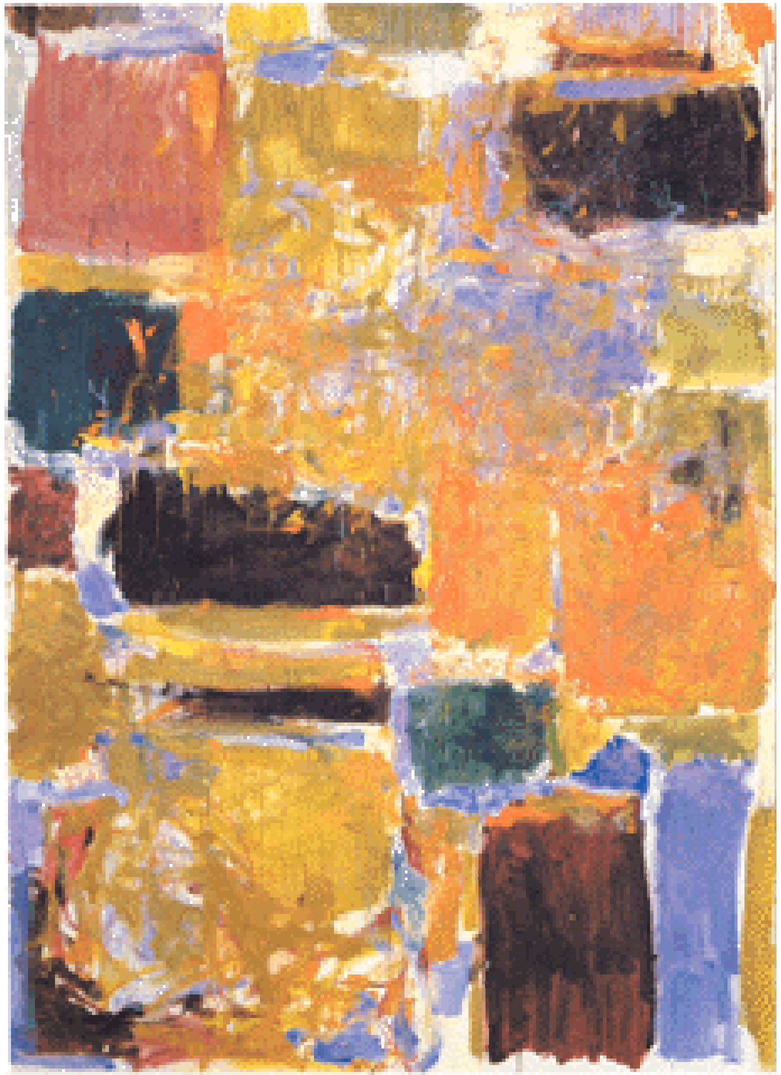


Digital Integrated Circuits
A Design Perspective

Jan M. Rabaey
Anantha Chandrakasan
Borivoje Nikolic

Semiconductor Memories

December 20, 2002

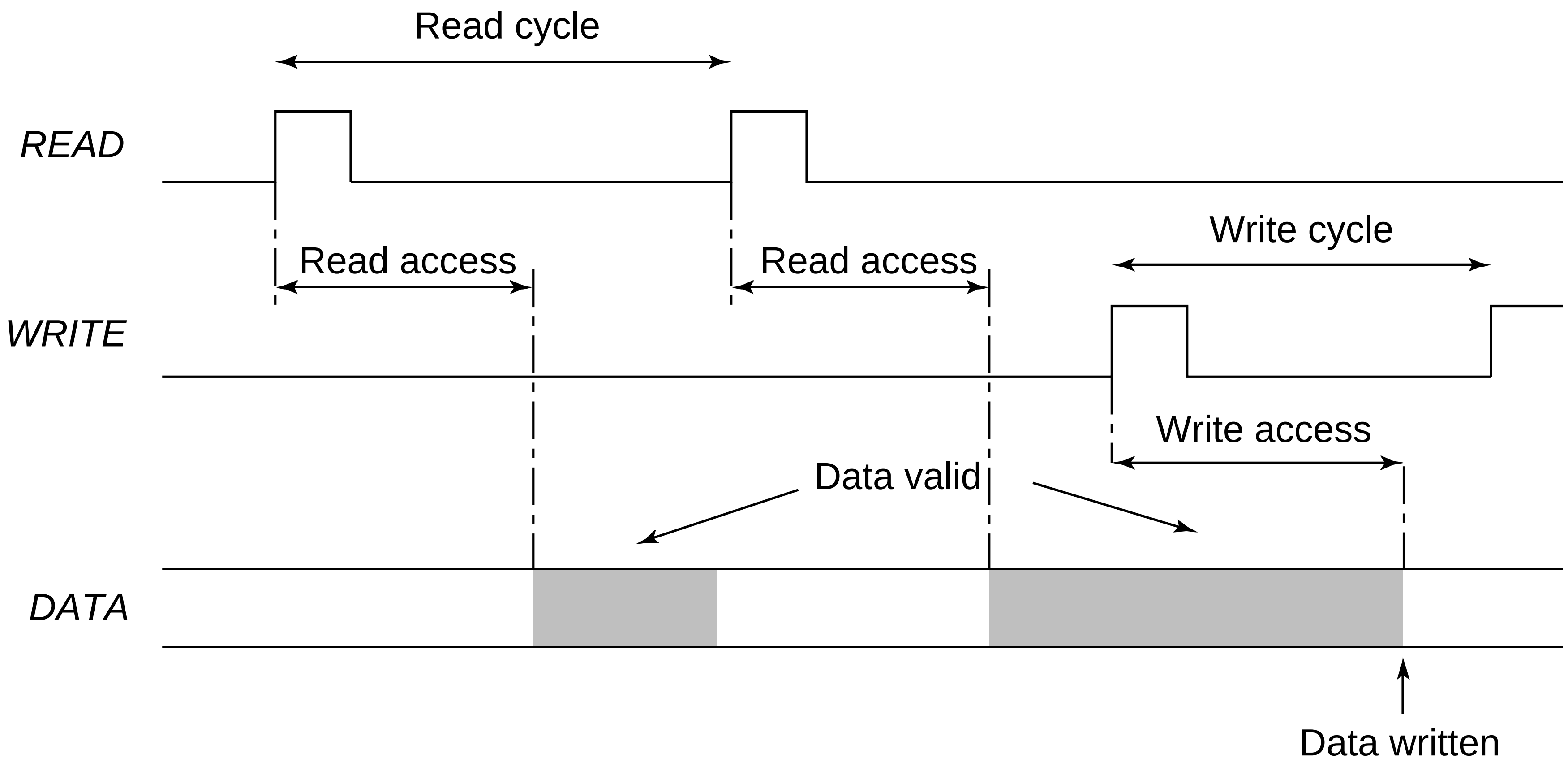
Chapter Overview

- □ Memory Classification
- □ Memory Architectures
- □ The Memory Core
- □ Periphery
- □ Reliability
- □ Case Studies

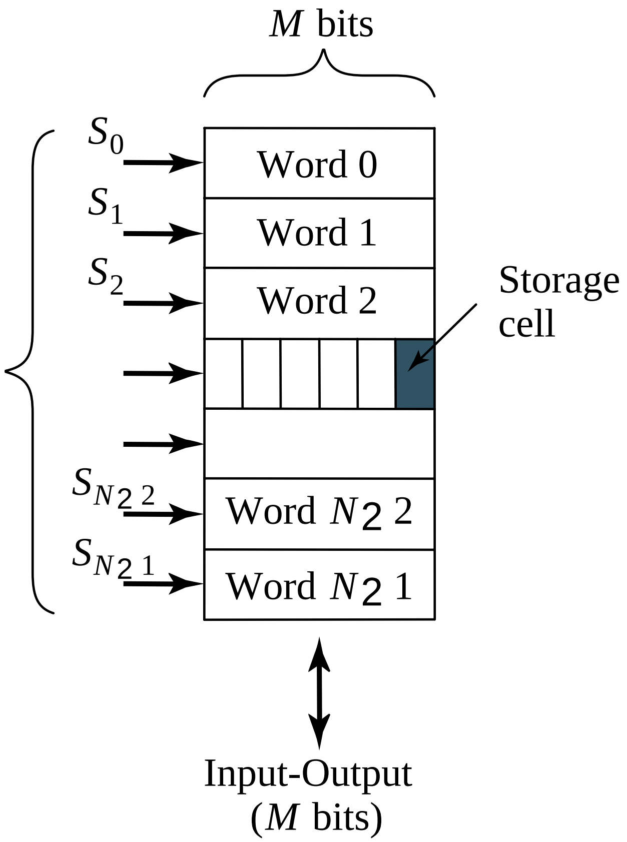
Semiconductor Memory Classification

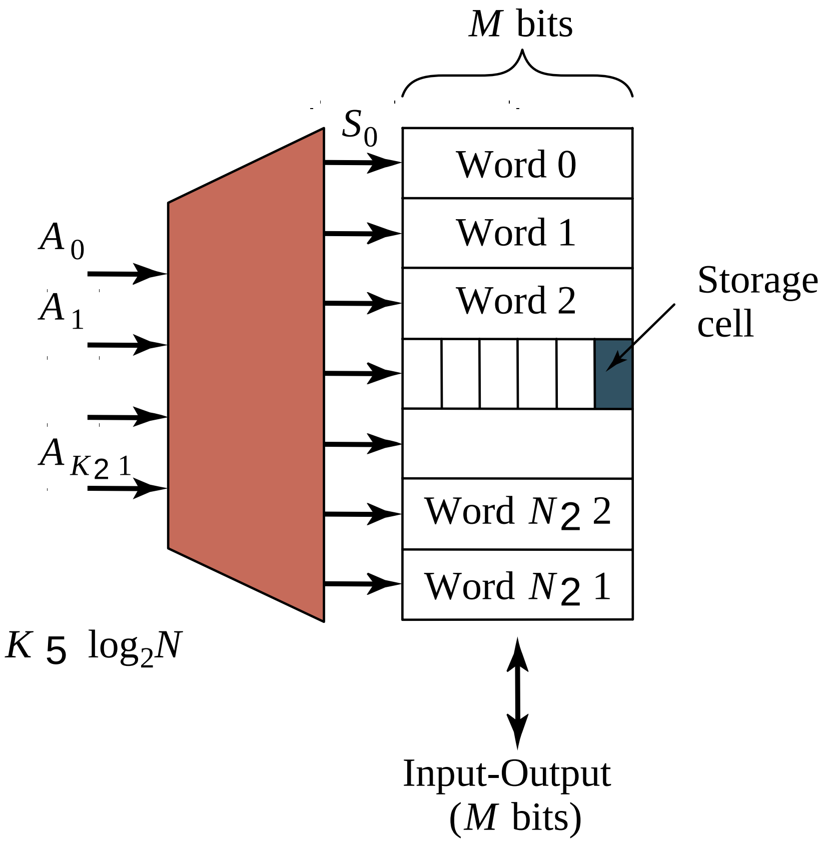
Read-Write Memory		Non-Volatile Read-Write Memory	Read-Only Memory
Random Access	**Non-Random Access**	EPROM E^{2}PROM FLASH	Mask-Programmed Programmable (PROM)
SRAM DRAM	FIFO LIFO Shift Register CAM		

Memory Timing: Definitions



Memory Architecture: Decoders





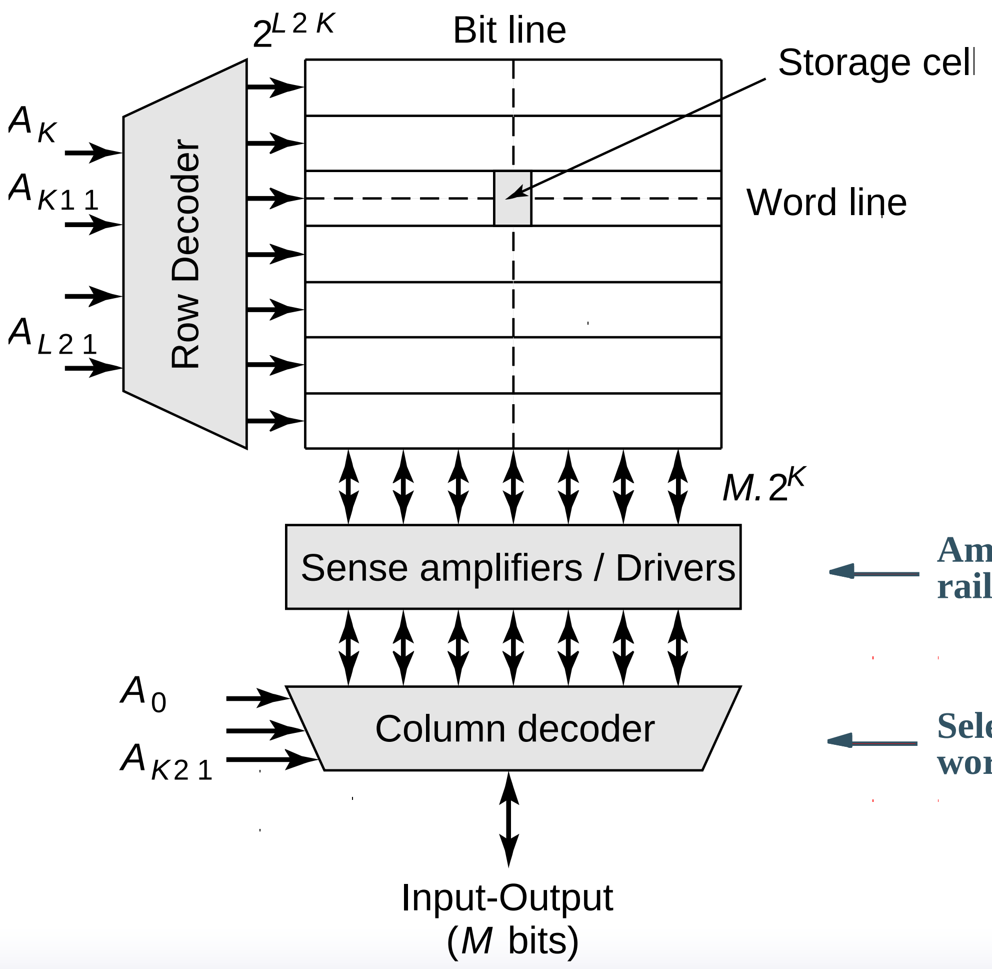
Intuitive architecture for N x M memory
Too many select signals:
N words == N select signals

Decoder reduces the number of select signals

$K = \log_2 N$

Array-Structured Memory Architecture

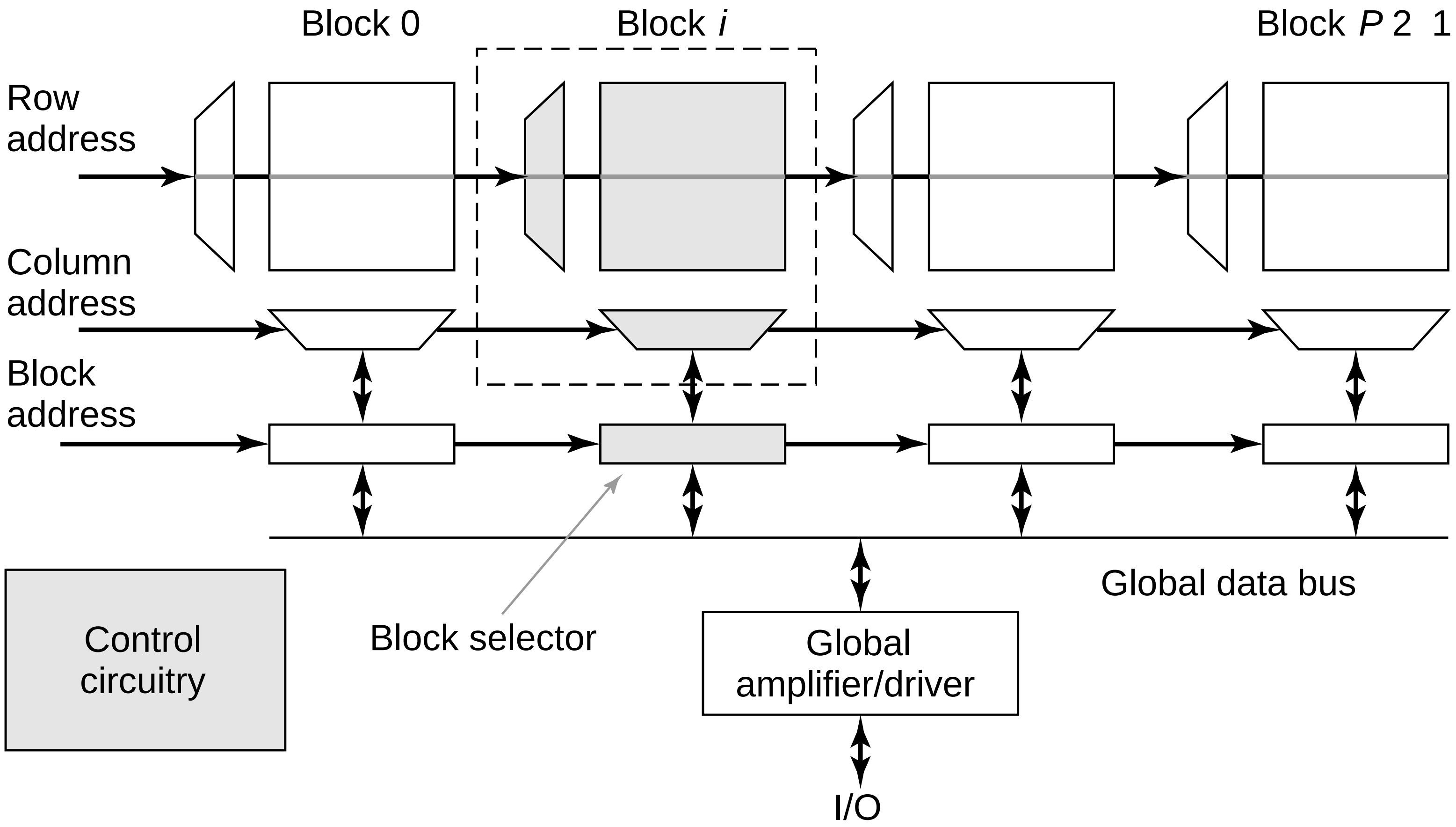
Problem: ASPECT RATIO or HEIGHT >> WIDTH



Amplify swing to rail-to-rail amplitude

Selects appropriate word

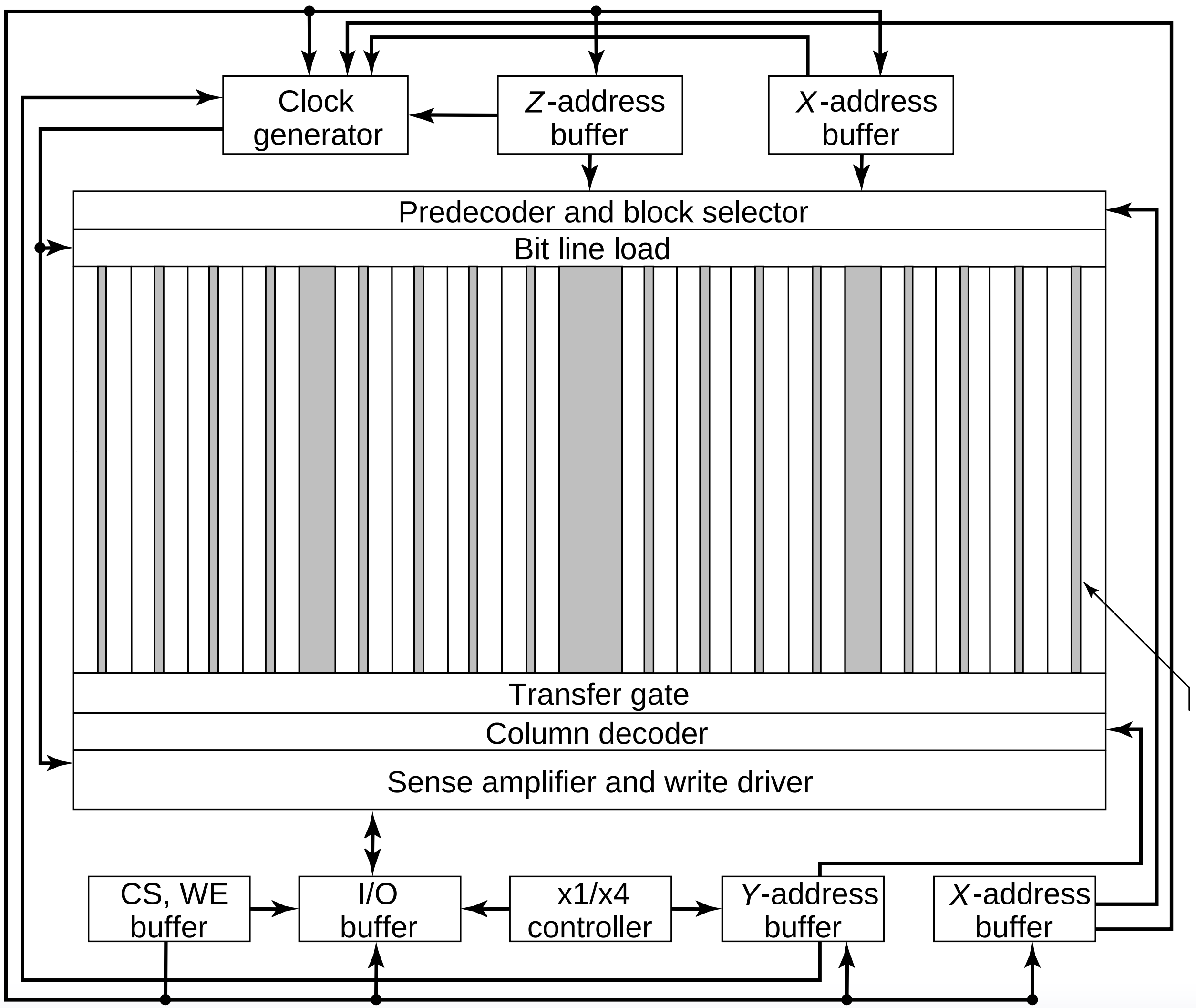
Hierarchical Memory Architecture



Advantages:

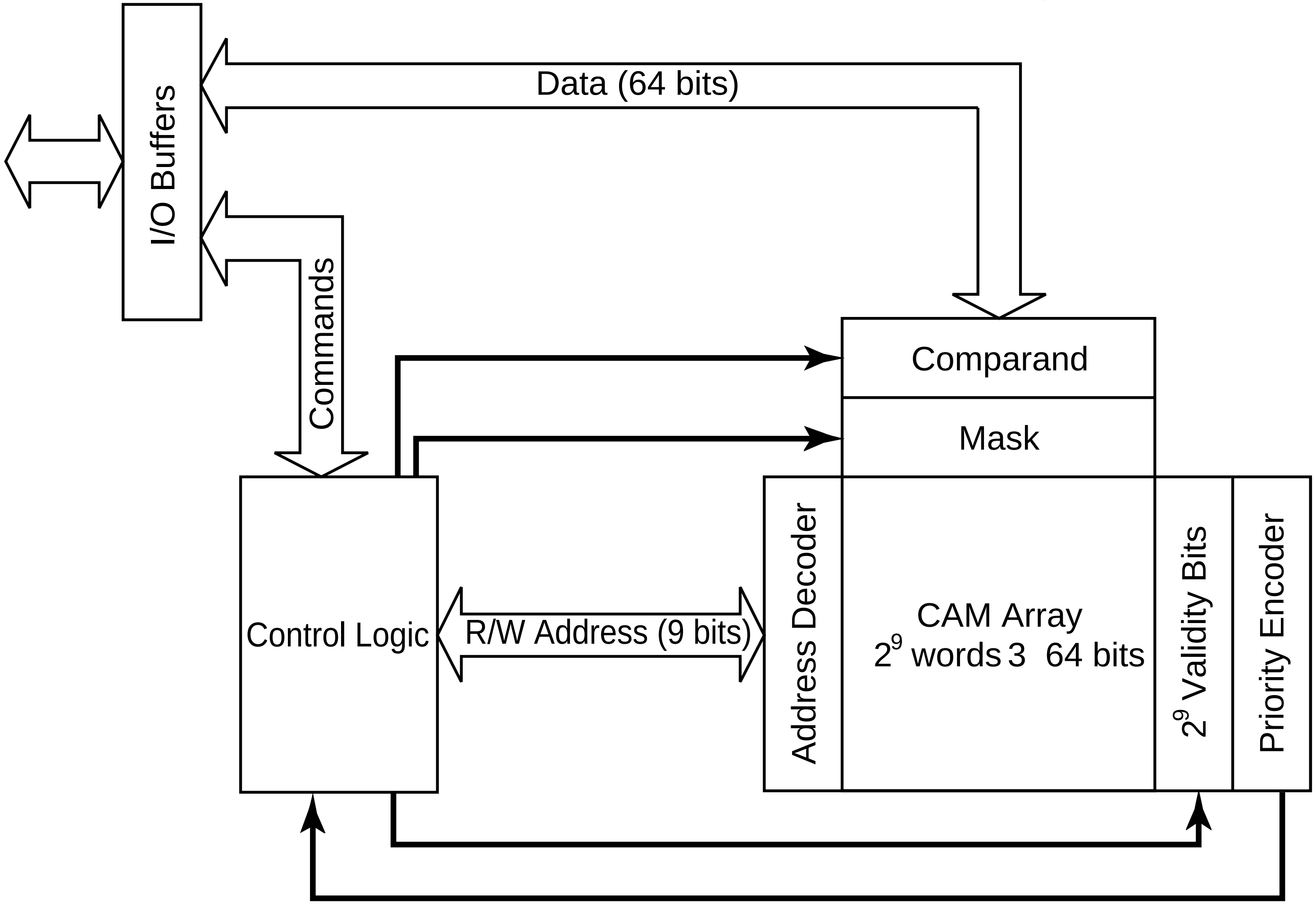
1. Shorter wires within blocks
2. Block address activates only 1 block => power savings

Block Diagram of 4 Mbit SRAM

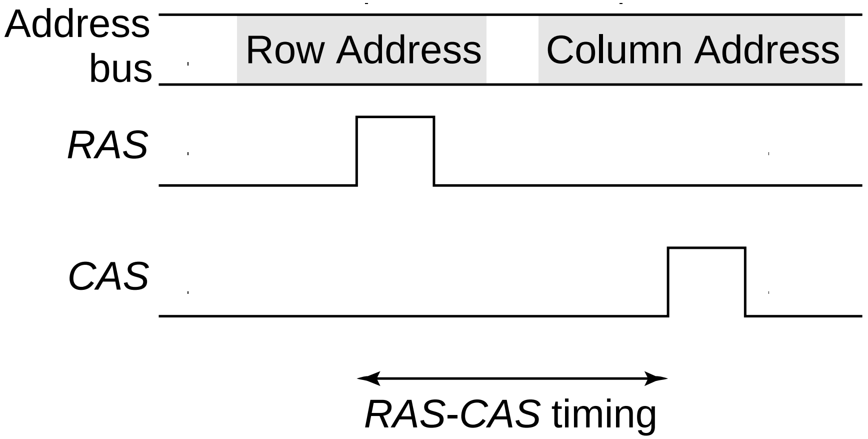


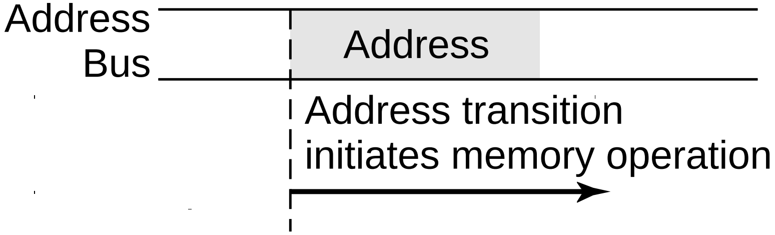
[Hirose90]

