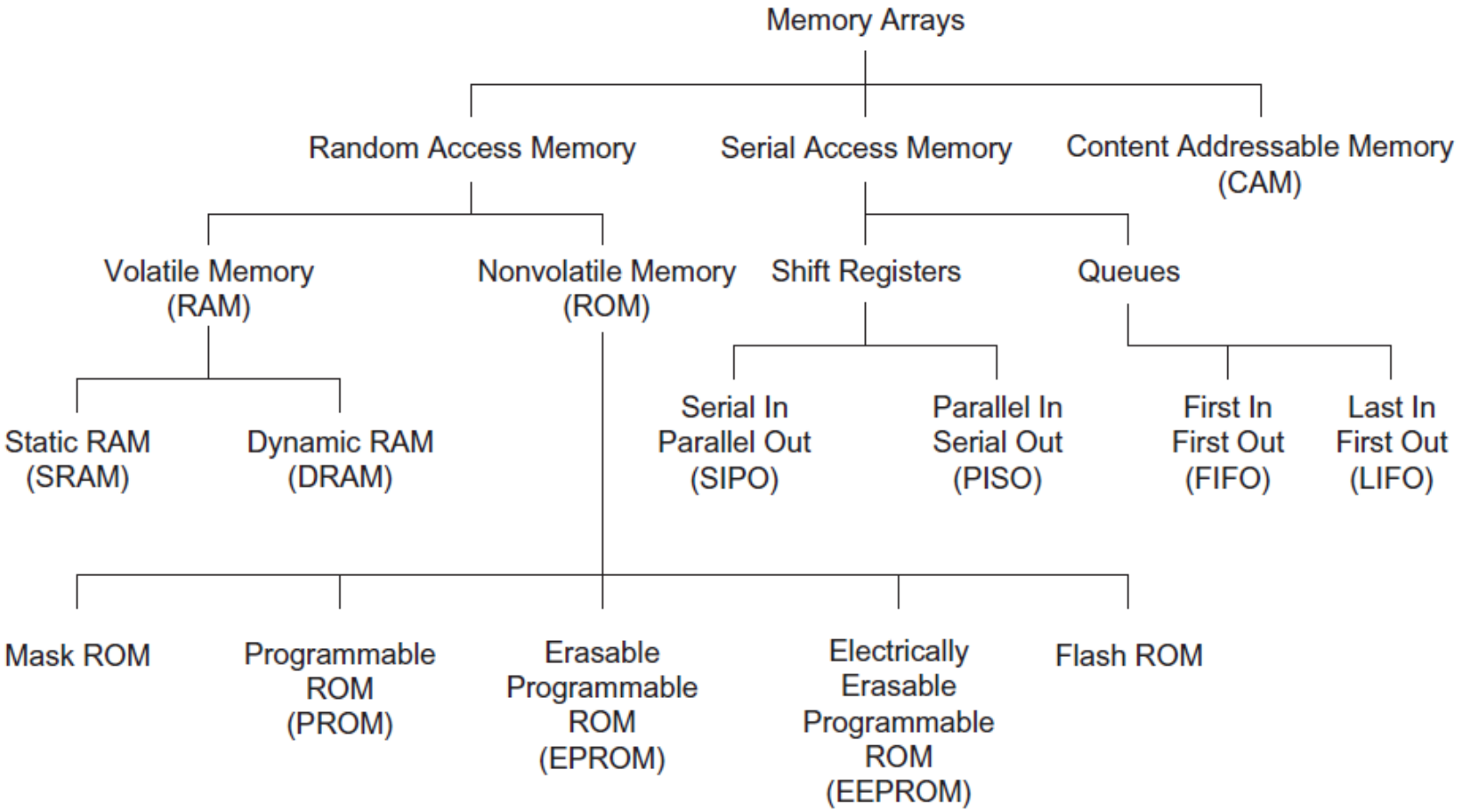
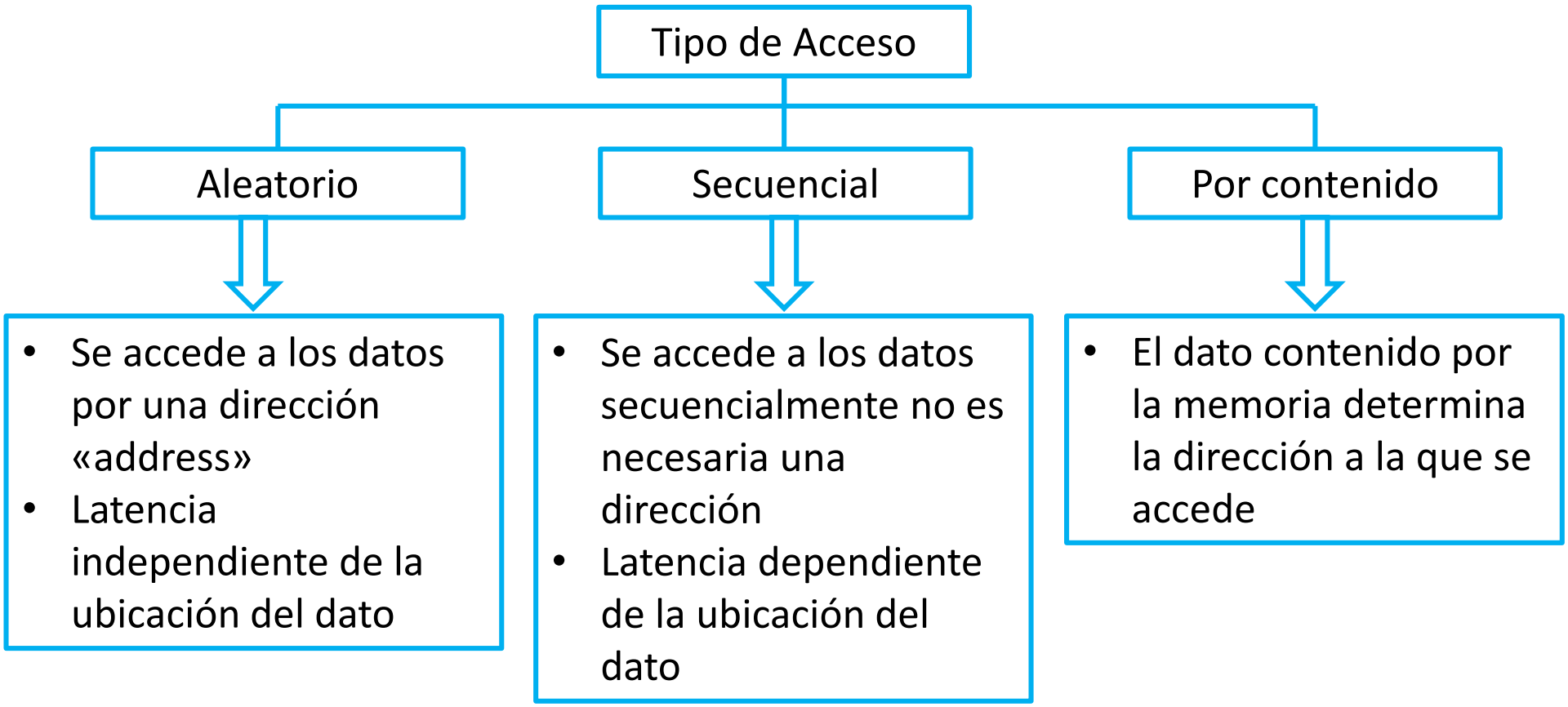
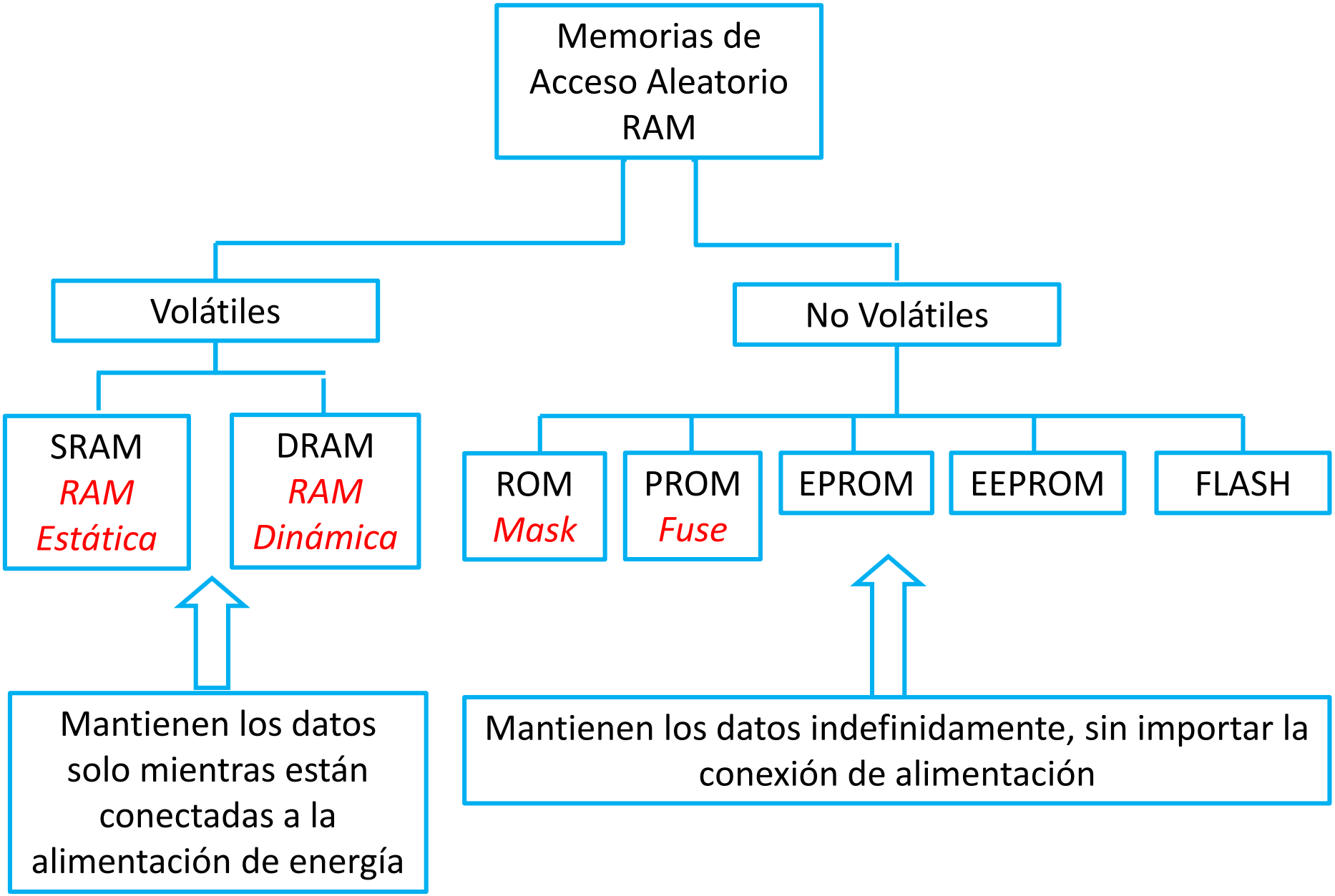
- El funcionamiento de una memoria se evalúa por:
  i. el área del chip por bit de almacenamiento
  ii. el tiempo de acceso R/W
  iii. Durabilidad
  iv. Confiabilidad
  v. Costo

MEMORIAS de SEMICONDUCTORES



MEMORIAS SEMICONDUCTORES




Memorias de Acceso Aleatorio RAM
Volátiles
No Volátiles
SRAM
RAM Estática
DRAM
RAM Dinámica
ROM
Mask
PROM
Fuse
EPROM
EEPROM
FLASH
Mantienen los datos solo mientras están conectadas a la alimentación de energía
Mantienen los datos indefinidamente, sin importar la conexión de alimentación

