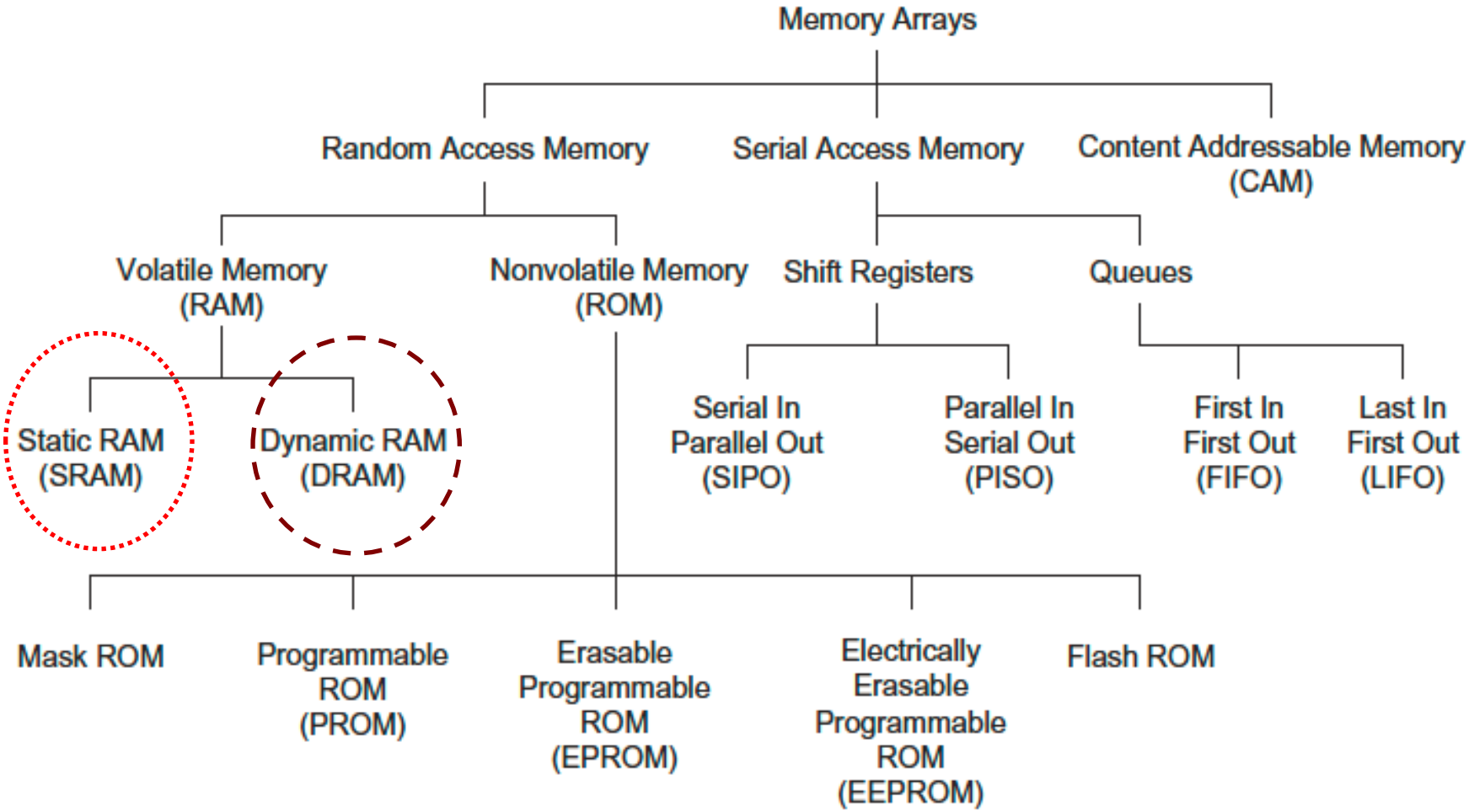
Learning Objectives for SRAM

- ❑ Articulate memory hierarchy and the value proposition of SRAMs in the memory chain + utilization in current processors
- ❑ Explain SRAM building blocks and peripheral operations and memory architecture (with physical arrangement)
- ❑ Articulate commonly used SRAM cells (6T vs 8T), their advantages and disadvantages
- ❑ Explain the operation of a non-conventional SRAM cells, and their limitations
- ❑ Explain commonly used assist methods
- ❑ Explain how variations impact memory cells

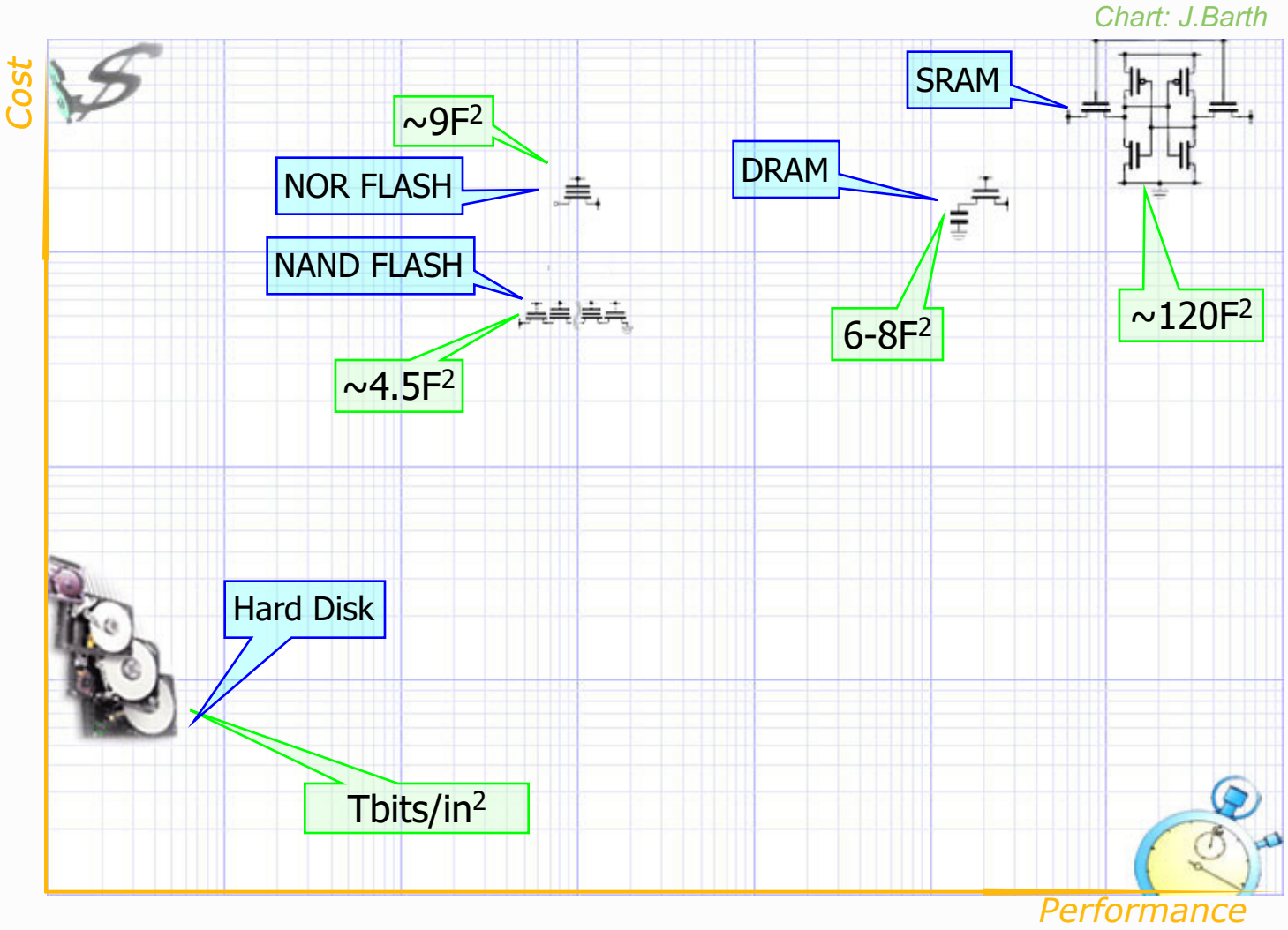
Topics

- Introduction to memory
- SRAM : Basic memory element
- Operations and modes of failure
- Cell optimization
- SRAM peripherals
- Memory architecture and folding

