CAL POLY
SAN LUIS OBISPO

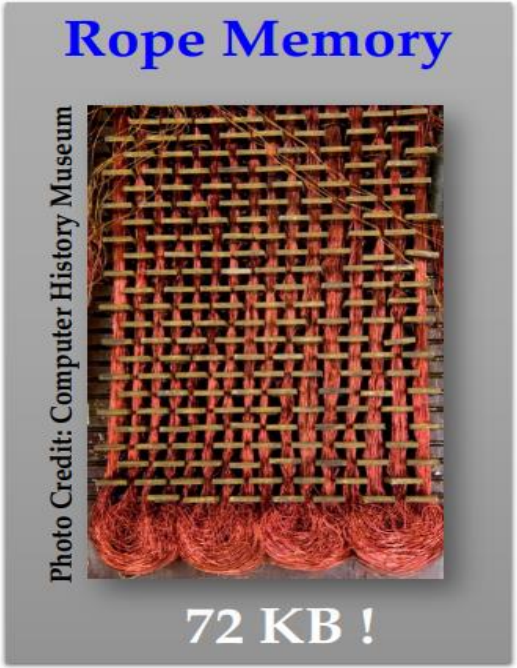
EE 431: COMPUTER-AIDED DESIGN OF VLSI DEVICES

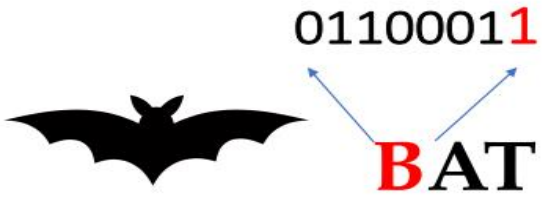
Memory Design

Nishith N. Chakraborty

November, 2024

DATA DRIVEN WORLD!

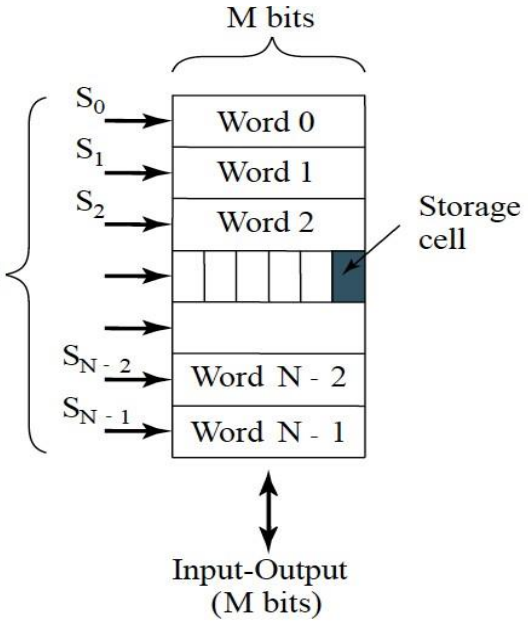




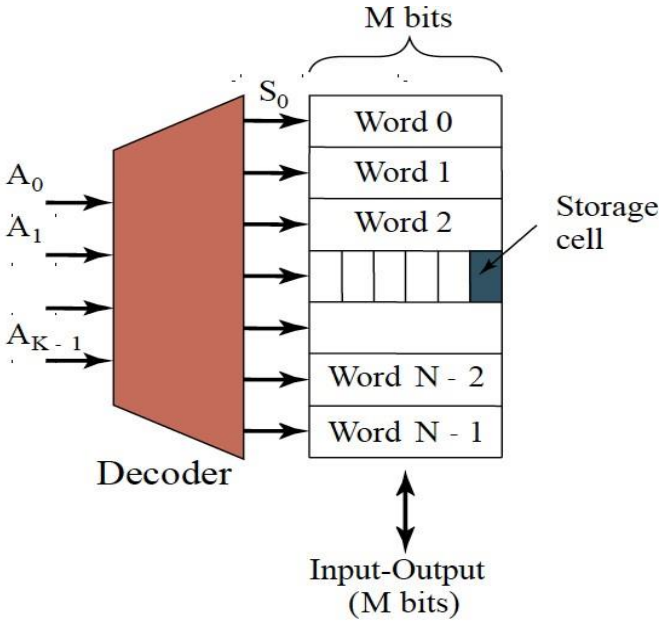
MEMORY CLASSIFICATION

<table>
<tr><td colspan="2">Read-Write Memory</td><td>Non-Volatile Read-Write Memory</td><td>Read-Only Memory</td></tr>
<tr><td>Random Access</td><td>Non-Random Access</td><td rowspan="2">EPROM
E^2PROM
FLASH</td><td rowspan="2">Mask-Programmed
Programmable (PROM)</td></tr>
<tr><td>SRAM
DRAM</td><td>FIFO
LIFO
Shift Register
CAM</td></tr>
</table>

MEMORY ARCHITECTURE: DECODERS



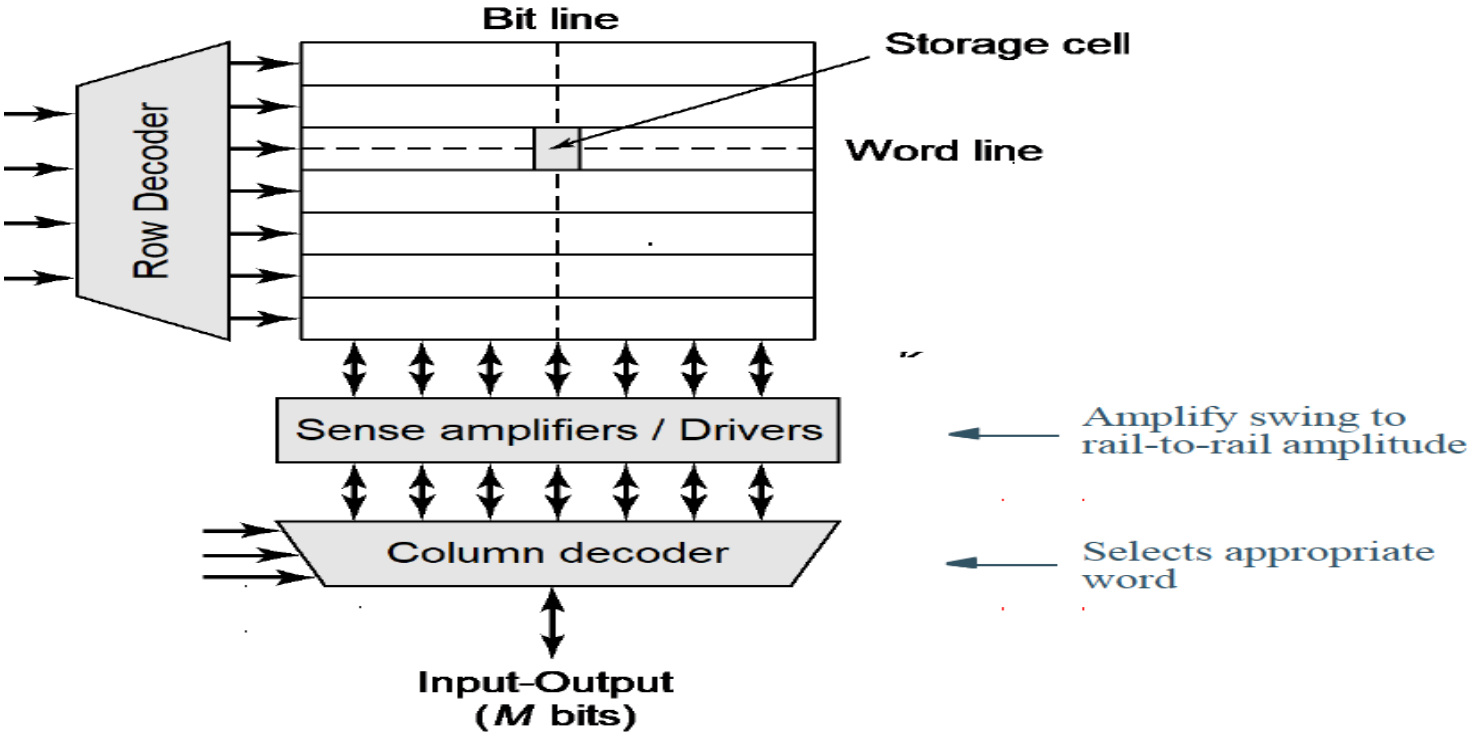
Intuitive architecture for N x M memory
Too many select signals:
N words == N select signals



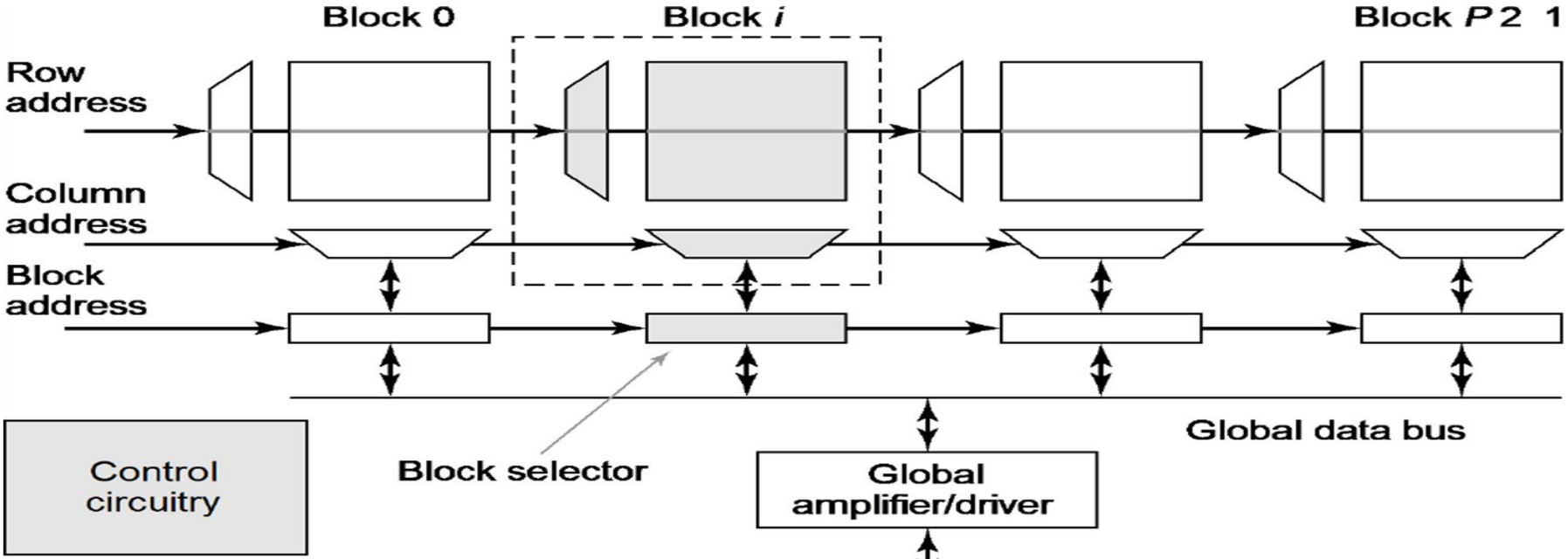
Decoder reduces the number of select signals

$K = \log_2 N$

ARRAY-STRUCTURED MEMORY ARCHITECTURE



HIERARCHICAL MEMORY ARCHITECTURE



Advantages:

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

READ-WRITE MEMORIES

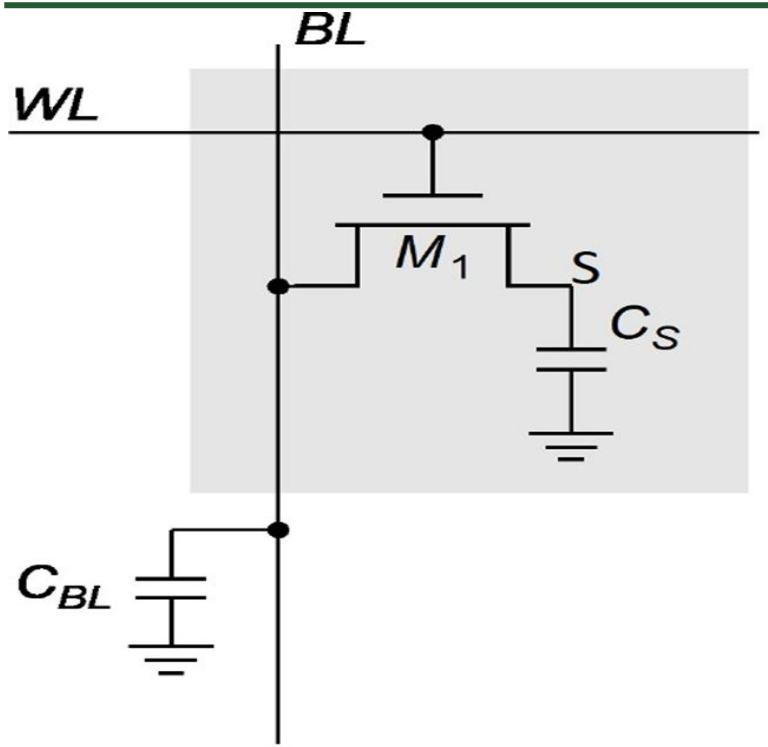
- STATIC (SRAM)

