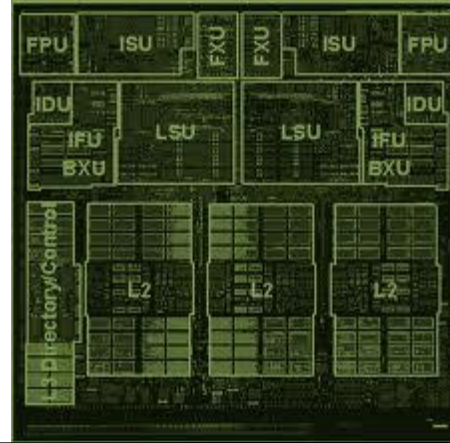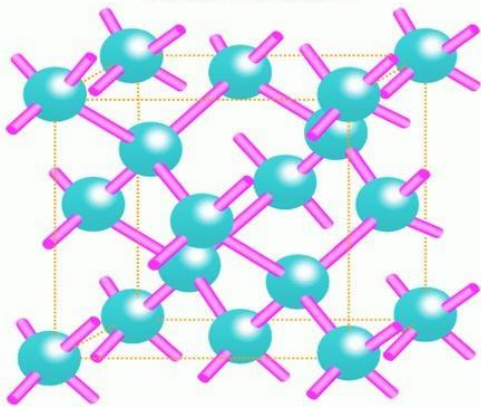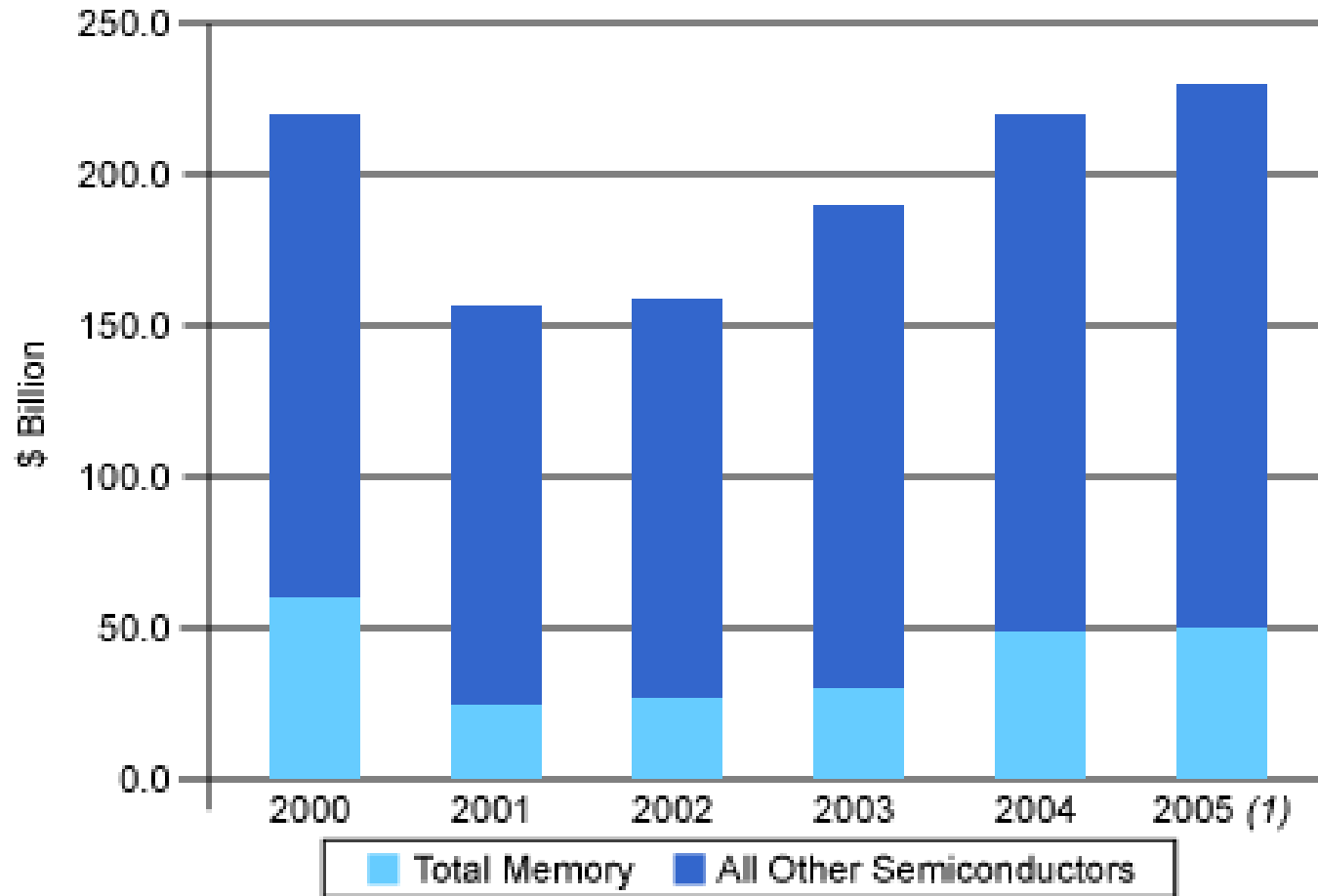# ECE 122A
VLSI Principles

Lectures 16

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

Kaustav Banerjee

Semiconductor Memories….

Kaustav Banerjee

Memory Design...

❑ **Increasing number of transistors in uprocessors are devoted to cache memories….more than 60%, see ITRS 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**

Kaustav Banerjee

Memory Classification

❑ Size

❑ Timing Parameters

❑ Function

❑ Access Pattern

Kaustav Banerjee

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

Kaustav Banerjee

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

Kaustav Banerjee

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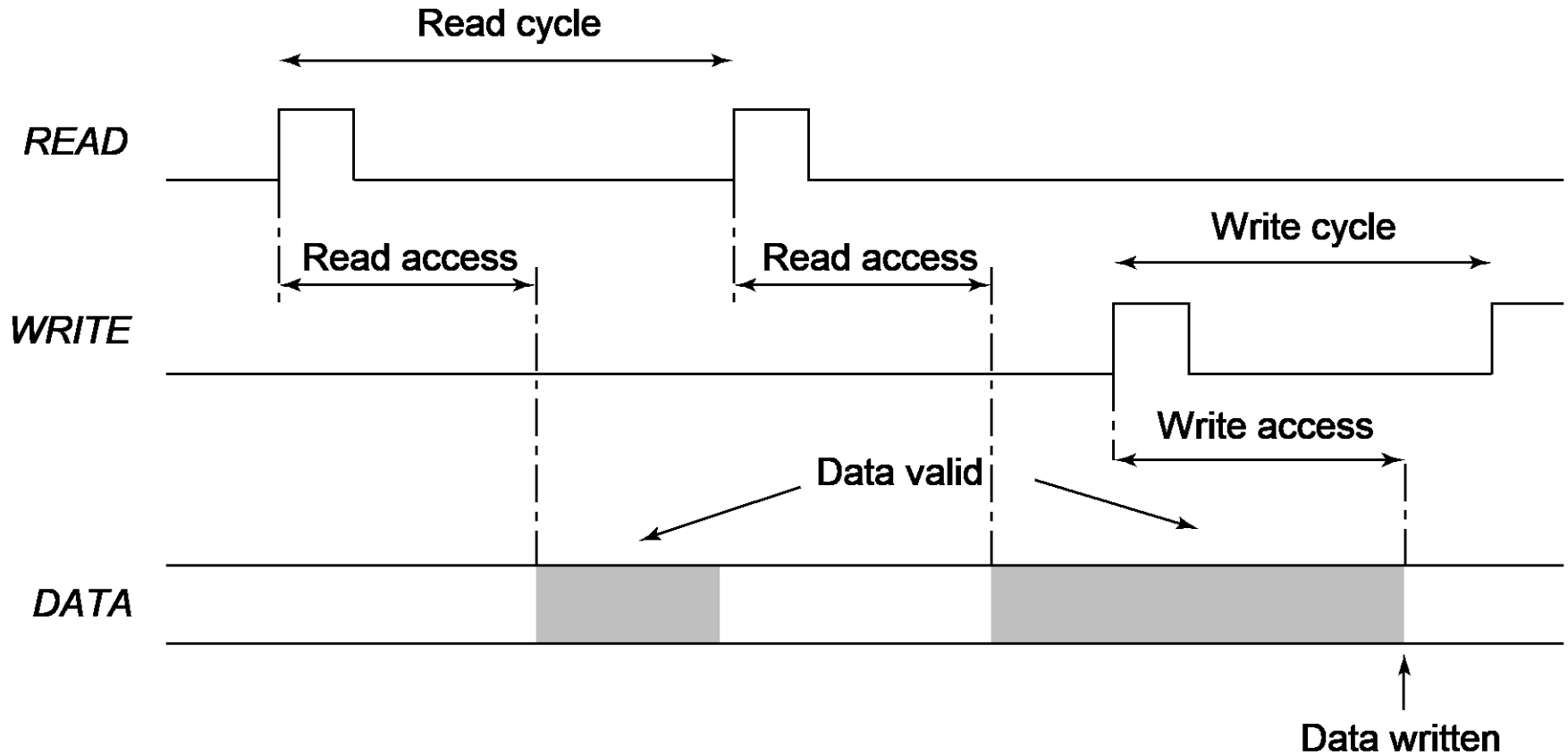
Note: Read and Write cycles do not necessarily have the same length but are considered to be equal for simplicity of system design.

Kaustav Banerjee

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.

Kaustav Banerjee

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^2PROM: Electrically Erasable and Programmable ROM

   ▪ Flash memory

Kaustav Banerjee

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

Kaustav Banerjee

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?

Kaustav Banerjee

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

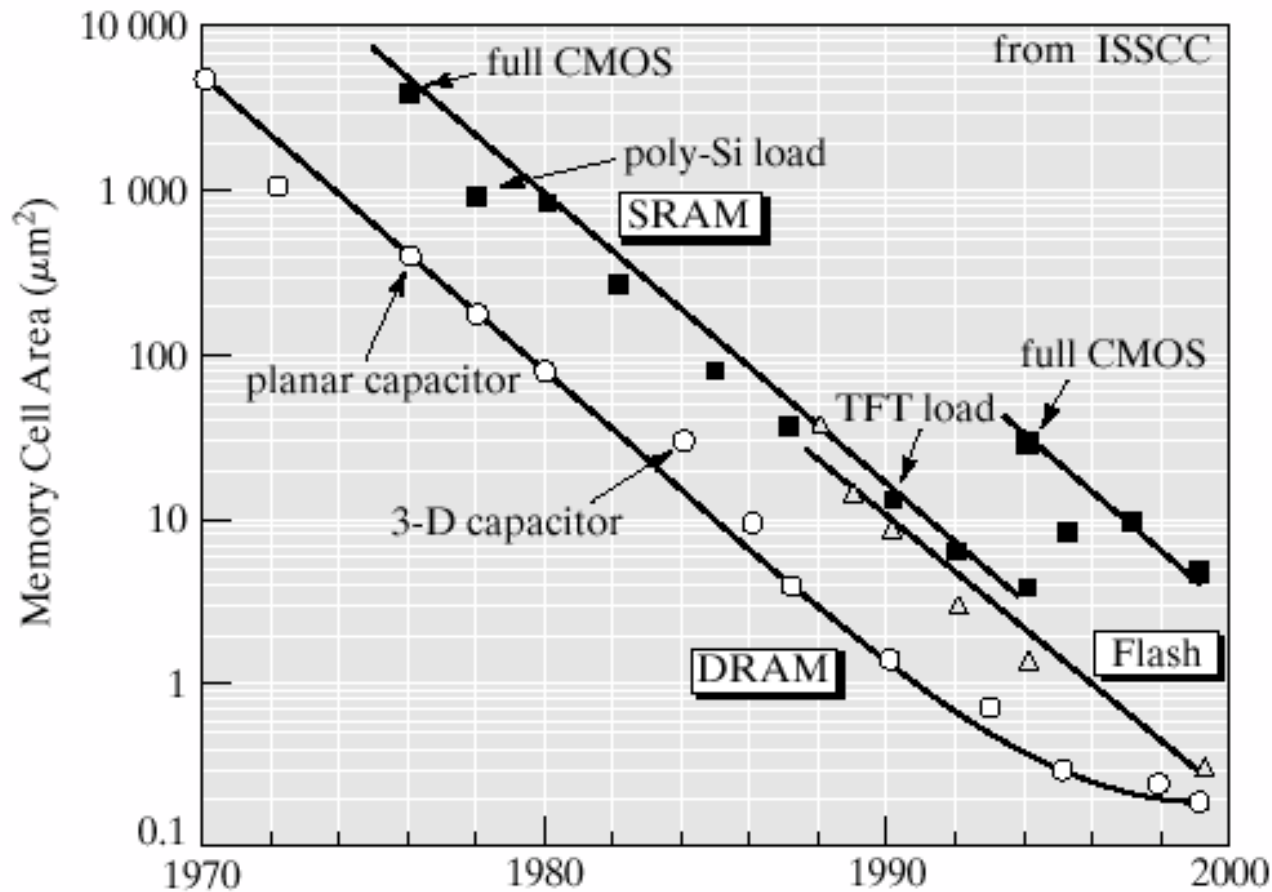Kaustav Banerjee

Semiconductor Memory Trends (up to the 90's)
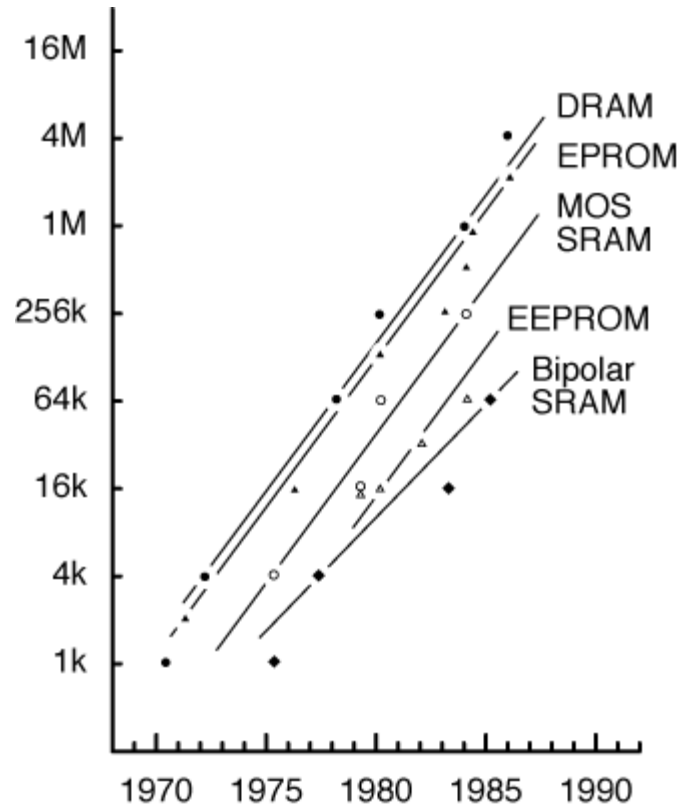
Memory Size as a function of time: x 4 every three years

