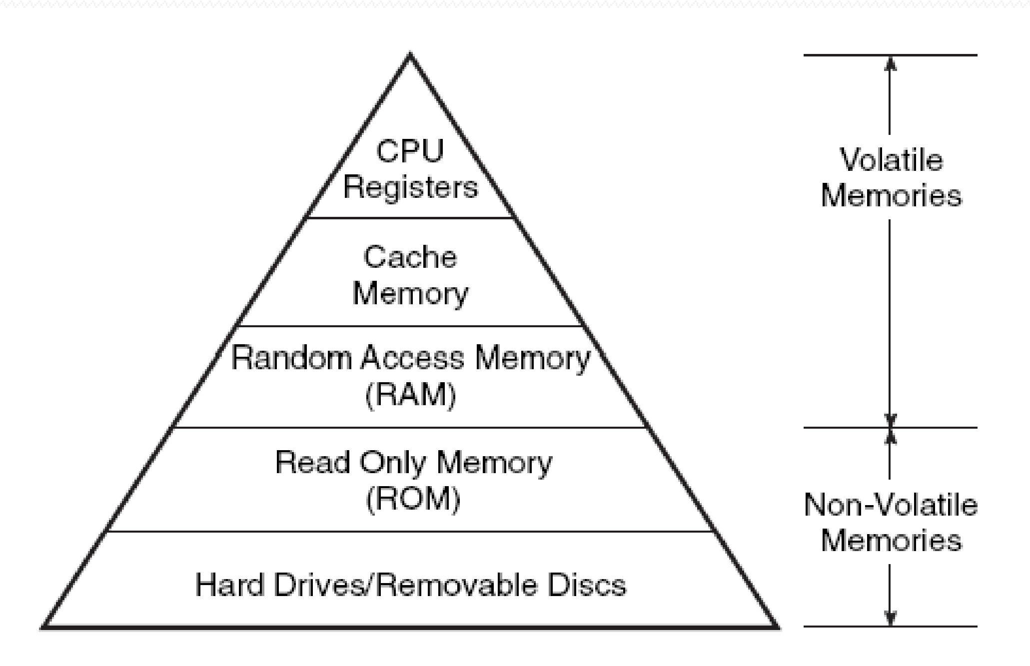
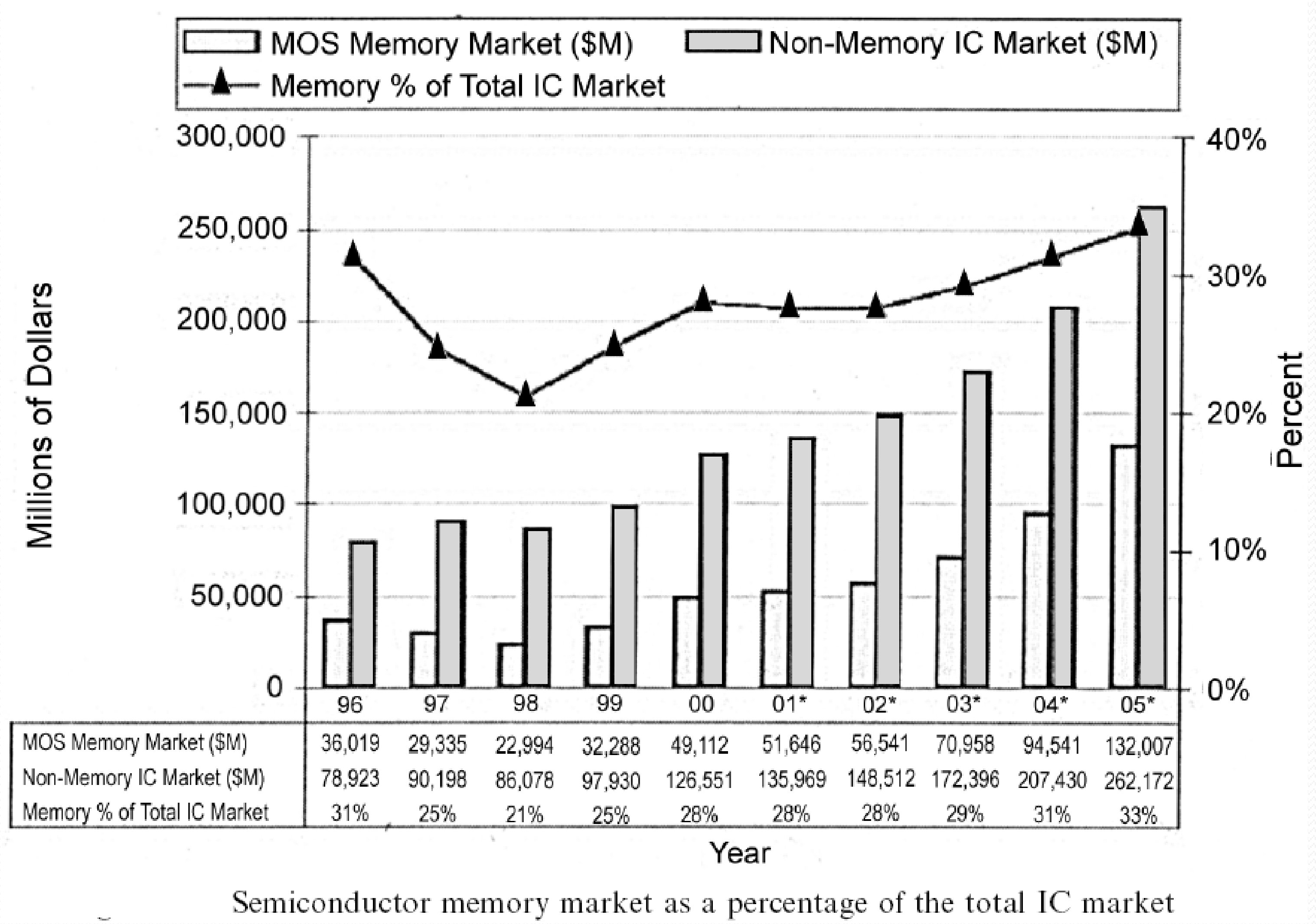
Memories

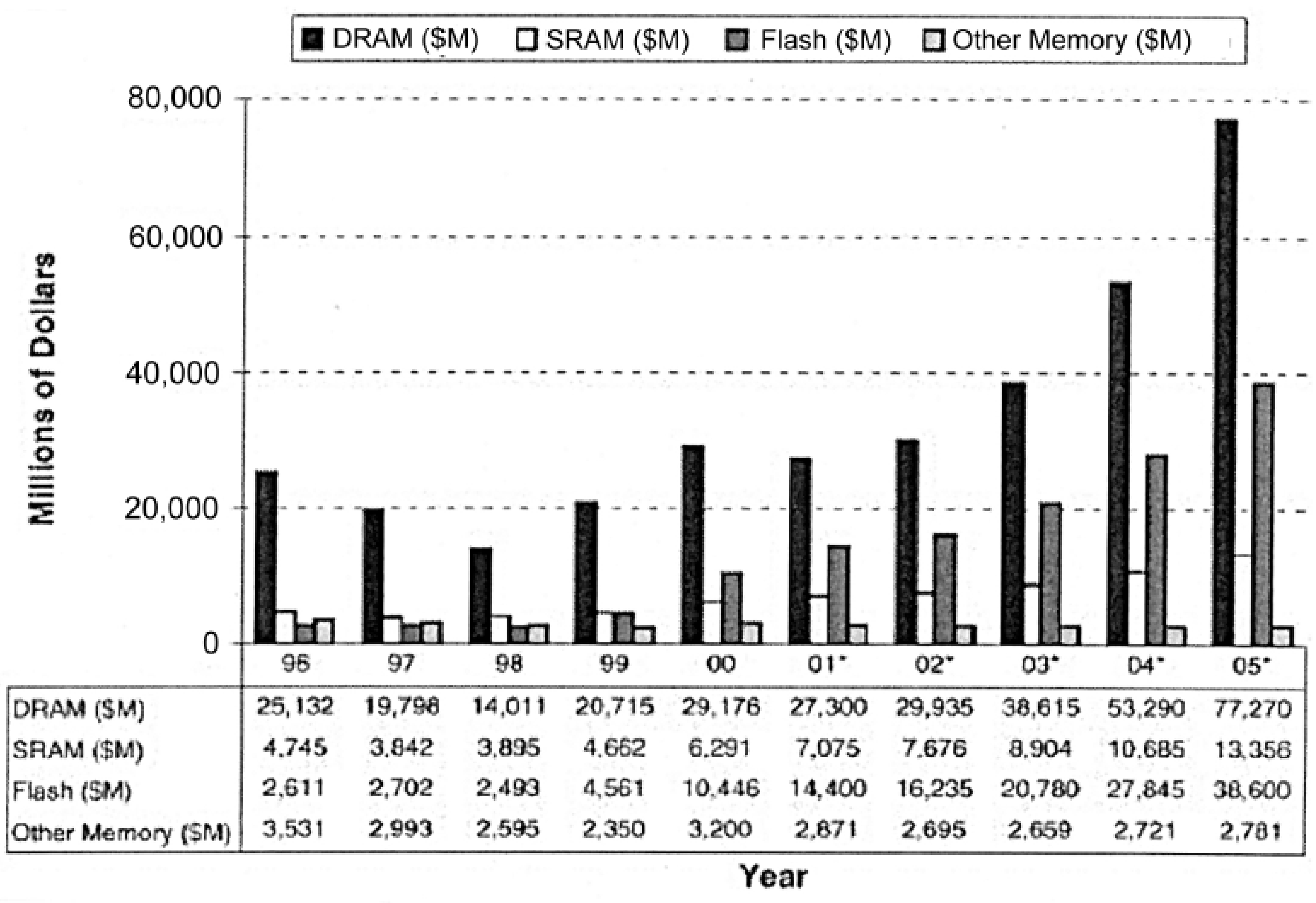


Various types of memory present in a typical computer system.