  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



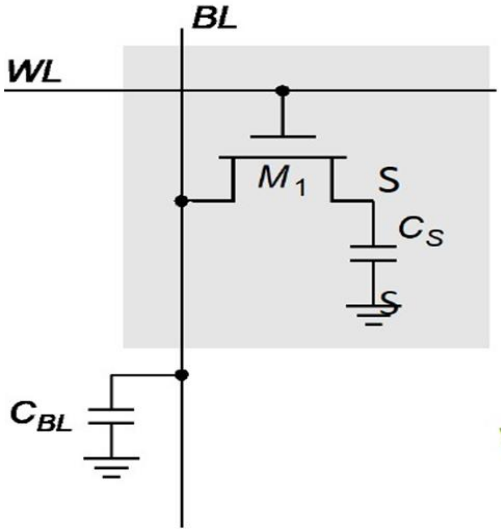
Read Operation

- BL precharged to $V_{DD}/2$
- WL asserted (M1 turned ON)
- BL charges if C_S stores '1'
- BL discharges if C_S stores '0'

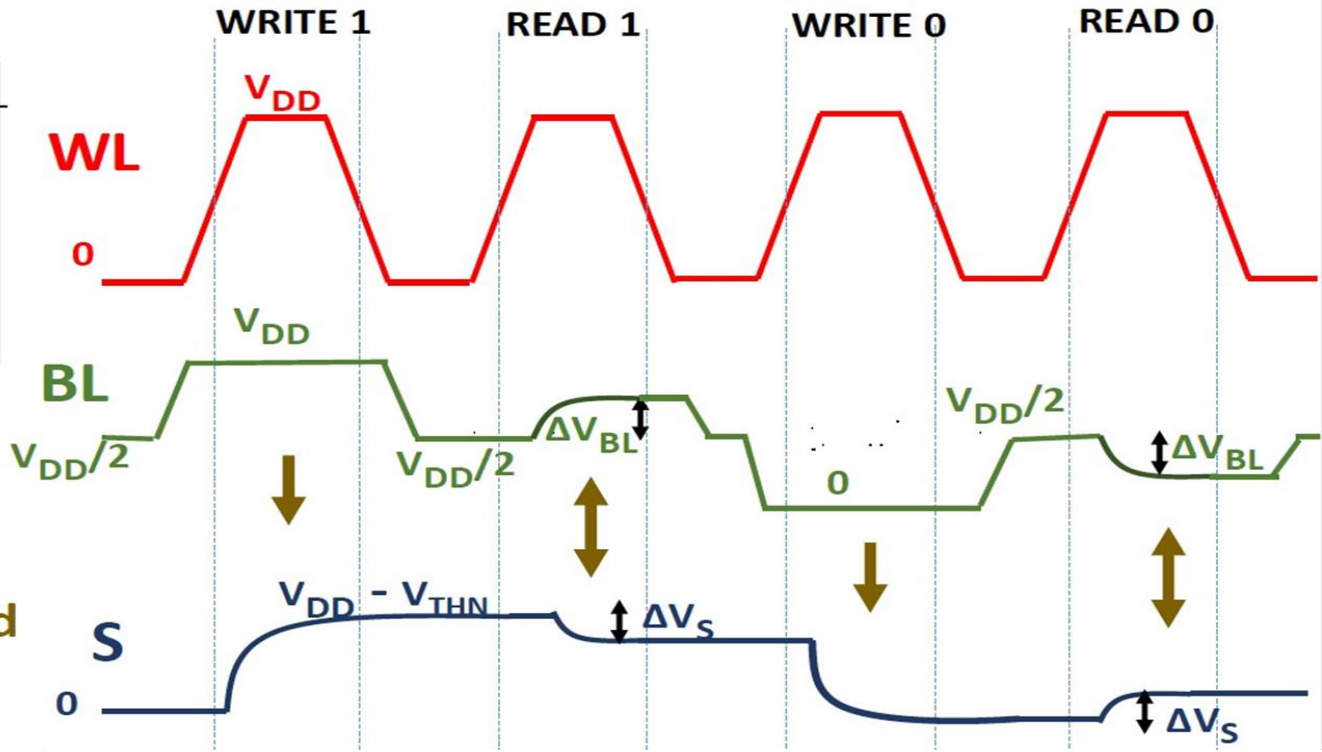
Write Operation

- BL driven to V_{DD} (for write 1) or 0 (for write 0)
- WL asserted
- C_S is charged to V_{DD}-V_{THN} or discharged to 0

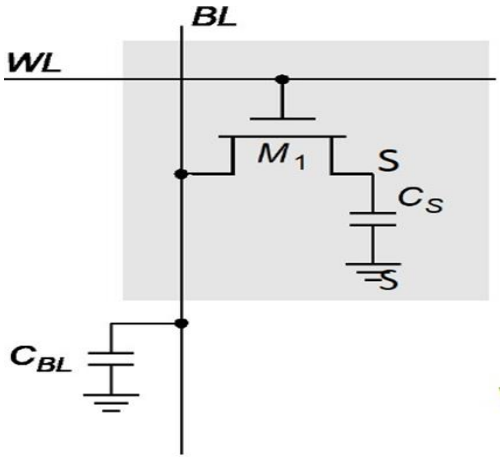
1-TRANSISTOR DRAM CELL



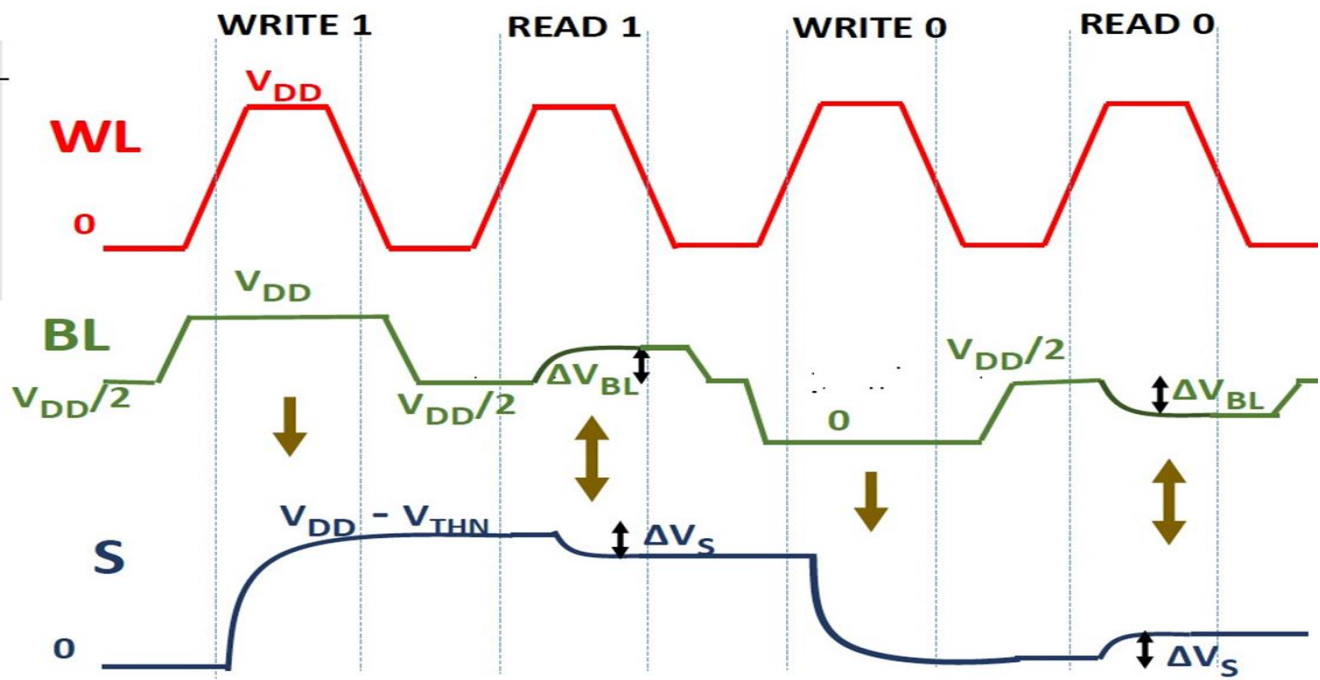
- Destructive Read
- Needs Refresh



CHARGE SHARING DURING READ



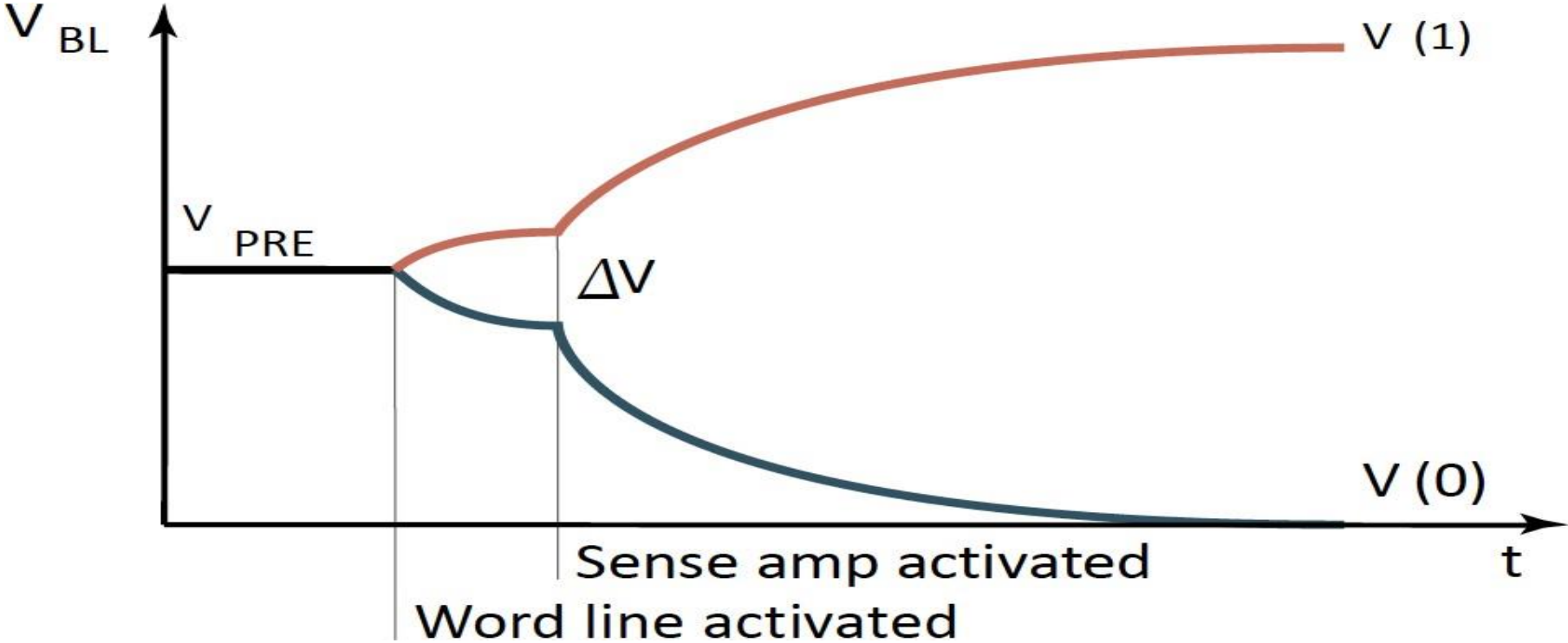
$V_{PRE} = V_{DD}/2$
$V_S = 0$ or $V_{DD}-V_{THN}$