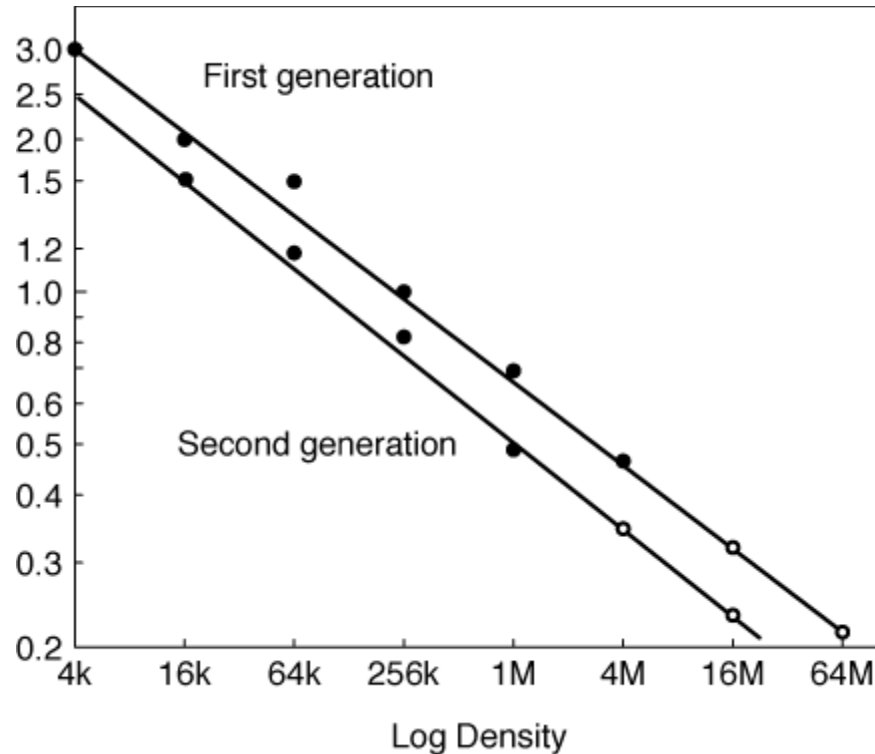Kaustav Banerjee

Semiconductor Memory Trends (more recent...)

From [Itoh01]

Kaustav Banerjee
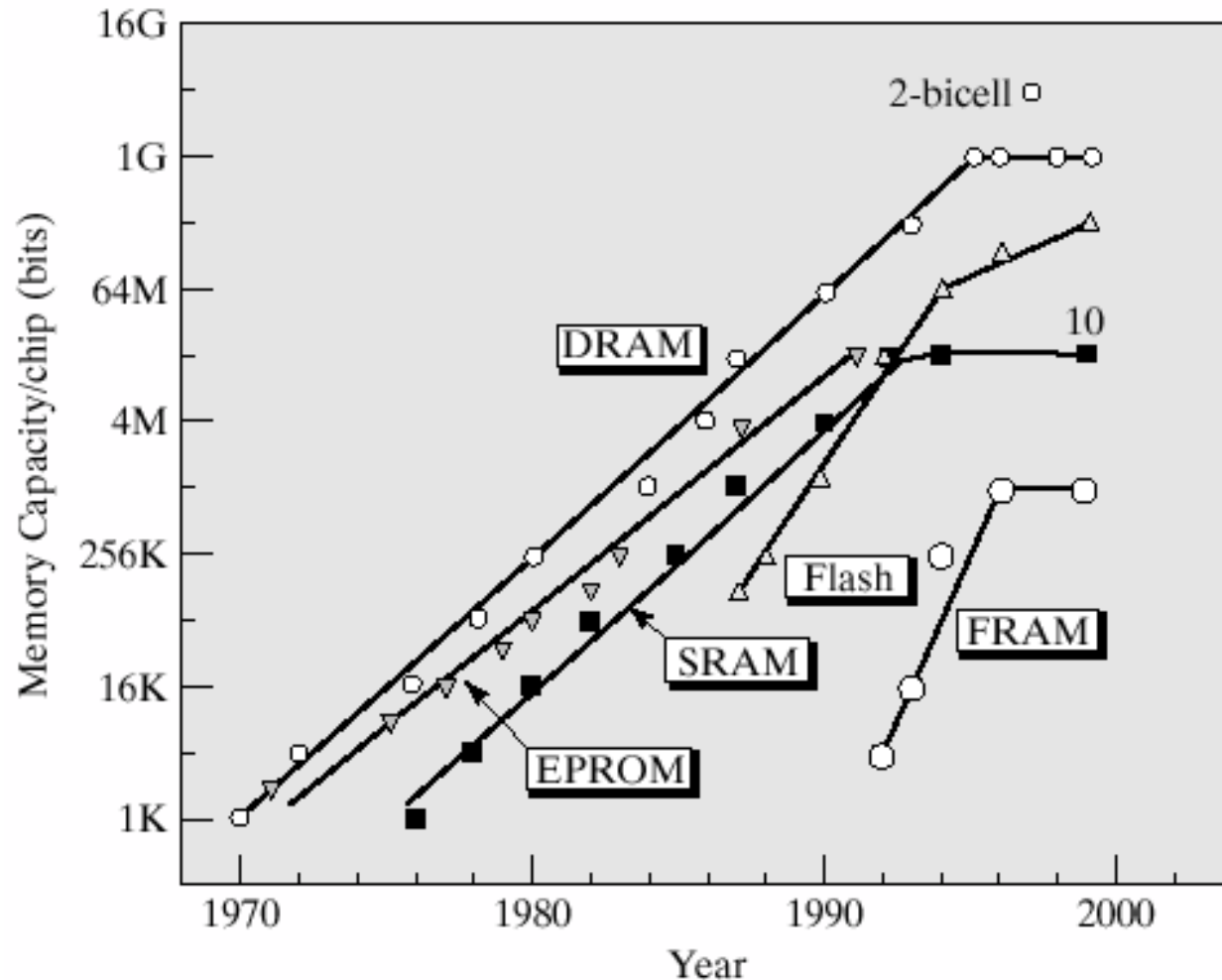
Trends in Memory Cell Area

From [Itoh01]

Kaustav Banerjee

Semiconductor Memory Trends

Technology feature size for different SRAM generations

Kaustav Banerjee

Memory Architecture: Decoders

M bits

S_0 → Word 0
S_1 → Word 1
S_2 → Word 2
→ (Storage cell)
→
S_{N-2} → Word N-2
S_{N-1} → Word N-1

N words

Input-Output
(M bits)

M bits

S_0

A_0 →
A_1 →
A_{K-1} →

DECODER

Word 0
Word 1
Word 2
(Storage cell)

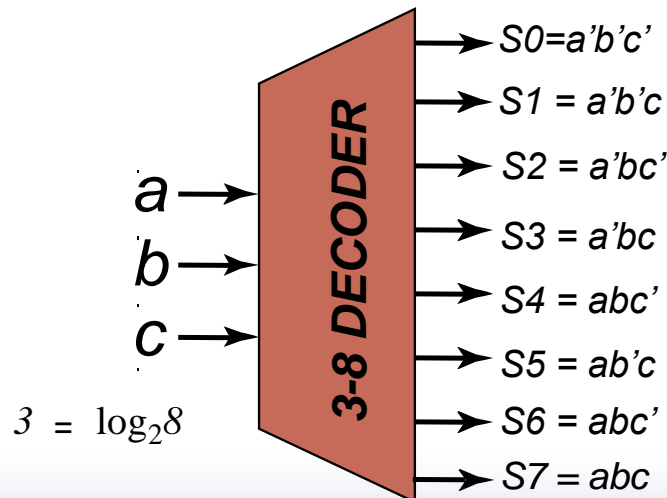Word N-2
Word N-1

$K = \log_2 N$
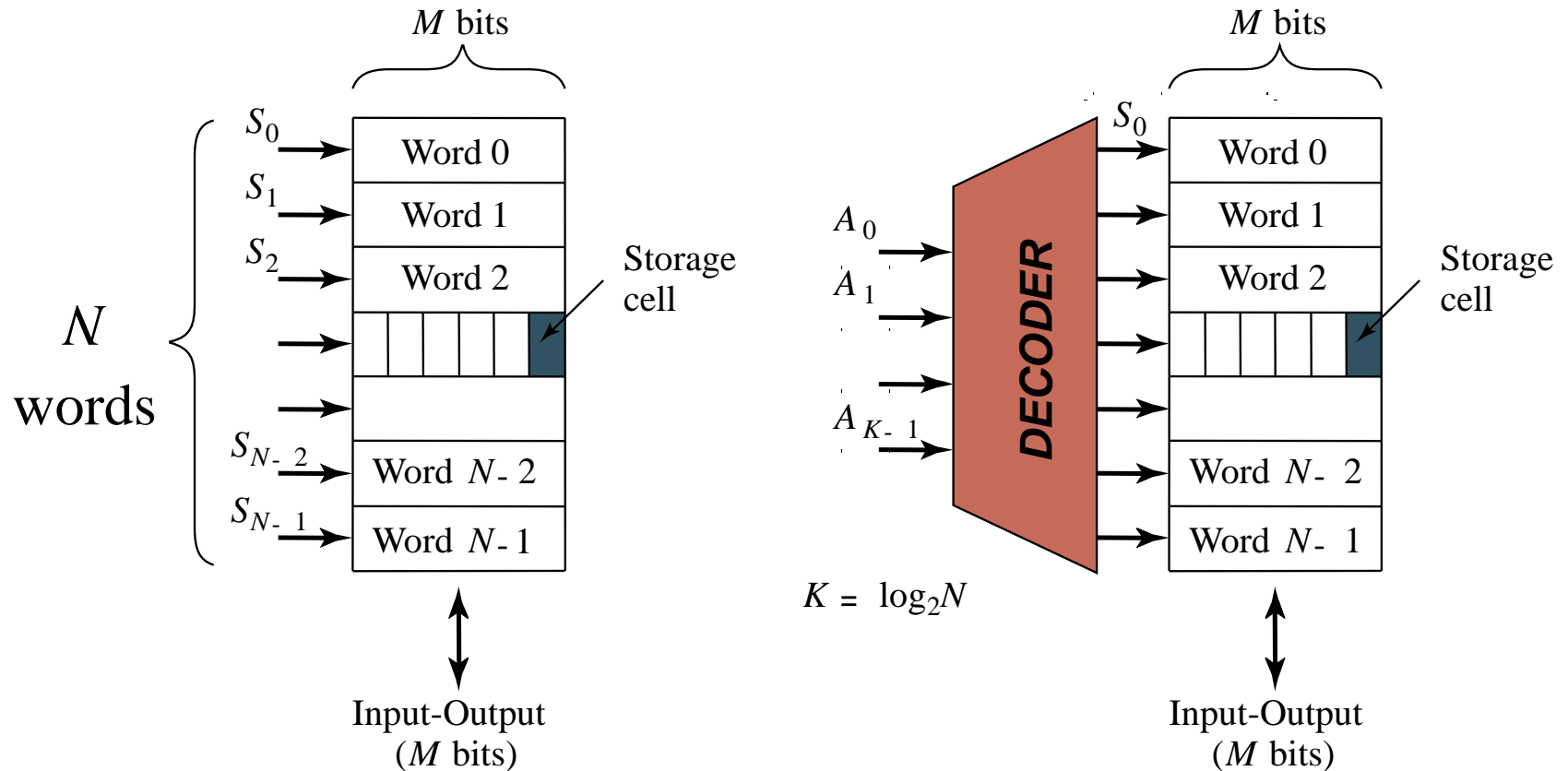
Input-Output
(M bits)

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

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

Kaustav Banerjee

Decoder Basic
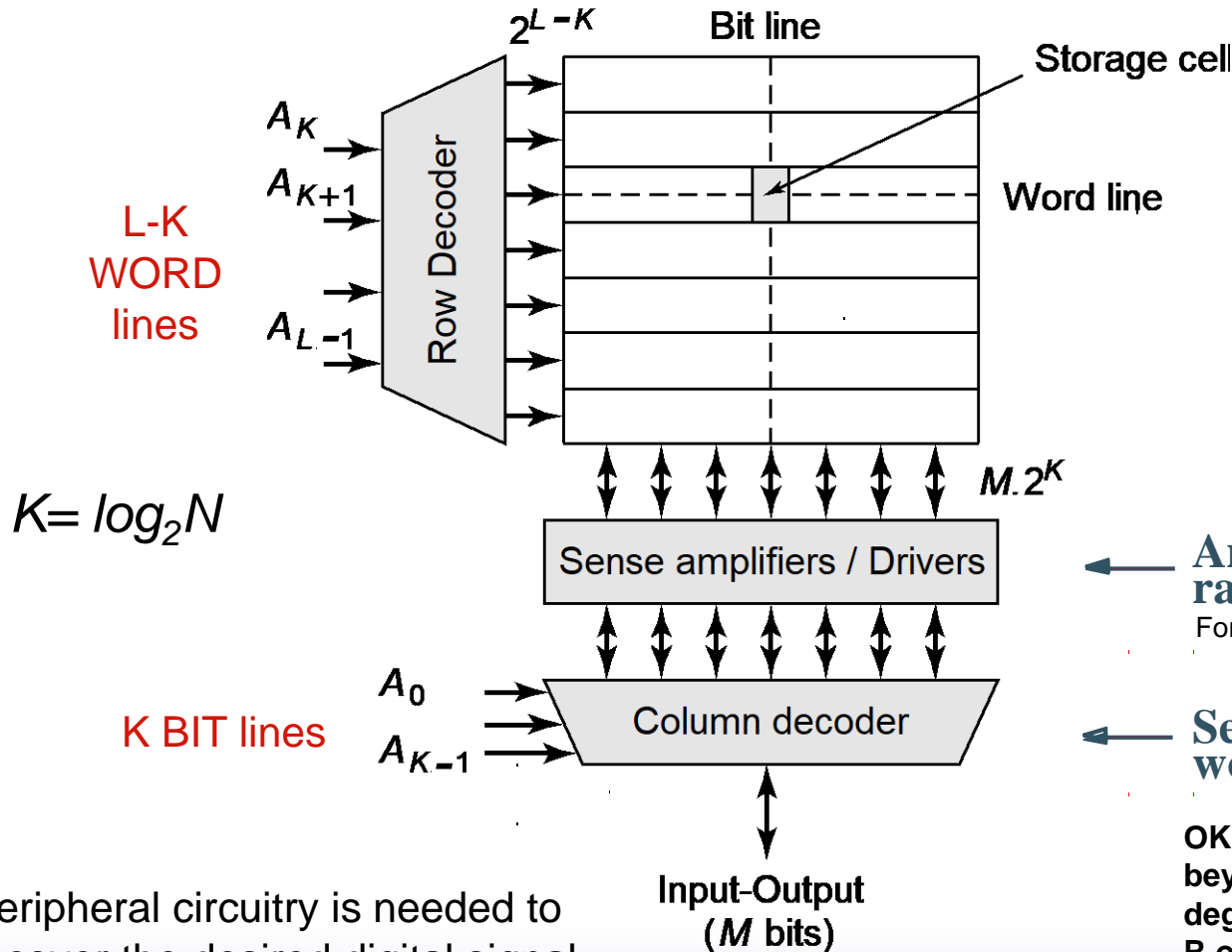
- ❏ Recall that a decoder is a combinational circuit with k inputs and at most 2^k outputs.

