione di 1.5 V tramite due resistenze R_5 e R_6; la tensione di alimentazione della cella è di 3.5 V. In stato di quiescenza, la tensio-

ne della riga W viene mantenuta al valore di 2.5 V; poiché questa è anche la tensione degli emettitori dei transistori Q_1 e Q_2 del bistabile, la cella è alimentata ancora a 1 V, e dissipa una potenza contenuta. Per mantenere la stessa convenzione sul dato in fase di lettura sulle linee dati, si assuma che la memorizzazione di un 1 corrisponda alla conduzione di Q_2; in assenza di comando di abilitazione i diodi D_1 e D_2 sono entrambi interdetti in quanto le tensioni dei punti A e B sono certamente superiori a 1.5 V.

In fase di abilitazione della cella la tensione della riga W si porta ad un valore più basso, 0.4 V. Supponendo che sia immagazzinato un 1 (conduzione di Q_2), la lettura avviene attraverso il diodo D_2 che va in conduzione, in quanto la tensione ai capi del diodo vale: $1.5 - (0.4 + 0.2) > V_\gamma$; la corrente di conduzione passerà quindi attraverso la resistenza R_6 e provocherà una diminuzione della tensione della linea $\overline{D}$, che potrà essere rilevata.

Per la scrittura di un 1, la linea D viene portata ad un valore alto e questa tensione, attraverso D_1, agisce sulla base di Q_2 portandolo in conduzione (l'inverso avviene se si vuole scrivere uno zero, come si vede dai diagrammi di Figura 13.26b).

13.6 Circuiti di lettura e scrittura

I circuiti di lettura e scrittura delle linee dati sono essenzialmente degli amplificatori abilitati dalle uscite del decodificatore di colonna, che nella fase di scrittura pilotano in uscita le tensioni di linea in funzione dei dati inviati ai loro ingressi, e per l'operazione di lettura amplificano lo sbilanciamento di tensioni sulle due linee dati e lo trasferiscono come segnale logico in uscita. Faremo riferimento per una descrizione sommaria del loro funzionamento ad una versione semplificata della loro realizzazione, per il caso di memorie a tecnologia MOS, ricordando che i circuiti effettivi sono in genere molto più sofisticati in quanto è alla loro prestazione che è affidata la capacità di buon funzionamento della memoria.

In Figura 13.27 è indicata la configurazione elementare degli amplificatori di lettura e scrittura inseriti al termine di ogni colonna della matrice di memoria, in accordo all'organizzazione generale della RAM indicata in Figura 13.15.

Sia per le operazioni di scrittura che per quelle di lettura, la colonna in cui vi è la cella di memoria a cui si deve accedere viene abilitata dal decodificatore di colonna che porta allo stato alto il segnale corrispondente alla colonna data; questo segnale viene applicato all'ingresso di due porte di trasmissione (in figura sono riportate porte NMOS ma possono essere anche porte PMOS o CMOS) che una volta abilitate pongono in connessione le due linee dati sia con gli amplificatori di scrittura che con quello di lettura. Lo schema di principio della Figura 13.27 si applica sia a RAM con linee dati precaricate che a linee con carico attivo. Se si deve scrivere l'informazione nella cella, il dato D da memorizzare viene inviato a stadi invertitori che forniscono rispettivamente il dato D e quello $\overline{D}$ alle due linee dati. I valori W/L degli stadi di uscita di questi invertitori sono scelti relativamente elevati in modo da fornire sufficiente corrente di uscita per il pilotaggio delle rela-

tivamente elevate capacità delle linee, e per accelerare l'operazione di scrittura nella cella di memoria.

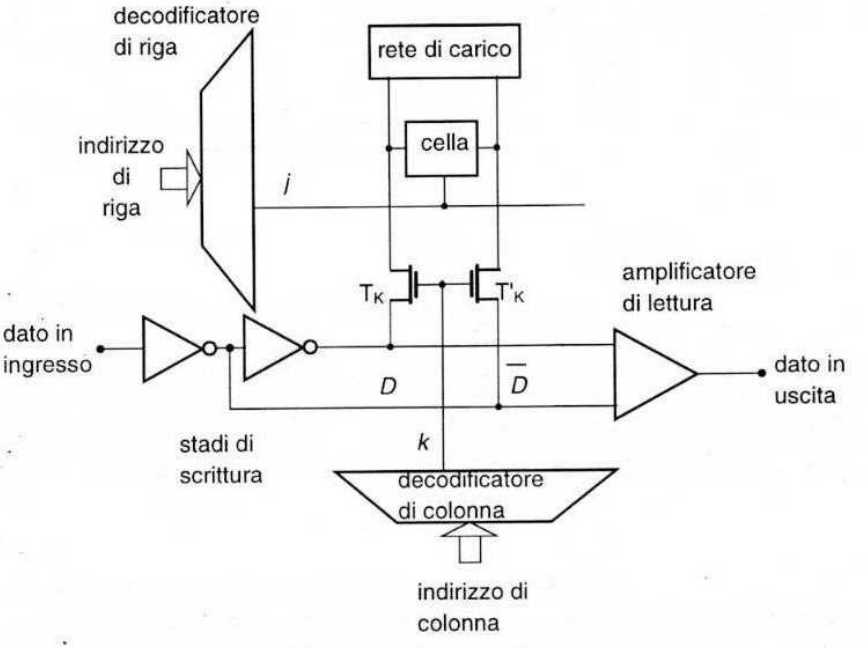


Figura 13.27 Amplificatori di lettura e scrittura in tecnologia MOS

In fase di lettura questi invertitori debbono essere disabilitati, per cui si utilizzano invertitori tri-state abilitati da un segnale di controllo.

Gli amplificatori di lettura (detti "*sense amplifier*") sono l'elemento più critico per il funzionamento delle memorie, e richiedono una progettazione accurata, in quanto funzionano con un segnale in ingresso di tipo "analogico", legato alla differenza di tensione tra le due linee dati, da cui debbono poter estrarre l'informazione dello stato logico contenuta nella cella letta. Questa differenza di tensione può essere relativamente bassa nel caso di celle di memoria connesse a linee caricate, come ad esempio nel caso di Figura 13.22, per cui occorre un circuito capace di riportare questo segnale di valore ridotto ai livelli logici nominali richiesti dai circuiti digitali. Anche nel caso di celle con linee precaricate, come nei casi delle Figure 13.18 e 13.20 rispettivamente per celle NMOS e CMOS, l'utilizzo di un amplificatore di lettura capace di rilevare piccole differenze tra le tensioni delle due linee dati permette di velocizzare l'operazione di lettura, in quanto non occorre attendere la scarica completa della capacità della linea che deve portarsi al livello logico V_{OL}, ma basta attendere il tempo ΔT nel quale la tensione della capacità si è ridotta della quantità misurabile ΔV, come è mostrato in Figura 13.28a.

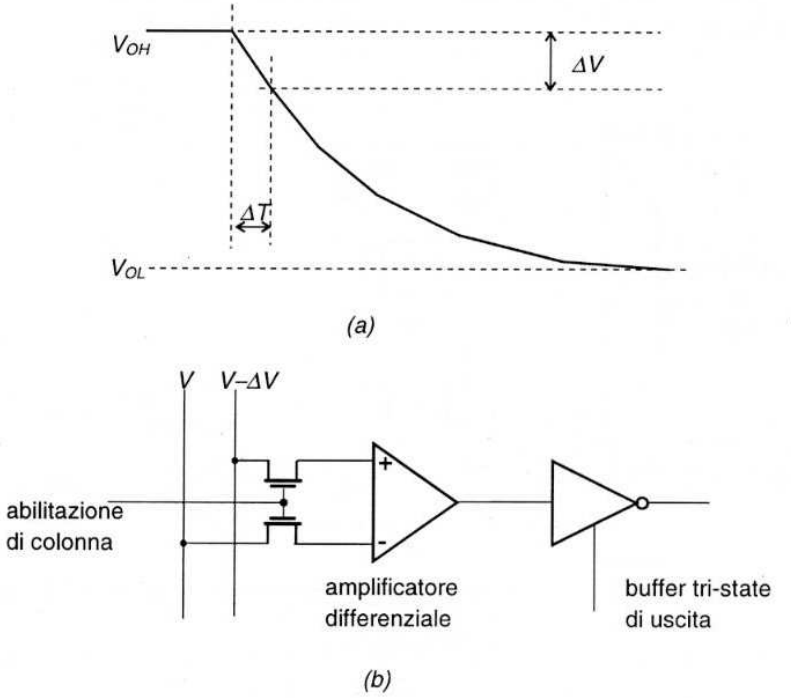


Figura 13.28 a) Lettura di una variazione ridotta della tensione sulla linea connessa al valore V_{OL}; b) schema di principo di un amplificatore di lettura

Per poter rilevare una piccola differenza di tensione tra le due linee dati, la configurazione più adatta è quella differenziale, indicata in Figura 13.28b; a valle dell'amplificatore il segnale di uscita viene applicato ad un buffer tri-state che sarà collegato al bus comune di uscita da cui vengono raccolti i dati di lettura delle celle di memoria. L'amplificatore di lettura è usualmente costituito da più stadi in cascata, e spesso per aumentare la sensibilità di lettura si utilizza una configurazione bistabile in modo da utilizzare l'instabilità che si instaura nel bistabile a seguito di uno sbilanciamento degli ingressi, per fornire in un tempo breve due segnali logici complementari e di livello voluto alle due uscite.

Un possibile schema, ancora di principio, per un amplificatore di lettura utilizzabile in memorie SRAM in tecnologia CMOS è quello riportato in Figura 13.29. L'amplificatore è costituito da tre stadi in cascata: il primo è costituito da una configurazione differenziale, con carichi attivi realizzati con PMOS, il secondo, pilotato dalle due uscite in opposizione di fase del primo, è un bistabile nel quale il carico di ognuno dei due rami, ancora costituito da PMOS, è modificato in funzione dei segnali di uscita del primo stadio, e il terzo da un invertitore CMOS (per ognuna delle due uscite del bistabile).

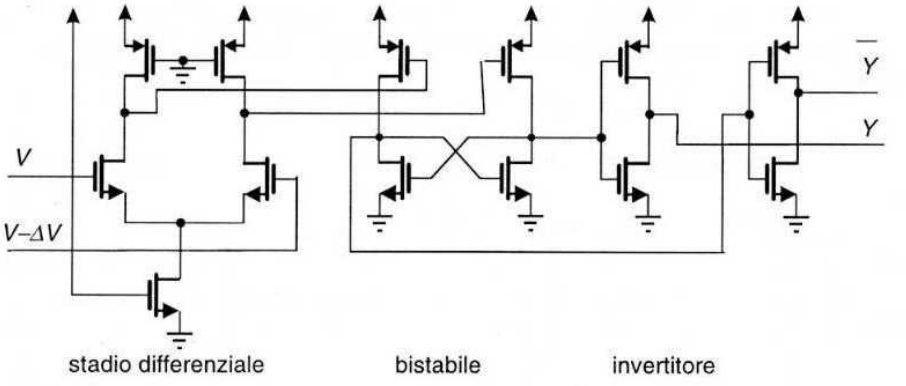


Figura 13.29 Schema di un amplificatore di lettura a tre stadi

La debole variazione tra le due tensioni di linea viene amplificata dal primo stadio e ingenera una commutazione del bistabile. Le due uscite sono portate ai livelli logici voluti (V_{DD} e 0) dall'invertitore CMOS, che ripristina anche i margini di rumore. Ritorneremo sugli amplificatori di lettura trattando delle memorie dinamiche.

13.7 Organizzazione delle memorie RAM

L'organizzazione delle RAM con un'unica matrice di celle presentata in Figura 13.14 non è più conveniente per le memorie ad elevata capacità (superiore a 128 kbit), in quanto l'aumento del numero di celle costringerebbe a realizzare linee e colonne di lunghezza maggiore, con aumento corrispondente delle capacità di linea C_L e con conseguente inevitabile peggioramento della dinamica, in particolare nell'operazione di lettura che, come si è visto nei paragrafi precedenti, è fortemente condizionata da questo valore.

Per memorie di elevata capacità si ricorre ad un'architettura a blocchi, in cui lo schema a matrice della RAM presentato in Figura 13.14 è ripetuto per m volte, dando luogo ad un sistema che opera in parallelo sugli m blocchi, mediante una suddivisione dell'indirizzo di accesso di n bit, che viene ora diviso in n_1 bit per l'indirizzo di riga, n_2 bit per l'indirizzo di colonna e n_3 bit per l'indirizzo di blocco. I dati in ingresso e in uscita vengono quindi convogliati su un unico bus che raccoglie i dati dei diversi blocchi, connessi a questo mediante stadi buffer tri-state. Lo schema di principio di questa organizzazione della memoria è rappresentato nella Figura 13.30, dove sono stati indicati i blocchi (tipicamente con capacità di memoria di 128 kbit) in cui è partizionata la memoria globale. Questa architettura permette anche di effettuare una migliore gestione della potenza utilizzata, mediante una riduzione della potenza di alimentazione dei blocchi che non vengono indirizzati. La riduzione della potenza viene anche realizzata riducendo il valore della tensione di alimentazione interna della RAM, rispetto a quella dei circuiti di ingres-

so/uscita, che debbono mantenere i livelli logici compatibili con il resto del sistema; tipicamente si adotta una tensione di 3.3 V per la tensione V_{DD} della matrice della memoria.

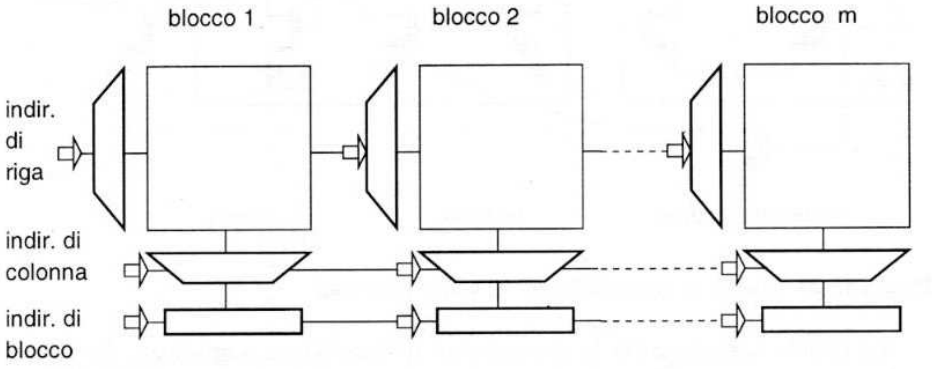


Figura 13.30 Organizzazione a blocchi delle memorie RAM ad elevata capacità

Nelle SRAM per cui è richiesta una velocità elevata di lettura e scrittura, può essere utilmente combinata la tecnologia BiCMOS, presentata nel Paragrafo 10.7, con quella CMOS, utilizzando sia la migliore capacità di pilotaggio di elevate capacità della tecnologia BiCMOS che la maggiore compattezza di area e l'elevata integrazione della tecnologia CMOS. Nella struttura delle SRAM vista precedentemente, la tecnologia CMOS viene ben impiegata per la realizzazione del nucleo (*core*) della memoria, ossia la matrice delle celle elementari di memoria, mentre i circuiti di decodifica, che debbono pilotare le linee di riga e di colonna con elevate capacità vengono realizzati con porte NOR (o NAND) in tecnologia BiCMOS. Anche in tecnologia BiCMOS vengono realizzati i circuiti di controllo e gli amplificatori di lettura; questi ultimi possono impiegare transistori bipolari per lo stadio differenziale in ingresso, con migliori prestazioni per quanto riguarda la velocità di risposta e l'amplificazione, rispetto alla versione con dispositivi MOS.

13.8 Memorie RAM dinamiche (DRAM)

Per aumentare la capacità di memoria disponibile a parità di area del chip occorre in ogni caso:

- ridurre il numero di transistori per cella elementare di memoria;
- ridurre il numero di interconnessioni, in particolare ridurre le dimensioni (e il numero) delle linee per l'alimentazione delle celle, e quelle per la lettura e la scrittura dei dati.

Una possibilità significativa per agire sia sul primo che sul secondo dei punti elencati, viene dall'impiego dei concetti della *logica dinamica*, già presentati per la

realizzazione di celle di memoria basate sulla conservazione della carica accumulata in una capacità, viste nel Paragrafo 12.10.
Le memorie RAM che utilizzano celle di memoria dinamiche vengono indicate come DRAM (*Dynamic Random Access Memory*), e permettono di ottenere, come vedremo, le più elevate capacità di memoria a parità di area utilizzata. Per tali memorie la tecnologia è obbligatoriamente quella MOS; con l'uso delle celle di memoria dinamiche si riducono i dispositivi necessari per la singola cella, e quindi l'occupazione di area a parità di bit incamerati (e quindi di celle della matrice).

