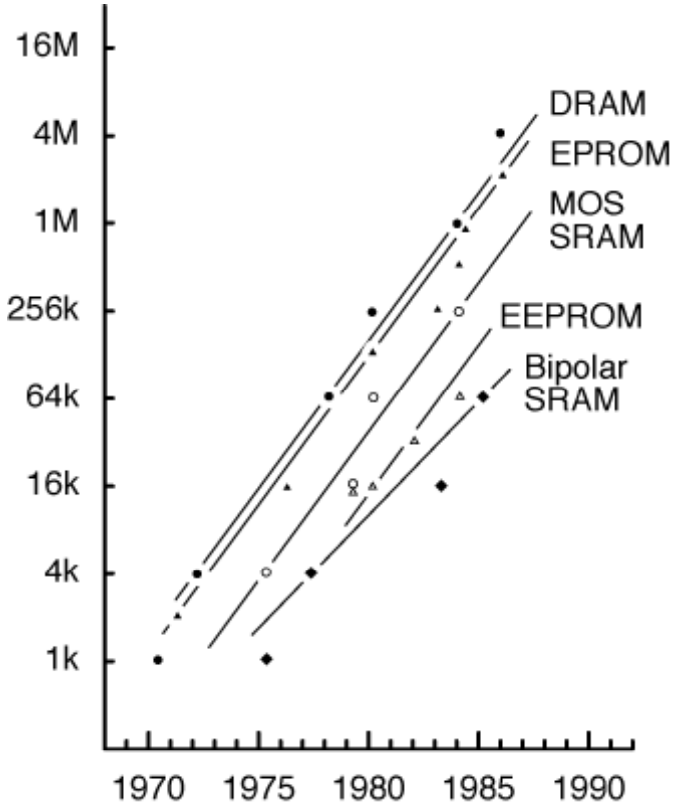
Semiconductor Memories: an Introduction

Talk Overview

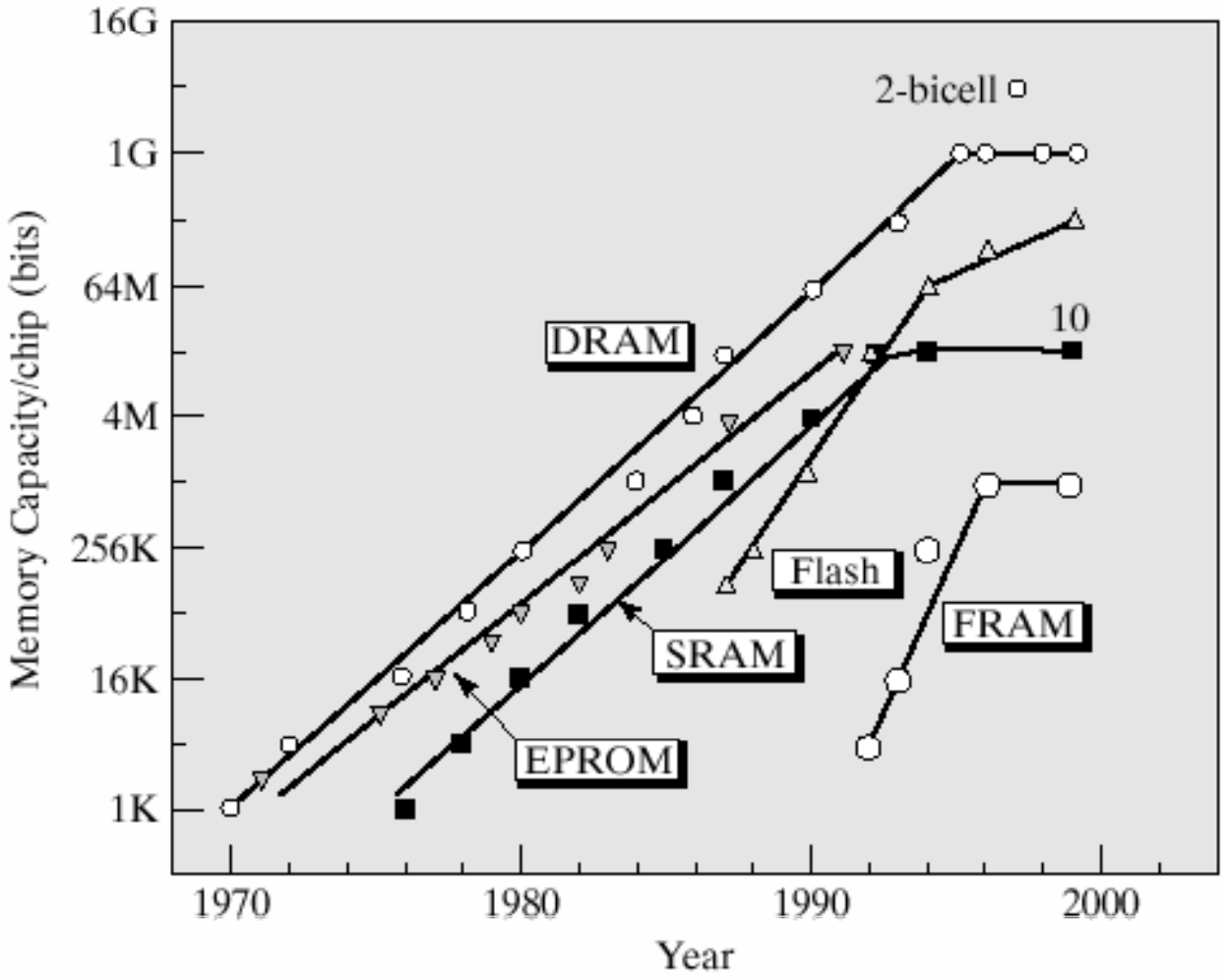
- ❑ Memory Trend
- ❑ Memory Classification
- ❑ Memory Architectures
- ❑ The Memory Core
- ❑ Periphery
- ❑ Reliability

Semiconductor Memory Trends (up to the 90's)

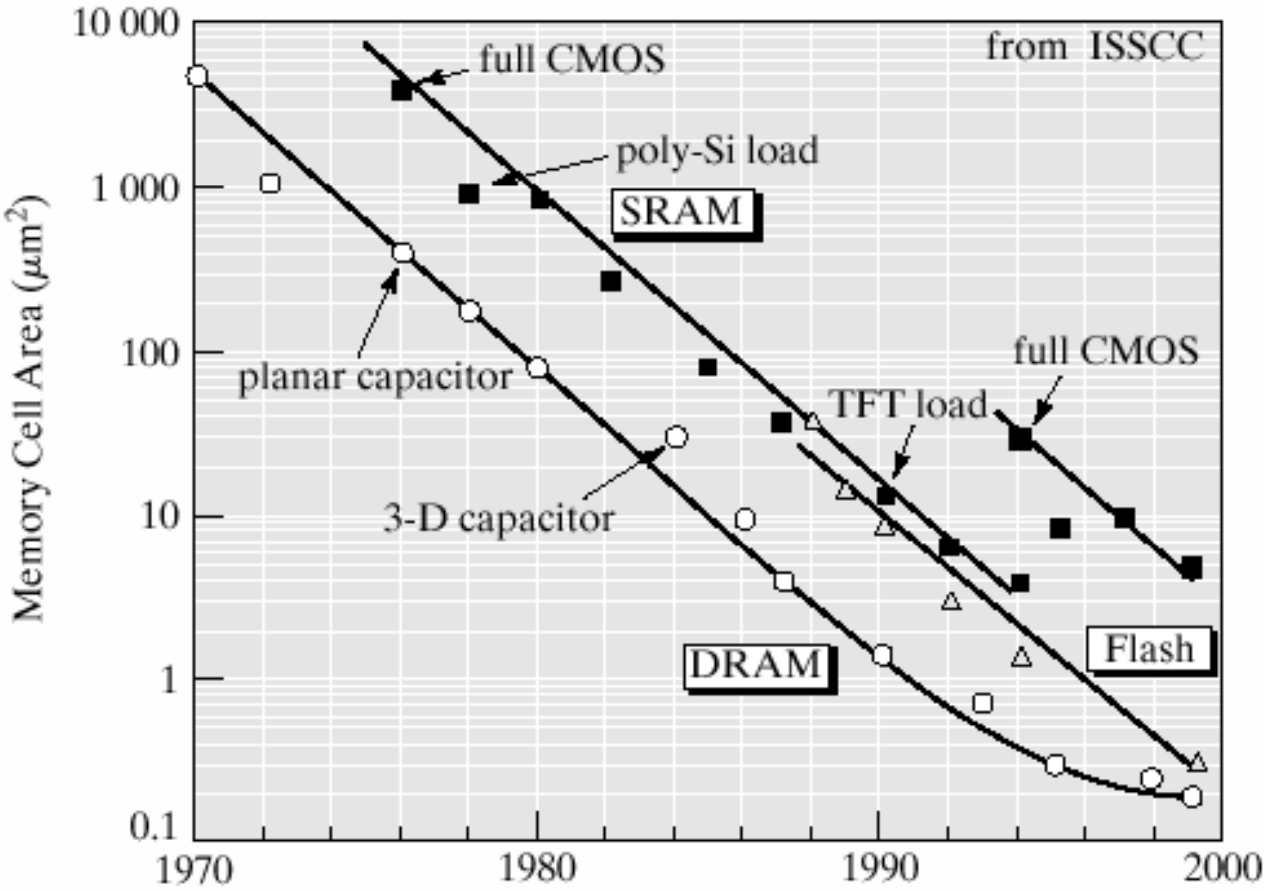


Memory Size as a function of time: x 4 every three years

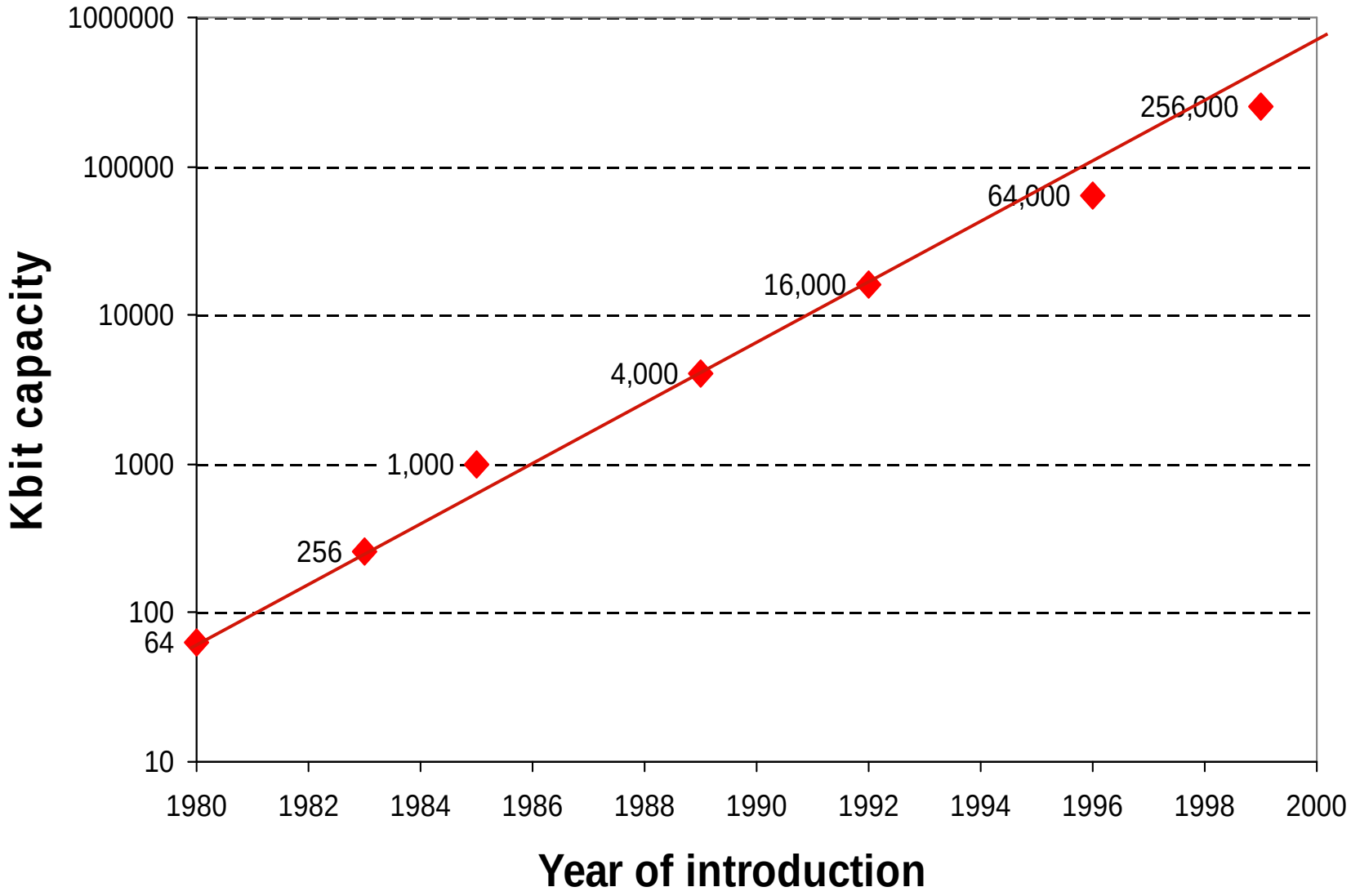
Semiconductor Memory Trends (updated)



Trends in Memory Cell Area



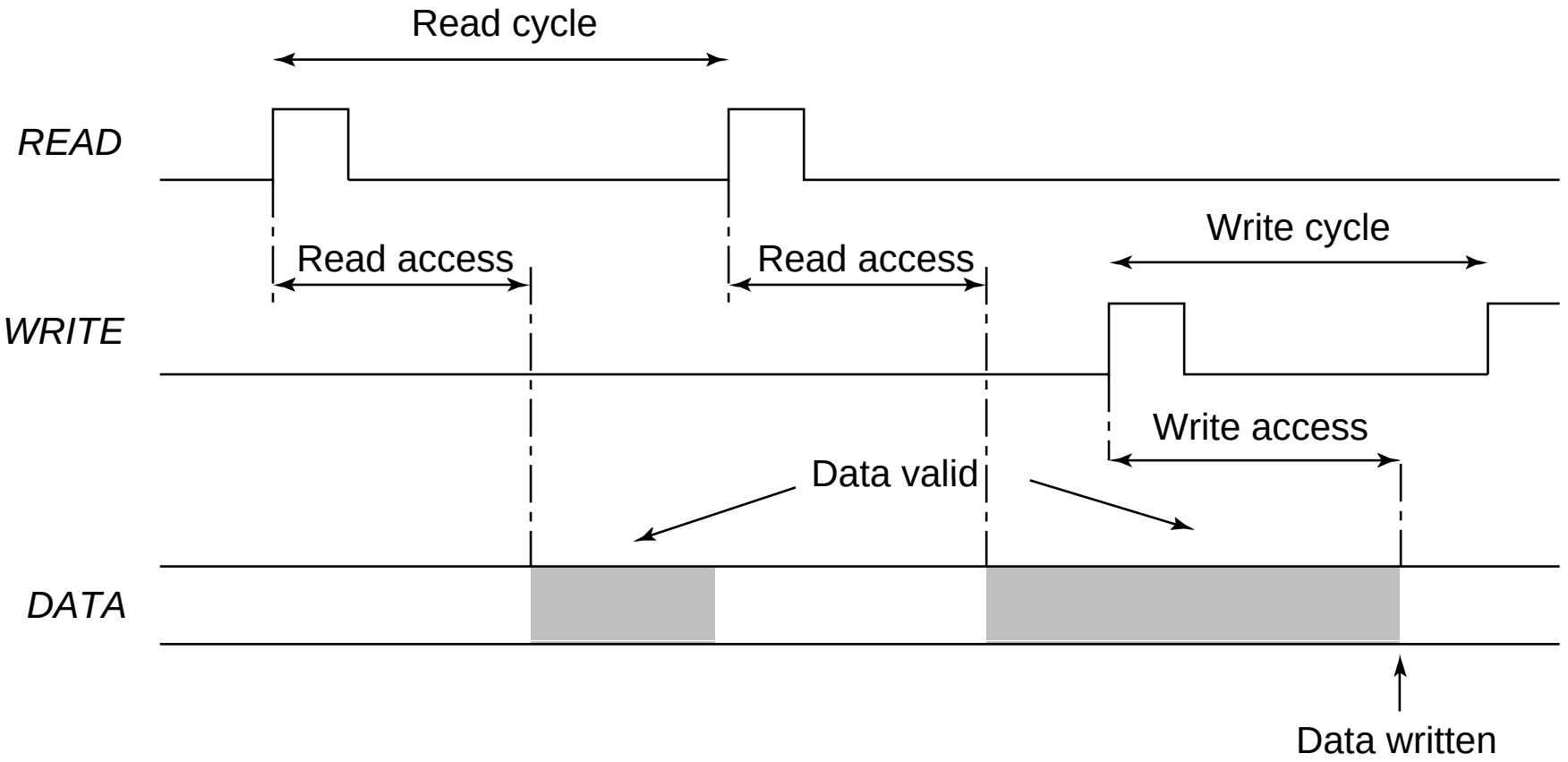
Growth in DRAM Chip Capacity



Semiconductor Memory Classification

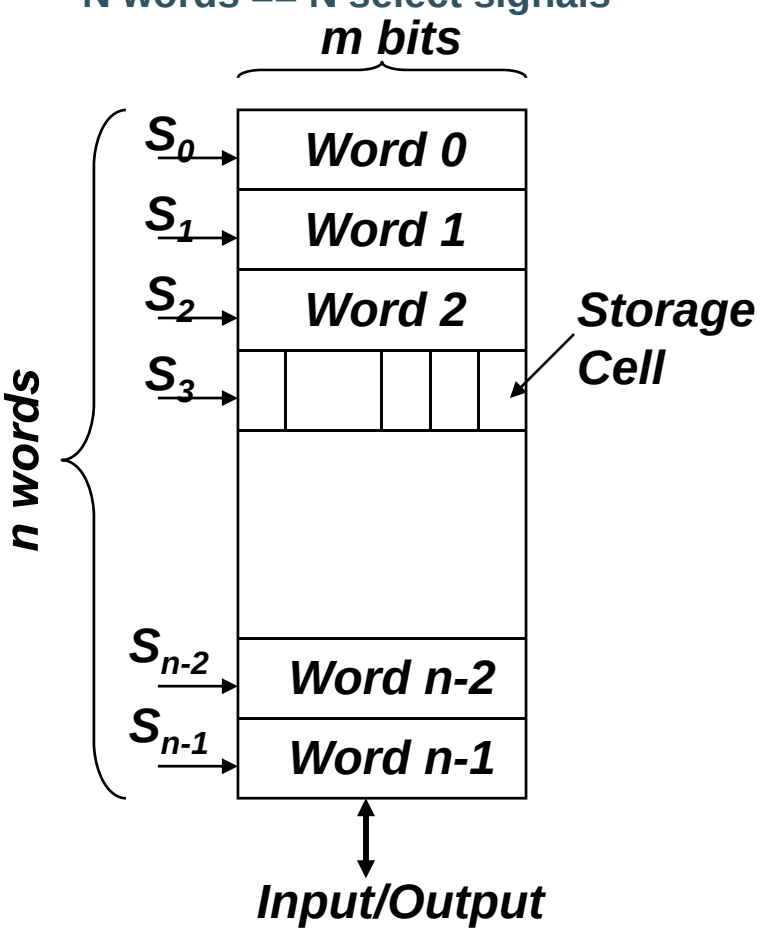
Read-Write Memory		Non-Volatile Read-Write Memory	Read-Only Memory
Random Access	**Non-Random Access**	EPROM E^2PROM FLASH MRAM PRAM RRAM	Mask-Programmed Programmable (PROM)
DRAM SRAM	FIFO LIFO Shift Register CAM		

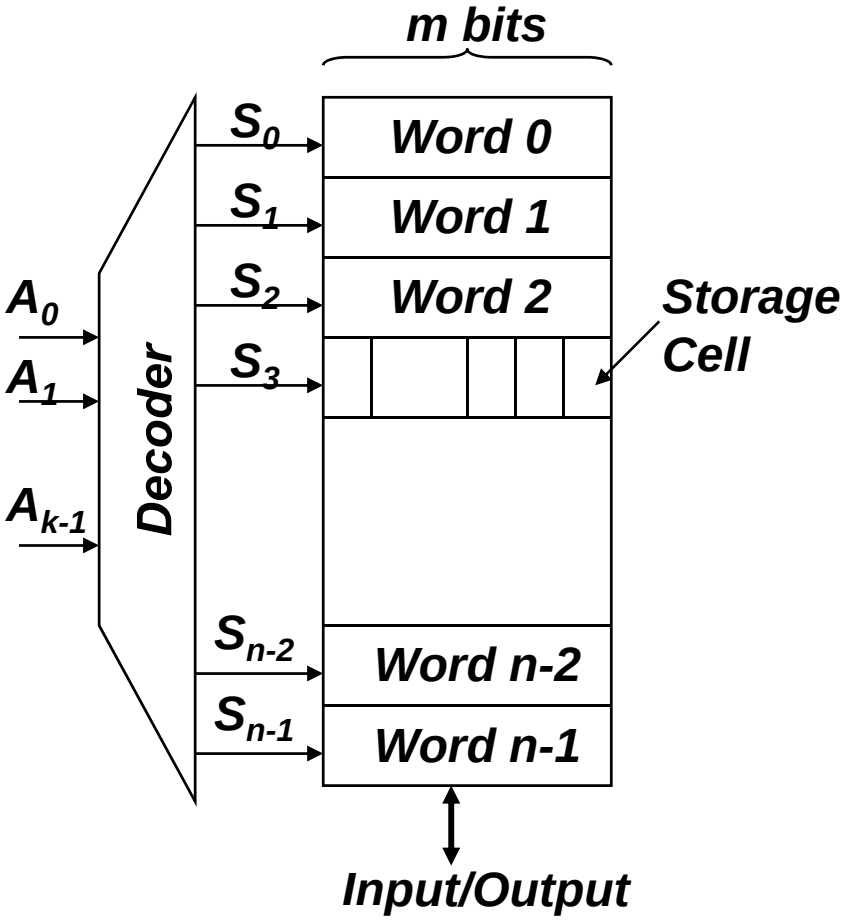
Memory Timing: Definitions



Memory Architecture: Decoders

Intuitive architecture for N x M memory
Too many select signals:
N words == N select signals