- ❏ 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

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

$3 = \log_2 8$

Kaustav Banerjee

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…..**

2^{L-K}

Bit line

Storage cell

A_K

A_{K+1}

Row Decoder

A_{L-1}

Word line

L-K WORD lines

$K = log_2 N$

$M.2^K$

Sense amplifiers / Drivers

A_0

A_{K-1}

K BIT lines

Column decoder

Input-Output (*M* bits)

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
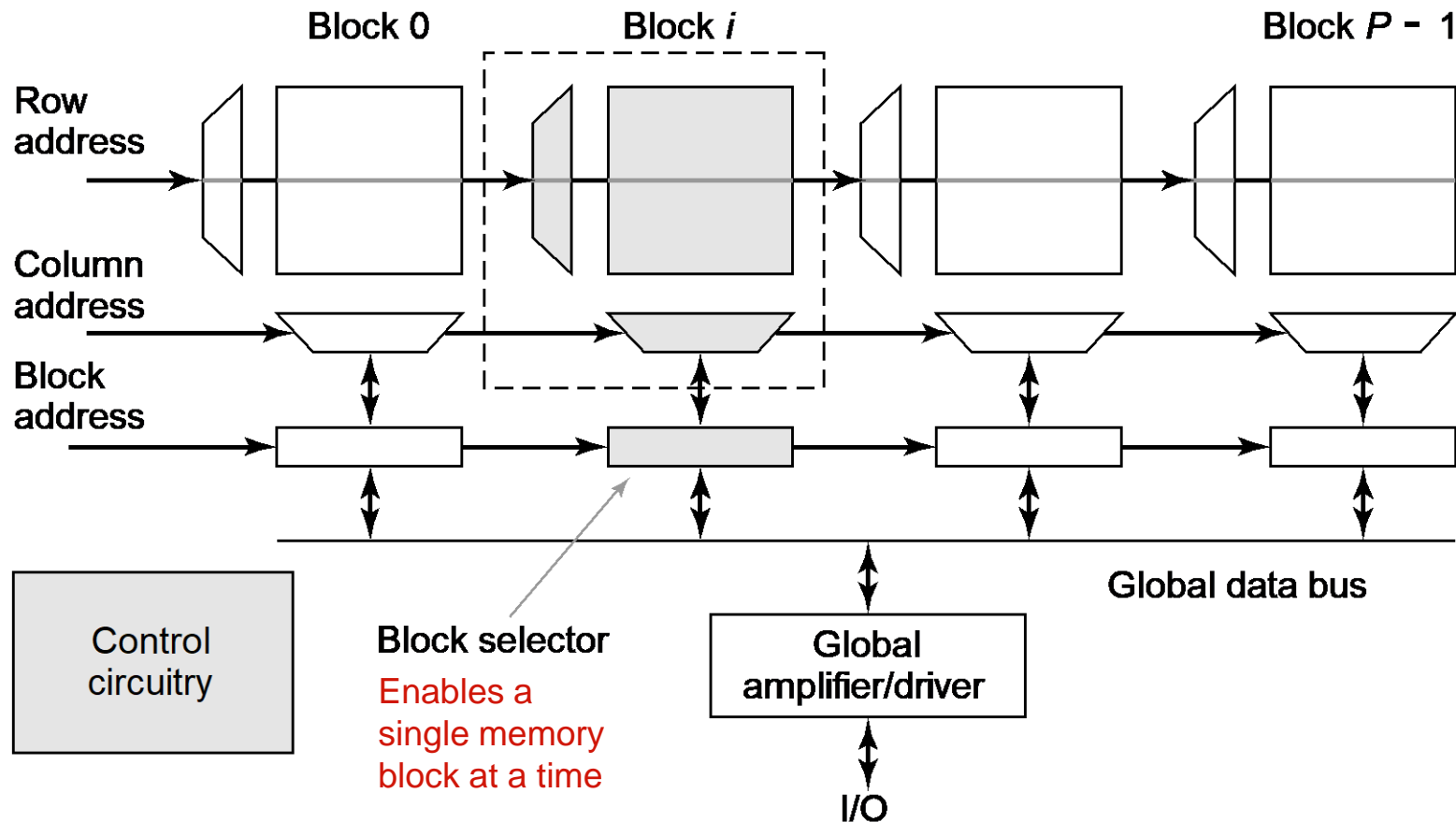For interfacing to the external world

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

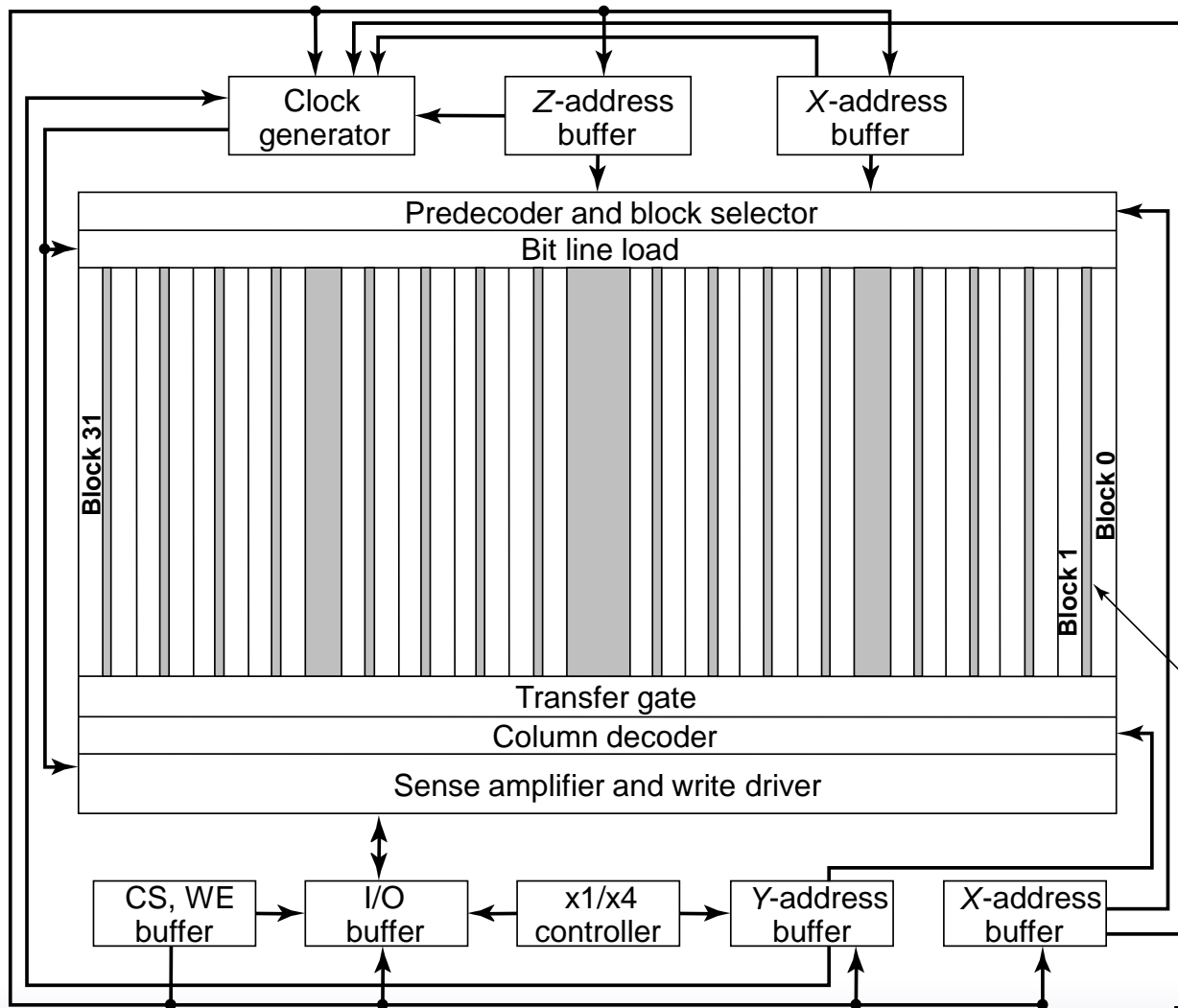Kaustav Banerjee

Hierarchical Memory Architecture

For Larger Memories….

Advantages:

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

Kaustav Banerjee

Block Diagram of 4 Mbit SRAM

| Clock generator | *Z*-address buffer | *X*-address buffer |

Predecoder and block selector

Bit line load

Block 31

Block 1

Block 0

32 blocks, each containing 128 Kbits

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

Transfer gate

Column decoder

Sense amplifier and write driver

Local Row Decoder

| CS, WE buffer | I/O buffer | x1/x4 controller | *Y*-address buffer | *X*-address buffer |

[Hirose90]

Kaustav Banerjee

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**

Kaustav Banerjee

6-transistor CMOS SRAM Cell

Should be minimum sized to achieve high memory density…..
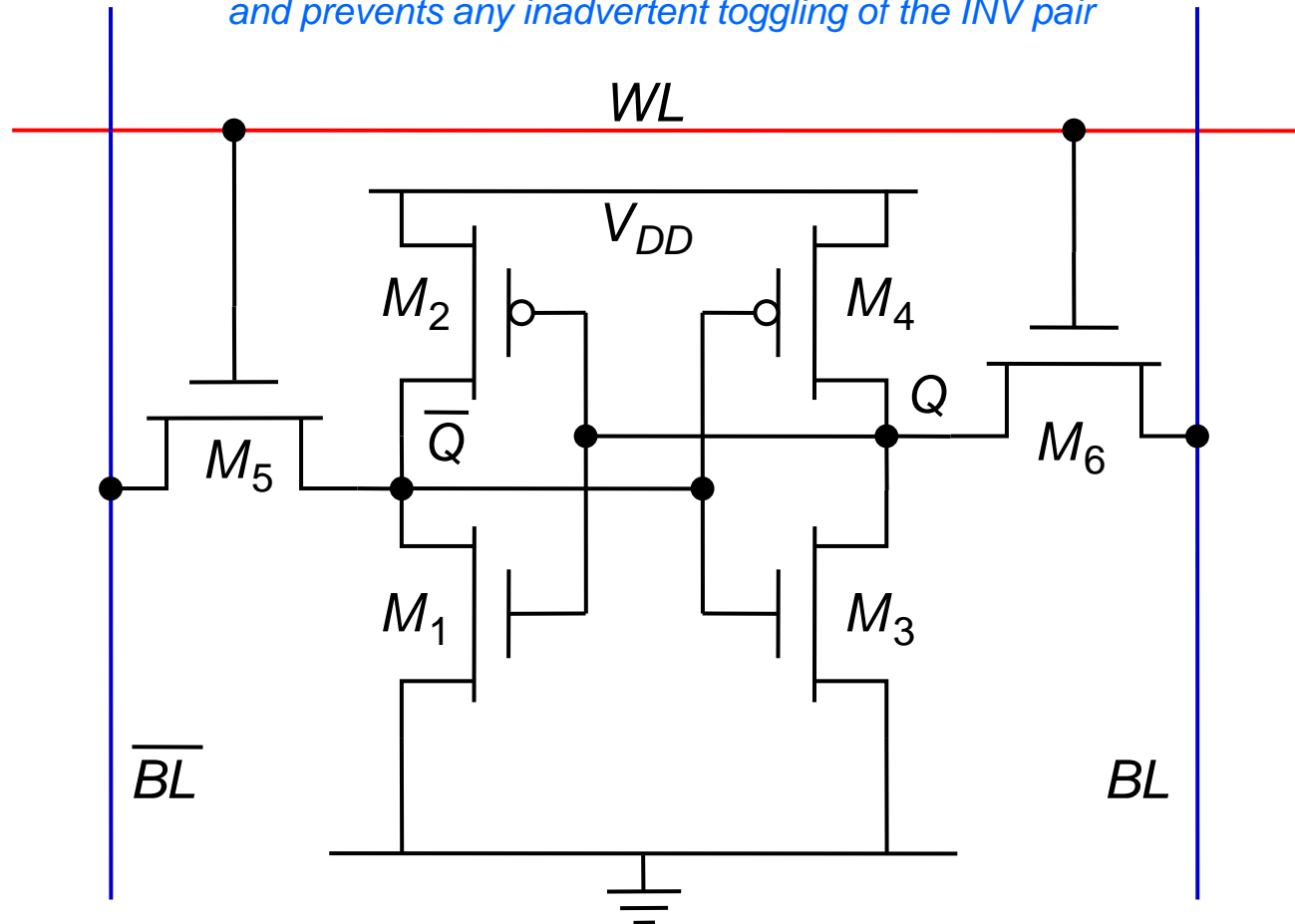
READ Operation:

