F. BONANI
S. DONATI GUERRIERI
G. MASERA
G. PICCININI

Dispositivi e tecnologie elettroniche

CLUT

C.L.U.T. Editrice S.c.r.l.
C.so Duca degli Abruzzi **24** 10129 Torino
ESEMPLARE FUORI **COMM**ERCIO PER IL DEPOSITO LEGALE AGLI EFFETTI DELLA LEGGE 15/04/04, N° 106 COME DA D.P.R. 252 DEL 03/05/06 ATR. 10, C. 2

F. BONANI
S. DONATI GUERRIERI
G. MASERA
G. PICCININI

Dispositivi e tecnologie elettroniche

CLUT

La messa a punto di un libro è un'operazione complessa ed articolata, che necessita di studi, progettualità grafica, nonché di numerosi controlli di testo, immagine, stili grafici e di stampa. È praticamente impossibile pubblicare un libro scevro da errori. La C.L.U.T. ringrazia sin d'ora i lettori che vorranno segnalare all'indirizzo dell'Editore eventuali errori riscontrati nella lettura del libro.

Classe LaTeX : clut.cls

Stampato in Italia da STAMPATRE - Torino

ISBN 978-88-7992-252-4

Edizioni C.L.U.T. - Torino
Corso Duca degli Abruzzi, 24 - 10129 Torino
Tel. 011.5647980 - Fax 011.542192
e-mail: clut@inrete.it - www.clut.it

Presentazione

Il panorama librario universitario italiano è già ricco di ottimi testi che riguardano i dispositivi elettronici a semiconduttore e le relative tecnologie costruttive. I contenuti di questi testi tuttavia risultano spesso non adeguati rispetto alla nuova organizzazione didattica nata con le lauree triennali, che richiedono una trattazione dei dispositivi elettronici meno approfondita e più immediatamente finalizzata alle applicazioni circuitali.

In questo libro i dispositivi elettronici sono trattati a un livello di approfondimento intermedio fra quello proposto nella maggior parte dei testi classici sull'argomento, originariamente concepiti per le lauree quinquennali, e quello puramente introduttivo fornito all'interno di alcuni testi di Elettronica Applicata. La trattazione è comunque ampia e i contenuti proposti comprendono i concetti e gli strumenti di base per lo studio dei semiconduttori, l'analisi dei principali dispositivi, la tecnologia costruttiva e alcune applicazioni significative, come gli stadi amplificatori a singolo transistore e le memorie a semiconduttore.

Il testo è quindi adatto per tutti i corsi di laurea triennali in Ingegneria Elettronica, Informatica e delle Telecomunicazioni, oltre che per i corrispondenti corsi di laurea a distanza.

Indice

Capitolo 1

Proprietà elettriche dei semiconduttori

1.1 Introduzione

Le proprietà elettriche dei materiali costituiscono i parametri di base che consentono di studiare e, quindi, di prevedere, le caratteristiche elettriche dei dispositivi utilizzati nei circuiti elettronici, per applicazioni sia analogiche sia digitali. In particolare, poiché nella maggior parte delle applicazioni più comuni i fenomeni di tipo magnetico possono essere ritenuti trascurabili, tra le varie proprietà quella più importante è la *conducibilità elettrica* σ (unità di misura: S/cm – Siemens / cm), in alcune occasioni sostituita dal suo reciproco, la *resistività elettrica* $\rho = 1/\sigma$ (unità di misura: Ω cm – Ohm cm). Qualora ad un materiale di conducibilità elettrica σ venga applicato un campo elettrico esterno $\boldsymbol{\mathcal{E}}$ (unità di misura: V/cm – Volt / cm), esso viene attraversato dalla *densità di corrente di conduzione* $\boldsymbol{J}_{\text{cond}}$ (unità di misura: A/cm^2 – Ampere / cm^2), data da:

$$\boldsymbol{J}_{\text{cond}} = \sigma \boldsymbol{\mathcal{E}}. \tag{1.1}$$

Utilizzando come riferimento la conduzione elettrica, i materiali possono essere classificati in tre categorie:

- *conduttori,* ovvero i materiali che non presentano alcuna opposizione al passaggio della corrente elettrica. Un conduttore *ideale* è caratterizzato dalla proprietà:

$$\rho = 0; \tag{1.2}$$

- *isolanti* o *dielettrici,* ovvero i materiali che, anche sotto l'applicazione di un campo elettrico, non sono attraversati da alcuna corrente di conduzione. Un dielettrico *ideale* è caratterizzato dalla proprietà:

$$\sigma = 0. \tag{1.3}$$

  Naturalmente, in accordo ai principi dell'elettromagnetismo, qualora al materiale isolante venga applicato un campo elettrico tempo-variante $\boldsymbol{\mathcal{E}}(t)$, il dielettrico viene

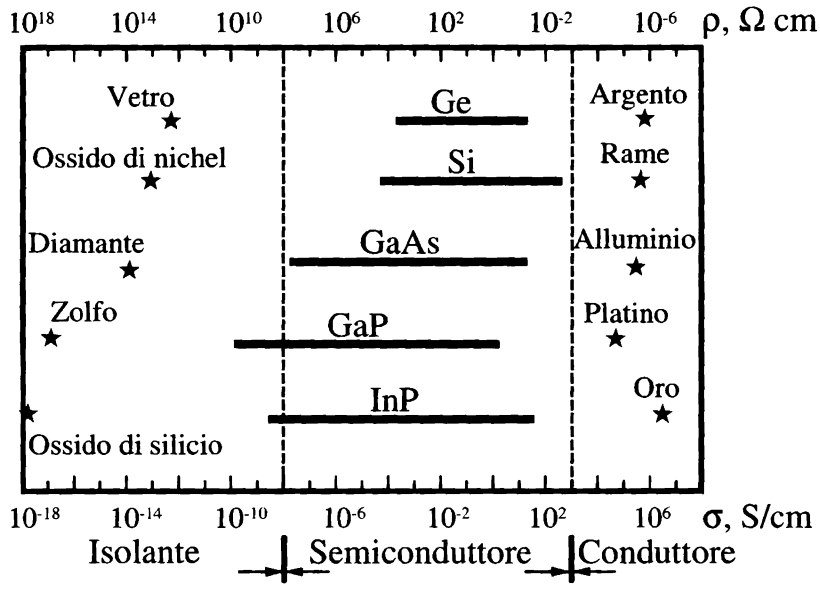


Figura 1.1 Valori di conducibilità (asse inferiore) e resistività elettrica (asse superiore) per alcuni materiali.

attraversato dalla *densità di corrente di spostamento* $\boldsymbol{J}_{\mathrm{sp}}$:

$$\boldsymbol{J}_{\mathrm{sp}} = \frac{\partial \boldsymbol{\mathcal{D}}}{\partial t}, \tag{1.4}$$

dove si è considerato il vettore *spostamento dielettrico* (unità di misura: $\mathrm{C/cm^2}$ – Coulomb / $\mathrm{cm^2}$):

$$\boldsymbol{\mathcal{D}} = \epsilon \boldsymbol{\mathcal{E}} \tag{1.5}$$

essendo ϵ la *costante* o *permittività dielettrica* del materiale (unità di misura: F/cm – Farad / cm). Tipicamente, la costante dielettrica viene espressa come multiplo della *costante dielettrica del vuoto* $\epsilon_0 = 8,854 \times 10^{-14}$ F/cm, definendo la *costante dielettrica relativa* ϵ_{r} secondo la relazione $\epsilon = \epsilon_{\mathrm{r}}\epsilon_0$;

▷ *semiconduttori*, caratterizzati da proprietà intermedie tra isolanti e conduttori.

Nella figura 1.1 sono rappresentati i valori di conducibilità e resistività elettrica dei principali materiali: naturalmente i valori ideali del parametro rappresentati da (1.2) e (1.3) sono solo approssimati dai materiali reali come quelli considerati nel grafico. Con riferimento alle tecnologie relative ai dispositivi elettronici, tra i conduttori i materiali più importanti sono l'oro, il platino, l'alluminio e il rame: la loro resitività ρ è compresa tra 10^{-5} e 10^{-6} Ω cm. Per quanto riguarda gli isolanti, il materiale più importante è certamente il biossido di silicio SiO_2, con conducibilità $\sigma \approx 10^{-18}$ S/cm. Dalla figura 1.1 si può notare come i valori di conducibilità elettrica di isolanti e conduttori siano separati da almeno 17 ordini di grandezza.

Nella stessa figura si trovano anche i dati relativi ai principali materiali semiconduttori, che si possono suddividere in:

- ▷ semiconduttori *elementari,* cioè costituiti da un solo elemento della tavola periodica, quali il germanio (Ge) ed il silicio (Si);
- ▷ semiconduttori *composti,* cioè risultanti da un legame chimico tra elementi diversi, quali l'arseniuro di gallio (GaAs), il fosfuro di indio (InP) e il fosfuro di gallio (GaP).

In questo caso, la resisitività elettrica non è rappresentata da un valore specifico, ma da una fascia di valori che copre diversi ordini di grandezza (circa 8 per il Si, 10 per il GaAs): la grande importanza pratica dei semiconduttori risiede proprio nella possibilità di modulare la conducibilità su un campo di valori così vasto, facendo uso della tecnologia del *drogaggio*, descritta nel paragrafo 1.6. Si noti che per alcuni semiconduttori composti, come il GaAs, il valore minimo di conducibilità elettrica raggiunge la regione degli isolanti: si parla, in questo caso, di semiconduttori *semi-isolanti.*

Le sezioni successive di questo capitolo hanno lo scopo di giustificare, sulla base di considerazioni di base legate alla fisica dei materiali, il valore delle proprietà elettriche dei materiali, con particolare attenzione ai materiali solidi *cristallini,* e con maggiore enfasi sui materiali semiconduttori del IV gruppo della tavola periodica, quali il Si.

1.2 Proprietà elettriche dei materiali

La fisica della materia permette di comprendere le proprietà dei materiali presenti in natura, studiando in primo luogo le proprietà fisiche delle singole specie atomiche e successivamente la natura dei legami che si instaurano quando gli atomi si combinano tra loro formando composti chimici, nelle varie forme possibili delle molecole (gas), dei liquidi e dei diversi tipi di solidi. Lo studio rigoroso delle proprità fisiche di un materiale richiede che si effettui la analisi a livello microscopico di un sistema molto complesso, che comprende l'insieme di tutti gli elettroni e dei protoni dei nuclei degli atomi che compongono il materiale stesso. Queste particelle interagiscono per effetto delle forze elettromagnetiche di attrazione tra gli elettroni e i nuclei, e di repulsione tra i nuclei e degli elettroni tra loro. L'analisi microscopica richiede, come è noto, di utilizzare le tecniche della meccanica quantistica, ovvero la soluzione della equazione di Schrödinger che permette di identificare gli stati energetici possibili per le varie particelle sottoposte alle rispettive forze di interazione. Considerando che i nuclei hanno massa molto maggiore degli elettroni, essi possono essere considerati a tutti gli effetti a riposo e il problema si semplifica riducendosi alla valutazione degli stati energetici permessi dei soli elettroni. Anche con questa semplificazione, tuttavia, l'analisi del materiale comprendente tutti gli elettroni sottoposti al potenziale di attrazione dei nuclei (e a quello corrispondente alla forza di repulsione tra di loro), rimane di difficile soluzione e può comunque essere condotto in modo rigoroso solo nei casi più semplici, quali per esempio nei singoli atomi o in alcune semplici molecole. Ad esempio, è nota la soluzione per la molecola biatomica H_2 che compone il gas di idrogeno, nella quale sono presenti solo due protoni e due elettroni [1]. Per materiali quali i solidi, costituiti da un numero molto elevato di atomi legati tra loro,[1] la soluzione diretta del

[1] Ad esempio una mole di materiale contiene un numero di atomi o molecole pari al numero di Avogadro $N_A = 6.022 \times 10^{23}\ \text{mol}^{-1}$. Ricordiamo anche che una mole di un materiale corrisponde ad

problema dal punto di vista quantistico risulta in genere troppo difficoltosa, e si può procedere cercando di identificare il tipo di legame che viene ad instaurarsi all'interno del solido stesso (ionico, covalente, metallico, ecc.) esaminando gli stati elettronici di semplici composti di pochi atomi, estendendo poi i risultati all'intero solido. Un caso sigificativo si ha quando i legami chimici portano all'interno del solido alla formazione di un sistema ordinato di atomi che si ripetono con una successione periodica nello spazio, ovvero alla formazione di un *cristallo*. In questo particolare caso l'equazione di Schödinger può essere risolta in maniera esatta, e gli stati elettronici sono descritti mediante la cosiddetta *teoria delle bande energetiche*. È proprio questo il contesto nel quale si possono comprendere le diverse caratteristiche elettriche dei materiali utilizzati in elettronica, e in particolare le differenze nei valori di conducibilità elettrica discusse nella sezione precedente, ovvero la classificazione dei materiali in isolanti, conduttori e semiconduttori. Nei paragrafi seguenti si cercerà di presentare la teoria delle bande di energia fornendone una trattazione alquanto semplificata e basata su considerazioni euristiche a partire dalle nozioni elementari della fisica atomica, poiché la trattazione completa dal punto di vista quantistico esula dagli scopi di questo testo. A tal fine nel paragrafo 1.2.1 si richiamano brevemente le principali nozioni di fisica atomica; nel paragrafo 1.2.2 si estendono le stesse al caso di una molecola biatomica come quella dell'idrogeno gassoso, e infine nel paragrafo 1.2.3 si arriva alla descrizione delle bande energetiche nelle strutture periodiche, e in particolare nei cristalli covalenti, di cui i semiconduttori quali il silicio e il germanio sono esempi.

1.2.1 Richiami di fisica atomica

Come anticipato, lo studio dei singoli atomi viene condotto nell'ipotesi che il nucleo abbia dimensioni molto piccole rispetto alla distanza media tra gli elettroni e il nucleo stesso, ed una massa molto maggiore rispetto a quella degli elettroni. Esso si considera quindi a riposo e, trascurando anche l'effetto della repulsione degli elettroni tra loro, ciascun elettrone risulta sottoposto ad un potenziale elettrico di tipo centrale

$$\varphi(r) = \frac{Zq}{4\pi\epsilon_0 r} \tag{1.6}$$

a cui corrisponde una energia potenziale:

$$U(r) = -\frac{Zq^2}{4\pi\epsilon_0 r} \tag{1.7}$$

dove Z è il numero atomico e r la distanza dell'elettrone dal nucleo. L'energia potenziale (1.7) è scritta supponendo l'energia dell'elettrone nulla a distanza infinita dal nucleo stesso, ovvero quando l'elettrone è libero, e il suo andamento in funzione della distanza dal nucleo è mostrato nella figura 1.2. È opportuno qui ricordare che l'unità di misura nel sistema internazionale (SI) per il potenziale elettrico è il V (Volt), mentre l'energia potenziale si esprime in J (Joule). Si nota però che un elettrone sottoposto a una differenza di potenziale d $\Delta\varphi = 1\,\mathrm{V}$ acquista una energia potenziale $U = -q\Delta\varphi$ pari a $-1,6\times10^{-19}\cdot 1\,\mathrm{C\,V} = 1,6\times10^{-19}\,\mathrm{J}$. Valori così piccoli di energia nascono dal fatto che

una quantità pari all'equivalente del suo peso atomico, espresso in *atomic mass units, a.m.u*

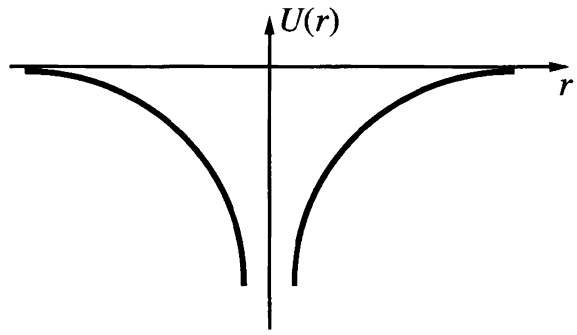


Figura 1.2 Andamento dell'energia potenziale a cui è sottoposto l'elettrone all'interno dell'atomo.

a livello microscopico le particelle in gioco (ad esempio gli elettroni) hanno carica molto piccola. Per evitare di dover utilizzare valori numericamente scomodi perchè troppo piccoli, si ricorre spesso ad una unità di misura alternativa di energia, che non è parte del SI, ma è più adatta nel campo della fisica atomica: l'*elettronvolt* eV. L'elettronvolt è definito come l'energia che un elettrone acquisisce quando posto al potenziale di 1 V, o meglio l'energia che un elettrone acquista quando accelerato mediante una differenza di potenziale di 1 Volt. Si ha ovviamente per definizione $1\,\text{eV} = 1,6 \times 10^{-19}\,\text{J}$. L'elettronvolt è l'unità di misura più utilizzata nella fisica della materia e il suo successo deriva anche dalla facilità nel passaggio dal potenziale elettrico all'energia potenziale elettrica: infatti un elettrone sottoposto ad un potenziale di x V, acquisisce una energia di x eV, ovvero il valore numerico delle due grandezze è esattamente lo stesso. Si faccia attenzione però a distinguere fisicamente i concetti di potenziale e di energia potenziale, anche quando per essi si usino gli stessi valori numerici.

In questa semplice descrizione si possono ricavare gli *stati elettronici* permessi per gli Z elettroni sottoposti all'azione del nucleo: questi sono come ben noto *discreti*, o quantizzati, e caratterizzati dai cosiddetti *numeri quantici* n, m, l, s detti rispettivamente numero quantico principale, numero quantico azimutale, numero quantico magnetico e numero quantico di spin. È noto [1] che il numero quantico principale n ($n = 0,1,2,\ldots$) corrisponde alla quantizzazione dell'energia, i numeri quantici l ($l = 0,\ldots,n-1$) e m ($m = -|l|,\ldots,|l|$) corrispondono alla quantizzazione del momento della quantità di moto mentre il momento angolare intrinseco, o spin, può assumere i due soli valori $s = \pm 1/2$. Uno stato elettronico, caratterizzato dall'insieme dei quattro numeri quantici, può essere occupato da un solo elettrone: per il principio di esclusione di Pauli è infatti vietato che ci possano essere nello stesso sistema due stati elettronici permessi con eguale insieme di numeri quantici. I diversi stati elettronici sono caratterizzati da una diversa densità di probabilità di trovare l'elettrone attorno al nucleo, ovvero da diversi *orbitali*. Nel caso in cui $l = 0$ la densità di probabilità ha simmetria sferica e il corrispondente orbitale è detto di tipo s; nel caso $l = 1$ si hanno tre orbitali, detti di tipo p, che corrispondono ai tre valori permessi del numero magnetico m, $m = -1,0,1$, e orientati nelle tre direzioni spaziali x, y, z; si possono poi definire cinque orbitali di tipo d in corrispondenza del numero quantico $l = 2$ e così via per i valori di l ancora crescenti. I livelli caratterizzati dal numero principale n hanno energia crescente al crescere di n, ed è quindi conveniente definire la *corteccia elettronica*, o in inglese

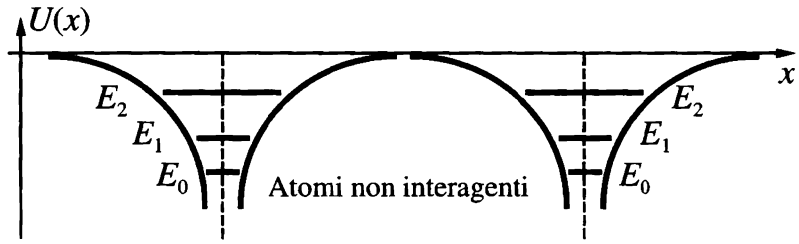


Figura 1.3 Andamento dell'energia potenziale e dei livelli energetici permessi nel sistema H_2 per una distanza interatomica molto elevata.

shell, come l'insieme degli stati elettronici con uguale numero atomico principale n. La prima corteccia con $n = 0$ è caratterizzata da $l = m = 0$, e quindi in questo caso si ha solo l'orbitale s, indicato con il simbolo $1s$: esso può ospitare al massimo due elettroni con spin opposti. La seconda corteccia, con $n = 1$, comprende sia l'orbitale $2s$ sia i tre orbitali $2p$ e ospita al massimo 8 elettroni. La terza e quarta corteccia ospitano poi rispettivamente 18 e 32 elettroni. Gli Z elettroni presenti nella specie atomica con numero atomico Z riempiono gli stati disponibili a partire da quello a energia minore fino a quelli a energia superiore. La corteccia con $n = 0$ viene quindi occupata per prima, seguita da quella con $n = 1$ e così via fino ad esaurire gli Z elettroni dell'atomo. Le cortecce che risultano occupate da un numero di elettoni pari al massimo numero di stati disponibili sono dette *complete*, mentre gli elettroni che riempiono parzialmente la corteccia più esterna sono detti *elettroni di valenza*. Sono questi gli unici elettroni che concorrono alla formazione dei legami chimici. È noto che il numero di elettroni presenti nella corteccia più esterna e il numero di cortecce completamente piene consente di catalogare le specie atomiche nella tavola periodica degli elementi.

1.2.2 Le molecole

La molecola H_2 di cui è composto il gas di idrogeno è il più semplice sistema multiatomico di cui sia nota la soluzione della equazione di Schrödinger: essa è infatti costituita solamente da due nuclei di idrogeno (due protoni) e da due elettroni. In questa sede non siamo certamente interessati alla molecola H_2 in quanto tale, ma utilizzeremo questo esempio per comprendere come si passi dalla conoscenza delle proprietà atomiche di un singolo atomo a quelle di una molecola composta da più atomi della stessa specie quando tra essi venga formarsi un legame chimico. Anche in questo caso si possono considerare i nuclei a riposo e gli elettroni sottoposti alla sovrapposizione di due potenziali centrali del tipo descritto nella equazione (1.7). Per semplificare la trattazione si considera in un primo momento che i nuclei siano molto distanti tra loro e che possano essere considerati non interagenti. In questo caso i livelli energetici di ciascuno dei due atomi singoli rimangono imperturbati: il potenziale a cui sono sottoposti gli elettroni, ciascuno nel suo atomo di partenza, è rappresentato nella figura 1.3, dove sono anche riportati i livelli energetici caratteristici dell'atomo di idrogeno isolato. Alla formazione della molecola i due atomi interagiscono e ciascuno degli elettroni viene ad essere sottoposto alla attrazione sempre più forte del nucleo dell'atomo vicino oltre che del proprio. Di conseguenza ciasun elettrone ha una probabilità non nulla di trovarsi indifferentemente attorno a ciascuno dei due nuclei e non è più possibile distinguere a

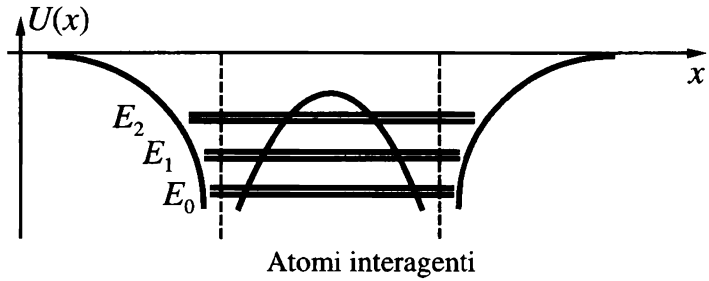


Figura 1.4 Andamento dell'energia potenziale e dei livelli energetici permessi nel sistema H_2 per atomi interagenti.

quale atomo ciascuno elettrone era inizialmente legato: si è formato un unico sistema quantistico in cui gli elettroni sono tra di loro indistinguibili, esattamente come gli Z elettroni che si trovano all'interno dello stesso atomo in un sistema monoatomico. In queste condizioni gli stati permessi per gli elettroni si estendono in tutto lo spazio che può essere occupato dall'elettrone stesso, e sono caratterizzati da nuovi orbitali, detti molecolari, che rappresentano la probabilità di trovare l'elettrone all'interno della molecola. Per quanto riguarda i livelli energetici, dalla soluzione esatta dal punto di vista quantistico si osserva un fenomeno di sdoppiamento di ciascuno dei livelli energetici del singolo atomo, come mostrato nella figura 1.4. Questo effetto può essere interpretato osservando che se i livelli restassero invariati ripetto al caso dei due atomi non interagenti si avrebbero due stati permessi con lo stesso insieme di numeri quantici, ovvero una situazione vietata dal principio di esclusione di Pauli. Sempre nella figura 1.4 i nuovi livelli energetici sono mostrati estesi a tutto lo spazio compreso tra i due nuclei, per indicare che a questi nuovi livelli sono associati i nuovi orbitali molecolari, che risultano quindi anch'essi in numero doppio rispetto agli orbitali di partenza. Ad esempio l'orbitale atomico $1s$, si sdoppia in due stati che possono contenere ciascuno 2 elettroni (con spin opposti), per un totale di 4 stati possibili, e così via per gli orbitali $2s$, $2p$ ecc. Si può anche dimostrare che questo risultato ha natura generale: quando N atomi interagiscono formando una molecola, i livelli energetici dell'atomo isolato si trasformano all'interno della molecola in un insieme ravvicinato di N livelli. Si osservi che il numero complessivo di stati permessi per l'elettrone viene conservato, ovvero rimane pari al numero di stati permessi per l'atomo singolo moltiplicato per il numero di atomi N che formano la molecola. Se $N = 2$, come nel caso delle molecole biatomiche, si ha proprio il caso mostrato nella figura 1.4. Gli elettroni presenti nella molecola riempiono poi gli stati molecolari con energia minore e via via gli stati con energia maggiore.

Dal punto di vista chimico l'esistenza di stati elettronici permessi che si estendono a tutta la molecola corrisponde proprio alla formazione di un legame chimico tra gli atomi della molecola stessa, e gli elettroni che si trovano all'interno degli orbitali molecolari sono elettroni di legame. In modo improprio si dice che gli oribitali originari dei due atomi separati *interagiscono* nella molecola a formare N orbitali molecolari, mantenendo il numero totale di stati permessi invariato: naturalmente la nozione di interazione degli orbitali non ha nulla di fisico, ma richiama in maniera intuitiva il fatto che il fenomeno della separazione dei livelli energetici e la formazione degli orbitali

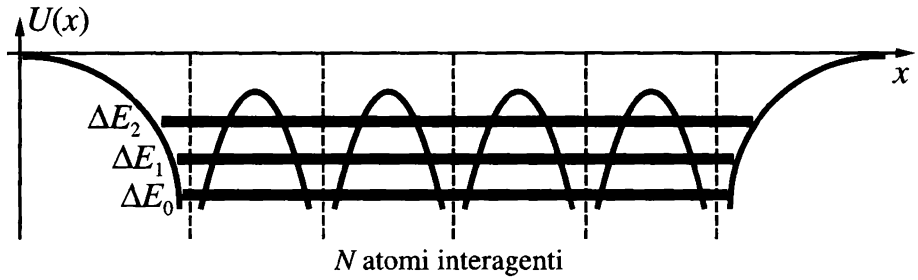


Figura 1.5 Andamento dell'energia potenziale e delle bande energetiche permesse in un solido ordinato con N atomi.

molecolari si ha effettivamente quando gli atomi originari interagiscono formando il legame molecolare.

1.2.3 La teoria delle bande energetiche

Si consideri ora di estendere il ragionamento svolto finora al caso di una schiera ordinata di N atomi identici uniti a formare un solido cristallino. Per quanto visto in precedenza ogni livello energetico della specie atomica di partenza viene a dividersi in N livelli vicini. Poiché però in un solido il numero di atomi è molto elevato, dell'ordine come visto del numero di Avogadro, gli N livelli che si formano determinano un insieme molto ravvicianto di stati permessi, fino a formare una cosiddetta *banda di energia permessa*. Un esempio qualitativo è mostrato nella figura 1.5. La trattazione rigorosa mediante la teoria quantistica permette di verificare che le bande di energia sono una caratteristica dei sistemi con *potenziale periodico*, e quindi in particolare dei solidi cristallini in cui si ha una successione regolare di atomi. Una banda energetica che si forma dall'interazione di N livelli energetici può ospitare un numero di elettroni pari a N volte il numero di elettroni possibili per il livello energetico originario. La rappresentazione in figura 1.5 è ovviamente semplificata. In realtà quando si viene a formare un solido formato dalla successione di N atomi con numero atomico Z, gli elettroni delle cortecce più interne non vengono a partecipare ai legami con gli atomi vicini e rimangono pressoché inalterati negli stati originari all'interno degli atomi di partenza. Gli elettroni che risentono della interazione con gli atomi vicini sono invece solo gli elettroni di valenza. Sono quindi solamente gli stati elettronici della corteccia più esterna a interagire, formando le bande di energia mostrate in figura 1.5 e a costituire gli elettroni di valenza totali del solido, in numero pari al prodotto di N per il numero di elettroni di valenza di un singolo atomo, che vanno a occupare tali bande energetiche a partire da quella a energia più bassa fino a quelle di energia superiore. L'ultima banda occupata da elettroni, ovvero la banda a energia più elevata, che risulta in genere parizialmente piena, è detta *banda di conduzione*.

La formazione delle bande di energia riveste notevole importanza nella comprensione delle proprietà elettriche dei materiali. Infatti gli stati elettronici sono adesso estesi a tutto il solido, come mostrato nella stessa figura 1.5, ovvero gli elettroni che hanno energia corrispondente ad una determinata banda energetica sono completamente *delocalizzati*. Essi possono quindi occupare una posizione qualunque all'interno del solido, indipendentemente dalla posizione dell'atomo di origine. Questo non vuol dire che sot-

to l'azione di una forza esterna, quale ad esempio un campo elettrico, questi possano sempre muoversi all'interno del solido formando un flusso di carica, ovvero una corrente di conduzione elettrica. Anzi, si può facilmente comprendere che una banda energetica completamente piena di elettroni non può dar luogo ad alcun trasporto di carica, in quanto gli elettroni, per muoversi sotto l'azione di una forza esterna, devono essere in grado di acquisire energia cinetica aumentando la loro energia totale. Per farlo, però, è necessario che sia disponibile nella banda uno stato vuoto che possa essere occupato dall'elettrone al variare della sua energia cinetica, cosa che non si realizza in un banda completamente occupata. Se ne deduce che gli elettroni che occupano le bande energetiche più profonde e completamente piene non partecipano alla conduzione di corrente e, anzi, non possono muoversi se non per scambiarsi di posto con altri elettroni della stessa banda, mantenendo la loro posizione e velocità mediamente immutata. Questo fatto corrisponde a supporre che gli elettroni delle bande a energia inferiore siano ancora fortemente legati agli atomi originari oppure siano impeganti a formare legami chimici di tipo fortemente localizzato con gli atomi vicini: è questo per esempio quanto si realizza nel caso dei solidi di tipo *covalente* di cui si discuterà in dettaglio nel paragrafo 1.2.4.

La banda di conduzione è invece, nel caso più generale, solo parzialmente occupata da elettroni, ed in questo caso essi possono muoversi sotto l'azione di una forza esterna all'interno del solido, occupando gli stati energetici liberi della banda al variare della loro energia cinetica. Si può avere in questo caso trasporto di carica, e quindi conduzione di corrente elettrica. In presenza di trasporto di carica è però anche necessario analizzare quali siano le leggi che descrivono il moto degli elettroni stessi. Essi sono sottoposti sia al potenziale di attrazione dei nuclei sia alle forze agenti dall'esterno sul materiale, così che lo studio del loro moto risulta complesso. Una semplificazione si ottiene supponendo che forze esterne siano in grado di esercitare sull'elettrone una variazione di energia piccola, rispetto alla sua energia di legame nel materiale. Le forze esterne sono perciò considerate come delle deboli perturbazioni dello stato elettronico delle cariche all'interno del solido. Questa ipotesi è ragionevole poiché, nelle situazioni di normale utilizzo dei dispositivi elettronici, i materiali non vengono sottoposti a forze esterne tali da modificare, se non in misura minima, le proprietà elettriche del materiale stesso con cui essi vengono realizzati, altrimenti si avrebbe un danneggiamento o addirittura la distruzione del dispositivo stesso. Nella ipotesi che le forze applicate esterne siano deboli, si può dimostrare una interessante proprietà: gli elettroni all'interno del solido si comportano come *particelle quasi-classiche,* ovvero il loro moto può essere descritto mediante le ben note leggi della meccanica classica, assegnando ad essi una posizione e una quantità di moto. Inoltre esse si comportano come *particelle quasi-libere* ovvero come particelle che, sebbene trovandosi all'interno del materiale siano sottoposte alla forza di legame con il solido, si possono descrivere come se questa forza non esistesse: in altre parole come particelle libere nel vuoto a cui venga applicata la sola forza esterna. Il moto degli elettroni si studia quindi secondo le leggi della meccanica classica come se essi fossero sottoposti alla azione delle *sole forze esterne.* Naturalmente questa descrizione non è realistica, e infatti si dimostra che essi si muovono nel solido effettivamente come se fossero liberi, ma con una massa equivalente, detta *massa efficace,* diversa da quella dell'elettrone nel vuoto e diversa a seconda della banda energetica in cui essi si trovano a muoversi. La massa efficace degli elettroni verrà in seguito indicata con il simbolo m_n^*. In conclusione si può affermare che gli elettroni all'interno delle bande di energia possono essere trattati mediante le leggi della meccanica classica, a

patto di sostituire la loro massa con la massa efficace. Questo comporta una notevole semplicazione nella trattazione delle proprietà di trasporto di carica sotto l'azione dei campi elettrici esterni, ovvero nello studio della conduzione elettrica, poiché non sarà necessario effettuare questa analisi mediante gli strumenti della meccanica quantistica.

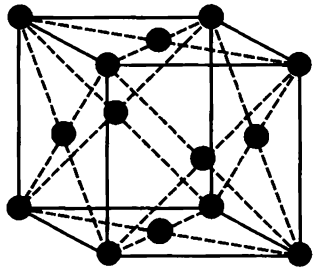


Figura 1.6 Cella elementare del cristallo di silicio.

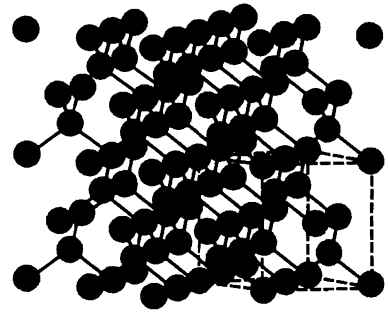


Figura 1.7 Struttura tetraedrica del cristallo di silicio.

1.2.4 Gli elementi del IV gruppo della tavola periodica

I cristalli formati da una successione ordinata di elementi del IV gruppo della tavola periodica, quali il carbonio C, il germanio Ge, il silicio Si e lo stagno Sn, sono in generale cristalli covalenti. Lo studio di questi materiali ci consente in particolare di comprendere le proprietà elettriche del silicio, ovvero del materiale che è maggiormente utilizzato nella fabbricazione di circuiti integrati.

Dal punto di vista chimico i cristalli formati dagli elementi del gruppo IV sono caratterizzati da una cella cristallina elementare di tipo cubico a facce centrate, come quella del silicio mostrata nella figura 1.6.

I legami che si formano sono di tipo covalente e sono caratterizzati da una orientazione tetraedrica nello spazio, come mostrato nella figura 1.7. Gli orbitali che formano i legami nascono dalla ibridizzazione dell'obitale s e dei tre orbitali p formando i quattro orbitali sp^3, con la tipica orientazione tetraedrica nello spazio. Ciascun legame covalente nasce dalla condivisione di una coppia (doppietto) di elettroni tra un atomo e ciascuno dei suoi quattro primi vicini, come mostrato schematicamente nella figura 1.8.

Questi stessi risultati si possono ricavare anche seguendo l'approccio della teoria delle bande visto nel paragrafo precedente, ovvero seguendo un approccio di tipo fisico. Passiamo quindi ad analizzare in dettaglio la struttura delle bande di energia che caratterizzano materiali del IV gruppo. La figura 1.9 mostra al variare della distanza interatomica tra gli atomi che formano il cristallo, l'andamento tipico delle bande energetiche che si formano nel solido. Essa rappresenta in modo qualitativo quanto si verica in tutti i solidi covalenti del IV gruppo della tavola periodica, anche se le energie coinvolte variano a seconda delle singole specie atomiche. Come già spiegato nel paragrafo precedente, non è necessario considerare i livelli energetiche delle cortecce interne complete, i cui elettroni non partecipano alla formazione di legami chimici e rimangono pressoché localizzati attorno agli atomi di origine. Si considerano quindi solamente le

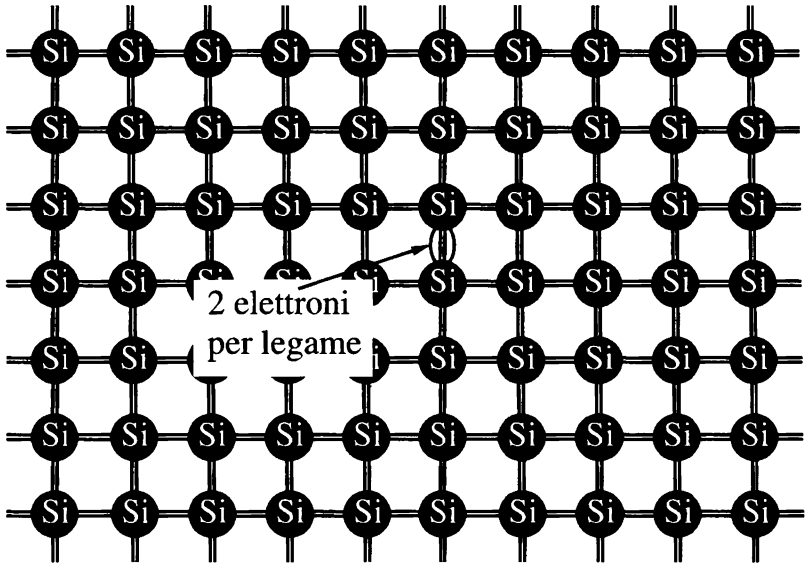


Figura 1.8 Struttura semplificata del legame chimico covalente.

bande di energia che nascono dalla interazione dei livelli energetici della corteccia più esterna, ovvero i livelli $2s$ e $2p$ per il C, i $3s$ e $3p$ per il Si, i $4s$ e $4p$ per il Ge e infine i $5s$ e $5p$ per lo Sn. Si noti che, essendo tutti questi elementi del quarto gruppo, la configurazione della corteccia più esterna degli atomi singoli è la stessa, caratterizzata dallo stato s occupato da due elettroni e lo stato p occupato da 2 elettroni e con 4 stati liberi disponibili. La figura 1.9 mostra l'andamento delle bande energetiche che si formano dalla interazione degli orbitali s e p al variare della distanza interatomica che caratterizza il solido covalente. Si nota che se la distanza interatomica è molto grande, i livelli energetici dell'atomo singolo non vengono perturbati e si hanno quindi $2N$ stati possibili nell'orbitale s completamente occupati da $2N$ elettroni e $6N$ stati possibili negli orbitali p, occupati però solamente da $2N$ elettroni. A mano a mano che la distanza interatomica si riduce, i livelli energetici formano le bande di energia, la cui forma varia al variare della distanza interatomica, saldandosi in un'unica banda per distanze interatomiche intermedie e infine dividendosi ulteriormente a formare due bande, quella inferiore, detta *banda di valenza* (BV) e quella superiore, detta *banda di conduzione* (BC). Ciascuna di queste bande può contenere al suo interno un numero di elettroni pari a $4N$. Osserviamo anche che il numero di elettroni di valenza nei materiali del IV gruppo è pari a quattro per ogni atomo, due dei quali originariamente nel livello s e due nei livelli p, corrispondenti ad un totale di $4N$ elettroni di valenza da disporre nelle bande. Tali elettroni occupano per prima la banda a energia inferiore, ovvero la banda di valenza, che risulta quindi completamente piena, mentre la banda di conduzione risulta completamente vuota. Il fatto poi che, al variare della distanza interatomica, le bande energetiche si uniscano prima e si dividano poi nelle due bande di valenza e conduzione, che possono ospitare ciascuna al suo interno lo stesso numero

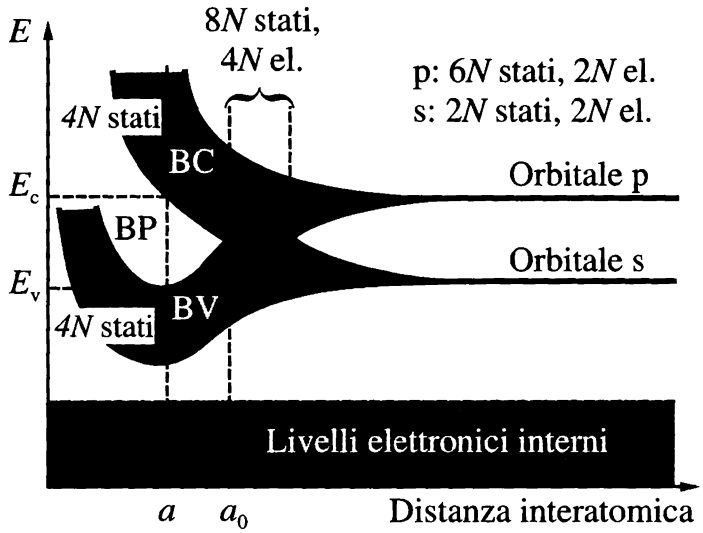


Figura 1.9 Formazione della banda di valenza e banda di conduzione in un cristallo covalente.

di elettroni, è legato al fenomeno della ibridizzazione degli orbitali sp^3. Questi sono infatti quattro orbitali identici che possono al loro interno contenere $8N$ stati, e che interagendo danno luogo a due bande identiche con $4N$ stati ciascuna. Si può quindi sostenere che solo i materiali caratterizzati da un passo reticolare minore del valore a_0 definito nella figura 1.9 formano legami effettivamente covalenti, come ad esempio il materiale mostrato con passo reticolare a. Tra questi si hanno i materiali finora analizzati, ovvero il carbonio, il germanio e il silicio: per tutti questi si possono quindi definire le bande di conduzione e di valenza.

L'intervallo di energia che separa la banda di valenza dalla banda di conduzione è detto *banda di energia proibita* (BP), sebbene comunemente si utilizzi anche la terminologia inglese *energy gap*. Il valore della ampiezza della banda proibita si denota con E_g. Si definiscono anche i valori dell'energia del bordo superiore della banda di valenza E_v, e del bordo inferiore della banda di conduzione E_c, per cui si ha $E_g = E_c - E_v$. Le energie comprese nella banda proibita costituiscono stati vietati per gli elettroni che, per passare eventualmente dagli stati permessi della banda di valenza a quelli della banda di conduzione, devono quindi acquisire (da cause esterne al cristallo), energia pari o superiore al valore di E_g. Nella figura 1.9 si nota che il valore di E_g è maggiore per i materiali caratterizzati da una minore distanza interatomica a, e minore per materiali con a maggiore. Si hanno quindi proprietà fisiche diverse al variare del passo reticolare: ad esempio il carbonio con passo reticolare $a = 3,57$ Å ha ampiezza della banda proibita $E_g = 5,47$ eV mentre il silicio, con passo reticolare $a = 5,43$ Å, è caratterizzato da $E_g = 1,124$ eV, molto minore del carbonio.

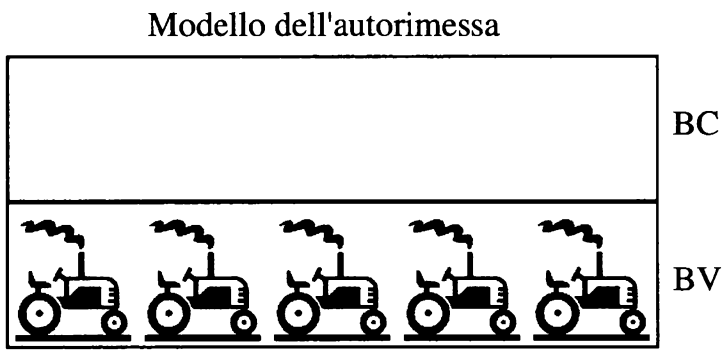


Figura 1.10 Modello dell'autorimessa per illustrare l'occupazione della banda di valenza e della banda di conduzione per $T = 0\,\mathrm{K}$.

1.3 Isolanti, semiconduttori e metalli

Come già illustrato nel paragrafo precedente una banda in cui tutti gli stati elettronici permessi sono occupati da elettroni non può dare origine a trasporto di carica elettrica. Questo è il caso ad esempio della banda di valenza, che viene riempita da tutti gli elettroni di valenza del cristallo. Essi non sono liberi di muoversi all'interno del solido, in quanto impegnati nei legami covalenti, fortemente direzionali e localizzati. Per cercare di illustrare meglio questo concetto si utilizza spesso il cosiddetto modello della autorimessa, riportato nella figura 1.10. La banda di valenza è equiparata al piano inferiore di una rimessa per automobili, in cui tutti i posti disponibili sono occupati da un'automobile (nel caso della figura 1.10 si tratta in realtà di trattori!). All'interno del piano inferiore le automobili non possono muoversi, poiché ogni posto disponibile è occupato, mentre per potersi muovere esse dovrebbero riuscire a raggiungere il piano superiore della autorimessa, ovvero la banda di conduzione, completamente vuoto. Per poter raggiungere il piano superiore è però necessario che le automobili abbiano energia sufficiente, ovvero, in termini fisici, è necessario che gli elettroni della banda di valenza acquisiscano energia sufficiente per effettuare il salto di banda, pari almeno ad E_g. Se il sistema è però isolato e privo di qualsiasi interazione con l'ambiente esterno ciò non è possibile. Naturalmente questa situazione si verifica a rigore solamente se il sistema è completamente isolato anche dal punto di vista termico, ovvero alla temperatura di congelamento $T = 0\,\mathrm{K}$. In queste condizioni si può quindi affermare che i materiali del IV gruppo sono dei perfetti isolanti, non essendo possibile alcun moto di carica e alcuna conduzione elettrica.[2]

Naturalmente un solido non mai è in realtà un sistema completamente isolato dal mondo esterno, poiché è almeno in grado di assorbire dall'ambiente in cui è immerso energia termica, portandosi all' equilibrio termico con esso. In altre parole le particelle

[2] Ovviamente il sistema potrebbe condurre anche per $T = 0\,\mathrm{K}$, qualora le forze esterne fossero cosìì intense da essere in grado di cedere agli elettroni della banda di valenza una energia pari ad E_g. Questo però contrasta con l'ipotesi che il sistema sia sottoposto a deboli forzanti etserne, ovvero a forze che non mutino significativamente le proprietà del materiale stesso. Si deve infatti ricordare che la presenza degli elettroni nella banda di valenza è equivalente al supporre che gli elettroni siano impegnati nei legami chimici, la loro rimozione dalla banda di valenza causerebbe quindi la rottua di un legame covalente, come spiegato in dettaglio più avanti.

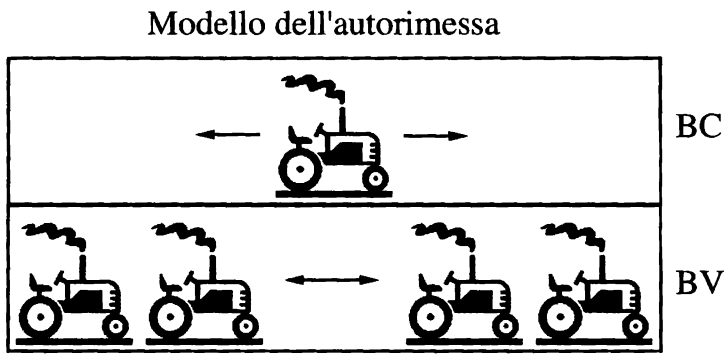


Figura 1.11 Modello dell'autorimessa per illustrare l'occupazione della banda di valenza e della banda di conduzione per $T \neq 0\,\mathrm{K}$.

che compongono il solido, trovandosi ad una temperatura $T \neq 0\,\mathrm{K}$, acquisiscono una energia non nulla di tipo termico, che si aggiunge a quella che essi hanno per effetto delle sole forze elettromagnetiche di legame all'interno del materiale. In queste condizioni è possibile che qualche elettrone all'interno della banda di valenza acquisisca una energia sufficiente per effettuare il salto di banda, lasciando uno stato elettronico permesso vuoto nella banda di valenza e occupando uno degli stati permessi nella banda di conduzione. Gli elettroni hanno in media energia di agitazione termica molto minore di E_g, dell'ordine di qualche frazione di eV, e una distribuzione statistica di energia descrescente per energie crescenti, per cui solo pochi di essi hanno probabilità significativamente diversa da zero di acquisire energia sufficiente per passare dalla banda di valenza alla banda di conduzione. Ne consegue che il numero di elettroni che si trovano nella banda di conduzione (e di posti vuoti nella banda di valenza) è in numero limitato. Dal punto di vista del modello dell'autorimessa la situazione è ora quella descritta nella figura 1.11.

Si osserva che è adesso possibile una conduzione elettrica. Infatti sia gli elettroni nella banda di valenza sia gli elettroni della banda di conduzione possono ora muoversi sotto l'azione di forze esterne, avendo a disposizione degli stati permessi liberi nelle rispettive bande. Poiché il numero di posti liberi nella banda di valenza è limitato, solo pochi elettroni sono effettivamente in grado di muoversi. Essi abbandonano lo stato occupato originariamente per occupare uno di quelli vuoti disponibili, lasciando a loro volta un posto vuoto dietro di sé. Il moto degli elettroni può quindi essere associato anche ad un equivalente spostamento dei posti vuoti. Da questa analogia si intuisce che è possibile introdurre una *particella fittizia*, detta *lacuna* (in inglese *hole*), che corrisponde ad un posto non occupato dall'elettrone, e che si muove in sintonia con esso, ovvero nella stessa direzione ma con verso opposto. Dal punto di vista chimico la lacuna può essere interpretata come la rottura di un legame covalente, come mostrato nella figura 1.12. Se infatti la banda di valenza fosse completamente piena tutti gli elettroni sarebbero impegnati nei legami covalenti. Quando invece un elettrone passa nella banda di conduzione, esso può muoversi all'interno del cristallo libero dal legame covalente originario, il quale risulta quindi incompleto. La presenza di legami incompleti rende possibile il moto degli elettroni anche all'interno della banda

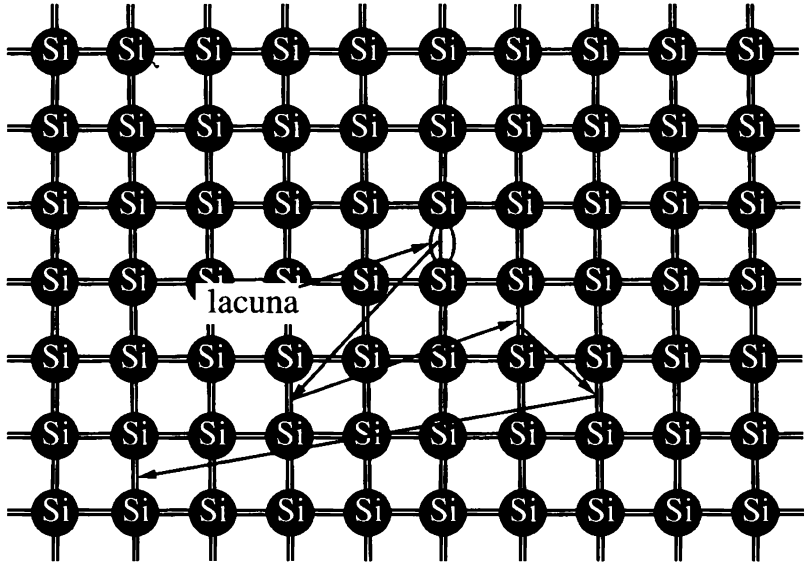


Figura 1.12 Rappresentazione chimica del concetto di lacuna.

di valenza: essi possono infatti muoversi per ricostituire il legame covalente, lasciando però a loro volta dietro di sé una mancanza nel legame stesso, ovvero una nuova lacuna.

Si noti che la lacuna può essere utilizzata per descrivere il trasporto di carica all'interno della banda di valenza in maniera del tutto equivalente alla descrizione mediante gli elettroni. Poiché però all'interno della banda di valenza il numero di elettroni è estremamente elevato e il numero di lacune è ridotto, risulta conveniente descrivere il moto delle cariche attraverso le lacune piuttosto che attraverso gli elettroni. Si può ancora dimostrare che la lacuna è effettivamente equiparabile ad una particella fittizia a cui è possibile assegnare sia una carica sia una massa equivalente. Per quanto riguarda la carica, è intuitivo convincersi che la lacuna, rappresentando uno stato non occupato da un elettrone, ha una carica tale da compensare quella di un eventuale elettrone che occupasse lo stato stesso, ovvero una carica positiva uguale e opposta alla carica dell'elettrone. Per quanto riguarda la massa, la lacuna si comporta come una particella quasi-libera all'interno della banda di valenza, e sarà quindi caratterizzata dalla sua *massa efficace* m_p^*.

In conclusione si può affermare che la conduzione elettrica si può descrivere mediante il moto delle lacune all'interno della banda di valenza e il moto degli elettroni all'interno della banda di conduzione. Nel solido le lacune si comportano come particelle libere sotto l'azione delle forze esterne, con carica pari a $q = +1,6 \times 10^{-19}$ C e massa m_p^* il cui valore dipende dal materiale in esame. Gli elettroni, a loro volta, si comportano come particelle con carica $-q = -1,6 \times 10^{-19}$ C e massa m_n^* il cui valore, oltre a dipendere dal materiale in esame, è in generale diverso dal valore di m_p^*. Solitamente le lacune nella banda di valenza hanno massa efficace maggiore degli elettroni nella banda di conduzione: ad esempio nel silicio $m_p^* \approx 3m_n^*$. D'ora in avanti ci si riferirà

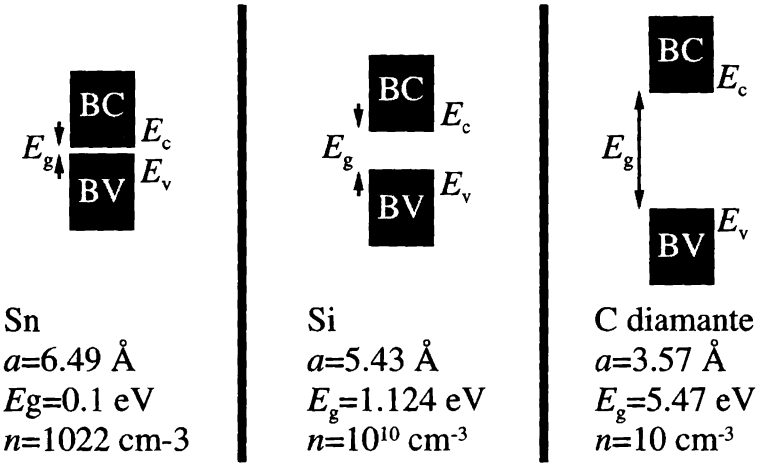


Figura 1.13 Differenza nella struttura a bande di metalli, semiconduttori e isolanti.

alle lacune sottintendendo che esse si trovano nella banda di valenza e, analogamente, con elettroni si indicheranno unicamente quelli presenti nella banda di conduzione.

Il numero di elettroni e di lacune nel materiale è chiaramente esattamente uguale. Ad esso è legata la conducibilità del materiale stesso, in quanto una maggiore concentrazione di carica all'interno delle bande dà origine a una maggiore intensità di corrente di conduzione elettrica. È quindi importante classificare i materiali in esame in base alla densità di portatori nelle bande, la quale a sua volta dipende in primo luogo dalla ampiezza della banda proibita. Tanto più questa è piccola tanto più gli elettroni saranno favoriti nell'effettuare il salto di banda per effetto della agitazione termica, dando luogo a una maggiore concentrazione sia di elettroni sia di lacune. Se invece essa è grande la concentrazione di carica nelle bande rimane trascurabile. Nella figura 1.13 si mostra la differenza nella ampiezza di banda proibita tra il carbonio, il silicio e lo stagno. Tra questi il carbonio (che nella forma cristallina covalente è detto diamante) ha ampiezza di banda maggiore e una concentrazione di elettroni n nella banda di conduzione pari a soli 10 elettroni al centimetro cubo ($10\,\mathrm{cm}^{-3}$): esso è quindi un ottimo isolante. Il caso opposto è quello dello stagno, dove le bande sono talmente ravvicinate da costituire quasi un'unica banda senza soluzione di continuità. Gli elettroni riescono già a temperatura ambiente a raggiungere in gran numero la banda di conduzione, distante solo qualche decimo di eV, dando luogo a una forte popolazione di elettroni ($n = 10^{22}\,\mathrm{cm}^{-3}$), tanto che lo stagno ha comportamento di tipo metallico ed è un buon conduttore di corrente elettrica. Una situazione intermedia si ha nel caso del silicio, in cui il valore della ampiezza della banda proibita è dell'ordine di 1 eV e la concentrazione di elettroni è dell'ordine di $10^{10}\,\mathrm{cm}^{-3}$. Una situazione analoga si ha nel germanio, in cui $E_g = 0.66\,\mathrm{eV}$. Questi materiali, avendo caratteristiche comprese tra gli isolanti e i conduttori, sono detti *semiconduttori:* essi hanno tipicamente conducibilità di basso valore ma non completamente trascurabili, poiché la densità di carica all'interno delle bande è comunque significativa.

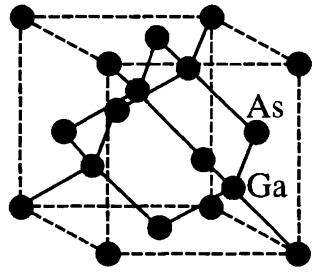


Figura 1.14 Struttura cristallina del semiconduttore composto GaAs (zincoblenda).

1.4 I semiconduttori composti

Nel paragrafo precedente si è visto che alcuni materiali del IV gruppo della tavola periodica sono semiconduttori. Le caratteristiche di questi materiali sono la struttura cristallina cubica a facce centrate, il legame puramente covalente e una ampiezza della banda proibita dell'ordine dell'elettronvolt. Esistono anche altri materiali che hanno caratteristiche simili al silicio e al germanio, pur non appartemendo al IV gruppo della tavola periodica. Essi sono comunque solidi cristallini ma la loro struttura non è costituita da atomi identici, ovvero si tratta di materiali composti. La cella cristallina di questi materiali non è in genere cubica a facce centrate e il legame tra i diversi atomi non è di tipo puramente covalente, presentando anzi spesso una componente di legame ionico. Nonostante queste differenze, è ancora possibile individuare una banda di valenza e una banda di conduzione, la cui distanza rimane dello stesso ordine di grandezza di quello visto per i semiconduttori monoatomici quali il silicio e il germanio. Di conseguenza questi materiali hanno anch'essi caratteristiche elettriche di tipo semiconduttore e si parla infatti di materiali *semiconduttori composti*. I principali semiconduttori composti sono miscele di sole due specie atomiche, nonostante sia possibile creare materiali semiconduttori anche con miscele di tre o quattro specie atomiche diverse. La più importante famiglia di semiconduttori composti è quella dei semiconduttori III-V, in cui si succedono nel cristallo atomi del III e atomi del V gruppo della tavola periodica. In questa famiglia troviamo ad esempio l'*arseniuro di gallio* GaAs, la cui cella elementare, rappresentata nella figura 1.14, è un esempio della struttura detta della zincoblenda, e il *fosfuro di indio*, InP. Esistono poi semiconduttori composti della famiglia II-VI, come il nitruro di gallio GaN e il tellururo di cadmio CdTe. Infine sono semiconduttori anche alcuni composti binari di atomi del IV gruppo della tavola periodica, quali il carburo di silicio SiC e il silicio-germanio SiGe.

Molti materiali composti non possono essere fabbricati in forma cristallina mediante un processo di produzione diretto, ma è possibile solamente realizzarne sottili strati per accrescimento a partire da un substrato di tipo cristallino di materiale diverso. Gli unici semiconduttori che possono essere fabbricati nella forma di substrati (in inglese *bulk*) sono il silicio, l'arseniuro di gallio, il fosfuro di indio, il carburo di silicio e il nitruro di gallio. Le dimensioni dei cristalli ottenibili e i costi di produzione variano però fortemente da un materiale all'altro. La tecnologia di fabbricazione del silicio resta quella più avanzata ed economica, mentre tecnologie quali quella del carburo di silicio e del nitruro di gallio sono in fase ancora sperimentale, e quindi molto costose. In generale nella produzione dei circuiti integrati viene utilizzato il silicio a meno che non ci sia

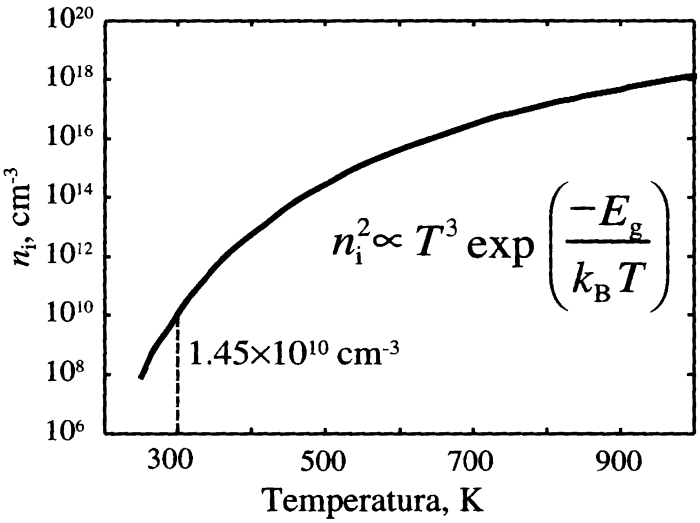


Figura 1.15 Andamento della concentrazione intrinseca al variare della temperatura nel silicio.

una reale necessità di utilizzare semiconduttori diversi per applicazioni particolari o di nicchia, quali l'optoelettronica, l'elettronica delle microonde o l'elettronica di potenza.

1.5 La concentrazione intrinseca

Il numero di elettroni per unità di volume che si trovano nella banda di conduzione di un semiconduttore è detto *concentrazione intrinseca* n_i e si esprime in cm^{-3}; questo è anche il numero di lacune per unità di volume p_i, ma quest'ultima notazione, in quanto superflua, non viene quasi mai utilizzata. La concentrazione intrinseca nei semiconduttori varia in funzione della ampiezza della banda proibita, diminuendo all'aumentare di questa. Poiché il numero di portatori nelle bande è legato alla energia di agitazione termica degli elettroni, se ne deduce facilmente che n_i dipende anche fortemente dalla temperatura. Infatti si dimostra nel paragrafo 1.7.1 che la concentrazione intrinseca cresce secondo la legge:

$$n_i^2 \propto T^3 \exp\left(-\frac{E_g}{k_B T}\right) \tag{1.8}$$

dove k_B è la costante di Boltzmann, il cui valore può essere espresso nel SI $k_B = 1.3807 \times 10^{-23}$ J/K oppure utilizzando come unità di energia l'elettronvolt, $k_B = 8.62 \times 10^{-5}$ eV/K. L'andamento della concentrazione intrinseca (1.8) è mostrato nella figura 1.15 nel caso del silicio.

Si osservi anche che nella (1.8) la dipendenza della temperatura è presente anche nel valore dell'ampiezza della banda proibita E_g, che a sua volta dipende da T come mostrato nella figura 1.16. Si osservi che al crescere della temperatura l'ampiezza di banda diminuisce in tutti i materiali semiconduttori, favorendo ulteriormente la crescita della concentrazione intrinseca. Si noti anche che l'ampiezza della banda proibita

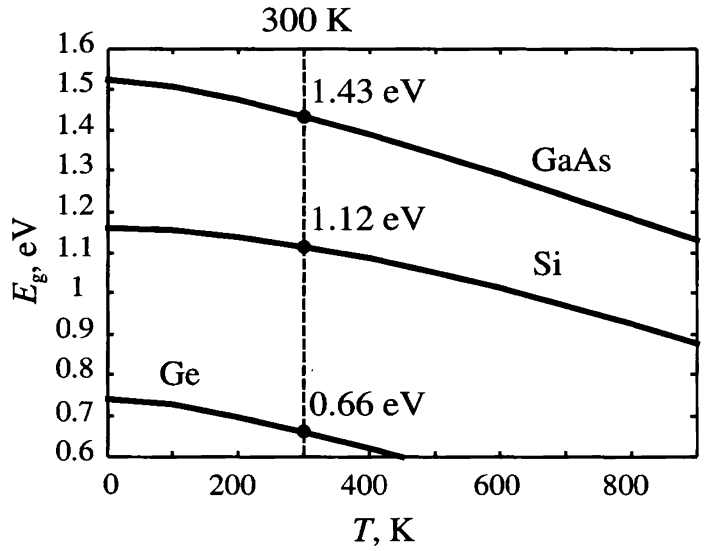


Figura 1.16 Andamento della ampiezza della banda proibita E_g al variare della temperatura in tre materiali semiconduttori.

è maggiore per tutte le temperature nell'arseniuro di gallio, seguita dal silicio e dal germanio. A questo corrisponde una concentrazione intrinseca minore nell'arseniuro di gallio e maggiore nel germanio. Il valore del silicio è intermedio tra i due. I valori della ampiezza di banda e della concentrazione intrinseca per i principali semiconduttori si trovano tabulati alla temperatura di 300 K, ovvero a temperatura ambiente, e possono poi essere riscalati per temperature diverse, come illustrato nell'esempio 1.1. È comunque opportuno ricordare almeno i valori per il silicio a temperatura ambiente: $n_i = 1,45 \times 10^{10}\,\mathrm{cm}^{-3}$ e $E_g = 1,12\,\mathrm{eV}$.

Esempio 1.1 Valutare la concentrazione intrinseca nel Si a $T_1 = 400\,\mathrm{K}$.
Un modello empirico che descrive la dipendenza dalla temperatura dell'ampiezza della banda proibita del Si ha l'espressione matematica

$$E_g(T) = E_{g0} - \frac{\alpha T^2}{T + \beta}$$

dove $E_{g0} = 1,165\,\mathrm{eV}$, $\alpha = 7,02 \times 10^{-4}\,\mathrm{eV/K}$ e $\beta = 1108\,\mathrm{K}$. Naturalmente, nell'espressione precedente T deve essere espressa in K. Alla temperatura T_1 richiesta, si ha quindi

$$E_g(T_1) = E_{g0} - \frac{\alpha T_1^2}{T_1 + \beta} = 1,09\,\mathrm{eV}$$

Per calcolare la concentrazione intrinseca a T_1, si utilizza la (1.8) che può essere espressa nella forma:

$$n_i(T) = AT^{3/2} \exp\left(-\frac{E_g(T)}{2k_B T}\right)$$

dove A è una costante indipendente dalla temperatura.[3] Poiché a $T_0 = 300\,\mathrm{K}$ è noto il valore $n_i(T_0) = 1,45 \times 10^{10}\,\mathrm{cm}^{-3}$, valutando il rapporto $n_i(T_1)/n_i(T_0)$ dall'espressione precedente si trova:

$$n_i(T_1) = n_i(T_0)\left(\frac{T_1}{T_0}\right)^{3/2} \exp\left(-\frac{E_g(T_1)}{2k_BT_1} + \frac{E_g(T_0)}{2k_BT_0}\right) = 7,5 \times 10^{12}\,\mathrm{cm}^{-3}$$

dove, allo stesso modo, si può valutare $k_BT_1 = (k_BT_0)(T_1/T_0) = 26 \times (4/3)\,\mathrm{meV} = 34,67\,\mathrm{meV}$.

1.6 Il drogaggio

Anche se la concentrazione intrinseca varia nei diversi materiali semiconduttori, il suo valore rimane comunque molto inferiore a quello della concentrazione di elettroni nei metalli. Si è visto ad esempio che nello stagno si arriva a concentrazioni di elettroni dell'ordine di $10^{22}\,\mathrm{cm}^{-3}$ e in generale nei materiali metallici si ha la banda di conduzione parzialmente piena di elettroni, che risultano in numero pari al prodotto del numero di atomi per il numero di elettroni di valenza per ogni atomo. Questi numeri così elevati ci fanno comprendere come i semiconduttori, per quanto abbiano concentrazioni di carica non trascurabili nelle bande, risultino essere caratterizzati comunque da conducibilità elettriche di diversi ordini di grandezza al di sotto di quelle dei metalli. L'interesse nei materiali semiconduttori è però legato alla possibilità di variare la concentrazione di portatori nelle bande (elettroni e lacune) mediante un processo tecnologico detto *drogaggio*, il cui fine è quindi essenzialmente quello di modificare in maniera controllata la conducibilità elettrica del materiale stesso. Dal punto di vista della nomenclatura, un semiconduttore puro e perfettamente cristallino viene detto *intrinseco*, mentre nel caso del materiale drogato si usa a volte anche l'attributo *estrinseco*.

La tecnologia del drogaggio consiste nel sostituire un certo numero di atomi di semiconduttore presenti nel reticolo cristallino con degli elementi di un gruppo della tavola periodica diverso dal IV. In particolare, si utilizzano tipicamente due tipologie dei atomi droganti:

- ▷ atomi aventi *più di quattro elettroni* sulla corteccia esterna, ad esempio elementi del V gruppo quali l'arsenico As;
- ▷ atomi aventi *meno di quattro elettroni* sulla corteccia esterna, ad esempio elementi del III gruppo quali il boro B.

Naturalmente il numero totale di atomi droganti che sostituiscono un atomo di semiconduttore deve essere trascurabile rispetto al numero di atomi di semiconduttore stesso, in modo che gli elementi droganti possano essere considerati una piccola perturbazione della struttura cristallina stessa.

Considerando inizialmente il drogaggio da parte di un elemento del V gruppo, ad esempio As nel Si, si verifica quanto segue. Dei cinque elettroni sulla corteccia esterna dell'As, quattro vengono impegnati nella costituzione dei legami covalenti con i quattro atomi di Si più vicini al drogante, mentre il quinto risulta trovarsi, se l'elemento drogante è stato scelto con oculatezza, su un livello energetico inferiore ma molto vicino al minimo della banda di conduzione E_c, intendendo con ciò una distanza dell'ordine di qualche decina di meV. Ciò implica che anche la sola energia termica fornita all'elettrone dall'ambiente esterno può essere sufficiente, almeno a temperatura ambiente, a liberare

[3] Per una espressione completa, si veda la successiva equazione (1.27).

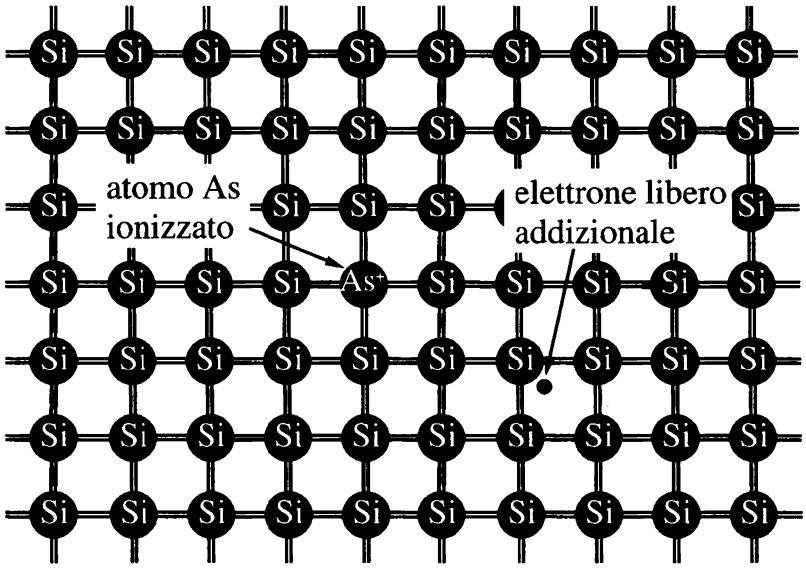


Figura 1.17 Effetto dell'introduzione di un atomo di As nel reticolo cristallino del Si.

l'elettrone verso la banda di conduzione: si parla, pertanto, di atomo drogante *donatore*, in quanto fornisce alla banda di conduzione degli elettroni liberi di muoversi nel cristallo (si veda la figura 1.17). Si dice anche che si è effettuato un drogaggio di *tipo n*, poiché in questo caso si incrementa il numero di elettroni liberi nel materiale. Naturalmente, se l'atomo donatore perde uno dei suoi elettroni, non è più elettricamente neutro: esso risulta essere una carica fissa *ionizzata*, e caratterizzato da una carica elettrica positiva pari a $+q$.

Il caso del drogaggio mediante un atomo del III gruppo è presentato nella figura 1.18, con riferimento ad un atomo di B inserito in un cristallo di Si. Un elemento del III gruppo è caratterizzato dalla presenza, sulla corteccia esterna, di tre elettroni. Essi, pertanto, risultano essere impegnati nella costituzione di tre legami covalenti che vincolano l'atomo di boro al reticolo cristallino. La scelta dell'elemento del III gruppo deve essere condotta in modo tale che esso, una volta inserito nel semiconduttore che si intende drogare, renda disponibile un livello energetico per un elettrone che sia vicino (ovvero, al di sopra e ad una distanza di qualche decina di meV) al massimo della banda di valenza E_V. In questo caso, a temperatura ambiente la sola energia termica può essere sufficiente perché l'atomo di B venga ionizzato assorbendo un elettrone dalla banda di valenza del semiconduttore: l'atomo diviene una carica fissa negativa pari a $-q$, e nella banda di valenza del semiconduttore viene liberata una lacuna in grado di muoversi nel cristallo. Si parla di atomi di tipo *accettatore*, in quanto assorbono un elettrone dalla banda di valenza, e di drogaggio di *tipo p*.

Dalla discussione precedente, emerge chiaramente come la caratteristica fondamentale richiesta ad un elemento del V o III gruppo per essere un buon atomo drogante è che ad esso, una volta inserito nel reticolo cristallino del semiconduttore, corrispondano dei livelli energetici permessi per gli elettroni nella banda proibita che siano il più vicino

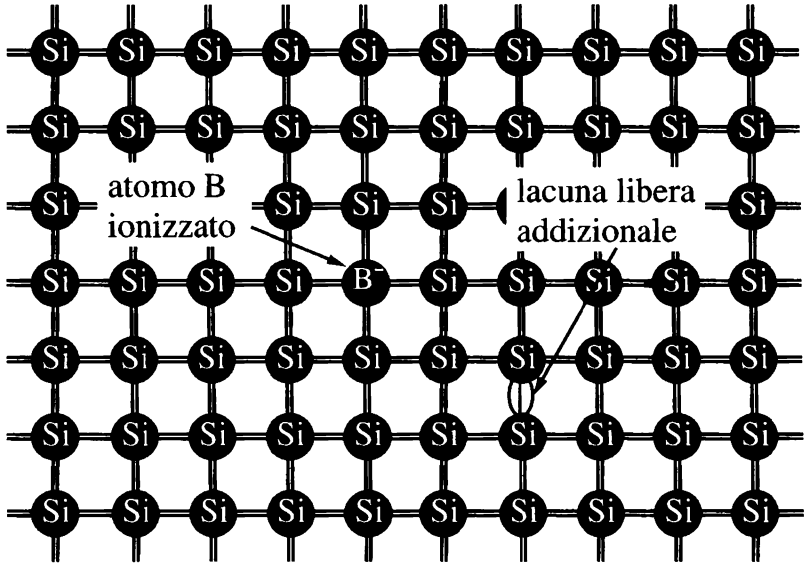


Figura 1.18 Effetto dell'introduzione di un atomo di B nel reticolo cristallino del Si.

possibile alla banda di conduzione, per gli atomi donatori, o a quella di valenza, per gli atomi accettatori. Nel caso del Si, alcuni casi importanti sono rappresentati nella figura 1.19: nel caso degli atomi droganti donatori, si vede che il fosforo e l'arsenico sono dei buoni elementi per il drogaggio di tipo n, mentre il boro e l'alluminio sono dei buoni atomi driganti accettatori. Si può anche notare come l'oro presenti alcune caratteristiche particolari: innanzitutto, ad esso corrispondono due livelli energetici permessi nella banda proibita, entrambi di tipo accettatore, cioè in grado di ospitare un elettrone. In particolare, di grande importanza pratica è il livello vicino al centro della banda proibita, detto anche livello *profondo,* in quanto è possibile dimostrare come tali livelli siano particolarmente efficienti nel favorire la ricombinazione delle cariche libere (si veda il successivo paragrafo 2.2.2). Per contrapposizione, gli atomi con livelli energetici vicini alla banda di riferimento (la banda di conduzione per i donatori, quella di valenza per gli accettatori) vengono anche detti livelli *superficiali.* Per ulteriori approfondimenti, si veda [2].

1.7 Calcolo della concentrazione di carica libera in equilibrio termodinamico

La determinazione della concentrazione di carica libera di muoversi in un semiconduttore, sia esso intrinseco o drogato, assume una grande rilevanza, in quanto è intuitivo che ad essa sono strettamente correlate le proprietà elettriche del materiale. In particolare, ci si attende che la conducibilità elettrica cresca all'aumentare della concentrazione di carica libera: questa affermazione verrà giustificata da un punto di vista quantitativo mediante la trattazione sviluppata nel paragrafo 2.1.

Sebbene nelle applicazioni pratiche ad un dispositivo a semiconduttore sono sempre

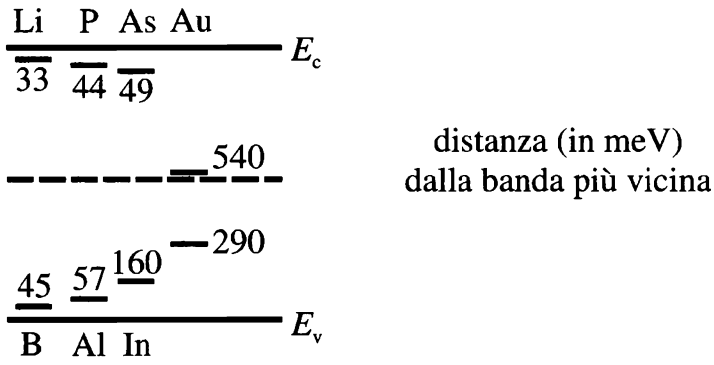


Figura 1.19 Posizione dei livelli energetici corrispondenti ad alcuni atomi droganti nel Si.

applicati dei segnali elettrici esterni, è opportuno iniziare a studiare i semiconduttori in condizioni di *equilibrio termodinamico,* ovvero assumendo che il materiale sia completamente isolato dal resto dell'ambiente: in altri termini, la condizione di equilibrio termodinamico consiste nell'assenza di qualunque scambio energetico con l'esterno del materiale stesso. Indicando la concentrazione volumica di elettroni in banda di conduzione, ovvero il numero di elettroni per unità di volume, con n, e la concentrazione volumica di lacune in banda di valenza con il simbolo p (in entrambi i casi l'unità di misura è cm^{-3}), è conveniente esprimere tali concentrazioni partendo dalla relativa *funzione di distribuzione in energia* $d(E)$, essendo E l'energia totale della particella. Il significato fisico di d è espresso dalla sua definizione matematica:

$$\mathrm{d}n = d_n(E)\mathrm{d}E \qquad \mathrm{d}p = d_p(E)\mathrm{d}E \tag{1.9}$$

ovvero, la funzione di distribuzione in energia rappresenta il numero di portatori liberi per unità di volume e di energia aventi energia totale compresa nell'intervallo di estremi E ed $E+\mathrm{d}E$. Dalla definizione (1.9) è immediato ricavare che le densità di carica libera si esprimono come l'integrale della rispettiva funzione di distribuzione in energia esteso a tutte le energie disponibili nella relativa banda di energie permesse, ovvero la banda di conduzione per gli elettroni e la banda di valenza per le lacune. Poiché, come si vedrà tra breve, la funzione di distribuzione decresce esponenzialmente verso lo zero mano a mano che ci si allontana dal limite della banda considerata, è intuibile come sia numericamente accettabile approssimare l'estensione della banda con una quantità infinita:

$$n = \int_{E_c}^{+\infty} d_n(E)\mathrm{d}E \qquad p = \int_{-\infty}^{E_v} d_p(E)\mathrm{d}E \tag{1.10}$$

Tale approssimazione consentirà di valutare gli integrali presenti nella (1.10) in modo un po' più semplice.

Per esplicitare le (1.10) occorre, naturalmente, valutare le funzioni densità in energia: queste si possono esprimere come il prodotto dei due fattori seguenti

- il numero di *stati* $N_n(E)$ e $N_p(E)$ che possono essere occupati nella relativa banda per unità di energia e di volume: tale funzione viene chiamata *densità degli stati*;

▷ la *probabilità di occupazione* $f_n(E)$ e $f_p(E)$ associata ad ognuno di tali stati, ovvero la probabilità che ha uno stato di energia E di essere occupato da un portatore

$$d_n(E) = N_n(E) f_n(E) \qquad d_p(E) = N_p(E) f_p(E) \tag{1.11}$$

La valutazione dettagliata dei due fattori che compongono la funzione di distribuzione in energia richiede l'uso della meccanica quantistica, e pertanto esula gli scopi di questo testo. I dettagli possono essere trovati, ad esempio, nel riferimento [2]; noi ci limiteremo a riportare i risultati.

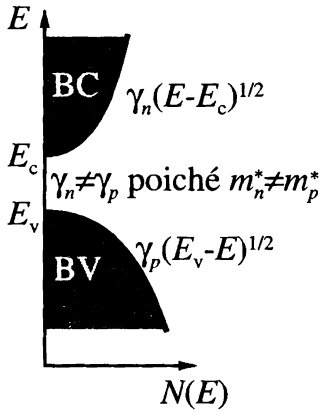


Figura 1.20 Rappresentazione delle densità degli stati in un semiconduttore.

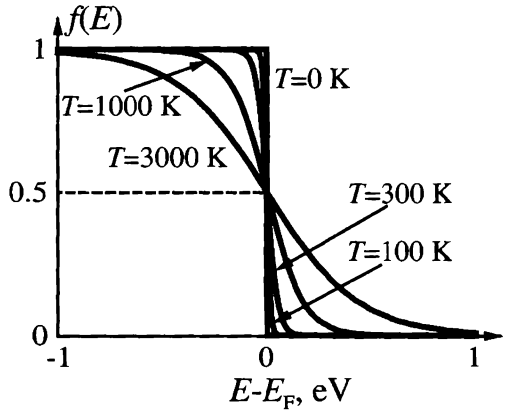


Figura 1.21 Distribuzione di Fermi Dirac per diversi valori di temperatura.

Per quanto riguarda la densità degli stati, essa risulta essere proporzionale alla radice quadrata dell'energia cinetica (l'energia totale meno quella potenziale) dei portatori liberi:

$$N_n(E) = \gamma_n \sqrt{E - E_c} \qquad N_n(E) = \gamma_p \sqrt{E_v - E} \tag{1.12}$$

dove i coefficienti γ dipendono dalla massa efficace di elettroni e lacune:

$$\gamma_n = \frac{4\pi}{h^3}(2m_n^*)^{3/2} \qquad \gamma_p = \frac{4\pi}{h^3}(2m_p^*)^{3/2} \tag{1.13}$$

essendo $h = 6,625 \times 10^{-34}$ J s $= 4,135 \times 10^{-15}$ eV s la costante di Planck. Si noti che nella (1.12), rappresentata graficamente nella figura 1.20, l'inversione di segno per le energie presente nella densità degli stati per le lacune è determinata dal fatto che l'asse delle energie E utilizzato è riferito agli elettroni, mentre le lacune sono cariche di segno opposto e quindi presentano un asse delle energie capovolto. Inoltre, la differenza tra le masse efficaci di elettroni e lacune[4] è responsabile della differente concavità delle due curve.

[4] Si ricorda che normalmente $m_p^* > m_n^*$.

Il secondo fattore nella (1.11), la probabilità di occupazione, può essere valutato in modo semplice solo in condizioni di equilibrio termodinamico. In questo caso, a partire dal fatto che gli elettroni sono particelle elementari dotate di un numero quantico di spin semi-intero (infatti possono assumere spin pari solo a $\pm 1/2$), è possibile dimostrare che si tratta di particelle quantistiche che seguono la *statistica di Fermi Dirac* $f(E)$

$$f_n(E) = f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} \tag{1.14}$$

dove E_F è un valore di energia caratteristico detto *energia di Fermi* o, più spesso nella teoria dei semiconduttori, *livello di Fermi*. Per quanto riguarda la probabilità di occupazione per le lacune, è sufficiente osservare come uno stato occupato da una lacuna corrisponda ad uno stato non occupato da un elettrone. Pertanto, sempre in equilibrio termodinamico:

$$f_p(E) = 1 - f(E) = \frac{\exp\left(\dfrac{E - E_F}{k_B T}\right)}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} = \frac{1}{1 + \exp\left(-\dfrac{E - E_F}{k_B T}\right)} \tag{1.15}$$

La funzione di Fermi Dirac (1.14) è rappresentata in figura 1.21 per diversi valori della temperatura:[5] dal grafico è immediato identificare il significato fisico del livello di Fermi E_F, ovvero il valore di energia che ha sempre probabilità del 50% di essere occupato, qualunque sia il valore di T. Inoltre, la probabilità di occupazione è una funzione monotona decrescente dell'energia, con valori che diminuiscono da 1, per basse energie, verso lo 0 per energie elevate. La rapidità con cui avviene tale transizione è tanto più elevata quanto più la temperatura è bassa. Una semplica analisi della (1.14), basata sul fatto che la temperatura assoluta T è una variabile non negativa, consente di dimostrare come per $T = 0\,\mathrm{K}$ $f(E)$ degeneri in una funzione costante a tratti:

$$f(E) = \begin{cases} 1 & \text{per } E < E_F \\ 1/2 & \text{per } E = E_F \\ 0 & \text{per } E > E_F \end{cases} \tag{1.16}$$

È così possibile dare una seconda interpretazione fisica ad E_F: si tratta del massimo valore di energia che un elettrone è in grado di occupare (in equilibrio termodinamico) alla temperatura dello zero assoluto. Si noti, infine, come alla temperatura ambiente $T = 300\,\mathrm{K}$ la transizione della funzione di Fermi Dirac, pur non essendo ideale come per $T = 0\,\mathrm{K}$, è comunque piuttosto rapida: ciò significa che, anche a temperatura ambiente, la gran parte degli elettroni occupano gli stati con energia $E \leq E_F$.

Una volta valutati i fattori che compongono la funzione di distribuzione in energia,

[5] Si osservi che in condizioni di equilibrio termodinamico l'assenza di scambi energetici tra il semiconduttore e l'ambiente implica anche che la temperatura del materiale debba essere spazialmente uniforme, e pari alla temperatura dell'ambiente esterno stesso.

è possibile sostituirli nella (1.10) ottenendo:

$$n = \int_{E_c}^{+\infty} \gamma_n \sqrt{E - E_c}\, f(E)\, \mathrm{d}E \tag{1.17a}$$

$$p = \int_{-\infty}^{E_v} \gamma_p \sqrt{E_v - E}\, [1 - f(E)]\, \mathrm{d}E \tag{1.17b}$$

I due integrali in (1.17) non possono essere valutati in forma chiusa, ma è comunque possibile esprimerli in funzione di una funzione speciale detta *integrale di Fermi di ordine* 1/2

$$\mathfrak{F}_{1/2}(x) = \int_0^{+\infty} \frac{\sqrt{\alpha}}{1 + \exp(\alpha - x)}\, \mathrm{d}\alpha \tag{1.18}$$

che può essere valutata solo numericamente. Come mostrato nel successivo approfondimento 1.1, le (1.17) si esprimono

$$n = \frac{2}{\sqrt{\pi}} N_c\, \mathfrak{F}_{1/2}\left(\frac{E_F - E_c}{k_B T}\right) \qquad p = \frac{2}{\sqrt{\pi}} N_v\, \mathfrak{F}_{1/2}\left(\frac{E_v - E_F}{k_B T}\right) \tag{1.19}$$

dove si sono definite le *densità efficaci degli stati* in banda di conduzione e di valenza (unità di misura: cm^{-3}):

$$N_c = 2\frac{(2\pi m_n^* k_B T)^{3/2}}{h^3} \qquad N_v = 2\frac{(2\pi m_p^* k_B T)^{3/2}}{h^3} \tag{1.20}$$

Sebbene le relazioni (1.19) non siano espresse in forma chiusa, ovvero in termini di funzioni elementari, esse esprimono come, in condizioni di equlibrio termodinamico, le concentrazioni di carica libera in un semiconduttore siano direttamente correlate con la posizione del livello di Fermi. Inoltre, poiché n, p ed E_F sono tre variabili poste in relazione, per il momento, dalle sole due relazioni (1.19), la loro determinazione esplicita richiede di definire una terza equazione indipendente dalle precedenti.

Una analisi dell'andamento della funzione integrale di Fermi, mostrato nella figure 1.22, suggerisce una possibile semplificazione alle (1.19): si nota, infatti, che per valori negativi dell'argomento x, la funzione integrale di Fermi di ordine 1/2 è ben approssimata da una funzione esponenziale:

$$\mathfrak{F}_{1/2}(x) \approx \frac{\sqrt{\pi}}{2} \exp(x) \qquad x < 0 \tag{1.21}$$

dove l'approssimazione è tanto migliore quanto più x è negativo ed elevato in valore assoluto. Esaminando le (1.19), si osserva che l'argomento di $\mathfrak{F}_{1/2}$ è negativo se $E_F - E_c < 0$, per gli elettroni, e $E_v - E_F < 0$, per le lacune: pertanto, nel caso in cui il livello di Fermi E_F si trovi *all'iinterno della banda proibita* del semiconduttore, è possibile

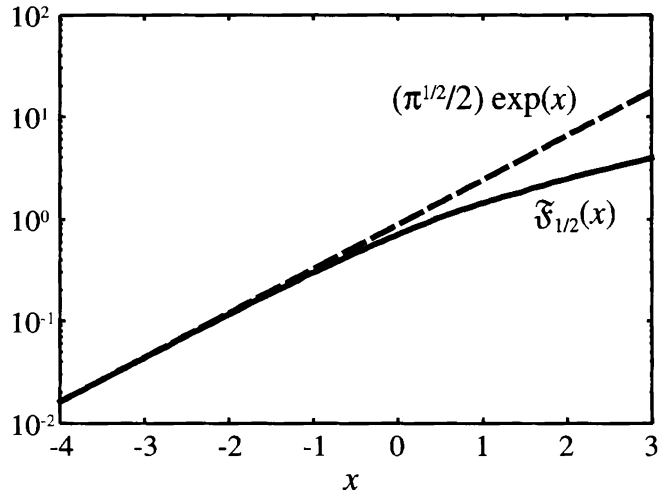


Figura 1.22 Rappresentazione della funzione integrale di Fermi di ordine 1/2, e confronto con la funzione $(\sqrt{\pi}/2)\exp(x)$.

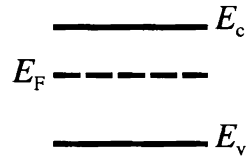


Figura 1.23 Posizione del livello di Fermi per un semiconduttore non degenere.

semplificare le (1.19) facendo uso della (1.21), ottenendo

$$n = N_c \exp\left(-\frac{E_c - E_F}{k_B T}\right) \qquad p = N_v \exp\left(-\frac{E_F - E_v}{k_B T}\right) \tag{1.22}$$

La (1.22), che esprime in forma chiusa la relazione tra la concentrazione di carica libera in banda di conduzione e di valenza e la posizione del livello di Fermi, viene anche chiamata *approssimazione di Boltzmann* per la (1.19): la ragione di questa denominazione verrà chiarita nell'approfondimento 1.1, dove si ricavano le (1.22) facendo uso di considerazioni di carattere fisico. Un semiconduttore nel quale il livello di Fermi si trova all'interno della banda proibita (si veda la figura 1.23), per il quale valgono quindi le (1.22), viene detto *non degenere*. Per contrapposizione, se E_F si trova molto vicino ad uno dei due limiti di banda, o addirittura all'interno di una delle due bande, si parla di *semiconduttore degenere:* in questo caso, occorre utilizzare le (1.19).

Approfondimento 1.1 In questo approfondimento, si dimostrano le relazioni (1.19) e si mostra come, partendo dall'assunzione di materiale non degenere, si derivino le (1.22) sulla base di considerazioni di carattere fisico.

Si esplicita la (1.17a), ottenendo grazie alla (1.14):

$$n = \gamma_n \int_{E_c}^{+\infty} \frac{\sqrt{E - E_c}}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} \, dE$$

Grazie al cambiamento di variabile $\alpha = (E - E_c)/k_B T$, cui corrisponde il differenziale $dE = k_B T d\alpha$, e alla relazione $E - E_F = E - E_c + E_c - E_F$, si trasforma la relazione precedente in

$$n = \gamma_n (k_B T)^{3/2} \int_0^{+\infty} \frac{\sqrt{\alpha}}{1 + \exp\left(\alpha - \dfrac{E_F - E_c}{k_B T}\right)} \, d\alpha$$

che, per la (1.13) e la (1.21), si esprime in

$$n = \frac{4\pi}{h^3} (2 m_n^* k_B T)^{3/2} \mathfrak{F}_{1/2}\left(\frac{E_F - E_c}{k_B T}\right)$$

Confrontando con la (1.19), è immediato ricavare la (1.20).
La derivazione della relazione per le lacune procede in modo del tutto analogo: si esplicita la (1.17b) facendo uso di (1.15)

$$p = \gamma_p \int_{-\infty}^{E_v} \frac{\sqrt{E_v - E}}{1 + \exp\left(-\dfrac{E - E_F}{k_B T}\right)} \, dE$$

e si effettua il cambiamento di variabile $\beta = (E_v - E)/k_B T$, cui corrispondono il differenziale $dE = -k_B T d\beta$ e la relazione $E_F - E = E_F - E_v + E_v - E$, ottenendo

$$p = \gamma_p (k_B T)^{3/2} \int_0^{+\infty} \frac{\sqrt{\beta}}{1 + \exp\left(\beta - \dfrac{E_v - E_F}{k_B T}\right)} \, d\beta$$

Sempre per la (1.13) e la (1.21), la precedente equazione diviene

$$p = \frac{4\pi}{h^3} (2 m_p^* k_B T)^{3/2} \mathfrak{F}_{1/2}\left(\frac{E_v - E_F}{k_B T}\right)$$

cioè la (1.20).
La determinazione delle equazioni per il caso non degenere può essere effettuata, invece che sulla base dell'approssimazione (1.21) della funzione integrale di Fermi, partendo da considerazioni diverse. Infatti, nel caso si abbia a che fare con un semiconduttore non degenere, è immediato verificare che $E - E_F > 0$ in tutta la banda di conduzione, mentre $E - E_F < 0$ in tutta la banda di valenza, ovvero sull'intero dominio di integrazione delle funzioni di distribuzione in energia presenti nelle (1.17). In particolare, ciò ha delle conseguenze significative sulle probabilità di occupazione, in quanto nelle (1.14) e (1.15) il termine esponenziale a denominatore risulta essere molto maggiore di 1 nelle rispettive bande, per cui:

$$f_n(E) \approx \exp\left(-\frac{E - E_F}{k_B T}\right) \quad \text{per } E - E_F > 0$$

$$f_p(E) \approx \exp\left(+\frac{E - E_F}{k_B T}\right) \quad \text{per } E - E_F < 0$$

Pertanto, per un semiconduttore non degenere in equilibrio termodinamico le probabilità di occupazione presentano una dipendenza esponenziale dall'energia totale (normalizzata grazie alla divisione per $k_B T$): si tratta, in altri termini, della probabilità di occupazione di Boltzmann, che viene seguita da un gas di particelle classiche non interagenti [2]. Per questo motivo le (1.22) sono anche chiamate approssimazione di Boltzmann delle (1.19). Infatti, sostituendo le relazioni approssimate per f_n ed f_p nelle (1.17) si ottiene

$$n = \gamma_n \int_{E_c}^{+\infty} \sqrt{E - E_c} \exp\left(-\frac{E - E_F}{k_B T}\right) \, \mathrm{d}E$$

$$p = \gamma_p \int_{-\infty}^{E_v} \sqrt{E_v - E} \exp\left(+\frac{E - E_F}{k_B T}\right) \, \mathrm{d}E$$

che, grazie ai cambiamenti di variabile $\alpha = (E - E_c)/k_B T$ e $\beta = (E_v - E)/k_B T$ già utilizzati in precedenza, divengono:

$$n = \gamma_n (k_B T)^{3/2} \exp\left(-\frac{E_c - E_F}{k_B T}\right) \int_0^{+\infty} \sqrt{\alpha} \exp(-\alpha) \, \mathrm{d}\alpha$$

$$p = \gamma_p (k_B T)^{3/2} \exp\left(-\frac{E_F - E_v}{k_B T}\right) \int_0^{+\infty} \sqrt{\beta} \exp(-\beta) \, \mathrm{d}\beta$$

Dalle tabelle di integrali notevoli (ad esempio si veda [3]) si ricava

$$\int_0^{+\infty} \sqrt{\alpha} \exp(-\alpha) \, \mathrm{d}\alpha = \frac{\sqrt{\pi}}{2}$$

pertanto le relazioni precedenti conducono direttamente alle (1.22).

1.7.1 Il livello di Fermi intrinseco

In un semiconduttore intrinseco, come discusso nel paragrafo 1.5, il numero di elettroni e di lacune libere di muoversi coincide con la concentrazione intrinseca del materiale n_i. Corrispondentemente, il livello di Fermi E_{Fi} nel materiale prende il nome di *livello di Fermi intrinseco*. Assumendo che il materiale intrinseco sia non degenere, dalle (1.22) e dalla condizione $n = p = n_i$ si ottiene una equazione nella sola incognita E_{Fi}:

$$N_c \exp\left(-\frac{E_c - E_{Fi}}{k_B T}\right) = N_v \exp\left(-\frac{E_{Fi} - E_v}{k_B T}\right) \tag{1.23}$$

che può essere facilmente esplicitata:

$$E_{Fi} = \frac{E_c + E_v}{2} - \frac{k_B T}{2} \ln \frac{N_c}{N_v} \tag{1.24}$$

In altri i termini, il livello di Fermi intrinseco si trova spostato rispetto al centro della banda proibita (espresso da $(E_c + E_v)/2$) di una quantità proporzionale al logaritmo del rapporto tra le due densità efficaci degli stati. Visto che, come si evince dalla (1.20), N_c ed N_v differiscono solo per il valore della massa efficace delle cariche libere, il loro rapporto è un numero molto prossimo ad uno, il cui logaritmo è un numero molto

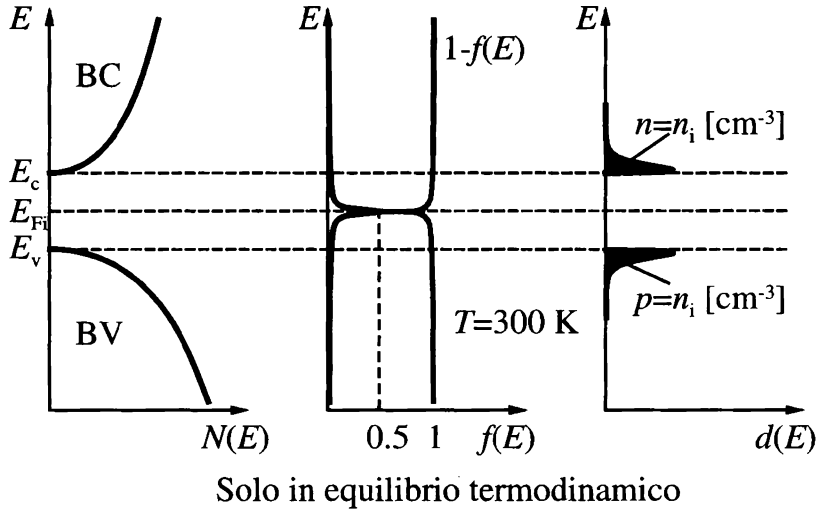


Figura 1.24 Densità degli stati (a sinistra), probabilità di occupazione (al centro) e funzione di distribuzione in energia (a destra) per un semiconduttore intrinseco in equilibrio termodinamico.

prossimo a zero. Pertanto, con buona approssimazione si ha

$$E_{Fi} \approx \frac{E_c + E_v}{2} \tag{1.25}$$

cioè il livello di Fermi intrinseco di un semiconduttore è, con buona approssimazione, situato al centro della banda proibita. Questa conclusione, inoltre, consente di validare l'assunzione fatta per la condizione di non degenerazione del materiale.

Una volta nota la posizione del livello di Fermi, è possibile rappresentare graficamente la densità degli stati disponibili, la probabilità di occupazione e la funzione di distribuzione in energia per un semiconduttore intrinseco in equilibrio termodinamico: tali grafici sono mostarti nella figura 1.24. Le probabilità di occupazione di elettroni e lacune, pari a $1/2$ per $E = E_{Fi}$ divengono molto piccole all'interno delle due bande, e quindi le curve rappresentate nella parte destra, la cui area corrisponde alle concentrazioni di carica libera in accordo alle (1.17), sono state rappresentate con una scala diversa da quella utilizzata per i fattori N ed f: in realtà, la gran parte dei portatori liberi è accumulata per valori di energia molto prossimi ad E_c, per gli elettroni, ed E_v, per le lacune.

Esempio 1.2 Nel Si alla temperatura $T = 300\,\mathrm{K}$ le densità efficaci degli stati nella banda di conduzione e di valenza valgono rispettivamente

$$N_c = 2,8 \times 10^{19}\,\mathrm{cm}^{-3} \qquad N_v = 1,04 \times 10^{19}\,\mathrm{cm}^{-3}$$

Sostituendo nella (1.24) si trova

$$E_{Fi} = \frac{E_c + E_v}{2} - 12,88\,\mathrm{meV}$$

Visto che l'ampiezza della banda proibita alla stessa temperatura è pari a $E_g = 1,12\,\text{eV}$, utilizzare l'approssimazione (1.25) corrisponde a commettere un errore pari a

$$\frac{12,88}{E_g/2} = \frac{12,88}{560} = 2,3\%$$

Avendo verificato che un semiconduttore intrinseco è effettivamente non degenere, moltiplicando tra loro le due espressioni (1.22) tenendo conto che $n = p = n_i$ conduce a

$$n_i^2 = N_c N_v \exp\left(-\frac{E_c - E_v}{k_B T}\right) \tag{1.26}$$

ovvero, per le (1.20) ed essendo $E_c - E_v = E_g$

$$n_i = 2\frac{(2\pi k_B T)^{3/2}}{h^3}(m_n^* m_p^*)^{3/4} \exp\left(-\frac{E_g}{2k_B T}\right) \tag{1.27}$$

1.7.2 Le equazioni di Shockley

Le equazioni (1.22), valide per un semiconduttore non degenere, possono essere espresse in una forma equivalente sulla base delle considerazioni che seguono. Nel caso particolare di un campione intrinseco dello stesso semiconduttore alla medesima temperatura, le (1.22) si scrivono:

$$n_i = N_c \exp\left(-\frac{E_c - E_{Fi}}{k_B T}\right) \qquad p_i = n_i = N_v \exp\left(-\frac{E_{Fi} - E_v}{k_B T}\right) \tag{1.28}$$

dalle quali si ricava

$$N_c = n_i \exp\left(+\frac{E_c - E_{Fi}}{k_B T}\right) \qquad N_v = n_i \exp\left(+\frac{E_{Fi} - E_v}{k_B T}\right) \tag{1.29}$$

Sostituendo queste espressioni nelle (1.22), si ottengono le *equazioni di Shockley*

$$n = n_i \exp\left(\frac{E_F - E_{Fi}}{k_B T}\right) \qquad p = n_i \exp\left(\frac{E_{Fi} - E_F}{k_B T}\right) \tag{1.30}$$

Le (1.30) hanno gli stessi limiti di validità delle approssimazioni di Boltzmann, e cioè possono essere utilizzate solo per semiconduttori non degeneri all'equilibrio termodinamico. Consentono, però, di dare una nuova interpretazione fisica al concetto di livello di Fermi. Nel caso di un semiconduttore drogato

- ▷ di tipo n, l'effetto del drogaggio è di aumentare la concentrazione n di elettroni liberi rispetto a n_i: dalla prima delle (1.30), ciò implica che $E_F > E_{Fi}$, ovvero che il livello di Fermi si sposta *nella metà superiore* della banda proibita, come mostrato nella figura 1.25. Contemporaneamente, la seconda delle (1.30) implica che $p < n_i$;
- ▷ di tipo p, in conseguenza del drogaggio p aumenta rispetto alla concentrazione intrinseca: per la seconda delle (1.30), ciò comporta che $E_F < E_{Fi}$, ovvero che

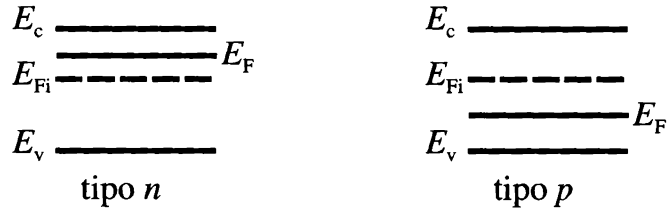


Figura 1.25 Posizione del livello di Fermi in un semiconduttore drogato di tipo n (a sinistra) e di tipo p (a destra).

il livello di Fermi si sposta *nella metà inferiore* della banda proibita (si veda la figura 1.25). Per la prima delle (1.30), inoltre, si ha che $n < n_i$.

In definitiva, il livello di Fermi costituisce una misura della quantità di elettroni presenti in equilibrio termodinamico nel materiale, essendo tanto più elevato quanto maggiore è n.

1.7.3 La legge dell'azione di massa

Si consideri un semiconduttore non degenere, per il quale valgono le equazioni di Shockley (1.30). Moltiplicando tra loro le due espressioni per n e p, si ha

$$np = n_i^2 \tag{1.31}$$

La (1.31), valida per un materiale non degenere in equilibrio termodinamico, viene detta *legge dell'azione di massa.* Essa stabilisce come non sia possibile, nelle condizioni precedenti, variare arbitrariamente le concentrazioni di elettroni e lacune libere: se, grazie al drogaggio, si aumenta il numero di elettroni, questo comporta una riduzione del numero di lacune, e viceversa.

1.8 Semiconduttori omogenei in equilibrio

Un caso particolarmente importante è quello che corrisponde ad un campione di semiconduttore drogato uniformemente, ovvero con una concentrazione costante nello spazio di atomi droganti. Consideriamo un semiconduttore omogeneo nel quale siano stati introdotti degli atomi droganti sia di tipo donatore, in concentrazione N_D (unità di misura: cm^{-3}), sia di tipo accettatore, in concentrazione N_A. Nel caso del Si, i valori più comuni per le concentrazioni di elementi droganti sono comprese tra circa $10^{14}\,\mathrm{cm}^{-3}$ e circa $10^{20}\,\mathrm{cm}^{-3}$.

Come discusso nel paragrafo 1.6, per un valore fissato della temperatura T ed in equilibrio termodinamico, parte degli atomi droganti risultano essere ionizzati: indicheremo con N_D^+ la concentrazione di atomi donatori ionizzati, ai quali corrisponde la carica positiva per unità di volume qN_D^+, e con N_A^- la densità di atomi accettatori ionizzati, ai quali corrisponde la densità di carica negativa $-qN_A^-$. Le altre cariche presenti nel sistema omogeneo sono quelle corrispondenti ai portatori liberi, ovvero una carica qp relativa alle lacune in banda di valenza ed una $-qn$ corrispondente agli elettroni in banda di conduzione.

In condizioni di equilibrio termodinamico, le (1.19) oppure, nel caso non degenere, le (1.22) costituiscono delle relazioni tra le concentrazioni di carica libera e il livello di Fermi E_F. Anche le concentrazioni di atomi droganti ionizzati possono essere messe in relazione ad E_F, infatti è possibile dimostrare che:

$$N_D^+ = \frac{N_D}{1 + g_D \exp\left(\frac{E_F - E_D}{k_B T}\right)} \qquad N_A^- = \frac{N_A}{1 + g_A \exp\left(\frac{E_A - E_F}{k_B T}\right)} \tag{1.32}$$

dove E_D ed E_A sono, rispettivamente, i livelli energetici corrispondenti agli atomi donatori e accettatori, mentre i coefficienti g_D e g_A vengono detti *fattore di degenerazione* per la corrispondente specie drogante.

Nel campione omogeneo in equilibrio termodinamico deve essere rispettata la condizione di *neutralità locale*, ovvero l'annullamento in ogni posizione nel campione della densità di carica totale ρ:

$$\rho = q\left(p - n + N_D^+ - N_A^-\right) = 0 \tag{1.33}$$

Sostituendo nella condizione di neutralità (1.19) e (1.32), si ottiene una equazione algebrica nella sola incognita E_F:

$$\frac{2}{\sqrt{\pi}} N_v \mathfrak{F}_{1/2}\left(\frac{E_v - E_F}{k_B T}\right) - \frac{2}{\sqrt{\pi}} N_c \mathfrak{F}_{1/2}\left(\frac{E_F - E_c}{k_B T}\right) + \frac{N_D}{1 + g_D \exp\left(\frac{E_F - E_D}{k_B T}\right)} - \frac{N_A}{1 + g_A \exp\left(\frac{E_A - E_F}{k_B T}\right)} = 0 \tag{1.34}$$

che, naturalmente, può essere risolta solo numericamente. A questo proposito, può essere opportuno notare come la soluzione numerica richieda di definire un riferimento per l'asse delle energie, in modo da poter assegnare alle variabili E_c, E_v, E_D, E_A, E_F un valore numerico. Come noto, la scelta del riferimento energetico è arbitraria. Pertanto, è possibile scegliere il punto nel quale $E = 0$ senza vincoli. Ad esempio, si potrebbe decidere di porre $E_v = 0\,\text{eV}$: con questa scelta, e con riferimento al Si a $T = 300\,\text{K}$, si ha di conseguenza $E_c = E_g = 1,12\,\text{eV}$. Inoltre, supponendo che l'elemento donatore sia As e quello accettatore sia B, dalla figura 1.19 si ricava $E_D = E_c - 49\,\text{meV} = 1,071\,\text{eV}$ ed $E_A = E_v + 45\,\text{meV} = 45\,\text{meV}$.

Una volta scelti il riferimento per le energie ed un valore per la temperatura T, la (1.34) può essere risolta numericamente, ottenendo il corrispondente valore di E_F che garantisce la condizione di neutralità locale. Noto E_F, sostituendo nelle (1.32) è possibile ricavare le concentrazioni di atomi droganti ionizzati o, equivalentemente, i coefficienti di ionizzazione N_D^+/N_D e N_A^+/N_A. Per dei valori tipici di drogaggio nel Si, la figura 1.26 mostra l'andamento del coefficiente di ionizzazione in funzione della temperatura. Si può notare come esso sia praticamente unitario per T maggiore di circa 250 K: si dice che si è in presenza di *ionizzazione completa* degli atomi droganti nel semiconduttore.

Naturalmente, conoscendo E_F è anche possibile, utilizzando (1.19), stimare le concentrazioni di carica libera. Assumendo di avere un campione di Si uniformemente

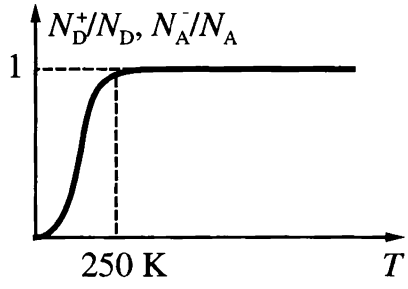


Figura 1.26 Andamento del coefficiente di ionizzazione degli atomi droganti in funzione della temperatura.

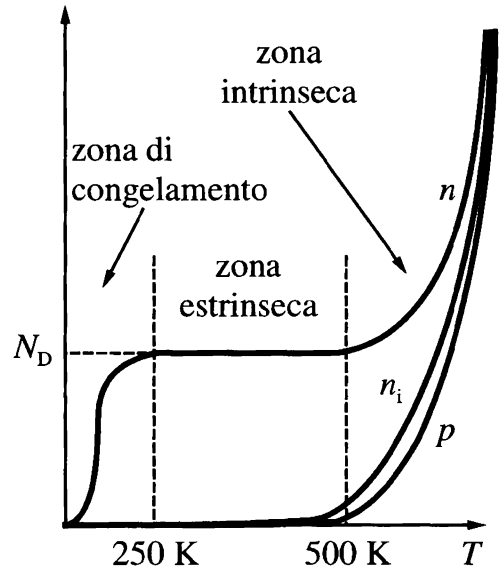


Figura 1.27 Andamento in funzione della temperatura delle concentrazioni di carica libera e della concentrazione intrinseca in un campione di Si drogato di tipo n.

drogato di tipo n con una concentrazione N_D, per valori tipici del drogaggio si ricava un andamento simile a quello riportato nella figura 1.27. In esso si possono identificare, in funzione della temperatura, tre regioni di funzionamento: la *zona di congelamento* per T inferiore a circa 250 K, la *zona intrinseca* per temperature elevate (superiori a $400 \div 500$ K), e la *zona estrinseca* per temperature intermedie. La regione di normale funzionamento è naturalmente la zona estrinseca, nella quale $n \approx N_D$ e la concentrazione di lacune è trascurabile.[6] Nella zona di congelamento, la ionizzazione degli atomi droganti non è completa, e questo giustifica la riduzione nel valore di n. Nella zona intrinseca, infine, n_i cresce fortemente, fino a superare il valore di N_D: le due concentrazioni di carica libera aumentano anch'esse tendendo al valore $n \approx p \approx n_i$, giustificando così il nome dato a questa regione di funzionamento.

1.8.1 Semiconduttore omogeneo non degenere

Se il semiconduttore omogeneo è anche non degenere, la validità della legge di azione di massa consente di semplificare il calcolo della concentrazione dei portatori liberi. Infatti, alla condizione di neutralità locale si aggiunge la (1.31):

$$\begin{cases} n - p = N_D^+ - N_A^- = N^+ \\ np = n_i^2 \end{cases} \tag{1.35}$$

[6] Assumendo di avere a che fare con un semiconduttore non degenere, dalla legge di azione di massa (1.31) segue $p = n_i^2/n$.

dove si è definito il drogaggio ionizzato netto N^+. Dalla prima equazione del sistema (1.35) segue un importantissimo risultato, indicato come *legge di compensazione:* la concentrazione di carica libera in equilibrio in un campione di semiconduttore omogeneo non dipende dai valori delle densità delle varie specie droganti, ma solo dalla loro somma algebrica. Ciò consente, ad esempio, di convertire una regione inizialmente drogata di tipo n in una drogata tipo p semplicemente aggiungendo un numero di atomi droganti accettatori superiore al drogaggio donatore iniziale. Alternativamente, è possibile ottenere un campione di materiale intrinseco introducendo una concentrazione uguale di atomi donatori ed accettatori.

Per risolvere il sistema (1.35) conviene considerare separatamente due casi:

- ▷ campione drogato n, per il quale $N^+ > 0$. Dalla legge di azione di massa si ricava $p = n_i^2/n$, per cui sostituendo nella prima equazione del sistema si ottiene una equazione algebrica di secondo grado nell'incognita n

$$n - \frac{n_i^2}{n} = N^+ \Longleftrightarrow n^2 - N^+ n - n_i^2 = 0 \tag{1.36}$$

Risolvendo l'equazione e prendendo la sola soluzione avente significato fisico, cioè quella che garantisce $n > 0$, si ricava

$$n = \frac{N^+}{2}\left[1 + \sqrt{1 + \left(\frac{2n_i}{N^+}\right)^2}\right] \tag{1.37}$$

- ▷ campione drogato p, per il quale $N^+ < 0$. Dalla (1.31) segue $n = n_i^2/p$, e sostituendo nella condizione di neutralità si ricava

$$\frac{n_i^2}{p} - p = N^+ = -|N^-| \Longleftrightarrow p^2 - |N^+|p - n_i^2 = 0 \tag{1.38}$$

da cui si trova

$$p = \frac{|N^+|}{2}\left[1 + \sqrt{1 + \left(\frac{2n_i}{|N^+|}\right)^2}\right] \tag{1.39}$$

Le due relazioni (1.37) e (1.39) possono essere significativamente semplificate per tutti quei valori di temperatura per i quali $n_i \ll |N^+|$. Infatti, in questo caso si ha:

$$\text{campione drogato } n \qquad n \approx N^+ \qquad p = \frac{n_i^2}{N^+} \tag{1.40a}$$

$$\text{campione drogato } p \qquad p \approx |N^+| \qquad n = \frac{n_i^2}{|N^+|} \tag{1.40b}$$

Le due relazioni precedenti dimostrano come in un campione dorgato con una concentrazione significativamente superiore alla concentrazione intrinseca del materiale, si

abbia

$$\text{campione drogato } n \qquad \frac{n}{p} = \left(\frac{N^+}{n_i}\right)^2 \gg 1 \tag{1.41a}$$

$$\text{campione drogato } p \qquad \frac{p}{n} = \left(\frac{|N^+|}{n_i}\right)^2 \gg 1 \tag{1.41b}$$

Nel caso tipico di un semiconduttore drogato con $10^{16}\,\text{cm}^{-3}$ atomi e con $n_i \approx 10^{10}\,\text{cm}^{-3}$, il rapporto è pari a circa 10^{12}: per questo motivo le cariche libere dello stesso tipo del drogaggio (elettroni per drogaggio n, lacune per drogaggio p) vengono chiamati *portatori maggioritari;* per contrapposizione, l'altra specie di cariche libere viene detta *portatori minoritari.*

Si noti che a temperatura ambiente, così come discusso in precedenza, la ionizzazione degli atomi droganti è praticamente completa, pertanto si presentano le seguenti possibilità che semplificano le (1.40):

▷ campione non degenere drogato n con $N_D \gg n_i$ e a temperatura ambiente:

$$n = N_D \qquad p = \frac{n_i^2}{N_D} \tag{1.42}$$

▷ campione non degenere drogato p con $N_A \gg n_i$ e a temperatura ambiente:

$$p = N_A \qquad n = \frac{n_i^2}{N_A} \tag{1.43}$$

Tenendo conto di questi risultati, e della discussione svolta nel paragrafo 1.7.1 sulla posizione del livello di Fermi, si possono facilmente ricavare gli andamenti qualitativi per la funzione di distribuzione in energia di elettroni e lacune in un campione non degenere drogato di tipo n (figura 1.28) e di tipo p (figura 1.29). Inoltre, la posizione del livello di Fermi può essere facilmente valutata se il campione soddisfa le condizioni che conducono alle (1.42) e (1.43):

▷ per un campione drogato n che soddisfi le condizioni per le (1.42), la prima delle (1.22) diviene

$$n = N_D = N_c \exp\left(-\frac{E_c - E_F}{k_B T}\right) \tag{1.44}$$

da cui si ricava

$$E_c - E_F = k_B T \ln\left(\frac{N_c}{N_D}\right) \tag{1.45}$$

▷ per un campione drogato p che soddisfi le condizioni per le (1.42), la seconda delle (1.22) diviene

$$p = N_A = N_v \exp\left(-\frac{E_F - E_v}{k_B T}\right) \tag{1.46}$$

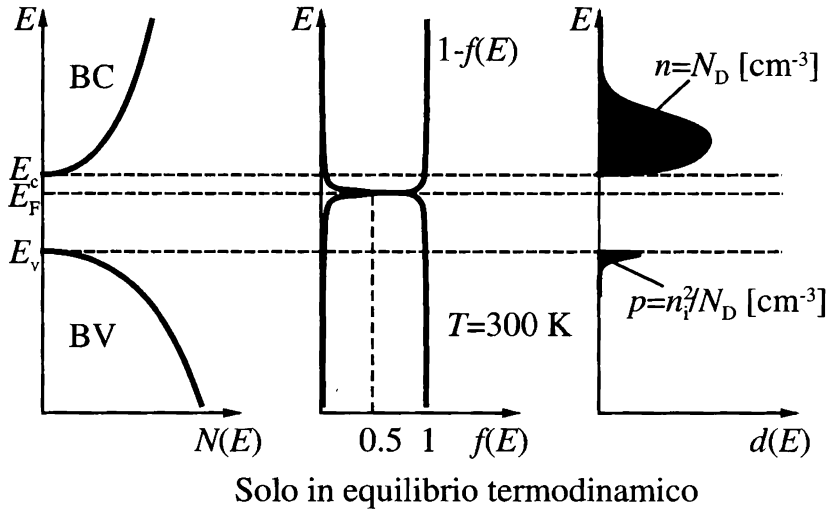


Figura 1.28 Densità degli stati (a sinistra), probabilità di occupazione (al centro) e funzione di distribuzione in energia (a destra) per un semiconduttore drogato n in equilibrio termodinamico.

da cui segue

$$E_F - E_v = k_B T \ln\left(\frac{N_v}{N_A}\right) \tag{1.47}$$

Le relazioni precedenti forniscono anche un criterio esplicito per valutare se un campione drogato sia o meno degenere: infatti, la condizione perché il livello di Fermi si trovi nella banda proibita risulta essere

$$\text{campione drogato } n \qquad N_D < N_c \tag{1.48a}$$

$$\text{campione drogato } p \qquad N_A < N_v \tag{1.48b}$$

ovvero, un limite superiore sul valore del drogaggio. Nel caso del Si a temperatura ambiente, le densità efficaci degli stati hanno un valore pari a qualche unità in $10^{19}\,\text{cm}^{-3}$ (si veda l'esempio 1.2), pertanto tale valore è anche il limite superiore al livello di drogaggio per poter considerare un campione di Si non degenere.