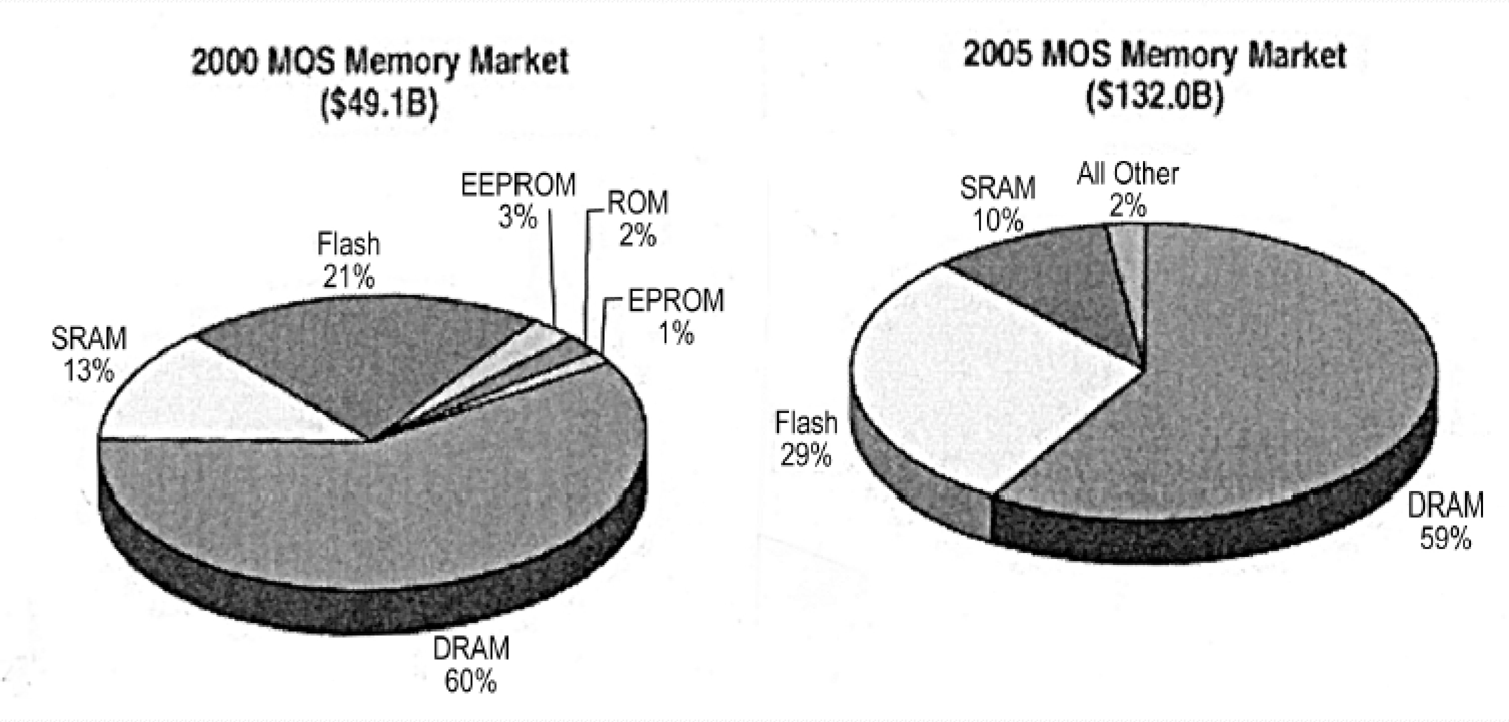


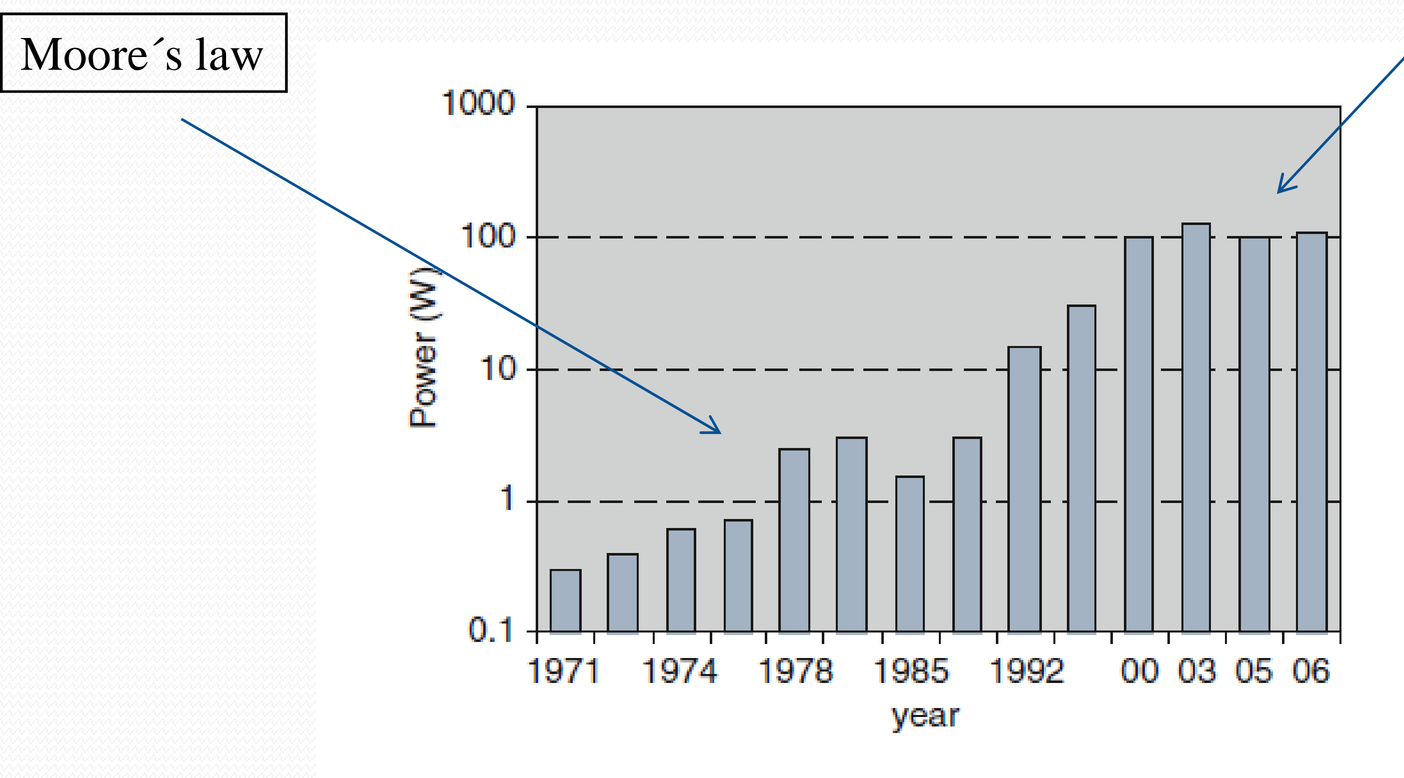
	96	97	98	99	00	01*	02*	03*	04*	05*
MOS Memory Market ($M)	36,019	29,335	22,994	32,288	49,112	51,646	56,541	70,958	94,541	132,007
Non-Memory IC Market ($M)	78,923	90,198	86,078	97,930	126,551	135,969	148,512	172,396	207,430	262,172
Memory % of Total IC Market	31%	25%	21%	25%	28%	28%	28%	29%	31%	33%