Memory Classification revisited

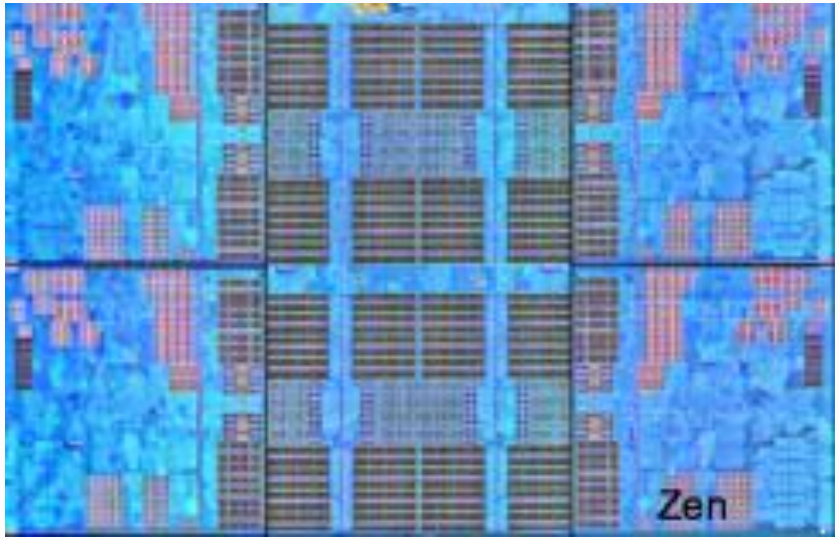


Technology choices for memory hierarchy

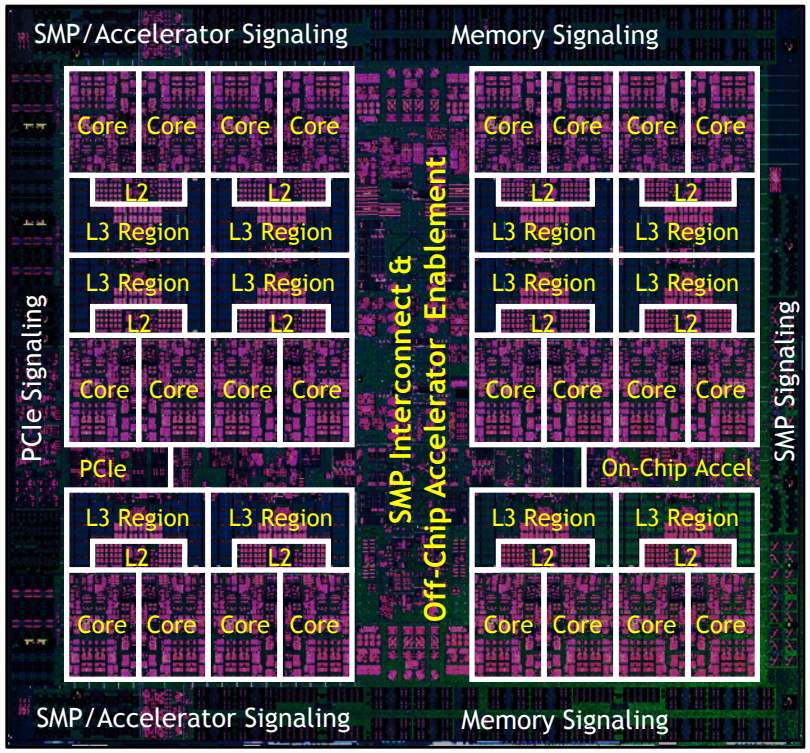


Cache Sizes

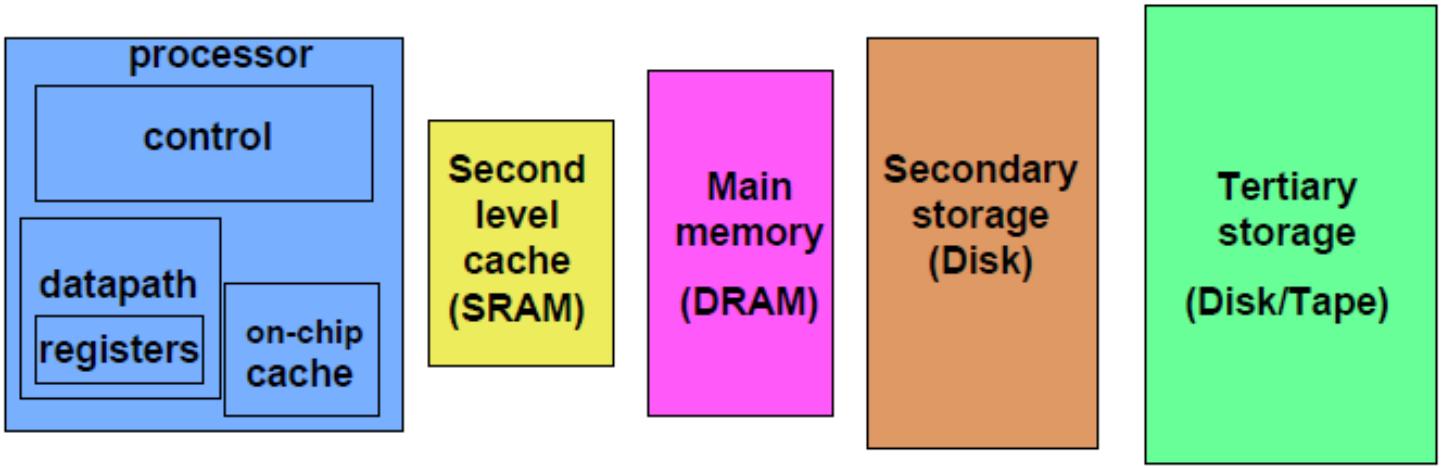
AMD Zen
ISSCC 2017
512kB L2$,
8MB L3$



IBM POWER9
Hotchips 2016
32kB L1 I$ and D$,
120MB e-DRAM L3$



Cache size impacts cycles-per-instruction



Speed	1ns	10ns	100ns	10ms	10sec
Size	B	KB	MB	GB	TB

For a 5GHz processor, scale the numbers by 5x

Several memory blocks in a typical processor core: I$, D$, Address translation tables, Branch history tables, all in the KB – low MB range

Question 1

- SRAMs are preferred for L1 Cache over eDRAMs

a) Refresh operation of e–DRAM causes undesirable performance penalty

b) e-DRAMs is too slow for high speed operation

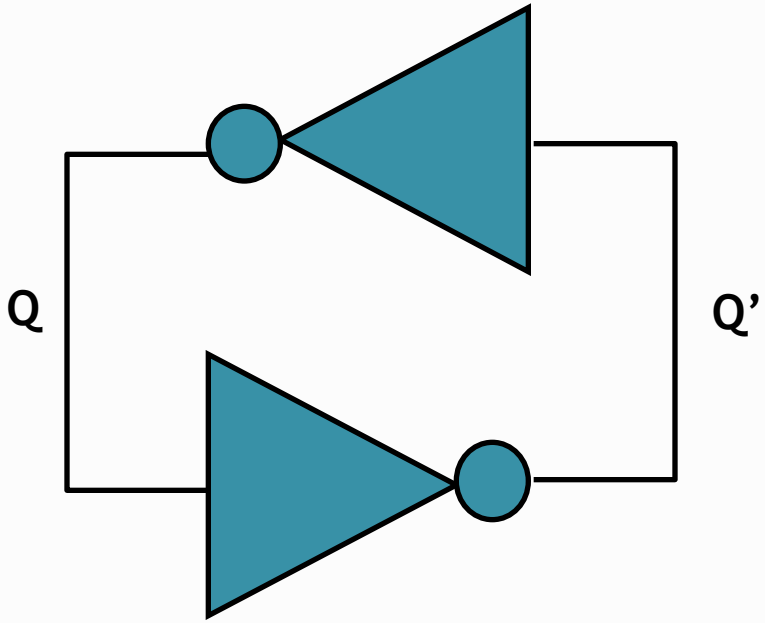
c) SRAMs are compact and can easily fit near functional blocks

d) e-DRAMs requires multiple supply voltages

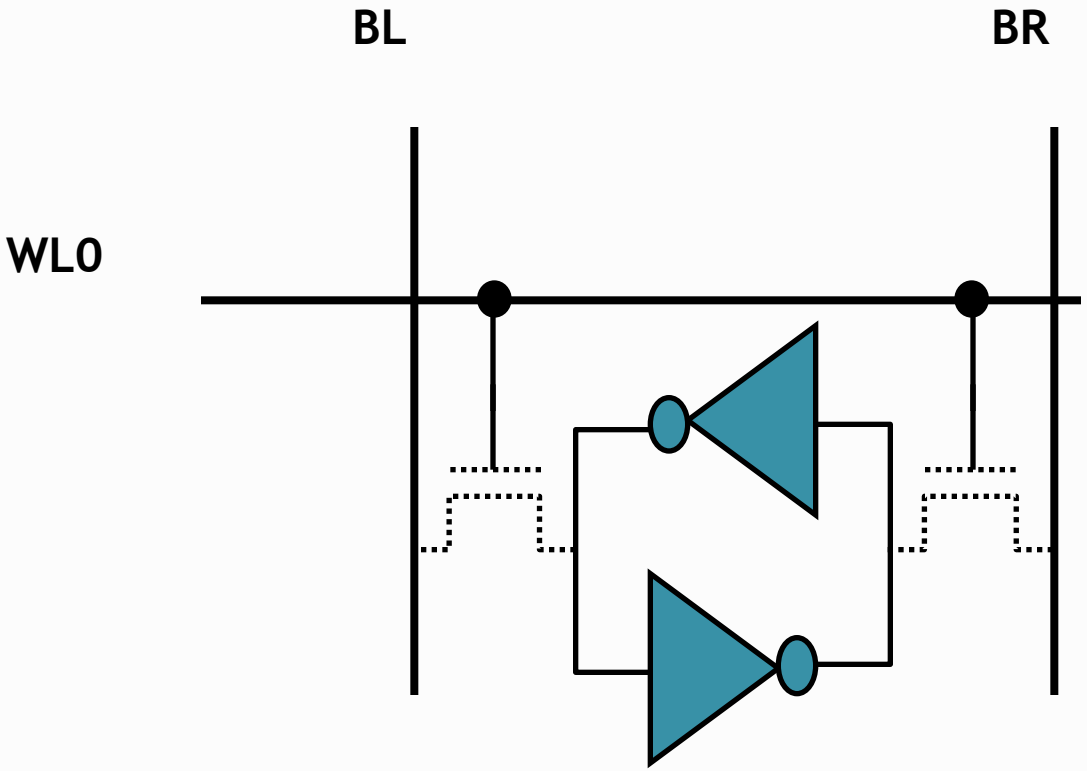
Topics

- Introduction to memory
- SRAM : Basic memory element
- Operations and modes of failure
- Cell optimization
- SRAM peripherals
- Memory architecture and folding