Contents-Addressable Memory



Memory Timing: Approaches

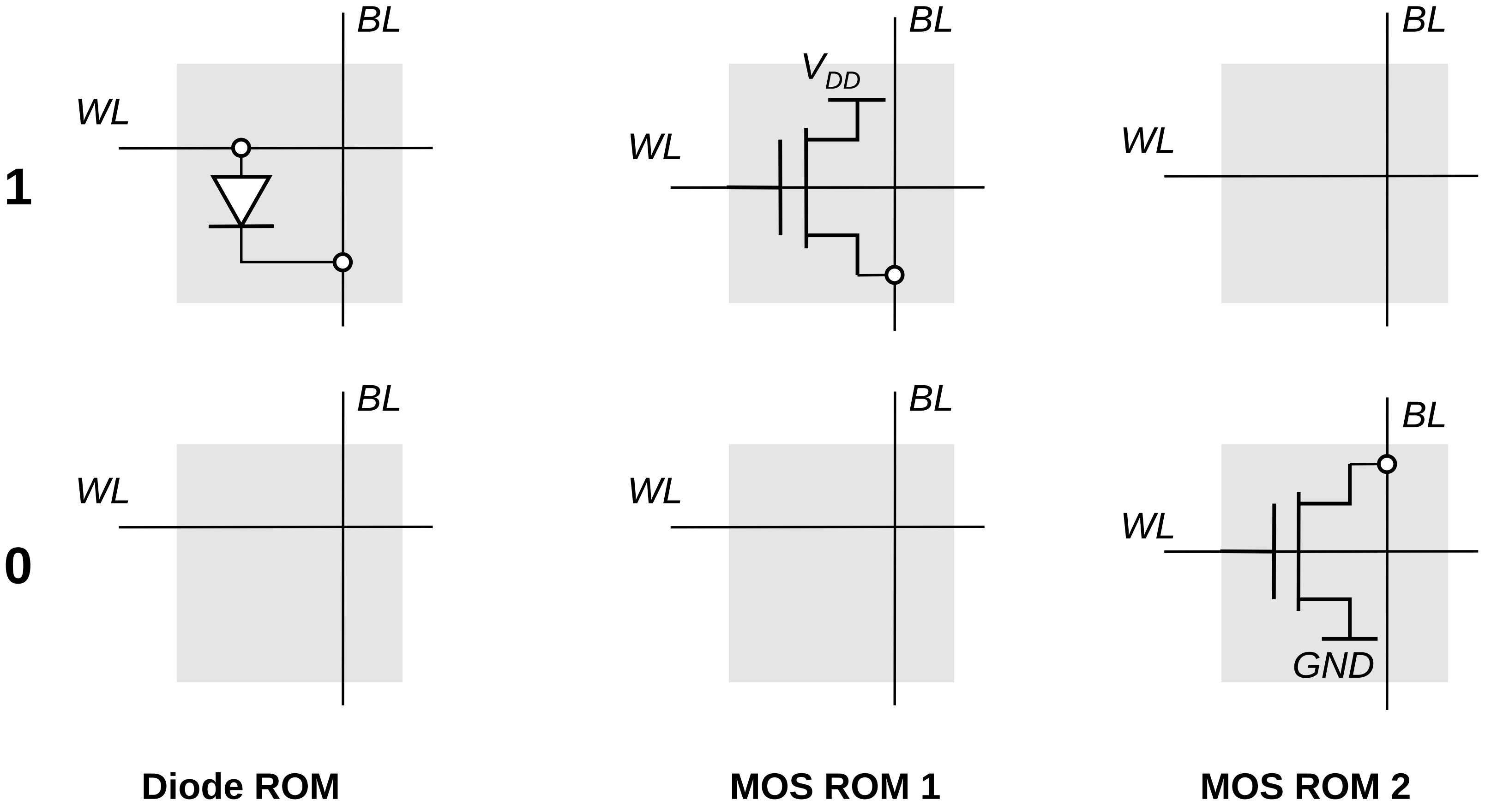




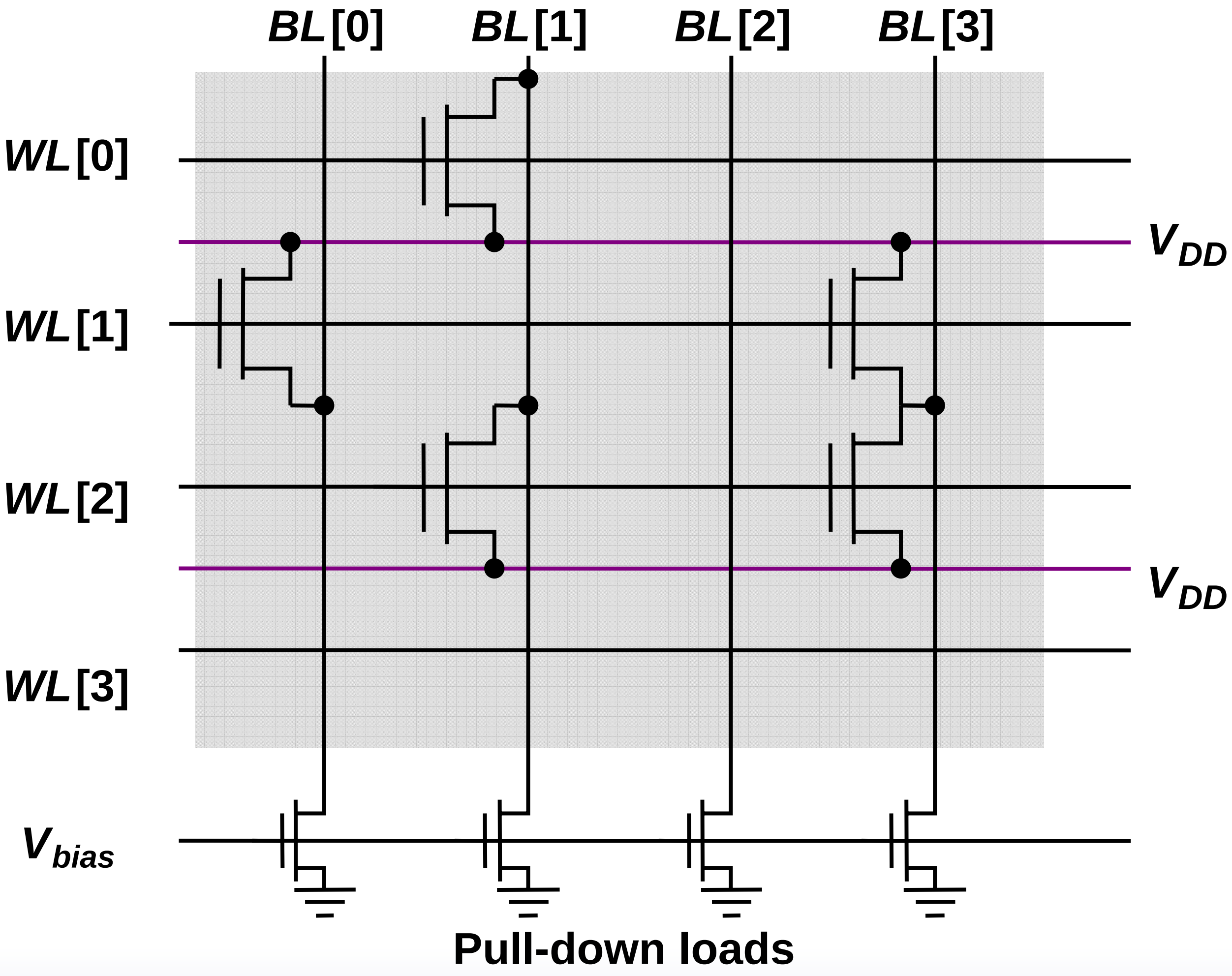
DRAM Timing
Multiplexed Adressing

SRAM Timing
Self-timed

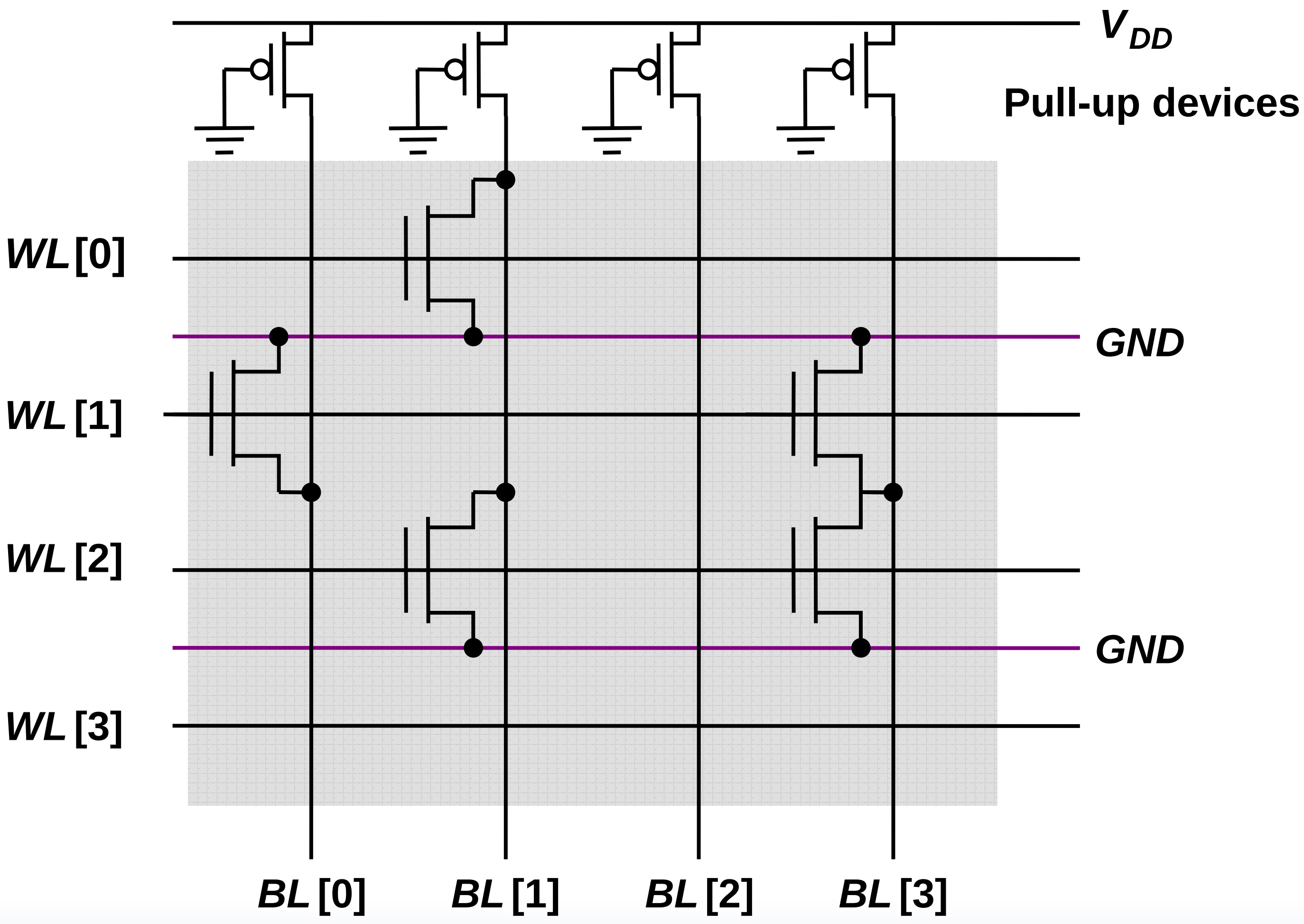
Read-Only Memory Cells



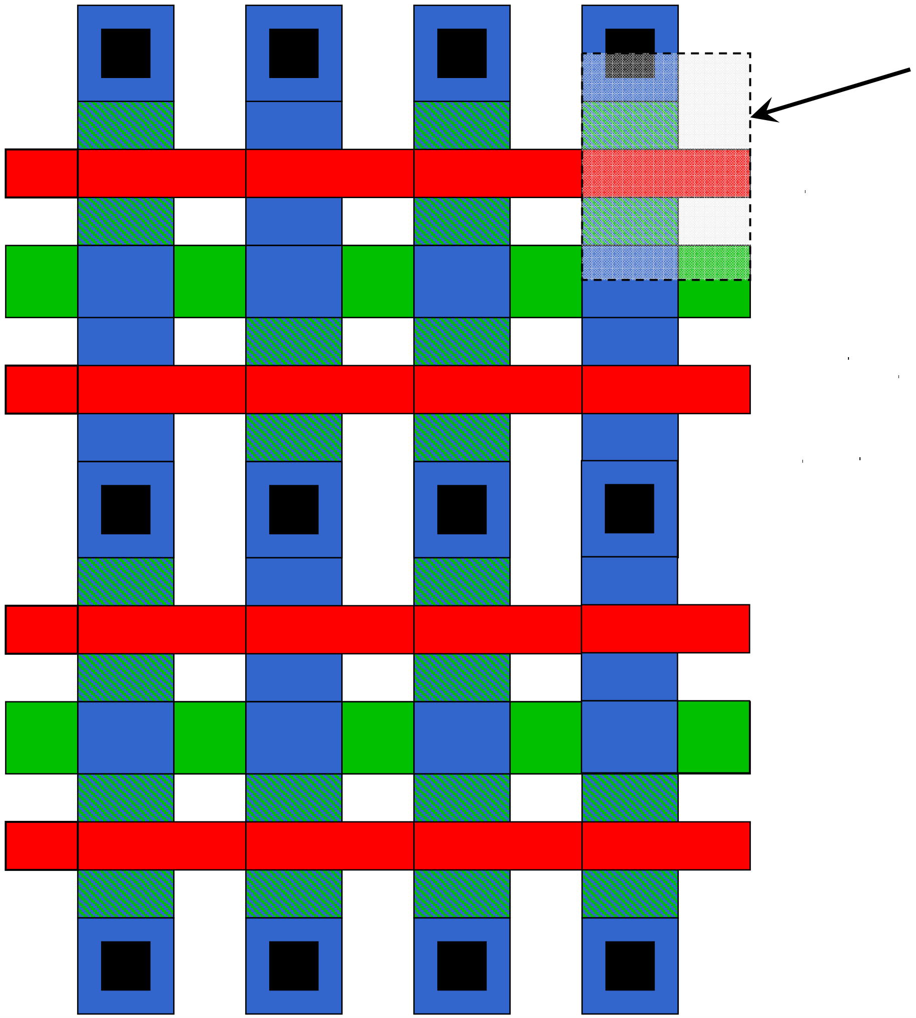
MOS OR ROM



MOS NOR ROM



MOS NOR ROM Layout

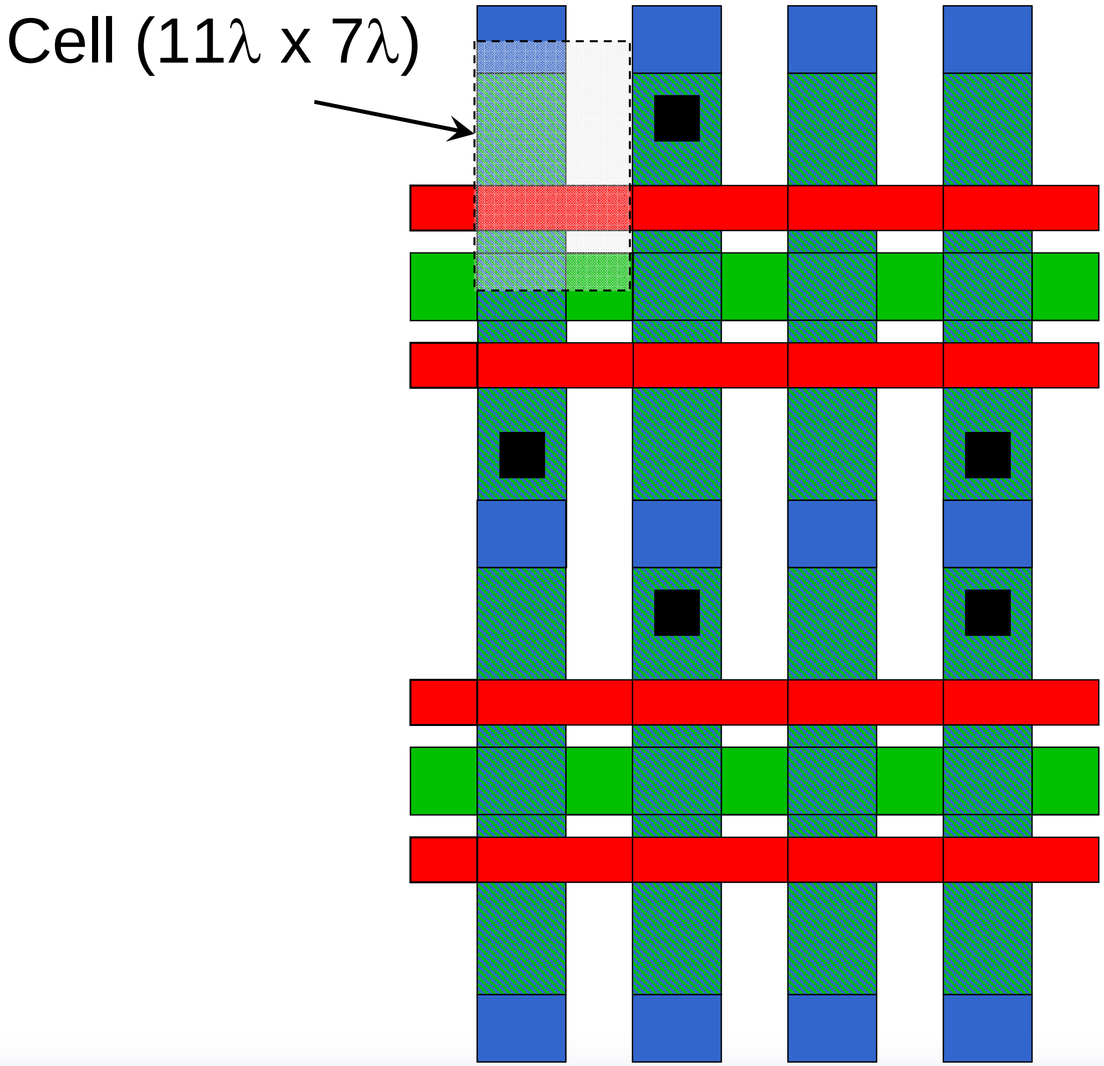


Cell (9.5λ x 7λ)

Programmming using the Active Layer Only

- 🟥 Polysilicon
- 🟦 Metal1
- 🟩 Diffusion
- ▨ Metal1 on Diffusion

MOS NOR ROM Layout



Cell (11λ x 7λ)

Programmming using the Contact Layer Only

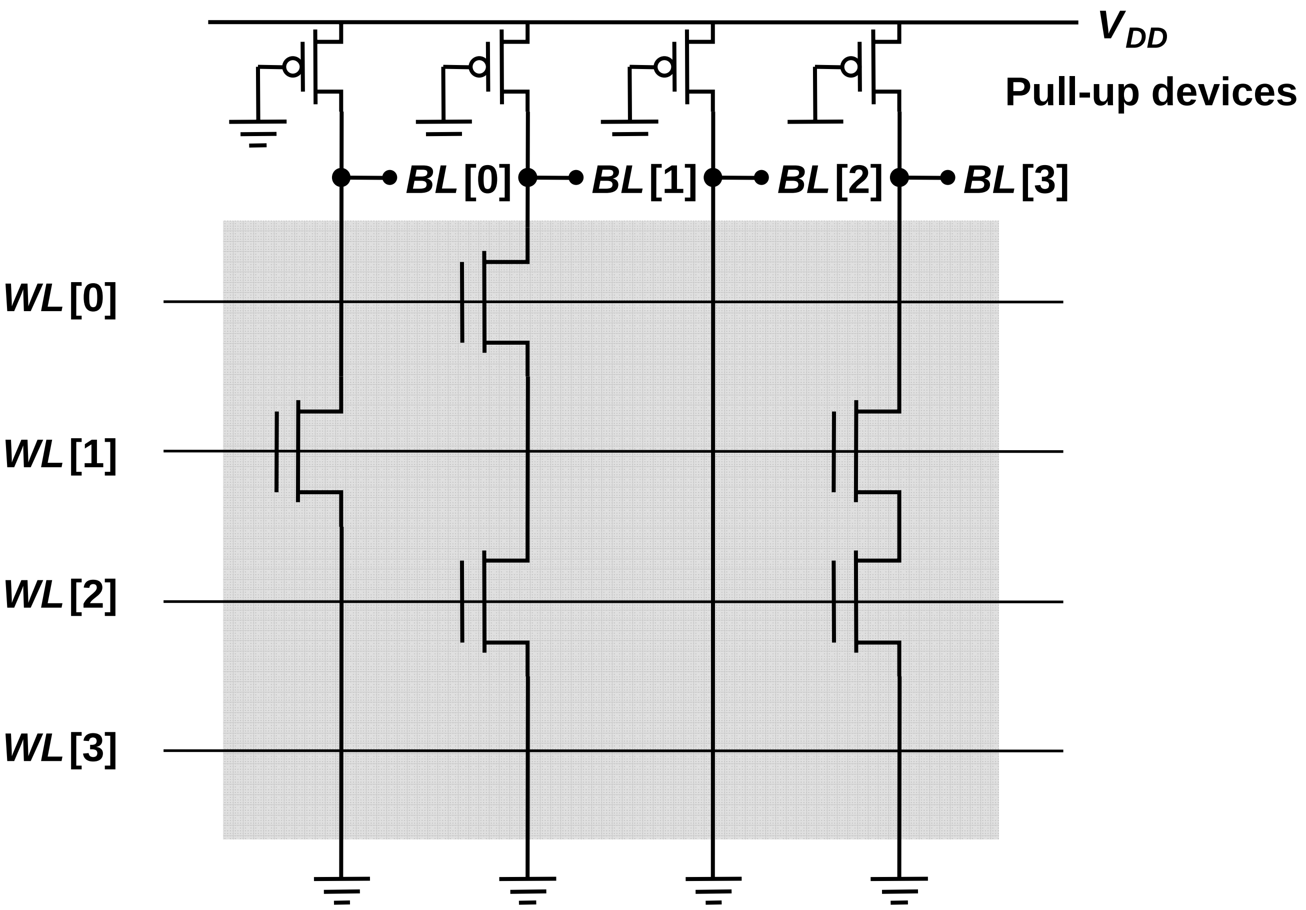
🟥 Polysilicon

🟦 Metal1

🟩 Diffusion

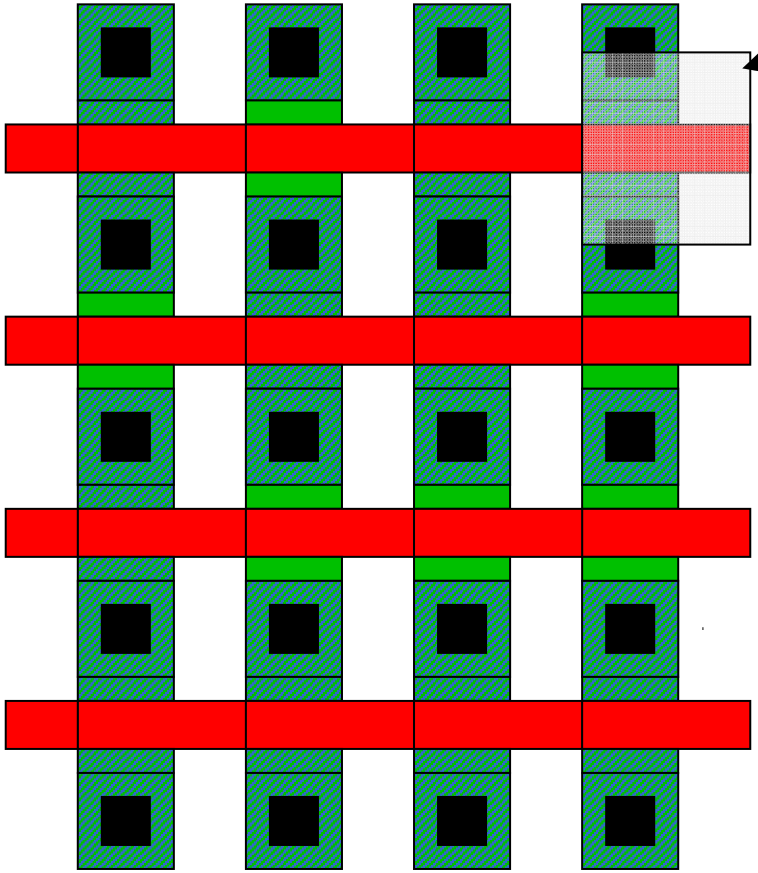
▨ Metal1 on Diffusion

MOS NAND ROM



All word lines high by default with exception of selected row

MOS NAND ROM Layout



Cell (8λ x 7λ)

Programmming using the Metal-1 Layer Only

No contact to VDD or GND necessary;
drastically reduced cell size
Loss in performance compared to NOR ROM

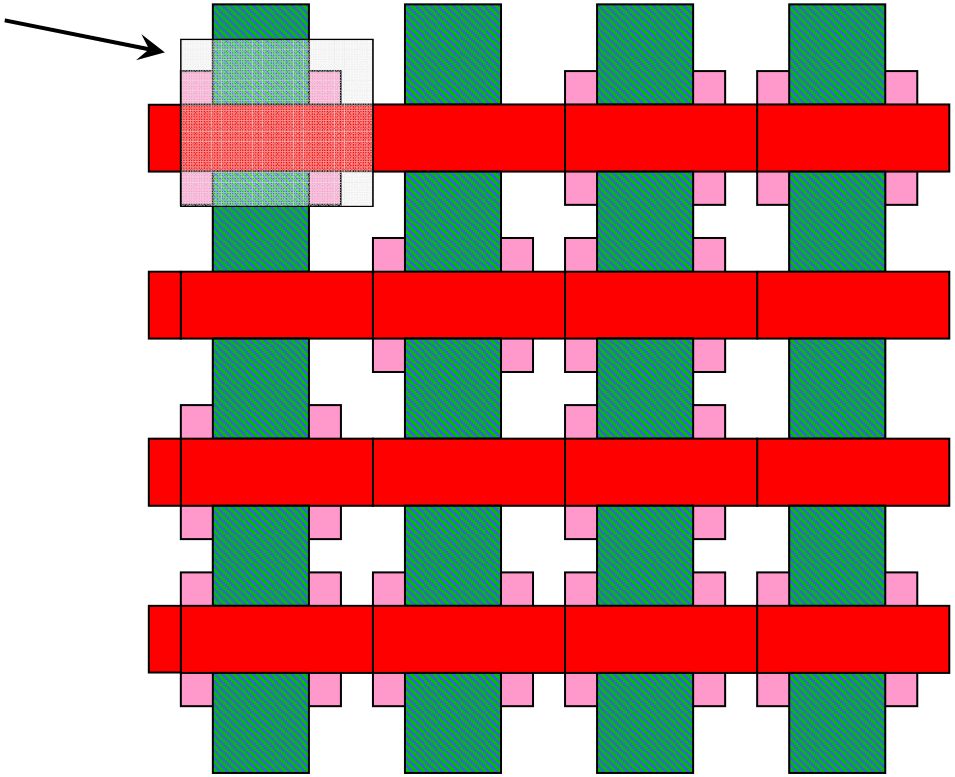
🟥 Polysilicon

🟩 Diffusion

▨ Metal1 on Diffusion

NAND ROM Layout

Cell (5λ x 6λ)



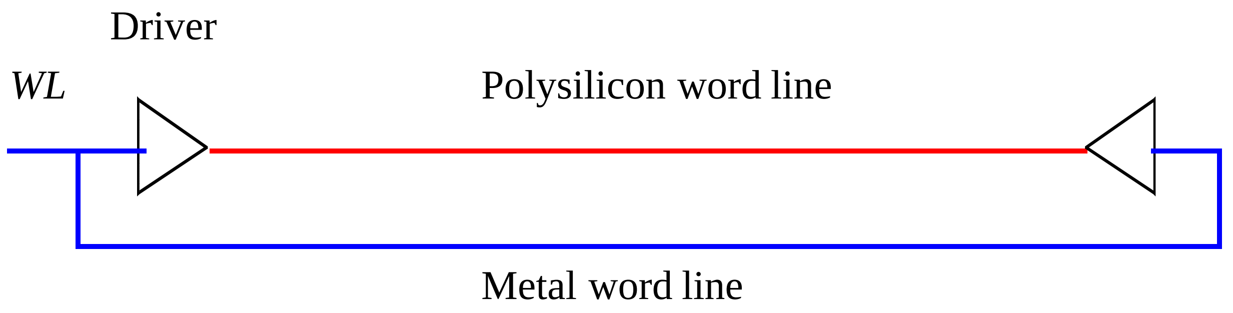
Programmming using Implants Only

🟥 Polysilicon

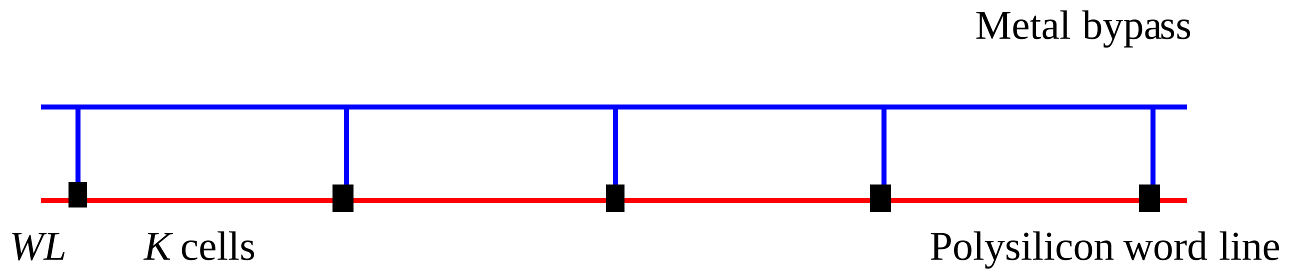
▭ Threshold-altering implant

▨ Metal1 on Diffusion

Decreasing Word Line Delay



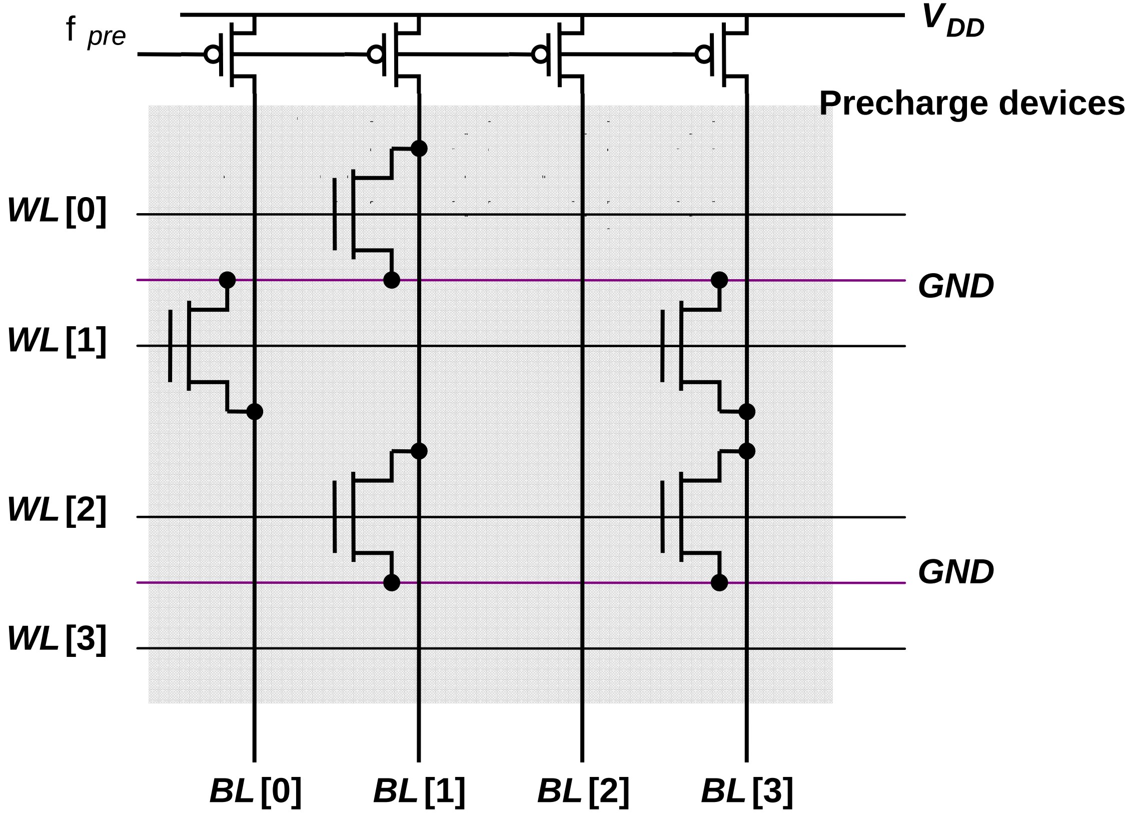
(a) Driving the word line from both sides



(b) Using a metal bypass

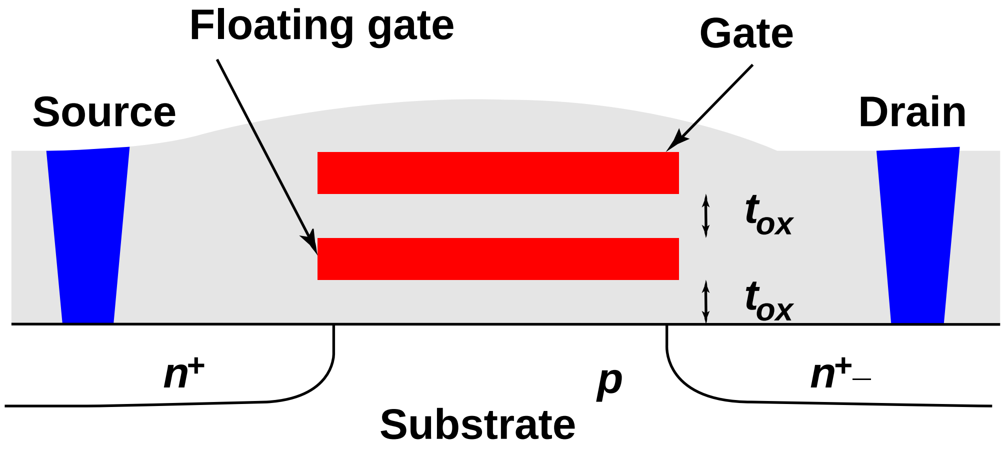
(c) Use silicides

Precharged MOS NOR ROM

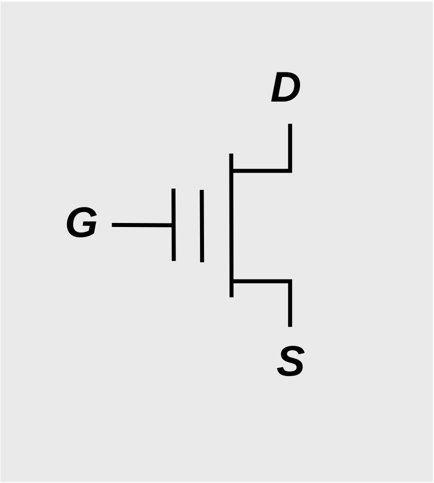


PMOS precharge device can be made as large as necessary, but clock driver becomes harder to design.

