

Appunti di Progettazione di Sistemi Microelettronici
Modulo Analogico



Università di Pisa – Ingegneria Elettronica 2024/2025

Prof.: Paolo Bruschi, Michele Dei

Autore: Gabriele Stefani

Indice

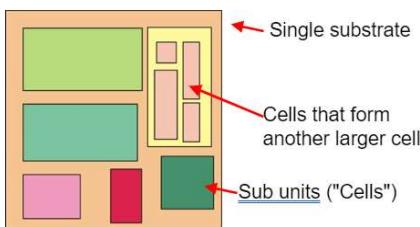
Progettazione di microsistemi elettronici analogici

Quando si parla di progettazione di circuiti integrati, l'analogico diventa una sorta di metodologia. La filosofia con cui si realizzano circuiti integrati analogici si ritrova nella realizzazione dei circuiti digitali base, quali flip-flop, porte logiche, ecc. In questo modulo daremo uno sguardo dal punto di vista dell'integrazione ai dispositivi passivi e attivi, concentrandoci sulle tecniche di realizzazione, per poi formalizzare e trattare i blocchi elementari della microelettronica integrata, tra cui amplificatori, specchi di corrente, coppie differenziali. È importante notare che la parola “progettazione” implica non più uno studio di analisi dei circuiti fine a sé stesso, bensì uno studio mirato alla loro sintesi concreta su silicio.

La progettazione integrata lascia molto spazio all'immaginazione e alla creatività personale; è un'arte. La necessità di progettare in questo campo va di pari passo alla progressiva miniaturizzazione dei circuiti, fenomeno che rende il settore molto competitivo, sia dal punto di vista dei software necessari alla progettazione, che dal punto di vista delle soluzioni hardware.

Celle

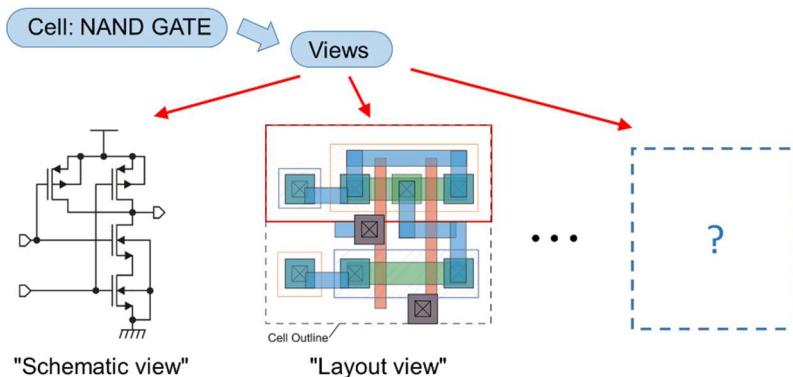
Un circuito integrato è un sistema elettronico composto da dispositivi realizzati sullo stesso substrato che, in assenza di particolari requisiti di velocità/potenza, è in silicio. All'interno il circuito integrato (o “chip”, “die”) sarà suddiviso in unità funzionali diverse.



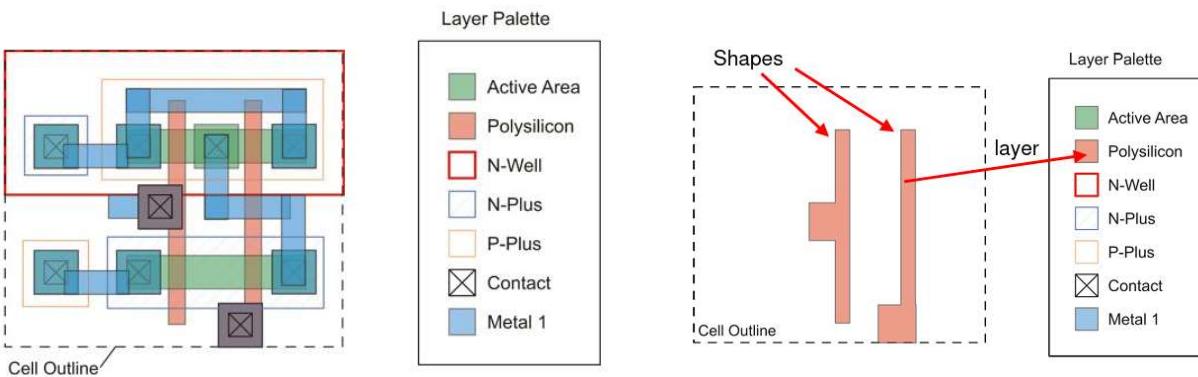
I blocchi che costituiscono PIC, circoscritti in fase di progetto (ampli, converter, ecc.) e messi assieme per implementare la funzione complessiva, prendono il nome di celle (o moduli). Se queste unità sono progettate e vendute da terzi, prendono il nome di IP (intellectual property). La stessa cella, con ottica gerarchica, può essere suddivisa in più parti, in più celle. Il chip stesso, nella sua interezza, è da considerarsi una cella.

La cella più complessa contiene tutte le altre e prende il nome di **top-cell**. Spesso la top-cell coincide al chip, ma non sempre. Se la progettazione si riferisce ad un circuito che solo in un secondo momento sarà integrato nel chip assieme ad altri moduli, allora la top-cell coinciderà a quel particolare circuito.

Una stessa cella può avere più **viste**. Consideriamo ad esempio una porta NAND in tecnologia CMOS. Una delle viste è lo schematico (schema elettrico), “schematic view”.



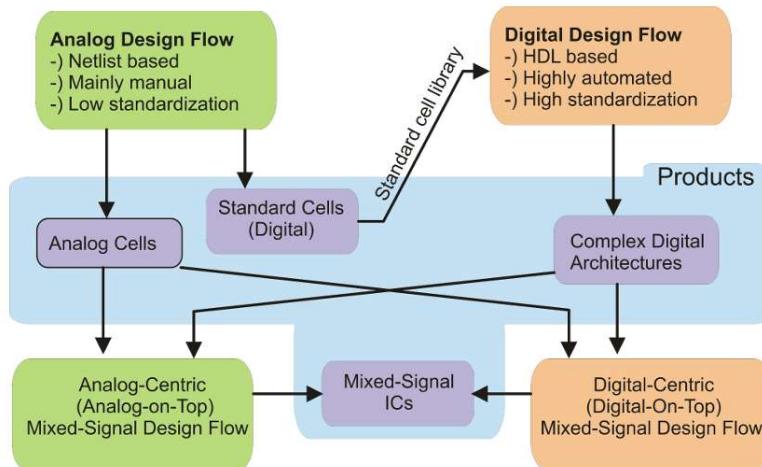
Un'altra vista, coerente con la prima, è la vista geometrica di layout. Le viste sono dei modi di rappresentare la cella finalizzati a descriverne la funzione in un certo contesto. In un ambiente professionale come Cadence possono esserci anche 10 viste, ognuna supportata da un tool diverso. Lo schematico elettrico fissa il comportamento del circuito in modo virtuale, specificando i componenti e le interconnessioni tra di essi. Un'architettura con un certo comportamento elettrico può essere disegnata come schematico in tanti modi diversi, per cui la schematic view è una vista abbastanza elastica. Per il layout scompare questo grado di libertà: la geometria, i distanziamenti, il posizionamento dei vari livelli determinano univocamente il comportamento fisico, reale, del circuito.



La vista layout è la parte di progetto che contiene le indicazioni per la realizzazione stratificata del circuito integrato. La fonderia tradurrà le informazioni contenute nel layout nei passi di processo da condurre sul wafer di silicio per realizzare il circuito integrato. Il layout è costituito da figure geometriche (**shapes**) la cui forma codifica per l'area del substrato di silicio interessata da una particolare sequenza di operazioni. Ogni forma è caratterizzata da un **layer** (strato) che fornisce l'informazione della qualità elettrica della geometria. I layer sono raccolti in una tavolozza (**palette**) e differenziati sulla base del colore e dello stile grafico. Processi diversi avranno layer diversi. Dunque, il layout contiene le informazioni sia della posizione geometrica che della tipologia delle operazioni fisiche da applicare al substrato per realizzare il circuito integrato. L'outline è l'ingombro complessivo della cella. Spesso l'outline è fissata da chi commissiona il progetto. La fonderia fornisce le informazioni per programmare l'ambiente di layout, fornendo, ad esempio, i colori delle zone di sovrapposizione tra i vari livelli.

Flusso di progetto

Con flussi di progetto intendiamo le serie di operazioni che ci portano da un'idea a un qualcosa che sia pronto ad essere realizzato.



Flusso di progetto analogico

Il flusso di progetto analogico è strettamente basato sul concetto di netlist. La netlist, “lista di nodi”, è la descrizione della rete elettrica che definisce l’oggetto della progettazione in termini di componenti e interconnessioni. La netlist di un circuito può essere costruita graficamente come schematico (schematic editor) oppure testualmente (text editor). La progettazione analogica è principalmente manuale, umana, artistica in un certo senso. Le istruzioni per portare avanti il flusso analogico sono costituite da poche regole generali e ciò porta ad una bassa standardizzazione del flusso: due progettisti analogici faranno due oggetti dalle stesse specifiche in modo diverso. Il prodotto di questo flusso è la cella analogica: l’intero chip, un ADC, un convertitore corrente-tensione o una porta logica stessa.

La metodologia con cui si sviluppa il flusso di progetto analogico si applica anche per le standard cells, ovvero celle digitali elementari: inverter, porte logiche, fino a full adder, flip flop. I circuiti analogici sono circuiti in cui l'informazione è attribuita all'intero andamento dei segnali in tensione o in corrente, che variano con continuità coprendo un numero infinito di valori tutti significativi all'interno di una certa dinamica. In realtà la presenza del rumore fa sì che anche un segnale analogico sia intrinsecamente quantizzato rispetto la capacità di trasportare informazione. Nel circuito digitale invece le grandezze in gioco vengono mappate su un insieme di valori significativi discreto, caso tipico quello binario. Anche la cella digitale in sé è analogica, ma si considerano solo tre stati rispetto le grandezze di uscita (1,0, indeterminato). Nel caso dell'elettronica analogica la grandezza elettrica rappresenta un'informazione con analogia rispetto la fonte primaria. Ad esempio, si potrebbe costruire un circuito la cui tensione di uscita sia proporzionale (analogica) alla temperatura. Un segnale digitale, invece, potrebbe rappresentare la temperatura soltanto a posteriori di una certa codifica, un'intermediazione, per cui senza analogia diretta.

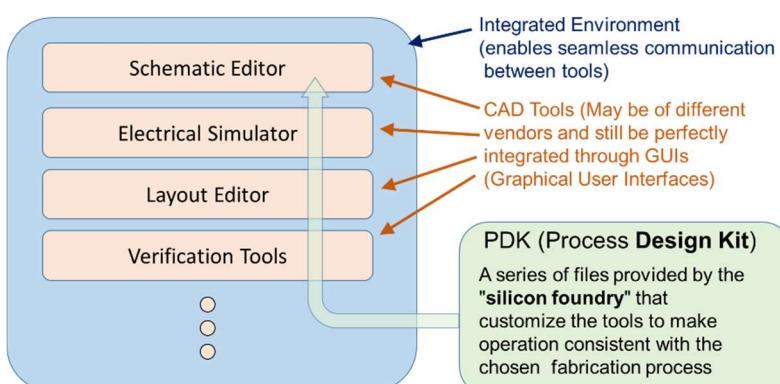
Flusso di progetto digitale

Il flusso di progetto digitale è basato su linguaggi HDL (Hardware Description Language) attraverso i quali si descrive la funzione da integrare su silicio con una sintassi specifica. I linguaggi HDL permettono anche di poter descrivere il circuito come insieme di primitive interconnesse tra loro, ma l'approccio più potente è quello di descrivere l'hardware in modo funzionale, specificandone il comportamento ad alto livello. In HDL si possono descrivere con facilità macchine a stati, processori, DSP, microcontrollori, protocolli di tlc, intelligenze artificiali. Il tutto è automatizzato e fortemente standardizzato tramite l'impiego di librerie, paradigmi precisi e di uso comune. Il flusso di progettazione digitale permette meno libertà al progettista, ma rimane comunque un certo grado di creatività. La descrizione HDL si traduce poi nella rappresentazione fisica del circuito digitale in termini di celle fondamentali attraverso il tool di sintesi logica automatica. A parità di funzionalità il layout risultante è pertanto predicibile, fissato, fortemente standardizzato. Questo fa sì che il flusso di progetto digitale debba servirsi delle standard cells realizzate nel flusso di progetto analogico e raccolte in un'apposita libreria.

Flusso mixed signal

Per comporre un circuito mixed signal, cioè con parti analogiche e parti digitali, i due flussi devono essere opportunamente combinati. A tal proposito esistono due possibili modalità. Con la modalità analog centric (analog on top) la composizione del circuito finale avviene con gli stessi strumenti CAD utilizzati per il flusso analogico. Se invece si opta per una modalità digital centric (digital on top) il progettista mixed signal assembla le parti analogiche e digitali con gli strumenti utilizzati nel flusso di progetto digitale. La scelta dipende sia dalla proporzione analogico/digitale del circuito finale e dalla preferenza del gruppo incaricato dell'integrazione finale.

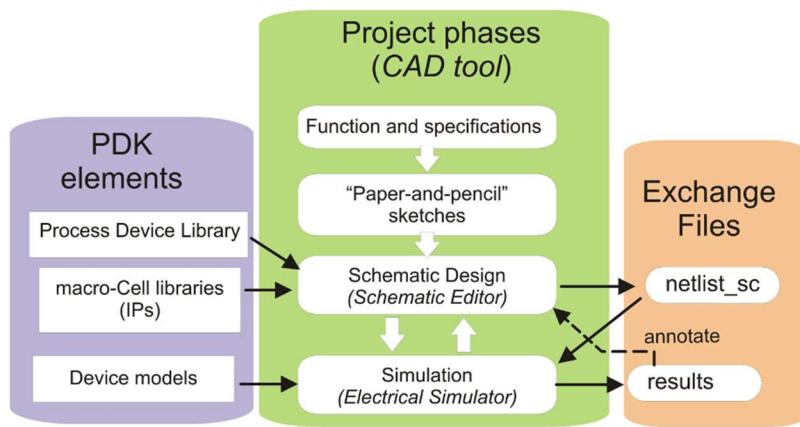
EDA



L'ambiente di progettazione viene spesso indicato come EDA (Electrical Design Automation Environment). Si tratta di un contenitore all'interno del quale si trova una collezione di strumenti con i quali si effettuano le operazioni di progettazione durante le varie fasi. Tali strumenti possono essere prodotti anche da aziende diverse, ma devono poter essere integrati nell'EDA in modo trasparente, fluido, senza interruzioni (seamless).

L'ambiente leader nel campo dei CAD per la progettazione analogica si chiama Cadence (San Giovanni Vincentelli). Gli strumenti contenuti nell'EDA devono essere personalizzati, indirizzati alla tecnologia di riferimento. Questo viene fatto per mezzo di file specifici riuniti in un PDK (Process Design Kit) forniti dalla fonderia. Ciò permette di poter progettare il circuito con le caratteristiche fisicamente implementabili nella tecnologia scelta.

Electrical design



La colonna centrale rappresenta le fasi del progetto e gli strumenti impiegati. Qualsiasi progetto analogico parte dalla funzione e le specifiche dell'oggetto. La funzione è una descrizione qualitativa e non quantitativa. Si può ad esempio identificare la funzionalità op amp con uscita single ended, ingresso differenziale e alto guadagno. Le specifiche caratterizzeranno lo stesso oggetto in modo quantitativo, con dei valori. Ad esempio, potremmo avere una specifica sul minimo guadagno accettabile. Le specifiche saranno date in forma di disequazioni o uguaglianze. Una volta stabilite le specifiche (consumo, ingombro, velocità, ecc.) il passo successivo è il design carta e penna. Si creano i primi abbozzi di schematici con conti di massima. Una volta che l'idea è caratterizzata in modo ragionevole, si può passare allo schematico elettrico.

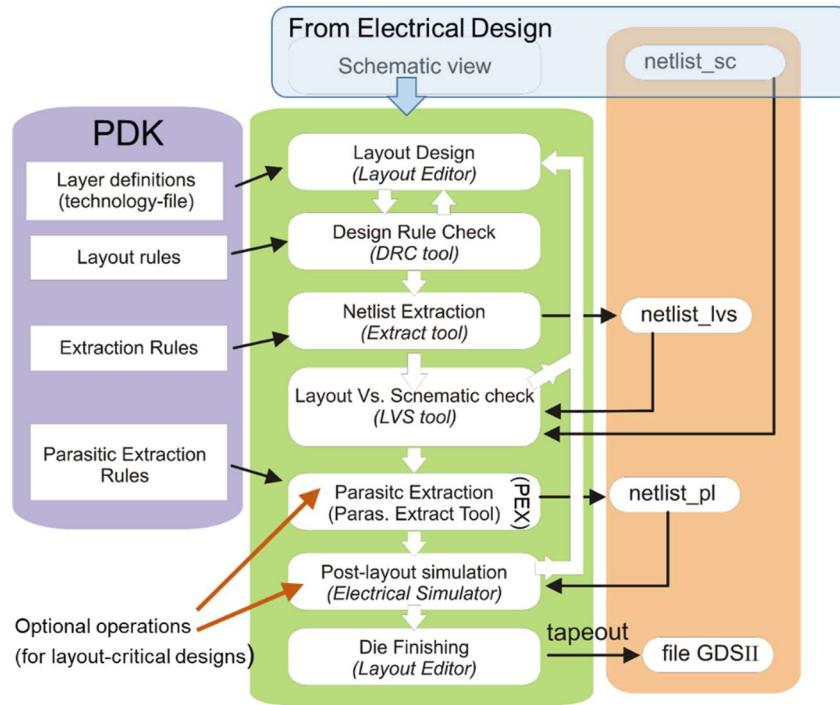
Il prodotto dello step di progettazione elettrica è lo schematico, che dovrà essere compatibile con il processo scelto. La compatibilità è garantita dal fatto di prelevare i componenti esclusivamente dalla libreria dei componenti di processo fornita nel PDK. In un processo CMOS avremo i transistori a canale n, a canale p, transistori a tensione di soglia ribassata, transistori ad alta tensione (ossido più spesso), resistori, condensatori. Compiuto lo schematico o alcune porzioni dello stesso si può passare a una verifica attraverso i tool di simulazione. Gli schematic editor di ambienti di progettazione professionali come Cadence mettono a disposizione anche componenti puramente ideali (generatori comandati, generatori indipendenti) che servono per arricchire la simulazione del circuito integrato, modellare i sistemi che interagiscono con il circuito integrato dall'esterno (es. capacità di uscita del carico).

In questo contesto gli strumenti di progettazione regolano la scelta del valore dei componenti, valore limitato dalla tecnologia di integrazione. Chi realizza il design kit può automatizzare il controllo sulla conformità dei valori specificati, prevedendo la comparsa di un errore nel caso in cui si scelgano valori non conformi al processo. Il simulatore avrà bisogno di un modello che descriva il comportamento dei dispositivi (effetto body, canale corto, canale stretto, ecc.).

Anche se l'ambiente di progettazione appare contiguo tutte le operazioni avvengono attraverso uno scambio di file nascosto. Per effettuare la simulazione, ad esempio, è necessario estrarre la netlist testuale dallo schematico (in linguaggio Spice). La fase del progetto elettrico si conclude quando lo schema elettrico è completamente disegnato, i transistori sono completamente dimensionati, le verifiche soddisfano le specifiche elettriche e si estrae la netlist. Uno schematico elettrico che rispetta le regole di processo è implementabile come layout nelle fasi successive del flusso di progetto.

Physical Design

La vista schematico procede entrando nel flusso alla progettazione fisica. Il layout editor è un ambiente di progettazione puramente grafico, vincolato in risoluzione, in forme e layer dal processo tecnologico.



All'interno del layout editor si disegnano i rettangoli dei layer sovrapponendoli in modo tale da codificare le operazioni di processo necessarie per realizzare su silicio il circuito descritto nello schematico. È importante controllare frequentemente la correttezza del layout rispetto le regole di layout con il tool di **DRC** (Design Rule Check). Il DRC segnala tutti gli errori di layout commessi nella progettazione fisica di una cella. Le regole di layout sono fissate dalla tecnologia e sono costituite principalmente da indicazioni di carattere geometrico rispetto le dimensioni assolute e le posizioni relative tra i vari layer. Tali indicazioni sono specificate all'interno del manuale di processo che, per processi spinti, può anche arrivare a 150-200 pagine. Si può programmare l'editor di layout per forzare le geometrie entro le regole, ma è meglio optare per la filosofia del warning. Oltre alle regole di layout il PDK specifica i layer che si possono usare, cioè quelli che il processo selezionato mette a disposizione, il nome delle varie aree e altre direttive inerenti al layout.

Il DRC controlla esclusivamente il rispetto delle regole di layout, la conformità rispetto alla tecnologia, ma non la corrispondenza elettrica tra il layout e lo schematico. Potrebbe esserci una non congruenza tra lo schematico ed il layout (ad esempio potrebbe mancare una connessione metallica laddove nello schematico era presente un filo). Errori di questo tipo, molto comuni, non sono di competenza della fonderia. Lo strumento che ha a disposizione il progettista per proteggersi da problematiche di questo tipo è l'**LVS** (Layout Versus Schematic). Si tratta di un tool che confronta il layout con lo schematico. In particolare, ad essere confrontate sono due netlist, una estratta dallo schematico, (netlist_sc) già disponibile dal flusso precedente, una estratta dal layout con un apposito tool estrattore (netlist_lvs). L'estrattore riconosce i componenti nel layout analizzando gli incroci e le dimensioni dei vari layer, nonché i drogaggi, extrapolando così i MOSFET, i resistori, i BJT. Inoltre, analizzando i contatti, riconosce i collegamenti tra vari componenti (ideali). Anche se le sorgenti di estrazione sono diverse, se il circuito è elettricamente equivalente tra le due viste, le netlist dovrebbero contenere la stessa informazione descrittiva della rete (componenti, nodi, collegamenti). Tuttavia, le due netlist non sono equivalenti a livello testuale; tra le due viste, anche se la rete elettrica rappresentata fosse la stessa, potrebbero esserci notevoli differenze a livello di descrizione, nomi, ecc. Questo fa sì che non sia possibile condurre un semplice confronto carattere per carattere tra i due file. L'LVS è in effetti un programma complesso che poggia su algoritmi di intelligenza artificiale, altamente "crittografato". Lo strumento LVS, oltre che controllare la corrispondenza topologica tra schematico e layout, controlla anche che i valori dei componenti siano congruenti (mismatch sui valori).

Ad essere confrontati per primi tra netlist_sc e netlist_lvs sono quei componenti equivalenti per caratteristiche, sia in termini di tipologia che di contesto relativo al resto del circuito, ad alto grado di complessità (es. alto numero di connessioni). Procedendo per confronti il programma scomponete la rete per graduale semplicità in classi, fino a che la classe si riduce ad una corrispondenza biunivoca tra due componenti. Lo stesso viene fatto per i nodi. Nel caso in cui l'LVS segnali una non equivalenza tra le due viste non viene fornita un'istruzione dettagliata per la risoluzione del problema. Una buona prassi è quella di iniziare a verificare una serie di congruenze elementari, come ad esempio il numero di nodi, la quantità di dispositivi. Un'incongruenza tra il numero di nodi può suggerire la presenza di cortocircuiti o collegamenti mancanti.

Superata la fase di LVS, non è detto che il circuito progettato si comporti in modo coerente rispetto allo schematico e alle specifiche di partenza. Questo perché l'estrazione della netlist_lvs non considera gli elementi parassiti. Ad esempio, le piste di metal danno luogo ad una moltitudine di capacità verso altri nodi, induttanze mutue, una resistenza e un'autoinduttanza parassita. Per conoscere il comportamento più vicino alla realtà del circuito occorre estrarre la rete che comprende anche gli elementi parassiti (netlist_pl – post layout).

Un'estrazione di questo tipo non potrebbe avvenire in fase LVS; la netlist ricavata non sarebbe più confrontabile con quella di schematico. I parassiti relativi ai dispositivi sono già compresi a livello di schema elettrico, poiché descritti nel modello dei dispositivi. Tuttavia, a tale livello è impossibile valutare i parassiti delle interconnessioni. Non conoscendo la loro implementazione fisica, determinata soltanto in fase di layout, l'estrazione dei parassiti delle interconnessioni può esser fatta soltanto a posteriori. L'estrazione PEX (Parasitic Extraction) non passa da alcun ulteriore confronto, si tratta di un file di tipo Spice pronto per essere simulato. La simulazione comprendente i componenti passiti del circuito viene fatta anche sui vari corner del processo per tener conto delle variabilità e dei mismatch tra i dispositivi (simulazione montecarlo).

Ci sono due motivi per evitare di estrarre i parassiti:

1. Complessità: la PEX è un passo molto pesante. Da un semplice specchio di corrente potrebbero essere estratti almeno un centinaio di componenti parassiti. Con l'esperienza si capisce se il circuito può funzionare senza questo ulteriore controllo
2. Mancanza di regole di estrazione PEX: qualche volta la fonderia non inserisce nel PDK le regole di estrazione dei parassiti. In particolare, potrebbero essere assenti le regole che riguardano i parassiti delle interconnessioni (i modelli dei componenti tengono già conto delle non idealità a livello di schematico, anche al variare del punto di riposo, dimensioni, ecc.). Tali regole sono fondamentali poiché ciò che determina i parassiti dipende dal tipo di processo (resistenza di strato, spessore delle interconnessioni, ecc.)

Chi progetta le celle di libreria è obbligato a condurre l'estrazione dei PEX e simulare il circuito in post-layout poiché i tempi di propagazione e le varie caratteristiche devono essere passate al processo digitale come specifiche accurate. Con lo scaling delle tecnologie i dispositivi diventano sempre più piccoli, i parassiti delle interconnessioni sempre più importanti. Oltre un certo livello di scaling diventa quasi indispensabile estrarre i parassiti e simulare il layout post-PEX.

Il flusso di progetto termina con la produzione di un file in formato standard (GDSII – Graphic Design System) che contiene le informazioni del layout per la produzione del chip. Il file editor è in grado di leggere un file .gds, oltre che a produrlo. Questa fase prende il nome di **tapeout**, termine che indica un retaggio dei vecchi layout, i quali venivano caricati da una workstation su una cassetta a nastro da spedire alla fonderia.

A volte si può ricorrere alla generazione automatica: in tal caso i componenti dello schematico vengono automaticamente sintetizzati nel layout editor. Tuttavia, il layout analogico è talmente personalizzato (aspetti di simmetria) che un rooting automatico può portare a malfunzionamenti, ingombri maggiori e peggiori comportamenti in temperatura, in frequenza. Meno critico è il rooting automatico nel processo digitale, in cui ad essere interconnesse sono unità standardizzate che offrono un livello di astrazione più elevato.

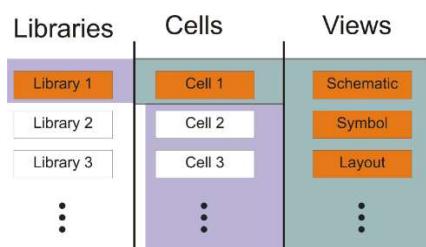
A chip concluso vengono compiute le operazioni di die finishing. Ad esempio, si inseriscono i dummies, delle porzioni di metal o di polisilicio isolate che vengono aggiunte con un tool automatico per far sì che la densità di alcuni layer sia costante sul chip. In questo modo si migliora l'uniformità del processo produttivo:

1. Uniformità della deposizione e dell'incisione: nei processi di fotolitografia, incisione chimica e deposito del metallo, la distribuzione non uniforme del materiale può causare problemi di planarità e difetti di fabbricazione. I dummies di metallo aiutano a mantenere una distribuzione più omogenea del materiale, riducendo variazioni nello spessore.
2. Controllo della planarità (CMP - Chemical Mechanical Planarization): durante la planarizzazione chimico-meccanica, la rimozione del materiale avviene in modo più uniforme se la superficie è bilanciata. Senza dummies, le aree con meno metallo possono subire una maggiore rimozione di materiale rispetto a quelle più dense.
3. Riduzione dello stress meccanico e termico: differenze nella densità di metallo possono causare stress termico durante il raffreddamento e il riscaldamento del wafer, portando a deformazioni o crepe. I dummies aiutano a ridurre queste differenze, migliorando l'affidabilità del circuito.
4. Miglioramento del rendimento di produzione: una distribuzione più uniforme del metallo e dei processi produttivi riduce la probabilità di difetti, migliorando la resa del wafer.

I dummies di metallo non sono connessi elettricamente al circuito, ma servono esclusivamente a scopi di stabilizzazione del processo produttivo. L'inserzione di porzioni di metal comporta accoppiamenti capacitivi non previsti, per cui può rivelarsi necessario estrarre i parassiti soltanto dopo aver inserito i dummies. Si possono aggiungere anche dummies a mano eventualmente ancorati a ground.

Altre operazioni sono finalizzate al taglio del chip. Con gli stessi processi produttivi vengono realizzate delle strutture che circondano il chip (guard ring) per evitare che il taglio del die propaghi delle crepe al circuito integrato.

Struttura di un progetto

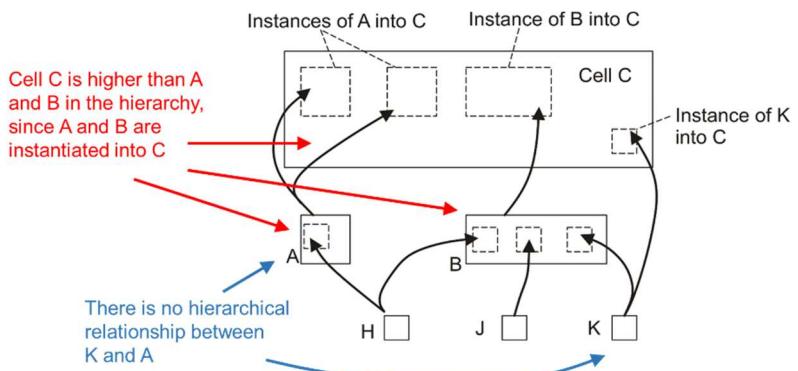


Le librerie sono delle collezioni di celle sottoforma di cartelle. Installato il PDK compariranno le librerie di processo (librerie tecnologiche con i dispositivi) e altre con layout prefabbricati (ad esempio per i bonding pad). Si ha anche una libreria di componenti ideali (generatori, analog lib., OP amp ideali). Altre librerie potrebbero contenere lavori altrui, circuiti base quali MUX analogici, OP amp, ecc. Se si tratta di un progetto digitale avremo le librerie di standard cell (NAND, NOT, OR, ecc.).

Il progetto stesso sarà una libreria, con celle progettate da noi. Una delle celle sarà la cella top e conterrà tutte le altre a gerarchia inferiore. A volte la struttura di un progetto comprende anche un altro campo, le categorie. Se abbiamo tanti tipi di MOSFET, resistori, condensatori, può far comodo raggrupparli in categorie a sé, senza le quali verrebbero organizzati in ordine alfabetico nel campo delle celle.

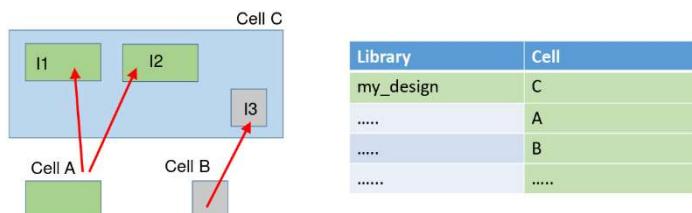
Gerarchia

La gerarchia è uno dei concetti fondamentali alla base di qualsiasi flusso di progettazione. Qualunque processo deve avere un'impostazione gerarchia. L'obiettivo finale della progettazione potrebbe essere quello di sintetizzare una cella da vendere come IP, un chip intero, un ADC. Il naturale approccio con cui si affronta la questione è quello di suddividere il problema in blocchetti funzionali più piccoli. Ad esempio, per la progettazione dell'ADC possiamo prima realizzare l'opamp, che nel progetto complessivo sarà condensato a un simbolo. Si scende di gerarchia per entrare nel triangolo, fino al singolo transistore. Un circuito complesso, un chip intero o magari una singola cella, ha tanti livelli di gerarchia, tante sotto-celle.

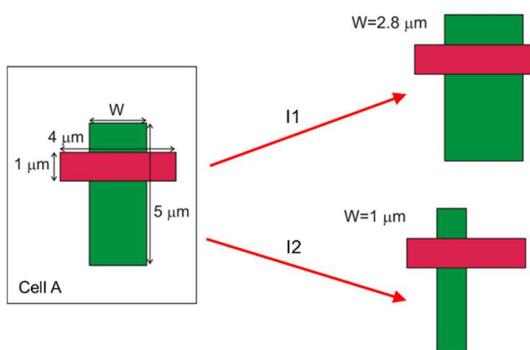


Si rappresentano le celle come rettangoli. Come si osserva, una cella può essere sintetizzata come insieme di altre celle. L'operazione con cui si preleva una cella più semplice e la si inserisce in una più complessa si chiama istanziazione; la cella, di pari passo, si chiama istanza. Istanziare più volte la medesima cella si riduce a un copia e incolla.

La gerarchia nasce nel momento in cui si istanzia una cella in un'altra sale automaticamente di livello gerarchico. Se due celle rimangono separate, anche se una è più complessa dell'altra, non c'è alcuna relazione gerarchica. Tutto ciò è presente già nelle prime fasi del flusso di progetto, a livello di pre-schematico, in una sorta di schema a blocchi mentale. Il programma non crea una struttura ad albero. In relazione all'esempio, la cella A e la cella B possono essere contenute da più celle diverse. Se non ci fosse K ad essere istanziata direttamente in C, quella in figura sarebbe una struttura a livelli. Non avendo A alcuna istanza di K e J al suo interno, tra A, J e K non c'è alcuna relazione di gerarchia (non è vero che K è ad un livello di gerarchia inferiore rispetto ad A). Anche se non ci fosse l'istanziazione diretta di K in C, K sarebbe comunque gerarchicamente inferiore a C poiché è istanziata in B, e B è a sua volta istanziata in C.



La gerarchia non è visibile nella libreria di progetto. Le celle appaiono come se fossero tutte allo stesso livello gerarchico. Un'altra proprietà delle istanze da tenere presente è che, come per l'esempio di sopra, due istanze di una stessa cella (I1 e I2) sono da considerarsi oggetti distinti, individuali. Anche a livello elettrico, nonostante rappresentino lo stesso circuito, avranno tensioni e correnti diversi. Tuttavia, l'effetto di una modifica della cella madre A (alla quale si può accedere selezionando qualunque sua istanza) si propaga a tutte le sue istanze. Questa è una limitazione. Potremmo aver bisogno di cambiare le caratteristiche di un'istanza della cella senza applicare la solita modifica a tutte le altre istanze. Spesso ciò impone di creare una nuova cella duplicando quella di base ed apportando le modifiche desiderate. Ci sono casi, però, in cui tale procedimento darebbe luogo ad un'eccessiva ridondanza. Ad esempio, se le celle sono dei transistori è impensabile creare ogni cella per tutti i diversi possibili dimensionamenti del medesimo transistor.

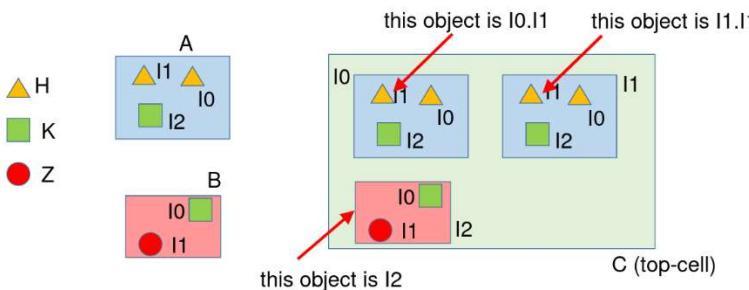


Con le **celle parametriche** (p-cell) è possibile ovviare a questo problema. La cella parametrica è una cella particolare che permette di differenziare le istanze che se ne fanno. Nella fase di progetto le istanze mantengono un certo legame con l'istanza madre. Cambiare qualcosa in un'istanza comporta cambiamenti in tutte le altre, ma questo in realtà succede perché la modifica, intrinsecamente, penetra nella cella madre, con conseguente modifica delle celle figlie. La corretta modifica selettiva delle celle figlie si ottiene soltanto per mezzo della parametrizzazione, spesso utilizzata per celle elementari (dispositivi).

Ad esempio, si può pensare di fare una instance parametrica di un MOSFET a livello di layout, la cui dimensione della larghezza di canale sia associata al parametro simbolico W. Al momento dell'istanziazione si specificherà il valore del parametro e l'istanza del MOSFET sarà creata con la larghezza di canale specificata. In questo modo è possibile creare più istanze personalizzate della stessa cella parametrizzata (p-cell).

La parametrizzazione delle celle è possibile sia nella vista layout che nella vista schematico. Si tratta di una soluzione meno adatta per celle complesse.

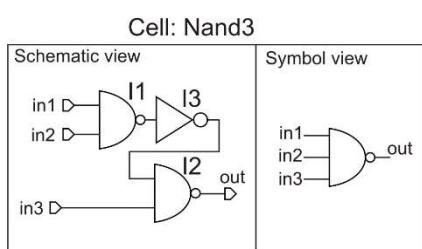
Terminologia



Tale organizzazione terminologica si rivede sia a livello di schematico che a livello di layout. Occorre sempre rispettare la gerarchia: mentre è possibile istanziare la cella Z nella cella K, non è possibile istanziare la cella B nella cella K poiché tra le due esiste un legame gerarchico. Con questa terminologia ogni oggetto ha un nome diverso.

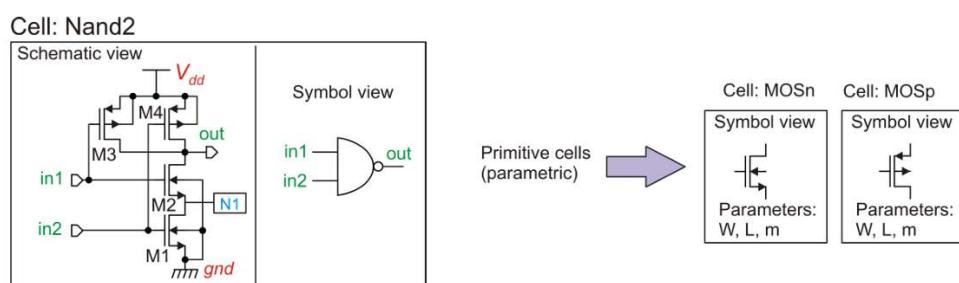
Schematic view

La vista schematico è una rappresentazione delle celle come interconnessioni di celle più semplici. Le celle che non possono essere decomposte in celle più semplici non sono associate a una vista schematico, e prendono il nome di celle primitive, o semplicemente **primitive**.



In questo esempio si esprime la cella NAND a 3 ingressi come interconnessione di tre istanze di celle più semplici. Per istanziare questa porta in un circuito a gerarchia superiore occorre che la cella complessiva abbia un suo simbolo, editabile in un symbol editor (symbol view), che a livello elettrico esprima soltanto i nodi utilizzati per interconnettere la cella stessa con il resto del circuito.

Scendendo di gerarchia, la NAND a due ingressi sarà descritta come interconnessione di transistori.

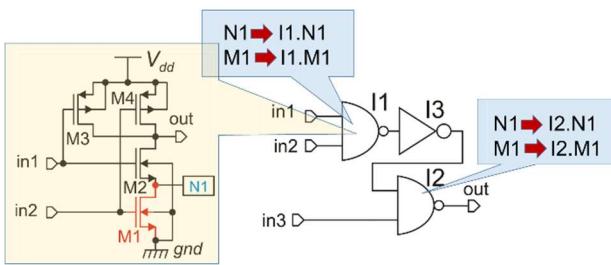


Le celle che compaiono nella schematic view, questa volta, sono le primitive costituite dai dispositivi elettronici; il MOSFET non può essere descritto come interconnessione di oggetti più semplici. Questa particolarità si riconosce nel momento in cui si naviga nella libreria e per i transistori troviamo soltanto la vista simbolo o la vista per la simulazione (vista Spectre), ma non quella schematica. Qualche volta può essere utile rappresentare anche i dispositivi come interconnessione di oggetti più semplici, ad esempio per modellare i diodi di substrato o i bipolar parassiti.

Symbol view

La vista simbolo costituisce il simbolo della cella e permette di poterne creare un'istanza a gerarchia superiore, che nasconde l'architettura interna e sia solo rappresentativa dei terminali della cella. Di default il simbolo di un'istanza custom è rettangolare. È comunque possibile modificare l'aspetto dei simboli per rendere più chiara ed intuitiva la funzionalità di quella cella.

Nodi



L'accesso ai nodi interni è regolato da una sintassi particolare basata sull'uso del separatore punto. Ogni entità, seguendo una filosofia gerarchica di assegnazione dei nomi, è univocamente identificata.

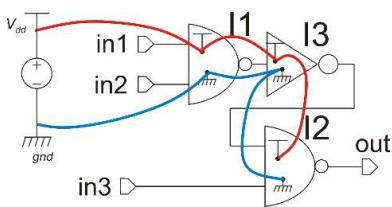
Con **nodo** intendiamo una regione di spazio equipotenziale (net). A livello elettrico la definizione è completa e non distingue diverse tipologie di nodi. A livello di schematico, tuttavia, si differenziano tre diverse tipologie di nodo:

- **Nodi terminali**: servono per interconnettere la cella a un livello gerarchico superiore, dall'esterno tramite un pin (triangolino)
- **Nodi globali**: valgono per qualsiasi istanza a tutti i livelli gerarchici, senza la necessità di aggiungere terminali per specificarne la connessione (es. Vdd, gnd). Anche se non vengono rappresentati graficamente al livello di gerarchia superiore, sono comunque presenti e condivisi globalmente. Nell'ambiente Cadence i nodi globali sono evidenziati dal punto esclamativo!
- **Nodi interni**: tutti gli altri, quei nodi che non sono né terminali né globali. I nodi interni hanno una visibilità solo internamente alla istanza della cella. Una volta istanziata la cella i nodi interni non sono accessibili. Lo sono soltanto in fase di simulazione, in cui possono essere analizzati e nominati.

I nodi che si selezionano come terminali (pin) per far comunicare la cella NAND2 con altre celle a livello gerarchico superiore, in questo caso, sono i nodi in1, in2, out. Quando si userà la cella in un progetto a livello gerarchico superiore, nella vista simbolo questi nodi coincideranno ai pin (input, output, bidirectional, ...) che supportano le interconnessioni. La simulazione non produce errori nel caso di conflitti elettrici tra i pin interconnessi, ma soltanto warning (es. out-out -> warning di corto circuito).

I nodi interni di solito vengono automaticamente nominati dal programma. Bisogna sempre tenere presente che, a meno di non assegnare tramite etichetta un nome specifico a un nodo interno, il programma potrebbe cambiarne il nome a seguito di modifiche della rete. Se un determinato nodo interno è di particolare interesse o per ragioni di simulazione o perché è associato ad una moltitudine di rami, è bene assegnargli un nome personalizzato. Questo permette, oltre che una facile identificazione nell'elenco nodi e una protezione dall'eventuale riassegnazione del nome, di poter collegare due nodi semplicemente ripetendo la label nella vista schematica.

I nodi globali, in questo caso Vdd e gnd, non compaiono nel simbolo.



Se gnd e Vdd delle celle sono trattati come nodi globali, in fase di simulazione sarà sufficiente assegnare ai capi di un generatore di tensione gli stessi nodi, senza necessità di portare i fili ai diversi livelli di gerarchia. A volte si usa un nodo globale per il clock nei circuiti digitali. In analogico spesso non si sfrutta la proprietà dei nodi globali; creando terminali appositi per Vdd, ad esempio, si può monitorare il consumo locale delle singole celle.

Il parametro *m* che figura nella vista simbolo dei MOSFET identifica la **molteplicità**. Con *m* = 2, ad esempio, il transistore è un equivalente di due transistori dello stesso tipo in parallelo. Collocare componenti attivi e passivi in parallelo sarà utile come capiremo più avanti.

Elementi del layout editor

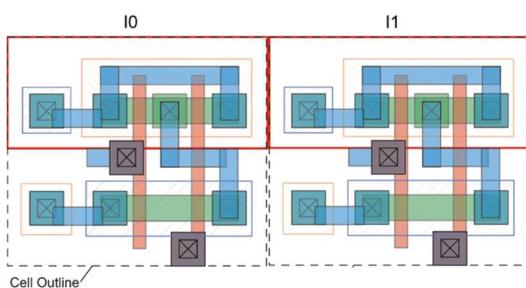
Gli oggetti che si gestiscono con il layout editor, fondamentalmente, sono forme geometriche. Una volta definita la forma di un oggetto se ne definisce la qualità elettronica tramite layer. I layer messi a disposizione dal design kit sono riportati nella tavolozza layer e si raggruppano in tre tipologie principali:

- Layer tecnologici (“tooling layers”): corrispondono ad operazioni fisiche che verranno riportate direttamente su silicio durante il processo di fabbricazione. Possono essere classificati come layer “drawing purpose”
- Layer derivati: sono il risultato di operazioni logiche tra altri layer. La regione di spazio più importante in un MOSFET è l’intersezione tra l’area attiva e il poly. Tale intersezione comporta “automaticamente” la formazione del gate, che può essere visto come layer derivato *area attiva AND poly*. I layer derivati offrono aiuto a tutto il comparto di progettazione, semplificando ad esempio le operazioni di estrazione. Le regole di layout a volte sono espresse proprio in funzione di questi layer.
- Layer di servizio: si impiegano per aspetti secondari, ad esempio per fare annotazioni, marcare oggetti come pin, definire il perimetro delle celle (outline), sollevare il DRC dal controllo di certe aree volutamente fuori regola, impedire la formazione automatica di dummies. I layer di servizio non si traducono direttamente in architetture su silicio, ma automatizzano certe procedure, controlli, in fase di progettazione.

Layer Palette	
■	Active Area
■	Polysilicon
■	N-Well
■	N-Plus
■	P-Plus
☒	Contact
■	Metal 1

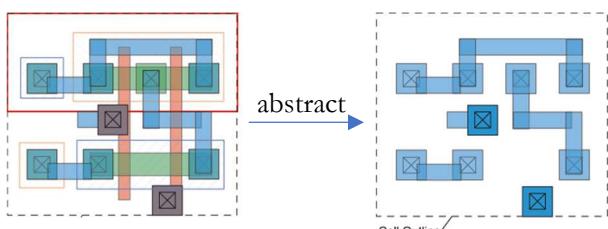
Istanze nel layout editor

L’istanza nello schematic editor avveniva per mezzo del simbolo. Non esiste un analogo del simbolo nel caso del layout. Vale comunque l’impronta gerarchica del progetto, per cui si faranno istanze di celle più semplici, ma queste verranno incollate così come sono, con tutti gli elementi del layout che le definiscono. Nel caso del layout, in effetti, è importante avere una visione completa dei layer onde evitare di violare le regole di DRC o di intersecare altri layer involutamente. Esistono particolari modalità, esclusivamente ai fini della visualizzazione, in cui si può temporaneamente nascondere l’interno del layout e lasciare solo l’outline quando altrimenti il circuito sarebbe troppo complesso.



Le istanze a livello di layout possono essere connesse al resto del circuito, spostate e ruotate, ma non possono essere modificate. A questo livello di gerarchia le interiora dell’istanza non sono interattive, sono fisse, si trovano a livello di gerarchia più basso. La gerarchia del layout è di fondamentale importanza: aiuta a progettare, a concettualizzare e permette di estendere una modifica a tutte le celle della stessa origine.

Abstract view

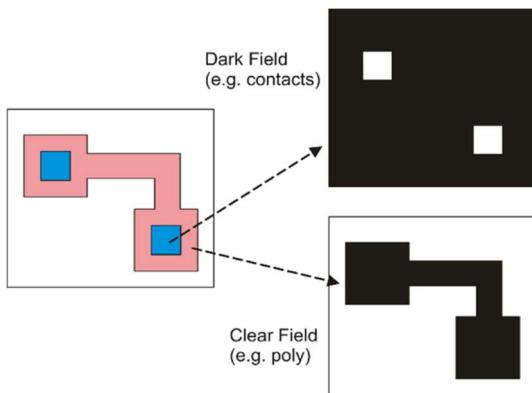


La vista abstract è una particolare vista layout riassuntiva, in cui si nascondono certi layer della cella. Viene utilizzata da chi vende IP per protezione dal plagio. All’interno della vista astratta è visibile soltanto il confine della cella, i contatti e i livelli di metal usati per le interconnessioni.

L’azienda che detiene l’IP fornirà la vista layout completa alla fonderia sotto un accordo commerciale e l’abstract view al cliente, con tanto di istruzioni per la simulazione. Per ogni cella inserita nel chip in vista astratta dal cliente progettista, la fonderia farà una sostituzione con il layout completo fornito dall’azienda. La presenza di celle in vista astratta limita l’LVS, che potrà esser portato avanti solo in modo parziale. Un altro utilizzo valido della vista astratta è quello di semplificare le operazioni di place & rout.

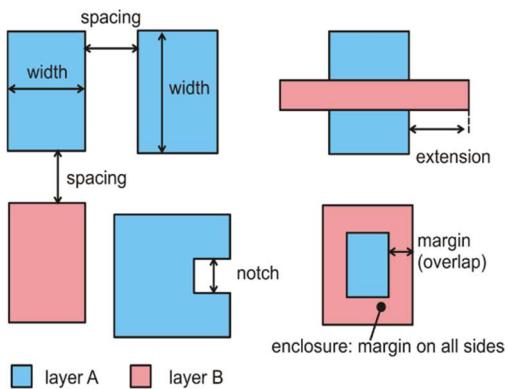
Layers, shapes and masks

I layer tecnologici costituiscono delle forme, le quali corrispondono alle geometrie che vorremmo riportare direttamente su silicio. Per compensare gli effetti di bordo nella fase di impressione litografica e la disomogeneità dell'attacco chimico nelle varie fasi del processo tecnologico, la geometria di una forma non si riporta tale e quale nella corrispondente geometria sulla maschera. La fonderia, tramite software appositi, estrae il set di maschere compensate a partire dal layout e le commissiona all'azienda specializzata nella realizzazione fisica delle maschere. Ciò che interessa al progettista non è la corrispondenza tra layout e maschera, ma tra layout e silicio. Talvolta può accadere che una forma dia luogo a più di una maschera. Oppure, forme costituite da più layer possono essere condensate in un'unica maschera o generare maschere differenti attraverso operazioni logiche (AND, OR, ...); le operazioni con cui si ricavano le maschere a partire dai layer si raccolgono sotto le “mask merging rules”. Un esempio è il caso in cui abbiamo bisogni di realizzare i due drogaggi complementari di tipo n+ e d i tipo p+. Per realizzare il droggaggio p+ si può omettere il droggaggio n+ nell'area attiva e lasciare che la fonderia ricavi la maschera p+ con l'operazione logica “NOT n+”. Oppure, può essere reso disponibile un layer che esclude ogni tipo di droggaggio (per realizzare resistori). Possiamo distinguere i layer in due categorie: dark field layer, clear field layer.



Le geometrie realizzate con un layer di tipo dark field verranno riportate in trasparenza sulla maschera, mentre la geometria complementare, lo sfondo, sarà scuro e impedirà il passaggio della luce durante la litografia. Un layer è di tipo clear field se, al contrario, costituisce quelle geometrie che nella maschera saranno opache; la geometria complementare sarà trasparente. Se si devono realizzare dei buchi nell'ossido di silicio, la tecnica è quella di proteggere l'isolante ovunque tranne in corrispondenza dei buchi, in cui il fotoresist verrà rimosso e l'attacco penetrerà nel silicio. Anche questo è un aspetto che interessa poco al progettista, più ai tecnologi.

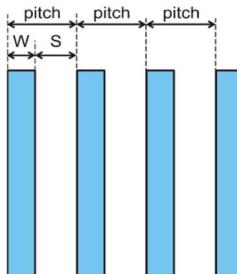
Regole di layout topologiche (TLR)



Le regole di layout forniscono delle limitazioni di carattere geometrico alle forme che si possono disegnare, in accordo con i limiti del processo tecnologico. Tali regole garantiscono che il layout sia fabbricabile e che lo sia con accuratezza tale da rispettare la funzionalità che ha. Rompere le regole di layout ha l'effetto certo di comportare un esito negativo del DRC e un effetto più aleatorio: la non garanzia che le forme si trasferiscano coerentemente sulla maschera, ovvero la non garanzia che la funzionalità progettata su layout si mantenga su silicio. La fonderia può comunque accettare un progetto che non rispetti le regole di layout, ma soltanto con l'assunzione di responsabilità da parte del cliente.

Le regole di layout sono riportate su un apposito manuale. Gli aspetti geometrici regolati dalle TLR sono molteplici: minima larghezza (vincolo risoluzione), minimo spacing (margini per le tolleranze), spacing delle tacche, minima estensione della sporgenza dei layer a intersezione (se in un MOSFET l'area attiva non è completamente coperta il dispositivo non si può spegnere), margini, enclosure (margini su tutti i lati per l'oggetto che deve contenere un altro), ecc. Più si va verso tecnologie avanzate, con risoluzione spinte, più le regole e le eccezioni diventano complesse. Tecnologie più antiche come la 180 nm sono comunque ancora molto utilizzate dato il loro costo ridotto.

Una regola che si deduce come combinazione di regole di layout è il **pitch** o passo. Il passo è il minimo distanziamento con cui si può ripetere e accostare uno stesso oggetto, come il filo di un bus, una cella di memoria, un pixel, un transistore senza violare le regole di layout.

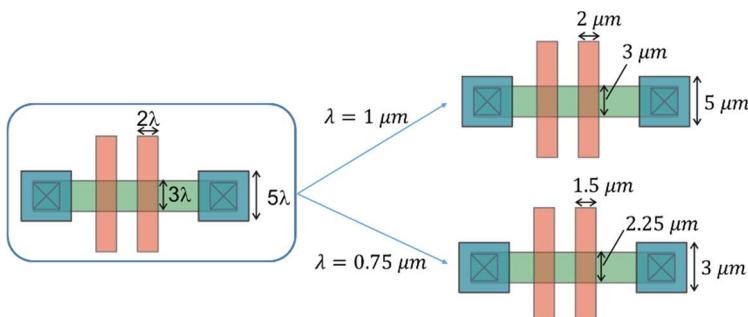


$$pitch = W + S$$

Immaginiamo di dover realizzare un bus a 64 bit. Per occupare il meno spazio possibile le linee dovrebbero essere larghe il meno possibile e lo spacing tra di esse il minimo possibile. Il pitch, così come tutte le altre regole di layout, è strettamente legato al tipo di processo con cui si realizza il progetto. L'estensione minima per un bus a n linee sarà data da $pitch_{min} \times n$, mentre il numero di linee massime in un certo spazio s sarà dato da $s/pitch_{min}$.

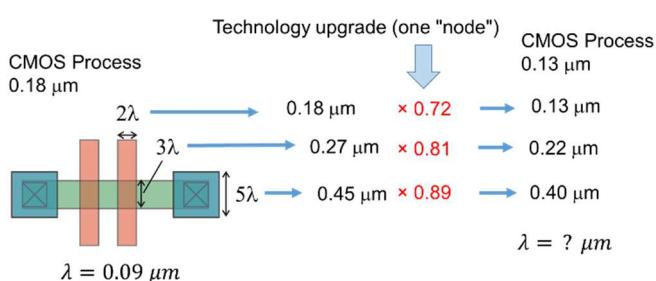
Le regole di layout possono essere fornite con due filosofie diverse:

1. **Micron rules**: in questo caso le regole sono fornite in unità di misura di lunghezza, tipicamente μm o nm , ed è questo lo standard industriale.
 2. **Lambda rules** (Mead and Conway): in questo caso le regole sono fornite in maniera adimensionale secondo multipli di una lunghezza elementare λ da valore noto.



Questo approccio è divenuto popolare negli anni '80 in cui sembrava che le regole di layout potessero scalare proporzionalmente al nodo tecnologico. Con questa filosofia è possibile adattare un layout a un processo più spinto semplicemente scalando il λ .

Le regole lambda erano pensate per il digitale e rendevano i progetti portabili.



Dal momento in cui le regole di layout hanno cessato di scalare uniformemente con il nodo tecnologico (a partire da $1 \mu m$ le regole hanno cominciato ad essere dettate da problemi fisici di altra natura rispetto alla litografia), l'utilizzo del λ si è ridotto. Si potrebbero salvare le regole lambda accettando un design a ingombro non ottimizzato, scalando tutte le TLR per il fattore di scala meno efficace, nell'esempio $\times 0.89$.

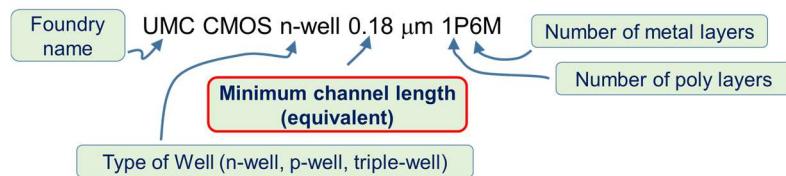
Esistono servizi particolari, come il servizio MOSIS, che permettono alle piccole imprese o università di accedere ai processi di varie fonderie e realizzare dei prototipi di chip a prezzi più contenuti, su wafer condivisi a più clienti (MPW: multi project wafers). MOSIS è basato sulle regole λ e mette a disposizione in modo pubblico le regole e i modelli dei dispositivi. Quando la tecnologia migliora il servizio aspetta anche più nodi tecnologici in modo tale da ritrovare un fattore di scala comune, magari stirando le regole in peggio, preservando così il vantaggio di poter scalare vecchi layout. In questo modo MOSIS mantiene la riservatezza delle vere regole di processo e fa sì che i progetti siano scalabili al costo di rinunciare alle caratteristiche più spinte teoricamente ottenibili dalla tecnologia.

Oltre ciò, quando i nodi tecnologici avanzano è possibile applicare uno shrinking della tecnologia. Ad esempio, quando la tecnologia è migliorata così tanto da passare dal 180 nm al 90 nm , la fonderia può scalare uniformemente il nodo 180 nm , magari a 160 nm (shrink factor), senza che il progettista lo sappia. Naturalmente, in tal caso tutti i modelli per le simulazioni devono essere aggiornati (esiste un parametro shrink factor), ma in modo del tutto trasparente il progettista continua a progettare a 180 nm .

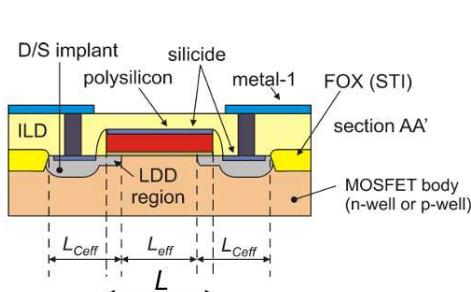
Dunque, la scalatura uniforme si applica ancora non per seguire l'evoluzione del nodo tecnologico, ma per migliorare un processo precedente prolungandone la sua compatibilità, senza dover riscrivere il design kit (di solito il DRC continua ad essere quello di prima) e senza comportare una riprogettazione da parte del progettista.

Ad ogni modo, nel mondo analogico non è detto che allo scalare delle regole di layout debba corrispondere un altrettanto scalare delle dimensioni del circuito. Mentre un circuito digitale beneficia sempre dalla miniaturizzazione, un circuito analogico come un amplificatore operazione rimpicciolito troppo funziona male (es. grandi offset). I processi analogici si basano ancora su processi planari (CMOS 28 nm), che supportano tensioni non troppo piccole, ecc.

Processo CMOS



Il processo CMOS è il processo fondamentale, il più utilizzato in tutte le sue sfumature. Fino a **28 nm** il processo era bulk planare, cioè, i dispositivi erano costruiti al di sopra del substrato di silicio. Si riesce a rimpicciolire ulteriormente (anche al di sotto di **18 nm**) mantenendo la planarità passando ad una tecnologia fully depleted SOI. Il flusso principale che ha portato a ridurre le dimensioni è stato quello di esser passati a dispositivi tridimensionali come i finFET e i GAAFET. I processi planari sono ancora usatissimi, soprattutto dai flussi di progettazione analogici. Il nodo tecnologico è un valore che rappresenta il grado di avanzamento della tecnologia. Fino a un certo grado di avanzamento tecnologico il nodo tecnologico rappresentava la lunghezza del canale. Il valore dà comunque un'indicazione di quanti dispositivi si riesce ad integrare in una determinata area.



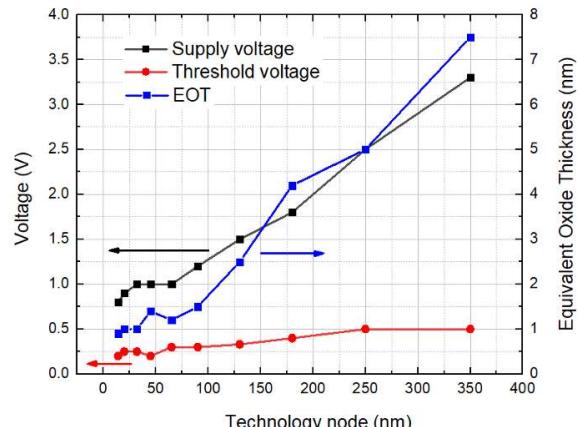
Year	Node (nm)	Half-pitch (nm)	Gate Length L_{eff} (nm)
1997	250	250	250
1999	180	239	140
2001	130	150	65
2003	90	90	37
2005	65	90	32
2007	45	68	38
2009	32	52	29

Con half pitch si intende la metà del minimo passo con cui si possono ripetere due MOSFET. Fino agli anni 2000 il nodo tecnologico coincideva all'estensione della minima lunghezza di canale, passando poi all'half pitch, ed infine ad un valore totalmente scollegato da una dimensione reale, stabilito dalla fonderia stessa, più legato alle performance di densità di integrazione dei dispositivi.



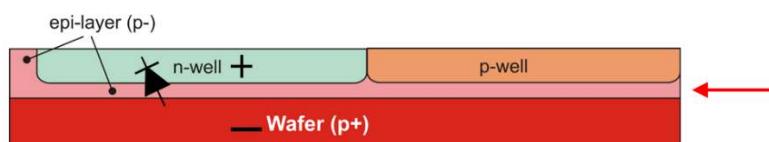
La tecnologia FD-SOI ha un minor costo, è più semplice e permette di ottenere buone performance. In questo caso si ha un sottile strato di silicio isolato dal bulk tramite uno strato ultra-thin di ossido sepolto. Il canale, in questo caso, è completamente svuotato all'origine e la conduzione avviene all'inversione. Il substrato sottostante può agire come un altro gate per modulare la tensione di soglia (ridurre i consumi, ridurre il leakage o aumentare la velocità dei circuiti digitali).

- Smaller W and L → Higher transistor density, higher f_T
- Lower Vdd and Vth → Lower dynamic power consumption, higher static power consumption
- Lower tox → Higher gate leakage (from SiO₂/PolySi to High-k materials/metal gate around 45 nm)

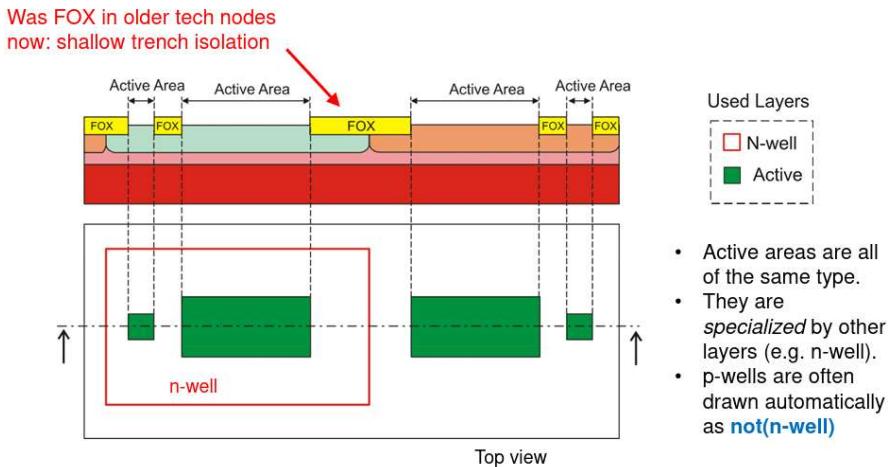


Lo scaling ha portato a dispositivi che sopportano tensioni sempre minori. La tensione di soglia, invece, ha subito una riduzione molto meno drastica. Anche se dal punto di vista della tendenza dell'alimentazione si vorrebbe ridurre il più possibile la tensione di soglia, il limite inferiore è dato dal margine di rumore: se basta una piccola fluttuazione per accendere il transistor, il livello logico risente fortemente del rumore. Con l'avanzare dello scaling si nota anche una diminuzione dell'EOT (Equivalent Oxide Thickness). Da un certo punto in poi per l'isolante di gate si è iniziato ad usare ossidi con elevata costante dielettrica anziché SiO₂. Questo perché per poter continuare a controllare il canale quando i transistori scalano è necessario ridurre lo spessore dell'ossido, ma oltre un certo spessore si determina corrente di perdita. Per ridurre questo effetto si sono introdotti materiali ad alta costante dielettrica, che permettono di ottenere la stessa configurazione di campo elettrico con uno spessore maggiore (equivalente a uno spessore minore di SiO₂).

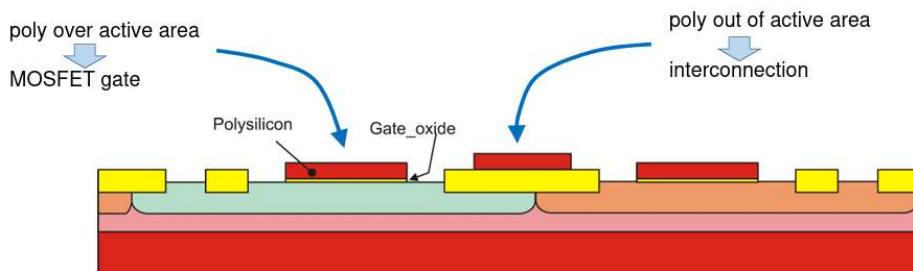
Trattiamo ora una semplificazione del processo CMOS, cercando di mantenere una visione ad alto livello in ottica di layout. Il processo ha inizio con un wafer di silicio, ad esempio drogato di tipo p+ (pseudo metallico), con uno strato p epitassiale poco drogato. Quest'ultimo ha una buona conducibilità (meno impurità → scattering minore) e permette di integrare bene le well di tipo p (per gli nMOSFET, sono necessarie per ottenere un droggaggio corretto che dia la giusta tensione di soglia) e le well di tipo n (per i pMOSFET). Con questo assetto stiamo parlando di un processo di tipo n (il substrato è di tipo p). Un processo del genere, in cui sia possibile controllare il droggaggio delle p well prende il nome di **twin tub**.



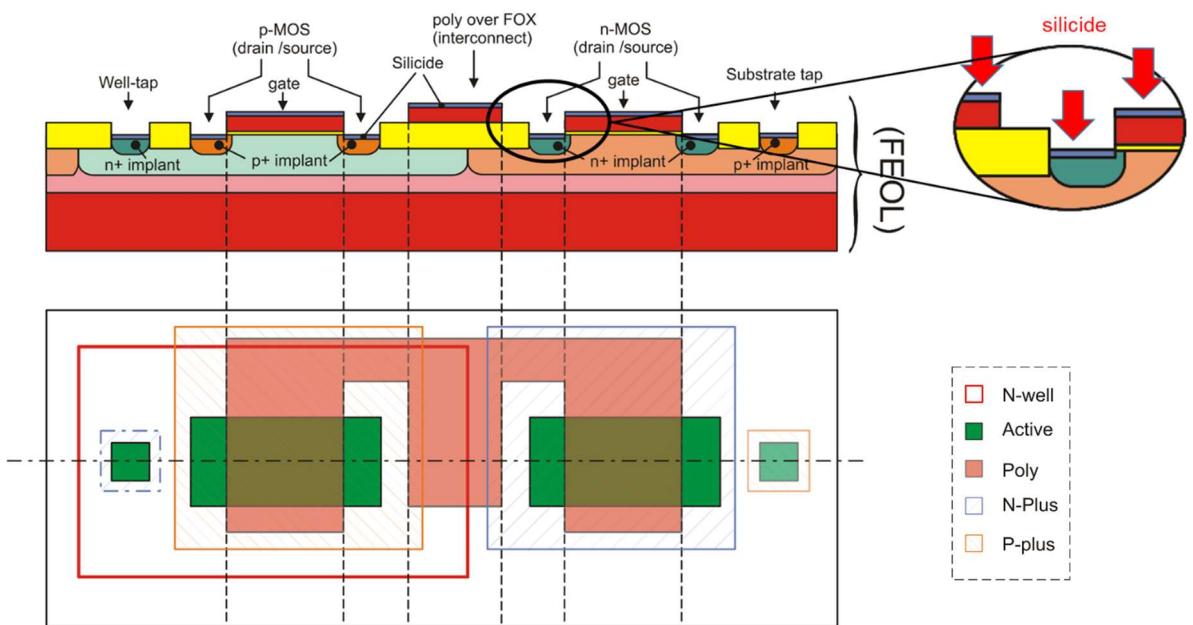
Le p well si trovano tutte in corto circuito con il substrato; questo fa sì che tutti i transistori di tipo n abbiano il body vincolato allo stesso potenziale elettrico. Le n well, invece, sono isolate rispetto al substrato con una giunzione che, se il substrato è posto al potenziale più basso, è polarizzata in inversa. Dunque, le n well (che coincidono al body dei pMOS) sono anche isolate tra loro, elettricamente indipendenti, mentre invece le p well (che corrispondono al body degli nMOS) sono tutte elettricamente in contatto con il substrato. Volendo si possono polarizzare i substrati dei transistori pMOS a potenziali tutti diversi tra loro, ma sempre e comunque la giunzione substrato-n well non deve andare in diretta. Quasi sempre le n well saranno poste al potenziale più alto disponibile e il substrato al potenziale più piccolo del circuito. Lo strato inferiore è comodo che sia di tipo p+ poiché offre un comportamento elettrico simile a quello di un metallo. Questo permette di ottenere un substrato il più possibile equipotenziale (essendo la resistenza molto minore, la corrente che scorre nel substrato causa minori differenze di potenziale), riduce il fenomeno del latch up e del rumore di substrato. Se avessimo considerato un processo CMOS complementare, con substrato di tipo n, i transistori di tipo p nelle n well avrebbero avuto il body cortocircuitato al substrato, mentre quelli di tipo n nelle p well avrebbero potuto avere il body a potenziale variabile, ma sempre e comunque tale da non mandare in diretta la giunzione di isolamento rispetto al substrato.



Le **aree attive**, che vengono progettate in fase di layout con l'apposito layer, corrispondono alle aperture in cui il substrato è accessibile per la creazione dei dispositivi e dei contatti metallici. Tutto ciò che è complementare alle aree attive corrisponde a zone di ossido (FOX: field oxide, ormai sostituito dal STI - Shallow Trench Isolation). L'ossido è sufficientemente spesso per reggere le tensioni operative che possono presentarsi nel circuito. Per il momento l'unica cosa che si può affermare dall'esempio sopra è che siano presenti quattro aperture nell'ossido di campo. Ciò che darà contenuto e funzionalità a un'area attiva è l'intersecarsi degli altri layer in corrispondenza della stessa. Precedentemente alle aperture vengono create le n well con l'apposito layer; il resto è automaticamente drogato di tipo p. I due drogaggi sono tali da determinare una certa tensione di soglia per i dispositivi. Ci sono dei casi in cui può essere comodo ricavare aree attive che rimangono con il drogaggio originale del substrato (un p solitamente molto leggero). Un eventuale MOSFET costruito in corrispondenza di queste aree attive sarebbe di tipo n con una tensione di soglia molto bassa, vicina a 0. La possibilità di realizzare questi transistori, detti **nativi**, è a carico della fonderia. I transistori nativi non sono inclusi nei CMOS; questi ultimi hanno soglie più robuste che permettono margini di rumore più ampi. Per non creare equivoci la convenzione è che l'assenza di layer corrisponda ad un drogaggio di tipo p per i CMOS, mentre invece è d'obbligo un layer apposito (p exclude) per lasciare il drogaggio di substrato. La fonderia ricaverà la maschera p well con la seguente operazione logica: (NOT n well) OR (NOT p exclude).

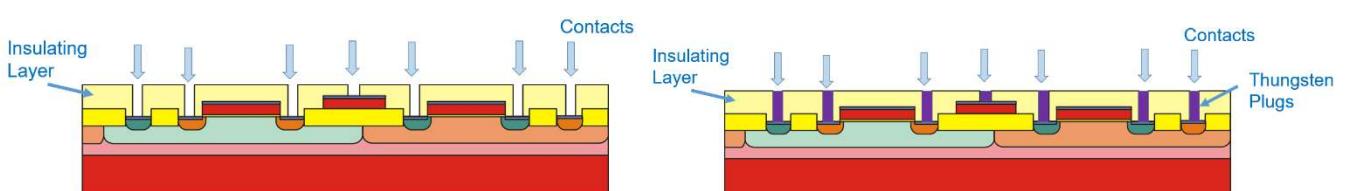


Si procede all'ossidazione di tutte le aree attive, attraverso cui si determina la formazione di un sottile film di ossido superficiale. A questo punto segue l'applicazione del polisilicio, che viene steso ovunque. Le geometrie in poly vengono ottenute dunque al negativo, rimuovendo il materiale dove non serve con un passo fotolitografico. È importante notare che la rimozione del poly si porta dietro la rimozione del film di ossido sottostante; similmente, laddove viene lasciato il poly è anche preservato il film di ossido. L'intersezione tra il poly e l'area attiva è il gate del MOSFET creato, in questo caso un MOSFET ad arricchimento. Il layer di polisilicio è utilizzato principalmente per realizzare il gate dei transistori, ma anche per interconnessioni locali (la resistività del polisilicio non è ottimale).

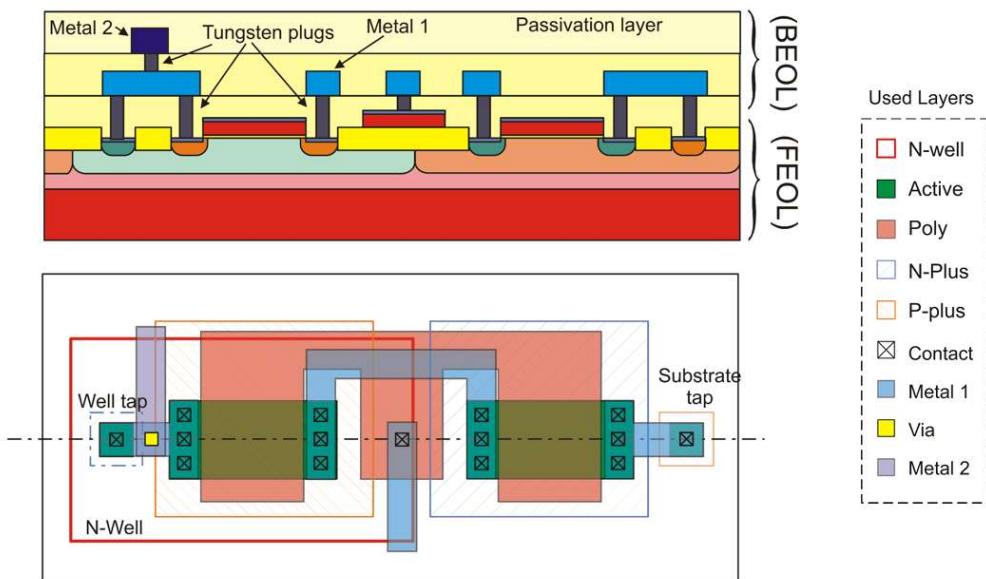


Le porzioni di area attiva rimaste libere dovranno ospitare i pozzetti e i contatti. All'inversione i portatori che supportano la conduzione nel dispositivo provengono in larga parte dai pozzetti, che devono avere drogaggi idonei a tal proposito. I layer di silicio n^+ e p^+ prendono il nome, rispettivamente, di N-Plus e P-Plus. Queste maschere devono obbligatoriamente includere tutta l'area attiva con un certo margine in eccesso. Bisogna tenere presente che fisicamente il droggaggio plus a questo punto del processo interesserà soltanto le zone di area attiva rimaste libere. Questa possibilità è dovuta alla presenza del polisilicio, che si mantiene intatto durante gli step tecnologici successivi e rende il processo di droggaggio plus autoallineato. Con gli stessi layer plus si realizzano le aree attive di contatto con il substrato (prese di substrato), necessarie per la polarizzazione della well. Fin da subito appare chiaro come i dispositivi con cui abbiamo a che fare sono rigorosamente a 4 terminali.

In un processo CMOS standard, a questo punto, tutte le aree attive e i livelli di poly vengono siliciurizzati, ricoperti da un silicio che ne abbassa la resistività superficiale, migliorando le prestazioni di velocità del dispositivo. Il polisilicio viene drogato nel momento in cui si applicano i drogaggi plus e questo permette di ottenere tensioni di soglia complementari. La siliciurizzazione evita che le eventuali giunzioni che si possono creare a seguito di questo step di droggaggio siano problematiche. Non esiste una maschera di siliciurizzazione; la fonderia applica tale processo indipendentemente a tutte le zone esposte. Con questo step si determina la chiusura della così detta fase FEOL (Front End Of the Line), termine con cui identifichiamo tutti e soli quei passaggi tecnologici che portano alla realizzazione dei dispositivi e non oltre. Segue la BEOL (Back End Of the Line), la fase di processo che raccoglie tutti i passaggi con cui si realizzano le interconnessioni tra i dispositivi. In realtà i layer di poly possono già costituire delle connessioni, come nell'esempio considerato, in cui una pista di poly contatta i due gate di due dispositivi complementari che realizzano un inverter CMOS. Quello in poly è un collegamento che si riserva esclusivamente per distanze limitate, interconnessioni locali. Per interconnessioni relativamente lunghe, sia locali che a grandi distanze, si usano i livelli di metal. FEOL e BEOL contengono entrambi l'espressione "of the line", la quale indica che stiamo sempre dentro il chip. Non si deve confondere il BEOL con la fase di "back end", la quale invece raccoglie tutti i passaggi successivi alla realizzazione del chip: taglio dei die, packaging, aggancio dei pin, saldature, ecc.



Il primo passo della BEOL è un’ossidazione completa di tutti i dispositivi. Vengono poi ricavati, tramite un layer contact apposito, dei buchi in corrispondenza dei contatti. Il poly di solito non si può contattare direttamente in corrispondenza del gate, area alquanto delicata. Si sceglie invece un altro punto dello stesso layer al di fuori delle aree attive. Le aperture vengono riempite con plug di tungsteno, i quali risultano molto uniformi, privi di disomogeneità come strozzature, le quali potrebbero comportare aumenti critici della resistenza elettrica. Segue una smerigliatura e un passo fotolitografico di definizione delle interconnessioni metalliche.

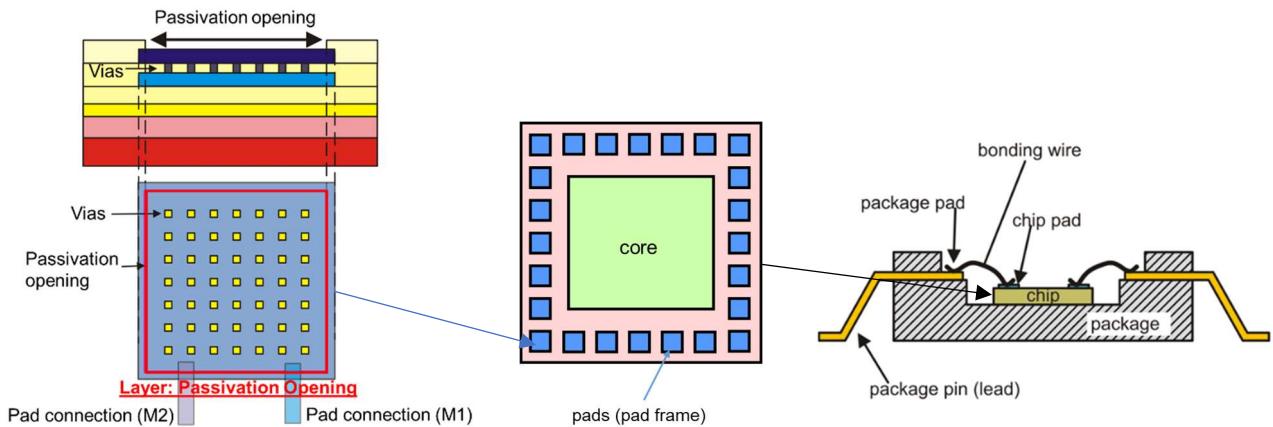


Dalla vista layout si osserva la presenza di contatti multipli sui source e drain dei dispositivi. Il contatto equivalente che ne risulta è un parallelo di tre contatti, per cui la resistenza equivalente è un terzo e ciò permette una portata maggiore di corrente elettrica. Spesso un unico contatto grande è proibito dalle regole di layout, poiché la fase di deposizione del plug di tungsteno è delicata e soggetta a limitazioni tecnologiche del processo (se è troppo grande la deposizione non è buona). Ecco perché il layer contact solitamente non ha una larghezza minima, ma esatta. In alcune tecnologie sofisticate sono possibili diverse dimensioni e forme per il contatto. Un’altra accortezza da prendere è quella di evitare di porre un plug senza che in corrispondenza vi sia una metal. Il contact mette in contatto la metal 1 o con l’area attiva o con il poly a seconda di quale coppia di layer forma l’intersezione. Il primo livello di metal realizza i collegamenti tra i due drain, tra il source dell’nMOSFET e ground e tra il source del pMOSFET e il potenziale più alto del circuito.

In un processo moderno vengono aggiunti livelli di metal isolati rispetto a quelli inferiori (inter-metal dielectrics) per mezzo dell’ossido, che si apre solo in corrispondenza dei siti in cui la metal superiore deve contattare la metal inferiore. Andare oltre due livelli di metal non è obbligatorio, ma certamente permette di risparmiare area e di sfruttare la dimensione verticale per creare contatti a minore resistenza. I contatti tra metal prendono il nome di **vie** e sono diversi dal primo livello contact. I processi vengono forniti con delle opzioni: più livelli di metal, l’ultimo livello in rame, alluminio, più o meno spesso. Due livelli di metal sono quasi indispensabili; più livelli di metal con le relative vie offrono uno sbroglimento più efficiente e contenuto in termini di area. Un livello di metal può essere posto in collegamento soltanto con il livello immediatamente sottostante; per ogni coppia di metal il layer via sarà diverso. Inoltre, gli spessori dei livelli di metal superiori crescono per offrire una maggiore portata di corrente, utile per portare segnali globali ad alto fan out come il clock (si risolve anche utilizzando più livelli di metal in parallelo).

Dopo aver definito l’ultimo livello di metal viene deposto uno strato di passivazione isolante che può avere varie composizioni. Spesso si tratta di una combinazione di ossido/nitruro al cui al di sopra si può depositare poliammide. Lo strato di passivazione protegge le strutture dall’umidità, graffi, sollecitazioni meccaniche.

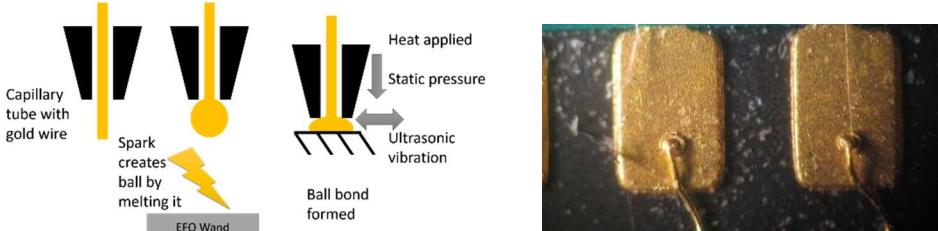
Infine, il circuito deve parlare con il mondo esterno tramite i pad:



Anche il package avrà dei pad corrispondenti, che andranno collegati a quelli del chip. Per realizzare un pad occorre una metal superficiale più spessa, resistente al processo di saldatura e alle vibrazioni meccaniche. Il bonding pad è descritto nelle librerie del PDK; è possibile realizzarlo su più di un livello di metal, ad esempio sugli ultimi tre (connessi da vie). Questo facilita le connessioni al pad e irrobustisce la struttura. Al di sotto dei pad si possono comunque integrare dispositivi che spesso, in quei pressi, sono di I/O (buffer tristate, inverter). Si tratta di dispositivi diversi, che reggono tensioni maggiori, la cui progettazione può indurre a riciclare qualche livello di metal dai modelli di libreria degli altri dispositivi attorno. La zona di passivazione non permetterebbe di accedere al pad, ragion per cui il contatto con il pad è possibile aprendo lo strato di passivazione per mezzo di un layer specifico (glass, passivate, pad, sono tutti nomi possibili). Solitamente la saldatura tra i pad del package e i pad del chip è di tipo wedge bonding, una tecnica che fa uso di ultrasuoni per rimuovere gli ossidi dalla superficie dei metalli e per saldare i due conduttori con l'addizione di una certa pressione.

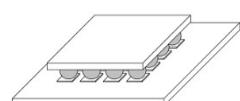


Quando i pad sono molto piccoli il fatto di dover stendere i fili per orizzontale e la deformazione dei pad dovuta agli ultrasuoni rende preferibile la tecnica del ball bonding. In quel caso il filo procede in verticale. Una scarica elettrica fonde localmente la punta del filo, che assume la forma di una sferetta. Quando la sfera è ancora semi-fusa viene applicata e, sempre per mezzo di una combinazione tra calore, pressione e ultrasuoni, si determina la saldatura.



Another bonding technique is flip-chip bonding. In order to mount the chip to external circuitry (a circuit board or another chip or wafer), it is flipped over so that its top side faces down, and aligned so that its pads align with matching pads on the external circuit, and then the solder is reflowed to complete the interconnect.

1. Integrated circuits are created on the wafer
2. Pads are metallized on the surface of the chips
3. A solder ball is deposited on each of the pads, in a process called wafer bumping
4. Chips are cut
5. Chips are flipped and positioned so that the solder balls are facing the connectors on the external circuitry
6. Solder balls are then remelted (typically using hot air reflow)
7. Mounted chip is “underfilled” using a (capillary, shown here) electrically-insulating adhesive

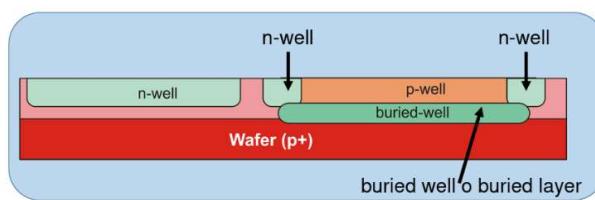


Il vantaggio del flip chip è che i pad possono essere posti da per tutto, non solo alla periferia. Questo è particolarmente comodo quando il chip ha un numero di pad molto elevato, come nel caso di processori. Non è detto che la fonderia supporti il bonding flip chip. Si tratta di un processo con passi particolari in più e non è detto che le bonding house abbiano l'attrezzatura per effettuarlo. Al progettista interessa la dimensione del pad (possono esserci pad diversi disponibili nel design kit), la distanza tra i pad (dipende dalla bonding house).