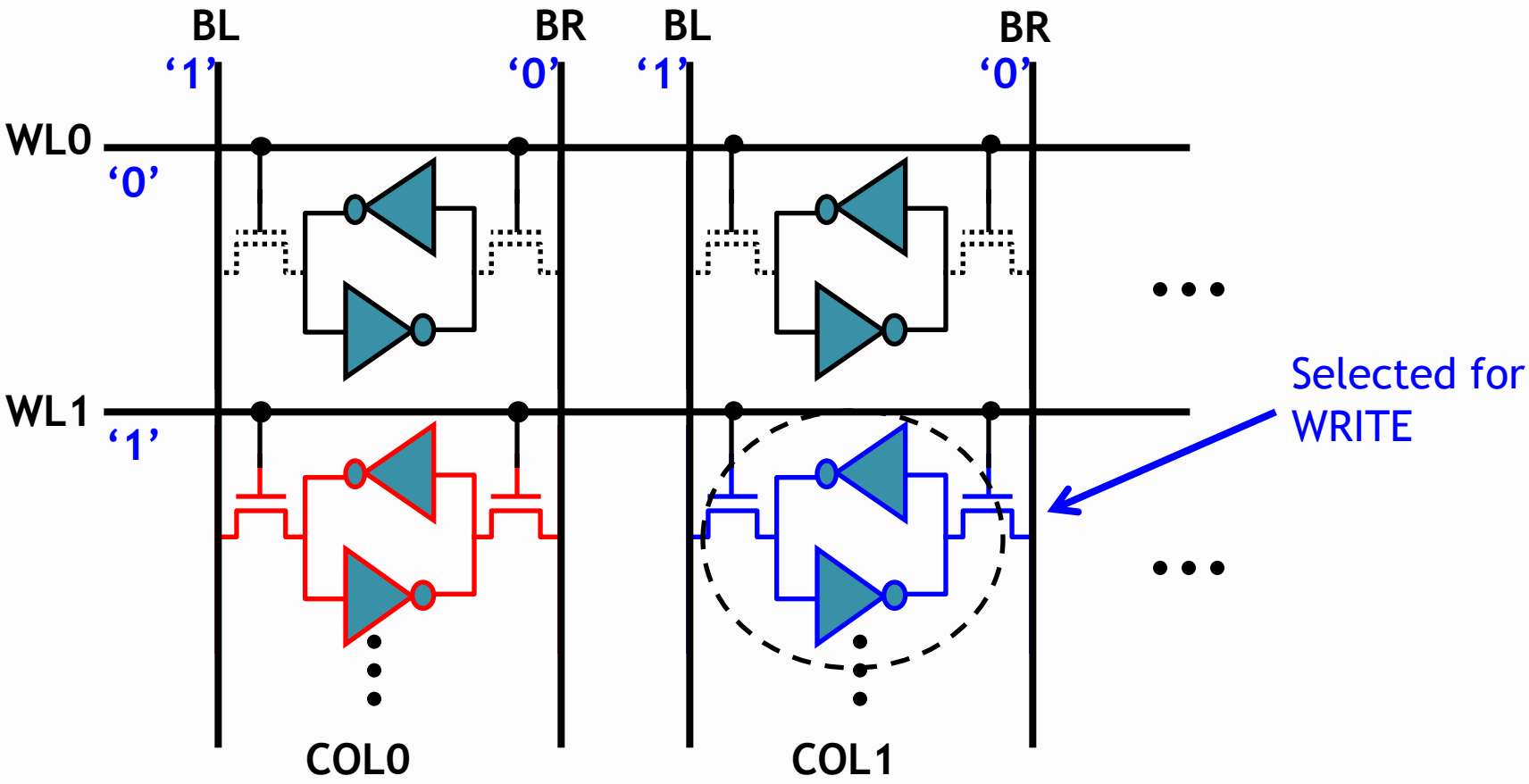
Basic Memory Element



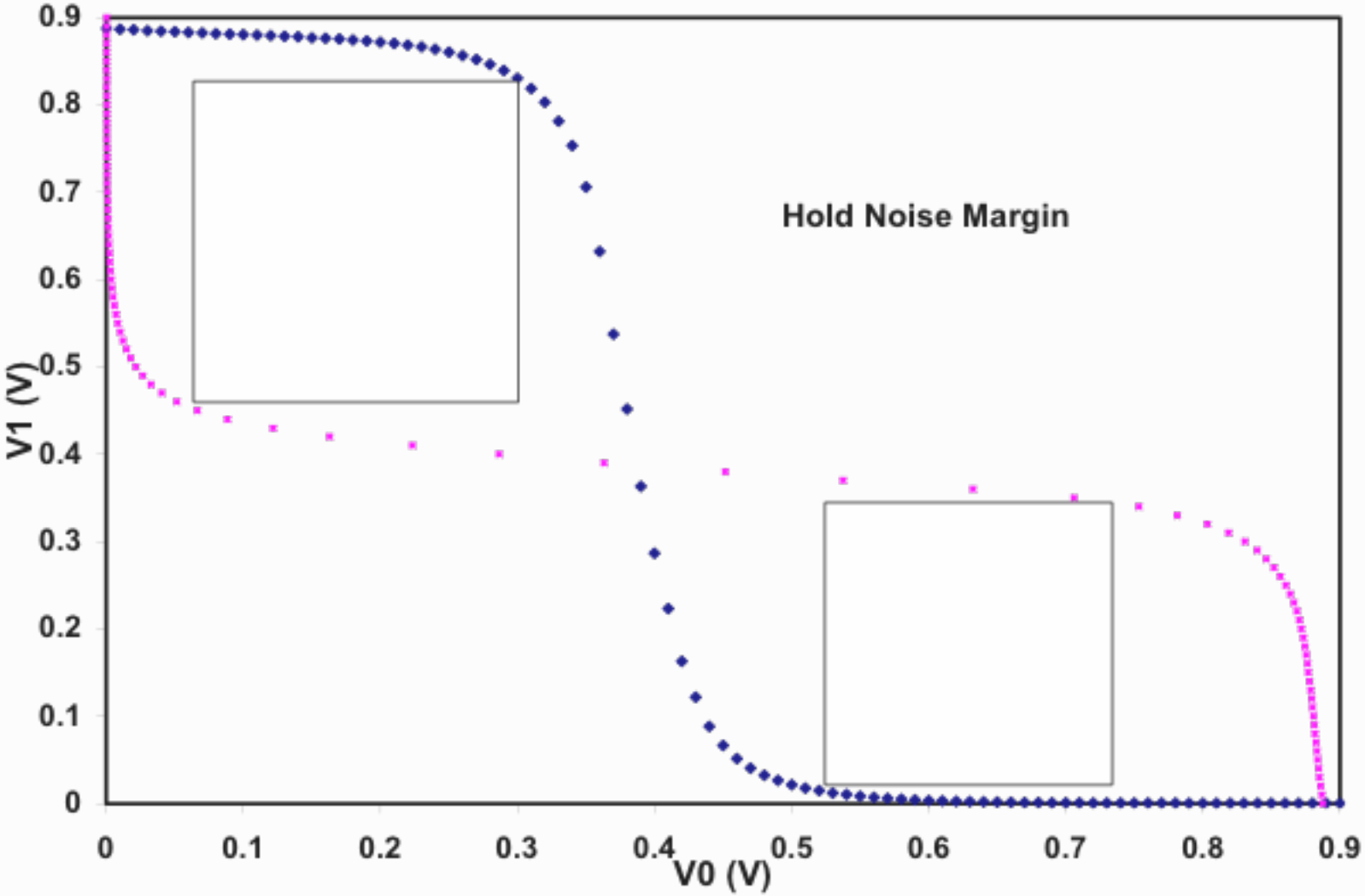
Access ways



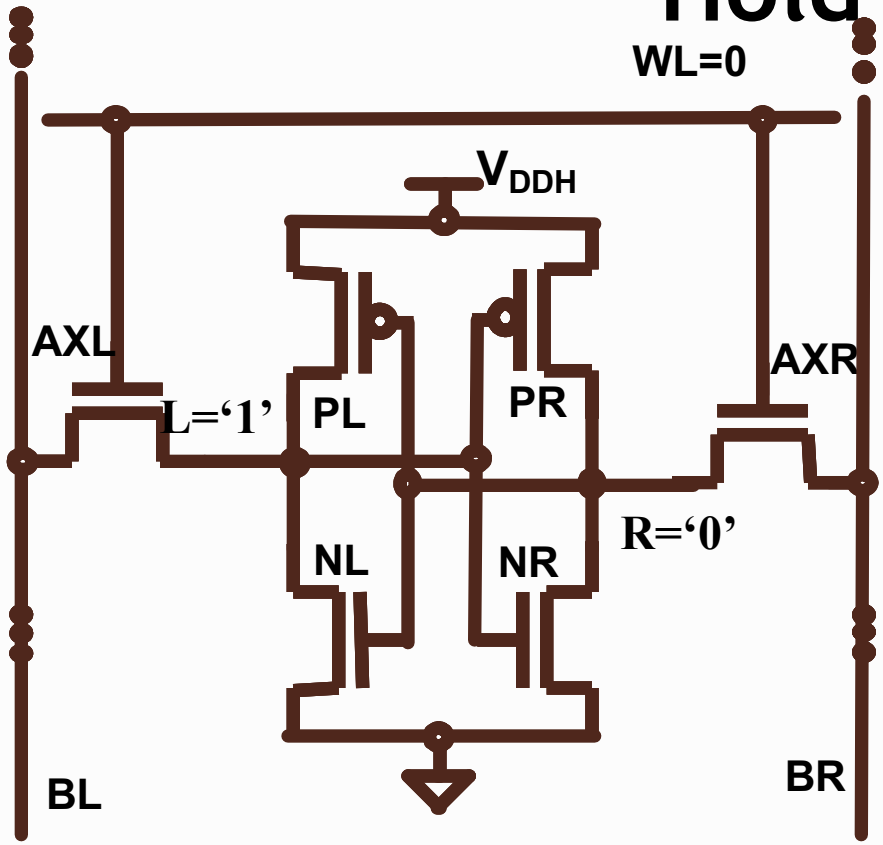
No Operation (Hold)



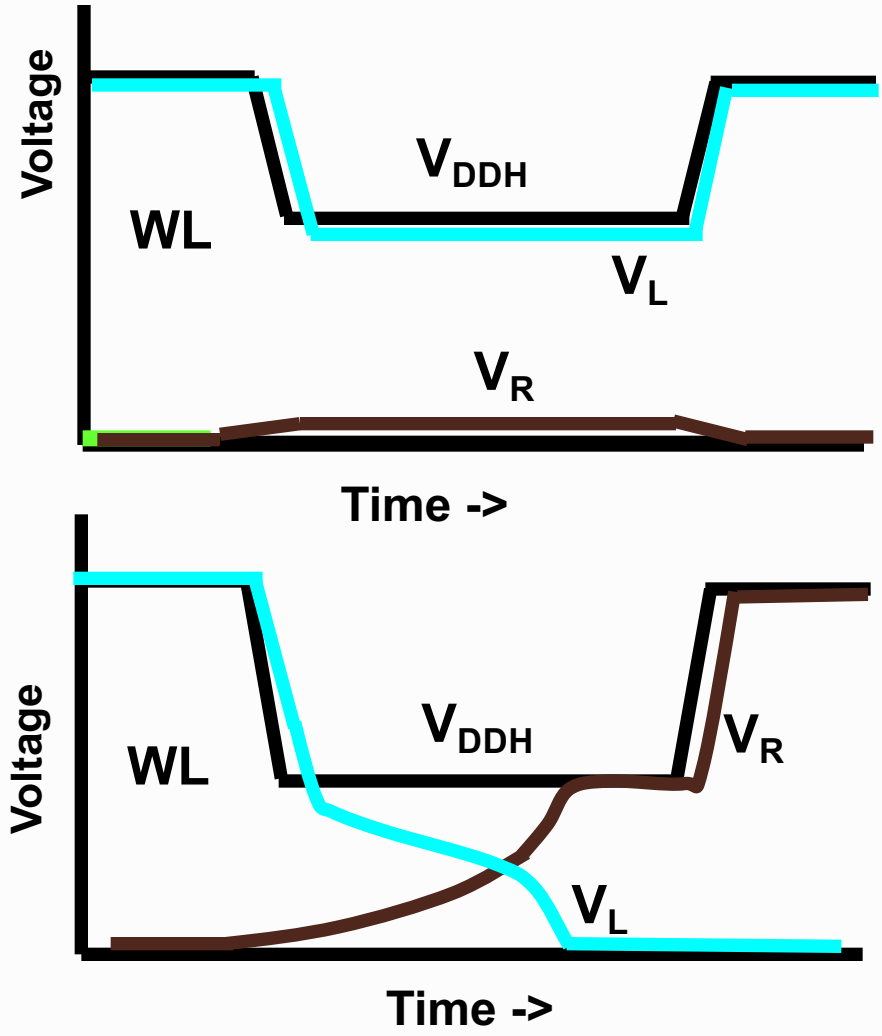
Hold Margin



Hold Failure



$$P_{HF} = P\left(V_{DDHmin} > V_{HOLD}\right)$$



Hold Failure => Flipping of cell data in the Hold mode with the application of a lower supply voltage.

