Memory

Outline

- Sense Amplifier
- DRAM
- FLASH

nMOS I-V Summary

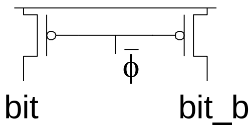
$$I_{ds} = \begin{cases} 0 & V_{gs} < V_t & \text{cutoff} \\ \beta\left(V_{gs} - V_t - {V_{ds}}/{2}\right)V_{ds} & V_{ds} < V_{dsat} & \text{linear} \\ \dfrac{\beta}{2}\left(V_{gs} - V_t\right)^2 & V_{ds} > V_{dsat} & \text{saturation} \end{cases}$$

Column Circuitry

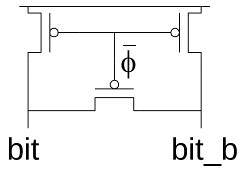
- Some circuitry is required for each column
  - Bitline conditioning
  - Sense amplifiers
  - Column multiplexing

Bitline Conditioning

- Precharge bitlines high before reads



- Equalize bitlines to minimize voltage difference when using sense amplifiers



Sense Amplifier: Why?

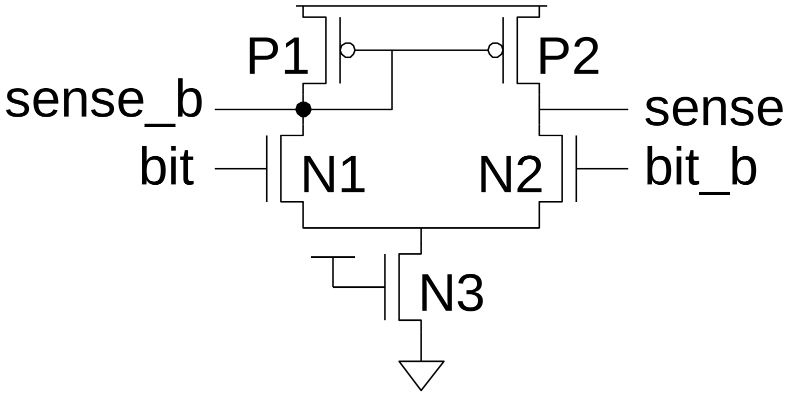
- **Bit line cap significant for large array**
  - **If each cell contributes 2fF,**
    - **for 256 cells, 512fF plus wire cap**
  - **Pull-down resistance is about 15K**
  - **RC = 7.5ns! (assuming $\Delta V = V_{dd}$)**
- **Cannot easily change R, C, or V_{dd}, but can change ΔV i.e. smallest sensed voltage**
  - **Can reliably sense ΔV as small as <50mV**

Sense Amplifiers

- Bitlines have many cells attached
  - Ex: 32-kbit SRAM has 256 rows x 128 cols
  - 256 cells on each bitline
- $t_{pd} \propto (C/I)\ \Delta V$
  - Even with shared diffusion contacts, 64C of diffusion capacitance (big C)
  - Discharged slowly through small transistors (small I)
- *Sense amplifiers* are triggered on small voltage swing (reduce ΔV)

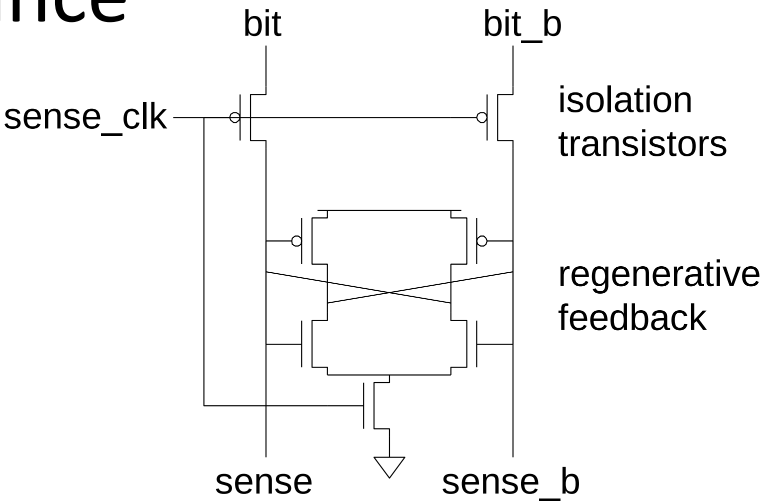
Differential Pair Amp

- Differential pair requires no clock
- But always dissipates static power



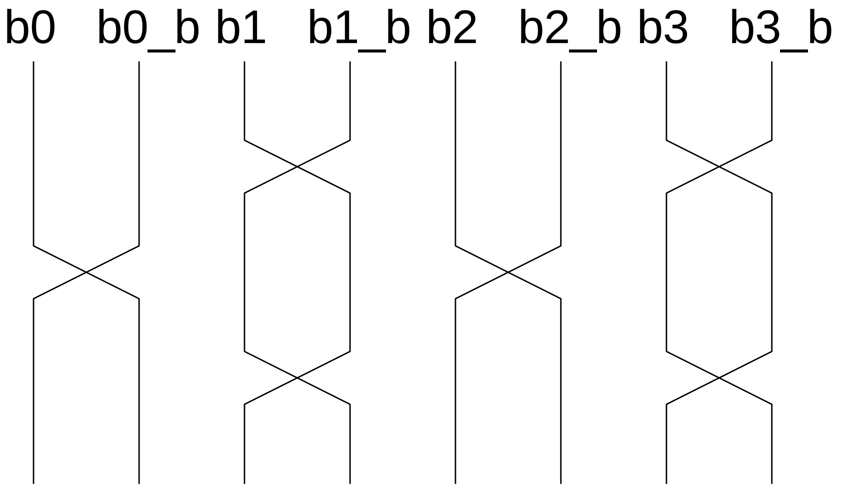
Clocked Sense Amp

- Clocked sense amp saves power
- Requires sense_clk after enough bitline swing
- Isolation transistors cut off large bitline capacitance



Twisted Bitlines

- Sense amplifiers also amplify noise
  - Coupling noise is severe in modern processes
  - Try to couple equally onto bit and bit_b
  - Done by *twisting* bitlines

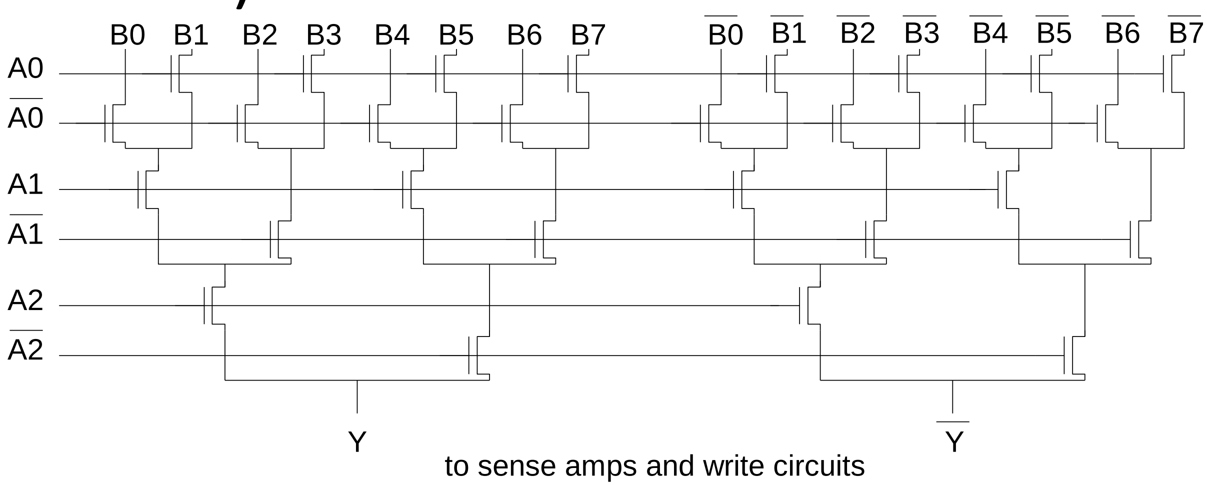


Column Multiplexing

- Recall that array may be folded for good aspect ratio
- Ex: 2 kword x 16 folded into 256 rows x 128 columns
  - Must select 16 output bits from the 128 columns
  - Requires 16 8:1 column multiplexers

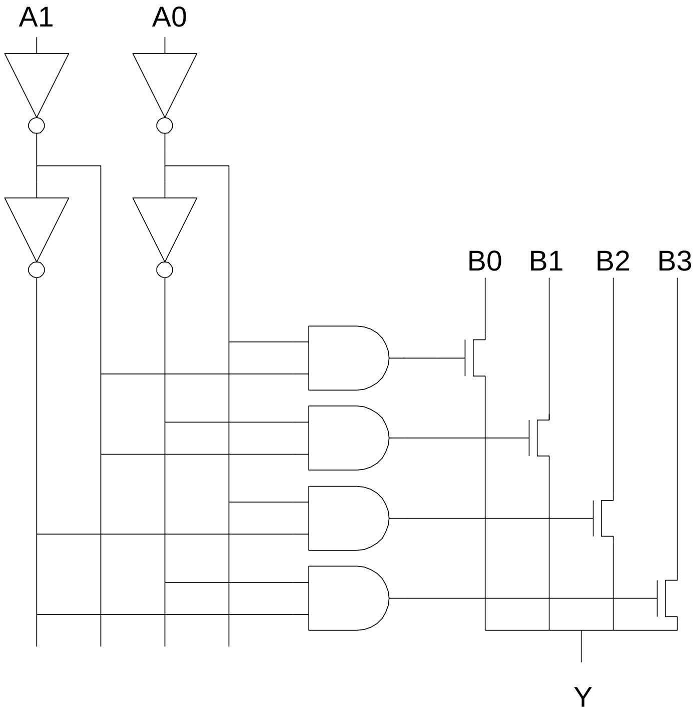
Tree Decoder Mux

- Column mux can use pass transistors
  - Use nMOS only, precharge outputs
- One design is to use k series transistors for 2^k:1 mux
  - No external decoder logic needed (big area reduction)



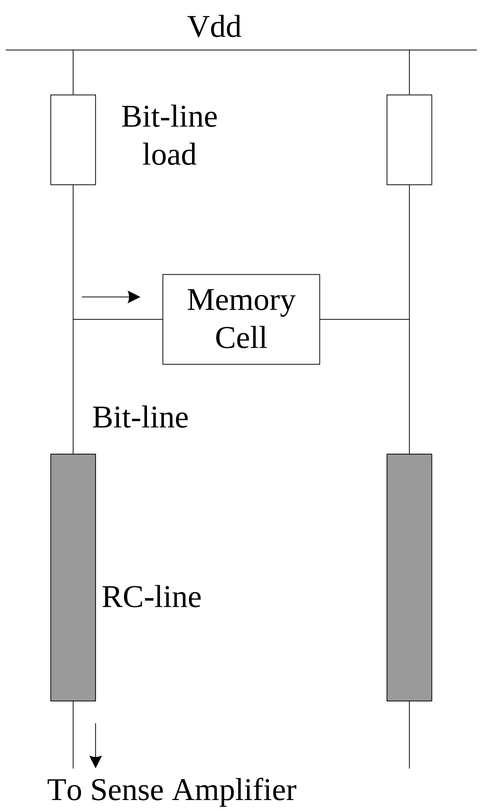
Single Pass-Gate Mux

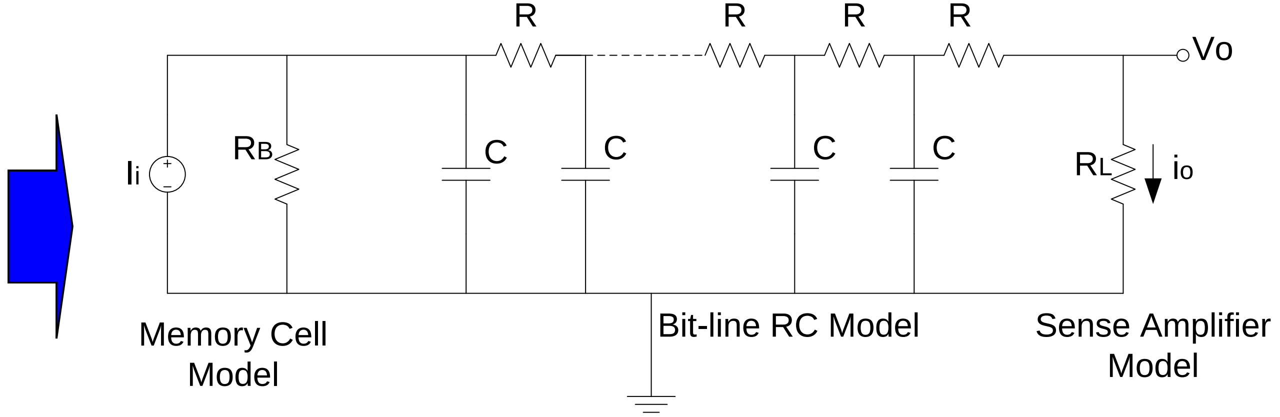
- Or eliminate series transistors with separate decoder



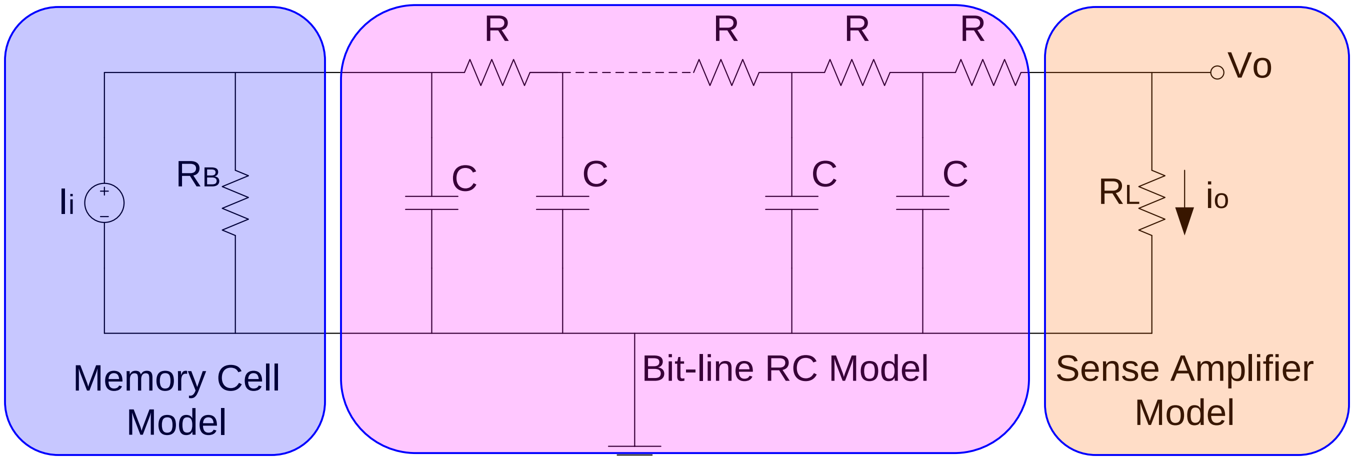
- Sense Amplifier Design Objective
  - Minimum sense delay
  - Required amplification
  - Minimum power consumption
  - Restricted layout area
  - High reliability and tolerance
- Classification

