

ECE 225

High-Speed Digital IC Design

Lecture 17

Semiconductor Memories, SRAM Design

Prof. Kaustav Banerjee
Electrical and Computer Engineering
E-mail: kaustav@ece.ucsb.edu

Semiconductor Memories

Memory Design...

- ❑ Increasing number of transistors in processors are devoted to cache memories....more than 60%, see ITRS for more details.....
- ❑ At the system level: high-performance workstations and desktops have several Gbytes of memory
- ❑ Audio (MP3), Video players (MPEG4) and GPUs require large amount of memory
- ❑ Can we store Memory using registers?yes but the area required will be excessive (need > 10 transistors/bit)
- ❑ Memory cells are therefore combined into large arrays, which minimizes the overhead caused by the peripheral circuits and increases storage density
- ❑ Memory design can be classified as high-performance, high density, low-power circuit design

Memory Classification

- ❑ Size
- ❑ Timing Parameters
- ❑ Function
- ❑ Access Pattern

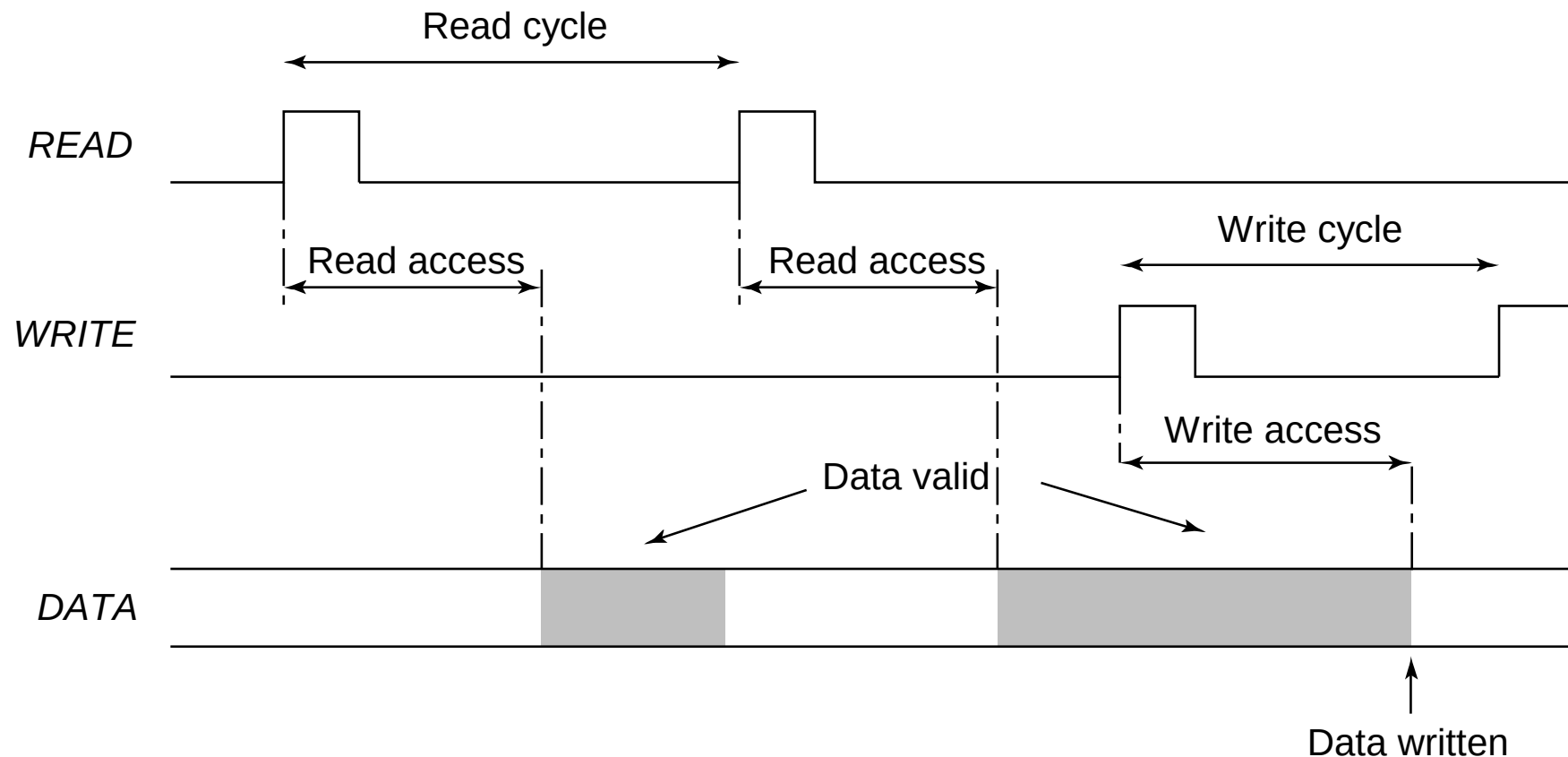
Memory Size

- ❑ Depends on the level of abstraction
- ❑ **Bits:** (used by circuit designers) are equivalent to the number of individual cells (FFs or Registers) to store data
- ❑ **Bytes:** (used by chip designers) are groups of 8 or 9 bits or their multiples: Kbyte, Mbyte, Gbyte, Tbyte
- ❑ **Words:** (used by system designers) represent a basic computational entity. For example, a group of 32 bits represent a word in a computer that operates on 32 bit data

Timing Parameters

- ❑ **READ-Access Time:** time it takes to retrieve (read) from the memory. This is equal to the delay between the read request and the moment the data becomes available at the O/P.
- ❑ **WRITE-Access Time:** time elapsed between a write request and the final writing of the input data into the memory
- ❑ **CYCLE Time:** minimum time required between successive reads or writes

Memory Timing: Definitions



Note: Read and Write cycles do not necessarily have the same length but are considered to be equal for simplicity of system design.

Function

❑ Read-Only Memory (ROM):

- encode the information into the circuit topology-by removing or adding transistors. The topology is hard wired and the data cannot be modified....can only be read.
- They belong to the class of **Non-volatile** memories. Disconnection of the supply voltage does not result in a loss of the stored data.

❑ Read-Write Memories (RWM): called as **RAM** (Random-Access Memories).

- **Static** (retains data if Vdd is retained): example SRAM
- **Dynamic** (needs periodic refreshing): example DRAM
- They use active circuitry to store information and belong to the class of **Volatile** memories.

Function....cont'd

❑ **Non-Volatile Read-Write (NVRWM):**

- Recent Non-Volatile Memories can read and write---although write function is substantially slower
- Novel, cheap and dense: Fastest growing among semiconductor memories

❑ **Examples:**

- EPROM: Electrically Programmable ROM
- E²PROM: Electrically Erasable and Programmable ROM
- Flash memory

Access Pattern

❑ Random-Access (RAM):

- memory locations can be read or written in a random manner
- Most ROMs and NVRWMs allow random access....but “RAM” is used for the RWMs only

❑ Serial Access:

- Restricts the order of access. Results in faster access times, smaller area, or allows special functionality
- Examples: (Video Memories)
 - FIFO (first-in first-out)
 - LIFO (last-in first-out)
 - Shift Register

❑ Content-Addressable Memory (CAM): (non-random access)

- Also known as associative memory
- Doesn't use an address to locate the data.....rather **uses a word of data itself as input**... when input data matches a data word stored in memory array, a MATCH flag is raised
- Important component of the cache architecture of most microprocessors

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		

Where does your brain's memory fit into these classification schemes?

More Classification

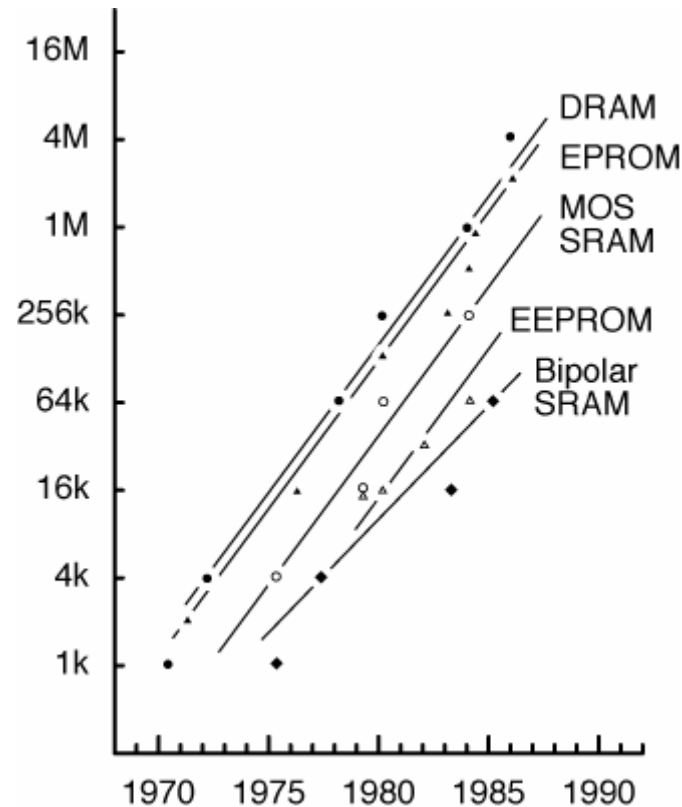
❑ I/O Architecture:

- Based on the number of data input and output ports
- Most memories use a single I/O port
- **Multiport memories** offer higher bandwidth
 - Example: register files used in RISC processors
 - Adds more complexity to the design

❑ Application:

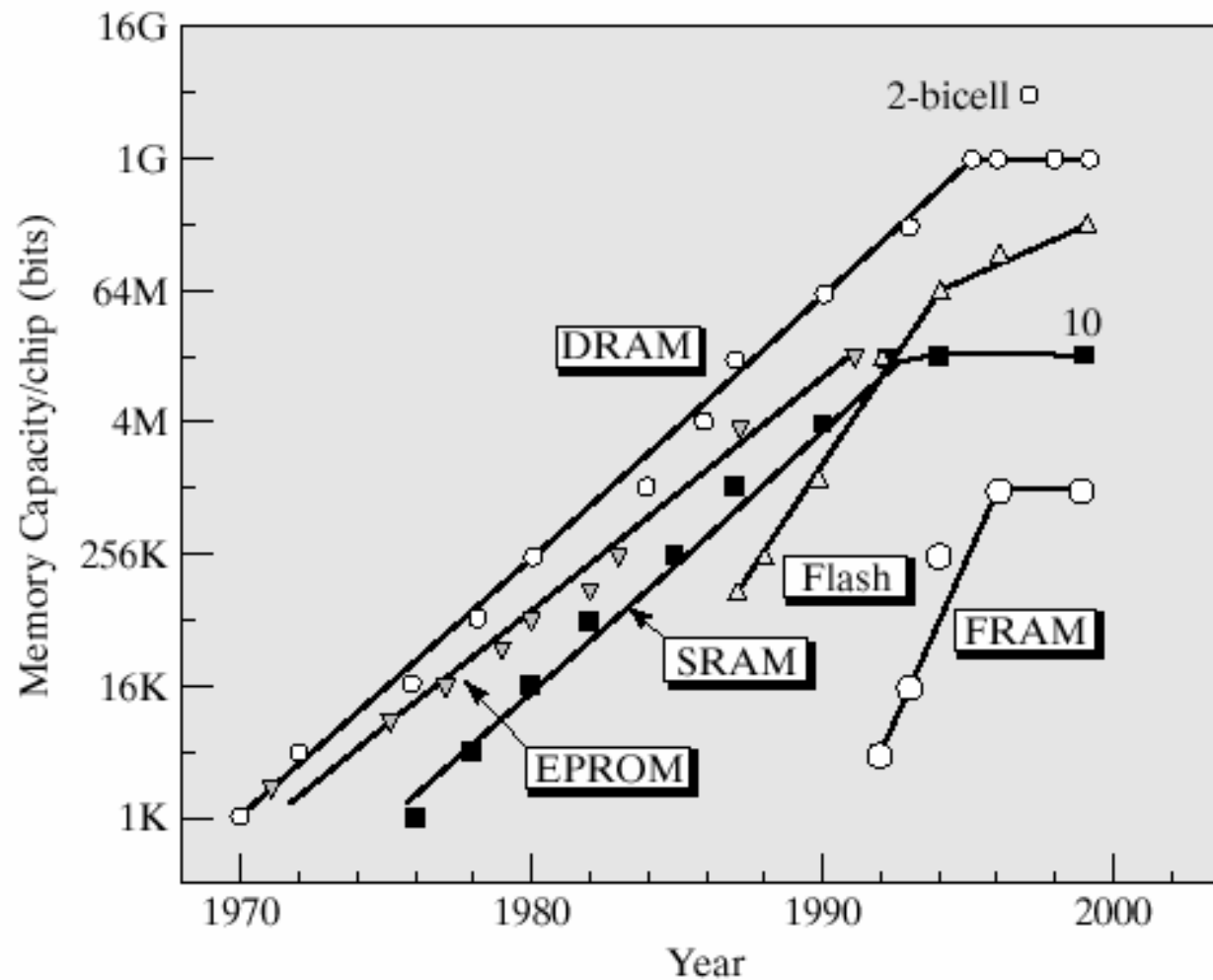
- Embedded Memories in SoCs
- For massive storage (multiples of Gbytes and beyond), more cost effective solutions are to use **magnetic tapes** and **optical disks**---they however, tend to be slower and provide limited access pattern