Semiconductor memory market as a percentage of the total IC market

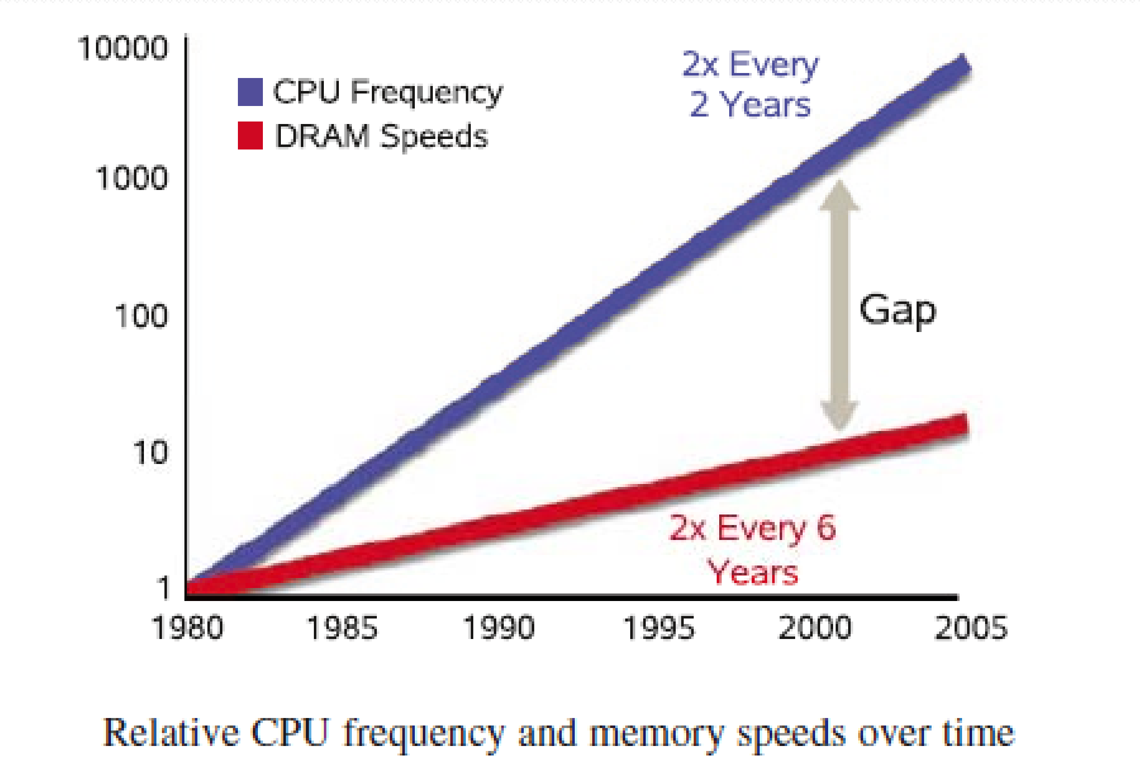


	96	97	98	99	00	01*	02*	03*	04*	05*
DRAM ($M)	25,132	19,798	14,011	20,715	29,176	27,300	29,935	38,615	53,290	77,270
SRAM ($M)	4,745	3,842	3,895	4,662	6,291	7,075	7,676	8,904	10,685	13,356
Flash ($M)	2,611	2,702	2,493	4,561	10,446	14,400	16,235	20,780	27,845	38,600
Other Memory ($M)	3,531	2,993	2,595	2,350	3,200	2,871	2,695	2,659	2,721	2,781

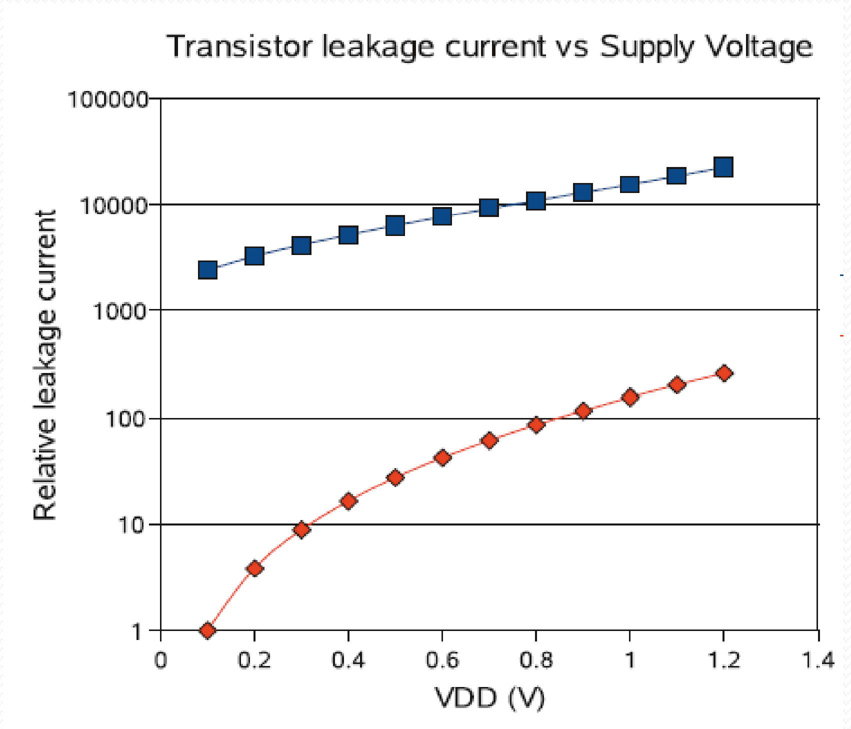

2000 MOS Memory Market
($49.1B)
EEPROM
3%
ROM
2%
Flash
21%
EPROM
1%
SRAM
13%
DRAM
60%
2005 MOS Memory Market
($132.0B)
SRAM
10%
All Other
2%
Flash
29%
DRAM
59%

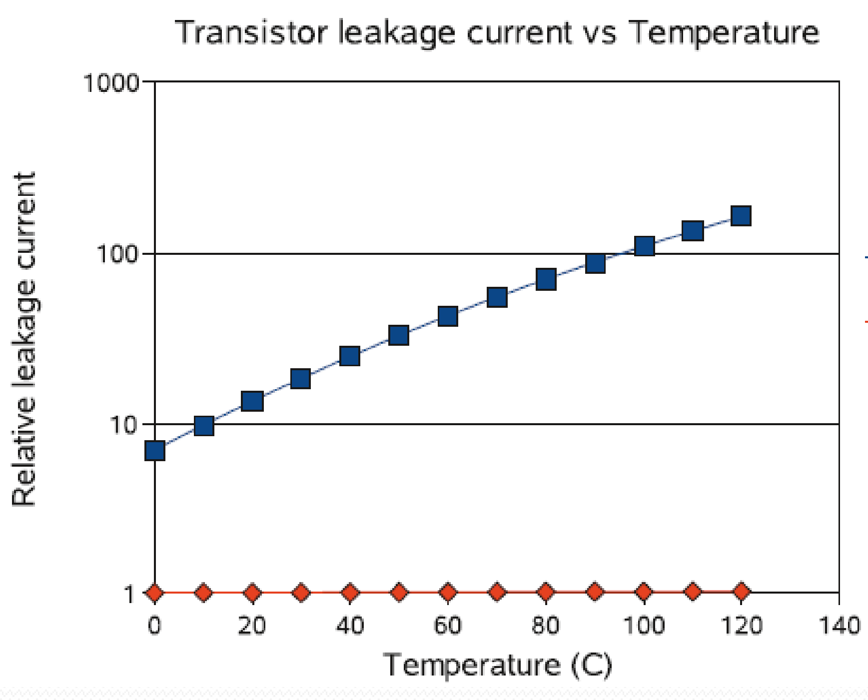


Power consumption of Intel microprocessors from 1970 to 2006;