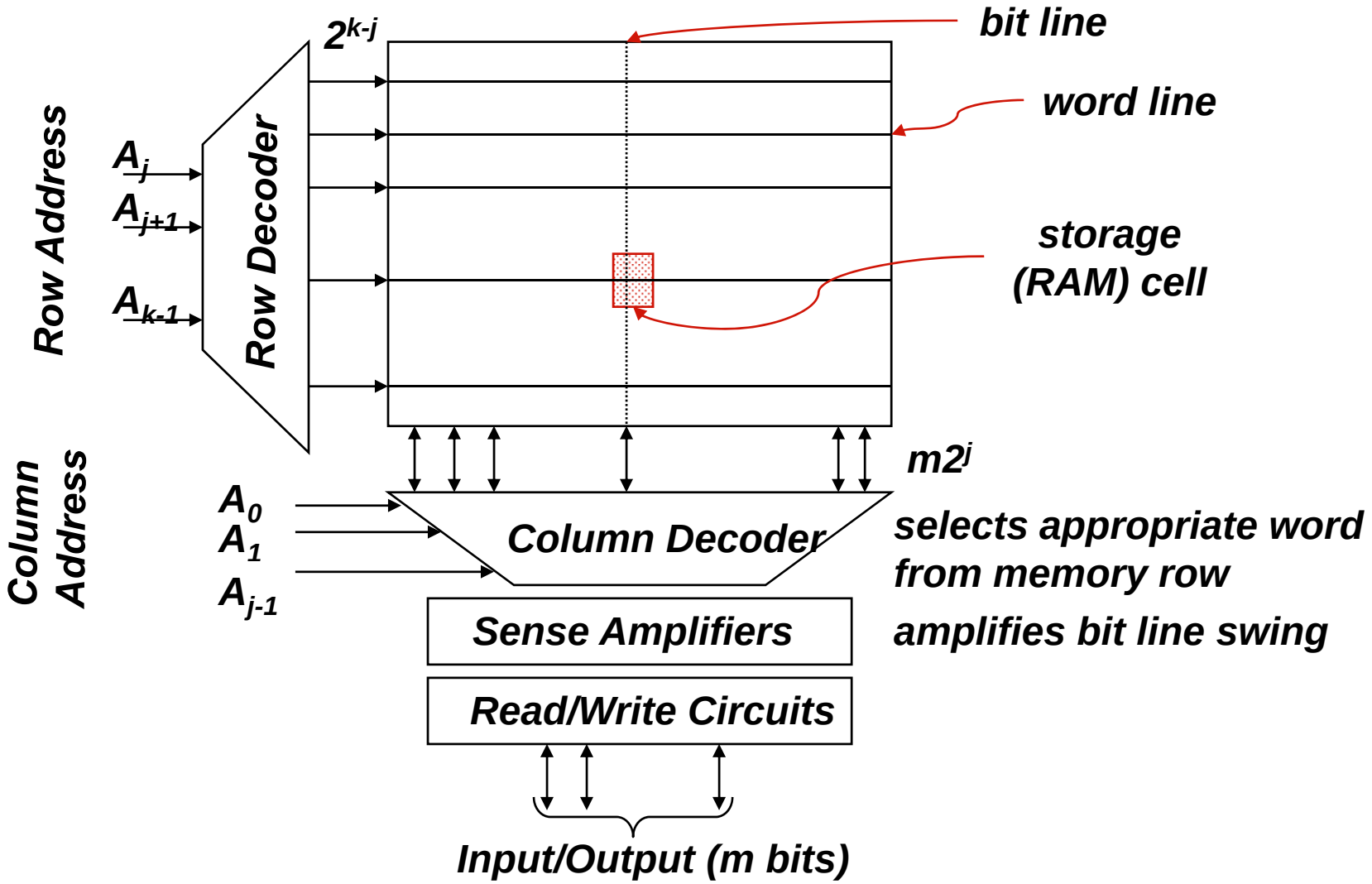




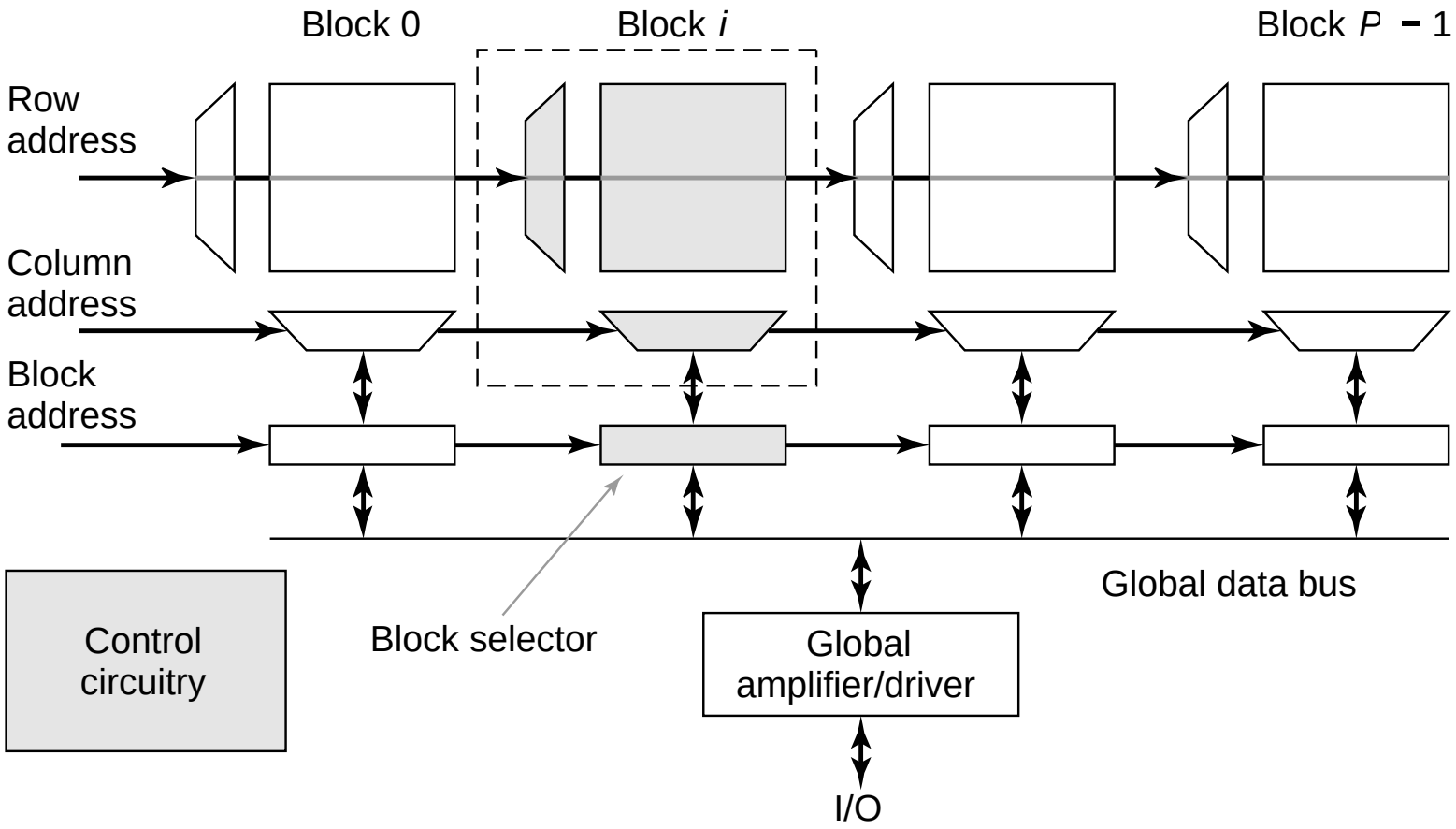
n words → n select signals

Decoder reduces # of inputs
$k = \log_2 n$

Array-Structured Memory Architecture



Hierarchical Memory Architecture

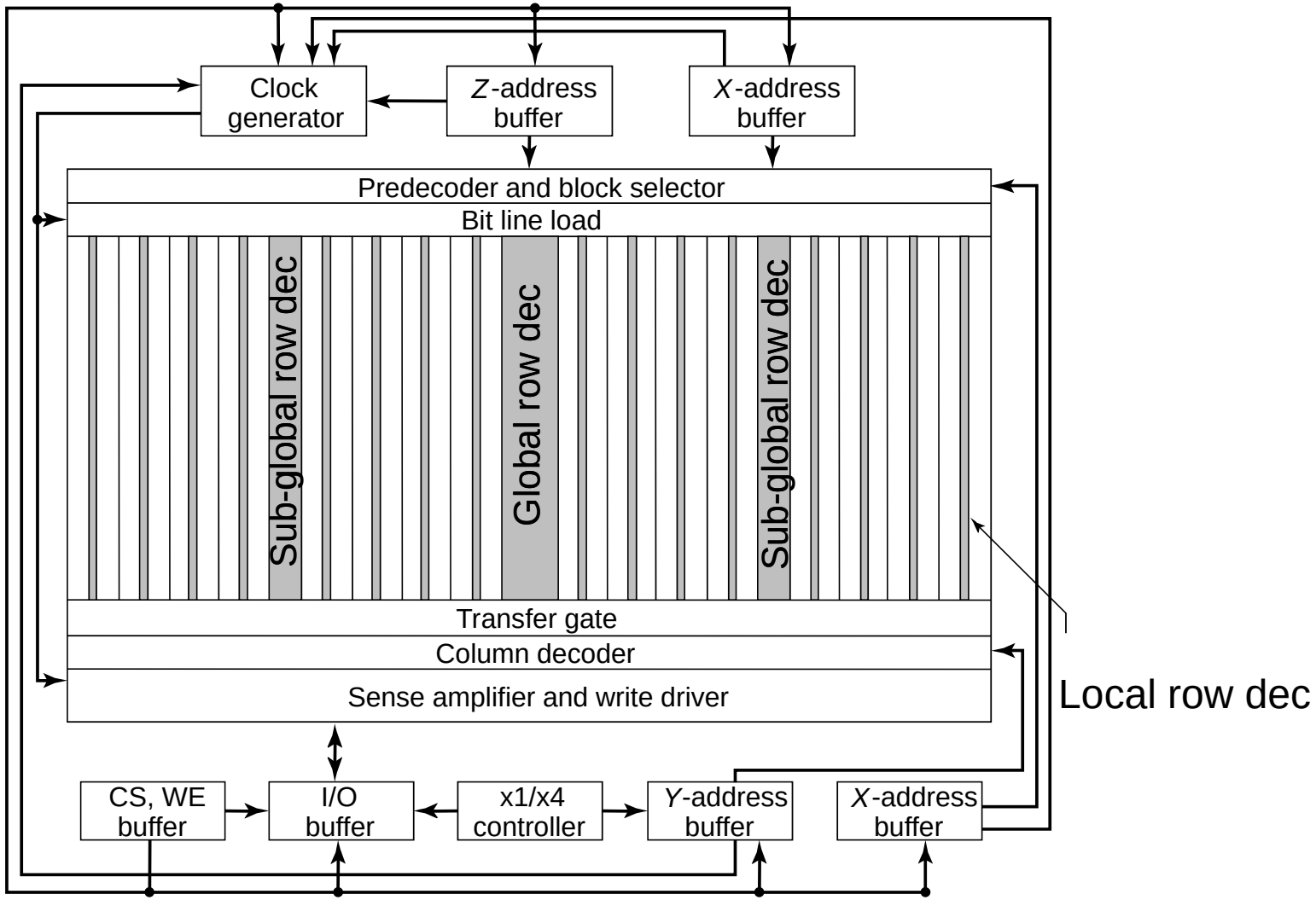


Advantages:

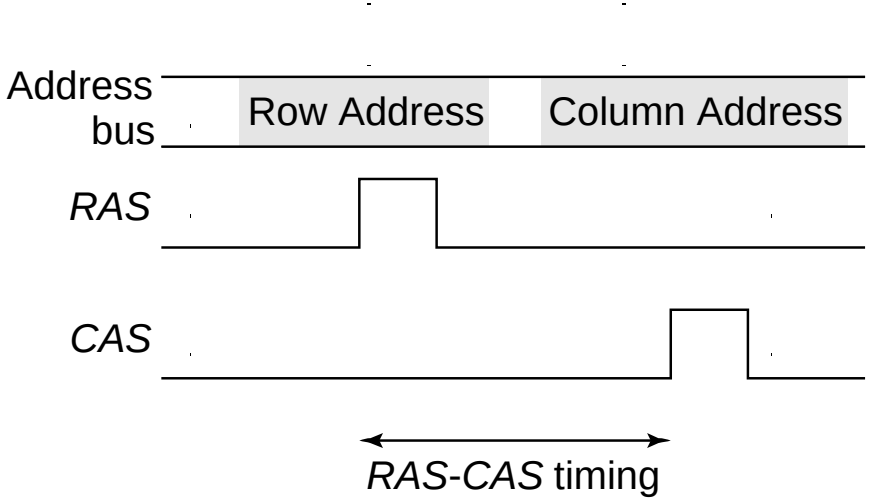
1. Shorter wires within blocks

2. Block address activates only 1 block => power savings

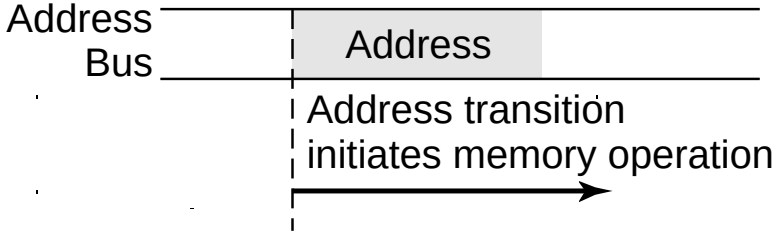
Block Diagram of 4 Mbit SRAM



Memory Timing: Approaches

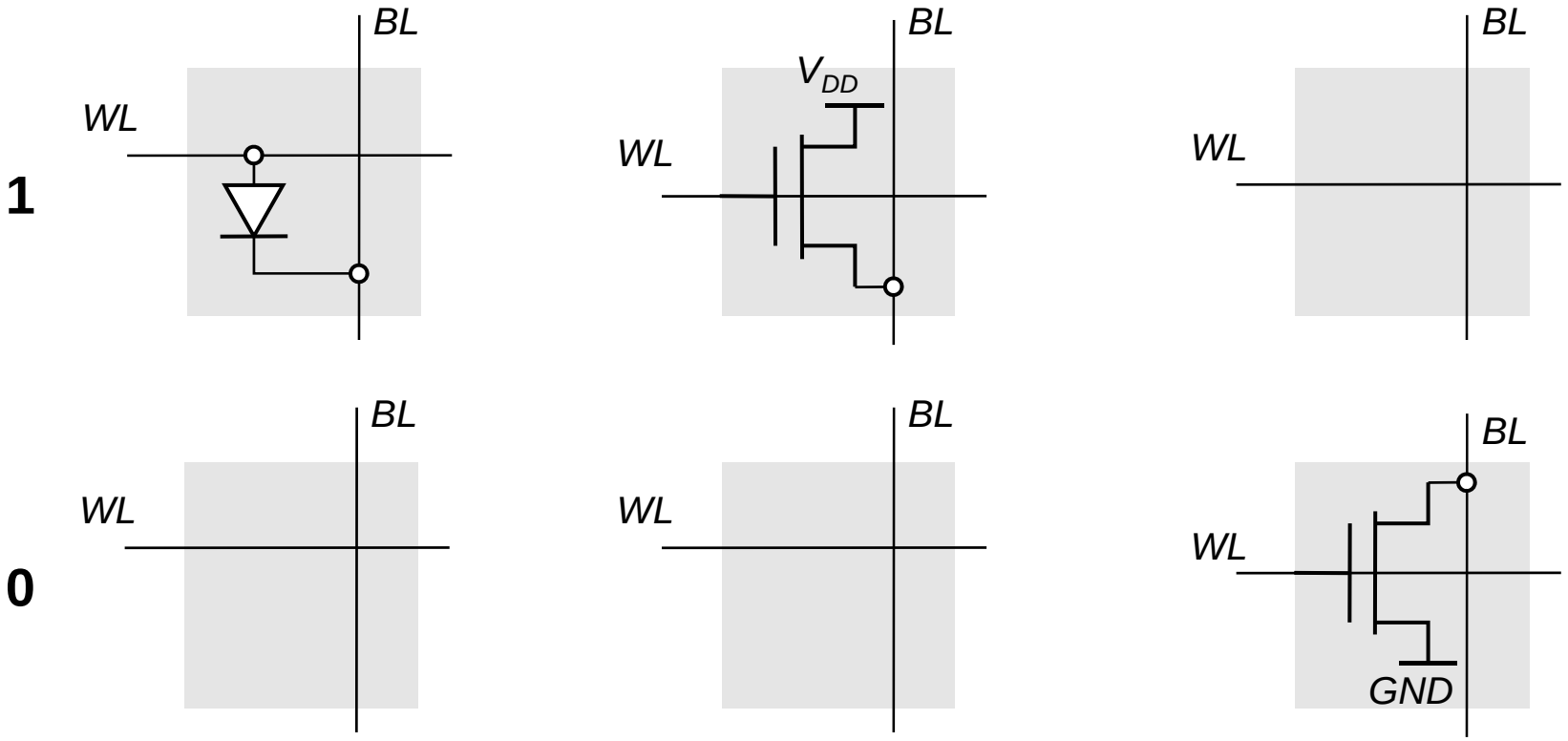


DRAM Timing
Multiplexed Adressing

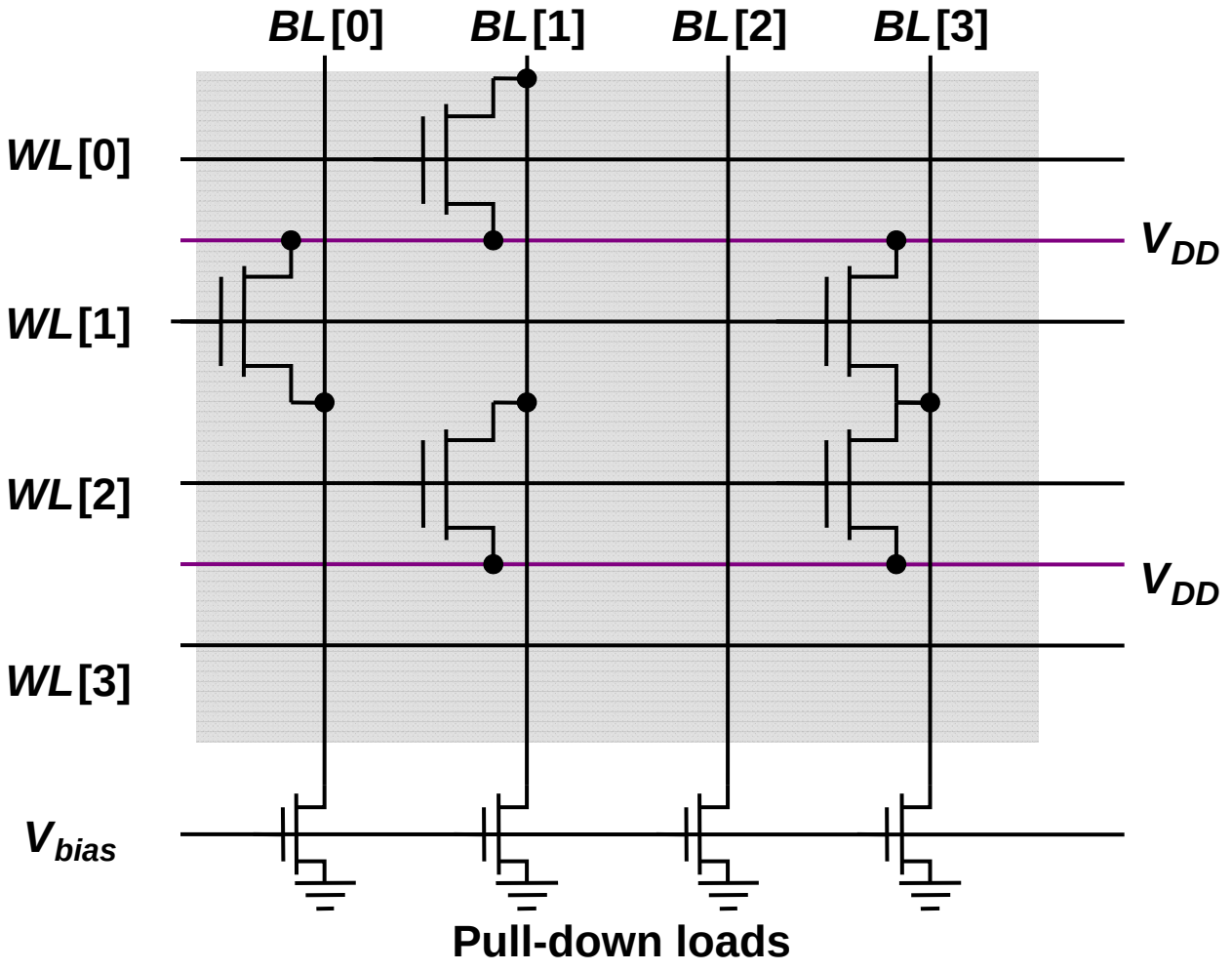


SRAM Timing
Self-timed

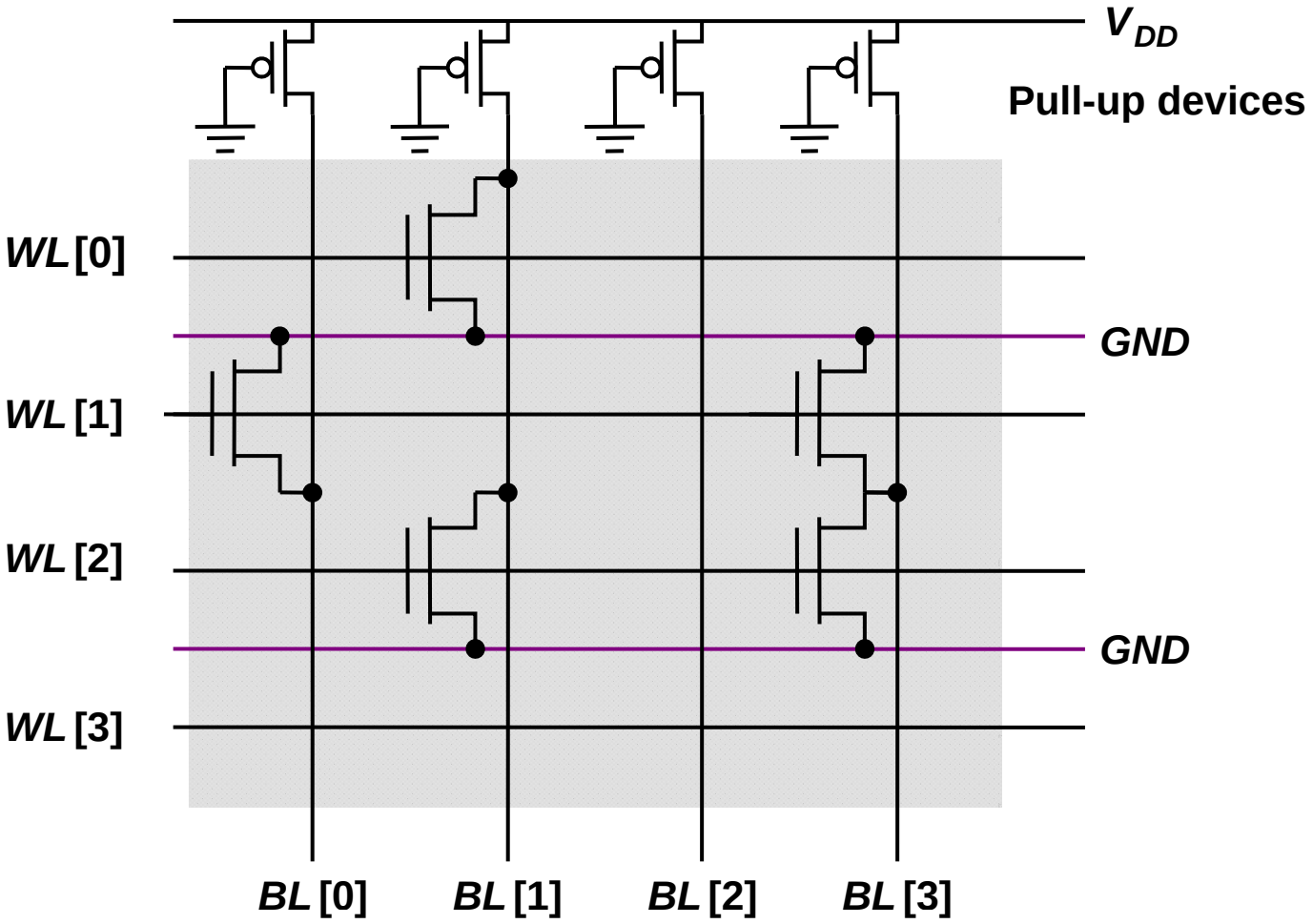
Read-Only Memory Cells



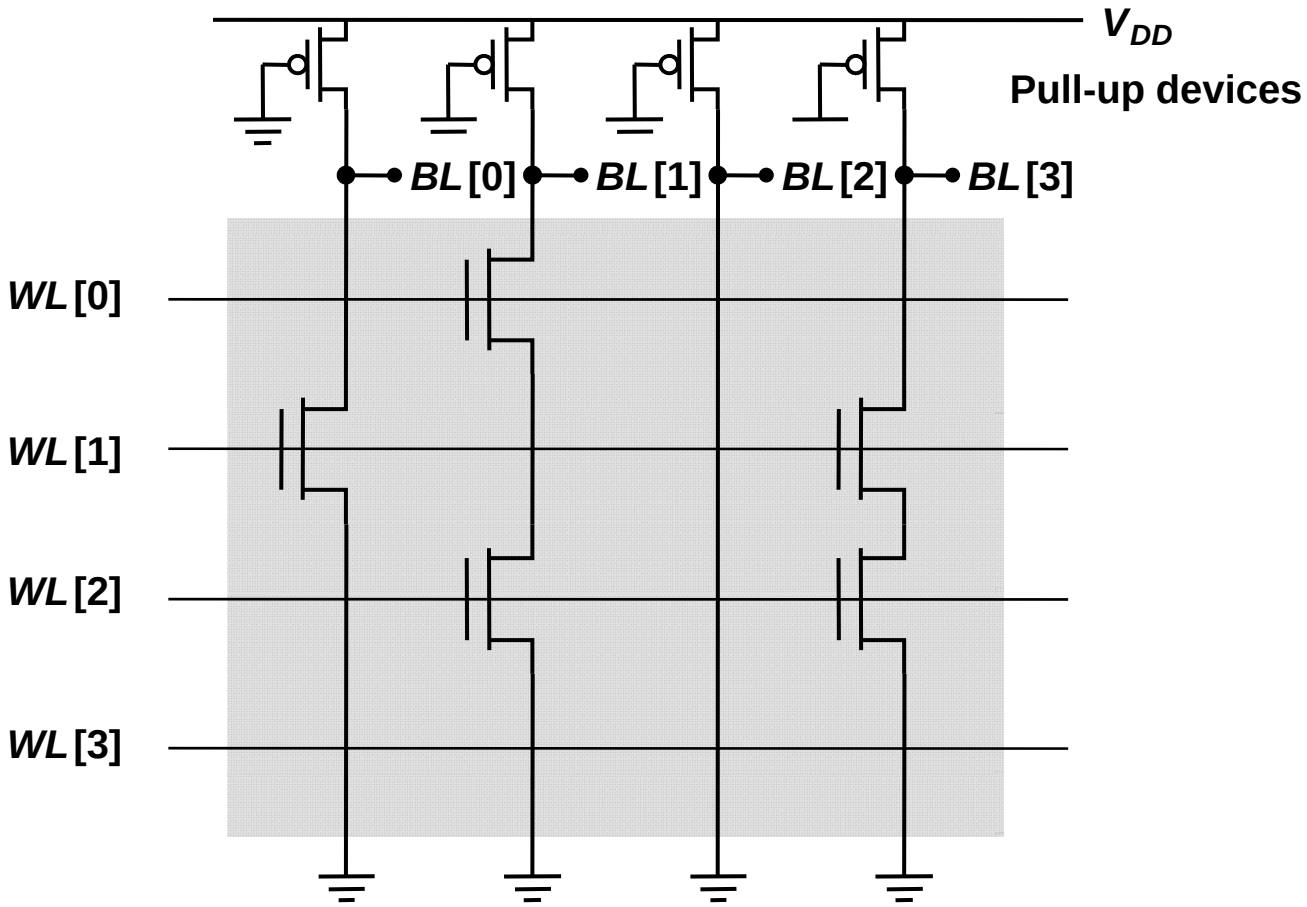
MOS OR ROM



MOS NOR ROM



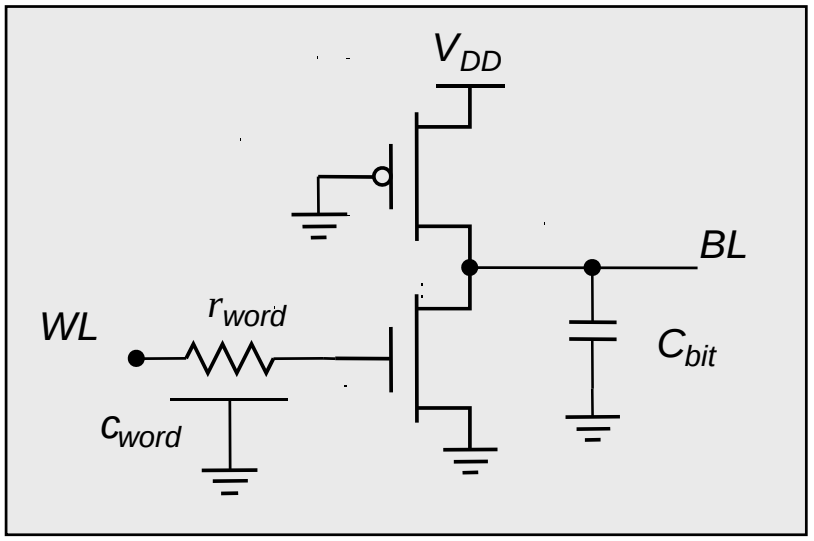
MOS NAND ROM



All word lines high by default with exception of selected row

Equivalent Transient Model for MOS NOR ROM

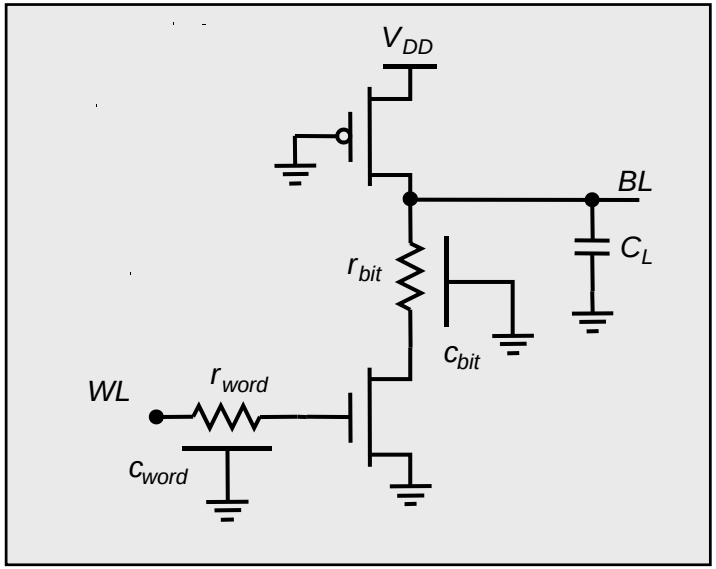
Model for NOR ROM



- Word line parasitics
  - Wire capacitance and gate capacitance
  - Wire resistance (polysilicon)
- Bit line parasitics
  - Resistance not dominant (metal)
  - Drain and Gate-Drain capacitance