Semiconductor Memory Trends (up to the 90's)



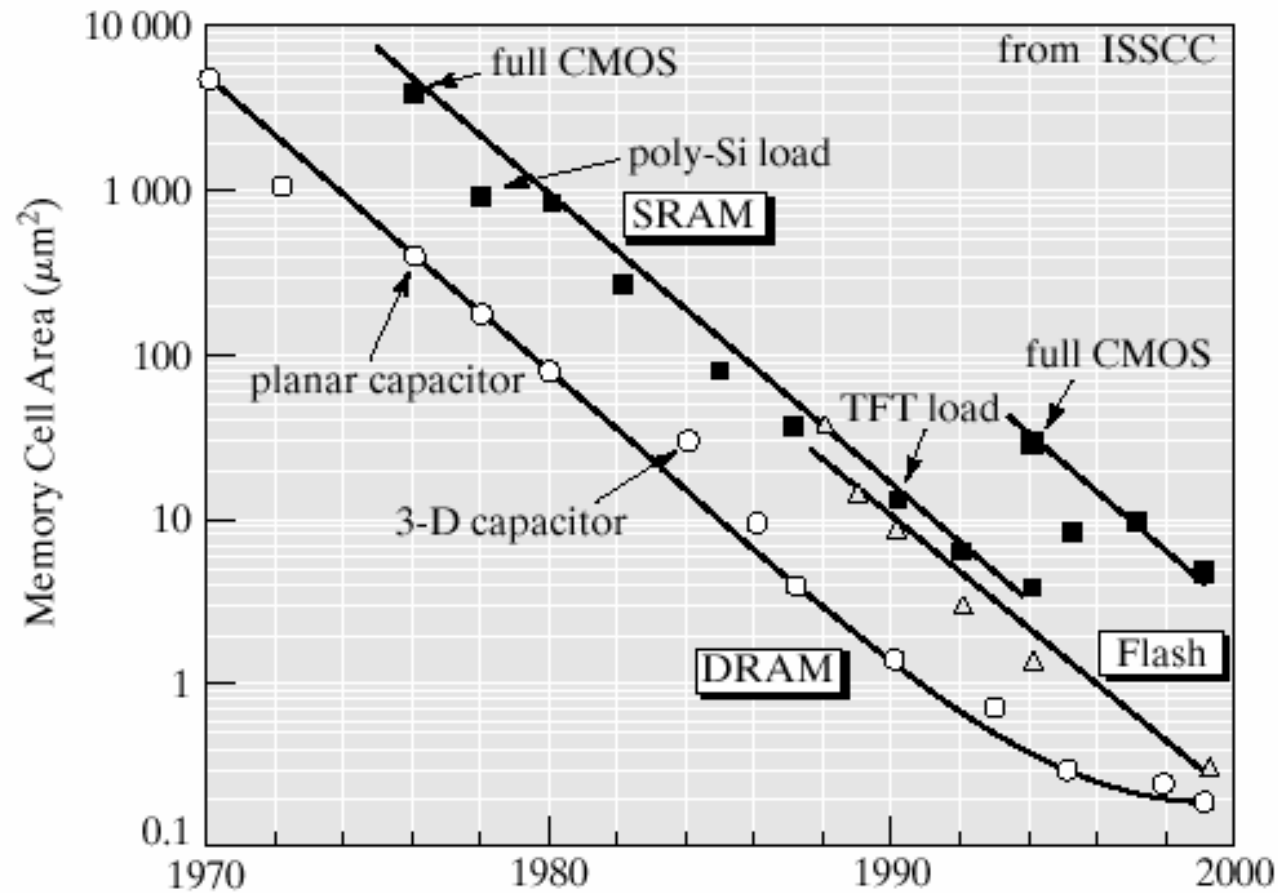
Memory Size as a function of time: x 4 every three years

Semiconductor Memory Trends (recent)



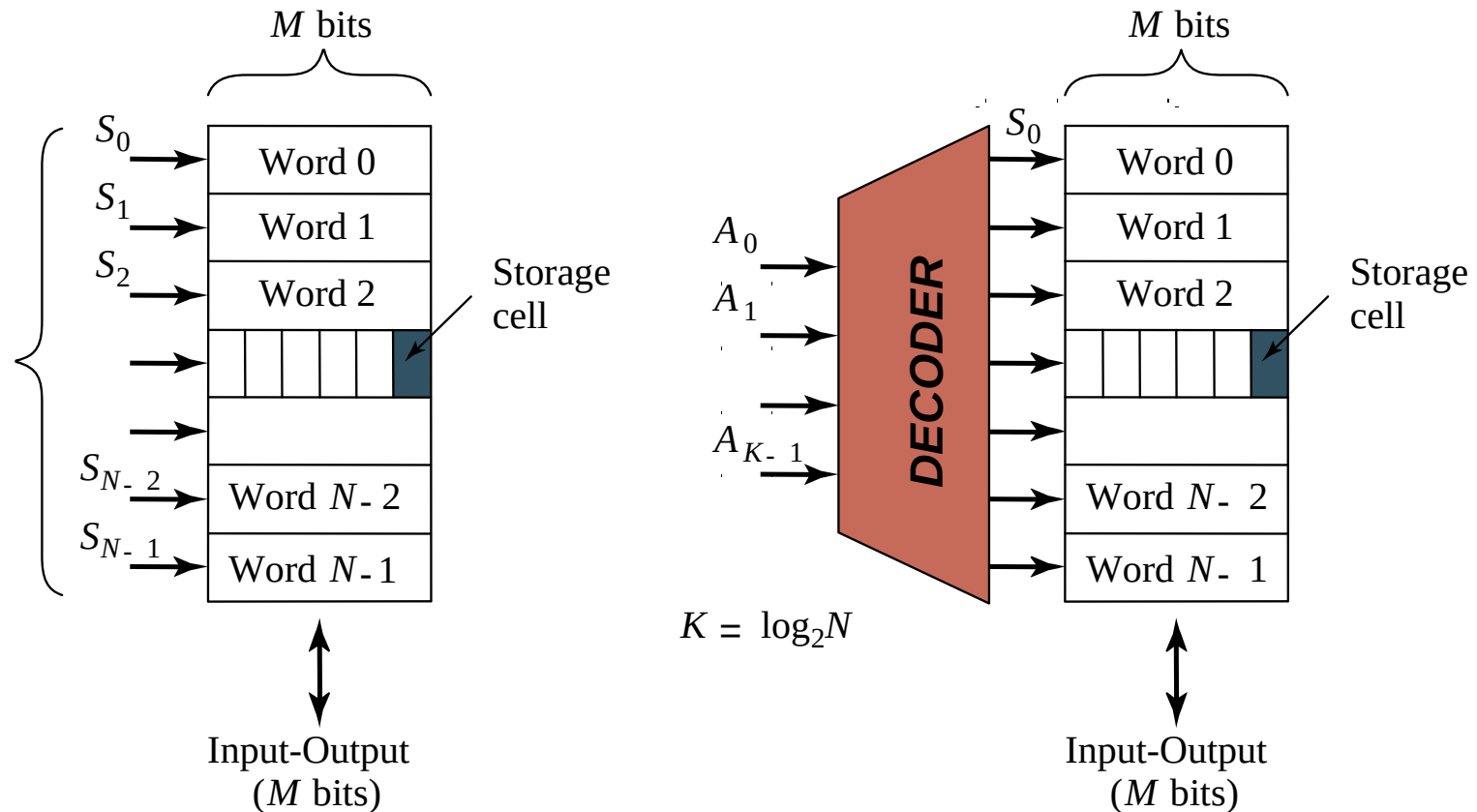
From [Itoh01]

Trends in Memory Cell Area



From [Itoh01]

Memory Architecture: Decoders

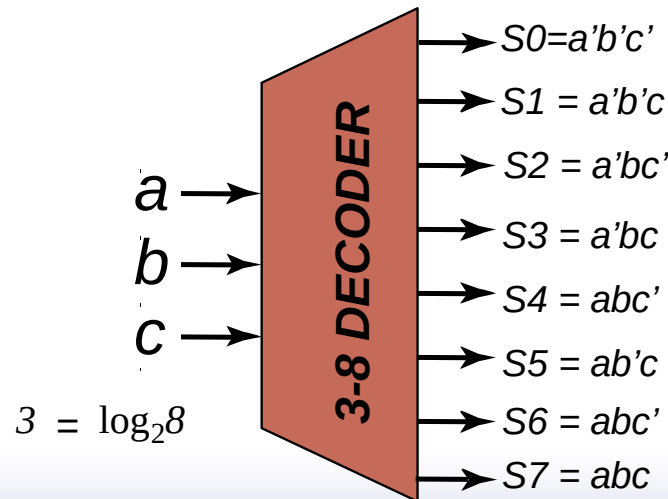


Intuitive architecture for $N \times M$ memory
Too many select signals:
 N words == N select signals

