

MICROELECTRÓNICA

TEMA 4: EL TRANSISTOR MOS COMO ELEMENTO CIRCUITAL DIGITAL

Desde la aparición del transistor MOS (TMOS) las áreas de aplicación del mismo se han ido ampliando y especializando. Actualmente las principales áreas de aplicación son:

- transistores de muy alta impedancia de entrada;
- transistores de potencia;
- circuitos integrados digitales de muy alta integración: lógicos y de memorias;
- circuitos integrados analógico-digitales.

El área de mayor impacto científico, tecnológico y económico a sido la de los circuitos integrados de muy alta integración, donde en la actualidad se alcanzan niveles de integración de 3 - 10 millones de transistores en un “chip” para los circuitos lógicos, y decenas de millones de transistores en las memorias. Esto ha sido posible por las reducidas dimensiones que pueden tener los TMOS, la fiabilidad de la tecnología de fabricación, los reducidos consumos de energía y las velocidades relativamente altas de operación, hasta decenas de GHz.

El desarrollo alcanzado es el resultado del desarrollo de múltiples ramas de la ciencia y la tecnología, como la obtención y preparación de los materiales de base, la purificación de los productos químicos de grado electrónico, el desarrollo de las técnicas fotolitográficas para dimensiones submicrométricas, y el desarrollo de equipamiento tecnológico y de medición altamente sofisticados y automatizados.

A continuación se expondrán algunos aspectos relacionados con la utilización de los TMOS como elementos de los circuitos integrados digitales, entre ellos el inversor, las compuertas lógicas básicas y las celdas de memoria.

4.1 EL INVERSOR

El inversor es el circuito digital básico, donde la salida niega lógicamente la entrada. A partir del mismo se pueden hacer los restantes bloques lógicos. En la Fig. 4.1 se muestran los principales tipos de inversores que se pueden conformar con TMOS. Normalmente el transistor de entrada es del tipo de enriquecimiento y se conoce como transistor de control (“*driver*”). También se conoce como transistor de “*pull down*”, por ser el que descarga los capacitores cargados a la salida. Actualmente se hacen los TMOS de control generalmente con canal N (NMOS) de enriquecimiento.

Como elementos de carga se pueden usar diferentes tipos de elementos, a los fines didácticos y como parte del desarrollo histórico se verán cinco casos: 1) carga resistiva; 2) carga con TMOS de enriquecimiento saturado; 3) carga con TMOS de enriquecimiento no saturado; 4) carga con TMOS de empobrecimiento (“*depletion type*”) y 5) carga con TMOS de enriquecimiento canal P (PMOS) formando una pareja complementaria, o sea un inversor CMOS. El TMOS de carga recibe también el nombre de “*pull up*”, por ser a través del cual se carga la capacitancia de salida.

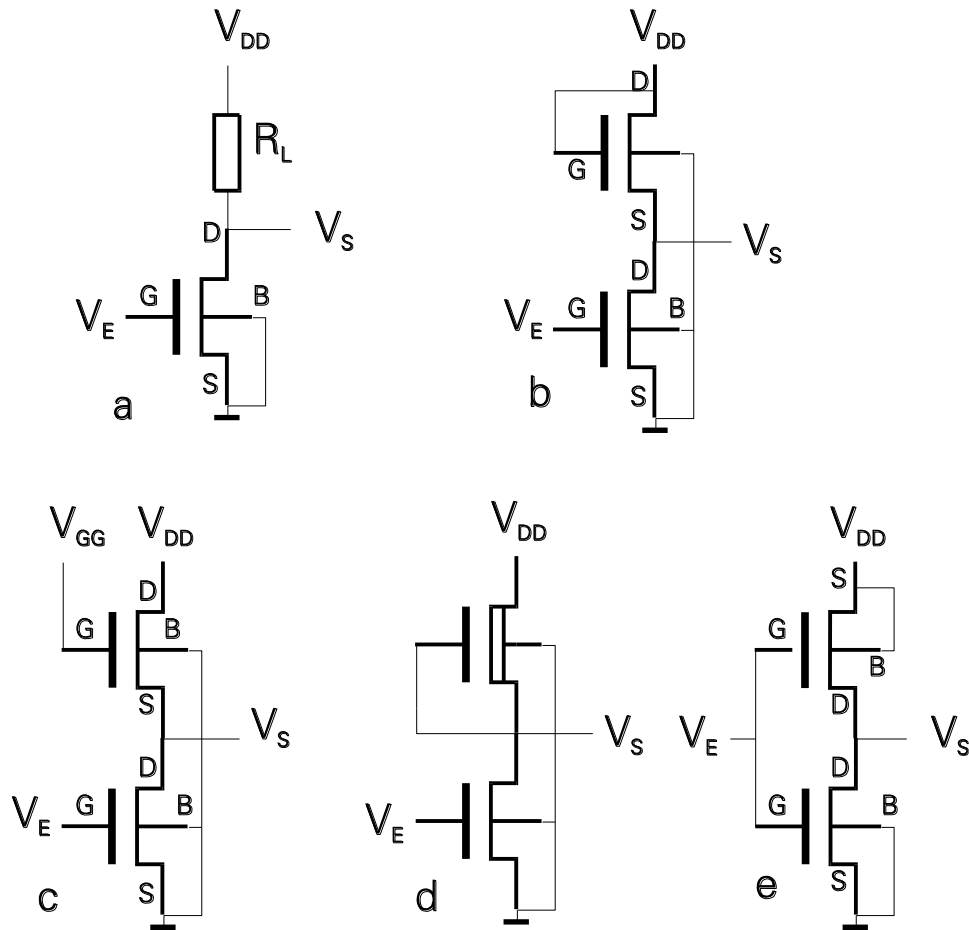


Fig. 4.1 Inversores típicos:
a) con carga resistiva;
b) con carga activa saturada;
c) con carga activa no saturada;
d) con carga de empobrecimiento;
e) con carga complementaria.

A continuación se precisaran algunas definiciones y denominaciones de uso generalizado en la técnica digital.

SÍMBOLO. - Cada elemento o bloque circuital tiene un símbolo normado a los fines de poder representarlo en los diagramas eléctricos de los circuitos. En la Fig. 4.2a se muestra el símbolo del inversor en una norma generalizada en los EE.UU. Hay otras normas de gran difusión donde el símbolo siempre corresponde con un rectángulo y algunas indicaciones interiores, Fig. 4.2b.

NIVELES LÓGICOS. - En la técnica digital se utilizan diferentes tipos de lógica, la más generalizada y que se utilizará en las presentes notas es la lógica positiva bivevaluada. En la Fig. 4.2c se muestra un diagrama de los niveles de voltaje que se utilizan en este caso. Se definen dos niveles lógicos, el alto igual a “H” o a “1”, y el nivel bajo “L” o “0”.

