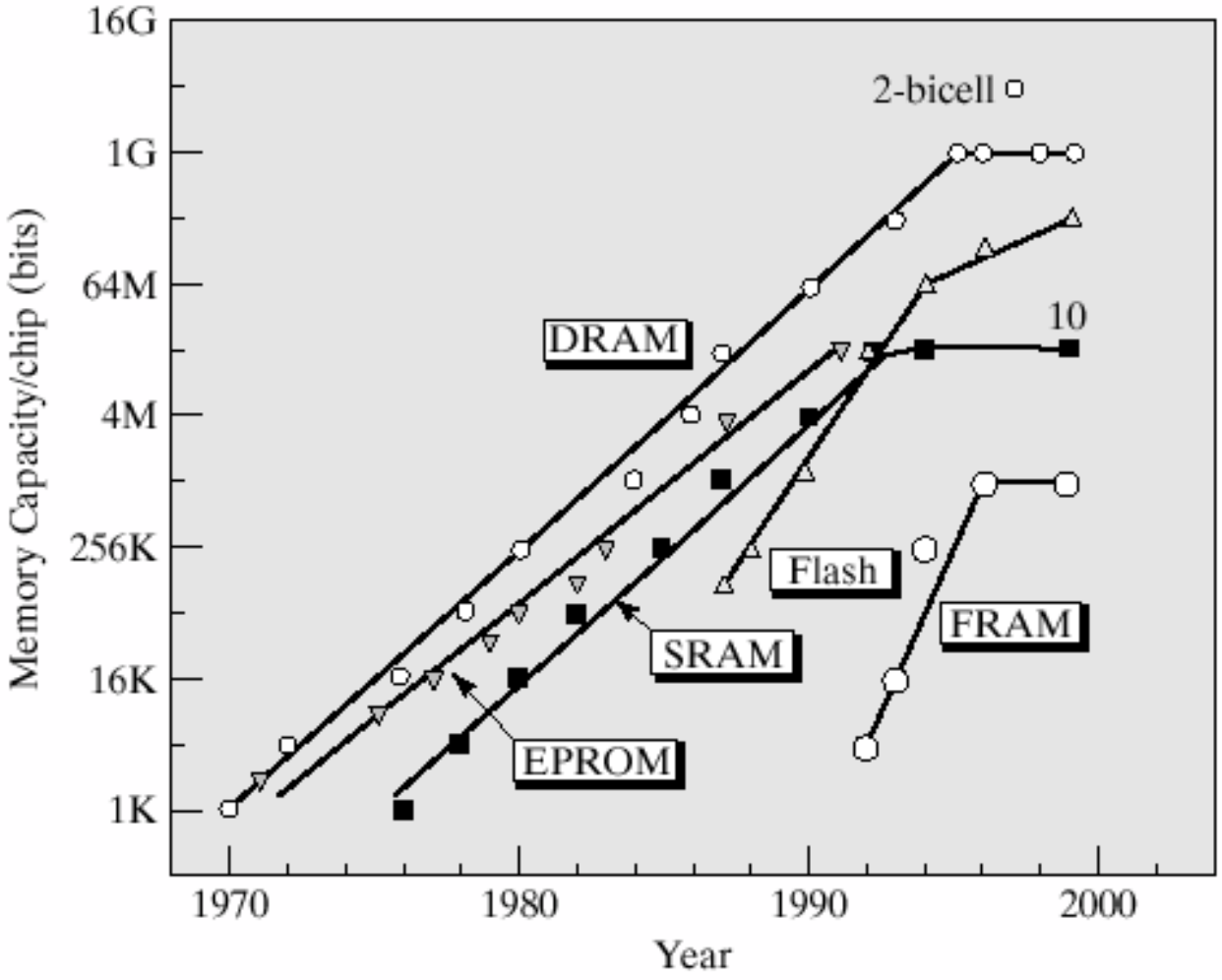
SRAM - Basics

Jaydeep P. Kulkarni

jaydeep@ecn.purdue.edu

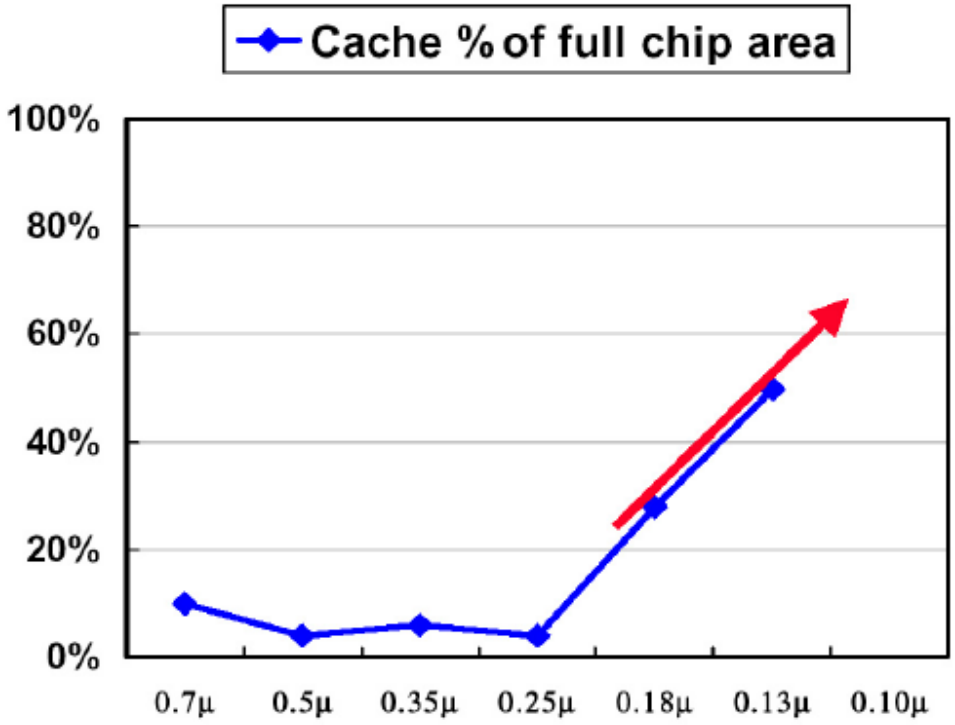
Slides are adopted from Prof. Kaushik Roy, Prof. Chris Kim , Prof. Saibal Mukhopadhyay, and Prof. Jan Rabaey's slides

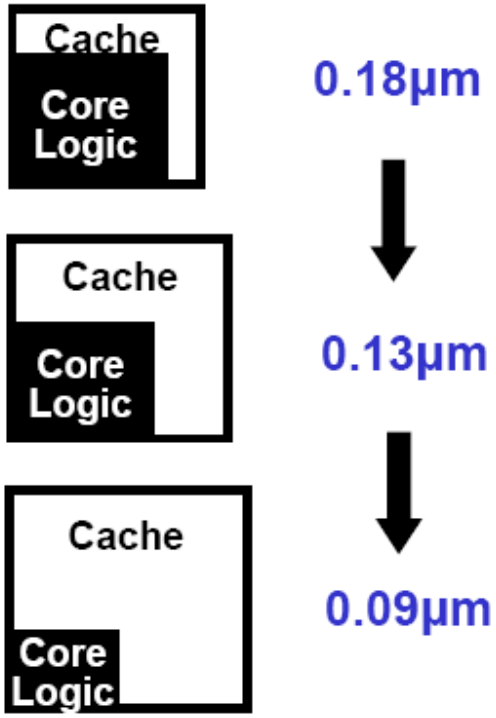
Semiconductor Memory Trends



From [Itoh01]

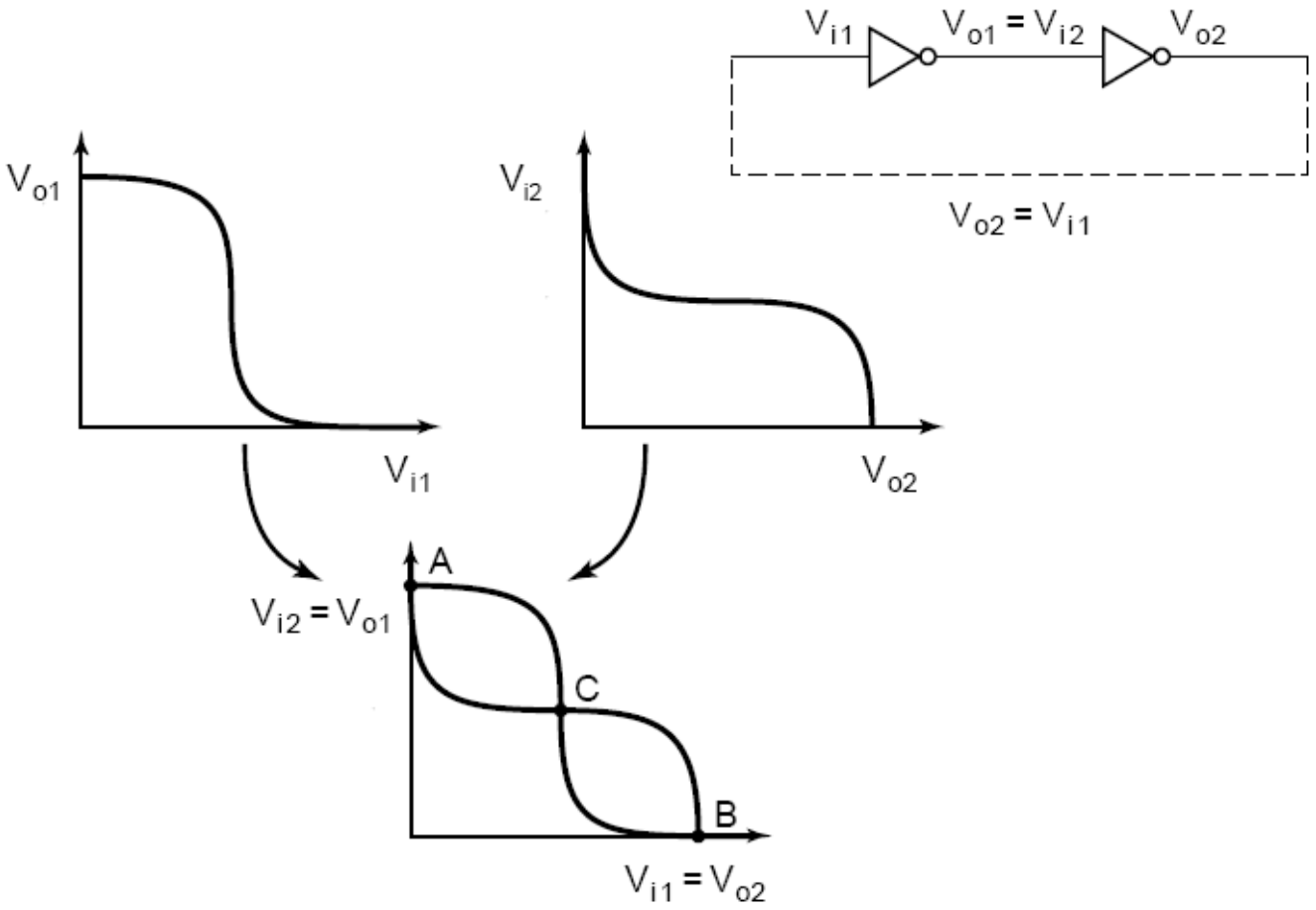
Why SRAMs are Important



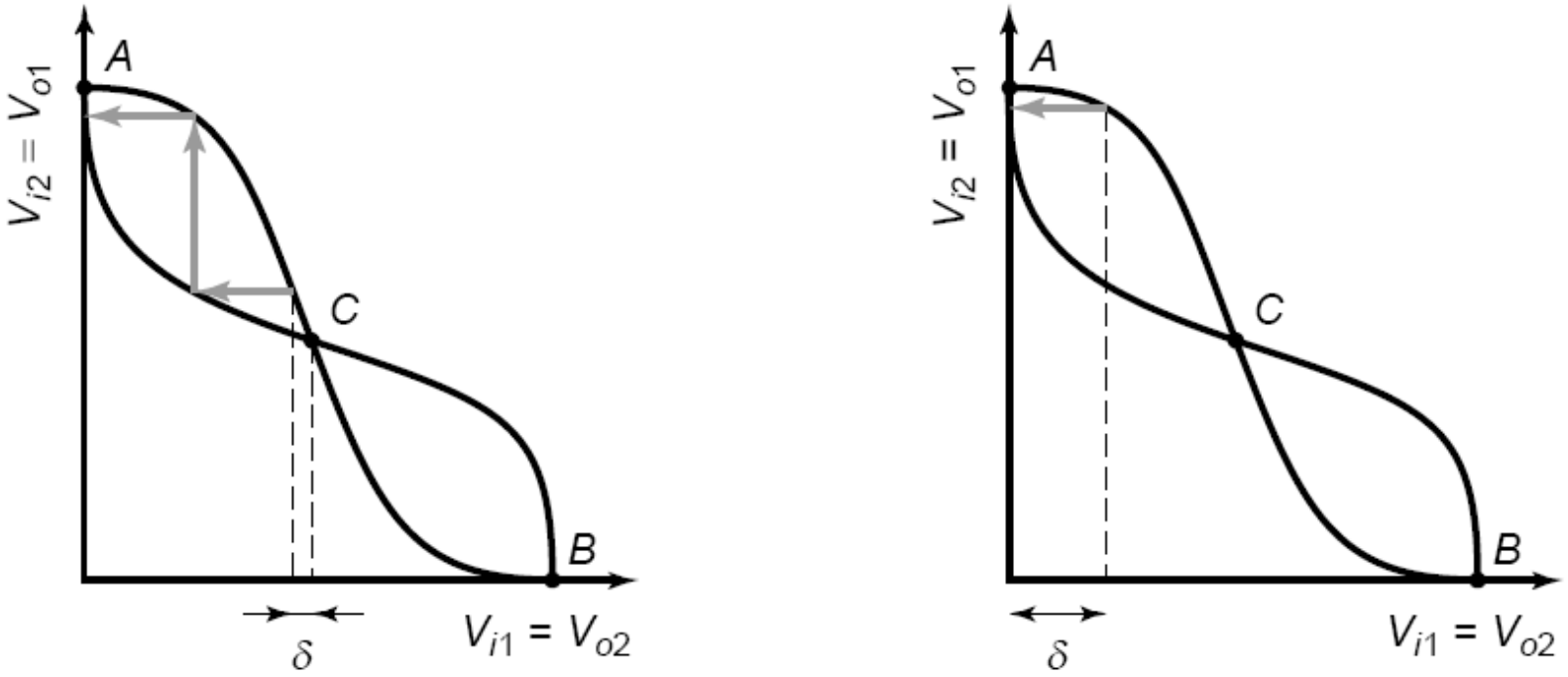


- Memories have better power efficiency compared to logic
- ~9.9B out of 10B transistors will be used for SRAMs
- Company with better SRAM design will dominate

Positive Feedback: Bi-Stability

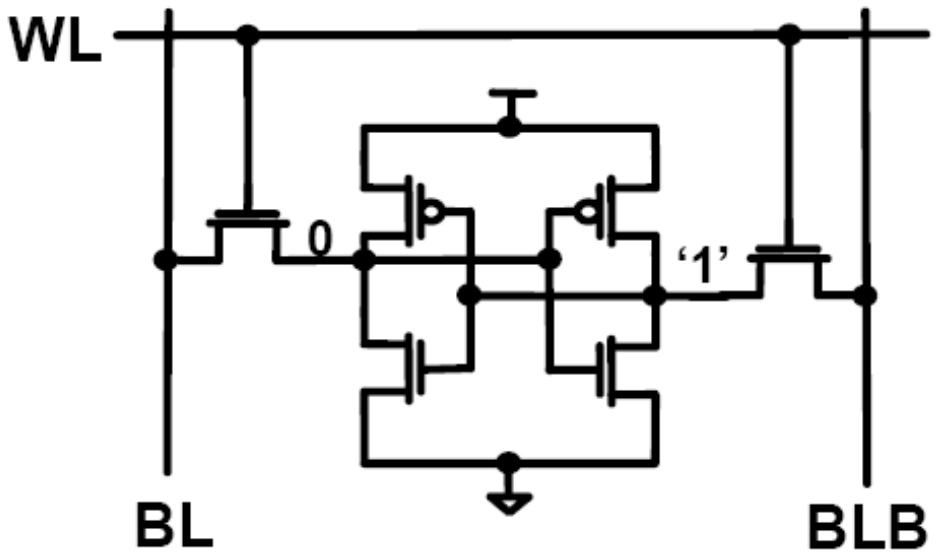


Meta-Stability



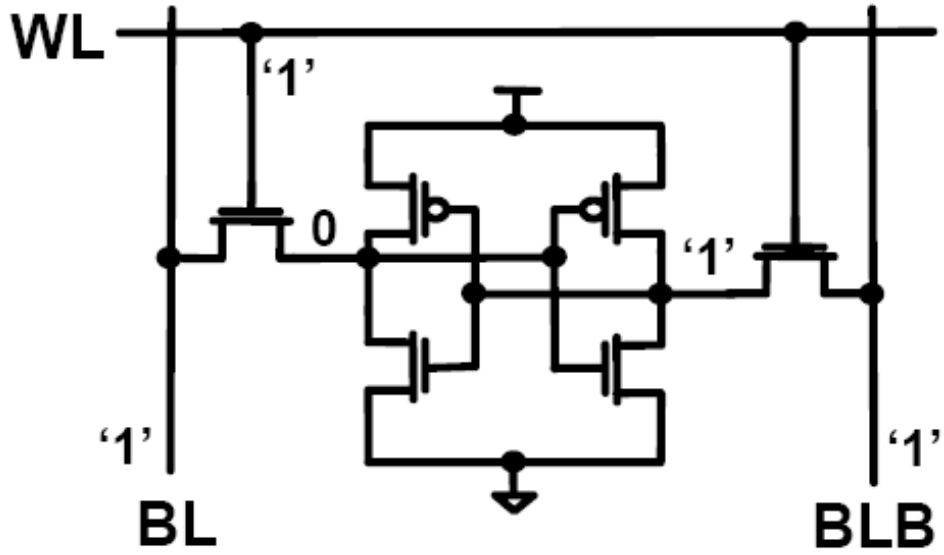
Gain should be larger than 1 in the transition region

SRAM Memory Cell



- NMOS access transistors
- Read and write uses the same port: need sufficient margins
- One wordline to access cell
- Two bit lines (BL, BLB) to carry the data
- Almost minimum size transistors for small cell area