Question 2

- Hold Margin for an SRAM cell

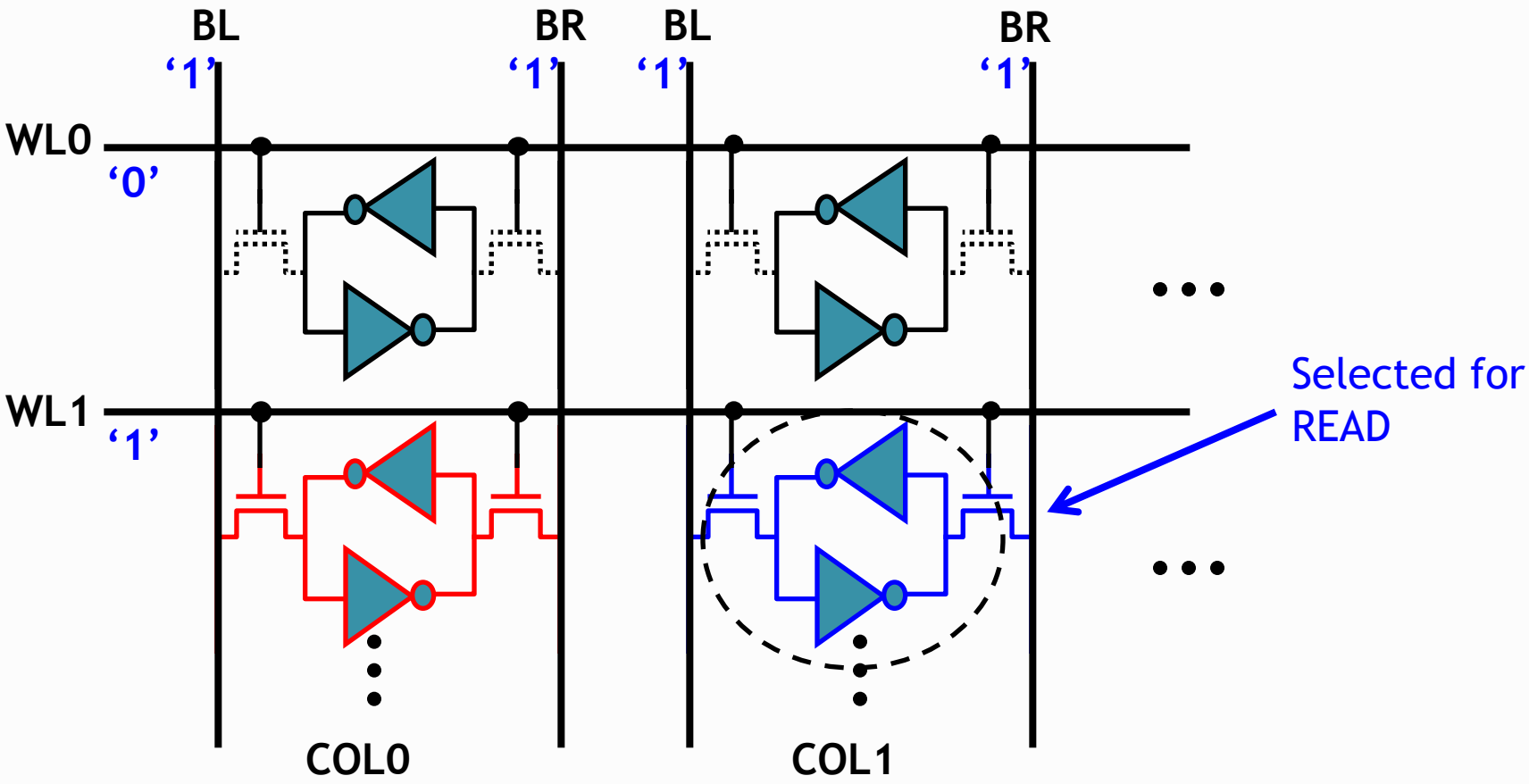
a) Is always greater than read margin

b) Can be improved by making the access transistor bigger

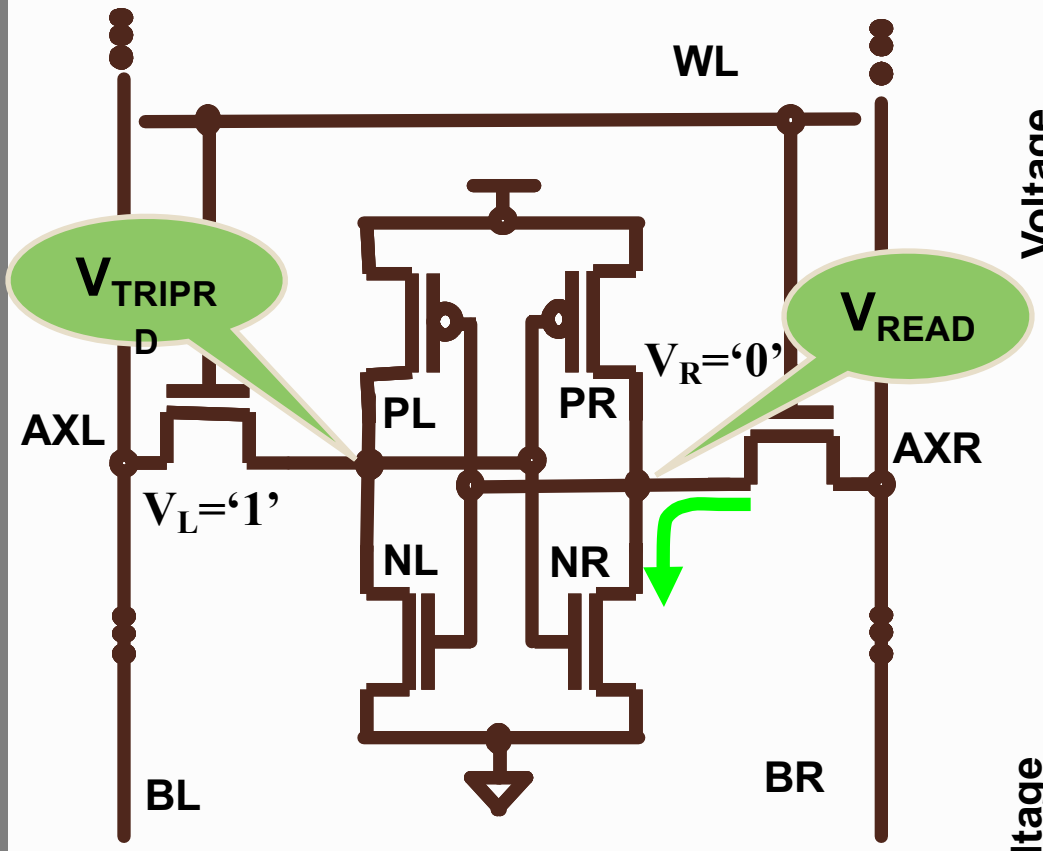
c) Defines the minimum supply voltage required to perform a read operation

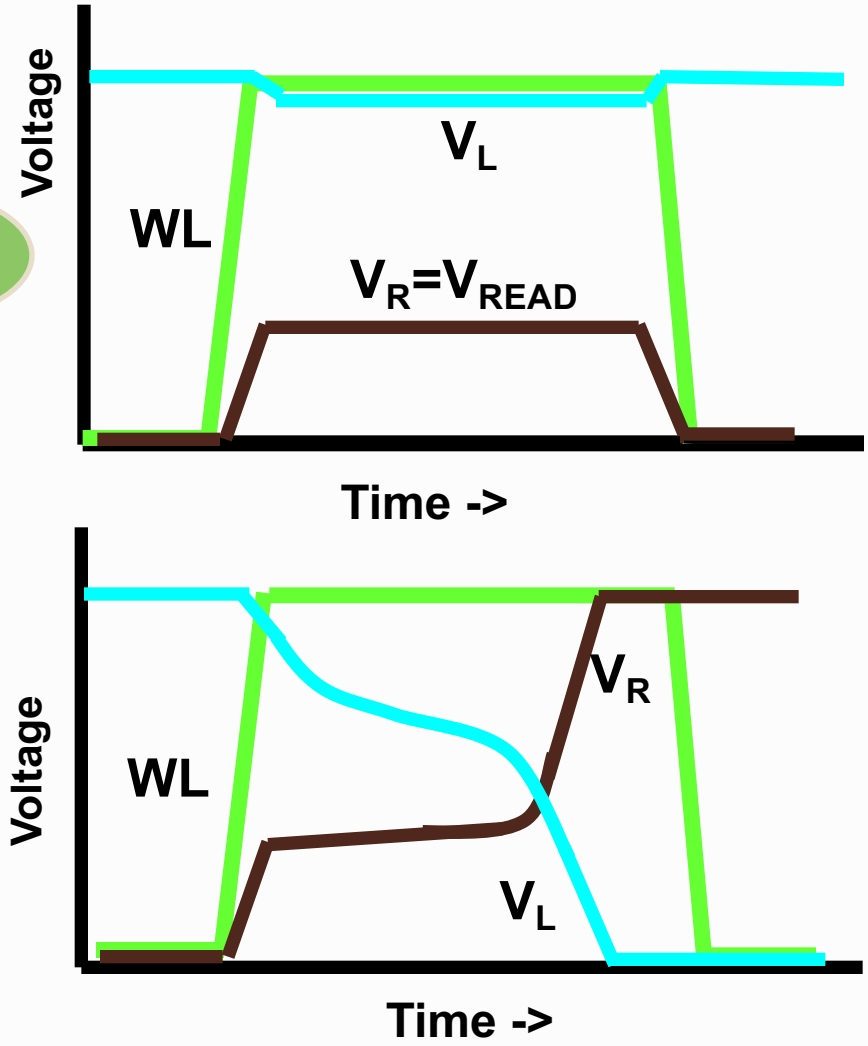
d) Depends on the number of cells on the bit line