Circuit Types	Operation Mode
Differential	*Voltage-mode*
Nondifferential	*Current-mode*

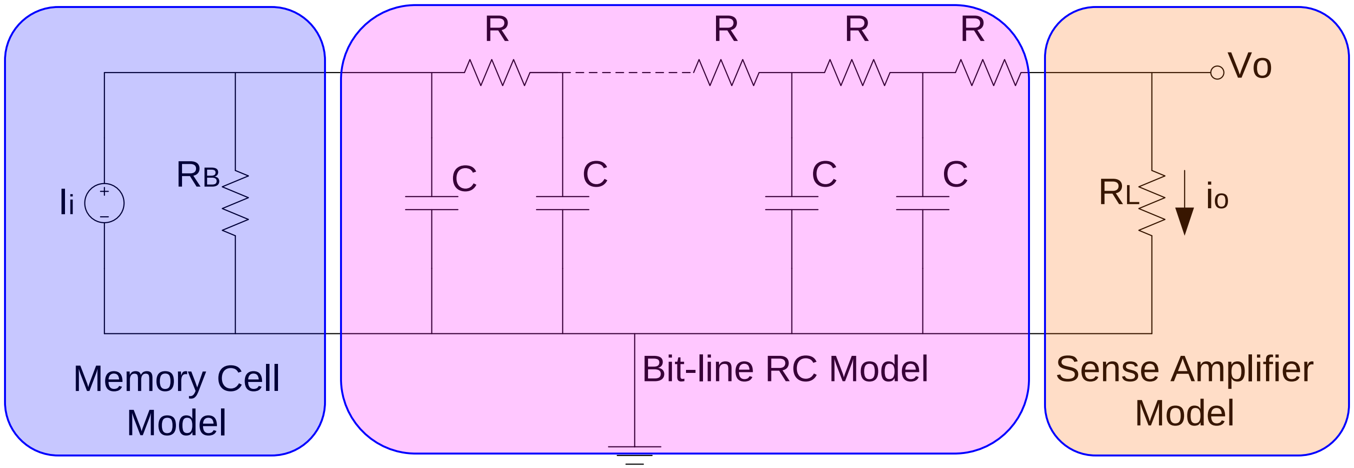




Bit-line Model



- $\boldsymbol{I_i}$ is the output current of the driving source, i.e. memory cell.
- $\boldsymbol{R_B}$ is the output resistance of the bit-line load in parallel with the drain resistance of the access transistor, which is the output device of the memory cell.
- The infinite ***RC*** ladder structure represents the interconnect line. The total resistance and capacitance of the line is given by $\boldsymbol{R_T}$ and $\boldsymbol{C_T}$.
- The output of the line is terminated by resistor $\boldsymbol{R_L}$.



- RC delay [1]

$$\delta t = \frac{R_T C_T}{2} \bullet \left(\frac{R_B + \frac{R_T}{3} + R_L}{R_B + R_T + R_L} \right) + R_B C_T \bullet \left(\frac{R_L}{R_B + R_T + R_L} \right)$$

- Delay for Voltage-Mode
  - For voltage-mode signals, R_L *is infinite* and the output signal is the *open-circuit* voltage V_o.
  - $$\boxed{\delta t_v = \frac{R_T C_T}{2} \bullet \left(1 + \frac{2R_B}{R_T}\right)}$$

- Delay for Current-Mode
  - For current-mode signals, R_L *is ideally zero* and the output signal is the *short-circuit* current i_o.
  - $$\boxed{\delta t_i = \frac{R_T C_T}{2} \bullet \left(\frac{R_B + \frac{R_T}{3}}{R_B + R_T}\right)}$$

- Example from [1], when
  - RB = 2500Ω
  - RT=250Ω
  - CT=2pf
- Voltage-mode
  - $\delta t_v = \frac{R_T C_T}{2} \bullet \left(1 + \frac{2R_B}{R_T}\right)$ = 5.25ns
- Current-mode
  - $\delta t_i = \frac{R_T C_T}{2} \bullet \left(\frac{R_B + \frac{R_T}{3}}{R_B + R_T}\right)$ 0.235ns

Traditional Voltage-Mode Sense Amplifier

- Traditional Difference Amplifier
- Full Complementary Positive-Feedback Sense Amplifier
- Enhanced Positive Feedback Sense Amplifier

Traditional Voltage-Mode Sense Amplifier (cont.)

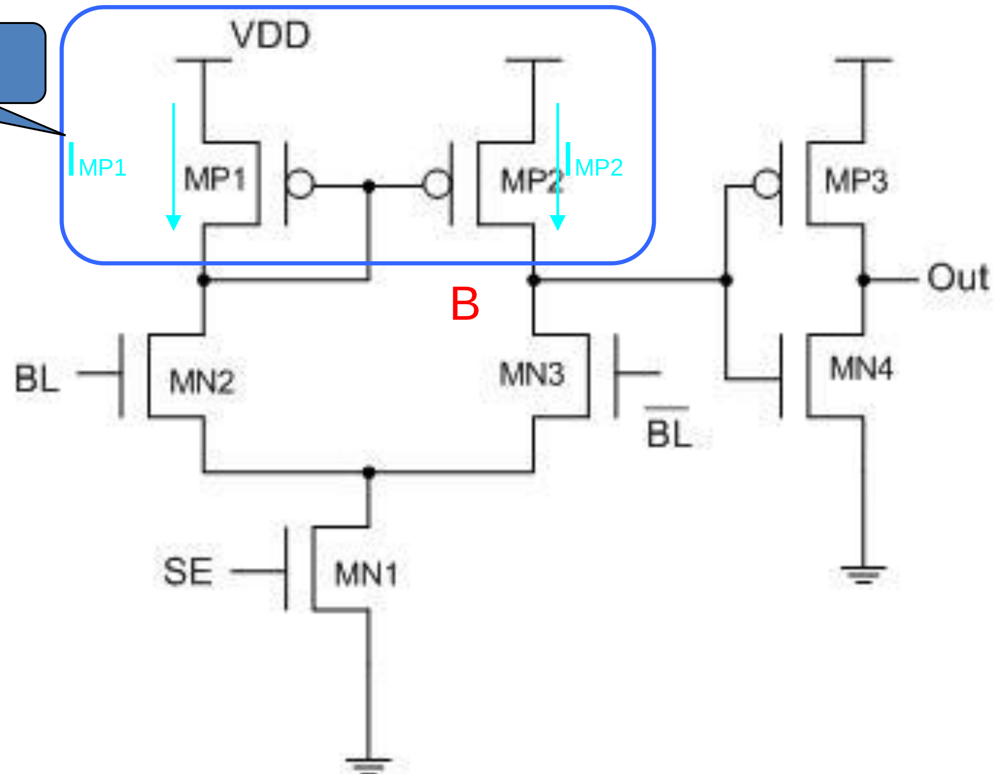
Traditional Difference Amplifier

Current mirror

$$I_{MP1} = \frac{1}{2}k_n^{'}\left(\frac{W}{L}\right)_{MP1}(V_{gs} - V_t)^2$$

$$I_{MP2} = \frac{1}{2}k_n^{'}\left(\frac{W}{L}\right)_{MP2}(V_{gs} - V_t)^2$$

$$\frac{I_{MP2}}{I_{MP1}} = \frac{\left(\frac{W}{L}\right)_{MP2}}{\left(\frac{W}{L}\right)_{MP1}}$$

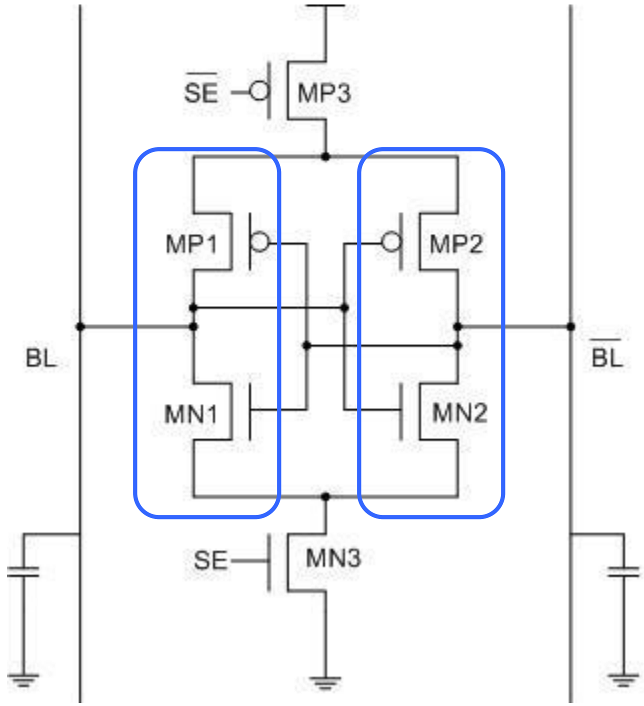


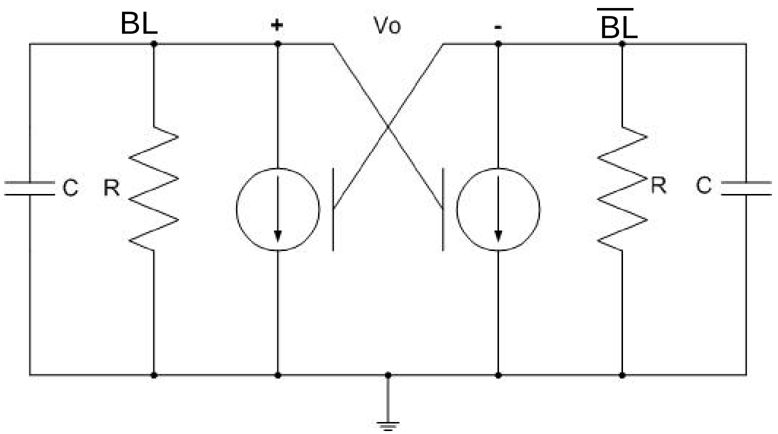
SE = 1 , Sense mode

SE = 0 , Standby mode

Traditional Voltage-Mode Sense Amplifier (cont.)

Full Complementary Positive Feedback Sense Amplifier





$$V_b(t) = V_o(0) \bullet e^{\frac{(g_m R - 1)t}{RC}}$$

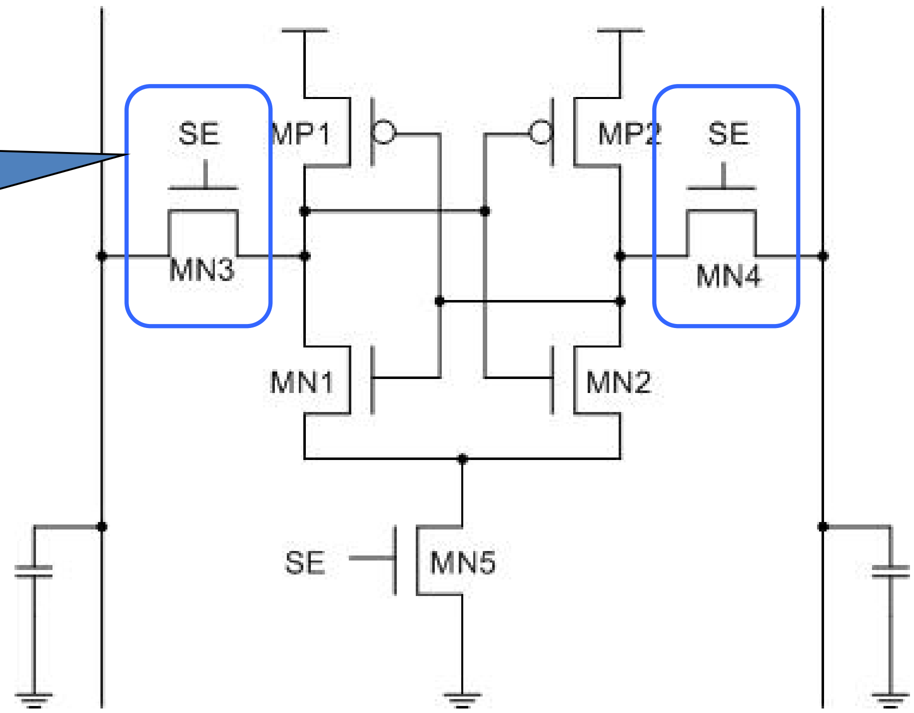
SE = 1 , Sense mode

SE = 0 , Standby mode

Traditional Voltage-Mode Sense Amplifier (cont.)