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

Decoder Basic

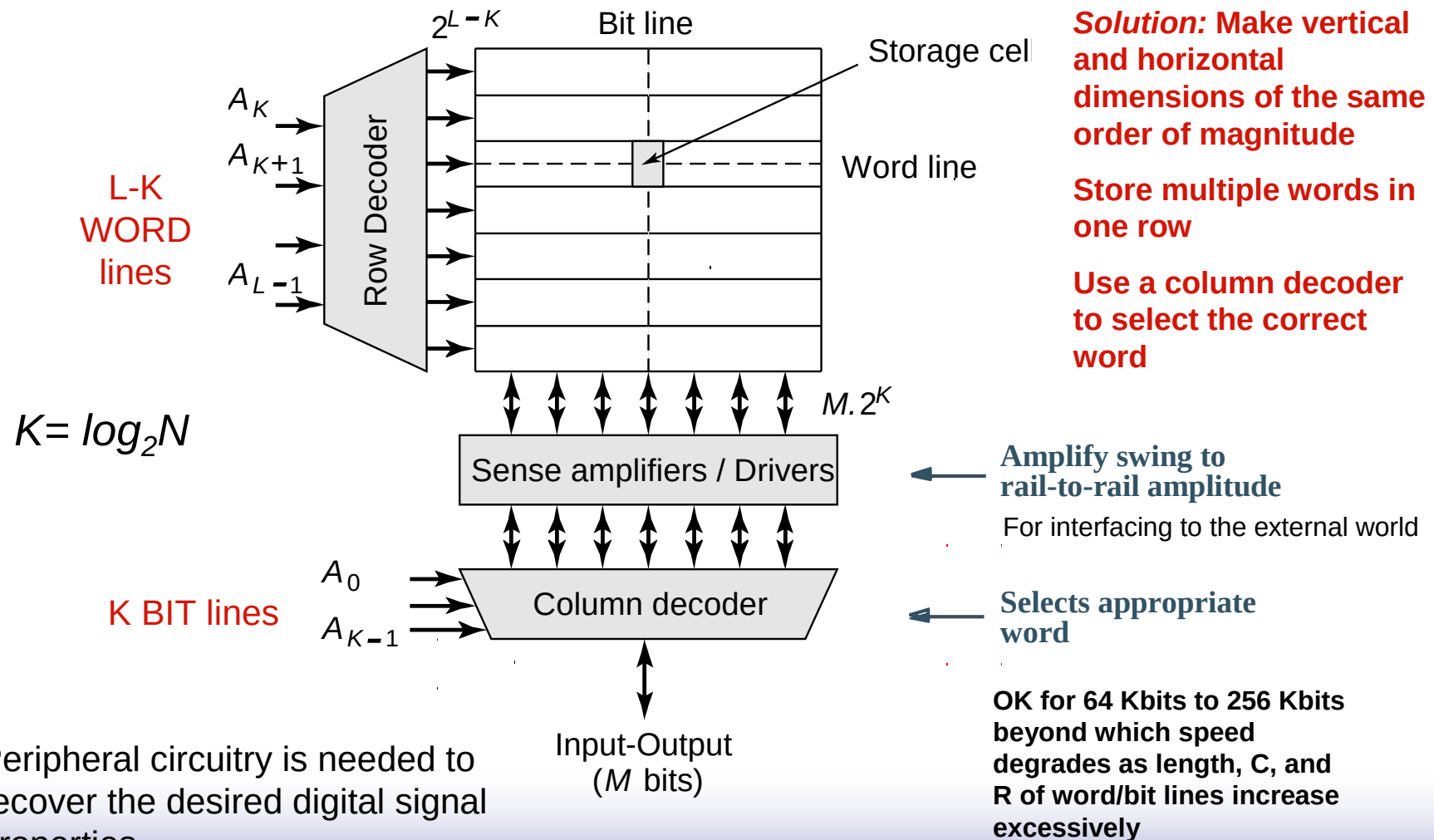
- ❑ Recall that a decoder is a combinational circuit with **k inputs** and at most **2^k outputs**.
- ❑ Its characteristics property is that for every combination of input values only **ONE** output =1 at the same time.
- ❑ Used to route input data to specific output line.



For example: for $a=b=c=0$, only $S0 = 1$

Array-Structured Memory Architecture

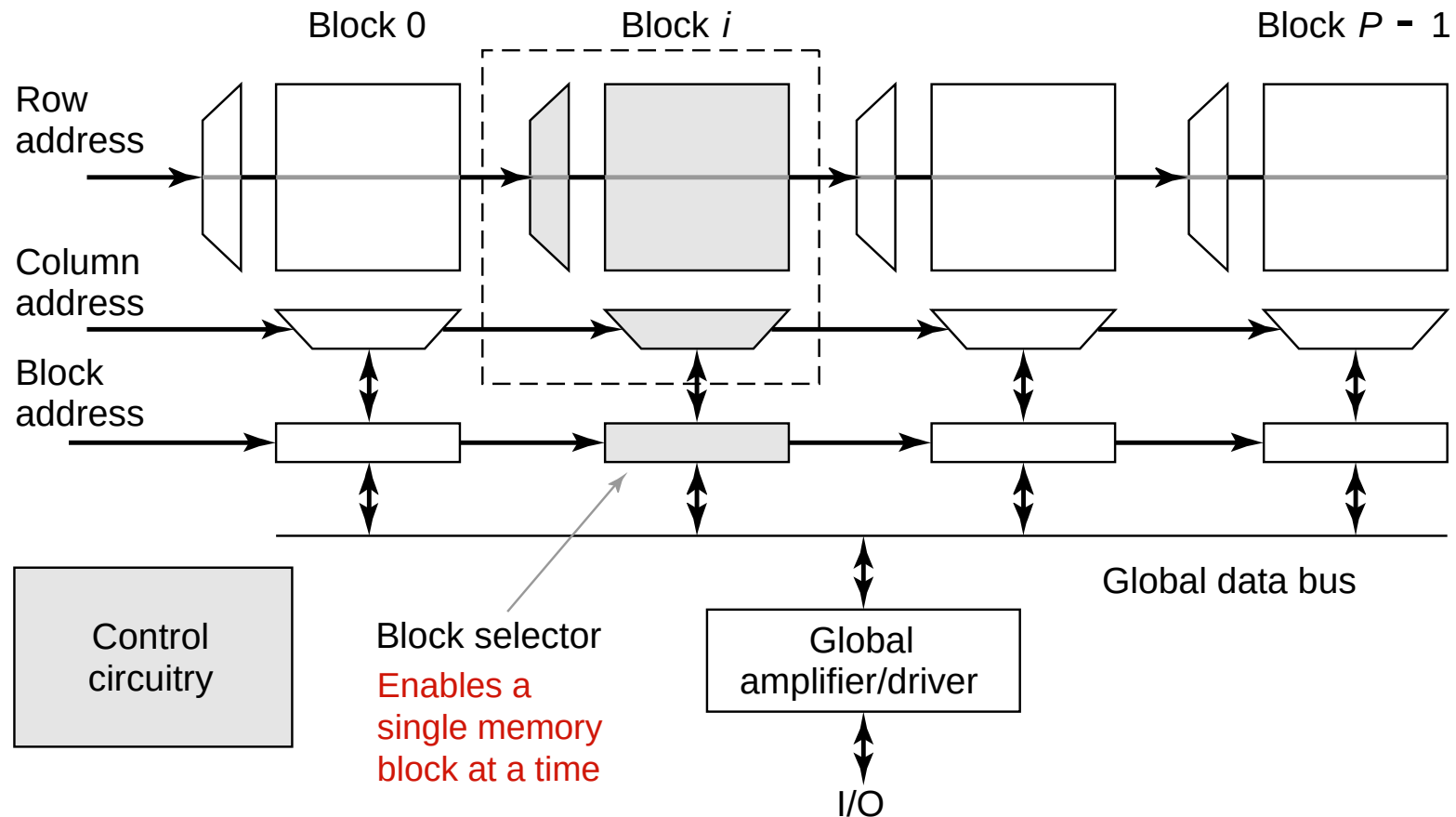
Problem: consider ~1 million ($N=2^{20}$) 8-bit ($M=2^3$) words, ASPECT RATIO is very large!!! or HEIGHT $\gg$ WIDTH, cannot be implemented and will result in very slow design.....



Peripheral circuitry is needed to recover the desired digital signal properties

Hierarchical Memory Architecture

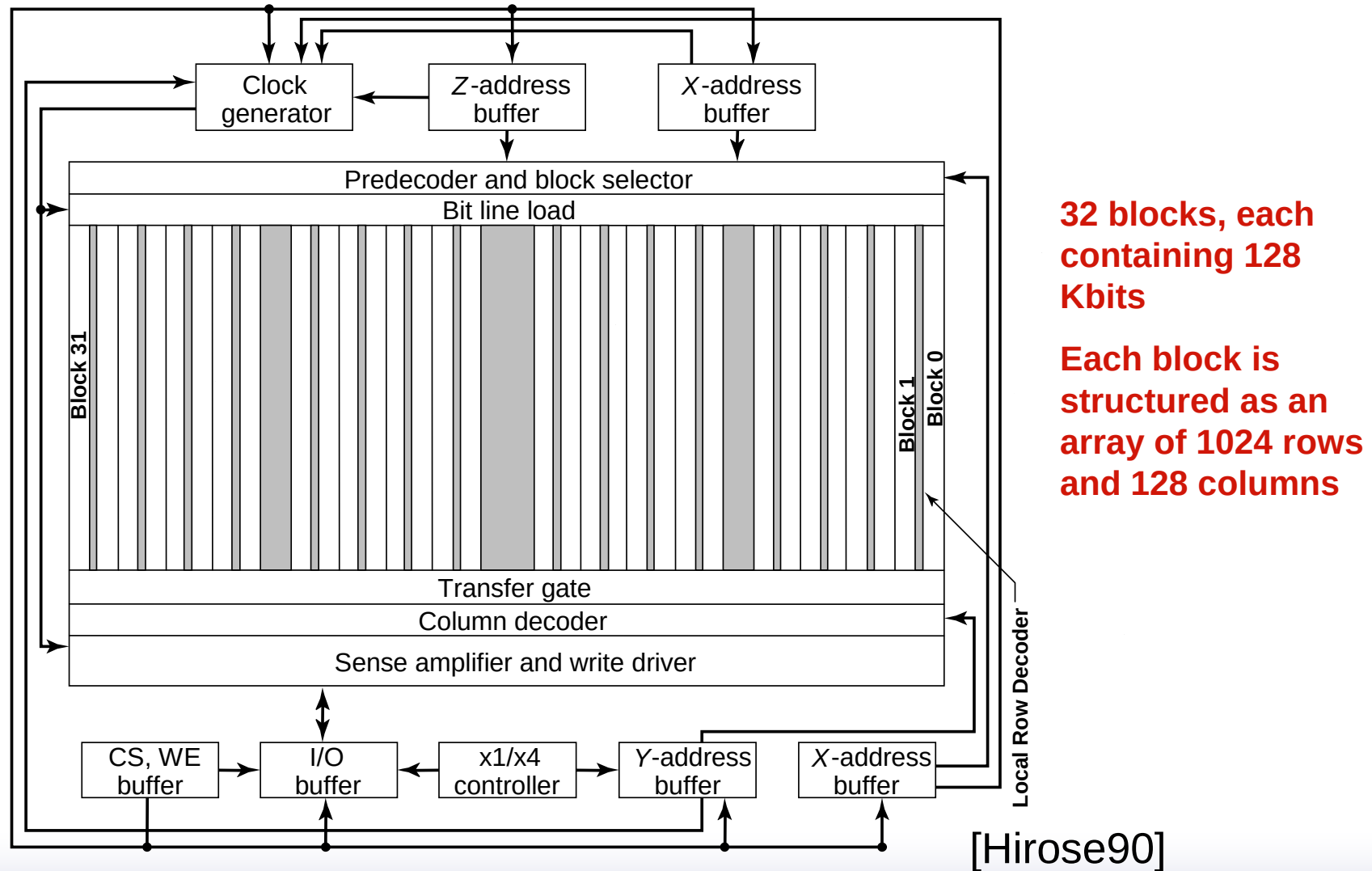
For Larger Memories....



Advantages:

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

Block Diagram of 4 Mbit SRAM



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

Should be minimum sized to achieve high memory density.....