Esempio 1.3 Si considerino due campioni di Si a temperatura ambiente $T = 300\,\text{K}$. Entrambi siano uniformemente drogati, uno di tipo n con concentrazione $N_D = 10^{17}\,\text{cm}^{-3}$, l'altro di tipo p con concentrazione $N_A = 5 \times 10^{16}\,\text{cm}^{-3}$. Si richiede di calcolare, in entrambi i casi, le concentrazioni di carica libera in equilibrio termodinamico e la posizione del livello di Fermi.
Poiché a $T = 300\,\text{K}$ è possibile assumere la ionizzazione completa degli atomi droganti, nei due casi si ha, rispettivamente

$$N_D^+ = N_D \qquad N_A^- = N_A$$

Inoltre, i due livelli di drogaggio sono di almeno due ordini di grandezza inferiori alle densità

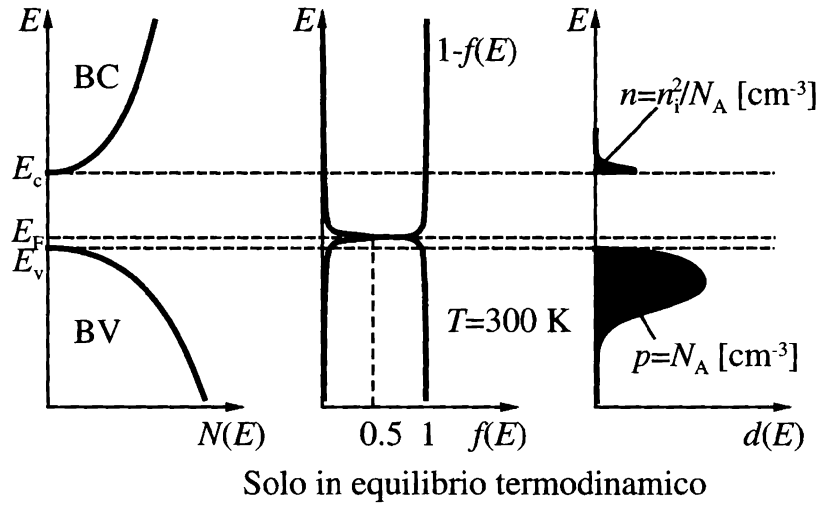


Figura 1.29 Densità degli stati (a sinistra), probabilità di occupazione (al centro) e funzione di distribuzione in energia (a destra) per un semiconduttore drogato p in equilibrio termodinamico.

efficaci degli stati nel Si a temperatura ambiente: di conseguenza, è possibile assumere che i due semiconduttori siano non degeneri. Grazie alla validità della legge di azione di massa, le concentrazioni di portatori maggioritari sono date dalla (1.37) per il campione di tipo n (dove $N^+ = N_D$), e dalla (1.39) per il semiconduttore drogato p (dove $|N^+| = N_A$). Inoltre, la concentrazione intrinseca a $T = 300\,\mathrm{K}$ vale $n_i = 1,45 \times 10^{10}\,\mathrm{cm}^{-3}$, per cui nei due casi

$$\left(\frac{2n_i}{N_D}\right)^2 = 8,4 \times 10^{-14} \ll 1 \qquad \left(\frac{2n_i}{N_A}\right)^2 = 3,4 \times 10^{-13} \ll 1$$

giustificando così l'uso delle (1.42) per il campione drogato n:

$$n = N_D = 10^{17}\,\mathrm{cm}^{-3} \qquad p = \frac{n_i^2}{N_D} = 2,1 \times 10^3\,\mathrm{cm}^{-3}$$

e delle (1.43) per il campione drogato p:

$$p = N_A = 5 \times 10^{16}\,\mathrm{cm}^{-3} \qquad n = \frac{n_i^2}{N_A} = 4,2 \times 10^3\,\mathrm{cm}^{-3}$$

Infine, per valutare la posizione del livello di Fermi si utilizzano le relazioni (1.45) e (1.47), ottenendo

$$E_c - E_F = k_B T \ln\left(\frac{N_c}{N_D}\right) = 146\,\mathrm{meV}$$

per il campione drogato n, mentre per il semiconduttore drogato p si ha

$$E_F - E_v = k_B T \ln\left(\frac{N_v}{N_A}\right) = 139\,\mathrm{meV}$$

Esempio 1.4 Si consideri un campione di GaAs a $T = 300\,\mathrm{K}$, per il quale $E_g = 1,42\,\mathrm{eV}$ e $n_i = 10^7\,\mathrm{cm}^{-3}$. Il semiconduttore è drogato uniformemente con una concentrazione $N_A = 10^{13}\,\mathrm{cm}^{-3}$ di atomi accettatori e $N_D = 10^{17}\,\mathrm{cm}^{-3}$ di atomi donatori. Assumendo $N_c \approx N_v$ e la completa ionizzazione degli atomi droganti, calcolare n, p e la posizione del livello di Fermi in equilibrio termodinamico.
Per la legge di compensazione, il semiconduttore risulta essere drogato di tipo n con una concentrazione netta $N^+ = N_D - N_A \approx N_D$, dove si è usata l'assunzione di completa ionizzazione e il fatto che N_A è quattro ordini di grandezza inferiore ad N_D. Pertanto, essendo il semiconduttore non degenere (per il livello di drogaggio) e $2n_i^2/N_D = 2 \times 10^{-3}$, si possono utilizzare le (1.42):

$$n = N_D = 10^{17}\,\mathrm{cm}^{-3} \qquad p = \frac{n_i^2}{N_D} = 10^{-3}\,\mathrm{cm}^{-3}$$

Per quanto riguarda la posizione del livello di Fermi, questa può essere valutata sia attraverso la (1.30), sia facendo uso della relazione (1.45). Utilizzando l'equazione di Shockley per la concentrazione di elettroni liberi, si ha

$$n = N_D = n_i \exp\left(\frac{E_F - E_{Fi}}{k_B T}\right)$$

da cui si ricava

$$E_F - E_{Fi} = k_B T \log\left(\frac{N_D}{n_i}\right) = 0,599\,\mathrm{eV}$$

Si noti, inoltre, che la condizione $N_c \approx N_v$ garantisce che il livello di Fermi intrinseco sia posizionato esattamente al centro della banda proibita, pertanto:

$$E_c - E_F = \frac{E_g}{2} - 0,599\,\mathrm{eV} = 0,111\,\mathrm{eV}$$

Allo stesso risultato si può arrivare utilizzando l'equazione (1.45), nella quale è però richiesto di conoscere il valore di N_c. La condizione $N_c \approx N_v$, unitamente alla (1.26), consente di stimare

$$N_c = n_i \exp\left(\frac{E_g}{2k_B T}\right) = 7,237 \times 10^{18}\,\mathrm{cm}^{-3}$$

Infine

$$E_c - E_F = k_B T \ln\left(\frac{N_c}{N_D}\right) = 0,111\,\mathrm{eV}$$

Esempio 1.5 Si consideri un campione di Si a $T = 300\,\mathrm{K}$, drogato uniformemente con $N_A = 10^{18}\,\mathrm{cm}^{-3}$ atomi accettatori e $N_D = 10^{17}\,\mathrm{cm}^{-3}$ atomi donatori. Sapendo che $n_i = 1,45 \times 10^{10}\,\mathrm{cm}^{-3}$ e che $N_c/N_v = 1,54$, determinare n, p, e la posizione del livello di Fermi.
Grazie alla legge di compensazione, si può concludere che si tratta di un campione drogato di tipo p con drogaggio netto $N^+ = N_D - N_A = -9 \times 10^{17}\,\mathrm{cm}^{-3}$. Il valore dei livelli di drogaggio permette di assumere che il semiconduttore sia non degenere, ed essendo $n_i \ll |N^+|$ si ha

$$p = |N^+| = 9 \times 10^{17}\,\mathrm{cm}^{-3} \qquad n = \frac{n_i^2}{|N^+|} = 233,6\,\mathrm{cm}^{-3}$$

Per quanto riguarda la posizione del livello di Fermi, si possono di nuovo seguire due strade. Una prima possibilità consiste nell'utilizzare l'equazione di Shockley (1.30)

$$p = |N^+| = n_i \exp\left(\frac{E_{Fi} - E_F}{k_B T}\right)$$

dalla quale segue

$$E_{Fi} - E_F = k_B T \log\left(\frac{|N^+|}{n_i}\right) = 0,467\,\text{eV}$$

Il livello di fermi intrinseco, d'altra parte, per la (1.24) si trova al di sotto del centro della banda proibita di $(k_B T/2)\log(N_c/N_v) = 5,6\,\text{meV}$. Pertanto,

$$E_F - E_v = \frac{E_g}{2} - 5,6\,\text{meV} - 467\,\text{meV} = 87,4\,\text{meV}$$

Per utilizzare direttamente la (1.43), invece, occorre valutare esplicitamente N_v. Sostituendo nella (1.26) la condizione $N_c = 1,54 N_v$, si ricava

$$N_v = \frac{n_i}{\sqrt{1,54}} \exp\left(\frac{E_g}{2k_B T}\right) = 2,64 \times 10^{19}\,\text{cm}^{-3}$$

Infine:

$$E_F - E_v = k_B T \ln\left(\frac{N_v}{|N^+|}\right) = 87,8\,\text{meV}$$

Naturalmente, la piccola differenza tra i due risultati dipende dalle approssimazioni numeriche introdotte nei calcoli.

Capitolo 2

Trasporto di carica nei semiconduttori

In questo capitolo verranno discussi i fenomeni relativi al trasporto della carica libera in un semiconduttore quando ad esso siano applicati dei campi elettrici esterni, ovvero in condizioni fuori dall'equilibrio termodinamico. Nel paragrafo 2.1 verranno trattate le possibili cause del movimento dei portatori liberi, il trascinamento da parte di un campo elettrico e la diffusione in caso di disuniformità spaziale della concentrazione di carica, definendo il valore delle corrispondenti densità di corrente di trascinamento (o conduzione) e di diffusione, e valutando la resistività di un campione uniformemente drogato. L'analisi delle cause che determinano, fuori equilibrio, le variazioni di concentrazione di carica libera nello spazio e nel tempo costituisce l'argomento del paragrafo 2.2, nella quale si definiscono, oltre alla rappresentazione a deriva-diffusione della corrente totale nel materiale, anche i concetti di concentrazione di carica libera in eccesso, e di alto e basso livello di iniezione. Infine, nel paragrafo 2.3 viene ricavata l'equazione di continuità dei portatori liberi e si descrive il modello matematico, detto a deriva-diffusione, che verrà utilizzato nel seguito per analizzare i dispositivi elettronici a semiconduttore.

2.1 Moto dei portatori liberi in un semiconduttore

L'analisi condotta nel capitolo 1 in condizioni di equilibrio termodinamico ha permesso di stimare, in un semiconduttore drogato, il numero di portatori liberi presenti nel materiale. Si sono così determinate le concentrazioni volumiche di elettroni n e di lacune p libere di muoversi nelle rispettive bande di conduzione e di valenza. Naturalmente, in condizioni di equilibrio termodinamico la densità di corrente totale attraverso una qualunque sezione nel materiale deve essere nulla. Poiché la corrente in un semiconduttore è determinata, almeno a frequenze sufficientemente basse da poter trascurare la corrente di spostamento, dal moto degli elettroni e delle lacune, parrebbe naturale dedurre dall'osservazione precedente che in condizioni di equilibrio termodinamico le cariche libere siano, individualmente, ferme. Nel caso si abbia a che fare con un volume V di semiconduttore tale da contenere un numero elevato di portatori liberi[1]

[1] Qualora si abbia a che fare con dispositivi estremamente miniaturizzati, con dimensioni caratteristiche dell'ordine del nanometro (nanoelettronica), il numero di portatori liberi è tanto ridotto da non consentire un'analisi statistica: si parla di dispositivi *mesoscopici*.

($N_n = nV$ elettroni e $N_p = pV$ lacune, nell'ipotesi che la densità di carica libera sia costante in V), tale affermazione è palesemente falsa, in quanto i due insiemi di cariche libere, sottoposte ad una temperatura costante $T \neq 0\,\mathrm{K}$, sono costituiti da particelle che individualmente si muovono con velocità istantanea $\boldsymbol{v}_{ni}$ e $\boldsymbol{v}_{pi}$ non nulla (determinata dall'agitazione termica associata a T). La condizione di corrente nulla è in realtà riferita al *valore medio* di tale corrente, dove la media viene valutata sui due insiemi di particelle; pertanto, ciò che risulta essere pari a zero è la *velocità media,* detta anche *velocità di trascinamento,* di elettroni e lacune:

$$\boldsymbol{v}_n = \frac{1}{N_n}\sum_{i=1}^{N_n}\boldsymbol{v}_{ni} \qquad \boldsymbol{v}_p = \frac{1}{N_p}\sum_{i=1}^{N_p}\boldsymbol{v}_{pi} \tag{2.1}$$

La discussione condotta nel capitolo 1 ha indicato come il moto degli elettroni e delle lacune nella struttura a bande del semiconduttore, che per essere studiato rigorosamente richiederebbe l'utilizzo della meccanica quantistica, possa essere, almeno approssimativamente, analizzato facendo uso delle leggi della cinematica classica, ovvero le equazioni di Newton, applicate ad una particella puntiforme *libera*[2] la cui massa, detta *massa efficace* m^*, dipende dal tipo di semiconduttore considerato. Senza entrare in ulteriori dettagli (si veda [2] per un approfondimento), si può dimostrare che l'approssimazione di massa efficace dipende strettamente dalla rigorosa periodicità spaziale del potenziale che gli elettroni subiscono, e quindi dalla natura cristallina del materiale. Sulla base di questa affermazione, è facile comprendere come qualunque interruzione della periodicità del cristallo vada a perturbare l'approssimazione di massa efficace, e quindi costituisca una interazione non descritta dall'approssimazione stessa. Tali interazioni vengono interpretate, nell'approssimazione di massa efficace, come *urti* della particella puntiforme: pertanto, il moto di una carica libera nel semiconduttore è costituito da una successione di *voli liberi,* che possono essere caratterizzati da un moto uniforme (se $\boldsymbol{\mathcal{E}} = \mathbf{0}$) o accelerato (se $\boldsymbol{\mathcal{E}} \neq \mathbf{0}$), e di *urti* contro tutto ciò che interrompe la periodicità del cristallo. In particolare, le interazioni principali che causano urti al moto delle particelle libere sono:

- ▷ le *imperfezioni reticolari,* ovvero i difetti strutturali del reticolo cristallino;
- ▷ gli *atomi di impurezze* eventualmente presenti nel cristallo, come gli atomi droganti o eventuali altri elementi presenti nel semiconduttore, siano essi stati introdotti intenzionalmente o no;
- ▷ le *vibrazioni reticolari* associate all'energia termica posseduta dagli atomi del cristallo. Le interazioni delle cariche libere con le vibrazioni reticolari possono essere interpretate [2] come degli urti tra le particelle libere stesse e delle particelle fittizie dette *fononi.*

[2] La particella è libera nel senso che non subisce più il potenziale del cristallo, il cui effetto è compreso in m^*, ma è sottoposta ad una forza nel caso nel materiale sia presente un campo elettrico.

2.1.1 Moto delle cariche libere per trascinamento da parte di un campo elettrico

Nel caso in cui nel materiale sia presente un campo elettrico,[3] le cariche libere vengono individualmente accelerate dalla forza esercitata su di loro da parte del campo. Poiché la forza $\boldsymbol{F}$ agente su una carica di valore Q sottoposta ad un campo elettrico $\boldsymbol{\mathcal{E}}$ è data da $\boldsymbol{F} = Q\boldsymbol{\mathcal{E}}$, le cariche verranno accelerate nella direzione del campo e con verso concorde per le lacune ($Q = +q$) e opposto per gli elettroni ($Q = -q$). È facile dimostrare (si veda l'approfondimento 2.1 più avanti) come la velocità di trascinamento dell'insieme di particelle sia data dalla relazione:

$$\boldsymbol{v}_n = -\mu_n \boldsymbol{\mathcal{E}} \qquad \boldsymbol{v}_p = \mu_p \boldsymbol{\mathcal{E}} \tag{2.2}$$

dove μ_n,$\mu_p > 0$ sono la *mobilità* (unità di misura: $\mathrm{cm^2\ V^{-1}\ s^{-1}}$) di elettroni e lacune.

Approfondimento 2.1 Si consideri la *quantità di moto* $\boldsymbol{q}_{ni} = m_n^* \boldsymbol{v}_{ni}$ e $\boldsymbol{q}_{pi} = m_p^* \boldsymbol{v}_{pi}$ associata, rispettivamente, al singolo elettrone e lacuna libera, e se ne definisca il valore medio per i due insiemi di cariche libere:

$$\boldsymbol{q}_n = \frac{1}{N_n}\sum_{i=1}^{N_n} \boldsymbol{q}_{ni} = m_n^* \boldsymbol{v}_n \qquad \boldsymbol{q}_p = \frac{1}{N_p}\sum_{i=1}^{N_p} \boldsymbol{q}_{pi} = m_p^* \boldsymbol{v}_p$$

Applicando un campo elettrico $\boldsymbol{\mathcal{E}}$ *stazionario*, cioè costante nel tempo, esso esercita sulle cariche una forza proporzionale a $\boldsymbol{\mathcal{E}}$. La dinamica di ogni particella soddisfa l'equazione di Newton per la quantità di moto:

$$\frac{\mathrm{d}\boldsymbol{q}_{ni}}{\mathrm{d}t} = -q\boldsymbol{\mathcal{E}} + \left.\frac{\mathrm{d}\boldsymbol{q}_{ni}}{\mathrm{d}t}\right|_{\text{urti}} \qquad \frac{\mathrm{d}\boldsymbol{q}_{pi}}{\mathrm{d}t} = +q\boldsymbol{\mathcal{E}} + \left.\frac{\mathrm{d}\boldsymbol{q}_{pi}}{\mathrm{d}t}\right|_{\text{urti}}$$

dove a destra si è indicata, oltre alla forza esercitata da $\boldsymbol{\mathcal{E}}$, la variazione di quantità di moto (per unità di tempo) determinata dai fenomeni di urto. Calcolando la media sull'insieme di particelle delle precedenti equazioni, e ricordando che:

$$\frac{1}{N_n}\sum_{i=1}^{N_n}(-q\boldsymbol{\mathcal{E}}) = -q\boldsymbol{\mathcal{E}} \qquad \frac{1}{N_p}\sum_{i=1}^{N_p}(+q\boldsymbol{\mathcal{E}}) = +q\boldsymbol{\mathcal{E}}$$

si ottiene l'equazione di evoluzione per la quantità di moto media dei due insiemi di particelle:

$$\frac{\mathrm{d}\boldsymbol{q}_n}{\mathrm{d}t} = -q\boldsymbol{\mathcal{E}} + \left.\frac{\mathrm{d}\boldsymbol{q}_n}{\mathrm{d}t}\right|_{\text{urti}} \qquad \frac{\mathrm{d}\boldsymbol{q}_p}{\mathrm{d}t} = +q\boldsymbol{\mathcal{E}} + \left.\frac{\mathrm{d}\boldsymbol{q}_p}{\mathrm{d}t}\right|_{\text{urti}}$$

In prima approssimazione [2], si può assumere per la variazione di quantità di moto nell'unità di tempo dovuta agli urti una relazione lineare con la quantità di moto media stessa:

$$\left.\frac{\mathrm{d}\boldsymbol{q}_n}{\mathrm{d}t}\right|_{\text{urti}} = -\frac{\boldsymbol{q}_n}{\tau_{qn}} \qquad \left.\frac{\mathrm{d}\boldsymbol{q}_p}{\mathrm{d}t}\right|_{\text{urti}} = -\frac{\boldsymbol{q}_p}{\tau_{qp}}$$

essendo τ_{qn} e τ_{qp} i *tempi di rilassamento della quantità di moto* per elettroni e lacune, rispettivamente. I tempi di rilassamento, in generale, dipendono dall'energia media posseduta dall'insieme di particelle cui si riferiscono, e sono tipicamente dell'ordine di qualche picosecondo.

[3] Tale campo elettrico può sia essere determinato da segnali elettrici applicati dall'esterno, sia essere presente anche in condizioni di equilibrio termodinamico a causa di una disomogeneità spaziale nel semiconduttore.

In condizioni stazionarie, tutte le variabili risultano essere indipendenti dal tempo, e quindi si può assumere nelle equazioni di Newton:

$$\frac{\mathrm{d}q_n}{\mathrm{d}t} = 0 \qquad \frac{\mathrm{d}q_p}{\mathrm{d}t} = 0$$

ovvero:

$$0 = -q\mathcal{E} - \frac{q_n}{\tau_{qn}} \qquad 0 = +q\mathcal{E} + \frac{q_p}{\tau_{qp}}$$

Ricordando la definizione di quantità di moto media ($q_n = m_n^* v_n$ e $q_p = m_p^* v_p$), dalle equazioni precedenti si ricava immediatamente l'espressione per la velocità di trascinamento:

$$v_n = -\frac{q\tau_{qn}}{m_n^*}\mathcal{E} = -\mu_n\mathcal{E} \qquad v_p = +\frac{q\tau_{qp}}{m_p^*}\mathcal{E} = +\mu_p\mathcal{E}$$

dove si sono definite le *mobilità* di elettroni e lacune:

$$\mu_n = \frac{q\tau_{qn}}{m_n^*} \qquad \mu_p = \frac{q\tau_{qp}}{m_p^*}$$

Come discusso in precedenza, tipicamente in un semiconduttore sono presenti diversi meccanismi di urto, ad esempio le interazioni con fononi, atomi di impurezza e imperfezioni reticolari. In generale ogni meccanismo di urto è caratterizzato da un proprio tempo di rilassamento $\tau_{q,k}$, e purché le interazioni si possano considerare indipendenti tra loro si ha:

$$\left.\frac{\mathrm{d}q_n}{\mathrm{d}t}\right|_{\text{urti}} = -\sum_k \frac{q_n}{\tau_{qn,k}} = -\frac{q_n}{\tau_{qn,\text{eq}}} \qquad \left.\frac{\mathrm{d}q_p}{\mathrm{d}t}\right|_{\text{urti}} = -\sum_k \frac{q_p}{\tau_{qp,k}} = -\frac{q_p}{\tau_{qp,\text{eq}}}$$

dove si sono definiti i tempi di rilassamento equivalenti:

$$\frac{1}{\tau_{qn,\text{eq}}} = \sum_k \frac{1}{\tau_{qn,k}} \qquad \frac{1}{\tau_{qp,\text{eq}}} = \sum_k \frac{1}{\tau_{qp,k}}$$

Utilizzando queste relazioni nell'espressione della mobilità, è immediato ricavare la *regola di Mathiessen* che esprime la mobilità equivalente totale determinata dalla coesistenza di diversi meccanismi di urto indipendenti in funzione della mobilità μ_k corrispondente alla sola presenza della k-esima interazione:

$$\frac{1}{\mu_n} = \sum_k \frac{1}{\mu_{n,k}} \qquad \frac{1}{\mu_p} = \sum_k \frac{1}{\mu_{p,k}}$$

Nella (2.2) la mobilità, come discusso nell'approfondimento 2.1, dipende dall'energia associata all'insieme delle particelle cui si riferisce. Pertanto, in presenza di un campo elettrico di modulo $\mathcal{E}$ che fornisca energia ad elettroni e lacune libere, è ragionevole attendersi una deviazione dalla relazione puramente lineare corrispondente alla (2.2) nella quale la mobilità sia una costante.

In effetti, una rilevazione sperimentale della cosiddetta *curva velocità-campo* del semiconduttore conduce ai grafici presentati nella figura 2.1, dove è rappresentato il modulo v della velocità di trascinamento in funzione di $\mathcal{E}$ per i principali semiconduttori, ad una temperatura di 300 K. Nei grafici, si possono osservare due regioni principali:

- ▷ per piccoli valori di $\mathcal{E}$, per entrambi i tipi di portatori liberi la relazione tra v ed $\mathcal{E}$ è effettivamente lineare: ciò significa che, per bassi valori del campo elettrico, la mobilità risulta essere una costante: si parla di *mobilità di basso campo;*

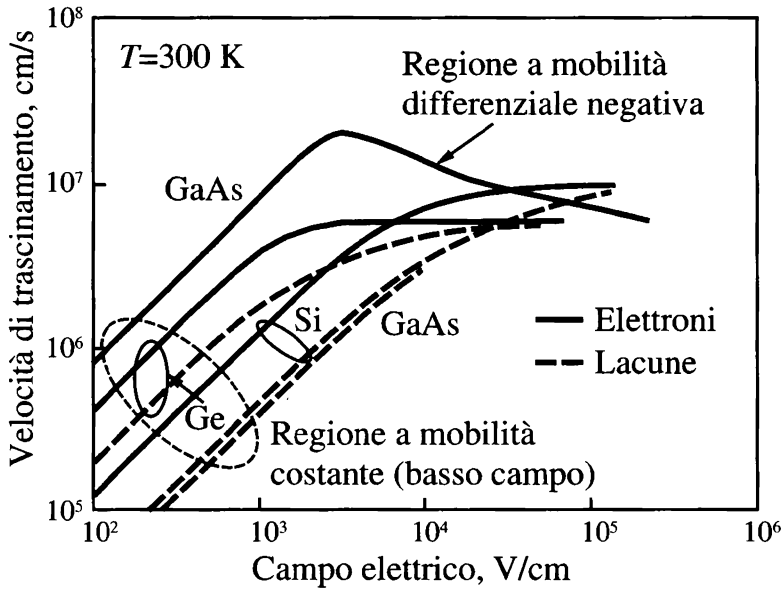


Figura 2.1 Andamento della relazione tra velocità di trascinamento e campo elettrico applicato per alcuni semiconduttori a temperatura ambiente ($T = 300\,\mathrm{K}$).

▷ per elevati valori di $\mathcal{E}$, la velocità di trascinamento tende ad un valore costante, indipendente dal campo applicato: si tratta del fenomeno della *saturazione di velocità*. Il valore della velocità di saturazione è, per tutti i principali semiconduttori, dell'ordine di grandezza di 10^7 cm/s.

Alcuni materiali, ad esempio il GaAs o l'InP, presentano per un certo intervallo di valori di campo elettrico una regione nella quale ad un aumento di $\mathcal{E}$ corrisponde una riduzione di velocità di trascinamento degli elettroni: si parla di *regione a mobilità differenziale negativa,* avendo definito la mobilità differenziale come:

$$\mu_{n,\mathrm{d}} = \frac{\mathrm{d}v_n}{\mathrm{d}\mathcal{E}} \tag{2.3}$$

Per contrapposizione, la mobilità $\mu_n = v_n/\mathcal{E}$ definita nella (2.2) viene anche detta *mobilità incrementale.*

Da un punto di vista quantitativo, si può osservare come, in generale, la mobilità di basso campo delle lacune risulti essere inferiore a quella degli elettroni, coerentemente con il fatto che la massa efficace m_p^* è, per questi materiali, superiore a m_n^*. Inoltre, la mobilità di basso campo per gli elettroni nel GaAs è di circa un ordine di grandezza superiore a quella nel Si, giustificando l'uso dell'arseniuro di gallio per realizzare dispositivi, basati sull'utilizzo della conduzione per elettroni, che lavorino ad elevate frequenze. Infine, è immediato verificare che la mobiltà di basso campo dei portatori liberi nel Ge è nettamente superiore a quella del Si: sebbene questo sia uno svantaggio per il Si, esso è ampiamente compensato dal maggiore sviluppo della tecnologia

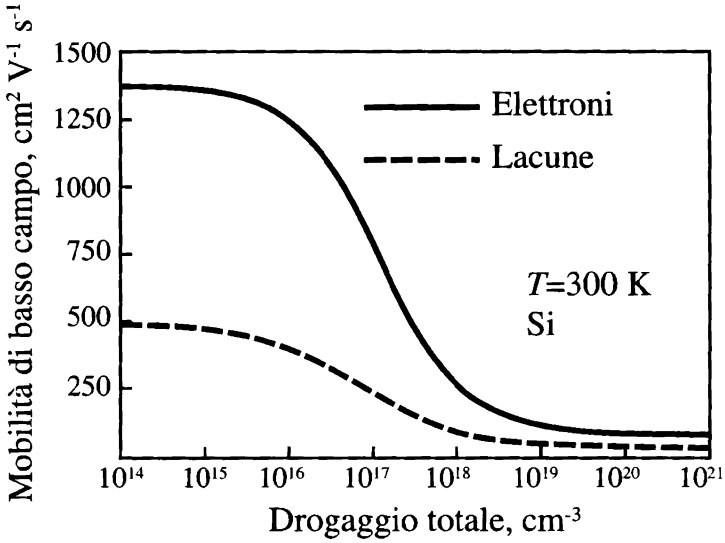


Figura 2.2 Dipendenza dal drogaggio totale della mobilità di basso campo di elettroni e lacune per il Si a $T = 300\,\mathrm{K}$.

per questo materiale e, soprattutto, dalla disponibilità di un *ossido nativo* (il biossido di silicio SiO_2) con eccellenti proprietà dielettriche, proprietà importantissima nella tecnologia dei circuiti integrati e per lo sviluppo del transistore MOS.

Dipendenza della mobilità dal drogaggio e dalla temperatura

Vista la dipendenza delle proprietà di trasporto dai meccanismi di urto delle particelle libere nel semiconduttore, è naturale attendersi una dipendenza della mobilità di basso campo da tutti quei parametri che possono influenzare le interazioni dei portatori liberi.

La presenza di atomi droganti nel semiconduttore determina, oltre alla variazione del numero di portatori liberi rispetto al materiale instrinseco, la possibilità di interazioni di elettroni e lacune con atomi di impurezze. Pertanto, ci si attende che la mobilità di basso campo sia una funzione decrescente della concentrazione di atomi droganti. Occorre però tenere presente che in caso si introducano diverse specie droganti, ad esempio una concentrazione N_D di atomi donatori e N_A di atomi accettatori, il valore della mobilità di basso campo viene determinato dal valore del drogaggio *totale* $N_t = N_D + N_A$, poiché le cariche libere interagiscono con *tutti* gli atomi di impurezze, indipendentemente dalla loro funzione nel semiconduttore. Si noti che questa proprietà rende peggiori le caratteristiche di trasporto di una regione drogata di semiconduttore ottenuta per *compensazione* tra specie droganti diverse: infatti, assumendo di avere una regione di Si drogata n a seguito di una compensazione tra drogaggi con $N_D > N_A$, dalla trattazione nel paragrafo 1.8 si ha, a temperatura ambiente, $n \approx N_D - N_A$, dove gli elettroni liberi hanno una mobilità di basso campo definita da N_t, minore di quella ottenibile drogando direttamente il campione con un drogaggio solo di tipo n con concentrazione $N'_D = N_D - N_A$. Sperimentalmente, su un campione di Si mantenuto alla temperatura ambiente $T = 300\,\mathrm{K}$ si può rilevare la dipendenza della mobilità di basso

campo di elettroni e lacune in funzione del drogaggio totale mostrata nella figura 2.2.

Un secondo meccanismo di interazione dei portatori liberi con il reticolo è costituito dagli urti con i fononi, che rappresentano le vibrazioni reticolari subite dagli atomi del reticolo a causa dello stato di agitazione termica nel quale si trovano a seguito della temperatura $T \neq 0\,\mathrm{K}$ alla quale sono sottoposti. Anche in questo caso, è ragionevole prevedere come la mobilità di basso campo sia una funzione decrescente di T, visto che ad un aumento della temperatura corrisponde un incremento delle vibrazioni reticolari stesse. Infatti, nel caso del Si la dipendenza dalla temperatura della mobilità di basso campo, mostrata nella figura 2.3 per diversi valori di drogaggio del campione, soddisfa tale aspettativa. Per temperature intorno al valore ambiente di 300 K e superiori, si può notare come la mobilità di basso campo segua un andamento proporzionale a $T^{-\alpha}$, con $\alpha \approx 2,2 \div 2,4$, ovvero:

$$\mu(T) \approx \mu(300\,\mathrm{K}) \left(\frac{T}{300\,\mathrm{K}} \right)^{-\alpha} \tag{2.4}$$

Legge di Ohm microscopica

In questo paragrafo si determina l'espressione della corrente di conduzione che attraversa un campione di semiconduttore, nel quale si trovi una concentrazione costante di carica libera, a seguito dell'applicazione di un campo elettrico costante.

Si consideri un volume elementare $\mathrm{d}V$ di semiconduttore, sottoposto ad un campo elettrico uniforme $\mathcal{E}$ e attraversato da una corrente stazionaria I nella direzione del campo (figura 2.4). Indicheremo con A la sezione trasversale al flusso di corrente, e con $\mathrm{d}s$ la lunghezza del campione elementare.

Per semplicità, si assumerà inizialmente che la corrente I sia determinata dal moto per trascinamento di una concentrazione n di elettroni liberi, costante nel volume elementare $\mathrm{d}V$. Per definizione, la corrente I è data dalla variazione $\mathrm{d}Q$ della carica contenuta nel volume riferita al tempo $\mathrm{d}t$ nel quale tale variazione avviene. Poiché la concentrazione di carica mobile è uniforme, si ha:

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}Q}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}Q}{\mathrm{d}V} A v_n \tag{2.5}$$

dove si sono usate la definizione di velocità di trascinamento degli elettroni $v_n = \mathrm{d}s/\mathrm{d}t$ e la relazione $\mathrm{d}V = A\mathrm{d}s$. Dalla (2.2) e dal fatto che la carica per unità di volume $\mathrm{d}Q/\mathrm{d}V$ coincide con $-qn$ segue:

$$I = -qnAv_n = qnA\mu_n\mathcal{E} \tag{2.6}$$

ovvero, essendo la densità di corrente $J = I/A$, la *legge di Ohm microscopica*:

$$J = qn\mu_n\mathcal{E} \tag{2.7}$$

Confrontando l'espressione ottenuta con la (1.1), è immediato identificare la conducibilità del materiale con:

$$\sigma = qn\mu_n. \tag{2.8}$$

Nel caso siano presenti, oltre a n elettroni per unità di volume, anche p lacune, si

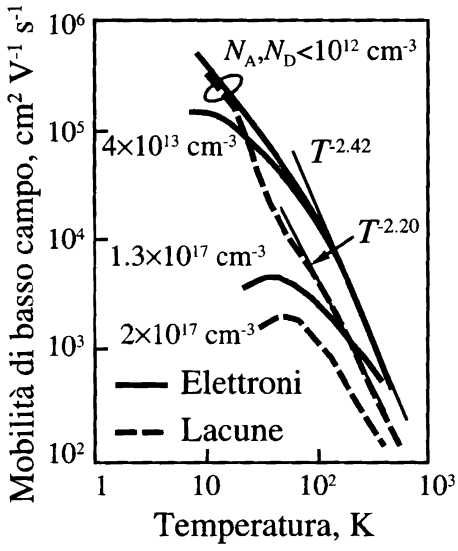


Figura 2.3 Dipendenza dalla temperatura della mobilità di basso campo di elettroni e lacune per il Si a $T = 300\,\mathrm{K}$.

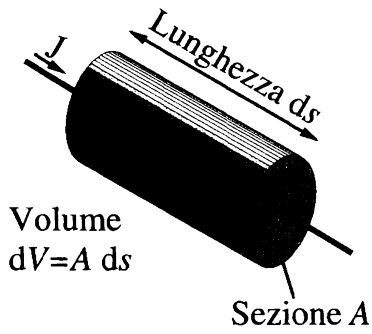


Figura 2.4 Volume elementare dV di semiconduttore per la determinazione della legge di Ohm microscopica.

verifica facilmente che la (2.7) si generalizza nella somma di una densità di corrente di conduzione di elettroni e di lacune:

$$J = qn\mu_n\mathcal{E} + qp\mu_p\mathcal{E} \tag{2.9}$$

permettendo così di valutare la conducibilità elettrica di un campione di semiconduttore uniformemente drogato secondo:

$$\sigma = qn\mu_n + qp\mu_p \tag{2.10}$$

La resistività elettrica, infine, è data dal reciproco della conducibilità:

$$\rho = \frac{1}{\sigma} = \frac{1}{qn\mu_n + qp\mu_p} \tag{2.11}$$

Visto che la resistività ρ dipende dalla concentrazione di carica libera e dalla relativa mobilità, una volta fissata la temperatura il valore di ρ in condizioni di basso campo (ovvero, quando la mobilità è costante e pari al valore di basso campo) dipende solo dalla concentrazione di atomi droganti introdotti nel semiconduttore. In particolare, nel caso in cui il drogaggio sia di un solo tipo (atomi di fosforo per il campione drogato n, di boro per il campione drogato p) si hanno, per il Si a temperatura ambiente, le curve rappresentate nella figura 2.5.

Esempio 2.1 Si consideri un campione di Si uniformemente drogato tipo n con una concentrazione $N_D = 5 \times 10^{16}\ \mathrm{cm}^{-3}$, nel quale si effettui un ulteriore processo di drogaggio di tipo p con

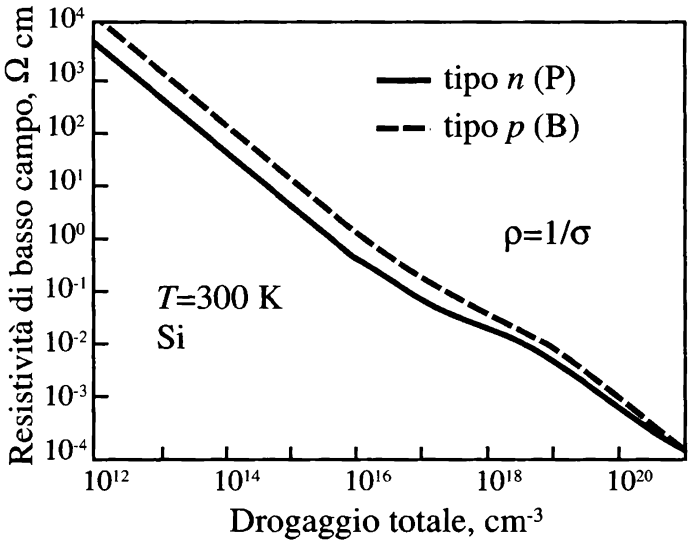


Figura 2.5 Dipendenza dal drogaggio totale della resistività elettrica di basso campo per un campione di Si drogato tipo n e tipo p a $T = 300\,\mathrm{K}$.

una concentrazione $N_\mathrm{A} = 2 \times 10^{17}\ \mathrm{cm}^{-3}$. Si richiede di calcolare:

- ▷ la resistività del campione *prima* dell'introduzione del drogante di tipo p;
- ▷ la resistività del campione *dopo* l'introduzione del drogante di tipo p.

In presenza del solo drogante di tipo n, a temperatura ambiente ($T = 300\,\mathrm{K}$) si ha

$$n = N_\mathrm{D} = 5 \times 10^{16}\,\mathrm{cm}^{-3} \qquad p = \frac{n_\mathrm{i}^2}{N_\mathrm{D}} = 1{,}04 \times 10^{4}\,\mathrm{cm}^{-3}$$

La resisitività del campione è data dalla (2.11), nella quale essendo $p/n \approx 2 \times 10^{-13}$ si può trascurare il contributo dei portatori minoritari, ottenendo così

$$\rho = \frac{1}{qn\mu_n + qp\mu_p} \approx \frac{1}{qn\mu_n}$$

Per calcolare la resistività, è necessario valutare la mobilità degli elettroni liberi. Sulla base del drogaggio totale $N_\mathrm{t} = N_\mathrm{D}$, si ha dalla figura 2.6 (si veda il successivo approfondimento 2.2 per i dettagli su come leggere un valore su una scala lineare)

$$\mu_n = 975\ \mathrm{cm}^2\ \mathrm{V}^{-1}\ \mathrm{s}^{-1}$$

Infine, sostituendo si ottiene $\rho = 0{,}128\,\Omega$ cm.
Dopo il secondo processo di drogaggio, per compensazione il semiconduttore diviene di tipo p con drogaggio efficace

$$N'_\mathrm{A} = N_\mathrm{A} - N_\mathrm{D} = 1.5 \times 10^{17}\,\mathrm{cm}^{-3}$$

A $T = 300\,\mathrm{K}$ si ha poi

$$p \approx N'_\mathrm{A} = 1{,}5 \times 10^{17}\,\mathrm{cm}^{-3} \qquad n = \frac{n_\mathrm{i}^2}{p} = 1{,}4 \times 10^{3}\,\mathrm{cm}^{-3}$$

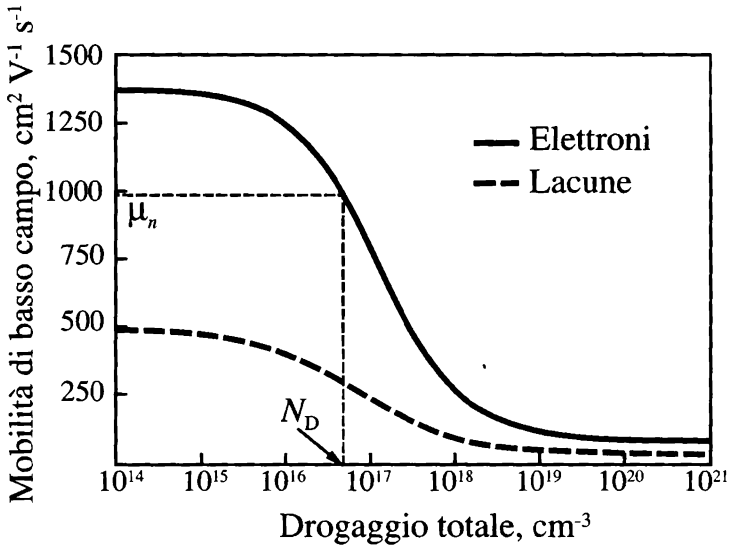


Figura 2.6 Determinazione della mobilità di basso campo per gli elettroni nell'esempio 2.1 dopo il primo passo di drogaggio.

Anche in questo caso si può trascurare il contributo dei portatori minoritari alla resistività

$$\rho = \frac{1}{qn\mu_n + qp\mu_p} \approx \frac{1}{qp\mu_p}$$

Per calcolare la resistività, è necessario valutare la mobilità delle lacune libere. Sulla base del drogaggio totale $N_t = N_D + N_A = 2.5 \times 10^{17}\ \text{cm}^{-3}$, si ha dalla figura 2.7

$$\mu_p = 175\ \text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$$

Sostituendo nell'espressione della resisitività si trova $\rho = 0,238\,\Omega$ cm.

Approfondimento 2.2 In questo approfondimento si richiama la tecnica necessaria per una lettura accurata delle scale graduate, siano esse di tipo lineare o logaritmico.
Nel caso di una scala lineare (si veda la parte sinistra della figura 2.8), si ha proporzionalità diretta tra la grandezza rappresentata sull'asse e la distanza tra i relativi punti. Volendo leggere il valore x compreso tra due estremi a e b distanti d, si ha

$$\frac{b-a}{x-a} = \frac{d}{d_x}$$

dove d_x è la misura della distanza di x da a. Infine, si ricava

$$x = a + (b-a)\frac{d_x}{d}$$

Analogamente, volendo leggere un valore su una scala logaritmica conviene notare come la scala logaritmica preveda una proporzionalità diretta tra il logaritmo in base 10 dei valori e la loro distanza. Pertanto, volendo leggere il valore x compreso nella decade tra a e $10a$, si ha

$$\frac{\log_{10} x - \log_{10} a}{\log_{10}(10a) - \log_{10} a} = \frac{d_x}{d}$$

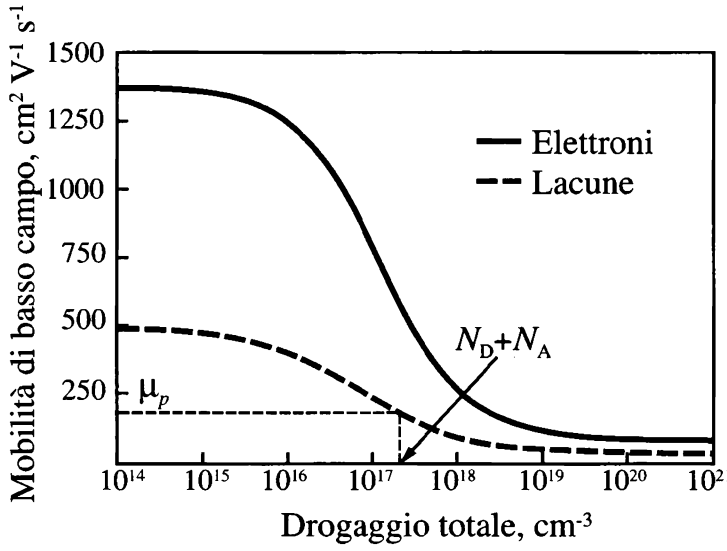


Figura 2.7 Determinazione della mobilità di basso campo per gli elettroni nell'esempio 2.1 dopo il secondo passo di drogaggio.

essendo d la distanza relativa ad una decade e d_x la distanza di x da a sulla scala (si veda la definizione nella parte destra della figura 2.8). Ricordando che per la funzione logaritmo vale a proprietà $\log_{10} x_1 - \log_{10} x_2 = \log_{10}(x_1/x_2)$ e che $\log_{10} 10 = 1$, si ricava

$$\log_{10}(x/a) = \frac{d_x}{d}$$

da cui segue

$$x = a10^{d_x/d}$$

2.1.2 Moto delle cariche libere per diffusione

Il fenomeno della *diffusione* si presenta naturalmente quando in un insieme qualunque di particelle (anche non interagenti) si abbia una disuniformità di concentrazione. In particolare, la tendenza naturale, dettata dalla termodinamica, consiste nello spostamento delle particelle in modo da opporsi alla disuniformità di concentrazione stessa: in altre parole, la diffusione determina un moto di particelle che corrisponde ad uno spostamento netto dalla regione dove ve ne sono molte verso quella dove ve ne sono poche. Da un punto di vista quantitativo, il flusso[4] netto $\boldsymbol{F}$ di particelle determinato dalla diffusione è posto in relazione con la concentrazione c dalla *legge di Fick*:

$$\boldsymbol{F} = -D\nabla c \tag{2.12}$$

[4] Cioè il numero di particelle che attraversano una sezione per unità di area e di tempo.

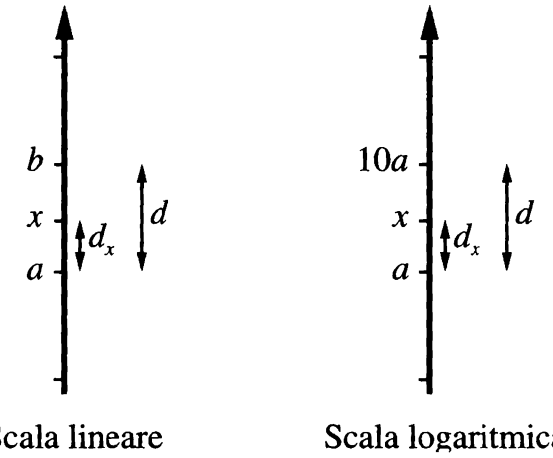


Figura 2.8 Definizione dei valori da leggere e delle distanze da misurare per una lettura accurata di una scala lineare (a sinistra) e logaritmica (a destra).

ovvero, nel caso monodimensionale (si veda la figura 2.9):

$$F = -D\frac{\partial c}{\partial x} \tag{2.13}$$

dove D è il *coefficiente di diffusione* o *diffusività* (unità di misura: cm^2/s) dell'insieme di particelle.

Naturalmente, qualora le particelle siano delle cariche, al loro movimento per diffusione corrisponde una densità di corrente detta *corrente di diffusione.* Poiché il verso della corrente è, per definizione, quello di un flusso di carica positiva, per i due tipi di portatori liberi in un semiconduttore si hanno le seguenti espressioni per le corrispondenti densità di corrente di diffusione:

$$\boldsymbol{J}_{n,\mathrm{diff}} = -q\boldsymbol{F}_n = +qD_n\nabla n \qquad \boldsymbol{J}_{p,\mathrm{diff}} = +q\boldsymbol{F}_p = -qD_p\nabla p \tag{2.14}$$

nel caso tridimensionale, mentre in un istema ad una dimensione spaziale si ha più semplicemente:

$$J_{n,\mathrm{diff}} = -qF_n = +qD_n\frac{\partial n}{\partial x} \qquad J_{p,\mathrm{diff}} = +qF_p = -qD_p\frac{\partial p}{\partial x} \tag{2.15}$$

In queste equazioni, si sono definite separatamente le diffusività D_n e D_p di elettroni e lacune. È possibile dimostrare che, in equilibrio termodinamico, la diffusività e la mobilità delle cariche libere sono tra loro legate dalla *relazione di Einstein:*

$$D_n = \frac{k_\mathrm{B}T}{q}\mu_n = V_T\mu_n \qquad D_p = \frac{k_\mathrm{B}T}{q}\mu_p = V_T\mu_p \tag{2.16}$$

dove $V_T = k_\mathrm{B}T/q$ è una quantità, avente le dimensioni di una tensione, che viene

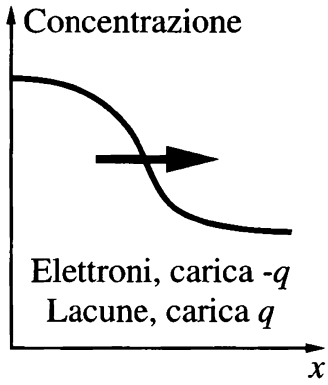


Figura 2.9 Fenomeno della diffusione di particelle in presenza di una concentrazione non uniforme.

spesso denominata *equivalente elettrico della temperatura.* Alla temperatura ambiente $T = 300\,\mathrm{K}$, V_T assume il valore notevole:

$$V_T = 26\,\mathrm{mV} \tag{2.17}$$

2.2 I semiconduttori fuori equilibrio termodinamico

La condizione di equilibrio termodinamico per un dispositivo elettronico a semiconduttore, definita come per qualunque sistema fisico sulla base della richiesta di non avere alcuno scambio di energia con l'esterno, non costituisce di certo la normale condizione di funzionamento per il dispositivo stesso: infatti, lo scopo di qualunque circuito elettronico, sia esso analogico o digitale, è di operare trasformazioni su segnali elettrici (amplificazione, commutazione tra livelli logici, ecc.), richiedendo quindi di operare fuori equilibrio.

In generale, in condizioni fuori equilibrio è naturale attendersi che le concentrazioni di carica libera siano differenti rispetto al valore di equilibrio: anche in un campione di semiconduttore omogeneo, quindi, le concentrazioni di elettroni e lacune saranno delle funzioni (incognite) della posizione e del tempo:

$$n = n(x,t) \qquad p = p(x,t) \tag{2.18}$$

Nella (2.18), si è scelto, per semplicità, di considerare variazioni spaziali in una sola direzione, indicata dall'asse x.

Le variazioni, nello spazio e nel tempo, delle concentrazioni sono imputabili ad una serie di cause:

- ▷ moto delle cariche libere per diffusione, rappresentato dalla *corrente di diffusione;*
- ▷ moto delle cariche libere per trascinamento, rappresentato dalla *corrente di trascinamento* o conduzione;
- ▷ corrente di *spostamento dielettrico:* come discusso nel paragrafo 1.1, questa componente di corrente, presente solo in caso si abbia nel materiale un campo elettri-

co tempo-variante, dà un contributo significativo solo se i segnali applicati hanno frequenze molto elevate, dell'ordine del centinaio di gigahertz;
- ▷ fenomeni di *generazione* e *ricombinazione* (GR) di cariche libere, ovvero creazione e distruzione di portatori liberi nelle bande di conduzione e valenza.

La corrente che attraversa una qualunque sezione di semiconduttore, quindi, può essere espressa dalla somma di due componenti,[5] una relativa agli elettroni ed una alle lacune:

$$J = J_n + J_p \tag{2.19}$$

Ognuna delle due correnti parziali, a sua volta, si esprime nella somma di una componente di trascinamento, data dalla (2.7) e dalla corrispondente equazione per le lacune, e di una di diffusione, data dalla (2.15). Si ha così il modello a *deriva-diffusione* per la densità di corrente in un semiconduttore

$$J_n = J_{n,\mathrm{tr}} + J_{n,\mathrm{diff}} = qn\mu_n\mathcal{E} + qD_n\frac{\partial n}{\partial x} \tag{2.20a}$$

$$J_p = J_{p,\mathrm{tr}} + J_{p,\mathrm{diff}} = qp\mu_p\mathcal{E} - qD_p\frac{\partial p}{\partial x} \tag{2.20b}$$

ovvero, generalizzando al caso tridimensionale:

$$\boldsymbol{J}_n = \boldsymbol{J}_{n,\mathrm{tr}} + \boldsymbol{J}_{n,\mathrm{diff}} = qn\mu_n\boldsymbol{\mathcal{E}} + qD_n\nabla n \tag{2.21a}$$

$$\boldsymbol{J}_p = \boldsymbol{J}_{p,\mathrm{tr}} + \boldsymbol{J}_{p,\mathrm{diff}} = qp\mu_p\boldsymbol{\mathcal{E}} - qD_p\nabla p \tag{2.21b}$$

2.2.1 Eccessi di carica e basso livello di iniezione

Indicando con un pedice 0 la condizione di equilibrio termodinamico, si definiscono le *concentrazioni in eccesso* secondo le relazioni:

$$n'(x,t) = n(x,t) - n_0(x) \qquad p'(x,t) = p(x,t) - p_0(x) \tag{2.22}$$

dove si è considerato il caso generico di concentrazione di carica libera in equilibrio dipendente dalla posizione: le concentrazioni in eccesso sono quindi definite come la variazione della densità di carica libera calcolata rispetto al valore di equilibrio termodinamico. Si noti che mentre le concentrazioni di carica sono quantità definite non negative, gli eccessi di carica possono avere segno positivo o negativo a seconda della particolare condizione di funzionamento e del punto del dispositivo nel quale vengono misurati. Per questo motivo, si è in presenza di un fenomeno di *iniezione* di carica se $n',p' > 0$, mentre si ha il fenomeno dello *svuotamento* di carica se $n',p' < 0$.

Nel seguito, in particolare nello studio dei dispositivi bipolari, qualora sia significativo evidenziare il tipo di drogaggio della regione di semiconduttore da prendere in esame, esso verrà aggiunto a pedice delle concentrazioni di carica:

$$n_p(x,t), p_p(x,t) \text{ in una regione tipo } p \qquad n_n(x,t), p_n(x,t) \text{ in una regione tipo } n \tag{2.23}$$

[5] A questa decomposizione occorrerebbe ancora aggiungere la componente di spostamento dielettrico che, come già più volte affermato, è trascurabile a meno di applicare al dispositivo campi elettrici tempo-varianti a frequenza elevata.

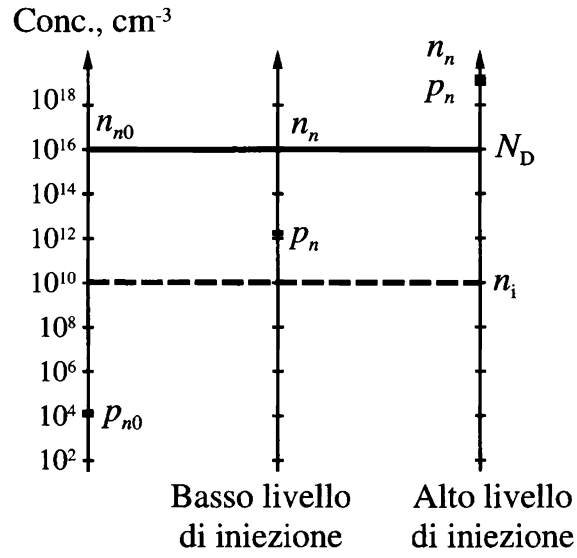


Figura 2.10 Condizioni di basso e alto livello di iniezione in un semiconduttore drogato n: si assume per questo grafico $n' \approx p' > 0$.

Corrispondentemente, i relativi eccessi di carica saranno indicati secondo:

$$n'_p = n_p - n_{p0}, p'_p = p_p - p_{p0} \text{ in una regione tipo } p \tag{2.24}$$

$$n'_n = n_n - n_{n0}, p'_n = p_n - p_{n0} \text{ in una regione tipo } n \tag{2.25}$$

Considerando il caso specifico di una regione di semiconduttore uniformemente drogata, la condizione di neutralità locale descritta nel paragrafo 1.8, valida in equilibrio termodinamico, si esprime:[6]

$$\rho_0 = q(p_0 - n_0 + N_D - N_A) = 0 \tag{2.26}$$

Si può dimostrare [2] che anche fuori equilibrio la condizione di neutralità locale è spesso, almeno approssimativamente, valida: si parla di approssimazione di *quasi-neutralià*. In questo caso, essendo

$$\rho = q(p - n + N_D - N_A) = q(p_0 + p' - n_0 - n' + N_D - N_A) \approx 0 \tag{2.27}$$

e facendo uso della (2.26), è facile mostrare come la condizione di quasi-neutralità implichi:

$$p' \approx n' \tag{2.28}$$

Si consideri ora un campione di Si drogato con una concentrazione costante N_D di atomi donatori. A temperatura ambiente, come discusso nel capitolo 1, e in equilibrio

[6] Si noti che si è anche utilizzata la condizione di completa ionizzazione.

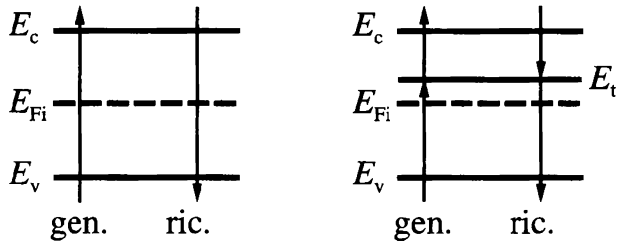


Figura 2.11 Classificazione dei processi di GR: transizioni dirette (a sinistra) e mediate da livelli trappola (a destra). Le frecce indicano lo spostamento di un elettrone.

termodinamico $n_{n0} \approx N_D$ e $p_{n0} \approx n_i^2/N_D$: tali valori sono rappresentati sull'asse di sinistra della figura 2.10. Si supponga, ora, di portare il campione fuori dall'equilibrio termodinamico, ottenendo al suo interno un eccesso uniforme di carica libera $n'_n \approx p'_n$ in condizioni di quasi-neutralità. Si definisce l'approssimazione di *basso livello di iniezione* come la condizione per la quale:

$$|n'_n|, |p'_n| \ll n_{n0} = N_D \tag{2.29}$$

che, più in generale, si può esprimere come segue: l'eccesso di carica libera deve essere trascurabile rispetto alla concentrazione di portatori maggioritari di equilibrio nel campione. Come si vede dall'asse centrale della figura 2.10, in basso livello di iniezione i portatori minoritari, in equilibrio in concentrazione di molti ordini di grandezza inferiore rispetto ai maggioritari, subiscono una *significativa variazione* rispetto al valore di equilibrio termodinamico, mentre i portatori maggioritari restano *praticamente invariati.* Ciò significa che, in termini matematici, la condizione di basso livello di iniezione può anche esprimersi:

$$p_p \approx p_{p0} \text{ in una regione tipo } p \qquad n_n \approx n_{n0} \text{ in una regione tipo } n \tag{2.30}$$

Pertanto, in basso livello di iniezione solo i portatori minoritari vengono perturbati in modo significativo rispetto al valore di equilibrio termodinamico: per questo motivo, tale condizione verrà spesso invocata nel seguito per poter semplificare le equazioni del modello matematico per lo studio dei dispositivi elettronici.

Per contrapposizione, si parla di *alto livello di iniezione* (asse di destra nella figura 2.10 quando l'eccesso di carica libera è almeno dello stesso ordine di grandezza dei portatori maggioritari in equilibrio termodinamico: in questo caso, entrambi i tipi di portatori risultano essere significativamente perturbati dall'eccesso di carica.

2.2.2 I fenomeni di generazione e ricombinazione

I fenomeni di *generazione* e *ricombinazione* (GR) sono, rispettivamente, eventi di creazione e distruzione di portatori liberi nella relativa banda del semiconduttore. All'interno dei semiconduttori, vi sono diversi meccanismi fisici che possono presiedere alla realizzazione dei fenomeni di GR. Senza entrare troppo nel dettaglio,[7] i fenomeni di

[7] Il lettore interessato può trovare in [2] una descrizione più dettagliata.

GR possono essere classificati secondo la tipologia del salto energetico che gli elettroni subiscono:

- meccanismi *diretti* (detti anche *banda-banda,* si veda la parte sinistra della figura 2.11), se si hanno transizioni direttamente tra la banda di valenza e quella di conduzione. Esempi di questi meccanismi sono quello termico, quello Auger, quello per ionizzazione da impatto, e quello radiativo: quest'ultimo costituisce il fenomeno fisico di base che viene utilizzato nei dispositivi optoelettronici;
- meccanismi *indiretti* (parte destra della figura 2.11), nei quali le transizioni sono mediate da livelli intermedi nella banda proibita detti *livelli trappola.* Le transizioni indirette sono spesso denominate processi Shockley Read Hall (SRH), dai nomi di coloro che per primi hanno proposto un modello interpretativo del fenomeno.

L'importanza relativa dei vari meccanismi di GR dipende da vari fattori, quali il tipo di semiconduttore, il livello di drogaggio, il valore del campo elettrico nel materiale, ecc. Nel caso del Si, le transizioni SRH sono quelle che rivestono, in generale, la maggiore importanza.

Da un punto di vista quantitativo, i fenomeni di GR vengono caratterizzati definendo, per ogni tipo di carica libera nel materiale, due quantità:

- il *tasso* o *velocità di generazione* per gli elettroni (simbolo: G_n, unità di misura: $\mathrm{cm}^{-3}/\mathrm{s}$) e le lacune (simbolo: G_p), cioè il numero di elettroni (lacune) per unità di volume che vengono generati in banda di conduzione (valenza) nell'unità di tempo;
- il *tasso* o *velocità di ricombinazione* per gli elettroni (simbolo: R_n, unità di misura: $\mathrm{cm}^{-3}/\mathrm{s}$) e le lacune (simbolo: R_p), cioè il numero di elettroni (lacune) per unità di volume che vengono ricombinati dalla banda di conduzione (valenza) nell'unità di tempo.

Con riferimento ad ogni insieme di cariche libere, si utilizza spesso anche il *tasso netto di ricombinazione:*

$$U_n = R_n - G_n \qquad U_p = R_p - G_p \tag{2.31}$$

che, in condizioni di equilibrio termodinamico quando non vi può essere aumento o riduzione netta di cariche libere, deve soddisfare le condizioni

$$U_n = 0 \qquad U_p = 0 \tag{2.32}$$

A seconda del tipo di meccanismo di GR considerato, è possibile ricavare un diverso modello matematico che descriva il valore del tasso netto di ricombinazione di elettroni e lacune. Tali espressioni sono in generale abbastanza complesse: ai fini di uno studio approssimato dei dispositivi, condotto il più possibile per via analitica, si fa uso di una descrizione semplificata, detta *approssimazione di tempo di vita,* che consiste nell'assumere:

$$U_n \approx \frac{n - n_0}{\tau_n} = \frac{n'}{\tau_n} \qquad U_p \approx \frac{p - p_0}{\tau_p} = \frac{p'}{\tau_p} \tag{2.33}$$

dove τ_n e τ_p sono detti, rispettivamente, *tempo di vita media* di elettroni e lacune.

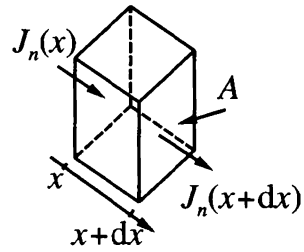


Figura 2.12 Volume elementare $dV = Adx$ utilizzato nella dimostrazione dell'equazione di continuità. Caso degli elettroni.

2.3 L'equazione di continuità e il modello matematico dei semiconduttori

Partendo dal principio di conservazione della carica, è possibile ricavare una equazione che definisca la dinamica delle concentrazioni di carica libera in un semiconduttore, ovvero la legge di evoluzione nello spazio e nel tempo di n e di p.

Si consideri un volume elementare dV di semiconduttore, di sezione A trasversale al flusso di corrente di elettroni J_n diretta secondo l'asse x, e si indichi con dx lo spessore del volume elementare, in modo da avere $dV = Adx$. Sulla base del principio di conservazione della carica, la variazione nell'unità di tempo $dV\partial n/\partial t$ del numero di elettroni contenuti nel volume è determinata da quattro possibili componenti:

1. gli elettroni entranti per unità di tempo dalla faccia nella posizione x ($J_n(x)$);
2. gli elettroni uscenti per unità di tempo dalla faccia nella posizione $x + dx$ ($J_n(x + dx)$);
3. gli elettroni generati nell'unità di tempo nel volume dV (G_n);
4. gli elettroni ricombinati nell'unità di tempo nel volume dV (R_n).

Sommando questi quattro contributi, si ha:

$$\frac{\partial n}{\partial t}Adx = \underbrace{\frac{J_n(x)}{-q}A}_{1.} - \underbrace{\frac{J_n(x+dx)}{-q}A}_{2.} + \underbrace{G_nAdx}_{3.} - \underbrace{R_nAdx}_{4.} \tag{2.34}$$

dove $J_n/(-q)$ rappresenta il numero di elettroni che attraversano la sezione A nell'unità di tempo e superficie. Utilizzando lo sviluppo in serie:

$$J_n(x+dx) \sim J_n(x) + \frac{\partial J_n}{\partial x}dx \quad dx \to 0 \tag{2.35}$$

si ottiene, per $dx \to 0$

$$\frac{\partial n}{\partial t}Adx \sim \frac{J_n(x)}{-q}A - \frac{A}{-q}\left[J_n(x) + \frac{\partial J_n}{\partial x}dx\right] + G_nAdx - R_nAdx \tag{2.36}$$

ovvero l'*equazione di continuità per gli elettroni* nel caso unidimensionale:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - U_n \tag{2.37}$$

Ripetendo il ragionamento nel caso di un flusso di lacune, l'unica differenza da tenere presente è relativa alla carica $+q$ corrispondente a tali cariche, per cui il numero di lacune che attraversano la sezione A nell'unità di tempo e superficie risulta essere espressa da $J_p/(+q)$. Si ottiene così l'*equazione di continuità per le lacune* nel caso unidimensionale:

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - U_p \tag{2.38}$$

Nelle precedenti espressioni, i tassi netti di ricombinazione sono, almeno nell'approssimaizone di tempo di vita (2.33), delle funzioni delle concentrazioni di carica libera, mentre le densità di corrente dipendono, oltre che da n e da p, anche dal campo elettrico (si veda la (2.20)). Pertanto, per rendere il problema matematicamente ben posto occorre aggiungere alle due equazioni di continuità una terza equazione indipendente. Essa è fornita dall'*equazione di Gauss* che lega il campo elettrico alla densità di carica totale:

$$\frac{\partial \mathcal{E}}{\partial x} = \frac{\rho}{\epsilon} \qquad \rho = q(p - n + N_{\mathrm{D}}^{+} - N_{\mathrm{A}}^{-}) \tag{2.39}$$

Poiché ai dispositivi elettronici vengono quasi sempre applicate differenze di potenziale, per semplificare le condizioni al contorno si preferisce utilizzare come incognita il *potenziale elettrostatico* φ al posto del campo elettrico, essendo le due grandezze legate dalla relazione

$$\mathcal{E} = -\frac{\partial \varphi}{\partial x} \tag{2.40}$$

In definitiva, quindi, si è determinato il *modello matematico dei semiconduttori* (caso unidimensionale) nelle tre incognite φ,n,p costituito dalle tre equazioni differenziali a derivate parziali:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n}{\partial x} - U_n \tag{2.41a}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p}{\partial x} - U_p \tag{2.41b}$$

$$\frac{\partial^2 \varphi}{\partial x^2} = -\frac{\rho}{\epsilon} \tag{2.41c}$$

completato dalle relazioni costitutive

$$J_n = qn\mu_n\mathcal{E} + qD_n\frac{\partial n}{\partial x} \tag{2.42a}$$

$$J_p = qp\mu_p\mathcal{E} - qD_p\frac{\partial p}{\partial x} \tag{2.42b}$$

$$\mathcal{E} = -\frac{\partial \varphi}{\partial x} \tag{2.42c}$$

dalle espressioni per i tassi netti di ricombinazione e per la carica netta, e dalle relative condizioni iniziali ed al contorno. Nel caso tridimensionale, le equazioni si generalizzano in:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \boldsymbol{J}_n - U_n \tag{2.43a}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \boldsymbol{J}_p - U_p \tag{2.43b}$$

$$\nabla^2 \varphi = -\frac{\rho}{\epsilon} \tag{2.43c}$$

e

$$\boldsymbol{J}_n = qn\mu_n \boldsymbol{\mathcal{E}} + qD_n \nabla n \tag{2.44a}$$

$$\boldsymbol{J}_p = qp\mu_p \boldsymbol{\mathcal{E}} - qD_p \nabla p \tag{2.44b}$$

$$\mathcal{E} = -\nabla \varphi \tag{2.44c}$$

Questo insieme di equazioni viene spesso denominato *modello a deriva-diffusione,* dalla approssimazione utilizzata per la descrizione della corrente nel materiale.

2.3.1 Approssimazioni del modello matematico

Con l'obiettivo di condurre un'analisi approssimata dei dispositivi a semiconduttore, occorre semplificare le equazioni del modello matematico per poterne ottenere una soluzione analitica. Tra le molte semplificazioni necessarie, le perincipali sono le seguenti:

- ▷ la *mobilità* di elettroni e lacune è assunta *costante e pari al valore di basso campo;*
- ▷ la *diffusività* viene valutata sulla base della *relazione di Einstein* ($D = V_T \mu$);
- ▷ i fenomeni di *generazione e ricombinazione* sono trattati nell'*approssimazione di tempo di vita;*
- ▷ la *ionizzazione* degli atomi droganti è *completa.*

In questo modo, le equazioni (2.41) si semplificano in:

$$\frac{\partial n}{\partial t} = \mu_n \frac{\partial (n\mathcal{E})}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} - \frac{n - n_0}{\tau_n} \tag{2.45a}$$

$$\frac{\partial p}{\partial t} = -\mu_p \frac{\partial (p\mathcal{E})}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} - \frac{p - p_0}{\tau_p} \tag{2.45b}$$

$$\frac{\partial^2 \varphi}{\partial x^2} = -\frac{q}{\epsilon}(p - n + N_\mathrm{D} - N_\mathrm{A}) \tag{2.45c}$$

Approssimazione di quasi-neutralità

Una regione di semiconduttore viene detta *quasi-neutra* se in essa si può assumere che la densità di carica totale sia nulla:

$$\rho = 0 \tag{2.46}$$

Nel caso unidimensionale, dall'equazione di Poisson (2.39) segue banalmente che in una regione quasi-neutra il *campo elettrico è costante* $\mathcal{E}(x) = \mathcal{E}_0$ e quindi, per la (2.40), il *potenziale elettrostatico è rettilineo*:

$$\varphi(x) = -\mathcal{E}_0 x + \varphi_0 \tag{2.47}$$

cosicché la differenza di potenziale tra due punti distanti d (ad esempio, di coordinate x_1 e $x_2 = x_1 + d$) è data, in valore assoluto, dalla quantità:

$$|\varphi(x_2) - \varphi(x_1)| = |\mathcal{E}_0|(x_2 - x_1) = |\mathcal{E}_0|d \tag{2.48}$$

Qualora le condizioni al contorno lo consentano, in una regione quasi-neutra il campo elettrico costante potrebbe essere nullo: $\mathcal{E}_0 = 0$. In questo caso, la caduta di potenziale ai capi della regione è anch'essa nulla, e si può trascurare la componente di trascinamento della densità di corrente, semplificando così il modello matematico poiché le due equazioni di continuità risultano essere disaccoppiate:

$$\frac{\partial n}{\partial t} = D_n \frac{\partial^2 n}{\partial x^2} - \frac{n - n_0}{\tau_n} \tag{2.49a}$$

$$\frac{\partial p}{\partial t} = D_p \frac{\partial^2 p}{\partial x^2} - \frac{p - p_0}{\tau_p} \tag{2.49b}$$

Esempio 2.2 Un tipico esempio di regione quasi-neutra nella quale non si ha necessariamente un campo elettrico nullo è la regione di dielettrico tra le armature di un condensatore a facce piane parallele. Indicando con $d = x_2 - x_1$ la distanza tra le due armature, ed applicando al condensatore una differenza di potenziale costante e pari a V_0, sulle due armature si accumulano due cariche uguali ed opposte, rispettivamente pari a $+Q = CV_0$ e $-Q = -CV_0$, come mostrato nella figura 2.13. La regione compresa tra le armature, definita da $x_1 \leq x \leq x_2$ è neutra, ovvero, assumendo il dielettrico ideale, presenta una densità di carica totale nulla. D'altra parte, il campo elettrico al suo interno non può essere nullo, essendo la differenza di potenziale tra le due armature fissata dall'esterno: sulla base della (2.48), si deduce che il campo elettrico presente nella regione neutra, corrispondente alla distribuzione di carica presente ai bordi della regione stessa, è costante e di valore pari a $\mathcal{E}_0 = V_0/d$.

Generalizzando i risultati dell'esempio 2.2, si può affermare come una regione quasi neutra unidimensionale possa essere considerata a campo elettrico nullo qualora si verifichi almeno una tra le condizioni seguenti:

- ▷ la regione *non è compresa* all'interno di un doppio strato di carica, ovvero non ha ai bordi una carica $+Q$ da un lato compensata da una carica uguale ed opposta $-Q$ dal lato opposto;
- ▷ la regione *presenta una caduta di potenziale trascurabile* ai capi, ad esempio perché viene attraversata da una corrente nulla.

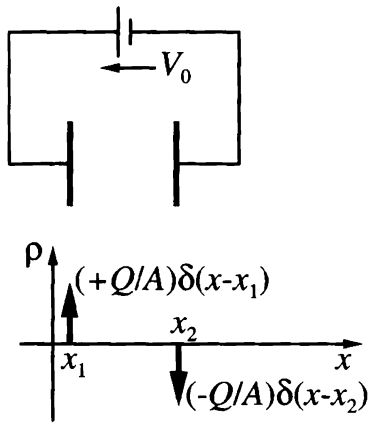


Figura 2.13 Rapresentazione di un condensatore a facce piane parallele al quale sia applicata una differenza di potenziale V_0 (sopra) e corrispondente distribuzione di densità di carica totale (sotto). Si indica con A l'area delle armature.

Approfondimento 2.3 In alcune applicazioni può essere conveniente utilizzare come incognite del modello matematico, al posto delle concentrazioni di carica libera, delle quantità aventi la dimensione di un'energia, definite estendendo alle condizioni fuori dall'equilibrio termodinamico le relazioni di Shockley (1.30). Si definiscono così i *quasi-livelli di Fermi* E_{Fn} e E_{Fp} (detti anche *IMREF*) per gli elettroni e le lacune postulando:

$$n = n_i \exp\left(\frac{E_{Fn} - E_{Fi}}{k_B T}\right) \qquad p = n_i \exp\left(\frac{E_{Fi} - E_{Fp}}{k_B T}\right)$$

Poiché in generale le concentrazioni di carica sono una funzione della posizione e del tempo, mentre il livello di Fermi intrinseco dipende solo dalla posizione all'interno del materiale, anche per i quasi-livelli di Fermi vale la dipendenza $E_{Fn} = E_{Fn}(x,t)$ e $E_{Fp} = E_{Fp}(x,t)$.
Visto che il potenziale elettrostatico nel materiale può essere definito a meno di una costante, e rappresenta l'energia potenziale alla quale sono sottoposte le cariche libere nel materiale, senza perdere di generalità è possibile assumere $E_{Fi} = -q\varphi$, in modo da ottenere

$$\mathcal{E} = -\frac{\partial \varphi}{\partial x} = \frac{1}{q}\frac{\partial E_{Fi}}{\partial x}$$

e quindi, per le espressioni delle correnti di trascinamento di elettroni e lacune

$$J_{n,tr} = qn\mu_n\mathcal{E} = n\mu_n\frac{\partial E_{Fi}}{\partial x} \qquad J_{p,tr} = qp\mu_p\mathcal{E} = p\mu_p\frac{\partial E_{Fi}}{\partial x}$$

Derivando le definizioni di quasi-livelli di Fermi, poi, si ottiene

$$\frac{\partial n}{\partial x} = \frac{n}{k_B T}\left(\frac{\partial E_{Fn}}{\partial x} - \frac{\partial E_{Fi}}{\partial x}\right) \qquad \frac{\partial p}{\partial x} = \frac{p}{k_B T}\left(\frac{\partial E_{Fi}}{\partial x} - \frac{\partial E_{Fp}}{\partial x}\right)$$

per cui, grazie alla relazione di Einstein, la corrente di diffusione assume l'espressione

$$J_{n,diff} = qD_n\frac{\partial n}{\partial x} = n\mu_n\left(\frac{\partial E_{Fn}}{\partial x} - \frac{\partial E_{Fi}}{\partial x}\right) \qquad J_{p,diff} = -qD_p\frac{\partial n}{\partial x} = p\mu_p\left(\frac{\partial E_{Fp}}{\partial x} - \frac{\partial E_{Fi}}{\partial x}\right)$$

Sommando le componenti di trascinamento e diffusione delle densità di corrente di elettroni e lacune, si ottengono le relazioni

$$J_n = n\mu_n \frac{\partial E_{Fn}}{\partial x} \qquad J_p = p\mu_p \frac{\partial E_{Fp}}{\partial x}$$

che identificano una importante proprietà dei quasi-livelli di Fermi: essi risultano essere *indipendenti dalla posizione* qualora la corrispondente densità di corrente totale sia nulla. Tali considerazioni, naturalmente, valgono nell'ambito delle stesse approssimazioni utilizzate per derivare le relazioni sulle quali sono basate, ovvero l'uso della relazione di Einstein per la valutazione della diffusività. Moltiplicando tra loro le due definizioni dei quasi-livelli di Fermi, si trova un generalizzazione della legge di azione di massa:

$$np = n_i^2 \exp\left(\frac{E_{Fn} - E_{Fp}}{k_B T}\right)$$

dalla quale segue immediatamente che, in equilibrio termodinamico

$$E_{Fn} = E_{Fp} = E_F$$

Regime stazionario

La più semplice condizione di funzionamento possibile per un dispositivo elettronico è quando ad esso vengano applicati segnali elettrico costanti nel tempo, e la loro applicazione sia avvenutà per un tempo sufficientemente lungo da poter considerare esauriti tutti i transitori possibilmente presenti nel dispositivo stesso. In queste condizioni, tutte le variabili presenti possono essere assunte costanti nel tempo, e si parla di funzionamento in *regime stazionario*. Dal punto di vista matematico, nel regime stazionario vale la condizione $\partial/\partial t = 0$ per qualunque variabile, per cui le equazioni del modello matematico (2.45) si semplificano in:

$$0 = \mu_n \frac{\mathrm{d}(n\mathcal{E})}{\mathrm{d}x} + D_n \frac{\mathrm{d}^2 n}{\mathrm{d}x^2} - \frac{n - n_0}{\tau_n} \tag{2.50a}$$

$$0 = -\mu_p \frac{\mathrm{d}(p\mathcal{E})}{\mathrm{d}x} + D_p \frac{\mathrm{d}^2 p}{\mathrm{d}x^2} - \frac{p - p_0}{\tau_p} \tag{2.50b}$$

$$\frac{\mathrm{d}^2 \varphi}{\mathrm{d}x^2} = -\frac{q}{\epsilon}(p - n + N_D - N_A) \tag{2.50c}$$

Esempio 2.3 Si considerino due campioni di Si uniformemente drogati di tipo p con una concentrazione $N_A = 5 \times 10^{16}\,\mathrm{cm}^{-3}$, lunghi rispettivamente $L_1 = 3\,\mu\mathrm{m}$ e $L_2 = 200\,\mu\mathrm{m}$. Sulla superficie posta in $x = 0$ si assuma di aver prodotto un *eccesso di portatori minoritari* $n'_p(0) = 10^{13}\,\mathrm{cm}^{-3}$, caratterizzati da un tempo di vita $\tau_n = 50$ ns. Si richiede di calcolare la distribuzione di portatori minoritari in condizioni stazionarie,assumendo valida l'ipotesi di quasi-neutralità.
Grazie all'assunzione di quasi-neutralità, si ha $n'_p(x) \approx p'_p(x)$. Inoltre, nel campione non si ha passaggio di corrente nè carica accumulata da ambo i lati, pertanto si può ritenere che le condizioni al contorno siano compatibili con una regione a campo elettrico nullo. Di conseguenza, la concentrazione di portatori minoritari dovrà seguire l'equazione (2.49a), qui riscritta tenendo conto della ulteriore condizione di regime stazionario

$$0 = D_n \frac{\mathrm{d}^2 n_p}{\mathrm{d}x^2} - \frac{n_p - n_{p0}}{\tau_n}$$

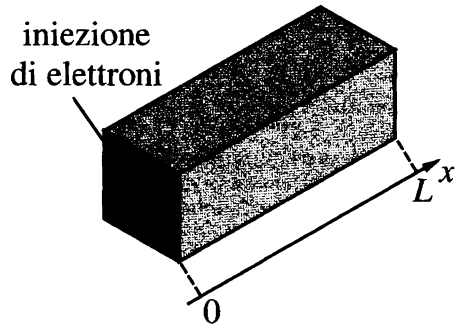


Figura 2.14 Campione di semiconduttore con iniezione di carica su una superficie per l'esempio 2.3.

Poiché il campione è uniformemente drogato, n_{p0} è indipendente dalla posizione, e quindi l'equazione precedente può essere riscritta nella forma

$$0 = D_n \frac{\mathrm{d}^2 n_p'}{\mathrm{d}x^2} - \frac{n_p'}{\tau_n}$$

L'equazione deve essere completata da due condizioni al contorno per definirne univocamente la soluzione. La prima condizione è il valore, fissato dal testo del problema, dell'eccesso di carica nella faccia in $x = 0$, ovvero $n_p'(0)$. La seconda condizione si ottiene imponendo che l'altro estremo del campione sia un *contatto ohmico*, ovvero che valga

$$n_p'(L) = 0$$

L'equazione differenziale è a coefficienti costanti, e può essere risolta, ad esempio, con il metodo dell'equazione caratteristica: si costruisce innanzitutto l'equazione algebrica (l'*equazione caratteristica*) associata all'equazione differenziale:

$$0 = D_n \lambda^2 - \frac{1}{\tau_n}$$

le cui radici sono

$$\lambda_{1,2} = \pm \frac{1}{\sqrt{D_n \tau_n}} = \pm \frac{1}{L_n}$$

dove si è definita la *lunghezza di diffusione dei portatori minoritari* $L_n = \sqrt{D_n \tau_n}$. La soluzione generale dell'equazione differenziale, infine, è data da una combinazione lineare di esponenziali aventi per esponente le radici dell'equazione caratteristica:

$$n_p'(x) = A \exp\left(\frac{x}{L_n}\right) + B \exp\left(-\frac{x}{L_n}\right)$$

dove A e B sono due costanti da determinarsi in funzione delle condizioni al contorno:

$$n_p'(0) = A + B$$

$$n_p'(L) = A \exp\left(\frac{L}{L_n}\right) + B \exp\left(-\frac{L}{L_n}\right) = 0$$

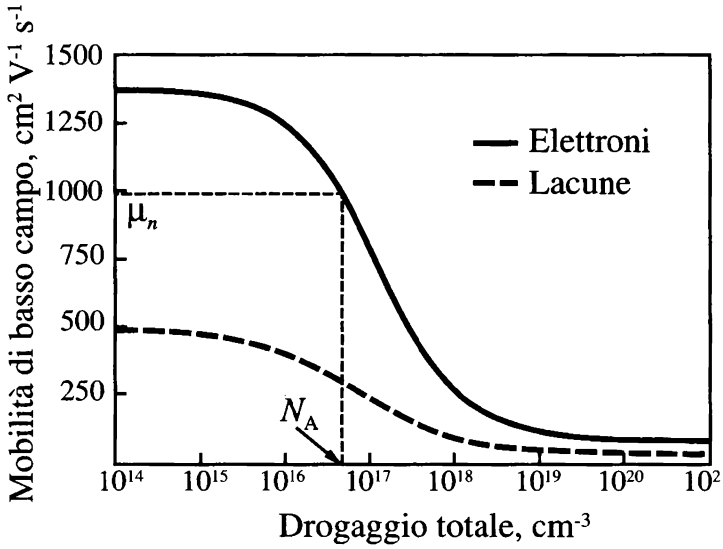


Figura 2.15 Determinazione della mobilità di basso campo per gli elettroni nell'esempio 2.3.

Risolvendo questo sistema lineare nelle due incognite A e B si ricava:

$$A = n'_p(0)\frac{-\exp\left(-\dfrac{L}{L_n}\right)}{\exp\left(\dfrac{L}{L_n}\right) - \exp\left(-\dfrac{L}{L_n}\right)}$$

$$B = n'_p(0)\frac{\exp\left(\dfrac{L}{L_n}\right)}{\exp\left(\dfrac{L}{L_n}\right) - \exp\left(-\dfrac{L}{L_n}\right)}$$

Sostituendo queste espressioni nella soluzione generale, dopo qualche manipolazione algebrica basata sulla definizione di seno iperbolico:

$$\sinh\alpha = \frac{\exp(\alpha) - \exp(-\alpha)}{2}$$

si ottiene la soluzione particolare richiesta nella forma

$$n'_p(x) = n'_p(0)\frac{\sinh\left(\dfrac{L-x}{L_n}\right)}{\sinh\left(\dfrac{L}{L_n}\right)}$$

Dalla figura 2.15, in corrispondenza del drogaggio totale N_A si ottiene la mobilità degli elettroni liberi nel campione $\mu_n = 975\,\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$, per cui a 300 K

$$D_n = V_T\mu_n = 25,35\,\text{cm}^2/\text{s} \qquad L_n = \sqrt{D_n\tau_n} = 11,26\,\mu\text{m}$$

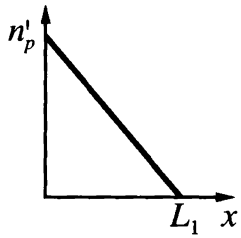


Figura 2.16 Dipendenza dalla posizione dei portatori minoritari iniettati nel caso di un campione corto rispetto alla lunghezza di diffusione.

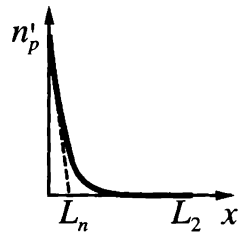


Figura 2.17 Dipendenza dalla posizione dei portatori minoritari iniettati nel caso di un campione lungo rispetto alla lunghezza di diffusione.

Per il primo campione si ha $L_1/L_n \ll 1$, e quindi si avrà anche $(L_1 - x)/L_n \ll 1$ essendo $0 \leq x \leq L_1$. In questo caso, si dice che si ha un *campione corto* rispetto alla lunghezza di diffusione dei portatori minoritari. Usando nella soluzione dell'equazione differenziale lo sviluppo asintotico $\sinh \alpha \approx \alpha$, valido per $\alpha \to 0$, si ottiene una relazione lineare per la dipendenza spaziale dei portatori minoritari iniettati (figura 2.16):

$$n'_p(x) = n'_p(0)\frac{L_1 - x}{L_1}$$

Per il secondo campione, invece, si ha $L_2/L_n \gg 1$: nella soluzione generale si può quindi assumere $A \approx 0$ e $B \approx n'_p(0)$, in modo da evitare la presenza di un esponenziale divergente nella soluzione. Si dice, in questo caso, che si ha un *campione lungo* rispetto alla lunghezza di diffusione, con dipendenza esponenziale dei portatori minoritari iniettati (figura 2.17)

$$n'_p(x) = n'_p(0)\exp\left(-\frac{x}{L_n}\right)$$

Capitolo 3

Diodo a giunzione *pn*

La giunzione *pn*, ovvero un campione di semiconduttore perfettamente cristallino nel quale si siano realizzate due regioni con drogaggio di tipo opposto, è il dispositivo a semiconduttore che verrà studiato in questo capitolo. Nel paragrafo 3.1 verrà introdotta la rappresentazione del diagramma a bande di energia nel dispositivo in condizioni di equilibrio termodinamico, definendo anche le regole sufficienti alla sua costruzione qualitativa in un generico dispositivo a semiconduttore. L'analisi quantitativa del diagramma a bande sarà invece presentata nel paragrafo 3.2. Il paragrafo 3.3, invece, è dedicato allo studio delle caratteristiche elettriche della giunzione in regime stazionario. Gli effetti capacitivi sono introdotti nel paragrafo 3.4, mentre la rappresentazione circuitale del comportamento elettrico della giunzione, compresa una introduzione al concetto di analisi di piccolo segnale, viene discussa nel paragrafo 3.5. Infine, nel paragrafo 3.6 si studia il comportamento della giunzione in condizioni di breakdown.

3.1 Diagramma a bande di energia

Il *diagramma a bande* di un dispositivo a semiconduttore è, convenzionalmente, una rappresentazione grafica della dipendenza spaziale dell'energia potenziale alla quale sono sottoposti gli elettroni nel materiale. Naturalmente, il diagramma a bande assume, per uno stesso dispositivo, una forma diversa a seconda che si consideri il sistema in condizioni di equilibrio termodinamico oppure no. Già nel caso più semplice, ovvero in equilibrio termodinamico, dalla forma del diagramma a bande di energia è possibile inferire, almeno in forma qualitativa, alcune caratteristiche del comportamento elettrico del dispositivo. Nell'ambito di questo paragrafo, si enunceranno le regole necessarie a costruire, almeno in termini qualitativi, il diagramma a bande di equilibrio termodinamico in un qualunque dispositivo, prendendo in considerazione il caso specifico di una giunzione *pn brusca*, ovvero caratterizzata da una transizione brusca tra la regione drogata p e quella drogata n.

Nella figura 3.1 si trova una rappresentazione della giunzione *pn* brusca (a sinistra), e della dipendenza dalla posizione del *drogaggio netto di tipo n* nel campione, ovvero della funzione $N_D - N_A$: dove il drogaggio netto ha segno positivo, il semiconduttore è drogato di tipo n, dove ha segno negativo il campione è drogato di tipo p. Il punto nel

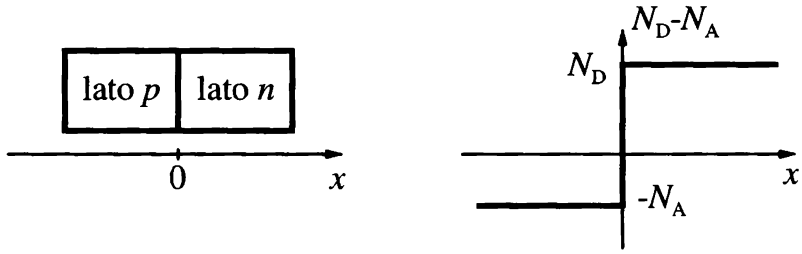


Figura 3.1 Definizione dell'asse x in una giunzione brusca (a sinistra), e rappresentazione grafica del drogaggio netto di tipo n nel dispositivo (a destra).

quale il drogaggio netto si annulla viene chiamato *giunzione metallurgica,* in quanto si tratta della posizione che definisce la transizione tra i due materiali di tipo diverso. La costruzione del diagramma a bande di energia presenta delle difficoltà nel caso il materiale sia non omogeneo spazialmente, come ad esempio in presenza di una giunzione *pn.* Tali difficoltà sono, naturalmente, ridotte se la disomogeneità spaziale è di tipo particolarmente semplice, ad esempio se la struttura è *omogenea a tratti,* come accade per la giunzione *pn* brusca, costituita da due regioni di semiconduttore al loro interno omogenee. La costruzione del diagramma a bande inizia dalla definizione, in equilibrio termodinamico, dei diagrammi a bande di energia parziali delle due regioni omogenee, assumendo che esse siano inizialmente isolate: si ha in questo caso il diagramma mostrato nella figura 3.2.

Nei due lati della figura 3.2, il diagramma a bande di energia è definito, oltre che dal drogaggio, anche da altri due parametri energetici che dipendono dal materiale semiconduttore utilizzato: l'*affinità elettronica,* normalmente indicata dal simbolo $q\chi_S$, e il *lavoro di estrazione* $q\Phi_S$. Queste quantità rappresentano delle differenze di energia misurate a partire dal *livello del vuoto* E_0, ovvero dall'energia minima che deve assumere un elettrone per essere libero di lasciare il cristallo. In particolare, $q\chi_S$ rappresenta la distanza tra E_0 e il minimo energetico della banda di conduzione:

$$q\chi_S = E_0 - E_c \tag{3.1}$$

e rappresenta, per un semiconduttore, l'energia minima che deve essere fornita ad un elettrone in banda di conduzione per poterlo allontanare dal materiale: per il Si, essa vale $q\chi_S = 5,05$ eV.

Il lavoro di estrazione, invece, rappresenta la distanza tra il livello del vuoto ed il livello di Fermi:

$$q\Phi_S = E_0 - E_F \tag{3.2}$$

Nel caso di un semiconduttore, quindi, $q\Phi_S$ dipende, oltre che dal materiale, dal tipo e dal livello del drogaggio (si veda la discussione nel paragrafo 1.7.2). Naturalmente, noto il valore del drogaggio le considerazioni del paragrafo 1.8 consentono di valutare la posizione del livello di Fermi rispetto ad E_c e E_v, e quindi di calcolare il lavoro di estrazione del materiale. Assumendo di trovarsi a temperatura ambiente e con livelli di drogaggio tali da avere campioni non degeneri, per i due lati p ed n si ha,

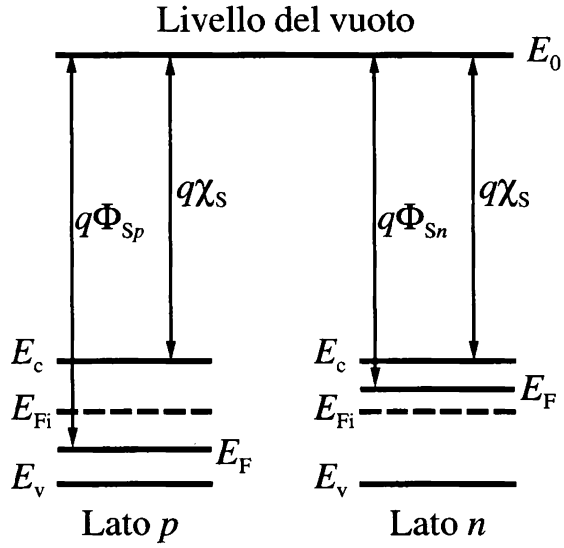


Figura 3.2 Diagrammi a bande di energia per i due lati della giunzione *pn* brusca assunti isolati.

rispettivamente:

$$q\Phi_{Sp} = q\chi_S + E_g - (E_F - E_v) = q\chi_S + E_g - k_B T \ln \frac{N_v}{N_A} \tag{3.3a}$$

$$q\Phi_{Sn} = q\chi_S + (E_c - E_F) = q\chi_S + k_B T \ln \frac{N_c}{N_D} \tag{3.3b}$$

Una volta definiti i diagrammi a bande di equilibrio dei lati isolati, il ragionamento che consente di determinare, almeno in forma qualitativa, il diagramma a bande del sistema complessivo in equilibrio termodinamico si basa sull'ipotesi di poter realizzare il seguente esperimento concettuale: formare, ad un certo istante, la giunzione ed osservare un transitorio che, dopo un tempo sufficientemente lungo, possa portare il dispositivo verso la condizione di equilibrio termodinamico. Ovviamente, un tale esperimento non ha significato pratico, poiché il funzionamento della giunzione dipende necessariamente dalla capacità di realizzare una transizione tra la regione n e quella p senza perdere la periodicità spaziale del reticolo cristallino, caratteristica non realizzabile con una operazione di saldatura, per quanto sofisticata, tra materiali separati.

Nell'approfondimento 3.1 si dimostra come, in equilibrio termodinamico, il livello di Fermi E_F debba essere *costante* in tutti i punti di qualunque dispositivo a semiconduttore. Da questa regola, segue che a seguito dell'instaurarsi della giunzione, durante il transitorio verso la condizione di equilibrio termodinamico, occorre che la distanza tra i livelli di Fermi dei due materiali isolati si annulli. Poiché, come discusso nel paragrafo 1.7.2, il livello di Fermi è una misura del numero di elettroni liberi in equilibrio nel materiale, per ridurre la differenza tra i livelli di Fermi nei due materiali isolati occorre che si realizzi uno *spostamento netto di elettroni dal materiale con livello di*

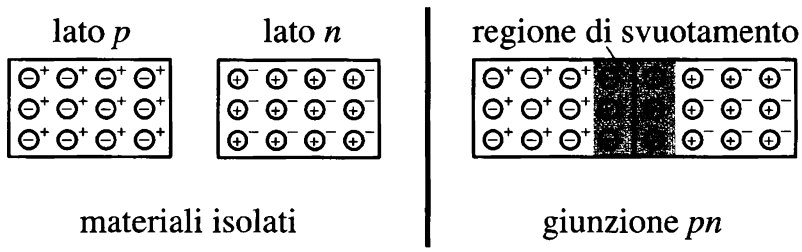


Figura 3.3 Formazione della regione di svuotamento a cavallo della giunzione *pn*: materiali isolati (a sinistra) e giunzione in equilibrio termodinamico (a destra). Nella rappresentazione delle cariche, si sono trascurati i portatori minoritari nei due lati. Le cariche fisse ionizzate sono rappresentate dai circoletti.

Fermi maggiore verso quello con livello di Fermi minore: nel caso della giunzione *pn*, ciò si traduce in uno spostamento, per diffusione, di elettroni dal lato n verso il lato p. Visto che i due materiali isolati sono localmente neutri, lo spostamento netto di elettroni determina il formarsi di (si veda la figura 3.3):

- una regione di carica positiva nel lato n, vicino alla giunzione, che corrisponde agli atomi droganti donatori ionizzati non più compensati dagli elettroni liberi;
- una regione di carica negativa nel lato p, vicino alla giunzione, che corrisponde agli atomi droganti accettatori ionizzati non più compensati dalle lacune libere.

Questa regione non neutra a cavallo della giunzione metallurgica viene detta *regione di carica spaziale* o *regione di svuotamento,* poiché corrisponde ad uno spazio nel quale le cariche libere di muoversi sono in numero trascurabile.

Una volta definita la regola della costanza del livello di Fermi, altre se ne possono enunciare per rendere possibile la costruzione qualitativa del diagramma a bande di un dispositivo omogeneo a tratti:

- è ragionevole attendersi come, ad una distanza sufficiente dalla giunzione, il diagramma a bande torni ad essere quello del *materiale isolato;*
- le caratteristiche proprie del semiconduttore, non dipendenti da drogaggio, sono *indipendenti dalla posizione,* come ad esempio l'affinità elettronica e l'ampiezza della banda proibita. Da questa considerazione, segue che le curve che rappresentano E_0, E_c, E_{Fi} ed E_v sono parallele tra loro, e distanti l'una dall'altra, rispettivamente, di $q\chi_S$, $E_g/2$ ed $E_g/2$.
- il livello energetico del vuoto E_0 è una *funzione continua* della posizione.

Approfondimento 3.1 In questo approfondimento si dimostra, seguendo la trattazione riportata ad esempio in [2], come il livello di Fermi in qualunque sistema fisico caratterizzato dalla presenza di elettroni e lacune libere di muoversi sia, in equilibrio termodinamico, indipendente dalla posizione.

Si consideri, per semplicità, un sistema fisico costituito da due materiali diversi dotati di elettroni e lacune libere. Considereremo esplicitamente solo il caso degli elettroni, poiché l'estensione del ragionamento alle lacune è ovvia. Siano, rispettivamente, $N_1(E)$ ed $f_1(E)$ le densità degli stati disponibili e la relativa probabilità di occupazione per il materiale 1, entrambe caratterizzate, in equilibrio, dalla temperatura e dal livello di Fermi. Analogamente, siano $N_2(E)$ ed $f_2(E)$ le

analoghe quantità per il materiale 2. Assumendo di aver messo in contatto i due sistemi (giunzione), e di trovarsi in equilibrio termodinamico, la condizione di assenza di scambi energetici con l'esterno del sistema fisico complessivo richiede che non si abbia uno spostamento netto di elettroni tra i due materiali costituenti, ovvero che la corrente che attraversa la giunzione sia nulla.
Fissato un valore E di energia, la probabilità associata all'attraversamento della giunzione di un elettrone da 1 verso 2, indicata dal simbolo $P_{1\to 2}(E)$, è proporzionale al prodotto del numero di elettroni presenti in 1 per tale energia, per il numero di posti disponibili, nel materiale 2, ad essere occupati dagli elettroni che hanno effettuato la transizione:

$$P_{1\to 2}(E) = K \times [N_1(E) f_1(E)] \times [N_2(E)(1 - f_2(E))]$$

Analogamente, la probabilità di transizione da 2 a 1 varrà:

$$P_{2\to 1}(E) = K \times [N_2(E) f_2(E)] \times [N_1(E)(1 - f_1(E))]$$

Poiché all'equilibrio termodinamico il flusso netto di elettroni deve essere nullo, dalla condizione $P_{1\to 2}(E) = P_{2\to 1}(E)$, valida qualunque sia E, segue

$$f_1(E) = f_2(E)$$

Visto che, in equilibrio termodinamico, la temperatura T deve essere la stessa in tutto il sistema complessivo, l'unico modo per soddisfare la precedente condizione indipendentemente dal valore di E è avere lo stesso livello di Fermi in tutti i lati della giunzione.
Questo ragionamento può essere facilmente esteso ad un numero arbitrario di giunzioni. Inoltre, poiché in equilibrio termodinamico elettroni e lacune sono caratterizzati dallo stesso valore di E_F, da queste considerazioni segue direttamente la costanza del livello di Fermi in qualunque punto di un sistema in equilibrio costituito da materiali, conduttori o semiconduttori, che formino una giunzione.
Si noti inoltre come, a rigore, la condizione di costanza del livello di Fermi non segua necessariamente dall'equilibrio termodinamico, ma dall'assenza di un flusso di carica tra i materiali costituenti la giunzione, purché non vi siano differenze di temperatura nel sistema complessivo.

3.1.1 Diagramma a bande in una regione di carica spaziale

La presenza di una regione non neutra nel sistema, ovvero una regione spaziale nella quale $\rho \neq 0$, ha una conseguenza diretta sulla forma del diagramma a bande di energia. Infatti, per l'equazione di Gauss (2.39), ad essa corrisponde una distribuzione di campo elettrico anch'essa non nulla, e quindi, per l'equazione (2.40), una distribuzione $\varphi(x) \neq 0$ di potenziale elettrostatico. Poiché il diagramma a bande è una rappresentazione dell'energia potenziale degli elettroni, e l'energia potenziale per un elettrone associata ad una distribuzione di potenziale è pari a $-q\varphi(x)$, la forma del diagramma a bande coincide con la forma di $-\varphi(x)$. In generale, una scelta possibile per la definizione di potenziale elettrostatico è quella che conduce all'espressione [2]:

$$E_{\mathrm{Fi}}(x) = -q\varphi(x) \tag{3.4}$$

Ricavando dalla (3.4) $\varphi(x) = -E_{\mathrm{Fi}}(x)/q$ e sostituendo nella equazione di Poisson (2.41c), si ottiene una relazione diretta tra il diagramma a bande e la densità di carica

$$\frac{\mathrm{d}^2 E_{\mathrm{Fi}}}{\mathrm{d}x^2} = \frac{q}{\epsilon}\rho \tag{3.5}$$

Questa relazione dimostra come il segno della derivata seconda del diagramma a bande, ovvero la sua curvatura, sia lo stesso della densità di carica, pertanto (si veda la figura 3.4)

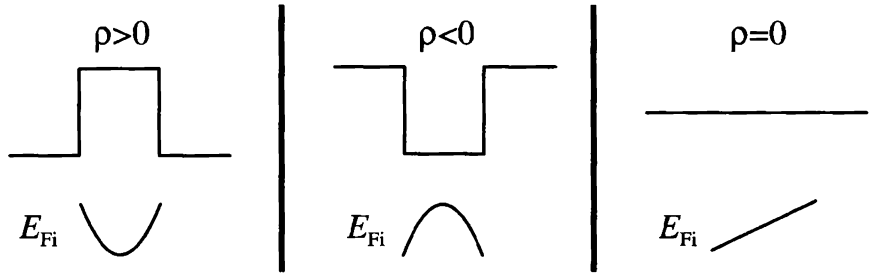


Figura 3.4 Relazione qualitativa tra il segno della carica netta in una regione di semiconduttore e la curvatura del diagramma a bande: a una carica positiva corrispondono bande con curvatura verso l'alto (a sinistra), a una carica negativa bande con curvatura verso il basso (al centro). In una regione neutra, le bande sono lineari (a destra).

- una regione di carica positiva $\rho > 0$, presenta bande con curvatura *verso l'alto;*
- una regione di carica negativa $\rho < 0$, presenta bande con curvatura *verso il basso;*
- una regione neutra $\rho = 0$, presenta bande con curvatura nulle, ovvero *rettilinee.*

In particolare, in una regione neutra si ha un campo elettrico costante:

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}x} = \frac{\rho}{\epsilon} = 0 \Longrightarrow \mathcal{E} = \mathcal{E}_0 = \mathrm{cost} \tag{3.6}$$

e quindi:

$$\frac{\mathrm{d}\varphi}{\mathrm{d}x} = -\mathcal{E}(x) = -\mathcal{E}_0 \Longrightarrow \varphi(x) = -\mathcal{E}_0 x + k \tag{3.7}$$

dove k è una costante di integrazione dipendente dalla scelta del riferimento per il potenziale. La (3.7) corrisponde alla terza delle precedenti osservazioni, infatti il diagramma a bande risulta essere rettilineo:

$$\rho = 0 \Longrightarrow E_{\mathrm{Fi}}(x) = -q\varphi(x) = q\mathcal{E}_0 x + qk \tag{3.8}$$

Se nella regione neutra il campo è nullo, $\mathcal{E}_0 = 0$ e il diagramma a bande risulta essere *orizzontale:* si parla in questo caso di *banda piatta.* Naturalmente, queste considerazioni valgono anche in senso opposto: una regione di bande rettilinee (e, in particolare, piatte) corrisponde, per la (3.5), ad una regione neutra ($\rho = 0$).

3.1.2 Costruzione qualitativa del diagramma a bande di equilibrio per una giunzione *pn* brusca

La costruzione del diagramma a bande di energia in equilibrio per una giunzione brusca viene utilizzata come paradigma per descrivere una metodologia più generale, che può essere applicata anche per altri dispositivi. Il primo passo, mostrato nella figura 3.5, consiste nel tracciare il livello di Fermi, costante in tutta la struttura, e i diagrammi a bande nelle due regioni neutre, lontano dalla giunzione, dove le bande sono piatte[1] e

[1] Infatti, in equilibrio la corrente che attraversa il sistema è nulla, e quindi tale è anche la caduta di potenziale sulle due regioni neutre, garantendo così un campo elettrico pari a zero.

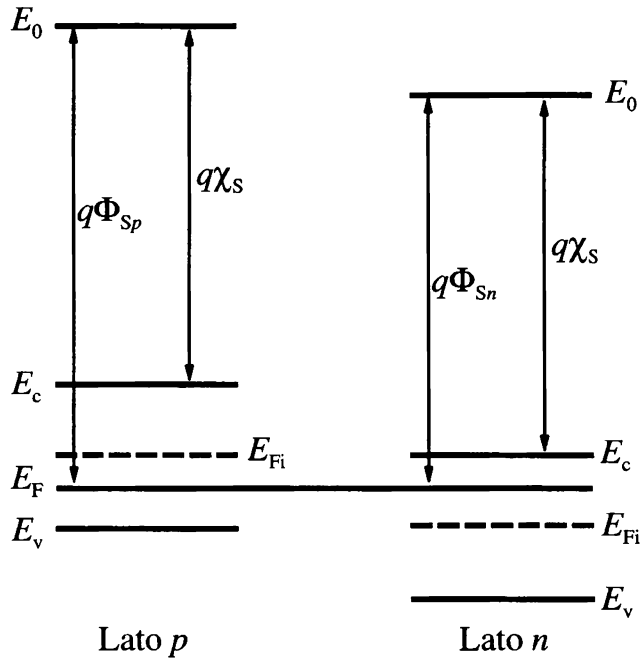


Figura 3.5 Primo passo nella costruzione qualitativa del diagramma abande di energia in equilibrio per una giunzione *pn* brusca: E_F è costante e le bande nelle regioni neutre sono piatte.

pari a quelle dei due materiali isolati.

Grazie alla discussione condotta precedentemente, è noto che a cavallo della giunzione si viene a formare una regione svuotata di portatori liberi, e pertanto non neutra. Assumendo che la transizione tra la regione svuotata e quella neutra nei due lati sia brusca, ovvero ipotizzando la cosiddetta *approssimazione di completo svuotamento*, la densità volumica di carica netta nel sistema risulta essere quella rappresentata nella figura 3.6. Conseguentemente, nel diagramma a bande si hanno due regioni con bande curve:

- ▷ nella regione svuotata nel lato p la curvatura è verso l'alto;
- ▷ nella regione svuotata nel lato n la curvatura è verso il basso.

Nella regione di svuotamento, la curvatura delle bande corrisponde alla presenza di un campo elettrico $\mathcal{E} \neq 0$, che quindi determinerà una corrente di trascinamento accelerando i pochi portatori liberi presenti nella giunzione: tale corrente di trascinamento avrà un valore ed un verso tali da compensare la corrente di diffusione che corrisponde alla tendenza dei portatori maggioritari nei due lati hanno a diffondere verso il lato opposto, come deve accadere per garantire la condizione di equilibrio termodinamico.

Grazie alla continuità del livello del vuoto E_0, e al fatto che questo è, in una

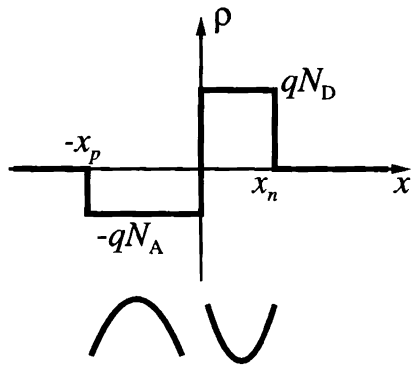


Figura 3.6 Andamento della densità di carica netta nella regione di svuotamento per una giunzione *pn* brusca all'equilibrio nell'approssimazione di completo svuotamento.

struttura costituita da un solo semiconduttore, parallelo al livello di Fermi intrinseco:

$$E_0(x) = E_{Fi}(x) + \frac{E_g}{2} + q\chi_S \tag{3.9}$$

si ha facilemente la forma qualitativa del diagramma a bande complessivo mostrata nella figura 3.7.

Da una analisi della figura 3.7 si vede come ai capi della regione svuotata si venga a formare una barriera di energia potenziale, tale da opporsi alla diffusione dei portatori maggioritari verso il lato opposto e corrispondente al campo elettrico che determina un contributo di trascinamento come descritto precedentemente. L'ampiezza della barriera di energia potenziale costituisce il cosiddetto *potenziale di contatto* o di *built-in* della giunzione:

$$qV_{bi} = E_0(-x_p) - E_0(x_n) = E_{Fi}(-x_p) - E_{Fi}(x_n) \tag{3.10}$$

Grazie alla costanza del livello di Fermi, è immediato verificare come il potenziale di contatto possa essere calcolato come differenza tra i lavori di estrazione nei due lati isolati

$$qV_{bi} = q\Phi_{sp} - q\Phi_{sn} = E_g - k_B T \log \frac{N_c N_v}{N_A N_D} \tag{3.11}$$

dove si è fatto uso delle relazioni (3.3). Poiché dalla (1.26) si ricava

$$E_g = k_B T \log \frac{N_c N_v}{n_i^2} \tag{3.12}$$

sostituendo nella (3.11) si ottiene il valore del potenziale di contatto

$$V_{bi} = V_T \log \frac{N_A N_D}{n_i^2} \tag{3.13}$$

essendo $V_T = k_B T/q$.

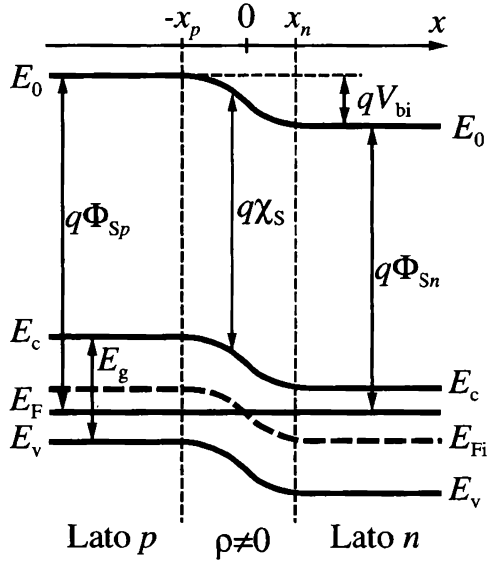


Figura 3.7 Diagramma a bande qualitativo di equilibrio termodinamico in una giunzione *pn* brusca.

3.2 Elettrostatica di equilibrio nella giunzione *pn*

La costruzione qualitativa del diagramma a bande descritta nel paragrafo 3.1 può essere resa quantitativa valutando esplicitamente la distribuzione di campo elettrico e di potenziale elettrostatico in condizioni di equilibrio termodinamico, ovvero per corrente nulla. La determinazione di $\phi(x)$ consente, inoltre, di calcolare le estensioni x_n e x_p delle regioni svuotate nei due lati: usando le equazioni (3.4) e (3.10), infatti, di ha una relazione tra le due incognite

$$V_{bi} = \varphi(x_n) - \varphi(-x_p) \tag{3.14}$$

dove il valore della tensione di built-in è dato, una volta noti il semiconduttore ed i livelli di drogaggio, dalla (3.13). Naturalmente, la per poter calcolare esplicitamente le due ampiezze occorre aggiungere all'equazione precedente un'altra relazione da essa indipendente. Questa può essere ricavata applicando la legge di Gauss in forma integrale ad un volume Ω che racchiuda completamente la regione di carica spaziale, come mostrato nella figura 3.8, ottenendo

$$\int_\Sigma \epsilon \boldsymbol{\mathcal{E}} \cdot \hat{n} \, \mathrm{d}\sigma = \int_\Omega \rho \, \mathrm{d}V \tag{3.15}$$

dove Σ è la superficie che racchiude il volume Ω, $\hat{n}$ il versore normale esterno a Σ, e $\boldsymbol{D} = \epsilon \boldsymbol{\mathcal{E}}$ è il *vettore spostamento dielettrico*. L'integrale a primo membro è certamente nullo in equilibrio termodinamico, poiché:

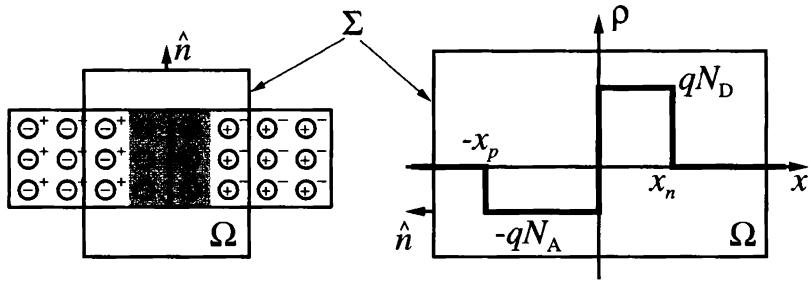


Figura 3.8 Volume Ω utilizzato per l'applicazione della legge di Gauss in forma integrale.

- ▷ nelle regioni neutre il campo è nullo, e quindi tale è il contributo al flusso uscente da Σ attraverso le facce ortogonali all'asse x;
- ▷ nella regione svuotata il campo elettrico non è nullo, ma è diretto lungo l'asse x mentre sulle facce di Σ comprese in tale regione $\hat{n}$ è diretto perpendicolarmente ad esso.

Si ricava così la *condizione di neutralità*

$$\int_{\Omega} \rho \, \mathrm{d}V = 0 \Longleftrightarrow N_{\mathrm{A}} x_p = N_{\mathrm{D}} x_n \tag{3.16}$$

che, insieme alla (3.14), costituisce un sistema algebrico nelle due incognite x_n e x_p.

La valutazione della distribuzione di campo elettrico $\mathcal{E}(x)$ può essere condotta risolvendo l'equazione di Gauss in forma differenziale (2.39), che può essere risolta tenendo conto della condizione al controrno corrispondente al fatto che il campo elettrico si annulla nella regione neutra:

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}x} = \frac{\rho}{\epsilon} \tag{3.17}$$

Nelle due regioni neutre, dove $\rho = 0$, il campo elettrico è costante. Nella regione svuotata nel lato p ($-x_p \leq x < 0$), si ha

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}x} = -\frac{qN_{\mathrm{A}}}{\epsilon} \tag{3.18}$$

ovvero

$$\mathcal{E}(x) = -\frac{qN_{\mathrm{A}}}{\epsilon} x + c_1 \tag{3.19}$$

Tenendo conto della condizione al contorno, si ricava il valore di c_1

$$\mathcal{E}(-x_p) = \frac{qN_{\mathrm{A}}}{\epsilon} x_p + c_1 = 0 \Longrightarrow c_1 = -\frac{qN_{\mathrm{A}}}{\epsilon} x_p \tag{3.20}$$

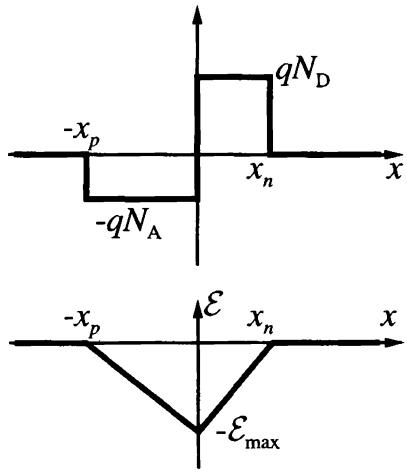


Figura 3.9 Rappresentazione della densità di carica netta e del campo elettrico in una giunzione *pn* brusca all'equilibrio termodinamico.

Analogamente, nella regione svuotata nel lato n ($0 \leq x < x_n$), l'equazione di Gauss diviene

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}x} = \frac{qN_D}{\epsilon} \tag{3.21}$$

che ha per soluzione generale

$$\mathcal{E}(x) = \frac{qN_D}{\epsilon}x + c_2 \tag{3.22}$$

La costante c_2 può essere ricavata imponendo la continuità di $\mathcal{E}(x)$ in $x = 0$

$$\mathcal{E}(0^-) = -\frac{qN_A}{\epsilon}x_p = \mathcal{E}(0^+) = c_2 \tag{3.23}$$

Per la condizione di neutralià (3.16) si ha anche

$$c_2 = -\frac{qN_A}{\epsilon}x_p = -\frac{qN_D}{\epsilon}x_n \tag{3.24}$$

Riunendo i risultati, si ottiene l'espressione complessiva per il campo elettrico

$$\mathcal{E}(x) = \begin{cases} 0 & x < -x_p \\ -\frac{qN_A}{\epsilon}(x + x_p) & -x_p \leq x < 0 \\ \frac{qN_D}{\epsilon}(x - x_n) & 0 \leq x < x_n \\ 0 & x \geq x_n \end{cases} \tag{3.25}$$

È interessante notare come, grazie alla condizione di neutralità, la continuità del campo elettrico in $x = 0$ garantisca anche l'annullarsi di $\mathcal{E}$ nella regione neutra nel lato n. La distribuzione di campo elettrico è mostrata nella figura 3.9, dove si osserva come il massimo di intensità per il campo elettrico si collochi esattamente alla giunzione metallurgica, con un valore

$$\mathcal{E}_{\max} = -\mathcal{E}(0) = \frac{qN_D}{\epsilon} x_n = \frac{qN_A}{\epsilon} x_p \tag{3.26}$$

Noto $\mathcal{E}(x)$, il potenziale viene valutato utilizzando la definizione

$$\frac{d\varphi}{dx} = -\mathcal{E}(x) \tag{3.27}$$

completata da una condizione al contorno, che corrisponde alla scelta del riferimento di potenziale. Essendo questo arbitrario, si può liberamente scegliere di porre a zero il potenziale in un punto qualunque.

Nelle due regioni neutre il potenziale è costante, e una scelta possibile è porre

$$\varphi(x) = 0 \qquad x < -x_p \tag{3.28}$$

Nella regione svuotata nel lato p si ha

$$\frac{d\varphi}{dx} = \frac{qN_A}{\epsilon}(x + x_p) \tag{3.29}$$

dalla quale segue

$$\varphi(x) = \frac{qN_A}{2\epsilon}(x + x_p)^2 + k_1 \tag{3.30}$$

dove la costante k_1 è definita dalla condizione di continuità del potenziale in $-x_p$

$$\varphi(-x_p^-) = 0 = \varphi(-x_p^+) = k_1 \tag{3.31}$$

Inoltre, nella regione svuotata nel lato n il potenziale soddisfa la condizione

$$\frac{d\varphi}{dx} = -\frac{qN_D}{\epsilon}(x - x_n) \tag{3.32}$$

dalla quale

$$\varphi(x) = -\frac{qN_D}{2\epsilon}(x - x_n)^2 + k_2 \tag{3.33}$$

dove k_2 è definita dalla condizione di continuità in $x = 0$

$$\varphi(0^-) = \frac{qN_A}{2\epsilon} x_p^2 = \varphi(0^+) = -\frac{qN_D}{2\epsilon} x_n^2 + k_2 \tag{3.34}$$

per cui

$$k_2 = \frac{qN_A}{2\epsilon} x_p^2 + \frac{qN_D}{2\epsilon} x_n^2 \tag{3.35}$$

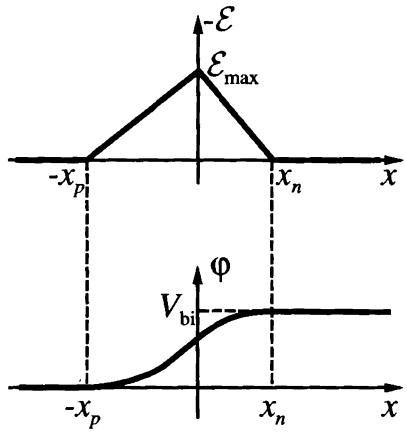


Figura 3.10 Rappresentazione di $-\mathcal{E}$ e del potenziale elettrostatico in una giunzione *pn* brusca all'equilibrio termodinamico.

Complessivamente, quindi, si ha l'andamento mostrato nella figura 3.10, ovvero

$$\varphi(x) = \begin{cases} 0 & x < -x_p \\ \frac{qN_A}{2\epsilon}(x + x_p)^2 & -x_p \le x < 0 \\ -\frac{qN_D}{2\epsilon}(x - x_n)^2 + \frac{qN_A}{2\epsilon}x_p^2 + \frac{qN_D}{2\epsilon}x_n^2 & 0 \le x < x_n \\ \frac{qN_A}{2\epsilon}x_p^2 + \frac{qN_D}{2\epsilon}x_n^2 & x \ge x_n \end{cases} \tag{3.36}$$

Grazie alla valutazione esplicita di $\varphi(x)$, è ora possibile identificare il sistema algebrico che consente di valutare l'ampiezza della regione svuotata, costituito dalle equazioni (3.14) e (3.16):

$$\begin{cases} V_{bi} = \frac{qN_A}{2\epsilon}x_p^2 + \frac{qN_D}{2\epsilon}x_n^2 \\ N_A x_p = N_D x_n \end{cases} \tag{3.37}$$

Per sostituzione, è facile risolvere il sistema precedente ottenendo le espressioni

$$x_n = \sqrt{\frac{2\epsilon}{qN_D}\,\frac{N_A}{N_A + N_D}\,V_{bi}} = \sqrt{\frac{2\epsilon}{q}\,\frac{N_{eq}}{N_D^2}\,V_{bi}} \tag{3.38a}$$

$$x_p = \sqrt{\frac{2\epsilon}{qN_A}\,\frac{N_D}{N_A + N_D}\,V_{bi}} = \sqrt{\frac{2\epsilon}{q}\,\frac{N_{eq}}{N_A^2}\,V_{bi}} \tag{3.38b}$$

$$x_d = x_n + x_p = \sqrt{\frac{2\epsilon}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)V_{bi}} = \sqrt{\frac{2\epsilon}{q}\,\frac{1}{N_{eq}}\,V_{bi}} \tag{3.38c}$$

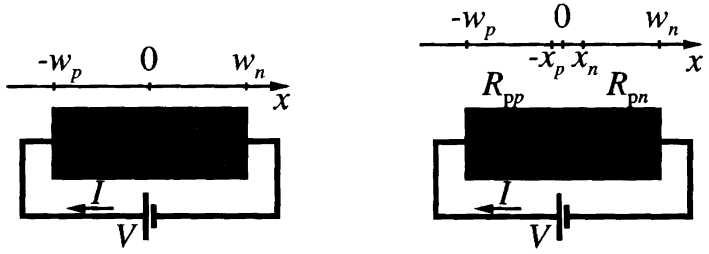


Figura 3.11 Convenzione adottata per la tensione applicata e la corrente nella giunzione *pn* (a sinistra), e definizione delle resistenze parassite corrispondenti alle regioni neutre (a destra).

dove si è definito il *drogaggio equivalente* della giunzione brusca

$$N_{\text{eq}} = \frac{N_{\text{A}} N_{\text{D}}}{N_{\text{A}} + N_{\text{D}}} = N_{\text{A}} \parallel N_{\text{D}} \tag{3.39}$$

Sempre dall'espressione di $\varphi(x)$, è facile verificare come la caduta di potenziale su una regione di carica costante sia proporzionale al quadrato dell'ampiezza di tale regione, infatti

$$\varphi(0) - \varphi(-x_p) = \frac{qN_{\text{A}}}{2\epsilon} x_p^2 \qquad \varphi(x_n) - \varphi(0) = \frac{qN_{\text{D}}}{2\epsilon} x_n^2 \tag{3.40}$$

3.3 La giunzione *pn* fuori equilibrio in condizioni stazionarie

Una volta studiata la giunzione *pn* brusca in condizioni di equilibrio termodinamico, il passo successivo consiste nell'analizzare il funzionamento del dispositivo fuori equilibrio termodinamico, ed in particolare applicando alla giunzione un generatore di tensione. Questo studio consentirà di ricavare le relazioni che caratterizzano il funzionamento elettrico della giunzione, ovvero quella che, nell'ambito della teoria dei circuiti, viene detta *relazione costitutiva* del bipolo. Il modello verrà inizialmente ricavato in condizioni statiche, per poi generalizzare l'analisi ai segnali elettrici tempo-varianti.

3.3.1 Corrente nella giunzione fuori equilibrio termodinamico

Nel caso più semplice, alla giunzione viene applicata una tensione V costante nel tempo. La convenzione di segno scelta per l'analisi del dispositivo è indicata nella parte sinistra della figura 3.11: la tensione viene applicata misurandola sul lato p rispetto al lato n. Corrispondentemente, la corrente stazionaria I viene misurata entrante nel lato p ed uscente dal lato n. Si indicano con w_p e w_n, rispettivamente, le lunghezze fisiche dei lati p ed n.

La parte destra della figura 3.11 evidenzia come, nella giunzione *pn*, si possano identificare tre regioni che concorrono al funzionamento elettrico del dispositivo: le due regioni quasi-neutre, esterne alla regione di svuotamento, e la regione di svuotamento stessa. Poiché la struttura del dispositivo è monodimensionale, la corrente I può solo essere costante in tutti i punti dell'asse x, e quindi attraversa tutte e tre le regioni che, quindi, si trovano ad essere collegate in serie. Visto che, in generale, la corrente I

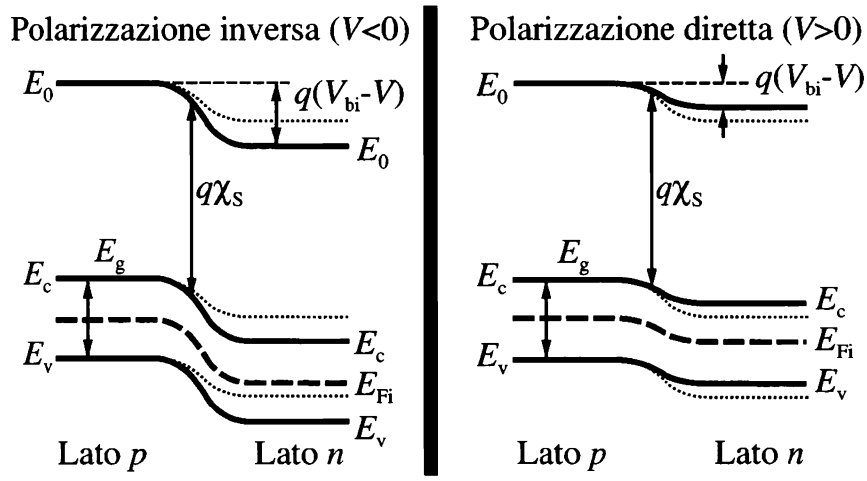


Figura 3.12 Diagrammi a bande fuori equilibrio termodinamico, nell'approssimazione di campo elettrico nullo nelle regioni quasi-neutre, per una giunzione *pn* in polarizzazione inversa (a sinistra) e diretta (a destra). Le curve a punti rappresentano il diagramma a bande di equilibrio termodinamico.

non è nulla fuori equilibrio, le due regioni quasi neutre non possono essere ritenute a campo esattamente nullo. In *basso livello di iniezione,* però, si può ritenere che il loro comportamento elettrico sia prevalentemente resistivo, essendo caratterizzate da un numero di portatori maggioritari pari al drogaggio e, di conseguenza, da una resistività sostanzialmente costante. Per questo motivo, si assume che le due regioni quasi-neutre siano caratterizzate da una resistenza R_{pp} e R_{pn}, rispettivamente, e che quindi ai loro capi si determini la caduta di tensione $(R_{pp} + R_{pn})I = R_p I$. L'ipotesi fondamentale che si effettua è che $R_p|I| \ll |V|$, ovvero che *tutta la tensione applicata V vada a modificare la caduta di potenziale sulla regione svuotata.* In queste condizioni, inoltre, le due regioni quasi-neutre sono caratterizzate da un campo elettrico molto piccolo, in prima approssimazione pari a zero.