Enhanced Positive Feedback Sense Amplifier

Decouple device: avoid the effect of the bit-line capacitance



SE = 1 , Sense mode

SE = 0 , Standby mode

Current-Mode Sense Amplifier

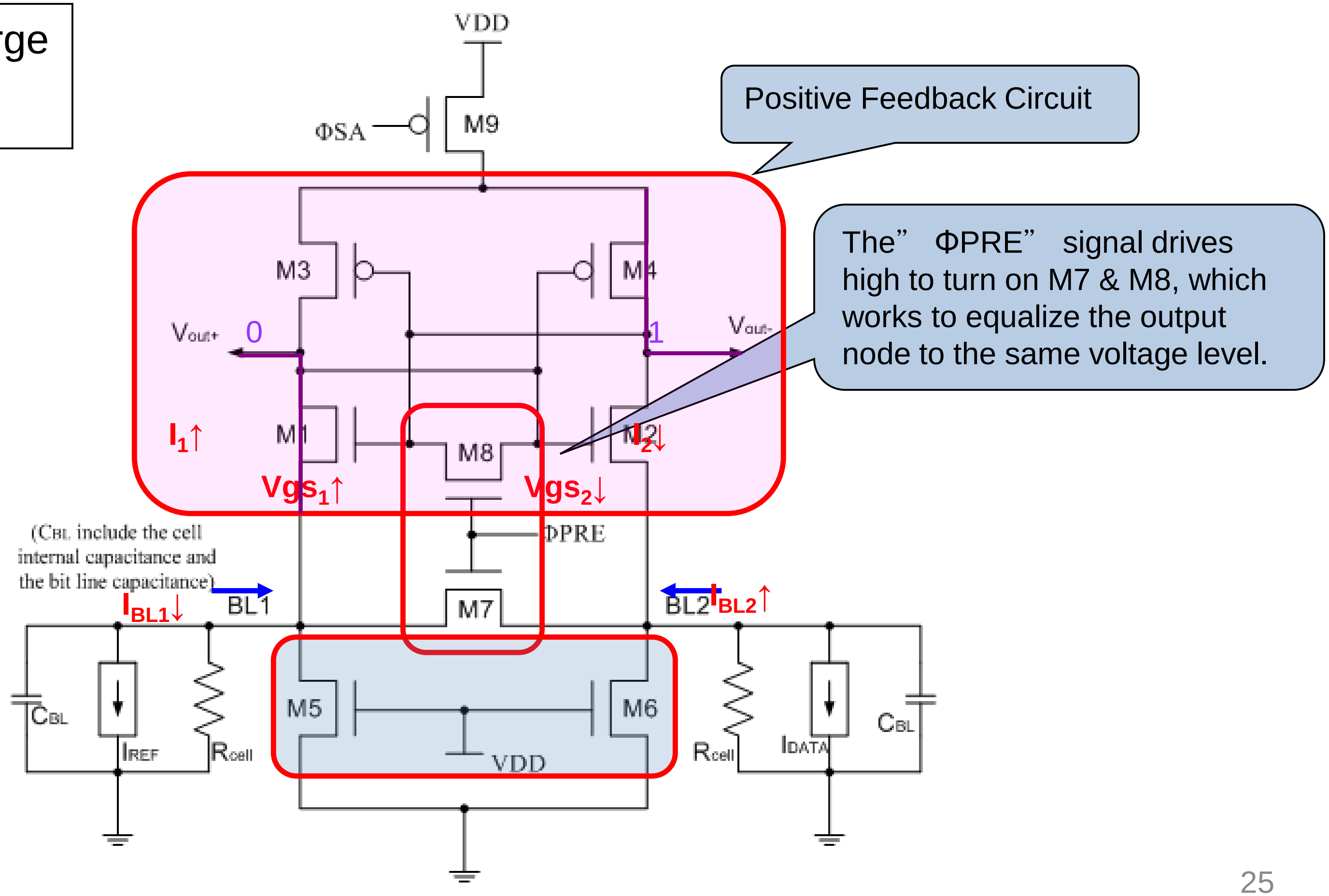
- Clamped Bit-Line Sense Amplifier
- Simple Four Transistor Sense Amplifier
- Hybrid Current Sense Amplifier
- New Hybrid Current Sense Amplifier

Current-Mode Sense Amplifier (cont.)

Clamped Bit-Line Sense Amplifier

Φ PRE = 1 , Pre-charge

Φ SA = 1 , Sense

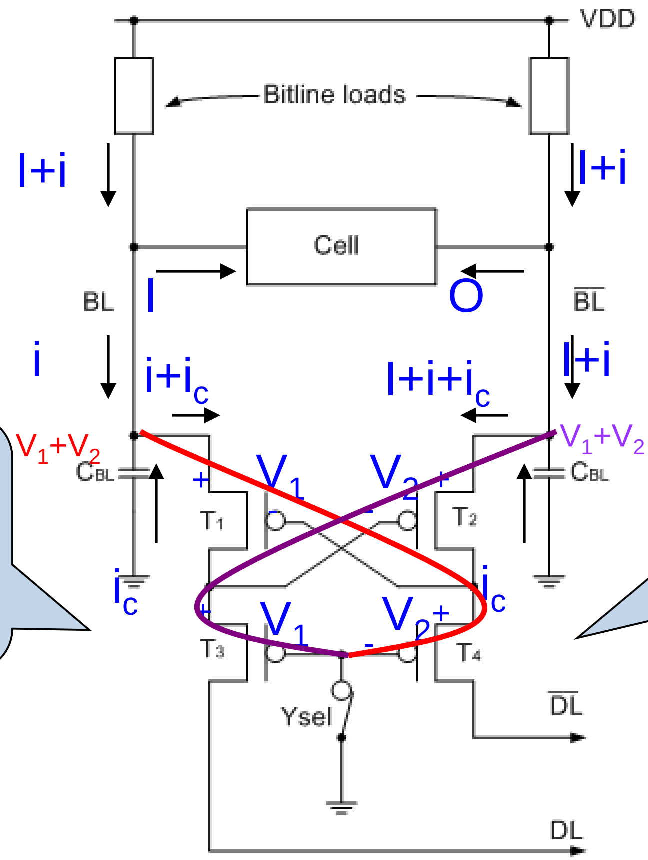


Advantages:

- The input nodes of the sense amplifier are low-impedance current sensitive nodes, the voltage swing of the highly capacitance bit lines change small.
- The output nodes of the sense amplifier are no longer with the bit-line capacitance and the sense amplifier is able to respond very rapidly.
- M1-M4 works as cross-coupled latch, its positive feedback effect can improve the driving ability of output nodes.

Current-Mode Sense Amplifier (cont.)

Simple Four Transistor Sense Amplifier (Current conveyor)



The gate-source voltage of T_1 will be equal to that of T_3, since their currents are equal, their size are equal, and both transistors are in saturation.

The gate-source voltage of T_2 will be equal to that of T_4, since their currents are equal, their size are equal, and both transistors are in saturation.

Advantages:

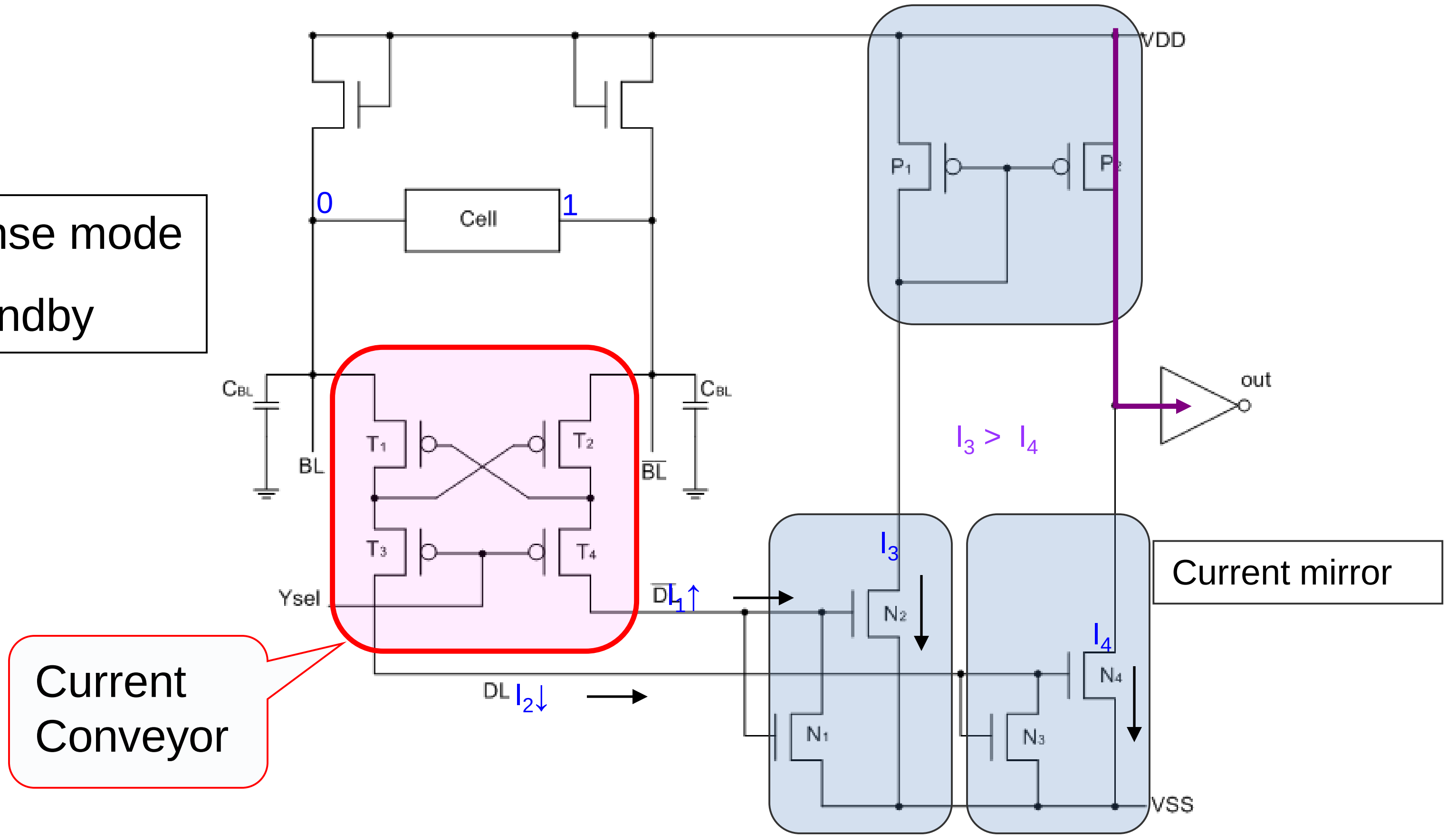
- In many cases it can fit in the column pitch, avoiding the need for column-select devices, thus again reducing propagation delay.
- There exists a virtual short circuit across the bit lines, therefore the potential of the bit lines will be equal independent of the current distribution.
- -The sensing delay is unaffected by the bit-line capacitance since no differential capacitor discharging is required to sense the cell data.
- Discharge current ic from the bit-line capacitors, effectively precharging the sense amplifier.

Current-Mode Sense Amplifier (cont.)

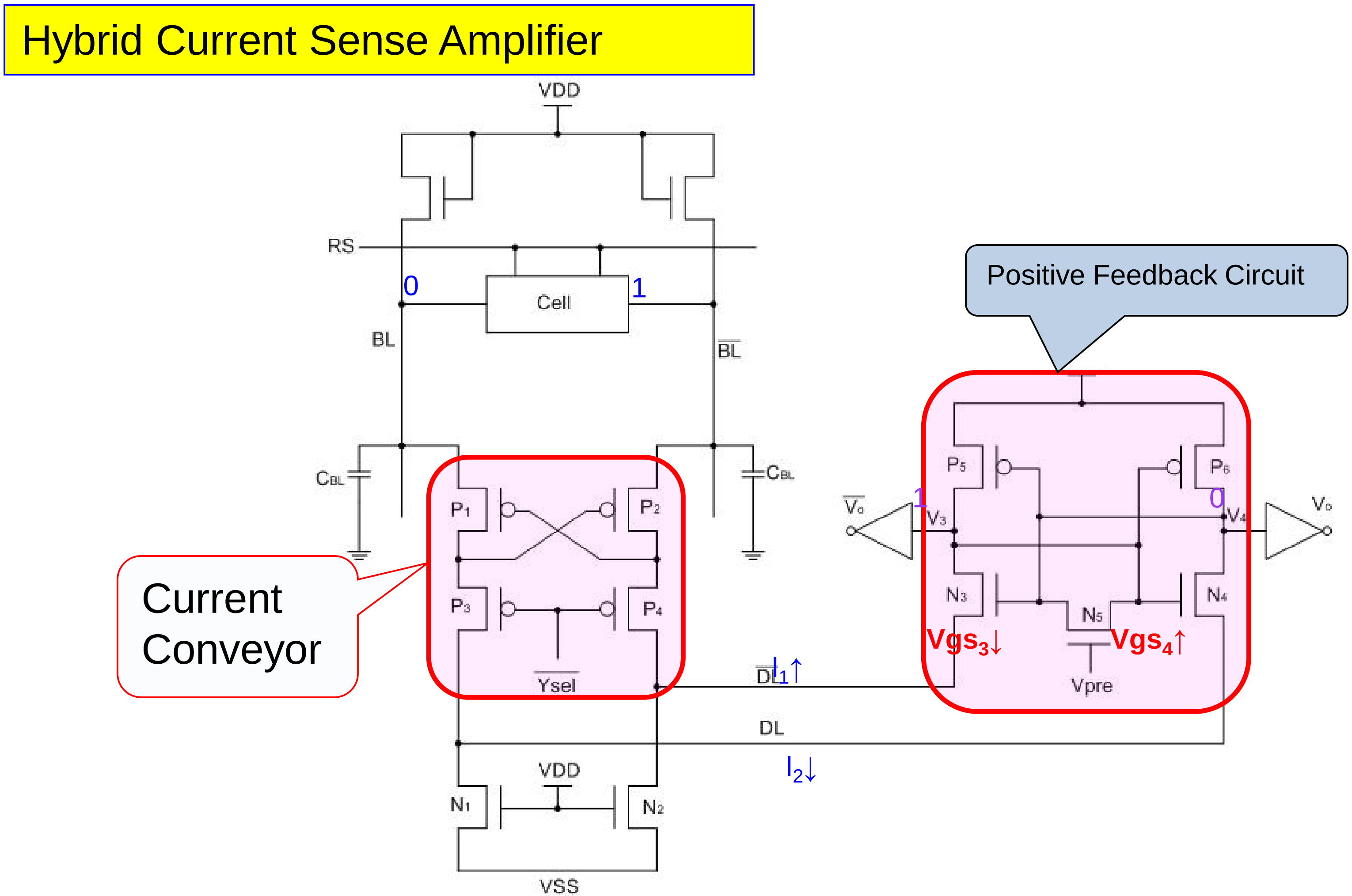
Simple Four Transistor Sense Amplifier (Current conveyor)