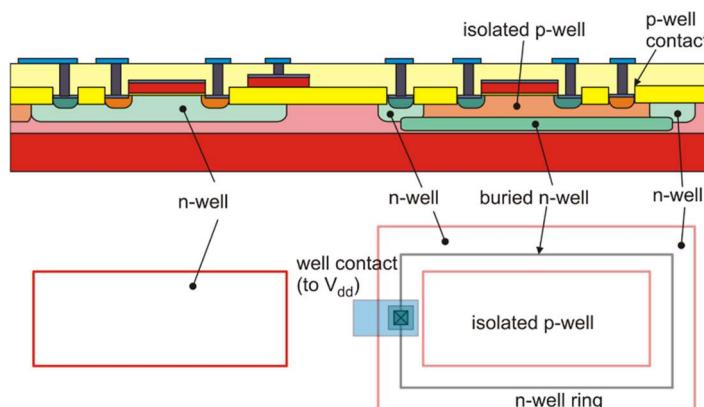
Varianti sul tema

Triple well

Il processo CMOS può presentare alcune varianti. Possono, ad esempio, essere previste più famiglie di dispositivi, come i nativi già discussi, oppure transistori a soglie diverse con well drogate diversamente (HVT, LVT). Per dispositivi di I/O possono essere disponibili ossidi di spessore diverso. Un ossido più spesso determina una tensione di soglia maggiore. La variante che prendiamo in considerazione è la **triple well**.



Un processo triple well offre la possibilità di racchiudere la p well all'interno di una sacca di tipo n. In questo modo, con un substrato di tipo p, entrambe le well sono isolate. Conseguentemente, entro i limiti, i body di entrambi i dispositivi possono essere posti a un potenziale a piacere. Ciò permette di poter controllare l'effetto body, dunque la tensione di soglia. Questo risolve il problema dell'equipotenzialità delle p well rispetto al substrato nel processo twin tub. Il passo di processo in più non è tanto la parte laterale della sacca, che può essere realizzata con il normale layer n-well, ma la buried well, cioè la parte profonda sul fondo della sacca. Tipicamente si ottiene con un'impiantazione ionica ad alta energia. L'utilizzo delle triple rende il processo robusto contro il fenomeno del latch up evitando l'insorgenza di transistori parassiti e contro le radiazioni ionizzanti presenti nello spazio. La sequenza p-well – buried-well – substrato, di fatto, forma due giunzioni back-to-back, una delle quali sarà necessariamente in inversa (solitamente la sacca è polarizzata al potenziale più alto).



Bipolar processes

Technology	Available Devices	Notes
Bipolar	Vertical NPN, Lateral PNP	Used for precision and/or fast amplifier. <u>Si-Ge</u> versions for RF applications
Complementary Bipolar	Vertical NPN, Vertical PNP	
BiFet	BJTs and JFETs	Used for precision / low bias current amplifiers

BiCMOS, BCD, SOI

Technology	Available Devices	Notes
BiCMOS	CMOS + BJTs	Mixed Signal ICs High speed digital line drivers
BCD	Bipolar, CMOS, DMOS	Smart Power
SOI Silicon on Insulator.	As CMOS, BiCMOS or BCD	High Voltage and Rad Hard (e.g. space applications)

I bipolari puri si usano sempre meno per integrare logica. Il problema principale è che nel caso della tecnologia bipolare il passo di ripetizione tra i dispositivi è molto maggiore rispetto alla tecnologia MOS, per cui la densità di integrazione è peggiore. Si utilizzano ancora per la maggior parte degli amplificatori operazionali e, nella versione Si-Ge, per circuiti ad alta frequenza (fino a 100 GHz). Trovano anche ampio spazio in circuiti di precisione o, nella versione discreta, per realizzare dispositivi di potenza.

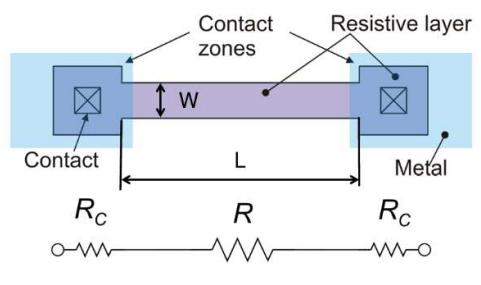
In generale il BJT è comodo quando in combinazione con il CMOS. Uno dei vantaggi del BJT per applicazioni analogiche è la capacità di portare più corrente a parità di area rispetto al MOSFET (per stadi di uscita anche digitale per il pilotaggio di linee a bassa impedenza/capacità molto grandi si risparmia area) e un g_m più grande a parità di corrente consumata. Altro vantaggio per l'analogico è un minore rumore flicker. Il processo DMOS permette di sopportare tensioni elevate, per cui è un'ottima scelta per alimentatori e dispositivi di potenza.

Il processo SOI si divide nell'SOI tradizionale, che consiste nel fatto che i dispositivi sono realizzati su uno strato di silicio relativamente spesso isolato rispetto al substrato con uno spesso strato di ossido. Nel caso dell'FD-SOI (Fully Depleted) l'ossido di isolamento è molto sottile, per cui l'effetto elettrostatico tra il substrato e il silicio soprastante è notevole. Si può modulare il droggaggio dello strato di silicio, anch'esso molto sottile, in modo elettrostatico svuotando completamente la zona in cui è realizzato il dispositivo. La tecnologia FD-SOI è utilizzata maggiormente per dispositivi avanzati ultrascalati come alternativa al finFET.

Elementi passivi IC

Nei circuiti integrati analogici è fondamentale realizzare componenti passivi: resistori, condensatori e induttori. I resistori possono servire per realizzare DAC di precisione (R-2R, resistor string), amplificatori tempo continui con guadagni accurati (resistori per il feedback a rapporto resistivo), interfacce per sensori, riferimenti di tensione. Anche se i resistori sono importanti si cerca di limitarne l'uso nell'elettronica integrata per ridurre il consumo di potenza il più possibile. Anche i condensatori hanno molteplici applicazioni in campo analogico: filtri analogici, circuiti a condensatori commutati (la maggioranza degli ADC sono switched cap), lettura di carica trasferita dai sensori di immagine, ecc. Gli induttori sono invece problematici nei circuiti integrati poiché occupano molto spazio, contro la tendenza della microelettronica di rendere tutto compatto e miniaturizzato. La conseguenza di ciò è che si riescono ad integrare al massimo induttanze nell'ordine del nH . Inoltre, il piano di massa influisce negativamente sul fattore di qualità Q (legato al rapporto L/R), che determina la qualità del filtraggio e il rumore di fase negli oscillatori. Di fatto gli induttori integrati vengono utilizzati solo per applicazioni ad alte frequenze, anche se si trovano alternative all'induttore passivo.

Resistori integrati



$$R = R_s \frac{L}{W} \quad \text{with} \quad R_s = \frac{\rho}{t} \quad R_{tot} = R + 2R_C$$

In fase di layout l'istanza di un resistore viene creata con una cella parametrica fornendo il valore di resistenza. Il resistore sarà contattato dalle metal con appositi contatti, i quali sono soggetti a regole di enclosure: devono essere circondati per una certa estensione sia dal materiale resistivo che dalla metal. Inevitabilmente il contatto determinerà l'insorgenza di una resistenza serie non voluta; la resistenza totale della geometria così costituita è data da $R + 2R_C$. La resistenza che si inserisce sul layout è la sola R , non tiene conto dei contatti. L'errore dovuto alle resistenze di contatto è tanto più impattante quanto la R è piccola.

Questo può essere un problema per la realizzazione dei rapporti di amplificazione. Sono rari i design kit che tengono conto di questo aspetto, ma ciò è ragionevole in quanto è sempre possibile aggiungere contatti per abbassare la resistenza complessiva.

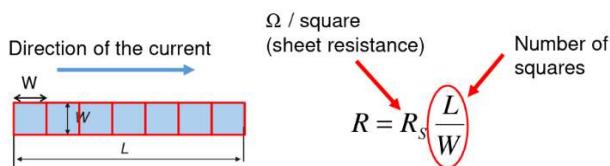


A parità di L , aumentando W si abbassa la resistenza, aumenta la portata di corrente per μm e aumenta l'area complessiva. Nella zona dei contatti potremmo anche lasciare un solo contatto, ma in tal caso le linee di flusso della corrente non sarebbero più parallele alla direzione della lunghezza, per cui si avrebbe una resistenza non più accurata.

Inoltre, inserire più contatti abbassa considerevolmente la resistenza di contatto complessiva. Il numero di contatti che si possono inserire dipende dai margini, dagli spacing. Spesso, l'inserzione dei contatti è automatica.

Parametri di merito

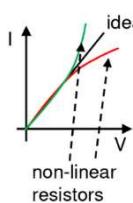
Resistenza di strato



Gli oggetti con cui abbiamo a che fare si sviluppano in pianta, superficialmente. Consideriamo una striscia di un materiale conduttivo, ad esempio un resistore o un'interconnessione. Una volta individuata la direzione della corrente, chiamiamo la dimensione lungo tale direzione L , la dimensione perpendicolare W . Il parametro R_s identifica la resistenza di strato, in Ω . La quantità L/W è il numero di quadrati di lato W che si possono costruire verso la lunghezza, lungo la direzione della corrente. Se nell'esempio ruotassimo di 90° la direzione della corrente, non avremmo nemmeno un quadro. Ciascun layer conduttivo sarà caratterizzato da una certa resistenza di strato. Si possono realizzare resistori in poly, poly ad alta resistività, oppure con strati diffusi. Il poly stesso può essere drogato di tipo n o tipo p. Per una resistenza grande è ragionevole scegliere un layer con R_s grande, così da minimizzare l'ingombro del resistore.

Dipendenza dalla tensione

Per alcuni resistori la resistenza dipende considerevolmente dalla tensione. Ad esempio, il filamento delle lampade a incandescenza all'aumentare della tensione applicata si scalda e la sua resistenza aumenta indirettamente tramite effetto Joule.



Se la resistenza dipende dalla tensione si osserva una legge IV non lineare. Per un amplificatore, se al variare della tensione di ingresso cambiano le resistenze impiegate nel feedback il guadagno è anch'esso dipendente dalla tensione, il che comporta distorsioni del segnale amplificato.

$$R(V) = R(0)[1 + \alpha_{V1}V + \alpha_{V2}V^2]$$

Dipendenza dalla temperatura

I circuiti con cui abbiamo a che fare non sono termostatati e si troveranno a una temperatura nominale di $T = 300\text{ K}$. In alcuni casi, ad esempio per resistenze il cui rapporto deve mantenersi costante, si può ignorare l'effetto della temperatura. In altri casi invece, come quello di una resistenza utilizzata in un convertitore tensione-corrente ($I = V_u/R$), la dipendenza della resistenza degrada l'accuratezza e corrompe la legge di conversione. Anche in questo caso possiamo approssimare l'andamento con un'espansione al secondo ordine:

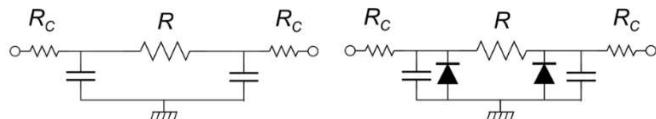
$$R(T) = R(T_0)[1 + \alpha_1(T - T_0) + \alpha_2(T - T_0)^2]$$

Il coefficiente α_1 è denominato TCR (Temperature Coefficient of Resistance) ed ha un valore tipico di 10^{-3} K^{-1} : se la temperatura varia di un grado, la resistenza varia dell'1 per mille.

$$\alpha_1 = TCR = \frac{1}{R} \left. \frac{dR}{dT} \right|_{T=T_0}$$

Componenti parassiti

Quando si introduce un resistore in un circuito compaiono sempre e comunque degli elementi parassiti.



Oltre alle resistenze dei contatti, non molto predicibili, si vengono a determinare capacità distribuite tra tutto il corpo del resistore e il substrato del chip. Solitamente si modella il fenomeno ponendo metà della capacità distribuita complessiva da una parte, metà dall'altra. Concentrare tutta la capacità su un nodo sarebbe scorretto in quanto quello stesso nodo potrebbe essere collegato a massa e dunque essere più immune al crosstalk capacitivo. Anche a livello di schematico i resistori sono caratterizzati da W ed L ; se il design kit è ben fatto i parassiti capacitivi, che dipendono dalle dimensioni geometriche del resistore, sono considerati in fase di simulazione.

Inoltre, per i resistori diffusi compaiono delle giunzioni parassite. Consideriamo ad esempio una diffusione di tipo n all'interno di uno strato p. Se il resistore è a potenziale più alto è isolato tramite un isolamento a giunzione (polarizzazione inversa), ben diverso da un isolamento tramite dielettrico. Le correnti di saturazione inversa che scorrono possono iniziare ad essere significative se il resistore è abbastanza ingombrante. Oltre ciò, anche se staticamente la giunzione si mantiene polarizzata in inversa, durante i transitori può accadere che una zona del resistore raggiunga un potenziale tale da polarizzare direttamente la giunzione.

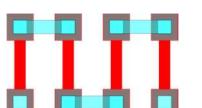
Tecniche di realizzazione dei resistori integrati

Serpentina

Per realizzare resistori di grande valore senza ingombrare troppo spazio si utilizza un corpo stretto (W piccola per minimizzare l'area) ripiegato a serpentina.



Questa soluzione è in realtà poco utilizzata. In corrispondenza delle curve la direzione della corrente non è rettilinea, per cui la resistenza per quadro non è facilmente calcolabile. Ciò rende il valore effettivo della resistenza poco accurato.

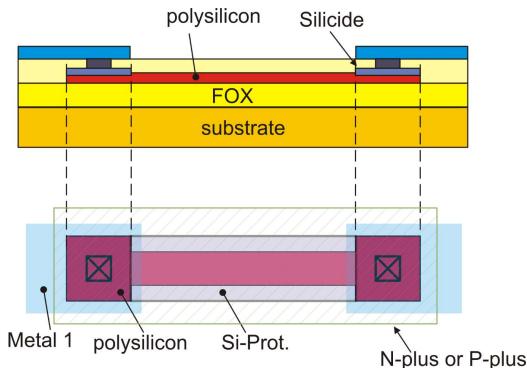


Si può fare affidamento alla funzione di placement array, che permette di suddividere il resistore in tante unità collegate in serie attraverso contatti, disposte sempre a serpentina. Generalmente è questa la soluzione preferita. Con questi metodi si riescono ad integrare resistenze fino al $M\Omega$.

In molti circuiti ciò che è importante sono i rapporti di resistenze. Se entrambe le resistenze sono realizzate nel solito modo, il rapporto si mantiene preciso. In generale, laddove le quantità importanti sono rapporti tra grandezze omogenee, effetti che producono incertezze sul singolo componente non contano più.

Resistori in polisilicio

La prima tipologia di resistori integrati è quella dei resistori in polisilicio:



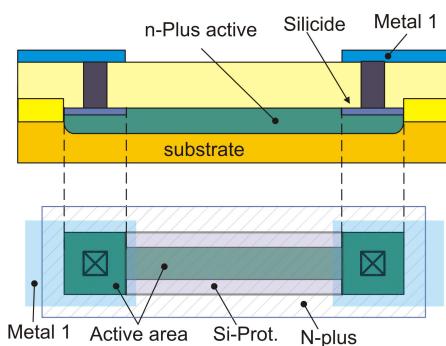
Il corpo del resistore non deve subire silicurizzazione, processo che ne abbasserebbe la resistenza di strato. A tal proposito, se il processo lo supporta, si utilizza il layer di protezione Si-Prot; la maschera corrispondente farà sì che in corrispondenza delle geometrie disegnate non venga applicato il film di siliciuro. Per il resto la silicurizzazione è voluta, in quanto migliora i contatti. Il poly viene drogato per mezzo degli stessi layer con cui si realizzano i pozetti, p-plus o n-plus.

In alcuni processi come quello digitale, pur essendo la stessa la tecnologia, non è possibile realizzare resistori in polisilicio. Questo perché lo step della protezione dal siliciuro può essere assente per questioni di economia di processo. Altre volte, anche se è possibile realizzare un determinato dispositivo nel layout, questo può non essere inserito in libreria, per cui può essere assente nello schematico e non caratterizzato in termini di comportamento in temperatura e componenti parassiti. Un esempio è il transistore nativo. In un processo analogico standard c'è sempre la possibilità di realizzare resistori in poly di tipo p o tipo n e poly silicurizzato per resistenze piccole. Per resistenze grandi alcuni processi permettono anche la realizzazione di resistori in poly intrinseco o compensato. Tecnologie che consentono maschere in più avranno un costo maggiore.

I resistori in polisilicio sono i meno dipendenti dalla temperatura, i più lineari con le tensioni applicate. Inoltre, c'è la possibilità di renderli molto resistivi su aree compatte.

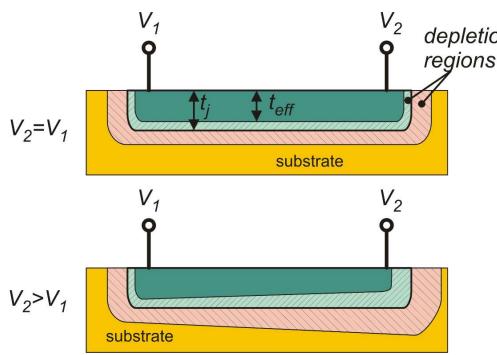
Resistor Type	Sheet resistance	TCR	Non linearity (a_{V1})
n-plus polysilicon	30-150 Ω	100-500 ppm/ $^{\circ}\text{C}$	50 ppm / V
p-plus polysilicon	50-400 Ω	250-1000 ppm/ $^{\circ}\text{C}$	-50 ppm / V
high-res polysilicon	400-4000 Ω	-1000 ppm/ $^{\circ}\text{C}$, -3000 ppm/ $^{\circ}\text{C}$	100 ppm / V
Salicidized polysilicon	5-10 Ω	2500-3500 ppm/C	-

Resistori diffusi



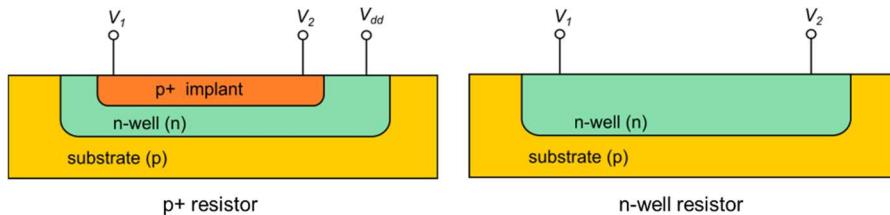
In questo caso, dovendo accedere direttamente al substrato per realizzare il resistore, occorre disporre di un'area attiva. Il droggaggio che interessa la zona diffusa è lo stesso utilizzato per i pozetti (plus). Come per il resistore in poly, si protegge il corpo del resistore dalla silicurizzazione, la quale è invece presente ai contatti per migliorarne la qualità in termini di comportamento ohmico. Per i resistori diffusi l'effetto della tensione sulla resistenza è significativo

Il resistore costituisce di per sé una giunzione rispetto al substrato. Impedendone l'accensione in diretta, la sua presenza è tollerabile.



Tuttavia, la regione di svuotamento che si crea alla giunzione comporta alcune problematiche. Anzitutto, se i potenziali V_1 e V_2 variano a modo comune, lo spessore efficace del resistore tende a diminuire. Applicando invece V_1 e V_2 diversi, laddove il potenziale è maggiore la tensione in inversa è più grande, la zona di svuotamento più estesa, lo spessore efficace compresso. Dunque, il modo differenziale fa sì che la geometria del resistore cambi in modo più complicato. Entrambi gli effetti fanno sì che la resistenza del resistore diffuso sia considerevolmente dipendente dalla tensione applicata ai capi.

A volte lo strozzamento comandato in tensione è voluto per ottenere grandi resistenze. Nel μ A741 i resistori sono tutti diffusi. Ad oggi è possibile che tali resistori non siano nemmeno disponibili in libreria. Altri resistori diffusi:



Alcune caratteristiche dei resistori diffusi:

Resistor Type	Sheet resistance	TCR	Non linearity (a_{V1})
n-plus on substrate	30-80 Ω	1000-1500 ppm/ $^{\circ}$ C	400 ppm / V
p-plus on n-well	50-150 Ω	1000-1500 ppm/ $^{\circ}$ C	400 ppm / V
n-Well on substrate	400-4000 Ω	-2000, -3000 ppm/ $^{\circ}$ C	3000 ppm / V

Condensatori integrati

Il valore di capacità di un condensatore integrato dipende molto meno dalla tensione applicata, dalla temperatura e dai componenti parassiti (capacità verso il substrato, giunzioni in inversa se utilizzate come isolamento), il che rende tale componente molto importante in elettronica integrata.

I parametri con cui si caratterizzano i condensatori integrati sono i seguenti:

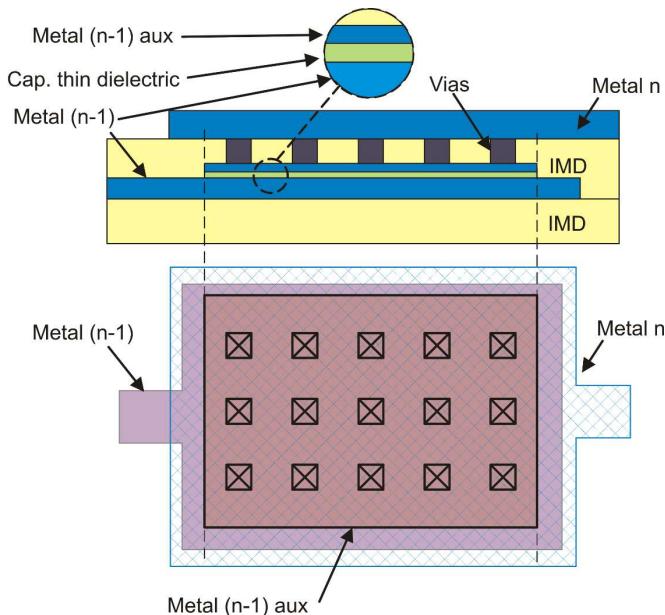
- Capacità per unità di area: le capacità per unità di area integrabili spaziano tra $1 \text{ fF}/\mu\text{m}^2$ e $8 \text{ fF}/\mu\text{m}^2$. Con tali valori per integrare un condensatore da 1 pF occorre occupare un'area di $1000 \mu\text{m}^2$. Un condensatore grande tanto quanto un chip di 1 mm^2 raggiungerebbe il valore di 1 nF .
- Linearità: la capacità non dovrebbe dipendere dalla tensione. Se la capacità cambia con la tensione, i circuiti switched cap non funzionano più come dovrebbero. In altri casi si sfrutta la non linearità di $C(V)$ per fare accordare gli oscillatori.

La dipendenza della temperatura è trascurabile per la maggior parte delle applicazioni.

Tecniche di realizzazione dei condensatori integrati

Condensatore MIM

Una prima tipologia di condensatori integrati è quella dei condensatori MIM (Metal-Insulator-Metal). Si tratta di un condensatore le cui armature sono realizzate con due livelli di metallizzazione adiacenti. Se come dielettrico si utilizzasse l'intermetallico, dato l'elevato spessore, la capacità per unità di area sarebbe bassa. Per la realizzazione dei MIM le tecnologie mettono a disposizione un ulteriore livello di metallizzazione (layer "cap" o "mim cap") in una zona intermedia, separato da quello inferiore da un ossido sottile apposito (high k). La fonderia, dopo aver deposto la metal che corrisponde all'armatura inferiore, depone al di sopra un ossido sottile, a volte plastico organico, al di sopra del quale viene poi deposta la metal ausiliaria.



L'accesso all'armatura superiore viene ottenuto aprendo tanti buchi nell'ossido intermetallico superiore, i quali vengono poi contattati con la metal superiore. La metal ausiliaria ha proprio la funzione di minimizzare la distanza tra le armature, con conseguente aumento della capacità per unità di area. Un'idea alternativa potrebbe essere quella di scavare una cavità nell'ossido che separa le due metal in corrispondenza del condensatore.

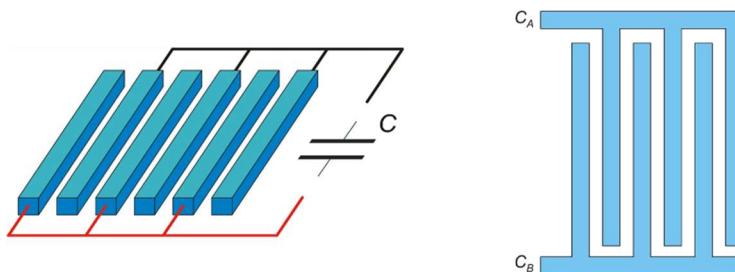


I condensatori MIM sono molto lineari, affidabili, precisi. Generalmente la loro capacità per unità di area è bassa ($1 \text{ fF}/\mu\text{m}^2$); per appena 1 pF è necessaria un'area di $33 \mu\text{m} \times 33 \mu\text{m}$. Altri svantaggi sono la necessità di processi tecnologici in più, possibile corrente di perdita, basse tensioni di rottura.

$$C = \epsilon \cdot \frac{A}{d}$$

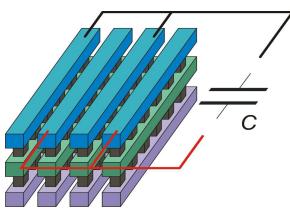
Condensatore MOM

Altra tipologia di condensatori integrabili è costituita dai condensatori MOM (Metal-Oxide-Metal)



La capacità d'interesse nei MOM è quella laterale tra le strisce di metallo standard (struttura interdigitata); un'armatura del condensatore è costituita da un pettine, l'altra dal pettine complementare. Anche le linee di campo ai bordi partecipano a determinare la capacità. I MOM sono più recenti dei MIM; l'efficacia di questa struttura in termini di capacità, infatti, dipende dal minimo spacing tra le geometrie permesso dal nodo tecnologico. Per nodi più spinti anche i MOM diventano efficienti in termini di capacità per unità di area e competitivi per il costo della tecnologia. Uno dei vantaggi più lampanti dei MOM è l'assenza di passi di processo aggiuntivi per la sua fabbricazione.

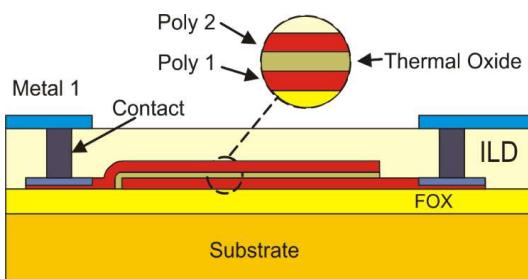
A parità di capacità l'ingombro è maggiore rispetto ai MIM, ma rimane la possibilità di stratificare i condensatori MOM (multi-layer MOM capacitors).



La tensione di rottura è superiore rispetto ai MIM. Uno svantaggio di una struttura MOM multistrato dal punto di vista del layout è che utilizzando tutti i livelli di metallizzazione impedisce il passaggio di interconnessioni metalliche, per cui può complicare il routing del circuito.

Il MOM multistrato è competitivo con il MIM in termini di ingombro per unità di capacità. Una buona tecnologia ha sia MIM che MOM disponibili e caratterizzati. Se il MOM non è caratterizzato, non necessitando di layer particolari è comunque possibile realizzarlo e se ne può stimare la capacità con le informazioni di contorno riguardanti la capacità interlinea.

Condensatore poly – poly

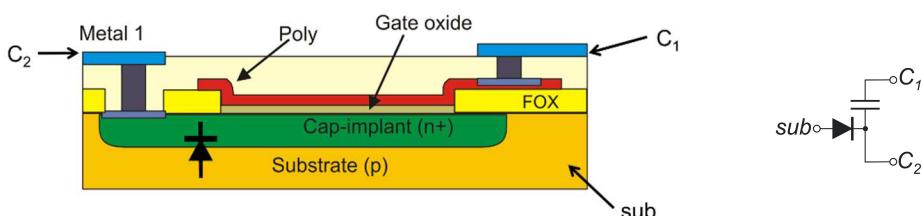


In alcune tecnologie (EPROM) si possono avere a disposizione due livelli di polisilicio con cui realizzare le armature di un condensatore a facce piane e parallele. Il dielettrico in questo caso è l'ossido termico di gate, che permette di poter integrare grandi capacità per unità di area ($8 \text{ fF}/\mu\text{m}^2$).

Il condensatore poly-poly soffre di una discreta non linearità: all'aumentare della tensione il campo elettrico penetra maggiormente nel polisilicio rispetto al metallo e determina l'insorgenza di una capacità in serie (zona di svuotamento da effetto di campo). Anche il condensatore poly-poly sopporta tensioni relativamente basse.

Condensatore poly-implant

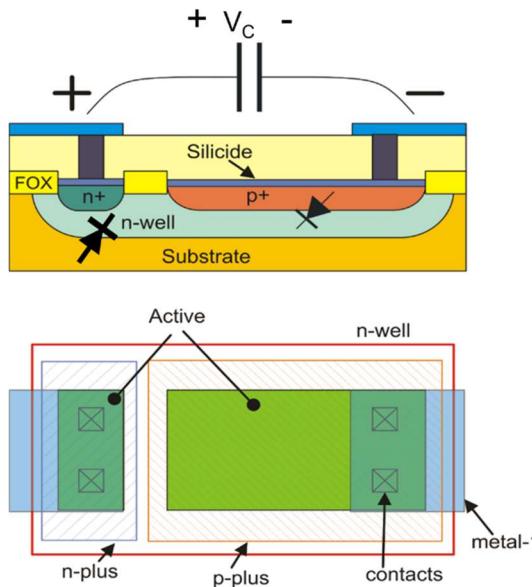
In assenza del doppio livello di poly è comunque possibile realizzare condensatori poly-implant:



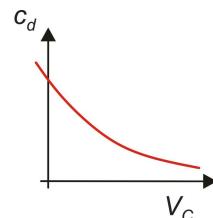
L'armatura inferiore può essere realizzata con un'impiantazione molto drogata. Il circuito equivalente del condensatore, dunque, comprende anche un diodo parassita rispetto al substrato a cui è associata una capacità a sua volta oltre che una corrente di saturazione inversa. Il condensatore poly-implant non è particolarmente ideale; può essere utilizzato quando non si necessita di un valore di capacità particolarmente preciso.

Condensatore a giunzione

Un’ultima tipologia di condensatore integrabile è costituita dai junction capacitor.



Soprattutto per applicazioni RF è possibile utilizzare come armature due zone di droggaggio opposto. Questo comporta inevitabilmente una forte dipendenza della capacità dalla tensione applicata.



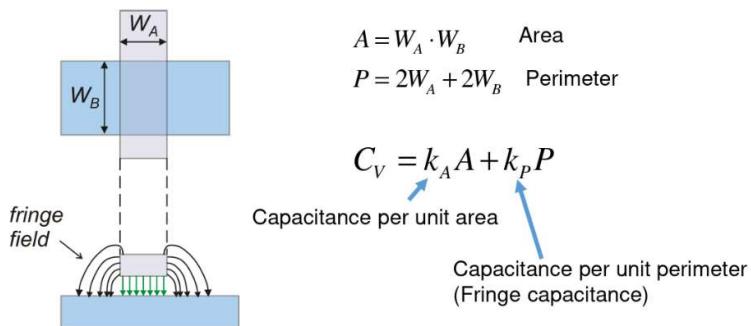
L’utilizzo per le RF si deve proprio a questa dipendenza, con cui è possibile realizzare varicap per accordare filtri LC. Oppure, si possono utilizzare come compensatori di frequenza o come condensatori di bypass.

Confronto

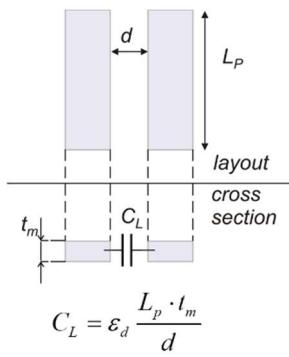
I condensatori integrati a confronto:

Capacitor Type	Capacitance per unit area	Linearity dependence (voltage)	Parasitic components
MIM	1 fF / μm^2	< 100 ppm/ V	Capacitance to substrate (bottom plate only)
MOM (flux) 1 metal layer	0.1 fF / μm^2	< 100 ppm/ V	Capacitance to substrate (both terminals only)
MOM (flux) 6 metal layer	1 fF / μm^2	< 100 ppm/ V	Capacitance to substrate (both terminals only)
poly-poly	6 fF / μm^2	100 - 1000 ppm/V	Capacitance to substrate (bottom plate only)
poly-diffusion	6 fF / μm^2	100 - 1000 ppm/V	Diode to Substrate (bottom plate only)
junction	1 fF / μm^2	very -high (up to 30 % / V)	Diode to Substrate (bottom plate). Diode between terminals.

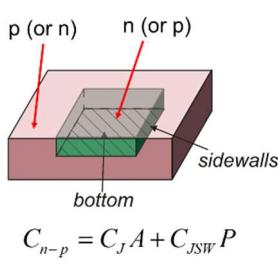
Capacità parassite



Quando si intrecciano strati conduttori su livelli diversi si formano condensatori parassiti. La capacità equivalente è determinata da un contributo di campo elettrico perpendicolare alle facce e da un contributo di campo elettrico di bordo, laterale. Rispetto alla zona di intersezione, il primo è proporzionale all'area, il secondo all'estensione dei bordi. Se si considerasse soltanto la capacità dovuta alle facce piane e parallele commetteremmo un'approssimazione per difetto. I parametri di processo con cui calcoliamo la capacità, “area capacitance” e “fringe capacitance”, saranno diversi dipendentemente dalla coppia di layer considerata. Cosa potremmo fare se nell'esempio la pista blu portasse un segnale analogico e quella grigia il segnale di clock? Se fossimo vincolati a due livelli di metal potremmo soltanto risolvere la questione a livello geometrico. Con tre livelli di metal, invece, potremmo interporre tra i due livelli uno schermo metallico posto a ground, il quale introduce un effetto gabbia di faraday. Il problema è sentito in quanto i segnali analogici risultano delicati nei confronti dei disturbi... uno sbilanciamento di appena 100 mV può essere deleterio. Nonostante il confronto diretto tra area e perimetro non abbia senso, si può comunque considerare il rapporto area/perimetro per avere un'idea di quale contributo alla capacità sia più pesante. Per il quadrato il rapporto $A/P = l/4$; se quindi il quadrato è grosso, ma più in generale se la geometria considerata è relativamente grande, il perimetro comporta una capacità più trascurabile. Le operazioni di estrazione PEX giocano proprio su questi aspetti.



È bene stimare sin dal principio quelle che possono essere le resistenze e le capacità parassite, a livello di layout, onde evitare di dover modificare un progetto ormai complesso. Le capacità parassite non si instaurano soltanto tra oggetti l'uno sopra l'altro, ma anche lateralmente tra due oggetti. Dovremmo anche in questo caso considerare gli effetti di bordo, ma per gli accostamenti laterali conta più la dipendenza con la lunghezza e la distanza tra gli oggetti. Ci sarebbe anche una forte dipendenza dallo spessore t_m dei layer, ma si tratta di un parametro che non viene sempre comunicato, un parametro dunque non sempre controllabile. Ecco perché talvolta C_L è fornita per unità di $t_m \cdot \epsilon_d$. La capacità laterale è importante per piste lunghe e vicine. Gli ordini di grandezza di queste capacità sono nel centinaio dei fF.

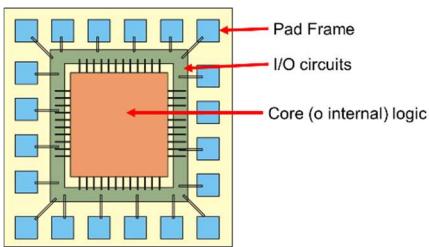


In casi particolari si ha un substrato e un'area con droggaggio opposto. La giunzione che si viene a formare la immaginiamo polarizzata in inversa (es. n well dentro il sub p o pozzetti di drain/source in un processo bulk). Oltre alle correnti di saturazione inversa ($fA/\mu m^2$), trascurabili a meno che non alterino una carica rappresentante di informazione, si osservano le capacità di giunzione inversa. Si possono stimare queste capacità tramite parametri dipendenti dall'area e al perimetro della sezione di giunzione. In realtà anche lateralmente avremmo un'area da considerare, ma lo spessore viene inglobato nel parametro C_{JSW} (“junction sidewalls”).

Con una geometria molto grande prevale il primo termine. La particolarità degna di nota è che la capacità di giunzione dipende dalla tensione, in particolare diminuisce all'aumentare della tensione in inversa. I parametri, specifici per ogni layer, sono riferiti a una tensione di polarizzazione nulla. Il manuale, dunque, dato che le giunzioni in questione non verranno polarizzate in diretta, fornisce il caso peggiore.

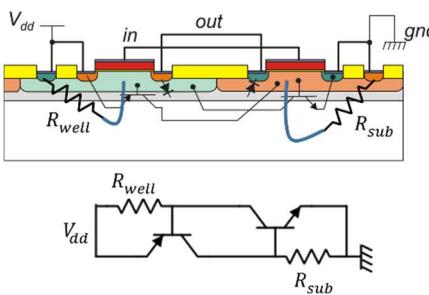
Latch up

Per contrastare il fenomeno del latch up si deve intervenire sulle regole di layout. Consideriamo un chip digitale, che ha una struttura meno anarchica rispetto a quella dei chip analogici.



Si ha una parte centrale detta core logic, che implementa le funzioni digitali (elaborazione, memorizzazione, esecuzione istruzioni, fetch, SRAM, periferiche I/O) e contiene i circuiti più veloci, più miniaturizzati e che reggono le tensioni più piccole (es. $\max(V_{dd}) = 1.8 \text{ V}$). La peculiarità dei dispositivi nel core è che non vedono il mondo esterno. Nessuna delle interconnessioni nel core è collegata direttamente ai pad, le piazzole sulle quali si saldano i fili da collegare ai pin esterni del package.

Il collegamento tra le due parti è regolato attraverso un'interfaccia di altri dispositivi che, simbolicamente, abbiamo raccolto nell'anello dei circuiti I/O. Si tratta di circuiti idonei ad essere collegati ad elementi esterni (carichi, piste, linee di comunicazione, altri moduli), realizzati con dispositivi appartenenti ad una famiglia particolare atta alla gestione di tensioni più alte (es. $V_{dd} = 3.3 \text{ V}$ nominalmente, anche un po' oltre) e al pilotaggio di carichi esterni. Questa differenza ha una ricaduta sul **latch up**.

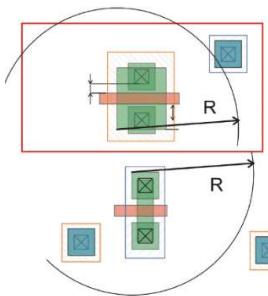


In un processo CMOS, i collegamenti riportati realizzano l'inverter CMOS. Anche per circuiti destinati all'analogico spesso compaiono drain/source in condivisione, attaccati a ground o a V_{dd} (S a V_{dd} per il p, S a gnd per l'n è un assetto comune). Possiamo guardare con un occhio diverso la struttura dell'inverter, individuando dispositivi passivi e giunzioni che possiamo rappresentare come diodi. Si nota che si vengono a creare dei BJT parassiti con resistenze di contatto tra le basi e le rail di alimentazione.

Il pnp ha l'emettitore collegato al pozetto p+ ed è fissato a V_{dd} . La base è la n well, che fa anche da collettore per l'npn, il collettore è la p well, che fa anche da base per l'npn. Questa struttura così alimentata non è attiva, non c'è alcun passaggio di corrente. L'npn ha la base a massa, la sua V_{be} è a 0, la corrente di collettore è nulla, per cui anche il pnp è spento (vale anche il viceversa). Se però c'è un impulso sulla base dell'npn e si accende appena la giunzione b-e, scorre una corrente collettore amplificata, la quale produce una caduta sulla R_{well} . Il potenziale sulla base del pnp è $V_{dd} - R_{well} \cdot I_{c_{npn}}$, il quale può andare sotto di una V_T rispetto all'emettitore, dunque il pnp può accendersi anch'esso. A questo punto si verifica latch up. A prescindere dalla causa che ha scaturito l'effetto, si innesta un anello di reazione positiva: aumenta la corrente dell'npn, più il pnp si accende più la sua hie diventa piccola, con conseguente maggiore prelevamento di corrente proveniente dall'npn. La corrente in gioco aumenta fino a distruzione del circuito.

Il passaggio di corrente nelle resistenze è la cosa critica. Può succedere che durante le transizioni vengano iniettate cariche attraverso le capacità tra giunzione e p well. Se la well fosse flottante, cioè se non fosse collegata a massa, anche un accoppiamento capacitivo minimo potrebbe farne cambiare il potenziale. Ad impedire questo fenomeno è proprio la resistenza, che forma un filtro con le capacità ed attenua tanto più quanto è bassa la resistenza. L'attenuazione sarebbe completa se la Rsub fosse nulla, per cui una regola che se ne deduce è quella di ridurre le resistenze in gioco.

Non tutta la p well partecipa all'effetto transistore. La parte critica, in questo senso, è quella che si trova in corrispondenza dell'area attiva, cioè laddove si forma l'emettitore. In prossimità delle aree attive occorre che la resistenza del substrato/well sia piccola, così da minimizzare le cadute agli eventuali passaggi di corrente (dallo schema equivalente si apprezza anche che se la resistenza è più piccola, è minore la frazione di corrente che polarizza la base del transistore). La regola di layout che cura l'aspetto del latch up è quella di inserire più prese di substrato il più vicino possibile a tutte le aree attive che formano giunzioni, in modo da appiattire e fissare il potenziale del substrato stesso evitando la polarizzazione diretta delle giunzioni. La regola, in particolare, pone una massima distanza R a cui può essere inserita la presa (distanza calcolata da ogni punto dell'area attiva a cui si riferisce).



Every part of a p-active area in an n-well must "see" at least one well-tap with a maximum distance R

Every part of an n-active area in the substrate (p-well) must "see" at least one substrate-tap with a maximum distance R

Se c'è spazio si mettono più prese di substrato, così da diminuire ancora la resistenza e rendere il substrato il più equipotenziale possibile (non è una regola, ma è buona prassi). Queste accortezze, in ambito mixed signal, riducono anche l'accoppiamento del substrato con circuiti di segnale, evitando che si formino canali di disturbo e il così detto "rumore di substrato". In ogni caso, almeno una presa entro la distanza massima è obbligatoria.

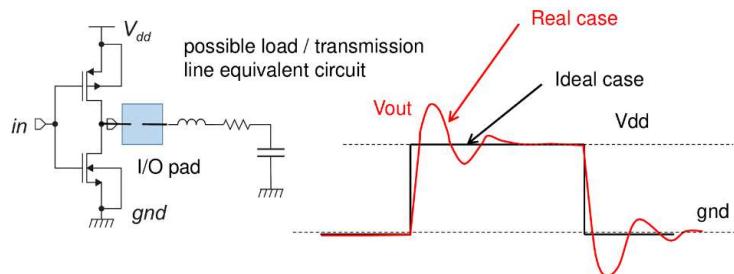
Lo stesso vale per la n well, che fa da base per il pnp, ma in generale per tutte le aree attive drogate p^+ : la well dovrà avere ad una distanza che non può superare R una presa di well fissata a V_{dd} . Questa regola riguarda quei circuiti che non si interfacciano con il mondo esterno, che non hanno alcun terminale connesso ai pad (core logic).

Il substrato tipico di un processo CMOS di tipo bulk molto spesso vede uno strato sottile epistassiale meno drogato ad alta resistenza, nel quale si realizzano i drogaggi di well. I drogaggi di well, anche se aumentano con lo scaling dei dispositivi, sono comunque medio bassi, ovvero abbastanza resistivi. La presenza di uno strato sottostante pesantemente drogato abbassa le resistenze. Questo è un accorgimento tecnologico, non progettuale.

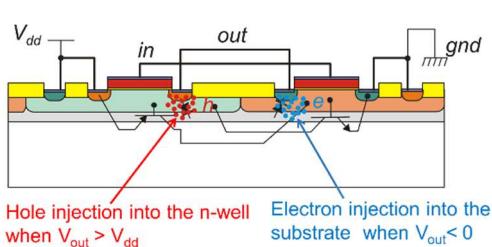
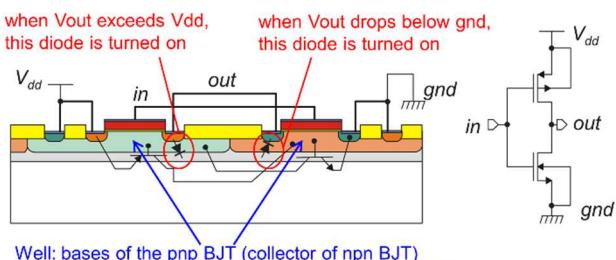
Sono presenti anche più bipolar parassiti. Ad esempio, ogni MOSFET è intrinsecamente anche una struttura p^+np^+ o n^+pn^+ . Questi bipolar, però, in genere sono spenti. Ad esempio, nel caso dell'nMOS la base p coincide al substrato, per cui è polarizzata al potenziale più basso del circuito. Questo fa sì che nessuna delle due giunzioni riuscirà ad accendersi rispetto alla base. Immaginiamo, però, che l'uscita dell'inverter sia a V_{dd} . A causa della resistenza parassita della p well, il potenziale della base non sarà a ground. La giunzione drain-substrato è polarizzata in inversa e la corrente inversa può polarizzare la base attraverso il collettore (corrispondente al drain). All'aumentare della tensione inversa la corrente di saturazione inversa aumenta. Inoltre, aumenta anche la zona di svuotamento, per cui l'estensione della base efficace del BJT parassita si riduce, con conseguente aumento del guadagno. Se questo bipolare si accende causa un passaggio di corrente da drain a source per un effetto diverso da quello controllato dal gate. Il fenomeno prende il nome di snap-back e determina la massima V_{ds} che può sopportare il MOSFET. Anche la fusione delle zone di svuotamento (punch-through), che causa un simil-corto tra drain e source, partecipa a determinare la V_{ds} massima.

Circuiti di input output

A differenza dei circuiti di core logic, i circuiti di input output vedono il mondo esterno, si interfacciano ai pad. Immaginiamo di avere un inverter di output (se fosse di input avrebbe i diodi di protezione). L'uscita è collegata ai pad, ma nell'andare verso il mondo esterno si vede sempre un carico RLC; il bonding wire o una generica interconnessione mostrano sempre un'autoinduttanza, una resistenza serie e una capacità verso massa (carico risonante).



La risposta all'impulso dipende dal Q del circuito. Se l'induttanza non è trascurabile si ha un po' di ringing. Queste oscillazioni sono classiche e presentano overshoot da L a H, e undershoot da H a L. La tensione può temporaneamente superare la V_{dd} o scendere al di sotto del ground. Occorre indagare le conseguenze di questo comportamento, ed evitare che a seguito di sovraelungazioni della tensione, o dovute dinamicamente a ringing o imposte dai pilotaggi del pin dall'esterno, parta il latch up.

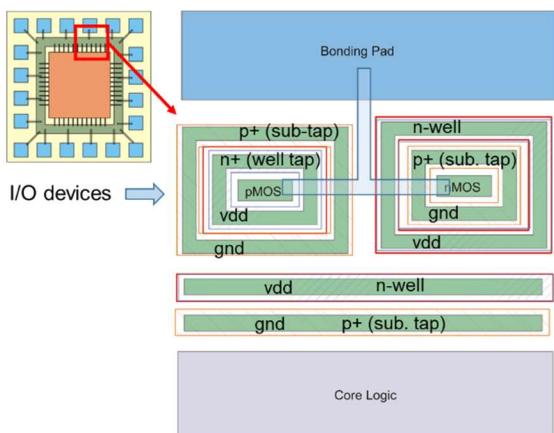


I pozetti di drain formano con il body del rispettivo MOSFET una giunzione pn. Di solito queste due giunzioni sono in inversa. Se $V_{out} > V_{dd}$ il drain del pMOSFET può andare in diretta nei confronti della n well, ed iniettare lacune nella stessa. Analogamente, se $V_{out} < 0$ la p well (substrato) può andare in diretta nei confronti del pozetto di drain dell'nMOSFET, il quale inietta elettroni nel substrato.

L'accensione dinamica di queste giunzioni può attivare le basi dei BJT parassiti accendendoli nelle zone in cui il $\beta \gg 1$. Gli elettroni iniettati dal pozetto di drain dell'nMOS nel substrato possono essere raccolti dalla n well accanto che fa da collettore per il pnp. Analogamente, le lacune iniettate dal pozetto di drain del pMOS nella n well possono essere raccolte dalla p well che fa da collettore per il npn.

Se l'inverter commuta rapidamente si può anche determinare un accumulo dei portatori iniettati. Questo ci dice che per i dispositivi di input/output non è sufficiente porre delle prese che fissino il potenziale substrato/well, perché il pozetto, che è in qualche modo affacciato all'esterno, può sempre e comunque mandare in diretta la giunzione. È più facile, in questo caso, che i transistori bipolari parassiti si accendano fortemente, cioè che il loro β diventi grande.

Sorgono delle regole particolari, specifiche per i dispositivi di input/output, che si occupano di gestire l'iniezione dei minoritari e far sì che non comporti latch up. Le regole valgono anche per i dispositivi di input. Per proteggere questi dispositivi dall'ESD si pongono dei diodi di protezione che, all'eventuale accensione, iniettano anch'essi minoritari nel substrato o nella well, con conseguente potenziale accensione di qualche BJT parassita.



Anzitutto, per evitare l'iniezione di minoritari dalla parte di input/output alla core logic quest'ultima viene circondata da una presa distribuita di well a V_{dd} e una di substrato a gnd. Per proteggere dal latch up la parte di input/output si separano i transistori nMOS e pMOS tra loro fisicamente. Si pongono poi degli anelli di guardia attorno ai dispositivi, cioè delle prese distribuite che li circondano completamente. Il primo anello di guardia fa sì che il potenziale delle well/del substrato sia uniforme e che le resistenze tra le aree attive e V_{dd} o ground siano più basse possibili. Il secondo anello di guardia impedisce che gli eventuali minoritari raggiungano l'altro bjt parassita nel dispositivo complementare.

Talvolta può bastare un solo anello di guardia; dipende dalle particolari regole di latch up esterno. Anche i dispositivi stessi possono essere realizzati con una struttura concentrica, con il drain al centro visto che è il più delicato (possibile emettitore, si affaccia all'esterno). Tecniche del genere rendono i dispositivi resistenti anche a latch up dovuti a radiazioni ionizzanti. Nel caso di circuiti mixed signal la parte digitale induce rumore nel substrato attraverso accoppiamenti capacitivi. Circondando tutta la zona rumorosa con un guard ring si inibisce la propagazione del rumore digitale al substrato; ancora meglio se si circondano anche i circuiti analogici a loro volta. Spesso la creazione degli anelli di guardia è parametrica e automatica.

Questi trattamenti si limitano alle sole parti di circuito che sono collegate ai pad di input/output. Proteggere con la solita tecnica anche i dispositivi interni sarebbe insostenibile a livello di ingombro.

Manuale di processo (Design Rule Manual PSM 025-MM)

Analizzeremo adesso un manuale di processo MM (“mixed mode”), valido sia per progettazione analogica che per progettazione digitale. Ciò significa che all’interno della libreria dei dispositivi si troveranno anche degli elementi passivi (condensatori, resistori) che di solito non servono per la progettazione digitale. Solitamente l’analogo ha bisogno di una caratterizzazione più precisa, fedele e ampia del dispositivo. Il manuale è in versione educational, ma è comunque molto simile per struttura ai manuali di processo veri.

Process description

The PSM025 is an *n*-well – single poly – double metal (1P2M) CMOS process with Al-BEOL (Aluminum Back End of Line). Minimum channel length for both *n*-MOS and *p*-MOS devices is $0.25 \mu\text{m}$. Polysilicon and diffusions are silicided for sheet resistance reduction. Use of tungsten plugs allows stacking of vias and contacts.

General rules

- All dimensions in this manual are expressed in microns.
- All dimensions are absolute minimum. Larger values should be used whenever possible. The only exception is represented by contacts and vias, which must be drawn at nominal values only.
- Notches are to be considered as spaces and should comply with the corresponding rules.
- Non-orthogonal lines are prohibited.
- The database must be digitalized on a 0.05-micron grid.

Il fatto di avere plug in tungsteno e la planarizzazione permettono lo stacking: i contatti che collegano i livelli di metal si possono impilare l’uno sopra l’altro. Nei processi vecchi si poteva salire per i livelli di metal solo con strutture a scala a chiocciola. Oggi si possono realizzare ascensori verticali di contatti/vie che portano direttamente i drain e i source all’ultima metallizzazione. Il database è l’ambiente di lavoro, il foglio di lavoro. La regola che riguarda il database ci dice che ogni distanza può essere incrementata rispetto alla precedente di 50 nm . La lunghezza minima di canale è 250 nm . Si intuisce una certa discretizzazione della risoluzione del layout. Un buon design kit approssima autonomamente i valori a quelli consentiti più vicini.

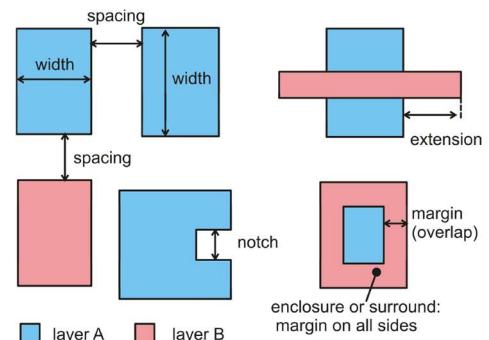
Tooling layers

Layer name	Description	GDS N.	Note
n-well	N-well implant	1	Standard
active	Active areas definition	2	Standard
poly	Polysilicon	3	Standard
n_plus	<i>n</i> + implant	4	Standard
p_plus	<i>p</i> + implant	5	Standard
contact	Contact layer for connecting Metal1 to Poly or Active	6	Standard
metal1	First metal interconnect layer (Al)	7	Standard
via	For contacting metal1 to metal2	8	Standard
metal2	Second metal interconnect layer (Al)	9	Standard
passivation	Passivation opening for bonding purposes	10	Standard
siprot	Silicide protection: inhibits silicide formation over active and poly	11	Optional
hires	Selects polysilicon areas with reduced doping for high value resistors	12	Optional
metal1_opt	Optional metal 1 layer for lower MIM capacitor plate (between M1 and M2)	13	Optional

Topological Layout Rules (TLR)

Conventions used in this manual: topological rule types.

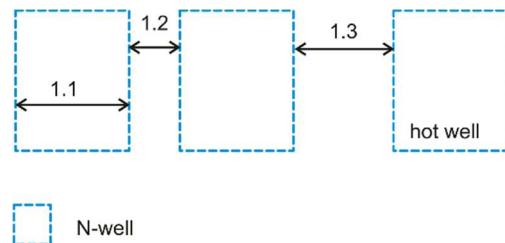
- Width
- Spacing
- Margin or overlap (enclosure when on all sides)
- Extension



Si passa alle regole di processo. Solitamente si parte dall'n well, dai layer che occorrono cronologicamente prima nel processo

N-well

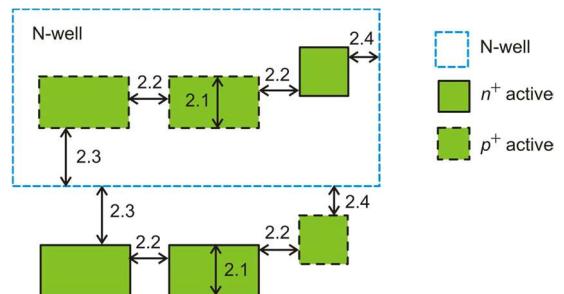
Rule	Description	Value (μm)
1.1	Minimum width	1.5
1.2	Minimum spacing between wells at same potential	1.0
1.3	Minimum spacing between wells at different potential (hot wells)	2.2



Vengono trattati due spacing, uno tra well allo stesso potenziale, per esempio entrambe a V_{dd} , e uno tra due well a potenziale diverso, ad esempio una a V_{dd} e l'altra a $V_{dd}/2$. Queste ultime prendono il nome di hot wells e sono problematiche se la zona p che le divide è troppo piccola. Infatti, con potenziali diversi delle well questa struttura potrebbe accendersi da BJT parassita, con la base che è proprio la zona p. L'allontanamento garantisce che la base sia grande abbastanza da avere un β sufficientemente piccolo. Di solito si considerano hot well tutte le well vicine che non sono collegate assieme.

Active

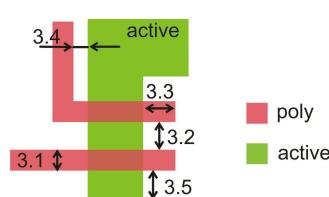
Rule	Description	Value (μm)
2.1	Minimum width	0.5
2.2	Minimum spacing	0.5
2.3	Source/drain active to well edge	0.8
2.4	Substrate/well contact active to well edge	0.5



Le regole fanno distinzione tra le aree attive aperte nelle well e quelle aperte nel substrato. Il layer non è diverso, saranno diversi i dispositivi conseguenti a seconda dei drogaggi. In ogni caso si anticipa tutto ciò con regole specifiche da rispettare. La regola che riguarda i contatti è più leggera rispetto a quella che riguarda le aree attive per i source/drain. La ragione è che se anche il contatto toccasse il bordo troverebbe comunque una giunzione in inversa.

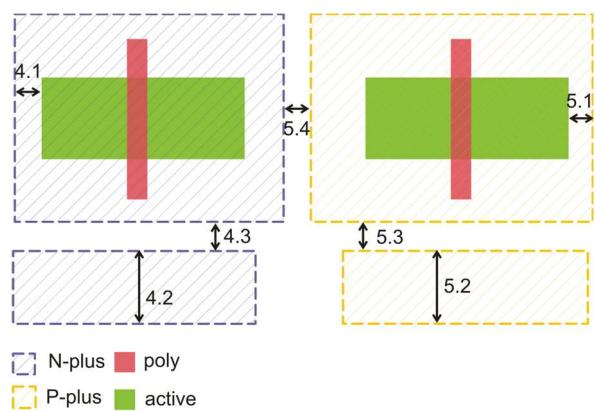
Poly (polysilicon)

Rule	Description	Value (μm)
3.1	Minimum width	0.25
3.2	Minimum spacing	0.5
3.3	Minimum gate extension of active	0.8
3.4	Minimum field poly to active spacing	0.2
3.5	Minimum active extension of poly	0.5



N-plus/P-plus Source/Drain implant

Rule	Description	Value (μm)
4.1	Minimum N-plus overlap of active	0.3
4.2	Minimum N-plus width	0.5
4.3	Minimum N-plus spacing (merge whenever possible)	0.5

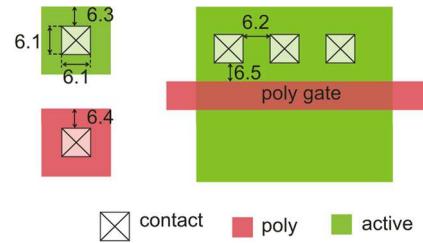


Rule	Description	Value (μm)
5.1	Minimum P-plus overlap of active	0.3
5.2	Minimum P-plus width	0.5
5.3	Minimum P-plus spacing (merge whenever possible)	0.5
5.4	Minimum N-plus to P-Plus spacing	0.5

Contacts

La realizzazione dei contatti è un passaggio di processo che può essere inserito nel FEOL o nel BEOL

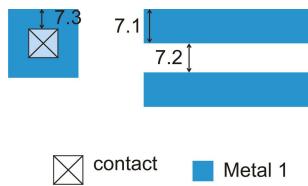
Rule	Description	Value (μm)
6.1	Exact contact size	0.3
6.2	Minimum spacing	0.4
6.3	Minimum margin to active area	0.2
6.4	Minimum margin to polysilicon area	0.2
6.5	Minimum spacing to polysilicon gate	0.25



Per contattare il poly con una metal (es. gate all'ingresso di un inverter) occorre impiegare un contatto circondato completamente dal margine 6.4. Per quanto riguarda l'area attiva il margine è regolato dalla 6.3. Pertanto, la geometria minima di un contatto ha un'estensione complessivamente di $0.7 \mu\text{m}$.

Metal 1

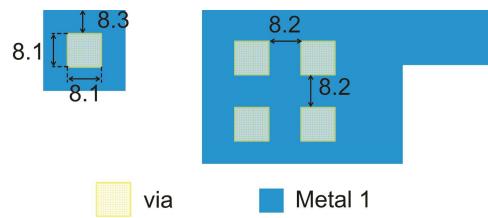
Rule	Description	Value (μm)
7.1	Minimum width	0.5
7.2	Minimum spacing	0.5
7.3	Minimum overlap of contact	0.2



Anche la metal deve circondare il contatto con un margine (7.3). Alla fine, un contatto tra poly e metal occuperà un quadrato $0.7 \times 0.7 \mu\text{m}$ con al centro un quadratino $0.3 \times 0.3 \mu\text{m}$, e un altro quadrato $0.7 \times 0.7 \mu\text{m}$ di metal al di sopra.

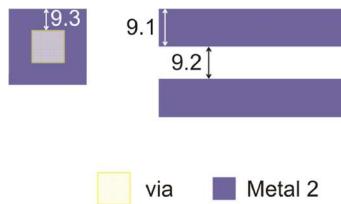
Via

Rule	Description	Value (μm)
8.1	Exact via size	0.3
8.2	Minimum spacing	0.4
8.3	Minimum margin to metal 1	0.2
8.4	Stacking of vias and contacts is allowed	



Metal 2

Rule	Description	Value (μm)
9.1	Minimum width	0.5
9.2	Minimum spacing	0.5
9.3	Minimum overlap of via	0.2



Latch-UP prevention rules

Core devices

Rule	Description	Value (μm)
20.1	Maximum distance between an n+ active area on substrate (n-mos source/drain diffusion) and the closest substrate contact (substrate tap)	25
20.2	Maximum distance between a p+ active area on n-well (p-mos source/drain diffusion) and the closest well contact (well tap)	25

Input/ output devices

NMOS I/O transistor

- Rule 20.3: a $p+$ guard ring covered by metal 1 and filled by as much as possible contacts should be placed around any single n -MOSFET. The guard ring should be connected to Vss.
- Rule 20.4: an $n+$ collector guard ring, embedded into an n -well ring must be placed around the $p+$ guard ring. The $n+$ guard ring must be covered by metal 1 and filled by as much as possible contacts. The $n+$ collector guard ring should be connected to Vdd.

PMOS I/O transistor

- Rule 20.5: an $n+$ guard ring covered by metal 1 and filled by as much as possible contacts should be placed around any single p -MOSFET inside the n-well that includes the p-MOSFET. The guard ring should be connected to Vdd.
- Rule 20.6: a $p+$ collector guard ring must be placed around the n-well that includes the p-MOSFET. The $p+$ collector guard ring must be covered by metal 1 and filled by as much as possible contacts. The $p+$ collector guard-ring must be connected to Vss.

Electromigration rules: current limits through interconnections and contacts

Le regole seguenti riguardano il fenomeno dell'**elettromigrazione**: nel tempo il passaggio di corrente porta alla rottura delle interconnessioni. I portatori, soprattutto tra i bordi dei grani dei materiali policristallini, possono cedere quantità di modo agli atomi e determinare zone di scarsità di materiale (incremento di resistenza), o zone di accumulo di materiale (decremento di resistenza). Le regole che contrastano il fenomeno di elettromigrazione si traducono in regole di massima per la densità di corrente attraverso piste, contatti e vie.

Interconnections

Linear current density ($mA/\mu m$)

Layer	Width (μm)	I max at 80 °C (W in μm)	I max at 100 °C (W in μm)	I max at 125 °C (W in μm)
Metal1	$0.5 \leq W \leq 1$	$1mA \times (W + 0.5)$	$0.5 mA \times (W + 0.5)$	$0.3mA(W + 0.33)$
	$W \geq 1$	$1.5 mA \times W$	$0.75 mA \times W$	$0.4 mA \times W$
	$W = 0.5$	$1mA$	$0.5 mA$	$0.25 mA$
Metal2	$0.5 \leq W \leq 1$	$1mA \times (W + 1)$	$0.5mA \times (W + 1)$	$0.25mA \times (W + 1)$
	$W \geq 1$	$2 mA/\mu m$	$1 mA/\mu m$	$0.5 mA \times W$
	$W = 0.5$	$1.5 mA$	$0.75 mA$	$0.375 mA$

Vias and contacts

Current for a single via / contact

Type	80 °C	100 °C	125 °C
Contact	0.7 mA	0.45 mA	0.3 mA
Via	1 mA	0.75 mA	0.5 mA

Più alta la temperatura, minore è la corrente che la pista può portare. Il fenomeno dell'elettromigrazione è spalmato nel tempo, non si discutono in questo caso margini di danneggiamento istantaneo. Se ad esempio si ha una pista di metal con $W = 1 \mu m$, $I_{max} = 1.5 mA \times W$ significa che la corrente massima per tale pista è $1.5 mA$. Contatti e vie possono essere posti in parallelo per aumentare la portata di corrente.

Mosfet electrical parameters

NMOS electrical parameters

Parameter	W / L ($\mu m/\mu m$)	Min	Typ	Max
Vt	10 / 10	0.32 V	0.38 V	0.44 V
Vt	10 / 0.25	0.48 V	0.53 V	0.58 V
Vt	0.5 / 0.25	0.4 V	0.45 V	0.51 V
μ_nC_{ox}	10 / 10	$180 \times 10^{-6} A/V^2$	$240 \times 10^{-6} A/V^2$	$280 \times 10^{-6} A/V^2$
γ (body effect factor)	10 / 10	–	0.44 V^{0.5}	–

PMOS electrical parameters

Parameter	W / L ($\mu m/\mu m$)	Min	Typ	Max
Vt	10 / 10	-0.48 V	-0.56 V	-0.64 V
Vt	10 / 0.25	-0.45 V	-0.5 V	-0.55 V
Vt	0.5 / 0.25	-0.4 V	-0.44 V	-0.5 V
μ_pC_{ox}	10 / 10	$40 \times 10^{-6} A/V^2$	$50 \times 10^{-6} A/V^2$	$60 \times 10^{-6} A/V^2$
γ (body effect factor)	10 / 10	–	0.6 V^{0.5}	–

Matching parameters (Pelgrom area parameters)

Parameter	NMOS	PMOS
C_{V_t}	$8.5 mV \cdot \mu m$	$8.5 mV \cdot \mu m$
C_β	$0.03 \mu m$	$0.03 \mu m$

Parameters of parasitic devices

Resistances

Type	Unit	Min (Fast)	Typ.	Max. (Slow)
N-Well Sheet Resistance	Ohm/square	400	500	600
N+ / P +Sheet Resistance	Ohm/square	3.5	5	8.5
N+ Sheet Resistance (non salicide)	Ohm/square	65	80	100
P+ Sheet Resistance (non salicide)	Ohm/square	85	120	150
Poly Sheet Resistance	Ohm/square	3.0	5.0	8.0
Poly N+ Sheet Resistance (non salicide)	Ohm/square	80	100	135
Poly P+ Sheet Resistance (non salicide)	Ohm/square	155	180	205
HR Poly	Ohm/square	800	1.0k	1.2k
Metal 1 Sheet Resistance	Ohm/square	0.06	0.08	0.1
Metal 2 Sheet Resistance	Ohm/square	0.055	0.065	0.085
N+ /Metal 1 contact Resistance.	Ohm	5.0	10	20
P+ /Metal 1 contact Resistance	Ohm	5.0	10	20
Poly /Metal 1 contact Resistance	Ohm	4.0	8.0	12
Metal1 /Metal 2 Via resistance	Ohm	2.5	5.0	9.0

Capacitances

Mosfet Capacitances

Type	Unit	Min (Fast)	Typ.	Max. (Slow)
Gate to substrate / Gate to well (area)	$fF/\mu m^2$	5.5	6.0	6.5
N+ / Substrate (area)	$fF/\mu m^2$	1.66	1.85	2.05
N+ / Substrate (edge)	$fF/\mu m^2$	0.35	0.38	0.42
P+/Well (area)	$fF/\mu m^2$	1.7	1.9	2.1
P+/Well (edge)	$fF/\mu m^2$	0.36	0.39	0.43

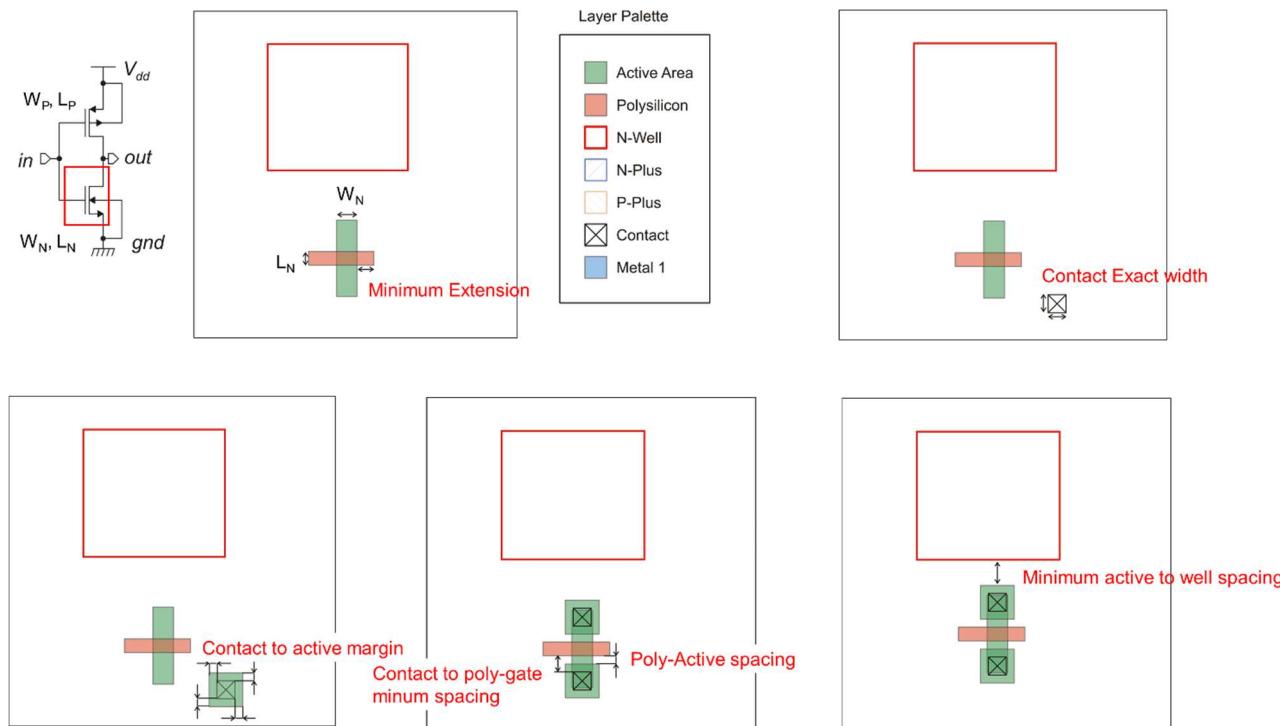
La capacità del MOS è riferita solo all'area attiva. Non viene fornita la capacità per unità di perimetro poiché l'ossido, in questo caso, è molto sottile.

Interconnect Related Capacitances

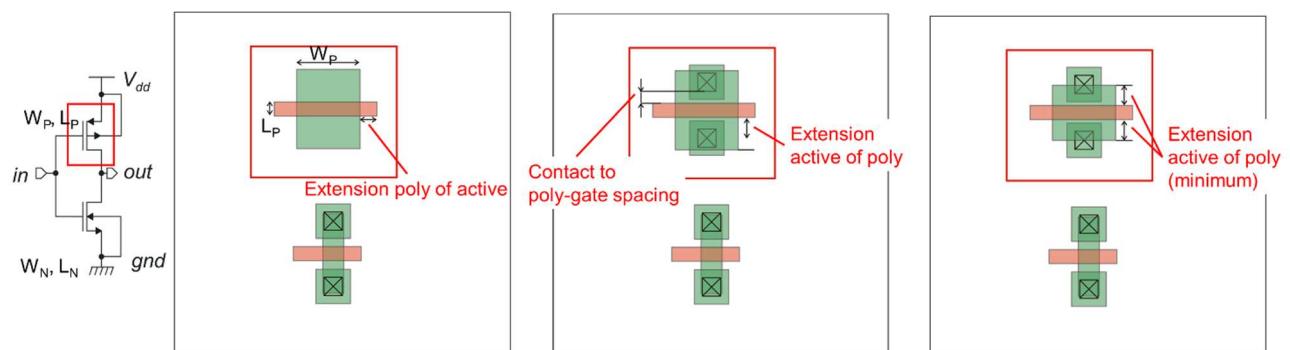
Type	Unit	Min (Fast)	Typ.	Max. (Slow)
Poly to substrate (area)	$fF/\mu m^2$	0.09	0.1	0.11
Poly to substrate (fringe)	$fF/\mu m^2$	0.075	0.08	0.09
Metal 1 to substrate (area)	$fF/\mu m^2$	0.027	0.03	0.033
Metal 1 to substrate (fringe)	$fF/\mu m^2$	0.035	0.04	0.045
Metal 1 to Poly (area)	$fF/\mu m^2$	0.053	0.06	0.065
Metal 1 to Poly (fringe)	$fF/\mu m^2$	0.059	0.065	0.071
Metal 2 to substrate (area)	$fF/\mu m^2$	0.012	0.015	0.018
Metal 2 to substrate (fringe)	$fF/\mu m^2$	0.063	0.07	0.078
Metal 2 to Poly (area)	$fF/\mu m^2$	0.025	0.03	0.035
Metal 2 to Poly (fringe)	$fF/\mu m^2$	0.072	0.08	0.088
Metal 2 to Metal 1 (area)	$fF/\mu m^2$	0.035	0.04	0.045
Metal 2 to Metal 1 (fringe)	$fF/\mu m^2$	0.054	0.06	0.066

Progettazione di un inverter CMOS

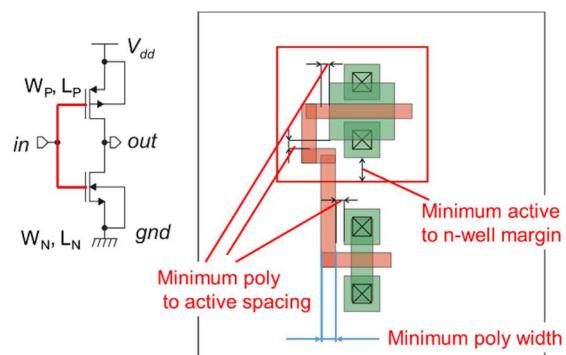
n-MOS



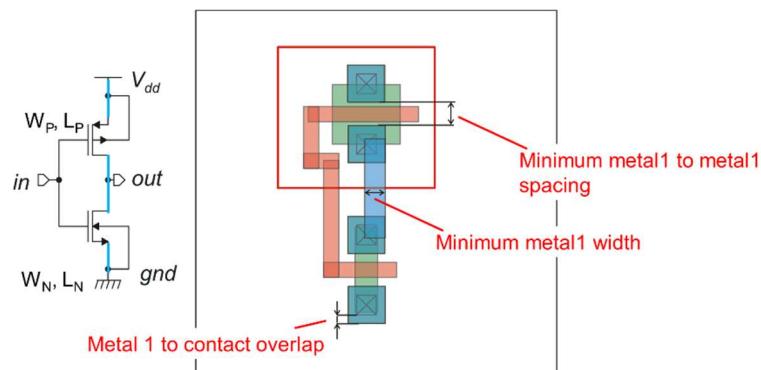
p-MOS



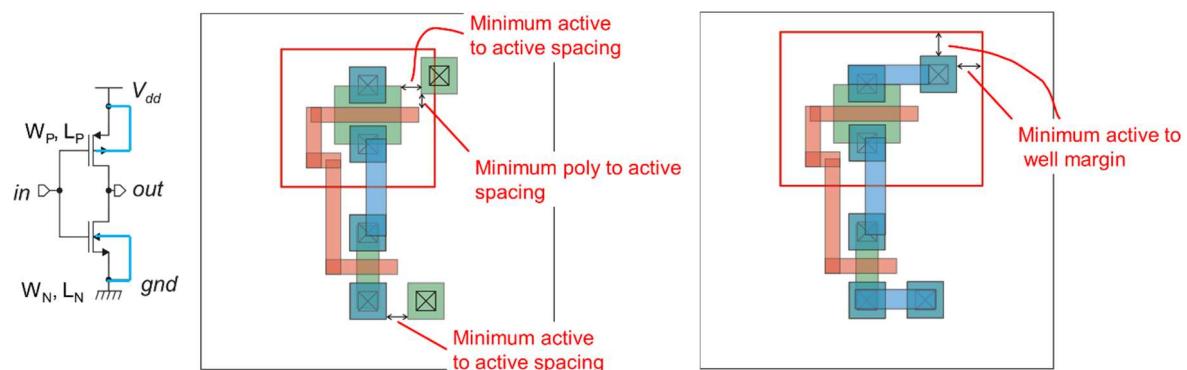
Gates



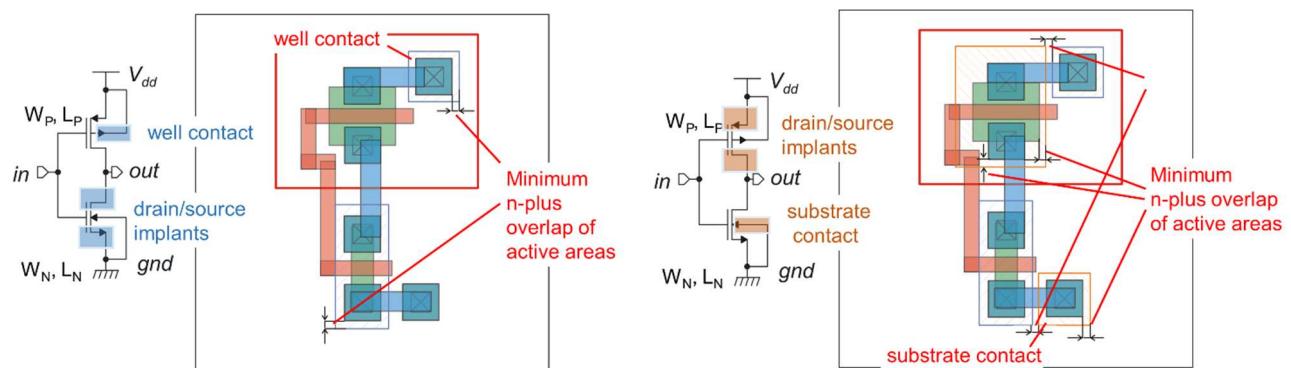
Metal 1 connection



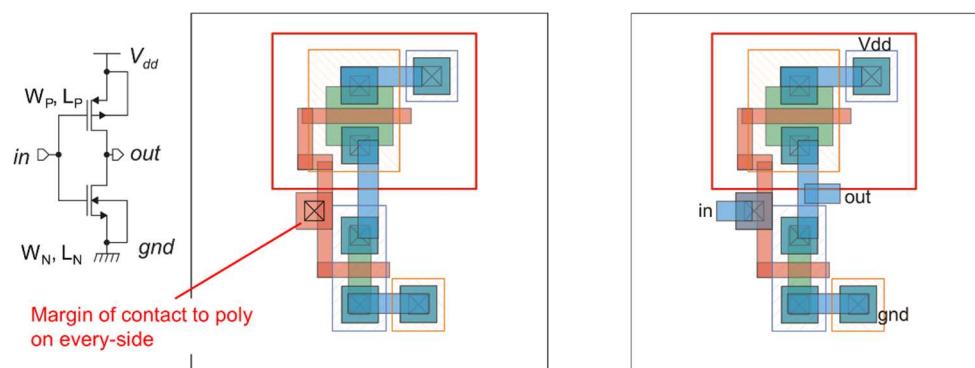
Substrate and n-well contacts



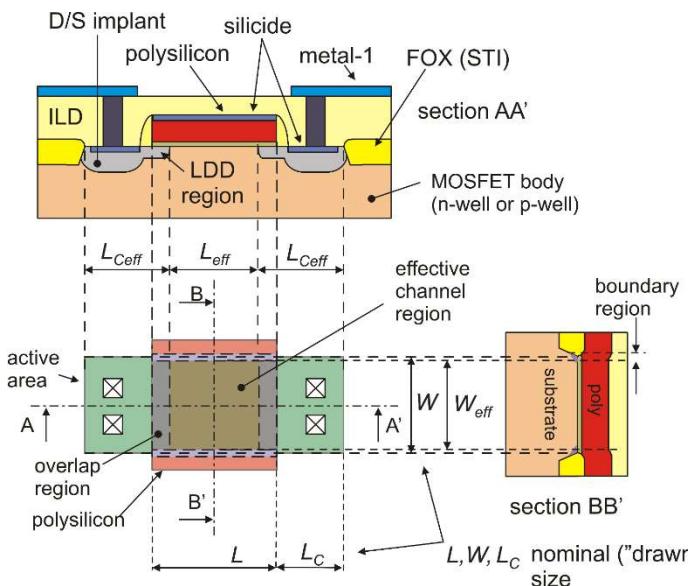
Plus implants



Poly-metal contact



Modello fisico del MOSFET



A sinistra si osserva un modello approssimato del MOSFET planare. Si nota che i pozzetti di drain e source si estendono al di sotto del gate con le regioni LDD (Light Doping Drain). Questo aspetto tecnologico cura la corrente di leakage dovuta agli elettroni caldi che scatterano e rimangono intrappolati nell'ossido. A causa delle sotto-diffusioni il canale ha una lunghezza efficace minore rispetto a quella di layout. Anche la larghezza del canale effettiva è minore a causa delle zone di svuotamento che si formano ai pozzetti.

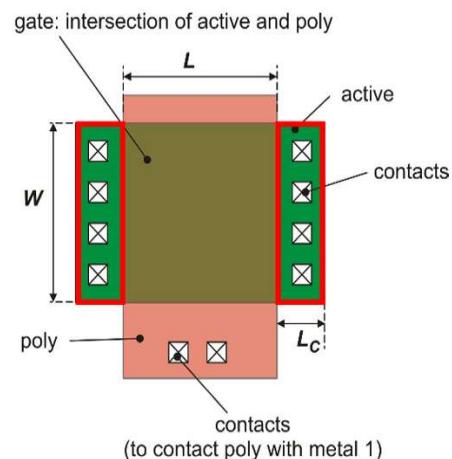
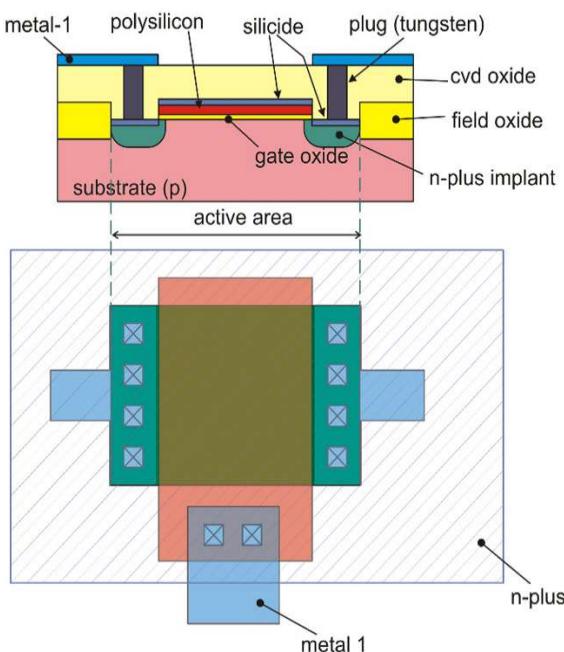
$$L_{eff} = L - 2L_D$$

$$W_{eff} = W - 2W_D$$

Le stesse non idealità di bordo affliggono anche le resistenze, che non risultano perfettamente uniformi. Il MOSFET è caratterizzato elettricamente con L_{eff} e W_{eff} , di cui si tiene conto nel modello. Gli effetti di bordo pesano di più quando gli oggetti sono più piccoli. Si definisce con L_C la lunghezza minima delle aree di drain e source. Se possibile si cerca di minimizzare L_C in modo da ridurre l'ingombro e le capacità parassite. La regola su L_C è una regola composita: dipende dal margine di area attiva rispetto al contatto, dalla dimensione minima del contatto stesso, la distanza dei contatti dal gate...

Visione semplificata del progettista

Si riporta una visione semplificata dell'nMOSFET:



$$A_D = A_S = WL_C$$

$$P_D = P_S = 2L_C + 2W$$

In questo caso le aree di source e drain sono uguali, ma non sempre è così. Area e perimetro dei pozzetti sono necessari per estrapolare le capacità parassite delle giunzioni verso il substrato/well che le alloggia.

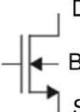
Modello elettrico MOSFET (grandi segnali)

Per modello del MOSFET si intende un sistema di equazioni e disequazioni che ne descrivono il comportamento. L'aspetto principale di un MOSFET è la variazione della corrente di drain con controllo in tensione al gate. Un modello inserito nell'ambito della progettazione microelettronica deve rappresentare tutti i dispositivi che il processo permette di realizzare cambiando L e W . Cioè, il modello deve tenere conto di come cambiano i parametri al variare delle dimensioni del dispositivo. L'altro problema che si incontra è quello delle regioni di funzionamento (debole inversione, forte inversione, saturazione); i modelli devono essere accurati anche in debole inversione. Rappresentare il comportamento di un MOSFET in qualunque zona e condizionarlo alle dimensioni è difficile.

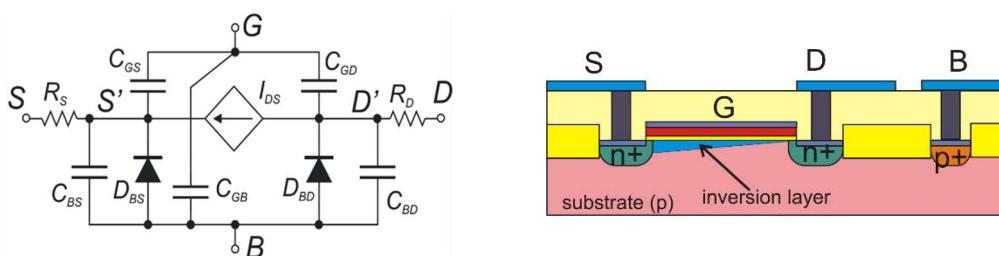
Da una parte si vorrebbe un modello accurato che possa prevedere con accuratezza il comportamento del circuito. D'altra parte, quanto più è accurato il modello tanto meno il comportamento del dispositivo risulta intuitivo e tanto più il numero di gradi di libertà e di variabili progettuali con cui raggiungere le specifiche del circuito aumenta. Detto altrimenti, l'elevata accuratezza si trasforma in equazioni che irrisolvibili oppure in molteplici approssimazioni che introducono molti vincoli nella validità dell'analisi.

Esistono tanti modelli, evoluti nel tempo, di cui possiamo distinguere alcune categorie. La prima è quella dei modelli per le simulazioni elettriche. Questi modelli sono molto complessi, fanno uso di centinaia di parametri che descrivono la fisica del dispositivo e parametri di fitting per avvicinare le caratteristiche a quelle misurate sperimentalmente. I modelli per le simulazioni più in uso sono BSIM (Berkeley Short-channel IGFET Model, curano la modellizzazione degli effetti di canale corto, debole inversione), EKV (Enz, Krummenacher, Vittoz è un modello più vicino alla realtà fisica), NXP; quelli più usati sono i BSIM. La fonderia stabilisce la tipologia di modello più accurata per la descrizione dei dispositivi che produce.

L'altra categoria è quella dei modelli di calcolo a mano. La progettazione non può prescindere da un dimensionamento di massima, da una stima generale fatta su carta. A tal proposito servono modelli molto semplici (il modello EKV è quasi utilizzabile a mano). I risultati che si ottengono dai calcoli a mano, poi, se hanno senso si ritoccano e si raffinano per mezzo del simulatore. Se non si è grado di dare una prima descrizione del circuito a mano, stipulando un primo dimensionamento a mano, il simulatore è del tutto inutile (garbage in – garbage out).

 D'ora in avanti caratterizzeremo il comportamento dell'nMOSFET, il complementare verrà di conseguenza. Il MOSFET integrato è rigorosamente un dispositivo a quattro terminali: Drain, Source, Gate, Body. Nei MOSFET discreti generalmente il body è connesso al source internamente.