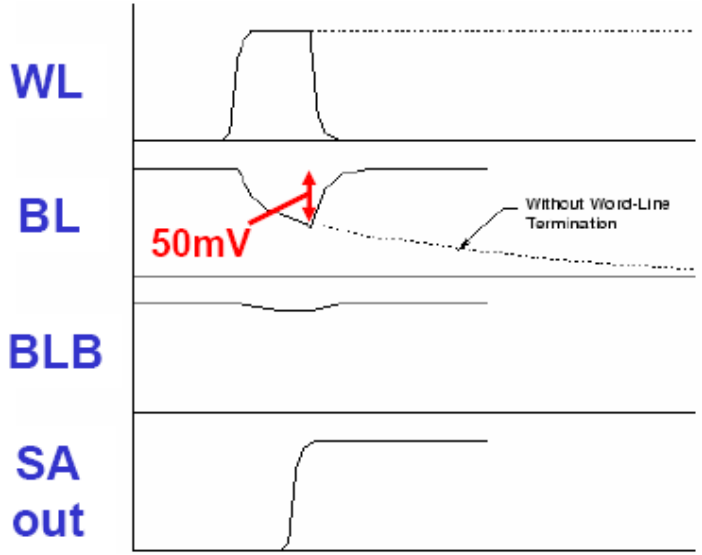
SRAM Read Operation



- Both bit lines are precharged to Vdd
- Wordline is fired for one of the cells on bit line
- Cell pulls down either BL or BLB
- Sense amp regenerates the differential signal
- Data should not flip after read access
- Driver TR must be stronger than access TR

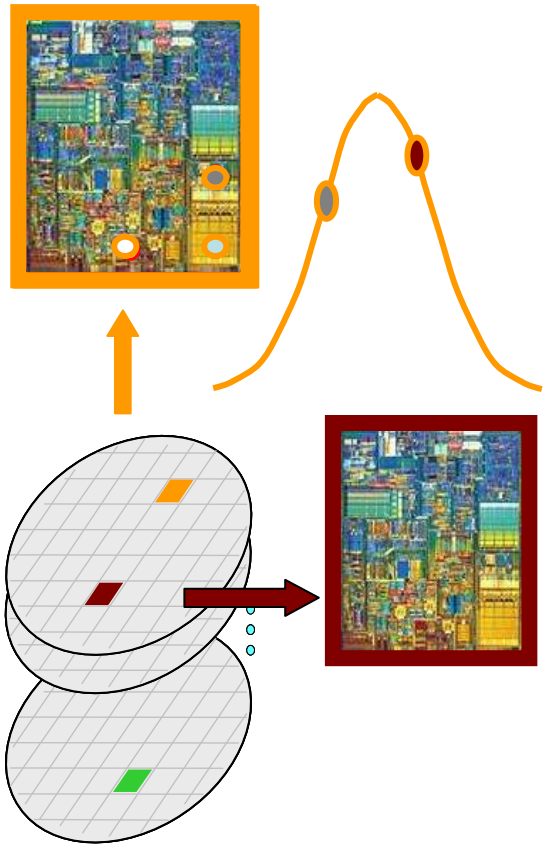
SRAM Read Operation

$$bitline\ delay = \frac{C_{bitline}\Delta V_{bitline}}{I_{cell}}$$

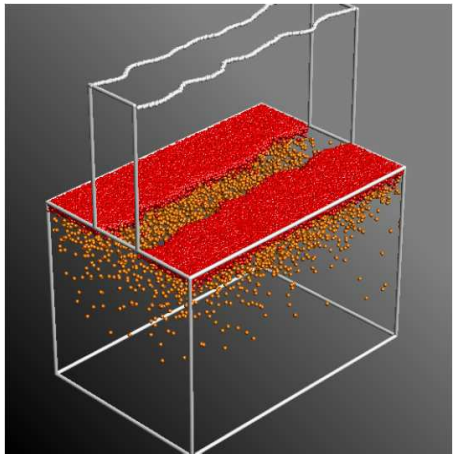


- $C_{bitline}$ is large due to large number of cells attached
- I_{cell} is small due to high density cells
- $\Delta V_{bitline}$ has to be minimized for high speed
  - < 100mV bitline voltage difference generated by SRAM cell
  - Let the sense amplifier finish the job
  - Increased noise sensitivity, circuit complexity

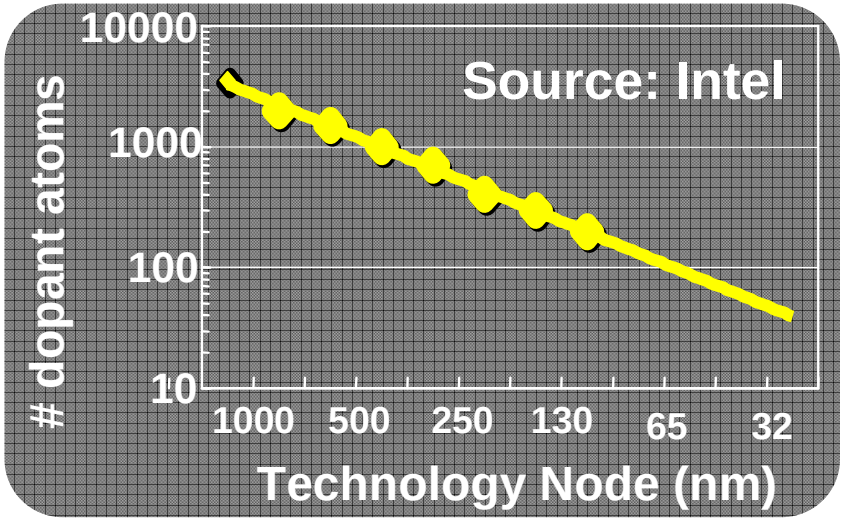
Variation in Process Parameters



Inter and
Intra-die Variations



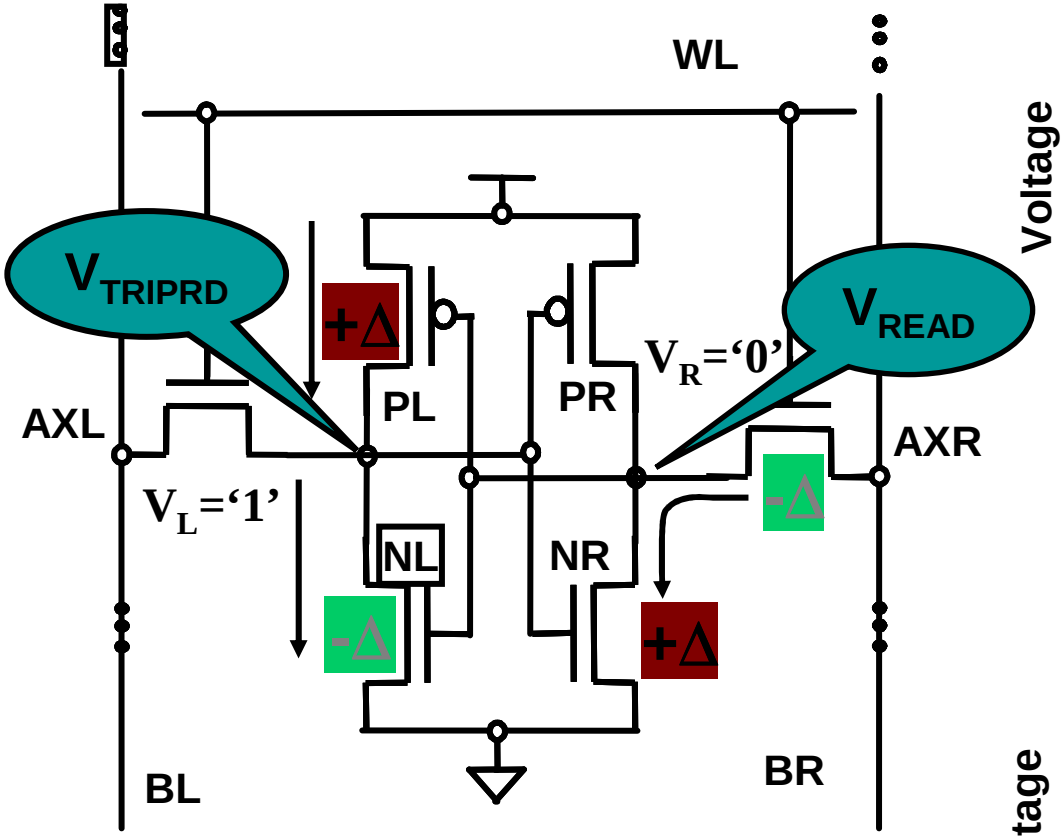
Line Edge Roughness
(LER)

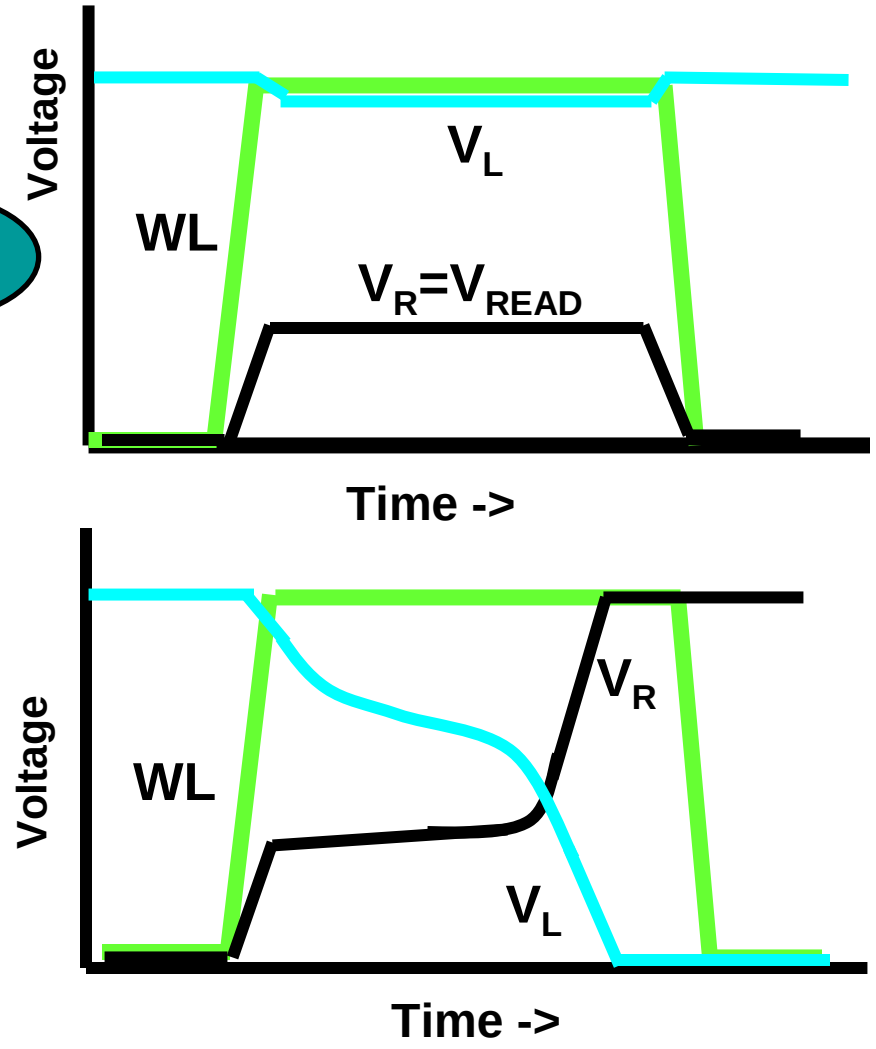


Random Dopant Fluctuation
(RDF)

- SRAM Transistors – Minimum geometry
- Inter-die and Intra-die variations
- LER, RDF induced device mismatch

Parametric Failures: Read Failure

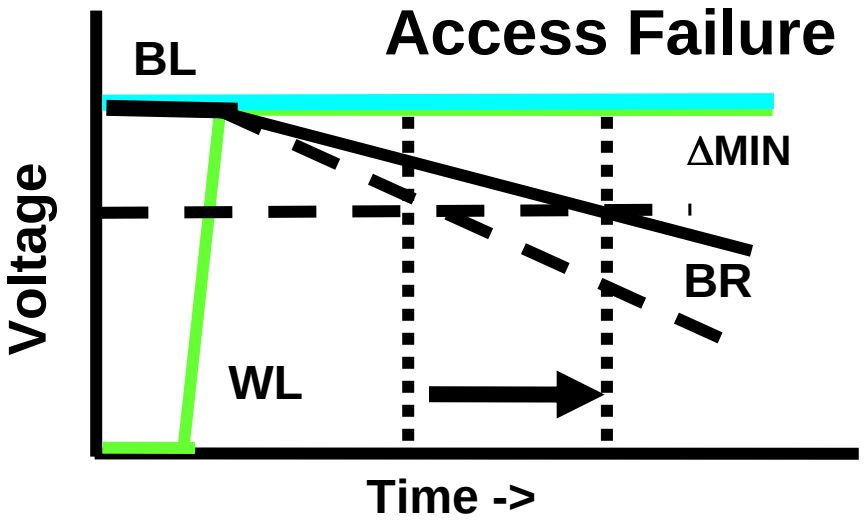


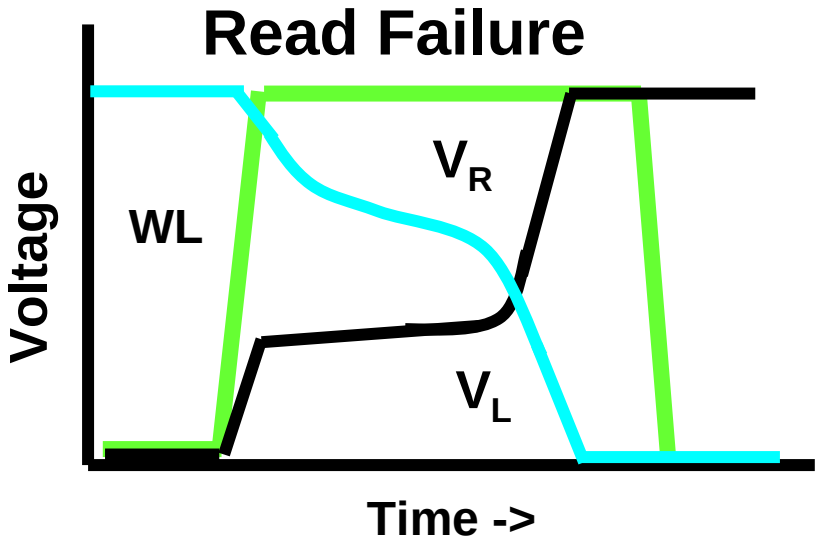


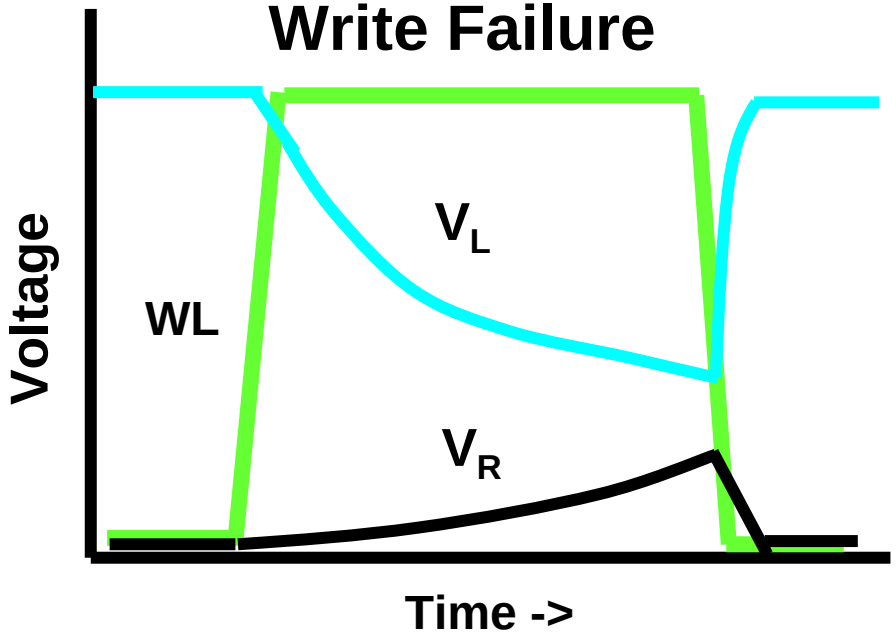
$$P_{RF} = P\left(V_{READ} > V_{TRIPRD}\right)$$

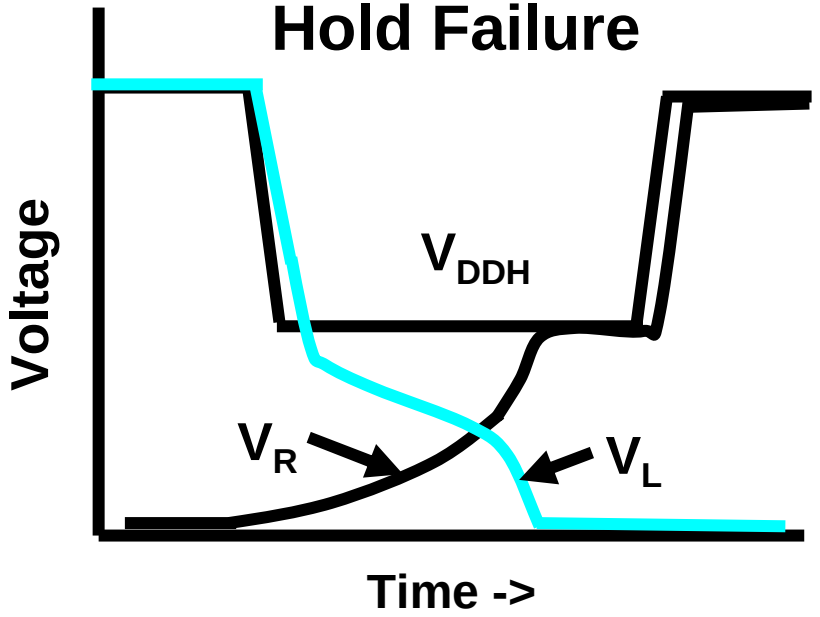
Read failure => Flipping of Cell Data while Reading

