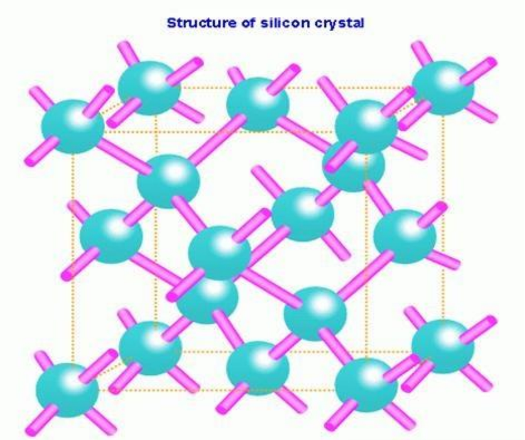
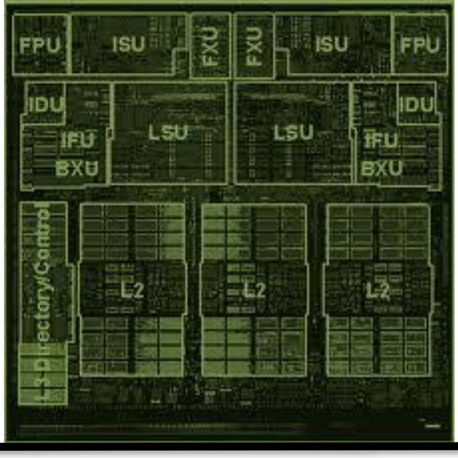
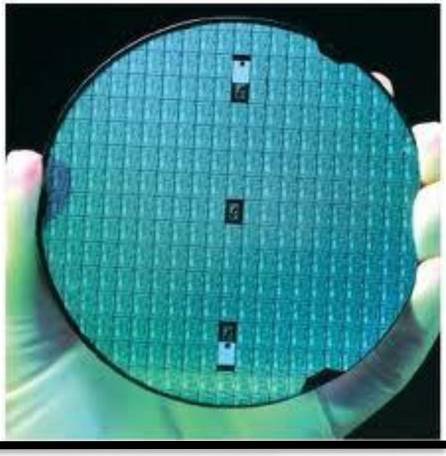






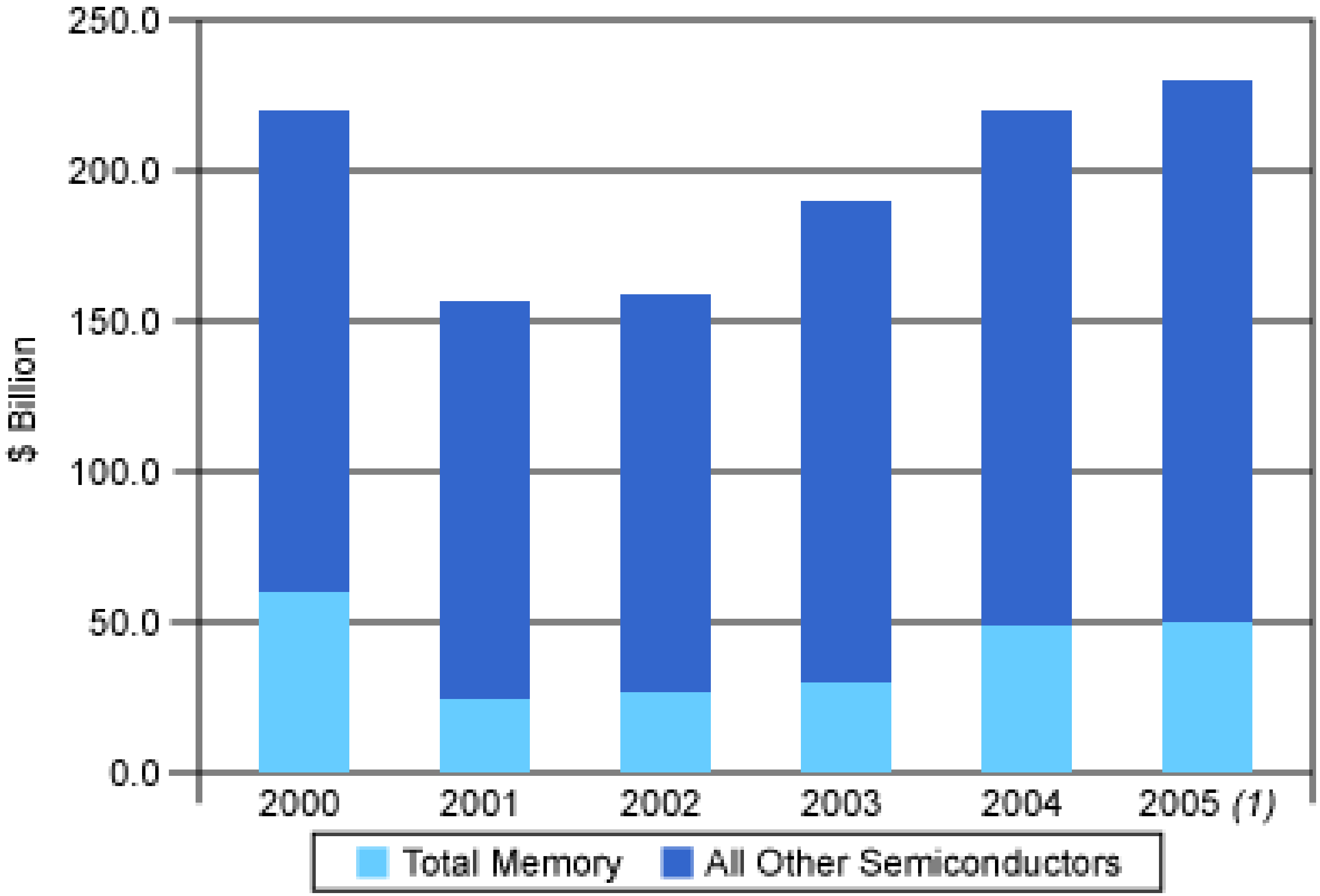
ECE 122A

VLSI Principles

Lecture 15

Prof. Kaustav Banerjee
Electrical and Computer Engineering
University of California, Santa Barbara
E-mail: kaustav@ece.ucsb.edu

Semiconductor Memories….



Memory Design...

- ❑ Increasing number of transistors in uprocessors are devoted to cache memories….more than 60%, see IRDS for more details…..

- ❑ At the system level: high-performance workstations and desktops have several Terra-bytes 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