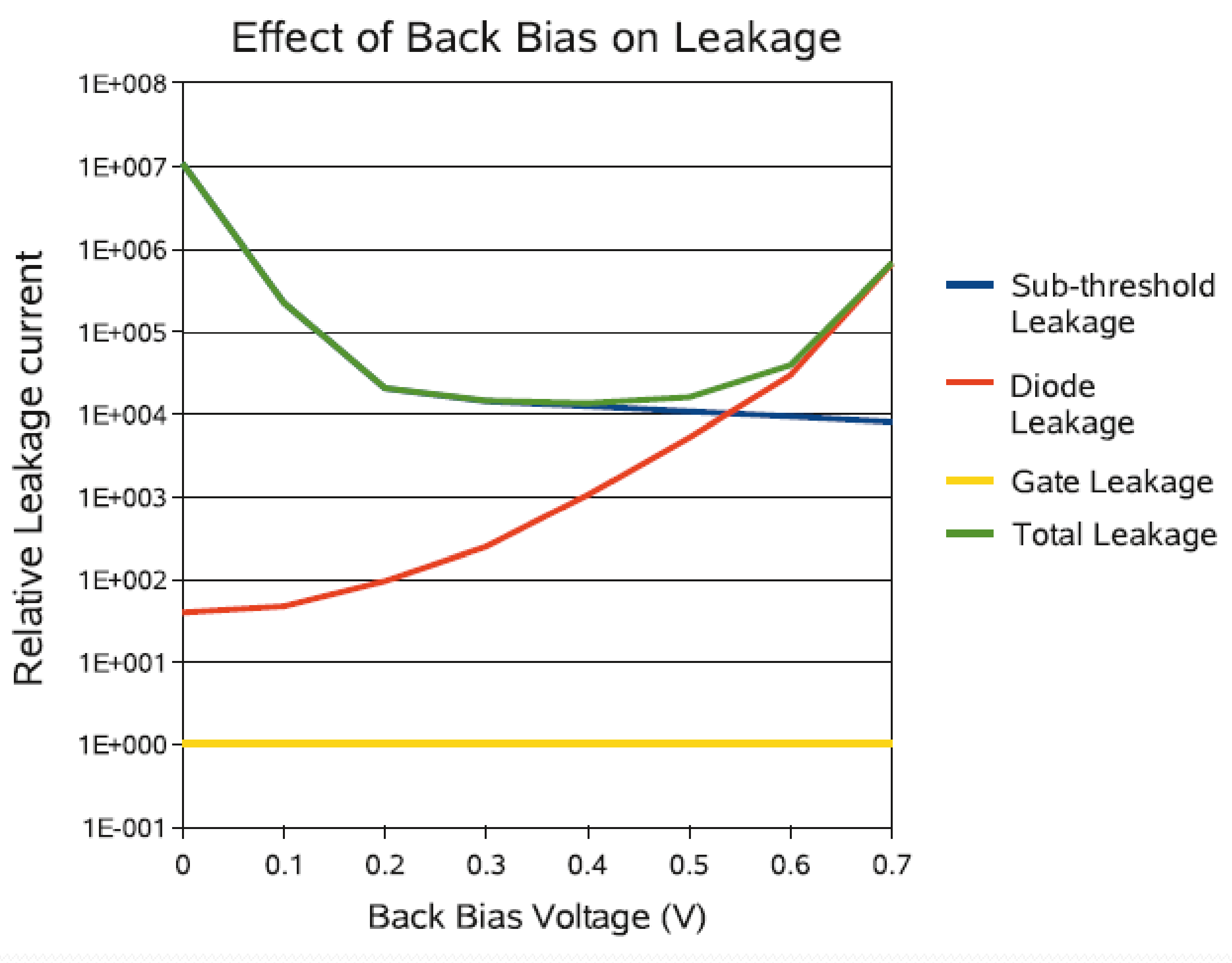
Relative CPU frequency and memory speeds over time





Relative Drain Leakage

Relative Gate Leakage

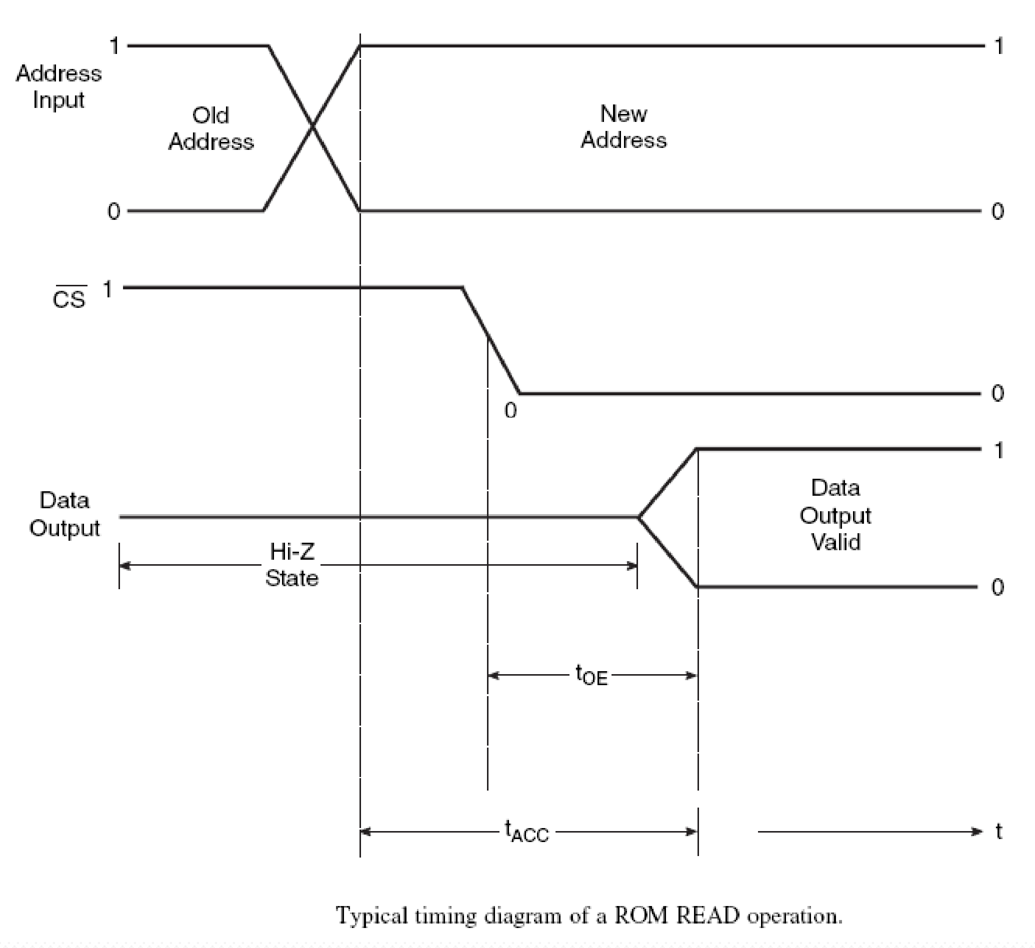


$$V_{th} = V_{th0} +/- \gamma * (|V_{SB}|)^{1/2}$$

Non-Volotile Memories (NVM´s)

	Cell Phone	Consumer	Automotive	Computer & Communication
EPROM	Analog, Residential	Games, Set Top Box	Engine Mgt	HDD, Copiers, Fax, Switching
FLASH	Digital (GSM)	Set Top Box PDA	All Power Train Car Navigation, ABS, GPS	HDD, PC Bios, CDROM
EEPROM	Digital (GSM)	Audio, Video	All Car Body	PC SPD, Graphic boards, Printers

Read-Only Memory (ROM)



Typical timing diagram of a ROM READ operation.

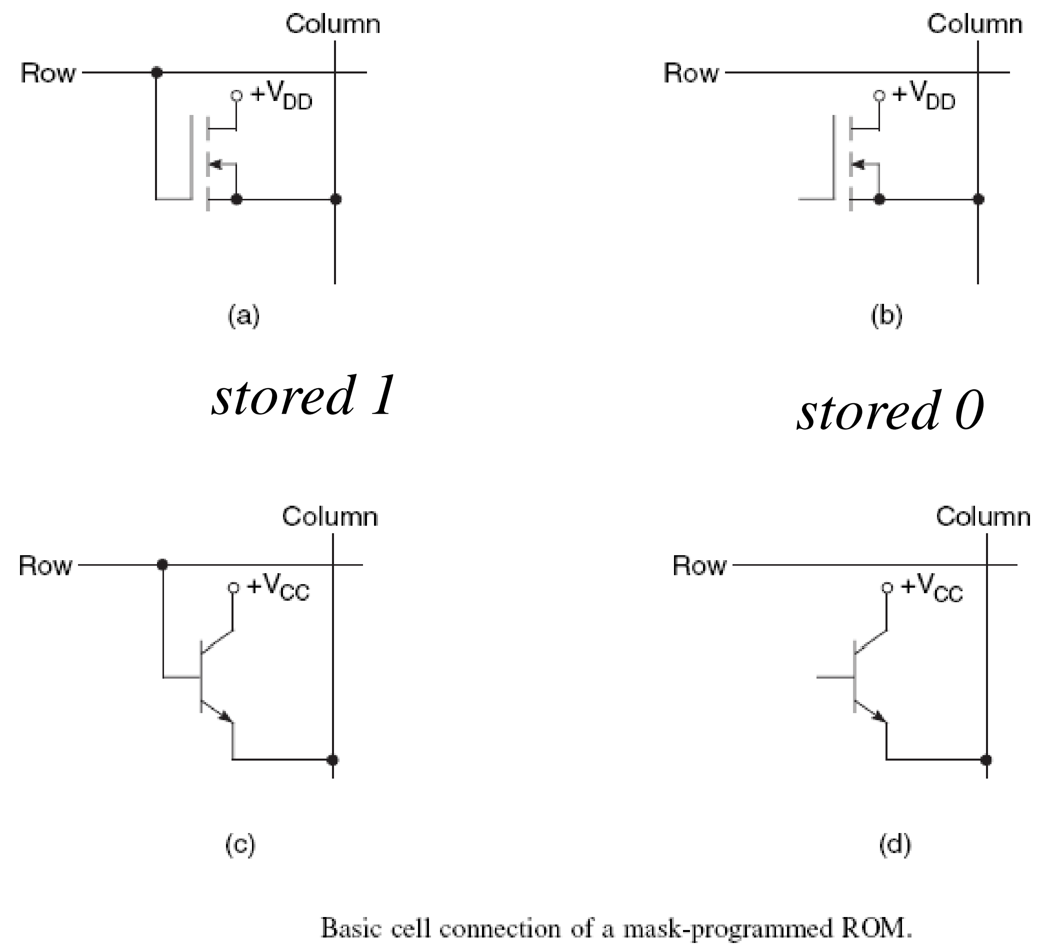
Mask-programmed ROM

- ROM is programmed at the manufacturer's site according to the specifications of the customer