La prima cella dinamica utilizzata nelle memorie DRAM prevede per la sua realizzazione 4 transistori MOS e viene quindi indicata come *cella a 4 transistori*. La cella è basata su due latch dinamici elementari (come quelli indicati sinteticamente in Figura 12.32), e la sua realizzazione circuitale è riportata in Figura 13.31.

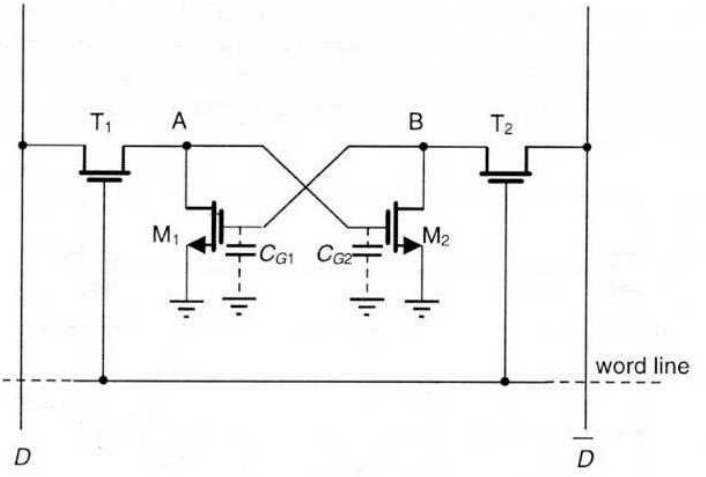


Figura 13.31 Cella di memoria dinamica a 4 transistori

La cella utilizza le due capacità di gate rispettivamente C_{G1} del MOS M_1 e C_{G2} del MOS M_2 per l'immagazzinamento dell'informazione (in effetti vengono immagazzinati sia il bit D che il suo negato $\overline{D}$ per una ridondanza di informazione). A seguito del collegamento in anello dei due MOS tra drain e gate, se C_{G1} è ad esempio nello stato alto ($\sim V_{DD} - V_T$), M_1 conduce e la capacità C_{G2} connessa al drain di M_1 si scarica portandosi allo stato basso (~ 0); a sua volta M_2 sarà interdetto e quindi si ritrova che C_{G1} rimane carica (ovviamente trascurando le correnti di perdita delle giunzioni drain/substrato di M_1 e source/substrato di T_1). I due MOS T_1 e T_2 agiscono come porte di trasmissione per la lettura e/o scrittura dell'informazione nelle capacità, e vengono abilitati (aperti) dall'abilitazione della riga (word line)

corrispondente, analogamente al caso delle celle MOS statiche. Le linee sono del tipo "a precarica", come già visto per le celle NMOS e CMOS, per effettuare l'operazione di lettura, che anche in questo caso è facilitata dalla precarica delle linee alla tensione alta. Infatti, supponendo, con riferimento al circuito di Figura 13.31, che il nodo A sia al livello alto (C_{G2} carica), quando si aprono le porte T_1 e T_2 questo valore di tensione non viene alterato, e quindi non viene modificata la carica immagazzinata sulla capacità C_{G2}, mentre la linea $\overline{D}$, che viene connessa alla capacità C_{G1} (caricata al valore basso V_{OL}) non altera lo stato di carica di C_{G2}, in quanto la capacità C_L della linea viene scaricata attraverso il MOS M_2 che è in conduzione.

Rispetto al caso delle celle statiche CMOS si risparmiano i due transistori PMOS per ogni cella; rispetto a quelle NMOS con polisilicio, si risparmia lo spazio della resistenza in polisilicio; ma si risparmia in ogni caso (e questo è ben più importante) la linea di alimentazione V_{DD} per la cella, in quanto questa versione non prevede alimentazione diretta, ma solo una carica delle capacità attraverso le porte di trasmissione.

La cella prevede, oltre alle operazioni di lettura e scrittura (che sono analoghe a quelle viste per la cella a 6 transistori), una fase di ripristino della carica immagazzinata nella capacità, detta "*refresh*", con frequenza di clock dell'ordine dei ms, fase necessaria, come si è visto, per tutti i circuiti a logica dinamica.

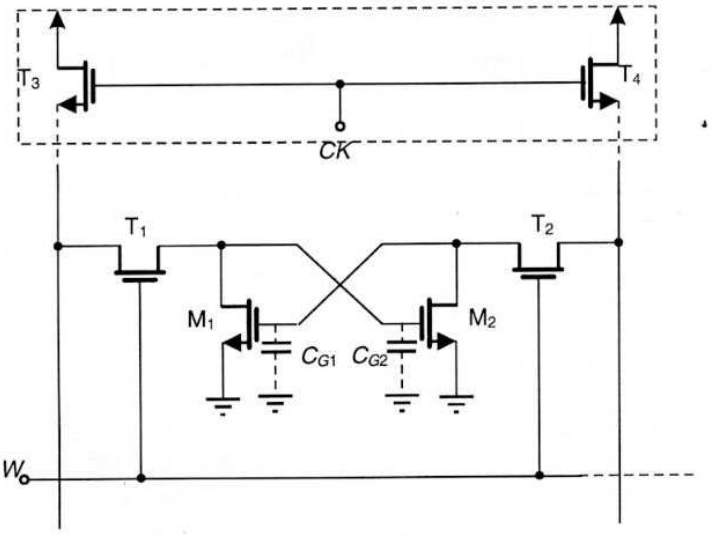


Figura 13.32 Rete di "*refresh*" (rettangolo tratteggiato) per la cella a 4 transistori

Le connessioni della cella alle linee D e $\overline{D}$ e alla rete di *refresh* (che è in comune a tutte le celle della colonna) sono indicate in Figura 13.32.

La rete di ripristino della carica è formata dai due MOS T_3 e T_4, che agiscono anch'essi come porte di trasmissione e vengono abilitati dal segnale di clock *CK*. L'operazione di ripristino richiede che siano contemporaneamente alti il clock *CK* ed il segnale *W* di abilitazione della riga, per cui in questa fase i transistori T_1 e T_3 corrispondono al carico di M_1 mentre T_2 e T_4 corrispondono al carico di M_2. Assumendo che la memorizzazione di un 1 corrisponda a C_{G1} caricato alla tensione alta e C_{G2} scarico, se la tensione ai capi di C_{G1} non è scesa fino al valore di V_T, M_1 è in conduzione e la tensione V_{D1} di drain sarà bassa, mantenendo C_{G2} scarico e M_2 interdetto; a sua volta l'interdizione di M_2 fa sì che la corrente del carico $T_2 + T_4$, ripristini la carica persa da C_{G1} fino a riportarlo al valore $V_{DD} - V_T$. L'operazione di "refresh" è effettuata per tutte le celle connesse alla stessa colonna in quanto la rete di refresh $T_3 \div T_4$ è comune a tutta la colonna, ed è effettuata sequenzialmente su tutte le colonne della matrice.

Il risparmio di transistori per memoria di questa soluzione per una DRAM da *n* bit, considerando che si risparmiano 2 MOS per cella rispetto alle SRAM è di $2n$, il che comporta un significativo risparmio di area.

13.8.1 Celle dinamiche a un transistore

La versione più ridotta di cella di memoria dinamica per DRAM è quella, detta *cella a un transistore*, che impiega solo un MOS e una capacità per bit di memoria, secondo lo schema di principio riportato in Figura 13.33.

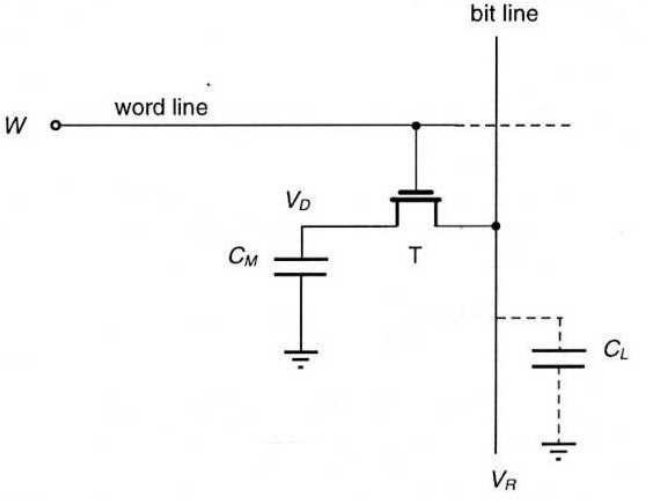


Figura 13.33 Cella dinamica ad un transistore

Questa cella è basata sulla considerazione che, per l'immagazzinamento dell'informazione in una capacità, la struttura minima necessaria è quella di una capacità connessa ad una porta di trasmissione, la quale, quando è aperta, connette la capacità al segnale in ingresso, e quando è chiusa permette di mantenere la carica stabilitasi sulla capacità anche se l'ingresso varia. In base a questa drastica semplificazione della cella di memoria, si può utilizzare una sola linea dati per colonna e ridurre ad un solo transistore la cella stessa. Questa soluzione comporta quindi un significativo risparmio di area, legato alle regole di progetto, per cui questa cella viene utilizzata in tutte le memorie RAM di elevata capacità. Naturalmente questa riduzione così drastica di componenti, congiunta all'eliminazione della ridondanza permessa dallo sdoppiamento delle linee dati per le operazioni di lettura e scrittura, comporta, come vedremo successivamente, una maggiore complessità di progettazione della memoria e in particolare dei circuiti di lettura.

L'operazione di scrittura della cella è molto semplice, in quanto, dopo aver selezionato la riga (word line), si applica il bit da memorizzare alla colonna (bit line) voluta, ed essendo aperta la porta di trasmissione T, il bit viene immagazzinato nella capacità di memoria C_M, che si porta ad una tensione V_D (V_{OH} o V_{OL}), in funzione del livello logico presente sulla colonna al momento della scrittura.

L'operazione di lettura è ancora in principio semplice, perché occorre abilitare la riga (e quindi la porta di trasmissione T) e porre in contatto la capacità C_M con la colonna, ossia la bit line, che assumiamo sia stata precaricata ad una tensione di riferimento V_R, per leggere la tensione presente ai capi di C_M. Questa operazione è però complicata dal fatto che l'operazione di lettura avviene ora attraverso una ridistribuzione di carica tra la capacità C_M e capacità di linea C_L nel momento in cui la porta T si apre, per cui la tensione V'_R ai capi della bit line dopo l'apertura di T sarà:

$$Q = C_M V_D + C_L V_R = (C_M + C_L) V'_R \Rightarrow V'_R = \frac{C_M}{(C_M + C_L)} V_D + \frac{C_L}{(C_M + C_L)} V_R \tag{13.19}$$

e la variazione di tensione da rilevare $\Delta V_R = V_R - V'_R$ sarà data da:

$$\Delta V_R = \frac{C_M}{C_M + C_L} (V_D - V_R) \tag{13.20}$$

Poiché $C_M << C_L$, la variazione di tensione ΔV_R sulla linea dati sarà in ogni caso molto bassa; ad esempio, assumendo un valore della capacità $C_M = 25$ fF, un valore di capacità di linea $C_L = 1$pF, un valore della tensione di riferimento $V_R = 5$ V, e un valore di tensione memorizzato $V_D = 0$ V, dalla (13.20) si ha una variazione di tensione $\Delta V_R = 125$ mV. Ciò richiede una particolare attenzione alla realizzazione degli amplificatori di lettura per rilevare queste variazioni relativamente ridotte della tensione di linea, come vedremo nel Paragrafo 13.9.2.

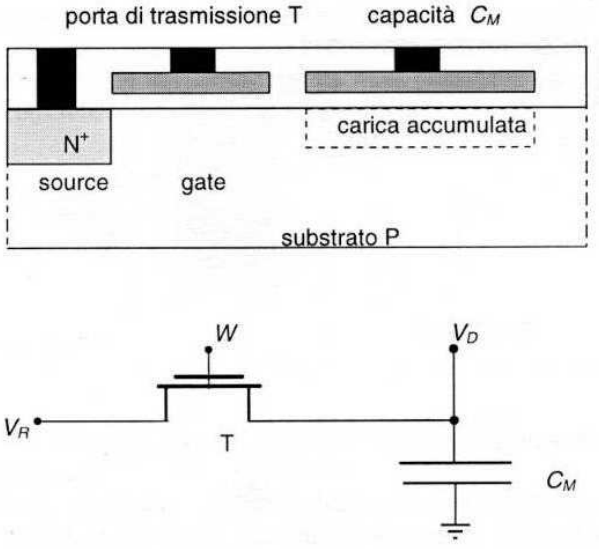


Figura 13.34 Cella a un transistore con capacità planare

La realizzazione di questa struttura prevede un transistore MOS per la porta e una capacità; quest'ultima viene realizzata utilizzando ancora l'ossido di gate come dielettrico, il polisilicio come armatura superiore e il substrato come armatura inferiore. La prima realizzazione è quella planare, riportata nella Figura 13.34. Il transistore NMOS della porta di trasmissione non ha il drain perché gli elettroni che percorrono il canale non debbono essere prelevati al drain ma vanno intrappolati nella regione di svuotamento che si crea alla superficie del silicio sottostante all'armatura in polisilicio della capacità.

Per aumentare il valore della capacità C_M occorre aumentare l'area utilizzata dalla capacità, e questo si traduce in un aumento dell'area della cella elementare; d'altra parte all'aumentare della capacità di memoria della DRAM occorre almeno mantenere inalterato il rapporto C_M / C_L, dal quale, come si è visto, dipende il segnale che deve essere rilevato dall'amplificatore di lettura. Occorre quindi ricorrere a soluzioni diverse per memorie ad elevata capacità, che contemperino la necessità di un aumento del valore di C_M con quello di una ridotta occupazione di area. La soluzione a questo problema è ancora una volta basata su una tecnologia innovativa sinteticamente riportata nella struttura di Figura 13.35. In questa struttura la superficie della capacità viene notevolmente incrementata, rispetto al caso planare, mantenendo nel contempo molto limitata l'occupazione dell'area della superficie del silicio, agendo in senso verticale alla superficie del silicio, e cioè scavando una trincea profonda nella quale viene depositato sia l'ossido sottile (dielettrico), che il polisilicio che costituisce l'armatura del condensatore. Questa struttura viene denominata "*trench capacitor*" con riferimento alla trincea scavata nel silicio con attacco selettivo in plasma.

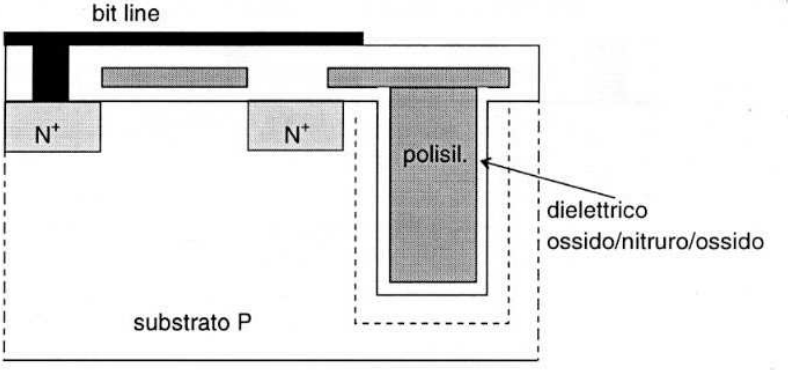


Figura 13.35 Cella di memoria con capacità "a trincea" (*trench capacitor*)

Un ulteriore incremento del valore della capacità, a parità di superficie utilizzata, viene dall'utilizzo di dielettrici a costante dielettrica maggiore di quella dell'ossido di silicio. Si utilizza a questo fine uno strato di nitruro di silicio tra due strati di ossido di silicio come dielettrico all'interfaccia tra la trincea nel silicio e il polisilicio di riempimento.

13.8.2 Circuiti di lettura per DRAM

L'operazione di lettura è particolarmente critica per le memorie DRAM a 1 transistore per i seguenti motivi:

a. la variazione di tensione che deve essere letta sulla bit line è molto ridotta (inferiore a 100 mV);
b. il pilotaggio della porta di trasmissione con un segnale di clock introduce un disturbo ulteriore sulla variazione di tensione che deve essere letta;
c. l'informazione contenuta in C_M viene distrutta quando la cella viene letta (perché viene modificata la carica immagazzinata nella capacità), per cui occorre ripristinare l'informazione nella cella ogni volta che viene effettuata una lettura.

Esaminiamo questi problemi e descriviamo le soluzioni impiegate per superare queste difficoltà.