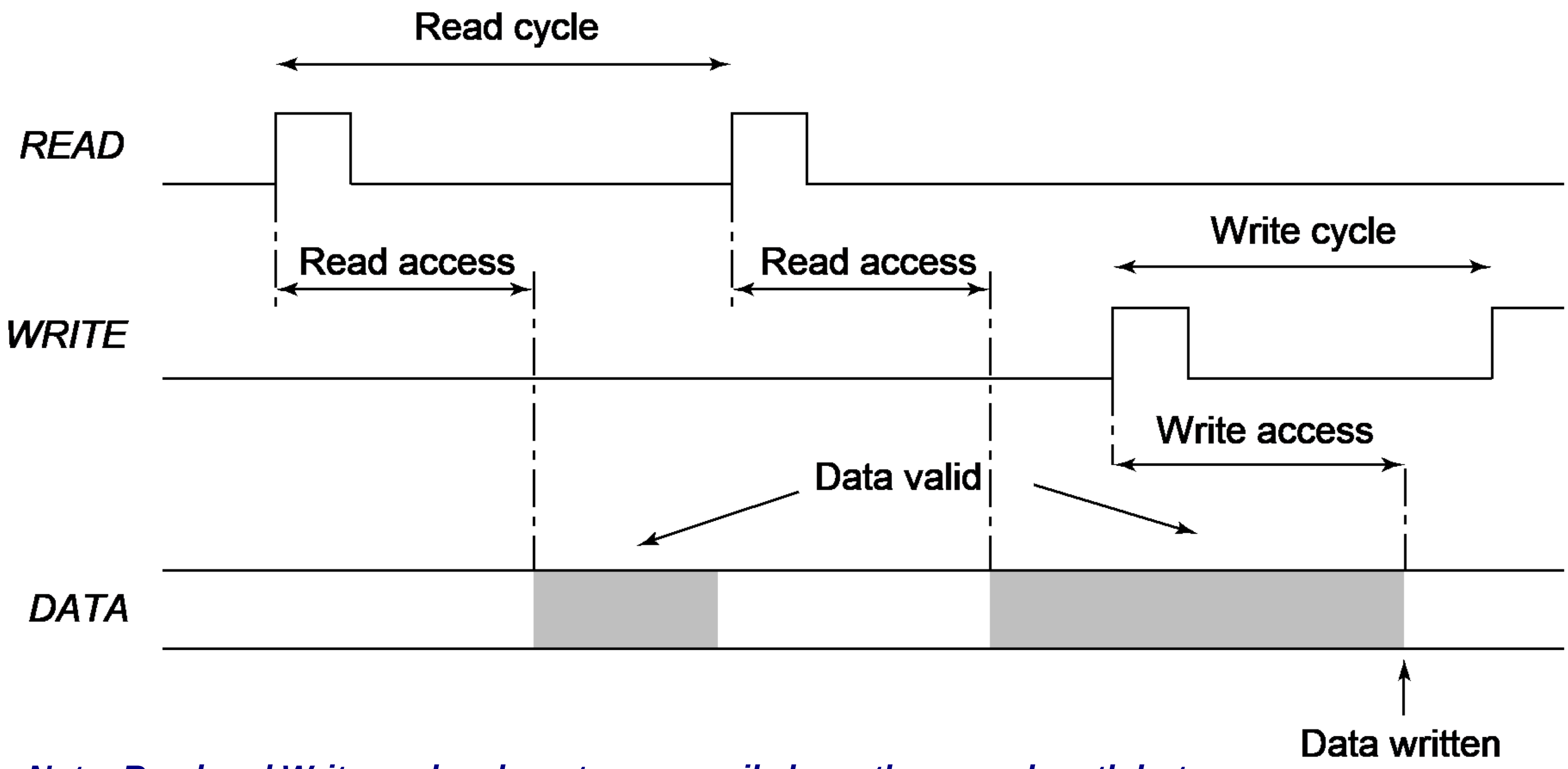
Memory Size

- 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
  - E^{2}PROM: 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

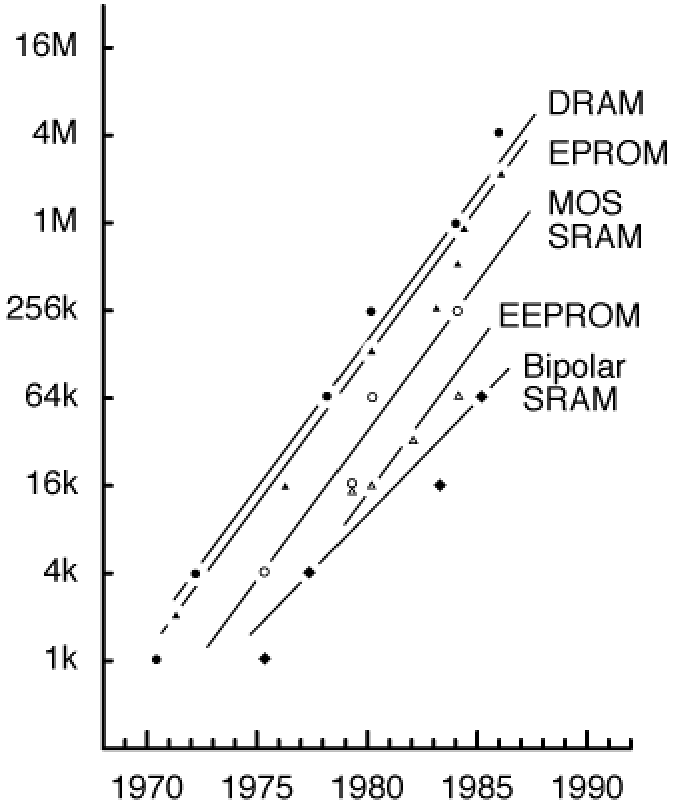
Read-Write Memory		Non-Volatile Read-Write Memory	Read-Only Memory
Random Access	**Non-Random Access**	EPROM E^2PROM FLASH	Mask-Programmed Programmable (PROM)
SRAM DRAM	FIFO LIFO Shift Register CAM		

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

More Classification

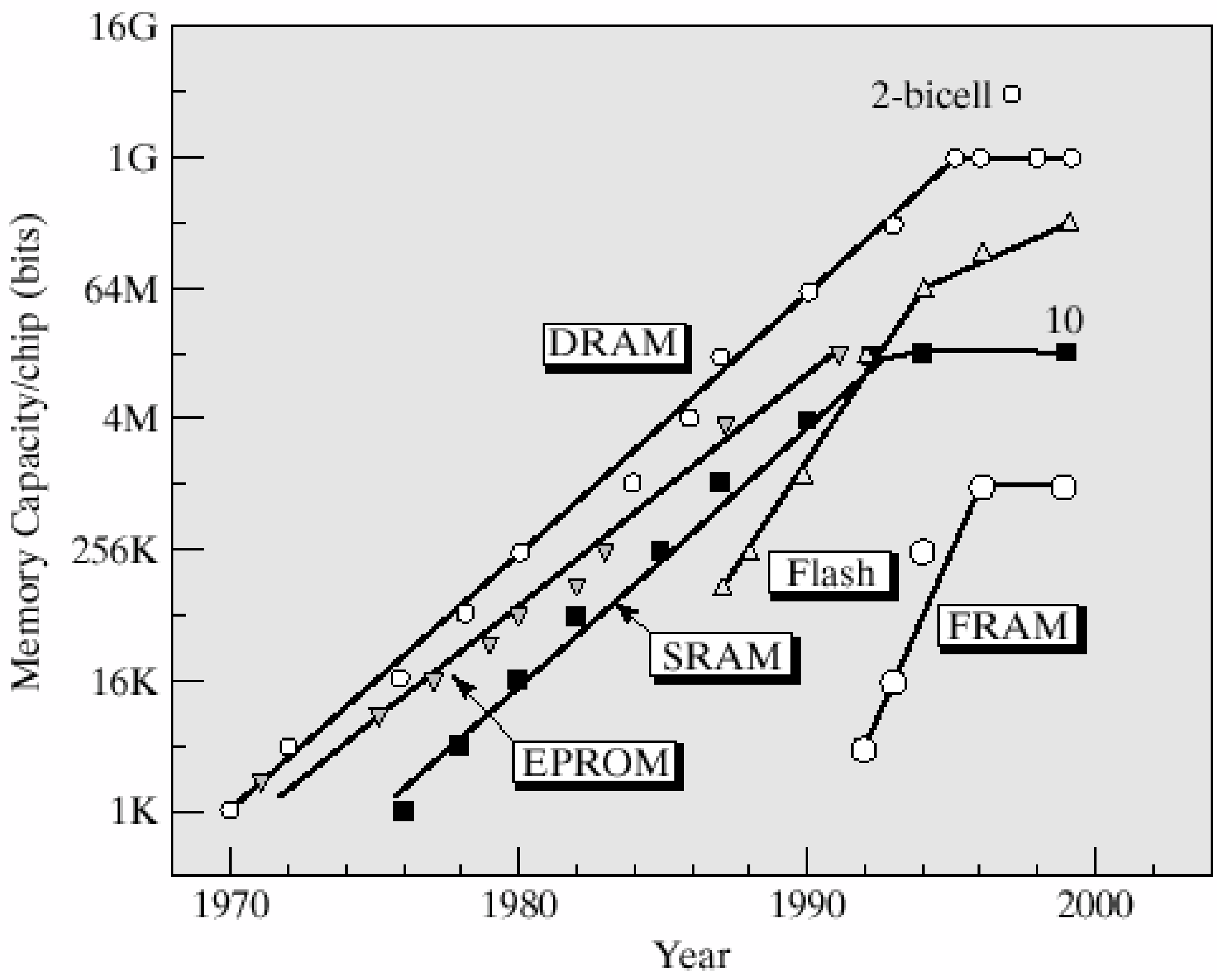
- **I/O Architecture:**
  - Based on the number of data input and output ports
  - Most memories use a single I/O port
  - **Multiport memories** offer higher bandwidth
    - Example: register files used in RISC processors
    - Adds more complexity to the design
- **Application:**
  - Embedded Memories in SoCs
  - For massive storage (multiples of Tbytes 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)