Arquitecturas de arreglos de memorias de acceso aleatorio RAM´S

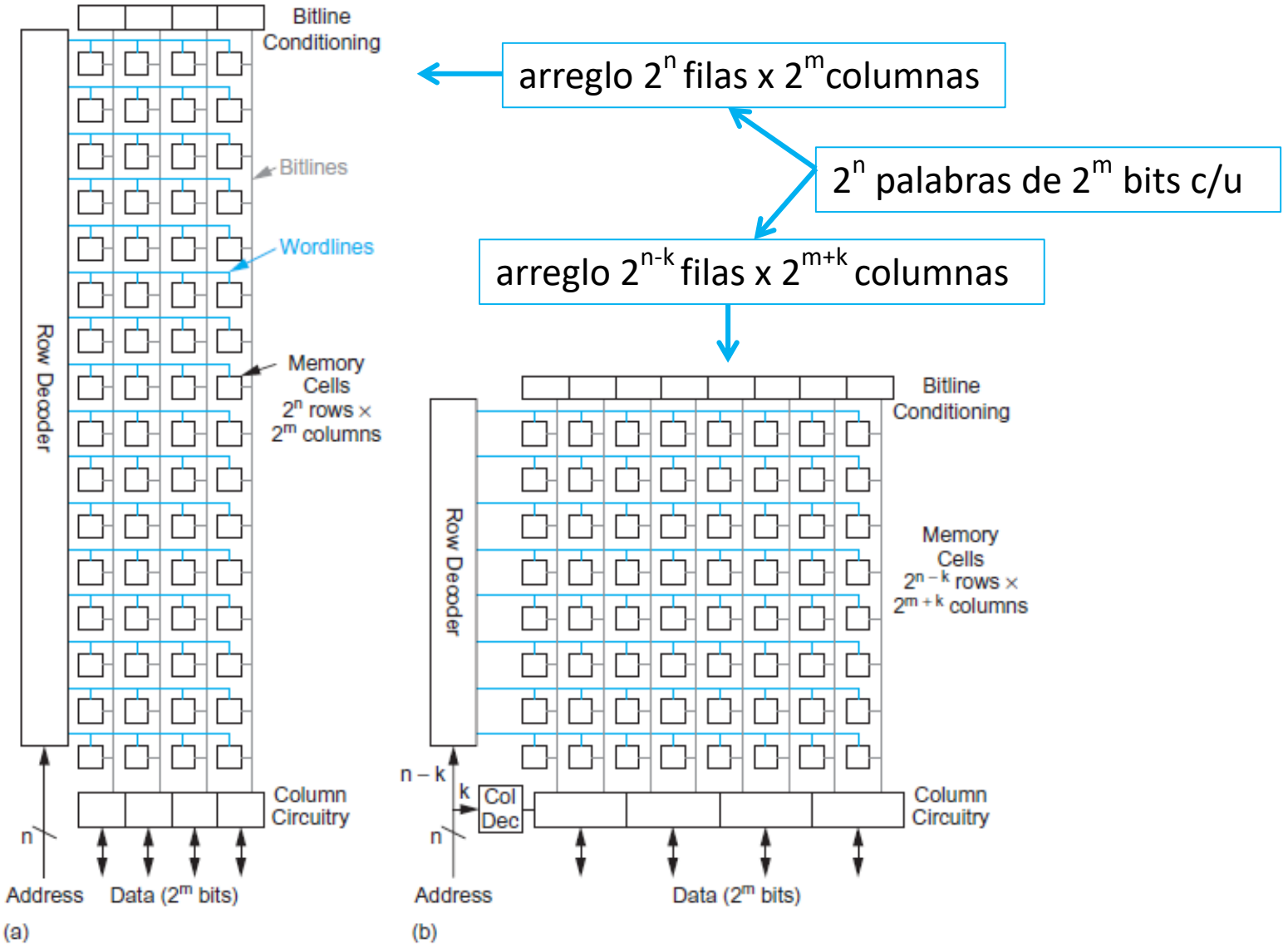


FIGURE 12.2 Memory array architecture

Memorias volátiles de acceso aleatorio

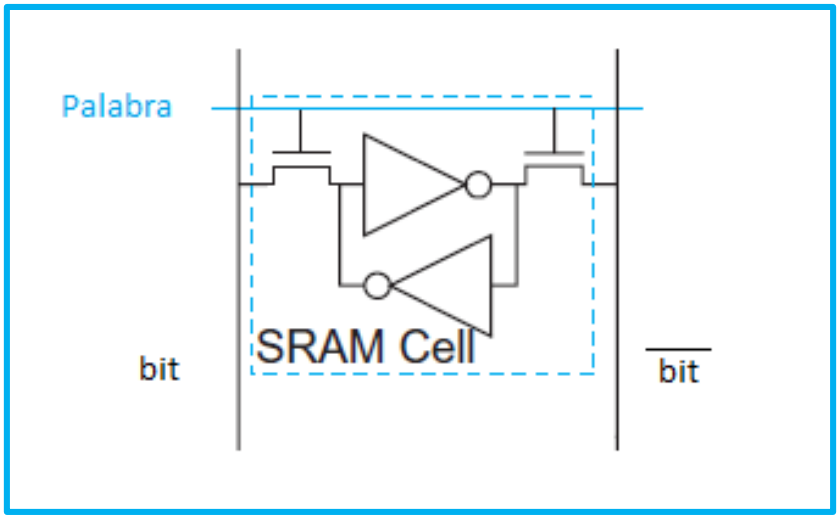
SRAM
RAM Estática

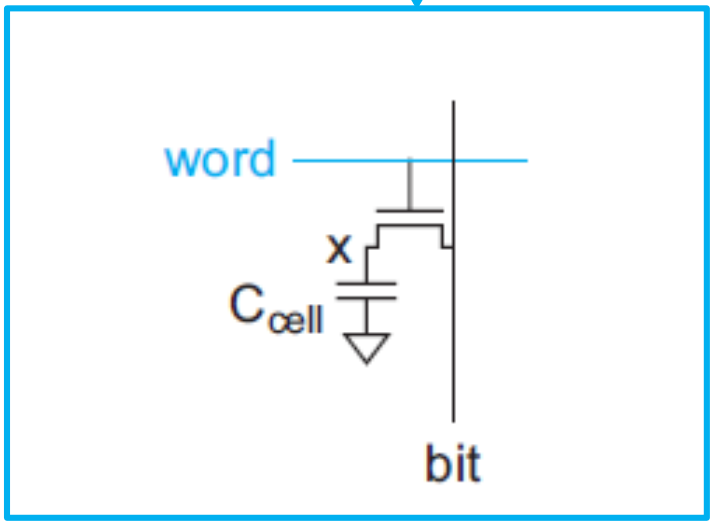
DRAM
RAM Dinámica

- Usan alguna realimentación para mantener el dato

- SRAM son mas rápidas y mas fáciles de usar
- DRAM requieren menos área por bit

- Usan la carga almacenada en un capacitor para mantener el dato

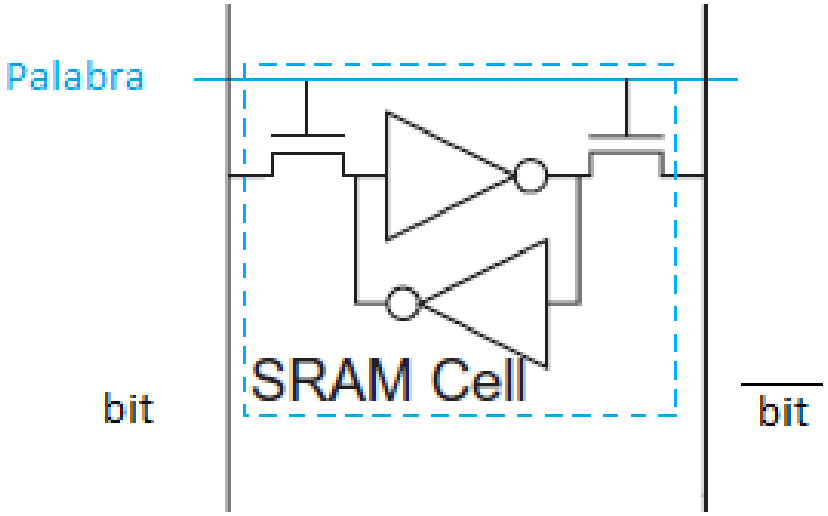


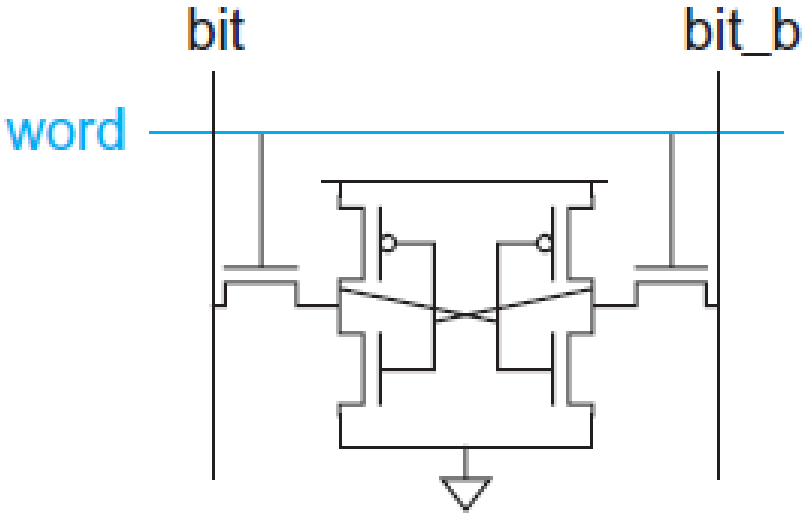


CELDA de SRAM

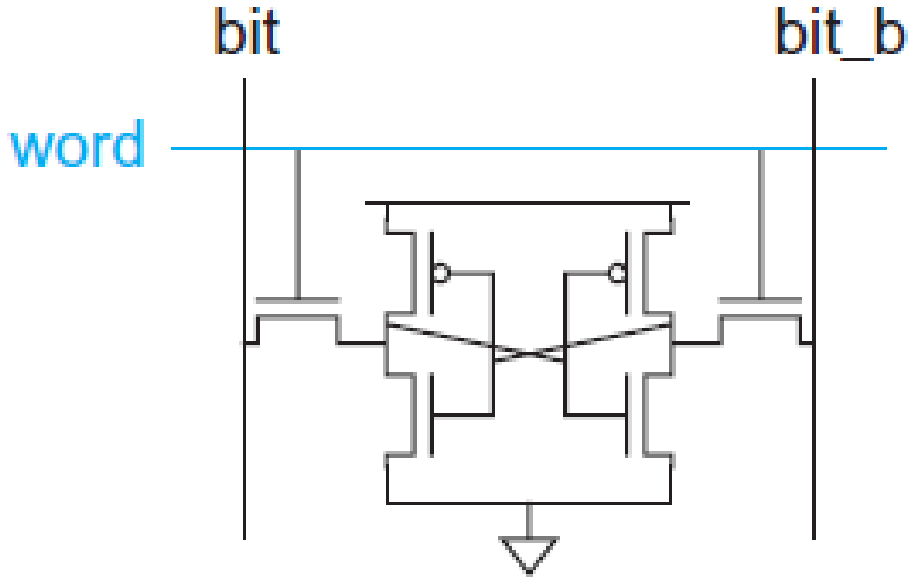
Celda SRAM debe poder:

- Mantener el dato mientras esta aplicada la alimentación
- Leer el dato «READ»
- Escribir el dato «WRITE»





CELDA de SRAM



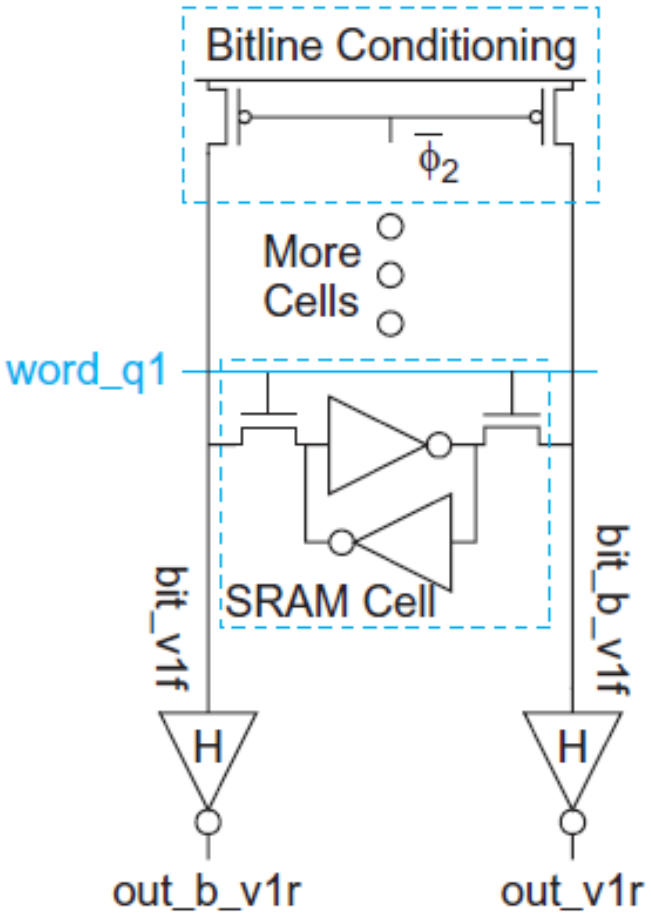
«WRITE»

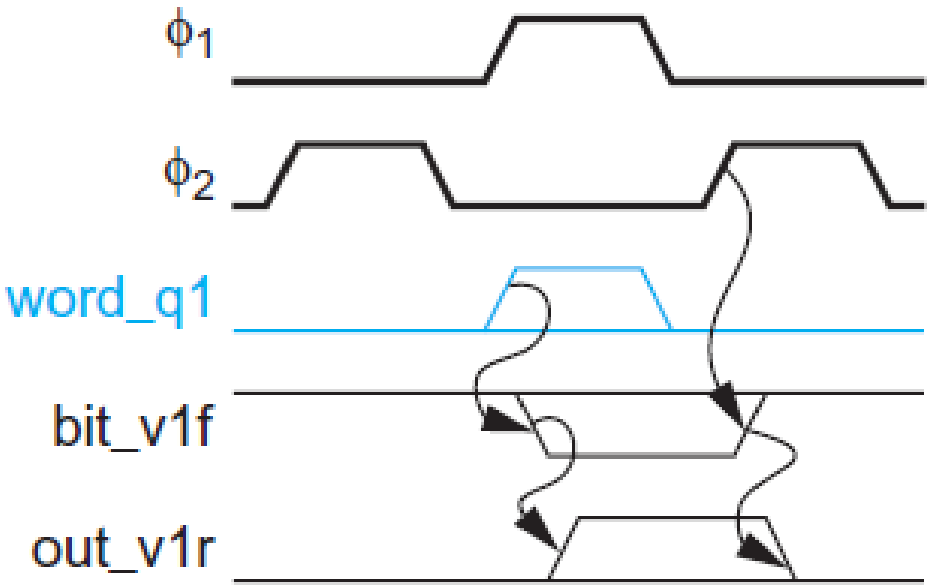
1. Cargar el valor del dato en bit y bit_b
2. Poner en «1» la línea «WORD»

«READ»

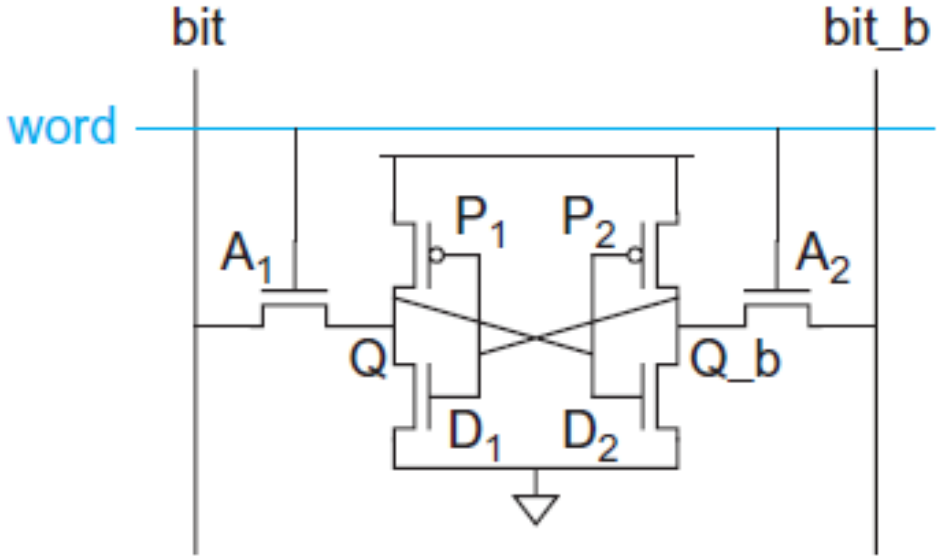
1. Cargar el valor «1» en bit y bit_b
2. Poner en «1» la línea «WORD»
3. La línea que este en «0» (bit o bit_b) impone el valor (pull-down) y se lee el dato