Ysel = 0 , Sense mode

Ysel = 1 , Standby mode



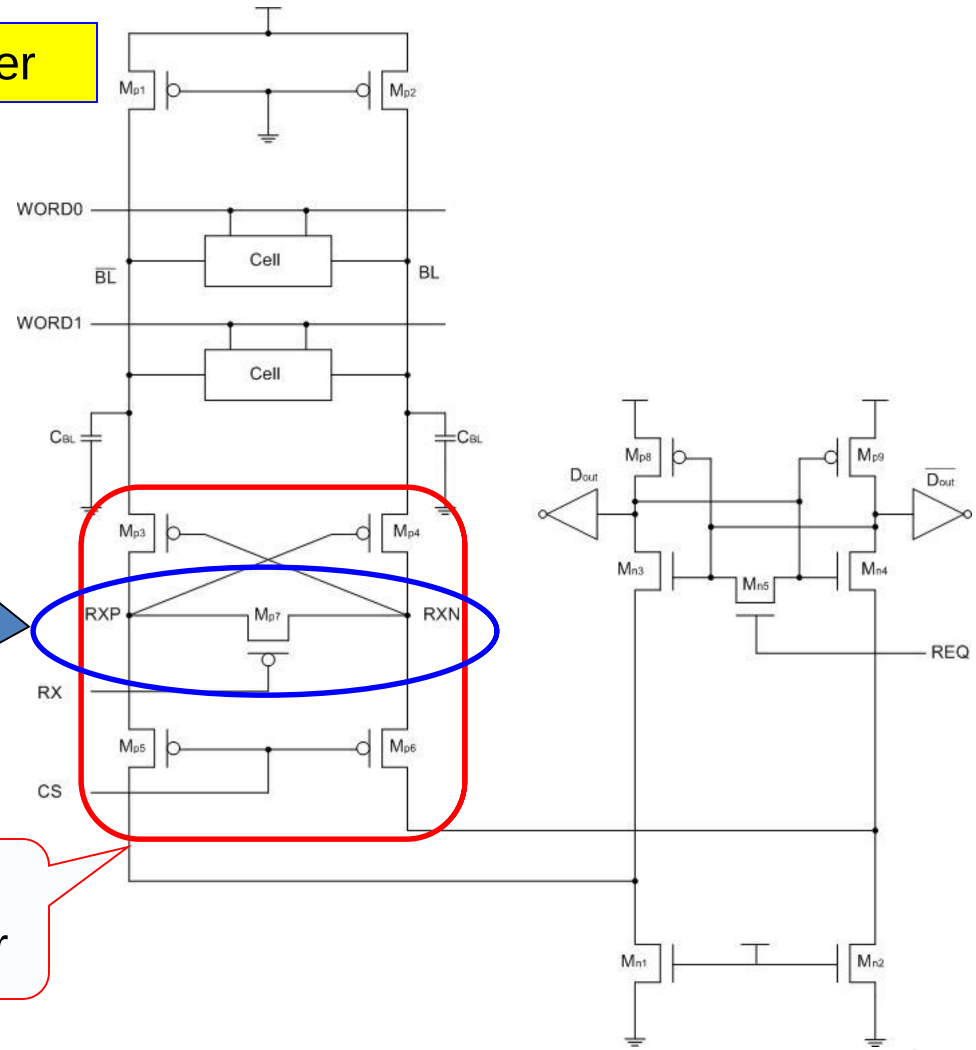
Current-Mode Sense Amplifier (cont.)



Current-Mode Sense Amplifier (cont.)

New Hybrid Current Sense Amplifier

Overcome the **pattern-dependent problem** of the conventional current conveyor. In conventional current conveyor, after each read operation, the nodes RXP and RXN get floated and there will exist a residual differential voltage between them. The pattern-dependent problem occurs when the same logic value is sequentially read out several times from the same column.



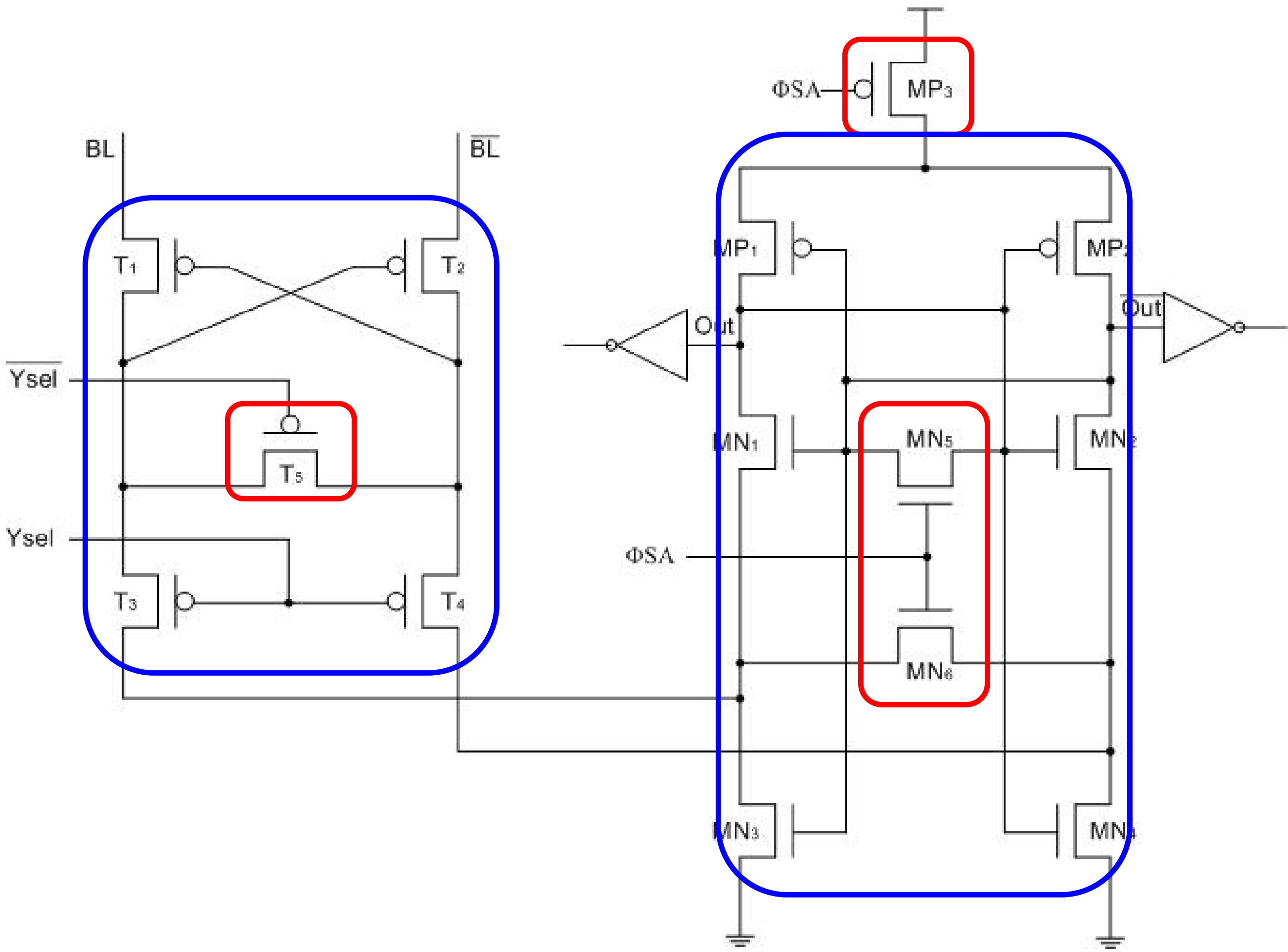
Current Conveyor

Sense Amplifier Design

- Current-mode is fast than voltage-mode sense amplifier.
- Positive feedback can fast amplify the small input signal to full supply voltage swing.
- Output nodes are separated from input nodes, that means the output nodes of the sense amplifier are no longer loaded with the bit-line capacitance.
- The input nodes of the sense amplifier are low-impedance current sensitive nodes to avoid the effect of the bit-line capacitance.
- Keeping the bit-line in small voltage swing to avoid power dissipation.

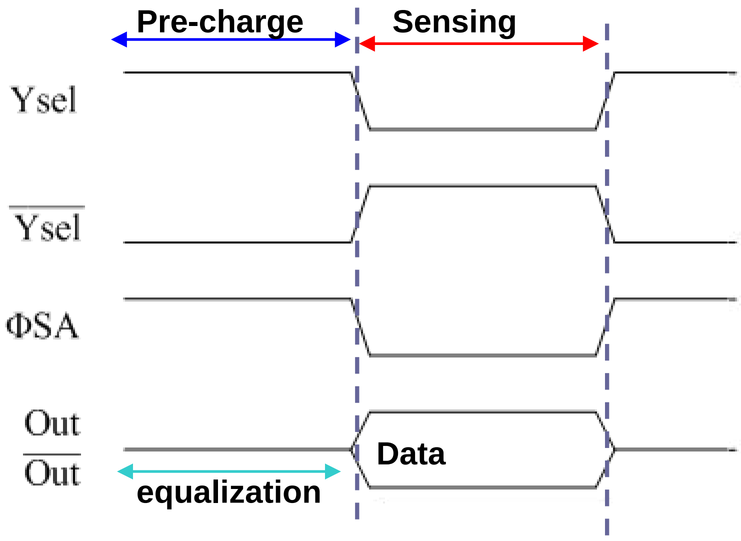
Sense Amplifier Design (cont.)

- **Switch circuit:**
  - T1~T4 → Current conveyor
  - T5 is used for equalization after sensing cycle to avoid pattern-dependent problem.
- **Positive feedback circuit:**
  - MP1~MP2 & MN1~MN4 construct a positive feedback circuit.
  - MN5 & MN6 is used for equalization after sensing cycle.
  - MP3 only turned on in sensing cycle to reduce power dissipation.



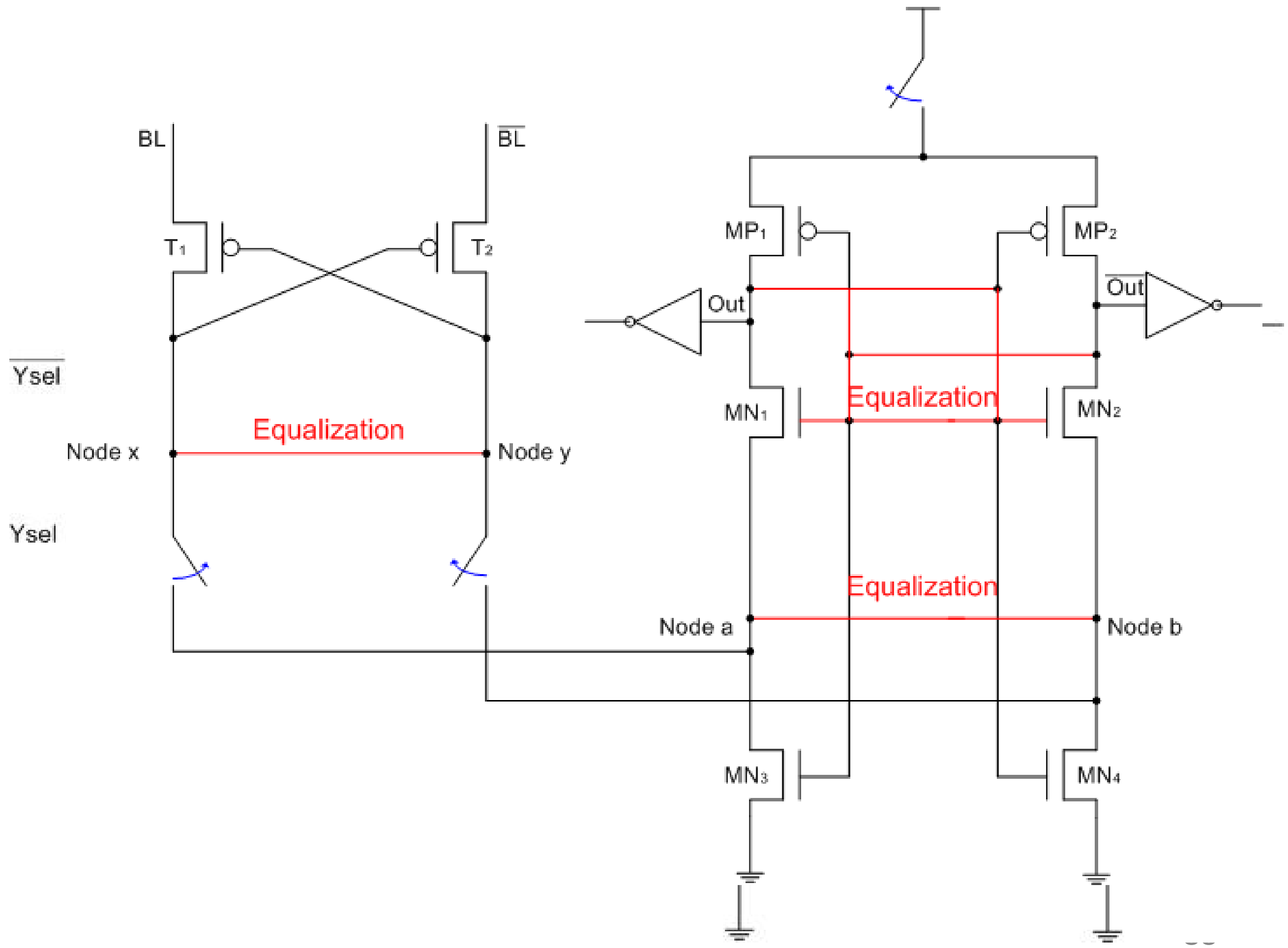
Sense Amplifier Design (cont.)

- Operating mode
  - Pre-charge
  - Sensing



Sense Amplifier Design (cont.)