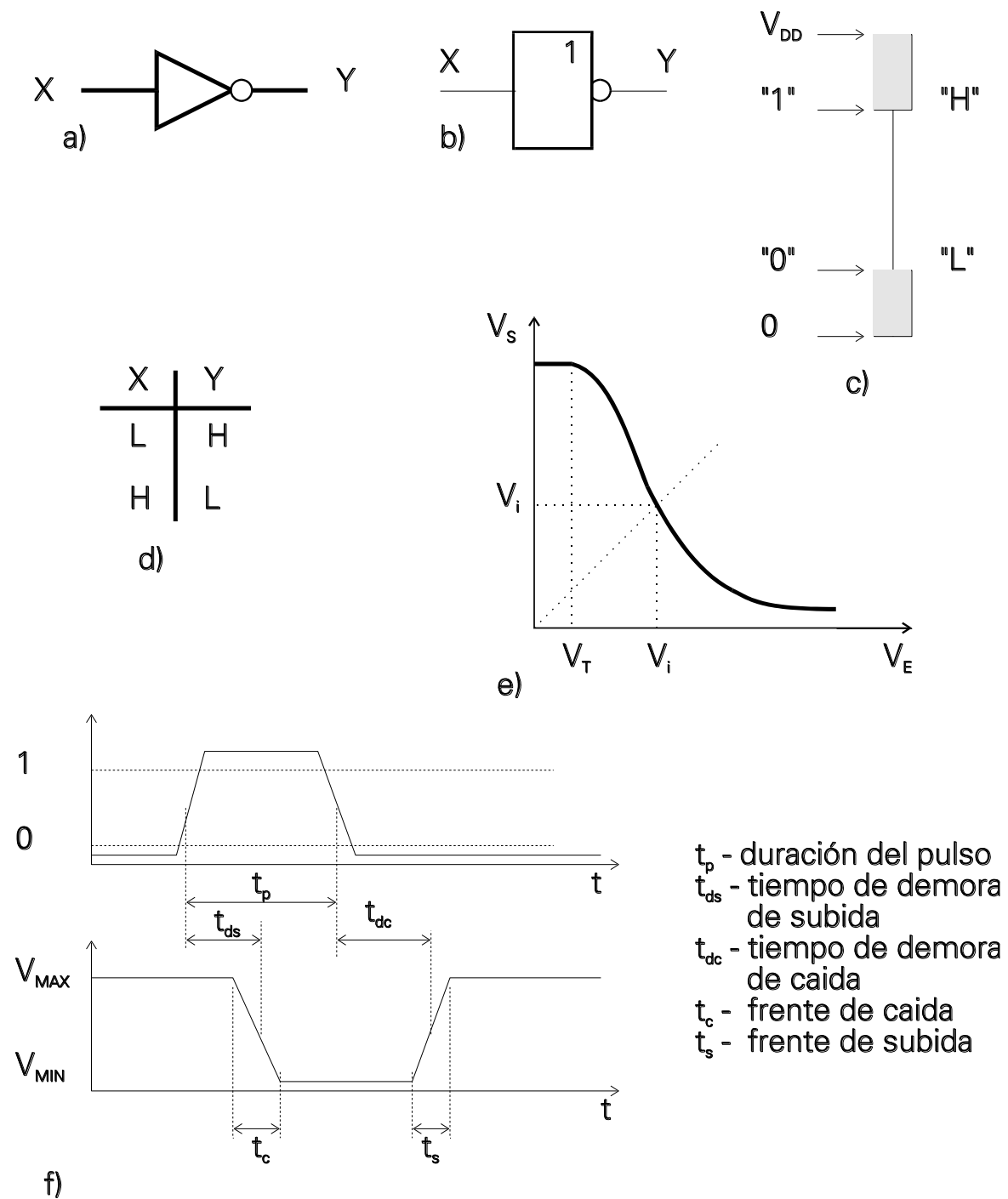


Fig. 4.2 Símbolos y características de los inversores:
 a) símbolo lógico; b) símbolo lógico;
 c) niveles lógicos; d) tabla de verdad;
 e) característica transferencial; f) diagramas de tiempo.

V_E - voltaje de entrada
 V_s - voltaje de salida
 V_i - voltaje para $V_E = V_s$

Se dice que el circuito está en el nivel lógico **H** cuando el voltaje de salida está entre el nivel **H** y el voltaje de la fuente V_{DD} . El estado que corresponde al nivel lógico **L** es cuando el voltaje de salida está entre el valor de tierra (0) y el valor del nivel **L**. En los casos que el voltaje de salida este entre estas dos regiones se dice que el circuito tiene un estado indefinido o esta en transición.

TABLA DE VERDAD. - Una forma precisa de definir el estado final posible en un circuito digital es el uso de la llamada Tabla de Verdad. En la misma es necesario definir todas las posibles condiciones en las entradas del circuito, y describiendo cuales serán los estados en cada una de las salidas del circuito. El ejemplo de un inversor se da en la Fig. 4.2d. Cada estado se define utilizando el símbolo **H** o **L**, y si no importa cual sea el estado, por una **X**.

CARACTERÍSTICAS TRANSFERENCIALES. - La característica transferencial de un circuito dado es la característica para corriente directa, que muestra la relación entre el voltaje de salida V_S y el de entrada V_E . En la Fig. 4.2e se muestra el ejemplo para un inversor. En la misma se indican algunos puntos significativos como son el voltaje umbral, a partir del cual comienza a caer el voltaje de salida, el voltaje V_i para el cual $V_S = V_D = V_i$.

DIAGRAMA DE TIEMPOS. - La descripción de la variación de los voltajes en las entradas y en las salidas de un circuito con el tiempo se realiza a través de los diagramas de tiempo. En la Fig. 4.2f se muestra el ejemplo de un inversor. Los momentos principales que se señalan en estos diagramas son los tiempos de subida y de caída de los voltajes, lo cuales se miden entre los niveles correspondientes al 10% y el 90% de los voltajes máximos, V_{max} y V_{min} . Los tiempos de demora de los frentes de subida y de caída. Los tiempos de demora de los frentes desde que se aplican a la entrada, hasta que se definen a la salida. Para tener una definición se utilizan los niveles de $0.5(V_{max} - V_{min})$ como referencias de tiempo.

Es totalmente claro que $V_{max} > \mathbf{H}$ y $V_{min} < \mathbf{L}$.

4.1.1 INVERSOR CON CARGA RESISTIVA

En la Fig. 4.1a se muestra el inversor MOS con carga resistiva. El TMOS de control puede ser tanto de canal P (PMOS), como de canal N (NMOS). Al comienzo de los años 70 se utilizaban casi exclusivamente los PMOS, por que la tecnología existente permitía obtener PMOS con voltaje umbral negativo, siendo también los voltajes de drenaje y compuerta negativos, de esta forma cuando $V_G = 0$, la corriente se hacía 0, Fig. 4.3. En el caso de los NMOS el voltaje umbral es también negativo, pero los otros voltajes son positivos, así para $V_G = 0$ hay corriente de drenaje. Cuando la tecnología permitió obtener NMOS con voltajes umbral positivos, casi automáticamente se pasó a usar los NMOS, ya que la movilidad de los electrones en el canal de estos es al menos el doble de la de los huecos.

Si el transistor de control tiene un voltaje umbral igual a V_{TD} , y la resistencia de carga es R_L , Fig. 4.4a, un cálculo estimado de los dos estados lógicos corresponde con:

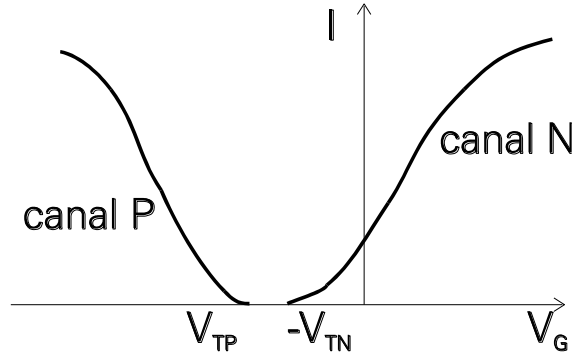


Fig. 4.3 Voltajes umbral para los transistores canal N y los canal P, sin ajustes.

- 1) $V_E = V_{\min} < V_{TD}$; $V_S = V_{DD}$; $I_D = 0$.
- 2) $V_E = V_{DD} \gg V_{TD}$; $V_S = V_{\min} < V_{TD}$; $I_D = I_0$, en la región lineal.

El consumo de energía se produce cuando el inversor está el segundo estado, y en las transiciones hacia o desde el mismo.