Mechanisms of Parametric Failures

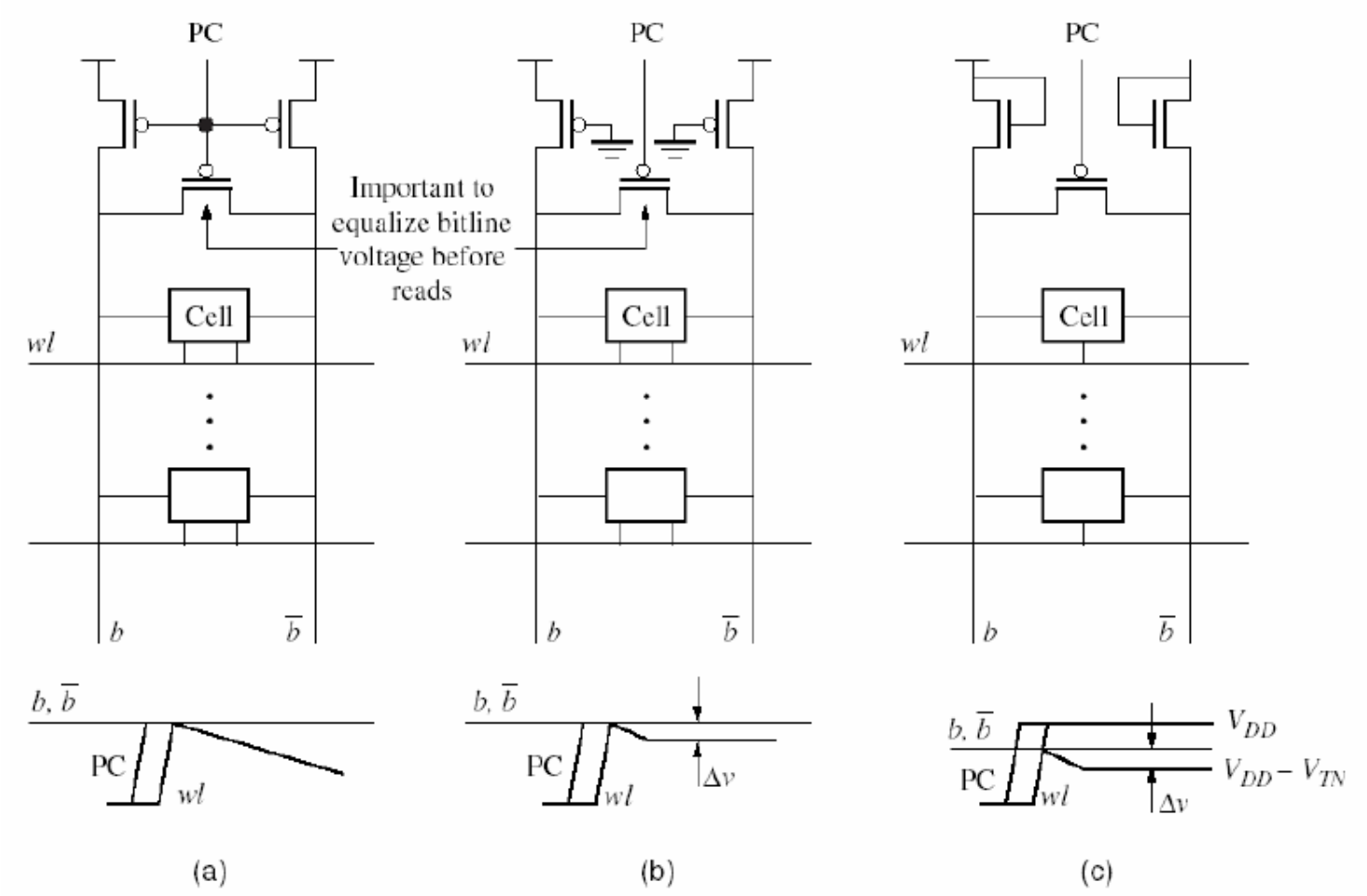




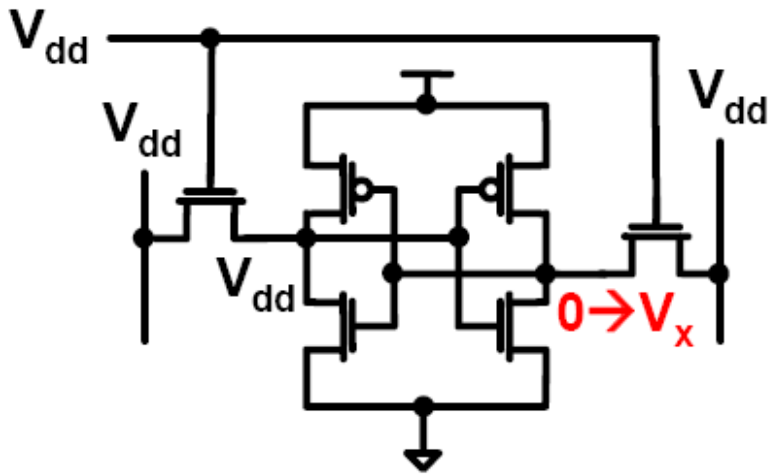




SRAM Read Operation: Precharge

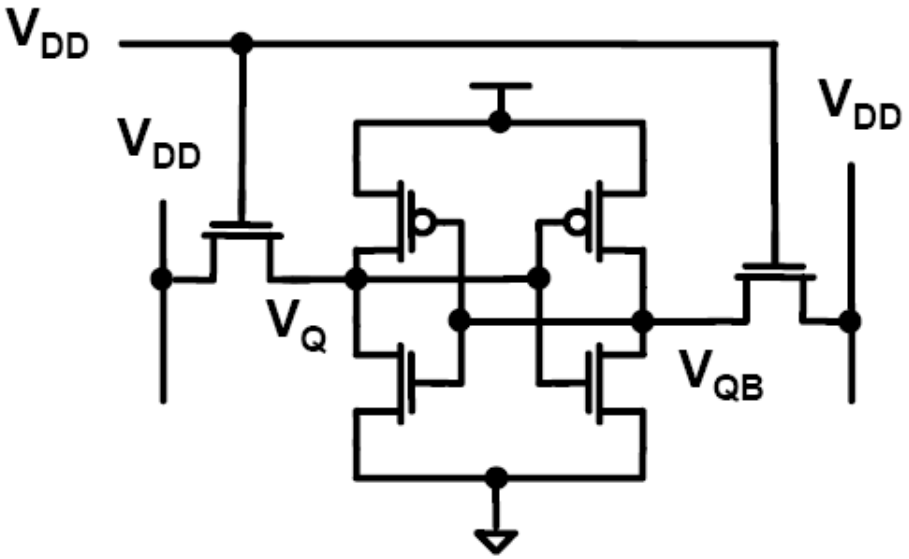


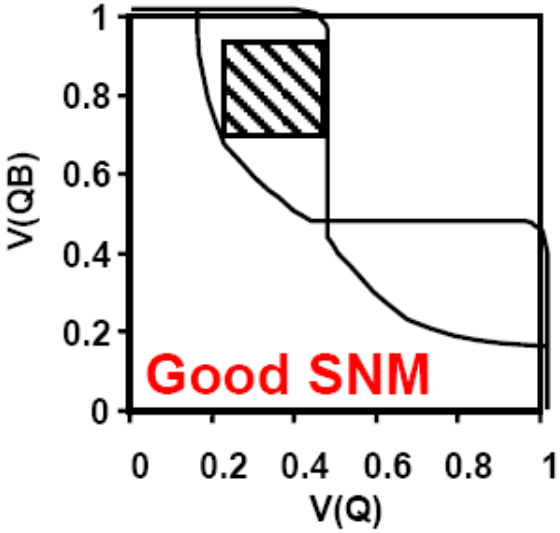
SRAM Cell Read Margin

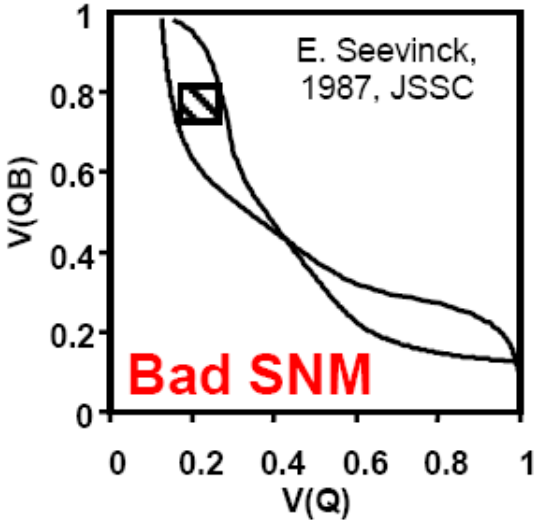


- When cell is not accessed (WL=0)
  - Data is safely kept inside the cell
  - High noise margin
- When cell is accessed (WL=V_{dd})
  - Access transistor acts as a noise source
  - Data '0' is pulled up to V_x
  - Cell data can flip if V_x rises above V_{tn}

Static Noise Margin

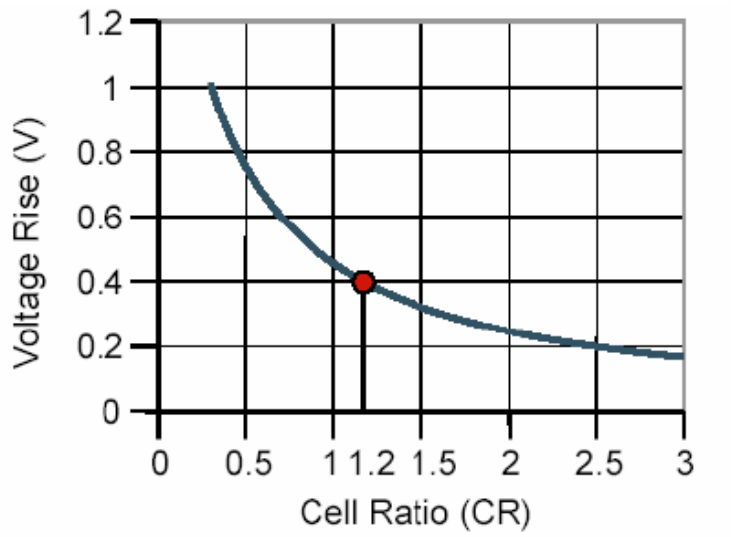






- Destructive read problem
- The size of the largest square enclosed in the butterfly curves = read static noise margin

Techniques to Improve Read Margin



Cell beta ratio =

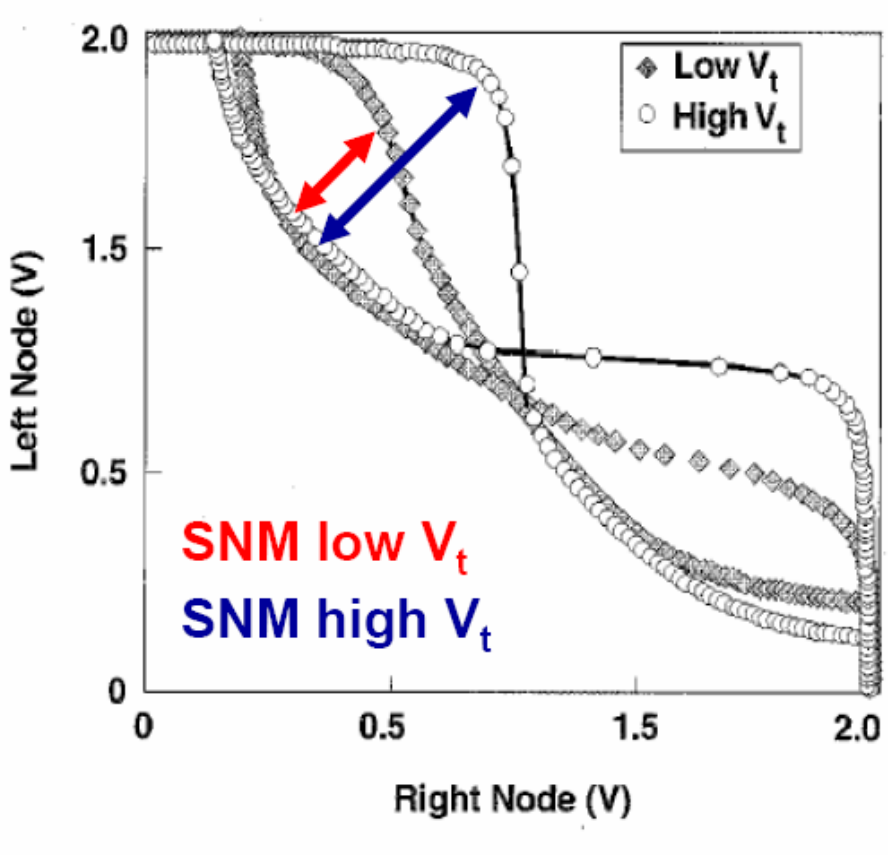
$$(W/L)_{drv} / (W/L)_{access}$$

J. Rabaey

- **Increasing the size of the driver NMOS improves read margin**
- **But remember, area is the number one constraint in memory design**
- **Increasing cell size a not a good trade off**

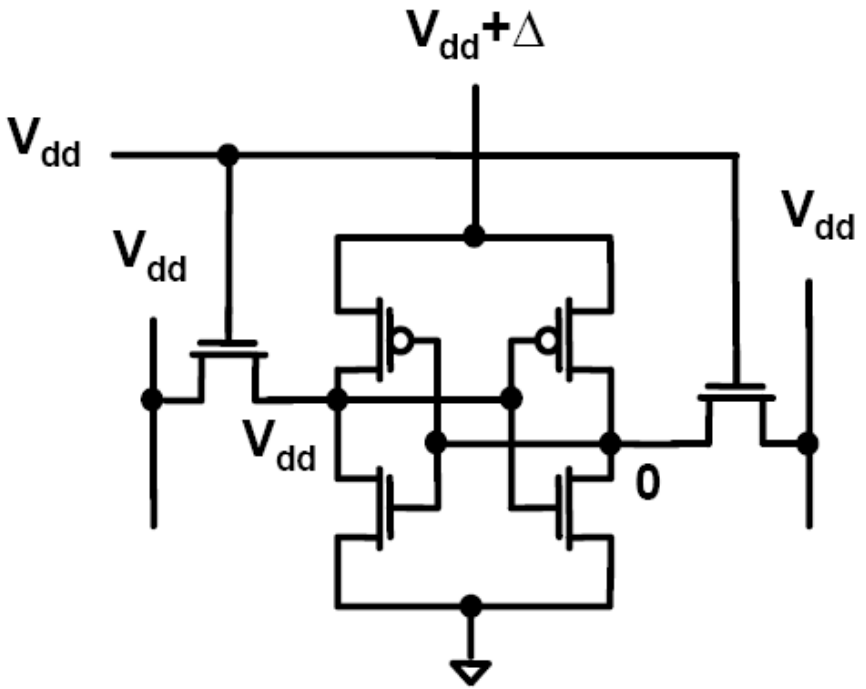
Techniques to Improve Read Margin

- High V_t transistors
  - Internal node on low side needs to rise to V_t or more
  - Virtually never happens when V_t is larger than half V_{dd}
  - Cell is extremely stable at ultra-low power design point
  - Beta ratio constraint is relaxed → smaller driver and larger access TR can be used for faster read and write

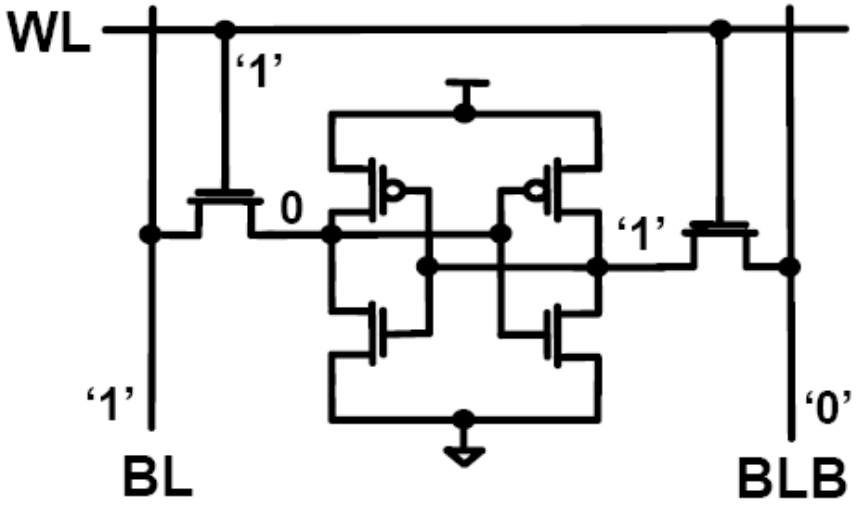


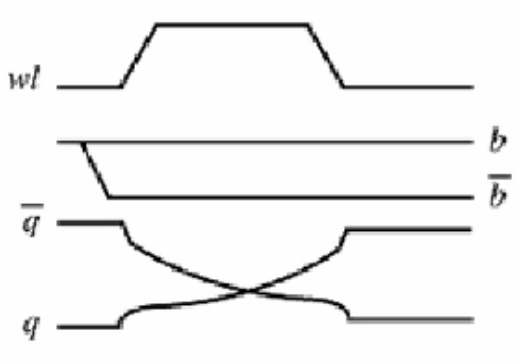
Techniques to Improve Read Margin

- **Boosted cell supply**
  - Supply voltage of SRAM cell is higher than outside
  - Makes driver stronger than access, suppressing the rise in the low side
  - Effectively improves the beta ratio
  - Driver NMOS can be downsized, decreasing cell size



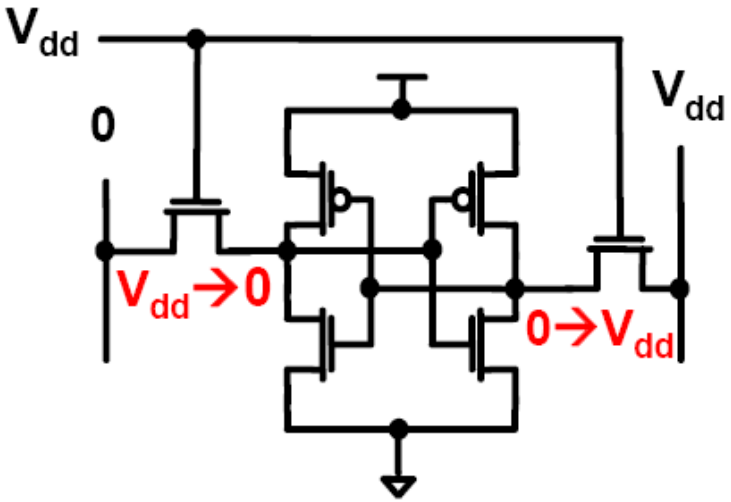
SRAM Write Operation

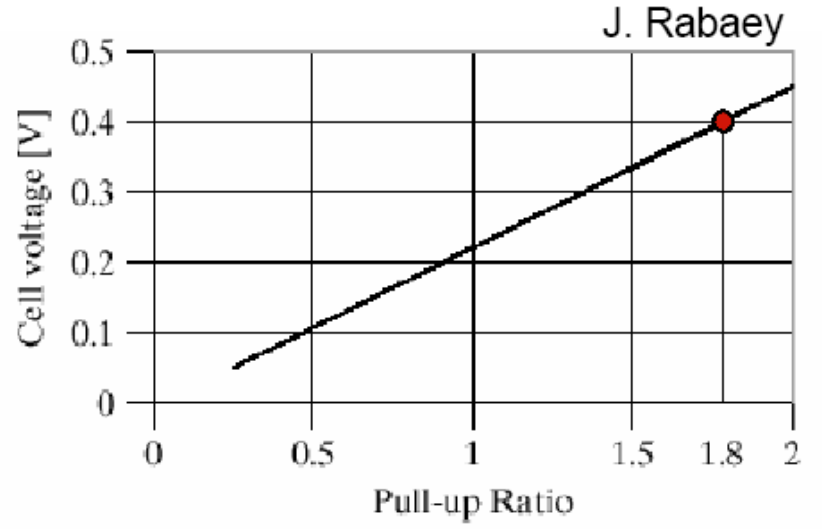




- Launch the write data on BL and BLB
- Word line signal is fired
- Low bit line value flips cell data
- Access TR must be stronger than PMOS load