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…

Kaustav Banerjee

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

$\overline{Q}$ must stay low enough so that there is no substantial current through M3-M4 INV
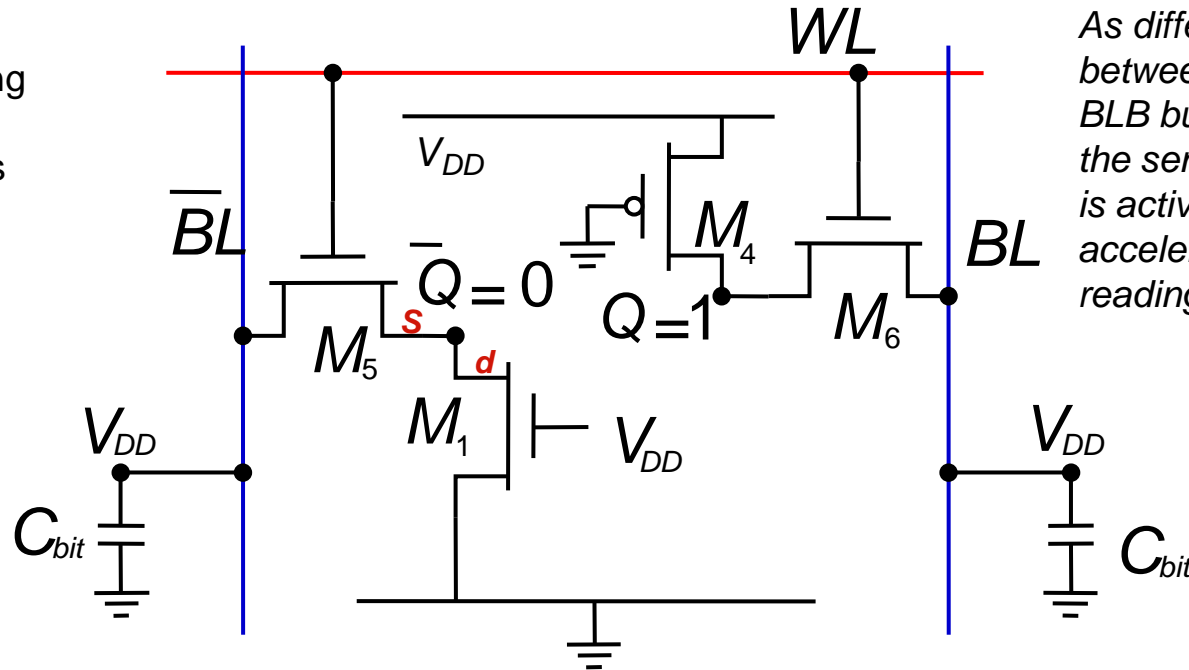
As difference between BL and BLB builds up, the sense amp. is activated to accelerate the reading process

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

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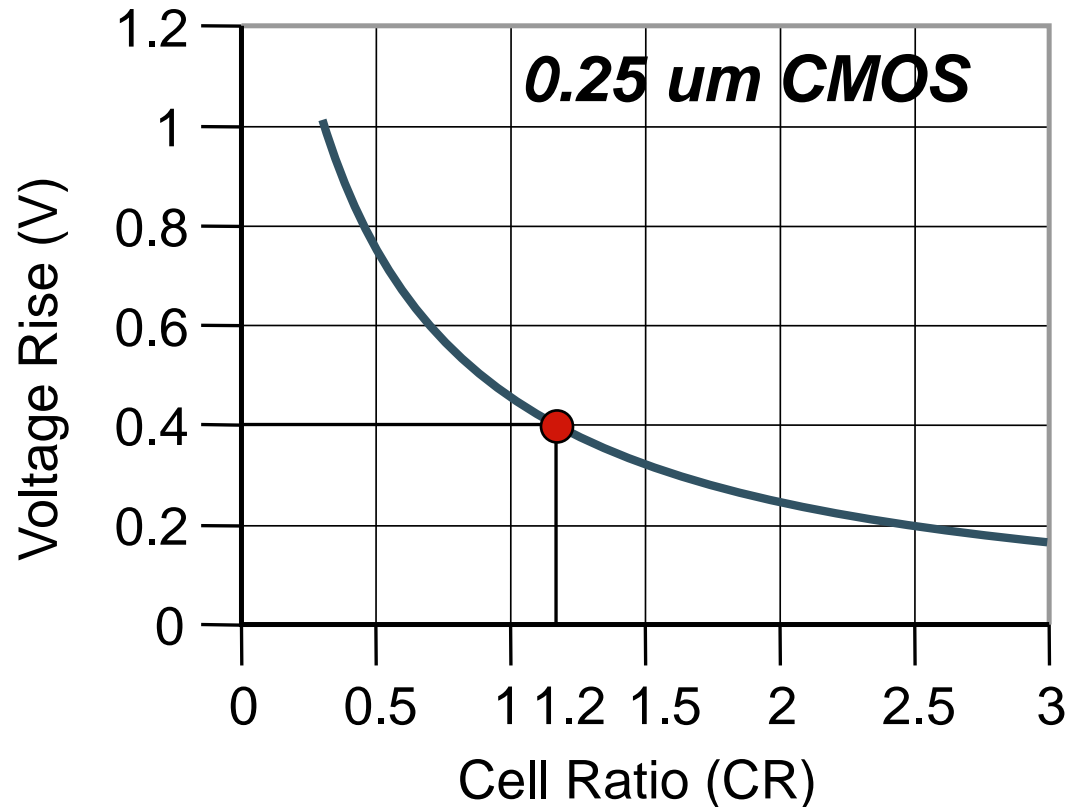
$$CR = \frac{W_1/L_1}{W_5/L_5}$$

Choose M5 to be minimum size and M1 > M5

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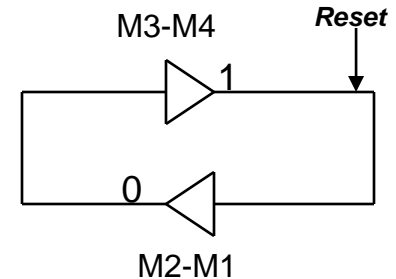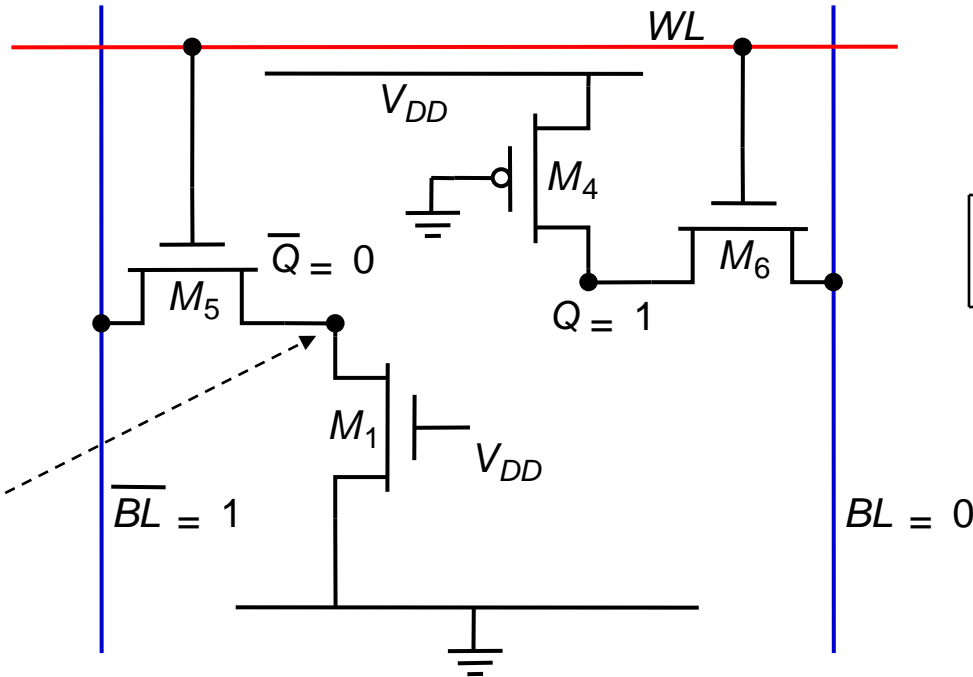
Kaustav Banerjee

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

WL

V_{DD}

M_4

M_5

$\overline{Q} = 0$

M_6

$Q = 1$

M_1

V_{DD}

$\overline{BL} = 1$

$BL = 0$

M3-M4

Reset

1

0

M2-M1

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

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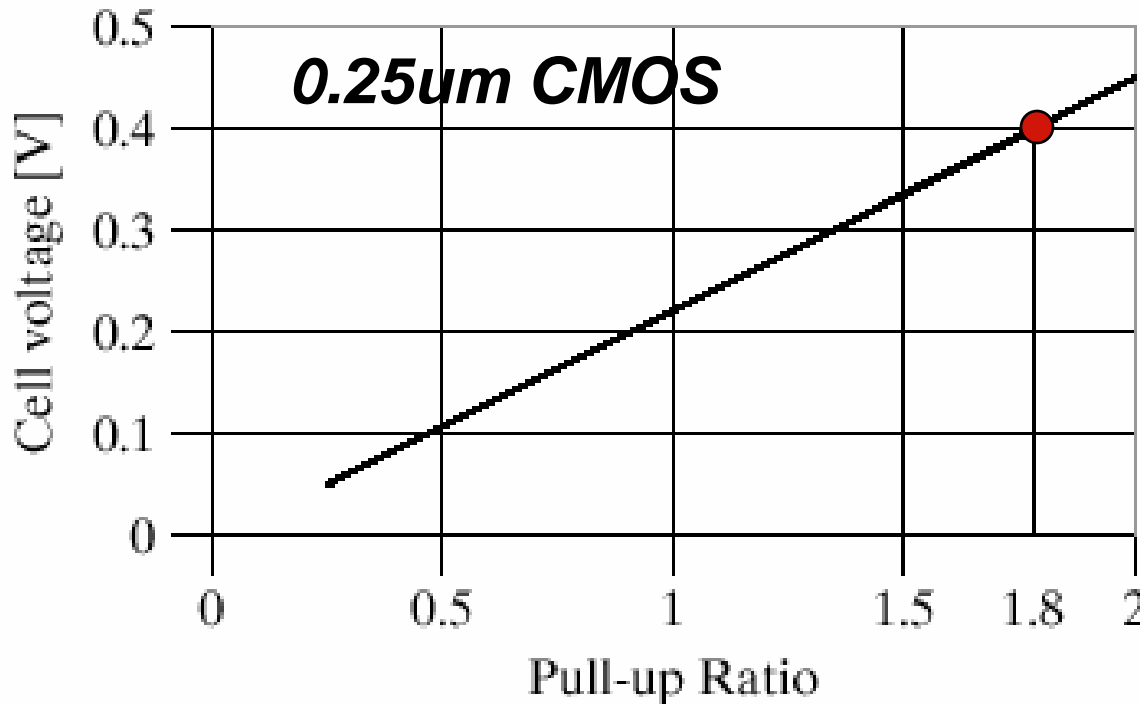
(M6 in linear)   *(M4 in saturation)*

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

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

Kaustav Banerjee

CMOS SRAM Analysis (Write)

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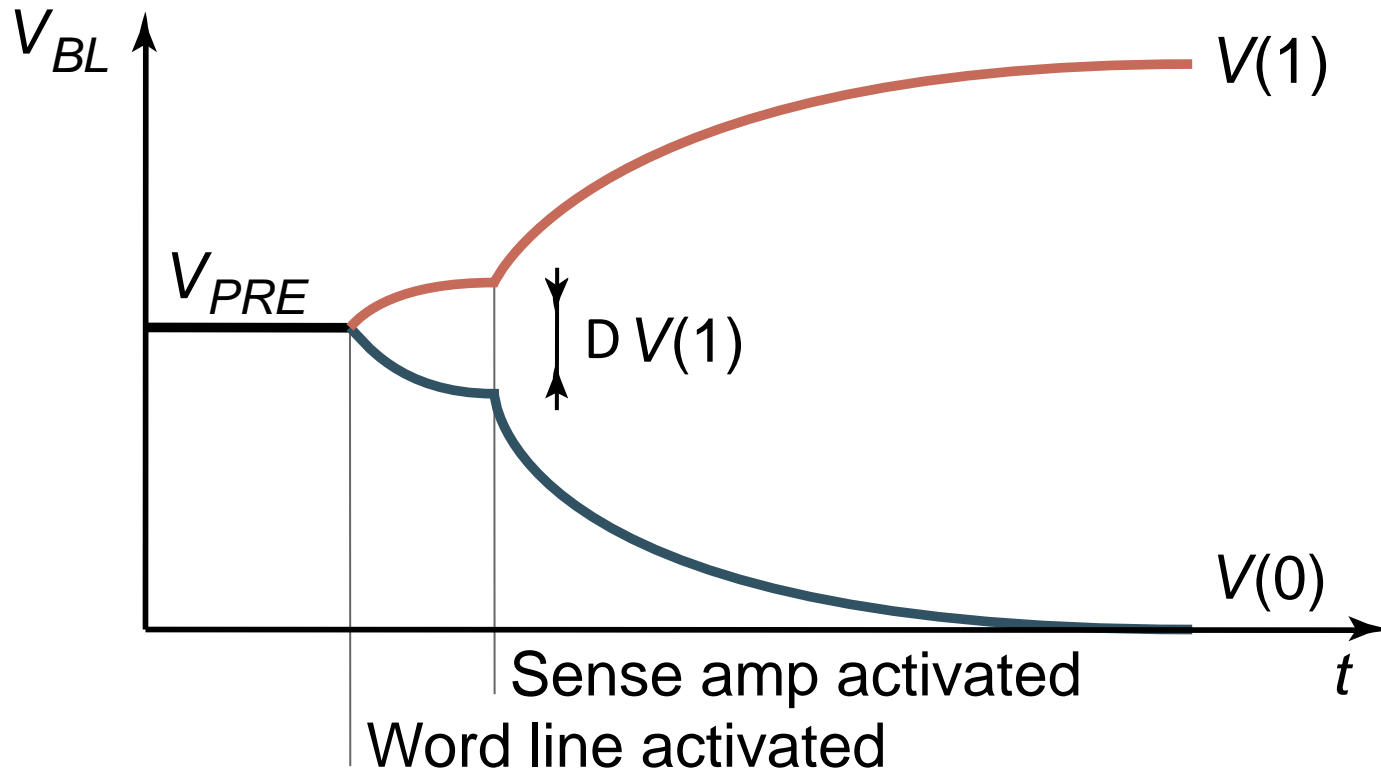