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

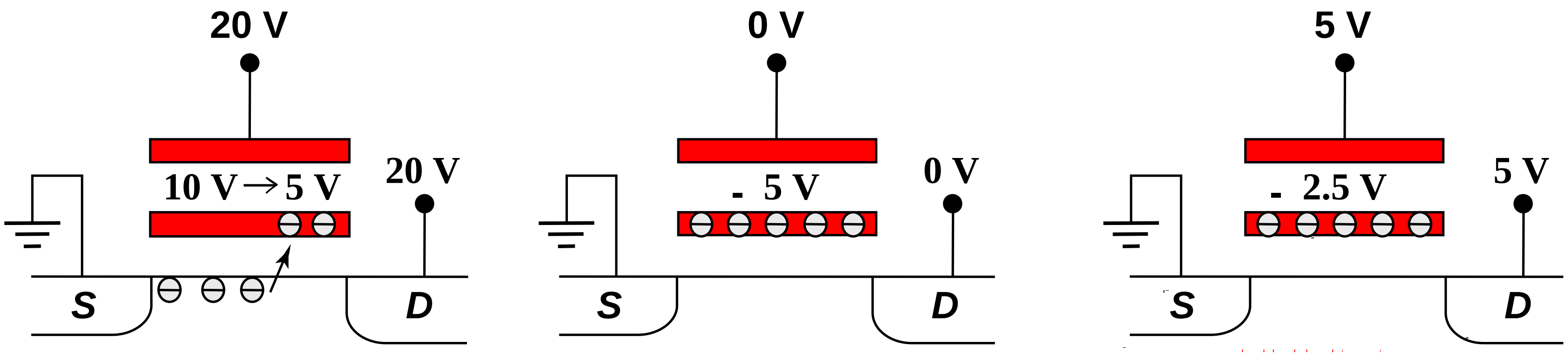


Device cross-section



Schematic symbol

Floating-Gate Transistor Programming

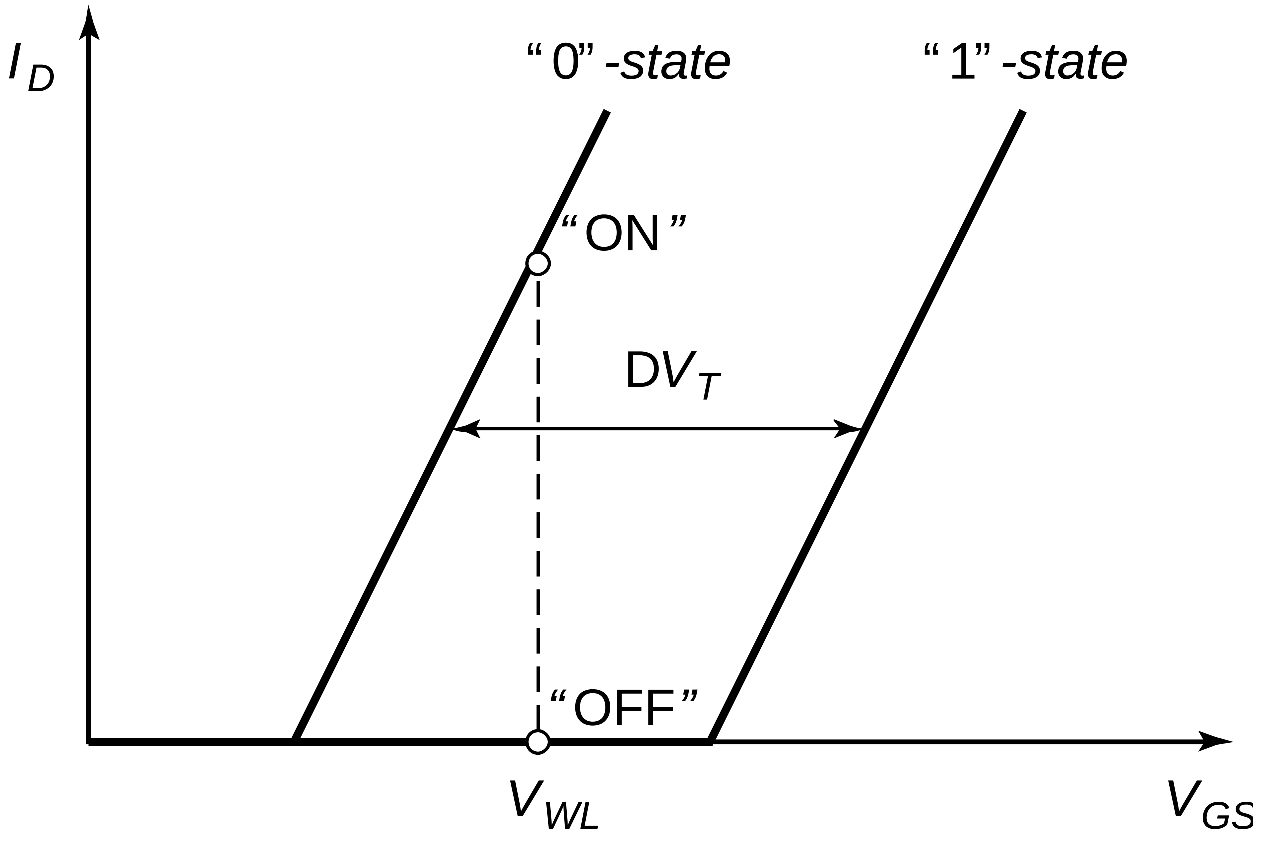


Avalanche injection

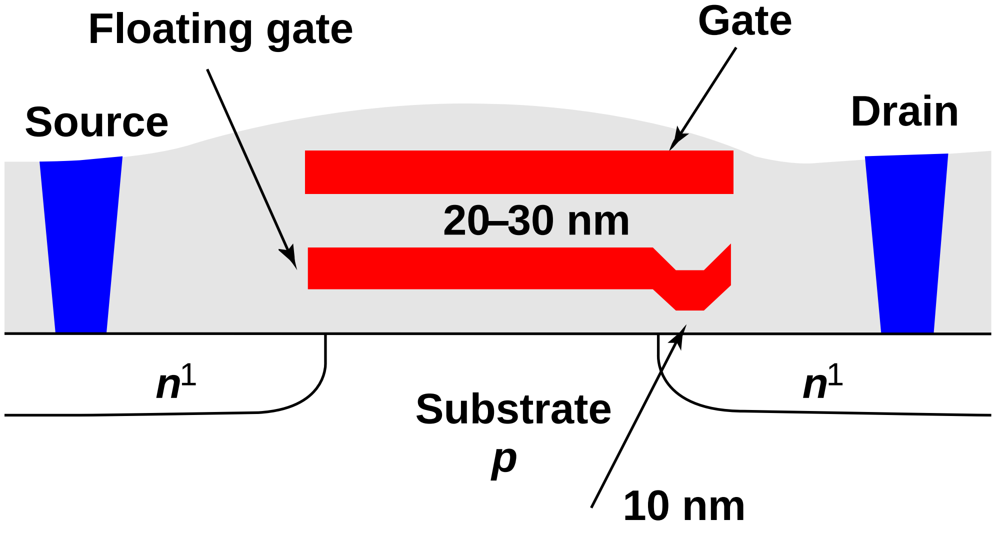
Removing programming voltage leaves charge trapped

Programming results in higher V_T.

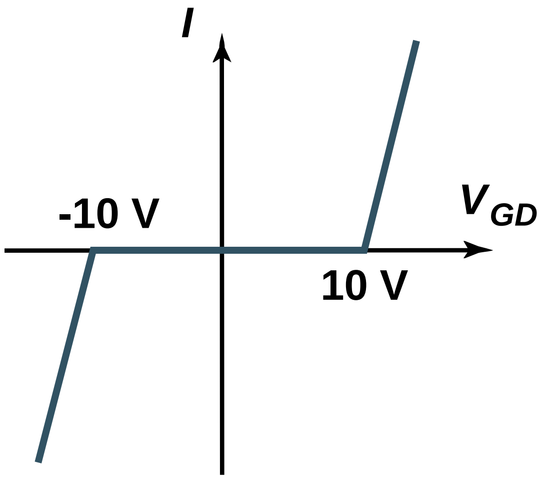
A "Programmable-Threshold" Transistor



FLOTOX EEPROM

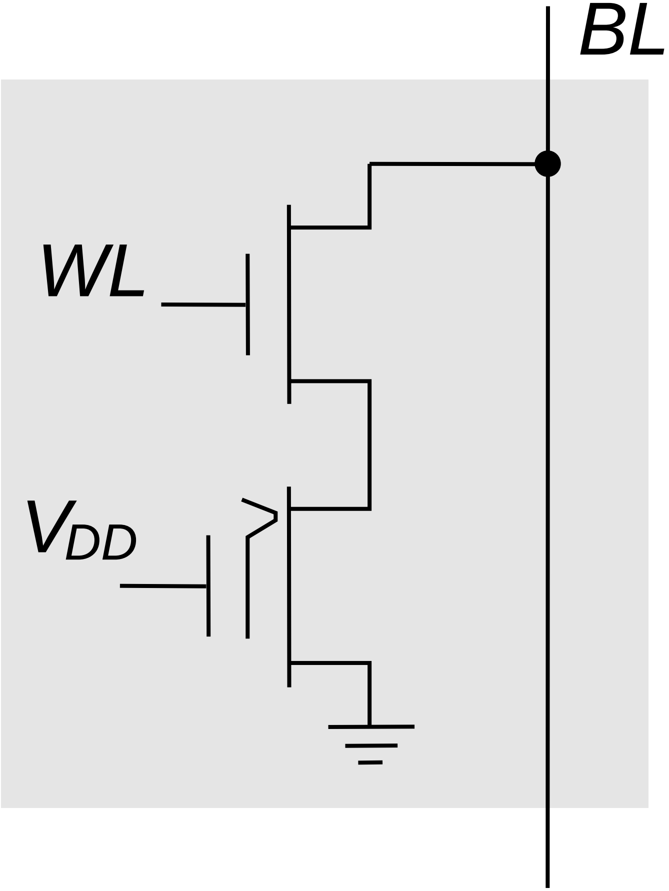


FLOTOX transistor



Fowler-Nordheim
I-V characteristic

EEPROM Cell

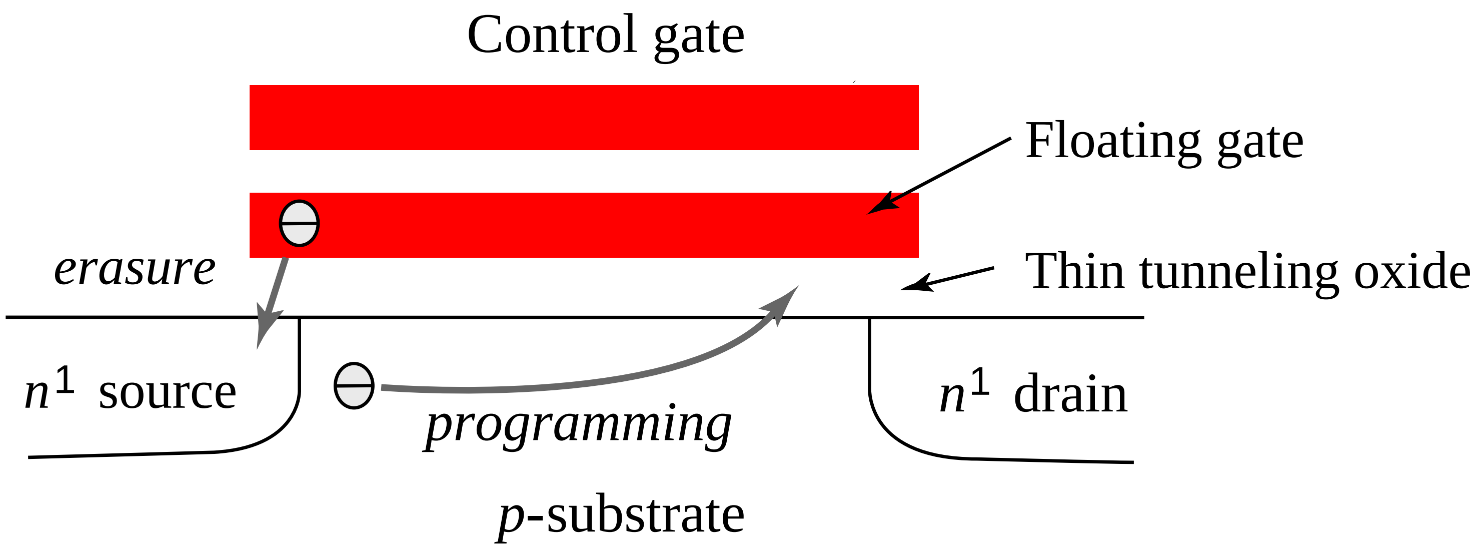


Absolute threshold control is hard

Unprogrammed transistor might be depletion

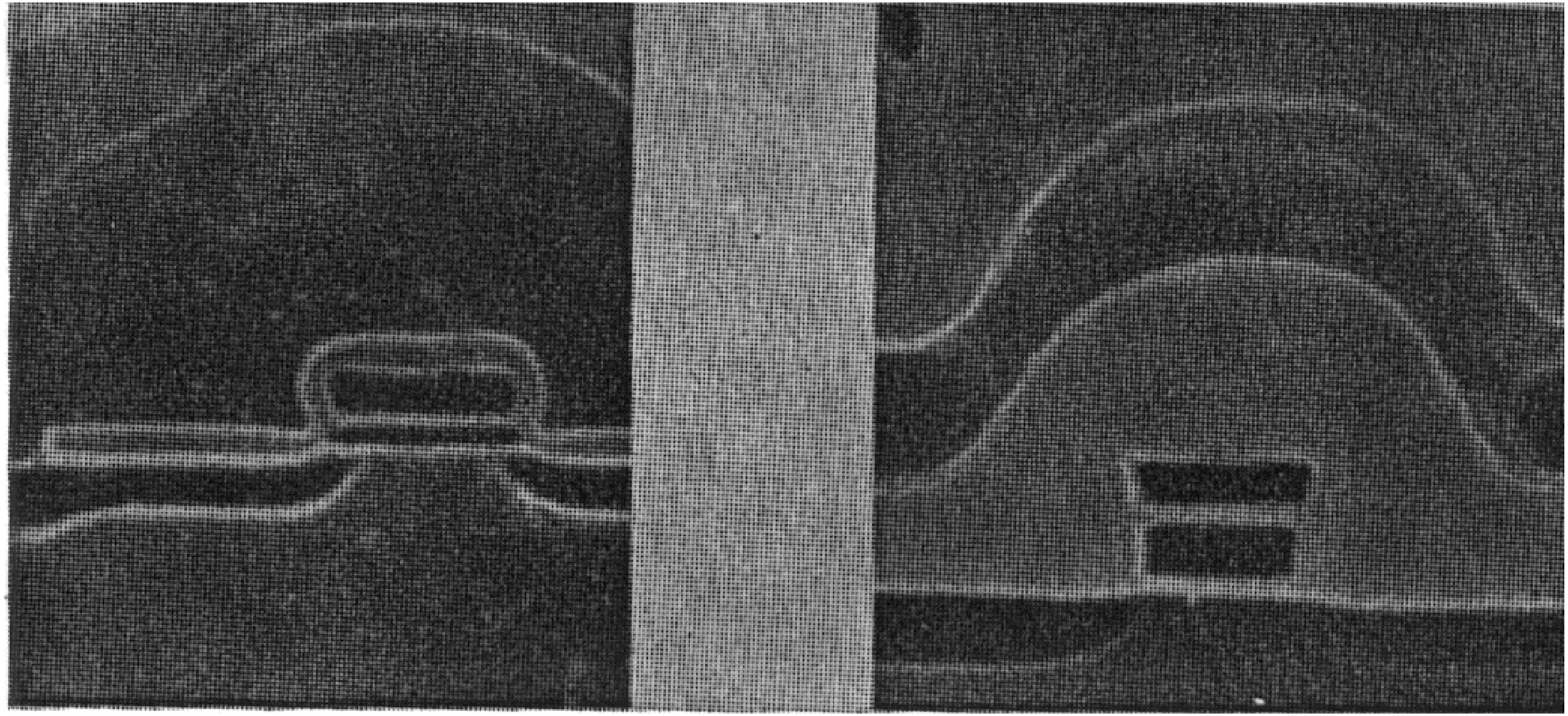
⇨ 2 transistor cell

Flash EEPROM



Many other options …

Cross-sections of NVM cells



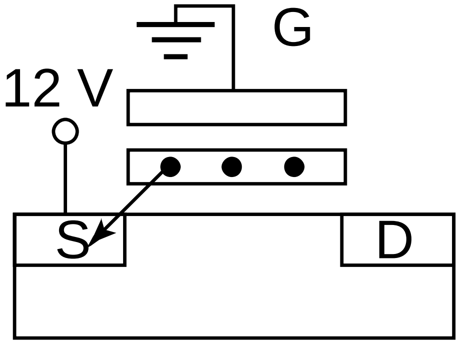
Flash

EPROM

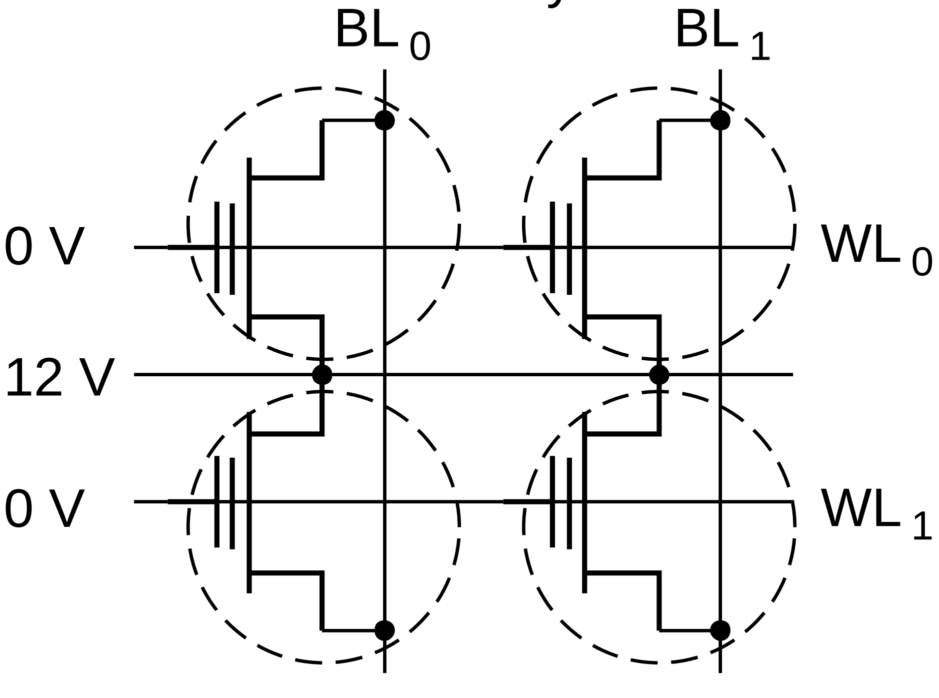
Courtesy Intel

Basic Operations in a NOR Flash Memory—Erase

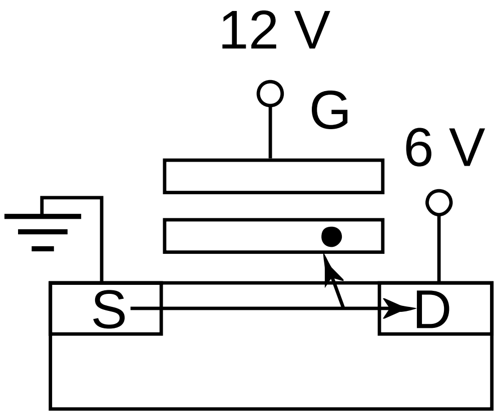
cell

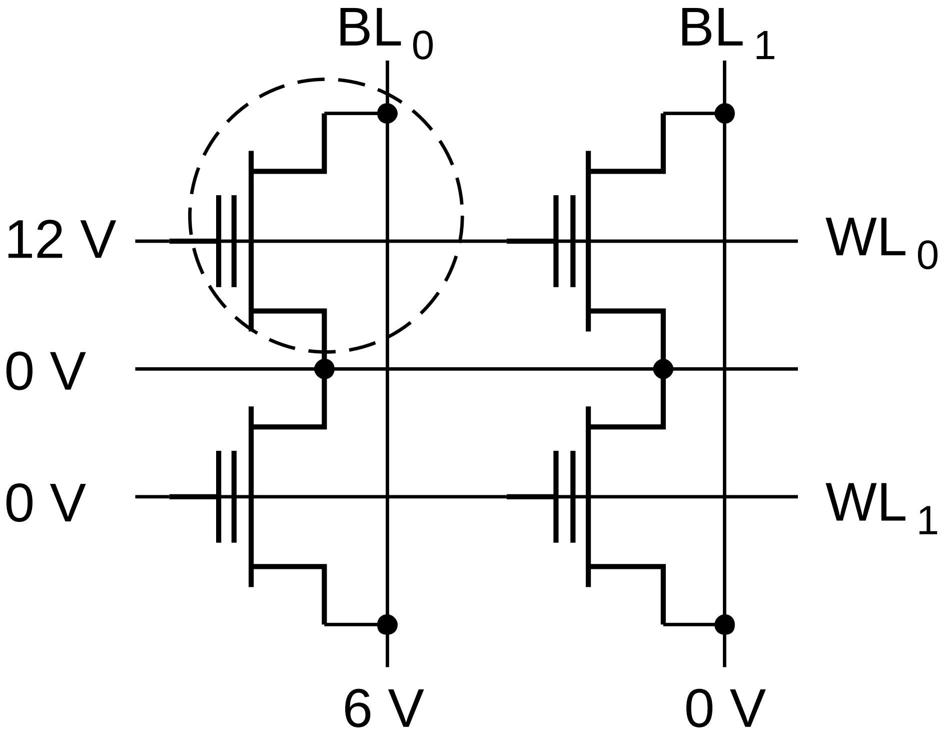


array

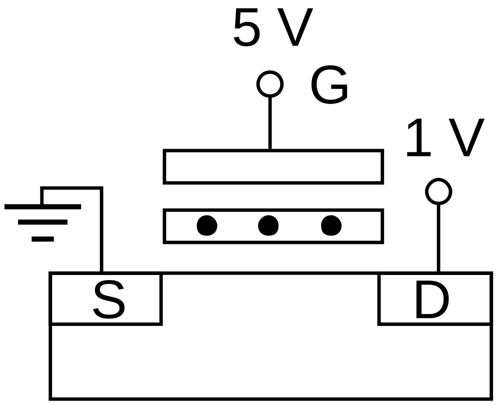


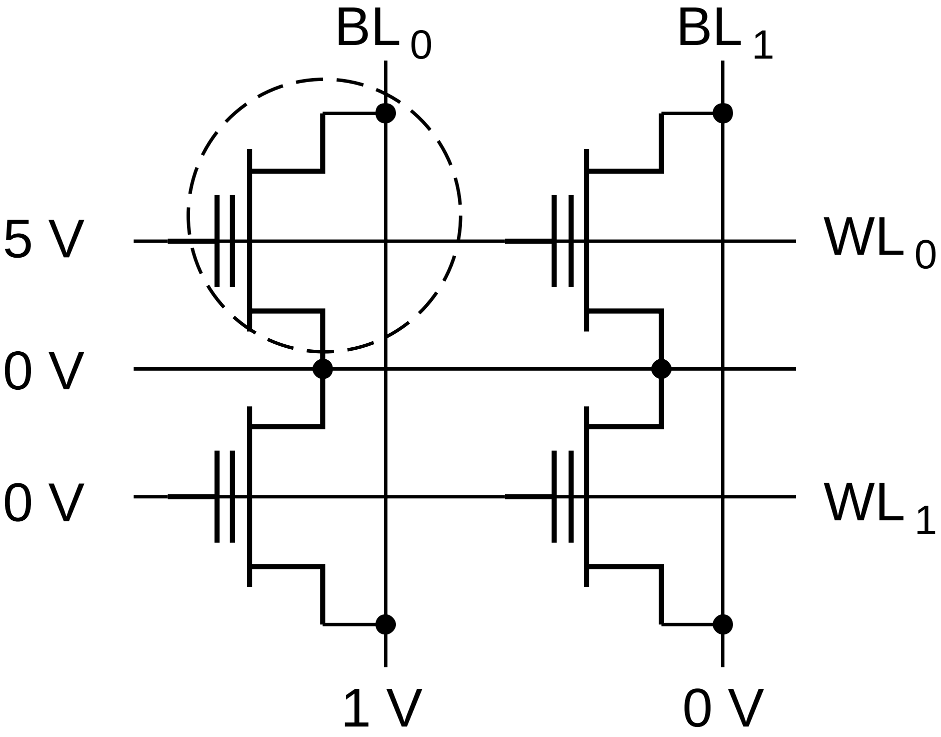
Basic Operations in a NOR Flash Memory—Write



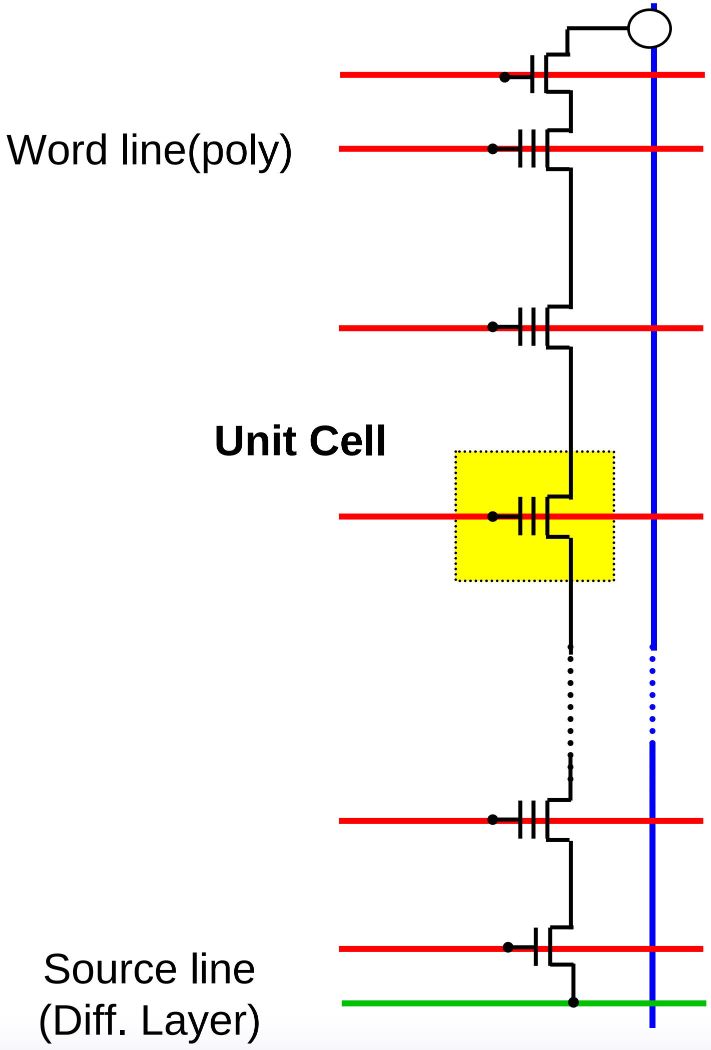


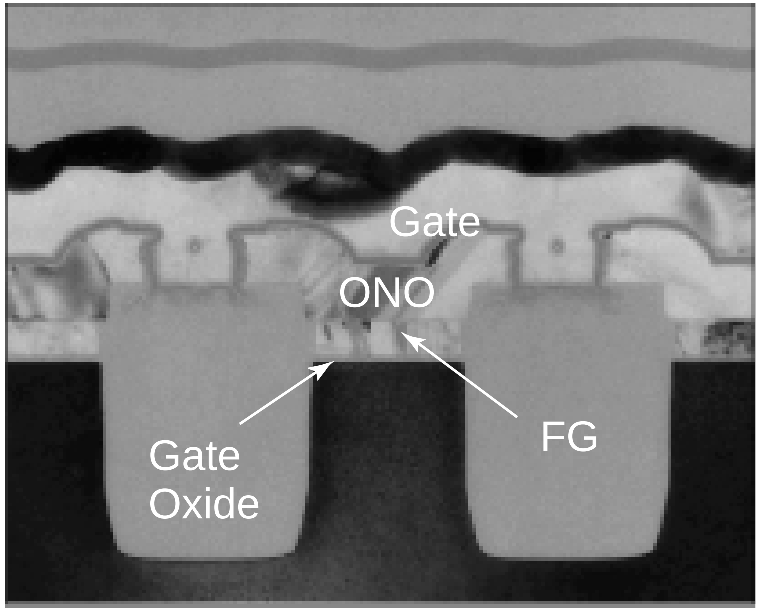
Basic Operations in a NOR Flash Memory—Read