SRAM Cell Write Margin



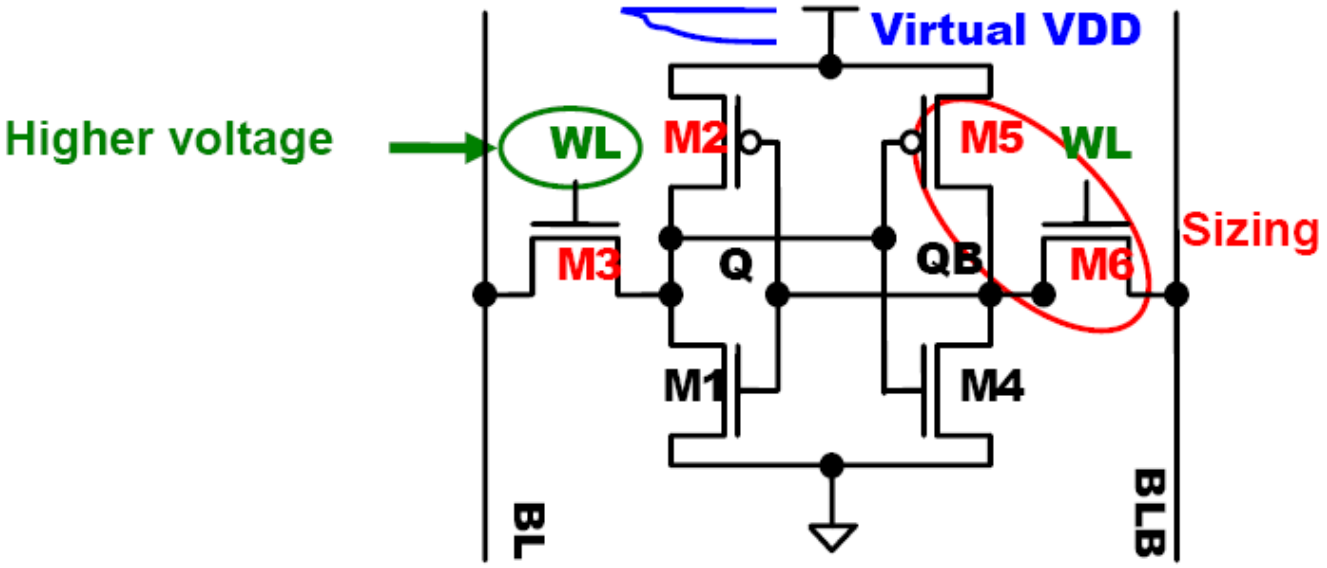


$= (W/L)_{pmos} / (W/L)_{access}$

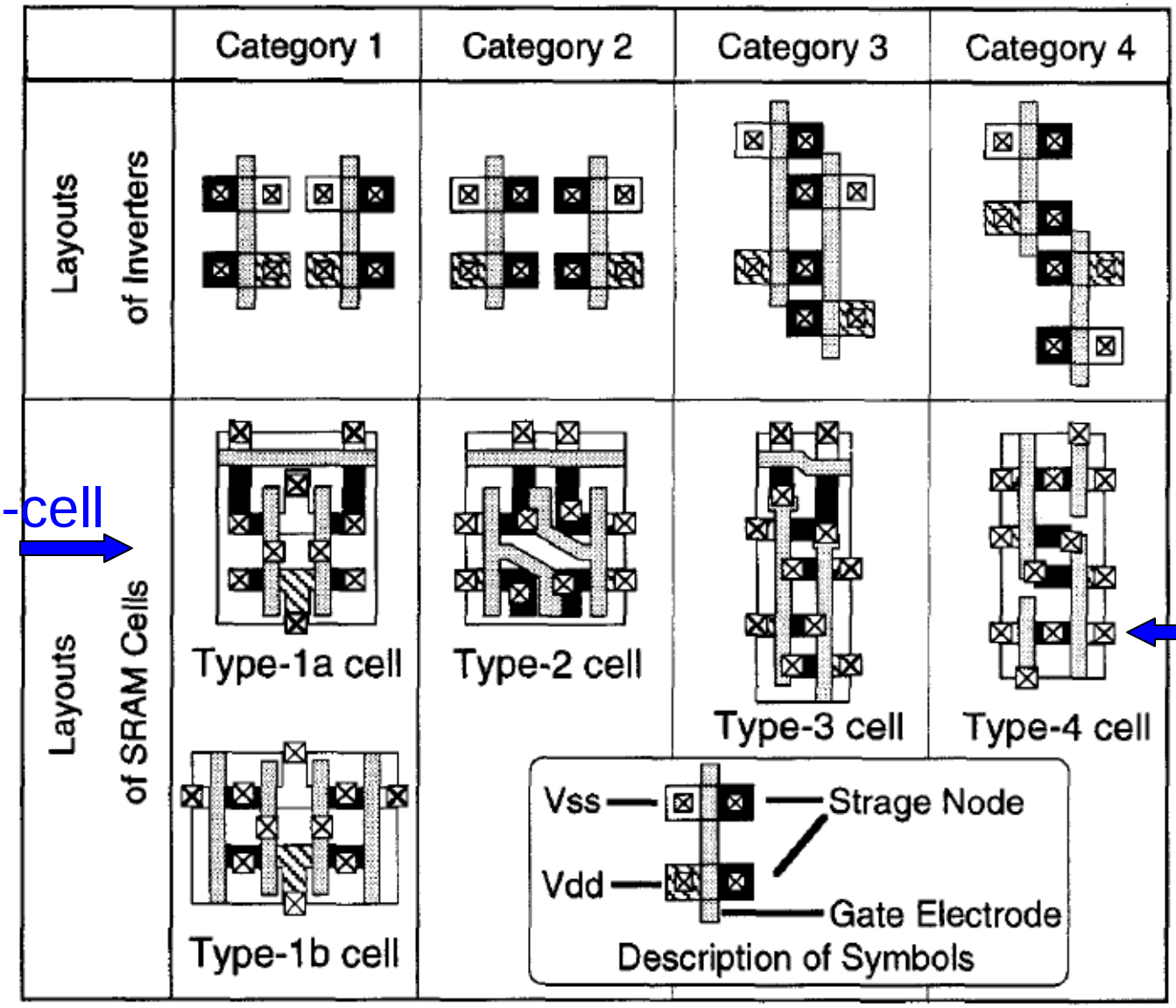
- Access transistor must be stronger than PMOS to pull the below the trip point (typical pull-up ratio ~ 1.5)
- To avoid cell size increase, correct pull-up ratio achieved by controlling V_{tn} and V_{tp}

Techniques to Improve Write Margin

- Sizing: access TR vs. PMOS in latch
- Higher WL voltage for access TR
- Virtual VDD



SRAM Bitcell Layouts

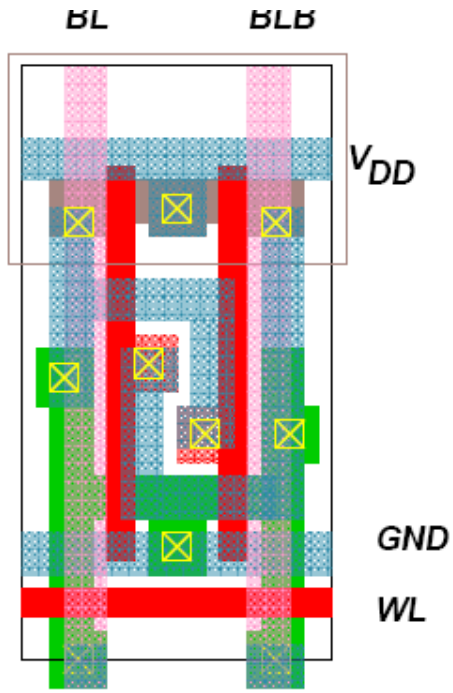


Tall-cell

Thin-cell

M. Ishida, IEDM 98

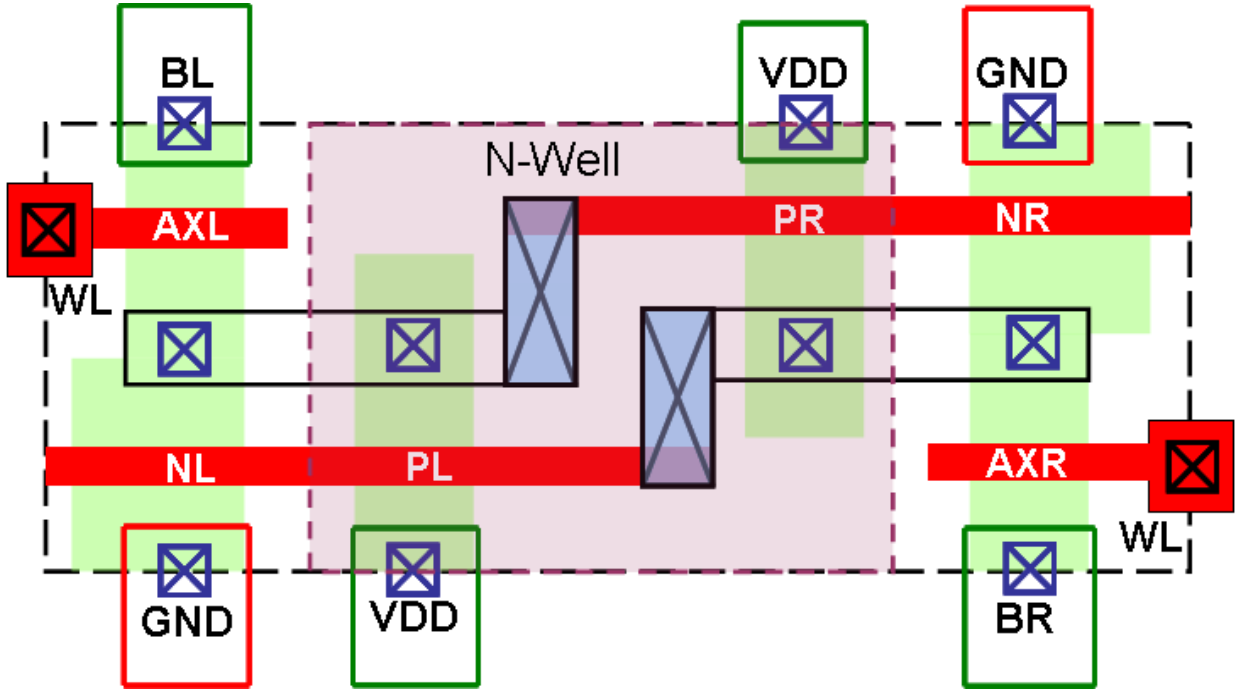
Tall Cell Layout



Compact cell

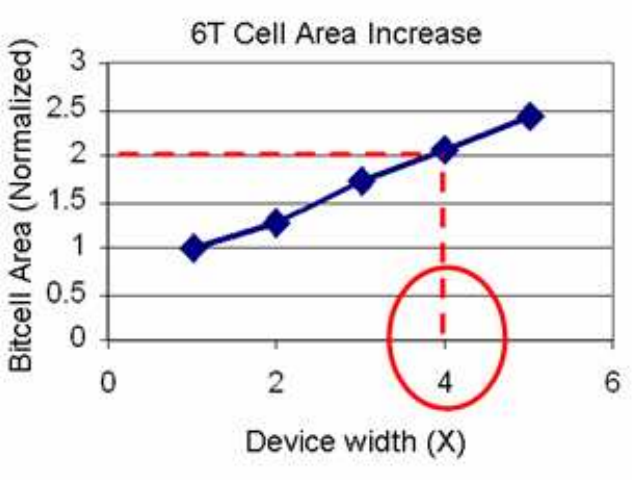
- **Bitlines: M2**
- **Wordline: strapped in M3**

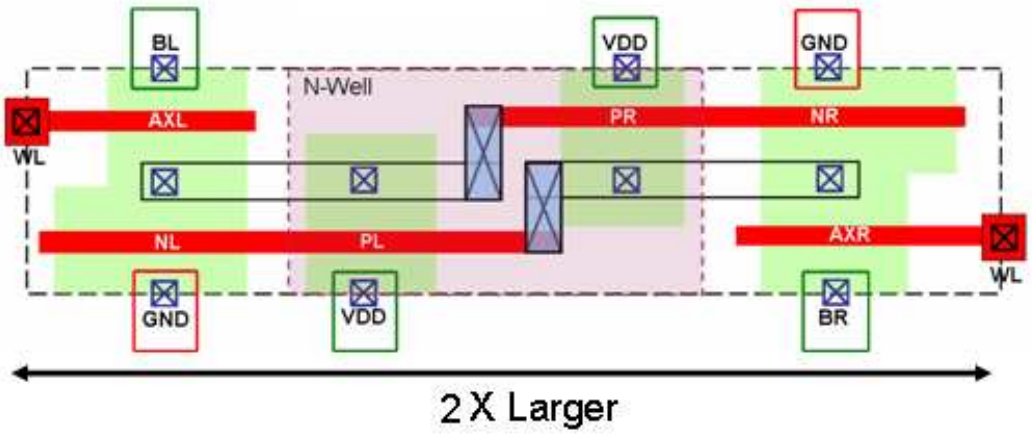
Thin Cell Layout



- 2 Poly-pitch, N-well continuous, Shared Contacts
- WL (Horizontal), BL(Vertical), Lower bitline capacitance

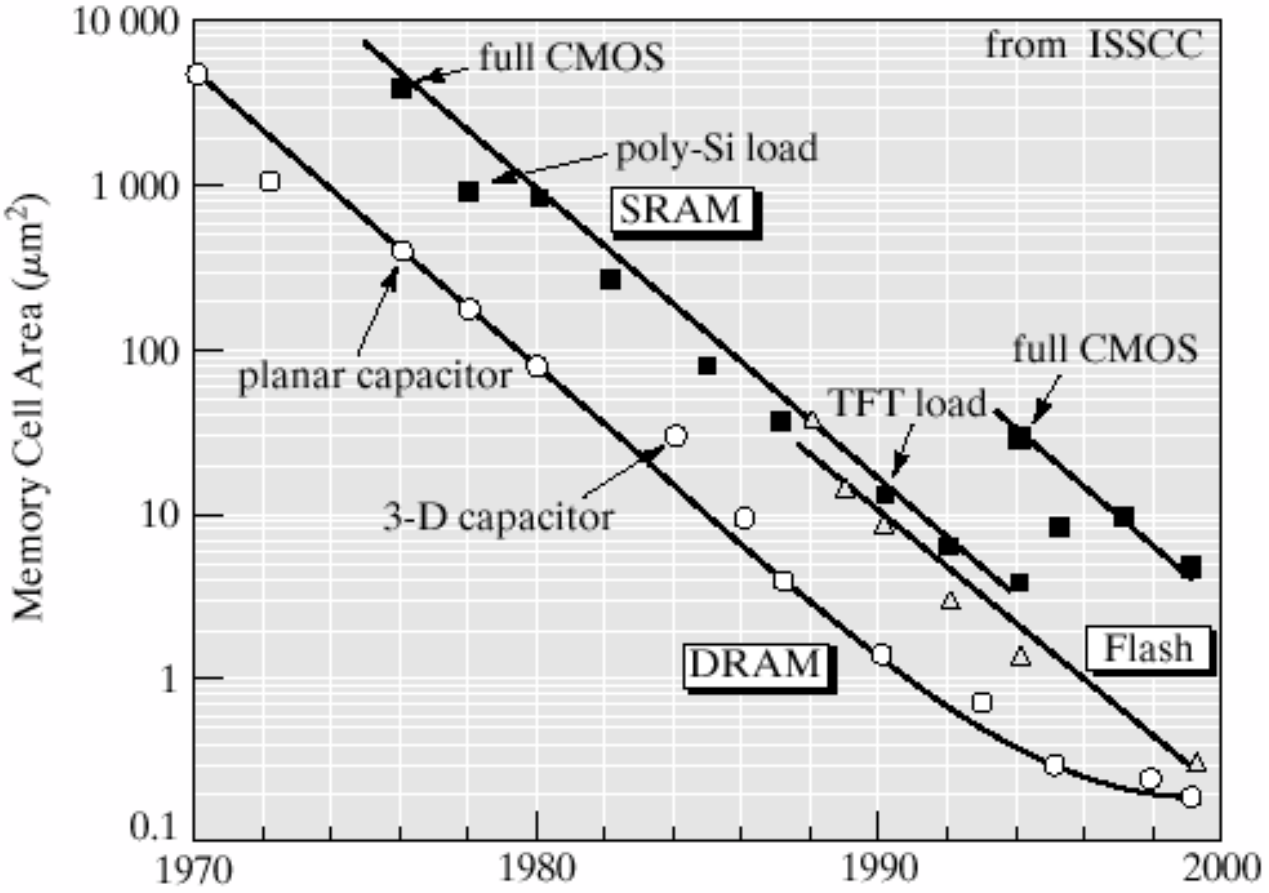
Upsizing Thin-cell size





Increasing cell area by 2X increase device widths by 4X !