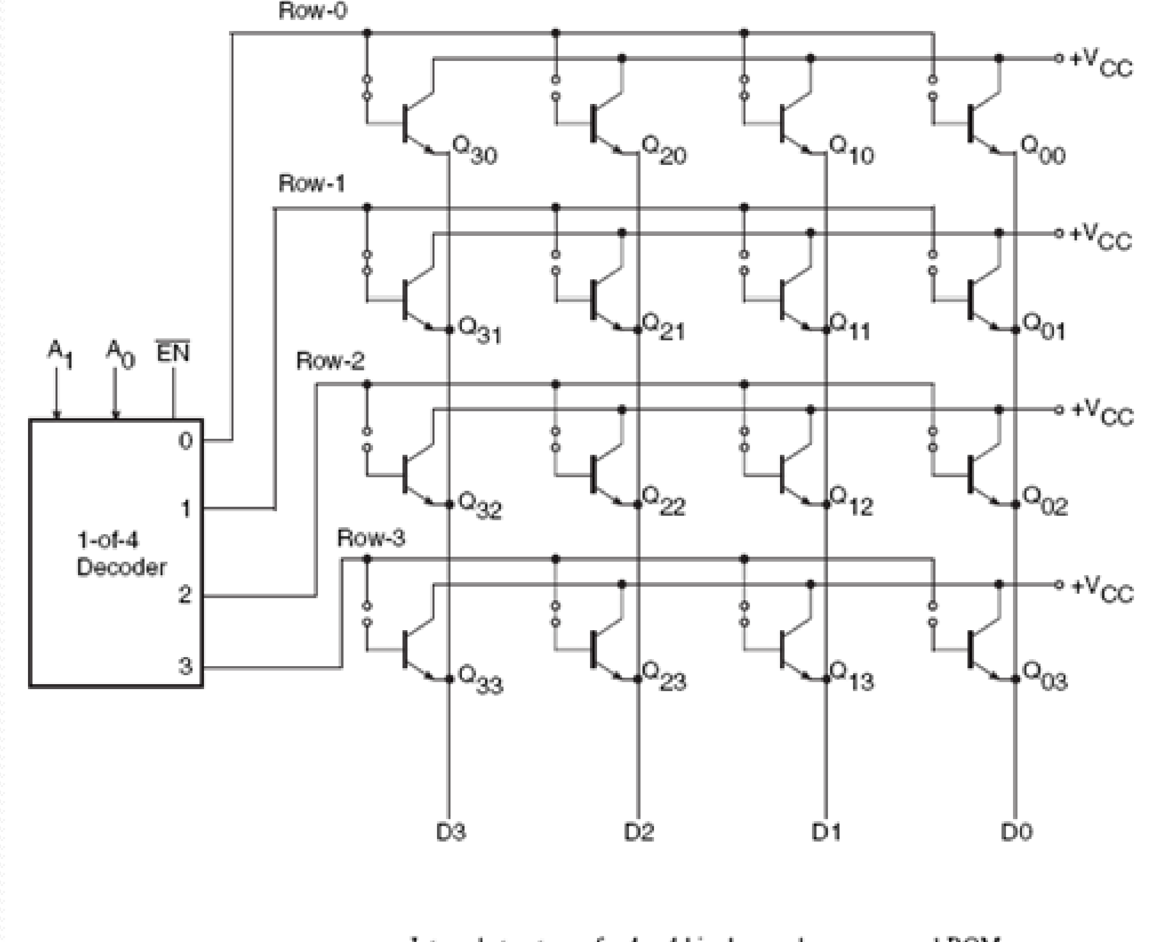
- economical only when manufactured in large quantities

- once programmed, it cannot be reprogrammed.

- The basic storage element is an NPN bipolar transistor, connected in common-collector configuration, or a MOSFET in common-drain configuration.



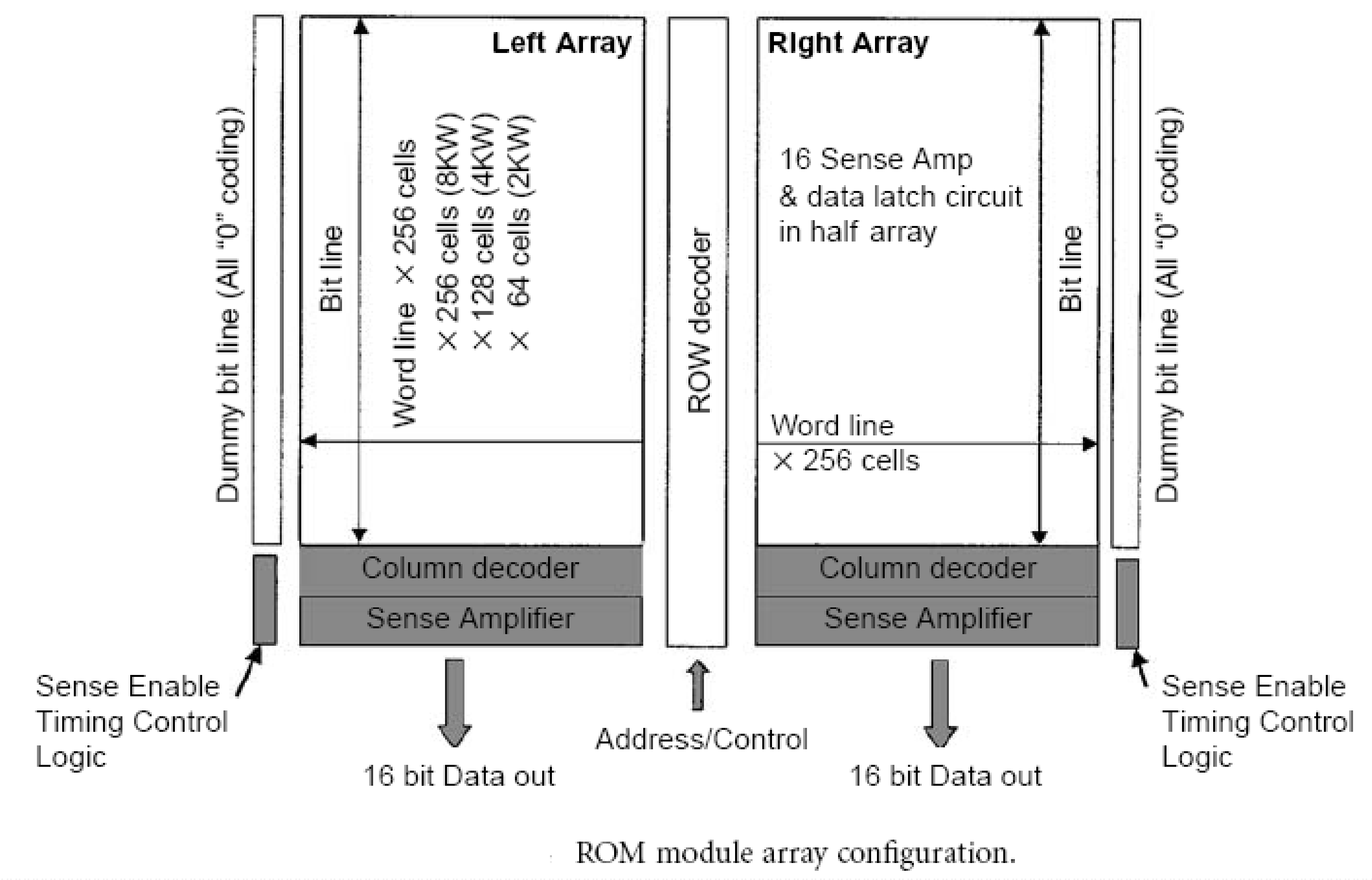
Basic cell connection of a mask-programmed ROM.