Visto che la tensione V è misurata positiva sul lato p rispetto a quello n, e che ad essa corrisponde per gli elettroni una energia potenziale pari a $-qV$, l'effetto dell'applicazione di V è di modificare la caduta di energia potenziale sulla regione svuotata portandola al valore $q(V_{bi} - V)$ (figura 3.12). L'effetto della tensione applicata, pertanto, dipende dal segno di V:

- per $V < 0$ si parla di *polarizzazione inversa* della giunzione, e la barriera di energia potenziale risulta essere di ampiezza maggiore rispetto all'equilibrio termodinamico. Pertanto, la diffusione di portatori maggioritari di ogni lato verso il lato opposto risulta essere sfavorita. Prevale, quindi, il trascinamento dei pochi portatori minoritari disponibili vicino alla regione svuotata, elettroni dal lato p verso n e lacune da n verso p: si ha pertanto una corrente I negativa e di piccolo valore;
- per $V > 0$ la giunzione è in *polarizzazione diretta,* e la barriera di energia potenziale ai capi della regione svuotata risulta essere ridotta, favorendo così la diffusione degli elettroni dal lato n verso il p, e delle lacune nella direzione opposta. Pertanto, la

corrente I risulta essere di valore positivo, ed in grado di crescere rapidamente con V a causa della grande disponibilità di portatori maggioritari nei due lati.

Per rendere quantitativa questa analisi, ovvero valutare la *caratteristica statica* $I = I(V)$ del dispositivo, si può procedere come segue. Poiché si sta considerando una struttura monodimensionale, la legge di conservazione della corrente richiede che la densità di corrente totale sia costante in ogni sezione x: indicando con A la sezione trasversale al flusso di corrente, supposta costante, si ha $I = JA$ indipendentemente dalla posizione x. Per stimare I, si assumono due ipotesi, coerenti con la discussione precedente sul diagramma a bande:

- ▷ nelle regioni quasi-neutre, la caduta di potenziale è abbastanza piccola da poter assumere $\mathcal{E} \approx 0$;
- ▷ nelle regioni quasi-neutre, vale l'assunzione di basso livello di iniezione, per la quale i portatori maggioritari risultano essere approssimativamente quelli di equilibrio termodinamico:

$$p_p(x) \approx p_{p0}(x) = N_A \qquad -w_p \leq x < -x_p \tag{3.41a}$$

$$n_n(x) \approx n_{n0}(x) = N_D \qquad x_n < x \leq w_n \tag{3.41b}$$

In una generica sezione x della giunzione, si può esprimere

$$J = J_{n,\mathrm{tr}}(x) + J_{n,\mathrm{diff}}(x) + J_{p,\mathrm{tr}}(x) + J_{p,\mathrm{diff}}(x) \tag{3.42}$$

A seconda della posizione x, però, è possibile effettuare delle semplificazioni. In particolare, nelle regioni quasi-neutre in basso livello di iniezione, dove $\mathcal{E} \approx 0$, è certamente possibile trascurare la corrente di trascinamento dei portatori minoritari rispetto a quella dei portatori maggioritari:

- ▷ nel lato p, dove $-w_p \leq x < -x_p$

$$J_{n,\mathrm{tr}}(x) + J_{p,\mathrm{tr}}(x) = qn_p(x)\mu_n\mathcal{E}(x) + qp_p(x)\mu_n\mathcal{E}(x) \approx qp_{p0}\mu_n\mathcal{E}(x) = J_{p,\mathrm{tr}}(x) \tag{3.43a}$$

- ▷ nel lato n, per $x_n < x \leq w_n$

$$J_{n,\mathrm{tr}}(x) + J_{p,\mathrm{tr}}(x) = qn_n(x)\mu_n\mathcal{E}(x) + qp_n(x)\mu_n\mathcal{E}(x) \approx qn_{n0}\mu_n\mathcal{E}(x) = J_{p,\mathrm{tr}}(x) \tag{3.43b}$$

Sui portatori maggioritari, invece, non è possibile effettuare semplificazioni, in quanto il campo elettrico, di piccolo valore, risulta essere moltiplicato per la concentrazione di portatori maggioritari. Complessivamente, quindi, si ha

$$J \approx J_{n,\mathrm{diff}}(x) + J_p(x) \qquad -w_p \leq x < -x_p \tag{3.44a}$$

$$J \approx J_n(x) + J_{p,\mathrm{diff}}(x) \qquad x_n < x \leq w_n \tag{3.44b}$$

Per valutare le correnti di diffusione dei portatori minoritari nei due lati quasi-neutri, occorre determinare la dipendenza dalla posizione delle due concentrazioni in eccesso $n'_p(x)$ per $-w_p \leq x < -x_p$, e $p'_n(x)$ per $x_n < x \leq w_n$. Poiché in corrispondenza

del contatto elettrico alle estremità del dispositivo si assume la condizione di contatto ohmico, per la stima delle concentrazioni in eccesso si può fare uso dei risultati presentati nell'esempio 2.3: nel caso di lati *lunghi* rispetto alla rispettiva lunghezza di diffusione dei portatori minoritari ($w_p \gg L_n$ e $w_n \gg L_p$), la distribuzione spaziale di portatori minoritari in eccesso ha andamento esponenziale, tendendo a zero verso il contatto ohmico. Tenendo conto della scelta effettuata per l'origine dell'asse x, è facile scrivere le seguenti relazioni per le due concentrazioni in eccesso

$$n'_p(x) = n'_p(-x_p)\exp\left(\frac{x+x_p}{L_n}\right) \qquad p'_n(x) = p'_n(x_n)\exp\left(-\frac{x-x_n}{L_p}\right) \tag{3.45}$$

Conseguentemente, applicando le relazioni (2.15) che definiscono le correnti di diffusione si ha

$$J_{n,\text{diff}}(x) = +qD_n\frac{\partial n'_p}{\partial x} = +\frac{qD_n}{L_n}n'_p(-x_p)\exp\left(\frac{x+x_p}{L_n}\right) \qquad -w_p \le x < -x_p \tag{3.46a}$$

$$J_{p,\text{diff}}(x) = -qD_p\frac{\partial p'_n}{\partial x} = +\frac{qD_p}{L_p}p'_n(x_n)\exp\left(-\frac{x-x_n}{L_p}\right) \qquad x_n < x \le w_n \tag{3.46b}$$

e quindi, al bordo della regione di svuotamento

$$J_{n,\text{diff}}(-x_p) = \frac{qD_n}{L_n}n'_p(-x_p) \tag{3.47a}$$

$$J_{p,\text{diff}}(x_n) = \frac{qD_p}{L_p}p'_n(x_n) \tag{3.47b}$$

Naturalmente, per completare la valutazione delle due correnti di diffusione è necessario determinare il valore degli eccessi di portatori ai bordi della regione di svuotamento, ovvero $n'_p(-x_p)$ e $p'_n(x_n)$.

Approfondimento 3.2 Nel caso i lati quasi-neutri siano corti rispetto alla relativa lunghezza di diffusione, ovvero se $w_p \ll L_n$ e $w_n \ll L_p$, la distribuzione di portatori minoritari in eccesso, come discusso nell'esempio 2.3, risulta essere lineare, ovvero

$$n'_p(x) = a_n x + b_n \quad -w_p \le x < -x_p \qquad p'_n(x) = a_p x + b_p \quad x_n < x \le w_n$$

Le costanti a e b possono essere determinate imponendo il valore delle concentrazioni in eccesso agli estremi delle due regioni di svuotamento:

$$\begin{cases} n'_p(-w_p) = 0 = -a_n w_p + b_n \\ n'_p(-x_p) = -a_n x_p + b_n \end{cases} \qquad \begin{cases} p'_n(w_n) = 0 = a_p w_n + b_p \\ p'_n(x_n) = a_p x_n + b_p \end{cases}$$

Risolvendo i due sistemi lineari, si trova

$$\begin{cases} a_p = n'_p(-x_p)/(w_p - x_p) \approx n'_p(-x_p)/w_p \\ b_p = n'_p(-x_p)w_p/(w_p - x_p) \approx n'_p(-x_p) \end{cases} \qquad \begin{cases} a_n = -p'_n(x_n)/(w_n - x_n) \approx p'_n(x_n)/w_n \\ b_n = p'_n(x_n)w_n/(w_n - x_n) \approx p'_n(x_n) \end{cases}$$

dove le approssimazioni corrispondono a trascurare l'ampiezza della regione di svuotamento rispetto alle lunghezze fisiche dei due lati, e quindi ad assumere $x_p \approx x_n \approx 0$. Infine, gli andamenti delle

concentrazioni in eccesso sono

$$n'_p(x) = n'_p(-x_p)\frac{w_p + x}{w_p - x_p} \approx n'_p(-x_p)\frac{w_p + x}{w_p} \qquad -w_p \le x < -x_p$$

$$p'_n(x) = p'_n(x_n)\frac{w_n - x}{w_n - x_n} \approx p'_n(x_n)\frac{w_n - x}{w_n} \qquad x_n < x \le w_n$$

cui corrispondono le correnti di diffusione

$$J_{n,\text{diff}}(x) = +qD_n\frac{\partial n'_p}{\partial x} = \frac{qD_n}{w_p - x_p}n'_p(-x_p) \approx \frac{qD_n}{w_p}n'_p(-x_p) \qquad -w_p \le x < -x_p$$

$$J_{p,\text{diff}}(x) = -qD_p\frac{\partial p'_n}{\partial x} = \frac{qD_p}{w_n - x_n}p'_n(x_n) \approx \frac{qD_p}{w_n}p'_n(x_n) \qquad x_n < x \le w_n$$

3.3.2 Legge della giunzione

La discussione nel precedente paragrafo ha evidenziato come la valutazione delle correnti di diffusione dei portatori minoritari nei due lati quasi-neutri richieda di stimare l'eccesso di portatori minoritari iniettati ai bordi della regione di svuotamento. Il calcolo di $n'_p(-x_p)$ e $p'_n(x_n)$ può essere condotto in diversi modi: lasciando all'approfondimento 3.3 una stima più rigorosa, si seguirà qui un approccio euristico. Si osserva che, in equilibrio termodinamico, si ha

$$n_{n0}(x_n) = N_\text{D}, \qquad p_{n0}(x_n) = n_\text{i}^2/N_\text{D} \tag{3.48}$$

$$p_{p0}(-x_p) = N_\text{A}, \qquad n_{p0}(-x_p) = n_\text{i}^2/N_\text{A} \tag{3.49}$$

e quindi l'espressione (3.13) per il potenziale di contatto può essere interpretata, equivalentemente, nei due modi seguenti

$$V_\text{bi} = V_T \log\frac{N_\text{A}N_\text{D}}{n_\text{i}^2} = V_T \log\frac{N_\text{D}}{n_\text{i}^2/N_\text{A}} = V_T \log\frac{n_{n0}(x_n)}{n_{p0}(-x_p)} \tag{3.50}$$

$$V_\text{bi} = V_T \log\frac{N_\text{A}N_\text{D}}{n_\text{i}^2} = V_T \log\frac{N_\text{A}}{n_\text{i}^2/N_\text{D}} = V_T \log\frac{p_{p0}(-x_p)}{p_{n0}(x_n)} \tag{3.51}$$

ricavando così

$$\frac{n_{p0}(-x_p)}{n_{n0}(x_n)} = \exp\left(-\frac{V_\text{bi}}{V_T}\right) \qquad \frac{p_{n0}(x_n)}{p_{p0}(-x_p)} = \exp\left(-\frac{V_\text{bi}}{V_T}\right) \tag{3.52}$$

In basso livello di iniezione, la caduta di potenziale sulla regione svuotata diviene $V_\text{bi} - V$ e, assumendo che i portatori liberi siano poco fuori equilibrio, si può ritenere che la relazione esponenziale (3.52) che lega le concentrazioni di carica alla barriera di energia potenziale tra i due lati continui a sussistere

$$\frac{n_p(-x_p)}{n_n(x_n)} \approx \exp\left(-\frac{V_\text{bi} - V}{V_T}\right) \qquad \frac{p_n(x_n)}{p_p(-x_p)} \approx \exp\left(-\frac{V_\text{bi} - V}{V_T}\right) \tag{3.53}$$

Sostituendo nella (3.53) le relazioni (3.41), corrispondenti all'ipotesi di basso livello di iniezione, e usando la (3.52) si ottiene

$$n_p(-x_p) \approx n_{n0}(\ \ x_n) \exp\left(-\frac{V_{\text{bi}}}{V_T}\right) \exp\left(\frac{V}{V_T}\right) = n_{p0}(-x_p) \exp\left(\frac{V}{V_T}\right) \tag{3.54a}$$

$$p_n(\ \ x_n) \approx p_{p0}(-x_p) \exp\left(-\frac{V_{\text{bi}}}{V_T}\right) \exp\left(\frac{V}{V_T}\right) = p_{n0}(\ \ x_n) \exp\left(\frac{V}{V_T}\right) \tag{3.54b}$$

Poiché gli eccessi di carica sono definiti come $n'_p = n_p - n_{p0}$ e $p'_n = p_n - p_{n0}$, si ottiene immediatamente la *legge della giunzione*

$$n'_p(-x_p) = n_{p0}(-x_p)\left[\exp\left(\frac{V}{V_T}\right) - 1\right] = \frac{n_{\text{i}}^2}{N_{\text{A}}}\left[\exp\left(\frac{V}{V_T}\right) - 1\right] \tag{3.55a}$$

$$p'_n(\ \ x_n) = p_{n0}(\ \ x_n)\left[\exp\left(\frac{V}{V_T}\right) - 1\right] = \frac{n_{\text{i}}^2}{N_{\text{D}}}\left[\exp\left(\frac{V}{V_T}\right) - 1\right] \tag{3.55b}$$

Approfondimento 3.3 La legge della giunzione può essere ricavata in termini più rigorosi facendo uso dei quasi-livelli di Fermi (si veda la discussione nell'approfondimento 2.3) di elettroni e lacune nelle due regioni quasi-neutre. In particolare, dall'ipotesi di basso livello di iniezione le (3.41) implicano che i quasi-livelli di Fermi dei portatori maggioritari nelle due regioni quasi-neutre siano costanti nelle regioni quasi-neutre[2]

$$E_{\text{F}p}(x) - E_{\text{v}}(x) = k_{\text{B}}T \log \frac{N_{\text{v}}}{N_{\text{A}}} \qquad -w_p \le x < -x_p$$

$$E_{\text{c}}(x) - E_{\text{F}n}(x) = k_{\text{B}}T \log \frac{N_{\text{c}}}{N_{\text{D}}} \qquad x_n < x \le w_n$$

Dal diagramma a bande mostrato nella figura 3.13, si ricava facilmente

$$\begin{aligned} E_{\text{F}n}(x_n) - E_{\text{F}p}(-x_p) &= [E_{\text{F}n}(x_n) - E_{\text{c}}(x_n)] + [E_{\text{c}}(x_n) - E_{\text{c}}(-x_p)] \\ &\quad + [E_{\text{c}}(-x_p) - E_{\text{v}}(-x_p)] + [E_{\text{v}}(-x_p) - E_{\text{F}p}(-x_p)] \\ &= -k_{\text{B}}T \log \frac{N_{\text{c}}}{N_{\text{D}}} - q\,(V_{\text{bi}} - V) + E_{\text{g}} - k_{\text{B}}T \log \frac{N_{\text{v}}}{N_{\text{A}}} \end{aligned}$$

che, per la (3.11), dimostra come la distanza tra di due quasi-livelli di Fermi degli elettroni nel lato n e delle lacune nel lato p sia esattamente pari (in basso livello di iniezione) alla tensione applicata

$$E_{\text{F}n}(x_n) - E_{\text{F}p}(-x_p) = qV$$

Per quanto riguarda i portatori minoritari, poiché come descritto in precedenza nelle regioni quasi-neutre la concentrazione in eccesso tende ad annullarsi mano a mano che si procede verso il contatto ohmico, il relativo quasi-livello, inizialmente differente da quello dei portatori maggioritari, tende a convergere verso quest'ultimo (si veda la figura 3.13). Inoltre, moltiplicando tra loro le definizioni di quasi-livello riportate nell'approfondimento 2.3 si ottiene:

$$n(x)p(x) = n_{\text{i}}^2 \exp\left(\frac{E_{\text{F}n}(x) - E_{\text{F}p}(x)}{k_{\text{B}}T}\right)$$

Pertanto, nelle due regioni quasi-neutre, per la (3.55), in caso di polarizzazione

[2] Naturalmente si tratta di una approssimazione, infatti se i quasi-livelli di Fermi dei maggioritari fossero esattamente costanti, la corrispondente corrente sarebbe identicamente nulla.

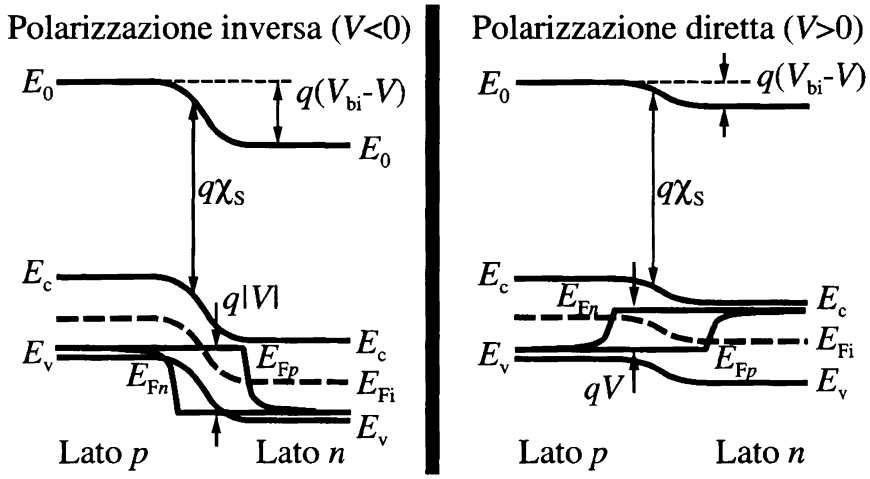


Figura 3.13 Diagrammi a bande fuori equilibrio termodinamico, nell'approssimazione di campo elettrico nullo nelle regioni quasi-neutre, per una giunzione *pn* in polarizzazione inversa (a sinistra) e diretta (a destra). Sono mostrati anche gli andamenti dei quasi-livelli di Fermi di elettroni e lacune.

- ▷ diretta, per la quale si ha un fenomeno di iniezioni di portatori minoritari ($n'_p(-x_p)$,$p'_n(x_n) > 0$), dovrà essere $E_{Fn} > E_{Fp}$
- ▷ inversa, per la quale si ha un fenomeno di svuotamento di portatori minoritari ($n'_p(-x_p)$,$p'_n(x_n) < 0$), dovrà essere $E_{Fn} < E_{Fp}$

Riunendo queste osservazioni, si arriva agli andamenti per i quasi-livelli di Fermi dei portatori minoritari e maggioritari nelle due regioni quasi neutre mostrati nella figura 3.13. Infine, nella regione di svuotamento si dimostra (si veda ad'esempio [2]) che, in basso livello di iniezione, i quasi-livelli di Fermi di elettroni e lacune sono approssimativamente costanti[3]. Di conseguenza, si verifica dalla figura 3.13 che

$$E_{Fn}(-x_p) - E_{Fp}(-x_p) = E_{Fn}(x_n) - E_{Fp}(x_n) = qV$$

e quindi

$$n_p(-x_p)p_p(-x_p) = n_i^2 \exp\left(\frac{V}{V_T}\right) \qquad n_n(x_n)p_n(x_n) = n_i^2 \exp\left(\frac{V}{V_T}\right)$$

Utilizzando nuovamente le condizioni (3.41) di basso livello di iniezione, si trova immediatamente la legge della giunzione (3.55).

3.3.3 Relazione tensione-corrente statica

Sostituendo la legge della giunzione (3.55) nelle (3.47), si trova un'espressione esplicita per le componenti di corrente di diffusione dei portatori minoritari al bordo della regione

[3] Come già ricordato, la costanza dei due quasi-livelli è solo approssimativamente vera, in quanto se essa fosse esatta la corrente totale nella regione svuotata sarebbe identicamente nulla.

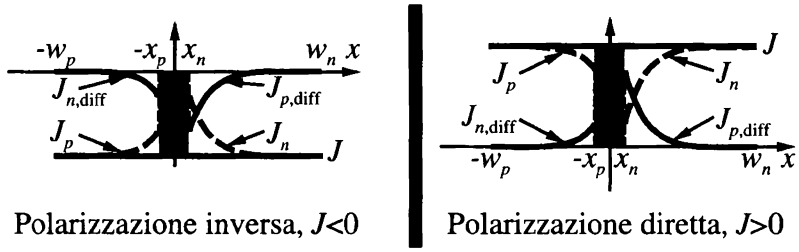


Figura 3.14 Rappresentazione della dipendenza dalla posizione delle correnti di portatori maggioritari e minoritari nelle due regioni neutre di una giunzione *pn* brusca: polarizzazione inversa (a sinistra) e diretta (a destra).

di svuotamento

$$J_{n,\mathrm{diff}}(-x_p) = \frac{qD_n}{L_n}\frac{n_\mathrm{i}^2}{N_\mathrm{A}}\left[\exp\left(\frac{V}{V_T}\right) - 1\right] \tag{3.56a}$$

$$J_{p,\mathrm{diff}}(x_n) = \frac{qD_p}{L_p}\frac{n_\mathrm{i}^2}{N_\mathrm{D}}\left[\exp\left(\frac{V}{V_T}\right) - 1\right] \tag{3.56b}$$

Pertanto, dalle (3.44) si può valutare

$$J = J_{n,\mathrm{diff}}(-x_p) + J_p(-x_p) = J_n(x_n) + J_{p,\mathrm{diff}}(x_n) \tag{3.57}$$

dove, per poter valutare esplicitamente la densità di corrente nella giunzione, occorre stimare le correnti di portatori maggioritari ai bordi della regione di svuotamento. Più in generale, dalle (3.44) è però possibile stimare graficamente l'andamento di $J_n(x)$ e $J_p(x)$ nelle due regioni neutre drogate n e p, rispettivamente, come il complementare delle correnti di diffusione dei portatori minoritari rispetto al valore totale J, indipendente dalla posizione: si veda la figura 3.14 per una rappresentazione grafica.

La determinazione delle densità di corrente $J_p(-x_p)$ e $J_n(x_n)$ richiede di introdurre una ulteriore ipotesi: gli effetti di generazione e ricombinazione nella regione di carica spaziale sono *trascurabili*. Dalle equazioni di continuità in condizioni stazionarie nel tempo, segue quindi

$$\frac{\mathrm{d}J_n}{\mathrm{d}x} = qU_n = 0 \qquad \frac{\mathrm{d}J_p}{\mathrm{d}x} = -qU_p = 0 \tag{3.58}$$

ovvero, le densità di corrente di elettroni e lacune sono costanti nella regione di svuotamento (figura 3.15). Ma allora, si ha immediatamente

$$J_n(x_n) = J_{n,\mathrm{diff}}(-x_p) \qquad J_p(-x_p) = J_{p,\mathrm{diff}}(x_n) \tag{3.59}$$

e quindi si può determinare la densità di corrente totale nella giunzione come somma delle correnti di diffusione dei portatori valutate agli estremi della regione di svuotamento

$$J = J_{n,\mathrm{diff}}(-x_p) + J_{p,\mathrm{diff}}(x_n) \tag{3.60}$$

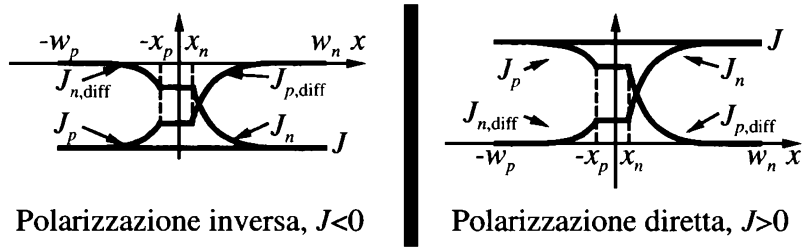


Figura 3.15 Rappresentazione della dipendenza dalla posizione delle correnti di portatori maggioritari e minoritari in una giunzione *pn* brusca: polarizzazione inversa (a sinistra) e diretta (a destra).

Sostituendo le relazioni (3.55) e ricordando che $I = JA$, si trova l'espressione della *caratteristica statica* della giunzione

$$I = I_\mathrm{s}\left[\exp\left(\frac{V}{V_T}\right) - 1\right] \tag{3.61}$$

dove I_s viene detta *corrente inversa di saturazione* della giunzione

$$I_\mathrm{s} = qA\left[\frac{D_n}{L_n}\frac{n_\mathrm{i}^2}{N_\mathrm{A}} + \frac{D_p}{L_p}\frac{n_\mathrm{i}^2}{N_\mathrm{D}}\right] \tag{3.62}$$

Approfondimento 3.4 Come discusso nell'approfondimento 3.2, la corrente di diffusione dei portatori minoritari nel caso di lati corti è approssimativamente costante in tutta la regione neutra, e proporzionale all'eccesso di portatori minoritari presente ai bordi della regione di svuotamento. Grazie alla legge della giunzione (3.55), e sulla base dell'ipotesi della trascurabilità della generazione e ricombinazione nella regione di carica spaziale, si ottiene

$$J = J_{n,\mathrm{diff}}(-x_p) + J_{p,\mathrm{diff}}(x_n) = q\left[\frac{D_n}{w_p}\frac{n_\mathrm{i}^2}{N_\mathrm{A}} + \frac{D_p}{w_n}\frac{n_\mathrm{i}^2}{N_\mathrm{D}}\right]\left[\exp\left(\frac{V}{V_T}\right) - 1\right]$$

Pertanto, anche in caso di lati corti vale la caratteristica statica (3.61), pur di definire la corrente inversa di saturazione

$$I_\mathrm{s} = qA\left[\frac{D_n}{w_p}\frac{n_\mathrm{i}^2}{N_\mathrm{A}} + \frac{D_p}{w_n}\frac{n_\mathrm{i}^2}{N_\mathrm{D}}\right]$$

Più in generale, quindi, si può dare una definizione della corrente inversa di saturazione valida se i lati sono o lunghi o corti rispetto alla relativa lunghezza di diffusione dei portatori minoritari

$$I_\mathrm{s} = qA\left[\frac{D_n}{l_n}\frac{n_\mathrm{i}^2}{N_\mathrm{A}} + \frac{D_p}{l_p}\frac{n_\mathrm{i}^2}{N_\mathrm{D}}\right]$$

dove i parametri l_n e l_p sono definiti come segue

$$l_n = \begin{cases} L_n & \text{se } w_p \gg L_n \\ w_p & \text{se } w_p \ll L_n \end{cases} \qquad l_p = \begin{cases} L_p & \text{se } w_n \gg L_p \\ w_n & \text{se } w_n \ll L_p \end{cases}$$

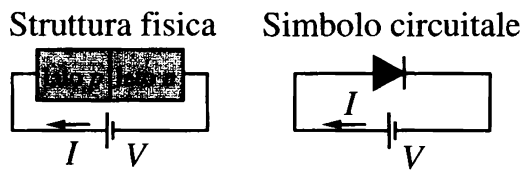


Figura 3.16 Struttura di una giunzione *pn* (a sinistra) e relativo simbolo circuitale (a destra).

Una analisi più accurata della giunzione *pn* [2], in particolare tenendo conto degli effetti di generazione e ricombinazione nella regione di svuotamento, consente di dimostrare come la caratteristica statica possa essere rappresentata dalla seguente relazione

$$I = I_s \left[\exp\left(\frac{V}{\eta V_T} \right) - 1 \right] \tag{3.63}$$

dove $\eta = \eta(V)$ viene detto *fattore di idealità* della giunzione. Il fattore di idealità risulta essere una quantità compresa tra 1 e 2, e in particolare, nel caso di diodi al Si, esso vale

- ▷ circa 2 in polarizzazione inversa e per bassa polarizzazione diretta (V inferiore a circa 0,3 V)
- ▷ circa 1 per una polarizzazione diretta superiore a circa 0,3 V

Esempio 3.1 Si consideri una giunzione brusca e simmetrica con drogaggio $N_A = N_D = 10^{17}$ cm^{-3}, sezione trasversale $A = 0.5$ mm^2 e lati lunghi rispetto alle lunghezze di diffusione dei portatori minoritari. Sapendo che le mobilità dei portatori minoritari nei due lati valgono $\mu_n = 1000$ cm^2 V^{-1} s^{-1} e $\mu_p = 400$ cm^2 V^{-1} s^{-1}, mentre i tempi di vita dei portatori minoritari sono $\tau_n = \tau_p = 1$ μs, si richiede di determinare la caratteristica statica della giunzione.
La richiesta del testo dell'esercizio consiste sostanzialmente nel valutare la corrente inversa di saturazione del dispositivo, utilizzando la (3.62) in quanto i due lati sono dichiarati lunghi nella formulazione del problema. Le diffusività dei portatori minoritari nei due lati sono legate alle mobilità dalla relazione di Einstein

$$D_n = V_T \mu_n = 26 \text{ cm}^2/\text{s} \qquad D_p = V_T \mu_p = 10,4 \text{ cm}^2/\text{s}$$

mentre le lunghezze di diffusione sono date dalle espressioni

$$L_n = \sqrt{D_n \tau_n} = 50,99 \ \mu\text{m} \qquad L_p = \sqrt{D_p \tau_p} = 32,25 \ \mu\text{m}$$

Pertanto, essendo per il Si $n_i = 1,45 \times 10^{10}$ cm^{-3}, si ottiene infine

$$I_s = qA \frac{n_i^2}{N_A} \frac{D_n}{L_n} + qA \frac{n_i^2}{N_D} \frac{D_p}{L_p} = 1,4 \times 10^{-14} \text{ A} = 14 \text{ fA}$$

3.3.4 Modello statico e concetto di punto di funzionamento

La caratteristica statica (3.63) della giunzione *pn* rappresenta una relazione tra i valori della tensione ai capi del dispositivo e della corrente che la attraversa assumendo che il dispositivo stia operando in regime stazionario nel tempo. Si tratta di una relazione *nonlineare* che rende la giunzione, dal punto di vista elettrico, un *bipolo nonlineare*

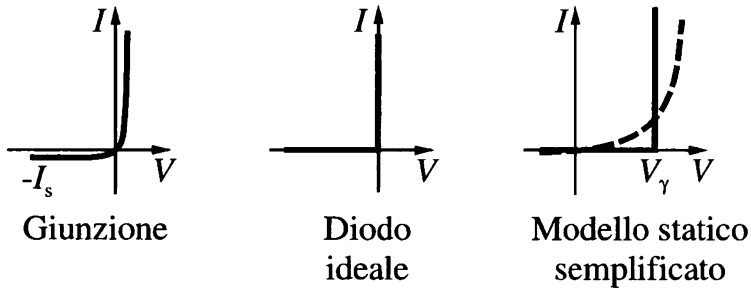


Figura 3.17 Caratteristica statica di una giunzione *pn* (a sinistra), di un diodo ideale (al centro), e caratteristica statica semplificata del diodo (a destra).

controllato in tensione. Il simbolo circuitale che rappresenta la giunzione *pn* è riportato nella figura 3.16.

Confrontando la caratteristica statica della giunzione *pn* con la relazione costitutiva di un *diodo ideale* (si vedano la parte sinistra e centrale della figura 3.17) si osserva che il diodo a giunzione ne costituisce una approssimazione, infatti

- ▷ in polarizzazione diretta, ad una piccola variazione di tensione applicata corrisponde una forte variazione di corrente, approssimando così il comportamento del *corto circuito*
- ▷ in polarizzazione inversa, la corrente che attraversa il diodo è costante e molto piccola, approssimando il comportamento del *circuito aperto*

Naturalmente, è immediato osservare come il comportamento statico del diodo a giunzione in condizioni di polarizzazione inversa sia effettivamente una buona approssimazione della caratteristica di un circuito aperto, ovvero corrente nulla per tensione arbitraria applicata. In polarizzazione diretta, invece, la caratteristica statica del diodo approssima in modo molto poco accurato il comportamento del corto circuito. Per questo motivo, si preferisce adottare una diversa approssimazione per la caratteristica statica (3.63), rappresentata graficamente nella parte destra della figura 3.17, detta *modello statico semplificato*

- ▷ in polarizzazione inversa, il diodo è un circuito aperto: si dice che è *interdetto*
- ▷ in polarizzazione diretta, la giunzione presenta una caduta di tensione costante V_γ: questo stato viene detto di *conduzione*

In termini analitici, ciò corrisponde ad approssimare la caratteristica statica $I(V)$ con una relazione lineare a tratti

$$\begin{cases} I = 0 & \text{se } V < 0 \\ V = V_\gamma & \text{se } I > 0 \end{cases} \tag{3.64}$$

Si noti che l'approssimazione in condizioni di polarizzazione diretta corrisponde a sostituire al diodo a giunzione un *generatore ideale di tensione,* invece del corto circuito che rappresenterebbe il comportamento del diodo ideale. Nel caso di una giunzione su

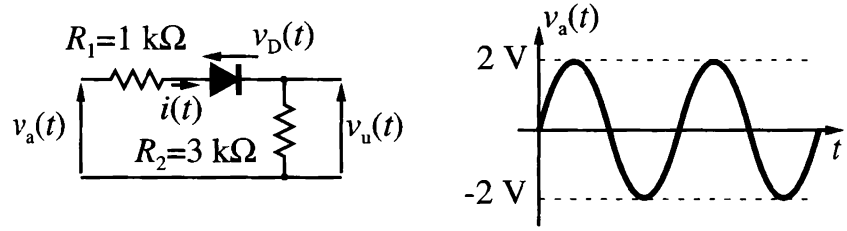


Figura 3.18 Circuito elementare per la definizione di punto di funzionamento (a sinistra), e rappresentazione grafica della retta di carico nel piano della caratteristica statica (a destra).

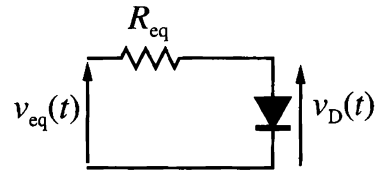


Figura 3.19 Circuito equivalente di Thevenin ai capi del diodo per il circuito di figura3.18.

Si, si può assumere $V_\gamma = 0,5 \div 0,6$ V. Più in generale, si può ritenere [2] che qV_γ sia dell'ordine di grandezza della metà dell'ampiezza della banda proibita del semiconduttore utilizzato per realizzare la giunzione.

Esempio 3.2 Si prenda in considerazione il circuito mostrato nella figura 3.18. Assumendo che $v_a(t)$ sia una forma d'onda sinusoidale di valore massimo pari a 2 V, determinare graficamente $v_u(t)$ utilizzando il modello statico semplificato del diodo. Si assuma inizialmente $V_\gamma = 0$ V, e poi si ripeta l'analisi per $V_\gamma = 0,5$ V.
Il modello statico semplificato (3.64) approssima la caratteristica statica nel diodo con una curva lineare a tratti: all'interno del dominio di validità dei due tratti lineari, quindi, il circuito risulta essere facile da studiare. La difficoltà da superare consiste nel determinare i valori di tensione di ingresso per i quali il dispositivo commuta da uno stato all'altro. Per fare ciò, nei circuiti contenenti un solo diodo come unico elemento nonlineare, conviene sostituire la parte lineare del circuito, vista ai capi del diodo, con il corrispondente circuito equivalente di Thevenin, mostrato nella figura 3.19. Per il circuito di figura 3.18 i parametri del circuito equivalente di Thevenin valgono

$$v_{eq}(t) = v_a(t) \qquad R_{eq} = R_1 + R_2 = 4 \text{ k}\Omega$$

Poiché quando il diodo è interdetto la corrente è nulla, la commutazione si ha negli istanti nei quali:

$$v_D(t) = V_\gamma \Longrightarrow v_{eq}(t) = v_a(t) = V_\gamma$$

In particolare, il diodo è interdetto per

$$v_D(t) \leq V_\gamma \Longrightarrow v_{eq}(t) = v_a(t) \leq V_\gamma$$

Noti gli istanti di commutazione, il circuito della figura 3.18 ammette due circuiti equivalenti (mostrati nella figura 3.20), uno valido quando il diodo è interdetto, l'altro quando il diodo è in conduzione. Studiando separatamente i due circuiti validi per i due stati del diodo, si valuta

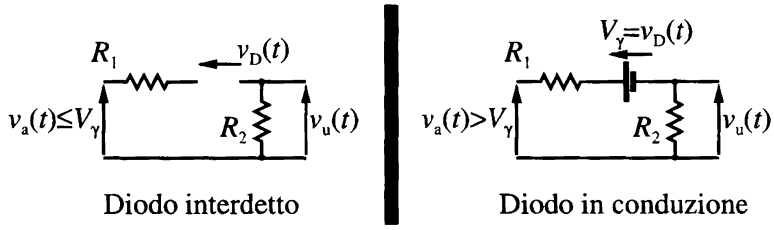


Figura 3.20 Circuito equivalente quando il diodo è interdetto (a sinistra) e in conduzione (a destra).

facilmente la tensione di uscita $v_u(t)$

$$v_u(t) = \begin{cases} 0 & v_a(t) \leq V_\gamma \\ [v_a(t) - V_\gamma]\dfrac{R_2}{R_1 + R_2} = 0,75[v_a(t) - V_\gamma] & v_a(t) > V_\gamma \end{cases}$$

La relazione $v_u(v_a)$, mostrata nella figura 3.21, viene detta *caratteristica di trasferimento* o *ingresso-uscita* del circuito. Nella parte inferiore della stessa figura sono rappresentati gli andamenti temporali delle due tensioni di ingresso ed uscita. Il circuito in esame viene anche detto *raddrizzatore ad una semionda*, in quanto traferisce all'uscita solo parte della semionda positiva della tensione di ingresso.

La caratteristica statica del diodo a giunzione può essere sfruttata per definire un concetto generale valido per tutti i circuiti elettronici che utilizzino dei componenti non lineari, ovvero l'idea di *punto di funzionamento a riposo* (pdf) del dispositivo. Si consideri il circuito elementare di figura 3.22, nel quale un diodo a giunzione è alimentato da un generatore costante reale di tensione di valore V_a, caratterizzato da una resistenza interna R. Indicando con I la corrente che attraversa la giunzione e con V la differenza di potenziale ai capi del diodo, entrambe misurate secondo la convenzione degli utilizzatori definita nel paragrafo 3.3.1, il pdf del dispositivo è definito dai valori I_0 e V_0 assunti da queste due variabili a seguito dell'applicazione della tensione V_a: in altri termini, si tratta dei valori di tensione e corrente nel bipolo definiti dal circuito nel quale il dispositivo è inserito. Il termine "a riposo" si riferisce al fatto che si assume di applicare al circuito dei generatori costanti nel tempo, e quindi giustifica l'utilizzo della caratteristica statica nel descrivere il comportamento elettrico del dispositivo. In generale, in presenza di generatori di segnale dipendenti dal tempo, si parla di *punto di lavoro* (tempo-variante): in questo caso, come si discuterà nel paragrafo 3.4, il comportamento elettrico del dispositivo deve essere descritto da un modello dinamico.

Da queste considerazioni, è evidente come il pdf si debba ricavare utilizzando sia le relazioni che consentono lo studio del circuito nel quale il dispositivo è inserito, sia la caratteristica statica del componente. In particolare, le due incognite I e V saranno definite da due equazioni indipendenti, che sono la legge di Kirchhoff di bilancio delle tensioni per la maglia presente nel circuito di figura 3.22 e la stessa caratteristica statica. La prima è una relazione lineare nel piano definito dalle variabili (V,I)

$$V = V_a - RI \tag{3.65}$$

e viene chiamata *retta di carico* per il diodo. Poiché la caratteristica statica è una

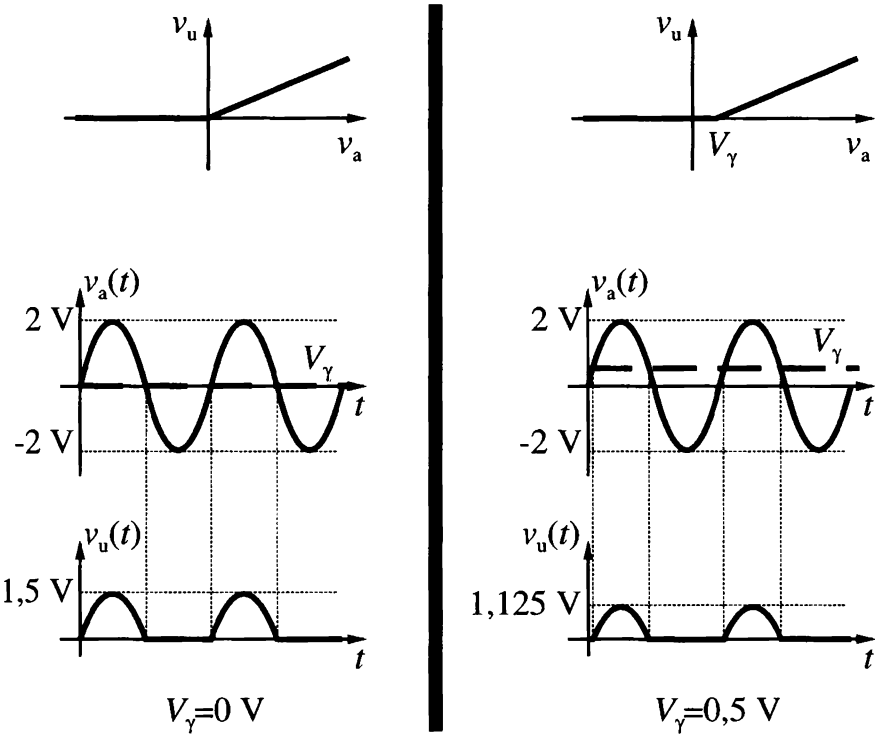


Figura 3.21 Rappresentazione della caratteristica di trasferimento (in alto) e degli andamenti temporali delle tensioni di ingresso ed uscita (in basso) per il circuito di figura 3.18. La colonna di sinistra si riferisce al caso $V_\gamma = 0$ V, quella di destra al caso $V_\gamma = 0,5$ V.

relazione nonlineare nello stesso piano (V,I), il pdf è definito dall'intersezione di quest'ultima con la retta di carico, come mostrato nella parte destra della figura 3.22. La valutazione numerica del pdf, ovvero di I_0 e V_0, richiede di risolvere numericamente il sistema algebrico nonlineare formato da (3.63) e (3.65). Una strategia alternativa, comunque approssimata, consiste nel sostituire alla caratteristica statica completa (3.63) il modello statico semplificato (3.64).

Effetto della temperatura sul pdf

La discussione presentata nel capitolo 1 dimostra come la temperatura influenzi pesantemente le proprietà dei materiali semiconduttori. È quindi ragionevole attendersi che anche le prestazioni elettriche dei dispositivi a semiconduttore siano a loro volta significativamente dipendenti da questo parametro. Nel caso della caratteristica statica (3.63), la temperatura influenza direttamente l'espressione sia attraverso l'equivalente elettrico della temperatura V_T, presente a denominatore della funzione esponenziale, sia attraverso la corrente inversa di saturazione I_s. Infatti, nella (3.62) le quantità dipendenti dalla temperatura sono la concentrazione intrinseca n_i, la mobilità μ e la lunghezza di diffusione dei portatori minoritari. In particolare, la forte dipendenza dal-

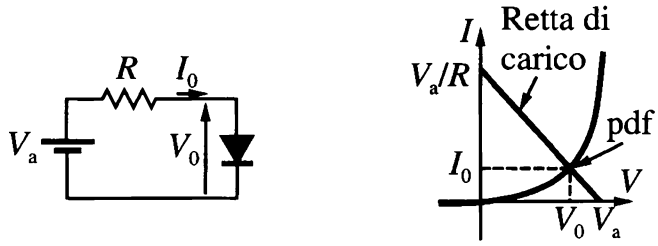


Figura 3.22 Circuito elementare per la definizione di punto di funzionamento (a sinistra), e rappresentazione grafica della retta di carico nel piano della caratteristica statica (a destra).

la temperatura di n_i descritta nella (1.8) costituisce la principale causa di dipendenza dalla temperatura della caratteristica statica del diodo. Trascurando le cause di variazione con la temperatura di I_s diverse dalla concentrazione intrinseca, e linearizzando tale dipendenza attorno ad un valore di temperatura nominale pari alla temperatura ambiente $T_0 = 300$ K, si può stimare la variazione relativa di I_s per unità di variazione di temperatura secondo la relazione

$$\frac{1}{I_s}\frac{\Delta I_s}{\Delta T} \approx \frac{1}{I_s}\frac{\mathrm{d}I_s}{\mathrm{d}T} \approx \frac{1}{I_s}\frac{\mathrm{d}I_s}{\mathrm{d}n_i^2}\frac{\mathrm{d}n_i^2}{\mathrm{d}T} \tag{3.66}$$

Dalla (1.27) si ricava

$$\frac{\mathrm{d}n_i^2}{\mathrm{d}T} = n_i^2\left[\frac{3}{T} - \frac{\mathrm{d}}{\mathrm{d}T}\left(\frac{E_g}{k_B T}\right)\right] \approx n_i^2\left[\frac{3}{T} + \frac{E_g}{k_B T^2}\right] \tag{3.67}$$

dove si è trascurato l'effetto della dipendenza dalla temperatura di E_g. Inoltre, la (3.62) consente di valutare

$$\frac{\mathrm{d}I_s}{\mathrm{d}n_i^2} = \frac{I_s}{n_i^2} \tag{3.68}$$

Sostituendo nella (3.66) si trova, infine

$$\frac{1}{I_s}\frac{\Delta I_s}{\Delta T} \approx \frac{3}{T} + \frac{E_g}{k_B T^2} \tag{3.69}$$

L'espressione (3.69) può essere utilizzata per stimare l'effetto sul pdf delle variazioni di caratteristica statica indotte da una variazione di temperatura. Il ragionamento viene condotto assumendo di mantenere la corrente I_0 nel pdf costante, con l'obbiettivo di stimare la variazione di tensione ΔV subita dalla componente V_0 del pdf richiesta per mantenere I al valore alla temperatura ambiente T_0. Supponendo che il diodo si trovi in polarizzazione diretta, ovvero $V > 0$, l'espressione della caratteristica statica

può essere invertita ottenendo

$$V = V_T \log\left(\frac{I + I_s}{I_s}\right) \approx V_T \log\left(\frac{I}{I_s}\right) \tag{3.70}$$

essendo, in polarizzazione diretta, $I \gg I_s$. La varizione di tensione per unità di variazione di temperatura viene stimata differenziando la (3.70) e utilizzando la (3.69)

$$\frac{\Delta V}{\Delta T} \approx \frac{\mathrm{d}V}{\mathrm{d}T} = \frac{V}{T} - \frac{V_T}{I_s}\frac{\mathrm{d}I_s}{\mathrm{d}T} \approx \frac{V - E_g/q - 3V_T}{T} \tag{3.71}$$

Nel caso di un diodo al Si ($E_g = 1,12$ eV) polarizzato con $V_0 = 0,6$ V a $T_0 = 300$ K, si ottiene $\Delta V/\Delta T = -2$ mV/K. L'evidenza sperimentale, nella quale non sono comprese le diverse approssimazioni condotte nel derivare la (3.71), indica un valore vicino a $-2,5$ mV/K.

3.4 Effetti capacitivi

La caratteristica statica ricavata nel paragrafo 3.3 rappresenta una relazione istantanea tra la tensione ai capi della giunzione e la corrente che la attraversa, e pertanto non tiene conto degli effetti di ritardo determinati dalla presenza di cariche accumulate all'interno del dispositivo, a loro volta dipendenti da tali variabili. In presenza di segnali applicati dipendenti dal tempo, le cariche accumulate determinano un effetto reattivo, che dal punto di vista circuitale corrisponde alla presenza di elementi capacitivi nel circuito equivalente che rappresenta il funzionamento del dispositivo.

Dal punto di vista fisico, le cariche accumulate nella giunzione dipendenti dai segnali elettrici esterni sono di due tipi:

- ▷ quella accumulata nel doppio strato di carica a cavallo della regione di svuotamento, costituita da cariche fisse. In questo caso, la dipendenza della carica dal punto di lavoro è determinata dalla dipendenza dalla tensione applicata dell'ampiezza della regione di svuotamento. Tenendo conto della condizione di neutralità, nei due lati della regione di svuotamento si trovano due cariche opposte: si indica con $Q_f[v(t)]$ la carica totale accumulata nella regione svuotata nel lato p
- ▷ la carica di iniezione o svuotamento, a seconda del segno di $v(t)$, di portatori minoritari nei due lati neutri: si denotano $Q_n[v(t)]$ e $Q_p[v(t)]$ le cariche totali corrispondenti, rispettivamente, agli elettroni in eccesso nel lato p, e alle lacune in eccesso nel lato n. Si tratta, in questo caso, di una carica costituita da portatori mobili. Si può dimostrare [2], a partire dalle equazioni di continuità, che la carica che controlla il corrispondente effetto capacitivo è data da $Q_m[v(t)] = Q_p[v(t)] - Q_n[v(t)]$

Alle due cariche Q_f e Q_m corrispondono due contributi di corrente che attraversano la giunzione, additivi rispetto al contributo istantaneo determinato dalla caratteristica statica

$$\frac{\mathrm{d}Q_f[v(t)]}{\mathrm{d}t} = \frac{\mathrm{d}Q_f}{\mathrm{d}v}\frac{\mathrm{d}v}{\mathrm{d}t} = C_s[v(t)]\frac{\mathrm{d}v}{\mathrm{d}t} \qquad \frac{\mathrm{d}Q_m[v(t)]}{\mathrm{d}t} = \frac{\mathrm{d}Q_m}{\mathrm{d}v}\frac{\mathrm{d}v}{\mathrm{d}t} = C_d[v(t)]\frac{\mathrm{d}v}{\mathrm{d}t} \tag{3.72}$$

dove sono definite la *capacità di svuotamento* C_s e la *capacità di diffusione* C_d

$$C_s[v(t)] = \frac{dQ_f}{dv} \qquad C_d[v(t)] = \frac{dQ_m}{dv} \tag{3.73}$$

Si tratta di due elementi capacitivi nonlineari controllati dalla tensione $v(t)$, che rendono conto in termini circuitali degli effetti di ritardo discussi in precedenza.

3.4.1 Capacità di svuotamento

La carica totale accumulata nella regione svuotata nel lato p della giunzione vale

$$Q_f[v(t)] = -qAN_A x_p[v(t)] \tag{3.74}$$

Assumendo che l'ampiezza della regione svuotata x_p sia una funzione istantanea della tensione $v(t)$ applicata alla giunzione, ovvero utilizzando una *approssimazione quasi-statica* [2] per la carica di svuotamento, si può ritenere che x_p sia ancora dato dalla relazione istantanea (3.38b), pur di sostituire $V_{bi} - v(t)$ al valore V_{bi} della caduta di potenziale di equilibruio ai capi della regione svuotata stessa

$$x_p[v(t)] = \sqrt{\frac{2\epsilon}{qN_A} \frac{N_D}{N_A + N_D} [V_{bi} - v(t)]} \tag{3.75}$$

Sostituendo la (3.74) nella (3.73), si ottiene l'espressione per la capacità di svuotamento

$$C_s[v(t)] = \frac{dQ_f}{dv} = -qAN_A \frac{dx_p}{dv} \tag{3.76}$$

che si esplicita, grazie alla (3.75), nella relazione

$$C_s[v(t)] = A\sqrt{\frac{q\epsilon N_{eq}}{2[V_{bi} - v(t)]}} \tag{3.77}$$

dove il drogaggio equivalente N_{eq} è definito dalla (3.39).

La (3.77) presenta, in funzione del valore di v, un asintoto verticale in $v = V_{bi}$ (si veda la figura 3.23): questo comportamente, pur essendo presente nell'espressione (3.77), non ha significato fisico, poiché per valori di tensione applicata positiva viene a cadere una delle ipotesi fondamentali che consente di determinare $x_p(v)$ mediante la (3.75), ovvero l'assunzione della trascurabilità della caduta di potenziale sulle due regioni neutre (si veda la discussione nel paragrafo 3.3.1). Infatti, per $v > 0$ la corrente che attraversa la giunzione diviene significativa, e quindi non si può più trascurare la caduta di potenziale $R_p i$ sulle due regioni neutre: la caduta di potenziale sulle regioni svuotate, quindi, si riduce a

$$V_{bi} - [v(t) - R_p i(t)] = V_{bi} - v(t) + R_p i(t) \tag{3.78}$$

e questa non si annulla mai perché $i(t)$ è una funzione esponenzialmente crescente di $v(t)$. In altri termini, l'espressione (3.77) vale solo in polarizzazione inversa e in debole

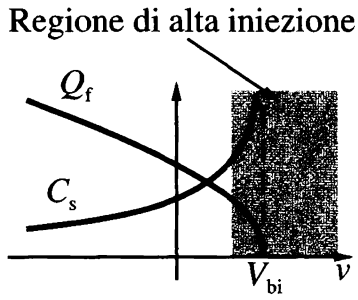


Figura 3.23 Dipendenza dalla tensione applicata alla giunzione della carica e della capacità di svuotamento.

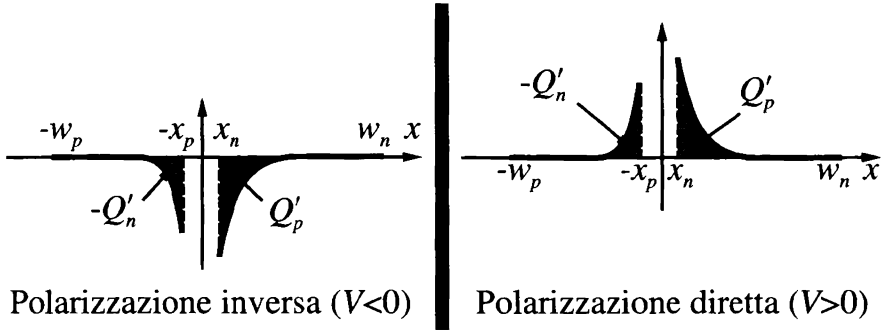


Figura 3.24 Carica mobile in eccesso nei due lati neutri in polarizzazione inversa (a sinistra) e diretta (a destra).

polarizzazione diretta, ovvero al di fuori della regione ombreggiata nella figura 3.23, che viene denominata *regione di alta iniezione* per distinguerla dalla condizione di funzionamento in basso livello di iniezione.

3.4.2 Capacità di diffusione

Per quanto riguarda la carica mobile in eccesso iniettata nei due lati neutri, essa può essere valutata integrando le concentrazioni in eccesso di elettroni e lacune nei due lati, come descritto nella figura 3.24

$$Q_n = -qA \int_{-w_p}^{-x_p} n'_p(x)\, dx \qquad Q_p = qA \int_{x_n}^{w_n} p'_n(x)\, dx \tag{3.79}$$

Nel caso di lati lunghi, si può approssimare il calcolo dei due integrali supponendo che la lunghezza fisica del lato sia infinita, ottenendo

$$Q_n \approx -qA \int_{-\infty}^{-x_p} n'_p(x)\, dx \qquad Q_p \approx qA \int_{x_n}^{+\infty} p'_n(x)\, dx \tag{3.80}$$

Anche in questo caso, nell'ambito di una approssimazione quasi-statica per la carica, si può assumere che le concentrazioni in eccesso rispondano istantaneamente alle variazioni di tensione applicata, in modo da poter continuare ad utilizzare le relazioni di basso livello di iniezione ricavate nel paragrafo 3.3.1. In particolare, per dei lati lunghi si possono utilizzare le(3.45) ottenendo

$$\begin{aligned} Q_n &= -qA\int_{-\infty}^{-x_p} n'_p(-x_p)\exp\left(\frac{x+x_p}{L_n}\right)\,\mathrm{d}x \\ &= -qAL_n n'_p(-x_p)\exp\left(\frac{x+x_p}{L_n}\right)\Bigg|_{-\infty}^{-x_p} = -qAL_n n'_p(-x_p) \end{aligned} \tag{3.81a}$$

$$\begin{aligned} Q_p &= qA\int_{x_n}^{+\infty} p'_n(x_n)\exp\left(-\frac{x-x_n}{L_p}\right)\,\mathrm{d}x \\ &= -qAL_p p'_n(x_n)\exp\left(-\frac{x-x_n}{L_p}\right)\Bigg|_{x_n}^{+\infty} = qAL_p p'_n(x_n) \end{aligned} \tag{3.81b}$$

Infine, sostituendo la legge della giunzione (3.55) si trova

$$Q_n = -qA\frac{n_\mathrm{i}^2}{N_\mathrm{A}}L_n\left[\exp\left(\frac{v(t)}{V_T}\right)-1\right] \qquad Q_p = qA\frac{n_\mathrm{i}^2}{N_\mathrm{D}}L_p\left[\exp\left(\frac{v(t)}{V_T}\right)-1\right] \tag{3.82}$$

da cui segue

$$Q_\mathrm{m} = Q_p - Q_n = qA\left[\frac{n_\mathrm{i}^2}{N_\mathrm{A}}L_n + \frac{n_\mathrm{i}^2}{N_\mathrm{D}}L_p\right]\left[\exp\left(\frac{v(t)}{V_T}\right)-1\right] \tag{3.83}$$

Facendo uso della (3.73), si può determinare l'espressione della capacità di diffusione

$$C_\mathrm{d}[v(t)] = \frac{\mathrm{d}Q_\mathrm{m}}{\mathrm{d}v} = qA\frac{n_\mathrm{i}^2}{V_T}\left[\frac{L_n}{N_\mathrm{A}}+\frac{L_p}{N_\mathrm{D}}\right]\exp\left(\frac{v(t)}{V_T}\right) \tag{3.84}$$

Vista la dipendenza esponenziale dalla tensione $v(t)$ sulla giunzione, la capacità di diffusione ha un valore trascurabile in polarizzazione inversa, mentre in polarizzazione diretta diviene tipicamente preponderante rispetto al contributo di svuotamento.

Approfondimento 3.5 Nel caso di una giunzione con lati quasi-neutri corti, le espressioni per le concentrazioni in eccesso sono quelle riportate nell'approfondimento 3.2. Si ha in questo caso

$$\begin{aligned} Q_n &= -qA\int_{-w_p}^{-x_p} n'_p(-x_p)\frac{w_p+x}{w_p-x_p}\,\mathrm{d}x = -qAn'_p(-x_p)\,\frac{(w_p+x)^2}{2(w_p-x_p)}\Bigg|_{-w_p}^{-x_p} \\ &= -qAn'_p(-x_p)\frac{w_p-x_p}{2} \approx -qAn'_p(-x_p)\frac{w_p}{2} \\ Q_p &= qA\int_{x_n}^{w_n} p'_n(x_n)\frac{w_n-x}{w_n-x_n}\,\mathrm{d}x = -qAp'_n(x_n)\,\frac{(w_n-x)^2}{2(w_n-x_n)}\Bigg|_{x_n}^{w_n} \end{aligned}$$

$$= qAp'_n(x_n)\frac{w_n - x_n}{2} \approx qAp'_n(x_n)\frac{w_n}{2}$$

Facendo uso della legge della giunzione (3.55) si trova

$$Q_n = -qA\frac{n_i^2}{N_A}\frac{w_p}{2}\left[\exp\left(\frac{v(t)}{V_T}\right) - 1\right] \qquad Q_p = qA\frac{n_i^2}{N_D}\frac{w_n}{2}\left[\exp\left(\frac{v(t)}{V_T}\right) - 1\right]$$

pertanto la carica mobile può essere stimata in

$$Q_m = Q_p - Q_n = qA\left[\frac{n_i^2}{N_A}\frac{w_p}{2} + \frac{n_i^2}{N_D}\frac{w_n}{2}\right]\left[\exp\left(\frac{v(t)}{V_T}\right) - 1\right]$$

Derivando questa espressione rispetto a v, si trova infine l'espressione della capacità di diffusione per un diodo con i lati corti

$$C_d[v(t)] = \frac{dQ_m}{dv} = qA\frac{n_i^2}{V_T}\left[\frac{w_p}{2N_A} + \frac{w_n}{2N_D}\right]\exp\left(\frac{v(t)}{V_T}\right)$$

Riunendo questo risultato con la (3.84), si ottiene l'espressione generale per la C_d, valida indipendentemente dalla lunghezza dei due lati neutri

$$C_d[v(t)] = qA\frac{n_i^2}{V_T}\left[\frac{l'_n}{N_A} + \frac{l'_p}{N_D}\right]\exp\left(\frac{v(t)}{V_T}\right)$$

dove, in analogia con l'approfondimento 3.4, si è posto

$$l'_n = \begin{cases} L_n & \text{se } w_p \gg L_n \\ w_p/2 & \text{se } w_p \ll L_n \end{cases} \qquad l'_p = \begin{cases} L_p & \text{se } w_n \gg L_p \\ w_n/2 & \text{se } w_n \ll L_p \end{cases}$$

3.5 Modello circuitale della giunzione *pn*

Dalla descrizione del comportamento elettrico della giunzione *pn* in condizioni stazionarie e dinamiche presentata nei paragrafi 3.3 e 3.4, si può concludere come la corrente $i(t)$ che attraversa la giunzione sia esprimibile come la somma di tre contributi. Il primo di questi consiste della risposta istantanea del dispositivo, rappresentata dalla caratteristica statica

$$i_{dc}[v(t)] = I_s\left\{\exp\left[\frac{v(t)}{\eta V_T}\right] - 1\right\} \tag{3.85}$$

Le altre due componenti di corrente corrispondono ai termini di ritardo espressi tramite le capacità di svuotamento e diffusione

$$C_s[v(t)]\frac{dv}{dt} \qquad C_d[v(t)]\frac{dv}{dt} \tag{3.86}$$

Complessivamente, si ha l'espressione

$$i(t) = i_{dc}[v(t)] + C_s[v(t)]\frac{dv}{dt} + C_d[v(t)]\frac{dv}{dt} \tag{3.87}$$

La (3.87) ammette l'interpretazione circuitale mostrata in figura 3.25, infatti la decomposizione della corrente totale nella somma di tre contributi corrisponde nel circuito

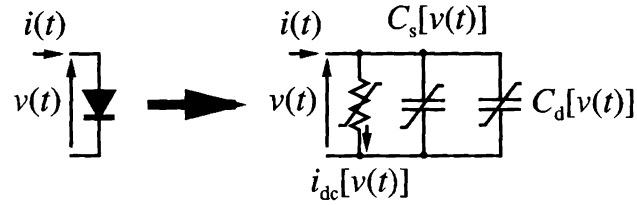


Figura 3.25 Rappresentazione circuitale della relazione (3.87).

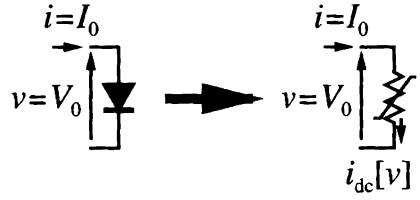


Figura 3.26 Circuito equivalente statico come caso particolare del modello di ampio segnale.

alla presenza di tre rami in parallelo. Poiché tutti e tre gli addendi sono controllati dalla stessa tensione $v(t)$, che è anche la tensione applicata alla giunzione, i tre rami sono costituiti da tre bipoli nonlineari: la componente istantanea corrisponde ad un bipolo resistivo, mentre i due contributi di ritardo sono rappresentati dalle due capacità nonlineari. Il circuito di figura 3.25 viene detto *circuito equivalente di ampio segnale* del diodo, in quanto vale per generici segnali applicati di tipo dinamico. Si noti che le espressioni che costituiscono le relazioni costitutive dei bipoli nonlineari presentano delle limitazioni, corrispondenti alle ipotesi effettuate per la loro derivazione: in particolare, l'assunzione fondamentale è quella di basso livello di iniezione, alla quale corrisponde anche la trascurabilità della caduta di tensione sulle regioni quasi-neutre rispetto alla tensione applicata totale. Ciò significa che il circuito equivalente costituisce un modello che approssima le caratteristiche elettriche del dispositivo, con delle ben precise limitazioni di validità: naturalmente, è in generale possibile pensare di sviluppare modelli più accurati della giunzione, per i quali le limitazioni precedentemente discusse possano essere ridotte. Tipicamente, e per qualunque dispositivo, ad una maggiore generalità del modello corrisponde un incremento della sua complessità, e quindi una minore efficienza nella sua implementazione in strumenti di progettazione circuitale assistita (ad esempio, SPICE). La scelta del modello, quindi, deve essere condotta mediando tra le necessità di accuratezza della rappresentazione del comportamento del dispositivo, e le necessità di efficienza numerica, ad esempio determinate dalla necessità di analizzare e/o progettare un circuito complesso costituito da un gran numero di dispositivi.

Il modello di ampio segnale, essendo valido in condizioni dinamiche, contiene come caso particolare il modello statico discusso nel paragrafo 3.3. Infatti, in condizioni stazionarie nel tempo il contributo di corrente corrispondente agli elementi capacitivi

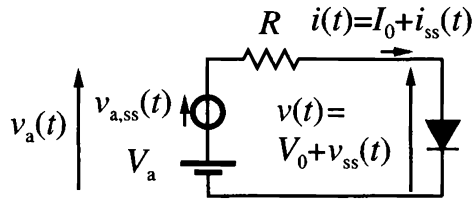


Figura 3.27 Circuito per la definizione del concetto di modello di piccolo segnale.

si annulla (essendo $dv/dt = 0$), e la (3.87) si riduce a

$$i(t) = I_0 = i_{dc}(V_0) = I_s \left\{ \exp\left[\frac{v(t)}{\eta V_T}\right] - 1 \right\} \tag{3.88}$$

ovvero alla caratteristica statica (3.63), come descritto nella figura 3.26. Naturalmente, in caso di necessità il modello statico completo può essere a sua volta semplificato con il modello semplificato introdotto nel paragrafo 3.3.4.

3.5.1 Modello circuitale di piccolo segnale

Il modello circuitale di ampio segnale, pur essendo di validità generale seppure nell'ambito di ben precise approssimazioni, fornisce una rappresentazione del comportamento elettrico del dispositivo che difficilmente consente di studiare il comportamento dei circuiti che lo contengano senza sfruttare metodi di analisi numerica, quali ad esempio il simulatore circuitale SPICE e i suoi derivati di tipo commerciale. D'altra parte, il regime di funzionamento dinamico è naturalmente la condizione di funzionamento più comune nelle applicazioni. È possibile definire un particolare regime di funzionamento dinamico, detto di *piccolo segnale,* che a sua volta consente di introdurre una metodologia di analisi circuitale approssimata, ma sufficientemente semplice da poter essere condotta anche senza l'ausilio di strumenti di tipo informatico.

L'analisi di piccolo segnale, pur essendo qui introdotta con riferimento ai diodi a giunzione, è una tecnica di analisi approssimata che può essere applicata a qualunque circuito contenente dispositivi nonlineari, quali ad esempio i transistori. Si prenda in cobnsiderazione il semplice circuito mostrato nella figura 3.27, nel quale la tensione $v_a(t)$ applicata come segnale esterno risulta essere decomposta nella sovrapposizione di una parte V_a costante nel tempo, e di un segnale tempo-variante $v_{a,ss}(t)$. Questa decomposizione non toglie nulla alla generalità dell'analisi, infatti qualunque segnale $x(t)$ può essere decomposto nella somma di un valore costante X_0 e di un termine variabile nel tempo $x_{ss}(t) = x(t) - X_0$. Nel caso dell'analisi circuitale, però, la decomposizione segue naturalmente dalla condizione tipica di funzionamento dei circuiti, almeno per quanto riguarda le applicazioni analogiche: la parte costante del generatore corrisponde all'alimentazione in continua (DC) del sistema, ovvero alla tensione[4] applicata al circuito nel momento nel quale questo viene posto in condizione di funzionare, in assenza cioè del segnale utile (tipicamente tempo-variante) che il circuito stesso deve processare. In altri termini, la parte costante V_a del segnale applicato corrisponde ai

[4] In termini generali, si può pensare di alimentare in DC un circuito anche con generatori di corrente, sebbene in pratica l'alimentazione sia tipicamente fornita con generatori di tensione.

generatori che definiscono il pdf dei vari dispositivi (si veda la discussione nel paragrafo 3.3.4): tali generatori vengono anche detti di *polarizzazione.* Da questa discussione, segue naturalmente la scelta di decomporre anche le tensioni e le correnti per tutti i dispositivi nel circuito secondo la stessa logica, ovvero nella somma di una parte costante, corrispondente al pdf di ogni dispositivo determinato dall'applicazione dei soli generatori DC, e di una parte tempo-variante, corrispondente alla perturbazione del pdf determinata dal segnale applicato $v_{a,ss}(t)$. Per il circuito di figura 3.27, nel caso del diodo ciò corrisponde ad assumere

$$v(t) = V_0 + v_{ss}(t) \qquad i(t) = I_0 + i_{ss}(t) \tag{3.89}$$

dove V_0 e I_0 sono le coordinate del pdf del diodo a seguito dell'applicazione della sola tensione di polarizzazione V_a. Analogamente, per il resistore si avrà la decomposizione

$$v_R(t) = V_{R0} + v_{R,ss}(t) \qquad i(t) = I_0 + i_{ss}(t) \tag{3.90}$$

essendo v_R la tensione ai capi della resistenza e i la corrente che la attraversa. Naturalmente, la linearità del resistore implica

$$V_{R0} = RI_0 \qquad v_R(t) = Ri(t) \tag{3.91}$$

e quindi, nel caso di questo circuito, la determinazione di tensioni e correnti nel diodo consente di valutare anche il punto di lavoro del resistore.

Per calcolare il punto di lavoro tempo-variante del diodo, ovvero i due segnali $v(t)$ e $i(t)$ si potrebbe pensare di sostituire al diodo un modello di ampio segnale, ad esempio il circuito equivalente descritto da (3.87), e di risolvere, eventualmente in modo numerico con l'ausilio di un simulatore circuitale, il sistema costituito da tale relazione e dall'equazione di bilancio delle tensioni alla maglia

$$v_a(t) = v_R(t) + v(t) = Ri(t) + v(t) \tag{3.92}$$

Si può, però, anche procedere in modo approssimato sfruttando la decomposizione dei segnali definita in precedenza, ed effettuando l'*ipotesi di piccolo segnale* che consiste nell'assumere che $v_{a,ss}(t)$ sia sufficientemente piccola da determinare un segnale $v_{ss}(t)$ sul diodo (ovvero, sull'elemento nonlineare) tale da soddisfare la condizione

$$|v_{ss}(t)| \ll |V_0| \tag{3.93}$$

In altri termini, ciò corrisponde ad assumere che la perturbazione della componente V_0 del pdf determinata dal segnale applicato sia trascurabile rispetto a V_0: il motivo per il quale si effettuta questa ipotesi sulla sola componente di tensione del punto di lavoro è legato al fatto che il diodo presenta delle nonlinearità *controllate in tensione.* Se le nonlinearità fossero controllate in corrente, occorrerebbe riformulare l'ipotesi di piccolo segnale in termini di tale variabile.

L'ipotesi di piccolo segnale (3.93) consente di semplificare significativamente le componenti nonlineari del modello dinamico di ampio segnale (3.87), infatti sulla base della (3.93) è possibile utilizzare uno sviluppo in serie, centrato nel pdf e arrestato al primo

ordine, delle tre quantità

$$i_{dc}[v(t)] = i_{dc}[V_0 + v_{ss}(t)] \approx i_{dc}(V_0) + \left.\frac{\partial i_{dc}}{\partial v}\right|_{v=V_0} v_{ss}(t) \tag{3.94a}$$

$$Q_f[v(t)] = Q_f[V_0 + v_{ss}(t)] \approx Q_f(V_0) + \left.\frac{\partial Q_f}{\partial v}\right|_{v=V_0} v_{ss}(t) \tag{3.94b}$$

$$Q_m[v(t)] = Q_m[V_0 + v_{ss}(t)] \approx Q_m(V_0) + \left.\frac{\partial Q_m}{\partial v}\right|_{v=V_0} v_{ss}(t) \tag{3.94c}$$

ovvero, linearizzare la dipendenza da v degli elementi nonlineari presenti nel modello di ampio segnale. Ricordando che, per definizione, $i_{dc}(V_0)$ coincide con la componente I_0 del pdf, nelle (3.94) sono definiti i tre *parametri differenziali* del diodo nel punto di lavoro:

▷ la *conduttanza differenziale* g_{d0} (misurata in S)

$$g_{d0} = \left.\frac{\partial i_{dc}}{\partial v}\right|_{v=V_0} = \frac{I_s}{\eta V_T} \exp\left(\frac{V_0}{\eta V_T}\right) = \frac{I_0 + I_s}{\eta V_T} \tag{3.95a}$$

ottenuta derivando la (3.85)

▷ la *capacità (differenziale) di svuotamento* C_{s0} (misurata in F)

$$C_{s0} = \left.\frac{\partial Q_f}{\partial v}\right|_{v=V_0} = A\sqrt{\frac{q\epsilon N_{eq}}{2[V_{bi} - V_0]}} \tag{3.95b}$$

ottenuta valutando in V_0 la (3.77)

▷ la *capacità (differenziale) di diffusione* C_{d0} (misurata in F)

$$C_{d0} = \left.\frac{\partial Q_m}{\partial v}\right|_{v=V_0} = qA\frac{n_i^2}{V_T}\left[\frac{L_n}{N_A} + \frac{L_p}{N_D}\right] \exp\left(\frac{V_0}{V_T}\right) \tag{3.95c}$$

ottenuta valutando in V_0 la (3.84)

Questi parametri sono costanti, essendo funzione della sola componente DC di $v(t)$. Sostituendo le tre approssimazioni nella (3.87) si trova

$$i(t) \approx i_{dc}(V_0) + g_{d0}v_{ss}(t) + C_{s0}\frac{dv_{ss}}{dt} + C_{d0}\frac{dv_{ss}}{dt} \tag{3.96}$$

ovvero, essendo $i_{ss}(t) = i(t) - I_0 = i(t) - i_{dc}(V_0)$, l'approssimazione di piccolo segnale linearizza il modello di ampio segnale trasformandolo in una relazione *lineare* tra le variazioni di corrente e di tensione rispetto al pdf

$$i_{ss}(t) = g_{d0}v_{ss}(t) + C_{s0}\frac{dv_{ss}}{dt} + C_{d0}\frac{dv_{ss}}{dt} \tag{3.97}$$

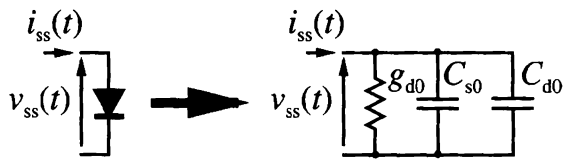


Figura 3.28 Circuito equivalente di piccolo segnale, o per le variazioni, del diodo.

Da un punto di vista circuitale, la (3.97) corrisponde al *circuito equivalente di piccolo segnale* del diodo mostrato nella figura 3.28. Si noti che il modello di piccolo segnale pone in relazione non la tensione e la corrente totali nel diodo, ma esclusivamente le *variazioni* di tensione e corrente rispetto al pdf: per questo motivo viene anche chiamato *circuito equivalente per la variazioni.*

Esempio 3.3 Si consideri una giunzione brusca e simmetrica con drogaggio $N_A = N_D = 10^{15}$ cm^{-3}, sezione trasversale $A = 1.5$ mm^2 e lati lunghi rispetto alle lunghezze di diffusione dei portatori minoritari. Le mobilità dei portatori minoritari nei due lati valgono $\mu_n = 1050$ cm^2 V^{-1} s^{-1} e cm^2 V^{-1} s^{-1}, mentre i tempi di vita sono $\tau_n = \tau_p = 1$ μs. Si richiede di determinare i parametri di piccolo segnale del diodo per la polarizzazione $V_0 = 0.6$ V.
Per determinare i parametri di piccolo segnale si applicano le relazioni (3.95), e quindi occorre conoscere il punto di funzionamento a riposo della giunzione. Pertanto, è necessario valutare la caratteristica statica del dispositivo. Il calcolo di I_s richiede di conoscere la diffusività dei portatori minoritari

$$D_n = V_T \mu_n = 27{,}3 \text{ cm}^2/\text{s} \qquad D_p = V_T \mu_p = 11{,}7 \text{ cm}^2/\text{s}$$

e le relative lunghezze di diffusione

$$L_n = \sqrt{D_n \tau_n} = 52{,}25 \ \mu\text{m} \qquad L_p = \sqrt{D_p \tau_p} = 34{,}21 \ \mu\text{m}$$

Inoltre, dal testo dell'esercizio si possono assumere lati lunghi per la giunzione, e quindi per del Si a temperatura ambiente

$$I_s = qA \frac{n_i^2}{N_A} \frac{D_n}{L_n} + qA \frac{n_i^2}{N_D} \frac{D_p}{L_p} = 4{,}36 \text{ pA}$$

Infine, assumendo $\eta = 1$, si può valutare la corrente nel pdf

$$I_0 = I_s \left[\exp\left(\frac{V_0}{\eta V_T} \right) - 1 \right] = 45{,}88 \text{ mA}$$

Per quanto riguarda i parametri di piccolo segnale, essendo il pdf in polarizzazione diretta si può ritenere trascurabile la capacità di svuotamento rispetto a quella di diffusione, in modo che il circuito equivalente di piccolo segnale si semplifichi nel parallelo della conduttanza differenziale e della capacità di diffusione. Sostituendo nelle (3.95) si trova

$$g_{d0} = \frac{I_0 + I_s}{\eta V_T} = 1{,}76 \text{ S} \qquad C_{d0} = qA \frac{n_i^2}{V_T} \left[\frac{L_n}{N_A} + \frac{L_p}{N_D} \right] \exp\left(\frac{V_0}{V_T} \right) = 1{,}77 \ \mu\text{F}$$

Esempio 3.4 Si consideri una giunzione brusca e asimmetrica, con drogaggi $N_A = 10^{15}$ cm^{-3} e $N_D = 10^{17}$ cm^{-3}, sezione trasversale $A = 0.2$ mm^2 e lati lunghi. Nei due lati, i portatori minoritari sono caratterizzati da una mobilità $\mu_n = 1050$ cm^2 V^{-1} s^{-1} e $\mu_p = 400$ cm^2 V^{-1} s^{-1}, e dai tempi di vita $\tau_n = \tau_p = 1$ μs. Si richiede di determinare, se esiste, l'intervallo di tensioni di

polarizzazione necessario a realizzare una capacità accordabile con valori compresi tra 5 pF e 10 pF.
La realizzazione di una capacità accordabile richiede di scegliere un punto di lavoro per il dido nel quale il suo comportamento elettrico approssimi quello di un condensatore. Il funzionamento in regime di piccolo segnale soddisfa questa condizione, come mostrato nel circuito equivalente riportato nella figura 3.28. Per avere un buon comportamento capacitivo, però, occorre che la conduttanza differenziale g_{d0} abbia un valore tale da influenzare il meno possibile il comportamento in piccolo segnale del diodo, ovvero sia il più possibile vicino a zero. Essendo g_{d0} proporzionale a $I+I_s$, questo risultato si ottiene solo scegliendo per il diodo un pdf in polarizzazione inversa. Si noti che questa scelta ha come conseguenza che anche la capacità di diffusione risulta essere trascurabile. In definitiva, sebbene C_{d0} possa assumere valori elevati come mostrato nell'esempio 3.3, la necessità di rendere trascurabile il comportamento resisitivo del diodo per realizzare un condensatore di buona qualità rende possibile sfruttare a questo scopo la sola capacità di svuotamento. Assumendo, quindi, che il diodo sia in polarizzazione inversa, si può ricavare dalla (3.95b) il valore di tensione di polarizzazione V_0 necessaria a garantire un valore C_{s0} alla capacità di svuotamento

$$V_0 = V_{bi} - A^2 \frac{qN_{eq}\epsilon_s}{2C_{s0}^2}$$

dove, trattandosi di una giunzione in Si a $T = 300$ K

$$V_{bi} = V_T \ln \frac{N_A N_D}{n_i^2} = 0,7 \text{ V}$$

Sostituendo i valori di capacità richiesti nel testo si trova

$$V_0(C_{s0} = 10 \text{ pF}) = -2,58 \text{ V} \qquad V_0(C_{s0} = 5 \text{ pF}) = -12,43 \text{ V}$$

Poiché entrambi gli estremi dell'intervallo di valori di V_0 sono negativi, e quindi in polarizzazione inversa, si può concludere che la giunzione considerata consente di realizzare una capacità accordabile nell'intervallo di valori richiesto.
In commercio sono disponibili diodi a giunzione appositamente progettati per essere utilizzati come capacità accordabili, detti *varactor*; per massimizzare la variazione di capacità in funzione del pdf si sfruttano normalmente giunzioni tra materiali non uniformemente drogati.

Si noti che il circuito equivalente per le variazioni del diodo descritto dalla (3.97) è una approssimazione, ottenuta per linearizzazione attorno al pdf, della modello di ampio segnale del diodo. Pertanto, per inserire correttamente tale approssimazione nel circuito complessivo, è necessario procedere con lo stesso tipo di approssimazione per tutti i dispositivi presenti nel circuito stesso. Si possono quindi considerare tre tipi di elementi da linearizzare:

- elementi circuitali *lineari*, quali ad esempio resistori, condensatori e induttori lineari, e generatori pilotati dipendenti linearmente dal pilota. Per questi dispositivi la relazione costitutiva è, per definizione, già lineare, quindi il loro circuito equivalente per le variazioni è identico al circuito equivalente di ampio segnale. Considerando, per fissare le idee, in resistore lineare di resistenza R inserito nel circuito di figura 3.27, la sua relazione costitutiva è $v_R(t) = Ri(t)$, pertanto la corrispondente relazione per le variazioni si esprime nella forma

$$v_{R,ss}(t) = Ri_{ss}(t) \tag{3.98}$$

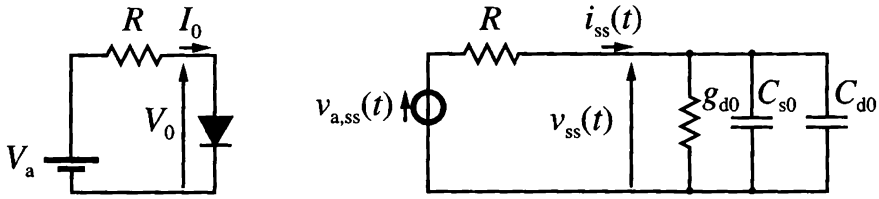


Figura 3.29 Circuito equivalente per l'analisi in DC (a sinistra) e per l'analisi di piccolo segnale (a destra) per il circuito di figura 3.27.

ovvero, in termini circuitali, il modello di piccolo segnale del resistore di resistenza R è ancora un resistore dello stesso valore. Lo stesso ragionamento può essere ripetuto per tutte le relazioni costitutive di elementi circuitali lineari

- ▷ *generatori indipendenti* di tensione o corrente costanti. Per un generatore ideale di tensione costante, la relazione costitutiva stabilisce che la tensione ai suoi capi è pari al valore nominale indipendentemente dalla corrente che lo attraversa, quindi la variazione di tensione di piccolo segnale ai capi del generatore è, per definizione, nulla. Di conseguenza, il circuito equivalente per le variazioni di un generatore di tensione costante è un corto circuito (variazioni di tensione nulla indipendentemente dalla variazione di corrente che lo attraversa). Ripetendo il ragionamento per un generatore ideale di corrente costante, si verifica facilmente che in questo caso è la variazione di corrente ad essere nulla indipendentemente dalla variazione di tensione ai capi del generatore. Pertanto, il circuito equivalente di piccolo segnale per un generatore ideale di corrente costante è un circuito aperto. In altri termini, i generatori in DC hanno un circuito equivalente per le variazioni corrispondente al loro spegnimento
- ▷ elementi circuitali *nonlineari,* quali tipicamente i dispositivi a semiconduttore. In questo caso, occorre sostituire al dispositivo il corrispondente circuito equivalente di piccolo segnale, ricavato per linearizzazione di un modello di ampio segnale. Naturalmente, la topologia e i valori dei componenti di tale circuito (comunque dipendenti dal pdf) dipendono dal tipo di dispositivo.

In definitiva, l'analisi approssimata di piccolo segnale consente di stimare le prestazioni di un circuito in condizioni dinamiche sulla base di due passi successivi:

1. si *valutano i pdf* di tutti i dispositivi presenti nel circuito, tipicamente i componenti nonlineari. Per effettuare questa valutazione, occorre alimentare il circuito con i soli generatori in continua (DC), spegnendo quindi tutti i generatori di segnale. Questo passo viene anche detto di *studio della polarizzazione* o *in DC*: nel caso del circuito di figura 3.27, il circuito per lo studio della polarizzazione è mostrato nella parte sinistra della figura 3.29
2. si *costruisce il circuito per le variazioni* ottenuto sostituendo a tutti i dispositivi il corrispondente modello di piccolo segnale (il circuito equivalente per gli elementi nonlineari, lo stesso dispositivo per gli elementi lineari) e spegnendo i generatori in continua. Si ottiene così un circuito lineare che pone in relazione i generatori di segnale alle variazioni di tensione e corrente nei vari dispositivi, mostrato nel caso

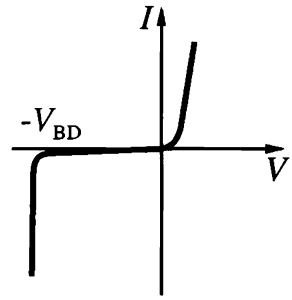


Figura 3.30 Caratteristica statica di una giunzione *pn* in presenza del breakdown per elevata polarizzazione inversa.

del semplice circuito con diodo e resistenza nella parte destra della figura 3.29.

3.6 Fenomeni di breakdown

In condizioni di forte polarizzazione inversa ($V \ll 0$), la caratteristica statica di una giunzione *pn* non mantiene l'andamento ideale previsto dalla (3.63) per $V < 0$, ovvero una corrente sostanzialmente indipendente da V e pari a $-I_s$. Quando la tensione applicata V raggiunge un valore critico $V = -V_{BD}$, detto *tensione di breakdown* della giunzione, la corrente I subisce un repentino aumento di valore come mostrato nella figura 3.30: si parla di *rottura* o *breakdown* della giunzione. In condizioni di breakdown, il valore assoluto della corrente cresce in modo molto rapido per piccoli aumenti del valore assoluto della tensione V, approssimando così il comportamento elettrico di un generatore ideale di tensione.

La causa dell'incremento di corrente inversa può risiedere in tre diversi fenomeni fisici. Di questi, il primo ha luogo nel caso in cui la tensione di polarizzazione inversa sia sufficientemente elevata in valore assoluto perché la regione svuotata si estenda fino a coprire l'intero volume del dispositivo: si parla, in questo caso, di *perforazione diretta* o *punch-through* del dispositivo. Infatti, estendendo la (3.38c) al caso fuori equilibrio assumendo che tutta la tensione applicata cada sulla regione di svuotamento, si ha

$$x_d = \sqrt{\frac{2\epsilon}{q}\frac{1}{N_{eq}}(V_{bi} - V)} \tag{3.99}$$

e quindi x_d è una funzione decresente di V, dimostrando come esista un valore di V per il quale $x_d = w_n + w_p$. Se questa condizione ha luogo, dalla figura 3.12 è immediato verificare come il diagramma a bande divenga una funzione monotona decrescente della posizione tra il contatto ohmico per il lato p e quello per il lato n. Pertanto, per gli elettroni disponibili nel materiale che costituisce il contatto ohmico sul lato p è molto facile poter raggiungere l'altro contatto ohmico, dando quindi luogo ad una elevata corrente. Questo fenomeno, naturalmente, è favorito quando la lunghezza fisica della giunzione è particolarmente corta e i livelli di drogaggio sono ridotti,[5] condizioni che

[5] Infatti, in questo caso $N_{eq} = N_A \parallel N_D$ si riduce.

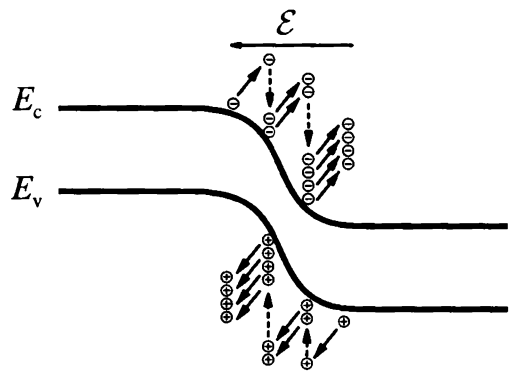


Figura 3.31 Moltiplicazione a valanga dei portatori liberi nella regione di svuotamento di una giunzione *pn*.

difficilmente si presentano per un diodo a giunzione *pn*. La perforazione diretta, però, si può presentare nella base di un transistore bipolare che, come si vedrà in seguito, deve essere la più corta possibile.

Gli altri due meccanismi di rottura, invece, dominano il breakdown dei diodi a giunzione. Essi sono:

- ▷ la generazione di portatori liberi nella regione di svuotamento per *moltiplicazione a valanga*
- ▷ il passaggio diretto di elettroni dalla banda di valenza alla banda di conduzione attraverso la regione svuotata per *effetto tunnel*. Questo fenomeno viene anche detto di breakdown per *effetto Zener*.

3.6.1 Breakdown per moltiplicazione a valanga

Il meccanismo della moltiplicazione a valanga è basato sul fenomeno fisico mostrato nella figura 3.31. Si consideri una giunzione *pn* in polarizzazione inversa, ed una coppia elettrone lacune che venga generata vicino all'inizio della regione neutra nel lato *p*: la lacuna viene accelerata dal campo elettrico verso il lato *p*, mentre l'elettrone viene accelerato nella direzione opposta: di conseguenza, entrambi incrementano in valore assoluto la corrente (inversa) nella giunzione. Se il campo elettrico è sufficientemente intenso, nel tratto di volo libero prima dell'urto successivo l'elettrone una energia sufficiente perché nella fase di urto questa possa essere utilizzata per generare un'altra coppia elettrone-lacuna. A questo punto, il fenomeno si ripete innescando un processo di crescita esponenziale delle coppie elettrone-lacuna generate, giustificando così il nome di *effetto valanga*. Naturalmente, il ragionamento può essere ripetuto a partire da una lacuna generata nelle vicinanze dell'inizio della regione neutra del lato *n*.

A seconda del materiale considerato, si definiscono un *campo elettrico critico* $\mathcal{E}_c$, corrispondente al valore minimo di campo elettrico necessario ad innescare il fenomeno della moltiplicazione a valanga, ed i *coefficienti di ionizzazione* per elettroni (α_n) e lacune (α_p), ovvero il numeri di elettroni o lacune generati per unità di lunghezza. Intuitivamente, è ragionevole attendersi che il fenomeno del breakdown a valanga sia favorito da lati poco drogati. Infatti, ad un drogaggio minore corrisponde un minor

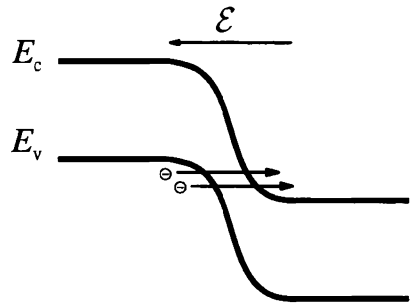


Figura 3.32 Passaggio di elettroni dalla banda di valenza a quella di conduzione per effetto tunnel.

numero di occasioni di urto dei portatori liberi, e quindi ad un maggior cammino percorso dalla carica tra due urti successivi; di conseguenza, in questo spazio, a parità di campo elettrico, i portatori potranno assumere una energia cinetica maggiore. Infine, si può dimostrare [2] che i coefficienti di ionizzazione, oltre ad essere una funzione rapidamente crescente del campo elettrico, diminuiscono con la temperatura.

In presenza di questo fenomeno, la tensione di breakdown viene definita (a parte il segno) come la tensione applicata alla giunzione che innesca il fenomeno della moltiplicazione a valanga. Come conseguenza della dipendenza dalla temperatura dei coefficienti di ionizzazione, si ricava immediatamente che V_{BD} cresce con la temperatura.

3.6.2 Breakdown per effetto Zener

L'effetto Zener si basa su un fenomeno fisico tipicamente quantistico, il passaggio attraverso una barriera di energia potenziale per effetto tunnel. Secondo le leggi della meccanica classica, una particella di energia E è in grado di passare oltre una barriera di energia di altezza E_0 se e solo se $E > E_0$: nel caso $E < E_0$, la particella può solo essere riflessa dalla barriera stessa. Se la particella ha dimensioni atomiche o subatomiche, invece, le leggi della meccanica classica cessano di valere, e la dinamica della particella viene descritta dalla meccanica quantistica. Senza entrare nei dettagli, si può affermare come in questo caso [1, 2] la particella con energia E abbia sempre, indipendentemente dal valore (purché finito) di E_0, una probabilità finita e non nulla R di essere riflessa dalla barriera di energia potenziale, e corrispondentemente una probabilità T di essere trasmessa. In particolare, ciò è vero nel caso $E < E_0$: nei casi nei quali la particella ha effettivamente attraversato la barriera, si dice che tale attraversamento è avvenuto per *effetto tunnel*.

Nel caso di una giunzione *pn* in polarizzazione inversa, la caduta di potenziale sulla regione di svuotamente, pari a $V_{bi} - V$, diviene sempre più elevata al crescere in valore assoluto di $V < 0$. Per $-V$ sufficientemente elevata, quindi, il valore di E_c nella regione neutra del lato n scende sotto il valore di E_v nella regione neutra del lato p, come mostrato nella figura 3.32. In questo modo, gli elettroni presenti nella banda di valenza nel lato p si trovano, sostanzialemente a parità di energia, un gran numero di posti disponibili ad essere occupati nella banda di valenza della regione neutra nel lato n: risulta quindi favorito un attraversamento della regione svuotata per effetto tunnel,

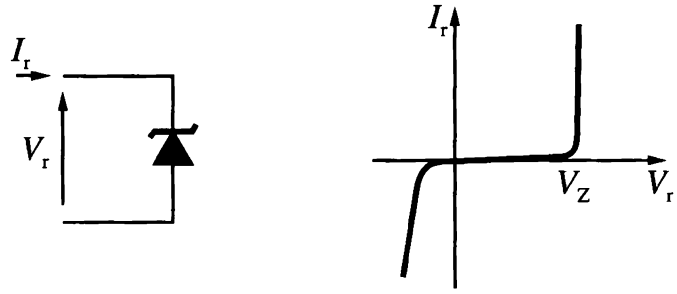


Figura 3.33 Simbolo circuitale (a sinistra) e caratteristica statica (a destra) del diodo Zener.

dando luogo ad un flusso di elettroni che si spostano da p verso n che costituisce un incremento del valore assoluto della corrente che attraversa la giunzione. Al crescere della polarizzazione inversa, il numero di posti occupati nella banda di valenza affacciati a stati disponibili nella banda di conduzione del lato opposto aumenta, giustificando quindi l'incremento del valore assoluto della corrente.

Chiaramente, il breakdown per effetto Zener è favorito da una zona di svuotamento sottile, e quindi da livelli di drogaggio elevati. Si può anche dimostrare che la tensione di breakdown per effetto tunnel decresce con la temperatura.

3.6.3 Diodo Zener

La sostanziale indipendenza dalla tensione della corrente che attraversa la giunzione *pn* in condizioni di breakdown suggerisce l'utilizzo di tale regime di funzionamento per approssimare il comportamento di un generatore ideale di tensione. Si trovano in commercio dei dispositivi a giunzione appositamente progettati per lavorare in condizioni di rottura, allo scopo di essere utilizzati come riferimento di tensione: si parla, in questo caso, di *diodi zener,* caratterizzati dal simbolo circuitale mostrato nella parte sinistra della figura 3.33. Poiché questi dispositivi lavorano sempre in polarizzazione inversa, di solito si preferisce cambiare i versi per i riferimenti di tensione e corrente rispetto a quanto usato finora per il diodo a giunzione, utilizzando le trasformazioni

$$I_r = -I \qquad V_r = -V \tag{3.100}$$

che conducono alla convenzione di segno mostrata nella figura 3.33. Con riferimento a queste variabili, la caratteristica statica del diodo zener assume la forma riportata nella parte destra della stessa figura. Il valore della tensione di breakdown costituisce la tensione nominale del diodo Zener, e per questo viene indicata con il simbolo V_Z.

Nei diodi Zener la giunzione viene di solito appositamente progettata perché il breakdown avvenga per una combinazione dei due effetti precedentemente descritti in quanto la dipendenza opposta dalla temperatura che li caratterizza può essere utilmente sfruttata allo scopo di rendere il più possibile V_Z indipendente anche da T, oltre che dalla corrente.

Una tipica applicazione del diodo Zener è nei circuiti stabilizzatori di tensione, per i quali un esempio di schema di principio è riportato nella parte destra della figura 3.34: se la tensione applicata V_a è sufficientemente elevata da fornire al diodo Zener una

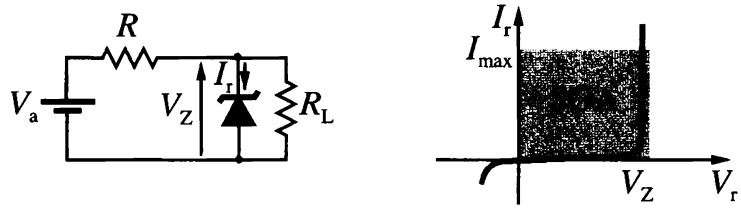


Figura 3.34 Circuito di principio per l'utilizzo del diodo Zener come stabilizzatore di tensione (a sinistra) e definizione della zona di funzionamento sicuro sulla caratteristica statica (a destra).

corrente I_r sufficiente a portarlo a lavorare nella zona di breakdown, la tensione ai capi del diodo, e quindi del carico R_L, risulta essere pari a V_Z (la tensione nominale del diodo Zener) indipendentemente dal valore di R_L. La funzione della resistenza R, che include al suo interno anche la resistenza interna del generatore di tensione reale, è quella di limitare la corrente nel diodo al di sotto di un valore limite I_{max}, fornito dal costruttore del dispositivo, al di sopra del quale il diodo smette di funzionare correttamente o, al limite, si distrugge a causa della eccessiva dissipazione di potenza. Infatti, la corrente nel diodo Zener è data da:

$$I_r \approx \frac{V_a - V_Z}{R} - \frac{V_Z}{R_L} \tag{3.101}$$

per cui la condizione $I_r \leq I_{max}$ implica

$$R \geq \frac{V_a - V_Z}{I_{max} + V_Z/R_L} \tag{3.102}$$

Queste considerazioni possono essere generalizzate a tutte le tipologie di dispositivo utilizzate nei circuiti elettronici: in generale, un qualunque componente (bipolo, dispositivo a due porte eccetera) presenta un insieme di limitazioni alle tensioni e alle correnti che gli possono essere applicate. Tali limitazioni possono essere dovute a cause diverse, come ad esempio la tensione massima prima del breakdown e la massima potenza dissipabile: il loro effetto, comunque, è di limitare ad una regione, detta di *funzionamento sicuro* o *safe operating area (SOA)*, le variabili che costituiscono la caratteristica statica. Nel caso del diodo Zener, assumendo una limitazione dovuta solo alla massima corrente sopportabile, la SOA assume la forma mostrata nella parte destra della figura 3.34.

Naturalmente, la caratteristica statica in regione di breakdown del diodo Zener non è perfettamente verticale, ovvero la tensione ai capi del diodo non è completamente indipendente dalla corrente che lo attraversa: questa è la motivazione sottesa al segno di approssimazione utilizzato nella (3.101), che è invece scritta nell'ipotesi $V_r = V_Z$. Una analisi più accurata del comportamento della curva $I_r(V_r)$ in questa regione di funzionamento, rivela come essa possa essere ben approssimata da una retta

$$V_r \approx V_Z + R_Z I_r \tag{3.103}$$

Per un diodo Zener ideale $R_Z = 0$, quindi per un dispositivo di qualità tale parametro deve essere il più piccolo possibile. La (3.103) ammette l'interpretazione circuitale

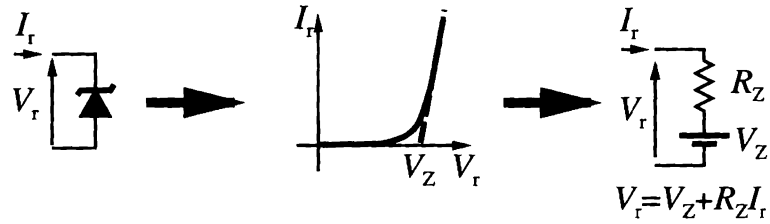


Figura 3.35 Approssimazione nella regione di breakdown della caratteristica statica di un diodo Zener con un circuito equivalente lineare.

mostrata nella parte destra della figura 3.35, che costituisce il circuito equivalente del diodo Zener nella regione di funzionamento normale: i parametri che lo caratterizzano sono V_Z e R_Z.

Capitolo 4

Transistore bipolare

I transistori sono dispositivi a semiconduttore generalmente a tre poli, in cui la corrente che fluisce tra due poli è controllata dalla corrente o tensione del terzo polo. Si osservi che un tripolo può anche essere sempre interpretato come un due-porte, mettendo in comune un terminale alla porta di ingresso e di uscita. In un transistore, quindi, si può equivalentemente affermare che la corrente alla porta di uscita dipende (è controllata) dalla corrente o tensione alla porta di ingresso.

I transistori sono utilizzati sia nelle applicazioni di tipo analogico sia in quelle digitali. Nel primo caso la forma d'onda del segnale di uscita viene controllata (ed eventualmente amplificata) mediante il segnale applicato alla porta di ingresso. Nelle applicazionio digitali, mediante la tensione o la corrente alla porta di ingresso, il dispositivo viene portato a commutare tra i due stati *off*, con corrente di uscita nulla, e *on*, con corrente diversa da zero, identificando i due stati logici necessari per il funzionamento dei circuiti digitali.

I transistori a semiconduttore sono classificati in due grandi famiglie

- ▷ i transistori bipolari a giunzione, detti anche transistori BJT dall'acronimo della denominazione inglese Bipolar Junction Transistor
- ▷ i transistori a effetto di campo, detti anche transistori FET dall'acronimo della denominazione inglese Field Effect Transistor

Nei BJT (che sono trattati in questo capitolo), la corrente alla porta di uscita è di tipo ambipolare (ovvero composta da corrente di elettroni e lacune) e il controllo (porta di ingresso) è prevalentemente in corrente. Nei FET, invece, la corrente alla porta di uscita è di tipo unipolare e il controllo è prevalentemente in tensione; più propriamente il controllo avviene mediante il campo elettrico indotto dalla tensione applicata alla porta di ingresso stessa, da cui il nome di transistori a effetto di campo.

I dispositivi FET si dividono a loro volta in

- ▷ JFET (Junction FET)
- ▷ MESFET (Metal Semiconductor FET)
- ▷ MOSFET (Metal Oxide Semiconductor FET)

Benché i primi due tipi di FET trovino utilizzo in particolari applicazioni di nicchia,

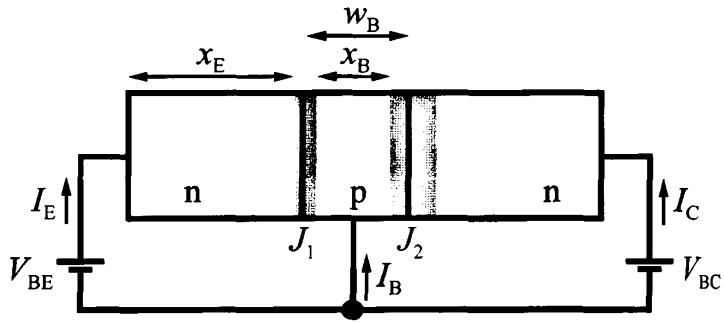


Figura 4.1 Struttura di principio di un transistore bipolare *npn*.

quali l'elettronica delle microonde o l'optoelettronica, il MOSFET è il dispositivo di gran lunga più utilizzato nei circuiti integrati su silicio. In questo capitolo viene presentato solo il funzionamento del MOSFET, rimandando a testi più avanzati la trattazione completa della famiglia dei FET [2].

4.1 Il transistore bipolare

Il transistore bipolare fu introdotto nel 1947 e ha subito una rapida evoluzione tecnologica, prima come componente discreto e successivamente anche come componente integrato. La minore complessità realizzativa del transistore bipolare rispetto al MOSFET ne ha determinato l'ampia diffusione, facendone per molti anni il principale dispositivo attivo usato per realizzare circuiti elettronici analogici e digitali. Le applicazioni del transistore bipolare negli anni hanno coperto la quasi totalità dei settori dell'elettronica, dalle porte logiche ai transistori di segnale, dai dispositivi di potenza alle memorie a semiconduttore, dai fotorilevatori ai circuit di conversione. Soltanto negli anni '80, quando l'affidabilità dei processi produttivi MOS ha raggiunto un livello sufficiente, l'uso del transistore bipolare nella realizzazione di circuiti digitali è andato progressivamente riducendosi a favore di soluzioni MOS complementari, in grado di offrire vantaggi superiori soprattutto in termini di consumo energetico e densità di integrazione. Negli anni successivi, anche in molti settori dell'elettronica analogica il transistore bipolare ha ceduto spazio ai MOSFET, che oggi trovano uso nella maggioranza delle applicazioni analogiche e nella totalità di quelle digitali.

Le motivazioni per mantenere lo studio del transistore bipolare in un corso universitario, sono legate non soltanto all'uso che si continua a fare di questo dispositivo in alcuni ambiti, come per esempio quello dell'elettronica di potenza e dei dispositivi a radio-frequenza (RF), ma anche alla sua importanza teorica e storica.

Il transistore bipolare è fondamentalmente composto da due giunzioni *pn*, realizzate sul medesimo substrato a formare una struttura *npn* oppure *pnp*; il caso *npn* è illustrato in figura 4.1, dove si possono distinguere le due giunzioni *pn*, J_1 e J_2, che dividono il dispositivo in tre regioni:

1. la base (B) al centro; di tipo p e larghezza W_B
2. l'emettitore (E), di tipo n, da un lato

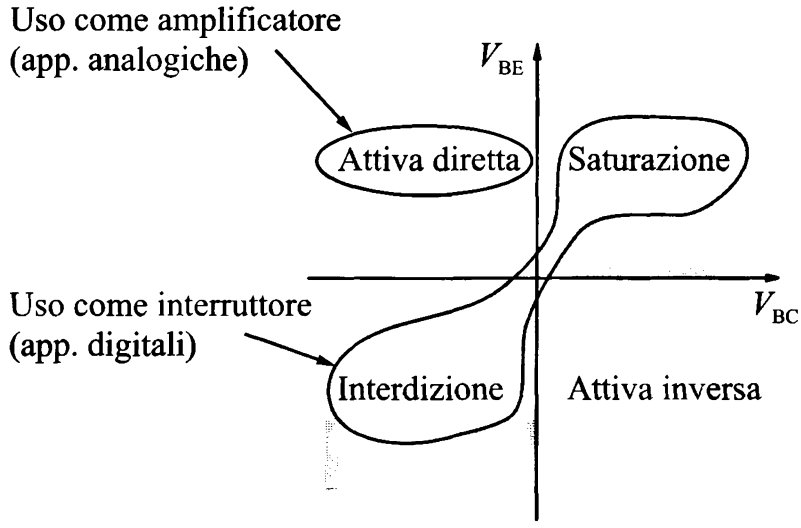


Figura 4.2 Regioni di funzionamento del transistore bipolare.

3. il collettore (C), anch'esso di tipo n, dal lato opposto

Esternamente al dispositivo si hanno tre terminali, che permettono di applicare tre tensioni e tre correnti; in virtù delle leggi di Kirchhoff, solo due delle tensioni e due delle correnti sono indipendenti.

Il comportamento del dispositivo cambia al cambiare delle tensioni applicate alle due giunzioni, indicate nella figura come V_{BE}, tra base e emettitore, e V_{BC}, tra base e collettore. Analogamente alla terminologia usata per il diodo, l'operazione di connettere il transistore a batterie esterne in modo da imporre ben precisi valori di V_{BE} e V_{BC} prende il nome di polarizzazione.

Poiché ogni giunzione può essere polarizzata direttamente o inversamente, variando le polarità applicate alle due giunzioni, si può polarizzare il transistore in quattro diverse regioni di funzionamento, come indicato in figura 4.2:

- la *regione attiva diretta*, ottenuta con la giunzione B-E polarizzata direttamente e la giunzione B-C polarizzata inversamente, è sfruttata soprattutto per ottenere un guadagno, quindi nella realizzazione di amplificatori di segnale o in altri circuiti analogici
- la *regione attiva inversa*, ottenuta con la giunzione B-C polarizzata direttamente e la B-E polarizzata inversamente, è caratterizzata da un funzionamento qualitativamente identico a quello della regione attiva diretta; tuttavia, la struttura reale di un transistore bipolare non è simmetrica e sambiando il ruolo di emettitore e collettore si ottengono guadagni piuttosto scadenti, che rendono questa regione di scarso interesse applicativo
- la *regione di saturazione* richiede tensioni dirette applicate a entrambe le giunzioni (in queste condizioni il dispositivo approssima il comportamento di un interruttore chiuso)

▷ la *regione di interdizione* richiede tensioni inverse applicate a entrambe le giunzioni e in queste condizioni il dispositivo approssima il comportamento di un interruttore aperto, quindi le regioni di saturazione e interdizione sono utili per realizzare porte logiche

In regione attiva diretta, la giunzione B-E è polarizzata direttamente e questo implica il passaggio di una corrente significativa dalla regione di base verso l'emettitore. Tale corrente, indicata come $-I_E$ in figura 4.1, è misurabile sul morsetto di emettitore ed è composta da due contributi, legati al flusso attraverso J_1 delle lacune e degli elettroni. In particolare il fenomeno di diffusione degli elettroni dall'emettitore di tipo n verso la base di tipo p prende il nome di iniezione di elettroni in base e tale fenomeno è modulato dalla tensione V_{BE} in modo esponenziale, come avviene in ogni giunzione polarizzata direttamente. In un diodo a giunzione, gli elettroni iniettati dal lato n subirebbero un processo di ricombinazione all'interno del lato p, come descritto nel capitolo 3, secondo una legge del tipo

$$n_B(x) = n_B(0) \exp\left(-\frac{x}{L_{nB}}\right)$$

dove $n_B(0)$ è la concentrazione di elettroni alla giunzione, x è la distanza di un punto della base misurata dalla giunzione J_1 e L_{nB} è la lunghezza di diffusione degli elettroni nella regione p di base.
Nel transistore bipolare, la lunghezza della base, W_B, ha dimensioni ridottissime, nettamente inferiori a L_{nB} e per questa ragione la ricombinazione degli elettroni iniettati è minima. Ne consegue che quasi tutti gli elettroni attraversano la regione di base, arrivando alla giunzione B-C, e vengono accelerati verso il collettore, che si trova al potenziale più basso.

La giunzione base collettore è polarizzata inversamente e, in assenza dell'emettitore o con una base particolarmente ampia, sarebbe percorsa soltanto dalla corrente inversa di saturazione, originata dalle scarse lacune presenti nel collettore di tipo n e dai pochi elettroni disponibili all'interno della base di tipo p. Tuttavia, nel caso del transistore, la base risulta per così dire allagata dagli elettroni che sono iniettati dal lato di emettitore e che non hanno spazio abbastanza per ricombinarsi; questi elettroni quindi sono facilmente raccolti dal collettore (corrente I_C).

È importante notare che, in regione attiva diretta, la corrente di collettore è poco influenzata dalla tensione di collettore, perché la giunzione B-C è polarizzata inversamente. La I_C è invece fortemente legata alla tensione (o corrente) della giunzione di emettitore. In particolare, troveremo più avanti che I_C è proporzionale a I_B:

$$I_C = \beta_F I_B$$

Questo comportamento può essere descritto mediante un modello elettrico equivalente molto semplice, costituito da un generatore di corrente pilotato in corrente, ovvero un componente che eroga al nodo C una corrente I_C proporzionale alla corrente presente su un nodo diverso (B).

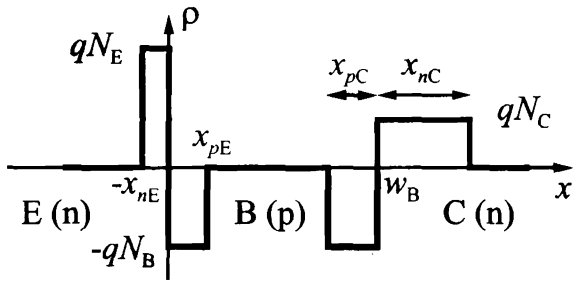


Figura 4.3 Densità di carica.

4.2 Diagramma a bande di energia

Come per il diodo a giunzione, anche nel caso del transistore bipolare il diagramma a bande di energia costituisce uno strumento grafico molto utile per comprendere alcune caratteristiche del funzionamento del dispositivo, in primo luogo l'andamento del potenziale elettrico nella struttura del transistore e le diverse componenti di corrente. In questo paragrafo quindi deriveremo il diagramma a bande in condizioni di equilibrio termodinamico, seguendo la metodologia illustrata nel paragrafo 3.1.

Innanzitutto, si risolve l'equazione di Poisson (o equazione di Gauss differenziale), nelle regioni di interesse, integrando la densità di carica $\rho(x)$.

$$\frac{\partial^2 \varphi}{\partial x^2} = -\frac{\rho(x)}{\epsilon_S}$$

dove $\varphi(x)$ è il potenziale elettrostatico nella direzione x, legato al campo elettrico dalla

$$\mathcal{E} = -\frac{\partial \varphi}{\partial x}$$

Poiché la struttura del transistore è formata da due giunzioni pn, possiamo assumere per ciascuna giunzione il medesimo andamento della densità di carica elettrica $\rho(x)$ già usato nel paragrafo 3.2 per una singola giunzione, nell'ipotesi di transizione brusca tra regione di carica spaziale e regioni neutre; la densità di carica è riportato in figura 4.3.

Nella regioni neutre all'interno della base, dell'emettitore e del collettore, la carica elettrica è nulla, $\rho = 0$, mentre l'approssimazione di completo svuotamento ci permette di assumere per la distribuzione di carica all'interno di ciascuna zona svuotata valori costanti e pari alle concentrazioni di impurità droganti. Si ottiene quindi un andamento di $\rho(x)$ costante a tratti e quindi particolarmente semplice da integrare. Le ampiezze delle quattro regioni svuotate, indicate in figura 4.3 come x_{nE} e x_{pE} per la giunzione di emettitore e x_{pC} e x_{nC} per quella di collettore, non sono note in questa fase e costituiscono incognite da determinare. Queste quattro grandezza non sono però indipendenti, in quanto la condizione di neutralità elettrica a ciascuna delle giunzioni impone

$$N_E x_{nE} = N_B x_{pE} \qquad N_C x_{nC} = N_B x_{pC} \tag{4.1}$$

Dal legame tra potenziale elettrico $\varphi(x)$ e campo $\mathcal{E}(x)$, si esplicita nell'equazione

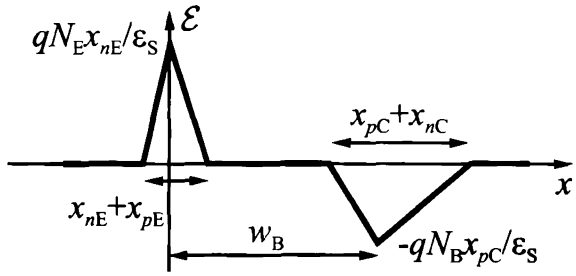


Figura 4.4 Campo elettrico.

di Gauss differenziale il campo elettrico

$$\frac{\partial \varphi}{\partial x} = -\mathcal{E} \qquad \longrightarrow \qquad \frac{\partial \mathcal{E}}{\partial x} = \frac{\rho(x)}{\epsilon_S}$$

La (4.1) implica che per ciascuna giunzione la carica di svuotamento dal lato n sia uguale e opposta a quella dal lato p. Se quindi immaginiamo di applicare la legge di Gauss a una porzione della struttura del transistore che includa completamente le due zone svuotate di una giunzione, il campo elettrico misurato sul contorno di tale porzione deve essere nullo, in quanto la carica inclusa è complessivamente nulla.
In definitiva, si può affermare che $\mathcal{E} = 0$ nelle regioni neutre. In corrispondenza delle regioni svuotate, invece, ρ è costante e quindi il campo avrà andamento lineare, con pendenza pari al valore della densità di carica, diviso per ϵ_S.
Complessivamente l'andamento del campo elettrico attraverso l'intera struttura è dato nella figura 4.4, dove si può notare come il valore massimo del campo si raggiunga sempre in corrispondenza della sezione di giunzione tra lato n e lato p.
Il campo è diverso da zero in due zone di forma triangolare, che hanno base di lunghezza pari all'ampiezza complessiva delle regioni svuotate delle giunzioni di emettitore e collettore, mentre le altezze sono pari alle aree delle zone p, o n, svuotate, riscalate del fattore ϵ_S.

Per la giunzione di emettitore, il valore massimo del campo elettrico si ottiene in $x = 0$ come

$$\mathcal{E}(0) = \frac{qN_E x_{nE}}{\epsilon_S} = \frac{qN_B x_{pE}}{\epsilon_S} \tag{4.2}$$

Alla giunzione di collettore, in $x = W_B$ il campo vale

$$\mathcal{E}(W_B) = -\frac{qN_B x_{pC}}{\epsilon_S} = -\frac{qN_C x_{nC}}{\epsilon_S} \tag{4.3}$$

Dal campo elettrico, per integrazione, possiamo trovare il potenziale:

$$\frac{\partial \varphi}{\partial x} = -\mathcal{E} \tag{4.4}$$

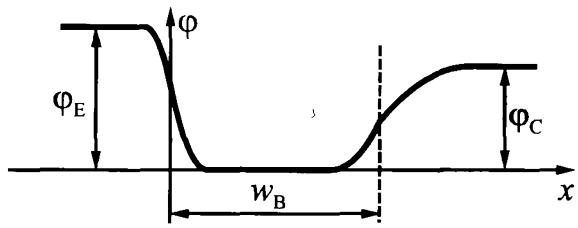


Figura 4.5 Potenziale elettrico.

Le condizioni al contorno della (4.4) devono imporre il valore del potenziale in un punto della struttura e questo può essere scelto arbitrariamente, perché il potenziale elettrico è definito a meno di una costante. Per ragioni di comodità, conviene fissare il valore di φ in una delle regioni neutre, per esempio nella base. Scegliamo quindi

$$\varphi(0) = 0 \tag{4.5}$$

e usiamo questa equazione come condizione al contorno nell'integrazione della(4.4).
I tratti diversi da zero della funzione integranda $-\mathcal{E}(x)$ hanno andamento lineare e l'integrazione fornisce quindi un comportamento di tipo quadratico per $\varphi(x)$ in ciascun tratto lineare di $\mathcal{E}$.

Come indicato nella figura 4.5, in corrispondenza della regione di svuotamento alla giunzione base emettitore occorre integrare un tratto di campo elettrico positivo e quindi, in base alla (4.4), il potenziale dovrà scendere di una quantità complessiva pari all'area sottesa dal campo nella zona triangolare di interesse:

$$\varphi_E = \frac{qN_E x_{nE}}{2\epsilon_s}(x_{nE} + x_{pE})$$

In modo analogo, il potenziale deve salire in corrispondenza della giunzione base collettore di una quantità

$$\varphi_C = \frac{qN_B x_{pC}}{2\epsilon_s}(x_{nC} + x_{pC})$$

Il diagramma a bande per il transistore può ora essere ricavato dall'andamento del potenziale, tenendo conto che convenzionalmente su tale diagramma si riporta l'energia per gli elettroni,

$$E = -q\varphi$$

Quindi applicando al potenziale elettrico il fattore di scala $-q$, si ottiene un andamento che a meno di una costante additiva, vale sia per il limite inferiore della banda di conduzione, E_C, sia per quello superiore della banda di valenza, E_V (figura 4.6).

Si noti che, scegliendo di esprimere le energie in eV e i potenziali in V, le differenze di energie sul diagramma a bande coincidono numericamente con le differenza di potenziale. In particolare, $q\varphi_E$, altezza della barriera di energia per gli elettroni che si trovano nell'emettitore, avrà lo stesso valore di φ_E, potenziale di contatto alla giunzione base emettitore. Inoltre $q\varphi_C$, altezza della barriera di energia per gli elettroni che si

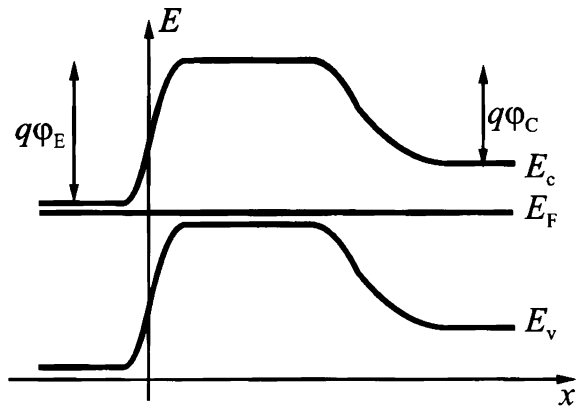


Figura 4.6 Diagramma a bande in condizioni di equilibrio termodinamico.

trovano nel collettore, avrà lo stesso valore di φ_C, potenziale di contatto alla giunzione base collettore.

Come per il diodo a giunzione, il diagramma a bande così ricavato per il caso di equilibrio termodinamico può facilmente essere adattato a polarizzazioni diverse del dispositivo; nell'ipotesi che le correnti generate alle due giunzioni non alterino sensibilmente la condizione di quasi neutralità che si ha nelle regioni di base, emettitore e collettore al di fuori delle zone svuotate, le tensioni applicate dall'esterno andranno semplicemente ad alterare le differenze φ_E e φ_C, aumentandole in caso di polarizzazione inversa e diminuendole per polarizzazione diretta.

4.3 Correnti nel transistore

La corrente di emettitore si misura al terminale di emettitore e coincide con la corrente che attraversa la giunzione base emettitore. Tale corrente è somma di due contributi, dovuti ai flussi delle due tipologie di portatori, elettroni e lacune:

$$I_E = -I_{En} - I_{Ep}$$

La convenzione adottata per il segno delle correnti è indicata nella figura 4.7: per la corrente complessiva di emettitore si è assunta la convenzione degli utilizzatori, mentre per le componenti I_{En} e I_{Ep} si è scelto come verso positivo quello di reale scorrimento nelle condizioni di polarizzazione in regiove attiva diretta.

In polarizzazione diretta,

- I_{En} è dovuta agli elettroni iniettati dall'emettitore nella base
- I_{Ep} è dovuta alle lacune iniettate dalla base nell'emettitore

Solitamente, in un buon transistore bipolare, il drogaggio di emettitore è nettamente superiore a quello di base, $N_E \gg N_B$, e quindi anche le due componenti della corrente

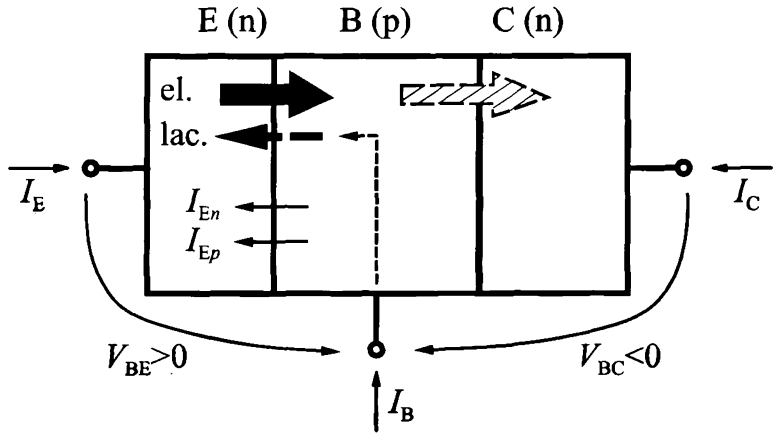


Figura 4.7 Correnti coinvolte nella definizione di efficienza di emettitore.

di emettitore risultano piuttosto sbilanciate

$$I_{En} \gg I_{Ep}$$

Questo significa che la I_E è dominata dal flusso di elettroni provenienti dall'emettitore; la relazione tra le due correnti, quella complessiva e quella legata ai soli elettroni, può allora essere espressa introducendo un coefficiente tecnologico γ minore di 1:

$$I_{En} = -\gamma I_E$$

γ prende il nome di *efficienza di emettitore* e numericamente è molto prossimo a 1, nei casi reali.