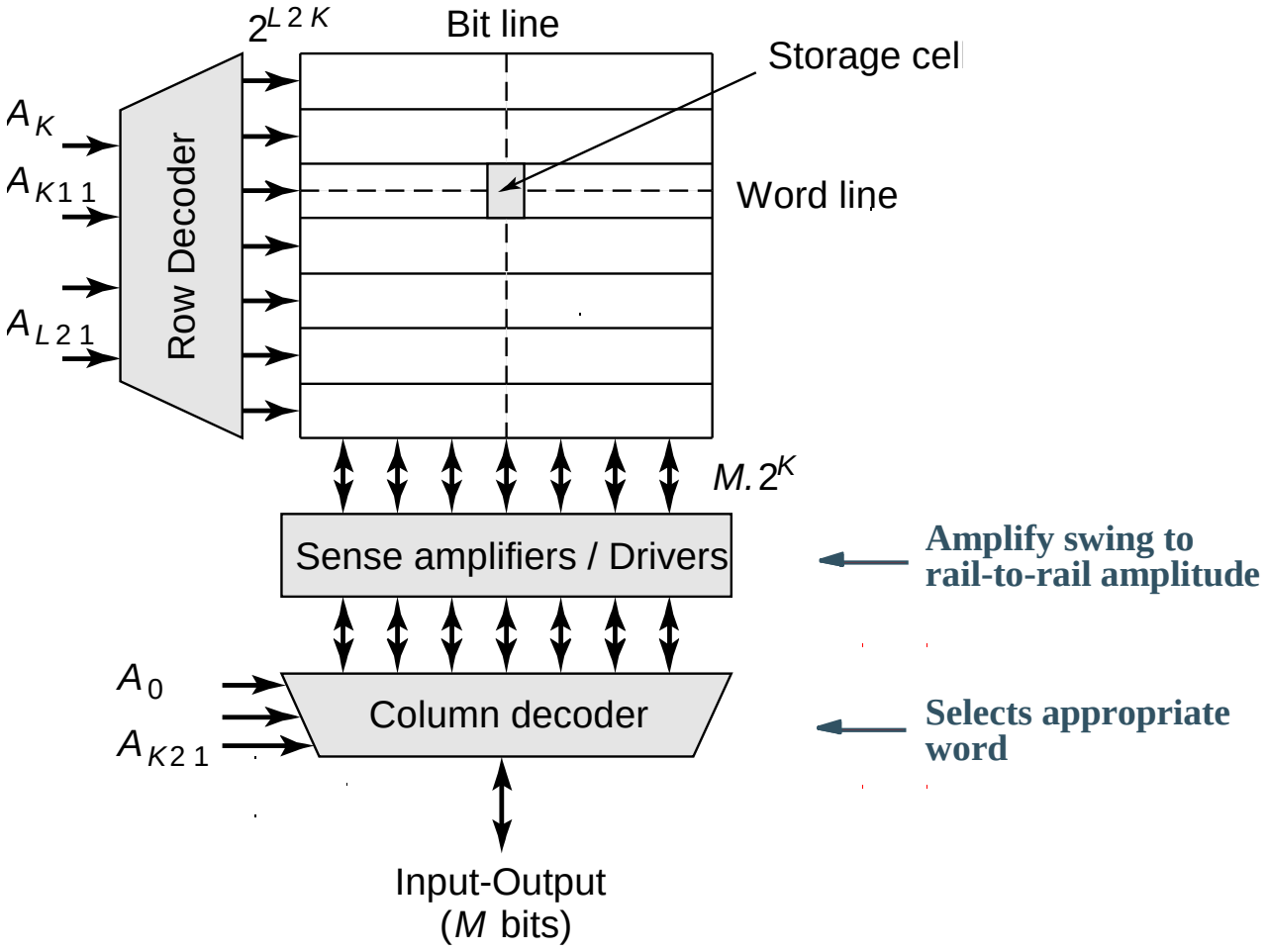
Trends in Memory Cell Area



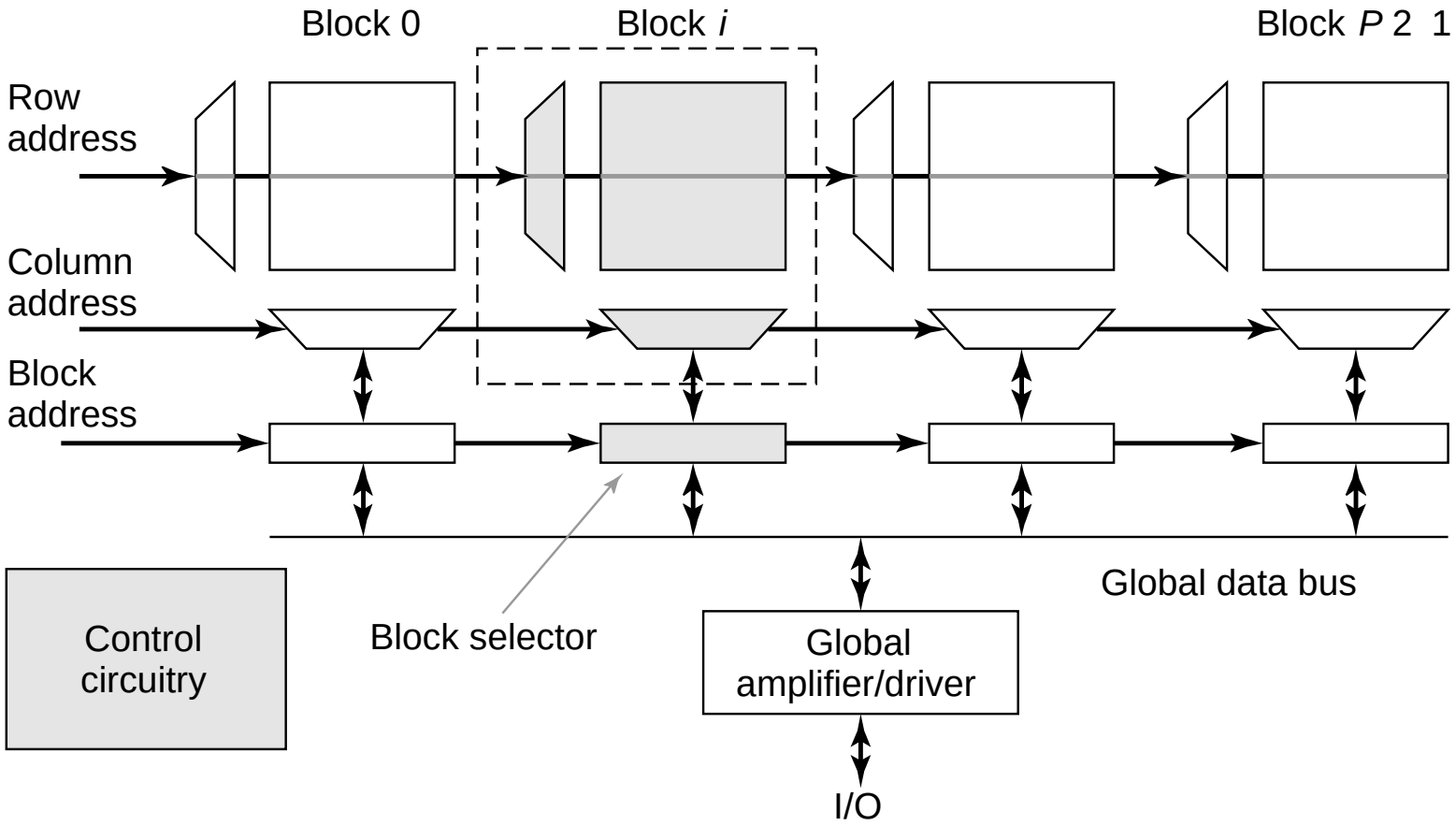
From [Itoh01]

Array-Structured Memory Architecture

Problem: ASPECT RATIO or HEIGHT >> WIDTH



Hierarchical Memory Architecture

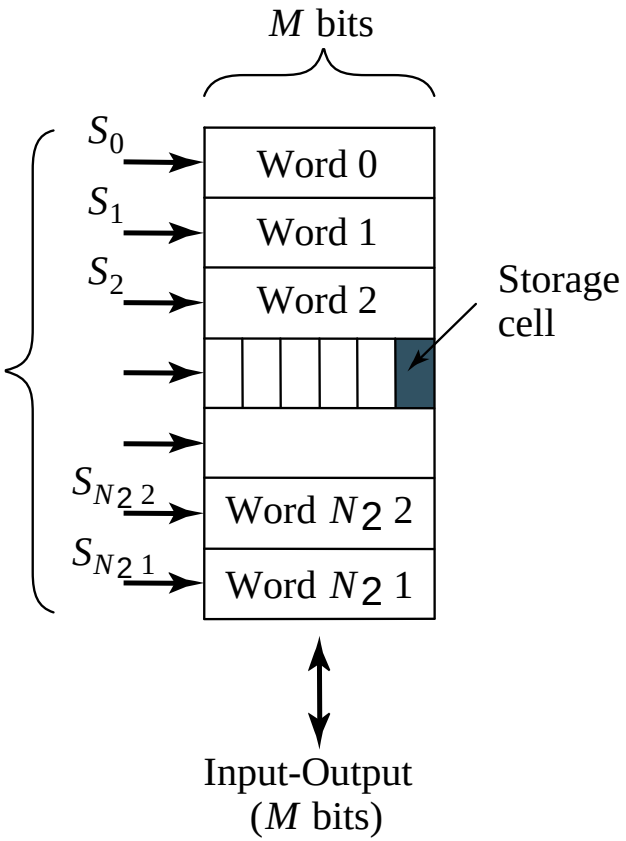


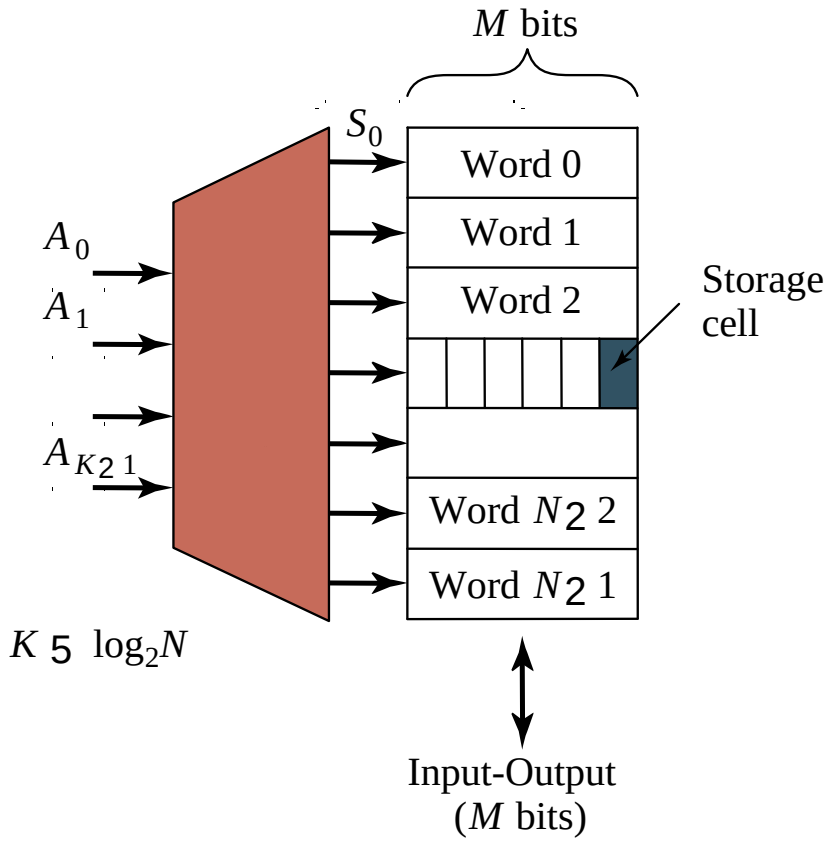
Advantages:

1. Shorter wires within blocks

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

Memory Architecture: Decoders



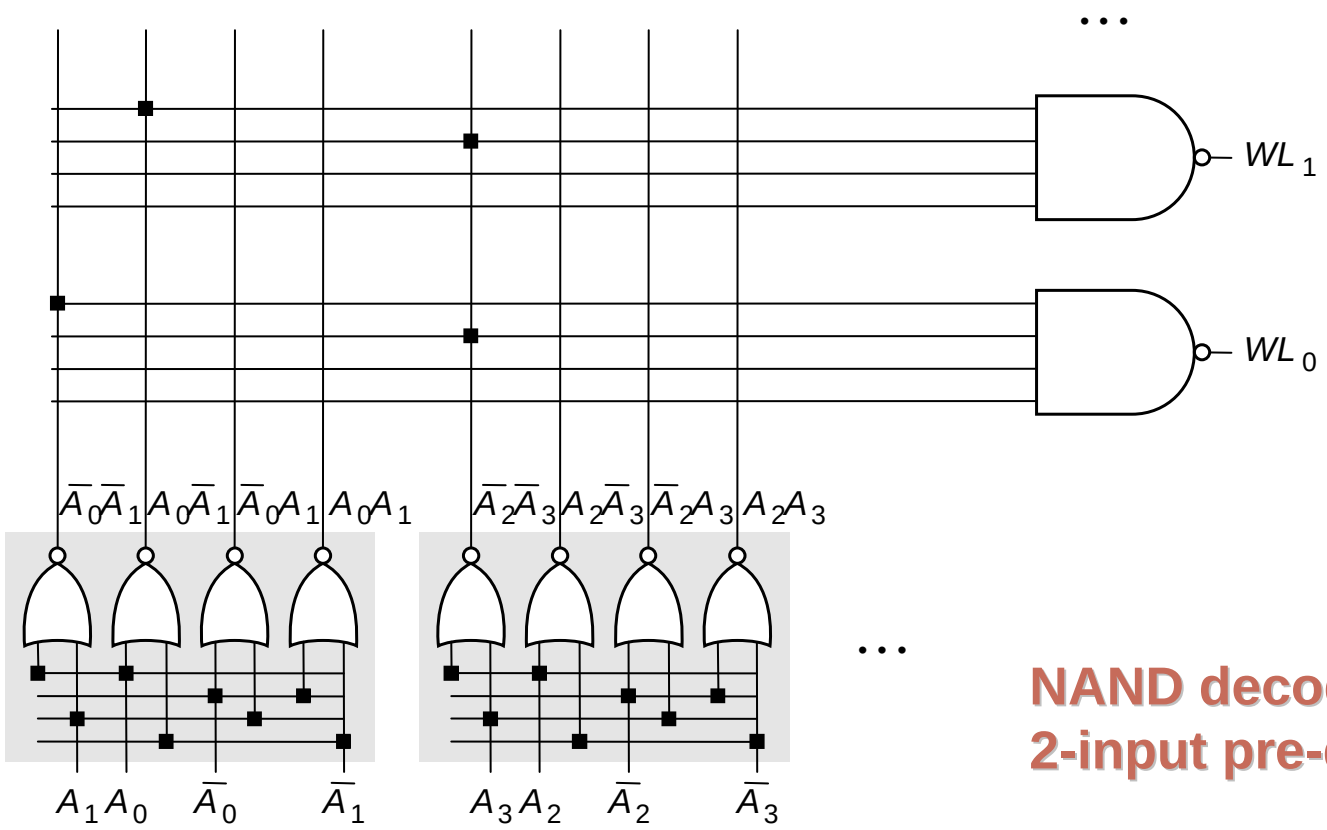


Intuitive architecture for N x M memory
Too many select signals:
N words == N select signals

Decoder reduces the number of select signals
$K = log_2 N$

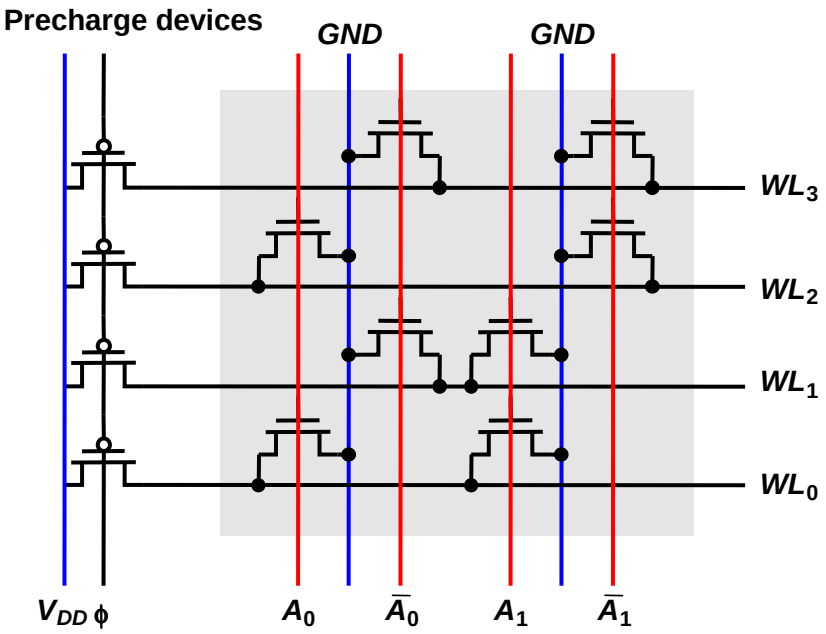
Hierarchical Decoders

Multi-stage implementation improves performance

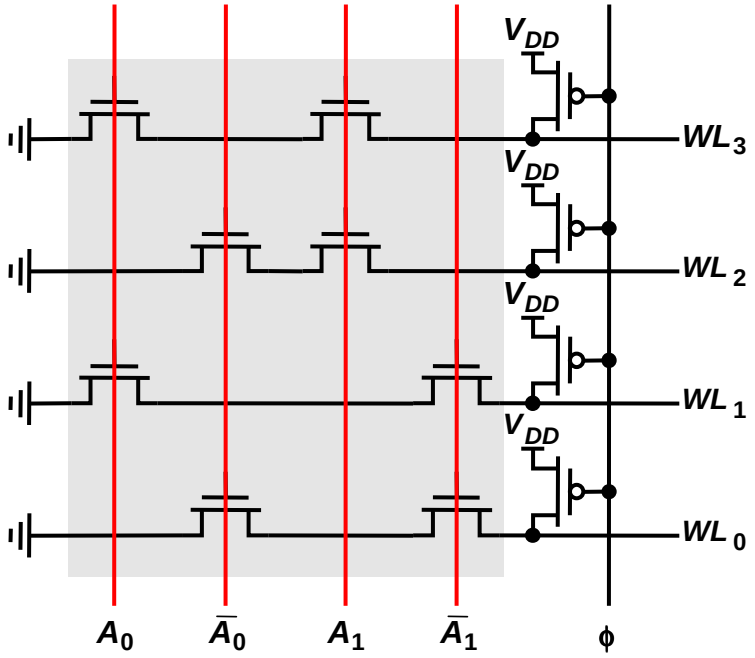


NAND decoder using 2-input pre-decoders

Dynamic Decoders

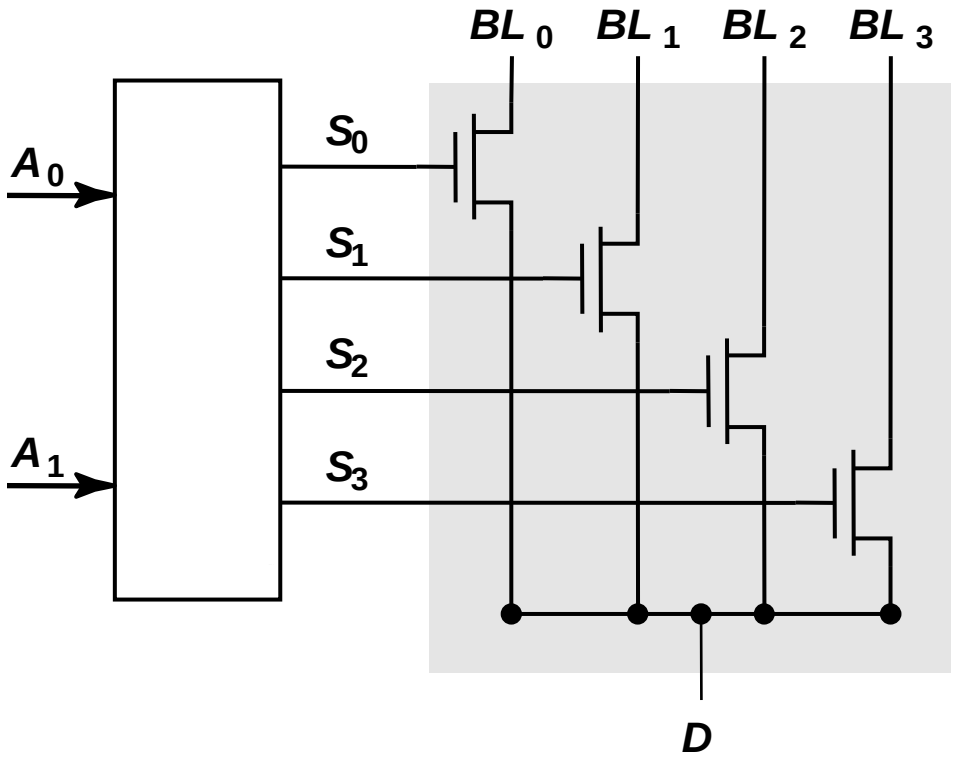


2-input NOR decoder



2-input NAND decoder

Column Decoder

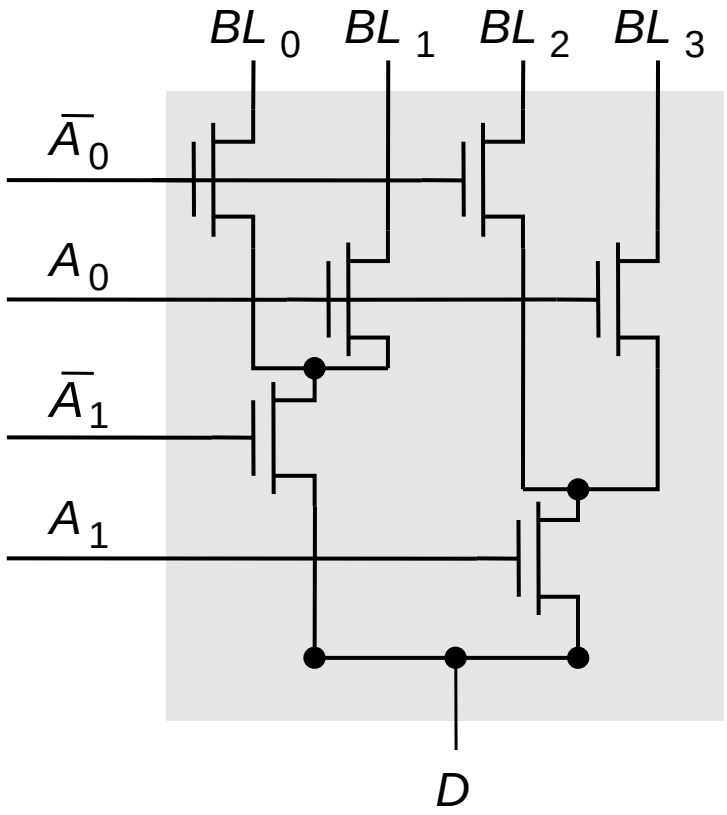


Advantages: speed (t_{pd} does not add to overall memory access time)