READ Operation:

Assume 1 is stored at Q

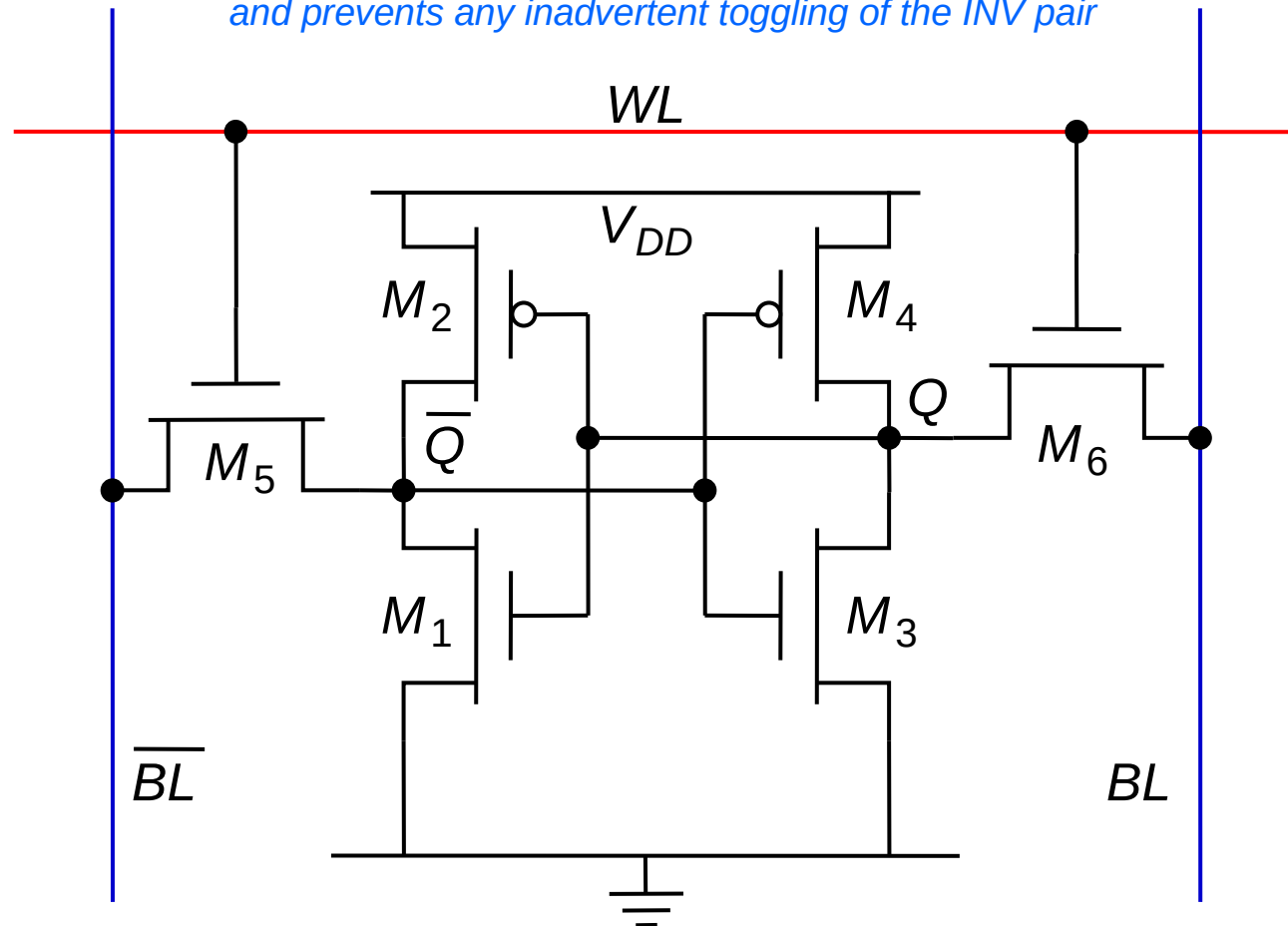
Assume both BLs are held high before the read.

Read cycle started by asserting the WL, enables PTs M5 and M6

During a correct read operation values stored in Q and $\overline{Q}$ are transferred to the bit lines leaving BL at its precharge value and by discharging $\overline{BL}$ through M1-M5

A "0" can be read in a similar manner (now BL gets discharged through M6 and M3)

Major advantage of dual BL: Q is clamped to V_{DD} by BL and prevents any inadvertent toggling of the INV pair



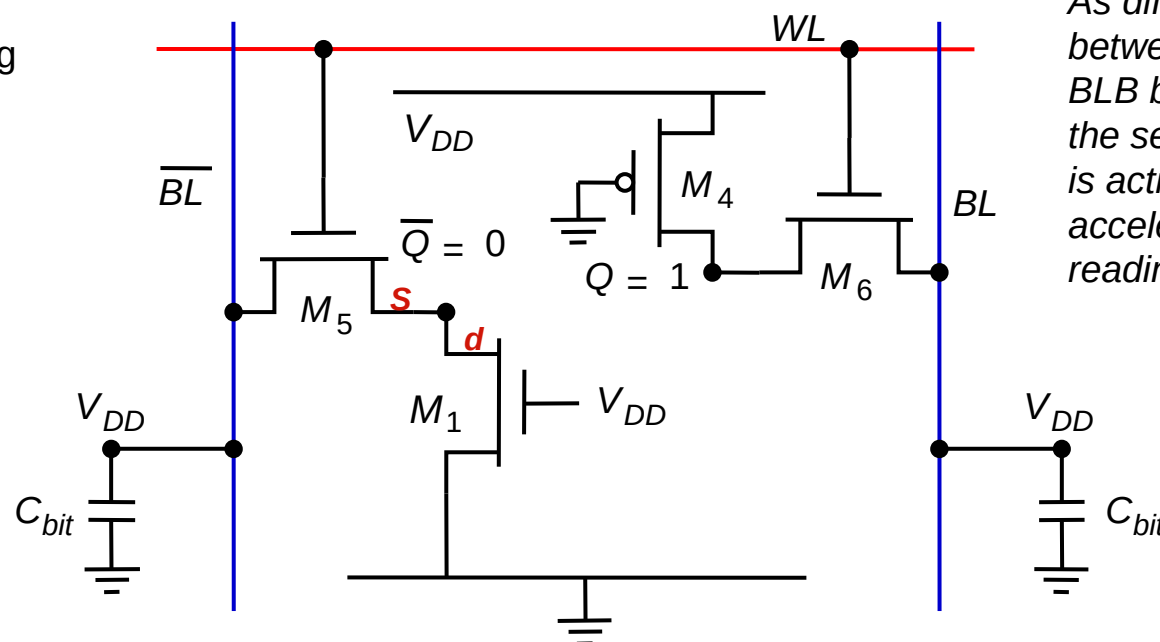
SRAM cell should be as small as possible.....but reliable operation requires careful sizing...

CMOS SRAM Analysis (Read “1” operation)

Transistor sizing is needed to avoid writing 1 accidentally, i.e., voltage at $\bar{Q}$ becomes $> V_M$ of Inv M3-M4

M1 must be stronger than M5

$\overline{Q}$ must stay low enough so that there is no substantial current through M3-M4 INV



As difference between BL and BLB builds up, the sense amp. is activated to accelerate the reading process

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

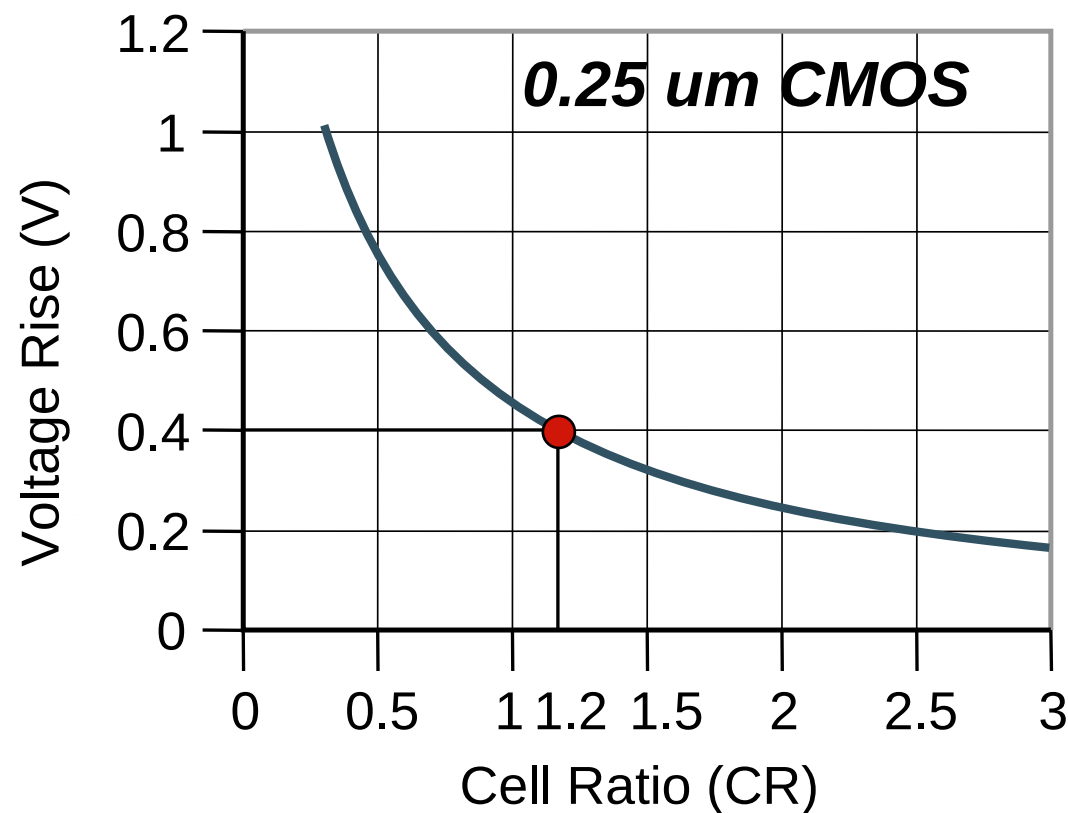
(M5 in saturation) (M1 in linear)

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

Value of the ripple voltage

CR = cell ratio = M1/M5

CMOS SRAM Analysis (Read)



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

**Choose
M5 to be
minimum
size and
M1 > M5**

Node voltage must stay below the V_{th} of M3: CR must be >1.2

CMOS SRAM Analysis (Write)

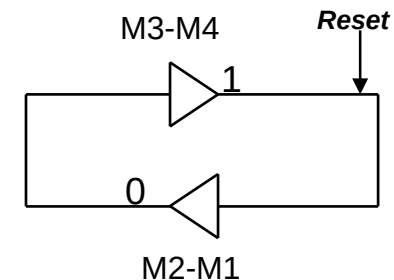
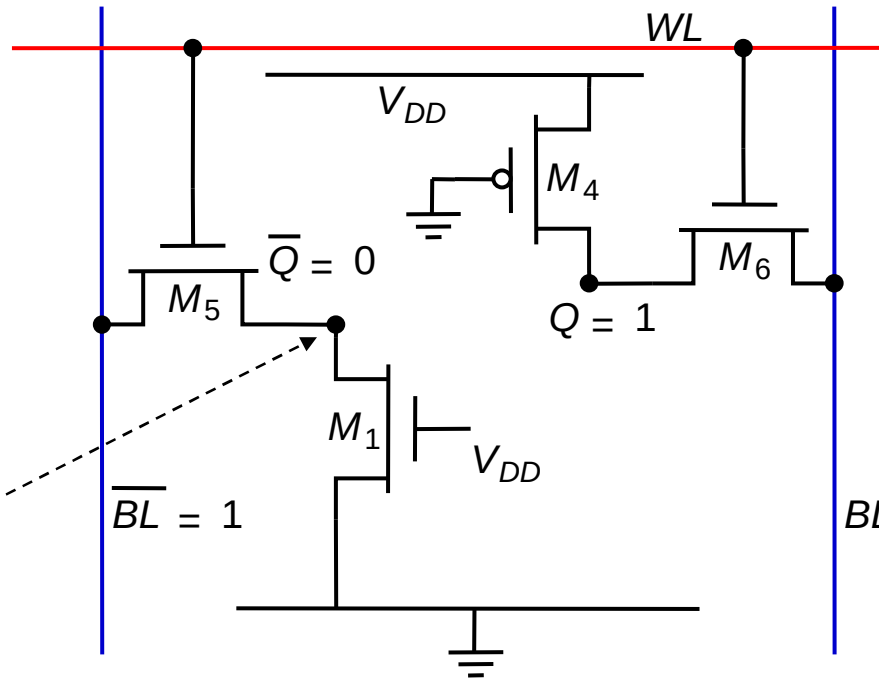
Assume that $Q=1$

To write a 0 in the cell: set $\overline{BL}=1$ and $BL=0$

Similar to applying a reset pulse to an SR latch. FF will change state if sized properly

$\overline{Q}$ cannot be pulled high due to the sizing of M5 and M1 already done for reading

New value must be written through M6



Reliable writing of the cell is ensured if we can pull node Q low enough—below the V_{th} of M1