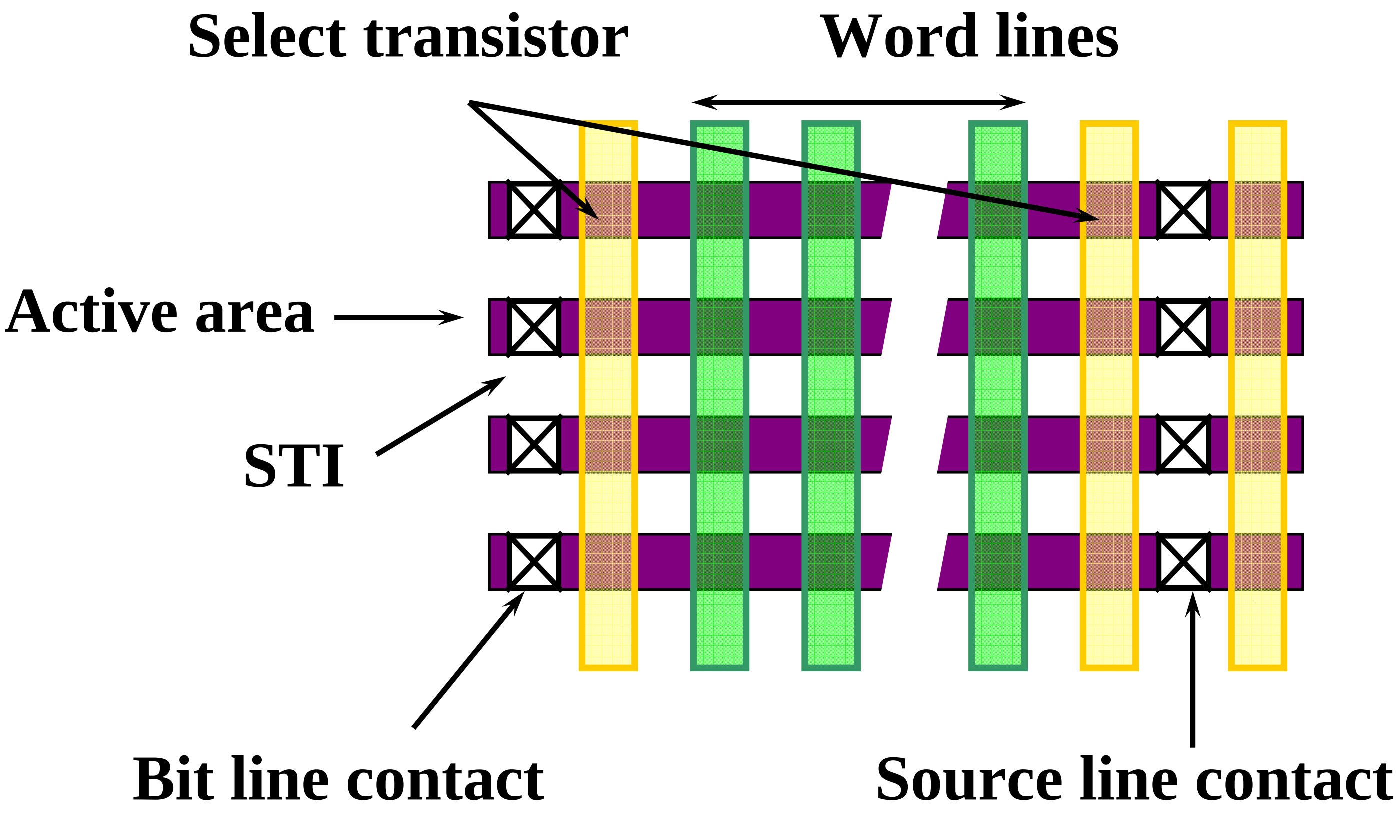
NAND Flash Memory





Courtesy Toshiba

NAND Flash Memory



Courtesy Toshiba

Characteristics of State-of-the-art NVM

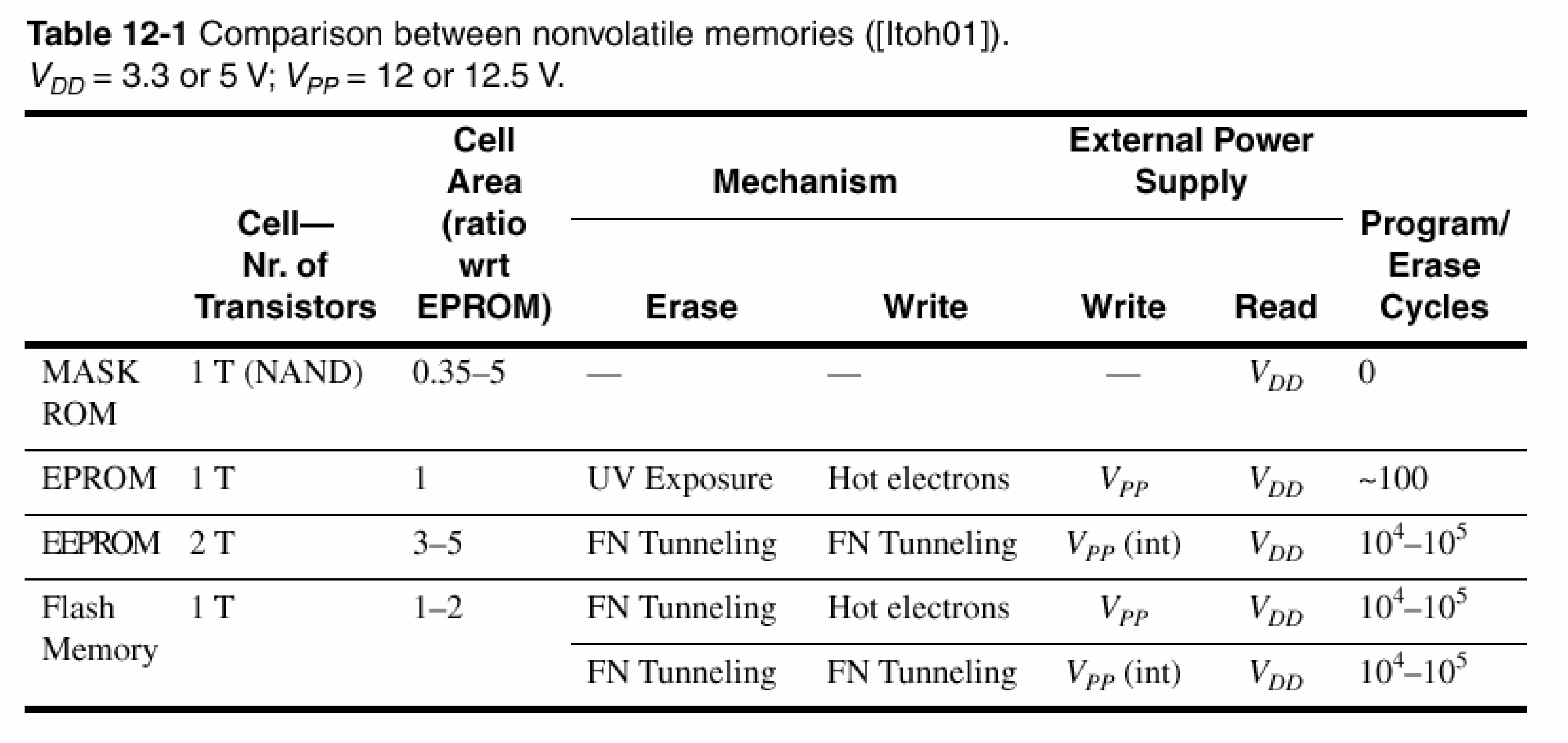


Table 12-1 Comparison between nonvolatile memories ([Itoh01]).
V_{DD} = 3.3 or 5 V; V_{PP} = 12 or 12.5 V.

	Cell— Nr. of Transistors	Cell Area (ratio wrt EPROM)	Mechanism		External Power Supply		Program/ Erase Cycles
			Erase	Write	Write	Read	
MASK ROM	1 T (NAND)	0.35–5	—	—	—	V_{DD}	0
EPROM	1 T	1	UV Exposure	Hot electrons	V_{PP}	V_{DD}	~100
EEPROM	2 T	3–5	FN Tunneling	FN Tunneling	V_{PP} (int)	V_{DD}	10^4–10^5
Flash Memory	1 T	1–2	FN Tunneling	Hot electrons	V_{PP}	V_{DD}	10^4–10^5
			FN Tunneling	FN Tunneling	V_{PP} (int)	V_{DD}	10^4–10^5

Read-Write Memories (RAM)

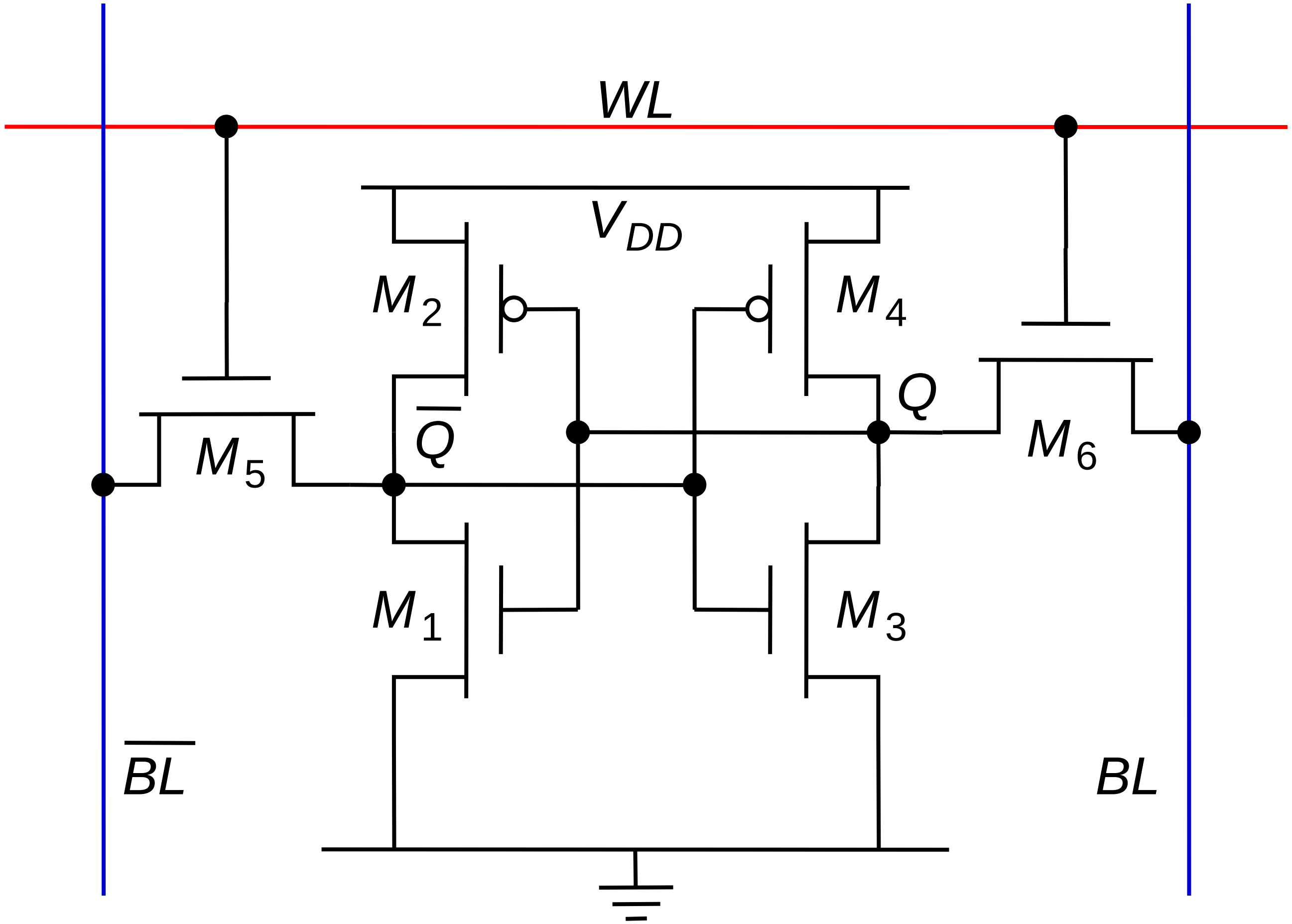
- ❑ STATIC (SRAM)

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

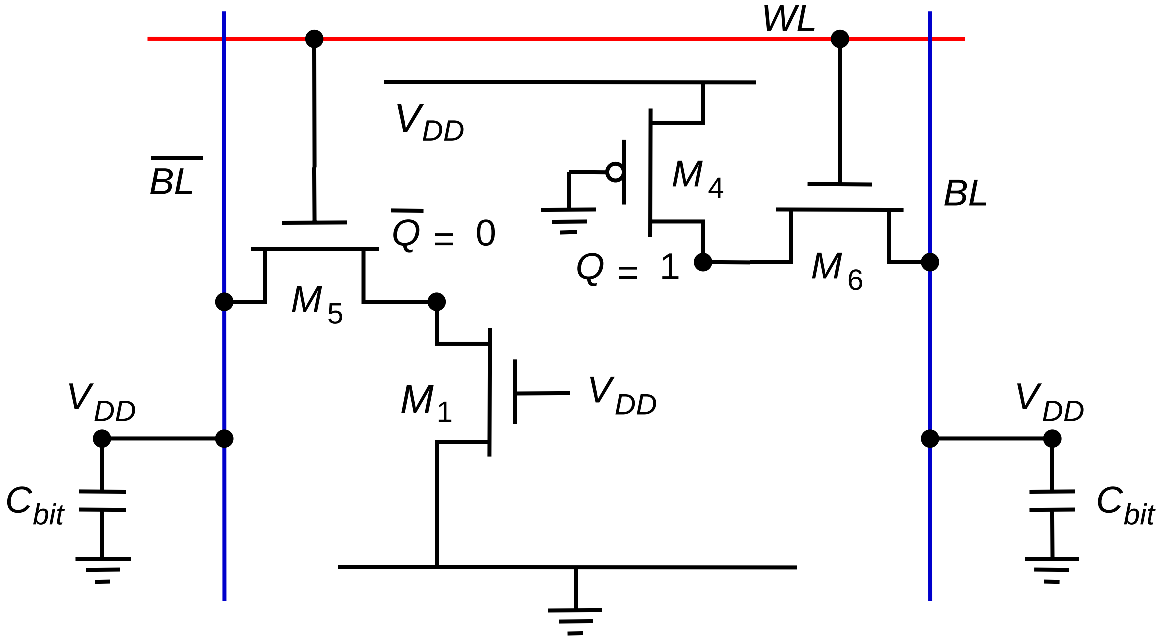
- ❑ DYNAMIC (DRAM)

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

6-transistor CMOS SRAM Cell



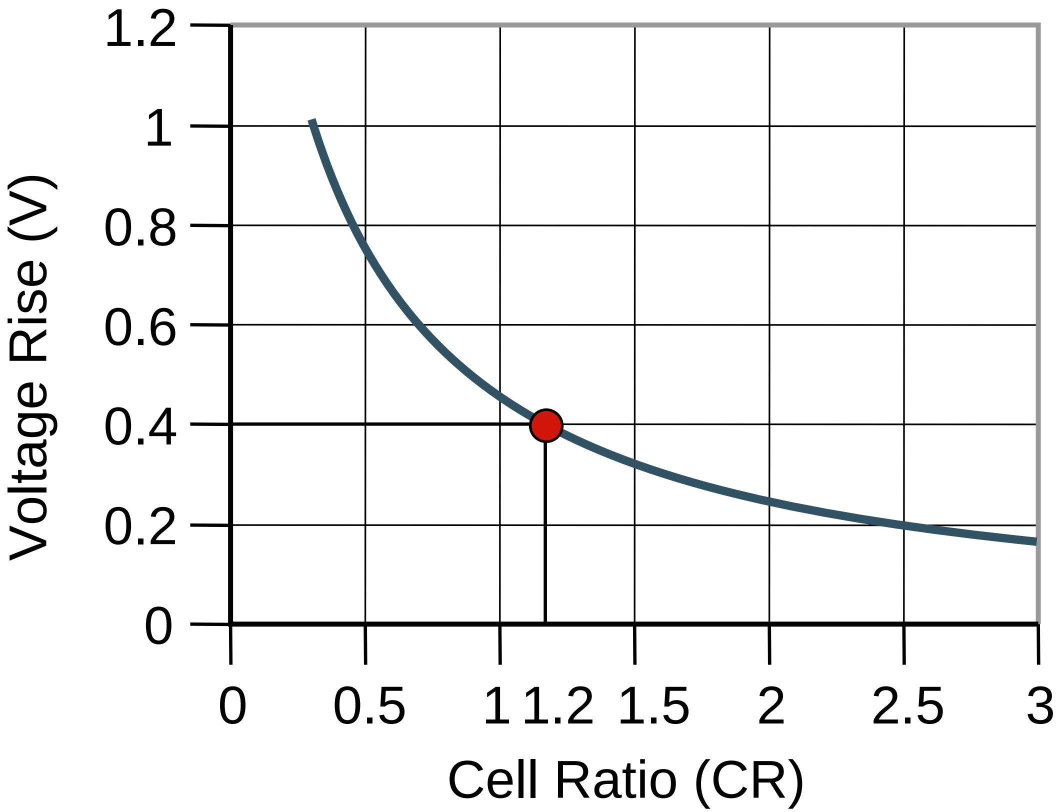
CMOS SRAM Analysis (Read)



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

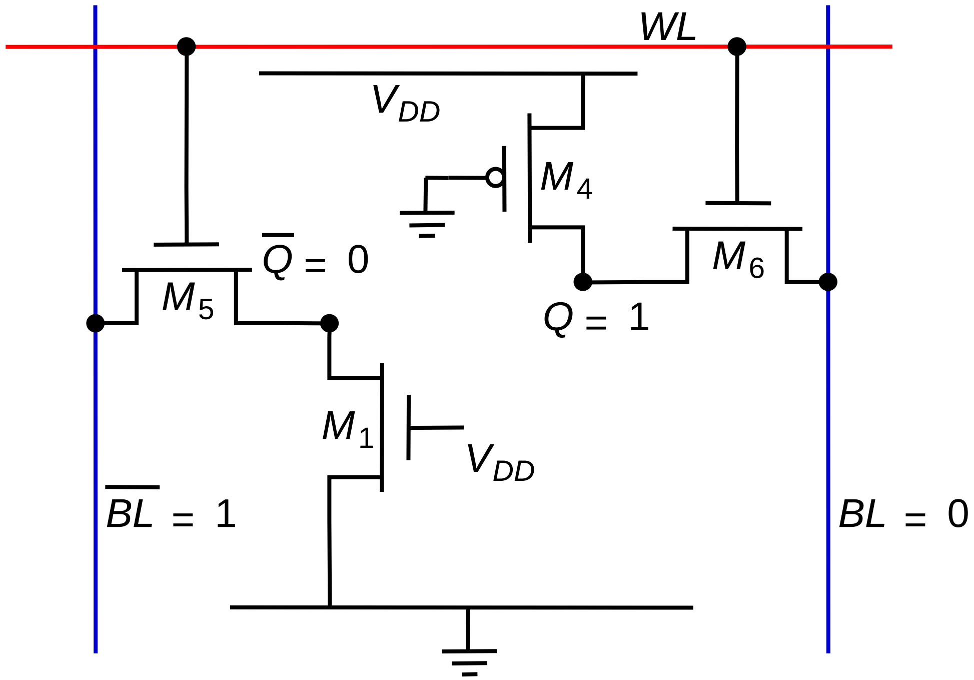
$$\Delta V = \frac{V_{DSATn}+CR(V_{DD}-V_{Tn})-\sqrt{V_{DSATn}^2(1+CR)+CR^2(V_{DD}-V_{Tn})^2}}{CR}$$

CMOS SRAM Analysis (Read)



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

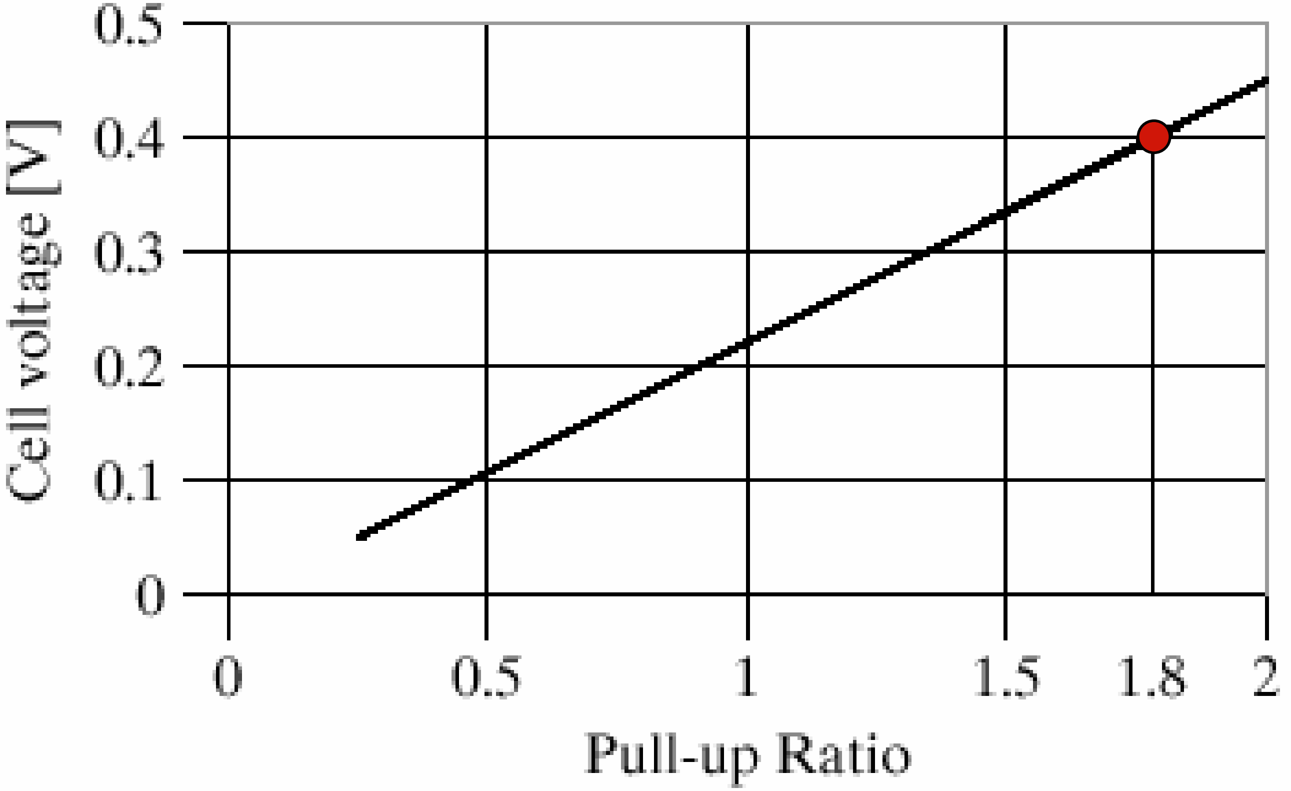
CMOS SRAM Analysis (Write)



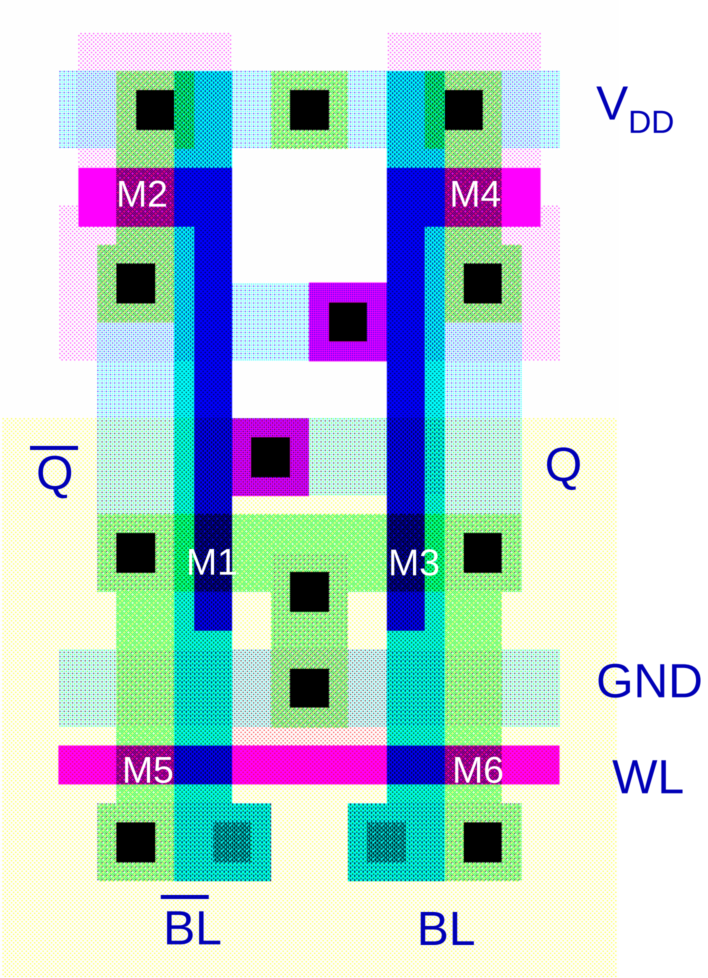
$$k_{n,M6}\left((V_{DD}-V_{Tn})V_Q-\frac{V_Q^2}{2}\right) = k_{p,M4}\left((V_{DD}-|V_{Tp}|)V_{DSATp}-\frac{V_{DSATp}^2}{2}\right)$$

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

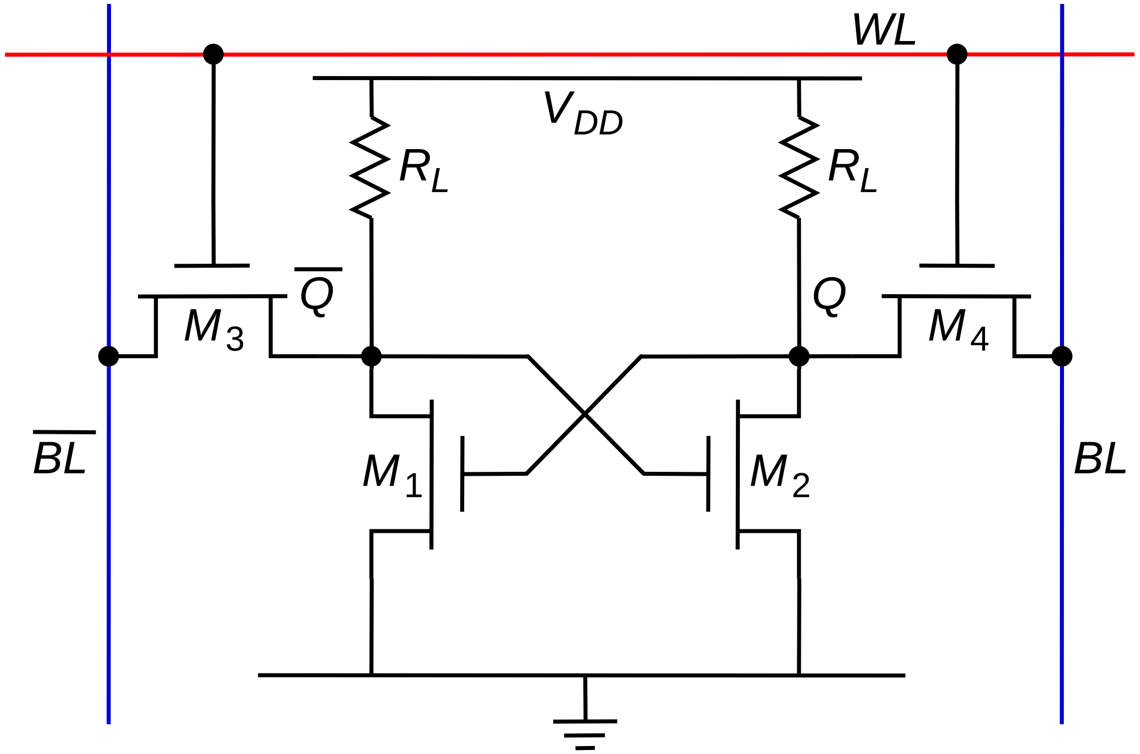
CMOS SRAM Analysis (Write)



6T-SRAM — Layout



Resistance-load SRAM Cell



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

SRAM Characteristics

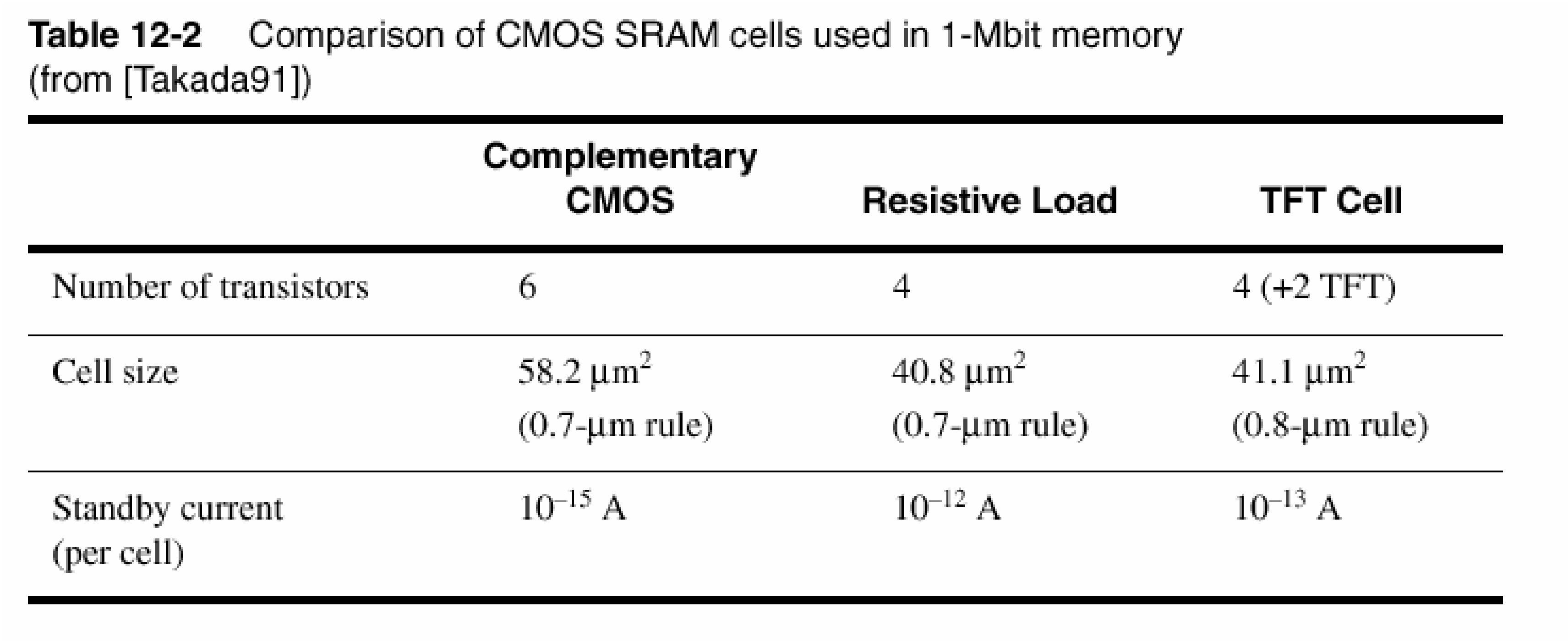
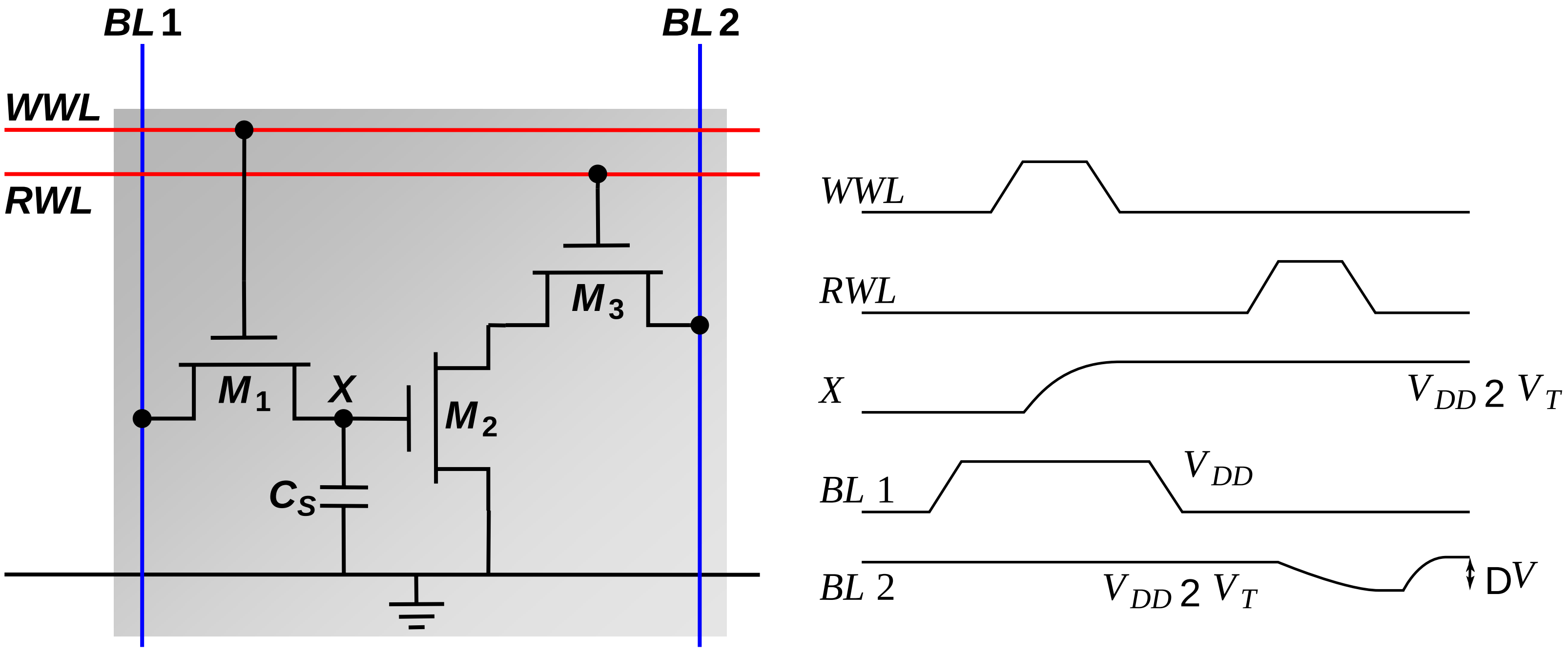