Nella figura 4.7 si noti come il percorso delle lacune si chiuda tra i terminali di emettitore e base, senza coinvolgere il collettore. Al contrario, gli elettroni provenienti dall'emettitore, raggiunta la base, hanno la possibilità di proseguire verso il collettore, dando così origine all'effetto transistore: come si vedrà nel seguito, l'effetto transistore consiste nella possibilità di ottenere una corrente di collettore sotto il controllo della tensione base-emettitore, ovvero un generatore di corrente pilotato. Da questo punto di vista, la I_{Ep} rappresenta una perdita, cioè una parte di corrente non utile a controllare il generatore pilotato e quindi da minimizzare.

Al terminale di collettore, la corrente è formata da due componenti (figura 4.8)

$$I_C = I_{Cn} + I_{C0} \tag{4.6}$$

- ▷ la corrente I_{Cn} è dovuta agli elettroni che attraversano l'intera regione di base e sono quindi raccolti al collettore
- ▷ la I_{C0} è la corrente inversa della giunzione base-collettore

La lunghezza della regione di base W_B deve essere molto più breve di L_{nB}, lunghezza di diffusione degli elettroni nella base: in queste condizioni, si ottiene il trasporto dal-

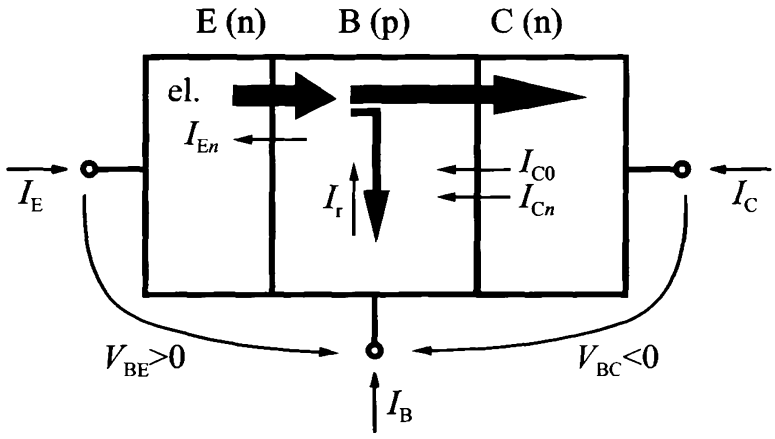


Figura 4.8 Componenti della corrente di collettore.

l'emettitore fino al collettore di un'elevata percentuale di portatori minoritari (elettroni per il dispositivo *npn*).

L'effetto della ricombinazione sui portatori iniettati attraverso una giunzione polarizzata direttamente è stato analizzato nel paragrafo 3.3. In particolare è stato evidenziato che la concentrazione in eccesso dei portatori minoritari, iniettati dal lato opposto della giunzione, diminuisce progressivamente spostandosi attraversa la regione quasi neutra. Tale diminuzione ha andamento diverso a seconda delle dimensioni relative dell'ampiezza della regione quasi neutra e della lunghezza di diffusione dei portatori. Come si è visto, nel caso del transistore bipolare, la regione di base ha sempre ampiezza molto piccola rispetto alla lunghezza di diffusione dei portatori:

$$W_B \ll L_{nB}$$

Si può quindi assumere una distribuzione lineare della concentrazione dei portatori, che implica la presenza di una corrente costante di diffusione associata ai portatori. Come si è visto nel paragrafo 3.3 e in particolare nell'approfondimento 3.2, tale corrente costante risulta inversamente proporzionale all'ampiezza della regione quasi neutra. Poiché la base di un transistore bipolare ha dimensioni piccole, la corrente associata alla diffusione degli elettroni attraverso la base è significativa e, per dispositivi reali, nettamente dominante rispetto alla corrente inversa di saturazione della giunzione base collettore.

Si approssima quindi la corrente di collettore con la sola prima componente indicata nella 4.6, ovvero I_{Cn}

$$I_C \approx I_{Cn}$$

questa corrente è legata agli elettroni residui che dalla base si riversano nel collettore, ovvero agli elettroni non ricombinati nell'attraversamento della base.

Se si tiene conto che la ricombinazione è un fenomeno statistico che coinvolge un numero di portatori proporzionale a quelli iniettati e che la corrente iniettata dall'emettitore, I_{En}, e quella portata al collettore, I_{Cn}, sono proporzionali alle concentrazioni

dei portatori alle giunzioni di emettitore e collettore rispettivamente, possiamo dedurre che I_{Cn} è proporzionale a I_{En} e la costante di proporzionalità è una misura dell'entità della ricombinazione verificatasi in base.

Si indica tale coefficiente con α_T, detto fattore di trasporto, e si esprime la corrente di collettore, in regione attiva diretta, come

$$I_C \approx I_{Cn} = \alpha_T I_{En}$$

Poiché la ricombinazione non può comunque essere completamente annullata, α_T è sempre minore di 1, ma può avvicinarsi molto all'unità in buoni transistori bipolari.

Infine la corrente di base si può ricavare indirettamente dalle correnti di emettitore e collettore applicando la legge di Kirchhoff per le tre correnti entranti nel dispositivo:

$$\begin{aligned} I_B + I_E + I_C &= 0 \\ I_B &= -I_E - I_C \end{aligned}$$

Si sostituiscono ora le I_E e I_C e si ottiene

$$\begin{aligned} I_B &= I_{En} + I_{Ep} - I_{Cn} - I_{C0} \\ I_B &= I_{Ep} + I_r - I_{C0} \end{aligned}$$

dove I_r è la corrente di ricombinazione, definita come

$$I_r = I_{En} - I_{Cn}$$

La corrente di base è quindi composta da tre contributi (figura 4.9):

- I_{Ep}, dovuta alle lacune iniettate dalla base nell'emettitore
- I_{C0}, la corrente inversa della giunzione base collettore
- I_r, corrente di ricombinazione nella base

Si noti che la corrente inversa di una giunzione è sempre molto piccola, in condizioni di temperatura ambiente, e anche I_{Ep} e I_r sono piccole in un buon transistore bipolare, con elevati coefficienti di emettitore e fattore di trasporto. Ne consegue che la corrente di base deve assumere valori nettamente inferiori a I_E e I_C.

Le grandezze introdotte nelle precedenti espressioni per le correnti di emettitore e collettore permettono di definire ora il guadagno di corrente. Trascurando la corrente di ricombinazione, I_{C0}, si scrivono I_C e I_E come:

$$\begin{aligned} I_C &\approx \alpha_T I_{En} \\ I_{En} &= -\gamma I_E \end{aligned}$$

Definito il guadagno α_F come prodotto dell'efficienza di emettitore e del fattore di trasporto,

$$\alpha_F = \frac{I_{En}}{I_E} = \gamma \alpha_T$$

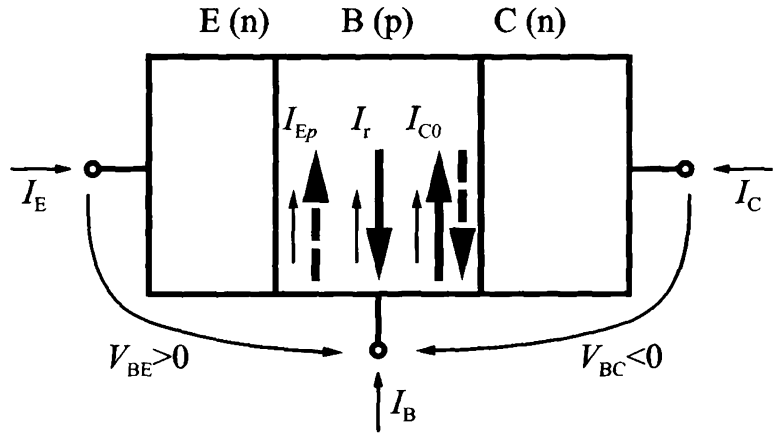


Figura 4.9 Componenti della corrente di base.

per sostituzione, si ha

$$I_C \approx -\alpha_T \gamma I_E = -\alpha_F I_E$$

e dalla legge di Kirchhoff:

$$I_B = -I_E - I_C = \left(\frac{1}{\alpha_F} - 1\right) I_C$$
$$I_C = \frac{\alpha_F}{1-\alpha_F} I_B = \beta_F I_B$$

dove il guadagno di corrente β_F è definito come

$$\beta_F = -\frac{I_C}{I_B} = \frac{\alpha_F}{1-\alpha_F} \tag{4.7}$$

Per un buon transistore bipolare, il fattore di trasporto è molto prossimo a uno e anche l'efficienza di emettitore è vicina all'unità; pertanto il guadagno α_F è anch'esso di poco inferiore a 1 e di conseguenza β_F risulta potenzialmente elevato

$$\alpha_F \approx 1 \qquad \rightarrow \qquad \beta_F = \frac{\alpha_F}{1-\alpha_F} \gg 1$$

I parametri tecnologici che determinano il valore di α_F, come le concentrazioni delle impurità droganti o le lunghezze di diffusione, sono caratterizzati da imprecisioni che si ripercuotono su α_F. La dipendenza di β_F da α_F è tuttavia tale da amplificare le imprecisioni tecnologiche e rendere il guadagno di corrente estremamente difficile da controllare.

Tale effetto può essere evidenziato calcolando l'errore relativo di β_F rispetto a α_F:

$$\frac{\Delta\beta_F}{\beta_F} = \frac{1}{1-\alpha_F} \times \frac{\Delta\alpha_F}{\alpha_F}$$

Per esempio, in un transistore con $\alpha_F = 0{,}999$ e errore relativo contenuto del 2% ($\frac{\Delta\alpha_F}{\alpha_F} = 0{,}02$), si ha un guadagno di corrente nominale β_F pari a 1000 ma affetto da un errore relativo molto elevato:

$$\frac{\Delta\beta_F}{\beta_F} = 20$$

4.3.1 Efficienza di emettitore

L'efficienza di emettitore è definita come fattore di proporzionalità tra la corrente di emettitore e la componente I_{En}

$$I_{En} = -\gamma I_E$$

In un transistore integrato, la profondità della giunzione di emettitore è piccola rispetto alla lunghezza di diffusione ("emettitore corto")

$$x_E \ll L_{pE}$$

e l'efficienza di emettitore si valuta come

$$\gamma = \left[1 + \frac{N_B w_B D_{pE}}{N_E x_E D_{nB}}\right]^{-1} \tag{4.8}$$

(l'espressione di gamma è ricavata nell'approfondimento 4.1).
Nel caso invece di emettitore lungo ($x_E \gg L_{pE}$), l'espressione di γ cambia in

$$\gamma = \left[1 + \frac{N_B w_B D_{pE}}{N_E L_{pE} D_{nB}}\right]^{-1}$$

Per i transistori integrati, γ è il fattore dominante nel limitare il guadagno complessivo del dispositivo.

Approfondimento 4.1 In questo approfondimento, si deriva la relazione (4.8) dell'efficienza di emettitore γ per un transistore n-p-n che presenta drogaggi con concentrazioni costanti nelle tre regioni.
In un transistore n-p-n la corrente utile è quella degli elettroni iniettati dall'emettitore nella base. Se si trascura la ricombinazione nella zona di carica spaziale della giunzione E-B[1], la frazione utile di corrente che attraversa tale giunzione è

$$\gamma = \frac{|I_{nE}|}{|I_{nE}| + |I_{pE}|} = \frac{|I_{nB}(0)|}{|I_{nB}(0)| + |I_{pE}(0)|}$$

[1] Con questa approssimazione, valida per il calcolo delle correnti, la larghezza della regione spaziale si può ritenere nulla. La sezione $x = 0$ che indica la regione di carica spaziale è anche quella di inizio della regione di base. Considerazioni analoghe sono valide per la giunzione base-collettore; la sezione $x = w_B$ è quella di inizio della regione di base.

dove $I_{nB}(0)$ e $I_{pE}(0)$ sono le correnti di elettroni e lacune misurate in $x = 0$, ovvero alla giunzione base–emettitore.
Per valutare γ si ricavano le correnti. Per la corrente di diffusione degli elettroni nella base si applica l'equazione di continuità, con le ipotesi di quasi neutralità ($\mathcal{E} = 0$) nella regione di base e che la generazione sia di tipo termico, $U = n'/\tau_n$. In condizioni di stazionarietà $\left(\frac{\partial}{\partial t} = 0\right)$ si ottiene

$$D_n \frac{\partial^2 n'}{\partial x^2} = \frac{n'}{\tau_n}\,; \quad L_{nB} = \sqrt{D_n \tau_n}$$
$$n'(x) = A \exp(-x/L_{nB}) + B \exp(x/L_{nB})$$

come già visto nell'esempio 2.3. Le condizioni al contorno pongono in relazione le costanti A e B con le concentrazioni agli estremi della base

$$\begin{aligned} n'(0) &= A + B \\ n'(w_B) &= A \exp(-w_B/L_{nB}) + B \exp(w_B/L_{nB}) \\ &= n'(0) \exp(-w_B/L_{nB}) + B\,[\exp(w_B/L_{nB}) - \exp(-w_B/L_{nB})] \end{aligned}$$

e consentono di ottenere

$$B = \frac{n'(w_B) - n'(0) \exp(-w_B/L_{nB})}{\exp(w_B/L_{nB}) - \exp(-w_B/L_{nB})}$$

$$\begin{aligned} A &= \frac{n'(0) \exp(w_B/L_{nB}) - n'(0) \exp(-w_B/L_{nB}) - n'(w_B) + n'(0) \exp(-w_B/L_{nB})}{\exp(w_B/L_{nB}) - \exp(-w_B/L_{nB})} \\ &= \frac{n'(0) \exp(-w_B/L_{nB}) - n'(w_B)}{\exp(w_B/L_{nB}) - \exp(-w_B/L_{nB})} \end{aligned}$$

Sostituendo si ha

$$\begin{aligned} n'(x) = {} & \frac{1}{\exp(w_B/L_{nB}) - \exp(-w_B/L_{nB})} \left\{ -n'(0) \left[\exp\left(\frac{x - w_B}{L_{nB}}\right) - \exp\left(-\frac{x - w_B}{L_{nB}}\right) \right] \right. \\ & \left. + n'(w_B)\,[\exp(x/L_{nB}) - \exp(-x/L_{nB})] \right\} \end{aligned}$$

$$n'(x) = -n'(0) \frac{\sinh\left(\frac{x - w_B}{L_{nB}}\right)}{\sinh\left(\frac{w_B}{L_{nB}}\right)} + n'(w_B) \frac{\sinh\left(\frac{x}{L_{nB}}\right)}{\sinh\left(\frac{w_B}{L_{nB}}\right)}$$

Nell'ipotesi di basso livello di iniezione, vale la legge della giunzione, che permette di esprimere la concentrazione in eccesso di portatori ai bordi della base in funzione delle tensioni applicate:

$$\begin{aligned} n'(0) &= n_{pB0}\,[\exp(V_{BE}/V_T) - 1] \\ n'(w_B) &= n_{pB0}\,[\exp(V_{BC}/V_T) - 1] \end{aligned}$$

La corrente di diffusione di elettroni dall'emettitore nella base è data da

$$I_{nB}(x) = q\,S\,D_n \frac{\mathrm{d}n'}{\mathrm{d}x} = \tag{4.9}$$

$$= q\,S\,D_n \left[-n'(0) \frac{1}{L_{nB}} \frac{\cosh\left(\frac{x - w_B}{L_{nB}}\right)}{\sinh\left(\frac{w_B}{L_{nB}}\right)} + n'(w_B) \frac{1}{L_{nB}} \frac{\cosh\left(\frac{x}{L_{nB}}\right)}{\sinh\left(\frac{w_B}{L_{nB}}\right)} \right] \tag{4.10}$$

(4.11)

$$I_{nB}(0) = -\frac{q\,S\,D_n}{L_{nB}}\,n'(0)\,\coth\left(\frac{w_B}{L_{nB}}\right) + \\ + \frac{q\,S\,D_n}{L_{nB}}\,\frac{1}{\sinh\left(\frac{w_B}{L_{nB}}\right)}\,n'(w_B)$$

dove S è l'area della giunzione BE. La corrente di lacune iniettate dalla base nell'emettitore, nell'ipotesi di diodo lungo, è

$$I_{pE}(x_E) = I_{pE}(0) = -q\,S\,D_p\frac{\mathrm{d}p'}{\mathrm{d}x} = -\frac{q\,S\,D_p}{L_p}\,p_{E0}\,[\exp(V_{BE}/V_T) - 1] \qquad (4.12)$$

Dato che la base è corta, $w_B \ll L_{nB}$, si approssima l'espressione di $I_{n_B}(0)$ sostituendo la coth con l'inverso del suo argomento e la sinh con l'argomento

$$I_{nB}(0) \simeq -\frac{q\,S\,D_n\,n_{p_{B0}}}{w_B}\,[\exp(V_{BE}/V_T) - 1] + \frac{q\,S\,D_n\,n_{p_{B0}}}{w_B}\,[\exp(V_{BC}/V_T) - 1] \qquad (4.13)$$

$$= -\frac{q\,S\,D_n}{w_B}\,\frac{n_i^2}{N_A}\,[\exp(V_{BE}/V_T) - \exp(V_{BC}/V_T)] \qquad (4.14)$$

In regione attiva diretta di funzionamento, la giunzione base collettore è polarizzata inversamente pertanto $V_{BC} < 0$ quindi $\exp(V_{BC}/V_T) \simeq 0$. Le due correnti cercate si riducono quindi a

$$I_{nB}(0) \simeq -\frac{q\,S\,D_n n_i^2}{w_B N_A}\,\exp(V_{BE}/V_T)$$

$$I_{pE}(0) \simeq -\frac{q\,S\,D_p n_i^2}{L_p N_D}\,\exp(V_{BE}/V_T)$$

Se l'emettitore è molto corto (come avviene nei circuiti integrati reali) $L_{pE} \gg x_E$, la corrente di lacune iniettate dalla base nell'emettitore si può esprimere come

$$I_{pE}(0) \simeq -\frac{q\,S\,D_p\,n_i^2}{x_E\,N_D}\,\exp(V_{BE}/V_T)$$

Si ottiene quindi per γ l'espressione

$$\gamma = \frac{|I_{nB}(0)|}{|I_{nB}(0)| + |I_{pE}(0)|} = \frac{1}{1 + \left|\frac{I_{pE}(0)}{I_{nB}(0)}\right|}$$

$$\left|\frac{I_{pE}(0)}{I_{nB}(0)}\right| = \frac{q\,D_p\,n_i^2}{x_E\,N_D}\,\frac{w_B\,N_A}{q\,D_n\,n_i^2} = \frac{D_p\,w_B\,N_A}{D_n\,x_E\,N_D}$$

$$\gamma = \frac{1}{1 + \frac{D_p\,w_B\,N_A}{D_n\,x_E\,N_D}}$$

Per i transistori integrati, il valore di γ è generalmente superiore a 0,98. Osservando l'espressione di γ,

$$\gamma = \left[1 + \frac{N_B w_B D_{pE}}{N_E x_E D_{nB}}\right]^{-1}$$

si vede che, in generale, per massimizzare l'efficienza di emettitore si possono imporre alcune condizioni sui parametri tecnologici coinvolti; per esempio è certamente utile

che

$$N_E \gg N_B$$

Aiuta a massimizzare γ anche scegliere x_E grande e ridurre la ricombinazione di lacune nell'emettitore. Infine è importante contenere l'ampiezza della base, scegliendo valori di w_B piccoli.

Esempio 4.1 Con $x_E \approx W_E = 1$ μm, $w_B \approx W_B = 5$ μm, $\mu_{nB} = 1500$ cm^{2}V^{-1}s^{-1}, $\mu_{pE} = 500$ cm^{2}V^{-1}s^{-1}, $\tau_n = \tau_p = 10$ μs, si ha

$$\gamma = 0{,}9983 \text{ per } N_E = 10^{18} \text{ cm}^{-3},\ N_B = 10^{15} \text{ cm}^{-3}$$
$$\gamma = 0{,}8571 \text{ per } N_E = 10^{17} \text{ cm}^{-3},\ N_B = 10^{16} \text{ cm}^{-3}$$

4.3.2 Fattore di trasporto

Il fattore di trasporto tiene conto del fenomeno di ricombinazione dei portatori minoritari nell'attraversamento della base: maggiore è α_T e minore risulta la percentuale di portatori iniettati nella base che si ricombinano prima di essere raccolti dal collettore. Dalla definizione, si ha per α_T

$$\alpha_T = -\frac{I_{Cn}}{I_{En}} = \frac{I_{En} - I_r}{I_{En}} = 1 - \frac{I_r}{I_{En}}$$

Il valore del fattore di trasporto dipende dalla dimensione della base w_B e dalla lunghezza di diffusione L_{nB}, secondo l'espressione seguente, che è ricavata nell'approfondimento 4.2:

$$\alpha_T = 1 - \frac{w_B^2}{2\tau_n D_{nB}} = 1 - \frac{w_B^2}{2L_{nB}^2} \tag{4.15}$$

Pertanto, al fine di contenere l'effetto di ricombinazione e ottenere un buon transistore bipolare, è indispensabile che sia $w_B \ll L_{nB}$, ovvero la base deve avere ampiezza molto ridotta rispetto alla lunghezza di diffusione dei portatori minoritari nella base.

Approfondimento 4.2 In questo approfondimento, si deriva prima la corrente di ricombinazione in base e poi l'espressione (4.15) del fattore di trasporto.
In condizioni di stazionarietà, la popolazione di elettroni in base è costante e deriva dall'equilibrio tra i fenomeni di iniezione dall'emettitore e ricombinazione nella base; tale popolazione è comunque molto più elevata della concentrazione di elettroni che si avrebbero in assenza di iniezione. Indichiamo con Q'_{nB} la carica di elettroni in eccesso nella base e con τ_n il tempo di vita medio, dipendente principalmente dal livello di drogaggio. Evidentemente, se Q'_{nB} è il livello aggiuntivo di carica raggiunto come effetto dell'iniezione e τ_n è l'intervallo di tempo medio entro il quale un elettrone iniettato si ricombina, la corrente originata dal fenomeno della ricombinazione, I_{rB}, si può esprimere come rapporto tra le due grandezze:

$$I_{rB} = \frac{Q'_{nB}}{\tau_n}$$

Assumendo come riferimento la posizione della giunzione base-emettitore e indicando con w_B la larghezza della base, si ha

$$\mathrm{d}Q_n = qA\,n'_B(x)\,\mathrm{d}x$$

La corrente di ricombinazione si calcola allora come

$$I_{rB} = \int_0^{Q_n} \frac{1}{\tau_n}\,\mathrm{d}Q_n = \frac{qA}{\tau_n}\int_0^{w_B} n'_B(x)\,\mathrm{d}x$$

quindi il valore di I_{rB} è legato all'area sottesa dalla curva di $n'_B(x)$; assumendo per $n'_B(x)$ un andamento lineare, si ha

$$n'_B(x) = \frac{n'_B(w_B)-n'_B(0)}{w_B}\,x + n'_B(0)$$

$$n'_B(0) = n_{B0}\left[\exp(V_{BE}/V_T)-1\right] \quad ; \quad n'_B(w_B) = n_{B0}\left[\exp(V_{BC}/V_T)-1\right]$$

$$I_{rB} = \frac{qA}{\tau_n}\int_0^{w_B}\left[\frac{n'_B(w_B)-n'_B(0)}{w_B}\,x + n'_B(0)\right]\mathrm{d}x$$

$$I_{rB} = \frac{2qA}{\tau_n}\,w_B\left(n'_B(0)+n'_B(w_B)\right)$$

In polarizzazione diretta, si ha

$$I_{rB} \simeq \frac{qA\,n_i^2\,w_B}{2\tau_n N_A}\,\exp(V_{BE}/V_T)$$

Il fattore di trasporto α_T si può quindi esprimere come

$$\alpha_T = \frac{|I_{nB}|-|I_{rB}|}{|I_{nB}|} = 1-\frac{|I_{rB}|}{|I_{nB}|} = 1-\frac{qn_i^2\,w_B}{2\tau_n\,N_A}\,\frac{w_B\,N_A}{qD_n\,n_i^2}$$

$$= 1-\frac{w_B^2}{2D_n\,\tau_n} = 1-\frac{w_B^2}{2L_{nB}^2}$$

Alternativamente, il fattore di trasporto può essere ricavato osservando che la corrente di portatori minoritari iniettata in base è $I_{nB}(0)$, mentre quella raccolta dal collettore è $I_{nB}(w_B)$. Il fattore di trasporto può quindi essere valutato anche come rapporto tra le due correnti:

$$\alpha_T = \frac{I_{nB}(w_B)}{I_{nB}(0)}$$

Le due correnti, valutate agli estremi della regione di base, possono essere ottentute dall'equazione (4.9) sostituendo $x=0$ e $x=w_B$:

$$I_{nB}(0) = -\frac{q\,S\,D_n}{L_{nB}}\,n'(0)\,\coth(\frac{w_B}{L_{nB}}) + \frac{q\,S\,D_n}{L_{nB}}\,n'(w_B)\,\frac{1}{\sinh\left(\frac{w_B}{L_{nB}}\right)}$$

$$I_{nB}(w_B) = -\frac{q\,S\,D_n}{L_{nB}}\,n'(0)\,\frac{1}{\sinh\left(\frac{w_B}{L_{nB}}\right)+\frac{q\,S\,D_n}{L_{nB}}\,n'(w_B)\,\coth(\frac{w_B}{L_{nB}})}$$

Il rapporto fornisce α_T:

$$\alpha_T = \frac{n'(0)-n'(w_B)\cosh\left(\frac{w_B}{L_{nB}}\right)}{n'(0)\cosh\left(\frac{w_B}{L_{nB}}\right)-n'(w_B)}$$

Poiché $w_B \ll L_{nB}$, si può approssimare il coseno iperbolico troncandone lo sviluppo in serie al termine di grado due:

$$\alpha_T \approx \frac{n'(0)-n'(w_B)\left(1+\frac{1}{2}\frac{w_B^2}{L_{nB}^2}\right)}{n'(0)\left(1+\frac{1}{2}\frac{w_B^2}{L_{nB}^2}\right)-n'(w_B)}$$

Considerando poi che, in regione attiva diretta,

$$n'(0) = n_{pB0}\,[\exp(V_{BE}/V_T) - 1] \approx n_{pB0}\,\exp(V_{BE}/V_T)$$
$$n'(w_B) = n_{pB0}\,[\exp(V_{BC}/V_T) - 1] \approx -n_{pB0}$$

il rapporto di correnti si semplifica ulteriormente:

$$\alpha_T \approx \frac{1}{1 + \frac{1}{2}\frac{w_B^2}{L_{nB}^2}} \approx 1 - \frac{1}{2}\frac{w_B^2}{L_{nB}^2}$$

Per i BJT moderni, l'ampiezza della base è inferiore al μm mentre la lunghezza di diffusione L_{nB} è superiore a 30 μm. Questa differenza è sufficiente a garantire un buon comportamento, portando a un guadagno $\alpha_T > 0{,}9994$, che non è quindi da considerarsi il principale fattore limitante di α_F.

Per esempio, con $\alpha_T = 0{,}9994$ e un'efficienza di emettitore $\gamma = 0{,}9983$, si ha

$$\alpha_F = 0{,}9977 \qquad \text{e} \qquad \beta_F = 433$$

Esempio 4.2 Si intende determinare il valore del fattore di trasporto, per un transistore caratterizzato dalle seguenti grandezze:

- ▷ lunghezza di base $w_B = 1$ μm
- ▷ costante di diffusione nella base $D_n = 36{,}8\,\text{cm}^2\text{s}^{-1}$
- ▷ tempo di vita medio $\tau_n = 2{,}5 \cdot 10^{-3}$ s

Il fattore di trasporto α_T è definito come

$$\alpha_T = \frac{|I_{nB}| - |I_{rB}|}{|I_{nB}|} = 1 - \frac{|I_{rB}|}{|I_{nB}|}$$

$$\alpha_T = 1 - \frac{q n_i^2 w_B}{2\tau_n N_A}\frac{w_B N_A}{q D_n n_i^2}$$

$$\alpha_T = 1 - \frac{w_B^2}{2D_n \tau_n} = 1 - \frac{w_B^2}{2L_{nB}^2}$$

Dai valori della costante di diffusione e del tempo di vita medio dei portatori minoritari in base, $D_n = 36{,}8\,\text{cm}^2\text{s}^{-1}$ e $\tau_n = 2{,}5 \cdot 10^{-3}$ s si ha

$$L_{nB}^2 = D_n \tau_n = 0{,}092\ \text{cm}^2 \quad L_{nB} = 0{,}3\ \text{cm}$$

Il fattore di trasporto risulta assai prossimo a 1 e vale precisamente in questo caso

$$\alpha_T = 0{,}9998$$

Nel transistore bipolare, il guadagno di corrente β_F è caratterizzato da una significativa dispersione: questo significa che il parametro β_F risente fortemente dalla variazione statistica delle grandezze tecnologiche che determinano γ e α_T e risulta

quindi difficilmente controllabile. Ne consegue che dispositivi nominalmente identici mostrano in realtà guadagni di corrente molto diversi, rendendo così problematica la realizzazione di circuiti con comportamento riproducibile.

La dispersione del guadagno β_F non può essere evitata e anche imprecisioni ridottissime nelle grandezze tecnologiche che lo influenzano ne determinano forti variazioni. Pertanto la stabilità e riproducibilità del comportamento dei circuiti a transistore bipolare si ottiene introducendo opportune retroazioni che compensano la variazione di β_F.

Esempio 4.3 In un transistore bipolare con $\beta_F = 100$, l'efficienza di emettitore γ è caratterizzata da un errore relativo pari a $\Delta\gamma/\gamma = 1\%$, e anche per il fattore di trasporto si ha $\Delta\alpha_T/\alpha_T = 1\%$

Si intende calcolare l'errore relativo sui guadagni α_F e β_F.

I due guadagni sono definiti come

$$\begin{aligned} \alpha_F &= \gamma\alpha_T \\ \beta_F &= \frac{\alpha_F}{1-\alpha_F} \end{aligned}$$

L'errore su α_F si può ottenere calcolando le derivate parziali

$$\begin{aligned} \Delta\alpha_F &= \frac{\partial\alpha_F}{\partial\gamma}\Delta\gamma + \frac{\partial\alpha_F}{\partial\alpha_T}\Delta\alpha_T \\ &= \alpha_T\Delta\gamma + \gamma\Delta\alpha_T \end{aligned}$$

L'errore relativo è dato dalla somma degli errori relativi su γ e α_T:

$$\frac{\Delta\alpha_F}{\alpha_F} = \frac{\alpha_T\Delta\gamma + \gamma\Delta\alpha_T}{\alpha_F}$$

$$\frac{\Delta\alpha_F}{\alpha_F} = \frac{\alpha_T\Delta\gamma + \gamma\Delta\alpha_T}{\gamma\alpha_T}$$

$$\frac{\Delta\alpha_F}{\alpha_F} = \frac{\Delta\gamma}{\gamma} + \frac{\Delta\alpha_T}{\alpha_T}$$

Con lo stesso metodo, si ottiene l'errore assoluto su β_F:

$$\Delta\beta_F = \frac{\partial\beta_F}{\partial\alpha_F}\Delta\alpha_F = \frac{1}{(1-\alpha_F)^2}\Delta\alpha_F$$

L'errore relativo è

$$\frac{\Delta\beta_F}{\beta_F} = \frac{1}{\beta_F}\frac{1}{(1-\alpha_F)^2}\Delta\alpha_F = \frac{1}{1-\alpha_F}\frac{\Delta\alpha_F}{\alpha_F}$$

Sostituendo i valori numerici, si ha

$$\frac{\Delta\alpha_F}{\alpha_F} = 2\%$$

$$\frac{\Delta\beta_F}{\beta_F} = 200\%$$

4.4 Equazioni di Ebers-Moll

I risultati ottenuti nei paragrafi precedenti in termini di equazioni che descrivono le correnti nel transistore in funzione delle tensioni applicate possono essere raccolti in un sistema di equazioni note come *equazioni di Ebers-Moll.* Nella loro forma generale, queste equazioni forniscono le correnti di emettitore e collettore:

$$\begin{aligned} I_E &= a_{11}[\exp(V_{BE}/V_T) - 1] + a_{12}[\exp(V_{BC}/V_T) - 1] \\ I_C &= a_{21}[\exp(V_{BE}/V_T) - 1] + a_{22}[\exp(V_{BC}/V_T) - 1] \end{aligned}$$

e i quattro coefficienti si ricavano come indicato per esempio nell'approfondimento 4.1:

$$\begin{aligned} a_{11} &= -\frac{q\,S\,D_{nB}\,n_i^2}{w_B\,N_B} - \frac{q\,S\,D_{pE}\,n_i^2}{L_{pE}\,N_E} \\ a_{12} &= \frac{q\,S\,D_{nB}\,n_i^2}{w_B\,N_B} = a_{21} \\ a_{22} &= -\frac{q\,S\,D_{nB}\,n_i^2}{w_B\,N_B} - \frac{q\,S\,D_{pC}\,n_i^2}{L_{pC}\,N_C} \end{aligned}$$

S è l' area della giunziona base-emettitore; la lunghezza di diffusione delle lacune nell'emettitore L_{pE} deve essere sostituita con la profondità dell'emettitore quando l'emettitore non può essere considerato lungo.

4.5 Modello di Ebers-Moll

Il modello di Ebers-Moll è un modello circuitale statico approssimato usabile in tutte le regioni di funzionamento e permette di determinare le correnti ai terminali del componente in funzione delle tensioni applicate.

Il modello si costruisce a partire da una coppia di diodi associati alle giunzioni base-emettitore e base-collettore del transistore. Le componenti di corrente attraverso questi diodi mostrano dipendenza di tipo esponenziale dalle tensioni applicate alle giunzioni, secondo il comportamento già ricavato per la giunzione pn:

$$\begin{aligned} I_F &= I_{ES}\left[\exp\left(\frac{V_{BE}}{V_T}\right) - 1\right] \\ I_R &= I_{CS}\left[\exp\left(\frac{V_{BC}}{V_T}\right) - 1\right] \end{aligned}$$

I_F è la corrente associata al diodo che modellizza la giunzione base-emettitore e il pedice F, iniziale del termine anglosassone *forward* fa riferimento proprio alla giunzione che risulta polarizzata direttamente in regione attiva diretta. Analogamente il pedice R rimanda al termine *reverse*, associato al diodo che modellizza la giunzione base-collettore, polarizzata inversamente nella regione attiva diretta.

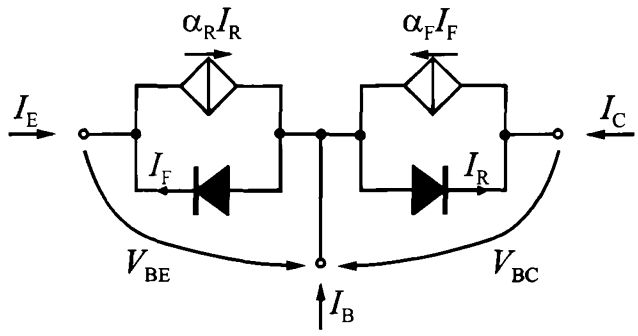


Figura 4.10 Modello di Ebers-Moll.

I due diodi tuttavia non sono in grado di riprodurre l'effetto transistore, che può essere modellizzato aggiungendo in paralleo ai diodi due generatori di corrente pilotati, come indicato nella figura 4.10. Il generatore pilotato posizionato dal lato di collettore eroga una corrente $\alpha_F I_F$, proporzionale alla corrente che si ha attraverso la giunzione base-emettitore nella regione attiva diretta, mentre l'altro generatore modellizza il controllo che la corrente di collettore esercita su quella di emettitore nella regione attiva inversa.

Dai contributi di corrente indicati nella figura, si possono ricavare le correnti ai terminali di emettitore e collettore

$$\begin{aligned} I_E &= -I_F + \alpha_R I_R \\ I_C &= -I_R + \alpha_F I_F \end{aligned}$$

e da queste ottenere la corrente di base

$$I_B = I_F(1 - \alpha_F) + I_R(1 - \alpha_R)$$

Il modello è abbastanza completo da includere la descrizione del funzionamento nelle quattro regioni di funzionamento; tuttavia è molto più frequente la necessità di utilizzare il modello di Ebers-Moll per descrivere, in forma approssimata, il funzionamento in una specifica regione di funzionamento e, per questo scopo, è possibile semplificare il modello generale dato nella figura 4.10.

Poiché, nella regione attiva diretta, è $V_{BC} < 0$, si può assumere $I_R \approx 0$ e trascurare quindi il diodo dal lato di collettore e il generatore da quello di emettitore. Le correnti di collettore e base in questo caso si possono scrivere

$$\begin{aligned} I_C &\approx \alpha_F I_F \\ I_B &\approx I_F(1 - \alpha_F) \end{aligned}$$

ed è poi immediato ottenere la relazione base del transistore:

$$I_C = \frac{\alpha_F}{1 - \alpha_F} I_B = \beta_F I_B$$

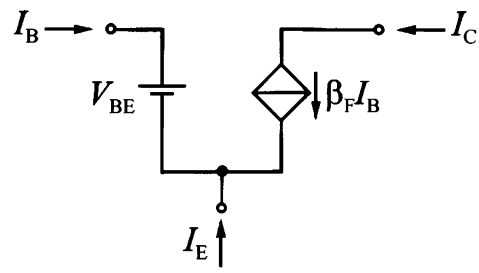


Figura 4.11 Modello di Ebers-Moll in regione attiva diretta.

Come ulteriore semplificazione circuitale, si può considerare che, se la giunzione B-E è polarizzata direttamente, la si può modellizzare con un semplice generatore di tensione di valore pari a $V_{BE} = 0{,}7$ V, ottenendo lo schema nella figura 4.11.

L'effetto transistore è riprodotto dal generatore di corrente pilotato, che eroga una corrente proporzionale alla I_B.

4.6 Effetto Early

Il comportamento descritto nelle sezioni precedenti riguarda il transistore ideale, mentre il comportamento del dispositivo reale presenta alcune deviazioni tra le quali una delle più significative è il cosiddetto effetto Early.

Nel transistore ideale, l'ampiezza della regione quasi neutra della base è considerata costante, invariante rispetto alla tensione applicata alla giunzione base-collettore. In realtà, poiché questa giunzione in regione attiva diretta è polarizzata inversamente, la zona svuotata è modulata dalla tensione applicata e questo ha un effetto sull'ampiezza della regione quasi neutra della base.

Questo fenomeno è anche noto come *odulazione della lunghezza di base*; se la tensione base-collettore, V_{BC}, cresce in modulo, la regione di svuotamento si allarga e quindi la larghezza della regione quasi neutra della base, W_B, si riduce.

La riduzione di w_B ha due conseguenze:

- si riduce il tasso di ricombinazione e quindi aumenta il fattore di trasporto α_T
- aumenta l'iniezione dei portatori minoritari in base, cioè migliora γ

Entrambe le conseguenze contribuiscono a migliorare il guadagno β_F.

L'effetto è ben visibile sulle curve caratteristiche del dispositivo, in particolare sulla caratteristica di uscita della corrente di collettore in funzione della tensione collettore emettitore, perché a parità di I_B, la corrente I_C cresce con $|V_{BC}|$.

È certamente possibile ricavare dall'analisi dispositivistica del fenomeno le espressioni per determinare l'effetto delle variazioni di V_{BC} sulla corrente; questo approccio appesantirebbe però molto il modello del transistore e quindi nelle simulazioni circuitali si preferisce adottare modelli semi-empirici.

La configurazione a emettitore comune prevede l'applicazione della tensione di ingresso tra base e emettitore e l'uscita disponibile sul morsetto di collettore, come indicato in figura 4.12. Poiché in questo caso il terminale di emettitore risulta comune alla

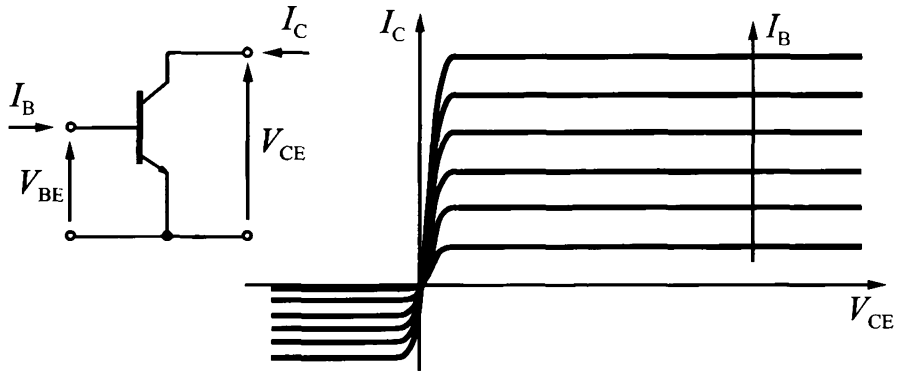


Figura 4.12 Caratteristica a emetitore comune.

maglia di ingresso e a quella di uscita, la configurazione prende il nome di *emettitore comune* o CE.

Per descrivere il comportamento statico del transistore in questa configurazione, si ricorre a due famiglie di curve, che riportano una la tensione base emettitore in funzione della corrente di base e l'altra la corrente di collettore al variare della tensione tra collettore e emettitore. Si descrivono quindi le due caratteristiche

$$\begin{aligned} V_{BE} &= V_{BE}(I_B, V_{CE}) \\ I_C &= I_C(I_B, V_{CE}) \end{aligned}$$

La seconda caratteristica, riportata nella figura 4.12, mostra una corrente di collettore costante in un'ampia fascia di valori di V_{CE}. Per il transistore reale invece, l'effetto Early evidenzia invece un progressivo aumento della I_C con V_{CE}, come indicato in figura 4.13.

Al fine di includere l'effetto Early nel comportamento del dispositivo, si adotta nella maggior parte dei casi un modello semi-empirico, che corregge la corrente di collettore con un contributo che dipende dalla tensione di collettore:

$$I_C = \beta_F I_B \left(1 + \frac{V_{CE}}{V_A}\right)$$

Il modello, ovviamente applicabile soltanto nella regione attiva diretta, include un parametro di calibrazione, V_A, avente le dimensioni di una potenziale e denominato tensione di Early: valori piccoli di V_A implicano un più marcato effetto Early.

Tale parametro ammette anche un'interapretazione grafica approssimativa ma intuitiva, descritta nella figura 4.13 e associata all'ascissa di confluenza del prolungamento delle curve di I_C nel semipiano delle tensioni negative. Nella regione di linearità, le

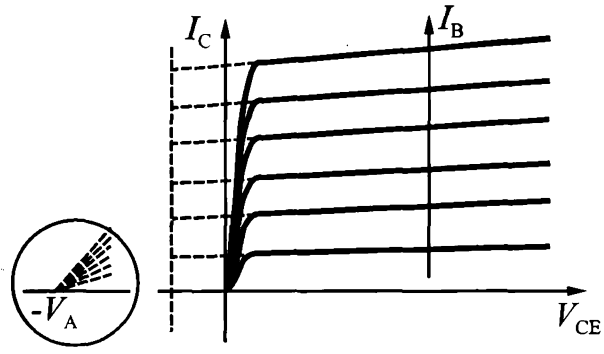


Figura 4.13 Caratteristica con effetto Early.

curve di corrente in figura 4.13 hanno pendenza

$$\frac{\partial I_C}{\partial V_{CE}} = \beta_F I_B \frac{1}{V_A}$$

Indicando con I_{C1} la tensione di collettore valutata all'inizio della regione attiva diretta e dunque per valori di V_{CE} molto prossimi a 0, il punto di convergenza sull'asse delle ascisse delle rette che prolungano le curve di corrente può essere identificato approssimativamente come rapporto tra I_{C1} e la pendenza

$$I_{C1} \frac{\partial V_{CE}}{\partial I_C} = \frac{I_{C1} V_A}{\beta_F I_B} \approx V_A$$

4.6.1 Caratteristica a base comune

Al variare del modo di utilizzo del transistore, cambiano le curve rappresentative del comportamento. Per esempio, una topologia alternativa a quella con emettitore comune è la configurazione a base comune, nella quale il terminale di base costituisce riferimento sia per la maglia di ingresso che per quella di uscita, come indicato nella parte sinistra della figura 4.14.

Il comportamento in queste condizione è descritto da due famiglie di curve:

$$V_{BE} = V_{BE}(I_E, V_{BC})$$
$$I_C = I_C(I_E, V_{BC})$$

Le curve di I_C in funzione della tensione base-collettore V_{BC}, con parametro I_E, sono dette caratteristiche di uscita e sono indicate nella parte destra della figura 4.14. La regione attiva diretta si estende lungo l'asse delle ascisse fino a circa 0,6 V oltre l'origine; in questa zona, la corrente di collettore è costante, a meno dell'effetto Early, e risulta

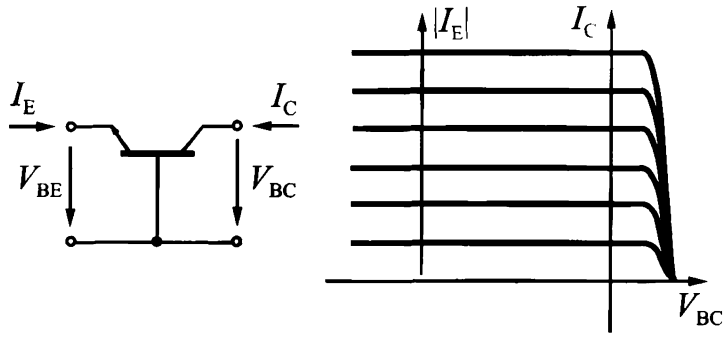


Figura 4.14 Caratteristica a base comune.

proporzionale alla corrente di emettitore I_E attraverso il guadagno α_F. Per $V_{BC} > 0{,}6$ V, la giunzione base-collettore risulta polarizzata direttamente e quindi la corrente di collettore cessa di essere controllata dalla I_E.

4.6.2 Altre regioni di funzionamento

Le altre regioni di funzionamento del dispositivo sono di minore interesse, almeno per le applicazioni analogiche, che tipicamente sfruttano l'effetto transistore e più precisamente il funzionamento come generatore pilotato che è tipico della regione attiva diretta.

Tuttavia, in alcuni casi, il transistore è utilizzato come interruttore, come avviene per esempio nella realizzazione di porte logiche elementari in tecnologia bipolare. In questi casi il dispositivo è alternativamente polarizzato nelle regioni di interdizione e saturazione, per ottenere il comportamento rispettivamente di un circuito aperto e di un circuito chiuso.

In regione di saturazione entrambe le giunzioni sono polarizzate direttamente; in genere $V_{BE} \geq 0{,}7$ V, mentre V_{BC} è prossima a 0,6 V. Per esempio, con $V_{BE} = V_{BEsat} = 0{,}8$ V e $V_{BC} = V_{BCsat} = 0{,}6$ V, si ha $V_{CE} = V_{CEsat} = 0{,}2$ V. Il modello circuitale equivalente si riduce quindi a una coppia di batterie, come indicato nella parte sinistra della figura 4.15.

In queste condizioni, il transistore non è in grado di imporre alcuna legge di dipendenza tra le correnti, che sono pertanto fissate dal circuito esterno.

In condizioni di interdizione, ovvero quando entrambe le giunzioni sono polarizzate inversamente, non si possono avere correnti apprezzabili e quindi, trascurando le correnti inverse di saturazione, il dispositivo si modellizza con una coppia di interruttori aperti (parte destra della figura 4.15).

Infine, nella regione attiva inversa i ruoli di emettitore e collettore sono scambiati e valgono le seguenti condizioni sulle tensioni:

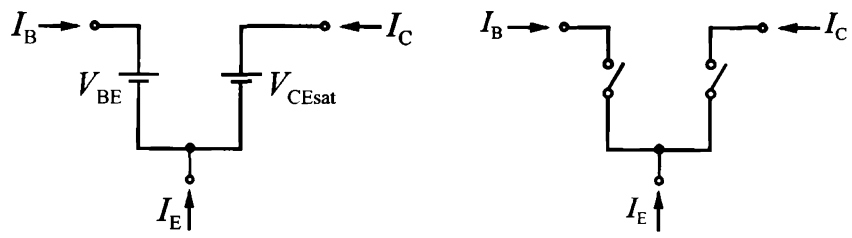


Figura 4.15 Regioni di saturazione e interdizione.

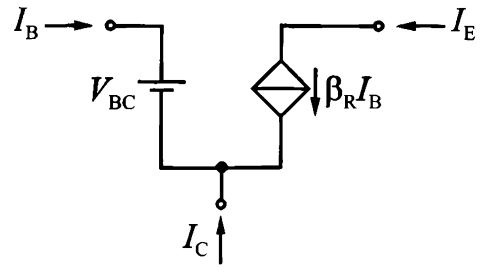


Figura 4.16 Modello in regione di funzionamento attiva inversa.

- la giunzione BE è polarizzata inversamente, $V_{BE} < 0$
- la giunzione BC è polarizzata direttamente, $V_{BC} \geq 0{,}7$

In questa configurazione si può ancora apprezzare l'effetto transistore, in linea di principio, perché la struttura del dispositivo appare simmetrica e quindi si può avere in questo caso una corrente di emettitore controllata dalla tensione base-collettore. Ripetendo i passaggi del paragrafo 4.3, si arriva anche in questo caso a determinare un guadagno di corrente β_R definito come

$$\beta_R = -\frac{I_E}{I_B}$$

in analogia a quanto descritto nella 4.7.

Nella realtà, alcune caratteristiche tecnologiche differenziano fortemente emettitore e collettore e questo cancella quasi completamente l'effetto transistore nella regione inversa. Per esempio, è stato visto in precedenza che il drogaggio dell'emettitore dev'essere di molto superiore a quello della base affinché si possa ottenere un elevato guadagno β_F; poiché la stessa condizione tecnologica non è verificata dal lato del collettore, ovvero non si può avere $N_C \gg N_B$ e quindi l'efficienza di collettore tende a essere nettamente inferiore all'efficienza di emettitore: il guadagno β_R risulta piuttosto basso e le conseguenti prestazioni sono tanto limitate da rendere questa regione completamente priva di interesse applicativo. Per completezza il modello equivalente nella regione attiva inversa è comunque riportato nella figura 4.16.

Esempio 4.4 Si intende ricavare l'andamento delle concentrazioni dei portatori minoritari nel

transistore, in condizioni di saturazione. In particolare, per le regioni di emettitore, base e collettore, si ricava e si disegna in modo qualitativo la concentrazione dei minoritari con $V_{BE} > V_{BC}$ e $V_{BE} < V_{BC}$.

Nella regione di base, la concentrazione dei portatori minoritari in eccesso è data in generale dall'equazione (4.9), ovvero

$$n'(x) = -n'(0)\frac{\sinh\left(\frac{x-w_B}{L_{nB}}\right)}{\sinh\left(\frac{w_B}{L_{nB}}\right)} + n'(w_B)\frac{\sinh\left(\frac{x}{L_{nB}}\right)}{\sinh\left(\frac{w_B}{L_{nB}}\right)} \tag{4.16}$$

Le concentrazioni agli estremi della regione di base, $n'(0)$ e $n'(w_B)$, dipendono dalle cadute di tensione sulle giunzioni (legge della giunzione):

$$\begin{aligned} n'(0) &= n_{pB0}\left[\exp(V_{BE}/V_T) - 1\right] \\ n'(w_B) &= n_{pB0}\left[\exp(V_{BC}/V_T) - 1\right] \end{aligned}$$

Poiché $w_B \ll L_{nB}$ e inoltre $0 < x < w_B$, nell'equazione (4.16) i quattro sinh possono essere approssimati con i rispettivi argomenti, per semplificare l'espressione di $n'(x)$:

$$n'(x) = -n'(0)\frac{x - w_B}{w_B} + n'(w_B)\frac{x}{w_B}$$

L'equazione ottenuta rappresenta una retta passante per le quote $n'(0)$ in $x = 0$ e $n'(w_B)$ in $x = w_B$; la pendenza della retta è proporzionale alla corrente che transita in base. Pertanto, con $V_{BE} > V_{BC}$, si ha $n'(0) > n'(w_B)$, la pendenza della retta risulta negativa e la corrente scorre nella direzione dal collettore verso l'emettitore; con $V_{BE} < V_{BC}$, invece, la corrente ha verso opposto e $n'(0) < n'(w_B)$, come indicato nella figura 4.17.
Nelle regioni di emettitore e collettore, la concentrazione dei portatori minoritari è data dall'equazione (3.45), qui riportata per entrambe le regioni con la semplificazione che l'ampiezza della regione svuotata sia trascurabile, $x_p = 0$; l'espressione vale nell'ipotesi di "diodo lungo", ovvero con lunghezza fisica dei lati di emettitore e collettore molto superiore alla lunghezza di diffusione delle lacune:

$$\begin{aligned} p'_{nE}(x) &= p'_{nE}(0)\exp(x/L_{pE}) = p_{nE0}[\exp(V_{BE}/V_T) - 1]\exp(x/L_{pE}) \\ p'_{nC}(x) &= p'_{nC}(w_B)\exp[-(x - w_B)/L_{pC}] = p_{nC0}[\exp(V_{BC}/V_T) - 1]\exp[-(x - w_B)/L_{pC}] \end{aligned}$$

Entrambe le concentrazioni hanno andamento esponenziale decrescente e i valori alle giunzioni dipendono dalle cadute di tensione V_{BE} e V_{BC}.

4.6.3 Componenti di piccolo segnale

Nel caso in cui le correnti e le tensioni mostrino variazioni di piccola intensità intorno a valori costanti, il comportamento non lineare del transistore bipolare può essere approssimato con uno lineare, con il vantaggio di semplificare l'analisi circuitale attraverso l'uso di equazioni e modelli lineari. Si parla in questo caso di *condizioni di piccolo segnale.*

Le tre correnti misurate all'ingresso dei tre terminali del dispositivo si esprimono separando la componente costante e quella variabile di piccolo segnale:

$$i_C(t) = I_C + i_c(t) \tag{4.17}$$
$$i_E(t) = I_E + i_e(t) \tag{4.18}$$
$$i_B(t) = I_B + i_b(t) \tag{4.19}$$

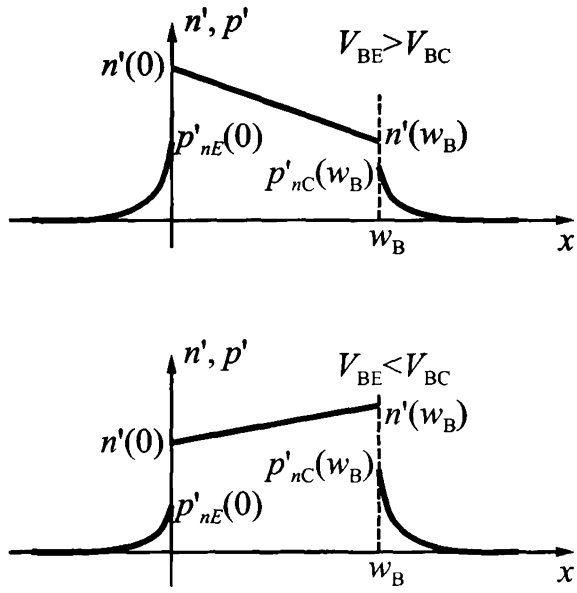


Figura 4.17 Andamento delle concentrazioni dei portatori minoritari nella regione di saturazione.

Nel caso del collettore, per esempio, I_C rappresenta la parte costante della corrente, $i_c(t)$ (con pedice c in minuscolo) è quella di piccolo segnale e $i_C(t)$ (con pedice C in maiuscolo) indica la corrente totale.

Anche per le tre differenze di potenziale misurabili tra coppie di terminali, si usa lo stesso tipo di notazione:

$$v_{BC}(t) = V_{BC} + v_{bc}(t) \tag{4.20}$$
$$v_{BE}(t) = V_{BE} + v_{be}(t) \tag{4.21}$$
$$v_{CE}(t) = V_{CE} + v_{ce}(t) \tag{4.22}$$

Ricavare un modello di piccolo segnale significa ottenere un modello approssimato del primo ordine che esprima il legame linearizzato tra le componenti di piccolo segnale delle correnti e delle tensioni. Ad esempio, dal modello di Ebers-Moll si può facilmente derivare un legame tra le correnti di collettore e di base e la tensione base-emettitore. in regione attiva diretta:

$$i_C = \alpha_F I_{ES} \exp\left(\frac{v_{BE}}{V_T}\right)$$
$$i_B = (1-\alpha_F) I_{ES} \exp\left(\frac{v_{BE}}{V_T}\right)$$

Poiché al transistore bipolare sono associate tre correnti e tre tensioni non indipendenti, si possono scegliere tra queste due correnti e due tensioni per descrivere

completamente lo stato del dispositivo. Scegliamo per esempio le correnti di base e collettore e le tensioni base-emettitore e collettore-emettitore:

$$i_B \quad , \quad i_C$$
$$v_{BE} \quad , \quad v_{CE}$$

Le correnti di base e collettore si esprimono in funzione delle tensioni v_{BE} e v_{CE}:

$$i_C = i_C(v_{BE}, v_{CE}) \tag{4.23}$$
$$i_B = i_B(v_{BE}, v_{CE}) \tag{4.24}$$

La corrente di emettitore e la tensione base-collettore si ricavano dalle leggi di Kirchhoff:

$$v_{CE} = v_{BE} - v_{BC}$$
$$i_E = -i_B - i_C$$

In condizioni di piccolo segnale, le espressioni delle correnti (4.24) si possono sviluppare al primo ordine intorno al punto di polarizzazione, corrispondente ai valori di corrente e tensione costanti definiti nelle equazioni (4.18)–(4.22):

$$i_C = i_C(V_{BE},V_{CE}) + v_{be} \cdot \left.\frac{\partial i_C}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} + v_{ce} \cdot \left.\frac{\partial i_C}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$$

$$i_B = i_B(V_{BE},V_{CE}) + v_{be} \cdot \left.\frac{\partial i_B}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} + v_{ce} \cdot \left.\frac{\partial i_B}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$$

Secondo il modello di Ebers-Moll, in regione attiva diretta la corrente di collettore è legata alla tensione base-emettitore da una semplice relazione esponenziale,

$$i_C = \alpha_F i_E = \alpha_F I_{ES} \exp\left(\frac{v_{BE}}{V_T}\right)$$

dalla quale si ricava facilmente il modello linearizzato:

$$i_C = i_C(V_{BE},V_{CE}) + v_{be} \cdot \left.\frac{\partial i_C}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} + v_{ce} \cdot \left.\frac{\partial i_C}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$$

I coefficienti di piccolo segnale si calcolano come

$$\left.\frac{\partial i_C}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} = \frac{\alpha_F I_{ES}}{V_T} \exp\left(\frac{V_{BE}}{V_T}\right) = \frac{I_C}{V_T} = \frac{\beta_0 I_B}{V_T}$$
$$\left.\frac{\partial i_C}{\partial v_{CE}}\right|_{V_{BE},V_{CE}} = 0$$

In modo analogo si può procedere per la corrente di base, legata anch'essa alla v_{BE} da una relazione esponenziale:

$$i_B = (1-\alpha_F)i_E = (1-\alpha_F)I_{ES}\exp\left(\frac{v_{BE}}{V_T}\right)$$

$$i_B = i_B(V_{BE},V_{CE}) + v_{be}\cdot\left.\frac{\partial i_B}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} + v_{ce}\cdot\left.\frac{\partial i_B}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$$

I coefficienti di piccolo segnale sono

$$\left.\frac{\partial i_B}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} = \frac{(1-\alpha_F)I_{ES}}{V_T}\exp\left(\frac{V_{BE}}{V_T}\right) = \frac{I_B}{V_T} = \frac{I_C}{\beta_0 V_T}$$

$$\left.\frac{\partial i_B}{\partial v_{CE}}\right|_{V_{BE},V_{CE}} = 0$$

Nei passaggi appena svolti, i coefficienti $\left.\frac{\partial i_C}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$ e $\left.\frac{\partial i_B}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$ risultano nulli soltanto perché il modello di Ebers-Moll è un modello semplificato, che non tiene conto dell'effetto Early. Per completare il modello di piccolo segnale, occorre includere nella trattazione l'effetto l'Early, che descrive la dipendenza della corrente di collettore dalla tensione collettore-emettitore.
Assumendo

$$\frac{\Delta i_C}{\Delta v_{CE}} = \frac{I_C}{V_A} \quad \text{e} \quad \Delta i_B = -\frac{\Delta i_C}{\beta_F}$$

si ottiene in forma approssimata

$$\left.\frac{\partial i_C}{\partial v_{CE}}\right|_{V_{BE},V_{CE}} = \frac{I_C}{V_A} \qquad \text{e} \qquad \left.\frac{\partial i_B}{\partial v_{CE}}\right|_{V_{BE},V_{CE}} = -\frac{I_C}{\beta_0 V_A} \approx 0$$

dove β_0 è il guadagno di corrente per piccolo segnale a emettitore comune (numericamente $\beta_F \approx \beta_0$)

Il modello ibrido a π

In generale, isolando le sole componenti di piccolo segnale, si possono esprimere in forma lineare le relazioni tra le correnti e le tensioni attraverso un sistema di equazioni del tipo:

$$i_b = y_{11}v_{be} + y_{12}v_{ce}$$
$$i_c = y_{21}v_{be} + y_{22}v_{ce}$$

dove i quattro parametri per le variazioni y hanno la dimensione di un'ammettenza e si ottengono come derivate parziali delle correnti rispetto alle tensioni, calcolate nel

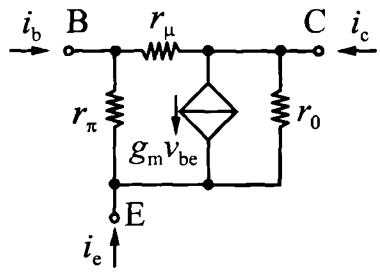


Figura 4.18 Modello a pi-greco.

punto di polarizzazione del dispositivo

$$y_{11} = \left.\frac{\partial i_B}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} \quad ; \quad y_{12} = \left.\frac{\partial i_B}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$$

$$y_{21} = \left.\frac{\partial i_C}{\partial v_{BE}}\right|_{V_{BE},V_{CE}} \quad ; \quad y_{22} = \left.\frac{\partial i_C}{\partial v_{CE}}\right|_{V_{BE},V_{CE}}$$

Il modello per le variazioni così ottenuto prende il nome di *modello ibrido a* π.

Con riferimento alla configurazione a emettitore comune del transistore, si può dare ai quattro parametri differenziali y un preciso significato circuitale:

- $1/y_{11} = r_\pi$ ha il significato di resistenza differenziale di ingresso, ovvero misurata alla coppia di terminali base-emettitore,
- $1/y_{12} = r_\mu \approx 0$ è una resistenza differenziale collocata tra i terminali di base e collettore,
- $y_{21} = g_m$ è la transconduttanza, che lega la corrente di collettore alla tensione base-emettitore,
- $1/y_{22} = r_0$ è infine la resistenza differenziale di uscita, valutatata tra i terminali di collettore e emettitore.

I parametri considerati permettono di definire un modello elettrico lineare equivalente, utile per l'analisi di circuiti a transistore. Il modello equivalente è dato nella figura 4.18.

Le relazioni per ottenere i parametri del modello si ricavano dalle definizioni date in precedenza:

$$r_\pi = \frac{\beta_0 V_T}{I_C} \qquad r_\mu \approx 0$$

$$g_m = \frac{I_C}{V_T} \qquad r_0 = \frac{V_A}{I_C}$$

Il modello è poi completato con la resistenza di base r_b, che tiene conto del comportamento ohmico del silicio attraversato da corrente.

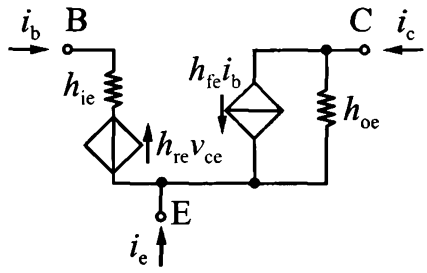


Figura 4.19 Modello a parametri h.

Modello a parametri h

Un modello alternativo, molto diffuso, per l'analisi di piccolo segnale del transistore bipolare è il *modello a parametri h*, rappresentato in figura 4.19

Per derivare i parametri del modello, applichiamo le leggi di Kirchhoff e scriviamo la corrente di collettore e la tensione base-emettitore come

$$i_c = h_{fe}\, i_b + h_{oe}\, v_{ce}$$

$$v_{be} = h_{ie}\, i_b + h_{re}\, v_{ce}$$

Il parametro h_{fe} esprime la dipendenza di i_c dalla corrente di base e si ottiene dalla derivata parziale di i_c rispetto a i_b, valutata nel punto di polarizzazione; poiché in regione attiva, il legame tra I_c e I_b è approssimativamente rettilineo, si ha

$$h_{fe} = \left.\frac{\partial i_C}{\partial i_B}\right|_{V_{CE0}} \simeq \beta_F$$

La corrente di collettore dipende anche dalla tensione di collettore, come descritto dall'effetto Early; pertanto il secondo contributo della corrente di collettore è proporzionale alla tensione collettore-emettitore

$$h_{oe} = \left.\frac{\partial i_C}{\partial v_{CE}}\right|_{I_{B0}} \simeq \frac{I_C}{V_A}$$

Il parametro h_{ie} è la resistenza differenziale della giunzione base-emettitore e si ottiene come

$$h_{ie} = \left.\frac{\partial v_{BE}}{\partial i_B}\right|_{V_{CE0}} = \frac{V_T}{I_B}$$

Infine, poiché l'effetto Early introduce una dipendenza della tensione base-emettitore, V_{BE}, da quella collettore-emettitore V_{CE}, nel modello si ha ancora un termine $h_{re}v_{CE}$ e il parametro h_{re} si ottiene imponendo $i_B = 0$ e studiando la dipendenza di i_C da

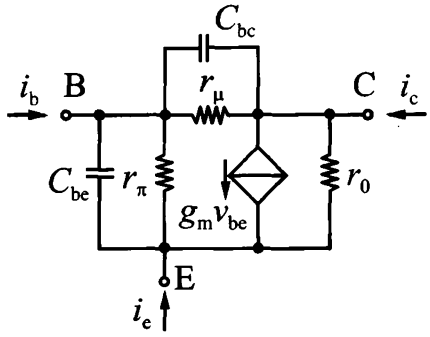


Figura 4.20 Modello per lo studio in frequenza.

v_{CE}:

$$h_{re} = \left.\frac{\partial v_{BE}}{\partial v_{CE}}\right|_{i_{B0}} = \frac{V_T}{V_A}$$

4.6.4 Comportamento in frequenza

Il modello circuitale sviluppato nei paragrafi precedenti per l'analisi delle variazioni in un transistore bipolare descrive soltanto il comportamento a bassa frequenza. Quando i segnali di corrente e tensione in gioco hanno componenti significative a frequenza elevata, il comportamento del dispositivo è influenzato anche dalle capacità localizzate alle giunzioni. Come si è visto nel paragrafo 3.4, ogni giunzione presenta due contributi di capacità, quella di svuotamento, C_s, legata alla carica accumulata a cavallo della giunzione e dovuta a cariche fisse, e la capacità di diffusione, C_d, associata ai portatori mobili iniettati nelle regioni quasi neutre; entrambe le capacità dipendono dalla tensione applicata alla giunzione e tipicamente C_d tende a prevalere in polarizzazione diretta, mentre C_s è superiore in polarizzazione inversa.

Il modello circuitale equivalente deve quindi essere completato con opportuni parametri capacitivi, in grado di descrivere l'andamento di tensioni e correnti al variare della frequenza. In figura 4.20 è dato un modello per l'analisi in frequenza, ottenuto completando il modello ibrido a π con la capacità delle giunzioni base-emettitore e base-collettore.

In regione attiva diretta, la capacità prevalente della giunzione base-emettitore è la capacità di diffusione, proporzionale alla corrente di base; poiché la giunzione base-collettore è polarizzata inversamente, la capacità più significativa in questo caso è invece la capacità di svuotamento A frequenza elevate, le due capacità tendono a cortocircuitare le giunzioni e quindi il guadagno diminuisce.

Per valutare le prestazioni ad alta frequenza, si calcola il guadagno di corrente di corto circuito: si chiude l'uscita su un corto circuito, si applica una corrente alla base e si misura la corrente sul corto circuito (terminale di collettore). Il guadagno di corrente

di corto circuito è il rapporto $\frac{i_c}{i_b}$. In questo caso si ha:

$$\frac{i_c}{i_b} = \beta(f) = \frac{\beta_0}{1 + j\frac{f}{f_0}}$$

dove $\beta_0 = g_m r_\pi$ e $f_0 = \frac{1}{2\pi r_\pi (C_{be}+C_{bc})}$ (frequenza di taglio a 3 dB)

A frequenze nettamente superiori a f_0, $f \gg f_0$, il guadagno può essere espresso come

$$\beta(f) \approx -j\beta_0 \frac{f_0}{f}$$

Si definisce *frequenza di taglio* f_T il valore di f per il quale il modulo di $\beta(f)$ si riduce a 1

$$|\beta(f)| = 1 \rightarrow f_T = \beta_0 f_0$$

f_T è quindi pari al prodotto della banda del transistore e del guadagno in continua. Nella figura 4.21 è mostrato l'andameno del guadagno β in funzione della frequenza f: il passaggio per $\beta = 1$ identifica la frequenza di taglio.

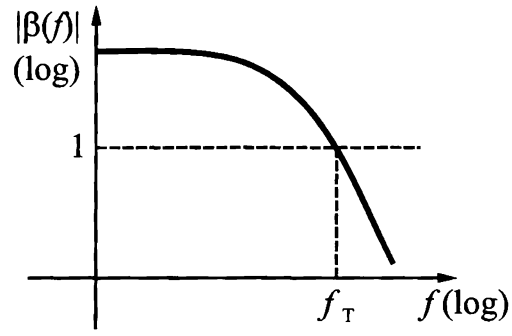


Figura 4.21 Frequenza di taglio del transistore bipolare.

La tabella 4.1 riassume l'evoluzione del transistore bipolare attraverso i valori tipici di alcuni parametri significativi in diverse generazioni tecnologiche. In particolare si evidenziano la crescita della frequenza di taglio e la corrispondente diminuzione dei ritardi di propagazione in una famiglia logica bipolare; tali miglioramenti accompagnano il progressivo scalamento delle dimensioni in termini di larghezza di base e di emettitore.

Parametri come quelli riassunti nella tabella 4.1 sono tipicamente riportati dai costruttori su fogli tecnici di documentazione, detti *data sheet*, che raccolgono le ca-

parametro	1980	1985	1990
larghezza di emettitore (μm)	3	1,5	0,8
larghezza di base (μm)	0,3	0,15	0,07
f_T (GHz)	1	10	30
ECL gate delay (ps)	500	100	30

Tabella 4.1 Evoluzione della tecnologia bipolare.

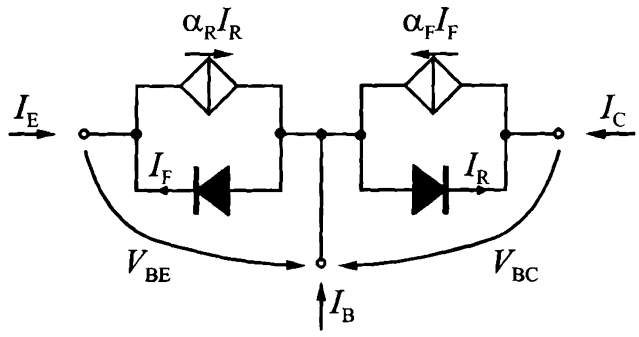


Figura 4.22 Modello di Ebers Moll.

ratteristiche tecniche di tipo elettrico, meccanico, termico e funzionale di componenti, circuiti e sistemi integrati.

Esempio 4.5 Usando il modello di Ebers-Moll, si ricavi l'espressione della caduta di tensione ai capi di un transistore in saturazione. Si calcoli inoltre $V_{CE\,sat}$ per $I_C/I_B = 10, \alpha_F = 0{,}985$ e $\alpha_R = 0{,}72$ (si assuma $\alpha_F I_{ES} = \alpha_R I_{CS}$).
Le equazioni del modello di Ebers-Moll sono:

$$\begin{cases} I_E = \alpha_R I_R - I_F \\ I_C = \alpha_F I_F - I_R \end{cases} \quad \text{con} \quad \begin{cases} I_R = I_{CS}\,(e^{V_{BC}/V_T} - 1) \\ I_F = I_{ES}\,(e^{V_{BE}/V_T} - 1) \end{cases}$$

Entrambe le giunzioni sono polarizzate direttamente, quindi le correnti si possono approssimare come:

$$I_R \simeq I_{CS}\, e^{V_{BC}/V_T} \quad ; \quad I_F \simeq I_{ES}\, e^{V_{BE}/V_T}$$

Si ricavano le tensioni

$$V_{BC} = V_T \ln \frac{I_R}{I_{CS}} \quad ; \quad V_{BE} = V_T \ln \frac{I_F}{I_{ES}}$$

e per differenza la tensione V_{CE} di saturazione:

$$V_{CE_{sat}} = V_{BE} - V_{BC} = V_T \ln \left(\frac{I_F}{I_{ES}} \frac{I_{CS}}{I_R} \right)$$

Per ottenere un'espressione utile, si devono calcolare i rapporti I_F/I_R e I_{CS}/I_{ES}.

La corrente di base è

$$I_B = -I_E - I_C = (1-\alpha_F)\, I_F + (1-\alpha_R)\, I_R$$

con $I_R = \alpha_F\, I_F - I_C$.
Si sostituisce $I_R = \alpha_F\, I_F - I_C$ e si ricava I_F:

$$\begin{aligned} I_B &= (1-\alpha_F)\, I_F + (1-\alpha_R)\, \alpha_F\, I_F - (1-\alpha_R)\, I_C \\ I_F &= \frac{1}{(1-\alpha_F\, \alpha_R)}\, [I_B + (1-\alpha_R)\, I_C] \end{aligned}$$

$$\begin{aligned} I_F &= \frac{1}{(1-\alpha_F\, \alpha_R)}\, [I_B + (1-\alpha_R)\, I_C] \\ I_R &= \alpha_F\, I_F - I_C \end{aligned}$$

Si sostituisce ora I_F in I_R:

$$\begin{aligned} I_R &= \alpha_F\, \frac{1}{1-\alpha_F\, \alpha_R}\, [I_B + (1-\alpha_R)\, I_C] - I_C \\ &= \frac{\alpha_F}{1-\alpha_F\, \alpha_R}\, I_B - \frac{1-\alpha_F}{1-\alpha_F\, \alpha_R}\, I_C \end{aligned}$$

e, per sostituzione diretta, si ricava il rapporto I_F/I_R:

$$\frac{I_F}{I_R} == \frac{\alpha_R}{\alpha_F}\, \frac{\frac{1}{\alpha_R} + \frac{1}{\beta_R}\, \frac{I_C}{I_B}}{1 - \frac{1}{\beta_F}\, \frac{I_C}{I_B}}$$

Infine, dalla definizione di I_{CS} e I_{ES} e dalla condizione $\alpha_F I_{ES} = \alpha_R I_{CS}$, si ha

$$\frac{I_{CS}}{I_{ES}} = \frac{\alpha_R\, I_{CS}}{\alpha_R\, I_{ES}} = \frac{\alpha_F\, I_{ES}}{\alpha_R\, I_{ES}} = \frac{\alpha_F}{\alpha_R}$$

$$V_{CE_{sat}} = V_T\, \ln\left(\frac{I_{CS}}{I_{ES}}\, \frac{I_F}{I_R}\right) = V_T\, \ln\left[\frac{\frac{1}{\alpha_R} + \frac{1}{\beta_R}\, \frac{I_C}{I_B}}{1 - \frac{1}{\beta_F}\, \frac{I_C}{I_B}}\right]$$

Sostituendo i numeri dell'esempio, si ottiene:

$$V_{CE_{sat}} = 0{,}048\, V$$

Il valore reale di $V_{CE_{sat}}$ dipende dalla polarizzazione, dalla tecnologia e dalla temperatura. Tipicamente da misure sperimentali si trovano valori tra 0,1 e 0,2 V: la differenza rispetto alla $V_{CE_{sat}}$ calcolata è principalmente dovuta alle cadute di potenziale sulle regioni quasi neutre, che non sono state incluse nel modello.

Capitolo 5

Il transistore MOS

In questo capitolo si descrivono la struttura e il funzionamento del transistore MOSFET, il cui nome deriva dall'acronimo dell' inglese Metal–Oxide–Semiconductor Field Effect Transistor, ovvero transistore metallo–ossido–semiconduttore a effetto di campo. Le caratteristiche di questo transistore ne fanno attualmente il dispositivo più utilizzato nella fabbricazione dei circuiti integrati su silicio: da questo si intuisce l'importanza di comprenderne correttamente il funzionamento e l'utilizzo. Poiché, però, si tratta di un dispositivo dalla struttura e dal funzionamento complesso, è necessario articolare il suo studio per gradi successivi, a partire da strutture più semplici ma di più immediata comprensione. Nel paragrafo 5.1 si analizza quindi per prima cosa il cosiddetto sistema MOS (Metal–Oxide–Semiconductor), che costituisce una sottoparte del transistore MOSFET e il cui studio risulta propedeutico alla comprensione del funzionamento del transistore stesso. Si passa quindi alla analisi del transistore MOSFET vero e proprio, paragrafi 5.2–5.5, ricavandone le caratteristiche statiche tensione–corrente. Nel paragrafo 5.6 vengono poi esaminati i principali effetti di non idealità mentre il ruolo del terminale di substrato è esaminato nel paragrafo 5.7. Vengono infine descritti il circuito equivalente di ampio segnale statico e il circuito equivalente di piccolo segnale nel paragrafo 5.8.

5.1 Il sistema MOS

Il sistema MOS è formato dalla giunzione di tre strati successivi di materiali diversi, ovvero

- ▷ metallo (M)
- ▷ ossido (OX)
- ▷ semiconduttore (S)

Il sistema MOS è formato a partire da un substrato di materiale semiconduttore, per deposizione successiva dell'ossido e del metallo. Se il drogaggio del semiconduttore è di tipo p, si parla di sistema MOS *su substrato di tipo p*, altrimenti di sistema MOS *su substrato di tipo n*, come mostrato nella figura 5.1.

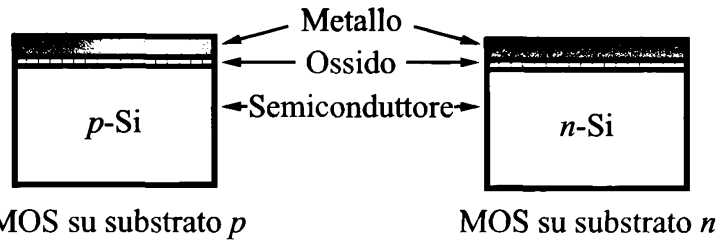


Figura 5.1 Struttura del sistema MOS su substrato p (a sinistra) e su substrato n (a destra). Si è ipotizzato di utilizzare il silicio come semiconduttore.

L'ossido e il metallo utilizzati variano a seconda del semiconduttore impiegato. In questo testo ci limiteremo a considerare i sistemi MOS fabbricati su substrato di silicio poiché questa rappresenta la tecnologia di gran lunga più diffusa. In questo caso, l'ossido è di norma costituito da ossido di silicio (SiO_2) e, come si spiegherà in seguito, dal punto di vista tecnologico è fattore di merito del sistema MOS che l'ossido sia il più sottile possibile, tanto che i processi di fabbricazione MOS sono progrediti negli anni fino a ridurne lo spessore a pochi nanometri. Nelle prime tecnologie MOS degli anni '70 per il metallo si è utilizzato alluminio, anche se a partire dagli anni '80 si è affermata una nuova tecnologia in cui il questo viene sostituito da un materiale non metallico ma ugualmente con buona conducibilità elettrica: il polisilicio. Questo materiale, detto anche *poly* dall'inglese polycristal silicon o polysilicon, è un policristallo formato da grani di silicio cristallino disallineati e, quando pesantemente drogato di tipo n, presenta una buona conducibilità elettrica; a differenza di un metallo, però, ha anche una buona resitenza meccanica ai cicli termici ad alta temperatura necessari nella realizzazione dei circuiti integrati a semiconduttore, caratteristica che ne ha determinato il crescente utilizzo.

Da quanto detto, risulta evidente che nella figura 5.1, gli spessori dei tre materiali non sono rappresentati in scala, anzi, il substrato ha spessore di gran lunga superiore rispetto agli altri due materiali.

Introducendo una seconda metallizzazione collegata al semiconduttore che forma il substrato del sistema MOS, si ottiene un dispositivo a due terminali, detto anche condensatore MOS. Il contatto formato dal metallo deposto al di sopra dell'ossido prende il nome di contatto di *gate* mentre l'altro è detto contatto di *substrato* o, in inglese, contatto di *bulk* o contatto di *body*. Il sistema MOS risultante è mostrato nella figura 5.2.

La struttura in esame è per costruzione una struttura planare a strati, e in prima approssimazione consideremo che i tre materiali con cui essa è formata siano uniformi nei rispettivi strati. In questo caso le uniche variazioni si hanno nella direzione perpendicolare alla interfaccia tra i diversi strati, mentre la struttura rimane uniforme nella sezione A del dispositivo, definita come il prodotto della larghezza W e della lunghezza L, come mostrato nella figura 5.3.

In questo caso è sufficiente restringere l'analisi del sistema MOS alla sola direzione perpendicolare alle superfici di interfaccia (eterogiunzioni) tra i vari materiali. Si definisca quindi l'asse y diretto nella direzione perpendicolare a tali interfacce, con verso dal metallo di gate al substrato, come mostrato nella figura 5.4. L'origine viene definita

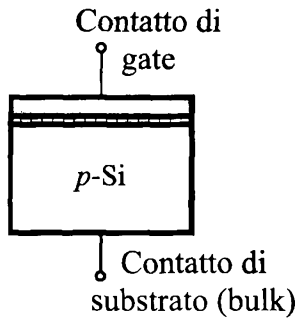


Figura 5.2 Contatti del sistema MOS su substrato *p*.

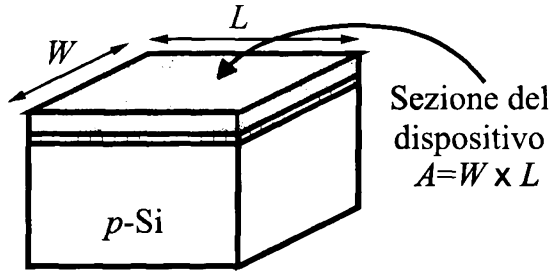


Figura 5.3 Sezione del sistema MOS su substrato *p*.

in corrispondenza del punto di interfaccia tra l'ossido e il substrato di semiconduttore (OX-S), mentre l'interfaccia tra il metallo e l'ossido (M-OX) si trova alla coordinata $y = -t_{ox}$, rimanendo pertanto definita la quantità t_{ox} pari allo spessore dello strato di ossido. Il substrato ha spessore complessivo pari a W_p mentre il metallo si estende nell'intervallo $-t_{\mathrm{M}} < y < -t_{ox}$.

5.1.1 Regioni di funzionamento del sistema MOS

Per analizzare il comportamento del sistema MOS ci si concentra in un primo momento sul sistema MOS su substrato di tipo *p*. Il sistema MOS su substrato *n* verrà analizzato nel paragrafo 5.1.7. Il sistema a MOS si può in prima approssimazione paragonare ad un condensatore a facce piane e parallele: uno strato di materiale dielettrico (o isolante – l'ossido di silicio nel nostro caso) si trova tra due "armature", l'una di tipo convenzionale costituita dal metallo del contatto di gate e l'altra, non convenzionale, costituita non già da uno strato metallico, ma da uno strato di materiale semiconduttore drogato (il substrato). La analogia con un condensatore convenzionale è particolarmente utile, poiché permette di effettuare una prima analisi qualitativa delle diverse regioni di funzionamento del sistema MOS in funzione della tensione applicata ai terminali. Si osservi peraltro che ovviamente esiste una sola tensione di controllo, ovvero la tensione tra le due armature, il gate e il substrato e, per semplificare il ragionamento, si può anche supporre di collegare il contatto di substrato al riferimento di potenziale (massa) e di applicare una tensione V_{G} al contatto di gate.

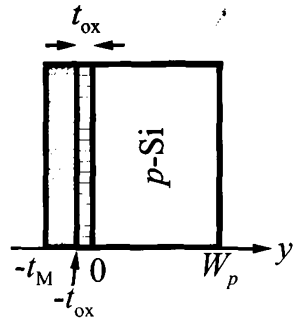


Figura 5.4 Definizione dell'asse x nel sistema MOS su substrato p.

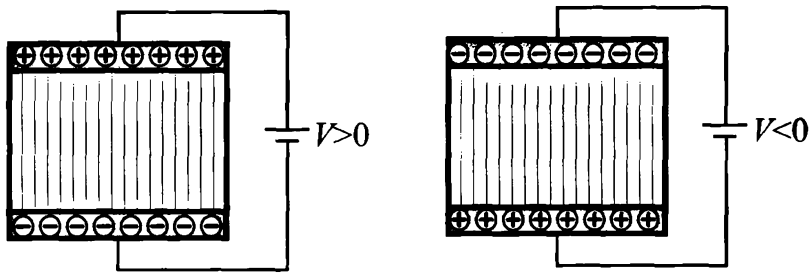


Figura 5.5 Regioni di funzionamento di un convenzionale condensatore a facce piane e parallele.

Richiamiamo prima i punti fondamentali relativi al condensatore convenzionale[1] per passare poi al MOS per analogia. Caratteristica di un condensatore è la capacità di variare la carica sulle armature mediante la tensione V applicata agli elettrodi. Una volta accumulata la carica, in condizioni strazionarie la corrente ai terminali è sempre nulla poiché essi sono isolati dal dielettrico che, idealmente, non supporta corrente di conduzione ma solo corrente di spostamento dielettrico. In funzione della tensione V applicata tra le due armature del condensatore si hanno due regioni di funzionamento (vedi figura 5.5):

- $V > 0$ la carica è positiva sulla armatura superiore e negativa su quella inferiore
- $V < 0$ la carica è negativa sulla armatura superiore e positiva su quella inferiore

Il campo elettrico è diretto sempre dalla carica positiva a quella negativa, ovvero dal potenziale superiore a quello inferiore. La transizione tra una regione di funzionamento e l'altra avviene in assenza di tensione applicata dall'esterno ($V = 0$), ovvero in equilibrio termodinamico, e in questo caso la carica sulle armature è nulla.

Analogamente al condensatore convenzionale, anche il sistema MOS è caratterizzato dal fatto che la carica sulle armature varia in funzione della tensione V_G applicata ai

[1] Indicheremo nel seguito con condensatore convenzionale un condensatore con entrambe le armature metalliche.

terminali mentre la corrente statica è nulla. Nonostante le analogie, il condensatore MOS presenta però alcune notevoli differenze rispetto al condensatore convenzionale, legate proprio al fatto che una armatura è costituita da un materiale semiconduttore invece che da un metallo. Le differenze principali sono:

- in condizioni di equilibrio le armature sono cariche e tra di esse è presente una differenza di potenziale (d.d.p.). Per annullare la carica delle armature è necessario invece applicare dall'esterno una tensione V_G non nulla e tale tensione è detta *tensione di banda piatta* V_{FB}[2]
- nel sistema MOS si hanno tre, e non due, regioni di funzionamento al variare di V_G.

Le affermazioni precedenti meritano una attenta analisi. A partire dalla prima, si può in altre parole dire che all'equilibrio termodinamico è presente tra i terminali del sistema MOS una d.d.p. non nulla esattamente pari a $-V_{FB}$ sostenuta dalla presenza di una carica non nulla. Applicando invece dall'esterno una tensione uguale e opposta alla d.d.p. presente all'equilibrio, ovvero per $V_G = V_{FB}$, la caduta di tensione complessiva presente tra le armature si annulla e, di conseguenza, si annulla anche la carica elettrica. Il fatto che nella struttura esista, in condizioni di equilibrio termodinamico, una differenza di potenziale non nulla tra il contatto del metallo e quello del substrato, ricorda quanto avviene per esempio in una giunzione *pn*, dove all'equilibrio è presente una tensione interna (o tensione di *built-in*) non nulla. La tensione di banda piatta piatta può infatti venire interpretata come la tensione interna del sistema MOS in equilibrio. Si ricordi, però, che in una giunzione *pn* la tensione esterna non è mai in grado di annullare completamente la tensione interna di *built-in* a causa delle cadute ohmiche che si instaurano sulle regioni resistive quando nella giunzione scorre corrente. Nel sistema MOS, invece, anche applicando una tensione (continua) ai terminali, la corrente è sempre identicamente nulla e la tensione interna può venire quindi completamente annullata.

Poiché per $V_G = V_{FB}$ si ha la condizione di carica e d.d.p. nulla nel sistema MOS, si intuisce che variando la tensione esterna al di sopra o al di sotto di V_{FB} si avrà una carica diversa da zero sulle armature. Come anticipato, si hanno *tre regioni di funzionamento*:

- $V_G < V_{FB}$ accumulo di lacune
- $V_G > V_{FB}$ svuotamento di lacune
- $V_G \gg V_{FB}$ inversione di popolazione

Mentre le prime due regioni di funzionamento corrispondono approssimativamente a quelle di un condensatore convenzionale, la regione di inversione di popolazione è invece tipica dei condensatori MOS e, proprio per questo, riveste particolare interesse nel funzionamento dei transistori MOSFET. Si osservi anche che la tensione di banda piatta di un sistema MOS su substrato di tipo *p* risulta normalmente negativa e, di conseguenza, la *condizione di equilibrio termodinamico* $V_G = 0$ *viene a verificarsi in corrispondenza della regione di svuotamento.*

[2] In inglese *flat band voltage*, da cui l'acronimo FB utilizzato nel pedice.

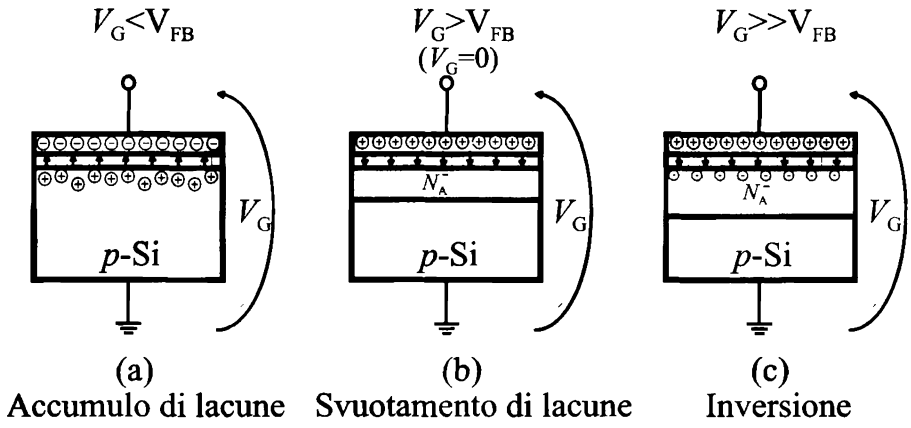


Figura 5.6 Regioni di funzionamento di un condensatore MOS.

Nella regione di accumulo di lacune, la tensione applicata al gate è negativa, e di conseguenza l'armatura del gate si carica negativamente. Dalla parte del semiconduttore le lacune sono attirate verso l'interfaccia tra l'ossido e il semiconduttore, dove si accumulano formando uno strato di carica positiva che costituisce la carica della armatura inferiore. Il campo elettrico è diretto dal semiconduttore al metallo e sostenuto in gran parte dallo strato di dielettrico (isolante). La situazione è illustrata nella figura 5.6 (a).

Nella regione di svuotamento la tensione applicata al gate è positiva, e di conseguenza l'armatura del gate si carica positivamente[3]. Dalla parte del semiconduttore le lacune sono respinte dall'interfaccia tra l'ossido e il semiconduttore, dove si crea quindi una regione *svuotata di lacune*: questa è carica per la presenza degli accettatori ionizzati negativamente. Lo strato svuotato forma anche la carica negativa dell'armatura inferiore del condensatore MOS. Il campo elettrico è diretto dal metallo al semiconduttore e sostenuto in parte dallo strato di dielettrico (isolante) e in parte dalla regione svuotata di semiconduttore (si ricordi infatti che a differenza di un metallo, il semiconduttore non è equipotenziale e, in particolare, presenta una d.d.p. non nulla in presenza di una regione svuotata carica). La situazione è illustrata nella figura 5.6 (b).

Si supponga poi di aumentare la tensione di gate fino a valori molto positivi raggiungendo la regione di inversione di popolazione (figura 5.6 (c)). Nel semiconduttore si ha un fenomeno nuovo: all'interfaccia tra l'ossido e il semiconduttore si forma *sottile strato di elettroni liberi.* Si dice che si ha, in un sottile strato posto corrispondenza della interfaccia tra l'ossido e il semiconduttore, una condizione di *inversione di popolazione*: mentre infatti nel semiconduttore drogato di tipo p le lacune sono normalmente maggioritarie, in condizione di inversione sono gli elettroni ad esserlo. Si osservi che in questo caso la carica negativa presente sul semiconduttore ha due componenti: una è costituita dalla *carica fissa* dovuta agli accettatori ionizzati presenti nello strato di semiconduttore svuotato; l'altra è costituita dalla *carica mobile* degli elettroni nello strato di inversione di popolazione. Gli elettroni, poi, sono attirati verso l'interfaccia

[3] Si ricordi che in un metallo si può avere sia carica positiva (lacune) sia negativa (elettroni).

OX-S dal campo elettrico stesso tanto che lo strato di inversione è estremamente sottile e concentrato: in molti casi esso viene considerato come uno strato di carica puramente superficiale, ovvero con spessore virtualmente nullo. La carica di elettroni presente nello strato inversione viene sfruttata, nel transisitore MOSFET, per consentire il passaggio di una corrente elettrica ed è quindi importante determinare in che modo essa dipenda dalla tensione applicata: si dimostrerà nei paragrafi successivi che la carica di elettroni cresce al crescere della tensione applicata al gate V_G seguendo una importante relazione, detta *legge di controllo di carica.*

5.1.2 Il sistema MOS all'equilibrio

Come anticipato nel paragrafo precedente, il sistema MOS su substrato di tipo p all'equilibrio si trova nella condizione di svuotamento di lacune. Per giustificare questa affermazione, in questo paragrafo si effettua lo studio dettagliato del sistema MOS all'equilibrio, determinando in maniera esatta l'andamento della carica, del campo elettrico, del potenziale elettrostatico e disegnando infine il diagramma a bande.

Per tracciare il diagramma a bande all'equilibrio si seguono le regole già viste nel caso della giunzione *pn*. Ricordiamo che, in estrema sintesi, esse sono:

- ▷ il livello di Fermi è costante nella struttura
- ▷ l'affinità elettronica e l'ampiezza della banda proibita sono costanti per ogni materiale
- ▷ lontano dalle giunzioni, la struttura a bande torna ad essere quella del materiale isolato
- ▷ il livello del vuoto E_0 è continuo

Poiché poi siamo ora in presenza di una *eterogiunzione*, ovvero la giunzione tra materiali diversi, si osserveranno nel diagramma a bande del sistema MOS alcune particolarità che verranno discusse nel seguito.

Procedendo come già visto nel caso della giunzione *pn*, si disegni per prima cosa il *diagramma a bande dei tre materiali separatamente* prendendo come riferimento per le energie il livello di vuoto E_0, come mostrato nella figura 5.7. Ricordiamo che con la scelta effettuata per l'asse x (figura 5.4), si ha a sinistra il metallo, al centro l'ossido e a destra il substrato di materiale semiconduttore drogato di tipo p. Si osservi nella figura 5.7 che i tre materiali presentano struttura a bande molto diversa tra loro.

Nel metallo (figura 5.7 a sinistra) non si distinguono banda di valenza o banda di conduzione, poiché il metallo è caratterizzato da una sola banda, la banda di conduzione, riempita di elettroni fino all'energia del livello di Fermi e vuota al di sopra di esso. In questo senso il metallo ha sia una concentrazione molto elevata di elettroni con energia inferiore al livello di Fermi, sia una concentrazione molto elevata di posti vuoti[4] al di sopra del livello di Fermi. Nel diagramma a bande è quindi sufficiente identificare la posizione del livello di Fermi E_F in relazione al riferimento E_0. La distanza tra E_F e E_0 è detta *lavoro di estrazione* ed è indicata con $q\Phi_M$[5].

Nell'ossido (figura 5.7 al centro) la banda di valenza e la banda di conduzione sono separate dalla banda proibita caratterizzata da una ampiezza $E_{g,OX}$ molto elevata. La

[4] Eventualmente interpretabili come lacune, anche se questa identificazione non viene comunemente utilizzata per i metalli.

[5] Si osservi che questa energia corrisponde al prodotto della carica elementare per il *potenziale di estrazione* Φ_M.

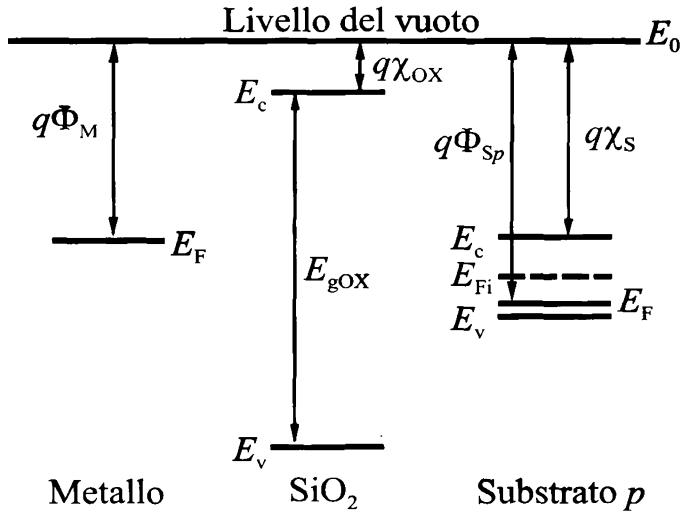


Figura 5.7 Costruzione del diagramma a bande all'equilibrio del sistema MOS su substrato di tipo p.

banda di valenza è completamente piena mentre quella di conduzione è vuota. Infatti un isolante è un materiale che, anche a temperatura ambiente, ha concentrazione intrinseca così bassa da essere considerata trascurabile. La distanza tra il bordo inferiore della banda di conduzione E_c e il riferimento energetico E_0 è la *affinità elettronica dell'ossido*, $q\chi_{OX}$. Nell'ossido il livello di Fermi si trova all'interno della banda proibita, ma qui non lo tracciamo esplicitamente poiché, come si vedrà in seguito, esso è fissato una volta che E_F sia noto sia nel metallo che nel semiconduttore.

Nel semiconduttore (figura 5.7 a destra) la struttura a bande è quella già nota per un semiconduttore drogato di tipo p, caratterizzata dalla affinità elettronica $q\chi_S$, dal lavoro di estrazione $q\Phi_S$ e dalla ampiezza di banda proibita E_g. Si ricordi che nel semiconduttore drogato p con concentrazione di accettatori N_A, la posizione del livello di Fermi e, di conseguenza, il lavoro di estrazione, dipendono dal livello di drogaggio. Infatti:

$$q\Phi_{Sp} = q\chi_S + E_g - (E_F - E_v) = q\chi_S + E_g - k_B T \ln \frac{N_v}{N_A} \tag{5.1}$$

Alla formazione della giunzione, si ha un trasferimento di cariche, tale da garantire che, passato un transitorio, il livello di Fermi sia allineato in tutti e tre i materiali. Poiché il livello di Fermi nel metallo è ad una energia maggiore di quello del semiconduttore[6], si ha un trasferimento di elettroni dal metallo al semiconduttore. Il metallo si carica quindi positivamente per effetto dello spopolamento di elettroni. Questi raggiungendo il semiconduttore vengono ad occupare i posti liberi nella banda di valenza. ovvero nel semiconduttore si forma uno strato svuotato di lacune e carico negativa-

[6] Questa situazione si verifica sia nel caso di contatto di gate in alluminio sia in *poly* n^+, cfr. il successivo esempio 5.1.

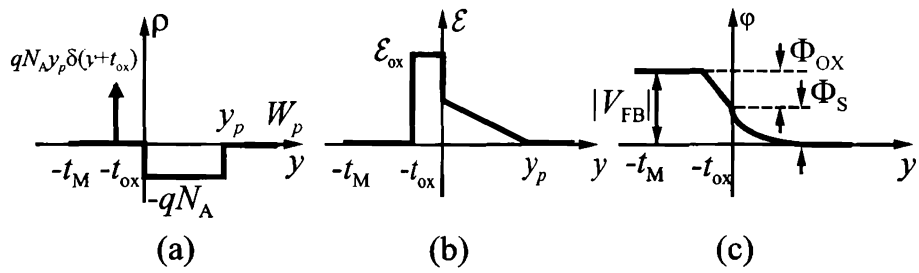


Figura 5.8 Andamento della concentrazione di carica ρ (a), del campo elettrico $\mathcal{E}$ (b) e del potenziale elettrostatico φ (c) all'interno del sistema MOS all'equilibrio.

mente per effetto della presenza degli accettatori ionizzati carichi negativamente[7]. La situazione che si determina è quindi in effetti del tutto analoga a quanto visto per la condizione di svuotamento di lacune nel paragrafo precendente (vedi anche la figura 5.6 (b)). Ovviamente il passaggio di carica dal metallo al semiconduttore non può avvenire direttamente attraverso l'ossido (se non in tempi molto lunghi): l'ossido infatti è un isolante e non permette passaggio di carica[8]. In realtà l'equilibrio viene raggiunto in tempi molto rapidi poiché nel processo di fabbricazione della giunzione stessa si vengono a creare, almeno temporaneamente (per esempio nella deposizione delle metallizzazioni), dei cammini ohmici tra il semiconduttore e il metallo che consento un rapido scambio di cariche tra questi.

In definitiva, la figura 5.8 (a) rappresenta la densità di carica all'interno del sistema MOS all'equilibrio. Nel semiconduttore si ha uno strato di materiale svuotato che si estende dalla coordinata $y = 0$ in corrispondenza della interfaccia con l'ossido fino alla coordinata y_p: in questa regione si ha una concentrazione di carica negativa pari a $-qN_A$ dovuta agli accettatori ionizzati. Al di fuori della regione svuotata ($y \geq y_p$) il semiconduttore è neutro e la densità di carica è nulla. Nell'ossido, isolante, la densità di carica è idealmente nulla. Nel metallo si forma uno strato superficiale di carica positiva posto all'interfaccia tra il metallo e l'ossido: si ricordi infatti che un metallo è idealmente un materiale equipotenziale (campo elettrico nullo) e non può quindi aveve alcuna carica al suo interno. Se quindi un metallo è carico, la carica si distribuisce solo sulla superficie. Nella figura 5.8 (a), la carica nel metallo è quindi rappresentata mediante una funzione δ di Dirac, centrata nel punto di interfaccia $y = -t_{ox}$. Poiché siamo in condizioni di equilibrio termodinamico, la carica complessiva presente all'interno del sistema deve essere comunuqe nulla, altrimenti si avrebbe al netto un campo elettrico diverso da zero al di fuori della struttura pur in assenza di forze applicate dall'esterno. Perché il sistema risulti *globalmente neutro*, la carica nel metallo deve essere uguale e opposta a quella nel semiconduttore.

Per effetto della distribuzione di carica elettrica, si avranno all'interno del sistema MOS un campo elettrico e un salto di potenziale non nulli. L'andamento del campo

[7] Si noti che il passaggio di elettroni dal metallo al semiconduttore si può anche interpretare come un passaggio di lacune dal semiconduttore al metallo.

[8] Dal punto di vista energetico questo fatto si può anche comprendere osservando il diagramma a bande: gli elettroni del metallo si trovano a dover superare una *barriera di potenziale* pari a $q\chi_{OX} - q\Phi_M$ per poter passare nella banda di conduzione (vuota) dell'ossido e raggiungere infine il semiconduttore.

elettrico si ottiene integrando una volta l'equazione di Gauss:

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}y} = \frac{\rho}{\epsilon} \tag{5.2}$$

dove ϵ è la costante dielettrica del materiale (semiconduttore, ossido o metallo). L'integrale si può effettuare a tratti nelle tre regioni separatamente applicando poi opportune condizioni di raccordo, come noto dai corsi di Fisica. Il campo elettrico risultante ha l'andamento mostrato nella figura 5.8 (b). Si osservi che per $y < -t_{\mathrm{ox}}$ e $y > y_p$ il sistema è neutro e anche a campo zero: il campo elettrico è infatti racchiuso all'interno del doppio strato di cariche uguali e opposte ai due lati dell'ossido. All'interno dell'ossido la carica è nulla ma non così il campo elettrico, che assume un valore $\mathcal{E}_{\mathrm{ox}}$ costante diverso da zero per la presenza delle cariche ai bordi. Questa situazione è peraltro del tutto analoga a quanto avviene in un condensatore convenzionale. Nella regione svuotata di semiconduttore $0 < y < y_p$ il campo è rettilineo con pendenza negativa, in analogia a quanto visto nel lato p della giunzione pn. Il campo elettrico presenta inoltre una discontinuità in corrispondenza del punto $y = -t_{\mathrm{ox}}$ per la presenza dello stato superficiale di carica nel metallo e anche in $y = 0$ per effetto della differenza della costante dielettrica tra l'ossido e il semiconduttore[9].

Infine si ricava l'andamento del potenziale elettrostatico integrando rispetto ad x il campo elettrico cambiato di segno:

$$\frac{\mathrm{d}\varphi}{\mathrm{d}y} = -\mathcal{E} \tag{5.3}$$

ottenedo l'andamento mostrato nella figura 5.8 (c). Il riferimento di potenziale è stato scelto in corrispondenza del lato neutro di semiconduttore per $y > y_p$. All'interno della regione carica di semiconduttore il potenziale presenta un andamento parabolico con concavità positiva. Nell'ossido (neutro) il potenziale è lineare mentre nel metallo (equipotenziale) esso è costante. Si osservi che il potenziale elettrostatico, a differenza del campo elettrico, è una funzione continua. Il salto di potenziale totale presente sul sistema MOS in condizioni di equilibrio è pari a $-V_{\mathrm{FB}}$. Poiché la grandezza V_{FB} è solitamente negativa, il salto di potenziale è pari a $|V_{\mathrm{FB}}|$. Esso si ripartisce nel salto di potenziale presente ai capi dell'ossido Φ_{ox} e la caduta di potenziale presente sulla regione svuotata di semiconduttore Φ_{S}, come mostrato nella figura 5.8 (c).

Siamo ora in grado di costruire il diagramma a bande complessivo del sistema MOS all'equilibrio. Si ricordi che l'energia potenziale è legata al potenziale elettrostatico mediante la relazione $-q\varphi = U$. Poiché siamo in presenza di una eterostruttura è necessario legare il potenziale elettrostatico U a un livello energetico esistente in tutti e tre i materiali come ad esempio il livello dell'energia del vuoto:

$$-q\varphi(y) = U_0(y) \tag{5.4}$$

Noto l'andamento dell'energia potenziale, si disegna l'andamento del livello di vuo-

[9] Si ricordi che in presenza di giunzione tra materiali diversi si ha la conservazione della componente normale alla superficie di interfaccia del vettore di spostamento dielettrico $\mathcal{D} = \epsilon\mathcal{E}$. Nel nostro caso quindi deve essere $\epsilon_{\mathrm{ox}}\mathcal{E}(y = 0^-) = \epsilon_{\mathrm{S}}\mathcal{E}(y = 0^+)$, e, poiché $\epsilon_{\mathrm{ox}} \neq \epsilon_{\mathrm{S}}$ se ne deduce che $\mathcal{E}(y = 0^-) \neq \mathcal{E}(y = 0^+)$, ovvero il campo elettrico in $y = 0$ è discontinuo.

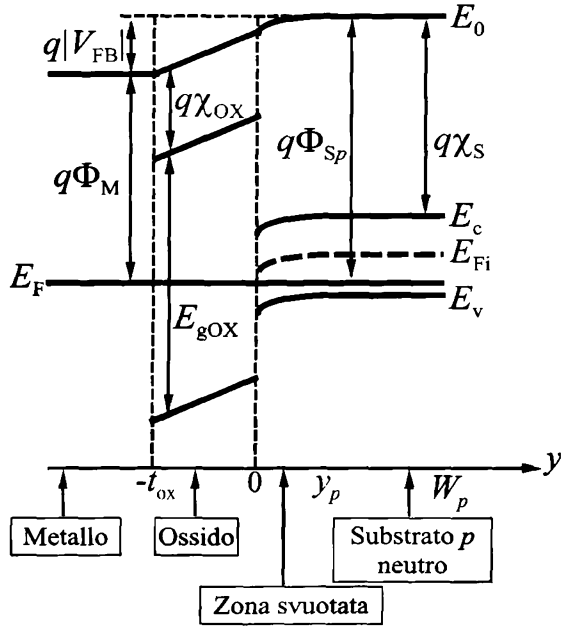


Figura 5.9 Diagramma a bande del sistema MOS all'equilibrio.

to $U_0(y)$ e tutti gli altri livelli energetici vengono di conseguenza fissati mantenendo le affinità elettroniche e le ampiezze di banda costanti nei rispettivi materiali. Il diagramma a bande risutante è mostrato nella figura 5.9. Si osservi che in coerenza con l'incurvamento del livello del vuoto $U_0(y)$ è stato possibile ottenere un livello di Fermi costante in tutto il sistema. Allineando i livelli di Fermi dal lato del semiconduttore e del metallo, resta anche definito lo stesso livello all'interno dell'ossido. Si osservi che a differenza della giunzione *pn*, il sistema MOS presenta un diagramma bande discontinuo sia per quanto riguarda il livello E_c che E_v (essi non sono nemmeno definiti nel metallo!): questo risultato è caratteristico delle eterostrutture, ovvero giunzioni di materiali con caratteristuche fisiche diverse, come, appunto, il sistema MOS.

Il salto complessivo di potenziale presente nella struttura si ricava dall'andamento del livello del vuoto U_0. Tra il metallo e il semiconduttore il salto energetico è pari a $\Delta U = U_0(y = -t_M) - U_0(y = W_p) = qV_{FB}$: esso risulta negativo poiché il metallo si trova ad una energia inferiore rispetto al semiconduttore. Per costruzione, però, si osserva che il salto di energia è anche dato dalla differenza del lavoro di estrazione dal lato del semiconduttore e del lavoro di estrazione dal lato del metallo: infatti per allineare i livelli di Fermi la struttura a bande del metallo è stata traslata verso il basso di una quantità pari alla differenza dei lavori di estrazione dei due materiali. Si ha quindi

$$qV_{FB} = q\Phi_M - q\Phi_{Sp} \tag{5.5}$$

Questa importante relazione permette di ricavare la tensione di banda piatta del sistema MOS noti i parametri fisici dei materiali che costituiscono il sistema stesso. Un esempio

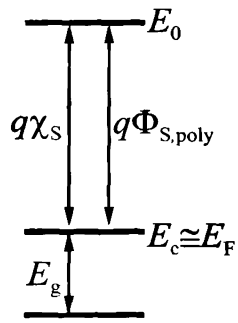


Figura 5.10 Diagramma a bande del polisilicio drogato n^+.

di calcolo della tensione di banda piatta è presentato nell'esempio 5.1.

Esempio 5.1 Si consideri un substrato di silicio drogato p con drogaggio $N_A = 10^{16}$ cm^{-3} e gate formato da

- ▷ alluminio
- ▷ polisilicio drogato (n^+-poly).

Si calcoli la tensione di banda piatta del sistema MOS nei due casi.
Dalla (5.5) si ricava l'espressione della tensione di banda piatta

$$V_{FB} = \Phi_M - \Phi_{Sp}$$

È necessario quindi calcolare separatamente il potenziale di estrazione nel semiconduttore Φ_{Sp} e quello nel metallo Φ_M.
Per quanto riguarda il semiconduttore, il lavoro di estrazione è dato dalla (5.1):

$$q\Phi_{Sp} = q\chi_S + E_g - k_B T \ln \frac{N_v}{N_A}$$

Poiché il substrato è formato da silicio, si sostituiscono alle costanti fisiche quelle relative al silicio, ovvero, $E_g = 1{,}12$ eV, $q\chi_S = 4{,}05$ eV, $N_v = 1{,}04 \cdot 10^{19}$ cm^{-3} mentre il drogaggio è assegnato dal problema. Si ottiene:

$$q\Phi_{Sp} = (4.05 + 1.12 - 0.18)\ \text{eV} = 4.99\ \text{eV}$$

Passando ad analizzare il gate si devono distinguere i due casi: gate in alluminio o in polisilicio drogato. Nei metalli il lavoro di estrazione è un valore caratteristico del materiale che di solito si valuta sperimentalmente: nel caso dell'alluminio, si trova tabulato $q\Phi_M \simeq 4.1$ eV.
Nel caso in cui il metallo venga sostituito con polisilicio, la struttura a bande è simile a quella del silicio cristallino. Poiché poi il polisilicio è pesantemente drogato di tipo n per aumentarne la conducibilità elettrica, il livello di Fermi si trova molto prossimo alla banda di conduzione e, come mostrato nella figura 5.10, $E_F \simeq E_c$. In questo caso la distanza del livello di Fermi dal livello di vuoto, ovvero il lavoro di estrazione del polisilicio, è praticamente uguale alla distanza di E_c dal livello di vuoto, ovvero alla affinità elettronica $q\chi_S$ del silicio. In definitiva nel caso in cui il gate del sistema MOS sia di polisilicio drogato si può approssimativamente supporre

$$q\Phi_M = q\chi_S = 4.05\ \text{eV}$$

Si può ora procedere con il calcolo della tensione di banda piatta V_{FB}. Nel caso del gate in alluminio

$$V_{FB} = \Phi_M - \Phi_{Sp} = (4.1 - 4.99)\ \text{V} = -0.89\ \text{V},$$

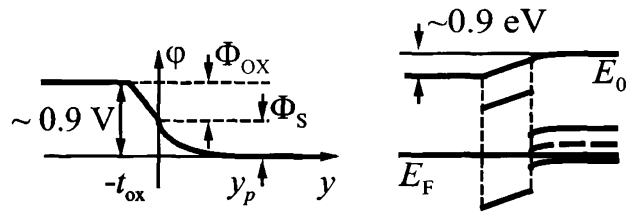


Figura 5.11 Potenziale elettrostatico e diagramma a bande all'equilibrio nei due sistemi MOS dell'esempio 5.1.

mentre nel caso del gate in n^+ –poly:

$$V_{FB} = \Phi_M - \Phi_{Sp} = (4.05 - 4.99)\ \text{V} = -0.94\,\text{V}.$$

Si osservi che in entrambi i casi si è ottenuto un valore della tensione di banda piatta negativo, come ci si aspetta in un sistema MOS su substrato di tipo p. Si invita il lettore a ripetere il calcolo con valori di drogaggio di substrato compresi tra $N_A = 10^{15} - 10^{17}$ cm^{-3} per convincersi che questa caratteristica continua a essere verificata in tutte le strutture MOS significative.
La tensione di banda piatta trovata (approssimativamente pari a -0.9 V in entrambi i casi) consente di disegnare l'andamento del potenziale e del diagramma a bande all'equilibrio del sistema MOS in esame, come mostrato nella figura 5.11.

5.1.3 Il sistema MOS fuori equilibrio

Si supponga ora di applicare al gate di un sistema MOS su substrato di tipo p una tensione $V_G \neq 0$, portandolo fuori equilibrio. Questa tensione viene a sommarsi alla d.d.p. interna presente già all'equilibrio e pari a $-V_{FB}$, cambiando sia il bilancio di carica sia la distribuzione di potenziale e di campo elettrico nel sistema stesso. Il salto di potenziale totale tra il metallo e il semiconduttore risulta ora dato da $V_G - V_{FB}$ e, di conseguenza, il salto energetico corrisponde a $-q(V_G - V_{FB})$. A seconda del valore della tensione applicata, quindi, la barriera di potenziale (o energia) può variare fino ad essere annullata o addirittura cambiare segno rispetto alla condizione di equilibrio.

Un caso particolare si ha se V_G è esattamente pari alla tensione di banda piatta V_{FB}: come anticipato, questa condizione è particolarmente significativa poiché corrisponde anche al passaggio dalla regione di accumulo a quella di svuotamento. Per comprendere quello che accade nel sistema MOS, si consideri l'andamento del diagramma a bande in questa condizione, riportato nella figura 5.12, dove per confronto è mostrato in tratteggio anche il livello di Fermi in equilibrio.

Si osservi per prima cosa che, poiché il sistema non è più all'equilibrio, non è ora possibile definire un unico livello di Fermi costante in tutto il sistema. Si può però dimostrare che poiché nella struttura non passa corrente, è ancora possibile definire dei quasi-livelli di Fermi costanti[10] rispettivamente nel metallo, E_{FM}, e nel semiconduttore, E_{Fp}. Dalla parte del metallo gli elettroni acquisiscono, sotto l'azione della tensione esterna applicata V_G, una energia potenziale pari a $-qV_G$ in aggiunta a quella che hanno in condizioni di equilibrio. I livelli energetici dalla parte del metallo, e quindi il livello di Fermi del metallo stesso, vengono quindi traslati rigidamente della quantità

[10] La definizione rigorosa dei quasi-livelli di Fermi in un semiconduttore si trova nell'approfondimento 2.3: nel sistema MOS la corrente statica è nulla e di conseguenza il quasi-livello di Fermi delle lacune E_{Fp} è costante.

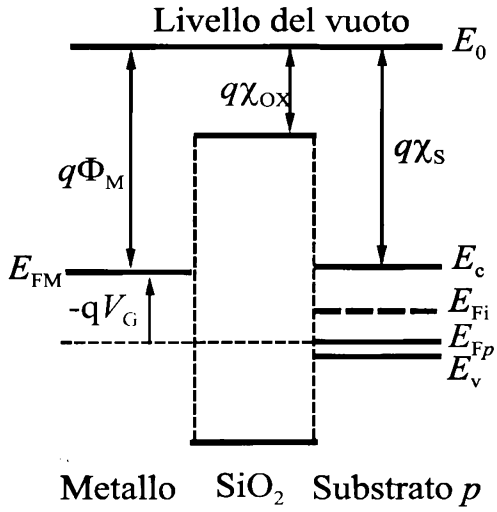


Figura 5.12 Diagramma a bande del sistema MOS nella condizione di banda piatta.

$-qV_G$ rispetto alla condizione di equilibrio. Nella figura 5.12 si osserva infatti che il livello di Fermi E_{FM} viene traslato verso l'alto (si ricordi che V_{FB} è una grandezza negativa e di conseguenza $-qV_G = -qV_{FB}$ risulta una variazione *positiva* di energia) rispetto all'equilibrio fino ad annullare completamente la barriera energetica[11]. In questa condizione la barriera di potenziale interna del sistema MOS viene completamente annullata e, di conseguenza, la carica interna del sistema è nulla. Poiché non è presente carica nè nel metallo nè nel semiconduttore le bande risultano rettilinee. Inoltre la differenza di potenziale totale è nulla, e quindi *le bande sono rette orizzontali.* Questa condizione prende il nome di *condizione di banda piatta* e, poiché essa si verifica quando viene applicata una tensione pari a V_{FB}, resta anche spiegato il motivo per cui questa particolare tensione è detta proprio tensione di banda piatta.

Si consideri ora il caso $V_G < V_{FB}$ (accumulo di lacune). Il livello di Fermi del metallo viene traslato ancora più verso l'alto rispetto alla condizione di banda piatta e il diagramma a bande risultante è rappresentato nella figura 5.13 (a), dove il tratteggio mostra, per confronto, anche l'andamento del livello di vuoto all'equilibrio. Si osservi che la d.d.p. ai capi del sistema MOS è ora negativa, opposta a quella all'equilibrio e per giustificare questo potenziale anche la carica presente nel sistema deve avere segno opposto: nel metallo si forma uno strato superficiale di elettroni mentre nel semiconduttore la carica è positiva e data dall'accumulo delle lacune all'interfaccia tra il semiconduttore e l'ossido (vedi la figura 5.6 (a)). Ovviamente anche la pendenza e la curvatura delle bande energetiche sono opposte rispetto all'equilibrio: ad esempio la pendenza delle bande nell'ossido è negativa, in corrispondenza del fatto che il segno del campo elettrico dell'ossido stesso è negativo (vedi ancora la figura 5.6 (a)).

[11] Si osservi che i due quasi-livelli di Fermi sono dislocati l'uno rispetto all'altro di una quantità pari alla tensione applicata tra il metallo e il semiconduttore, in maniera analoga a quanto accade ai quasi-livelli di Fermi della giunzione *pn* fuori equilibrio (cfr. l'approfondimento 3.3).

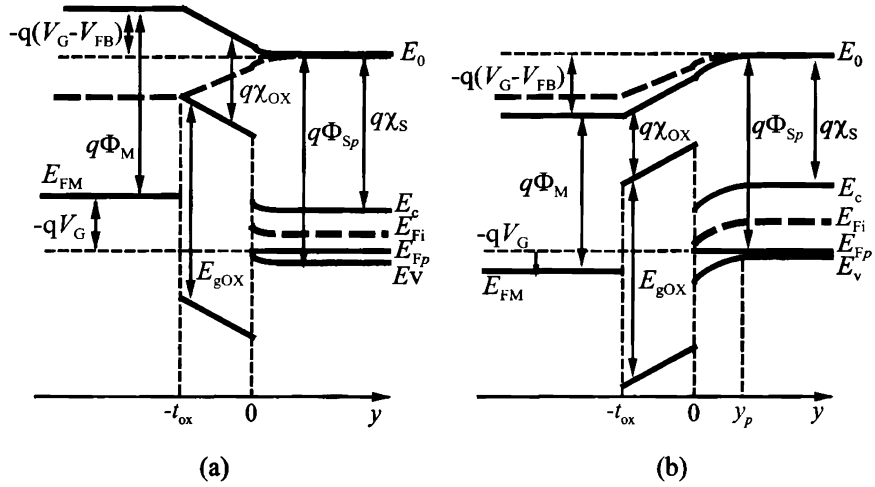


Figura 5.13 Diagramma a bande del sistema MOS nella condizione di accumulo (a) e svuotamento (b).

Se invece $V_G > V_{FB}$, si ha la regione di svuotamento. Il livello di Fermi del metallo viene questa volta traslato verso il basso rispetto alla condizione di banda piatta, come mostrato nella figura 5.13 (b). Sul metallo è presente una carica positiva mentre nel semiconduttore una carica negativa in maniera analoga a quanto avviene nella condizione di equilibrio stesso: si ricordi infatti che il sistema MOS all'equilbrio si trova proprio nella condizione di svuotamento. All'aumentare della tensione applicata aumentano la carica nella regione svuotata di semiconduttore, l' incurvamento delle bande e il salto energetico totale.

Si osservi infine il diagramma a bande nella condizione di inversione, ovvero per $V_G \gg V_{FB}$, riportato nella figura 5.14: all'aumentare della tensione applicata al gate la curvatura delle bande nel semiconduttore diventa sempre più pronunciata a causa dell'aumento della carica negativa nella regione svuotata di semiconduttore. Per effetto di questa curvatura, in un sottile strato in prossimità della interfaccia tra l'ossido e il semiconduttore *il livello di Fermi intrinseco interseca il livello di Fermi* del semiconduttore, portandosi al di sotto di esso. Poiché un semiconduttore in cui il livello di Fermi intrinseco si trovi al di sotto del livello di Fermi è di tipo n e gli elettroni sono maggioritari[12], se ne conclude che al di sotto della interfaccia il semiconduttore ha subito una *inversione di popolazione.* Si forma quindi uno strato di elettroni, o strato di inversione, con spessore estrememente sottile tanto da poter essere considerato uno strato di carica puramente superficiale. La concentrazione di elettroni nello strato di inversione aumenta all'aumentare della tensione applicata al gate, mentre al diminuire di questa la concentrazione di elettroni diminuisce, tanto che il sistema MOS esce dalla condizione di inversione e si riporta nella condizione di svuotamento.

[12] Si ricordino le equazioni di Shockley che legano le concentrazioni di elettroni e lacune alla posizione del livello di Fermi relativamente al livello di Fermi intrinseco.

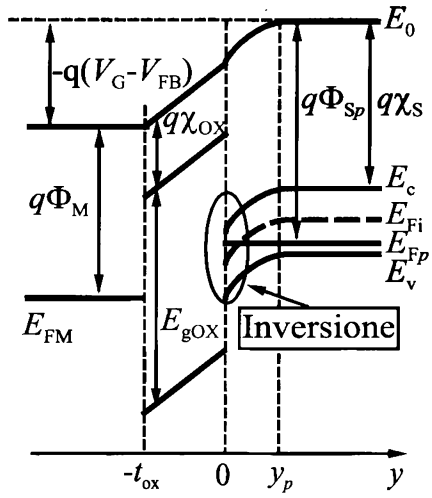


Figura 5.14 Diagramma a bande del sistema MOS nella condizione di inversione.