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

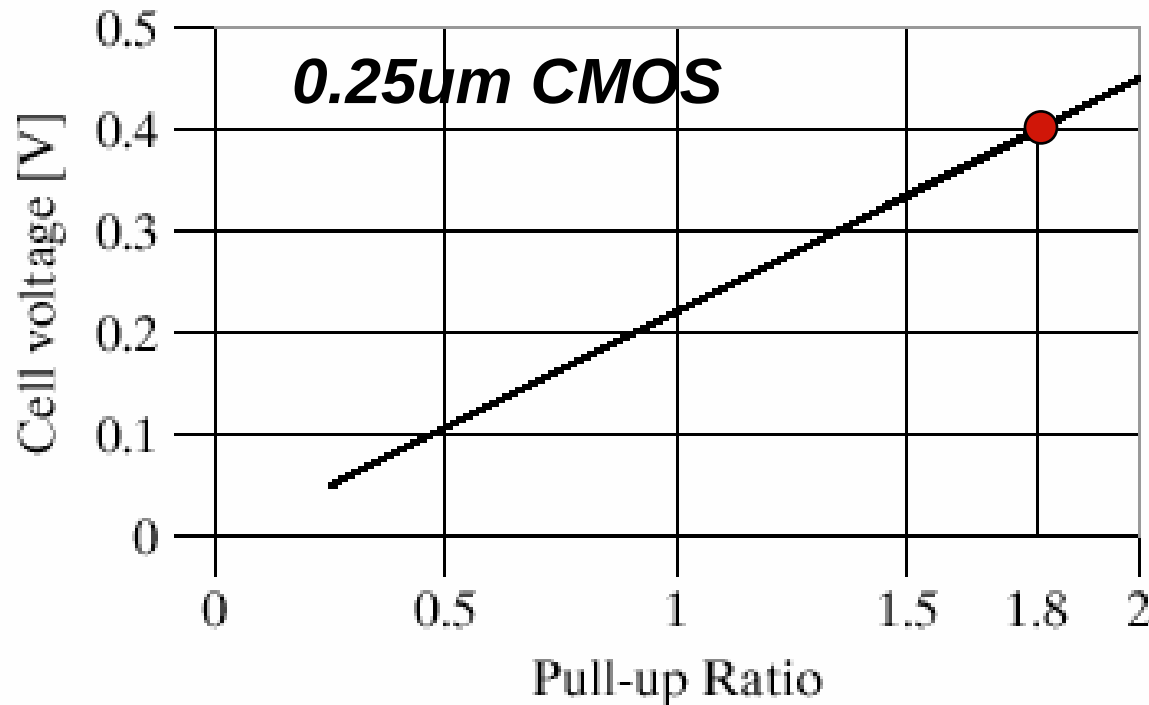
(M6 in linear) (M4 in saturation)

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

PR = pull-up ratio of cell = $M4/M6$

CMOS SRAM Analysis (Write)

Dependence of V_Q on Pull-up Ratio.....lower PR gives lower V_Q



$$PR = \frac{W_4/L_4}{W_6/L_6}$$

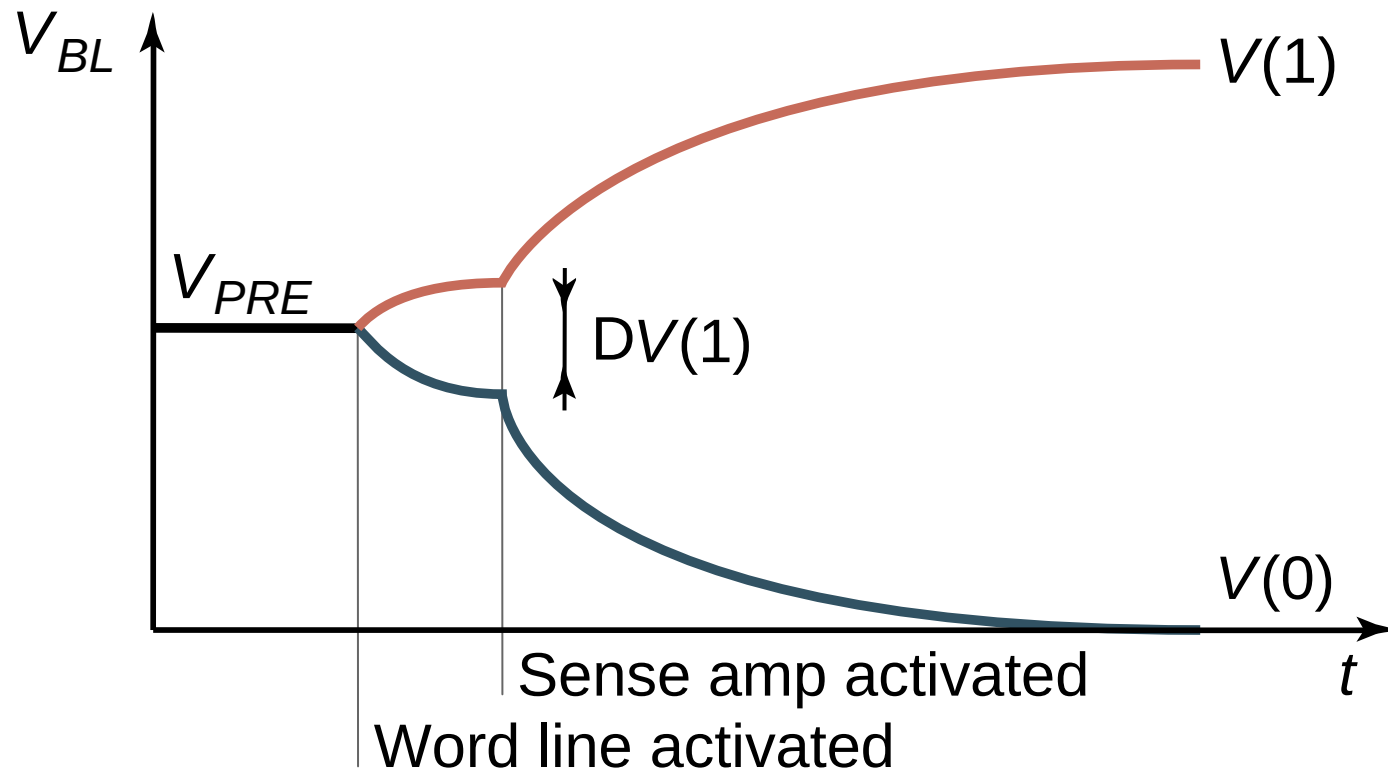
Should
be low
to keep
 V_Q low

PR between the PMOS (M4) pull-up and the NMOS (M6) Pass Transistor must be < 1.8 to keep $V_{tn} < 0.4$ V

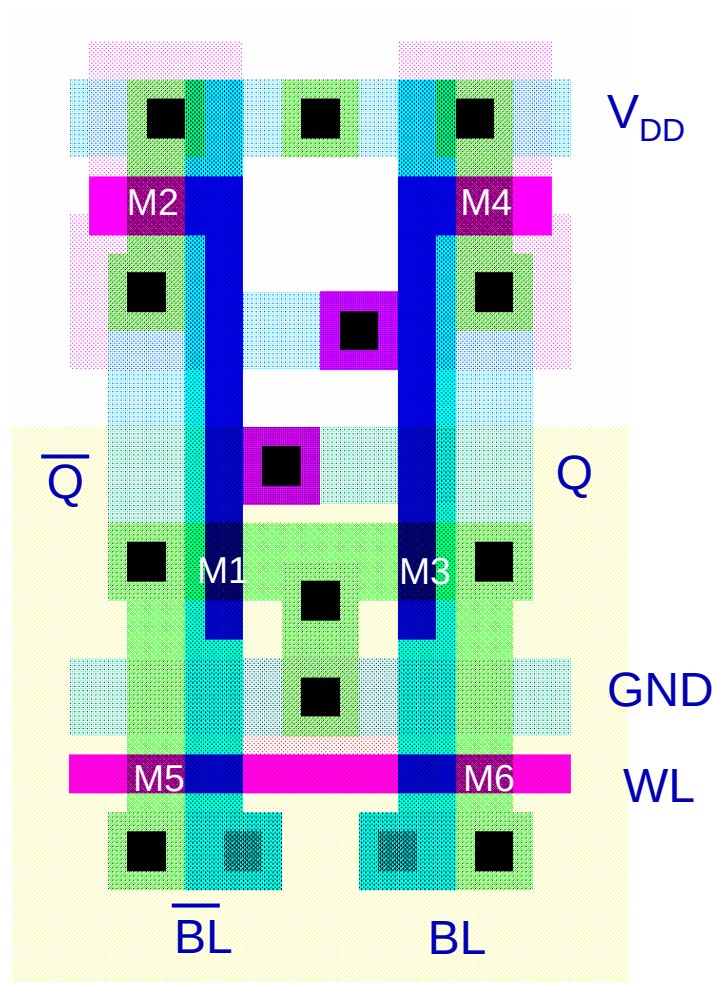
Performance of SRAM

- ❑ Read operation is more critical. It requires discharging of the large bit line capacitance through the stack of 2 transistors (M1-M5)
- ❑ Write time is dominated by the propagation delay of the cross-coupled inverter pair, since the drivers that set BL and $\overline{BL}$ can be large
- ❑ Sense amplifiers used to accelerate Read time....as the difference between BL and $\overline{BL}$ builds up, sense amplifier is activated, and it discharges one of the bit lines

Sense Amp Operation



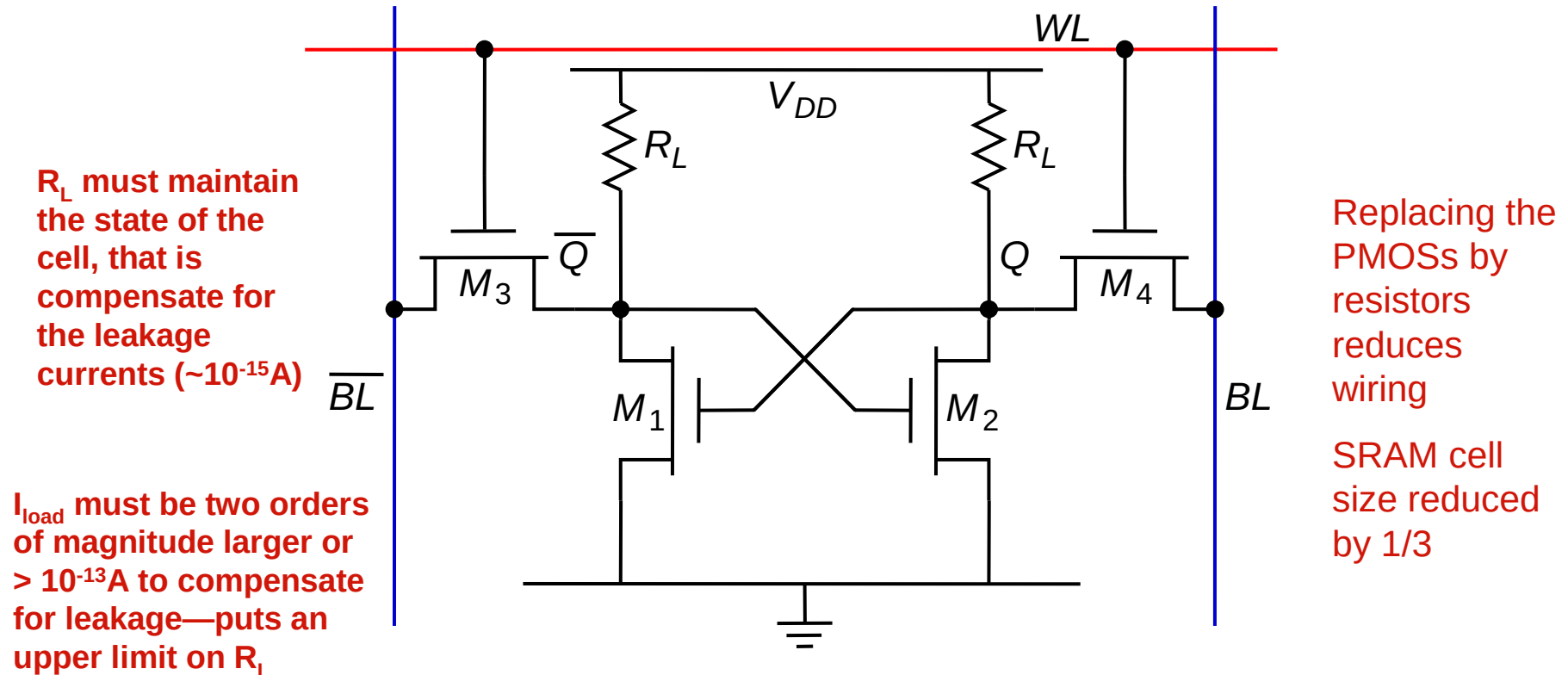
6T-SRAM — Layout



6T SRAM
Takes
significant
area...the two
PMOS need
n-wells

Resistive-load (4T) SRAM Cell

Reduce area using resistive load inverters...simplifies writing



Static power dissipation -- Want R_L large (use undoped poly)
Bit lines precharged to V_{DD} to address t_p problem