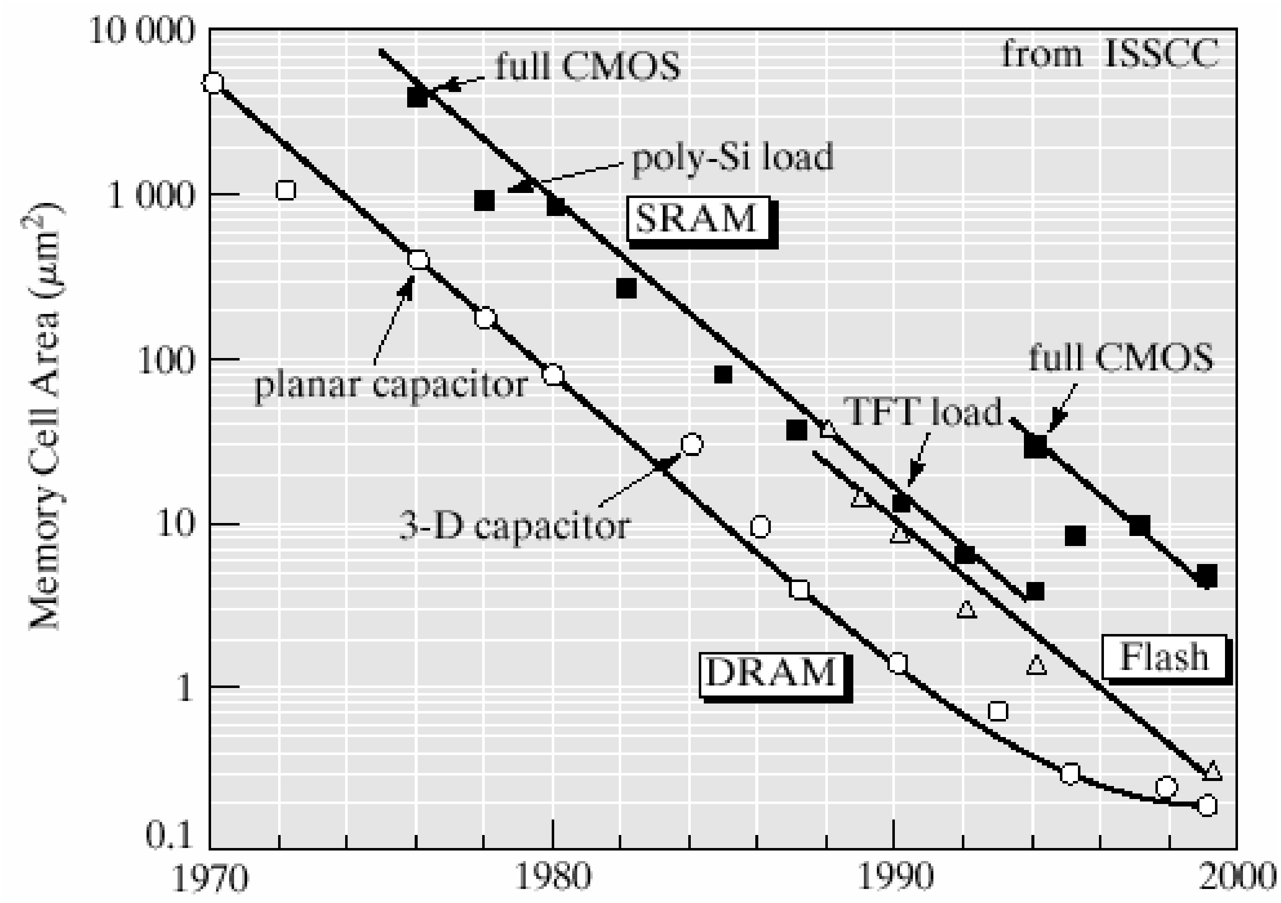
Memory Size as a function of time: x 4 every three years

Semiconductor Memory Trends (more recent…)



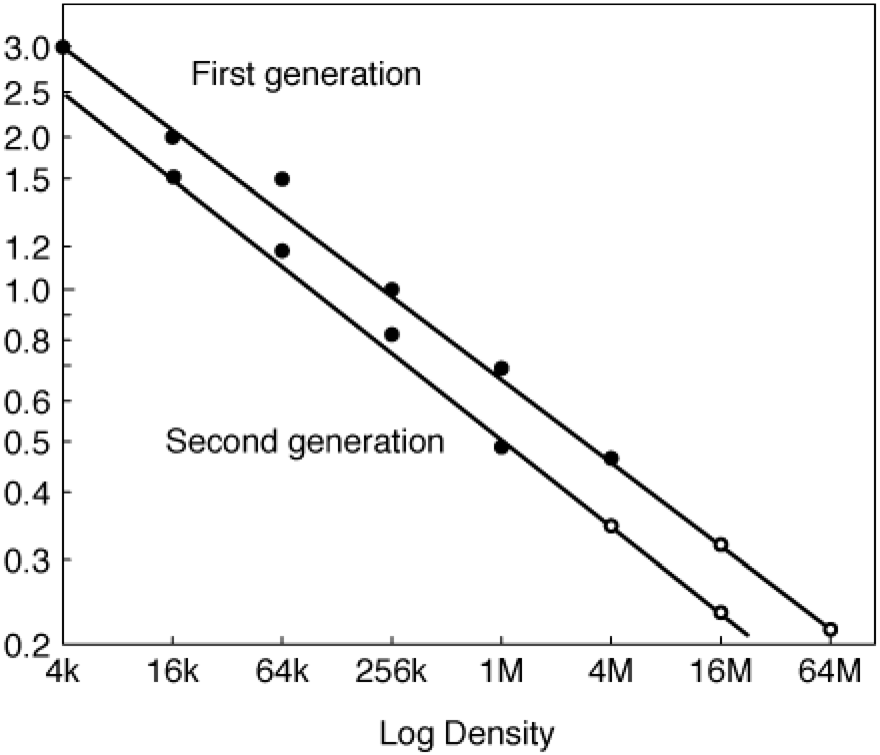
From [Itoh01]