Por ser un elemento de gran importancia se calculará en este caso el nivel de V_{\min} para este caso. Para ello se considerará la corriente a través de la resistencia y del TMOS, por lo cual

$$I_0 = \frac{V_{DD} - V_{\min}}{R_L} = \beta_D \left[(V_G - V_{TD}) V_D - \frac{V_D^2}{2} \right]; \quad (4.1)$$

donde $\beta_D = \mu C_0 \frac{Z}{L}$;

$$V_{DD} - V_{\min} = \beta_D R_L \left[(V_{DD} - V_{TD}) V_{\min} - \frac{V_{\min}^2}{2} \right]; \quad (4.2)$$

$$V_{\min} = \frac{1 + \beta_D R_L (V_{DD} - V_{TD})}{\beta_D R_L} \left[1 - \sqrt{1 - \frac{2 V_{DD} \beta_D R_L}{[1 + \beta_D R_L (V_{DD} - V_{TD})]^2}} \right]. \quad (4.3)$$

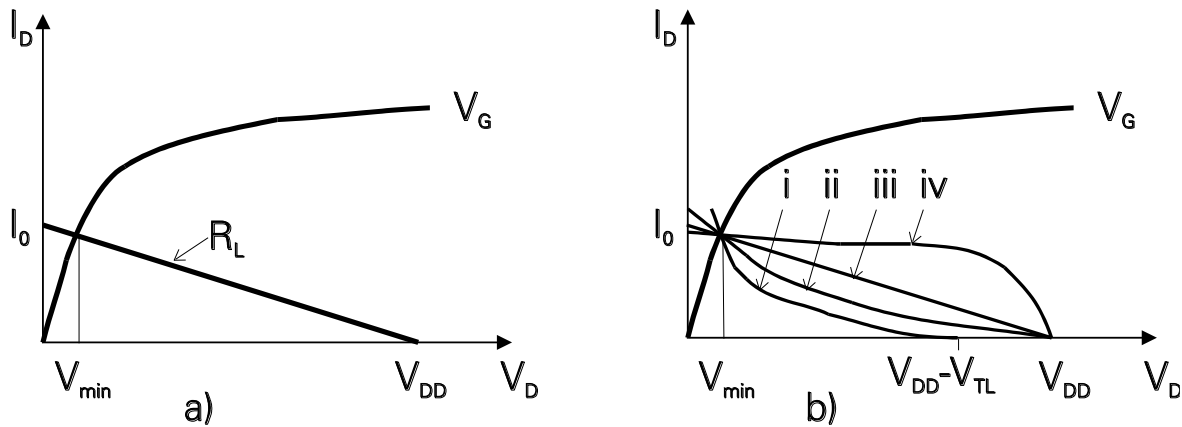


Fig. 4.4 Características de salida de transistor de control con diferentes tipos de carga.
a) carga resistiva R_L ;
b) diferentes cargas: i) TMOS saturado; ii) TMOS no saturado;
iii) carga resistiva; iv) carga de empobrecimiento.

Es conveniente analizar el siguiente ejemplo. Para un NMOS con los siguientes parámetros: $x_0 = 60 \text{ nm}$; $Z/L = 5$; $m_0 = 600 \text{ cm}^2/\text{Vs}$; $q = 0.3 \text{ 1/V}$; $V_T = 1 \text{ V}$; $R = 1000 \text{ } \Omega/\square$; $N=30$.

De acuerdo con estos datos la resistencia de carga será igual a $30 \text{ k}\Omega$, por lo tanto aplicando la Ec. 4.1 y después la Ec. 4.3 resulta que $V_{\min} = 0.508 \text{ V}$. En este caso todo es normal ya que $V_{\min} < V_T$, lo cual garantiza que no conduzca el transistor en este estado. El problema es que para lograr esto se requiere una resistencia de carga alta, la cual es posible de obtener, por ejemplo, si se tiene una resistencia implantada de $10 \text{ } \mu\text{m}$ de ancho y $300 \text{ } \mu\text{m}$ de largo, o sea $3000 \text{ } \mu\text{m}^2$ de resistencia, mientras que el área activa del NMOS analizado es de $500 \text{ } \mu\text{m}^2$, 6 veces menos. Por este uso excesivo de área esta solución fue desechada rápidamente, prefiriéndose el uso de cargas activas de mucha menor área efectiva.

4.1.2 INVERSOR CON CARGA SATURADA

Históricamente el primer inversor con carga activa es de carga saturada, Fig. 4.1b. La carga consiste en un NMOS de enriquecimiento con la compuerta conectada a la fuente V_{DD} , con lo cual $V_G = V_D = V_{DD}$, lo que implica que siempre estará en la región de saturación. El voltaje umbral del transistor de control será V_{TD} y el del transistor de carga V_{TL} . Los dos estados lógicos en este caso en que para el TMOS de carga el voltaje compuerta-fuente es $V_{GS} = V_{DD} - V_S > V_{TL}$, se definen según:

- 1) $V_E = V_{\min} < V_{TD}$; $V_S = V_{DD} - V_{TL}$; $I_D = 0$;
- 2) $V_E = V_{DD} - V_{TL}$; $V_S = V_{\min}$; $I_D = I_{\min}$.

En la Fig. 4.4b se muestra la característica de salida con la línea de carga resistiva y con la línea de carga correspondiente con este transistor saturado. La corriente en todo momento está por debajo de la de carga resistiva, lo que reduce la velocidad de operación al ser los niveles de corriente menores. Otro momento importante es que el voltaje máximo a alcanzar a la salida es $V_{DD} - V_{TL}$, lo cual también reduce la velocidad de conmutación.

Debe de considerarse que como el sustrato de ambos TMOS es común, el de carga tendrá un voltaje entre compuerta y sustrato adicional dado que $V_{BS} = -V_S$, lo que resulta en un aumento de V_{TL} .

Por ejemplo, para el TMOS antes considerado si la carga equivalente en el óxido es de $5 \times 10^{10} \text{ cm}^{-2}$ y se hace una implantación de ajuste del voltaje umbral hasta $2 \times 10^{16} \text{ cm}^{-3}$ resultando un $V_T = 0.9 \text{ V}$. Al subir el voltaje de salida hasta unos 4V el V_{TL} aumenta unos 2V.

Por todo lo señalado este tipo de inversor es susceptible de mejorarse y su utilización fue limitada.

4.1.3 INVERSOR CON CARGA NO SATURADA

El poder alcanzar el voltaje de la fuente se resuelve poniendo al transistor de carga en el régimen de no saturación. Para ello es necesario aplicarle un voltaje independiente a la compuerta, que se denomina V_{GG} y es mayor que V_{DD} al menos en V_{TL} . Como regla cuando $V_{DD} = 5\text{V}$, $V_{GG} = 12 \text{ V}$. La línea de carga en este caso mejora, ya que la salida llega hasta V_{DD} , Fig. 4.4, pero aún la corriente se mantiene por debajo de la de la carga resistiva.

A pesar del inconveniente de requerir de una fuente de voltaje adicional, este tipo de inversor fue ampliamente utilizado en los años 70 para el diseño de los circuitos integrados, fundamentalmente utilizando transistores del tipo canal P (PMOS).

En la Fig. 4.5 se muestra la característica transferencial de este inversor, donde se pueden señalar tres regiones. Mientras el TMOS de carga está siempre en la región lineal, el de control pasa por todas las regiones: si $V_E < V_{TD}$ está cortado; para el caso de $V_S > V_E - V_{TD}$ estará en saturación y si $V_S < V_E - V_{TD}$, se encontrará en la región lineal.

Otro momento en el desarrollo histórico de los inversores fue la utilización de la polarización de sustrato negativa para lograr ajustar el V_T usando un sustrato de alta resistividad. Este voltaje V_{BB} era del orden de -2 V . Su uso también fue temporal, por la necesidad de otra fuente de voltaje adicional.

El uso de estos inversores se fue eliminando a medida que la tecnología permitió realizar otro tipo de transistor de carga más ventajoso.

4.1.4 INVERSOR CON CARGA DE EMPOBRECIMIENTO

El uso de la compuerta de polisilicio, el ajuste de las características del canal con ayuda de la implantación iónica y el aislamiento de campo del tipo LOCOS permitió la utilización simultánea de transistores MOS canal N del tipo de enriquecimiento TMOS/E y del tipo de empobrecimiento (“normally ON”) TMOS/D. De esta forma aparecieron los inversores del tipo enriquecimiento - empobrecimiento (“enhancement-depletion”), o del tipo E/D.