SRAM Characteristics

Instead of PMOS devices, use parasitic devices on top of cell structure using thin-film transistors (TFTs)

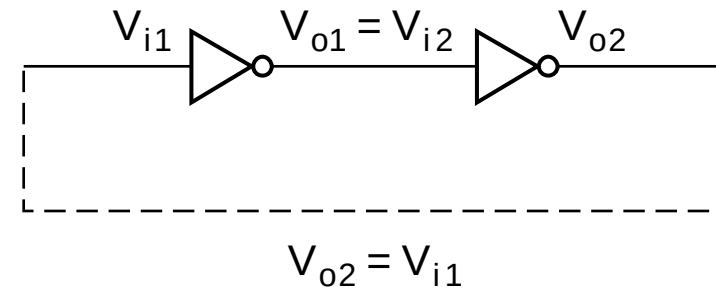
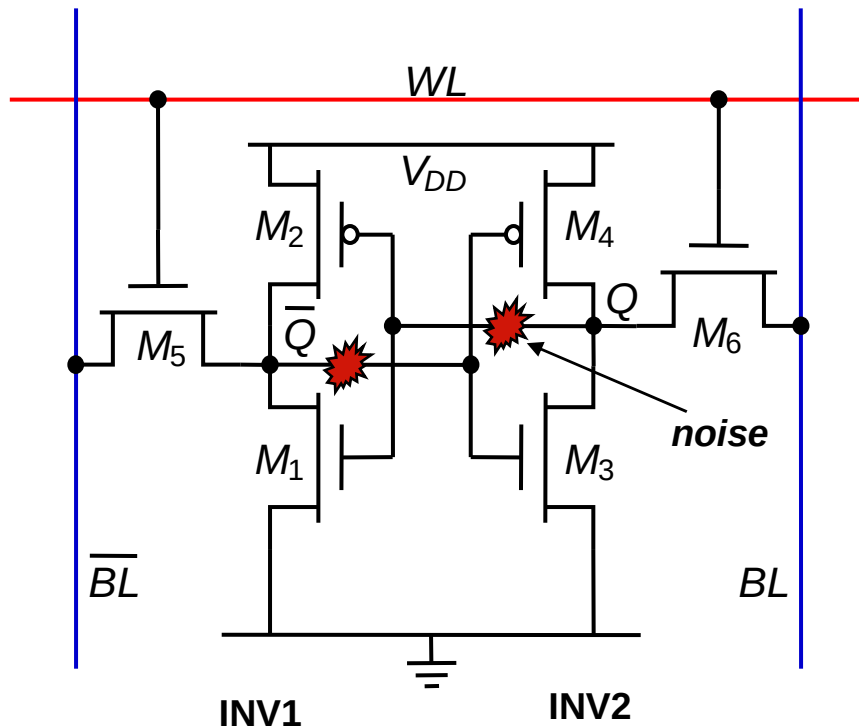
Table 12-2 Comparison of CMOS SRAM cells used in 1-Mbit memory (from [Takada91])

	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

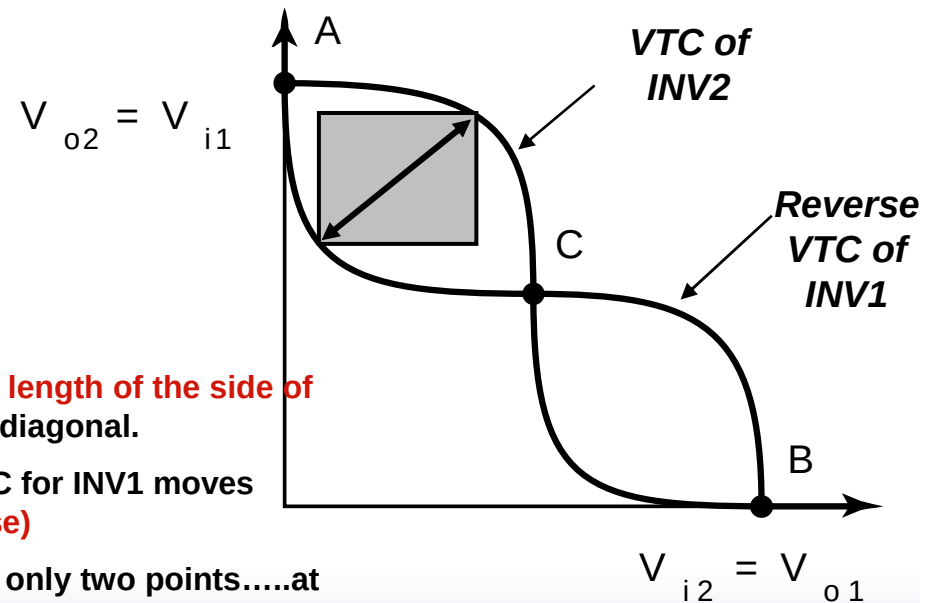
Use high V_t

However, embedded SRAM cells---used in processor caches, employ 6T cells.

6-T CMOS SRAM Cell: Static Noise Margin



Butterfly Curve



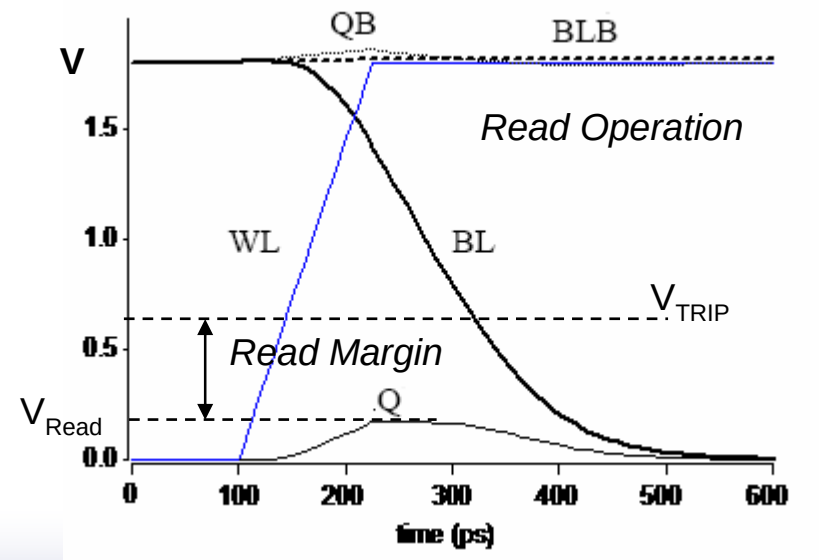
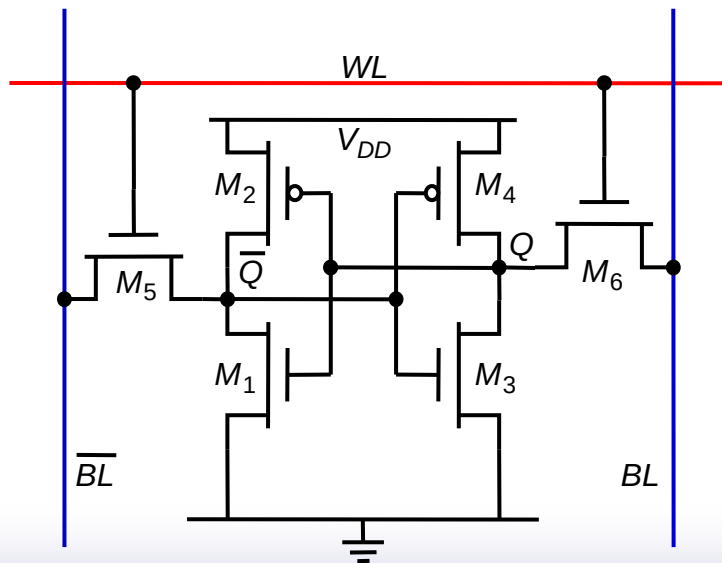
The SNM (hold margin) can be estimated graphically by the **length of the side of the square** fitted between the VTCs and having the longest diagonal.

As noise increases at the two nodes above, the reverse VTC for INV1 moves upward, while the VTC for INV2 moves to the left (**worst case**)

Once they both move by the SNM value, the curves meet at only two points.....at A' and B'..... and any further noise flips the data.

Static Noise Margin (SNM)

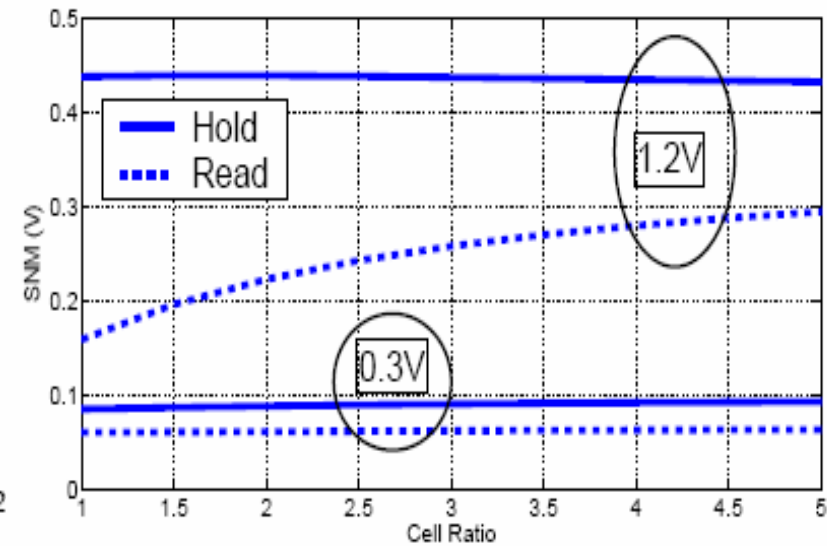
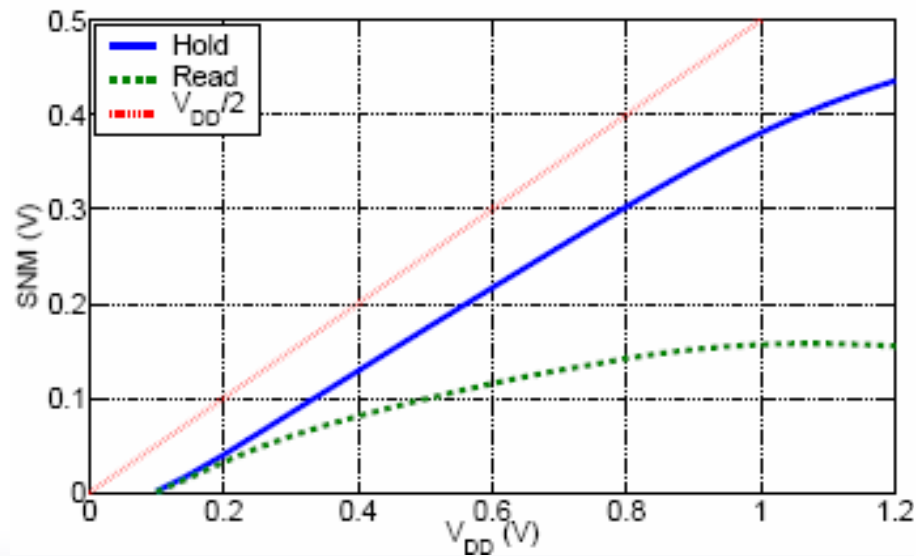
- ❑ **Hold Margin:** How strongly the node storing '1' and the node storing '0' are coupled to V_{DD} and V_{SS} respectively.
- ❑ **Read Margin:** The difference between V_{TRIP} and V_{READ} (max. voltage at Q)
- ❑ **Write Margin:** The maximum voltage on a bit-line that allows writing to the cell, while the other bit-line is at V_{DD} . (not determined by the butterfly curve)