Only one extra transistor in signal path

Disadvantage: Large transistor count

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

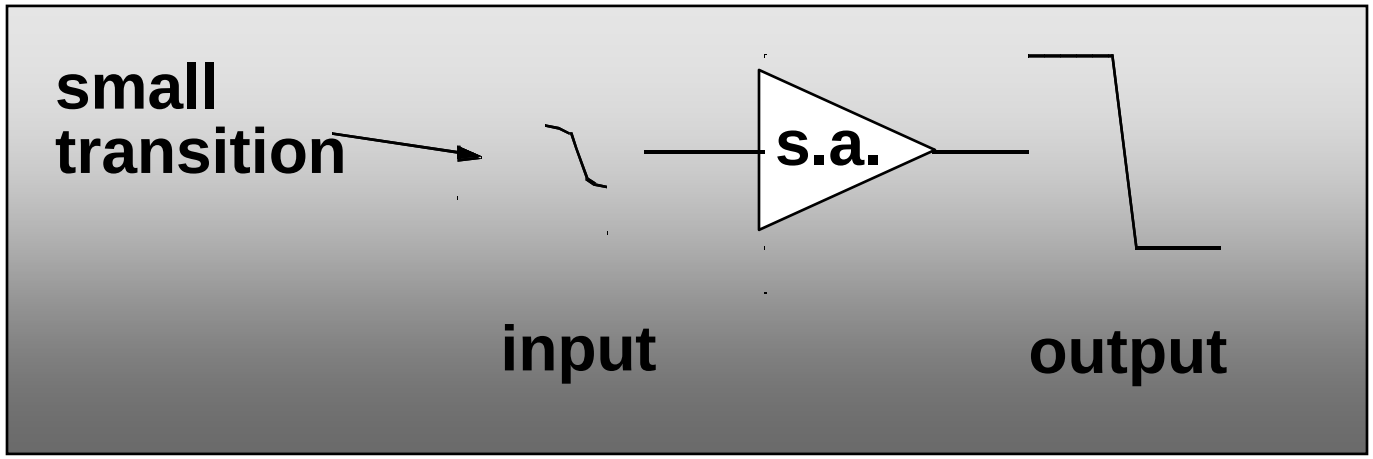
$$t_p = \frac{C \times \Delta V}{I_{av}}$$

make ΔV as small as possible

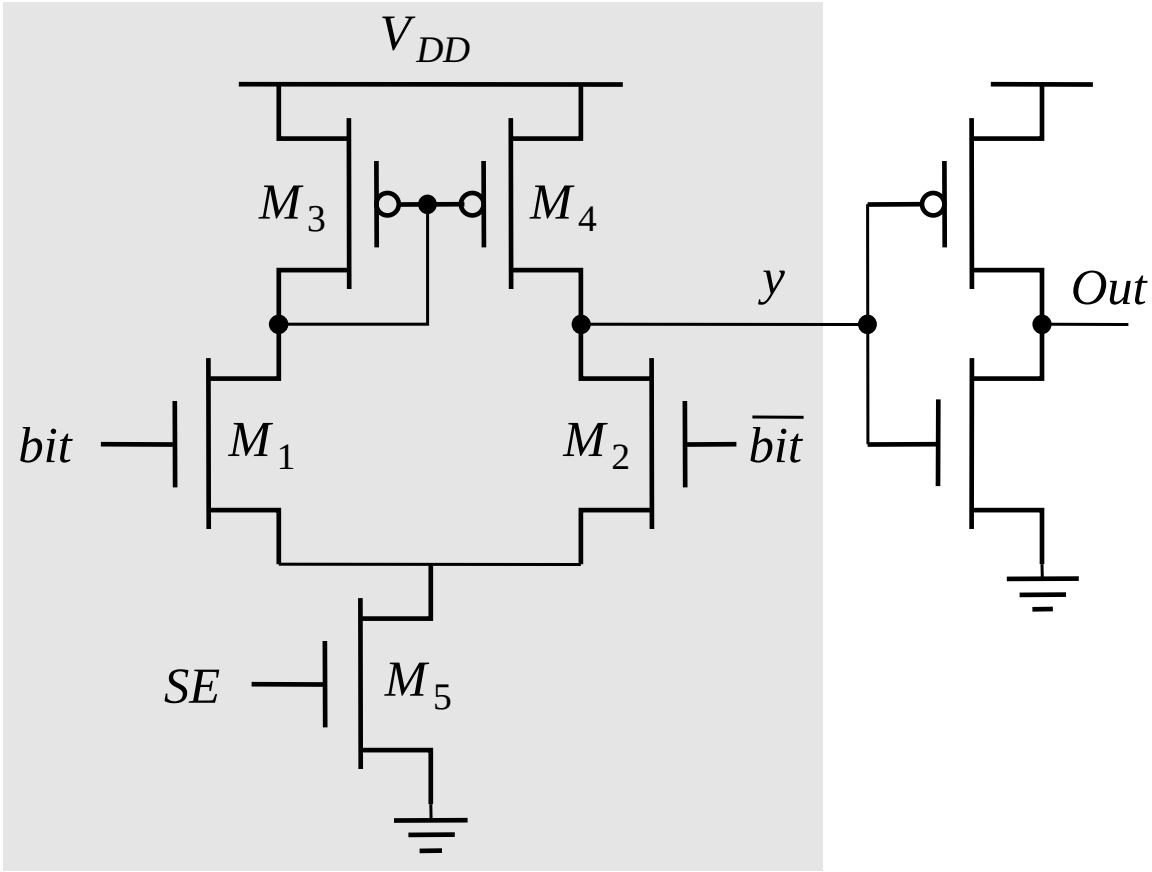
large

small

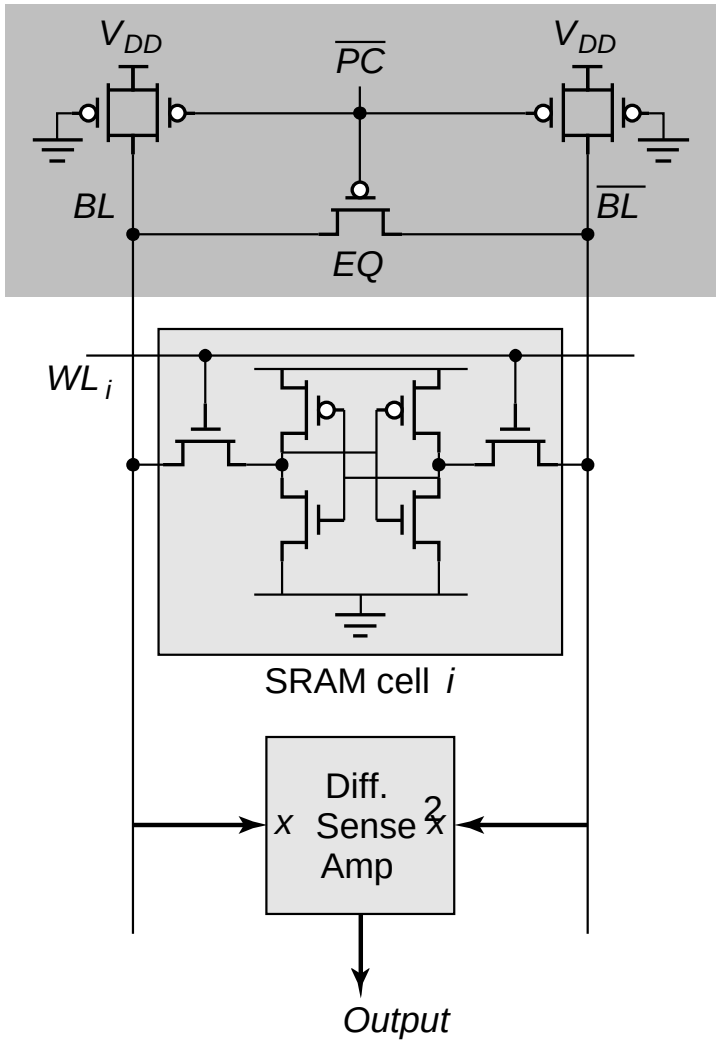
Idea: Use Sense Amplifer



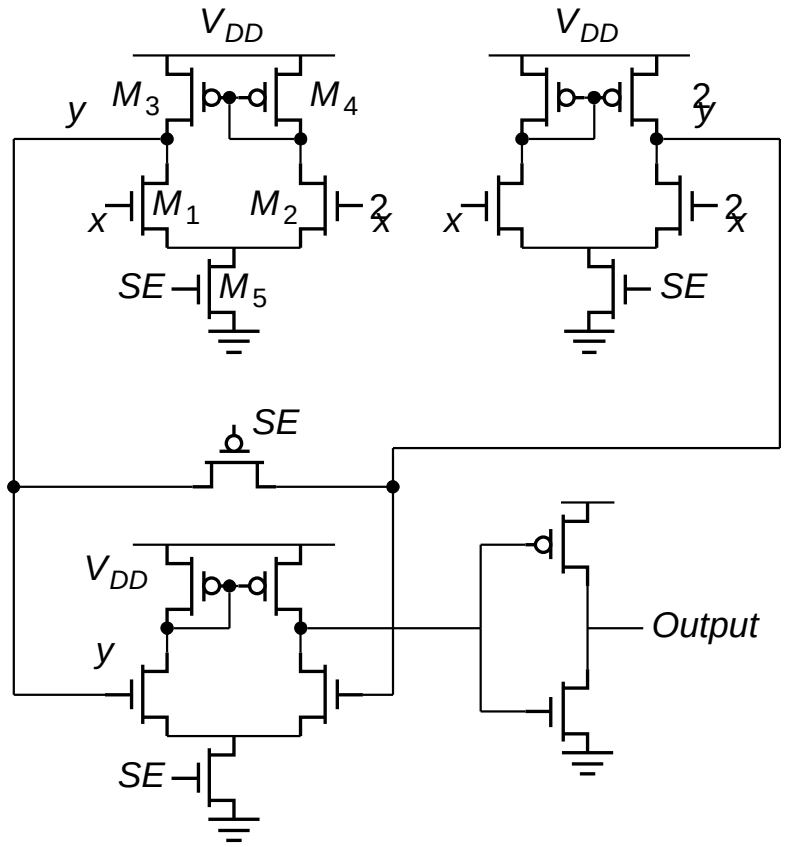
Differential Sense Amplifier



Differential Sensing — SRAM

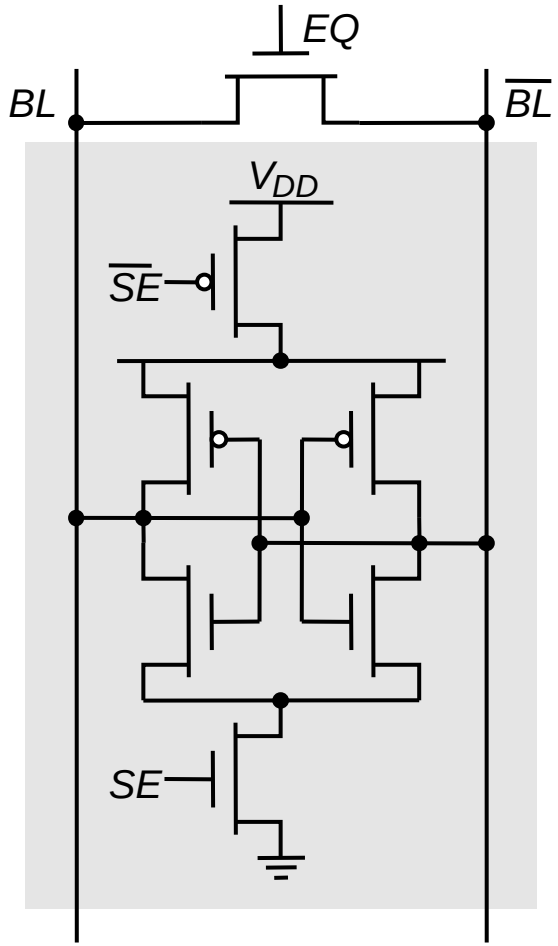


(a) SRAM sensing scheme



(b) two stage differential amplifier

Latch-Based Sense Amplifier



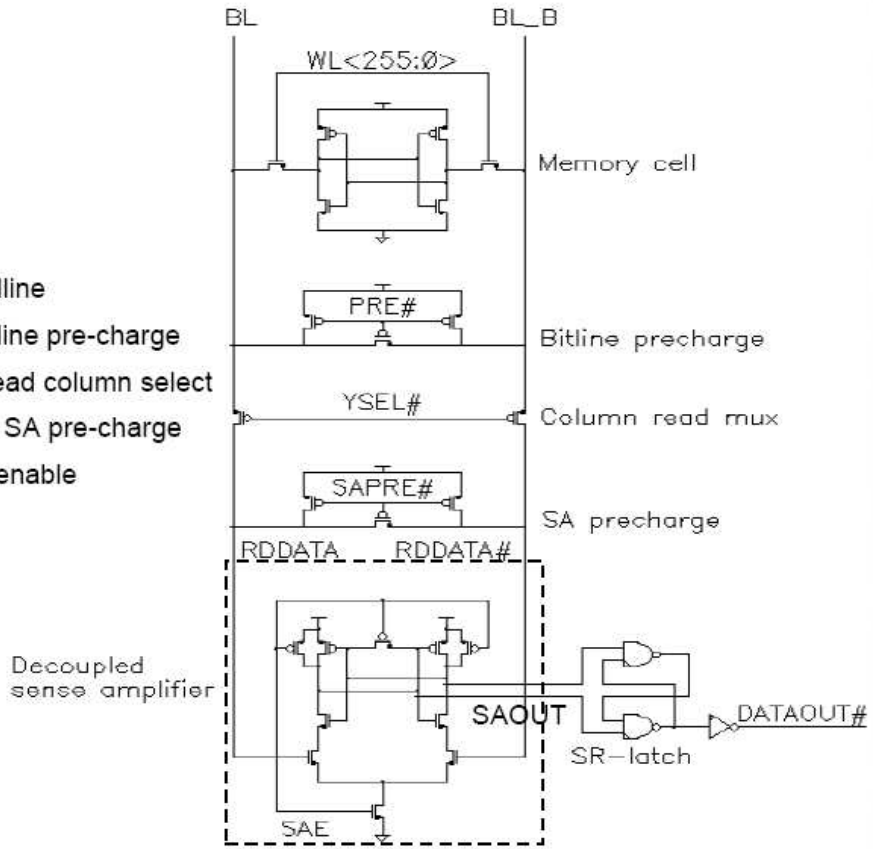
Initialized in its meta-stable point with EQ

Once adequate voltage gap created, sense amp enabled with SE

Positive feedback quickly forces output to a stable operating point.