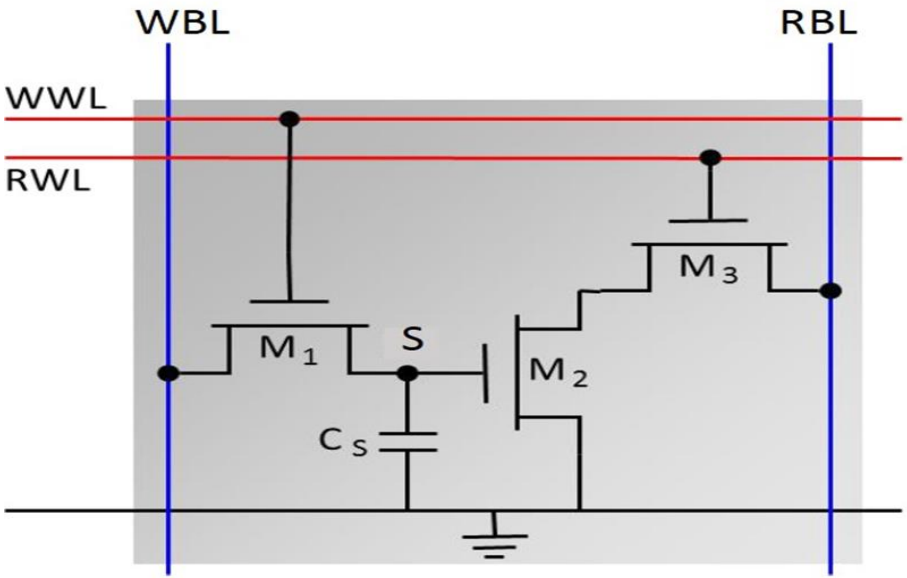
Initial Q = $C_S V_S + C_{BL} V_{PRE}$

Final Q = $(C_S + C_{BL})\, V_{FINAL}$

SENSE AMPLIFIER OPERATION



3-T DRAM CELL



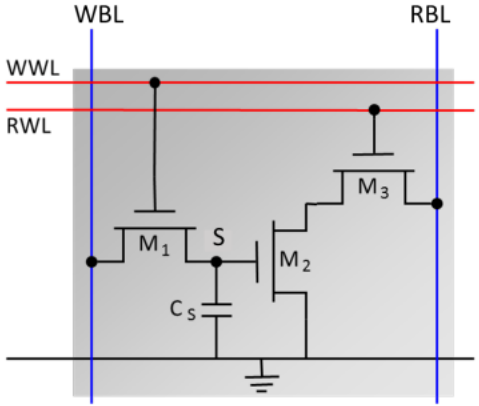
Read Operation

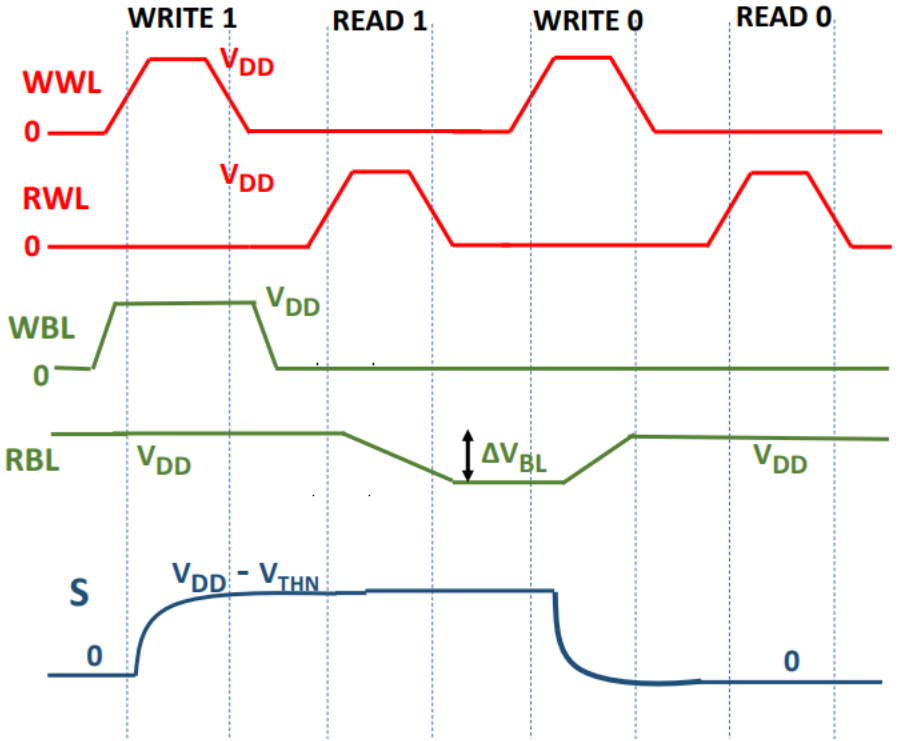
- RBL precharged to V_{DD}
- RWL asserted (M3 turned ON)
- BL remains at V_{DD} if C_S stores '0' (M2 is OFF)
- BL discharges if C_S stores '1' (M2 is ON)

Write Operation

- WBL driven to V_{DD} (for write 1) or 0 (for write 0)
- WWL asserted (M1 turned ON)
- C_S is charged to V_{DD}-V_{THN} or discharged to 0

3-T DRAM CELL

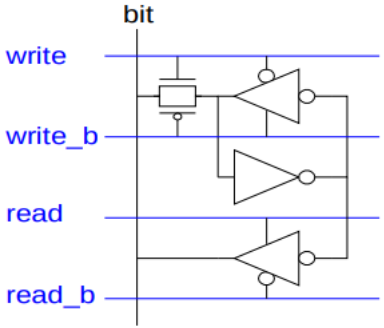


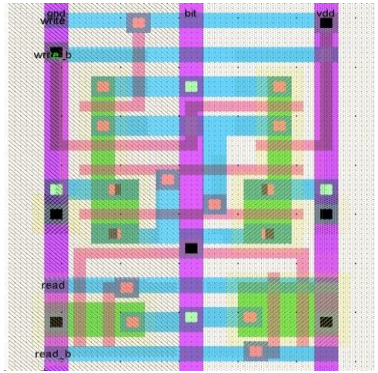


- Non-Destructive Read
- No Refresh
- Incomplete swing at S. (WL overdrive solves this issue)

12T SRAM CELL

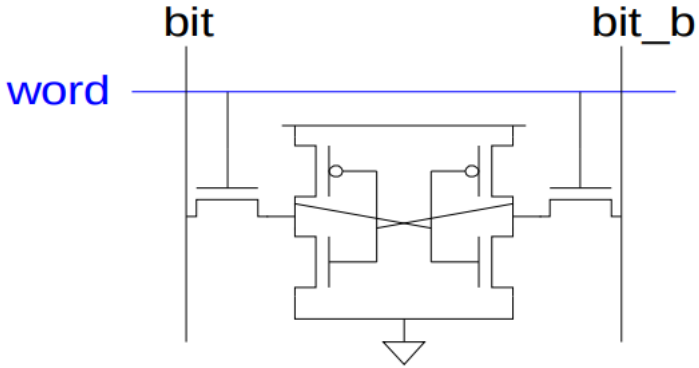
- Basic building block: SRAM Cell
  - Holds one bit of information, like a latch
  - Must be read and written
- 12-transistor (12T) SRAM cell
  - Use a simple latch connected to bitline
  - 46 x 75 λ unit cell





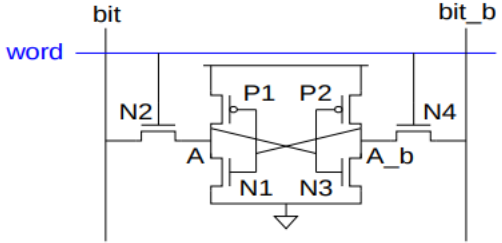
6T SRAM CELL

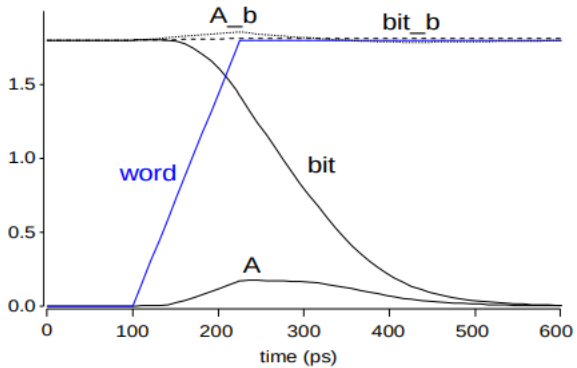
- Cell size accounts for most of array size
  - Reduce cell size at expense of complexity
- 6T SRAM Cell
  - Used in most commercial chips
  - Data stored in cross-coupled inverters
- Read:
  - Precharge bit, bit_b
  - Raise wordline
- Write:
  - Drive data onto bit, bit_b
  - Raise wordline



SRAM READ

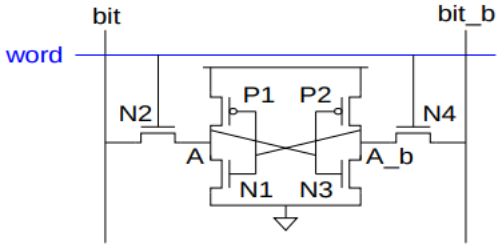
- Precharge both bitlines high
- Then turn on wordline
- One of the two bitlines will be pulled down by the cell
- Ex: A = 0, A_b = 1
  - bit discharges, bit_b stays high
  - But A bumps up slightly
- Read stability
  - A must not flip
  - N1 >> N2

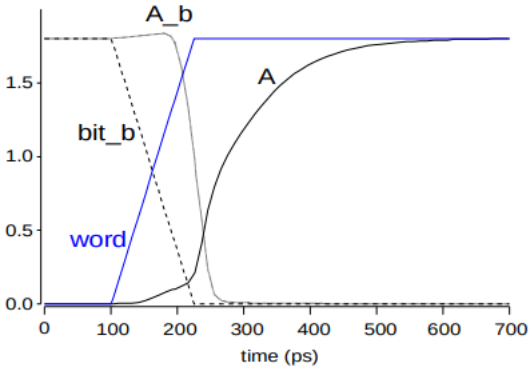




SRAM WRITE

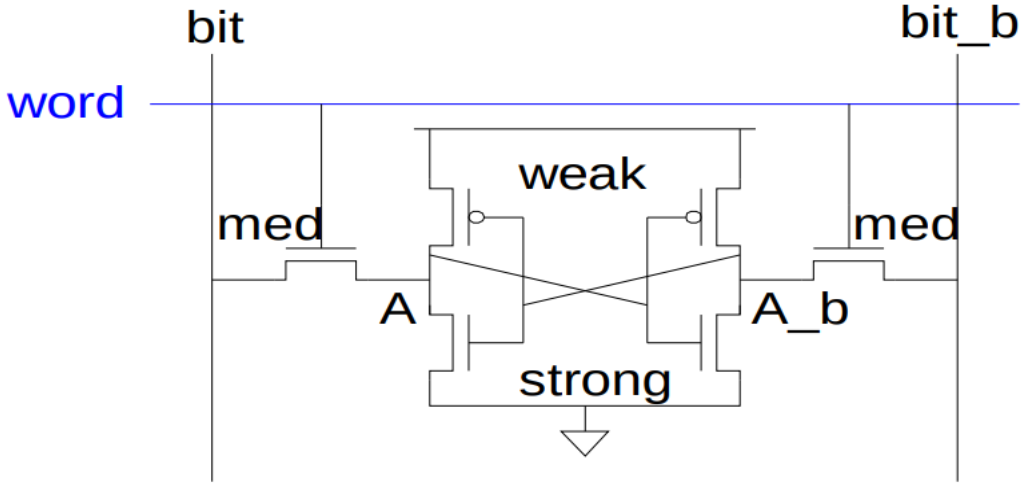
- Drive one bitline high, the other low
- Then turn on wordline
- Bitlines overpower cell with new value
- Ex: A = 0, A_b = 1, bit = 1, bit_b = 0
  - Force A_b low, then A rises high
- Writability
  - Must overpower feedback inverter
  - N2 >> P1





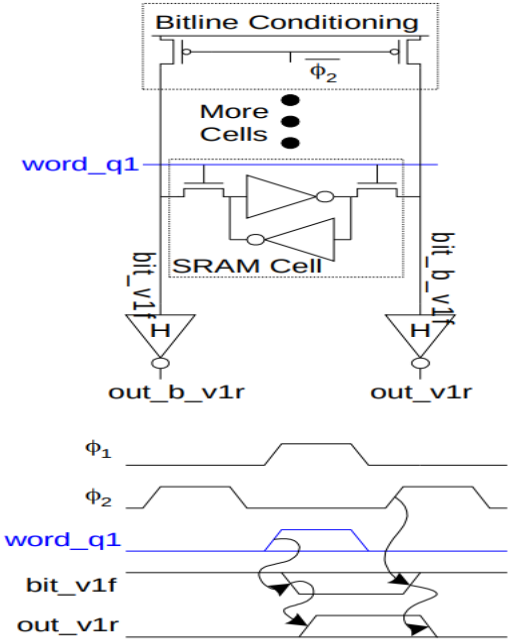
SRAM SIZING

- High bitlines must not overpower inverters during reads
- But low bitlines must write new value into cell