5.1.4 La legge di controllo di carica

Obiettivo di queto paragrafo è ricavare la legge di controllo di carica, che esprime la concentrazione degli elettroni nello strato di inversione del sistema MOS in funzione della tensione applicata al gate. Poiché lo strato di inversione ha spessore t_{inv} molto piccolo. gli elettroni in esso presenti sono approssimabili ad una carica superficiale: in questo caso più che la concentrazione di elettroni è interessante calcolare la carica di elettroni per unità di superficie Q_n, definita come l'integrale della concentrazione degli elettroni all'interno dello strato di inversione nella direzione perpendicolare all'interfaccia:

$$Q_n = -q \int_0^{t_{inv}} n \, dy \tag{5.6}$$

dove n è la concentrazione di elettroni. Si noti che dimensionalmente Q_n risulta effettivamente una carica per unità di superficie. Poiché n è trascurabile al di fuori dello strato di inversione, l'integrale precedente può anche venire esteso a tutto il substrato di semiconduttore

$$Q_n = -q \int_0^{W_p} n \, dy \tag{5.7}$$

Oltre alla carica Q_n, nel sistema MOS in inversione è poi presente la carica dovuta agli accettatori ionizzati nella regione svuotata di semiconduttore e la carica sul metallo: poiché si tratta comunque di strati omogenei di carica, anche per essi è opportuno ricorrere al concetto di carica per unità di superficie. In particolare definiamo la carica Q_d nella regione svuotata di semiconduttore, che si ricava integrando nella direzione perpendicolare alla interfaccia la densità di carica $\rho = -qN_A$, dovuta agli accettatori

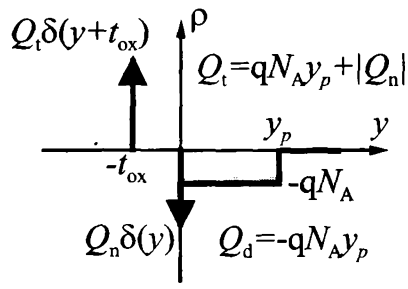


Figura 5.15 Densità di carica nel sistema MOS in inversione.

ionizzati:

$$Q_d = \int_0^{y_p} \rho \, dy = -qN_A y_p \tag{5.8}$$

Infine si ha la carica per unità di superficie nel metallo, positiva e uguale e opposta alla somma delle altre due. Essa verrà quindi indicata come carica totale Q_t. Infatti, anche al di fuori dell'equilibrio, nel condensatore MOS deve essere verificata la condizione di neutralità:

$$Q_t = -Q_n - Q_d = |Q_n| + qN_A y_p \tag{5.9}$$

In definitiva le diverse cariche presenti nel sistema MOS all'inversione sono rappresentate nella figura 5.15. Le cariche Q_n e Q_t sono rappresentate mediante una funzione δ di Dirac centrata rispettivamente nel punto di interfaccia tra l'ossido e il semiconduttore $y = 0$ e nel punto di interfaccia tra il metallo e l'ossido $y = -t_{ox}$.

Per calcolare la relazione di *controllo di carica*, ovvero il legane tra Q_n e V_G nella condizione di inversione, procediamo a valutare separatamente Q_t e Q_d ottenendo Q_n per differenza dalla (5.9). Per calcolare le prime due quantità ci si pone però un problema: in quale intervallo di tensioni V_G va effettuato questo calcolo, ovvero quale è l'intervallo di valori di V_G che corrisponde esattamente alla condizione di inversione a cui siamo interessati? In realtà questa regione è stata per ora definita in maniera del tutto qualitativa mediante la relazione $V_G \gg V_{FB}$, da intendersi nel senso che al crescere di V_G prima o poi il sistema passa dalla regione di svuotamento a quella di inversione. Ma quale è, se esiste, il valore di V_G per cui questo avviene?

In realtà il confine tra la regione di svuotamente e quella di inversione non è facilmente definibile. All'aumentare di V_G (ovvero passando dalla situazione della figura 5.13 (b) a quella della 5.14) il livello di Fermi si avvicina gradualmente al livello di Fermi intrinseco fino ad intersecarlo. Quando i due livelli coincidono esattamente, il materiale all'interfaccia è intrinseco, ovvero sia la concentrazione di elettroni sia quella di lacune sono pari a n_i. Aumentando ancora la tensione, il livello di Fermi intrinseco si sposta al di sotto del livello di Fermi e la concentrazione di elettroni comincia ad aumentare gradualmente. Si passa quindi ad una condizione di debole inversione finché la concentrazione di elettroni rimane modesta per passare man mano a una condizone di forte inversione in cui il materiale è molto ricco di elettroni. Ma per quale V_G si ha forte inversione, ovvero una quantità di carica significativa nello strato di inversione? Definiremo in maniera arbitraria, ma ragionevole, che nella condizione di forte inver-

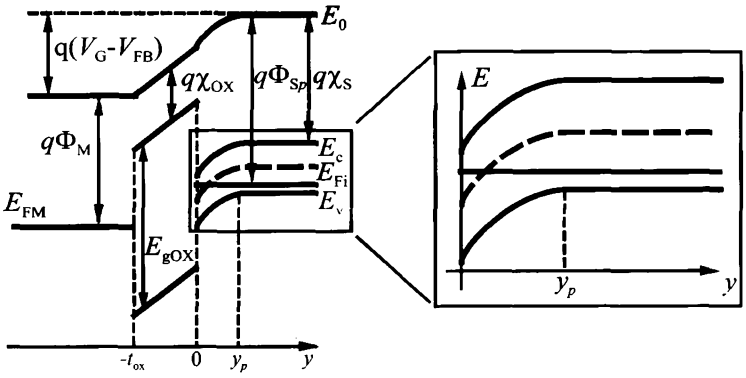


Figura 5.16 Dettaglio della struttura a bande nel semiconduttore del sistema MOS.

sione si abbia *la concentrazione di elettroni nello strato di inversione almeno pari alla concentrazione di lacune nel lato neutro del semiconduttore.* Tradotto in formule:

$$n(y=0) = p(y_p < y < W_p) \approx N_A \tag{5.10}$$

dove si è utilizzato il fatto che il semiconduttore nell'intervallo $y_p < y < W_p$ è neutro e la concentrazione di lacune in questa regione è costante e pari ad N_A. La condizione precedente corrisponde a richiedere che la concentrazione di elettroni nello strato di inversione sia arrivata ad un livello pari alla concentrazione originaria di lacune nel semiconduttore stesso, ovvero che la *inversione di popolazione sia completa.*

Utilizzando le relazioni di Shockley, la relazione che definisce la condizione di forte inversione comporta anche:

$$E_F - E_{Fi}(y=0) = E_{Fi}(y_p < y < W_p) - E_F = k_B T \ln \frac{N_A}{n_i} \tag{5.11}$$

Concentriamoci ora sul diagramma a bande del semiconduttore (cfr. la figura 5.16), il cui dettaglio è mostrato nella figura 5.17. Per la (5.11), il salto di energia complessivo sulla regione svuotata di semiconduttore tra $y = 0$ e $y = y_p$ è:

$$E_{Fi}(y_p < y < W_p) - E_{Fi}(y=0) = 2\, k_B T \ln \frac{N_A}{n_i} \tag{5.12}$$

Definendo infine la variabile ϕ_p

$$\phi_p = \frac{E_{Fi}(y_p < y < W_p) - E_F}{q} = V_T \ln \frac{N_A}{n_i} \tag{5.13}$$

si conclude che nella condizione di forte inversione il salto di energia complessivo è pari a $2q\phi_p$, come risulta evidente sempre dal grafico della figura 5.17.

La condizione di forte inversione può anche essere tradotta in termini di potenziale elettrostatico, piuttosto che in termini di energia potenziale. La d.d.p. ai capi della

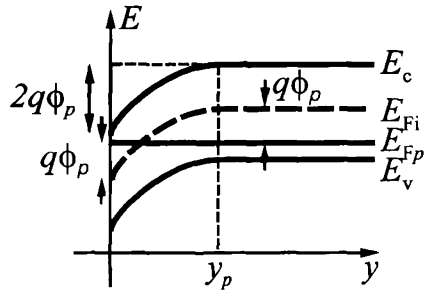


Figura 5.17 Struttura a bande nel semiconduttore del sistema MOS nella condizione di forte inversione.

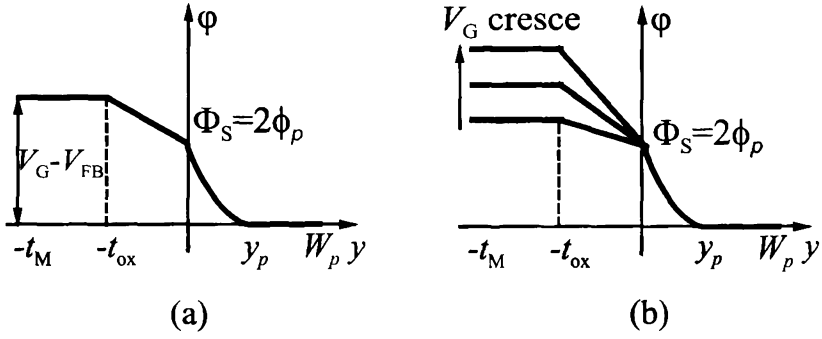


Figura 5.18 Potenziale elettrostatico del sistema MOS nella condizione di forte inversione.

regione svuotata di semiconduttore in condizione di forte inversione è pari a $2\phi_p$ e di conseguenza l'andamento del potenziale nel sistema MOS risulta quello mostrato nella figura 5.18 (a). Definendo il *potenziale superficiale* Φ_S come il valore del potenziale in $y = 0$, si ha

$$\Phi_S = 2\phi_p \tag{5.14}$$

La condizione di forte inversione è caratterizzata da una importante proprietà: se si aumenta la tensione V_G al di sopra del valore di innesco di tale condizione, la d.d.p. ai capi della regione svuotata di semiconduttore rimane approssimativamente costante mentre aumenta solamente la d.d.p. (e il campo elettrico) sull'ossido. Questa proprietà è mostrata qualitativamente nella figura 5.18 (b) dove si osserva che *il potenziale superficiale* Φ_S *è costante e pari a* $2\phi_p$ mentre la tensione ai capi dell'ossido $\Phi_{ox} = V_G - V_{FB} - \Phi_S$ aumenta linearmente con V_G. In altre parole il semiconduttore rimane sostanzialmente "congelato" al variare di V_G. La dimostrazione di questa proprietà non è banale e si rimanda il lettore a testi più avanzati [2]. In estrema sintesi si dimostra che la carica Q_n di elettroni nello strato di inversione dipende esponenzialmente dal valore del potenziale Φ_S mentre la carica Q_d dipende da $\sqrt{\Phi_S}$. Si possono quindi avere grosse variazioni della carica Q_n pur avendo piccole variazioni di Φ_S, tanto che in prima approssimazione si può supporre che Φ_S rimanga costante. In questo caso

anche la carica Q_d è costante. La relazione esponenziale esplicita della dipendenza di Q_n da Φ_S non verrà utilizzata in questo testo, poiché Q_n verrà calcolata non direttamente a partire da Φ_S ma per differenza *utilizzando la relazione di neutralità di carica* (5.9)

$$Q_n = -Q_t - Q_d \tag{5.15}$$

Si procede quindi a calcolare separatamente Q_d e Q_t per valutare Q_n. Dalla (5.18) la carica fissa nella regione svuotata è:

$$Q_d = -qN_A y_p \tag{5.16}$$

dove y_p si calcola risolvendo l'equazione di Poisson nel semiconduttore. Come mostrato nell'approfondimento 5.1, si ottiene:

$$\Phi_S = \frac{qN_A y_p^2}{2\epsilon_S} \Longrightarrow y_p = \sqrt{\frac{2\epsilon_S \Phi_S}{qN_A}} \tag{5.17}$$

Sostituendo poi $\Phi_S = 2\phi_p$ in forte inversione si ottiene:

$$Q_d = -\sqrt{2q\epsilon_S N_A}\sqrt{2\phi_p} \tag{5.18}$$

Si osservi che nella condizione di forte inversione Q_d non dipende dalla tensione applicata V_G.

Approfondimento 5.1 Nel substrato del sistema MOS all'inversione la densità di carica ρ è rapprentata nella figura 5.8 (a). Il campo elettrico $\mathcal{E}(y)$ nel semiconduttore si ottiene integrando l'equazione di Gauss in forma differenziale:

$$\frac{d\mathcal{E}}{dy} = \frac{\rho}{\epsilon_S} \tag{5.19}$$

Nella regione neutra $y_p \leq y \leq W_p$, dove $\rho = 0$, il campo elettrico è costante e in particolare nullo. Nella regione svuotata $0 < y < y_p$, si ha

$$\frac{d\mathcal{E}}{dy} = -\frac{qN_A}{\epsilon_S} \tag{5.20}$$

ovvero

$$\mathcal{E}(y) = -\frac{qN_A}{\epsilon_S} y + c_1 \tag{5.21}$$

Tenendo conto della condizione al contorno in $y = y_p$, si ricava c_1

$$\mathcal{E}(y_p) = -\frac{qN_A}{\epsilon_S} y_p + c_1 = 0 \Longrightarrow c_1 = \frac{qN_A}{\epsilon_S} y_p \tag{5.22}$$

In definitiva:

$$\mathcal{E}(y) = \begin{cases} -\frac{qN_A}{\epsilon_S}(y - y_p) & 0 < y < y_p \\ 0 & y \geq y_p \end{cases} \tag{5.23}$$

Noto $\mathcal{E}(y)$, il potenziale viene valutato utilizzando la definizione

$$\frac{d\varphi}{dy} = -\mathcal{E}(y) = \begin{cases} \frac{qN_A}{\epsilon_S}(y - y_p) & 0 < y < y_p \\ 0 & y \geq y_p \end{cases} \tag{5.24}$$

completata da una condizione al contorno, che corrisponde alla scelta del riferimento di potenziale. Nel caso del sistema MOS si è scelto come riferimento di potenziale il terminale di substrato, per cui $\varphi(y) = 0$ per $y \geq y_p$. Integrando invece la funzione $-\mathcal{E}(y)$ per $0 < y < y_p$ si ottiene:

$$\varphi(y) = \frac{qN_A}{2\epsilon_S}(y - y_p)^2 + k_1 \tag{5.25}$$

dove la costante k_1 è definita dalla condizione di continuità del potenziale in y_p

$$\varphi(y_p) = k_1 = 0 \tag{5.26}$$

Il potenziale superficiale ϕ_S nel sistema MOS è quindi dato da:

$$\Phi_S = \varphi(y = 0) = \frac{qN_A}{2\epsilon_S} y_p^2 \tag{5.27}$$

La carica Q_t nel metallo si calcola come la carica presente sull'armatura superiore del condensatore MOS. Si definisce quindi la *capacità per unità di superficie dell'ossido* C_{ox} in maniera analoga ad un condensatore convenzionale a facce piane e parallele:

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{5.28}$$

La carica totale per unità di area Q_t vale allora:

$$Q_t = C_{ox}\Phi_{ox} \tag{5.29}$$

dove Φ_{ox} è la tensione applicata all'ossido:

$$\Phi_{ox} = V_G - V_{FB} - \Phi_S = V_G - V_{FB} - 2\phi_p \tag{5.30}$$

Si ottiene quindi:

$$Q_t = C_{ox}(V_G - V_{FB} - 2\phi_p) \tag{5.31}$$

La carica Q_n nello strato di inversione si ricava adesso per differenza utilizzando la (5.9), sostituendo le espressioni di Q_d (5.18) e Q_t (5.31), ottenendo:

$$Q_n = -C_{ox}(V_G - V_{FB} - 2\phi_p) + \sqrt{4q\epsilon_S N_A \phi_p} \tag{5.32}$$

La (5.32) è nota come la *legge di controllo di carica* del sistema MOS su substrato di tipo p e fornisce un legame esplicito tra la carica per unità di superficie dello strato di inversione e la tensione di controllo applicata al gate. Si osservi che la legge di controllo di carica è ovviamente definita solamente nella condizione di inversione, nella quale la carica Q_n è diversa da zero.

5.1.5 La tensione di soglia

Si definisce *tensione di soglia* V_{th0} del sistema MOS la tensione V_G per cui $Q_n = 0$. Imponendo, nella legge di controllo di carica (5.32), $Q_n = 0$ e ricavando V_G si ottiene:

$$V_{th0} = V_{FB} + 2\phi_p + \frac{\sqrt{4q\epsilon_S N_A \phi_p}}{C_{ox}} \tag{5.33}$$

Si definisce poi il *coefficiente di substrato* (o coefficiente di *effetto body*) γ_B

$$\gamma_B = \frac{\sqrt{2q\epsilon_S N_A}}{C_{ox}} \tag{5.34}$$

per cui la tensione di soglia si riduce a:

$$V_{th0} = V_{FB} + 2\phi_p + \gamma_B\sqrt{2\phi_p} \tag{5.35}$$

La tensione di soglia di un sistema MOS su substrato di tipo p risulta solitamente *positiva*.

Sostituendo ora l'espressione di V_{th0} nella relazione di controllo di carica (5.32), essa assume una forma particolarmente semplice:

$$Q_n = -C_{ox}(V_G - V_{th0}) \tag{5.36}$$

In questa forma risulta particolarmente evidente la caratteristica fondamentale della legge di controllo di carica: la carica Q_n *dipende linearmente dalla tensione di controllo* V_G e cresce a a mano a mano che questa si porta al di sopra della tensione di soglia. In questa condizione la carica Q_n risulta *negativa* come deve essere, essendo carica di elettroni. Se invece $V_G = V_{th0}$ allora Q_n si annulla per definizione. La tensione di soglia ha in definitiva il particolare significato di indicare per quale tensione applicata V_G, il sistema MOS non ha più elettroni nello strato di inversione, ovvero lo strato di inversione non è presente nel sistema MOS. V_{th0} rappresenta quindi la tensione di gate per cui il sistema MOS passa dalla condizione di svuotamento $Q_n = 0$ e la condizione di inversione in cui $Q_n \neq 0$. La relazione di controllo di carica vale quindi solo per tensioni di gate al di sopra della soglia, mentre al di sotto di essa la carica Q_n è nulla:

$$Q_n = \begin{cases} -C_{ox}(V_G - V_{th0}) & V_G \geq V_{th0} \\ 0 & V_G < V_{th0} \end{cases} \tag{5.37}$$

Il sistema MOS è quindi un sistema *a soglia*: esso si "accende" solo se la tensione supera il valore della soglia, altrimenti è "spento".

Naturalmente come abbiamo già spiegato la transizione tra la regione di svuotamento e la regione di inversione non è in realtà un fenomeno brusco e la carica Q_n non si annulla ad una ben precisa tensione: essa piuttosto diminuisce gradualmente finché la concentrazione di elettroni nella regione svuotata del semiconduttore diventa trascurabile, come normalmente in un semiconduttore drogato di tipo p. La (5.37) rappresenta quindi una approssimazione dell'andamento reale della carica Q_n, come mostrato nella figura 5.19. Il motivo di questa discrepanza è legato al fatto che la relazione di controllo di carica è stata ricavata nella *ipotesi di forte inversione*, ovvero quando la carica nello stato di inversione raggiunge valori elevati, ed essa non è quindi accurata quando la carica Q_n è molto piccola e il sistema MOS si spegne. Si osservi che le uniche differenze rilevanti si hanno però solo in prossimità della tensione di soglia stessa, ovvero quando il sistema è vicino alla condizione di spegnimento.

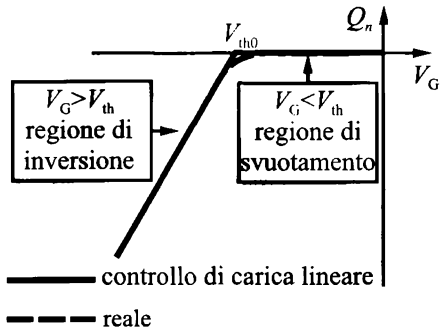


Figura 5.19 Andamento della carica Q_n in funzione di V_G. Linea continua: andamento ideale con legge di controllo di carica lineare (5.37). Linea tratteggiata: andamento reale.

5.1.6 L'effetto di substrato

Si supponga di poter polarizzare il terminale di substrato del sistema MOS con una tensione V_B indipendente[13]. Rimandando il lettore interessato all'approfondimento 5.2, si dimostra che la tensione di soglia varia con V_B seguendo la relazione:

$$V_{th} = V_{FB} + 2\phi_p + \gamma_B \sqrt{(2\phi_p - V_B)} \tag{5.38}$$

Indicando poi con V_{th0} la tensione di soglia con $V_B = 0$, dalla relazione precedente si può scrivere:

$$V_{th} = V_{th0} + \gamma_B \left[\sqrt{(2\phi_p - V_B)} - \sqrt{2\phi_p} \right] \tag{5.39}$$

ovvero la variazione della tensione si soglia è proporzionale al coefficiente γ_B definito nella (5.34), che è infatti denominato coefficiente di effetto di substrato. Poiché γ_B aumenta all'aumentare del drogaggio N_A del substrato, questo effetto è tanto più pronunciato quanto più il semiconduttore è drogato. Esso può invece essere in prima approssimazione trascurato se il drogaggio del substrato è sufficientemente basso.

Approfondimento 5.2 Applicando una tensione non nulla al substrato, i quasi-livelli di Fermi nel metallo E_{FM} e nel semiconduttore E_{Fp} si separano di una quantità proporzionale alla differenza di potenziale tra i due terminali, come dimostrato nell'approfondimento 3.3.

$$E_{FM}(-t_{ox}) - E_{Fp}(y_p) = q(V_B - V_G) \tag{5.40}$$

Per quanto riguarda l'energia del livello di vuoto, nel metallo vale

$$E_0(-t_{ox}) = E_{FM}(-t_{ox}) + q\Phi_M \tag{5.41}$$

mentre per il lato neutro del semiconduttore è dato da

$$E_0(y_p) = E_{Fp}(y_p) + q\Phi_{Sp} \tag{5.42}$$

[13] Nella analisi precedente si è supposto $V_B = 0$, ovvero il terminale di substrato ha rappresentato il riferimento di potenziale per le tensioni. In questo momento si suppone invece di poter riferire entrambe le tensioni V_B e V_G ad un riferimento di potenziale diverso. Questa analisi è particolarmente importante per il seguito, dove il sistema MOS è inserito all'interno del transistore MOS.

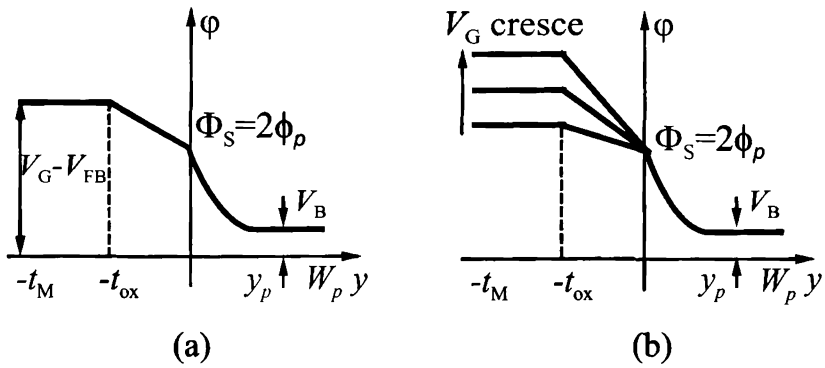


Figura 5.20 Potenziale elettrostatico del sistema MOS nella condizione di forte inversione e con potenziale di substrato non nullo.

Ricordando che $V_{FB} = \Phi_M - \Phi_{Sp}$ il salto di energia complessivo sul sistema MOS risulta:

$$E_0(-t_{ox}) - E_0(y_p) = qV_B - qV_G + q\Phi_M - q\Phi_{Sp} = -q(V_G - V_{FB}) + qV_B \tag{5.43}$$

Poiché il potenziale elettrostatico è messo in relazione con il livello del vuoto E_0 dalla relazione $-q\varphi = E_0$, si ha anche che il salto di tensione è adesso

$$\varphi(-t_{ox}) - \varphi(y_p) = (V_G - V_{FB}) - V_B \tag{5.44}$$

L'andamento del potenziale elettrostatico deve essere modificato rispetto alla figura 5.18, ed è ora mostrato nella figura 5.20
Si osservi che perché il sistema MOS sia nella condizione di forte inversione il potenziale superficiale deve raggiungere il valore $\Phi_S = 2\phi_p$, ma questo potenziale non è più riferito al substrato neutro ma ad un riferimento indipendente, lo stesso a cui sono riferiti anche V_B e V_G. La caduta di potenziale ai capi della regione svuotata del substrato è ora pari a $2\phi_p - V_B$ e di conseguenza viene anche modificata la carica fissa Q_d della (5.18):

$$Q_d = -\sqrt{2q\epsilon_S N_A}\sqrt{2\phi_p - V_B} \tag{5.45}$$

La caduta di potenziale ai capi dell'ossido, invece, non viene modificata dalla applicazione della tensione di substrato e di conseguenza la carica Q_t sul metallo non varia. Sostituendo adesso le espressioni di Q_d (5.45) e Q_t (5.31) all'interno della (5.7), si ottiene che la *legge di controllo di carica dipende anche dalla tensione* V_B:

$$Q_n = -C_{ox}(V_G - V_{FB} - 2\phi_p) + \sqrt{2q\epsilon_S N_A (2\phi_p - V_B)} \tag{5.46}$$

Se a partire da questa nuova relazione di controllo di carica si ricava ora la tensione di soglia ovvero la tensione V_G per cui $Q_n = 0$, si ottiene l'espressione riportata nella (5.38).

Esempio 5.2 Si consideri un sistema MOS costituito da un substrato di Si con drogaggio $N_A = 10^{16}$ cm^{-3}, uno strato di ossido di silicio SiO_2 e il gate di silicio policristallino drogato n^-
Determinare

- ▷ lo spessore di ossido necessario ad ottenere una tensione di soglia di 1.3 V
- ▷ la tensione di substrato necessaria per innalzare la tensione di soglia del 30%

Ricordando l'espressione della tensione di soglia (5.35) e l'espressione della carica Q_d (5.18) si può scrivere:

$$V_{th} = V_{FB} + 2\phi_p + \gamma_B\sqrt{2\phi_p} = V_{FB} + 2\phi_p - \frac{Q_d}{C_{ox}} \tag{5.47}$$

Utilizzando questa espressione si calcolerà prima il valore della capacità dell'ossido C_{ox} e successivamente, dalla (5.28), lo spessore di ossido richiesto.
In primo luogo si calcola ϕ_p definito in (5.13), utilizzando le costanti fisiche del silicio (cfr. esempio 5.1):

$$\phi_p = V_T \ln\frac{N_A}{n_i} = 0.35\,\text{V}$$

Si passa poi a calcolare la tensione di banda piatta seguendo il procedimento illustrato nell'esempio 5.1, ottenendo:

$$V_{FB} = \Phi_M - \Phi_{Sp} = \chi_S - \left(\chi_S + E_g/q - V_T \ln\frac{N_v}{N_A}\right) = -0.94\,\text{V}$$

Utilizzando la (5.18), poi, la carica nella regione svuotata vale:

$$Q_d = -\sqrt{2q\epsilon_S N_A}\sqrt{2\phi_p} = -4.817 \times 10^{-8}\,\text{C/cm}^2$$

Sostituendo i valori trovati nella (5.47) e imponendo $V_{th} = 1.3\,\text{V}$ come richiesto dal problema, è ora possibile determinare C_{ox}:

$$C_{ox} = \frac{-Q_d}{V_{th} - V_{FB} - 2\phi_p} = 3.12 \times 10^{-8}\,\text{F/cm}^2$$

Infine dalla (5.28) si ottiene lo spessore dell'ossido richiesto al primo punto:

$$t_{ox} = \frac{\epsilon_{ox}}{C_{ox}}$$

Per effettuare il calcolo si ricordi che $\epsilon_{ox} = \epsilon_0\epsilon_{r,ox}$, dove ϵ_0 è la costante dielettrica del vuoto pari a $8{,}85 \times 10^{-14}$ F/cm^2 mentre $\epsilon_{r,ox}$ è la *costante dielettrica relativa dell'ossido*, che nel caso dell'ossido di silicio vale 3.9. Sostituendo i valori si ricava

$$t_{ox} = \frac{\epsilon_{ox}}{C_{ox}} = 110.6\,\text{nm}$$

Per quanto riguarda la seconda richiesta del problema, si utilizza la (5.39) per ottenere il valore della *variazione della tensione di soglia* in funzione della tensione di substrato V_B.

$$\Delta V_{th} = \gamma_B\left[\sqrt{(2\phi_p - V_B)} - \sqrt{2\phi_p}\right]$$

Invertendo la relazione precedente

$$V_B = -\left(\frac{\Delta V_{th}}{\gamma_B} + \sqrt{2\phi_p}\right)^2 + 2\phi_p$$

Poiché si vuole una variazione della tensione di soglia del 30%, si deve imporre $\Delta V_{th} = 0.3\,V_{th} = 0.39$ V. Si sostituisce poi il coefficiente di substrato calcolato dalla (5.34)

$$\gamma_B = \frac{\sqrt{2q\epsilon_S N_A}}{C_{ox}} = 1.845\ \text{V}^{1/2}$$

ottenendo

$$V_B = -0.398\ \text{V}$$

Si osservi che per *aumentare la tensione di soglia*, V_B deve essere *negativa*.

5.1.7 Il sistema MOS su substrato di tipo *n*

Il sistema MOS su substrato di tipo n, la cui struttura è stata introdotta nella figura 5.1, è del tutto analogo al sistema MOS su substrato p, scambiando il tipo di drogaggio del semiconduttore. Si intuisce che il comportamento del sistema MOS su substrato n risulta in qualche modo *speculare* all'analoga stuttura di tipo p, a patto di sostituire il ruolo delle lacune con quello degli elettroni e viceversa. Più in generale, il sistema MOS su substrato n si comporta ancora come un condensatore MOS, le cui armature hanno però carica di segno opposto rispetto al substrato p. Cambiando i segni delle cariche, poi, è anche necessario cambiare i segni delle tensioni di pilotaggio, ed è quindi facile convincersi che si avranno anche in questo caso tre regioni di funzionamento così definite:

- ▷ la regione di accumulo di elettroni, per $V_G > V_{FB}$, in cui gli elettroni sono attirati all'interfaccia tra l'ossido e il semiconduttore e il metallo si carica positivamente
- ▷ la regione di svuotamento di elettroni, per $V_G < V_{FB}$, in cui si forma uno strato svuotato di elettroni al di sotto dell'interfaccia tra l'ossido e il semiconduttore, mentre il metallo si carica positivamente. Lo strato svuotato presenta una carica *positiva* Q_d formata dalla concentrazione di donatori ionizzati
- ▷ la regione di inversione di popolazione, per $V_G \ll V_{FB}$, in cui si crea un sottile strato di *lacune libere* in corrispondenza della interfaccia tra l'ossido e il semiconduttore

Indicando con Q_p la concentrazione di lacune libere per unità di area nella condizione di inversione, e procedendo in maniera analoga a quanto dimostrato per Q_n nel sistema MOS su substrato p, si ricava la *legge di controllo di carica* del sistema su substrato n:

$$Q_p = -C_{ox}(V_G - V_{FB} + 2\phi_n) - \sqrt{4q\epsilon_S N_D \phi_n} \tag{5.48}$$

dove si è definito ϕ_n in maniera analoga a $\phi_p$nella (5.13)

$$\phi_n = V_T \ln \frac{N_D}{n_i} \tag{5.49}$$

La tensione di soglia è definita come la tensione V_G per la quale la carica di lacune nello strato di inversione si annulla, ovvero $Q_p = 0$. Essa vale:

$$V_{th0} = V_{FB} - 2\phi_n - \gamma_B \sqrt{2\phi_n} \tag{5.50}$$

dove

$$\gamma_B = \frac{\sqrt{2q\epsilon_S N_D}}{C_{ox}} \tag{5.51}$$

sostituendo la (5.50) nella (5.48), la legge di controllo di carica assume la forma semplificata:

$$Q_p = -C_{ox}\,(V_G - V_{th0}) \cdot \tag{5.52}$$

Si osservi che il sistema MOS su substrato n si "accende" quando la tensione V_G viene portata *al di sotto* della tensione di soglia, mentre è nulla altrimenti, così che Q_p nella

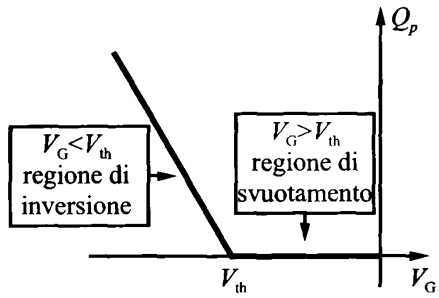


Figura 5.21 Andamento della carica Q_p in funzione di V_G in un sistema MOS su substrato di tipo n.

(5.52) risulta *positiva*, come deve essere la carica formata da lacune. Inoltre, la tensione di soglia è di solito *negativa*. Come mostrato nella figura 5.21, il comportamento del sistema MOS su substrato n risulta quindi del tutto speculare, o più precisamente *complementare*, a quello del sistema MOS su substrato di tipo p.

Anche in questo caso applicando una tensione al substrato la tensione di soglia varia. Analogamente alla (5.39), si ha

$$V_{th} = V_{th0} - \gamma_B \left[\sqrt{(2\phi_n - V_B)} - \sqrt{2\phi_n}\right] \tag{5.53}$$

dove si è indicata con V_{th0} la tensione di soglia per $V_B = 0$.

5.2 I transistori MOSFET

La struttura MOS a due terminali vista nel paragrafo 5.1 costituisce la struttura base del MOSFET. Esso è però una struttura più complessa, in cui ai due terminali di *gate* e *substrato* già visti, se ne aggiungono altri due, il *source* e il *drain*, cosicché il MOSFET è in definitiva un dispositivo a quattro terminali. Questo potrebbe portare a dubitare che il MOSFET sia un transistore, visto che di solito un transistore è un tripolo. Si tenga però presente che il terminale di substrato non viene utilizzato come un terminale di controllo, non essendo applicato intenzionalmente ad esso alcun segnale; di conseguenza il MOSFET ha tre terminali di controllo effettivi, come deve essere in un transistore convenzionale. In particolare, nel MOSFET la corrente che scorre tra il terminale di source e quello di drain, che costituiscono normalmente la porta di uscita, è controllata dalla tensione di controllo sul terminale di gate, che di solito costituisce il teminale di ingresso. Il controllo è puramente in tensione poiché come già visto nello studio del sistema MOS, il gate costituisce una delle armature del condensatore MOS, ed è quindi idealmente *un terminale isolato*, ovvero con corrente statica nulla. Se ne deduce che in condizioni statiche il transistore MOSFET non assorbe corrente alla porta di ingresso.

Poiché la corrente tra il source e il drain nei FET è di tipo unipolare si hanno due casi:

- ▷ se di elettroni si parla di transistore MOSFET a *canale n*, o nMOS
- ▷ se di lacune si parla di transistore MOSFET a *canale p*, o pMOS

A loro volta i transistori a canale n si dividono

- ▷ nMOS *ad arricchimento* (di tipo enhancement) o *normalmente* off
- ▷ nMOS *a svuotamento* (di tipo depletion) o *normalmente* on

e, a loro volta, i transistori a canale p si dividono in

- ▷ pMOS *ad arricchimento* (di tipo enhancement) o *normalmente* off
- ▷ pMOS *a svuotamento* (di tipo depletion) o *normalmente* on

Come si vedrà più avanti, i transistori ad arricchimento si differenziano da quelli a svuotamento per la struttuta fisica ma, indipendentemente dalle differenze, entrambi i transistori a canale n sono costruiti a partire da un sistema MOS su *substrato di tipo p* mentre i pMOS si costruiscono su *substrati di tipo n*. Inltre i due transistori a canale n si portano nello stato *on*, ovvero lo stato in cui possono condurre corrente, quando la tensione di gate si porta *al di sopra della tensione di soglia*; questa poi è positiva per i transistori ad arricchimento e negativa per quelli a svuotamento. I due nMOS hanno quindi comportamento qualitativamente identico e si differenziano solo per il segno della tensione di soglia. Al contrario, i pMOS si portano nello stato *on* quando la tensione di gate si porta *al di sotto della tensione di soglia*; questa è poi negativa nel caso dei pMOS ad arricchimento e positiva nel caso dei dispositivi a svuotamento. Anche in questo caso i due pMOS si differenziano solo per il segno della tensione di soglia.

È immediato anche osservare che i transistori pMOS hanno comportamento del tutto complementare a quello dei corrispondenti nMOS.

Si analizzerà per primo lo nMOS ad arricchimento nel prossimo paragrafo 5.3. per passare poi allo nMOS a svuotamento nel paragrafo 5.4, mettendo in luce quali particolari della struttura fisica sono in grado di modificarne la tensione di soglia. I pMOS verranno infine trattati nel paragrafo 5.5, sfruttando largamente i risultati già ottenuti per i dispositivi nMOS e la complementarietà di comportamento rispetto a questi ultimi.

5.3 Il transistore *n*MOS ad arricchimento

La struttura fisica del transistore nMOS ad arricchimento è mostrata nella figura 5.22 (a

In verticale si riconosce la presenza di un sistema MOS su substrato di tipo p, caratterizzato dalle presenza del contatto di *gate*, realizzato al di sopra di uno strato d ossido e di un substrato di semiconduttore drogato di tipo p. Il substrato ha una metallizzazione che forma il contatto di *substrato* o, in inglese, *bulk* o *body*: analogament a quanto visto nello studio del sistema MOS, questo contatto viene contraddistint con la lettera B. A sinistra e a destra della struttura del gate, sono realizzate n semiconduttore due regioni di tipo n, con drogaggio sufficientemente elevato sia pe compensare la presenza del drogaggio p del substrato, sia per aumentare quanto poss bile la conducibilità elettrica di queste regioni del dispositivo. La regione a sinistra d gate è denominata *source*, e su di essa viene deposta la metallizzazione che costituis il contatto di source stesso, identificato con la lettera S. A destra del gate si ha l regione di *drain* e il contatto di drain, D[14]. Per l'ottimo funzionamento del dispositiv

[14] Si osservi che il MOSFET è spesso una struttura *simmetrica*, tanto che il source e il drain posson in realtà essere scambiati tra loro.

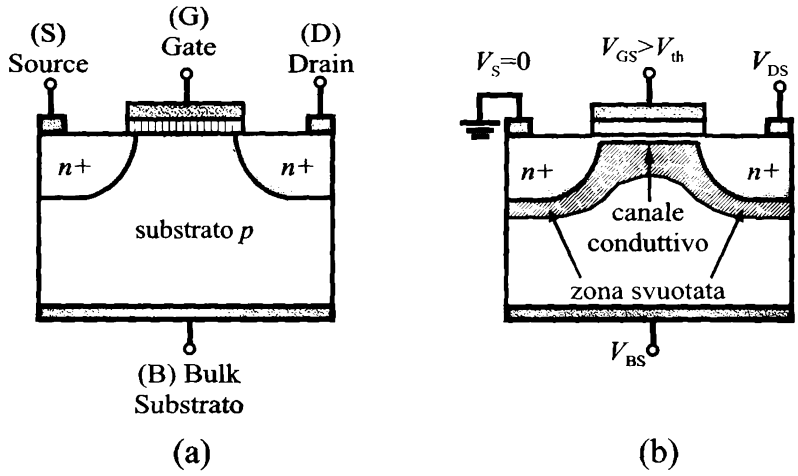


Figura 5.22 (a) Struttura fisica del transistore *n*MOS ad arricchimento. (b) Formazione del canale conduttivo.

le regioni drogate n di source e drain devono essere il più possibile *allineate* ai lati del gate, anche se questa richiesta è difficile da soddisfare dal punto di vista tecnologico e. come mostrato nella figura 5.22 (a), a volte si ha un piccolo *overlapping*, ovvero le regioni n si estendono un poco anche al di sotto dell'ossido del gate.

Essendo il dispositivo a quattro terminali, si riconoscono tre tensioni di pilotaggio indipendenti. Normalmente si riferiscono le tensioni alla tensione del contatto di source, che viene portato a massa: restano quindi definite la *tensione gate-source* V_{GS}, la *tensione drain-source* V_{DS} e la *tensione bulk-source* V_{BS}. Quando la tensione V_{GS} viene portata al di sopra della tensione di soglia V_{th} del sistema MOS formato dal contatto di gate e il substrato, in corrispondenza della interfaccia tra l'ossido e il semiconduttore si forma uno strato di elettroni libero, che corrisponde allo strato di inversione del sistema MOS stesso. Questo strato di carica di elettroni libera si salda a sinistra con la regione di source e a destra alla regione di drain, formando un *canale conduttivo* tra il source e il drain, come mostrato nella figura 5.22 (b). Tra i due terminali di source e drain si è quindi formata una unica regione *ohmica*, ovvero una regione dove si può avere una corrente di trascinamento non nulla[15].

Al di sotto del canale conduttivo si ha una regione di semiconduttore svuotata di lacune e carica per la presenza di accettatori ionizzati negativamente. Si osservi, peraltro, che le due regioni di source e di drain formano due giunzioni *pn* con il substrato di semiconduttore drogato p e a cavallo della giunzione si forma una regione svuotata, che si estende principalmente dalla parte del lato p, essendo il drogaggio di tale regione più debole del drogaggio di tipo n del source e del drain. In definitiva, attorno alla regione ohmica, ricca di elettroni, disegnata in grigio nella figura 5.22 (b), si ha una unica regione svuotata di lacune, carica negativamente, disegnata con tratteggio.

[15] Per regione ohmica si intende una regione in cui è presente una carica elettrica libera, che in presenza di un campo elettrico applicato può dare luogo a una corrente di trascinamento. Una regione ohmica ha una conducibilità elettrica proporzionale alla densità di portatori liberi, come visto nel paragrafo 2.1.1.

Una volta formato il canale conduttivo si può avere un flusso di elettroni dal source, attraverso il canale, fino al drain, dando luogo ad una corrente elettrica che scorre tra il source ed il drain. Perché si abbia questo moto di cariche è necessario che si generi una differenza di potenziale tra il source ed il drain: questa a sua volta origina un campo elettrico all'interno del canale conduttivo, *parallelo alla interfaccia tra l'ossido e il semiconduttore*, in grado di mettere in moto gli elettroni per *trascinamento*. Se la tensione applicata tra il drain e il source, V_{DS} è positiva, allora gli elettroni si muovono effettivamente dal source (in inglese *la sorgente*) al drain (in inglese *il collettore*) attraverso il canale e la corrente I_{DS} generata da questo flusso di carica è positiva entrante nel drain. Perché il dispositivo funzioni correttamente si richiede che, in condizioni statiche, la corrente I_{DS} sia l'unica corrente non nulla, ovvero che siano nulle sia la corrente al terminale di gate sia a quello di substrato. Per quanto riguarda il terminale di gate, la corrente è certamente nulla poiché, come già osservato, questo terminale è isolato. Per garantire che la corrente di substrato sia nulla, è invece necessario che le due giunzioni *pn* formate tra il source e il substrato e tra il drain e il substrato siano polarizzate inversamente o al più non polarizzate. Poiché il terminale di source è a massa, mentre il terminale di drain è ad una tensione V_{DS} positiva, è sufficiente garantire che non conduca la giunzione source-substrato, mantenendo la tensione V_{BS} negativa o al più nulla. Si osservi che il terminale di substrato non è quindi un terminale di controllo nel MOSFET: la tensione in continua eventualmente applicata a questo ternimale è utile solo a garantire il corretto funzionamento del dispositivo stesso. Si osservi anche che se la tensione V_{BS} è non nulla e negativa, la tensione di soglia del dispositivo *aumenta*, come visto nel paragrafo 5.1.6[16]. Il ruolo del terminale di substrato verrà analizzato in dettaglio nel paragrafo 5.7.

5.3.1 Il canale conduttivo

Come illustrato nel paragrafo precedente, il transisitore MOSFET è caratterizzato, in condizioni stazionarie, da una sola corrente che scorre tra il terminale di source e il terminale di drain: si tratta di una corrente di trascinamento di elettroni, che si muovono dal source al drain per effetto di una tensione V_{DS} positiva. È noto che la corrente di trascinamento è proporzionale sia alla densità di carica che partecipa al flusso sia all'intensità del campo elettrico che trascina le cariche stesse. Per calcolare il valore della corrente I_{DS} è quindi necessario calcolare sia la quantità di carica all'interno del canale conduttivo, sia l'entità del campo elettrico. Si noti che la prima richiesta corrisponde in definitiva a determinare una *legge di controllo di carica* all'interno del canale conduttivo, che leghi la densità di carica nello strato di inversione del semiconduttore. il canale conduttivo, appunto, *ad entrambe le tensioni di gate e di drain*, estendendo quindi la legge di controllo di carica già ottenuta nel caso del sistema MOS, dove l'unica tensione di controllo è la tensione di gate. Prima di procedere al calcolo della legge di controllo di carica è opportuno fissare un riferimento di assi all'interno della struttura. Si sceglie l'asse x orientato parallelamente alla interfaccia ossido-semiconduttore e l'asse y perpendicolare a tale interfaccia e orientato verso la profondità del substrato. L'origine dell'asse x viene fissata in corrispondenza dell'inizio del canale conduttivo dalla parte del source, come mostrato nella figura 5.23, mentre l'origine dell'asse y s:

[16] Nelle tecnologie MOS meno recenti l'utilizzo della tensione di substrato per modificare la soglia del MOSFET era comune. Attualmente la tensione di soglia viene regolata mediante impiantazione ionica superficiale, come verrà illustrato nel paragrafo 5.4.

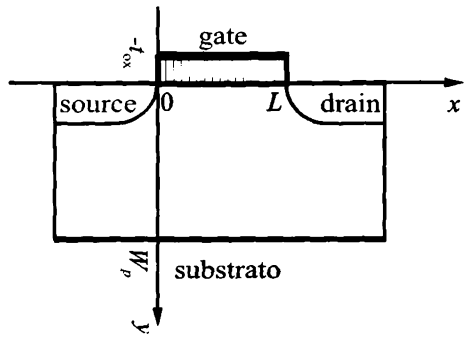


Figura 5.23 Riferimento di assi nel transistore nMOS.

trova in corrispondenza dell'interfaccia ossido-semiconduttore stessa. Si utilizzano le stesse notazioni usate per il sistema MOS: lo spessore dell'ossido è pari a t_{ox} mentre lo spessore del substrato è pari a W_p. Lungo l'asse x invece, il canale conduttivo si estende dall'origine degli assi fino alla regione n del drain, e la sua lunghezza viene indicata con L, come mostrato nella figura 5.23. Si osservi che se le regioni di source e drain sono allineate ai lati del gate, la lunghezza L del canale conduttivo corrisponde anche alla lunghezza fisica della metallizzazione di gate. In prima approssimazione è quindi possibile *identificare la lunghezza del canale conduttivo con la lunghezza fisica del contatto di gate.*

Si consideri ora un elettrone all'interno del canale conduttivo del transistore. Come mostrato nella figura 5.24 (a), esso è sottoposto ad un campo elettrico complessivo formato dalla componente parallela all'asse x, $\mathcal{E}_x$, e quella perpendicolare $\mathcal{E}_y$. Queste due compenenti hanno interpretazioni fisiche molto diverse tra loro. La componente $\mathcal{E}_y$ corrisponde al campo elettrico che nel sistema MOS si genera perpendicolarmente alle interfacce principalmente per effetto della applicazione della tensione V_{GS}; sostanzialmente esso regola la quantità di carica presente nel canale conduttivo stesso. La componente $\mathcal{E}_x$, invece, è indotta nel canale conduttivo dalla applicazione della tensione V_{DS}, che a sua volta induce una caduta di potenziale nella direzione longitudinale al canale stesso. In realtà in una struttura bidimensionale, non è possibile attribuire in maniera rigorosa la presenza del campo $\mathcal{E}_y$ alla *sola tensione* V_{GS} e quella di $\mathcal{E}_x$ alla *sola tensione* V_{DS}; si dimostra, però, che se il contatto di gate è sufficientemente lungo allora questa approssimazione è ben verificata e si parla di *approssimazione di canale graduale.* All'interno di questa approssimazione la analisi del comportamento fisico del canale conduttivo si semplifica, poiché è possibile sfruttare una specie di sovrapposizione degli effetti per la quale ogni elettrone nel canale è sottoposto alla azione delle tensioni di gate e drain in maniera *disaccoppiata*, per cui la tensione V_{DS} agisce nella direzione parallela all'asse x e la tensione V_{GS} nella direzione perpendicolare, sommando alla fine i risultati.

Analogamente a quanto visto per il campo elettrico, anche il potenziale elettrostatico, mostrato nella figura 5.24 (b), ha in linea di principio un andamento molto complesso, indotto dalla *applicazione contemporanea* delle tensioni di gate e di drain

$$\varphi(x,y) = \varphi(x,y,V_{GS},V_{DS}). \tag{5.54}$$

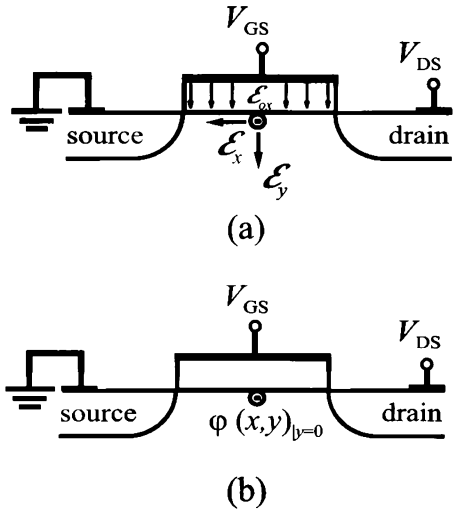


Figura 5.24 Campo elettrico (a) e potenziale elettrostatico (b) a cui sono sottoposti gli elettroni nel canale conduttivo del MOSFET a canale n ad arricchimento.

Il potenziale è un campo bidimensionale e gli elettroni nel canale sono sottoposti al potenziale $\varphi(x,y)$ valutato nella coordinata $y = 0$, che corrisponde alla interfaccia tra l'ossido e il semiconduttore, dove il canale stesso è concentrato. Nella approssimazione di canale graduale si può anche in questo caso operare una notevole semplificazione: il potenziale a cui sono sottoposti gli elettroni si scompone in due componenti ϕ_{ch} e ϕ dipendenti, rispettivamente, dalla sola tensione V_{DS} e V_{GS}:

$$\varphi(x,y,V_{GS},V_{DS}) = \phi_{ch}(x,V_{DS}) + \phi(y,V_{GS}) \tag{5.55}$$

Il potenziale ϕ_{ch} è detto *potenziale di canale*: esso dipende dalla tensione V_{DS} e varia con la posizione x lungo il canale conduttivo. Ad esso è dovuta la componente $\mathcal{E}_x$ del campo elettrico:

$$\mathcal{E}_x = -\frac{d\phi_{ch}}{dx}. \tag{5.56}$$

Viceversa, il potenziale ϕ varia solo nella direzione dell'asse y essendo legato alla applicazione della tensione V_{GS} e ad esso è legato $\mathcal{E}_y$. Grazie alla approssimazione di canale graduale, l'andamento del potenziale elettrostatico $\phi(y)$ in funzione di V_{GS} non dipende dalla applicazione della tensione al drain e si può quindi ricavare direttamente dalla analisi effettuata per il sistema MOS. Per comprendere invece il sognificato fisico del potenziale di canale, si osservi la figura 5.25.

In questa figura si considera il solo effetto della applicazione della tensione V_{DS} che si ripartisce in maniera graduale lungo tutto il percorso ohmico della corrente. crescendo man mano che ci si sposta dal source verso il drain. È come se il canale potesse essere suddiso in tanti tratti elementari di lunghezza infinitesima Δx, o tante "fettine", ciascuna caratterizzata da una sua resistenza elementale R_{ch}. Quando la corrente attraversa la regione ohmica, essa attraversa le sezioni elementari, dando luogo

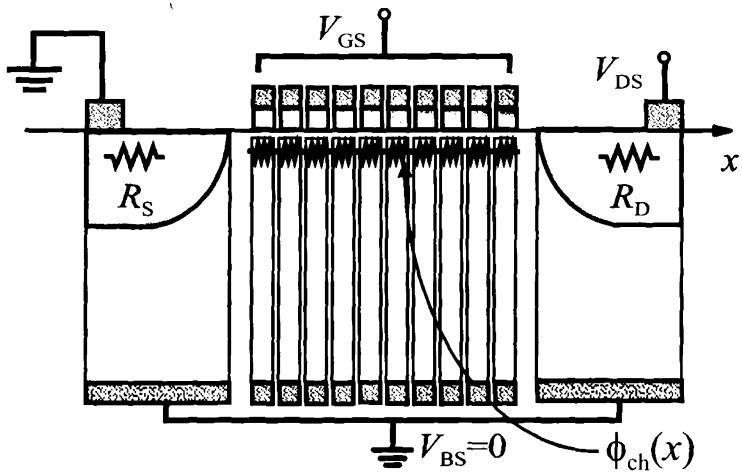


Figura 5.25 Modello distribuito del canale nel transistore *n*MOS.

in ciascuna ad una piccola caduta di tensione. Poiché ciascun elemento è in serie con gli altri, le cadute di tensione si sommano fino a dare origine ad una caduta complessiva di potenziale tra l'inizio e la fine del canale conduttivo, posti rispettivamente in $x = 0$ e $x = L$. Il potenziale che si origina è proprio il potenziale di canale. Ricordando, poi, che il canale è concentrato alla interfaccia tra semiconduttore e ossido, il potenziale ϕ_{ch} è definito solo per $y = 0$. Si osservi che per raggiungere il canale conduttivo gli elettroni entranti dal contatto di source devono percorrere la regione n del source stesso fino al punto iniziale del canale posto in $(x,y) = (0,0)$. Questo tratto di semiconduttore è molto drogato e, sebbene presenti una resistenza serie R_S non nulla al passaggio della corrente di elettroni, questa resistenza è in realtà molto piccola e il suo effetto può essere in prima approssimazione trascurato. Analogamente la resistenza R_D, che corrisponde alla regione di drain che gli elettroni attraversano una volta usciti dal canale conduttivo per raggiungere il contatto di drain stesso, può essere in prima approssimazione trascurata. Allora la caduta di tensione sulla regione di source è approssimativamente nulla e il potenziale di canale nel punto iniziale $x = 0$ corrisponde alla stessa tensione applicata al terminale di source. Nello stesso modo il potenziale di canale in $x = L$, ovvero alla fine del canale, è pari alla tensione applicata al contatto di drain.

$$\phi_{ch}(x = 0) = 0 \qquad \phi_{ch}(x = L) = V_{DS} \tag{5.57}$$

È come se la tensione V_{DS} fosse in realtà applicata, non tanto ai terminali esterni, quanto direttamente a cavallo del canale. Il potenziale ϕ_{ch} varia poi tra i due valori estremi a mano a mano che ci si sposta dal source verso il drain.

Vediamo ora di sommare l'effetto del potenziale di canale a quello del potenziale ϕ dovuto alla presenza della tensione V_{GS}. Ricordiamo innanzitutto dalla figura 5.18 l'andamento del potenziale elettrostatico nella direzione perpendicolare alla interfaccia ossido-semiconduttore nel sistema MOS in forte inversione: all'interfaccia il potenziale superficiale vale $\Phi_S = 2\phi_p$. Adesso, però, in ciascuna sezione del canale conduttivo

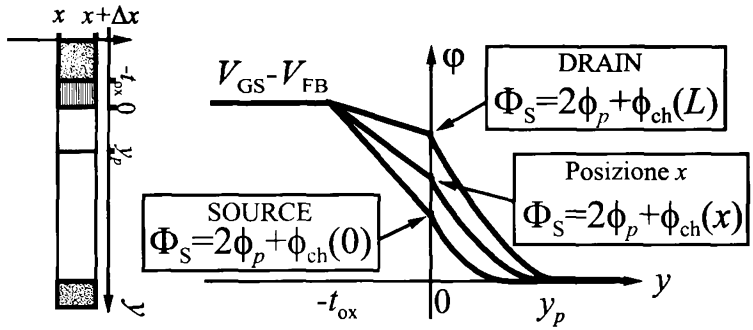


Figura 5.26 Andamento del potenziale elettrostatico in ciascuna sezione del transistore *n*MOS.

gli elettroni sono sottoposti anche alla presenza del potenziale di canale. Il potenziale risultante è mostrato nella figura 5.26, dove si riporta φ in funzione di y per una generica sezione del dispositivo posizionata tra x e $x + \Delta x$. Sono mostrati anche i due casi particolari della prima sezione dalla parte del source ($x = 0$) e dell'ultima sezione dalla parte del drain ($x = L$).

Si osservi che il potenziale superficiale Φ_S cresce dal valore $2\phi_p$ fino al valore $2\phi_p + V_{DS}$ passando dal source al drain. Il potenziale superficiale nel generico punto del canale posizionato alla coordinata x vale

$$\Phi_S = 2\phi_p + \phi_{ch}(x). \tag{5.58}$$

Anche la d.d.p. ai capi dell'ossido varia adesso con la posizione lungo il canale e, nel generico punto x

$$\Phi_{ox} = V_{GS} - V_{FB} - \Phi_S = V_{GS} - V_{FB} - 2\phi_p - \phi_{ch}(x). \tag{5.59}$$

Questa tensione dipende ora sia dalla tensione V_{GS} sia dal potenziale di canale e, quindi, dalla V_{DS}.

Siamo a questo punto in grado di effettuare il calcolo della legge di controllo di carica all'interno del canale del transistore *n*MOS. Si procede in maniera del tutto simile a quanto visto nel caso del sistema MOS, calcolando la carica Q_n per differenza a partire dalla carica Q_d nella regione svuotata di semiconduttore e della carica Q_t sul metallo. Poiché dalla (5.58) il potenziale superficiale varia con la posizione x lungo il canale conduttivo, anche le cariche possono variare con x, e la legge di controllo di carica va quindi valutata separatamente in ciascuna sezione del transistore, come mostrato nella figura 5.27.

Utilizzando la (5.58), la (5.18) si estende come:

$$Q_d(x) = -\sqrt{2q\epsilon_S N_A}\sqrt{2\phi_p + \phi_{ch}(x)} \tag{5.60}$$

Si osservi che Q_d dipende adesso dal potenziale di canale e quindi da V_{DS}. La carica

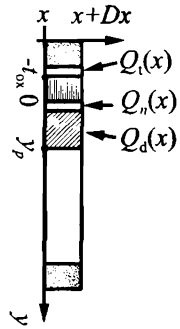


Figura 5.27 Cariche Q_n, Q_d e Q_t nella sezione del transistore posizionata nella coordinata x.

totale alla sezione x, $Q_t(x)$, vale ora

$$Q_t(x) = C_{ox}\Phi_{ox}(x) \tag{5.61}$$

e, sostituendo il valore di Φ_{ox} (5.59), si ottiene

$$Q_t(x) = C_{ox}(V_{GS} - V_{FB} - 2\phi_p - \phi_{ch}(x)) \tag{5.62}$$

Si osservi che, a differenza del caso della carica Q_d, Q_t dipende sia dalla tensione V_{GS} sia dalla tensione V_{DS}.

Infine la carica di elettroni nello strato di inversione si determina dalla condizione di neutralità $Q_n(x) = -Q_t(x) - Q_d(x)$. Sostituendo le espressioni di Q_d (5.60) e Q_t (5.62) si ricava la legge di controllo di carica al variare della posizione x lungo il canale:

$$Q_n(x) = -C_{ox}(V_{GS} - V_{FB} - 2\phi_p - \phi_{ch}(x)) + \sqrt{2q\epsilon_S N_A\,(2\phi_p + \phi_{ch}(x))} \tag{5.63}$$

Ricordando l'espressione della tensione di soglia di un sistema MOS con substrato di tipo p (5.35), l'espressione della carica di canale diventa:

$$Q_n(x) = -C_{ox}(V_{GS} - \phi_{ch} - V_{th}) + \gamma_B C_{ox}\left(\sqrt{2\phi_p + \phi_{ch}} - \sqrt{2\phi_p}\right) \tag{5.64}$$

Il primo addendo corrisponde al *controllo di carica lineare*. Il secondo addendo, non-lineare, è globalmente *proporzionale a* γ_B, ovvero al coefficiente di effetto body, e può essere trascurato se il drogaggio del substrato è sufficientemente basso.

Naturalmente la carica $Q_n(x)$ dipende dalle entrambe le tensioni di pilotaggio. In particolare si osservi che al crescere della tensione di gate V_{GS} si ha un aumento sia della carica Q_t sia della carica Q_n in ogni punto del canale conduttivo, ovvero $\forall x$. Se ne deduce che al crescere di V_{GS} il canale conduttivo è più popolato di elettroni in ogni sua parte.

Se invece si analizza cosa accade al crescere della tensione di drain V_{DS}, si osserva un comportamento molto diverso. Il potenziale $\phi_{ch}(x)$ cresce in ogni punto x del canale e

questo innalzamento è più significativo dalla parte del drain che dalla parte del source, ovvero nella parte terminale del canale conduttivo. Al crescere di $\phi_{ch}(x)$, poi, Q_d aumenta per la (5.60) mentre Q_t diminuisce per la (5.62), portando complessivamente ad una diminuzione (in modulo) della carica Q_n: in pratica la popolazione di elettroni nel canale conduttivo varia con x ed è maggiore dalla parte del source e minore dalla parte del drain. Questo effetto è tanto più marcato quanto più è elevata la tensione V_{DS}. Si dice infatti che il canale conduttivo tende a *strozzarsi* dalla parte del drain per tensioni di drain crescenti. Proprio a questo fenomeno si deve il caratteristico andamento della corrente che scorre nel canale conduttivo in funzione di V_{DS}. Allo studio della corrente del transistore è dedicato il paragrafo 5.3.2.

5.3.2 La corrente di canale

Nota dalla relazione di controllo di carica (5.64) la quantità di carica presente nel canale conduttivo al variare della posizione x, siamo in grado di studiare l'andamento della corrente del transistore nMOS in funzione delle tensioni di pilotaggio. Ricordiamo che in condizioni statiche l'unica corrente non nulla nel dispositivo è la corrente I_{DS} tra il source e il drain: per convenzione essa sarà considerata *positiva entrante* nel contatto di drain. Procederemo prima ad una analisi di tipo qualitativo per poi ricavare esplicitamente la relazione analitica che lega I_{DS} alle tensioni di pilotaggio V_{GS} e V_{DS}: $I_{DS} = I_{DS}(V_{GS},V_{DS})$. Tale relazione cosituisce la cosiddeta *caratteristica tensione–corrente* del transistore nMOS stesso. Si osservi inoltre che se la tensione di gate è al di sotto della tensione di soglia, allora il canale conduttivo non è formato e I_{DS} è comunque identicamente nulla, indipendentemente dal valore della tensione V_{DS}. Quindi:

$$\begin{cases} I_{DS}(V_{GS},V_{DS}) \neq 0 & V_{GS} > V_{th}, \forall\, V_{DS} \neq 0 \\ I_{DS}(V_{GS},V_{DS}) = 0 & V_{GS} \leq V_{th}, \forall\, V_{DS} \end{cases}$$

Come anticipato nel paragrafo 5.3, se la tensione applicata al contatto di drain è positiva, allora la corrente I_{DS} è entrante nel drain e, quindi, con la convenzione adottata, si ha $I_{DS} > 0$. Se si avesse invece $V_{DS} < 0$, si sarebbero scambiati i tra loro i contatti di source e drain: gli elettroni in questo caso attraverserebbero il canale conduttivo dal drain verso il source e si avrebbe una corrente I_{DS} negativa, ovvero entrante nel contatto di source. In realtà, in tutti i casi significativi, è sufficiente limitare l'analisi al solo intervallo $V_{DS} \geq 0$. Infatti, come si è già messo in evidenza nel paragrafo 5.3, il transistore MOS è solitamente una struttura simmetrica, almeno in prima approssimazione, per cui il drain e il source possono essere scambiati a piacere. Da questo punto di vista basta analizzare la corrente per $V_{DS} > 0$, poiché per $V_{DS} < 0$ si avrebbe semplicemente una corrente uguale ma con segno opposto. Se poi il transistore non è simmetrico, allora il contatto di drain è chiaramente identificato e il MOSFET per funzionare correttamente *deve* essere utilizzato in modo che V_{DS} sia positiva e I_D sia effettivamente entrante nel drain.

Analisi qualitativa della corrente di canale

Fissiamo ora la tensione di gate ad un valore costante maggiore della tensione di soglia in modo che il canale conduttivo esista. Aumentiamo poi in maniera graduale la tensione di drain per determinare l'andamento della corrente in funzione di V_{DS}. Ripetendo

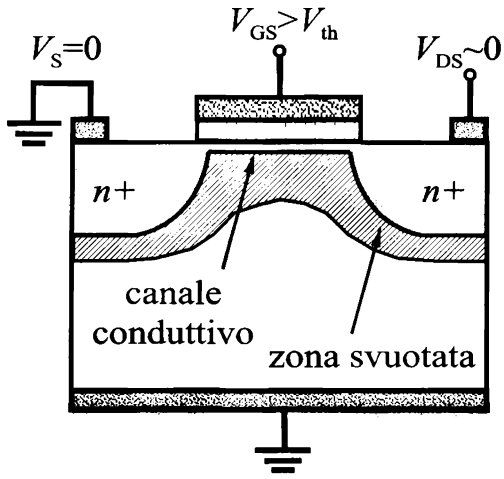


Figura 5.28 Canale conduttivo del transistore *n*MOS ad arricchimento per bassi valori della tensione V_{DS}, ovvero nella regione lineare.

l'analisi per diversi valori di V_{GS} saremo in grado di determinare l'andamento globale delle caratteristiche tensione–corrente del MOSFET. Sia quindi $V_{GS} > V_{th}$ e $V_{DS} > 0$.

Supponiamo un primo momento che la tensione applicata al drain sia diversa da zero ma molto piccola, come rappresentato nella figura 5.28.

In questo caso la tensione V_{DS} induce un potenziale di canale ϕ_{ch} trascurabile poiché $\phi_{ch} \leq V_{DS} \simeq 0$. Utilizzando la (5.64) si osserva che la carica di canale è praticamente indipendente da ϕ_{ch} e, di conseguenza, da V_{DS}: il canale è popolato in maniera pressoché uniforme dal source al drain e il valore della concentrazione di carica dipende unicamente dalla tensione V_{GS}. Ricordando che il canale conduttivo si può paragonare ad una serie di resistenze elementari in serie (cfr. la figura 5.25), allora queste resistenze sono tra loro tutte identiche e in definitiva il canale si comporta come una unica resistenza elettrica. Il transistore si comporta allora come un *resistore* e la corrente I_{DS} che attraversa il canale è proporzionale alla tensione V_{DS} attraverso la *conduttanza di canale* G_{ch}, definita come il reciproco della resistenza del canale:

$$I_{DS} = G_{ch}(V_{GS}) \cdot V_{DS} \tag{5.65}$$

Il valore della conduttanza di canale dipende poi dalla quantità di carica presente nel canale stesso, ovvero è *pilotato dalla tensione* V_{GS}: $G_{ch} = G_{ch}(V_{GS})$. Al crescere di V_{GS} la carica aumenta così come G_{ch}.

L'andamento della corrente in funzione delle tensioni di pilotaggio è quello rappresentato nella Fig. 5.29, dove le diverse curve corrispondono a diversi valori (crescenti) di V_{GS}. Poiché la relazione tra la corrente e V_{DS} è lineare, le curve sono semplicemente delle rette con pendenza variabile in funzione di V_{GS}. La regione di funzionamento del transistore caratterizzata da questo comportamento prende di nome di *regione lineare*.

Si dimostra che nella regione lineare la conduttanza di canale è legata ai parametri

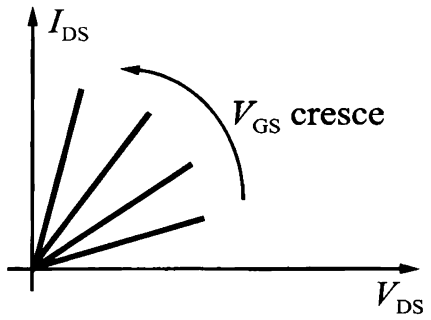


Figura 5.29 Caratteristiche statiche del transistore *n*MOS ad arricchimento per bassi valori della tensione V_{DS}, ovvero nella regione lineare.

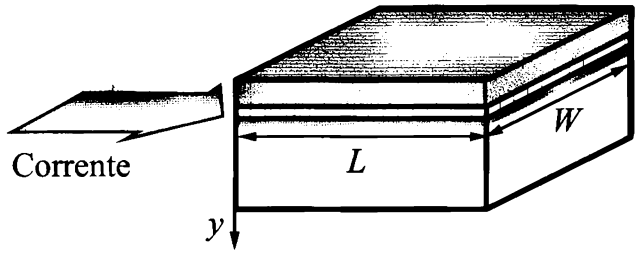


Figura 5.30 Calcolo della conduttanza di canale.

fisici del transistore *n*MOS mediante la relazione:

$$G_{ch} = \frac{W}{L}\mu_n C_{ox}(V_{GS} - V_{th}) \tag{5.66}$$

dove W e L sono rispettivamente la larghezza e la lunghezza del contatto di gate (cfr. la figura 5.30). Questa relazione è ricavata esplicitamente nell'approfondimento 5.3.

Approfondimento 5.3 Osservando la figura 5.30, si osserva che la corrente del transistore è concentrata in uno strato laminare, parallelo alla interfaccia tra il semiconduttore e l'ossido, di larghezza pari alla dimensione trasversale del contatto di gate, indicata con W in figura, e lunghezza pari alla lunghezza del gate L. Ricordando che in una regione di tipo n la conducibilità vale $\sigma_n = qn\mu_n$ (cfr. la (2.11)), la conduttanza dello strato carico è:

$$G_{ch} = \frac{W}{L}\int_0^{W_p} \sigma_n \mathrm{d}y = q\frac{W}{L}\mu_n \int_0^{W_p} n \, \mathrm{d}y = \frac{W}{L}\mu_n |Q_n| \tag{5.67}$$

dove si è utilizzata la definizione di Q_n della equazione (5.7). Si sostituisce poi a Q_n la legge c controllo di carica (5.64) che per $\phi_{ch} \approx 0$ ($V_{DS} \approx 0$) si riduce a $Q_n = -C_{ox}(V_{GS} - V_{th})$, ottenenc: infine l'espressione (5.66) della conduttanza di canale .

Supponiamo ora di aumentare la tensione V_{DS}. Al crescere della tensione drain-source la caduta di potenziale lungo il canale conduttivo diventa non trascurabile -

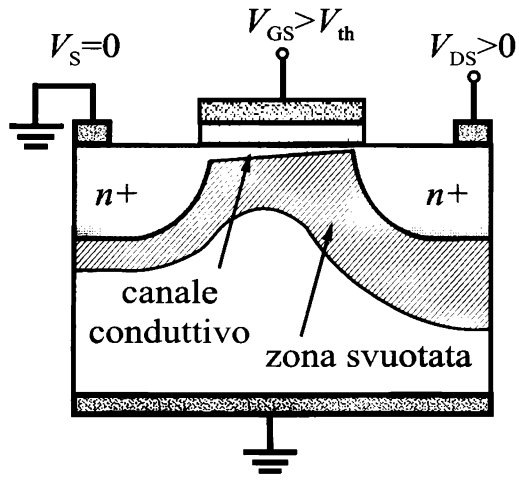


Figura 5.31 Canale conduttivo del transistore *n*MOS ad arricchimento nella regione quadratica.

di conseguenza il potenziale di canale ϕ_{ch} aumenta dal source al drain. Secondo la relazione (5.64), allora, la carica di elettroni presente nel canale diminuisce passando dal source al drain, ovvero il canale tende a strozzarsi dalla parte del drain. Nella figura 5.31 questo è mostrato con uno spessore maggiore della linea che rappresenta il canale al source rispetto al drain: si noti che si tratta solo di un espediente grafico, poiché il canale conduttivo non ha in ogni caso spessore, essendo concentrato in uno strato laminare all'interfaccia tra semiconduttore e ossido. Poiché inoltre il substrato è posto al riferimento di potenziale, all'aumentare della tensione di drain la giunzione *pn* formata tra il drain e il substrato è polarizzata sempre più inversamente e la regione svuotata a cavallo di tale giunzione si estende sempre di più all'interno del substrato, come mostrato nella figura 5.31.

Per quanto riguarda la corrente, si osserva che poiché il canale si strozza al drain, la resistenza equivalente del canale è maggiore dalla parte del drain e la resistenza complessiva del canale aumenta rispetto al caso lineare. Le caratteristiche del transistore sono ora raffigurate nella figura 5.32, dove si osserva che a parità di tensione applicata V_{GS} si ha ora un aumento meno che lineare della corrente di drain al crescere di V_{DS}. Si osservi però che il canale rappresenta comunque una regione di tipo ohmico, percorsa da una corrente di trascinamento di elettroni sotto l'azione del campo elettrico longitudinale al canale. Tale campo elettrico, poi, è più intenso nella parte terminale del canale (ovvero al drain) poiché per definizione esso è, a meno del segno, la derivata del potenziale di canale. Tale derivata, a sua volta, cresce dal source al drain, poiché la caduta di potenziale è maggiore dove la resistenza elettrica del canale è maggiore, ovvero nella parte del canale più strozzata. Dalla figura 5.32 si osserva che la corrente di drain ha ora un andamento di tipo parabolico, con concavità rivolta verso il basso, in funzione di V_{DS}[17]: per questo motivo questa regione di funzionamento è detta *regione quadratica* o anche, per via della crescita meno che lineare, *regione di ginocchio.*

[17] Questo risultato verrà dimostrato rigorosamente nel paragrafo seguente.

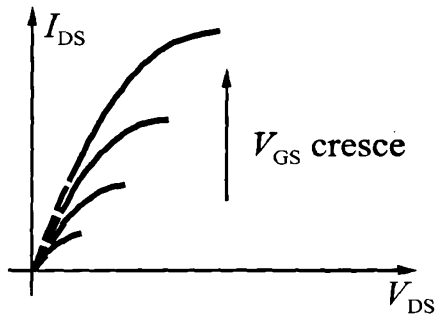


Figura 5.32 Caratteristiche statiche del transistore *n*MOS ad arricchimento nella regione quadratica.

Aumentando ancora la tensione di drain si raggiunge una condizione del tutto nuova, rappresentata nella figura 5.33 (a). Ad una ben precisa tensione V_{DS}, detta *tensione di saturazione* V_{DSS}, il canale conduttivo si strozza completamente nell'ultimo punto dalla parte del drain[18]. Nello stesso punto la carica Q_n si annulla completamente ed il canale non presenta più carica libera. La tensione di saturazione può essere calcolata proprio imponendo nella legge di controllo di carica (5.64) $Q_n = 0$ nel punto $x = L$, ovvero nel punto terminale del canale conduttivo dalla parte del drain:

$$Q_n(x = L) = -C_{ox}(V_{GS} - \phi_{ch}(x = L) - V_{th}) = 0 \tag{5.68}$$

Poiché $\phi_{ch}(x = L) = V_{DS}$ e, allo strozzamento del canale, $V_{DS} = V_{DSS}$, la relazione precedente comporta:

$$V_{GS} - V_{DSS} - V_{th} = 0 \tag{5.69}$$

e in definitiva:

$$V_{DSS} = V_{GS} - V_{th} \tag{5.70}$$

Si noti che lo strozzamento avviene a tensioni V_{DSS} più alte al crescere di V_{GS}.

Il punto di strozzamento rappresenta una regione svuotata ad alto campo elettrico. ma l'annullarsi della carica in un solo punto del canale non significa che sia nulla anche la corrente all'interno dell'intero canale conduttivo. Al contrario, il canale conduttivo è ora diviso in una parte ohmica, dove la carica Q_n è non nulla e in cui si ha corrente di trascinamento di elettroni, ed una parte strozzata di tipo non ohmico. La corrente nella parte non ohmica non è comunque necessariamente nulla, anzi, si può avere una iniezione di carica in una regione svuotata ad alto campo elettrico, come ad esempio nel caso della corrente che attraversa la regione svuotata di una giunzione *pn*. Del resto si ricordi che per l'equazione di continuità della corrente di elettroni, in condizioni stazionarie e in assenza di fenomeni GR, si ha:

$$\frac{dJ_n}{dx} = 0 \tag{5.71}$$

[18] Sebbene dal punto di vista matematico si possa parlare di un singolo punto del canale, dal punto di vista fisico si tratta comunque di una porzione, benché di dimensioni estremamente piccole, del canale stesso, come mostrato nell'ingrandimento della figura 5.33 (a).

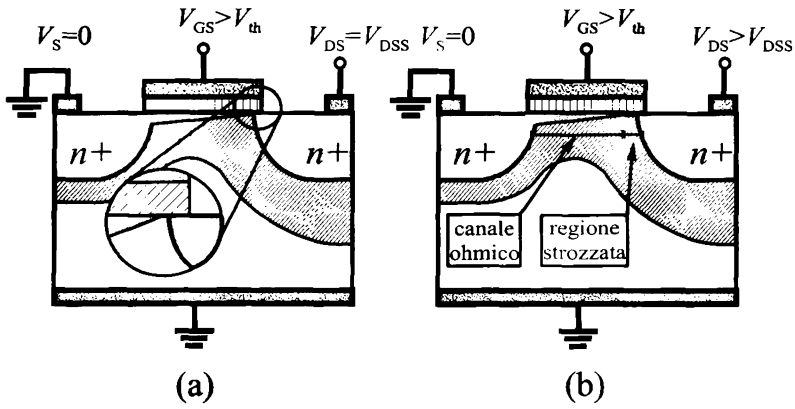


Figura 5.33 Canale conduttivo del transistore *n*MOS ad arricchimento alla saturazione.

ovvero la densità di corrente di elettroni deve essere costante in ogni punto x del canale. La corrente non può quindi essere diversa da zero nella parte del canale non strozzata e nulla altrove. Essa è invece non nulla in tutto il canale: quello che cambia è solo la sua natura che al variare della posizione nel canale può essere di tipo ohmico o meno. In pratica gli elettroni lasciano il source, attraversano la regione ohmica del canale raggiungendo il punto di strozzamento e qui vengono iniettati nella regione del drain mantenendo inalterato il flusso complessivo della corrente nel canale.

Al crescere della tensione di drain al di sopra del valore di saturazione ($V_{DS} > V_{DSS}$) la situazione è rappresentata nella figura 5.33 (b). Il canale conduttivo si divide nella regione ohmica e nella parte strozzata altamente resistiva. Quest'ultima del resto ha dimensioni comunque estremamente piccole e in prima approssimazione la lunghezza della parte di canale non strozzata rimane inalterata. La tensione di drain applicata in eccesso al valore V_{DSS} (ovvero, $V_{DS} - V_{DSS}$) cade prevalentemente nella piccola porzione di canale strozzata, priva di carica e quindi molto resistiva, aumentandone il campo elettrico, mentre la tensione sulla parte ohmica rimane approssimativamente pari a V_{DSS}, indipendentemente dalla V_{DS} applicata. In pratica, una volta che il canale si strozza per $V_{DS} = V_{DSS}$, anche se si aumenta ancora V_{DS} la d.d.p. sulla parte ohmica di canale rimane come congelata al valore V_{DSS}. Di conseguenza, anche la corrente che caratterizza la parte ohmica del canale rimane inalterata e, poiché come detto, la corrente nel canale è costante in ogni punto, allora la corrente I_{DS} rimane fissa al valore raggiunto per $V_{DS} = V_{DSS}$. In altre parole, quando $V_{DS} > V_{DSS}$ la corrente I_{DS} *non dipende più della tensione* V_{DS}, ma dipende solo dalla tensione V_{GS}. Questa regione di funzionamento è detta *regione di saturazione* e le caratteristiche del transistore *n*MOS in questa regione sono rette orizzontali, come mostrato nella figura 5.34.

La regione di saturazione è per molti aspetti la regione di funzionamento più importane del transistore. Nelle tipiche applicazioni in cui il transistore è utilizzato come doppio bipolo, si sfrutta proprio la caratteristica del MOSFET di avere la corrente alla porta di uscita (di solito la porta drain-source) pilotata unicamente dalla tensione alla porta di ingresso (normalmente la porta gate-source) *indipendentemente dalla tensione alla porta di uscita.*

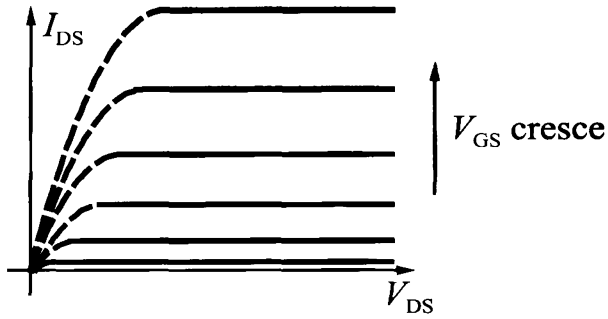


Figura 5.34 Caratteristiche statiche del transistore *n*MOS ad arricchimento nella regione di saturazione.

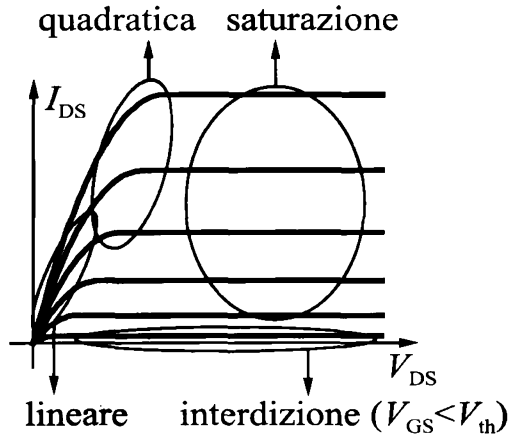


Figura 5.35 Regioni di funzionamento del transistore *n*MOS ad arricchimento..

Quanto detto conclude la analisi qualitativa della corrente nel transistore *n*MOS ad arricchimento. Si sono illustrate le regioni di funzionamento lineare, quadratica e di saturazione. A queste va ovviamente aggiunta la regione di interdizione, in cui $V_{GS} < V_{th}$ e la corrente è identicamente nulla. Si noti infine che nel transistore *n*MOS la corrente I_{DS} è comunque una *funzione crescente di* V_{GS}. Le regioni di funzionamento sono riassunte nella figura 5.35.

Valutazione analitica della corrente di canale

Per determinare la relazione analitica di I_{DS} in funzione delle tensioni di pilotaggio V_{GS} e V_{DS}, si valuterà la corrente nella parte non strozzata del canale, ovvero nella regione ohmica, dove questa è puramente corrente di trascinamento di elettroni[19]. Poiché la densità di corrente di trascinamento degli elettroni è $J_{DS} = qn\mu_n\mathcal{E}_x$, la corrente I_{DS} si

[19] Sebbene le correnti di diffusione siano sempre presenti all'interno di qualsiasi dispositivo a semiconduttore, nel funzionamento del transistore MOSFET in conduzione queste sono generalmente trascurabili.

calcola come l'integrale di J_{DS} nella sezione laminare attraverso cui la corrente scorre (ovvero moltiplicando per la dimensione trasversale W e integrando nella direzione verticale lungo y; si faccia ancora riferimento alla figura 5.30):

$$I_{DS} = -W\mu_n \left(q \int_0^{W_p} n \, dy \right) \mathcal{E}_x = W\mu_n Q_n \mathcal{E}_x \tag{5.72}$$

Nella relazione precedente il primo segno − è legato al fatto che per la nostra convenzione la densità di corrente (positiva entrante nel drain) è in effetti antiparallela all'asse x. Si osservi che Q_n e $\mathcal{E}_x$ variano lungo il canale (dipendono da x), ma la corrente I_{DS} come già spiegato è invece costante lungo il canale. Ricordando che per definizione:

$$\mathcal{E}_x = -\frac{d\phi_{ch}}{dx} \tag{5.73}$$

si ricava:

$$I_{DS} = -W\mu_n Q_n \frac{d\phi_{ch}}{dx} \tag{5.74}$$

Integriamo ora entrambi i membri lungo il canale tra $x = 0$ e $x = L$:

$$\int_0^L I_{DS} dx = -W\mu_n \int_0^L Q_n \frac{d\phi_{ch}}{dx} dx \tag{5.75}$$

dove L è la lunghezza del contatto di gate. Nel secondo integrale si possono cambiare la variabile e gli estremi di integrazione utilizzando le relazioni $\phi_{ch}(x = 0) = 0$ e $\phi_{ch}(x = L) = V_{DS}$:

$$\int_0^L I_{DS} dx = -W\mu_n \int_0^{V_{DS}} Q_n \, d\phi_{ch} \tag{5.76}$$

Poiché I_{DS} è costante, a primo membro ottiene semplicemente $I_{DS} \cdot L$.

Sostituendo infine a Q_n l'espressione esplicita in funzione di ϕ_{ch} della relazione di controllo di carica (5.64), l'integrale si risolve in modo analitico:

$$I_{DS} = \frac{W}{L}\mu_n C_{ox} \left[(V_{GS} - V_{th}) V_{DS} - \frac{1}{2} V_{DS}^2 \right] + \\ - \gamma_B \frac{W}{L}\mu_n C_{ox} \left\{ \frac{2}{3} \left[\sqrt{(2\phi_p + V_{DS})^3} - \sqrt{(2\phi_p)^3} \right] - \sqrt{2\phi_p} V_{DS} \right\}$$

Si osservi che il primo addendo deriva dalla parte lineare della relazione di controllo di carica e fornisce un contributo di tipo quadratico (parabolico) della corrente I_{DS} in funzione di V_{DS}. Il secondo addendo è invece proporzionale a γ_B e può quindi venire trascurato se il substrato ha drogaggio sufficientemente basso. È molto comune omettere completamente il secondo addendo, tanto che spesso la corrente del MOSFET riportata nei libri di testo si limita alla sola parte quadratica, trascurando il fatto che nei casi in cui il coefficiente di effetto body sia elevato, la parte non quadratica può

comunque risultare significativa. Sapendo che si tratta di una semplificazione, comunque, anche in qesto testo nel seguito si utilizzerà prevalentemente la sola caratteristica quadratica.

Ovviamente la relazione trovata

$$I_{DS} = \frac{W}{L}\mu_n C_{ox}\left[(V_{GS} - V_{th})\,V_{DS} - \frac{1}{2}V_{DS}^2\right] \tag{5.77}$$

non ha comunque valore nella regione di saturazione per $V_{DS} > V_{DSS}$, ovvero per $V_{DS} > V_{GS} - V_{th}$: a tale tensione il canale è strozzato in $x = L$ e di conseguenza la tensione che cade sulla parte ohmica del canale non è più pari a V_{DS}, bensì a V_{DSS}, e l'integrale nella (5.76) è mal definito negli estremi di integrazione. Se si utilizzasse la (5.77) si otterrebbe un risultato assurdo, ovvero che la corrente di drain decrescerebbe al crescere della tensione V_{DS}[20] mentre si è visto che nella regione di saturazione la corrente I_{DS} rimane costante, saturando al valore raggiunto quando V_{DS} è esattamente pari a V_{DSS}. Tale valore prende il nome di *corrente di saturazione* I_{DSS} e si calcola sostituendo $V_{DS} = V_{DSS} = V_{GS} - V_{th}$ nella (5.77), ottenendo:

$$I_{DSS} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\,(V_{GS} - V_{th})^2 \tag{5.78}$$

In definitiva la relazione analitica per il calcolo della corrente I_{DS} deve essere divisa a seconda che ci si trovi a tensioni di drain inferiori (regione lineare e quadratica) o superiori (regione di saturazione) al valore V_{DSS}:

$$I_{DS} = \begin{cases} \frac{W}{L}\mu_n C_{ox}\left[(V_{GS} - V_{th})\,V_{DS} - \frac{1}{2}V_{DS}^2\right] & V_{GS} > V_{th} \text{ e } V_{DS} < V_{DSS} \\ \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\,(V_{GS} - V_{th})^2 & V_{GS} > V_{th} \text{ e } V_{DS} > V_{DSS} \end{cases} \tag{5.79}$$

Si noti anche che per valori di tensioni di drain molto piccoli ($\lim_{V_{DS}\to 0} I_{DS}$) dalla relazione precedente si riottiene la relazione (5.65) tipica della regione lineare.

Le caratteristiche statiche del nMOS (5.79) forniscono un legame esplicito della corrente in funzione delle tensioni di pilotaggio. Quando si voglia disegnare in un grafico la (5.79), è necessario in realtà disegnare un insieme, più propriamente una famiglia, di curve. Ad esempio si può disegnare la corrente in funzione della tensione V_{DS}, disegnando una curva diversa per ogni valore di V_{GS}, come già mostrato nella figura 5.35. In questo caso si hanno le cosiddette *caratteristiche di uscita* del transistore: il nome deriva dal fatto che la corrente I_{DS} e la tensione V_{DS} sono spesso le grandezze che caratterizzano la porta di uscita del MOSFET utilizzato come doppio bipolo. Quando però il transistore lavora nella regione di saturazione, la corrente di drain non dipende più da V_{DS} ed è allora più interessante disegnarla in *funzione della tensione* V_{GS}. La relazione che lega I_{DS} a V_{GS} non è altro che la (5.78): essa è anche detta *transcaratteristica statica* del MOSFET poiché fornisce la caratteristica di trasferimento tra la porta di ingresso (gate-source) e la porta di uscita (drain-source). Si osservi che la transcaratteristica è una funzione paraboidale (quadratica) della variabile V_{GS}: il minimo si raggiunge per $V_{GS} = V_{th}$ dove la corrente si annulla mentre il solo ramo della

[20] Si noti infatti che per $V_{DS} = V_{GS} - V_{th}$ la parabola della (5.77) raggiunge il suo valore massimo.

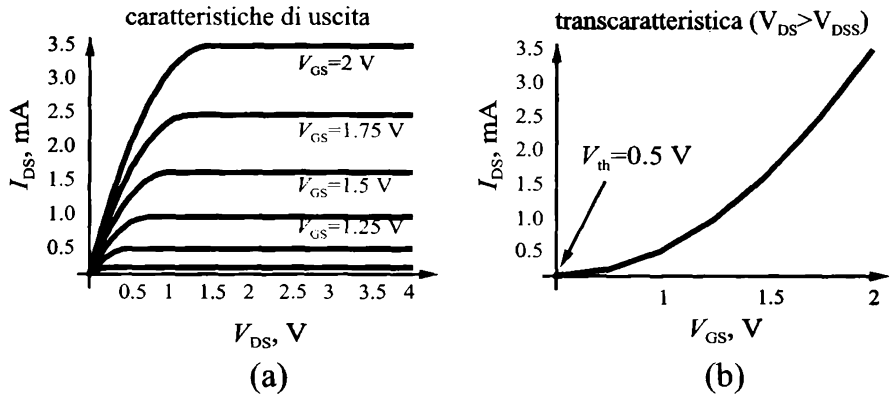


Figura 5.36 Caratteristiche di uscita e transcaratteristica statica del transistore *n*MOS ad arricchimento.

parabola corrispondente a $V_{GS} > V_{th}$ ha significato fisico. È interessante osservare che dal grafico della transcaratteristica è possibile determinare immediatamente il valore della tensione di soglia come quel valore per cui la corrente si annulla. Un esempio di caratteristiche di uscita e transcaratteristica di un *n*MOS ad arricchimento è riportato nella figura 5.36.

È consuetudine introdurre il fattore $\beta_n = \frac{W}{L}\mu_n C_{ox}$ (detto semplicemente β del MOS) per cui le caratteristiche si portano nella forma più compatta:

$$I_{DS} = \begin{cases} \beta_n \left[(V_{GS} - V_{th})\, V_{DS} - \frac{1}{2}V_{DS}^2\right] & V_{GS} > V_{th} \text{ e } V_{DS} < V_{DSS} \\ \frac{1}{2}\beta_n (V_{GS} - V_{th})^2 & V_{GS} > V_{th} \text{ e } V_{DS} > V_{DSS} \end{cases} \tag{5.80}$$

In questa forma si mette in evidenza il fatto che la corrente dipende, oltre che dalle tensioni di pilotaggio, dal fattore β_n moltiplicativo, che è quindi un fattore di merito del transistore. Infatti, a parità di tensioni applicate, se si massimizza β_n la corrente del transistore è maggiore e di conseguenza il comportamento del dispositivo migliora. Il fattore β_n dipende *unicamente dalla tecnologia di fabbricazione* ed è costituito dal fattore $\frac{W}{L}$, detto *fattore di forma*, e da $\mu_n C_{ox}$, detto *fattore tecnologico*. Esso è quindi massimizzato se

- ▷ L decresce (*canale corto*)
- ▷ C_{ox} cresce, ovvero t_{ox} decresce (*ossido sottile*)

Le tecnologie di fabbricazione dei dispositivi MOS a semiconduttore sono guidate costantemente da questi due requisiti: attualmente lo spessore dell'ossido di gate è dell'ordine di qualche decina di angstrom mentre le dimensioni minime del contatto di gate sono dell'ordine di qualce decina di nanometri.

Esempio 5.3 Il fattore β_n del transistore può essere ricavato note le dimensioni e i parametri di fabbricazione del dispositivo. Molto spesso alcuni parametri quali la mobilità degli elettroni

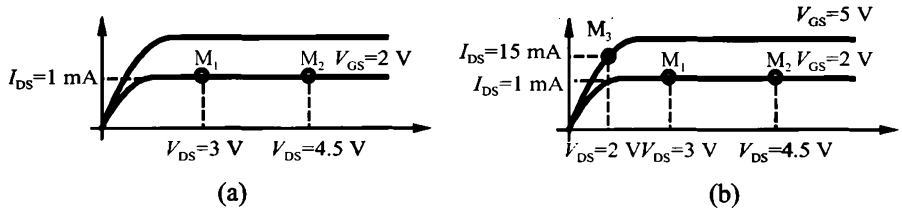


Figura 5.37 Misure effettuate sul dispositivo *n*MOS dell'esempio 5.3.

nel canale conduttivo o la tensione di soglia non sono noti se non in maniera approssimativa e si preferisce allora estrarre queste informazioni a partire da misure. Si consideri a questo fine il seguente esercizio.
Su un dispositivo MOSFET a canale n sono eseguite tre misure riassunte nella seguente tabella

$$\begin{array}{llll} \mathrm{M_1}\text{:} & I_{DS1} = 1\ \mathrm{mA} & V_{GS1} = 2.0\,\mathrm{V} & V_{DS1} = 4.5\,\mathrm{V} \\ \mathrm{M_2}\text{:} & I_{DS2} = 1\ \mathrm{mA} & V_{GS2} = 2.0\,\mathrm{V} & V_{DS2} = 3.0\,\mathrm{V} \\ \mathrm{M_3}\text{:} & I_{DS3} = 15\,\mathrm{mA} & V_{GS3} = 5.0\,\mathrm{V} & V_{DS3} = 2.0\,\mathrm{V} \end{array}$$

Sapendo che la mobilità degli elettroni nel canale è $\mu_n = 1400\,\mathrm{cm^2/Vs}$ e $W/L = 1$, calcolare

- ▷ la tensione di soglia
- ▷ lo spessore dell'ossido del dispositivo

Osserviamo innanzitutto che le prime due misure sono in regione di saturazione poiché I_{DS} è costante a parità di V_{GS} e per diversi valori di VDS. La figura 5.37 (a) riporta i due punti misurati sulla caratteristica di uscita del transistore corrispondente a $V_{GS} = 2$ V.
La terza misura, invece, non può essere in saturazione poiché se così fosse dovrebbe essere

$$V_{DS3} > V_{GS3} - V_{th} \Rightarrow V_{th} > 3\,\mathrm{V}$$

in contrasto con le prime due misure: infatti per $V_{GS1} = 2\,\mathrm{V}$ le prime due misure riportano una corrente non nulla e quindi il transistore è sopra soglia, ovvero $V_{th} < 2\,\mathrm{V}$.
Ricordando ora le caratteristiche (5.79), è possibile sostituire i valori della misura $\mathrm{M_1}$ nella espressione della corrente in saturazione e la misura $\mathrm{M_3}$ in quella della corrente nella regione quadratica, ottenendo il seguente sistema di equazioni:

$$\begin{cases} I_{DS1} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\left(V_{GS1} - V_{th}\right)^2 \\ I_{DS3} = \frac{W}{L}\mu_n C_{ox}\left[\left(V_{GS3} - V_{th}\right)V_{DS3} - \frac{1}{2}V_{DS3}^2\right] \end{cases}$$

Posto $\alpha = I_{DS1}/I_{DS3}$, dividendo membro a membro si ottiene una equazione di secondo grado in V_{th}:

$$V_{th}^2 - 2\left(V_{GS1} - \alpha V_{DS3}\right)V_{th} + \left[V_{GS1}^2 + \alpha\left(V_{DS3}^2 - 2V_{DS3}V_{GS3}\right)\right] = 0$$

Delle due soluzioni

$$V_{th1} = 1.12\,\mathrm{V} \quad \text{e} \quad V_{th2} = 2.61\,\mathrm{V}$$

solo V_{th1} è accettabile, poiché come si è detto deve essere $V_{th} < 2\,\mathrm{V}$.
Si verifica che effettivamente $\mathrm{M_1}$ e $\mathrm{M_2}$ sono in saturazione poiché $V_{DS1,2} > V_{GS1,2} - V_{th} = 0.88$ V; al contrario la misura $\mathrm{M_3}$ non è in saturazione poiché $V_{DS3} < V_{GS3} - V_{th} = 3.88$ V.

Ora dalla condizione

$$I_{DS1} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\left(V_{GS1} - V_{th}\right)^2$$

noti V_{th}, $\frac{W}{L}$ e μ_n si ricava direttamente C_{ox}:

$$C_{ox} = \frac{2\,I_{DS1}}{\mu_n \frac{W}{L}\left(V_{GS1} - V_{th}\right)^2} = 0.0186\ \mathrm{F/m^2}$$

Infine, supponendo che nel disposisitivo l'ossido sia SiO_2, si calcola t_{ox}:

$$t_{ox} = \frac{\epsilon_{ox}}{C_{ox}} = 185.5\ \mathrm{nm}$$

Esempio 5.4 Per comprendere l'effetto delle dimensioni fisiche e della tecnologia sul funzionamento dei transistori MOS si consideri questo problema.
Un transistore MOS a canale *n* con gate in *poly* costruito in tecnologia integrata ha dimensioni minime imposte dalla tecnologia di fabbricazione e in particolare dalla fotolitografia. Se per effetto di un miglioramento della tecnologia le dimensioni del dispositivo vengono scalate, ovvero ridotte, di un fattore K, allora nella stessa superficie del circuito integrato occupata dal dispositivo originario sarà possibile realizzare K^2 dispositivi più piccoli. Ci si chiede se questo scalamento non comporti problemi nel funzionamento del circuito integrato stesso. Se infatti come si è visto la riduzione della lunghezza di gate consente di ottenere, a parità di tensioni applicate, delle correnti maggiori, è anche vero che la potenza dissipata per unità di superficie nel circuito integrato aumenta, data la maggiore densità di integrazione, ovvero il maggiore numero di dispositivi presenti per unità di area. Si pone quindi il problema di come sfruttare al meglio le potenzialità dei nuovi dispositivi riscalati senza incorrere in un eccesso di dissipazione di potenza con conseguente degrado di prestazioni dei dispositivi stessi se non, addirittura, la loro rottura.
A questo tipo di problematiche rispondono i cosiddetti *schemi di riscalamento*, che si occupano di determinare in che modo i dispositivi debbano essere riscalati e, in caso, come siano da riscalare anche le tensioni di pilotaggio dei dispositivi stessi. A titolo di esempio si consideri il seguente *case-study*, che costituisce un semplice esempio del cosiddetto riscalamento a *campo costante*.
Un *n*MOS è realizzato in un circuito integrato alimentato con tensione V_{DD} è caratterizzato da:

- $N_A = 5 \times 10^{16}\ \mathrm{cm^{-3}}$, $L = 6\,\mu\mathrm{m}$, $W = 4\,\mu\mathrm{m}$, $t_{ox} = 50\,\mathrm{nm}$, $\mu_n = 1417\,\mathrm{cm^2/Vs}$
- $V_{DD} = 5\,\mathrm{V}$

Esso viene riscalato del fattore $K = 2$ in modo che:

- $N'_A = K \times N_A$, $L' = L/K$, $W' = W/K$, $t'_{ox} = t_{ox}/K$
- $V'_{DD} = V_{DD}/K$

Si osservi che per il dispositivo più piccolo (riscalato), si scalano oltre alle dimensioni W e L, anche la tensione di alimentazione e il drogaggio del substrato. Si discuta come varia la massima potenza dissipata per unità di area[21].
Prima del riscalamento
Seguendo la falsariga del calcolo della tensione di soglia di un sistema MOS su substrato di tipo p e gate in *poly* mostrata nell'esempio 5.1, si ottengono i seguenti valori

[21] Poiché si riscalano dello stesso fattore sia le dimensioni geometriche sia le tensioni, allora i campi elettrici interni al dispostivo saranno almeno in prima approssimazione costanti, da cui si spiega il nome di questo schema di scalamento.

$$V_{FB} = -0.98\,\text{V}$$
$$\phi_p = 0.39\,\text{V}$$
$$C_{ox} = 6.9 \times 10^{-8}\,\text{F/cm}^2$$
$$Q_d = -1.14 \times 10^{-7}\,\text{C/cm}^2$$
$$\gamma_B = 1.87$$

Utilizzando questi valori nella espressione della tensione di soglia (5.35) si ricava $V_{th} = 1.44\,\text{V}$. Per valutare la potenza massima dissipata si scelgono le massime tensioni di pilotaggio del dispositivo. Essedo il circuito alimentato alla tensione V_{DD}, si deve imporre $V_{DS} = V_{GS} = V_{DD}$. Poiché in questa condizione $V_{DS} = V_{GS} > V_{GS} - V_{th}$, il MOS è in saturazione e la corrente vale:

$$I_{DS} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\left(V_{DD} - V_{th}\right)^2 = 413\,\mu\text{A}$$

La massima potenza dissipata dal dispositivo *prima* dello scalamento è allora:

$$P_D = I_{DS} \times V_{DD} = 2.1\,\text{mW}$$

Dopo il riscalamento

Procedendo in maniera analoga al caso precedente, per il dispositivo riscalato si ha:

$$V_{FB} = -0.999\,\text{V}$$
$$\phi_p = 0.41\,\text{V}$$
$$C_{ox} = 13.8 \times 10^{-8}\,\text{F/cm}^2$$
$$Q_d = -1.65 \times 10^{-7}\,\text{C/cm}^2$$
$$\gamma_B = 1.32$$

e la tensione di soglia risulta $V'_{th} = 1.03\,\text{V}$. Si noti che la tensione di soglia *non scala del fattore* K, rompendo così la simmetria nel riscalamento delle dimensioni fisiche e delle tensioni del circuito. Per valutare la massima potenza dissipata dal dispositivo riscalato si impone anche in questo caso

$$V'_{DS} = V'_{GS} = V'_{DD} = 2.5\,\text{V}$$

per cui la corrente (in saturazione) è:

$$I'_{DS} = \frac{1}{2}\frac{W'}{L'}\mu_n C'_{ox}\left(V'_{DD} - V'_{th}\right)^2 = 141\,\mu\text{A}$$

e la potenza dissipata vale:

$$P'_D = I'_{DS} \times V'_{DD} = 0.35\,\text{mW}$$

Si osservi che:

$$P'_D < \frac{P_D}{4}$$

Poiché in questo caso la potenza dissipata da ciascun dispositivo riscalato è minore di $1/K^2$ volte quella del dispositivo originario, allora K^2 dispositivi avranno una potenza dissipata non superiore, anzi di poco inferiore, a quella dissipata dal dispositivo originario. In questo caso quindi lo scalamento del dispositivo rappresenta uno schema plausibile che non comporta rischi nel funzionamento del circuito integrato. Si noti però che lo scalamento proposto comporta anche un certo numero di problematiche, quali ad esempio il fatto che il drogaggio del substrato deve venir aumentato (aumentando l'effetto di substrato), e lo spessore dell'ossido deve essere diminuito. Inoltre scalare drasticamente la tensione di alimentazione del circuito consente di abbattere la potenza dissipata ma diminuisce i margini di rumore delle porte logiche. È chiaro quindi che occorrono strategie di scalamento più raffinate del semplice esempio qui riportato.

5.4 Il transistore *n*MOS a svuotamento

Il transistore *n*MOS a svuotamento si differenzia dal corrispondente *n*MOS ad arricchimento unicamente per il valore della tensione di soglia, che è negativa invece che positiva. Per il resto, il comportamento del transistore è del tutto analogo a quanto visto per il dispositivo ad arricchimento, ed in particolare le caratteristiche statiche hanno forma esattamente identica alla (5.79).

Poiché si è visto che la tensione di soglia di un sistema MOS su substrato di tipo *p* dipende unicamente dai parametri fisici dei materiali utilizzati e dalle caratteristiche geometriche (ad esempio lo spessore dell'ossido di gate), il problema di come fissare la tensione di soglia di un dispositivo ad un valore predeterminato (positivo o, come in questo caso, negativo) è eminentemente di tipo tecnologico[22]. Esso riveste importanza non solo per la fabbricazione di dispositivi di tipo depletion, ma anche in generale per regolare la tensione di soglia dei dispositivi enhancement ad un preciso valore assegnato. Poiché non è praticabile la soluzione di scegliere materiali o spessori dell'ossido in maniera arbitraria, ci si chiede se dal punto di vista tecnologico esista qualche altra possibilità per variare V_{th}. La soluzione a questo problema si è avuta mediante lo sviluppo della tecnica della impiantazione ionica di drogante. Essa è stata via via perfezionata fino a rendere possibile impiantare strati di drogante in quantità ben controllata e con range di penetrazione molto ridotti. In questo modo è possibile deporre uno strato molto sottile di drogante posto esattamente al di sotto della interfaccia tra l'ossido e il semiconduttore, come mostrato nella figura 5.38. Nel caso in cui il drogaggio impiantato sia di tipo *n* (figura 5.38 (a)), allora è come se si *preformasse un canale conduttivo di elettroni* al di sotto dell'ossido di gate: se il drogaggio è sufficientemente intenso, in questo dispositivo anche in assenza di tensione applicata al gate, il passaggio di corrente dal source al drain è possibile mediante la sola applicazione della tensione drain-source. Quando la dose impiantata è sufficiente ed il canale conduttivo è preformato si ha il tipico dispostivo *n*MOS a svuotamento. Esso è "acceso", in grado di condurre, per $V_{GS} = 0$, e per questo motivo è anche detto *normalmente on*. Al crescere di V_{GS} il canale conduttivo si popolerà sempre più di elettroni poiché, oltre a quelli del canale pre–deposto, si andranno ad aggiungere quelli indotti mediante il meccanismo della inversione di popolazione del sistema MOS. Per spegnere il dispositivo è invece necessario che la tensione di gate diventi negativa, in modo da contrastare (svuotare, da cui il nome a svuotamento o depletion) il canale preformato portando il sistema MOS fuori dalla regone di inversione. È quindi evidente che in questa struttura la tensione di soglia del dispositivo è negativa.

Si noti che più in generale la tecnica della impiantazione di drogante *n* consente di regolare ad un valore desiderato la tensione di soglia anche per un dispositivo di tipo enhancement: infatti, se anche la dose impiantata non è sufficiente a preformare il canale conduttivo, si ha comunque un aumento della carica di canale e la tensione di soglia del MOS impiantato risulta inferiore a quella della corrispondente struttura senza impiantazione.

Analogamente, se invece di drogante di tipo *n* si impianta un drogante di tipo *p*, allora la tensione di soglia del sistema MOS viene innalzata (figura 5.38 (b)). Come si vedrà in seguito, l'impiantazione di drogante di tipo *p* è in particolare utilizzata per la

[22] Variare la tensione di soglia per via elettrica mediante la applicazione di una tensione di substrato è in questo caso impossibile: per diminuire V_{th} fino a renderla negativa si dovrebbe applicare una tensione V_{BS} positiva, ma questo porterebbe in polarizzazione diretta le giunzioni source-bulk e drain-bulk.

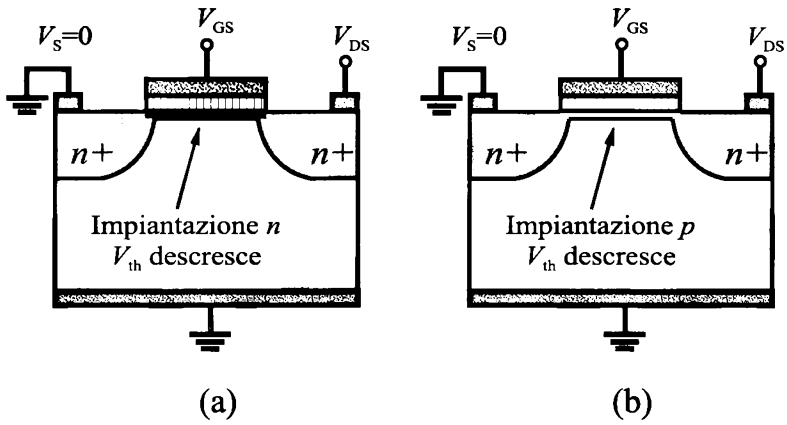


Figura 5.38 Regolazione della tensione di soglia mediante impiantazione ionica di drogante.

fabbricazione dei dispositivi *p*MOS a svuotamento.

Ovviamente perché tutta la discussione effettuata abbia senso, è necessario garantire che gli impianti di drogante siano estremamente superficiali e in concentrazione tale da variare la popolazione del canale conduttivo del sistema MOS senza stravolgerne completamente il funzionamento. In questo senso la tecnica della impiantazione ionica applicata a questo tipo di dispositivi richiede una precisione estrema.

Per valutare di quanto viene modificata la tensione di soglia mediante una impiantazione di drogante con dose[23] di donatori N_{dD} o di accettatori N_{dA}, si osservi che la carica che viene deposta è presente come carica aggiuntiva del sistema MOS già all'equilibrio: essa modifica quindi il bilancio di cariche dell'equilibrio e, in definitiva, la tensione per la quale il sistema si porta nella condizione di carica nulla, ovvero la tensione di banda piatta. Si dimostra facilmente che la tensione di banda piatta si modifica secondo la relazione:

$$V'_{FB} = V_{FB} - \frac{qN_{dD}}{C_{ox}} + \frac{qN_{dA}}{C_{ox}} \tag{5.81}$$

e di conseguenza la tensione di soglia del sistema MOS varia come:

$$\Delta V_{th} = -\frac{qN_{dD}}{C_{ox}} + \frac{qN_{dA}}{C_{ox}} \tag{5.82}$$

Tornando al dispositivo *n*MOS a svuotamento, le caratteristiche di uscita e la transcaratteristica sono del tutto analoghe a quelle del dispositivo di tipo enhancement, con la sola differenza che per $V_{GS} = 0$ la corrente di drain è non nulla. Si noti che comunque il dispositivo si porta in conduzione per $V_{GS} > V_{th}$ e che la corrente di drain è una funzione crescente di V_{GS}. Un esempio di caratteristiche statiche di un *n*MOS di tipo

[23] Con dose di atomi droganti si intende il numero di atomi inseriti nel semiconduttore per unità di area.

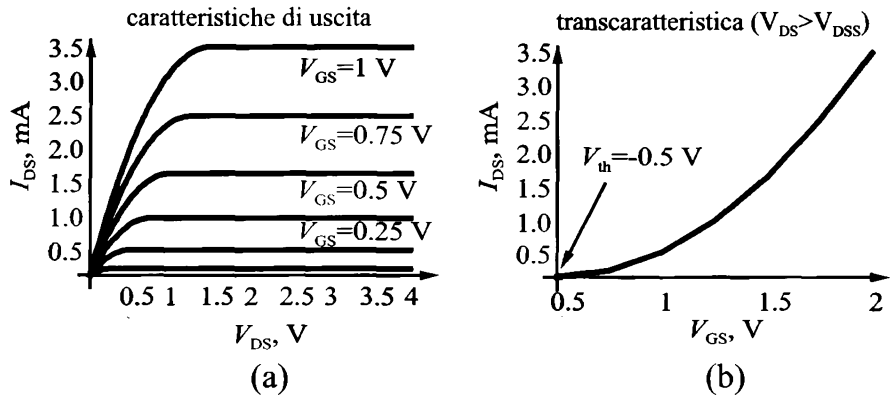


Figura 5.39 Caratteristiche di uscita e transcaratteristica statica del transistore *n*MOS a svuotamento.

depletion è riportato nella figura 5.39.

5.5 I transistori *p*MOS

La struttura fisica del transistore *p*MOS ad arricchimento è mostrata nella figura 5.40 (a). Essa è del tutto simmetrica rispetto al dispositivo *n*MOS poiché sono scambiati i drogaggi: il *p*MOS è costruito su un substrato drogato di tipo *n* mentre le regioni di source e drain sono di tipo *p*. La simmetria della struttura induce a pensare che esista anche una simmetria nel comportamento elettrico, a patto di scambiare i segni delle cariche e, in effetti, si può affermare che il funzionamento del transistore *p*MOS è del tutto analogo a quello del corrispondente *n*MOS a patto di scambiare il ruolo degli elettroni con quello delle lacune. Se poi le cariche in gioco hanno segno opposto allora anche le tensioni di pilotaggio del dispositivo dovranno avere segno opposto e così via.

Applicando al gate del *p*MOS una tensione V_{GS} *inferiore alla tensione di soglia* V_{th} (cfr. la figura 5.40 (b)) il sistema MOS su substrato di tipo *n* si porta nella regione di inversione e si viene quindi a formare un canale conduttivo di *lacune* all'interfaccia tra il semiconduttore e l'ossido. Il canale conduttivo forma una regione ohmica popolata di lacune saldandosi a sinistra con la regione del source e a destra con quella di drain. È allora possibile instaurare una corrente di lacune tra il source e il drain mediante la applicazione di una tensione V_{DS}: perché si abbia effettivamente un moto di lacune dal source al drain la tensione V_{DS} deve essere negativa (opposta al caso dello *n*MOS) e il campo elettrico indotto da questa tensione è diretto dal source al drain. La corrente I_{DS} che si sviluppa è *negativa entrante nel drain*, anche in questo caso con segno opposto alla I_{DS} del dispositivo *n*MOS. Per completare la simmetria si ricordi anche che la tensione di soglia del sistema MOS su substrato *n* è solitamente *negativa*: per portare un *p*MOS in conduzione, quindi, la tensione V_{GS} deve essere portata ad un valore ancora più negativo della tensione di soglia. A mano che V_{GS} decresce (aumenta in modulo) la quantità di lacune nel canale conduttivo aumenta e la corrente I_{DS} è più intensa (aumenta in modulo, diventando più negativa).

Da quanto detto si intuisce che l'andamento della corrente del *p*MOS si può ricavare

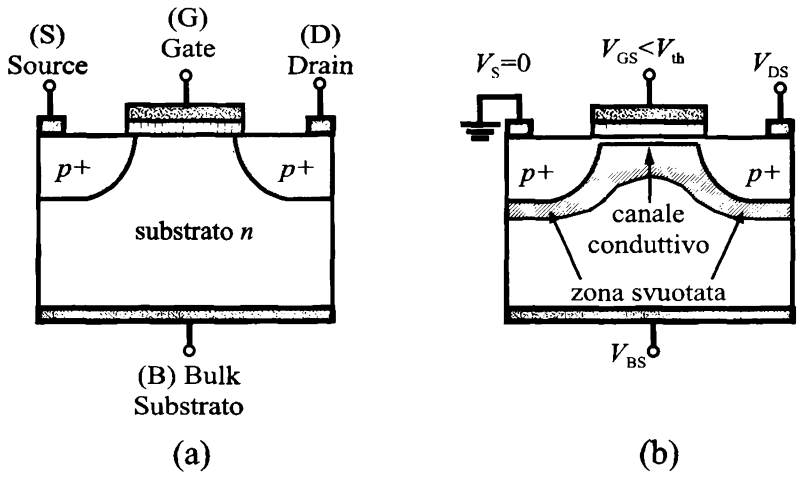


Figura 5.40 (a) Struttura fisica del transistore *p*MOS ad arricchimento. (b) Formazione del canale conduttivo.

a partire da quello del dispositivo a canale n (figura 5.36) scambiando il segno delle tensioni e delle correnti, ovvero ribaltando il grafico rispetto all'origine degli assi. Le caratteristiche di uscita risultanti sono mostrate nella figura 5.41, che mostra anche le regioni di funzionamento del *p*MOS. Per piccoli valori di V_{DS} il dispositivo si comporta come un resistore pilotato dalla tensione V_{GS} e la corrente dipende linearmente dalla tensione di drain (zona lineare). Diminuendo V_{DS} si passa nella regione quadratica dove la corrente cresce meno che linearmente, e infine quando si raggiunge il valore $V_{DSS} = V_{GS} - V_{th}$, detto tensione di saturazione, il canale conduttivo si strozza al drain. Per $V_{DS} < V_{DSS}$ la corrente di drain non dipende più dalla tensione V_{DS} e rimane costante al valore raggiunto allo strozzamento del canale. Si è quindi nella regione di saturazione. In questa regione la corrente di drain dipende dalla sola tensione V_{GS} secondo la transcaratteristica statica, il cui andamento è quadratico in V_{GS}. Un esempio di caratteristiche di uscita e transcaratteristica di un *p*MOS ad arricchimento è riportato nella figura 5.42.

Dal punto di vista analitico le caratteristiche statiche del *p*MOS ad arricchimento possono essere ricavate direttamente a partire dalla (5.79) sostituendo la mobilità delle lacune a quella degli elettroni e cambiando il segno a tensioni e correnti. Si ricava:

$$I_{DS} = \begin{cases} -\frac{W}{L}\mu_p C_{ox}\left[(V_{GS} - V_{th})\,V_{DS} - \frac{1}{2}V_{DS}^2\right] & V_{GS} < V_{th} \text{ e } V_{DS} > V_{DSS} \\ -\frac{1}{2}\frac{W}{L}\mu_p C_{ox}\,(V_{GS} - V_{th})^2 & V_{GS} < V_{th} \text{ e } V_{DS} < V_{DSS} \end{cases} \tag{5.83}$$

dove la tensione di saturazione $V_{DSS} = V_{GS} - V_{th}$ è negativa. Si noti che avendo cambiato segno a tensioni e correnti è stato anche scambiato il verso delle diseguaglianze.

Se in un transistore *p*MOS si effettua un impiantazione ionica superficiale di drogante di tipo p, la tensione di soglia aumenta secondo la (5.82). Se la concentrazione di drogante impiantato è sufficiente, la tensione di soglia del *p*MOS può diventare po-

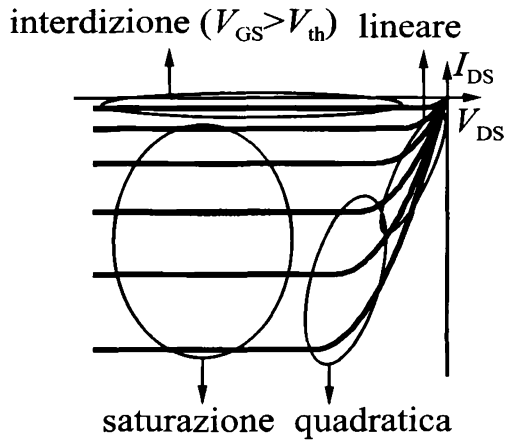


Figura 5.41 Regioni di funzionamento del transistore *p*MOS ad arricchimento.

sitiva e questo caso il canale conduttivo di lacune è *preformato*: si ha un dispositivo *p*MOS a svuotamento. Rispetto al *p*MOS ad arricchimento la sola differenza è nel segno della tensione di soglia, mentre l'andamento della corrente in funzione delle tensioni di pilotaggio rimane invariato e la forma analitica della corrente è ancora data dalla (5.83).

Nella figura 5.43 sono infine mostrate, per confronto, le caratteristiche di uscita di tutti e quattro i transistori MOS.

5.6 Effetti di non idealità del transistore MOSFET

La descrizione del comportamento dei transistori MOSFET effettuata nei paragrafi precedenti non tiene conto di alcuni effetti di non idelaità che modificano lievemente l'andamento della corrente di canale. Per fissare le idee ci limitiamo ad analizzare ancora il dispositivo *n*MOS ad arricchimento di cui si è studiato il funzionamento in maniera più approfondita, lasciando al lettore l'estensione agli altri tre dispositivi.

Il primo fenomeno di non idealità che esaminiamo è il cosiddetto effetto di *modulazione della lunghezza di canale.* Facendo riferimento alla figura 5.33 (b), nella condizione di saturazione la regione di canale del MOSFET si divide in una parte ohmica e in una piccola parte strozzata. Questa è stata fino ad ora considerata di lunghezza molto inferiore rispetto alla parte ohmica, tanto da approssimarla al solo punto terminale del canale ($x = L$). In realtà, a mano a mano che la tensione di drain cresce, la parte strozzata del canale si estende leggermente in modo da sostenere il campo elettrico crescente, e il punto in cui il canale si strozza tende ad arretrare verso il source, come mostrato nella figura 5.44. Si dice che si ha un effetto di modulazione della lunghezza del canale conduttivo da parte della tensione V_{DS}: la parte ohmica del canale risulta avere una lunghezza L' diversa dalla lunghezza fisica del contatto di gate L e decrescente con V_{DS}. Si noti che questo effetto è comunque di entità limitata poiché la differenza tra L' e L si mantiene comunque molto piccola. Inoltre esso è più significativo nei dispositivi a gate molto corto, dove l'incidenza relativa della differenza

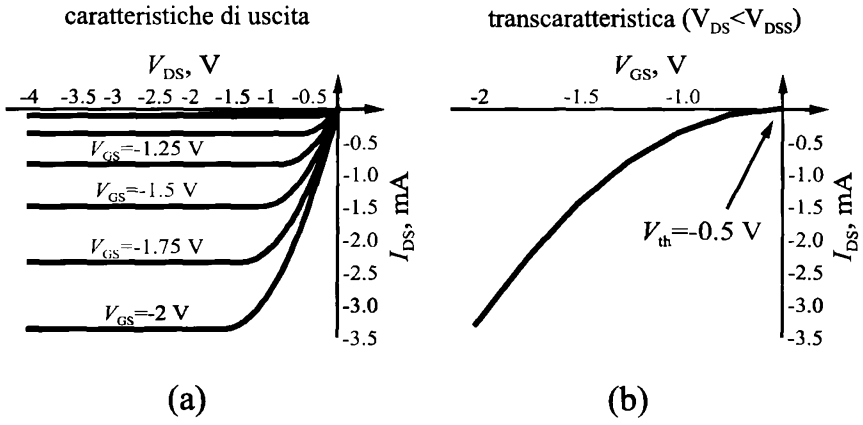


Figura 5.42 Caratteristiche di uscita e transcaratteristica statica del transistore *p*MOS ad arricchimento.

$L - L'$ è maggiore.

L'effetto principale della modulazione della lunghezza del canale consiste nel fatto che la corrente del MOSFET nella regione di saturazione *non è più indipendente dalla tensione* V_{DS}, ma cresce, sebbene debolmente, con essa. Infatti al crescere della tensione di drain il canale conduttivo effettivo *tende ad accorciarsi* e la corrente cresce. Nelle caratteristiche di uscita compare quindi una pendenza non nulla della corrente in funzione di V_{DS} nella regione di saturazione, come mostrato nella figura 5.45.

Dal punto di vista analitico, questo andamento viene descritto in maniera empirica. inserendo un fattore correttivo nella seconda delle (5.79) :

$$I_{DS} = \begin{cases} \frac{W}{L}\mu_n C_{ox}\left[(V_{GS} - V_{th})\,V_{DS} - \frac{1}{2}V_{DS}^2\right] & V_{GS} > V_{th} \text{ e } V_{DS} < V_{DSS} \\ \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\,(V_{GS} - V_{th})^2\,[1 + \lambda\,(V_{DS} - V_{DSS})] & V_{GS} > V_{th} \text{ e } V_{DS} > V_{DSS} \end{cases} \tag{5.84}$$

Il fattore correttivo è globalmente proporzionale al parametro λ, di solito determinato per via sperimentale.

Si noti che la dipendenza della corrente di drain da V_{DS} nella regione di saturazione è considerato un effetto di non indealità: se infatti I_{DS} dipende anche da V_{DS}, il pilotaggio della corrente alla porta di uscita da parte della tensione alla porta di ingresso V_{GS} viene deteriorato.