Table 12-2 Comparison of CMOS SRAM cells used in 1-Mbit memory (from [Takada91])

	Complementary CMOS	**Resistive Load**	**TFT Cell**
Number of transistors	6	4	4 (+2 TFT)
Cell size	58.2 μm^2 (0.7-µm rule)	40.8 μm^2 (0.7-µm rule)	41.1 μm^2 (0.8-µm rule)
Standby current (per cell)	10^{-15} A	10^{-12} A	10^{-13} A

3-Transistor DRAM Cell

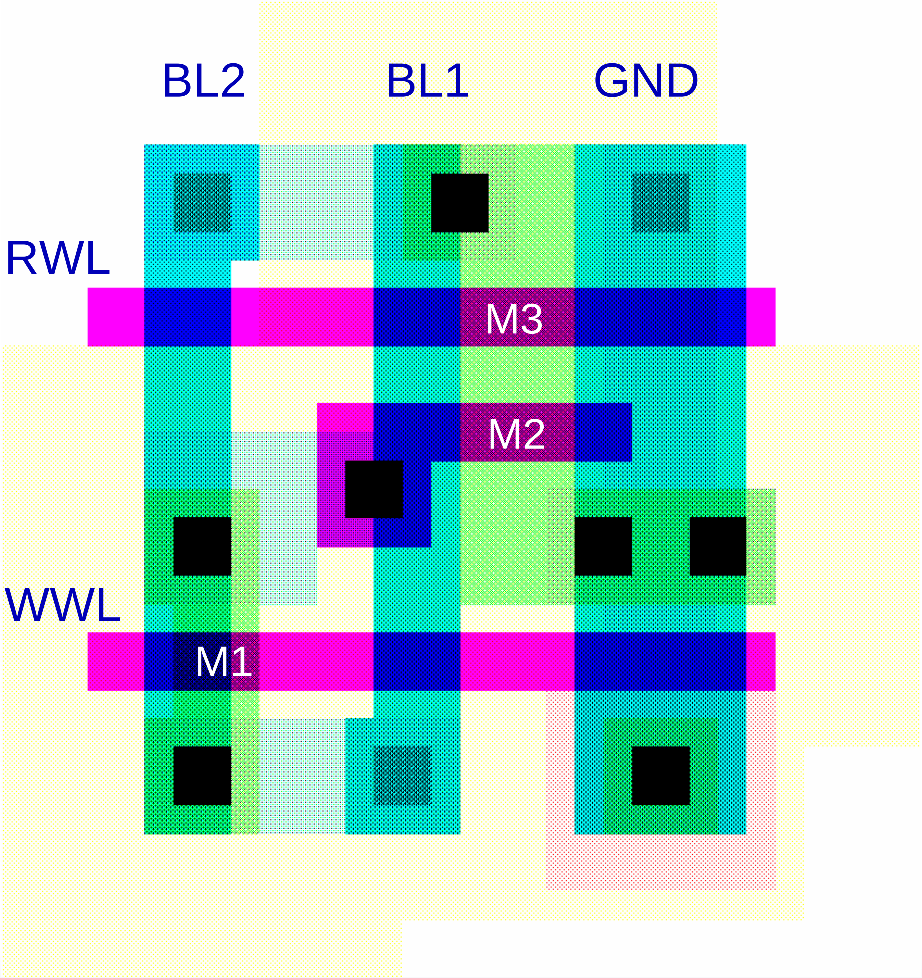


No constraints on device ratios

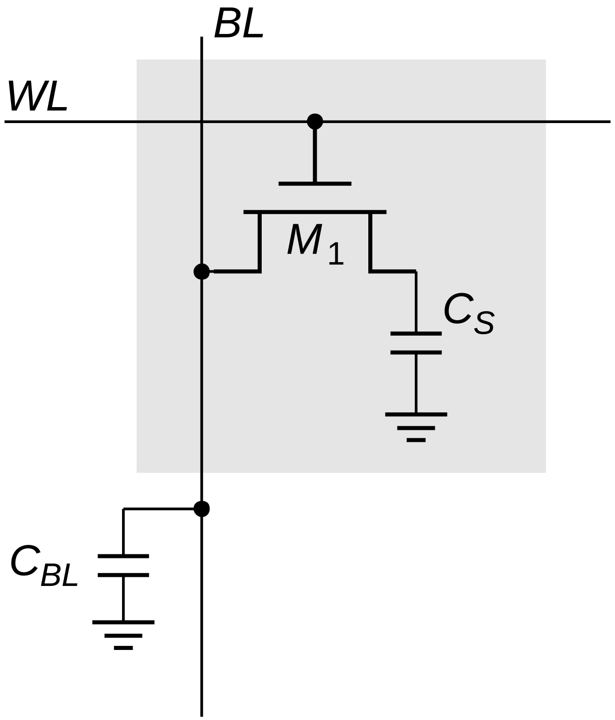
Reads are non-destructive

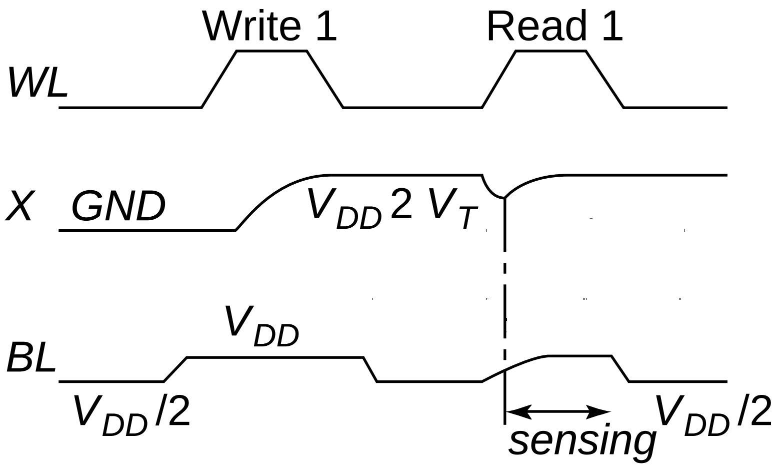
Value stored at node X when writing a “1” = V_{WWL}-V_{Tn}

3T-DRAM — Layout



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.

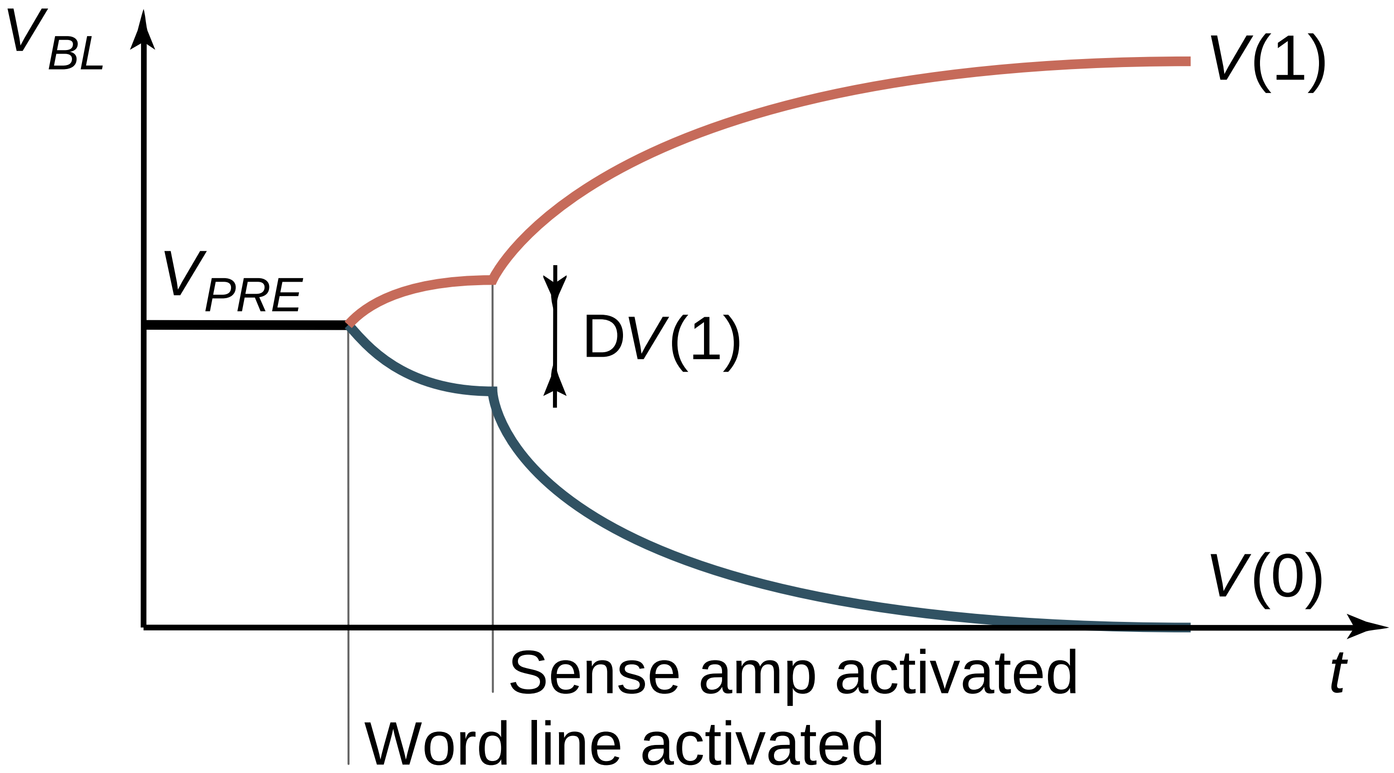
❑ DRAM memory cells are single ended in contrast to SRAM cells.

❑The read-out of the 1T DRAM cell is destructive; read and refresh operations are necessary for correct operation.

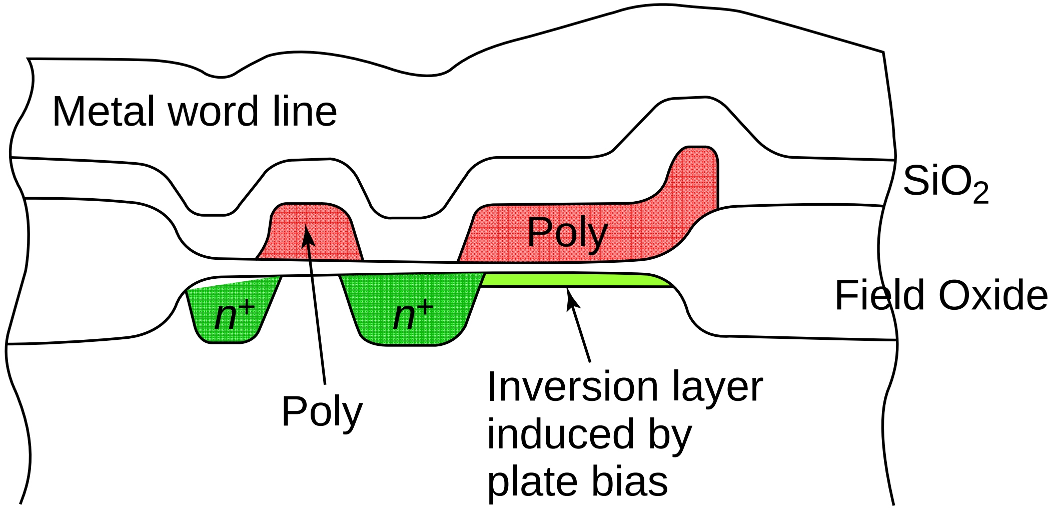
❑ Unlike 3T cell, 1T cell requires presence of an extra capacitance that must be explicitly included in the design.

❑ 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}

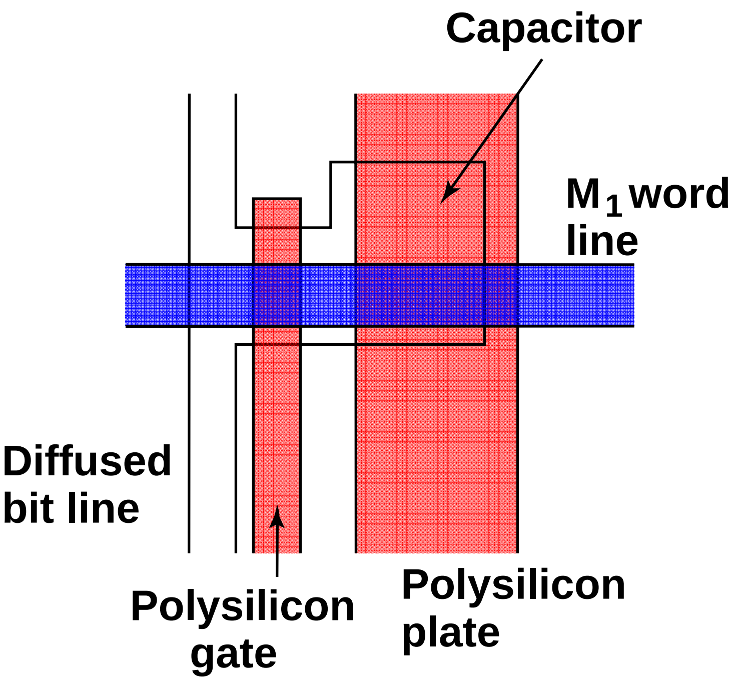
Sense Amp Operation



1-T DRAM Cell



Cross-section

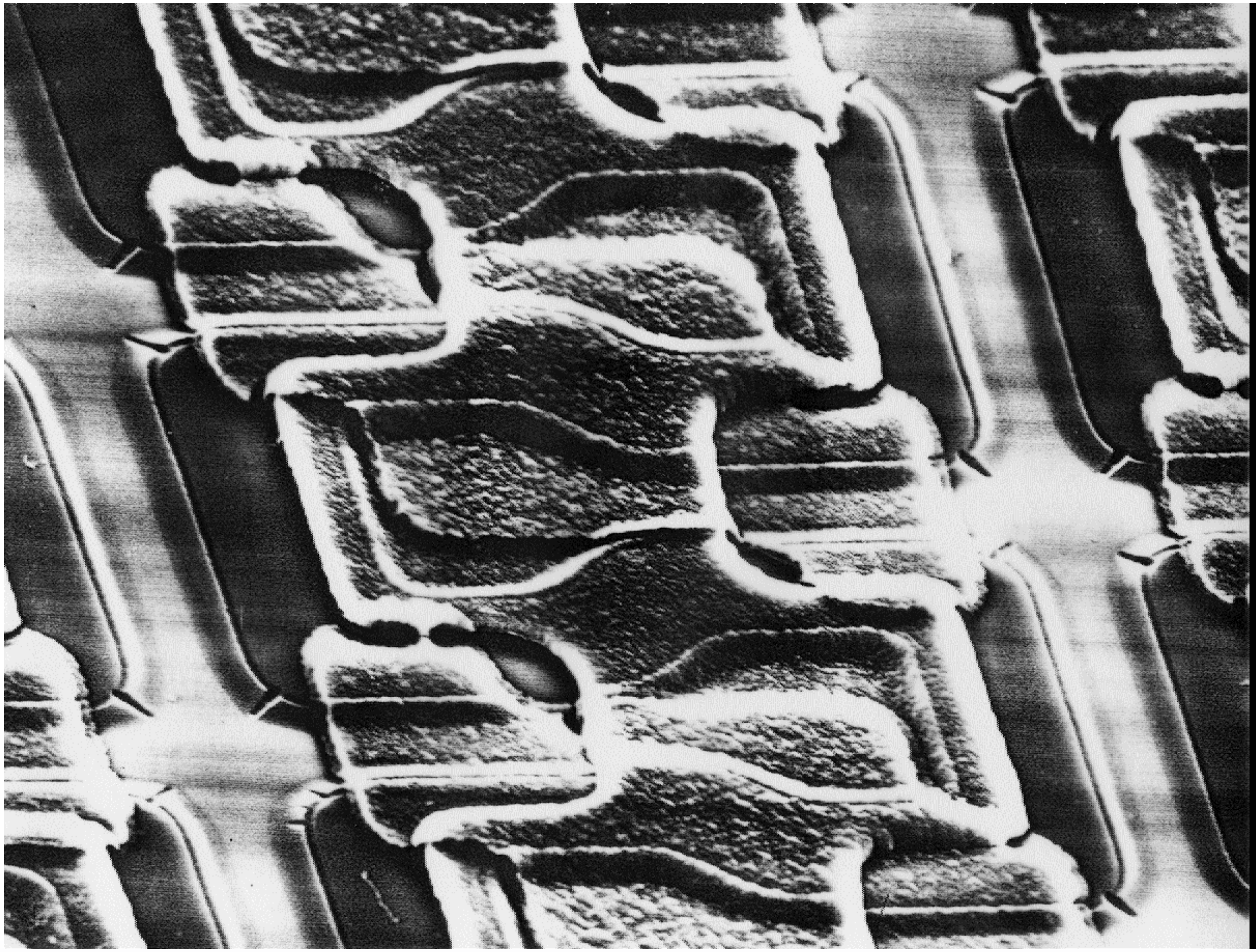


Layout

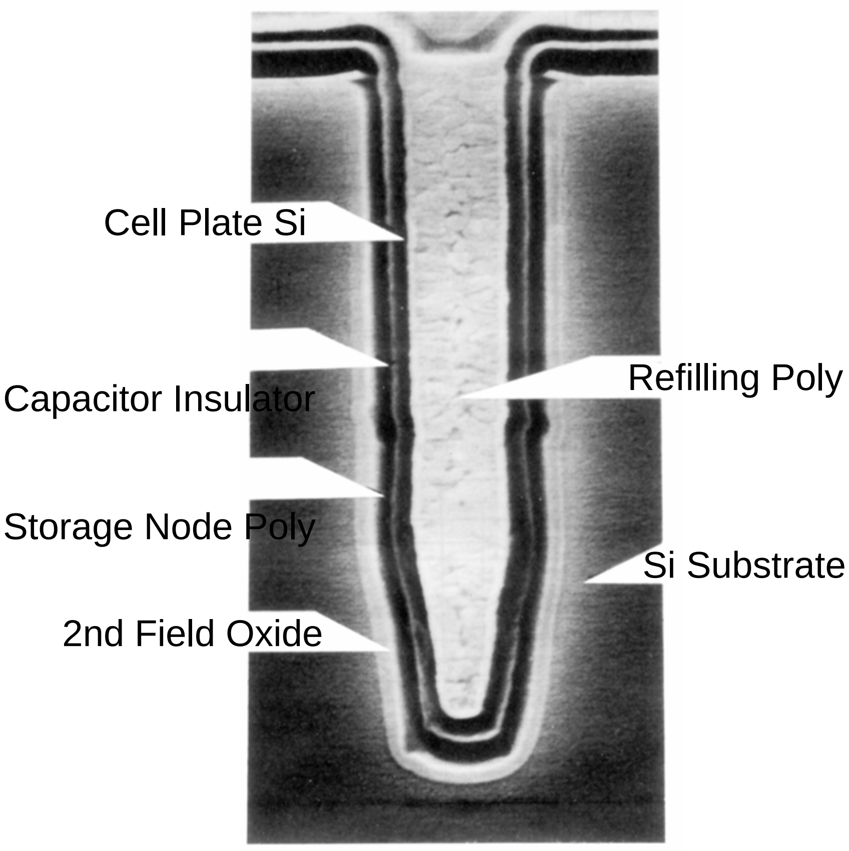
Uses Polysilicon-Diffusion Capacitance

Expensive in Area

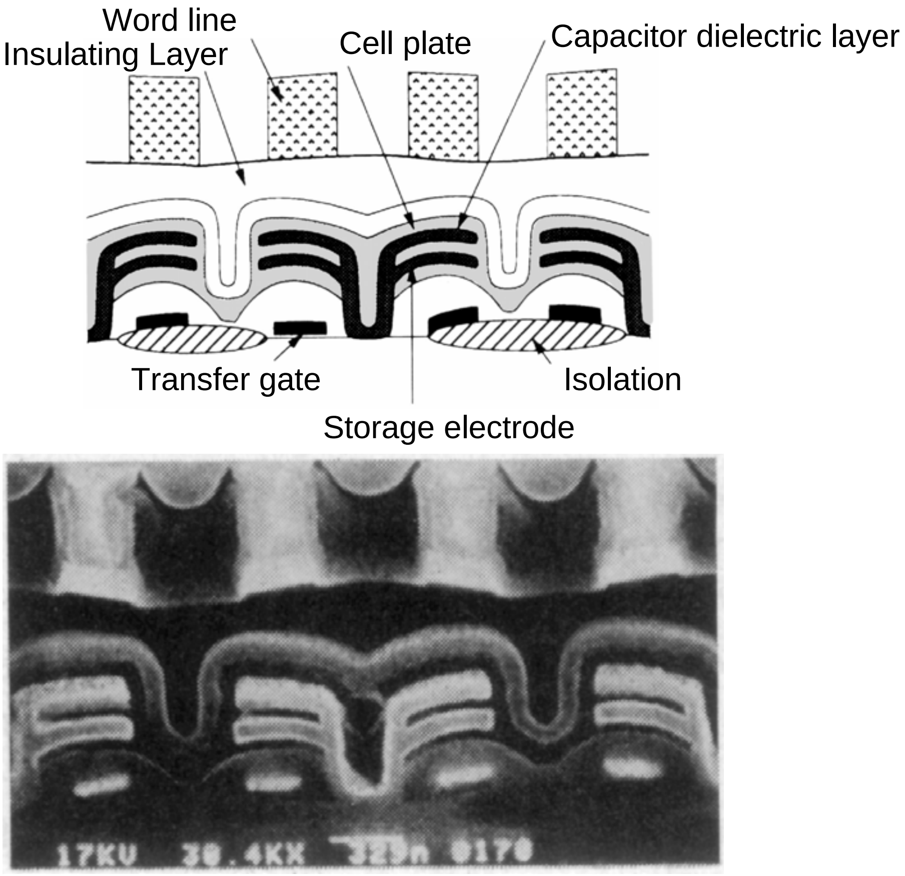
SEM of poly-diffusion capacitor 1T-DRAM



Advanced 1T DRAM Cells

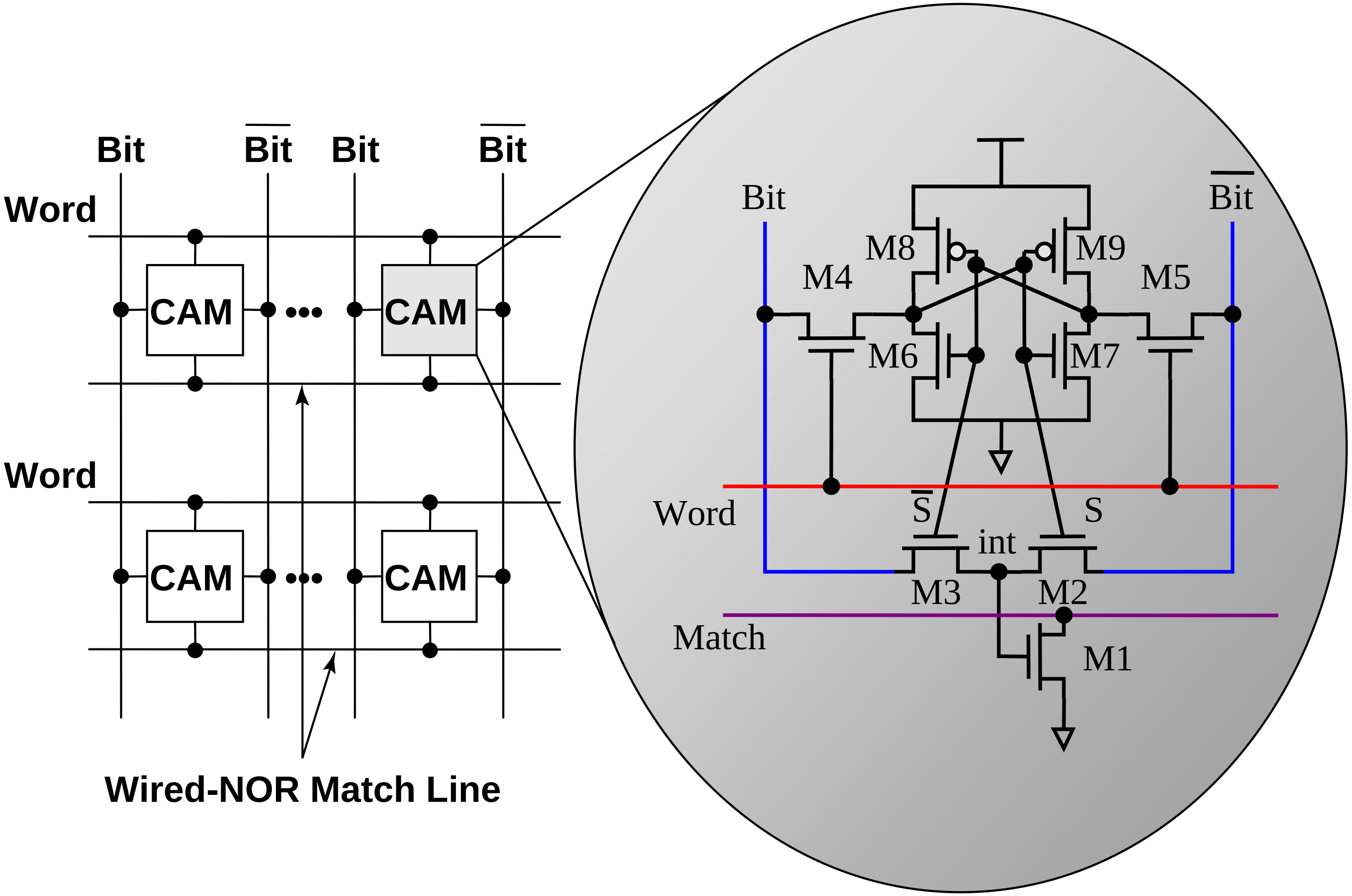


Trench Cell

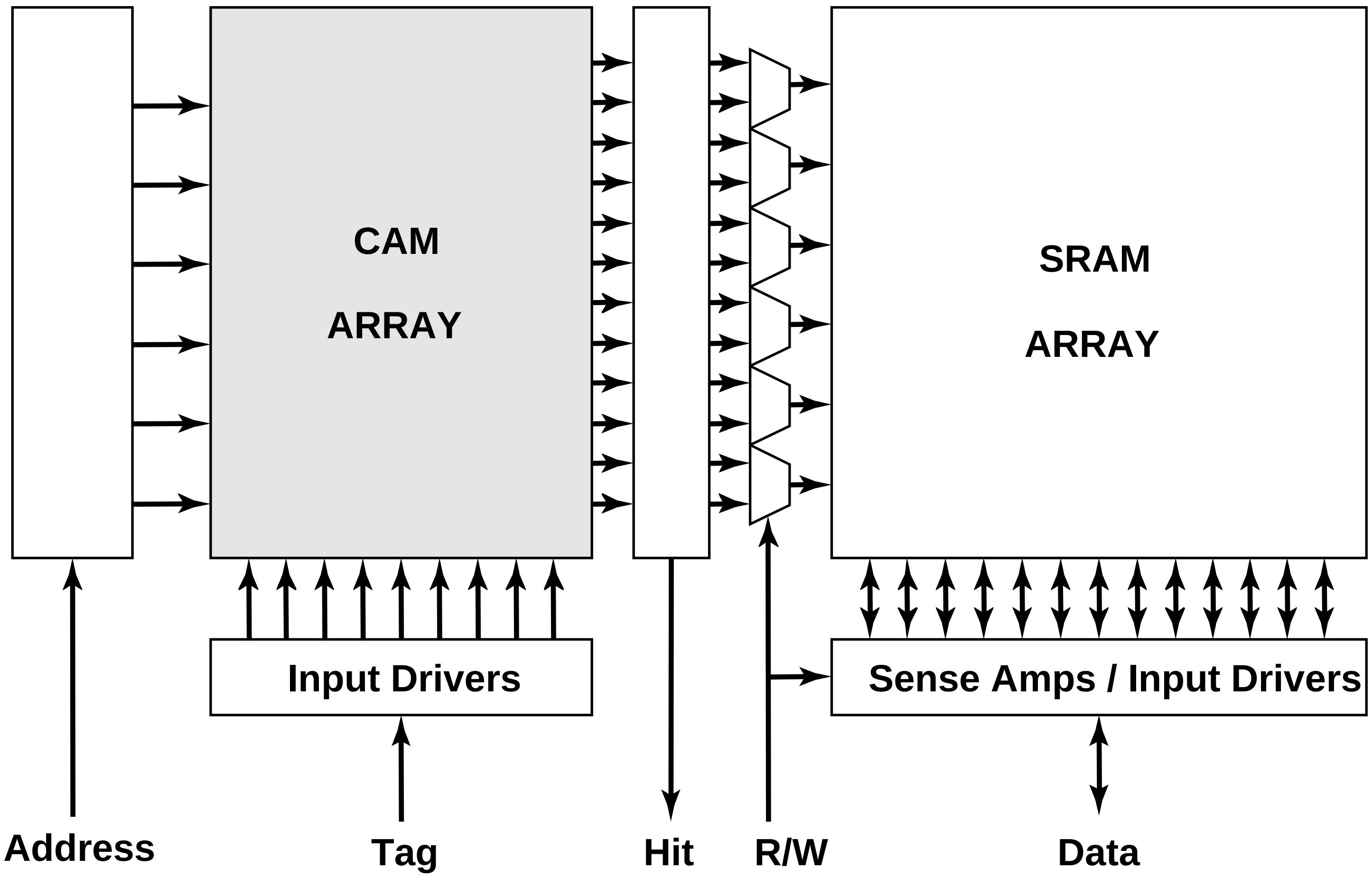


Stacked-capacitor Cell

Static CAM Memory Cell



CAM in Cache Memory



Periphery

- □ Decoders
- □ Sense Amplifiers
- □ Input/Output Buffers
- □ Control / Timing Circuitry