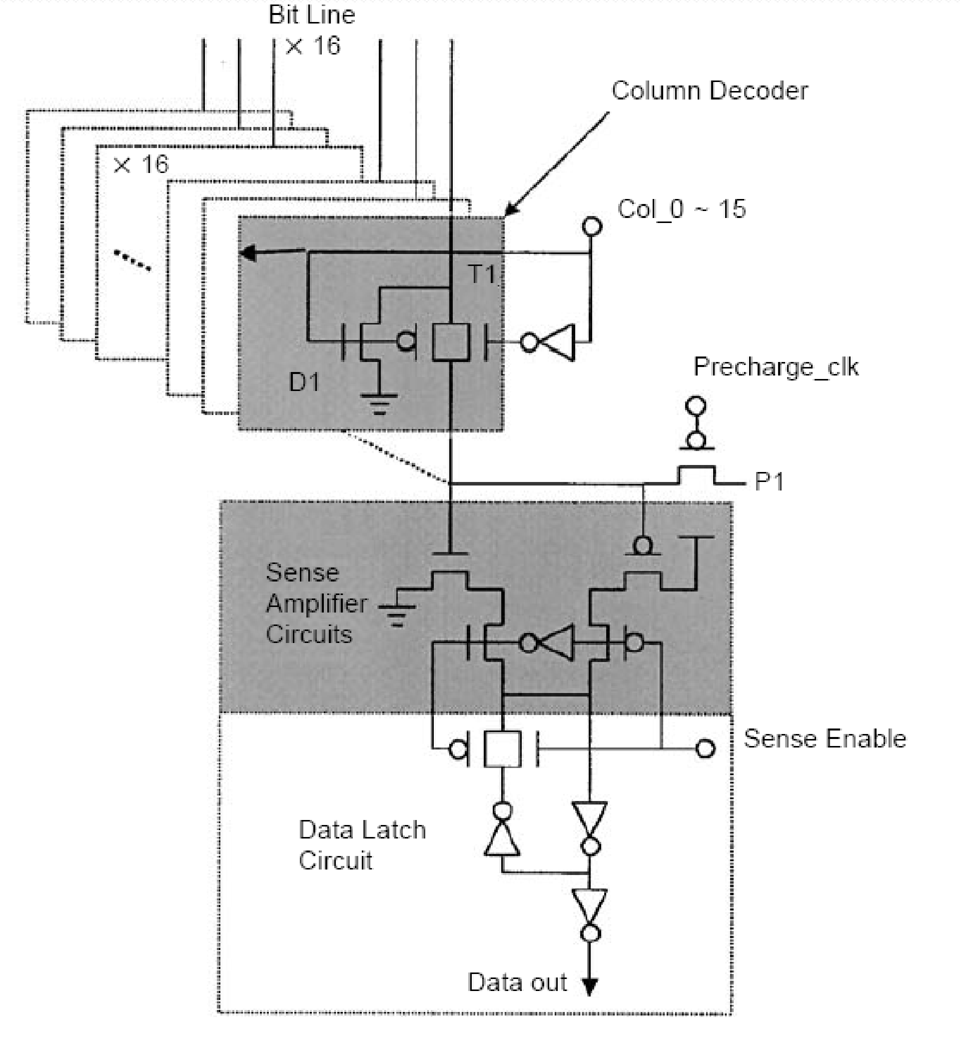
Internal structure of a 4 × 4 bipolar mask-programmed ROM.

Truth Table

Address		Data			
A_1	A_0	D_3	D_2	D_1	D_0
0	0	1	0	1	0
0	1	1	0	0	0
1	0	0	1	1	1
1	1	0	1	1	0

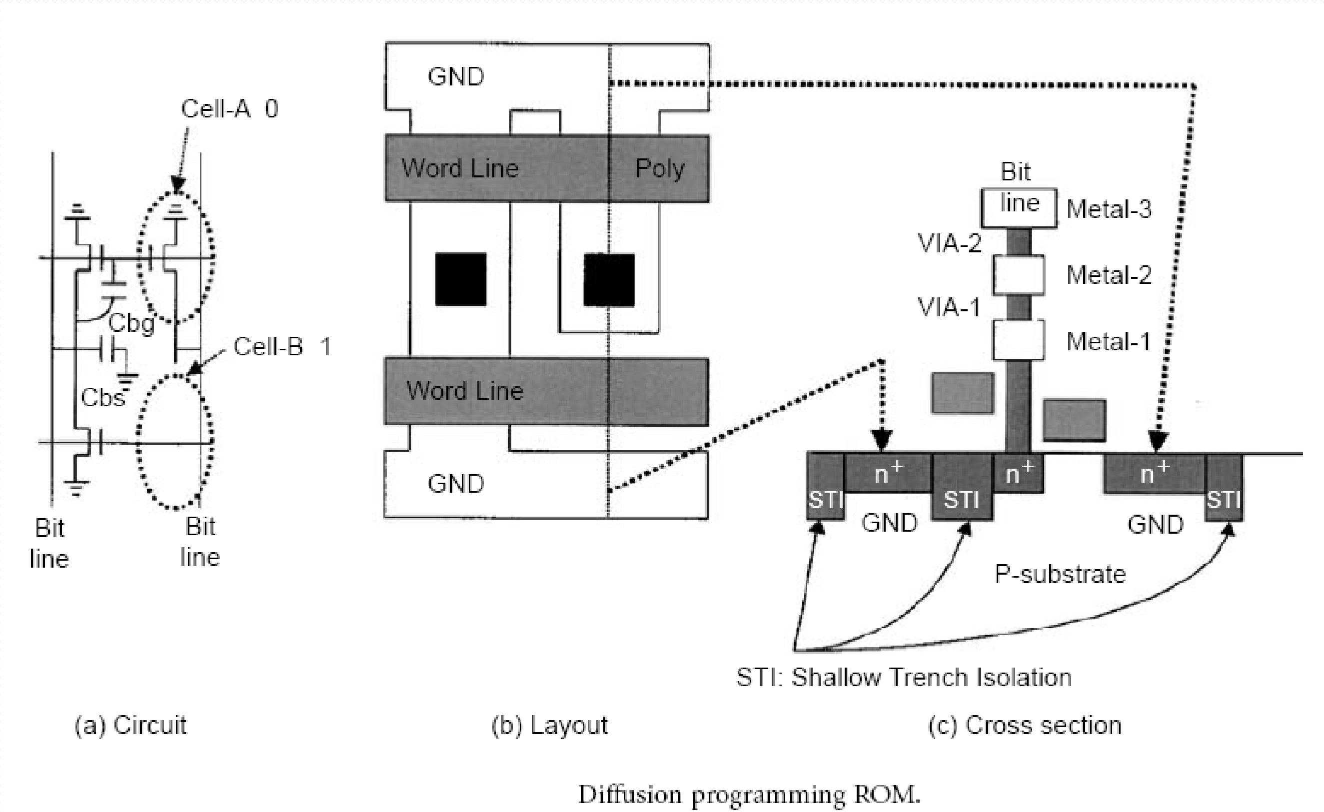


ROM module array configuration.



Detail of low power selective bit line precharge and sense amplifier circuits.

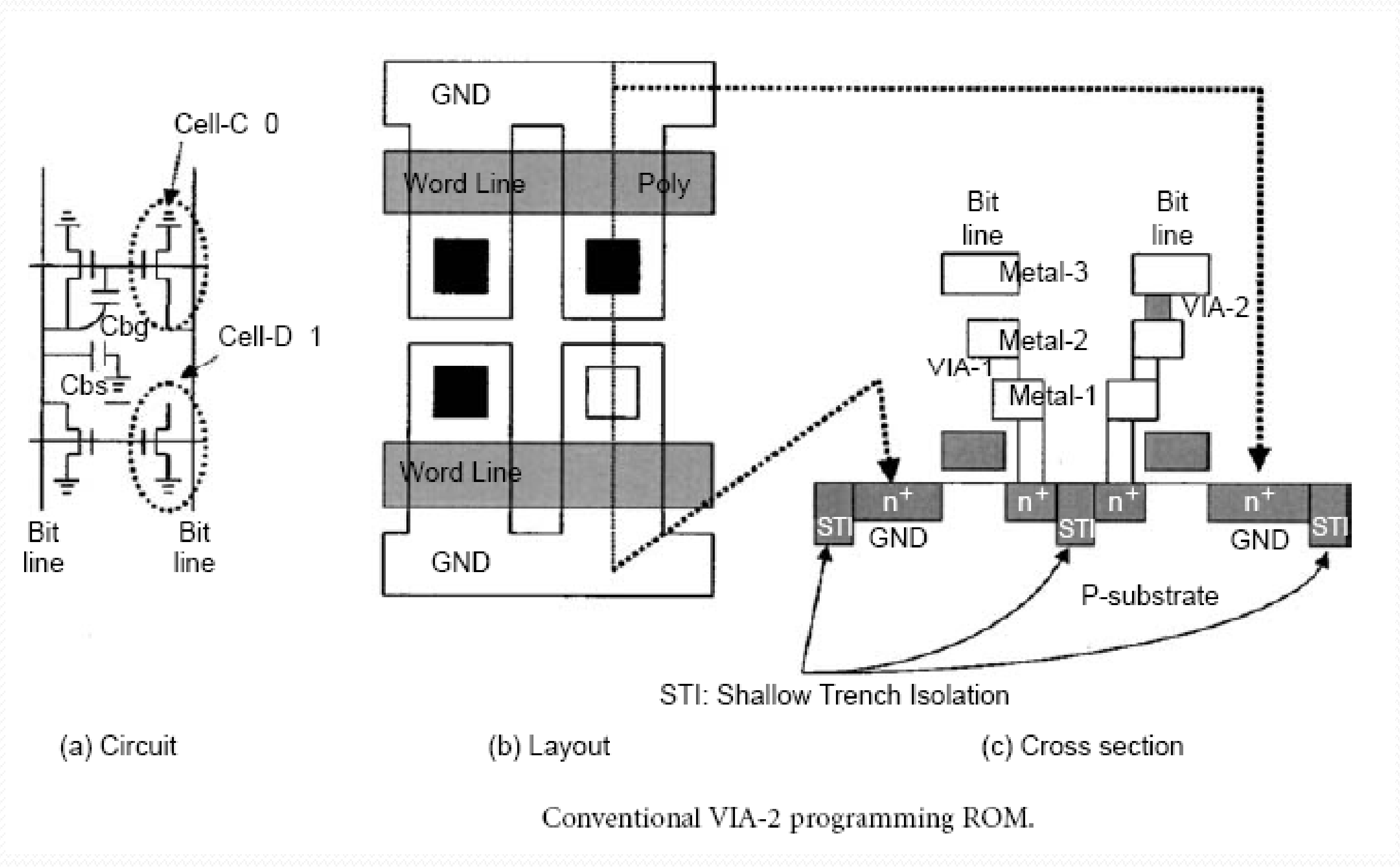
Diffusion-programming ROM



Diffusion programming ROM.