Passando ad analizzare altri effetti di non idealità del transistore, non si possono trascurare eventuali fenomeni di breakdown. Infatti all'interno del dispositivo sono presenti giunzioni *pn* polarizzate inversamente e queste possono presentare fenomeni di rottura. La giunzione più soggetta al breakdown e la giunzione drain-bulk: essa infatti è polarizzata più inversamente della source-bulk e la tensione inversa sulla giunzione cresce al crescere della tensione V_{DS}. Se ne deduce che se la V_{DS} aumenta troppo la giunzione drain-bulk entra in breakdown. In queste condizioni la corrente di drain cresce bruscamente poiché non è costituita dalla sola componente di elettroni provenienti dal canale conduttivo, ma ha un forte contributo legato alla corrente inversa

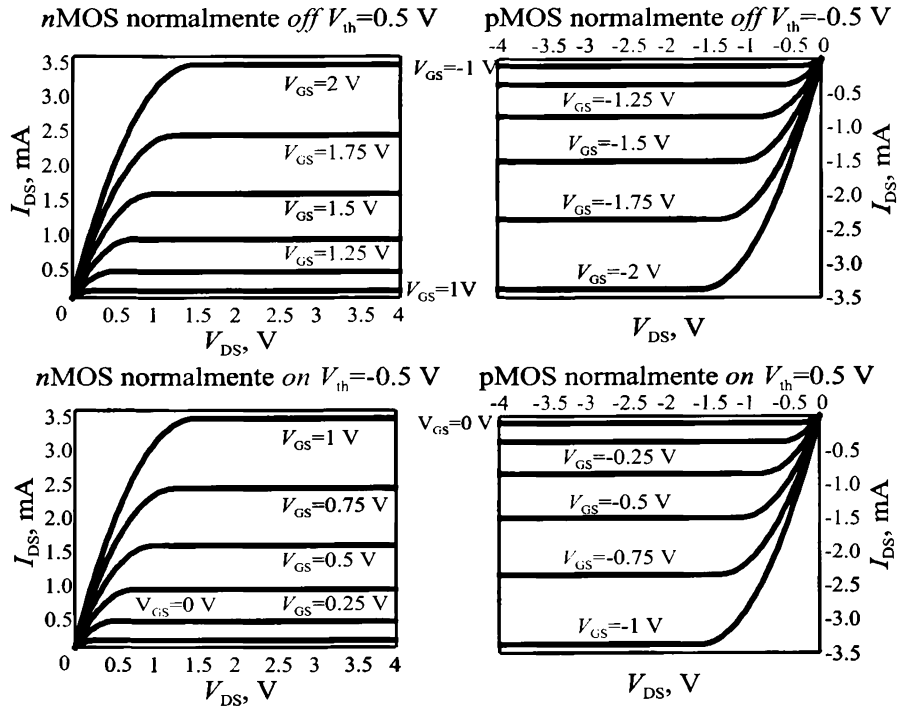


Figura 5.43 Sintesi delle caratteristiche di uscita dei quattro transistori MOS.

della giunzione stessa.

Altri possibili effetti di breakdown si hanno nella parte terminale (strozzata) del MOSFET in saturazione: in questa regione svuotata il campo elettrico raggiunge valori molto intensi e crescenti con la tensione di drain; questi a loro volta possono innescare il fenomeno della moltiplicazione a valanga di coppie elettroni-lacune portando, anche in questo caso, ad un brusco innalzamento della corrente di drain.

L'effetto del breakdown nelle caratteristiche di uscita del transistore è mostrato nella figura 5.46: esso è evidentemente limitato alla regione di saturazione per tensioni V_{DS} elevate. Si osservi il brusco aumento della corrente di drain in questa regione, pur essendo la tensione di gate invariata.

I fenomeni di rottura descritti, se non accuratamente controllati, possono portare ad un danneggiamento del dispositivo per effetto della notevole dissipazione di potenza elettrica con un conseguente aumento della temperatura interna del dispositivo stesso.

Anche una elevata tensione di gate può portare al breakdown, che si verifica in questo caso per via della perforazione dielettrica dell'ossido di gate. Il breakdown dell'ossido è un fenomeno sempre distruttivo per il MOSFET, e purtroppo anche molto frequente per via del suo spessore estremamente ridotto, tanto che particolare cura viene posta nel proteggere il contatto di gate da sovratensioni, disturbi o segnali indesiderati: oltre a particolari circuiti di protezione è sempre opportuno evitare di maneggiare i dispositivi senza le opportune precauzioni per evitare scariche elettriche.

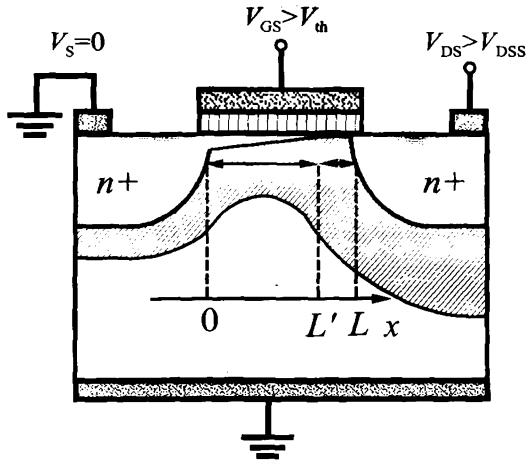


Figura 5.44 Modulazione della lunghezza di canale nel transistore *n*MOS ad arricchimento.

5.7 Il terminale di substrato

Come anticipato nel paragrafo 5.3 il terminale di substrato non è mai utilizzato come terminale attivo nei MOSFET. Esso deve però essere polarizzato in maniera da garantire che la corrente in esso sia nulla, ovvero che le giunzioni source-substrato e drain-substrato siano polarizzate inversamente. In generale è sufficiente che la tensione V_{BS} sia non positiva nel caso dei dispositivi a canale n e non negativa in quelli canale p. La scelta più ovvia in entrambi i casi è quella di non polarizzare il substrato e portarlo direttamente alla stessa tensione del riferimento di potenziale. In alcuni casi e in particolare per i dispositivi in package, la metallizzazione del bulk viene messa fisicamente in corto circuito con il source, come mostrato nella figura 5.47, e il MOSFET si presenta come un dispositivo a tre terminali piuttosto che a quattro.

Nel caso di dispositivi integrati il contatto di substrato è invece da considerare in fase di progetto. In alcuni circuiti complessi, e in particolare in molti circuiti digitali, i MOSFET non hanno sempre almeno un terminale collegato al riferimento di potenziale e in questo caso il contatto di substrato va collegato in maniera indipendente in modo che la sua corrente sia comunque nulla in qualsiasi condizione di utilizzo del circuito. Ad esempio, nei circuiti CMOS (cfr. il successivo approfondimento 5.4) nè il source nè il drain del *p*MOS sono normalmente collegati a massa, anzi, il source è spesso collegato alla massima tensione del circuito, la tensione di alimentazione, in modo da garantire una tensione V_{DS} negativa. In questo caso la scelta più conservativa per il contatto di substrato è quella di collegare anch'esso alla tensione di alimentazione. In questo modo per qualunque tensione di drain compresa tra zero e la tensione di alimentazione, la giunzione drain-substrato si mantiene in polarizzazione inversa.

Come si mostrerà nel paragrafo 5.8, un eventuale segnale elettrico applicato al substrato viene trasferito anche al contatto di drain, come variazione della corrente di drain stessa. Questo effetto è di solito indesiderato, in quanto al contatto di substrato non vengono mai applicati intenzionalmente seganli elettrici: è però possibile che su questo terminale siano presenti disturbi o segnali spuri che compromettono l'o-

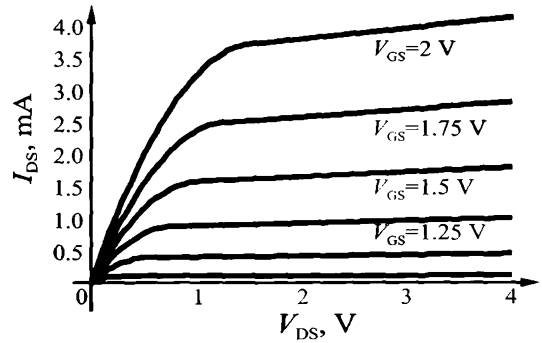


Figura 5.45 Caratteristiche di uscita del transistore *n*MOS ad arricchimento con modulazione della lunghezza di canale.

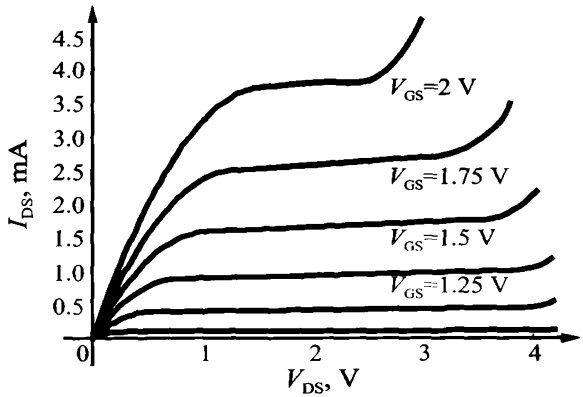


Figura 5.46 Caratteristiche di uscita del transistore *n*MOS ad arricchimento con effetto di breakdown.

timo funzionamento del dispositivo. Per evitare questi effetti è opportuno collegare sempre il contatto di substrato ad una tensione continua, preferibilmente la massa o l'alimentazione, mediante una metallizzazione il più possibile corta in modo da evitare disturbi.

Sebbene non sia comune, è anche possibile utilizzare il terminale di substrato per variare la tensione di soglia del dispositivo. Quando la dipendenza della tensione di soglia dalla tensione V_{BS} non sia nota per via teorica, essa può essere estratta sperimentalmente come mostrato nella figura 5.48. Polarizzando il gate e il drain alla stessa tensione si assicura che il MOSFET lavori nella regione di saturazione e che la sua dipendenza da $V_{GS} = V_{DS}$ sia parabolica. Disegnando allora la radice quadrata della corrente in funzione della tensione applicata si ha una retta, di cui è facile determinare per via grafica l'intersezione con l'asse delle ascisse. In tali intersezioni si determina il valore della tensione di soglia per ogni fissato valore di V_{BS}.

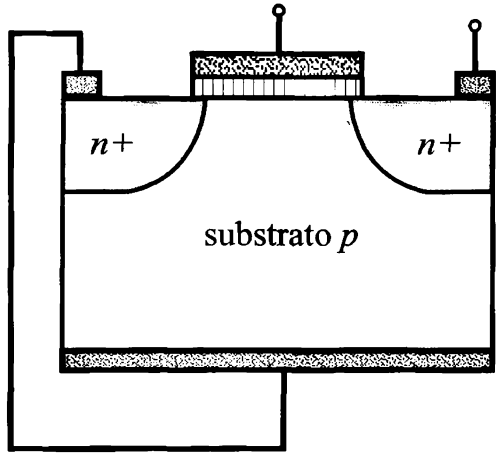


Figura 5.47 MOSFET con contatto di substrato cortocircuitato al source.

5.8 I modelli circuitali del transistore MOSFET

Come si è già visto nel caso della giunzione *pn* e del transistore bipolare, una volta capito il comportamento elettrico di un dispositivo a partire dai pincipi fisici, è opportun passare ad una descrizione dello stesso comportamento mediante modelli di tipo circuitale: questi sono più adatti ad essere utilizzati all'interno dei simulatori CAD nell progettazione dei circuiti elettronici. Prima di passare ad analizzare i modelli circuita che possono essere utilizzati per descrivere il transistore MOSFET, però, vediamo nell figura 5.49 alcuni dei suoi simboli circuitali più utilizzati.

I simboli in alto possono rappresentare indifferentemente dispositivi ad arricchimento o a svuotamento. Sono rappresentati i terminali di gate (simboleggiato da un doppia barra che ricorda l'ossido di gate), di drain e di source (identificato da un freccia che indica il verso della corrente nel terminale stesso – entrante nel *p*MOS uscente nello *n*MOS). Nei circuiti digitali, è anche consuetudine distinguere il *p*MO con un "pallino" sul gate. In questi simboli il terminale di substrato non è esplicitame te evidenziato: essi sono utili quando il terminale di bulk è cortocircuitato con il sour oppure quando esso non interviene in nessun modo nel funzionamento del circuito. p cui evidenziare il quarto terminale del MOSFET sarebbe una inutile complicazione. I simboli in basso mostrano invece come si indica il terminale di substrato, quand necessario: una freccia orientata verso il gate per il dispositivo *n*MOS e uscente per *p*MOS. Nella stessa figura si mostra anche come eventualmente distinguere i dispositi a svuotamento da quelli ad arricchimento, caratterizzati da una delle linee della dop barra del gate tratteggiata, a simbolizzare che il canale non è presente in condizio equilibrio.

5.8.1 Il modello di ampio segnale

Il transistore MOS è normalmente utilizzato nella configurazione a doppio bipolo. in cui un terminale viene posto in comune alla porta di ingresso e di uscita. La figur 5.50 mostra ad esempio la configurazione a source comune (CS, dall'inglese *com*

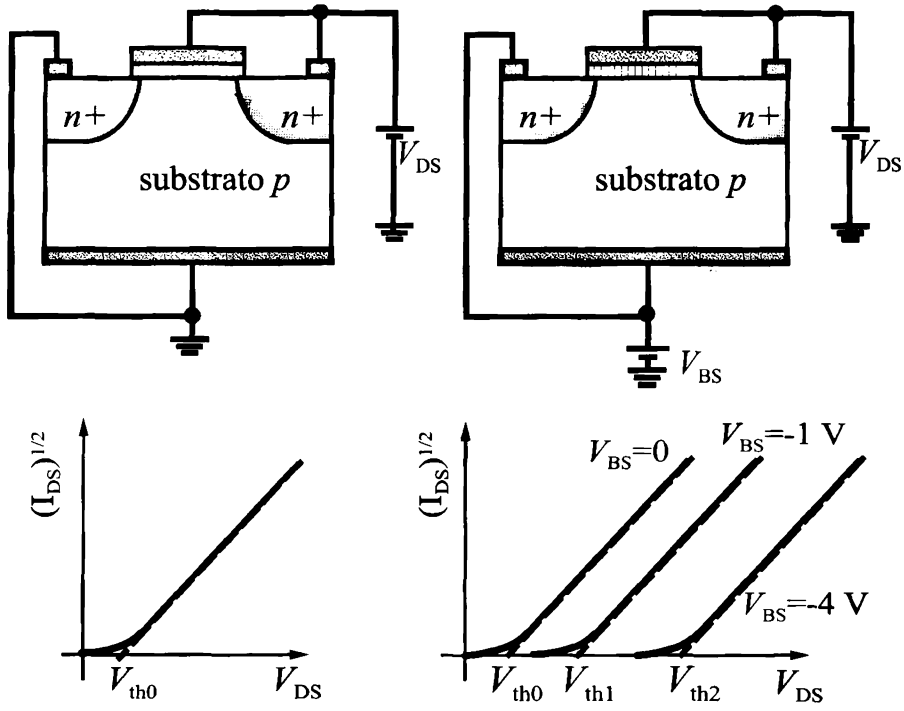


Figura 5.48 Misura della tensione di soglia in funzione della tensione di substrato.

source), per alcuni aspetti la più utilizzata, anche se molto comuni sono anche la configurazione a drain comune (CD, *common drain*) e la configurazione a gate comune (CG, *common gate*). Il terminale di substrato non è mai utilizzato come terminale attivo e, di conseguenza, non entra nella definizioni di porta di ingresso o di uscita: il suo ruolo è già stato discusso nel paragrafo 5.7.

È noto che un 2-porte è completamente caratterizzato dal legame tra le correnti e le tensioni di porta, ovvero dalle caratteristiche di ingresso e di uscita. Come si è visto nel capitolo 3, per determinare queste caratteristiche in condizioni generalmente tempovarianti (ad esempio con segnali a gradino, sinusoidi ecc.) è necessario il *modello di ampio segnale dinamico* del dispositivo, nella fattispecie del MOSFET, che fornisce il legame tra le correnti entranti alle porte $i_{GS}(t)$ ed $i_{DS}(t)$ in funzione dalle tensioni di controllo $v_{GS}(t)$ ed $v_{DS}(t)$ e del tempo[24]. Questo tipo di modello include sia la dipendenza istantanea delle correnti dalle tensioni (ovvero il modello *di ampio segnale statico*) sia eventuali effetti di tipo reattivo (ad esempio capacitivo) presenti nel dispositivo. Mentre per i dispositivi visti nei capitoli precedenti, giunzione *pn* e transistore bipolare, si è fornito sia il modello di ampio segnale statico sia quello dinamico, nel

[24] In questo capitolo adotteremo la convenzione già introdotta nei capitoli 3 e 4: si indicheranno con *simboli minuscoli* e *pedici maiuscoli* le generiche tensioni e correnti associate al dispositivo in condizioni di ampio segnale; con *simboli maiuscoli* e *pedici maiuscoli* le tensioni continue e con *simboli minuscoli* e *pedici minuscoli* le tensioni di piccolo segnale.

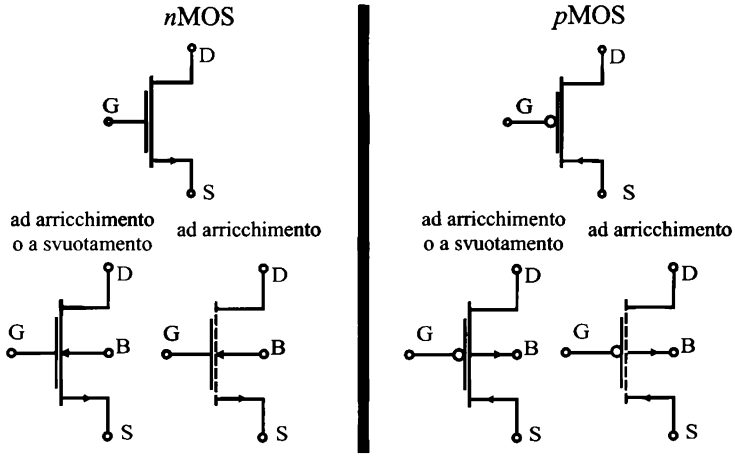


Figura 5.49 Simboli circuitali dei dispositivi MOSFET.

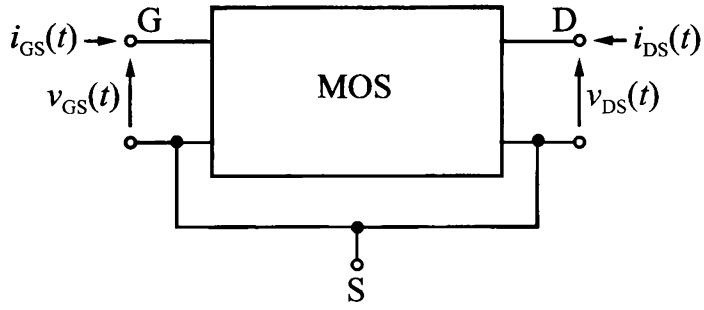


Figura 5.50 Collegamento del MOSFET come un 2-porte a source comune.

caso del transitore MOSFET ci limiteremo al solo caso statico. Infatti, lo studio dei effetti capacitivi, sebbene in linea di principio estremamente importante nelle applicazioni sia analogiche sia digitali ad alta frequenza, è molto complesso ed esula scopi di questo testo. Il modello di ampio segnale statico, del resto, può comunque essere utilizzato anche in condizioni dinamiche, a patto che le tensioni e correnti in gioco abbiano una variazione molto lenta e tale per cui gli effetti reattivi di sfasamento tra tensioni e correnti siano trascurabili.

Il modello di ampio segnale statico del MOSFET è più semplice dell'analogo modello del transistore bipolare: infatti la corrente alla porta di ingresso $i_{GS}(t)$, supponendo di trascurare, in condizioni quasi-statiche, la corrente che scorre nella capacità dell'ossido di gate, è nulla e la caratteristica di ingresso si semplifica. La caratteristica di uscita si utilizza poi il legame tra la corrente di drain e le tensioni di pilotaggio già determinato in condizioni statiche (cfr. la (5.79)), estendendolo al caso dinamico. Si ottiene:

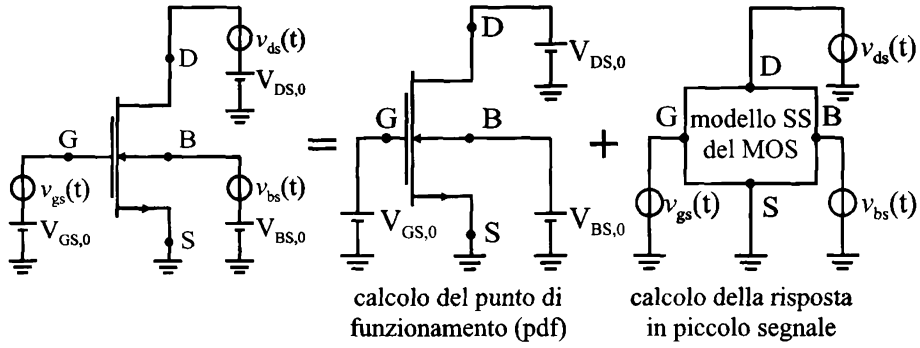


Figura 5.51 Soluzione del circuito a source comune nella approssimazione di piccolo segnale.

▷ porta di ingresso

$$i_{GS}(t) \equiv 0$$

▷ porta di uscita

$$i_{DS}(t) = \begin{cases} \frac{W}{L}\mu_n C_{ox}\left[(v_{GS}(t) - V_{th})\, v_{DS}(t) - \frac{1}{2}v_{DS}(t)^2\right] & v_{DS} < V_{DSS} \\ \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\,(v_{GS}(t) - V_{th})^2\,[1 + \lambda\,(v_{DS}(t) - V_{DSS})] & v_{DS} > V_{DSS} \end{cases} \tag{5.85}$$

Si osservi che esiste in effetti anche una dipendenza nascosta dalla *tensione* $v_{BS}(t)$ attraverso la tensione di soglia:

$$V_{th}(v_{BS}(t)) = V_{th,0} + \gamma_B\left(\sqrt{2\Phi_p - v_{BS}(t)} - \sqrt{2\Phi_p}\right)$$

ma $v_{BS}(t)$ non è in genere un segnale intenzionale applicato quanto piuttosto un disturbo dovuto ad esempio ad un imperfetto collegamento del terminale di substrato a massa o alla alimentazione del circuito (cfr. la discussione nel paragrafo 5.7).

5.8.2 Il modello di piccolo segnale

Come già visto nei capitoli precedenti un caso significativo si ha quando le tensioni e le correnti ai teminali del dispositivo si possono scomporre in una componente in continua (o *di polarizzazione*) e una componente di piccolo segnale. Se infatti le componenti di segnale si mantengono di ampiezza molto minore delle corrispondenti grandezze in contina, si parla di regime di piccolo segnale. La figura 5.51 mostra in modo schematico il procedimento utilizzato nella soluzione dei circuiti in regime di piccolo segnale: si noti che in questo paragrafo si darà solo una breve sintesi di questo concetto, rimandando alla discussione già effettuata nei capitoli 3 e 4.

Le tensioni e le correnti del circuito si scompongono nella parte in continua e nella

parte di piccolo segnale

$$\begin{aligned} v_{GS}(t) &= V_{GS,0} + v_{gs}(t) & \qquad i_{GS}(t) &= I_{GS,0} + i_{gs}(t) \\ v_{DS}(t) &= V_{DS,0} + v_{ds}(t) & \qquad i_{DS}(t) &= I_{DS,0} + i_{ds}(t) \\ v_{BS}(t) &= V_{BS,0} + v_{bs}(t) & \qquad i_{BS}(t) &= I_{BS,0} + i_{bs}(t) \end{aligned} \tag{5.86}$$

dove l'insieme $\{V_{GS,0},V_{DS,0},V_{BS,0},I_{GS,0},I_{DS,0},I_{BS,0}\}$ costituisce il *punto di funzionamento a riposo* (pdf) del dispositivo. Si noti che poiché non è possibile in genere garantire che sia nullo il segnale al substrato, nel seguito continueremo a considerare esplicitamente l'effetto della $v_{bs}(t)$; ove questa non sia presente, basterà porre $v_{bs}(t) = 0$.

Se le componenti tempovarianti sono di piccola ampiezza, allora la loro applicazione è una piccola perturbazione del circuito rispetto all'applicazione delle sole tensioni in continua: in questo caso il pdf del circuito non varia e può essere calcolato separatamente. Nel regime di piccolo segnale, quindi, la analisi dei circuiti si semplifica. poiché può essere scomposta in due passi successivi (cfr. la figura 5.51): il calcolo delle correnti e tensione in continua (il pdf, appunto), effettuato mediante il modello di ampio segnale statico, e il calcolo delle componenti di piccolo segnale. A sua volta. quest'ultimo può essere efficacemente effettuato mediante il *modello di piccolo segnale (small-signal, SS) del dispositivo*, che si ottiene a partire dal modello di ampio segnale attraverso la sua linearizzazione attorno al punto di funzionamento statico.

Dal punto di vista circuitale, il modello di piccolo segnale è particolarmente utile nelle applicazioni analogiche nelle quali, come risulterà chiaro nel successivo capitolo 6, il MOSFET è polarizzato in maniera che il punto di funzionamento risulti nella regione di suturazione. Per estrarre il modello di piccolo segnale del MOSFET, allora. linearizziamo le (5.85) attorno al punto di funzionamento a riposo limitandoci alla sola regione di saturazione ($V_{DS,0} > V_{DSS}$). Ovviamente la corrente di gate è nulla anche nel caso del modello di piccolo segnale:

$$i_{gs} = 0$$

mentre la corrente di drain è:

$$\begin{aligned} i_{DS}[v_{GS}(t),v_{DS}(t),v_{BS}(t)] \approx\ & i_{DS}[V_{GS,0},V_{DS,0},V_{BS,0}] \\ & + \left.\frac{\partial i_{DS}}{\partial v_{GS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \cdot v_{gs}(t) \\ & + \left.\frac{\partial i_{DS}}{\partial v_{DS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \cdot v_{ds}(t) \\ & + \left.\frac{\partial i_{DS}}{\partial v_{BS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \cdot v_{bs}(t) \end{aligned} \tag{5.87}$$

Poiché:

$$i_{DS}(t) = I_{DS,0} + i_{ds}(t)$$

e:

$$i_{DS}[V_{GS,0},V_{DS,0},V_{BS,0}] = I_{DS,0}$$

la componente di piccolo segnale $i_{ds}(t)$ della corrente di drain è:

$$\begin{aligned} i_{ds}(t) = & \left.\frac{\partial i_{DS}}{\partial v_{GS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \cdot v_{gs}(t) \\ & + \left.\frac{\partial i_{DS}}{\partial v_{DS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \cdot v_{ds}(t) \\ & + \left.\frac{\partial i_{DS}}{\partial v_{BS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \cdot v_{bs}(t) \end{aligned} \tag{5.88}$$

Definendo i parametri differenziali:

- *transconduttanza* [S]

$$g_m = \left.\frac{\partial i_{DS}}{\partial v_{GS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}}$$

- *conduttanza di uscita* [S]

$$g_o = \left.\frac{\partial i_{DS}}{\partial v_{DS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}}$$

- *tranconduttanza di substrato* [S]

$$g_{mb} = \left.\frac{\partial i_{DS}}{\partial v_{BS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}}$$

la corrente di piccolo segnale al drain si scrive come:

$$i_{ds}(t) = g_m \cdot v_{gs}(t) + g_o \cdot v_{ds}(t) + g_{mb} \cdot v_{bs}(t) \tag{5.89}$$

La corrente di piccolo segnale è ovviamente lineare nelle tensioni e la (5.89) ammette allora una interpretazione mediante un circuito lineare, il cosiddetto circuito equivalente di piccolo segnale, riportato nella figura 5.52. Esso si costruisce banalmente a partire dalla (5.89) per *sovrapposizione degli effetti.*

Dei due generatori di corrente pilotati, quello pilotato dalla tensione V_{gs} è di gran lunga il più significativo e rappresenta in sintesi la capacità di pilotaggio della corrente alla porta di uscita mediante la tensione alla porta di ingresso. Il secondo generatore pilotato, rappresenta invece un possibile effetto di disturbo sulla corrente di uscita ad opera della tensione di substrato.

È molto significativo osservare l'analogia del circuito equivalente di piccolo segnale del MOSFET con quello del transistore bipolare (cfr. ad esempio la figura 4.18). Questa analogia è soprattutto relativa alla corrente alla porta di uscita, che ha anche nel caso del bipolare una componente fornita da un generatore pilotato dalla tensione di ingresso e una componente resistiva legata alla conduttanza di uscita non nulla del

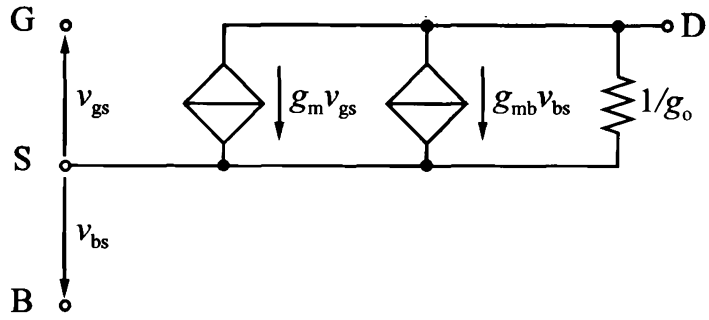


Figura 5.52 Circuito equivalente di piccolo segnale statico del MOSFET.

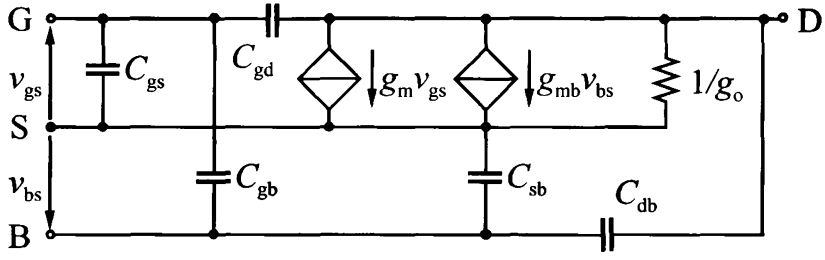


Figura 5.53 Circuito equivalente di piccolo segnale per le alte frequenze del MOSFET.

transistore. La differenza fondamentale tra il transistore bipolare e il MOSFET è invec legata alla porta di ingresso: mentre nel MOSFET, almeno nel caso quasi–statico. terminale di gate è *un terminale completamente isolato* e la corrente di ingresso è nulla nel transistore bipolare si ha una impedenza di ingresso finita e una corrente di ingres divesa da zero. Da questo punto di vista il MOSFET presenta un comportamento pi "ideale" rispetto al bipolare, poiché il controllo alla porta di uscita avviene puramen in tensione, senza dissipazione di potenza in ingresso.

Sebbene, come si è detto, in questo testo non si affronti lo studio dei modelli dinam ci del transisotre MOS, accenniamo a una possibile estensione del circuito equivalen di piccolo segnale per le alte frequenze. In questo caso è necessario aggiungere gli el menti reattivi e in particolare le capacità interne del dispositivo come mostrato nel figura 5.53.

Naturalmente quanto detto finora riguardo al circuito di piccolo segnale e ai parametri differenziali si applica ai soli dispositivi a canale n. È tuttavia immedia generalizzare la trattazione al caso dei pMOS con ovvie modifiche. Il modello di amp segnale statico del pMOS si ottiene a partire dalle (5.83) mentre il modello di picc segnale si ricava dalla loro linearizzazione attorno al punto di funzionamento a rip Il circuito equivalente di piccolo segnale del pMOS è uguale a quello del dispositivo canale n, ma il generatore pilotato ha corrente *uscente dal drain* invece che entrante I parametri differenziali del circuito di piccolo segnale del pMOS hanno espressio analoga a quelli dello nMOS, a patto di scambiare le mobilità degli elettroni con quel

delle lacune.

Le applicazioni dei modelli circuitali del MOSFET saranno affrontate nel capitolo 6 con particolare attenzione ai circuiti analogici. In questi circuiti si utilizza estesamente il circuito equivalente di piccolo segnale del MOSFET e, per familiarizzare il lettore con i suoi parametri, in questo capitolo forniamo un esempio di estrazione dei parametri del circuito equivalente di piccolo segnale di un dispositivo a canale n a partire da misure (esempio 5.5). Come applicazione del modello di ampio segnale statico, invece, gli approfondimenti 5.4 e 5.5 affrontano un esempio di circuito digitale: l'inverter CMOS.

5.8.3 Parametri differenziali del circuito equivalente di piccolo segnale

Il circuito equivalente della figura 5.52 è caratterizzato dai tre parametri differenziali g_m, g_o e g_{mb}. Questi a loro volta si ottengono dalla linearizzazione delle equazioni del modello di ampio segnale e non sono altro che derivate parziali della corrente rispetto alle tre tensioni v_{GS}, v_{DS}, v_{BS}, valutate nel punto di funzionmento a riposo.

Transconduttanza

Partendo dalla espressione della corrente del MOSFET in saturazione nel modello di ampio segnale (5.85) si ottiene:

$$g_m = \left.\frac{\partial i_{DS}}{\partial v_{GS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} = \frac{W}{L}\mu_n C_{ox}\left(V_{GS,0} - V_{th}\right)\left[1 + \lambda\left(V_{DS,0} - V_{DSS}\right)\right] \quad (5.90)$$

Poiché il termine di non idealità proporzionale a λ è in genere molto piccolo, si può approssimare l'espressione di g_m come:

$$g_m \simeq \frac{W}{L}\mu_n C_{ox}\left(V_{GS,0} - V_{th}\right) = \beta_n\left(V_{GS,0} - V_{th}\right) \quad (5.91)$$

Infine, osservando che $I_{DS,0} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}\left(V_{GS,0} - V_{th}\right)^2$ (sempre a meno del termine correttivo in λ), l'espressione precedente si può anche scrivere nella forma:

$$g_m = \sqrt{2\frac{W}{L}\mu_n C_{ox} I_{DS,0}} \quad (5.92)$$

dove si mette in evidenza la dipendenza dalla corrente nel punto di funzionamento piuttosto che dalla tensione.

La transconduttanza g_m è un parametro di merito fondamentale del MOSFET. Ad essa è legata l'intensità del generatore pilotato presente nel circuito equivalente (cfr. la figura 5.52) e in definitiva la capacità di pilotaggio della porta di uscita da parte di quella di ingresso. Per massimizzare la transconduttanza valgono le stesse considerazioni già viste per la corrente di drain: occorre minimizzare la lunghezza di gate L e massimizzare la capacità per unità di area dell'ossido, ovvero ridurre lo spessore dell'ossido stesso.

Conduttanza di uscita

Sempre nell'ipotesi che il MOSFET sia in saturazione, derivando la (5.85) rispetto a v_{DS} si ottiene:

$$g_o = \left.\frac{\partial i_{DS}}{\partial v_{DS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}(V_{GS,0} - V_{th})^2 \lambda \qquad (5.93)$$

Utilizzando ancora l'espressione della corrente in continua:

$$I_{DS,0} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}(V_{GS,0} - V_{th})^2 [1 + \lambda(V_{DS,0} - V_{DSS})] \simeq \frac{1}{2}\frac{W}{L}\mu_n C_{ox}(V_{GS,0} - V_{th})^2$$

si può anche semplicemente scrivere:

$$g_o = \lambda I_{DS,0} \qquad (5.94)$$

Si osservi che nella regione di saturazione il comportamento del MOSFET è tanto migliore quanto più g_o è piccola. In questo modo infatti la corrente di drain in condizioni di piccolo segnale è pilotata dalla sola tensione di ingresso e non dipende dalla tensione alla porta di uscita. Per avere quindi un MOSFET ottimo è necessario che il parametro λ sia molto piccolo ed eventualmente polarizzare il dispositivo a basse correnti. Quest'ultima scelta, però, non è sempre praticabile poiché riduce la dinamica degli stadi amplificatori, come verrà spiegato nel capitolo 6.

Transconduttanza di substrato

Come si è già avuto modo di osservare, la corrente del MOSFET dipende da v_{BS} attraverso la tensione di soglia

$$i_{DS} = \frac{1}{2}\frac{W}{L}\mu_n C_{ox}(v_{GS} - V_{th}(v_{BS}))^2$$

dove per semplicità si è trascurata la dipendenza di i_{DS} dalla tensione v_{DS} ($\lambda = 0$). Derivando rispetto a v_{BS}, la transconduttanza di substrato risulta:

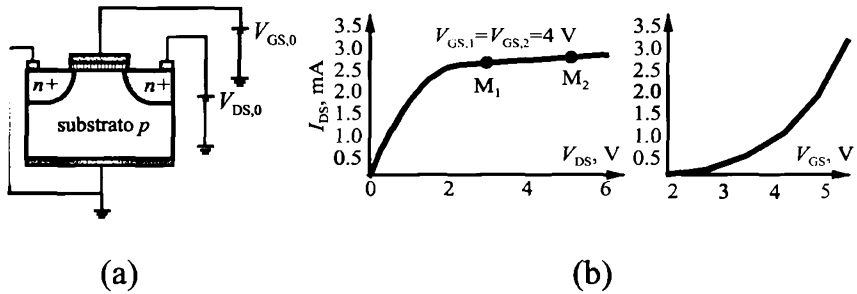
$$g_{mb} = \left.\frac{\partial i_{DS}}{\partial v_{BS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} = \left.\frac{\partial i_{DS}}{\partial V_{th}} \cdot \frac{\partial V_{th}}{\partial v_{BS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} \qquad (5.9$$

Poiché:

$$\frac{\partial i_{DS}}{\partial V_{th}} = -\frac{W}{L}\mu_n C_{ox}(v_{GS} - V_{th})$$

e, derivando la (5.39):

$$\left.\frac{\partial V_{th}}{\partial v_{BS}}\right|_{V_{GS,0},V_{DS,0},V_{BS,0}} = \frac{-\gamma_B}{2\sqrt{2\Phi_p - V_{BS,0}}}$$



(a) (b)

Figura 5.54 Esempio 5.5: (a) Set-up sperimentale dell'esperimento e (b) misure per la caratterizzazione del MOSFET.

Dalla (5.95) si ha:

$$g_{mb} = \frac{W}{L}\mu_n C_{ox}\left(v_{GS} - V_{th}\right)\frac{\gamma_B}{2\sqrt{2\Phi_p - V_{BS,0}}} = \frac{\gamma_B g_m}{2\sqrt{2\Phi_p - V_{BS,0}}} \tag{5.96}$$

Naturalmente è fattore di merito del transistore che la transconduttanza di substrato sia minima, così che il dispositivo sia più immune possibile ai disturbi. Dalla prima uguaglianza della (5.96), però, si osserva che per minimizzare g_{mb} è necessario minimizzare il fattore γ_B, a meno di non deteriorare anche la transconduttanza g_m. Il fattore γ_B, a sua volta, si minimizza riducendo il drogaggio del substrato, anche se è possibile dimostrare che questo comporta numerosi problemi soprattutto nei dispositivi più moderni con dimensioni molto scalate (cfr. ad esempio l'esempio 5.4) ed è spesso necessario accettare un valore di compromesso.

Esempio 5.5 Si consideri un MOS a canale *n* polarizzato come nella figura 5.54 (a). Si effettuano misure al variare di V_{GS} ottenedo la transcaratteristica mostrata nella figura 5.54 (b). Inoltre le due misure M_1 e M_2, effettuate per $V_{GS} = 4$ V forniscono i dati riportati nella caratteristica di uscita, sempre nella figura 5.54 (b).

- M_1: $V_{DS,1} = 3$ V, $I_{DS,1} = 2.5$ mA
- M_2: $V_{DS,2} = 5$ V, $I_{DS,2} = 2.7$ mA

Fissate $V_{GS,0} = 4$ V, $V_{DS,0} = 4$ V, si chiede di determinare il circuito equivalente di piccolo segnale del MOSFET in questo punto di funzionamento..

In primo luogo dalla transcaratteristica della figura 5.54 (b) si osserva che la tensione di soglia vale $V_{th} = 2$ V.

Poiché si chiede di determinare il circuito equivalente di piccolo segnale nel pdf caratterizzato da $V_{GS,0} = 4$ V, la tensione di saturazione vale $V_{DSS} = V_{GS,0} - V_{th} = 2$ V. Le misure M_1 e M_2 sono allora nella regione di saturazione, poiché sono effettuate ad una tensione di drain maggiore di V_{DSS}.

Sostituendo i valori delle tensioni delle misure M_1 e M_2 nella (5.84), si può scrivere il sistema di equazioni

$$\begin{cases} I_{DS,1} = \frac{1}{2}\beta_n\left(V_{GS,1} - V_{th}\right)^2\left[1 + \lambda\left(V_{DS,1} - V_{DSS}\right)\right] \\ I_{DS,2} = \frac{1}{2}\beta_n\left(V_{GS,2} - V_{th}\right)^2\left[1 + \lambda\left(V_{DS,2} - V_{DSS}\right)\right] \end{cases}$$

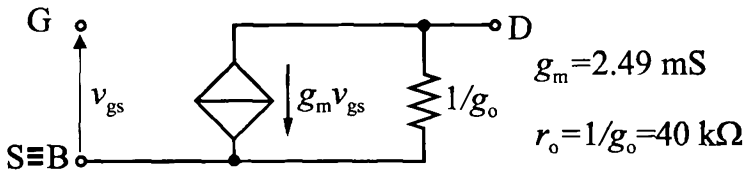


Figura 5.55 Esempio 5.5: circuito di piccolo segnale del MOSFET nel punto di funzionamento a riposo $V_{GS,0} = 4$ V, $V_{DS,0} = 4$ V.

Dividendo membro a membro si ottiene:

$$\lambda = 0.01 \text{ V}^{-1}$$

e infine:

$$\beta_n = 1.2375 \text{ mAV}^{-2}$$

Noti i parametri caratteristici del transistore, si calcolano i parametri differenziali. La transconduttanza vale:

$$g_m = \beta_n (V_{GS,0} - V_{th}) = 2.49 \text{ mS}$$

mentre la conduttanza di uscita è:

$$g_o = \frac{1}{2}\beta_n (V_{GS,0} - V_{th})^2 \lambda = 2.5 \times 10^{-5} \text{ S}$$

Si noti che la conduttanza di uscita è comunque un valore molto piccolo, che idealmente dovrebbe addirittura annullarsi. La transconduttanza di substrato in questo caso può essere trascurata poiché il source è direttamente cortocircuitato con il substrato ($v_{bs} \equiv 0$) ed entrambi sono collegati a massa.
Il circuito di piccolo segnale risultante è mostrato nella figura 5.55.

Approfondimento 5.4 L'inverter CMOS

Il circuito della figura 5.56 rappresenta un inverter nella cosiddetta tecnologia CMOS (dall'inglese *complementary MOS* ovvero a MOS complemantari). Questa tecnologia è caratterizzata dall'utilizzo di dispositivi sia a canale n sia a canale p: i circuiti sviluppati in questo modo sfruttano più possibile la simmetria, o meglio la complementarietà, nel comportamento delle due specie di dispositivi MOSFET.
L'inverter CMOS è il più semplice esempio di circuito digitale in tecnologia CMOS. Esso è costituito da due MOSFET collegati in serie, l'uno a canale n con il source collegato alla massa del circuito e l'altro a canale p con il source collegato alla alimentazione. In questo modo la tensione $V_{DS,n}$ dello *n*MOS è positiva e quella $V_{DS,p}$ del *p*MOS è negativa[25]. Come mostrato nella figura 5.56, l'ingresso del circuito è costituito dai due gate collegati insieme mentre l'uscita dai due drain anch'essi collegati tra loro. Inoltre poiché lo stadio è considerato a vuoto e l'ingresso sui gate non assorbe corrente in condizioni stazionarie, le uniche correnti possibili sono le correnti drain–source dei due dispositivi: per costruzione, poi, deve essere $I_{DS,n} = -I_{DS,p}$. In pratica la corrente scorre dall'alimentazione verso massa attraversando prima il canale conduttivo del *p*MOS dal source al drain e poi lo *n*MOS dal drain al source. Per motivi che sarranno chiari una volta compreso il funzionamento del circuito, il transistore *p*MOS è anche detto transistore di *pull–up* e lo *n*MOS detto di *pull–down*.
Il fatto di studiare lo stadio a vuoto non pregiudica la comprensione complessiva del funzionamento dell'inverter. Infatti è caratteristica generale dei circuiti in tecnologia CMOS che gli ingressi siano

[25] In questo approfondimento si utilizzerà il pedice n per le grandezze relative allo nMOS e il pedice p per il pMOS.

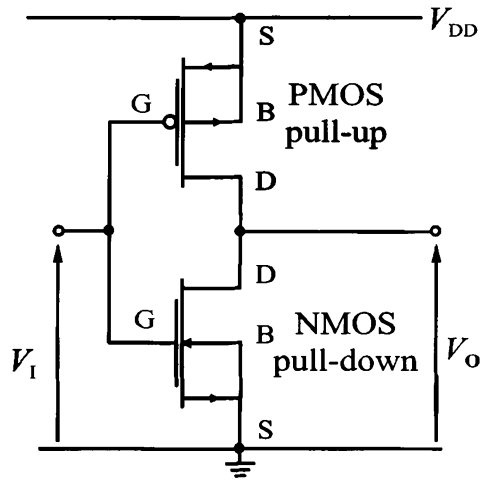


Figura 5.56 Stadio inverter in tecnologia CMOS.

costituiti da una coppia di terminali di gate (di un *p*MOS e di un *n*MOS) collegati insieme: questi, poi, non assorbono corrente in condizioni statiche. A valle dell'inverter in esame, allora, ci si deve immaginare l'ingresso di uno stadio digitale successivo che in continua rappresenta un circuito aperto (corrente nulla) così che è corretto studiare lo stadio a monte non caricato. Se si dovesse studiare invece il comportamento dinamico del circuito, allora si dovrebbe supporre l'inverter caricato in uscita con una capacità equivalente, che rappresenta la capacità di ingresso dei due gate dello stadio a valle. In questa trattazione si considererà il solo comportamento statico, per cui si trascureranno gli effetti capacitivi e di ritardo. Quando poi si utilizzeranno espressioni del tipo "variare la tensione", si intenderà comunque una variazione così lenta da poter considerare la risposta del circuito istantanea.

L'inverter realizza l'inversione del segnale digitale dall'ingresso all'uscita ovvero, indicando con "0" lo zero logico e con "1" l'uno logico, si ha:

$$V_I = \text{"1"} \Rightarrow V_O = \text{"0"}$$
$$V_I = \text{"0"} \Rightarrow V_O = \text{"1"}$$

La connotazione principale dell'inverter è la sua caratteristica di trasferimento: da essa si ricavano ad esempio il valore della tensione minima con uscita all'uno logico (V_{OH}), massima con uscita allo zero logico (V_{OL}) e così via, fino a determinare i margini di rumore. La caratteristica di trasferimento tipica dell'inverter CMOS è fornita nella figura 5.57, dove sono anche mostrati i margini di rumore (definiti mediante il punto in cui la pendenza della tangente è pari a 1) e il punto di inversione. Idealmente lo zero logico corrisponde alla tensione più bassa del circuito (il riferimento zero di potenziale) mentre l'uno logico corrisponde alla tensione di alimentazione V_{DD}: in questo modo si ottiene il massimo intervallo di tensione possibile tra le due uscite alta e bassa, ovvero si massimizza lo *swing logico*. Come si può osservare nella figura 5.57, l'inverter CMOS ha una caratteristica di trasferimento praticamente ideale dal punto di vista delle uscite logiche: lo "0" logico è esattamente alla tensione nulla e lo "1" alla tensione V_{DD}. È anche fattore di merito dell'inverter che esso sia il più simmetrico possibile, ovvero che i due margini di rumore per l'uscita alta e bassa siano il più ampi possibil, tra loro uguali, e che la tensione di inversione V_{inv} sia al centro dello swing logico. L'inverter CMOS può essere progettato in maniera da avere un comportamento ottimo: si dimostrerà nel seguito che i margini di rumore sono legati l'uno alla tensione soglia del transistore *p*MOS e l'altro a quella del transistore *n*MOS: essi si possono

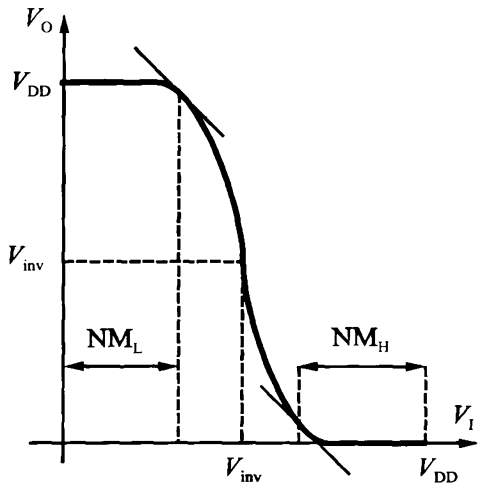


Figura 5.57 Caratteristica di trasferiemento dello stadio inverter in tecnologia CMOS.

quindi rendere uguali progettando i due transistori in maniera da avere la stessa tensione di so
(a meno del segno): $V_{th,p} = -V_{th,n}$. Si dimostrerà anche che per avere la tensione di invers
al centro dello swing logico, si deve avere $\beta_n = \beta_p$. In definitiva lo stadio è ottimo se i disposit
sono perfettamente *complementari*.
Passiamo ora a comprendere il funzionamento dello stadio. Si osservi che, perché almeno uno
due MOS dello stadio possa entrare in conduzione dove serva, deve essere comunque $V_{DD} > V_{th}$
quindi $-V_{DD} < V_{th,p}$): in questo modo lo *n*MOS può conduerre quando la sua tensione gate-so
è pari a V_{DD} e il *p*MOS conduce se invece la sua tensione gate-source è uguale a $-V_{DD}$.
Osserviamo anche che per costruzione:

$$\begin{aligned} V_{GS,n} &= V_I \\ V_{GS,p} &= V_I - V_{DD} \\ V_{DS,n} &= V_O \\ V_{DS,p} &= V_O - V_{DD} \end{aligned}$$

e, come si è detto, poiché lo stadio va studiato a vuoto (o con carico capacitivo):

$$I_{DS,n} = -I_{DS,p}.$$

Si supponga in un primo momento che in ingresso sia presente la tensione corrispondente a
logico e vediamo come lo stadio realizza in uscita lo zero logico (inversione "1"-"0").
$V_I = V_{DD} =$ "1", allora

- il *p*MOS (pull–up) è interdetto ($V_{GS,p} = 0$)
- lo *n*MOS (pull–down) conduce ($V_{GS,n} = V_{DD} > V_{th,n}$)

Se il transistore *p*MOS è interdetto allora deve essere $I_{DS,p} = 0$ e, di conseguenza, anche I_{DS}
Pur essendo lo *n*MOS in conduzione ($V_{GS,n} > V_{th,n}$), esso ha però corrente di drain nulla.
modo per realizzare contemporaneamente queste due richieste è che *anche la tensione*

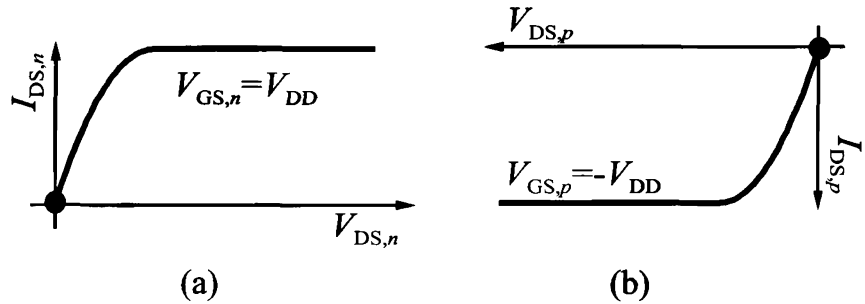


Figura 5.58 (a) Condizione di lavoro dello *n*MOS (pull–up) con ingresso "1". (b) Condizione di lavoro del *p*MOS (pull–down) con ingresso "0".

nulla, come mostrato nella figura 5.58 (a), ovvero che lo *n*MOS si trovi nel punto di funzionamento corrispondente all'origine della caratteristica corrispondente a $V_{GS,n} = V_{DD}$.
Se poi $V_{DS,n} = 0$, allora è anche $V_O = 0$, ovvero l'uscita si trova allo zero logico. Si osservi che la tensione di uscita è tenuta bassa dal fatto che lo *n*MOS è in conduzione pur avendo corrente nulla: per questo esso prende il nome di transistore di pull-down (in inglese pull–down significa tirare giù). Dal punto di vista fisico si può dire che il canale dello *n*MOS crea un collegamento ohmico tra l'uscita e la massa del circuito, e su questo collegamento ohmico la caduta di tensione è nulla: il transistore *n*MOS si comporta quindi come un corto circuito mentre il transistore *p*MOS è un circuito aperto.
Analogamente si supponga che in ingresso sia presente la tensione corrispondente allo zero logico. Allora:

- lo *n*MOS (pull–down) è interdetto ($V_{GS,n} = 0$)
- il *p*MOS (pull–up) conduce ($V_{GS,p} = -V_{DD} < V_{th,p}$)

In questo caso $I_{DS,n} = 0$ e, di conseguenza, anche $I_{DS,p} = 0$. L'unico modo perché il *p*MOS conduca pur avendo corrente di drain nulla è che *anche la tensione* $V_{DS,p}$ *sia nulla*, come mostrato nella figura 5.58 (b), ovvero che il *p*MOS si trovi nel punto di funzionamento corrispondente all'origine della caratteristica corrispondente a $V_{GS,p} = -V_{DD}$. Essendo $V_{DS,p} = 0$, In questo caso si ha $V_O = V_{DD}$, ovvero l'uscita è all'uno logico. Il transistore *p*MOS mantiene l'uscita alla tensione alta creando un collegamento ohmico con la tensione di alimentazione, ovvero comportandosi come un corto circuito: esso è quindi chiamato transistore di pull–up (in inglese pull–up significa tirare su).
Dal punto di vista circuitale l'inverter CMOS si può assimilare ad una serie di due interruttori in communtazione, pilotati dalla tensione gate-source dei due transistori, che non sono mai entrambi chiusi o aperti (cfr. la figura 5.59). In questo modo la corrente statica che attraversa lo stadio è sempre nulla, ovvero in condizioni statiche nello stadio non viene dissipata potenza elettrica.

Approfondimento 5.5 Transcaratteristica dell'inverter CMOS

Nell'approfondimento 5.4 si è studiato il comportamento dell'inverter solamente nel caso in cui gli ingressi sono segnali logici integri, ovvero pari alla tensione V_{DD} per l'uno logico e alla tensione 0 per lo zero logico. Come valutare però la tensione di inversione V_{inv} dello stadio? È necessario estendere l'analisi dello stadio ai valori della tensione di ingresso variabili tra 0 e V_{DD}, ovvero occorre *valutare la transcaratteristica dell'inverter.*
Dimostreremo ora che per V_I tra 0 e V_{DD} l'inverter passa attraverso cinque regioni di funzionamento, caratterizzate nel modo seguente e rappresentate nella figura 5.60:

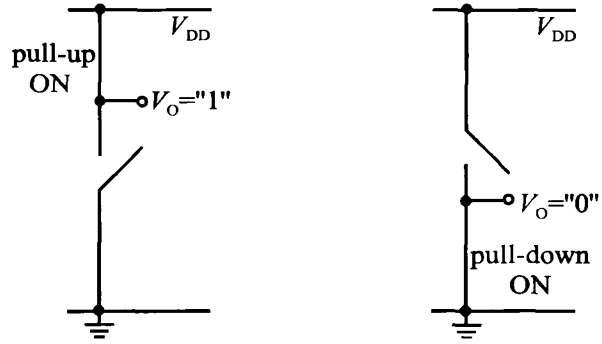


Figura 5.59 Interperetazione dell'inverter CMOS come un doppio interruttore comandato.

1. il *p*MOS è in zona lineare, lo *n*MOS in saturazione (uscita circa pari a "1" logico)
2. il *p*MOS è in zona triodo, lo *n*MOS in saturazione (inizio della inversione)
3. il *p*MOS e lo *n*MOS sono entrambi in saturazione (inversione)
4. il *p*MOS è in saturazione, lo *n*MOS in zona triodo (fine della inversione)
5. il *p*MOS è in saturazione, lo *n*MOS in zona lineare (uscita circa pari a "0" logico)

La regione 1. e la regione 5. corrispondono ai due estremi della transcaratteristica, già vist nell'approfondimento precedente.
Per dimostrare questa affermazione ricordiamo innanzitutto che per lo stadio in esame valgo le relazioni costitutive (5.97) e (5.98). Si consideri per primo il comportamento del disposit *n*MOS con tensione di ingresso variabile tra 0 e V_{DD}. Poiché $V_{GS,n} = V_I$, al crescere della tensio di ingresso il dispositivo *n*MOS viene portato dallo stato di interdizione ($V_{GS,n} = 0$) allo stato ($V_{GS,n} = V_{DD}$). Le caratterisitche dello *n*MOS sono riportate nella figura 5.61 (a).
Passando al dispositivo *p*MOS si osserva un comportamento complementare. Poiché $V_{GS,p} = V_I - V_{DD}$, al crescere della tensione di ingresso il dispositivo *p*MOS viene portato dallo stato ($V_{GS,p} = -V_{DD}$) allo stato di interdizione ($V_{GS,p} = 0$). Le caratterisitche del *p*MOS sono riporta nella figura 5.61 (b) in funzione della tensione di drain $V_{DS,p} = V_O - V_{DD}$.
Prima di passare alla valutazione della transcaratteristica, si osservi che la corrente del disposit *p*MOS è uguale e opposta alla corrente dello *n*MOS: la figura 5.61 (b) rappresenta quindi a l'opposto della corrente di drain dello *n*MOS e può essere disegnata sul grafico della $I_{DS,n}$ funzione della tensione di uscita dello stadio $V_O = V_{DS,n}$, ovvero sugli stessi assi della figu 5.61 (a). Per fare questo si osservi che, poiché $V_{DS,p} = V_O - V_{DD}$, il grafico va traslato ver destra della quantità V_{DD} e infine ribaltato rispetto all'asse x per tener conto del segno oppos della corrente. Si ottiene il grafico di figura 5.62.
Per determinare la tensione di uscita dello stadio al variare della tensione di ingresso (transcara teristica), si devono ora *intersecare le due famiglie di curve* rappresentate nelle figure 5.61 5.62. In particolare per ogni determinato valore della tensione di ingresso, si deve scegliere la spondente curva nelle figure 5.61 (a) e 5.62 e intersecarle. Si osservi che al crescere della tens di ingresso V_I lo *n*MOS entra in conduzione sempre più forte mentre il *p*MOS si spegne. Le cu vanno intersecate quindi a partire da quella a corrente minore del *n*MOS con quella a corre corrente massima per il *p*MOS fino ad arrivare ad intersecare quella a corrente massima p *n*MOS e quella a corrente minore per il *p*MOS. Il grafico delle possibili intersezioni è mos nella figura 5.63.
Si riconoscono le cinque regioni che vengono attraversate nella transizione da "0" a "1" in ingres

1. Per tensioni di ingresso vicine a zero, il *p*MOS è in zona lineare, lo *n*MOS in saturazione

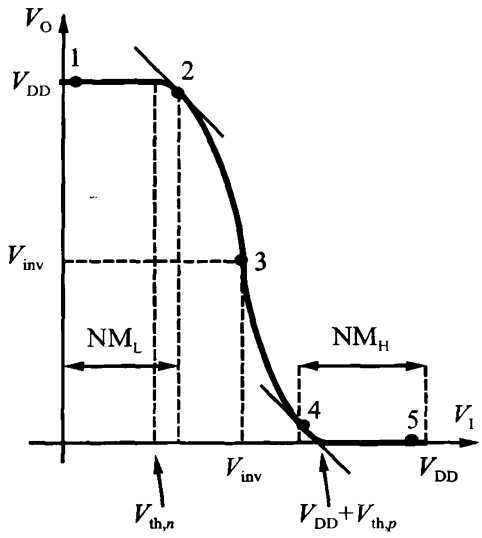


Figura 5.60 Regioni di funzionamento dell'inverter CMOS.

corrente in entrambi è quasi nulla. Questa condizione è ancora simile alla condizione "0"–"1" con ingresso esattamente pari a zero e tensione di uscita pari a "1" logico.

2. Al crescere della tensione di ingresso il *p*MOS entra in zona triodo, lo *n*MOS è ancora in saturazione. Poiché la corrente che scorre nei due dispositivi non è più trascurabile, aumenta la caduta di tensione tra il drain e il source del *p*MOS e, di conseguenza, la tensione di uscita inizia a calare. Si ha la condizione di inizio della inversione.
3. All'inversione il *p*MOS e lo *n*MOS sono entrambi in saturazione.
4. Aumentando ancora la tensione di ingresso V_I, il *p*MOS entra in saturazione e tende a spegnersi, mentre lo *n*MOS entra in zona triodo. La tensione di uscita continua a scendere per effetto della crescente tensione tra il drain e il source del *p*MOS (fine della inversione).
5. Quando la tensione di ingresso si avvicina al valore V_{DD}, il *p*MOS è in saturazione, mentre lo *n*MOS in zona lineare. In questa condizione la tensione di uscita è praticamente nulla e lo stadio opera in condizioni simili alla condizione "1"–"0".

La transcaratteristica risultante è quella già riportata nella figura 5.60. Si osservi che all'aumentare della tensione di gate la transione tra la regione 1 e la regione 2 è graduale, e la tensione di uscita si mantiene attorno al valore logico "1", garantendo un buon margine di rumore. Inoltre fintanto che $V_I = V_{GS,n} < V_{th,n}$, l'uscita si mantiene *esattamente* all'uno logico; analogamente per tutte le tensioni $V_{GS,p} > V_{th,p}$, ovvero per $V_I > V_{DD} + V_{th,p}$ l'uscita si mantiene *esattamente* a zero: questo dimostra che il valore della tensione di soglia dei MOS è legato ai margini di rumore e che se la tensione di soglia viene diminuita si degrada il comportamento complessivo dello stadio[26]. La transizione tra la zona 2 e la zona 4 (ovvero la vera e propria inversione dell'inverter) è invece molto brusca, e lo stadio si porta rapidamente dallo stato "1" allo stato "0" in uscita, come mostrato sempre nel grafico della figura 5.60. In definitiva l'inverter CMOS ha caratteristiche quasi ideali: consumo virtualmente nullo negli stati "1" e "0" in uscita, poiché la corrente è nulla, transizione molto rapida nella inversione e margini di rumore controllabili per via tecnologica mediante la

[26] Ovviamente per il corretto funzionamento dello stadio deve essere $V_{th,n} < V_{DD}/2$. La scelta della tensione di alimentazione è quindi legata anche al valore delle tensioni di soglia.

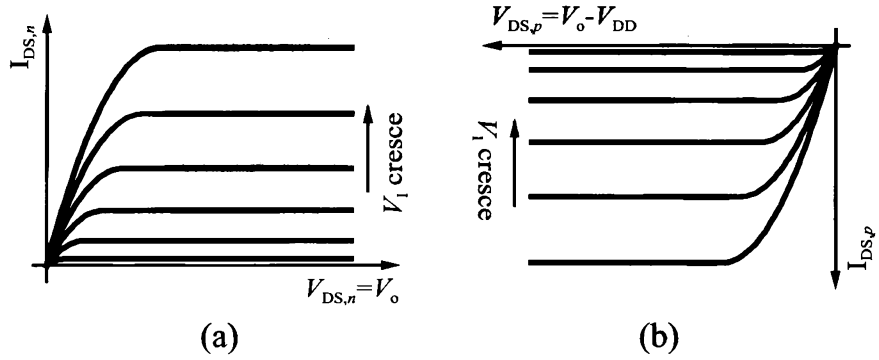


Figura 5.61 (a) Condizione di lavoro dello *n*MOS (pull–up) con ingresso V_I crescente. (b) Condizione di lavoro del *p*MOS (pull–down) con ingresso V_I crescente.

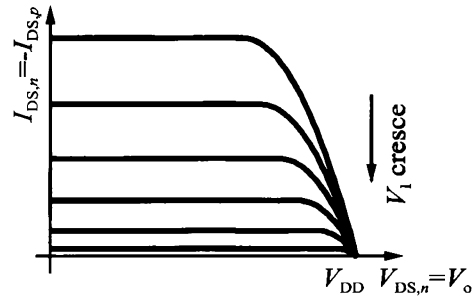


Figura 5.62 Caratteristiche del *p*MOS riportate sugli assi dello *n*MOS.

tensione di soglia dei dispositivi.

Esempio 5.6 Dimensionamento dell'inverter CMOS

Si consideri sempre l'inverter CMOS dell'approfondimento 5.4. Supponendo che

$$V_{th,p} = -V_{th,n}; \qquad t_{ox} = 50\,\text{nm}$$
$$\mu_n = 1294\,\text{cm}^2/\text{Vs}; \qquad \mu_p = 435\,\text{cm}^2/\text{Vs}$$

si chiede di determinare quali dimensioni devono avere i due MOS dello stadio perché si abbia $V_{inv} = V_{DD}/2$.

All'inversione ($V_I = V_O = V_{DD}/2$) l'inverter opera nella regione 3 della transcaratteristica (vedi la figura 5.60) in cui *entrambi i MOS sono in saturazione*. Definendo allora $\beta_n = \dfrac{W_n \mu_n C_{ox}}{L_n}$ e $\beta_p = \dfrac{W_p \mu_p C_{ox}}{L_p}$, e uguagliando le correnti in saturazione nei due dispositivi, deve essere:

$$\frac{1}{2}\beta_n (V_{GS,n} - V_{th,n})^2 = \frac{1}{2}\beta_p (V_{GS,p} - V_{th,p})^2$$

Sostituendo poi $V_{GS,n} = V_I$ e $V_{GS,p} = V_I - V_{DD}$ si ha:

$$\frac{1}{2}\beta_n (V_I - V_{th,n})^2 = \frac{1}{2}\beta_p (V_I - V_{DD} - V_{th,p})^2$$

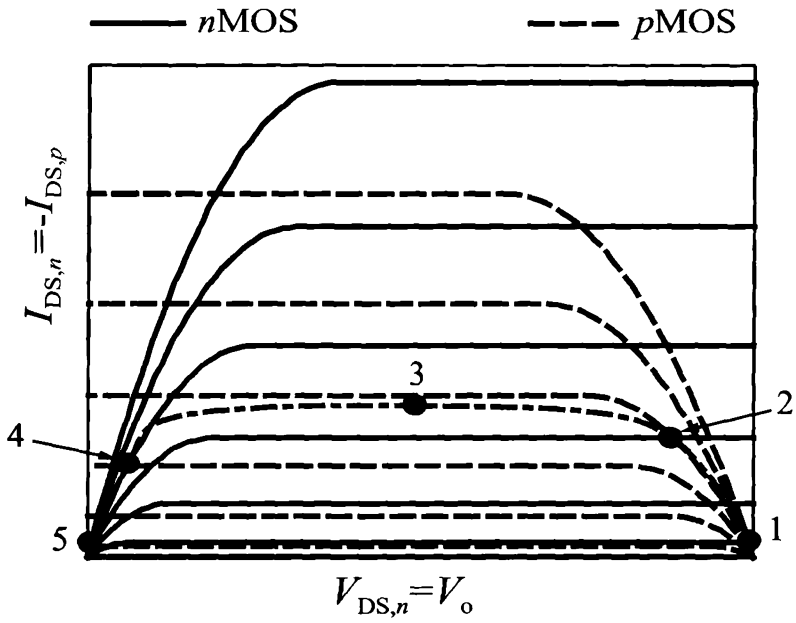


Figura 5.63 Intersezione delle caratteristiche del *p*MOS e dello *n*MOS.

Risolvendo rispetto a $V_I = V_{inv}$ si ottiene:

$$V_{inv} = \frac{V_{DD} + V_{th,p} + V_{th,n}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

Poiché i due transistori sono progettati in modo che $V_{th,p} = -V_{th,n}$ si ha $V_{inv} = V_{DD}/2$ se, come anticipato:

$$\beta_n = \beta_p$$

Per realizzare questa condizione, è necessario che le dimensioni fisiche dei dispositivi siano scelte appositamente. Supponendo per prima cosa che i due MOS siano realizzati nello stesso circuito integrato, essi avranno uguale spessore di ossido e quindi C_{ox} *è uguale in entrambi.* Per ottimizzare la corrente e la transconduttanza dei due dispositivi è anche opportuno scegliere per entrambi i MOS la *lunghezza di canale pari alla minima dimensione* L_{min} consentita dalla tecnologia $L_n = L_p = L_{min}$. Allora la condizione $\beta_n = \beta_p$ è verificata se:

$$W_n \mu_n = W_p \mu_p$$

ovvero

$$\frac{W_p}{W_n} = \frac{\mu_n}{\mu_p} = 2.97 \simeq 3$$

Dalla relazione trovata si deduce che per avere una vera simmetria nello stadio il *pull-up deve avere larghezza pari a 3 volte il pull-down*, in modo da compensare la differenza tra le mobilità degli

elettroni e delle lacune nei due canali conduttivi. Poiché i due MOS devono rispettare un preciso *rapporto di periferia* si parla di circuiti logici a rapporto o, semplicemente, logiche a rapporto.

Capitolo 6

Il MOSFET come amplificatore

6.1 Caratteristiche di un amplificatore e VTC

Utilizzare dispositivi con caratteristiche fortemente non lineari, quali i transistori MOS, per realizzare degli stadi amplificatori richiede un'attenta analisi dei parametri che caratterizzano un amplificatore ideale: in questo modo diventa possibile attuare delle opportune strategie circuitali per riuscire a garantire una linearità accettabile. Un amplificatore deve essere innanzitutto in grado di trasferire potenza dall'alimentazione al carico dove il segnale (in uscita) deve essere una replica fedele ed "amplificata" del segnale in ingresso. Il trasferimento di potenza può essere attuato ricorrendo al guadagno presente nei dispositivi attivi e non ad esempio in un trasformatore. Può essere conveniente a seconda dell'applicazione dell'amplificatore ricorrere ad un modello elettrico equivalente sia della sorgente del segnale che si vuole amplificare sia dello stadio amplificatore connesso al carico sul quale si vuole trasferire, tramite il segnale, la potenza.

L'amplificazione inoltre deve essere il più possibile *indipendente* dal generatore associabile al segnale in ingresso e dal carico applicato sull'uscita. Supponendo di utilizzare in ingresso o un segnale in tensione o in corrente e di pilotare un carico o in tensione o in corrente, il comportamento ideale dell'amplificatore può essere ottenuto con opportuni valori della resistenza in ingresso dell'amplificatore R_{in} (ovvero la resistenza che il segnale "vede" all'ingresso dell'amplificatore) e della resistenza in uscita R_{out} (ovvero la resistenza presente nel circuito equivalente dell'amplificatore connesso al carico):

- ingresso in corrente (eq. Norton) $R_{\text{in}} \to 0$
- ingresso in tensione (eq. Thevenin) $R_{\text{in}} \to \infty$
- uscita in corrente (eq. Norton) $R_{\text{out}} \to \infty$
- uscita in tensione (eq. Thevenin) $R_{\text{out}} \to 0$

Combinando la configurazione di ingresso con quella di uscita sono quindi possibili *quattro* tipi di amplificatore:

1. *Tensione* $v_{\text{in}} - v_{\text{out}}$;
2. *Transconduttanza* $v_{\text{in}} - i_{\text{out}}$;

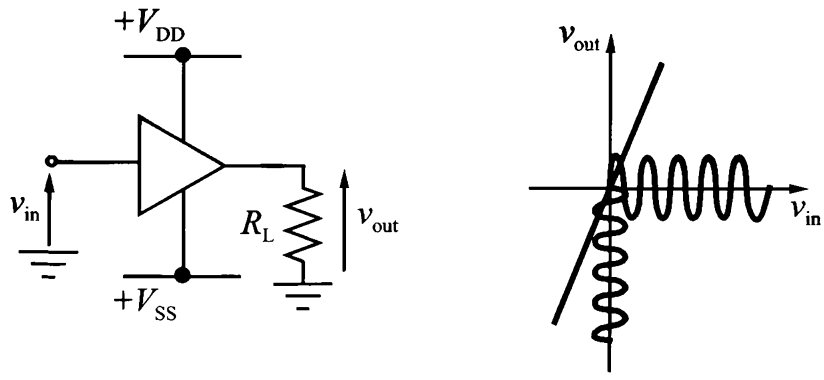


Figura 6.1 Schema di uno stadio amplificatore connesso al carico e della sua transcaratteristica idealmente esercitata da un segnale in ingresso.

3. *Corrente* $i_{in} - i_{out}$;
4. *Transresistenza* $i_{in} - v_{out}$.

Seppur generali, i concetti che saranno sviluppati faranno riferimento dapprima ad un amplificatore di tensione. Il funzionamento di un amplificatore di tensione è descrivibile sul piano (v_{in},v_{out}) tramite la sua transcaratteristica detta anche VTC (*V*oltage *T*ransfer *C*haracteristic) ovvero $v_{out} = f(v_{in})$. La transcaratteriscita dovrebbe essere descritta da una relazione quanto più possibile *lineare* dove la tensione di uscita risulti A_v volte la tensione di ingresso come illustrato nella figura 6.1: infatti la linearità della VTC garantisce che qualora all'amplificatore sia applicato un segnale in ingresso questo venga trasferito sull'uscita fedelmente.

$$v_{out} = A_v \cdot v_{in} \tag{6.1}$$

In realtà la transcaratteristica di un amplificatore reale realizzata con dispositivi a semiconduttore può essere quasi lineare solo su un intervallo di tensioni limitato e molto spesso l'aspetto complessivo della VTC reale è lontano da quello auspicato. Le principali trasformazioni che la VTC reale subisce possono essere analizzate in termini di:

1. *inversione* $v_{out} = -A_v \cdot v_{in}$, se la VTC risulta invertente in presenza di una tensione positiva in ingresso sull'uscita si viene ad avere un valore di tensione negativo;
2. *offset* $v_{out} = V_{off} + A_v \cdot v_{in}$, se la VTC ha un offset in assenza di una tensione in ingresso sull'uscita si viene ad avere un valore costante (positivo o negativo) di tensione ovvero la VTC risulta traslata del valore V_{off};
3. *saturazione*, la presenza di saturazione si evidenzia sulla VTC con una tensione $v_{out} = V_{max}$ costante qualora la tensione in ingresso abbia superato un valore massimo, il fenomeno si può presentare in modo simile anche nel semipiano delle tensioni negative.

Tali non idealità possono presentarsi *simultaneamente* come evidenziato nella figura 6.2, inoltre la VTC reale è sostanzialmente *non lineare* e alla luce di tutte queste

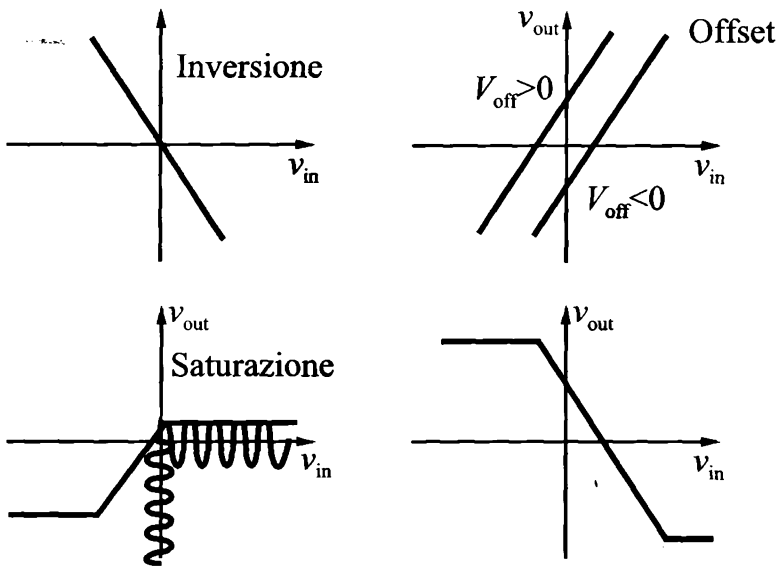


Figura 6.2 Esempi di VTC reale: con inversione, con offset, con saturazione e con inversione, offset e saturazione insieme.

considerazioni appare evidente come solamente tratti di ampiezza limitata possano essere considerati lineari.

Data l'importanza della VTC nel valutare la funzionalità di uno stadio amplificatore è necessario stabilire dapprima in quale regioni di funzionamento il transistore MOSFET si venga a trovare al variare della tensione in ingresso. Si deve ancora precisare che la VTC fino ad ora descritta viene anche detta statica in quanto può essere ottenuta facendo variare in modo quasi statico la tensione di ingresso e valutando la corrispondente tensione di uscita. In tal senso la VTC di un amplificatore può essere ottenuta utilizzando il *modello statico* per i dispositivi a semiconduttore utilizzati nello stadio amplificatore. Da un punto di vista operativo la VTC viene calcolata supponendo di avere un ingresso in *continua* e utilizzando per il MOSFET l'equazione congrua con la regione di funzionamento nella quale si viene a trovare.

Riscrivendo per chiarezza le equazioni che costituiscono il modello statico per le regioni di funzionamento di un n-MOSFET si deve sottolineare come l'effetto delle tensioni di alimentazione e dell'ingresso v_{in} sia quello di determinare i valori delle tre tensioni V_{GS}, V_{DS} e V_{BS} come illustrato nella figura 6.3.

▷ *interdizione* ($V_{GS} < V_{th}$)

$$I_D = 0$$

▷ *quadratica* ($V_{GS} > V_{th}, V_{DS} < (V_{GS} - V_{th})$)

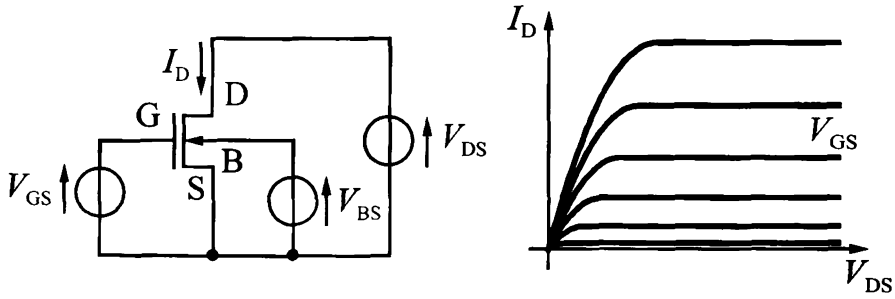


Figura 6.3 MOSFET polarizzato utilizzando tre generatori di tensione indipendenti: V_{GS}, V_{DS} e V_{BS}.

$$I_D = \frac{W}{L}\mu_n C_{ox}\left[(V_{GS} - V_{th})\,V_{DS} - \frac{V_{DS}^2}{2}\right]$$

▷ *saturazione* $(V_{GS} > V_{th}, V_{DS} > (V_{GS} - V_{th}))$

$$I_D = \frac{W}{2L}\mu_n C_{ox}\,(V_{GS} - V_{th})^2\,[1 + \lambda\,(V_{DS} - V_{DSsat})]$$

Inoltre la V_{th} a causa del *body effect* dipende dalla V_{BS}:

$$V_{th}(V_{BS}) = V_{th0} + \gamma\left(\sqrt{2\Phi_p - V_{BS}} - \sqrt{2\Phi_p}\right)$$

$$\gamma = \frac{\sqrt{2\epsilon_S q N_A}}{C_{ox}}$$

In realtà dato che generalmente le tensioni di alimentazioni sono fisse le tre tensioni che insistono sul MOSFET sono solo funzione della v_{in}, e la valutazione della VTC deve avvenire al variare della tensione continua in ingresso e della corrispondente *regione di funzionamento* nella quale si trova il MOSFET. Una strategia relativamente semplice per il calcolo della VTC consiste nel suddividere il piano (v_{in},v_{out}) in funzione delle regioni di funzionamento del MOSFET.

Si vedrà che nei principali stadi amplificatori la regione dove il MOSFET risulta *saturo* è anche quella che presenta la maggior *linearità* della VTC. Inoltre lo studio della VTC dell'amplificatore permette di determinare in modo ottimale il *punto di lavoro* nel quale deve essere posto l'amplificatore in assenza del segnale in ingresso: in altri termini il segnale sarà sempre visto come sovvrapposizione di una componente variabile al punto di lavoro in continua. La scelta del punto di lavoro dovrà essere fatta sul tratto *lineare* della VTC, come in figura 6.4, per garantire la *massima dinamica*: l'ampiezza del segnale sull'uscita deve esercitare solamente il tratto a maggior linearità della transcaratteristica per evitare una progressiva distorsione del segnale, quindi in presenza di un segnale in ingresso con dinamica simmetrica è necessario posizionare il punto di lavoro nel tratto centrale della regione lineare della VTC. Come illustrato

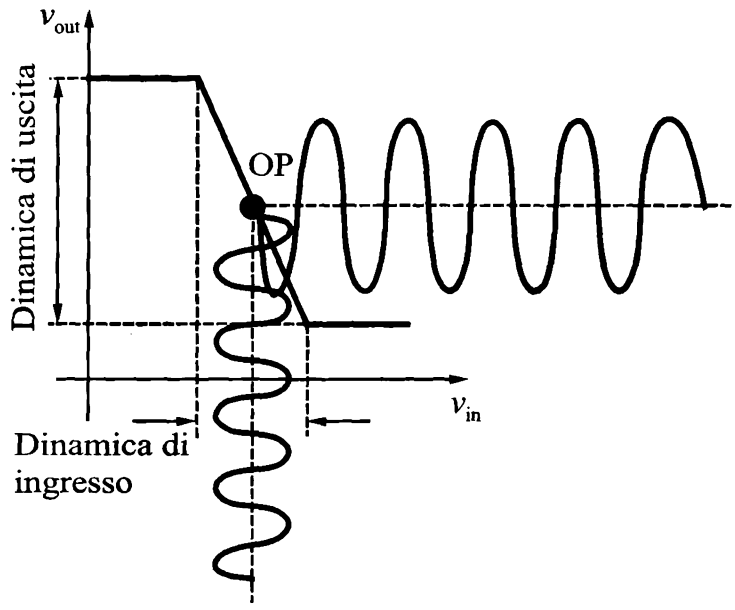


Figura 6.4 VTC di un amplificatore nella quale il punto di lavoro è posizionato sul tratto lineare in modo da garantire la massima dinamica.

figura 6.5 una scelta errata del punto di lavoro riduce la dinamica utile dell'amplificatore a causa della saturazione della tensione di uscita che porta a tagliare superiormente o inferiormente il segnale in uscita all'amplificatore.

6.2 Amplificatore a Source Comune

Il primo stadio che verrà studiato è quello a Source Comune (CS) il cui schema è illustrato in figura 6.6 in assenza di carico (a vuoto). Il circuito viene alimentato tramite le due tensioni $V_{SS} < 0$ e $V_{DD} > 0$ generalmente simmetriche. Il MOSFET a canale n viene connesso con $V_B = V_S$ in modo da inibire l'effetto di substrato e con il Drain connesso all'uscita. Inizialmente si analizza il circuito per determinare la *VTC* definita la Vv_{in} e la v_{out}.

Al fine di rendere più chiara l'analisi di tale circuito si definisce dapprima come tensione continua in ingresso $v_{in} = V_{GG} - V_{SS}$, e successivamente si determina la regione sul piano (v_{out},v_{in}) dove il MOSFET è saturo. In questo modo diventa possibile definire la regione di saturazione imponendo che il MOSFET sia in conduzione ovvero che $v_{in} > V_{th}$ e che simultaneamente valga:

$$V_{DS} > V_{GS} - V_{th}$$

$$v_{out} - V_{SS} > V_{GG} - V_{SS} - V_{th}$$

$$v_{out} > v_{in} + (V_{SS} - V_{th})$$

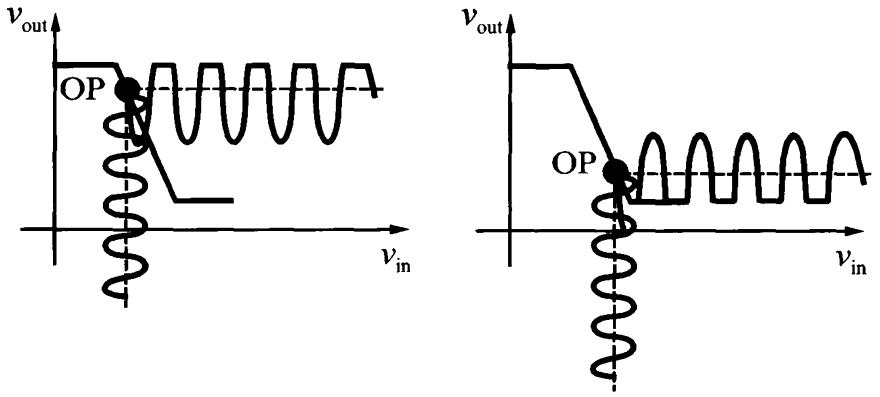


Figura 6.5 Esempio di errato posizionamento del punto di lavoro sulla VTC: nel primo caso il segnale viene tagliato superiormente, nel secondo inferiormente.

La retta definita dai seguenti punti (V_{th},V_{SS}) e $(V_{DD} - V_{SS} + V_{th},V_{DD})$ divide la parte del piano (v_{out},v_{in}) con $v_{in} > V_{th}$ tra la regione dove in MOSFET è saturo e quella dove ha caratteristica quadratica come illustrato nella figura 6.7.

Suddiviso il piano della VTC nelle tre regioni nelle quali il MOSFET è rispettivamente interdetto, saturo e quadratico, è possibile ottenere la VTC notando che con l'amplificatore a vuoto, la corrente nella resistenza R_D eguaglia la corrente del MOSFET e cioè, indipendentemente dalla condizione del transistore, $I_D = I_R$. L'analisi procede valutando la VTC come funzione definita a tratti:

1. MOSFET *interdetto*: in questa regione non scorre corrente nella resistenza R_D facendo sì che la tensione sull'uscita raggiunga il valore V_{DD}

$$v_{in} < V_{th}$$

$$I_R = 0$$

$$v_{out} = V_{DD}$$

2. MOSFET *saturo*: in questa condizione la tensione di uscita si riduce rapidamente a crescere della tensione v_{in}; per ottenere la VTC reale la corrente in saturazione andrebbe corretta con il termine dovuto alla modulazione della lunghezza di canale ma per semplicità nel seguito questo non verrà fatto

$$v_{in} > V_{th}$$

$$v_{out} > v_{in} + (V_{SS} - V_{th})$$

$$\frac{V_{DD} - v_{out}}{R_D} \approx \frac{W}{2L}\mu_n C_{ox}(v_{in} - V_{th})^2$$

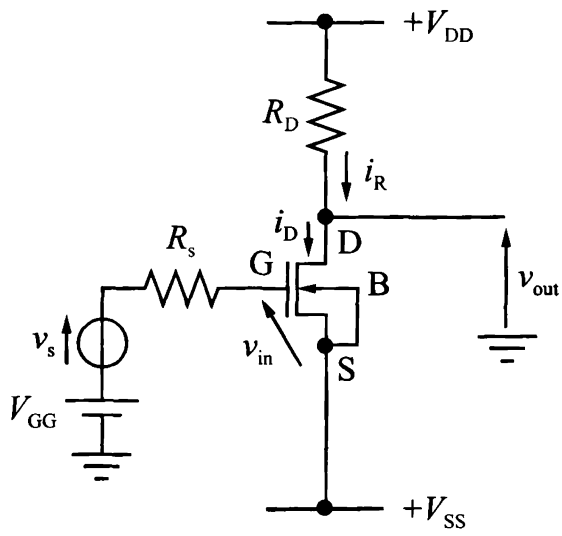


Figura 6.6 Stadio a Source Comune a vuoto.

3. MOSFET *quadratico*:

$$v_{in} > V_{th}$$

$$v_{out} < v_{in} + (V_{SS} - V_{th})$$

$$\frac{V_{DD} - v_{out}}{R_D} = \frac{W}{L}\mu_n C_{ox}\left[(v_{in} - V_{th})(v_{out} - V_{SS}) - \frac{(v_{out} - V_{SS})^2}{2}\right]$$

Tracciando la transcaratteristica sul piano (v_{out},v_{in}) come in figura 6.8 si osserva che la regione di maggior linearità è quella dove il MOSFET è saturo.

Alternativamente è possibile procedere *graficamente* per determinare la VTC, notando che nello stadio a source comune valgono le seguenti relazioni:

$$I_R = I_D$$

$$v_{out} = V_{DS} + V_{SS}$$

$$I_D = \frac{(V_{DD} - v_{out})}{R_D}$$

$$v_{in} = V_{GG} - V_{SS}; \quad I_G = 0$$

Infatti dato che le correnti nella resistena R_D e nel MOSFET sono uguali risulta possibile tracciare sullo stesso piano (v_{out},I_D) la retta di carico per resistenza R_D e le

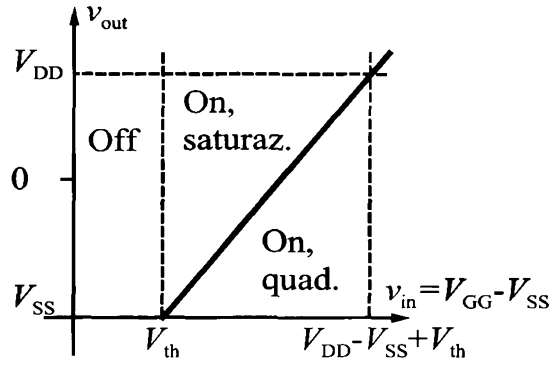


Figura 6.7 Regioni di funzionamento del MOSFET sul piano (v_{in}, v_{out}).

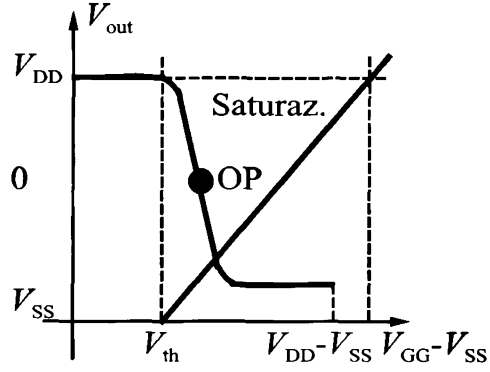


Figura 6.8 Transcaratteristica dello stadio a source comune e regioni di funzionamento del MOS.

caratteristiche di uscita del MOSFET. Ora al variare di $V_{GG} - V_{SS}$ si valuta l'intersezione fra la retta di carico e la corrispondente caratteristiva di uscita del transistore ottenendo graficamente la funzione $v_{out} = f(v_{in})$ come indicato nella figura 6.9.

6.2.1 Polarizzazione stadio CS

Fissate le tensioni V_{DD} e V_{SS} la scelta del punto di lavoro è determinata dalla scelta del valore del generatore V_{GG} e dato che si vuole posizionare il punto di lavoro nel tratto lineare della VTC si può ipotizzare il MOSFET in regione di saturazione.

In questa condizione con l'amplificatore a vuoto le due correnti diventano:

$$I_D = \frac{W}{2L}\mu_n C_{ox} (V_{GG} - V_{SS} - V_{th})^2$$

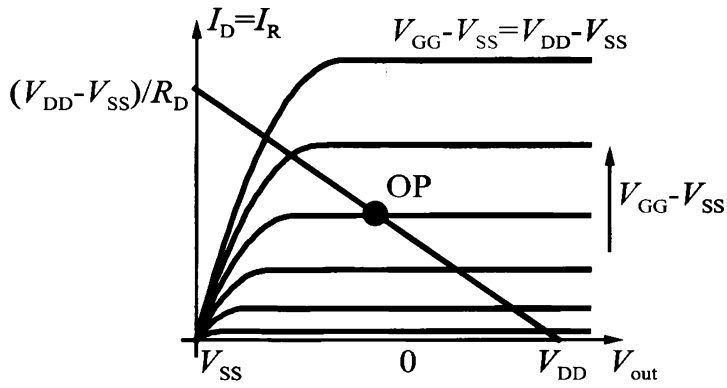


Figura 6.9 Punti di lavoro dell'amplificatore ottenuti come intersezioni della retta di carico e delle caratteristiche di uscita del MOSFET.

$$I_R = \frac{V_{DD} - v_{out}}{R_D}$$

Da un'analisi qualitativa della VTC è possibile posizionare il punto di lavoro nel tratto centrale della zona lineare: in questo modo si garantisce un'*ampia* dinamica del segnale d'uscita. Inoltre dato che la v_{out} nella VTC può variare tra V_{DD} e V_{SS} il punto con $v_{out} = 0$ rappresenta una scelta opportuna anche perché in assenza di segnale in ingresso l'uscita v_{out} è senza offset. Imporre $v_{out} = 0$ permette di determinare il valore di V_{GG} che garantisca il punto di lavoro prefissato:

$$I_D = I_R = \frac{W}{2L}\mu_n C_{ox}\left(V_{GG} - V_{SS} - V_{th}\right)^2 = \frac{V_{DD}}{R_D}$$

$$V_{GG} = \sqrt{\frac{2V_{DD}}{R_D \frac{W}{L}\mu_n C_{ox}}} + V_{SS} + V_{th}$$

6.2.2 Modello per piccolo segnale

Per valutare come l'amplificatore polarizzato con il valore di V_{GG} calcolato si comporti in presenza di un segnale in ingresso risulta utile ricorrere ad un modello linearizzato del MOSFET che consenta l'utilizzo della sovvrapposizione degli effetti. A tal fine come già presentato nel capitolo 5 se si considera un segnale *piccolo* sovvrapposto alla polarizzazione (ampio segnale) si ottiene un comportamento *lineare* dell'amplificatore, ovvero anche in presenza della forte non linearità del MOSFET vale la sovvrapposizione degli effetti. Matematicamente la corrente totale che scorre nel MOSFET può essere vista come la *somma* del contributo della polarizzazione e del piccolo segnale:

$$I_D\left(V_{GS},V_{DS},V_{BS};v_{gs},v_{ds},v_{bs}\right) \approx$$

$$I_D\left(V_{GS},V_{DS},V_{BS}\right) + i_d\left(v_{gs},v_{ds},v_{bs}\right)$$

La corrente dovuta al piccolo segnale può essere calcolata con la *sovrapposizione degli effetti*:

$$i_d = g_m v_{gs} + g_o v_{ds} + g_{mb} v_{bs}$$

Si definiscono quindi nell'intorno del punto di lavoro OP = (V_{GS},V_{DS},V_{BS}):

▷ *transconduttanza* [S]

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{OP} = \frac{W}{L}\mu_n C_{ox}\,(V_{GS} - V_{th})$$

$$g_m = \sqrt{2\frac{W}{L}\mu_n C_{ox} I_D}$$

▷ *conduttanza di uscita* [S]

$$g_o = \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{OP} = \frac{W}{2L}\mu_n C_{ox}\,(V_{GS} - V_{th})^2\,\lambda$$

$$g_o \approx \lambda I_D$$

▷ *transconduttanza di body* [S]

$$g_{mb} = \left.\frac{\partial I_D}{\partial V_{BS}}\right|_{OP}$$

$$g_{mb} = \frac{\gamma g_m}{2\sqrt{2\Phi_p - V_{BS}}}$$

Utilizzando il circuito equivalente complessivo a *bassa frequenza* per piccolo segnale illustrato nella figura 6.10 che si ottiene come sovvrapposizione delle sorgenti di piccolo segnale è possibile studiare il comportamento dello stadio amplificatore.

A tal fine si sostituisce il circuito equivalente per piccolo segnale nello stadio amplificatore nel quale si tengono conto solo delle componenti di segnale rispetto ai valori in continua, lo schema risultante è presente in figura 6.11. Dato che l'amplificatore è a vuoto la tensione in uscita può essere calcolata direttamente come tensione sulla resistenza R_D e da questa relazione ottenere il guadagno.

$$v_{out} = -g_m v_{in}\,(r_o||R_D)$$

$$A_{v0} = \frac{v_{out}}{v_{in}} = -g_m\,(r_o||R_D)$$

$$A_{v0} \approx -g_m R_D$$

Può essere utile a questo punto, calcolare la dinamica di uscita dell'amplificato-

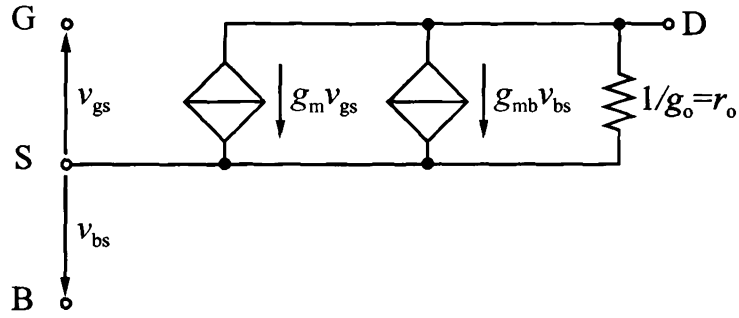


Figura 6.10 Schema equivalente del modello di piccolo segnale.

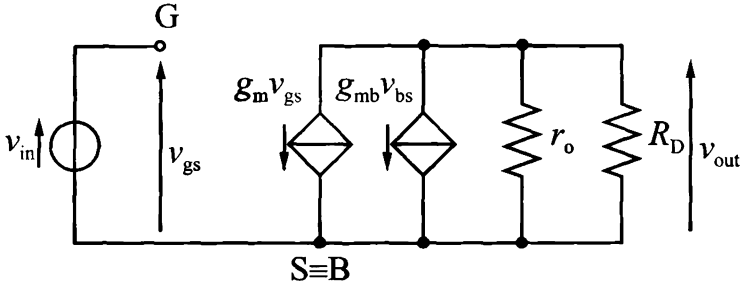


Figura 6.11 Modello di piccolo segnale dello stadio a source comune a vuoto.

re: a vuoto la dinamica è limitata superiormente dall'interdizione del MOSFET e inferiormente dall'uscita dalla saturazione:

$$v_{out,max} = V_{DD}$$

$$v_{out,min} - V_{SS} = V_{GG} - V_{SS} - V_{th}$$

$$v_{out,min} = V_{GG} - V_{th}$$

Se ora proviamo a valutare l'effetto di un carico sull'uscita dell'amplificazione come in figura 6.12, questa determina una riduzione sia dell'amplificazione sia del limite superiore della dinamica lasciando invece inalterato il limite inferiore:

$$v_{out,max} = V_{DD} \frac{R_L}{R_L + R_D}$$

$$A_v = -g_m (r_0 || R_D || R_L)$$

Volendo ora caratterizzare lo stadio a source comune come doppio bipolo è necessario calcolarne la resistenza di ingresso e la resistenza di uscita. A tal fine si applica

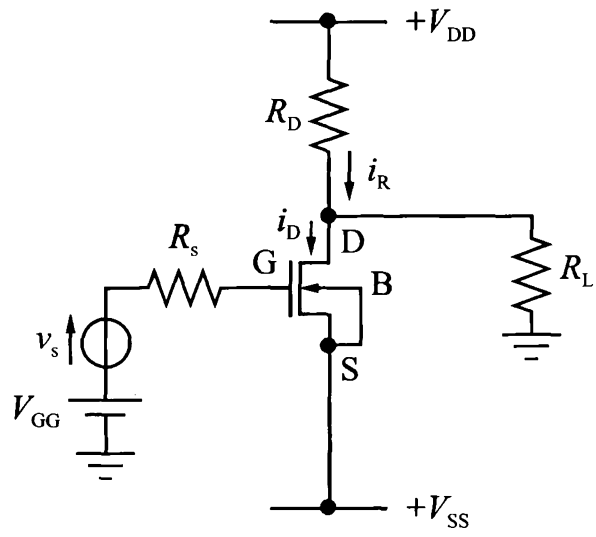


Figura 6.12 Schema dell'amplificatore a source comune con carico R_L.

un generatore di test sull'ingresso come indicato in figura 6.13 e successivamente si calcola il rapporto tra la tensione applicata e la corrente assorbita dall'amplificatore. Sapendo che il gate è isolato si ottiene:

$$i_t = 0 \qquad R_{in} = \frac{v_t}{i_t} = \infty$$

Analogamente per il calcolo della resistenza di uscita si applica un generatore di test sull'uscita come illustrato in figura 6.14 e si calcola il rapporto tra tensione applicata e la corrente erogata:

$$v_{gs} = 0 \qquad v_t = i_t \, (r_o || R_D)$$

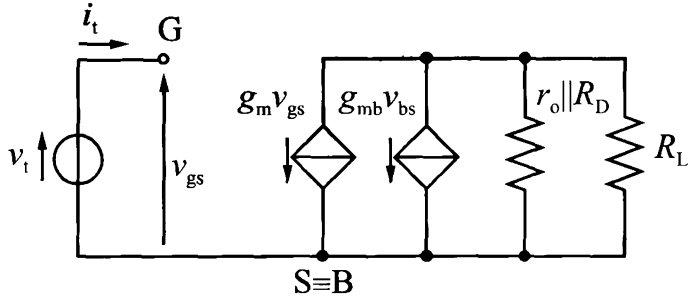


Figura 6.13 Circuito equivalente per il calcolo della resistenza di ingresso.

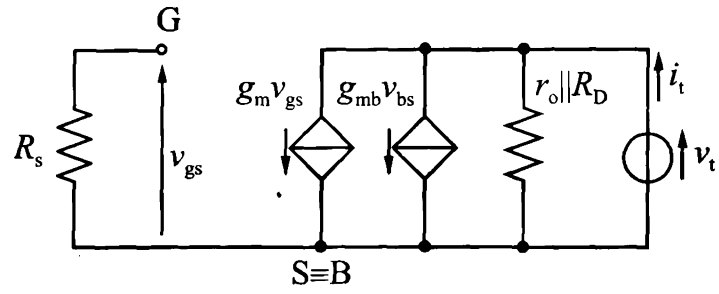


Figura 6.14 Circuito equivalente per il calcolo della resistenza di uscita.

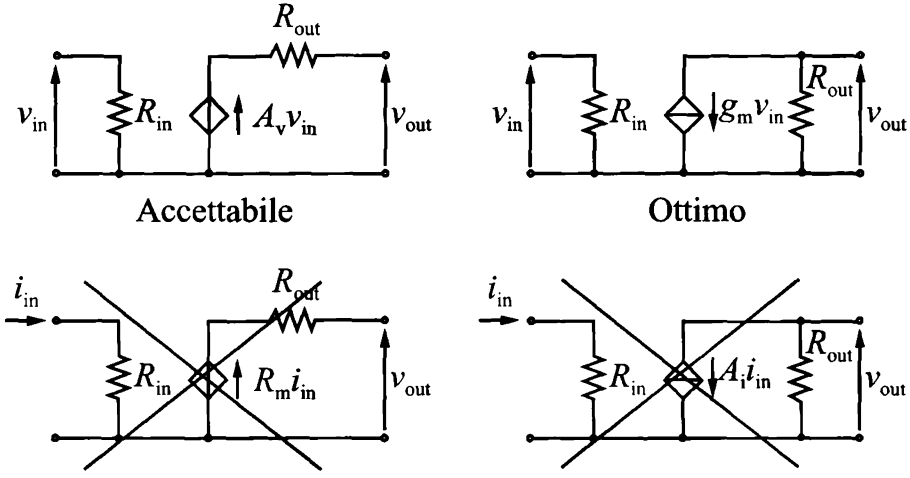


Figura 6.15 Circuito equivalente di uno stadio amplificatore CS nelle quattro possibili configurazioni.

$$R_{out} = \frac{v_t}{i_t} = r_o || R_D$$

Noti i parametri di amplificazione, resistenza di ingresso e di uscita è possibile costruire il doppio bipolo equivalente nelle configurazioni di amplificatore di tensione, di trasconduttanza, di transresistenza e di corrente: si utilizza sull'uscita l'equivalente Thevenin per l'amplificatore di tensione e di transresistenza, e quello di Norton per l'amplificatore di trasconduttanza e di corrente. Come illustrato nella figura 6.15 lo stadio a source comune risulta inutilizzabile sia come amplificatore di transresistenza che come amplificatore di corrente a causa della resistenza di ingresso troppo alta, è invece utilizzabile come amplificatore di tensione anche se la resistenza di uscita è relativamente elevata, trova la sua miglior applicazione come amplificatore di transconduttanza grazie all'elevata resistenza di ingresso e ad una conduttanza di uscita che può essere resa molto bassa.

In sintesi:

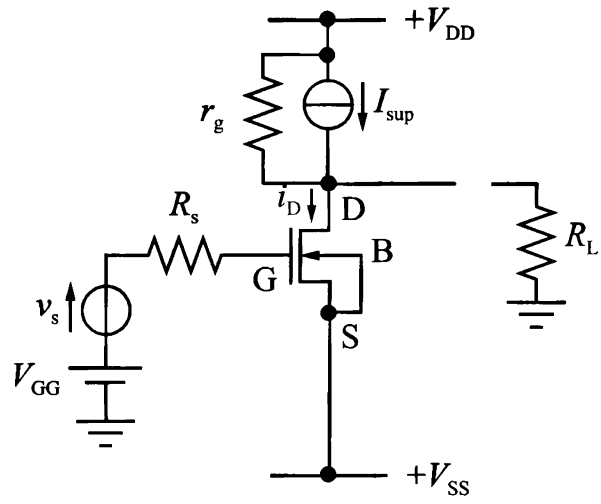


Figura 6.16 Schema dell'amplificatore a source comune con resistenza di drain sostituita da un carico attivo.

- ▷ lo stadio CS è accettabile come *amplificatore di tensione*
- ▷ lo stadio CS è ottimo come *amplificatore di trasconduttanza*
- ▷ lo stadio CS non può essere utilizzato nè come amplificatore di transresistenza nè come amplificatore di corrente

6.2.3 Carico attivo

Lo stadio a source comune presenta un guadagno che è inversamente proporzionale alla corrente di polarizzazione I_D, infatti a vuoto:

$$|A_{v0}| = g_m (r_o || R_D) \approx g_m R_D$$

$$|A_{v0}| \approx g_m R_D = \sqrt{2 I_D \frac{W}{L} \mu_n C_{ox}} \frac{V_{DD}}{I_D} \propto \frac{V_{DD}}{\sqrt{I_D}}$$

Praticamente è possibile *aumentare* il guadagno *aumentando* R_D: in forma integrata questa soluzione è sconveniente per motivi di area in quanto realizzare resistenze di grande valore utilizzando degli strati diffusi richiede un rapporto L/W molto elevato. Inoltre un valore troppo elevato di R_D può alterare in modo inaccettabile il punto di lavoro diminuendo la dinamica di uscita e portando il transistore fuori dalla regione di saturazione. Nelle realizzazioni integrate è possibile sostituire la resistenza R_D con un generatore di corrente con resistenza interna r_g elevata pervenendo allo schema illustrato in figura 6.16.

La valutazione dei parametri dell'amplificatore può essere fatta ricorrendo al modello di piccolo segnale nel quale il carico attivo viene descritto semplicemente dalla sua resistenza di uscita r_g come indicato nella figura 6.17.

Con l'amplificatore a vuoto il calcolo dell'amplificazione segue la stessa procedu-

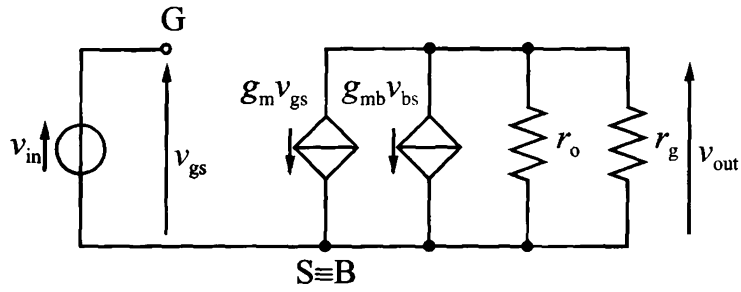


Figura 6.17 Circuito equivalente per piccolo segnale di uno stadio amplificatore CS con carico attivo.

ra utilizzata in precedenza con la sola differenza che la resistenza di drain R_D viene sostituita dalla resistenza r_g del generatore di corrente:

$$|A_{v0}| = g_m (r_o || r_g)$$

$$R_{in} = \infty$$

$$R_{out} = r_o || r_g$$

In tecnologia CMOS dove sono disponibili sia transistori a canale n sia a canale p il generatore di corrente può essere realizzato con un pMOS con il gate connesso ad un potenziale di polarizzazione V_B, in tal modo l'elevato valore della resistenza di uscita del MOSFET permette di aumentare il guagagno dello stadio senza alterarne il punto di lavoro come invece avverrebbe aumentando il valore della resistenza R_D. Lo schema in figura 6.18 presenta quindi uno stadio nel quale il transitore di tipo p viene connesso ad una tensione costante V_B in modo da garantirne la corretta polarizzazione.

Nel modello per piccolo segnale il transistore di tipo p si viene a trovare con una V_{GS} in continua annullando il termine dovuto alla trasconduttanza e rendendo necessaria solo la resistenza di uscita r_{op} come schematizzato in figura 6.19, il guadagno può essere calcolato come:

$$|A_{v0}| = g_m (r_o || r_{op})$$

6.3 Stadio a Drain Comune

Si è verificato come lo stadio a source comune presenti una resistenza di uscita troppo elevata per essere convenientemente utilizzato come amplificatore di tensione. Lo stadio a Drain Comune (CD) il cui schema è illustrato in figura 6.20 rende possibile ottenere un amplificatore di tensione con una resistenza di uscita ridotta rispetto allo stadio a source comune. Anche in questo caso la polarizzazione deve garantire il funzionamento del transistore in saturazione e avviene con un'opportuna scelta della tensione V_{GG}. Tralasciando questo aspetto, assai simile a quanto fatto per lo stadio a source comune, risulta più utile concentrarsi sui parametri dell'amplificatore. Il modello per piccolo

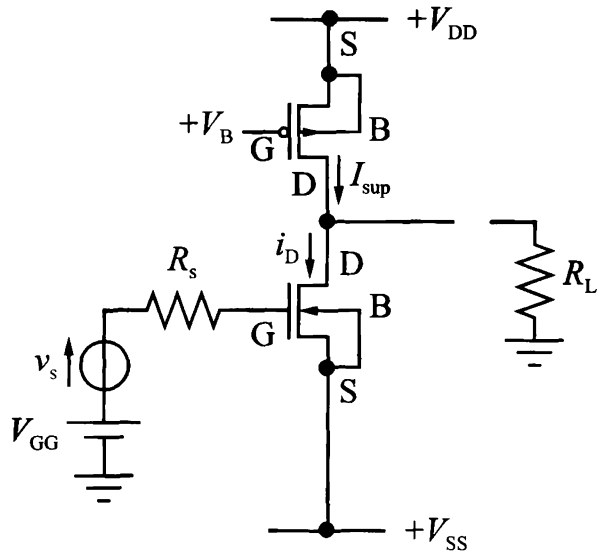


Figura 6.18 Schema di uno stadio amplificatore CS con carico attivo costituito da un transistore pMOS.

segnale presente nella figura 6.21 deve in questo caso tenere in conto della transconduttanza di substrato in quanto il MOSFET si trova con il substrato polarizzato a V_{SS} e la v_{bs} viene a coincidere con $-v_{out}$.

Con l'amplificatore a vuoto, definita $r_1 = r_o || r_g$

$$v_{gs} = v_{in} - v_{out}$$

$$v_{out} = (g_m v_{gs} - g_{mb} v_{out})\, r_1$$

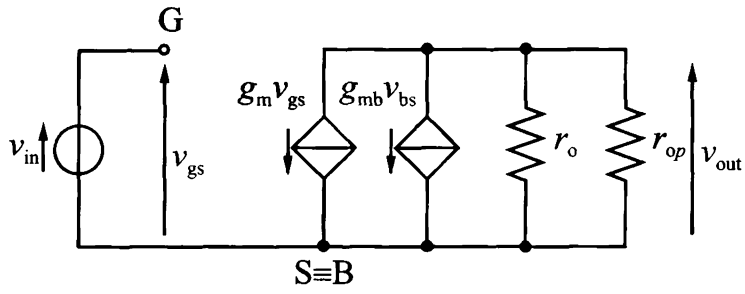


Figura 6.19 Circuito equivalente per piccolo segnale di uno stadio amplificatore CS con carico attivo costituito da un transistore pMOS.

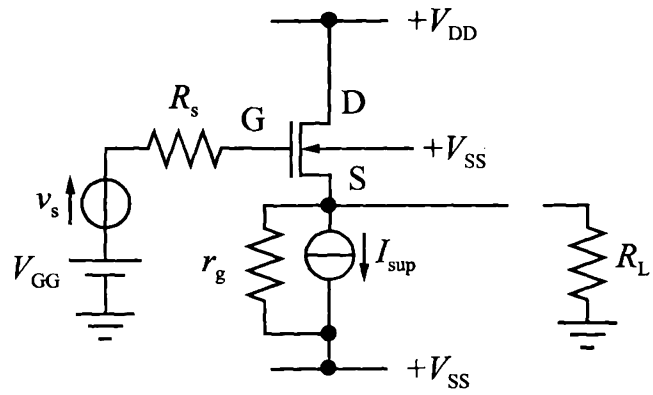


Figura 6.20 Schema di uno stadio a drain comune nel quale la resistenza sul source viene sostituita da un carico attivo.

$$v_{out} = \frac{g_m v_{gs} r_1}{1 + g_{mb} r_1}$$

Sostituendo l'espressione di v_{out} nella v_{gs} si ottiene:

$$v_{gs} = v_{in} \frac{1 + g_{mb} r_1}{1 + (g_m + g_{mb})\, r_1}$$

$$v_{out} = \left(\frac{g_m r_1}{1 + g_{mb} r_1} \right) \left(\frac{1 + g_{mb} r_1}{1 + (g_m + g_{mb}) r_1} \right) v_{in}$$

$$\frac{v_{out}}{v_{in}} = \frac{g_m}{(g_m + g_{mb}) + 1/r_1}$$

Dato che il parallelo tra r_o e r_g è generalmente un valore elevato (soprattutto nel caso di un carico attivo integrato) l'espressione del guagagno può essere semplificata e si constata che essendo $g_m > g_{mb}$ il guadagno di tensione diventa unitario.

$$A_{v0} = \frac{g_m}{g_m + g_{mb} + (r_o || r_g)^{-1}} \approx \frac{g_m}{g_m + g_{mb}}$$

$$R_{in} = \infty$$

$$R_{out} = \frac{1}{g_m + g_{mb} + (r_o || r_g)^{-1}} \approx \frac{1}{g_m + g_{mb}}$$

La resistenza d'ingresso non subisce variazioni, al contrario la resistenza di uscita può raggiungere valori relativamente piccoli se la trasconduttanza del MOSFET risulti

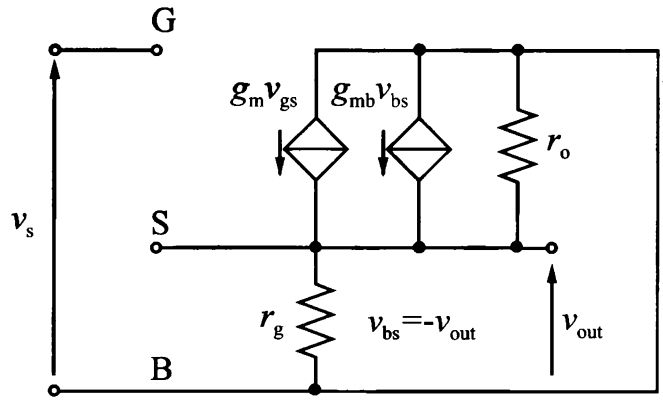


Figura 6.21 Circuito equivalente per piccolo segnale di uno stadio a drain comune con carico attivo.

sufficientemente elevata.

Nel suo complesso, dato che il guadagno di tensione è unitario, lo stadio a drain comune può essere utilizzato come "voltage buffer" ovvero per disaccoppiare il carico da uno stadio amplificatore che presenti una resistenza di uscita troppo grande.

6.4 Stadio a Gate Comune

Questo stadio amplificatore il cui schema è in figura 6.22 rende possibile ottenere un amplificatore di corrente con una resistenza di ingresso molto ridotta e quindi adatto ad amplificare un segnale in corrente. Infatti un segnale in corrente viene utilmente rappresentato dal suo equivalente Norton e la presenza di una conduttanza pari a $1/R_S$ rende inutilizzabile qualsiasi stadio che presenti una resistenza di ingresso elevata. Lo schema dello stadio a Gate Comune (CG) è illustrato in figura 6.22 dove sia la resistenza sul Drain sia quella sul Source sono state sostituite da un carico attivo rappresentato da un generatore di corrente con resistenza di uscita pari a r_g. Osservando tale schema, dato che il Gate del MOSFET viene collegato ad una tensione fissa (ad esempio il riferimento di massa) il segnale "entra" dal source in quanto la corrente del segnale determinerà una variazione della tensione v_{gs} modificando la tensione al source. Questo meccanismo di funzionamento diviene evidente se si osserva il modello per piccolo segnale presente nella figura 6.23. Dato che si vuole caratterizzare questo stadio come amplificatore di corrente si vuole calcolare il rapporto tra la corrente in uscita i_{out} e il segnale in ingresso i_s.

Il calcolo del guadagno in corrente viene fatto con l'uscita in corto circuito dato che si vuole determinare l'equivalente Norton dello stadio. Scrivendo le correnti nel nodo di source:

$$0 = i_s + \frac{v_{gs}}{r_g || R_S} + (g_m + g_{mb}) v_{bs} + \frac{v_{gs}}{r_o}$$

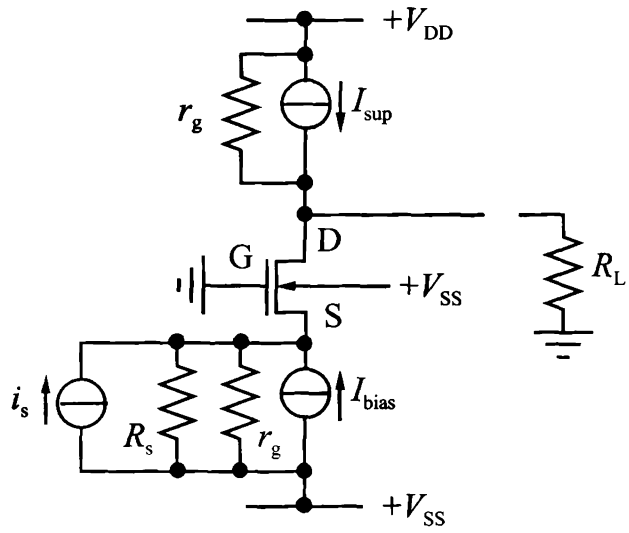


Figura 6.22 Schema di uno stadio a gate comune nel quale sia la resistenza di drain sia quella di source sono sostituite da un carico attivo.

$$v_{gs} = -\frac{i_s}{\frac{1}{r_o} + \frac{1}{r_g||R_S} + g_m + g_{mb}}$$

$$i_{out} = \left[(g_m + g_{mb}) + \frac{1}{r_o}\right] v_{gs}$$

Sostituendo l'espressione di v_{gs} si ottiene il guadagno di corrente che essendo circa unitario permette di utilizzare lo stadio come "current buffer":

$$A_{i0} = \frac{i_{out}}{i_s}$$

$$= -\frac{1 + (g_m + g_{mb})r_o}{1 + (g_m + g_{mb})r_o + \frac{r_o}{r_g||R_S}}$$

$$A_{i0} \approx -1$$

Determinante per poter pilotare questo stadio in corrente è conoscere il valore della resistenza in ingresso. Tale valore può essere ottenuto direttamente come rapporto tra la tensione v_{gs} e la corrente i_s sfruttando lo stesso circuito equivalente utilizzato per il

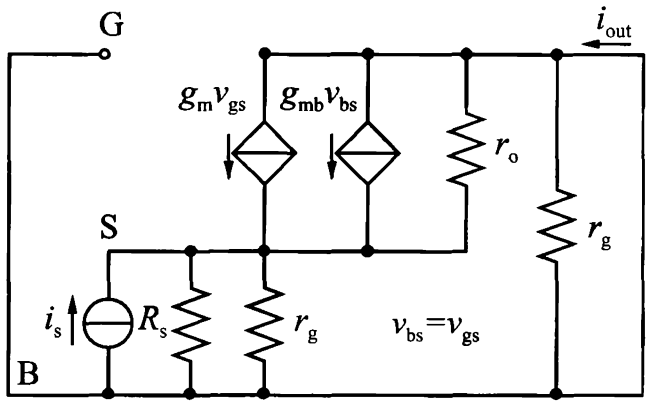


Figura 6.23 Circuito per piccolo segnale di uno stadio a gate comune.

guadagno in corrente.

$$R_{in} = \frac{1}{\frac{1}{r_o} + \frac{1}{R_S||r_g} + (g_m + g_{mb})}$$

$$R_{in} \approx \frac{1}{g_m + g_{mb}}$$

Come si puo' notare che rendendo la transconduttanza sufficientemente elevata si può ridurre il valore della resistenza di ingresso a valori anche di pochi Ω.
Il valore della resistenza di uscita può essere ottenuta applicando un generatore di corrente i_{out} sull'uscita e calcolando v_{out} (ponendo $i_s = 0$)

$$\begin{aligned} v_{gs} &= v_{bs} = -R_S||r_g i_{out} \\ i_{r_o} &= i_{out} - (g_m + g_{mb}) v_{gs} \\ i_{r_o} &= i_{out}\left[1 + (g_m + g_{mb})\, R_S||r_g\right] \\ v_{out} &= r_o i_{r_o} + R_S||r_g i_{out} \\ R_{out} &= r_o\left[1 + (g_m + g_{mb})\, R_S||r_g\right] + R_S||r_g \end{aligned}$$

Esempio 6.1 Si consideri il circuito in figura 6.24 per la polarizzazione dello stadio MOSFET a source comune.
Si calcoli:

- il punto di lavoro e la sua stabilità al variare dei parametri del MOSFET (V_{thn} e β_n);
- la dinamica di uscita;
- il guadagno v_{out}/V_s;
- la resistenza di uscita;

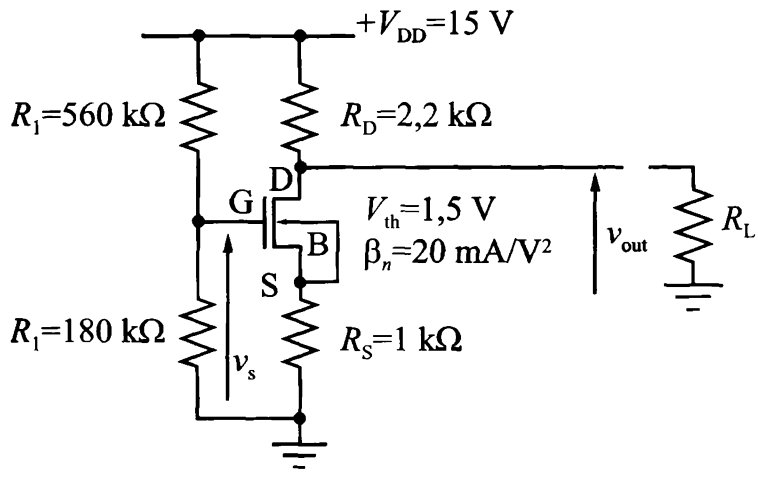


Figura 6.24 Schema per la polarizzazione di uno stadio a source comune.

- ▷ la resistenza di ingresso;
- ▷ l'effetto sui precedenti punti dell'inserzione della capacità $C_S = 33$ nF in parallelo alla resistenza di source;
- ▷ la risposta in frequenza dello stadio.

La struttura utilizzata in questo circuito di polarizzazione viene detta di "autopolarizzazione" in quanto la presenza della resistenza R_S svolge un ruolo di stabilizzazione del punto di lavoro introducendo una reazione negativa dovuta alla riduzione della V_{GS} in presenza di un aumento della corrente I_{DS}.
Calcolando l'equivalente Thevenin del partitore in ingresso e sostituendolo nello schema si ottiene quello presente in figura 6.25, infatti:

$$\begin{aligned} R_{GG} &= R_1 || R_2 \\ V_{GG} &= V_{DD} \frac{R_1}{R_1 + R_2} \end{aligned}$$

Scrivendo l'equazione alla maglia d'ingresso si ottiene:

$$\begin{aligned} V_{GG} &= V_{GS} + R_S I_D \\ V_{GG} &= V_{GS} + R_S \frac{\beta_n}{2} (V_{GS} - V_{th})^2 \\ V_{GG} &= V_{GS} - V_{th} + R_S \frac{\beta_n}{2} (V_{GS} - V_{th})^2 + V_{th} \end{aligned}$$

L'ultima equazione può essere scritta come equazione di secondo grado in $(V_{GS} - V_{th})$:

$$\frac{R_S \beta_n}{2} (V_{GS} - V_{th})^2 + (V_{GS} - V_{th}) - V_{GG} + V_{th} = 0$$

Risolvendo e accettando come valida solo la soluzione positiva si ha:

$$V_{GS} - V_{th} = \frac{-1 + \sqrt{1 + 2 R_S \beta_n (V_{GG} - V_{th})}}{R_S \beta_n}$$

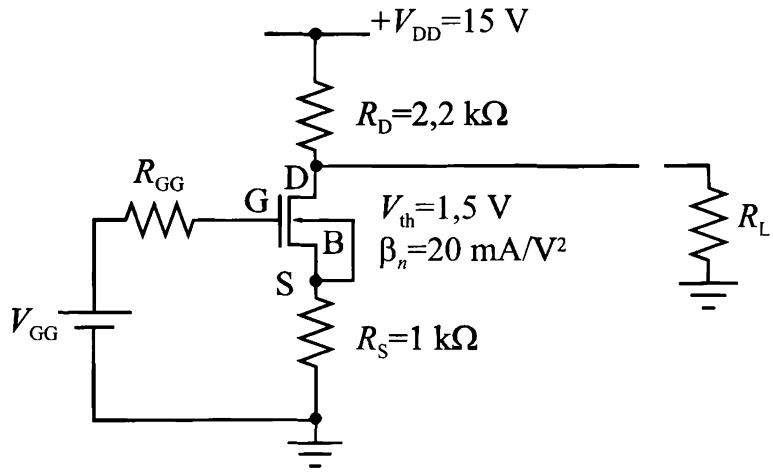


Figura 6.25 Schema con l'equivalente Thevenin in ingresso per la polarizzazione di uno stadio a source comune.