Il modello per grandi segnali:



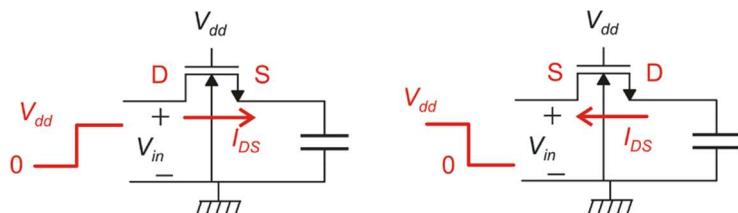
Affinché le giunzioni tra il substrato e pozetti non si accendano in diretta è necessario che i potenziali di source e drain siano sempre maggiori del potenziale di body. La corrente che scorre effettivamente nel canale I_{DS} , a causa della presenza delle capacità e delle giunzioni parassite non sarà uguale alle correnti di drain e di source I_D , I_S . Tuttavia, in continua si assumerà sempre e comunque che:

$$I_D \cong I_S \cong I_{DS} \quad R_S = R_D \cong 0$$

Gli elementi parassiti sono trascurabili dipendentemente dalle applicazioni. Nell'ambito della progettazione ultra low power le correnti di leakage attraverso le giunzioni possono essere significative. Oltre ai terminali discussi si notano dei terminali con l'apice. Il fatto di avere pozzi di drain e source di materiali dotati di una certa resistività rende necessario considerare dei drain e source ideali, D' e S' , ai quali accediamo attraversando in serie le resistenze dei materiali. Ad alte frequenze queste resistenze possono porre delle singolarità, poli non trascurabili. Per alte correnti diventano importanti anche in continua. Nei MOSFET moderni con aree attive siliciumizzate possiamo trascurare queste resistenze. Le capacità, invece, iniziano a contare nel caso di progettazione RF. La corrente di drain è la corrente che entra nel drain, la corrente di source è quella che entra nel source. Le correnti positive, teoricamente, sono quelle entranti. Nella nostra trattazione adotteremo i versi naturali, cioè quelli che rendono le correnti positive. La corrente nel drain, quindi, è considerata entrante, mentre quella di source uscente (per il pMOSFET è al contrario).

Source e drain effettivi

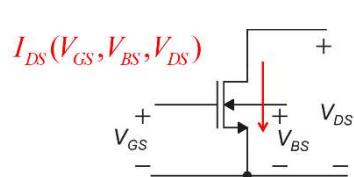
Il MOSFET planare è perfettamente simmetrico, per cui drain e source sono interscambiabili a seconda del contesto. Anche in fase di simulazione, scambiando drain e source, si ottiene lo stesso risultato. Tuttavia, ci sono equazioni che utilizzano il source come terminale di riferimento per le tensioni (V_{GS} , V_{DS} , V_{BS}); in tal caso occorre rispettare i terminali assegnati. In particolare, il source effettivo per un nMOSFET è il pozzetto che in un dato momento si trova a un potenziale più basso rispetto l'altro pozzetto (drain effettivo). Dunque, queste equazioni necessitano che si abbia conoscenza a priori di quale tra i due terminali sta funzionando da source. Ci sono anche equazioni che non pongono il riferimento al source, ma a uno dei terminali rispetto cui vale la simmetria geometrica del dispositivo. Ad esempio, il modello EKV è costituito da equazioni riferite al terminale di body. In tal caso non importa stabilire a priori drain e source. Per il pMOSFET il source effettivo è quello che a un dato momento si comporta come tale, che cioè è a potenziale maggiore rispetto al drain. Con questa definizione in situazioni transitorie i drain e source efficaci possono scambiarsi a seconda del potenziale raggiunto dai terminali.



Partendo dal condensatore scarico, applicando un gradino positivo sul terminale di sinistra il pozzetto rimane al potenziale più basso per inerzialità del condensatore. Dunque, inizialmente il terminale di source è quello connesso al condensatore mentre quello di drain è il terminale stimolato dal gradino. Rimuovendo la tensione V_{dd} ancora a gradino, sempre per inerzialità del condensatore, il potenziale più basso è quello del pozzetto a sinistra, per cui source e drain si scambiano. Per i due transitori si dovrebbero scrivere le equazioni in modo diverso.

Overdrive

In un editor di schematico, per fissare le tensioni V_{GS} e V_{DS} viene comunque fissato un source convenzionale di schematico. Se il circuito ha un punto di riposo chiaro, definito, per convenzione si marca come source effettivo il terminale che in quel punto di riposo opera da source. Mentre nel modello per grandi segnali la corrente di canale è pompata da un solo generatore comandato, formalmente le tensioni di controllo sono V_{GS} , V_{DS} , V_{BS} .



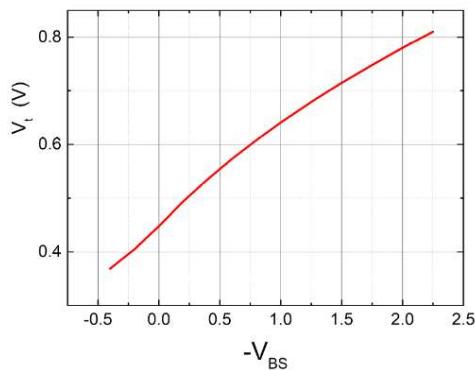
Quando il MOSFET è utilizzato da amplificatore (generatore comandato di corrente) il controllo che interessa è quello primario della V_{GS} . I controlli secondari delle altre tensioni appaiono come effetti parassiti, effetti collaterali. Un MOSFET la cui corrente di canale si controlla principalmente con la V_{GS} è un buon MOSFET.

La componente della V_{GS} che rappresenta la tensione di comando utile, quella che determina la formazione del canale, ovvero l'accensione o lo spegnimento in corrente del MOSFET, è la tensione di overdrive:

$$V_{OV} = V_{GS} - V_t = V_{DS_{sat}}$$

Se la tensione tra gate e source è superiore alla tensione di soglia, ovvero se l'overdrive è positivo, in primissima approssimazione il dispositivo è alla soglia di accensione. La corrente è poi gradualmente controllata dalla sovratensione della V_{GS} rispetto alla soglia.

Effetto body



La tensione di soglia dipende da molte variabili, tra cui la tensione tra body e source V_{BS} . La dipendenza $V_t(V_{BS})$ prende il nome di effetto body. Per $V_{BS} = 0$ la tensione di soglia V_t è definita V_{t_0} o V_{th} .

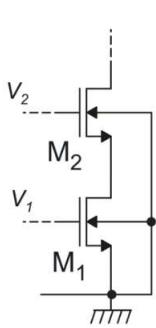
$$V_t(V_{BS}) = V_{t_0} + \gamma(\sqrt{\phi_s - V_{BS}} - \sqrt{\phi_s})$$

$$V_{t_0} = V_t(V_{BS} = 0)$$

Dove γ è il coefficiente dell'effetto body, ϕ_s è il potenziale di superficie

Questo fa sì che la V_{BS} abbia un impatto sull'overdrive, dunque sul controllo della corrente di canale. In un MOSFET discreto source e body sono fisicamente collegati assieme, per cui in tal caso la tensione di soglia è fissa a V_{t_0} . Per un MOSFET di tipo n, dato che il substrato deve essere a potenziale più basso rispetto al source, si riporta $-V_{BS}$. Fissato il terminale di body a 0, se il potenziale di source sale la giunzione può rimanere in inversa, ma si ha una modulazione della tensione di soglia, la quale subisce un accrescimento.

I parametri λ e ϕ_s sono specifici del processo, del dispositivo. Si osserva che per ottenere una tensione di soglia piccola nei circuiti low voltage una possibilità è quella di porre il body a un potenziale leggermente superiore a quello del source. In tal caso scorrerà una piccola corrente in diretta, che rimane piccola nella misura in cui $V_{BS} < V_t$. Analizziamo la seguente struttura cascode:



$$V_{B1} = V_{B2} = 0$$

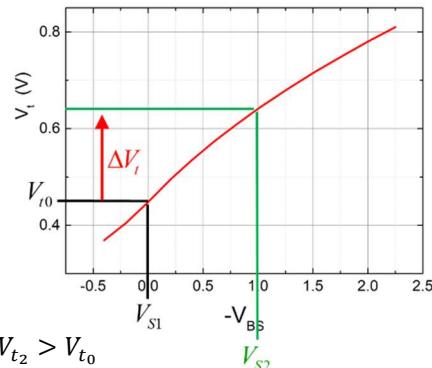
Per quanto riguarda M1:

$$V_{S1} = 0 \rightarrow V_{BS1} = 0 \rightarrow V_{t1} = V_{t0}$$

Per quanto riguarda M2:

$$V_{S2} = V_{D1} > 0 \text{ (hp M2 in sat.)}$$

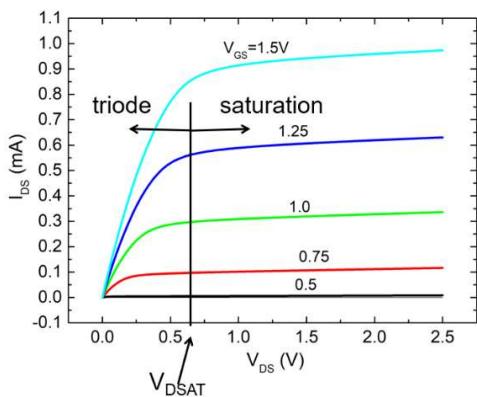
$$-V_{BS2} = -(V_{B2} - V_{S2}) = V_{S2} - V_{B2} = V_{DS1} > 0 \rightarrow V_{t2} > V_{t0}$$



Entrambi i substrati del transistore a source comune e del transistore a gate comune si trovano a ground. Si fa riferimento a un processo CMOS n well, in cui i transistori di tipo n hanno tutti i body in collegamento al substrato, il quale viene posto al potenziale più basso del circuito V_{SS} , in questo caso $V_{SS} = gnd$.

La tensione di source può evolvere dinamicamente per effetto dei segnali, quindi anche la tensione di soglia, dinamicamente, può variare. Se non si considera l'effetto della V_{BS} sull'overdrive si possono commettere seri errori progettuali.

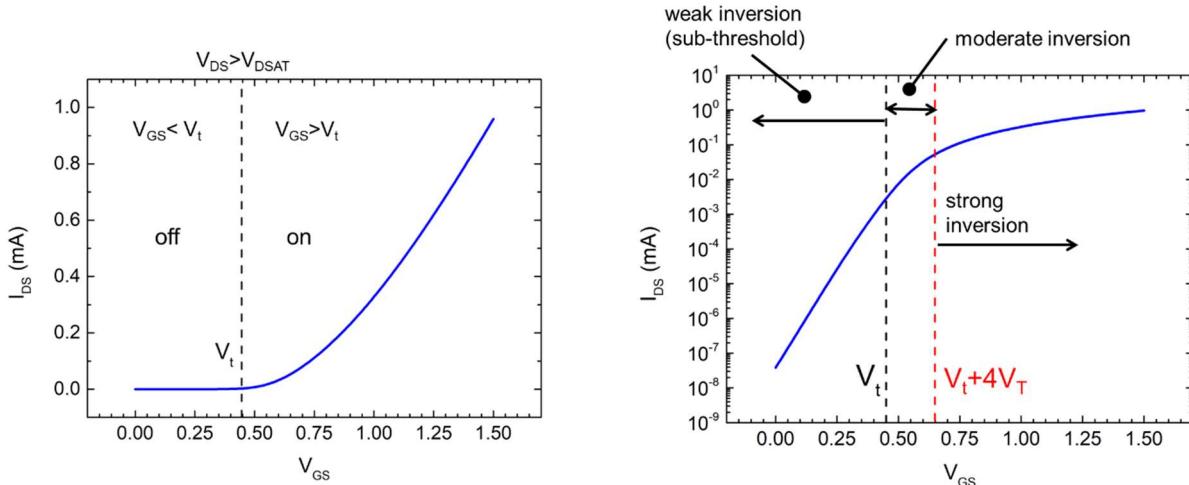
Zone di funzionamento



Si osservano a sinistra le caratteristiche di uscita di un nMOSFET. In zona triodo la corrente I_{DS} dipende molto da V_{DS} . Se vogliamo che il transistore si comporti come un generatore di corrente controllato dalla V_{GS} , la sensibilità della corrente rispetto la V_{DS} non è gradita. In saturazione abbiamo invece si osserva una dipendenza debole e quasi lineare della I_{DS} dalla V_{DS} . Spesso, in ambito analogico, il MOSFET si trova ad operare in regime di saturazione. La $V_{DS_{sat}}$ è una funzione della $V_{GS} - V_t$, e diminuisce al diminuire dell'overdrive

Quando la V_{DS} è piccola, le caratteristiche di uscita sono lineari attorno all'origine, per cui descrivono un comportamento simil resistivo. In alcuni casi, quindi, si può utilizzare il MOSFET come resistore variabile controllato con la V_{GS} .

La funzione che descrive la corrente I_{DS} in funzione della V_{GS} fissata la V_{DS} a un valore $> V_{DS_{sat}}$ è detta transcaratteristica.



A sinistra si ha la transcaratteristica in scala lineare. A sinistra della linea tratteggiata l'overdrive è negativo e si assume che il transistore sia spento. Per overdrive positivo, invece, si considera il transistore acceso. La raffigurazione acceso-spento del MOSFET può essere eccessivamente approssimativa in alcuni design.

Passando al diagramma in scala semilogaritmica, sulla destra, si apprezzano alcuni dettagli. In particolare, si individuano tre zone di lavoro.

- Forte inversione: se la tensione V_{GS} supera la V_t di un certo numero di volte V_T (tensione termica), con $V_T = kT/q$, si individua la zona di forte inversione. A temperatura ambiente $V_T \cong 25mV$, per cui si assume che quando l'overdrive supera di $100 mV$ la tensione di soglia il transistore operi in forte inversione. Si tratta di una soglia convenzionale, che può essere estesa fino ai $200mV$. Considerando che la transcaratteristica vale in regime di saturazione, in forte inversione si applicano le equazioni paraboliche, per cui l'andamento in scala semilogaritmica è logaritmico.
- Debole inversione: al di sotto della tensione di soglia il transistore si trova in zona di debole inversione (sotto-soglia) e, contrariamente a quanto si può dedurre dalla transcaratteristica in scala lineare, la corrente non è nulla. L'andamento lineare in scala semilogaritmica suggerisce che la corrente, in debole inversione, ha una dipendenza esponenziale dalla V_{GS} . Assumiamo che in questa zona valgano le equazioni esponenziali.

- Moderata inversione: in mezzo alle due zone si ha una zona intermedia di raccordo. In corrispondenza della moderata inversione il canale si sta formando gradualmente all'aumentare della V_{GS} ; le equazioni paraboliche non hanno ancora valenza.

La zona operativa del dispositivo, nel complesso, è una combinazione della zona di funzionamento nel diagramma delle caratteristiche di uscita e della zona di funzionamento nel diagramma della transcaratteristica.

	$V_{GS} - V_t \leq 0$	$0 \leq V_{GS} - V_t \leq 4V_T$	$V_{GS} - V_t \geq 4V_T$
$V_{DS} \leq V_{DS_{SAT}}$	Triode – Weak Inversion	Triode – Moderate Inversion	Triode – Strong Inversion
$V_{DS} \geq V_{DS_{SAT}}$	Saturation – Weak Inversion	Saturation – Moderate Inversion	Saturation – Strong Inversion

La tensione $V_{DS_{sat}}$ dipende dalla condizione di inversione del canale. In forte inversione $V_{DS_{sat}} = V_{GS} - V_t$, mentre in moderata e debole inversione si ha $V_{DS_{sat}} = 4V_T$. Per come sono state definite le zone di funzionamento nella transcaratteristica, tra i due regimi della $V_{DS_{sat}}$ c'è continuità.

Equazioni “a mano”

Per $V_{GS} - V_t > 4V_T$, cioè in forte inversione, vale il modello parabolico:

- In zona triodo, per $V_{DS} \leq V_{DS_{sat}}$: $I_{DS} = \beta_n \left(V_{GS} - V_t - \frac{V_{DS}}{2} \right) V_{DS}$
- In zona di saturazione, per $V_{DS} \geq V_{DS_{sat}}$: $I_{DS} = \beta_n \frac{(V_{GS}-V_t)^2}{2} [1 + \lambda(V_{DS} - V_{DS_{sat}})]$

Dove:

- $\beta_n = \mu_n C_{ox} \frac{W_{eff}}{L_{eff}}$
- $V_{DS_{sat}} = V_{GS} - V_t$
- $\lambda^{-1} = k_\lambda L_{eff}$

Spesso per l'equazione in zona di saturazione non si trova la $V_{DS_{sat}}$, ma l'omissione di questo termine comporta una non continuità della funzione tra le due zone. Accettando queste equazioni, pur non mantenendosi continuità della derivata, si mantiene la continuità della funzione.

Il λ tiene conto della dipendenza dalla V_{DS} , dalla modulazione della lunghezza di canale. Si sottolinea la dipendenza dal canale perché è un parametro controllabile. Idealmente vorremmo $\lambda = 0$, con completa reiezione dell'effetto modulante della V_{DS} in saturazione. Dall'espressione del λ si osserva come MOSFET piccoli, con L_{eff} piccola, risentano molto della V_{DS} . Ecco perché in analogico si cerca di stare nell'ordine dei μm piuttosto che dei nm (la dipendenza dalla V_{DS} ha un impatto negativo anche sull'amplificazione).

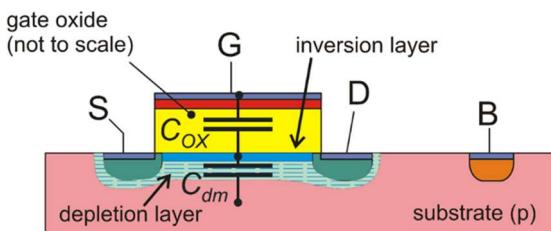
In debole inversione ($V_{GS} - V_t$ ragionevolmente piccola) vale il modello esponenziale. Il comportamento esponenziale del MOSFT in debole inversione, per certi aspetti, è simile a quello del BJT; un MOSFET sottosoglia potrebbe essere utile per costruire riferimenti di tensione, amplificatori esponenziali/logaritmici.

$$I_{DS} = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{-\frac{V_{DS}}{V_T}} \right) [1 + \lambda(V_{DS} - V_{DS_{sat}})]$$

La parte dell'equazione che ci interessa è $I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}}$, perché è quella in cui figura il controllo dell'overdrive. Il resto rappresenta l'effetto nocivo della V_{DS} . Quando $V_{DS} > 100mV$, che in debole inversione significa essere in zona di saturazione, il termine $(1 - e^{-V_{DS}/V_T})$ si può trascurare, mentre il fattore $[1 + \lambda(V_{DS} - V_{DS_{sat}})]$ conta.

Invece, al diminuire della V_{DS} , al di sotto della $V_{DS,sat}$, il termine esponenziale in V_{DS} conta sempre di più e per $V_{DS} = 0$ annulla la corrente.

$$I_{SM} = \mu_n C_{dm} \frac{W_{eff}}{L_{eff}} V_T^2 = \mu_n C_{ox} (m - 1) V_T^2 \frac{W_{eff}}{L_{eff}}$$



Nella I_{SM} compare la capacità dello strato di svuotamento C_{dm} . Il fattore m (sub-threshold slope factor) tiene conto del fatto che fisicamente la carica che si sviluppa alla superficie dipende da una partizione capacitiva tra C_{ox} e C_{dm} .

$$m = 1 + \frac{C_{dm}}{C_{ox}}$$

In effetti nell'overdrive compare la V_t , la quale risente di V_{BS} attraverso l'effetto body. La carica superficiale, in sostanza, è controllata sia dal gate che dal substrato. Nella pratica, l'azione dell'overdrive sul controllo di carica nel canale è meno efficiente: se $m > 1$, il termine esponenziale in V_{GS} con cui si controlla la corrente in debole inversione è più piccolo, a parità di overdrive, rispetto al caso $m = 1$ (l'overdrive è rapportato a mV_T).

Tra le due capacità quella ad offrire più controllo è la più grande, cioè C_{ox} . Il fattore m , che per questa ragione sta nell'ordine di 1.1 – 1.3, è anche la pendenza della $I_{DS}(V_{GS})$ in scala semilogaritmica nella zona di sottosoglia. Fisicamente ciò che succede è che all'aumentare del potenziale positivo sul gate si determina uno svuotamento più profondo (maggiore inversione) e la carica mobile che si forma ha un effetto schermante nei confronti di ulteriore svuotamento (C_{dm} è una capacità di depletion). Si riconosce che:

$$I_{SM} = \beta_n (m - 1) V_T^2$$

Ciò che interessa, a parte il fattore di pendenza sottosoglia, è che la I_{SM} sia controllabile dal progettista tramite il rapporto tra le dimensioni. Nonostante la dipendenza della corrente di canale dalle tensioni sia esponenziale in debole inversione, c'è comunque una proporzionalità con il fattore di progetto W/L , denominato aspect ratio. A parità di tensioni, per qualsiasi zona di funzionamento, la corrente è sempre proporzionale all'aspect ratio.

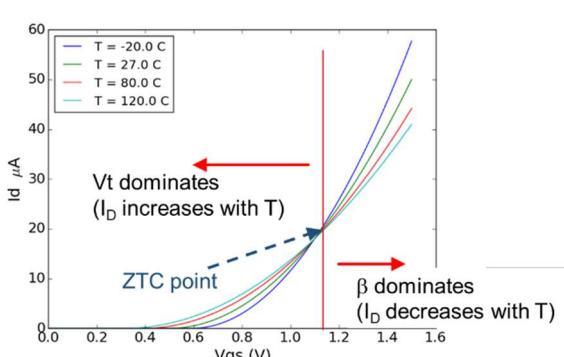
Effetti della temperatura

I parametri fondamentali che regolano la corrente in un MOSFET sono la tensione di soglia V_t e il β_n . Dobbiamo cercare di capire come la temperatura influenza questi parametri.

$$\beta_n(T) = \beta_n(T_0) \left(\frac{T}{T_0} \right)^{-\alpha_\mu} \text{ con } \alpha_\mu = 1.2 - 2.4 \text{ (typical 1.5)}$$

$$V_t(T) = V_t(T_0) - \alpha_{V_t}(T - T_0) \text{ con } 1mV/K \leq \alpha_{V_t} \leq 4 mV/K$$

T_0 è una temperatura di riferimento, fissata solitamente a 300K. Si tratta della temperatura a cui vengono misurati i parametri caratteristici del dispositivo. Il pedice del parametro α_μ fa riferimento al fatto che il fattore a rendere il β dipendente dalla temperatura, più che l'aspect ratio, è la mobilità $\mu(T)$. Polarizzando in saturazione:

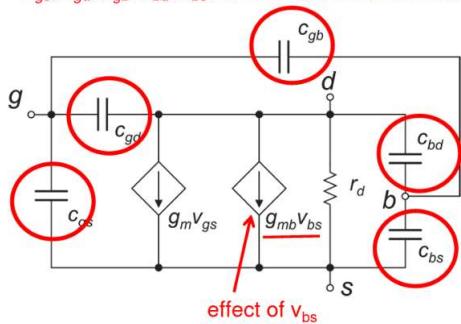


All'aumentare della temperatura β diminuisce, per cui, a parità del resto, la corrente nel MOSFET tende a diminuire. Al tempo stesso la tensione di soglia diminuisce con la temperatura, fenomeno che tenderebbe a far aumentare la corrente nel dispositivo a parità del resto. Si hanno quindi due effetti contrapposti. Quale dei due effetti prevalga dipende dal punto di riposo, dall'overdrive. Se consideriamo un transistore con V_{GS} fissata e $V_{DS} > V_{DS,sat}$ fissata, si individua un punto in cui i due effetti si bilanciano e la corrente non varia al variare della temperatura.

Questo, in realtà, succede solo se α_μ ha un certo valore. Tale punto, detto ZTC (Zero Temperature Coefficient), suddivide in ascissa la transcaratteristica in due zone: una in cui al variare della temperatura prevale l'effetto del β (a destra) e la corrente diminuisce con la temperatura, una in cui al variare della temperatura prevale l'effetto della V_t (a sinistra) e la corrente aumenta con la temperatura. Questo comportamento si può spiegare con il fatto che per overdrive piccoli un cambiamento della tensione di soglia ha un effetto relativo grande. Più è grande l'overdrive, più l'effetto relativo della variazione della tensione di soglia sull'overdrive è piccolo. L'effetto sul β , invece, è sempre lo stesso a prescindere dal punto di lavoro. In ogni caso si individua, se non un punto singolo, una regione in cui la dipendenza dalla temperatura è piccola. Dal punto di vista della progettazione, però, questa zona non è così interessante in quanto costringerebbe ad avere una V_{GS} più o meno fissa. Nei circuiti analogici ci troviamo spesso a lavorare sulla sinistra rispetto al punto ZTC, mentre nei circuiti digitali a destra.

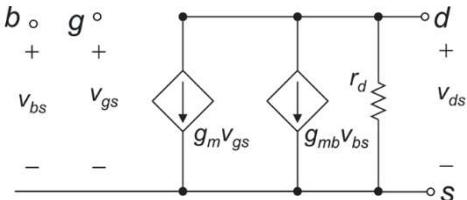
Modello di piccolo segnale MOSFET

$c_{gs}, c_{gd}, c_{gb}, c_{bd}, c_{bs}$: small signal capacitances



La capacità tra drain e source, a meno che non si abbiano transistori nanometrici, è trascurabile. Laddove ci sono giunzioni in inversa, invece, non possiamo trascurare la capacità. La variazione della corrente di drain dipende anche dalla variazione della V_{BS} : se varia il potenziale del body cambia la tensione di soglia, per cui non si può trascurare il generatore comandato da v_{BS} .

Immaginiamo di essere a frequenze sufficientemente piccole da poter trascurare le capacità parassite:



Avremmo potuto inserire un generatore comandato anche per la v_{DS} , ma i generatori comandati dalla stessa tensione che hanno ai capi equivalenti a resistenze/conduittanze.

Ai grandi segnali, in generale, $I_D = I_D(V_{GS}, V_{BS}, V_{DS})$

Ai piccoli segnali si fa una linearizzazione al prim'ordine: $i_d = g_m v_{gs} + g_{mb} v_{bs} + g_d v_{ds}$

$$g_m = \left. \frac{i_d}{v_{gs}} \right|_{v_{ds}, v_{bs}=0} = \left(\frac{\partial I_D}{\partial V_{GS}} \right)_{V_{DS}=\text{const}, V_{BS}=\text{con}} \quad g_{mb} = \left. \frac{i_d}{v_{bs}} \right|_{v_{ds}, v_{gs}=0} = \left(\frac{\partial I_D}{\partial V_{BS}} \right)_{V_{DS}=\text{con}, V_{GS}=\text{cons}}$$

$$\frac{1}{r_d} = g_d = \left. \frac{i_d}{v_{ds}} \right|_{v_{gs}, v_{bs}=0} = \left(\frac{\partial I_D}{\partial V_{DS}} \right)_{V_{GS}=\text{const}, V_{BS}=\text{cons}}$$

I parametri dinamici potrebbero essere ricavati graficamente; cerchiamo espressioni semplici in cui sia chiara la dipendenza dal punto di riposo. Quando poi applicheremo grandi segnali, questi saranno abbastanza lenti, per cui potranno essere considerati come successioni di punti di riposo e piccole variazioni sullo stesso circuito di piccolo segnale DC. Durante queste transizioni il transistore può passare dalla zona triodo a quella di saturazione.

Transconduttanza di substrato g_{mb}

Partiamo dall'analisi del g_{mb} , cercando di scrivere la I_D come funzione solo di due tensioni. In particolare, si ingloba la dipendenza della V_{BS} all'interno della tensione di soglia V_t :

$$I_D(V_{GS}, V_{BS}, V_{DS}) \cong I_D[(V_{GS} - V_t), V_{DS}]$$

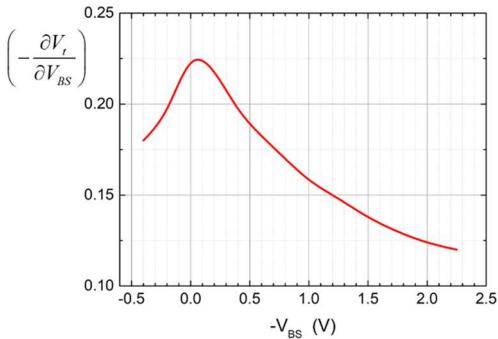
Richiamiamo la definizione del g_m

$$g_m = \left(\frac{\partial I_D}{\partial V_{GS}} \right)_{V_{DS}=\text{cons}, V_{BS}=\text{cons}} = \left(\frac{\partial I_D}{\partial (V_{GS} - V_t)} \right)_{V_{DS}} \left(\frac{\partial (V_{GS} - V_t)}{\partial V_{GS}} \right)^{-1}_{V_{BS}} = \left(\frac{\partial I_D}{\partial (V_{GS} - V_t)} \right)_{V_{DS}}$$

$$g_{mb} = \left(\frac{\partial I_D}{\partial V_{BS}} \right)_{V_{GS}=\text{const}, V_{DS}=\text{con}} = \left(\frac{\partial I_D}{\partial (V_{GS} - V_t)} \right)_{V_{DS}} \left(\frac{\partial (V_{GS} - V_t)}{\partial V_{BS}} \right)_{V_{GS}} = g_m \left(-\frac{\partial V_t}{\partial V_{BS}} \right)_{V_{DS}}$$

Si ottiene che il g_{mb} è una versione scalata del g_m . Il fattore di scala è la derivata $\partial V_t / \partial (-V_{BS})$

Example from simulation



Si può dimostrare che $g_{mb} = g_m(m - 1) \cong 0.2g_m$

Questa espressione è utile anche da un punto di vista progettuale; a spanne, moltiplicando la variazione della V_{BS} per 0.2 si ottiene la variazione della tensione di soglia. Conoscendo g_m in ogni punto di riposo, il g_{mb} si ottiene rapidamente come versione scalata del g_m .

Di solito nel caso di un amplificatore si sottopone il gate alle variazioni v_{gs} e si ottiene una variazione di corrente in uscita, che, se convogliata in una resistenza, produce una tensione amplificata. Esistono amplificatori “bulk driven”, pilotati dal bulk. La variazione della corrente, in generale, è data dalla transconduttanza moltiplicata per la variazione della tensione di comando corrispondente. Gli amplificatori bulk driven, essendo $g_{mb} < g_m$, sono meno efficienti. Non c’è differenza topologica, funzionale, tra i controlli v_{gs} e v_{bs} , ma è come se il generatore controllato da v_{gs} avesse una maggiore efficacia.

Espressione dei parametri differenziali

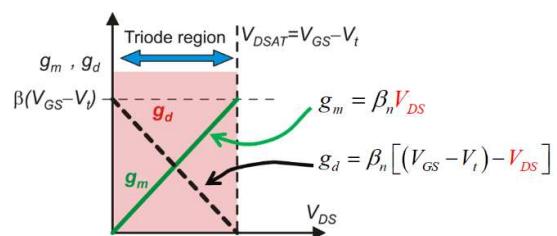
In forte inversione

Vogliamo capire come cambiano i parametri dinamici rispetto al punto di riposo. In forte inversione:

Per $V_{DS} < V_{DS_{sat}}$ (zona triodo): $I_{DS} = \beta_n \left(V_{GS} - V_t - \frac{V_{DS}}{2} \right) V_{DS}$

$$g_m = \left(\frac{\partial I_{D_{tri}}}{\partial V_{GS}} \right)_{V_{DS}, V_{BS}} = \beta_n V_{DS}$$

$$\frac{1}{r_d} = g_d = \left(\frac{\partial I_{D_{tri}}}{\partial V_{DS}} \right)_{V_{GS}, V_{BS}} = \beta_n [(V_{GS} - V_t) - V_{DS}]$$



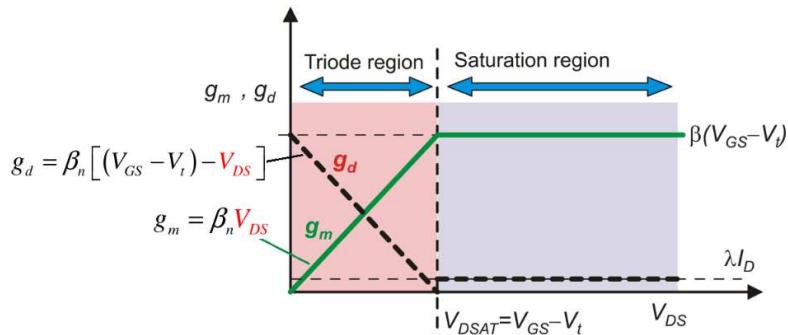
In realtà la tensione di soglia dipenderebbe anche da V_{DS} (drain induced barrier lowering). In zona triodo, però, la V_{DS} è piccola, il DIBL ha poco effetto. In saturazione, invece, si ingloba l’effetto nel parametro λ . Studiare i parametri di piccolo segnale in zona triodo è importante non tanto perché andremo a lavorare in questa zona, ma perché il dispositivo potrebbe finirci per grandi segnali.

Sempre in forte inversione, ma per $V_{DS} > V_{DS_{sat}}$ (saturazione): $I_{D_{sat}} = \beta_n \frac{(V_{GS} - V_t)^2}{2} [1 + \lambda(V_{DS} - V_{DS_{sat}})]$

$$g_m = \left(\frac{\partial I_{D_{sat}}}{\partial V_{GS}} \right)_{V_{DS}, V_{BS}} \stackrel{*}{\cong} \beta_n (V_{GS} - V_t) [1 + \lambda(V_{DS} - V_{DS_{sat}})] \cong \beta_n (V_{GS} - V_t) \quad \text{si trascura che } V_{DS_{sat}}(V_{GS})$$

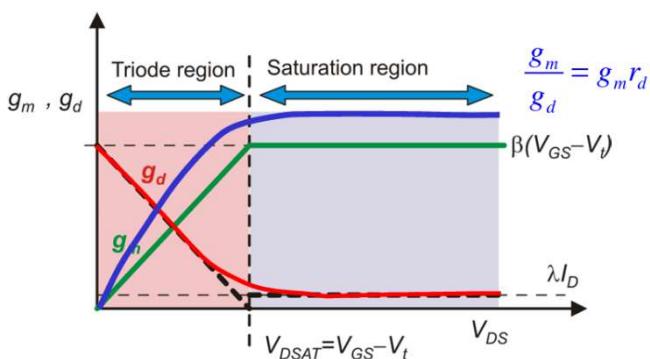
$$\frac{1}{r_d} = g_d = \left(\frac{\partial I_{D_{sat}}}{\partial V_{DS}} \right)_{V_{GS}, V_{BS}} = \lambda \frac{\beta_n}{2} (V_{GS} - V_t)^2 \cong \lambda I_{D_{sat}}$$

Quando si fa l'operazione di derivata per il calcolo della g_d non può essere trascurata la dipendenza dalla V_{DS} ; è proprio quella che rende il parametro non nullo. Ricapitolando, in forte inversione e per un overdrive costante:



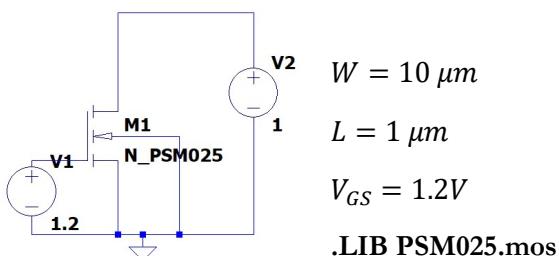
Al diminuire della V_{DS} , cioè verso la progressiva entrata in triodo, g_m diminuisce, e ciò significa che diventa più scarso il controllo della corrente tramite V_{GS} . Al contempo g_d , che rappresenta il controllo nocivo da parte della V_{DS} , da un valore piccolo in saturazione aumenta verso la zona triodo. Il g_m al limite della saturazione ha lo stesso valore di g_d per $V_{DS} = 0$. Dunque, da un punto di vista di controllo, per un amplificatore a sinistra le cose vanno piuttosto male.

Si nota che per g_d appare una discontinuità della derivata. Questo in realtà deriva da un'eccessiva approssimazione delle equazioni nella zona triodo affacciata alla saturazione. Le curve reali sono, come ci si può aspettare, più sinuose. Ad ogni modo, fissato un certo overdrive in zona di saturazione i parametri differenziali sono buoni e costanti con la V_{DS}



Si evidenzia un parametro importante del MOSFET che valuta la bontà del controllo: il rapporto tra g_m e g_d , equivalente al prodotto $g_m r_d$. Si tratta del guadagno intrinseco del MOSFET, ed è tanto migliore quanto più grande è. Il guadagno intrinseco del MOSFET è massimo in zona di saturazione. Una volta che il dispositivo è saturo, il prodotto $g_m r_d$ non migliora ulteriormente all'aumentare della V_{DS} .

Esercitazione personale: g_m e g_d su LTSpice



Simulazione $g_d(V_{DS})$: sweep DC primario sulle sorgente $V2$ (V_{DS}). Si traccia $\frac{d(I_D(M1))}{dV_2} = g_d$

Comando: **.DC V2 0 3 0.001**

Track: **d(Id(M1))**

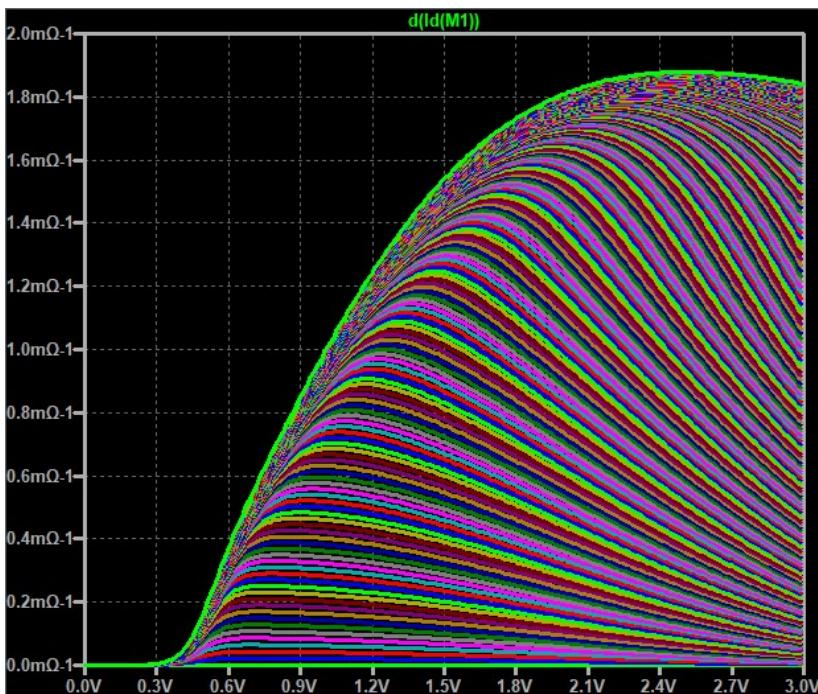
Simulazione $g_m(V_{DS})$: si tratterebbe di tracciare il valore della derivata $\partial I_D / \partial V_{GS}$ in $V_{GS} = 1.2V$ e in funzione della V_{DS} . Non so se la simulazione di $g_m(V_{DS})$ sia integralmente possibile da ottenere su LTSpice.

La variabile indipendente è V_{DS} , per cui sembrerebbe essere necessario uno sweep su V_{DS} . Il problema è che, fissato un certo valore di V_{DS} , c'è bisogno di calcolare la derivata di $\partial I_D / \partial V_{GS}$, operazione che Spice non riesce a fare “da sé in segreto”, ma che bensì necessita di un altro sweep sulla V_{GS} (è una mia ipotesi). Già qui sorgerebbe un problema: se lasciassi lo sweep primario su V_{DS} e quello secondario su V_{GS} , non avrei modo di vedere la $\partial I_D / \partial V_{GS}$, perché la variabile indipendente (quella rispetto a cui si possono tracciare le funzioni) sarebbe la V_{DS} . Quindi in quel caso vedrei soltanto le caratteristiche di uscita per diverse V_{GS} . L'unica cosa che si può fare è quindi fare uno sweep primario sulla V_{GS} e uno sweep secondario sulla V_{DS} . In questo modo si possono tracciare le varie curve $\partial I_D / \partial V_{GS}$ parametrizzate per diverse V_{DS} . Rimangono comunque due problemi:

1. Non ho bisogno di $\partial I_D / \partial V_{GS}$ in tutti i valori di V_{GS} , come invece appare in questo caso, ma solo per $V_{GS} = 1.2V$
2. La V_{DS} non è più la variabile indipendente primaria sulle ascisse, è diventata un parametro secondario

Comando: **.DC V1 0 3 0.01 V2 0 3 0.01**

Track: **d(Id(M1))**



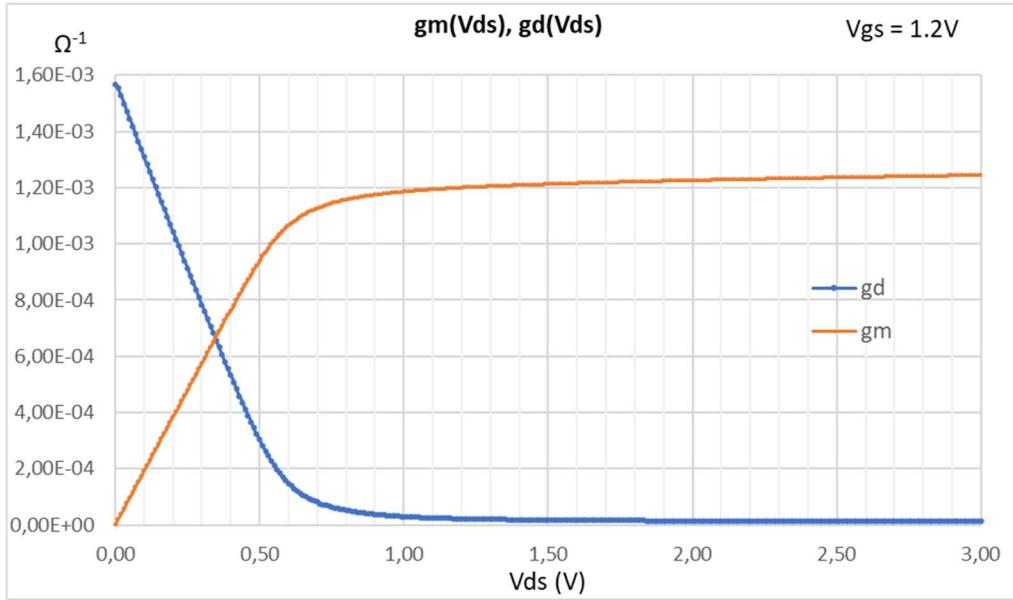
Si trattarebbe ora di posizionarsi in $V_{GS} = 1.2V$ e, curva per curva, riportare sulle ordinate il valore della deriva, sulle ascisse la V_{DS} corrispondente. Questa non è tanto una soluzione, ma un escamotage; in è possibile risolvere il problema esportando il file di testo dei campioni. La struttura del file è tabellare e così composta:

VDS=0 (V)	
Vgs(V)	d(Id(M1))
0	...
0.01	...
...	...
VDS=0.01 (V)	
Vgs(V)	d(Id(M1))
0	...
0.01	...
...	...
	...

Assicurandosi di trasformare i punti in virgole, si può copiare il file su Excel e filtrare per la voce $V_{GS} = 1.2V$. Basterà aggiungere una colonna delle V_{DS} a mano:

A	B	C
1	v1	D(Id(M1))
123	1,2	0
425	1,2	1,91269E-05
727	1,2	3,82575E-05
1029	1,2	5,73915E-05
1331	1,2	7,65289E-05
1633	1,2	9,56694E-05
1935	1,2	0,000114812
2237

Arrivati a questo punto si crea un grafico e si assegna la terza colonna alle ascisse, la seconda alle ordinate. Per completezza riporto su Excel anche il grafico della $g_m(V_{DS})$, anche se sarebbe ricavabile direttamente da LTSpice. Ultima cosa: sweep DC con step che non possono essere troppo piccoli fanno sì che sia le derivate (rapporti incrementali) rispetto la V_{GS} che l'interpolazione rispetto alla V_{DS} non siano precissime.



A meno di errori non sembra che, in realtà, il g_m in saturazione sia pari al g_d per $V_{DS} = 0V$. Inoltre, si vede che g_m continua a dipendere da V_{DS} , se pur poco, anche in saturazione.

Espressione “progettuale” del g_m in forte inversione e saturazione

In forte inversione, per $V_{DS} > V_{DS_{sat}}$ (saturazione):

$$I_{DS} = \beta_n \frac{(V_{GS} - V_t)^2}{2} [1 + \lambda(V_{DS} - V_{DS_{sat}})] \cong \beta_n \frac{(V_{GS} - V_t)^2}{2}$$

Da questa espressione, che vale solo in forte inversione e in saturazione, si ricava che:

$$(V_{GS} - V_t) = \sqrt{\frac{2I_D}{\beta_n}} \rightarrow g_m = \beta_n (V_{GS} - V_t) = \sqrt{2I_D \beta_n}$$

Se in un MOSFET si impone la corrente (ad esempio per un MOSFET montato a diodo) e la si fa aumentare linearmente, il g_m aumenta come la radice quadrata della corrente. Del resto, la corrente è proporzionale al quadrato dell'overdrive. Si può arrivare a un'espressione in cui sia assente il β_n , un'espressione indipendente dalla tecnologia. Dall'espressione della corrente in forte inversione e saturazione si ricava:

$$\beta_n = \frac{2I_{DS}}{(V_{GS} - V_t)^2} \rightarrow g_m = \beta_n (V_{GS} - V_t) = \frac{2I_D}{(V_{GS} - V_t)}$$

Si tratta di una formula molto utile in fase di progetto; permette di ricavare il g_m a prescindere dalle dimensioni del dispositivo, a partire dai soli corrente e overdrive.

Tale formula permette di rendere il progetto abbastanza indipendente dalla tecnologia. A parità di corrente, si può aumentare il g_m diminuendo l'overdrive. Tuttavia, la formula è valida esclusivamente in forte inversione, per cui l'overdrive deve comunque essere almeno $4V_T$. La cascata di passi progettuali potrebbe essere la seguente:

1. Specifica di progetto I_{DS}
2. Si ricava l'overdrive necessario per ottenere il g_m desiderato
3. Si ricava il β_n con le altre formule
4. Si ricavano le dimensioni del MOSFET che permettono di rispettare le specifiche

In debole inversione

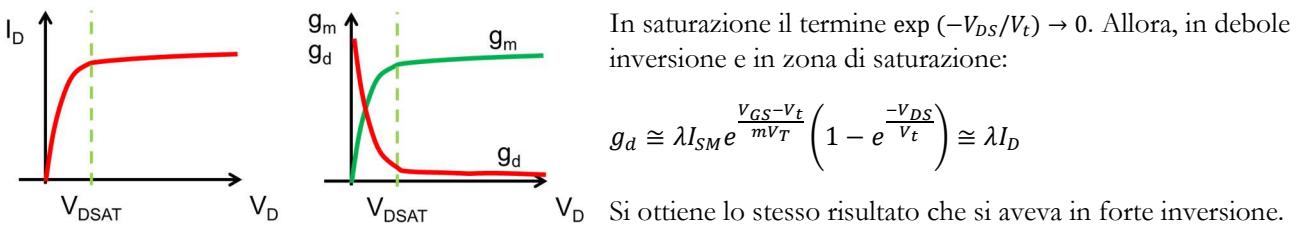
La corrente in debole inversione $I_{DS} = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{\frac{-V_{DS}}{V_t}} \right) [1 + \lambda(V_{DS} - V_{DS_{sat}})]$

$$g_m = \left(\frac{\partial I_D}{\partial V_{GS}} \right)_{V_{DS}, V_{BS}} = \frac{1}{mV_T} I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{\frac{-V_{DS}}{V_t}} \right) [1 + \lambda(V_{DS} - V_{DS_{sat}})] = \frac{I_D}{mV_T}$$

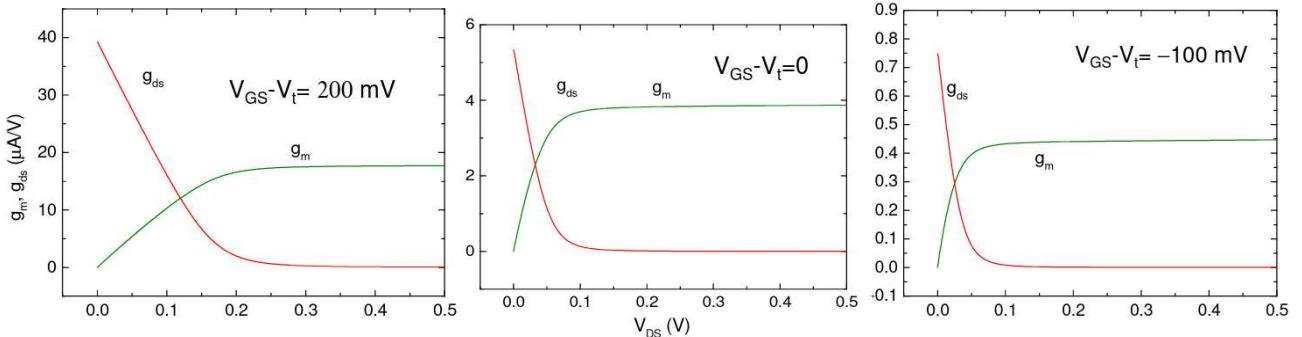
La tensione $V_{DS_{sat}}$ adesso non dipende più dalla V_{GS} , si assesta attorno ai 100mV e lì rimane. L'espressione del g_m ci dice che in debole inversione il MOSFET mima il bipolare.

$$\frac{1}{r_d} = g_d = \left(\frac{\partial I_D}{\partial V_{DS}} \right)_{V_{GS}, V_{BS}} = \frac{I_{SM}}{V_T} e^{\frac{V_{GS}-V_t}{mV_T}} e^{\frac{-V_{DS}}{V_t}} [1 + \lambda(V_{DS} - V_{DS_{sat}})] + \lambda I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{\frac{-V_{DS}}{V_t}} \right)$$

A livello qualitativo, g_m è una versione scalata della $I_D(V_{DS})$, mentre g_d è la derivata della I_D fatta rispetto alla V_{DS} . Per cui, qualitativamente:



Sia in forte inversione che in debole inversione gli andamenti qualitativi del g_m e g_d sono gli stessi. Di seguito sono riportate delle simulazioni di $g_m(V_{DS})$ e $g_d(V_{DS})$ per diversi overdrive:



Tra i tre casi cambia la $V_{DS_{sat}}$, ma non l'andamento qualitativo delle curve. Il transistore su cui sono state condotte le simulazioni è lo stesso. Cambiando l'overdrive cambia la corrente in saturazione. Diminuendo l'overdrive diminuisce la corrente di saturazione, per cui diminuisce anche il g_m , il quale è proporzionale alla corrente in debole inversione e proporzionale alla radice della corrente in forte inversione. Per mantenere lo stesso g_m in saturazione, al diminuire dell'overdrive occorre impiegare un aspect ratio maggiore.

Modello unificato della transconduttanza in zona di saturazione

Cerchiamo adesso un'espressione del g_m in saturazione che raccordi valida sia in debole inversione che in forte inversione. Per la forte inversione si considera la formula indipendente dalle dimensioni del MOSFET:

$$\text{Strong inversion: } g_m = \frac{2I_D}{(V_{GS}-V_t)} = I_D / \frac{(V_{GS}-V_t)}{2}$$

Weak inversion: $g_m = I_D/mV_T$

BJT: $g_m = I_C/V_T$

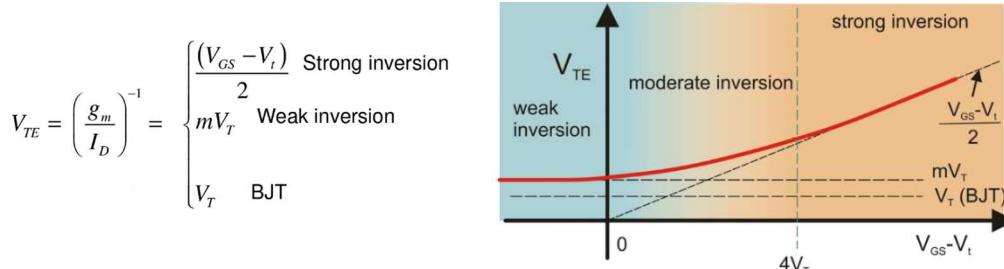
Le tre formule hanno tutte la stessa struttura. Possiamo allora condensarle in un'unica formula, indipendente sia dal tipo di dispositivo che dagli aspetti tecnologici:

$$g_m = \frac{I_D}{V_{TE}}$$

dove con V_{TE} individuiamo la tensione termica efficace (o equivalente). Si tratta di un parametro inventato che unifica le espressioni del g_m in un'unica formula e porta su di sé la differenziazione dei tre casi. In generale, anche per altri dispositivi si ottiene che:

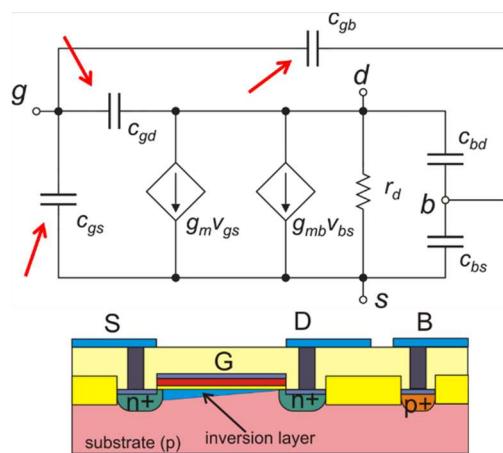
$$V_{TE} = \left(\frac{g_m}{I_D} \right)^{-1}$$

Lo scriviamo così anziché $V_{TE} = I_D/g_m$ perché il parametro g_m/I_D è un parametro noto detto efficienza di transconduttanza (“gmoverid” in Spice) Per i dispositivi con cui abbiamo a che fare:



L'importanza del parametro g_m/I_D deriva dal fatto che g_m elevato è sinonimo di amplificatore più veloce e meno rumoroso. Avere un g_m elevato, però, costa corrente. Il parametro g_m/I_D valuta quindi una questione di economia: a parità di corrente, un dispositivo con g_m/I_D maggiore (o similmente V_{TE} minore) ha un g_m più grande, per cui è migliore. Il dispositivo con la transconduttanza più efficiente da questo punto di vista è il BJT, che a parità di corrente è meno rumoroso e più veloce.

Modello delle capacità del MOSFET



Quelle riportate sono capacità differenziali (dQ/dV) per unità di lunghezza. Le più importanti sono quelle che si vedono dal gate, che contribuiscono alla capacità di ingresso del dispositivo. Queste si dividono, singolarmente, in una componente estrinseca e una componente intrinseca.

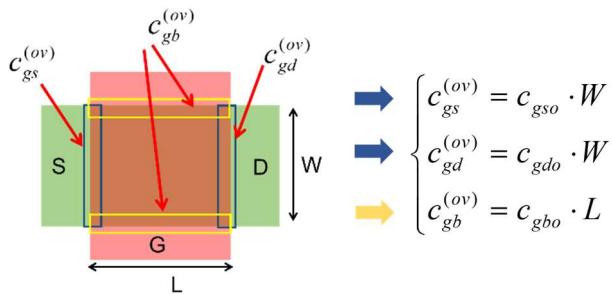
estrinsic cap.

$$\begin{aligned} c_{gs} &= c_{gs}^{(ov)} + c_{gs}^{(i)} \\ c_{gd} &= c_{gd}^{(ov)} + c_{gd}^{(i)} \\ c_{gb} &= c_{gb}^{(ov)} + c_{gb}^{(i)} \end{aligned}$$

intrinsic cap..

La componente intrinseca è quella legata alla capacità del potenziale di gate di indurre carica, imprimere lo svuotamento all'inversione, spostare la carica mobile all'interno del canale. Anche qualora si trattasse un dispositivo che non presenta effetto transistor, si avrebbero comunque capacità estrinseche per effetto elettrostatico, le quali si sommano a quelle intrinseche grazie alla linearizzazione del circuito di piccolo segnale.

Capacità estrinseche: overlap

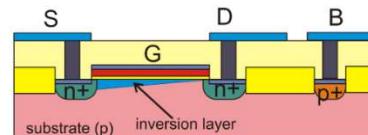


Le capacità estrinseche sono dovute alle linee di campo elettrico che ci sono tra i conduttori di drain, source, body e il gate a causa di sovrapposizioni geometriche (ov per “overlap”). Si tratta di capacità differenziali per unità di lunghezza.

La capacità di sovrapposizione tra il gate e i pozzi è proporzionale a W ; in una struttura simmetrica si ha $c_{gso} = c_{gdo}$. Per la capacità di sovrapposizione tra il gate e il body bisogna escludere la zona del canale; lì, infatti, si forma il foglio di carica mobile per effetto transistore, il quale scherma il campo verticale impedendo ulteriore estrazione di carica. Tuttavia, il gate può indurre carica anche a canale acceso ai bordi, dove è assente la carica mobile. Di tale effetto si tiene conto proprio con la capacità estrinseca tra gate e body, che risulta proporzionale a L tramite c_{gbo} . L'overlap tra il gate e i pozzi è significativo, origina capacità estrinseche che talvolta non possono essere trascurate. L'overlap tra il gate e il substrato si instaura laddove il gate si trova in prossimità del FOX, oltre cui iniziano le capacità tra le interconnessioni. Spesso c_{gbo} si considera nulla. Le capacità estrinseche sono lineari, non cambiano con la tensione.

Capacità intrinseche: modello di Meyer semplificato

	Off ($V_{GS} \ll V_t$)	Triode	Saturation
$C_{gs}^{(i)}$	0	$\frac{1}{2}C_{ox}WL$	$\frac{2}{3}C_{ox}WL$
$C_{gd}^{(i)}$	0	$\frac{1}{2}C_{ox}WL$	0
$C_{gb}^{(i)}$	$\left(\frac{1}{C_{ox}WL} + \frac{1}{C_{dm}}\right)^{-1}$ *	0	0

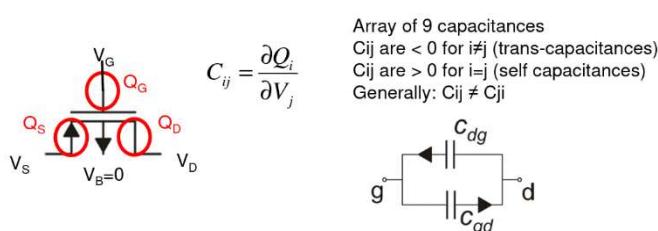


*Series of the oxide and depletion layer capacitances. Can be approximated with only the oxide cap $C_{ox}WL$

Quando il MOSFET è spento il gate vede il substrato senza schermaggio, per cui la capacità è la serie tra quella di ossido e quella di svuotamento. Al tempo stesso source e drain non possono avere controllo sul canale, non si vedono capacitivamente dal punto di vista della carica mobile, per cui le capacità intrinseche che coinvolgono drain e source sono nulle. In triodo si ha una continuità di carica mobile tra il source e il drain. Il potenziale dello strato di inversione è sotto controllo sia da parte del drain che da parte del source. Si può pensare che, se il dispositivo è simmetrico, la capacità totale $C_{ox}WL$ sia spartita equamente tra source e drain. Questo, in realtà, sarebbe rigorosamente vero solo per $V_{ds} = 0$. Infine, in saturazione avviene lo strozzamento: il drain non è più in grado di controllare efficacemente il potenziale dello strato di inversione e si riduce anche il controllo da parte del source. Quindi, la capacità relativa al drain si considera essere nulla, mentre quella relativa al source si considera essere due parti su tre di quella totale. Il modello di Meyer semplificato non garantisce la conservazione della carica e fa sì che le capacità intrinseche siano reciproche tra loro; questo comporta notevoli errori nel descrivere MOSFET come interruttori.

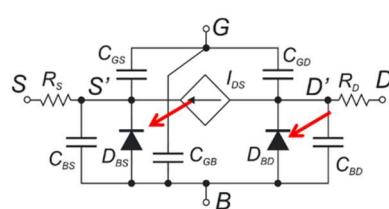
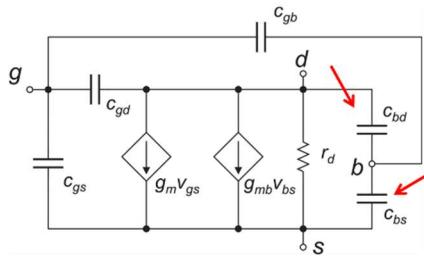
Modello di Dutton e Ward

Il simulatore si basa su un modello radicalmente diverso, più complesso, il modello di Dutton e Ward. Tale modellazione rispetta il principio di conservazione della carica e fa sì che le simulazioni convergano.



In questo caso le capacità intrinseche sono ricavate, fissando il body a ground, come derivata della carica a un terminale i -esimo Q_i rispetto alla tensione di un terminale j -esimo V_j . La matrice di capacità che si ottiene non è simmetrica: la capacità tra un terminale e un altro è diversa nei due sensi.

Capacità di giunzione



$$c_{bd} = \frac{C_J A_D}{\left(1 + \frac{V_{SB}}{V_0}\right)^{m_j}} + \frac{C_{JSW} P_D}{\left(1 + \frac{V_{SB}}{V_0}\right)^{m_{jsw}}}$$

Queste capacità sono relative alle giunzioni al drain, source e body. In quanto capacità di giunzione dipendono dalla tensione di polarizzazione. Nella formula, V_0 è il potenziale di build in, C_J è la componente di capacità per unità di area della giunzione, C_{JSW} è la componente di capacità per unità di perimetro della giunzione (“junction side wall”).

Altre non idealità del comportamento del MOSFET

Le equazioni paraboliche del MOSFET sono in realtà delle approssimazioni, sempre più pesanti per i MOSFET in tecnologie più avanzate. Siamo affezionati ai nostri modelli perché ci permettono di dimensionare i circuiti e lasciare i piccoli aggiustamenti al simulatore. Le equazioni parabole in forte inversione sono quelle più idonee, sono semplici. Rispetto a quanto fissato, però, in realtà, ci sono alcune deviazioni:

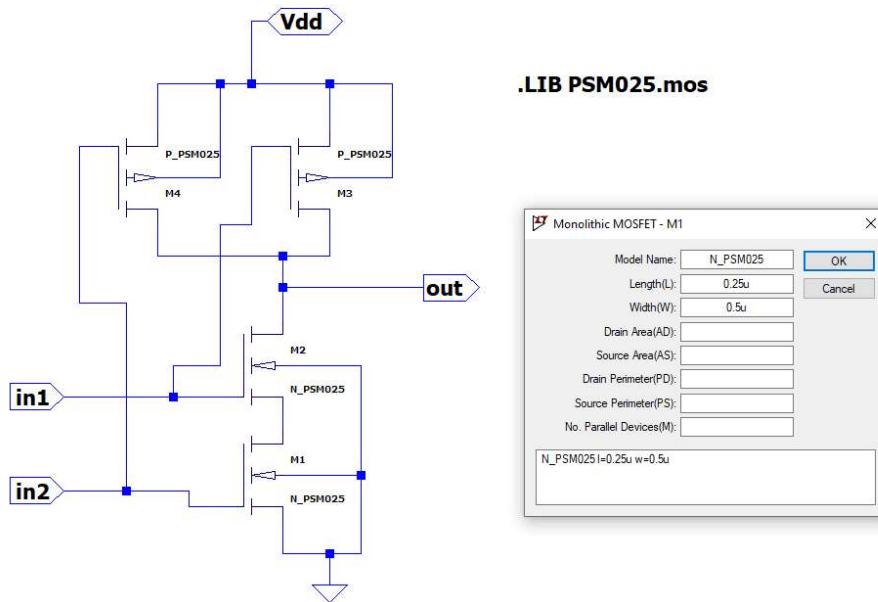
1. Gate-bias dependent mobility: la mobilità, oltre che con la temperatura, varia anche con il punto di riposo del transistore. All'aumentare della V_{GS} il campo verticale deflette sempre più portatori vicini verso l'interfaccia, regione in cui la concentrazione di difetti reticolari è maggiore e si ha più scattering. La conseguenza è che la mobilità si riduce con la V_{GS} , dunque la corrente cresce meno rispetto a quanto previsto dalla legge parabolica.
2. Carrier velocity saturation: al crescere della V_{DS} si determina un aumento di campo elettrico longitudinale, lungo il canale, il quale determina un aumento degli urti e quindi una diminuzione di mobilità. Più è corto il canale, a parità di tensione, maggiore il campo elettrico. L'effetto è quindi più importante per dispositivi molto piccoli in regime di saturazione della velocità. In forte inversione si osserva una dipendenza $I_D(V_{GS})$ lineare anziché quadratica.
3. Gate current: nei processi più spinti, anche se vengono utilizzati dielettrici ad alta costante dielettrica, gli spessori in gioco rendono difficile evitare correnti di riflusso attraverso il gate (o per effetto tunneling, o per i portatori caldi). Per contrastare la I_G si cerca di ridurre l'area di gate.
4. Effective dimensions: le dimensioni geometriche che figurano nelle equazioni discusse non sono quelle efficaci.
5. Threshold voltage: la tensione di soglia non è una costante. Oltre che cambiare per l'effetto body, risente delle dimensioni fisiche del transistore. Nel manuale di processo venivano fornite tensioni di soglia diverse per transistori con dimensioni diverse. Ciò è dovuto ai fenomeni di canale corto (L piccolo) e di canale stretto (W piccolo). Ci sono casi in cui la progettazione a minime lunghezze è d'obbligo (opamp per i pixel, PGB elevati es. 100GHz).

In analogico, di solito, questi problemi sono un po' meno importanti.

Esercitazione: progettazione di una NAND2

Una volta creata la cartella NAND2 in LTSpice: File -> new schematic -> save as -> “NAND2.asc”

Si inseriscono i componenti con tasto f2. Si può scegliere se prelevare i componenti dalla libreria standard di LTSpice, oppure dalla cartella in cui è salvato il progetto. Con ESC si termina la finestra per la scelta dei componenti. Con tasto destro sul dispositivo si impostano i parametri dimensionali; sceglieremo la L e la W minime. La molteplicità M è per default pari a 1. In caso contrario indica il numero di dispositivi delle stesse caratteristiche posti in parallelo, raccolti sotto un unico simbolo. Con tasto f6 si può duplicare un oggetto.



Il source è quello più vicino al gate. Per i pMOSFET il source va posto a potenziale più alto, verso la V_{dd} . Possiamo ruotare e specchiare il dispositivo con le combinazioni di comandi cntrl+R, cntrl+E. Tali operazioni possono esser fatte soltanto quando l'oggetto è in movimento (f7) o quando è inserito nello schematico. Con il tasto f3 si inseriscono i fili (wire).

GND è un nodo globale. Per V_{dd} si inserisce un terminale bidirezionale con una label net (f4). Il prossimo passo è quello di specificare il modello dei dispositivi: edit -> spice directive -> “.LIB PSM025”. In questo modo il simulatore cercherà i modelli dei dispositivi nel file “PSM025.mos”.

In previsione della fase di LVS si salva la netlist estratta dallo schematico, passaggio che in realtà andrebbe fatto dopo le simulazioni e il layout: view -> spice netlist -> edit as independent netlist. A questo punto la netlist diventa editabile. Per abilitare il confronto con la netlist di layout si aggiunge la formula “.SCALE METER”, salvando la modifica. Si tratta di un'espressione che viene considerata solo in fase LVS ed ha l'effetto di conservare le dimensioni dei dispositivi dal layout editor. La prima riga viene sempre considerata come commento da LTSpice.

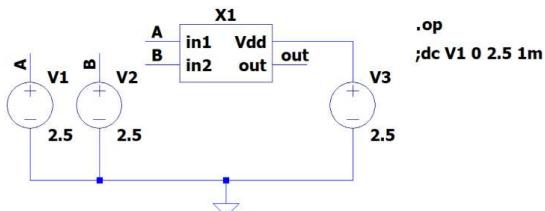
```
* C:\Users\gabri\Desktop\Ese_24_marzo_CAD\Ese_24_marzo_CAD\LTspice\nand2\nand2.asc
.SCALE METER
M1 N001 in2 0 N_PSM025 l=0.25u w=0.5u
M2 out in1 N001 0 N_PSM025 l=0.25u w=0.5u
M3 out in1 Vdd Vdd P_PSM025 l=0.25u w=1.5u
M4 out in2 Vdd Vdd P_PSM025 l=0.25u w=1.5u
.model NMOS NMOS
.model PMOS PMOS
.lib C:\Users\gabri\AppData\Local\LTspice\lib\cmp\standard.mos
.LIB PSM025.mos
.backanno
.end
```

È bene non avviare simulazioni su questa cella. Inserire i generatori altererebbe la netlist e la renderebbe non più confrontabile con quella estratta dal layout. Possiamo però creare un workbench, un file a livello di astrazione maggiore in cui si istanzia la NAND2 e collociamo i generatori.

Per creare un simbolo della NAND2: hierarchy -> open symbol -> crea un simbolo automatico (.asy).

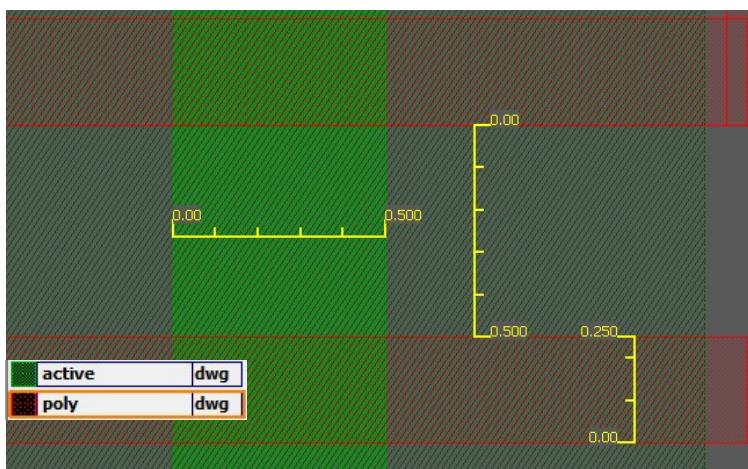
Alternativamente si può creare un simbolo con semplici tool grafici. Per realizzare il workbench si crea un nuovo schematico e con f2 si istanzia la porta NAND2. Per simulare la porta si inseriscono i generatori dalla libreria std LTspice.

Procediamo calcolando il punto di riposo. Prima ci assicuriamo che vengano salvate le tensioni e le correnti interne al sottocircuito: simulate -> control panel -> save defaults -> save subcircuit voltages/ currents. Si passa poi alla simulazione: simulate -> edit simulation command -> ".op"



Ci aspettiamo che l'uscita sia bassa. Per vedere il nome dei nodi, basta guardare in basso a sinistra. Portiamo un ingresso a 1, facendo uno sweep. Durante la simulazione cliccando sui terminali appare un voltmetro/amperometro.

Passiamo a questo punto al layout editor Glade. Per prima cosa associamo al progetto il file tecnologico: file -> new library -> technology file -> "psm025.tch". Il parametro database units/micron settato a **1000** indica che mille unità del database equivalgono a $1\mu m$, per cui un'unità è pari a 1nm . Creiamo la cella NAND2 con file -> new cell, e progettiamo gli nMOSFET come incroci tra area attiva e poly. Possiamo partire dalla realizzazione dell'area attiva. Si seleziona il layer corrispondente nella sezione LSW a destra, e per disegnare un rettangolo si preme tasto R. A questo punto, si seleziona l'oggetto e si preme il tasto Q per impostarne le dimensioni. Per aggiustare l'altezza, dato che non si tratta di una dimensione certa a priori, si può agire graficamente con uno stretch. Glade, così come Virtuoso di Cadence, ha due modalità di selezione di una geometria: full, con cui si selezionano tutti i bordi, parziale, con cui si seleziona un lato del bordo. Per passare da una modalità all'altra si preme f4. A questo punto, selezionando il bordo corretto in modalità parziale, con S si può stretchare la geometria in un'unica direzione. Per deselectare cntrl+D o click al di fuori.



Possiamo passare al poly; la lunghezza, adesso, è la dimensione caratteristica L del MOSFET. Scegliamo $W = 0.5\mu m$, $L = 0.25\mu m$. Per mettere in serie i due nMOSFET possiamo fare un copia incolla del poly e deporlo sulla stessa area attiva; il drain di un MOSFET verrà così a coincidere con il source dell'altro. La distanza minima tra i layer di poly è pari a $0.5\mu m$.

Per il controllo delle distanze possiamo utilizzare lo strumento righello con tasto K (per deselectare shift+K). Riduciamo l'extension del poly al minimo ($0.8\mu m$).

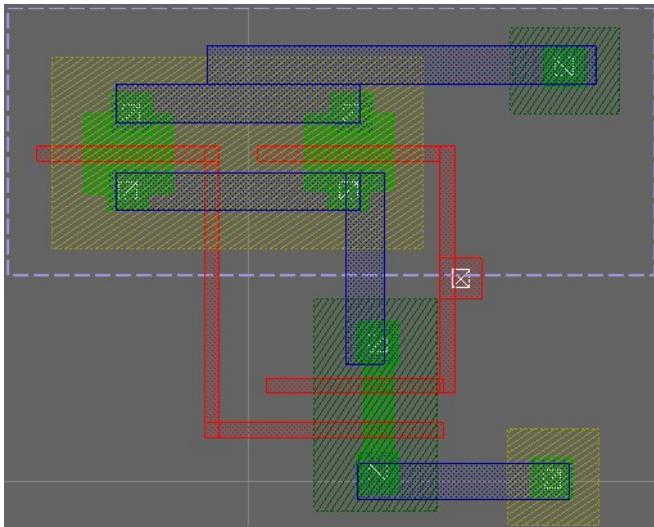


L'inserimento di un contatto si può fare con tasto V. Con f3 si può impostare la tipologia di contatto (M1-active, M1-poly, M1-M2). Attenzione: non scrivere niente in Net Name.

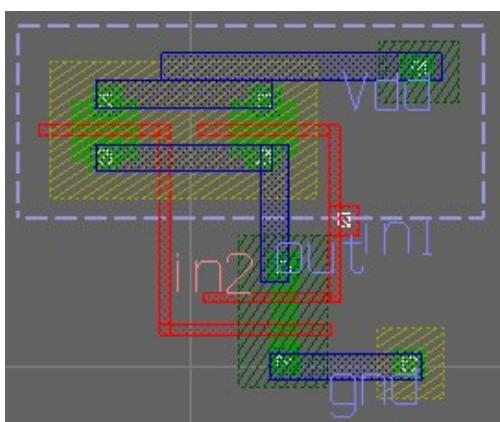
Al di sopra della tavolozza dei layer è possibile rendere invisibili tutti i layer (NV "non visible"), e poi renderne visibili alcuni con tasto destro, oppure riattivarne la

visualizzazione completa (AV “all visible”). Il tasto f3, in generale, apre una finestra di dialogo relativa al comando selezionato, permettendo di personalizzare l’operazione. Ad esempio, premendo f3 con tasto copia si possono specchiare gli oggetti, e copiarli con una molteplicità maggiore di 1.

NAND2 layout:

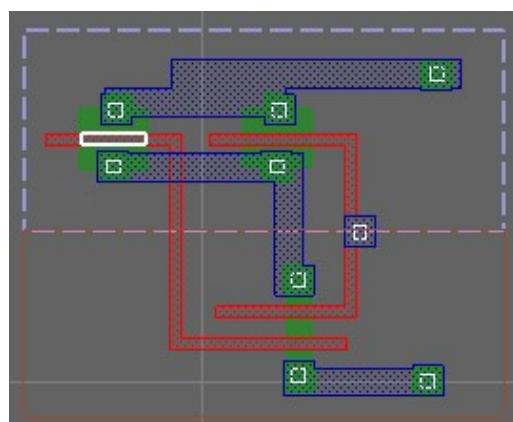


DRC -> run DRC. Occorre inserire il file delle regole alla voce “DRC rules file”, assicurandosi di lasciare in bianco la voce sottostante. Inseriamo il file “drc.py”, e verifichiamo l’assenza di errori con la funzione verify -> DRC -> view errors. Prima dell’LVS assegniamo i nomi ai pin. In Cadence è necessario che ci sia corrispondenza tra i pin dello schematico e i pin del layout, in Glade non sarebbe obbligatorio. I pin sono specifici dei layer, e si identificano proprio con il nome del layer corrispondente e un suffisso “txt”. Per far corrispondere un pin a un layer si preme il tasto T con la geometria in selezione. Possiamo ora passare all’estrazione: verify -> extract -> run LPE. Occorre ancora una volta linkare un file, il file di estrazione, “extract.py”. Il programma apre automaticamente la cella estratta (figura a destra).



n_well	dwg
active	dwg
poly	dwg
n_plus	dwg
p_plus	dwg
contact	dwg
metal1	dwg
via	dwg

Una volta concluso il layout possiamo passare a verificarne la conformità con il DRC tool: verify ->



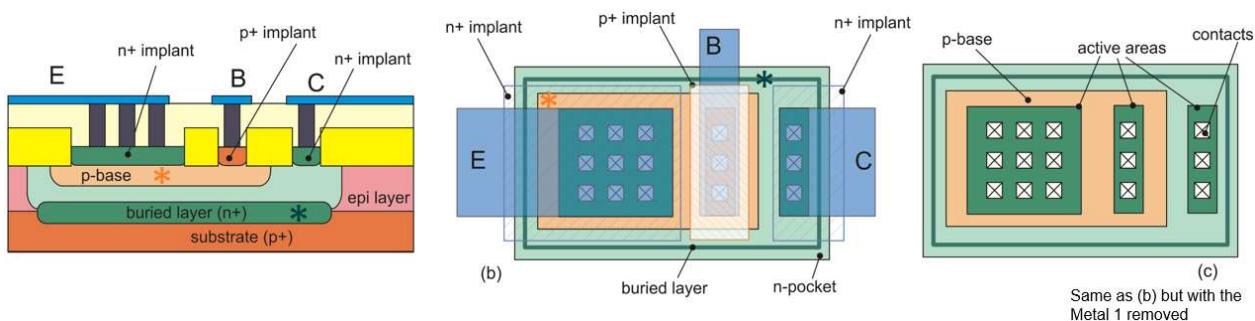
Con un po’ di cura si riesce a selezionare un layer di gate; premendo Q possiamo dunque visualizzare le dimensioni del MOSFET estratto. Nella stessa finestra di dialogo si può navigare in properties -> PCell properties -> properties, e si possono visualizzare le dimensioni L , W estratte.

Ultimo passo è l’LVS: verify -> LVS -> run LVS. Sulla sinistra inseriremo la vista estratta da layout, sulla destra dobbiamo inserire lo schematico di confronto (attenzione: non premere l’opzione “schematic” perché vale soltanto nel caso di aver prodotto lo schematico con glade). Cerchiamo quindi la netlist dello schematico. Su “gemini options” spuntiamo l’opzione che impedisce ai MOSFET in serie di ridursi, e una tolleranza all’1%. Se va tutto bene deve comparire nella message window un’espressione del tipo: 0 devices and 0 nets written to C:\...\nand2.err. L’LVS distingue tra errori topologici ed errori dimensionali. Gli errori topologici sono meno interpretabili, ma per avere un’idea di cosa può esser successo si possono ricercare nell’error log il numero di nodi e il numero di dispositivi delle due viste, e confrontarli.

BJT integrati

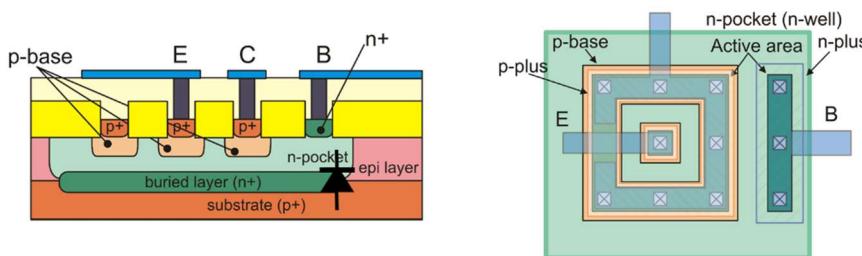
I bipolari possono servire sia in un processo analogico che digitale. A parità di ingombro su silicio il BJT è in grado di portare più corrente. Tuttavia, mentre un MOSFET miniaturizzato e polarizzato alla massima V_{GS} e alla massima V_{DS} resiste, l'analogo bipolare si rovina. Nel mondo del digitale i BJT sono utili per pilotare con una porta logica carichi capacitivi molto grandi esterni al chip, come driver di linea, bus. Un MOSFET, per portare la stessa corrente, dovrebbe essere reso più grande, per cui aggiungerebbe ulteriore carico capacitivo e sarebbe più lento. In effetti, il BJT è il dispositivo che ha la migliore efficienza di transconduttanza: $V_{TE} = V_T$ è la minima possibile, per cui si ottiene più g_m con meno corrente. Nei BJT, tra le altre cose, il rumore flicker normalizzato alla frequenza è più basso di ordini di grandezza. Inoltre, realizzare stadi amplificatori con il BJT comporta offset ordini di grandezza più piccoli rispetto agli stessi circuiti realizzati con MOSFET. Per concludere, i BJT hanno frequenze di transizioni maggiori rispetto al MOSFET (nella pratica si tratta della frequenza oltre cui il dispositivo non serve più a nulla, eccetto per realizzare oscillatori), per cui sono più indicati alle radiofrequenze.

Vertical BJT npn



Una tecnologia che supporta sia BJT che CMOS prende il nome di BCMOS. Per un processo bipolare occorre uno step in più, ovvero un droggaggio leggero per la costruzione della base. Per aumentare il β il droggaggio della base deve essere minore rispetto a quello di emettitore. Al contempo bisogna far sì che, per ridurre l'effetto Early, il droggaggio del collettore sia minore rispetto a quello della base. Questo aspetto, che permette al dispositivo di sopportare tensioni maggiori, ha un effetto negativo in termini di resistenza di collettore. Ecco perché il processo mette a disposizione un buried layer n+, il quale diminuisce la resistenza superficiale del collettore. La regione di collettore è chiamata tasca, sacca o n pocket. All'interno si trova la base con il relativo contatto e all'interno della base si trova l'emettitore. I contatti vengono realizzati per mezzo degli stessi layer CMOS p-plus, n-plus. Il buried layer, al di sotto della tasca, è l'analogo della buried well o triple well. Quindi, il minimo per poter realizzare un BJT con un processo CMOS triple well è l'aggiunta del solo passo di processo del droggaggio che realizza la base. Il BJT integrato è da considerarsi a quattro terminali. Infatti, per ottenere l'isolamento del dispositivo, è presente anche il substrato, il quale non corrisponde ad alcun terminale utile e non influenza le equazioni delle correnti. L'esistenza della giunzione di isolamento, però, fa comparire una capacità di giunzione che potrebbe rallentare il sistema riducendone la banda. Un dispositivo verticale permette di estendere le aree efficaci con facilità.

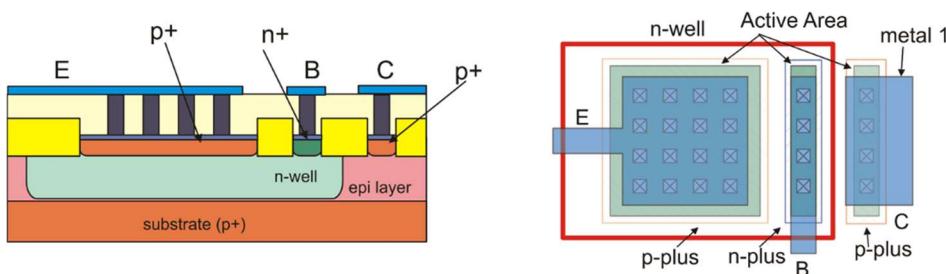
Lateral PNP BJT



Il BJT complementare si può realizzare senza bisogno di ulteriori passi di processo utilizzando il droggaggio di base per l'emettitore e il collettore, il droggaggio n-pocket per la base. Il collettore circonda l'emettitore per raccogliere al meglio gli elettroni.

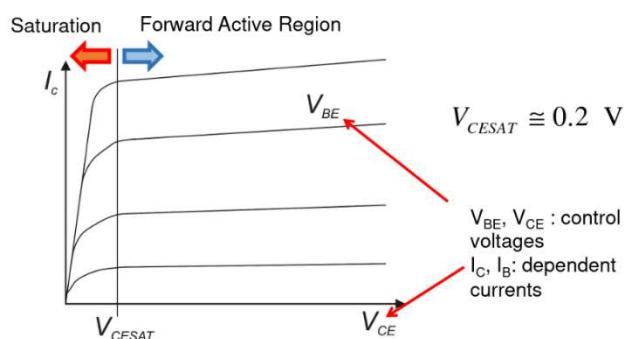
Nel dispositivo laterale la direzione principale è quella orizzontale, per cui le aree effettive sono quelle laterali di affacciamento, le quali non si riescono a fare abbastanza grandi. L'emettitore emette in più direzioni, compresa quella verticale. Questo aumenta la probabilità di ricombinazione dei portatori e comporta una conseguente diminuzione del β . Il dispositivo laterale presenta anche problemi in frequenza. La sacca n ha una resistenza serie e una capacità parassita verso il substrato elevate. Mentre prima la sacca era il collettore in questo caso è la base, il terminale di ingresso più delicato del dispositivo. La squadra RC in ingresso rallenta molto il dispositivo, che quindi non può essere usato alle radiofrequenze, ma al più per polarizzare altri stadi. Anche la tensione di Early è scadente: per avere un'elevata V_A il collettore dovrebbe essere significativamente meno drogato della base. A parità di corrente, rispetto ai dispositivi verticali, quelli laterali occupano più spazio, il che li rende non adatti ad applicazioni di potenza. Esistono processi complementari che prevedono passi di processo in più per la realizzazione del pnp verticale.

BJT di substrato (pnp)



Nel BJT pnp di substrato a fare da collettore è il substrato p stesso (a seconda di quanto va in profondità la base si può trovare un collettore p leggero o p plus). La base è una n well (sacca), mentre l'emettitore è una zona drogata p+. Si ritorna ad una conduzione di corrente verticale e ciò rappresenta un vantaggio in termini di prestazioni in potenza. Tuttavia, il β risulta scarso a causa delle dimensioni della base e il collettore è vincolato al potenziale di substrato, il più basso del circuito in questo caso. Se si vuole avere un buon BJT verticale, in generale, bisogna fare affidamento alla tecnologia bipolare complementare, la quale permette di realizzare npn e pnp verticali con layer in più. In ogni caso, il BJT di substrato è sempre disponibile in un processo CMOS standard n well, senza bisogno di altri layer. La possibilità di utilizzarlo davvero, però, è soggettata al fatto di averne o meno una caratterizzazione da parte della fonderia (parametric cell, modelli).

Caratteristiche di uscita BJT



La corrente di base I_B dipende principalmente dalla V_{BE} , per cui le caratteristiche di uscita possono essere parametrizzate con la V_{BE} stessa. Idealmente la corrente di collettore non dovrebbe essere controllata dalla tensione V_{CE} . La dipendenza dalla V_{CE} in ZAD è data principalmente dall'effetto Early: maggiore l'effetto Early, minore l'intercetta tra i prolungamenti delle caratteristiche in ZAD (tensione di Early), più forte la dipendenza della I_C rispetto alla V_{CE} .