Il primo punto è stato già valutato precedentemente, e discende dall'essere $C_M << C_L$. Per ottimizzare il valore di lettura ΔV_R, dato dalla (13.20), sia nel caso di $V_D = V_{OH} = V_{DD} - V_T$, che in quello di $V_D = V_{OL} = 0$ V, la scelta migliore per la tensione di riferimento della linea è $V_R = (V_{OH} + V_{OL})/2$, perché in tal caso si hanno valori di ΔV_R uguali ed opposti nei due casi. Per variazioni di tensioni così piccole sulla linea dati occorre utilizzare un amplificatore di lettura di tipo differenziale, come quelli già visti per le memorie SRAM. In questo caso però non vi sono due linee dati che presentano due segnali in opposizione,

come nel caso delle RAM con linee dati D e $\overline{D}$, per cui la soluzione è quella di dividere in due la linea dati e di inserire l'amplificatore di lettura tra le due metà della linea in modo da poter leggere lo squilibrio tra le tensioni che si produce quando una delle celle della linea (nella metà superiore o inferiore) viene letta, come è indicato schematicamente in Figura 13.36.

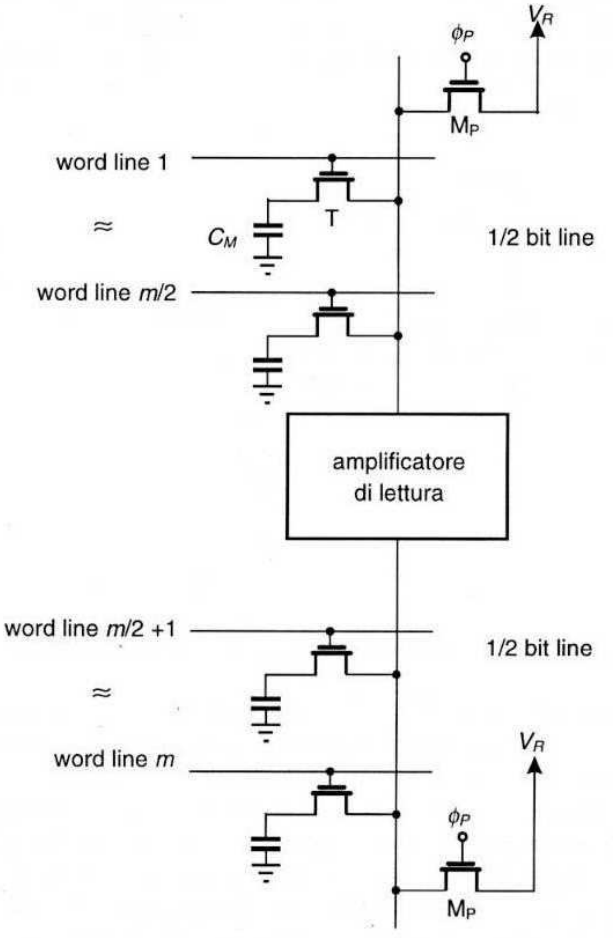


Figura 13.36 Amplificatore di lettura per singola bit line e cella a un transistore

Per la lettura si utilizza un circuito bistabile, che è sensibile a sbilanciamenti anche molto piccoli tra i due ingressi, e che a causa dell'effetto degenerativo della reazione positiva passa rapidamente ad una delle due condizioni stabili con livelli logici ben definiti alle uscite; le due metà della bit line sono entrambe precaricate a V_R mediante il comando di precarica ϕ_P applicato ai

transistori M_P, e la lettura della variazione ΔV_R viene effettuata abilitando il bistabile mediante un segnale di abilitazione. Vedremo che questa soluzione permette di risolvere anche il problema della distruzione dell'informazione durante la fase di lettura, punto c).

Per quanto riguarda il punto b), ricordiamo che si è già considerato questo effetto che è schematizzato nella Figura 13.37, parlando delle logiche a porte di trasmissione; si è visto (Equazione 11.9) che la variazione di tensione indotta sull'uscita (in questo caso la bit line) dal comando di fase applicato alla gate è legata al valore del rapporto tra la capacità $C_{GD,S}$ e quella C_L di carico della bit line.

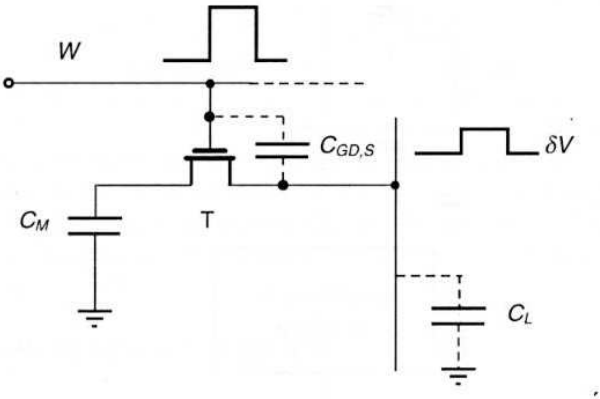


Figura 13.37 Effetto del segnale di abilitazione della porta sulla tensione della bit line

La capacità $C_{GD,S}$ della porta T, anche se è molto piccola rispetto a quella C_L della linea, non è molto inferiore alla capacità C_M di memoria, per cui la variazione δV indotta sulla linea a causa dell'accoppiamento del segnale di abilitazione non si può considerare trascurabile rispetto a ΔV_R, che è a sua volta molto bassa. Per depurare la lettura di ΔV_R da questo disturbo si utilizza ancora la configurazione differenziale dell'amplificatore di lettura introducendo volutamente sull'altra metà della linea, nella quale non viene abilitata nessuna cella, un disturbo uguale δV attraverso un pilotaggio di una *cella fittizia* (*dummy cell*) aggiunta ad ognuna delle due metà di bit line.

La struttura della bit line con l'amplificatore di lettura e le celle fittizie diventa quindi quella riportata in Figura 13.38. Assumendo m word line della DRAM, in ogni metà della bit line trovano posto le $m/2$ celle elementari a un transistore, più una cella fittizia che viene pilotata dallo stesso segnale W che abilita una delle porte T dell'altra metà della linea. Il bistabile utilizzato per la lettura viene abilitato da un segnale ϕ_S applicato al MOS M_S posto in serie verso massa.

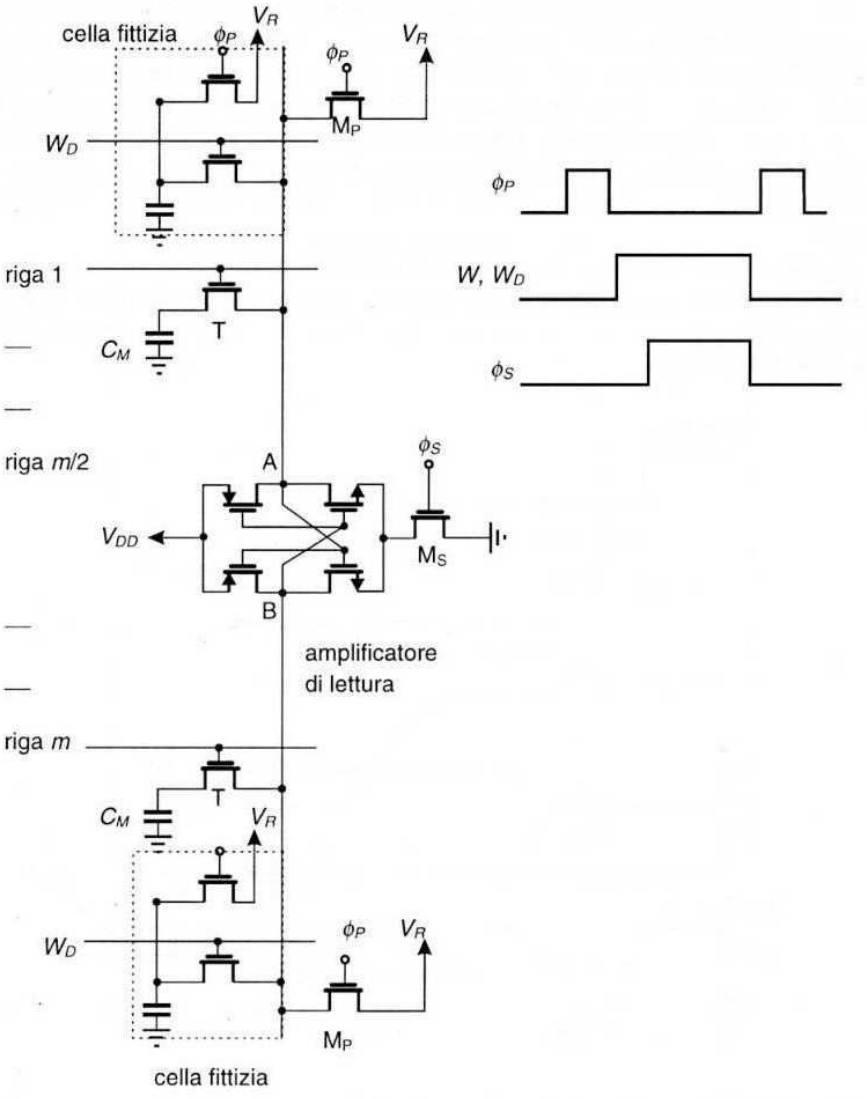


Figura 13.38 Schema elettrico di una bit line per una DRAM con celle a 1 transistore

L'operazione di precarica delle due metà della bit line alla tensione V_R è effettuata dalle porte M_P al cui ingresso si ritrova la tensione V_R, pilotate dal comando di precarica ϕ_P. Quando il segnale di precarica diventa basso, inizia la fase di lettura, che è articolata in più sottofasi, indicate nei diagrammi di temporizzazione dei segnali di Figura 13.39. Durante tutta la fase di lettura il segnale di abilitazione W di una delle word line (e il segnale W_D applicato alla cella fittizia) è alto. Supponiamo che sia stata abilitata una delle word line ($m/2 + 1 \div m$) della metà inferiore della bit

e che la cella di memoria corrispondente contenga memorizzato un 1 logico; all'apertura della porta T corrispondente avviene la ridistribuzione di carica tra la capacità di memoria C_M e quella della bit line C_L, generando la variazione di tensione ΔV; a questa si somma la variazione δV indotta dal comando sulla gate di T, che viene però compensata dalla uguale variazione indotta sull'altra metà della bit line dal pilotaggio della cella fittizia attraverso il segnale W_D, per cui la tensione di sbilanciamento tra i due ingressi del bistabile è ancora la differenza di potenziale ΔV.

Quando viene abilitato il transistore M_S in serie al bistabile mediante il segnale ϕ_S, questo può esplicare la sua azione rigenerativa portando, in seguito allo sbilanciamento iniziale, la tensione del nodo A al valore 0, e quello del nodo B al valore V_{DD}.

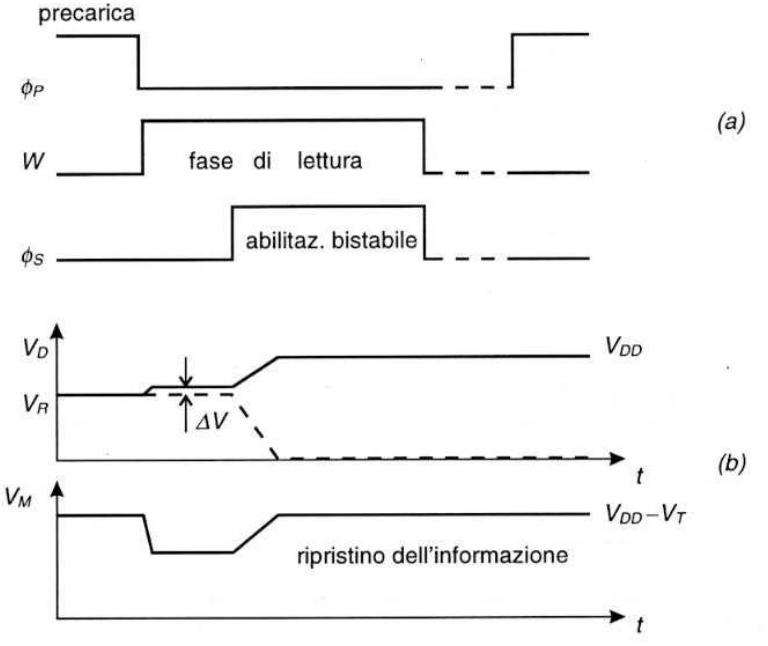


Figura 13.39 a) Temporizzazione dei segnali nel circuito di Figura 13.38; b) andamenti delle tensioni V_D delle due metà della bit line e V_M della capacità C_M

Il bistabile mantiene i suoi livelli logici dello stato stabile anche successivamente, per cui può essere utilizzato sia per eseguire la lettura dell'informazione memorizzata nella capacità C_M, che per il ripristino dell'informazione nella capacità stessa (che si era persa nella prima fase della lettura, in quanto la ridistribuzione della carica aveva portato la tensione V_M della capacità al valore V_R). Infatti, nella seconda parte della fase di lettura, sia la tensione della metà della bit line in cui vi è la cella indirizzata, sia la tensione della capacità della

cella stessa, si portano al valore alto (la bit line a V_{DD} e la capacità a $V_{DD} - V_T$), mentre la tensione della metà della bit line non indirizzata si porta a 0 (linea tratteggiata in Figura 13.39).

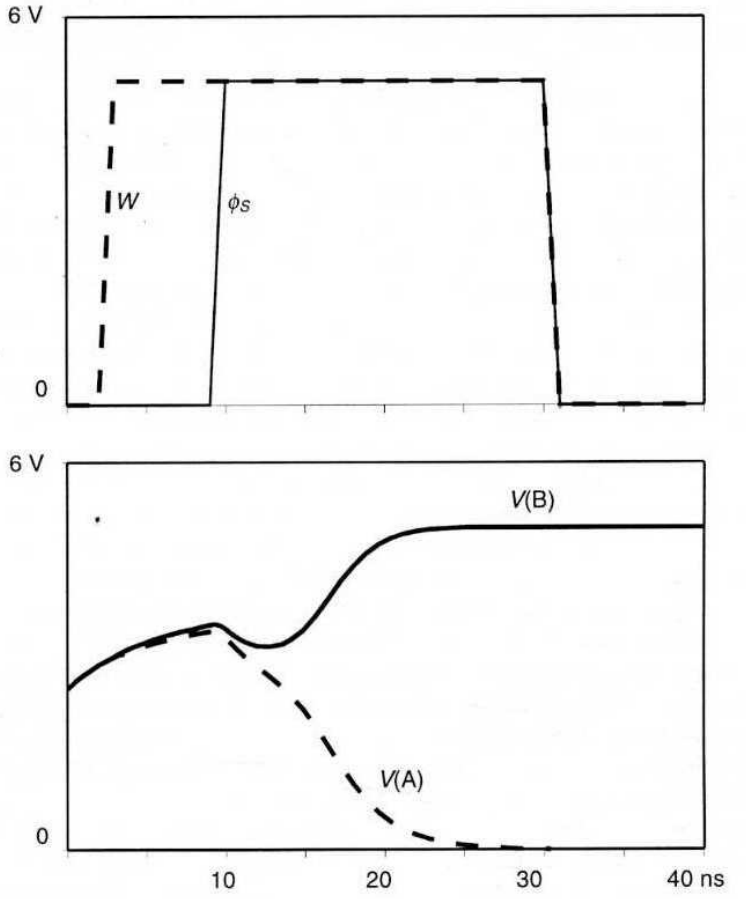


Figura 13.40 Simulazione SPICE dell'operazione di lettura di una cella di memoria a un transistore mediante un bistabile CMOS; a) segnali di lettura e di abilitazione del bistabile; b) andamenti delle tensioni ai due nodi A e B del bistabile

In Figura 13.40 è presentato il risultato di una simulazione SPICE per l'operazione di lettura dell'informazione di una cella di memoria relativa al circuito di Figura 13.39. La cella di memoria indirizzata è nella metà inferiore della bit line, e la capacità C_M è supposta carica al valore alto, pari a 3.6 V; la capacità C_L della metà della bit line è di 0.5 pF. Si vede come le tensioni dei due nodi A e B del bistabile, che viene abilitato portando in conduzione il MOS M_S, si portino rispettivamente alla tensione 0 e a V_{DD} in un tempo dell'ordine di 15 ns (il tempo relativa-

mente elevato è dovuto alla presenza delle capacità C_L ai due nodi del bistabile), a partire da uno squilibrio iniziale di un centinaio di millivolt.