Sostituendo i valori si ottengono:

$$V_{GG} = 3{,}64 \text{ V}$$

$$V_{GS} - V_{th} = 0{,}41 \text{ V}$$

Dal valore ottenuto per $V_{GS} - V_{th}$ e ipotizzando che il transistore sia in saturazione:

$$I_{D0} = \frac{\beta_n}{2}(V_{GS} - V_{th})^2 = 1{,}73 \text{ mA}$$

$$V_{out0} = V_{DD} - R_D I_D = 11{,}78 \text{ V}$$

$$V_{DS0} = 9{,}45 \text{ V}$$

Essendo $V_{DS0} > V_{GS0} - V_{th}$ viene confermata l'ipotesi di trovarci in *saturazione*.
La condizione di saturazione può anche essere verificata imponendo la condizioni di transistore saturo e in conduzione:

$$V_D - V_S > V_{GG} - V_S - V_{th}$$

$$V_{DD} - I_D R_D > V_{GG} - V_{th}$$

$$I_D < \frac{V_{DD} - V_{GG} + V_{th}}{R_D} = 5{,}84 \text{ mA}$$

La condizione di MOSFET in conduzione:

$$V_{GS} > V_{th}$$

$$V_{GG} - I_D R_S > V_{th}$$

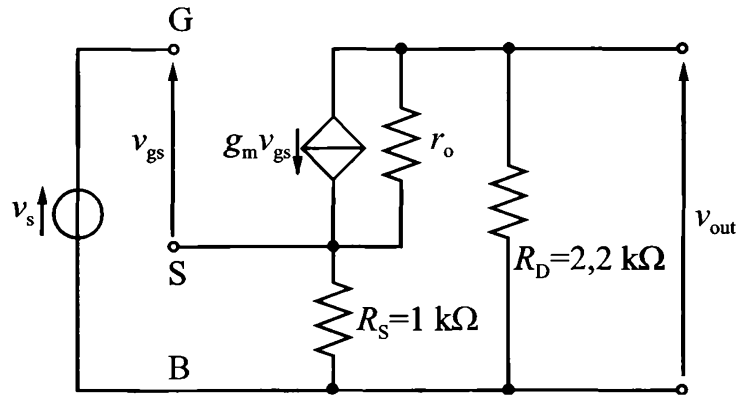


Figura 6.26 Circuito per il piccolo segnale dello stadio a source comune.

$$I_D < \frac{V_{GG} - V_{th}}{R_S} = 2{,}14 \text{ mA}$$

Come già accennato il circuito utilizzato per la polarizzazione risulta particolarmente stabile grazie all'effetto di *controreazione* dovuto a R_S: un aumento della corrente I_D per una riduzione di V_{th} o per un aumento di β_n tende a ridurre la V_{GS}. Questo comportamento può essere valutato quantitativamente tramite la ***sensibilità*** del punto di lavoro alla variazione della V_{th}.

$$S_{V_{th}} = \left.\frac{\partial (V_{GS} - V_{th})}{\partial V_{th}}\right|_{OP}$$

$$= \frac{(1 + 2R_S\beta_n (V_{GS} - V_{th}))^{-1/2}}{2R_S\beta_n}$$

risultato che dimostra che un valore sufficientemente elevato di R_S può ridurre la variazione del punto di lavoro a causa di una variazione (ad esempio con la temperatura) della tensione di soglia. Passando a valutare la dinamica di uscita questa è limitata superiormente dall'*interdizione* del MOSFET e inferiormente dall'uscita dalla *saturazione*

$$V_{out} < V_{DD}$$

$$V_{out} > V_{GG} - V_{th}$$

Il modello equivalente per piccolo segnale dell'amplificatore è presentato nella figura 6.26 nella quale viene indicata anche la resistenza di uscita del MOS r_o che nella successiva analisi verrà trascurata per semplicità.
I parametri di piccolo segnale si riducono alla valutazione della g_m dato che r_o verrà trascurata ed essendo $V_{BS} = 0$ risulta inutile valutare la g_{mb}.

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{OP} = \beta_n (V_{GS0} - V_{th})$$

$$g_m = 8{,}2 \text{ mS}$$

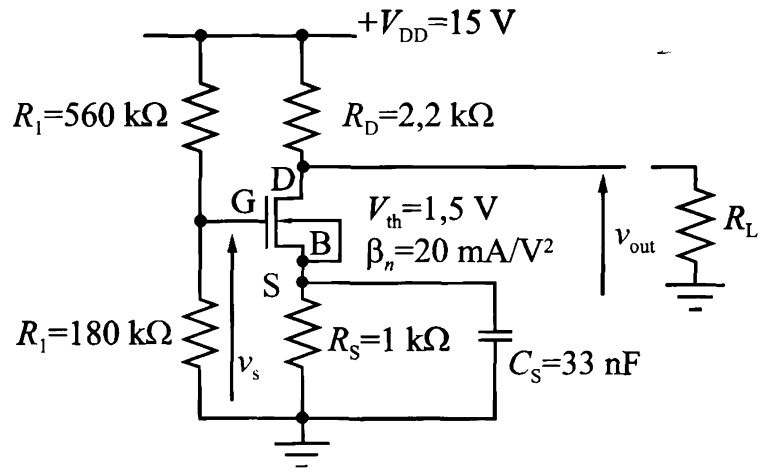


Figura 6.27 Schema di uno stadio a source comune con gruppo parallelo RC sul source.

Scrivendo l'equazione alla maglia d'ingresso si ottiene:

$$v_s = v_{gs} + R_S g_m v_{gs}$$

$$v_{gs} = \frac{v_s}{1 + R_S g_m}$$

$$\frac{v_{out}}{v_s} = -\frac{R_D g_m}{1 + R_S g_m} \approx \frac{R_D}{R_S} = -2{,}2$$

La presenza della resistenza R_S porta insieme all'aumento della stabilità del punto di lavoro un riduzione del guadagno dello stadio. Tale situazione può essre risolta introducendo un condensato C_S da 33 nF in parallelo alla resistenza R_S come indicato in figura 6.27, che senza alterare comportamento in continua a frequenza sufficientemente elevata può aumentare sensibilmente guadagno.
Ovviamente la polarizzazione non è alterata dalla presenza del condensatore C_S ma viene modifica la funzione di trasferimento:

$$\frac{v_{out}}{v_s} = -\frac{R_D g_m}{1 + Z_S g_m}$$

$$Z_S = \frac{R_S}{1 + sR_S C_S}$$

$$\frac{v_{out}}{v_s} = -\frac{R_D g_m (1 + sR_S C_S)}{sR_S C_S + g_m R_S + 1}$$

La funzione di trasferimento presenta uno *zero* e un *polo* e analizzando il comporatemento in continua e a frequenza infinita si può determinare il guadagno alle basse frequenze come $-R_D/R_S = -2.2$ equivalente a 6,84 dB, e alle alte frequenze come $-g_m R_D = -18$ equivale

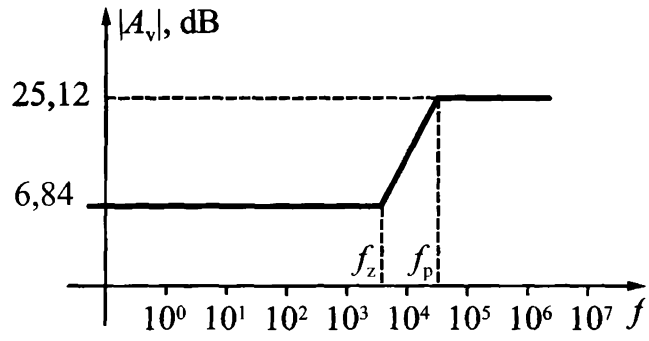


Figura 6.28 Diagramma di Bode del guadagno.

25,12 dB. Lo zero ed il polo si trovano alle frequenze:

$$f_z = \frac{1}{2\pi R_S C_S} = 4{,}82 \text{ kHz}$$

$$f_p = \frac{g_m R_S + 1}{2\pi R_S C_S} = 44{,}37 \text{ kHz}$$

Il corrispondente diagramma di Bode asintotico è presentato in figura 6.28.

Capitolo 7

Tecnologia dei semiconduttori

7.1 Leggi di Moore

La pervasività delle tecnologie elettroniche è il risultato della crescita dei livelli di integrazione nei circuiti integrati. Le leggi di Moore permettono di prevedere con buona precisione come, nel tempo, i miglioramenti tecnologici influenzino i principali parametri di sistema. La più nota fra le leggi di Moore è sicuramente quella relativa al numero di transistori integrabili sul singolo dispositivo che segue una legge esprimibile come:

$$\#Tr(t) = \#Tr(t_0) \cdot 2^{\frac{t - t_0}{1.5}}$$

con t_0 e t espressi in anni. Tale legge afferma quindi che il numero di transitor per circuito integrato raddoppia circa ogni 1.5 anni. La rappresentazione grafica di questa legge trova la sua maggiore evidenza su di un piano semilogaritmico dove sulle ascisse sono posizionati gli anni(su scala lineare) e sulle ordinate il log_{10} del numero di transistori per circuito integrato. Si ricorda che una qualsiasi relazione di tipo esponenziale rappresentata su di un piano semilogaritmico (indipendentemente dalla base) ha un andamento lineare. Infatti

$$\begin{aligned}
\log_{10}(\#Tr(t)) &= \log_{10}\left(\#Tr(t_0) \cdot 2^{\frac{t - t_0}{1.5}}\right) \\
&= \log_{10}(\#Tr(t_0)) + \log_{10}(2) \cdot \log_2\left(2^{\frac{t - t_0}{1.5}}\right) \\
&= \log_{10}(\#Tr(t_0)) + \log_{10}(2) \cdot \frac{t - t_0}{1.5} \\
&= \log_{10}(\#Tr(t_0)) + \frac{\log_{10}(2)}{1.5} \cdot (t - t_0)
\end{aligned}$$

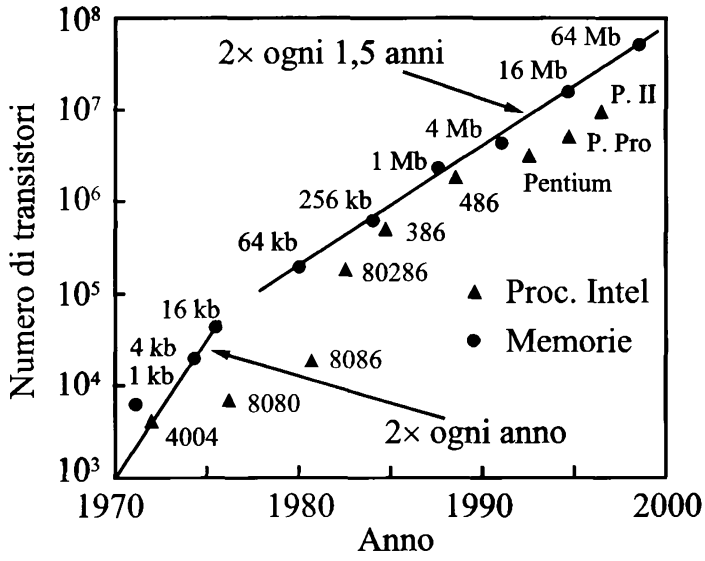


Figura 7.1 Numero di transistori per circuito integrato per le principali famiglie di microprocessori e di memorie.

$$= \log_{10}(\#Tr(t_0)) + 0.2 \cdot (t - t_0)$$

Il grafico in figura 7.1 presenta i dati relativi al numero di transistore per circuito integrato per i principali microprocessori e per i più comuni tagli di memoria dinamic che si sono susseguiti in produzione. Si può osservare piuttosto chiaramente come a partire dagli anni '80 il rateo di crescita del livello di integrazione si sia mantenut costante per entrambi le famiglie di dispositivi, nonostante il livello di integrazione pe le memorie risulti favorito da una maggior regolarià di disegno.

In realtà questa legge può essere approssimativamente confermata a partire da analoghe leggi empiriche correlate però a parametri tecnologici fondamentali. Infat la riduzione delle minime dimensioni integrabili segue una legge del tipo:

$$L_{\min}(t) = L_{\min}(t_0) \cdot 0.7^{\frac{t - t_0}{3}}$$

che può essere equivalentemente scritta come

$$L_{\min}(t) = L_{\min}(t_0) \cdot 2^{-\frac{t - t_0}{6}}$$

ovvero che le dimensioni minime si dimezzano approssimativamente ogni 6 anni. Parallelamente la miglior qualità dei processi di fabbricazione permette di ottenere cir

di maggior dimensioni con una resa (Yield) accettabile. Si definisce

$$\text{Yield} = \frac{\#IC_{\text{funzionanti}}}{\#IC_{\text{fabbricati}}}$$

La legge di Moore che permette di prevedere la massima area di un IC (circuito integrato) è del tipo:

$$\text{Area}_{\text{IC}}(t) = \text{Area}_{\text{IC}}(t_0) \cdot 1.5^{\frac{t-t_0}{3}}$$

È possibile fare la seguente valutazione sull'area occupata dal singolo transistore

$$\text{Area}_{\text{Tr}} \propto L_{\text{min}}^2$$

e quindi

$$\text{Area}_{\text{Tr}}(t) = \text{Area}_{\text{Tr}}(t_0) \cdot 2^{-\frac{2(t-t_0)}{6}}$$

e il numero di transistori integrabili per un IC di massima dimensione può essere valutato come

$$\begin{aligned}
\frac{\text{Area}_{\text{IC}}(t)}{\text{Area}_{\text{Tr}}(t)} &= \frac{\text{Area}_{\text{IC}}(t_0)}{\text{Area}_{\text{Tr}}(t_0)} \cdot \frac{1.5^{\frac{t-t_0}{3}}}{2^{-\frac{(t-t_0)}{3}}} \\
&= \#Tr(t_0) \cdot 1.5^{\frac{t-t_0}{3}} \cdot 2^{\frac{t-t_0}{3}} \\
&= \#Tr(t_0) \cdot 3^{\frac{t-t_0}{3}} \\
&= \#Tr(t_0) \cdot \sqrt{3}^{\frac{t-t_0}{1.5}}
\end{aligned}$$

Confrontabile, nonostante le approssimazioni fatte, con la regola empirica inizialmente studiata.

Passando a considerare alcune leggi di Moore di particolare utilità nella valutazione dei parametri di sistema la previsione della frequenza di clock in un IC digitale può essere ottenuta a partire dalle leggi di variazione dei parametri tecnologici e viene espressa come

$$f_{\text{CK}}(t) = f_{\text{CK}}(t_0) \cdot 1.5^{\frac{t-t_0}{3}}$$

si lascia come esercizio la verifica di tale relazione facendo riferimento ai sistemi MOS.

Concludiamo l'analisi delle leggi di Moore considerando due parametri che sembrano, apparentemente, subire un tred contraddittorio, infatti il costo per singolo transistore fabbricato può essere stimato come

$$\text{costo}_{\text{TR}}(t) = \text{costo}_{\text{TR}}(t_0) \cdot 2^{-\frac{t-t_0}{3}}$$

e il costo di un singolo impianto produttivo come

$$\text{costo}_{\text{IP}}(t) = \text{costo}_{\text{IP}}(t_0) \cdot 2.3^{-\frac{t-t_0}{3}}$$

In realtà queste due tendenza trovano ragione nel concetto di "economia di scala" dove la concentrazione in pochi insediamenti produttivi di grandissima dimensione ha permesso di aumentare enormemente la produttività e di ridurre i costi del singolo dispositivo anche in presenza di processi di fabbricazione sempre più complessi e costosi.

Nella tabella vengono raccolte le leggi di Moore che sono state descritte affiancate dai valori tipici nel 2004 dei parametri tecnologici.

	Legge	2004
Area IC	1.5x ogni 3 anni	6.5 cm^2
Minima dim.	-30% ogni 3 anni	0.13 μm
Trans/IC	2x ogni 1.5 anni	4Gb Dram
CK fr	1.5x ogni 3 anni	1GHz micro
Costo/tran	-50% ogni 3 anni	0.01 cent
Fab costo	2.3x ogni 3 anni	8B

7.2 Il processo planare

La realizzazione nel 1947 grazie al lavoro di Bardeen, Brattain e Schockley del primo transistore rappresenta la nascita delle moderne tecnologie elettroniche. Anche se la struttura originaria risulta assai lontana dalla successiva evoluzione del BJT, ad essa si deve la dimostrazione della possibilità di controllare una corrente in un semiconduttore drogato tramite le tensioni applicate alle due giunzioni di base-emettitore e di base-collettore.

La soluzione tecnologica del primo transistore $p - n - p$ definita "a punta" era caratterizzata da scarsa ripetibilità e scadenti parametri elettrici. Le due giunzioni base-emettitore, base-collettore erano in realtà ottenute mettendo a contato la base di germanio con due sottile lamine d'oro separate tramite l'asportazione del vertice di triangolo di supporto come si può osservare in figura 7.2.

I successivi studi di Schockley sui portatori minoritari permisero la realizzazione del transitore con meccanismi di funzionamento coincidenti con quelli del BJT attuale Le due tecniche inizialmente di tipo non planare utilizzate per la fabbricazione dei transitori erano dette:

- a *giunzione* (junction)

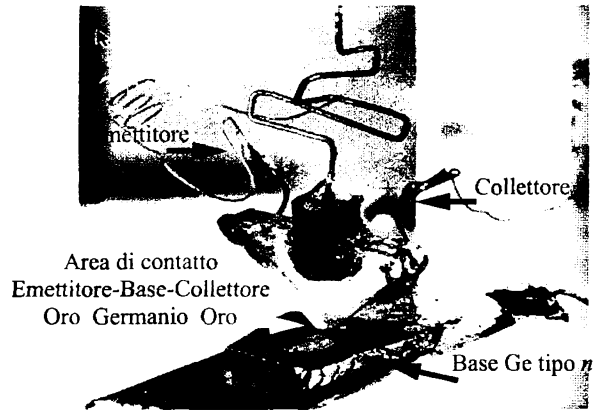


Figura 7.2 Struttura del primo transistor.

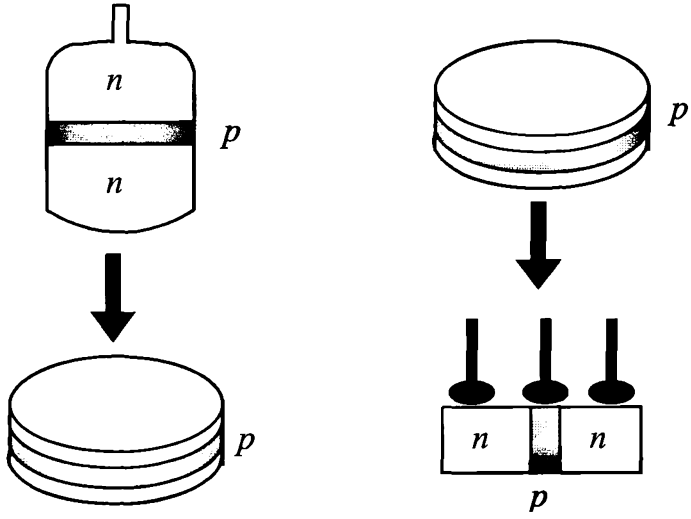


Figura 7.3 Fasi di fabbricazione del transistore a giunzione.

▷ a *lega-giunzione* (alloy-junction)

La prima tecnica richiede la realizzazione di un monocristallo di tipo n con la crescita di una sottile regione drogata di tipo p. La successiva lavorazione del wafer rende le giunzioni $n-p$ e $p-n$ accessibili dall'esterno tramite i terminali di emettitore di base e di collettore. Lo schema di fabbricazione seguito è a grandi linee illustrato nella figura 7.3.

Alternativamente l'utilizzo della diffusione da fase solida da due contatti metallici di indio su di un substrato di tipo n permetteva il controllo dello spessore della regione di base come indicato in figura 7.4

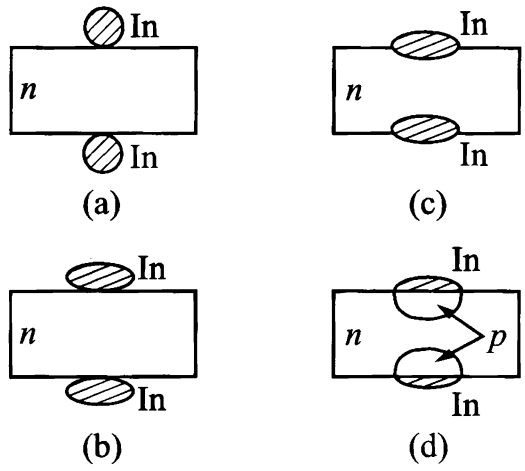
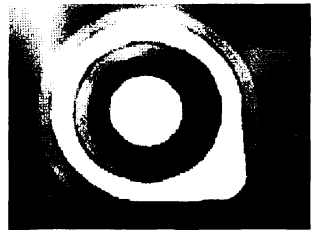


Figura 7.4 Fasi di fabbricazione del transistore a lega.



Fonte: National Semiconductor

Figura 7.5 Struttura del primo transistore bipolare planare.

Le precedenti tecniche di fabbricazione pur permettendo la realizzazione di strutture bipolari sono caratterizzate da uno scarso controllo dei profili di drogaggio e da una scarsa efficienza produttiva rispetto ai successivi processi di tipo planare. Infatti la definizione di un reale processo planare richiese, negli anni '50, la messa a punto di tre procedimenti tecnologici che ancora oggi costituiscono il nucleo dei processi di fabbricazione dei circuiti integrati. Prime fra tutte si resero necessarie le tecniche per la fabbricazione di monocristalli di silicio anzichè di germanio e insieme alle tecniche per l'ossidazione termica del silicio e alle tecniche fotolitografiche per la rimozione selettiva dell'ossido dal silicio.

È proprio con l'utilizzo congiunto di un substrato di silicio e delle tecniche di drogaggio selettivo nel 1958 Jean Hoerni(Fairchild) realizzò il primo BJT planare del quale nella figura 7.5 risulta evidente la struttura concentrica delle regioni di emettitore, base e collettore.

Volendo entrare maggiormente nel dettaglio delle fasi del processo planare, possiamo constatare che la fabbricazione avviene con la ripetizione di una sequenza stan-

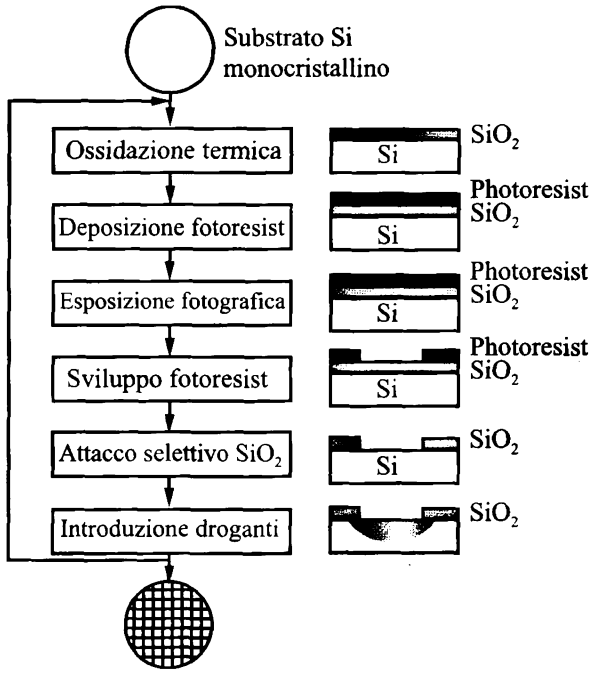


Figura 7.6 Fasi nella realizzazione planare di un transistore bipolare.

come indicato in figura 7.6: la possibilità di introdurre in modo selettivo impurità droganti all'interno di un substrato di silicio si basa sul capacità dell'ossido di ostacolare la diffusione di impurità e di resistere alle alte temperature presenti all'interno di un forno per la diffusione. Grazie a questi progressi fu possibile realizzare un transistore bipolare in modo planare: un doppio processo di mascheratura e di drogaggio permette dapprima la definizione della regione di base e successivamente di quella di emettitore. Come illustrato nella figura 7.7 la sequenza di fabbricazione mette in evidenza come la definizione delle geometrie planari per il transistore bipolare sia del tutto indipendente dal controllo della larghezza di base che rappresenta il parametro più determinante sul funzionamento del BJT. La larghezza di base risulta definita infatti dal controllo dei profili di drogaggio dell'emettitore e della base e non dalle dimensioni delle aperture nell'ossido. In tal modo le larghezze di base ottenibili potevano essere di pochi decimi di micron anche quando la fotolitografia non poteva sperare di risolvere che le decine di micron rappresentando un grande vantaggio delle tecnologie bipolari planari.

La possibilità dei processi planari di operare simultaneamente sull'intero substrato rese estremamente naturale passare dalla fabbricazione di singoli disposiivi alla realizzazione in forma integrata di un intero circuito elettronico. Nel 1959 Robert Noyce alla Fairchild realizzò il primo circuito integrato monolitico con tecnologia planare (figura 7.8).

Le enormi potenzialità del processo planare si resero evidenti a partire dagli anni '60 quando prese inizio la realizzazione di semplici porte logiche integrate SSI (small

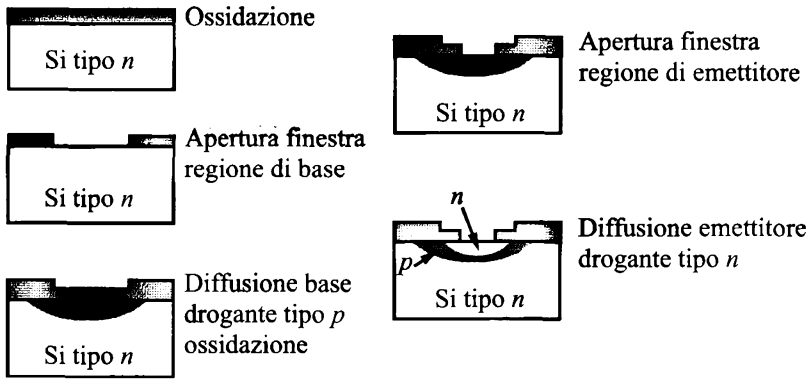
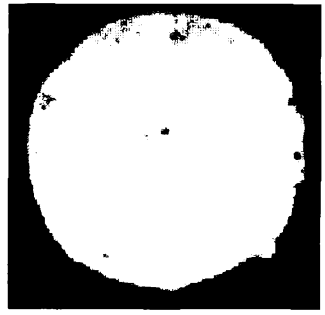


Figura 7.7 Doppia diffusione per le regioni di base e di emettitore di un transistore bipolare.



Fonte: National Semiconductor

Figura 7.8 Fasi di fabbricazione del transistore a lega.

scale integration), come quella in figura 7.9

I continui progressi della tecnologia planare unitamente alle tecnologie MOS a partire dagli anni '70 resero possibile, in pochi anni, passare da semplici porte logiche ai microprocessori integrati quali l'Intel 4004 (figura 7.10),8086 (figura 7.11), Pentium (figura 7.12), dove il numero dei transistori integrati passò rapidamente da alcune migliaia ad alcuni milioni.

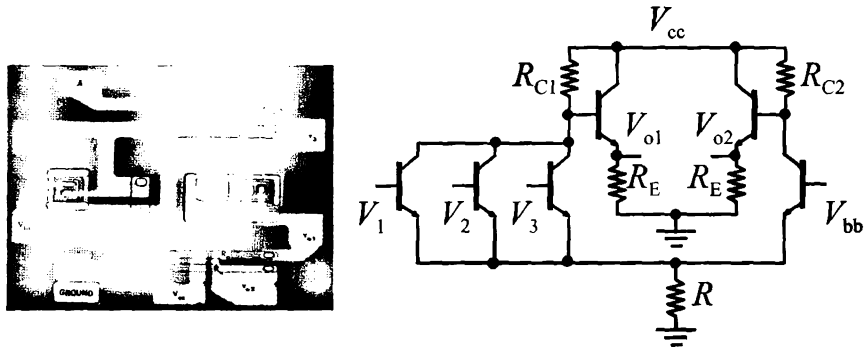
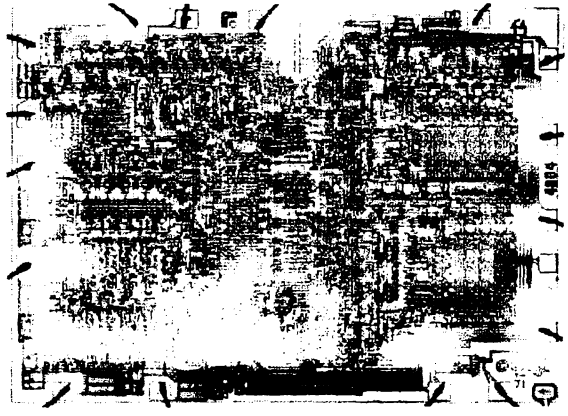
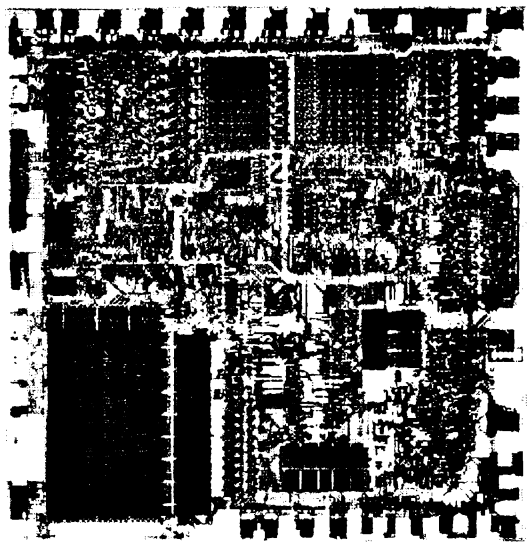


Figura 7.9 Microfotografia e corrispondente schema di una semplice porta logica ECL.



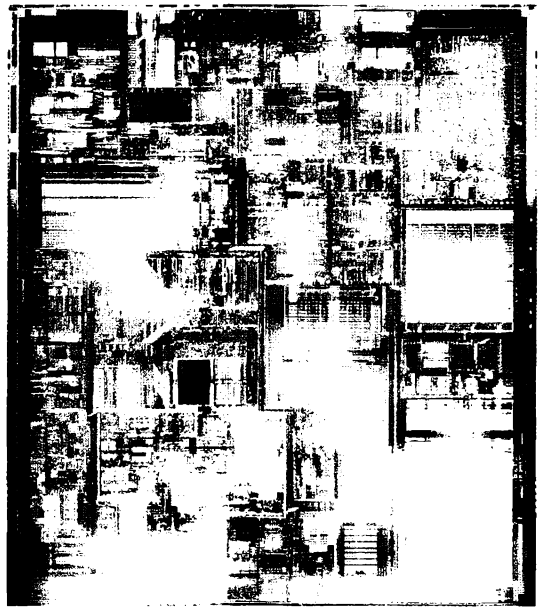
Fonte: Intel Museum - Online Exhibit

Figura 7.10 Microfotografia dell'Intel 4004: 1971, 2300 transistor(10 micron), 108 KHz bus a 4 bit.



Fonte: Intel Museum - Online Exhibit

Figura 7.11 Microfotografia dell'Intel 8086: 1978, 29000 transistor(3 micron), 5 MHz bus a 16 bit.



Fonte: Intel Museum - Online Exhibit

Figura 7.12 Microfotografia dell'Intel Pentium: 2000, 7.5 Ml transistor(0.35 micron), 300 MHz bus a 64 bit.

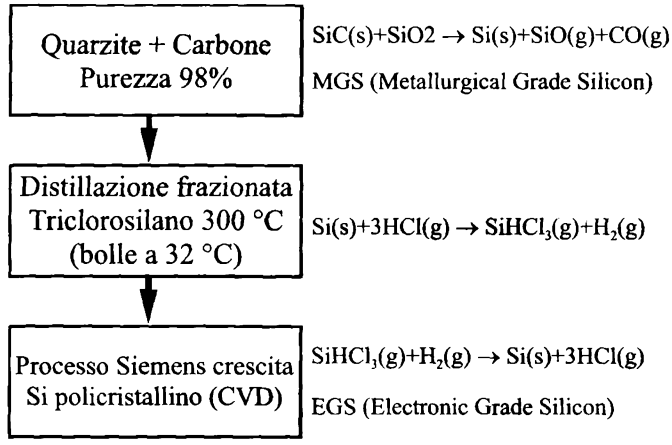


Figura 7.13 Principali fasi della purificazione del silicio per la realizzazione dei substrati

7.3 Crescita dei monocristalli e preparazione dei substrati

Strategico nella tecnologia planare è disporre di substrati di silicio monicristallino di grandi dimensioni. La preparazione di monocristalli di grandi dimensioni richiede silicio di elevatissima purezza: si definisce *Electronic Grade Silicon* (EGS) quando la densità relativa di impurità sia inferiore a 10^{-9}. La preparazione dell' EGS avviene in tre fasi volte ad aumentare progressivamente la purezza del silicio. Inizialmente la quarzite (SiO_2) viene ridotta in fornace con carbone portando a silicio con una purezza del 98% (Metallurgical Grade Silicon (MGS)). Il silicio così ottenuto viene fatto reagire con HCl in modo da ottenere triclorosilanno ($SiHCl_3$) che è liquido a temperatura ambiente e ha punto di ebollizione a 32 °C, si procede quindi alla distillazione frazionata del triclorosilano. L'ultima fase è quella della crescita del silicio policristallino a partire dal triclorosilano con una crescita di tipo CVD (Chemical Vapor Deposition) facendolo reagire con H_2. Lo schema complessivo della preparazione dell'EGS è illustrato in figura 7.13

Partendo da polisilicio EGS la tecnica di crescita dei monocristalli più utilizzata è detta *CZ* (Czochralski).

All'interno di un crogiolo di quarzo (figura 7.14), in modo da garantire una bassa contaminazione del fuso (al più verranno rilasciate molecole di O_2), il silicio policristallino (EGS) drogato con B o P viene portato alla temperatura di fusione (1420 °C). Sulla testa di un mandrino un seme cristallino con una opportuna orientazione reticolare viene messo a contatto del silicio fuso: il gradiente di temperatura esistente tra il fuso e il seme determina la solidificazione progressiva degli strati del liquido che vengono a contato del seme, strati che solidificando ripetono la struttura cristallina della gemma. Una lenta trazione/rotazione che può durare alcune ore, permette l'accrescimento di un lingotto di grandi dimensioni fino a 300 mm di diametro. Il raffreddamento deve avvenire in modo molto lento e continuo e la trazione deve essere effettuata in modo tale che le tensioni meccaniche restino limitate, diversamente la densità dei difetti nel monocristallo può crescere: difetti di linea e di volume potrebbero infatti compromettere l'utilizzo del substrato in quanto i successivi processi di fabbricazione, quali il

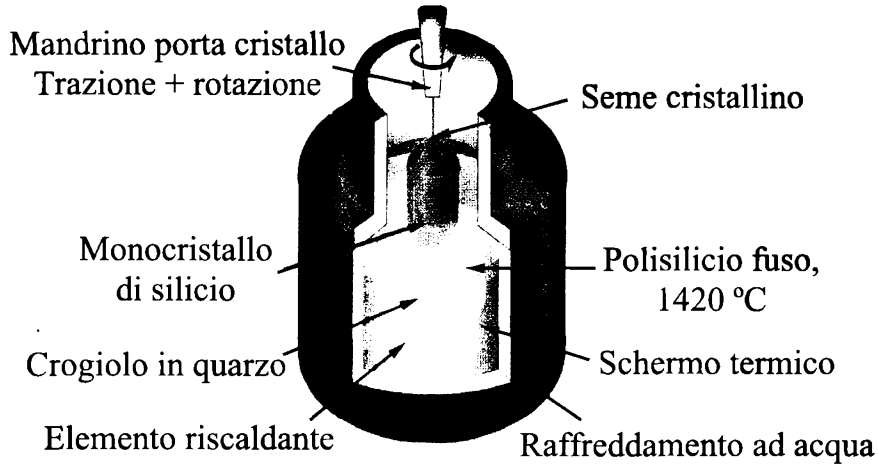


Figura 7.14 Struttura del crogiolo utilizzato per la crescita dei monocristalli con il metodo CZ.

drogaggio, risulterebbero pesantemente alterati dall'irregolarità del cristallo.

Nella figura 7.15 vengono illustrate le 4 fasi della tiratura di un monocristallo di silicio con il metodo CZ: 1) il silicio policristallino EGS viene posto nel crogiolo, 2) il seme cristallino viene posto a contatto del menisco del fuso, 3) con una lenta rotazione e trazione si accresce il collo del monocristallo, 4) raggiunto il diametro voluto inizia la tiratura.

Al termine del processo di tiratura e di raffreddamento il monocristallo è ultimato come appare in figura 7.16.

Dato che nonostante l'estrema cura durante la crescita il monocristallo non presenta una cilindricità perfetta, è necessario una fase di tornitura per rendere perfettamente cilindrico il lingotto (figura 7.17).

Dopo la fase di tornitura possono essere realizzati i "Flat" rimuovendo delle sezioni di calotte cilindriche del monocristallo. Questi flat, generalmente detti "primary" e "secondary", permetteranno durante le successive fasi di lavorazione dei substrati di identificarne l'orientazione reticolare e il tipo di drogaggio presente come indicato in figura 7.18. Attualmente per lingotti di grandi dimensioni, per limitare lo spreco di superficie utile, i flat vengono sostituiti con una codifica "bar code" sulle singole fette contenente molte informazioni sulle caratteristiche del wafer.

Per arrivare ad ottenere dei substrati di opportuno spessore il monocristallo deve essere "affettato" con l'utilizzo di una sega a filo diamantato (figura 7.19) la quale permette il taglio simultaneo di centinaia di fette (wafer) di uno spessore generalmente di circa 800-1000 μm.

Il procedimento seppur eseguito con cura determina una superficie dei wafer poco planare e inadatta alle lavorazioni successive, per questo motivo sono necessarie due ulteriori fasi di planarizzazione. La prima avviene sottoponendo le fette ad un attacco chimico (etching) in una soluzione acida (figura 7.20).

La planirazzione vera e propria avviene utilizzando un processo di tipo CMP (Che-

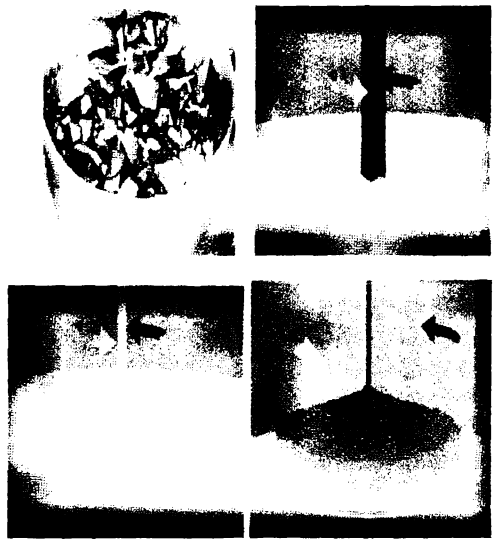


Figura 7.15 Fasi di crescita del monocristallo con il metodo CZ.

mical Mechanical Polishing) dove le fette vengono lappate utilizzando paste abrasive con granularià via via più fine (figura 7.21).

In questo modo si ottiene la riduzione della rugosità superficiale entro pochi nm rendendo la superficie perfettamente speculare. Molto frequentemente per ottenere sulla superficie uno strato di spessore controllato (generalmente alcuni micron) con livello di drogaggio molto accurato si procede ad una crescita di un ulteriore strato monocristallino tramite epitassia. A questo punto dopo una caratterizzazione dei parametri elettrici e fisici dei wafer, si può considerare conclusa l'intera fase di preparazione dei substrati. Negli ultimi decenni il miglioramento nella produzione di monocristalli di grandi dimensioni ha permesso l'utilizzo di wafer di 300 mm di diametro con enormi vantaggi dal punto di vista della produttività.

La realizzazione di wafer di grande diametro permette di ridurre i costi di processo per ogni singolo dispositivo, infatti facendo riferimento ad un microprocessore di 15 mm di lato realizzato utilizzando un wafer di 200 mm di diametro come illustrato in figura 7.22 se ne ottengono 88 per wafer, ma questo numero passa a 232 se si utilizza un wafer da 300 mm. Si ricorda che nei processi planari i tempi e i costi delle varie fasi di fabbricazione non dipendono, se non in modo marginale, dalle dimensioni dei wafer, in tal modo utilizzando wafer di grandi dimensioni si ottiene una riduzione dei costi per singolo dispositivo fabbricato.

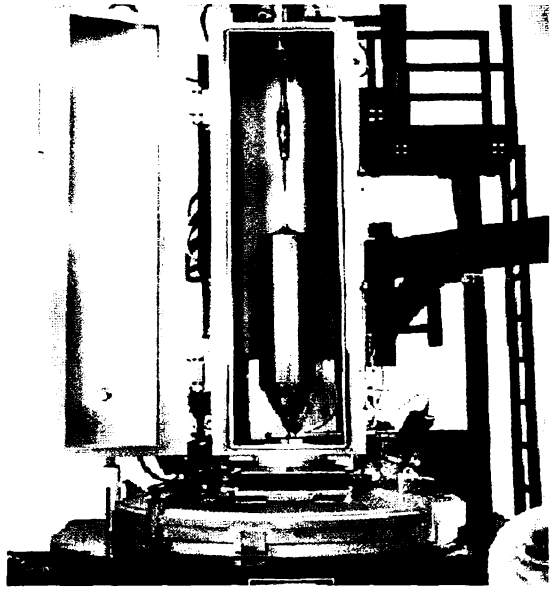


Figura 7.16 Il lingotto carota al termine della fase di tiratura.

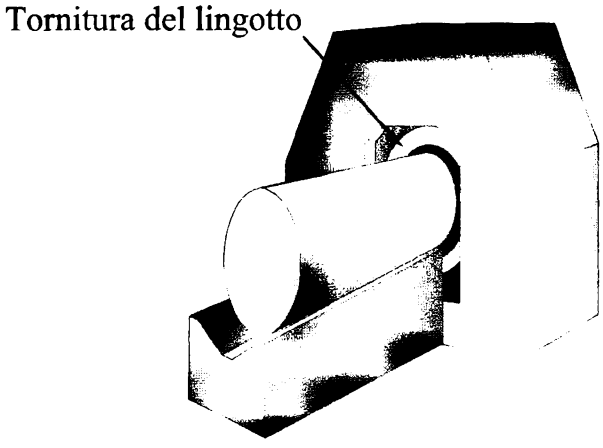


Figura 7.17 Il lingotto dopo la fase di tornitura.

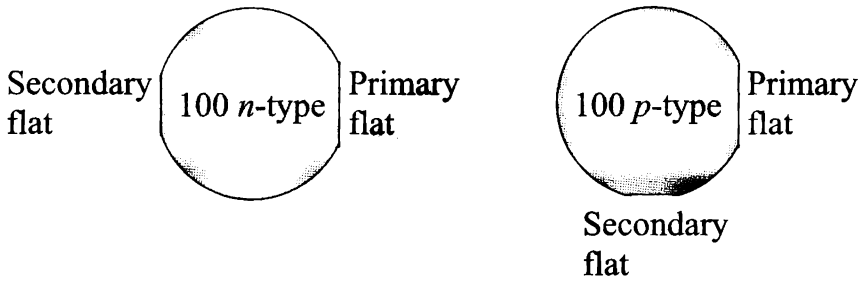


Figura 7.18 Esempi di flat.

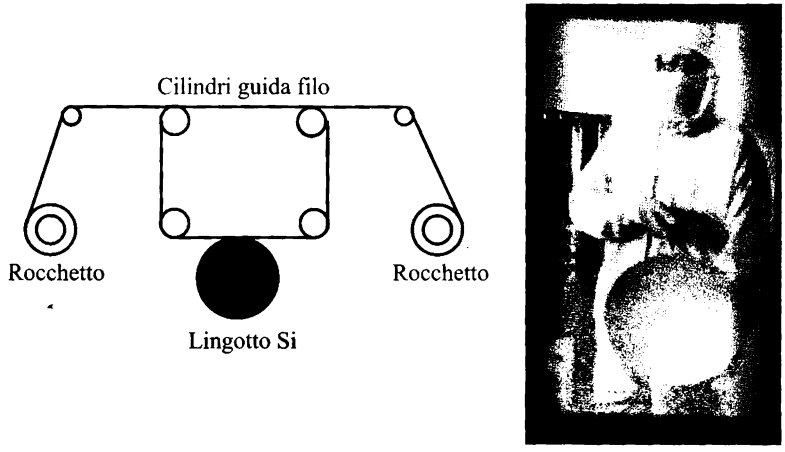


Figura 7.19 Sega a filo per il taglio dei wafer.

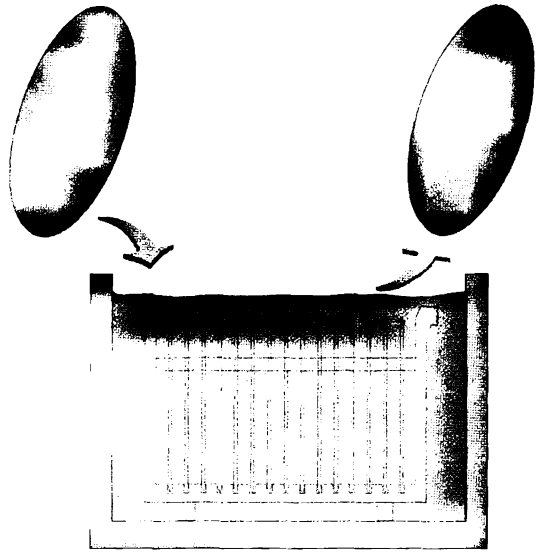


Figura 7.20 Etching superficiale dei wafer.

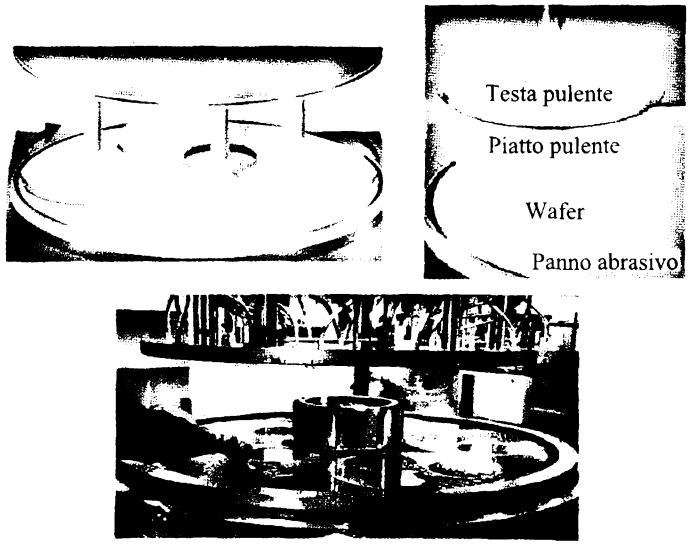


Figura 7.21 Fase di panarizzazione dei wafer tramite CMP.

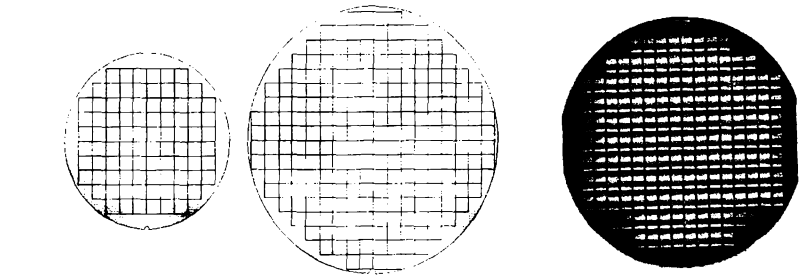


Figura 7.22 Realizzazione di microprocessori su dei wafer da 200 e 300 mm.

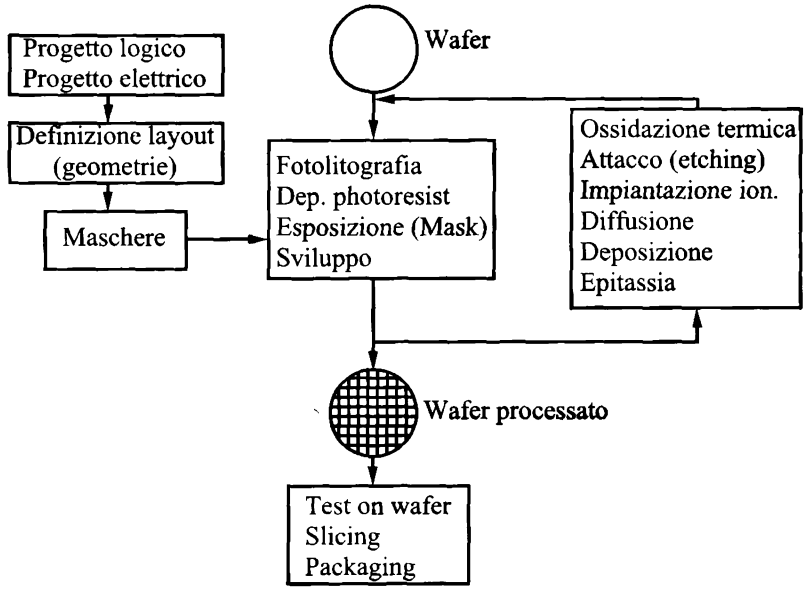


Figura 7.23 Ciclo di fabbricazione di un circuito integrato.

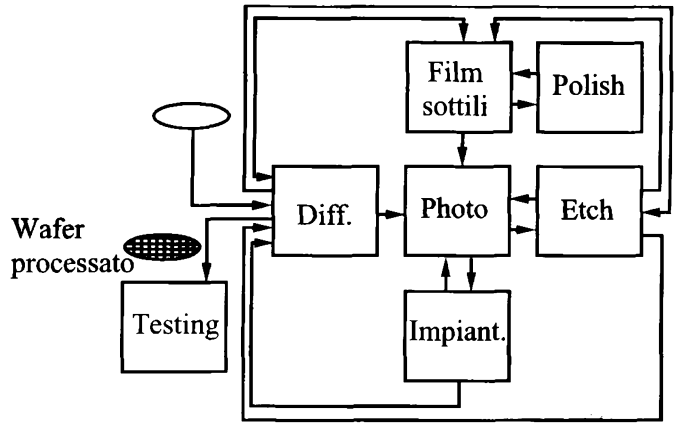


Figura 7.24 Organizzazione di una linea di fabbricazione di circuiti integrati.

7.4 Tecniche fotolitografiche

Come si era già delineato nell'evoluzione della tecnologia dei semiconduttori fu strategica insieme alla capacità di fabbricare substrati di grandi dimensioni, la messa a punto di tecniche fotolitografiche al fine di trasferire le geometrie di progetto sulla superficie del wafer. Ancora oggi la fotolitografia è centrale in qualsiasi processo planare. Seguendo il ciclo di fabbricazione di un circuito integrato IC, presentato in figura 7.23, il wafer subisce una sequenza di procedimenti sull'intera superficie (ossidazioni, drogaggi...) i quali solo grazie alle tecniche fotolitografiche, vengono selettivamente definiti entro delle opportune aree.

In un certo senso l'intero progetto di un sistema elettronico si traduce in un insieme di geometrie (layout) che definiscono la struttura dei dispositivi che si verranno a realizzare. Tali geometrie vengono quindi trasferite sulle maschere fotolitografiche e da qui riportare sulla superficie del wafer rendendo selettivi i procedimenti tecnologici sulle aree d'interesse. Ogni volta che nel ciclo di fabbricazione si rende necessaria la fotolitografia devono essere eseguiti tre passi principali: deposizione del film fotosensibile (photoresist) sulla superficie del wafer, esposizione del photoresist tramite le zone chiare della maschera e sviluppo del photoresist con la rimozione delle zone non selezionate durante l'esposizione. Il numero di passi fotolitografici può, in processi complessi, arrivare ad alcune decine, terminati i quali il wafer deve essere caratterizzato (testing on wafer), suddiviso e infine incapsulato nell'opportuno package. Il ciclo appena descritto determina la struttura organizzativa di una linea di fabbricazione, in figura 7.24 ne è illustrata la struttura a sezioni di lavorazione.

È necessario premettere che la fabbricazione di dispositivi integrati richiede che i differenti processi avvengano ciascuno in ambienti puliti, ovvero ambienti privi di contaminanti e in generale privi di polveri che potrebbero determinare dei difetti di fabbricazione. Si definisce come "classe", classe 10000, 1000, 100, 10, un ambiente dove vengano tollerate al più 10000, 1000, 100, 10 particelle di polveri con diametro inferiore al micron per piede cubo. Per raggiungere valori di classe molto bassi è necessario progettare l'intera linea di fabbricazione in modo da racchiudere le sezioni a maggior

pulizia entro ambienti di classe più elevata. La fotolitografia richiede, per evitare di introdurre difetti "fatali" sulle geometrie il massimo livello di pulizia, per questo motivo e per l'alto numero di procedimenti fotolitografici richiesti, la sezione fotolitografica risulta centrale all'interno di una linea di fabbricazione. Nel suo complesso una linea risulta organizzata sulle seguenti sezioni:

1. diffusione (diffusione termica di droganti, ossidazione termica);
2. impiantazione (introduzione delle impurità con impiantazione ionica);
3. film sottili (crescite epitassiali e non epitassiali);
4. polishing (planarizzazione);
5. etch (attacchi chimici per la rimozione selettiva degli strati);
6. *fotolitografia*

Entrando in maggior dettaglio nelle fasi del processo fotolitografico, illustrato in figura 7.25, si possono identificare

1. Preparazione del wafer utilizzando come primer (aggrappante) l'Hexamethyldisilazane HMDS il quale viene deposto sull'intero wafer;
2. Deposizione del Photoresist tramite rotazione del substrato (spin);
3. Essicazione soft del photoresist in modo da eliminare il solvente;
4. Allineamento della maschera al wafer ed esposizione a luce U.V.;
5. Ulteriore essicazione del resist;
6. Sviluppo del photoresist con un opportuno solvente (acetone..)
7. Essicazione hard in modo da consolidare il resist che non è stato rimosso dallo sviluppo;
8. Ispezione del photoresist in modo da garantirne il corretto funzionamento.

La definizione fotoligrafica e la realizzazione delle strutture sul wafer coinvolge quindi tre aspetti principali:

1. l'utilizzo di un film fotosensibile (photoresist positivo o negativo);
2. il passaggio delle geometrie dal "layout" alle maschere o ai reticoli (fabbricazione delle maschere);
3. l'esposizione dell'intero wafer per trasferire le geometrie dalla maschera al resist utilizzando una lunghezza d'onda in grado di "risolvere" la geometria di dimensione minima.

La definizione di photoresist positivo o negativo è legata al differente effetto che la radiazione ha sul film polimerico: nel caso di resist positivo i fotoni spezzando delle catene polimeriche riducono il peso molecolare del resit esposto rispetto a quello rimasto in ombra e ne facilitano la rimozione con il solvente di sviluppo, nel resist negativo al contrario i fotoni facilitano la coesione fra le differenti molecole rendendole meno solubili in fase di sviluppo. Dal punto di vista applicativo il photoresist si dice positivo in quanto dove la maschera è chiara la superficie del wafer risulta scoperta dal resit (figura 7.26), al contrario nel caso di resist negativo la superficie del wafer non viene protetta dal resist dove la maschera è scura (figura 7.27).

Nei processi fotolitografici risulta fondamentale disporre delle maschere a meno che si voglia ricorrerre a tecnologie a scrittura diretta. Una maschera fotolitografica riproduce su di un vetro (quarzo) le geometrie (1:1) necessarie a impressionare l'intero

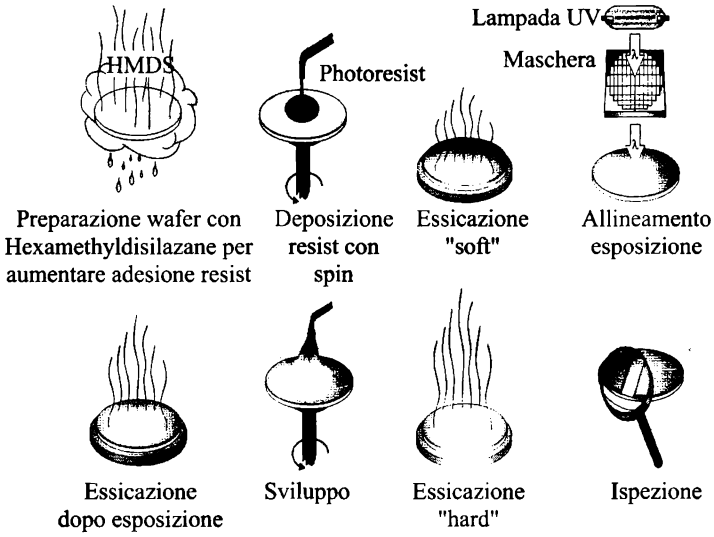


Figura 7.25 Fasi della procedura fotolitografica.

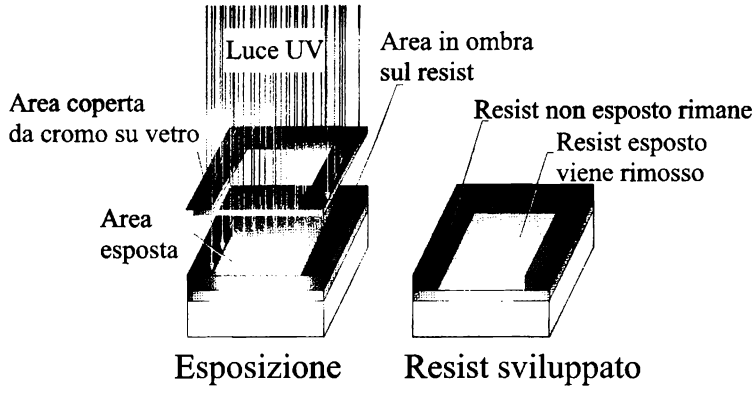


Figura 7.26 Funzionamento photoresit positivo.

wafer. Le geometrie vengono definite utilizzando un sottile strato metallico (generalmente cromo) nelle zone dove deve essere oscurato il fotoresist durante l'esposizione. La fabbricazione delle maschere si basa su di un procedimento nuovamente di natura fotolitografica, dove però questa volta le geometrie vengono scritte direttamente sul fotoresist utilizzando o un pennello laser o un pennello elettronico posizionato seguendo le geometrie descritte nel layout del dispositivo da realizzare. In particolare

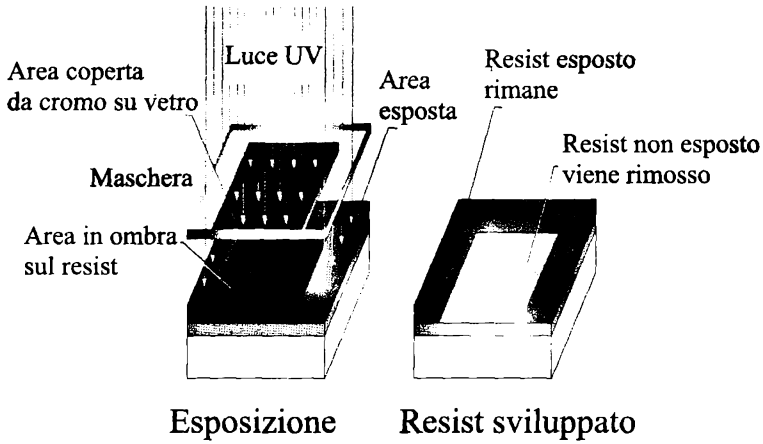


Figura 7.27 Funzionamento photoresit negativo

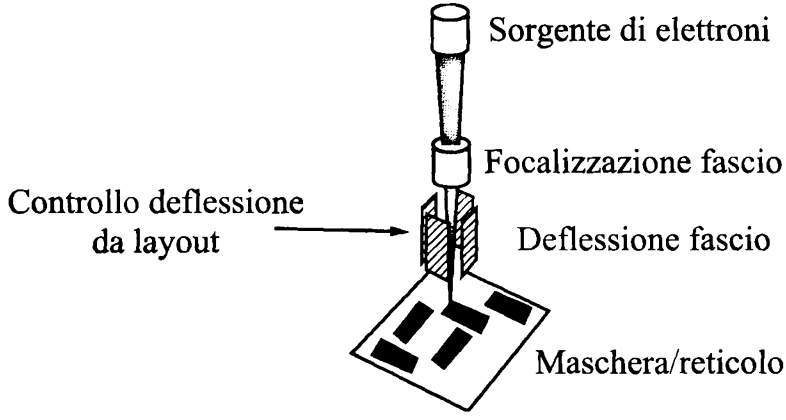


Figura 7.28 Fabbricazione delle maschere tramite scrittura diretta con fascio elettronico.

viene deposto sulla maschera precedentemente ricoperta di cromo uno strato di resist sensibile al fascio elettronico, che viene scritto utilizzando una tecnologia Electron Beam (*EBT*) illustrata in figura 7.28. Successivamente allo sviluppo del resist, tramite attacco chimico, viene rimosso il cromo nelle regioni non protette dal resist.

Dato che una maschera dovrebbe coprire l'intero wafer nel caso di substrati di grandi dimensioni l'utilizzo di maschera può risultare estremamemnte complicato e costoso, a questo scopo si preferisce l'uso dei reticoli visibili in figura 7.29.

Un reticolo è una maschera che presenta una dimensione inferiore rispetto al wafer e le geometrie sono spesso trasferite con rapporto 5:1 o 10:1 rispetto alle dimensioni finali. Oggi prevale l'utilizzo di reticoli rispetto alle tradizionali maschere in quanto

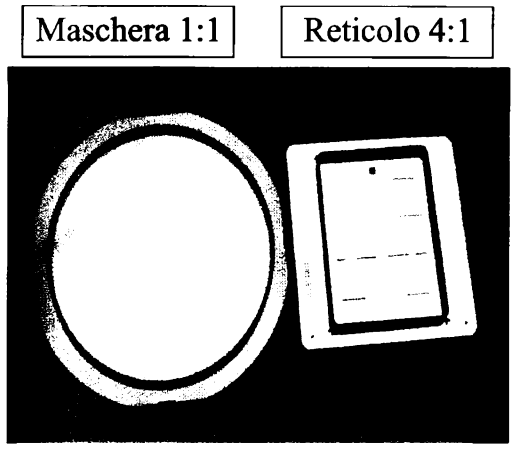


Figura 7.29 Maschera e retcolo a confronto.

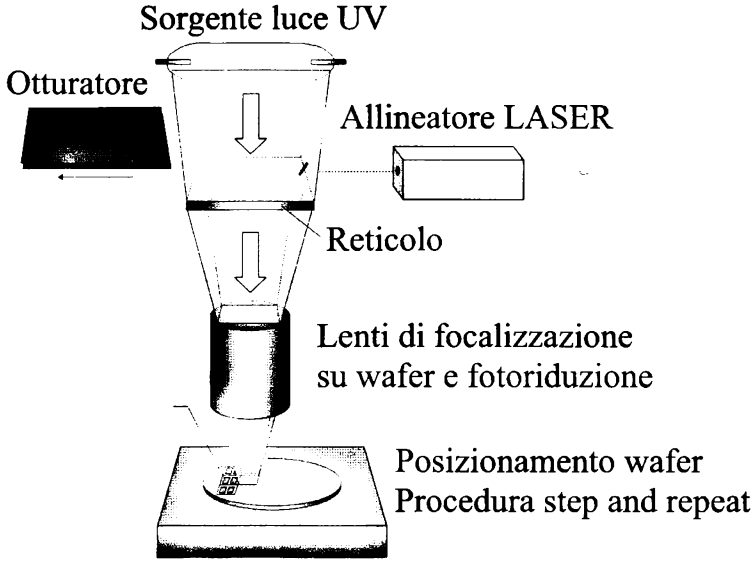


Figura 7.30 Esposizione di un intero wafer tramite ripetizione del reticolo.

sono di più semplice produzione e permetteno (grazie ad un sistema di fotoriduzione) la realizzazione di strutture di dimensioni inferiori. L'esposizione avviene con un sistema di messa a fuoco e di ripetizione dell'immagine del reticolo proiettata sulla superficie del wafer (figura 7.30). I reticoli, essendo utilizzati ripetitivamente per esporre tutto il wafer, replicano più volte gli eventuali difetti e allungano i tempi di esposizione.

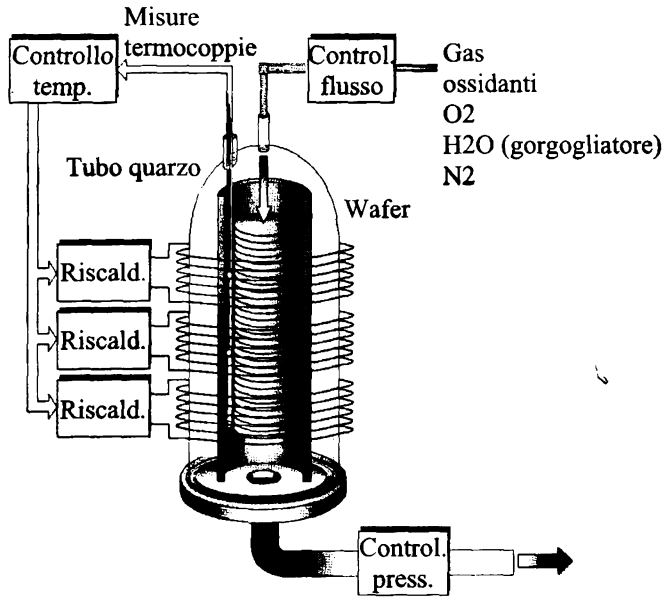


Figura 7.31 Struttura di un forno per l'ossidazione dei wafer.

7.5 Ossidazione Termica

La realizzazione di ossidi nativi richiede che la superficie del silicio sia raggiungibile dalle molecole ossidanti, in questo senso risulta possibile ossidare un wafer solo quando questo non sia già stato ricoperto con altri materiali quali strati di nitruro di silicio e di metalli. Perché lo strato di ossido possa avere un rateo di crescita compatibile con i tempi e i costi di processo è necessario effettuare l'ossidazione a temperature che vanno dai 1000 ai 1300 °C. Per rendere possibile ciò i wafer devono essere inseriti in un forno di ossidazione realizzato con un tubo in quarzo (figura 7.31) e successivamente riscaldati in modo da raggiungere la temperatura prevista e quindi messi a contatto con un'atmosfera ossidante costituita o da molecole di O_2 o da molecole di H_2O.

Dato che la superficie del silicio, tranne che nell'istante iniziale, si trova ad essere ricoperta da uno strato di ossido, la crescita di nuovo ossido richiede la diffusione delle molecole ossidanti all'interno dell'ossido già cresciuto. È possibile quindi definire un flusso di diffusione F_1 attraverso lo strato di ossido già presente alla superficie , e un secondo flusso F_2 assorbito dal silicio che si sta ossidando. L'equazione di ossidazione può essere quindi ottenuta equagliando il flusso nell'ossido delle molecole ossidanti con quello consumato all'interfaccia con il silicio

$$F_1 = -D\frac{\mathrm{d}C}{\mathrm{d}x} \approx D\frac{C_0 - C_{os}}{x}$$

$$F_2 = KC_{os} = F_1 = F$$

$$F = \frac{DC_0}{D/K + x} = C_{ox}\frac{dx}{dt}$$

Integrando l'equazione differenziale dopo aver definito $x(0) = d_0$

$$C_{ox}\int_{d_0}^{x}(D/K + x)\ dx = \int_0^t DC_0\ dt$$

$$x^2 + \frac{2D}{K}x = \frac{2DC_0}{C_{ox}}\left(t + \frac{C_{ox}}{2DC_0}d_0^2 + \frac{C_{ox}}{kC_0}d_0\right)$$

$$A = \frac{2D}{K},\quad B = \frac{2DC_0}{C_{ox}},\quad \tau = \frac{C_{ox}}{2DC_0}d_0^2 + \frac{C_{ox}}{KC_0}d_0$$

Sostituendo le costanti nell'equazione si ottiene

$$x^2 + Ax = B(t + \tau)$$

L'equazione può essere risolta sostituendo i valori delle costanti A e B che sono tabulati in funzione dei parametri chimico-fisici ai quali avviene l'ossidazione. Il tempo τ può essere interpretato come il tempo equivalente necessario a crescere nelle attuali condizioni di ossidazione lo strato di ossido iniziale (d_0), se presente:

$$\tau = \frac{d_0^2}{B} + \frac{d_0}{B/A}$$

Sono quindi possibili due approssimazioni dell'equazione dell'ossidazione a seconda se prevalga il termine lineare o il termine quadratico: il primo caso è reso possibile quando il processo di crescita è dominato dalla cinetica di superficie (ossidi sottili), il secondo quando la crescita di nuovo ossido è invece limitata dalla diffusività delle speci ossidanti nell'ossido già presente (ossidi spessi).

$$x \approx B/A\,(t + \tau) \quad - \text{ sottili-lineare}$$

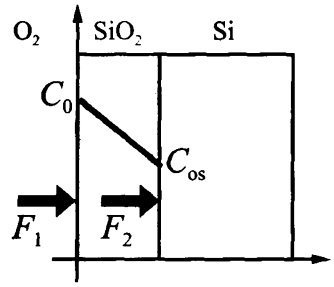
$$x \approx \sqrt{B(t + \tau)} \quad - \text{ spessi-parabolica}$$

Quindi dal punto di vista pratico per ossidi spessi si utilizza l'espressione approssimata di tipo $\sqrt{t}$ dove B viene detto *coefficiente parabolico* ed è ottenibile dal grafico in figura 7.33.

Quando inveve sia necessario studiare fenomeni di ossidazione relativamente rapidi il rateo di crescita dell'ossido è lineare e B/A viene detto *coefficiente lineare* ed è ottenibile dalla figura 7.34.

In qualsiasi processo planare l'ossidazione termica viene utilizzata più volte rendendo necessario studiare gli impatti dell'ossidazione termica sulle strutture coinvolte.

1. Durante la crescita di nuovo ossido viene progressivamente consumato il silicio all'interfaccia ossido silicio. Il *consumo* di silicio può essere stimato come lo 0,44 dello spessore finale del film di ossido cresciuto.



C_0: conc. atmosferica ossidante
C_{os}: conc. superf. ossidante
C_{ox}: conc. ossidante "incluso" nell'ossido
D: diffusività nell'ossido
K: costante cinetica chimica

Figura 7.32 Schema della cinetica di ossidazione dove sono indicati: C_0 la concentrazione dell'atmosfera ossidante, C_{os} la concentrazione della specie ossidante all'interfaccia ossido-silicio, C_{ox} concentrazione dell'ossidante nell'ossido cresciuto, D Diffusività nell'ossido, k costante di cinetica chimica.

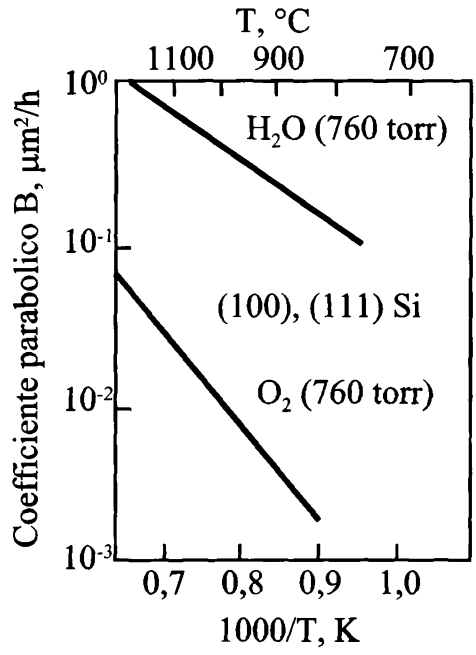


Figura 7.33 Coefficiente parabolico in funzione della temperatura.

2. La presenza di impurità droganti nel silicio che viene ossidato può portare o all'inclusione delle impurità nell'ossido oppure in alcuni casi alla loro parziale esplusione. Questo fenomeno viene indicato come *segregazione* dei droganti all'interfaccia tra ossido e silicio e può portare a concentrazioni di impurità nel silicio sottostante all'ossido maggiori di quelle iniziali.
3. Se sul wafer è presente uno strato di nitruro di silicio si osserva l'inibizione della

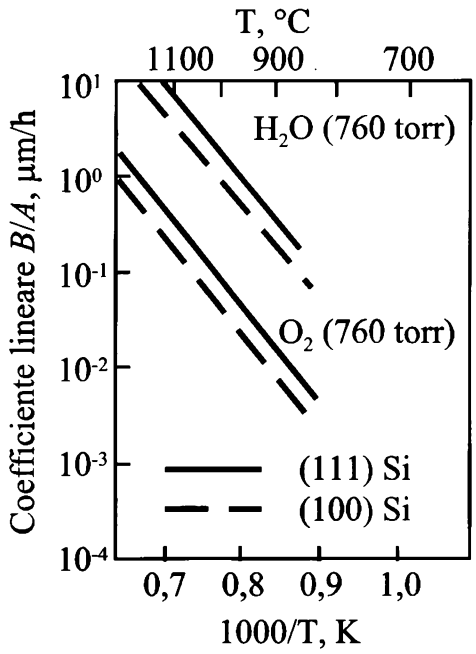


Figura 7.34 Coefficiente lineare in funzione della temperatura.

crescita dell'ossido a causa della scarsa diffusività delle speci ossidanti nel nitruro. Questa caratteristica del nitruro di silicio può essere utilmente sfruttato nella tecnica detta *LOCOS* nella quale la crescita dell'ossido viene resa selettiva.

Le fasi della tecnica LOCOS (*LOC*al *O*xidation of *S*ilicon) sono illustrate nella figura 7.35.

1. Sopra un sottile strato di ossido (buffer oxide), presente per adattare il coefficiente di dilatazione termica del silicio a quello del nitruro, vengono deposti circa 50 nm di nitruro di silicio con tecnica CVD.
2. Il nitruro viene rimosso dove si vuole crescere l'ossido utilizzando un processo fotolitografico standard e una fase di etching.
3. Il wafer viene quindi posto in un forno di ossidazione dove le regioni coperte dal nitruro non subiscono l'ossidazione.

Il vantaggio della tecnica LOCOS rispetto ad una sequenza tradizionale di ossidazione e successivo etching consiste nel fatto che il passaggio dalle zone ossidate a quelle protette dal nitruro è graduale infatti grazie alla diffusione laterale delle speci ossidanti nelle regioni sottostanti al perimetro di apertura del nitruro si produce una riduzione progressiva dello spessore dell'ossido, formando le zone dette a becco d'uccello, come ben visibile nella microfotografia di figura 7.36. Questo fenomeno garantisce una maggiore planarità della superficie e una resa migliore dei successivi procedimenti di deposizione di strato (ad esempio metallici).

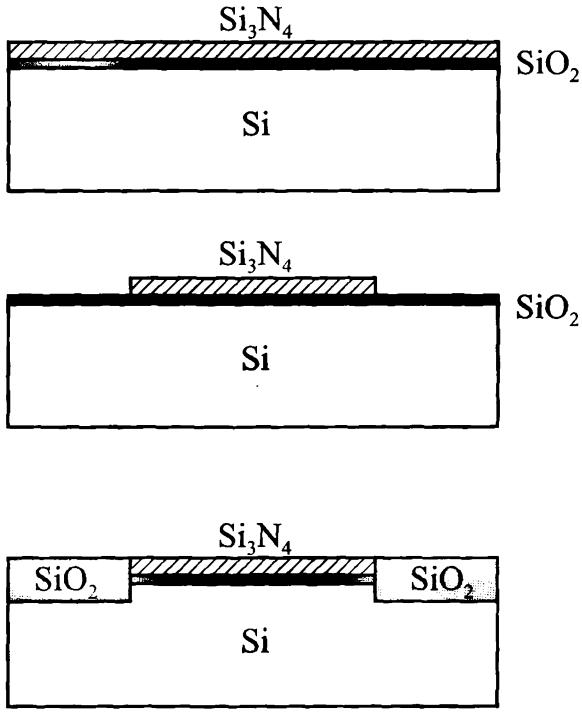


Figura 7.35 Sequenza nella tecnica LOCOS

Esempio 7.1 L'ossidazione termica del silicio trova applicazione nella realizzazione di importanti strutture quali ad esempio le trincee di isolamento per la separazione elettrica di dispositivi adiacenti sul medesimo substrato. In un substrato una trincea larga 1 μm e profonda 3 μm (ottenuta con un attacco chimico) come illustrato in figura 7.37, viene ossidata in ambiente di vapore acqueo a 1 atm e a 1100 °C in modo da riempirla e creare una trincea isolante.

1. Quanto è larga la striscia di biossido di silicio che si ha quando la trincea sia completamente riempita?
2. Quanto tempo è necessario per riempire la trincea con biossido di silicio?

Il rapporto tra il volume del Si consumato dall'ossido e il volume dell'ossido formatosi è il 44%. Pertanto lo spessore finale della trincea

$$\frac{1}{2}(d-l) = \frac{1}{2}\,0{,}44d$$

$$d-l = 0{,}44\,d \rightarrow d = \frac{l}{0{,}56} = 1{,}78\,\mu\text{m}$$

Il tempo per crescere $d/2 = 0{,}89$ μm di ossido viene calcolato usando l'equazione di ossidazione

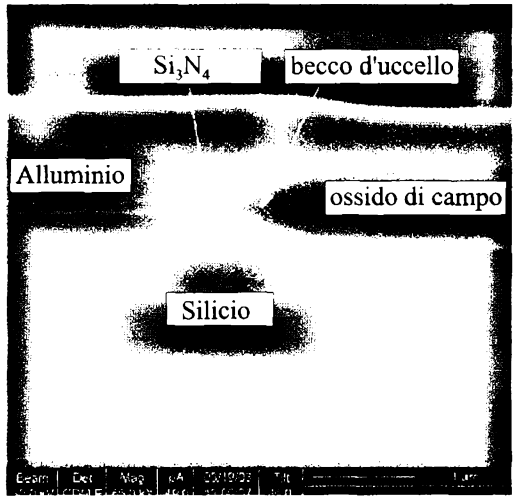


Figura 7.36 Microfotografia di una sezione ottenuta con la tecnica LOCOS

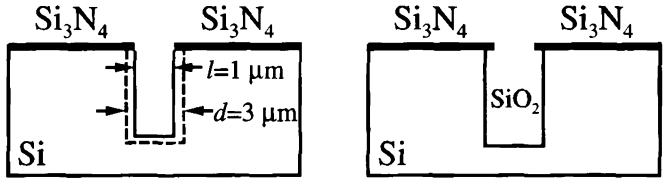


Figura 7.37 Creazione di una trincea tramite ossidazione termica

con i parametri B e B/A tabulati in figura 7.38.

$$t = \frac{A^2}{4B}\left[\left(\frac{2x_{ox}}{A}+1\right)^2 - 1\right]$$

$$t = \frac{A^2}{4B}\left(\frac{4x_{ox}^2}{A^2} + \frac{4x_{ox}}{A}\right)$$

$$t = \frac{x_{ox}^2}{B} + \frac{x_{ox}}{B/A}$$

con $P = 1$ atm, $T = 1100$ °C si trova

$$B = 0{,}5\ \frac{\mu\ \mathrm{m}^2}{\mathrm{h}} \qquad \frac{B}{A} = 3\ \frac{\mu\ \mathrm{m}}{\mathrm{h}}$$

da cui si ottiene il tempo di ossidazione necessario a chiudere la trincea $t = 1{,}88$ h.

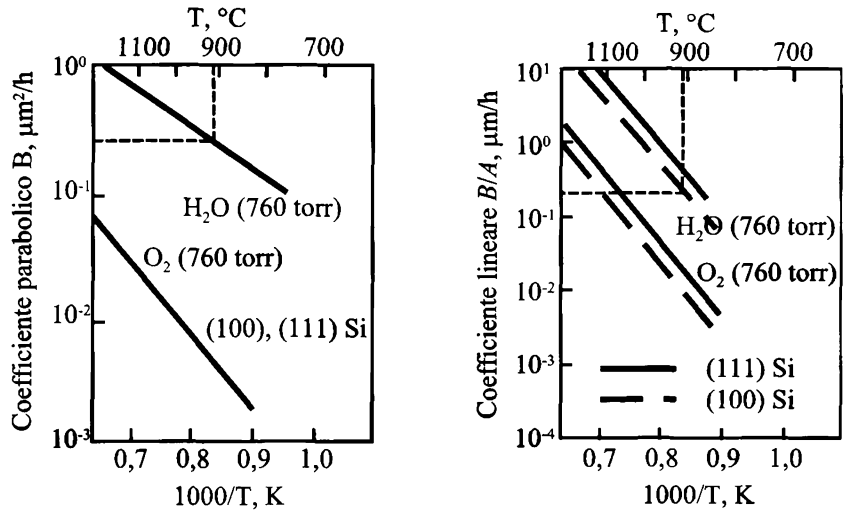


Figura 7.38 Utilizzo dei coefficienti di crescita parabolica e lineare.

7.6 Diffusione Termica

Attualmente le due tecniche più comuni per introdurre impurità droganti in un semiconduttore sono la diffusione termica e l'impiantazione ionica. Nella diffusione termica si distinguono due procedimenti:

- ▷ diffusione da fase gassosa dove il substrato viene posto in un forno e il drogante in forma gassosa raggiunge la superficie del semiconduttore e penetra all'interno.
- ▷ diffusione da fase solida dove il semiconduttore viene ricoperto tramite deposizione da un materiale contenente un'elevata concentrazione di impurità che nel processo termico diffondono nel semiconduttore.

In generale nella diffusione termica si sfrutta l'aumento della capacità dei droganti di diffondere nel silicio all'aumentare della temperatura. Per questo motivo il processo di diffusione deve avvenire tipicamente a temperature elevate ($\approx$ 1000 °C) all'interno di un forno di diffusione (figura 7.39).

Nel forno costituito da un tubo di quarzo vengono poste le fette che portate alla temperatura prevista vengono messe a contatto dei gas droganti quali ad esempio arsina AsH_3 e fosfina PH_3, lo schema è presente in figura 7.40.

Nei processi di drogaggio si vuol conoscere il profilo di drogaggio $C(x,t)$, noti i parametri (tipo di drogante, temperatura, durata) della diffusione. L'equazione che descrive la diffusione è l'equazione di *Fick*.

$$\frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial x^2}$$

dove la diffusività $D(T)$ nel silicio per i diversi droganti deve esserre ottenuta a partire da opportuni grafici quali quello indicato in figura 7.41.

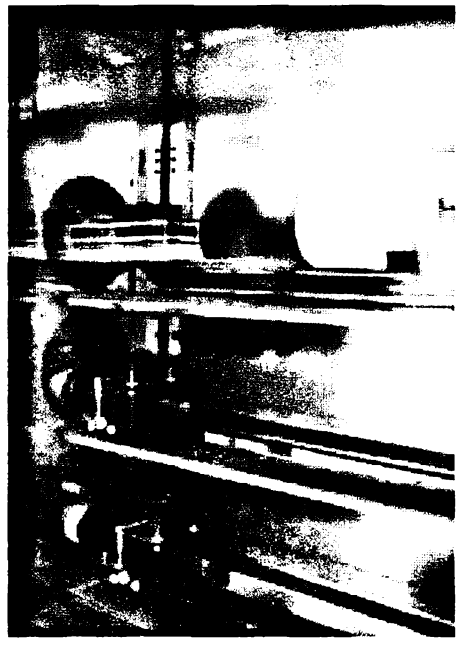


Figura 7.39 Fotografia di un forno di diffusione con slitte portafette.

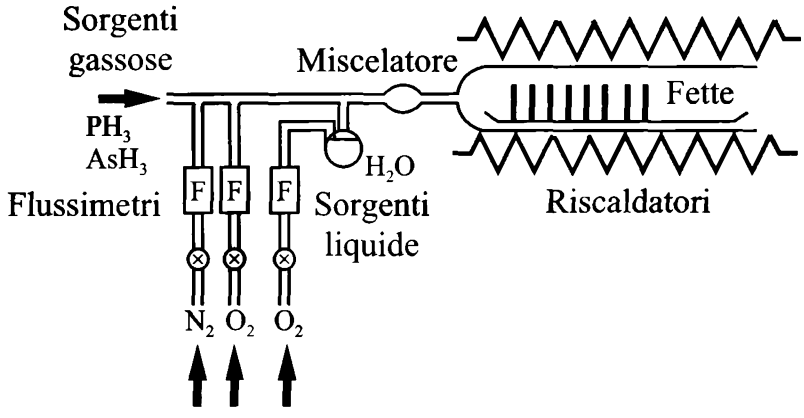


Figura 7.40 Schema di un forno per la diffusione termica di impurità droganti.

La soluzione dell'equazione di Fick richiede la definizione di condizioni al contorno consistenti, che nelle situazioni più comuni possono essere ricondotte ai seguenti due casi:

1. concentrazione superficiale costante;

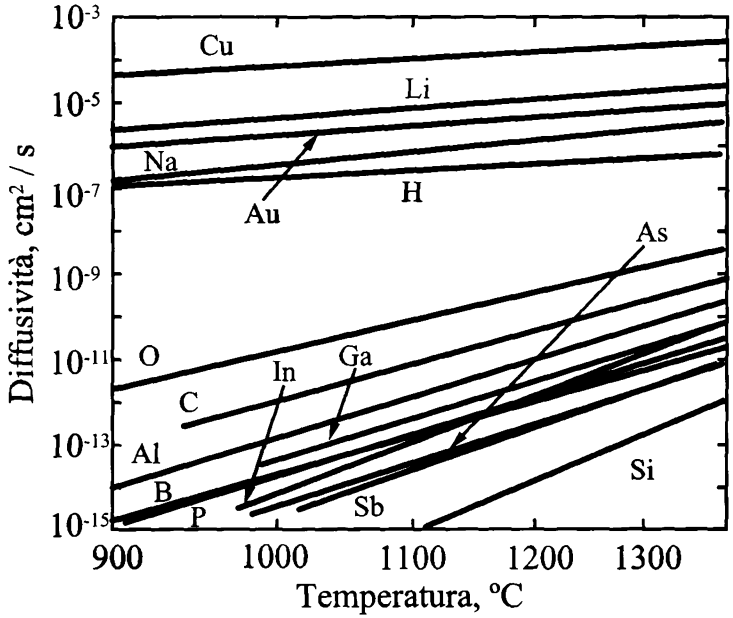


Figura 7.41 Diffusività delle principali impurità droganti nel silicio in funzione della temperatura.

2. dose totale di drogante costante.

Nel caso di concentrazione superficiale di drogante fissa $C(x = 0, t) = C_S$ il profilo di drogaggio che soddisfa l'equazione di Fick è del tipo:

$$C(x,t) = C_S \operatorname{erfc}\left(\frac{x}{L}\right)$$

$$= C_S \frac{2}{\sqrt{\pi}} \int_{\frac{x}{L}}^{\infty} \mathrm{e}^{-v^2}\, \mathrm{d}v$$

Dove $L = \sqrt{4Dt}$ è detta *lunghezza di diffusione* calcolata in funzione del tempo di diffusione t e della Diffusività D. I profili di drogaggio rappresentati su scala lineare o semilogaritmica sono illustrati in figura 7.42 al crescere del tempo nel quale il wafer è stato esposto all'atmosfera drogante con $t_1 > t_2 > t_3$. La funzione erfc è la funzione d'errore complementare che può essere calcolata in forma grafica utilizzando il valore x/L o utilizzando approssimazioni numeriche.

Integrando il profilo di drogaggio dalla superficie fino in profondità è possibile calcolare il numero totale di atomi introdotti durante la diffusione termica: tale valore è detto "dose". Il numero totale di atomi introdotti per unità di area cresce con il tempo

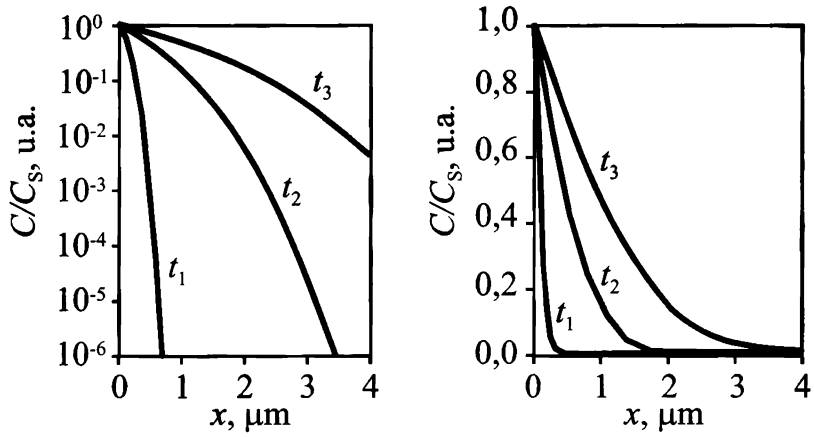


Figura 7.42 Profilo della concentrazione delle impurità nel caso di concentrazione superficiale costante.

t ed è calcolabile come:

$$N'(t) = \int_0^\infty C(x,t)\,dx = 2\sqrt{\frac{Dt}{\pi}}\,C_S$$

La seconda possibilità di risolvere l'equazione di Fick consiste nel porre la dose totale costante, a partire da una concentrazione impulsiva superficiale. Tale condizione è formalizzabile come

$$N'(t) = \int_0^\infty C(x,t)\,dx = \text{costante}$$

Risolvendo si ottiene una distribuzione per le impurità droganti di tipo *gaussiano*

$$C(x,t) = \frac{N'}{\sqrt{\pi}\sqrt{Dt}} \cdot e^{-\frac{x^2}{4Dt}} = \frac{2N'}{\sqrt{\pi}L} \cdot e^{-\frac{x^2}{L^2}}$$

dove nuovamente $L = \sqrt{4Dt}$ è ancora la lunghezza di diffusione In questo caso la concentrazione superficiale del drogante diminuisce con il tempo t come si può osservare dalla figura 7.43.

$$C(0,t) = \frac{N'}{\sqrt{\pi}\sqrt{Dt}}$$

Dato che l'intero wafer nei processi di fabbricazione può subire un numero relativamente grande di cicli termici ad alta temperatura è fondamentale saper valutare l'effetto che tali cicli hanno sui profili di drogaggio. Infatti quando il wafer subisce processi termici prolungati questi modificano inevitabilmente i profili di drogaggio esistenti

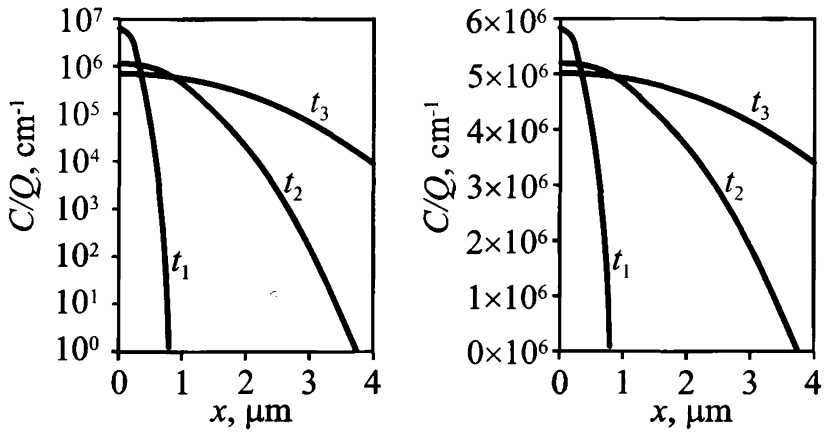


Figura 7.43 Profilo della concentrazione delle impurità nel caso di dose costante.

a causa della ridiffusione termica. Per poter ricalcolare i profili è necessario distinguere se il profilo originario derivi da una diffusione a concentrazione superficiale costante o a dose costante. Se il profilo deriva da una prediffusione gassosa a concentrazione superficiale costante questo profilo erfc può essere approssimato con un profilo gaussiano equivalente. Il profilo approssimante avrà una dose corrispondente a quella introdotta dalla prediffusione di durata t_0 e una lunghezza di diffusione L'. Supponendo quindi di indicare con D_0 e D_1 le costanti di diffusione dell'impurità rispettivamente nelle condizioni associate al processo iniziale e quello di ridiffusione, e con t_0 e t_1 i tempi di diffusione rispettivamente nella fase iniziale e in quella di ridiffusione si ottiene:

$$C(x,t) = \frac{2N'}{\sqrt{\pi}L'} \cdot e^{-\frac{x^2}{L'^2}}$$

$$N'(t) = 2\sqrt{\frac{D_0 t_0}{\pi}}\, C_S$$

$$L' = \sqrt{4\, D_0\, t_0 + 4\, D_1\, t_1}$$

Differentemente se il profilo iniziale deriva da una diffusione a dose costante e quindi il profilo è già gaussiano, la ridiffusione ha come unico effetto quello di modificare la lunghezza di diffusione come nel caso precedente lasciando inalterata la forma gaussiana del profilo.

7.7 Impiantazione Ionica

A partire dagli anni '80 la diffusione termica per l'introduzione di impurità droganti è stata progressivamente sostituita dall'impiantazione ionica. L'impiantazione ionica

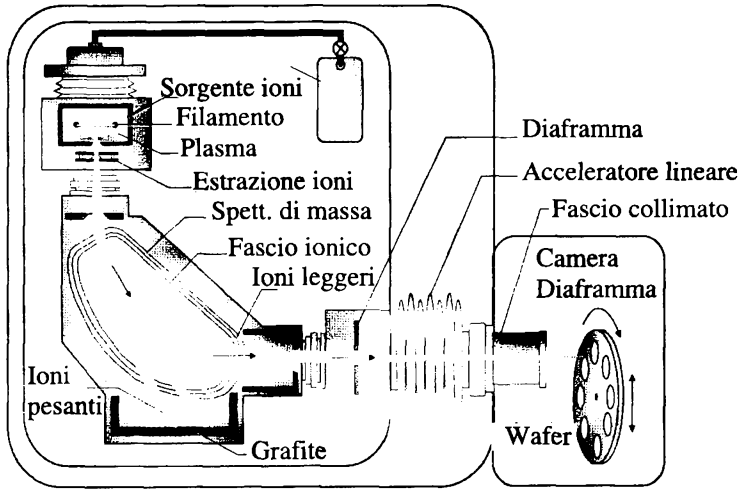


Figura 7.44 Schema costruttivo di un impiantatore ionico.

viene effettuta accelerando degli ioni della specie drogante tramite un campo elettrico fino a portare l'energia cinetica del singolo ione intorno ai 100 keV, in tal modo si rende possibile una penetrazione, anche di alcuni micron, all'interno del substrato da parte del drogante. La struttura di un impiantatore ionico è illustrata in figura 7.44 e alcune sue parti sono visibili nelle figure 7.45 e 7.46. ed è organizzata su alcuni moduli funzionali: in una camera a vuoto viene creato un plasma con gli ioni della specie che si vuole utilizzare, gli ioni vengono quindi estratti e inviati ad uno spettrografo di massa che separa gli ioni in fuzione del loro rapporto carica massa. In tal modo si ottiene una purificazione delle impurità, in quanto solo quelle della specie drogante avranno subito un raggio di curvatura tale da far attraversare il diaframma posto a valle dello spettrografo di massa. A questo punto un acceleratore lineare aumenta l'energia degli ioni che una volta collimati da una serie di lenti elettrostatiche, vengono diretti sul wafer da drogare.

I vantaggi dell'impiantazione ionica rispetto alla diffusione termica sono principalmente correlati al maggior controllo del profilo di drogaggio, infatti nell'impiantazione ionica vengono determinate con estrema precisione:

1. la *purezza* del drogante tramite lo spettrografo di massa;
2. l'*energia* degli ioni tramite il controllo dl campo elettrico applicato nell'acceleratore;
3. la *dose* totale misurata come integrale della corrente ionica indirizzata verso il wafer.

La determinazione del profilo di drogaggio ottenibile con l'impiantazione ionica richiede che vengano fatte alcune ipotesi sulla natura del reticolo cristallino colpito dagli ioni. Infatti grazie ad un'opportuna orientazione reticolare i danni indotti dalle prime fasi di impiantazione permettono di considerare il substrato "quasi amorfo", in altri termini viene meno la struttura cristallina che condizionerebbe la penetrazione degli ioni. Nel caso di materiale amorfo è possibile approssimare il profilo $C(x)$ ottenibile con

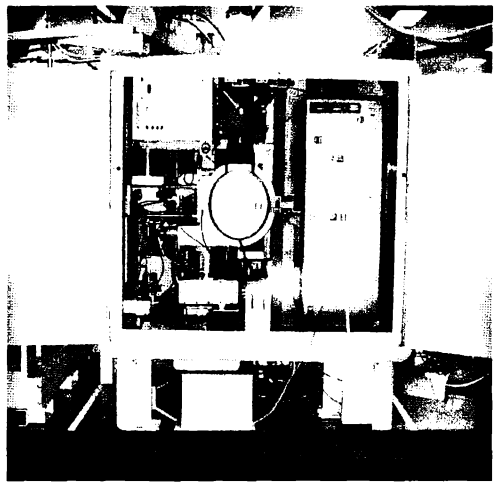


Figura 7.45 Fotografia della sezione sorgente di un impiantatore Varian VIIsion 80.

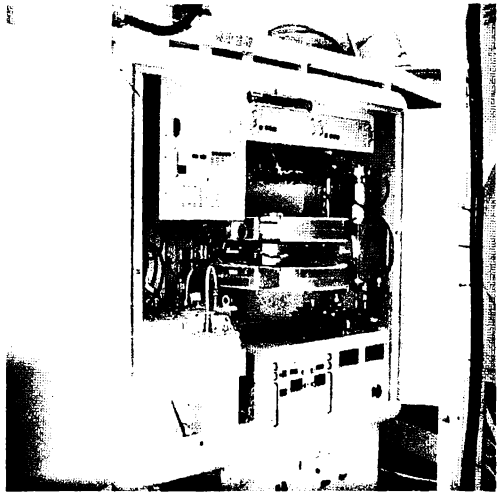


Figura 7.46 Fotografia della sezione di analisi di un impiantatore Varian VIIsion 80.

l'impiantazione ionica con una ditribuzione *gaussiana* con valore medio R_p e varianza ΔR_p:

$$C(x) = C_p \cdot \mathrm{e}^{-\frac{(x-R_p)^2}{L^2}}$$

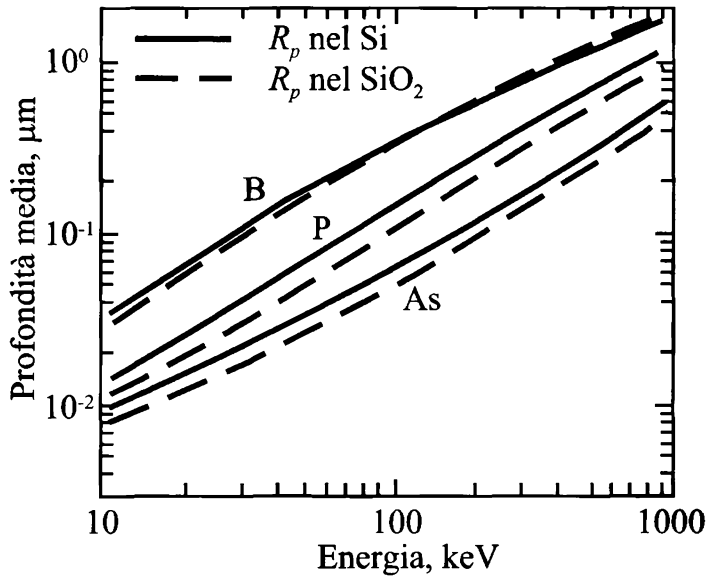


Figura 7.47 Profondità media di Impiantazione (Range Projected) R_p in funzione dell'energia.

Dove si è definita come lunghezza di diffusione equivalente

$$L = \sqrt{2\,\Delta R_p^2}$$

e con C_p la concentrazione di picco. I parametri R_p e ΔR_p vengono tabulati per i principali droganti in funzione dell'energia e del materiale nel quale avviene l'impiantazione (silicio, ossido, nitruro) come illustrato rispettivamente nelle figure 7.47 e 7.48

Il profilo di drogaggio può essere riscritto in funzione della dose N' e riportarlo in forma analoga a quella utilizzata per la diffusione termica. Infatti l'integrale della distribuzione può essere eguagliato alla dose N' permettendo di valutare la concentrazione di picco in funzione della semplice dose:

$$C_p \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{(x-R_p)^2}{L^2}}\,\mathrm{d}x = C_p\sqrt{\pi}\,L = N'$$

$$C_p = \frac{N'}{\sqrt{\pi}\,L}$$

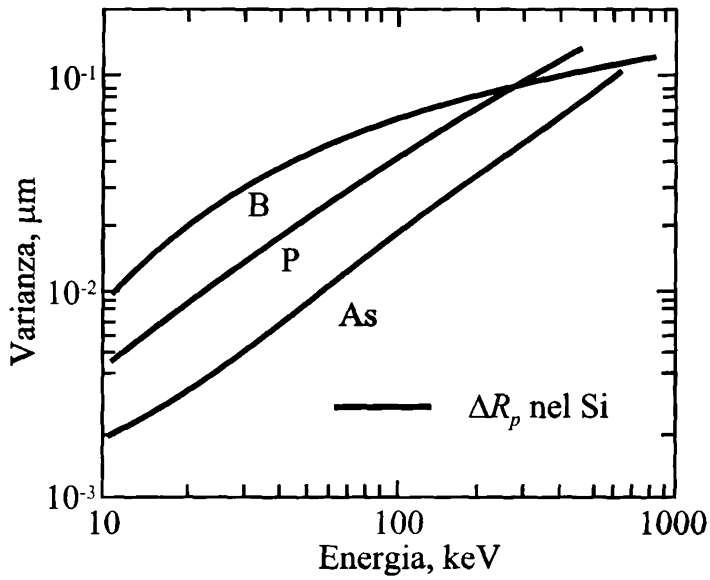


Figura 7.48 Varianza ΔR_p in funzione dell'energia di impiantazione.

La distribuzione può essere riscritta nella classica forma:

$$C(x) = \frac{N'}{\sqrt{\pi}\,L}\, e^{-\frac{(x-R_p)^2}{L^2}}$$

Soprattutto quando le dosi di drogante risultano molto elevate i danni al reticolo sono rilevanti a causa dei ripetuti urti degli ioni con il reticolo. I danni reticolari possono essere quindi riparati tramite un processo termico detto di *annealing* (o rinvenimento) che avviene riscaldando a T_a per un tempo t_a l'intero wafer, durante questo procedimento il silicio danneggiato ricristallizza e il drogante viene attivato. Il processo termico di annealing può essere analizzato in termini di diffusione termica a dose costante: la distribuzione resta *gaussiana* con una nuova lunghezza di diffusione.

$$C(x,t) = \frac{N'}{L'\sqrt{\pi}} \exp\left[-\left(\frac{x-R_p}{L'}\right)^2\right]$$

$$L' = \sqrt{2\,\Delta R_p{}^2 + 4\,D(T_a)\,t_a}$$

Esempio 7.2 In una fetta di silicio di tipo n con resistività di 5 Ω cm viene impiantata una dose di boro di 10^{12} cm^{-2} con un'energia di 100 keV, quindi si provoca un processo di ridiffusione con un annealing di 2 ore a 1000 °C come illustrato nella figura 7.49.
Si valuti l'evoluzione del profilo di drogaggio e la profondità della giunzione che si viene a creare.

Subito dopo l'impiantazione la distribuzione del drogante è gaussiana:

$$C(x) = C_p \cdot \mathrm{e}^{-\frac{(x-R_p)^2}{2\cdot\Delta R_p^2}}$$

$$C_p = \frac{N'}{\sqrt{\pi}\cdot\sqrt{2}\cdot\Delta R_p}$$

Dai grafici relativi al boro un'impiantazione con un'energia di 100 keV ha profondità media $R_p = 290\,\mathrm{nm}$ e varianza $\Delta R_p = 70$ nm
Nota la dose di boro che durante l'impiantazione raggiunge il substrato e la lunghezza di diffusione equivalente è possibile calcolare la concentrazione di picco:

$$N' = 10^{12}\ \mathrm{cm}^{-2}$$

$$L = \sqrt{2}\,\Delta R_p = 9{,}9\cdot 10^{-6}\ \mathrm{cm}$$

$$C_p = \frac{N'}{\sqrt{\pi}\cdot L} = 5{,}7\cdot 10^{16}\ \mathrm{cm}^{-3}$$

Inoltre dalla resistività, si ricava il valore del drogante nel silicio di tipo n

$$N_d = 9{,}2\times 10^{14}\ \mathrm{cm}^{-3}$$

Eguagliando la concentrazione del substrato con il profilo del boro e calcolando i valori di x che soddisfano l'equazione si ottiene la profondità di giunzione

$$C(x_j) = C_p \cdot \mathrm{e}^{-\frac{(x_j-R_p)^2}{2\,\Delta R_p^2}} = N_d$$

$$\frac{(x_j-R_p)^2}{2\cdot\Delta R_p^2} = \ln\frac{C_p}{N_d}$$

$$x_j = R_p \mp \sqrt{2}\cdot\Delta R_p\cdot\sqrt{\ln\frac{C_p}{N_d}}$$

Con i valori del profilo di drogaggio ottenuti subito dopo l'impiantazione si ottengono due valori di x che soddisfano l'equazione ad indicare che il profilo gaussiano ha realizzato due giunzioni. Infatti la regione di tipo p immediatamente dopo l'impiantazione è ampia $x_{j1} - x_{j2}$ come appare evidente nella figura 7.50.

$$x_{j1} = 491\ \mathrm{nm}$$

$$x_{j2} = 88.9\ \mathrm{nm}$$

$$\Delta x = 402\ \mathrm{nm}$$

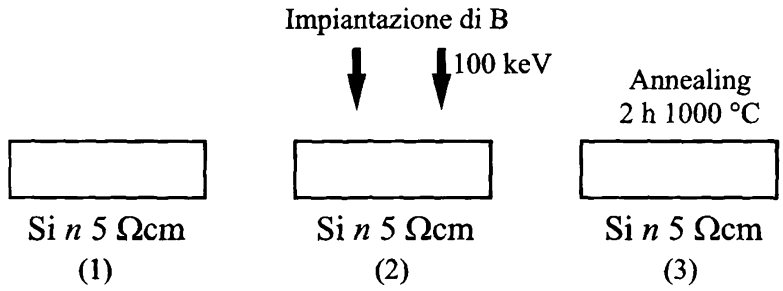


Figura 7.49 Impiantazione di Boro in un substrato di tipo n per la formazione di una giunzione.

Il processo di annealing determina una ridiffusione del drogante introdotto con l'impiantazione ionica. Il profilo iniziale per $t = t_0 = 0$ è gaussiano con dose nota e costante:

$$C(x)\,|_{t=t_0} = \frac{N'}{\sqrt{\pi}\cdot L}\cdot \mathrm{e}^{-\frac{(x-R_p)^2}{L^2}}$$

$$L = \sqrt{2}\,\Delta R_p$$

Dopo un tempo t di diffusione, la distribuzione è ancora gaussiana con lo stesso valore medio e con lunghezza equivalente di diffusione pari a

$$L' = \sqrt{2\,\Delta R_p^2 + 4\,D\cdot t} = 2{,}6\times 10^{-5}\ \text{cm}$$

Alle condizioni di annealing la diffusività del boro è pari a $D = 2\cdot 10^{-14}\text{cm}^2\,\text{s}^{-1}$

$$C(x)\,|_{t>t_0} = \frac{N'}{\sqrt{\pi}\cdot L'}\cdot \mathrm{e}^{-\frac{(x-R_p)^2}{(L')^2}}$$

$$C_p' = N'\sqrt{\pi}\cdot L' = 2{,}17\times 10^{16}\ \text{cm}^{-3}$$

La ridiffusione porta a ridurre la concentrazione di picco ma simultaneamente ad aumentare la varianza della gaussiana, e questo determina che la seconda giunzione si sposti alla profondità $x_j = 752$ nm, mentre la prima scompaia.

7.8 Crescite epitassiali

Utilizzando le tecniche epitassiali è possibile realizzare su di un substrato monocristallino strati di semiconduttore con tipo e profilo di drogaggio arbitrari utilizzando le tecniche epitassiali. La condizione necessaria per questo tipo di crescite è che la superficie del monocristallo possa venire a contatto dei materiali utilizzati per la crescita, in quanto solo in questo modo i nuovi strati potranno ricristallizare seguendo la struttura cristallina sottostante.

Le principali tecniche di crescita epitassiali sono:

1. *LPE*: Liquid Phase Epitaxy ovvero epitassia da fase liquida;

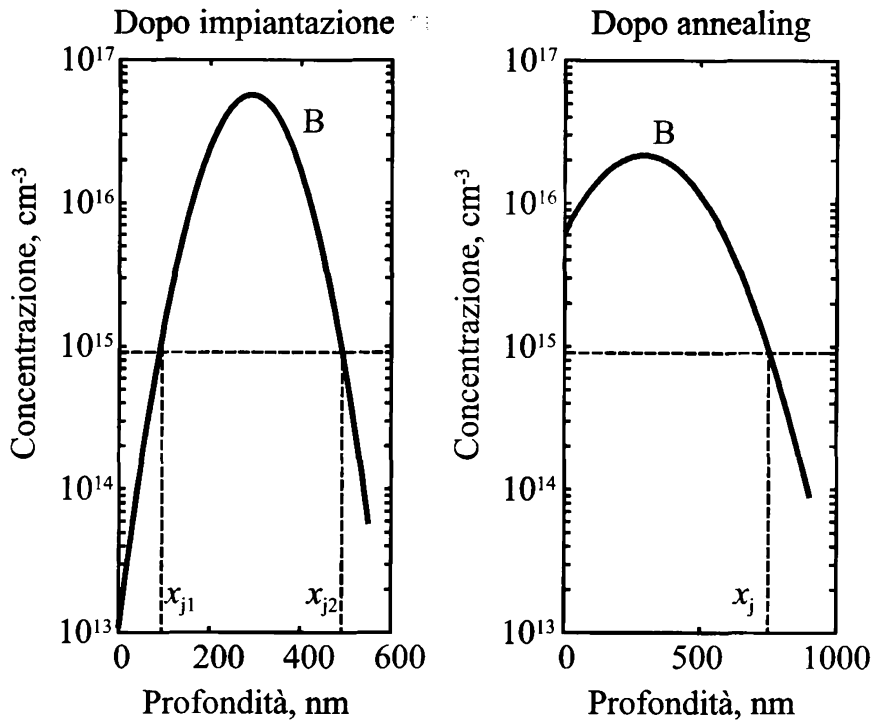


Figura 7.50 Profili di drogaggio del Boro prima dopo il processo di annealing.

2. *VPE*: Vapor Phase Epitaxy ovvero epitassia da fase vapore;
3. *MBE*: Molecular Beam Epitaxy ovvero epitassia da fascio molecolare.

La crescita LPE di nuovi strati monocristallini avviene mettendo il substrato riscaldato a contatto con soluzioni liquide: la crescita avviene in modo simile al metodo CZ. La struttura che permette di ottenere una crescita LPE è illustrara in figura 7.51 dove risulta visibile la navicella in grafita nella quale i wafer ad alta temperatura vengono a contatto con i materiali per la crescita presenti in forma liquida. Spostando la navicella è possibile accrescere strati di materiali differenti o semplicemente con drogaggio differente.

Nel caso di crescita epitassiale da fase vapore i substrati vengono inseriti in un reattore (figura 7.52) dove i differenti composti reagendo determinano la crescita di strati monocristallini sui wafer disposti in modo da garantire la maggior conformità possibile degli strati cresciuti. Durante la fase di crescita possono essere variati tipi e concentrazioni dei droganti realizzando strutture non ottenibili con altre tecniche di drogaggio. I substrati vengono mantenuti a circa 1000 °C.

Nella tecnologia del silicio la VPE è ampiamente utilizzata facendo ricorso alle seguenti reazioni di crescita:

$$SiCl_4(g) + 2H_2(g) \rightarrow Si(s) + 4HCl(g)$$

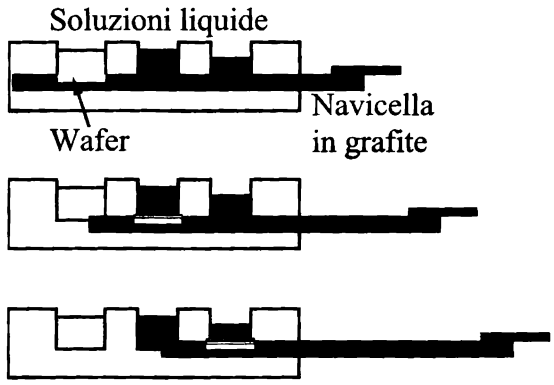


Figura 7.51 Schema per la crescita epitassiale da fase liquida.

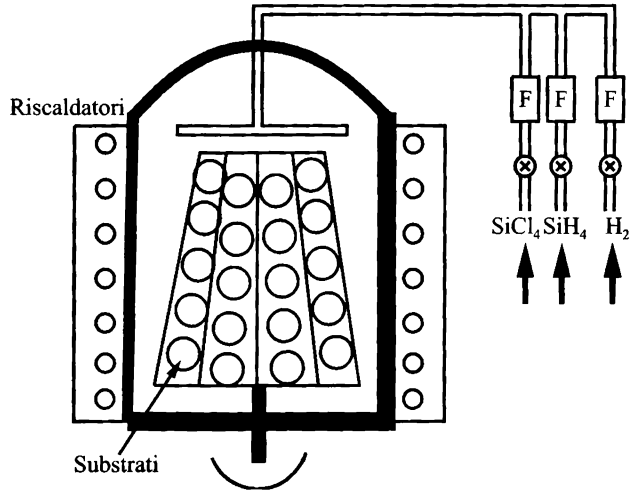


Figura 7.52 Schema per la crescita epitassiale da fase vapore.

$$SiHCl_3(g) + H_2(g) \rightarrow Si(s) + 4HCl(g)$$

$$SiH_2Cl_2(g) \rightarrow Si(s) + 2HCl(g)$$

$$SiH_4(g) \rightarrow Si(s) + 2H_2(g)$$

7.9 Crescite non epitassiali

È possibile realizzare strati di differenti materiali quali *ossido*, *nitruro*, *silicio policristallino* indipendentemente dal fatto che la superficie del semiconduttore sia accessibile,

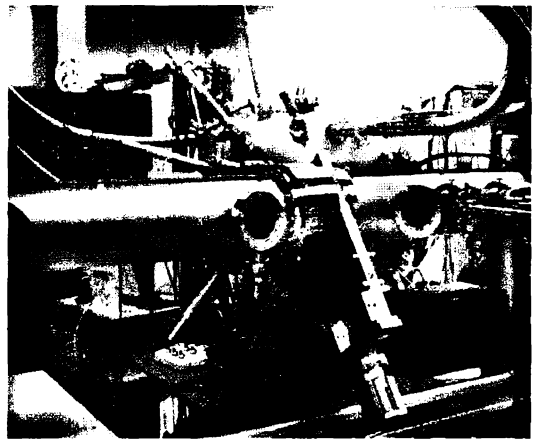


Figura 7.53 Fotografia di un reattore per epitassia.

differentemnte dall'epitassia. Nei processi MOS la crescita di silicio policristallino tramite CVD (Chemical Vapor Deposition) ha giocato un ruolo chiave nello sviluppo delle tecnologie integrate di tipo silicon gate. La crescita non epitassiale è inoltre di estrema importanza nella realizzazione di tutte quelle strutture quali le interconnessioni che richiedono di poter realizzare un buon isolamento tra piste adiacenti. In questi ultimi anni si è molto lavorato per arrivare a deporre strati con caratteristiche elettriche (bassa costante dielettrica per ridurre le capacità parassite) e elevata planarità compatibili con processi di fabbricazioni dove i livelli di interconnessione sovrapposte possono essre anche una decina.

In generale la procedura richiede che i substrati vengano introdotti all'interno di reattori per la deposizione da fase vapore (CVD) dove a partire da alcuni composti vengono realizzati degli strati che possono raggiungere anche alcuni micron di spessore.

Le principali reazioni utilizzate per la crescita di ossido, nitruro e silicio policristallino sono:

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$$

$$Si(OC_2H_5)_4 \rightarrow SiO_2 + \text{composti}$$

$$SiCl_2H_2 + 2N_2O \rightarrow SiO_2 + 2N_2 + 2HCl$$

$$3SiCl_2H_2 + 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H_2$$

$$SiH_4 \rightarrow Si + 2H_2(g)$$

7.10 Metallizzazioni

La realizzazione di strati metallici ha assunto un'enorme importanza nei processi planari sia per accedere direttamente ai dispositivi integrati sia per permettere l'interconne-

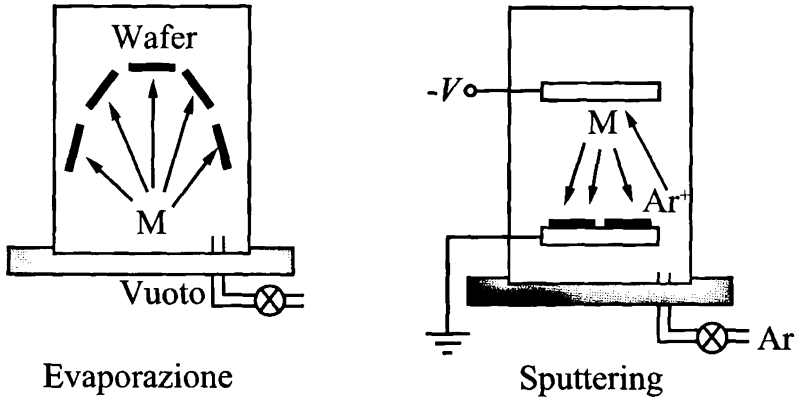


Figura 7.54 Schemi per la deposizione di film metallici per evaporazione e per sputtering.

sione tra celle e blocchi più o meno delocalizzati nel circuito integrato. Le realizzazione di strati metallici si basa su tecniche o di deposizione o di crescita elettrolitica (electroplating). Quest'ultima viene ampiamente utilizzata per la realizzazione di interconnessioni in rame e si basa sulla possibilità di crescere degli strati metallici utilizzando un bagno galvanico nei quali i wafer, precedentemente ricoperti da un sottile strato iniziale (seed layer), vengono posti. Le tecniche di deposizione dei metalli maggiormente utilizzate sono:

1. *evaporazione*: il metallo viene riscaldato in una campana a vuoto come illustrato in figura 7.54 fino a che le molecole metalliche evaporando si depositano sui substrati, tale tecnica risulta attuabile solo per metalli che presentino una tensione di vapore sufficientemente elevata a temperature tecnologicamente raggiungibili, ovvero per metalli che abbiano una bassa temperatura di fusione;
2. *sputtering*: il metallo viene colpito con ioni ad alta energia di gas inerti che determinano il distacco di atomi dal metallo (anche se esso presenta un'elevata temperatura di fusione) che vanno a quindi raggiungere i substrati sui quali formano il film metallico.

Capitolo 8

Memorie a semiconduttore

Le memorie rappresentano uno dei settori più importanti dell'industria dei semiconduttori: negli ultimi 30 anni, la diffusione delle memorie a semiconduttore è sempre andata crescendo, seguendo l'affermarsi dei personal computer, dell'elettronica di consumo, della telefonia cellulare. I prezzi delle memorie subiscono tipicamente forti fluttuazioni legate alla domanda; tuttavia, tra il 1995 e oggi, le memorie hanno coperto una percentuale del mercato mondiale complessivo dei circuiti integrati che, in termini di fatturato, non è mai scesa al di sotto del 20% e ha raggiunto per alcuni periodi valori anche superiori al 40%, per un valor medio attestato intorno al 30%.

Dal punto di vista tecnologico, il settore delle memorie si è dimostrato tra i più vivaci, con un'evoluzione molto rapida, che ha contribuito notevolmente al successo di molti prodotti e applicazioni dell'informatica e delle telecomunicazioni. Un buon esempio di questo tipo di evoluzione è dato nella figura 8.1, dove è illustrato l'andamento della capacità di memoria e della lunghezza di canale per il caso particolare della tecnologia DRAM, relativa alle memorie RAM dinamiche, che sono utilizzate in tutti i tipi di computer.

Il modo più tradizionale di classificare le memorie a semiconduttore fa riferimento al comportamento rispetto all'assenza della tensione di alimentazione: si distigue quindi tra memorie volatili, che conservano l'informazione immagazzinata soltanto in presenza di alimentazione, e memorie non volatili, che permettono di mantenere l'informazione anche in assenza di alimentazione. Lo sviluppo di quest'ultimo tipo di memoria ha consentito il diffondersi di numerosi prodotti digitali portabili privi di supporti magnetici o ottici (telefoni cellulari, lettori MP3, PDA, ...).

Una classificazione più completa delle memorie a semiconduttore di maggiore importanza è data nella tabella 8.1. Le memorie a sola lettura sono le più semplici, ma il loro utilizzo è limitato a poche applicazioni, per le quali il contenuto possa essere definito all'atto della realizzazione fisica ("mask programmed ROM") o al momento della programmazione ("programmable ROM") e non debba mai essere modificato. Alla categoria delle memorie non volatili appartengono un gran numero di tipologie, con diverse caratterististiche di costo e prestazioni: nella tabella 8.1 sono elencate le principali, EPROM, EEPROM e Flash. Infine la terza colonna della tabella riporta le memorie RAM, di tipo statico e dinamico, strutturate in modo da garantire tempi di accesso in lettura e scrittura confrontabili. Le sezioni seguenti presenteranno le caratteristiche

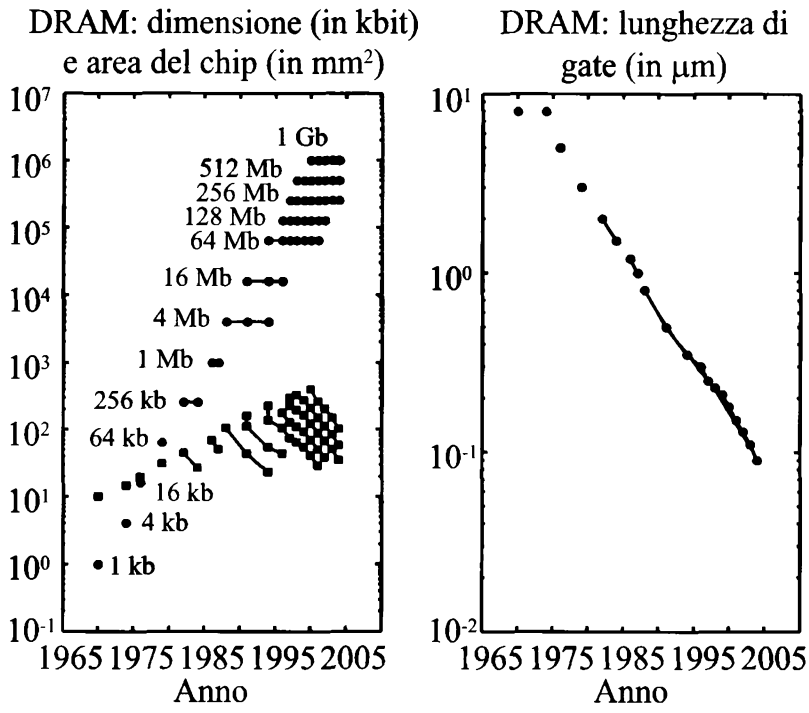


Figura 8.1 Evoluzione della tecnologia delle memorie RAM dinamiche: capacita' di memoria in Kbit e lunghezza di canale del transistore in μm tra il 1970 e il 2005 (fonte: http://www.icknowledge.com).

salienti di questi tipi di memorie a semiconduttore.

Memorie a sola lettura ROM	Memorie non volatili NVRWM	Memorie a lettura e scrittura RAM
- Mask programmed ROM - Programmable ROM (PROM)	- EPROM - EEPROM o E^2PROM - Flash	- statiche (SRAM) - dinamiche (DRAM)

Tabella 8.1 Classificazione delle memorie a semiconduttore.