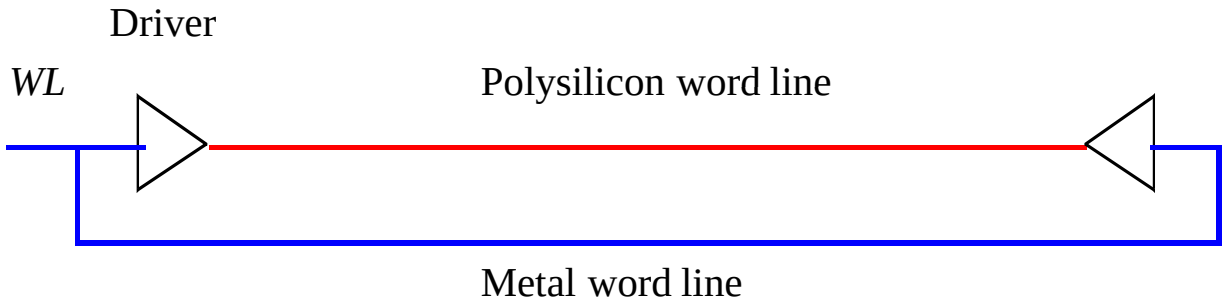
Equivalent Transient Model for MOS NAND ROM

Model for NAND ROM

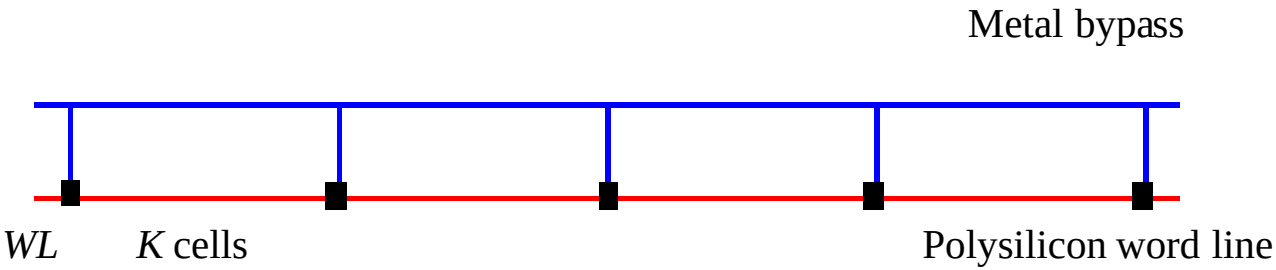


- Word line parasitics
  - Similar to NOR ROM
- Bit line parasitics
  - Resistance of cascaded transistors dominates
  - Drain/Source and complete gate capacitance

Decreasing Word Line Delay

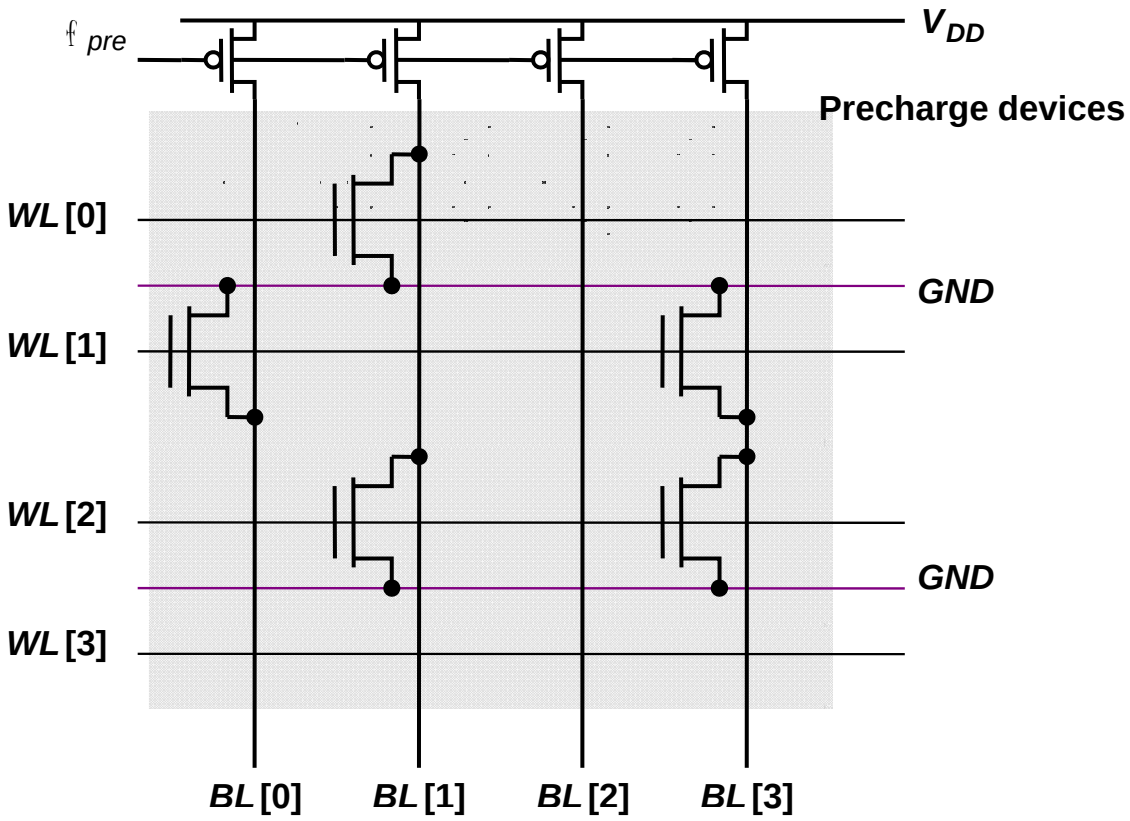


Driving the word line from both sides



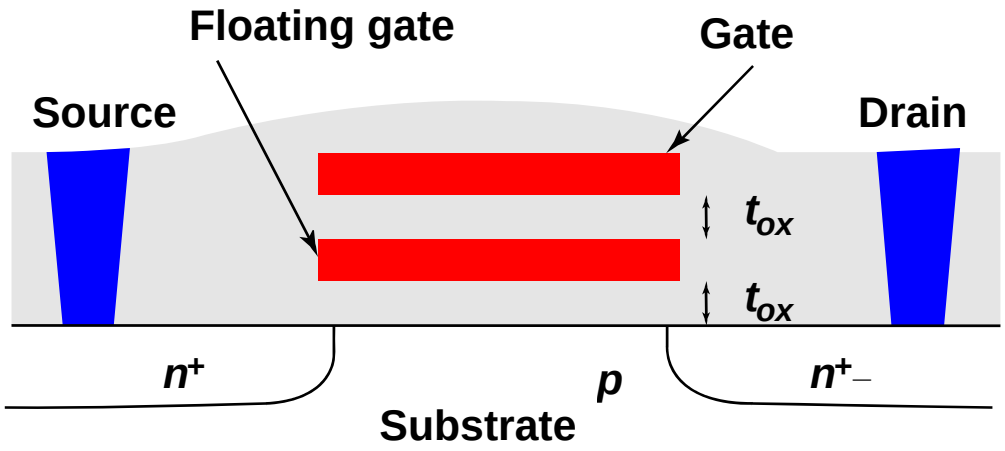
Using a metal bypass

Precharged MOS NOR ROM

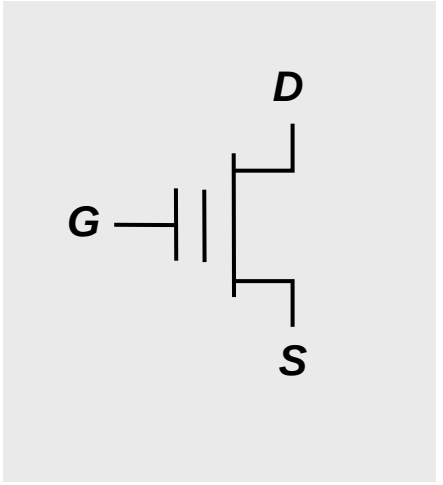


PMOS precharge device can be made as large as necessary, but clock driver becomes harder to design.

Non-Volatile Memories
The Floating-gate transistor (FAMOS)

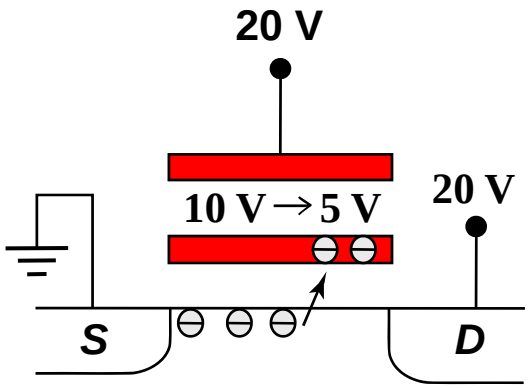


Device cross-section

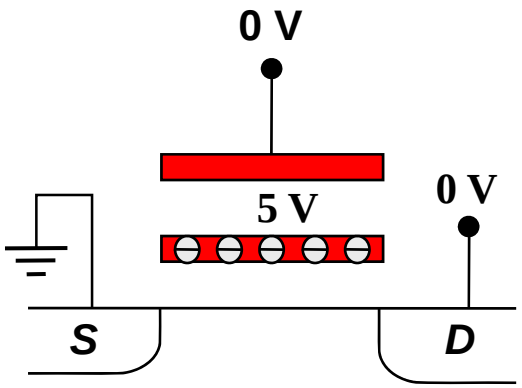


Schematic symbol

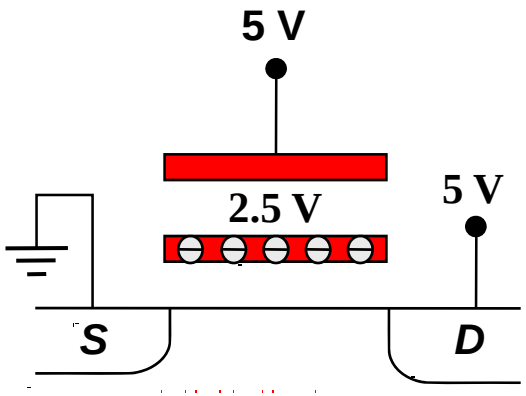
Floating-Gate Transistor Programming



Avalanche injection

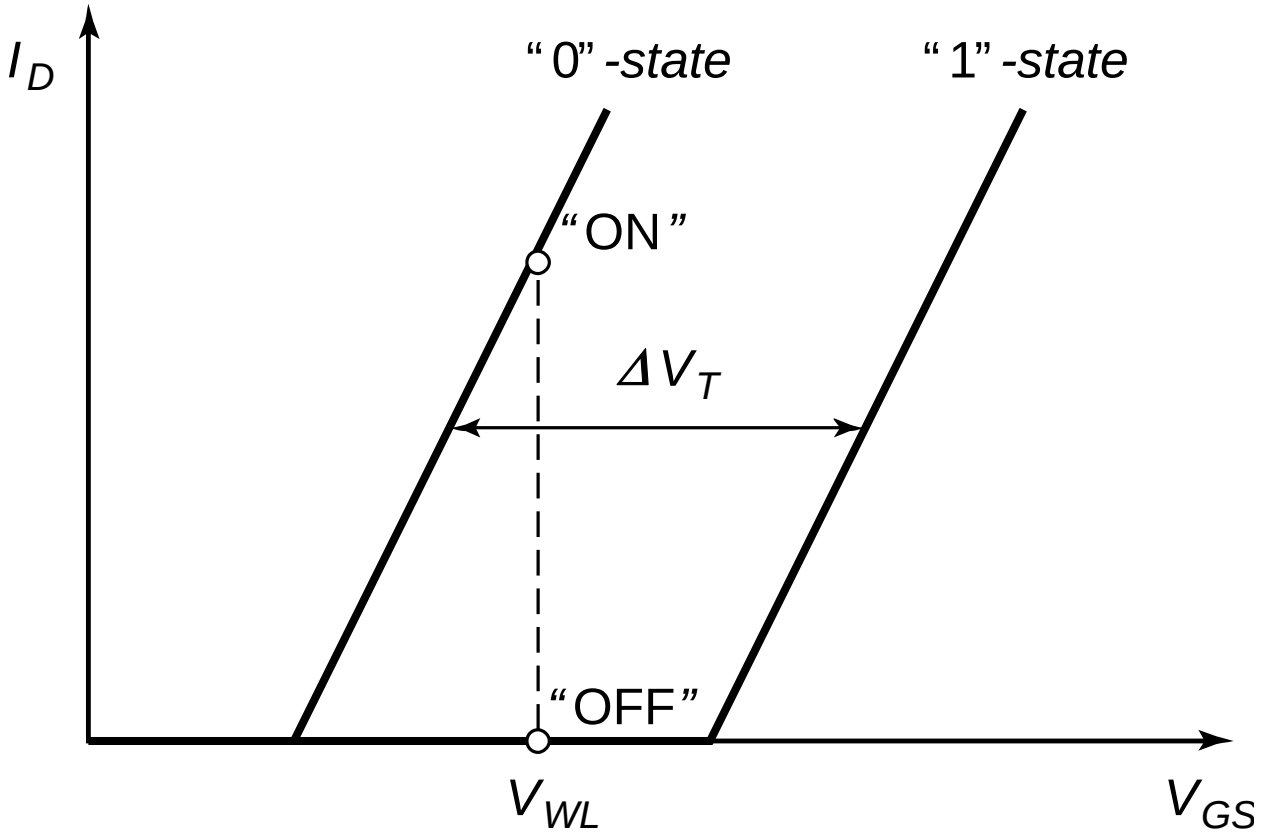


Removing programming voltage leaves charge trapped

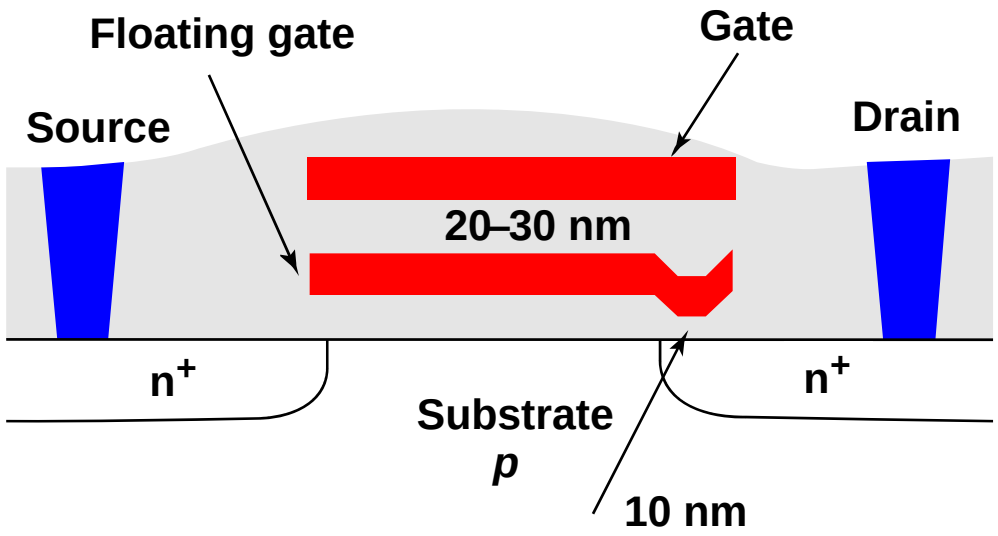


Programming results in higher V_T.

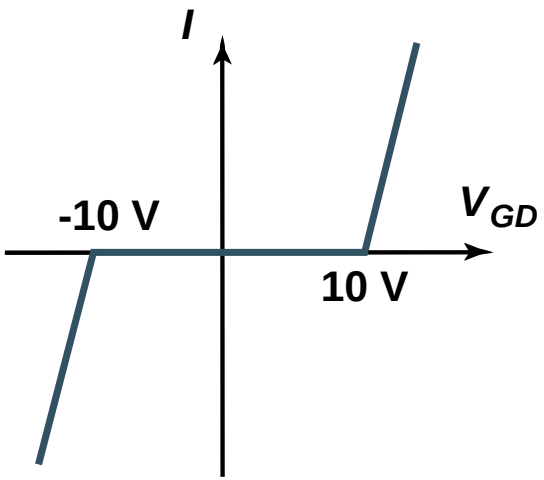
A “Programmable-Threshold” Transistor



FLOTOX EEPROM

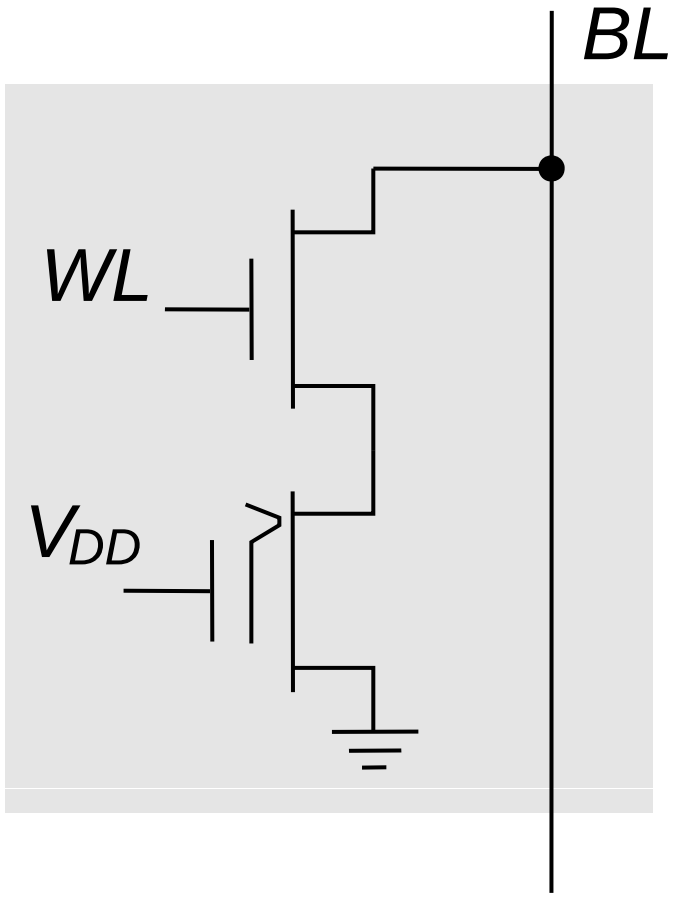


FLOTOX transistor



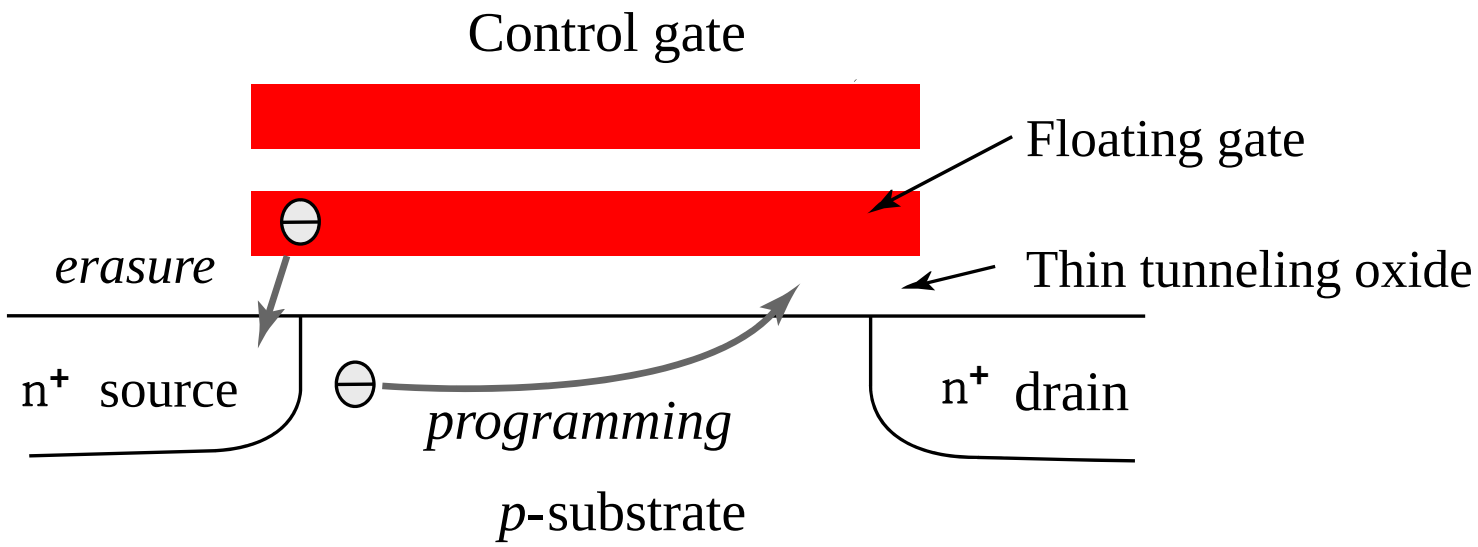
Fowler-Nordheim
I-V characteristic

EEPROM Cell



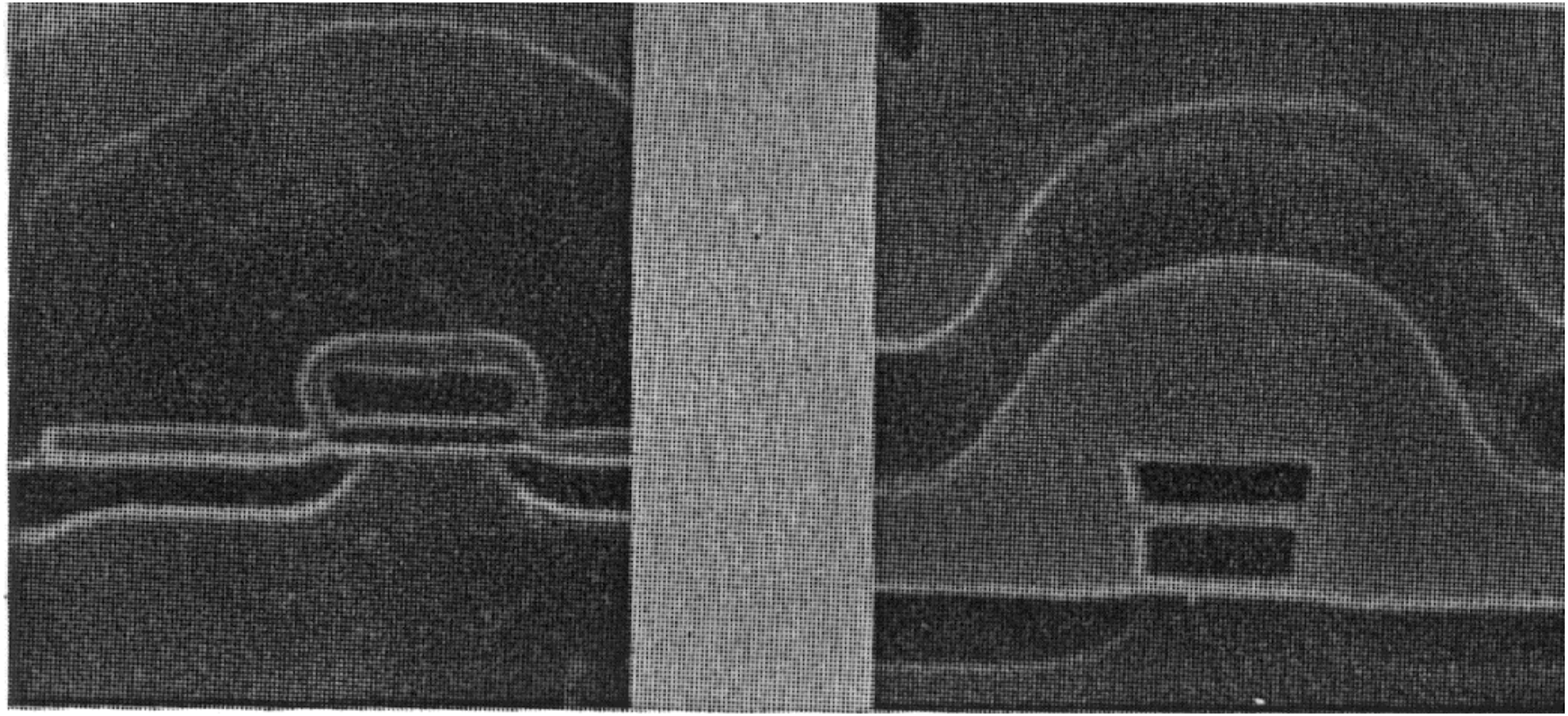
Absolute threshold control is hard
Unprogrammed transistor might be depletion
⇨ 2 transistor cell