Read Operation



Read Failure

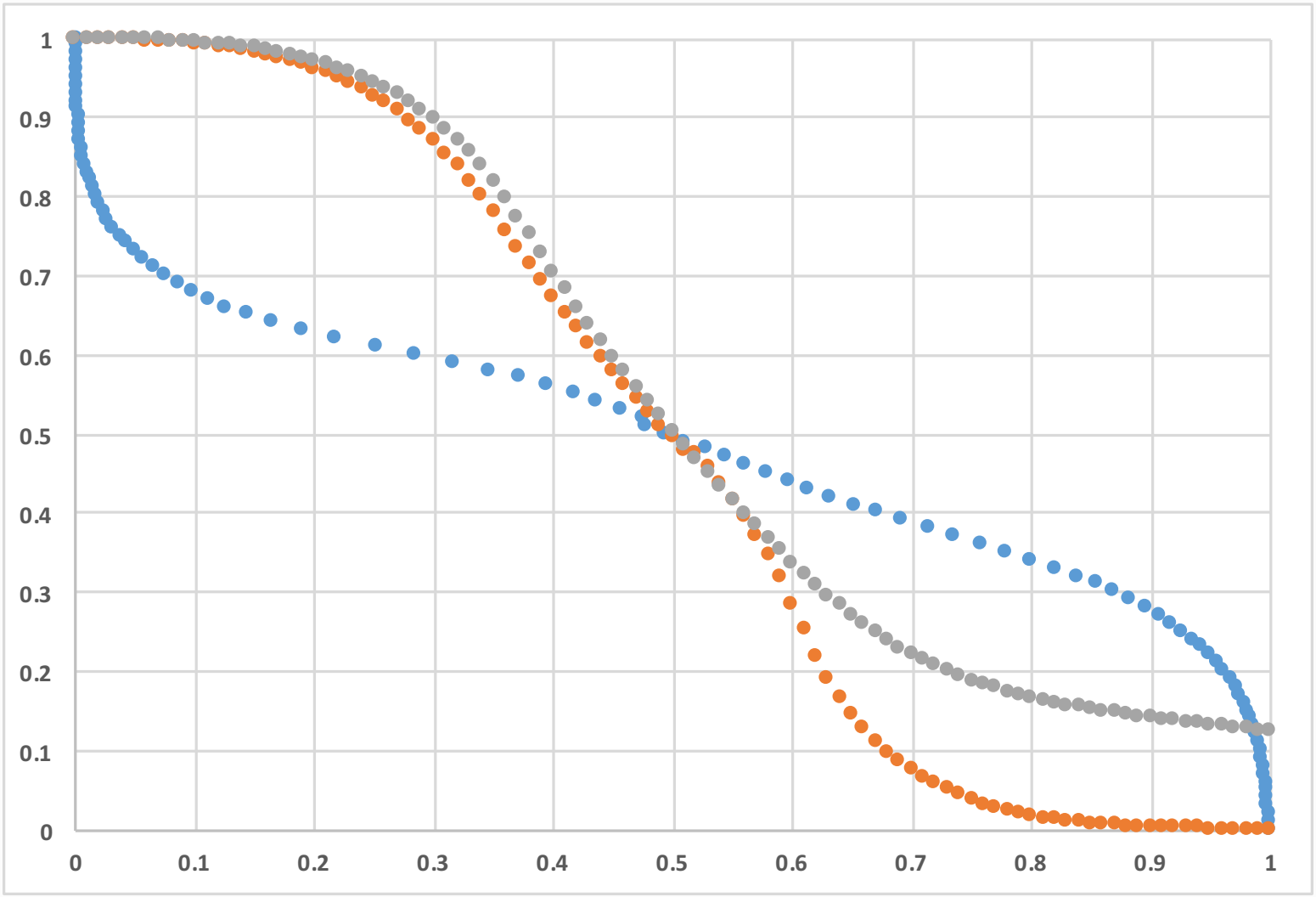




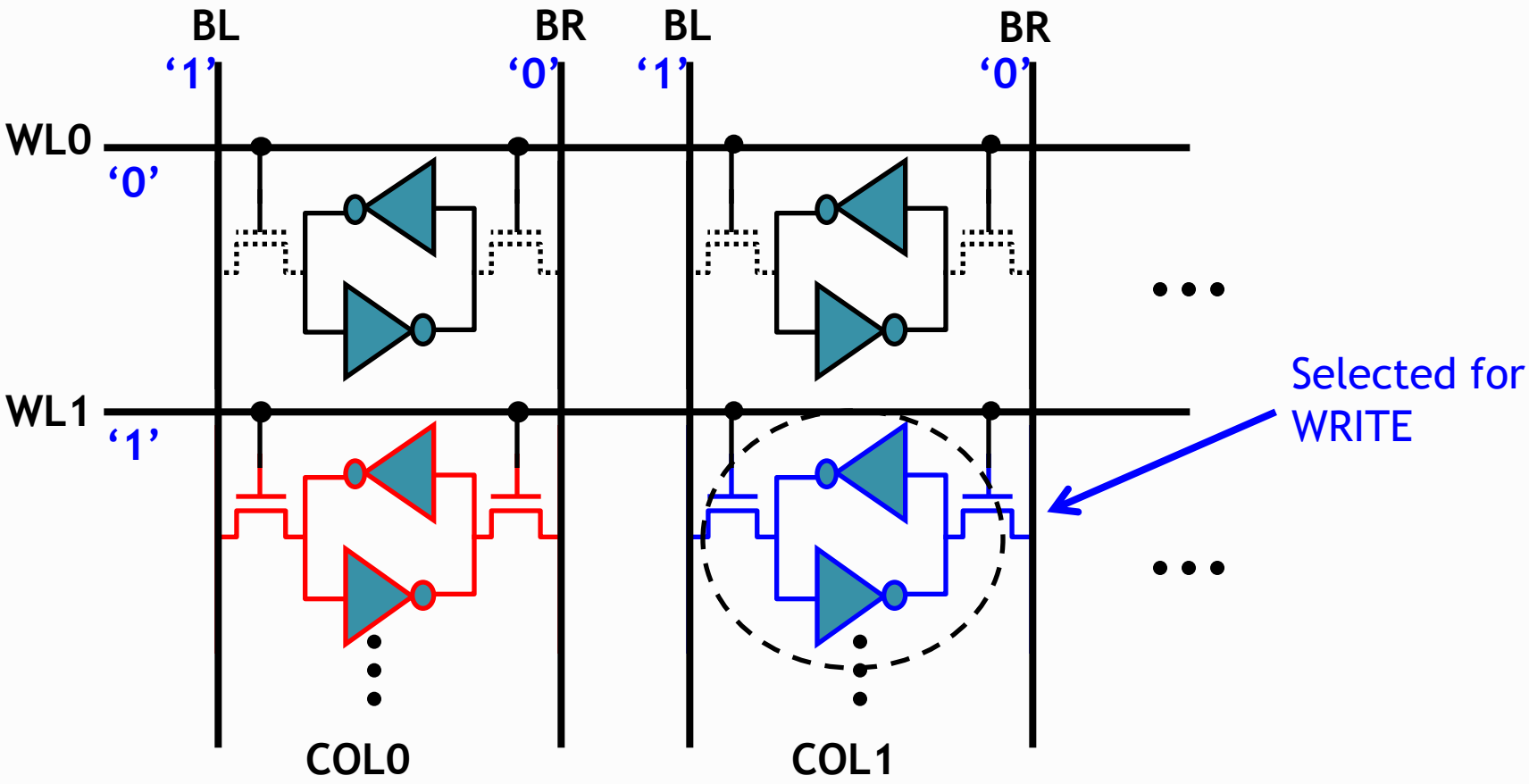
$$P_{RF} = P\left(V_{READ} > V_{TRIPRD}\right)$$