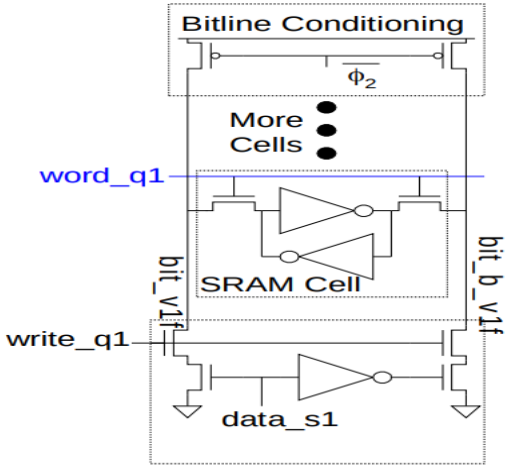


SRAM COLUMN EXAMPLE

Read

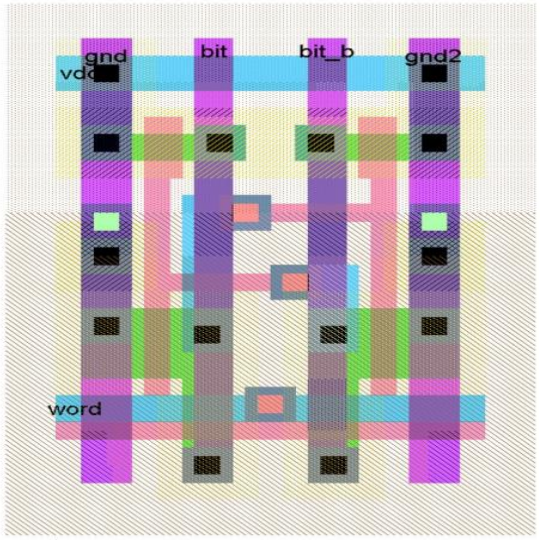


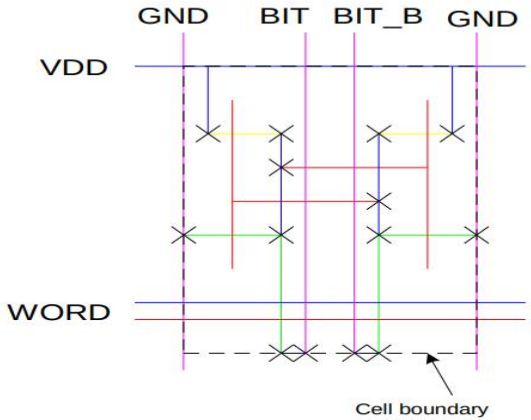
Write

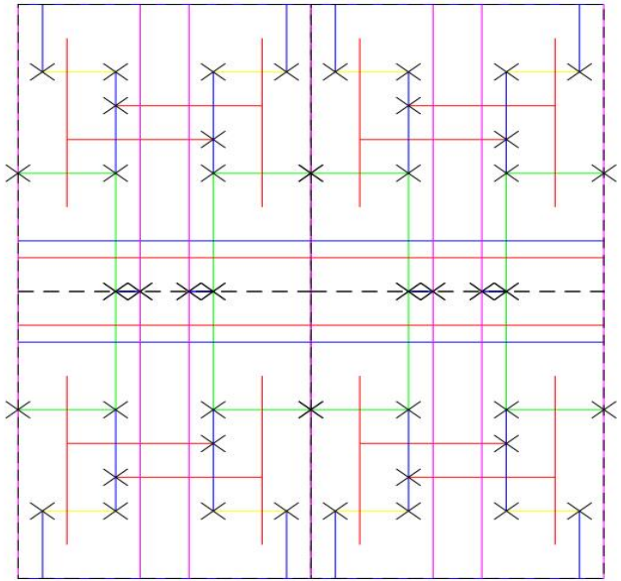


SRAM LAYOUT

- Cell size is critical: 26 x 45 λ (even smaller in industry)
- Tile cells sharing VDD, GND, bitline contacts



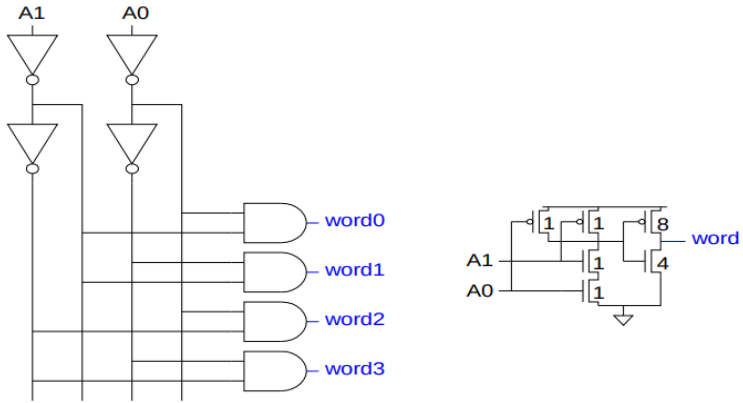




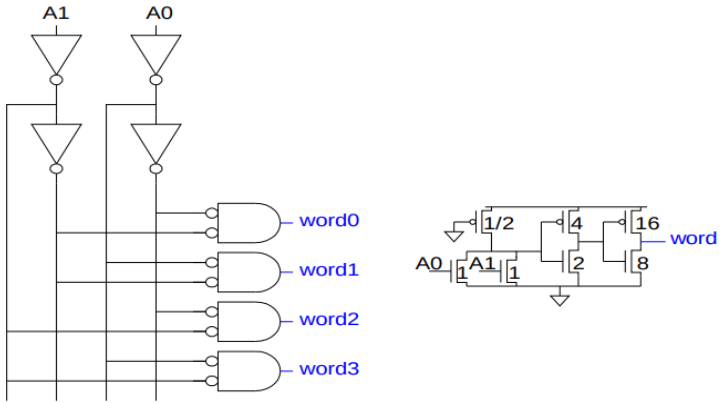
DECODERS

- n:2^n decoder consists of 2^n n-input AND gates
  - One needed for each row of memory
  - Build AND from NAND or NOR gates

Static CMOS

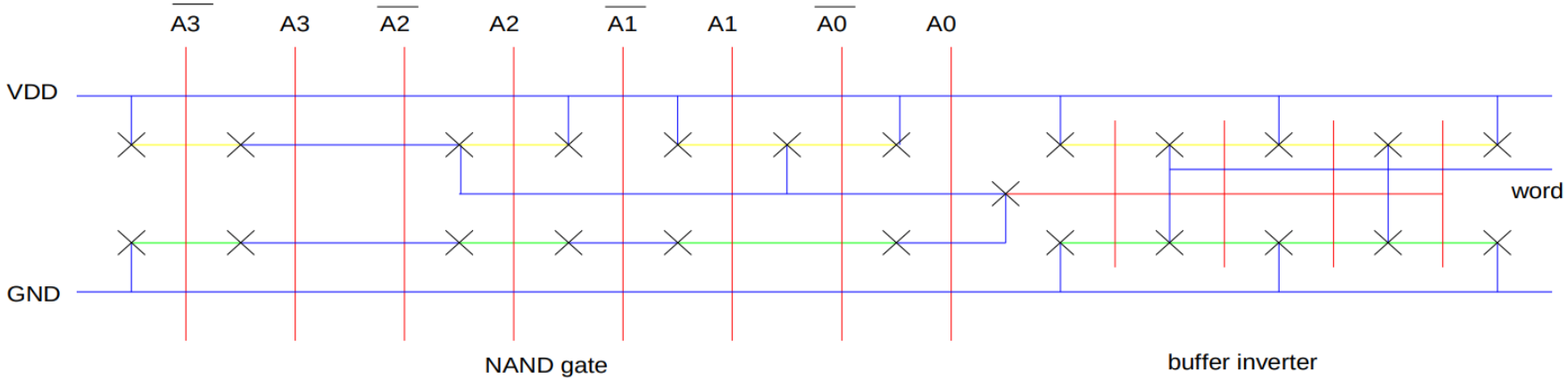


Pseudo-nMOS



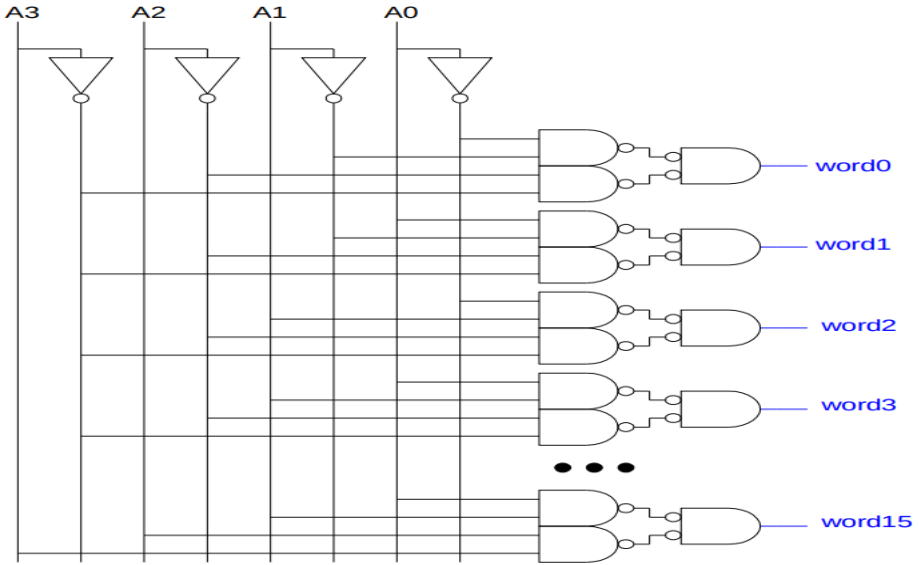
DECODER LAYOUT

- Decoders must be pitch-matched to SRAM cell
  - Requires very skinny gates



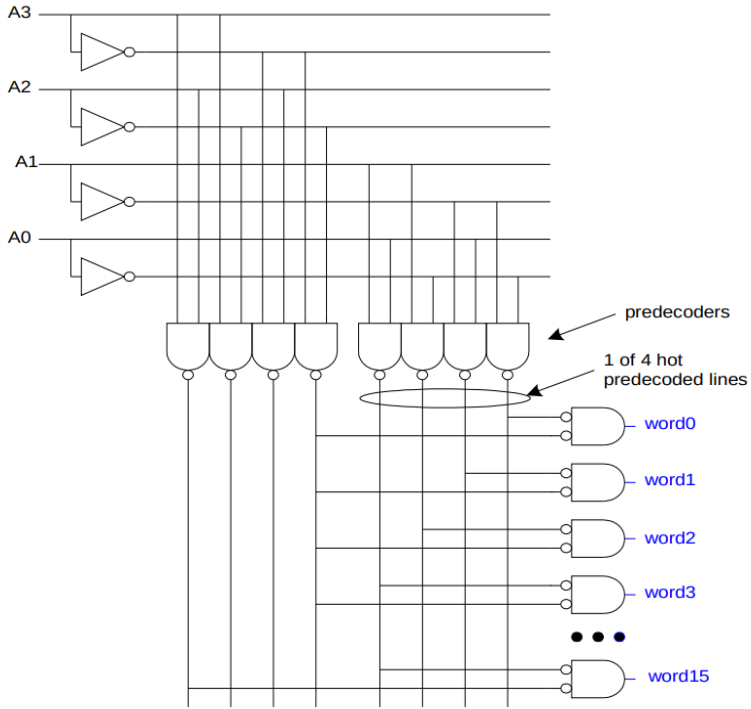
LARGE DECODERS

- For n > 4, NAND gates become slow
  - Break large gates into multiple smaller gates



PREDECODING

- Many of these gates are redundant
  - Factor out common gates into predecoder
  - Saves area
  - Same path effort