Row Decoders

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

(N)AND Decoder

$$WL_0 = A_0A_1A_2A_3A_4A_5A_6A_7A_8A_9$$

$$WL_{511} = \bar{A}_0A_1A_2A_3A_4A_5A_6A_7A_8A_9$$

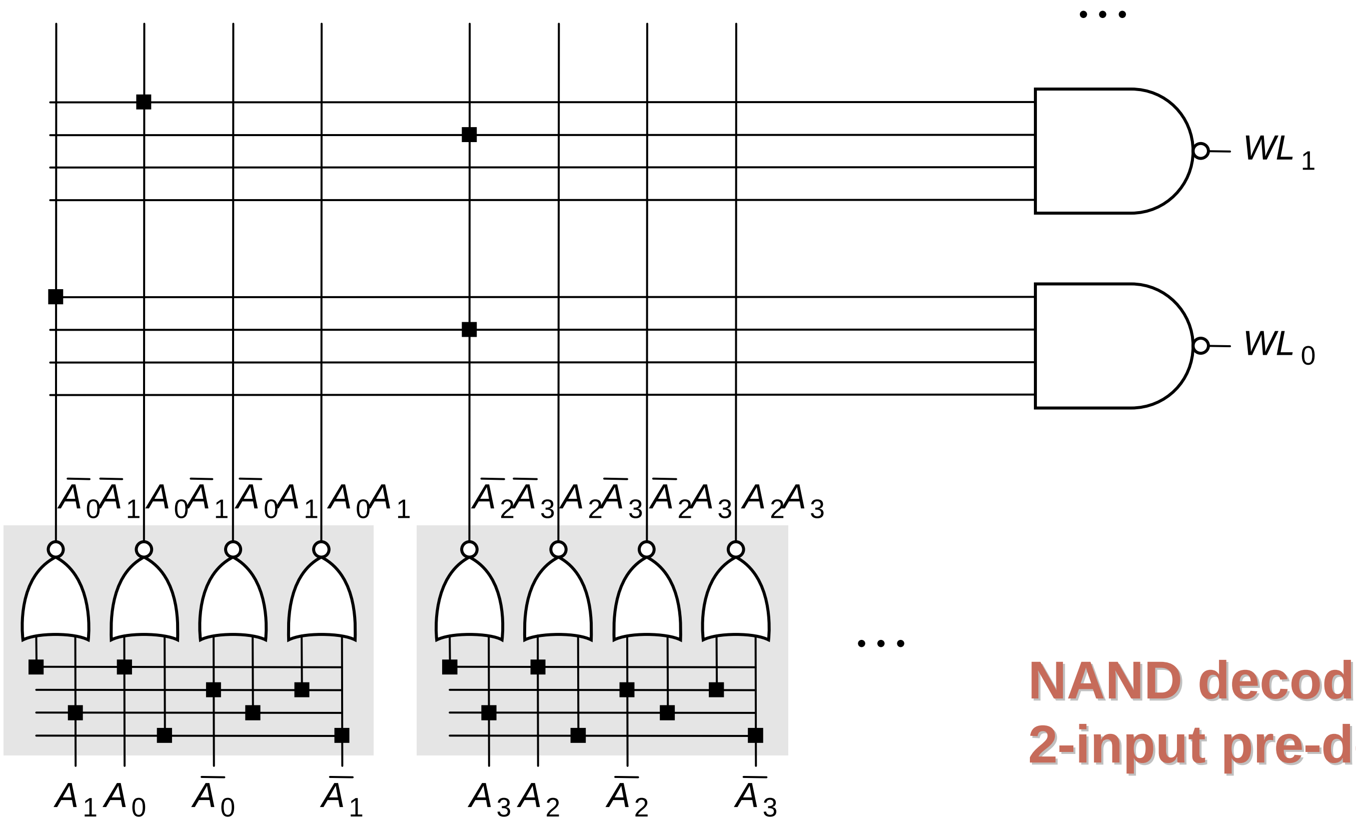
NOR Decoder

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

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

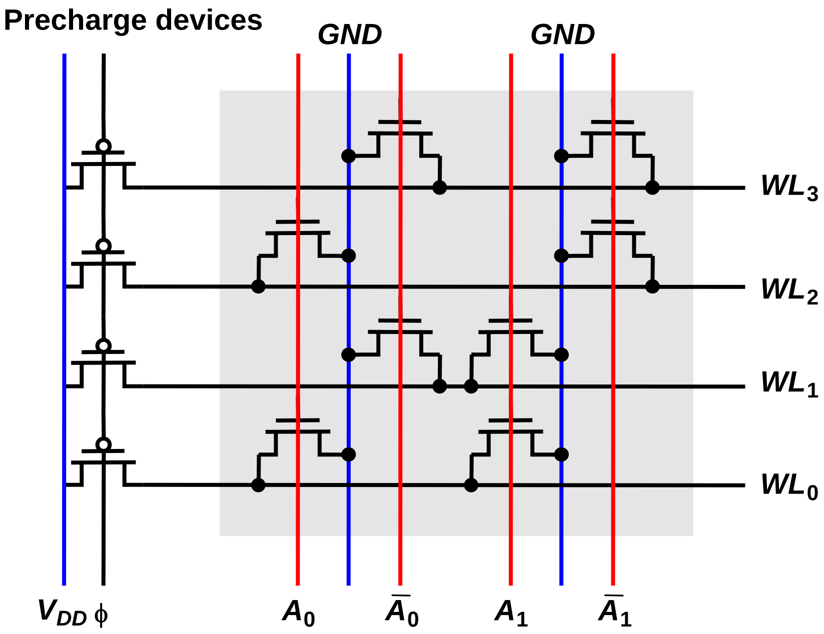
Hierarchical Decoders

Multi-stage implementation improves performance

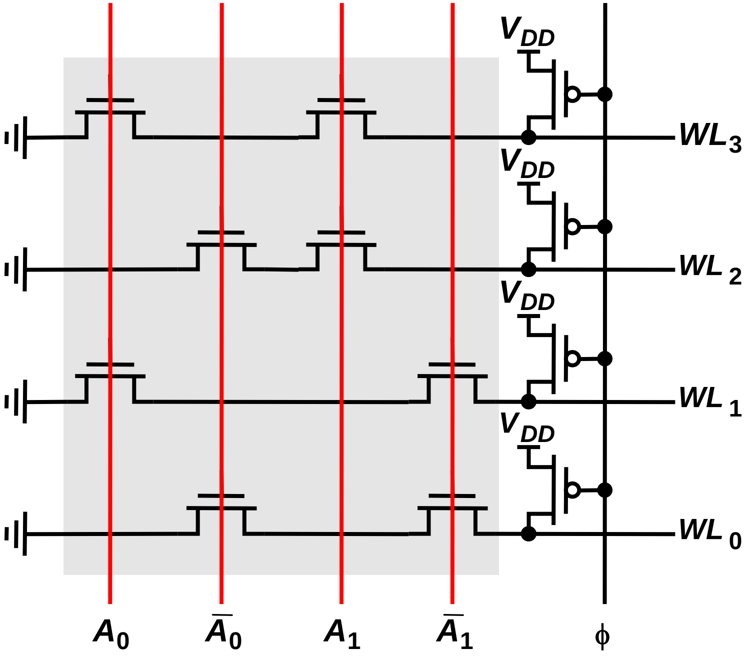


NAND decoder using 2-input pre-decoders

Dynamic Decoders

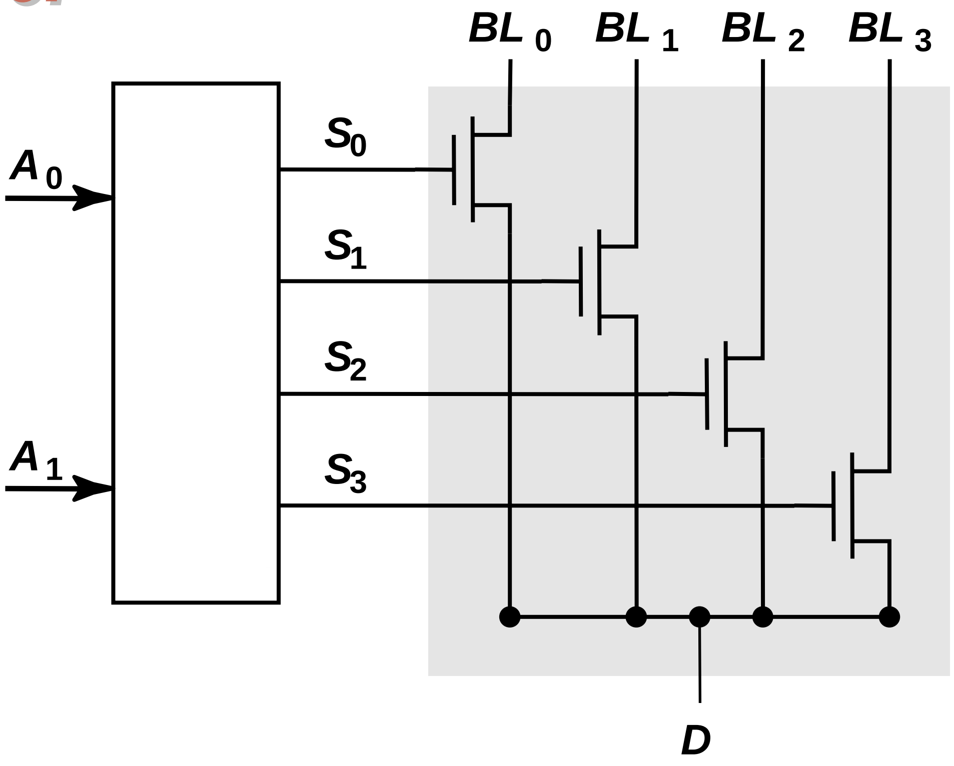


2-input NOR decoder



2-input NAND decoder

4-input pass-transistor based column decoder

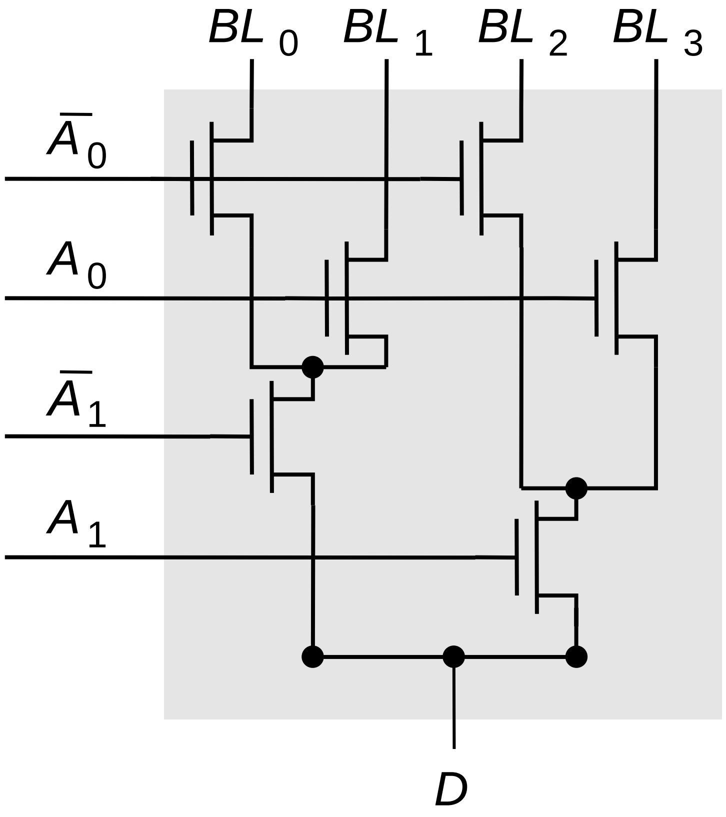


Advantages: speed (**t_{pd} does not add to overall memory access time)**

Only one extra transistor in signal path

Disadvantage: Large transistor count

4-to-1 tree based column decoder

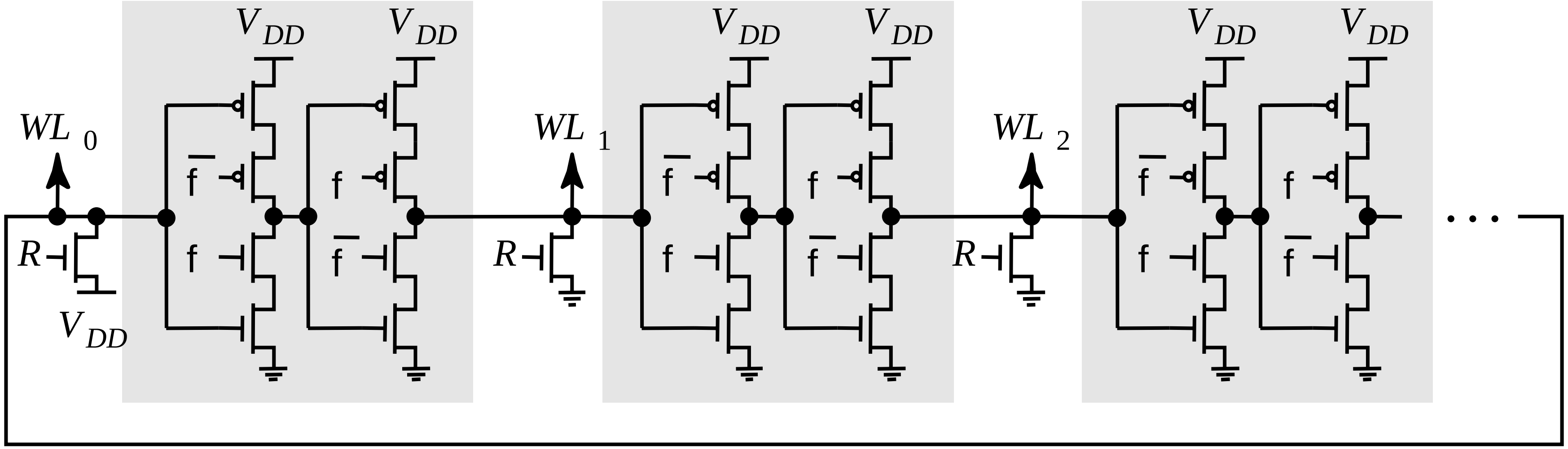


Number of devices drastically reduced

Delay increases quadratically with # of sections; prohibitive for large decoders

Solutions: buffers
progressive sizing
combination of tree and pass transistor approaches

Decoder for circular shift-register



Sense Amplifiers

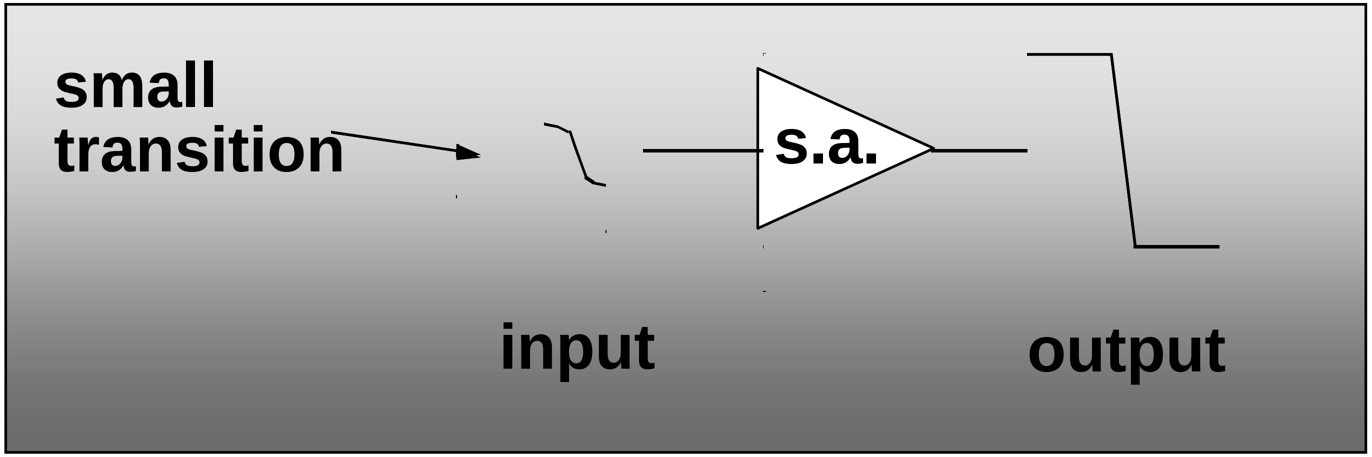
$$t_p = \frac{C \cdot \Delta V}{I_{av}}$$

make ΔV as small as possible

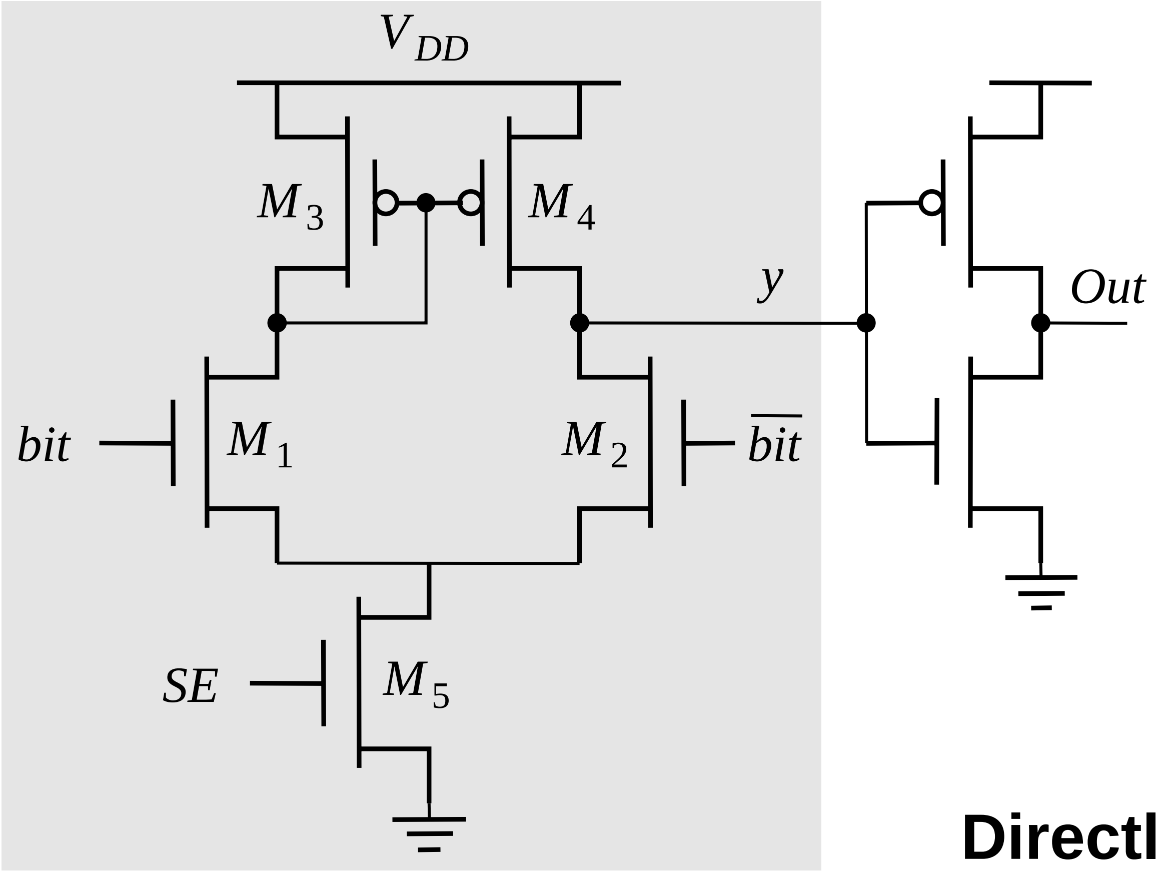
large

small

Idea: Use Sense Amplifer

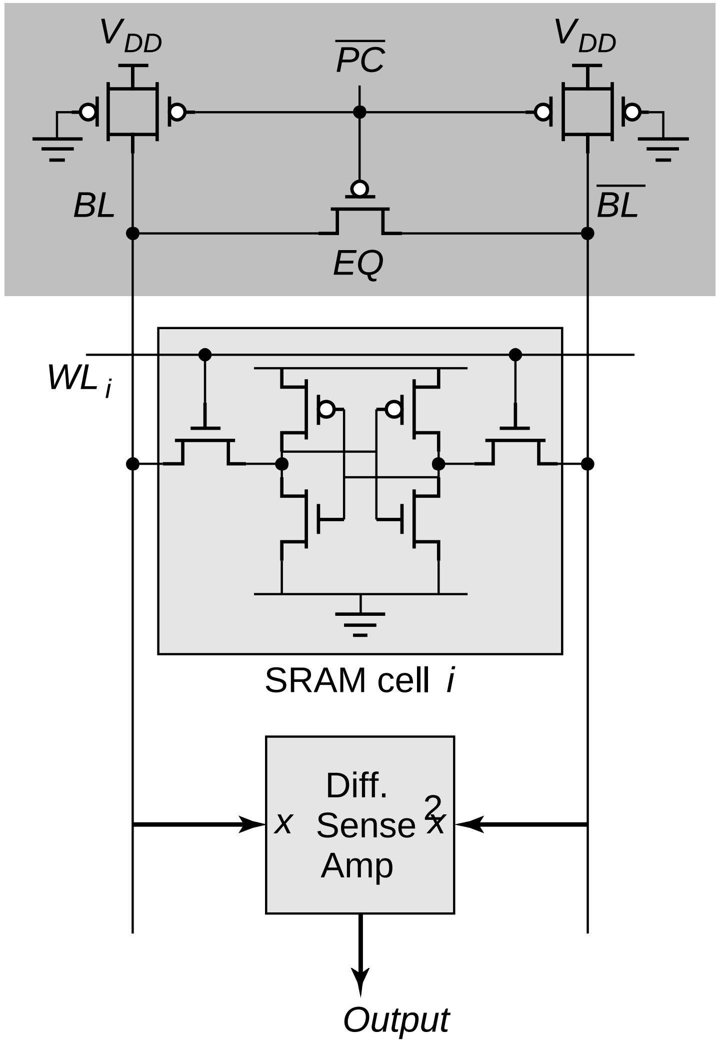


Differential Sense Amplifier

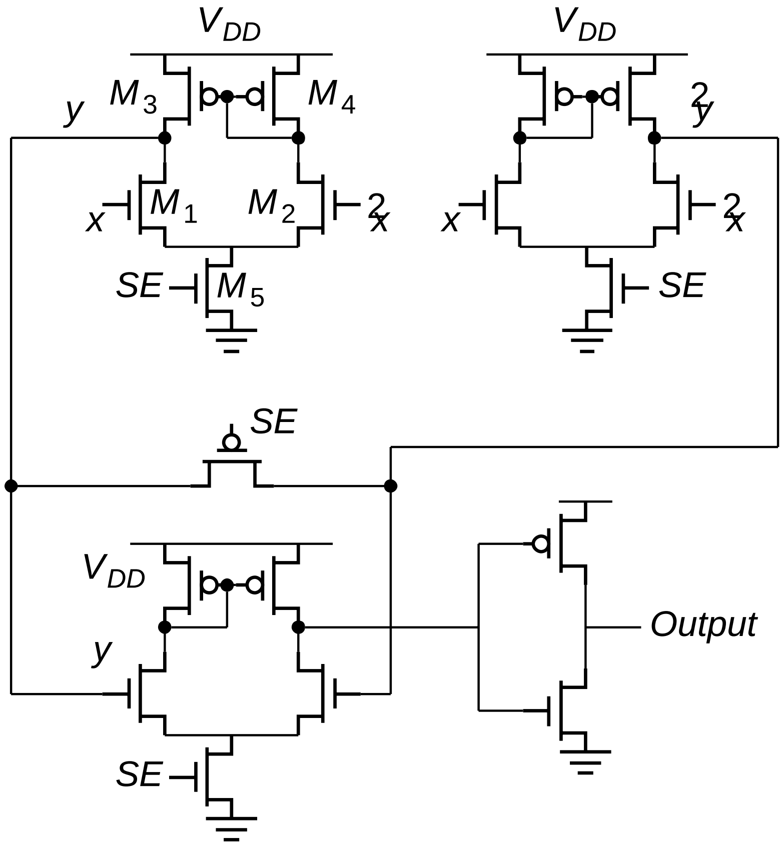


Directly applicable to SRAMs

Differential Sensing – SRAM

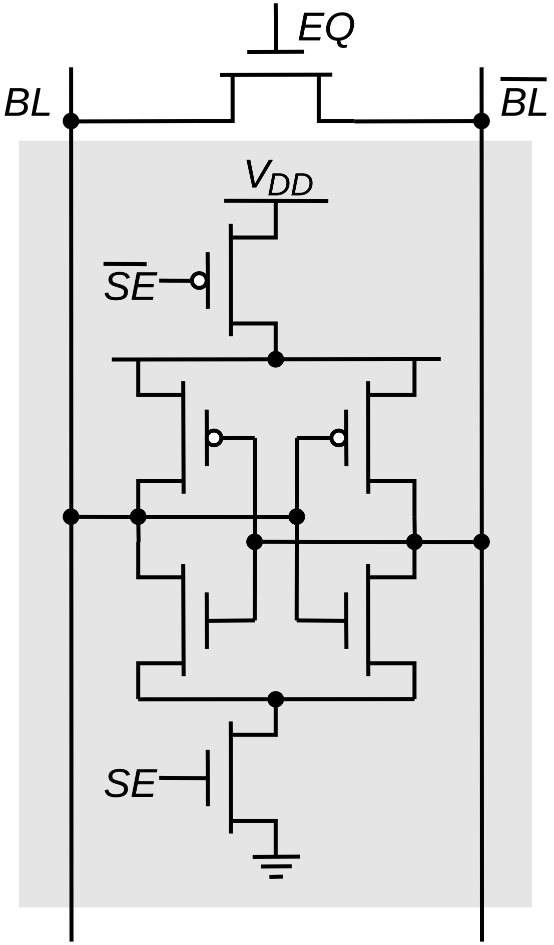


(a) SRAM sensing scheme



(b) two stage differential amplifier

Latch-Based Sense Amplifier (DRAM)



Initialized in its meta-stable point with EQ

Once adequate voltage gap created, sense amp enabled with SE

Positive feedback quickly forces output to a stable operating point.

Reliability and Yield

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

⇒ Highly Sensitive to Noise (Crosstalk, Supply Noise)

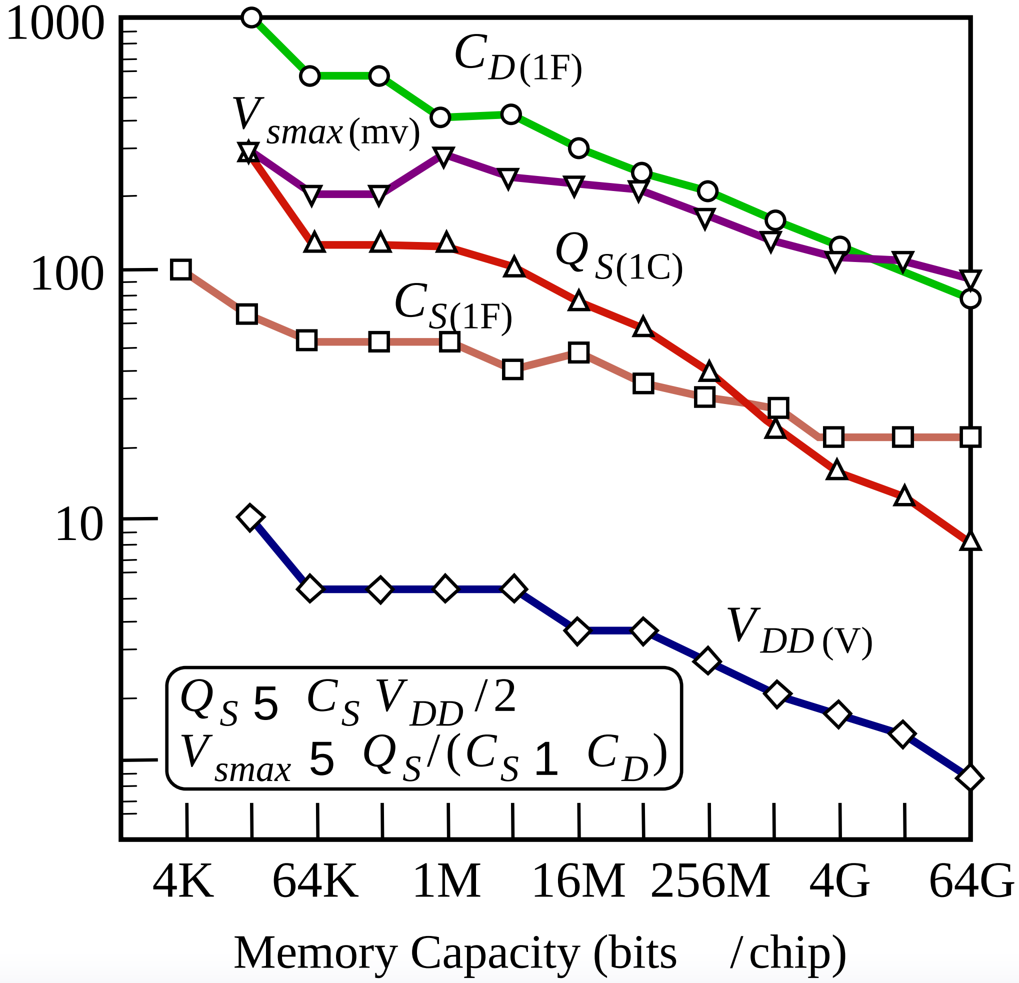
- High Density and Large Die size cause Yield Problems

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

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

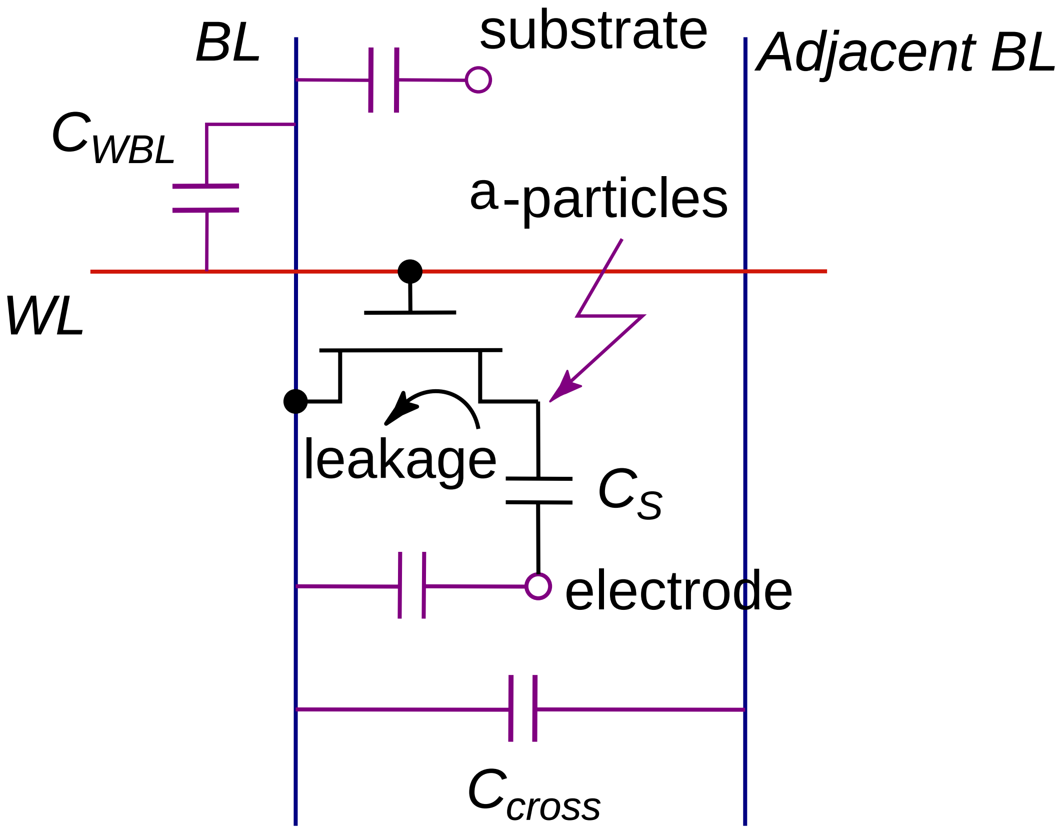
Increase Yield using Error Correction and Redundancy