- highest density: bit line contact to discharge transistor can be shared by two-bit cells
- very long fabrication cycle time: diffusion programming at early process stage

VIA-2 Contact Programming ROM (conventional)



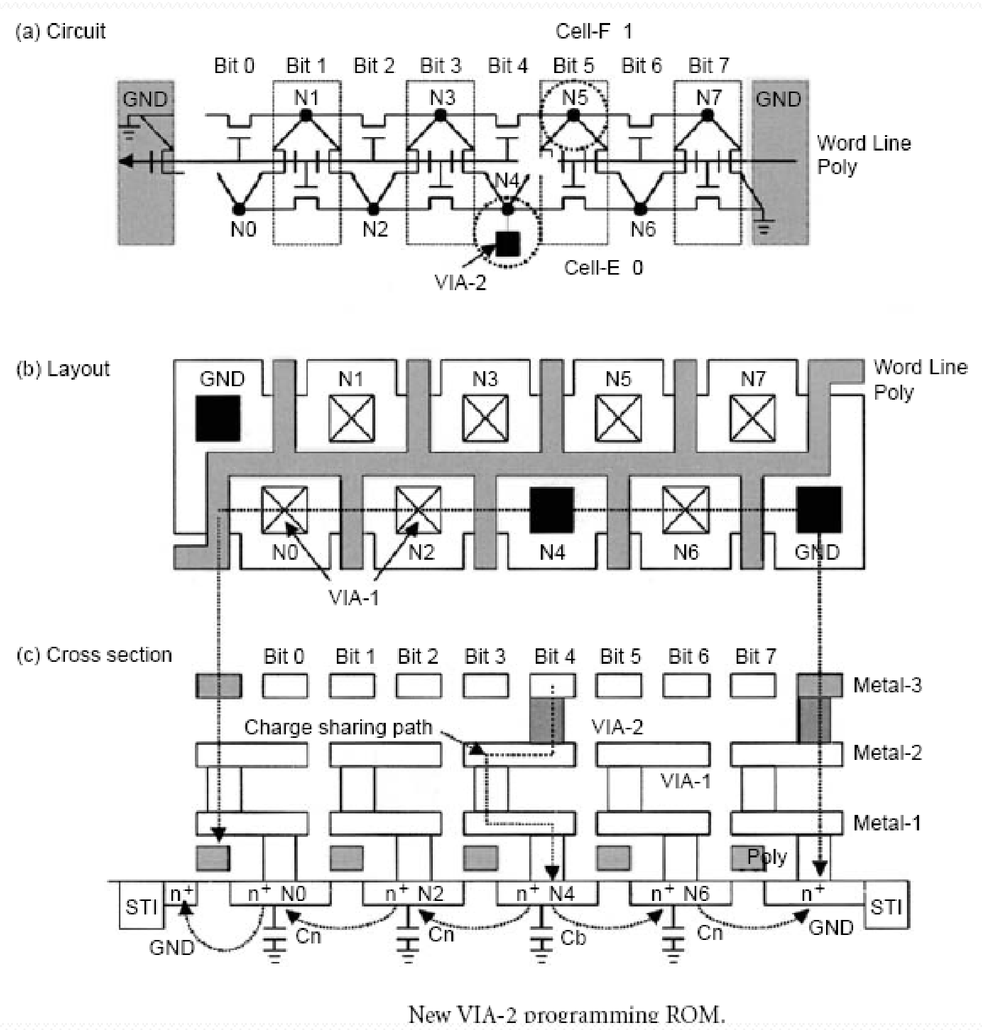
Conventional VIA-2 programming ROM.

- The VIA-2 is final stage of process and base process can be completed just before VIA-2 etching and remaining process steps are quite few

- VIA-2 ROM fabrication cycle time is about 1/5 of the diffusion ROM

- Drawback: poorer density ⇒ diffusion area and contact must be separated in each ROM bit cell

VIA-2 Contact Programming ROM (new)



New VIA-2 programming ROM.

- 8-bit block with GND on each side
- higher density

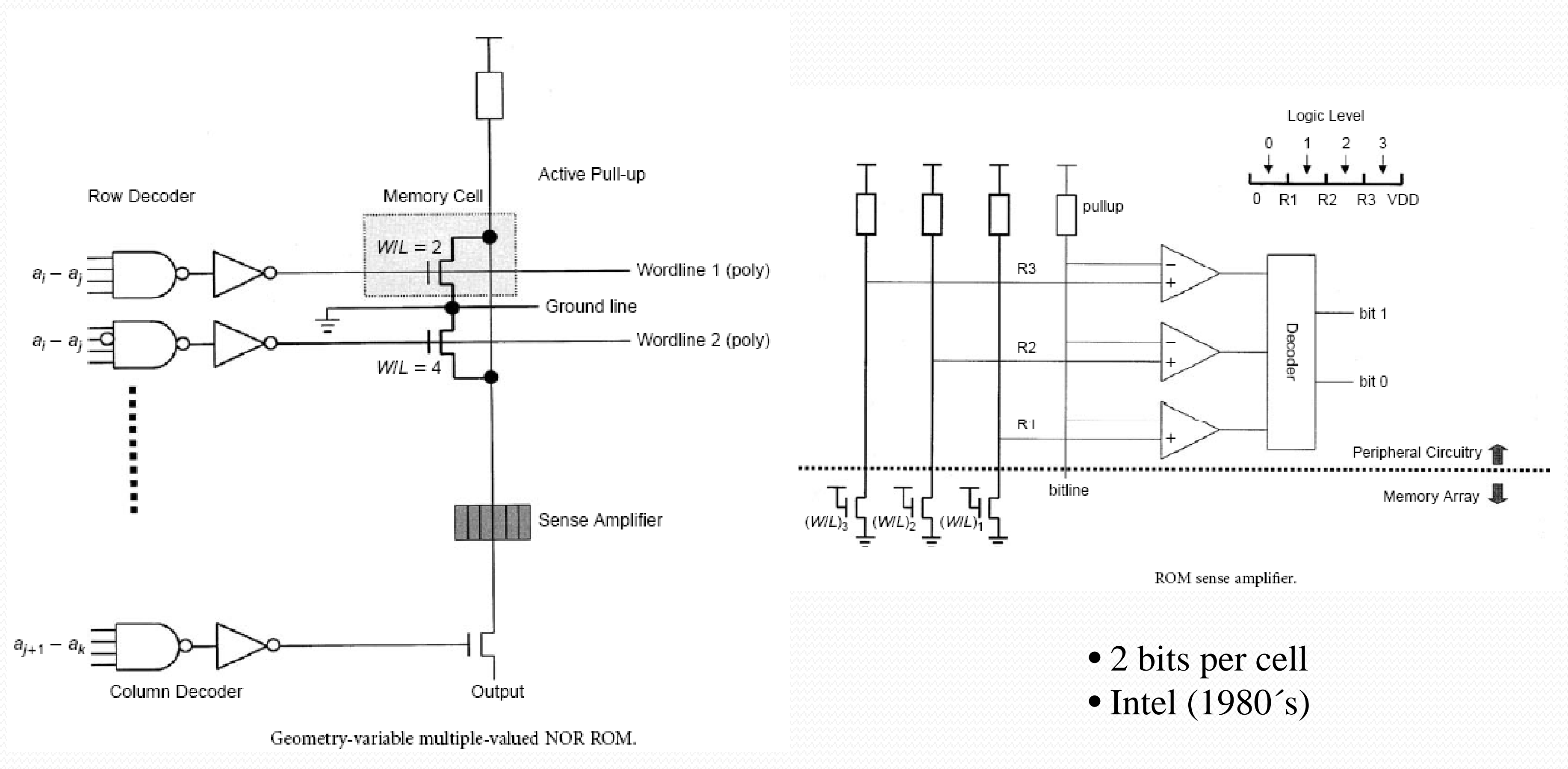
TABLE 51.1 Comparison of ROM Performance

Comparison Item	Diffusion ROM	Conventional VIA-2 ROM	New VIA-2 ROM
8KW (Area ratio)	1.0	1.2	1.04
TAT (Day ratio)	1.0	0.2	0.2
Speed @ 2.13 V, 125dc. Weak.	83 MHz	86 MHz	123 MHz
Speed @ 2.13 V, 125dc. Typical.	166 MHz	98 MHz	149 MHz
Speed @ 2.81 V, –40dc. Strong.	277 MHz	179 MHz	294 MHz
Speed @ 1.66 V. 125dc, Typical.	103 MHz	75 MHz	106 MHz
Power@2.81 V, –40dc. Strong. 100 MHz. (16-bit single access)	15.6 mW	19.3 mW	2 UrnW
Power@2.81 V@40dc. Strong. 100 MHz. (32-bit dual access)	29.6 mW	37.1 mW	401 mW

Performance was measured with worst coding (all coding "1").

