CURSO DE MICROELECTRÓNICA
PARTE II

TEMA 2B
LA FÍSICA DEL TRANSISTOR MOS
ANÁLISIS AMPLIADO

Dr. Yoanlys Hernández Barrios

Sección de Electrónica del Estado Sólido (SEES)

Departamento de Ingeniería Eléctrica (DIE)

CINVESTAV

EL TRANSISTOR MOS

ANÁLISIS AMPLIADO

TMOS Complementos del modelo

El modelo núcleo para canales largos, para que este completo, debe de complementarse al menos con tres efectos más:

1. La movilidad variable; μ_{var}
2. La resistencia en serie; R_S
3. La corriente subumbral; I_{sub} y la S.

Movilidad de los portadores en el canal de un TMOS

TMOS Movilidad de los portadores

Una de las magnitudes que ha sido de mayor objeto de estudio en los transistores MOS es la *movilidad de los portadores* en el canal, por la diversidad de factores que influyen en la misma.

Sobre esta magnitud física influyen:

1. la concentración de portadores y la temperatura;
2. la dispersión en el movimiento debido al campo longitudinal a lo largo del canal;
3. la dispersión de los portadores por el efecto del campo transversal en el canal que define los choques, o interacciones, de los portadores en la superficie (interfaz óxido-semiconductor);
4. las irregularidades de la superficie donde se crean centros de dispersión;
5. el efecto de la compresión de la red por las difusiones de ***D*** y ***S***, en general la presión.

TMOS Movilidad de los portadores

Una aproximación física bastante general para la movilidad en el canal, que se conoce como movilidad superficial, es la siguiente:

$$\mu_s = \frac{\mu_o}{\left(1+\frac{E_y}{E_{yc}}\right)\left(1+\frac{E_x}{E_{xc}}\right)}$$

donde E_{yc} - campo eléctrico crítico longitudinal;
E_{xc} - campo eléctrico crítico transversal;
μ_0 - movilidad superficial máxima.

Si los campos se incrementan, la movilidad disminuye

De las múltiples aproximaciones existentes escogeremos una semiempírica, de amplia utilización, donde las dependencias del campo eléctrico, que no pueden medirse externamente, se sustituyen por los voltajes externos aplicados.

TMOS Movilidad de los portadores

EFECTO DEL CAMPO TRANSVERSAL (DEPENDENCIA DE V_G)

$$\mu_s = \frac{\mu_0}{1 + \theta\,(V_G - V_T)}$$

El parámetro θ incluye todos los factores que afectan la caída de la movilidad con el campo transversal.

EFECTO DEL CAMPO LONGITUDINAL (DEPENDENCIA DE V_D)

En el caso del **campo longitudinal**, se modela a través de la velocidad de saturación de los portadores v_{max}. Esto es particularmente importante en los canales muy cortos y con campos intensos.

$$E_y = \frac{V_D}{L} \qquad E_{yc} = \frac{v_{max}}{\mu_s}$$

$$\mu_{eff} = \frac{\mu_s}{\left[1 + \left(\frac{\mu_s V_{ef}}{v_{sat} L}\right)^m\right]^{\frac{1}{m}}}$$

m = 1 for p-channel
m = 2 for n-channel

Resistencias en serie

TMOS Resistencias en serie

Entre el contacto de un electrodo externo, fuente y/o drenaje, y el canal del TMOS existen dos resistencias:

1. Resistencia de contacto metal-semiconductor;
2. Resistencia del material dentro de las regiones de fuente y/o drenaje.

La suma de ambas resistencias es lo que se conoce como ***resistencia en serie***, en la fuente R_S y en el drenaje R_D.

El voltaje aplicado al electrodo externo y al extremo del canal son diferentes y se determinan por las siguientes expresiones:

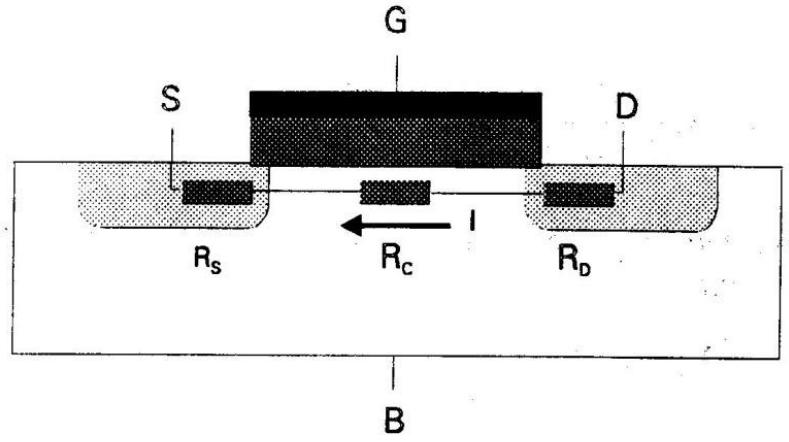


Fig. 3.2 Resistencias en un transistor MOS. R_s - resistencia de contacto y de la región del surtidor; R_D - resistencia del contacto y de la región del drenador.

$$V_{DS} = V_D - I \cdot (R_S + R_D)$$

$$y$$

$$V_{GS} = V_G - I \cdot R_S$$

TMOS Resistencias en serie

Usando la expresión aproximada en la región lineal:

$$K = \frac{W}{L} C_o \mu_{ef} \qquad R = R_S + R_D$$

$$I = \frac{K \cdot \left[(V_G - V_T)\, V_D - \left(\frac{1+n}{2} \right) V_D^2 \right]}{1 + K \cdot R \left[(V_G - V_T) - \left(\frac{1}{2} + n \right) V_D \right]}$$

Afectación por la R_S y R_D

NOTA: Se confunde con la reducción de la corriente debido a la movilidad,

pero no afecta la movilidad!!!

Corriente subumbral

TMOS Corriente subumbral

La corriente calculada hasta ahora es válida para $V_G \geq V_T$

Cuando $V_G < V_T$ $(u_s < 2u_f)$ la corriente se conoce como ***corriente subumbral (subthreshold current)***.

Condiciones:

1. La superficie cambió de conductividad, pero la concentración es menor que la del substrato, o sea, está en inversión débil primero y luego en empobrecimiento.

2. El campo eléctrico longitudinal es relativamente pequeño.

3. La concentración junto al ***S*** es de n_s y junto al ***D*** es cerca de 0 por eso la corriente viene dada por **la difusión de portadores** de **S** a **D**. ***Se calculará a continuación.***

TMOS Corriente subumbral

Densidad de carga

$$\begin{aligned}\rho(u,v) &= -qn_i e^{u_f}\left[(1-e^{-u})+e^{-2u_f}(e^{u-v}-1)\right]\\ &= -qN_a\left[(1-e^{-u})+e^{-2u_f}(e^{u-v}-1)\right]\\ &= -qN_a\left[1-e^{-u}+e^{u-2u_f-v}-e^{-2u_f}\right]\\ &\approx -qN_a\left[1-e^{-u}+e^{u-2u_f-v}\right]\end{aligned}$$

Resolviendo la ecuación de Poisson y considerando que

$$Es = \sqrt{\frac{2qNa\varphi t}{\varepsilon_s}}\sqrt{u_s + e^{u_s-2u_f-v}}$$

TMOS Corriente subumbral

$$Q_n = \varepsilon_s E_s - Q_B = \sqrt{2qN_a\varepsilon_s\varphi t u_s}\left[\sqrt{1+\frac{e^{u_s-2u_f-v}}{u_s}}-1\right]$$

Considerando que en subumbral $u_s<2u_f$ y desarrollando la raíz se obtiene:

$$Q_n = \varepsilon_s E_s - Q_B \approx \sqrt{\frac{qN_a\varepsilon_s\varphi t}{2u_s}}\cdot e^{u_s-2u_f}\cdot e^{-v}$$

TMOS Corriente subumbral

$$Q_n = \varepsilon_s E_s - Q_B \approx \sqrt{\frac{qN_a\varepsilon_s\varphi t}{2u_s}} \cdot e^{u_s - 2u_f} \cdot e^{-v} = \sqrt{\frac{qN_a\varepsilon_s}{2\phi_s}}\varphi t \cdot e^{\frac{\phi_s - 2\phi_f}{\varphi t}} \cdot e^{-\frac{V}{\varphi t}}$$

La corriente de difusión entre S y D se describe por:

$$I = -AD_n\frac{d(qn)}{dy} = -WD_n\frac{d(qnx_c)}{dy} = -WD_n\frac{Q_n^D - Q_n^S}{L-0} = \frac{Z}{L}\varphi t\mu_n(Q_n^S - Q_n^D)$$

$$si \quad D_n = \varphi t \cdot \mu_n; \quad V(0) = 0 \quad y \quad V(L) = V_D$$

$$I = \frac{W}{L}\mu_n\sqrt{\frac{qN_a\varepsilon_s}{2\phi_s}}\varphi t^2 \cdot e^{\frac{\phi_s - 2\phi_f}{\varphi t}}\left[1 - e^{-\frac{V_D}{\varphi t}}\right]$$

TMOS Corriente subumbral

$$\frac{d(\log(I))}{dV_G} = \frac{1}{2.3}\frac{d(\ln(I))}{dV_G}$$

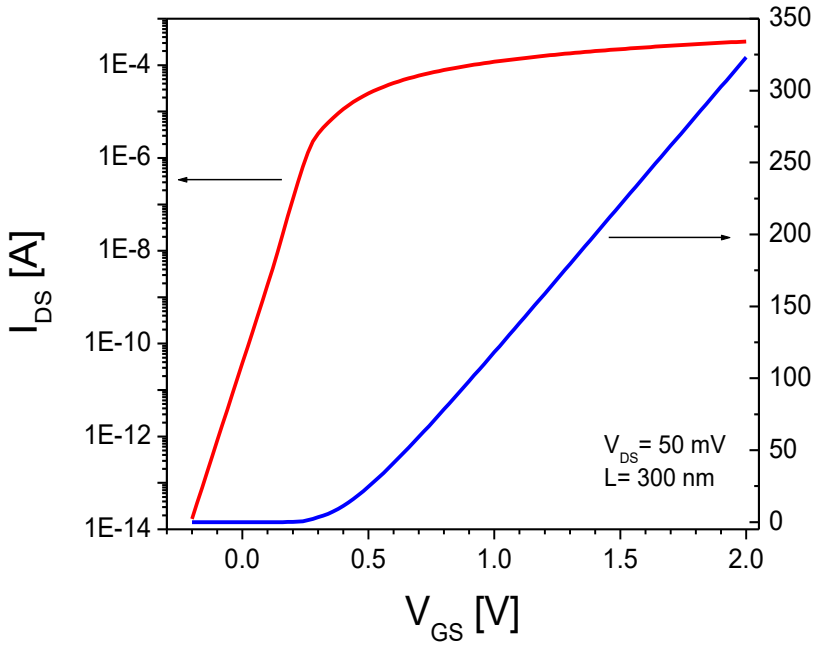
$$S = \frac{1}{\frac{d(\log(I))}{dV_G}} = 2.3\frac{1}{\frac{d(\ln(I))}{dV_G}}$$

$$V_G = V_{FB} + \phi_s + \gamma\sqrt{\phi_s}$$

$$\frac{dV_G}{d\phi_s} = 1 + \frac{\gamma}{2\sqrt{\phi_s}} = 1 + \frac{\sqrt{2qN_a\varepsilon_s}}{2C_o\sqrt{\phi_s}} = 1 + \frac{\varepsilon_s\sqrt{\frac{qN_a}{2\varepsilon_s\phi_s}}}{C_o} = 1 + \frac{\frac{\varepsilon_s}{x_d}}{C_o}$$

$$= 1 + \frac{C_s}{C_o}$$

$$\frac{d(\ln(I))}{dV_G} = \frac{d(\ln(I))}{d\phi_s}\frac{d\phi_s}{dV_G} = \frac{1}{\varphi t}\frac{1}{1 + \frac{C_s}{C_o}}$$

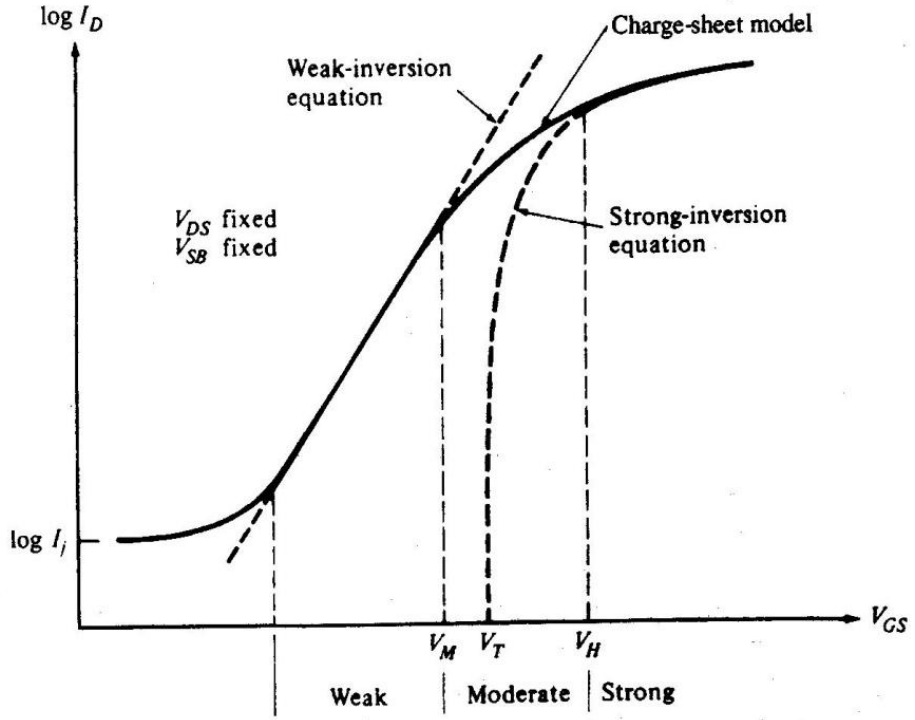


TMOS Corriente subumbral

$$S = 2.3\varphi t\left(1+\frac{C_s}{C_o}\right) \approx 60\left(1+\frac{C_s}{C_o}\right)\ \frac{mV}{decada}$$

$$S \approx 60\ \left(1+\frac{\varepsilon_s}{\varepsilon_o}\frac{x_o}{x_d}\right) \approx 60\ \left(1+3\frac{x_o}{x_d}\right)$$

$$x_d = \sqrt{\frac{2\varepsilon_s}{qN_a}\phi_s}$$



S nos da en la región exponencial la variación de V_G necesaria para que la corriente varíe un orden (10 veces).