SNM Dependencies

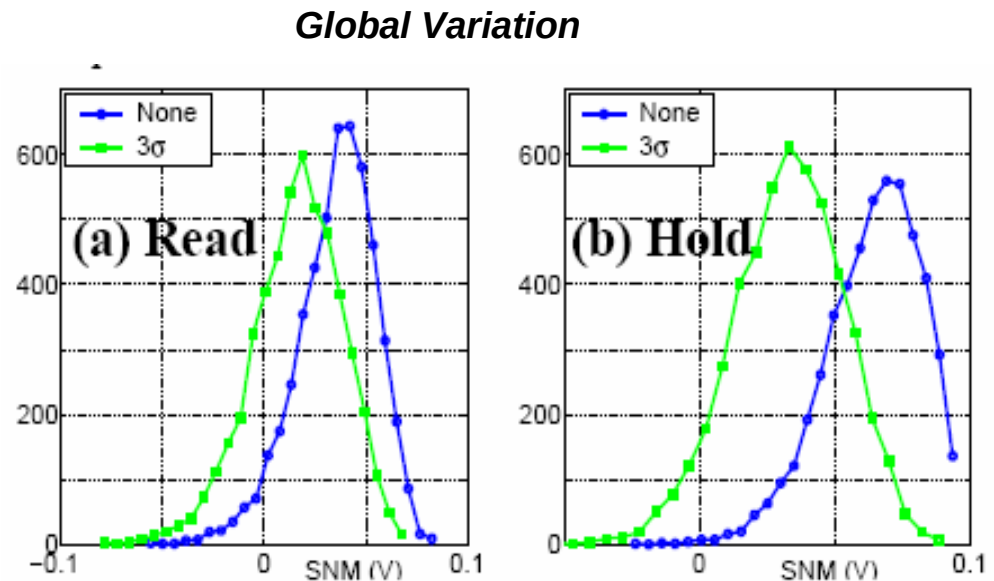
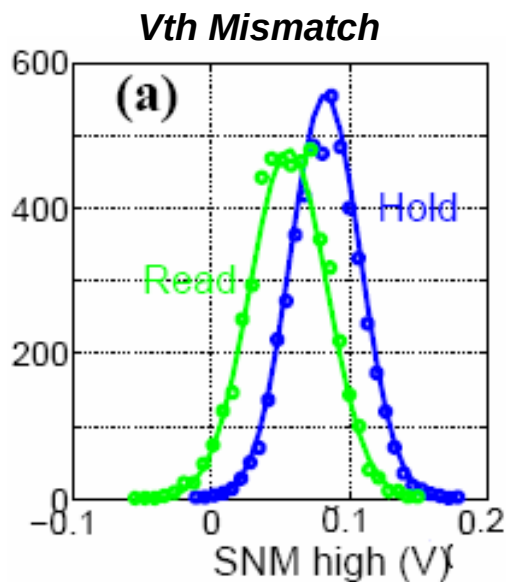
- Dependence on V_{DD} : SNM for a bitcell with ideal VTCs is still limited to $V_{DD}/2$
- Dependence on sizing

Here, Cell ratio = size of PD device over size of access device



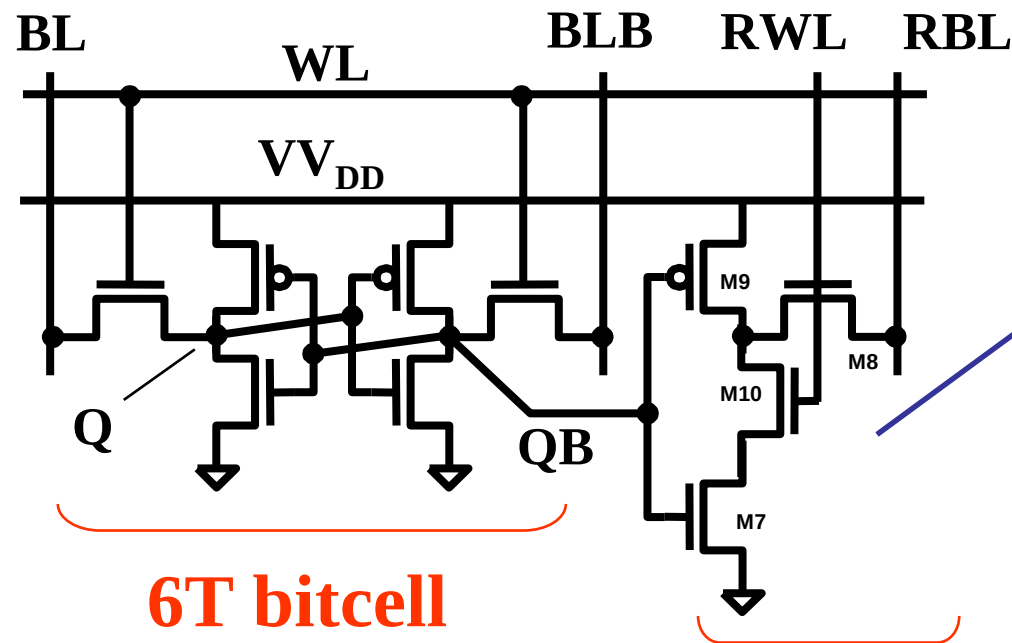
SNM Dependencies

- ❑ Dependence on random doping variation (V_{TH} mismatch)
- ❑ Dependence on global variation



10-T SRAM

- The idea is to add a 4T buffer at one side:



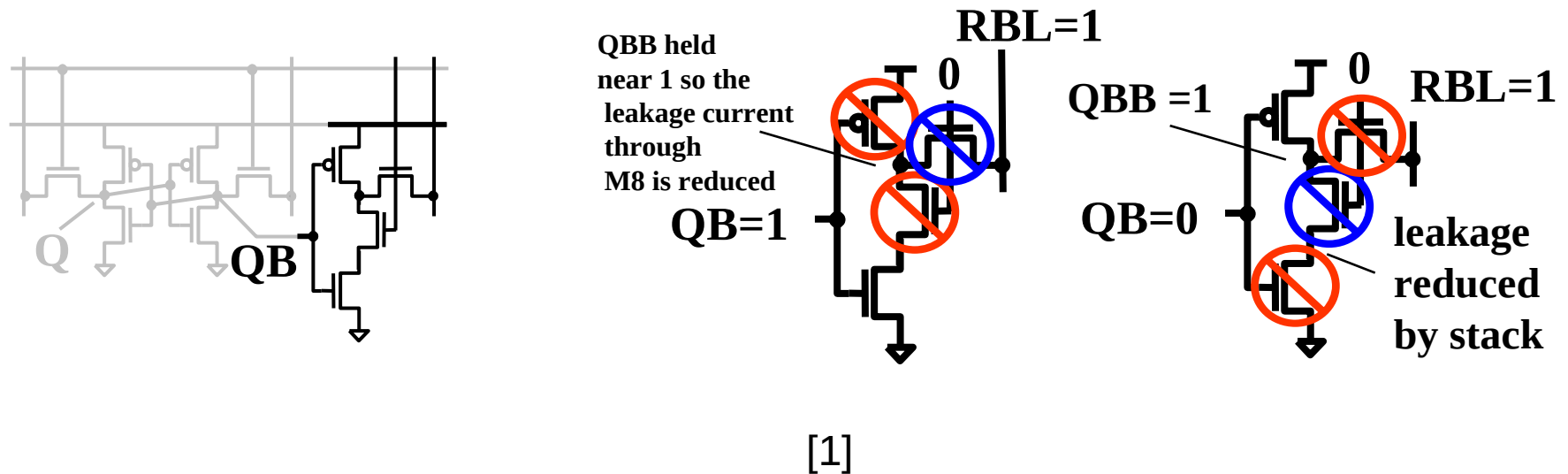
M7 to M10 to remove the problem of Read SNM by buffering the stored data during a read access.

Thus, the worst-case SNM for this bitcell is the Hold SNM related to M1 to M6, which is the same as the 6T Hold SNM for same sized M1 to M6

Proposed 10-T bitcell for Sub-VT[1]

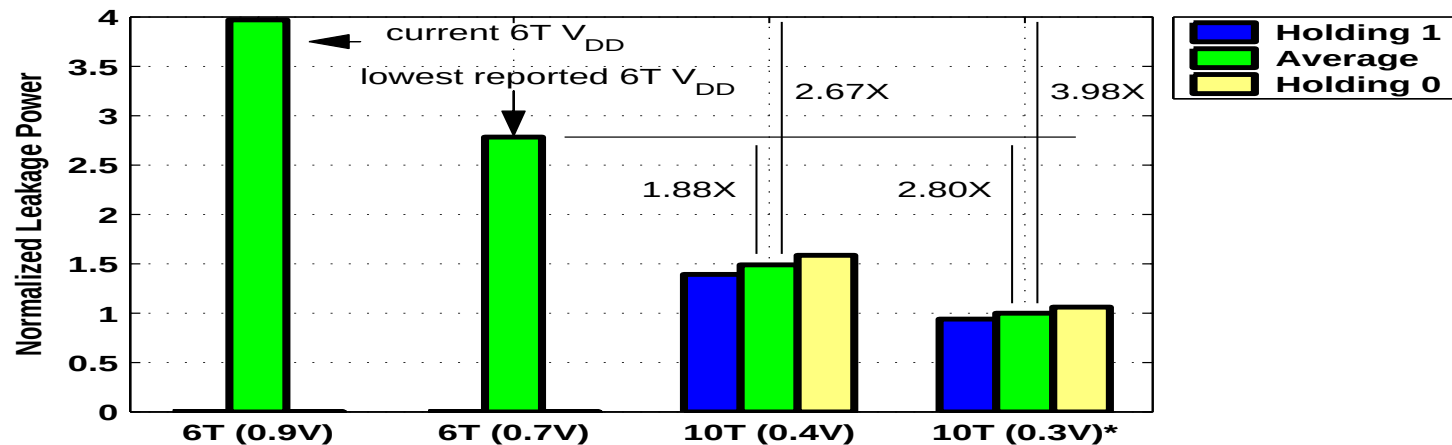
4T buffer

10T Bitcell Reduces Bitline Leakage



for iso-VDD, the 10T cell without M10 (a 9T cell) has 50% higher leakage current than the 6T, but adding M10 drops the overhead to 16%.