Modello del BJT in zona attiva diretta

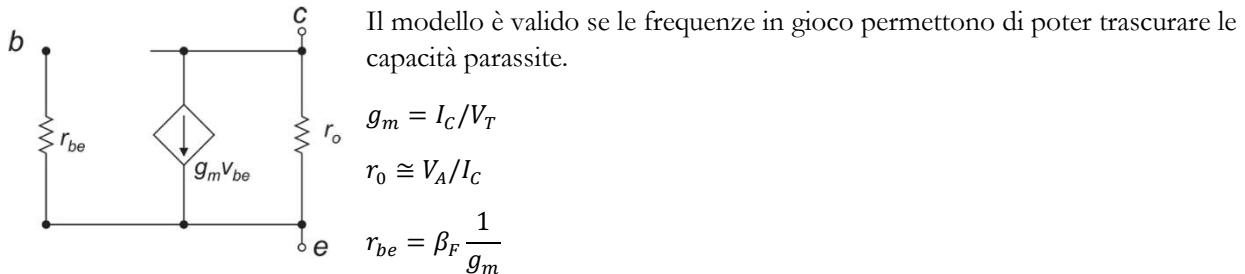
$$I_C = I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB}}{V_A} \right)$$

Con $V_{CB} = V_{CE} - V_{BE}$ (possiamo vedere la V_{CB} come l'analogo della $V_{DS} - V_{DS_{sat}}$). Esclusivamente all'interno delle parentesi la V_{BE} , in quanto figura con dipendenza lineare, si può considerare costante e pari a V_Y . Il parametro V_A è la tensione di Early (in realtà sarebbe quella forward V_{A_f}).

$$I_B = \frac{I_C}{\beta_F}$$

La corrente di base è estremamente semplificata. Per la determinazione di I_C e I_B in tutte le zone di funzionamento (saturazione, cut-off, zona attiva inversa, zona attiva diretta) occorrerebbe usare il modello di Ebers-Moll.

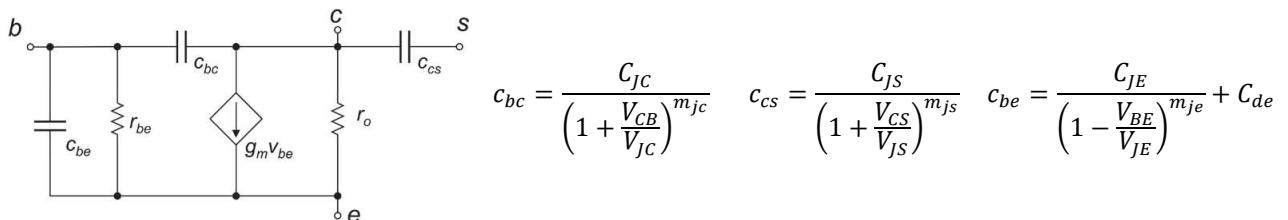
Modello di piccolo segnale BJT



I parametri sono analoghi a quelli del MOSFET: $V_T \leftrightarrow V_{TE}$, $V_A \leftrightarrow \lambda^{-1}$

Il vantaggio di lasciare $g_m \cdot v_{be}$ anziché $hfe \cdot i_b$ è che in tal modo si possono confrontare circuiti a BJT e circuiti a MOSFET, a meno che la I_B non possa essere trascurata.

Capacità di giunzione in ZAD (vertical NPN)



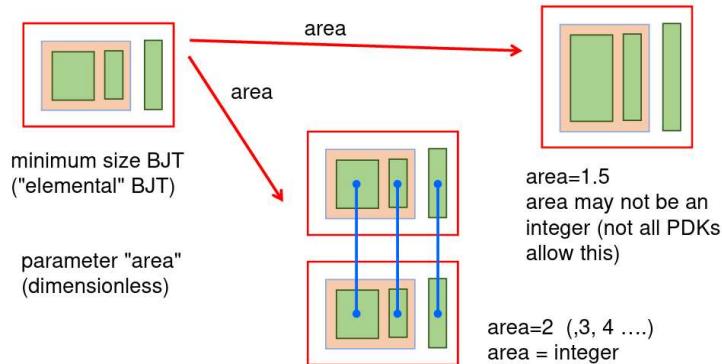
In ZAD la giunzione BC è in inversa o leggermente in diretta; fino a $V_{BC} = 0.5V$, essendo $V_{CE_{sat}} = 0.2V$, la I_C è ancora abbastanza immune ai cambiamenti della V_{CE} . Se non sono coinvolte correnti importanti possiamo modellare le capacità come capacità associate a zone di svuotamento (capacità di transizione). Le giunzioni tra collettore e substrato e tra base e collettore, di norma, sono in inversa. La giunzione base emettitore invece è in diretta, per cui si aggiunge una componente di diffusione che tende a dominare con la tensione, crescendo esponenzialmente. Dato che la base è a comune tra l'emettitore e il collettore, possiamo riferirci alle giunzioni BC e BE come $J*$: C_{J*} è la capacità per tensioni nulle applicate, V_{J*} è il potenziale di contatto, m^{J*} è il coefficiente di grading (0.5 giunzione brusca, 0.3 giunzione lineare). Per quanto riguarda la capacità di diffusione:

$$C_{de} = \tau_F \cdot g_m \quad f_T = \frac{1}{2\pi\tau_F}$$

Se si interrompe il flusso di elettroni provenienti dall'emettitore, c'è un certo numero di minoritari in base che si ricombinano entro un certo tempo di ricombinazione. La sopravvivenza dei minoritari nella base comporta una capacità di diffusione, la quale domina nel determinare la frequenza di transizione del transistore.

Progettazione di un BJT

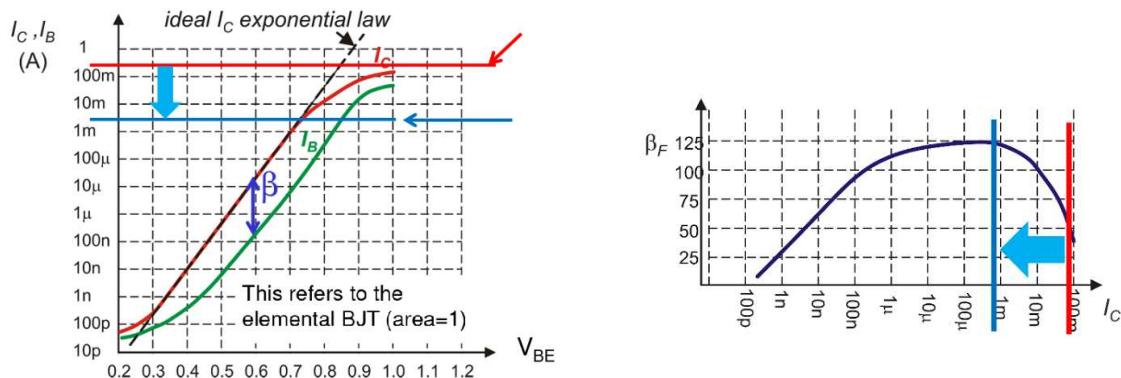
Quando si istanzia un MOSFET i parametri di progetto sono, soprattutto, le dimensioni W e L . Per i bipolari la situazione è diversa. La fonderia fornisce modelli e layout di un dispositivo elementare a dimensioni minime. Le possibilità a disposizione del progettista sono due: utilizzare quel dispositivo o personalizzarlo con il parametro “area”. Si tratta di un parametro adimensionale maggiore dell’unità che corrisponde ad un fattore scalare applicato all’area di emettitore, il quale applica uno stretch in una direzione al dispositivo. Il resto viene ingrandito coerentemente alle regole di layout, i contatti quanto più possibile.



Ci sono casi in cui non è possibile specificare aree frazionarie (dunque stretching dei dispositivi), per esempio i casi in cui contano gli effetti di bordo oppure per i dispositivi laterali. Quando si hanno fattori di area interi vengono posti più dispositivi in parallelo. Rispetto al dispositivo elementare, il ridimensionamento col parametro area può avere o meno effetto sui parametri:

$$I_S \rightarrow \text{area} \times I_S, V_A \rightarrow V_A, \beta \rightarrow \beta, C_{JE} \rightarrow \text{area} \times C_{JE}, R_{CS} \rightarrow R_{CS} \div \text{area}$$

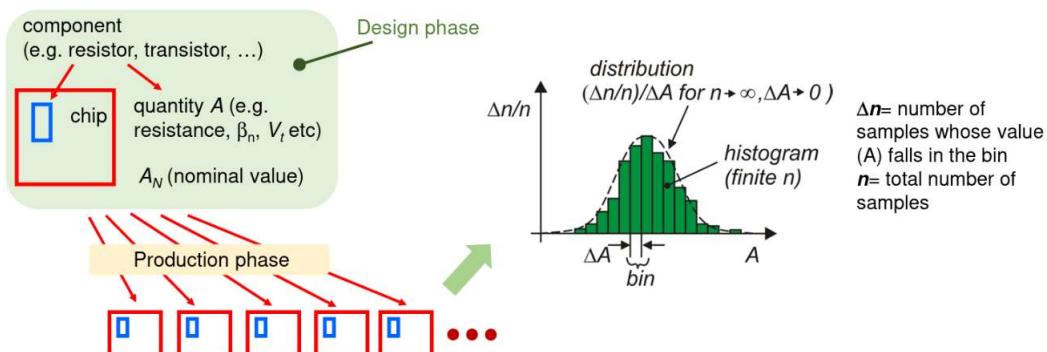
Gummel plot e beta plot



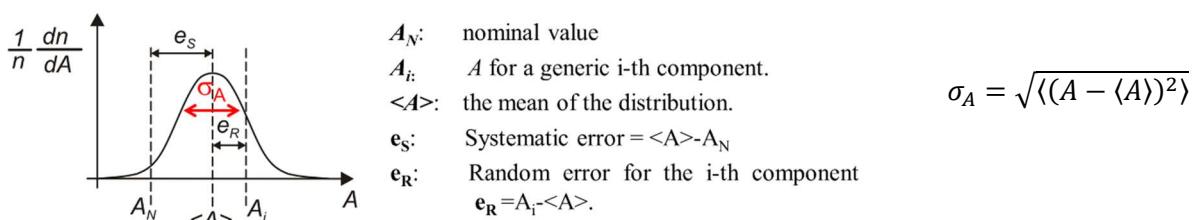
La V_{CE} è tale che il transistore operi in ZAD. Il Gummel Plot chiarisce la regione in cui l’andamento della corrente rispetto alla V_{BE} è veramente esponenziale (in scala semilogaritmica lineare). Agli estremi si hanno delle deviazioni; per correnti basse si fa sentire la ricombinazione, per correnti alte l’effetto Kirk e le cadute sulle resistenze serie. La regione in cui la corrente di collettore e la corrente di base corrono parallele è quella in cui tra loro è costante il rapporto β . Agli estremi le curve si avvicinano, il che significa che β diventa più piccolo. Supponiamo di dover utilizzare un BJT che opera a $I_C = 200$ mA (linea rossa). Non solo la caratteristica non è esponenziale, ma la curva della I_C non arriva nemmeno a quel valore. Da ciò si conclude che il dispositivo si romperebbe. Se però si imposta il parametro area pari a 100, ciascun transistore elementare in parallelo opera portando 1/100 della corrente (linea blu) in una zona in cui il funzionamento è ottimo in termini di β_F (beta plot). Continuando ad aumentare il fattore area, pur rimanendo la corrente totale la stessa, non solo diminuisce la corrente portata dai singoli dispositivi, ma diminuisce anche la V_{BE} con cui sono polarizzati, utile per circuiti low voltage.

Errori di processo

Tutto quello che si progetta, in qualsiasi campo della tecnica, è soggetto a non essere realizzato per come si vorrebbe a causa di errori di fabbricazione. Dobbiamo riuscire a quantificare e predire gli errori di processo. Immaginiamo di avere un chip e un componente generico su di esso, fissato a una certa tensione. Di questo componente valutiamo una sua caratteristica A , che consideriamo come parametro di progetto: ad esempio, il β_n per un MOSFET, R per un resistore, ecc. Il valore che assume il parametro in fase di progetto lo indichiamo con il pedice N , ad indicare che si tratta di un valore nominale. In fase di produzione si costruiranno tanti di questi chip, per cui si avranno tante copie del componente.



Se si procedesse alla misura di A in tutti i chip, l'effetto degli errori di processo sarebbe quello di distribuire il parametro A su un istogramma descritto da una certa statistica. L'asse del valore che può assumere A viene suddiviso in tanti intervalli (bin) di cui si stabilisce un'estensione tale da contenere un numero ragionevole di campioni: includerne troppi renderebbe la statistica meno efficace, includerne pochi renderebbe la curva frastagliata, piena di buchi. Per ciascun bin si conta il numero di chip Δn il cui parametro A è contenuto nel range del bin stesso, normalizzando al numero totale di chip analizzati n . Se si potesse far tenere $n \rightarrow \infty$ e proporzionalmente $\Delta A \rightarrow 0$, l'istogramma tenderebbe ad una distribuzione. Ciò che è importante capire è che questa valutazione statistica è rigorosamente fatta a posteriori della produzione dei campioni e della loro misurazione. Ragionare a priori produrrebbe una distribuzione non più statistica, ma di probabilità. Tuttavia, a priori occorre progettare in modo da ottenere una certa distribuzione statistica, per cui in un certo senso si progetta con "cognizione di probabilità". Ottenuta la distribuzione:



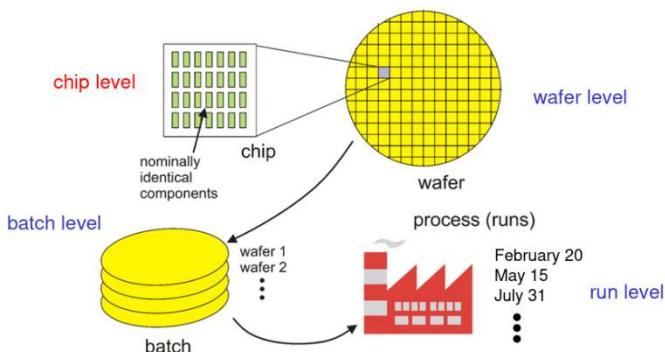
Il valore di progetto A_N , quello che si specifica quando si estrapola e si piazza la cella dal design kit, idealmente dovrebbe essere il valore assunto dal parametro A in tutti i chip. Un po' più realisticamente dovrebbe quanto meno coincidere al valor medio $\langle A \rangle$ della distribuzione, ma anche questa è una condizione statistica ideale. Ciò che allontana il valore nominale da quello medio è l'errore sistematico. Ad esempio, il valore nominale che si dà alle resistenze estrapolate dal design kit non tiene conto dei contatti, che aggiungono sistematicamente la loro resistenza in serie a quella del corpo del resistore. Gli errori sistematici possono anche derivare da defezioni dei modelli simulativi. Ad esempio, progettando un amplificatore da guadagno 1000, potremmo trovare sistematicamente dei guadagni inferiori, in media, a causa di non accuratezze dei modelli. A far disperdere i valori A_i del parametro rispetto al valore medio $\langle A \rangle$ è invece l'errore casuale. A tal proposito, un altro parametro statistico di interesse è la varianza, cioè il valore quadratico medio degli scarti. In realtà, poi, si considererà la radice quadrata della varianza, cioè la deviazione standard. Pur continuando a rappresentare la dispersione della distribuzione, la deviazione standard ha il vantaggio di conservare le stesse dimensioni della grandezza di interesse.

Non sempre la distribuzione del parametro è una gaussiana. Si fa riferimento a questa poiché in natura, quando un fenomeno è perturbato da una moltitudine di cause indipendenti, anche se queste si comportano ognuna statisticamente in maniera diversa la loro combinazione lineare origina una distribuzione gaussiana (teorema del limite centrale).

Max deviation from the mean value	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$	$\pm 4\sigma$
Fraction of data within the interval	68.3 %	95.4 %	99.7 %	99.994 %
Fraction of data outside the interval	31.7 %	4.6 %	0.3 %	0.006%

Progettare “a tre sigma” significa progettare in modo tale che siano accettabili tutti i campioni che ricadono dentro un intervallo di $\pm 3\sigma$. Ad esempio, progettare un amplificatore operazionale con offset 1 mV a tre sigma significa far sì che il 99.7% degli esemplari prodotti abbiano tensione di offset compresa tra -1 mV e $+1 \text{ mV}$. Nominalmente l’amplificatore si progetta per avere tensione di offset nulla. Se il progetto è ben fatto il valore nominale coinciderà a quello medio. Altro esempio, progettare un amplificatore con guadagno 10 e un errore dell’1% a tre sigma significa far sì che il 99.7% degli amplificatori avrà amplificazione compresa tra 9.99 e 10.01. Dovremo essere in grado di prevedere gli errori di processo e far sì che anche in presenza di errori i valori effettivi rientrino all’interno di una fascia accettabile. D’ora in poi considereremo che l’errore sistematico sia piccolo abbastanza da poter considerare che valore nominale e valor medio coincidano. Tutto ciò si tradurrà nella progettazione della deviazione standard.

Errori di processo nei circuiti integrati



Considerando un chip, immaginiamo che al suo interno siano integrati tanti componenti nominalmente identici di cui si vuole valutare il parametro A statisticamente. In tal caso si osserverebbe comunque una distribuzione dovuta ad errori che agiscono all’interno del chip stesso. Poiché la fabbricazione di un integrato è scomposta cronologicamente e tecnologicamente su diversi livelli produttivi, si determina una certa specificità degli errori di processo.

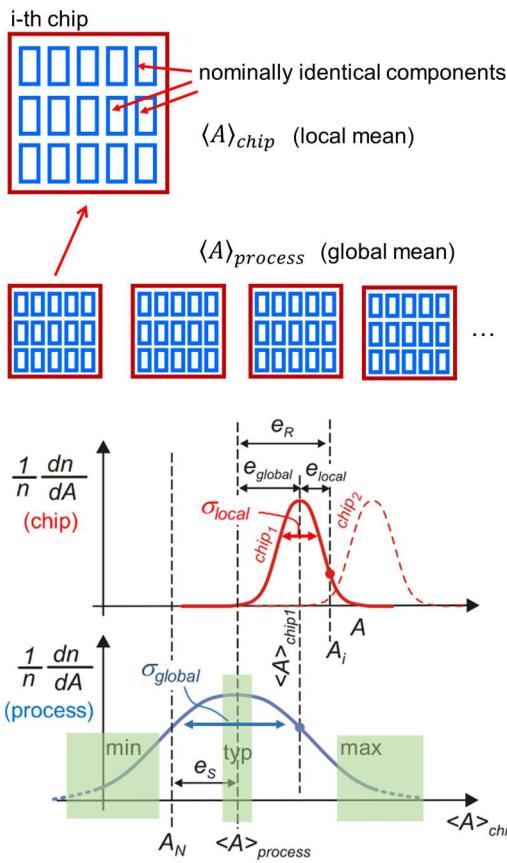
L’errore su chip può essere determinante, ad esempio, nella realizzazione di un amplificatore a retroazione resistiva. Immaginiamo di dover progettare un amplificatore invertente basato su operazionale con guadagno -1 . Il guadagno è affidato al fatto che i due resistori, all’interno dello stesso chip, siano uguali. Se non lo sono, anche il guadagno risentirà di un certo discostamento dal valore nominale.

Esistono poi errori a livello del wafer. Passando da un chip ad un altro nominalmente identico sullo stesso wafer agiscono errori di natura diversa, che possono rendere componenti nominalmente identici tra due chip anche molto diversi tra loro. A questo livello, ad esempio, può pesare particolarmente il diverso droggaggio che si ha nei vari punti del wafer.

All’aumentare del volume di produzione, il componente che abbiamo sotto esame non solo sarà replicato più volte all’interno dello stesso chip e più volte nei chip ricavati sullo stesso wafer, ma in più wafer. Si aggiunge allora un ulteriore livello a cui possono agire errori di processo di natura ancora diversa sul batch dei wafer. Se il volume è ancora maggiore il chip è prodotto in run periodici, cioè in batch ripetuti più volte nell’arco temporale della produzione. Si aggiungono dunque errori di livello ancora superiore, magari dovuti all’usura dei macchinari, allineamenti ottici diversi, caratteri metereologici.

Errori locali ed errori globali

Ai nostri fini, si distinguono soltanto due tipologie di errori di processo: errori locali ed errori globali



Considerando l'i-esimo chip di una produzione, si immagina di avere all'interno tanti componenti nominalmente identici. Del parametro A che caratterizza il componente e che vogliamo analizzare possiamo, a questo livello, calcolare una media statistica sul singolo chip $\langle A \rangle_{chip_{i-th}}$. Si tratta di una media locale. Se si considera il processo di fabbricazione, cioè la produzione di tanti di questi chip, possiamo mediare le medie locali tra il totale dei chip prodotti, ottenendo una media globale del parametro A, $\langle A \rangle_{process}$.

In rosso sono rappresentate le distribuzioni statistiche del parametro A tra i vari componenti, che immaginiamo di far tendere ad una quantità infinita, localmente al singolo chip. Ogni distribuzione locale avrà un valor medio, il quale è anch'esso descritto da una certa statistica globale. Distribuendo le medie locali per tutto il processo si ottiene una distribuzione di processo, in blu, la quale ha anch'essa una media, la media di processo $\langle A \rangle_{process}$. A questo livello, anzitutto si osserva la possibilità che il processo per intero sia affetto da un certo errore sistematico, che sposta la media di A dal valore nominale. Per entrambe le distribuzioni è possibile valutare la deviazione standard.

Consideriamo un singolo componente su un singolo chip. Ciò che causa un discostamento del parametro A misurato in questo singolo caso da $\langle A \rangle_{process}$ è l'errore casuale, il quale è originato da due componenti:

- Errore globale: si tratta della componente dell'errore casuale che causa la deviazione della media locale rispetto alla media globale del processo. Si tratta cioè della componente di errore che descrive statisticamente la dispersione di tutti i valori del parametro A di quel chip rispetto alla media del processo. Se il componente in questione fosse ad esempio un resistore e il parametro A la resistenza, la parte di errore globale ci direbbe con che incidenza in quel chip, per errori che agiscono a livello di processo, le resistenze sono tutte più grandi o più piccole rispetto alla media sul processo.
- Errore locale: si tratta della componente dell'errore casuale che, localmente al chip, causa la deviazione del parametro A misurato rispetto alla media locale. Sempre nell'esempio della resistenza, questa componente valuterebbe con che incidenza statistica i valori delle resistenze di quel chip si disperdonano rispetto a un valore di media locale.

Consideriamo ad esempio di dover realizzare un resistore con valore nominale di resistenza di $1\text{ k}\Omega$. Misurando il resistore in un chip si ottiene un valore di $1.2\text{ k}\Omega$. Il fenomeno è scomponibile in due aspetti diversi:

- Il chip, ad esempio, è caratterizzato da un droggaggio più piccolo poiché ritagliato dalla periferia del wafer, oppure risente maggiormente dell'invecchiamento dei macchinari in quanto è l'ultimo prodotto. Si tratta di fenomeni globali: parte dell'errore si deve al fatto che tutto il chip ha subito un aumento di resistenza per motivi fisici di processo. Rientra in questa categoria anche l'errore sistematico, come quello di aver trascurato le resistenze di contatto.
- Nella casualità di questo valore gioca anche il fatto che resistori nominalmente uguali sul chip, pur essendo in media più resistivi, sono anche diversi tra loro localmente al chip.

Rispetto la distribuzione degli errori globali sul processo, si distinguono tre casi importanti:

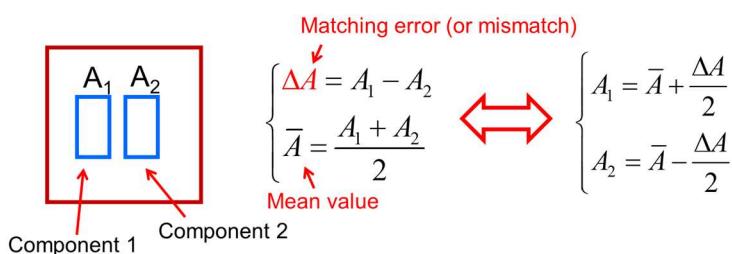
- Caso tipico: il parametro del chip analizzato è vicino al valore medio di processo
- Caso minimo: il parametro del chip analizzato è sulla coda sinistra della distribuzione di processo
- Caso massimo: il parametro del chip analizzato è sulla coda destra della distribuzione di processo

In fase di simulazione dei dispositivi si hanno a disposizione tutti e tre i modelli globali, che prendono anche il nome di corners (corner simulation). Con i corner di processo si possono solo esplorare gli errori globali, cioè osservare il cambiamento dei parametri sotto analisi applicando la stessa modifica a tutti i dispositivi, a tutto il silicio in maniera uniforme. Se si volesse invece simulare l'errore locale, cioè applicare una modifica casuale delle caratteristiche dei dispositivi all'interno dello stesso chip, dovremmo fare una simulazione montecarlo.

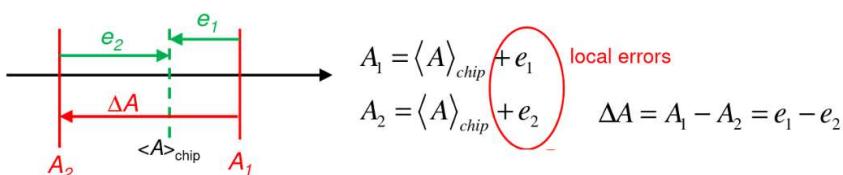
Si osserva una caratteristica molto interessante delle due distribuzioni: $\sigma_{global} \gg \sigma_{local}$. Questo significa che nei processi integrati si ha un pessimo controllo della dispersione sul valore esatto che il parametro del singolo componente avrà. Tra il valore misurato e quello nominale possono esserci anche errori del 20% su diverse run. Tuttavia, localmente, l'incidenza degli errori locali è più stretta, meno dispersa. Ciò significa che nonostante il parametro non sia centrato rispetto quello nominale, localmente al chip quello stesso parametro potrebbe distribuirsi sui diversi componenti con discrepanze di appena l'1%. Detto altrimenti, le differenze dei parametri sono grandi tra chip e chip, mentre su singolo chip sono molto piccole. Tornando al problema dell'amplificatore invertente: se a contare è il rapporto delle resistenze, l'errore globale commesso non è significativo; ciò che importa è che localmente le resistenze siano quanto più uguali possibile tra loro.

I processi di fabbricazione dei componenti discreti non sono migliori. I componenti discreti vengono testati e separati in post-produzione sulla base dei valori dei parametri misurati. Questo spiega perché appare che per i discreti le tolleranze siano piccolissime a basso prezzo. Nel circuito integrato ciò non è ovviamente possibile; si possono solo smistare e scegliere i chip a livello globale, ma non i componenti integrati a livello locale.

Errori di matching



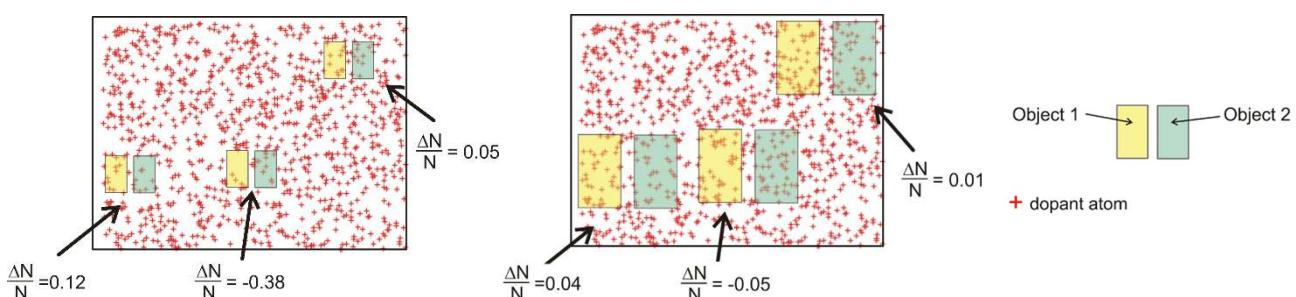
Consideriamo due componenti nominalmente identici sullo stesso chip. L'errore di matching sul parametro A è dato dalla differenza del parametro misurato nel primo componente e il parametro misurato nel secondo. Si definisce anche un valore medio tra le due misurazioni \bar{A} (si tratta di una media aritmetica, non statistica), che permette di scrivere i parametri locali in funzione del parametro medio e l'errore di matching. Tra le cause che rendono diversi i due oggetti ci sono irregolarità microscopiche, gradienti dei parametri. Il mismatch può avere anche una componente sistematica, oltre a quella casuale. Ad esempio, il componente preso come primario potrebbe avere sistematicamente $A_1 > A_2$ in tutti i chip analizzati. In tal caso la media dell'errore di matching fatta su più chip sarebbe non nulla. Ciò potrebbe esser dovuto, ad esempio, da un posizionamento dei componenti tale per cui uno si trovi sistematicamente a una temperatura maggiore/minore dell'altro (prossimità rispetto a elementi di potenza). Generalmente gli errori di matching sono dovuti a errori di design. Si trascurano, d'ora in poi, gli errori di matching sistematici e considereremo solo quelli casuali. L'errore di matching e l'errore locale sono strettamente legati, il fenomeno alla base è lo stesso. Si può estrarre una relazione tra i due:



L'errore di matching misura la discrepanza del parametro tra due componenti individuati all'interno del chip, mentre l'errore locale misura la discrepanza del parametro tra un componente e il valor medio del parametro rispetto tutti i componenti. Gli errori di matching sono causati dagli errori locali, i quali dipendono dalle non uniformità dei parametri fisico-chimici sull'area del chip.

Irregolarità microscopica

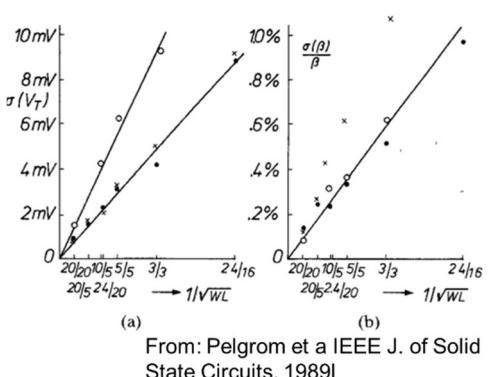
Una prima causa degli errori di matching è l'irregolarità microscopica (granularità locale). Immaginiamo, ad esempio, che il parametro A dipenda dalla concentrazione efficace di drogante in corrispondenza del componente (un altro esempio potrebbe essere una dipendenza dallo spessore fisico del FOX). Un resistore potrebbe presentare resistenza diversa a seconda di quanti atomi di dopante include. Il numero di atomi droganti influenza tante altre proprietà di interesse, tra cui la tensione di soglia di un MOSFET.



Di solito, parlando di matching tra due componenti, non interessa la differenza del parametro in assoluto, ma in relativo rispetto alla media aritmetica. In questo modo è possibile pesare l'errore di matching. Se ad esempio abbiamo un resistore che raccoglie 1000 atomi di drogante e l'altro 990, relativamente parlando l'errore di matching tra i due è piccolo. Una grossa fluttuazione tra gli errori di matching relativi può derivare dall'aver reso i componenti piccoli. Il rimedio naturale che consigue è quello di fare oggetti più grandi. Con rapporto L/W costante la resistenza rimane uguale, ma si può far aumentare il prodotto LW per ottenere area maggiore diminuendo così gli errori di matching relativi.

Modello di Pelgrom

Si può descrivere matematicamente il problema mediante il modello di Pelgrom



$$\text{MOSFET: } \sigma_{\Delta V_t} = \frac{C_{V_t}}{\sqrt{WL}}, \sigma_{\Delta \beta} = \frac{C_\beta}{\sqrt{WL}}$$

$$\text{Resistore: } \sigma_{\Delta R} = \frac{C_R}{\sqrt{WL}}$$

Dove C_{V_t} , $C_{\Delta \beta}/\beta$, $C_{\Delta R}/R$ sono costanti che dipendono dal processo, i cui valori sono forniti dalla fonderia nel manuale di processo, talvolta con nomi diversi. Le unità di misura sono generalmente $mV \cdot \mu m$ per C_{V_t} e μm per C_β e C_R .

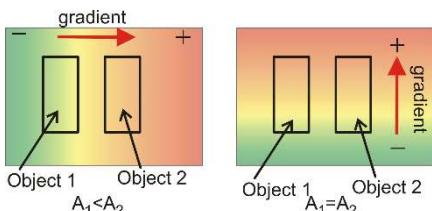
σ_{V_t} è la deviazione standard della distribuzione degli errori di matching sulle tensioni di soglia (o sullo stesso chip per più coppie analoghe di componenti o su più chip per la stessa coppia) di due MOSFET nominalmente identici. Si riporta anche l'errore relativo sul β , adimensionale. Anche per i resistori la σ è valutata su una distribuzione di errore relativo. Tali errori dipendono in modo inversamente proporzionale dalla radice dell'area: se si realizzano MOSFET con W ed L grandi ci si aspetta meno discrepanze tra i parametri. A livello progettuale il modello di Pelgrom permette di ottenere informazioni, per formula inversa, sulle dimensioni dei dispositivi che garantiscono un certo matching fissato.

Gradienti macroscopici

L'altra causa degli errori di matching sono i gradienti macroscopici. Le grandezze che determinano le proprietà dei dispositivi possono variare in modo graduale sull'area del chip, come ad esempio il droggaggio da un punto all'altro del wafer, la temperatura, lo strain. Si tratta di errori puramente casuali, in quanto i gradienti dipendono dal processo, dalla run, dalle condizioni delle macchine, dalle condizioni ambientali. Alcune quantità soggette a gradiente:

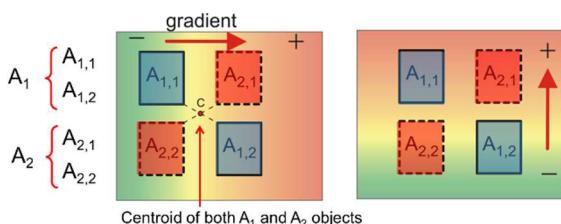
1. Densità di dopante (macroscopicamente)
2. Spessore dell'ossido (es. massimo al centro, gradualmente diverso ai bordi)
3. Bias geometrici (disuniformità dei processi litografici e degli attacchi chimici)
4. Temperatura (dispositivi di potenza presenti sul chip)
5. Stress meccanico (durante il packaging il chip è incollato a caldo su una struttura che, raffreddandosi, a causa del diverso coefficiente di dilatazione termica scarica stress al die)

Sfortunatamente, nella maggior parte dei casi la direzione lungo cui si sviluppa il gradiente non è predicibile. Il gradiente può essere relativo ad una qualche caratteristica fisica che influenza quella di interesse, non necessariamente coincidente alla stessa.



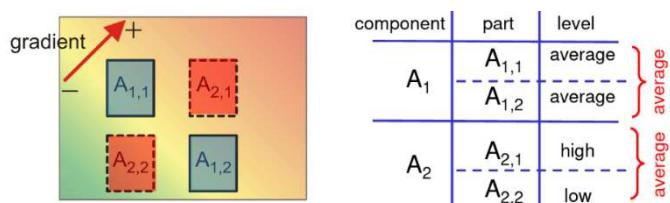
Nel caso a destra il gradiente agisce alla stessa maniera sui componenti, per cui non comporta alcun errore di matching tra i due componenti considerati. Nel caso di mismatch dovuti a gradienti aumentare le dimensioni dei componenti peggiorerebbe il mismatch.

La prima regola per contrastare l'effetto dei gradienti è quella di collocare i due oggetti il più vicino possibile (regola meno efficace per dispositivi grandi). Un'altra tecnica è la disposizione degli oggetti “a baricentro comune” (common centroid), che consiste nel dislocare gli oggetti in più parti e fare in modo che i baricentri delle geometrie coincidano.



L'oggetto A1 è stato suddiviso in due porzioni, così come l'oggetto A2. È possibile fare in modo che il baricentro dei due componenti sia a comune e al centro della struttura complessiva. La simmetria rispetto al gradiente fa sì che questo impatti in egual modo i due componenti lungo entrambe le direzioni ortogonali.

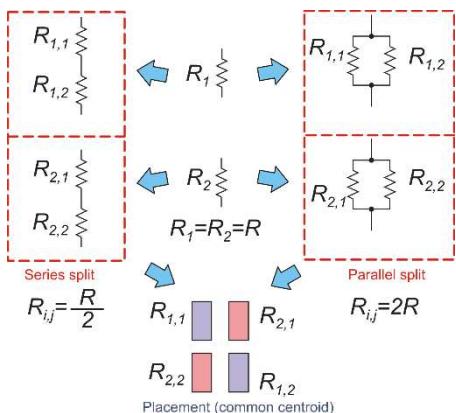
Si ha una compensazione efficace anche con gradiente lungo una direzione obliqua



La configurazione common centroid è una delle tante possibili tecniche di layout che permettono di avere una compensazione dei gradienti.

Resistori a baricentro comune

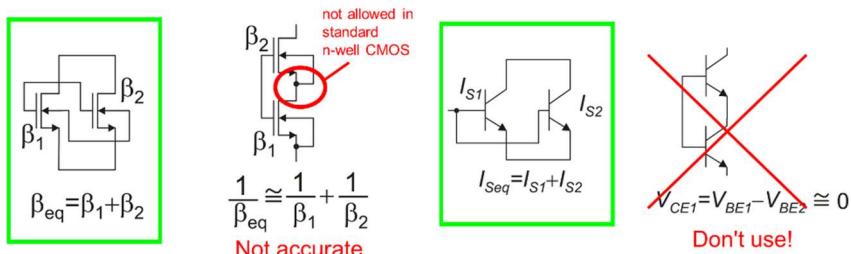
Supponiamo di voler realizzare due resistori nominalmente identici, utilizzati ad esempio nell'anello di reazione di un amplificatore basato su operazionale.



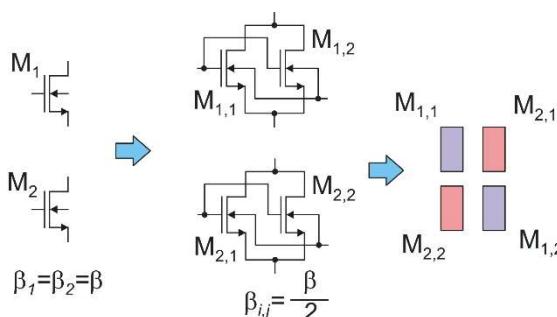
Si tratta sempre e comunque di spezzare gli oggetti in due porzioni. Per ottenere una configurazione common centroid per due resistori ci sono due possibilità: si può suddividere ciascun resistore in due porzioni connesse in serie oppure in due porzioni connesse in parallelo. Rispetto alla resistenza nominale le sotto-porzioni dovranno avere metà della resistenza nel primo caso, il doppio nel secondo. La configurazione che sfrutta il collegamento serie, quindi, permette di risparmiare sull'ingombro di un fattore 4. Se il valore nominale è molto piccolo, tale da non poter essere dimezzato, diventa preferibile la configurazione in parallelo.

Esistono anche altre possibili configurazioni common centroid. Si possono suddividere i componenti ciascuno in più di due unità, alternando poi le sotto-unità a scacchiera anche in modo casuale per migliorare ulteriormente la cancellazione dell'effetto dei gradienti. Per interconnettere gli oggetti tra loro, nella pratica sono necessari almeno due livelli di metal. Spesso anche se si ottiene la simmetria dei componenti non si ottiene comunque quella delle connessioni.

MOSFET e BJT a baricentro comune

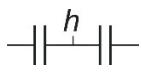


I parametri del MOSFET che possono risentire di mismatch dovuti a gradienti sono il β e la tensione di soglia V_t . L'unica configurazione possibile per ottenere una configurazione common centroid è quella “in parallelo”, che si riduce al connettere gli elettrodi del medesimo tipo assieme a due a due. Due transistori in parallelo mostrano rigorosamente un β_{eq} che è la somma dei rispettivi β . Potremmo pensare di connettere i MOSFET in serie; in tal caso si collegherebbero i gate assieme e ciascun source al relativo body (cosa non possibile in un processo CMOS standard). In tal caso, però, la relazione $\beta_{eq}^{-1} = \beta_1^{-1} + \beta_2^{-1}$ sarebbe meno accurata e, per di più, è facile che i due dispositivi finiscano a lavorare in zone di funzionamento diverse (in particolare, quello sotto in triodo e quello sopra in saturazione). Affinché la configurazione common centroid funzioni è necessario che le sotto-unità in cui è suddiviso un componente abbiano lo stesso ruolo, lo stesso funzionamento. Per il BJT, con la configurazione serie il primo transistore sarebbe forzato ad avere $V_{CE} = 0$. È da preferirsi la configurazione parallelo per entrambi i dispositivi. Consideriamo di voler realizzare due MOSFET M1 e M2 il più uguali possibile per realizzare uno specchio accurato a guadagno unitario o una coppia differenziale a basso offset.



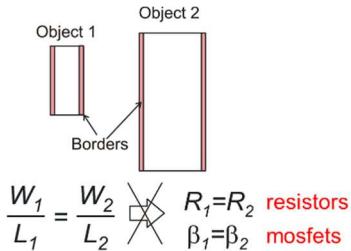
Dato che i β si sommano con la configurazione in parallelo, sarà sufficiente dimezzare i β delle due sotto unità dei MOSFET (privilegiando, spesso, il dimezzamento della larghezza al raddoppiamento della lunghezza). Per il BJT si fa lo stesso: poiché le correnti di saturazione si sommano nel determinare quella del transistore composto, ciascuna sotto-unità dovrà avere la metà della corrente di saturazione finale (se non si può dimezzare dovremo accettare una I_{Seq} doppia).

Condensatori a baricentro comune

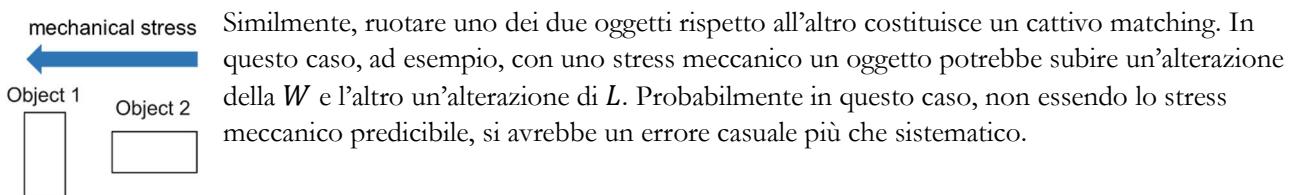
 Per le capacità il collegamento serie è da evitarsi. La ragione sta nel fatto che il nodo centrale tra i due è inaccessibile e flottante.

Il nodo h in continua ha una tensione indeterminata che dipende dalla carica iniziale dei condensatori. Se si accumulasse carica, il nodo flottante potrebbe raggiungere un potenziale tale da distruggere i condensatori. Dunque, anche una coppia di condensatori nominalmente uguali si scompongono in parallelo per ottenere una configurazione a baricentro comune.

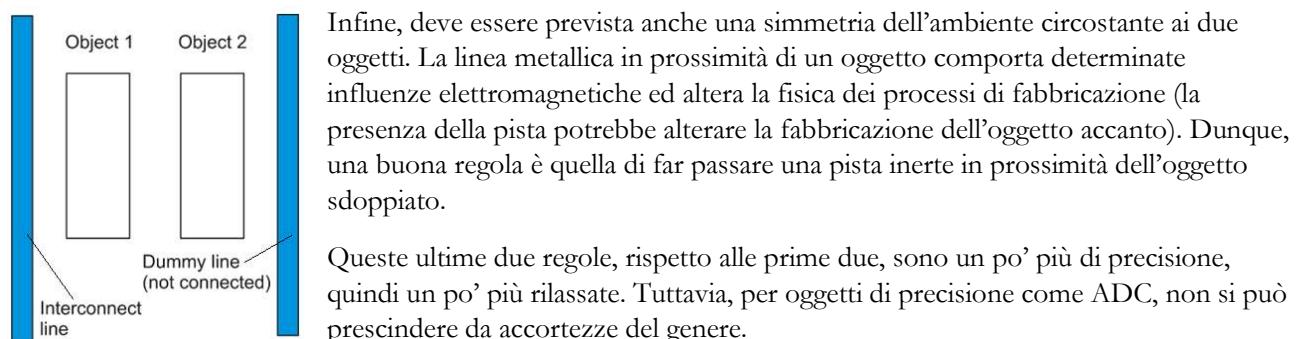
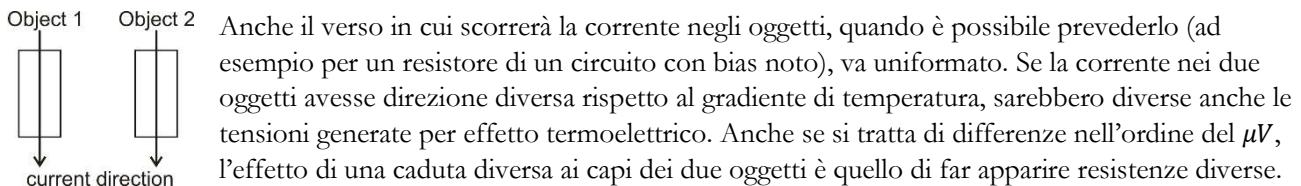
Regole pratiche per evitare errori di matching sistematici



Consideriamo due oggetti nominalmente identici. Dato che molte proprietà sono date dal rapporto W/L , potremmo pensare che mantenendo lo stesso rapporto si conservino i parametri di interesse. In realtà non è così a causa degli effetti di bordo. Per l'oggetto più piccolo i bordi contano di più, per quello più grande di meno. Per il MOSFET subentrano altri fenomeni: con L cambia la tensione di soglia per effetti di canale corto. Per evitare errori sistematici, in definitiva, gli oggetti devono essere ottenuti come copia e incolla.



La buona prassi per realizzare dispositivi nominalmente identici tra di loro si può riassumere affermando che: due dispositivi nominalmente uguali, per un buon matching, devono essere sovrapponibili per traslazione.

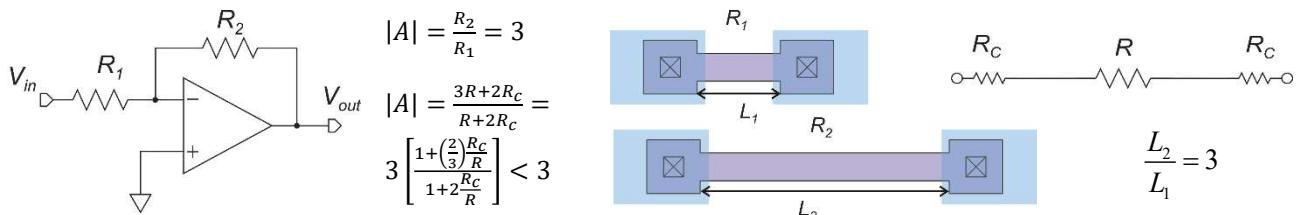


In conclusione, le regole per un buon matching:

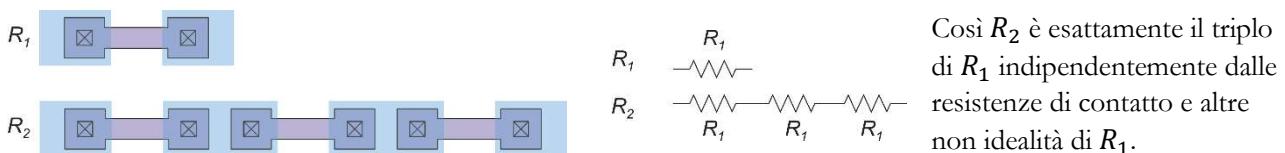
- I dispositivi devono essere nominalmente identici anche per dimensione e orientazione nello spazio
- I dispositivi devono occupare la più grande area possibile (modello di Pelgrom)
- I dispositivi devono essere disposti il più vicino possibile
- Utilizzare la configurazione a baricentro comune
- Uniformare la direzione della corrente nei due dispositivi
- I dispositivi devono “vedere” lo stesso ambiente

Rapporti precisi

Trattiamo ora le tecniche impiegate per realizzare rapporti precisi, ad esempio un rapporto resistivo preciso per ottenere un guadagno preciso. Supponiamo ad esempio di voler progettare un amplificatore invertente con guadagno -3 . Avere guadagni accurati è fondamentale, ad esempio, quando l'amplificatore è direttamente affacciato a un sensore; dividendo per l'amplificazione si ottiene la misura della tensione in uscita del sensore, da cui poi si risale all'entità della grandezza fisica misurata.

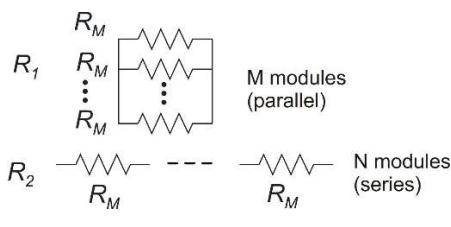


Le resistenze di contatto fanno sì che il rapporto sia sistematicamente minore di 3. L'effetto, a livello globale tra più chip, è un errore sistematico. La soluzione è la configurazione modulare: si spezza il resistore più grande in tre resistori in serie.



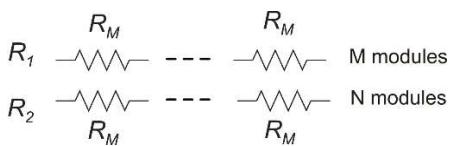
Il matching che conta, adesso, è tra le varie combinazioni di moduli a coppia. Occorre progettare in modo che tutti i moduli siano uguali tra di loro. In questo caso può diventare abbastanza impegnativo, anche se non impossibile, disporre i moduli a baricentro comune (per un rapporto 1:2 in realtà è semplice, basta porre i due moduli della resistenza maggiore agli estremi e l'unico modulo della resistenza minore al centro).

La configurazione modulare è possibile anche in modalità parallelo. Questa possibilità è utile per rapporti molto grandi/molto piccoli.



$$\frac{R_2}{R_1} = \frac{N R_M}{\frac{1}{M} R_M} = N \cdot M$$

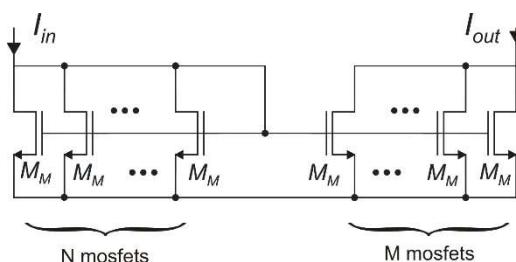
Per un rapporto 1:100 in modalità serie occorrerebbero 101 moduli. Sfruttando il parallelo si può realizzare con 10 moduli una resistenza da valore $\frac{1}{10} R_M$ e con altri 10 una resistenza da valore $10 R_M$.



Si possono anche realizzare rapporti frazionari realizzando entrambi i resistori come serie di più moduli. Nell'esempio a fianco si avrebbe:

$$\frac{R_2}{R_1} = \frac{N}{M}$$

Con la stessa soluzione modulare parallelo si possono ottenere rapporti precisi tra i β dei MOSFET, utili per realizzare specchi di corrente con guadagno preciso. Se si modificasse la W dei transistori per modificare il loro rapporto dei β si commetterebbe un errore dovuto agli effetti di bordo.



In questo esempio si ha uno specchio di corrente ottenuto con due blocchi modulari parallelo. Di fatto, è come avere un unico master con $\beta_1 = N\beta_M$ e un unico slave con $\beta_2 = M\beta_M$. Il rapporto tra I_{out}/I_{in} risulta accurato e pari a M/N .

$$\frac{I_{out}}{I_{in}} = \frac{\beta_2}{\beta_1} = \frac{M\beta_M}{N\beta_M} = \frac{M}{N}$$

Elementi di teoria di propagazione dell'errore

Consideriamo un rapporto di resistenze preciso, realizzato modularmente in modo che $R_2/R_1 = N/M$. Gli errori locali fanno sì che il valore misurato sia diverso da quello nominale. Studiamo in che modo si propagano gli errori su silicio all'errore relativo sul rapporto. L'errore relativo sul rapporto:

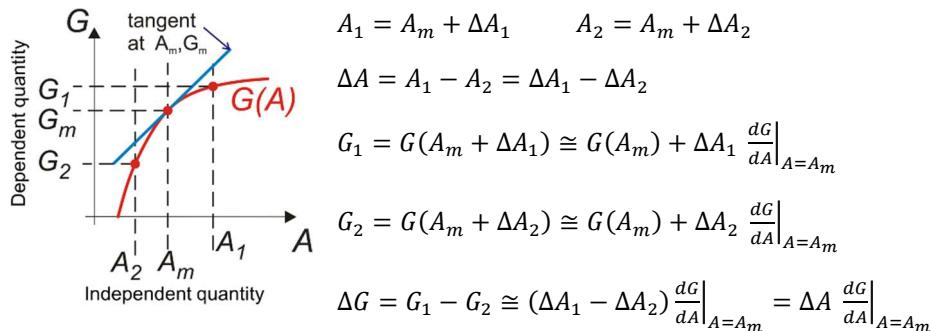
$$\frac{\sigma_{\Delta_r}}{r_{nom}} \cong \sigma_{\delta R} \sqrt{\frac{1}{N} + \frac{1}{M}}$$

$\sigma_{\Delta_r}/r_{nom}$, la deviazione standard del rapporto diviso il rapporto nominale stesso, è equivalente alla deviazione standard dell'errore relativo sul rapporto. Nella formula figurano il numero di moduli utilizzati e l'errore di matching tra una coppia di resistenze $\sigma_{\delta R}$, che si ottiene utilizzando il modello di Pelgrom. Se $N = 1, M = 3$ (come slide), il fattore moltiplicativo dell'errore di matching è $\sqrt{1.333}$. Si deduce che con N ed M molto grandi siamo avvantaggiati.

Nel modello di Pelgrom gli errori, sottoforma di deviazioni standard, sono funzioni di parametri di processo (costanti) e dell'area dei dispositivi. Questi errori si riferiscono a parametri elementari, come l'errore di matching sulla tensione di soglia, l'errore di matching percentuale sul β . Molte volte l'interesse è quello di studiare come gli effetti degli errori di matching dei parametri elementari si propagano su altre grandezze derivate, come ad esempio il guadagno, l'offset. Questa disciplina prende il nome di teoria della propagazione degli errori.

Stima dell'errore

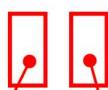
Consideriamo una grandezza G funzione soltanto di una grandezza A indipendente (caso monodimensionale).



I valori A_1 e A_2 possono essere valori diversi che assume la grandezza A nello stesso sistema, nello stesso componente (come due valori di resistenza per uno stesso resistore a temperature diverse), due valori in uscita da un sensore a seguito di un cambiamento nell'ambiente. Nel caso del matching, che si applica anch'esso alla solita trattazione, si considera che A_1 e A_2 siano due valori della stessa grandezza A in due componenti diversi nominalmente identici. Il valore A_m è un punto di riferimento tra A_1 e A_2 , il quale potrebbe coincidere al valor medio o addirittura essere all'esterno dell'intervallo $[A_1, A_2]$. Ciò a cui siamo interessati è $\Delta G(\Delta A)$, come cambia la grandezza derivata a seguito di un cambiamento (anche dovuto a errore) della grandezza elementare. La formula ottenuta per ΔG è tanto più accurata quanto ΔA è piccola. Il posizionamento di A_m , che è arbitrario, rende più o meno accurata l'approssimazione di Taylor a uno dei due estremi; diminuire ΔA rende l'approssimazione migliore a prescindere dalla scelta di A_m . Possiamo risolvere il problema anche nel caso in cui G sia funzione di più variabili elementari.

$$G = G(A, B, C)$$

Anche se ci focalizziamo sull'errore di matching, per cui si considerano due oggetti distinti, i diversi valori delle variabili indipendenti potrebbero essere assunti nello stesso oggetto in due condizioni diverse, per esempio a temperature diverse.



$$(A_1, B_1, C_1) \quad (A_2, B_2, C_2)$$

$P_m = (A_m, B_m, C_m)$ è un punto di riferimento nello spazio a tre dimensioni



$$G_1 = G(A_m + \Delta A_1, B_m + \Delta B_1, C_m + \Delta C_1) \quad G_2 = G(A_m + \Delta A_2, B_m + \Delta B_2, C_m + \Delta C_2)$$

$$\begin{cases} G_1 = G(A_m, B_m, C_m) + \Delta A_1 \frac{\partial G}{\partial A} \Big|_{P_m} + \Delta B_1 \frac{\partial G}{\partial B} \Big|_{P_m} + \Delta C_1 \frac{\partial G}{\partial C} \Big|_{P_m} \\ G_2 = G(A_m, B_m, C_m) + \Delta A_2 \frac{\partial G}{\partial A} \Big|_{P_m} + \Delta B_2 \frac{\partial G}{\partial B} \Big|_{P_m} + \Delta C_2 \frac{\partial G}{\partial C} \Big|_{P_m} \end{cases} \rightarrow \Delta G = G_1 - G_2 \cong \Delta A \frac{\partial G}{\partial A} \Big|_{P_m} + \Delta B \frac{\partial G}{\partial B} \Big|_{P_m} + \Delta C \frac{\partial G}{\partial C} \Big|_{P_m}$$

La corrente in un MOSFET potrebbe essere una $G(A, B)$ con $A = V_t, B = \beta$. P_m è definito in modo che A_m sia compreso in senso lato tra A_1 e A_2 , B_m tra B_1 e B_2 , C_m tra C_1 e C_2 . Vorremmo che il valore nominale dei parametri sia a metà degli estremi.

L'accuratezza della formula, ancora una volta, cambierà a seconda del punto scelto. Tuttavia, fare un errore sulla stima dell'errore può essere meno critico. Se la stima dell'errore è esageratamente in eccesso o in difetto, anche il progetto ne risentirà negativamente. Nel primo caso si introdurrebbe inutilmente una complessità maggiore, nel secondo caso si rischia un malfunzionamento del sistema. Però, commettere un errore anche del 50% sulla stima dell'errore può essere del tutto accettabile.

Espressione posinomiale

$$G(A, B, C) = A^\alpha B^\beta C^\gamma \quad \text{con: } \alpha, \beta, \gamma \text{ reali, } A, B, C \text{ reali positivi}$$

$$\begin{aligned} \frac{\partial G}{\partial A} \Big|_{P_m} &= \alpha A_m^{\alpha-1} B_m^\beta C_m^\gamma & \frac{\partial G}{\partial B} \Big|_{P_m} &= \beta A_m^\alpha B_m^{\beta-1} C_m^\gamma & \frac{\partial G}{\partial C} \Big|_{P_m} &= \gamma A_m^\alpha B_m^\beta C_m^{\gamma-1} \\ \Delta G &= \Delta A \frac{\partial G}{\partial A} \Big|_{P_m} + \Delta B \frac{\partial G}{\partial B} \Big|_{P_m} + \Delta C \frac{\partial G}{\partial C} \Big|_{P_m} = \alpha A_m^{\alpha-1} B_m^\beta C_m^\gamma \Delta A + \beta A_m^\alpha B_m^{\beta-1} C_m^\gamma \Delta B + \gamma A_m^\alpha B_m^\beta C_m^{\gamma-1} \Delta C \end{aligned}$$

Spesso si riescono ad esprimere le leggi dei dispositivi nella forma di un posinomio. Il nome deriva dal fatto che gli esponenti possono essere anche frazionari. Spesso non conta tanto l'errore assoluto, quanto l'errore relativo:

$$\begin{aligned} \frac{\Delta G}{G_m} &= \frac{\Delta G}{G(P_m)} = \frac{\Delta G}{A_m^\alpha B_m^\beta C_m^\gamma} = \frac{\alpha A_m^{\alpha-1} B_m^\beta C_m^\gamma \Delta A + \beta A_m^\alpha B_m^{\beta-1} C_m^\gamma \Delta B + \gamma A_m^\alpha B_m^\beta C_m^{\gamma-1} \Delta C}{A_m^\alpha B_m^\beta C_m^\gamma} \\ \frac{\Delta G}{G_m} &= \alpha \frac{\Delta A}{A_m} + \beta \frac{\Delta B}{B_m} + \gamma \frac{\Delta C}{C_m} \end{aligned}$$

Dunque, se la grandezza di interesse è espressa in forma posinomiale, l'errore relativo è la somma degli errori relativi sulle grandezze indipendenti pesati per i rispettivi coefficienti esponenziali. Se abbiamo bisogno dell'errore assoluto, conviene calcolare quello relativo e moltiplicarlo per G_m .

Esempi

$$P = \frac{V^2}{R} = V_2 R^{-1} \rightarrow \frac{\Delta P}{P_m} = 2 \frac{\Delta V}{V_m} - \frac{\Delta R}{R_m} \quad R = R_s \frac{L}{W} \rightarrow \frac{\Delta R}{R_m} = \frac{\Delta R_s}{R_{s_m}} + \frac{\Delta L}{L_m} - \frac{\Delta W}{W_m}$$

Logaritmo di un posinomio

Immaginiamo ora che la funzione G sia un logaritmo di un posinomio.

$$G(A, B, C) = \ln(A^\alpha B^\beta C^\gamma)$$

Definendo $Z = (A^\alpha B^\beta C^\gamma), Z_m = (A_m^\alpha B_m^\beta C_m^\gamma)$:

$$G = \ln(Z) \rightarrow \Delta G = \Delta Z \frac{dG}{dZ} \Big|_{Z=Z_m} = \Delta Z \frac{d[\ln(Z)]}{dZ} \Big|_{Z=Z_m} = \frac{\Delta Z}{Z_m} = \alpha \frac{\Delta A}{A_m} + \beta \frac{\Delta B}{B_m} + \gamma \frac{\Delta C}{C_m}$$

L'errore assoluto di un logaritmo di un posinomio è pari all'errore relativo dell'argomento, il quale è a sua volta somma degli errori relativi delle variabili indipendenti pesati per i coefficienti esponenziali.

Esempi

$$V_D = V_T \ln\left(\frac{I_D}{I_S}\right) \rightarrow \Delta V_D = V_T \left(\frac{\Delta I_D}{I_D} - \frac{\Delta I_S}{I_S}\right)$$

$$I_C = I_S e^{\frac{V_{be}}{V_T}} \rightarrow V_{be} = V_T \cdot \ln\left(\frac{I_C}{I_S}\right) \rightarrow \Delta V_{be} = V_T \left(\frac{\Delta I_C}{I_C} - \frac{\Delta I_S}{I_S}\right)$$

Queste trattazioni sono accurate per variazioni non troppo grandi, poiché sono basate sull'approssimazione di Taylor al prim'ordine. Per quanto riguarda gli errori di matching, A_m spesso coincide al valor medio \bar{A} , ma può essere il valore nominale o uno a piacere tra gli estremi. In realtà considerare il valore medio non ha alcun riferimento pratico in quanto A_1 e A_2 non sono conosciute a priori, sono anch'esse variabili aleatorie (oltre che la loro differenza). Ecco perché per la scelta di A_m c'è abbastanza libertà. Gli errori di matching godono anche di una proprietà di linearità

$$G = A + B \rightarrow \Delta G = \Delta A + \Delta B$$

$$G = kA \rightarrow \Delta G = k\Delta A \quad \frac{\Delta G}{G} = \frac{\Delta A}{A}$$

Finora abbiamo fatto calcoli deterministici, immaginando di conoscere le variazioni delle grandezze indipendenti. In realtà per gli errori di matching si conoscono solo le deviazioni standard. Quindi il problema sarebbe quello di calcolare la σ_G in funzione di σ_A , σ_B , σ_C . Possiamo considerare gli errori di matching indipendenti (lo sono davvero se le grandezze derivano da componenti diversi); le variazioni microscopiche non sono correlate, i parametri elementari dipendono da una moltitudine di parametri fisici scorrelati.

$$\Delta G = k_1 \Delta A + k_2 \Delta B + k_3 \Delta C \rightarrow \sigma_{\Delta G} = \sqrt{k_1^2 \sigma_{\Delta A}^2 + k_2^2 \sigma_{\Delta B}^2 + k_3^2 \sigma_{\Delta C}^2}$$

dove k_i sono le derivate parziali. Grazie all'indipendenza statistica possiamo calcolare la deviazione standard della ΔG con la regola di Pitagora. Se siamo in grado di avere controllo sulle derivate k_i , sulle deviazioni standard degli errori di matching elementari, possiamo manipolare la deviazione standard della grandezza derivata.

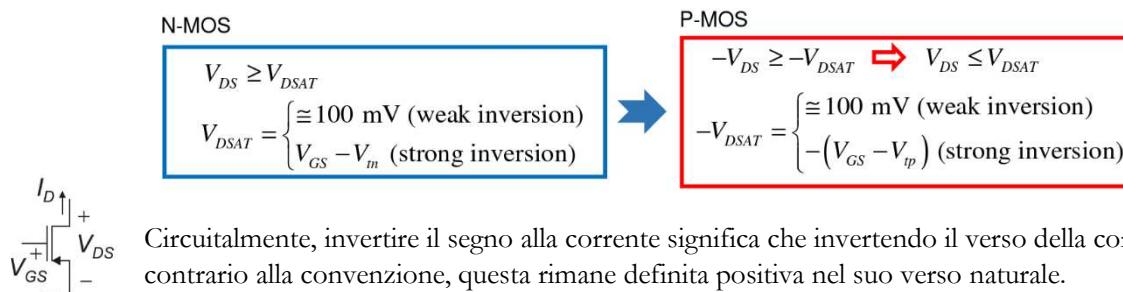
Progettazione analogica di circuiti

Dispositivi complementari

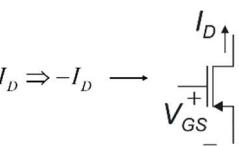
Dall'nMOSFET al pMOSFET

Il circuito di piccolo segnale per i MOSFET complementari è identico. Ai grandi segnali:

N-MOS	P-MOS	
		Conoscendo le equazioni dell'nMOSFET, le equazioni del pMOSFET si ottengono applicando le seguenti trasformazioni:
$V_{DS} \geq 0$	$V_{DS} \leq 0$	$V_{DS_n} \rightarrow -V_{DS_p}$ ($V_{DS_{satn}} \rightarrow -V_{DS_{satp}}$)
$V_{GS} - V_m \geq 0$	$V_{GS} - V_{tp} \leq 0$	$(V_{GS_n} - V_{tn}) \rightarrow - (V_{GS_p} - V_{tp})$
$I_D \geq 0$	$I_D \leq 0$	$I_{Dn} \rightarrow -I_{Dp}$



In qualunque zona di funzionamento il pMOSFET, sia enhancement che depletion, si trova con $V_{DS} < 0$; il terminale di source, in questo caso, è quello a potenziale più alto.


Se per il pMOSFET si considera come verso della corrente di drain quello uscente dal drain, cioè quello contrario al verso convenzionale, tale operazione coincide ad invertire il segno della corrente. Per cui, considerando il verso naturale della corrente non occorre invertirne il segno nelle equazioni.

Un altro modo di vedere le cose per trattare i pMOSFET:

$$V_{DS} \leq 0 \rightarrow -V_{DS} = |V_{DS}|$$

In forte inversione per pMOSFET ad arricchimento:

$$V_{t_p} < 0 \rightarrow -V_{t_p} = |V_{t_p}|, \quad V_{GS} - V_{t_p} \leq 0 \rightarrow V_{GS} \leq V_{t_p} < 0$$

In forte e moderata inversione:

$$(V_{GS} - V_{t_p}) \leq 0 \rightarrow -(V_{GS} - V_{t_p}) = |V_{GS} - V_{t_p}|$$

$$\begin{aligned} -V_{GS} &= |V_{GS}| & + \\ & - \quad -V_{DS} & = |V_{DS}| \\ & I_D \downarrow & \end{aligned}$$

Passing from the gate to the source the voltage is increased by $|V_{GS}|$
 The voltage drop across the mosfet, measured along the natural direction of I_D is $|V_{DS}|$

Scrivere l'equazione in funzione dei moduli significa fare riferimento alle intensità delle grandezze. In questa forma pMOSFET ed nMOSFET hanno equazioni identiche. Con una visione un po' più circuitale possiamo lavorare con V_{DS} e V_{GS} come cadute di tensione.

Dall'npn al pnp

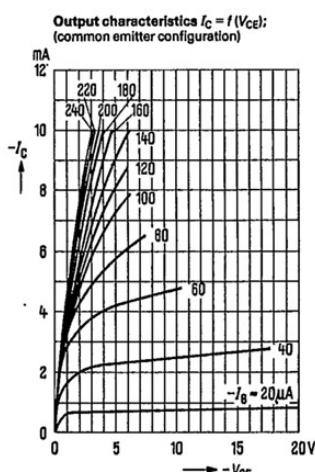
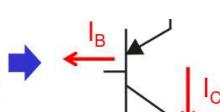


Trasformazioni:

$$V_{CE_{npn}} \rightarrow -V_{CE_{pnp}} \quad (V_{CE_{sat\,npn}} \rightarrow -V_{CE_{sat\,pnp}})$$

$$V_{BE_{npn}} \rightarrow -V_{BE_{pnp}}$$

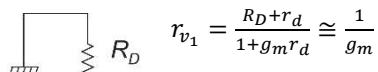
$$\left. \begin{array}{l} I_{C_{npn}} \rightarrow -I_{C_{pnp}} \\ I_{B_{npn}} \rightarrow -I_{B_{pnp}} \end{array} \right\} \text{These two can be avoided if the opposite convention for the current direction is used}$$



Resistenze viste

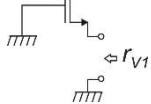
Per semplificare l'analisi dei circuiti in cui figurano un alto numero di dispositivi è fondamentale memorizzare le espressioni delle resistenze viste da un terminale del dispositivo verso massa nelle diverse configurazioni. Le espressioni complete sono eccessivamente complicate, per cui è sufficiente memorizzare le espressioni semplificate e le ipotesi di semplificazione.

Per il MOSFET, trascurando l'effetto body:



$$r_{v1} = \frac{R_D + r_d}{1 + g_m r_d} \cong \frac{1}{g_m}$$

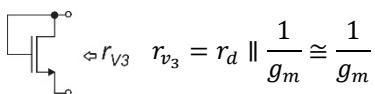
Condizioni per l'approssimazione: $g_m r_d \gg 1$, $R_D \ll r_d$, substrato collegato al source:



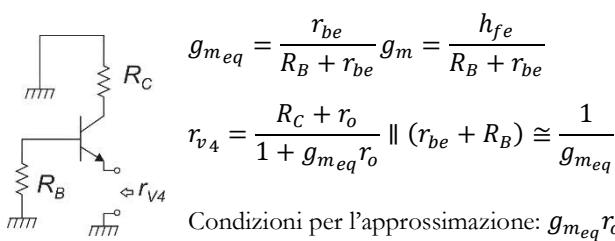
$$r_{v2} = R_S + r_d(1 + g_m R_S)$$

$$r_{v3} = r_d \parallel \frac{1}{g_m} \cong \frac{1}{g_m}$$

Se $R_S = 0$ si vede solo la r_d . Altrimenti, la r_d è amplificata di un bonus di $(1 + g_m R_S)$. L'operazione di degenerazione di source è importante per far aumentare la resistenza di uscita.



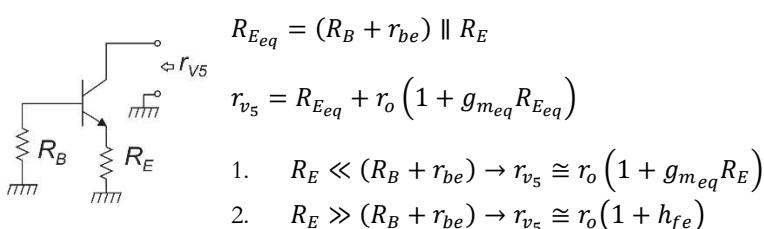
Per il BJT:



$$g_{m_{eq}} = \frac{r_{be}}{R_B + r_{be}} g_m = \frac{h_{fe}}{R_B + r_{be}}$$

$$r_{v4} = \frac{R_C + r_o}{1 + g_{m_{eq}} r_o} \parallel (r_{be} + R_B) \cong \frac{1}{g_{m_{eq}}}$$

Condizioni per l'approssimazione: $g_{m_{eq}} r_o \gg 1$, $R_C \ll r_o$, $\frac{1}{g_{m_{eq}}} \ll (R_B + r_{be})$

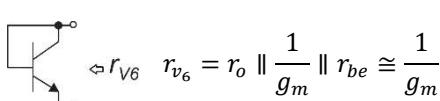


$$R_{E_{eq}} = (R_B + r_{be}) \parallel R_E$$

$$r_{v5} = R_{E_{eq}} + r_o (1 + g_{m_{eq}} R_{E_{eq}})$$

$$1. \quad R_E \ll (R_B + r_{be}) \rightarrow r_{v5} \cong r_o (1 + g_{m_{eq}} R_E)$$

$$2. \quad R_E \gg (R_B + r_{be}) \rightarrow r_{v5} \cong r_o (1 + h_{fe})$$



Progettazione del guadagno intrinseco $g_m r_d$

Possiamo classificare queste resistenze relativamente tra loro in maniera qualitativa, in riferimento a dispositivi polarizzati nelle stesse condizioni (correnti simili, punti di lavoro simili).

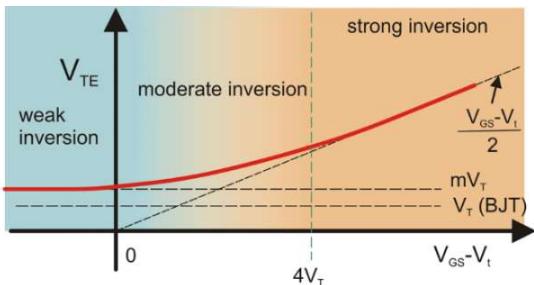
	Small	Medium-large	Large	Very large
MOSFETs	$1/g_m$	-	r_d	$(g_m r_d) r_d$

Il g_m dipende più che altro dalla corrente (la V_t non cambia molto). La resistenza $1/g_m$ cresce al diminuire della corrente. È illegittimo affermare che $1/g_m$ sia una resistenza piccola in assoluto. Per circuiti low power $1/g_m$ potrebbe essere anche nell'ordine dei $M\Omega$. Il confronto è lecito soltanto se fatto su dispositivi analoghi polarizzati con correnti simili. Questi confronti sono utili per capire dove e come, approssimativamente, scorre la maggior parte della corrente alle variazioni. Il prodotto $g_m r_d$ è un fattore importante; si tratta di una cifra di merito statica del dispositivo, che dipende dal transistore e dal suo punto di lavoro. Il g_m e la r_d crollano in zona triodo. Nelle condizioni operative usuali, $g_m r_d$ è nell'ordine del centinaio. Per il BJT:

BJTs	$1/g_m$	$r_{be} (h_{ie})$	r_o	$h_{fe} r_o$

Per il bipolare l'analogico del prodotto $g_m r_d$ è $g_m r_0$ (tipicamente più grande di $g_m r_d$, anche fino a un migliaio). Il $g_m r_d$ rappresenta il massimo guadagno teorico che si può avere in un amplificatore a source comune a singolo stadio; in generale, compare spesso nelle espressioni del guadagno di tensione per diverse topologie circuituali ad alto guadagno. Inoltre, $g_m r_d$ compare anche come bonus moltiplicativo nella resistenza di uscita dei generatori di corrente. In saturazione:

$$g_m = \frac{I_D}{V_{TE}}, r_d = \frac{1}{g_d} = \frac{\lambda^{-1}}{I_D} \rightarrow g_m r_d = \frac{g_m}{I_D} \cdot \frac{1}{\lambda} = \frac{1}{V_{TE}} \cdot \frac{1}{\lambda}$$



Per massimizzare il $g_m r_d$ dobbiamo anzitutto cercare di rendere piccola la V_{TE} . Quindi, per migliorare il $g_m r_d$ occorre polarizzare i dispositivi con overdrive il più piccolo possibile. Da questo punto di vista, la regione della debole inversione è la migliore. Se il dispositivo opera in forte inversione, la condizione migliore è quella in cui l'overdrive si arresti a $4V_T$.

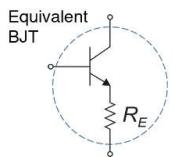
L'altro fattore che influenza il $g_m r_d$ è il λ . Avere un λ piccolo significa che le caratteristiche di uscita in saturazione sono pressoché orizzontali. Questo succede per lunghezze di canali efficaci non troppo piccole: $\lambda^{-1} \propto L_{eff}$. Dunque, la ricetta finale per avere $g_m r_d$ grande è basso overdrive e grandi (non troppo piccole) lunghezze di canale.

Supponiamo che la corrente di bias nel MOSFET sia nota da specifica. Rimane libero l'overdrive, che in forte inversione è pari a $\sqrt{2I_D/\beta}$. Per avere overdrive piccolo a parità di corrente occorre aumentare il β , cioè aumentare W/L . Se volessimo salvare l'ingombro dovremmo diminuire L , ma questo cozzerebbe con la richiesta di avere λ piccolo. Quindi, ciò costringerebbe ad aumentare la W , con conseguente aumento della capacità parassita del dispositivo. Ecco perché ha poco senso cercare di diminuire la V_{TE} passato un certo limite.

Per i bipolar:

$$g_m = \frac{I_D}{V_T}, r_o = \frac{V_A}{I_D} \rightarrow g_m r_o = \frac{V_A}{V_T}$$

Questo risultato è corretto nella regione in cui regge la dipendenza esponenziale della I_C dalla V_{BE} . La tensione termica V_T non è un parametro progettabile e la tensione di Early è fissata dal processo tramite i drogaggi, la forma dei transistori elementari. In effetti $g_m r_o$ si legge come parametro nella model; per $V_A = 25 V$ $g_m r_o$ raggiunge un valore di $\cong 1000$. Dunque, a differenza del MOSFET, $g_m r_o$ non dipende né dalle dimensioni, né dal punto di lavoro del BJT. In generale, le performance dei BJT nei circuiti integrati si alterano con meno gradi di libertà rispetto a quelle dei MOSFET e ciò, da un punto di vista progettuale, non è apprezzato. Una delle ragioni di ciò è che per i BJT il rapporto g_m/I_C è costante e inchiodato a $1/V_T$. Visto che la V_T è il minimo valore che può raggiungere la V_{TE} , dal punto di vista del guadagno intrinseco il BJT è favorito. L'unico modo per variare questo rapporto per il BJT è la degenerazione di emettitore.



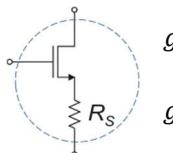
- Lower effective g_m for the same I_D (I_C)
- Higher equivalent r_o
- Higher input resistance (BJT)

$$g_{mrid} \cong \frac{g_m}{1+g_m R_E} \rightarrow \frac{g_{mrid}}{I_C} \cong \frac{1}{V_T} \cdot \frac{1}{1+g_m R_E}$$

$$g_m R_E = \frac{I_C R_E}{V_T}$$

A parità di corrente, il g_m del transistore equivalente è minore. Peggiorare volutamente il rapporto g_m/I_C può essere utile negli stadi di ingresso degli amplificatori operazionali per ridurre lo slew rate a parità di consumo. Il fattore $g_m R_E$ è dato dal rapporto della caduta a riposo sulla R_E e la V_T . Si può ridurre il g_m tanto più quanto si riesce ad aumentare $g_m R_E$ (il quale, se $R_E \ll hie$, fa da bonus anche alla r_o) e ciò comporta di dover aumentare la caduta sulla resistenza R_E . Immaginando di avere una caduta di 1V su R_E : $r_o \rightarrow r_o \cdot 40$, $g_m \rightarrow g_m/40$. Se la corrente I_C è da specifica, il progettista gioca sulla R_E .

Lo stesso si può fare per il MOSFET degenerando il source:



$$g_{mrid} \cong \frac{g_m}{1+g_m R_S} \rightarrow \frac{g_{mrid}}{I_D} \cong \frac{1}{V_{TE}} \cdot \frac{1}{1+g_m R_S}$$

$$g_m R_S = \frac{I_D R_S}{V_{TE}}$$

Nel MOSFET avere $g_m R_S$ sufficientemente grande è un po' più difficile, perché la V_{TE} è sempre più grande di V_T .

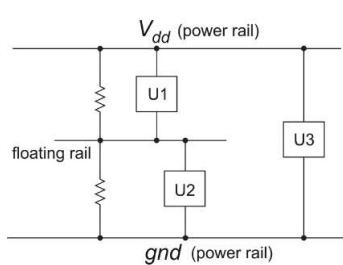
Power rails and floating rails

Le power rails sono le linee di alimentazione del circuito. Si tratta dei collegamenti che, oltre a rappresentare i nodi di riferimento rispetto cui si esprimono le tensioni, portano energia al circuito, fornendo le tensioni e le correnti di alimentazione. Nei casi più comuni si hanno solo due power rail, ma in un caso generale in cui sia presente anche un modulo digitale la situazione potrebbe essere la seguente:

V_{dd1}	16 V	
gnd	0 V	V_{dd} 3.3 V
V_{ss1}	-16 V	

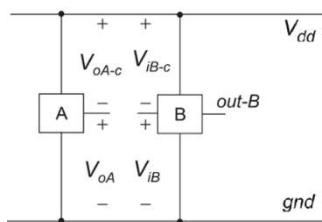
Quelle che sono indicate simbolicamente come linee, nel circuito sono fisicamente dei nodi conduttori. Nell'esempio a fianco si può individuare un dominio ad alta tensione, costituito da tutti i circuiti alimentati tra V_{ss1} e V_{dd_1} , e un dominio a bassa tensione, costituito da tutti i circuiti alimentati tra gnd e V_{dd} .

Spesso si è costretti a lavorare con una single supply. In tal caso si possono generare comunque tensioni utili intermedie (rail intermedio) a cui agganciare circuiti, ad esempio con un partitore resistivo.



I nodi a cui compaiono queste tensioni sono detti floating rail. A differenza dei power rail, se sono utilizzati da circuiti con effetto caricante il loro potenziale può variare. Piuttosto che rendere il partitore pesante, cioè abbassare le resistenze di partizione con conseguente aumento di consumo statico, la soluzione è quella di bufferizzare la tensione intermedia.

Sempre con single supply, consideriamo due blocchi A e B:



La tensione di uscita del blocco A può essere riferita a *gnd*, $V_{oA} = V_{oA} - gnd$, oppure rispetto a V_{dd} in modo complementare:

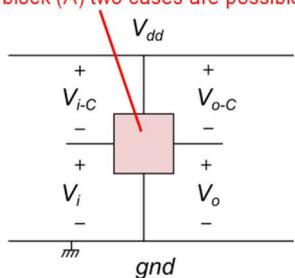
$$V_{oA-c} + V_{oA} = V_{dd} \rightarrow V_{oA-c} = V_{dd} - V_{oA}$$

Se cambia la tensione di alimentazione e una delle due tensioni è costante, l'altra cambia.

Non si possono rendere entrambe le tensioni, quella riferita a *ground* e quella riferita a V_{dd} , entrambe indipendenti dall'alimentazione, ma soltanto una dipendentemente dal circuito. La V_{dd} non è una tensione costante, è soggetta a variazioni: la batteria con cui è generata si scarica e i regolatori non sono perfetti nella loro funzione. Se V_{dd} subisce queste variazioni in maniera aleatoria, ad esempio per spunti di corrente improvvisi, la presenza della V_{dd} nella relazione in-out rende inaffidabile la stessa relazione e produce un disturbo in ciò che consideriamo essere l'uscita. È la topologia del circuito a fissare quale delle rail di alimentazione si può prendere come riferimento per un nodo in quanto invariante rispetto al nodo stesso.

Si può dire la stessa cosa per l'ingresso del circuito B: possiamo considerare una tensione rispetto a *gnd* o V_{dd} , $V_{iB} = V_{iB} - gnd$, $V_{iB-c} = V_{dd} - V_{iB}$, complementari rispetto alla dinamica dell'alimentazione. A seconda di come sono progettati A e B, l'invarianza di input e output rispetto all'alimentazione si ha soltanto per quella delle due componenti che, se fissata costante, mantiene il punto di riposo del circuito.

Depending on the topology of a given block (A) two cases are possible:



Invarianza dell'input:

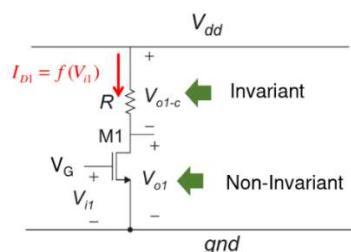
Tra le due espressioni possibili dell'input, V_i o V_{i-c} , l'input "vero" è quello che fa sì che, una volta reso costante, al variare della V_{dd} non si osservino cambiamenti delle correnti interne o delle tensioni di uscita (al I° ordine).

Invarianza dell'output:

Tra le due espressioni possibili dell'output, V_o o V_{o-c} , l'output "vero" è quello che, con input "vero" costante, non cambia al variare della V_{dd} .

Esempio

Consideriamo uno stadio source comune con carico resistivo. La corrente di drain è funzione della tensione riferita a *ground* in quanto, per natura del dispositivo, la corrente dipende dalla V_{GS} .



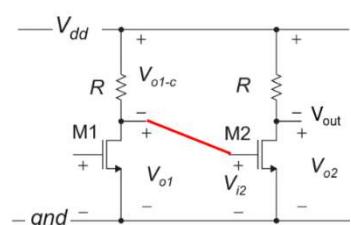
L'ingresso vero è V_{i1} , questo perché $V_{i1} = V_{GS}$ e la I_D principalmente è funzione di V_{GS} . Se considerassimo l'uscita rispetto a *gnd*:

$$V_{o1} = V_{dd} - RI_{D1} = V_{dd} - Rf(V_{i1})$$

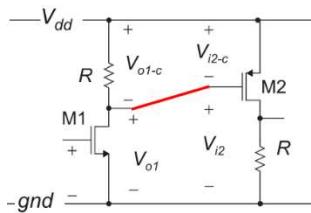
Al variare della V_{dd} cambia la V_{o1} . Invece V_{o1-c} :

$$V_{o1-c} = RI_{D1} = Rf(V_{i1})$$

Dunque, questo stadio amplificatore a source comune, propriamente, ha l'ingresso riferito a *ground* e l'uscita riferita a V_{dd} . Tutto ciò diventa importante quando si hanno due stadi in cascata; in un circuito integrato non si separano gli stadi con un condensatore, i circuiti sono accoppiati direttamente. Consideriamo ad esempio una cascata tra due stadi n-mos a source comune.



Questo collegamento è errato! La tensione V_{o1} riferita a *ground* non è invariante rispetto alla V_{dd} e viene riportata all'ingresso vero del secondo stadio V_{i2} . Variando la V_{dd} varia la V_{o1} , di conseguenza varia V_{i2} , per cui cambia anche la $I_{D2} = f(V_{GS2}) = f(V_{dd} - Rf(V_{i1}))$, per cui cambia l'uscita della cascata V_{out} . Il collegamento corretto prevede l'utilizzo di uno stadio di tipo p.

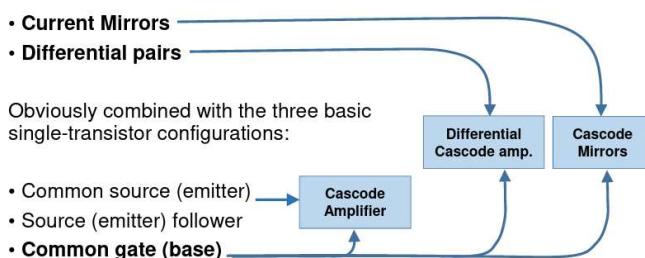


Adesso l'ingresso vero del secondo stadio è quello riferito alla V_{dd} . L'uscita del primo stadio è ora intrinsecamente riferita a V_{dd} , rispetto cui è invariante, ed è collegata ad un ingresso riferito alla V_{dd} . La tensione di uscita prelevata sulla R al secondo stadio risulta invariante rispetto a ground.

In conclusione, l'uscita di uno stadio e l'ingresso dello stadio successivo accoppiato in continua devono essere invarianti rispetto allo stesso rail di alimentazione. La complementarità risolve la questione (è valida anche la configurazione complementare, con primo stadio di tipo p e secondo stadio di tipo n). L'influenza dell'alimentazione si valuta con il PSRR (power supply rejection ratio), migliore se maggiore. In questo esempio, in realtà la V_{o1} rimane un po' sensibile alla V_{dd} a causa della modulazione della V_{ds} sulla corrente, ma è sufficiente porre una degenerazione al source.

Circuiti fondamentali della microelettronica analogica

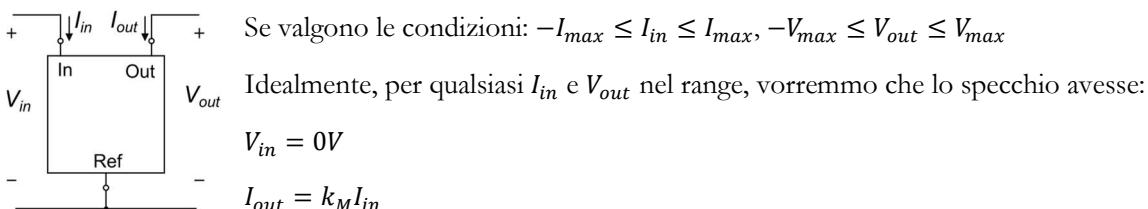
Iniziamo ad analizzare i principali blocchi elementari che combinati opportunamente permettono di realizzare la quasi totalità circuiti dei circuiti analogici integrati.