0.25um CMOS

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

Should
be low
to keep
V_Q low

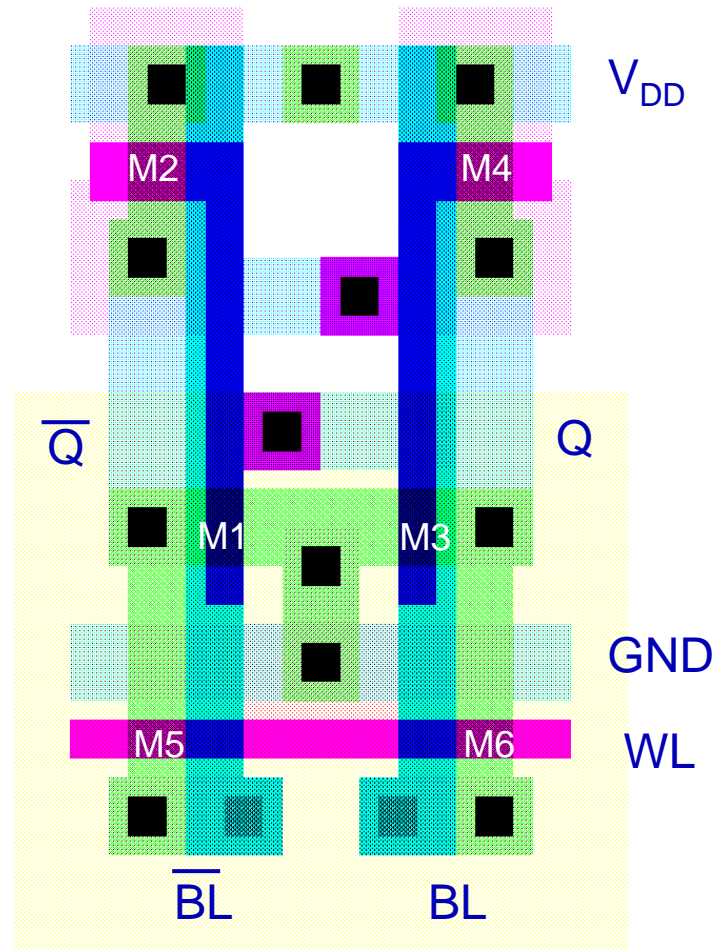PR between the PMOS (M4) pull-up and the NMOS (M6) Pass Transistor must be < 1.8 to keep Vtn < 0.4 V

Kaustav Banerjee

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
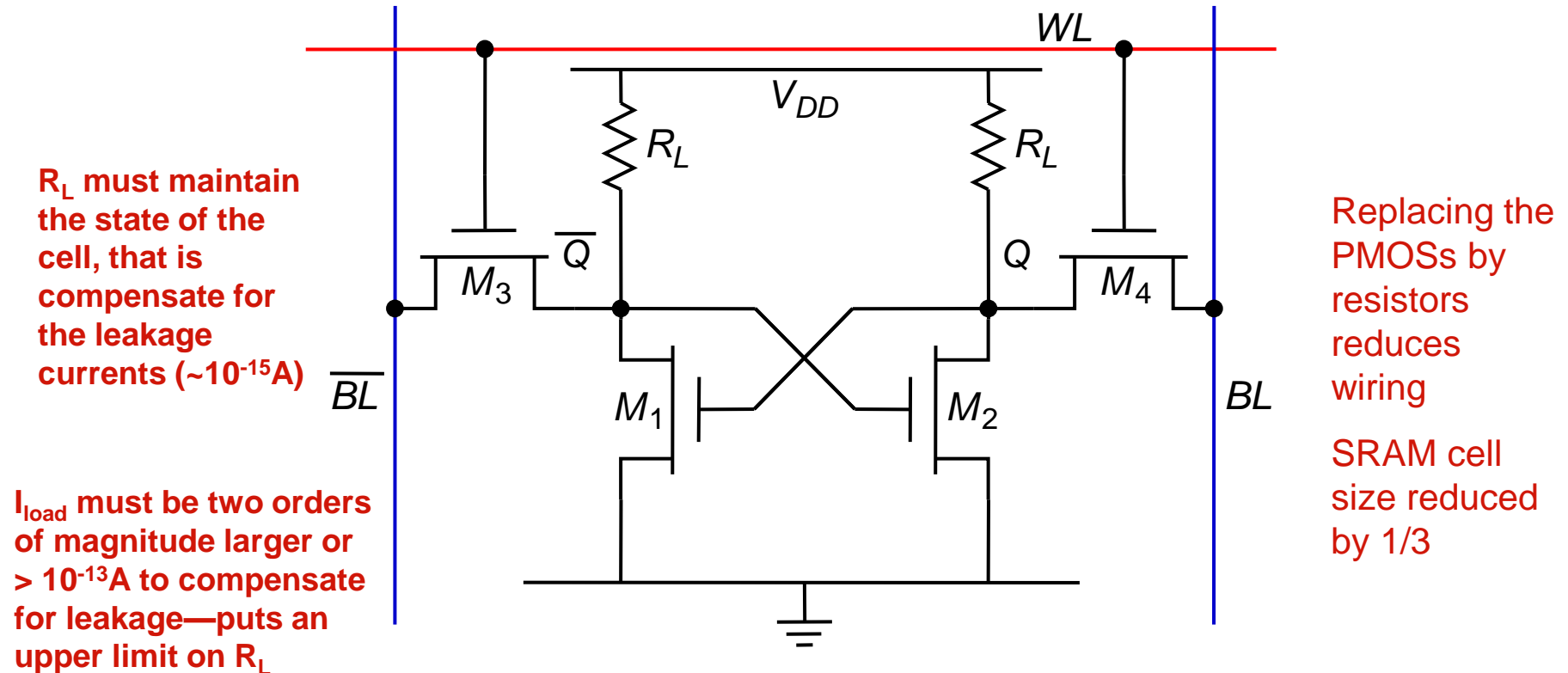
Kaustav Banerjee

Sense Amp Operation

Kaustav Banerjee

6T-SRAM — Layout

V_{DD}

M2   M4

$\overline{Q}$ Q

M1   M3

GND

M5   M6   WL

$\overline{BL}$ BL

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

Kaustav Banerjee

Resistive-load (4T) SRAM Cell

Reduce area using resistive load inverters…simplifies writing

R_L must maintain the state of the cell, that is compensate for the leakage currents ($\sim 10^{-15}$A)

I_{load} must be two orders of magnitude larger or $> 10^{-13}$A to compensate for leakage—puts an upper limit on R_L

Replacing the PMOSs by resistors reduces wiring

SRAM cell size reduced by 1/3

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

Kaustav Banerjee

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	$58.2 \ \mu m^2$ (0.7-μm rule)	$40.8 \ \mu m^2$ (0.7-μm rule)	$41.1 \ \mu m^2$ (0.8-μm rule)
Standby current (per cell)	10^{-15} A	10^{-12} A	10^{-13} A

Use high Vt

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

Kaustav Banerjee

6-T CMOS SRAM Cell: Static Noise Margin
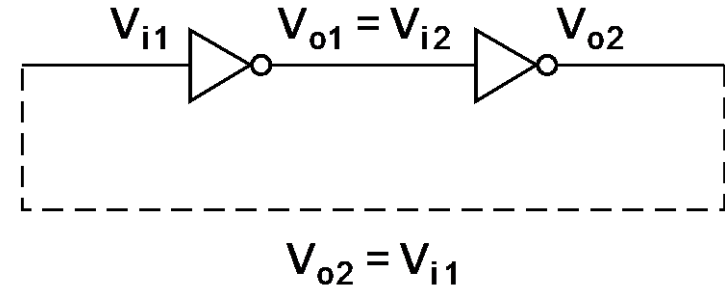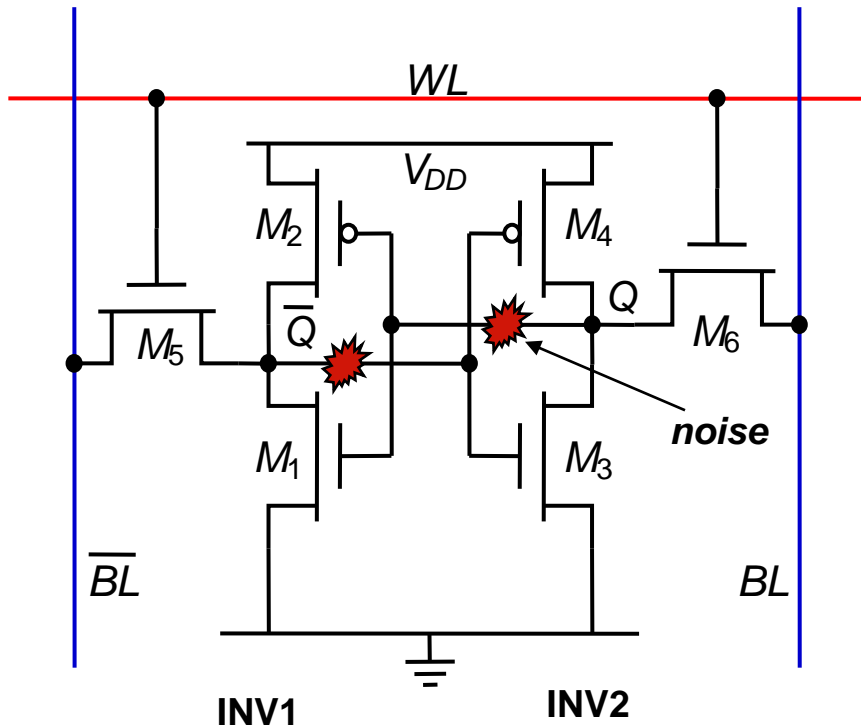

Butterfly Curve
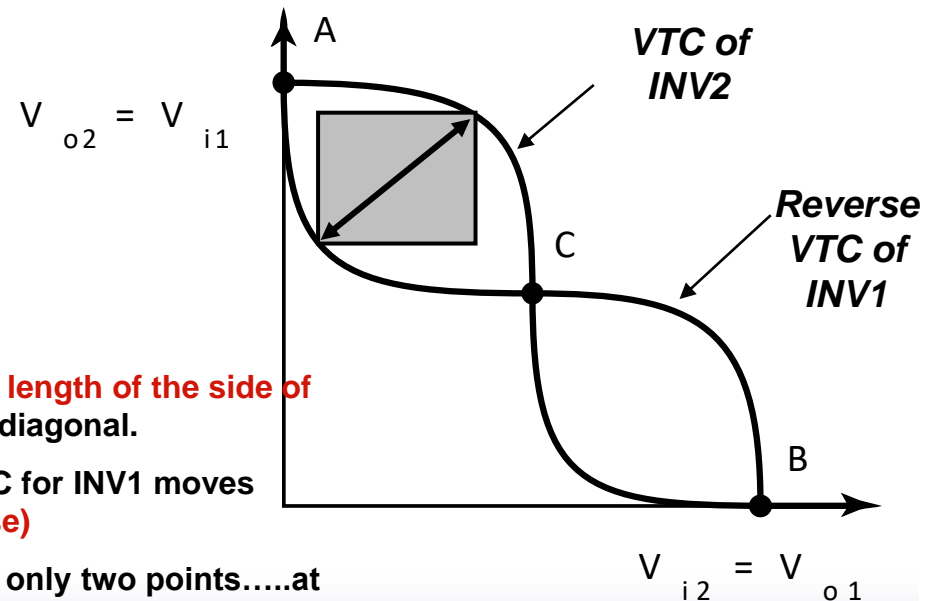
The SNM (hold margin) can be estimated graphically by the **length of the side of the square** fitted between the VTCs and having the longest diagonal.

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.

Kaustav Banerjee

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)
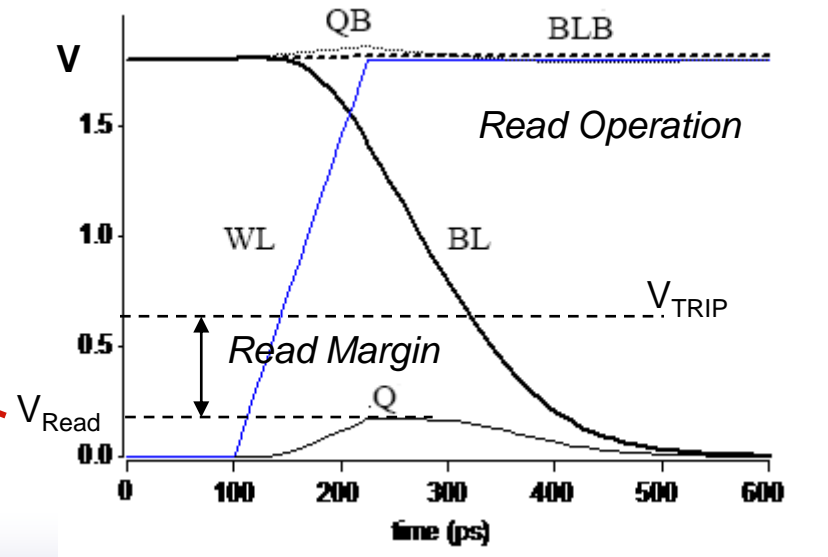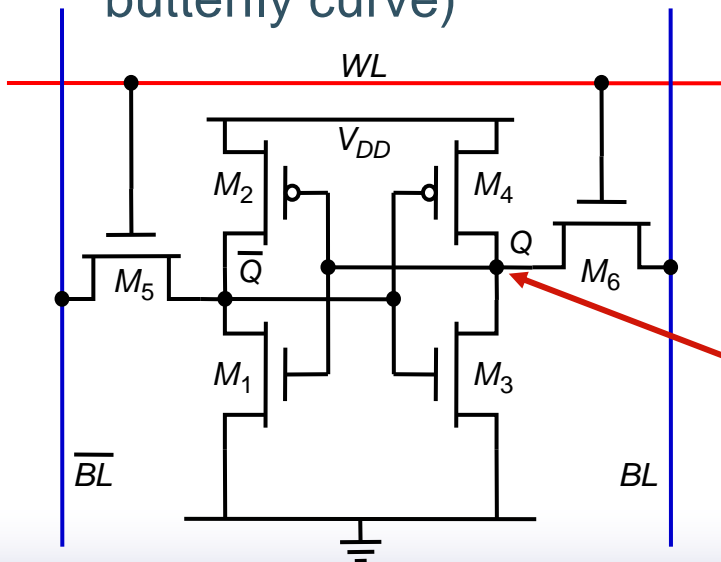
SNM Dependencies