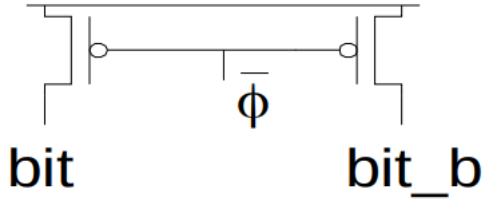


COLUMN CIRCUITRY

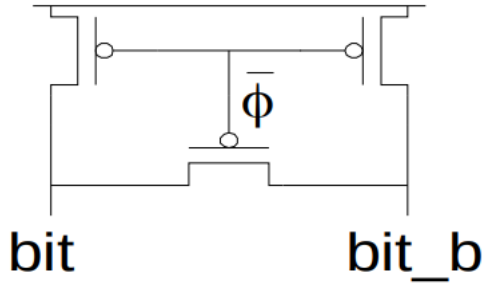
- Some circuitry is required for each column
  - Bitline conditioning
  - Sense amplifiers
  - Column multiplexing

BITLINE CONDITIONING

- Precharge bitlines high before reads



- Equalize bitlines to minimize voltage difference when using sense amplifiers

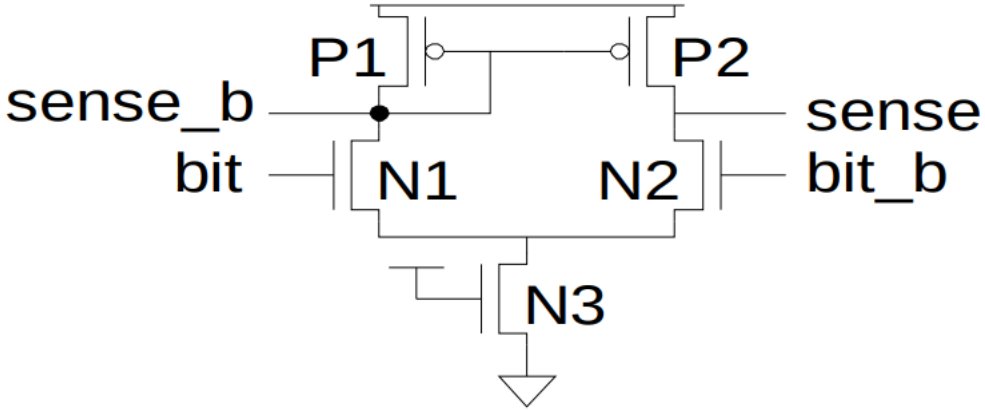


SENSE AMPLIFIERS

- Bitlines have many cells attached
  - Ex: 32-kbit SRAM has 256 rows x 128 cols
  - 128 cells on each bitline
- $t_{pd} \approx (C/I)\ \Delta V$
  - 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)

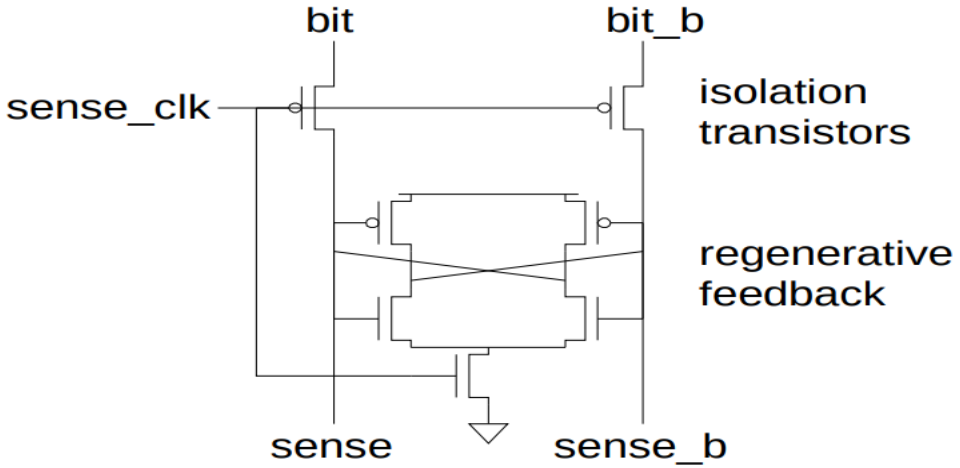
DIFFERENTIAL PAIR AMP

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



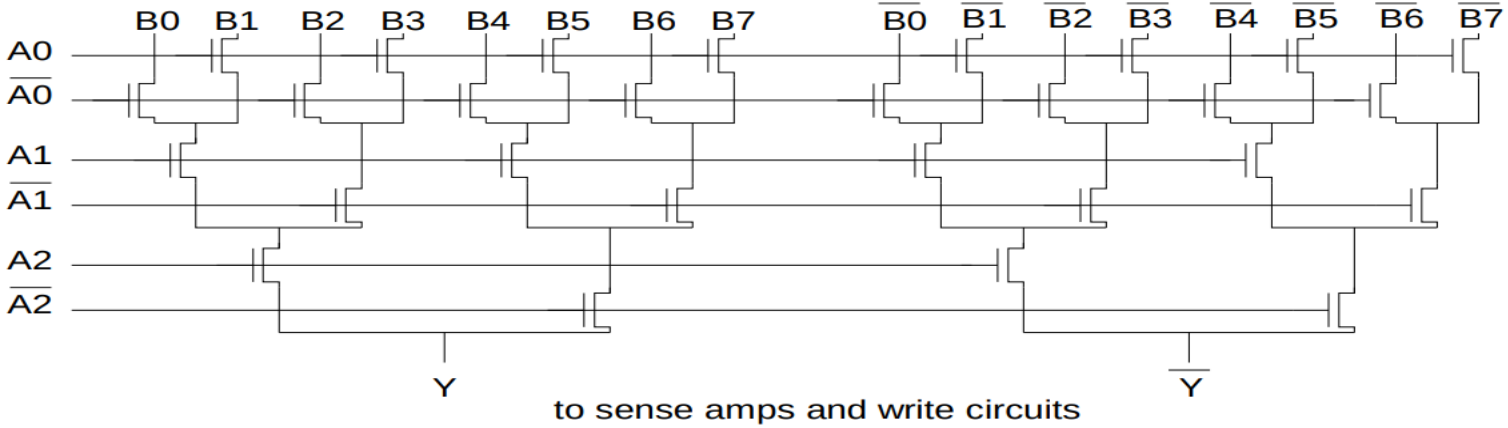
CLOCKED SENSE AMP

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



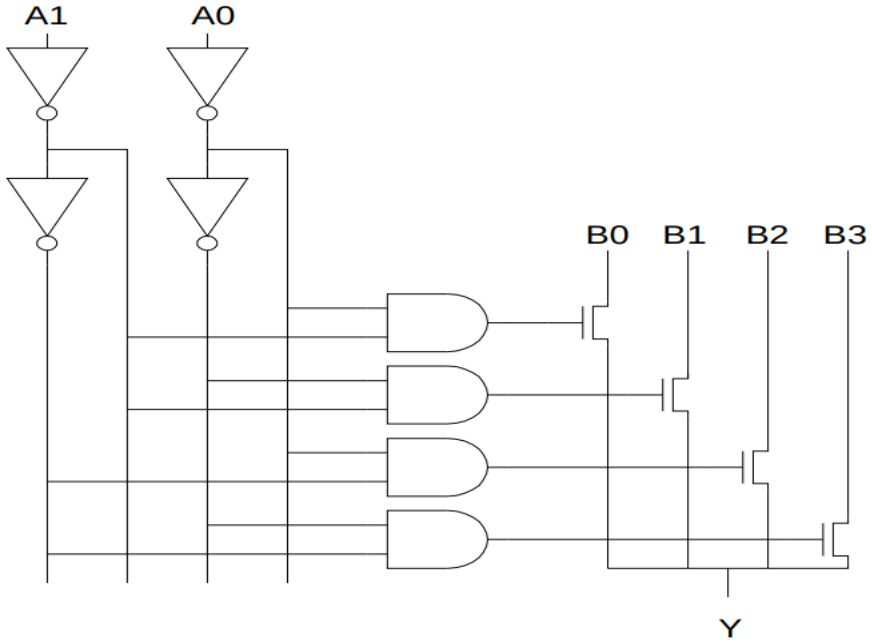
COLUMM MULTIPLEXING: TREE DECODER MUX

- 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

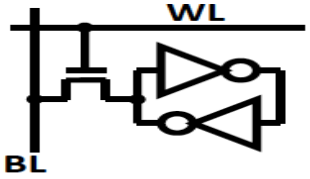


SINGLE PASS-GATE MUX

- Or eliminate series transistors with separate decoder

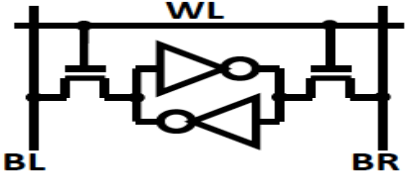


OTHER SRAM BIT-CELLS

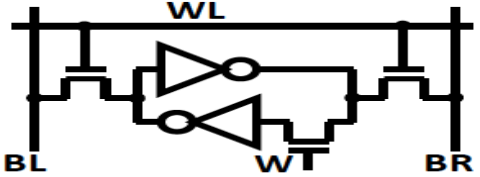


5T

I. Calson et. al., ESSCIRC05

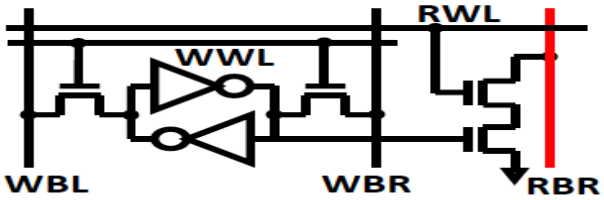


6T



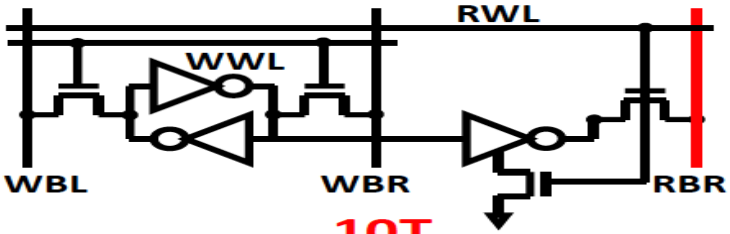
7T

K. Takeda et. al., ISSCC05



8T (Register File)

L. Chang et. al., VLSI Tech.'05



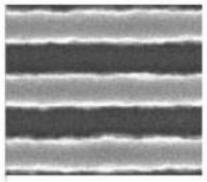
10T

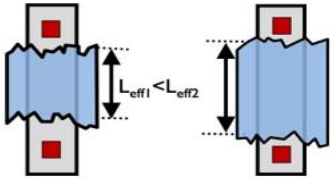
B. Calhoun et. al., ISSCC06

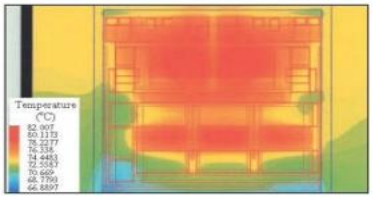
- 5T, 8T, 10T cells - single ended.
- 8T/10T decoupled read and write operation.
- No in-built process variation tolerance