Combinando common source e common gate si ottengono gli amplificatori cascode, che si possono poi ottenere in versione differenziale, versione specchi.

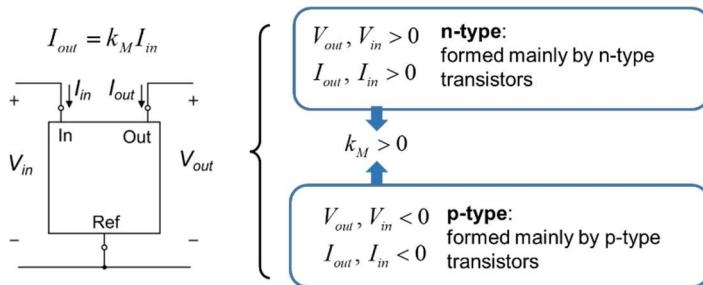
Specchi di corrente

Uno specchio di corrente, in generale, è un sistema a 3 terminali: riferimento, input, output. Le tensioni di ingresso e uscita sono riferite al terminale di riferimento, il quale fa anche da nodo di richiusura per le correnti.



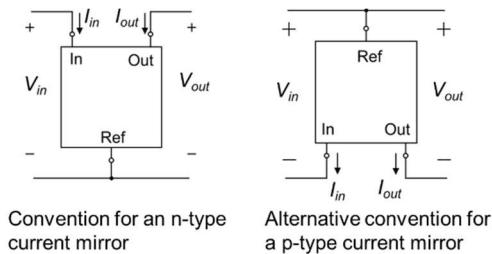
La condizione ideale in ingresso allo specchio sarebbe un cortocircuito, che convoglia la corrente di ingresso senza sviluppare cadute di tensione ai capi; in questo modo non c'è corrente all'interno della resistenza interna del generatore di ingresso. Questo si rifà alla condizione ideale degli amplificatori di corrente, a cui gli specchi aspirano in un certo senso, cioè quella di avere $R_{in} \rightarrow 0$. Altra condizione ideale è che la corrente di uscita sia una replica proporzionale della corrente di ingresso tramite un certo fattore di guadagno in corrente k_M . Vorremmo che le due condizioni fossero rispettate per un certo range di correnti e tensioni anche negative. La tensione V_{out} non è prodotta dallo specchio, il quale si limita a comportarsi come generatore di corrente. È il carico all'uscita, ad esempio una resistenza o un circuito intero, che determina la tensione V_{out} . Queste condizioni ideali non si possono ottenere per uno specchio di corrente elementare, che deve necessariamente essere semplice per poter esser definito tale e per poter essere utilizzato come tale. Esistono amplificatori di corrente più vicini a questi comportamenti, ma sono oggetti complessi che comprendono più blocchi elementari.

Tuttavia, possiamo limitarci a due versioni elementari di specchi di corrente, di tipo p e di tipo n; in tal modo si rispetta la relazione proporzionale $I_{out} = k_M I_{in}$, ma su range di correnti e tensioni dimezzati.

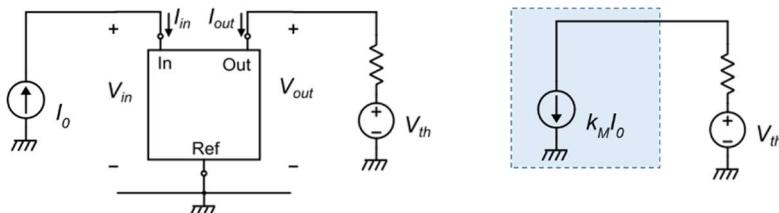


Essendo le correnti entrambe positive o entrambe negative, il coefficiente $k_M > 0$. Soltanto negli amplificatori di corrente si può ottenere un'inversione di segno della corrente tra ingresso e uscita.

Per lo specchio di tipo p si può adottare una convenzione tale per cui correnti e tensioni si mantengano positive:

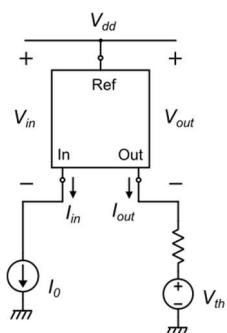


In questo modo le equazioni saranno formalmente identiche e anche per il p-mirror le tensioni e le correnti saranno positive. Gli specchi sono il blocco più pratico e semplice da impiegare per la progettazione di circuiti analogici. Tramite specchi è possibile fare elaborazione di segnale in corrente grazie alla KCL che permette di ottenere somme/differenze di correnti come semplice confluenza di rami in nodi.



Consideriamo la sorgente di corrente ideale sulla sinistra. Il verso della corrente è quello naturale: la corrente sgorga in uscita dal generatore (sourcing). L'unico specchio che può accettare tale corrente in entrata è lo specchio di tipo n. In uscita possiamo schematizzare l'utilizzatore con un equivalente di Thévenin che imposta la tensione V_{out} . Da questo esempio si vede come anche lo specchio, in realtà, imponendo la I_{out} partecipa alla sua tensione di uscita: $V_{out} = V_{th} - RI_{out}$.

L'utilizzatore vede lo specchio come un generatore di corrente, ma all'uscita dello specchio il generatore di corrente equivalente, assorbendo corrente, non si comporta più come una sorgente ma come un pozzo di corrente (sinking). Se il generatore a sinistra fosse stato collegato direttamente all'utilizzatore, la corrente sarebbe stata emessa verso l'utilizzatore e non assorbita dallo stesso. Uno specchio di corrente, quindi, non solo può essere utilizzato per magnificare/demagnificare la corrente, ma anche creare un'inversione di corrente, rispetto ad una sorgente di corrente a monte, nell'utilizzatore a valle. L'inversione di corrente al carico è un possibile impiego per specchi di corrente con k_M unitario. Gli specchi di corrente possono anche essere utilizzati al piccolo segnale. L'importante è che la variazione non sia mai così grande da invertire il verso della corrente di polarizzazione. Il fattore moltiplicativo della corrente viene applicato anche alla variazione: lo specchio si può usare per invertire il senso del segnale in corrente nell'utilizzatore. Gli specchi di corrente sono in effetti blocchi usatissimi per l'elaborazione di segnale, che soffrono di offset, rumore, distorsione lineare.



Mentre uno specchio n trasforma current source → current sink, lo specchio p fa il contrario.

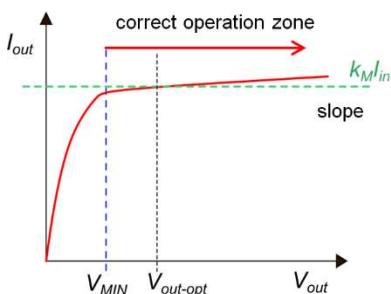
Il generatore di corrente all'ingresso dello specchio assorbe corrente, mentre lo specchio si comporta nei confronti del carico come un generatore di corrente di tipo source.

Parametri di merito di uno specchio di corrente

Per far sì che la relazione $I_{out} = k_M I_{in}$ rimanga buona con sufficiente approssimazione nei quadranti di corrente/tensione giusti occorre che:

1. k_M dipenda soltanto da semplici rapporti geometrici dei dispositivi
2. V_{in} sia pressoché costante (nulla alle variazioni) e in modulo la più piccola possibile
3. V_{out} sia maggiore in modulo di un certo valore V_{min} il più piccolo possibile
4. Per $V_{out} > V_{min}$ la dipendenza $I_{out}(V_{out})$ sia la più piccola possibile

Il punto 4 si può esprimere in funzione di un altro parametro di merito, la resistenza di uscita dello specchio R_{out} , che dovrebbe essere la più grande possibile. I parametri statici individuati, V_{in} , V_{min} , R_{out} , chiariscono in quali condizioni è vera la legge proporzionale tra corrente di ingresso e corrente di uscita.



A fianco si ha la caratteristica di uno specchio di corrente. Si assume che la I_{in} sia costante e pari a un certo valore. La caratteristica ideale è quella verde tratteggiata, quella reale è la curva rossa. Per $V_{out} < V_{min}$ la corrente di uscita cala al diminuire della V_{out} . In questa regione la resistenza di piccolo segnale R_{out} è bassa. In effetti la V_{min} può essere determinata simulando la R_{out} in funzione della V_{out} arrestandosi laddove la R_{out} va sotto il minimo valore accettabile. La zona di corretto funzionamento dello specchio prende anche il nome di dinamica dello specchio (range, swing).

Immaginando che la tensione V_{out} massima sia fissata, ad esempio da limiti fisici di rottura, l'ampiezza della dinamica dipende da quanto più piccola si riesce a rendere la V_{min} . Gli specchi con V_{min} piccola si dicono a larga dinamica (wide swing mirror).

Approssimando la caratteristica nella zona di corretto funzionamento a una retta, questa non sarà a pendenza nulla, cioè R_{out} non sarà infinita. Una R_{out} grande rende accurato il valore della corrente quando la si usa per fare misurazioni, aumenta il guadagno degli amplificatori alla cui uscita si ha uno specchio di corrente e rende lo specchio preciso. Si osserva che spesso (non sempre) esiste una tensione di uscita particolare $V_{out-opt}$ in corrispondenza della quale la legge $I_{out} = k_M I_{in}$ è esatta con un k_M funzione di rapporti geometrici. La richiesta sarebbe quella di avere $V_{out} = V_{out-opt}$ sempre, ma la tensione di uscita dipende dal contesto e spesso l'utilizzatore è un amplificatore, non un carico statico.

Alternativamente alla R_{out} si può derivare un altro parametro di merito degli specchi di corrente

$$V_{th} = I_{out} R_{out}$$

Si definisce V_{th} tensione di Thévenin dello specchio. Il generatore di tensione nell'equivalente di Thévenin di un generatore reale di corrente produrrebbe proprio una tensione data dal prodotto della corrente per la resistenza parallelo.

$$\Delta I_{out} = \Delta V_{out} \cdot \frac{1}{R_{out}}$$

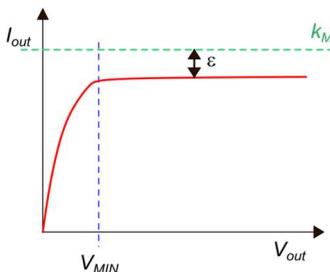
Immaginiamo di avere uno specchio progettato per avere una corrente di riposo di 10 mA e un altro progettato per avere corrente di riposo di $1\text{ }\mu\text{A}$. Da questa relazione si vede che una stessa R_{out} potrebbe essere buona relativamente al primo specchio, ma pessima per il secondo specchio. Ad esempio, ponendo $R_{out} = 1\text{ M}\Omega$, $\Delta V_{out} = 1\text{ V}$, avremmo $\Delta I_{out} = 1\text{ }\mu\text{A}$, variazione che sul primo specchio non pesa nulla, ma sul secondo rappresenta un errore del 100 %. La R_{out} trasferisce il disturbo della V_{out} in una variazione assoluta di corrente di uscita che, se non normalizzata alla corrente operativa dello specchio, è inutile per confrontare in maniera equa specchi tra loro.

Un confronto equo si basa quindi sull'errore relativo:

$$\frac{\Delta I_{out}}{I_{out}} = \Delta V_{out} \cdot \frac{1}{R_{out} I_{out}} = \frac{\Delta V_{out}}{V_{th}}$$

Ecco che al denominatore compare la tensione di Thévenin dello specchio. Questo parametro, quindi, ci permette di confrontare tra loro diverse topologie di specchi di corrente anche quando sono progettati per portare correnti diverse di ordini di grandezza. Se $V_{th} = 100 V$, una variazione di V_{out} di $1V$, a prescindere dal dimensionamento, dalla corrente di riposo, l'errore sulla corrente di uscita è dell'1 %.

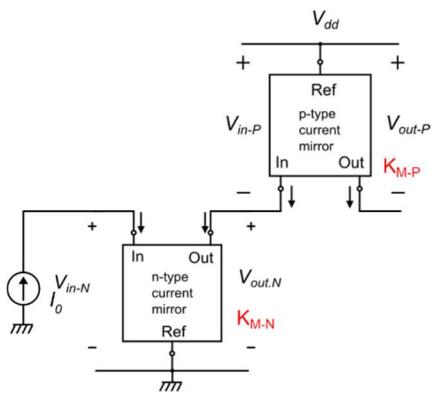
Errori sistematici



In questo caso, anche se la caratteristica di uscita è più orizzontale nella zona buona rispetto a prima, si osserva che non c'è alcun valore della V_{out} tale per cui la legge di proporzionalità tra le correnti sia valida. Qualche volta la cosa è accettabile, certi sistemi funzionano bene se è garantita stabilità più che precisione. Altre volte, come nel caso dei DAC che convertono il codice proprio sommando correnti con degli specchi, interessa avere uno specchio preciso.

In un caso del genere è lecito ridefinire il k_M della legge a un valore più basso. La correzione su silicio è più delicata: il k_M è un parametro pensato per essere progettabile in modo semplice attraverso i rapporti tra gli aspect ratio dei dispositivi (o parametro area dei BJT). Per ottenere un'espressione che tenga conto anche dell'errore sistematico renderemmo il k_M un parametro meno interessante da un punto di vista progettuale, meno efficace ed intuitivo.

Ruolo della V_{in} e V_{min}



Supponiamo di voler moltiplicare una corrente sourcing per un fattore M senza invertirne il verso all'utilizzatore. Lo specchio di ingresso è necessariamente uno specchio di tipo n. Per ottenere il risultato basterà ribaltare di nuovo la corrente di uscita dallo specchio n, sinking, con uno specchio di tipo p. Si viene a determinare un limite sull'alimentazione:

$$V_{out-n} + V_{in-p} = V_{dd}$$

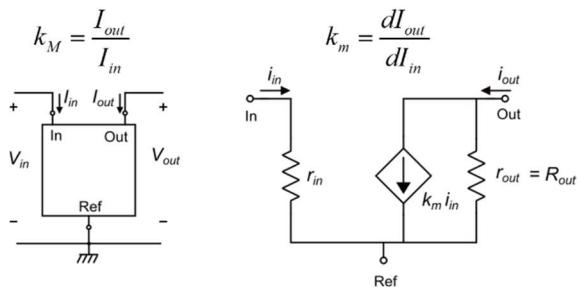
La tensione di ingresso dello specchio p, V_{in-p} , dipende dalla corrente che si richiede allo specchio stesso. Ad essere libera in modo complementare rispetto alla V_{dd} è la tensione di uscita dello specchio n; se cambia la V_{dd} , ad esempio diminuendo, diminuisce la tensione di uscita V_{out-n} .

Fintanto che la $V_{out-n} > V_{min-n}$ il comportamento dello specchio n non è alterato gravosamente, anche se è d'obbligo accertarsi che in uscita si continui ad avere una corrente accettabile. Al di sotto, invece, il sistema non funziona più correttamente come generatore di corrente. La minima tensione di alimentazione che si può avere prima di tale condizione:

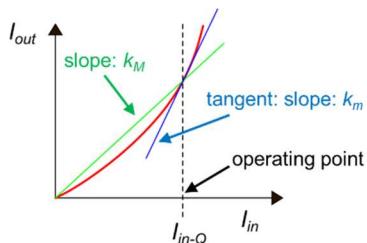
$$\min(V_{dd}) = V_{min-n} + V_{in-p}$$

Più piccole sono la V_{min-n} e la V_{in-p} , più piccola può essere resa la minima tensione di alimentazione del circuito. Tra le due tensioni la più critica è la V_{min-n} ; in alcuni casi è solo questa a porre il limite inferiore alla tensione di alimentazione.

Specchio di corrente ai piccoli segnali

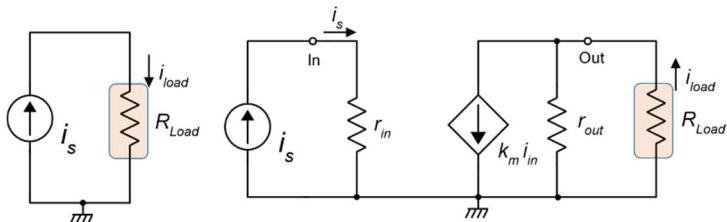


A sinistra lo specchio ai grandi segnali, a destra lo specchio ai piccoli segnali. I guadagni in corrente per grande e piccolo segnale saranno simili, ma diversi in generale. Il motivo è che la caratteristica $I_{out}(I_{in})$ potrebbe non essere esattamente lineare nella zona operativa dello specchio, per cui, scelto un punto, non è detto che il rapporto I_{out}/I_{in} sia pari alla derivata dI_{out}/dI_{in} in quello stesso punto.



Generalmente gli specchi di corrente sono circuiti molto lineari nella loro caratteristica in corrente. Alcuni specchi, come la sorgente di corrente di Widlar, sono volutamente non lineari. In ogni caso, per qualsiasi specchio, non va confuso il guadagno statico k_M con il guadagno dinamico k_m anche se in valore sono spesso simili.

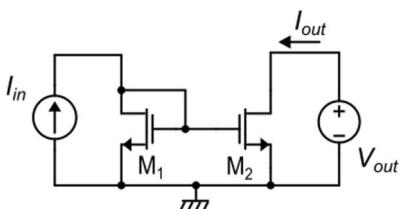
Inversione della corrente alle variazioni



La variazione di corrente nel carico è opposta rispetto a quella che si avrebbe applicando la variazione della corrente di ingresso direttamente al carico. Per cui, anche alle variazioni uno specchio di corrente inverte il senso della corrente.

Questa proprietà può essere utilizzata per calcolare una differenza tra correnti, sia al piccolo che al grande segnale.

Specchio semplice



La corrente di ingresso $I_{in} = I_{D_1}$, la corrente di uscita $I_{out} = I_{D_2}$. I due MOSFET condividono i gate al solito potenziale. Essendo entrambi i body connessi ai source (a massa) non si ha effetto body. Per cui:

$$V_{t_1} = V_{t_2}, V_{GS_1} = V_{GS_2} \rightarrow V_{ov_1} = V_{ov_2}$$

In realtà, se si progetta lo specchio per avere $k_M \neq 1$, le dimensioni dei dispositivi possono essere diverse ed effetti di canale corto potrebbero rendere le tensioni di soglia significativamente diverse. Nel caso in cui si volesse ottenere $k_M \neq 1$ si può comunque evitare il problema con la modularità; ogni modulo avrà la stessa V_t . L'altra causa di tensioni di soglia diverse tra i due dispositivi è l'errore di matching. Assumeremo, d'ora in poi, il caso nominale.

La corrente di ingresso è forzata in M1 grazie al montaggio a diodo. In assenza del lacchetto, all'accensione la V_{GS} è nulla, la corrente di ingresso entra nelle capacità parassite facendo salire il potenziale di drain fino a che non si blocca la sorgente di ingresso. Nel ramo di uscita si osserva corrente nulla. Con il montaggio a diodo, invece, la corrente fa sempre salire il potenziale di drain attraverso le capacità parassite, ma il drain è equipotenziale al gate. Il transistore, quindi, comincia a condurre. La quota di corrente che passa dal transistore è sottratta a quella che fa alzare il potenziale di gate, per cui si rallenta il transitorio, ma alla fine si arriva ad un equilibrio e in M_1 scorre tutta la I_{in} . Si tratta di un meccanismo di retroazione negativa. Per gli specchi in cui è assente il lacchetto, dovrà comunque esser previsto un fenomeno del genere.

In forte inversione:

$$I_D = \beta \frac{(V_{GS} - V_t)^2}{2} [1 + \lambda(V_{DS} - V_{DS_{sat}})]$$

$$\beta = \mu C_{ox} \frac{W_{eff}}{L_{eff}} \cong \mu C_{ox} \frac{W}{L}$$

In debole inversione:

$$I_D = I_{SM} e^{\frac{V_{GS}-V_t}{mV_T}} \left(1 - e^{\frac{-V_{DS}}{V_T}}\right) [1 + \lambda(V_{DS} - V_{DS_{sat}})]$$

$$I_{SM} = \mu_n C_{ox} (m-1) V_T^2 \frac{W_{eff}}{L_{eff}} = \beta (m-1) V_T^2$$

In entrambi i casi $I_D = \beta f(V_{GS} - V_t, V_{DS})$. L'overdrive $V_{GS} - V_t$ è uguale per i due MOSFET, per cui, possiamo scrivere che:

$$I_{out} = \beta_2 f(V_{GS} - V_t, V_{DS_2}) = \beta_2 f(V_{GS} - V_t, V_{out})$$

$$I_{in} = \beta_1 f(V_{GS} - V_t, V_{DS_1}) = \beta_1 f(V_{GS} - V_t, V_{in})$$

Nel fare il rapporto tra le due correnti, idealmente, si dovrebbe ottenere il k_M dipendente soltanto dalle geometrie di layout. Tra i parametri in ingresso alla funzione, l'unico che cambia tra la I_{out} e la I_{in} è la V_{DS} . Allora, a prescindere dalla zona di funzionamento, nel caso particolare in cui:

$$V_{out} = V_{in} \rightarrow \frac{I_{out}}{I_{in}} = \frac{\beta_2}{\beta_1} = \frac{W_2/L_2}{W_1/L_1} = k_M$$

Questo si può esprimere anche dicendo che la $V_{out-opt} = V_{in}$. Non appena cambia la corrente in ingresso cambierà la V_{in} e la corrente di uscita sarà una versione specchiata della corrente in ingresso con tramite un k_M diverso (le V_{DS} saranno diverse, per cui la f non è semplificabile nel rapporto tra le correnti).

Parametri statici

$$V_{in} = V_{DS_1} = V_{GS_1} = V_{GS} = V_t + (V_{GS} - V_t)$$

Esprimere la V_{GS} in questo modo permette di dividerla in due componenti: la tensione di soglia, di norma non progettabile e fissata dalla tecnologia, e l'overdrive, progettabile. In forte inversione:

$$V_{GS} - V_t = \sqrt{2I_{in}/\beta_1}$$

Fissata la V_{GS} (la corrente di ingresso fissa la V_{GS_1} che poi coincide alla V_{GS_2}), la I_{out} al variare della V_{out} corrisponde a una delle caratteristiche di uscita del M2. Quindi, fin tanto che M2 rimane in saturazione la corrente $I_{D_2} = I_{out}$ non dipende molto dalla tensione $V_{DS_2} = V_{out}$; dunque, la zona di saturazione le M2 corrisponde alla zona di corretto funzionamento dello specchio. Se invece M2 entra in triodo, $V_{out} < V_{DS_{sat_2}}$ (essendo le V_{GS} uguali anche le $V_{DS_{sat}}$ sono uguali tra i due dispositivi), la corrente di uscita comincia a dipendere molto dalle variazioni della tensione di uscita. Allora, in generale per lo specchio semplice:

$$V_{min} = V_{DS_{sat_2}}$$

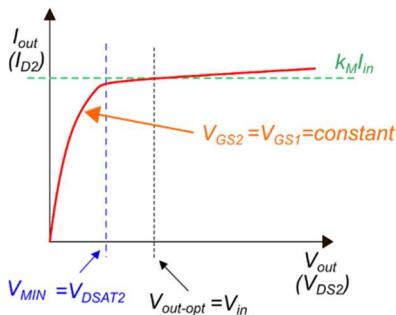
La tensione V_{in} all'incirca, considerando tensioni di soglia di 0.5 V e overdrive minimi di 100 mV è circa pari a una V_y . La V_{min} può essere resa piccola fino a 100 mV. Ciò significa che, combinando uno specchio n ed uno specchio p come nell'esempio precedente, un circuito del genere potrebbe essere alimentato anche a tensioni di $V_{dd} = 0.7$ V.

Parametri dinamici

Per valutare i parametri dinamici dobbiamo ragionare sul circuito alle variazioni. In tal caso i gate sono a massa e guardando dal drain di M2 con source a massa si vede la r_{d_2} . Per cui:

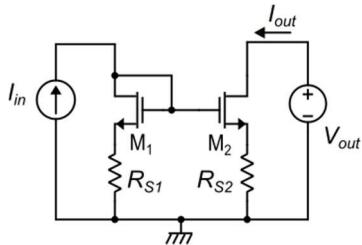
$$R_{out} = r_{d_2} = \frac{1}{\lambda_2 I_{out}} \rightarrow V_{th} = R_{out} I_{out} = \frac{1}{\lambda_2}$$

Si deduce che, scegliendo L più grandi, diminuendo λ_2 , aumenta la R_{out} e quindi la V_{th} .



Lo specchio semplice ha V_{in} e V_{min} ottime... sarà difficile fare di meglio con altre topologie. Inoltre, si tratta di uno specchio accurato dato che esiste una $V_{out-opt}$. Per quanto riguarda la resistenza di uscita, invece, anche se r_{d_2} è classificata come resistenza grande, è ampiamente insufficiente. Aumentare la lunghezza del MOSFET è una cura ragionevole, ma, soprattutto quando lo specchio è utilizzato per processare segnale, non si può eccedere in tal senso. Si può però aumentare la resistenza di uscita degenerando il source.

Specchio semplice con degenerazione di source



Per far sì che le V_{GS} dei due MOSFET siano uguali e che quindi lo specchio si comporti linearmente in corrente, occorre aggiungere una resistenza R_{S_1} anche al source di M1. Inoltre, occorre far sì che la caduta su di essa sia pari a quella su R_{S_2} ($V_{S_1} = V_{S_2} \rightarrow V_{GS_1} = V_{GS_2}$ se i gate sono collegati tra loro).

$$R_1 I_{D_1} = R_2 I_{D_2} \rightarrow \frac{R_1}{R_2} = \frac{I_{out}}{I_{in}} = k_M$$

$$R_{out} = R_{S_2} + r_{d_2}(1 + g_{m_2} R_{S_2}) \cong r_{d_2}(1 + g_{m_2} R_{S_2})$$

Tenendo presente che $g_{m_2} R_{S_2} = I_{D_2} R_{S_2} / V_{TE}$, per guadagnare davvero occorrerebbe fare R_{S_2} grande, ma ciò nasconde un'insidia. Con degenerazione, la tensione di uscita

$$V_{out} = V_{DS_2} + R_{S_2} I_{out}$$

Quando la $V_{DS_2} = V_{DS_{2sat}}$ il M2 entra in triodo e lo specchio esce dalla zona di corretto funzionamento.

Continuando a diminuire V_{out} , anche in ragione del fatto che la resistenza di uscita è più grande, è proprio la V_{DS_2} ad assorbire la variazione, mentre il termine $R_{S_2} I_{out}$ rimane pressoché costante. L'entrata in triodo di M2, contestualizzata nella caratteristica dello specchio, non accade più per $V_{out} = V_{DS_{2sat}}$ come prima, bensì per:

$$V_{out} = V_{DS_{2sat}} + R_{S_2} I_{out} = V_{min}$$

La tensione V_{min} aumenta, riducendo la dinamica dello specchio, ed aumenta tanto più quanto aumenta la R_{S_2} a parità di corrente specchiata. Non solo aumenta la V_{min} , ma aumenta anche la V_{in}

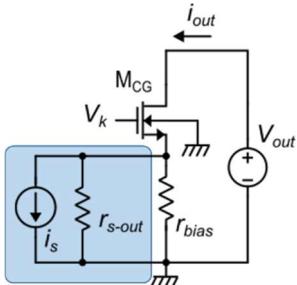
$$V_{in} = V_{GS_1} + R_{S_1} I_{in} = V_t + (V_{GS} - V_t) + R_{S_1} I_{in}$$

Supponiamo di poter sopportare una caduta massima su R_2 di 250 mV. Considerando di essere al limite della forte inversione, $V_{TE} = (V_{GS} - V_t)/2 = 50mV$. Il miglioramento sulla resistenza è di un fattore 6; oltre un fattore 10 la dinamica diventa inaccettabile. Questa tecnica, però, nasconde un altro problema intrinseco, tecnologico. Per metterla in atto occorrono resistori, difficili da integrarsi se non sono presenti nel design kit. Immaginiamo di voler specchiare una corrente di 1 μA ; per ottenere la cauta di 250 mV occorrerebbe una resistenza da $R_2 = 250 k\Omega$, enorme. Tale tecnica è utile solo per applicare un fattore 2 o 3 alla resistenza di uscita; oltre si rischia di peggiorare troppo gli altri parametri dello specchio.

Configurazione specchio CASCODE

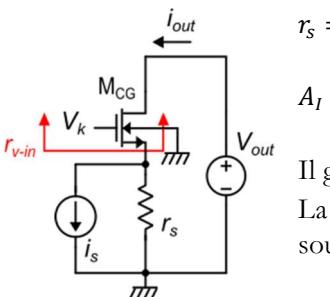
Stadio a gate comune

La configurazione cascode è una cascata tra uno stadio a source comune e uno stadio a gate comune. Prima di trattare il cascode, studiamo un semplice stadio a gate comune alle basse-medie frequenze.



Il gate è fissato a un potenziale fisso V_k , per cui alle variazioni sarà a massa. Il segnale è applicato al source ed è schematizzato come una sorgente equivalente di Norton. Potrebbe trattarsi di una sorgente vera e propria, oppure del circuito equivalente dello stadio a monte. La scelta di una schematizzazione Norton deriva dal fatto che tra le configurazioni di amplificatori, l'amplificatore di corrente è quello che si addice di più a rappresentare uno stadio a gate comune. Il segnale utile è costituito, quindi, dalla corrente di uscita.

Solitamente la sorgente è insufficiente a fornire il percorso della corrente verso ground per la polarizzazione. La r_{bias} rappresenta, in generale, il circuito di polarizzazione alle variazioni. La funzione ideale del gate comune è far sì che la variazione della corrente in ingresso i_s si rifletta in uscita come una stessa variazione della corrente di uscita i_{out} , in modo tale che l'impedenza di uscita appaia alta. A differenza di uno specchio, si nota che non c'è inversione di corrente; in questo caso, sia lo stimolo che l'uscita sono sinking. Se ciò avviene con guadagno unitario, lo stadio a gate comune è ideale. Compattiamo la resistenza interna della sorgente con quella di bias.

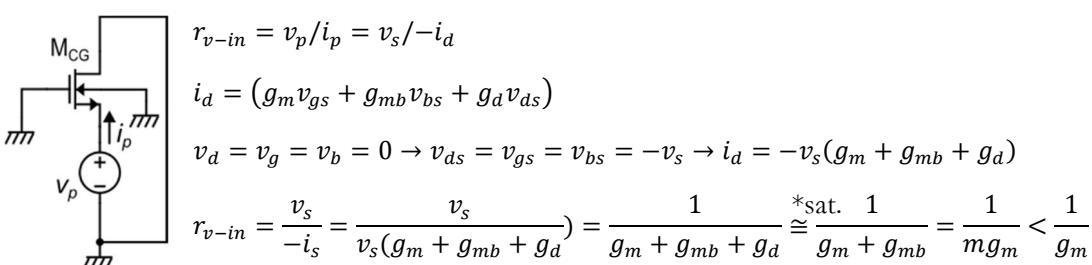


$$r_s = r_{s-out} \parallel r_{bias}$$

$$A_I = \frac{i_{out}}{i_s} = \frac{r_s}{r_s + r_{v-in}}$$

Il guadagno di corrente è riferito alla corrente di cortocircuito in uscita (V_{out} costante). La resistenza r_{v-in} è una delle resistenze notevoli: è quella che si vede entrando dal source con gate e drain a massa.

Ricalcoliamo precisamente questa resistenza tenendo conto dell'effetto body. Poniamo quindi un generatore di prova sul source:

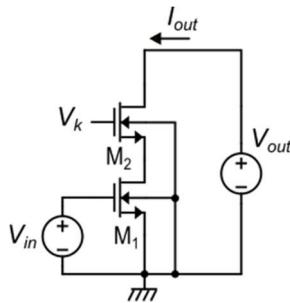


La resistenza di ingresso, tenendo conto l'effetto body, è ancora più piccola di quella che ci potevamo aspettare. Ciò è positivo: in ingresso ad un amplificatore di corrente idealmente vorremmo $R_{in} = 0$ così che tutta la corrente utile entrasse nell'amplificatore. Più piccola è la resistenza, più il guadagno di corrente è vicino a quello ideale unitario. Il guadagno in corrente:

$$A_I = \frac{r_s}{r_s + r_{v-in}} \cong \frac{r_s}{r_s + 1/m g_m} = \frac{m g_m r_s}{1 + m g_m r_s} \cong 1 \quad hp: (m g_m r_s) \gg 1$$

Un buon progetto di uno stadio a gate comune prevede che $g_m r_s$ sia molto maggiore dell'unità. Attaccando in ingresso altre sorgenti di correnti, purché la r_s non diventi troppo piccola rispetto a quella di ingresso, il gate comune si comporta da sommatore di correnti verso l'uscita.

Configurazione cascode (stadio cascode)



Si fa riferimento ad un cascode per circuiti integrati in un processo CMOS n well (i body coincidono tutti al substrato e sono ancorati al potenziale più basso).

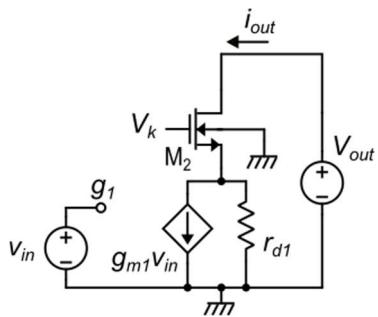
M1: source comune

M2: gate comune

Il transistore M1 riceve un segnale in tensione in ingresso (bias DC + variazioni) e lo converte in corrente di canale. Le variazioni in corrente di drain di M1 sono poi riportate in uscita dal gate comune M2. Dal punto di vista del segnale, la cosa migliore che può fare lo stadio a gate comune è “non fare nulla”, cioè passare il segnale inalterato.

$$V_{DS_1} = V_k - V_{GS_2}$$

La V_k è impostata dall'esterno ed ha la funzione di far sì che la V_{DS_1} sia sufficientemente grande da far operare il transistore M1 in zona di saturazione. La tensione V_{DS} di M1 è fissata, oltre che da V_k , da V_{GS_2} , la quale si può esprimere in funzione della soglia e dell'overdrive. La soglia risentirà un po' dell'effetto body, l'overdrive di M2, in forte inversione, è funzione di $I_{D_2} = I_{D_1}$ e di β_2 . Ciò significa che, una volta impostata la corrente con M1 e stabilite le dimensioni di M2, lo stadio a gate comune ritoccherà quella corrente come effetto secondario e la passerà in uscita. Al contempo, il gate comune “proteggerà” il source comune fissandone la V_{DS_1} . Le variazioni della V_{out} non si ripercuotono direttamente sul drain di M1, il quale ha il gate fissato dal segnale e il drain protetto dal gate comune. Dimostreremo che questa funzionalità corrisponderà ad una elevata resistenza di uscita. Passiamo al circuito di piccolo segnale:



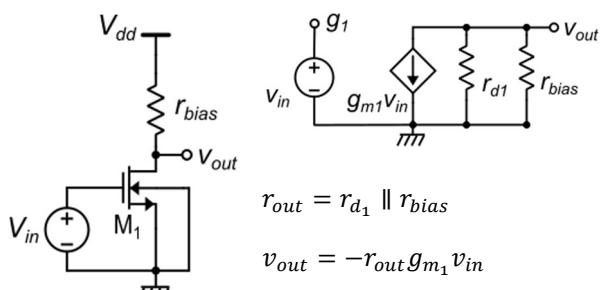
Al posto di una sorgente generica di corrente, adesso il gate comune è stimolato dal M1. È assente la r_{bias} perché M1 in questo caso funge sia da bias che da generatore di segnale.

$$i_{out} = i_{d_1} \cdot A_{I_{CG}} = A_{I_{CG}} g_{m_1} v_{in}$$

Dove $A_{I_{CG}}$ è il guadagno in corrente dello stadio gate comune.

$$A_{I_{CG}} = \frac{m_2 g_{m_2} r_{d_1}}{1 + m_2 g_{m_2} r_{d_1}} \rightarrow \frac{i_{out}}{v_{in}} = g_{m_1} \frac{m_2 g_{m_2} r_{d_1}}{1 + m_2 g_{m_2} r_{d_1}} \cong g_{m_1}$$

I due transistori hanno la stessa corrente di polarizzazione. Se operano anche con lo stesso overdrive (stessa V_{TE}) e hanno lunghezze confrontabili, rispettivamente, $g_{m_1} \cong g_{m_2}$, $r_{d_1} \cong r_{d_2}$. Dunque, è lecito assumere che $g_{m_2} r_{d_1}$ sia comunque un fattore grande, che rende il guadagno in corrente del common gate prossimo a 1. Di conseguenza $i_{out}/i_{in} = i_{out}/g_{m_1} v_1 \rightarrow i_{out}/v_{in} \cong g_{m_1}$, come se dall'uscita si vedesse soltanto il source comune. Per capire i vantaggi della configurazione, mettiamo a confronto uno stadio a source comune e uno stadio cascode usati entrambi come amplificatori di tensione



r_{bias} non necessariamente rappresenta un resistore.

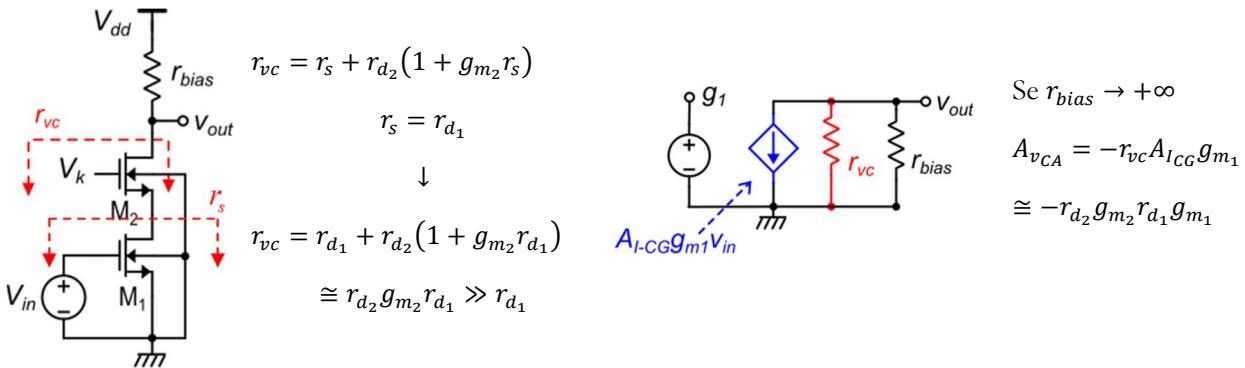
Quando la V_{dd} è bassa, mettere un resistore comporterebbe guadagni troppo bassi. La resistenza r_{bias} sarà la resistenza differenziale di un circuito attivo.

$$r_{out} = r_{d_1} \parallel r_{bias}$$

$$v_{out} = -r_{out} g_{m_1} v_{in}$$

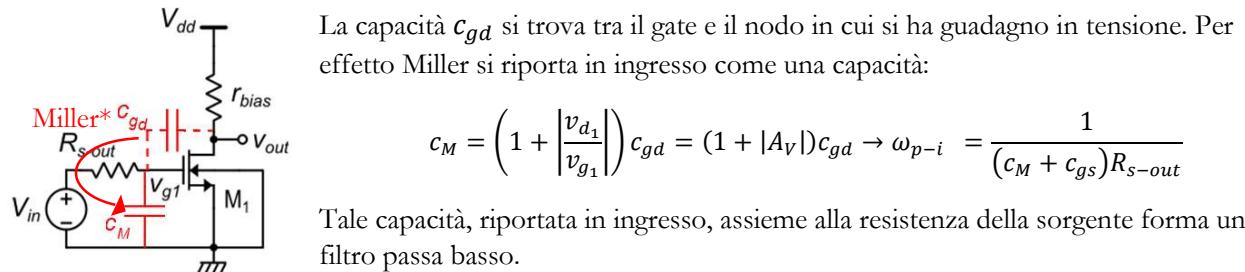
$$\text{Se } r_{bias} \rightarrow +\infty, A_{v_{CS}} = \frac{v_{out}}{v_{in}} = -r_{d_1} g_{m_1}$$

Lo stadio cascode:

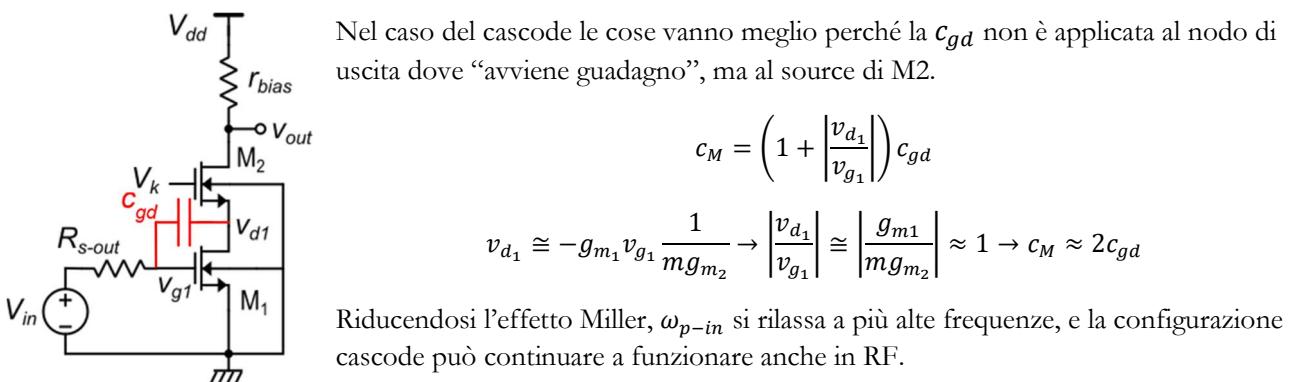


L'effetto benefico di M₂ è quello di aver incrementato la resistenza di uscita. Con la configurazione cascode, in generale, si ottengono guadagni asintotici molto alti (se nel parallelo si può trascurare la r_{bias}), nell'ordine di $(g_m r_d)^2$. Ciò, però, è da considerarsi come una conseguenza dell'aumento di resistenza di uscita. Un amplificatore di tensione può esser visto come un oggetto che converte una tensione di ingresso in una corrente e fornisce la propria uscita in tensione convertendo quella corrente in caduta di tensione su un carico resistivo. A parità di corrente, maggiore è la resistenza del carico all'uscita, maggiore è la tensione sviluppata. Si rivede come lo stadio a gate comune applichi una trasformazione di impedenza: senza alterare troppo la corrente i_{out} siamo passati, all'uscita, dal vedere $r_{d_1}g_{m_2}r_{d_2}$. Altro vantaggio si vede nel comportamento in frequenza.

Per un common source:

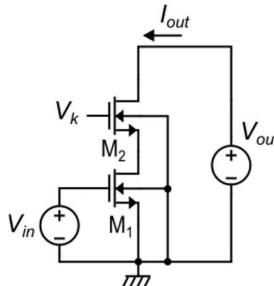


Maggiore è l'effetto Miller, maggiore è c_M , minore è la frequenza entro cui può funzionare il circuito. L'effetto Miller si può comprendere anche con un ragionamento fisico. Immaginiamo di avere un generatore di tensione che vede un condensatore reale. Ad una variazione di tensione corrisponderà una variazione di carica sul condensatore, proporzionalmente alla capacità: $C\Delta V = \Delta Q$. Nel caso in cui l'altra estremità del condensatore sia a un “un nodo di guadagno”, si osserva un ΔQ maggiore per la stessa sollecitazione, per cui appare una capacità maggiore.



L'uso del gate comune nei circuiti analogici integrati è un pilastro. Quando si arriva a tensioni di alimentazione così basse che il cascode non funziona più, è un dramma. Per fortuna, con tensioni basse ma non bassissime, l'uso del cascode è possibile.

Approfondiamo ulteriormente lo studio della struttura cascode. In particolare, studiamo gli effetti della variazione della tensione di uscita v_{out} sulla corrente di uscita i_{out} . Infine, cercheremo di capire come varia la tensione V_{ds_1} al variare della V_{out} .



In questa analisi è attivo il solo generatore v_{out} ; quindi, $v_{gs_1} = v_{bs_1} = 0$.

$$i_{d_1} = g_{m_1}v_{gs_1} + g_{mb_1}v_{bs_1} + g_{d_1}v_{ds_1} = g_{d_1}v_{ds_1}$$

$$i_{d_2} = g_{m_2}v_{gs_2} + g_{mb_2}v_{bs_2} + g_{d_2}v_{ds_2}$$

$$v_{ds_1} = v_{d_1} = v_{s_2}$$

$$v_{gs_2} = v_{bs_2} = v_{ds_2} = v_{out} = -v_{s_2}$$

La variazione di corrente $i_{out} = i_{d_1} = i_{d_2}$ è pari a $g_{d_1}v_{ds_1}$. Se M2 protegge M1, quello che succede è che le variazioni di v_{out} non sono in grado di indurre grosse variazioni di v_{ds_1} , per cui la variazione i_{out} rimane piccola rispetto la variazione di v_{out} (sintomo di una resistenza di uscita alta). Dimostriamo ciò:

$$i_{d_1} = g_{d_1}v_{s_2}$$

$$i_{d_2} = -g_{m_2}v_{s_2} - g_{mb_2}v_{s_2} + g_{d_2}(v_{out} - v_{s_2})$$

$$i_{out} = i_{d_2} = i_{d_1} \rightarrow v_{out}g_{d_2} = v_{s_2}(g_{m_2} + g_{mb_2} + g_{d_2} + g_{d_1})$$

$$v_{ds_1} = v_{s_2} = g_{d_2} \frac{v_{out}}{g_{m_2} + g_{mb_2} + g_{d_2} + g_{d_1}} = \frac{v_{out}}{\frac{g_{m_2}}{g_{d_2}} + \frac{(m_2 - 1)g_{m_2}}{g_{d_2}} + \frac{g_{d_2}}{g_{d_2}} + \frac{g_{d_1}}{g_{d_2}}} = \frac{v_{out}}{m_2 \frac{g_{m_2}}{g_{d_2}} + 1 + \frac{g_{d_1}}{g_{d_2}}} \rightarrow$$

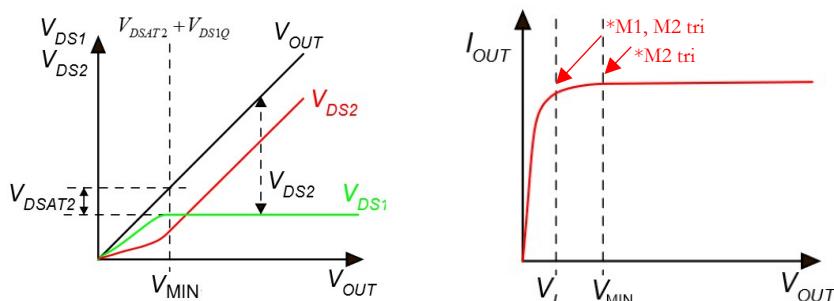
$$v_{ds_1} \approx \frac{v_{out}}{m_2 g_{m_2} r_{d_2}} \ll v_{out}$$

Abbiamo dimostrato che le variazioni v_{ds_1} sono molto attenuate (di un fattore 100) rispetto alle variazioni della v_{out} . Ciò significa che, essendo

$$v_{out} = v_{ds_1} + v_{ds_2} \rightarrow v_{out} \cong v_{ds_2}$$

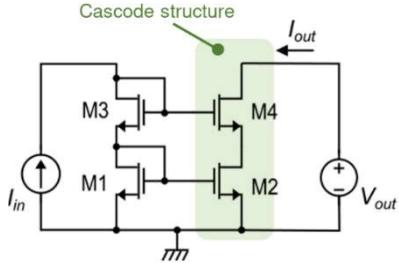
è il M2 ad assorbire le variazioni della tensione di uscita sulla sua V_{DS2} . Il transistore M1 rimane nelle condizioni impostate dal bias di M2: V_{GS2} , che ha una parte poco progettabile di soglia (con effetto body), e la parte di overdrive determinabile impostando le dimensioni e V_k , con cui maggiormente si fissa la polarizzazione della struttura cascode. Il limite inferiore della V_{out} è quello per cui M2 entra in triodo. Quando succede questo, per $V_{DS_2} = V_{DS_{2sat}}$, g_{m_2} ed r_{d_2} crollano, per cui crolla la protezione di M2 nei confronti della V_{DS_1} . Se chiamiamo la V_{DS_1} di riposo, quella impostata da M2 a funzionamento normale, $V_{DS_{1Q}}$, si deduce che

$$V_{min} = V_{DS_{2sat}} + V_{DS_{1Q}}$$



Appena M2 entra in triodo, è vero che le variazioni di V_{out} si ripercuotono significativamente sulla V_{DS_1} , ma M1 è ancora in saturazione, per cui la corrente non risente esageratamente della v_{ds_1} . Soltanto quando anche M1 entra in triodo, con $V_{DS_1} = V_{DS_{sat_1}}$, si ha un crollo delle prestazioni.

Specchio cascode standard



La corrente che scorre sul ramo di sinistra, a prescindere dalla parte destra che non ha effetto caricante, produce due potenziali rispettivamente sui gate di M3 e M1. Consideriamo tutti i body collegati al substrato, cioè a ground:

$$V_{B_1} = V_{B_2} = V_{B_3} = V_{B_4} = sub \equiv gnd$$

$$I_{out} = I_{D_4} = I_{D_2} = \beta_2 f[(V_{GS} - V_t)_2, V_{DS_2}] \quad I_{in} = I_{D_3} = I_{D_1} = \beta_1 f[(V_{GS} - V_t)_1, V_{DS_2}]$$

$$V_{GS_1} = V_{GS_2}, V_{BS_1} = V_{BS_2} \rightarrow V_{t_1} = V_{t_2} \rightarrow (V_{GS} - V_t)_1 = (V_{GS} - V_t)_2$$

$$hp: V_{DS_1} = V_{DS_2} \rightarrow \frac{I_{out}}{I_{in}} = \frac{\beta_2}{\beta_1} = k_M$$

Il rapporto delle correnti è impostato dai transistori di sotto. M2 gioca il ruolo del source comune, per cui imposta la corrente ed M4, che funziona da gate comune, la trasferisce in uscita adattando la r_{out} . Il transistore M3 ha la funzione di fornire la tensione di polarizzazione a M4 (che dipende da V_{in}). Nello specchio semplice la condizione sulle V_{DS} era valida soltanto nel caso in cui $V_{out} = V_{in}$. In questo caso, invece, la condizione è possibile da mantenersi con buona approssimazione su una larga dinamica di V_{out} . Infatti, la V_{DS_1} è fissata dalla I_{in} , e se tramite M4 si riesce a progettare la $V_{DS_2} = V_{DS_1}$, M2, protetto da M4 nella struttura cascode, mantiene quella stessa V_{DS_2} pressoché costante al variare della V_{out} . In questo specchio, quindi, la condizione per cui il rapporto tra le correnti è quello nominale viene mantenuta con buona approssimazione al variare della V_{out} . Anche qui si avrà matematicamente una sola $V_{out-opt}$, ma finché $V_{out} > V_{min}$ si ha l'ottimo in tutta la dinamica di corretto funzionamento. Cerchiamo di progettare V_{DS_2} in modo che sia pari a V_{DS_1}

$$V_{DS_2} = V_{G_4} - V_{GS_4}, V_{DS_1} = V_{G_3} - V_{GS_3}$$

Dato che $V_{G_4} = V_{G_3}$, per far sì che $V_{DS_1} = V_{DS_2} \Leftrightarrow V_{GS_3} = V_{GS_4} \rightarrow V_{t_3} + (V_{GS} - V_t)_3 = V_{t_4} + (V_{GS} - V_t)_4$

Dunque, per fare in modo da avere le V_{DS} uguali tra M1 ed M2 occorre rendere uguali le V_{GS} di M3 ed M4. Il problema si può scomporre in due: rendere uguali le tensioni di soglia, rendere uguali gli overdrive. I source di M3 e M4 non sono a comune, per cui le V_{BS} potrebbero essere diverse e dar luogo a tensioni di soglia diverse. Iniziamo dall'uguaglianza tra gli overdrive:

$$V_{OV_3} = V_{OV_4}$$

Occorre applicare una forzatura, accettabile se i transistori sono in saturazione:

$$I_{D_4} = \beta_4 f[(V_{GS} - V_t)_4, V_{DS_4}] \cong \beta_4 f[(V_{GS} - V_t)_4] \rightarrow f[(V_{GS} - V_t)_4] \cong \frac{I_{D_4}}{\beta_4}$$

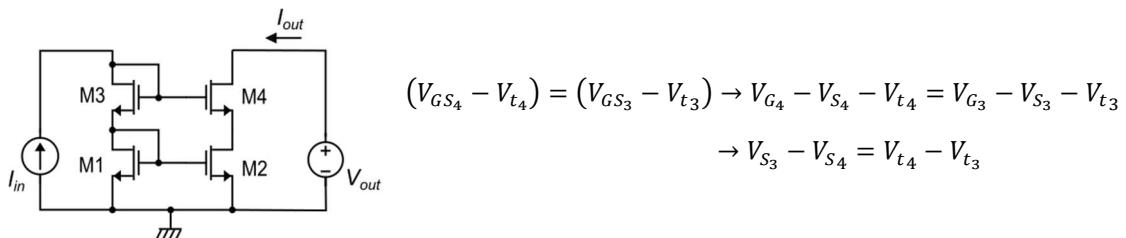
$$I_{D_3} = \beta_3 f[(V_{GS} - V_t)_3, V_{DS_3}] \cong \beta_3 f[(V_{GS} - V_t)_3] \rightarrow f[(V_{GS} - V_t)_3] \cong \frac{I_{D_3}}{\beta_3}$$

Dato che cercheremo di progettare le cose imponendo $f_4 = f_3$, rigorosamente dovremmo avere V_{DS_3} pari a V_{DS_4} per ricavare l'uguaglianza tra gli overdrive. Ma se i transistori operano in saturazione la dipendenza della f da V_{DS} è debole e due V_{DS} simili non hanno grande potere di rendere diverse le f . Allora, a ritroso, possiamo dimenticarci del loro effetto sulle singole funzioni. Dunque, approssimativamente, per rendere uguali tra loro gli overdrive si devono rendere uguali tra loro le funzioni degli overdrive (monotone):

$$V_{GS_3} = V_{GS_4} \Leftrightarrow f[(V_{GS} - V_t)_3] = f[(V_{GS} - V_t)_4] \rightarrow \frac{I_{D_4}}{\beta_4} = \frac{I_{D_3}}{\beta_3} \rightarrow \frac{\beta_4}{\beta_3} = \frac{I_{D_4}}{I_{D_3}} = \frac{I_{D_2}}{I_{D_1}} = \frac{I_{out}}{I_{in}} \rightarrow \frac{\beta_4}{\beta_3} = \frac{\beta_2}{\beta_1} = k_M$$

Dunque, con β_2/β_1 si imposta il guadagno di corrente dello specchio k_M ; sono M1 e M2 con le loro V_{GS} a far sì che le correnti stiano in rapporto dei beta impostando. Per far sì che lo specchio sia anche preciso ($V_{DS_1} = V_{DS_2}$) si fa in modo che $\beta_4/\beta_3 = \beta_2/\beta_1$; M3 e M4 conducono le medesime correnti impostando le V_{DS} dei transistori sotto. Se non si rispetta tale progettazione lo specchio cascode ha un'accuratezza inferiore e potrebbe non avere una $V_{out-opt}$ (quella tale per cui si ha, oltre che $V_{DS_1} = V_{DS_2}$, anche $V_{DS_4} = V_{DS_3}$).

Dobbiamo anche garantire che $V_{t_3} = V_{t_4}$.



Procediamo ora con una dimostrazione per assurdo.

- Ipotizzando che $V_{S_3} > V_{S_4}$ si otterrebbe $V_{t_4} > V_{t_3}$. Poiché $V_{B_3} = V_{B_4}$, se $V_{S_3} > V_{S_4}$ si avrebbe $V_{BS_3} > V_{BS_4}$ e questo implicherebbe $V_{t_3} > V_{t_4}$, il che nega l'ipotesi di partenza.
 - Ipotizzando che $V_{S_3} < V_{S_4}$ si otterrebbe $V_{t_4} < V_{t_3}$. Poiché $V_{B_3} = V_{B_4}$, se $V_{S_3} < V_{S_4}$ si avrebbe $V_{BS_3} < V_{BS_4}$ e questo implicherebbe $V_{t_3} < V_{t_4}$, il che nega l'ipotesi di partenza.

Dunque, si dimostra per assurdo che $V_{t_3} = V_{t_4}$.

Studiamo i parametri dello specchio cascode; a confronto, sulla destra, i parametri dello specchio semplice:

$$V_{in} = V_{GS_1} + V_{GS_3} \cong 2V_{GS}$$

$$V_{in} = V_{GS}$$

$$V_{min} = V_{DS_2} + V_{DS_{sat4}} = V_{GS_1} + V_{DS_{sat4}}$$

$$V_{min} = V_{DSsat}$$

$$R_{out} \cong r_{d_4}(g_{m_4}r_{d_2}) = \frac{1}{\lambda_{4I_{out}}}(g_{m_4}r_{d_2})$$

$$R_{out} = \frac{1}{\lambda I_{out}}$$

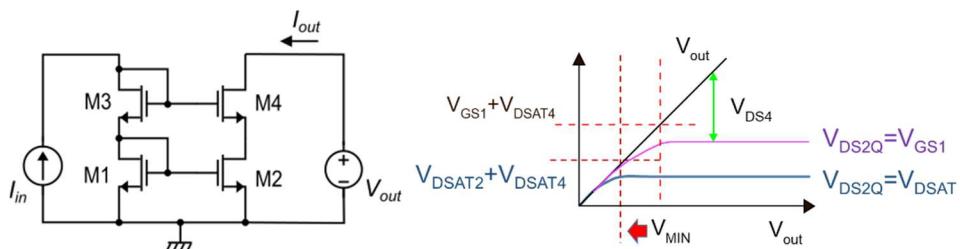
$$V_{th} = R_{out} I_{out} = \frac{1}{\lambda_4} (g_{m_4} r_{d_2})$$

$$V_{th} = \frac{1}{\lambda}$$

La minima tensione di alimentazione sarà maggiore per lo specchio cascode standard rispetto allo specchio semplice. La V_{min} , considerando che V_{GS} contiene al suo interno la tensione di soglia, subisce un notevole peggioramento. Dal punto di vista dinamico, invece, si ha un miglioramento di un fattore g_{mrq} .

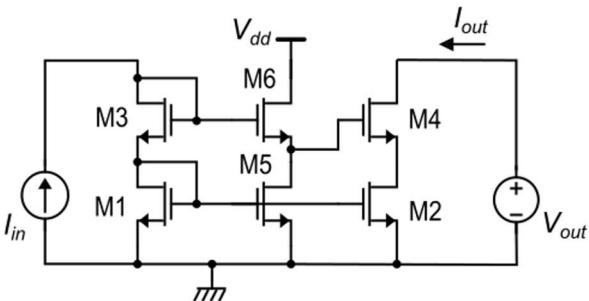
Specchio cascode a larga dinamica

Cerchiamo di sviluppare una configurazione cascode con una V_{min} migliore di quella che offre la configurazione cascode standard.



Quando la V_{out} raggiunge la $V_{DS_2} + V_{DS_{sat_4}} = V_{GS_1} + V_{DS_{sat_4}}$ lo specchio esce dalla zona di funzionamento ottimo. Non potendo migliorare la $V_{DS_{sat_4}}$ abbassandola oltre i 100 mV, la chiave per migliorare al V_{min} è quella di far sì che la V_{DS_2} non sia più pari alla V_{GS_1} e che, in particolare, possa essere resa più piccola fino a $V_{DS_{sat_2}}$. In questo modo, la V_{min} sarà nell'ordine di $2V_{DS_{sat}}$, cioè 200 mV contro i 700 mV della soluzione precedente.

Quando lo specchio non importa che sia preciso, si può utilizzare la seguente configurazione:



Si nota la presenza di un ramo ausiliario. Come prima, le correnti di ingresso e di uscita sono date da M1 e M2.

L'obiettivo è quello di fissare la V_{DS_2} non a V_{GS_1} , ma a una tensione minore. Nella struttura cascode a determinare maggiormente la V_{DS_2} è la V_k . Allora, invece di attaccare il gate di M4 direttamente a M3 in modo che $V_{G_3} = V_{G_4} \rightarrow V_{DS_2} = V_{DS_1}$, si attacca al drain di M5, il quale si comporta da traslatore di livello abbassando la V_{G_3} di V_{GS_3} .

$$V_{DS_2} = V_{GS_1} + V_{GS_3} - V_{GS_6} - V_{GS_4}$$

Non vogliamo che la V_{DS_2} sia nulla, per cui si fa sì che una di queste V_{GS} sia un po' più grande delle altre, che invece sono circa uguali. In particolare, le $V_{GS_1}, V_{GS_6}, V_{GS_4}$ saranno circa uguali, mentre la V_{GS_3} sarà maggiorata quanto basta per avere $V_{DS_2} = V_{DS_{sat_2}}$.

$$V_{DS_2} = (V_{GS} - V_t)_1 + (V_{GS} - V_t)_3 - (V_{GS} - V_t)_6 - (V_{GS} - V_t)_4 + V_{t_1} + V_{t_3} - V_{t_6} - V_{t_4}$$

$$V_{t_1} + V_{t_3} - V_{t_6} - V_{t_4} \cong 0 \rightarrow V_{DS_2} = (V_{GS} - V_t)_1 + (V_{GS} - V_t)_3 - (V_{GS} - V_t)_6 - (V_{GS} - V_t)_4$$

In forte inversione:

$$V_{DS_2} = (V_{GS} - V_t)_1 + (V_{GS} - V_t)_3 - (V_{GS} - V_t)_6 - (V_{GS} - V_t)_4 = \sqrt{\frac{2I_{D_1}}{\beta_1}} + \sqrt{\frac{2I_{D_3}}{\beta_3}} - \sqrt{\frac{2I_{D_6}}{\beta_6}} - \sqrt{\frac{2I_{D_4}}{\beta_4}}$$

Poiché $I_{D_4} = I_{out} = k_M I_{in} = k_M I_{D_1}$, per far sì che l'overdrive del M4 sia lo stesso del M1 si rende $\beta_4 = k_M \beta_1$. Poi, per far sì che l'overdrive del M6 sia uguale a quello di M1 si può scegliere, ad esempio, $I_{D_6} = I_{D_1}$ e $\beta_6 = \beta_1$. Infine, per far sì che la V_{DS_2} sia circa una $V_{DS_{sat}}$, si fa in modo che l'overdrive del M3 sia il doppio più grande rispetto agli altri. Considerando che $I_{D_3} = I_{D_1}$, si sceglie $\beta_3 = \beta_1/4$.

$$V_{DS_2} = \sqrt{\frac{2I_{D_1}}{\beta_1}} + \sqrt{4\frac{2I_{D_1}}{\beta_1}} - \sqrt{\frac{2I_{D_1}}{\beta_1}} - \sqrt{\frac{2k_M I_{D_1}}{k_M \beta_1}} = \sqrt{\frac{2I_{D_1}}{\beta_1}} = (V_{GS_1} - V_t) = V_{DS_{sat2}}$$

Se la corrente in ingresso fosse relativamente grande, si può risparmiare sulla I_{D_6} rendendola più piccola compensando con un β_6 più piccolo. Il MOSFET M2 ha una V_{DS} di riposo al limite del triodo, pari a $V_{DS_{sat2}}$. Quando M4 entra in triodo, la tensione minima dello specchio V_{min} è data da $\cong 2V_{DS_{sat}}$. Nella pratica non si imposta la V_{DS_2} al limite del triodo in quanto la V_{DS_2} cambia con V_{out} , per cui la $V_{DS_{2Q}}$ (di riposo) si farà leggermente maggiore con l'aiuto del simulatore. Non solo, se si tiene conto dell'effetto Body si ottiene che l'overdrive del transistore M3 deve essere reso ancora maggiore.

Il limite di questa soluzione, a parte il maggiore consumo, è che non valendo più la condizione $V_{DS_1} = V_{DS_2}$, il rapporto tra le correnti non è più accuratamente pari al rapporto geometrico; adesso la V_{DS_2} è sistematicamente più piccola della V_{DS_1} (M1 montato a diodo $\rightarrow V_{DS_1} = V_{GS_1}$).

$$\frac{I_{out}}{I_{in}} = \frac{I_{D_2}}{I_{D_1}} = \frac{\beta_2}{\beta_1} \cdot \frac{f((V_{GS} - V_t)_2, V_{DS_2})}{f((V_{GS} - V_t)_1, V_{DS_1})}$$

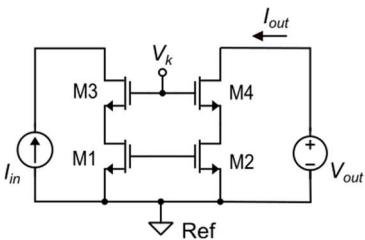
Mentre gli overdrive di M2 e M1 sono uguali e i rapporti dei β sono quelli giusti la $V_{DS_1} = V_{GS_1} > V_{DS_2} = V_{DS_{sat_2}}$, per cui il rapporto tra le funzioni, che sono monotone rispetto alla V_{DS} , è sistematicamente < 1 . Di conseguenza, sistematicamente, si avrà:

$$\frac{I_{out}}{I_{in}} < \frac{\beta_2}{\beta_1} @ k_M$$

Lo specchio che ne risulta è buono in termini di R_{out} e di V_{min} , ma non è uno specchio preciso; non esiste alcun valore ottimale della V_{out} tale per cui la $I_{out} = k_M I_{in}$. Se si considera una coppia differenziale, a meno di non avere un circuito particolare, un errore anche del 10% sulla corrente di polarizzazione non ha alcun impatto significativo sull'amplificatore. Per cui, in tal caso, questo specchio è una valida scelta. (offre comunque una corrente costante in uscita stabile al variare della V_{out}).

Specchio cascode ad alta precisione e ampia dinamica

Si può fare di meglio. L'obiettivo è quello di ottenere elevata resistenza di uscita, V_{min} piccola nell'ordine di $2V_{DS_{sat}}$ e $V_{DS_2} = V_{DS_1}$ in modo che lo specchio sia anche accurato.



La configurazione è simile a quella dello specchio semplice, ma M1 ed M3 non sono montati a diodo. La tensione di polarizzazione del cascode V_k deve essere prodotta da un circuito ausiliario. Si ha ancora $V_{GS_1} = V_{GS_2}$; se si ha anche $V_{DS_1} = V_{DS_2}$ il rapporto delle correnti è uguale a quello nominale. Si ottiene cioè la stessa regola di progetto che avevamo visto per il cascode semplice.

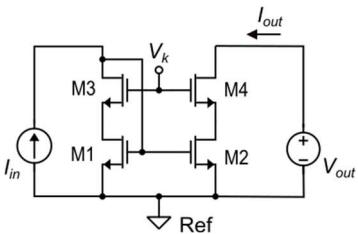
Senza laccio, giocando su V_k si possono fare V_{DS_1} e V_{DS_2} piccole, con un certo margine rispetto alla $V_{DS_{sat}}$ (nello specchio semplice M1 era montato a diodo, per cui aveva $V_{DS_1} = V_{GS_1}$). La V_{min} sarà pari a $V_{DS_{sat_2}} + V_{DS_{sat_4}} +$ margine di sicurezza.

$$V_{DS_1} = V_k - V_{GS_3} \quad V_{DS_2} = V_k - V_{GS_4}$$

La condizione $V_{DS_1} = V_{DS_2}$, come per lo specchio cascode standard, si ottiene per la stessa regola di progetto:

$$V_{GS_3} = V_{GS_4} \rightarrow \frac{\beta_4}{\beta_3} = \frac{\beta_2}{\beta_1}$$

Il problema è che nonostante $V_{GS_1} = V_{GS_2}$, i gate di M1 e M2 sono flottanti. Il gate di M1 deve essere al potenziale tale che M1 porti la corrente di ingresso. Si risolve con un collegamento in più:



Immaginiamo che il circuito sia spento con correnti nulle e capacità parassite scariche. Accendendo la corrente di ingresso a gradino, per inerzia delle capacità i potenziali ai nodi rimangono nulli. Se la $V_{GS_1} = 0 \rightarrow I_{D_1} = 0$, per cui anche $I_{D_3} = 0$. La corrente di ingresso, quindi, inizialmente entra nelle capacità caricandole. Si arriva all'equilibrio nel momento in cui le capacità parassite non entrano corrente e i potenziali dei nodi sono tali per cui $I_{D_1} = I_{D_3} = I_{in}$.

Dunque, il laccio fa sì che il potenziale del gate di M1 all'equilibrio sia tale proprio da far scorrere la corrente di ingresso attraverso il canale. Lo stesso meccanismo varrebbe per un MOSFET montato a diodo.

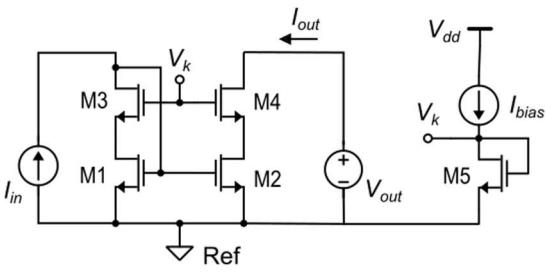
Una volta che le correnti e le dimensioni dei transistori sono fissate, sono fissate anche V_{GS_3} e V_{GS_4} , uguali tra loro. A questo punto, le V_{DS} dei transistori M1 ed M2 sono modulabili attraverso la tensione V_k . Immaginiamo, quindi, di voler fissare le $V_{DS_1} = V_{DS_2}$ a un valore a piacere tale da mantenere M1 e M2 in saturazione.

$$V_{GS_4} = V_{t_4} + (V_{GS} - V_t)_4 \cong V_{t_0} + (m_4 - 1)V_{SB_4} + (V_{GS} - V_t)_4 = V_{t_0} + (m_4 - 1)V_{S_4} - (m_4 - 1)V_{B_4} + (V_{GS} - V_t)_4$$

Per l'aumento della tensione di soglia con la tensione V_{SB} si fa un'approssimazione al prim'ordine. Immaginiamo che tutti i substrati siano allo stesso potenziale V_B . Possiamo procedere espandendo l'espressione di V_{DS_2} :

$$\begin{aligned} V_{DS_2} &= V_k - V_{GS_4} = V_k - V_{t_0} - (m_4 - 1)V_{S_4} + (m_4 - 1)V_B - (V_{GS} - V_t)_4 \\ V_{S_4} &= V_{DS_2} \rightarrow V_{DS_2} = V_k - V_{t_0} - (m_4 - 1)V_{DS_2} + (m_4 - 1)V_B - (V_{GS} - V_t)_4 \rightarrow \\ m_4 V_{DS_2} + V_{t_0} + (V_{GS} - V_t)_4 - (m_4 - 1)V_B &= V_k \end{aligned}$$

Da questa equazione è possibile ricavare la tensione di polarizzazione V_k che permette di ottenere una certa V_{DS_2} nel momento in cui si conosce l'overdrive di M4 e il potenziale del substrato. Gli overdrive solitamente sono determinati antecedentemente per rispettare altri criteri, il potenziale del substrato rispetto al riferimento si conosce, la tensione di soglia V_{t_0} è nota, per cui diventa semplice determinare V_k . In particolare, gli overdrive dei transistori M1 ed M2 non potranno essere schiacciati minimo poiché, essendo questi i transistori che impostano le correnti, sono soggetti a più compromessi. Invece gli overdrive di M3 ed M4, che fissano le V_{DS} dei transistori di sotto ritoccandone appena la corrente, saranno minimizzati in modo da minimizzare le $V_{DS_{sat}}$ allargando così la dinamica dello specchio. Per generare la tensione V_k :



La tecnica è quella di far scorrere una corrente costante in un MOSFET montato a diodo prelevando la tensione V_{GS} . È fondamentale che il source di M5 sia connesso a V_{ref} in modo che la sua V_{GS} sia V_k riferita a V_{ref} .

$$\begin{aligned} V_{GS_5} &= V_{t_5} + (V_{GS} - V_t)_5 \cong V_{t_0} + (m_5 - 1)V_{SB_5} + (V_{GS_5} - V_t)_5 \\ &\rightarrow V_{GS_5} \cong V_{t_0} - (m_5 - 1)V_{B_5} + (V_{GS_5} - V_t)_5 \end{aligned}$$

Possiamo ora eguagliare l'espressione della tensione V_k in funzione della V_{DS_2} ottenuta precedentemente alla V_{GS_5} appena determinata.

$$V_k = V_{GS_5} \rightarrow m_4 V_{DS_2} + V_{t_0} + (V_{GS} - V_t)_4 - (m_4 - 1)V_B = V_{t_0} + (m_5 - 1)V_B + (V_{GS_5} - V_t)_5$$

L'unica incognita è l'overdrive di M5. La parte V_{t_0} che figura nelle tensioni di soglia di M4 e di M5 è la stessa, per cui si cancella. Approssimando $m_5 \cong m_4$ (il coefficiente m dei transistori dipende, tra le altre cose, dal grado di inversione; i transistori M4 ed M5, da questo punto di vista, operano in maniera differente) si cancellano anche i termini dipendenti da V_B . In conclusione, si ottiene:

$$(V_{GS} - V_t)_5 = m_4 V_{DS_2} + (V_{GS} - V_t)_4$$

Tale espressione definisce un obiettivo progettuale: dimensionare il transistore M5 in termini di overdrive in modo che la V_k che genera polarizzi lo specchio cascode con una certa $V_{DS_2} = V_{DS_1}$ scelta. Operando M5 in forte inversione, l'overdrive dipenderà dalla corrente e dal β_5 ; poiché le incognite sono due si può scegliere arbitrariamente, ad esempio, la corrente. A quel punto l'unica incognita è il β_5 .

$$V_{min} = V_{DS_2} + V_{DS_{sat4}}$$

Il meglio che si può fare è rendere $V_{DS_1} = V_{DS_2} = V_{DS_{sat}}$, in modo che la $V_{min} = 2V_{DS_{sat}} (\cong 200 \text{ mV})$. Se si riesce, lo specchio è sia accurato che ad ampia dinamica. La minima V_{DS} si può ottenere facendo l'overdrive pari a $4V_T$ al limite inferiore della forte inversione oppure andando ad operare in moderata o debole inversione. Scegliendo $V_{DS_2} = V_{DS_{sat2}}$, in forte inversione si ottiene:

$$(V_{GS} - V_t)_5 = m_4 V_{DS_{sat2}} + (V_{GS} - V_t)_4 \rightarrow (V_{GS} - V_t)_5 = m_4 (V_{GS} - V_t)_2 + (V_{GS} - V_t)_4$$

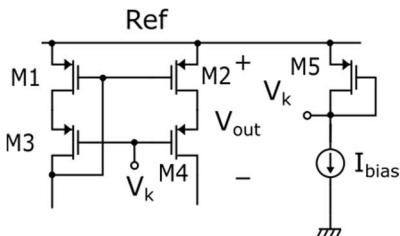
L'overdrive del transistore ausiliario M5 deve essere la somma, in sostanza, degli overdrive di M2 e M4.

Per questo specchio si ha, come per lo specchio semplice:

$$V_{in} = V_{GS}$$

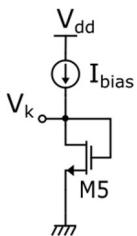
La R_{out} è grande nei limiti in cui M2 e M4 sono in saturazione ($R_{out} = r_{d_2}(g_{m_4}r_{d_4})$). Poiché $V_{DS_1} = V_{DS_2}$ questa configurazione di specchio cascode è anche di precisione. L'unico peggioramento è sulla V_{min} , ma è contenuto e nell'ordine di 100 mV. Poi, nella realtà, diminuire la V_{DS_2} verso il limite della saturazione degrada gradualmente la resistenza di uscita, la quale, però, sarà sempre comunque molto maggiore di quella di uno specchio semplice. Si tratta quindi di un ottimo specchio di corrente, il cascode più utilizzato in assoluto nei circuiti moderni.

Si potrebbe realizzare, con la stessa configurazione, uno specchio di precisione a larga dinamica di tipo p.



In questo caso la V_k controlla ancora le V_{DS} dei transistori M1 ed M2, ma aumentando ne causa una diminuzione. Il modo corretto per generare la V_k riferendola alla V_{ref} con un transistore montato a diodo è quello di collegare il source dello stesso al riferimento. Se la corrente I_{bias} cambia, si ha un effetto su V_{DS_1} e V_{DS_2} , le quali però rimangono uguali tra loro, per cui lo specchio continua ad operare nel modo corretto.

Uno dei modi sbagliati di generare la V_k in questo circuito è il seguente:

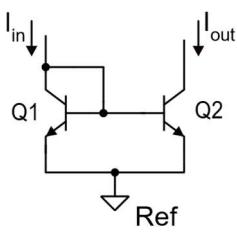


In questo caso, anzitutto non si può semplificare il termine V_{t_0} nelle espressioni poiché è significativamente diverso tra transistori di tipo p e transistori di tipo n. Il fatto grave è che in questo caso la V_k non sarebbe più riferita alla tensione di riferimento V_{ref} . Immaginando che la $V_{ref} = V_{dd}$, se cambia la tensione di alimentazione la V_k è costante rispetto a ground e quindi non rispetto alla V_{dd} . Se quindi si impiegasse questa soluzione, il circuito complessivo funzionerebbe nella maniera prevista per una sola tensione di alimentazione stabile.

Per far sì che la V_k sia invariante rispetto al riferimento, alla V_{dd} in questo caso, la soluzione corretta è quella di sopra. In conclusione, questa configurazione di specchio cascode è buona. L'unico difetto è che è più difficile da progettare poiché richiede la generazione di V_k . Inoltre, mentre gli specchi di prima sono molto tolleranti rispetto alla dinamica della corrente in ingresso, in questo caso non è così. Se la V_k è costante, all'aumentare di I_{in} le V_{GS} di M3 ed M4 aumentano, per cui le V_{DS} di M1 e M2 diminuiscono. Aumentando troppo la I_{in} oltre un certo limite il circuito smette di funzionare nella maniera corretta. **Nello specchio tradizionale standard anche la tensione del gate M4 variava, decresceva adattandosi. Si può fare adattiva la V_k , ci saranno le parti n, le parti p. Però il range di tensioni è migliore rispetto a quello tradizionale.**

Specchi a bipolar

Specchio semplice



Per l'espressione delle correnti si utilizza il modello semplificato di Ebers-Moll, includendo un termine correttivo funzione della V_{CB} per l'effetto Early.

$$I_{C_2} = I_{S_2} e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB_2}}{V_A} \right) = I_{S_2} e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CE_2}}{V_A} - \frac{V_{BE}}{V_A} \right)$$

$$I_{C_1} = I_{S_1} e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB_1}}{V_A} \right) = I_{S_1} e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CE_1}}{V_A} - \frac{V_{BE}}{V_A} \right)$$

Ad impostare la corrente in modo importante è la tensione V_{BE} ; la V_{CE} applica un solo ritocco. Per questo specchio la corrente di uscita I_{out} coincide con la I_{C_2} , ma la corrente di ingresso $I_{in} = I_{C_1} + I_{B_1} + I_{B_2}$. Dunque, la corrente di ingresso, rispetto alla versione a MOSFET, ha anche una parte che serve a polarizzare i dispositivi. Per il momento si semplifica l'analisi trascurando le correnti di base. Se $V_{CE_1} = V_{CE_2}$ il rapporto tra le correnti di collettore è pari al k_M nominale.

$$k_M = \frac{I_{S_2}}{I_{S_1}} = \frac{A_{E_2}}{A_{E_1}} = \frac{\text{area}_2}{\text{area}_1}$$

Il guadagno dello specchio dipende dal rapporto delle aree. Si può esprimere il rapporto o in termini delle aree di emettitore, vere aree, o in termini del parametro adimensionale area (se l design kit lo consente). Quest'ultimo non rappresenta un parametro fisico del dispositivo, ma è utilizzato dal CAD per personalizzare l'inserimento di un BJT; in particolare, il parametro area indica quante volte è più “grande” il dispositivo che si inserisce rispetto a quello elementare della medesima famiglia. Ad esempio, $\text{area} = 5 \rightarrow I_S = 5I_{S_{BJT_{elem}}}$.

Poiché:

$$V_{out} = V_{CE_2} \quad V_{in} = V_{CE_1}$$

La stessa condizione per ottenere il rapporto nominale k_M tra le correnti di collettore si può esprimere egualando $V_{out} = V_{in}$. Esiste cioè una sorta di $V_{out-opt}$, anche se in questo caso più un analogo in quanto il rapporto tra le correnti di collettore, sempre a causa delle correnti di base, non è pari al rapporto I_{out}/I_{in} .

$$V_{out} = V_{in} \rightarrow \frac{I_{C_2}}{I_{C_1}} = k_M$$

La corrente di uscita è indipendente dalla tensione di uscita quando Q2 opera in ZAD. Per $V_{out} \leq V_{CE_{sat_2}}$ il transistore è in saturazione e la corrente dipende fortemente dalla V_{CE} . Dunque:

$$V_{min} = V_{CE_{sa_2}} \cong 200 \text{ mV}$$

Per quanto riguarda la tensione di ingresso:

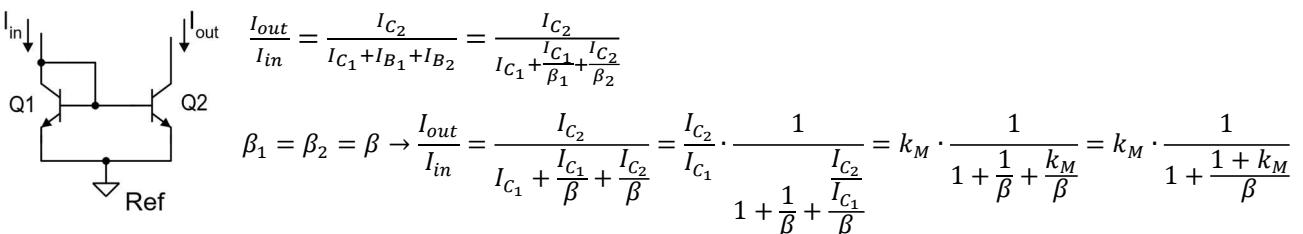
$$V_{in} = V_{CE_1} = V_{BE_1} \cong V_\gamma$$

La resistenza di uscita è pari alla r_o del Q2:

$$R_{out} = r_{o_2} = \frac{V_A}{I_{C_2}} = \frac{V_A}{I_{out}}$$

$$V_{th} = R_{out} I_{out} = V_A$$

Consideriamo adesso l'impatto delle correnti di base sul rapporto tra la corrente di uscita e la corrente di ingresso.



Come si osserva, il k_M reale è dato da quello di progetto affatto da un certo termine di errore sistematicamente minore dell'unità. Applicando l'espansione in serie di Taylor al prim'ordine:

$$\frac{1}{1+x} \approx 1-x \rightarrow \frac{I_{out}}{I_{in}} \cong k_M \left(1 - \frac{k_M+1}{\beta} \right)$$

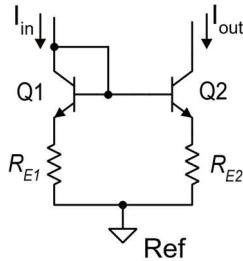
L'espressione è valida se $x = \frac{k_M+1}{\beta} \ll 1$. Si ottiene, dunque, che l'errore relativo che si commette rispetto il k_M nominale è pari a:

$$|\epsilon_R| \cong \frac{k_M+1}{\beta}$$

A parità di β dei transistori, specchi progettati per specchiare con alto guadagno soffrono di un errore relativo maggiore.

Specchio semplice con degenerazione di emettitore

Come per lo specchio semplice a MOSFET è possibile cercare di aumentare la resistenza di uscita applicando degenerazione agli emettitori di Q1 e Q2. In alcuni casi la resistenza di uscita r_{o_2} è troppo bassa.



Per ottenere $\frac{I_{C_2}}{I_{C_1}} = \frac{I_{S_2}}{I_{S_1}} = k_M$, oltre che a rendere $V_{CE_1} = V_{CE_2}$ (che poi implica $V_{in} = V_{out}$), occorre far sì che le V_{BE} dei due transistori siano uguali:

$$V_{BE_1} = V_{BE_2} \rightarrow I_{E_1}R_{E_1} = I_{E_2}R_{E_2} \rightarrow \frac{R_{E_2}}{R_{E_1}} = \frac{I_{E_1}}{I_{E_2}} = \frac{I_{CE_1}}{I_{CE_2}} = \frac{1}{k_M}$$

Per avere la stessa caduta occorre che le resistenze siano tra loro nel rapporto delle correnti di collettore.

Per il calcolo della resistenza di uscita, mentre per il MOSFET si ha un'unica formula valida, nel caso del BJT si ha una scrittura più complicata che si divide in due casi limiti

1. $R_E \ll hie$
2. $R_E \gg hie$

Per progetto si utilizzeranno R_E non troppo grandi onde evitare cadute esagerate, per cui vale la prima condizione. Dunque, per la resistenza di uscita si ottiene:

$$R_{out} \cong r_{o_2} \left(1 + g_{m_2} R_{E_2} \right) \cong r_{o_2} \left(1 + \frac{I_{C_2} R_{E_2}}{V_T} \right)$$

La formula dovrebbe in realtà contenere il g_{meq} , il quale però dipendeva dal rapporto tra l' hie_2 e la resistenza che si vede dalla base verso massa. Quando tale resistenza è trascurabile rispetto all' hie , come in questo caso, si può approssimare il g_{meq} con il g_m del transistore. Poiché al denominatore figura la V_T , rispetto al MOSFET si ottiene lo stesso bonus con meno caduta sulle resistenze di degenerazione (o a parità di caduta si ha un bonus maggiore). Purtroppo, mentre con i MOSFET è possibile migliorare drasticamente le cose con il cascode, per il BJT la configurazione cascode è molto più problematica e ha performance più scadenti. Ecco perché la degenerazione è molto comune per specchi a BJT ma non per specchi a MOSFET. Aumentare il bonus (aumentando la caduta sulle resistenze di degenerazione q.b.) contrasta gli altri parametri dello specchio.

$$V_{in} = V_{BE_1} + R_{E_1} I_{E_1}$$

$$V_{min} = V_{CE_{sat_2}} + R_{E_2} I_{E_2}$$

Si ha invece un miglioramento sui parametri dinamici

$$R_{out} \cong r_{o_2} \left(1 + g_{m_2} R_{E_2} \right) \cong r_{o_2} \left(1 + \frac{I_{C_2} R_{E_2}}{V_T} \right) = \frac{V_A}{I_{out}} \left(1 + \frac{I_{C_2} R_{E_2}}{V_T} \right)$$

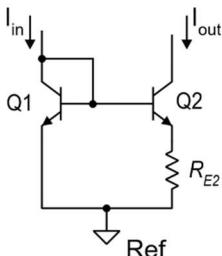
$$V_{th} = R_{out} I_{out} \cong V_A \left(1 + \frac{I_{C_2} R_{E_2}}{V_T} \right)$$

Se ad esempio si ha una caduta di 225 mV sulla resistenza R_{E_2} , il fattore di accrescimento della r_{o_2} è precisamente 10. Per quanto riguarda l'effetto delle correnti di base:

$$\frac{I_{out}}{I_{in}} = \frac{I_{C_2}}{I_{C_1} + I_{B_1} + I_{B_2}} \rightarrow |\epsilon_R| \cong \frac{k_M + 1}{\beta}$$

L'errore è lo stesso di prima. I BJT, rispetto ai MOSFET, fanno molto meno rumore flicker. Possono essere utilizzati ad esempio negli specchi di corrente in un amplificatore operazionale, con conseguente aumento dell'accuratezza.