Decoupled Sense Amplifier

WL: Wordline
PRE#: bitline pre-charge
YSEL#: read column select
SAPRE#: SA pre-charge
SAE: SA enable

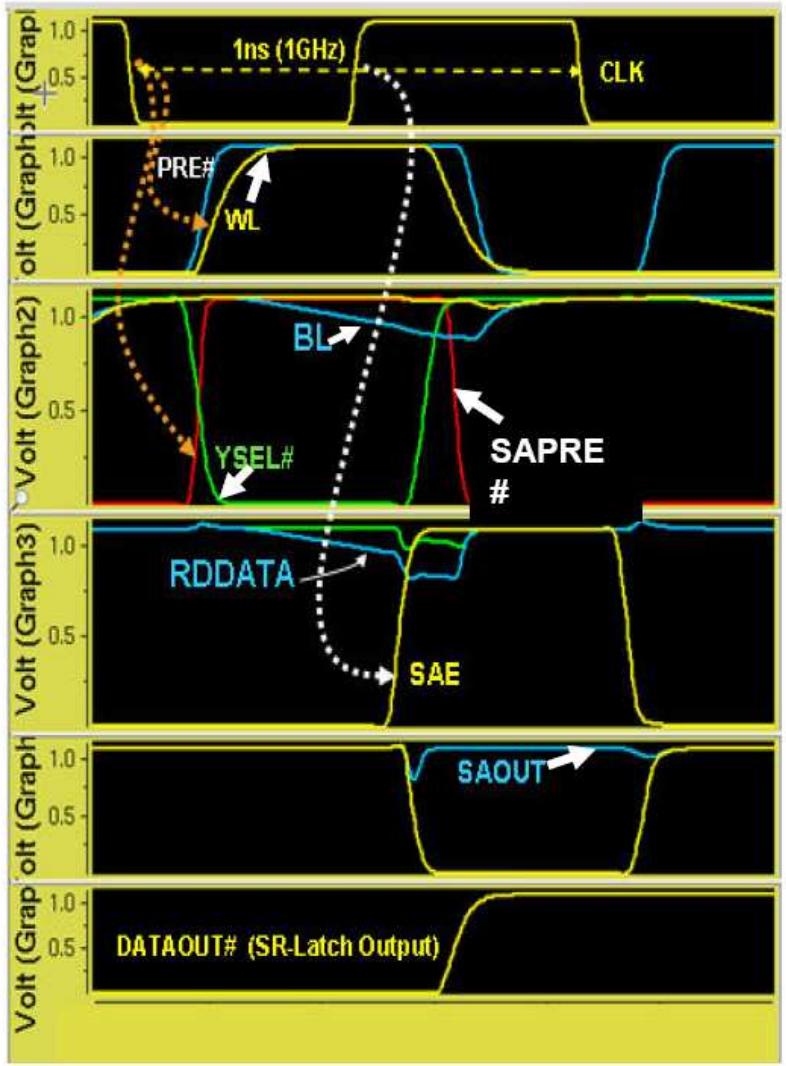


Low Phase

Input Latching
Address Decoding
WL Assertion
BL Development

High Phase

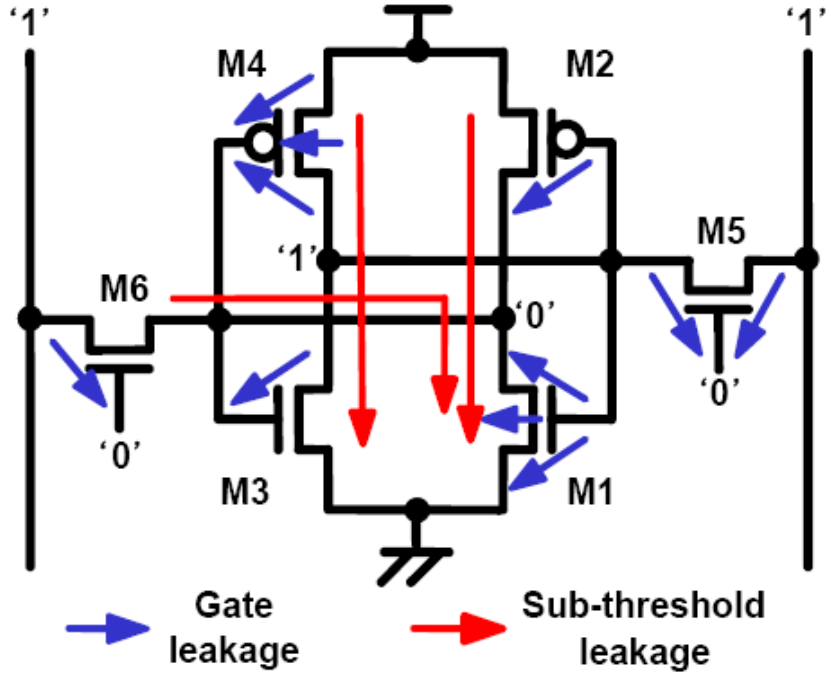
Sensing
Output Latching
Dataout Muxing
BL/SA Precharge

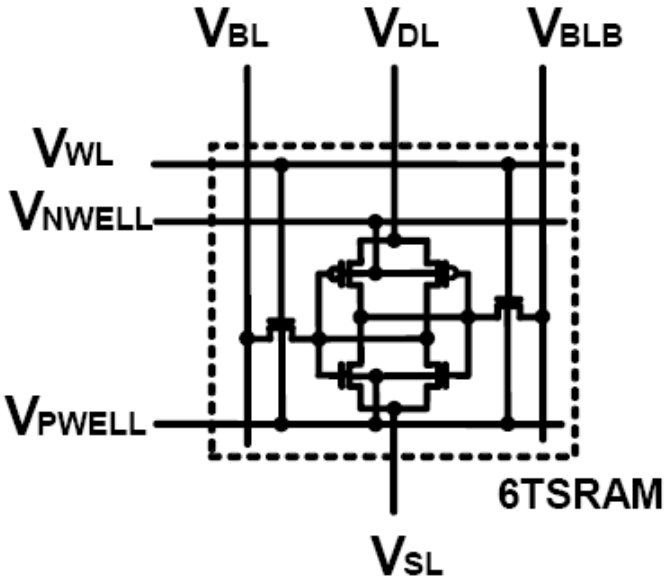


Simulation timing diagram

Y. Wang, Intel, ISSCC'07

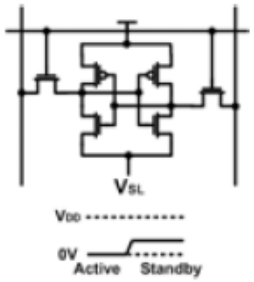
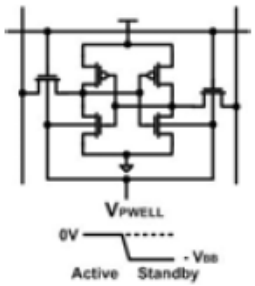
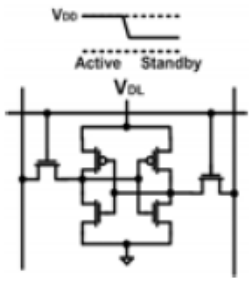
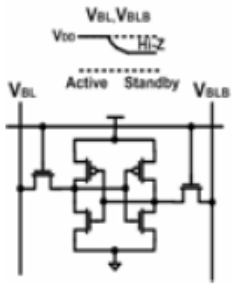
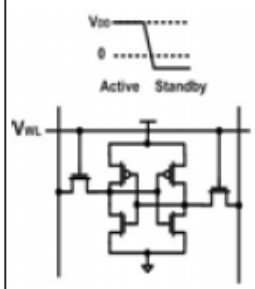
SRAM Cell Leakage



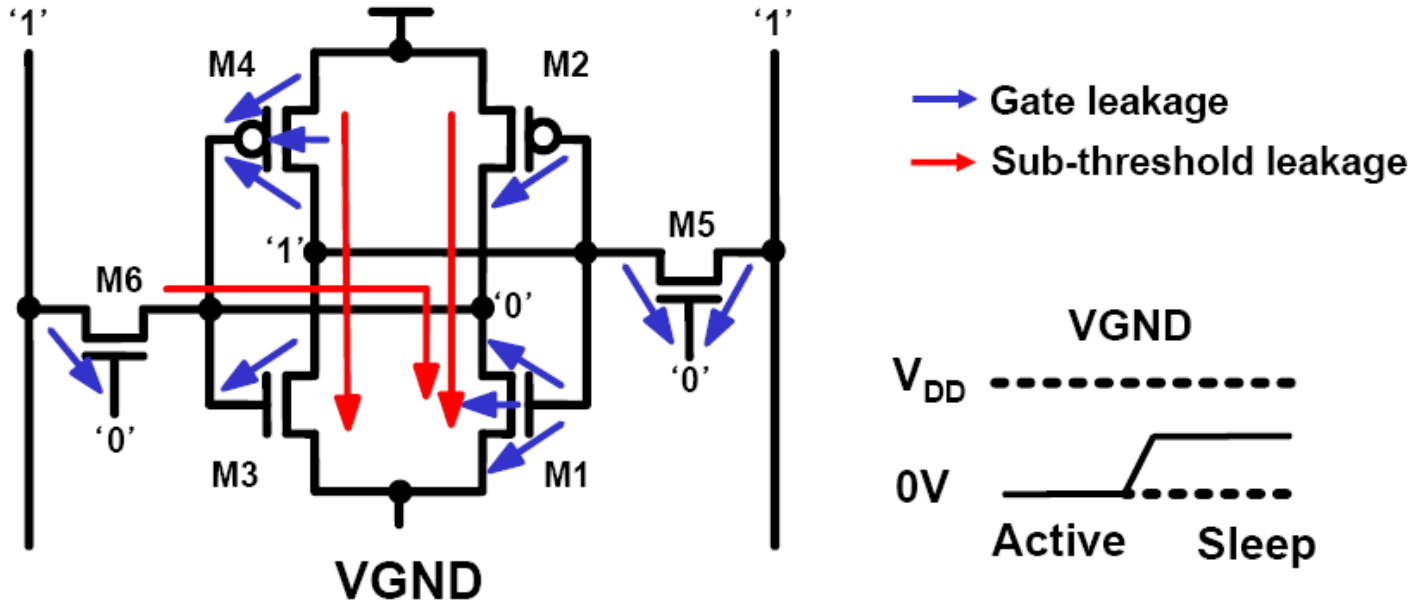


- Switch off cells into a low leakage state when they are not used
- 7 degrees of freedom to reduce 6T SRAM leakage

Leakage Reduction Schemes

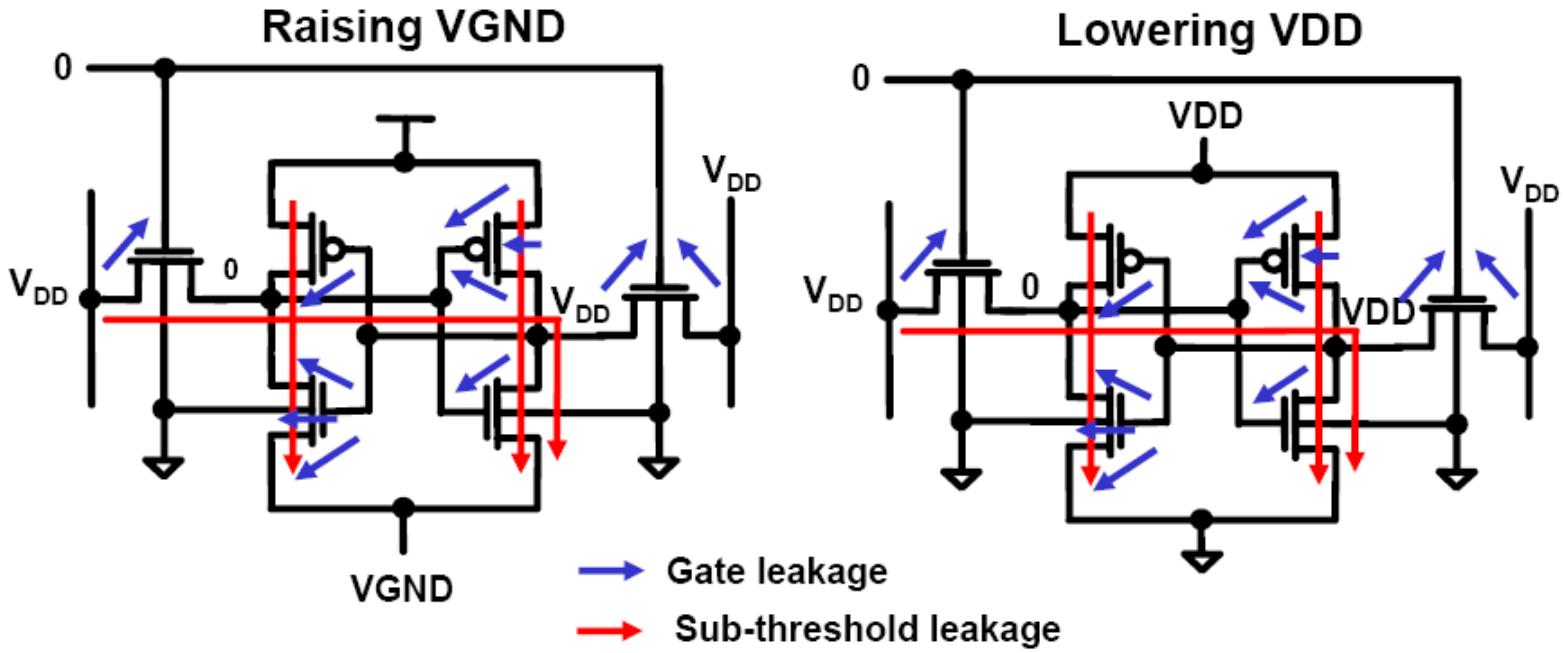
Schemes	Source Biasing (V_{SL})	Reverse Body-Biasing (V_{PWELL}, V_{NWELL})	Dynamic V_{DD} (V_{DL})	Floating Bitlines (V_{BL}, V_{BLB})	Negative Word Line (V_{WL})
					
Leakage reduction	Sub: ↓↓ Gate: ↓↓	Sub: ↓↓ *BTBT: ↑	Sub, gate: ↓ *Bitline leak: -	Sub: ↓ Gate: ↓	Sub: ↓ *Gate: ↑
Delay	*Delay increase	No delay increase	No delay increase	No delay increase	No delay increase
Overhead	Low transition overhead	Large transition overhead	Large transition overhead	*Precharge latency overhead	*Low charge pump efficiency
Stability	Impact on SER	No impact on SER	*Worst SER	No impact on SER	No impact on SER, voltage stress

Source Biasing SRAM Scheme



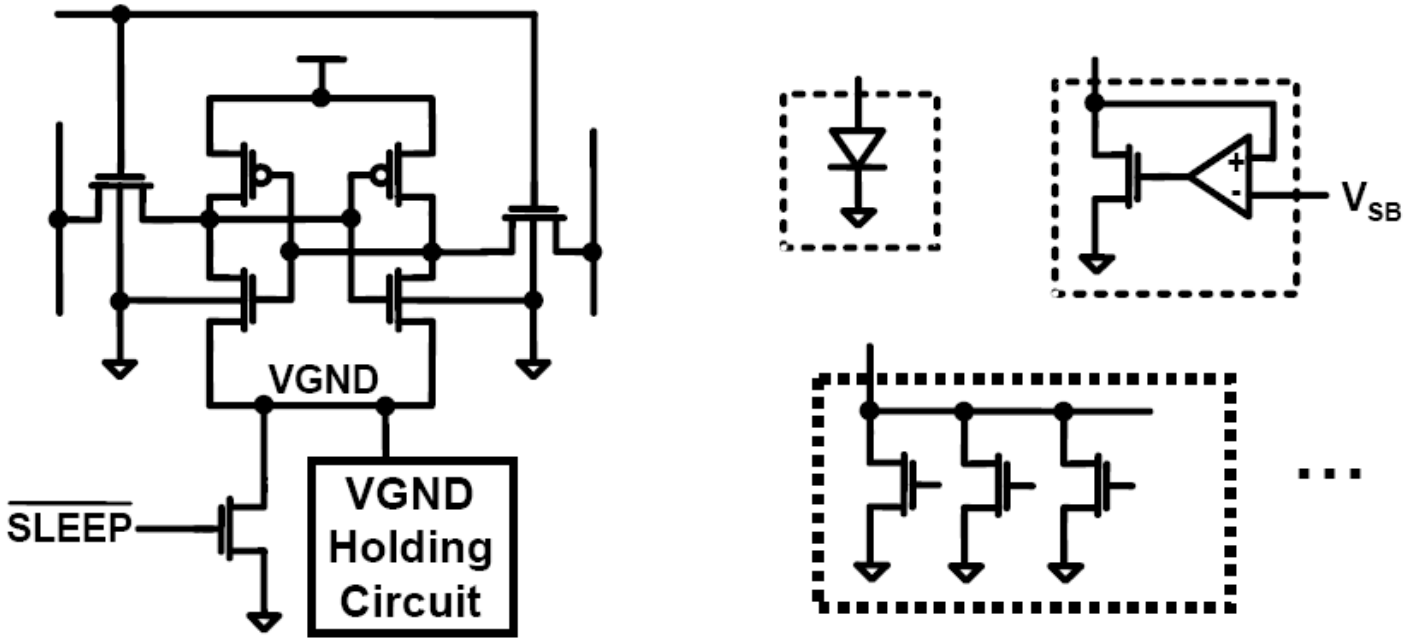
- Negative V_{GS} in access transistors
- Lower DIBL, body effect, less voltage stress
- Reduces both sub-threshold and gate leakage

Comparison with Dynamic V_{dd}



- Negative V_{GS} in access transistors
- Lower DIBL, body effect, less voltage stress
- Reduces both sub-threshold and gate leakage

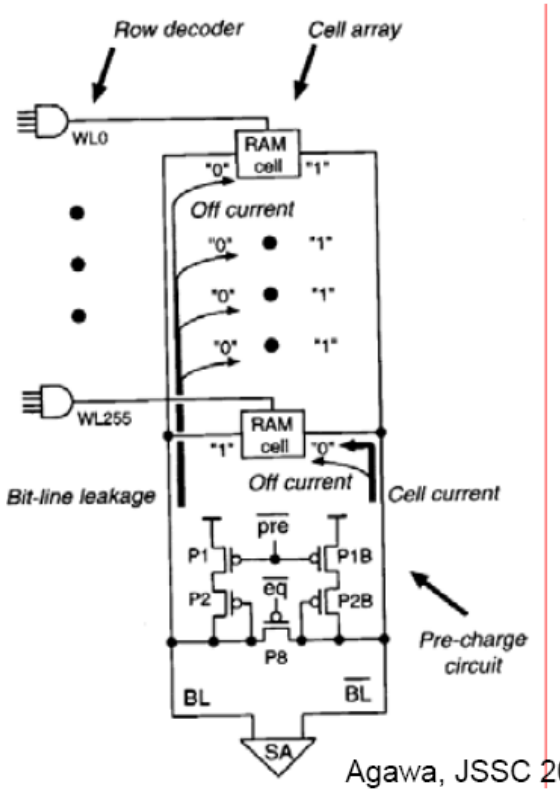
Data Retention During Sleep Mode



- Sleep transistor cuts off VGND from ground during sleep mode
- VGND is strapped using different circuit schemes

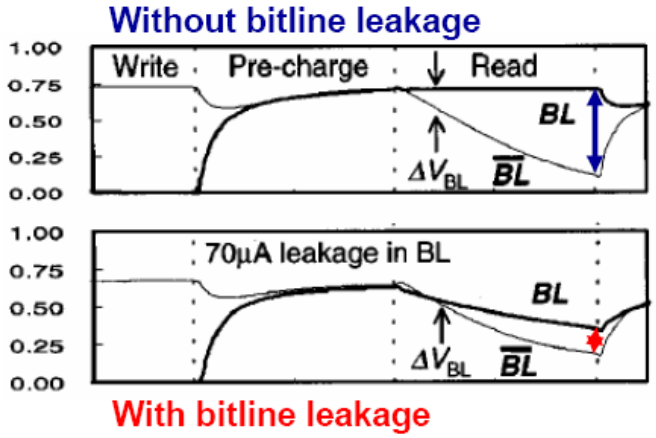
Bitline Leakage Problem

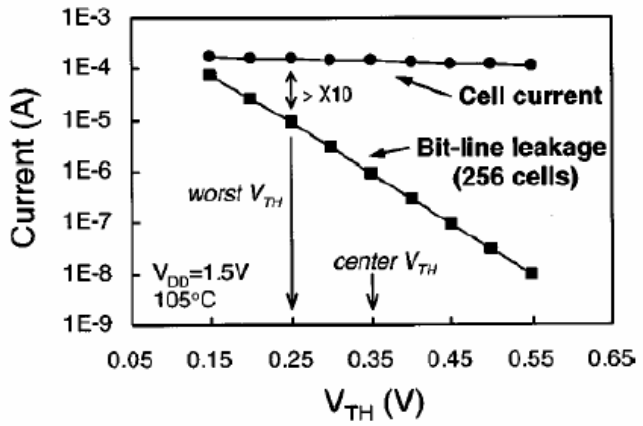
- Access TR leakage acts as noise current for sense amp
- Slow or incorrect read/write operation
- This is a functionality issue caused by increasing leakage
- Must consider worst case process and data condition
- Prohibits the use of long bitlines → impacts SRAM density