L'impiego del bistabile come sense amplifier potrebbe a prima vista permettere la lettura di differenze di potenziale anche ridottissime, in quanto dalla teoria del bistabile sappiamo che bastano squilibri minimi rispetto alla condizione di equilibrio instabile ($V_A = V_B$) per instaurare il meccanismo rigenerativo che porta ad uno dei due stati stabili; questa possibilità potrebbe permettere un aumento della capacità di memoria della DRAM in quanto si potrebbe ridurre l'area occupata dalla C_M. In effetti le analisi sviluppate nel Paragrafo 12.3 non tengono in conto le inevitabili differenze dei parametri elettrici dei MOS che costituiscono il bistabile, né delle variazioni che questi parametri possono avere con la temperatura, la tensione e l'invecchiamento. Ad esempio nei circuiti VLSI non si riescono ad avere uguali valori di k' e della tensione di soglia V_T per tutti i dispositivi (parecchi milioni nel singolo chip), per inevitabili piccole variazioni dello spessore dell'ossido di gate e della resistività del materiale da punto a punto. Quindi se, ad esempio, le tensioni di soglia dei due invertitori del bistabile variano di qualche decina di mV tra loro (variazioni realistiche per la tecnologia attuale), eventuali variazioni di tensione ΔV inferiori alle differenze tra i valori di V_T non potranno essere rilevate, in quanto il bistabile, a causa dello squilibrio già esistente dovuto alle differenze di V_T, si porterà in una posizione preferenziale dipendente da questo squilibrio e non dal valore e dal segno di ΔV. In definitiva la minima differenza di tensione rilevabile dipende ancora da considerazioni tecnologiche sull'uniformità del processo realizzato.

L'operazione di ripristino periodico (*refresh*) della carica persa dalla capacità di memoria C_M per effetto delle correnti inverse delle giunzioni drain/substrato delle porte a cui questa è collegata (problema analogo a tutte le celle dinamiche considerate), viene effettuata, come nel caso del ripristino dell'informazione persa in fase di lettura, leggendo e riscrivendo l'informazione di tutte le celle connesse ad una singola riga (word line), abilitando sia una data riga che tutti i sense amplifier connessi alle celle di memoria presenti su quella riga, e cioè i sense amplifier di tutte le colonne (bit line). Questa operazione viene ciclicamente effettuata per tutte le word line, con intervallo di ciclo di refresh dell'ordine del ms.

Per una memoria DRAM da 128 kbit, l'architettura può essere quella indicata nella Figura 13.41. Il decodificatore di riga (word line) utilizza un indirizzo da 10 bit e fornisce 1024 word line; il decodificatore di colonna (bit line) utilizza un indirizzo di 7 bit e alimenta 128 colonne. Al centro di ciascuna colonna sono posti i 128 amplificatori di lettura, mentre ad ognuno dei due estremi sono poste le due linee di celle fittizie (dummy cells).

Le architetture delle memorie DRAM ad elevata capacità utilizzano anch'esse un'organizzazione a blocchi, basata su un'unità base di 128 kbit, secondo lo schema già presentato in Figura 13.30. I decodificatori di riga e di colonna vengono inseriti negli spazi tra i singoli moduli, in modo da utilizzare una struttura regolare anche nella realizzazione di memorie a maggiore capacità a partire da quelle a capacità inferiore.

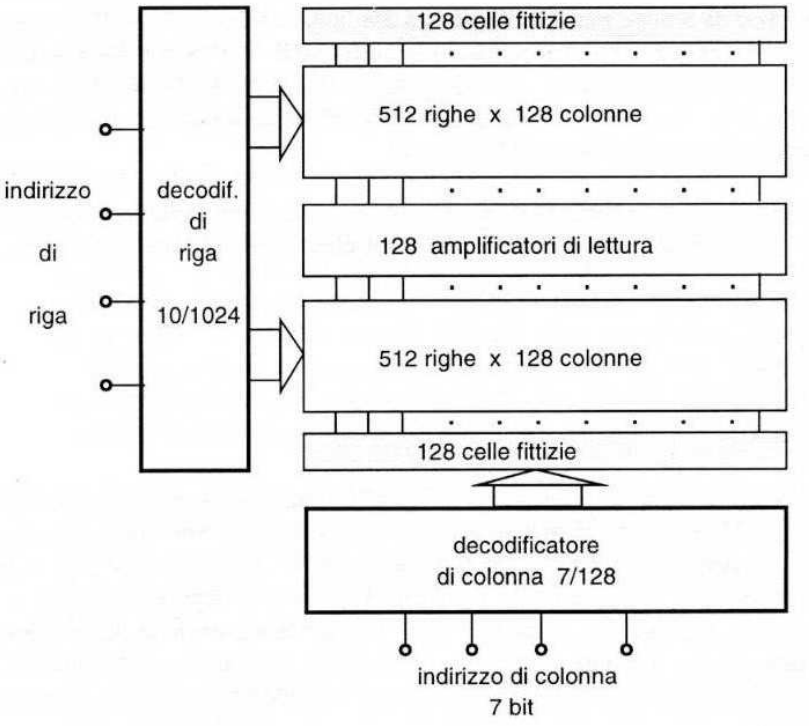


Figura 13.41 Organizzazione di una DRAM da 128 kbit con celle a un transistore

La descrizione dettagliata delle memorie DRAM, che costituiscono a tutti gli effetti dei sistemi digitali complessi, esula dai limiti di questo libro; per una trattazione più dettagliata di questi, come di altri sistemi digitali complessi, si suggerisce l'approfondimento su uno dei numerosi testi sui sistemi VLSI presenti in letteratura.

Esercizi di riepilogo

13.1 In un registro a scorrimento a 16 bit, realizzato con flip-flop D, determinare il massimo e il minimo tempo di lettura di una parola da 16 bit memorizzata, assumendo un segnale di clock con frequenza di 100 MHz.

13.2 Per una memoria ROM indirizzabile con parole da 16 bit, quante parole da 8 bit possono essere ottenute alle uscite? Volendo realizzare la memoria con matrici NOR in tecnologia pseudo-NMOS quanti dispositivi sono necessari per la realizzazione della memoria (non considerando i circuiti di disaccoppiamento e inversione per gli ingressi e le uscite)?

13.3 Determinare la dissipazione di potenza statica di una memoria ROM con indirizzi da 8 bit, e matrice di codifica 256 × 8, in tecnologia NMOS con carico a svuotamento, dimensionando le porte NOR NMOS con $K_R = 4$, e utilizzando i seguenti valori: $k'_N = 50\ \mu A/V^2$, $V_{TN} = 0.8$ V, $V_{TD} = -3$ V, $V_{DD} = 3.3$ V; ripetere il calcolo con una matrice 64 × 32: quale dei due casi presenta la minore dissipazione di potenza?

13.4 Disegnare a) lo schema elettrico e b) il tracciato di una matrice di codifica con porte NOR pseudo-NMOS per ROM che fornisca le seguenti parole in uscita:
1 0 1 0 0
0 1 1 0 0
0 0 0 1 1
1 0 1 1 1
0 1 0 0 0

13.5 Valutare il tempo di propagazione t_{PLH} all'uscita di un decodificatore 8/256 per ROM realizzato in tecnologia pseudo-NMOS con word line di polisilicio; si assumano i seguenti valori: $k'_N = 50\ \mu A/V^2$, $k'_P = 20\ \mu A/V^2$, $V_{TN} = 0.8$ V, $V_{TD} = -3$ V, $W/L|_N = W/L|_P = 2/1\ \mu m$, $V_{DD} = 3.3$ V, Rpoly = 50 Ω/□, e si consideri la word line come una linea *RC* per la valutazione del ritardo di propagazione. Valutare il miglioramento che si ottiene se si cortocircuita la linea di polisilicio con una di metallo con connessioni tra le due linee ogni 16 celle.

13.6 Con riferimento all'Esercizio 13.3, determinare i tempi di lettura per le due configurazioni della matrice di codifica proposte, utilizzando per i transistori i valori riportati nell'Esercizio 13.5 e per le capacità unitarie quelli di Tabella 3.2.

13.7 Disegnare lo schema elettrico in logica domino di una memoria ROM che associ le seguenti parole agli indirizzi:

00	1 0 1 0 0
01	0 1 1 0 0
10	0 0 0 1 1
11	0 1 1 1 0

13.8 Per una SRAM da 16 Mbit con celle di memoria NMOS con carico in polisilicio e alimentazione a 3.3 V, determinare la minima resistenza di carico compatibile con una dissipazione di potenza statica del chip di 1W; determinare per questa cella i valori di V_{OH} e V_{OL}, utilizzando i valori dei parametri riportati nell'Esercizio 13.3.

13.9 Per la cella NMOS di Figura 13.18a, valutare, mediante simulazioni SPICE, il dimensionamento massimo delle porte di trasmissione T_1 e T_2, che per-

metta il più basso tempo di scarica della capacità di linea C_L, compatibilmente con una sovratensione sull'uscita bassa V'_{OL} del bistabile che non superi il valore di soglia V_T. Utilizzare i valori di dimensionamento per i transistori del bistabile utilizzati per le simulazioni riportate in Figura 13.19, e $C_L = 1$ pF, $V_{DD} = 3.3$ V.

13.10 Dimensionare, mediante le formule analitiche approssimate, i transistori della cella di memoria CMOS con linee caricate da PMOS, della Figura 13.22, in modo da soddisfare alle condizioni imposte dalle (10.17) e (10.18), utilizzando i valori dei parametri dell'Esercizio 13.5, tenendo inoltre in conto, nel dimensionamento dei PMOS di carico, della costante di tempo relativa alla capacità C_L della bit line, che deve essere inferiore a 20 ns. Effettuare inoltre delle simulazioni SPICE del circuito così dimensionato per valutare l'influenza delle approssimazioni di comportamento in regime lineare dei dispositivi sui risultati ottenuti.

13.11 Per la cella di memoria bipolare di Figura 13.25, assumendo una dissipazione statica della cella in condizioni di quiescenza di 0.1 mW, determinare i valori delle resistenze R_1 e R_2 di carico. Determinare i valori delle resistenze R_3 e R_4 affinché la lettura di un 1 sulla linea D corrisponda ad una variazione ΔV di 0.5 V.

13.12 Per la cella di memoria dinamica a 4 transistori di Figura 13.31, assumendo una capacità di bit line $C_L = 1$pF, $k'_N = 50$ μA/V^2, $V_{TN} = 0.8$ V, $W/L|_N = 2/1$ μm, $V_{DD} = 3.3$ V, dimensionare le porte T_1 e T_2 affinché all'atto dell'apertura delle porte la tensione al nodo del latch che si trova a livello logico basso, non superi il valore di 0.6 V.

13.13 Con riferimento alla cella di memoria dinamica dell'Esercizio 13.12, determinare il tempo necessario affinché le capacità C_G si scarichino del 10% del valore massimo, supponendo una corrente inversa delle giunzioni drain (source)/substrato di 1 pA; determinare inoltre il tempo necessario per ripristinare il livello logico alto (approssimando la carica della capacità a corrente costante).

13.14 Determinare il valore della tensione V_D che si stabilisce sulla capacità C_M della cella a un transistore, se la bit line è portata a V_{DD}, utilizzando i seguenti parametri per il MOS della porta di trasmissione: $k'_N = 50$ μA/V^2, $V_{TN} = 0.8$ V, $W/L|_N = 2/1$ μm, $\gamma = 0.6$ V, $\phi* = 0.6$ V.

13.15 Per la cella ad un transistore di Figura 13.33, utilizzando i parametri dell'Esercizio 13.14, determinare il tempo necessario alla ridistribuzione della carica per avere la variazione ΔV sulla bit line, dopo l'apertura della porta T, (si consideri il transitorio estinto dopo 4 costanti di tempo).

13.16 Effettuare una simulazione SPICE del circuito di lettura per celle a un transistore di Figura 13.38, schematizzando una sola cella di memoria nella metà inferiore della bit line, i due transistori di precarica M_P, e il bistabile CMOS con $W/L|_N = 2/1$ μm, $W/L|_P = 5/1$ μm, e con $V_R = 2.5$ V, $V_{DD} = 5$ V. Assumere una tensione $V_D = 0$ ai capi di C_M prima dell'apertura della porta T, e valutare il tempo necessario allo stabilirsi dei livelli logici V_{DD} e 0 ai due nodi B e A del bistabile. Ripetere le simulazioni per diversi valori delle dimensioni W/L dei transistori del bistabile rispetto a quelli di base per avere tensione di soglia pari a $V_{DD}/2$, valutando per ogni caso: a) il tempo di lettura, b) la potenza dissipata durante la fase di abilitazione.

Riferimenti bibliografici

H. Taub, D. Schilling, *Elettronica Integrata Digitale*, Jackson, Milano, 1981.

D.A. Hodges, H.G. Jackson, *Analisi e progetto dei circuiti integrati digitali*, Bollati Boringhieri, Torino, 1991.

H.E. Weste, K. Eshraghian, *Principles of CMOS VLSI design: A system perspective*, 2nd ed., Addison-Wesley, 1993.

J. Millman, A. Grabel, *Microelettronica*, McGraw-Hill Libri Italia, Milano, 1994.

J.F. Wakerly, *Digital design, principles and practices*, 2nd ed., Prentice Hall, Englewood Cliffs, 1994.

R.C. Jaeger, *Microelectronic Circuit Design*, McGraw-Hill Int., New Jork, 1997, tr. it. di prossima pubblicazione.

A

Richiami sul simulatore SPICE

A.1 Premessa

Il software di simulazione dei circuiti integrati SPICE è diventato di fatto lo strumento CAD maggiormente impiegato per l'analisi e il progetto di circuiti integrati sia analogici che digitali. Per tale ragione il simulatore è stato ampiamente utilizzato nell'analisi e nel progetto dei circuiti presentati nel libro, sia al fine di una migliore comprensione del loro comportamento statico e dinamico, che per effettuare analisi dettagliate e verificare la bontà delle approssimazioni impiegate nei calcoli manuali.

In questa Appendice verranno fornite alcune note sull'uso e sulle caratteristiche del programma di simulazione SPICE, rimandando, per ulteriori istruzioni, al manuale d'uso del programma PSPICE™. Benché PSPICE sia un prodotto commerciale, la versione dimostrativa del programma è di pubblico dominio; nel seguito faremo riferimento alla versione dimostrativa PSPICE 5.0 DEMO o alle versioni equivalenti.

Il programma SPICE permette di effettuare analisi in continua ed in transitorio di circuiti elettronici lineari e non lineari e l'analisi in regime sinusoidale di circuiti lineari. Il programma ha un menu di controllo, indicato con PS, che permette di gestire le diverse operazioni, come la scrittura dei file di circuito, l'editing degli stessi, il salvataggio e l'apertura dei file, l'analisi, la lettura dei risultati, la rappresentazione grafica delle funzioni di uscita, e così via.

I comandi e la descrizione del circuito da analizzare vengono forniti tramite la specificazione di un file d'ingresso in formato ASCII, il cui nome ha come estensione `.CIR`. I risultati dell'analisi sono restituiti in un file di uscita (anch'esso in formato ASCII) con estensione `.OUT` ed un file con estensione `.DAT`; quest'ultimo contiene le informazioni circa le evoluzioni delle grandezze elettriche (tensione e corrente), che possono essere visualizzate graficamente mediante il programma di visualizzazione ed elaborazione PROBE.

A.2 Descrizione del circuito

Nel file d'ingresso al simulatore SPICE, indicato con l'estensione `.CIR`, si definiscono la topologia del circuito e gli elementi che lo compongono; inoltre si specifica il tipo di analisi che si intende effettuare, la temperatura a cui riferire la simulazione e il formato dei dati in uscita. Il file di ingresso può essere scritto con un qualsiasi editor in formato ASCII, esso è organizzato a linee, quindi ad ogni riga corrisponde un comando. È possibile proseguire una linea sulla riga successiva, se questa viene fatta iniziare in prima colonna con il carattere '+'. La prima riga viene interpretata come titolo della simulazione e non viene presa in considerazione per l'analisi. L'ultima riga *deve* contenere l'istruzione `.END` per segnalare la fine dei comandi. Fra queste due righe si inseriscono le righe contenenti la descrizione dei dispositivi componenti il circuito, le modalità di analisi e le scelte sul formato delle uscite.

Ogni elemento componente il circuito deve essere specificato con (nell'ordine) nome, nodi dei terminali, valore o modello ed eventuali opzioni. In particolare, la prima lettera del nome (in cui solo le prime otto lettere sono considerate come distintive) denota il tipo di dispositivo. Questa lettera vale:

R per le resistenze
C per le capacità
L per le induttanze
V per i generatori di tensione
I per i generatori di corrente
Q per i transistori bipolari
M per i transistori MOS
D per i diodi

Separati da uno o più spazi, dopo il nome, si indicano i nodi del circuito (tipicamente descritti con numeri) a cui devono essere connessi i terminali dei componenti o dispositivi, che possono essere in numero di due o più a seconda del dispositivo in considerazione. Per i dispositivi a due terminali, quando è il caso, il primo nodo rappresenta il polo positivo ed il secondo il negativo. I transistori presentano tre (eventualmente quattro) nodi; per i transistori bipolari l'ordine di definizione dei terminali è: *collettore*, *base*, *emettitore* e *substrato*. Per i transistori MOS l'ordine è: *drain*, *gate*, *source* e *body*. Sia per i transistori bipolari che per i MOS quando il quarto nodo viene omesso, il terminale viene implicitamente considerato a massa. La scelta dei numeri da associare ai nodi è libera, tranne per il numero 0 che per convenzione rappresenta la massa.