Sorgente di corrente di Widlar



Molti circuiti, tra cui questo, si basano sulla differenza delle V_{BE} e prendono il nome di circuiti a “delta V_{BE} ”.

$$V_{BE_1} = V_{BE_2} + R_{E_2} I_{E_2}$$

$$\Delta V_{BE} = V_{BE_1} - V_{BE_2} = R_{E_2} I_{E_2} \cong R_{E_2} I_{C_2}$$

$$I_C = I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB}}{V_A} \right) \cong I_S e^{\frac{V_{BE}}{V_T}} \rightarrow V_{BE} \cong V_T \ln \left(\frac{I_C}{I_S} \right)$$

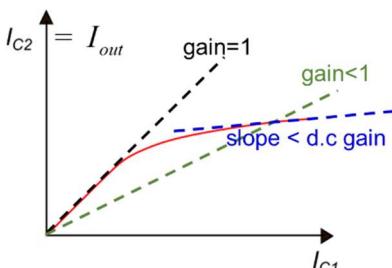
Nel calcolo della V_{BE} si trascura l'effetto Early.

$$\Delta V_{BE} \cong V_T \ln \left(\frac{I_{C_1}}{I_{S_1}} \right) - V_T \ln \left(\frac{I_{C_2}}{I_{S_2}} \right) = V_T \ln \left(\frac{I_{C_1} I_{S_2}}{I_{S_1} I_{C_2}} \right)$$

Facendo sì che le correnti di saturazione siano uguali per Q1 e Q2:

$$I_{S_1} = I_{S_2} \rightarrow \Delta V_{BE} \cong V_T \ln \left(\frac{I_{C_1}}{I_{C_2}} \right) \cong R_{E_2} I_{C_2} \rightarrow \frac{I_{C_2}}{I_{C_1}} \cong e^{-\frac{R_{E_2} I_{C_2}}{V_T}}$$

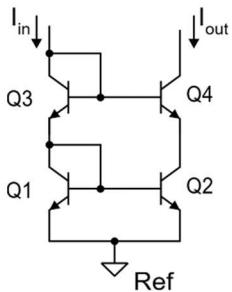
Il rapporto I_{C_2}/I_{C_1} non è più pari al rapporto delle aree, ma dipende esponenzialmente dalla corrente di uscita. Non essendoci più proporzionalità tra la corrente di uscita e la corrente di ingresso, la sorgente di corrente di Widlar sarebbe uno specchio non lineare.



Per $I_{in} = 0$ sia Q1 che Q2 sono spenti. Inizialmente il circuito si comporta linearmente poiché per I_{C_2} piccole il fattore esponenziale è circa unitario. Per tutte le correnti di ingresso tali per cui la caduta sulla resistenza R_{E_2} è $\ll V_T$ il comportamento rimane lineare. Quando la caduta diventa confrontabile e più grande della V_T la I_{C_2} cresce a un ritmo inferiore. In questo modo si può ottenere una corrente di uscita molto più piccola di quella di ingresso.

Il modo alternativo per ottenere una corrente più piccola sarebbe quello di specchiare in discesa. Supponendo di dover specchiare con coefficiente 1/10, utilizzando i rapporti si dovrebbero mettere 10 bipolari (ingombranti) in parallelo dal lato dell'ingresso. L'altro pregio è che il guadagno dinamico (la derivata delle curva dI_{out}/dI_{in}) nella regione in cui si ha demagnificazione è minore del guadagno statico (la pendenza statica della retta che collega il punto di lavoro all'origine, cioè il rapporto tra le correnti I_{out}/I_{in}). In questo modo la variazione percentuale della corrente di uscita $\Delta I_{out}/I_{out}$ è molto minore di quella della corrente di ingresso $\Delta I_{in}/I_{in}$. Questo circuito, grazie a questa proprietà, era molto utile per i primi circuiti integrati con pochi transistori per ottenere una polarizzazione quasi costante, stabile. Un modo di ottenere una corrente di polarizzazione (es. per opamp) è quello di agganciare all'alimentazione una resistenza e un MOSFET montato a diodo che fa da ingresso per uno specchio di corrente. Se però I_{in} è proporzionale all'alimentazione, nel caso di uno specchio lineare una variazione della tensione di alimentazione si riflette sulla corrente di bias. Con la sorgente di corrente di Widlar, invece, nella regione di saturazione della caratteristica all'aumentare della tensione di alimentazione la corrente di polarizzazione aumenta di poco. In effetti, tale circuito può assumere il comportamento di un generatore di corrente indipendente dalle tensioni di alimentazione (ecco perché “current source”). Ad oggi si utilizzano circuiti a bandgap, altre tecniche che con l'utilizzo della retroazione positiva permettono di ottenere correnti indipendenti dall'alimentazione.

Specchio cascode classico



Come per il MOSFET, la configurazione cascode permette di poter aumentare la resistenza di uscita senza impattare troppo sulla dinamica dell'alimentazione.

$$V_{CE_2} = V_{CE_1} + V_{BE_3} - V_{BE_4}$$

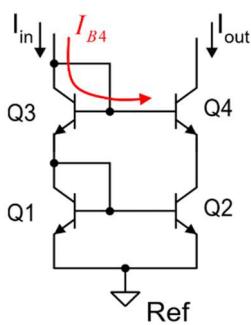
$$V_{BE_3} = V_{BE_4} \rightarrow V_{CE_2} = V_{CE_1}$$

$$V_{BE_3} = V_{BE_4} \rightarrow V_T \ln\left(\frac{I_{C_3}}{I_{S_3}}\right) = V_T \left(\frac{I_{C_4}}{I_{S_4}}\right) \rightarrow \frac{I_{C_4}}{I_{S_4}} = \frac{I_{C_3}}{I_{S_3}} \rightarrow \frac{I_{S_4}}{I_{S_3}} = \frac{I_{C_4}}{I_{C_3}} \cong \frac{I_{C_2}}{I_{C_1}} = \frac{I_{S_2}}{I_{S_1}} = k_M$$

Lo stadio sopra a gate comune deve impostare la $V_{CE_2} = V_{CE_1}$. Le variazioni della V_{out} sono assorbite da Q4 e la V_{CE_2} di Q2 rimane abbastanza costante. Tale robustezza si vede anche dall'espressione di V_{CE_2} : V_{CE_1} dipende più o meno, essendo pari a V_{BE_1} , solo dalla corrente di ingresso, V_{BE_3} e V_{BE_4} sono circa uguali a V_T e si cancellano, la V_{out} non comare. Un'indagine più approfondita mostrerebbe una piccola dipendenza delle V_{BE} da V_{out} . Dall'uguaglianza tra V_{BE_3} e V_{BE_4} si ottiene la seguente regola di design, analoga a quella del MOSFET.

$$\frac{I_{S_4}}{I_{S_3}} = \frac{I_{S_2}}{I_{S_1}} = k_M$$

Le I_S in questo caso giocano come i β dei MOSFET, poiché determinano la corrente moltiplicando una funzione della tensione di controllo. Sintetizzando, si ottiene che il fattore moltiplicativo in area di Q2 rispetto a Q1 deve essere lo stesso per Q4 rispetto a Q3. Valutiamo la resistenza di uscita:



A differenza del cascode a MOSFET, a causa delle correnti di base si ha una contaminazione tra il ramo di ingresso e il ramo di uscita. Il comportamento di Q4 influenza la corrente che scorre nel ramo di ingresso. In particolare, se con V_{out} cambia la I_{B_4} parte della corrente viene drenata da Q1; se la corrente di Q1 cambia, varia anche la tensione del transistore Q4. Si dovrebbe in effetti considerare anche il generatore comandato di Q2; la corrente che arriva a Q2 dipende anche dalla I_{B_4} . Questa struttura ha una sorta di reazione: se la corrente tendesse a diminuire, il potenziale di emettitore tenderebbe a diminuire, aumenterebbe la V_{BE} e questo compenserebbe l'effetto di variazione di corrente.

A livello fisico, diminuendo la V_{out} diminuisce la V_{CE} sul Q4, il che comporta un aumento di I_{B_4} . Se I_{B_4} aumenta, al ramo di sinistra arriva meno corrente, per cui anche Q2 viene pilotato con meno corrente. In questo caso, quindi, non si può aprire il generatore di corrente equivalente di Q2 come per un MOSFET; c'è un effetto di V_{out} sul generatore di corrente di Q2. Questo effetto, complessivamente, causa l'abbassamento della resistenza di uscita: se diminuisce corrente sul ramo di sinistra il generatore di corrente ha una variazione che contribuisce a diminuire la corrente. Si può dimostrare che:

$$R_{out} \cong r_{o_4} \left(1 + \frac{h_{fe_4}}{2}\right) = \frac{V_A}{I_{out}} \left(1 + \frac{h_{fe_4}}{2}\right)$$

Calcoliamo l'errore relativo rispetto al rapporto nominale:

$$I_{out} = I_{C_4} = I_{E_4} - I_{B_4} = I_{C_2} - I_{B_4}$$

$$I_{in} = I_{C_1} + I_{B_1} + I_{B_2} + I_{B_4}$$

$$\frac{I_{out}}{I_{in}} = \frac{I_{C_2} - I_{B_4}}{I_{C_1} + I_{B_1} + I_{B_2} + I_{B_4}} \cong \frac{I_{C_2} - \frac{I_{C_2}}{\beta}}{I_{C_1} + \frac{I_{C_1}}{\beta} + \frac{I_{C_2}}{\beta} + \frac{I_{C_4}}{\beta}} \cong \frac{I_{C_2} - \frac{I_{C_2}}{\beta}}{I_{C_1} + \frac{I_{C_1}}{\beta} + 2 \frac{I_{C_2}}{\beta}} = \frac{I_{C_2}}{I_{C_1}} \cdot \frac{1 - \frac{1}{\beta}}{1 + \frac{1}{\beta} + 2 \frac{k_M}{\beta}}$$

$$\frac{1-y}{1+x} \approx (1-y)(1-x) = 1 - (x+y) + xy \rightarrow \frac{I_{out}}{I_{in}} \cong k_M \left(1 - 2 \frac{1+k_M}{\beta}\right) \rightarrow \epsilon_R = -2 \frac{1+k_M}{\beta}$$

In conclusione:

$$V_{in} = V_{BE_1} + V_{BE_3} \cong 2V_\gamma$$

$$V_{min} = V_{BE_1} + V_{CE_{sat4}} \cong V_\gamma + V_{CE_{sat}}$$

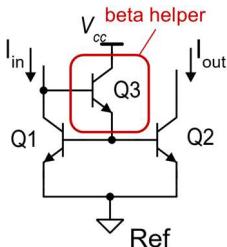
$$R_{out} \cong r_{o_4} \left(1 + \frac{h_{fe4}}{2} \right) = \frac{V_A}{I_{out}} \left(1 + \frac{h_{fe4}}{2} \right)$$

$$\epsilon_R = -2 \frac{1 + k_M}{\beta}$$

L'incremento della resistenza di uscita è appena di un fattore $h_{fe}/2$ (50 contro $g_m r_d$ che arriva anche a 200), che si paga con un incremento di V_{in} e V_{min} . L'errore relativo sul rapporto di specchio raddoppiato rispetto allo specchio semplice. A causa di questi svantaggi, sono preferite altre soluzioni quali lo specchio con degenerazione di emettitore o lo specchio di Wilson.

Beta helper

Si può migliorare l'errore dato dalle correnti di base con un beta helper che sostituisce il lacchetto.



L'ideale sarebbe togliere il lacchetto. In quel caso la corrente di ingresso sarebbe pari alla I_{C_1} , ma si perderebbe l'autopolarizzazione delle basi; la corrente di ingresso aggiusta la tensione V_{BE} in modo tale che il bipolare consumi proprio la I_{in} . Inserendo un bipolare montato a diodo, inizialmente la corrente carica la capacità associata alla base di Q3. Quando Q3 si accende, si accendono anche Q1 e Q2. Il vantaggio è che:

$$I_{B_3} = \frac{I_{C_3}}{\beta_3} \cong \frac{I_{B_1} + I_{B_2}}{\beta_3}$$

La quota di corrente che si spilla dalla corrente di ingresso per polarizzare Q1 e Q2 è abbattuta di un fattore β_3 .

$$I_{in} = I_{C_1} + I_{B_3} = I_{C_1} + \frac{I_{E_3}}{\beta_3 + 1}$$

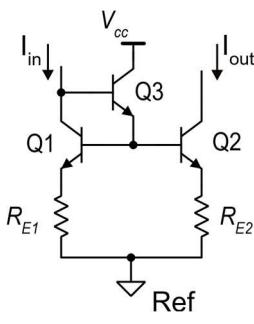
$$I_{E_3} = I_{B_1} + I_{B_2} \rightarrow I_{in} = I_{C_1} + \frac{I_{B_1} + I_{B_2}}{\beta_3 + 1} = I_{C_1} + \frac{1}{\beta_3 + 1} \left(\frac{I_{C_1}}{\beta} + \frac{k_M I_{C_1}}{\beta} \right)$$

$$\frac{I_{out}}{I_{in}} = \frac{I_{C_2}}{I_{C_1} + I_{C_1} \frac{1 + k_M}{(\beta_3 + 1)\beta}} = k_M \frac{1}{1 + \frac{1 + k_M}{(\beta_3 + 1)\beta}}$$

$$\epsilon_R = -\frac{1 + k_M}{(\beta_3 + 1)\beta} \rightarrow |\epsilon_R| \cong \frac{1}{\beta^2}$$

Per migliorare ancora si deve sfruttare la reazione. Con il beta helper la V_{min} rimane inalterata, mentre la V_{in} peggiora rispetto allo specchio semplice, il che impedisce di poter usare il circuito per basse alimentazioni.

$$V_{in} = 2V_\gamma$$



Si può combinare il beta helper alla degenerazione di emettitore. Così facendo si migliora sia la resistenza di uscita che l'errore dovuto alla corrente di base:

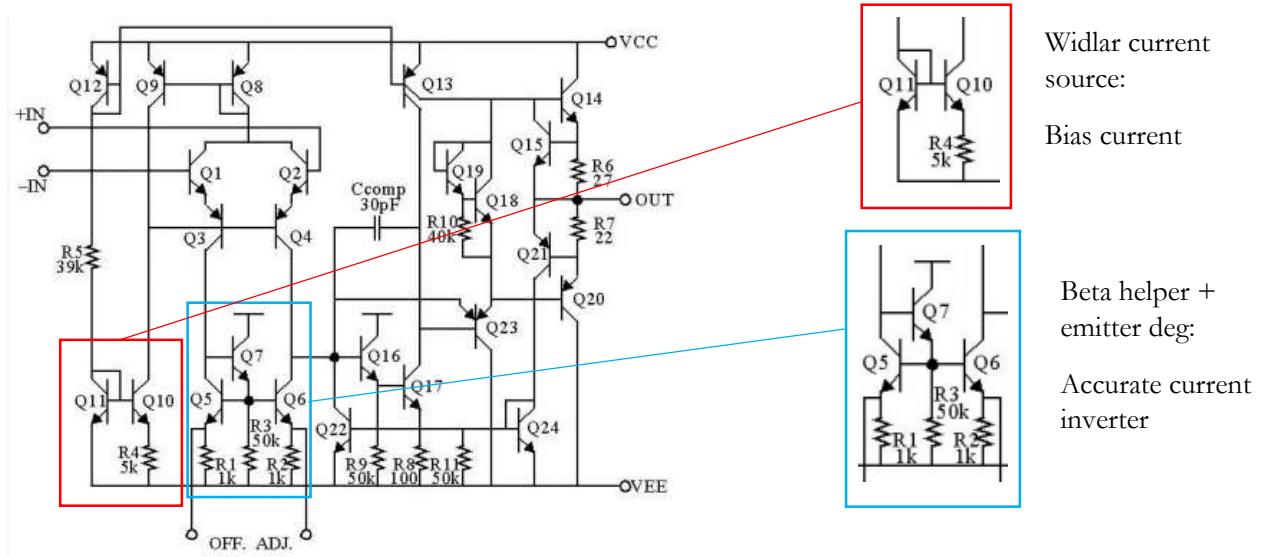
$$R_{out} \cong r_{o_2} (1 + g_{m2} R_{E_2})$$

$$V_{in} = 2V_\gamma + R_{E_1} I_{E_1}$$

$$V_{min} = V_{CE_{sat}} + R_{E_2} I_{E_2}$$

$$\epsilon_R = \frac{1 + k_M}{(\beta_3 + 1)\beta} \rightarrow |\epsilon_R| \cong \frac{1}{\beta^2}$$

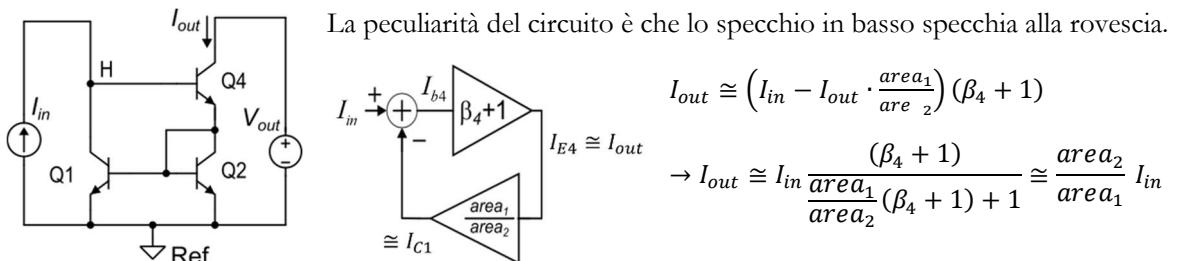
Osservando lo schema del $\mu A741$:



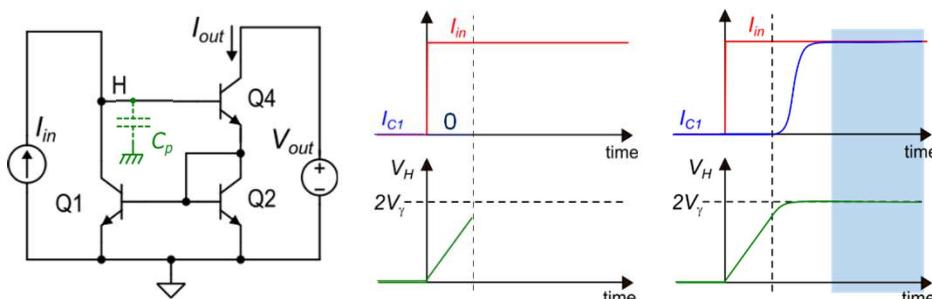
Senza R_3 il beta helper Q_7 avrebbe una corrente di collettore/emettitore pari a $I_{B_5} + I_{B_6}$, per cui rischierebbe di operare nella zona del Gummel Plot in cui il beta crolla per correnti di collettore troppo piccole. Si ha un miglioramento anche nella risposta in frequenza, con un peggioramento dell'errore dovuto alle correnti di base.

Specchio di Wilson

Lo specchio di Wilson si trova principalmente nella versione a bipolar. Si può trovare anche a MOSFET, ma non proprio come specchio di corrente. L'obiettivo è quello di ottenere uno specchio che abbia la stessa resistenza di uscita del cascode, ma con un errore nell'ordine di $1/\beta^2$.



Studiamo l'accensione del circuito



Fintanto che il nodo H è a tensione inferiore a $2V_T$, i transistori Q2 e Q4 sono spenti. Se è spento Q2, lo è anche Q1.

$$V_H \ll 2V_T \rightarrow Q2, Q4: off$$

$$I_{B_4} = 0, I_{C_4} = 0, I_{C_2} = 0 \rightarrow I_{C_1} = 0$$

La corrente che scorre nella capacità parassita C_p risulta pari a:

$$I_{C_p} = I_{in} - I_{C_1} - I_{B_4}$$

Dunque, inizialmente si ha $I_{C_p} = I_{in}$. Poiché la capacità si carica a corrente costante, la tensione al nodo H cresce linearmente. Appena il nodo H si avvicina a $2V_\gamma$ iniziano ad accendersi Q4 e Q2. Nel momento in cui si accendono la I_{C_1} sale e si innesca la retroazione. La corrente nella capacità diminuisce e si raggiunge la stabilità quando la tensione al nodo H smette di crescere, la corrente nella capacità è nulla e $I_{in} = I_{C_1} + I_{B_4}$.

Determiniamo l'errore sul rapporto di specchio

$$I_{in} = I_{C_1} + I_{B_4}$$

$$I_{out} = I_{C_4} = I_{E_4} - I_{B_4} = I_{C_2} + I_{B_1} + I_{B_2} - I_{B_4}$$

Facendo il rapporto tra la I_{out} e la I_{in} entrambi i termini hanno una sola corrente di base che si somma alle correnti di collettore, per cui l'effetto delle correnti di base tende a compensarsi.

$$\frac{I_{out}}{I_{in}} = \frac{I_{C_2} + I_{B_1} + I_{B_2} - I_{B_4}}{I_{C_1} + I_{B_4}} \cong \frac{I_{C_2}}{I_{C_1}} \cong \frac{\text{area}_2}{\text{area}_1} = k_M$$

Facendo i conti in modo più accurato:

$$\frac{I_{out}}{I_{in}} = \frac{\frac{I_{C_2}}{k_M \beta_1} + \frac{I_{C_2}}{\beta_2} - \frac{I_{C_2}}{\beta_4}}{\frac{k_M I_{C_1}}{\beta_4}} = \frac{I_{C_2}}{I_{C_1}} \frac{1 + \frac{1}{k_M \beta_1} + \frac{1}{\beta_2} - \frac{1}{\beta_4}}{1 + \frac{k_M}{\beta_4}}$$

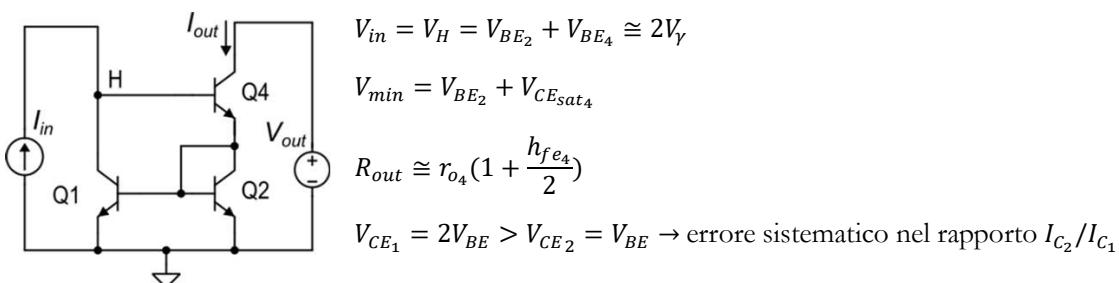
Si è approssimata la corrente I_{C_4} con la I_{C_2} . Tale approssimazione si ottiene trascurando le correnti di base, per cui comporta un errore nell'ordine di $1/\beta$. Tuttavia, le correnti di base vengono divise nuovamente per β . Dunque, approssimando I_{C_4}/β_4 con I_{C_2}/β_4 introduce al più errori nell'ordine di $1/\beta^2$.

L'errore dovuto alla corrente di base è nell'ordine di $1/\beta^2$ soltanto quando il guadagno di specchio è progettato per essere unitario. Questo perché le correnti di base sono simili tra loro soltanto quando lo sono quelle di emettitore, e ciò accade quando $k_M = 1$. Approssimando con Taylor al prim'ordine:

$$\frac{I_{out}}{I_{in}} \cong k_M \left(1 + \frac{1}{k_M \beta_1} - \frac{k_M}{\beta_4} + \frac{1}{\beta_2} - \frac{1}{\beta_4} \right)$$

$$\frac{1}{\beta_1} - \frac{1}{\beta_4} \approx \frac{1}{\beta^2} \quad k_M = 1 \rightarrow \begin{cases} \frac{I_{out}}{I_{in}} = k_M(1 + \epsilon_R) \\ |\epsilon_R| \approx \frac{1}{\beta^2} \end{cases}$$

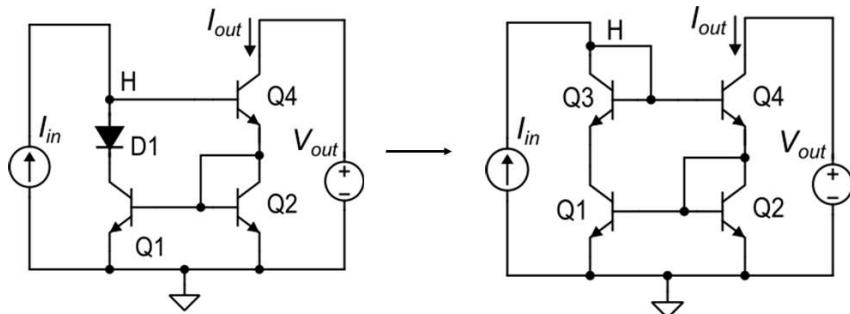
Per guadagni di corrente diversi da quello unitario l'errore relativo dovuto alle correnti di base tende a $1/\beta$. Dunque, lo specchio di Wilson funziona bene soltanto nel caso di guadagno in corrente unitario. In definitiva:



Lo specchio di Wilson si comporta bene nel far sì che I_{out}/I_{in} sia vicino al rapporto I_{C_2}/I_{C_1} , che è tanto più vicino al rapporto tra le aree quanto più V_{CE_1} e V_{CE_2} sono uguali. In questo caso $V_{CE_1} > V_{CE_2}$, il che introduce un errore sistematico $I_{in} > I_{out}$ tanto peggiore quanto la tensione di Early dei transistori è scarsa.

Specchio di corrente di Wilson a 4 transistori

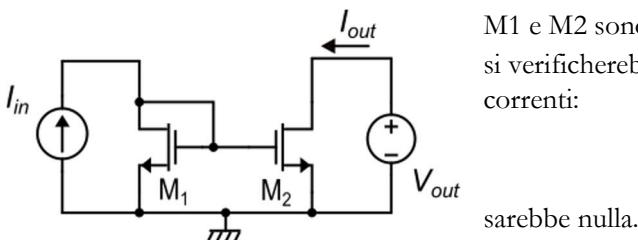
Si può risolvere la problematica aggiungendo una traslazione di V_V attraverso una giunzione BE in più, in modo tale che le V_{CE} di Q1 e Q2 tornino ad essere uguali.



La V_H è la stessa di prima, ed è fissata dal meccanismo di reazione che stabilizza la corrente in Q1 circa uguale a I_{in} . In questo modo le V_{CE} di Q1 e Q2 sono uguali e lo specchio è accurato.

Errori di matching negli specchi di corrente

Finora si sono trattati alcuni aspetti circuitali che influiscono sull'accuratezza degli specchi, quali la resistenza di uscita, l'uguaglianza delle V_{ds} (o V_{ce}) nelle configurazioni cascode, le correnti di base nel caso dei BJT. Le analisi fatte sono tutte analisi nominali, che non tengono conto degli errori di processo. Consideriamo uno specchio standard a MOS con guadagno unitario. Nel caso di specchi con guadagni non unitari si potrebbero fare considerazioni analoghe studiando l'accuratezza che affligge il rapporto delle correnti.



M1 e M2 sono nominalmente identici. Se lo fossero anche fisicamente, si verificherebbe la condizione $I_{in} = I_{out}$, per cui la differenza tra le correnti:

$$\Delta = I_{out} - I_{in} = I_{D_2} - I_{D_1} = \Delta I_D$$

sarebbe nulla.

In questo caso si considerano soltanto gli errori di matching. Trascurando tutte le altre sorgenti di errore, quali la discrepanza tra le V_{ds} , in forte inversione vale che:

$$I_D \cong \frac{\beta}{2} (V_{GS} - V_t)^2 = kA^2B$$

Dove $k = 1/2$, $A = (V_{GS} - V_t)$ e $B = \beta$. Con questa trasformazione l'espressione della corrente di drain è in forma posinomiale. In questo modo:

$$\frac{\Delta I}{I} = \frac{\Delta I_D}{I_D} = \frac{2\Delta A}{A} + \frac{\Delta \beta}{\beta} = \frac{\Delta \beta}{\beta} - \frac{2\Delta V_t}{(V_{GS} - V_t)}$$

L'errore di matching che affligge la I_D , che è una grandezza derivata, è dato dalla somma pesata degli errori di matching relativi sulle grandezze A e B . La quantità $\Delta A = \Delta(V_{GS} - V_t) = -\Delta V_t$ poiché le V_{GS} dei due transistori sono uguali a prescindere da condizioni di processo, per cui la differenza ΔV_{GS} è sempre nulla. Dunque, l'errore di matching sull'overdrive dipende esclusivamente dall'errore di matching tra le due tensioni di soglia. L'errore relativo sulla corrente dipende direttamente dall'errore relativo sul β . La componente di errore legata alla tensione di soglia è invece abbattuta dall'overdrive: se lo specchio opera con overdrive maggiore è più robusto rispetto all'effetto dell'errore di matching sulle tensioni di soglia. Tuttavia, aumentando l'overdrive si degradano altri parametri dello specchio, tra cui V_{in} e V_{min} , con conseguente peggioramento della minima tensione di alimentazione. Talvolta può rivelarsi necessario ridurre la V_{min} indipendentemente dalla V_{in} . Dunque, la scelta dell'overdrive è frutto di un compromesso; per la massima dinamica si deve ridurre l'overdrive, anche portando i transistori ad operare in regime di sottosoglia.

Ipotizzando che l'errore relativo sul β e l'errore assoluto sulla tensione di soglia siano variabili aleatorie indipendenti, possiamo calcolare la deviazione standard sull'errore relativo delle correnti:

$$\sigma_{\frac{\Delta I_D}{I_D}} = \sqrt{\sigma_{\frac{\Delta \beta}{\beta}}^2 + \left(\frac{2\sigma_{\Delta V_t}}{V_{GS} - V_t} \right)^2}$$

Le deviazioni standard dell'errore relativo sul β e dell'errore assoluto della tensione di soglia sono controllabili attraverso l'area dei transistori. In particolare:

$$\sigma_{\frac{\Delta \beta}{\beta}} = \frac{C_\beta}{\sqrt{WL}} \quad \sigma_{\Delta V_t} = \frac{C_{V_t}}{\sqrt{WL}}$$

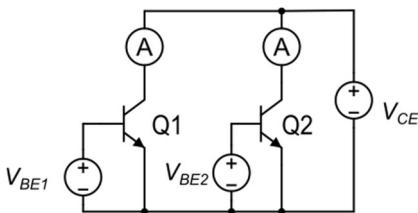
Esempi

	$\frac{\sigma_{\Delta \beta}}{\beta}$	$\frac{2\sigma_{\Delta V_t}}{V_{GS} - V_t}$	$\frac{\sigma_{\Delta I_D}}{I_D}$
$L = W = 1 \mu m \quad V_{GS} - V_t = 100 mV$	0.03	0.17	0.173
$L = W = 1 \mu m \quad V_{GS} - V_t = 500 mV$	0.03	0.034	0.0453
$L = W = 10 \mu m \quad V_{GS} - V_t = 100 mV$	0.003	0.0034	0.00453

$$C_\beta = 0.03 \mu m$$

$$C_{V_t} = 8.5 mV \cdot \mu m$$

Nel primo caso i transistori si trovano ad operare al confine tra la forte e la moderata inversione. Se si conosce la corrente nominale a cui deve operare lo specchio, l'overdrive è solo una conseguenza dell'aspect ratio. Come si osserva, poiché il termine legato alla tensione di soglia è molto maggiore di quello legato al β , la deviazione standard dell'errore relativo sulle correnti è praticamente interamente determinato dal mismatch delle tensioni di soglia. Si può tentare di migliorare l'errore aumentando l'overdrive. Per mantenere la stessa corrente con lo stesso prodotto WL si può moltiplicare e dividere la L e la W rispettivamente per lo stesso fattore. A parità di corrente è come se il “transistor fosse più resistivo” e facesse cadere un maggiore overdrive. Nel secondo caso aumenta l'overdrive e le dimensioni dei transistori rimangono invariate. Dunque, a parte l'effetto collaterale di una corrente 25 maggiore, si nota come l'aumento dell'overdrive con area dei transistori costante determini un abbassamento della deviazione standard di $\Delta I_D/I_D$ fino al 4.53%. In uno specchio di corrente, più grande è l'overdrive migliore è l'accuratezza in termini di bassa sensibilità agli errori di processo. Questo va in contrasto con altri parametri, quali la V_{in} e la V_{min} . Per migliorare le cose si può anche giocare con le aree dei transistori. Aumentando dello stesso fattore L e W l'aspect ratio si conserva e le deviazioni standard su β e V_t diminuiscono con la radice quadrata di WL . Ecco uno dei motivi per i quali l'elettronica analogica non scala come quella digitale. I processi attuali sono migliori, i parametri con cui si modellano gli errori di matching sono migliori, ma non così tanto da poter scalare troppo i dispositivi. Per uno specchio cascode, l'effetto dell'errore di matching maggiore è da parte dei transistori di sotto. Il mismatch tra i transistori di sopra determina un discostamento tra le V_{DS} , ma i transistori sono molto più sensibili alle variazioni della V_{GS} che non alle variazioni della V_{DS} . Ecco perché i transistori di sopra possono essere resi più piccoli.



Per i bipolari, la fonderia non fornisce direttamente un errore di matching sulle correnti di saturazione. Bensì, seleziona due bipolari nominalmente identici e li polarizza con una $V_{CE} \gg V_{CE_{sat}}$ in modo che operino entrambi in zona attiva diretta. Dopo di che, misurando le correnti di collettore, fa in modo che queste siano uguali variando le V_{BE} dei bipolari.

Trascurando l'effetto della V_{CE} sulla modulazione della corrente di collettore, in zona attiva diretta:

$$I_C = I_S e^{\frac{V_{BE}}{V_T}} \left(1 + \frac{V_{CB}}{V_A} \right) \cong I_S e^{\frac{V_{BE}}{V_T}}$$

Se la temperatura a cui si trovano i bipolari è la stessa, l'unica fonte di mismatch è la corrente di saturazione I_S .

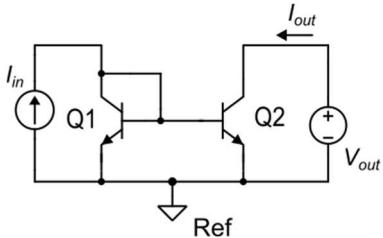
Poiché i transistori sono diversi a causa del mismatch, per avere la stessa corrente avranno bisogno di V_{BE} diverse. La fonderia fornisce proprio la differenza di V_{BE} che si deve applicare a una coppia di bipolari nominalmente identici affinché, con stessa V_{CE} , conducano una stessa corrente di riferimento.

$$V_{BE} = V_T \ln\left(\frac{I_C}{I_S}\right) \rightarrow \Delta V_{BE} = V_T \left(\frac{\Delta I_C}{I_C} - \frac{\Delta I_S}{I_S} \right)$$

In questo caso la differenza tra la V_{BE} è calcolata nella particolare condizione in cui le correnti di collettore dei bipolari sono uguali:

$$I_{C_1} = I_{C_2} \rightarrow \Delta I_C = 0 \rightarrow \Delta V_{BE}^* \triangleq \Delta V_{BE}|_{\Delta I_C=0} = -V_T \left(\frac{\Delta I_S}{I_S} \right)$$

La fonderia, dopo aver effettuato molte misure di questo tipo, ottenendo una statistica della ΔV_{BE}^* ne fornisce la deviazione standard per il bipolare elementare. Applichiamo ciò allo specchio di corrente a BJT:



Tale specchio ha un errore sistematico rispetto alla differenza $I_{out} - I_{in}$ dovuto alle correnti di base. Quindi, per studiare l'effetto degli errori di matching si prendono in esame quantità che avevamo ipotizzato essere nominalmente identiche. Nel caso specifico, si considera la differenza tra le correnti di collettore:

$$\Delta I_C = I_{C_2} - I_{C_1}$$

Nel considerare gli errori di matching, il termine e^{V_{BE}/V_T} può essere ignorato. Infatti, le V_{BE} dei transistori sono uguali tra loro per topologia circuitale e le V_T sono uguali per ipotesi di uniformità di temperatura. Dunque, si ottiene che:

$$I_C \cong I_S e^{\frac{V_{BE}}{V_T}} \rightarrow \frac{\Delta I_C}{I_C} = \frac{\Delta I_S}{I_S}$$

L'errore relativo sulle correnti di saturazione può essere ricavato in funzione della ΔV_{BE}^* :

$$\Delta V_{BE}^* = -V_T \left(\frac{\Delta I_S}{I_S} \right) \rightarrow \frac{\Delta I_S}{I_S} = -\frac{\Delta V_{BE}^*}{V_T}$$

Dunque, passando ad una valutazione di tipo statistico:

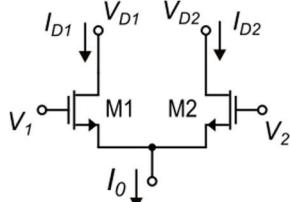
$$\sigma_{\frac{\Delta I_C}{I_C}} = \sigma_{\frac{\Delta I_S}{I_S}} = \frac{\sigma_{\Delta V_{BE}^*}}{V_T}$$

Valori tipici per $\sigma_{\Delta V_{BE}^*}$ sono $100 \div 300 \mu V$. Prendendo $V_T = 25 mV$ si ottiene che la deviazione standard sull'errore relativo $\Delta I_C/I_C$ risulta compresa tra $0.4 \cdot 10^{-2} \div 1.2 \cdot 10^{-2}$. Anche se il confronto con i MOSFET è difficile dato che il bipolare elementare potrebbe avere un'area considerevole, tipicamente a parità di area i bipolaris risentono del matching in modo molto più attenuato. Per migliorare ulteriormente la performance dello specchio si possono utilizzare bipolaris più grandi aumentando il parametro area (o disponendo più transistori in parallelo).

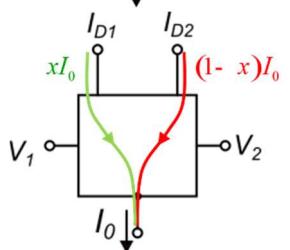
$$\sigma_{\Delta V_{BE}^*} = \frac{\sigma_{\Delta V_{BE}^* elem}}{\sqrt{area}}$$

Coppie differenziali

Un altro blocco fondamentale per la realizzazione di circuiti analogici integrati è la coppia differenziale. Si tratta di una coppia di transistori che condividono il terminale di source (source coupled MOSFET pair/differential MOSFET pair).



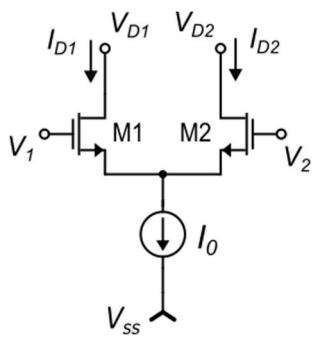
Al circuito potremmo assegnare tre input: le tensioni V_1 , V_2 e la corrente di polarizzazione (corrente di coda) I_0 . Tuttavia, la parte utile dei potenziali V_1 e V_2 è la differenza $V_D = V_1 - V_2$. Dunque, gli input del circuito, nella pratica, sono la tensione differenziale e la corrente di polarizzazione. L'uscita potrebbe essere costituita dalle singole correnti I_{D1} , I_{D2} o, più spesso, la loro differenza $I_{D1} - I_{D2}$.



La funzione caratteristica della coppia differenziale è quella di prendere la corrente di input I_0 e, attraverso il controllo della tensione differenziale V_D , smistarla in due frazioni variabili x e $(1 - x)$ tra le due correnti I_{D1} e I_{D2} . Se la $V_D = 0$ il sistema risulta essere simmetrico, per cui la corrente I_0 si suddivide a metà tra la parte di sinistra e la parte di destra, ovvero $x = 1/2$. In generale si ha che:

$$x = f(V_D)$$

Grazie a questa funzione la coppia differenziale può essere impiegata in tantissime applicazioni. L'obiettivo della seguente analisi è quello di ricavare la funzione che lega la tensione differenziale V_D alla frazione x con cui vengono smistate le correnti dei due rami.



Si considera ideale il generatore che produce la corrente di polarizzazione I_0 .

$$I_{D1} = xI_0 \rightarrow x = I_{D1}/I_0$$

Si effettuano le seguenti ipotesi:

- M1 e M2 sono nominalmente identici
- M1 e M2 operano in regime di saturazione; si trascurano gli effetti delle V_{DS}
- Si possono applicare le equazioni di forte inversione
- M1 e M2 condividono lo stesso substrato ($V_{B1} = V_{B2}$)
- La corrente I_0 non dipende dai potenziali V_1 , V_2

Se valgono le ipotesi elencate:

$$I_D = \frac{\beta}{2} (V_{GS} - V_t)^2 \rightarrow V_{GS} = V_t + \sqrt{\frac{2I_D}{\beta}}$$

$$\begin{cases} V_1 = V_{S1} + V_{GS1} = V_S + V_{GS1} \\ V_2 = V_{S2} + V_{GS2} = V_S + V_{GS2} \end{cases} \rightarrow V_D = V_1 - V_2 = V_{GS1} - V_{GS2} = V_{t1} + \sqrt{\frac{2I_{D1}}{\beta_1}} - \left(V_{t2} + \sqrt{\frac{2I_{D2}}{\beta_2}} \right)$$

Poiché l'analisi è condotta in condizioni nominali, cioè per M1 e M2 identici, le tensioni di soglia e i β si possono assumere uguali. Quindi:

$$V_D = V_t + \sqrt{\frac{2I_{D1}}{\beta}} - \left(V_t + \sqrt{\frac{2I_{D2}}{\beta}} \right) = \sqrt{\frac{2I_{D1}}{\beta}} - \sqrt{\frac{2I_{D2}}{\beta}} = \sqrt{\frac{2}{\beta}} \left(\sqrt{I_{D1}} - \sqrt{I_{D2}} \right)$$

$$\rightarrow V_D = \sqrt{\frac{2}{\beta}} \left(\sqrt{I_{D1}} - \sqrt{I_{D2}} \right) = \sqrt{\frac{2}{\beta}} \left(\sqrt{xI_0} - \sqrt{(1-x)I_0} \right) = \sqrt{\frac{2I_0}{\beta}} \left(\sqrt{x} - \sqrt{(1-x)} \right) = V_{Dmax} \left(\sqrt{x} - \sqrt{(1-x)} \right)$$

$$\rightarrow \frac{V_D}{V_{Dmax}} = \left(\sqrt{x} - \sqrt{(1-x)} \right)$$

Elevando a quadrato l'espressione si può ottenere una soluzione, ma si aggiungono anche soluzioni non valide. Per cui, occorre aggiungere alcune condizioni a contorno:

$$V_D > 0 \rightarrow \sqrt{x} - \sqrt{(1-x)} > 0 \rightarrow \sqrt{x} > \sqrt{(1-x)} \rightarrow x > (1-x) \rightarrow x > 1/2$$

$$\left(\frac{V_D}{V_{D_{max}}}\right)^2 = x + (1-x) - 2\sqrt{x(1-x)} = 1 - 2\sqrt{x(1-x)} \rightarrow \left(\frac{V_D}{V_{D_{max}}}\right)^2 - 1 = -2\sqrt{x(1-x)}$$

Il secondo membro è necessariamente negativo, dunque anche il primo deve essere negativo. È proprio questa una delle altre condizioni che il sistema dovrà rispettare affinché le soluzioni siano accettabili. Elevando ancora al quadrato:

$$\left[\left(\frac{V_D}{V_{D_{max}}}\right)^2 - 1\right]^2 = 4x(1-x) \rightarrow \frac{1}{4}\left[\left(\frac{V_D}{V_{D_{max}}}\right)^2 - 1\right]^2 = x - x^2 \rightarrow x^2 - x + \frac{1}{4}\left[\left(\frac{V_D}{V_{D_{max}}}\right)^2 - 1\right]^2 = 0$$

$$x = \frac{1}{2} \pm \frac{1}{2} \sqrt{1 - \left[\left(\frac{V_D}{V_{D_{max}}}\right)^2 - 1\right]^2} = \frac{1}{2} \pm \frac{1}{2} \sqrt{2\left(\frac{V_D}{V_{D_{max}}}\right)^2 - \left(\frac{V_D}{V_{D_{max}}}\right)^4} = \frac{1}{2} \pm \frac{1}{2} \left(\frac{V_D}{V_{D_{max}}}\right) \sqrt{2 - \left(\frac{V_D}{V_{D_{max}}}\right)^2}$$

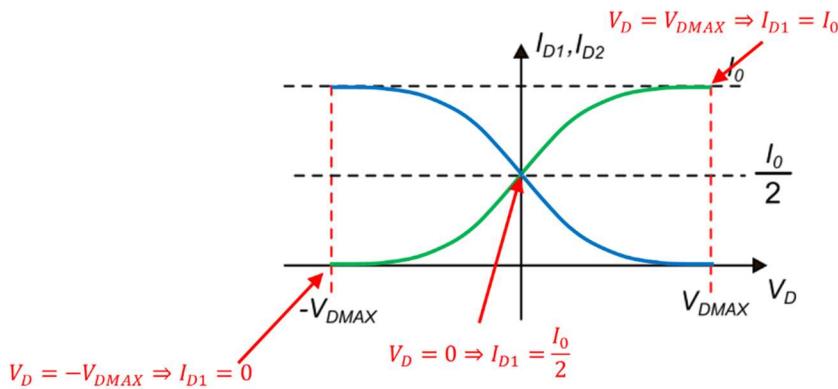
Utilizzando la condizione $V_D > 0 \rightarrow x > \frac{1}{2}$ si ottiene che:

$$\begin{cases} x = \frac{I_{D_1}}{I_0} = \frac{1}{2} + \frac{1}{2}\left(\frac{V_D}{V_{D_{max}}}\right)\sqrt{2 - \left(\frac{V_D}{V_{D_{max}}}\right)^2} \\ 1-x = \frac{I_{D_2}}{I_0} = \frac{1}{2} - \frac{1}{2}\left(\frac{V_D}{V_{D_{max}}}\right)\sqrt{2 - \left(\frac{V_D}{V_{D_{max}}}\right)^2} \end{cases} \rightarrow \begin{cases} I_{D_1} = \frac{I_0}{2} + \frac{I_0}{2}\left(\frac{V_D}{V_{D_{max}}}\right)\sqrt{2 - \left(\frac{V_D}{V_{D_{max}}}\right)^2} \\ I_{D_2} = \frac{I_0}{2} - \frac{I_0}{2}\left(\frac{V_D}{V_{D_{max}}}\right)\sqrt{2 - \left(\frac{V_D}{V_{D_{max}}}\right)^2} \end{cases}$$

La seconda condizione trovata non pone limitazioni direttamente sull'espressione della x , ma sulla validità delle equazioni rispetto l'asse V_D :

$$\left(\frac{V_D}{V_{D_{max}}}\right)^2 - 1 \leq 0 \rightarrow \left|\frac{V_D}{V_{D_{max}}}\right| \leq 1 \rightarrow -V_{D_{max}} \leq V_D \leq V_{D_{max}}$$

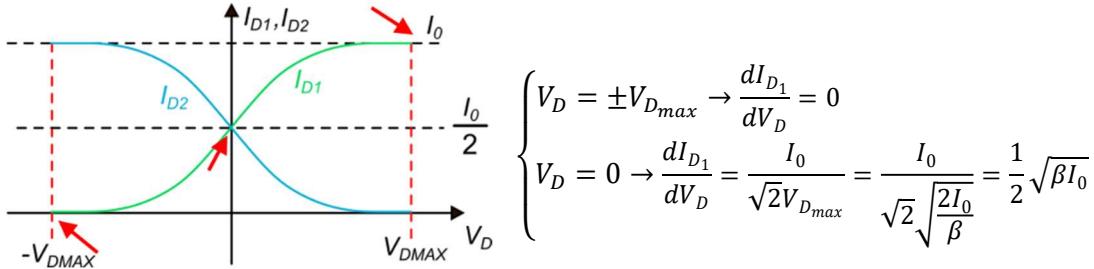
Dunque, la $V_{D_{max}}$ è effettivamente il massimo valore della V_D in modulo entro cui vale questa analisi.



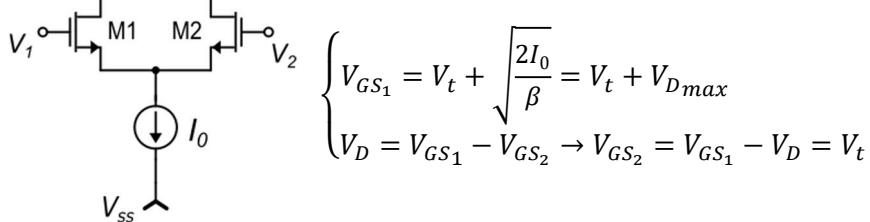
Affinché la somma $I_{D_1} + I_{D_2}$ sia ovunque I_0 , l'andamento di I_{D_2} è simmetrico a quello di I_{D_1} rispetto all'asse y e viceversa. Vale anche la simmetria rispetto all'asse x : scambiando di segno la V_D una corrente diventa l'altra. Per dimostrare che l'andamento sia effettivamente quello raffigurato si può effettuare uno studio di funzione.

$$z = \left(\frac{V_D}{V_{D_{max}}} \right) \rightarrow I_{D_1} = \frac{I_0}{2} + \frac{I_0}{2} \left(\frac{V_D}{V_{D_{max}}} \right) \sqrt{2 - \left(\frac{V_D}{V_{D_{max}}} \right)^2} = \frac{I_0}{2} \left[1 + z \sqrt{2 - z^2} \right]$$

$$\frac{dI_{D_1}}{dV_D} = \frac{I_0}{2V_{D_{max}}} \left[\sqrt{2 - z^2} - \frac{2z^2}{2\sqrt{2 - z^2}} \right] = \frac{I_0}{2V_{D_{max}}} \frac{2 - z^2 - z^2}{\sqrt{2 - z^2}} = \frac{I_0}{V_{D_{max}}} \frac{1 - z^2}{\sqrt{2 - z^2}}$$

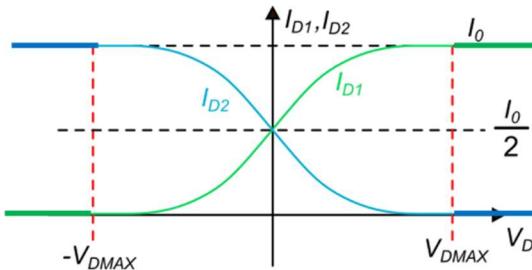


Per capire cosa accade all'infuori dell'intervallo $[-V_{D_{max}}, V_{D_{max}}]$ si passa ad un'analisi più fisica, circuitale. Per $V_D = V_{D_{max}}$



$$\begin{cases} V_{GS_1} = V_t + \sqrt{\frac{2I_0}{\beta}} = V_t + V_{D_{max}} \\ V_D = V_{GS_1} - V_{GS_2} \rightarrow V_{GS_2} = V_{GS_1} - V_D = V_t \end{cases}$$

Dunque, si dimostra che le curve saturano a I_0 dell'intervallo $[-V_{D_{max}}, V_{D_{max}}]$. Se V_D aumenta V_{GS_1} non può aumentare perché altrimenti la $I_{D_1} > I_0$. Dunque, V_{GS_2} diventa minore di V_t , perciò la I_{D_1} si fissa a I_0 mentre la I_{D_2} si fissa a 0. Il caso opposto ma analogo si verifica quando $V_D < -V_{D_{max}}$.



La coppia differenziale effettua lo smistamento controllato della corrente tra i due rami entro l'intervallo $[-V_{D_{max}}, V_{D_{max}}]$, all'infuori del quale satura sbilanciandosi completamente da una parte senza produrre più risposte a variazioni della tensione differenziale. Dunque, la $V_{D_{max}}$ rappresenta la dinamica di ingresso a modo differenziale della coppia differenziale. La $V_{D_{max}}$ è progettabile facilmente:

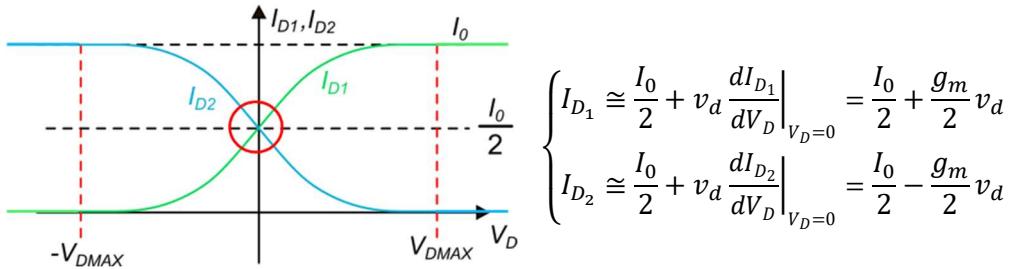
$$V_{D_{max}} = \sqrt{\frac{2I_0}{\beta}}$$

Per $V_D = 0$, che è il tipico punto di riposo della coppia differenziale:

$$I_{D_1} = I_{D_2} = I_{DQ} = \frac{I_0}{2} \rightarrow V_{D_{max}} = \sqrt{\frac{2I_0}{\beta}} = \sqrt{2} \sqrt{\frac{2I_{DQ}}{\beta}} = \sqrt{2}(V_{GS} - V_t)_Q$$

$$\left. \frac{dI_{D_1}}{dV_D} \right|_{V_D=0} = \frac{1}{2} \sqrt{2\beta I_{DQ}} = \frac{g_m}{2}$$

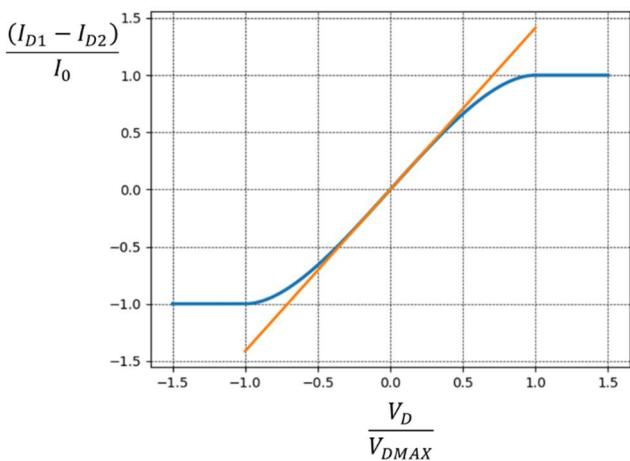
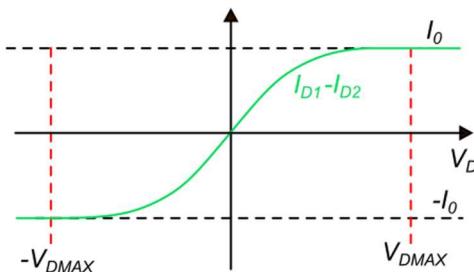
La dinamica di ingresso della coppia differenziale V_{Dmax} è direttamente proporzionale all'overdrive di riposo dei due transistori. Se si è progettata la coppia al limite tra la forte e la moderata inversione, la V_{Dmax} è circa 150 mV. Non è detto che la dinamica sia utile, però è utile che sia programmabile. Attorno all'origine, il comportamento per piccoli segnali:



Spesso può essere utile prelevare dal circuito una corrente differenziale:

$$\begin{cases} I_{D_1} = \frac{I_0}{2} + \frac{I_0}{2} \left(\frac{V_D}{V_{D_{max}}} \right) \sqrt{2 - \left(\frac{V_D}{V_{D_{max}}} \right)^2} \\ I_{D_2} = \frac{I_0}{2} - \frac{I_0}{2} \left(\frac{V_D}{V_{D_{max}}} \right) \sqrt{2 - \left(\frac{V_D}{V_{D_{max}}} \right)^2} \end{cases} \rightarrow I_{D_1} - I_{D_2} = I_0 \left(\frac{V_D}{V_{D_{max}}} \right) \sqrt{2 - \left(\frac{V_D}{V_{D_{max}}} \right)^2}$$

$$i_d = i_{d_1} - i_{d_2} = g_m v_d$$



In arancione si raffigura l'approssimazione lineare calcolata attorno all'origine. L'errore di non linearità rispetto alla curva reale:

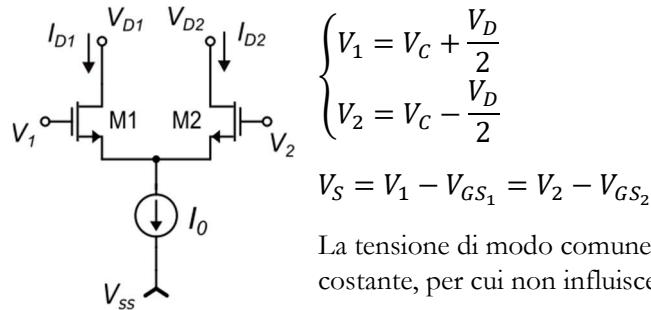
$$|V_D| < \frac{V_{D_{MAX}}}{5} : e_R < 1\%$$

$$|V_D| < \frac{V_{D_{MAX}}}{2\sqrt{2}} = \frac{V_{GS} - V_t}{2} : e_R < 3\%$$

$$|V_D| < \frac{V_{D_{MAX}}}{2} : e_R < 7\%$$

$$|V_D| < \frac{V_{D_{MAX}}}{\sqrt{2}} = (V_{GS} - V_t) : e_R < 15\%$$

Applicando una tensione solo differenziale si possono considerare i source a massa solo al piccolo segnale. Per grandi segnali questa assunzione diventa falsa.



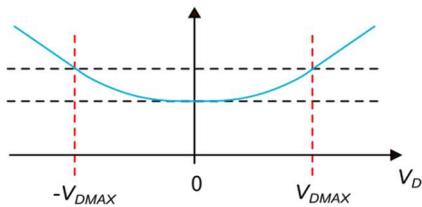
La tensione di modo comune serve a polarizzare la coppia, ma può essere mantenuta costante, per cui non influisce sull'analisi.

Per l'analisi si considera la seconda uguaglianza, ovvero $V_S = V_2 - V_{GS_2}$:

$$V_S|_{V_D=0} = V_2|_{V_D=0} - V_{GS_2}|_{V_D=0} = V_C - V_t - (V_{GS} - V_t)_Q$$

$$V_S|_{V_D=V_{Dmax}} = V_2|_{V_D=V_{Dmax}} - V_{GS_2}|_{V_D=V_{Dmax}} = V_C - \frac{V_{Dmax}}{2} - V_t = V_C - V_t - \frac{(V_{GS} - V_t)_Q}{\sqrt{2}}$$

La tensione di source sale nel passare da $V_D = 0$ a $V_D = V_{Dmax}$. Si può ripetere la stessa analisi per V_D che scende da 0 fino a $-V_{Dmax}$, trovando un comportamento simmetrico.



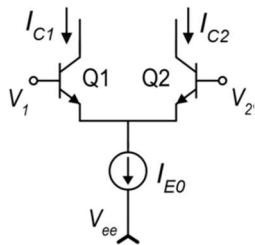
Si verifica che la derivata della tensione di source rispetto alla tensione differenziale è nulla attorno all'origine. Questo conferma che al piccolo segnale, quando la coppia opera a riposo a tensione differenziale nulla, il source si trova effettivamente a massa.

Quando la V_D esce dal range $[-V_{Dmax}, V_{Dmax}]$ la tensione di source è sostanzialmente decisa dal transistore che porta la totalità della corrente. In particolare, poiché la corrente totale massima è I_0 , andando oltre l'intervallo il transistore che conduce I_0 avrà un overdrive fissato. Pertanto, la tensione di source segue quella di gate traslando rigidamente. Ad esempio, per $V_D > V_{Dmax}$:

$$V_S = V_C + \frac{V_D}{2} - V_{GS_1} = V_C + \frac{V_D}{2} - (V_{Dmax} + V_t)$$

La tensione di gate, per $V_D > V_{Dmax}$, aumenta linearmente con la V_D con pendenza $1/2$. Lo stesso accade per $V_D < -V_{Dmax}$, per cui è il M2 a condurre tutta la corrente della coppia e a determinare la tensione di source come conseguenza del fatto che $I_{D2} = I_0$ fissa.

Coppia differenziale a bipolari – emitter coupled pair



Lo schema è analogo a quello della versione a MOSFET.

$$I_{E_1} + I_{E_2} = I_{E_0} \rightarrow \frac{\beta + 1}{\beta} I_{C_1} + \frac{\beta + 1}{\beta} I_{C_2} = I_{E_0}$$

Moltiplicando per β e dividendo per $\beta + 1$ si ottiene che:

$$I_{C_1} + I_{C_2} = \frac{\beta}{\beta + 1} I_{E_0} = \alpha I_{E_0} \rightarrow I_{C_1} + I_{C_2} \triangleq I_0$$

Se β è sufficientemente grande, α è circa unitario e la somma delle correnti di collettore I_0 è circa pari alla corrente di tail I_{E_0} . La tensione differenziale:

$$\begin{cases} V_1 = V_E + V_{BE_1} \\ V_2 = V_E + V_{BE_2} \end{cases} \rightarrow V_D = V_1 - V_2 = V_{BE_1} - V_{BE_2}$$

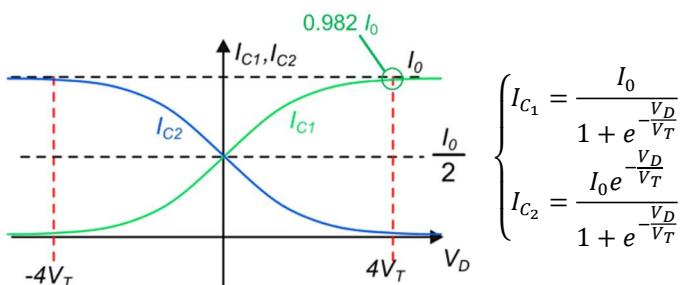
Trascurando l'effetto delle V_{CE} e considerando di operare in zona attiva diretta:

$$\begin{cases} I_{C_1} = I_S e^{\frac{V_{BE_1}}{V_T}} \\ I_{C_2} = I_S e^{\frac{V_{BE_2}}{V_T}} \end{cases} \rightarrow \frac{I_{C_2}}{I_{C_1}} = e^{\frac{V_{BE_2} - V_{BE_1}}{V_T}} = e^{\frac{-V_D}{V_T}}$$

L'analisi è condotta in condizioni nominali: si assumono Q1 e Q2 identici, per cui le correnti di saturazione dei due dispositivi sono le stesse e si cancellano a vicenda.

$$\frac{I_{C_2}}{I_{C_1}} = e^{\frac{-V_D}{V_T}} \rightarrow I_{C_2} = I_{C_1} e^{\frac{-V_D}{V_T}} \rightarrow I_{C_1} + I_{C_2} = I_{C_1} + I_{C_1} e^{\frac{-V_D}{V_T}} = I_{C_1} \left(1 + e^{\frac{-V_D}{V_T}}\right) = I_0$$

Dunque, si ottiene che:



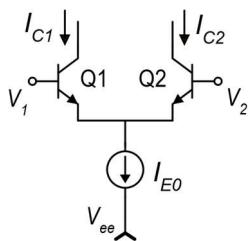
Per $V_D \rightarrow -\infty$, la I_{C_1} tende asintoticamente a 0. Viceversa, per $V_D \rightarrow +\infty$ la I_{C_1} tende a I_0 . La I_{C_2} , per mantenere la somma costante a I_0 , si comporta in modo complementare rispetto a $I_0/2$. A differenza del MOSFET, non si ha una tensione differenziale per cui la corrente è esattamente nulla in un ramo. Tuttavia, poiché il comportamento asintotico è esponenziale, se la tensione differenziale raggiunge il valore di $4V_T$ uno dei due rami porta il 98.2% della corrente, l'altro la restante parte. In realtà anche nel caso dei MOSFET la corrente di tail non è completamente sbilanciata da un lato agli estremi del range operativo. Questo perché uno dei due MOSFET in tale condizione opera con $V_{GS} = V_t$, per cui non è del tutto spento.

In molti casi la corrente utile della coppia è quella differenziale:

$$I_{C_1} - I_{C_2} = I_0 \frac{1 - e^{-\frac{V_D}{V_T}}}{1 + e^{\frac{V_D}{V_T}}} = I_0 \frac{e^{-\frac{V_D}{2V_T}} \left(e^{\frac{V_D}{2V_T}} - e^{-\frac{V_D}{2V_T}} \right)}{e^{-\frac{V_D}{2V_T}} \left(e^{\frac{V_D}{2V_T}} + e^{-\frac{V_D}{2V_T}} \right)} = I_0 \tanh \left(\frac{V_D}{2V_T} \right)$$

La corrente differenziale varia tra $-I_0$ e $+I_0$.

Comportamento di piccolo segnale



Applicando una piccola variazione della tensione differenziale rispetto al caso di riposo $V_D = 0$, al prim'ordine:

$$I_{c_1} = \frac{I_0}{2} + \frac{g_m}{2} v_d$$

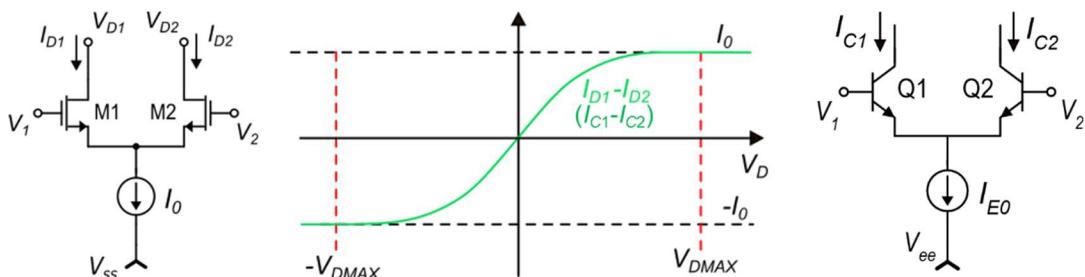
$$I_{c_2} = \frac{I_0}{2} - \frac{g_m}{2} v_d$$

Per piccolo segnale si possono considerare gli emettitori a massa. Continuando ad alzare il potenziale da un lato a modo differenziale, la tensione di emettitore si sposta verso l'alto fino a che non trasla rigidamente seguendo la V_{BE} del BJT che conduce maggiormente. La differenza rispetto al MOSFET è che il BJT, presentando anche una giunzione tra base e collettore, anche il potenziale del terminale di collettore viene trascinato rigidamente oltre una certa tensione differenziale.

$$I_{c_1} - I_{c_2} = g_m v_d$$

$$g_m = \frac{I_{cQ}}{V_T} = \frac{I_0}{2V_T}$$

Differenze tra la coppia a MOSFET e la coppia a BJT

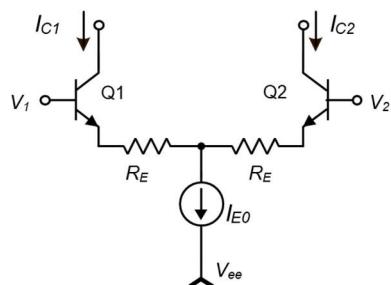


A livello qualitativo, le curve della differenza tra le correnti in funzione della tensione differenziale sono analoghe. Le differenze:

- Mentre nel caso della coppia a MOSFET la $V_{D_{max}}$ può essere variata, entro certi margini, modificando il β . In forte inversione $V_{D_{max}} = \sqrt{2}(V_{GS} - V_t)$. Per la coppia a bipolari, invece, la $V_{D_{max}}$ è fissa a $4V_T$.
- Il g_m nel caso dei MOSFET è pari a $\sqrt{\beta I_0}$, per cui, anche qualora la I_0 fosse fissata da specifica, può essere variato con β . Il g_m nel caso del bipolare, invece, è bloccato a $I_0/2V_T$.

Se per la coppia a MOSFET si considera che i dispositivi partono a riposo in debole inversione, il loro comportamento è dominato da equazioni esponenziali. In particolare, le equazioni sono le stesse del bipolare ma con la sostituzione $V_T \rightarrow mV_T$. In tal caso la coppia si comporta come quella a bipolare: il g_m dipende solo dalla corrente e la $V_{D_{max}}$ è fissata a $4mV_T$.

Per cambiare la dinamica differenziale di ingresso di una coppia differenziale a bipolare si può sfruttare la degenerazione di emettitore.

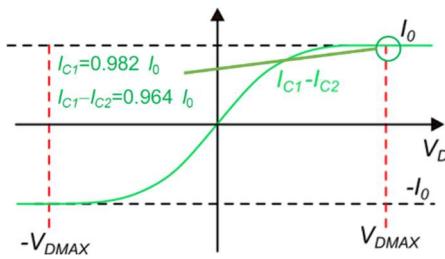


$$V_{D_{max}} \cong V'_{D_{max}} + I_{E_0} R_E = 4V_T + I_{E_0} R_E$$

Infatti:

$$V_D = V_{BE_1} + I_{E_1} R_E - (V_{BE_2} + I_{E_2} R_E) = V_{BE_1} - V_{BE_2} + R_E (I_{E_1} - I_{E_2})$$

$$\frac{I_{c_2}}{I_{c_1}} = e^{\frac{-V_D}{V_T}}$$



Considerando il punto evidenziato come riferimento, si può considerare la coppia come completamente sbilanciata. Si possono ripetere gli stessi calcoli, determinando le I_C in funzione delle V_{BE} . In questo caso la differenza delle V_{BE} per ottenere questa condizione sarà sempre $4V_T$.

Poiché in questa condizione il transistore Q1 conduce tutta la corrente di tail, si assume $I_{E_1} = I_{E_0}$ e $I_{E_2} = 0$. Dunque:

$$V_D = V_{BE_1} - V_{BE_2} + R_E(I_{E_1} - I_{E_2}) = 4V_T + R_E I_{E_0}$$

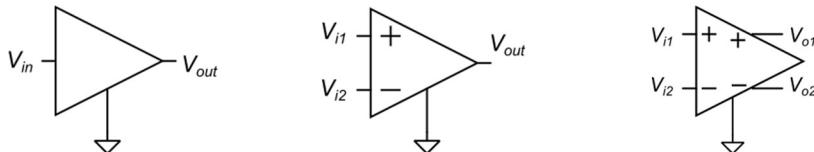
Si è ricavata così la tensione differenziale per cui lo specchio è quasi completamente sbilanciato. La differenza tra le V_{BE} in questa configurazione non è più la tensione differenziale di ingresso. Aumentare la $V_{D_{max}}$ ha l'effetto di stirare le curve delle correnti rispetto alla tensione V_D lungo la direzione x . Pertanto, la derivata attorno all'origine diminuirà. Tale effetto si traduce in un g_m equivalente inferiore.

$$g_{m-rid} = \frac{d(I_{C_1} - I_{C_2})}{dV_D} \Big|_{V_D=0} = \frac{g_m}{1 + g_m R_E} < g_m$$

Qualche volta la degenerazione viene impiegata per estendere la zona di comportamento lineare della caratteristica. Coppie differenziali a dinamica estesa sono utilizzate nella cella di Gilbert originaria. Ridurre il g_m attraverso una degenerazione con resistenze è una tecnica molto utilizzata in amplificatori operazionali veloci.

Amplificatori di tensione

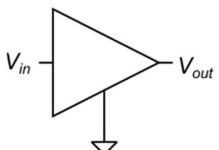
Una prima classificazione che si fa degli amplificatori di tensione riguarda il numero di terminali in ingresso e in uscita.



L'amplificatore più semplice è quello unipolare, che presenta un singolo ingresso e una singola uscita. Quando si parla di amplificatore differenziale ci si riferisce invece a un amplificatore con ingresso differenziale a due terminali e un singolo terminale di uscita (single ended). Gli amplificatori fully differential, invece, hanno sia ingresso che uscita differenziali.

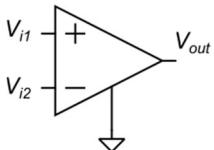
Guadagni

Il guadagno è un coefficiente che lega l'uscita all'ingresso in un'ipotesi di linearità, per cui ci riferisce a guadagni di piccolo segnale.



Per l'amplificatore unipolare si considera un unico guadagno A_v :

$$v_{out} = A_v v_{in}$$

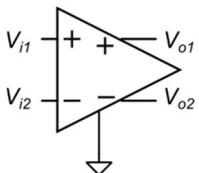


Nel caso differenziale l'ingresso è determinato oltre che dalla tensione differenziale, il segnale utile, anche dalla tensione a modo comune. Dunque, si identificano due coefficienti nel legame tra ingresso e uscita: un guadagno a modo differenziale e un guadagno a modo comune.

$$\begin{cases} v_d = v_{i1} - v_{i2} \\ v_c = \frac{v_{i1} + v_{i2}}{2} \end{cases} \rightarrow \begin{cases} A_d = \frac{v_{out}}{v_d} \Big|_{v_c=0} \\ A_c = \frac{v_{out}}{v_c} \Big|_{v_d=0} \end{cases} \rightarrow v_{out} = A_d v_d + A_c v_c$$

Il caso ideale è quello in cui l'amplificatore è sensibile solo alla tensione differenziale e reietta completamente il modo comune. Il parametro che valuta la bontà dell'amplificatore nel reiettare la componente a modo comune rispetto a quella a modo differenziale è il CMRR (Common Mode Rejection Ratio):

$$CMRR = \left| \frac{A_d}{A_c} \right|$$



Nel caso fully differential l'amplificatore ha sia due tensioni in ingresso che due tensioni in uscita. Entrambi ingresso e uscita possono essere scomposti come due tensioni, una a modo comune e l'altra a modo differenziale. Si distinguono quattro guadagni che legano le possibili combinazioni tra ingresso e uscita.

$$\begin{cases} v_{id} = v_{i1} - v_{i2} \\ v_{ic} = \frac{v_{i1} + v_{i2}}{2} \end{cases} \quad \begin{cases} v_{od} = v_{o1} - v_{o2} \\ v_{oc} = \frac{v_{o1} + v_{o2}}{2} \end{cases}$$

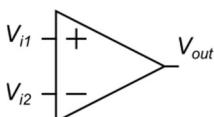
$$A_{dd} = \frac{v_{od}}{v_{id}} \Big|_{v_{ic}=0} \quad A_{cd} = \frac{v_{od}}{v_{ic}} \Big|_{v_{id}=0} \quad A_{dc} = \frac{v_{oc}}{v_{id}} \Big|_{v_{ic}=0} \quad A_{cc} = \frac{v_{oc}}{v_{ic}} \Big|_{v_{id}=0}$$

L'amplificazione utile è A_{dd} . Tra gli altri guadagni, quello più nocivo è A_{cd} , perché trasforma un disturbo a modo comune in ingresso in un modo differenziale all'uscita, confondendolo con l'uscita utile. I guadagni che trasformano in modo comune sono meno nocivi perché il modo comune è reiettabile da altri stadi differenziali in cascata. Tuttavia, è comunque importante che il modo comune in uscita non alteri troppo il comportamento dell'amplificatore in termini di distorsione o saturazione. Dunque, in questo caso, il CMRR:

$$CMRR = \left| \frac{A_{dd}}{A_{cd}} \right|$$

Dinamiche

Consideriamo un amplificatore differenziale single ended. Quando è possibile portare un amplificatore single ended a un'analogia versione fully differential, gli elementi passivi diventano il doppio e il circuito diventa molto più ingombrante.



Al comportamento ideale dell'amplificatore si aggiunge l'offset:

$$V_{out} = A_d(V_{id} - V_{io})$$

Un altro comportamento non ideale dell'amplificatore è che il comportamento lineare non è rispettato per qualunque tensione differenziale in ingresso. Genericamente, ci sarà una tensione $V_{D_{max}}$ oltre la quale, in modulo, l'amplificatore perde linearità. La dinamica differenziale in ingresso (input differential range):

$$-V_{D_{max}} \leq V_D \leq V_{D_{max}}$$

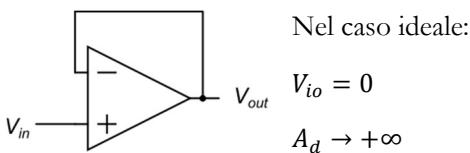
Ancora più importante della dinamica differenziale in ingresso, che può essere tenuta sotto controllo dalla reazione, è il range di modo comune in ingresso (input common-mode range). Oltre una certa dinamica di modo comune in ingresso, l'amplificatore differenziale non funziona più correttamente:

$$V_{C_{min}} \leq V_C \leq V_{C_{max}}$$

Si deve anche considerare la dinamica di uscita (output voltage range o output voltage swing). Variando la tensione di ingresso si può far variare la tensione di uscita. Superato un certo valore dell'uscita, o in alto o in basso, iniziano a comparire non linearità. Il migliore comportamento in questo senso è quello rail-to-rail; la tensione di alimentazione determina limite superiore e inferiore della tensione di uscita.

Input Common Mode range

Consideriamo un amplificatore operazionale chiuso a buffer.



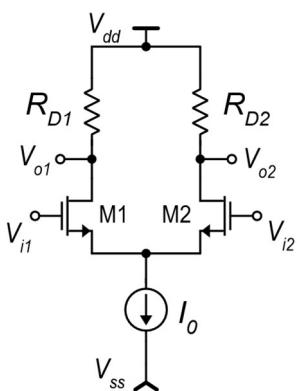
Se valgono queste ipotesi si instaura in ingresso il corto circuito virtuale e $V_{in} = V_{out}$. Dunque, la tensione differenziale è circa nulla, mentre quella di modo comune è $V_C = V_{in}$. Cioè, per un operazionale chiuso a buffer la tensione di modo comune varia esattamente come la tensione di ingresso. Pertanto, il comportamento da inseguitore di tensione si ha soltanto nel caso in cui la tensione di ingresso V_{in} appartiene al range di tensioni di modo comune in ingresso. Inoltre, poiché la $V_{out} = V_{in}$, la tensione di ingresso V_{in} dovrà appartenere anche al range di tensioni di uscita. Dunque:

$$V_{out} = V_{in} \leftrightarrow \begin{cases} V_{C_{min}} \leq V_{in} \leq V_{C_{max}} \\ V_{O_{min}} \leq V_{out} \leq V_{O_{max}} \end{cases}$$

Un inseguitore ideale dovrebbe inseguire tensioni che vanno dalla tensione di alimentazione negativa a quella positiva, soprattutto per applicazioni moderne in cui le alimentazioni sono basse. Pertanto, anche il range di modo comune in ingresso e il range di uscita dovrebbero essere entrambi rail to rail (per il modo comune ci si riferisce a rail to rail di input).

Amplificatore differenziale con carichi resistivi

L'amplificatore differenziale con carichi resistivi si usano pochissimo, ma sono didatticamente interessanti e mostrano alcuni pregi.



La tensione di ingresso differenziale: $V_{ID} = V_{i_1} - V_{i_2}$

La tensione di uscita differenziale: $V_{OD} = V_{o_1} - V_{o_2}$

Le resistenze ai drain dei MOSFET non hanno lo stesso nome perché a causa di errori di matching potrebbero risultare effettivamente diverse. Faremo un'analisi nominale per i transistori, che sono considerati identici, e non nominale per i resistori, che sono considerati potenzialmente diversi.

Ai grandi segnali:

$$V_{o_1} = V_{DD} - R_{D_1} I_{D_1} \quad V_{o_2} = V_{DD} - R_{D_2} I_{D_2} \quad V_{o_d} = R_{D_2} I_{D_2} - R_{D_1} I_{D_1}$$

Ai piccoli segnali:

$$v_{o_1} = -R_{D_1} i_{D_1} \quad v_{o_2} = -R_{D_2} i_{D_2} \quad v_{o_d} = R_{D_2} i_{D_2} - R_{D_1} i_{D_1}$$

A modo differenziale le variazioni delle correnti di drain:

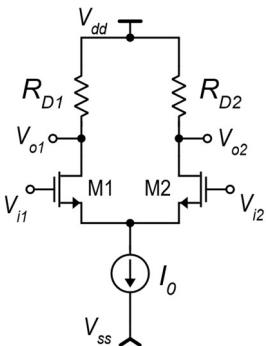
$$i_{d_1} = \frac{g_m}{2} v_{id} \quad i_{d_2} = -\frac{g_m}{2} v_{id}$$

Questo vale solo trascurando la dipendenza di i_d dalla v_{ds} (si trascura la r_d). La semplificazione di trascurare l'effetto delle r_d è abbastanza buona per $R_D \ll r_d$. Sostituendo:

Nel caso single ended: $v_{out} = v_{o_1} = -R_{D_1} \frac{g_m}{2} v_{id} \rightarrow A_d = -R_{D_1} \frac{g_m}{2}$

Nel caso fully differential: $v_{out} = v_{od} = -\frac{g_m}{2} (R_{D_1} + R_{D_2}) v_{id} \rightarrow A_{dd} = -g_m \frac{(R_{D_1} + R_{D_2})}{2}$

Se le resistenze sono identiche, nel caso fully differential l'amplificazione è doppia rispetto al caso single ended. Questo circuito può essere utilizzato sia in versione single-ended che in versione fully differential. Avendo un ingresso differenziale non fa più importanza distinguere tra terminale invertente e non invertente; scambiando gli ingressi si scambia anche l'inversione dell'uscita. Procediamo con un'analisi a modo comune:



Se si trascura l'effetto delle v_{ds} : $v_{id} = 0 \rightarrow i_{d_1} = i_{d_2}$

Operare a modo comune significa: $v_{i_1} = v_{i_2} = v_c$

Nella realtà, al posto del generatore di corrente si avrà uno specchio di corrente. Facendo l'analisi per piccolo segnale, si dovrà sostituire il generatore di corrente con l'equivalente dello specchio, che coprende anche la resistenza di uscita r_{os} (output source). Per l'analisi differenziale non occorre fare questa considerazione in quanto la tensione differenziale ai source rimane a massa, per cui non scorrebbe corrente differenziale all'interno di r_{os} .

Se la sorgente fosse ideale non si avrebbe variazione della corrente di polarizzazione. Per una sorgente reale come uno specchio, invece, si ha una variazione:

$$i_0 = i_{d_1} + i_{d_2} = \frac{v_s}{r_{os}}$$

La variazione della tensione di source:

$$v_s = v_{i_1} - v_{gs_1} = v_c - v_{gs_1}$$

Vogliamo in qualche modo dimostrare che la variazione della tensione di source è praticamente pari alla variazione di modo comune. Per farlo, esprimiamo v_{gs_1} in modo diverso.

$$i_{d_1} \cong g_m v_{gs_1} \rightarrow v_{gs_1} = \frac{i_{d_1}}{g_m} \rightarrow v_s = v_c - \frac{i_{d_1}}{g_m}$$

Possiamo legare poi la corrente i_{d_1} a quella della sorgente e quindi alla tensione v_s . Poiché la tensione di ingresso differenziale è nulla:

$$i_{d_1} + i_{d_2} = 2i_{d_1} = i_0 \rightarrow i_{d_1} = \frac{i_0}{2} = \frac{v_s}{2r_{os}}$$

Dunque:

$$v_s = v_c - \frac{i_{d_1}}{g_m} = v_c - \frac{v_s}{2r_{os}g_m} \rightarrow v_s = v_c \frac{1}{1 + \frac{1}{2g_m r_{os}}} \cong v_c$$

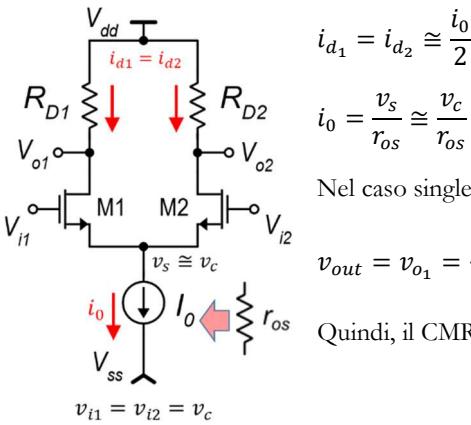
Nel caso peggiore la sorgente di corrente è uno specchio di corrente semplice, con resistenza di uscita pari a r_d . Dunque, il termine $2g_m r_{os} \gg 1$ e l'inverso può essere trascurato rispetto a 1.

Alle variazioni il potenziale di source segue il modo comune; se il modo comune sale o si abbassa di 1 Volt, il potenziale di source fa lo stesso. Ai grandi segnali, invece:

$$V_S = V_C - V_{GS}$$

Se il modo comune scende e di conseguenza il potenziale di source arriva alla V_{SS} , lo specchio di corrente che genera la corrente di polarizzazione si trova a tensione di uscita nulla, per cui non può erogare corrente. Il malfunzionamento si ha ancor prima, quando la $V_S - V_{SS}$ scende fino alla V_{min} dello specchio. Questo determina il limite inferiore della tensione di modo comune, che vale per qualunque amplificatore basato sulla coppia differenziale polarizzata da un generatore di corrente di tail. La relazione al piccolo segnale, inoltre, ci dice che le V_{GS} sono circa costanti (circa perché le variazioni non sono esattamente uguali). In realtà, a causa dell'effetto Body, che abbiamo trascurato, le V_{GS} subiscono una certa variazione.

Valutiamo il CMRR di questo amplificatore. A modo comune:



$$i_{d1} = i_{d2} \cong \frac{i_0}{2}$$

$$i_0 = \frac{v_s}{r_{os}} \cong \frac{v_c}{r_{os}}$$

Nel caso single ended a modo comune, sostituendo nell'espressione dell'uscita:

$$v_{out} = v_{o1} = -R_{D1}i_{d1} = R_{D1} \frac{i_0}{2} \cong -R_{D1} \frac{v_c}{2r_{os}} \rightarrow A_c \cong -\frac{R_{D1}}{2r_{os}}$$

Quindi, il CMRR dell'amplificatore differenziale a carico resistivo:

$$CMRR = \left| \frac{A_d}{A_c} \right| = \frac{g_m R_{D1}}{2} \cdot \frac{2r_{os}}{R_{D1}} = g_m r_{os}$$

Considerando che generalmente la corrente di polarizzazione è prodotta da uno specchio di corrente semplice la $r_{os} = r_d$, per cui il CMRR si attesta attorno a 40 dB . Tale valore di CMRR per un amplificatore differenziale è insufficiente per molte applicazioni. Questo anche perché molte volte il modo comune ha variazioni molto più grandi del modo differenziale all'ingresso. Un CMRR che inizia ad essere accettabile è 80 dB .

Se invece si considera un'uscita fully differential:

$$v_{out} = v_{od} = \frac{i_0}{2} R_{D2} - \frac{i_0}{2} R_{D1} = \frac{i_0}{2} (R_{D2} - R_{D1}) = \frac{v_s}{2r_{os}} (R_{D2} - R_{D1}) \cong \frac{v_c}{2r_{os}} (R_{D2} - R_{D1})$$

Quindi, il guadagno di modo comune nel caso di uscita fully diff:

$$A_{cd} = \left. \frac{v_{od}}{v_{ic}} \right|_{v_{id}=0} = -\frac{1}{2r_{os}} (R_{D2} - R_{D1})$$

E il CMRR:

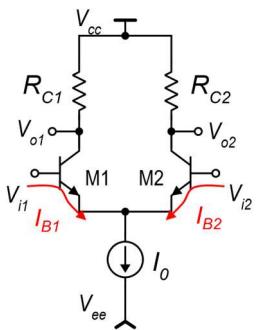
$$CMRR = \left| \frac{A_{dd}}{A_{cd}} \right| = \frac{g_m}{2} (R_{D2} + R_{D1}) \frac{2r_{os}}{(R_{D2} - R_{D1})} = g_m \bar{R}_D \frac{2r_{os}}{\Delta R_D}$$

Dove \bar{R}_D è il valore medio delle resistenze e ΔR_D è l'errore di matching. Il rapporto $\Delta R_D / \bar{R}_D$ è l'errore di matching relativo.

$$CMRR = 2g_m r_{os} \left(\frac{\Delta R_D}{\bar{R}_D} \right)^{-1}$$

Se le resistenze fossero perfettamente uguali, il CMRR risulterebbe infinito. Senza troppi sforzi a livello di layout, un errore relativo dell'1% di matching tra le resistenze si riesce ad ottenere. In tal caso il CMRR risulta circa uguale a 80 dB , adeguato per la maggior parte delle applicazioni in cui si richiede un amplificatore differenziale.