Agawa, JSSC 2001

Bitline Leakage Problem

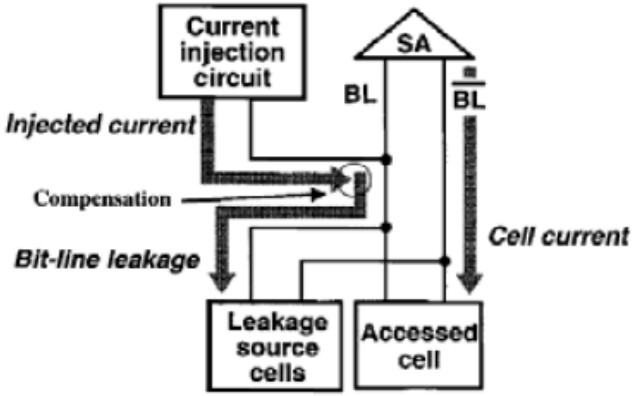




Agawa, JSSC 2001

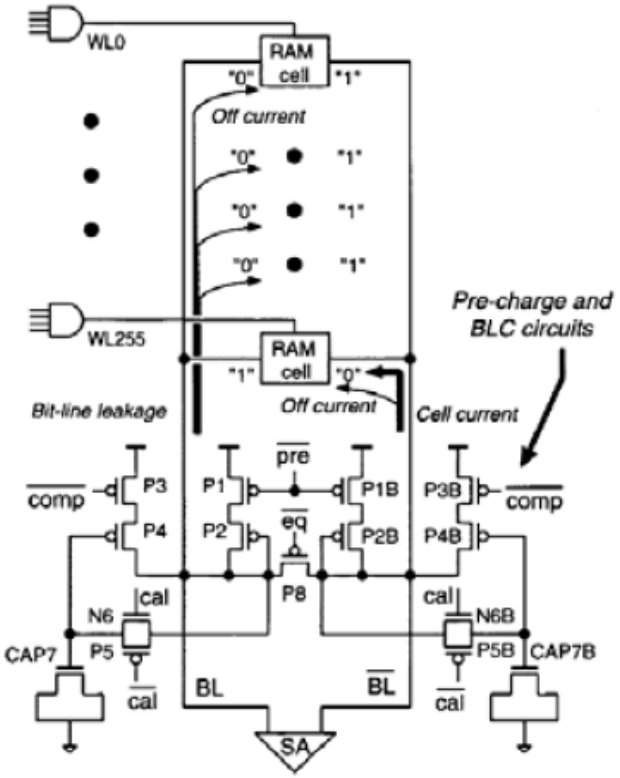
- Worst case bitline difference voltage is reduced
- This would mean that unlike in random logics, lower Vt can slow down performance in SRAMs
- Circuit solutions
  - Compensate leakage, reduce leakage, or equalize leakage

Bitline Leakage Compensation



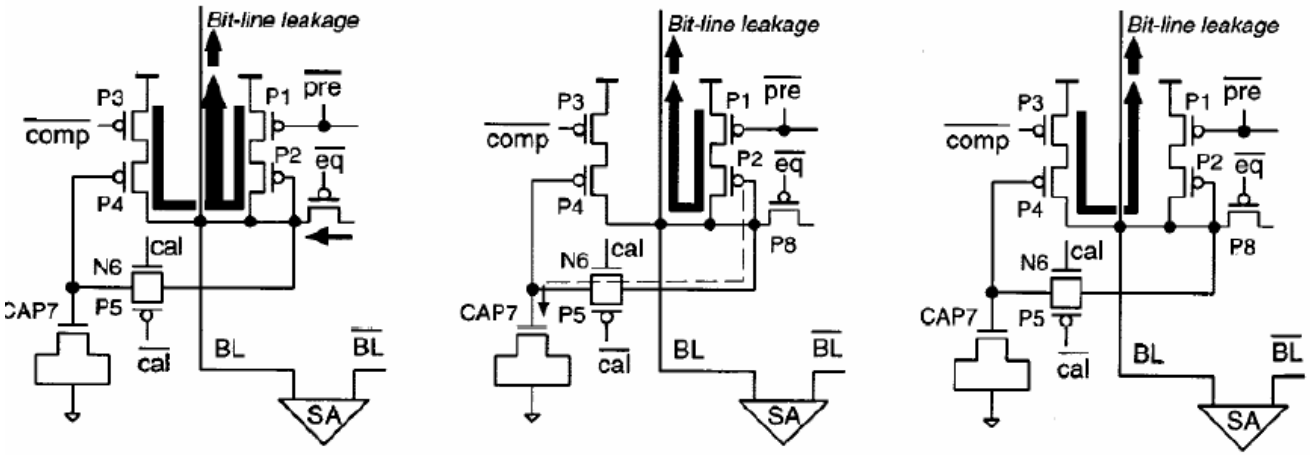
Concept

- Precharge phase: Measure the amount of bitline leakage during
- Evaluation phase: Duplicate and inject the same amount of current into the bitline

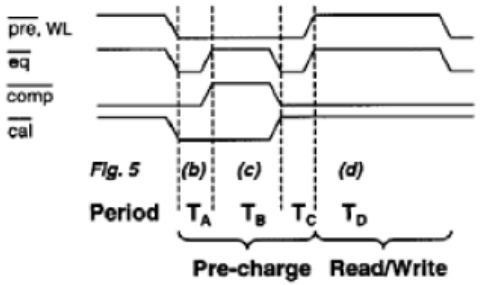


Agawa, JSSC 2001

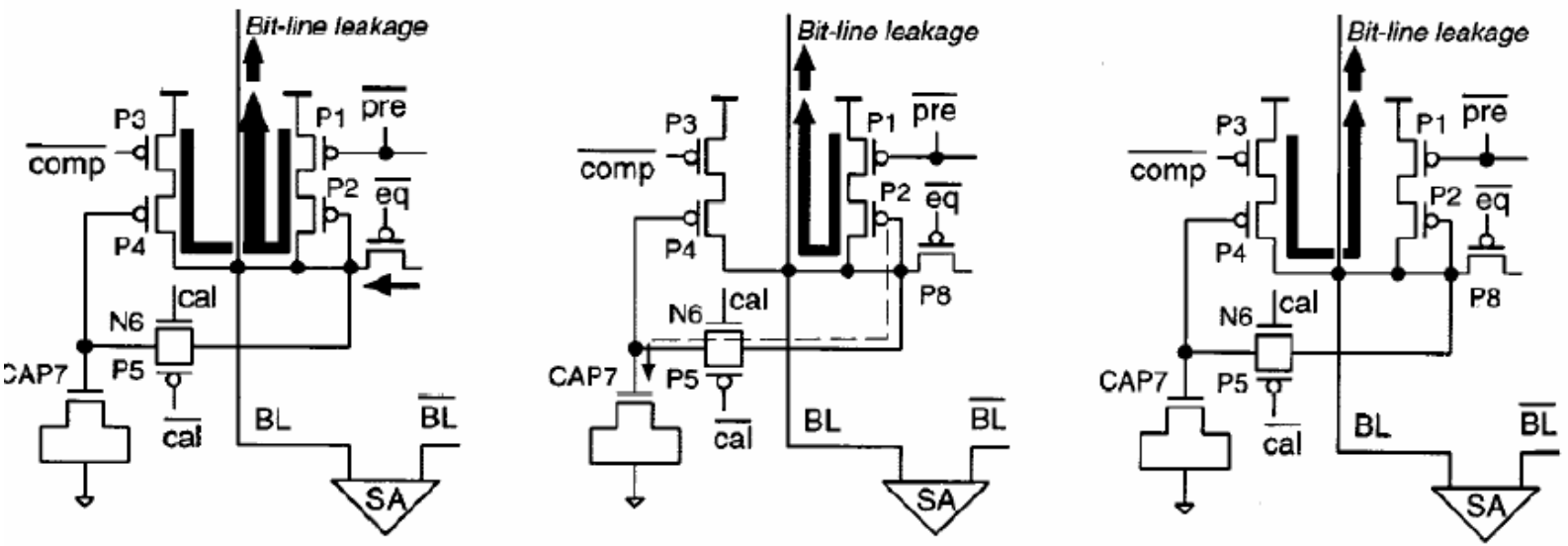
Bitline Leakage Compensation (cont'd)



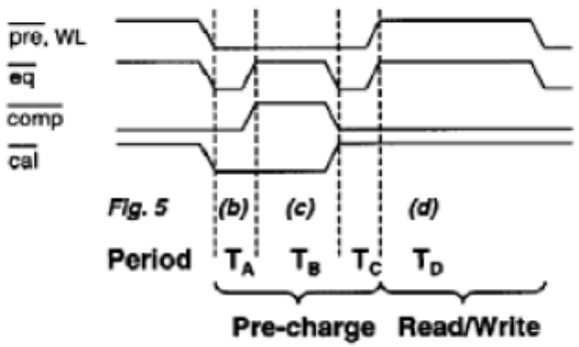
- **Precharge using P2, P4**
- **Measure leakage using P2**
- **Compensate during read/write cycle using P4**



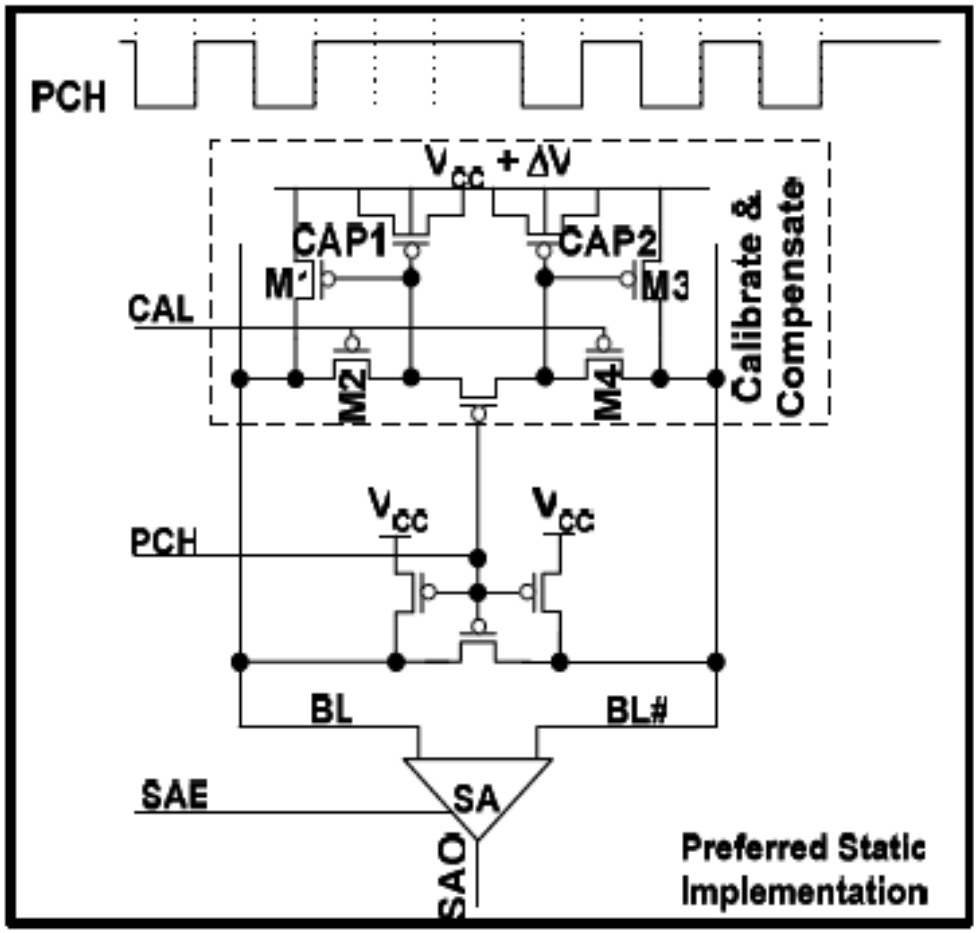
Bitline Leakage Compensation (cont'd)



- Precharge using P2, P4
- Measure leakage using P2
- Compensate during read/write cycle using P4



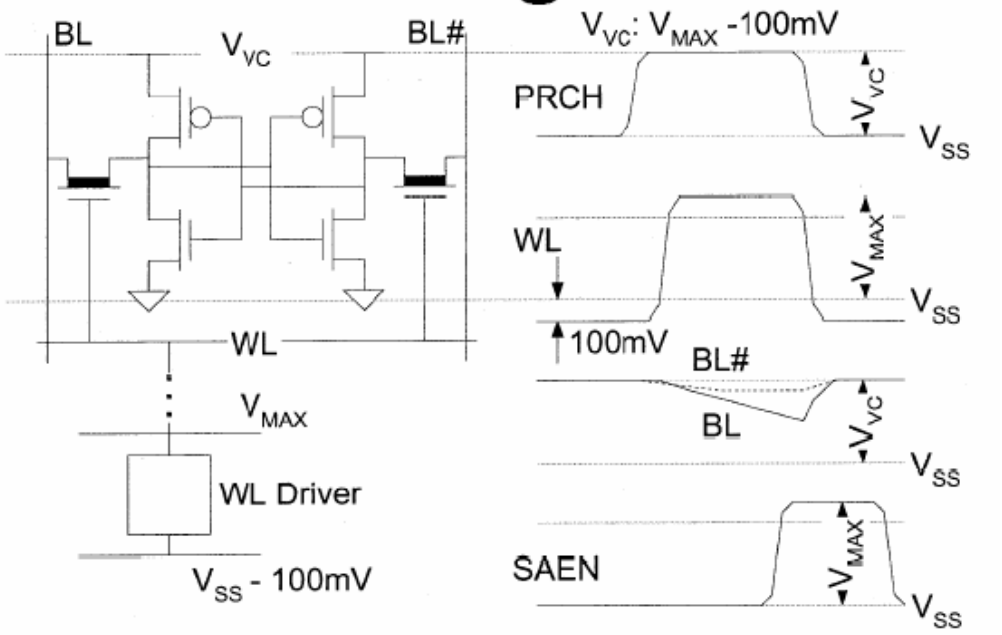
Bitline Leakage Compensation - II



- Less transistor count
- Easy to implement in within column pitch
- No complicated timing
- Calibration after every write

M. Khellah, Intel, VLSI'05

Bitline Leakage Reduction

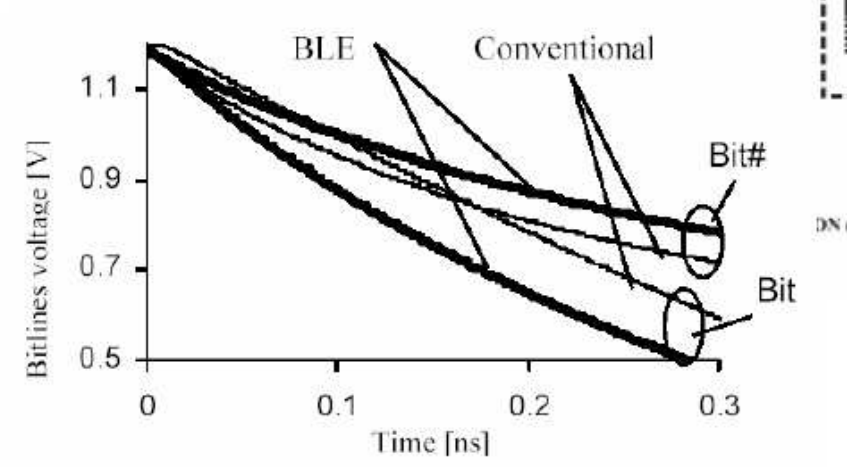


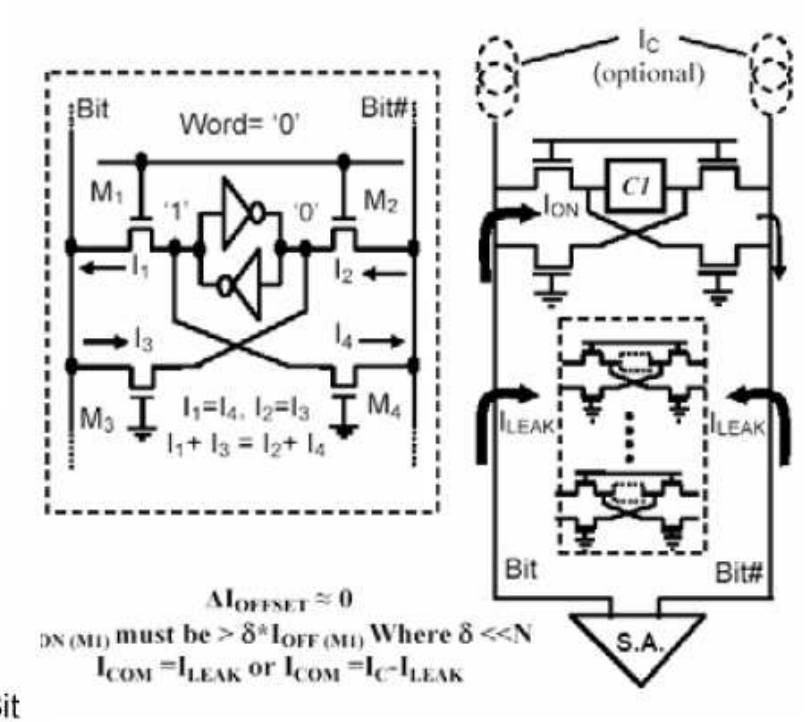
Ye, VLSI 2003

- **Wordline underdrive during idle mode to reduce leakage through access TR**
- **Low V_t access transistor for fast read**
- **Cell and precharge voltage lower than global V_{cc} (=V_{max})**
  - **This ensures that V_{gs} and V_{ds} doesn't exceed V_{max}**

Bitline Leakage Equalization

- 8T SRAM cell with 40% area hit
- Dummy transistors M3 and M4 inject same amount of leakage to each bitlines
- Both bitlines discharge together due to equalized bitline leakage
- Less practical due to area impact





Alvandpour, ESSCIRC 2003