I componenti elettrici, quali le resistenze, le capacità, le induttanze, i generatori costanti di tensione e di corrente, sono descritti completamente assegnando il valore di resistenza [Ohm], di capacita [F], di induttanza [H], di tensione [V], e di corrente [A]. I valori possono essere interi, a virgola mobile con notazione decimale o scientifica; si possono utilizzare fattori di scala mnemonici come:

G	per 10^9
MEG	per 10^6

M	per 10^{-3}
U	per 10^{-6}
N	per 10^{-9}
P	per 10^{-12}

È possibile, come vedremo, specificare generatori di tensione e di corrente variabili nel tempo, che di solito rappresentano i segnali d'ingresso nelle analisi in funzione del tempo; si possono simulare generatori lineari a tratti (PWL), ad impulsi trapezoidali (PULSE), sinusoidali (SIN) ed esponenziali. Successivamente alla specifica del tipo di generatore si definiscono i parametri caratteristici.

Esempi per i generatori:

Un generatore di tensione (o di corrente) con un generico andamento della grandezza nel tempo si può definire mediante la descrizione PWL basata sulla definizione di un diagramma temporale a spezzate (*piecewise linear*), in cui ogni punto della spezzata è identificato dall'ascissa e dall'ordinata. Ad esempio, per simulare un generatore di tensione (Vin) posto fra i nodi 0 e 1, che ha tensione nulla da 0 a 2 ns, che aumenti la tensione linearmente raggiungendo 5 V a 3 ns, che resti costante a 5 V fino a 20 ns, e infine decresca linearmente da 5 V raggiungendo 0 V a 21 ns, per poi permanere a questo valore di tensione fino a 30 ns, si fornisce la seguente istruzione:

```
Vin 1 0 PWL(0ns, 0V2ns, 0V3ns, 5V20ns, 5V21ns, 0V30ns, 0V)
```

Se si vuole simulare un generatore di tensione (o di corrente) periodico impulsivo si utilizza la descrizione PULSE, in cui si definisce il livello minimo e massimo, il ritardo del fronte di salita, la durata del fronte di salita, quella del fronte di discesa, la durata della parte alta e il periodo di ripetizione. Ad esempio, se si vuole simulare un segnale di tensione (Vin) posto fra i nodi 0 e 2, che eroghi una tensione variabile fra 0.2 V e 5 V, con ritardo del primo fronte di salita di 5 ns, fronte di salita di 1 ns, di discesa pari a 2 ns, durata dell'impulso al valore alto di 50 ns e periodo 100 ns :

```
Vin 2 0 PULSE(0.2 V 5 V 5 lns 2 ns 50 ns 100 ns)
```

Un generatore di tensione (o di corrente) sinusoidale è descritto dall'istruzione SIN, in cui vanno definiti i parametri dell'onda sinusoidale come offset, ampiezza, frequenza di oscillazione, e (eventualmente) costante di attenuazione e sfasamento iniziale. Ad esempio se si vuole simulare un generatore di tensione sinusoidale (Vin), posto fra i nodi 0 e 5, con andamento dell'ampiezza decrescente con legge esponenziale (sinusoide smorzata), che eroghi una tensione con offset di 0 V, ampiezza iniziale di 2 V, frequenza di oscillazione di 100 kHz, ritardo iniziale 10 μs, costante di tempo 10^{-5} s e fase iniziale di 30 gradi, si utilizza la seguente descrizione:

```
Vin 5 0 SIN(0V 2V 100K 10us 1E5 30)
```

Simulazione dei dispositivi a semiconduttore

In SPICE i dispositivi a semiconduttore vengono simulati facendo riferimento a diversi modelli elettrici. Per simulare correttamente le diverse caratteristiche elettriche si devono specificare i parametri caratteristici; a tale scopo si usano nel file `*.CIR` le istruzioni `.MODEL` nelle quali si definisce il nome mnemonico del dispositivo, il tipo di dispositivo ed i parametri caratteristici. Tali schede possono essere comuni a più di un dispositivo utilizzato nel circuito, purché di tipo analogo. Quindi per dispositivi come transistori bipolari, transistori MOS e diodi, la sintassi per la loro descrizione nel file `.CIR` è la seguente: nome, nodi dei terminali, nome mnemonico della scheda `.MODEL` a cui ci si riferisce ed eventuali parametri geometrici.

Riguardo le schede `.MODEL`, ci si può riferire al manuale d'uso del programma SPICE per una trattazione più approfondita. Nelle nostre simulazioni faremo uso delle istruzioni `.MODEL` dei dispositivi MOS e bipolari raccolte in una libreria di dispositivi appositamente creata, come verrà specificato in Appendice B.

Durante le diverse analisi tutti i parametri dei dispositivi a semiconduttore sono valutati alla temperatura di 27 °C. Mediante il comando `.TEMP` è possibile assegnare una diversa temperatura (espressa in gradi centigradi) a cui far effettuare le simulazioni; se si specifica più di un valore di temperatura verranno effettuate tante simulazioni quanti sono i valori specificati.

Dopo la descrizione del circuito e dei modelli dei dispositivi, si specifica il tipo di analisi che si intende effettuare con i comandi `.DC`, `.AC`, `.TRAN`, in cui si specificano di seguito i campi di analisi.

A.3 L'analisi statica

L'analisi del funzionamento statico di un circuito si effettua inserendo il comando `.DC` nel file di descrizione del circuito `*.CIR`; essa permette di calcolare il punto di funzionamento a riposo. Tutte le induttanze e le capacità presenti nel circuito (o nei modelli dei dispositivi specificati) vengono sostituite rispettivamente da corti circuiti o da circuiti aperti. Se sono stati specificati generatori variabili nel tempo, SPICE considera il valore all'istante $t = 0$. È possibile ottenere anche la valutazione dei parametri dei modelli per piccoli segnali dei dispositivi non lineari presenti nel circuito con l'inserimento dell'istruzione `.OP` (che comunque effettua un'analisi in continua e che quindi, quando specificata, permette di omettere l'istruzione `.DC`). Si può ottenere un'analisi statica per diversi valori di una grandezza elettrica (ad esempio una caratteristica di trasferimento in funzione della tensione di ingresso), indicando nell'istruzione `.DC` la grandezza che deve variare, i valori minimo e massimo, e il passo di variazione.

L'analisi in continua viene sempre effettuata prima dell'analisi in transitorio per valutare le condizioni iniziali, e prima dell'analisi sinusoidale per la valutazione dei parametri a piccoli segnali. I risultati dell'analisi in continua sono riportati nel file `*.OUT`, che contiene lo stesso nome del circuito analizzato con il file `*.CIR`; il file `*.OUT` è leggibile attivando il comando *<Browse output>* del menu *files* di PS.

Esempio di analisi statica

Se si vuole effettuare l'analisi statica di un circuito, a diversi valori della tensione di un generatore indipendente Vin (fra 0 V e 5 V con passo 0.5 V), si utilizza il comando `.DC` in questo modo:

```
.DC Vin 0V 5V 0.5V
```

In tal modo si otterranno i valori statici di tutte le tensioni in ogni nodo del circuito e di tutte le correnti in ogni ramo, nonché dei parametri dei dispositivi dipendenti da tali grandezze, per tensioni del generatore Vin variabili tra 0 V e 5 V con passo 0.5 V. I valori delle grandezze elettriche di ogni punto del circuito saranno disponibili anche per la visualizzazione grafica mediante il menu PROBE.

A.4 L'analisi in frequenza

Quando si vuole studiare la risposta in frequenza di una rete si utilizza il comando `.AC`; l'analisi parte da una frequenza *fstart* e termina ad una frequenza *fstop* con una variazione lineare o logaritmica, e in quest'ultimo caso con una scelta di variazione per decade o per ottava. Occorre anche definire il numero di punti nella variazione totale di frequenza se lineare o il numero di punti rispettivamente per decade o per ottava negli altri due casi. Gli unici generatori di segnale indipendenti ammessi per l'analisi a.c. sono generatori a.c. di corrente o di tensione indicati con il comando:

```
<v (i)> x  y  ac N  M
```

dove `x` e `y` indicano i nodi a cui è connesso il generatore `V` (o `I`), `N` indica l'ampiezza del segnale e `M` lo sfasamento.

A.5 L'analisi in transitorio

Quando si vuole studiare l'evoluzione in transitorio di un circuito si utilizza il comando `.TRAN`; l'analisi parte dall'istante $t = 0$ e si conclude all'istante finale specificato nel comando, in cui si indica anche il passo temporale in cui ottenere i dati di uscita. Il passo di integrazione viene scelto dal programma in maniera automatica, ma rimane comunque sempre inferiore o uguale al passo imposto nella scheda.

Effettuando l'analisi in transitorio è possibile assegnare delle condizioni iniziali adoperando il comando `.IC` che permette di imporre tensioni ai nodi o correnti ai dispositivi diverse da quelle che si avrebbero dalla preventiva analisi in continua.

Esempio di analisi in transitorio

Se si vuole effettuare l'analisi in transitorio di un circuito da 0 ns fino a 50 ns, con passo di integrazione massimo di 0.02 ns, si utilizza il comando `.TRAN` in questo modo:

```
.TRAN 0.02 ns 50 ns
```

Si otterranno, così, le evoluzioni temporali di tutte le tensioni e correnti del circuito, che potranno essere visualizzate mediante il programma PROBE.

A.6 Sottocircuiti e librerie

Quando una stessa rete elettrica si ripete più volte nel circuito da studiare, al fine di non ripetere più volte le stesse righe di descrizione, è possibile definire un sottocircuito. Esso inizia e finisce con le istruzioni `.SUBCKT` e `.ENDS` rispettivamente. L'istruzione `.SUBCKT`, oltre al nome, da usare per richiamare il sottocircuito durante la descrizione del circuito, è composta dalla lista dei nodi dei terminali accessibili all'esterno. Il sottocircuito viene quindi trattato come un normale dispositivo che presenta tanti morsetti quanti quelli definiti nella lista dei nodi dell'istruzione `.SUBCKT`. Quando, nella descrizione del circuito, si vuole inserire la parte di circuito definita da un sotto circuito, lo si nomina con una stringa che inizi con la lettera `X` poi si specificano i nodi a cui il sottocircuito è connesso (con riferimento all'ordine espresso nell'istruzione `.SUBCKT`) ed infine si indica il nome del sottocircuito da considerare.

Le istruzioni `.MODEL` e `.SUBCKT` (con le relative linee di comandi fino al comando `.ENDS` incluso) possono essere anche scritte in un file diverso da quello utilizzato per la descrizione del circuito da studiare. Infatti è possibile scrivere un file (sempre in formato ASCII) con suffisso `.LIB` in cui descrivere i modelli dei dispositivi a semiconduttore e i sottocircuiti di interesse. Durante la stesura del file `.CIR`, inserendo il comando `.LIB` seguito dalla specifica del nome del file (con il suo suffisso `.LIB`) si ottiene l'inclusione dei modelli e dei sottocircuiti descritti nella libreria come se fossero stati scritti nel file `.CIR`.

Nell'Appendice B si riporta il file della libreria di dispositivi utilizzati per le analisi successive, file che verrà indicato come: `DISPO.LIB`. Esso contiene le schede `.MODEL` di transistori NMOS ad arricchimento e a svuotamento, transistori PMOS, transistori bipolari NPN e diodi.

A.7 Analisi multiple

Si possono analizzare più versioni dello stesso circuito con differenti dispositivi, componenti, o tensioni di alimentazione, e presentare i dati di uscita contemporaneamente, inserendo uno dopo l'altro i file che specificano i differenti circuiti in un unico file `*.CIR`. Ogni file che specifica un dato caso o circuito deve essere

terminato dall'istruzione .END; l'analisi viene fatta in successione per i diversi casi, e tutti i dati (come anche le singole sezioni) sono disponibili per essere visualizzati dal menu PROBE, se il comando .PROBE è inserito in ogni caso da analizzare.

I diversi circuiti possono anche contenere un numero di nodi e di componenti diverso per ogni circuito, nel qual caso il menu PROBE avverte l'utente che alcuni nodi possono non corrispondere tra i diversi circuiti analizzati, ed è possibile visualizzare una o più sezioni del file .CIR, scegliendo i casi da visualizzare mediante una loro identificazione fornita dalla prima linea di istruzione (quella di inizio o quella immediatamente seguente il comando .END); conviene quindi in tale posizione inserire un nome che permetta di identificare il caso analizzato.

A.8 Rappresentazioni delle uscite

È possibile scegliere il formato dei risultati in uscita; esso può essere selezionato mediante le schede .PRINT e .PLOT, che vengono seguite dall'elenco delle grandezze che si vuole stampare o disegnare.

La modalità più utilizzata per visualizzare i risultati delle simulazioni è quella del comando .PROBE, che permette di salvare tutte le grandezze elettriche del circuito in un file con suffisso .DAT che verrà utilizzato come ingresso del menu di elaborazione grafica PROBE. Quest'ultimo permette di visualizzare, elaborare, e stampare in diversi modi le grandezze d'interesse. Il menu permette la scelta delle grandezze da porre sull'asse x e su quello y, come anche di presentare i dati su uno o più grafici, di effettuare operazioni tra le variabili, come somma, sottrazione, integrale, derivata, media e altro, e di aggiungere linee, testo e commenti ai grafici. È inoltre possibile porre dei cursori sulle curve in modo da visualizzare in apposite finestre i valori delle ascisse e ordinate di ogni punto, come anche le differenze di ascisse e ordinate tra due punti, specificati mediante due cursori distinti.

Indipendentemente dalla scelta del formato di uscita, SPICE genera un file .OUT in cui sono riportati, oltre alla descrizione del circuito che è stato analizzato (includendo anche le parti che sono state definite nei file .LIB), l'elenco e i valori dei parametri dei modelli utilizzati ed altri dati, come ad esempio la potenza dissipata, il punto di funzionamento, i valori delle correnti fornite dai generatori di tensione, e altro. Inoltre, se si è verificato un errore sintattico, nel file .OUT viene fornita l'indicazione della sua possibile causa.

B

Schede .MODEL dei dispositivi

B.1 Premessa

In questa appendice è riportato il file `DISPO.LIB`, contenente le schede `.MODEL` dei dispositivi che possono essere utilizzate per le simulazioni SPICE delle porte e dei circuiti logici riportate in Appendice C.

Si rimanda al manuale SPICE per la sintassi e la descrizione delle schede `.MODEL`; si ricorda che le schede `.MODEL` dei dispositivi possono essere inserite direttamente nei file `*.CIR` che descrivono il circuito da simulare, oppure possono essere raggruppate in un'apposita libreria, e richiamate nel file `*.CIR` mediante il comando *.lib**.lib*, dove * è il nome dato al file `.LIB` utilizzato.

Nei file `*.CIR` presentati in Appendice C è stata adottata quasi generalmente questa seconda modalità per l'introduzione dei singoli dispositivi, mediante la definizione di un file `DISPO.LIB`. Questo file riporta i modelli di dispositivi con tecnologia attuale e tolleranze minime inferiori al micron; per i diversi dispositivi si sono adottati i seguenti nomi:

MOS a canale N ad arricchimento	= MN
MOS a canale P ad arricchimento	= MP
MOS a canale N a svuotamento	= MND
Transistore bipolare NPN con $\beta_F = 50$	= QN1
Transistore bipolare NPN con $\beta_F = 20$	= Qi
Transistore bipolare NPN ad area larga	= QNbig
Diodo P/N	= DX
Diodo Schottky	= DSchttky

Nelle Tabelle B.1÷B.3 sono riportati il nome, il significato fisico e le relazioni con le espressioni utilizzate nel testo, per i parametri utilizzati nelle schede `.MODEL` dei diversi dispositivi.

Per i transistori MOS, il simulatore SPICE permette di scegliere fra differenti modelli, indicati come *level* 1, *level* 2, *level* 3, *level* 4. Per le simulazioni si farà riferimento al modello indicato come *level* 1 e ai parametri indicati in Tabella B.1. Le dimensioni geometriche W e L della regione di gate, e le aree A_D e A_S delle regioni di drain e source vengono indicate nel file `.CIR` nella linea che identifica il MOS specifico nel circuito in esame. I valori delle capacità di giunzione sono sempre specificati per tensione nulla ai capi della giunzione; la dipendenza dalla tensione di contropolarizzazione è definita dal coefficiente m (che sostituisce il coefficiente 1/2 della Equazione (2.11) nel caso più generale di giunzioni diffuse).