- Pre-charge mode
  - $\overline{\text{Ysel}} = 0$
  - Ysel = 1
  - SAen = 1

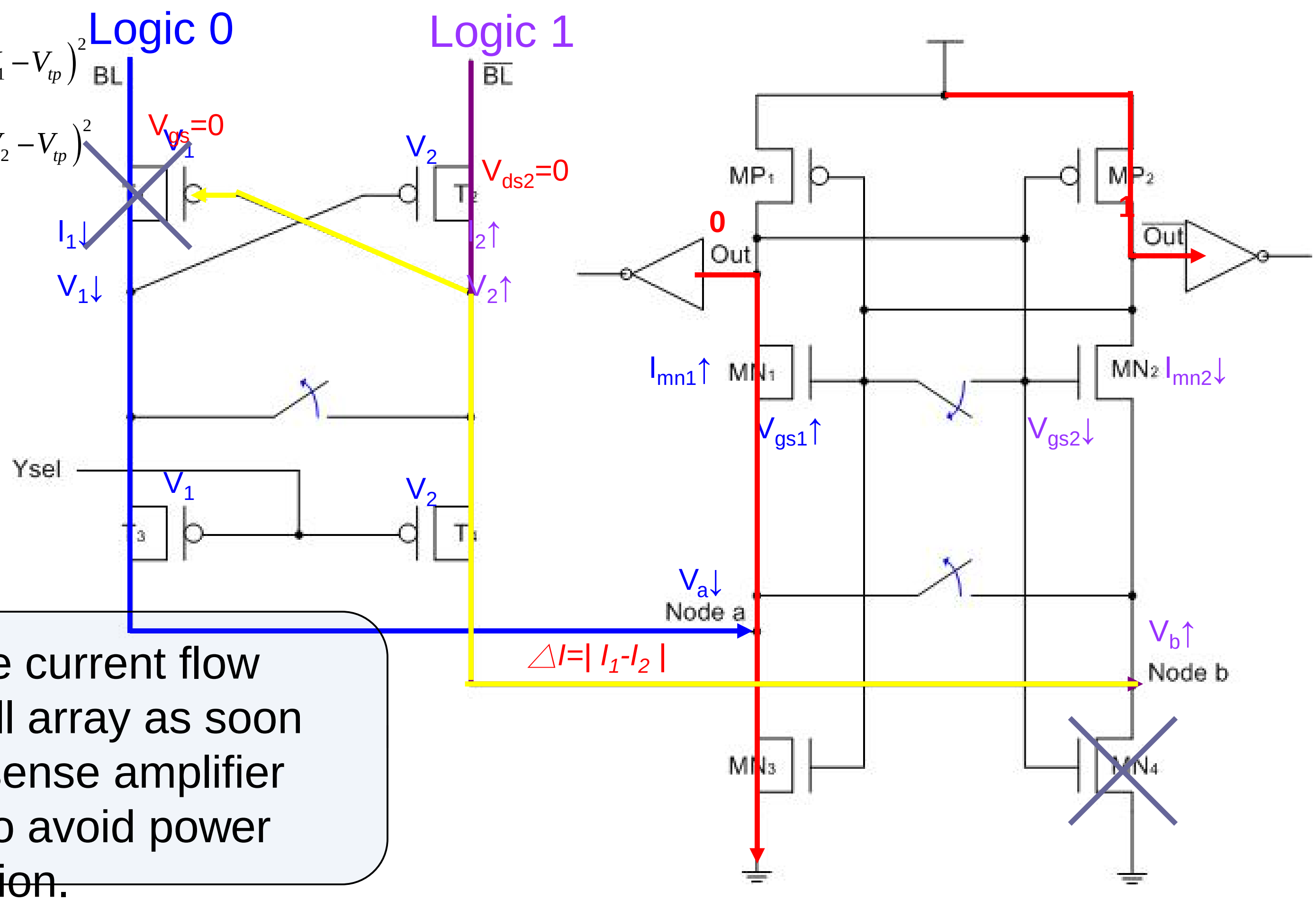


Sense Amplifier Design (cont.)

- Sensing mode

$$I_1 = \frac{1}{2} k_p' \left(\frac{W}{L}\right)_{T1} (V_1 - V_{tp})^2$$

$$I_2 = \frac{1}{2} k_p' \left(\frac{W}{L}\right)_{T2} (V_2 - V_{tp})^2$$



Stop the current flow from cell array as soon as the sense amplifier stable to avoid power dissipation.

Read-Write Memories (RAM)

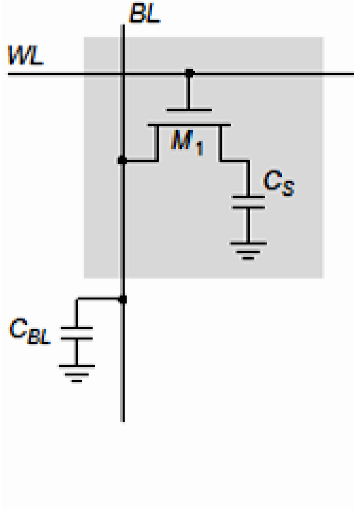
- Static (SRAM)
  - Data stored as long as supply is applied
  - Large (6 transistors/cell)
  - Fast
  - Differential
- Dynamic (DRAM)
  - Periodic refresh required
  - Small (1-3 transistors/cell)
  - Slower
  - Single Ended

1-Transistor DRAM Cell

– Write: Cs is charged or discharged by asserting WL and BL

– Read: Charge redistribution takes place between bit line and storage capacitance

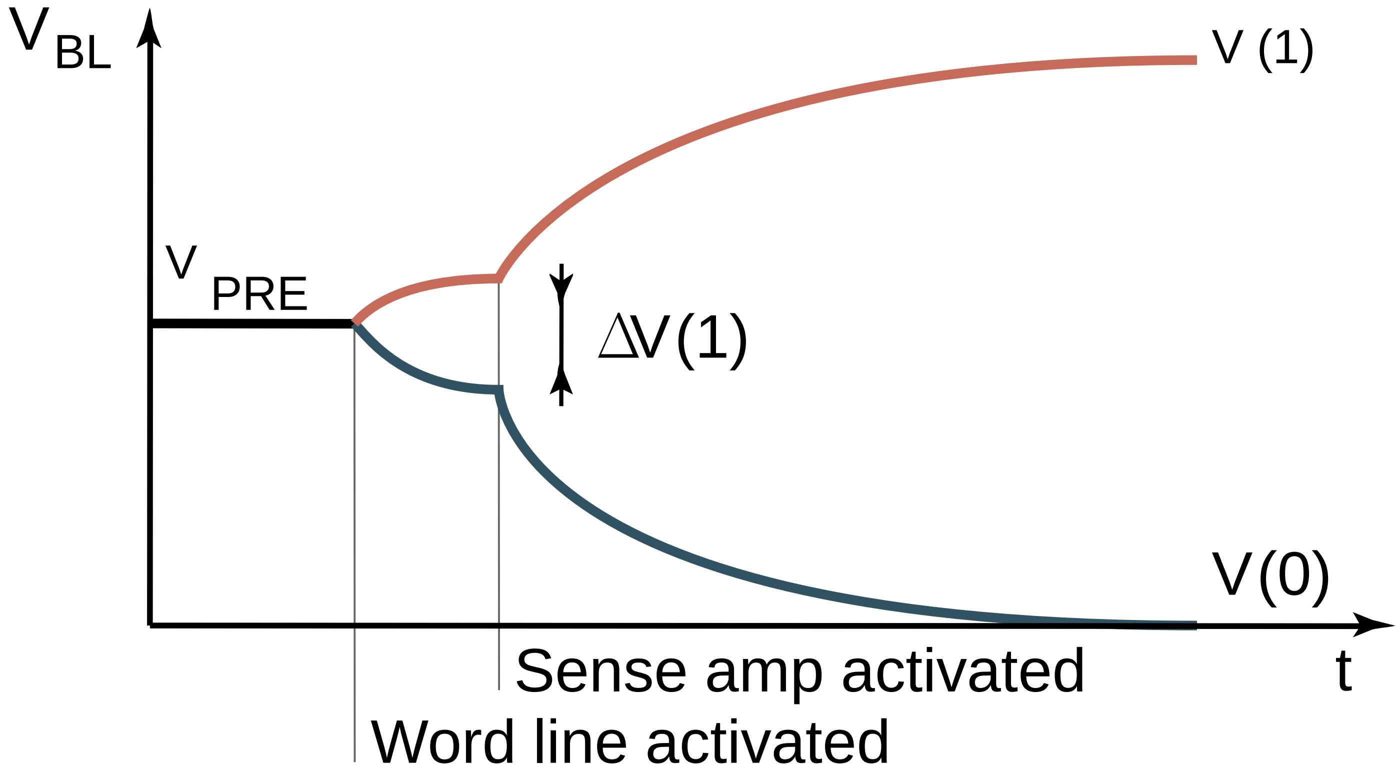
– Voltage swing is small; typically around 250 mV



DRAM Cell Observations

- 1T DRAM requires a sense amplifier for each bit line, due to charge redistribution read-out.
- DRAM memory cells are single ended in contrast to SRAM cells.
- The read-out of the 1T DRAM cell is destructive; read and refresh operations are necessary for correct operation.
- 1T cell requires presence of an extra capacitance that must be explicitly included in the design.
- When writing a "1" into a DRAM cell, a threshold voltage is lost. This charge loss can be circumvented by bootstrapping the word lines to a higher value than V_{DD}

Sense Amp Operation



DRAM Read

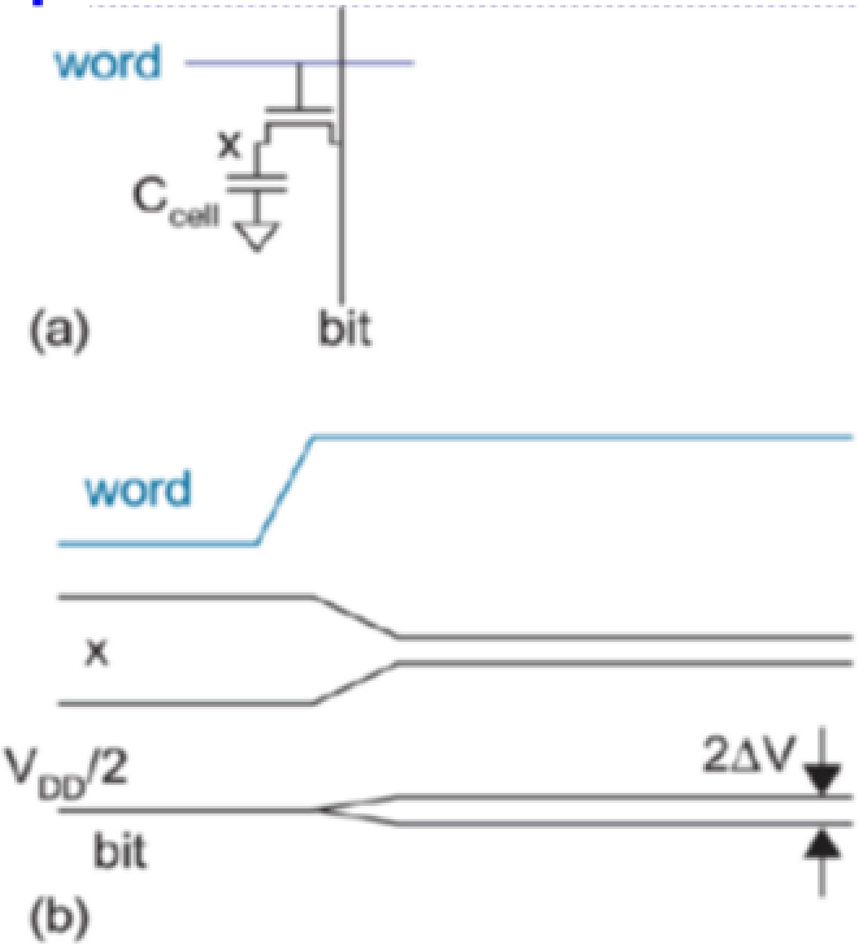


FIG 11.26 DRAM cell read operation

1. bitline precharged to $V_{DD}/2$

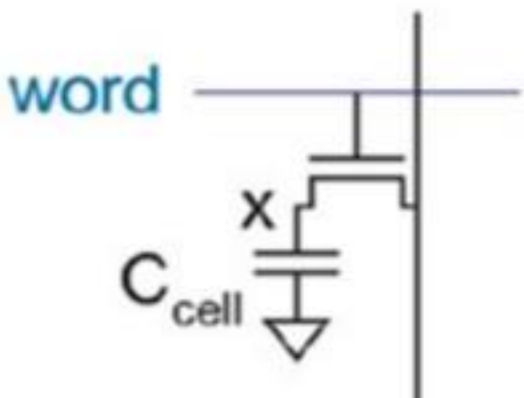
2. wordline rises, cap. shares it charge with bitline, causing a voltage ΔV

3. read disturbs the cell content at x, so the cell must be rewritten after each read

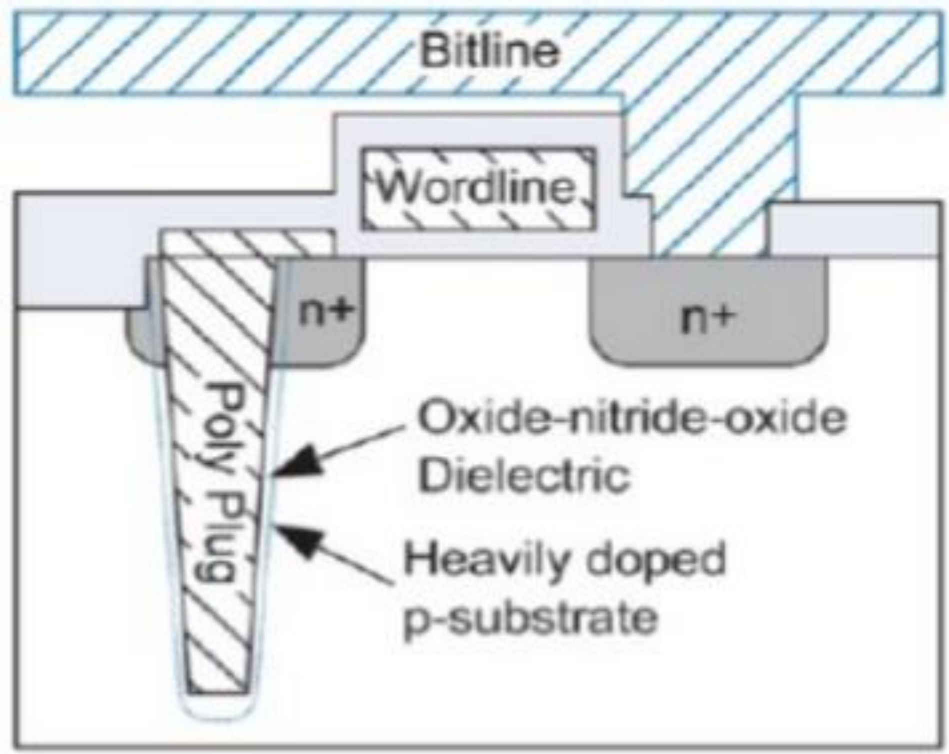
$$\Delta V = \frac{V_{DD}}{2} \frac{C_{cell}}{C_{cell} + C_{bit}}$$