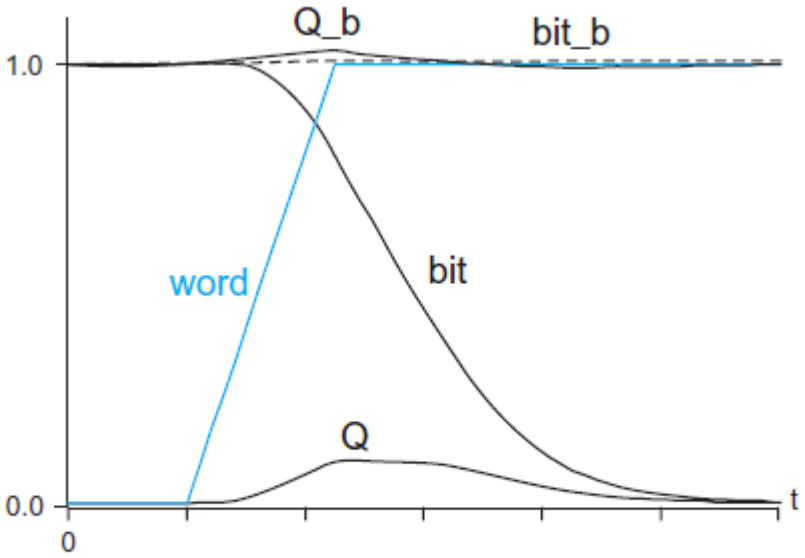
Para las operaciones de «READ» y «WRITE» se utilizan dos fases de reloj (Ø1 y Ø2), en una se realiza la precarga y en la otra se ejecuta la operación de escritura o lectura.





R/W CELDA SRAM





(b)

FIGURE 12.4 Read operation for 6T SRAM cell

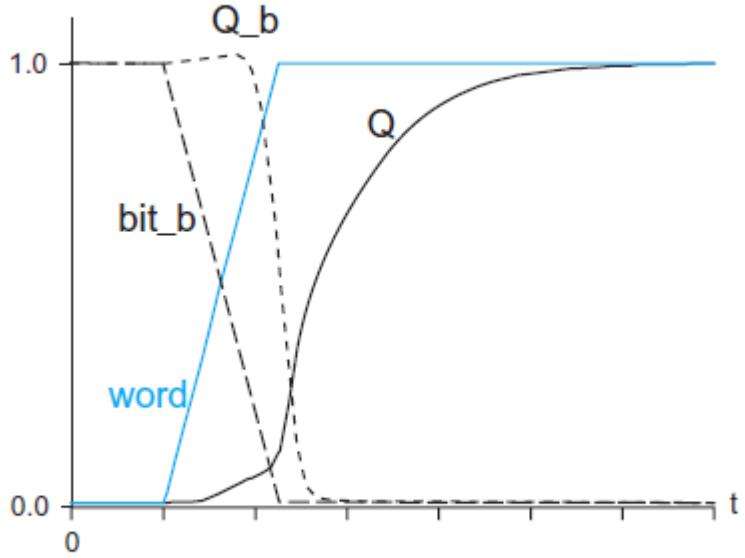
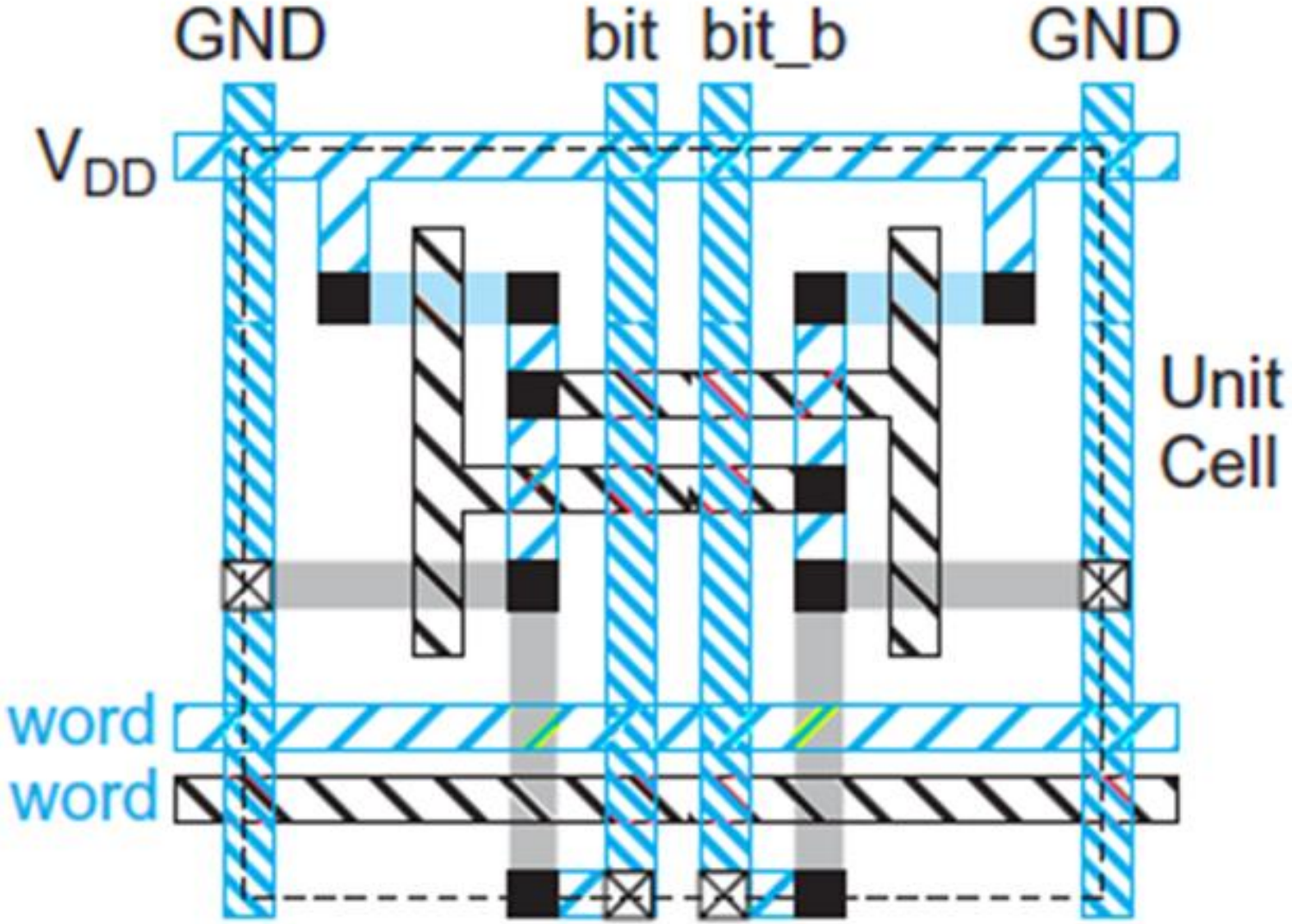
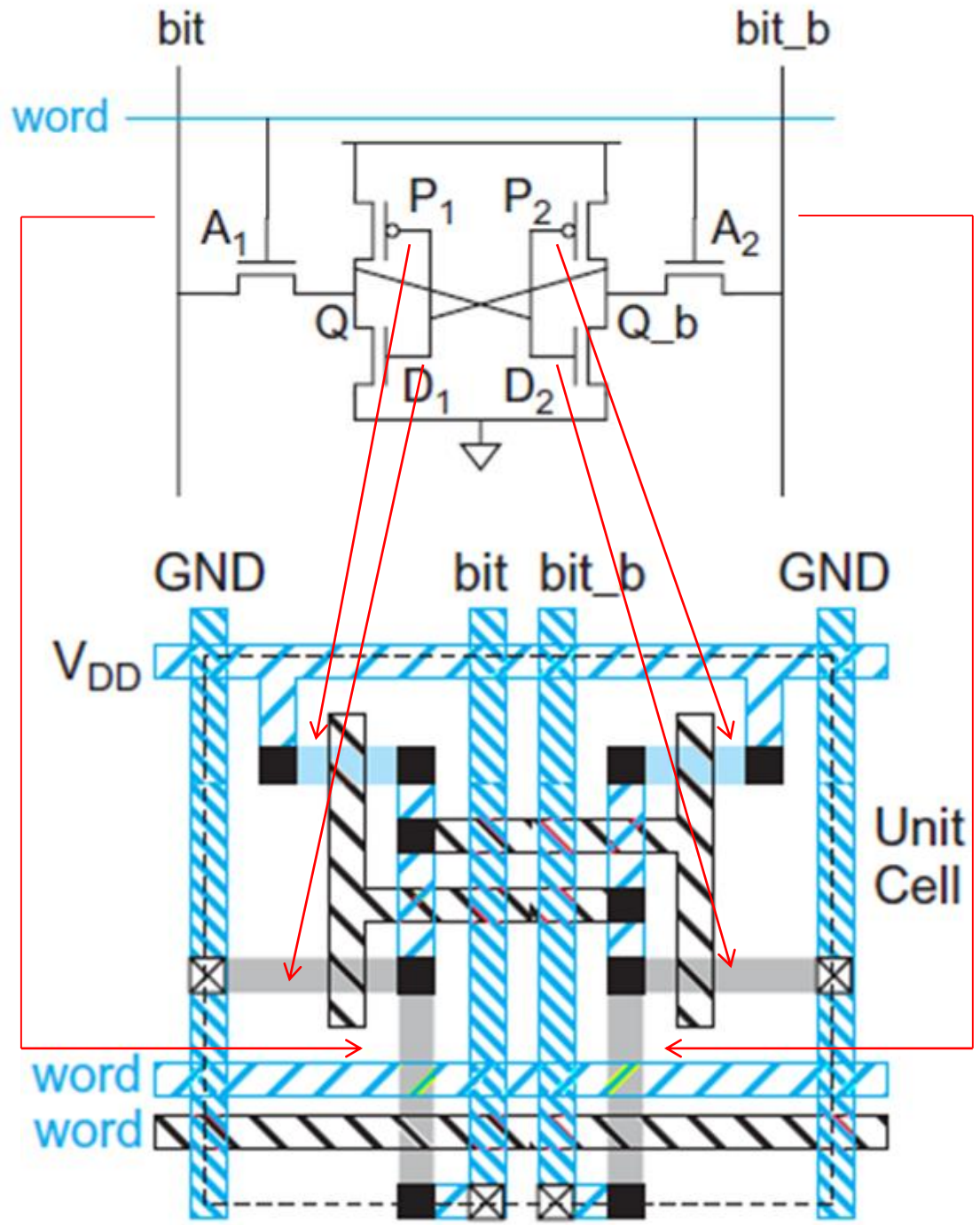