Flash EEPROM



Many other options …

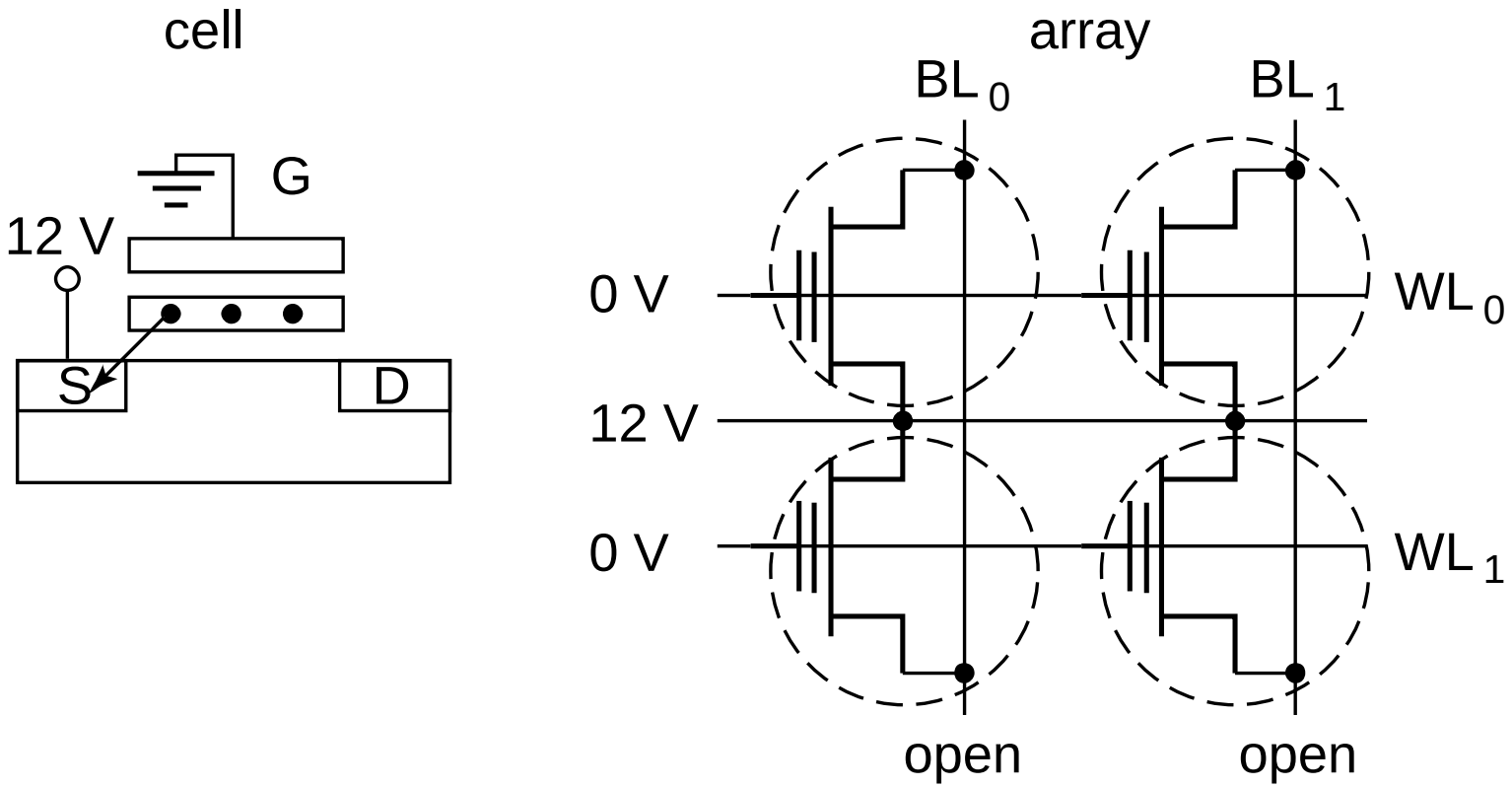
Cross-sections of NVM cells



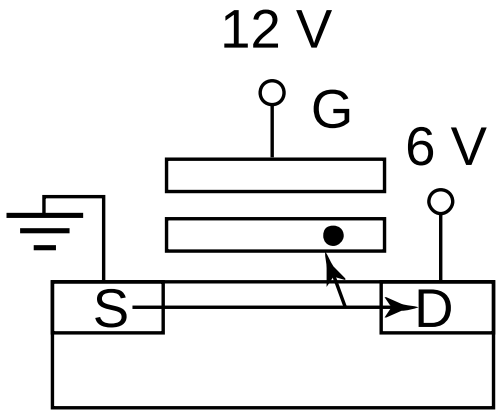
Flash **EPROM**

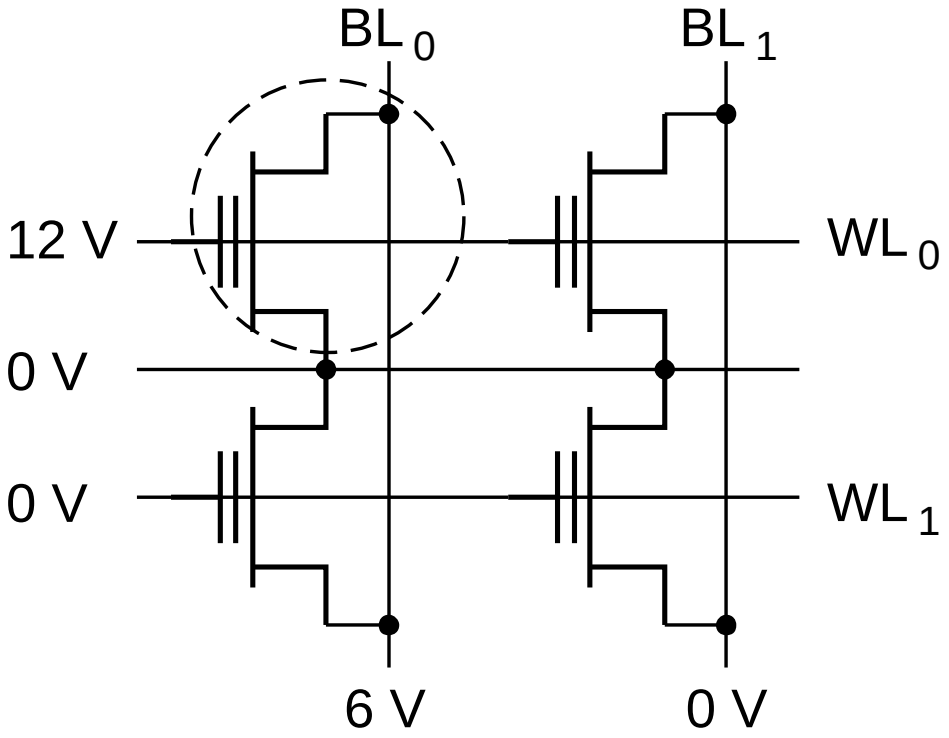
Courtesy Intel

Basic Operations in a NOR Flash Memory—Erase

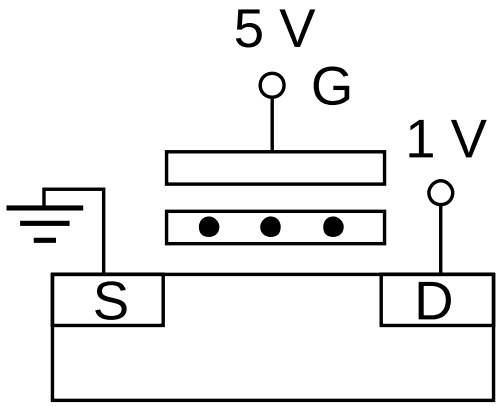


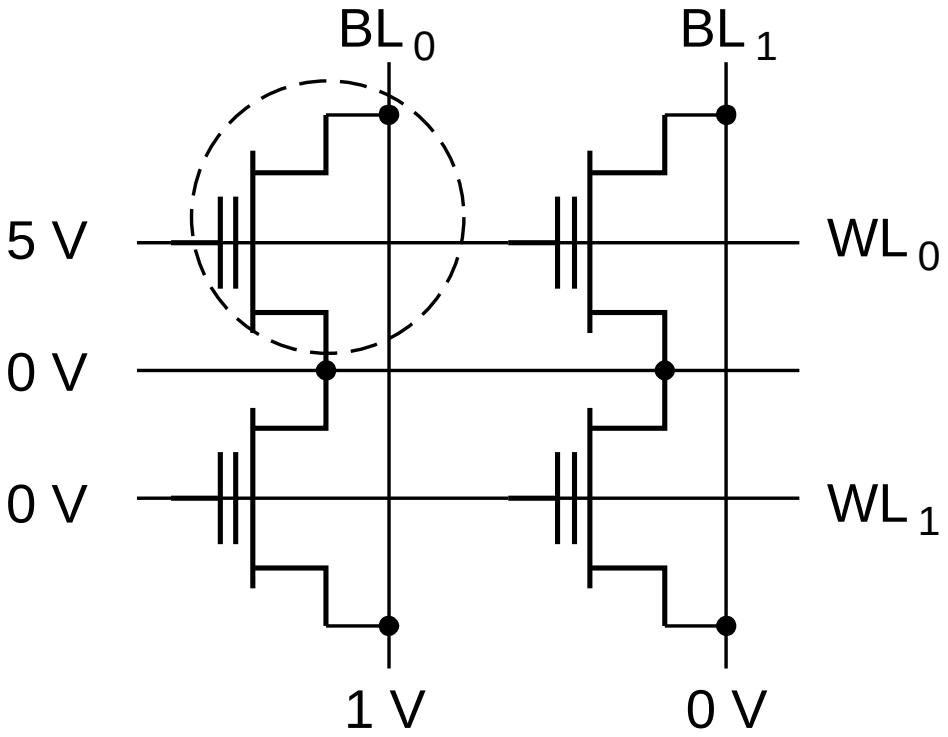
Basic Operations in a NOR Flash Memory—Write



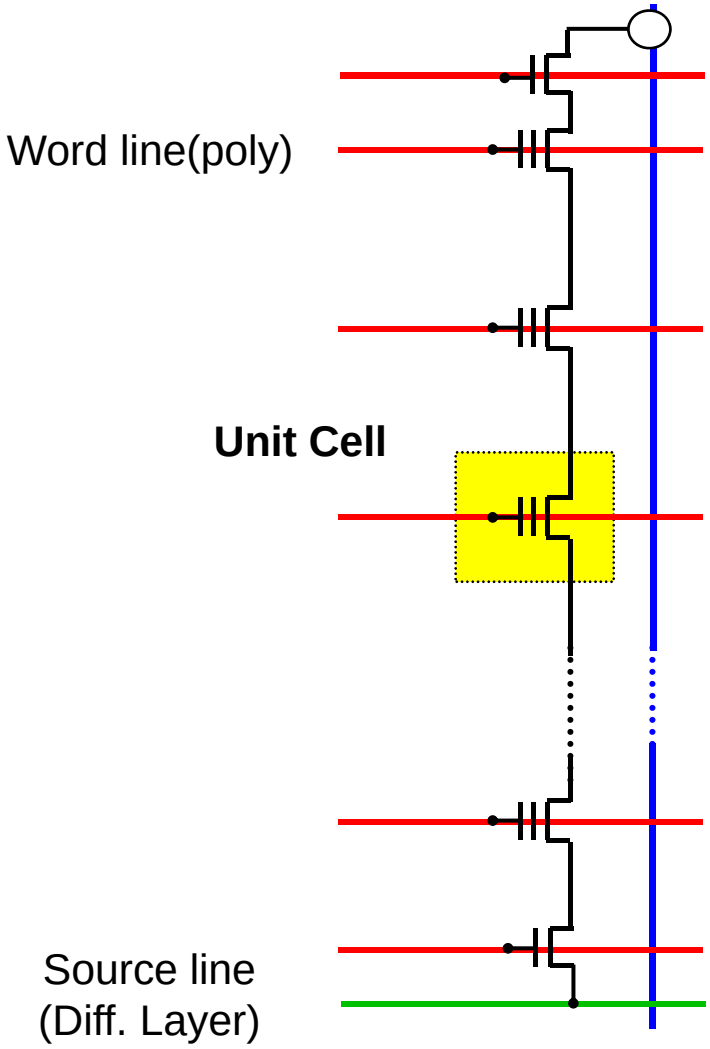


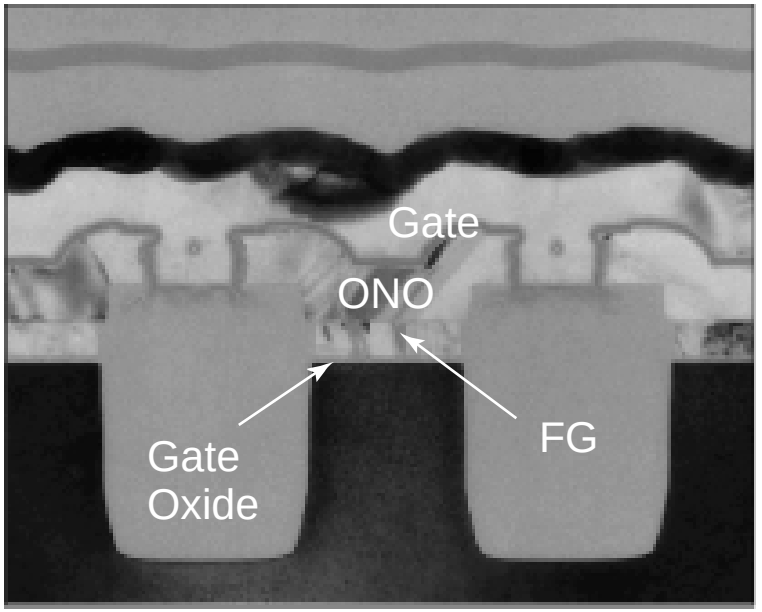
Basic Operations in a NOR Flash Memory—Read





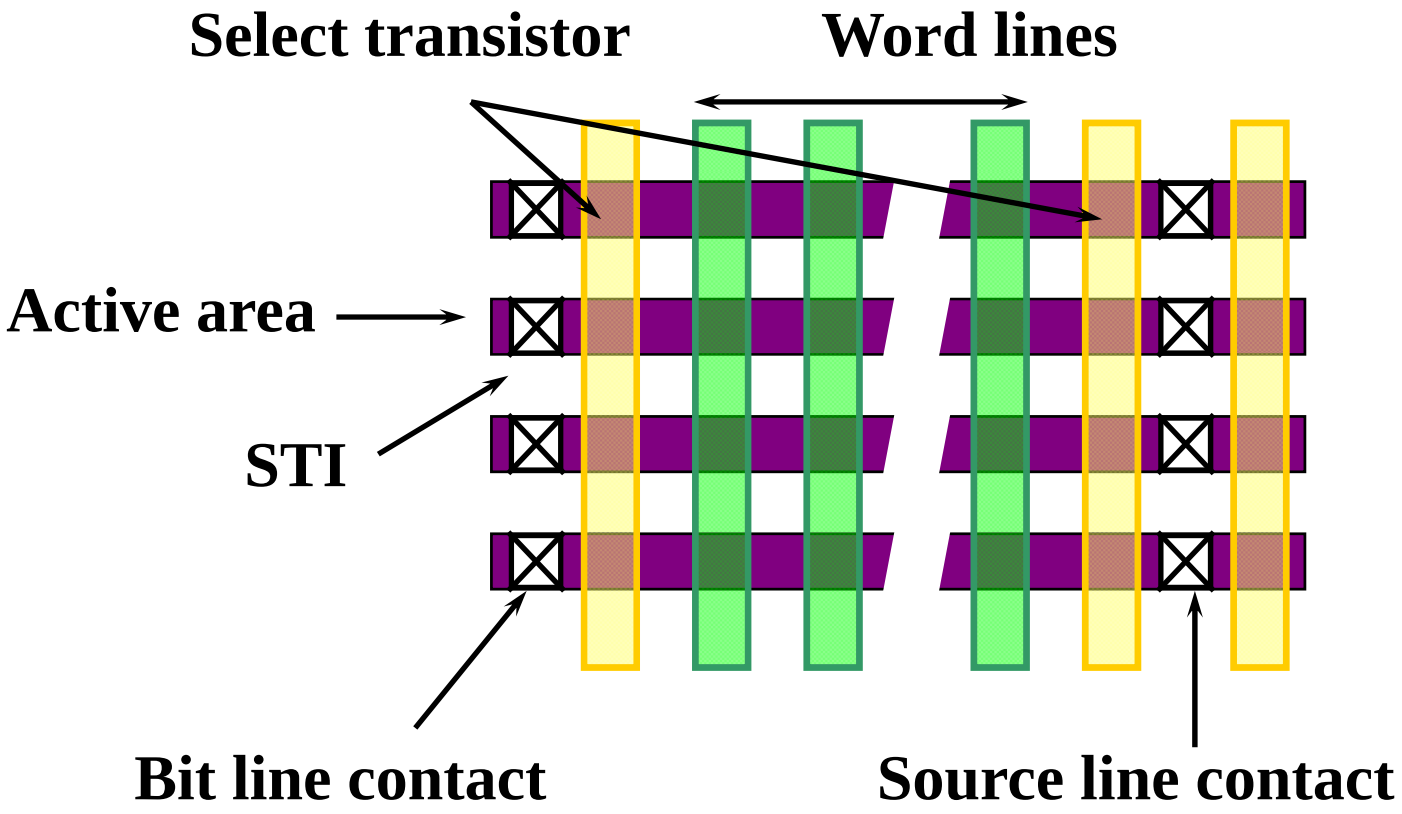
NAND Flash Memory





Courtesy Toshiba

NAND Flash Memory



Courtesy Toshiba

Read-Write Memories (RAM)

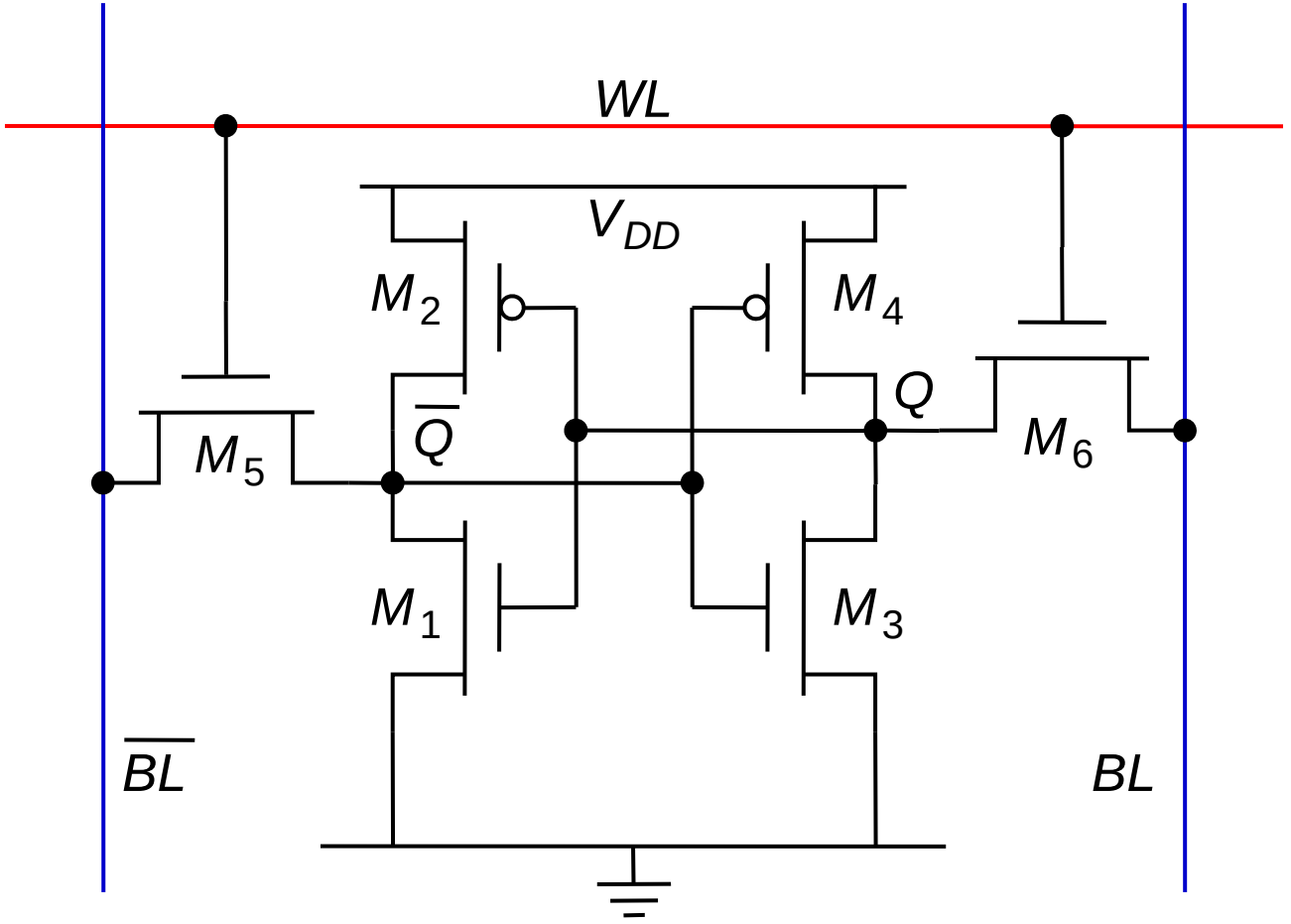
- **STATIC (SRAM)**

  Data stored as long as supply is applied
  Large (6 transistors/cell)
  Fast
  Differential

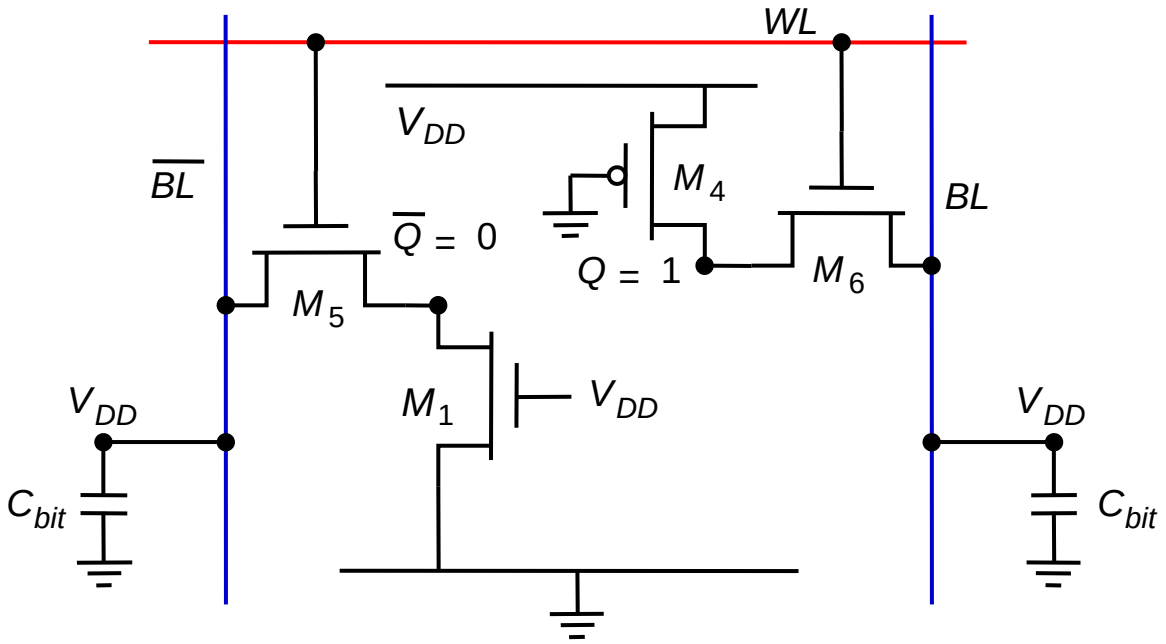
- **DYNAMIC (DRAM)**

  Periodic refresh required
  Small (1-3 transistors/cell)
  Slower
  Single Ended

6-transistor CMOS SRAM Cell



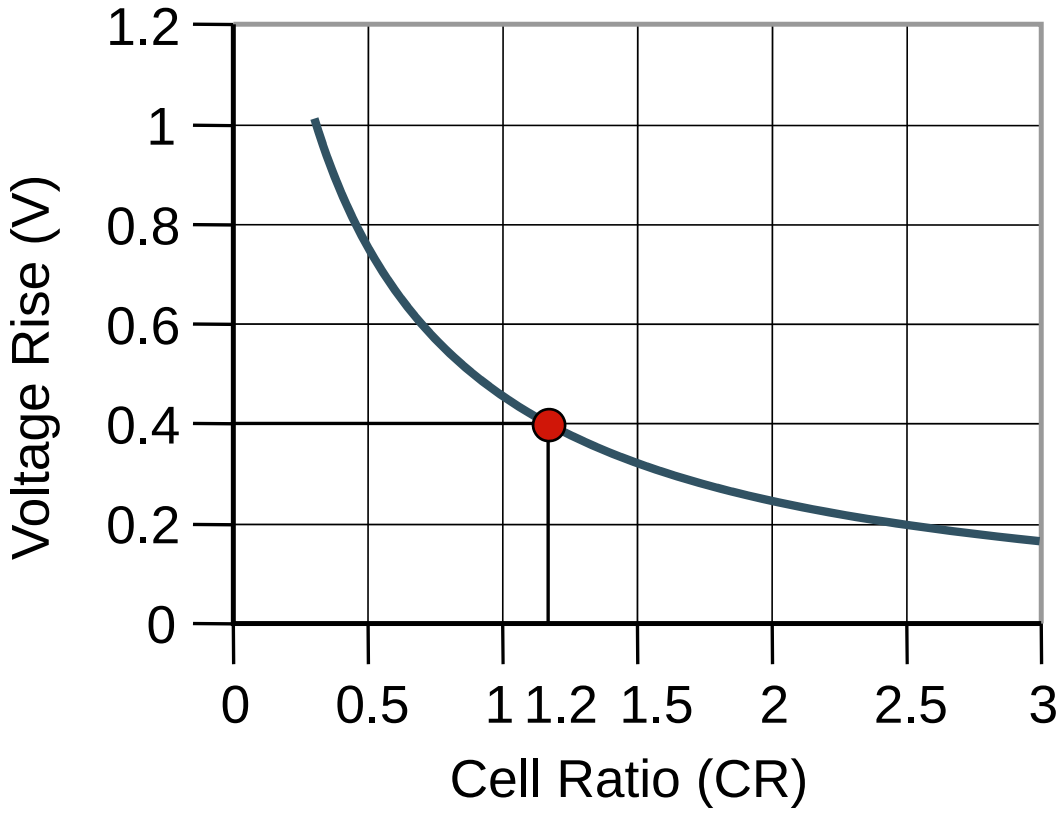
CMOS SRAM Analysis (Read)



$$k_{n,M5}\left((V_{DD}-\Delta V-V_{Tn})V_{DSATn}-\frac{V_{DSATn}^2}{2}\right) = k_{n,M1}\left((V_{DD}-V_{Tn})\Delta V-\frac{\Delta V^2}{2}\right)$$

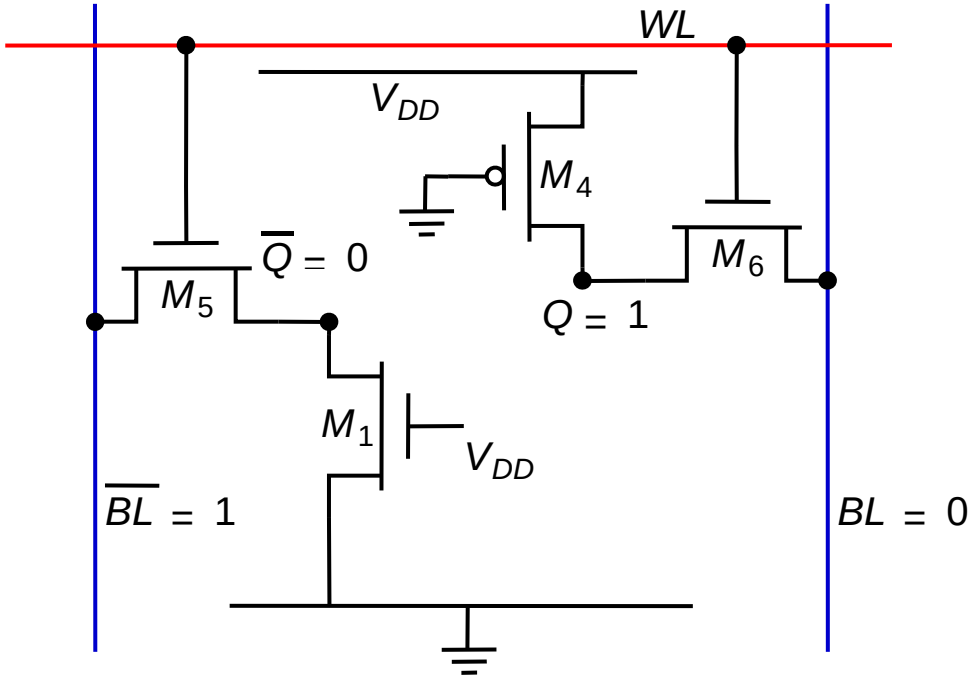
$$\Delta V = \frac{V_{DSATn}+CR(V_{DD}-V_{Tn})-\sqrt{V_{DSATn}^2(1+CR)+CR^2(V_{DD}-V_{Tn})^2}}{CR}$$