Leakage Power Savings with 10T Bitcell



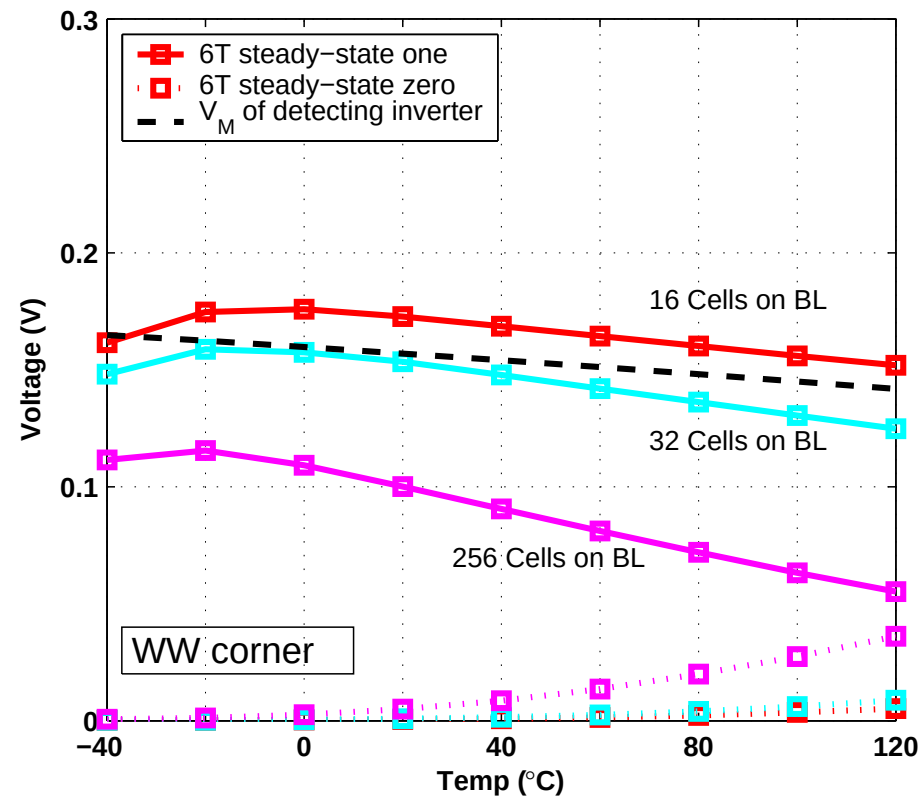
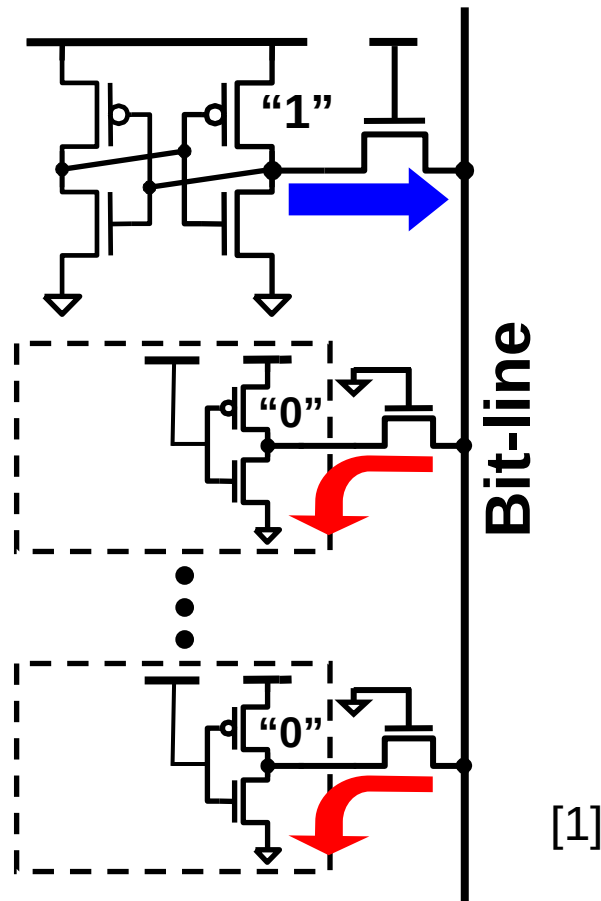
❑ 6T memories in 65nm usually at 0.9V or greater (lowest reported is 0.7V)

[1]

❑ 10T bitcell allows scaling to lower voltages

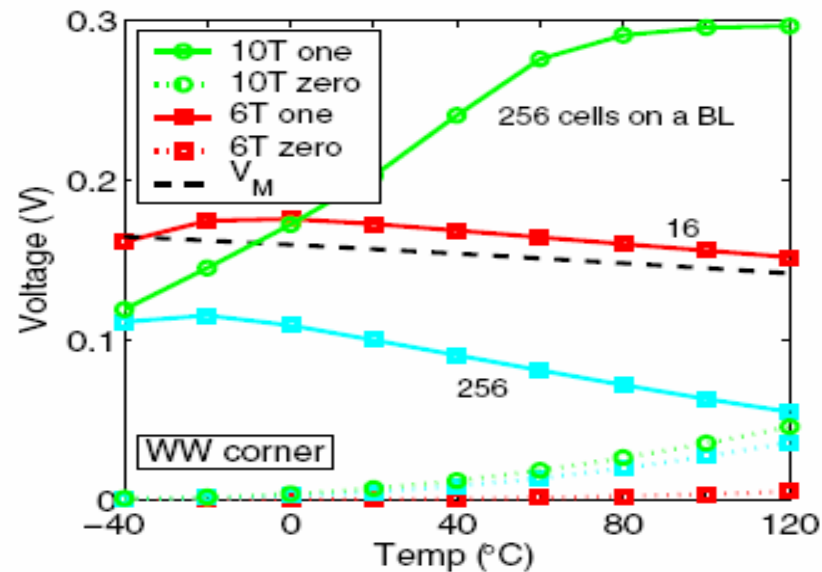
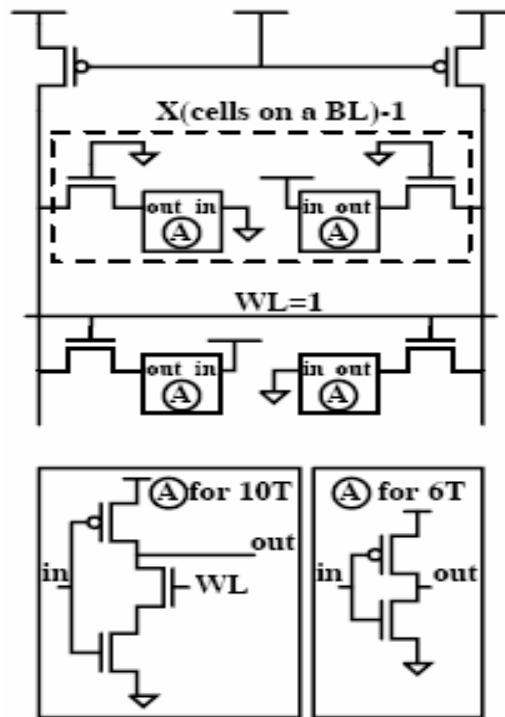
❑ Lower voltage operation reduces leakage power dramatically for unaccessed cells

Bitline Leakage Limits Integration Level



16 bitcells on bitline is best can hope for standard 6T

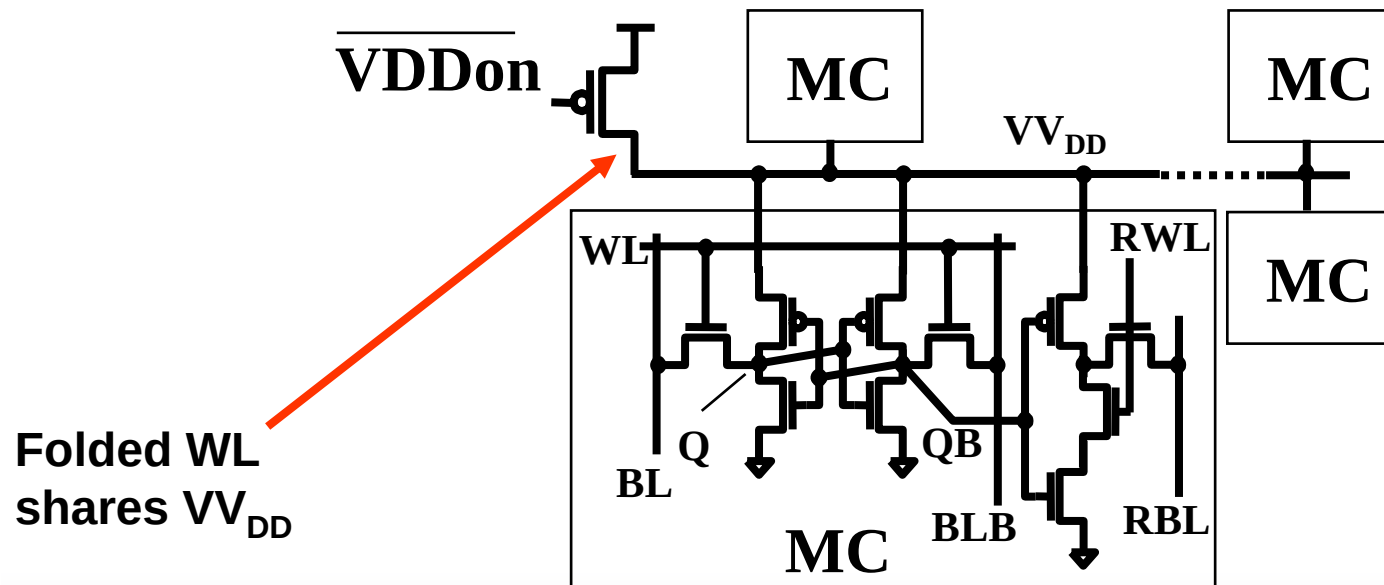
Cont'd



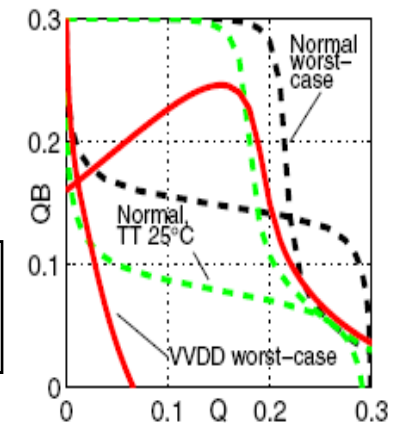
BL leakage limits the number of cells on a BL. The 10T bitcell can sustain 256 cells/BL at 0.3V compared to 16 without M10 (6T or 9T). higher level of integration allowed by the 10T cell reduces the peripheral circuits and slightly mitigates the bitcell area overhead[1].

10T Bitcell Allows Sub-VT Write

To achieve write in sub-threshold, the virtual supply (VV_{DD}) to the selected cells floats during the write operation

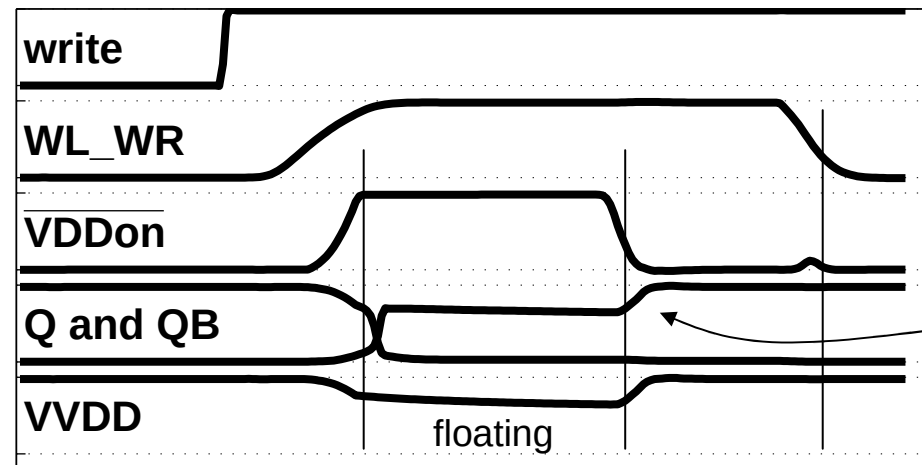


[1]



Cont'd

Floating V_{DD} weakens feedback and allows Write.



feedback restores '1' to V_{DD}

A virtual supply voltage (VVDD) that floats during write allows robust write operation into sub-VT (mono-stable butterfly curve). VVDD stops floating while WL_WR remains asserted to restore the '1' value to full VDD[1].