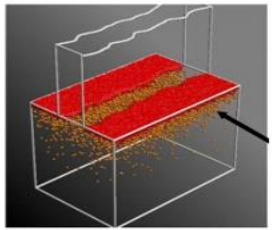
VARIABILITY: SOURCES & AN EXAMPLE

- Line Edge Roughness
- Random Dopant Fluctuations
- Non-uniform Temperature

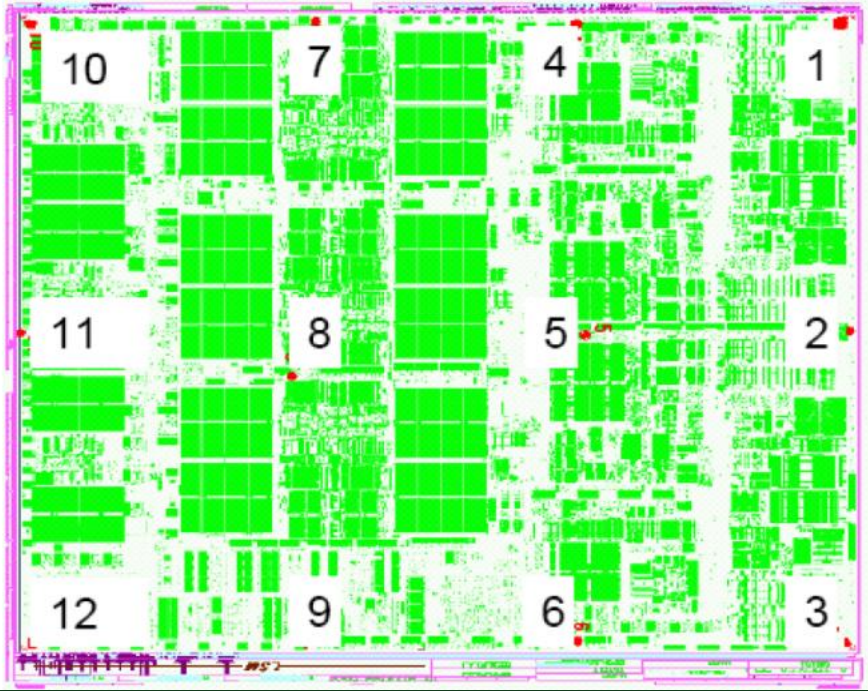




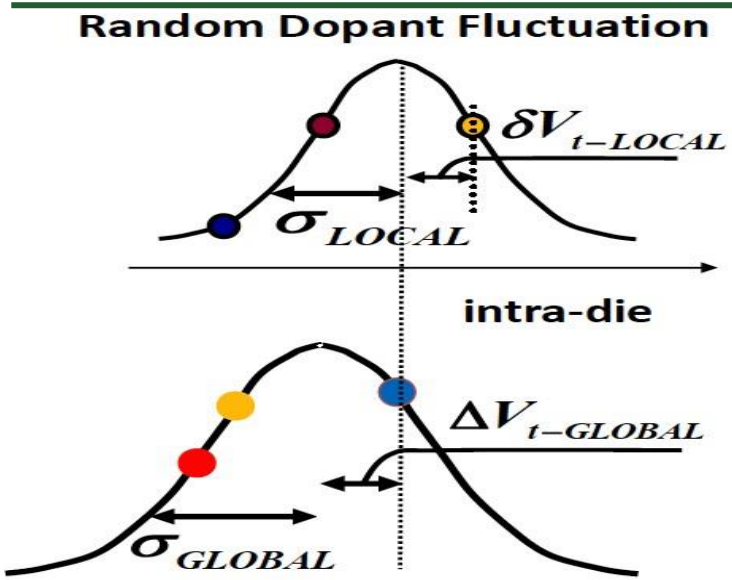




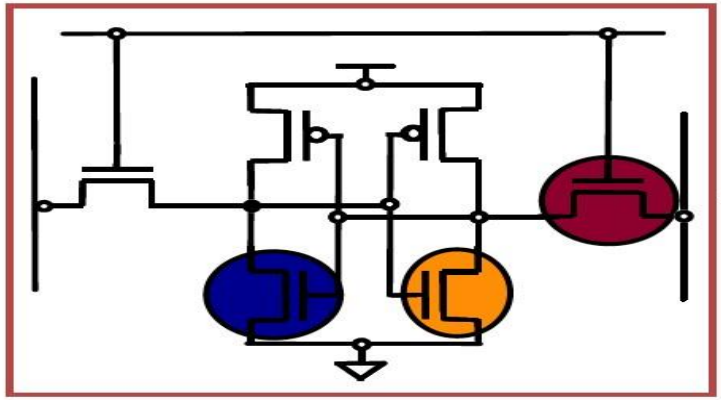
12 identical ring oscillators placed across 250 mm^2 chip

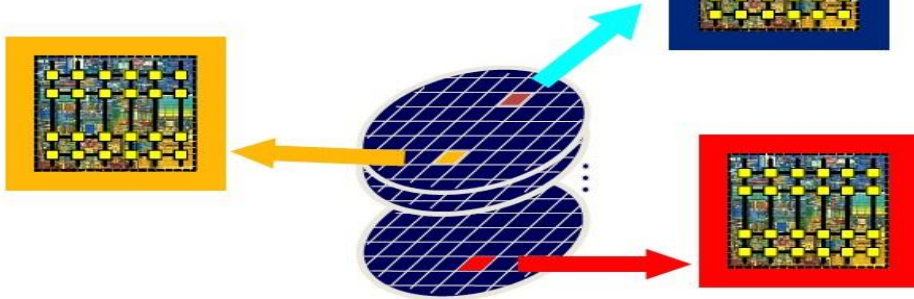


IMPACT OF PARAMETER VARIATION

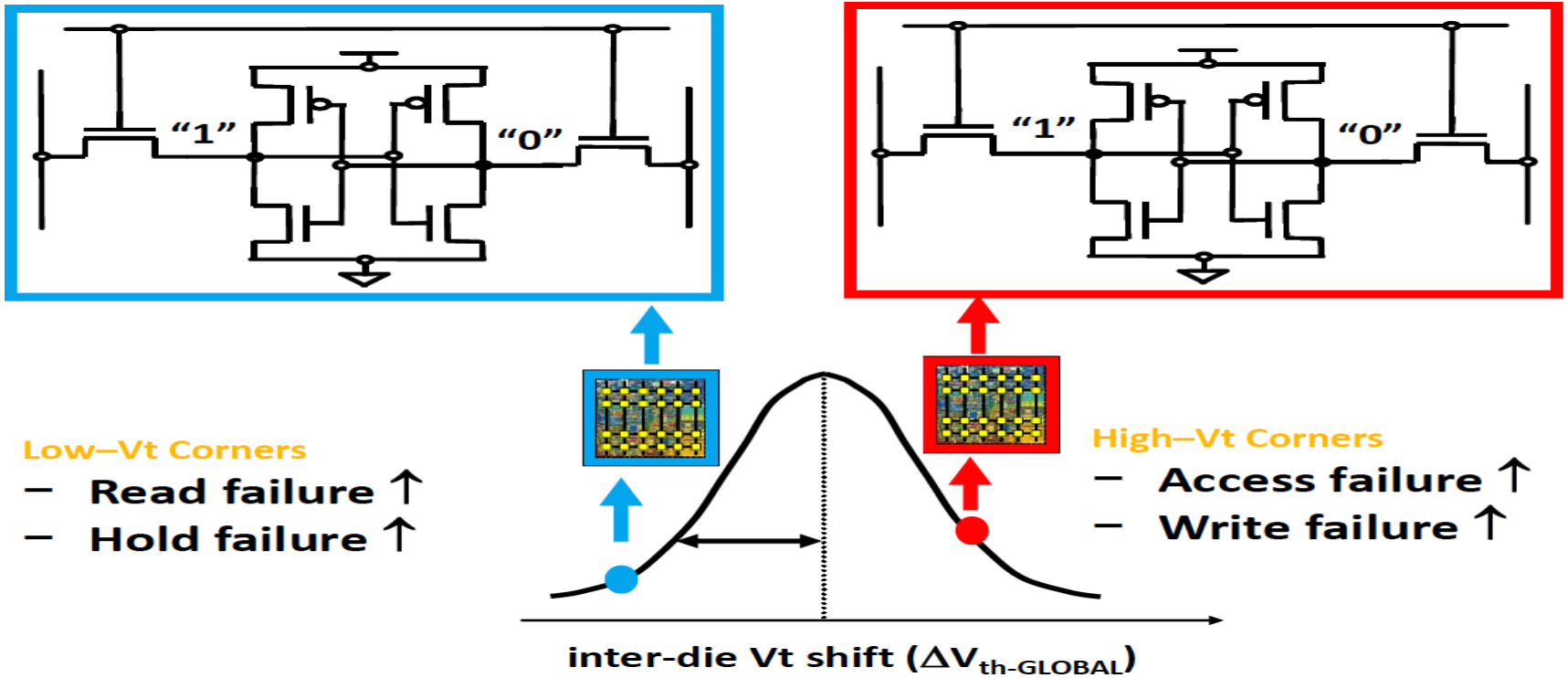


$$\delta V_t = \Delta V_{t-GLOBAL} + \delta V_{t-LOCAL}$$

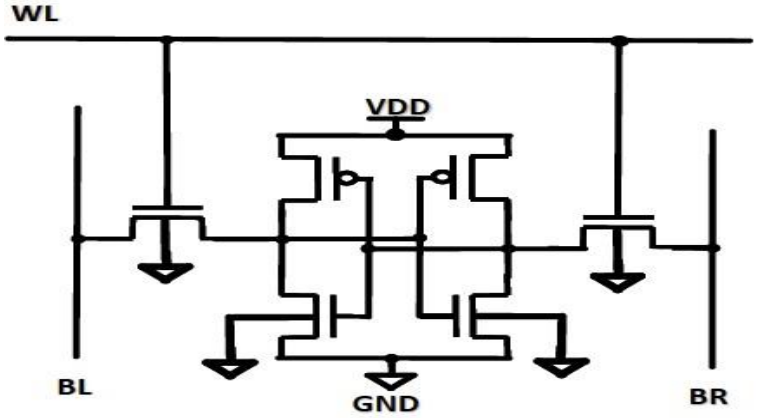


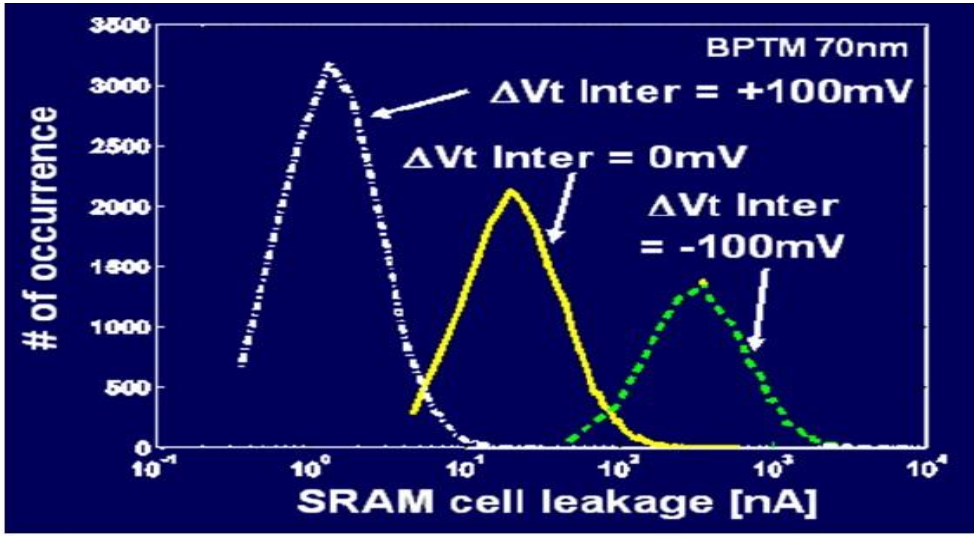


INTER-DIE VARIATION & CELL FAILURES



IDENTIFYING THE VT CORNERS

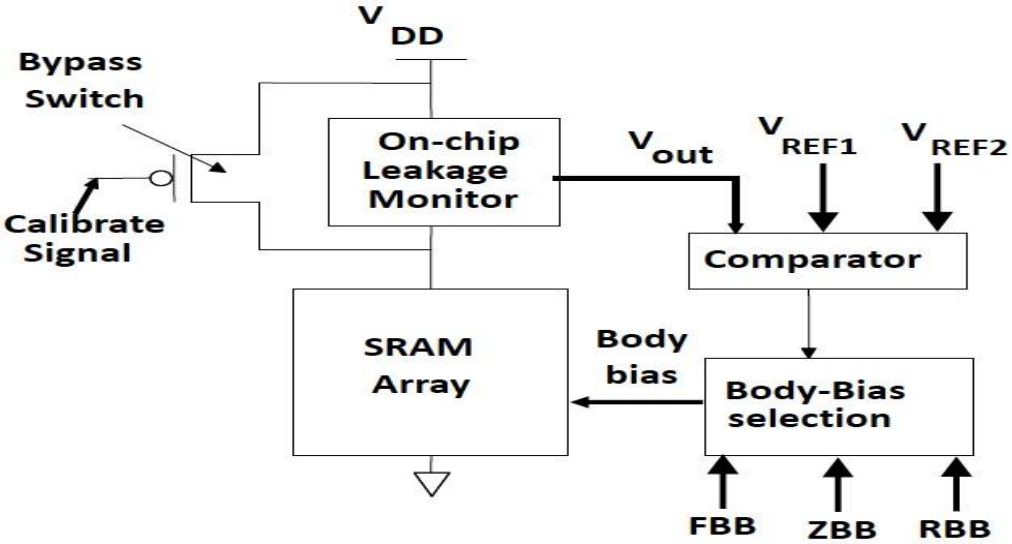




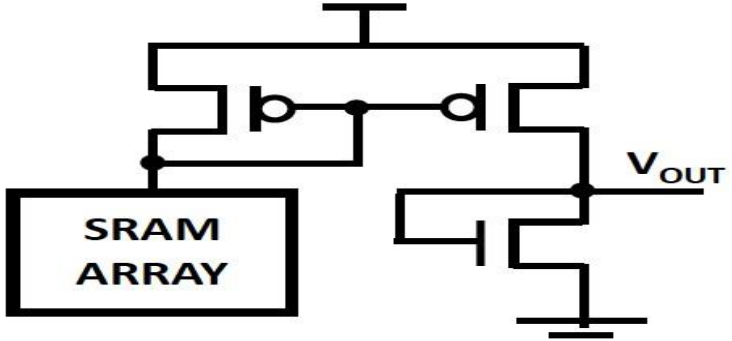
Monitor circuit parameters, e.g. leakage current