DRAM: Dynamic RAM

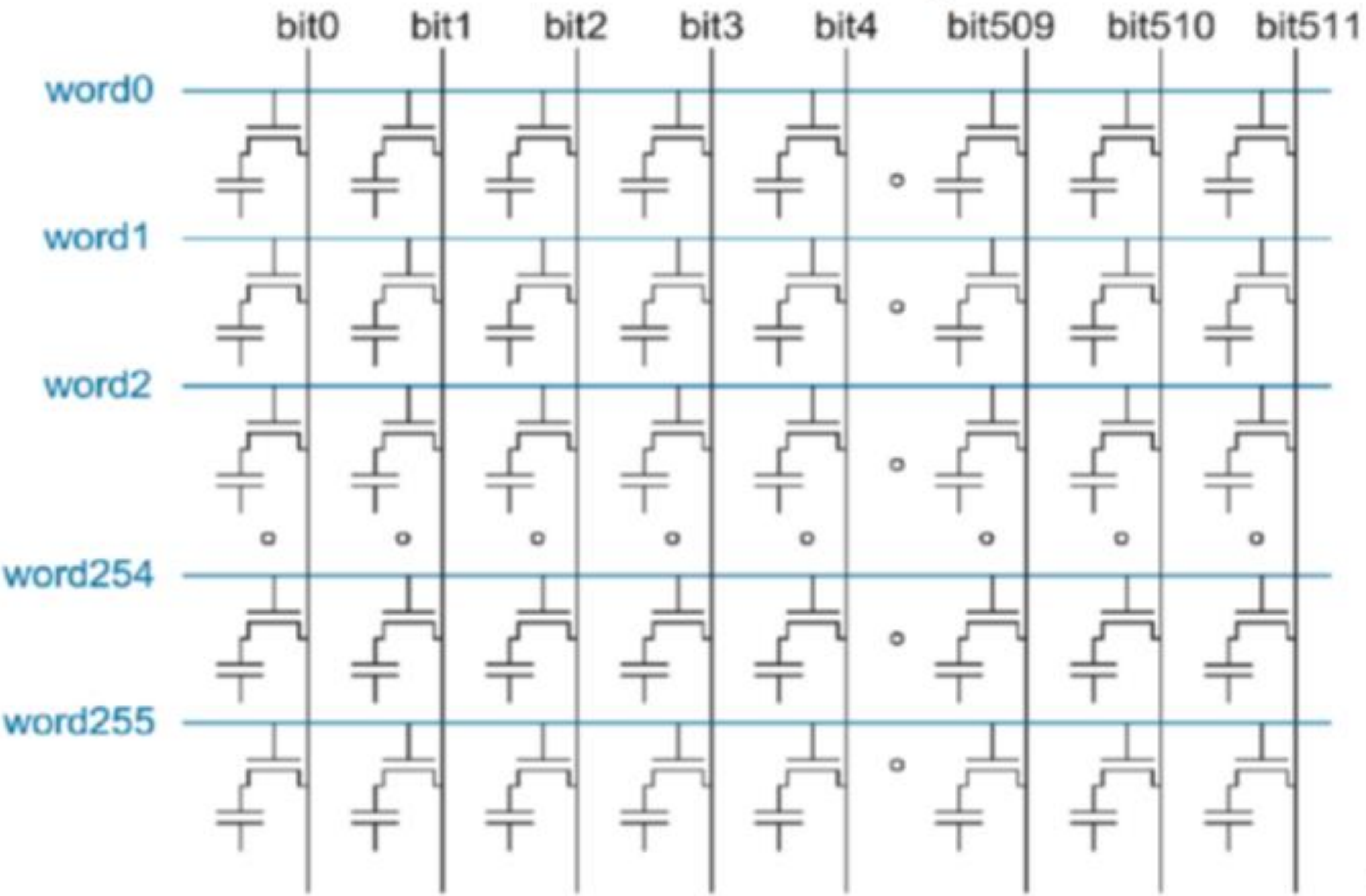
- Store their contents as charge on a capacitor rather than in a feedback loop.
- 1T dynamic RAM cell has a transistor and a capacitor



(a)



DRAM Array



bit0
bit1
bit2
bit3
bit4
bit509
bit510
bit511
word0
word1
word2
word254
word255

DRAM

- With large size, the bitline cap is an order of magnitude higher than in the cell, causing very small voltage swing.
- A sense amplifier is used.
- Three different bitline architectures, open, folded, and twisted, offer different compromises between noises and area.

DRAM Timing

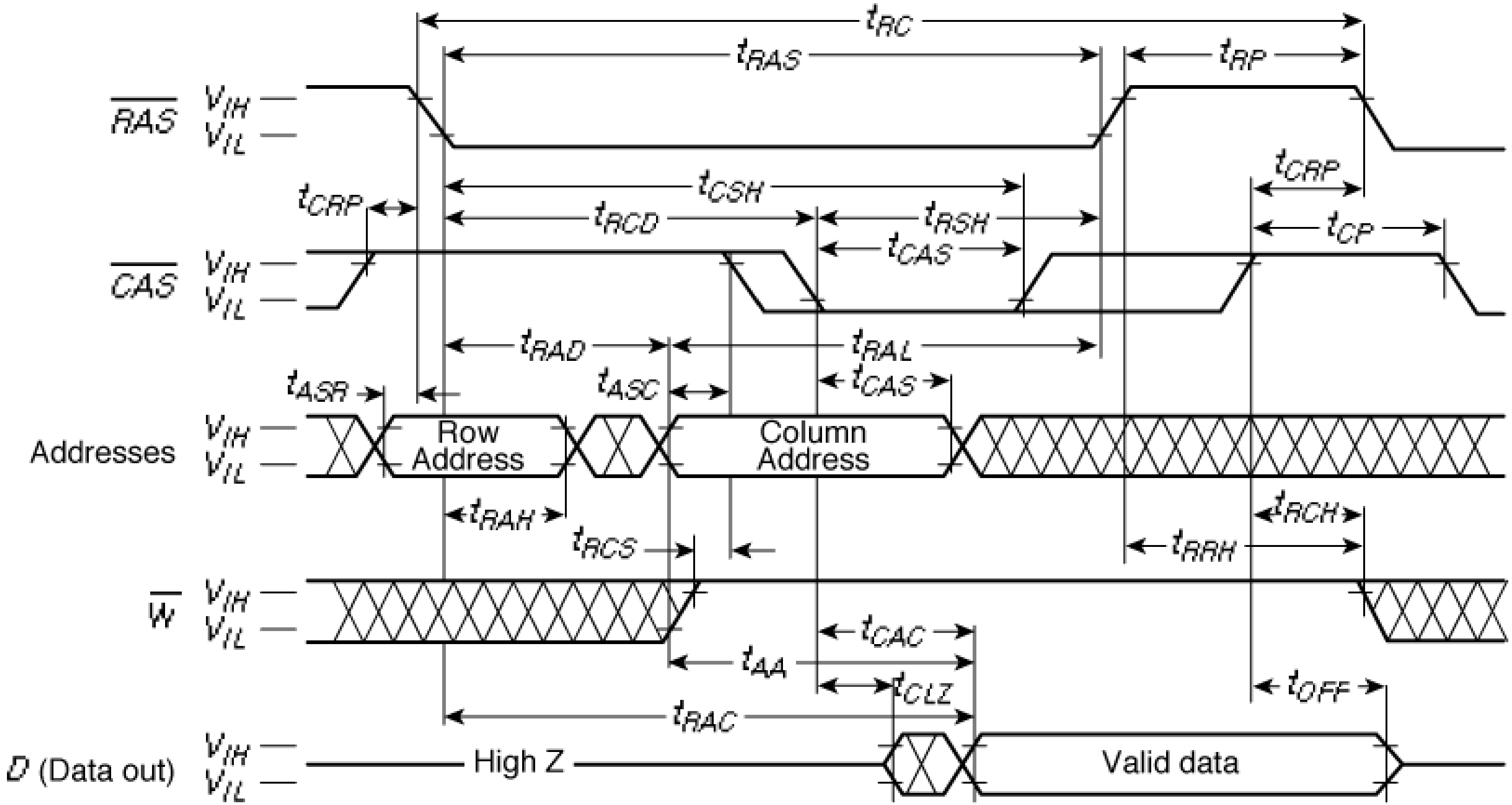
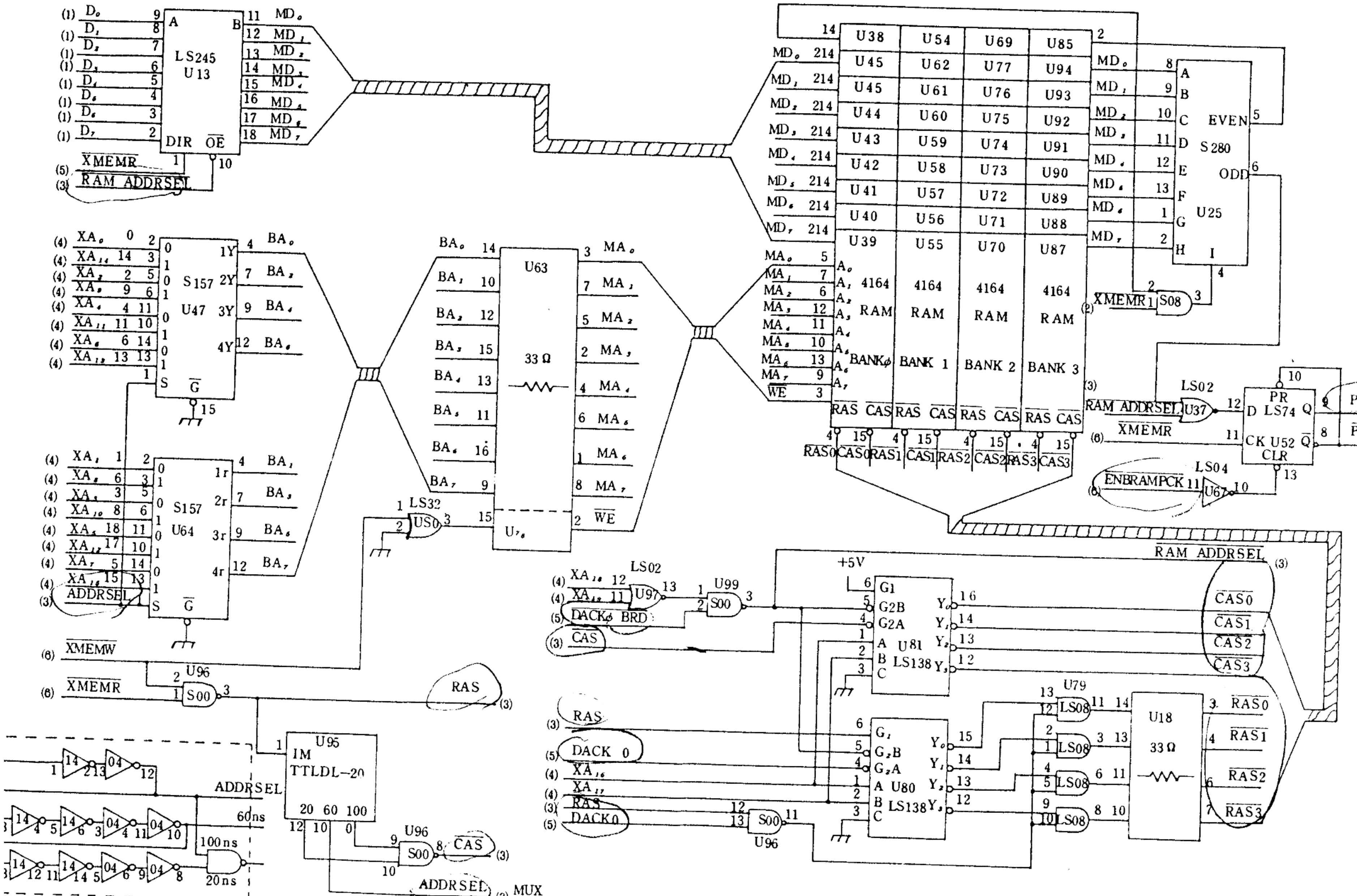

t_{RC}
t_{RAS}
t_{RP}
$\overline{RAS}$
V_{IH}
V_{IL}
t_{CSH}
t_{CRP}
t_{RCD}
t_{RSH}
t_{CAS}
t_{CP}
$\overline{CAS}$
t_{RAD}
t_{RAL}
t_{ASR}
t_{ASC}
Addresses
Row Address
Column Address
t_{RAH}
t_{RCS}
t_{RCH}
t_{RRH}
$\overline{W}$
t_{CAC}
t_{AA}
t_{CLZ}
t_{OFF}
t_{RAC}
D (Data out)
High Z
Valid data

Fig 2


LS245
U13
DIR
OE
XMEMR
RAM ADDRSEL
MD0 MD1 MD2 MD3 MD4 MD5 MD6 MD7
S157
U47
S157
U64
ADDRSEL
U63
33 Ω
WE
LS32
U50
XMEMW
XMEMR
U96
S00
RAS
U95
IM
TTLDL-20
20 60 100
CAS
ADDRSEL
MUX
60ns
100ns
20ns
U38 U54 U69 U85
U45 U62 U77 U94
U45 U61 U76 U93
U44 U60 U75 U92
U43 U59 U74 U91
U42 U58 U73 U90
U41 U57 U72 U89
U40 U56 U71 U88
U39 U55 U70 U87
4164 RAM
BANKφ
BANK 1
BANK 2
BANK 3
RAS0 CAS0 RAS1 CAS1 RAS2 CAS2 RAS3 CAS3
EVEN
S280
ODD
U25
S08
LS02
U37
PR
LS74
U52
CK
CLR
LS04
U67
ENBRAMPCK
+5V
LS02
U97
U99
DACKφ BRD
G1
G2B
G2A
U81
LS138
U80
LS138
DACK 0
U79
LS08
U18
33 Ω
RAS0
RAS1
RAS2
RAS3