Amplificatore differenziale con carichi resistivi a BJT

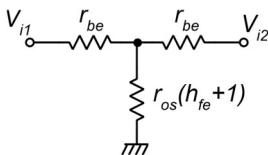


Per la versione a BJT le equazioni sono formalmente le stesse. L'unica differenza sostanziale è data dalla presenza delle correnti di base, che impattano sui generatori che producono V_{i_1} e V_{i_2} .

$$I_{B1} = \frac{I_{C1}}{\beta} \quad I_{B2} = \frac{I_{C2}}{\beta}$$

Poiché l'ingresso differenziale è prodotto da un generatore reale di tensione, le correnti di base di polarizzazione scorrono attraverso la resistenza interna del generatore. La caduta $R_s I_B$ fa sì che alle basi dei transistori non arrivi tutta la tensione del generatore.

Dunque, la presenza delle correnti di bias (l'effetto caricante del BJT) può alterare la tensione che si intende processare attraverso l'amplificatore. Dal punto di vista del piccolo segnale, la corrente di base causa una resistenza di ingresso finita che può causare una significativa attenuazione delle sorgenti di segnale V_{i_1} e V_{i_2} se la loro resistenza interna è significativa.

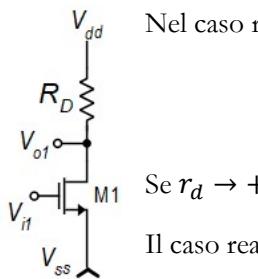


A fianco il circuito equivalente dai terminali di ingresso. Il generatore reale di tensione differenziale aggancia i suoi terminali a V_{i_1} e V_{i_2} . Chiamando R_V la resistenza che il generatore vede guardando verso l'amplificatore e R_{out} la sua resistenza interna:

$$V_{in} = V_{i_1} - V_{i_2} = V_D \frac{R_V}{R_V + R_{out}}$$

Guadagno massimo

Consideriamo un semplice stadio a source comune (amplificatore unipolare):



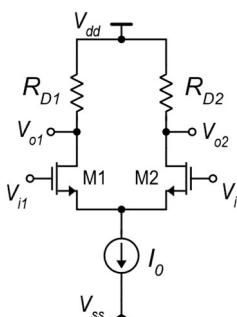
Nel caso reale in cui non si tracura la r_d il guadagno:

$$A_v = \frac{V_{o_1}}{V_{i_1}} = -g_m R_D \parallel r_d$$

Se $r_d \rightarrow +\infty$ (per MOSFET con L grandi $\rightarrow \lambda$ piccoli), il guadagno è approssimabile a $-g_m R_D$

Il caso reale con r_d pone un limite a quanto grande si può fare R_D per aumentare il guadagno.

Per l'amplificatore differenziale single-ended e fully differential, passando al caso nominale $R_{D1} = R_{D2} = R_D$ si ha rispettivamente:



$$A_d = -R_{D1} \frac{g_m}{2} \rightarrow |A_d| = R_D \frac{g_m}{2}$$

$$A_{dd} = -g_m \frac{(R_{D1} + R_{D2})}{2} \rightarrow |A_{dd}| = g_m R_D$$

Consideriamo il guadagno $g_m R_D$:

$$g_m = \frac{I_{DQ}}{V_{TE}} \rightarrow g_m R_D = \frac{R_D I_{DQ}}{V_{TE}}$$

Si ottiene che il fattore di guadagno $g_m R_D$ è il rapporto tra due tensioni: al numeratore compare la caduta di tensione statica attraverso la resistenza R_D , al denominatore compare la V_{TE} . Considerando di dover lasciare un margine per la V_{min} dello specchio che genera la I_0 e per la V_{DS} in modo tale che il dispositivo continui ad operare in saturazione durante tutta l'escursione del segnale di uscita, la massima caduta attraverso la R_D :

$$R_D I_D \leq V_{DD} - V_{SS} - V_{DS} - V_{min} \rightarrow R_D I_D < V_{DD} - V_{SS}$$

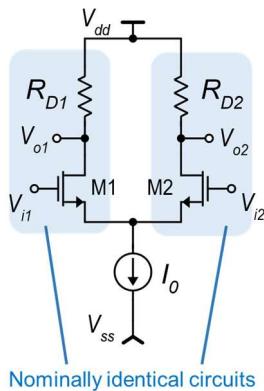
La V_{TE} al limite della forte inversione vale circa 50 mV ($V_{TE} = (V_{GS} - V_t)/2$). Per migliorare ulteriormente il guadagno si potrebbe arrivare fino alla debole inversione, in cui $V_{TE} = mV_T \cong 40$ mV. Con il bipolare, che ha la V_{TE} più piccola in assoluto, pari alla V_T , si può migliorare ulteriormente. Ad ogni modo, dalla precedente disequazione si deduce che l'amplificazione è limitata dall'alimentazione disponibile. Per l'amplificatore differenziale fully diff si ottiene che:

$$|A_{dd}| < \frac{V_{DD} - V_{SS}}{V_{TE}}$$

Se si considera $V_{DD} - V_{SS} = 3V$ e $V_{TE} = 50$ mV, il guadagno $|A_{dd}| < 60$. Utilizzando i BJT si ottiene $V_{TE} = V_T = 25$ mV, per cui si può spingere il guadagno fino a $|A_{dd}| < 120$. Per di più, il caso più comune è quello di scegliere la caduta sulla R_D pari alla metà dell'escursione della tensione di alimentazione. Ad ogni modo, il guadagno diminuisce all'aumentare dell'overdrive e al diminuire delle tensioni di alimentazione. Soprattutto a causa della tendenza della diminuzione della tensione di alimentazione nei circuiti integrati, gli amplificatori differenziali con carico resistivo non sono spesso utilizzati proprio per questo limite.

Tensione di offset di un amplificatore differenziale con carico resistivo

Consideriamo l'amplificatore fully differential a MOSFET e con carico resistivo (in forte inversione)



La tensione di offset è un parametro fondamentale per un amplificatore. Si definisce tensione di offset la tensione differenziale da applicare in ingresso per annullare l'uscita:

$$V_{io} = V_D |_{V_{OD}=0}$$

Se il circuito fosse perfettamente simmetrico, per avere uscita differenziale nulla l'ingresso differenziale da applicare dovrebbe anch'esso essere nullo, per cui non si avrebbe offset. L'offset nasce dal mismatch tra i carichi R_{D1} e R_{D2} e tra i dispositivi M1 e M2, a due a due progettati per essere nominalmente identici tra loro.

La tensione di uscita:

$$V_{out} = V_{o1} - V_{o2}$$

La tensione differenziale in ingresso:

$$V_D = V_{GS_1} - V_{GS_2} = \Delta V_{GS}$$

Nel caso reale la condizione di annullamento della tensione differenziale in uscita si verifica per una tensione V_D di ingresso non nulla. L'errore di matching tra i due rami, quindi, si manifesta anche come differenza tra le V_{GS} nel caso di annullamento dell'uscita. In forte inversione:

$$V_{GS} = V_t + \sqrt{\frac{2I_D}{\beta}}$$

$$\begin{cases} A = V_t \\ B = \sqrt{\frac{2I_D}{\beta}} = \sqrt{2} I_D^{1/2} \beta^{-1/2} \end{cases} \rightarrow V_{GS} = A + B \rightarrow \Delta V_{GS} = \Delta A + \Delta B$$

$$\begin{cases} \Delta A = \Delta V_t \\ \Delta B = B \cdot \frac{\Delta B}{B} = B \left(\frac{1}{2} \frac{\Delta I_D}{I_D} - \frac{1}{2} \frac{\Delta \beta}{\beta} \right) \end{cases}$$

Dunque, riscrivendo la tensione differenziale di ingresso V_D :

$$V_D = \Delta V_{GS} = \Delta V_t + \sqrt{\frac{2I_D}{\beta}} \left(\frac{1}{2} \frac{\Delta I_D}{I_D} - \frac{1}{2} \frac{\Delta \beta}{\beta} \right)$$

La tensione ricavata non è ancora la tensione di offset. Questo perché non si è ancora applicata la condizione di annullamento della tensione differenziale di uscita. Inoltre, si nota la presenza della I_D , che non è un vero parametro tecnologico, quanto una soluzione del circuito. Quindi, l'errore $\Delta I_D/I_D$ rappresenta un parametro libero con cui specializzare l'equazione appena ottenuta. La stessa equazione potrebbe essere utilizzata per calcolare la tensione di offset speciale anziché quella classica, che corrisponde alla tensione differenziale da porre in ingresso per ottenere correnti uguali nei due rami. In tal caso si dovrebbe fissare $\Delta I_D/I_D = 0$. Oppure, si potrebbe voler calcolare la tensione differenziale da porre in ingresso per avere un discostamento tra le correnti nei due rami del 10%. Poiché l'equazione è frutto di un'approssimazione di Taylor al prim'ordine, si possono considerare soltanto piccole variazioni per mantenere una buona accuratezza del risultato. In altri termini, la formula rappresenta la tensione differenziale da applicare in ingresso, con un certo errore di matching tra i dispositivi, per avere un certo errore tra le correnti nei rami. Per ottenere la vera tensione di offset occorre prima capire qual è il rapporto $\Delta I_D/I_D$.

La condizione che la tensione differenziale di uscita sia nulla con correnti uguali, infatti, è vera soltanto se anche le resistenze R_{D_1} e R_{D_2} sono uguali: a parità di correnti si avrebbe la stessa caduta su entrambi i lati. La $\Delta I_D/I_D$ giusta da inserire nell'espressione è quella che annulla la tensione di uscita tenendo conto anche del mismatch tra le resistenze. Dunque, cerchiamo di esprimere $\Delta I_D/I_D$ in funzione del mismatch tra le resistenze.

$$V_{out} = V_{OD} = R_{D_2} I_{D_2} - R_{D_1} I_{D_1}$$

Per grandi segnali, l'uscita è la differenza tra le cadute sulle resistenze. Anche questa espressione ha il carattere di un errore di matching tra due grandezze analoghe del tipo $R_D I_D$. Definendo:

$$Z = R_D I_D \rightarrow \Delta Z = V_{out}$$

Annulare la tensione di uscita differenziale equivale a ricercare l'annullamento dell'errore di matching ΔZ . Dunque:

$$\Delta Z = Z \cdot \frac{\Delta Z}{Z} = R_D I_D \left(\frac{\Delta R_D}{R_D} + \frac{\Delta I_D}{I_D} \right) = 0$$

Poiché $R_D I_D$ è il valore nominale, Z non può essere nullo, per cui l'equazione ha soluzione quando $\frac{\Delta Z}{Z} = 0$.

$$\frac{\Delta Z}{Z} = 0 \rightarrow \frac{\Delta R_D}{R_D} + \frac{\Delta I_D}{I_D} = 0 \rightarrow \frac{\Delta I_D}{I_D} = -\frac{\Delta R_D}{R_D}$$

Si trova così l'espressione di $\Delta I_D/I_D$ che sostituita nell'equazione della V_D precedente fa sì che la V_D coincida effettivamente alla tensione di offset dell'amplificatore. Applicando la sostituzione:

$$V_{io} = \Delta V_t + \sqrt{\frac{2I_D}{\beta}} \left(-\frac{1}{2} \frac{\Delta R_D}{R_D} - \frac{1}{2} \frac{\Delta \beta}{\beta} \right)$$

Sa i valori con i quali si effettua lo sviluppo di Taylor per la determinazione degli errori sono i valori nominali, la radice quadrata esprime la tensione di overdrive in forte inversione. Pertanto:

$$V_{io} = \Delta V_t - \frac{V_{GS} - V_t}{2} \left(\frac{\Delta R_D}{R_D} + \frac{\Delta \beta}{\beta} \right)$$

Infine, si può determinare la deviazione standard della tensione di offset:

$$\sigma_{V_{io}} = \sqrt{\sigma_{V_t}^2 + \left[\frac{\sigma_{\Delta\beta}}{\beta} \frac{(V_{GS} - V_t)}{2} \right]^2 + \left[\frac{\sigma_{\Delta R_D}}{R_D} \frac{(V_{GS} - V_t)}{2} \right]^2}$$

In fase di progetto possiamo valutare la sola distribuzione statistica della tensione di offset. In forte inversione i termini $(V_{GS} - V_t)/2$ equivalgono alla tensione termica efficace V_{TF} . Le varianze che compaiono sotto la radice sono tutte funzioni dell'area del dispositivo, per cui, aggiustando l'area dei dispositivi, si riesce a determinare una certa statistica della tensione di offset. Anche nel caso della tensione di offset si ottiene un risultato tanto migliore quanto la tensione di overdrive è piccola. Nello specchio di corrente, invece, per avere una buona accuratezza tra la corrente di ingresso e la corrente di uscita si deve impiegare $V_{GS} - V_t$ grandi.

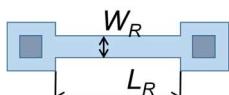
Ruolo dei parametri

Cerchiamo di studiare il ruolo dei diversi parametri che compongono la formula della deviazione standard della tensione di offset.

Resistori

I resistori hanno un valore nominale pari a R_D . La resistenza R_D è determinata dal rapporto tra la larghezza e la lunghezza del corpo del resistore, W_R/L_R . Solitamente Cadence imposta la W_R al valore minimo e calcola la L_R in funzione del valore di resistenza per minimizzare le dimensioni del resistore.

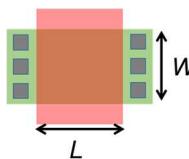
L'area del corpo del resistore determina la deviazione standard dell'errore di matching relativo tra i resistori.



$$\sigma_{\frac{\Delta R_D}{R_D}} = \frac{C_R}{\sqrt{W_R L_R}}$$

Scegliere la W_R minima per minimizzare la L_R e minimizzare l'ingombro totale del resistore può comportare una deviazione standard $\sigma_{\frac{\Delta R_D}{R_D}}$ troppo elevata, dunque una tensione di offset inaccettabile.

Transistori

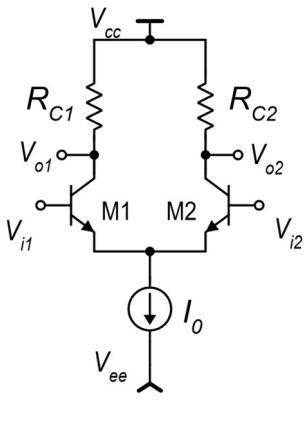


Il rapporto W/L determina il β dei transistori il quale, se è nota la I_0 , determina l'overdrive $V_{GS} - V_t$. Le deviazioni standard della tensione di soglia e dell'errore relativo sul β dipendono entrambi dall'area del MOSFET, ovvero dal prodotto WL .

$$\begin{cases} \sigma_{V_t} = \frac{C_{V_t}}{\sqrt{WL}} \\ \sigma_{\frac{\Delta\beta}{\beta}} = \frac{C_\beta}{\sqrt{WL}} \end{cases}$$

Determinati i valori dei rapporti W_R/L_R e W/L per ottenere la resistenza desiderata e l'overdrive più piccolo possibile, la varianza della tensione di offset dipende soltanto dall'area dei resistori e dall'area dei MOSFET. Il problema del determinare singolarmente l'area dei resistori e l'area dei transistori non si può risolvere esattamente poiché si dispone di una sola equazione con due incognite. Ingegneristicamente, si può dare la stessa area ai resistori e ai MOSFET, per ridurre l'incognita a una soltanto. La soluzione ottima che minimizza l'area complessiva non per forza coincide a quella con aree uguali. In generale, un amplificatore a basso offset beneficia di aree maggiori; si tratta di un'altra prova del fatto che i circuiti analogici non beneficiano della miniaturizzazione spinta come i circuiti digitali.

Tensione di offset dell'amplificatore differenziale a carico resistivo a BJT



Analogamente a prima:

$$V_{io} = V_{ID}|_{V_{OD}=0}$$

$$V_{OD} = V_{o_1} - V_{o_2} = R_{c_1}I_{c_1} - R_{c_2}I_{c_2}$$

$$V_{i_D} = V_{i_1} - V_{i_2} = V_{BE_1} - V_{BE_2} = \Delta V_{BE}$$

La V_{BE} ha la seguente espressione:

$$V_{BE} = V_T \ln\left(\frac{I_C}{I_S}\right)$$

Ipotizzando che Se Q1 e Q2 operino alla stessa temperatura la loro V_T è la stessa, per cui:

$$V_{id} = \Delta V_{BE} = V_T \left(\frac{\Delta I_C}{I_C} - \frac{\Delta I_S}{I_S} \right)$$

Come per l'analisi del MOSFET si impone la tensione differenziale di uscita a 0:

$$V_{od} = 0 \rightarrow \Delta(R_C I_C) = 0 \rightarrow \frac{\Delta I_C}{I_C} = -\frac{\Delta R_C}{R_C}$$

Pertanto, si ottiene:

$$v_{io} = V_T \left(-\frac{\Delta R_C}{R_C} - \frac{\Delta I_S}{I_S} \right) \rightarrow \sigma_{V_{io}} = \sqrt{\left(V_T \sigma_{\frac{\Delta R_C}{R_C}} \right)^2 + \left(V_T \sigma_{\frac{\Delta I_S}{I_S}} \right)^2}$$

Dove:

$$\begin{cases} \sigma_{\frac{\Delta R_C}{R_C}} = \frac{C_R}{\sqrt{W_R L_R}} \\ V_T \sigma_{\frac{\Delta I_S}{I_S}} = \sigma_{\Delta V_{BE-elem}}^* = \frac{\sigma_{\Delta V_{BE-elem}}^*}{\sqrt{\text{area}}} \end{cases} \rightarrow \sigma_{V_{io}} = \sqrt{\left(V_T \sigma_{\frac{\Delta R_C}{R_C}} \right)^2 + \sigma_{\Delta V_{BE-elem}}^*}$$

Drift termico della tensione di offset (BJT)

È necessario valutare l'effetto della temperatura sulla tensione di offset.

$$V_{io} = \frac{kT}{q} \left(-\frac{\Delta R_C}{R_C} - \frac{\Delta I_S}{I_S} \right)$$

Nella parentesi si trovano rapporti di grandezze che hanno approssimativamente lo stesso coefficiente di temperatura; se varia la temperatura, il rapporto rimane lo stesso. Quindi, il termine nella parentesi, complessivamente, può essere considerato indipendente dalla temperatura. L'entità del drift della tensione di offset rispetto la temperatura si può valutare con la derivata rispetto alla temperatura:

$$\frac{dV_{io}}{dT} = \frac{k}{q} \left(-\frac{\Delta R_C}{R_C} - \frac{\Delta I_S}{I_S} \right)$$

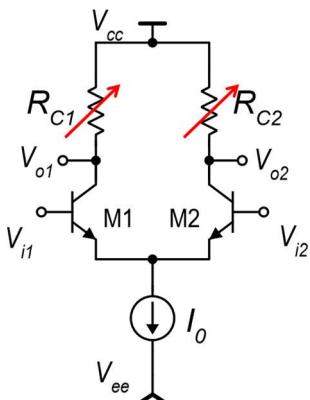
Moltiplicando e dividendo per T :

$$\frac{dV_{io}}{dT} = \frac{kT}{q} \left(-\frac{\Delta R_C}{R_C} - \frac{\Delta I_S}{I_S} \right) \frac{1}{T} = \frac{V_{io}}{T}$$

La deriva della tensione di offset con la temperatura si ottiene facilmente misurando la tensione di offset e dividendo per la temperatura di esercizio. Considerando ad esempio una tensione di offset di $V_{io} = 1mV$ e una temperatura $T = 300 ^\circ K$ si ottiene una deriva di:

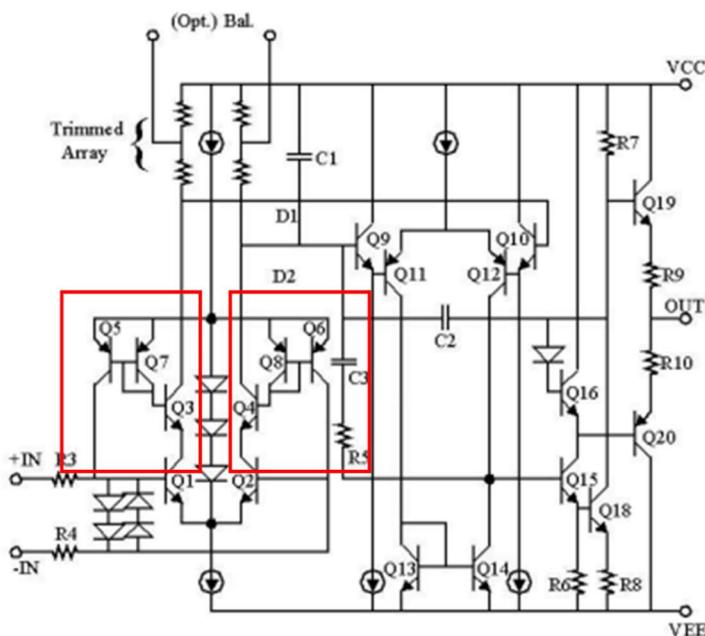
$$\frac{dV_{io}}{dT} = \frac{V_{io}}{T} = \frac{1 \cdot 10^{-3}}{300} = 3.3 \mu V/K$$

Se si riuscisse ad ottenere una tensione di offset $V_{io} = 0$, anche la deriva termica dell'offset sarebbe nulla. Per ridurre la tensione di offset ulteriormente avvicinandola al valore nullo si possono regolare i resistori con la tecnica del laser trimming.



La tecnica del laser trimming consiste nel consumare i resistori con un laser in modo tale da compensare l'effetto dell'offset. Un'altra tecnica consiste nel connettere dinamicamente resistori in serie o in parallelo ai resistori dei due rami della coppia attraverso interruttori (anche digitali) in modo tale da regolare la resistenza equivalente affinché l'offset si aggiusti. Mentre per il BJT annullare l'offset comporta anche un annullamento della sua deriva in temperatura, per un amplificatore differenziale a MOSFET non vale lo stesso.

Amplificatore operazionale OP07



Lo stadio di uscita del OP07 è abbastanza simile a quello del $\mu A741$. Lo stadio di ingresso, invece, è diverso e nel caso dell'OP07 presenta un amplificatore differenziale con carico resistivo. I resistori sono regolati con il laser trimming per ottenere una bassa tensione di offset.

Le correnti di base di dispositivi di ingresso Q1 e Q2 sono le stesse di Q3 e Q4. Le correnti di base di Q3 e Q4 vengono poi specchiate dagli specchi Q5 Q7 e Q6 Q8 e reimessa nei terminali di ingresso. In questo modo si riducono le correnti di polarizzazione in ingresso. Si parla di circuito di cancellazione delle correnti di polarizzazione. La sorgente esterna dovrà fornire soltanto la parte mancante di corrente.

Confrontando l'OP07 con il uA741:

Table 1.

Parameter	Symbol	Conditions	Min	Typ	Max	Unit
INPUT CHARACTERISTICS						
$T_A = 25^\circ\text{C}$						
Input Offset Voltage ¹	V_{os}		30	75		μV
Long-Term V_{os} Stability ²	V_{os}/Time		0.3	1.5		$\mu\text{V}/\text{Month}$
Input Offset Current	I_{os}		0.5	3.8		nA
Input Bias Current	I_B		± 1.2	± 4.0		nA
$0^\circ\text{C} \leq T_A \leq 70^\circ\text{C}$						
Input Offset Voltage ¹	V_{os}		45	130		μV
Voltage Drift Without External Trim ⁴	TCV_{os}		0.3	1.3		$\mu\text{V}/^\circ\text{C}$
Voltage Drift with External Trim ³	TCV_{osN}	$R_p = 20 \text{ k}\Omega$	0.3	1.3		$\mu\text{V}/^\circ\text{C}$

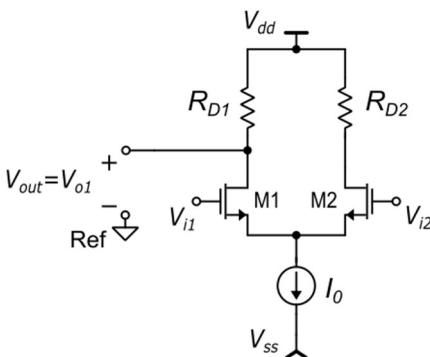
OP07

μA741

PARAMETER	TEST CONDITIONS ⁽¹⁾	MIN	Typ	MAX	UNIT
V_{io} Input offset voltage	$V_o = 0$	25°C	1	6	mV
		Full range		7.5	
$\Delta V_{io(\text{adj})}$ Offset voltage adjust range	$V_o = 0$	25°C		± 15	mV
		25°C	20	200	
I_{io} Input offset current	$V_o = 0$	25°C		300	nA
		Full range			
I_B Input bias current	$V_o = 0$	25°C	80	500	nA
		Full range		800	

La tensione di offset dell'amplificatore OP07 è significativamente migliore di quella del uA741.

Tensione di offset nel caso di amplificatore differenziale single ended

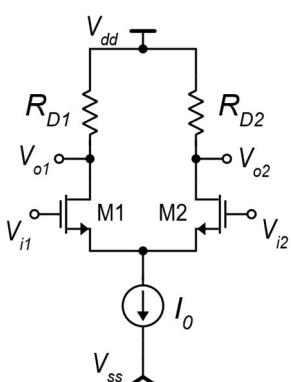


Nel caso di uscita single ended nell'espressione dell'uscita i parametri dei dispositivi non compaiono come differenze di matching.

$$V_{out} = V_{o1} = V_{dd} - R_1 I_{D1}$$

Questo fa sì che la tensione di uscita sia principalmente influenzata da errori globali piuttosto che da quelli di matching. Pertanto, l'amplificatore differenziale single ended è soggetto a un offset più grande rispetto all'amplificatore differenziale fully diff.

Limitazioni dell'amplificatore differenziale a MOSFET con carico resistivo



L'amplificatore differenziale a MOSFET con carichi resistivi presenta alcune problematiche:

- La versione single-ended presenta un guadagno di modo comune elevato (basso CMRR) e una tensione di offset elevata. Occorre una versione alternativa di amplificatore differenziale single-ended che si comporti meglio da questi punti di vista.
- Entrambe le configurazioni single ended e fully differential non permettono di raggiungere guadagni di tensione elevati quando operano con tensione di alimentazione ridotta.

Soluzione alla prima problematica

Nel caso nominale si ha $R_{D1} = R_{D2} = R_D$, per la configurazione fully differential si ha:

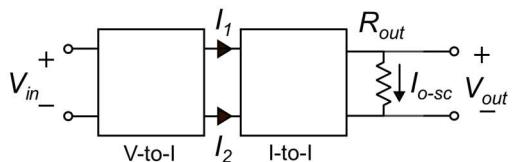
$$V_{oD} = R_D (I_{D2} - I_{D1})$$

La tensione differenziale in uscita è data dal prodotto tra R_D e la differenza tra le correnti. Questo fa sì che sia nel caso di offset, in cui per $V_D = 0$ la tensione di uscita dipende da soli errori di matching, che nel caso del modo comune, in cui le correnti nei due rami tendono ad essere uguali tra loro, si cancelli la tensione di uscita. Per riciclare questo vantaggio ed applicarlo ad una nuova versione single ended si può prima fare la differenza tra le correnti e poi veicolarla attraverso una resistenza. In questo modo si ottengono amplificatori differenziali single-ended con performance di offset e CMRR paragonabili a quelle dei fully differential.

Soluzione alla seconda problematica

Per risolvere questa problematica, anziché mandare la corrente all'interno di una resistenza statica la si manda dentro una resistenza differenziale elevata. In questo modo si ottengono guadagni molto più elevati.

Amplificatori a singolo stadio



La definizione di amplificatore a singolo stadio è una definizione operativa, il cui risultato si apprezza nel momento in cui si assemblano più stadi.

Nella trattazione che faremo si considera un amplificatore a singolo stadio come la cascata di due blocchi funzionali. Il primo dei due opera una conversione della tensione di ingresso in una o più correnti; tipicamente è una coppia differenziale. Le due correnti, che nelle configurazioni studiate andavano nelle resistenze, ora entrano all'interno di una rete con ingresso e uscita in corrente che può farne la somma o la differenza.

Considerando un comportamento linearizzato del circuito, la tensione in ingresso viene convertita in una differenza di correnti:

$$I_1 - I_2 = G_{m_1} V_{in}$$

Dove G_{m_1} è la funzione di trasferimento del blocco complessivo. Esistono anche stadi unipolari che convertono una tensione in una singola corrente anziché una differenza tra correnti. In questo caso, uno dei terminali di ingresso e/o uscita potrebbe essere anche ground; il circuito è del tutto generale. Tale cascata potrebbe essere utilizzata per sottrarre la corrente di riposo $I_0/2$ lasciando solo la componente di variazione. Tipicamente, per capire qual è la corrente prodotta in uscita si pone un cortocircuito all'uscita del blocco I-I:

$$I_{o-sc} = k_1(I_1 - I_2) = k_1 G_{m_1} V_{in}$$

Si potrebbe anche avere una somma tra correnti. Dopo di che, la corrente di uscita viene condotta nella resistenza ai terminali di uscita per produrre la tensione di uscita. Non è detto che la resistenza di uscita debba essere implementata da un resistore fisico. Potrebbe trattarsi invece della resistenza vista di un oggetto attivo ed è questo il caso più comune.

$$V_{out} = I_{o-sc} R_{out} = k_1 G_{m_1} V_{in} R_{out}$$

Si definisce il rapporto tra la corrente di cortocircuito e la V_{in} :

$$G_m = k_1 G_{m_1} \rightarrow I_{o-sc} = G_m V_{in}$$

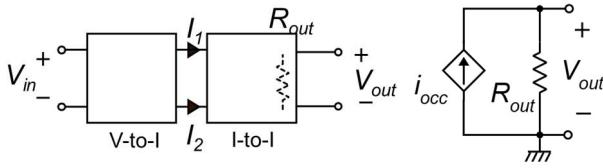
Il parametro G_m è il parametro di conversione della tensione di ingresso nella corrente di cortocircuito. Se internamente ci fosse un'altra conversione da corrente a tensione e poi tensione corrente, non si tratterebbe più un amplificatore single stage. La tensione di uscita:

$$V_{out} = G_m V_{in} R_{out} \rightarrow A = \frac{V_{out}}{V_{in}} = G_m R_{out}$$

Gli unici nodi ad alta impedenza rispetto ground, a parte gli ingressi, sono quelli di uscita. Poiché i poli della rete sono grosso modo dati da R_V per la capacità, con un unico nodo ad alta impedenza si ha un unico polo dominante. Un amplificatore a due stadi ha due nodi ad alta impedenza: l'uscita del primo stadio e l'uscita del secondo. Quindi, un amplificatore a due stadi presenta due poli singoli. Un amplificatore operazionale a singolo stadio ha vantaggi notevoli per quanto riguarda la compensazione, è stabile in reazione anche se più lento.

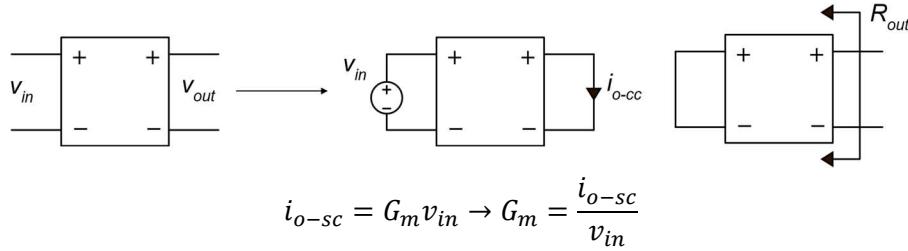
Metodo per il calcolo del guadagno di un amplificatore a singolo stadio

Spesso, nei circuiti di interesse, la resistenza che produce la tensione di uscita è la resistenza di uscita della rete di conversione della corrente (resistenza differenziale).



Generalmente, in un amplificatore a singolo stadio è semplice calcolare la corrente di cortocircuito in funzione della tensione di ingresso. Dunque, valutando anche la resistenza differenziale all'uscita del blocco di conversione della corrente, si può ricorrere a un'equivalente di Norton.

Per calcolare il guadagno di un amplificatore a singolo stadio, dal circuito alle variazioni si ricava G_m dalla relazione lineare tra la corrente e la tensione di ingresso.



Determinando la R_{out} il calcolo del guadagno è concluso.

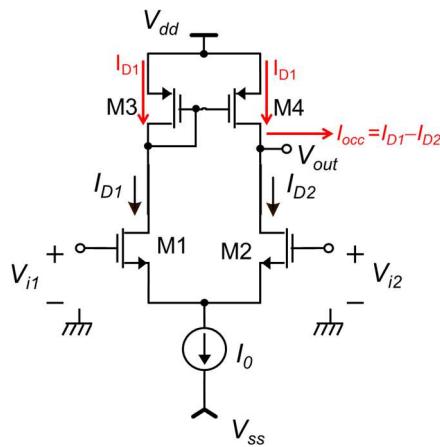
$$v_{out} = R_{out} i_{o-sc} = R_{out} G_m v_{in} \rightarrow A = \frac{v_{out}}{v_{in}} = G_m R_{out}$$

Tale approccio è valido per qualsiasi amplificatore, ma risulta particolarmente vantaggioso nel caso dell'amplificatore a singolo stadio perché solitamente è semplice calcolarne separatamente la corrente di cortocircuito e la resistenza di uscita, individuarne il convertitore V-I e la rete di trasporto della corrente.

Amplificatore differenziale a MOSFET con carico a specchio

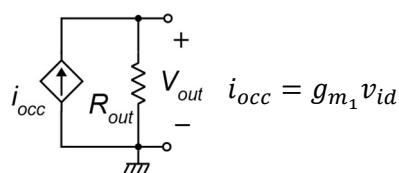
Si vogliono centrare le seguenti specifiche:

- Configurazione single-ended
- Alto CMRR ($> 80 dB$)
- Alto guadagno ($\sim 40 dB$) anche per basse tensioni di alimentazione $V_{dd} - V_{ss}$

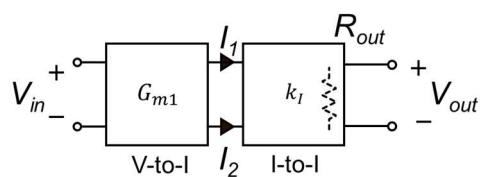
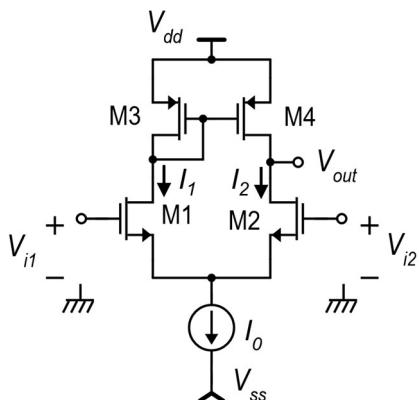


Per realizzare la differenza $I_{D1} - I_{D2}$, rispetto all'amplificatore differenziale con carico resistivo, si può impiegare uno specchio di corrente di tipo p.

Analizziamo il circuito con un equivalente di Norton:



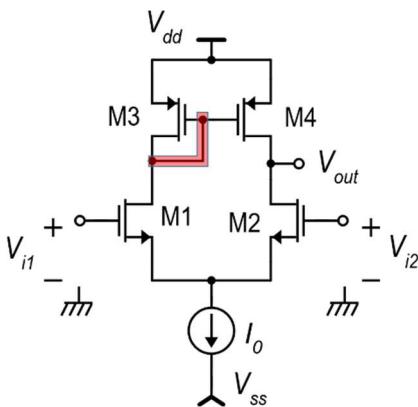
Lo specchio inverte il segno della corrente I_{D_1} in modo da poter effettuare la differenza tra le correnti dei rami. Non si aggiunge alcuna resistenza all'uscita perché finirebbe in parallelo alla resistenza di uscita dell'amplificatore, abbattendone il guadagno. In questo caso il G_m coincide al g_m . Analizziamo le singole sotto-unità che compongono il circuito.



La coppia differenziale costituisce la rete V-I, mentre lo specchio costituisce la rete I-I di trasporto. Come dimostreremo in seguito:

$$G_{m1} = g_{m1} = g_{m2} \quad k_I = 1$$

Studio del punto di riposo



Per l'analisi del punto di riposo si assume tensione differenziale nulla:

$$V_{id} = 0$$

Sopravvive però una componente di modo comune non nulla atta alla polarizzazione di M1 e M2.:

$$V_{i_1} = V_{i_2} = V_C$$

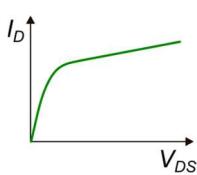
Lo stimolo a riposo, quindi, è elettricamente simmetrico, ma il circuito non è topologicamente simmetrico a causa del lacchetto su M3. Tuttavia, per la soluzione elettrica del circuito vale la simmetria: i transistori M1/M2 e M3/M4 hanno le stesse tensioni e le stesse correnti.

Consideriamo per prima un'analisi nominale dei transistori M1 e M2. Se si riesce a dimostrare che M1 e M2 hanno le stesse V_{GS} , V_{BS} , V_{DS} , a prescindere dalla simmetria topologica le correnti nei due MOSFET saranno le stesse. Per M1 e M2:

$$\begin{cases} V_{iD} = 0 \rightarrow V_{GS_1} = V_{GS_2} \\ V_{S_1} = V_{S_2} \rightarrow V_{BS_1} = V_{BS_2} \\ V_{DS_1} ? V_{DS_2} \end{cases}$$

Per i transistori M3 e M4:

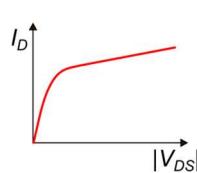
$$\begin{cases} V_{GS_3} = V_{GS_4} \\ V_{BS_3} = V_{BS_4} \\ V_{DS_1} ? V_{DS_2} \end{cases}$$



Inoltre, vale sempre che:

$$\begin{cases} I_{D_3} = I_{D_1} \\ I_{D_4} = I_{D_2} \end{cases}$$

Con V_{GS} e V_{BS} fissate, la I_D è una funzione monotona della V_{DS} (o della $|V_{DS}|$ per un pMOS). Dunque, il transistore dei due che ha la V_{DS} maggiore è quello che conduce la corrente maggiore.



Applicando il secondo principio di Kirchoff:

$$V_{DS_1} + |V_{DS_3}| = V_{DS_2} + |V_{DS_4}|$$

L'obiettivo è quello di dimostrare che $I_{D_1} = I_{D_2}$. Procediamo con una dimostrazione per assurdo, assumendo che $I_{D_1} > I_{D_2}$. Per quanto affermato in precedenza:

$$I_{D_1} > I_{D_2} \rightarrow V_{DS_1} > V_{DS_2}$$

Analogamente, per la solita ipotesi si avrebbe:

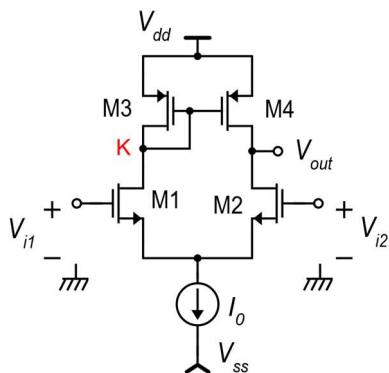
$$I_{D_3} > I_{D_4} \rightarrow |V_{DS_3}| > |V_{DS_4}|$$

Dunque, dall'ipotesi per assurdo segue che:

$$V_{DS1} + |V_{DS_3}| > V_{DS_2} + |V_{DS_4}|$$

Poiché la relazione è in contraddizione con quella ricavata tramite la seconda legge di Kirchoff, ciò costituisce un assurdo della dimostrazione. Se si ripete lo stesso in modo duale ipotizzando per assurdo che $I_{D_1} < I_{D_2}$ si ottiene comunque una contraddizione. Quindi, l'unica soluzione che si può avere (nel caso ideale, senza considerare i mismatch) è la seguente:

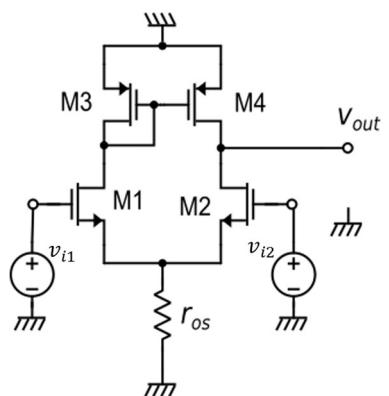
$$\begin{cases} I_{D_1} = I_{D_2} \\ I_{D_3} = I_{D_4} \end{cases} \rightarrow \begin{cases} V_{DS_1} = V_{DS_2} \\ |V_{DS_3}| = |V_{DS_4}| \end{cases}$$



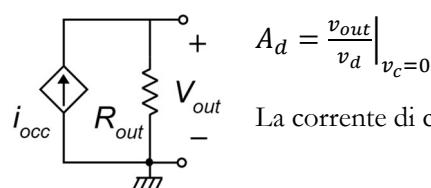
$$\begin{cases} V_K = V_{dd} - |V_{DS_3}| \\ V_{out} = V_{dd} - |V_{DS_4}| \end{cases} \rightarrow V_{out} = V_K$$

Il lacchetto non disturba la simmetria della soluzione elettrica del circuito, ma al piccolo segnale si rivelerà significativo. Poiché a riposo la tensione di uscita è uguale alla tensione al nodo K si dimostra così che malgrado la non simmetria topologica, la soluzione elettrica del circuito a riposo è perfettamente simmetrica, per cui $g_{m_1} = g_{m_2}$.

Guadagno a modo differenziale



Applichiamo la tecnica precedentemente discussa precedentemente che consente nel ricavare l'equivalente di Norton alle variazioni dell'uscita.



La corrente di cortocircuito dipende da altre quantità:

$$i_{occ} = G_m v_d$$

Calcolo del G_m

$$v_{out} = i_{occ} R_{out}$$

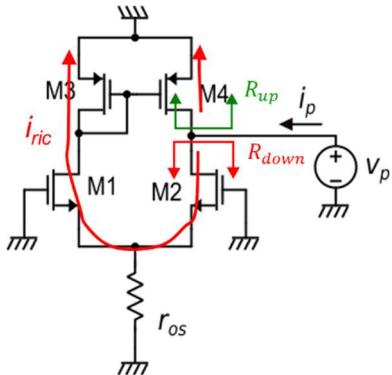
$$i_{occ} = i_{d_1} - i_{d_2} \cong i_{d_1} - i_{d_2} = g_{m_1} v_d \rightarrow G_m = g_{m_1}$$

Il guadagno a modo differenziale:

$$v_{out} \equiv g_m, v_d R_{out} \rightarrow A_d \equiv g_m, R_{out}$$

Resta di calcolare la resistenza di uscita R_{out} .

Calcolo della resistenza di uscita



Si collega un generatore di prova all'uscita. A prima vista potremmo pensare che la resistenza di uscita sia determinata da:

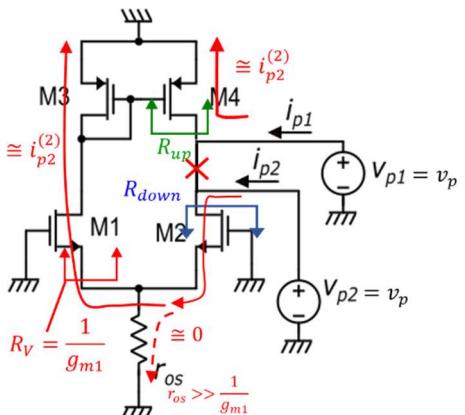
$$R_{out} = R_{up} \parallel R_{down}$$

Tale assunzione è però falsa. Infatti, la corrente differenziale di sotto, uscita dal source di M2 entra principalmente nel source di M1 (da cui si vede $1/g_{m_1}$). Da qui la corrente entra in M3 e viene specchiata sul ramo di destra. Dunque, la corrente di prova i_p non è dovuta soltanto alla corrente in R_{up} e alla corrente in R_{down} , ma anche a questa corrente di ricircolo.

La corrente i_p :

$$i_p = \frac{v_p}{R_{up}} + \frac{v_p}{R_{down}} + i_{ric}$$

Quindi, non è sempre vero che la resistenza vista dal generatore è data dal parallelo delle resistenze viste dai vari rami. I rami potrebbero interagire l'uno con l'altro: la corrente di uno potrebbe essere riportare nell'altro. Il metodo migliore per determinare la resistenza di uscita è quello di porre due generatori di prova v_{p_1} e v_{p_2} , entrambi pari a v_p .



I terminali positivi dei due generatori di prova hanno lo stesso potenziale, per cui si può tagliare il ramo che li connette. Poiché il circuito è lineare alle variazioni si può applicare il principio di sovrapposizione degli effetti: si calcolano le i_{p_1} e le i_{p_2} prima nel caso v_{p_1} acceso e v_{p_2} spento, poi nel caso v_{p_1} spento e v_{p_2} acceso, sommando tutti i contributi per risalire alla i_p .

- $v_{p_1} = v_p, v_{p_2} = 0 \rightarrow i_{p_1} = i_{p_1}^{(1)}, i_{p_2} = i_{p_2}^{(1)}$
 - $v_{p_1} = v_p, v_{p_2} = 0 \rightarrow i_{p_1} = i_{p_1}^{(2)}, i_{p_2} = i_{p_2}^{(2)}$

$$i_p = i_{p_1}^{(1)} + i_{p_2}^{(1)} + i_{p_1}^{(2)} + i_{p_2}^{(2)}$$

Con v_{p_1} acceso e v_{p_2} spento:

Si ha un unico ramo che richiude la corrente verso ground, che è l'uscita dello specchio. Tale corrente non causa una corrente i_{p_2} nell'altro generatore poiché entra nello specchio dallo slave. La R_{up} è pari alla resistenza di uscita dello specchio semplice, quindi:

$$i_{p_1}^{(1)} = \frac{v_p}{R_{up}} = \frac{v_p}{r_{d_1}} \quad i_{p_2}^{(1)} = 0$$

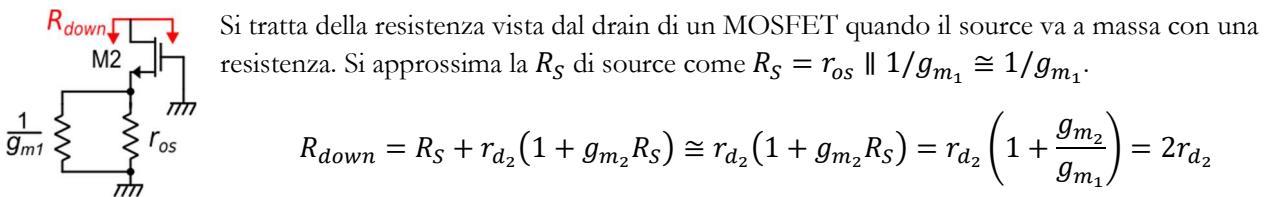
Con v_{p_1} spento e v_{p_2} acceso:

La corrente che esce dal source di M2 vede una resistenza pari a $r_{os} \parallel 1/g_{m_1}$. Approssimativamente tutta la i_{p_2} entra nel drain di M3 e viene specchiata dalla parte di M4 con lo stesso verso di i_{p_1} .

$$i_{p_1}^{(2)} \cong i_{p_2}^{(2)}$$

Cioè, i due generatori, quando agisce v_n , sono attraversati dalla stessa corrente.

Per valutare la R_{down} :



Il risultato ottenuto vale per la coppia differenziale in generale: quando si guarda dal drain di uno dei sue MOSFET della coppia verso ground si vede sempre $2r_d$. Sommando i termini, la corrente di prova totale:

$$i_p = \frac{v_p}{r_{d_4}} + 0 + \frac{v_p}{2r_{d_2}} + \frac{v_p}{2r_{d_2}} = v_p \left(\frac{1}{r_{d_4}} + \frac{1}{r_{d_2}} \right)$$

Dunque, la resistenza di uscita:

$$R_{out} = \frac{v_p}{i_p} = \left(\frac{1}{r_{d_2}} + \frac{1}{r_{d_4}} \right)^{-1} = r_{d_2} \parallel r_{d_4}$$

Per cui l'amplificazione a modo differenziale:

$$A_d = g_{m_1} R_{out} = g_{m_1} (r_{d_2} \parallel r_{d_4})$$

Per capire l'ordine di grandezza di questa quantità, che non dipende dalla tensione di alimentazione, possiamo assumere le resistenze r_{d_2} e r_{d_4} uguali a r_d . Per cui, scegliendo dei MOSFET con lunghezza non troppo piccola, si riesce a raggiungere:

$$A_d = \frac{g_m r_d}{2} \cong 50$$

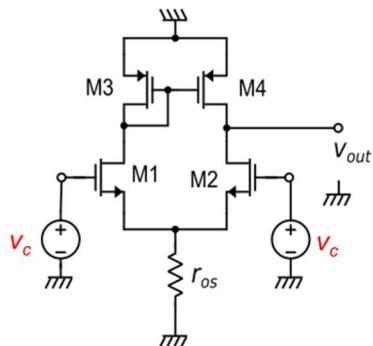
Cerchiamo un'espressione che metta in risalto quali sono i parametri su cui agire per esprimere il massimo guadagno da questo amplificatore.

$$A_d = g_{m_1} \left(\frac{1}{\frac{1}{r_{d_2}} + \frac{1}{r_{d_4}}} \right) = g_{m_1} \left(\frac{1}{\lambda_2 I_{D_2} + \lambda_4 I_{D_4}} \right)$$

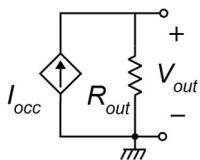
$$I_{D_2} = I_{D_4} = I_{D_1} \rightarrow A_d = \frac{g_{m_1}}{I_{D_1}} \left(\frac{1}{\lambda_2 + \lambda_4} \right) = \frac{1}{V_{TE}} \left(\frac{1}{\lambda_2 + \lambda_4} \right)$$

L'espressione finale contiene tutte le manopole progettuali con cui spingere al massimo il guadagno di questo amplificatore. Spesso, l'amplificatore differenziale con carico a specchio è utilizzato come primo stadio di un amplificatore operazionale. Per ottenere il massimo guadagno è necessario mantenere la V_{TE} al minimo e dotare i MOSFET di una buona lunghezza in modo da minimizzarne il λ . Se la corrente di riposo è fissata da qualche compromesso, aumentando la L per aumentare il guadagno bisogna aggiustare anche la W . Altrimenti, oltre che alterarsi la corrente, diminuisce il β e quindi aumenta l'overdrive, con possibilità di peggiorare il guadagno. A patto che la tensione di alimentazione sia almeno quella minima, il guadagno non dipende dalla tensione di alimentazione, il che è un vantaggio enorme rispetto ai comportamenti degli amplificatori con carico resistivo.

Guadagno a modo comune



Si può utilizzare l'equivalente di Norton anche quando lo stimolo del circuito è il modo comune.



La resistenza di uscita R_{out} è la stessa di prima:

$$v_{out} = i_{occ}R_{out} \quad R_{out} = r_{d_2} \parallel r_{d_4}$$

Tra la configurazione a modo differenziale e la configurazione a modo comune cambia la corrente di cortocircuito.

La corrente di cortocircuito:

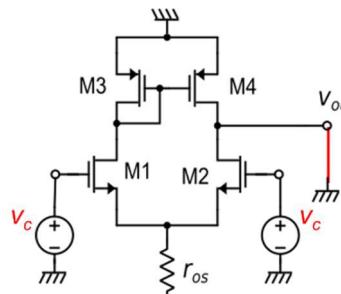
$$i_{occ} = i_{d_1} - i_{d_2}$$

$$v_d = 0 \rightarrow i_{d_1} = i_{d_2} = \frac{v_c}{2r_{os}} \rightarrow i_{occ} = 0$$

Come per la coppia a carico resistivo, la variazione del modo comune si ritrova praticamente identica come variazione del potenziale di source. Dunque, produce una corrente v_c/r_{os} che si divide in parti uguali tra i due rami. Essendo le correnti nei due rami uguali, la corrente di cortocircuito risulta nullo, per cui il guadagno di modo comune risulta anch'esso nullo.

$$v_{out}|_{v_d=0} = 0 \rightarrow A_c = 0 \rightarrow CMRR = +\infty$$

Il risultato non è in realtà corretto. Il problema deriva dall'aver considerato uguali i_{d_3} e i_{d_4} , cioè dall'aver considerato lo specchio ideale. Inoltre, l'assunzione $i_{d_1} = i_{d_2}$ sarebbe corretta soltanto in caso di comportamento perfettamente simmetrico del circuito. Nel momento in cui si applica l'equivalente di Norton, il cortocircuito all'uscita rompe la simmetria del circuito alle variazioni, per cui non è più vero che le correnti in M1 e M2 sono esattamente uguali.



Affinché M3 e M4 abbiano la stessa corrente alle variazioni, devono condividere stesse v_{gs} , v_{bs} e v_{ds} . Le prime due sono uguali per costruzione, mentre le v_{ds} non sono uguali:

$$v_{ds_4} = 0 \quad v_{ds_3} = v_{d_3} = -i_{d_1} \cdot \frac{1}{g_{m3}}$$

Similmente, M1 e M2, a causa del cortocircuito in uscita, hanno v_{ds} diverse:

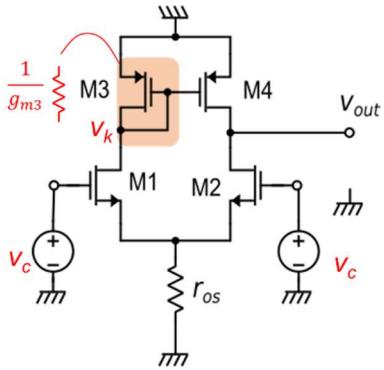
$$v_{ds_1} = -\frac{1}{g_{m3}} i_{d_1} - v_{s_1} \quad v_{ds_2} = -v_{s_1} \neq v_{ds_1}$$

Essendo $v_{ds_3} \neq v_{ds_4}$ M3 e M4 non specchiano nel modo ideale, per cui non saranno attraversati dalla stessa corrente. Ma allora, la corrente i_{d_1} che scorre in M3 non viene riportata esattamente in M4, per cui $i_{occ} \neq i_{d_1} - i_{d_2}$. E, anche se fosse vero, la i_{d_1} è di per sé diversa dalla i_{d_2} , per cui la differenza sarebbe comunque non nulla. Tenendo conto delle discrepanze date dalla non simmetria del circuito, si può calcolare la i_{occ} in modo preciso, scoprendo che non è davvero nulla.

$$i_{d_1} \neq i_{d_2} \quad i_{d_3} \neq i_{d_4} \quad i_{occ} = i_{d_4} - i_{d_2} \neq 0$$

Nell'analisi a modo differenziale le approssimazioni sulla simmetria elettrica sono accettabili perché in quel caso le due correnti i_{d_1} e i_{d_2} hanno segno diverso e si sommano nel produrre la corrente di cortocircuito. L'errore relativo che si commette nella somma tra due quantità simili è del tutto trascurabile rispetto al caso della differenza.

Per il calcolo dell'amplificazione di modo comune, quindi, si rinuncia all'equivalente di Norton del circuito. Rimuovendo il cortocircuito in uscita il circuito, essendo stimolato a modo comune, torna a comportarsi in modo simmetrico.



In particolare, cerchiamo di ricavare un'espressione per V_k . Questo perché $V_k = V_{out}$, ma è più semplice da calcolare.

Il transistore M₃, poiché è montato a diodo, nel circuito equivalente di piccolo segnale può essere sostituito da una resistenza $1/g_{m_3}$. Dunque:

$$v_k = -\frac{1}{g_{m_3}} i_{d_1}$$

Per cui, il guadagno a modo comune:

$$A_c = \frac{v_{out}}{v_c} \Big|_{v_d=0} = -\frac{1}{2r_{os}g_{m_3}}$$

CMRR

Disponendo di entrambi i guadagni di modo differenziale e modo comune, si può procedere al calcolo del CMRR dell'amplificatore differenziale con carico a specchio.

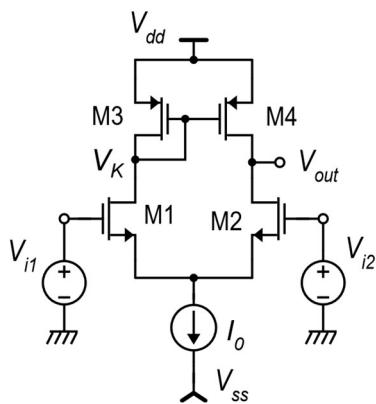
$$CMRR = \left| \frac{A_d}{A_c} \right| = \frac{g_{m_1}(r_{d_2} \parallel r_{d_4})}{\left(\frac{1}{2r_{os}g_{m_2}} \right)} = 2g_{m_3}r_{os}g_{m_1}(r_{d_2} \parallel r_{d_4})$$

Se il generatore della corrente di tail è uno specchio di corrente semplice, si può assumere $r_{os} \cong r_d$, per cui:

$$CMRR \approx (g_m r_d)^2$$

L'amplificatore differenziale con carico a specchio raggiunge facilmente valori di CMRR di 80 dB.

Caratteristica ingresso differenziale – uscita s.e.



A modo differenziale, sul nodo k c'è pochissimo segnale. Infatti, la variazione di corrente $g_m \frac{v_d}{2}$ sul lato sinistro entra nella bassa resistenza $1/g_{m_3}$ del transistore M3 montato a diodo, per cui le variazioni di tensione sono piccole. Il segnale V_k è di fase opposta a V_{out} ed è piccolo, ma ha la stessa componente di riposo della V_{out} . Questo comportamento, qualche volta, viene sfruttato per ottenere un'uscita quasi differenziale.

Applichiamo in ingresso un grande segnale a modo differenziale, che deve necessariamente comprendere anche una componente di modo comune costante.

$$V_{i_1} = V_C + \frac{V_D}{2} \quad V_{i_2} = V_C - \frac{V_D}{2}$$

Per $V_D = 0$, nel punto di riposo, il circuito è perfettamente simmetrico. Quindi, per determinare la V_{out} si può determinare la V_k .

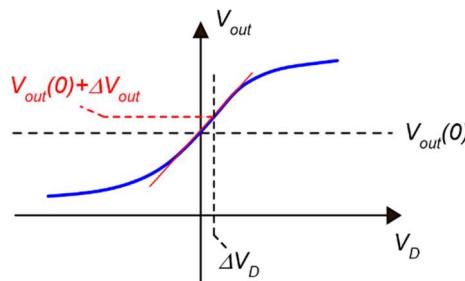
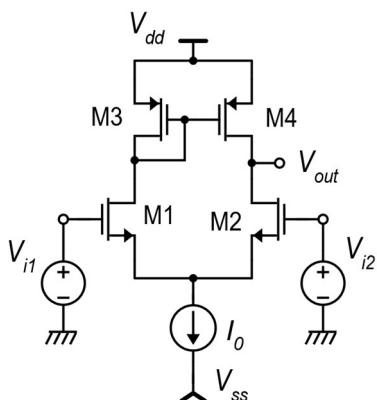
$$V_{out}(V_D = 0) = V_k(V_D = 0) = V_{dd} - |V_{GS_3}|$$

$$|V_{GS_3}| = |V_{tp}| + |V_{GS_3} - V_{tp}|_3$$

Se M3 opera in forte inversione:

$$|V_{GS_3}| = |V_{tp}| + \sqrt{\frac{2I_{D3}}{\beta_3}}$$

Dunque, la V_{GS_3} è determinata e progettabile modificando l'overdrive del transistore M3 giocando sul β_3 .

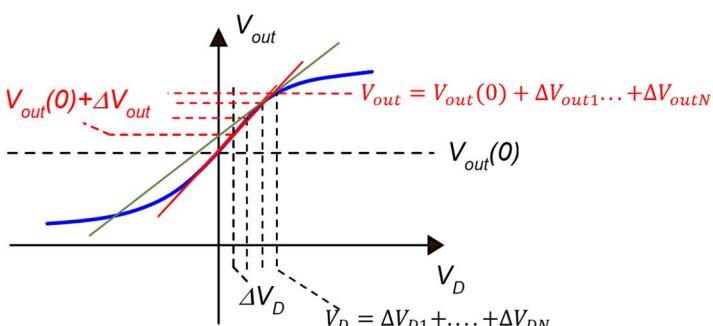


Dal punto di riposo $V_D = 0$, una variazione ΔV_D causa una variazione ΔV_{out} dell'uscita $V_{out}(0)$. Eseguendo un'approssimazione lineare:

$$\Delta V_{out} = R_{out} \Delta I_{occ}$$

L'approssimazione di Taylor è valida solo se quelle indicate con Δ sono piccole variazioni; la formula, in effetti, potrebbe anche essere scritta come $v_{out} = R_{out} i_{occ}$. A sua volta, la variazione della corrente di cortocircuito sarà data da:

$$\Delta I_{occ} = \Delta V_D \left(\frac{dI_{occ}}{dV_D} \right)_{V_D=0}$$

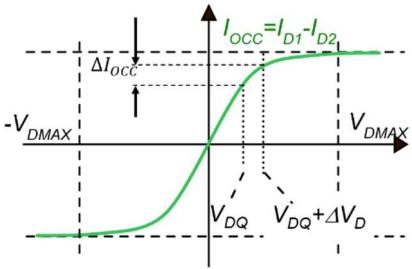


Applicando consecutivamente più variazioni, aggiornando il punto di riposo al termine di ognuna, si traccia l'intera caratteristica come successione di tratti lineari. Inizialmente la caratteristica viene percorsa lungo la tangente calcolata nell'origine. Man mano che V_D e V_{out} si allontanano dal riposo iniziale, i parametri di piccolo segnale iniziano a cambiare, alterando la pendenza dei tratti che compongono la curva.

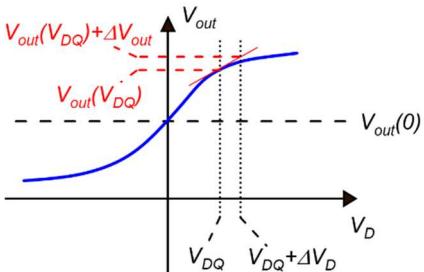
$$V_{out_N} = V_{out_{N-1}} + \Delta V_{D_N} \left. \frac{\partial V_{out}}{\partial V_D} \right|_{V_{D_{N-1}}}$$

Dove la derivata che compare nell'espressione coincide al guadagno a modo differenziale che si ha nel punto di riposo precedente alla variazione N -esima. Per ogni variazione, la derivata che imposta la pendenza del tratto lineare rappresenta il guadagno al punto di riposo precedente e cambierà in base ai nuovi valori dei parametri di piccolo segnale. In generale:

$$\Delta V_{out} = R_{out}(V_{DQ}) \Delta I_{occ} = R_{out}(V_{DQ}) \Delta V_D \left. \frac{\partial I_{occ}}{\partial V_D} \right|_{V_D=V_{DQ}}$$



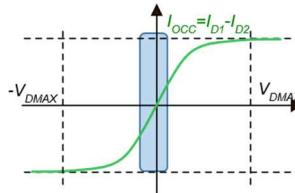
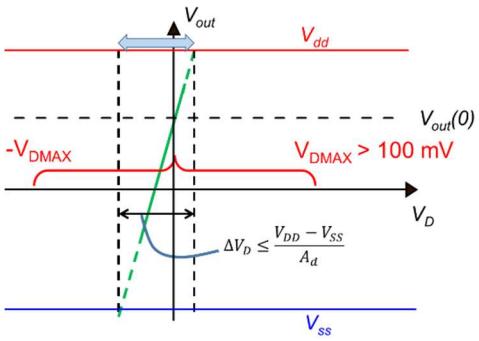
A fianco si trova rappresentata la caratteristica di $I_{D_1} - I_{D_2}$ in funzione della tensione differenziale per una coppia differenziale. Nel caso dell'amplificatore differenziale con carico a specchio $I_{D_1} - I_{D_2} = I_{occ}$. È proprio questa la caratteristica da derivare nel punto di riposo V_{DQ} per poter ottenere la ΔV_{out} .



$$\Delta V_{out} = \Delta V_D R_{out}(V_{DQ}) \frac{\partial I_{occ}}{\partial V_D} \Big|_{V_D=V_{DQ}}$$

$$\frac{\partial V_{out}}{\partial V_D} = R_{out}(V_{DQ}) \frac{\partial I_{occ}}{\partial V_D} \Big|_{V_D=V_{DQ}}$$

Dunque, per determinare l'andamento della caratteristica di ingresso – uscita dell'amplificatore, è sufficiente determinare come la resistenza differenziale di uscita e la derivata della corrente di cortocircuito variano rispetto alla tensione differenziale V_D . In condizioni normali, a contare è solo uno dei due termini, ovvero la R_{out} ; la derivata della corrente di cortocircuito rimane più o meno pari al g_m .



Si approssima che la derivata della corrente di cortocircuito sia ovunque pari a g_m e che la caratteristica dell'amplificatore si comporti in modo altrettanto lineare fino alla saturazione data dalla tensione di alimentazione.

In queste ipotesi, ci chiediamo quale variazione della tensione differenziale in ingresso sarebbe in grado di far fare alla tensione di uscita V_{out} l'intera escursione da V_{ss} a V_{dd} .

La saturazione dell'amplificatore interverrà prima di quanto raffigurato, per cui quella che troveremo è una sovrastima della tensione di ingresso differenziale entro cui l'amplificatore si comporta linearmente.

$$\Delta V_D \leq \frac{V_{dd} - V_{ss}}{\frac{\partial V_{out}}{\partial V_D}} = \frac{V_{dd} - V_{ss}}{A_d}$$

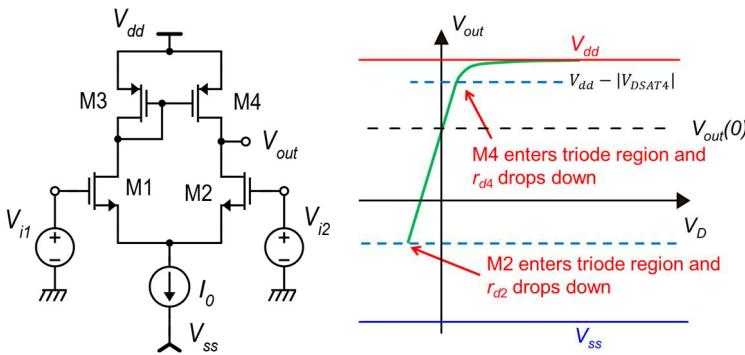
Se si considera un guadagno differenziale $A_d = 100$ e un'escursione della tensione di alimentazione di $V_{dd} - V_{ss} = 5V$ si ottiene che $\Delta V_D < 50 mV$. La V_{Dmax} è sicuramente maggiore di $100 mV$, per cui la regione di linearità della caratteristica in-out dell'amplificatore avviene sicuramente in una zona molto più piccola dell'intervallo $[-V_{Dmax}, V_{Dmax}]$. Applicando una tensione differenziale $V_D > V_{Dmax}$ l'amplificatore esce sicuramente dalla zona lineare. Quindi, si verifica che l'amplificatore satura prima che la coppia differenziale esca dalla sua zona lineare. Pertanto, è ragionevole assumere, nella zona operativa in cui l'amplificatore non è in saturazione, che la derivata della corrente di cortocircuito rispetto alla tensione di ingresso differenziale sia costante e pari a g_m . Dunque, per vedere come varia la derivata della caratteristica dell'amplificatore, è sufficiente determinare la relazione $R_{out}(V_D)$. Infatti, da prima:

$$\frac{\partial V_{out}}{\partial V_D} = R_{out}(V_{DQ}) \frac{\partial I_{occ}}{\partial V_D} \Big|_{V_D=V_{DQ}} \cong R_{out}(V_{DQ}) g_m$$

Anche la resistenza di uscita R_{out} , nella zona lineare, non subisce grandi cambiamenti

$$R_{out} = r_{d_2} \parallel r_{d_4}$$

Le resistenze differenziali r_d sono nell'ordine di $1/\lambda I_D$ e non subiscono grandi variazioni fintanto che i transistori si trovano in saturazione. Si osserverà una variazione importante di r_{d_2} e r_{d_4} , quindi della R_{out} dell'amplificatore, quando rispettivamente M2 e M4 inizieranno ad entrare in zona triodo a causa di un'eccessiva tensione differenziale di ingresso.



M4 esce dalla zona di saturazione quando la V_{out} oltrepassa in salita $V_{dd} - |V_{DSat4}|$. Poiché la zona operativa della V_D è abbastanza più piccola dell'intervallo $[-V_{Dmax}, V_{Dmax}]$, le correnti di riposo non cambiano così tanto da alterare le r_d quando i transistori si trovano in saturazione, per cui la R_{out} rimane abbastanza costante. Tuttavia, quando M4 passa in zona triodo la sua r_{d4} inizia a crollare fortemente, quindi crolla la R_{out} e il guadagno dell'amplificatore di conseguenza.

Il source di M1 e M2 rimane più o meno costante applicando una tensione a modo differenziale. Si era dimostrato che per $V_D = V_{Dmax}$ la tensione del source era cambiata di una frazione di $V_{GS} - V_t$. Dunque, quando la V_{out} diminuisce si abbassa progressivamente la V_{DS2} di M2. La tensione minima a cui può arrivare la V_{out} prima che M2 entri in zona triodo sarà pari a

$$V_S + V_{DSsat2} \cong V_S(V_D = 0) + V_{DSsat2} = V_C - V_{GS2}(V_D = 0) + V_{DSsat2}$$

Chiamando $V_{GS2}(V_D = 0) = V_{GS2Q}$, l'amplificatore differenziale con carico a specchio rimane nella regione di funzionamento lineare se:

$$V_C - V_{GS2Q} + V_{DSsat2} \leq V_{out} \leq V_{dd} - |V_{DSat}|_4$$

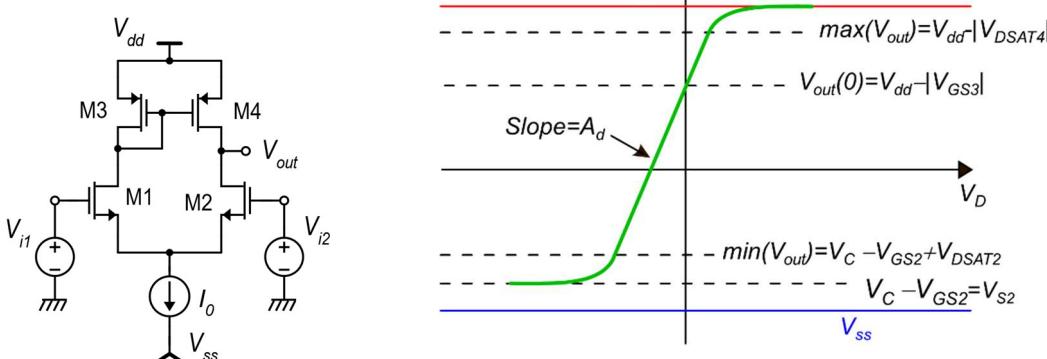
Per quanto riguarda la tensione di uscita minima per il funzionamento lineare, si considerano due casi:

Forte inversione: $V_{DSsat2} = V_{GS2} - V_{tn} \cong V_{GS2Q} - V_{tn} \rightarrow \min(V_{out}) \cong V_C - V_{GS2Q} + V_{GS2Q} - V_{tn} = V_C - V_{tn}$

Debole inversione: $V_{DSsat2} \cong 100 \text{ mV}$, $V_{GS2} \cong V_{tn} \rightarrow \min(V_{out}) \cong V_C - V_{tn} + 100 \text{ mV}$

Se M2 entra in triodo la r_{d2} crolla e, finendo questa in parallelo con r_{d4} nel determinare la R_{out} , determina il crollo di R_{out} e quindi il crollo del guadagno dell'amplificatore, con conseguente appiattimento della caratteristica.

In conclusione:



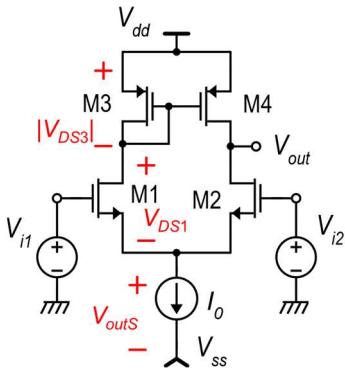
L'asintoto inferiore della caratteristica è dato dalla condizione $V_{DS2} = 0$, che si verifica quando la V_{out} raggiunge la tensione del source, ovvero quando:

$$V_{out} = V_C - V_{GS2Q} = V_{S2}$$

Il limite superiore della dinamica è pari a V_{dd} . Quando la coppia è totalmente sbilanciata per forti tensioni differenziali positive, la corrente va tutta in M1, per cui M3 (diodo) si accende e si accende M4 di conseguenza. Tuttavia, in M2 non scorre corrente a causa dello sbilanciamento, per cui M4 è acceso senza corrente; quindi, M4 si comporta come un resistore di pull-up. Questa condizione si verifica per $V_D = V_{D_{max}}$. Viceversa, se la coppia è sbilanciata dall'altra parte M2 prende tutta la corrente.

Minima tensione di alimentazione

Determiniamo la minima tensione di alimentazione $V_{dd} - V_{ss}$ accettabile per questo circuito.



Con V_{outS} si identifica la tensione di uscita dello specchio.

$$V_{dd} - V_{ss} = V_{outS} + V_{DS1} + |V_{DS3}| = V_{outS} + V_{DS1} + V_{GS3}$$

La minima tensione di alimentazione:

$$\min(V_{dd} - V_{ss}) = \min(V_{outS}) + \min(V_{DS1}) + \min(V_{GS3})$$

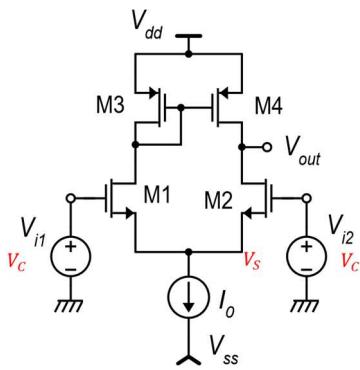
$$\rightarrow \min(V_{dd} - V_{ss}) = V_{min} + V_{DS_{sat1}} + |V_{GS3}|$$

Se ad esempio $V_{min} = 100 \text{ mV}$, $V_{DS_{sat1}} = 100 \text{ mV}$, $|V_{GS3}| = 0.5 \text{ V}$ si ottiene una minima tensione di alimentazione di **0.7 V**. Pertanto, l'amplificatore differenziale con carico a specchio si può caratterizzare come “ultra low voltage amplifier”.

Range di modo comune

Parlando di grandi segnali, abbiamo caratterizzato il range di uscita. La massima V_{out} per il funzionamento lineare è V_{dd} a meno di una $V_{DS_{sat}}$, mentre in basso la minima V_{out} per il funzionamento lineare è limitata dalla tensione di modo comune. Se si adotta una tensione di modo comune elevato la dinamica dell'amplificatore si restringe. Ad ogni modo, anche la tensione di modo comune è soggetta a una certa dinamica per il corretto funzionamento dell'amplificatore.

Limite inferiore



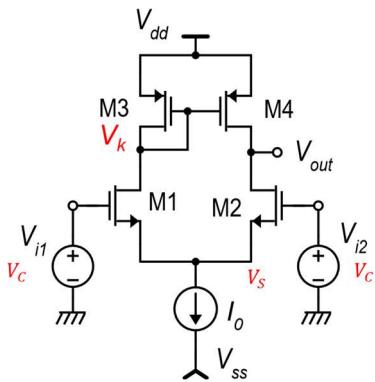
Abbassando progressivamente la tensione di modo comune, la tensione di source scende alla stessa velocità, fintanto che raggiunge la V_{min} dello specchio che genera la corrente di polarizzazione I_0 . Scendendo ulteriormente, la tensione di uscita dello specchio diminuisce oltre quella minima e la corrente I_0 diminuisce. Quando la V_s diventa nulla, la corrente di polarizzazione diventa nulla. Il limite inferiore della tensione di modo comune:

$$\min(V_C) = V_{ss} + V_{min} + V_{GS1}$$

Come si nota, la dinamica della tensione di modo comune, non avendo limite inferiore pari alla V_{ss} , non è rail to rail.

Oltre al fatto che la corrente dello specchio diminuisce quando $V_s < V_{min}$, andando in triodo il MOSFET verso l'uscita dello specchio la resistenza di uscita r_{os} crolla, per cui anche il CMRR dell'amplificatore si degrada.

Limite superiore



Quando la tensione di modo comune V_C sale, la tensione di source V_S sale anch'essa allo stesso ritmo. Poiché il drain di M1 (V_k) è bloccato a $V_{dd} - |V_{GS_3}|$, se sale il source di M1 si riduce la V_{DS_1} . Poiché a riposo il circuito è simmetrico, lo stesso accade per M2 e la V_{DS_2} .

$$\begin{aligned} V_{DS_1} &= V_k - V_{S_1} \geq V_{DS_{sat_1}} \\ \rightarrow V_{dd} - |V_{GS_3}| - (V_C - V_{GS_1}) &\geq V_{DS_{sat_1}} \\ \rightarrow V_{dd} - |V_{GS_3}| + V_{GS_1} - V_{DS_{sat_1}} &\geq V_C \end{aligned}$$

Espandendo le V_{GS} :

$$\begin{cases} |V_{GS_3}| = |V_{t_{p_3}}| + |V_{GS_3} - V_{t_{p_3}}| \\ V_{GS_1} = V_{t_{n_1}} + (V_{GS_1} - V_{t_{n_1}}) \end{cases} \rightarrow \max(V_C) = V_{dd} - |V_{t_{p_3}}| + V_{t_{n_1}} - |V_{GS_3} - V_{t_{p_3}}| + (V_{GS_1} - V_{t_{n_1}}) - V_{DS_{sat_1}}$$

Le tensioni di overdrive di M1 e M3 possono essere rese uguali da progetto, per cui si cancellano. Mentre la tensione del substrato di M1 è a ground, la tensione di source sale con il modo comune, per cui M1 e M2 sono influenzati dall'effetto body e la loro tensione di soglia è maggiore rispetto alla $V_{t_{n_0}}$. Invece, i transistori M3 ed M4 hanno il substrato e il source collegati assieme alla V_{dd} . Questo fa sì che solitamente:

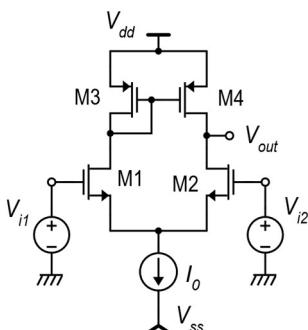
$$V_{t_{n_1}} - |V_{t_{p_3}}| > V_{DS_{sat_1}} \rightarrow \max(V_C) > V_{dd}$$

Dunque, la dinamica del modo comune include e può superare (di poco) il rail di alimentazione positivo.

Amplificatore telescopico

Migliorare il guadagno dell'amplificatore differenziale con carico a specchio

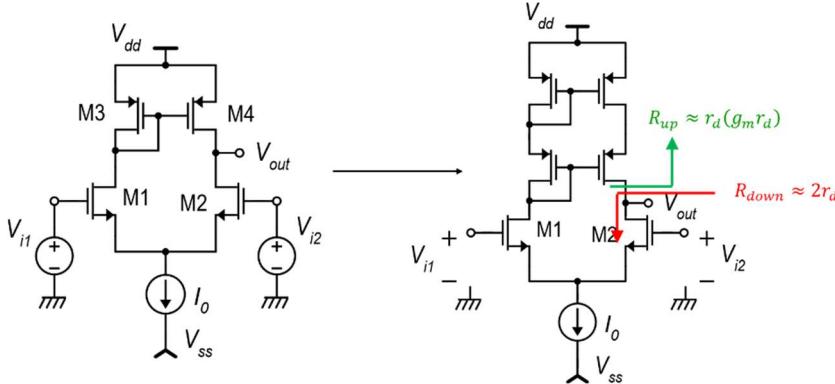
L'obiettivo è quello di migliorare il guadagno dell'amplificatore differenziale con carico a specchio rimanendo sul singolo stadio.



$$A_d = G_m R_{out}$$

$$G_m = g_{m_1} = \frac{I_{D_1}}{V_{TE}} \quad R_{out} = r_{d_2} \parallel r_{d_4} = \frac{1}{I_{D_1}} \left(\frac{1}{\lambda_2 + \lambda_4} \right)$$

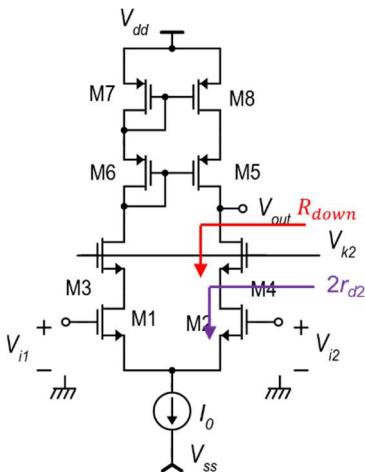
Per migliorare il g_{m_1} , una volta che V_{TE} è al minimo l'unica opzione è quella di aumentare la I_{D_1} aumentando la I_o . Contemporaneamente, però, la I_{D_1} si trova al denominatore della R_{out} , per cui l'effetto di miglioramento viene abbattuto. Il miglioramento del guadagno si può ottenere migliorando la R_{out} lasciando inalterato il G_m . Per farlo, si impiega la configurazione cascode.



La resistenza R_{up} corrisponde alla resistenza di uscita dello specchio. Se quindi si passa dallo specchio semplice allo specchio cascode, la R_{up} è amplificata di un fattore $g_m r_d$. Tuttavia, la R_{down} non cambia e rimane nell'ordine di $2r_d$ come nel caso precedente. Facendo nuovamente i conti si ottiene che:

$$R_{out} \cong r_{d_2}$$

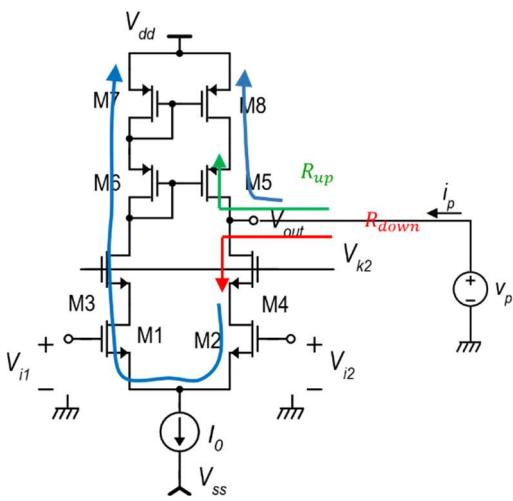
Il miglioramento rispetto al caso dell'amplificatore con carico a specchio semplice è solo di un fattore 2. Allora, si cerca di aumentare anche la R_{down} aggiungendo un cascode anche di sotto.



Si aggiungono M3 e M4 che realizzano degli stadi a gate comune, trasformando la struttura sotto in un amplificatore cascode differenziale (alle variazioni e a modo differenziale il source è a ground). Entrambi i rami della parte di sotto sono formati da una cascata gate comune – source comune. L'aggiunta dei gate comune non cambia la funzionalità del circuito: le correnti di M1 e M2 vengono rispettivamente passate da M3 allo specchio e da M4 all'uscita inalterate. La corrente di cortocircuito rimane la stessa ($I_{D_1} - I_{D_2}$) poiché la I_{D_1} viene trasferita da M3 allo specchio cascode che la ribalta sul ramo di destra verso l'uscita. La R_{down} :

$$R_{down} \cong 2r_{d_2} + r_{d_4}(1 + g_{m_4}2r_{d_2})$$

La resistenza R_{down} , all'incirca, è accresciuta di un fattore $g_m r_d$. Aver incrementato anche la resistenza R_{down} comporta complessivamente una resistenza di uscita maggiore.



Anche in questo caso si ha un ricircolo di corrente: la corrente in R_{down} ricircola indietro all'uscita a causa dello specchio di corrente.

$$R_{up} \cong r_{d_8} + r_{d_5}(1 + r_{d_8}g_{m_5}) \cong r_{d_5}(r_{d_8}g_{m_5})$$

$$R_{down} \cong 2r_{d_2} + r_{d_4}(1 + g_{m_4}2r_{d_2}) \cong 2r_{d_4}(r_{d_2}g_{m_4})$$

A causa del ricircolo:

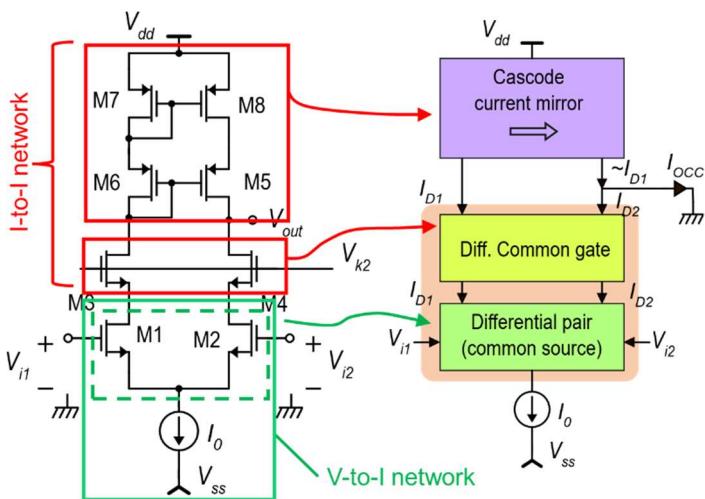
$$i_p = \frac{v_p}{R_{up}} + 2 \frac{v_p}{R_{down}}$$

La resistenza di uscita:

$$R_{out} = R_{up} \parallel \left(\frac{R_{down}}{2} \right)$$

La resistenza di uscita dell'amplificatore differenziale così ottenuto è il parallelo di due resistenze cascode:

$$R_{out} \cong r_{d_5}(r_{d_8}g_{m_5}) \parallel r_{d_4}(r_{d_2}g_{m_4})$$



La coppia differenziale si comporta come stadio source comune (source a ground per le variazioni). Lo stadio comune differenziale è polarizzato ai gate con tensione costante V_{k_2} passa il segnale da source a drain. La coppia differenziale e lo stadio a gate comune differenziale formano lo stadio cascode differenziale (che realizza l'amplificatore cascode). La coppia differenziale trasforma la tensione differenziale di ingresso in una corrente differenziale. Le correnti I_{D_1} e I_{D_2} vengono passate in alterate dal gate comune allo specchio cascode, il quale le trasporta in uscita facendone la differenza. L'impedenza di uscita, rispetto all'amplificatore differenziale con carico a specchio, è molto più grande.

La corrente di cortocircuito:

$$I_{occ} = G_m v_d \cong I_{D_1} - I_{D_2} \cong g_{m_1} v_d$$

Il G_m complessivo continua ad essere pari al g_m della coppia. L'amplificazione differenziale totale:

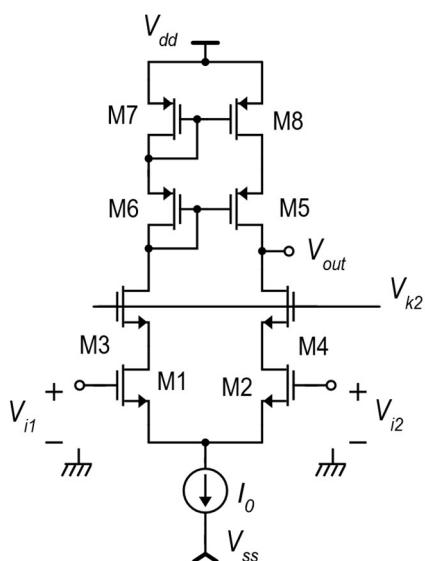
$$A_d = G_m R_{out} \cong g_{m_1} [(r_{d_2} g_{m_4} r_{d_4}) \parallel (r_{d_8} g_{m_5} r_{d_5})]$$

Se si considerano i dispositivi con r_d e g_m uguali si ottiene che l'amplificazione di modo differenziale, come ordine di grandezza:

$$A_d \cong \frac{(g_m r_d)^2}{2}$$

Raddoppiando l'ingombro dell'amplificatore differenziale con carico a specchio semplice, senza modificare le L dei transistori, il guadagno è aumentato di un fattore 100. Nel caso dell'amplificatore differenziale con carico a specchio semplice, per raddoppiare il guadagno si dovrebbero raddoppiare le L e le W, di conseguenza, per mantenere gli overdrive costanti, quadruplicandone l'ingombro. All'interno di un circuito integrato l'amplificatore differenziale cascode ha un guadagno tale da poter funzionare da amplificatore operazionale, ad esempio per bufferizzare una tensione.