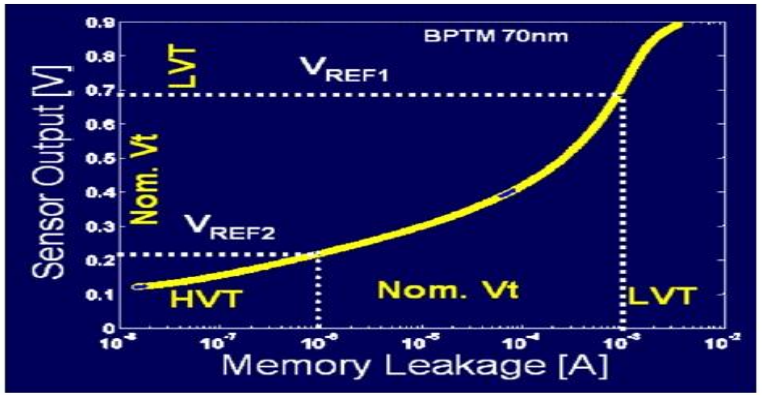
Effect of inter-die variation can be masked by intra-die variation

SELF-REPAIR TECHNIQUE

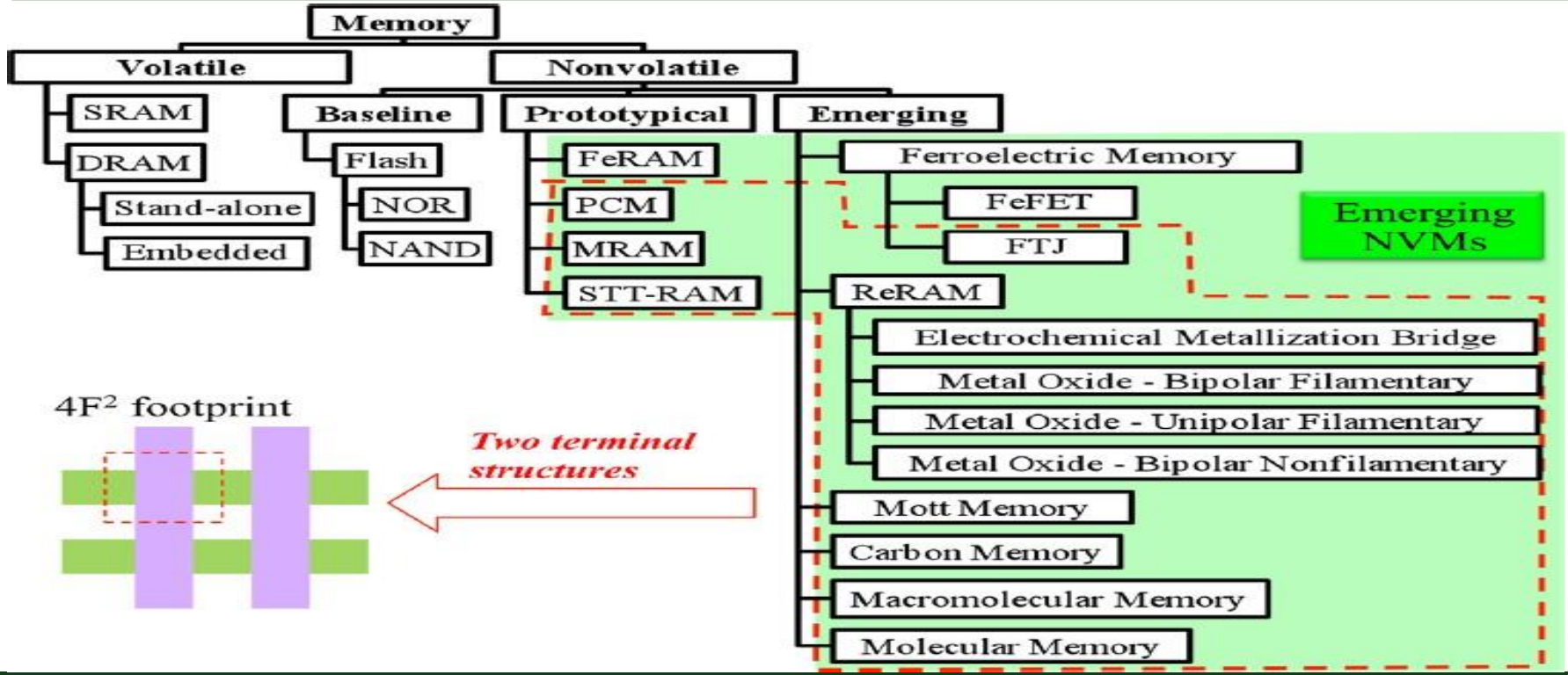


Entire array leakage is monitored to detect inter-die corner and proper body-bias is selected

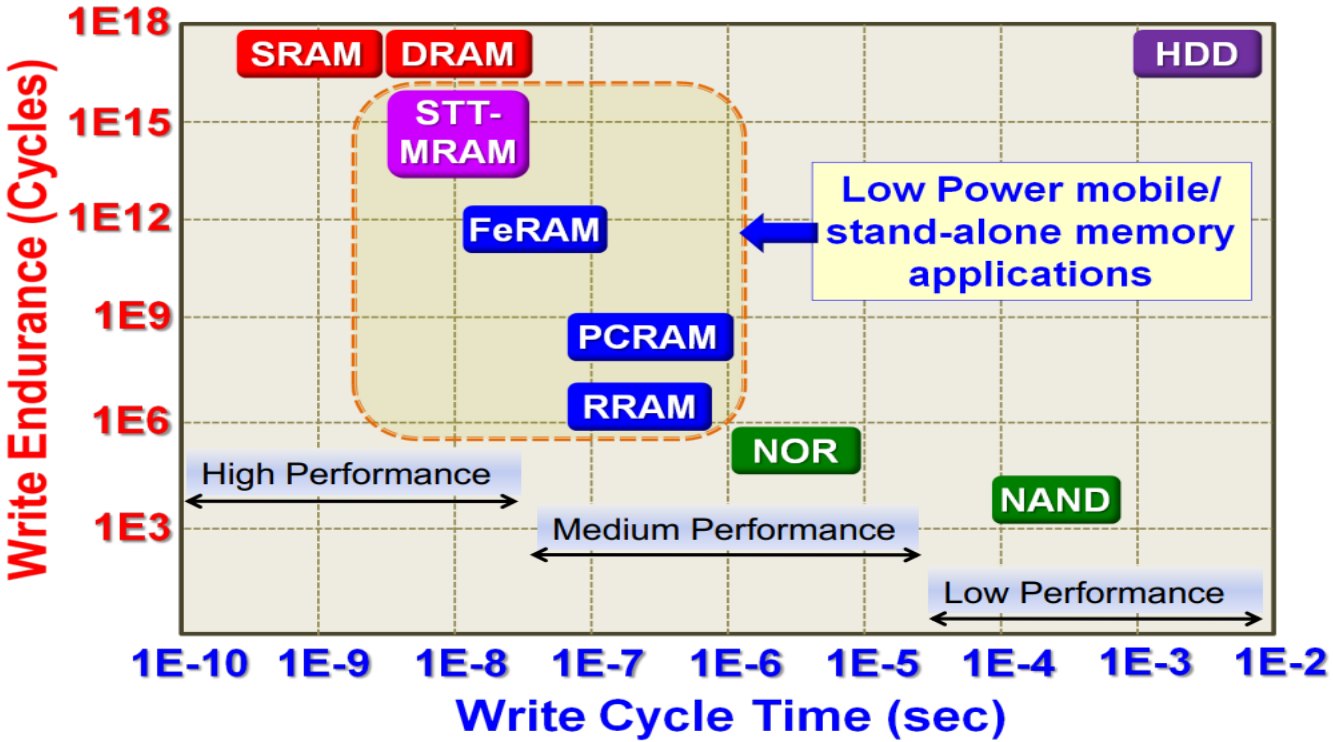




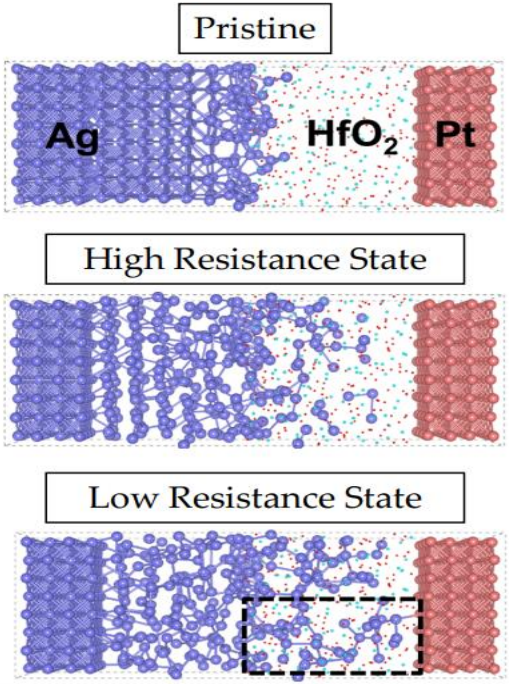
NON-VOLATILE MEMORY TECHNOLOGIES



NON-VOLATILE MEMORY TECHNOLOGIES



RESISTIVE RAM (RRAM)



N. Shukla, R. K. Ghosh, B. Grisafe and S. Datta, "Fundamental mechanism behind volatile and non-volatile switching in metallic conducting bridge RAM," *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017.

RESISTIVE RAM (RRAM): EVALUATION

Positive Features

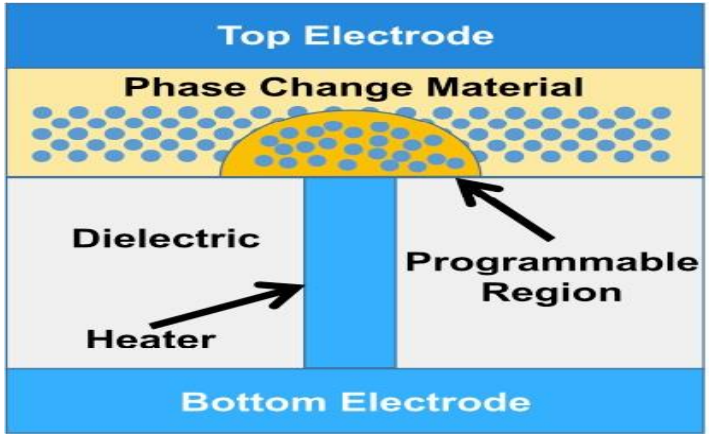
- High ON-OFF ratio
- Highly scalable
- CMOS compatibility

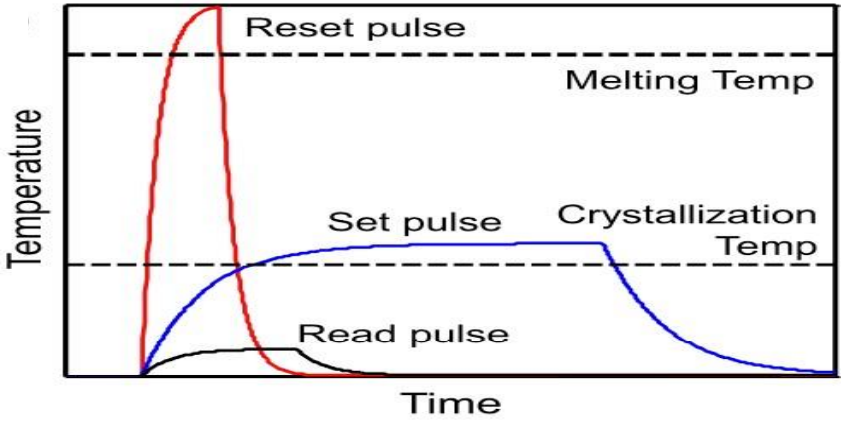
Limitations

- Limited endurance
- Variations
- Low retention time (in case of thin filament)

S. Yu and P. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," in *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, Spring 2016.

PHASE CHANGE RAM (PCRAM)





- Usually made of ‘GeSbTe’, *or in short* GST
- Changes the crystal structure from amorphous to crystalline (and vice versa) to switch resistive states

S. Yu and P. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," in *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, Spring 2016.