Tabella B.1 Parametri per la simulazione SPICE di transistori MOS

parametro	*nome*	*unità*	*espressione*
modello adottato	LEVEL	–	–
tensione di soglia	VTO	V	V_{T0}
spessore dell'ossido di gate	TOX	m	t_{OX}
drogaggio del substrato	NSUB	cm^{-3}	N_S
modulazione della lunghezza di canale	LAMBDA	V^{-1}	λ
coefficiente dell'effetto body	GAMMA	$V^{1/2}$	γ
potenziale di inversione di superficie	PHI	V	ϕ^*
potenziale di barriera	PB	V	ϕ_0
mobilità superficiale	UO	cm^2/Vs	μ
capacità di overlapping gate/source	CGSO	F/m	C_{GSO}
capacità di overlapping gate/drain	CGDO	F/m	C_{GDO}
capacità di overlapping gate/body	CGBO	F/m	C_{GBO}
corrente di saturazione delle giunzioni	IS	A	I_S
resistenza per quadro source/drain	RSH	ohm	$\Omega/\square$
capacità di giunzione (per area)	CJ	F/m^2	C_{J0}
capacità di giunzione (per perimetro)	CJSW	F/m	C_{JW}
coefficiente per la capacità CJ	MJ	–	m
coefficiente per la capacità CJSW	MJSW	–	m'

Per i transistori bipolari, il modello utilizzato da SPICE è quello di Gummel-Poon modificato. I parametri per la descrizione dei dispositivi mediante questo modello, indicati nella scheda `.MODEL` relativa, sono riportati in Tabella B.2. Anche in questo caso, le capacità di giunzione sono valutate per tensione nulla; la dipendenza dalla tensione viene specificata attraverso i coefficienti *me*, *mc*, *ms* in accordo con le Equazioni (6.27), (6.28), (6.29).

Tabella B.2 Parametri per la simulazione SPICE di transistori bipolari

parametro	*nome*	*unità*	*espressione*
guadagno di corrente in modo diretto	BF	–	β_F
guadagno di corrente in modo inverso	BR	–	β_R
resistenza ohmica di collettore	RC	ohm	R_C
resistenza ohmica di base	RB	ohm	R_B
resistenza ohmica di emettitore	RE	ohm	R_E
tensione di Early in modo diretto	VAF	V	V_{AF}
tensione di Early in modo inverso	VAR	V	V_{AR}
tempo di transito in modo diretto	TF	s	τ_F
tempo di transito in modo inverso	TR	s	τ_R
capacità base/emettitore	CJE	F	C_{JE0}
capacità base/collettore	CJC	F	C_{JC0}
capacità di collettore/substrato	CJS	F	C_{JS0}
potenziale di barriera base/emettitore	VJE	V	ϕ_0
potenziale di barriera base/collettore	VJC	V	ϕ_0
potenziale di barriera di substrato	VJS	V	ϕ_0
coefficiente per la capacità CJE	MJE	–	*me*
coefficiente per la capacità CJC	MJC	–	*mc*
coefficiente per la capacità CJS	MJS	–	*ms*
corrente di saturazione inversa	IS	A	I_S
corrente critica per la riduzione di β	IKR	A	–

Si può definire un valore di AREA nella linea che definisce lo specifico transistore del circuito, analogamente: ai MOS, questo parametro non va considerato come l'area effettiva di una delle regioni del transistore, ma come un fattore moltiplicativo (>1 o <1) che modifica sia le correnti che le resistenze e le capacità indicate nella scheda `.MODEL`; questi parametri quindi assumono il valore indicato nella scheda `.MODEL` per AREA = 1.

I transistori bipolari utilizzati nelle simulazioni sono indicati con tre nomi diversi: QN1 corrisponde al transistore NPN standard, utilizzato nelle porte bipolari; Qi rappresenta il transistore di ingresso di porte TTL, per il quale si è effettuato un controllo del lifetime nella base per avere $\beta_R << 1$, e QNbig è un transistore ad area maggiore, utilizzato per gli stadi di uscita.

Per i diodi, i parametri da specificare nella scheda `.MODEL` sono un sottoinsieme di quelli validi per i transistori bipolari. La scheda `.MODEL` dei diodi Schottky discende da quella dei diodi p/n, con la sola modifica della corrente IS, che deve essere superiore a quella del diodo p/n di parecchi ordini di grandezza per presentare cadute di tensione dell'ordine di 0.4-0.5 V, e del tempo di transito che deve essere nullo per eliminare la capacità di diffusione dovuta ai portatori minoritari (che non sono presenti nel diodo Schottky).

Tabella B.3 Parametri per la simulazione SPICE di diodi

parametro	*nome*	*unità*	*espressione*
resistenza ohmica	RS	ohm	R_S
tempo di transito	TT	s	τ_F
capacità di giunzione	CJO	F	C_{J0}
potenziale di barriera	VJ	V	ϕ_0
coefficiente per la capacità CJO	M	–	m
corrente di saturazione inversa	IS	A	I_S

Qui di seguito viene indicato il file `DISPO.LIB` contenente le schede `.MODEL` dei diversi dispositivi:

DISPO.LIB

```
.MODEL  MN   NMOS (LEVEL=1 VTO=0.8 TOX=0.02e-6 NSUB=1e16
+        LAMBDA=0.02 GAMMA=0.37 PB=0.87 PHI=0.65 UO=600
+        CGSO=2e-10 CGDO=2e-10 CGBO=2e-10 IS=1e-15
+        RSH=10 CJ=3e-4 CJSW=3e-10 MJ=0.5 MJSW=0.33)
*
*
.MODEL  MND  NMOS (LEVEL=1 VTO=-3 TOX=0.02e-6 NSUB=1e16
```

```
+       LAMBDA=0.02 GAMMA=0.37 PHI=0.65
+       IS=1e-15 UO=600 PB=0.86
+       CGSO=2e-10 CGDO=2e-10 CGBO=2e-10 UCRIT=5e4
+       RSH=10 CJ=3e-4 CJSW=3e-10 MJ=.5 MJSW=0.33)
*
*
.MODEL  MP   PMOS (LEVEL=1 VTO= -0.8 TOX=0.02e-6 NSUB=1e16
+       LAMBDA=0.02 GAMMA=0.37 PHI=0.65
+       IS=1e-15 UO=600 PB=0.86
+       CGSO=2e-10 CGDO=2e-10 CGBO=2e-10 UCRIT=5e4
+       RSH=10 CJ=3e-4 CJSW=3e-10 MJ=0.5 MJSW=0.33)
*
*
.MODEL  QN1   NPN (BF=50 BR=1 RB=100 RC=10 RE=1 VAF=50
        VAR=50
+       TF=60p TR=10n CJE=90f CJC=45f CJS=100f
+       VJE=1 MJE=0.5 VJC=0.75 MJC=.36 IS=5e-16)
*
*
.MODEL  Qi   NPN (BF=20 BR=0.02 RB=100 RC=10 RE=1 VAF=50 VAR=50
+       TF=60p TR=10n CJE=90f CJC=45f CJS=100f
+       VJE=1 MJE=0.5 VJC=0.75 MJC=.36 IS=5e-16)
*
*
.MODEL  Qnbig   NPN (BF=50 BR=1 RB=50 RC=1 RE=0.3 VAF=50
        VAR=50
+       IKF=0.5 TF=60p TR=10n CJE=270f CJC=135f CJS=300f
+       VJE=1 MJE=0.5 VJC=0.75 MJC=.36 IS=1.5e-15)
*
*
.MODEL  DX   D (IS=5e-16 Rs=10 Cjo=0.3p M=0.5 VJ=0.69
        TT=0.2n)
*
*
.MODEL  Dschttky D (IS=3e-10, RS=150, CJO=0.045p, M=0.5,
        VJ=0.69, TT=0)
```

C

Analisi SPICE di circuiti digitali

C.1 Premessa

In questa appendice sono presentate le analisi di alcuni dei circuiti logici discussi nel testo, partendo dai rispettivi file .CIR che li descrivono, e riportando in figura gli schemi elettrici con l'indicazione dei nodi utilizzati nei file stessi. Le analisi riguardano il comportamento sia statico che dinamico delle principali porte logiche elementari, e di qualche circuito combinatorio e sequenziale, come i latch e le celle di memoria per RAM.

Le prime analisi verranno anche utilizzate per una esemplificazione delle possibili articolazioni dei file `.CIR`, e dei principali comandi. Le analisi susseguenti verranno presentate in maniera molto più succinta, ritenendo ormai il lettore familiarizzato con le sintassi dei principali comandi.

C.2 Porte logiche NMOS

Per le analisi di questa famiglia logica faremo riferimento allo schema elettrico dell'invertitore elementare NMOS con dispositivo NMOS a svuotamento come carico attivo, riportato in Figura C.1, in cui sono indicati i numeri dei nodi a cui si fa riferimento nel file `NMOS1.CIR` (il nodo 0 di massa non verrà più indicato negli schemi successivi).

C.2.1 Analisi statica dell'invertitore NMOS

Il file `NMOS1.CIR` effettua l'analisi statica dell'invertitore NMOS, da cui si può ottenere in particolare la caratteristica di trasferimento e i livelli logici nominali. L'analisi è effettuata per due casi del rapporto K_R: 4 e 16, in modo da evidenziare l'effetto del valore di K_R sulle prestazioni statiche del circuito stesso. Si utilizzerà

questo primo file `.CIR` per esemplificare l'organizzazione del file stesso (per tale scopo, solo in questo file si sono numerate le righe per identificarle nei commenti che seguono, mentre nel file reale queste non vanno numerate).

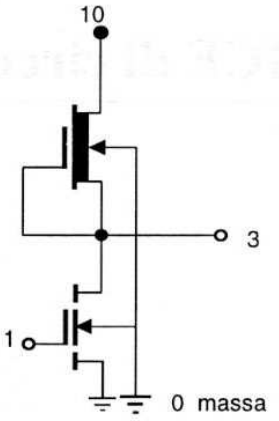


Figura C.1 Circuito dell'invertitore NMOS analizzato con il file `MOSD1.CIR`

Il file `MOSD1.CIR` viene scritto nel modo seguente:

```
1    MOSD1.CIR
2    * Caratteristica trasferimento invertitore NMOS E-D Kr=4
3    .opt nomod nopage reltol=.001
4    .lib dispo.lib
5    ma 3 1 0 0 mn l=1u w=2u
6    ml 10 3 3 0 mnd l=2u w=1u
7    vdd 10 0 dc 5v
8    vi 1 0 dc 2.5
9    .dc lin vi 0 5 0.05
10   .probe
11   .end
12   Kr=16
13   * Caratteristica trasfer. invertitore Kr = 16
14   .opt nomod nopage reltol=.001
15   .lib dispo.lib
16   ma 3 1 0 0 mn l=1u w=4u
17   ml 10 3 3 0 mnd l=4u w=1u
18   vdd 10 0 dc 5v
19   vi 1 0 dc 3
20   .dc lin vi 0 5 0.05
21   .probe
22   .end
```

Il file `MOSD1.CIR` è in realtà costituito da due file in sequenza, il primo (righe da 1 a 11) analizza il caso con $K_R = 4$, il secondo (righe da 12 a 21) quello con $K_R = 16$.

- la riga 1 contiene il titolo del file
- la riga 2 contiene un commento (inizia con un *)
- la riga 3 definisce le opzioni e le tolleranze
- la riga 4 chiama la libreria di dispositivi `DISPO.LIB` che contiene le schede `.MODEL` dei NMOS ad arricchimento MN e a svuotamento MND, che vengono chiamati in questo circuito *ma* e *ml*, e per i quali vengono specificati i parametri contenuti nella parentesi
- le righe 5 e 6 definiscono i nodi a cui vengono connessi i due MOS e i parametri geometrici (in questo caso *L* e *W*). I nodi vanno definiti in questa sequenza: drain gate source body: si verifichi l'esatta connessione dei due dispositivi nel circuito
- la riga 7 definisce i nodi (10 0) a cui viene applicato il generatore di tensione chiamato *vdd* (tensione continua, 5 V)
- la riga 8 definisce i nodi (1 0) a cui viene applicata la tensione di ingresso *vi* (dc, 2.5 V)
- la riga 9 definisce il tipo di analisi effettuata: si effettua un'analisi in continua (dc) facendo variare in maniera lineare (lin) il valore di *vi* da 0 a 5 V con passo 0.05 V
- la riga 10 attiva il modulo `.PROBE` che permette la visualizzazione dei risultati
- la riga 11 termina il primo caso.

Il caso con $K_R = 16$ è ottenuto ricopiando le linee precedenti (inserendo un nuovo titolo alla linea 12 e un diverso commento alla linea 13) e modificando solo le linee 16 e 17 rispetto al caso precedente.

L'analisi viene effettuata in sequenza, attivando il modulo "*analysis*" del menu PS; successivamente viene automaticamente attivato il modulo `.PROBE` che permette la visualizzazione di tutte le grandezze elettriche al variare dell'ingresso *Vi*.

Mediante PROBE è possibile visualizzare contemporaneamente i risultati delle due analisi o uno alla volta: per visualizzare la caratteristica di trasferimento si sceglie come variabile per l'asse *Y* la grandezza *V*(3), che rappresenta la tensione nel nodo 3 (uscita).

C.2.2 Analisi dinamica dell'invertitore NMOS

Lo schema elettrico per l'analisi è quello di un invertitore NMOS E-D caricato da un uguale invertitore E-D, ed è riportato in Figura C.2.

Il file `SPICE` che descrive questo circuito è chiamato `MOSD2.CIR`, ed effettua anche in questo caso due analisi, relative ai casi con due differenti valori di K_R: $K_R = 4$ e $K_R = 16$. In questo file sono state specificate le aree di source e di drain dei MOS in quanto nell'analisi dinamica occorre definire le capacità di source e di drain in base alle aree relative. Il generatore di tensione Vi è di tipo

impulsivo (PULSE). L'analisi è quella in transitorio (.TRAN) e dura 30 ns con passo massimo di 0.1 ns.

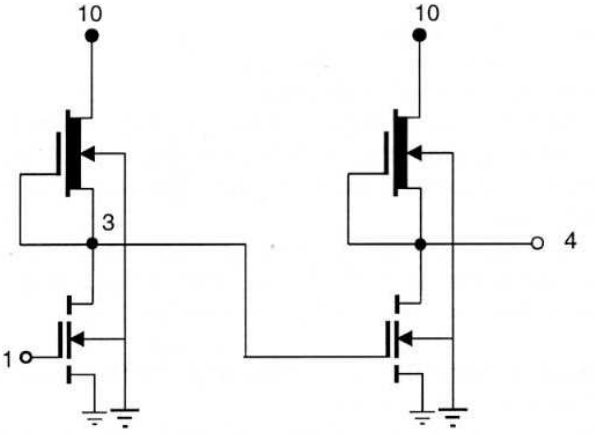


Figura C.2 Schema elettrico relativo al file MOSD2.CIR

```
MOSD2
* Transitorio invertitore NMOS Kr=4
.opt nomod nopage reltol=.001
.lib dispo.lib
ma 3 1 0 0 mn l=1u w=2u as=14e-12 ad=14e-12
ml 10 3 3 0 mnd l=2u w=1u as=1e-12 ad=14e-12
ma2 4 3 0 0 mn l=1u w=2u as=14e-12 ad=14e-12
ml2 10 4 4 0 mnd l=2u w=1u as=1e-12 ad=14e-12
vdd 10 0 dc 5v
vi 1 0 pulse(.2 5 1n 1n 1n 8n 30n)
.tran .1n 30n
.probe
.end
KR=16
* transitorio NMOS Kr=16
.opt nomod nopage reltol=.001
.lib dispo.lib
ma 3 1 0 0 mn l=1u w=4u as=28e-12 ad=28e-12
ml 10 3 3 0 mnd l=4u w=1u as=1e-12 ad=14e-12
ma2 4 3 0 0 mn l=1u w=4u as=28e-12 ad=28e-12
ml2 10 4 4 0 mnd l=4u w=1u as=1e-12 ad=14e-12
vdd 10 0 dc 5v
vi 1 0 pulse(.2 5 1n 1n 1n 8n 30n)
.tran .1n 30n
.probe
.end
```

C.3 Porte logiche CMOS

C.3.1 Analisi statica dell'invertitore CMOS

Il circuito per l'analisi statica dell'invertitore CMOS è quello riportato nella Figura C.3, con la numerazione dei nodi utilizzata nel file `CMOS1.CIR`; questo viene riportato di seguito:

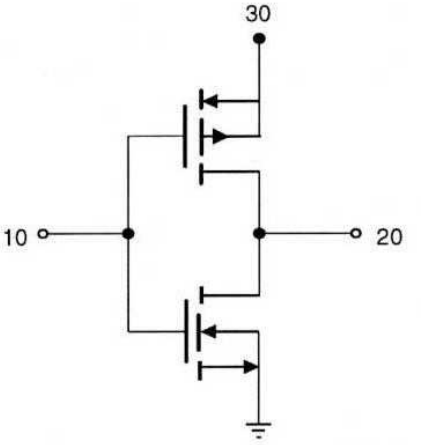


Figura C.3 Circuito dell'invertitore CMOS analizzato con il file `CMOS1.CIR`

```
CMOS1
* caratteristica trasferimento invertitore CMOS
.opt nomod nopage reltol=.001
.lib dispo.lib
* descrizione del circuito
mn1 20 10  0  0  mn l=1u w=2u
mp1 20 10 30 30  mp l=1u w=5u
vdd 30 0 dc 5v
vi  10 0  2.5
.dc vi 0 5 .05
.probe
.end
```

Si noti che nel file `.CIR` il MOS `mp1` è un MOS a canale P, descritto nella libreria `DISPO.LIB` con la scheda `.MODEL MP`.