El circuito de este inversor es el que se muestra en la Fig. 4.1d. Cuando el TMOS/D tiene un voltaje de compuerta 0, la corriente es alta y con una amplia región de saturación, lo cual lo hace una carga casi ideal. Si se ve la característica de salida del TMOS/E de control con una carga del tipo TMOS/D, Fig. 4.4, se nota que en este caso la corriente es superior a la de la carga resistiva durante toda la variación del voltaje de salida, lo cual acelera la operación del inversor. Además, se obtiene una mayor resistencia de carga dinámica con menos área y la variación del voltaje de salida va de V_{\min} a V_{DD} .

Los dos niveles lógicos corresponderán a los siguientes estados de la entrada y la salida:

- 1) $V_E = V_{\min} < V_{TD}$; $V_S = V_{DD}$; $I_D = 0$;
- 2) $V_E = V_{DD}$; $V_S = V_{\min}$; $I_D = I_0$.

El gasto de energía se produce cuando se encuentra en el estado de “0” lógico a la salida, y en las transiciones desde y hacia este estado.

Se conoce que en saturación la corriente del TMOS/D es igual a:

$$I \approx \beta_L (-V_{TL})^2, \quad \text{donde} \quad \beta_L = \frac{\mu_D C_0 Z_L / L_L}{1 + \frac{C_0}{C_I} \left(\frac{C_0}{2C_I} + 1 \right)}, \quad (4.4)$$

donde μ_D es la movilidad en la zona del canal construido, generalmente superior a la del canal inducido, y C_I es la capacitancia del canal implantado, con una profundidad d .

En el caso del TMOS/E, cuando se calcula V_{\min} se encuentra en la región lineal, por eso se puede escribir que

$$\beta_D \left[(V_{DD} - V_{TD}) V_{\min} - V_{\min}^2 / 2 \right] = \beta_L (-V_{TL} / 2)^2, \quad (4.5)$$

$$V_{\min} = \frac{V_{DD} - V_{TD}}{2} \left[1 - \sqrt{1 - 4 \frac{\beta_L}{\beta_D} \left(\frac{-V_{TL}}{V_{DD} - V_{TD}} \right)^2} \right] \quad (4.6)$$

Si se desea calcular el llamado voltaje de umbral lógico V_i , que corresponde con el momento en que se igualan los voltajes de entrada y salida, y considerando que ambos transistores están en saturación, resulta:

$$\frac{\beta_D}{2} (V_i - V_{TD})^2 = \frac{\beta_L}{2} (-V_{TL})^2, \quad (4.7)$$

de donde

$$V_i = V_{TD} - \sqrt{\frac{\beta_L}{\beta_D}} (V_{TL}). \quad (4.8)$$

A los fines de los estimados prácticos para este inversor se considera que $V_{TD} = 0.2V_{DD}$ y $V_{TL} = -0.6V_{DD}$. Así, si se quiere que $V_i = 0.5$, la relación $b_L/b_D = 0.25$. Si se considera que las movilidades son aproximadamente iguales resultara que las relaciones de aspectos deben ser:

$$Z_D/L_D = 4Z_L/L_L. \quad (4.9)$$

Para $Z_D = 40 \mu\text{m}$ y $L_D = 10 \mu\text{m}$, el TMOS de carga debe ser de $Z_D = 10 \mu\text{m}$ y $L_D = 10 \mu\text{m}$.

Si se calcula el valor de V_{\min} en este caso resulta que

$$V_{\min} = 0.13V_{DD} = 0.65 \text{ V} < V_{TD} = 0.2V_{DD} = 1 \text{ V}.$$

La forma de la característica transferencial es similar a la de la Fig. 4.5.

A continuación, veremos en el ejemplo de este inversor algunos parámetros complementarios que son generales a todos los inversores y que sirven para caracterizarlos.

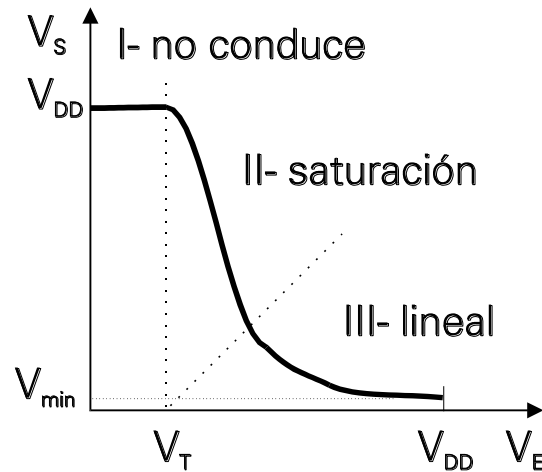


Fig. 4.5 Característica transferencial para el caso de carga no saturada, mostrando las tres regiones de operación del TMOS de control.

MÁRGENES DE RUIDO. - Una de los parámetros que se definen para los inversores es el llamado margen de ruido. Este parámetro se define del siguiente modo; en la característica transferencial, Fig. 4.6, se definen los dos momentos (voltajes) en los cuales el factor de ganancia $g = V_s/V_E = 1$, V_1 y V_2 . La zona activa de transición se considera la zona en la cual $g > 1$, donde se supone que el inversor pase a un estado u otro. En la zona desde V_{min} hasta V_1 , $g < 1$, es lo que se llama margen de ruido del "0", ya que cualquier variación del "0" que no sobrepase V_1 no debe causar una transición. El margen de ruido del "1" será en este caso del V_2 hasta V_{DD} . Ambas zonas aparecen rayadas en la figura. Se consideran mejores aquellos inversores en que los márgenes de ruido son mayores.

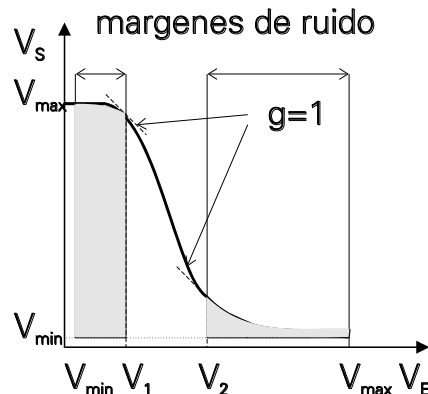


Fig. 4.6 Margenes de ruido de un inversor:
- para el "0" es de $V_{min}-V_1$;
- para el "1" es de V_2-V_{max} .

TIEMPOS DE RESPUESTA. - Siempre hay asociado un capacitor equivalente a la salida de un inversor que tiene en cuenta la capacitancia de entrada de la etapa posterior, de interconexión, etc. Los tiempos de carga y de descarga de estos capacitores definen el llamado tiempo de respuesta del inversor.

En el caso de que el inversor esté en **H**, Fig. 4.7a, el capacitor de carga C tiene un voltaje igual a V_{DD} . Cuando se aplica un pulso de voltaje a la entrada, Fig. 4.7b, el transistor de control comenzará a conducir en la región de saturación con corriente cuasiconstante, y como se muestra en la Fig. 4.7c, C se comienza a descargar a través del TMOS. Durante este tipo de descarga el voltaje cae casi linealmente hasta t_1 , donde se alcanza la condición para pasar a la región lineal, $V_S = V_{DD} - V_{TD}$. A partir de este momento la descarga sigue una ley exponencial con una constante de tiempo t_c .

La carga del capacitor C se muestra en la Fig. 4.8, donde el inversor está en **L** y se le aplica a la entrada un pulso que va desde V_{DD} a V_{min} , con lo cual se cierra el TMOS de control, produciéndose la carga del C a través del elemento de carga. Si la carga fuera una resistencia R_L , curva *ii*, el aumento de voltaje en el C es exponencial. Para los casos de cargas con TMOS/E las corrientes son menores, curva *iii* no saturado, curva *iv* saturada, por lo cual la carga es más lenta. En el caso de la carga con TMOS/D, curva *i*, las corrientes son superiores realizando una carga de C más rápida, lo cual es una ventaja notable de este tipo de inversor.