PHASE CHANGE RAM (PCRAM): EVALUATION

Positive Features

- High ON-OFF ratio
- High retention
- Multilevel possible
- CMOS compatible

Limitations

- High switching energy
- Slow transitions
- Resistance drift → Low retention
  (due to relaxation of amorphous phase)

S. Yu and P. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," in *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, Spring 2016.

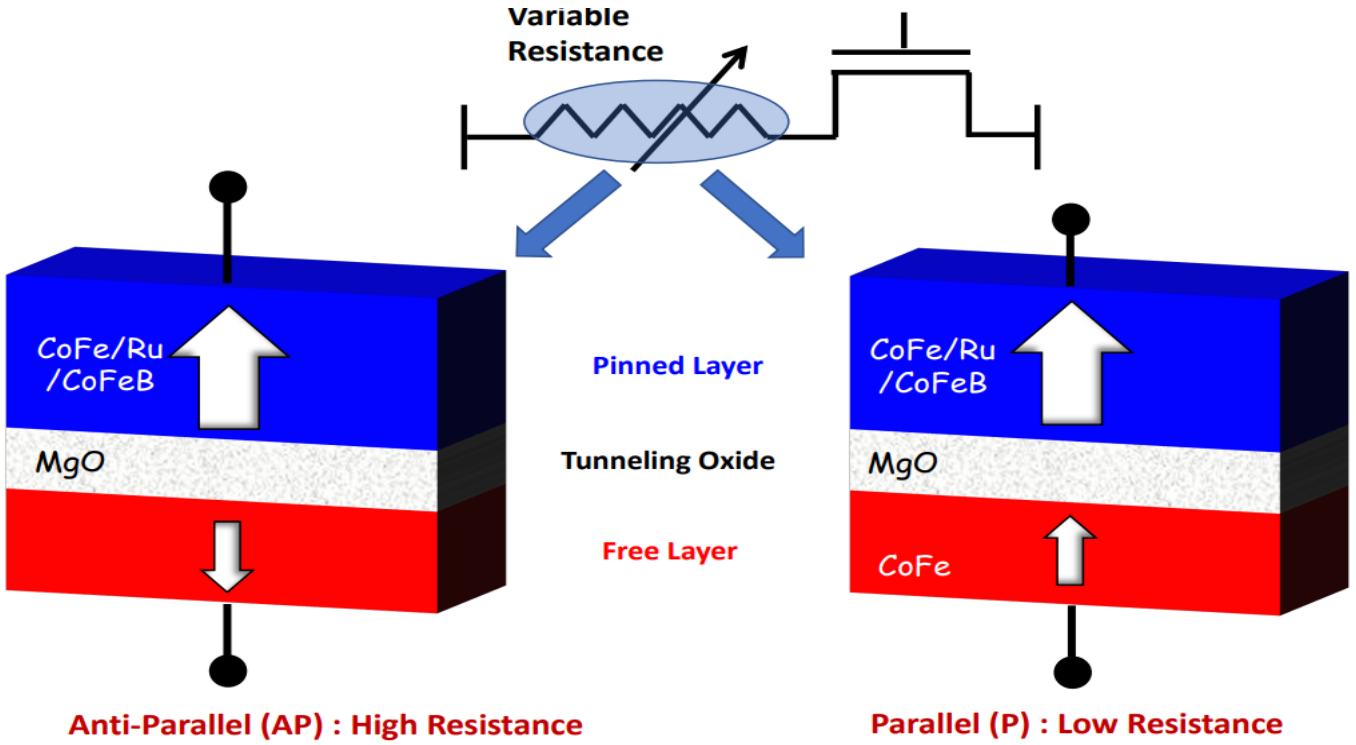
MEMORY BENCHMARK

	MAINSTREAM MEMORIES				EMERGING MEMORIES		
			FLASH				
	SRAM	DRAM	NOR	NAND	STT-MRAM	PCRAM	RRAM
Cell area	>100 F^2	6 F^2	10 F^2	<4F^2 (3D)	6~50F^2	4~30F^2	4~12F^2
Multibit	1	1	2	3	1	2	2
Voltage	<1 V	<1 V	>10 V	>10 V	<1.5 V	<3 V	<3 V
Read time	~1 ns	~10 ns	~50 ns	~10 µs	<10 ns	<10 ns	<10 ns
Write time	~1 ns	~10 ns	10 µs–1 ms	100 µs–1 ms	<10 ns	~50 ns	<10 ns
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	>10 y
Endurance	>1E16	>1E16	>1E5	>1E4	>1E15	>1E9	>1E6~1E12
Write energy (J/bit)	~fJ	~10fJ	~100pJ	~10fJ	~0.1pJ	~10pJ	~0.1 pJ

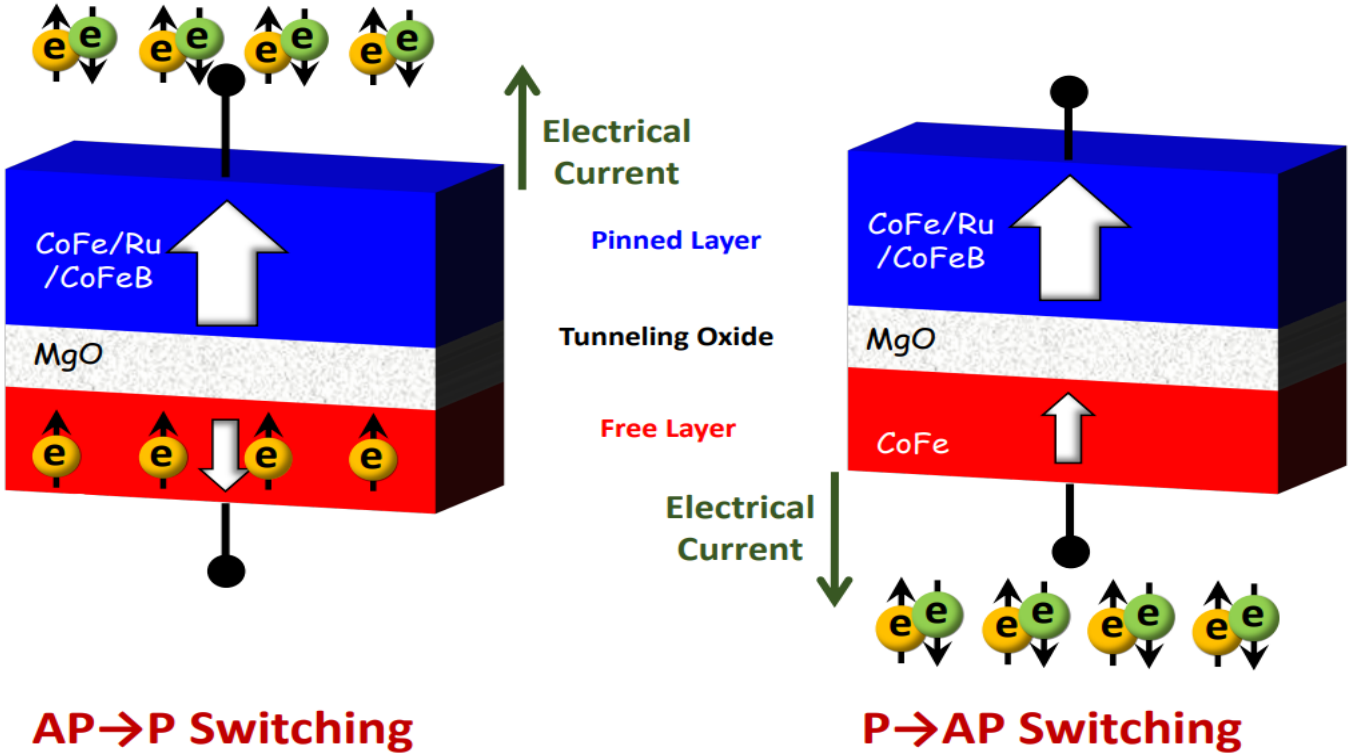
Notes: F: feature size of the lithography. The energy estimation is on the cell-level (not on the array-level). PCRAM and RRAM can achieve less than 4F^2 through 3D integration. The numbers of this table are representative (not the best or the worst cases).

S. Yu and P. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," in *IEEE Solid-State Circuits Magazine*, vol. 8, no. 2, pp. 43-56, Spring 2016.

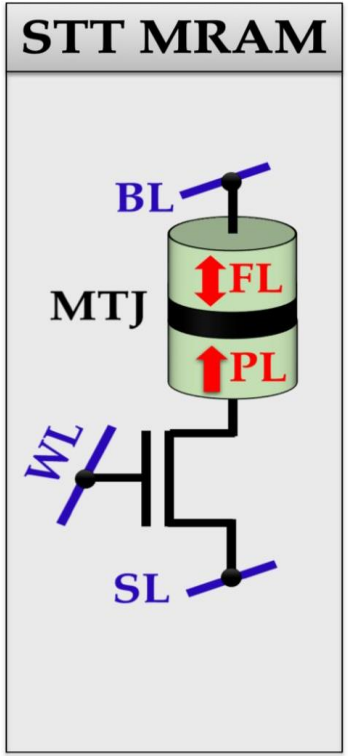
STANDARD STT MRAM

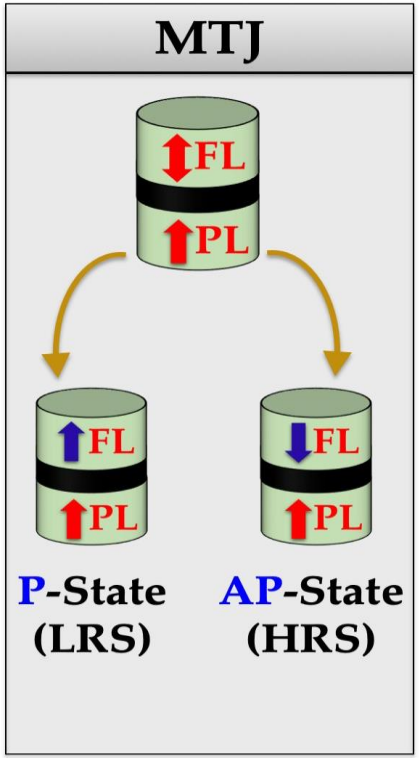


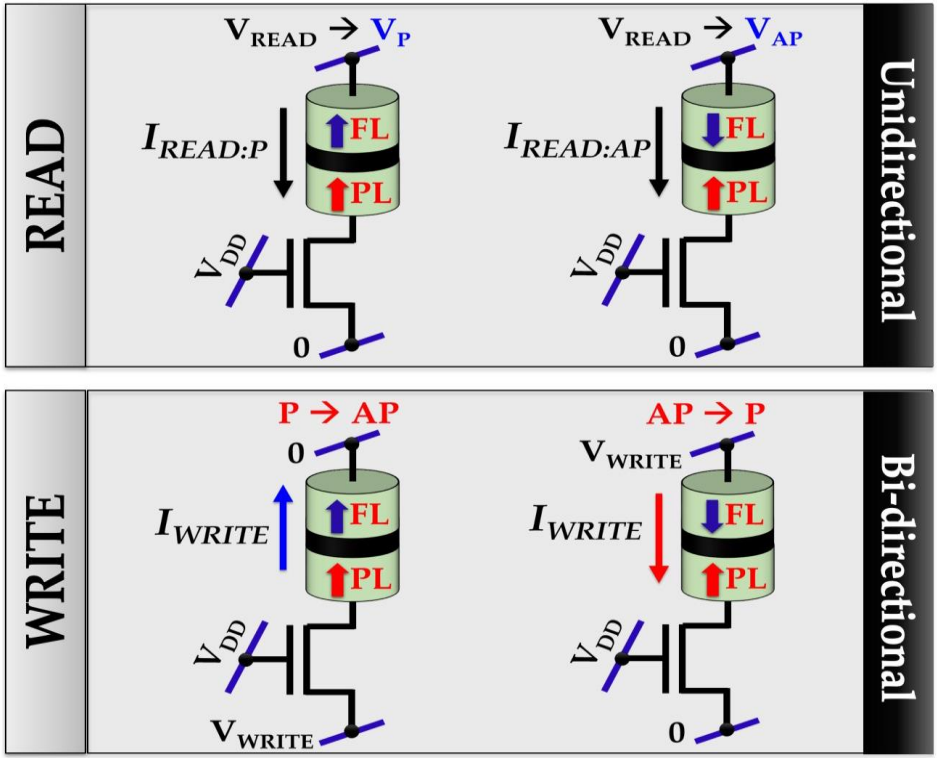
STANDARD STT MRAM: CURRENT BASED WRITE



SPIN BASED MEMORY







Thank you!