Sensing Parameters in DRAM

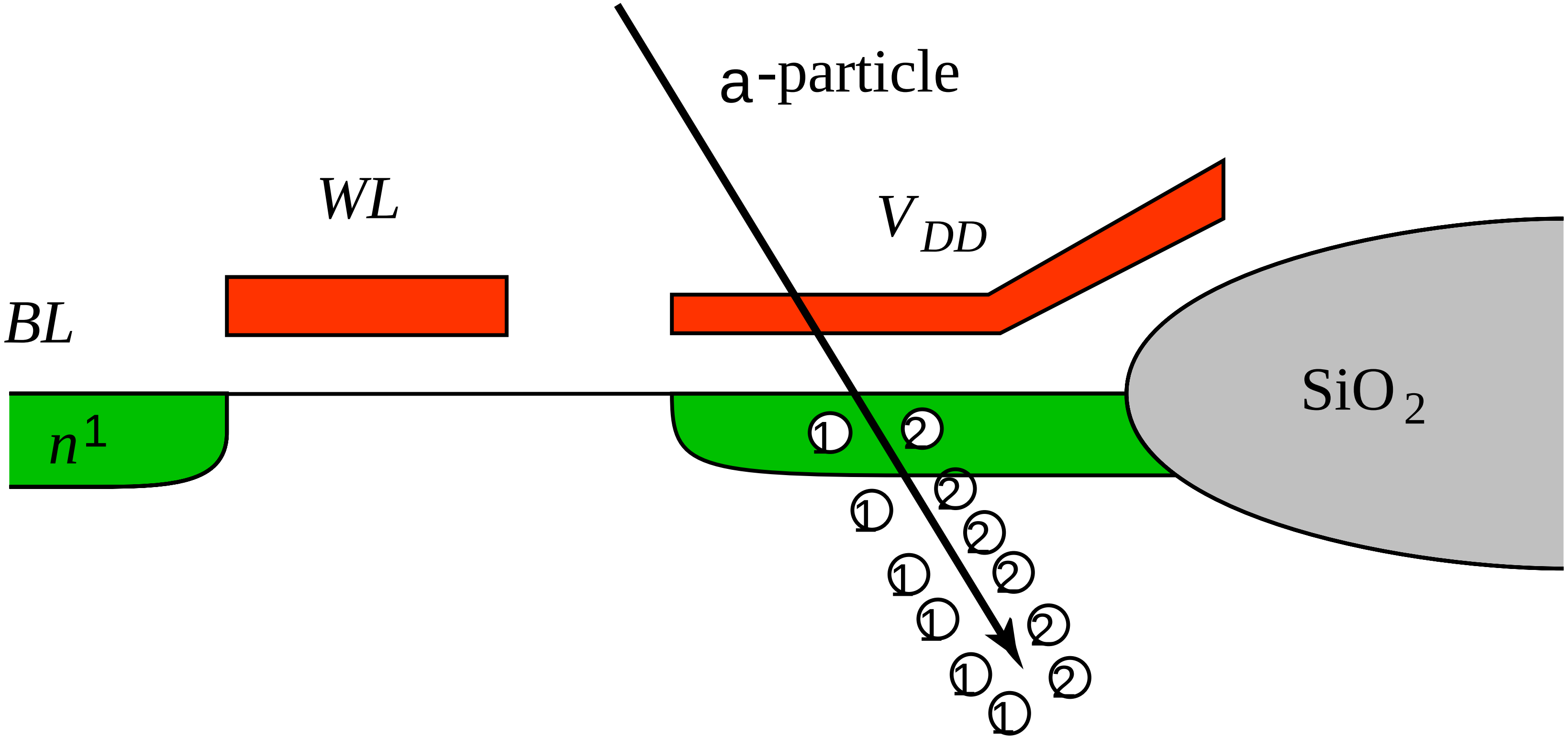


From [Itoh01]

Noise Sources in 1T DRam

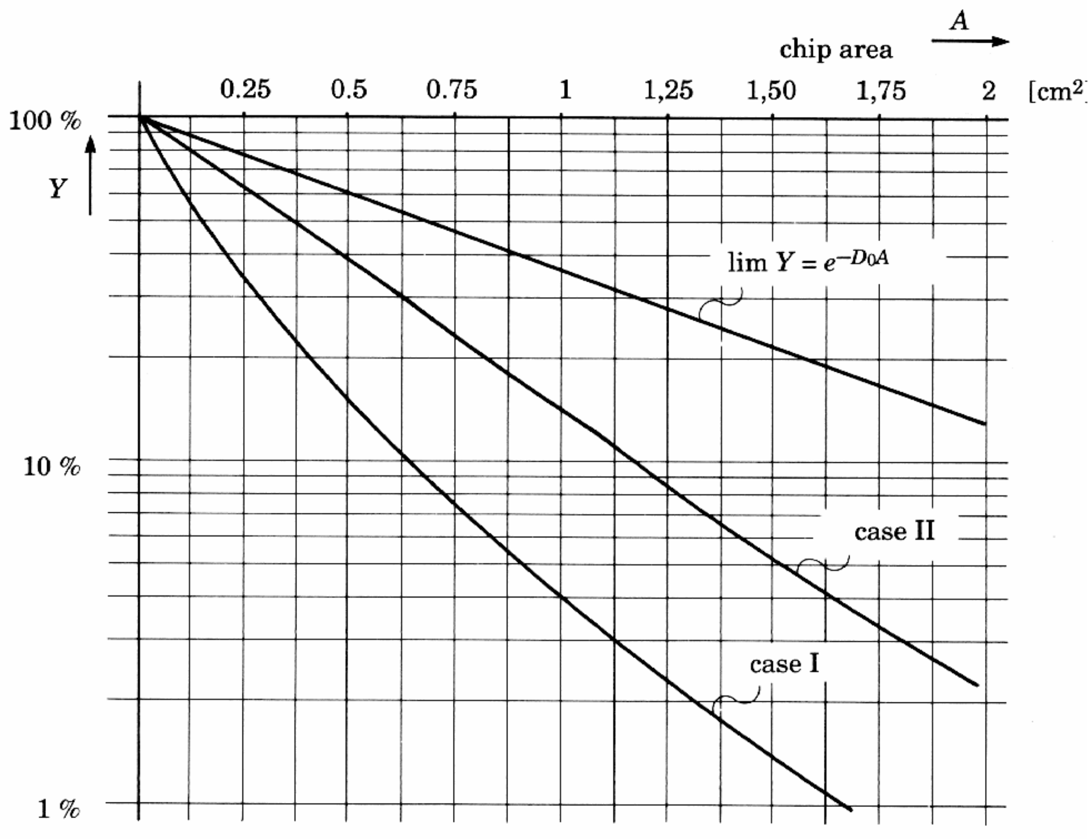


Alpha-particles (or Neutrons)



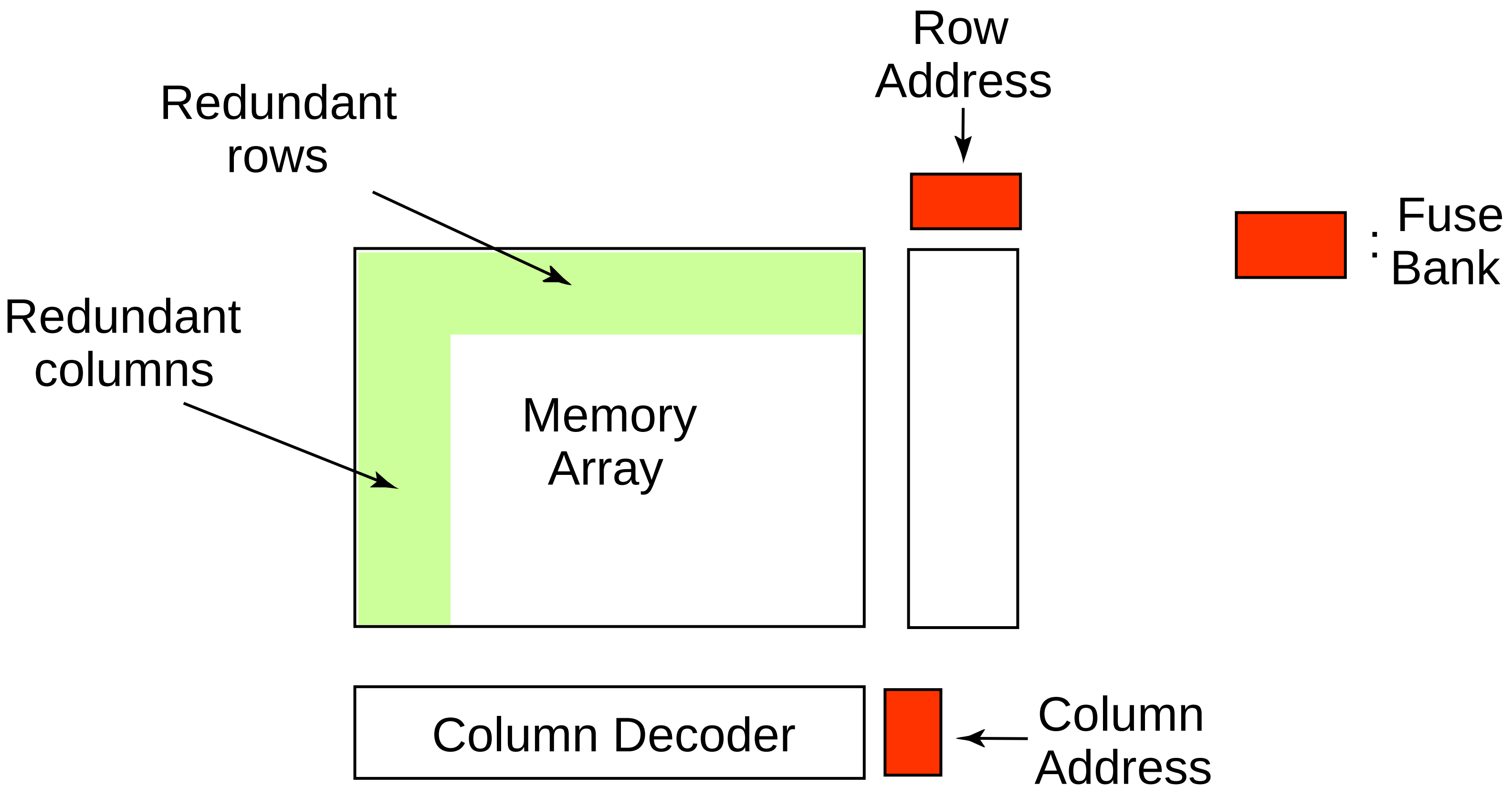
1 Particle ~ 1 Million Carriers

Yield



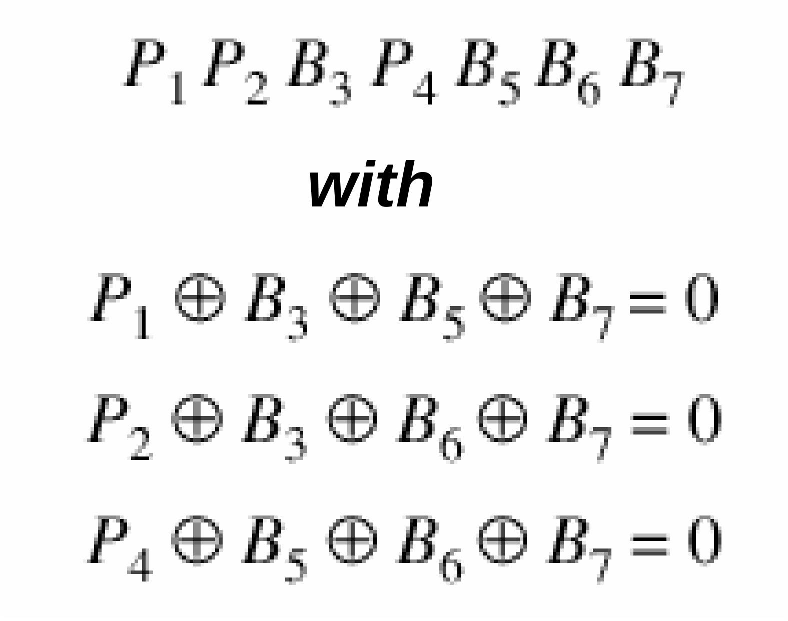
Yield curves at different stages of process maturity (from [Veendrick92])

Redundancy


Row
Address
Redundant
rows
Fuse
Bank
Redundant
columns
Memory
Array
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

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

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

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

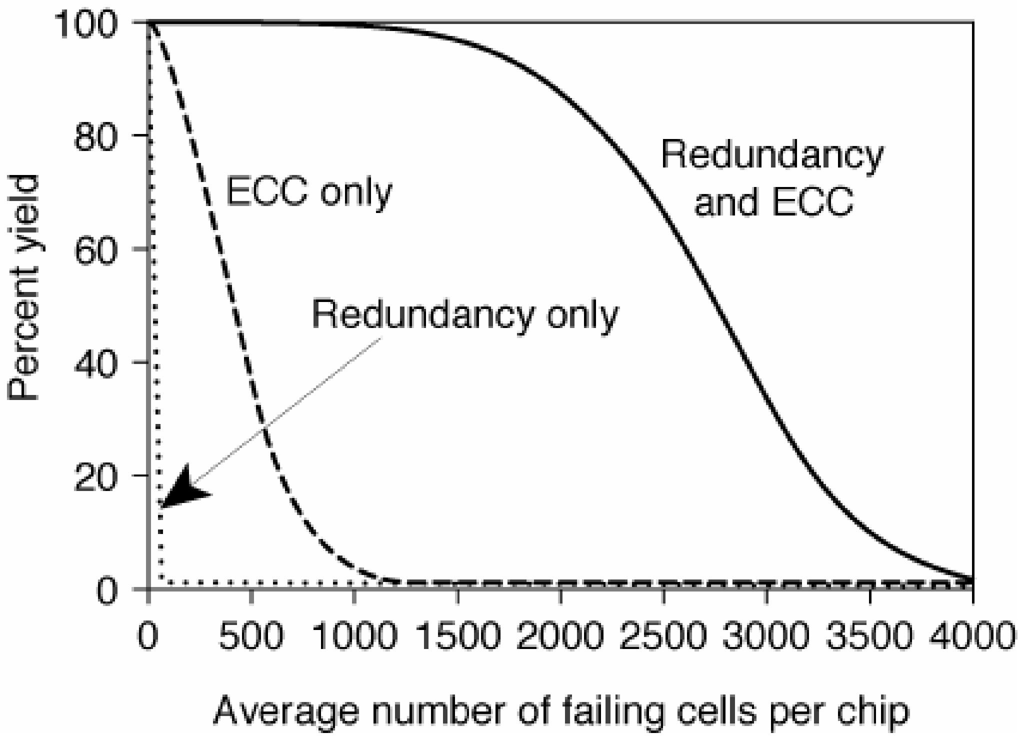
e.g. B3 Wrong

1

1 ***= 3***

0

Redundancy and Error Correction



Case Studies

- ❑ Programmable Logic Array
- ❑ SRAM
- ❑ Flash Memory

PLA versus ROM

- □ **Programmable Logic Array**

  structured approach to random logic
  “two level logic implementation”
    NOR-NOR (product of sums)
    NAND-NAND (sum of products)

IDENTICAL TO ROM!

- □ **Main difference**

  ROM: fully populated
  PLA: one element per minterm

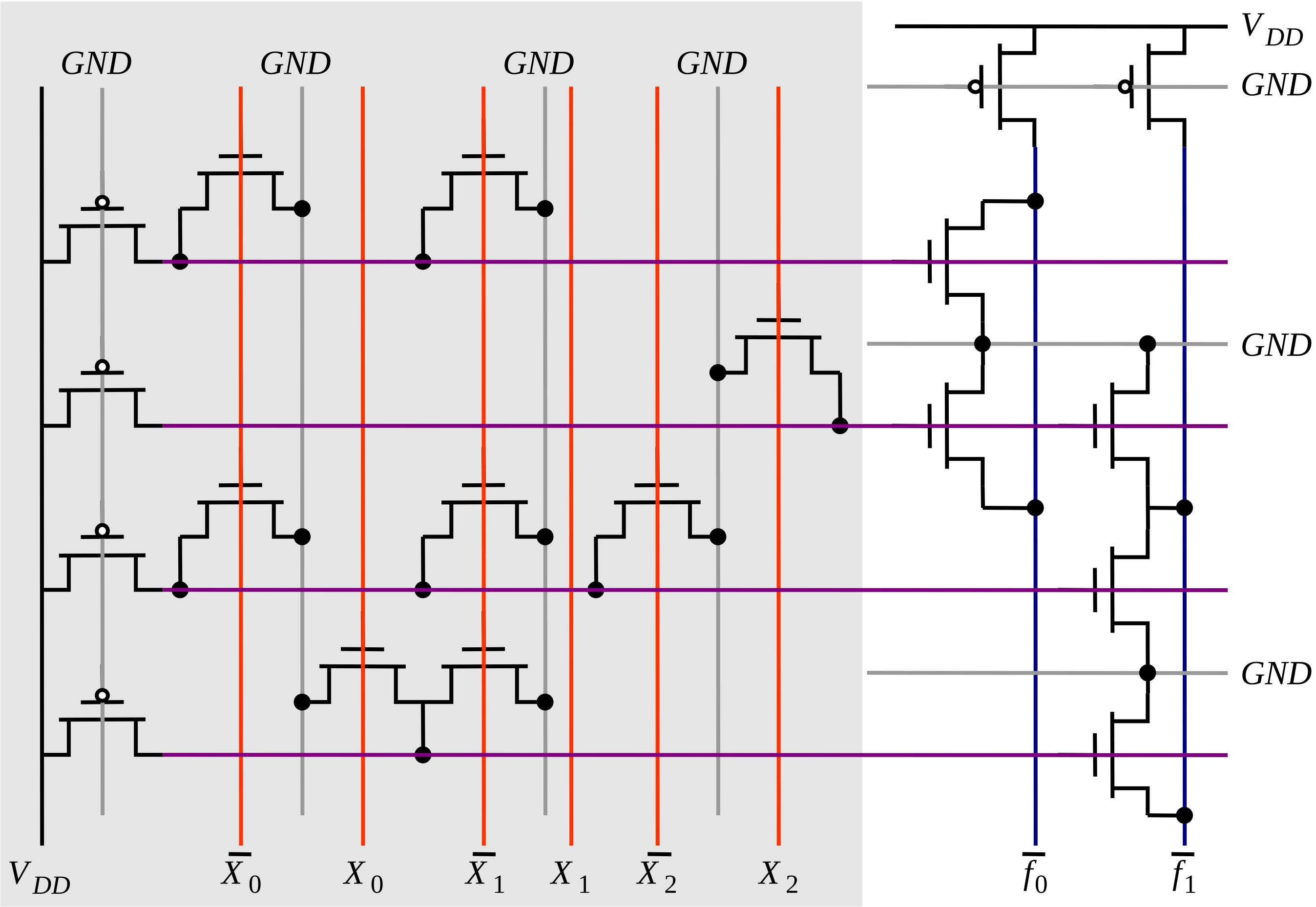
Note: Importance of PLA’s has drastically reduced

1. **slow**
2. **better software techniques (mutli-level logic synthesis)**

But …

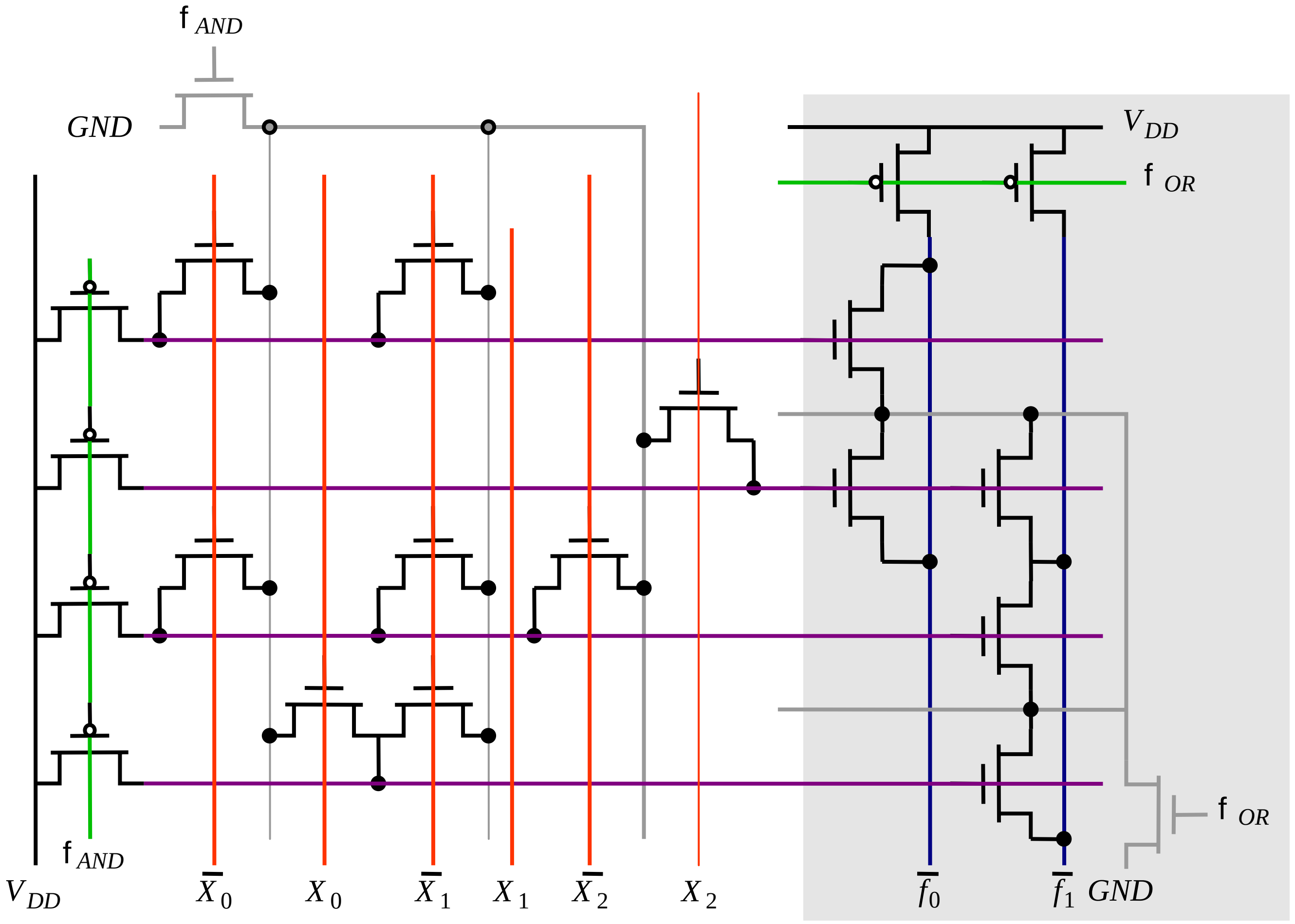
Programmable Logic Array

Pseudo-NMOS PLA



AND-plane **OR-plane**

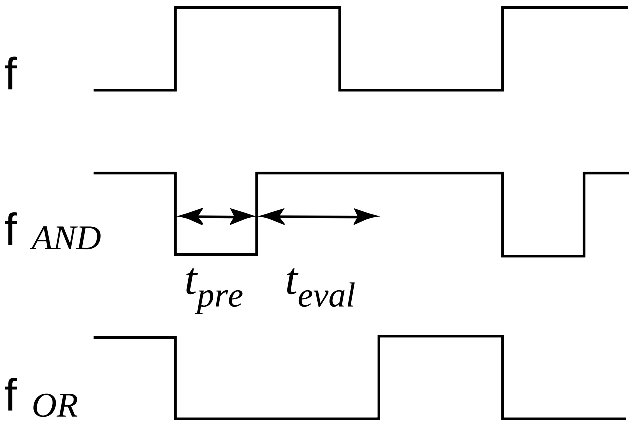
Dynamic PLA



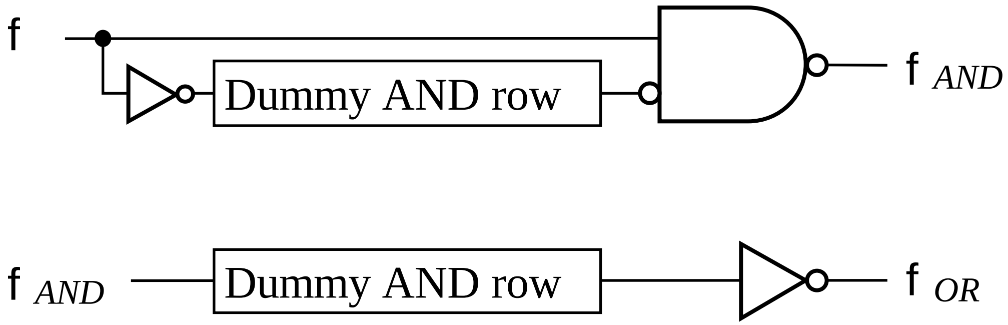
AND-plane

OR-plane

Clock Signal Generation for self-timed dynamic PLA

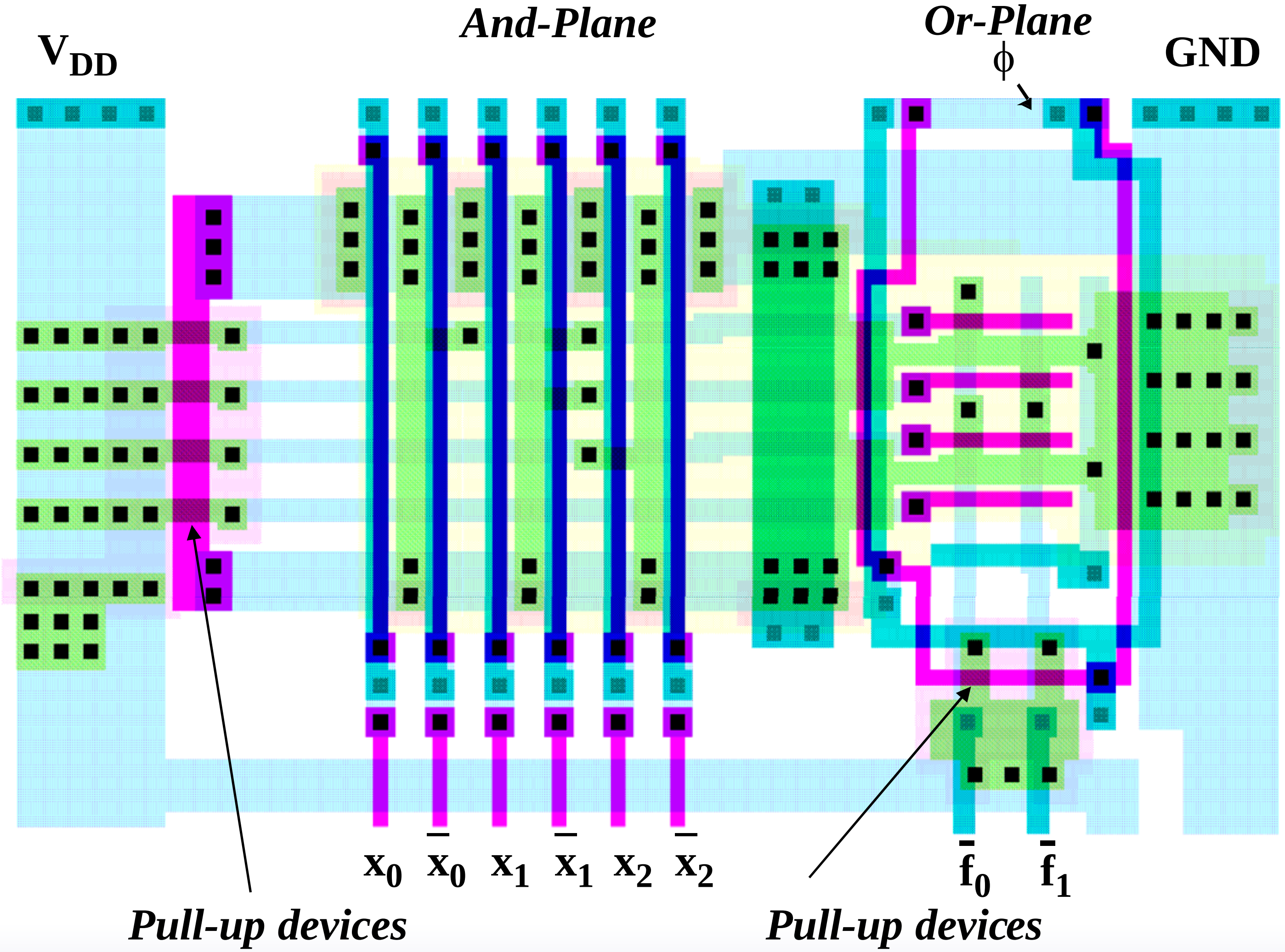


(a) Clock signals

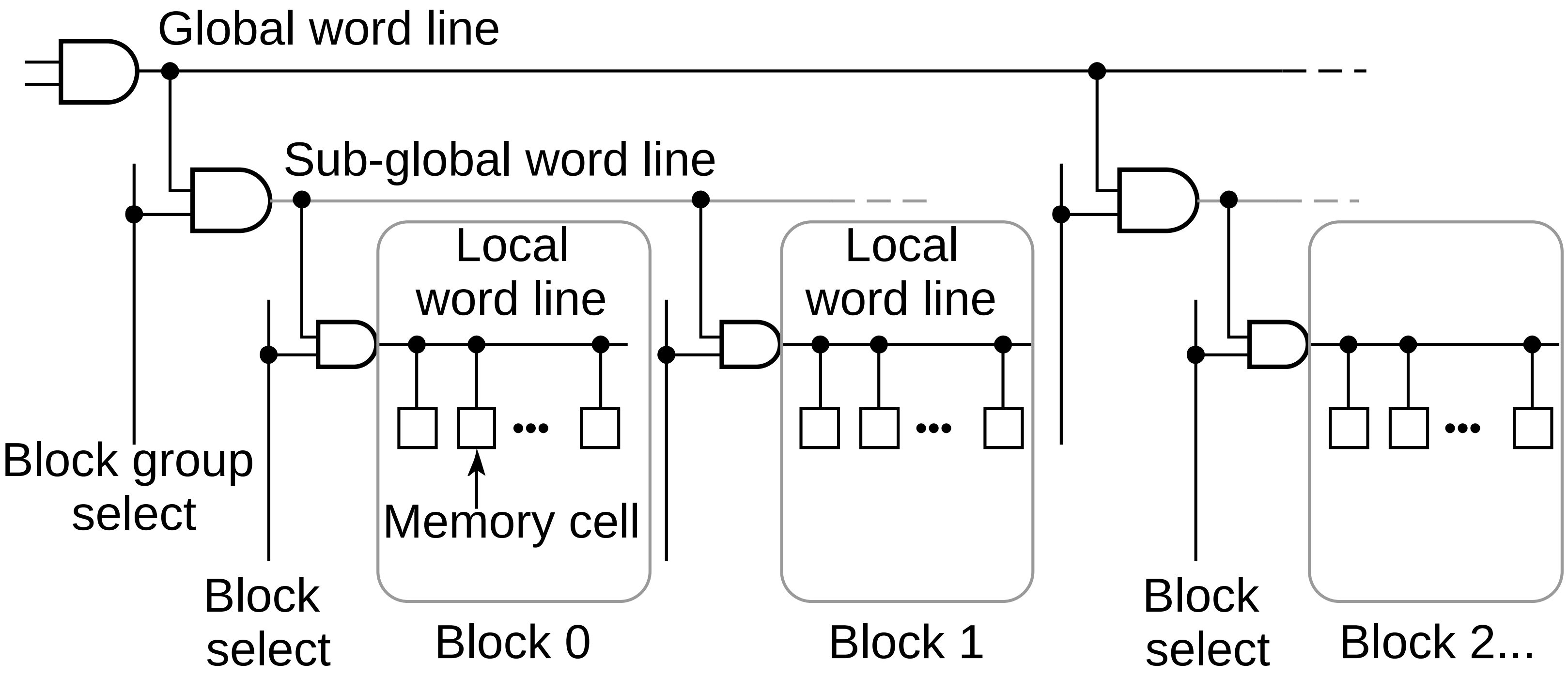


(b) Timing generation circuitry

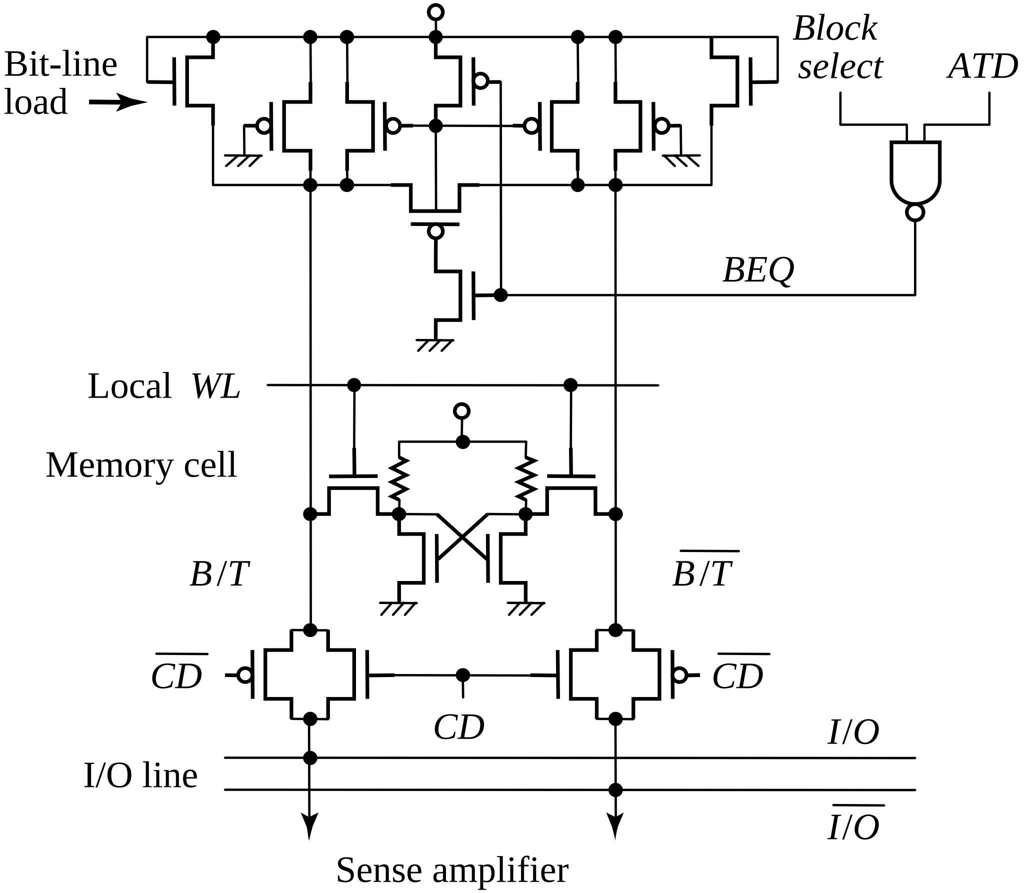
PLA Layout


And-Plane
Or-Plane
ϕ
V_{DD}
GND
x_0
$\overline{x}_0$
x_1
$\overline{x}_1$
x_2
$\overline{x}_2$
$\overline{f}_0$
$\overline{f}_1$
Pull-up devices
Pull-up devices