Trends in Memory Cell Area



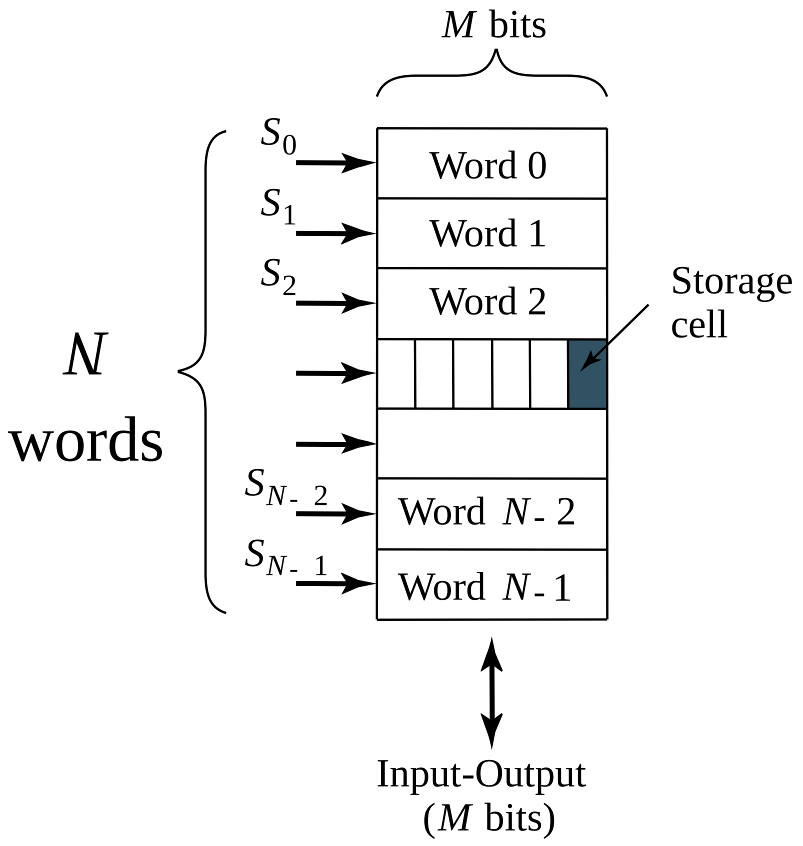
From [Itoh01]

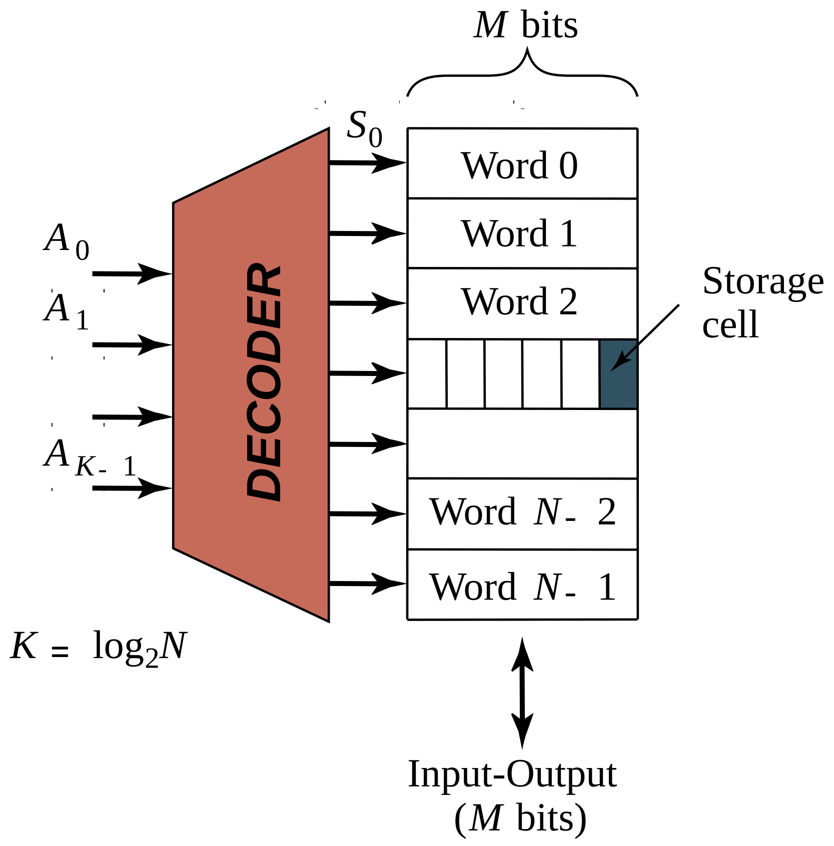
Semiconductor Memory Trends



Technology feature size for different SRAM generations

Memory Architecture: Decoders



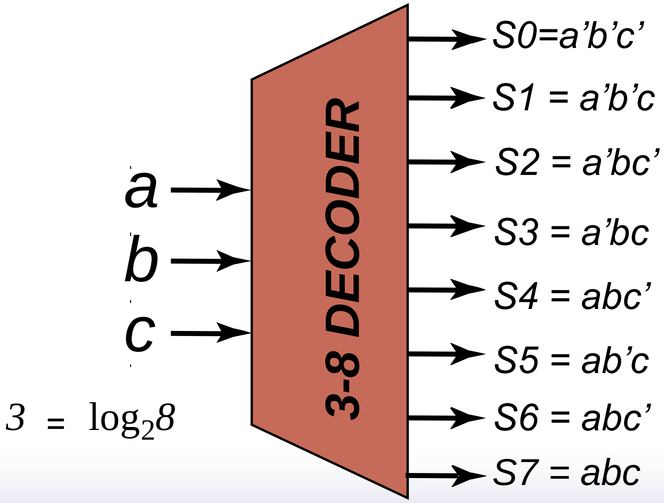


Intuitive architecture for N x M memory
Too many select signals:
N words == N select signals

Decoder reduces the number of select signals
$K = log_2N$

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.



a

b

c

3-8 DECODER

S0=a'b'c'

S1 = a'b'c

S2 = a'bc'

S3 = a'bc

S4 = abc'

S5 = ab'c

S6 = abc'