Read failure => Flipping of Cell Data while Reading

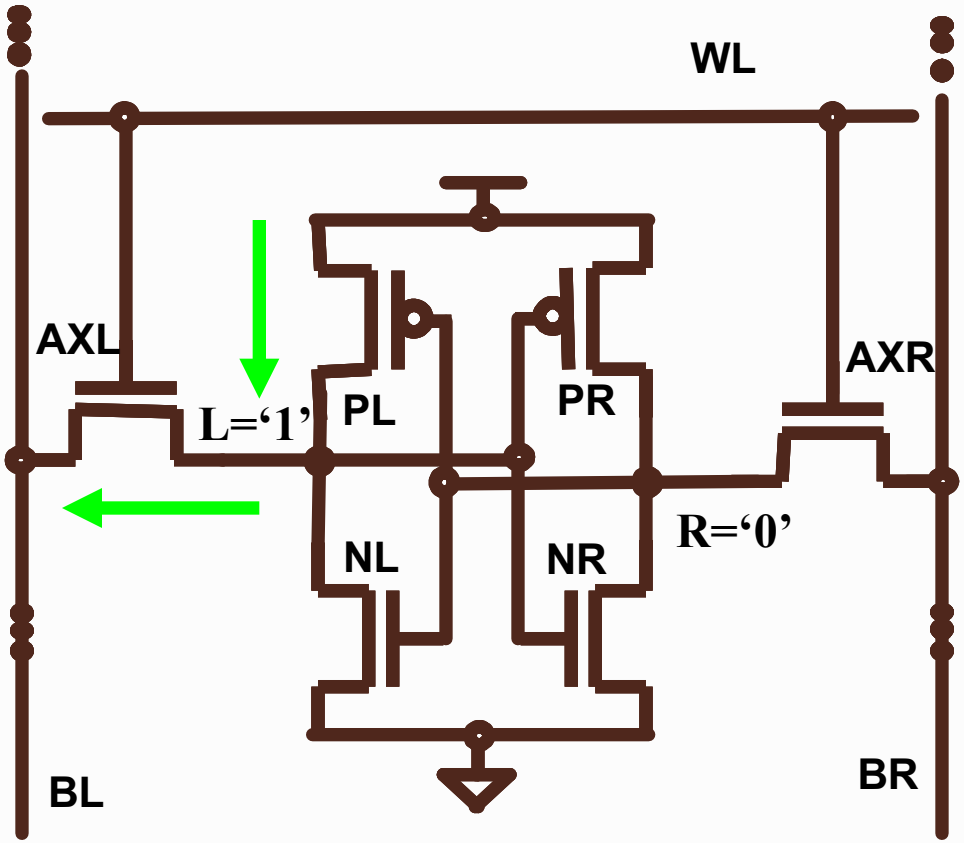
Hold vs Read Margin

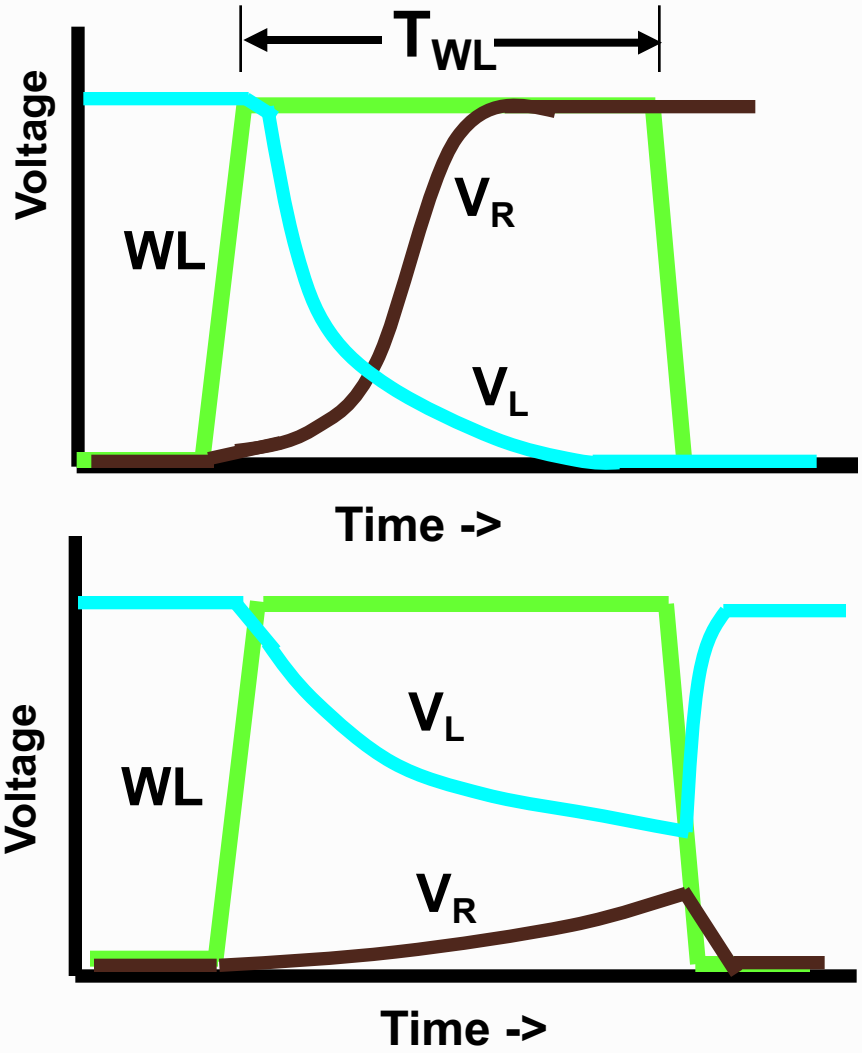


Write Operation



Write Failure

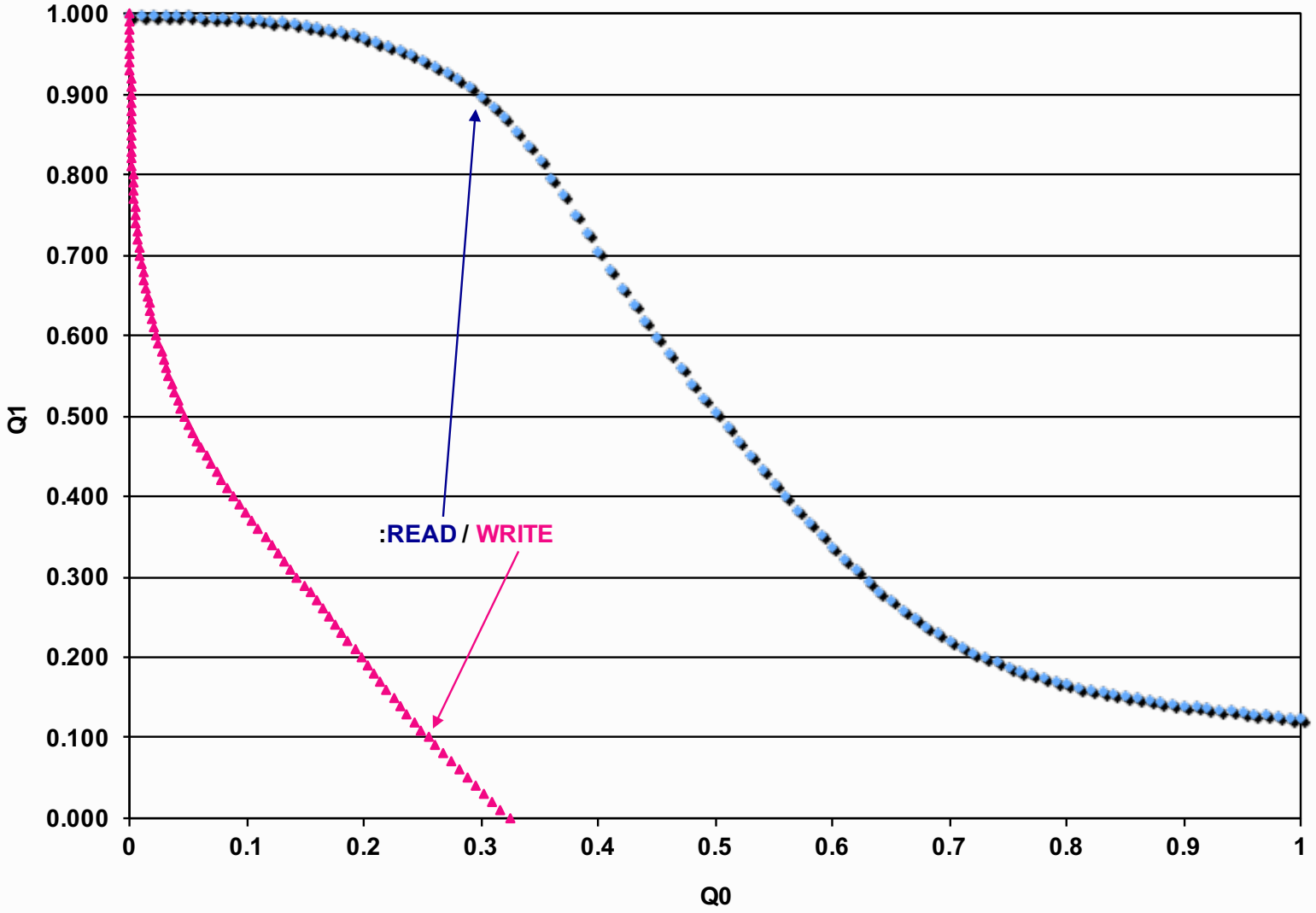




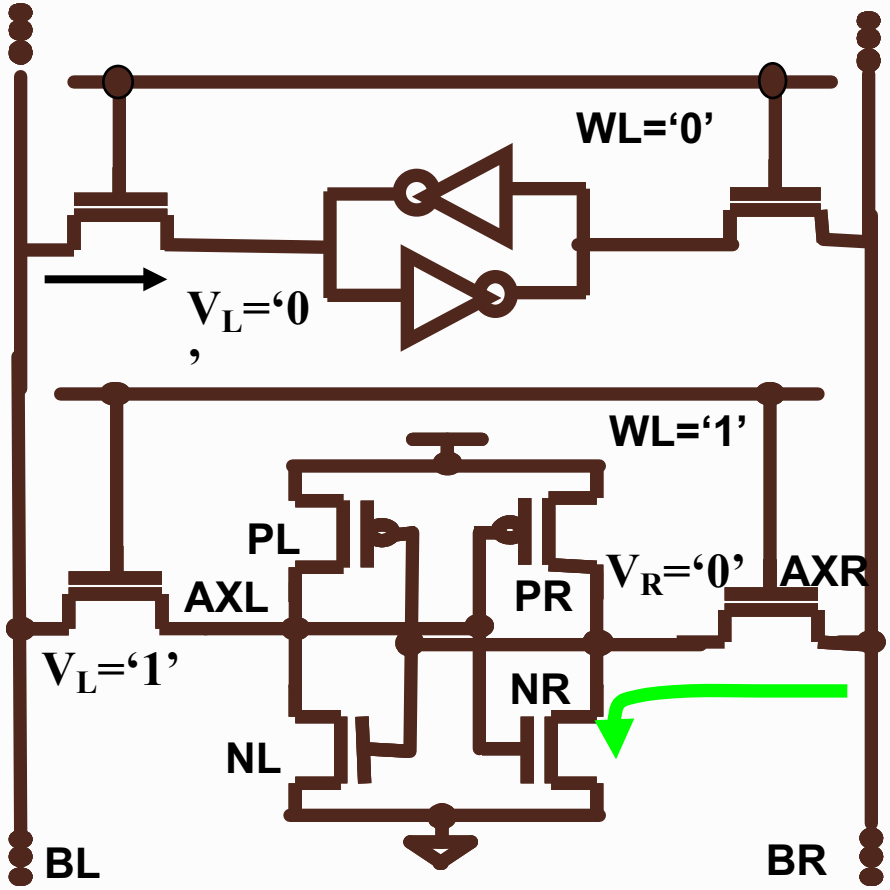
$$P_{WF} = P\left(T_{WRITE} > T_{WL}\right)$$

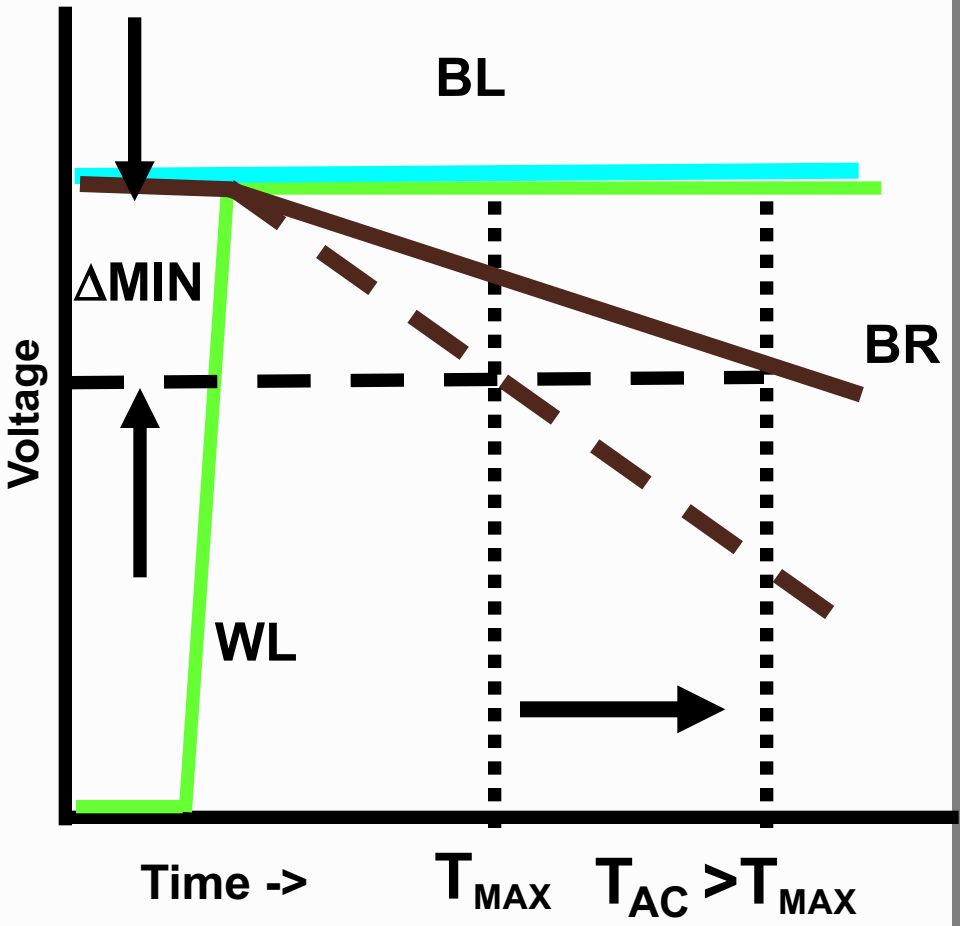
Write Failure => Unsuccessful write to the cell.

Write Margin



Access Failure





$$P_{AF} = P(T_{ACCESS} > T_{MAX})$$

Access Failure : Time required to produce a pre-specified bit-differential is higher than a maximum allowed time.

Question 2b

- Write Margin for an SRAM cell

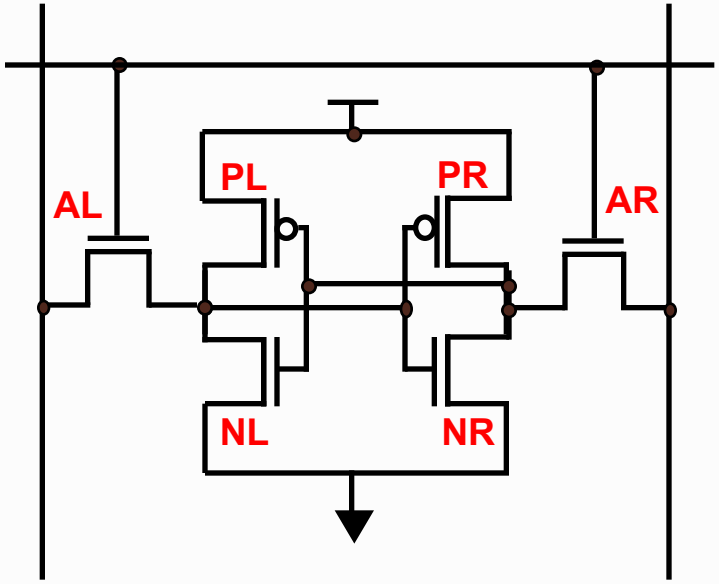
a) Is always greater than read margin

b) Can be improved by making the access transistor bigger

c) Defines the minimum supply voltage required to perform a read operation

d) Depends on the number of cells on the bit line

The Balancing Act



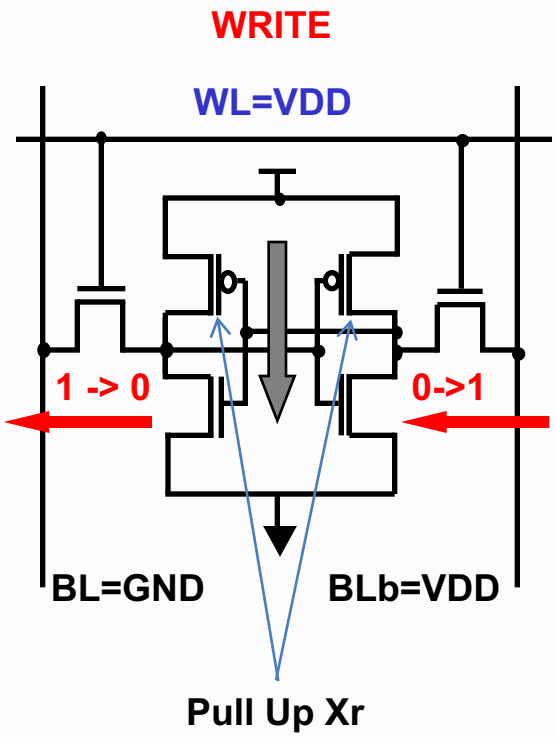
Large N: Better READ performance. If too large, trip voltage of inverter becomes so low that cell becomes unstable.

Large A: Better Performance. If too large, storage node voltage goes high during READ, causing cell flip

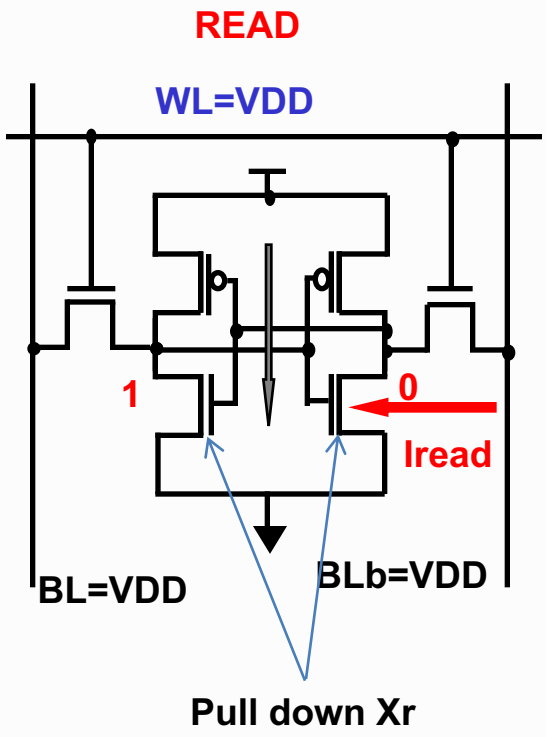
Large P: Increase stability. If too large, hard to WRITE.