LÓGICA DINÁMICA. - Hasta el momento todo lo que se mostrado corresponde a la llamada lógica estática, en la cual los niveles lógicos se conservan en el tiempo, en el estado de los transistores. En su momento se buscaron soluciones circuitales para aumentar la velocidad de operación y disminuir el consumo de energía, desarrollándose lo que se llamó lógica dinámica. Los principios en los que se basa son:

- 1) la alimentación no es continua, sino en forma de pulsos;
- 2) un estado dado se mantiene por la carga de los capacitores de interconexión, por lo tanto como esa carga se pierde en el tiempo, se requiere recargar esos capacitores para mantener el estado.

Si en la operación se utilizan dos pulsos desfasados entre sí, se llama lógica de 2 fases, y si se utilizan cuatro pulsos, lógica de 4 fases. Esta lógica se conoce también como “ratioless logic”.

4.1.5 EL INVERSOR CMOS

El inversor MOS complementario (CMOS), Fig. 4.1e, consiste en un NMOS/E como transistor de control y en un PMOS/E como transistor de carga. En este caso la fuente y el sustrato del TMOS de carga de canal P están conectados a V_{DD} , mientras que la fuente y el sustrato del transistor de control, canal N, están conectados a tierra. De esta forma para ambos circuitos el $V_{GS} = 0$ en todo momento, lo cual elimina el efecto de sustrato. Ambos transistores están conectados por sus respectivos drenajes, siendo este nodo la salida del inversor. La entrada es la conexión común de ambas compuertas.

La característica de salida del TMOS con la carga del PMOS/E, se muestra en la Fig. 4.9. Los dos estados lógicos corresponderán en este caso a:

- 1) $V_E = 0$; NMOS no conduce; PMOS conduce; $V_S = V_{DD}$;
- 2) $V_E = V_{DD}$; NMOS conduce; PMOS no conduce; $V_S = 0$.

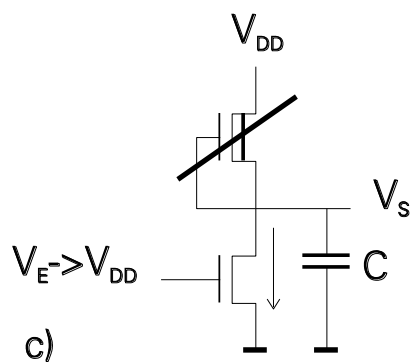
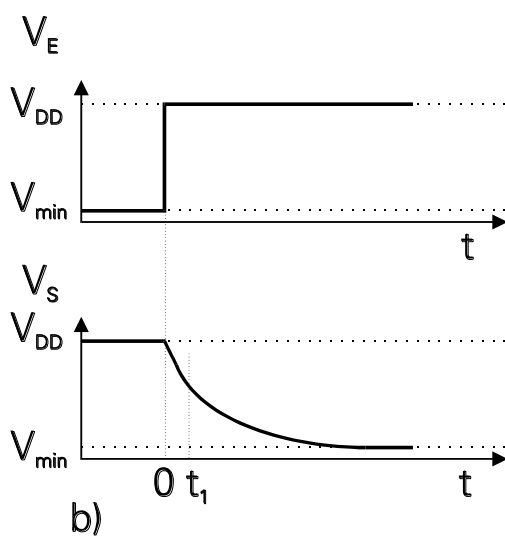
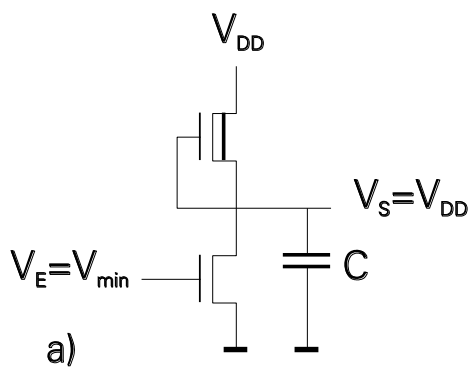


Fig. 4.7 Descarga del capacitor de salida C.

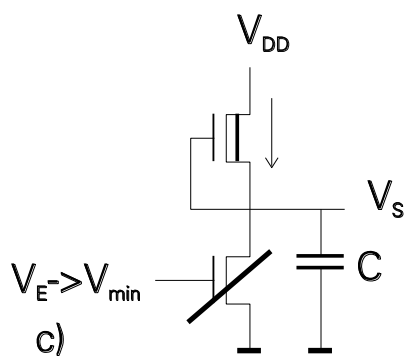
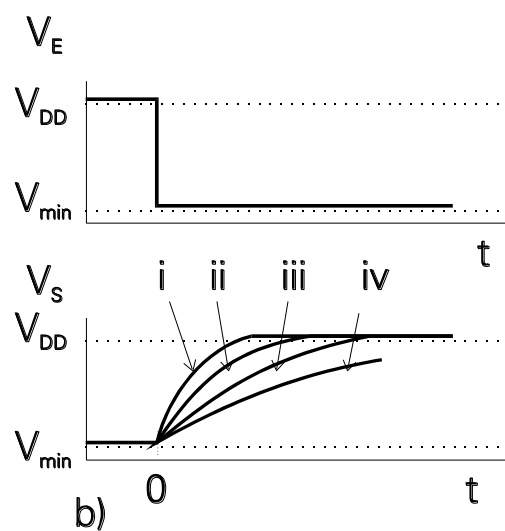
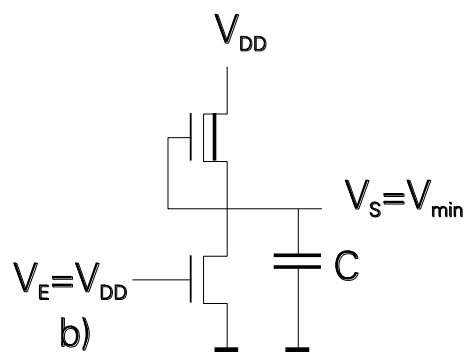


Fig. 4.8 Carga del capacitor de salida C.

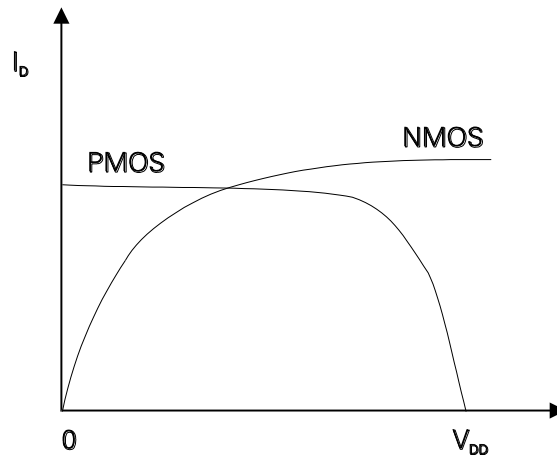


Fig. 4.9 Característica de salida de un NMOS con una carga complementaria PMOS.

Para este inversor solamente un solo transistor conduce en cada estado, con lo cual en los estados finales no hay corriente. La corriente fluye solamente en el caso de las transiciones, o sea de las cargas y descargas de los capacitores. Esto es un gran ahorro de energía en el régimen estático, donde las corrientes de fuga prevalecen.

El consumo depende linealmente de la frecuencia de conmutación al aumentar el tiempo de la conducción simultánea de ambos transistores. También se aumenta por la carga y descarga del capacitor de salida y es un factor dominante, que aumenta con el cuadrado del voltaje del capacitor.

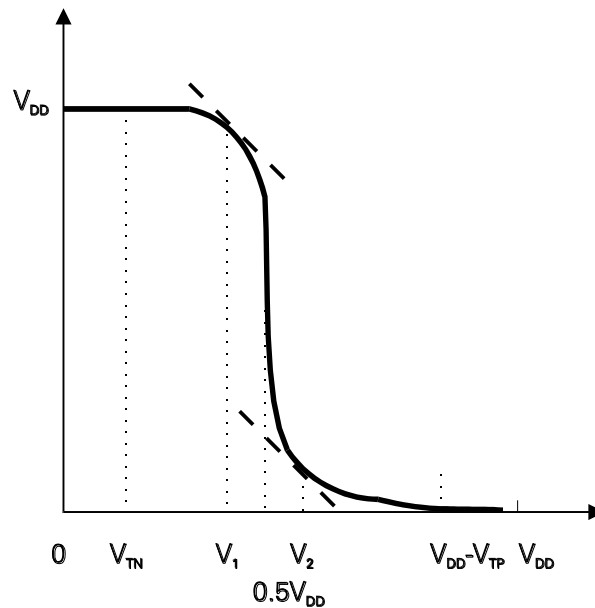


Fig. 4.10 Característica transferencial de inversor CMOS donde se muestra el voltaje de transición cerca del centro y los márgenes de ruido que se amplían.