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

❑ Dependence on sizing

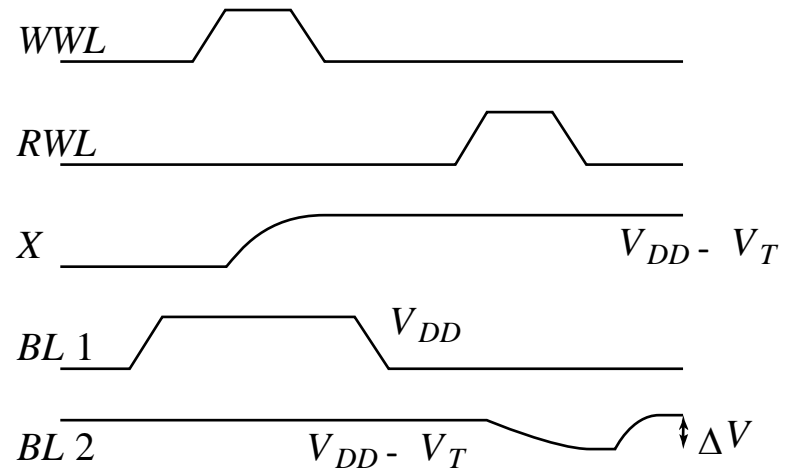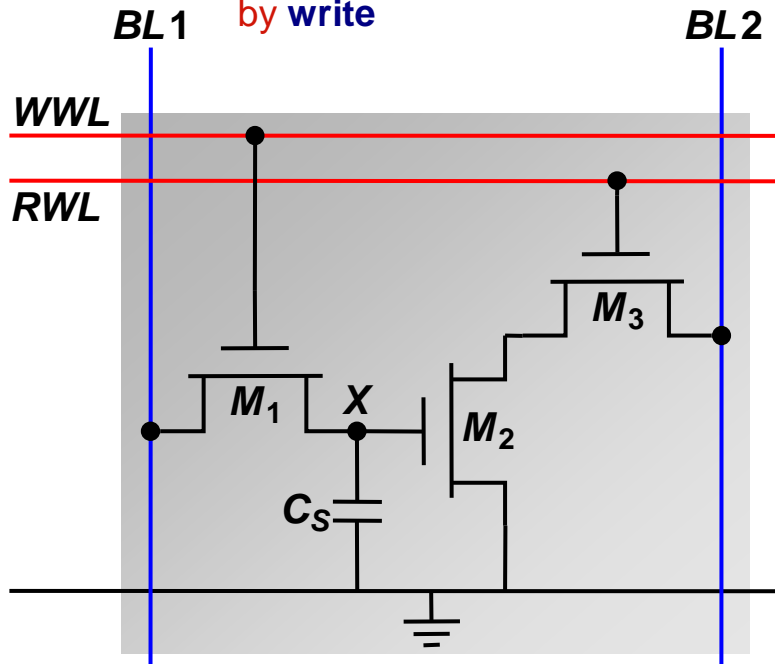*Here, Cell ratio = size of PD device over size of access device*

Upper limit of SNM = Vdd/2
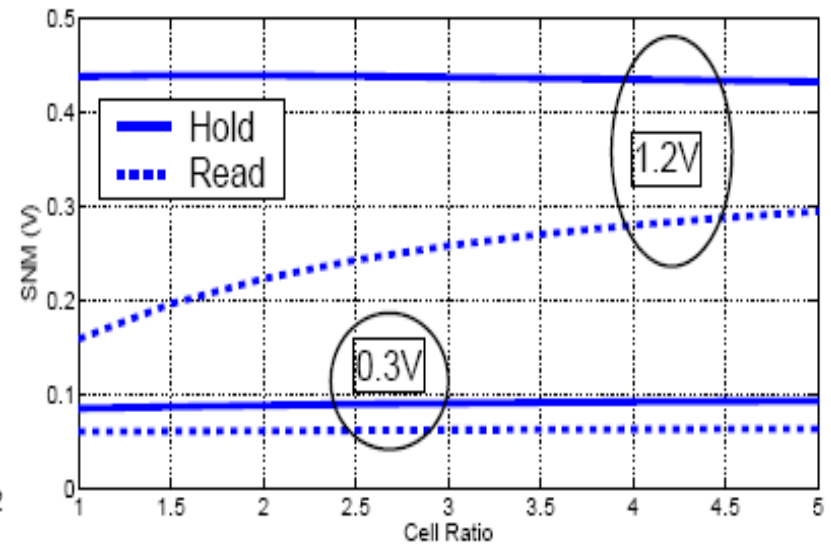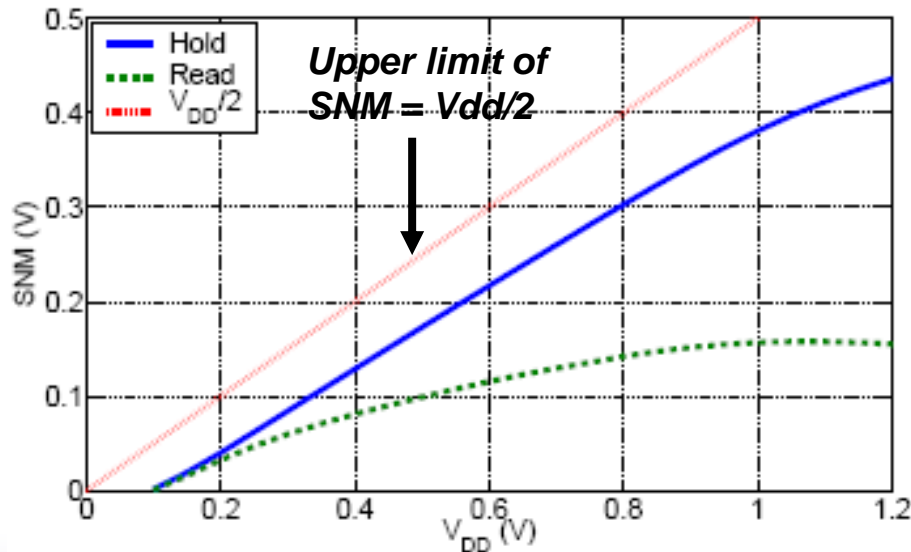
3-Transistor DRAM Cell (Early Days)

Periodic refresh operation consists of **read** followed by **write**

1 Kbit memory: Intel

Still used in some ASICs

Dynamic: since it involves charge storage on a capacitor

Cell is written by placing value on BL1 and asserting Write Word Line (WWL=1)

Data retained as charge stored on Cs once WWL=0
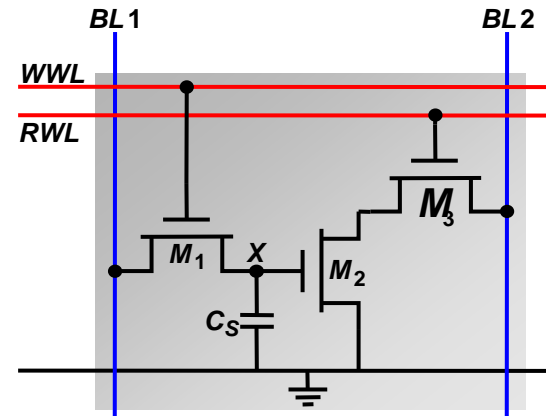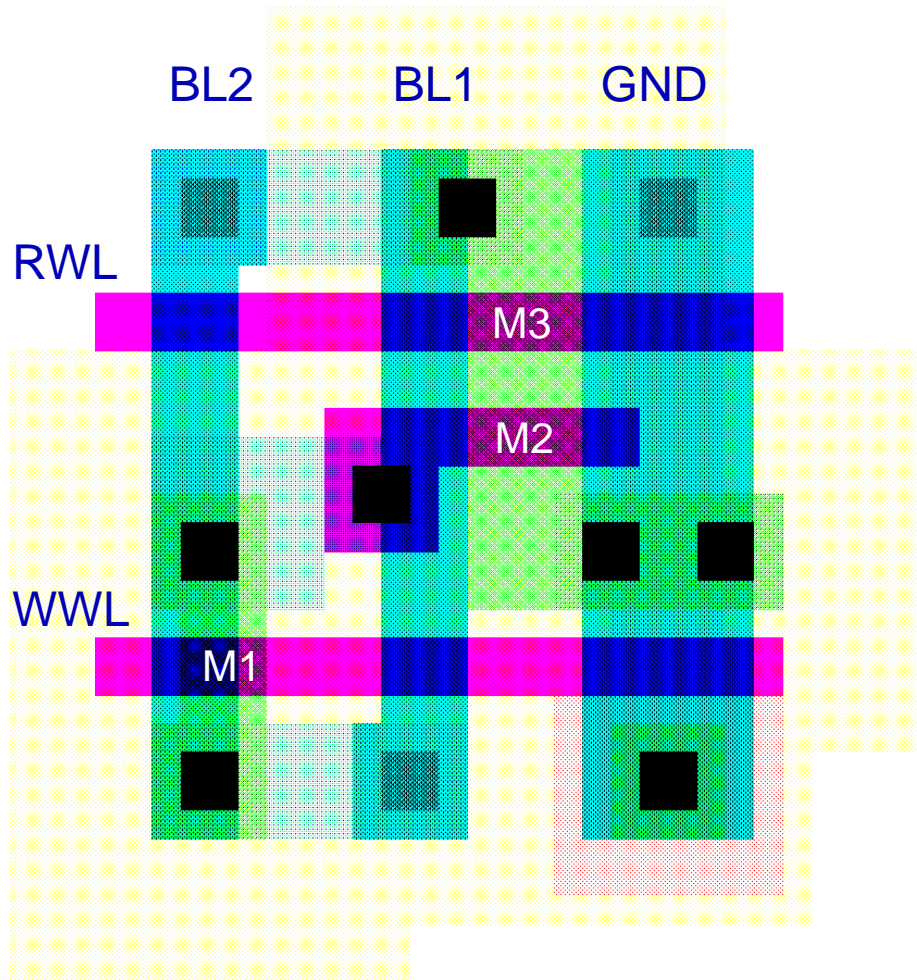
For reading the cell, RWL=1

M2 can be on or off depending on stored value

BL2 is either clamped to Vdd or is precharged to either Vdd or Vdd-Vt

M2-M3 pulls BL2 low when X=1, otherwise BL2 remains high (cell is inverting: senses the inverse value of the stored signal)

Kaustav Banerjee

3T-DRAM — Layout

BL2   BL1   GND

RWL

M3

M2

WWL

M1

Unlike SRAM, no constraint on device sizes
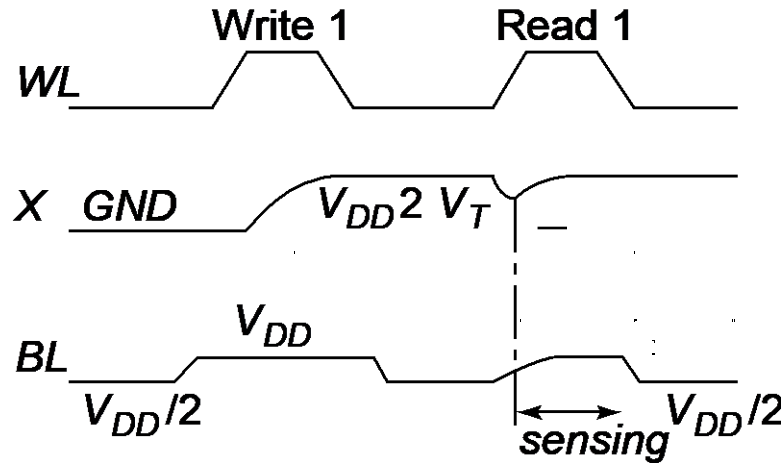
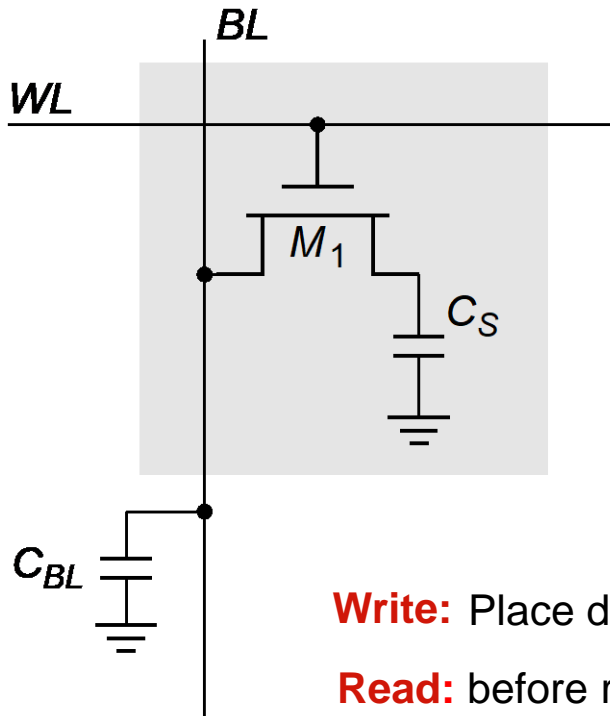**Read operation is non-destructive**

No special process steps needed

Value at node X = V_{WWL} - V_{tn}

This reduces the current through M2 during read operation and increases read access time: can use a higher value of V_{WWL} to avoid this

Kaustav Banerjee

1-Transistor DRAM Cell

Most pervasive in commercial memory design

Write 1   Read 1

WL

X GND V_{DD} $2 V_T$ —

V_{DD}

BL

$V_{DD}/2$ *sensing* $V_{DD}/2$

Write: Place data on BL and assert WL, depending on data value, Cs is 1 or 0

Read: before read, precharge BL to V_{PRE}

After WL=1, charge redistribution takes places between bit line and storage capacitance resulting in a voltage change on BL

$$\Delta V = V_{BL} - V_{PRE} = (V_{BIT} - V_{PRE}) \frac{C_S}{C_S + C_{BL}}$$

Charge transfer ratio (1-10%)

V_{BIT} is initial voltage on C_s. V_{BL} is final voltage on BL after charge redistribution. Voltage swing is small since $C_s \ll C_{BL}$; typically around 250 mV.

Note: From charge conservation, $C_S V_{BIT} + C_{BL} V_{PRE} = (C_S + C_{BL}) V_{final}$, and V_{final} (= $\mathbf{V_{BL}}$)

Kaustav Banerjee

DRAM Cell Observations

❑ 1T DRAM requires a sense amplifier for each bit line, due to charge redistribution read-out.

❑ DRAM memory cells are single ended in contrast to SRAM cells.

❑The read-out of the 1T DRAM cell is destructive; read and refresh operations are necessary for correct operation.

❑ Unlike 3T cell, 1T cell requires presence of an extra capacitance that must be explicitly included in the design.