FIGURE 12.6 Write operation for 6T SRAM cell

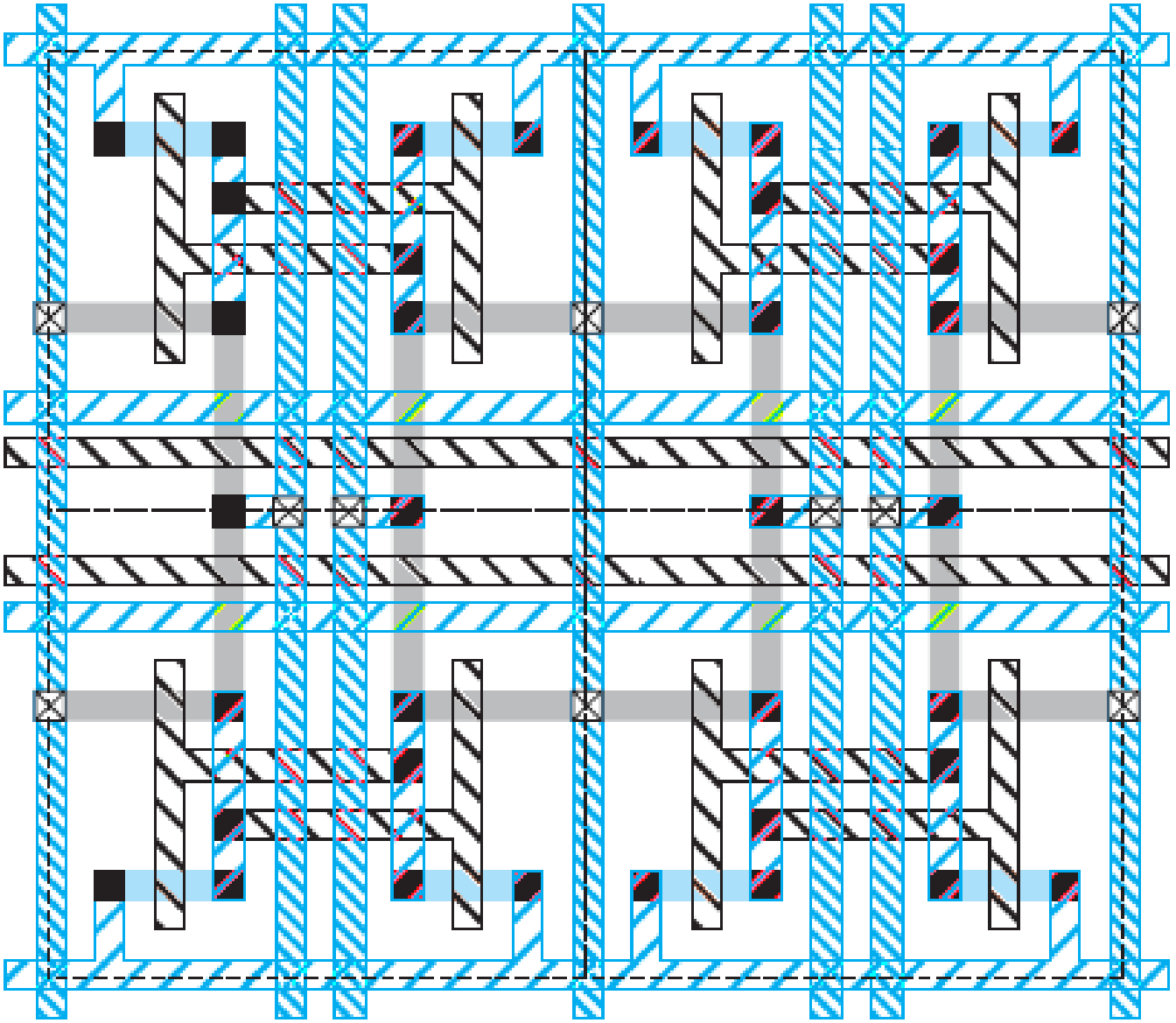
CELDA SRAM

Layout

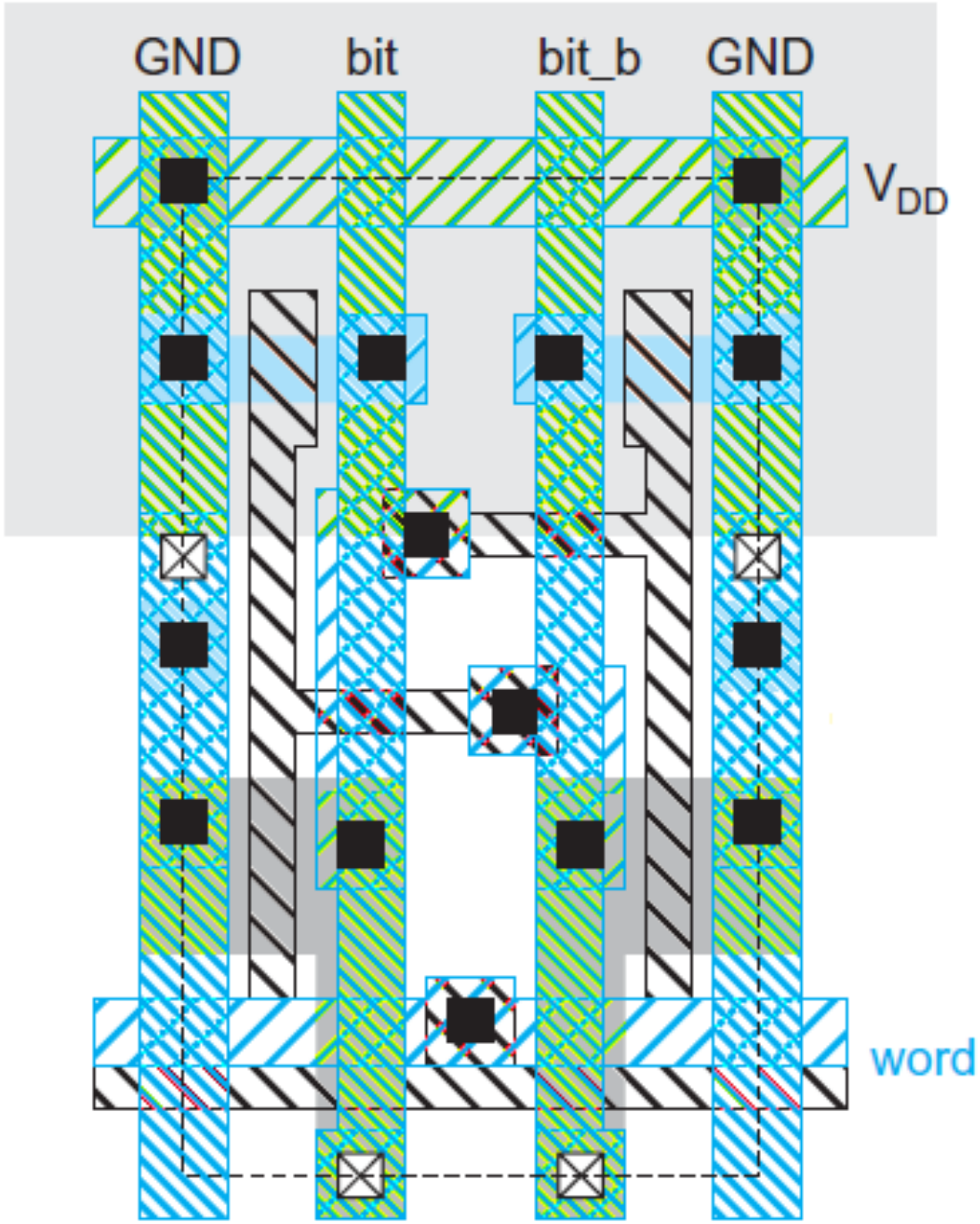



bit
bit_b
word
P1
P2
A1
A2
Q
Q_b
D1
D2
GND
bit
bit_b
GND
VDD
Unit
Cell
word
word

Layout



Layout



El desafío central del diseño de una celda SRAM es:

1. Minimizar su tamaño
2. Asegurar que la estabilidad del estado lógico almacenado sea lo suficientemente débil como para
   a) Cambiar durante la escritura
   b) No variar durante la lectura

CELDA SRAM

Layout

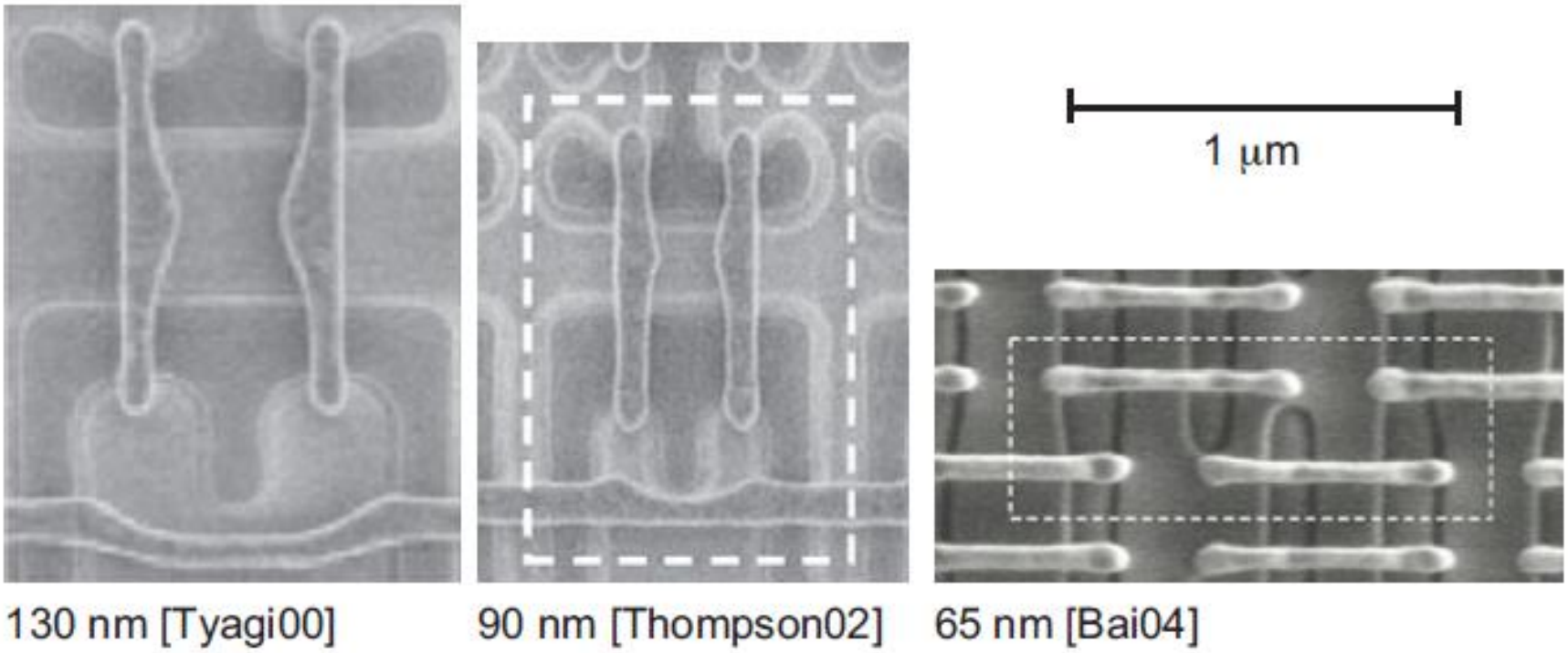
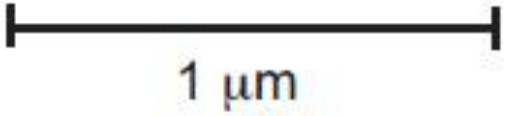
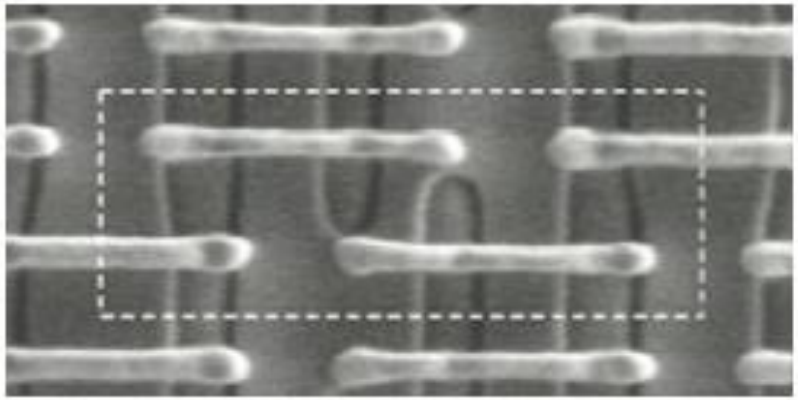


FIGURE 12.16 SRAM scaling (© 2000–2008 IEEE.)

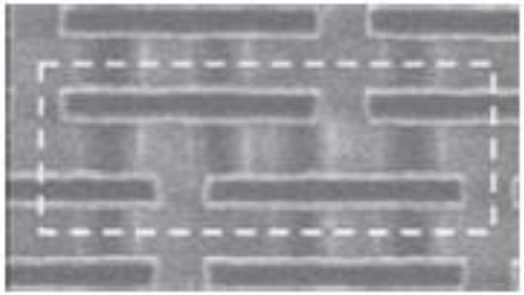
CELDA SRAM

Layout

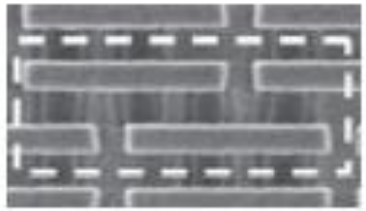




65 nm [Bai04]



45 nm [Mistry07]



32 nm [Natarajan08]

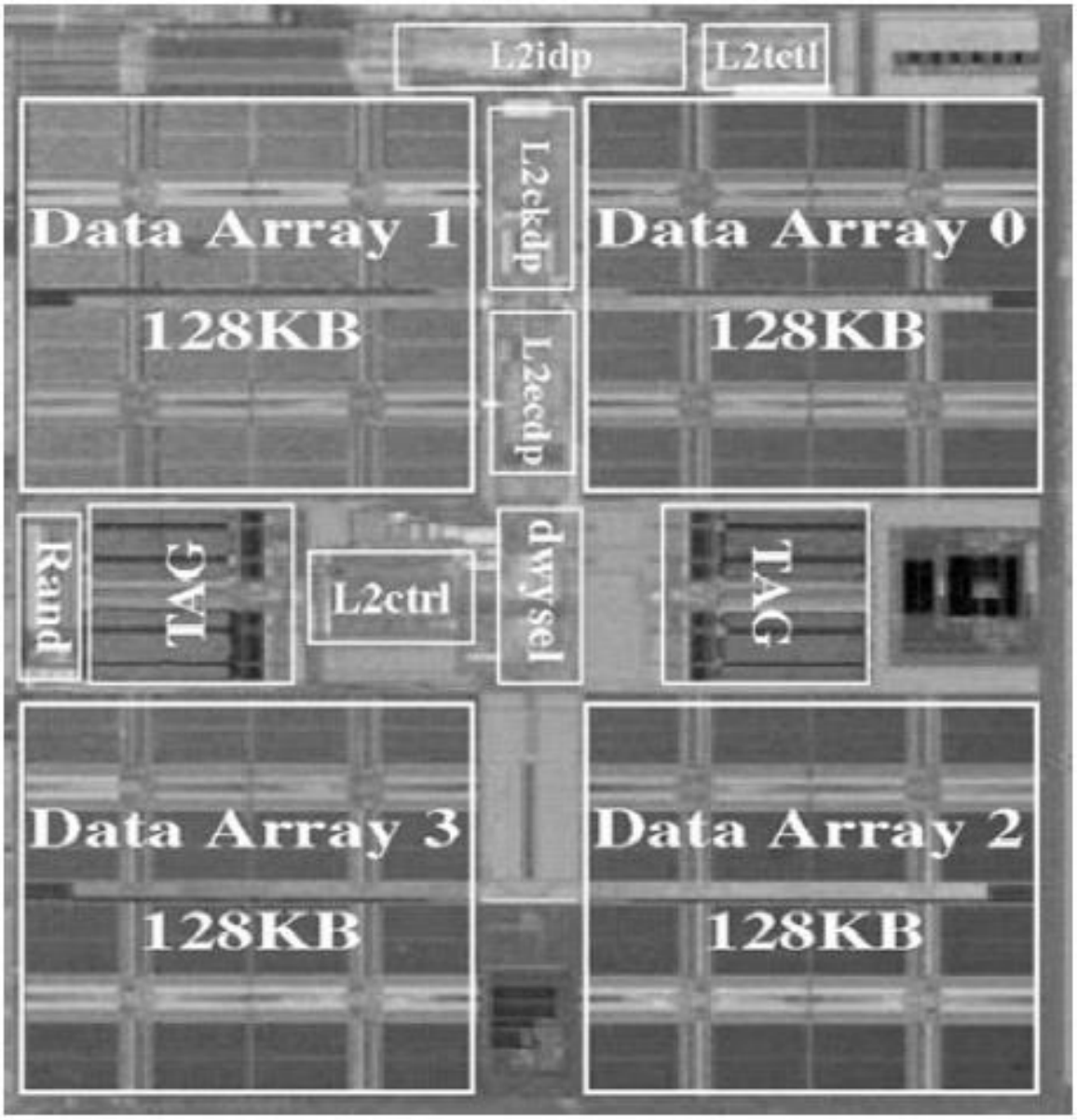
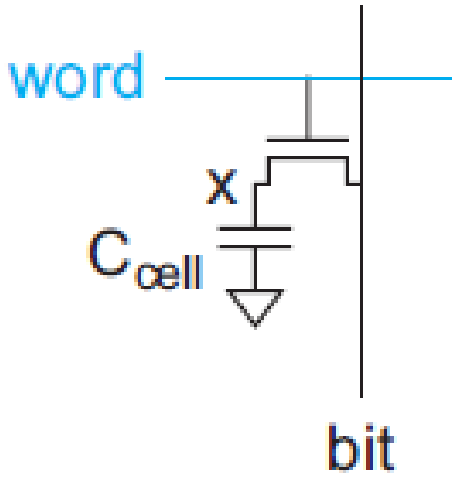


FIGURE 12.36 512 KB cache array (© 2005 IEEE.)