La característica transferencial en este caso es bastante simétrica, Fig. 4.10. Los niveles lógicos del “0” y del “1” no dependen de la relación de aspecto, correspondiendo al 0 y al V_{DD} , por lo cual es “*ratioless*”. También hay un aumento en los márgenes de ruido de ambos niveles, que es algo muy útil. Finalmente, un resultado de esta forma de conexión es que el voltaje de V_{DD} puede variar en un rango amplio, normalmente entre 3 y 15 V. Es la tecnología en que el inversor trabaja más independiente de los voltajes de alimentación. Los niveles superiores lo dan la inversión de los transistores de campo, o el *punch-through*, y los inferiores la velocidad de conmutación.

Si se calcula que la transición se realice para $V_{DD}/2$, donde ambos transistores se encuentran en saturación resulta:

$$\mu_n C_0 \frac{Z_n}{L_n} \frac{(V_E - V_{TN})^2}{2} = \mu_p C_0 \frac{Z_p}{L_p} \frac{(V_E - V_{TP})^2}{2}, \quad (4.10)$$

o sea

$$\frac{\mu_n Z_n L_p}{\mu_p Z_p L_n} \frac{(V_E - V_{TN})^2}{(V_E - V_{TP})^2} = 1. \quad (4.11)$$

Si $V_{TN} \approx |V_{TP}|$, y $L_p = L_n = L_{min}$, entonces

$$\mu_n Z_n = \mu_p Z_p. \quad (4.12)$$

Como la movilidad para el canal N es aproximadamente el doble que la movilidad en el canal P, entonces el ancho del transistor de carga será el doble del de control, con iguales longitudes de canal. Esta es la condición óptima desde el punto de vista de la inmunidad de ruido y la velocidad de conmutación. Pero si estas características no son críticas, se puede obtener transistores de dimensión mínima con resultados de ahorro de espacio y de capacitancia de entrada del inversor. Cambiando la relación de aspecto se varía el voltaje de conmutación entre estados.

El tiempo de demora o retardo del inversor se puede calcular aproximadamente. La corriente de descarga del capacitor C_L pasa a través del transistor de control, variando esta corriente desde 0 hasta I_{sat} , por lo cual

$$C_L V_{DD} = t_f I_{sat} / 2 = t_f \left[\mu_n C_0 \frac{Z_n}{L_n} (V_{DD} - V_{TN})^2 / 2 \right] / 2 \quad (4.13)$$

de donde se obtiene el tiempo de descarga (caída) t_f según:

$$t_f = \frac{4 C_L V_{DD} L_n}{\mu_n C_0 Z_n (V_{DD} - V_{TN})^2}. \quad (4.14)$$

Para el tiempo de carga, cuando la salida va de 0 a V_{DD} , que también se conoce como tiempo de subida (“*rise time*”) t_r , se obtiene de forma similar la expresión:

$$t_r = \frac{4C_L V_{DD} L_p}{\mu_p C_0 Z_p (V_{DD} - |V_{TP}|)^2} . \quad (4.15)$$

Si se igualan $\mu_n Z_n / L_n = \mu_p Z_p / L_p$, entonces ambos tiempos serán iguales, considerando el retardo total t_d como el promedio de ambos resulta que

$$t_d = \frac{t_r + t_f}{2} . \quad (4.16)$$

Si se quiere optimizar la demora se puede calcular que la misma es mínima cuando $Z_p = \sqrt{\mu_n / \mu_p} Z_n$.

Capacitancia de carga. Las capacitancias de carga de un inversor CMOS se muestran en la Fig. 4.11. Las mismas son: C_{gn} - la capacitancia de la compuerta del NMOS; C_{gp} - la capacitancia de la compuerta del PMOS; C_m - la capacitancia de Miller (retroalimentación). C_m esta compuesta por las capacitancias parásitas entre las compuertas de ambos y la región del drenaje de ellos mismos que forma la llamada penetración lateral de las junturas x_{jl} , o sea

$$C_m = Z_n x_{jl} C_{gdn} + Z_p x_{jl} C_{gdp} ; \quad (4.17)$$

también se tiene que añadir la capacitancia de interconexión C_s , la cual en los dispositivos reales es de gran importancia.

Es necesario considerar que la capacitancia de Miller se carga desde $-V_{DD}$ hasta $+V_{DD}$, por lo cual su valor efectivo es el doble de C_m , tal y como se presenta en todos los demás dispositivos electrónicos de vacío o semiconductores.

Como resultado se puede resumir que la capacitancia total de carga C_L será igual a:

$$C_L = C_{gn} + C_{gp} + 2C_m + C_s . \quad (4.18)$$

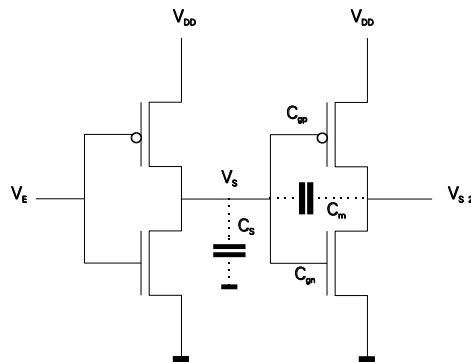


Fig. 4.11 Capacitancias de carga de un inversor CMOS además de las de compuerta de cada transistor.

4.2 LA COMPUERTA DE TRANSMISIÓN

Las compuertas de transmisión son elementos básicos de los circuitos, que permiten el paso o no de una señal a lo largo de una línea de conexión. Si se utiliza un NMOS en esa función, Fig. 4.12a, al ser el pulso f igual a V_{DD} es posible transmitir la información que está en V_E a la salida V_S , la cual se guarda en el capacitor C_L . Por supuesto que cuando $V_E = V_{DD}$ entonces la salida solamente llegará hasta $V_{DD} - V_{TN}$. Otro problema es el efecto de sustrato, el cual hace que el V_{TN} se incremente en forma considerable. Utilizando un par de transistores complementarios se pueden superar algunos de estos problemas. En la Fig. 4.12b se muestra la compuerta de transmisión resultante en este caso, donde se conectan conjuntamente las regiones de fuente y de drenaje, quedando las compuertas conectadas a dos pulsos complementarios entre sí, de modo que ambos transistores se encuentren conduciendo a un mismo tiempo. A diferencia del caso anterior el transistor PMOS se encuentra en una zona de conducción permitiendo que la salida alcance el valor de V_{DD} . Durante la descarga del capacitor se produce el proceso contrario y se puede descargar completamente. Esta es la gran ventaja de esta estructura, que es ampliamente utilizada en los circuitos CMOS dinámicos, con los cuales incluso se implementan circuitos lógicos simplificados.

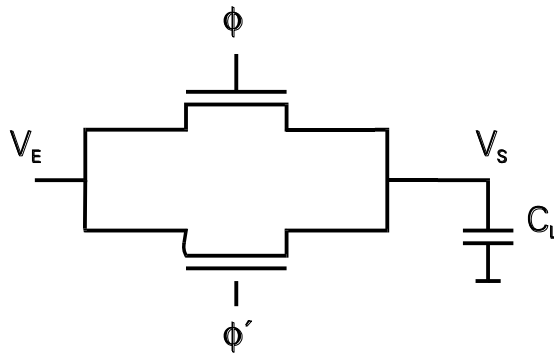


Fig. 4.12 Compuerta de transmisión con CMOS

4.3 LAS COMPUERTAS LÓGICAS

Fig. 4.13 Símbolo y esquema en las tecnologías CMOS y NMOS de las dos compuertas básicas de mayor uso: a) NAND; b) NOR.