❑ When writing a "1" into a DRAM cell, a threshold voltage is lost. This charge loss can be circumvented by bootstrapping the word lines to a higher value than V_{DD}

Kaustav Banerjee
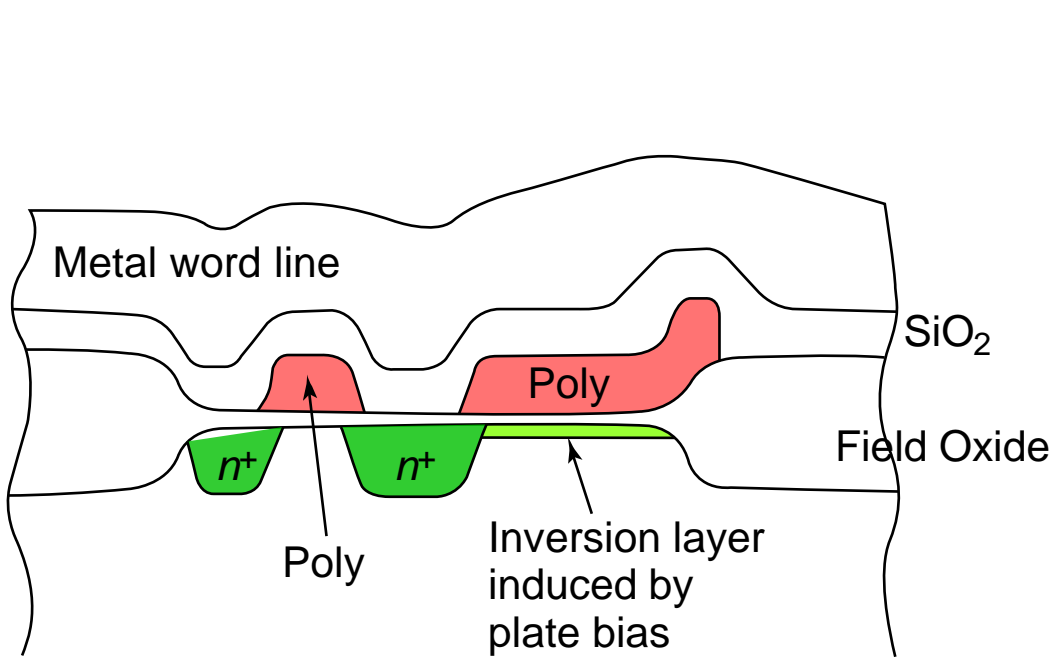
1-T DRAM Cell

Cross-section
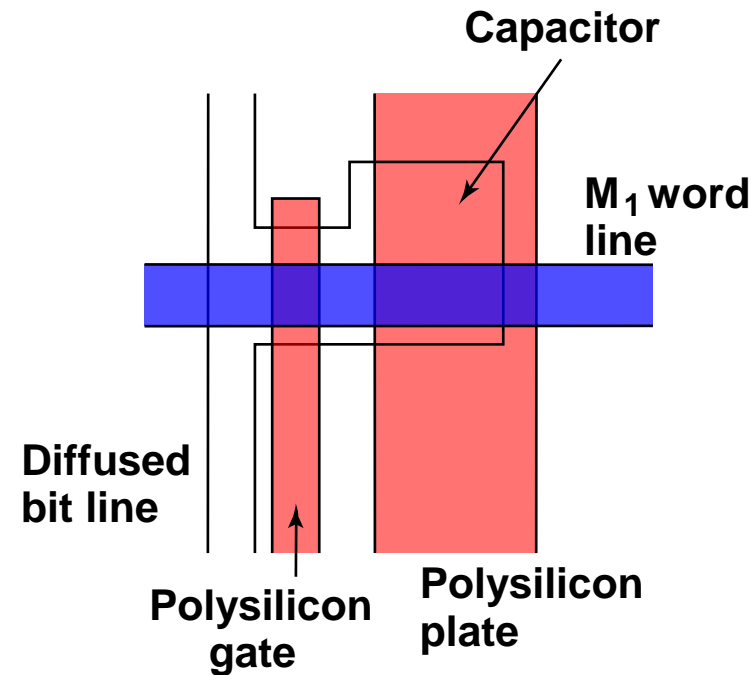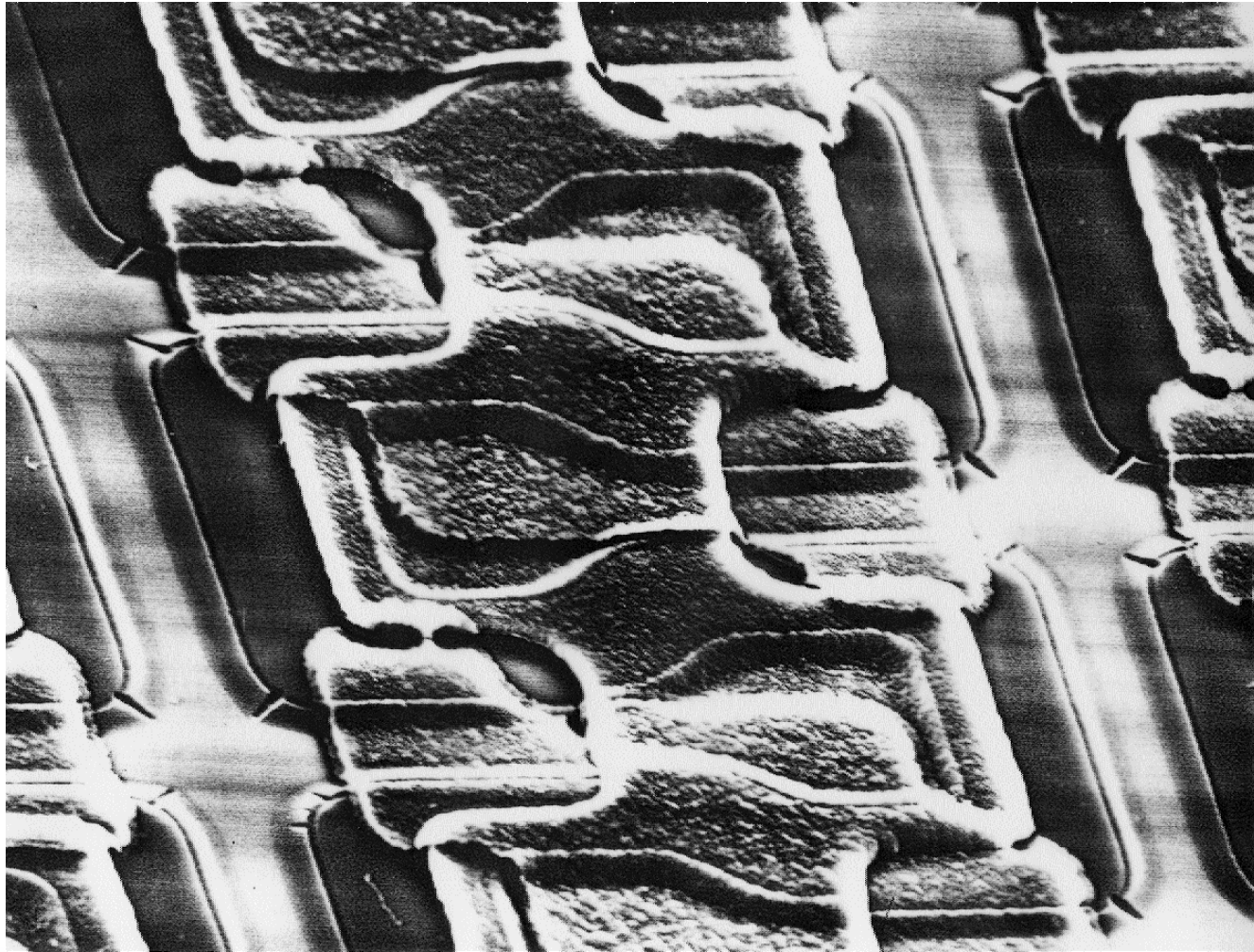
Layout

Uses Polysilicon-Diffusion Capacitance

Expensive in Area

Kaustav Banerjee

SEM of poly-diffusion capacitor 1T-DRAM

Kaustav Banerjee

Advanced 1T DRAM Cells

Cell Plate Si

Capacitor Insulator

Storage Node Poly

2nd Field Oxide

Refilling Poly

Si Substrate

Trench Cell

Word line
Insulating Layer

Cell plate

Capacitor dielectric layer

Transfer gate

Isolation

Storage electrode

17KV 30.4KX 329n #170
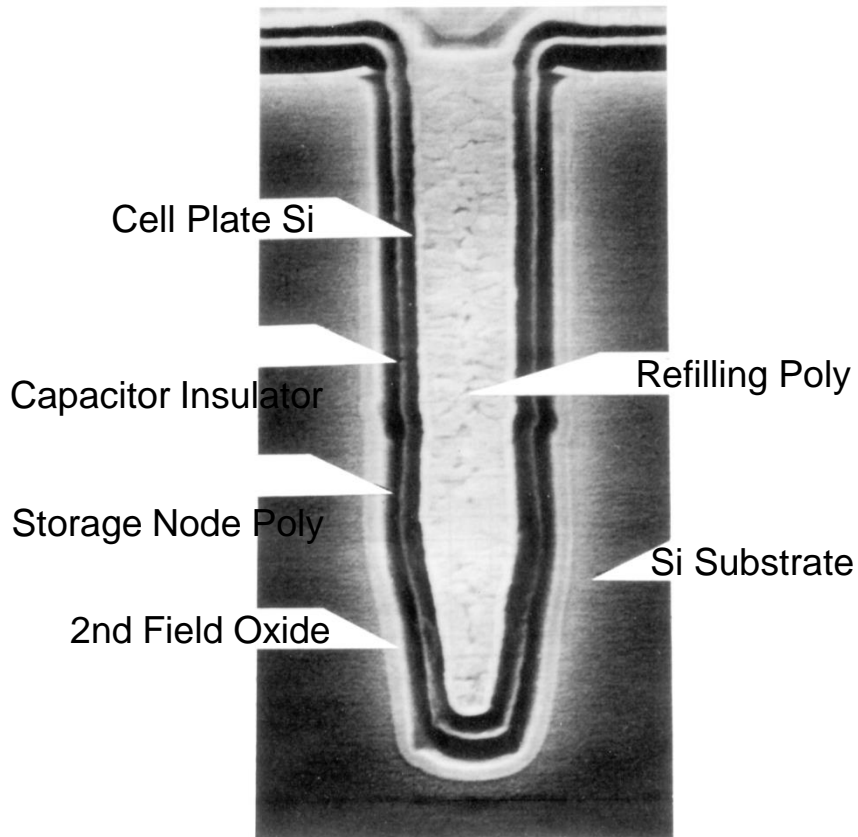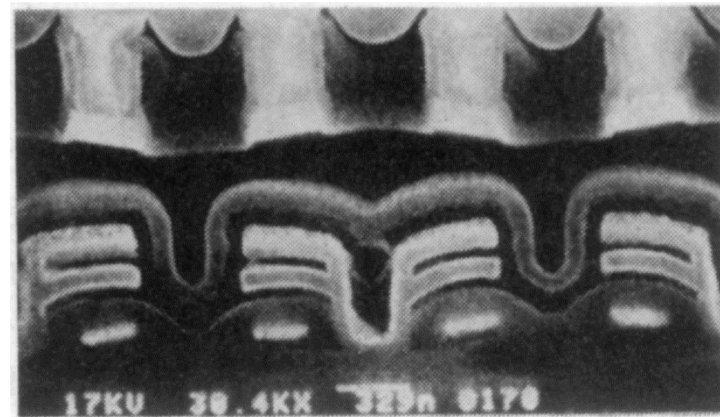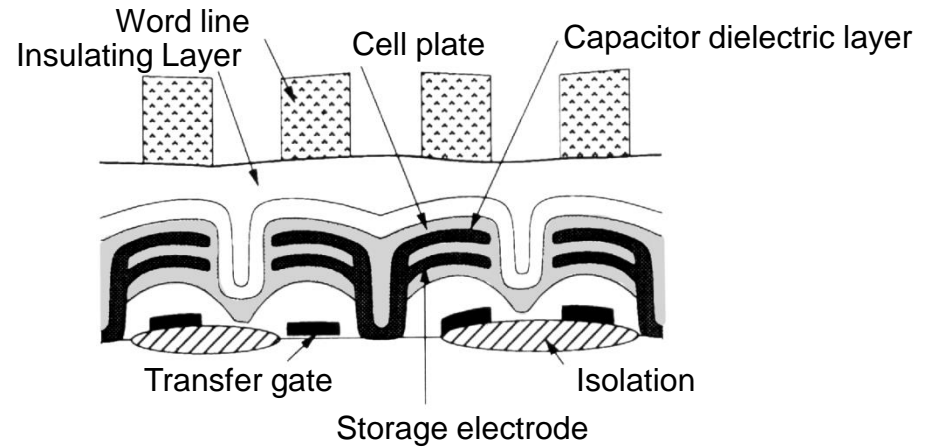
Stacked-capacitor Cell

Kaustav Banerjee

Non-Volatile Memory

Kaustav Banerjee

Flash Memory (with Floating Gate (FG) Transistor)