CMOS SRAM Analysis (Read)



$$CR = \frac{W_1/L_1}{W_5/L_5}$$

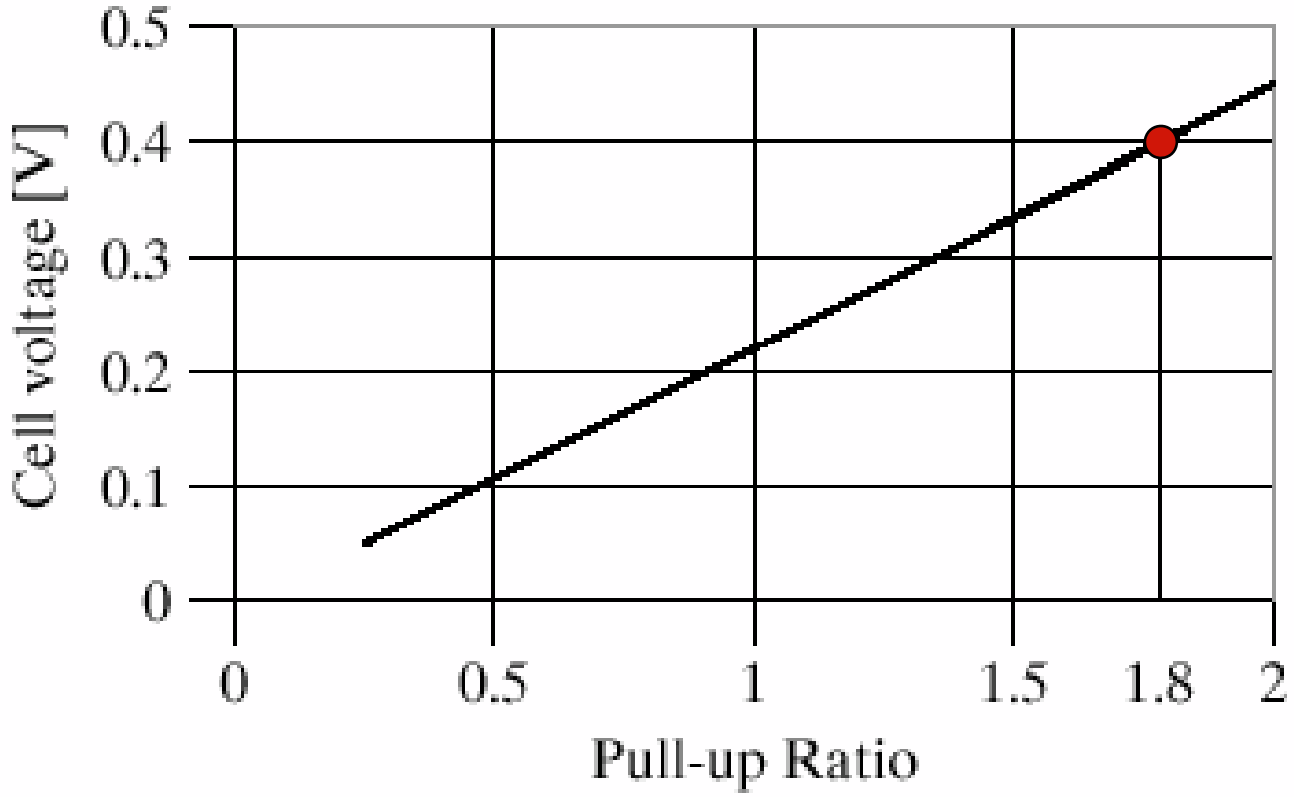
CMOS SRAM Analysis (Write)



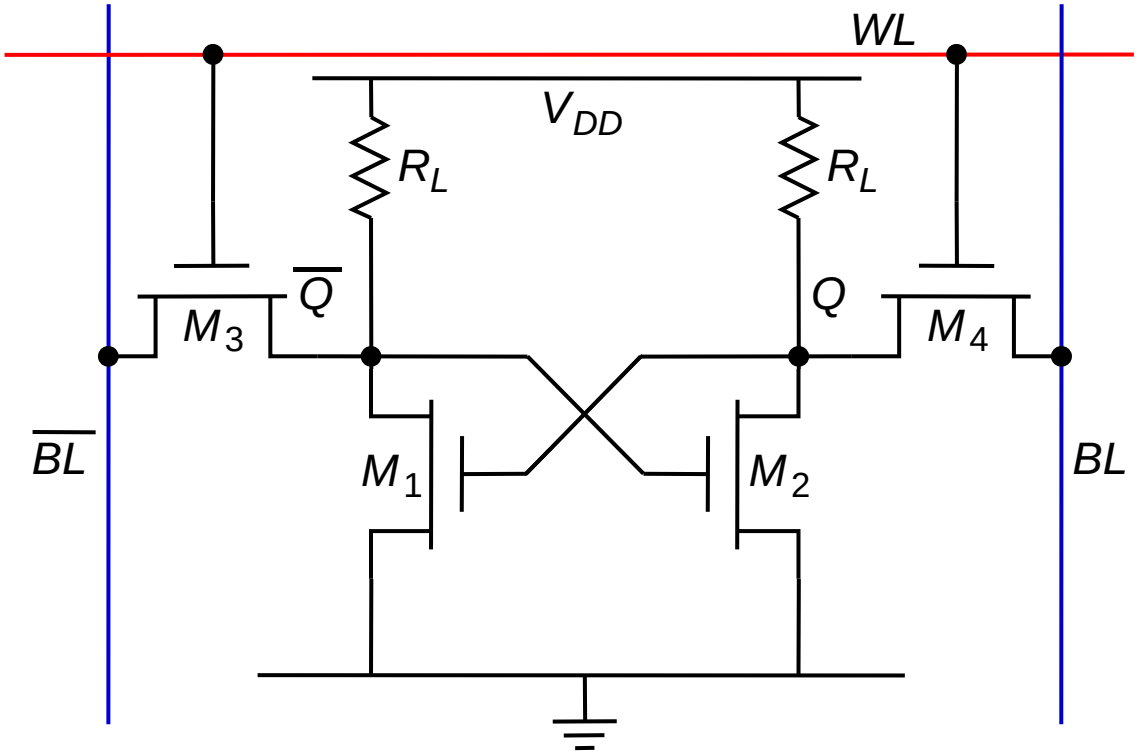
$$k_{n,M6}\left((V_{DD}-V_{Tn})V_Q-\frac{V_Q^2}{2}\right) = k_{p,M4}\left((V_{DD}-|V_{Tp}|)V_{DSATp}-\frac{V_{DSATp}^2}{2}\right)$$

$$V_Q = V_{DD}-V_{Tn}-\sqrt{(V_{DD}-V_{Tn})^2-2\frac{\mu_p}{\mu_n}PR\left((V_{DD}-|V_{Tp}|)V_{DSATp}-\frac{V_{DSATp}^2}{2}\right)},$$

CMOS SRAM Analysis (Write)

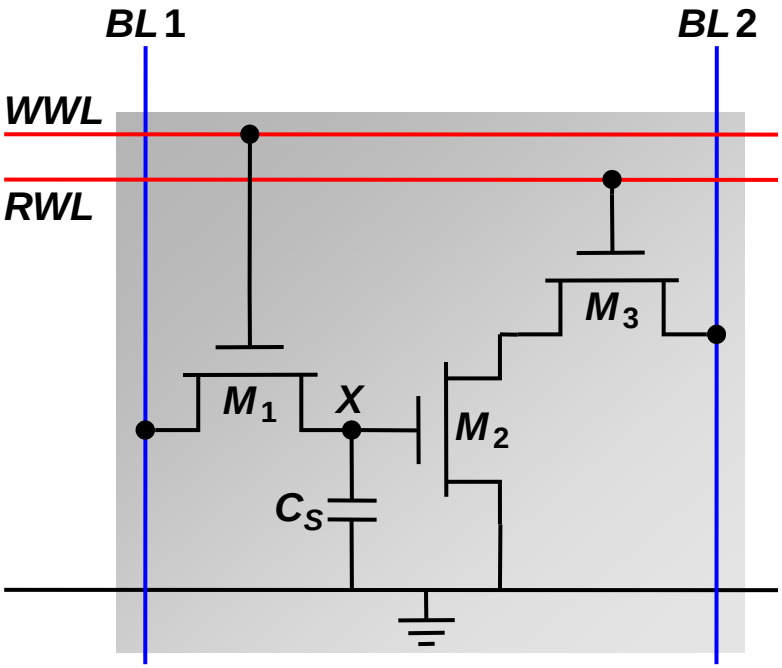


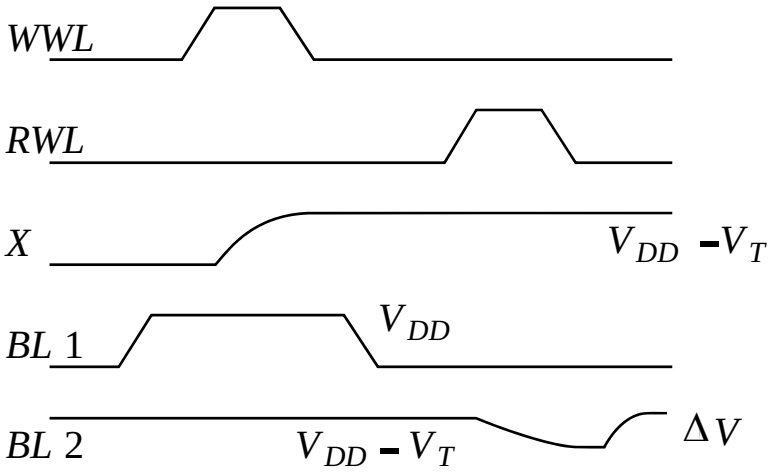
Resistance-load SRAM Cell



Static power dissipation -- Want R_L large
Bit lines precharged to V_{DD} to address t_p problem

3-Transistor DRAM Cell



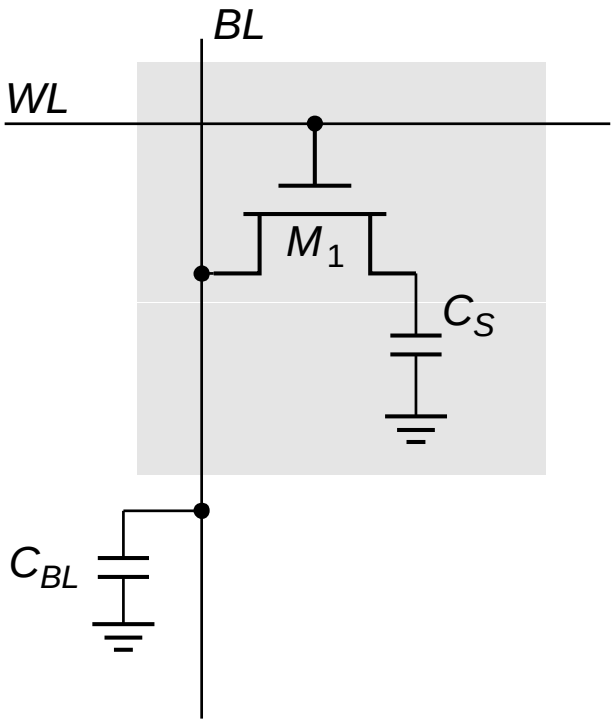


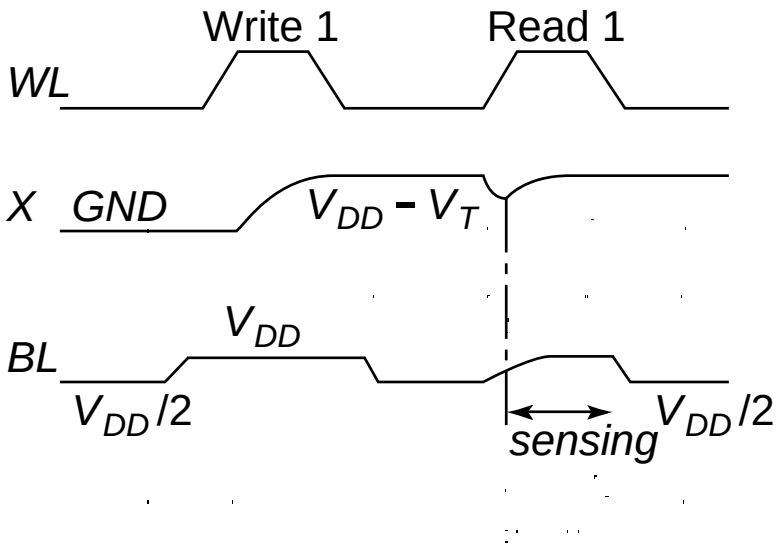
No constraints on device ratios

Reads are non-destructive

Value stored at node X when writing a “1” = V_{WWL}-V_{Tn}

1-Transistor DRAM Cell





Write: C_S is charged or discharged by asserting WL and BL.
Read: Charge redistribution takes places between bit line and storage capacitance

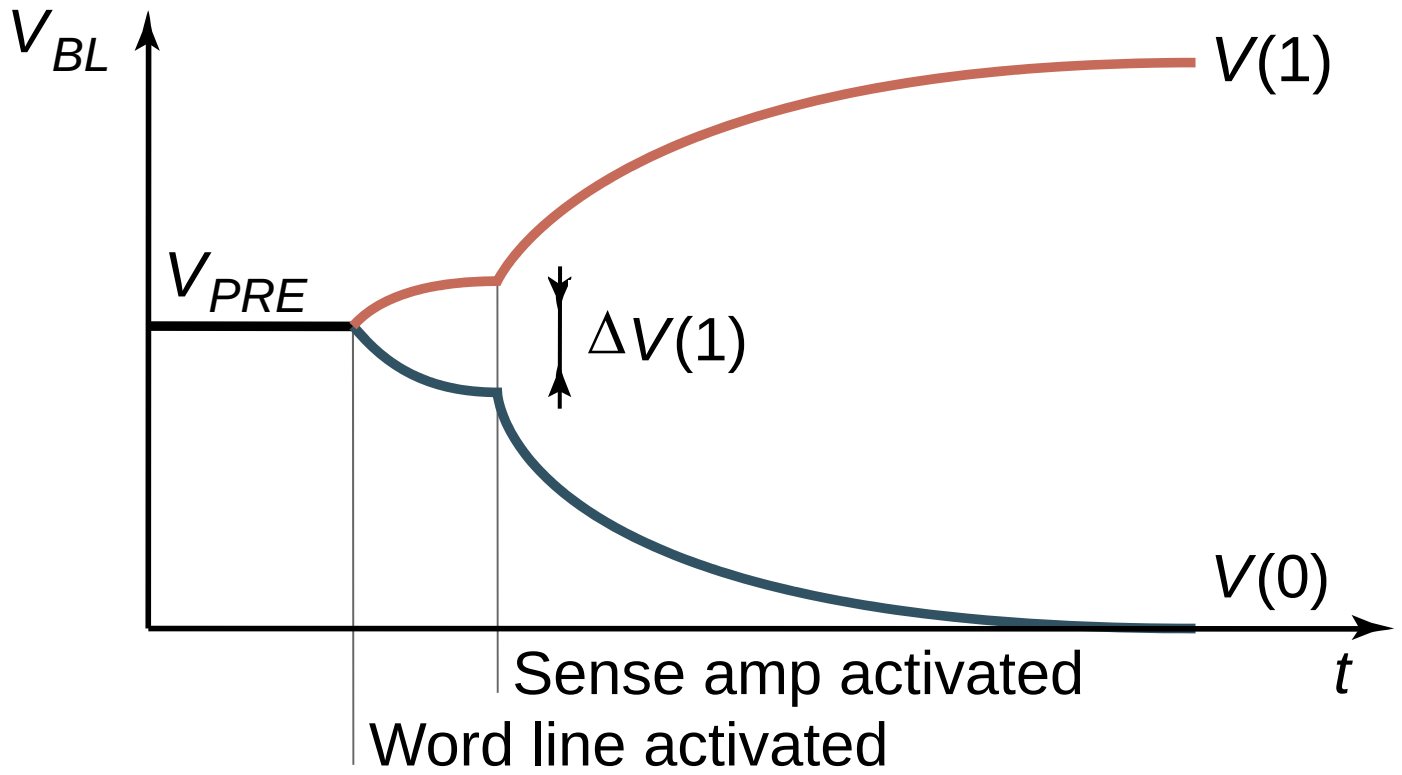
$$\Delta V = V_{BL} - V_{PRE} = V_{BIT} - V_{PRE} \frac{C_S}{C_S + C_{BL}}$$

Voltage swing is small; typically around 250 mV.

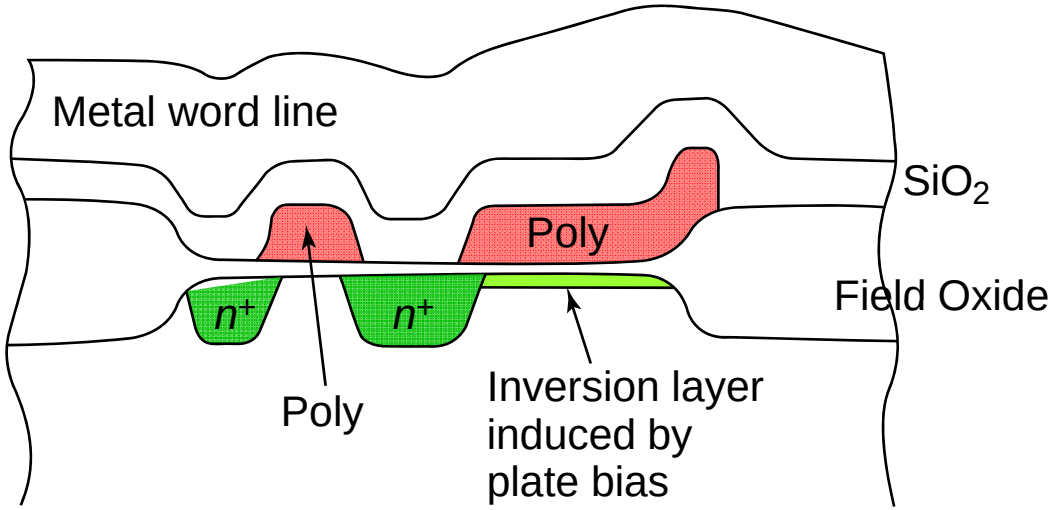
DRAM Cell Observations

- 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.
- Unlike 3T cell, 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}

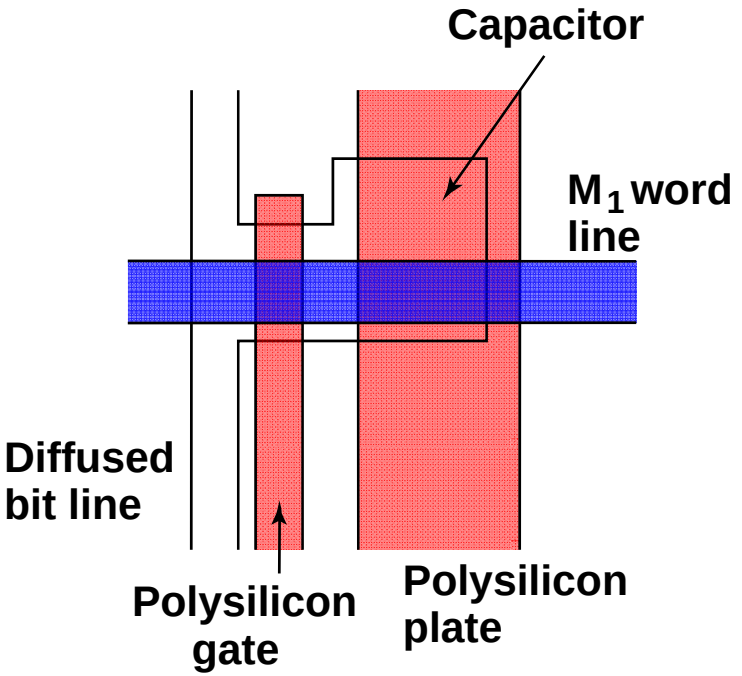
Sense Amp Operation



1-T DRAM Cell



Cross-section

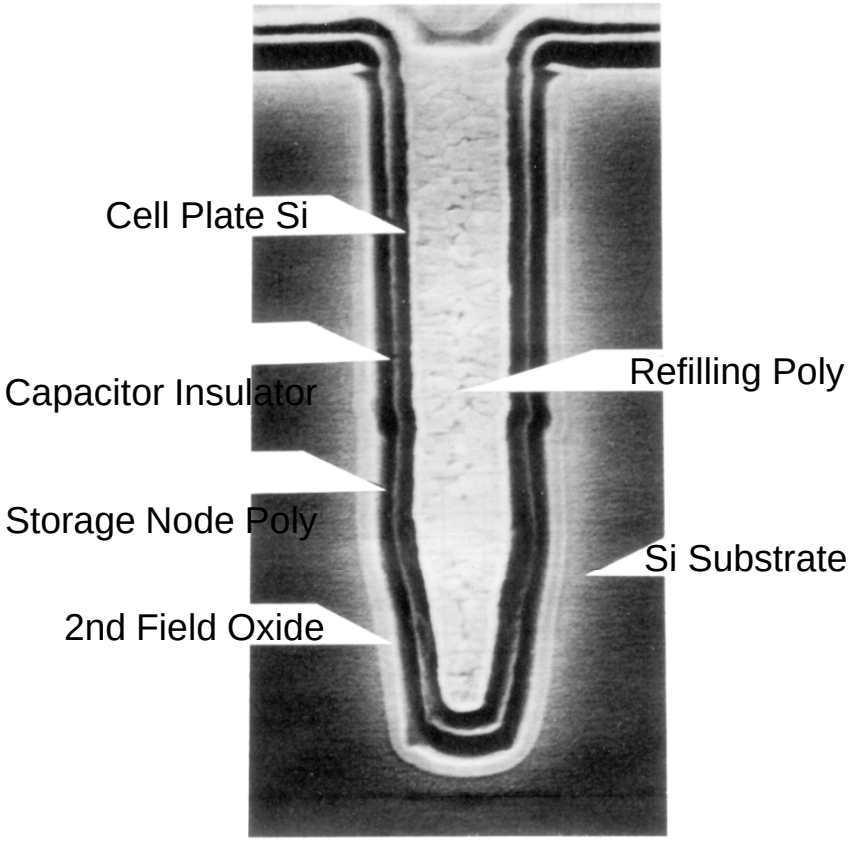


Layout

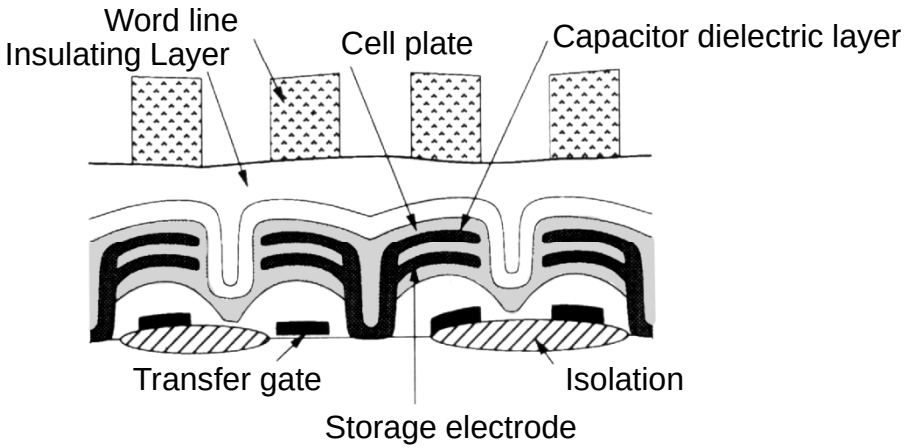
Uses Polysilicon-Diffusion Capacitance

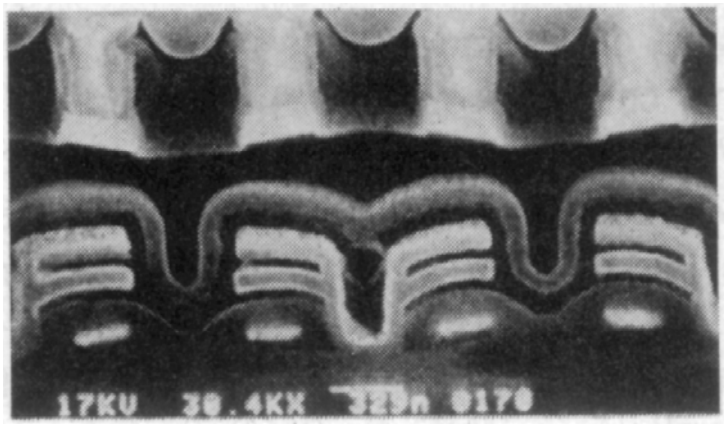
Expensive in Area

Advanced 1T DRAM Cells



Trench Cell





Stacked-capacitor Cell

Row Decoders

Collection of 2^M complex logic gates
Organized in regular and dense fashion

(N)AND Decoder

$$WL_0 = A_0 A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8 A_9$$

$$WL_{511} = \bar{A}_0 A_1 A_2 A_3 A_4 A_5 A_6 A_7 A_8 A_9$$

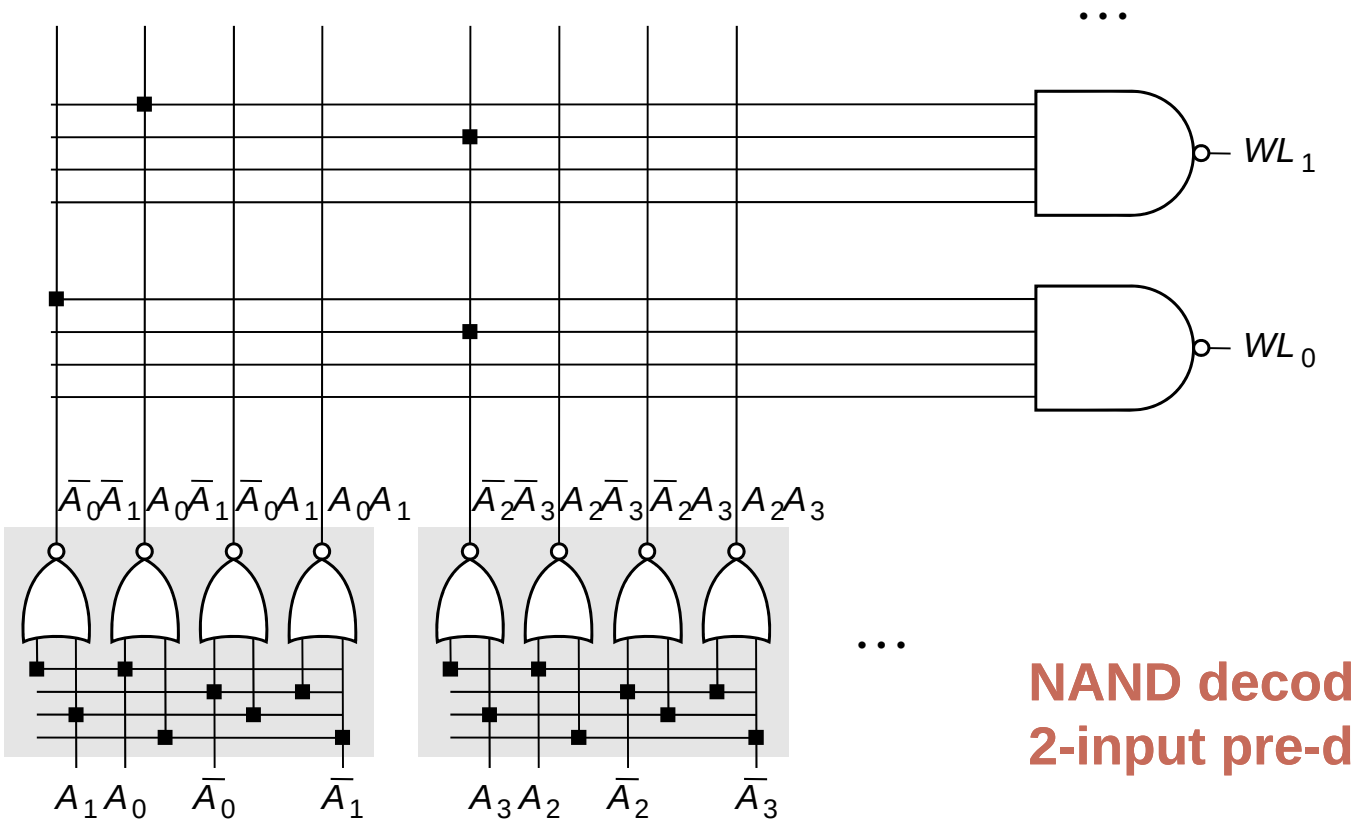
NOR Decoder

$$WL_0 = \overline{A_0 + A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8 + A_9}$$

$$WL_{511} = \overline{A_0 + \bar{A}_1 + \bar{A}_2 + \bar{A}_3 + \bar{A}_4 + \bar{A}_5 + \bar{A}_6 + \bar{A}_7 + \bar{A}_8 + \bar{A}_9}$$