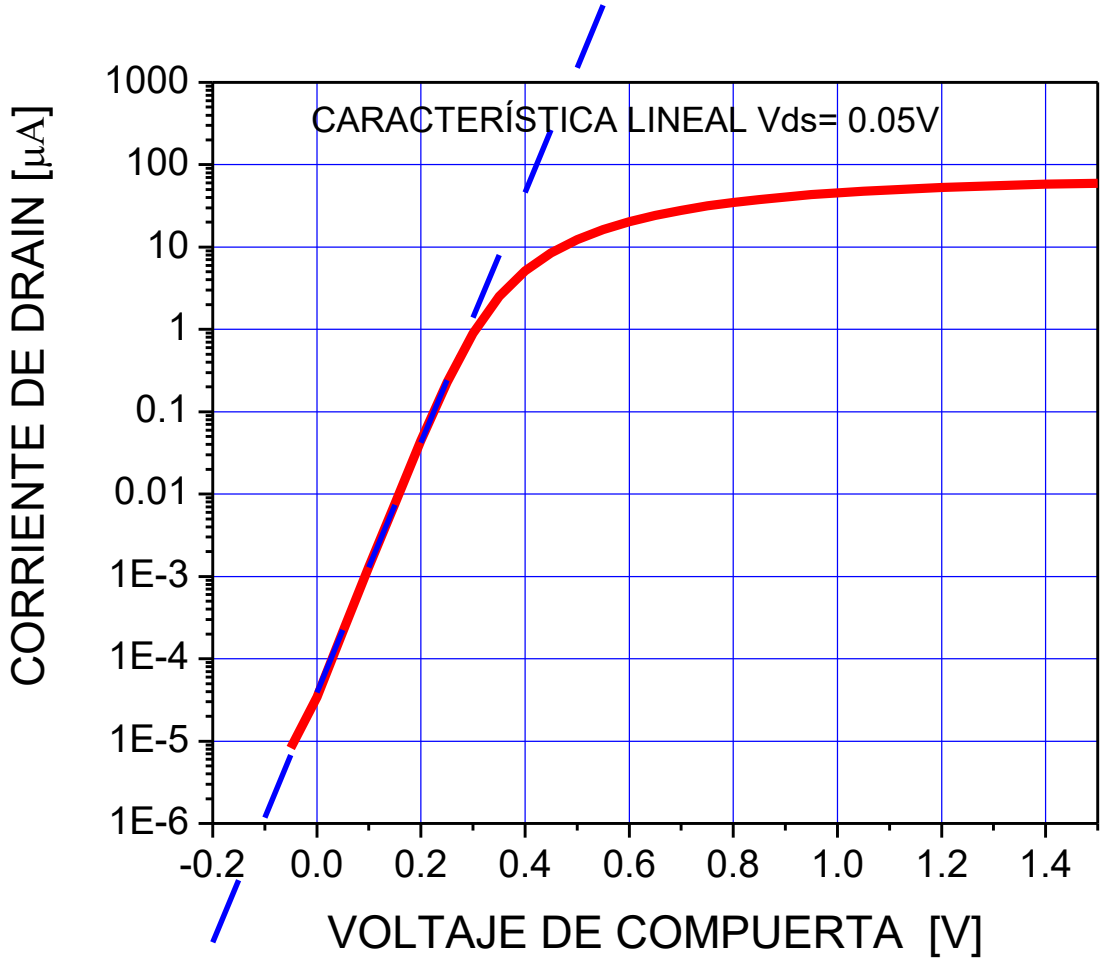
TMOS Corriente subumbral

Ejemplo medido:

Transistor FinFET de 50 nm de longitud de canal

S = 66 mV/dec

Para el Si a temperatura ambiente el límite inferior de S es de 60 mV/dec



TMOS Corriente subumbral

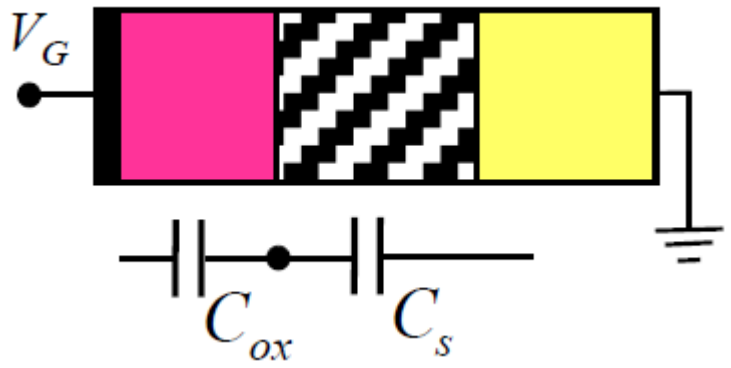
Si en la interfaz hay una densidad de cargas en trampas D_{it}, se tendrá una capacitancia asociada a las trampas que llamaremos C_{it} la cual estará en paralelo con la capacitancia C_s, o sea, ahora tenemos C_S+C_{it}

En este caso el parámetro ***S*** se vera alterado según la expresión:

$$S = 2.3\varphi t\left(1+\frac{C_s}{C_o}\right) \Rightarrow 2.3\phi t\left(1+\frac{C_s+C_{it}}{C_o}\right) = 2.3\phi t\left(1+\frac{C_s}{C_o}\right)\left(1+\frac{C_{it}}{C_o+C_s}\right)$$

El parámetro ***S*** se incrementa con la existencia de trampas cargadas en la interfaz

TMOS** Capacitancias del Capacitor MOS ideal



$$C_{tot} = \frac{dQ_G}{dV_G} = -\frac{dQ_s}{d(V_{ox} + \varphi_s)} = \frac{1}{-\frac{dV_{ox}}{dQ_s} - \frac{d\varphi_s}{dQ_s}}$$

$$= \frac{1}{1/C_{ox} + 1/C_s} = \frac{C_{ox}}{1 + C_{ox}/C_s}$$

En general, la carga en el semiconductor es representada como la suma de la densidad de carga en la capa de inversión (Q_N), la densidad de carga en la capa de empobrecimiento (Q_B), y la densidad de carga en la capa de acumulación (Q_P).

where:

- C_{ox} is the oxide capacitance
- C_s is the SC capacitance

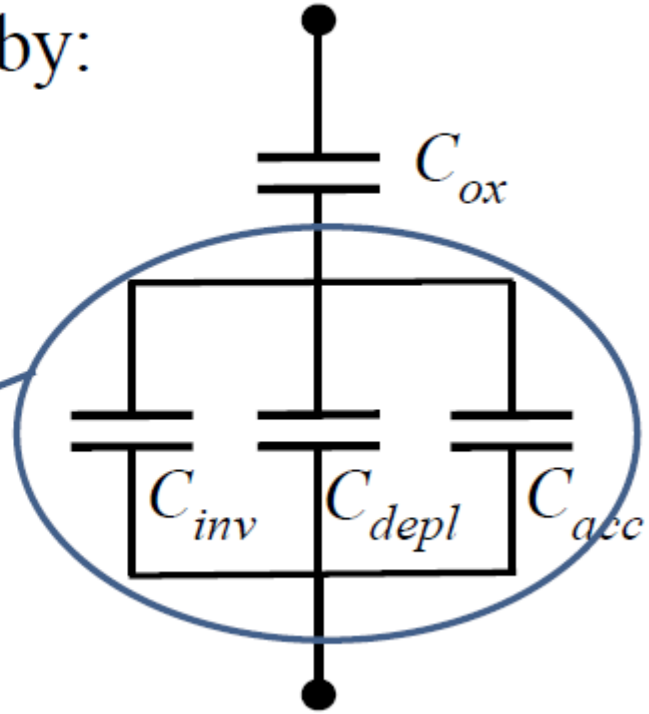
TMOS** Capacitancias del Capacitor MOS ideal

$$C_s = -\frac{dQ_s}{d\varphi_s} = -\frac{dQ_N}{d\varphi_s} - \frac{dQ_B}{d\varphi_s} - \frac{dQ_P}{d\varphi_s} = C_{inv} + C_{depl} + C_{acc}$$

The total gate capacitance is, thus, given by:

$$C_{tot} = \frac{C_{ox}}{1 + C_{ox}/C_s} = \frac{C_{ox}}{1 + \dfrac{C_{ox}}{C_{inv} + C_{depl} + C_{acc}}}$$

$$C_{ox} = \frac{k_{ox}\varepsilon_0}{d_{ox}}$$



TMOS** Capacitancias del Capacitor MOS ideal

$$C_s = -\frac{dQ_s}{d\varphi_s} = Cso\frac{\left|1 - e^{-\varphi_s/V_T} + \frac{n_{po}}{p_{po}}\left(e^{\varphi_s/V_T} - 1\right)\right|}{\sqrt{2}f(\varphi_s)}$$

$$f(\varphi_s) = \left[e^{-\varphi_s/V_T} + \frac{\varphi_s}{V_T} - 1 + \frac{n_{po}}{p_{po}}\left(e^{\varphi_s/V_T} - \frac{\varphi_s}{V_T} - 1\right)\right]^{1/2}$$

$$C_{so} = \frac{k_s \varepsilon_0}{L_D} \rightarrow \text{Flat-band capacitance}$$

Caso: Acumulación

$$\varphi_s < 0 \rightarrow f(\varphi_s) \propto \exp\left(-\varphi_s / 2V_T\right) \Big\} \rightarrow C_{tot} \approx C_{ox}$$
$$dQ_N = 0, \quad dQ_B = 0 \Big\}$$

La capacitancia total del gate es aproximadamente igual a Cox

TMOS** Capacitancias del Capacitor MOS ideal

Caso: Empobrecimiento

La carga de inversión es despreciable en comparación a la carga de empobrecimiento

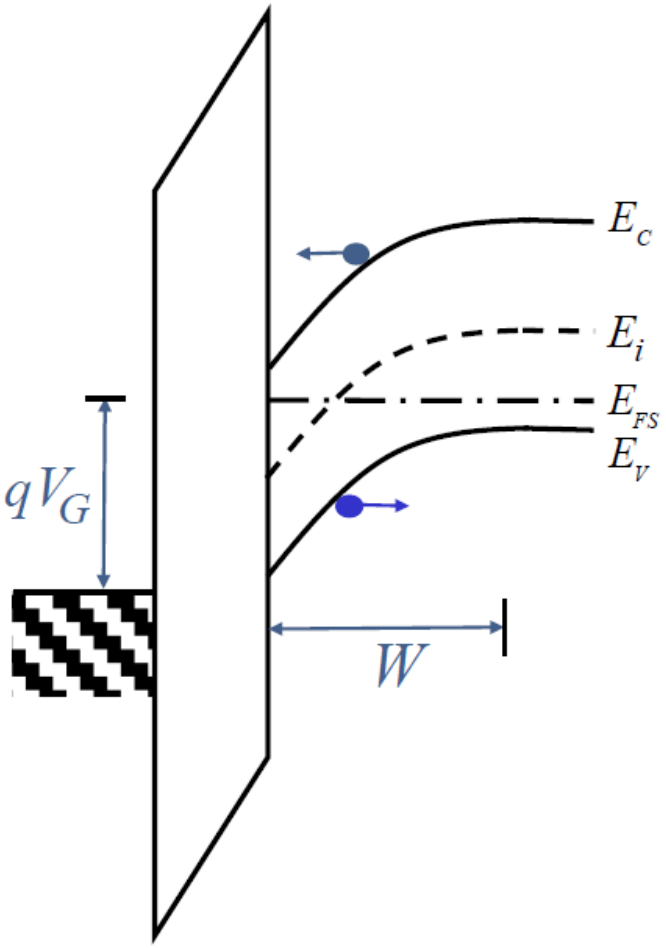
$$0 < \varphi_s < 2\varphi_F \rightarrow f(\varphi_s) \propto \sqrt{\varphi_s / V_T} \\ dQ_N = 0, \quad dQ_P = 0 \Big\} \rightarrow C_s = \frac{Cso}{\sqrt{2\varphi_s / V_T}} = \sqrt{\frac{k_s \varepsilon_0 q N_A}{2\varphi_s}}$$

$$C_{tot} = \frac{C_{ox}}{1 + \frac{C_{ox}}{C_s}} = \frac{C_{ox}}{1 + \frac{C_{ox}}{C_{depl}}} = \frac{k_{ox}\varepsilon_0}{d_{ox} + k_{ox}\varepsilon_0 \sqrt{\frac{2\varphi_s}{k_s \varepsilon_0 q N_A}}}$$