C.3.2 Analisi dinamica dell'invertitore CMOS

Il circuito prevede un invertitore CMOS caricato da un secondo invertitore, come nello schema di Figura C.4. Il circuito è descritto dal file `CMOS2.CIR`:

```
CMOS2
* transitorio invertitore CMOS
.opt nomod nopage reltol=.001
.lib dispo.lib
* descrizione del circuito
mn1 20 10  0  0  mn l=1.2u w=1.8u as=14e-12 ad=14e-12
+                                  ps=15.6e-6 pd=15.6e-6
mp1 20 10 30 30  mp l=1.2u w=4.8u as=20.2e-12 ad=17.3e-12
+                                    ps=18e-6 pd=17e-6
mn12 40 20  0  0  mn l=1.2u w=1.8u as=14e-12 ad=14e-12
+                                  ps=15.6e-6 pd=15.6e-6
mp2 40 20 30 30  mp l=1.2u w=4.8u as=20.2e-12 ad=17.3e-12
+                                    ps=18e-6 pd=17e-6
vdd 30 0 dc 5v
vin 10 0 PWL(0n,0V 3n,0V 3.5ns,5V 10ns,5V 10.5ns,0v 15n, 0v)
.tran 0.2n 15n
.probe
.end
```

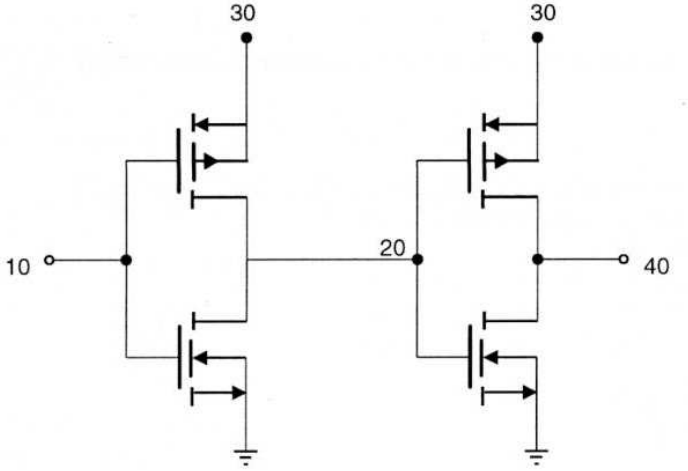


Figura C.4 Circuito dell'invertitore CMOS analizzato con il file `CMOS2.CIR`

In questo file viene effettuata l'analisi dinamica di un invertitore CMOS con $K_N \cong K_P$ e con $W_N = 1.8$ µm, $W_P = 4.8$ µm, $L_N = L_P = 1.2$ µm. Nelle linee relative ai MOS *mn* e *mp* sono stati indicati anche i perimetri delle regioni di source e drain, per valutare le capacità perimetriche C_{JW}. Il generatore di segnale Vin è descritto come una forma d'onda a spezzate lineari (*piecewise linear*).

C.4 Porte logiche TTL

Le esercitazioni seguenti trattano delle porte logiche TTL. Faremo sempre riferimento all'invertitore TTL e confronteremo il comportamento nelle versioni standard e con reti di pull-up e pull-down

C.4.1 Analisi statica

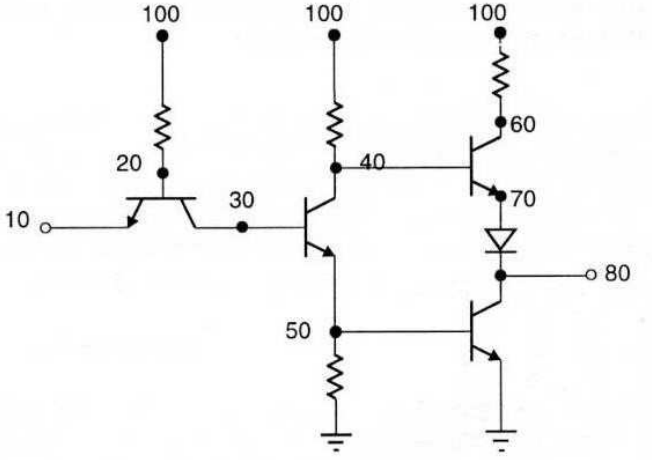


Figura C.5 Schema dell'invertitore TTL standard analizzato nel file `TTL1.CIR`

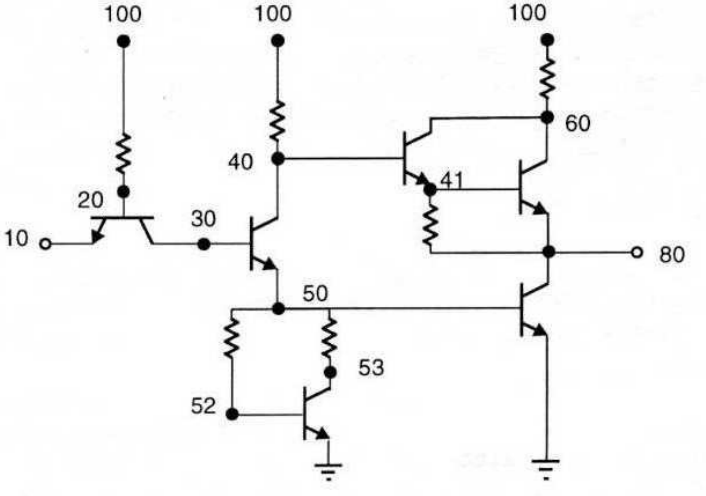


Figura C.6 Schema dell'invertitore TTL con reti di pull-up e pull-down analizzato nel file `TTL1.CIR`

Per l'analisi statica dell'invertitore TTL faremo riferimento allo schema riportato in Figura C.5 per l'invertitore TTL standard e in Figura C.6 per quello con le reti di pull-up e pull-down, indicando sugli schemi relativi la numerazione dei nodi adottata nel file .CIR per l'analisi con SPICE.

Il file per l'analisi statica è chiamato `TTL1.CIR`; esso contiene l'analisi sia dell'invertitore di Figura C.6 che di quello standard di Figura C.5. In entrambi i casi è stata considerata una resistenza di carico di 50 kΩ, per semplificare il circuito; questo carico non rappresenta correttamente quello di una porta TTL a valle, che ha un comportamento non lineare: si provi ad identificare la rete più semplice da porre all'uscita per effettuare una simulazione più corretta.

```
TTL1.CIR
*Caratteristica di trasferimento TTL con pull-up e pull-down
.opt nomod nopage reltol=.001
.width out=80
.temp 25
.lib dispo.lib
* descrizione del circuito
qi    30   20   10   0   Qi
qd    40   30   50   0   QN1
q3b   60   40   41   0   QN1
q4b   53   52   0    0   QN1
q3    60   41   80   0   QNbig
q4    80   50   0    0   QNbig
Rb    20  100   4k
Rc    40  100   1.6k
R     41   80    3.5K
R1    50   52   500
R2    50   53   250
Rt    60  100   120
Vcc   100  0    5V
RL    80   0    50k
V1    10   0    .2V
.dc   v1  0  5v  .05v
.probe
.end
* Caratteristica trasferimento TTL standard
.opt nomod nopage reltol=.001
.width out=80
.temp 25
.lib dispo.lib
* descrizione del circuito
qi    30   20   10   0   Qi
qd    40   30   50   0   QN1
q3    60   40   70   0   QNbig
q4    80   50   0    0   QNbig
```

```
d1    70   80   DX
Rb    20  100   4k
Rc    40  100   1.6k
Re    50   0    1K
Rt    60  100   120
Vcc   100  0    5V
RL    80   0    50k
V1    10   0    .2v
.dc   V1  0 5  .05V
.probe
.end
```

Si può notare una novità rispetto ai file `.CIR` delle analisi precedenti: sono qui definiti tre tipi di transistori NPN di cui sono riportate le schede `.MODEL` nella libreria `DISPO.LIB` (vedi Appendice B). Il transistore QI è caratterizzato da un β_F e un β_R più bassi di quelli dei NPN usuali, il transistore QN1 è un NPN usuale; il transistore QNbig rappresenta un transistore con area maggiore e quindi con ridotti valori delle resistenze interne di emettitore, base e collettore.

C.4.2 Analisi dinamica

L'analisi dinamica relativa a questa esercitazione prevede il confronto tra la porta TTL standard e quella con reti di pull-up e pull-down, porte per le quali si sono già valutate le caratteristiche di trasferimento.

L'analisi dinamica richiederebbe di caricare la porta in esame con un'altra porta uguale; tuttavia nel caso dell'invertitore TTL con reti di pull-up e pull-down questo richiederebbe l'analisi di un circuito con 12 dispositivi e quindi troppo oneroso per la versione Demo di Spice. Si analizzerà quindi l'invertitore caricato da un semplice parallelo di capacità e resistenza che simula l'ingresso dell'invertitore di carico.

Il file che descrive il circuito fa quindi riferimento agli schemi degli invertitori di Figura C.5 e C.6 rispettivamente, entrambi caricati da un parallelo CL = 0.1 pF e RL = 10 kΩ. Il file, chiamato `TTL2.CIR`, per l'analisi dei due circuiti TTL, è il seguente:

```
TTL2.CIR
*transitorio TTL con rete pull-up e pull-down
.opt nomod nopage reltol=.001
.width out=80
.temp 25
.lib dispo.lib
* descrizione del circuito
qi    30   20   10   0   Qi
qd    40   30   50   0   QN1
q3b   60   40   41   0   QN1
q4b   53   52   0    0   QN1
```

```
q3    60   41   80   0   QNbig
q4    80   50   0    0   QNbig
Rb    20  100   4k
Rc    40  100   1.6k
R     41    0   3.5K
R1    50   52   500
R2    50   53   250
Rt    60  100   120
Vcc   100  0    5V
CL    80   0    .1p
RL    80   0    10k
V1    10   0    pulse(0.2V 3.8V 5ns .5ns .5ns 20ns 50ns)
.tran .2ns 50ns
.probe
.end
* transitorio TTL standard
.opt nomod nopage reltol=.001
.width out=80
.temp 25
.lib dispo.lib
* descrizione del circuito
qi    30   20   10   0   Qi
qd    40   30   50   0   QN1
q3    60   40   70   0   QNbig
q4    80   50   0    0   QNbig
Rb    20  100   4k
Rc    40  100   1.6k
R     50    0   1k
Rt    60  100   120
d1    70   80   DX
Vcc   100  0    5V
CL    80   0    1p
RL    80   0    10k
V1    10   0    pulse(0.2V 3.8V 5ns .5ns .5ns 20ns 50ns)
.tran .2ns 50ns
.probe
.end
```

C.5 Porte logiche ECL

C.5.1 Analisi statica

Per la porta ECL faremo riferimento allo schema elettrico con l'indicazione dei nodi riportato in Figura C.7; l'analisi statica, effettuata nel file `ECL1.CIR`, permette di ottenere la caratteristica di trasferimento a tre differenti temperature: 0°,

30° e 60°, in modo da valutare l'effetto sulla variazione dei livelli logici effettuato dalla rete di compensazione termica.

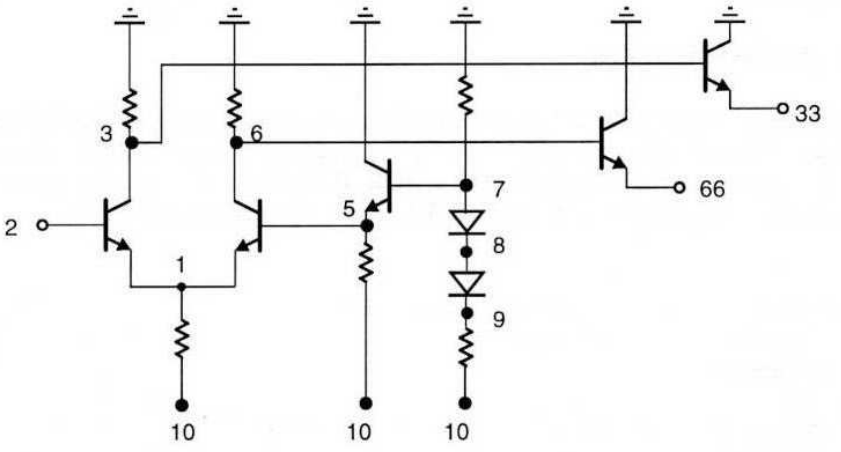


Figura C.7 Circuito della porta ECL analizzato nel file `ECL1.CIR`

```
ECL1
*caratteristica di trasferimento al variare della temperatura
.OPT nomod nopage reltol=.001
.width out=80
.lib dispo.lib
RC2 0 6  245
RC1 0 3  230
RE  1 10 780
q1  3 2 1 0 qn1
q2  6 5 1 0 qn1
qr  0 7 5 0 qn1
RR  5 10 6k
R1  0 7 900
R2  9 10 6k
d1  7 8 dx
d2  8 9 dx
q3  0 3 33 0 qn1
q4  0 6 66 0 qn1
R3  33 10 5k
R6  66 10 5k
VE  10 0 -5.2v
V1  2 0 -1v
.temp 0 30 60
.dc v1 0 -5.2 .05V
.probe
.end
```

L'analisi è effettuata in forma parametrica, utilizzando il comando per analisi multiple TEMP A B C. Le soluzioni relative ai tre casi sono presentate in uscita in un singolo grafico.