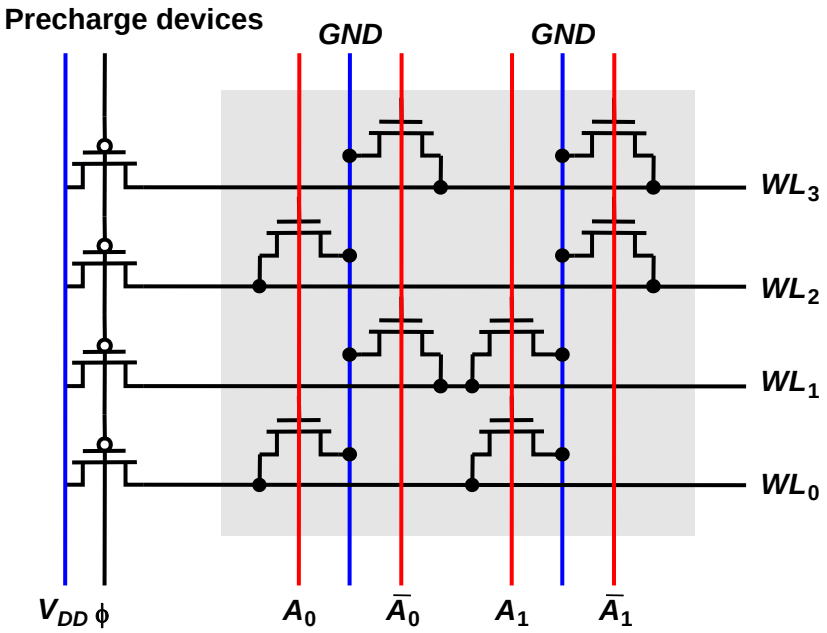
Hierarchical Decoders

Multi-stage implementation improves performance

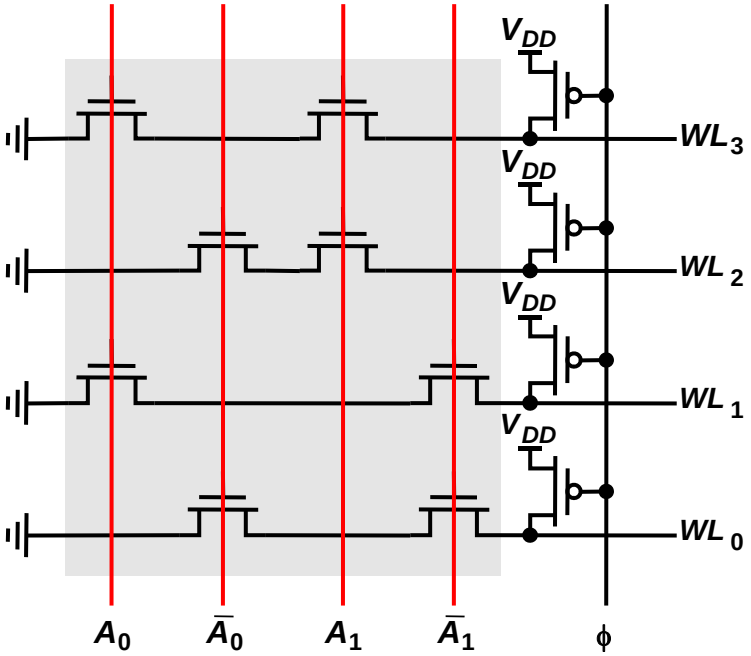


NAND decoder using 2-input pre-decoders

Dynamic Decoders

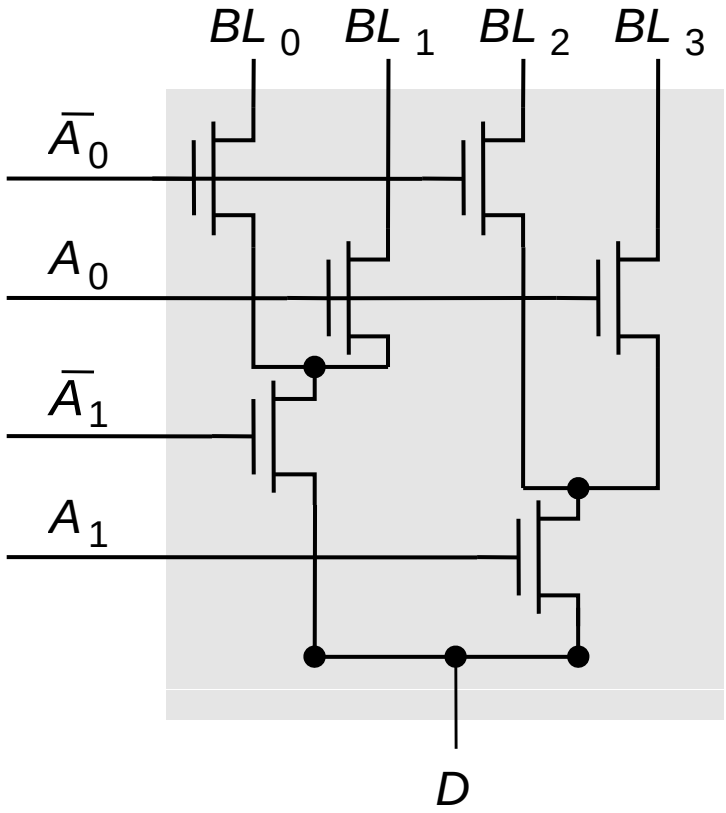


2-input NOR decoder



2-input NAND decoder

4-to-1 tree based column decoder



Number of devices drastically reduced
Delay increases quadratically with # of sections; prohibitive for large decoders
Solutions: buffers
progressive sizing
combination of tree and pass transistor approaches

Sense Amplifiers

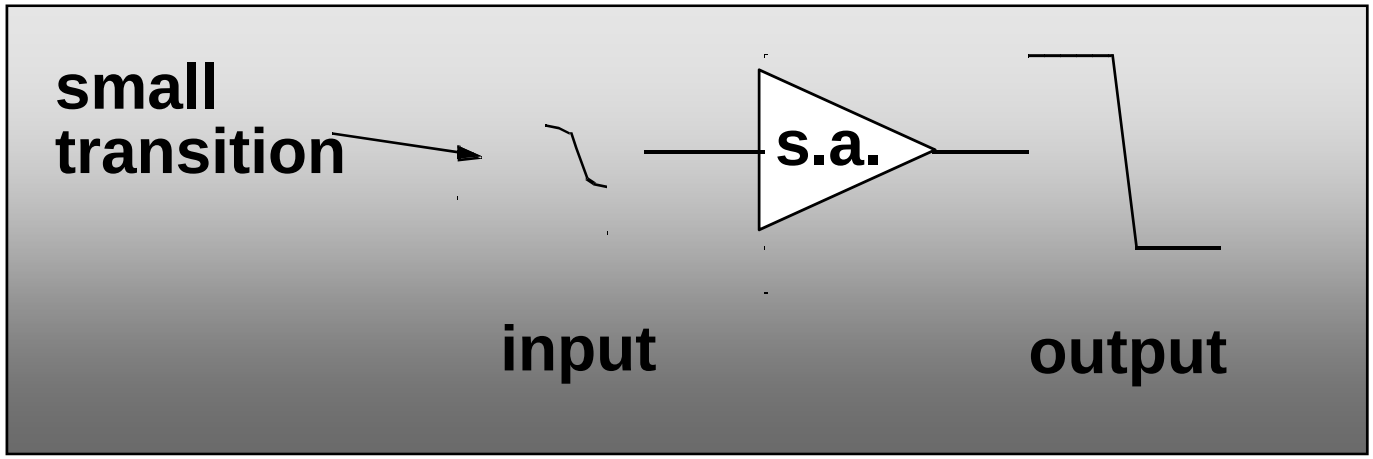
$$t_p = \frac{C \times \Delta V}{I_{av}}$$

make ΔV as small as possible

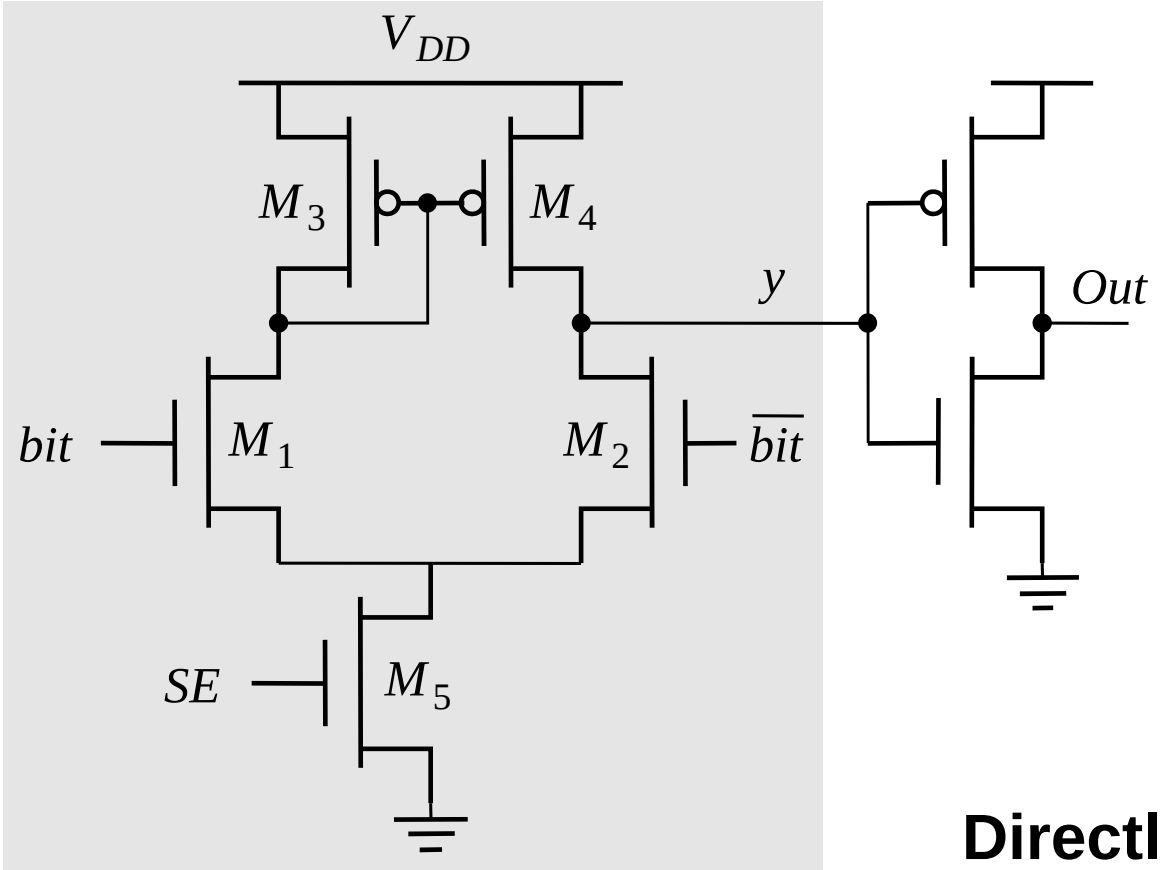
large

small

Idea: Use Sense Amplifer

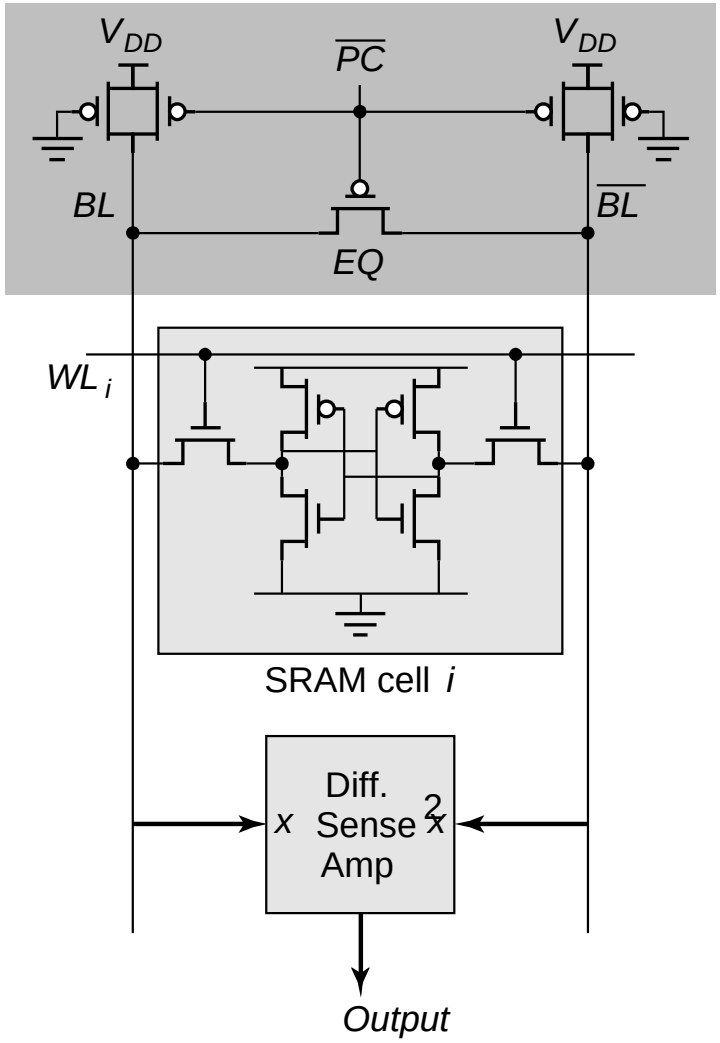


Differential Sense Amplifier

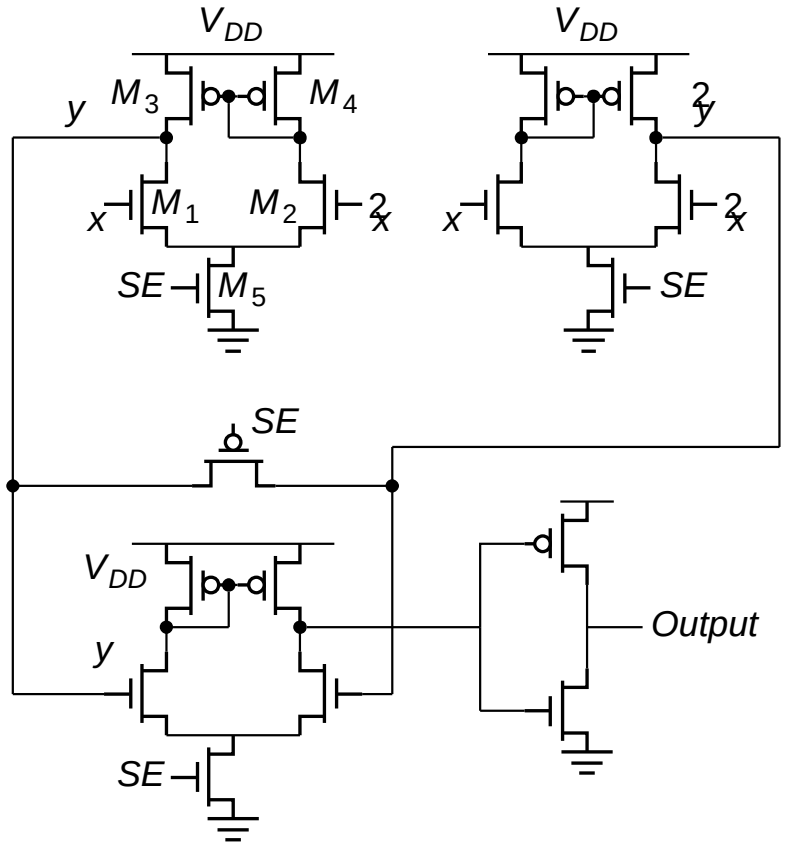


Directly applicable to SRAMs

Differential Sensing — SRAM

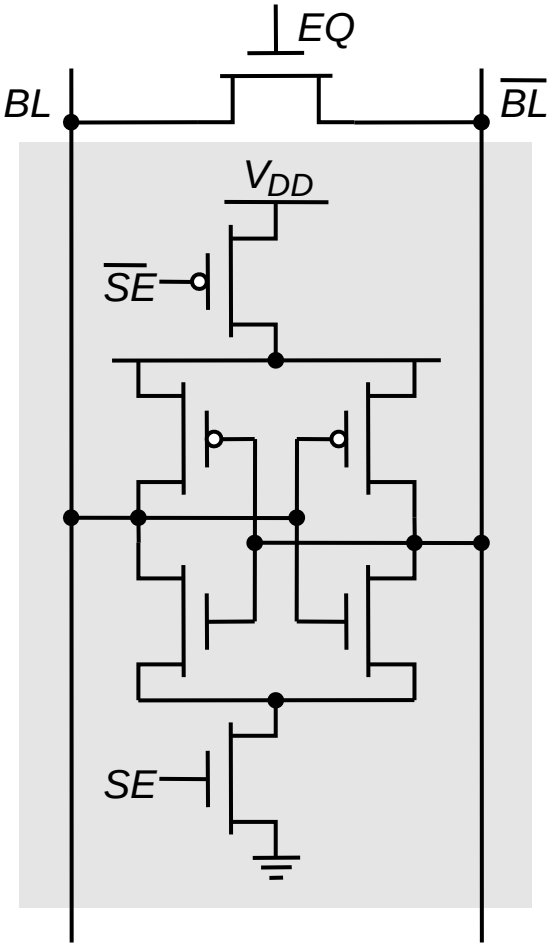


(a) SRAM sensing scheme



(b) two stage differential amplifier

Latch-Based Sense Amplifier (DRAM)

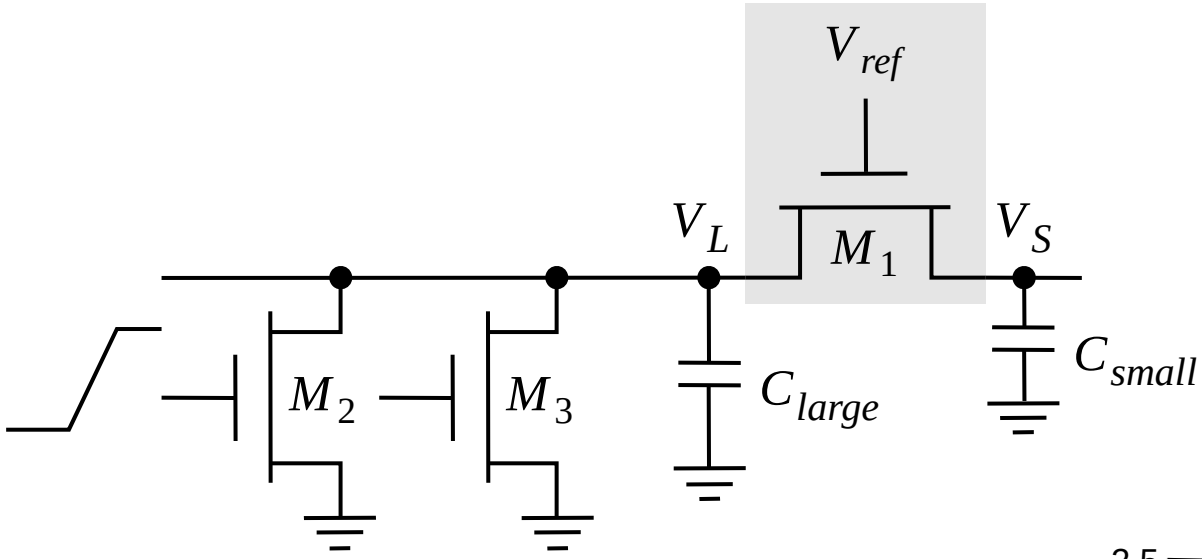


Initialized in its meta-stable point with EQ

Once adequate voltage gap is created, sense amp is enabled with SE

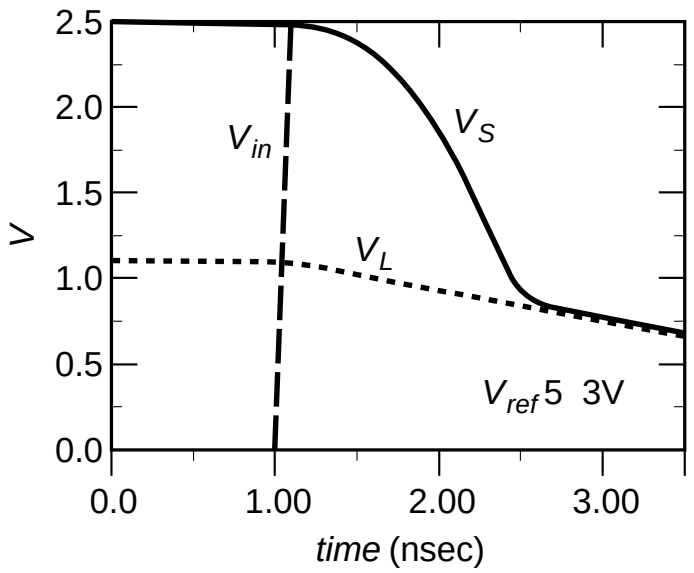
Positive feedback quickly forces output to a stable operating point.