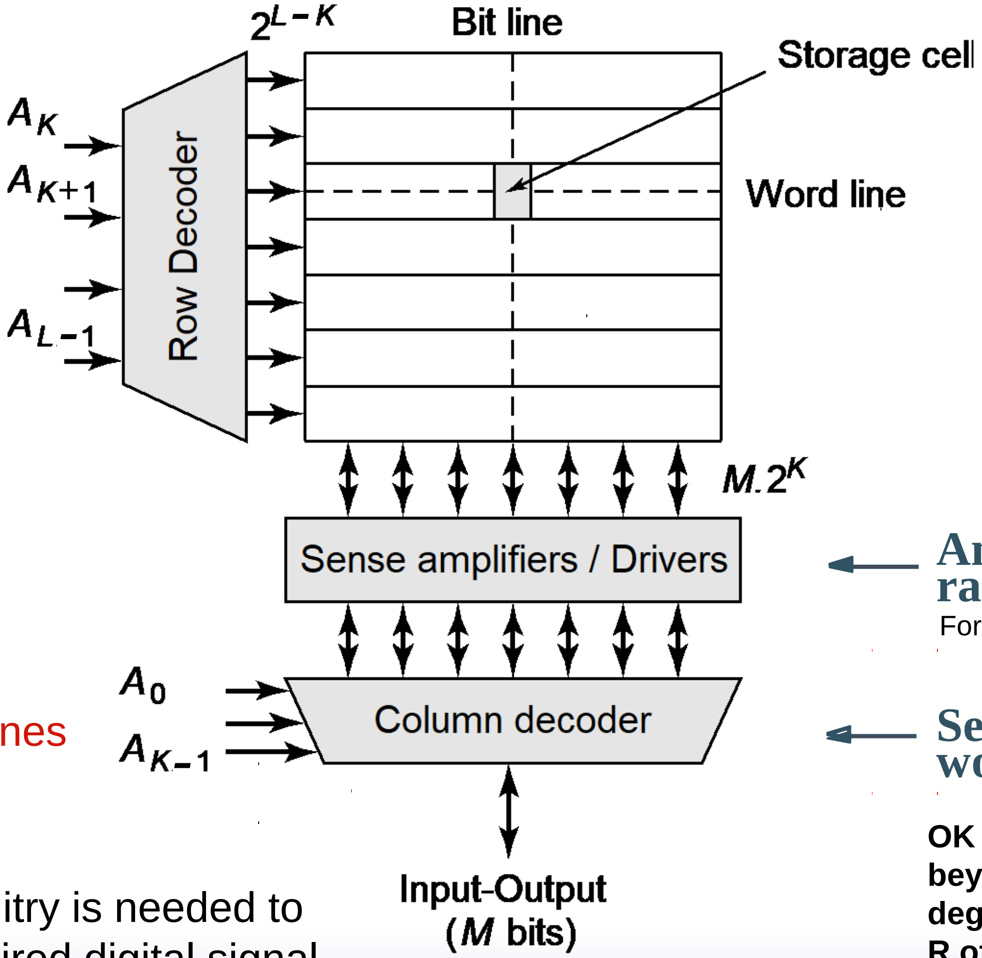
S7 = abc

$3 = \log_2 8$

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 >> WIDTH, cannot be implemented and will result in very slow design…..**



L-K WORD lines

$K = log_2N$

K BIT lines

Solution: **Make vertical and horizontal dimensions of the same order of magnitude**

Store multiple words in one row

Use a column decoder to select the correct word

Amplify swing to rail-to-rail amplitude

For interfacing to the external world

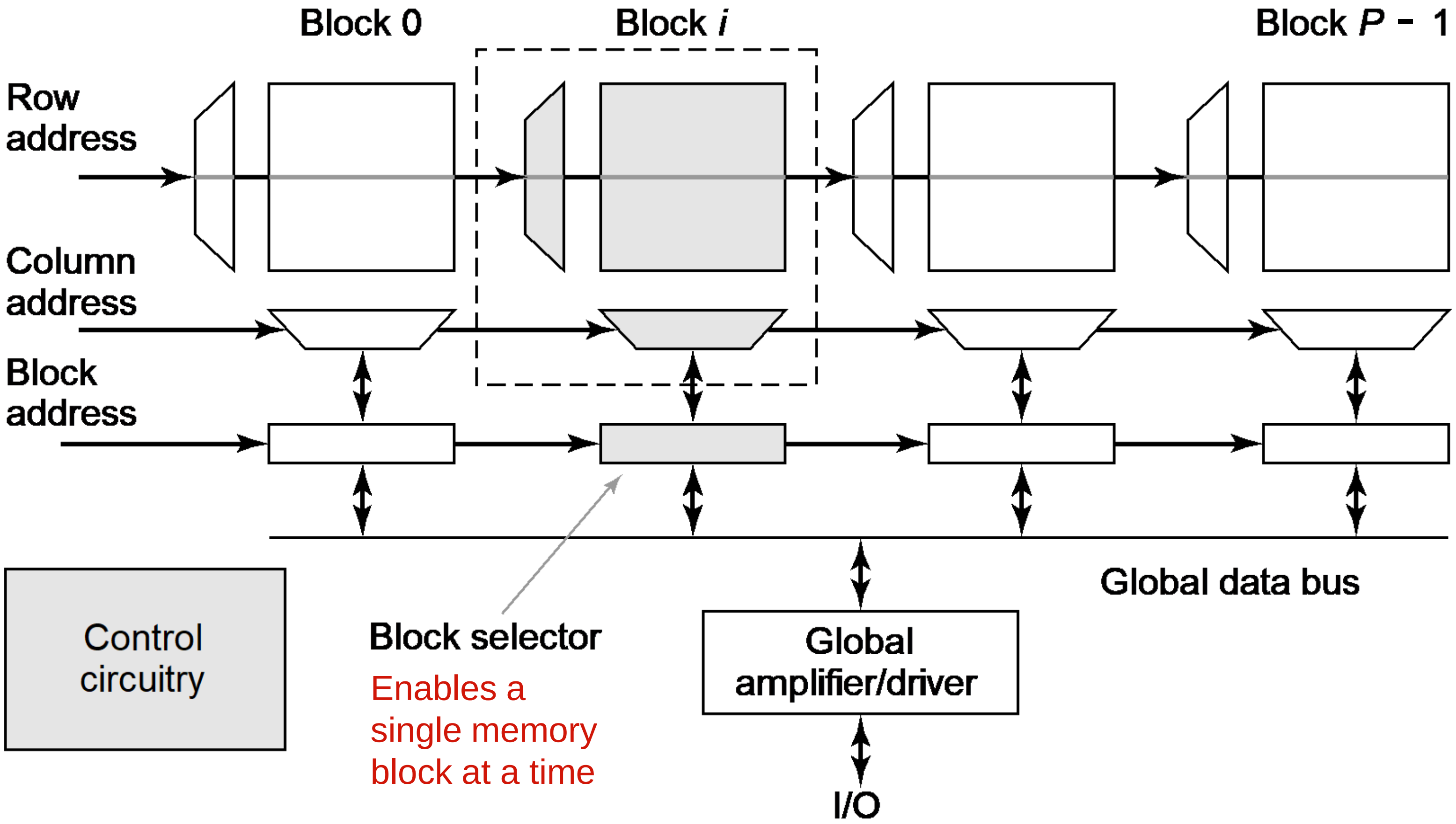
Selects appropriate word

OK for 64 Kbits to 256 Kbits beyond which speed degrades as length, C, and R of word/bit lines increase excessively

Peripheral circuitry is needed to recover the desired digital signal properties

Hierarchical Memory Architecture

For Larger Memories….