Need to balance all : NR:XR:PR ~ 2:1:1

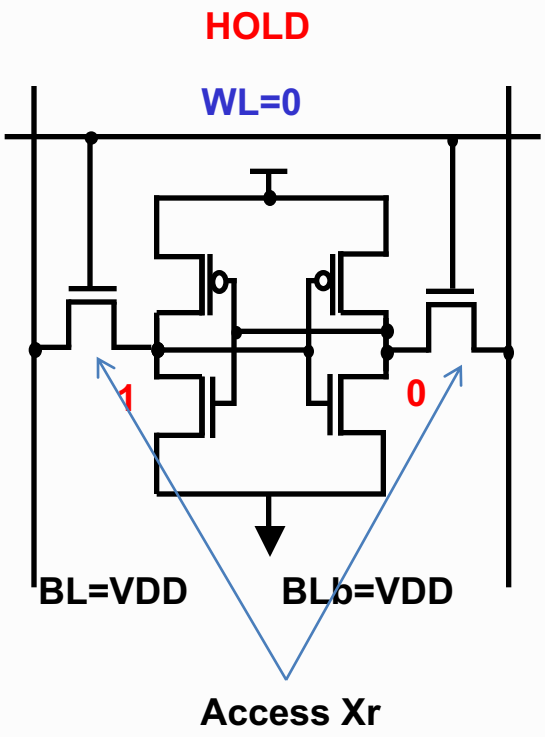
Workhorse 6T-Cell



Access Xr: On
Data driven on bit - lines
Data Flipped by over-coming pull-up / pull – down Xrs



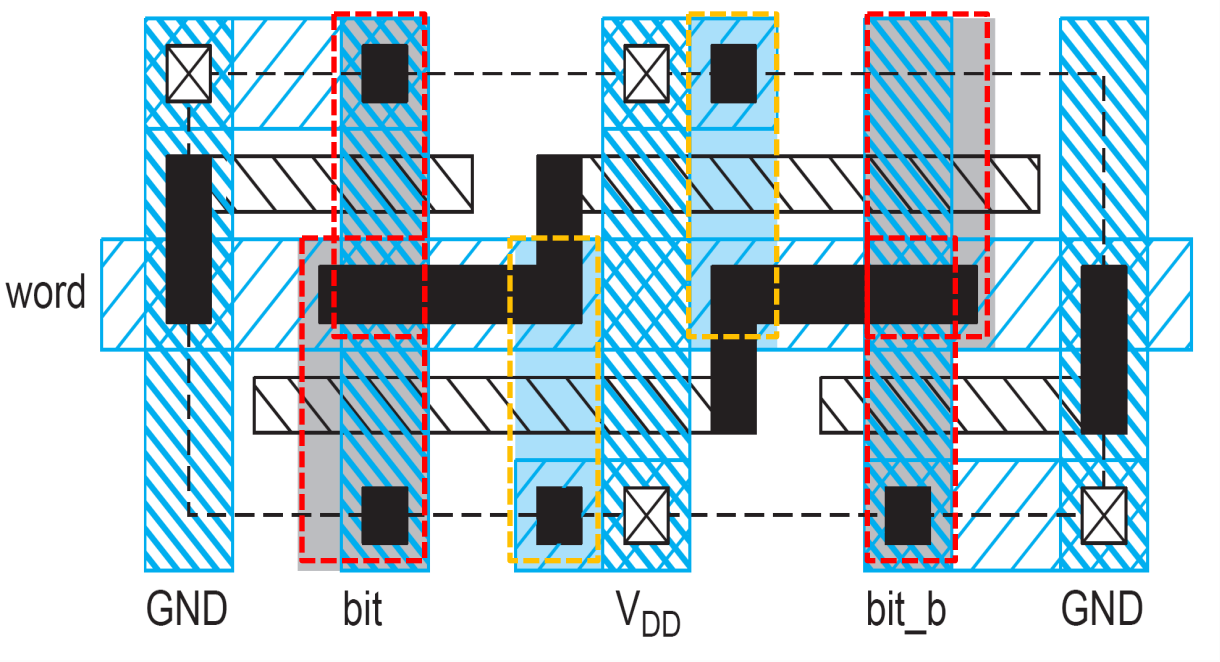
Acess Xr: On
BL, Blb pre-conditioned, and then floated, one line discharges thru the cell (Iread), voltage sensed, Data Retained



Access Xr : Off
Data Retained, due to back-to-back inverters

Thin Cell (Litho-Friendly)

Cell area vital for density



NMOS PMOS

Flip

R0	MY
MX	R180

Flip

Flip

Question 2c

❑Decreasing the size of only one side of NFET transistors will improve the cell

a) Cell density

b) Read margin

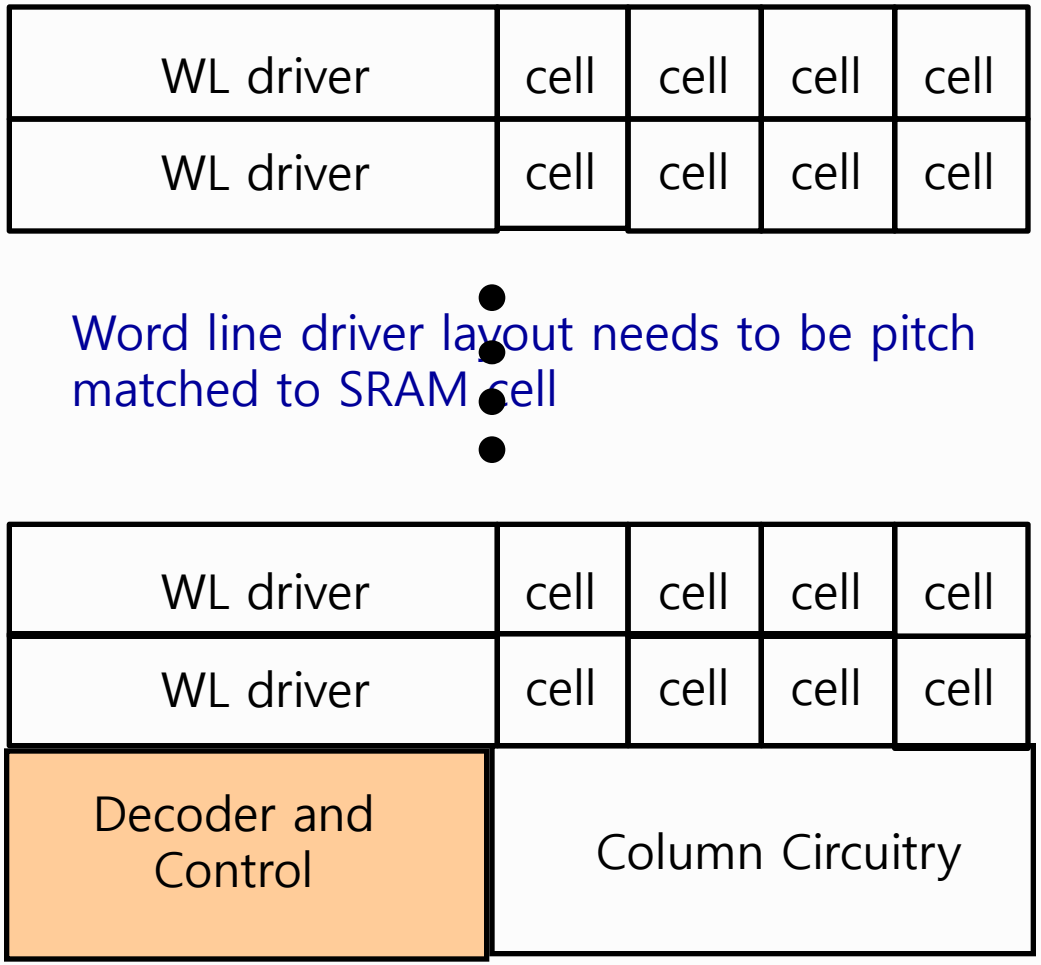
c) Write margin

d) Hold margin

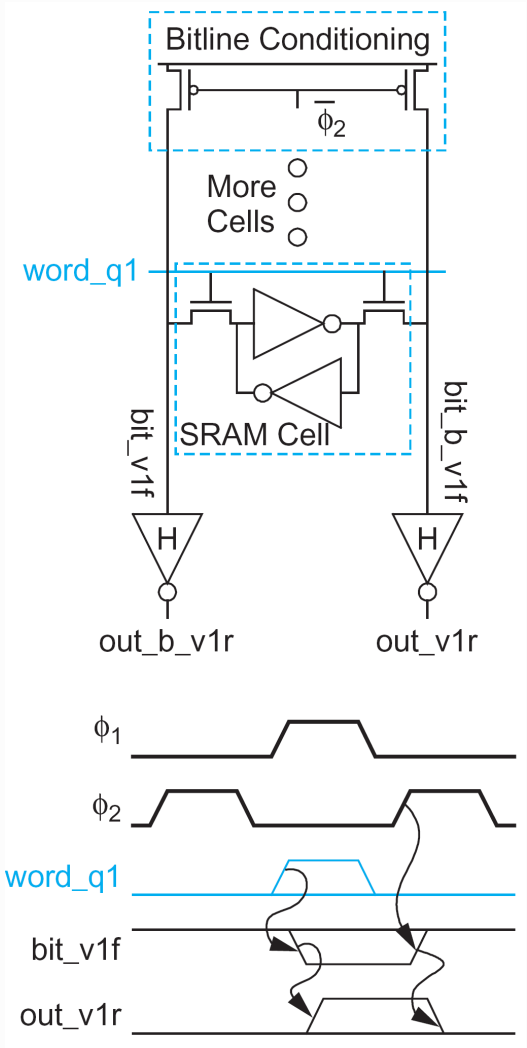
Topics

- Introduction to memory
- SRAM : Basic memory element
- Operations and modes of failure
- Cell optimization
- SRAM peripherals
- Memory architecture and folding