Charge-Redistribution Amplifier

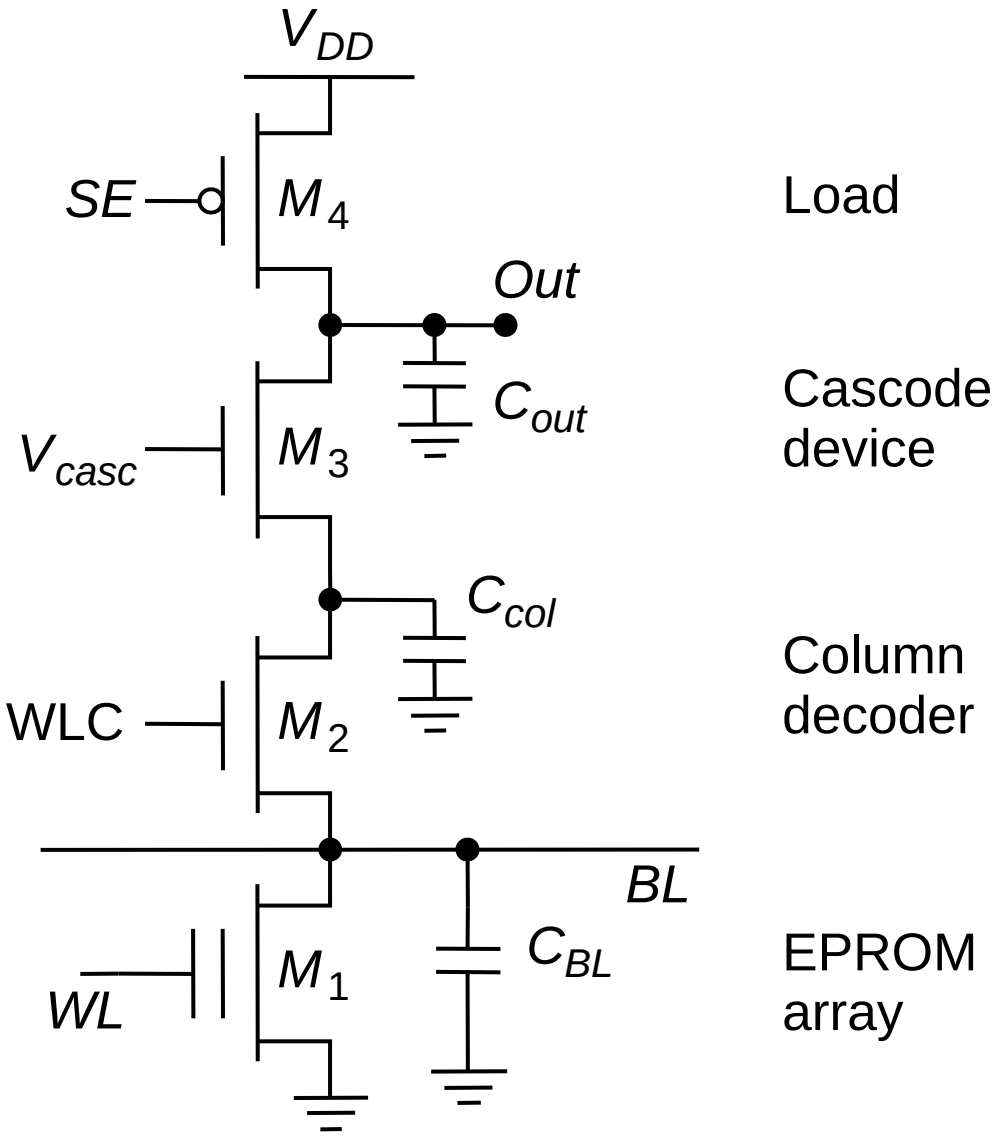


Concept

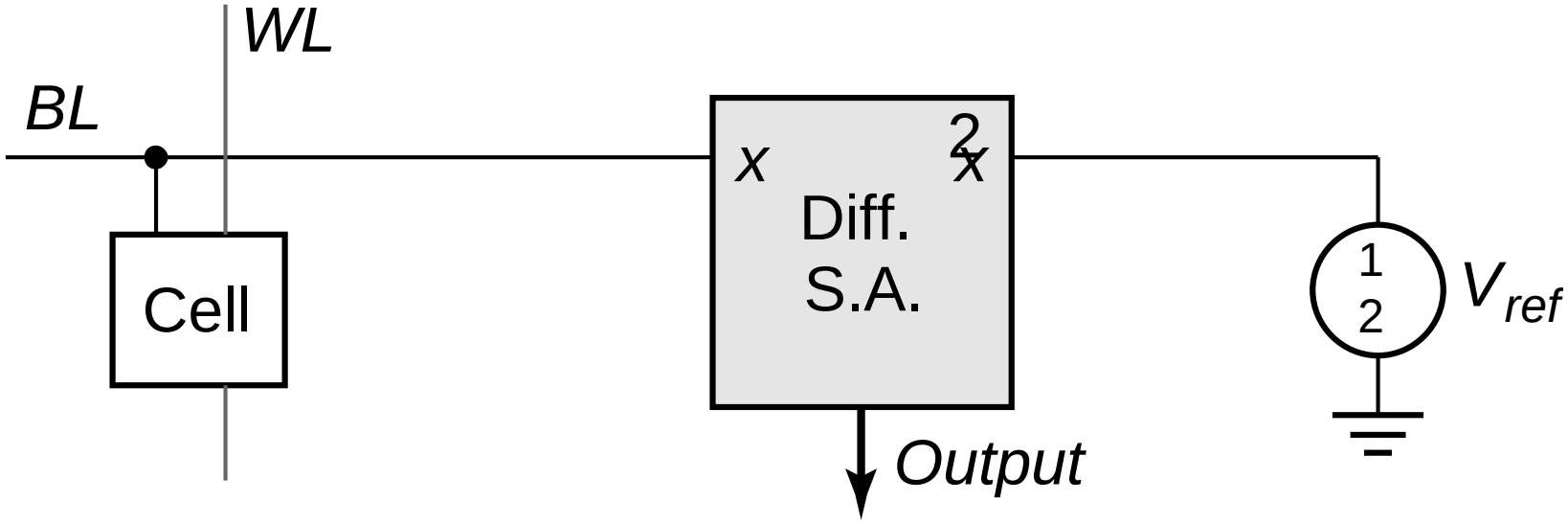
Transient Response



Charge-Redistribution Amplifier—EPROM

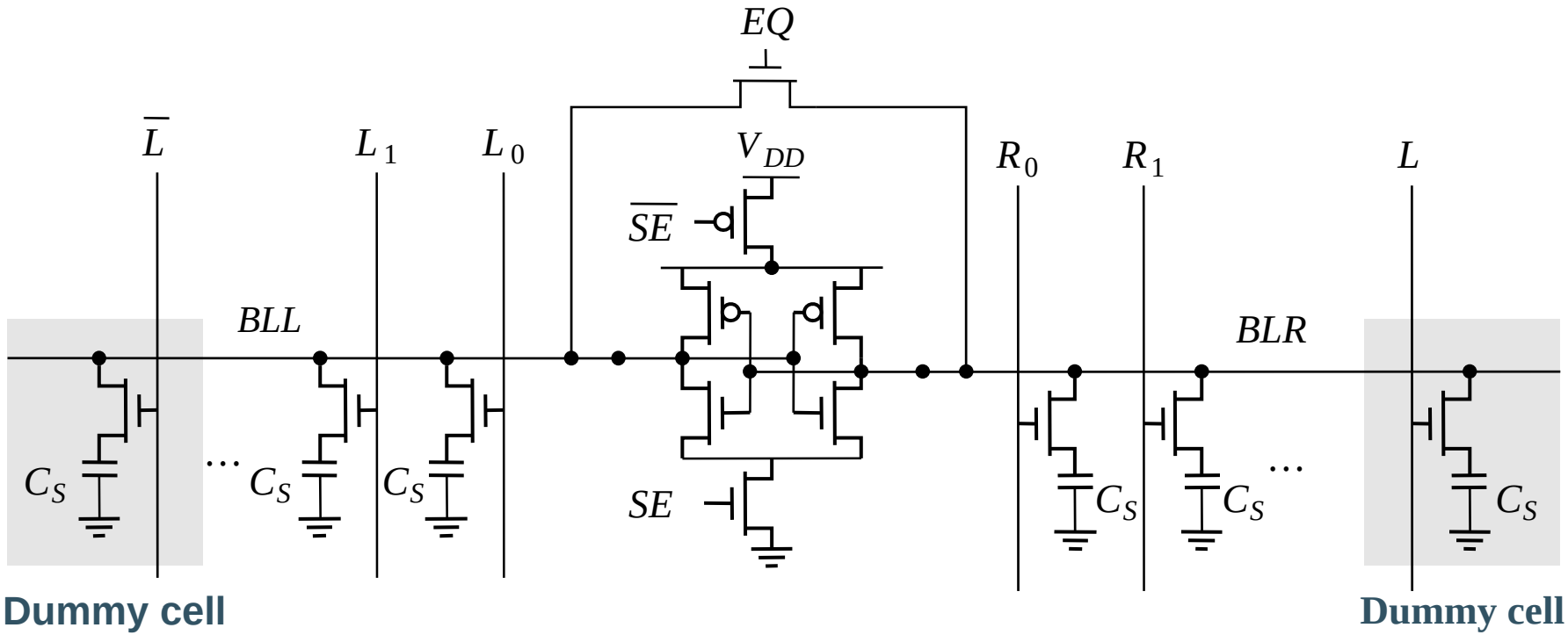


Single-to-Differential Conversion

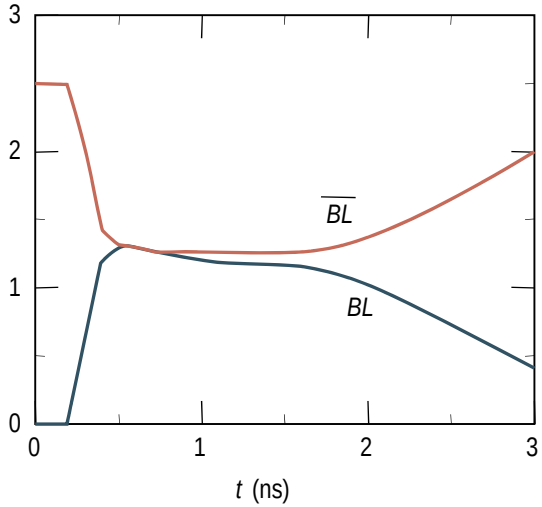


How to make a good V_{ref}?

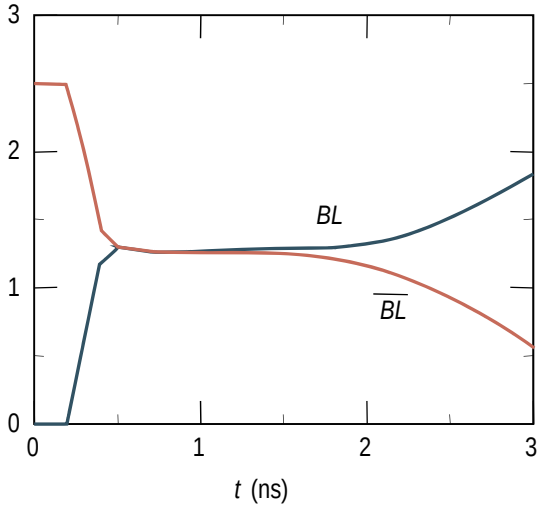
Open bitline architecture with dummy cells



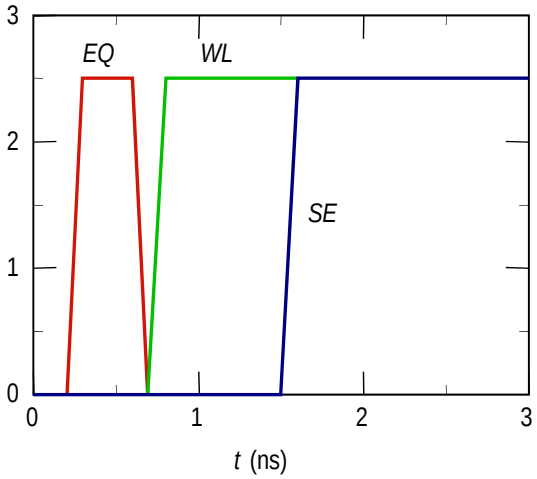
DRAM Read Process with Dummy Cell



reading 0

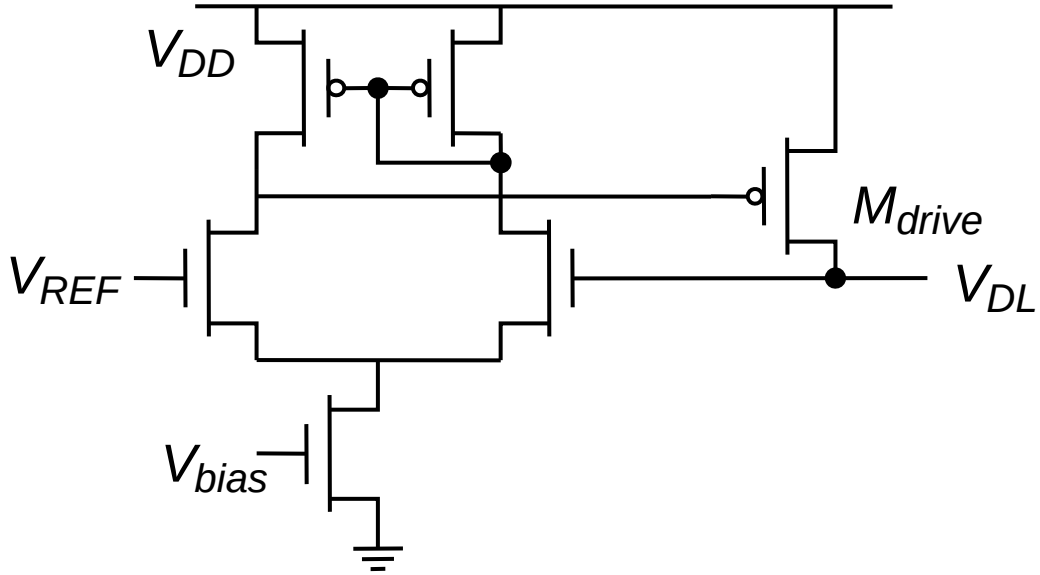


reading 1

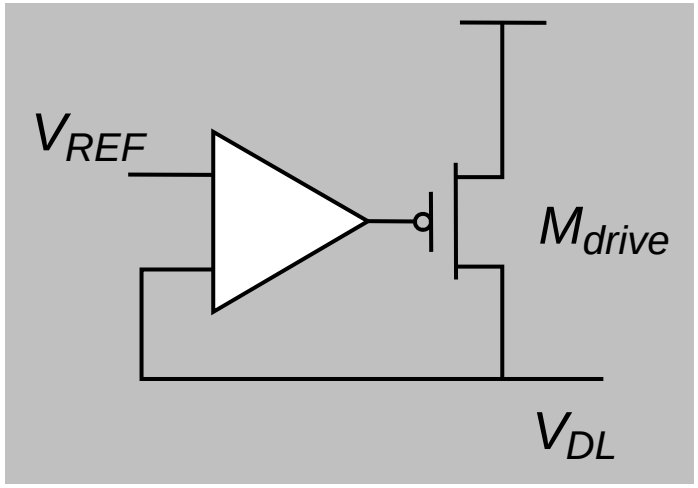


control signals

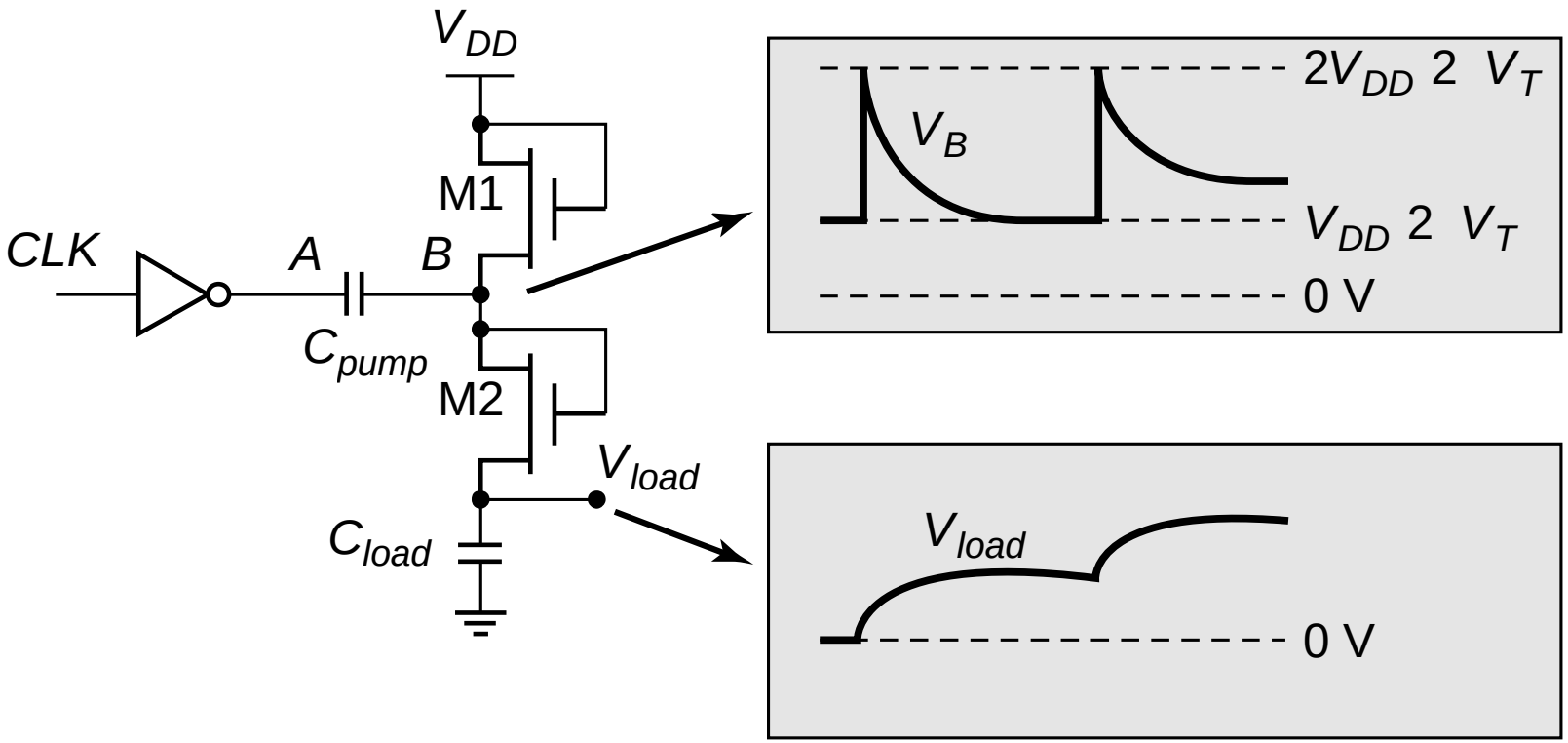
Voltage Regulator



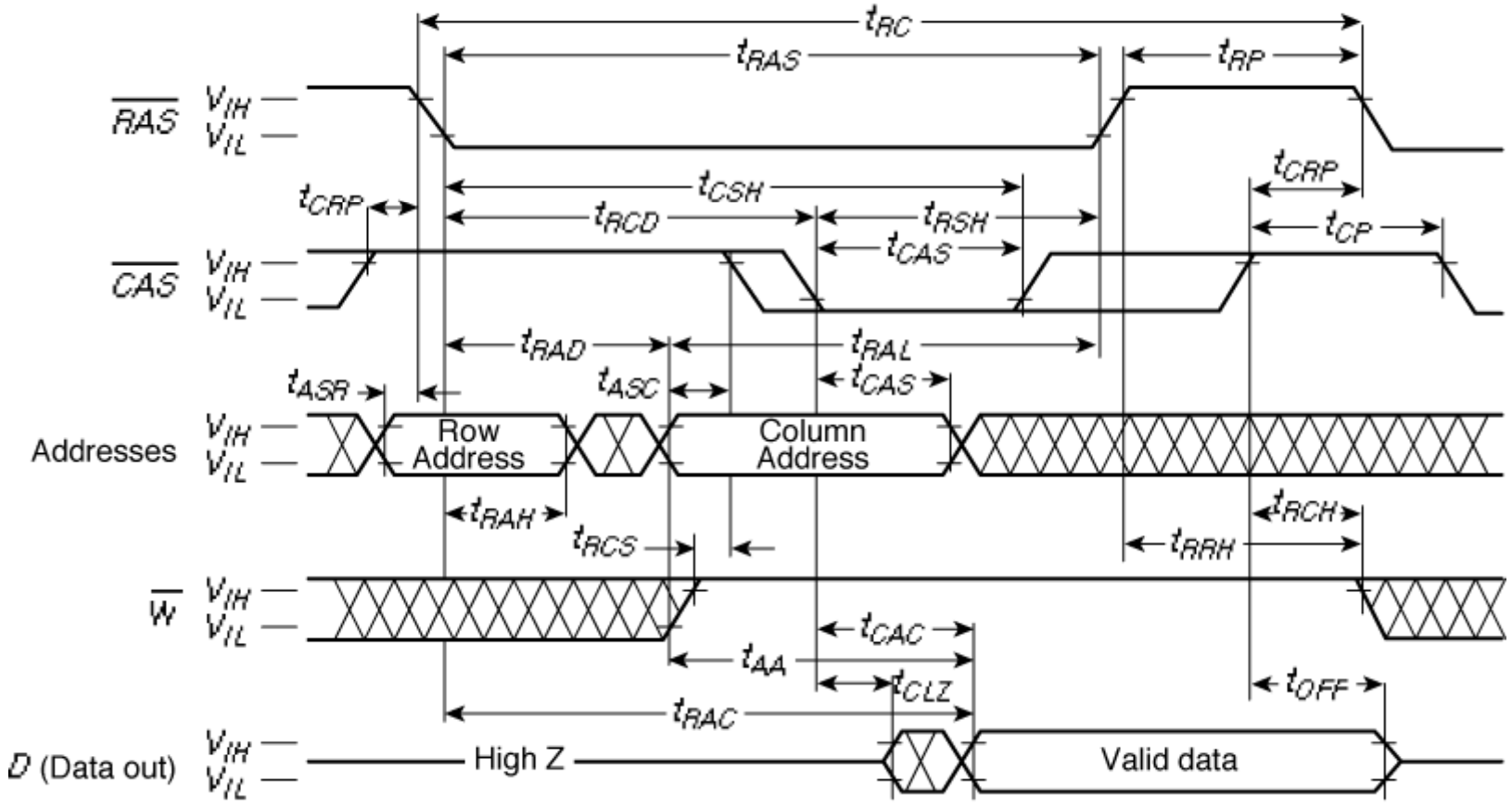
Equivalent Model



Charge Pump

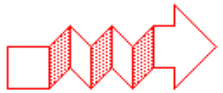

V_{DD}
M1
CLK
A
B
C_{pump}
M2
V_{load}
C_{load}
$2V_{DD}$ 2 V_T
V_B
V_{DD} 2 V_T
0 V
V_{load}
0 V

DRAM Timing


t_{RC}
t_{RAS}
t_{RP}
$\overline{RAS}$
V_{IH}
V_{IL}
t_{CRP}
t_{CSH}
t_{RCD}
t_{RSH}
t_{CAS}
t_{CRP}
t_{CP}
$\overline{CAS}$
t_{RAD}
t_{RAL}
t_{ASR}
t_{ASC}
t_{CAS}
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

Reliability and Yield

- Semiconductor memories trade off noise-margin for density and performance



Highly Sensitive to Noise (Crosstalk, Supply Noise)

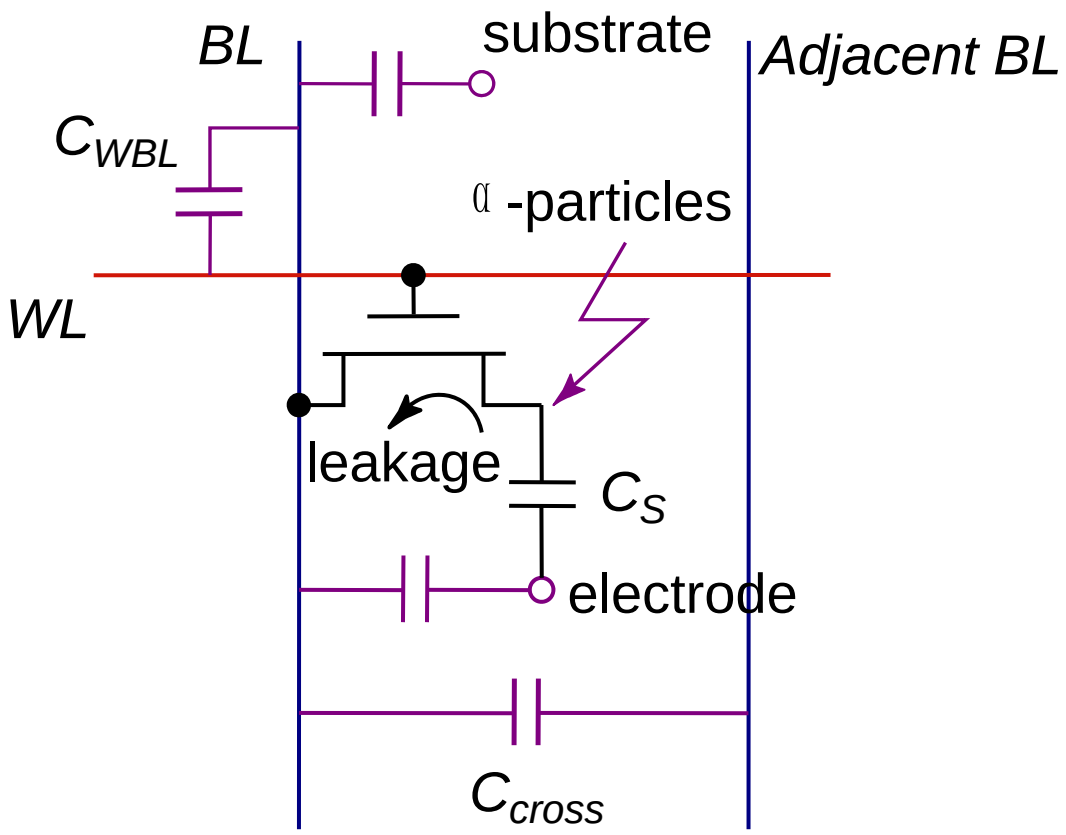
- High Density and Large Die size cause Yield Problems

$$Y = 100\frac{Number\text{""}of\text{""}Good\text{""}Chips\text{""}on\text{""}Wafer}{Number\text{""}of\text{""}Chips\text{""}on\text{""}Wafer}$$

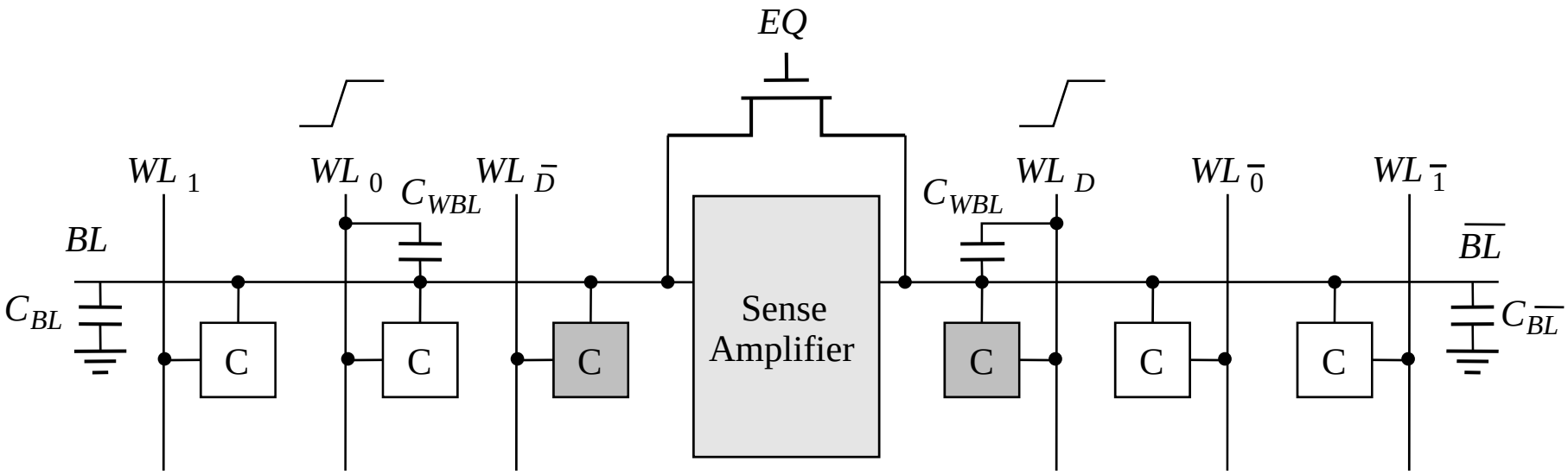
$$Y = \left[\frac{1 - e^{-AD}}{AD}\right]^2$$

Increase Yield using Error Correction and Redundancy

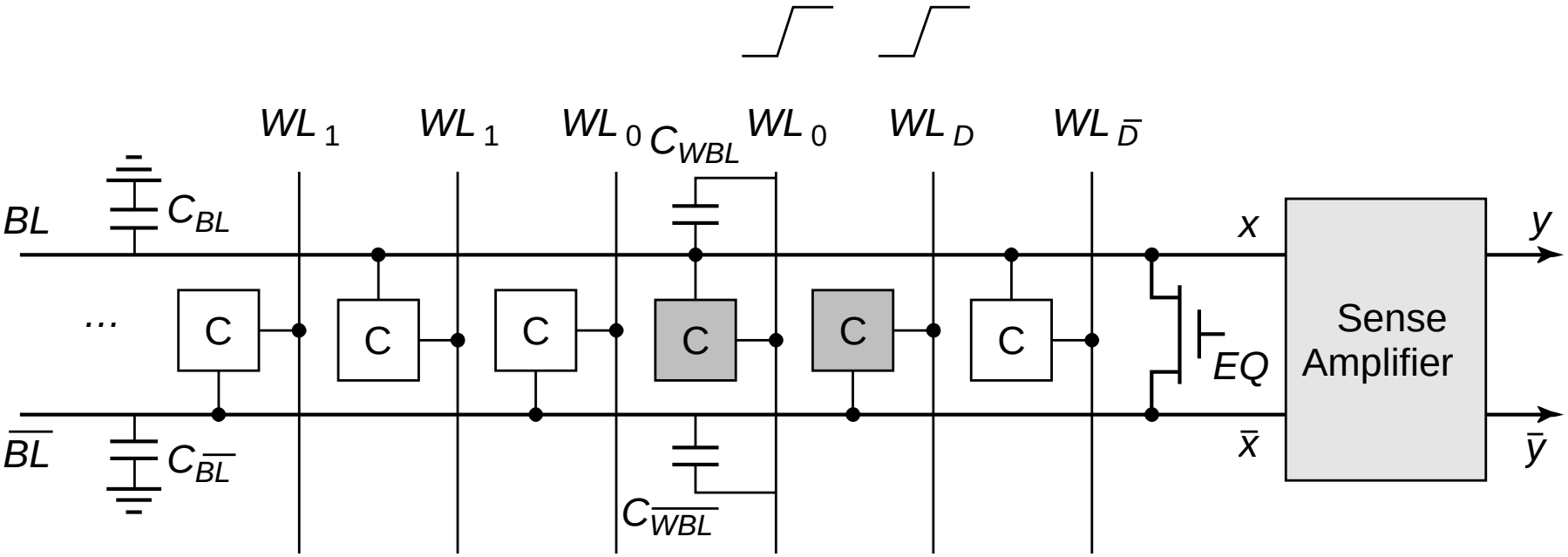
Noise Sources in 1T DRam



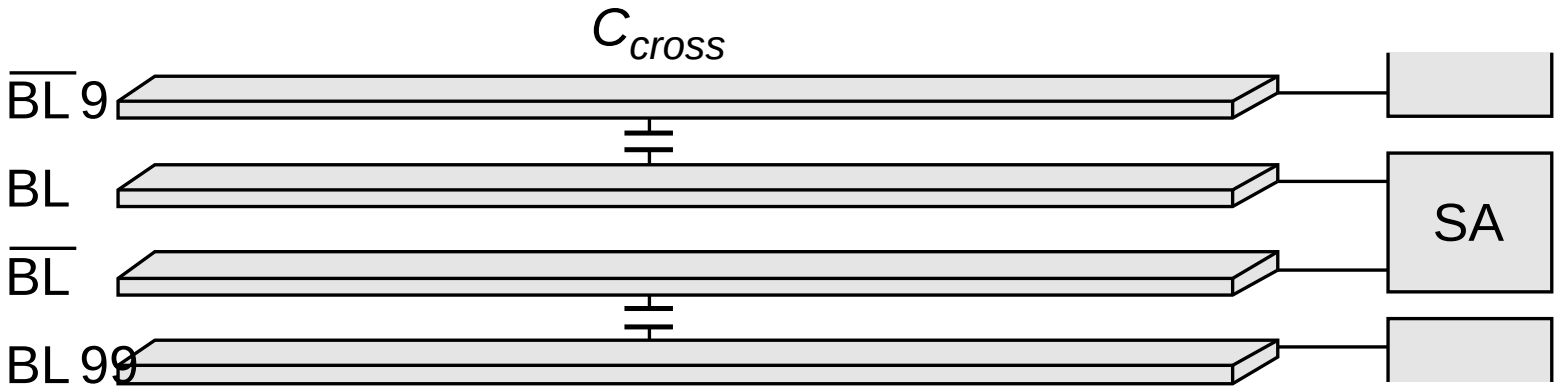
Open Bit-line Architecture —Cross Coupling