1. Si N_A aumenta, la capacitancia total aumenta.
2. Si el espesor de óxido aumenta, la capacitancia disminuye

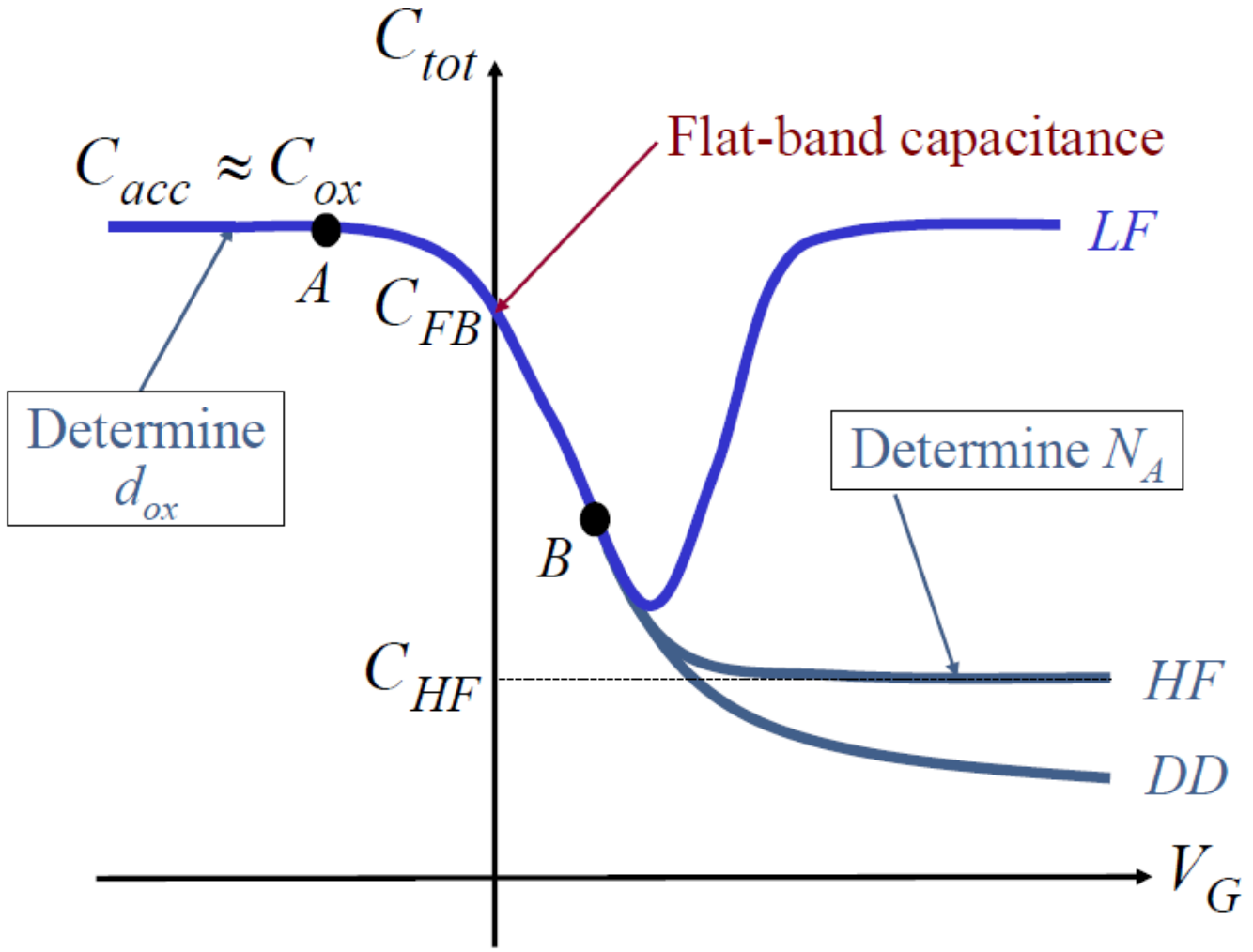
TMOS** Capacitancias del Capacitor MOS ideal

Caso: Inversión



1. La mayor parte de la carga inducida en la interfaz SC-óxido proviene de la generación de pares electrón-hueco (a través de centros de generación de recombinación).
2. La acumulación de portadores minoritarios se produce a un ritmo limitado por el proceso de generación de pares electrón-hueco.
3. Por lo tanto, dependiendo de la frecuencia de la señal aplicada y la velocidad de barrido del voltaje de compuerta, se puede medir:
   A. curva C-V a baja frecuencia (LF)
   B. curva C-V a alta frecuencia (HF)
   C. curva C-V en empobrecimiento profundo

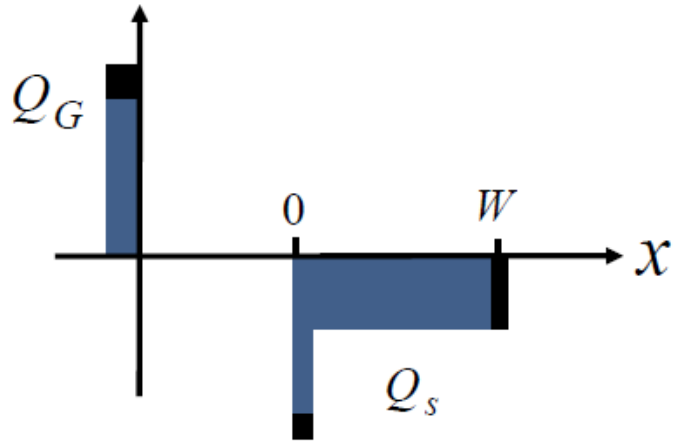
TMOS** Capacitancias del Capacitor MOS ideal



TMOS** Capacitancias del Capacitor MOS ideal

Caso: curva C-V a baja frecuencia y velocidad de barrido bajo

Permiten la generación de electrones en la capa de inversión y su respuesta a una señal AC aplicada



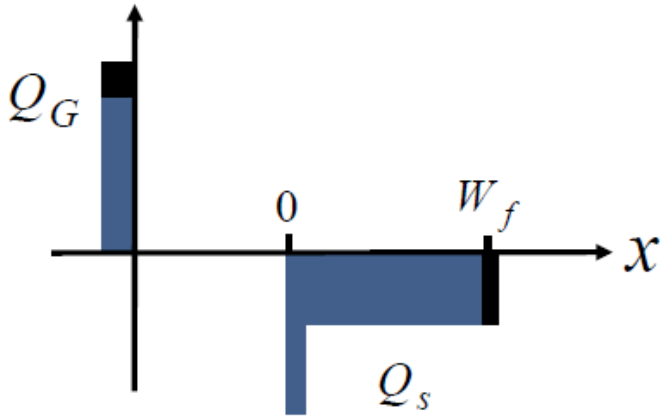
$$\left.\begin{array}{l}\varphi_s > 2\varphi_F \rightarrow f(\varphi_s) \propto \exp\left(\varphi_s / 2V_T\right) \\ dQ_P = 0\end{array}\right\} \rightarrow C_s \approx C_{inv} \approx C_{so}\sqrt{\frac{n_{po}}{2p_{po}}}e^{\varphi_s / 2V_T}$$

$$C_{tot} = \frac{C_{ox}}{1 + C_{ox}/C_s} = \frac{C_{ox}}{1 + C_{ox}/C_{inv}} \approx C_{ox}$$

TMOS** Capacitancias del Capacitor MOS ideal

Caso: curva C-V alta frecuencia (1) y velocidad de barrido bajo (2)

(1) Previene la respuesta de portadores minoritarios
(2) Permite la generación de electrones en la capa de inversión



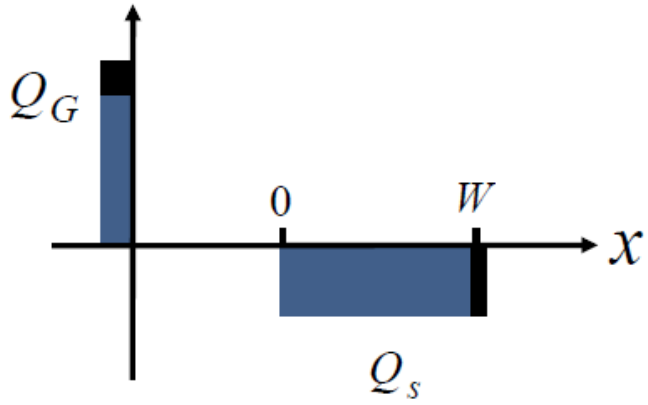
$$\left.\begin{array}{r} \varphi_s \approx 2\varphi_F \rightarrow f(\varphi_s) = \sqrt{2\varphi_F / V_T} \\ dQ_N = 0, \quad dQ_P = 0 \end{array}\right\} \rightarrow C_s \approx C_{depl} \approx \sqrt{\frac{k_s \varepsilon_0 q N_A}{2(2\varphi_F)}}$$

$$C_{tot} = \frac{C_{ox}}{1 + C_{ox}/C_{depl}} = \frac{C_{ox}}{1 + C_{ox}\sqrt{\dfrac{2(2\varphi_F)}{k_s \varepsilon_0 q N_A}}} \approx const$$

TMOS** Capacitancias del Capacitor MOS ideal

Caso: curva C-V alta frecuencia (1) y velocidad de barrido alta (2)

(1) Previene la respuesta de portadores minoritarios
(2) Previene la generación de electrones en la capa de inversión



$$\left.\begin{array}{r} f(\varphi_s) = \sqrt{\varphi_s / V_T} \\ dQ_N = 0, \quad dQ_P = 0 \end{array}\right\} \rightarrow C_s \approx C_{depl} \approx \sqrt{\frac{k_s \varepsilon_0 q N_A}{2\varphi_s}}$$

$$C_{tot} = \frac{C_{ox}}{1 + \frac{C_{ox}}{C_{depl}}} = \frac{C_{ox}}{1 + C_{ox}\sqrt{\frac{2\varphi_s}{k_s \varepsilon_0 q N_A}}}$$

TMOS** Capacitancias del Capacitor MOS ideal

Pero: ¿Qué es Baja Frecuencia?

En un MOS capacitor (o MOSFET):

- el gate se excita con una señal AC
- el óxido conduce solo corriente de desplazamiento
- el semiconductor debe suministrar portadores reales

👉 Para que exista inversión en baja frecuencia, los portadores minoritarios deben poder generarse a tiempo.

$$J_{SCR} = qn_i W / \tau_g$$

$$J_D = C_{ox} dV / dt$$

TMOS** Capacitancias del Capacitor MOS ideal

1. Corriente de generación en la región de carga espacial (SCR):

$$J_{SCR} = qn_i \frac{W}{\tau_g}$$

Significado físico:

J_{SCR} :corriente generada en la región de empobrecimiento

q: carga del electrón

n_i :concentración intrínseca

W: ancho de la región de empobrecimiento

τ_g :tiempo de generación

Esta corriente representa la velocidad con la que se pueden crear portadores minoritarios.

2. Corriente que circula por el óxido

El óxido no tiene portadores → solo puede circular:

$$J_D = C_{ox} \frac{dV}{dt}$$

Esto es:

- corriente de desplazamiento
- causada por la variación del voltaje del gate

Cuanto más rápido cambie el voltaje:

- mayor corriente exige el óxido

TMOS** Capacitancias del Capacitor MOS ideal

3. Condición clave de baja frecuencia

Para que la carga de inversión pueda seguir la señal AC, debe cumplirse:

corriente disponible ≥ corriente requerida

Es decir:

$$C_{ox}\frac{dV}{dt} \leq qn_i\frac{W}{\tau_g}$$

Interpretación física

- el óxido "pide" corriente cuando el voltaje cambia
- el semiconductor debe generar portadores minoritarios
- si no puede hacerlo suficientemente rápido → no hay inversión AC

Límite sobre la rapidez del voltaje

De ahí se obtiene:

$$\frac{dV}{dt} \leq \frac{qn_iW}{C_{ox}\tau_g}$$

Esto impone un límite de frecuencia.

Por eso "baja frecuencia" no significa un valor arbitrario, sino:

una frecuencia lo suficientemente baja como para que la generación de portadores pueda seguir la señal.

TMOS** Capacitancias del Capacitor MOS ideal

Example: d_{ox}=100 nm, W=1 μm, C_{ox}=3.45×10^{-8} F/cm^2 :

τ_g=10 μs, $dV/dt \leq 0.65$ V/s, f_{eff}=45 Hz (not a severe constraint)

τ_g=1 ms, $dV/dt \leq 6.5$ mV/s, f_{eff}=0.4 Hz (severe constraint)

Caso 1: generación rápida

$$\tau_g = 10\ \mu s$$

Resultados:

- $dV/dt \leq 0.65$ V/s
- $f_{eff} \approx 45$ Hz

👉 No es una restricción fuerte
👉 Inversión puede seguir la señal

Caso 2: generación lenta

$$\tau_g = 1\ \text{ms}$$

Resultados:

- $dV/dt \leq 6.5$ mV/s
- $f_{eff} \approx 0.4$ Hz

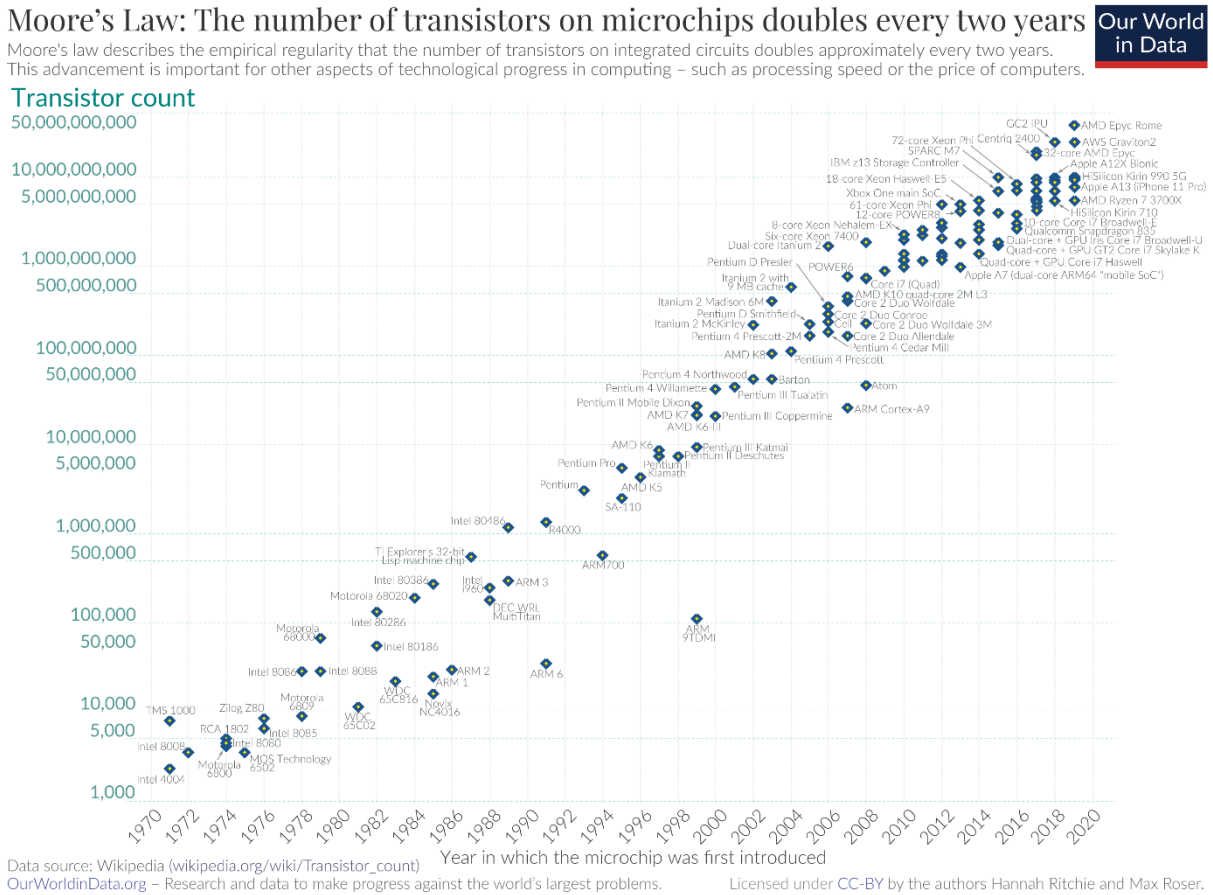
👉 Restricción severa
👉 Solo frecuencias muy bajas permiten inversión

ESCALADO

Reducción de las dimensiones del TMOS y su efecto en las características del transistor

Ley de Moore

El número de transistores en un chip se duplica cada dos años



PRACTICAMENTE SE HA CUMPLIDO DURANTE 60 AÑOS

TMOS Escalado

La reducción constante de las dimensiones de los transistores en los circuitos integrados se ha mantenido en los pasados 60 años llegando a las fantásticas cifras de decenas de miles de millones en un solo “chip”. Este proceso se conoce como ***escalado***.

El parámetro que se usa como referencia es la anchura de una línea en el circuito integrado y la distancia entre dos líneas. Actualmente las tecnologías de fabricación, que se basan en estas anchuras mínimas, se definen como **nodos** de tantos nm.

Anteriormente el concepto de nodo correspondía a la longitud del canal de los transistores ***L***, pero después del nodo de 22 nm el concepto de nodo varío. Actualmente la ***L*** es mayor que la dimensión de los nodos.

TMOS Escalado

Esta reducción de dimensiones se realizaba tratando que:

1. Las características de los transistores fueran lo más cercano posible a la de los TMOS de canal largo o mejores;
2. Se mantuviera el voltaje de alimentación constante, o el campo eléctrico constante, o un factor intermedio.

Este procedimiento tuvo validez hasta que las longitudes de canal comenzaron a bajar de 1 μm.

Para los llamados ***transistores submicrométricos*** se requiere ajustar cada parámetro en forma independiente. Mas aun para los ***nanométricos***.

TMOS Escalado. Versión inicial hasta los 90.

Table 2 MOSFET Scaling

Parameter	Scaling factor: Constant-$\mathscr{E}$	Scaling factor: Actual	Limitation
L	$1/\kappa$	/	/
$\mathscr{E}$	1	> 1	/
d	$1/\kappa$	$> 1/\kappa$	Tunneling, defects
r_j	$1/\kappa$	$> 1/\kappa$	Resistance
V_T	$1/\kappa$	$\gg 1/\kappa$	Off current
V_D	$1/\kappa$	$\gg 1/\kappa$	System, V_T
N_A	κ	$< \kappa$	Junction breakdown

In ideal constant-field scaling parameters are scaled by the same factor. In reality the scaling factors are limited by other reasons and skewed.

Con el escalado del MOSFET:

1. Diseño adecuado para preservar el comportamiento de canal largo lo más que se pueda.
2. Si disminuye L, el x_d(ZCE) del S y el D, es comparable con L, el "pinch-off" entre S y D ocurrirá eventualmente.
3. Esto requiere un canal más dopado.
4. Un canal más dopado incrementa el VT, y para controlar el VT en un rango razonable, se necesitará un espesor de óxido mas fino.

Se puede notar, que los parámetros del dispositivo están relacionados y se deben usar ciertas reglas de escalado para optimizar el desempeño del dispositivo.

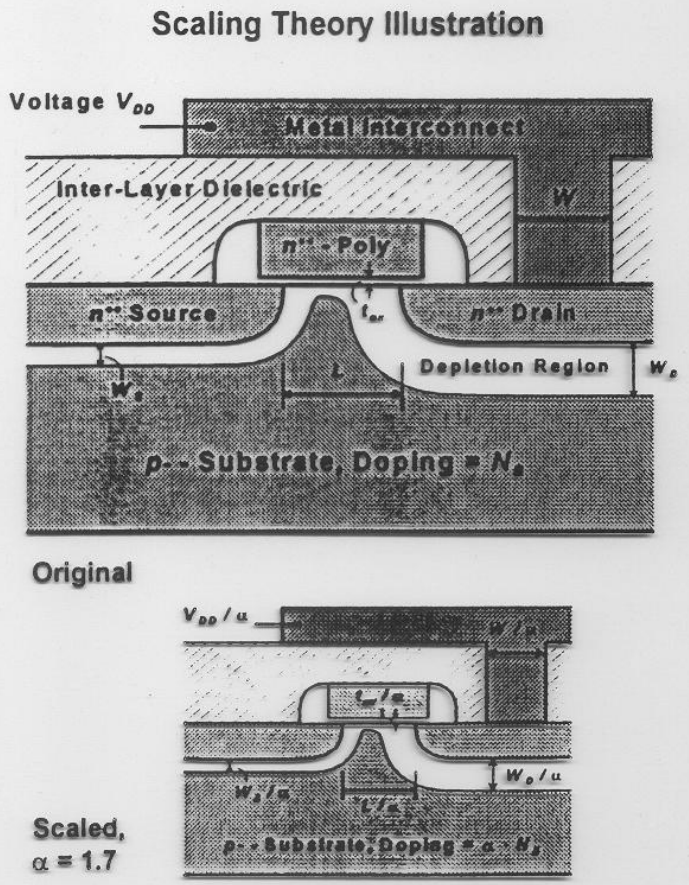
TMOS Escalado. Versión inicial hasta los 90.

Scaling Theory Revisited

- **Essence of scaling theory:**

  If everything is scaled by the rules, it is possible to build well-behaved (i.e., "long-channel") MOSFETs at any level of miniaturization

- **Areas of scaling rule violations:**
  - Gate oxide
  - Surface mobility
  - Source/drain junctions
  - Threshold voltage
  - Power supply voltage



TMOS Escalado -> Efectos de canal corto

El escaldo se realizó hasta los 90 con un factor de variación fijo, o factor ***α*** o ***λ***, con el cual se modificaban las distintas magnitudes del transistor.

Cuando la ***L*** bajo por debajo de los 100 nm se requirió que para cada nueva longitud de canal se realizara un ajuste para cada magnitud y parámetro, surgiendo el concepto de nodo, de 35 nm, 28 nm, 22 nm, 15 nm, 10 nm, 9 nm, 8 nm, 7 nm, 5 nm y 3 nm.

Cuando se reduce la longitud de canal ***L*** de un TMOS se producen algunas variaciones en sus principales parámetros que afectan a las corrientes, lo que se conocen como
Efectos de Canal Corto (*Short Channel Efects - SCE*)

TMOS Escalado -> Efectos de canal corto

Aún con las mejores reglas de escalado, según se reduce L, es inevitable que el comportamiento ya no sea de canal largo. Aparecen los Efectos de Canal Corto como resultado de una distribución de potencial en dos dimensiones y altos campos eléctricos en la región del canal. En otras palabras, la aproximación de canal gradual ($\varepsilon_x >> \varepsilon_y$) ya no es más válida. Este distribución de potencial en dos dimensiones resulta en muchas formas de comportamientos eléctricos indeseables.

1. **Más campo eléctrico, la movilidad en el canal se vuelve más dependiente del campo, y eventualmente ocurre la velocidad de saturación.**
2. **Aún más campo, los portadores se multiplican junto al D, lo que resulta en corriente a través del sustrato y acción parásita de transistor-bipolar.**
3. **Altos campos también causan inyección de portadores calientes dentro del óxido resultando en óxido cargado y desplazamiento del VT y degradación de la transconductancia.**

TMOS Efectos de canal corto

¿CUÁLES SON LOS EFECTOS DE CANAL CORTO (SCE)?

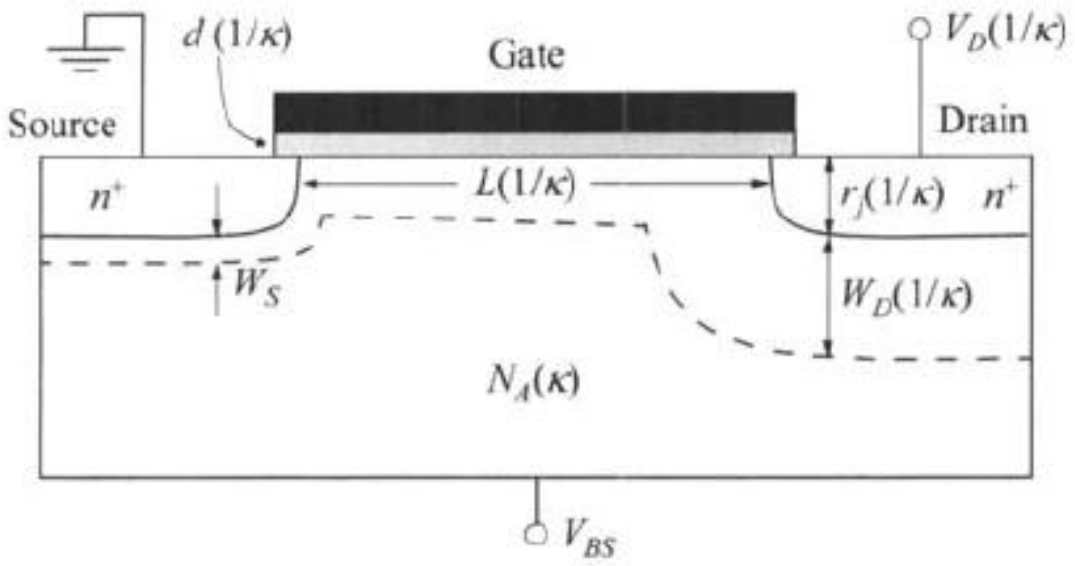
Se consideran los siguientes efectos:

1. Saturación de la velocidad de los portadores en el canal.
2. Variación del voltaje umbral.
3. DIBL – Reducción del potencial $\mathbf{V_S}$, en función del $\mathbf{V_D}$.
4. Incremento de **S**
5. Portadores calientes.
6. Rupturas: avalancha y “punch through”.

En otras palabras: (1) V_T no es constante con L; (2) I_{DS} no satura con V_D; (3) I_{DS} no es proporcional a 1/L; (4) las características del dispositivo se degradan con el tiempo.

TMOS Efectos de canal corto

La regla de escalado más ideal para evitar los efectos de canal corto es simplemente escalar todas las dimensiones y voltajes de un MOS de canal largo para que los campos eléctricos internos se mantengan igual.



1. L; W; D; rj: se reducen por el mismo factor k.
2. El nivel de dopaje se aumenta en k
3. Los voltajes son reducidos en 1/k, conduciendo a una reducción del ancho de empobrecimiento en 1/k.
4. La S permanece igual ya que es proporcional a la relación entre las capacitancias C_{ox} y C_D que son escaladas por el mismo factor.

Desafortunadamente, esta regla de escalado ideal se ve obstaculizada por otros factores que fundamentalmente no son escalables.

TMOS Efectos de canal corto

1. En primer lugar, el voltaje incorporado en la unión y el potencial de superficie para el inicio de la inversión débil no escalan (solo un cambio del 10 % para un aumento de 10 veces en los dopajes).
2. El rango de voltaje de compuerta entre el agotamiento y la inversión fuerte es de aproximadamente 0,5 V. Estas limitaciones se deben al hecho de que tanto la brecha energética como la energía térmica kT permanecen constantes.
3. El espesor del óxido de compuerta presenta la dificultad tecnológica de presentar defectos a medida que se acerca a la escala de nm bajos. La tunelización a través del óxido es otra limitación fundamental.
4. La resistencia en serie de fuente y drenaje aumenta cuando rj disminuye.
5. El dopaje del canal no se puede aumentar indefinidamente debido a la ruptura de la unión p-n.
6. El voltaje de umbral no se puede escalar debido a la consideración de la corriente I_{off}, incluso con una S fija.

Con estas limitaciones, el campo ya no se mantiene igual, y aumenta con longitudes de puerta más pequeñas.

TMOS Efectos de canal corto

Los MOSFET construidos sobre una estructura tridimensional con un cuerpo ultrafino eliminan eficazmente la mayor parte de la ruta de conducción para la perforación, lo que permite flexibilizar el requisito de dopaje del canal.

En segundo lugar, se ha realizado una intensa investigación en la búsqueda de dieléctricos de puerta con constantes dieléctricas altas. Un dieléctrico de puerta de alta k de este tipo puede flexibilizar el espesor físico, mejorando la densidad de defectos y reduciendo el campo para la tunelización.

Ambas tecnologías pueden ayudar a evitar o retrasar los efectos de canal corto para una generación específica de longitud de canal.