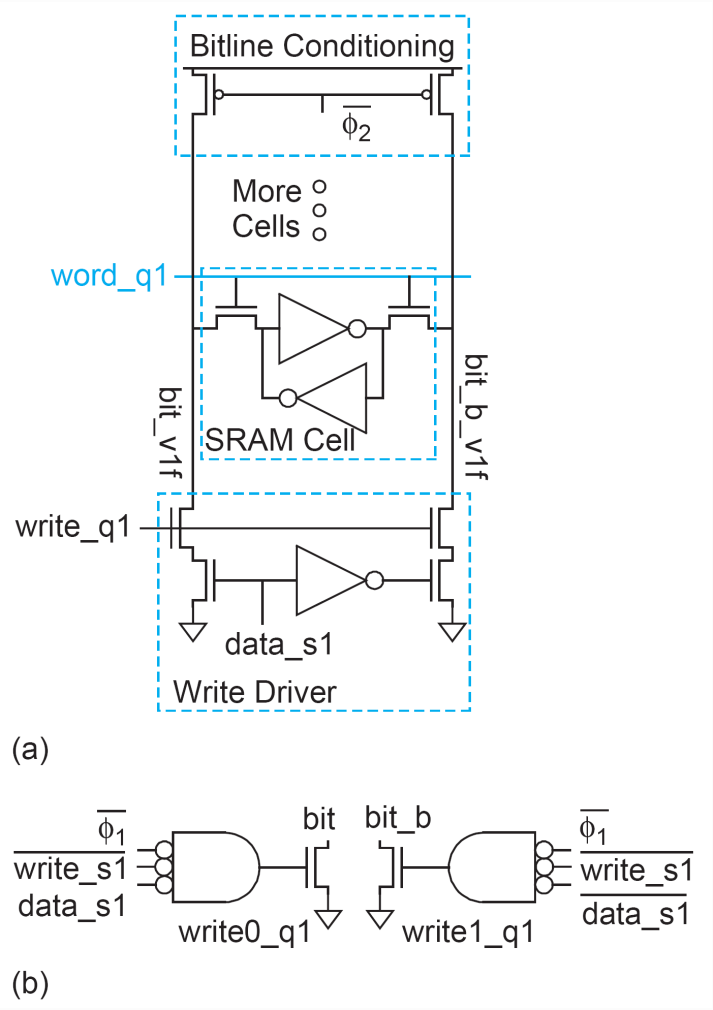
Decoders and Drivers



Column Circuitry for Read and Write

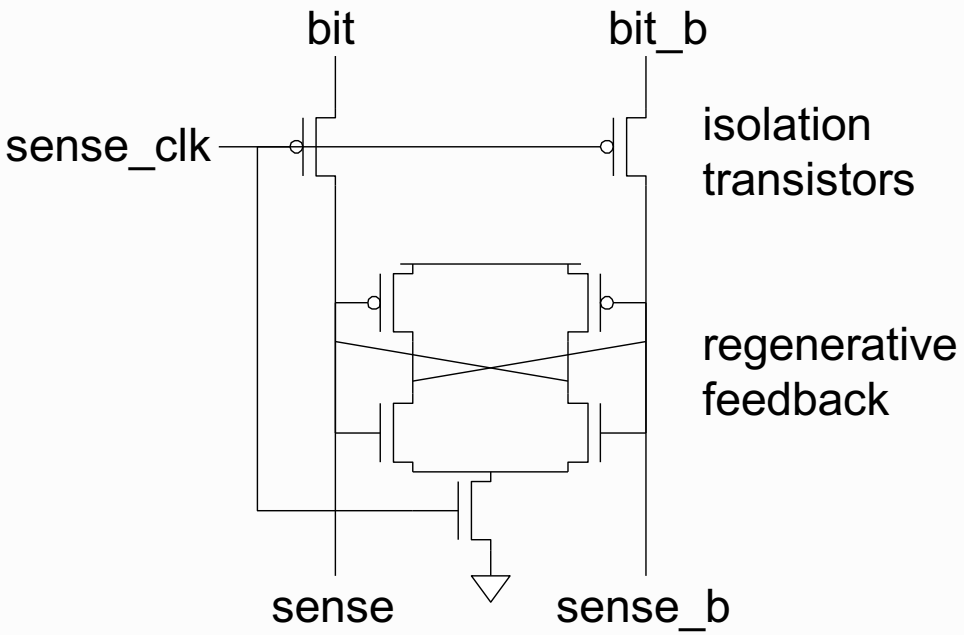


READ



WRITE

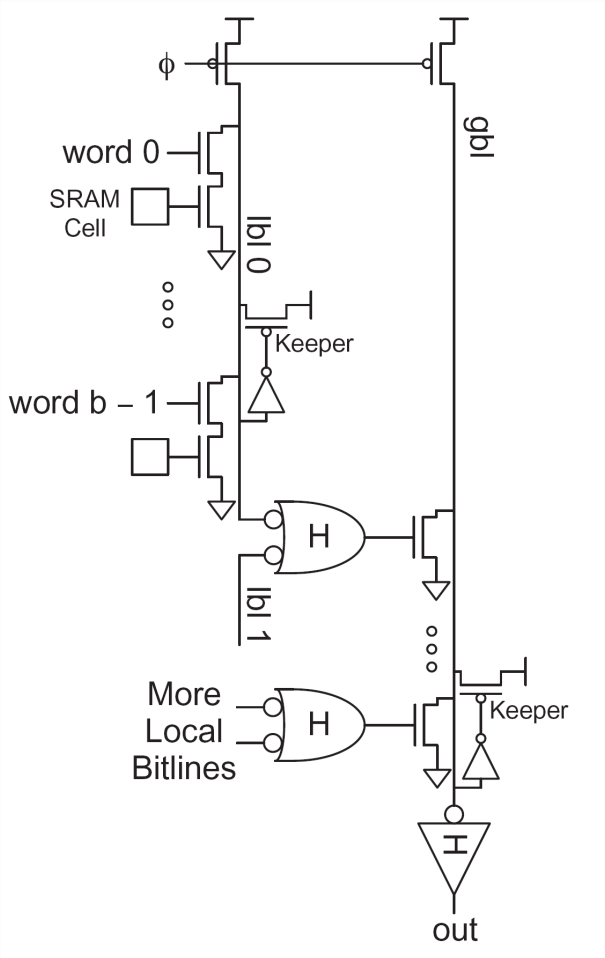
Sense Amplifiers



Sense-amp provide necessary gain (small input → large output) for read
If sense_clk arrives too early ➔ False read may happen due to too small difference
If sense_clk arrives too late ➔ Too slow

Isolation transistors: Disconnects sense amp to cutoff large bit line capacitance once sensing starts

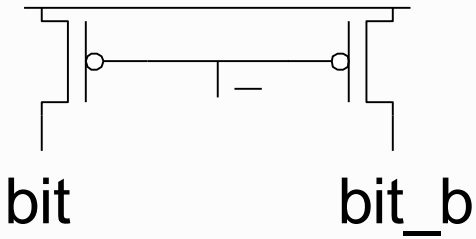
Hierarchical Bit-lines



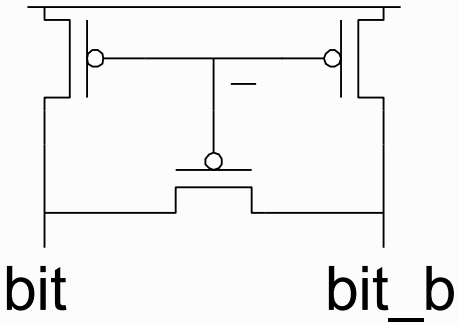
Hierarchical Bit-lines

Pre-Conditioning

Precharge



Equalizer



Pre-Conditioning

Question 3

- Equalizer is required

a) Both the bitlines are pre-charged at the same time

b) To ensure that the bit-lines are conditioned suitably for write operation

c) To minimize offset between the bitlines prior to read operation

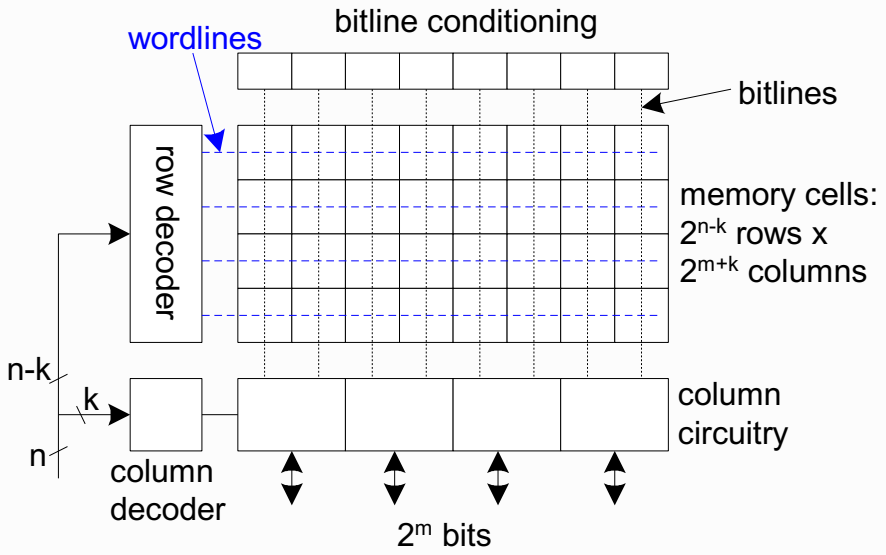
d) To improve the hold and read margin of the memory cells

Topics

- Introduction to memory
- SRAM : Basic memory element
- Operations and modes of failure
- Cell optimization
- SRAM peripherals
- Memory architecture and folding

Memory Architecture

- 2^n *words* of 2^m *bits* each, If n >> m, fold by 2^k into fewer *rows* of more *columns*



- Good regularity – easy to design
- Utilization = Cell Area / (Cell + Periphery Area)

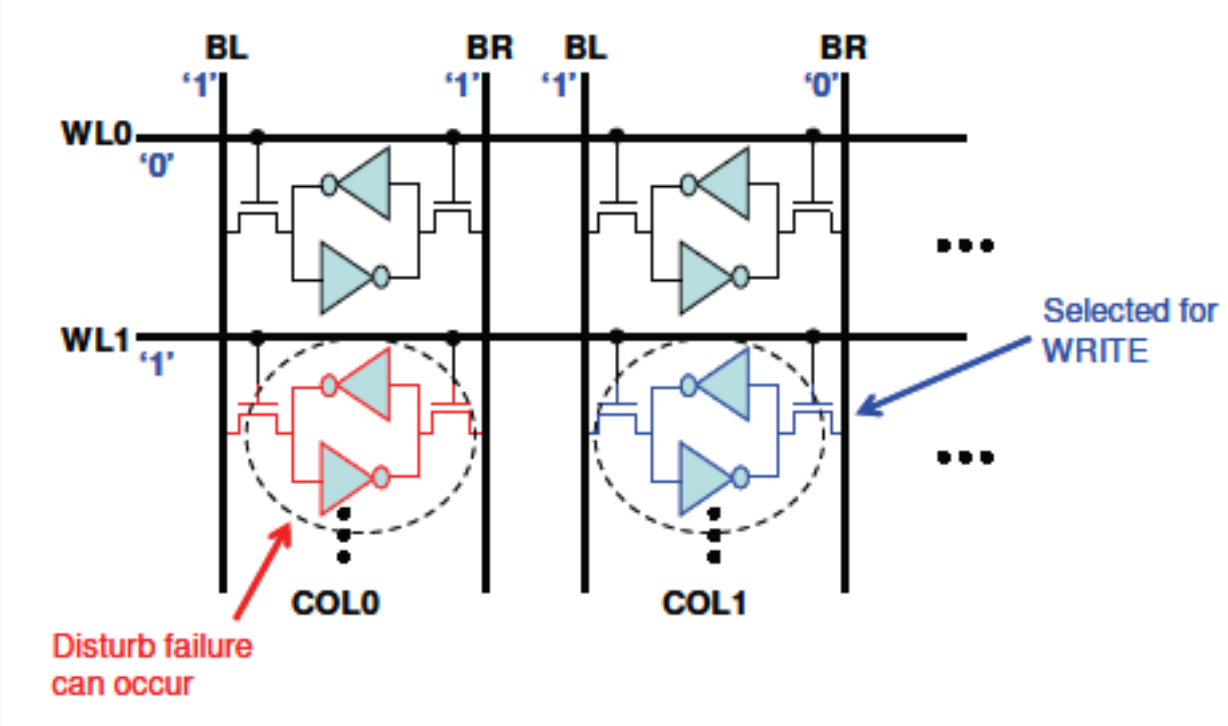
Why memory is folded?

- To improve aspect ratio by column multiplexing
  - E.g.: 2Kword x 16 can be arranged as 256 rows and 128 columns
  - 8:1 MUX are used to read 16 desired columns out of 128
- To improve soft-error immunity
  - bits of a word are not placed next to each other
    - Single-bit soft-error can be corrected by ECC

Column Select and Half-Select Issue

Prevents multiple-bit soft error

Better aspect ratio



Class Exercise

- Build a schematic of a 6T – SRAM cell with minimum sized PFETs, Pull down = 3*PFET size, and Access transistor = 2* PFET size. Simulate it and plot butterfly curve for margins
- Change Pull down size to 4*PFET size and re-simulate
- Change Access transistor size to 3*PFET size and re-simulate
- Change pull up device size to 2 original size and re-simulate

Next Class

- Alternative Cell Types (6 to 10T), Asymmetric Cells, Sub-threshold Cells, Low – leakage cells