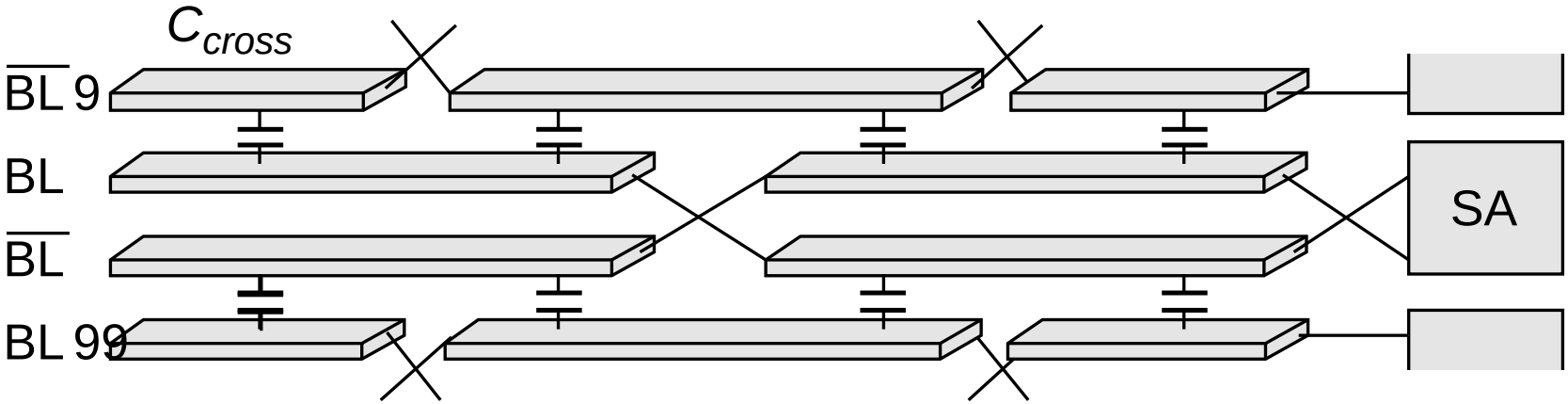
Folded-Bitline Architecture



Transposed-Bitline Architecture

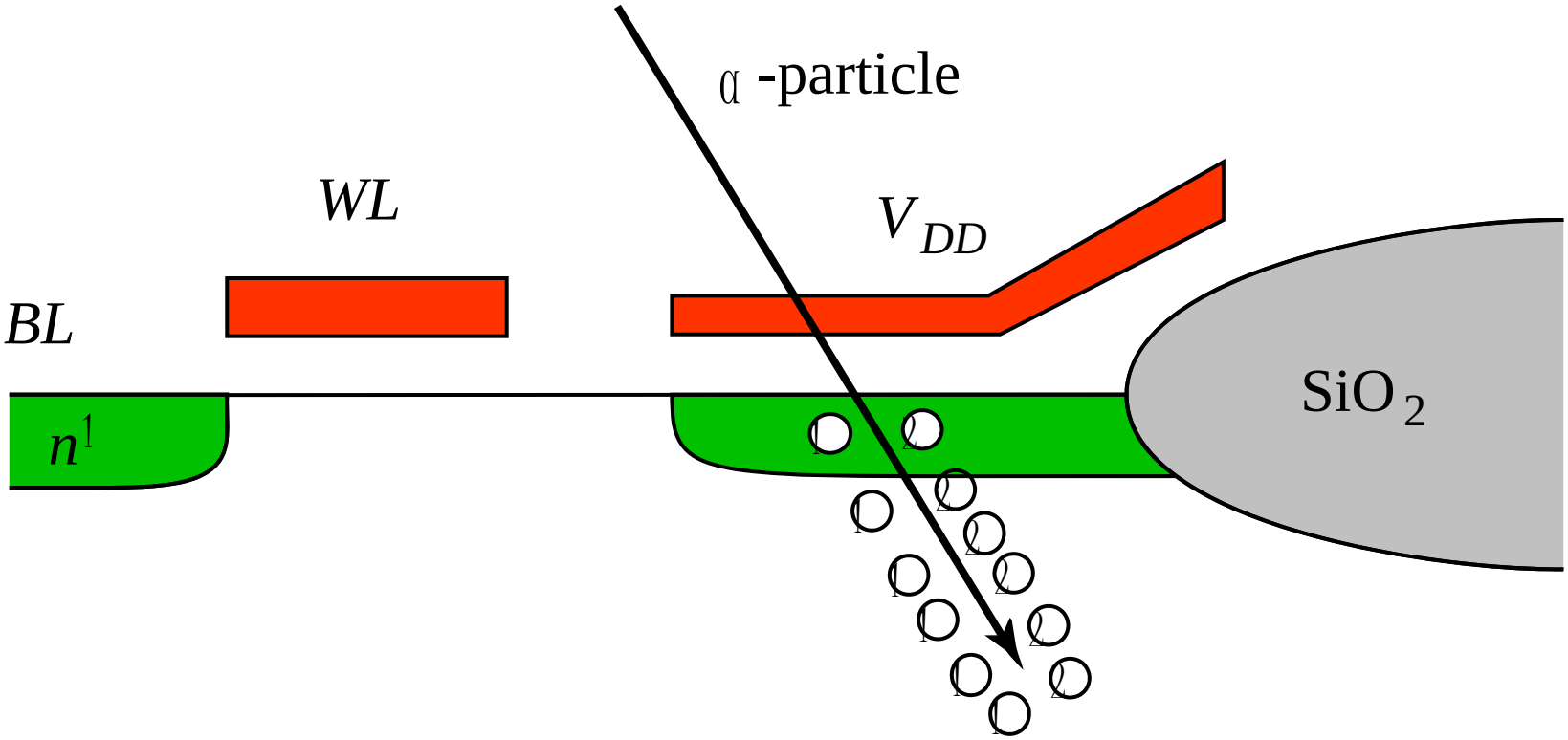


(a) Straightforward bit-line routing



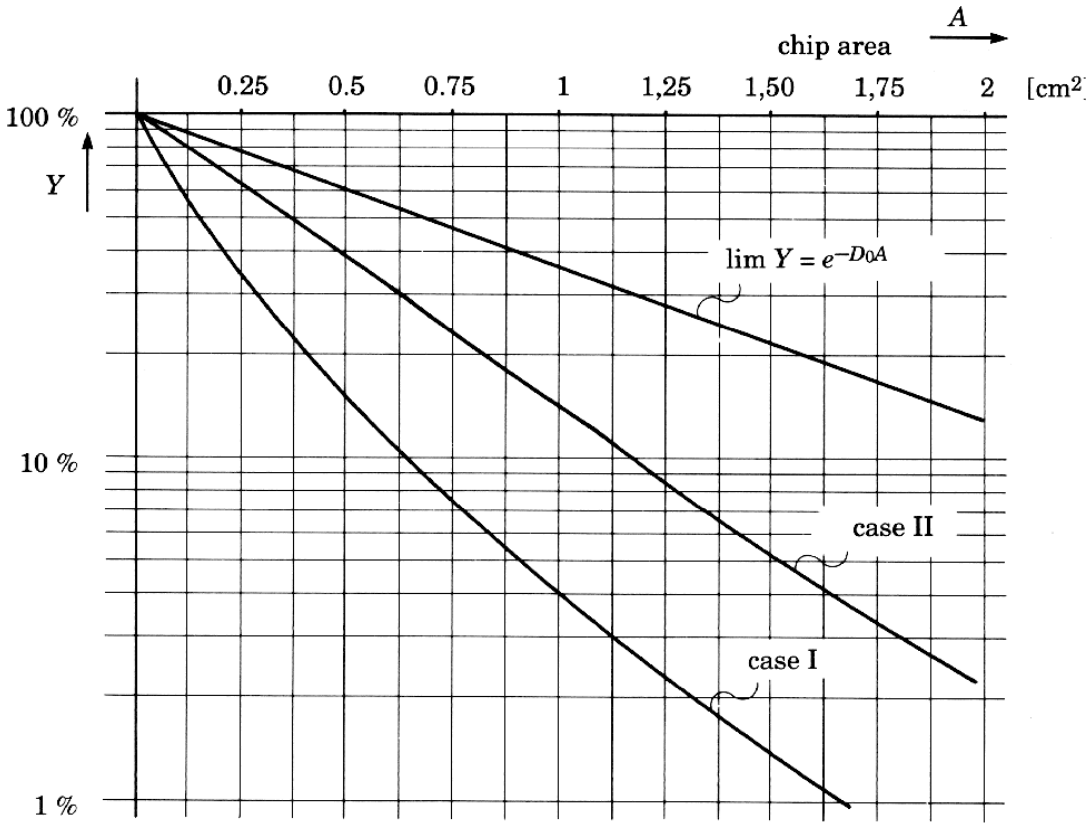
(b) Transposed bit-line architecture

Alpha-particles (or Neutrons)



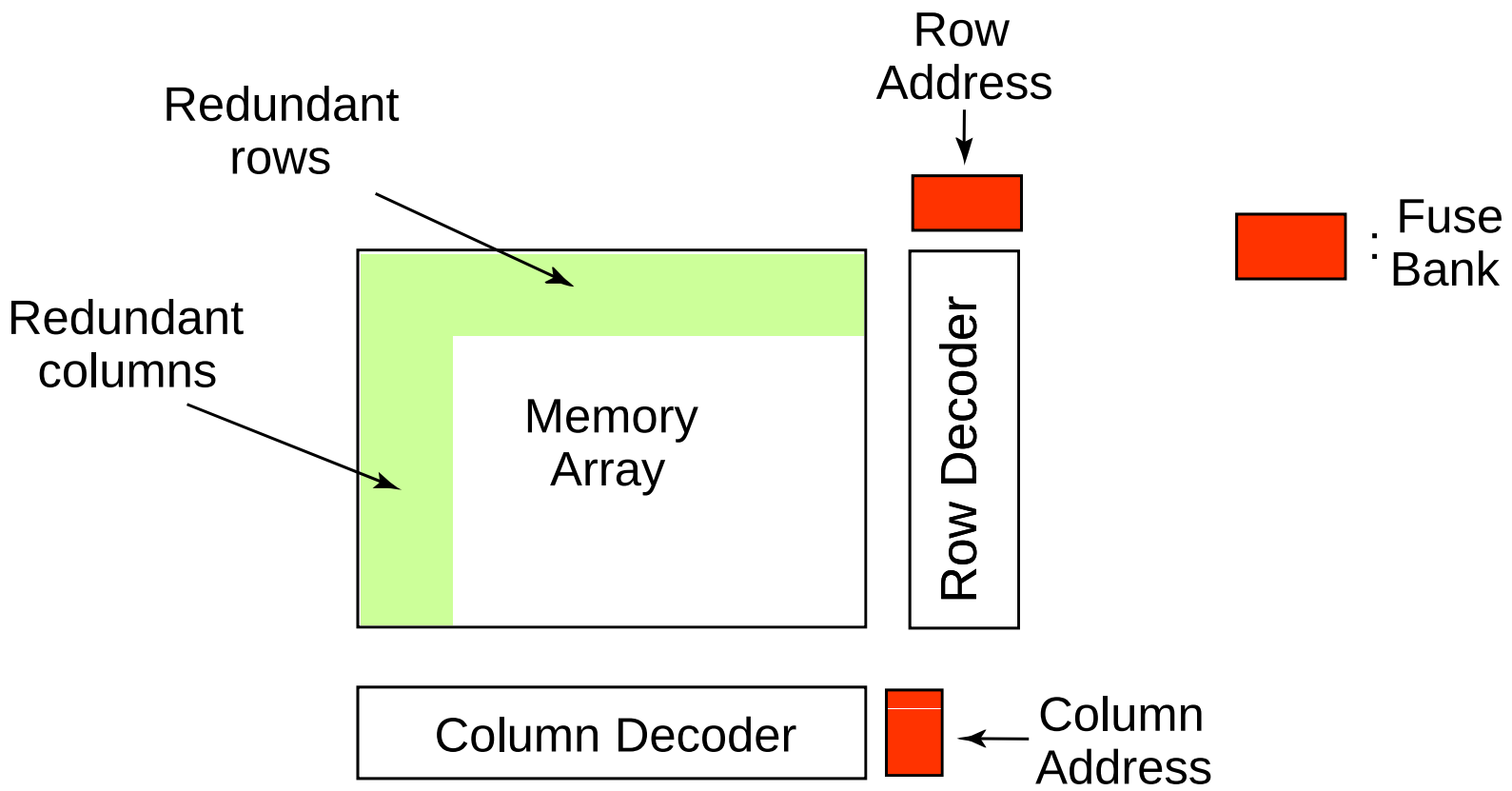
1 Particle ~ 1 Million Carriers

Yield



Yield curves at different stages of process maturity

Redundancy


Row
Address
Redundant
rows
Redundant
columns
Memory
Array
Row Decoder
Fuse
Bank
Column Decoder
Column
Address

Error-Correcting Codes

Example: Hamming Codes

$$P_1\, P_2\, B_3\, P_4\, B_5\, B_6\, B_7$$

with

$$P_1 \oplus B_3 \oplus B_5 \oplus B_7 = 0$$

$$P_2 \oplus B_3 \oplus B_6 \oplus B_7 = 0$$

$$P_4 \oplus B_5 \oplus B_6 \oplus B_7 = 0$$

e.g. B3 Wrong

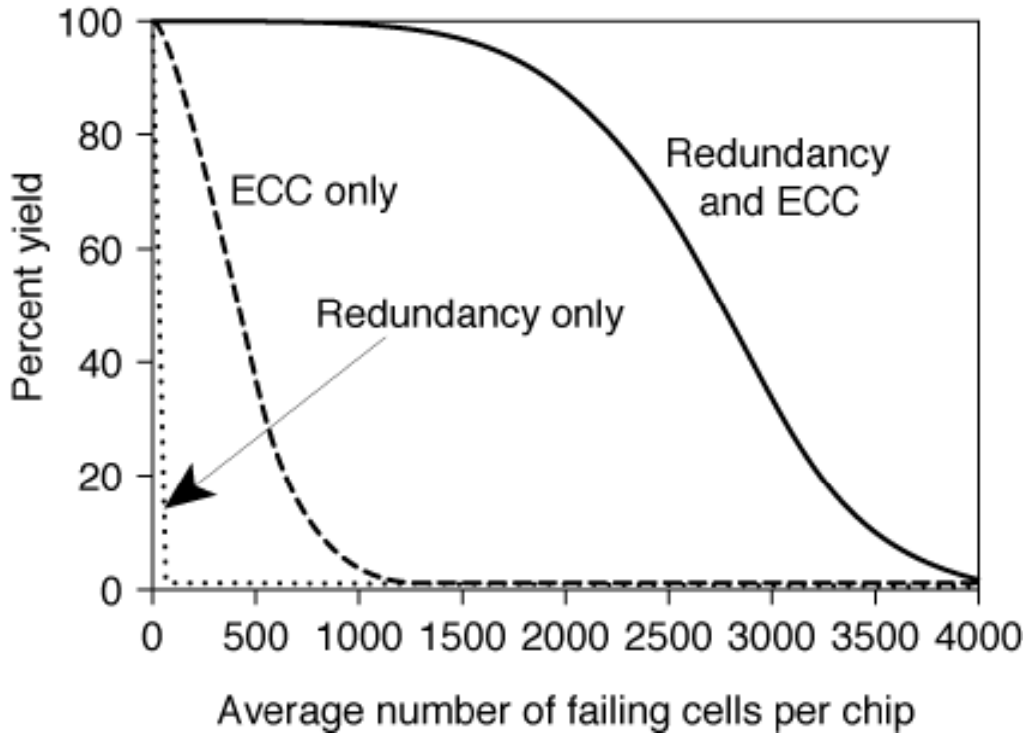
1

1

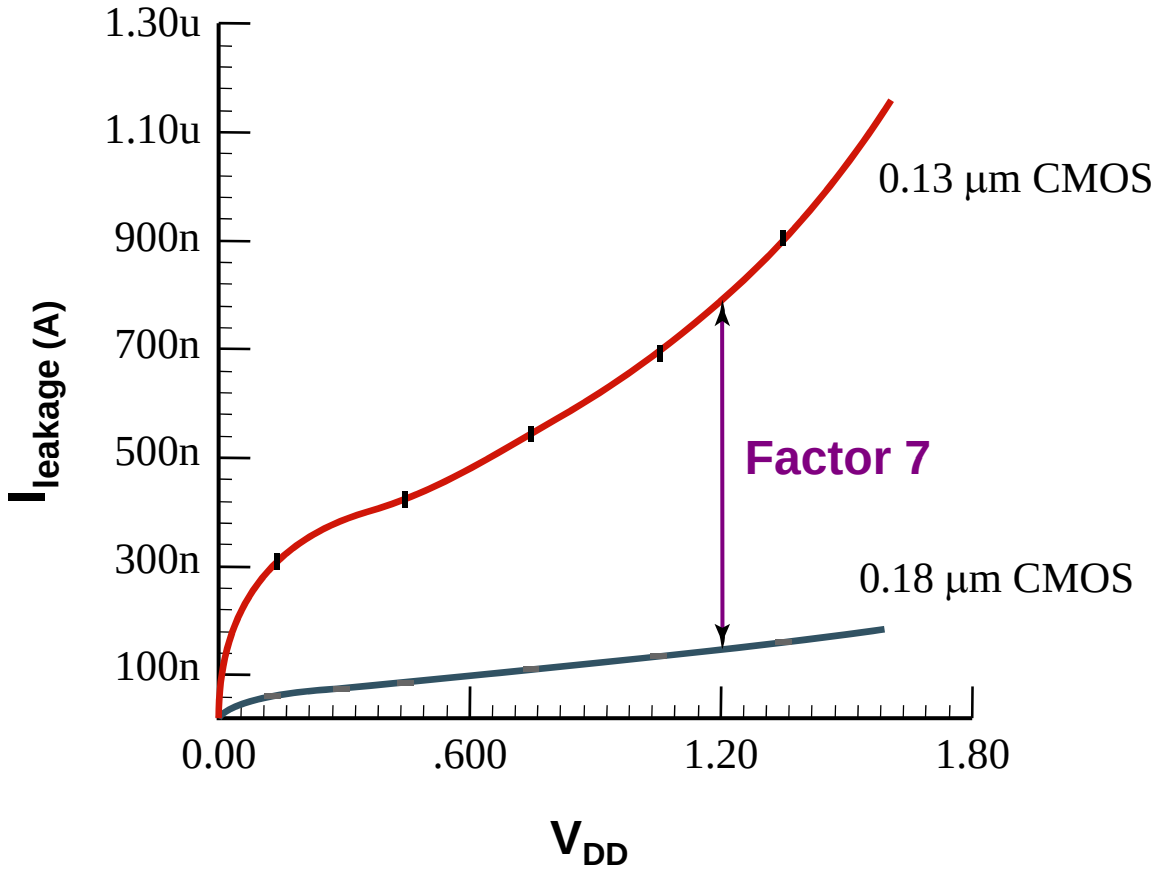
0

= 3

Redundancy and Error Correction

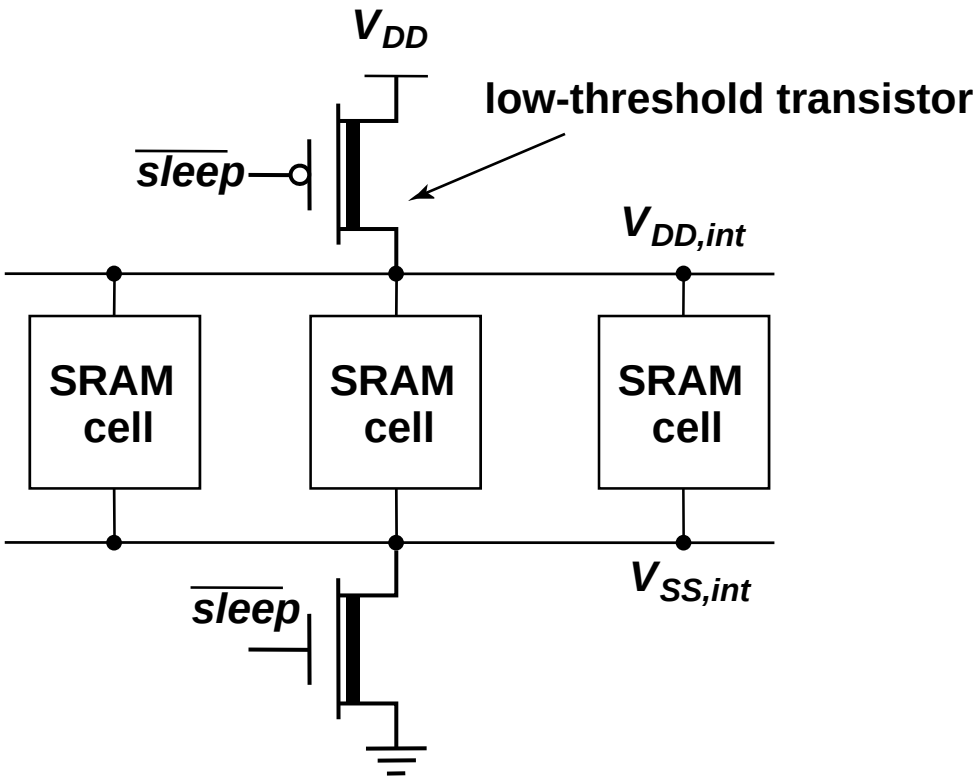


Data Retention in SRAM

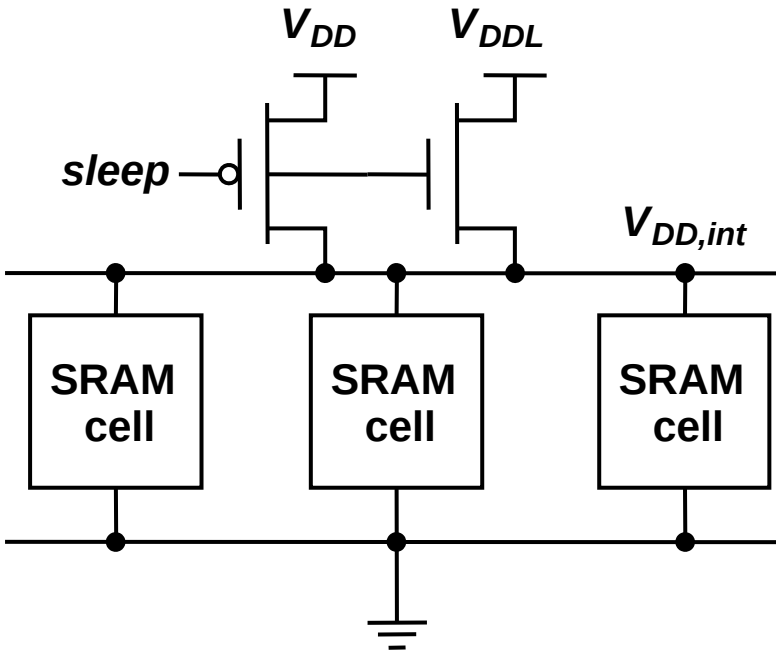


SRAM leakage increases with technology scaling

Suppressing Leakage in SRAM



Inserting Extra Resistance



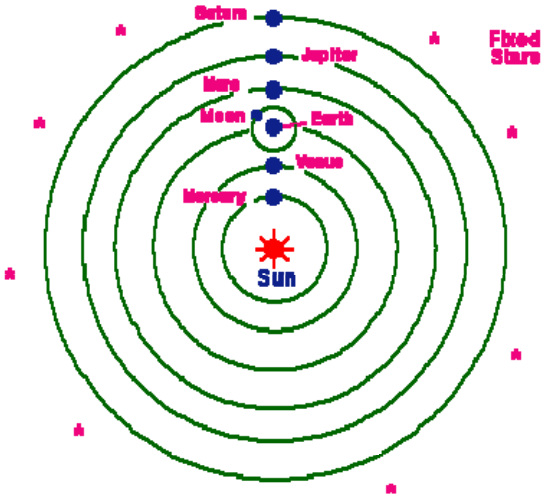
Reducing the supply voltage

Conclusions

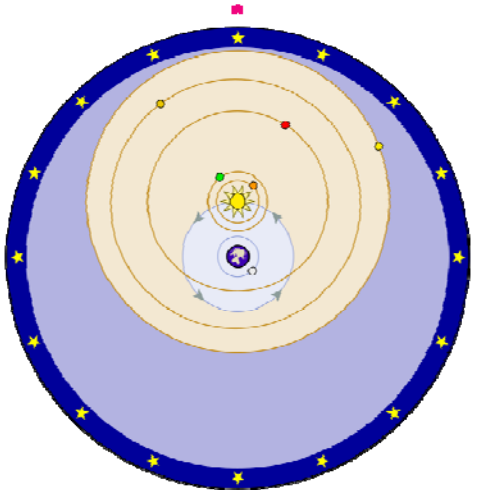
- The field of memory design is a dynamic and exciting specialty:
  - Coordinated efforts from marketing and planning, process design, device design, circuit design, test & production engineering, and software engineering are all needed.
  - Many innovative approaches are possible in every design stage.
  - The market competition is fierce, but the winner is awarded with a big prize.
  - The leading memory technologies are being pioneered by domestic companies/engineers.

Conclusions

- Think flexible and think “big.”



- Succeed as an engineer.



Engineer =

sElf-Motivated

eNergetic

Self-manaGed

Insightful

iNnovative

Eye on data

Execute

Rewards