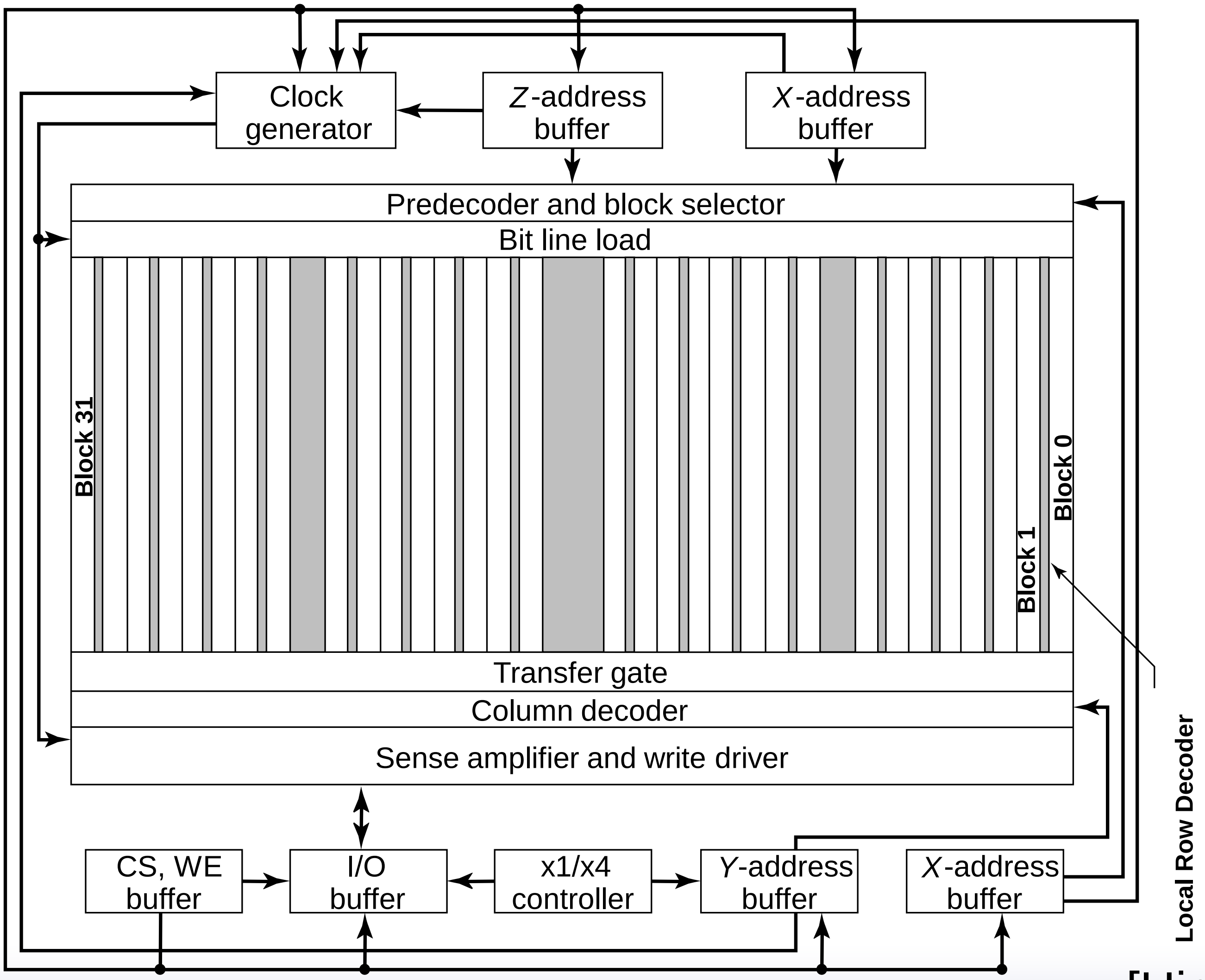
Enables a single memory block at a time

Advantages:

1. Shorter wires within blocks: faster access times

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

Block Diagram of 4 Mbit SRAM



[Hirose90]

32 blocks, each containing 128 Kbits

Each block is structured as an array of 1024 rows and 128 columns

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

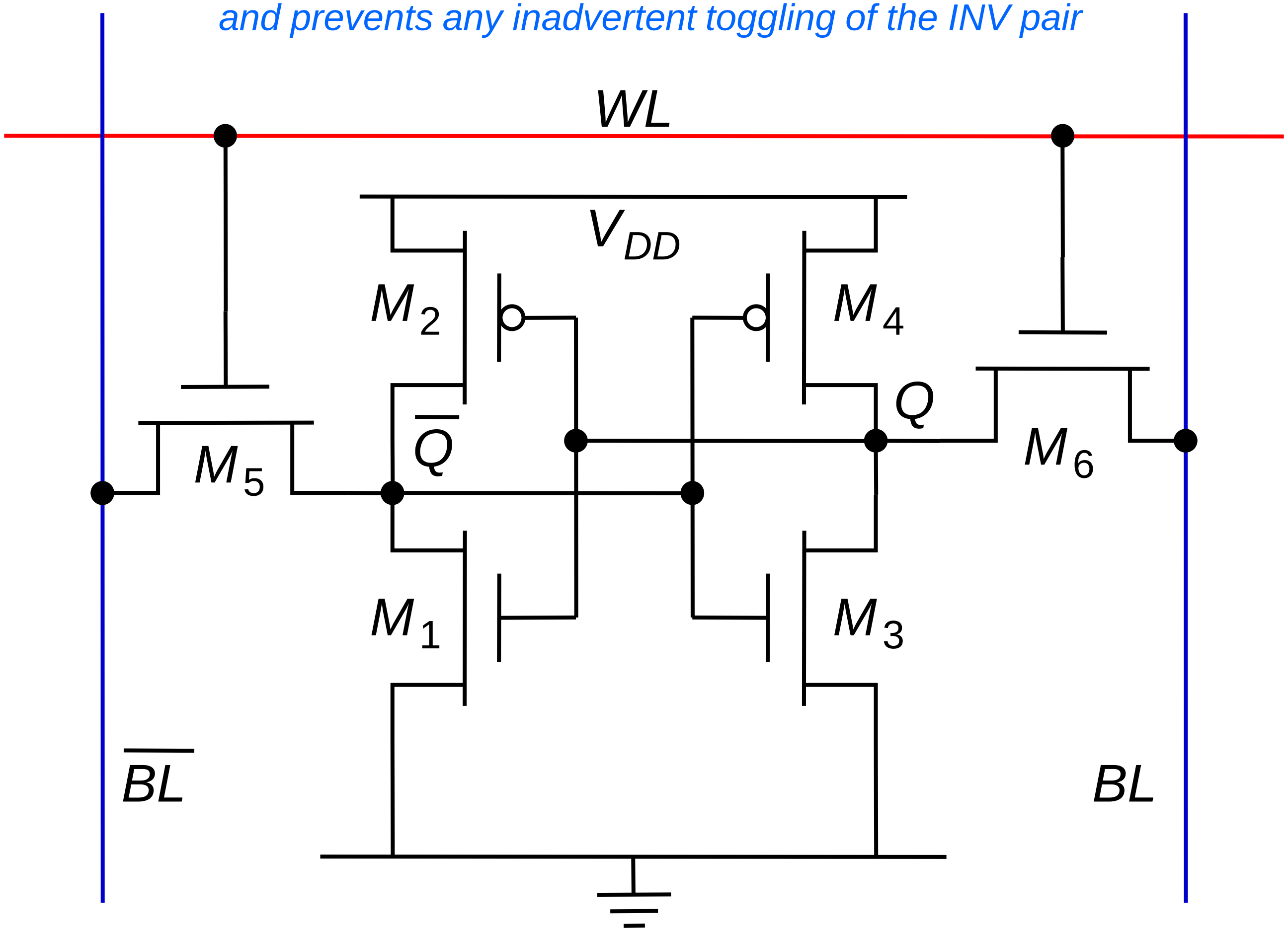
Assume both BLs are held high before the read.

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 Vdd by BL and prevents any inadvertent toggling of the INV pair

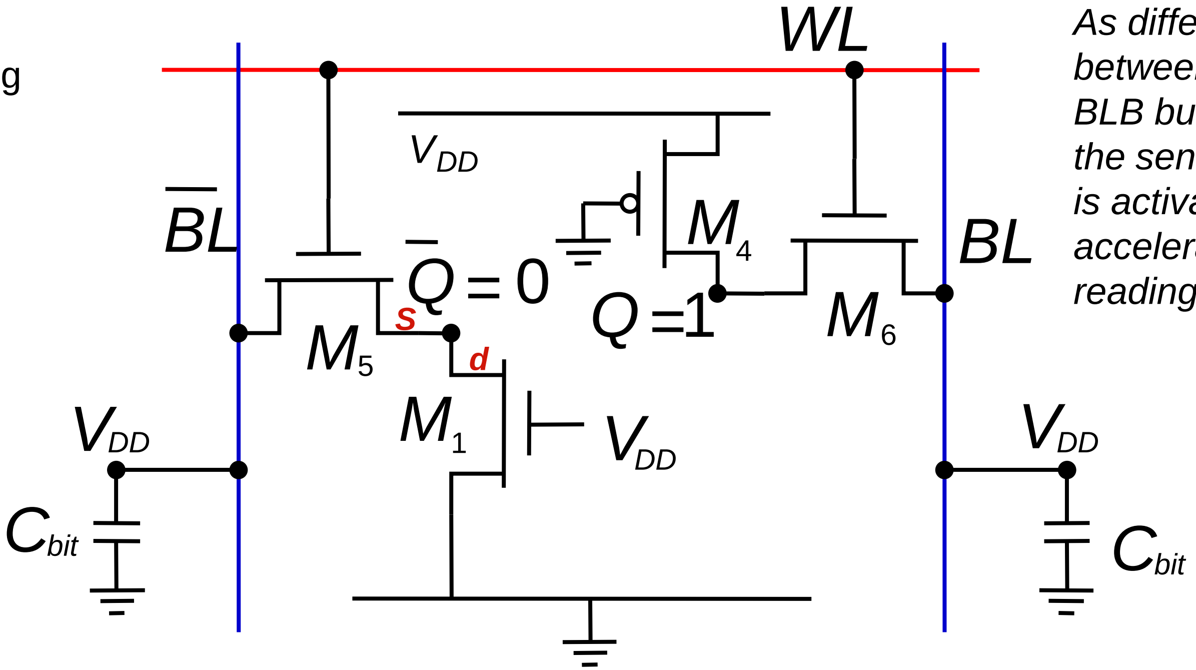


SRAM cell should be as small as possible…..but reliable operation requires careful sizing…