CELDA DRAM



Almacena el dato en la carga de un capacitor

Ocupa menor área que una celda SRAM

La densidad de almacenamiento [bits/cm^2] es 10 a 20 veces mayor que en una SRAM

Debe ser periódicamente leída y reescrita para mantener el contenido «refrescar»

La latencia es mayor que para una SRAM

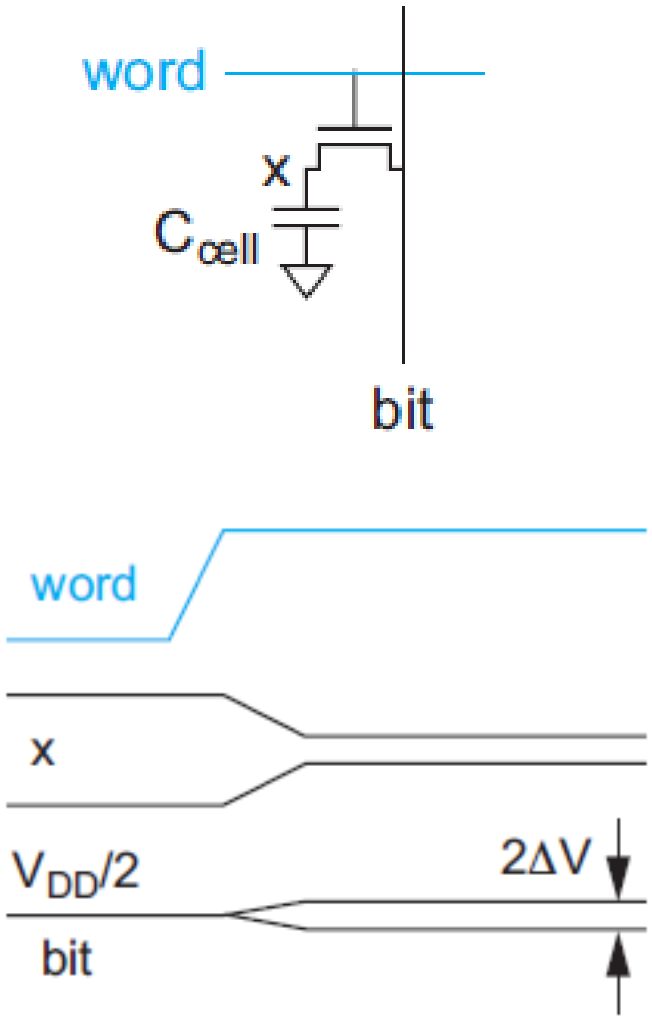


FIGURE 12.41
1T DRAM cell read operation

Para acceder al contenido de la celda hay que conectar el capacitor con la línea «bit»

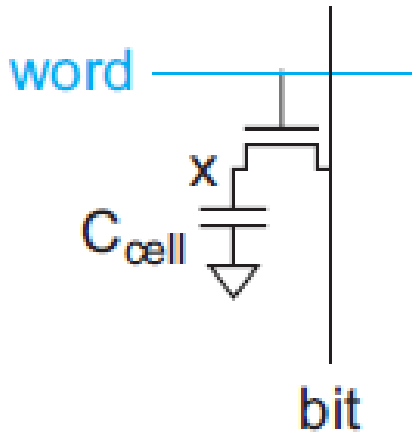
Para ello debemos cerrar el MOS poniendo «word» en «1»

«READ»

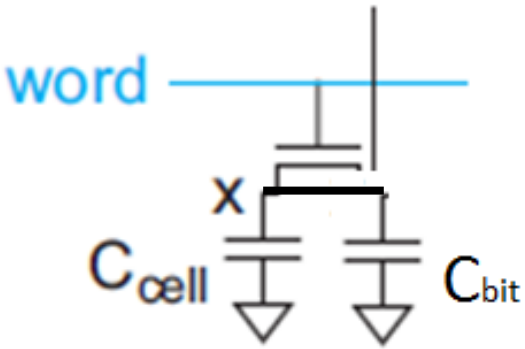
1. Cargar el valor «VDD/2» en bit
2. Poner en «1» la línea «WORD»
3. Se producirá un ΔV que puede ser medido
4. La lectura modifica la tensión de «x» por lo que se debe reescribir el dato

«WRITE»

1. Cargar el valor del dato en bit
2. Poner en «1» la línea «WORD»
3. El valor del dato se almacenara en C_{cell}



En la operación de «WRITE» algunas DRAM, cuando almacenan un valor lógico «1», colocan VDD+Vth en «word» para evitar que se degrade el valor del dato



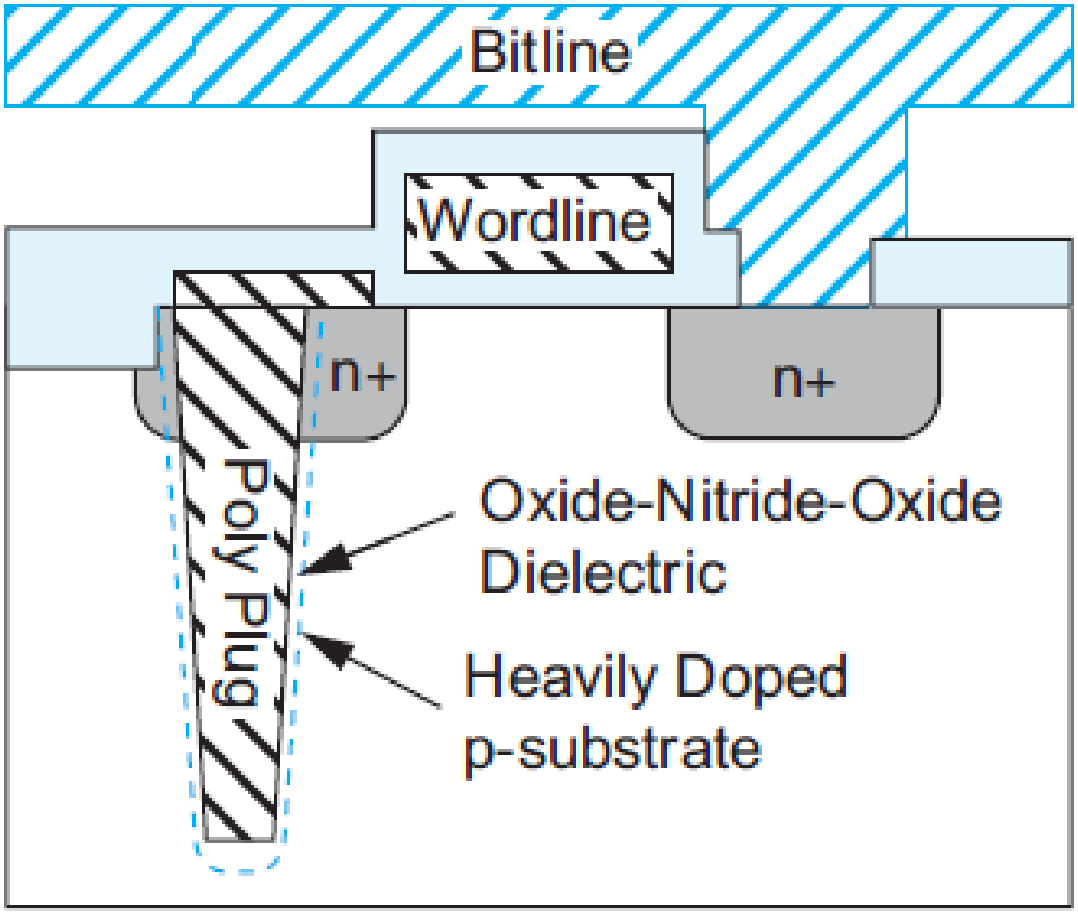
$$\Delta V = \frac{V_{DD}}{2} \frac{C_{cell}}{C_{cell} + C_{bit}}$$

El capacitor Ccell debe ser lo mas pequeño posible para obtener alta densidad [bits/cm^2]

La operación de «READ» depende de ΔV y este valor esta determinado por la relacion C_{cell}/C_{bit}

CELDA DRAM

Layout



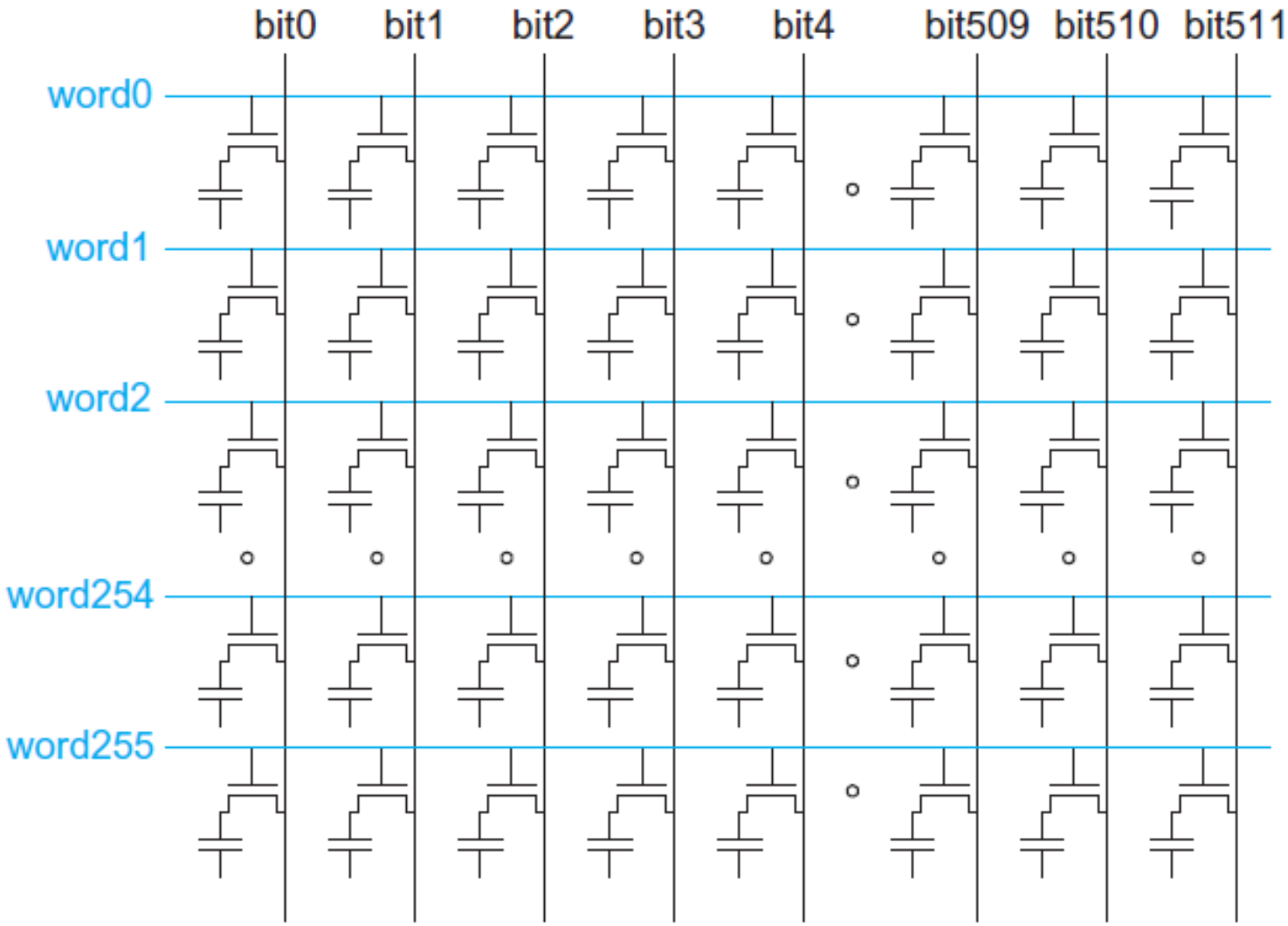
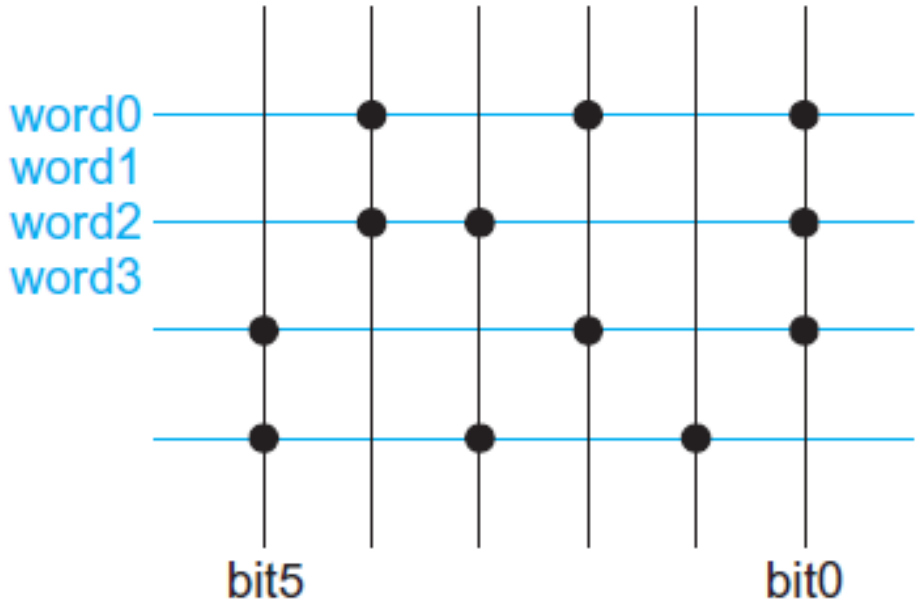