TMOS Velocidad de saturación de la corriente

Energía cinética de los portadores

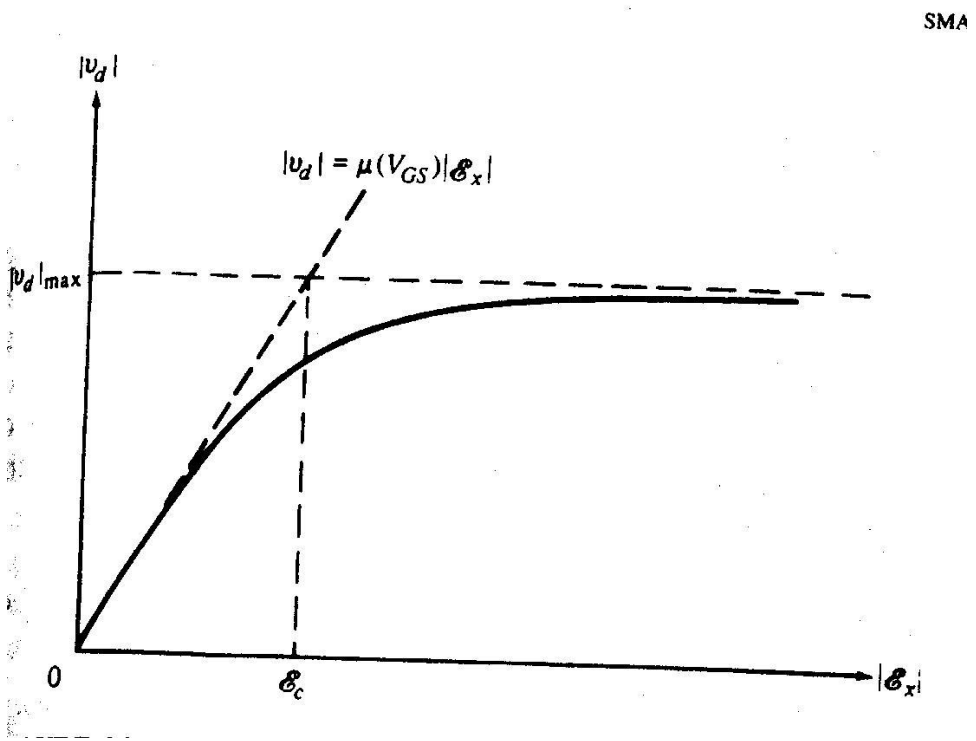
$$\frac{1}{2}m_n^* v_t^2 = \frac{3}{2}kT$$

$$v_d = -\frac{q\tau_{ch}E}{m_n^*} = -\mu_n E_y$$

$$v_d = \frac{|\mu E_y|}{\left(1+\left(\frac{E_y}{E_c}\right)^{\alpha}\right)^{\frac{1}{\alpha}}}$$

α=2 para electrones y

1 para huecos



$$v_{d\max} = v_{sat} = \mu E_c; \quad E_c = \frac{V_D}{L}$$

TMOS Velocidad de saturación de la corriente

Cuando incrementando el voltaje de drenaje se alcanza la velocidad de saturación la corriente no se incrementa, se satura.

Ese voltaje, que se conoce como voltaje de saturación V_{Dsatv} es menor que el voltaje de saturación de "pinch-off" que habíamos calculado para el caso de canal largo.

Para la primera aproximación

$$I_D = \frac{Z}{L} C_o \frac{\mu}{1 + \left|\frac{\mu V_D}{L v_{sat}}\right|} \left[(V_G - V_T)V_D - \frac{V_D^2}{2}\right]$$

$$I_D dy = -Z\mu Q_n dV_D$$

$$E_y = -\frac{dV}{dy} \qquad 1 + \left|\frac{V_D}{LE_c}\right| \Rightarrow 1 + \left|\frac{\mu V_D}{L v_{sat}}\right|$$

$$I_D = \mu Z Q_n E_y = Z Q_n \frac{\mu E_y}{1 + \frac{E_y}{E_c}}$$

$$= -\frac{Z\mu}{1 + \left|\frac{\mu V_D}{v_{dsat} L}\right|} \left(-\frac{dV}{dy}\right) Q_n$$

$$I_D \int_0^L dy = -\frac{Z\mu}{1 + \left|\frac{\mu V_D}{v_{dsat} L}\right|} \int_0^{V_D} Q_n dV$$

$$I_D = -\frac{Z}{L} \frac{\mu}{1 + \left|\frac{\mu V_D}{v_{dsat} L}\right|} \int_0^{V_D} Q_n dV$$

TMOS Corriente de saturación

En este caso el voltaje de saturación V_{Dsatv} se puede calcular igualando la derivada de la corriente a cero.

$$\frac{dI_D}{dV} = \frac{\mu Z Q_n}{L\left(1+\left|\frac{\mu V_D}{Lv_{sat}}\right|\right)} + \frac{Z\mu\int_0^{V_D} Q_n dV}{L^2 E_c\left(1+\left|\frac{\mu V_D}{Lv_{sat}}\right|\right)^2} = 0 \qquad LE_c = \frac{Lv_{sat}}{\mu}$$

$$V_{Dsatv}' = |LE_c|\left[\sqrt{1+\frac{2(V_G-V_T)}{|LE_c|}}-1\right] = \left|\frac{Lv_{sat}}{\mu}\right|\left[\sqrt{1+\frac{2(V_G-V_T)}{\left|\frac{Lv_{sat}}{\mu}\right|}}-1\right]$$

En este caso cuando se alcanza la v_{sat}, resulta que hay una carga junto al D diferente de cero:

$Q_n(V_D) > 0$ $y=L$ $V=V_D$

Caso de "pich-off" ← $V_{Dsat} = V_G - V_T > V_{Dsatv}$

TMOS Corriente de saturación

Efecto directo sobre la corriente

En canal corto:

$I_D \approx W\, C_{ox}(V_{GS} - V_T)\, v_{sat}$

→la corriente pasa de ser cuadrática a casi lineal con V_{GS}.

Consecuencia:

- la corriente ya no crece como se esperaba
- el transistor entrega menos corriente de la predicha por el modelo de canal largo

- Saturación temprana de corriente
- El MOSFET entra en saturación para valores menores de V_{DS}:
- $V_{DSsat} \approx \frac{L\, v_{sat}}{\mu}$
- →cuanto más corto el canal, antes se satura la corriente.

Menor escalabilidad del rendimiento

Aunque se reduzca L:

- la corriente no aumenta proporcionalmente
- el retardo no mejora tanto como se espera

Esto limita la mejora de velocidad en tecnologías profundas (nanométricas).

TMOS Corriente de saturación

Reducción de la ganancia

- menor transconductancia g_m
- menor ganancia analógica

Degradación del control del gate

Como el canal se vuelve muy corto:

- el drenaje influye más sobre el canal
- el gate pierde control electrostático

Esto favorece:

- DIBL (Drain Induced Barrier Lowering)
- reducción del voltaje umbral

TMOS Voltaje de umbral. Variación.

Variación del voltaje umbral con la reducción de la longitud y la anchura del canal

Modelo trapezoidal de Poon y Yau

$$BC = x_{dg} \quad AO = x_j + x_{dj}$$
$$DB = AC \quad x_{dj} \approx x_{dg}$$

$$(x_j + x_{dj})^2 = x_{dj}^2 + (AC + x_j)^2$$

$$AC = x_j \left[\sqrt{1 + \frac{2x_{dj}}{x_j}} - 1 \right]$$

Área rayada $$x_{dj} \cdot x_j \left[\sqrt{1 + \frac{2x_{dj}}{x_j}} - 1 \right]$$

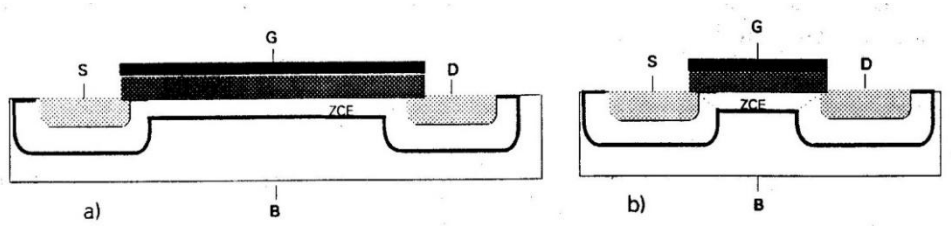


Fig. 3.6 Geometría trapezoidal del canal bajo la compuerta: a) canal largo; b) canal corto.

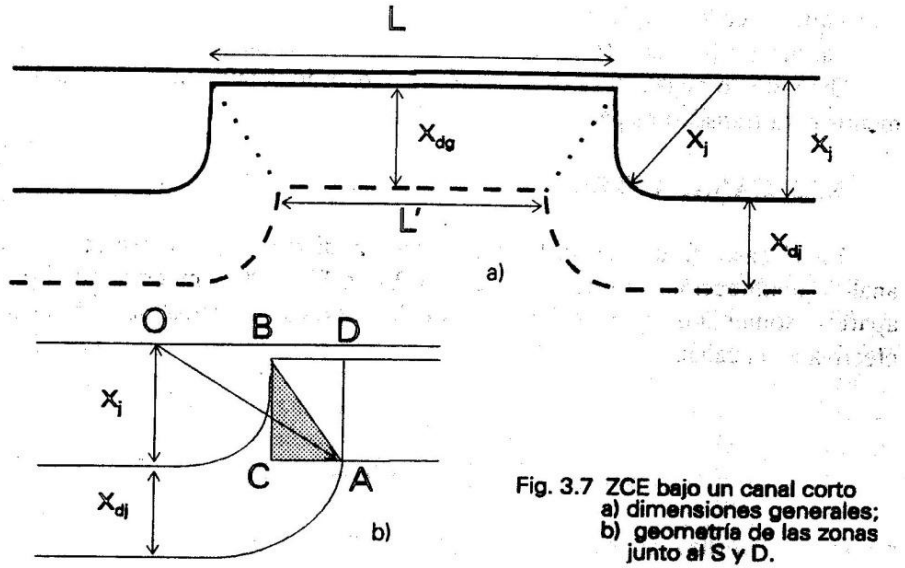


Fig. 3.7 ZCE bajo un canal corto
a) dimensiones generales;
b) geometría de las zonas junto al S y D.

TMOS Variación del voltaje umbral

Carga Q_B resultante:

$$Q_B = qN_aZLx_{dg}$$

$$Q_B^{'} = Q_B - qN_aZ\left[x_{dg}x_j\left(\sqrt{1+\frac{2x_{dj}}{x_j}}-1\right)\right] = Q_B\left[1-\frac{x_j}{L}\left(\sqrt{1+\frac{2x_{dj}}{x_j}}-1\right)\right]$$

$$V_T^{'} = V_{FB} + 2\phi_f + \frac{Q_B^{'}}{C_o}$$

Reducción de V_T depende de:

L ***x_j*** ***x_d*** ***Z=W***

$$\frac{\Delta V_T}{V_T} = \frac{V_T - V_T^{'}}{V_T} = 1 - \frac{Q_B^{'}}{Q_B} = \frac{x_j}{L}\left(\sqrt{1+\frac{2x_{dj}}{x_j}}-1\right)$$

TMOS Variación del voltaje umbral

Variación del V_T cuando se reduce la anchura del canal W.

$$Q_B = WLqN_a x_{dg}$$

$$Q_B^{''} = qN_a WLx_{dg}\left(1 + \frac{\pi x_{dg}}{2W}\right)$$

$$Q_B^{''} = Q_B + \left(2\frac{\pi x_{dg}^2}{4}\right) qN_a L$$

$$\frac{\Delta V_T}{V_T} = +\frac{\pi}{2}\frac{x_{dg}}{W}$$

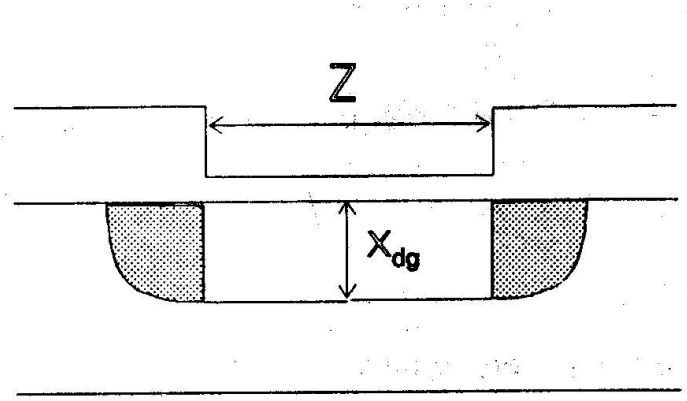


Fig. 3.8 Anchura del canal de un TMOS mostrando la ZCE que se prolonga mas allá de los bordes.

Z=W

TMOS DIBL – Drain Induced Barrier Lowering

- Hemos señalado que cuando las regiones de agotamiento de la fuente y el drenador representan una fracción sustancial de la longitud del canal, comienzan a producirse efectos de canal corto. En casos extremos, cuando la suma de estos anchos de agotamiento se aproxima a la longitud del canal (Wsm + Wdm = L), se producen efectos más graves. Esta condición se denomina comúnmente "punch-through". El resultado neto es una gran corriente de fuga entre la fuente y el drenador, y esta corriente es una función importante de la polarización del drenador.
- El origen de la penetración es la disminución de la barrera cerca de la fuente, comúnmente conocida como DIBL (disminución de la barrera inducida por el drenaje). Cuando el drenaje está cerca de la fuente, la polarización del drenaje puede influir en la barrera en el extremo de la fuente, de modo que la concentración de portadores del canal en esa ubicación deja de ser fija.