Fig. 4.14 Esquema de la topología de las compuertas NAND y NOR con tecnologías CMOS y NMOS.

Tomando el inversor como base se pueden obtener diferentes compuertas lógicas como las NAND o las NOR, añadiendo transistores. En la Fig. 4.13 se muestran el caso a) de compuertas NAND de dos entradas implementadas en tecnología CMOS y NMOS, también se muestra el símbolo lógico de las NAND. Dependiendo del número de transistores que se conecten en serie en la parte de control, será el número de entradas que tenga el circuito. La variación del estado de la salida se logra cuando todos los transistores de entrada están conduciendo. Según la tecnología la carga varia, si se usa la NMOS se requiere solamente un transistor de carga, normalmente del tipo de empobrecimiento. Para el CMOS la carga la constituyen un transistor complementario por cada transistor de entrada. Aunque se aumenta el área, tiene las ventajas de los CMOS antes consideradas.

Existe un problema común en ambas tecnologías referente al efecto de substrato en los transistores de control. Resulta que al estar conectados encadena los diferentes transistores de control, y tener un potencial de substrato común, el voltaje V_{BS} de cada transistor es diferente, al ser diferente la caída de voltaje en cada transistor. Así el efecto de substrato es diferente en cada transistor y va aumentando a medida que se incrementa en número de transistores de la tierra hacia arriba. Normalmente el número máximo es de 8, aunque se trata de no subir de 4. Es indispensable calcular este efecto en las compuertas NAND.

Como se ve, el circuito mas simple que se obtiene es la NAND, si se quiere obtener una compuerta AND se requiere añadir un inversor a la salida.

Fig.4.15 Compuerta OR-exclusivo (XOR), símbolo y ecuación lógica: a) CMOS convencional; b) CMOS simplificado; c) CMOS con compuertas de transmisión.

En el caso de los circuitos lógicos NOR, Fig. 4.1b, la situación es que los transistores de control se conectan en paralelo, siendo el estado diferenciado cuando todas las entradas están en L, con lo cual la salida está en H. La carga para el NMOS sigue siendo un transistor, aunque debe tenerse en cuenta la corriente en paralelo por los otros transistores. Para el CMOS ahora los transistores de carga están en serie. El número mayor de entradas que se usa en la práctica es de 8, aunque no es recomendable.

El problema del efecto del substrato no está presente en este caso. Al igual que en las NAND, para obtener la función OR se requiere poner un inversor a la salida.

Los tiempos de subida y caída del voltaje de salida dependerán de las relaciones de aspecto de los transistores y entre los de control y las cargas, siendo el principal objetivo en un diseño. Las

compuertas NOR y NAND tienen parámetros de tiempo que difieren entre sí, los criterios de optimización son distintos en ambos casos.

En la Fig. 4.14 se muestra en forma esquemática un ejemplo de las topologías de estas compuertas.

Al utilizar las estructuras CMOS no siempre hay que utilizar una solución ortodoxa cuando se conforma un circuito de mayor complejidad. Conociendo las funciones finales es posible en ocasiones reducir el número de transistores. Un ejemplo se tiene en el caso de la compuerta llamada OR-exclusivo (XOR) que se muestra en la Fig. 4.15. De los 10 transistores iniciales requeridos se puede pasar a 8, y con una topología mas favorable. También se muestra el caso de utilizar lógica con compuertas de transmisión, donde el número de transistores se reduce a 4.

4.4 EL EFECTO “LATCH-UP”

En las estructuras CMOS ocurre un fenómeno que limita su funcionamiento y se conoce en la literatura como “*latch-up*”. En la estructura típica de un inversor CMOS, Fig. 4.16, hay una conexión de las zonas P y N, tanto los cajones como las zonas de S y D, de forma tal que se forma una conexión entre dos transistores bipolares que conforman un tiristor. El resultado es que si se dan las condiciones propicias se produce el disparo del tiristor equivalente, por lo cual se cortocircuitan los transistores de carga y control, interrumpiendo el funcionamiento del inversor. Este fenómeno ha sido uno de los mas estudiados en los últimos 15 años, y se han propuesto gran cantidad de soluciones para reducirlo o eliminarlo. De hecho, en los circuitos comerciales se ha logrado eliminar. Para ello algunas de las reglas elementales son: no poner ninguna de las junturas en directa en ningún momento de la operación del CMOS; que las resistencias de los cajones (pozos) y de las zonas difundidas sean lo menor posible; alejar los transistores entre sí, aumentar el ancho de las bases equivalentes, para reducir la ganancia de los transistores parásitos que se forman.

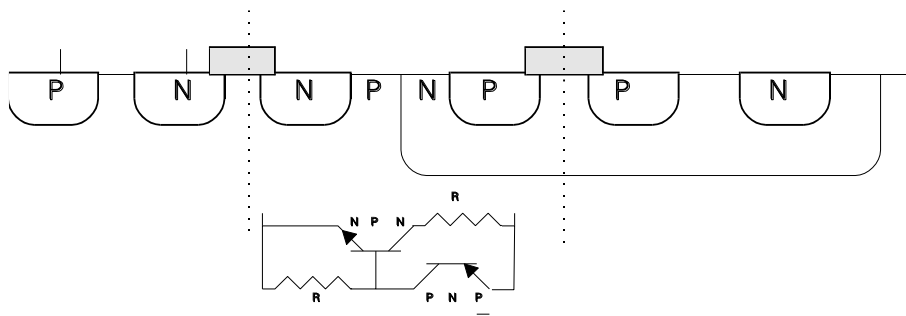


Fig. 4.16 Efecto "latch-up". Esquema del tiristor NPNP que se forma entre los drenajes, el pozo y el sustrato en una estructura CMOS. De acuerdo con los potenciales el tiristor puede dispararse.

4.5 LAS MEMORIAS RAM

Las memorias de acceso aleatorio (“*Random Access Memory*”-RAM) son la fuerza motriz de la tecnología de la microelectrónica por dos causas:

1) se requieren en cantidades muy grandes, obligando a aumentar constantemente el número de bits por chip de silicio, o sea la integración;

2) para lograr ese mayor grado de integración con rendimientos altos se requiere de un elevado desarrollo tecnológico. Por estas causas las memorias RAM son el acicate de la industria microelectrónica permitiéndole duplicar el nivel de integración aproximadamente cada dos años desde 1970. En ese año estaba en producción la memoria de 1 Kbit y ya en 1990 se producía la de 1 Mbit. En la actualidad se cuenta con memorias de varios Gbit y continúa creciendo.

Debe recordarse que por definición de la lógica binaria cuando se habla de 1 Kbit corresponde a 1024 bits, y 1 Mbit son 1048576 bits.

Por supuesto que, aunque los costos por bit disminuyen drásticamente con este aumento de la integración, los costos de producción han subido astronómicamente en los últimos años haciéndose costosa la producción solamente de muchos millones de unidades al mes.

Se diferencian dos tipos de memorias RAM de acuerdo con la forma en que la celda de memoria guarda la información: las RAM dinámicas y la RAM estáticas.

En el caso de la memoria RAM dinámicas la información se guarda en forma de la carga de un capacitor MOS, por lo cual se requiere de un refrescamiento de la información para mantenerla vigente. Para la memoria RAM estática la celda se conforma por un circuito biestable ("*flip-flop*"), manteniéndose la información mientras esté energizado el circuito.

4.5.1 LA CELDA DE LA MEMORIA RAM DINÁMICA

La celda de la memoria RAM dinámica se conformó inicialmente por 3 transistores, después por dos transistores y finalmente se llegó a un transistor y un capacitor, ambos de tecnología MOS. Esta última estructura, en diferentes variantes de realización es la que permitió el crecimiento impetuoso de la integración y se ha utilizado en los últimos 20 años.