FIGURE 12.43 DRAM subarray

ROM

- Es una estructura de memoria no volátil
- Puede retener indefinidamente la información, aun sin alimentación
- La celda de memoria ROM puede implementarse con solo un transistor por bit
- El contenido de una ROM puede ser representado simbólicamente con un diagrama de puntos, donde la presencia de un punto indica una un «1» lógico
- Los datos son grabados por el fabricante y no pueden cambiarse




```
word0: 010101
word1: 011001
word2: 100101
word3: 101010
```

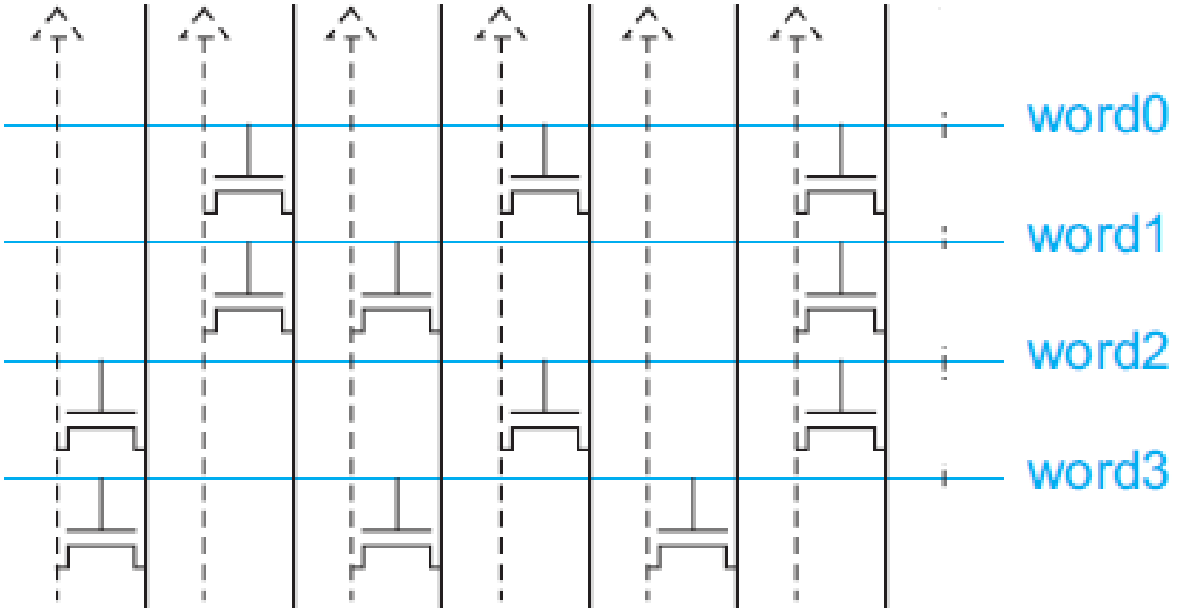
ROM programada por mascara

Se configura por presencia o ausencia

- un Transistor MOS
- un contacto

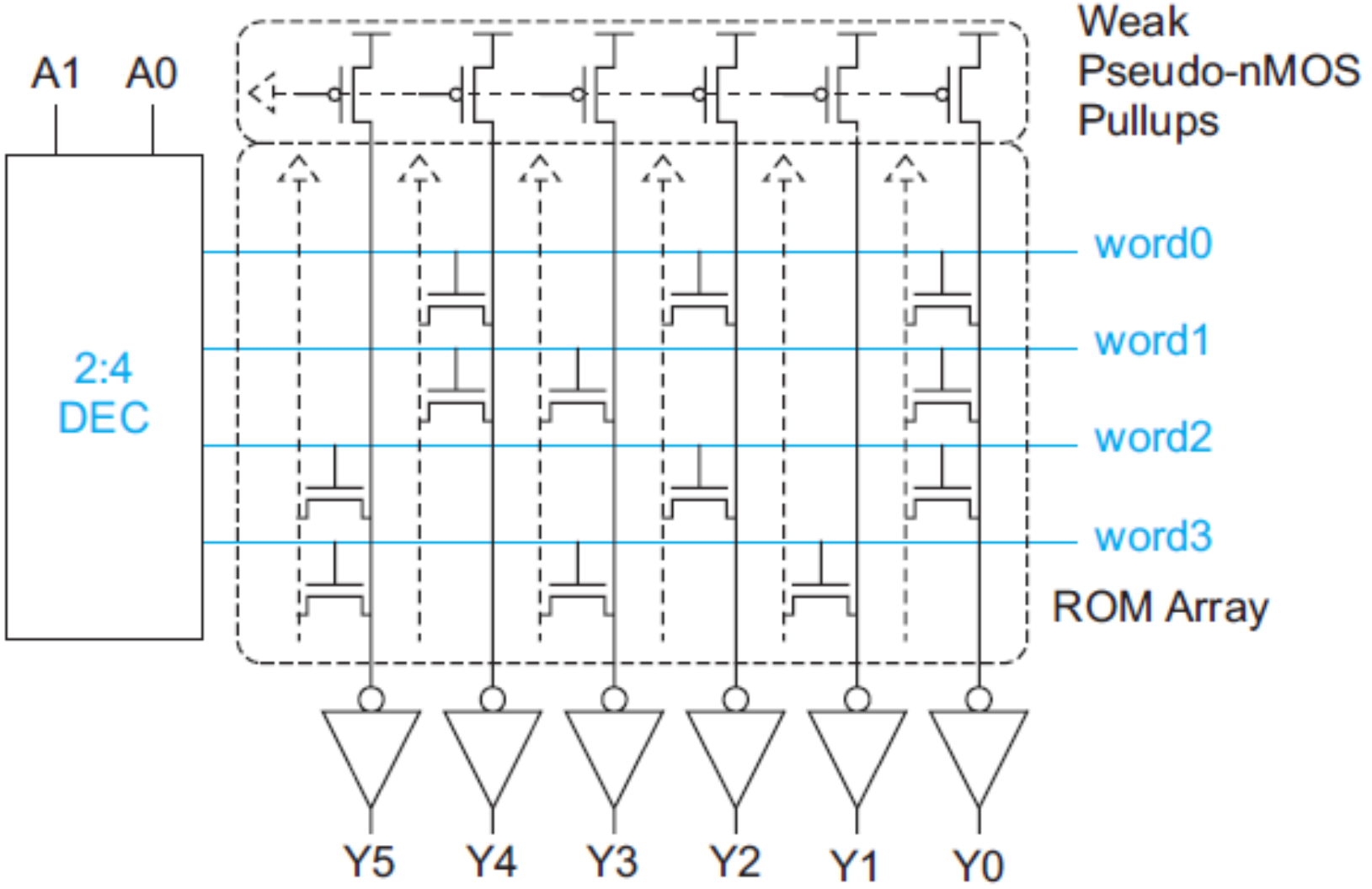
→ un implante que coloca el VTH del MOS en permanente OFF

- El método de omitir el MOS de contacto reduce la capacitancia de la línea «word» y el consumo de potencia