TMOS DIBL – Drain Induced Barrier Lowering

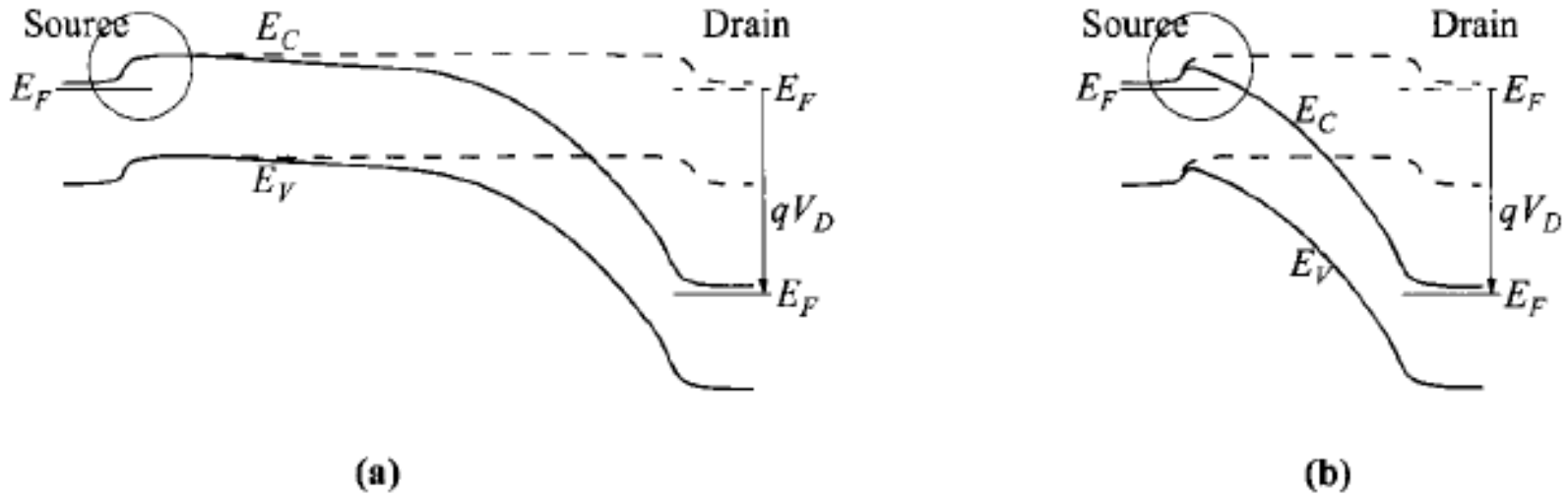


Fig. 28 Energy-band diagram at the semiconductor surface from source to drain, for (a) long-channel and (b) short-channel MOSFETs, showing the DIBL effect in the latter. Dashed lines $V_D = 0$. Solid lines $V_D > 0$.

En un dispositivo de canal largo, una polarización de drenador puede modificar la longitud efectiva del canal, pero la barrera en el extremo de la fuente permanece constante. En un dispositivo de canal corto, esta misma barrera deja de ser fija. La reducción de la barrera de la fuente provoca la inyección de portadores adicionales, lo que aumenta considerablemente la corriente. Este aumento de corriente se observa tanto en regímenes por encima del umbral como por debajo del umbral.

TMOS DIBL – Drain Induced Barrier Lowering

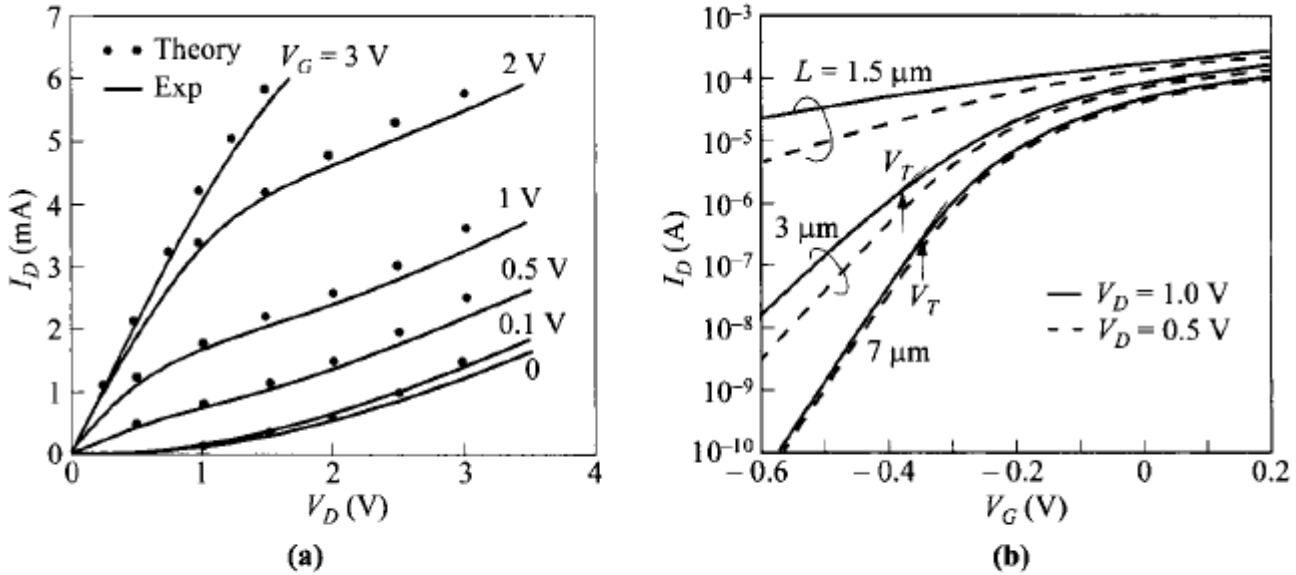


Fig. 29 Drain characteristics of MOSFETs showing DIBL effect. (a) Above threshold. L = 0.23 μm. d = 25.8 nm. $N_A = 7\times10^{16}$ cm^{-3}. (b) Below threshold. d = 13 nm. $N_A = 10^{14}$ cm^{-3}. (After Ref. 50.)

$$I_{DS} \approx \frac{9\varepsilon_s \mu_n A {V_D}^2}{8L^3}$$

VOLTAJES DE RUPTURA DEL TRANSISTOR MOS

TMOS Voltaje de ruptura

Voltaje de ruptura en la característica de salida.

Se entiende por voltaje de ruptura cuando hay un **incremento brusco** de la corriente inversa.

Si los portadores tienen una energía superior a la energía cinética media de la red se dice que son portadores calientes.

Si la energía cinética es muy alta se pueden producir choques junto al drenador ***D*,** que generen pares electrón-hueco.

Si los pares generados a su vez generan nuevos pares, se tiene un efecto multiplicador que puede dar lugar a un efecto de avalancha, o sea, incremento brusco de los portadores, y por lo tanto de la corriente.

TMOS Voltaje de ruptura

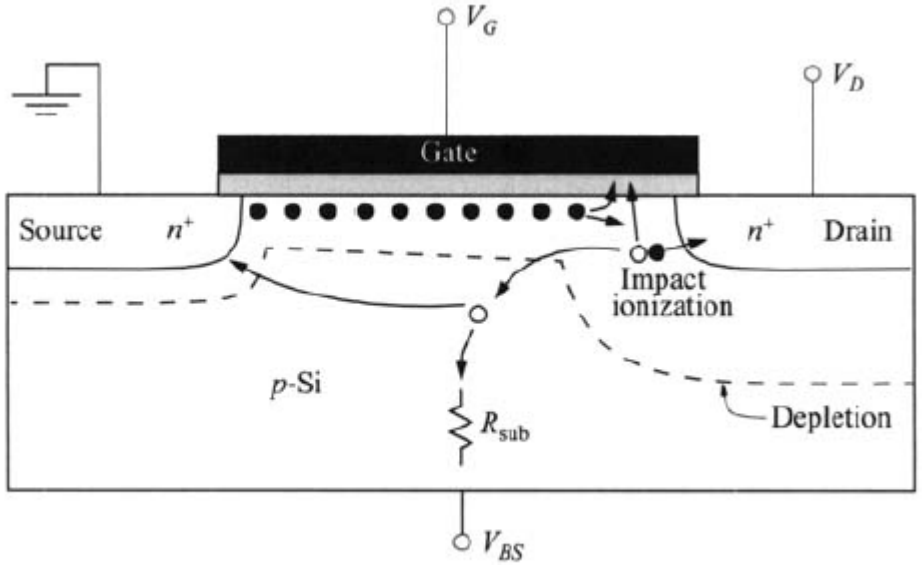


Fig. 30 Current components of a MOSFET under high fields.

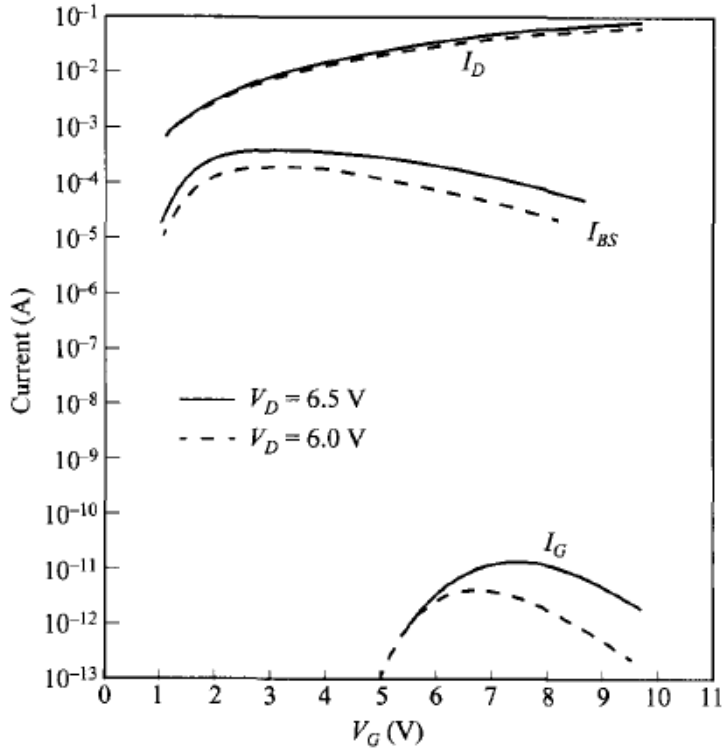


Fig. 31 Drain current, substrate current, and gate current vs. gate voltage of a MOSFET. L/W = 0.8/30 μm. (After Ref. 51.)

TMOS Corriente de ruptura por avalancha

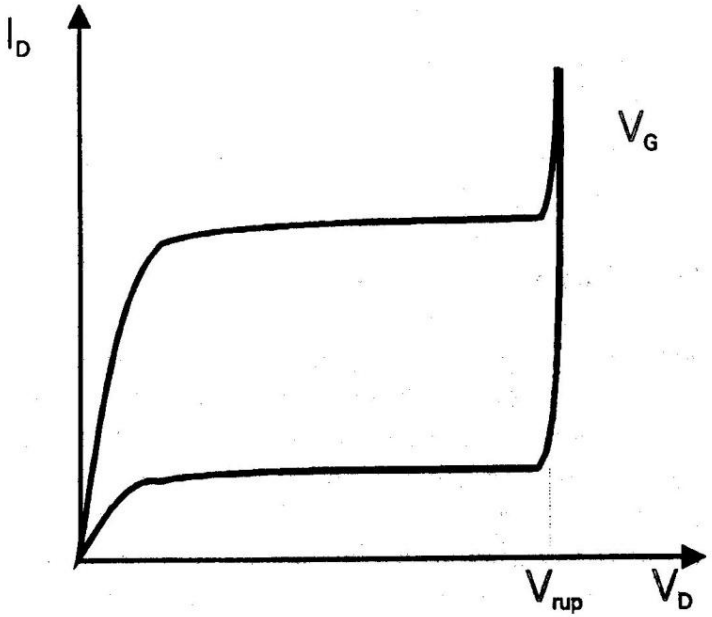


Fig. 3.9 Zona de ruptura en la característica I-V del TMOS

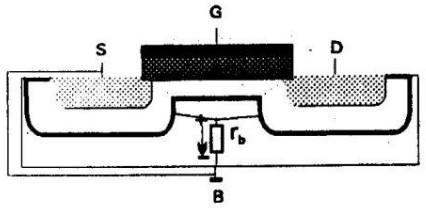


Fig. 3.10 Polarización de la juntura de la fuente por efecto de la corriente de substrato.

Avalancha –

Si se produce una avalancha, la corriente de drenaje se incrementa bruscamente y aparece la ruptura. Este efecto puede ser reversible si se controla la corriente.

Retroalimentación: Como se ve en la figura, la generación de pares conlleva una corriente por el substrato que produce una retroalimentación positiva al reducir la caída de voltaje en la unión P-N del ***S***.

TMOS Corriente de ruptura por avalancha

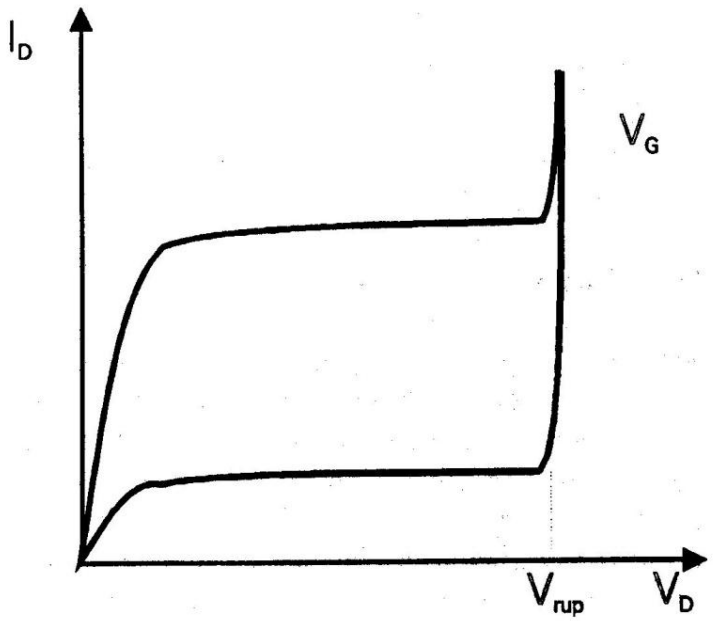


Fig. 3.9 Zona de ruptura en la característica I-V del TMOS

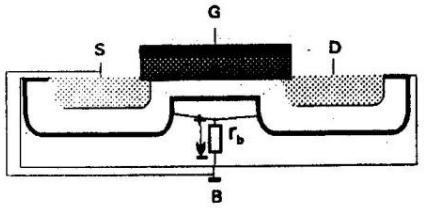


Fig. 3.10 Polarización de la juntura de la fuente por efecto de la corriente de substrato.