CMOS SRAM Analysis (Read "1" operation)

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

M1 must be stronger than M5



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

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

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

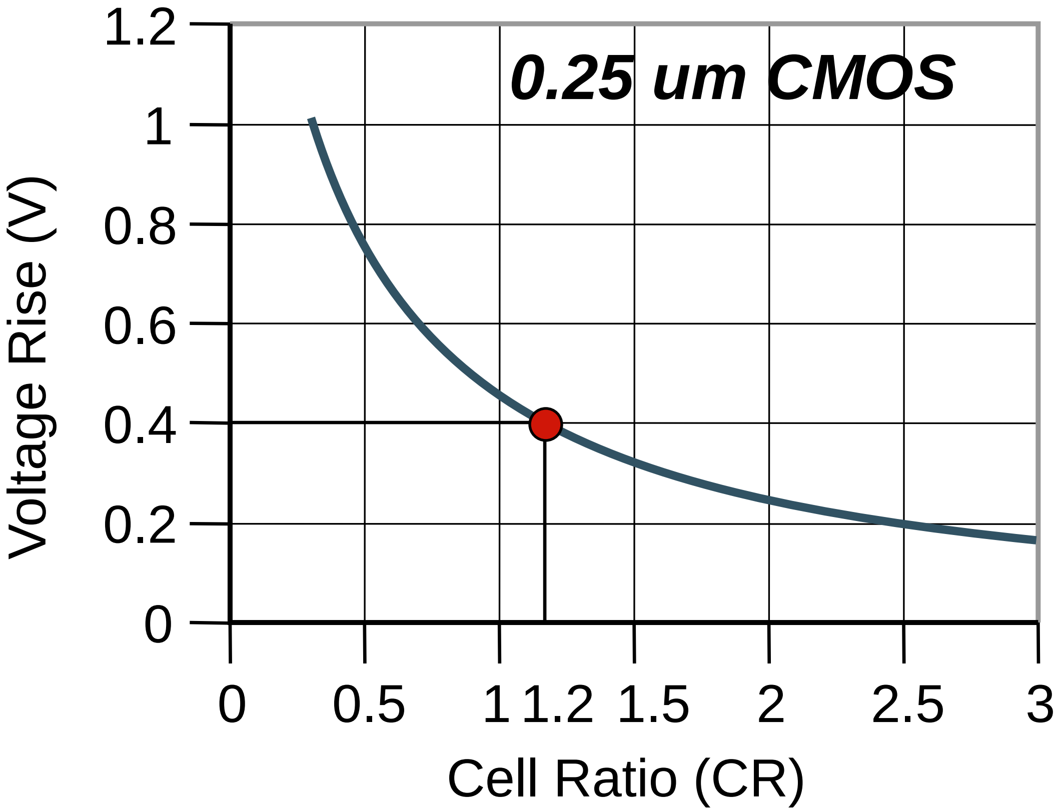
(M5 in saturation) *(M1 in linear)*

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

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

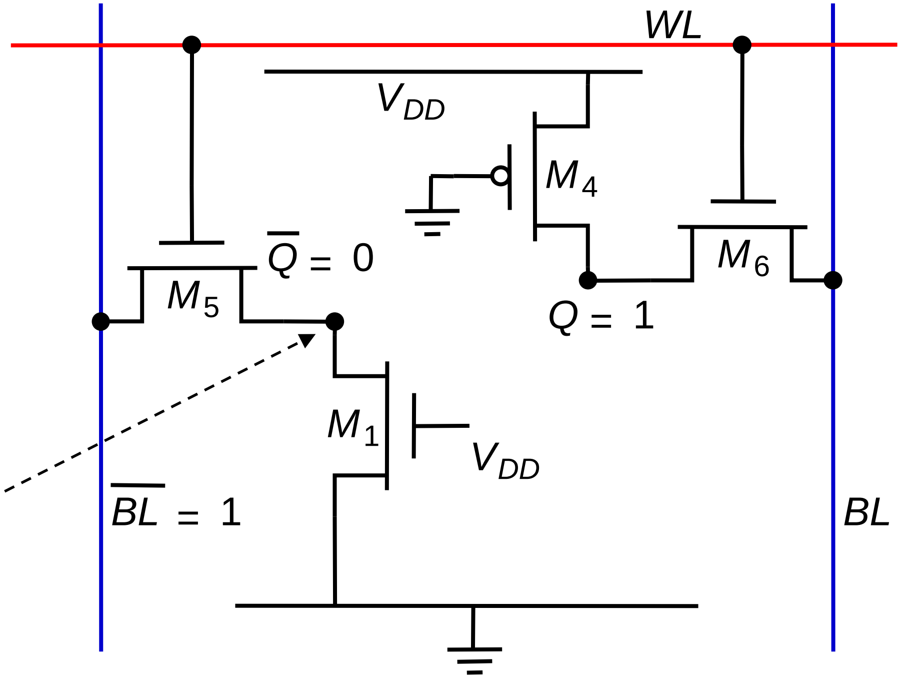
Node voltage must stay below the Vth of M3: CR must be >1.2

CMOS SRAM Analysis (Write)

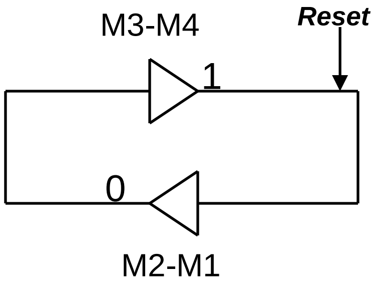
Assume that Q=1

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



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



Reliable writing of the cell is ensured if we can pull node Q low enough—below the Vth of M1