C.5.2 Analisi dinamica

Per l'analisi dinamica, il file `ECL2.CIR` analizza il comportamento con due diversi valori della capacità di carico, pari a 0.05 pF e 1 pF, per valutare l'effetto di una capacità di carico elevata sulle uscite OR e NOR. Il circuito a cui fare riferimento è ancora quello di Figura C.7, con un carico RLCL per ognuna delle due uscite.

```
ECL2
*transitorio ECL con carico CL=0.05pF
.OPT nomod nopage reltol=.001
.width out=80
.temp 25
.lib dispo.lib
RC2 0 6  245
RC1 0 3  230
RE  1 10 780
q1 3 2 1 0 qn1
q2 6 5 1 0 qn1
qr 0 7 5 0 qn1
RR 5 10 6k
R1 0 7 900
R2 9 10 6k
d1 7 8 dx
d2 8 9 dx
q3 0 3 33 0 qn1
q4 0 6 66 0 qn1
R3 33 10 5K
R6 66 10 5K
CLN 33 0 .05p
CLO 66 0 .05p
VE 10 0 -5.2v
v1 2 0 PWL ( 0n,-1.7v 5n,-1.7v 5.5n,-.7v 14.5n,-.7V 15n,-1.7v)
.tran .1n 25n
.probe
.end
* transitorio con carico  CL=1pF
.OPT nomod nopage reltol=.001
.width out=80
.temp 25
.lib dispo.lib
RC2 0 6  245
RC1 0 3  230
```

```
RE  1 10 780
q1 3 2 1 0 qn1
q2 6 5 1 0 qn1
qr 0 7 5 0 qn1
RR 5 10 6k
R1 0 7 900
R2 9 10 6k
d1 7 8 dx
d2 8 9 dx
q3 0 3 33 0 qn1
q4 0 6 66 0 qn1
R3 33 10 5K
R6 66 10 5K
CLN 33 0 1p
CLO 66 0 1p
VE 10 0 -5.2v
v1 2 0 PWL ( 0n,-1.7v 5n,-1.7v 5.5n,-.7v 14.5n,-.7V 15n,-1.7v)
.tran .1n 25n
.probe
.end
```

C.6 Invertitore BiCMOS

C.6.1 Analisi statica

L'analisi statica per l'invertitore BiCMOS, il cui schema elettrico è riportato nella Figura C.8, è effettuata nel file `BICMOS1.CIR`.

```
BICMOS1
* caratteristica di trasferimento invertitore Bicmos
.opt nomod nopage reltol=.001
.lib dispo.lib
* descrizione del circuito
mn1 2 1 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp1 2 1 3 3 mp l=1u w=5u as=30e-12 ad=30e-12
mn2 5 1 4 0 mn l=1u w=2u as=10e-12 ad=10e-12
mn3 4 2 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
Qb 3 2 5  QN1
Qa 5 4 0  QN1
vdd 3 0 5v
v1 1 0 1v
.dc V1 0 5 .05
.probe
.end
```

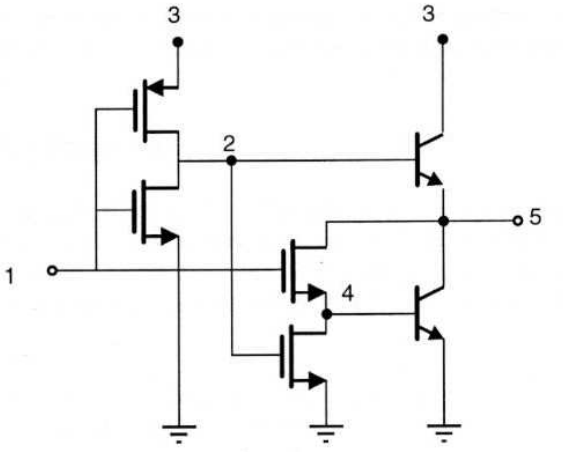


Figura C.8 Schema elettrico dell'invertitore BiCMOS

C.6.2 Analisi dinamica

L'analisi dinamica utilizza il file `BICMOS2.CIR` e analizza il comportamento con due diversi valori della capacità di carico, rispettivamente di 0.5 pF e 5 pF, per valutare l'effetto di una capacità di carico elevata in uscita. Il circuito a cui fare riferimento è ancora quello di Figura C.8, con un carico CL al nodo 5.

```
BICMOS2
* transitorio invertitore Bicmos CL=0.5pF
.opt nomod nopage reltol=.001
.lib dispo.lib
* descrizione del circuito
mn1 2 1 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp1 2 1 3 3 mp l=1u w=5u as=30e-12 ad=30e-12
mn2 5 1 4 0 mn l=1u w=2u as=10e-12 ad=10e-12
mn3 4 2 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
Qb 3 2 5  QN1
Qa 5 4 0  QN1
Cl 5 0 0.5pf
vdd 3 0 5v
vin 1 0 PWL(0n,5v 1n,5v 1.5n,0v 5n,0V 5.5n,5V 10n,5v
10.5n,0v 15n,0v)
.tran 0.1n 15n
.probe
.end
* transitorio CL=5pF
.opt nomod nopage reltol=.001
```

```
.lib dispo.lib
* descrizione del circuito
mn1 2 1 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp1 2 1 3 3 mp l=1u w=5u as=30e-12 ad=30e-12
mn2 5 1 4 0 mn l=1u w=2u as=10e-12 ad=10e-12
mn3 4 2 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
Qb 3 2 5  QN1
Qa 5 4 0  QN1
Cl 5 0 5pf
vdd 3 0 5v
vin  1  0  PWL(0n,5v  1n,5v  1.5n,0v  5n,0V  5.5n,5V  10n,5v
10.5n,0v 15n,0v)
.tran 0.1n 15n
.probe
.end
```

C.7 Circuiti sequenziali

C.7.1 Latch SR con porte NOR

Il circuito elettrico del latch SR con porte NOR, che è analizzato con il file `SRCMOS.CIR` è riportato in Figura C.9.

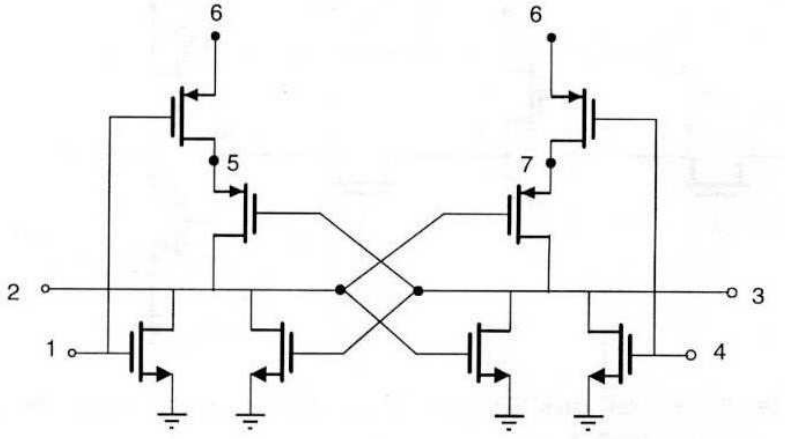


Figura C.9 Schema elettrico del latch SR a porte NOR analizzato nel file `SRCMOS.CIR`

Il file che effettua l'analisi dinamica del circuito è il file `SRCMOS.CIR`; viene applicato prima un segnale di reset e poi uno di set e infine un altro segnale di reset, rispettivamente ai due ingressi S e R del latch:

```
SRCMOS.CIR
*latch SR CMOS con porte NOR
.opt nomod nopage reltol=.001
.lib dispo.lib
m1 2 1 0 0 mn l=2u w=4u as=4e-11 ad=4e-11
mp1 2 3 5 6 mp l=2u w=10u as=1e-10 ad=1e-10
mp2 5 1 6 6 mp l=2u w=10u as=1e-10 ad=1e-10
m2 2 3 0 0 mn l=2u w=4u as=4e-11 ad=4e-11
m3 3 2 0 0 mn l=2u w=4u as=4e-11 ad=4e-11
m4 3 4 0 0 mn l=2u w=4u as=4e-11 ad=4e-11
mp3 3 2 7 6 mp l=2u w=10u as=1e-10 ad=1e-10
mp4 7 4 6 6 mp l=2u w=10u as=1e-10 ad=1e-10
vdd 6 0 dc 5v
vs 1 0 pwl(0,0 5n,0 5.5n,5 8n,5 8.5n,0 15n,0)
vr 4 0 pwl(0,5 2.5n,5 3n,0 12n,0 12.5n,5 15n,5)
.tran .05n 15n
.probe
.end
```

C.7.2 Cella dinamica per registro a scorrimento

Nel file REGISTRO.CIR viene analizzato il comportamento dinamico di un latch di tipo D per registro a scorrimento; lo schema elettrico è riportato in Figura C.10.

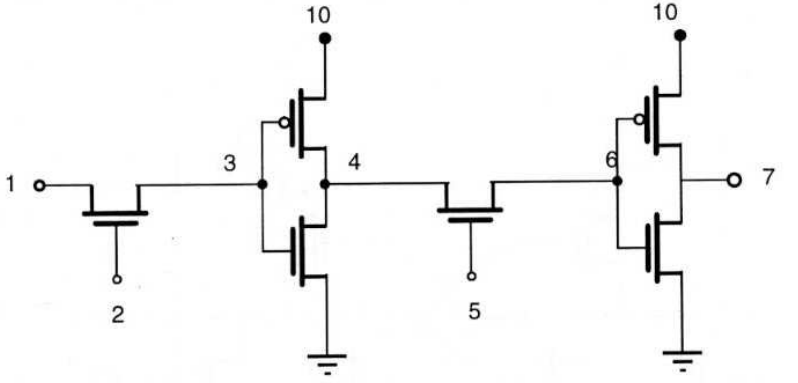


Figura C.10 Circuito del latch dinamico D per registro a scorrimento del file REGISTRO.CIR

```
REGISTRO.CIR
*Registro dinamico con porte NMOS
.opt nomod nopage reltol=.001
.lib dispo.lib
m1 1 2 3 0 mn l=1u w=2u as=10e-12 ad=10e-12
```

```
mn3 4 3 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp3 4 3 10 10  mp l=1u w=5u as=30e-12 ad=30e-12
m2 4 5 6 0 mn l=1u w=2u as=10e-12 ad=10e-12
mn4 7 6 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp4 7 6 10 10  mp l=1u w=5u as=30e-12 ad=30e-12
vdd 10 0 dc 5v
v1 2 0 pulse(0 5 2n .1n .1n 4n 10n)
v2 5 0 pulse(0 5 7n .1n .1n 4n 10n)
Vi 1 0 pulse(0 5 10n .1n .1n 7n 40n)
.tran .05n 40n
.probe
.end
```

C.8 Celle di memoria

C.8.1 Cella di memoria NMOS

Il file `MEMNMOS.CIR` analizza il comportamento dinamico, in fase di lettura, di una cella di memoria NMOS con carico in polisilicio; il circuito analizzato è riportato in Figura C.11.

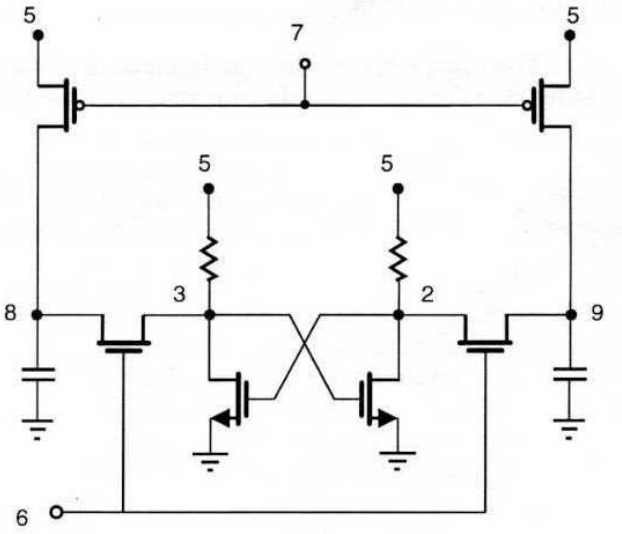


Figura C.11 Schema elettrico della cella NMOS analizzata nel file `MEMNMOS.CIR`

```
MEMNMOS.CIR
*cella memoria NMOS
.opt nomod nopage reltol=.001
.lib dispo.lib
m1 2 3 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
m2 3 2 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mT1 8 6 3 0 mn l=4u w=2u as=10e-12 ad=10e-12
mT2 9 6 2 0 mn l=4u w=2u as=10e-12 ad=10e-12
mp3 8 7 10 10 mp l=1u w=10u as=30e-12 ad=30e-12
mp4 9 7 10 10 mp l=1u w=10u as=30e-12 ad=30e-12
CL1 9 0 1p
CL2 8 0 1p
RP1 5 2 10000k
RP2 5 3 10000k
Vline 10 0 5v
vdd 5 0 dc 5v
vl 7 0 pwl(0,5 50n,5 51n,0 89n,0 90n,5 200n,5)
vd 6 0 pwl(0,0 99n,0 100n,5 170n,5 171n,0 200n,0)
.ic v(8)=0 v(9)=0 V(3)=5V V(2)=0.05V
.tran .2n 200n
.probe
.end
```

C.8.2 Cella di memoria CMOS

Il file `MEMCMOS.CIR` analizza il comportamento dinamico di fase di lettura di una cella di memoria CMOS; il circuito analizzato è riportato in Figura C.12.

```
MEMCMOS.CIR
*cella memoria CMOS
.opt nomod nopage reltol=.001
.lib dispo.lib
* descrizione del circuito
mn1 2 3 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp1 2 3 5 5 mp l=1u w=2u as=20e-12 ad=20e-12
mn2 3 2 0 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp2 3 2 5 5 mp l=1u w=2u as=20e-12 ad=20e-12
mt1 8 6 3 0 mn l=2u w=2u as=10e-12 ad=10e-12
mt2 9 6 2 0 mn l=2u w=2u as=10e-12 ad=10e-12
mp3 8 7 5 5 mp l=1u w=10u as=30e-12 ad=30e-12
mp4 9 7 5 5 mp l=1u w=10u as=30e-12 ad=30e-12
cl1 9 0 1p
cl2 8 0 1p
vdd 5 0 dc 5v
vl 7 0 pwl(0,5 49n,5 50n,0 89n,0 90n,5 200n,5)
vd 6 0 pwl(0,0 99n,0 100n,5 160n,5 161n,0 200n,0)
```

```
.ic v(8)=0 V(9)=0 V(3)=5V V(2)=0
.tran .2n 200n
.probe
.end
```

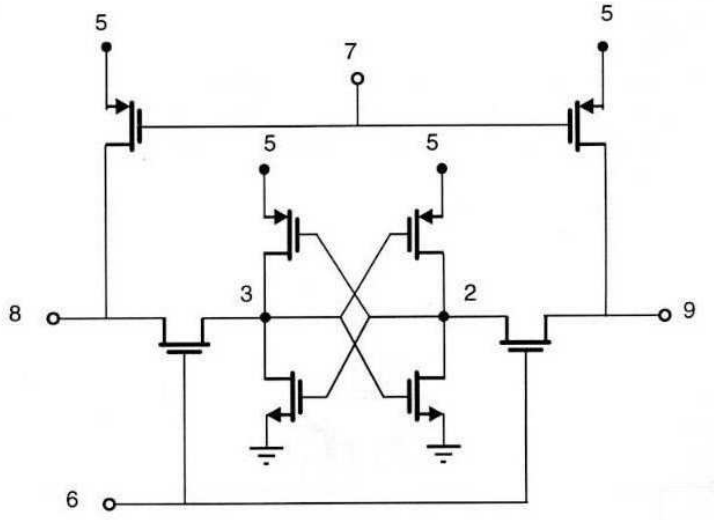


Figura C.12 Circuito della cella di memoria CMOS analizzato dal file `MEMCMOS.CIR`

C.8.3 Lettura di una cella ad un transistore

Il file `SENSE1T.CIR` analizza l'operazione di lettura di un 1 logico memorizzato in una cella di memoria a 1 transistore, mediante un sense amplifier di tipo bistabile. Il circuito semplificato a cui si riferisce l'analisi è quello riportato in Figura C.13. Si è assunta una capacità CM per la cella di 0.02 pF, e si sono assunti come valori iniziali di tensione, 3.6 V per CM1, 2.5 V per CM2, 2.5 V per i nodi 2 e 3. L'effetto della cella fittizia è simulato dall'apertura della porta connessa alla cella CM2, che è caricata a 2.5 V. Le due metà della bit line sono simulate con due capacità CL1 e CL2 pari a 0.5 pF.

```
SENSE1T.CIR
*lettura di cella a un transistore
.opt nomod nopage reltol=.001
.lib dispo.lib
* descrizione del circuito
mn1 2 3 11 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp1 2 3 5 5 mp l=1u w=5u as=30e-12 ad=30e-12
mn2 3 2 11 0 mn l=1u w=2u as=10e-12 ad=10e-12
mp2 3 2 5 5 mp l=1u w=5u as=30e-12 ad=30e-12
mt1 8 6 3 0 mn l=1u w=1u as=10e-12 ad=10e-12
```

```
mt2 9 6 2 0 mn l=1u w=1u as=10E-12 ad=10E-12
ms 11 7 0 0 mn l=1u w=4u as=30e-12 ad=30e-12
CL1 2 0 0.5p
CL2 3 0 0.5p
vdd 5 0 dc 5v
CM1 8 0 0.02p
CM2 9 0 0.02p
vd 6 0 pwl(0,0 2n,0 3n,5 30n,5 31n,0 40n,0)
vs 7 0 pwl(0,0 9n,0 10n,5 30n,5 31n,0 40n,0)
.ic V(3)=2.5V V(2)=2.5v V(8)=3.6V V(9)=2.5V
.tran .05n 40n
.probe
.end
```

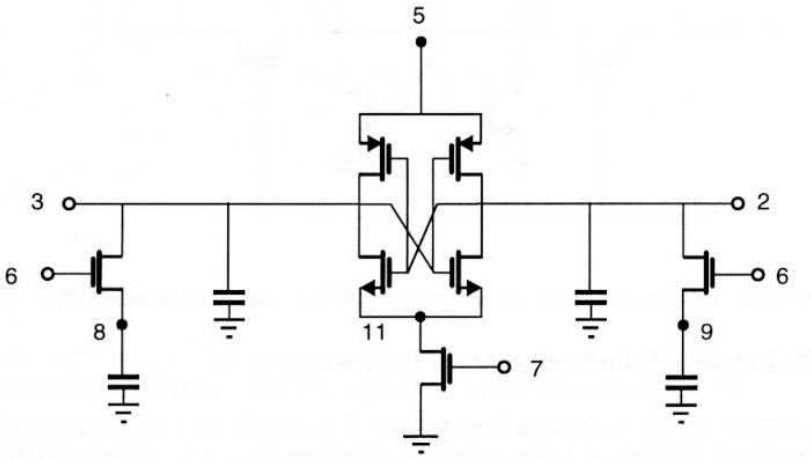


Figura C.13 Circuito di lettura della cella a 1 transistore analizzato nel file `SENSE1T.CIR`