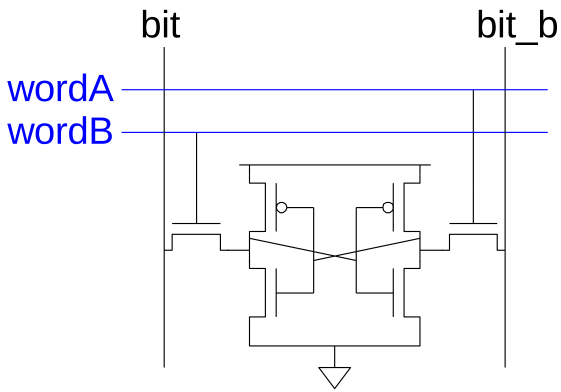
系統主機板(3)

Multiple Ports

- We have considered single-ported SRAM
  - One read or one write on each cycle
- *Multiported* SRAM are needed for register files
- Examples:
  - Multicycle MIPS must read two sources or write a result on some cycles
  - Pipelined MIPS must read two sources and write a third result each cycle
  - Superscalar MIPS must read and write many sources and results each cycle

Dual-Ported SRAM

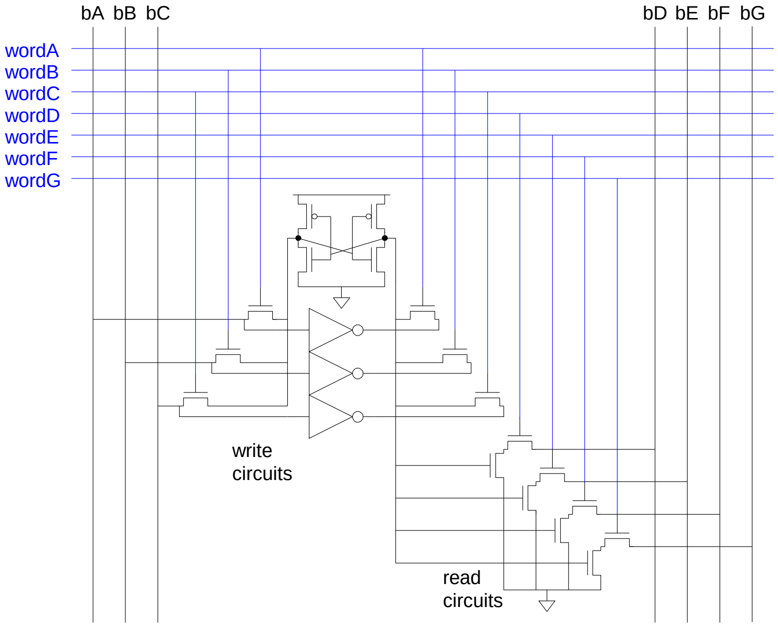
– Simple dual-ported SRAM
  - Two independent single-ended reads
  - Or one differential write



– Do two reads and one write by time multiplexing
  - Read during ph1, write during ph2

Multi-Ported SRAM

- Adding more access transistors hurts read stability
- Multiported SRAM isolates reads from state node
- Single-ended design minimizes number of bitlines

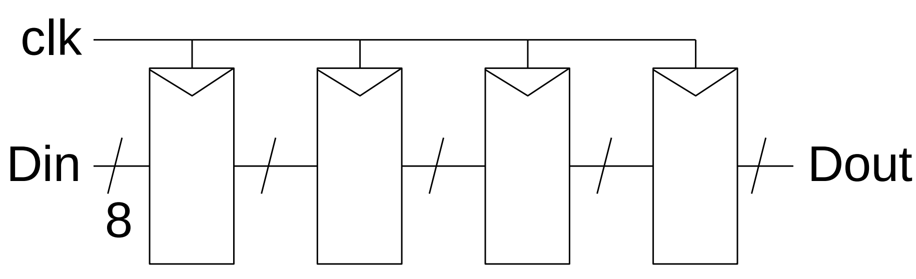


Serial Access Memories

- Serial access memories do not use an address
  - Shift Registers
  - Tapped Delay Lines
  - Serial In Parallel Out (SIPO)
  - Parallel In Serial Out (PISO)
  - Queues (FIFO, LIFO)

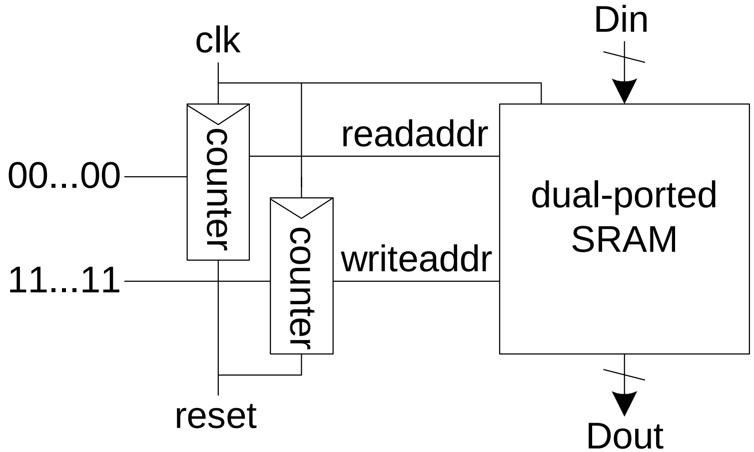
Shift Register

- *Shift registers* store and delay data
- Simple design: cascade of registers
  – Watch your hold times!



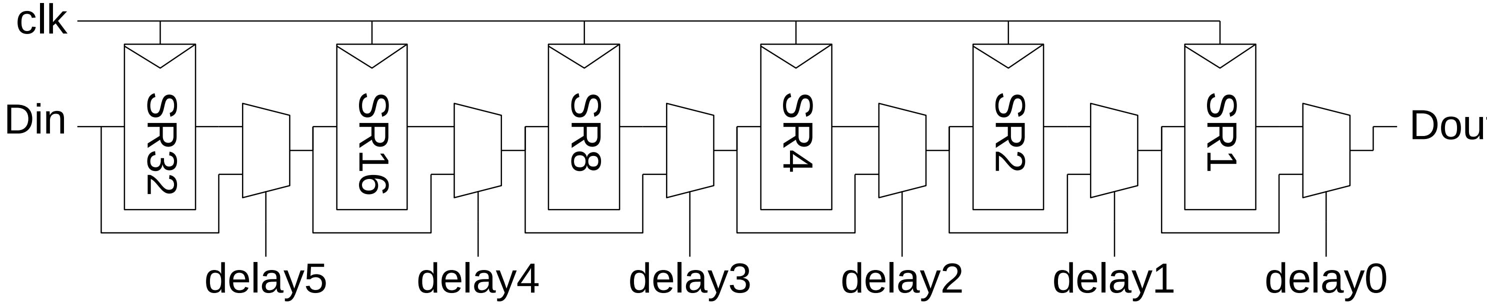
Denser Shift Registers

– Flip-flops aren’ t very area-efficient
– For large shift registers, keep data in SRAM instead
– Move read/write pointers to RAM rather than data
  - Initialize read address to first entry, write to last
  - Increment address on each cycle



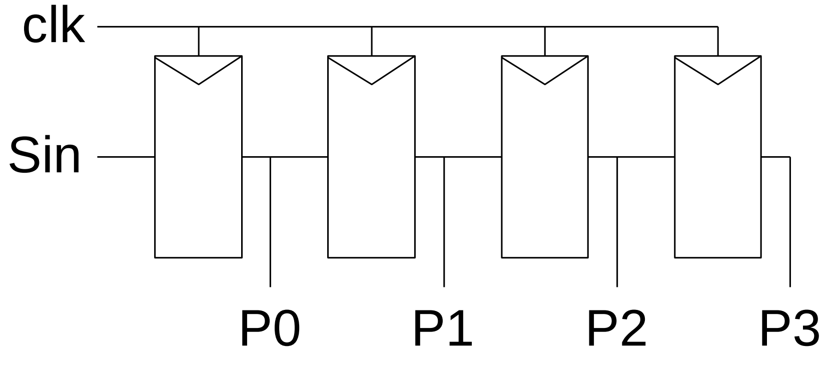
Tapped Delay Line

- A *tapped delay line* is a shift register with a programmable number of stages
- Set number of stages with delay controls to mux
  - Ex: 0 – 63 stages of delay



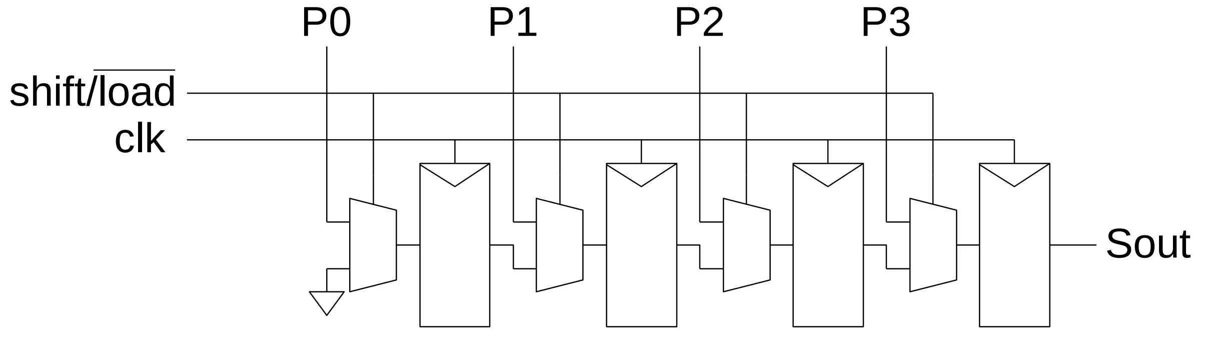
Serial In Parallel Out

- 1-bit shift register reads in serial data
  - After N steps, presents N-bit parallel output



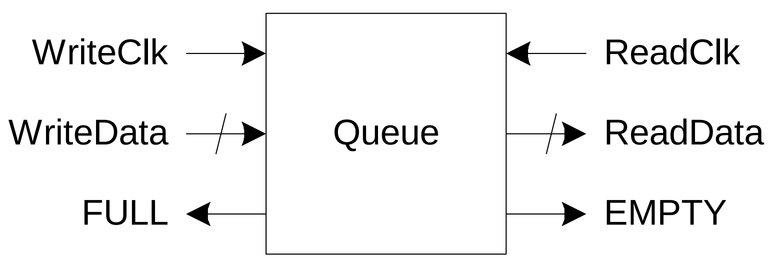
Parallel In Serial Out

- Load all N bits in parallel when shift = 0
  - Then shift one bit out per cycle



Queues

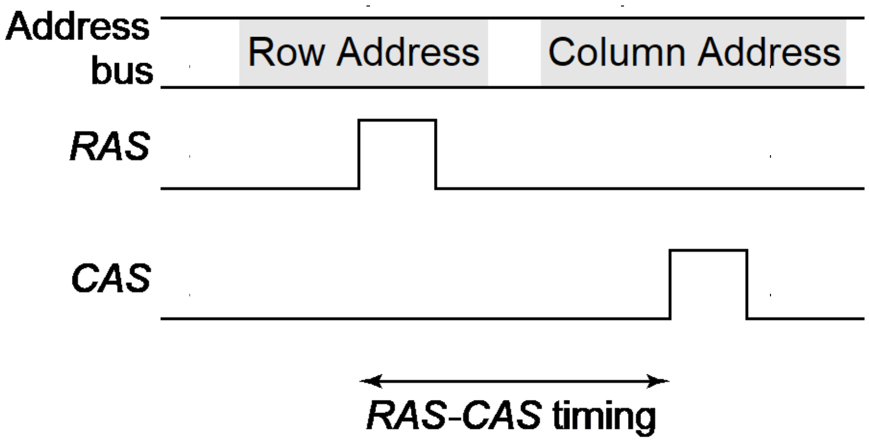
- *Queues* allow data to be read and written at different rates.
- Read and write each use their own clock, data
- Queue indicates whether it is full or empty
- Build with SRAM and read/write counters (pointers)



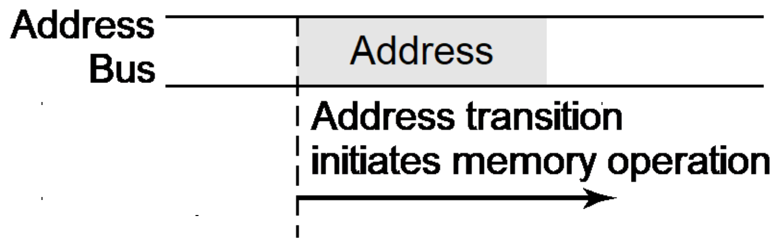
FIFO, LIFO Queues

- *First In First Out* (FIFO)
  - Initialize read and write pointers to first element
  - Queue is EMPTY
  - On write, increment write pointer
  - If write almost catches read, Queue is FULL
  - On read, increment read pointer
- *Last In First Out* (LIFO)
  - Also called a *stack*
  - Use a single *stack pointer* for read and write