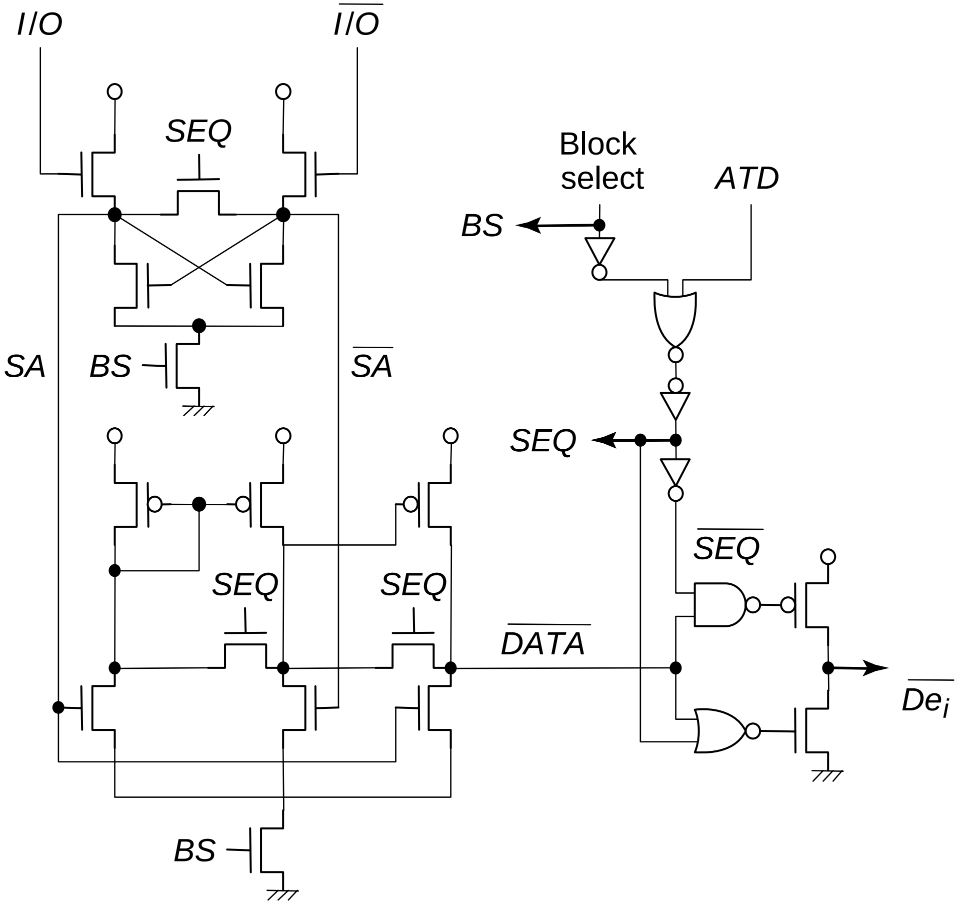
4 Mbit SRAM
Hierarchical Word-line Architecture

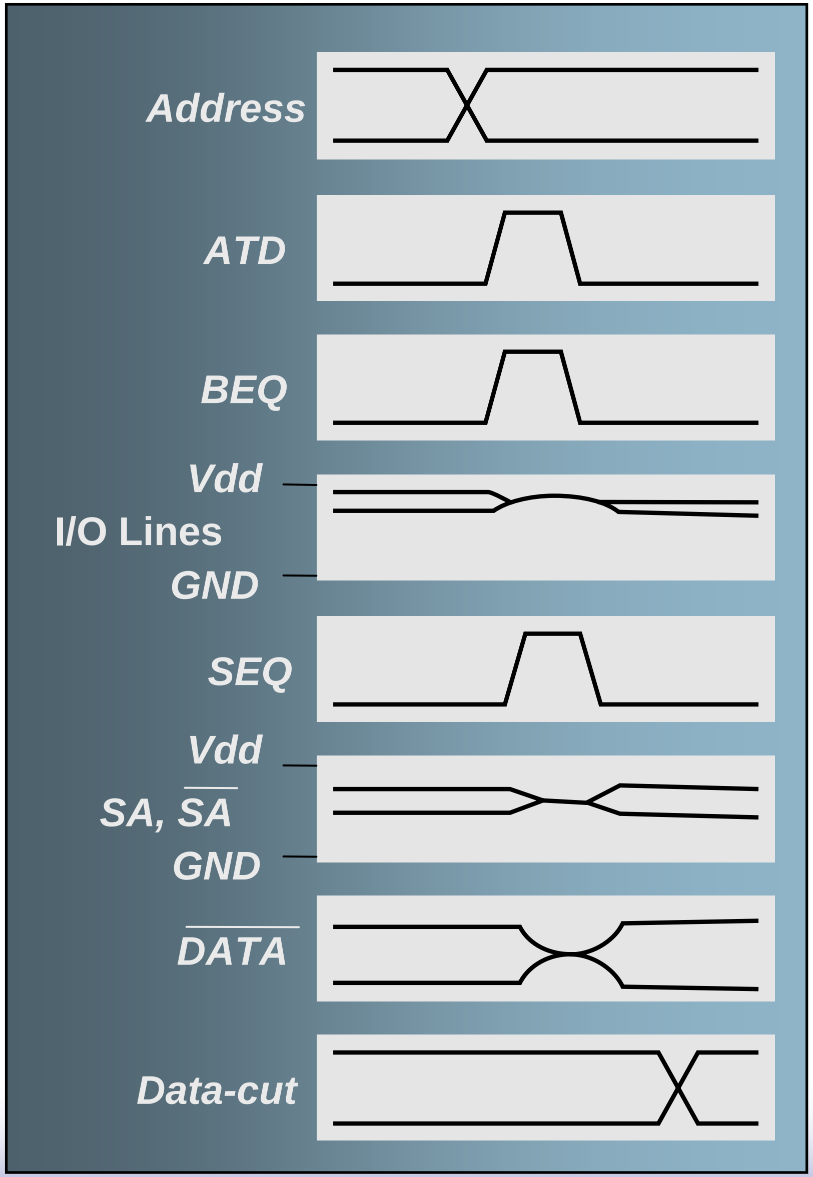


Bit-line Circuitry

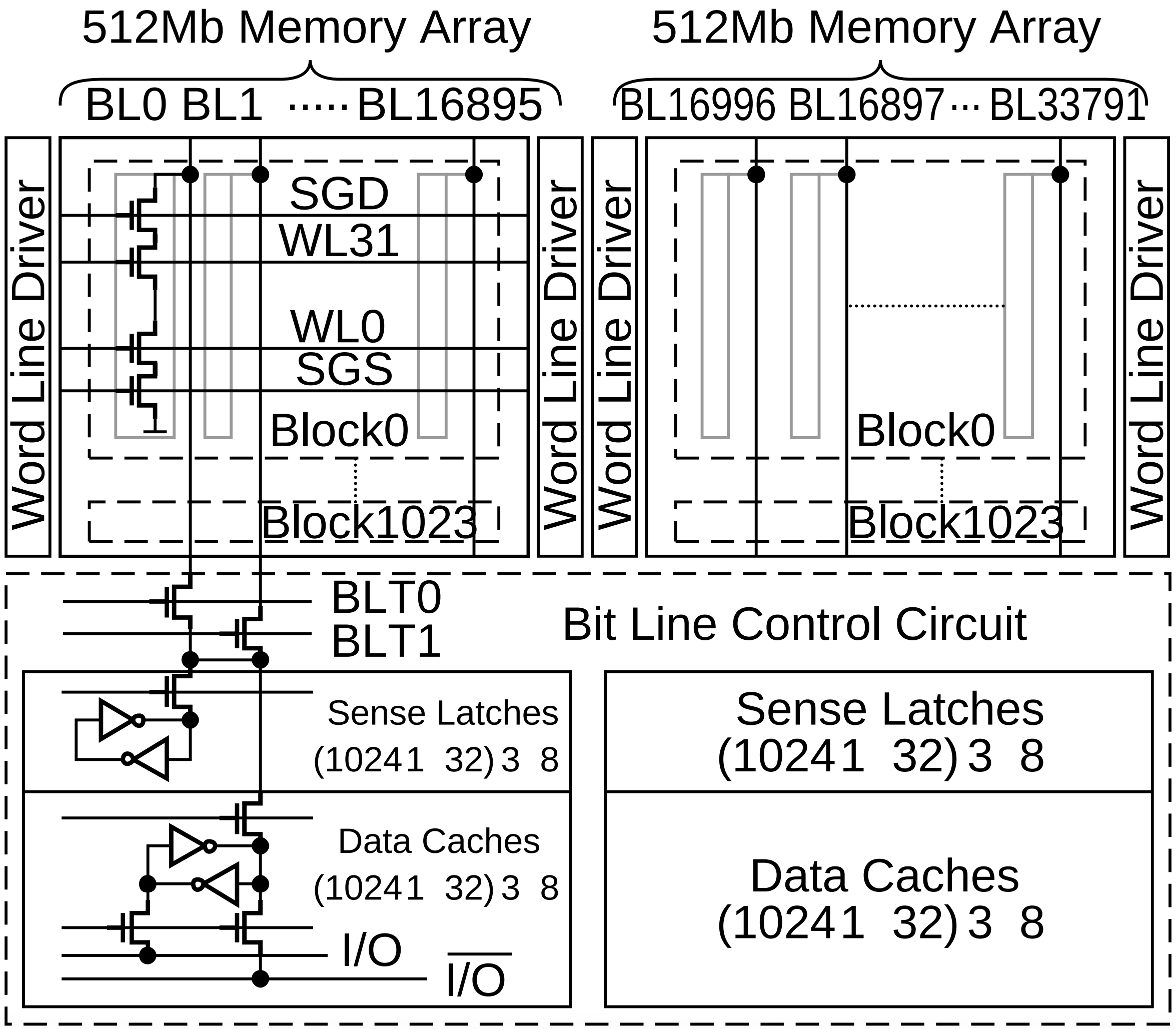


Sense Amplifier (and Waveforms)



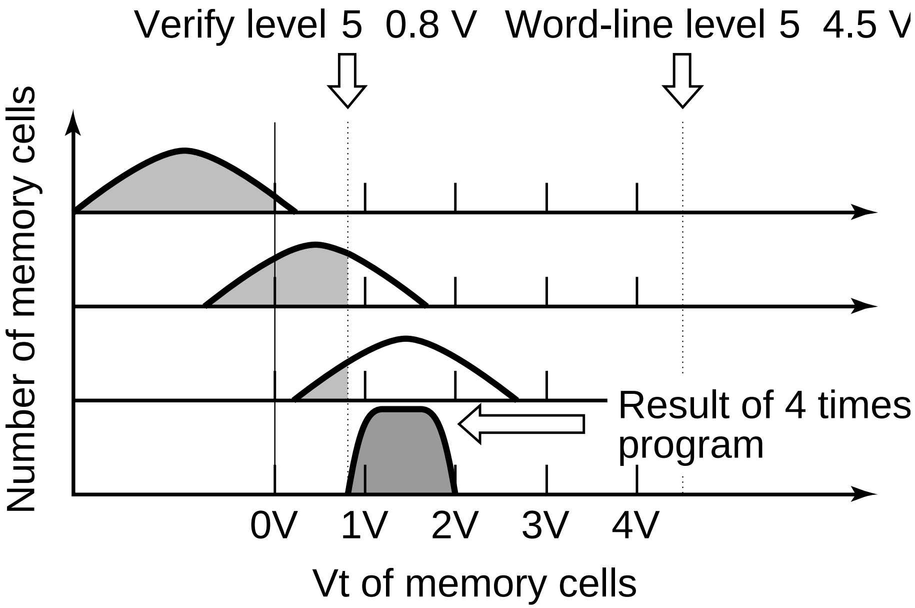


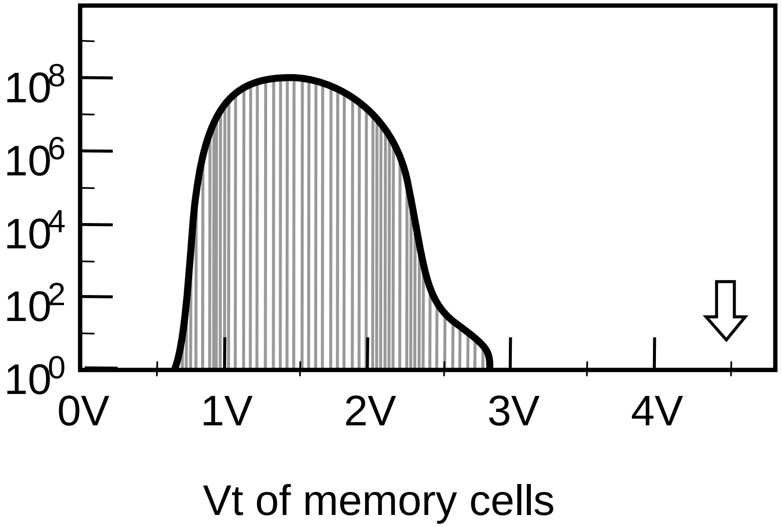
1 Gbit Flash Memory



From [Nakamura02]

Writing Flash Memory



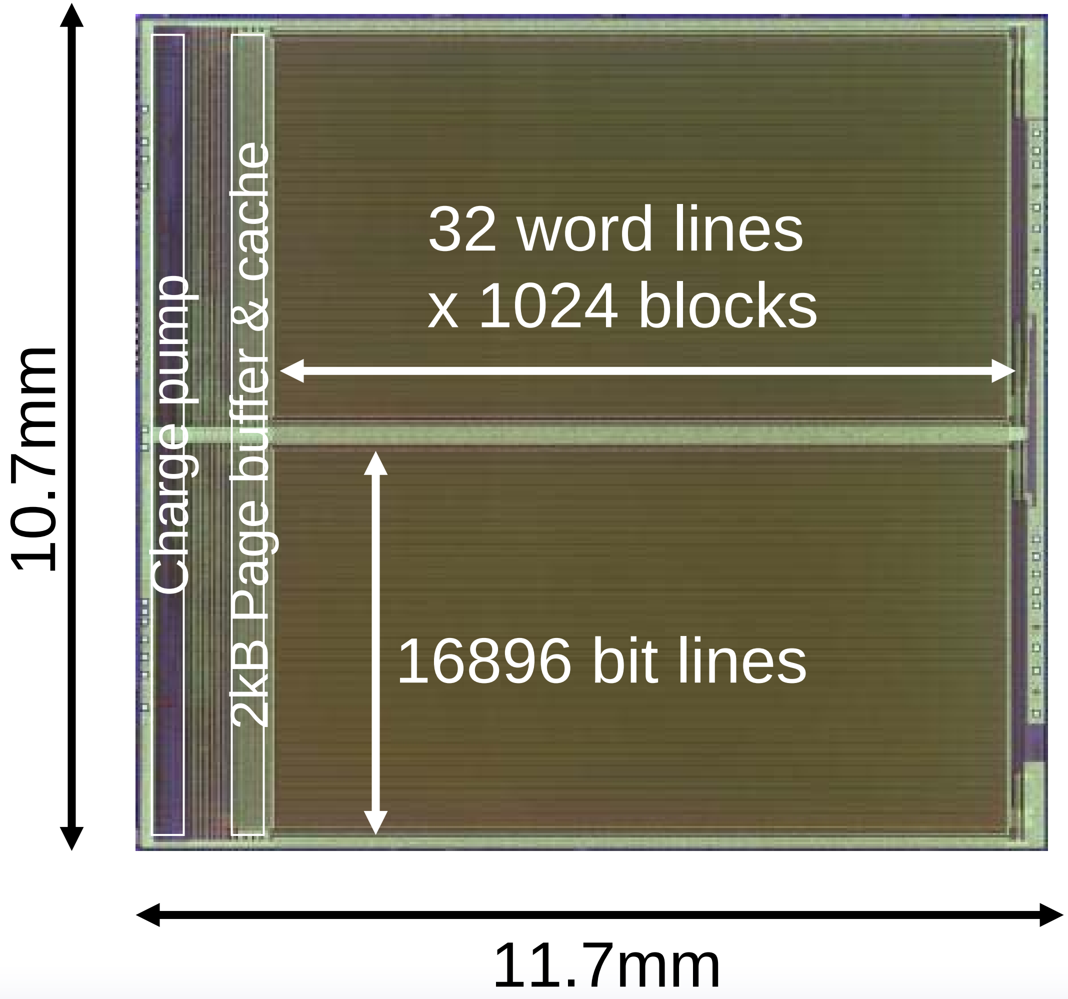


Evolution of thresholds

Final Distribution

From [Nakamura02]

$125mm^2$ 1Gbit NAND Flash Memory



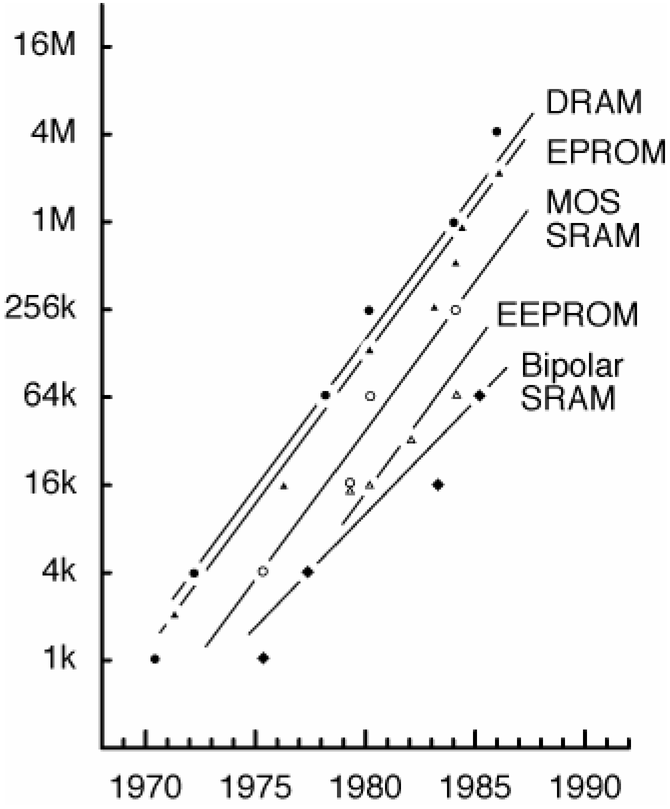
From [Nakamura02]

$125mm^2$ 1Gbit NAND Flash Memory

Technology	0.13μm p-sub CMOS triple-well 1poly, 1polycide, 1W, 2Al
Cell size	0.077μm2
Chip size	125.2mm2
Organization	2112 x 8b x 64 page x 1k block
Power supply	2.7V-3.6V
Cycle time	50ns
Read time	25μs
Program time	200μs / page
Erase time	2ms / block

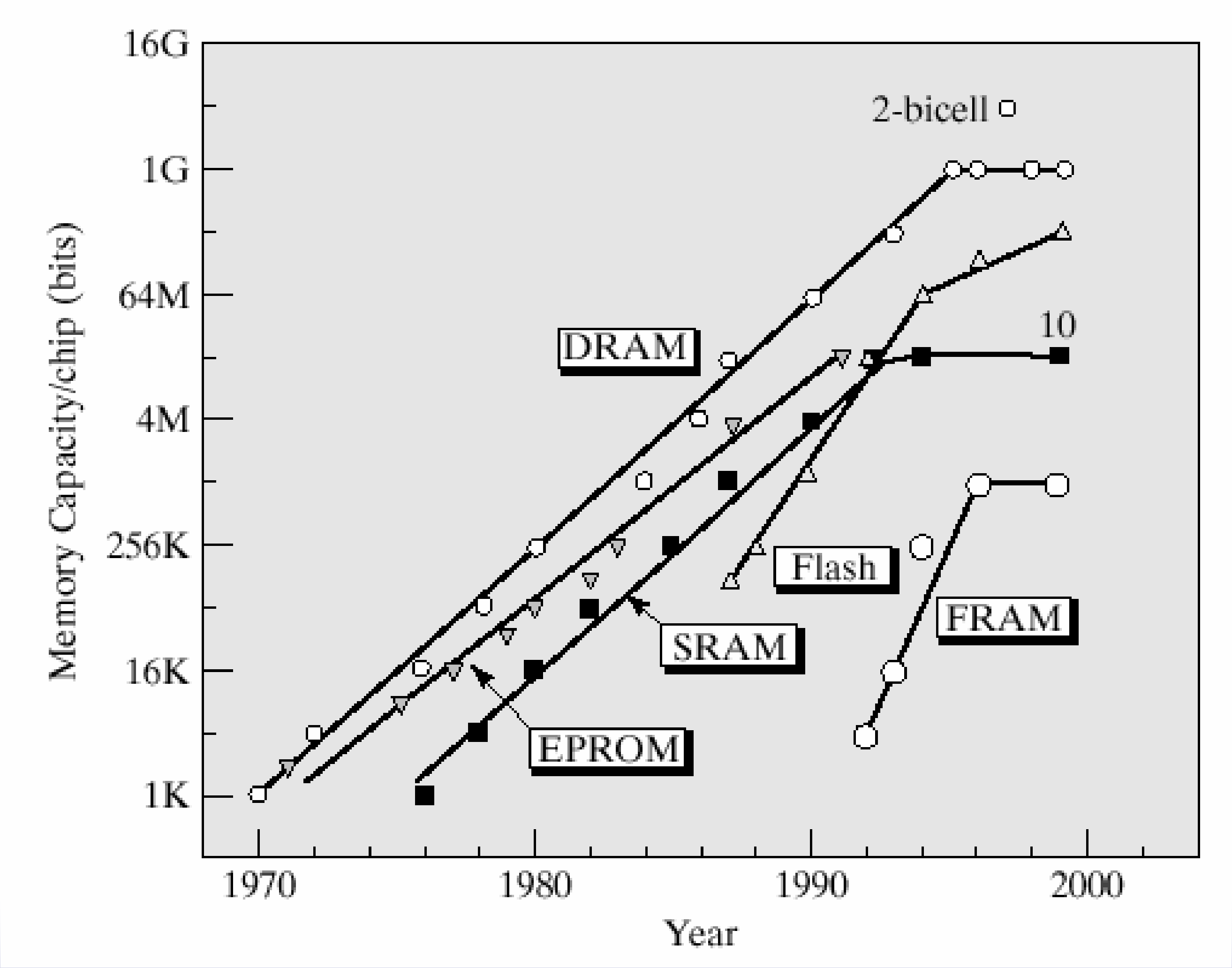
From [Nakamura02]

Semiconductor Memory Trends (up to the 90's)



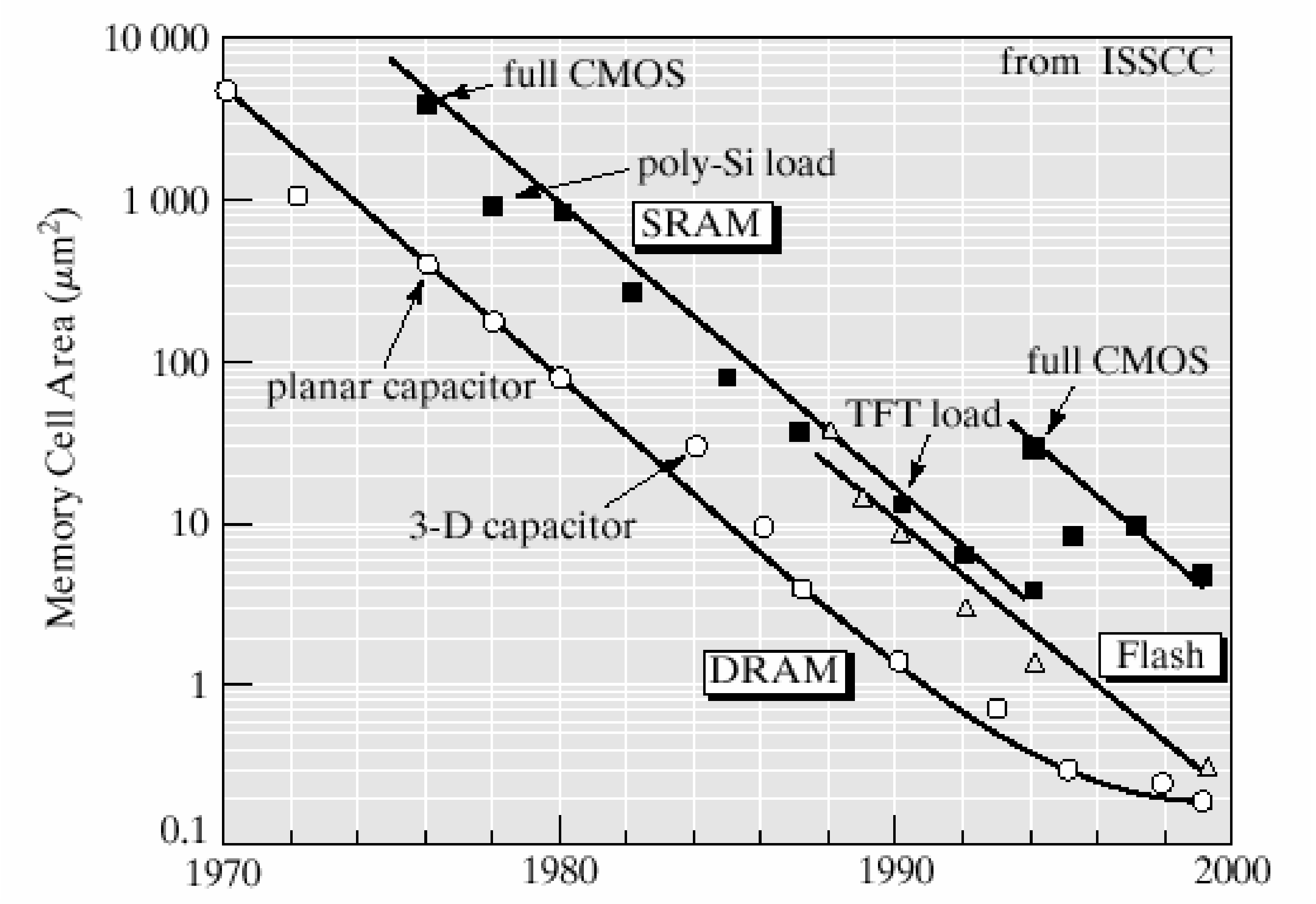
Memory Size as a function of time: x 4 every three years

Semiconductor Memory Trends (updated)



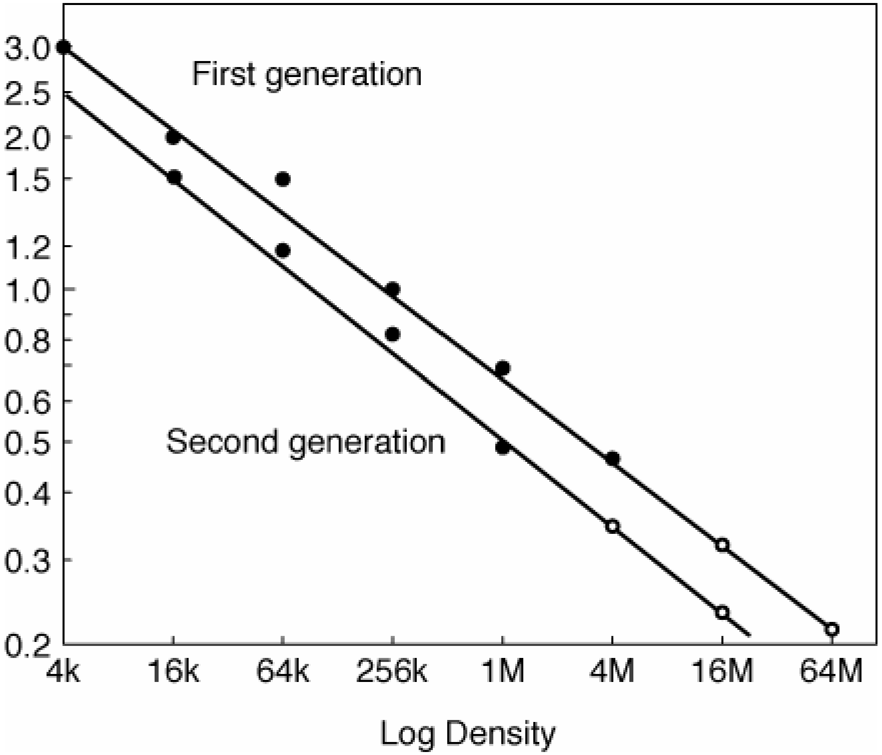
From [Itoh01]

Trends in Memory Cell Area



From [Itoh01]

Semiconductor Memory Trends



Technology feature size for different SRAM generations