En la Fig. 4.17 se muestra un esquema clásico de este tipo de celda de 1 transistor y 1 capacitor, con un solo nivel de polisilicio. Esta estructura consta de un capacitor C_A , que se conecta a una línea de bit a través de un transistor MOS controlado por la llamada línea de palabra.

La escritura de información se realiza poniendo la información en la línea de bit, un nivel de voltaje alto o bajo según se quiera escribir un 1 o un 0. Este dato es común a todas las celdas conectadas a esa línea de bit. La celda precisa se selecciona a través de la línea de palabra correspondiente, escribiéndose la información solamente en la celda (bit) que está en la intersección de la línea de bit y la línea de palabra seleccionadas.

La lectura de la información se realiza precargando la línea de bit a un valor predeterminado, luego al conectar la celda seleccionada, como se hizo anteriormente, este valor varía según la carga en el capacitor corresponda a un 0 o a un 1. Un amplificador de sensado amplifica esta información que en varias etapas se pone en la salida del circuito. Como se ve una memoria requiere, además de las celdas de memorias (bits), de los amplificadores de sensado y de los decodificadores para la selección de la celda correspondiente.

La dimensión del capacitor MOS es de mayor importancia, pues de ella depende la cantidad de carga acumulada. En la Fig. 4.17 el capacitor es de 0.2 pF. La posibilidad de reducir este capacitor está en que siempre debe ser mayor que la capacitancia parásita de la línea, para diferenciar la señal del ruido.

Durante los últimos años hay muchas proposiciones de reducir el área que ocupa este capacitor, o sea su topología, sin reducir considerablemente el valor de la capacitancia. También se ha logrado simplificar la estructura del transistor utilizando dos niveles de polisilicio.

Un elemento importante en la operación de estas celdas es que la carga acumulada en el capacitor tiende a recombinarse en el tiempo, por lo que la información se puede perder. Por este motivo se requiere que se refresque la información en cada celda cada cierto tiempo, este actualmente es de 2 ms. Para simplificar este proceso el refrescamiento se realiza en todos los bits de una columna, línea de palabra, a un mismo tiempo.

Hay otro evento en las memorias RAM dinámicas que se presenta cuando una partícula alfa atraviesa la zona del capacitor, generando una carga que puede ser equivalente a la carga de la información escrita, por lo cual se puede falsear la información en el momento de la lectura. Como este error no es permanente se conoce como error suave (“*soft error*”). Resultó que los propios materiales del encapsulado podían tener residuos radiactivos que generaran estas radiaciones alfa. Así, se realizó un gran trabajo en la selección de los materiales para el encapsulado de las memorias, además de cubrir los dispositivos con una capa que pudiera absorber estas radiaciones.

La estructura y características de las celdas de las memorias RAM siguen siendo objeto de un estudio intenso como puede verse en la literatura especializada.

Fig. 4.17 Celda de memoria RAM dinámica con un solo nivel de polisilicio.

a) Diagrama eléctrico de la celda; b) Topología de una celda; c) Corte transversal de la celda donde se ven el transistor T_A , el capacitor C_A y la línea de bit.

4.5.2 LA CELDA DE LA MEMORIA RAM ESTÁTICA

Fig. 4.18 Celda de memoria RAM estática: A) Celda de 6 transistores; B) Diagrama de la topología de esta celda; C) Amplificador diferencial de sensado.

Las memorias RAM estáticas se basan en guardar la información en un circuito del tipo biestable (“*flip-flop*”) consistente en dos inversores interconectados salidas con entradas. En la Fig. 4.18 se muestra la celda realizada con un biestable CMOS, formado por 4 transistores y dos transistores de acceso, que comunican la celda con la línea de bit. O sea la celda consta de 6 transistores.

Para escribir la información en esta celda se pone la misma en la línea de BIT a un lado y el valor negado en el otro. Cuando se selecciona la línea de palabra (WORD) entonces se abren los transistores de acceso y se fuerza a que el biestable tome el estado que se impone externamente.

Para la lectura solamente se precargan las líneas de bit a un valor preestablecido y cuando se selecciona una celda las líneas de bit toman el valor de los nodos de salida del biestable.

En la Fig. 4.18b se muestra el diagrama de la topología de este tipo de celdas.

En este caso la información se mantiene estable todo el tiempo en que la alimentación este conectada, no es necesario reescribir o refrescar la información.

Las memorias que mantienen la información mientras están conectadas a la alimentación se llaman volátiles, mientras que las que conservan la información aún sin alimentación se llaman no-volátiles.

Para realizar la lectura se utiliza el llamado amplificador de sensado, que permite pasar de la pequeña señal de la celda a una señal detectable externamente. En la Fig. 4.18c se muestra un amplificador de sensado, que es un amplificador diferencial con CMOS, y un transistor N3 que funciona como fuente de corriente, donde $I_d = I_{d1} + I_{d2}$. Si la señal en BIT baja ligeramente, I_{d1} bajará en D. En esa misma cantidad debe subir I_{d2} para mantener el valor total. De esta forma aparece una diferencia de potencial V_0 entre los dos nodos de salida, y esta es la señal útil que detectar. Generalmente las líneas de bit están precargadas a un valor alto, lo que permite acelerar la detección.

Este tipo de circuitos es la base de los amplificadores operaciones que se utilizan en los circuitos CMOS analógicos.

Es posible realizar celdas estáticas mas compactas utilizando cargas resistivas realizadas con una segunda capa de polisilicio de muy alta resistividad, llegando a resistencias de carga de $G\Omega$ y utilizando solo 4 transistores por celda.

4.6 LA CELDA DE LA MEMORIA NO-VOLATIL

Como se señaló las memorias no-volátiles son aquellas en que la información se mantiene aún sin conectar la fuente de alimentación. Las memorias más simples de estas son la de lectura solamente (“*Read Only Memory*” -ROM) que se conforman por una matriz de diodos, que se conectan o no en forma definitiva, dando lugar a los 0 y 1. Son muy baratas y se programan solamente en las fábricas.

Una memoria mas flexible es la ROM que se puede programar por el usuario en su laboratorio y que se conoce como EPROM (*Electrical Programmable ROM*). La celda de esta memoria se muestra en la Fig. 4.19, que consta de una compuerta flotante. El principio de funcionamiento de la celda de esta memoria se basa en aplicar un campo intenso en la compuerta de la celda seleccionada, esto genera electrones calientes que se quedan atrapados en la compuerta flotante intermedia cambiando considerablemente el voltaje umbral de este transistor. Así habrá dos transistores uno con voltaje umbral alto y otro con voltaje umbral bajo, el 1 y el 0.

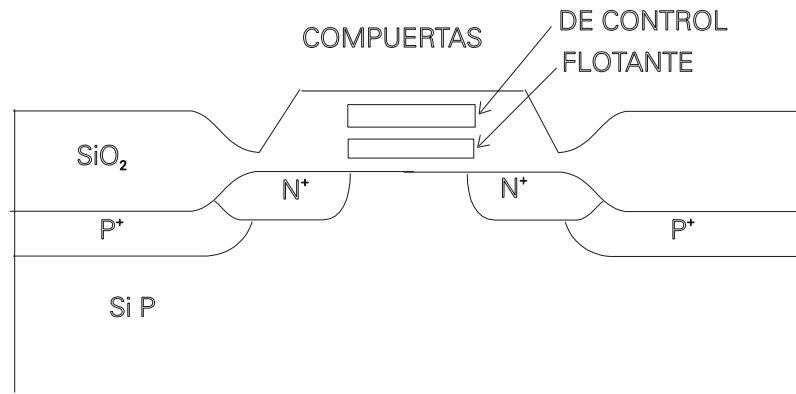


Fig. 4. 19 Estructura de una celda de memoria no volátil - EPROM

La información se conserva por unos cuantos años en este tipo de transistores con fugas muy pequeñas. Para borrar la información se puede acelerar el proceso de recombinación con una radiación intensa de luz ultravioleta, por lo cual estas memorias tienen una ventana de cuarzo.

Los nuevos desarrollos ya ofrecen memorias eléctricamente programables y que se borran también eléctricamente, la EEPROM.