Si danno anche altri criteri di classificazione delle memorie, legati per esempio alle modalità di accesso: le memorie volatili possono essere ad accesso casuale (si indica con questo termine, "random access memory", un accesso che richiede lo stesso tempo indipendentemente dalla posizione della singola locazione di memoria) oppure

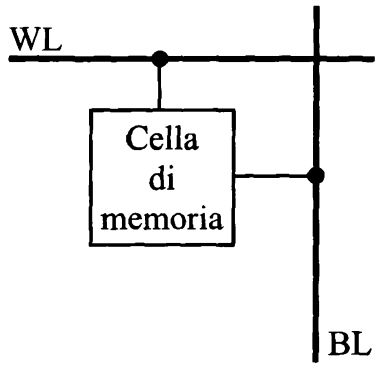


Figura 8.2 Cella di memoria per 1 bit di informazione. WL: *Word Line*, BL: *Bit Line*.

possono prevedere un ordinamento implicito, come quello di tipo FIFO ("First-In, First-Out"); possono avere comportamento sincrono o asincrono nelle operazioni di lettura e scrittura; possono infine essere a singola o multipla porta: nel primo caso, è fattibile una sola operazione di accesso per volta, mentre nel secondo, l'architettura della memoria permette di eseguire più operazioni di scrittura/lettura simultanee in locazioni distinte.

8.1 Architettura di una memoria

Nonostante la grande varietà di memorie a semiconduttore oggi disponibili e la conseguente dispersione nelle caratteristiche tecnologiche e di funzionamento, è possibile ricondurre l'organizzazione generale di gran parte delle memorie a una medesima architettura di massima. Ogni tecnologia ha poi proprie peculiarità nella struttura della singola cella di memoria, in grado di conservare un bit, e tipicamente introduce alcuni elementi di variazione anche nell'architettura generale. In questo paragrafo sarà quindi presentata la struttura complessiva alla quale sono riconducibili quasi tutte le memorie a semiconduttore, mentre le sezioni successive descriveranno in maggiore dettaglio le specificità delle principali memorie.

Fondamentalmente l'organizzazione generale, comune a tutti i tipi di memoria, è matriciale e richiede prima la selezione della cella, o delle celle, da leggere o scrivere e successivamente l'operazione vera e propria di trasferimento del dato; cambiano invece da memoria a memoria la struttura e le caratteristiche della singola cella di informazione.

Consideriamo una cella di memoria in grado di immagazzinare 1 bit di informazione (figura 8.2) Sono sufficienti due linee per connettere la cella con il mondo esterno:

- la *Bit Line* **BL** trasporta il bit da o verso la cella
- la *Word Line* **WL** trasporta il segnale di selezione della cella, derivato dal segnale indirizzo.

La cella risulta connessa alla *Bit Line* solo se il segnale di selezione è attivo, altrimenti la cella non è accessibile dalla *Bit Line* e rimane isolata.

Con un'organizzazione a matrice, si possono collocare celle identiche disposte su più righe e più colonne: ad ogni riga deve corrispondere una *Word Line* distinta, che

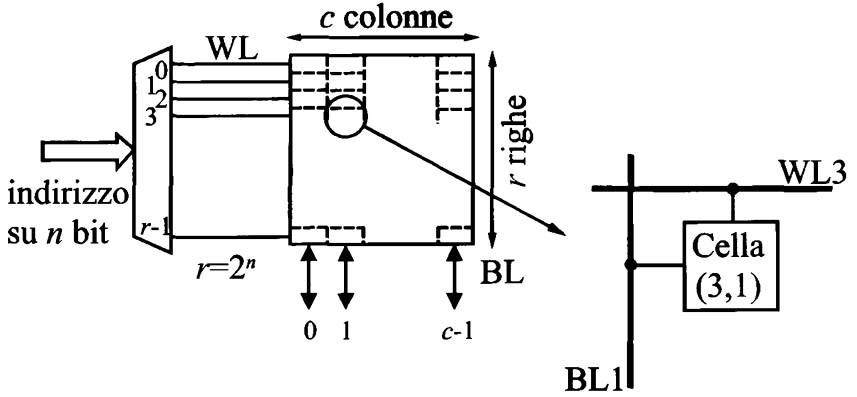


Figura 8.3 Architettura a matrice per una memoria.

permette di selezionare tutte e sole le celle di una riga. L'organizzazione a matrice è del tipo indicato nella figura 8.3. Se la matrice di celle ha dimensione $r \times c$, ovvero r righe e c colonne, ad ogni lettura o scrittura una soltanto per volta delle r linee di selezione può essere attivata: questo garantisce che ogni *Bit Line* sia connessa a una sola cella della colonna corrispondente. L'accesso alla memoria si traduce nel trasferimento in parallelo di un dato composto da c bit.

Al fine di selezionare una delle $r = 2^n$ *Word Line*, la memoria deve ricevere n bit di indirizzo, che sono convertiti nei segnali di selezione delle righe da un "decoder". La struttura circuitale a livello logico di un "decoder" è semplice, sebbene in memorie reali tale componente sia oneroso in termini di dimensioni e richieda l'adozione di opportune tecniche circuitali volte a semplificarlo.

Le *Bit Line* della matrice terminano in opportuni "buffer" che pilotano le linee di dato esterne durante le operazioni di lettura; nelle operazioni di scrittura, le *Bit Line* ricevono i valori da immagazzinare nella riga selezionata.

L'organizzazione matriciale descritta è da considerarsi solo come schema logico di funzionamento della memoria, mentre la struttura fisica è piuttosto diversa. La differenza principale nasce dal fatto che i tagli tipici di interesse applicativo corrispondono a valori di r e c piuttosto sbilanciati ($r >> c$), tali da determinare un rapporto d'aspetto molto lontano da 1. Per esempio, nell'ipotesi che la singola cella di memoria sia di forma circa quadrata, la matrice di una memoria di 1M byte, per la quale $n = 20$, $r = 2^{20}$ e $c = 8$, avrebbe un rapporto d'aspetto di 131072 a 1, cioè sarebbe 131072 volte più alta che larga! Una tale forma geometrica comporterebbe parecchi problemi, relativi per esempio al packaging e al ritardo di propagazione dei segnali lungo le *Bit Line*.

Nell'architettura fisica di una memoria (figura 8.4), si preferisce allora allocare più di una parola sulla medesima riga, fino a ottenere un rapporto d'aspetto circa unitario e quindi una matrice di forma circa quadrata. Con riferimento all'esempio precedente, si può scegliere $r = 2^{12}$ e $c = 2^{11}$, ottenendo una matrice di $r \times c = 2^{23}$ celle. Naturalmente sarà necessario selezionare tra le celle della riga abilitata quelle che compongono la parola da leggere o scrivere: questo si può ottenere con un "multiplexer"

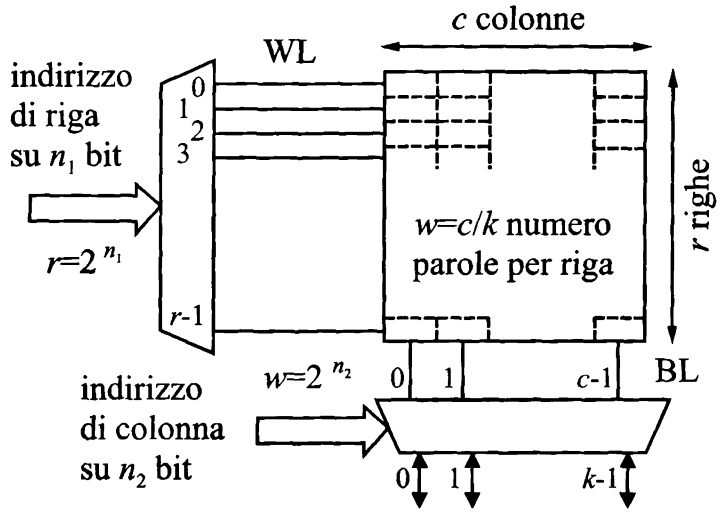


Figura 8.4 Architettura fisica di una memoria.

a più vie. L'indirizzo sarà diviso in due parti, la prima (*indirizzo di riga*) servirà per abilitare una delle righe della matrice, la seconda (*indirizzo di colonna*) permetterà di selezionare mediante il multiplexer la parola all'interno della riga. Nella figura 8.4, il numero complessivo di bit di indirizzo è suddiviso in due contributi,

$$n = n_1 + n_2$$

il primo, n_1, determina il numero di righe della matrice, $r = 2^{n_1}$, e la dimensione del decoder di riga; il secondo contributo, n_2, è legato alla molteplicità w delle parole di informazione memorizzate sulla medesima riga della matrice:

$$\frac{c}{k} = w \quad w = 2^{n_2}$$

8.2 Tempistiche di accesso

L'accesso in lettura o scrittura a un dispositivo di memoria avviene utilizzando i segnali di dato, indirizzo e controllo secondo precisi protocolli, che vengono specificati dal costruttore all'interno della documentazione tecnica (*data sheets*). Tra i numerosi tipi di protocollo usati a questo scopo, quelli più diffusi sono:

- il protocollo sincrono, che si appoggia a un segnale di sincronismo (*clock*) per temporizzare le sequenze di lettura e scrittura
- il protocollo asincrono, che sfrutta alcuni fronti dei segnali di controllo per temporizzare le operazioni di accesso.

In entrambi i casi, la procedura corretta di attivazione dei segnali e i tempi corrispondenti devono essere specificati dal costruttore.

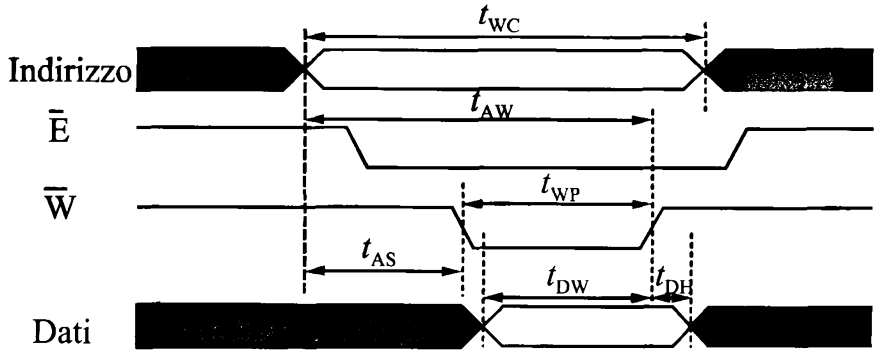


Figura 8.5 Sequenza di scrittura.

parametro	*simbolo*	*min*	*max*	*unità*
Write Cycle Time	t_{WC}	25	-	ns
Address Setup Time	t_{AS}	0	-	ns
Write Pulse Width	t_{WP}	20	-	ns
Data Hold Time	t_{DH}	0	-	ns
Address Valid to End of Write	t_{AW}	20	-	ns
Data Valid to End of Write	t_{DW}	10	-	ns

Tabella 8.2 Definizioni e valori esemplificativi dei ritardi associati ai segnali coinvolti nell'operazione di scrittura.

Consideriamo come esempio una RAM di tipo asincrono e vediamone le procedure tipiche di accesso in lettura e scrittura.

I segnali fondamentali coinvolti sono:

- i dati, **Q**, segnali bidirezionali in numero pari al parallelismo interno della memoria
- gli indirizzi, **A**, segnali che codificano la posizione del dato richiesto all'interno della memoria
- il *chip select*, $\overline{E}$, segnale di abilitazione del componente, utile quando il sistema comprenda più dispositivi di memoria, selezionabili indipendentemente (attivo a 0 in questo caso)
- il *write enable*, $\overline{W}$, segnale di abilitazione alla scrittura (attivo a 0)
- l'*output enable*, $\overline{G}$, segnale di abilitazione alla lettura (attivo a 0)

Al fine di effettuare un'operazione di scrittura, è richiesta una ben precisa sequenza di attivazione dei segnali; prima di tutto, occorre forzare sulle linee dati i bit che si intendono memorizzare, poi si deve selezionare la locazione di memoria desiderata e infine bisognerà attivare il comando di scrittura vero e proprio, secondo la seguente successione di operazioni:

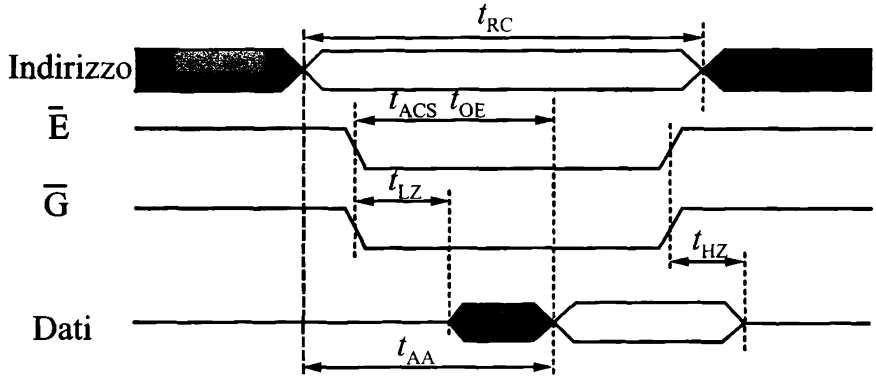


Figura 8.6 Sequenza di lettura.

1. *attivazione dell'indirizzo e e dei dati dati*
2. *attivazione del chip select*
3. *attivazione del write enable*

Questa sequenza di attivazioni è illustrata nelle forme d'onda di figura 8.5, dove sono state adottate le usuali convenzioni per indicare le transizioni dei segnali. La figura riporta anche i principali ritardi che caratterizzano l'operazione di scrittura e che sono riassunti anche nella tabella 8.2. L'informazione temporale più importante è costituita dal *Write Cycle Time*, t_{WC}, che rappresenta la minima durata dell'intera operazione di scrittura e che può essere misurata come il periodo di tempo minimo per il quale gli indirizzi della locazione selezionata devono restare stabili. L'*Address Setup Time*, t_{AS}, è invece il minimo anticipo con il quale gli indirizzi devono essere stabili prima dell'attivazione del segnale di scrittura ($\overline{W}$). Il *Write Pulse Width*, t_{WP}, è definito come la durata minima di attivazione del comando di scrittura. *Data Hold Time*, t_{DH}, è il tempo minimo concesso per porre le linee di dato della memoria nella condizione di alta impedenza dopo che il comando di scrittura è stato disattivato. Questa informazione è rilevante, perché le linee di dato sono sempre controllate da più di un circuito di pilotaggio: per esempio, in un sistema a microprocessore, queste linee sono controllate almeno dalla memoria stessa, che controlla le linee di dato durante le letture, e dal processore, che pilota le linee di dato durante gli accessi in scrittura. Il *Data Hold Time* permette di conoscere il primo istante di tempo in cui per un altro circuito di pilotaggio è lecito iniziare a comandare le linee di dato. *Address Valid to End of Write*, t_{AW}, indica il tempo minimo entro il quale il comando di scrittura può essere disattivato a partire dall'istante in cui gli indirizzi sono stabili; analogamente *Data Valid to End of Write*, t_{DW}, definisce il tempo minimo per il quale i segnali di dato devono restare stabili prima della corretta conclusione dell'operazione di scrittura.

Il protocollo attivato nell'accesso in lettura prevede la seguente sequenza di operazioni, anche descritta nelle forme d'onda di figura 8.6:

1. *indirizzo valido*
2. *chip select attivo*
3. *write enable non attivo*

parametro	*simbolo*	*min*	*max*	*unità*
Read Cycle Time	t_{RC}	25	-	ns
Address Access Time	t_{AA}	-	25	ns
Chip Enable Access Time	t_{ACS}	-	25	ns
Output Enable Access Time	t_{OE}	-	12	ns
Chip Enable Low to Output Active	t_{LZ}	5	-	ns
Chip Enable High to Output High-Z	t_{HZ}	0	10	ns

Tabella 8.3 Definizioni e valori esemplificativi dei ritardi associati ai segnali coinvolti nell'operazione di lettura.

Le definizioni dei tipici ritardi misurati sulle forme d'onda relative all'accesso in lettura sono riportate nella tabella 8.3. Anche in questo caso, la durata minima dell'accesso in lettura è data dal *Read Cycle Time*, t_{RC}.
L'*Address Access Time*, t_{AA}, è il massimo ritardo con il quale la memoria restituisce i dati letti validi a partire dall'istante in cui si sono stabilizzati gli indirizzi.
Il *Chip Enable Access Time*, t_{ACS}, e l'*Output Enable Access Time*, t_{OE}, misurano invece il massimo ritardo con il quale i dati validi sono ottenuti a partire dall'attivazione del *chip select* e dell'*output enable* rispettivamente e, nel caso della figura 8.6, coincidono.
Il *Chip Enable Low to Output Active*, t_{LZ}, indica il minimo ritardo con il quale il dispositivo inizia a pilotare le linee di dato dopo la ricezione del comando di lettura (*output enable*).
Infine il *Chip Enable High to Output High-Z*, t_{HZ}, permette di conoscere il tempo massimo che il dispositivo di memoria impiega per isolarsi dalle linee di dato al termine dell'operazione di lettura, ovvero quando il segnale di *output enable* ritorna inattivo.

8.3 La cella ROM

Una ROM è una memoria a sola lettura ("Read Only Memory"), un dispositivo cioè per il quale il contenuto di informazione è noto al momento della realizzazione fisica e non può più essere alterato successivamente. Poichè una memoria ROM può essere oggetto soltanto di operazioni di lettura, la struttura del componente e in particolare della singola cella risulta fortemente semplificata e compatta rispetto a quanto si vedrà per le memorie a lettura e scrittura: le ROM sono quindi componenti molto densi e potenzialmente molto veloci. Le applicazioni, d'altra parte, sono limitate a pochi casi, per i quali si rendono necessari elevati volumi di produzione per memorie di contenuto identico; un esempio può essere la memoria con il codice di "boot" di sistemi a processore.

L'idea fondamentale di una cella ROM è illustrata nella figura 8.7. Si consideri l'incrocio di una *Word Line*, pilotata da un "decoder" di riga, con una *bit line*, che si suppone connessa a massa attraverso una resistenza: se all'incrocio le due linee sono unite da un diodo, l'eventuale selezione della cella, effettuata applicando una tensione di qualche Volt, V_{WL}, sulla *Word Line*, porterà il diodo in conduzione e alzerà la tensione sulla *bit line* fino a $V_{WL} - 0{,}6$ V; se invece le due linee sono isolate l'una dall'altra,

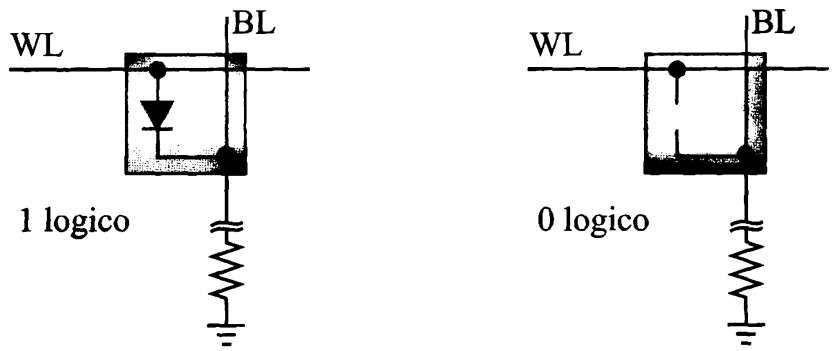


Figura 8.7 Cella di memoria ROM a diodo.

ovvero manca il diodo, la selezione della cella non potrà alterare la tensione della *bit line*, che resterà bassa per effetto della resistenza di *pull down* collegata a massa. In fase di costruzione della memoria, il diodo sarà realizzato soltanto in corrispondenza delle celle destinate a conservare un 1.

La cella di memoria, che comprende l'incrocio tra *bit line* e *Word Line* e eventualmente il diodo, è certamente molto compatta e permette di conseguire livelli molto spinti di densità di integrazione. Una limitazione della soluzione a diodo è data dal fatto che, con diodo in conduzione, la *Word Line* deve fornire la corrente di pilotaggio necessaria a caricare la capacità parassita associata alla *bit line*.

La figura 8.8 descrive la struttura delle celle ROM in tecnologia bipolare e MOS. In entrambi i casi, il controllo della *Word Line* sulla *bit line* è ottenuto mediante un transistore, evitando così connessioni dirette; le *bit line* poi sono collegate tramite una resistenza di *pull up* alla tensione di alimentazione e perciò la tensione di riposo in questo caso è alta. Quando viene selezionata una cella priva di transistore, la tensione sulla *bit line* non cambia e resta quella imposta dalla resistenza *pull up*: tale condizione corrisponde alla memorizzazione di un valore logico alto. Alla selezione invece di una cella con transistore, la tensione V_{WL} attiva il transistore, che a sua volta scarica a massa la capacità parassita associata alla *bit line*, e questo corrisponde a uno 0 memorizzato.

Tradizionalmente si associa la soluzione bipolare ad una maggiore velocità della memoria, mentre l'uso del transistore MOS garantisce un isolamento elettrico migliore tra *bit line* e *Word Line* e anche la minima dissipazione di potenza. Tuttavia i progressi della tecnologia MOS hanno ridotto di molto la penalizzazione in termini di tempo di accesso e le memorie MOS sono oggi largamente le più diffuse in tutte le applicazioni.

8.3.1 ROM programmabili (PROM)

Le applicazioni che richiedono memorie a sola lettura, ovvero la disponibilità di un codice particolare da usare più volte ma da non modificare mai, sono molteplici, ma poche di queste implicano volumi di produzione così elevati da giustificare l'approccio realizzativo proprio delle tecnologie ROM bipolare o MOS descritte nel paragrafo precedente.

È quindi estremamente utile la possibilità di lasciare la definizione del contenuto di

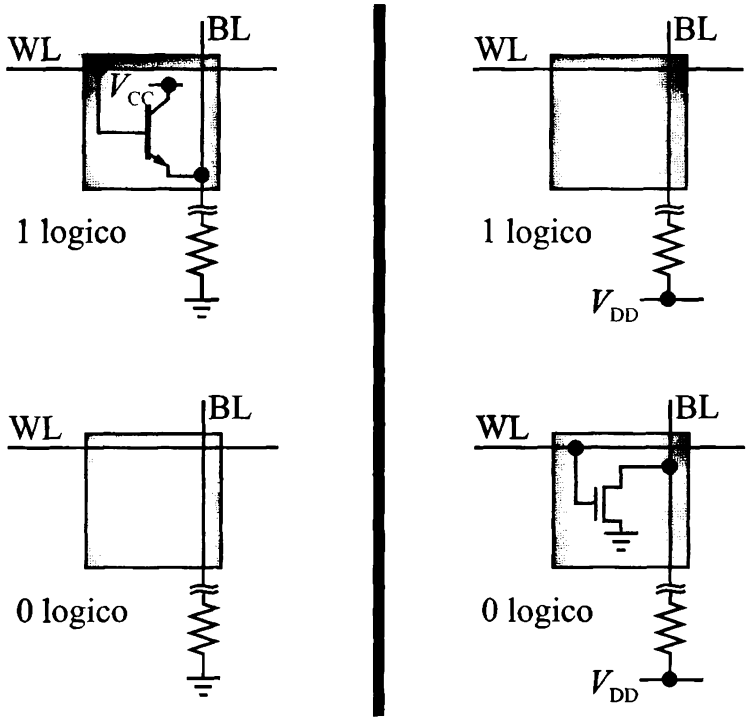


Figura 8.8 Celle di memoria in tecnologia bipolare e MOS.

memoria all'utilizzatore del dispositivo: l'idea è quella di produrre memorie identiche e inizialmente prive di contenuto di informazione e di affidare all'utilizzatore il compito di scrivere (una sola volta) questi dispositivi, personalizzandoli per una specifica applicazione. Tale operazione di scrittura, effettuata una sola volta, prende il nome di programmazione e i dispositivi di memoria di questo tipo sono noti come PROM ("Programmable Read Only Memory").

Questa possibilità di personalizzazione può essere ottenuta facilmente realizzando matrici di celle tutte identiche, ciascuna dotata di un elemento attivo (bipolare o MOS) collocato tra *bit line* e *Word Line*; in serie all'elemento attivo viene inserito un fusibile, che può essere selettivamente bruciato in fase di programmazione, rendendo inutilizzabile l'elemento attivo corrispondente e trasformando quindi l'informazione della cella da 0 a 1 logico.

Ovviamente anche un solo errore nel processo di programmazione rende il componente del tutto inutilizzabile: questa non è però da ritenersi una limitazione, in quanto la programmazione si avvale di strumenti automatici, detti programmatori, che sequenzialmente e selettivamente applicano alle possibili coppie *Word Line - bit line* una tensione di programmazione, più elevata delle normali tensioni di esercizio della memoria e sufficiente per bruciare i fusibili; le informazioni fornite al programmatore per questa operazione sono prodotte mediante strumenti CAD appositi che permet-

tono di verificare la correttezza della programmazione mediante simulazione prima di effettuarla.

Una limitazione più significativa delle PROM è invece rappresentata dall'impossibilità di collaudare questi componenti: verificare il corretto funzionamento di una PROM prima di immetterla sul mercato significherebbe sottoporla ad una fase di programmazione, ma questo renderebbe inutilizzabile il componente.

I fusibili possono essere realizzati mediante tre principali tecnologie.

- ▷ Una linea metallica viene assottigliata per un tratto, provocando al passaggio di corrente durante la programmazione un aumento locale della temperatura per effetto Joule; la conseguente evaporazione del metallo interrompe il collegamento.
- ▷ In alternativa al metallo, si possono usare sottili tratti di linea in silicio policristallino.
- ▷ Il fusibile può anche essere formato da una giunzione *pn*, che durante la programmazione viene prima portata in "breakdown" e poi cortocircuitata per azione dell'Al che diffonde nel Si a temperatura elevata.

Infine, molto diffusa è anche la tecnologie degli "anti-fusibili": questi componenti si comportano all'origine come circuiti aperti, ma dopo la programmazione risultano essere dei discreti corto circuiti; una PROM con anti-fusibili richiederà quindi una mappa di programmazione che è l'esatto inverso bit a bit di quella utilizzata per una PROM a fusibili.

8.4 Memorie non volatili riscrivibili

Le PROM sono programmabili una volta soltanto; nelle memorie a lettura-scrittura non volatili (*Non Volatile Read-Write Memories, NVRWM*), la possibilità di riprogrammare gli elementi di memoria è garantita dalla loro conduttività, che può essere alterata in modo non distruttivo.

L'elemento chiave è un transistore avente una tensione di soglia, V_t, modulabile, con valori diversi nei due stati possibili di "0" e "1".

Per un transistore n-MOS, V_t dipende dalla quantità di carica presente tra l'elettrodo di gate ed il canale, secondo l'equazione

$$V_t = K - Q/C_{ox}$$

dove C_{ox} è la capacità dell'elettrodo di gate e Q è la carica elettrica associata, mentre K è una grandezza indipendente da Q.

Il dispositivo "Floating Gate" (FG) permette di immagazzinare gli elettroni in uno strato conduttore, circondato da un dielettrico, posto tra l'elettrodo di gate e il canale; la struttura di questo tipo di dispositivo è data nella figura 8.9, dove è anche rappresentato il comportamento in termini di caratteristica corrente-tensione, al variare della carica Q. Qualitativamente, se FG è scarico ($Q = 0$), la tensione di soglia assume un valore inferiore e simile a quello di un normale transistore MOS: in queste condizioni, il dispositivo può essere pilotato mediante le normali tensioni di esercizio come interruttore aperto o chiuso. Se invece FG è carico ($Q < 0$), la tensione di soglia cresce a valori nettamente superiori a quello di riposo e il dispositivo risulta interdetto con qualunque tensione di normale esercizio applicata al gate.

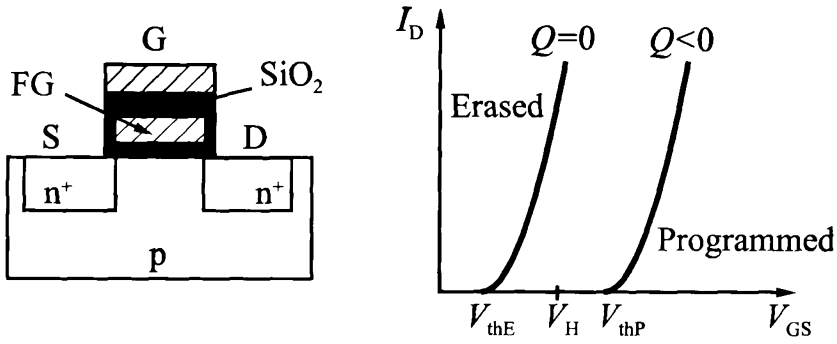


Figura 8.9 Struttura e comportamento del transistore MOSFET *floating gate*.

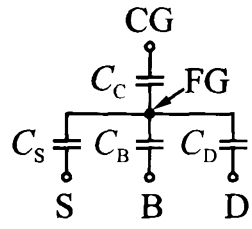


Figura 8.10 Modello capacitivo del gate isolato.

Un semplice modello del comportamento del transistore a gate isolato si può ottenere a partire dallo schema in figura 8.10, dove sono indicati per il nodo centrale (FG) i vari contributi capacitivi del gate isolato verso i terminali di gate (G), drain (D), source (S) e substrato (B). Nel modello,

- ▷ C_C è la capacità tra i due gate
- ▷ C_S, C_B e C_D sono le capacità di source, substrato e drain

Indicando con Q la carica immagazzinata nel gate isolato e con C_T la capacità totale, si ha

$$\begin{aligned} Q &= C_C(V_F - V_C) + C_S(V_F - V_S) + \\ &\quad + C_B(V_F - V_B) + C_D(V_F - V_D) \end{aligned}$$

La tensione al gate isolato è quindi data da

$$V_F = \frac{C_C}{C_T}V_C + \frac{C_S}{C_T}V_S + \frac{C_D}{C_T}V_D + \frac{C_B}{C_T}V_B + \frac{Q}{C_T}$$

Ponendo i terminali di source e substrato al potenziale di massa, si ha

$$V_{FS} = \frac{C_C}{C_T}V_{CS} + \frac{C_D}{C_T}V_{DS} + \frac{Q}{C_T}$$

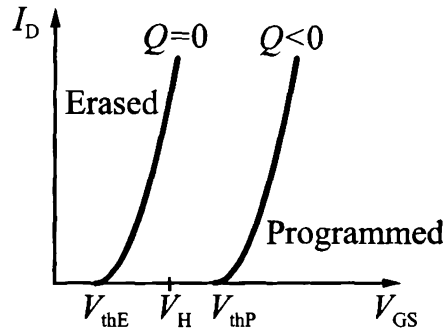


Figura 8.11 Soglia programmabile di un MOS a gate isolato.

$$= \alpha_C(V_{CS} + fV_{DS} + \frac{Q}{C_C})$$

con $\alpha_C = C_C/C_T$ e $f = C_D/C_C$ fattori di accoppiamento.

Indichiamo con V_{TFS} il potenziale da applicare al gate isolato con $V_{DS} = 0$ per ottenere l'inversione di popolazione; a tale potenziale corrisponde una tensione V_{TCS} applicata al terminale di controllo:

$$V_{TCS} = \frac{1}{\alpha_C}V_{TFS} - \frac{Q}{C_C}$$

Le due tensioni hanno significato molto diverso: V_{TFS} dipende dalla tecnologia, mentre V_{TCS} dipende anche dalla carica Q nel gate isolato.

Per una cella cancellata, nella quale sia $Q = 0$, la tensione V_{TCS} si può esprimere come (figura 8.11)

$$V_{TE} = \frac{1}{\alpha_C}V_{TFS}$$

La cella programmata presenta invece una Q non nulla, e la V_{TCS} è data da

$$V_{TP} = \frac{1}{\alpha_C}V_{TFS} - \frac{Q}{C_C}$$

Nella figura 8.12, sono mostrate le condizioni di cella programmata a "1" e "0" logico, in termini di diagramma a bande nella direzione trasversale.

Ci sono più meccanismi di iniezione di carica nel gate isolato e i principali sono i seguenti:

1. **Iniezione di elettroni caldi (HEI)**. Con un campo elettrico laterale superiore a $\approx$ 100 kV/cm, alcuni elettroni raggiungono energia superiore alla barriera di potenziale ossido-silicio: tali elettroni prendono il nome di elettroni caldi. Quando la ionizzazione da impatto al drain genera coppie elettroni-lacune ad alta energia, gli

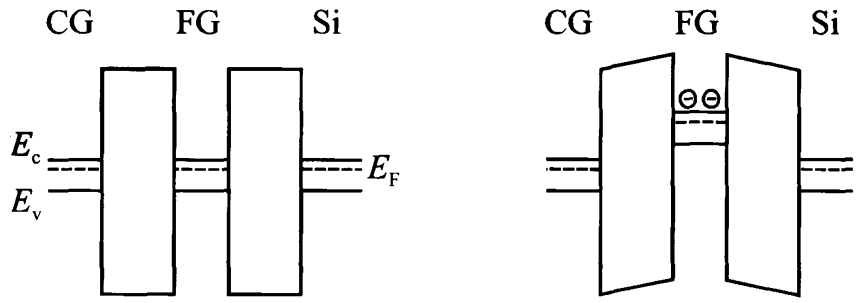


Figura 8.12 Diagramma a bande del dispositivo a gate isolato, nelle condizioni di programmazione a "0" e "1" logico.

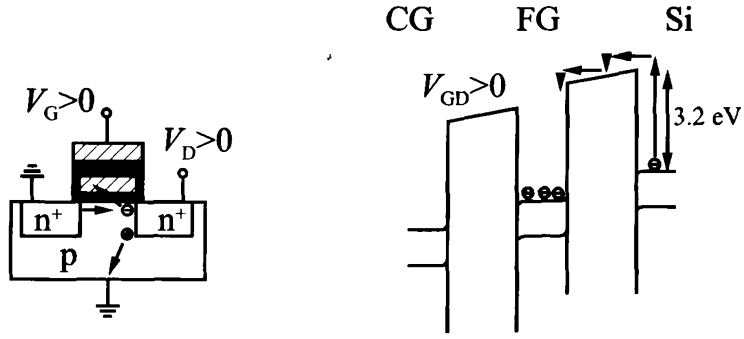


Figura 8.13 Iniezione da elettroni caldi.

elettroni sono accelerati verso il gate isolato e in parte riescono a raggiungerlo, rimanendovi intrappolati (figura 8.13). L'attuazione di questo meccanismo di iniezione richiede quindi l'applicazione simultanea di una tensione laterale molto elevata e di una tensione trasversale. In condizioni ideali, la carica elettrica trasferita nel gate isolato può rimanervi intrappolata per tempi misurabili in decine di anni, anche se le radiazioni incidenti hanno l'effetto di generare una corrente di scarica attraverso l'ossido equindi di cancellare il dispositivo.

2. **Tunneling Fowler-Nordheim (FN)**. Si tratta in questo caso di un effetto quantistico, in base al quale alcuni elettroni possono attraversare la barriera di potenziale localizzata tra canale e gate isolato; a tale fenomeno, denominato *tunneling*, è associata una probabilità dipendente dalle caratteristiche della barriera (in particolare da Φ_B, altezza della barriera) e dal campo elettrico nell'ossido (figura 8.14). Per facilitare il fenomeno di tunneling, il gate isolato viene esteso al di sopra del drain e nella zona di sovrapposizione lo spessore dell'ossido è ridotto a valori inferiori a 10 nm.

I principali tipi di memoria a semiconduttore che sfruttano i meccanismi descrit-

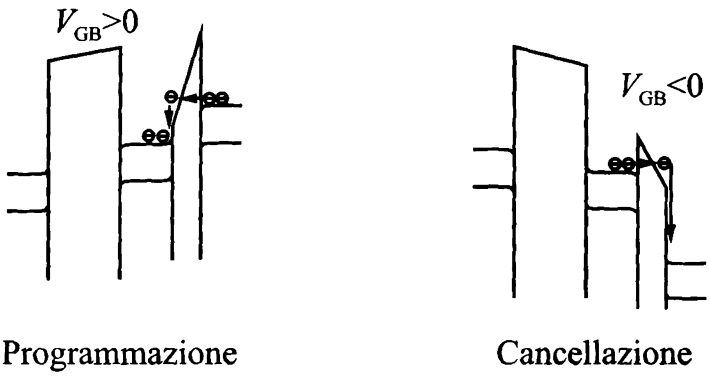


Figura 8.14 Tunneling di tipo Fowler-Nordheim.

Memoria	Elettronica di consumo	Autoveicolo	Computer e telecomunicazioni
EPROM	*giochi, Set top box*	*controllo motore*	*driver per hard disk, fotocopiatrici, fax*
Flash	*Set top box, PDA, lettori e dispositivi multimediali*	*power train, ABS navigazione, ABS*	*driver per hard disk e CDROM, PC Bios cellulari*
EEPROM	*audio e video*	*tutti i controlli*	*schede grafiche, stamponti*

Tabella 8.4 Esempi di applicazione delle memorie non volatili.

ti di iniezione di carica sono note come EPROM, EEPROM e Flash. Queste memorie, descritte nei paragrafi seguenti, hanno numerose applicazioni, riassunte nella tabella 8.4.

8.4.1 La cella EPROM

Il meccanismo di moltiplicazione a valanga degli elettroni al drain è sfruttato nelle memorie EPROM (*Erasable Programmable ROM*) per la programmazione delle celle di memoria.

Il più noto dispositivo di questo tipo porta il nome **FAMOS**, che sta per *Floating-gate Avalanche-injection MOS*. In questo dispositivo, la programmazione della cella avviene elettricamente per moltiplicazione a valanga degli elettroni del canale, che vengono poi iniettati nel gate isolato.

Il meccanismo di iniezione dei portatori non è però invertibile. Quindi, per cancellarne il contenuto, il modulo EPROM deve essere rimosso dal resto del circuito per poi venire esposto a radiazione ultravioletta, il cui effetto è di consentire la ricombinazione degli elettroni presenti nel gate isolato.

L'operazione di scrittura avviene selettivamente, cella per cella e richiede, per tempi dell'ordine dei μs, l'applicazione di tensioni elevate (tipicamente nell'intervallo 10-20 V) ai morsetti di gate e drain del dispositivo, attraverso le *word* e *bit line*.

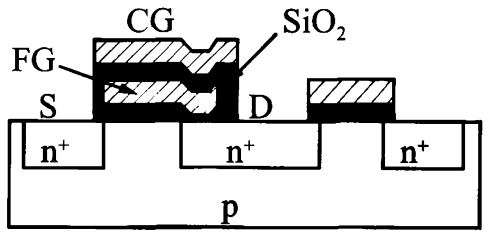


Figura 8.15 Cella di una memoria EEPROM.

La cancellazione è invece effettuata indiscritinatamente per l'intero chip, che deve quindi essere contenuto in un package dotato di una finestra di quarzo, per permettere l'esposizione alla radiazione ultra-violetta dell'intera regione di silicio occupata dalle celle EPROM. L'esposizione può durare svariati secondi o minuti, in funzione della intensità della sorgente UV. Il tempo di accesso in lettura è invece dell'ordine di 100 ns.

L'affidabilità di una memoria EPROM, in seguito ai danni reticolari provocati dalla radiazione UV, è tipicamente garantita per un numero massimo di riprogrammazioni (da 100 a 1000).

8.4.2 Cella EEPROM

La necessità di rimuovere il componente di memoria dal proprio circuito e di utilizzare radiazioni UV per la cancellazione è superata nei dispositivi EEPROM, *Electrically Erasable PROM*, la cui cella è programmata e cancellata per via elettrica.

Il funzionamento dell'elemento di memoria è basato sull'effetto tunnel Fowler-Nordheim, che è bidirezionale, ovvero può essere sfruttato anche per rimuovere carica elettrica dal gate isolato. In questi dispositivi, lo spessore dell'ossido dal lato del drain è ridotto da 100 a meno di 10 nm (figura 8.15), allo scopo di massimizzare l'efficienza dell'effetto tunnel quantistico. La figura 8.16 illustra qualitativamente il comportamento degli elettroni durante l'uso della memoria. La prima immagine in alto rappresenta la struttura della cella EEPROM letta in senso trasversale, a partire dal gate di controllo, che è separato dal gate isolato (FG) mediante un primo strato di ossido; un secondo strato di ossido separa il FG dal silicio di substrato. Le immagini successive in figura 8.16 mostrano invece l'andamento del diagramma a bande nelle tre condizioni di programmazione, memoria e cancellazione.

La reversibilità dell'operazione di tunneling pone però problemi di selettività, nel senso che risulta impossibile agire sia in scrittura che in cancellazione su una cella indipendentemente da tutte le altre. Al fine di comprendere il problema, si consideri la matrice di quattro celle in figura 8.17.

Si consideri la programmazione della cella di posizione *(1,1)*: questa operazione richiede l'applicazione di una tensione alta alla *Word Line* WL1 e bassa alla *Bit Line* BL1, al fine di indurre iniezione di carica verso il gate isolato della cella *(1,1)*:

- $WL_1 = V_H$
- $BL_1 = 0$

Per impedire la programmazione indesiderata della cella di posizione *(1,2)*, è necessario

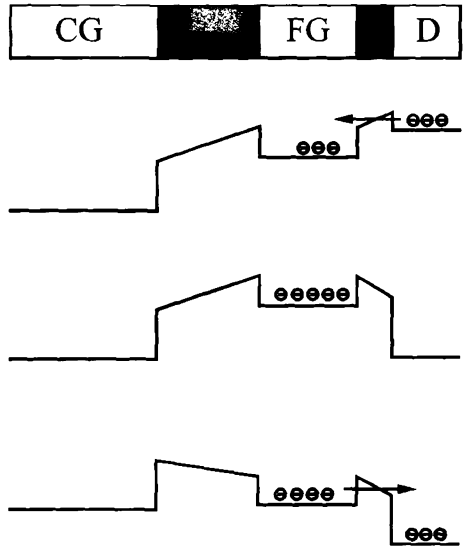


Figura 8.16 Struttura e diagramma a bande della cella EEPROM.

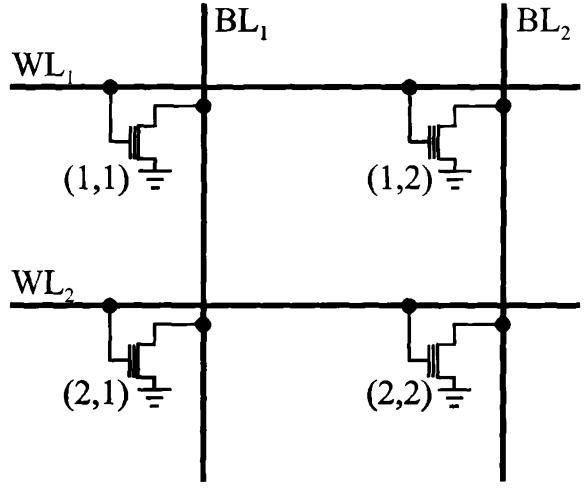


Figura 8.17 Porzione di memoria con quattro celle EEPROM, senza transistore di selezione.

che anche la tensione applicata alla *Bit Line* BL2 sia alta:

- $BL_2 = V_H$

Con questo valore sulla *Bit Line* BL_2, se $WL_2 = 1$, allora la cella *(2,1)* risulterà anch'essa programmata, mentre se $WL_2 = 0$, la stessa cella di posizione *(2,2)* sarà

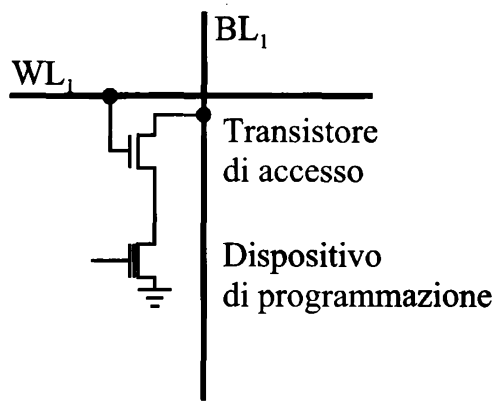


Figura 8.18 Cella di memoria EEPROM, con transistore di accesso e dispositivo di memoria.

cancellata.

Non è quindi possibile evitare un'operazione indesiderata sulla cella *(2,2)*, mentre si effettua la programmazione della *(1,1)*. Un altro problema associato all'uso del tunneling in entrambe le operazioni di programmazione e cancellazione è la difficoltà di controllare con precisione la carica trasferita. Mentre nel caso delle EPROM la cancellazione mediante esposizione a radiazioni UV garantisce il raggiungimento di una carica netta nulla nel gate isolato, questo risultato non è automatico nel caso della cancellazione EEPROM, a meno di usare complessi circuiti di misura della carica. L'esito di una imperfetta cancellazione della carica nel gate isolato può essere un valore di tensione di soglia scorretto e quindi in definitiva un errore di lettura. Al fine di risolvere entrambi i problemi, nella EEPROM si aggiunge ad ogni cella un secondo transistore MOS in serie, che operi semplicemente come dispositivo di selezione. Come indicato in figura 8.18, il transistore MOS riceve sul gate il segnale di selezione della cella: se questo è basso, la cella rimane isolata, indipendentemente dalla carica contenuta nel gate isolato del secondo dispositivo. Quando invece la cella è selezionata, una carica negativa nel gate isolato garantisce ancora l'isolamento della cella, mentre una carica nulla o anche positiva permette comunque di connettere al potenziale di massa la *Bit Line*.

Tutte le implementazioni basate sull'effetto tunnel richiedono la presenza di un ulteriore transistore per effettuare la selezione del bit e questo aumenta la dimensione della cella, riducendo la densità della memoria.

Dal confronto tra memorie EPROM e EEPROM risulta

- ▷ l'area della cella EEPROM è 3-4 volte superiore rispetto a quella della struttura FAMOS;
- ▷ la procedura di programmazione è lenta (circa 8 msec/word) e complessa (scrittura dell'intero chip seguita da una cancellazione selettiva dei bit).
- ▷ il numero massimo di riprogrammazioni è più elevato per le memorie EEPROM (10^5).

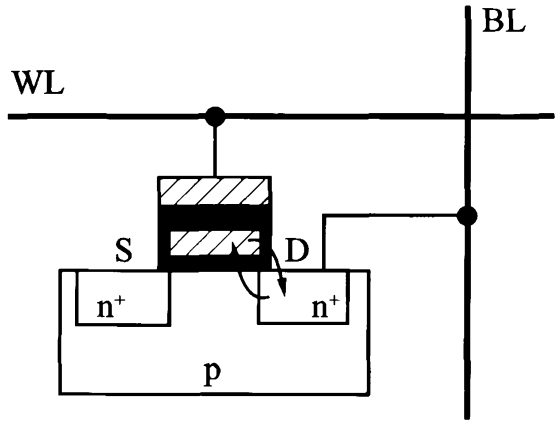


Figura 8.19 Programmazione e cancellazione di una cella Flash.

8.5 La cella Flash

Le memorie Flash sono state presentate la prima volta nel 1984 dalla Toshiba e commercializzate dalla Intel nel 1988. Da allora hanno subito uno sviluppo tecnologico molto rapido, che ne ha accresciuto enormemente la diffusione.

La cella Flash combina la versatilità della EEPROM con la compattezza della EPROM. Come nel caso delle EEPROM, le operazioni avvengono "on-system", cioè senza la necessità di rimuovere il dispositivo dal sistema che lo ospita, e per via elettrica, senza uso di radiazioni UV. Per contro non è richiesto il transistore di selezione e la densità è quindi prossima a quella delle EPROM. Per superare il problema della selettività, la cancellazione non è effettuata cella per cella, ma a gruppi di celle, detti settori o pagine. Come si vedrà nel seguito, tale caratteristica richiede l'intervento di un microcontrollore integrato nella memoria, che gestisce sequenze ripetute di verifica della tensione di soglia e applicazione di impulsi di cancellazione. La complessità aggiuntiva che deriva dal microcontrollore è comunque contenuta e mantiene la densità della memoria ben al di sopra di quanto ottenibile nel caso dele EEPROM.

La programmazione avviene o per effetto tunnel o per moltiplicazione a valanga, mentre la cancellazione avviene sempre per effetto tunnel (figura 8.19). Per esempio, nella Flash note come ETOX (*EPROM Tunnel Oxide*)

- ▷ la programmazione avviene per moltiplicazione a valanga dei portatori di carica (elettroni caldi) presenti nel canale conduttivo e richiede $V_{DS} = 6$ V, $V_{GS} = 12$ V
- ▷ la cancellazione avviene invece ricorrendo all'effetto Fowler-Nordheim: gli elettroni intrappolati nel FG vengono catturati dal source n^+ ($V_{SG} = 12$ V).

La non volatilità della cella dipende dalla qualità del dielettrico, che deve isolare la carica elettrica immagazzinata nel FG; tuttavia la possibilità di sfruttare il *tunneling* è legata allo spessore del dielettrico, che non può quindi essere eccessivo. Si utilizzano spesso materiali diversi per realizzare i diversi strati di isolamento, per esempio

- ▷ *Tunnel oxide*, posto tra canale e FG, di spessore spessore 9-10 nm

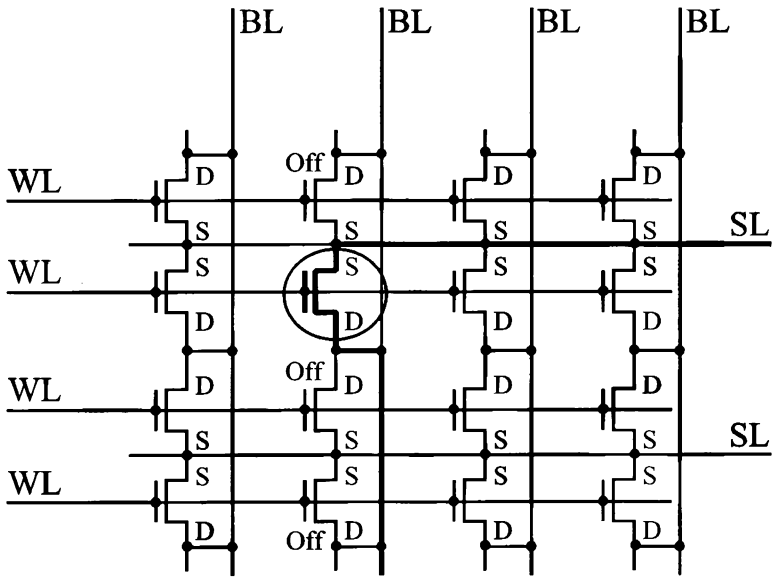


Figura 8.20 Architettura di una Flash NOR.

▷ *ONO layer* (oxide-nitride-oxide), posto tra FG e CG

Esistono due tecniche fondamentali per organizzare la struttura di una memoria flash, che differiscono principalmente per come avvengono gli accessi alle celle.

1. Nella NOR flash, le celle sono disposte a matrice, con tutti i terminali di gate delle celle appartenenti alla medesima riga connessi a una *Word Line*, tutti i terminali di drain delle celle della medesima colonna connessi a una *Bit Line* e tutti i terminali di source delle celle di un settore connessi a una linea comune ("source line").
2. Nella NAND flash, le celle sono raggruppate in catene di 8 elementi, unendo il drain di un transistore con il source di quello successivo; questa organizzazione seriale implica un accesso di tipo sequenziale e quindi piuttosto lento alle informazioni, ma in compenso la dimensione della unitaria della cella è inferiore a quella del caso NOR.

8.5.1 Architettura NOR

La struttura di tipo NOR, usata per applicazioni riferite sia a dati che a codice, è esemplificata nella figura 8.20, dove è evidenziata la connessione comune di tutti i terminali di source. La programmazione avviene a byte o a word, mentre la cancellazione è comune a un intero settore, che può avere dimensione di 64 kByte.

Per chiarire l'operazione di lettura, si immagini che sia indirizzata la cella evidenziata con un cerchietto nella figura e si assuma che tutti i transistori delle altre celle della colonna siano interdetti. Se la cella selezionata è programmata, allora tra *source line* e *Bit Line* non si può avere corrente; al contrario, se la cella non è programmata, sulla

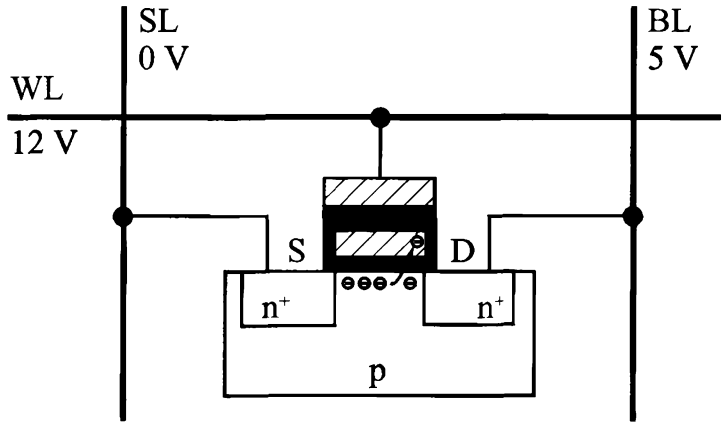


Figura 8.21 Programmazione di una Flash NOR.

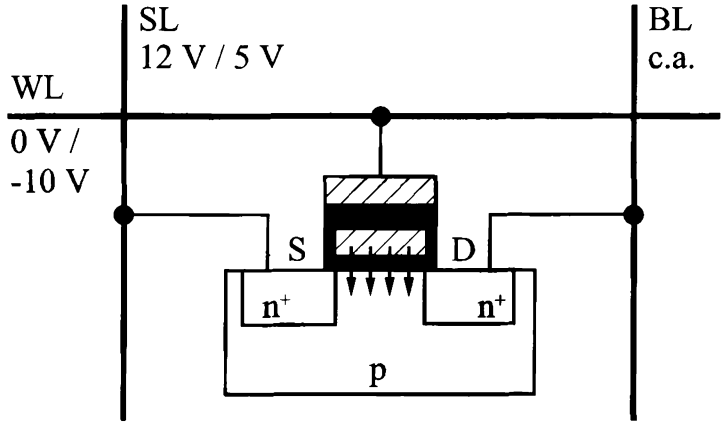


Figura 8.22 Cancellazione di una flash NOR.

Bit Line si può rilevare una corrente. È quindi semplice discriminare l'informazione binaria immagazzinata sulla base della corrente rilevata.

La programmazione, anch'essa selettiva, avviene per *Fowler-Nordheim tunneling* e richiede una tensione elevata (5 o 6 V) tra drain e source, per generare elettroni caldi; è inoltre necessaria l'applicazione della tensione di programmazione (per esempio 12 V) tra gate di controllo e canale, per indurre il superamento della barriera di ossido (figura 8.21.

La cancellazione avviene invece a settori ed è simultanea per tutte le celle di un settore. L'effetto tunnel è innescato dalla tensione elevata applicata tra source (comune a più celle) e gate (connesso alla *Word Line*), come indicato in figura 8.22.

Una delle limitazioni principali delle memorie Flash è dunque la necessità di cancellare il contenuto a blocchi, mentre le operazioni di lettura e scrittura possono avvenire

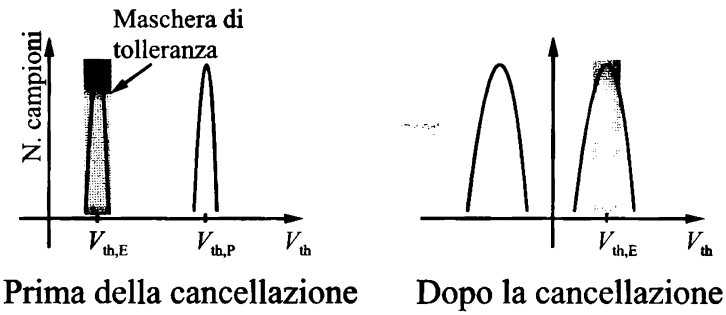


Figura 8.23 Distribuzione delle tensioni di soglia in una Flash prima e dopo l'operazione di cancellazione.

a byte o a word, con un indirizzamento di tipo "random", come avviene nelle normali memorie RAM trattate nel paragrafo 8.7. Occore tuttavia precisare che la cancellazione ha l'effetto di inizializzare tutte le celle del blocco a 1; ogni successiva operazione di scrittura può modificare il contenuto di una o più celle da 1 o 0 e una nuova cancellazione è richiesta soltanto se il contenuto di una cella deve essere portato da 0 a 1. Sebbene quindi il meccanismo di cancellazione a blocchi non permetta una gestione della memoria con lo stesso livello di flessibilità di una RAM, sono tipicamente possibili più operazioni di lettura e scrittura tra due cancellazioni.

Poiché prima della cancellazione alcune celle del settore possono essere programmate e altre no, l'operazione di cancellazione deve sempre essere preceduta da una fase di programmazione, in modo da agire su celle che si trovino nello stesso stato di partenza. Se questa programmazione preliminare fosse omessa, le celle non programmate risulterebbero sovra-cancellate, ovvero la carica elettrica nel gate isolato risulterebbe positiva, ad indicare che sono stati rimossi elettroni da un gate non programmato e quindi neutro.

In realtà, il valore della tensione di soglia per un gruppo di celle nominalmente identiche e nel medesimo stato logico (programmato o no) non è mai esattamente uniforme, a causa della dispersione per parametri tecnologici che controllano la tensione di soglia in un MOSFET a gate isolato: lievi differenze in termini di spessore dell'ossido o concentrazione locale di drogante determinano la non uniformità delle tensioni di soglia e questa situazione è raffigurata nella parte sinistra di figura 8.23: i valori delle tensioni di soglia di una popolazione di celle sono raccolti in due distribuzioni poste intorno ai due valori nominali che corrispondono alle condizioni di cella programmata e non programmata. Si può pensare che tali distribuzioni siano inizialmente piuttosto strette, a indicare che per ciascun dispositivo lo scostamento dai valori nominali sia molto limitato e contenuto entro limiti accettabili, evidenziati per la tensione bassa (cella non programmata) nella figura 8.23 dalla maschera rettangolare grigia. Tale maschera corrisponde alla massima dispersione tollerata per la tensione di soglia, tale cioè da non produrre, nelle operazioni di programmazione e cancellazione successive, celle sovra-cancellate. Purtroppo però le operazioni di cancellazione tendono progressivamente ad allargare le distribuzioni, imponendo l'adozione di meccanismi correttivi.

Se ipotizziamo che prima della cancellazione alcune celle siano programmate e altre no, l'effetto della cancellazione, ovvero della rimozione di elettroni da tutti i gate isolati

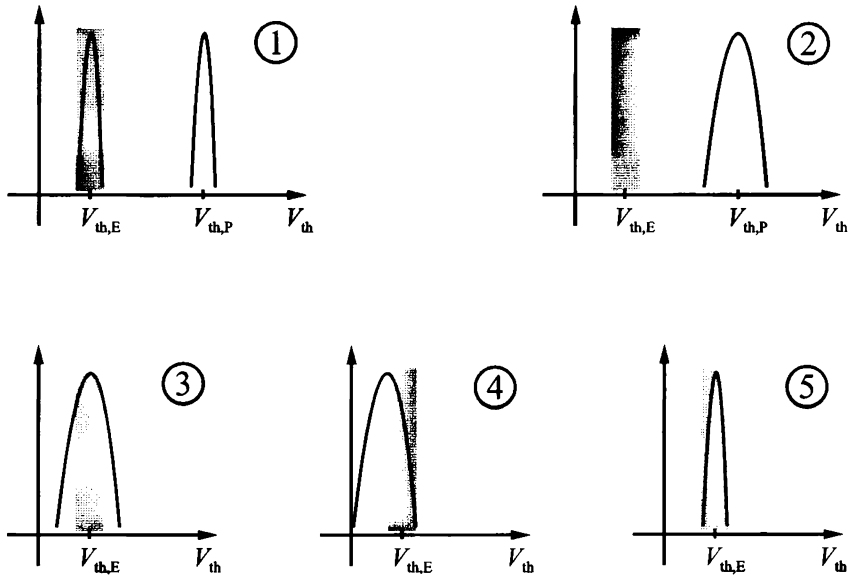


Figura 8.24 Procedura di cancellazione di una Flash.

del settore, sarà di abbassare la tensione di soglia delle celle programmate intorno al valore nominale di una cella non programmata e, nel contempo, di abbassare a valori negativi le soglie delle celle inizialmente non programmate, come indicato nella parte destra di figura 8.23. Le celle con tensione di soglia negativa non possono ovviamente essere utilizzate, in quanto non sono controllabili con le normali tensioni di esercizio della memoria. Inoltre, come ulteriore effetto della dispersione dei parametri tecnologici, le due distribuzioni ottenute dopo la cancellazione sono più ampie di quelle iniziali e possono uscire dai limiti di accettabilità.

Per evitare questo problema, si realizza l'operazione di cancellazione come procedura articolata su più fasi, indicate in figura 8.24. Inizialmente si sottopongono tutte le celle del settore da cancellare a programmazione, in modo da poter lavorare su una singola distribuzione di soglie: nella figura 8.24, si passa dalla fase 1, con due distribuzioni strette e ben distinte, alla fase 2, nella quale la distribuzionepiù ampia e si colloca intorno alla tensione di soglia superiore.

La fase successiva è di cancellazione e ha l'effetto di traslare la distribuzione di soglie iniziali intorno al valore nominale della cella non programmata (fase 3). L'ampiezza della distribuzione ottenuta è però eccessiva e non rientra nella maschera di valori consentiti. Occorre quindi correggere questa distribuzione, riducendone la deviazione standard fino a rientrare entro i limiti di sicurezza; questo obiettivo è tipicamente conseguito ricorrendo a una sequenza di impulsi alternati di programmazione e cancellazione. Ad ogni impulso di cancellazione, della durata compresa tra 1 e 10 μs, le tensioni di soglia tendono a essere ridotte, traslando verso sinistra la distribuzione (fase 4); i successivi impulsi di programmazione agiscono selettivamente sulle celle dotate di soglia troppo bassa, restringendo così la distribuzione (fase 5).

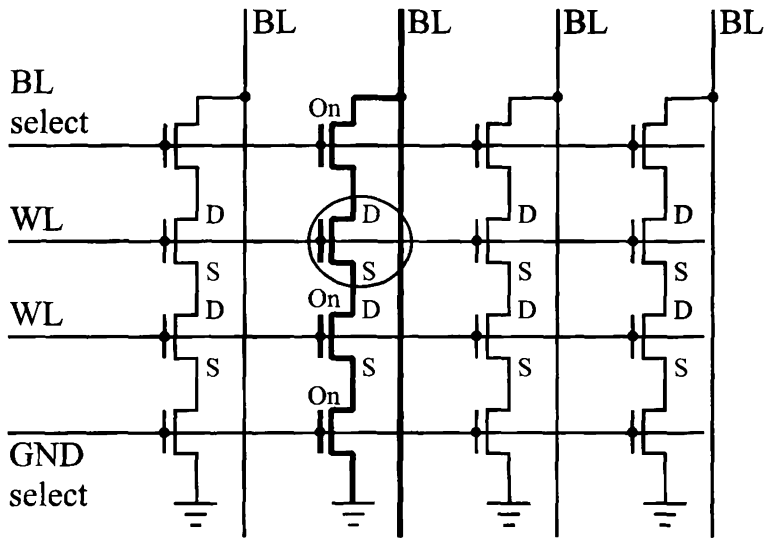


Figura 8.25 Architettura NAND Flash.

Questa sequenza di operazioni è gestita da un microcontrollore interno, che esegue l'algoritmo di correzione in modo adattativo, ovvero misurando le tensioni di soglia prima di sottoporle a correzione.

Poiché la procedura di cancellazione è piuttosto complessa e richiede l'accesso sequenziale e ripetuto alle celle del settore cancellato per poter controllare le tensioni di soglia, la durata dell'operazione di cancellazione è tipicamente considerevole: tale durata dipende fortemente dalle dimensioni del settore, ma difficilmente è inferiore ai 100 ms. Ne consegue che le memorie Flash con architettura NOR sono veloci nell'accesso in lettura, ma piuttosto lente in scrittura.

8.5.2 Architettura NAND

Un esempio di memoria Flash con architettura NAND è dato in figura 8.25. Le celle sono disposte secondo una compatta struttura serie, senza le connessioni verso le *source line* e *Bit Line* che sono presenti nell'architettura NOR per ogni singolo transistore. Le sequenza di transistori a gate isolato connessi in serie sono tipicamente di lunghezza 8, 16 o 32; in cima e in fondo a questa catena di dispositivi sono collocati due transistori di selezione, che abilitano il collegamento rispettivamente con la *Bit Line* e la *source line*. Nonostante la presenza di questi transistori aggiuntivi, la struttura seriale delle Flash NAND si ripercuote favorevolmente sulla densità di integrazione, superiore nel caso NAND rispetto a quello NOR tipicamente del 40%.

Il meccanismo fisico di *tunneling*, oltre che per la programmazione, è anche sfruttato per la cancellazione; nell'architettura NAND, non essendo disponibile un contatto di source per ogni transistore, il *tunneling* è attivato tra gate e substrato (figura 8.26), applicando una tensione elevata al substrato e portando a massa la *Word Line*, mentre i terminali di source e drain sono lasciati isolati. Questa condizione di cancellazione è

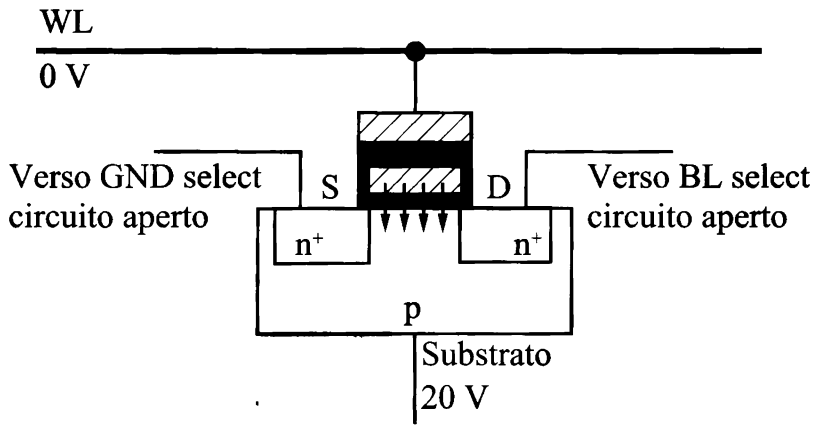


Figura 8.26 Cancellazione di una cellaFlash NAND.

estesa all'intero settore selezionato, mentre nei settori non selezionati per la cancellazione, il substrato continua ad essere polarizzato a tensione elevata, ma tutti gli altri terminali sono lasciati isolati. Come effetto della cancellazione, tutte le celle del settore risultano dotate di una tensione di soglia negativa e si comportano come dispositivi a svuotamento, ovvero normalmente accesi e in grado di spegnersi solo con una tensione negativa applicata al terminale di gate.

Le operazioni di programmazione avvengono mediante *tunneling FN* tra gate e canale. Per programmare una cella a 1, la tensione alta di programmazione (15 - 20 V) è applicata alla *Word Line* della cella da programmare e la *Bit Line* viene portata a massa, attivando il corrispondente transistore di selezione: gli elettroni si spostano per effetto tunnel nel gate isolato e provocano l'aumento della tensione di soglia, che invece resta negativa per le celle da programmare a 0. Tutti gli altri dispositivi della colonna devono essere accesi durante la programmazione. La programmazione è effettuata dopo la cancellazione e perciò la tensione di soglia iniziale della cella da programmare è negativa. Un singolo impulso di programmazione potrebbe non essere sufficiente a dare alla tensione di soglia un adeguato margine di sicurezza al di sopra dello 0 (solitamente si richiedono valori dell'ordine di 0,7-0,8 V) e quindi il controllo interno alla memoria esegue un ciclo di programmazione e lettura sulle celle da programmare, fino a che tutte le tensioni di soglia sono soddisfacenti. Nonostante ciò, l'operazione di programmazione è piuttosto rapida e richiede tempi tipici dell'ordine di 200-400 ns.

La lettura avviene connettendo l'estremità inferiore (*source line*) della catena di celle a massa, applicando un potenziale nullo alla *Word Line* della cella selezionata e misurando la corrente sulla *Bit Line*: nell'ipotesi che tutte le altre celle della colonna siano forzate nello stato ON, la corrente nella *Bit Line* risulterà governata dalla cella selezionata. In altri termini, occorre che le *Word Line* siano governate dal decoder di riga in logica negata: tutte le celle non selezionate per l'accesso ricevono una tensione alta sulla *Word Line* e quindi si portano nello stato attivo, mentre la tensione di gate per la cella selezionata è nulla. La lettura di uno 0 logico, corrisponde al caso in cui la tensione di soglia della cella selezionata sia negativa: in questa situazione, la cella selezionata è comunque attiva e può quindi scorrere corrente tra *Bit Line* e *source*

NOR	NAND
Capacità: capacità medio alta (dai kbit ai Gbit)	**Capacità:** capacità alta (attualemte fino a 32 Gbit)
Accesso: lettura veloce (50-100 ns) scrittura lenta (10 μs)	**Accesso:** lettura lenta (1 μs) scrittura veloce (1 μs)
Applicazioni: codice e dati	**Applicazioni:** dati
Affidabilità: migliore affidabilità (oltre 10^5 letture/scritture)	**Affidabilità:** affidabilità peggiore (oltre 10^4 letture/scritture)
Dimensioni: cella più grande	**Dimensioni:** cella più piccola (-40%)

Tabella 8.5 Confronto fra architetture NOR e NAND.

line, perché tutti i transistori della catena sono in conduzione. Quando invece la cella selezionata memorizza un 1 logico, la tensione di soglia è positiva: ne consegue che il transistore risulta in interdizione e non si può avere corrente tra *Bit Line* e *source line*.

Un confronto sintetico tra le architetture NOR e NAND è dato in tabella 8.5.

Tutti i tipi di dispositivi Flash sono pensati per capacità elevate; tuttavia le memorie di tipo NOR, meno compatte delle NAND e a indirizzamento "random", risultano molto convenienti per l'esecuzione di codice e meno per la memorizzazione di grandi quantità di dati. Per contro, le memorie NAND sono caratterizzate da un minore costo di produzione per bit e sono quindi la scelta migliore per applicazioni *mass storage*, come le PC e memory cards oggi disponibili in varie forme e dimensioni.

Per quanto riguarda i tempi di accesso in lettura e scrittura, questi sono determinati sia dall'architettura che dalla dimensione dei settori: settori più grandi implicano tempi di cancellazione più lunghi, in quanto occorre ripetere le operazioni iterative di lettura, programmazione e cancellazione su un numero maggiore di celle. Generalmente le architetture NAND utilizzano settori più piccoli e quindi risultano caratterizzate da tempi di accesso in scrittura più brevi rispetto alle memorie NOR. Per contro, la struttura ad accesso seriale delle Flash NAND comporta tempi di accesso in lettura più elevati rispetto alle NOR. Queste differenze rendono più adatte le memorie NAND in applicazioni di gestione dati, nelle quali letture e scritture avvengano con frequenza confrontabile, mentre le Flash NOR trovano preferibilmente uso in applicazioni dove la lettura sia l'operazione decisamente più frequente, come la lettura di codice da eseguire, per esempio il fimware di un Set top box o il BIOS di un PC.

Infine, l'affidabilità, misurata in termini di numero tollerato di letture/scritture prima di avere un deterioramento inaccettabile del funzionamento della memoria, è superiore nei dispositivi di tipo NOR, in quanto le tensioni di programmazione sono generalmente inferiori a quelle usate dalla Flash NAND. Per contro, le Flash NOR utilizzano l'iniezione da elettroni caldi per la programmazione e questo implica la disponibilità di

correnti ben superiori a quelle tipiche delle architetture NAND.

Un aspetto interessante delle memorie Flash è legato ai meccanismi di protezione e correzione dell'informazione. A causa dell'elevato livello di integrazione di questi dispositivi e delle numerose sorgenti interne di disturbo e invecchiamento, è indispensabile dotare le memorie Flash di meccanismi di correzione e rilevazione dell'errore. A questo scopo, l'uso delle Flash, specialmente nel caso delle NAND, è associato a codici a correzione di errore, che, aggiungendo ridondanza all'informazione memorizzata, permettono di correggere automaticamente alla lettura fino a un certo numero di errori per blocco. In aggiunta, i controlli di parità permettono di identificare (ma non correggere) ulteriori errori e sono sfruttati per poter annotare internamente alla memoria stessa i blocchi nei quali tali errori non correggibili si sono verificati: i blocchi così identificati sono esclusi dalle successive operazioni di accesso e possono essere rimpiazzati da blocchi di ridondanza o semplicemente inducono una progressiva riduzione della capacità del dispositivo. Questa tecnica di gestione dei blocchi prende il nome di "bad block management". Le memorie NAND spesso prevedono locazioni apposite (di solito il primo blocco del dispositivo, o blocco 0) a elevata affidabilità destinate a ospitare la tabella dei blocchi funzionanti e di quelli guasti.

8.5.3 Affidabilità delle memorie Flash

I principali effetti che riducono l'affidabilità delle celle Flash sono:

- ▷ la distribuzione ampia e asimmetrica delle tensioni di soglia che si ottiene dopo la cancellazione per via elettrica ("endurance" della cella flash)
- ▷ i disturbi generati dalla programmazione di una cella sulle celle della stessa *Bit Line* o *Word Line*, in termini di stress elettrico
- ▷ la perdita di carica dal gate isolato, perdita che può alterare l'informazione immagazzinata (*data retention*); in genere si richiedono 10 anni
- ▷ la riduzione della tensione di soglia indotta dalla generazione di trappole e stati superficiali ad ogni operazione di programmazione e cancellazione.

Ad ogni ciclo di programmazione e cancellazione, le tensioni di soglia V_{TE} e V_{TP} del dispositivo programmato e non tendono progressivamente ad avvicinarsi, con la diminuzione di V_{TP} e l'aumento di V_{TE}. Questo fenomeno è legato alla creazione di trappole nello strato di ossido e può essere compensato mediante cicli aggiuntivi: per esempio, le celle che a seguito di una cancellazione assumano una tensione troppo bassa (si parla in questo caso di "over erasing"), possono essere sottoposte a un ciclo di riprogrammazione, in modo da ottenere una tensione di soglia finale compresa entro un intervallo accettabile. Questa soluzione tuttavia va a discapito dei tempi di cancellazione e programmazione, che tendono a crescere velocemente mano a mano che con l'invecchiamento del dispositivo l'alterazione delle tensioni di soglia diventa più evidente.

La programmazione di una cella, richiede l'applicazione di tensioni elevate al gate di controllo. Con riferimento all'architettura NOR di figura 8.20, si può notare come questa tensione sia applicata all'intera riga di celle connesse alla medesima WL. Si parla in questo caso di "gate stress", perché queste celle subiscono uno stress elettrico il cui effetto nel tempo è un progressivo innalzamento della tensione di soglia V_{TE}, come indicato nella figura 8.27. figura 8.25.

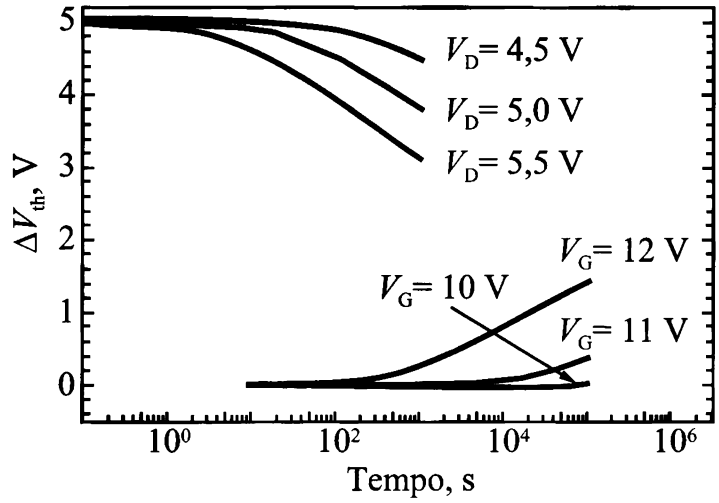


Figura 8.27 Effetti dei disturbi di programmazione.

Analogamente, la propagazione della tensione di programmazione lungo le *Bit Line*, induce un "drain stress", che tende ad abbassare la tensione di soglia delle celle programmate a "1" logico.

Per valutare la "data retention" di un dispositivo, si può valutare la perdita di carica dal gate isolato. Supponiamo che $V_G = 4$ V e si debba considerare perso il dato di una cella la cui tensione sia scesa al di sotto di 1 V. Se la capacità del gate è $C_G = 0{,}76$ fF, la carica elettrica del transistore programmato risulta $Q_P = -V_G \cdot C_G = -3$ fC, equivalente a circa $1{,}9 \cdot 10^4$ elettroni. La diminuzione di tensione da 4 V a 1 V equivale alla perdita del 75% degli elettroni, ovvero $Q_{leak} = 0{,}75$ fC, pari a $1{,}2 \cdot 10^4$ elettroni. Per garantire una "data retention" di 10 anni, la corrente di perdita deve quindi essere inferiore a $2{,}4 \cdot 10^{-24}$ A, cioè la perdita deve essere non superiore a 3,3 elettroni al giorno. Questo esempio dimostra la difficoltà tecnologica legata alla protezione dell'informazione per tempi lunghi. Poiché il numero di elettroni conservati in una cella è funzione della capacità, proporzionale alle dimensioni, e della tensione, la perdita di carica dal gate isolato introduce di fatto una limitazione allo scalamento sia delle dimensioni dei dispositivi, sia delle tensioni in gioco.

8.5.4 Circuiti di programmazione

Le tensioni elevate richieste per le operazioni di programmazione e cancellazione vengono generate all'interno del dispositivo per mezzo di circuiti detti pompe di carica che, alimentati dalla tensione di alimentazione disponibile, generano una tensione continua di valore superiore a quella di alimentazione. Questi circuiti sono composti da capacità e interruttori, connessi in modo da spostare iterativamente carica da una capacità all'altra.

Un esempio di circuito pompa di carica è dato in figura 8.28. Il circuito è composto da una successione di diodi (N in generale), che agiscono come interruttori, e da altrettante capacità, alternativamente connesse a una coppia di segnali periodici in

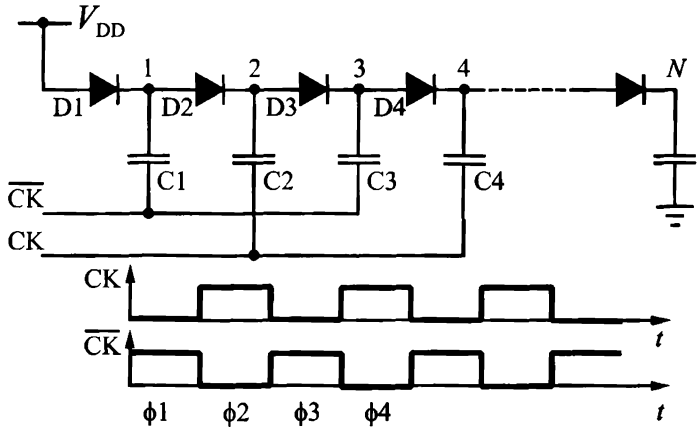


Figura 8.28 Esempio di circuito pompa di carica.

controfase, CK e $\overline{CK}$; l'ultima capacità della serie è la capacità di carico, sulla quale si intende trasferire la tensione finale generata dalla pompa di carica.

Per comprendere il funzionamento del circuito, si immagini che al tempo $t = 0$ tutti i diodi siano alla soglia di conduzione e che quindi le tensioni iniziali ai nodi del circuito siano

- $V(1) = V_{DD} - V_{th}$
- $V(2) = V_{DD} - 2V_{th}$
- ...
- $V(i) = V_{DD} - i \times V_{th}$

dove V_{th} è la caduta di tensione sul diodo polarizzato direttamente. Durante la fase Φ_1, $CK = 0$ e $\overline{CK} = 1$; quindi all'inizio di Φ_1, $\overline{CK}$ commuta da 0 a 1 e la corrispondente tensione passa da 0 a V_{DD}. Per effetto della capacità C_1, anche la tensione al nodo 1, $V(1)$, subisce un incremento della stessa quantità, salendo da $V_{DD} - V_{th}$ a $2V_{DD} - V_{th}$. Come effetto di tale variazione di tensione, il diodo D_1 si spegne e il diodo D_2 si accende, permettendo la ridistribuzione di carica tra le capacità C_1 e C_2 e facendo così salire anche $V(2)$. Un effetto simile si verifica per tutti i nodi dispari della catena di diodi.

All'inizio della successiva fase Φ_2, sarà CK a commutare da 0 a 1, portando D_2 all'interdizione e D_3 alla conduzione: la tensione $V(2)$ sale come effetto dell'incremento della tensione associata a CK e si ha ridistribuzione di carica tra i nodi 2 e 3, 4 e 5, e così via per tutte le altre coppie di nodi del tipo $2i$ e $2i + 1$.

Il circuito funziona quindi trasferendo progressivamente carica verso le capacità dei nodi di indice più elevato; poiché le capacità sono tutte uguali, l'aumento di carica è proporzionale all'aumento della tensione. La massima tensione al nodo finale si ottiene asintoticamente quando le variazioni di tensione indotte dalle transizioni di CK e $\overline{CK}$ non sono più sufficienti a portare in conduzione alcun diodo. Questo significa che ogni nodo deve essersi portato a una tensione superiore di $V_{DD} - V_{th}$ rispetto alla tensione del nodo precedente, a partire dal nodo 1, che si trova al potenziale $V_{DD} - V_{th}$. Quindi

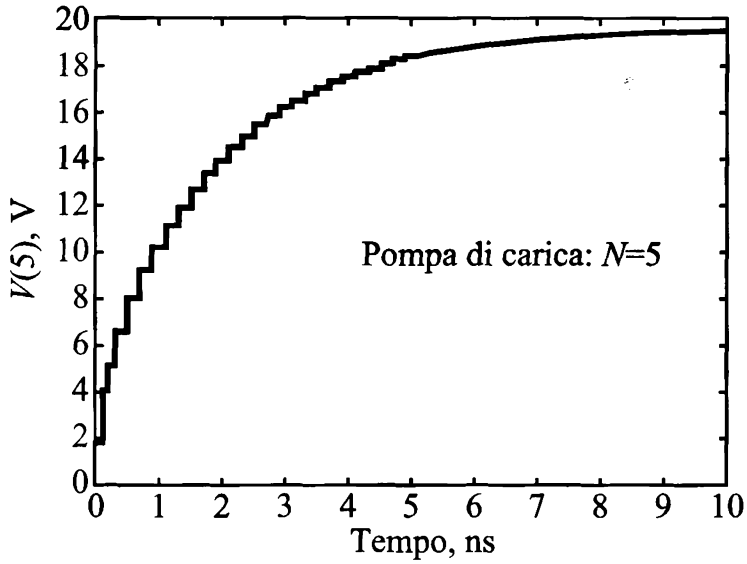


Figura 8.29 Andamento della tensione di uscita del circuito pompa di carica.

per il nodo finale si ha:

$$N\,(V_{DD} - V_{th})$$

L'andamento della tensione finale per il caso $N = 5$ è mostrato in figura 8.29; si noti che il valore asintotico previsto non è raggiunto nell'esempio perché occorrerebbero più iterazioni.

Circuiti di questo tipo permettono abbastanza agevolmente di generare tensioni anche molto elevate a partire da tensioni molto più basse. Tuttavia la tensione di uscita è generata dalla carica accumulata iterativamente su una serie di capacità e quindi il circuito non può erogare molta corrente, perché il prelievo di carica dal nodo di uscita tende ad abbassare rapidamente la tensione prodotta. Pertanto i circuiti a pompa di carica sono indicati soprattutto per le Flash che sfruttano solo il *tunneling Fowler Nordheim*, perché questo meccanismo di programmazione richiede tensioni elevate, ma assorbe correnti relativamente piccole. Al contrario, l'iniezione da portatori caldi richiede anche l'applicazione di tensioni alte tra drain e source, per generare portatori ad alta energia e questo implica un notevole consumo di corrente; in questo caso, i circuiti a pompa di carica sono meno efficaci e può essere necessario disporre di una seconda tensione di alimentazione.

8.5.5 Flash multilivello

Le architetture di tipo NOR e NAND discusse in precedenza sono solo due delle soluzioni studiate e proposte dai costruttori di memorie Flash, nel tentativo di perfezionare aspetti tecnologici diversi, come l'affidabilità, la velocità di accesso e la densità di integrazione. La ricerca nel settore delle memorie Flash è tutt'ora molto attiva e un approccio molto promettente è rappresentato dalle strutture multilivello.

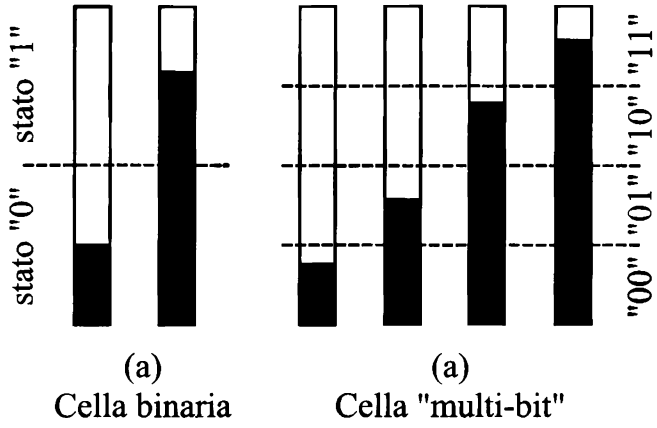


Figura 8.30 Corrispondenza tra la quantità di carica accumulata nel gate isolato e il valore logico della cella flash binaria e multi-livello.

L'elemento fondamentale della cella di memoria Flash è il transistore a gate isolato, come per le memorie EEPROM: la quantità di carica immagazzinata sul gate isolato determina il contenuto di informazione della cella. Nella tradizionale cella binaria, in grado di memorizzare un solo bit, una singola soglia è sufficiente a discriminare tra i due possibili stati logici: se la carica accumulata è inferiore alla soglia, il valore logico memorizzato è lo 0, altrimenti la cella memorizza un 1 logico (figura 8.30-a). Le celle *multi-bit*, nate per accrescere la densità delle memorie flash, permettono di memorizzare più bit per cella, usando più soglie e modulando opportunamente la quantità di carica presente sul gate isolato (figura 8.30-b). Per memorizzare b bit per singola cella, occorrono $2^b - 1$ tensioni di soglia distinte. Inoltre, per consentire i confronti multi-livello, la massima quantità di carica conservata nel gate isolato deve essere maggiore che nel caso a singolo bit, oppure, a parità di carica, i confronti con le soglie devono essere più precisi. I circuiti di controllo interni della memoria (per esempio riferimenti di tensione, comparatori, circuiti di scrittura e lettura) sono quindi nelle Flash multi-livello nettamente più complessi.

In figura 8.31 è mostrato un esempio di circuito di lettura per celle multilivello. Al comando di lettura, tre comparatori confrontano l'uscita della cella selezionata con altrettante celle di riferimento, programmate a tre diverse tensioni di soglia; le uscite binarie dei comparatori sono poi usate da un semplice encoder per ottenere la codifica su due bit del contenuto della cella indirizzata.

8.6 Memorie ferroelettriche

L'industria delle memorie è costantemente alla ricerca di soluzioni tecnologiche innovative, in grado di offrire dispositivi efficienti, caratterizzati da tempi ridotti di accesso in lettura e scrittura, comportamento non volatile, affidabilità prolungata e costo di produzione competitivo con quello delle migliori memorie oggi in uso.

In questo settore di ricerca, si possono citare tra le molte proposte le *memorie ferroelettriche* o *FERAM*, che sfruttano la polarizzazione permanente di un materiale

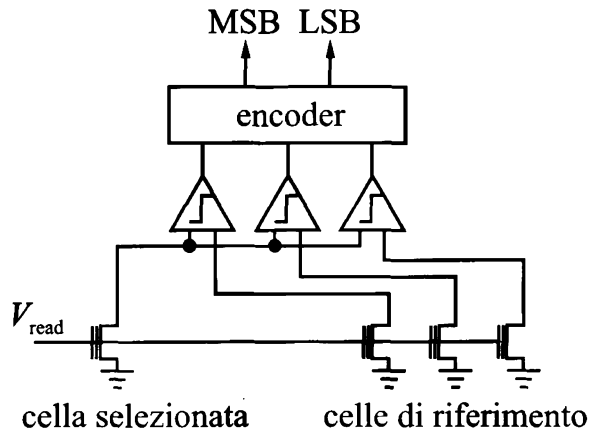


Figura 8.31 Circuito di lettura per celle multi-livello.

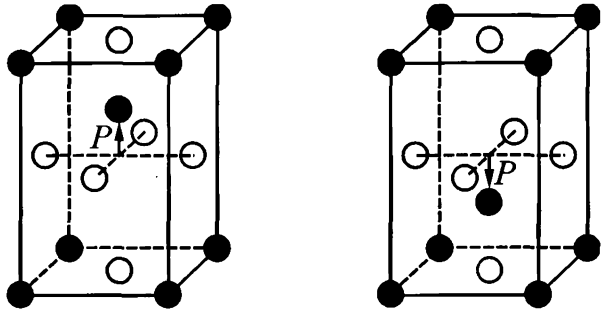


Figura 8.32 Polarizzazione dei materiali ferroelettrici.

ferroelettrico come meccanismo di conservazione dell'informazione.

I materiali ferroelettrici hanno polarizzazione elettrica spontanea (figura 8.32) e tale caratteristica può essere sfruttata per realizzare una cella di memoria non volatile. Infatti la direzione di polarizzazione può essere alterata mediante applicazione di un campo esterno, ma tende a conservarsi molto bene in assenza di campo. È quindi possibile associare il contenuto di informazione alla direzione di polarizzazione del materiale. La struttura più comune per i materiali ferroelettrici è del tipo ABO_3, dove A e B sono ioni metallici. Esempi di materiali con questa struttura sono: $BaTiO_3$, $BaSrTiO_3$, $PbTiO_3$.

Ad esempio, nel caso della struttura cristallina in figura 8.32, i nodi neri possono essere ioni Pb^{2+}, quelli bianchi ioni di ossigeno O^{2-} e quello grigio, con due siti cristallini possibili, uno ione Ti^{4+}.

Alla base delle memorie ferroelettriche (note anche come FERAM) vi è il condensatore ferroelettrico, la cui struttura è mostrata in figura 8.33 (FECAP); il condensatore ferroelettrico è dotato di due armature conduttive separate da uno strato ferroelettri-

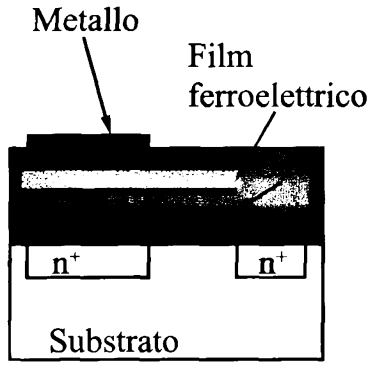


Figura 8.33 Il condensatore ferroelettrico.

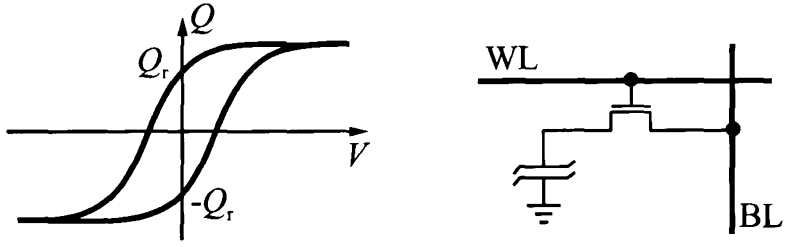


Figura 8.34 La caratteristica $V(Q)$.

Memoria	*EEPROM*	*Flash*	*FERAM*
area (relativa)	2	1	1,2
tempo di accesso (lettura)	50 ns	50 ns	100 ns
tempo di accesso (scrittura)	10 μs	100 ns	100 ns
Tensione di programmazione	10-18 V	10-18 V	2 V
Energia di programmazione	1 pJ	10 nJ	1 pJ
numero di scritture	10^5	10^5	10^{12}
numero di letture	∞	∞	10^{12}

Tabella 8.6 Confronti tra memoria non volatili.

co. La variazione di polarizzazione nel dielettrico induce una variazione di carica sugli elettrodi.

La caratteristica $V - Q$ (tensione - carica elettrica) del FECAP presenta isteresi, come indicato in figura 8.34. La carica residua, Q_r o $-Q_r$, è associata all'informazione conservata.

Le FERAM (tabella 8.6) sono caratterizzate da:

- ▷ bassissimo consumo, soprattutto in scrittura, rispetto alle altre memorie non volatili, poiché non richiedono tensioni elevate di cancellazione e programmazione
- ▷ ottima affidabilità, corrispondente a un numero permesso di ri-scritture che è di ordini di grandezza superiore rispetto ai valori tipici delle EEPROM e delle Flash
- ▷ facilità di integrazione in dimensioni molto contenute, inferiori a quelle tipiche delle Flash e delle RAM dinamiche.

La principale limitazione delle FERAM è probabilmente rappresentata dalla attuale difficoltà a realizzare memorie delle stesse dimensioni comunemente ottenibili per altri tipi di memorie.

8.7 Memorie a lettura/scrittura

Le memorie non volatili descritte nei paragrafi precedenti consentono numerose operazioni di programmazione e cancellazione sul medesimo dispositivo e le soluzioni tecnologiche più recenti sono ormai molto distanti dalle prime EPROM, arrivando ad offrire ottime prestazioni in termini sia di tempi di accesso in scrittura, che di affidabilità. Tuttavia nella classificazione introdutta all'inizio del capitolo, le memorie a lettura e scrittura rappresentano dispositivi nei quali le operazioni di accesso in lettura e scrittura siano del tutto equivalenti in termini di semplicità di attuazione, tempi di completamento e affidabilità. Per ottenere al meglio queste caratteristiche, si deve rinunciare alla non volatilità delle EEPROM e Flash, riorganizzando completamente la struttura della cella di memoria.

Le memorie a lettura e scrittura, anche dette tradizionalmente memorie ad *accesso casuale*, si dividono in due ampie classi, sulla base del principio sfruttato per immagazzinare l'informazione:

- ▷ RAM statiche (*SRAM*), nelle quali l'elemento base di memoria è costituito da un circuito con retroazione positiva, formato da due inverter chiusi in anello; il livello di tensione presente su uno dei due nodi dell'anello è associato all'informazione.
- ▷ RAM dinamiche (*DRAM*), nelle quali l'elemento base di memoria è rappresentato da una capacità e l'informazione è associata alla carica elettrica immagazzinata.

8.7.1 Memorie RAM statiche

Due inverter chiusi in retroazione mantengono i livelli di tensione inizialmente forzati ai due nodi finché il circuito è alimentato. Tale struttura, rappresentata in figura 8.35, è nettamente più onerosa in termini di area di silicio occupata rispetto a quanto visto in precedenza per le celle delle memorie non volatili; tuttavia la struttura a retroazione garantisce un elevatissimo grado di robustezza ai disturbi esterni, in quanto i due inverter sono in grado di rigenerare efficacemente i livelli corretti di tensione ai due nodi.

Per ragioni che saranno chiarite nel seguito, in questo caso ogni colonna della matrice di celle di memoria è connessa a una coppia di *Bit Line*, anziché a una sola, come avviene per tutte le altre memorie; delle due *Bit Line*, una trasporta sempre il dato della cella selezionata e l'altra il suo complemento. Per consentire la connessione con le *Bit Line*, la cella è completata da una coppia di *pass transistor*, ovvero transitori a

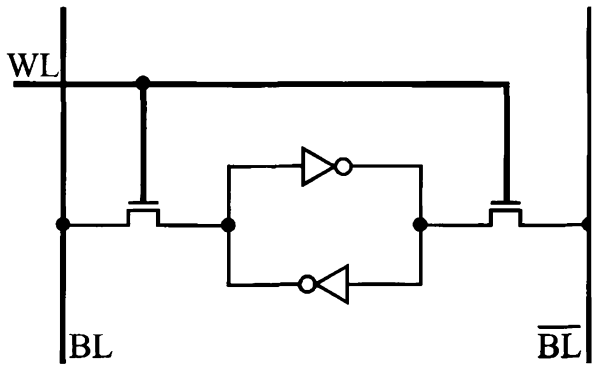


Figura 8.35 Cella di memoria RAM statica.

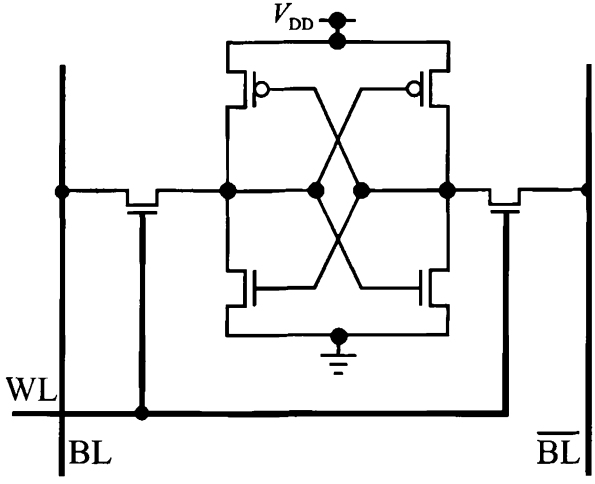


Figura 8.36 Cella a 6 transistori.

canale n, usati come interruttori: la tensione alta sulla *Word Line* *pass transistor* e connette quindi la cella indirizzata alle due *Bit Lin*

La struttura completa della cella SRAM, mostrata in figura 8.36, 6 transistori ed è nota come "cella 6-T". Le caratteristiche più importa sono:

1. La dinamica di tensione è molto ampia, $0 \div V_{DD}$: infatti, bench non possano propagare livelli di tensione superiori alla tensione diminuita della tensione di soglia, $V_{DD} - V_{th}$, gli inverter della r grado di rigenerare il livello massimo della dinamica.
2. Il consumo di potenza statico è approssimativamente nullo, gra complementare degli inverter, nei quali i due transistori lavorar

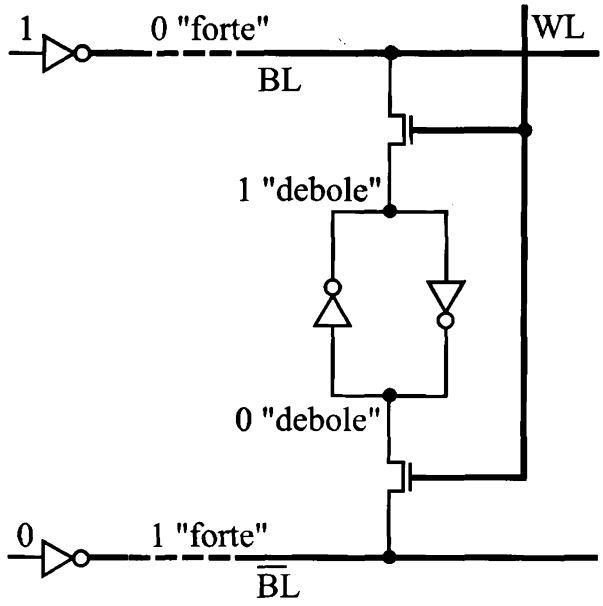


Figura 8.37 Accesso alla cella SRAM in scrittura.

trofase, non permettendo quindi, in condizioni statiche, il passaggio di corrente tra alimentazione e massa. In realtà, nelle tecnologie più recenti, le correnti sotto-soglia dei transistori introducono un effetto di dissipazione statica di crescente rilevanza; tale consumo rimane però ad oggi trascurabile rispetto alla dissipazione dinamica, legata alla commutazione degli inverter.

3. Per massimizzare la densità di integrazione, occorre ridurre il più possibile l'occupazione di area della singola cella e a questo scopo si ricorre all'isolamento a trincea di ossido tra i dispositivi.

Analizziamo ora la procedura di accesso in scrittura. Inizialmente, prima dell'effettiva selezione della cella indirizzata e quindi prima dell'attivazione della *Word Line*, le *Bit Line* sono pilotate al valore che si intende scrivere; tale funzione è svolta da appositi circuiti di pilotaggio, uno per ogni *Bit Line*, che ricevono il dato dall'esterno e lo propagano internamente alla memoria, applicandolo alle *Bit Line*. Nel momento in cui la *Word Line* è attivata, i *pass transistor* si accendono e i due nodi interni alla cella sono messi in contatto elettrico con le due *Bit Line* esterne (figura 8.37). Tale situazione produce potenzialmente un conflitto, in quanto i livelli che sono stati precedentemente immagazzinati nella cella e che continuano ad essere pilotati dagli inverter della cella possono essere opposti a quelli forzati dall'esterno durante la scrittura. Come esempio, la figura 8.37 mostra uno 0 logico forzato sulla *Bit Line* BL dal circuito di pilotaggio di colonna e un 1 logico forzato sul medesimo nodo dalla cella stessa. Il conflitto, quando si è verificato, viene risolto spontaneamente e senza ritardi apprezzabili dal dispositivo complessivo, grazie alle dimensioni dei circuiti di pilotaggio in gioco. Infatti, per ragioni di densità di integrazione, gli inverter della cella di memoria sono tipicamente di

dimensioni minime e quindi hanno poca capacità di pilotaggio: questo significa che la corrente erogata, direttamente proporzionale alla larghezza W dei transistori,

$$I_{DS} = \frac{W}{L}\mu C_{ox}\,(V_{GS} - V_{th})$$

è comunque molto piccola; analogamente si può dire che la resistenza equivalente dei transistori è molto elevata:

$$R_{eq} = \frac{V_{DS}}{I_{DS}}$$

Al contrario, i circuiti di pilotaggio delle *Bit Line* corrispondenti alle colonne sono caratterizzati da dimensioni ben superiori, in quanto debbono pilotare linee metalliche che attraversano in altezza l'intera matrice di memoria e sono connesse a tutte le celle di una colonna; d'altra parte il numero di circuiti di pilotaggio di *Bit Line* è molto più basso del numero di inverter delle celle interne a una memoria e quindi le maggiori dimensioni dei transistori che compongono i primi hanno un effetto trascurabile sulla densità di integrazione. Nella figura 8.37, questa differenza in termini di capacità di pilotaggio è rappresentata indicando con gli aggettivi *strong* e *weak* i livelli logici prodotti da circuiti di pilotaggio rispettivamente forti e deboli.

Nell'operazione di scrittura quindi i circuiti di pilotaggio delle *Bit Line* forzano i nuovi valori da memorizzare sui nodi interni della cella attraverso i *pass transistor* accesi; quando si verifica un conflitto tra circuito di pilotaggio esterno e inverter della cella, questo viene comunque risolto a favore del primo.

Più problematico risulta l'accesso in lettura. In questo caso, i circuiti di pilotaggio esterni alla cella non intervengono e sono gli inverter della cella stessa che devono imporre una tensione alle *Bit Line*. Come già ricordato, per ragioni di densità di integrazione, i transistori di ogni cella sono di dimensioni minime e forniscono quindi correnti limitate. Inoltre le *Bit Line*, lunghe in proporzione alle dimensioni della matrice di memoria, sono associate a capacità parassite elevate. Poiché il tempo necessario per variare di una quantità ΔV la tensione di una capacità C è dato da

$$\Delta t = \Delta V \frac{C}{I}$$

ovvero Δt è direttamente proporzionale alla capacità e inversamente proporzionale alla corrente, è chiaro che la transizione pilotata dagli inverter della cella sulle *Bit Line* tende a compiersi in tempi estremamente lunghi.

Per abbreviare la transizione, gli inverter della cella selezionata devono essere appoggiati da un circuito dotato di maggiore capacità di pilotaggio, cioè di erogare la corrente necessaria a caricare o scaricare la capacità associata alla *Bit Line* in un tempo molto più breve. Tale ruolo è svolto nelle memorie SRAM da un circuito noto come *sense amplifier*: si tratta di un amplificatore di tipo differenziale, in grado, quando attivato, di "sentire" la differenza tra le tensioni di due nodi e amplificarla molto rapidamente alla massima dinamica. Il modo nel quale il *sense amplifier* è connesso con il resto della memoria è mostrato in figura 8.38, dove sono riprodotti la cella di memoria indirizzata e il *sense amplifier* della colonna corrispondente. Si noti che, poiché una sola cella di memoria per colonna può essere selezionata, è sufficiente un solo *sense amplifier* per ogni coppia di *Bit Line*. Nella figura, il segnale SE abilita il *sense amplifier*

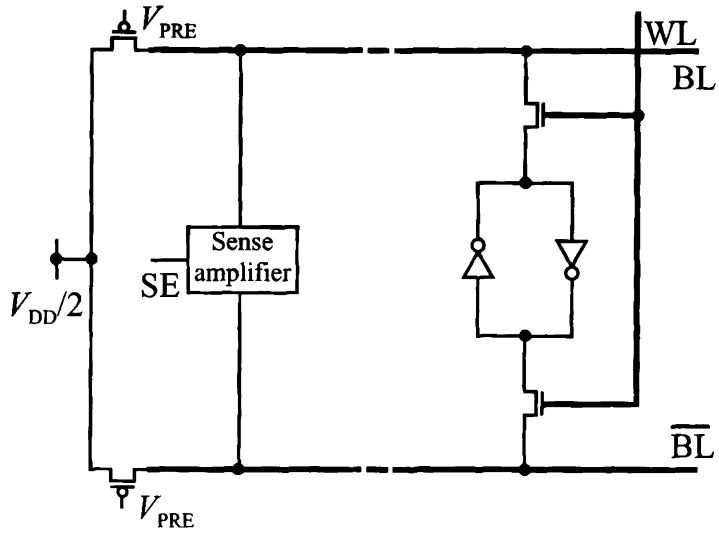


Figura 8.38 Uso del *sense amplifier* nelle SRAM.

e la scelta dell'istante di tempo nel quale tale segnale viene inviato è estremamente critica: il *sense amplifier* deve infatti essere attivato il più presto possibile, al fine di ridurre il tempo di lettura, ma non prima che la cella di memoria selezionata abbia indotto autonomamente sulle *Bit Line* una variazione di tensione sufficientemente ampia da superare le fluttuazioni casuali dovute al rumore e alla dispersione dei parametri tecnologici. In caso contrario, il *sense amplifier* potrebbe rilevare una differenza di tensione tra le due *Bit Line* dovuta a fenomeni indipendenti dal valore logico memorizzato nella cella e quindi determinare un errore di lettura.

Una tipica sequenza di operazioni previste durante l'accesso in lettura è la seguente:

1. Le *Bit Line* sono precaricate alla medesima tensione iniziale, di solito $V_{DD}/2$, in modo da presentare una differenza di potenziale nulla.
2. Subito dopo, la cella di memoria indirizzata è selezionata attivando la *Word Line*; gli inverter interni alla cella iniziano ad agire come circuiti di pilotaggio per le *Bit Line*, i cui livelli di tensione incominciano a muoversi in direzioni opposte, ma molto lentamente.
3. Il *sense amplifier* viene attivato, abilitando il segnale *SE*: l'effetto risultante è semplicemente di accelerare le transizioni sulle due *Bit Line*, come indicato nella figura 8.39, che rappresenta l'andamento delle tensioni di due *Bit Line* durante un'operazione di lettura.

8.7.2 Il *sense amplifier*

Un *sense amplifier* è un circuito che permette di ridurre il tempo di accesso alla locazione indirizzata, convertendo rapidamente livelli logici arbitrari presenti sulle *Bit*

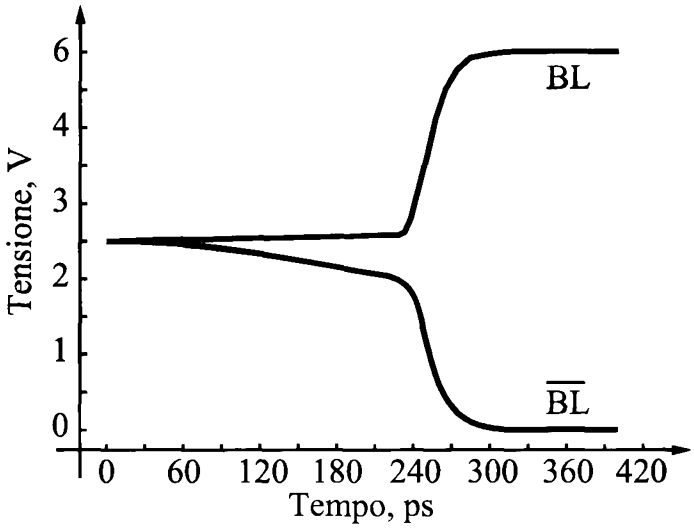


Figura 8.39 Transizioni di tensione sulle *Bit Line* e effetto del *sense amplifier*.

Line in livelli validi, riconoscibili correttamente dai dispositivi perifierici della memoria e all'esterno della memoria stessa.

Il progetto di un *sense amplifier*, e in particolare il relativo guadagno e ritado, è fortemente vincolato in termini di tolleranza alla condizioni ambientali (come le variazioni di temperatura e tensione di alimentazione) e alla dispersione dei parameti tecnologici. Il progetto è anche fortemente vincolato in termini di area e di rapporto d'aspetto: infatti, è necessario collocare un *sense amplifier* in corrispondenza di ogni *Bit Line* della memoria e quindi la larghezza del *sense amplifier* è limitata a quella delle celle di memoria.

La tipologia di amplificatore più diffusa per il *sense amplifier* è quella differenziale, soprattutto per i noti vantaggi offerti in termini di reiezione del modo comune, cioè capacità di attenuare componenti di disturbo che agiscano in modo bilanciato tra i due rami dell'amplificatore differenziale; anche la reiezione dei disturbi legati all'alimentazione è importante.

Un esempio di *sense amplifier* è riprodotto in figura 8.40: si tratta di uno stadio complementare con retroazione positiva, con due terminali, i nodi 1 e 2, che agiscono simultaneamente come ingressi e uscite del circuito. La retroazione positiva, introdotta dalle connessioni incrociate tra drain e gate dei transistori, fornisce rispetto a versioni senza reazione vantaggi di guadagno maggiore e ritardo inferiore.

I transistori superiore, $MP3$, e inferiore, $MN3$, permettono di accendere l'amplificatore all'attivazione del segnale di controllo SE, che viene generato internamente alla memoria con un piccolo ritardo rispetto all'inizio dell'operazione di lettura, in modo da garantire la misura e l'amplificazione di una differenza di tensione già salita ben sopra le fluttuazioni derivanti dal rumore ambientale a dalle variazioni di processo.

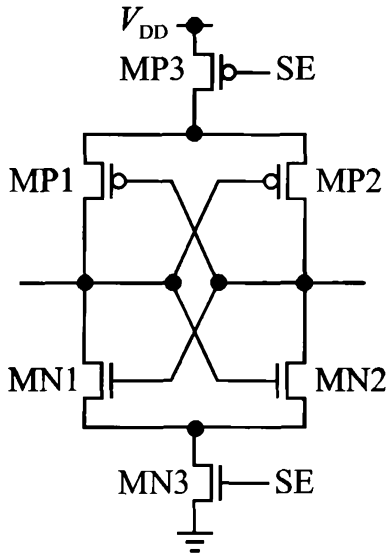


Figura 8.40 Esempio di *sense amplifier*.

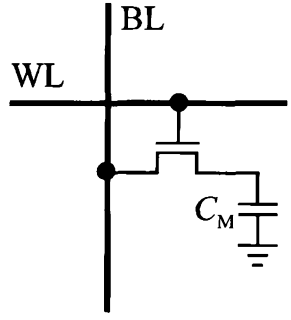


Figura 8.41 Struttura di una cella DRAM.

8.7.3 Memorie RAM dinamiche

L'elemento base di memoria nelle RAM dinamiche (DRAM) è la capacità C_M, in grado di immagazzinare (temporaneamente) una carica elettrica, cui è associata l'informazione da conservare. Leggere il contenuto della cella significa quindi semplicemente verificare la quantità di carica accumulata, confrontando il livello di tensione corrispondente, mentre scrivere la cella significa caricare o scaricare la capacità.

Il funzionamento è concettualmente molto semplice è l'area occupata estremamente ridotta, visto che per ogni cella non è richiesto altro che una capacità e un transistore di accesso alla cella (figura 8.41). Proprio questa estrema semplicità della cella genera però alcuni problemi di uso della memoria che possono essere superati grazie a opportuni

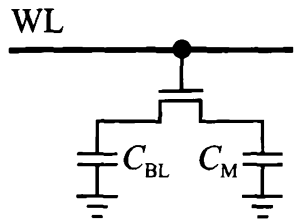


Figura 8.42 Ridistribuzione di carica per una cella DRAM.

accorgimenti.

Analizziamo innanzi tutto l'operazione di lettura. Durante la lettura, il transistore di accesso connette la capacità interna della cella, C_M, con quella esterna associata alla *Bit Line*, C_{BL}. Similmente a quanto già visto per le memorie statiche, la necessità di minimizzare le dimensioni della cella e massimizzare il numero di celle per dispositivo di memoria, induce un'ampia differenza tra le capacità C_M e C_{BL} e tale differenza rende di fatto impossibile per la cella pilotare la *Bit Line*.

Nell'operazione di lettura infatti, la carica inizialmente presente sulle capacità interna, C_M, e esterna, C_{BL}, si ridistribuisce tra le due capacità all'atto dell'attivazione del transistor di accesso (figura 8.42): la tensione finale alla quale si portano cella e *Bit Line* dipende dalle dimensioni relative di C_M e C_{BL}.

Tale fenomeno è noto come "ridistribuzione di carica", o *charge sharing*, e, nell'ipotesi di capacità e interruttori ideali, la tensione finale può essere determinata applicando il principio di conservazione della carica. Prima della selezione della cella, la carica elettrica complessivamente presenta sulle due capacità è data da:

$$Q_{tot} = Q_{BL} + Q_M = C_{BL}V_{BL} + C_M V_M$$

dove V_{BL} e V_M sono le tensioni iniziali rispettivamente presenti sulla *Bit Line* e all'interno della cella. Tali tensioni possono essere uguali, e in questo caso non ci sarà *charge sharing*, oppure diverse. Dopo la selezione della cella e al termine del possibile transitorio, la carica vale

$$Q_{tot} = (C_{BL} + C_M)V_f$$

dove V_f è la tensione finale alla quale di portano entrambe le capacità. Imponendo la conservazione della carica, si può risolvere l'eguaglianza rispetto all'incognita V_f, trovando:

$$V_f = \frac{C_{BL}V_{BL} + C_M V_M}{C_{BL} + C_M}$$

Se $C_{BL} >> C_M$, come avviene solitamente, allora la tensione finale è di poco modificata rispetto a quella iniziale sulla *Bit Line*, cioè

$$V_f \approx V_{BL}$$

e non è possibile acquisire questa tensione all'esterno della memoria come risultato della lettura. Inoltre lo stesso livello di tensione V_f, al termine della lettura, quando il transistore di accesso viene disattivato, rimane memorizzato all'interno della cella,

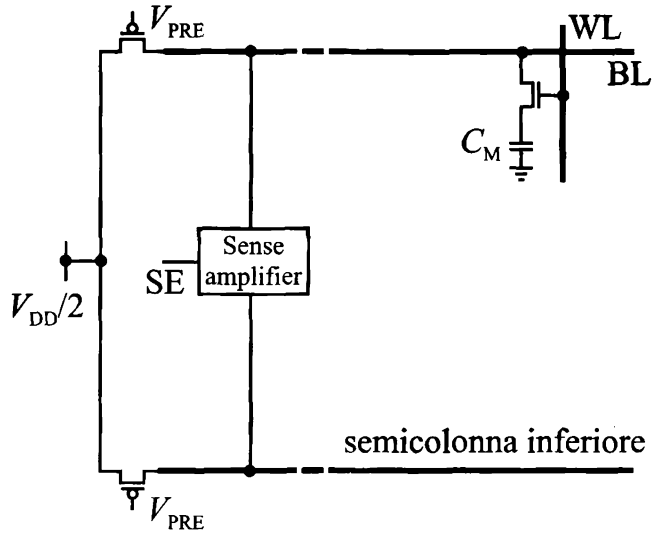


Figura 8.43 Uso del *sense amplifier* nelle DRAM.

distruggendo di fatto il contenuto di informazione.

Un esempio numerico può aiutare a comprendere meglio il problema. Si assuma che $C_{BL} = 10\ C_M$ e che, subito prima della lettura, la *Bit Line* sia precaricata alla tensione $V_{BL} = 0{,}5\ V_{DD} = 2{,}5$ V. La tensione finale della cella e della *Bit Line* sarà

- $V_f = 2{,}73$ V, nel caso $V_M = 5$ V, e
- $V_f = 2{,}27$ V, nel caso $V_M = 0$

Riassumendo, i problemi indotti dalla ridistribuzione di carica nelle DRAM sono due:

1. la piccola variazione di tensione sulla *Bit Line* conseguente alla selezione della cella e relativo *charge sharing* deve essere amplificata alla dinamica completa 0-V_{DD},
2. il valore di tensione originale V_M deve essere ripristinato nella cella di memoria letta.

Come già visto nel caso delle SRAM, il *sense amplifier* è efficace nel completare una transizione parzialmente attuata da altri circuiti. Mentre nel caso delle SRAM il *sense amplifier* è necessario per velocizzare la transizione, in questo caso esso permette di risolvere i due problemi citati, ovvero portare a termine una transizione che non potrebbe essere completata dalla cella selezionata e ripristinare il corretto valore di tensione sulla capacità della cella letta. Il *sense amplifier* inoltre gioca un ruolo fondamentale nella velocità dell'accesso in lettura.

Tipicamente ogni *Bit Line* è divisa in due semi-colonne precaricate a $V_{DD}/2$ prima di ogni lettura e l'amplificatore è collocato in mezzo, come indicato in figura 8.43. La tensione sulla semi-colonna che ospita la cella selezionata verrà lievemente alterata come effetto del *charge sharing* tra cella e *Bit Line*, nella direzione di salire o scendere

SRAM	DRAM
tempo di accesso < 10 ns	60 – 80 ns
densità 4-6 volte inferiore	necessita di "refresh"
costo 10 volte superiore	interfaccia più complessa
tagli tipici: 256-512 kB	64 MB
uso: cache (interna o esterna)	memoria principale

Tabella 8.7 Confronto tra memoria RAM statiche e dinamiche.

rispetto al valore di precarica, a seconda del valore logico immagazzinato. La tensione sull'altra semi-colonna non è invece alterata e può essere presa come riferimento per il *sense amplifier*.

Un secondo problema generato dalla estrema semplicità della cella DRAM è rappresentato dalle correnti di perdita che insistono sulle piccole capacità interne delle celle e che tendono ad alterarne la carica accumulata.

Per esempio, con una corrente di perdita $I_l = 10^{-14}$ A, una capacità $C_M = 1$ fF e una tensione iniziale $V_M = 1{,}8$ V, la tensione della cella scende a 1 V in appena 80 ms!

$$I_l = C_M \frac{\Delta V}{\Delta t} \quad \longrightarrow \quad \Delta t = \frac{C_M}{I_l}(1{,}8 - 1) = 0{,}08\,\mathrm{s}$$

Per evitare questo effetto, tutte le celle devono essere periodicamente lette (*refresh*): infatti, grazie al *sense amplifier*, l'operazione di lettura rigenera i valori di tensione corretti anche sulla capacità interna della cella. Non è necessario che l'operazione di lettura sia portata a compimento, rendendo disponibili i dati letti all'esterno del dispositivo, ma è sufficiente indirizzare periodicamente tutte le celle. Le memorie DRAM sono dotate di un controllore interno (*refresh controller* che provvede autonomamente all'operazione di *refresh*, senza richiedere alcun intervento dall'esterno della memoria.

Un confronto riassuntivo tra le memorie SRAM e DRAM è dato in tabella 8.7. I principali vantaggi offerti dalle memorie dinamiche sono:

- ▷ l'elevata densità di integrazione, che deriva dalla semplicità della singola cella,
- ▷ la possibilità di raggiungere capacità di memoria molto alte per singolo dispositivo

Le memorie statiche invece sono caratterizzate da una maggiore semplicità di uso, in quanto non richiedono *refresh*, e da tempi di accesso nettamente inferiori.

La DRAM trovano quindi applicazione quando si richiedono grandi disponinbilità di memoria, come nel caso della memoria principale di un computer; delle SRAM invece interessa sfruttare soprattutto l'elevata velocità: per esempio, la tecnologia SRAM è adottata nella realizzazione di memorie cache, di I e II livello, per le quali si richiedono in genere tempi di accesso inferiori ai 10 ns e capacità inferiori ai 10 Kbyte.

Dato che il basso costo per bit conseguibile con le RAM dinamiche rende questa tecnologia adatta alla realizzazione di dispositivi di memoria di grandi dimensioni, le DRAM sono comunemente configurate in modo da contenere il numero di pin, usando in particolare le linee di indirizzo in *multiplexing*, ovvero condividendo le stesse linee fisiche per trasportare separatamente prima una metà e poi l'altra dell'indirizzo. A questo

scopo, si sfrutta la struttura già presentata in figura 8.4, dove i bit di indirizzo sono divisi in due parti, n_1 e n_2, applicate rispettivamente al decoder di riga e di colonna. Si supponga che $n_1 = n_2 = n/2$, dove 2^n è il numero di locazioni della memoria. Inizialmente, l'indirizzo di riga è caricato attraverso le $n/2$ linee fisiche del dispositivo per selezionare una riga della matrice di memoria; le DRAM usano tradizionalmente un segnale di controllo apposito, *RAS* (*Row Address Strobe*), per attivare questa fase dell'accesso. Successivamente, le medesime $n/2$ linee sono usate per ricevere la seconda metà dell'indirizzo, che permette al decoder di colonna di completare l'accesso alla locazione richiesta; in questa II fase, è attivo il segnale di *CAS* (*Column Address Strobe*). Il *multiplexing* degli indirizzi comporta ovviamente un aumento del tempo complessivo di accesso alla memoria, ma tale aumento è abbastanza contenuto perché le operazioni di selezione della riga e della colonna non potrebbero comunque essere svolte in parallelo.

Per aumentare la velocità di accesso alle DRAM, si ricorre diffusamente alla tecnica del *pipelining*. L'idea base è molto semplice e consiste nel scomporre l'operazione di accesso in due o più sotto-operazioni da eseguire in successione, ciascuna di durata pari alla metà o meno del tempo di accesso complessivo; anziché portare a termine tutte le sotto-operazioni previste per un singolo accesso prima di accettare un nuovo indirizzo e comando di lettura o scrittura, si consente alla memoria di lavorare simultaneamente su due o più accessi distinti. Mentre la seconda sotto-operazione è eseguita per accedere a un dato, la prima sotto-operazione viene eseguita in parallelo per l'accesso al dato successivo. In questo modo, la separazione temporale tra due dati successivamente letti (o scritti) è pari alla durata di una singola sotto-operazione. Tecniche di questo tipo sono applicate per esempio nelle DRAM sincrone (SDRAM), memorie dinamiche a elevata velocità, dotate di un'interfaccia sincrona, tramite la quale avviene l'accesso dati.

8.7.4 Memorie sincrone SDRAM

Le memorie SDRAM (Synchronous Dynamic RAM) sono tra le più diffuse e efficienti memorie a semiconduttore e si differenziano dalla memorie asincrone per alcune caratteristiche fondamentali.

Innanzitutto i comandi di controllo, gli indirizzi e i dati sono forniti in modo sincrono, ovvero allineato temporalmente a un segnale di sincronizzazione o clock fornito dall'esterno. La struttura interna è poi costituita da 2 o più banchi di memoria attivati in sequenza, secondo una modalità di esecuzione di tipo *pipeline*: per esempio, nel caso di due banchi, mentre il primo si trova nella condizione di precarica, il secondo opera sui dati, effettuando l'accesso vero e proprio al dato, in lettura o scrittura. Questa struttura permette quindi di realizzare una forma di parallelismo interno che aumenta il numero di locazioni lette o scritte nell'unità di tempo. Infine, nelle SDRAM l'indirizzo di colonna tipicamente non viene fornito dall'esterno, ma generato internamente e incrementato automaticamente, per accedere in sequenza a locazioni adiacenti. Questo modo di funzionamento implica l'uso di una macchina a stati finiti che generi internamente al dispositivo di memoria gli indirizzi e i comandi necessari; per contro l'accesso sincrono e sequenziale (noto anche come *burst mode*) permette di aumentare notevolmente la velocità di lettura e scrittura.

Queste caratteristiche si traducono nel tipo di diagramma temporale indicato nella figura 8.44, dove si evidenzia come tutti gli eventi siano allineati ai fronti del segnale di

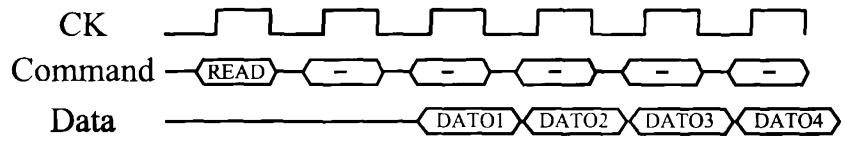


Figura 8.44 Tempistiche nelle memorie SDRAM.

clock e i dati letti siano disponibili uno per ciclo di clock. Tra il comando di lettura e la disponibilità del primo dato trascorre un intervallo di tempo tipicamente superiore al ciclo di clock e detto latenza: questo tempo è necessario perché la macchina a stati interna generi gli indirizzi e attivi i controlli interni per l'accesso vero e proprio alla matrice di memoria.

Bibliografia

[1] P. Caldirola, R. Cirelli, G. M. Prosperi, *Introduzione alla fisica teorica,* UTET, Torino, 1982.
[2] G. Ghione, *Dispositivi per la microelettronica,* McGraw-Hill, Milano, 1998.
[3] I. S. Gradshteyn, I. M. Ryzhik, *Table of integrals, series and products,* Academic Press, San Diego, 1994.