"Punch-through" –

Si ambas **ZCE** del ***S*** y del ***D*** se tocan desaparecen las uniones P-N y se produce otro tipo de ruptura que también pudiera ser reversible y/o controlado.

TMOS SOI y Thin-Film Transistor (TFT)

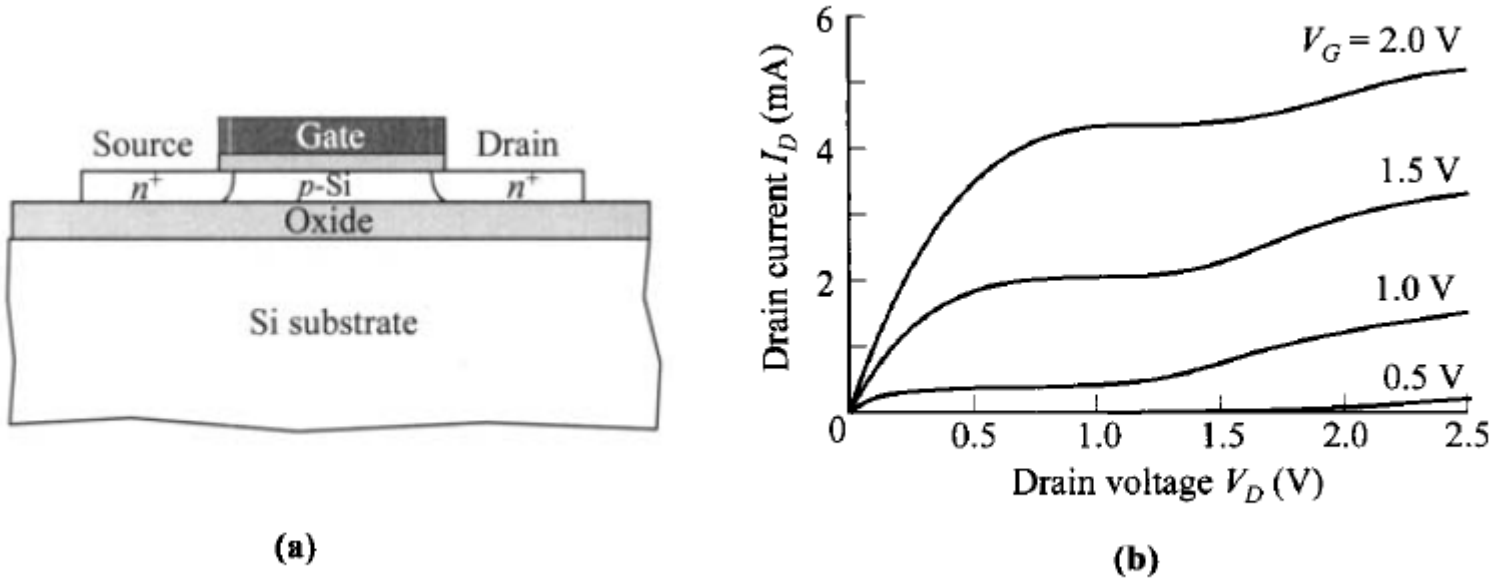
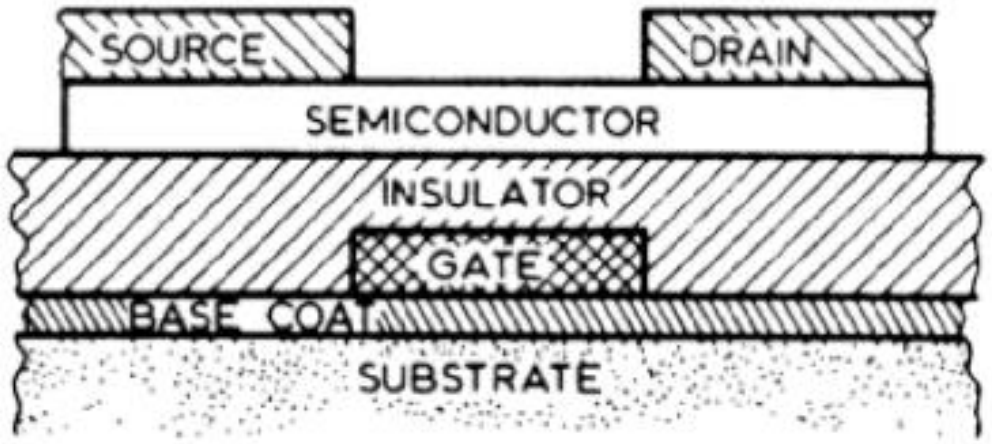
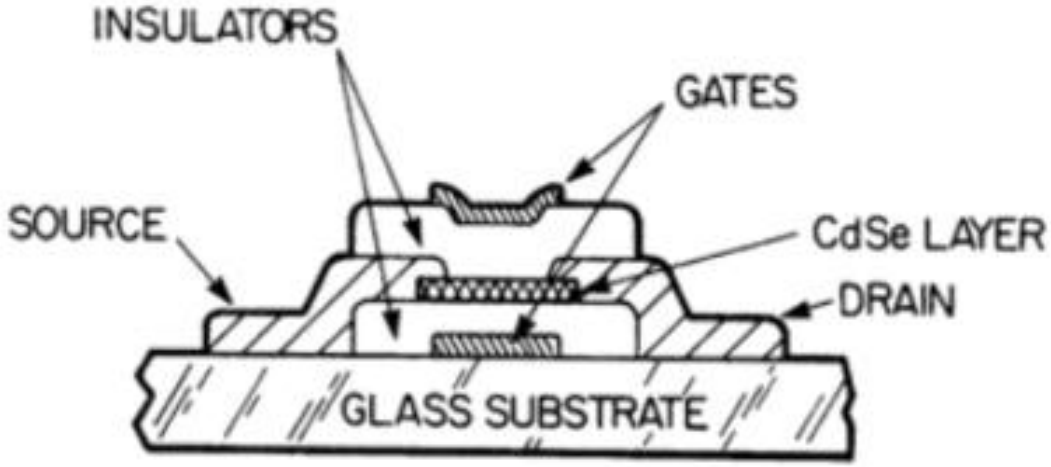


Fig. 37 (a) Typical structure of MOSFET on SOI wafer, and (b) its drain characteristics.

Las ventajas del sustrato SOI mejora el escalado del MOSFET debido a su delgado “cuerpo”. Este alivia la mayoría de los problemas del “punch through” a que el canal puede ser ligeramente dopado. Se mejora S. El óxido enterrado sirve como buen aislamiento y reduce la capacitancia del sustrato, dando mejor rapidez. El aislamiento del dispositivo mejora, simplemente removiendo la capa que lo rodea, mejorando densidad de circuito

TMOS SOI y Thin-Film Transistor (TFT)

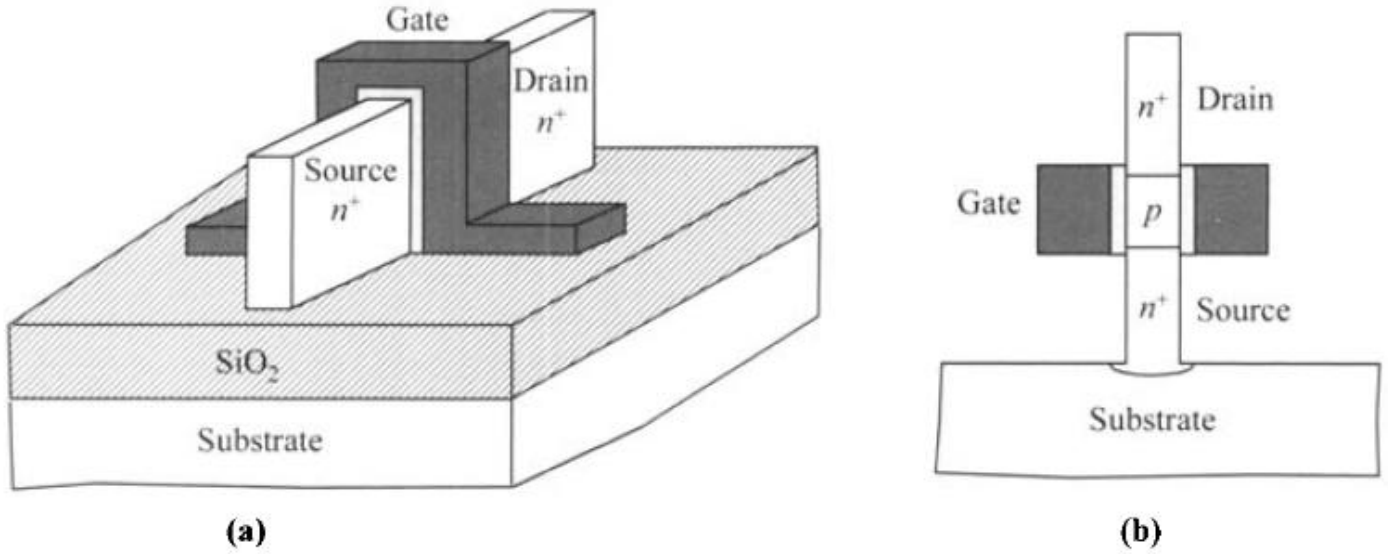




“Si el canal solo necesita una película delgada de semiconductor sobre un aislante… ¿por qué debe ser silicio monocristalino?”

Como la capa semiconductora es formada por depósito, un material amorfo tiene más defectos e imperfecciones que uno cristalino, resultando en procesos de transporte más complejos. En los TFTs la corriente siempre es muy limitada debido a baja movilidad, corrientes de fuga. Fueron usados para aplicaciones de área grande o sustratos flexibles donde el desempeño del dispositivo no es tan crítico.

TMOS Estructuras de tres dimensiones



En el escalado de dispositivos, el diseño óptimo consiste en un MOSFET construido sobre un cuerpo de capa ultrafina, de modo que este se agote completamente en todo el rango de polarización. Un diseño para lograr esto de forma más eficiente consiste en una estructura de puerta envolvente que encierra la capa del cuerpo por al menos dos lados.

CARACTERIZACIÓN DEL TRANSISTOR MOS

TMOS Caracterización de un transistor MOS

Consideramos que los parámetros eléctricos importantes para caracterizar un transistor MOS son:

- El voltaje de umbral, V_T;
- La movilidad máxima, μ_0;
- El voltaje de saturación, V_{dsat};
- El voltaje de ruptura, $V_{rup.}$

Existen otros parámetros muy importantes para caracterizar al TMOS que deben de tenerse en cuenta para el diseño de los circuitos integrados con estos transistores, estos son:

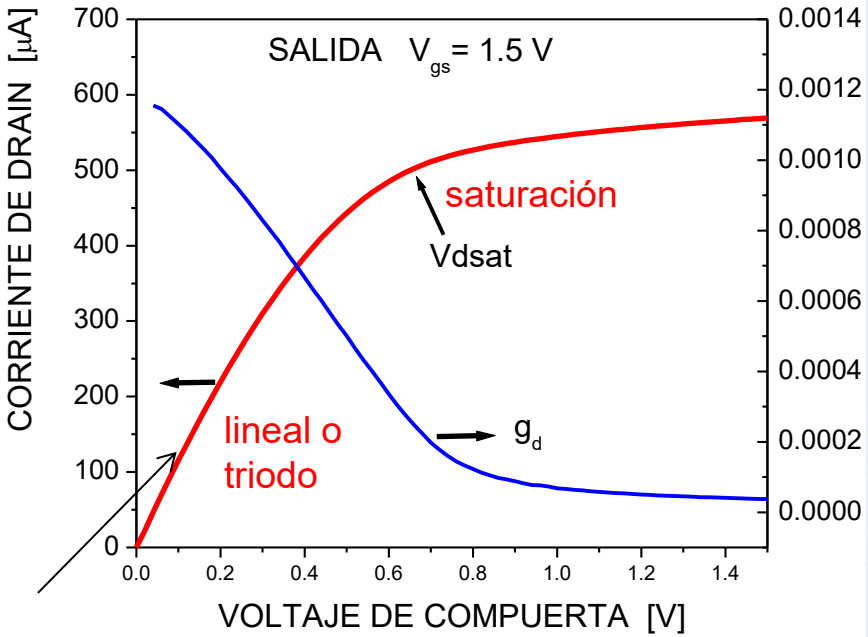
- En la característica de salida el parámetro R_{ON} y el factor de incremento de la corriente en saturación λ;
- La trasconductancia, g_m;
- La relación g_m/I_D;
- El voltaje de Early;
- La ganancia de voltaje a circuito abierto.

TMOS Característica de salida

En la característica de salida de un TMOS se ven tres regiones bien definidas:

1. Una zona lineal para $\mathbf{V_D}$ pequeños, que define la resistencia en el origen $\mathbf{R_{ON}}$.
2. La llamada región lineal o de triodo, desde ***0 a*** $\boldsymbol{V_{Dsat}}$.
3. La región de saturación para $\boldsymbol{V_D > V_{Dsat}}$.