Memory Timing: Approaches



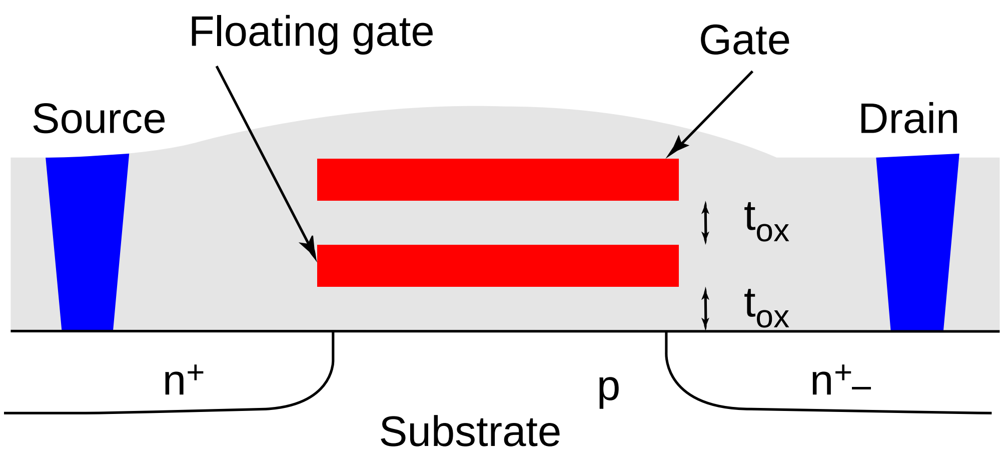
DRAM Timing
Multiplexed Adressing



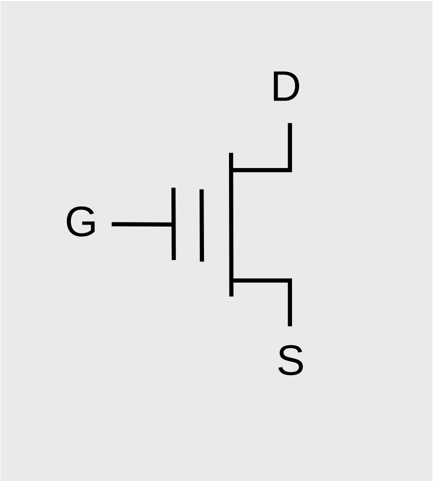
SRAM Timing
Self-timed

Non-Volatile Memories

- Floating-gate transistor

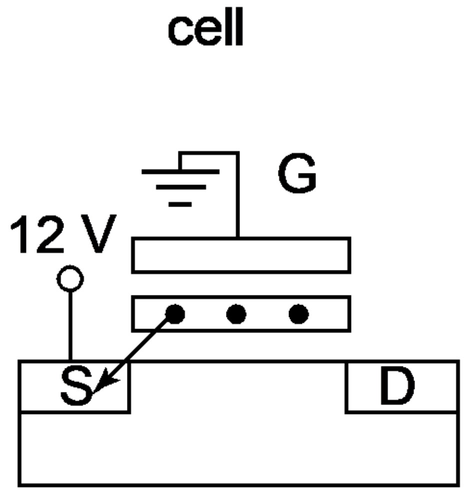


Device cross-section

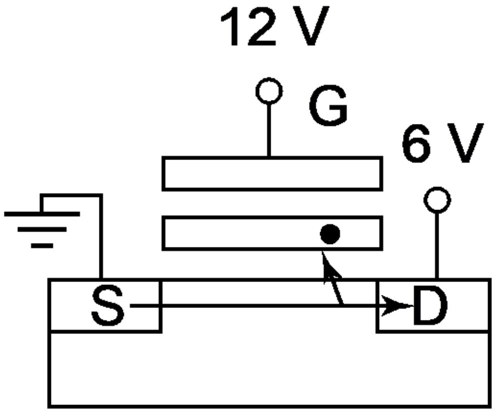


Schematic symbol

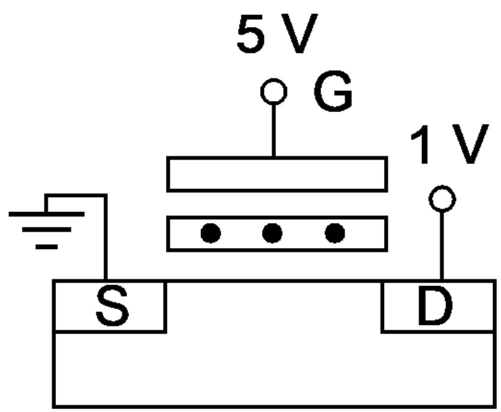
NOR Flash Operations —Erase



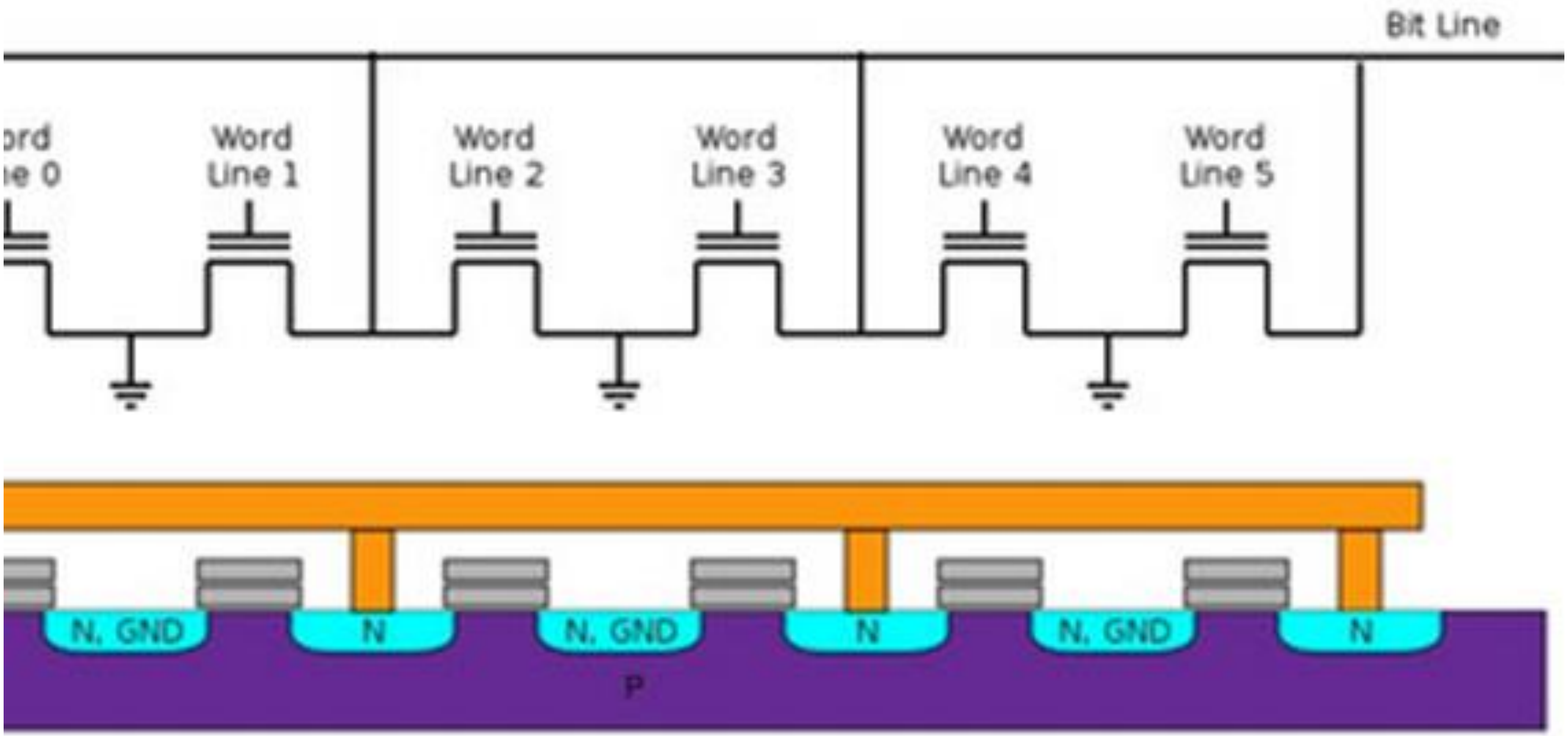
NOR Flash Operations —Program



NOR Flash Operations —Read

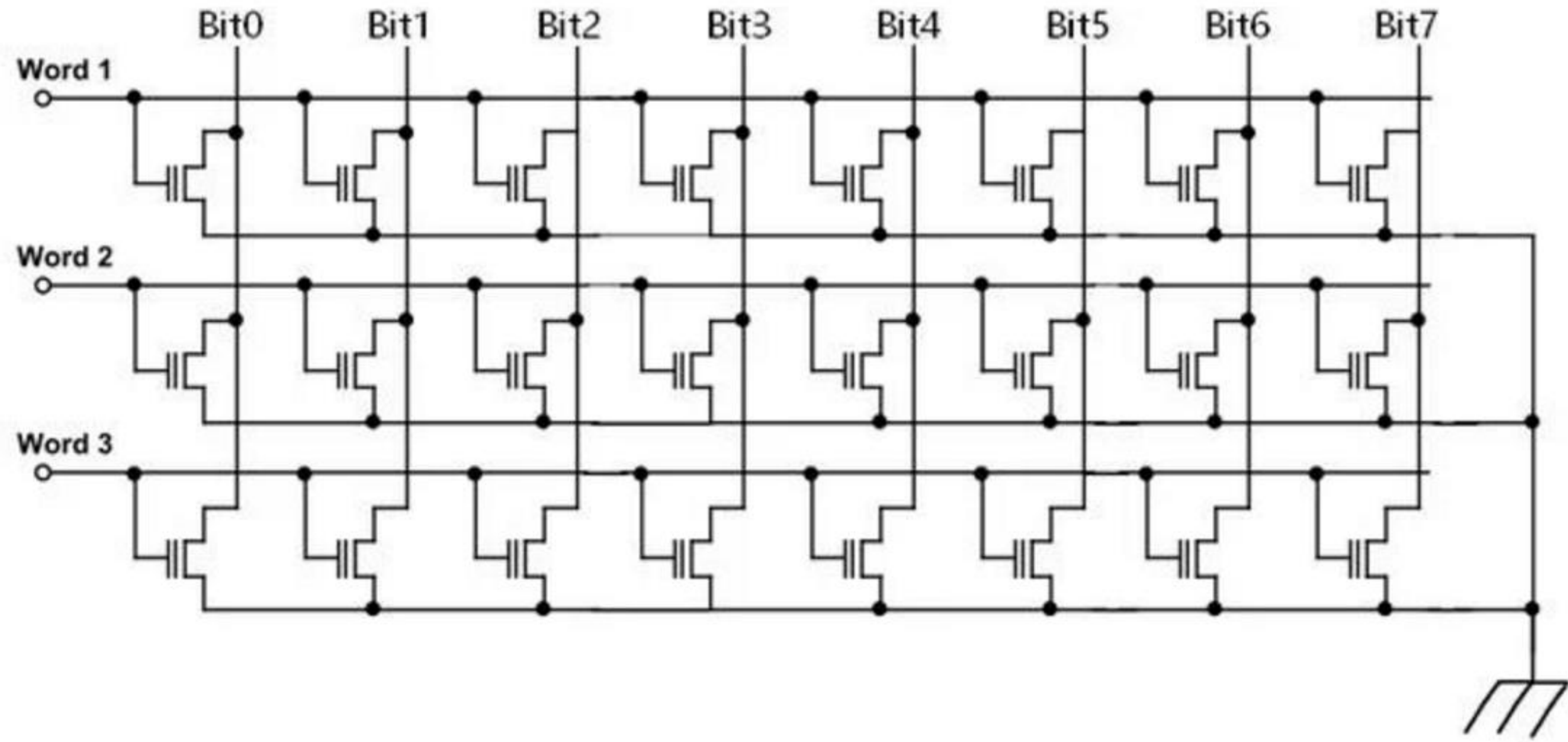


NOR Flash

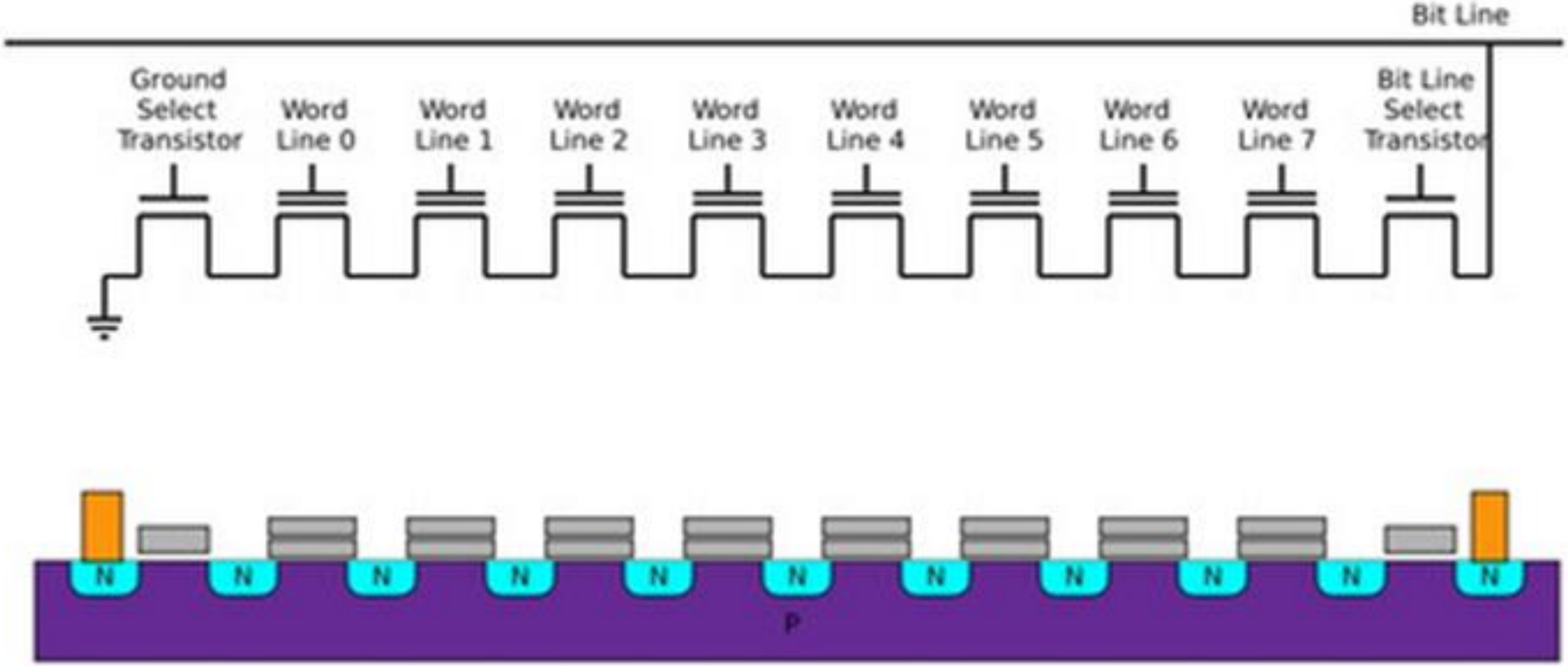


Bit Line
ord
e 0
Word
Line 1
Word
Line 2
Word
Line 3
Word
Line 4
Word
Line 5
N, GND
N
N, GND
N
N, GND
N
P

NOR Flash



NAND Flash


Bit Line
Ground Select Transistor
Word Line 0
Word Line 1
Word Line 2
Word Line 3
Word Line 4
Word Line 5
Word Line 6
Word Line 7
Bit Line Select Transistor
N
N
N
N
N
N
N
N
N
N
N
P

NAND Flash