New value must be written through M6

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

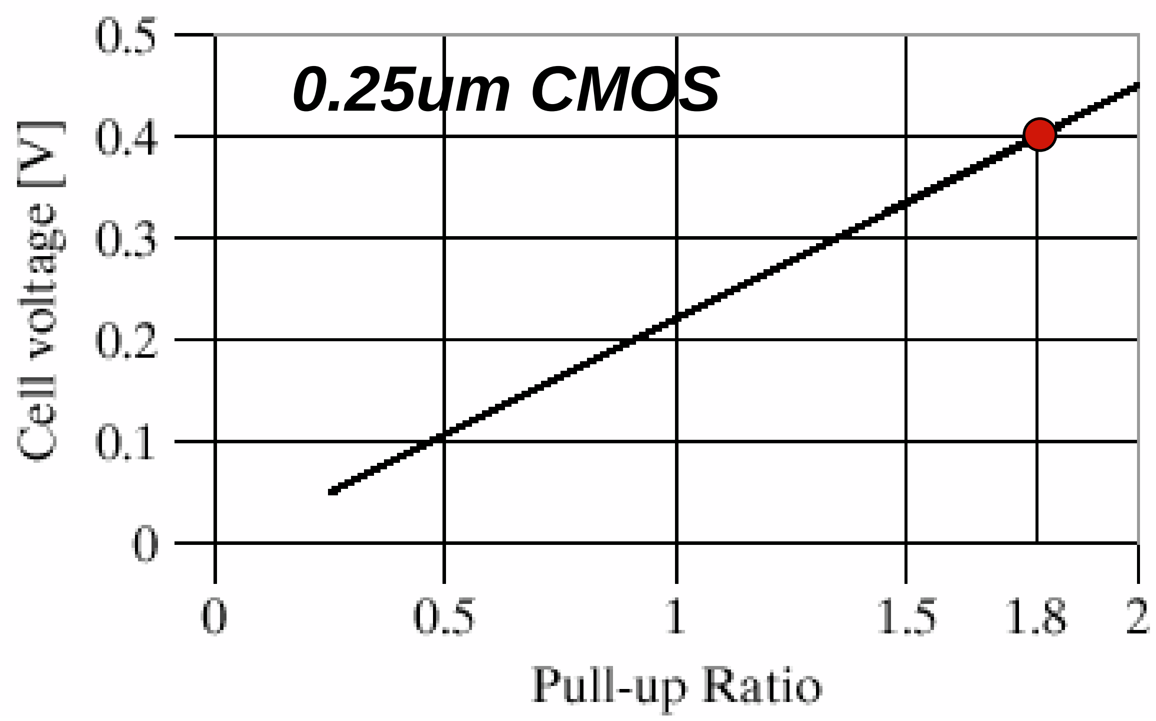
(M6 in linear) *(M4 in saturation)*

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

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



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

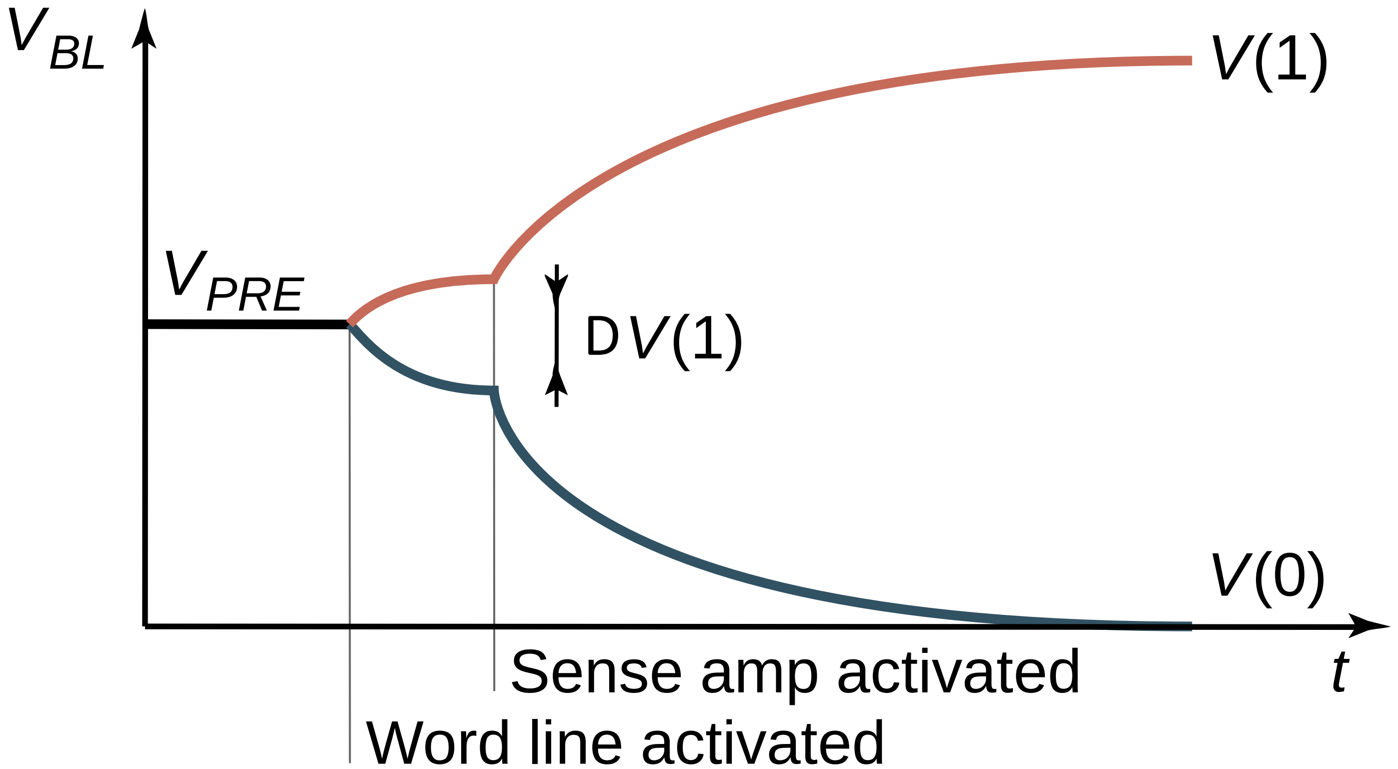
Should be low to keep V_Q low

PR between the PMOS (M4) pull-up and the NMOS (M6) Pass Transistor must be < 1.8 to keep Vtn < 0.4 V

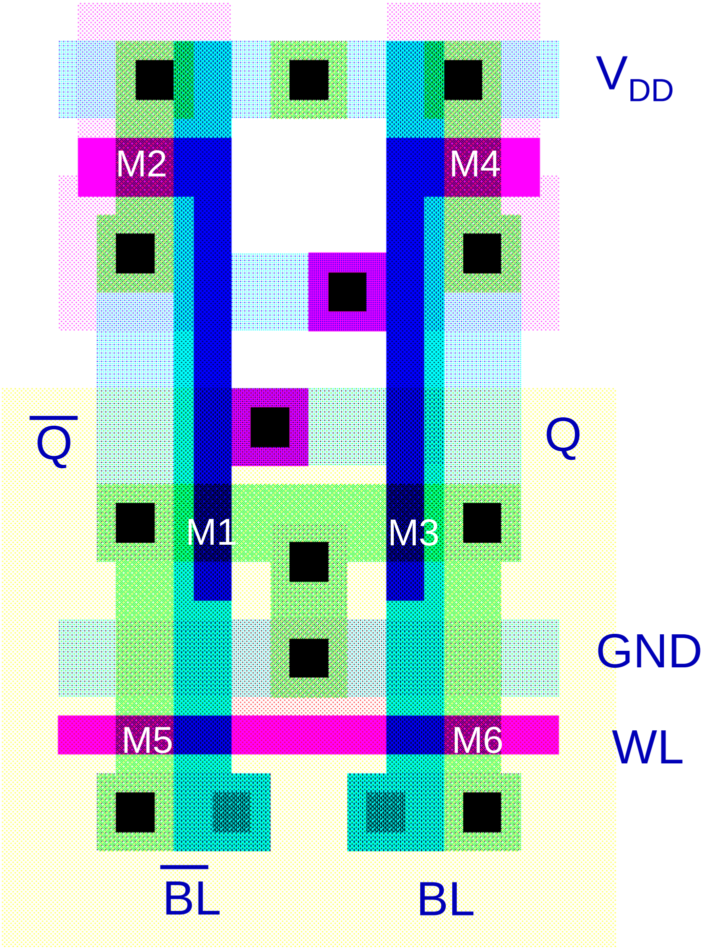
Performance of SRAM

- 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 $\overline{\text{BL}}$ can be large
- Sense amplifiers used to accelerate Read time….as the difference between BL and $\overline{\text{BL}}$ 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