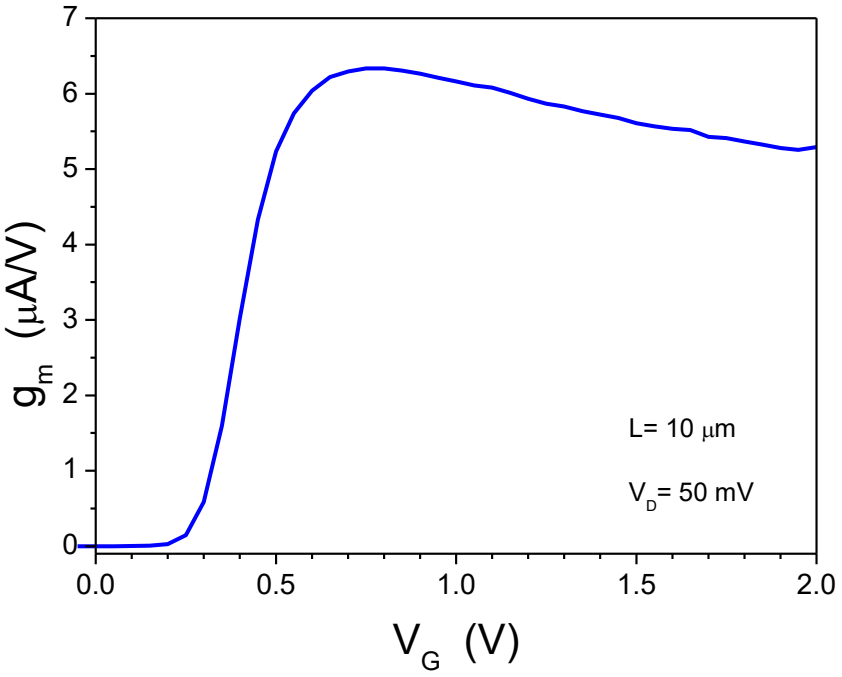
   En esta región la corriente se incrementa con una pendiente λ.

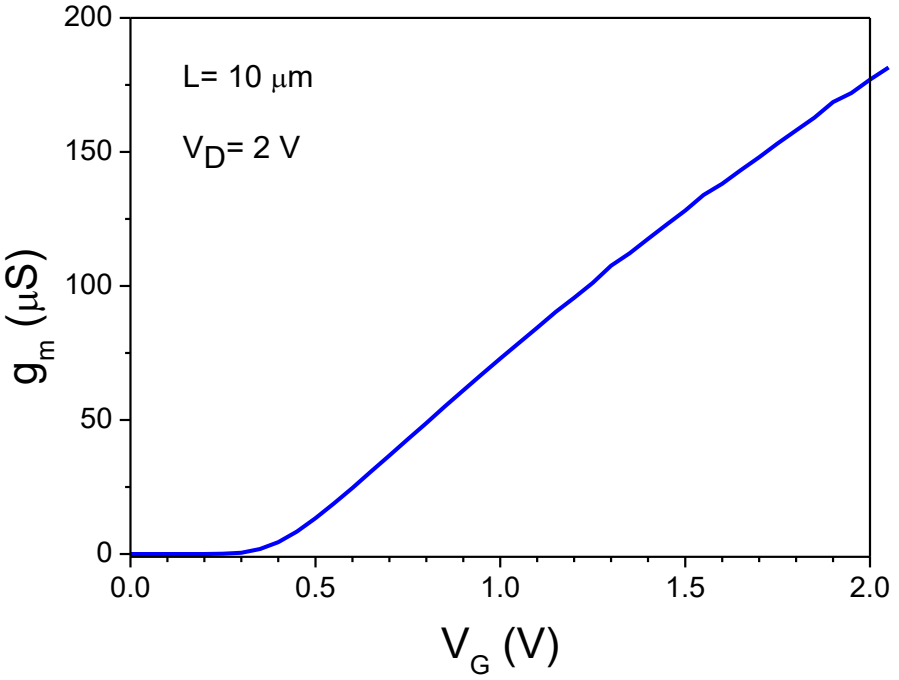


$\mathbf{R_{ON}} = 1/g_d$
en $\mathbf{V_D}$= 0V

TMOS Transconductancia

$$g_m = \left.\frac{\partial I_d}{\partial V_g}\right|_{V_d const}$$





REGIÓN LINEAL

REGIÓN DE SATURACIÓN

TMOS g_m/I_D

g_m/I_D es una magnitud muy importante para el diseño de circuitos analógicos. Se han desarrollado un procedimiento de diseño de amplificadores y otros circuitos basados en esta relación.

Se muestra generalmente en función del logaritmo de la corriente, para expandir el eje de corrientes.

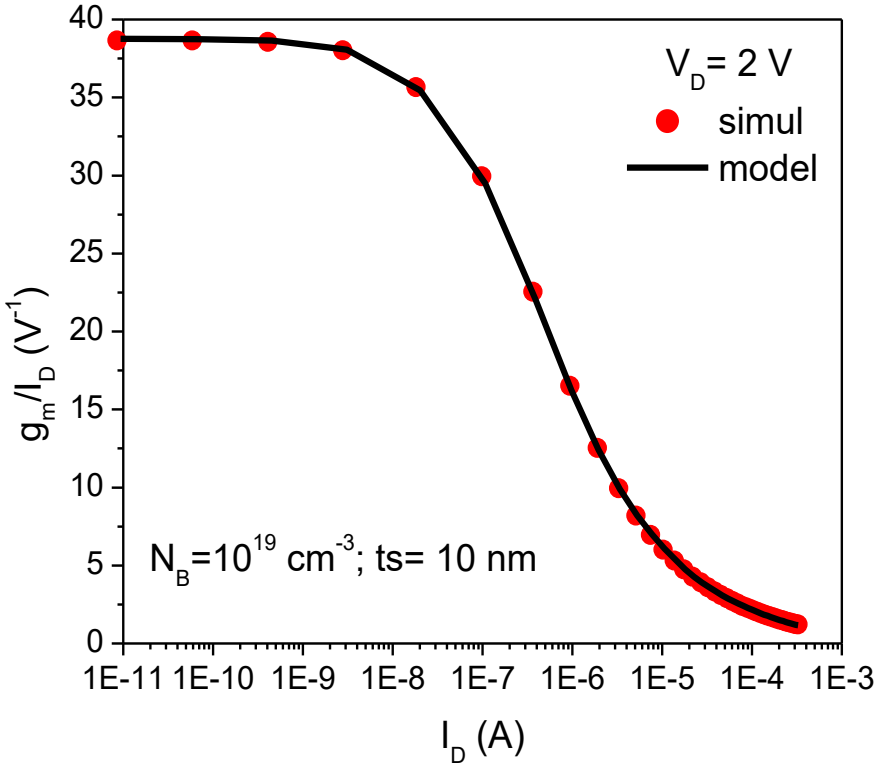
Para estos modelos las expresiones aquí deducidas son válidas a partir de un voltaje de compuerta ligeramente superior al voltaje de umbral.

TMOS g_m/I_D

Actualmente la tendencia es usar menores voltajes y Corrientes, o sea, menor potencia y obtener mayores ganancias, lo que se obtiene incrementando g_m/I_D.

En la región de transición es difícil modelar y de utilizar, debido a que se produce una variación brusca del parámetro.

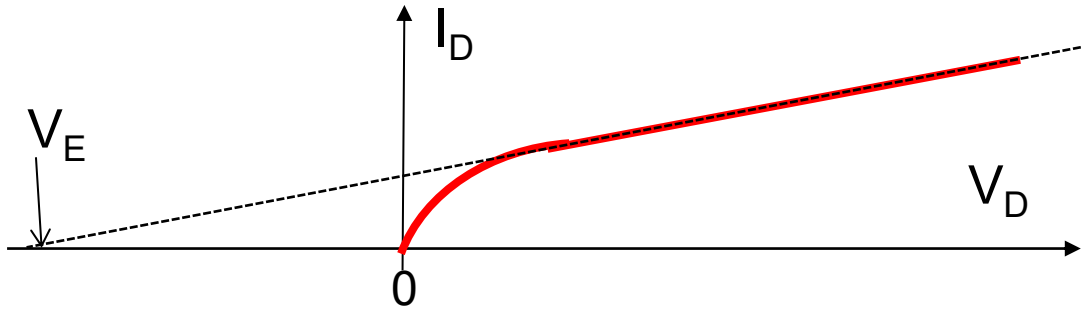
Típicamente se calcula este parámetro en saturación para el diseño de amplificadores.



TMOS Voltaje de Early

Se introduce la definición del Voltaje de Early, V_E, en el transistor MOS como la traza de la característica de salida en el eje del voltaje.

A mayor V_E menor conductancia y mayor ganancia del transistor.



$$V_E \ crece \Rightarrow \quad A_{V0} crece$$

TMOS Ganancia de voltaje

Otro parámetro importante del TMOS para aplicaciones analógicas es la ganancia de voltaje a circuito abierto $\boldsymbol{A_{V0}}$.

$$A_{V0} = \frac{\Delta V_D}{\Delta V_G} = \frac{\Delta V_D}{\Delta I_D} \cdot \frac{\Delta I_D}{\Delta V_G} = \frac{g_m}{g_d} = \frac{g_m}{I_D} \cdot V_E$$

VELOCIDAD DE OPERACIÓN DE UN TRANSISTOR MOS

TMOS Velocidad de respuesta

La velocidad de respuesta, o de conmutación, de un TMOS depende de diferentes factores:

1. El tiempo de tránsito de los portadores por el canal
2. Las capacitancias propias de la estructura MOS
3. Las capacitancias parásitas

Para la evaluación del tiempo de tránsito se usará la corriente en saturación con las expresiones de primera aproximación.

TMOS Velocidad de respuesta

Ley de Ohm diferencial

$$j = \sigma E = qn\mu E = -qn\mu \frac{dV}{dy} = \frac{i}{x_c W}, \quad Q_n = qnx_c$$

Carga móvil

$$Q_n = -C_o[V_G - V_T - V(y]$$

resulta

$$dV(y) = \frac{(V_G - V_T)^2 dy}{2L(V_G - V_T - V(y))}$$

Cuando y=L, V(L)=V_{Dsat}

$$E_y(y) = -\frac{V_G - V_T}{2L} \frac{1}{\sqrt{1 - \frac{V}{V_{Dsat}}}} = -\frac{V_G - V_T}{2L} \frac{1}{\sqrt{1 - \frac{y}{L}}}$$

$$v(y) = \frac{dy}{dt} = -\mu E(y) \longrightarrow t_{tr} = \int_0^L \frac{dy}{\mu E_y(y)}$$

TMOS Velocidad de respuesta

Integrando se obtiene el tiempo de tránsito entre S y D

$$t_{tr} = \frac{4}{3}\frac{L^2}{\mu_n (V_G - V_T)}$$

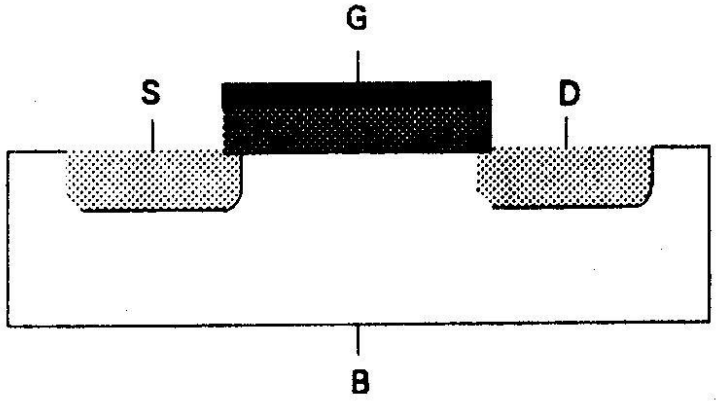
Si L= 3 μm; μ_n= 600 cm^2/Vs; V_G-V_T= 5 V; t_{tr}= 40 ps

L= 0.1 μm; μ_n= 400 cm^2/Vs V_G-V_T= 1 V; t_{tr}= 0.25 ps

Período (ps)	Frecuencia (Hz)
40	25 GHz
0.25	4 THz

TMOS Capacitancias

Capacitancias intrínsecas y parásitas de un TMOS



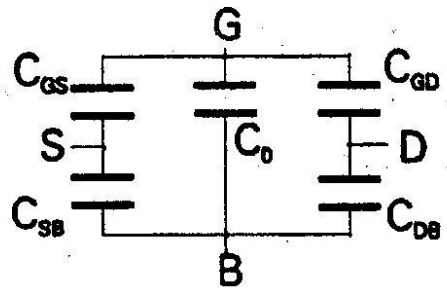


Fig. 3.1 Capacitancias en un transistor MOS, intrínsecas y parásitas.

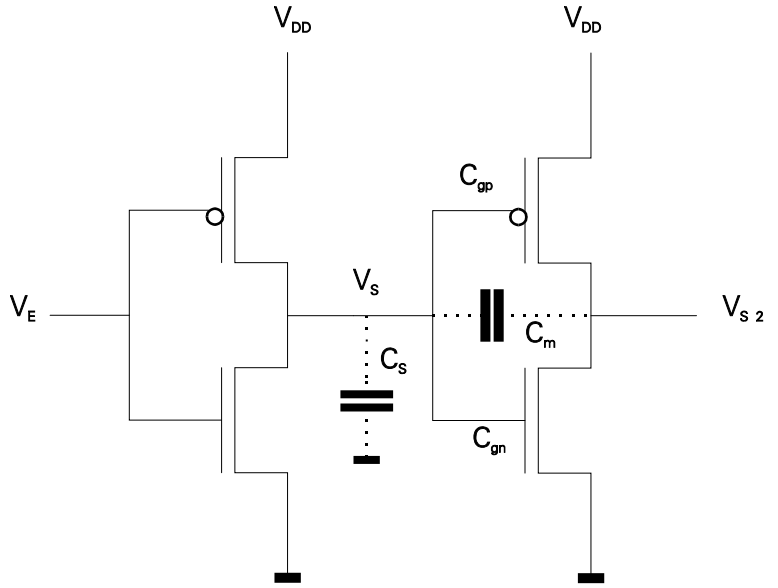


Fig. 4.11 Capacitancias de carga de un inversor CMOS además de las de compuerta de cada transistor.

FIN DEL TEMA 2B