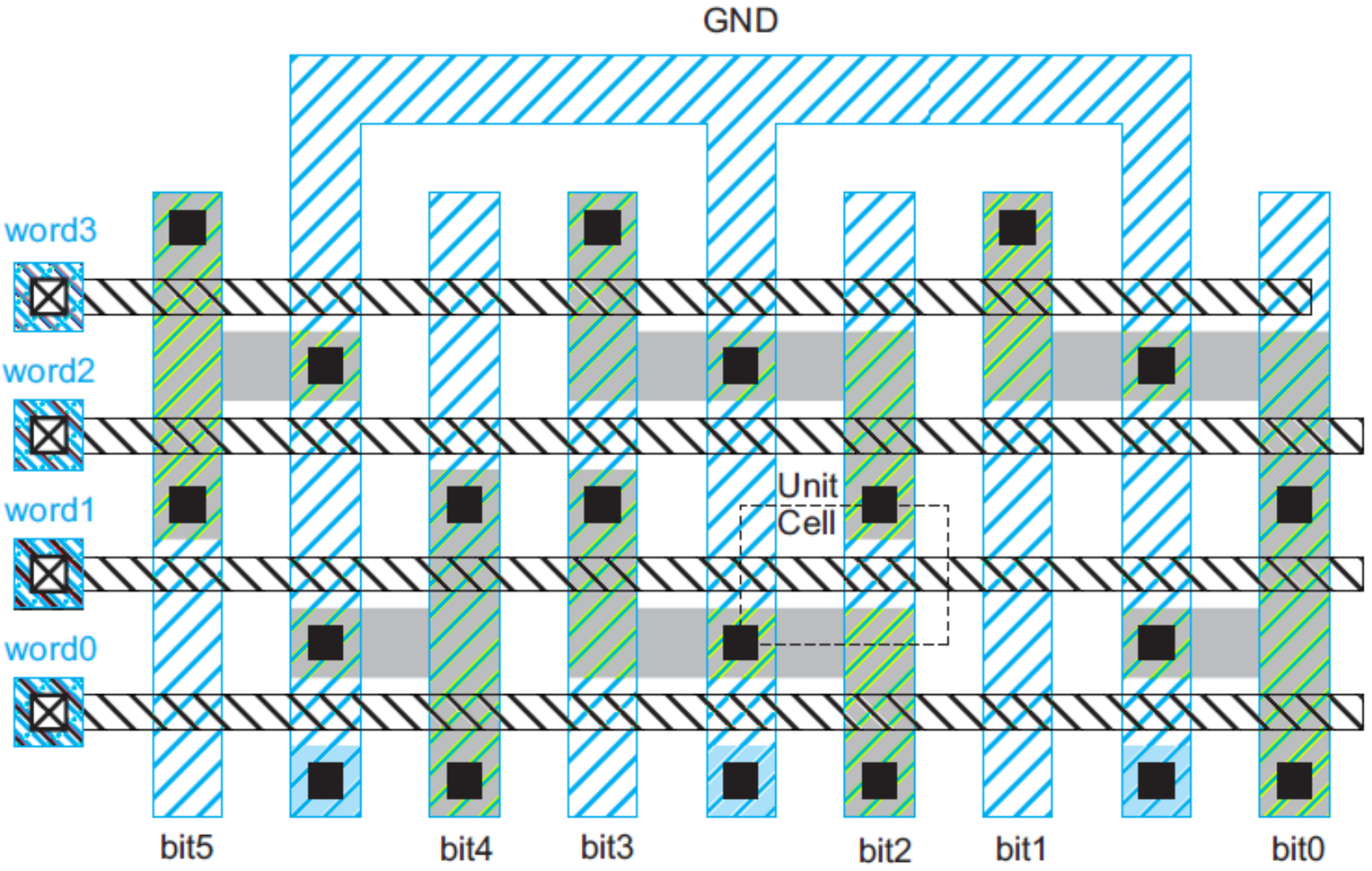
Arreglo seudo nMOS de ROM

NOR



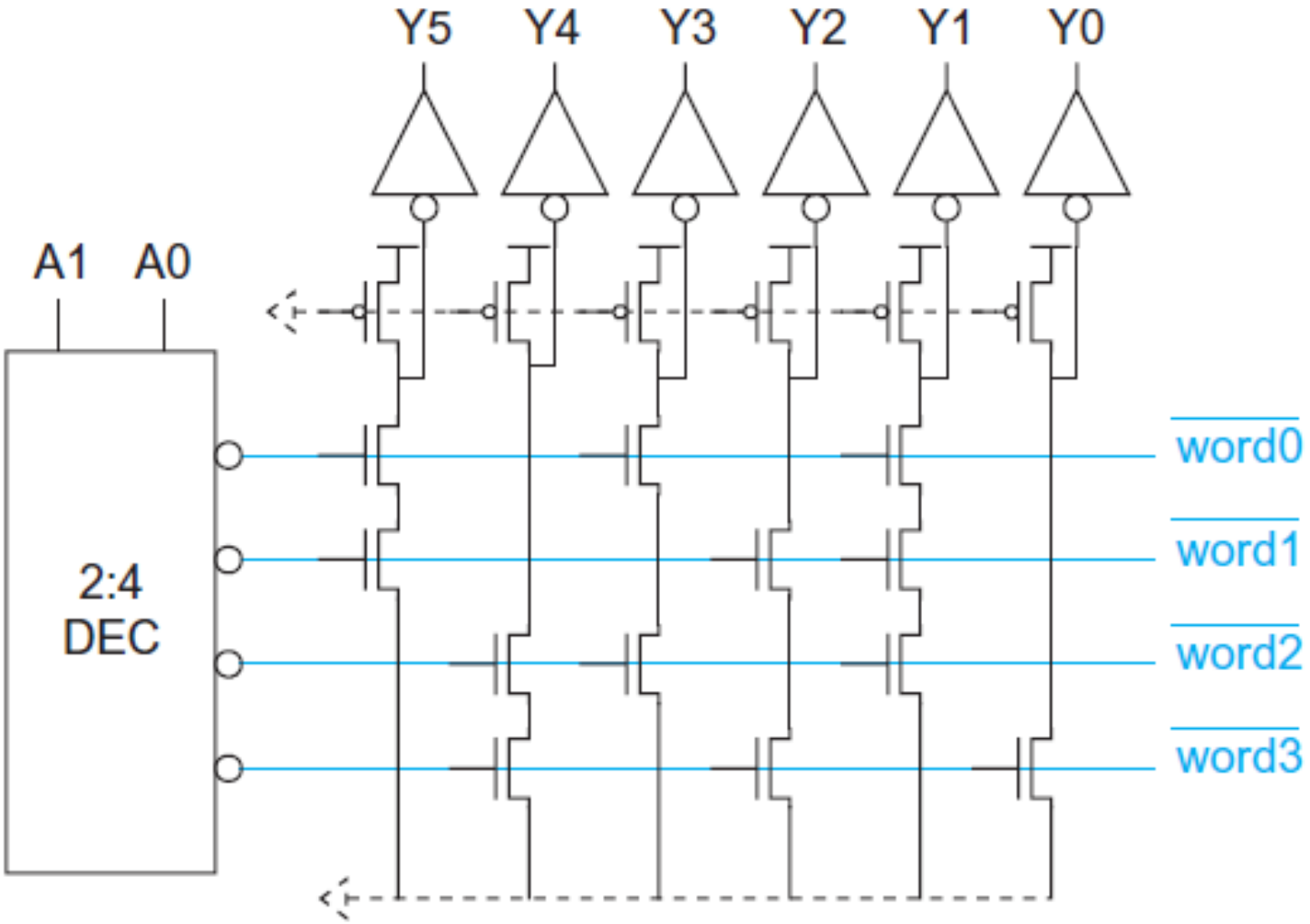
ROM

Layout NOR



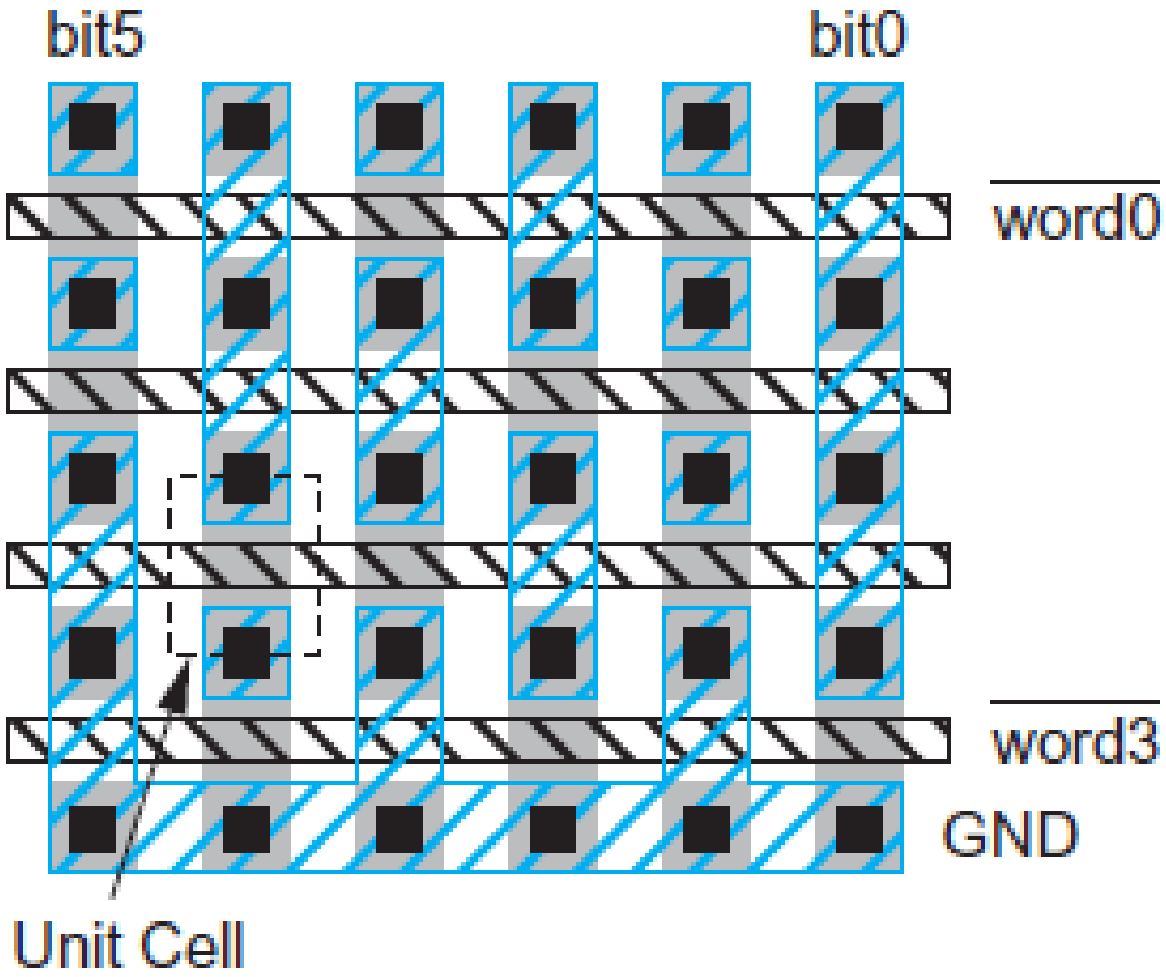
Arreglo seudo nMOS de ROM

NAND



ROM

Layout NAND



PROM

- Es una estructura de memoria no volátil

- Puede retener indefinidamente la información, aun sin alimentación

- Las memorias PROM pueden ser programadas por el usuario en un proceso de escritura que es lento

- Una vez programadas no pueden borrarse, el grabado se asimila a quemar fusibles

- Se fabrican con la misma configuración que las ROM de mascara usando un transistor de «pull-down» en cada bit
- Cada transistor esta en serie con un fusible fabricado de polysilicio, nicromo, o algún otro conductor el cual se quema aplicando alta corriente durante el proceso de grabado
- Una vez quemados los fusible no pueden volver a conectarse por lo que se configura una «One time programmable memorie»

EPROM-EEPROM-FLASH

- Las memorias PROMS utilizan el quemado de fusibles para grabar un dato

- Las memorias EPROMS, EEPROM y FLASH utilizan carga eléctrica almacenada en la compuerta flotante de un MOS

- Se utiliza una segunda capa de poli silicio entre el canal y la compuerta de control.

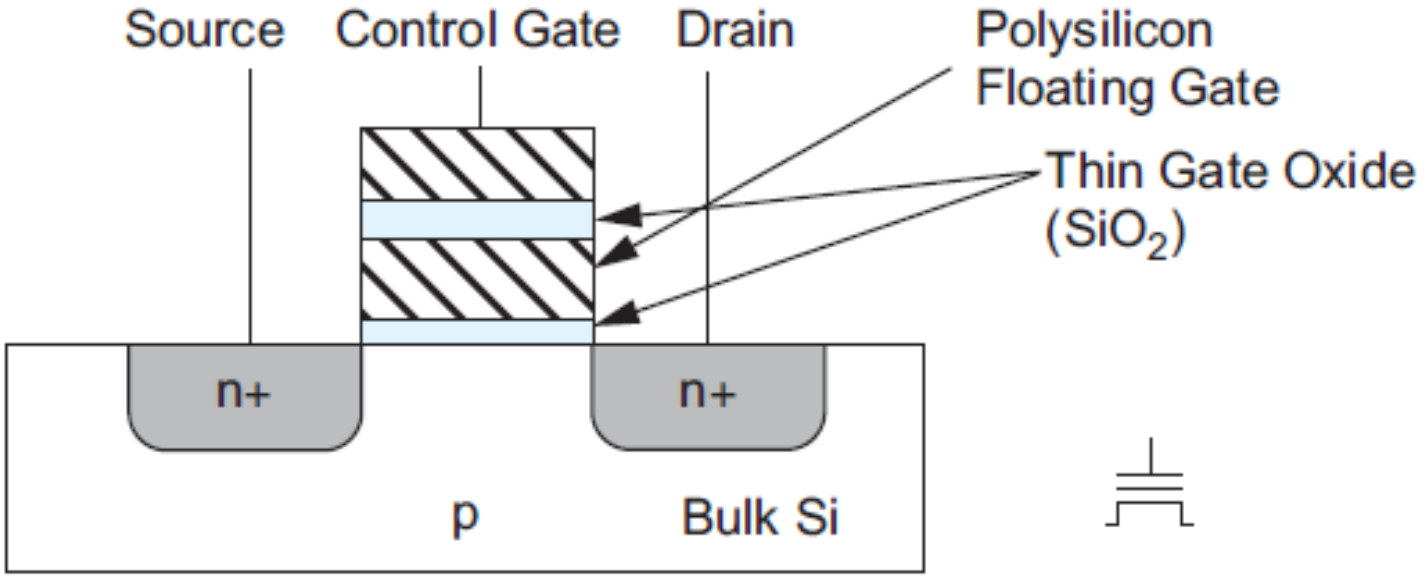
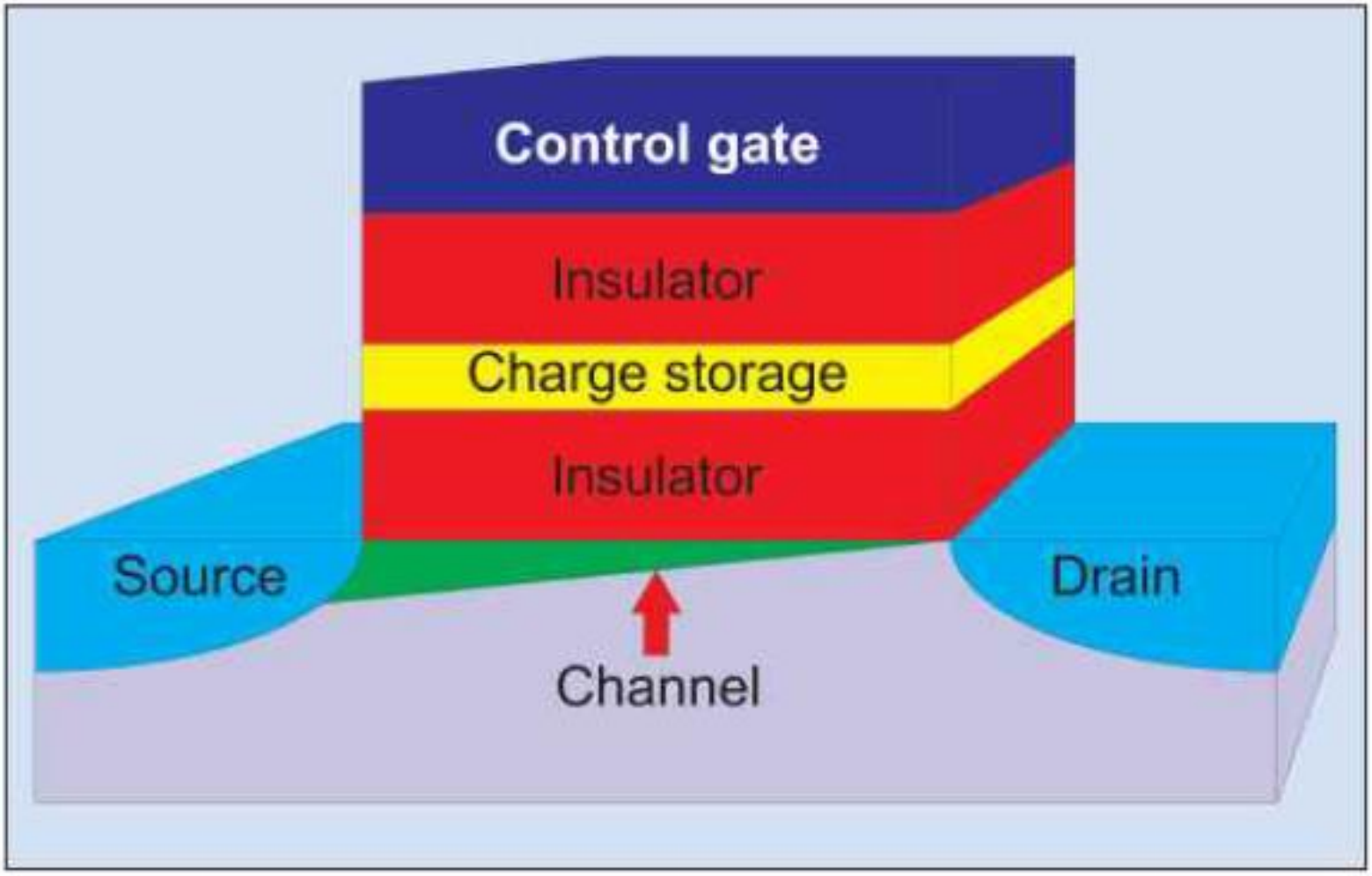


FIGURE 12.57 Cross-section of floating gate nMOS transistor

MOS para celda de memoria EEPROM o FLASH



- La compuerta flotante es un buen conductor pero no esta eléctricamente conectada

- Aplicando una tensión elevada en la compuerta de control se produce el salto de electrones, a través del oxido, a la compuerta flotante (Proceso de tunelización de Fowler-Nordheim)

- Los electrones inyectados aumentan el VTH del MOS y permanece siempre cortado «OFF»

- La programación se hace eléctricamente, aplicando una tensión elevada en la compuerta de control

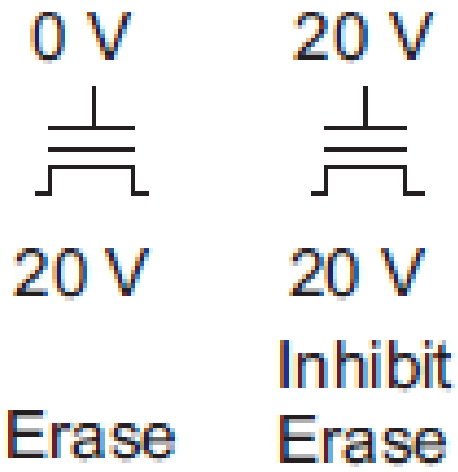
- Para el proceso de borrado (sacar los electrones de la compuerta flotante):
    - EPROM usan luz ultravioleta
    - EEPROM y FLASH usan tensión