Range di uscita



Limite superiore:

Man mano che la tensione di uscita V_{out} sale, la tensione di uscita dello specchio cascode scende. La più alta tensione di uscita possibile è quella per cui lo specchio cascode si trova ad operare con una tensione di uscita pari a V_{min} .

$$V_{out} = V_{dd} - |V_{out-mirro}| \rightarrow \max(V_{out}) = V_{dd} - V_{min-cascode}$$

Andando oltre questa tensione, la resistenza di uscita dello specchio cascode, che è l' R_{up} , diminuisce abbattendo anche la R_{out} e quindi il guadagno dell'amplificatore.

Limite inferiore

Quando invece la tensione di uscita scende verso il basso, M4 può entrare in zona triodo.

La V_{DS} :

$$V_{DS_1} = V_{out} - V_{S_1} \geq V_{DS_{11}}$$

La tensione al source di M4:

$$V_S \equiv V_{k_2} - V_{CS}$$

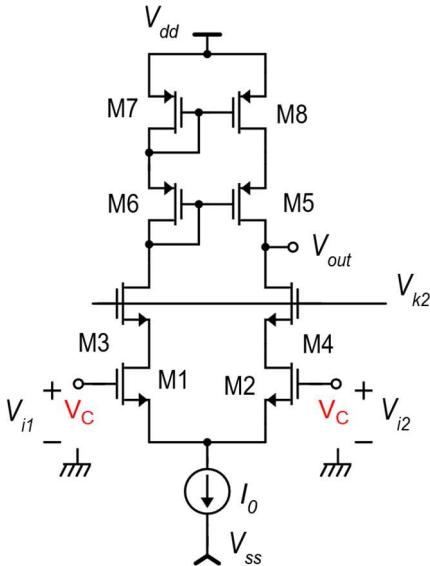
La V_{GS_4} rimane più o meno costante anche al variare del segnale di ingresso. Principalmente, la V_{GS} dipende dalla corrente che scorre nel MOSFET, la quale non cambia molto. Le variazioni delle correnti che imprime la coppia differenziale rispetto all'equipartizione di riposo sono piccolissime: appena una delle due correnti I_{D_1} o I_{D_2} diventa poco più grande dell'altra, la loro differenza (la corrente di cortocircuito) entra in una resistenza di uscita molto grande producendo una rapida saturazione della tensione di uscita. Rispetto all'amplificatore differenziale con carico a specchio semplice, in questo caso l'amplificazione è 100 volte più grande, per cui il range operativo della V_D è ancora più piccolo.

$$V_{out} - V_{k_2} + V_{GS_4} \geq V_{DS_{sat4}} \rightarrow \min(V_{out}) = V_{k_2} - V_{GS_4} + V_{DS_{sat4}}$$

In forte inversione: $V_{DS_{SG_4}} = V_{GS_4} - V_{t_4} \rightarrow \min(V_{out}) = V_{k_2} - V_{t_4}$

A limitare la minima tensione di uscita è V_{k_2} : a differenza del limite superiore, modulando la V_{k_2} si può modificare la minima tensione di uscita $\min(V_{out})$. Scendendo con la V_{out} oltre il limite inferiore M4 entra in zona triodo: la sua resistenza alle variazioni diminuisce e comporta una diminuzione della R_{out} e quindi una diminuzione dell'amplificazione differenziale.

Range di modo comune di ingresso



Il limite inferiore per la tensione di modo comune è sempre lo stesso: scendendo la tensione di modo comune scende quella di source di pari passo e strozza l'uscita dello specchio che genera la corrente di tail.

$$\min(V_C) = V_{ss} + V_{min-tail} + V_{GS_1}$$

Per il limite superiore della tensione di modo comune, si nota che i drain di M1 e M2 sono bloccati rispettivamente a $V_{k_2} - V_{GS_3}$ e $V_{k_2} - V_{GS_4}$. Quindi, aumentando la tensione di modo comune, aumenta la tensione di source di M1 e M2, ma essendo bloccati i relativi drain si riducono le V_{DS_1}, V_{DS_2} .

$$V_{DS_1} = V_{D_1} - V_{S_1} \geq V_{DS_{sat_1}}$$

$$\begin{cases} V_{D_1} = V_{k_2} - V_{GS_3} \\ V_{S_1} = V_C - V_{GS_1} \end{cases} \rightarrow V_{DS_1} = V_{k_2} - V_{GS_3} - V_C + V_{GS_1} \geq V_{DS_{sat_1}}$$

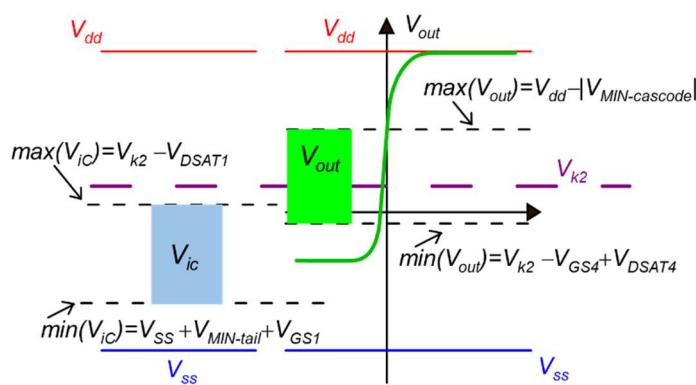
Quindi:

$$\max(V_C) = V_{k_2} - V_{GS_3} + V_{GS_1} - V_{DS_{sat_1}}$$

$$V_{GS_1} \cong V_{GS_3} \rightarrow \max(V_C) = V_{k_2} - V_{DS_{sat_1}}$$

Si possono rendere da progetto $V_{GS_1} \cong V_{GS_3}$. I transistori M1 e M3 portano la stessa corrente. M3 ha il source a un potenziale un po' più alto del body, che si suppone essere a ground, per cui risentirà di un certo effetto body. Se V_C passa il limite superiore, M1 ed M2 vanno in zona triodo. Oltre che impattare negativamente sulla R_{out} a causa del crollo della r_{d_2} , il crollo del loro g_m causa anche un crollo del G_m globale.

Range amplificatore cascode

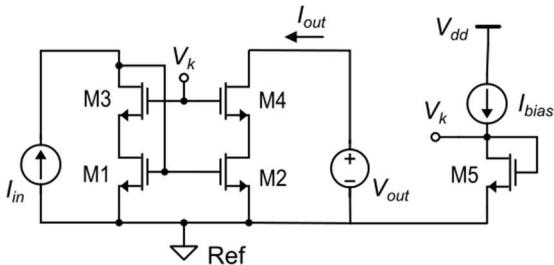


Si riportano i range della tensione di modo comune e della tensione di uscita (la zona lineare della tensione di uscita è quella in cui vale l'amplificazione nominale). I limiti superiore e inferiore della tensione di modo comune e della tensione di uscita dipendono dalla V_{k_2} . Per avere più dinamica in uscita si dovrebbe portare in basso la V_{k_2} , ma ciò avrebbe l'effetto di ridurre la dinamica del modo comune.

A inizio progetto si dovrebbe valutare la V_{k_2} come compromesso, ad esempio valutando la massima tensione di modo comune che ci si aspetta avere da parte della sorgente di segnale che pilota l'amplificatore. Se si vuole realizzare un buffer chiudendo l'uscita sul terminale invertente, il range di tensioni che l'inseguitore riesce ad inseguire è l'intersezione tra il range delle tensioni di uscita e il range delle tensioni di modo comune in ingresso. Come si nota, nel caso dell'amplificatore cascode tale intersezione è molto stretta.

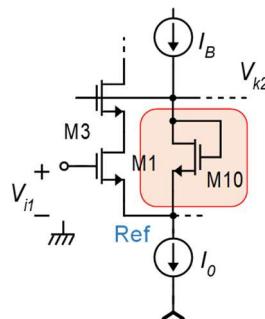
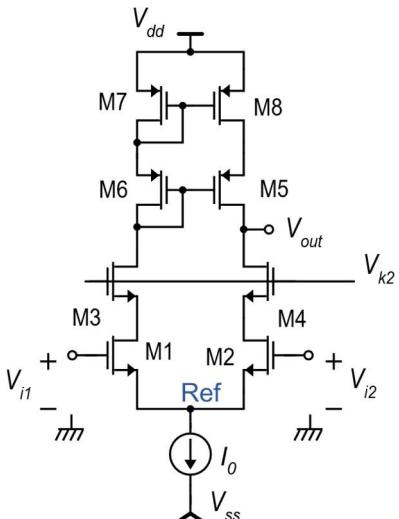
V_{k_2} adattiva – amplificatore telescopico

La scelta della V_{k_2} è un problema. La soluzione proposta non è una soluzione del problema, ma consente di ottenere un compromesso adattivo/dinamico, senza dover fissare scelte iniziali. In particolare, la V_{k_2} viene alzata e abbassata dinamicamente in base al modo comune: se il modo comune è alto la V_{k_2} si alza di pari passo, se il modo comune è basso la V_{k_2} si abbassa di pari passo. Rimane un'intersezione tra il range di modo comune e il range di uscita che si sposta in alto o in basso, permettendo l'inseguimento della tensione di ingresso. Per ottenere il risultato bisogna ottenere una V_{k_2} in funzione del modo comune.



Si utilizza lo stesso approccio impiegato nello specchio di precisione a larga dinamica cascode. Per produrre la V_{k_2} rispetto a un riferimento si forzava una corrente in un MOSFET montato a diodo.

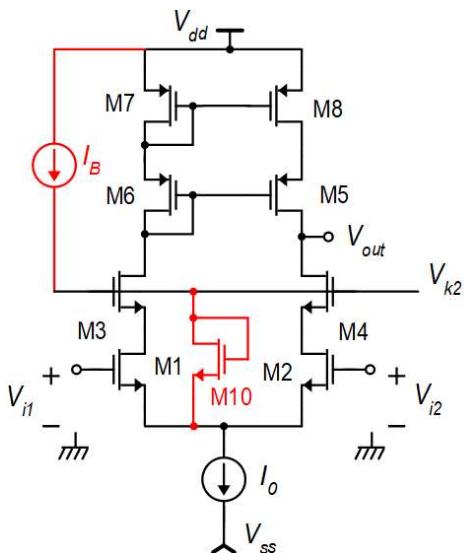
Nel caso dell'amplificatore cascode il riferimento è ai source di M1 e M2.



La V_{DS_1} è fissata da progetto; sicuramente è maggiore della $V_{DS_{sat1}}$.

$$V_{GS_{10}} = V_{DS_1} + V_{GS_3}$$

Si ottiene così la configurazione del così detto amplificatore telescopico:



Scegliendo $V_{DS_1} = V_{DS_{sat1}}$ si ottiene:

$$V_{GS_{10}} = V_{DS_{sat1}} + V_{GS_3}$$

Rifacendo gli stessi conti visti per lo specchio a larga dinamica:

$$(V_{GS} - V_t)_{10} = m_3 V_{DS_{sat1}} + (V_{GS} - V_t)_3$$

La corrente I_B viene sostenuta dal tail, quindi la corrente I_0 efficace che rimane per polarizzare la coppia:

$$I_{0-eff} = I_0 - I_B$$

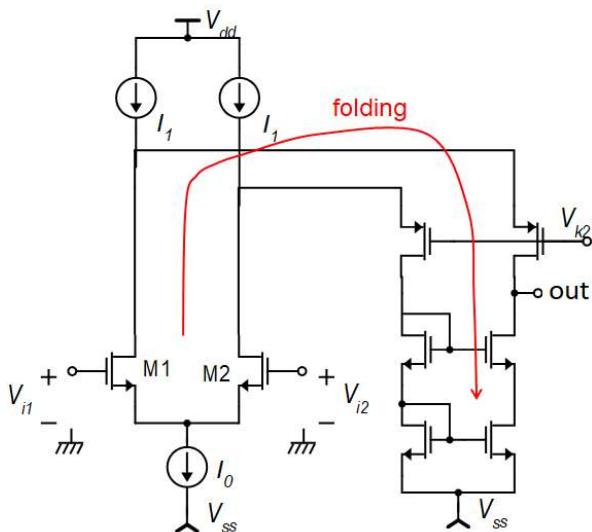
Quindi, la corrente del tail dovrà essere più grande di quella progettata per polarizzare la coppia differenziale.

La tensione adattiva non risolve completamente il problema, si tratta più di un palliativo. Permette di non dover fissare la V_{k_2} rigidamente, il che da un vantaggio nell'utilizzo dell'amplificatore come buffer. Utilizzando l'amplificatore per altri scopi, in ogni caso, se la tensione di modo comune sale si strozza la dinamica della tensione di uscita. Anche nell'amplificatore differenziale con carico a specchio semplice il modo comune influenzava il limite inferiore della tensione di uscita. In questo caso va un po' peggio; a causa delle strutture cascode il limite superiore dista dalla V_{dd} di uno stacco pari a tutta la V_{min} dello specchio cascode. Quindi, il peggioramento del limite inferiore è ancora più penalizzante sulla dinamica della tensione di uscita.

Un altro dei limiti di questa configurazione è che avendo molti transistori impilati, ciascuno dei quali deve avere una V_{DS} consona, la minima tensione di alimentazione non è così bassa (almeno 2V sono necessari). L'utilizzo dell'amplificatore telescopico si adatta quindi per applicazioni in cui la tensione di alimentazione non è troppo limitata e non si richiedono grandi escursioni della tensione di uscita.

Amplificatore folded cascode

L'amplificatore folded cascode rimuove questi limiti. Uno degli obiettivi che si cerca di raggiungere è eliminare l'interazione tra la dinamica di uscita e la dinamica di ingresso. Si impiega ancora una volta una configurazione single stage il cui convertitore V-I è ancora una volta una coppia differenziale polarizzata da una corrente di tail.



Per ottenere un amplificatore cascode, allo stadio a source comune rappresentato dalla coppia differenziale segue uno stadio a gate comune, questa volta di tipo p. Alle variazioni pMOS e nMOS si comportano allo stesso modo.

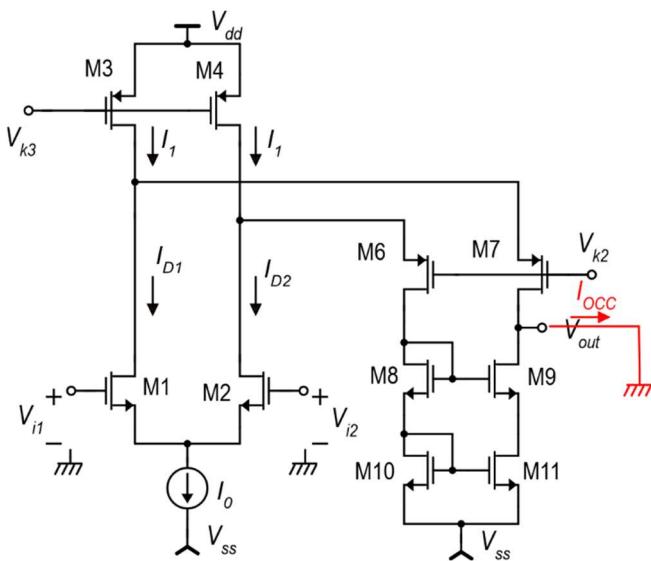
La rete di trasporto della corrente è ancora uno specchio cascode, che permette di riflettere e invertire la variazione di corrente da un lato all'altro, in modo tale da portare la differenza $i_{d_1} - i_{d_2}$ in uscita. In continua lo specchio deve accettare correnti entranti, per cui ne occorre uno di tipo n.

Mentre alle variazioni vale l'interscambiabilità dei pMOS e degli nMOS, in continua, per sostenere correnti entranti dai drain della coppia differenziale e dai source dello stadio a gate comune, è necessario fornire corrente a quei nodi in cui altrimenti non varrebbe il primo principio di Kirchoff.

Ecco perché il circuito, per funzionare, necessita dei generatori di corrente I_1 . A riposo, i transistori M1 e M2 richiedono una corrente verso il basso di $I_0/2$. La parte che rimane dalla differenza $I_1 - I_0/2$ scorre anch'essa verso il basso polarizzando il gate comune e lo specchio cascode.

Il ripiegamento consiste nel fatto che alle variazioni le correnti della coppia differenziale salgono e, vedendo una resistenza differenziale molto elevata dalla parte dei generatori di corrente, ripiegano sui source dello stadio comune, dai quali si vede una resistenza molto più bassa, nell'ordine di $1/g_m$. Dopo di che, le variazioni delle correnti proseguono verso il basso (verso la V_{ss}) dallo specchio cascode. Il ripiegamento avviene grazie alla presenza contemporanea di transistori p e transistori n.

Analisi delle correnti



Polarizzando i gate di M3 e M4 con una tensione costante, la loro V_{GS} è costante, per cui si comportano come generatori di corrente. Per produrre la V_{k_3} , anziché utilizzare un partitore, si ricorrerà ancora una volta a un transistore montato a diodo polarizzato da una corrente di bias (possiamo vedere M3 e M4 come due rami di uscita di uno specchio di corrente di tipo p). Le correnti che entrano nel gate comune in continua:

$$\begin{cases} I_{D_6} = I_1 - I_{D_2} \\ I_{D_7} = I_1 - I_{D_1} \end{cases}$$

Una condizione affinché il circuito funzioni:

$$I_{D_6} \geq 0 \quad I_{D_7} \geq 0$$

La corrente di cortocircuito:

$$I_{occ} = I_{D_7} - I_{D_9} \cong I_{D_7} - I_{D_6} = (I_1 - I_{D_1}) - (I_1 - I_{D_2}) = -(I_{D_1} - I_{D_2})$$

La corrente di cortocircuito è ancora una volta la differenza delle correnti dei rami della coppia. Considerando come tensione differenziale:

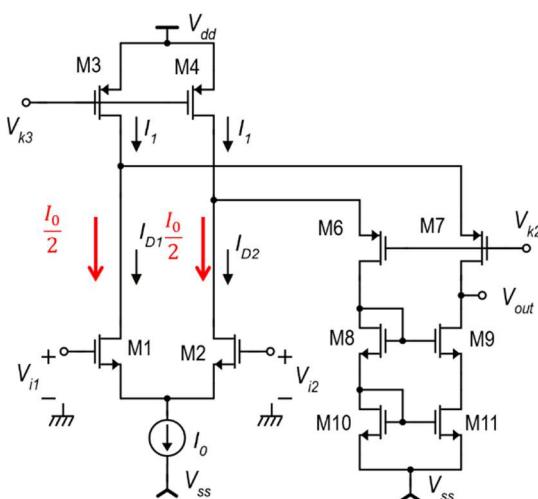
$$V_D = V_{i_1} - V_{i_2}$$

Alle variazioni:

$$-(i_{d_1} - i_{d_2}) \cong -g_{m_1} v_d$$

Impostazione di I_1

La scelta di I_1 è una scelta di progetto.



$$\begin{cases} I_{D_6} = I_1 - I_{D_2} = I_1 - \frac{I_0}{2} \\ I_{D_7} = I_1 - I_{D_1} = I_1 - \frac{I_0}{2} \end{cases}$$

$$\begin{cases} I_{D_6} > 0 \\ I_{D_7} > 0 \end{cases} \rightarrow I_1 > \frac{I_0}{2}$$

La condizione trovata è la minima per far funzionare lo stadio a gate comune. Nella realtà, la I_1 dovrà essere abbastanza maggiore con un certo margine. Immaginiamo di applicare una tensione differenziale di ingresso grande, anche oltre la $V_{D_{max}}$ in modulo. Le correnti della coppia differenziale possono sbilanciarsi completamente da un lato o da un altro:

$$V_{id} > V_{D_{max}} \rightarrow I_{D_1} = I_0$$

$$V_{id} < -V_{D_{max}} \rightarrow I_{D_2} = I_0$$

Quindi, se si considera che in uno dei due rami può scorrere anche tutta la corrente di tail, occorre che $I_1 \geq I_0$. Solitamente si sceglie $I_1 = I_0$; si accetta che con sbilanciamento completo uno dei due rami sia interdetto.

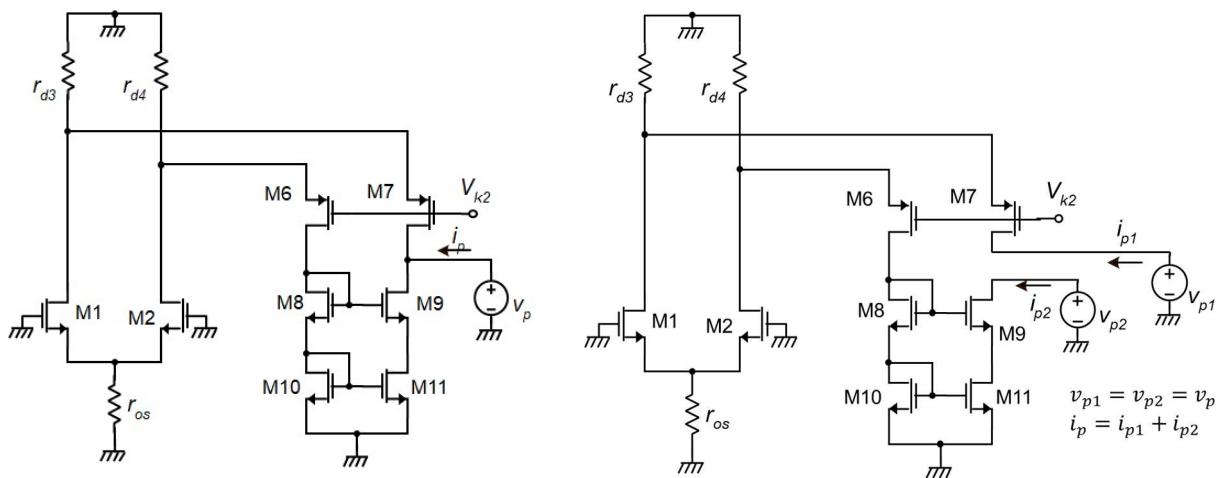
Se c'è margine di corrente, la corrente di cortocircuito in uscita può spaziare tutta la sigmoide della differenza $I_{D_1} - I_{D_2}$ della coppia differenziale, da $-I_0$ a $+I_0$. Se un ramo della parte ripiegata si secca di corrente, lo stadio ripiegato non è in grado di portare all'uscita tutta la corrente che vorrebbe. Poiché molte volte in uscita si avrà un carico capacitivo, maggiore è la corrente che si riesce a erogare in uscita, più veloce è l'amplificatore. Scgliendo $I_1 = I_0$ a riposo i rami del gate comune sono polarizzati anch'essi con $I_0/2$.

Guadagno di modo differenziale

La strategia per il calcolo dell'amplificazione è la stessa per i single stage e si basa sull'equivalente di Norton.

$$G_m = \frac{i_{occ}}{v_d} = \frac{-g_{m_1} v_d}{v_d} = -g_{m_1} \rightarrow A_d = G_m R_{out} = -g_{m_1} R_{out}$$

Poiché l'amplificazione è pari a $G_m R_{out}$ e $G_m = -g_{m_1}$, non resta che calcolare la resistenza di uscita.

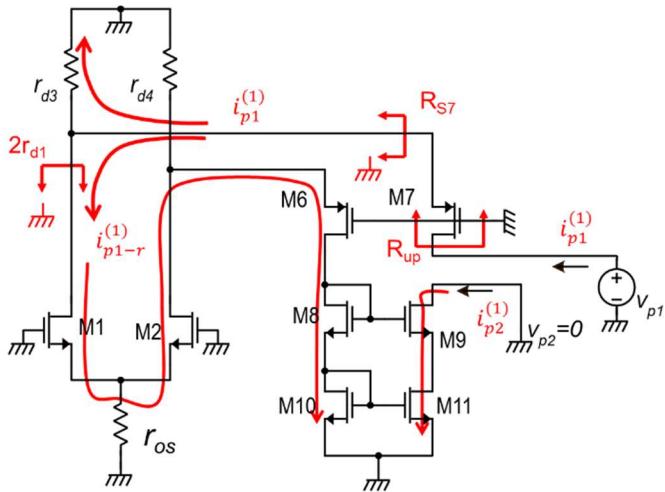


Poiché si ha un ricircolo di corrente, per un calcolo rigoroso si divide il generatore di prova in due generatori di prova, come per l'amplificatore con carico a specchio. Per l'amplificatore telescopico il ricircolo è identico a quello dell'amplificatore differenziale con carico a specchio semplice, per cui è stato sufficiente aggiornare i valori di R_{up} e R_{down} . In questo caso il ricircolo è diverso. Si può applicare il principio di sovrapposizione degli effetti, facendo agire i generatori in modo alternato.

$$\begin{cases} v_{p_1} = v_p \\ v_{p_2} = 0 \end{cases} \rightarrow i_{p_1}^{(1)}, i_{p_2}^{(1)} \quad \begin{cases} v_{p_2} = v_p \\ v_{p_1} = 0 \end{cases} \rightarrow i_{p_1}^{(2)}, i_{p_2}^{(2)}$$

$$i_{p_1} = i_{p_1}^{(1)} + i_{p_1}^{(2)} \quad i_{p_2} = i_{p_2}^{(1)} + i_{p_2}^{(2)}$$

Effetto del generatore 1



La corrente $i_{p_1}^{(1)}$:

$$i_{p_1}^{(1)} = \frac{v_p}{R_{up}}$$

La R_{up} è la resistenza che si vede dal drain di M7 con source a massa attraverso R_{S7} .

$$R_{up} \cong r_{d_7}(1 + g_{m_7} R_{S7})$$

$$R_{S7} = r_{d_3} \parallel 2r_{d_1} \rightarrow R_{up} \cong r_{d_7} g_{m_7} (r_{d_3} \parallel 2r_{d_1})$$

La corrente $i_{p_1}^{(1)}$ si divide in due porzioni, una verso l'alto attraverso r_{d_3} , una verso il basso attraverso $2r_{d_1}$.

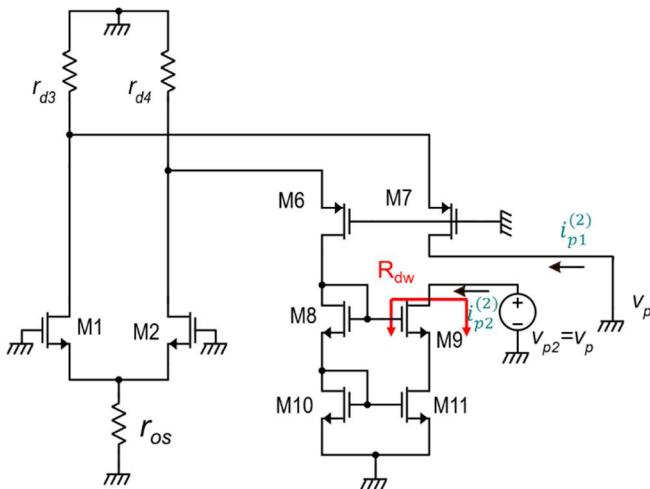
La corrente di ricircolo è quella che entra nella coppia verso il basso, $i_{p_1-r}^{(1)}$. Applicando il partitore di corrente:

$$i_{p_1-r}^{(1)} = i_{p_1}^{(1)} \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}}$$

Dal source di M2 si vede $1/g_{m_2}$, per cui quasi tutta la variazione prosegue all'interno di M2 anziché defluire in r_{os} . Di nuovo, in uscita dal drain di M2, la variazione prosegue verso M6 poiché dal source si vede $1/g_{m_6}$ che è bassa rispetto alla r_{d_4} di richiusura a massa. Dunque, la variazione prosegue entrando nello specchio di corrente e viene specchiata dall'altro lato. Quindi, quando agisce il primo generatore si trova una corrente non nulla anche attraverso il secondo generatore (che è spento).

$$i_{p_2}^{(1)} = i_{p_1-r}^{(1)} = i_{p_1}^{(1)} \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}}$$

Effetto del generatore di prova 2



In questo caso non si ha alcun ricircolo. La R_{down} è la resistenza di uscita di uno specchio cascode, all'incirca:

$$R_{dw} \cong r_{d_9} g_{m_9} r_{d_{11}}$$

Dunque:

$$i_{p_2}^{(2)} = \frac{v_p}{R_{dw}}$$

$$i_{p_1}^{(2)} = 0$$

Sommendo tutti gli effetti:

$$i_p = \frac{v_p}{R_{up}} \left(1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}} \right) + \frac{v_p}{R_{dw}} = v_p \left(\frac{1}{R_{up}} \left(1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}} \right) + \frac{1}{R_{dw}} \right)$$

Dunque, la resistenza di uscita:

$$R_{out} = \frac{v_p}{i_p} = \left(\frac{1}{R_{up}} \left(1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}} \right) + \frac{1}{R_{dw}} \right)^{-1} = \left(\frac{\frac{1}{R_{up}} \left(1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}} \right)}{1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}}} + \frac{1}{R_{dw}} \right)^{-1}$$

La forma è quella del parallelo di due resistenze. Una è la R_{dw} , l'altra è la R_{up} corretta di un fattore che tiene conto del ricircolo. Per cui:

$$\frac{R_{up}}{1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}}} = R_{up-r} \rightarrow R_{out} = R_{up-r} \parallel R_{dw}$$

Il ricircolo ha sempre l'effetto di diminuire la corrispondente resistenza vista. Nel caso precedentemente studiato il ricircolo era completo, per cui la resistenza vista veniva dimezzata. In questo caso il ricircolo è parziale, per cui si ottiene un fattore di abbattimento minore di 2.

$$\begin{cases} R_{dw} = r_{d_9} g_{m_9} r_{d_{11}} \\ R_{up-r} = \frac{r_{d_7} g_{m_7} (r_{d_3} \parallel 2r_{d_1})}{1 + \frac{r_{d_3}}{r_{d_3} + 2r_{d_1}}} \end{cases}$$

Anche la R_{up-r} è nell'ordine della resistenza di uscita di uno stadio cascode. Quindi, la R_{out} complessiva dell'amplificatore folded cascode è ottima. Se si considerano le r_d e i g_m tutti uguali:

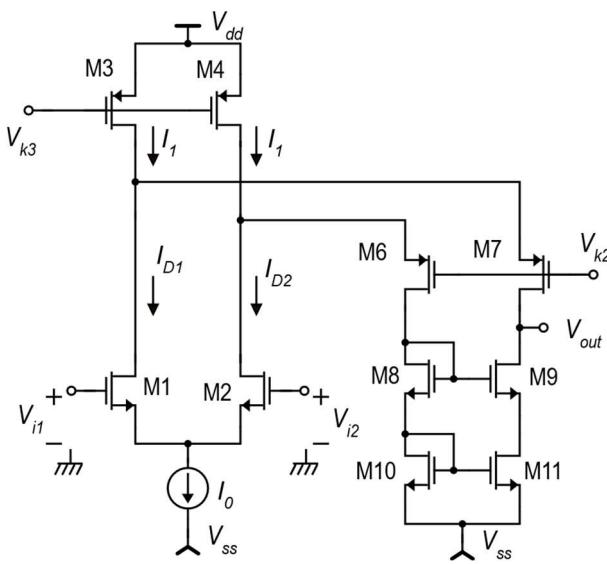
$$\begin{cases} r_{d_2} = r_{d_3} = r_{d_7} = r_{d_9} = r_{d_{11}} = r_d \\ g_{m_7} = g_{m_9} = g_m \end{cases} \rightarrow \begin{cases} R_{dw} = r_d(g_m r_d) \\ R_{up-r} = \frac{r_d(g_m r_d)}{2} \end{cases} \rightarrow R_{out} = \frac{r_d(g_m r_d)}{3}$$

L'amplificazione di modo differenziale, sempre nell'approssimazione di g_m tutti uguali:

$$A_d = -g_{m_1} R_{out} = -g_m R_{out} = -\frac{(g_m r_d)^2}{3}$$

Rispetto all'amplificatore telescopico, il guadagno è più piccolo di un fattore $3/2$. Ad ogni modo, il miglioramento del guadagno rispetto all'amplificatore con carico a specchio semplice è notevole. In particolare, il folded cascode guadagna di più di una cascata di due amplificatori con carico a specchio semplice (il cui guadagno è $g_m r_d/2$).

Metodo per la determinazione dei terminali invertente/non invertente

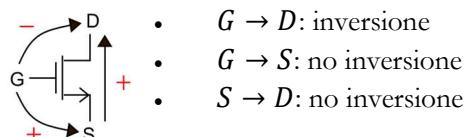


$$v_d = v_{i_1} - v_{i_2}$$

$$A_d = -g_{m_1} R_{out} \rightarrow v_{out} = g_{m_1} R_{out} (v_{i_2} - v_{i_1})$$

Arrivati all'espressione della tensione di uscita, il terminale invertente è quello con segno negativo. C'è però un altro metodo più topologico e più semplice per determinare quale dei due ingressi è quello invertente.

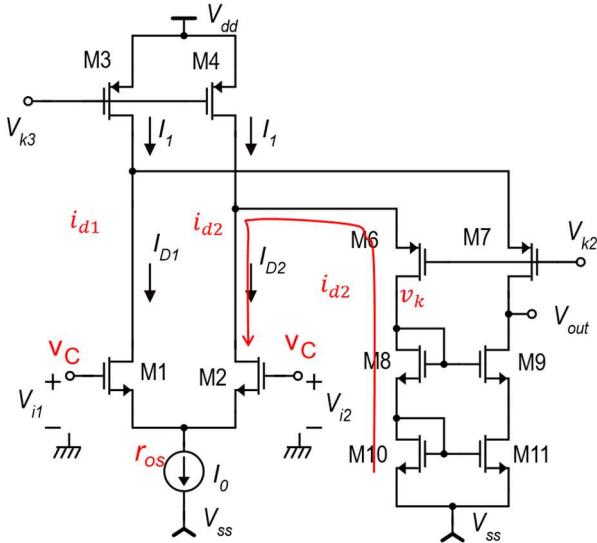
Considerando un transistore, la tensione che controlla è la V_{GS} . L'input, quindi, può essere o il gate o il source. Gli output possono essere o il drain o il source.



Analizzando il percorso del segnale tra i terminali dei vari transistori, si determinano tutte le inversioni tra un ingresso e l'uscita. Quando ci sono più percorsi tra l'ingresso e l'uscita occorre maggiore attenzione nell'applicare il metodo.

Amplificazione di modo comune e CMRR

L'amplificatore folded cascode, come l'amplificatore con carico a specchio, è simmetrico in termini di soluzione elettrica in caso di stimolo simmetrico.



$$i_{d_1} = i_{d_2} \cong \frac{v_c}{2r_{os}}$$

$$v_{out} = v_k = -i_{d_2} \left(\frac{1}{g_{m_8}} + \frac{1}{g_{m_{10}}} \right) = -\frac{v_c}{2r_{os}} \left(\frac{1}{g_{m_8}} + \frac{1}{g_{m_{10}}} \right)$$

Considerando ancora una volta i g_m dei transistori uguali:

$$v_{out} \cong -\frac{v_c}{g_m r_{os}}$$

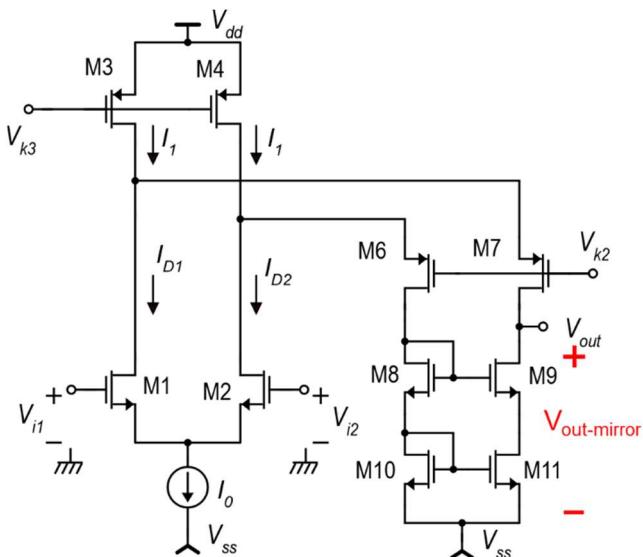
Dunque, l'amplificazione di modo comune:

$$A_c = -\frac{1}{g_m r_{os}}$$

$$CMRR \approx \frac{(g_m r_d)^2}{3} g_m r_{os}$$

Il CMRR è nell'ordine di $(g_m r_d)^3$ se si assume $r_{os} \cong r_d$. Se per $g_m r_d$ si assume un valore di 100, il CMRR del folded cascode si attesta approssimativamente a 333333, cioè 110 dB.

Dinamica della tensione di uscita



Limite inferiore

Se la tensione di uscita oltre la minima tensione di uscita dello specchio cascode, il funzionamento dell'amplificatore si degrada (crolla la resistenza di uscita, l'amplificazione dipende dal segnale → distorsione).

$$\min(V_{out}) = V_{ss} + V_{min-cascode}$$

Esplicitando la V_{min} dello specchio cascode:

$$\min(V_{out}) = V_{ss} + V_{GS_{11}} + V_{DS_{sat}}$$

Quindi, l'uscita dell'amplificatore cascode non può raggiungere il rail di alimentazione basso, soprattutto a causa della $V_{GS_{11}}$.

Limite superiore

La V_{out} coincide al potenziale di drain di M7. Il source di M7 si trova a un potenziale quasi costante. Infatti, fissata la V_{k_2} si fissa $|V_{GS_7}|$. Le variazioni di corrente in tutti i rami, essendo l'amplificazione così grande, sono piccole nella zona di funzionamento lineare dell'amplificatore, per cui la $|V_{GS_7}|$ non cambierà significativamente. Dunque, aumentando la V_{out} si schiaccia la $|V_{DS_7}|$ e quando questa arriva alla $|V_{DS_{sat}}|_7$ si riduce la resistenza di uscita. Pertanto, il limite superiore della tensione di uscita sarà dato dall'entrata in triodo di M7.

$$|V_{DS_7}| = V_{S_7} - V_{D_7} = V_{k_2} + |V_{GS_7}| - V_{out} \geq |V_{DS_{sat}}|_7$$

$$V_{k_2} + |V_{GS_7}| - |V_{DS_{sat}}|_7 \geq V_{out} \rightarrow \max(V_{out}) = V_{k_2} + |V_{GS_7}| - |V_{DS_{sat}}|_7$$

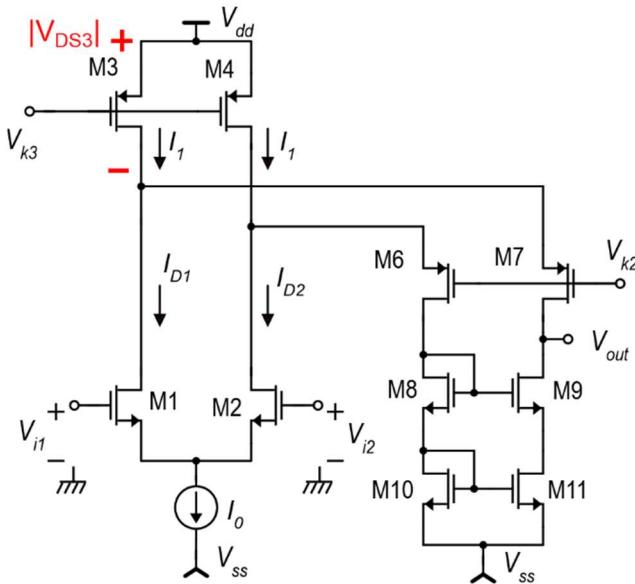
In forte inversione:

$$|V_{DS_{sat_7}}| = |V_{GS_7}| - |V_{t_7}|$$

$$\max(V_{out}) = V_{k_2} + |V_{t_7}|$$

Massima tensione V_{k_2}

Il limite superiore della tensione di uscita non contiene la V_{dd} , per cui non può essere confrontato con il rail superiore di alimentazione. Quindi, ricerchiamo il massimo della tensione V_{k_2} rispetto alla tensione V_{dd} .



Se V_{k_2} sale, i source di M6 e M7 salgono poiché la loro V_{GS} è costante. I source di M6 e M7 sono anche i drain di M3 e M4 e di M1 e M2. Il limite superiore della V_{k_2} , quindi, è quello per cui le V_{DS} dei transistori M3 e M4, i cui source sono fissi a V_{dd} , raggiungono la $V_{DS_{sat}}$. Concentrando l'analisi su M3:

$$|V_{DS_3}| = V_{dd} - V_{D_3}$$

$$\begin{aligned} V_{D_3} &= V_{S_7} = V_{k_2} + |V_{GS_7}| \rightarrow |V_{DS_3}| \\ &= V_{dd} - V_{k_2} - |V_{GS_7}| \end{aligned}$$

$$|V_{DS_3}| \geq |V_{DS_{sat_3}}| \rightarrow V_{dd} - V_{k_2} - |V_{GS_7}| \geq |V_{DS_{sat_3}}|$$

Quindi, la massima tensione V_{k_2} :

$$\max(V_{k_2}) = V_{dd} - |V_{DS_{sat_3}}| - |V_{GS_7}|$$

Quindi, la massima tensione di uscita:

$$\max(V_{out}) = \max(V_{k_2}) + |V_{t_7}| = V_{dd} - |V_{DS_{sat_3}}| - |V_{DS_{sat_7}}|$$

Con una struttura cascode meglio di così non si può fare. La massima V_{out} si ottiene soltanto quando anche la V_{k_2} è al massimo.

Range di modo comune in ingresso

La minima tensione di modo comune in ingresso è sempre la stessa:

$$\min(V_C) = V_{ss} + V_{min-tail} + V_{GS_1}$$

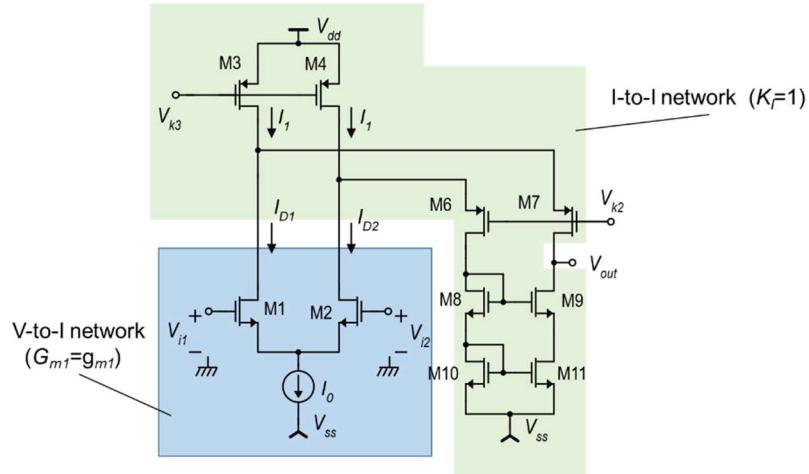
All'aumentare della tensione di modo comun, mentre i drain di M1 e M2 sono bloccati poiché la V_{k_2} è costante, il loro source aumenta. Dunque, il limite superiore della tensione di modo comune è quello per cui M1 e M2 entrano in zona triodo.

$$V_{DS_1} = V_{D_1} - V_{S_1} = V_{k_2} + |V_{GS_7}| - (V_C - V_{GS_1}) \geq V_{DS_{sat_1}}$$

$$\max(V_C) = \max(V_{k_2}) + |V_{GS_7}| + V_{GS_1} - V_{DS_{sat_1}} = V_{dd} - |V_{DS_{sat_3}}| + V_{GS_1} - V_{DS_{sat_1}}$$

In questo caso la V_{k_2} influenza il limite superiore della tensione di uscita e il limite superiore della tensione di modo comune. Per l'amplificatore folded cascode non c'è un compromesso sulla V_{k_2} : più grande è meglio è.

Il limite superiore della tensione di modo comune supera il rail positivo della tensione di alimentazione. Inoltre, ottenere la massima dinamica per la tensione di modo comune non peggiora la dinamica della tensione di uscita. L'amplificatore folded cascode, grazie a queste caratteristiche, può fare da amplificatore operazionale. Guardando alla struttura del folded cascode come single stage:



La rete di trasporto contiene lo specchio cascode, lo stadio a gate comune e i generatori di corrente I_1 .

Sommario delle proprietà

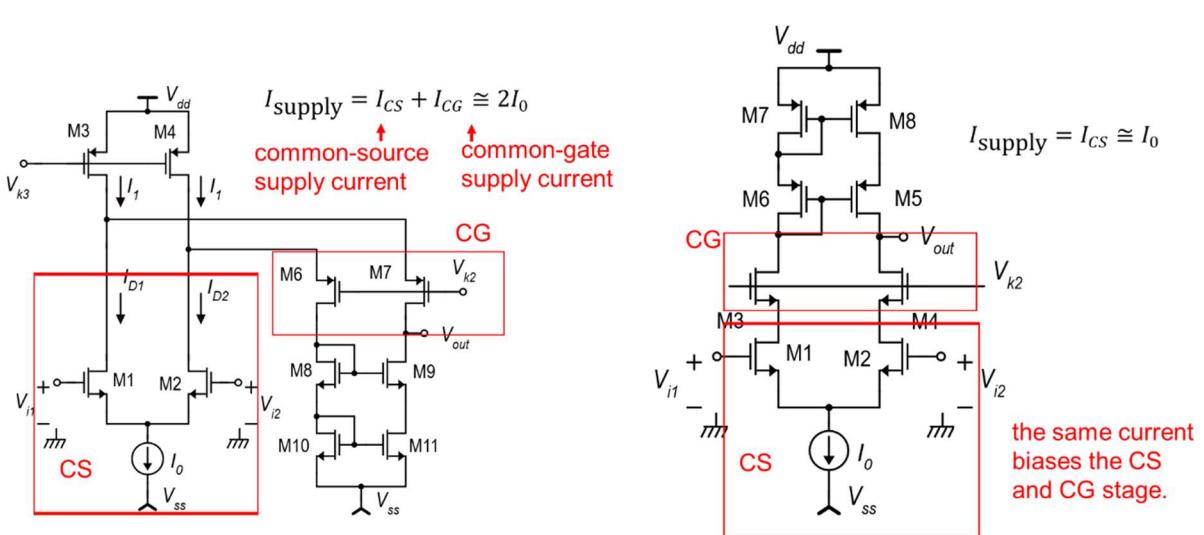
- Guadagno DC: il guadagno del folded cascode è leggermente inferiore a quello dell'amplificatore telescopico. Operando a bassi overdrive e aumentando la lunghezza dei MOSFET il folded cascode arriva a guadagnare anche 10^4 (80 dB).
- Range: per l'amplificatore cascode il range di uscita non è affatto dal range di modo comune di ingresso. Il risultato è che i range di modo comune in ingresso e di tensione di uscita sono molto più ampi di quelli dell'amplificatore telescopico.
- Range di uscita: $\begin{cases} \max(V_{out}) = V_{dd} - |V_{DS_{sat3}}| - |V_{DS_{sat7}}| \\ \min(V_{out}) = V_{ss} + V_{GS_{11}} + V_{DS_{sa_9}} \end{cases}$

La massima tensione di uscita si avvicina molto al rail V_{dd} . La minima tensione di uscita potrebbe avvicinarsi al rail inferiore V_{ss} se come generatore di corrente di tail si utilizzasse uno specchio a larga dinamica di precisione.

- Range di modo comune in ingresso: $\begin{cases} \min(V_{ic}) = V_{ss} + V_{min-tail} + V_{GS_1} \\ \max(V_{ic}) = V_{dd} - |V_{DS_{sat3}}| + V_{GS_1} - V_{DS_{sat1}} \end{cases}$

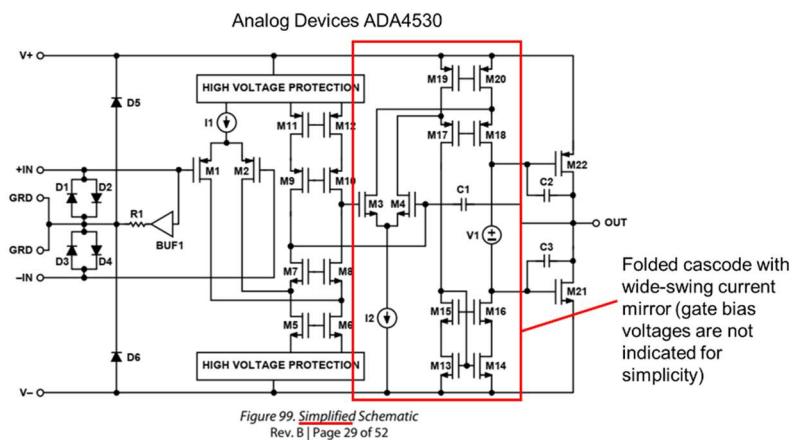
Mentre la massima tensione di modo comune di ingresso supera il rail positivo di alimentazione, quella minima è abbastanza lontana dal rail inferiore ed è questo l'unico vero limite di questo amplificatore.

Consumo di corrente



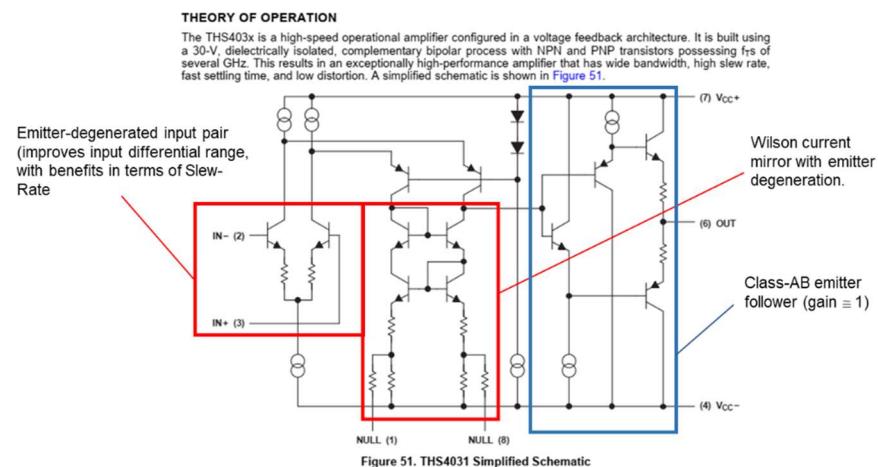
L'amplificatore cascode necessita di una corrente continua sia per la parte common source che per la parte common gate. Complessivamente, ponendo $I_1 = I_0$, il consumo di corrente è circa $2I_0$. Per l'amplificatore telescopico, invece, la stessa corrente che polarizza lo stadio a source comune polarizza anche lo stadio a gate comune (current re-use); il consumo totale è limitato a I_0 . A parità di corrente di polarizzazione nei transistori M1 e M2, l'amplificatore folded cascode consuma il doppio di corrente statica.

Esempi commerciali



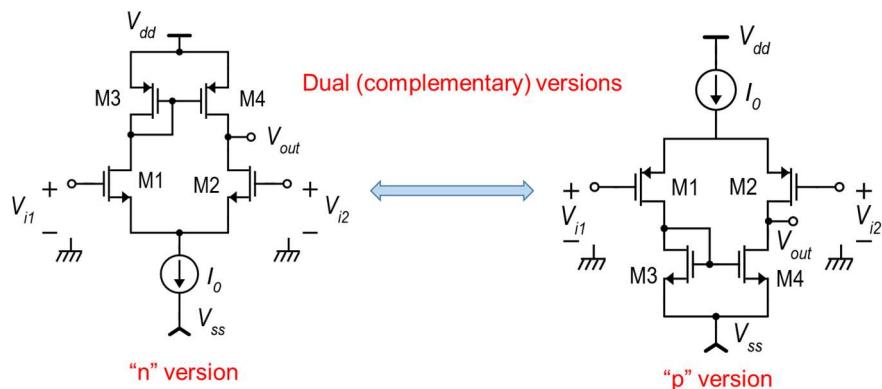
Si osserva la presenza di un folded cascode anche in ingresso nella versione fully differential duale: coppia differenziale a pMOS, gate comune di tipo n.

Un esempio di amplificatore folded cascode a bjt:

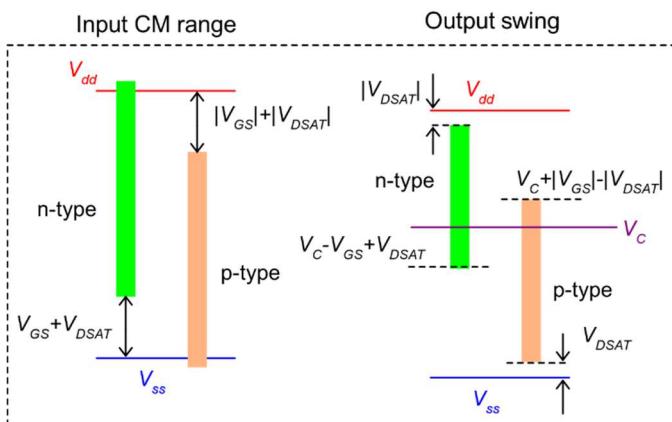


Lo stadio di uscita ha la funzione di adattare l'impedenza. Il suo guadagno è unitario ma offre una bassa impedenza di uscita in modo tale da avvicinare il comportamento del circuito a quello di un amplificatore di tensione ideale.

Versioni duali dei circuiti CMOS



Dell'amplificatore differenziale con carico a specchio semplice esiste una versione perfettamente equivalente ai piccoli segnali a transistori p. La coppia di ingresso è di tipo p ed è polarizzata da una corrente di tail di tipo sourcing. Una delle due correnti che esce dalla coppia dovrà essere ribaltata e portata in uscita. Lo specchio che accetta correnti di tipo sourcing è di tipo n. I risultati nell'analisi di piccolo segnale sono identici. A cambiare sono i range.

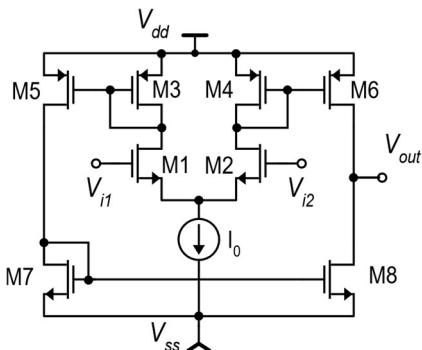


I range di modo comune in ingresso e della tensione di uscita sono simmetrici. A seconda dei range in cui deve lavorare il circuito si può scegliere una versione o la complementare. Ad esempio, molte volte le sorgenti di segnale differenziale non sono in grado di produrre una tensione di modo comune più grande della V_{ss} ; in tal caso, la versione p potrebbe essere avvantaggiata.

Un altro motivo di scelta tra le due versioni è il rail di riferimento della tensione di uscita e della tensione di ingresso. Nel caso dell'amplificatore di tipo n la tensione di uscita è riferita alla V_{dd} : variando la tensione di alimentazione, la differenza di potenziale rispetto alla V_{dd} non cambia. Dunque, uno stadio di tipo n si accoppia bene in uscita con uno stadio di tipo p. Nel caso dell'amplificatore di tipo p la tensione di uscita, a riposo, è fissata non più rispetto alla V_{dd} ma rispetto alla V_{ss} . Tutti gli amplificatori finora studiati possono essere realizzati in versione p.

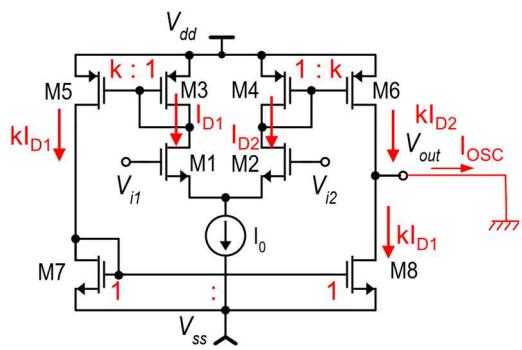
OTA – operational transconductance amplifier

Tutti gli amplificatori visti finora vengono in realtà caratterizzati come OTA. Un amplificatore transconduttivo, in generale, è un amplificatore il cui segnale di ingresso è una tensione e il segnale di uscita è una corrente. Gli amplificatori differenziali trattati hanno una chiara relazione tra la tensione differenziale di ingresso e la corrente di uscita e, guardati dall'uscita, si comportano bene come generatori di corrente (elevata resistenza di uscita – equivalente di Norton). Anche se poi il segnale di interesse è la tensione di uscita, prodotta dalla corrente di uscita attraverso la resistenza di uscita intrinseca dell'amplificatore, il termine OTA potrebbe comunque essere utilizzato.



Storicamente, il termine OTA è stato utilizzato per identificare la topologia di amplificatore a fianco. Il vantaggio principale di questa tipologia è quello di rimuovere l'interazione tra la dinamica di uscita e il modo comune (con tecnica diversa dal folding).

$$A_d \cong G_m R_{out}$$



Si immagina di porre un cortocircuito in uscita alle variazioni. La corrente di cortocircuito:

$$I_{osc} \cong k(I_{D_2} - I_{D_1}) \cong -k \cdot g_{m_1} v_d$$

Quindi il G_m complessivo può essere modulato progettando il fattore k :

$$G_m = -k g_{m_1} = -k \frac{I_{D_1}}{V_{TE_1}}$$

Si procede calcolando la resistenza di uscita R_{out} :

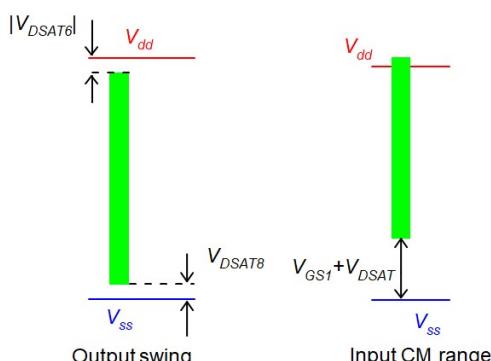
$$R_{out} = r_{d_6} \parallel r_{d_8} = \frac{1}{\lambda_6 I_{D_6} + \lambda_8 I_{D_8}}$$

$$I_{D_6} = I_{D_8} = kI_{D_1} \rightarrow R_{out} = \frac{1}{kI_{D_1}(\lambda_6 + \lambda_8)}$$

L'amplificazione a modo differenziale:

$$A_d = -k \frac{I_{D_1}}{V_{TE_1}} \frac{1}{kI_{D_1}(\lambda_6 + \lambda_8)} = - \frac{1}{V_{TE_1}(\lambda_6 + \lambda_8)}$$

Il k non ha influenza sull'amplificazione di uscita, ma influisce soltanto sulla corrente di cortocircuito in uscita.



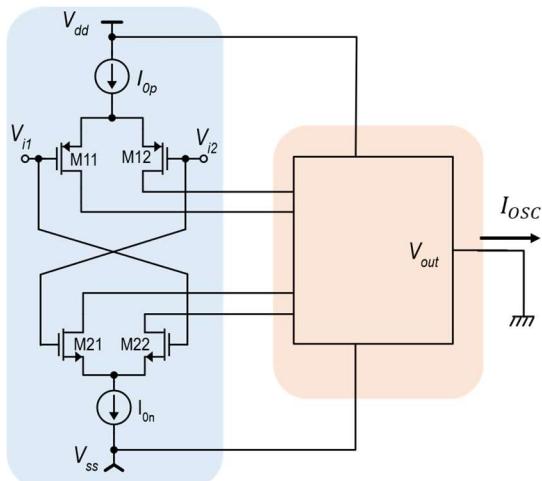
Il pregio di questo amplificatore è il range della tensione di uscita, che raggiunge i rail dell'alimentazione a meno delle V_{DSsat} dei transistori M6 e M8 (rail to rail output). L'uscita potrebbe essere anche forzata ai rail di alimentazione accettando una perdita di guadagno. Per quanto riguarda il modo comune di ingresso, il range è sempre vincolato alla presenza di una coppia e ad un generatore di tail, per cui non si ottengono notevoli miglioramenti in tal senso.

Questo circuito, data la variabilità del g_m con la temperatura, con il processo e con la corrente di bias, non può comportarsi da convertitore tensione-corrente accurato.

Anche l'utilizzo che ne considera la tensione di uscita è valido. Ponendosi in quest'ottica, rispetto all'amplificatore con carico a specchio, il guadagno è sempre nell'ordine di $g_m r_d / 2$. Tuttavia, il range della tensione di uscita è notevolmente migliore. Rispetto all'amplificatore con carico a specchio, uno svantaggio di questa topologia è un maggiore consumo di corrente statica dovuto alla presenza di più rami. Inoltre, le correnti passano da una rete più complessa prima di giungere all'uscita. Ogni volta che la corrente passa da uno specchio di corrente subentrano degli errori dovuti al matching. Si hanno anche errori sistematici: in uscita, a riposo, non si ha una corrente di cortocircuito nulla, il che si traduce in una tensione di offset. Gli errori di matching, soprattutto se le aree sono piccole, possono anche essere molto grandi rispetto a quelli sistematici.

OTA con dinamica di ingresso rail-to-rail

Dell'OTA discusso esiste anche la versione p, che presenta la stessa dinamica di uscita e una dinamica di modo comune di ingresso simmetrica. L'obiettivo è quello di ottenere una versione di OTA che presenti anche una dinamica di modo comune in ingresso rail-to-rail. In questo modo sarebbe possibile utilizzare l'amplificatore come buffer di tensione. Spesso per testare i prototipi di circuiti integrati si portano all'esterno, oltre che ingressi e uscite, anche nodi intermedi per diagnosi. Poiché questi nodi potrebbero in linea teorica raggiungere qualsiasi potenziale rispetto all'alimentazione, anche a causa di malfunzionamenti, è importante che l'amplificatore che si interfaccia con questi nodi bufferizzandone la tensione all'uscita del package abbia una dinamica di ingresso rail-to-rail. La tecnica con cui si ottiene un range di ingresso rail-to-rail è generale e si applica a più circuiti.



La rete di conversione V-I è formata da due coppie differenziali complementari. Le correnti prodotte dalle coppie vengono elaborate da una rete di trasporto di corrente I-I e convertite nella corrente di cortocircuito I_{OSC} . L'incrocio dei terminali di ingressi è necessario per far sì che le correnti si sommino in fase anziché in controfase.

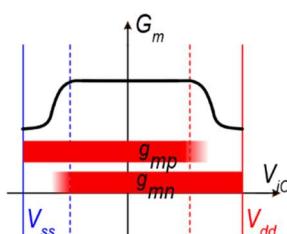
$$I_{OSC} = (I_{D_{22}} - I_{D_{21}}) + (I_{D_{12}} - I_{D_{11}})$$

Al piccolo segnale:

$$I_{OCC} \cong g_{m_n} v_d + g_{m_p} v_d$$

Il G_m complessivo sarà dato dalla somma dei due g_m :

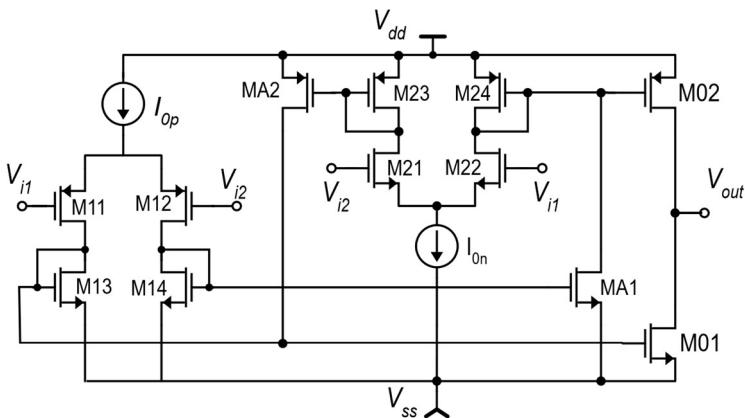
$$G_m = g_{m_n} + g_{m_p}$$



Quando la tensione di modo comune è insufficiente per polarizzare il tail di una coppia, non essendoci corrente di polarizzazione, il g_m di quella coppia crolla. Tuttavia, nelle regioni in cui una coppia smette di funzionare rimane l'altra e contribuisce al G_m complessivo. Questo fa sì che qualunque tensione di modo comune sia presente in ingresso, almeno una coppia è attiva nel trattamento del segnale.

Il potenziale problema di questo amplificatore è che nelle zone in cui cambia il G_m con il modo comune cambia anche il guadagno. Se l'amplificatore è utilizzato come buffer, il modo comune viene a coincidere con il segnale utile, per cui si altera il guadagno del buffer al variare del segnale utile. Tuttavia, quando una coppia si spegne sono disattivate anche le componenti continue oltre che quelle di riposo. Quindi, annullandosi le correnti di quella coppia verso la rete di trasporto I-I, i transistori della rete I-I, in particolare quelli di uscita, vedono dimezzarsi la loro corrente di riposo, per cui la resistenza di uscita raddoppia. Quindi, il guadagno tende a rimanere costante al variare della tensione di modo comune su tutto il range rail-to-rail.

Un esempio di OTA con dinamica di ingresso rail-to-rail:



La corrente $I_{D_{22}}$ viene specchiata da M24-M02 e si ritrova in uscita con segno positivo. Analogamente, la corrente I_{D_1} viene specchiata da M13-M01 e si ritrova in uscita con segno negativo. La corrente $I_{D_{21}}$ viene specchiata da M23-MA2 ed entra nello specchio M13-M01, per cui si ritrova in uscita con segno negativo. La corrente $I_{D_{12}}$ viene specchiata da M14-MA1 e da M24-M02, confluendo all'uscita con segno positivo.

$$I_{occ} = (I_{D_{22}} - I_{D_{21}}) + (I_{D_{12}} - I_{D_{11}})$$

Un prodotto commerciale:

OPA354, OPA2354, OPA4354
SBOS233G – MARCH 2002 – REVISED APRIL 2018



www.ti.com

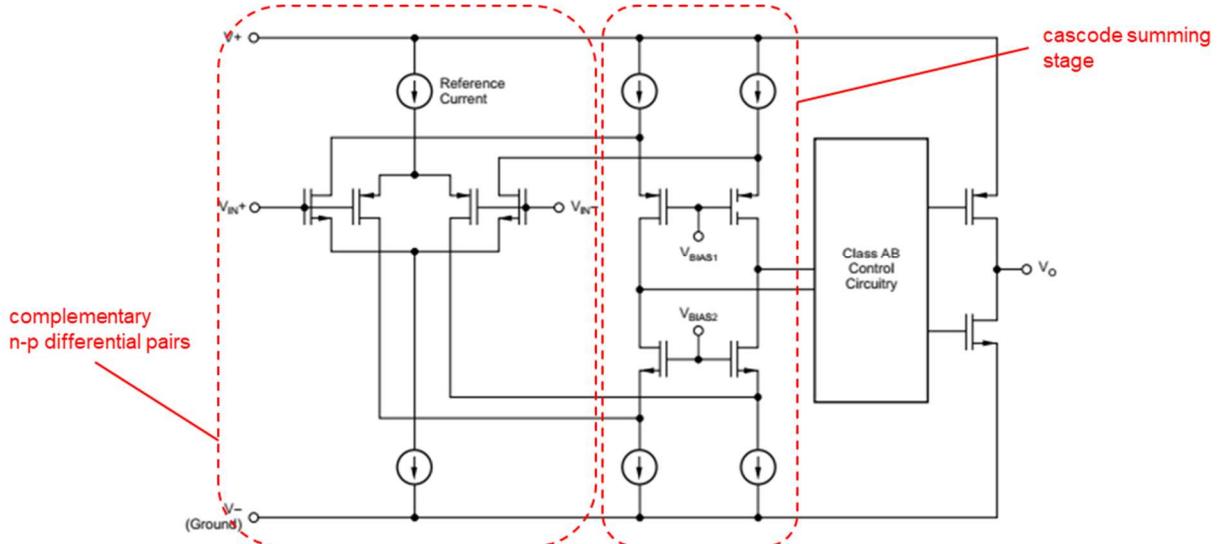
8 Detailed Description

8.1 Overview

The OPAx354 is a CMOS, rail-to-rail I/O, high-speed, voltage-feedback operational amplifier designed for video, high-speed, and other applications. It is available as a single, dual, or quad op amp.

The amplifier features a 100-MHz gain bandwidth, and 150-V/μs slew rate, but the amplifier is unity-gain stable and can operate as a 1-V/V voltage follower.

8.2 Functional Block Diagram



Nei circuiti in cui è problematico il fatto che il g_m cambia esistono circuiti denominati “constant g_m ” che aumentano artificialmente la corrente di tail della coppia sta funzionando mentre l'altra sta morendo.

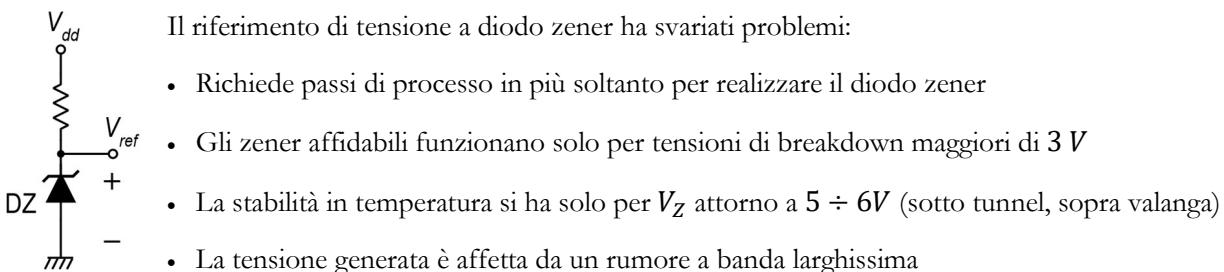
Riferimenti di tensione

I riferimenti di tensione (o di corrente) producono una tensione indipendente dal “PVT”:

- “P”: processo
- “V”: tensione di alimentazione
- “T”: temperatura

Spesso l'invarianza rispetto al processo è più difficile da ottenere, ma si può contrastare aggiungendo meccanismi di compensazione. L'utilizzo più comune dei riferimenti di tensione è quello di produrre la V_{ref} di ADC e DAC. Un altro utilizzo è quello di generare un riferimento stabile per tensioni di stimolo di sensori o altri sistemi che richiedono controlli di tensione precisi. Oppure, si possono utilizzare riferimenti di tensione direttamente per polarizzare circuiti che richiedono tensioni di bias precise.

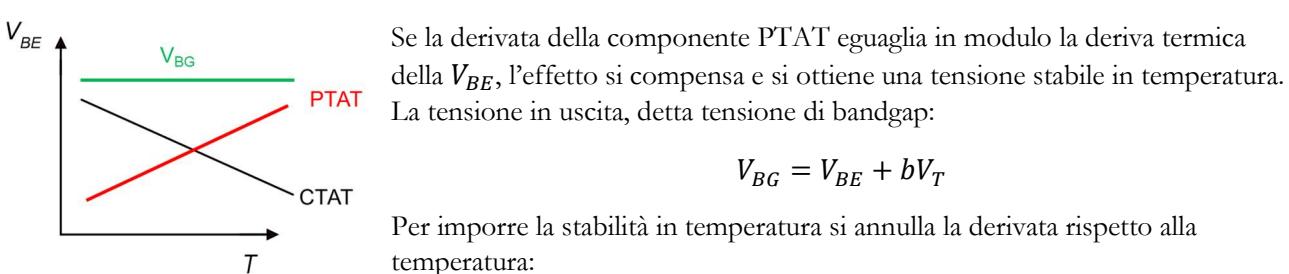
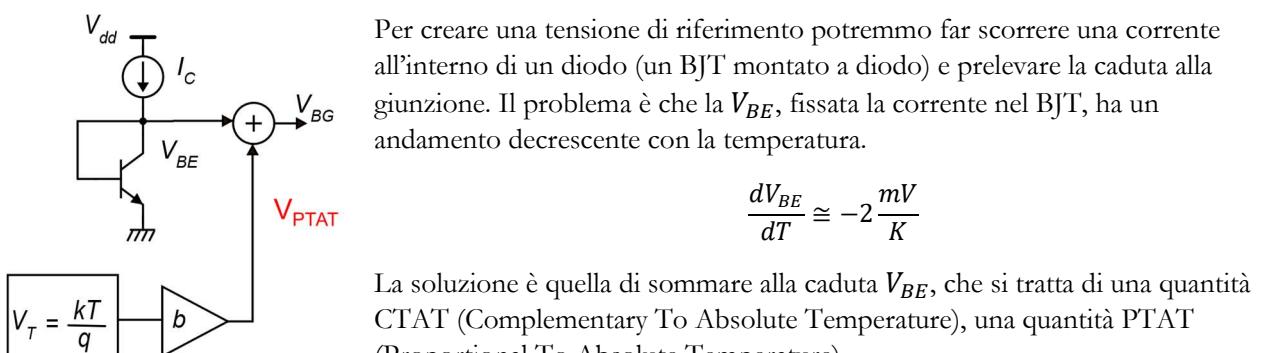
Riferimento di tensione a diodo zener



Nei processi integrati moderni gli zener sono utilizzati soltanto in processi ad alte tensioni per protezione.

Riferimento di tensione a bandgap

L'obiettivo è quello di ottenere un riferimento di tensione stabile in temperatura e rispetto alle variazioni della tensione di alimentazione.



$$\frac{dV_{BG}}{dT} = \frac{dV_{BE}}{dT} + b \frac{dV_T}{dT} = 0 \rightarrow b = -\frac{\frac{dV_{BE}}{dT}}{\frac{dV_T}{dT}}$$

Ad esempio, $\frac{dV_{BE}}{dT} = -2 \frac{mV}{K}$ e $\frac{dV_T}{dT} = \frac{k}{q} = 8.56 \cdot 10^{-5} \frac{V}{K}$ si ottiene $b \cong 23 \rightarrow V_{BG} \cong 0.65 + 0.025 \cdot 23 = 1.225V$.

La tensione stabilizzata è abbastanza piccola, per cui è compatibile con circuiti alimentati a bassa tensione (3.3 V , 2.5 V , 1.8 V).

Teoria del riferimento bandgap

La tensione V_{BE} :

$$V_{BE} = V_T \ln\left(\frac{I_C}{I_S}\right)$$

La corrente inversa di saturazione:

$$I_S = \frac{qA_E n_i^2 D_n}{Q_B} = F \cdot n_i^2 D_n$$

Dove F è un termine che raccoglie i parametri costanti con la temperatura. I termini che dipendono dalla temperatura:

$$\begin{cases} n_i^2 \propto T^3 e^{-\frac{E_{g_0}}{kT}} \\ D_n = \mu_n \frac{KT}{q} \\ \mu_n \propto T^{-\alpha_\mu} \end{cases}$$

Riscrivendo la corrente di saturazione inversa:

$$I_S = BT^\gamma e^{-\frac{E_{g_0}}{kT}} = BT^\gamma e^{-\frac{V_{g_0}}{V_T}}$$

Dove $\alpha_\mu \cong 1.5$, $\gamma = 4 - \alpha_\mu$

Inoltre, non è detto che la corrente che polarizza la giunzione sia costante in temperatura. In generale:

$$I_C = GT^\alpha$$

Dove B e G sono costanti che non dipendono dalla temperatura. Dunque, la V_{BE} :

$$\begin{aligned} V_{BE} &= V_T \ln\left(\frac{GT^\alpha}{BT^\gamma \exp\left(-\frac{V_{g_0}}{V_T}\right)}\right) = V_{g_0} + V_T \left[\ln\left(\frac{G}{B}\right) - (\gamma - \alpha) \ln(T) \right] \\ E &= \frac{1}{B} \rightarrow V_{BE} = V_{g_0} + V_T [\ln(G \cdot E) - (\gamma - \alpha) \ln(T)] \end{aligned}$$

La tensione di bandgap:

$$V_{BG} = V_{BE} + bV_T \rightarrow V_{BG} = V_{g_0} + V_T [\ln(G \cdot E) + b - (\gamma - \alpha) \ln(T)] = V_{g_0} + \frac{kT}{q} [\ln(G \cdot E) + b - (\gamma - \alpha) \ln(T)]$$

Il nome “bandgap” deriva dal fatto che il termine dominante della tensione di uscita è la tensione equivalente all’energia del bandgap del silicio V_{g_0} . Per ottenere il valore di b che permette la stabilità in temperatura si deriva l’espressione rispetto alla temperatura.

$$\frac{dV_{BG}}{dT} = \frac{k}{q} [\ln(G \cdot E) + b - (\gamma - \alpha) \ln(T)] - (\gamma - \alpha) \frac{kT}{q} \frac{1}{T} = \frac{k}{q} [\ln(G \cdot E) + b - (\gamma - \alpha) - (\gamma - \alpha) \ln(T)]$$

Si scopre che in realtà anche la derivata dipende dalla temperatura, per cui la tensione V_{BG} non sarà realmente costante in temperatura. Potremmo usare l’espressione, quindi, per trovare quel valore di b che annulla la derivata ad una certa temperatura $T = T_0$.

$$\frac{dV_{BG}}{dT} = \frac{k}{q} [\ln(G \cdot E) + b - (\gamma - \alpha) - (\gamma - \alpha) \ln(T)] = 0 \rightarrow \ln(G \cdot E) + b - (\gamma - \alpha) - (\gamma - \alpha) \ln(T_0) = 0$$

Rielaborando:

$$\ln(G \cdot E) + b = (\gamma - \alpha) + (\gamma - \alpha) \ln(T_0)$$

Il termine a sinistra dell'equazione si trova anche nell'espressione della tensione di bandgap. Per cui, sostituendo:

$$V_{BG} = V_{g_0} + V_T [\ln(G \cdot E) + b - (\gamma - \alpha) \ln(T)] = V_{g_0} + V_T [(\gamma - \alpha) + (\gamma - \alpha) \ln(T_0) - (\gamma - \alpha) \ln(T)]$$

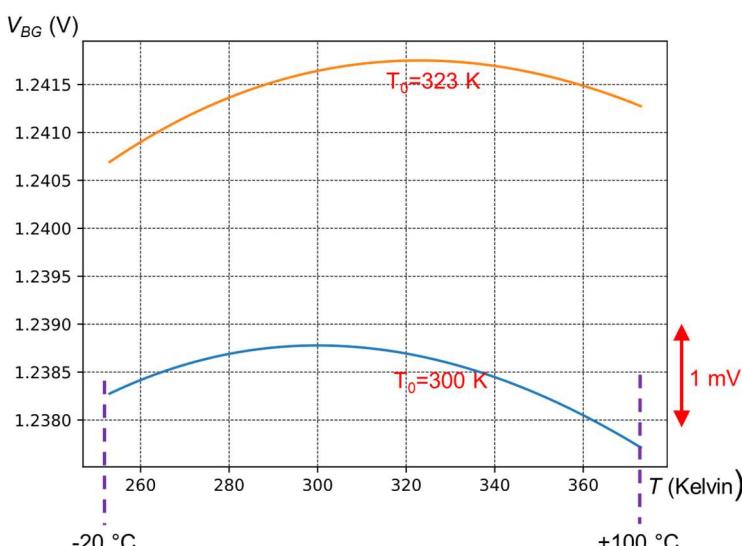
Dunque:

$$V_{BG} = V_{g_0} + V_T (\gamma - \alpha) \left(1 + \ln\left(\frac{T}{T_0}\right) \right)$$

La formula appena ottenuta permette di poter calcolare la tensione di bandgap in funzione della temperatura, fissata la temperatura T_0 a cui si annulla la derivata.

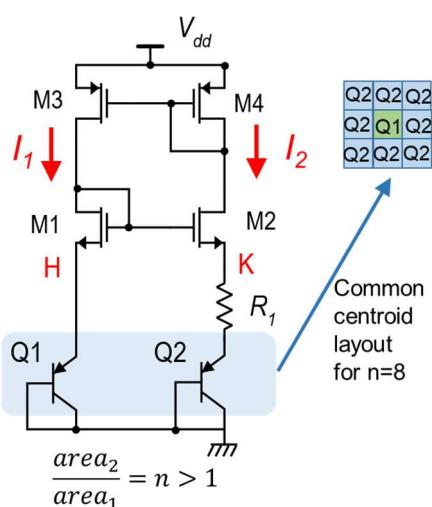
$$V_{BG}(T_0) = V_{g_0} + \frac{kT_0}{q} (\gamma - \alpha)$$

Nel punto in cui la derivata si annulla, la tensione di bandgap dipende dalla T_0 stessa.



Come si osserva a fianco, anche se la derivata è nulla solo per una certa temperatura, la dipendenza della tensione di bandgap dalla temperatura assoluta è molto lieve su un ampio intervallo di temperature, da -20°C a $+100^{\circ}\text{C}$. Come si nota, un riferimento di tensione progettato in modo tale da avere una T_0 più alta avrà anche una tensione di uscita maggiore.

Riferimento di tensione a bandgap CMOS – compatibile



Il circuito sulla sinistra ha la funzione di generare una corrente proporzionale alla temperatura assoluta (PTAT current generator).

I bipolari sono pnp di substrato (disponibili in tutti i processi CMOS) montati a diodo. Il Q2 è più grande di Q1 in termini di area di un fattore 8. M3 e M4 nominalmente identici e realizzano uno specchio con coefficiente di specchio unitario:

$$I_1 = I_2 = I$$

M1 e M2 sono nominalmente identici; essendo percorsi dalla stessa corrente:

$$V_{GS_1} = V_{GS_2}$$

$$V_H - V_K = (V_{G1} - V_{GS_1}) - (V_{G_2} - V_{GS_2}) = V_{GS_2} - V_{GS_1} \rightarrow V_H = V_K$$

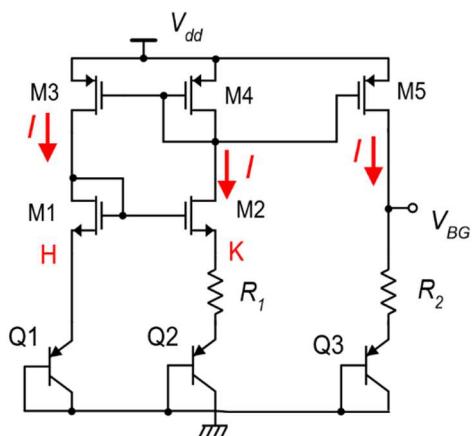
I nodi H e K:

$$V_H = |V_{BE_1}| \quad V_K = |V_{BE_2}| + R_1 I$$

Dunque:

$$R_1 I = |V_{BE_1}| - |V_{BE_2}| = V_T \ln \left(\frac{I_{C_1} I_{S_2}}{I_{S_1} I_{C_2}} \right) = V_T \ln \left(\frac{I_{S_2}}{I_{S_1}} \right) = V_T \ln(n) \rightarrow I = \frac{1}{R_1} \frac{kT}{q} \ln(n)$$

Si ottiene che la corrente I è proporzionale alla temperatura e indipendente dalla tensione di alimentazione (a patto che la resistenza si possa considerare abbastanza costante con la temperatura). Per ottenere il riferimento di tensione a bandgap si aggiunge un altro ramo:



La tensione di bandgap:

$$V_{BG} = |V_{BE_3}| + IR_2$$

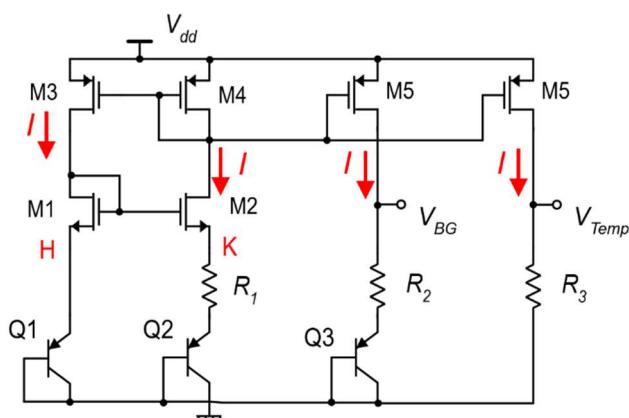
La giunzione di Q3 è polarizzata con una corrente PTAT, per cui nella trattazione precedente si avrebbe $\alpha = 1$. Sostituendo l'espressione della corrente I :

$$V_{BG} = |V_{BE_3}| + \frac{R_2 kT}{R_1 q} \ln(n) = |V_{BE_3}| + V_T \frac{R_2}{R_1} \ln(n)$$

Si ottiene così la forma classica della tensione di bandgap:

$$b = \frac{R_2}{R_1} \ln(n) \rightarrow V_{BG} = V_{BE} + bV_T$$

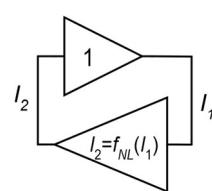
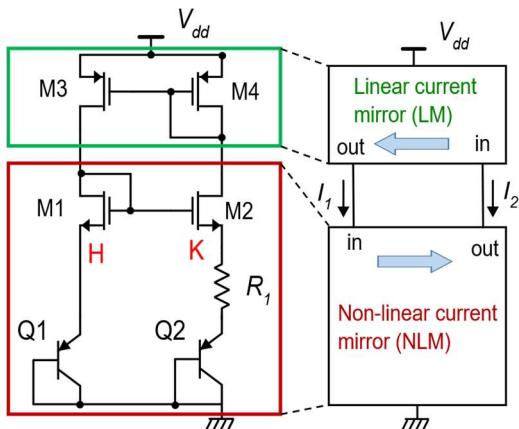
Cambiando il rapporto R_2/R_1 si influisce su b , che determina la T_0 del riferimento. Specchiando la corrente PTAT su una resistenza si ottiene una tensione PTAT che funziona da sensing di temperatura accurato.



L'accuratezza è data dal fatto che si riescono a fare rapporti di resistenze precise e anche il parametro n può essere reso accurato (il logaritmo ne addolcisce anche l'errore).

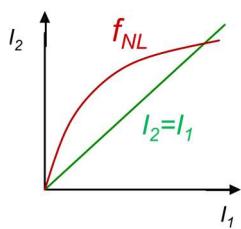
$$V_{temp} = R_3 I = \frac{R_3 kT}{R_1 q} \ln(n)$$

In realtà, il generatore di corrente PTAT presentato non funziona così com'è a causa della presenza di due stati stabili.

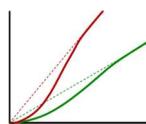


Il circuito presenta una retroazione positiva.
Lo specchio di sotto è uno specchio non lineare.

Un possibile stato del circuito è corrente nulla. L'altro stato stabile è la soluzione trovata prima.



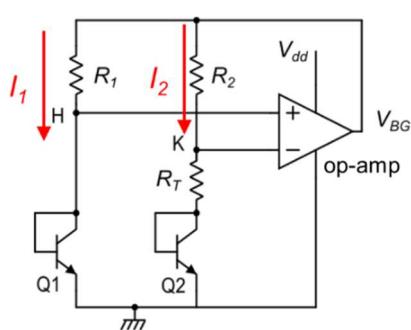
Una prima osservazione farebbe credere che la prima intersezione corrisponda in realtà ad una soluzione instabile. La derivata della curva non lineare è maggiore dell'unità, per cui una variazione di corrente verrebbe amplificata ad ogni ciclo fintanto che il circuito non raggiunge la soluzione stabile.



In realtà osservando vicino all'origine ci si accorge che i guadagni dei due specchi sono entrambi minori dell'unità, per cui anche la condizione di circuito spento è stabile. Pertanto, per far funzionare il generatore di corrente PTAT bisogna abbinare un circuito di startup che fornisca un po' di corrente per sbloccare il circuito dallo stato di off.

Altro circuito bandgap

Un circuito per la generazione di un riferimento di tensione bandgap ancora più utilizzato è il seguente:



Grazie al cortocircuito virtuale:

$$V_H = V_k \rightarrow V_{R1} = V_{R2}$$

Scegliendo $R_1 = R_2 \rightarrow I_1 = I_2$

$$V_H = V_k \rightarrow V_{BE1} = V_{BE2} + I_2 R_T$$

$$I_2 R_T = V_{BE1} - V_{BE2} = V_T \ln \left(\frac{I_{C1}}{I_{S1}} \frac{I_{S2}}{I_{C2}} \right) \cong V_T \ln \left(n \frac{I_1}{I_2} \right) = V_T \ln(n)$$

Quindi:

$$I_2 = I_1 = \frac{V_T \ln(n)}{R_T} \rightarrow V_{BG} = V_{BE1} + I_1 R_1 = V_{BE1} + \frac{V_T \ln(n)}{R_T} R_1$$