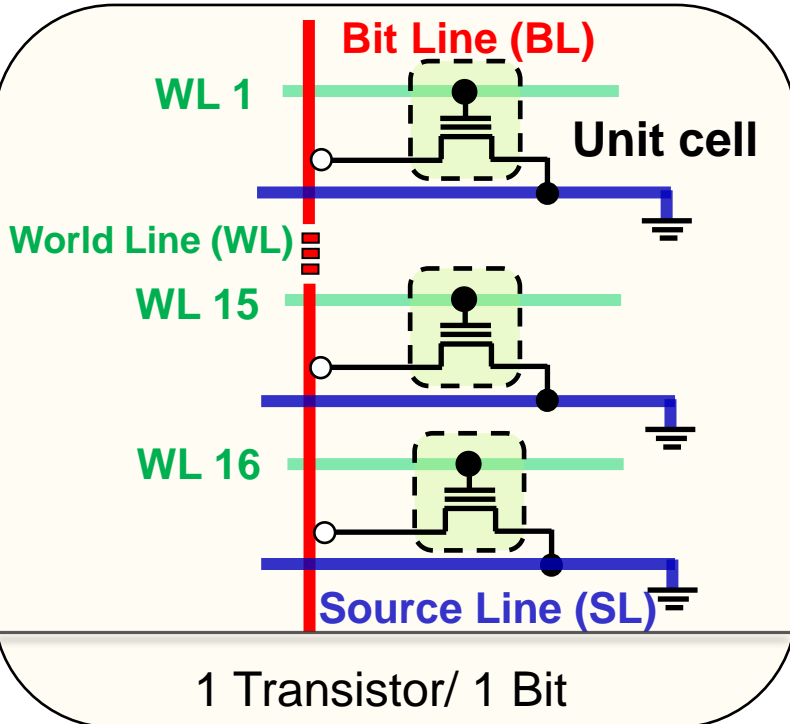|	NAND	NOR
Programming	Vg (>8 V) — CG, COX, FG, TOX, Channel, Source, Drain, Substrate — *Tunneling*	Vg (>8 V), Vd (>4 V) — CG, COX, FG, TOX, Channel, Source, Drain, Substrate — *Hot Carrier Injection*
Erasing	Vg (<-8 V) — CG, COX, FG, TOX, Channel, Source, Drain, Substrate — *Tunneling*	Vg (<-8 V) — CG, COX, FG, TOX, Channel, Source, Drain, Substrate — *Tunneling*
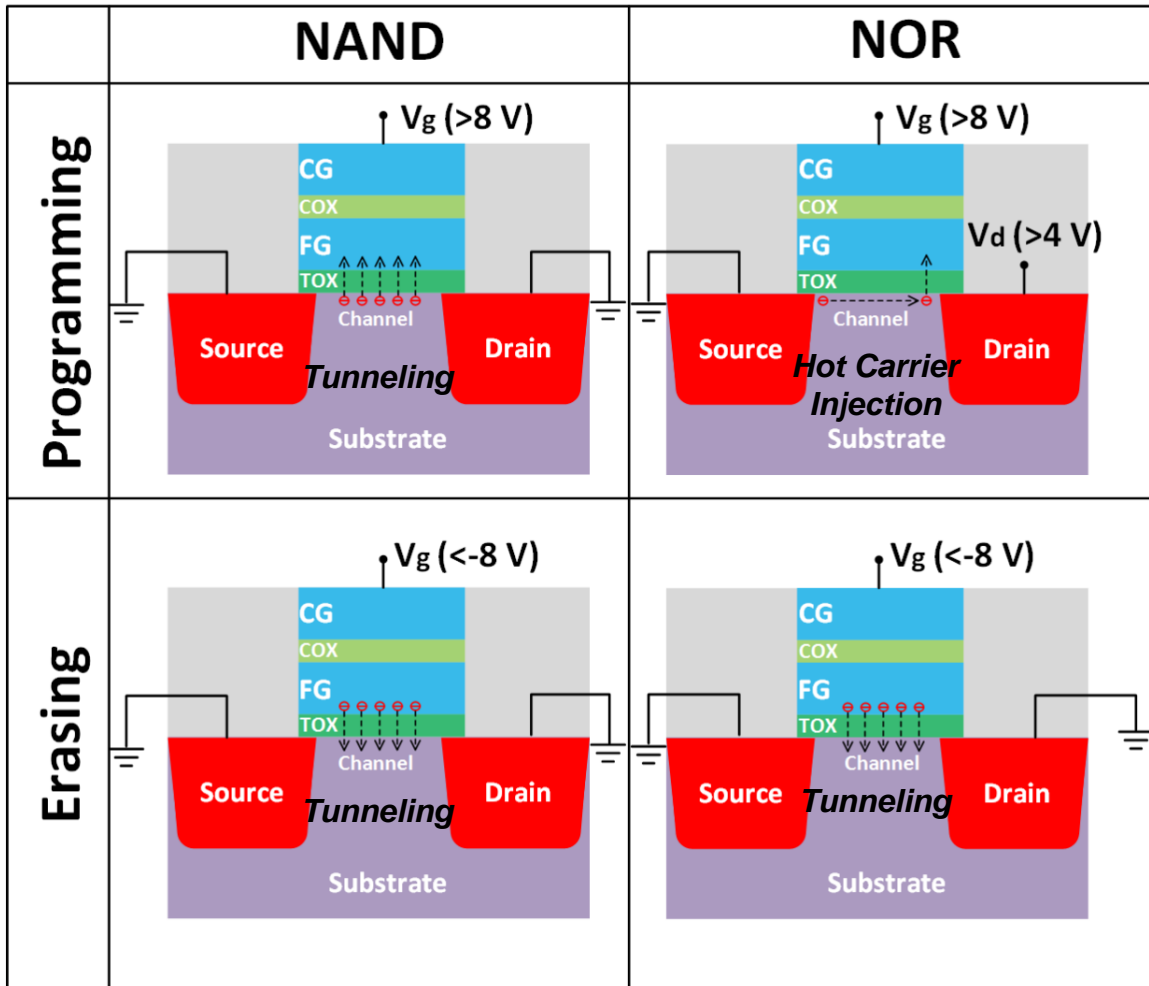
N-type device
CG: control gate
COX: control oxide
FG: floating gate
TOX: tunnel oxide

Read-out: by applying a small read voltage

cell current — read current — Unprogrammed FG — Programmed FG — high V_T — low V_T — read voltage — Control-gate voltage

Kaustav Banerjee
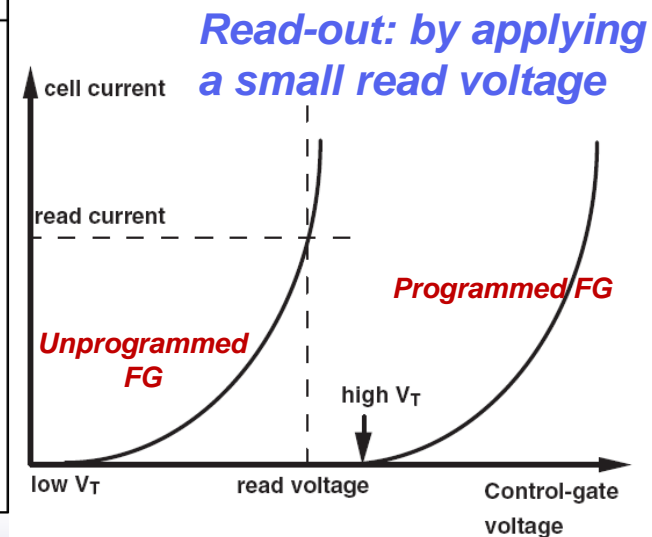
NAND vs. NOR circuit

NOR-EEPROM

Control Gate

Floating Gate

Drain   Source

Bit Line (BL)

WL 1

Unit cell

World Line (WL)

WL 15

WL 16

Source Line (SL)

1 Transistor/ 1 Bit
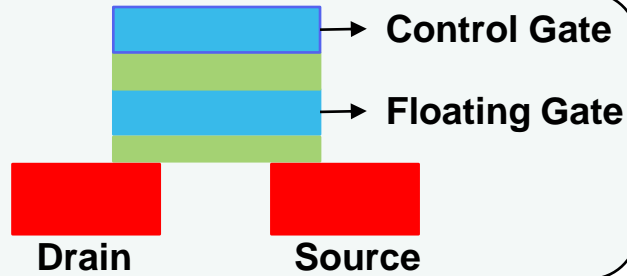
Devices can be Randomly accessed

NOR: It's available for random access

NAND-EEPROM

Select Gate SG (D)   Control Gate   Select Gate SG (S)

WL1 WL2 WL3   WL15 WL16

Drain   Source

SG (D)   Bit Line (BL)

WL 1

WL 2   Unit cell

World Line (WL)

WL 15

WL 16

SG (S)   Source Line (SL)

18 Transistor/ 16 Bit
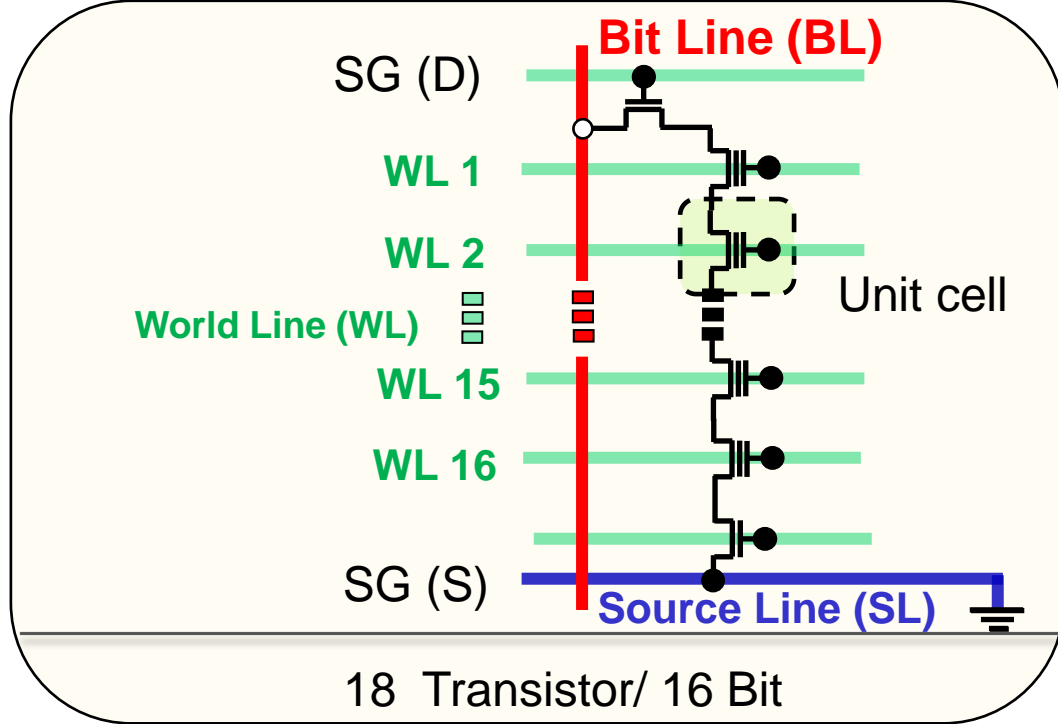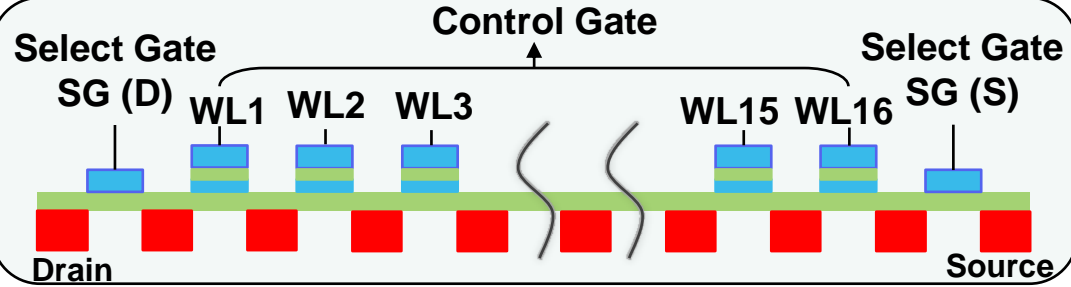
Both SG devices needed for read out

NAND: It's allowed to access one bit only per readout operation

Lectures 16, ECE 122A, VLSI Principles

Kaustav Banerjee

NAND vs. NOR

Merits of NAND

★ *Tunneling rate > Hot carrier inj.*

① **High speed programming** (★)
② **High speed erasing**
③ **High capacity**
④ **Low cost per bit**

Merits of NOR

① **High speed random access**
② **Byte programming**

(Individual bit cell for programming/erasing process)

Demerits of NAND

① **Slow random access**
② **Byte programming cannot be performed**

  (1 Byte= 8 bit)

(Need of bit by bit programming/erasing process / pre cell)

③ **Slow read operations**

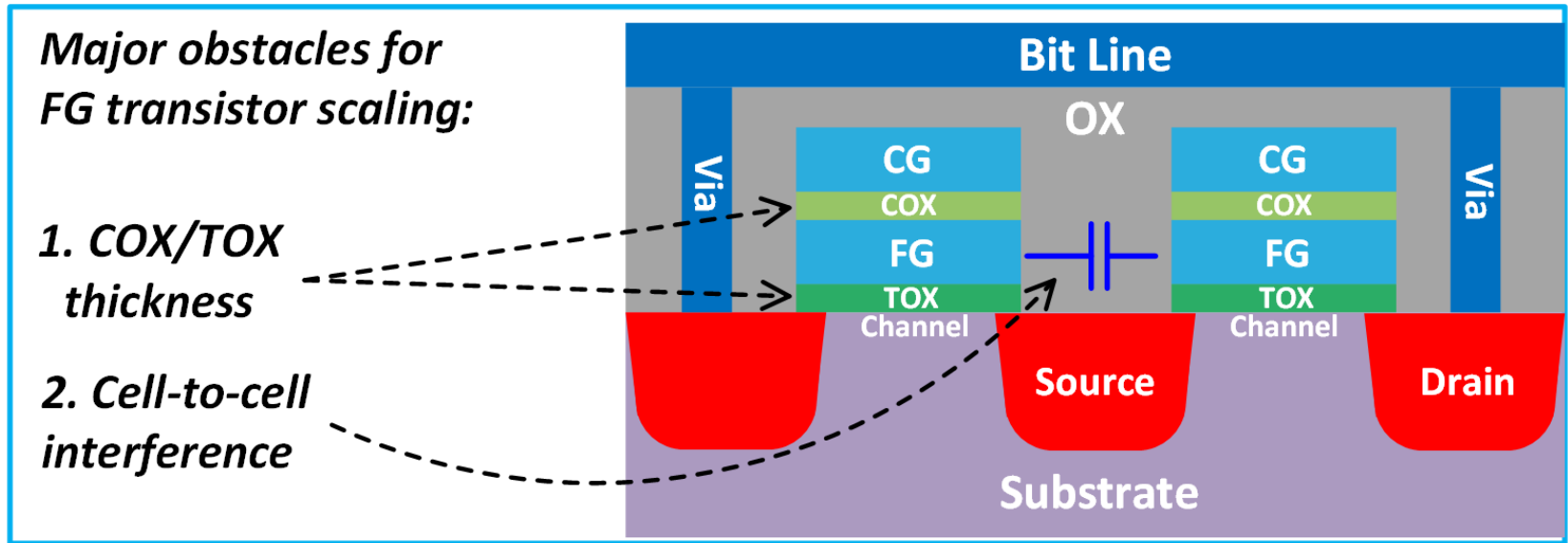*(Serial access => sequential data readout)*

Demerits of NOR

① **Slow programming**
② **Slow speed erasing**
③ **Low capacity**
④ **High cost per bit**

(NOR must be erased in large "chunks". Those chunks can often be subdivided and erased in smaller subunits resulting in a performance penalty)

- Applications -

• **Suitable for Data Memory (Handy terminal, Voice recorder, DSC, Fax modem, etc)**

- Applications -

• **Suitable for replacement of EPROM**

• **Suitable for control memory (BIOS, Cellular, HDD, etc)**

Kaustav Banerjee

Scaling issues

Major obstacles for FG transistor scaling:

1. COX/TOX thickness

2. Cell-to-cell interference

Bit Line · OX · Via · CG · COX · FG · TOX · Channel · Source · Drain · Substrate

➤ thin COX/TOX -> leakage current -> short retention time

➤ small cell-to-cell distance -> Vth perturbation by adjacent cells

Promising solution: 2D materials….

Kaustav Banerjee

3-D ICs : Multiple Active Si Layers

K. Banerjee et al., Proceedings of the IEEE, 2001

- ## Advantages

  - Reduce Interconnect Length by Vertically Stacking Multiple Si Layers

  - Reduce Chip Area, power dissipation and improve Chip Performance

  - Heterogeneous integration possible, e.g., memory, digital, analog, optical, etc. using different substrates (Si, III-V etc)

Optical I/O

Analog / RF

DRAM

Distributed Memory

Logic

Kaustav Banerjee