- EEPROM se borra bit a bit
- FLASH se borra en grupos de bits

EEPROM – FLASH

BORRADO

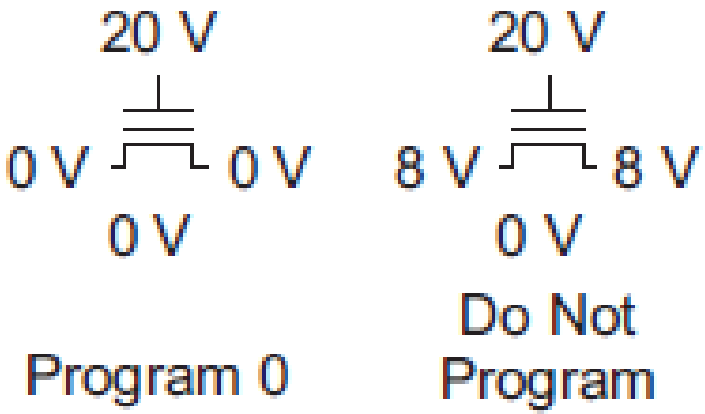
- Para el proceso de borrado (sacar electrones de la compuerta flotante) se conecta el sustrato a 20 V y la compuerta de control a 0 V
- En las celdas que no se quieren borrar se conecta la compuerta de control también a 20 V
- El proceso es lento, del orden de los milisegundos
- Terminado el proceso todos los bits quedan en «1»



EEPROM – FLASH

WRITE

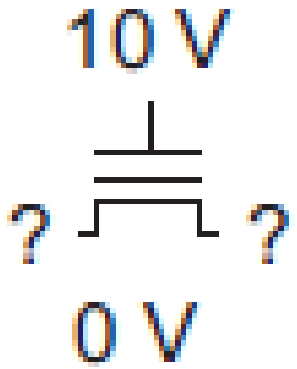
- Para el proceso de escritura «write» (colocar electrones en la compuerta flotante) se conecta la compuerta de control a 20 V y el sustrato a 0 V
- En las celdas que no se quieren escribir se conecta el drenador y la fuente a 8 V
- Una vez finalizado el proceso quedan en «0» las celdas seleccionadas y en «1» las otras



EEPROM – FLASH

READ

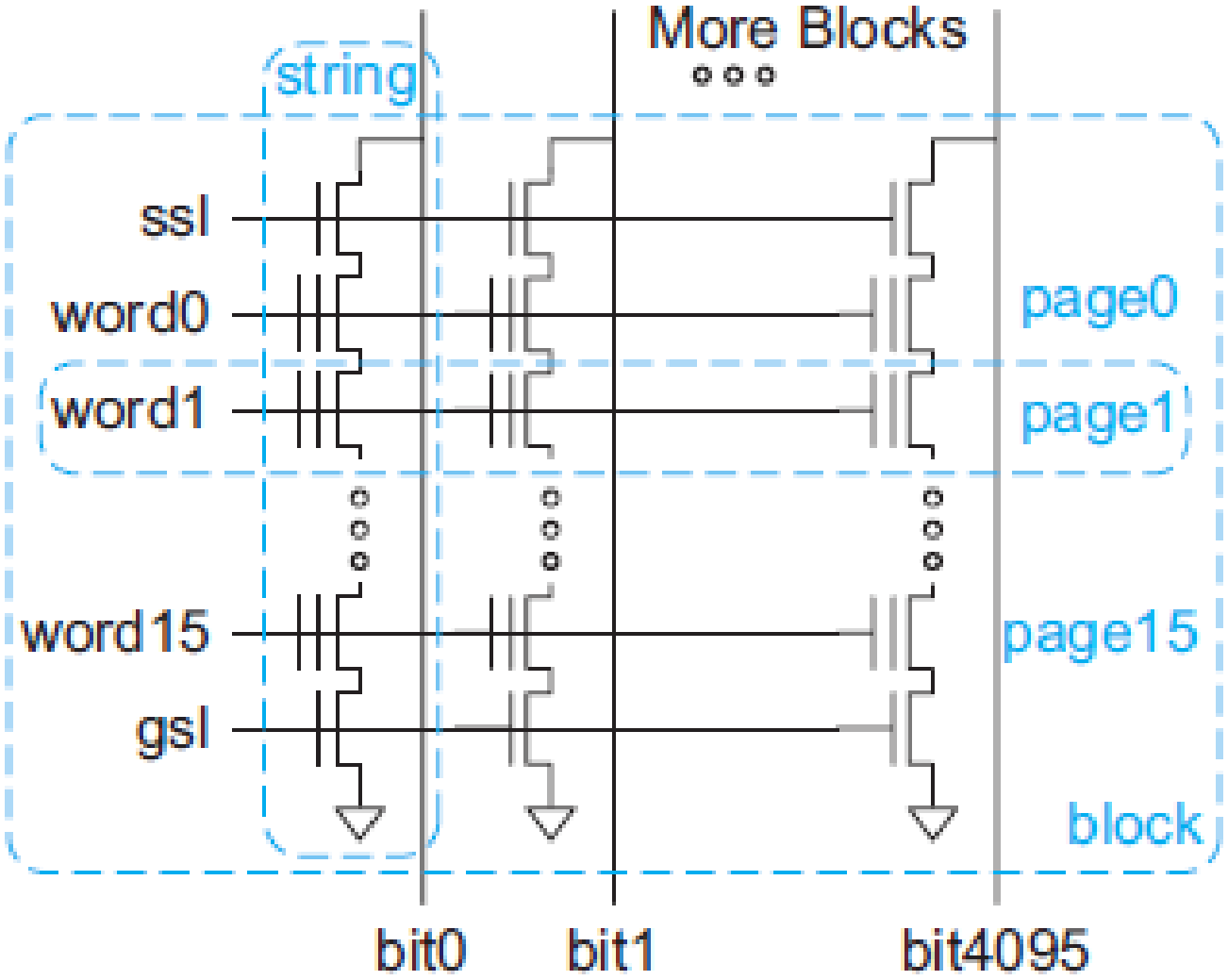
- Para el proceso de lectura «read» se conecta la compuerta de control a 10 V
- Las celdas con electrones en la compuerta flotante tendrán desconectado el drenador de la fuente
- Las celdas sin electrones en la compuerta flotante tendrán conectado el drenador con la fuente



Memoria FLASH → Arquitectura NAND

WRITE → Por palabra

BORRADO → Por bloque



Two reliability metrics for Flash memories are retention time and endurance. The *retention time* is the duration for which a Flash cell will hold its value. Under normal conditions, the charge on the floating gate would take thousands or millions of years to leak off. However, defects in the oxide may increase leakage for some cells. Manufacturers typically specify a 10 year retention time. *Endurance* is the number of times that a cell can be erased and reprogrammed. The high voltages stress the oxide and can eventually cause it to wear out. Endurance of 100,000 erase-program cycles are typical, but some multilevel Flash cells have endurance as low as 5000 cycles.

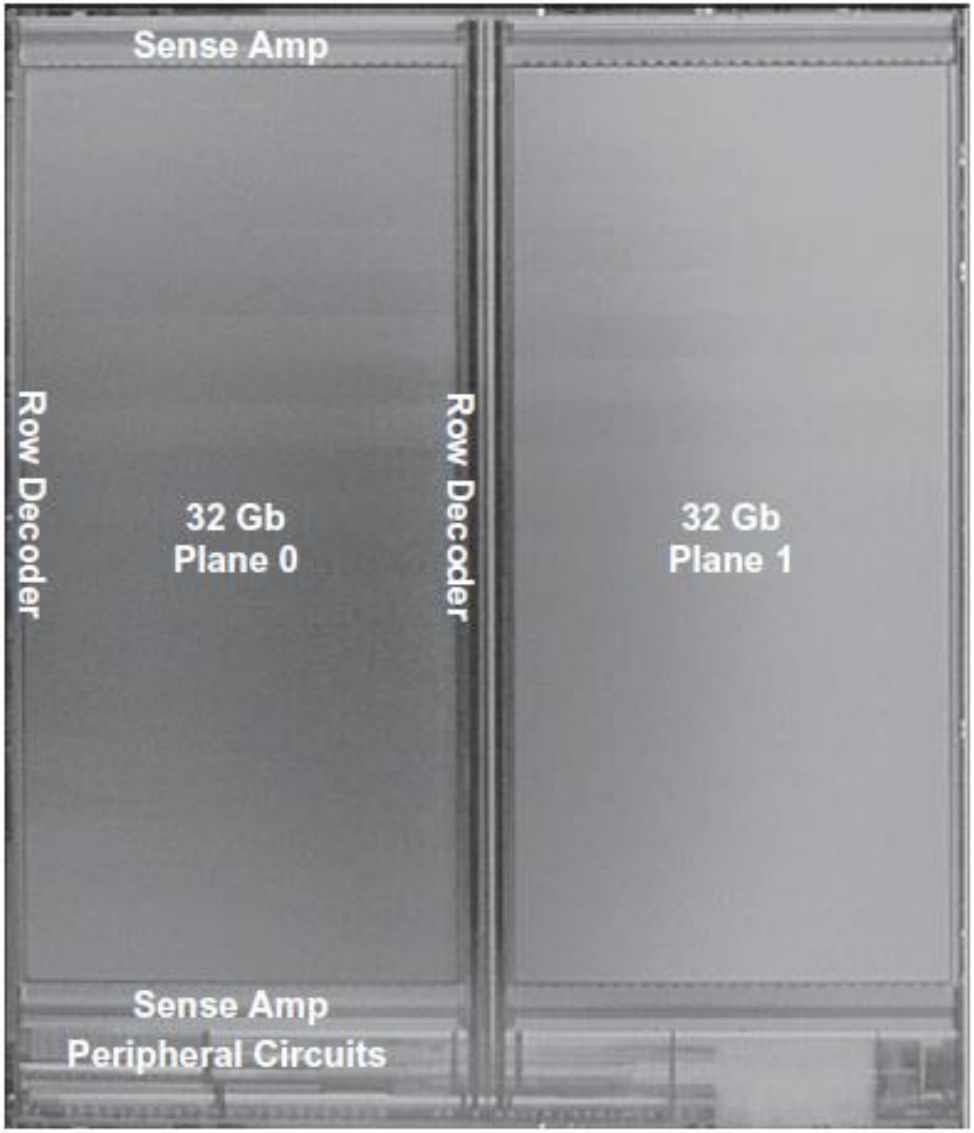


FIGURE 12.62 64 Gb NAND Flash (© 2009 IEEE.)

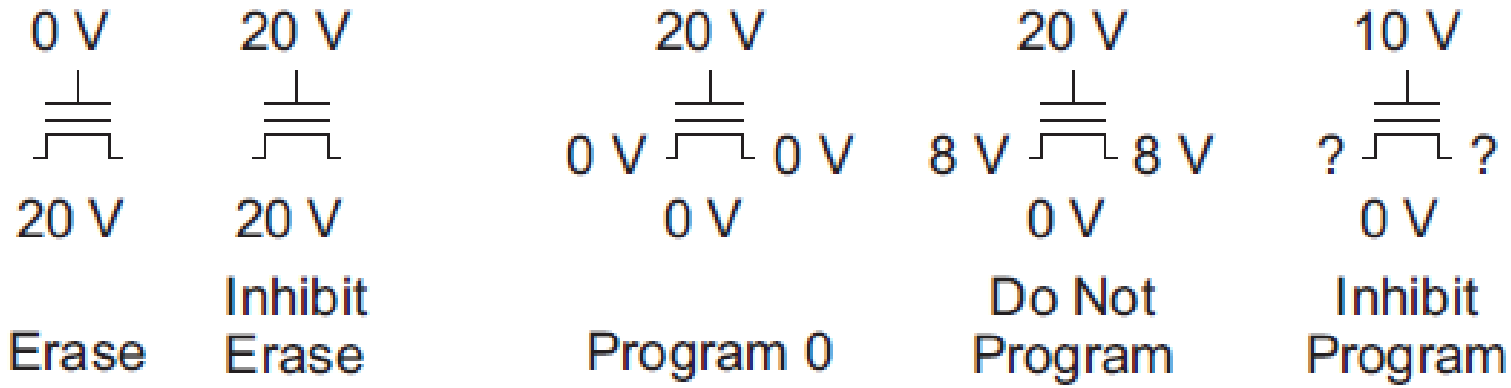


Figure 12.61 shows the operation of the Flash memory using voltages representative of a multimegabit design. The block is erased by setting all of the control gates to GND and raising the substrate to 20 V. The high voltage across the gate oxide induces FN tunneling, causing the electrons to flow from the floating gate to the substrate. At the end of the erase step, all the floating gate transistors have a negative V_t and thus represent 1. Tunneling is a slow process, so block erase takes on the order of a millisecond. The wordlines for other blocks on the chip are set to the same voltage as the substrate to inhibit erasing. An on-chip charge pump (see Section 13.3.7) is used to generate the high voltages.

A cell is programmed (written) to 0 by tunneling electrons onto the floating gate. The programming cannot restore 1 values, so the block must be erased before any cell is reprogrammed. An entire page is programmed at once. To program a page, the bitlines are driven with the data values: 0 V for a logic 0 and 8 V for a logic 1. The substrate is held at ground. The wordline is set to 20 V for the page being programmed and 10 V for the other pages in the block. The ground select line (*gsl*) is left OFF but the string select line (*ssl*) for the block is turned ON, passing the voltage on the bitline to the channels of all the transistors being programmed. Thus, cells being programmed to 0 see 20 V on the control gate and 0 V on the channel. This high voltage difference induces FN tunneling that drives electrons onto the floating gate, raising V_t to a positive voltage. The other cells see a smaller voltage that is insufficient to cause tunneling.