Multi-level ROM

- transistor-cell (W/L) adjusted according to its logic state



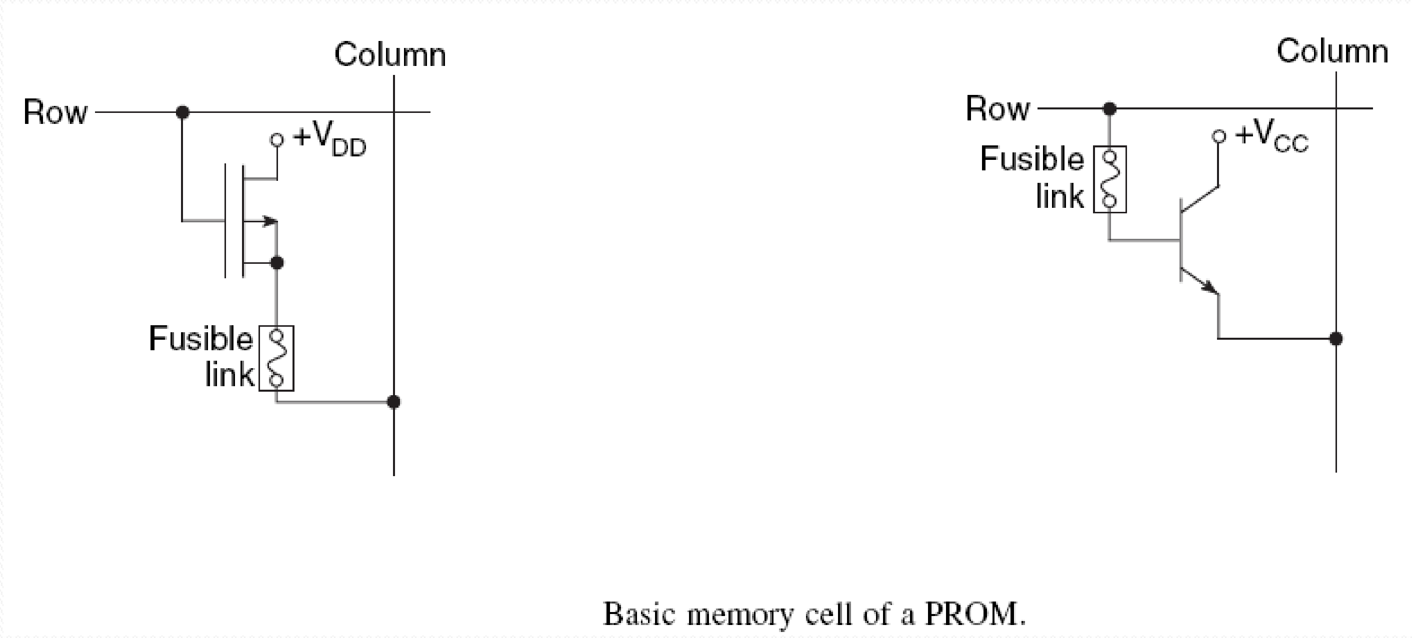
Geometry-variable multiple-valued NOR ROM.

ROM sense amplifier.

- 2 bits per cell
- Intel (1980´s)

Programmable ROM (PROM)

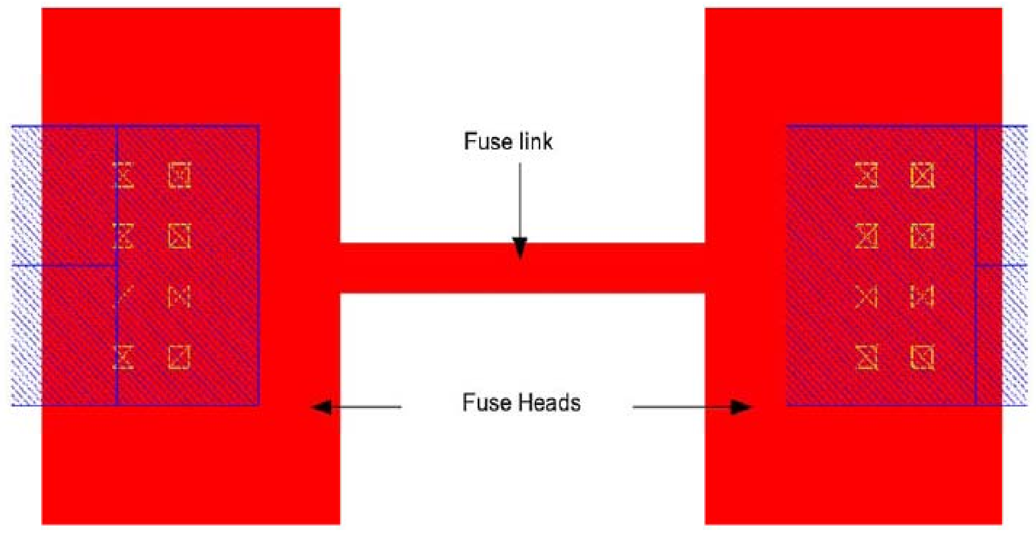
- programmed by the customer
- fuse links: metal or poly-Si



Basic memory cell of a PROM.

poly-Si/metal fuses

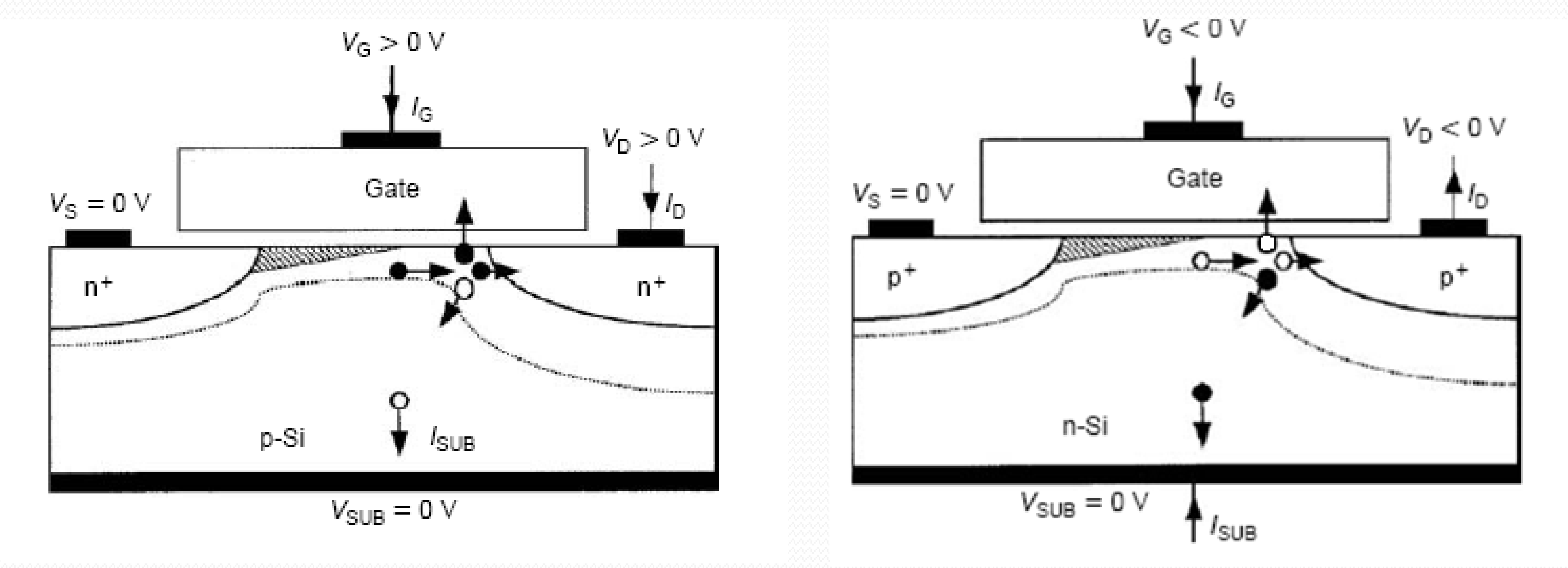
- line heats up to fusion by self-heating mechanism due to current intensity during programming mode
- fusion increases its resistance and eventually opens the link



polyfuse/metalfuse layout

Stacked-Gate Non-Volatile Memory (EPROM/EEPROM)

• **DAHC: Drain-Avalanche Hot Carrier**

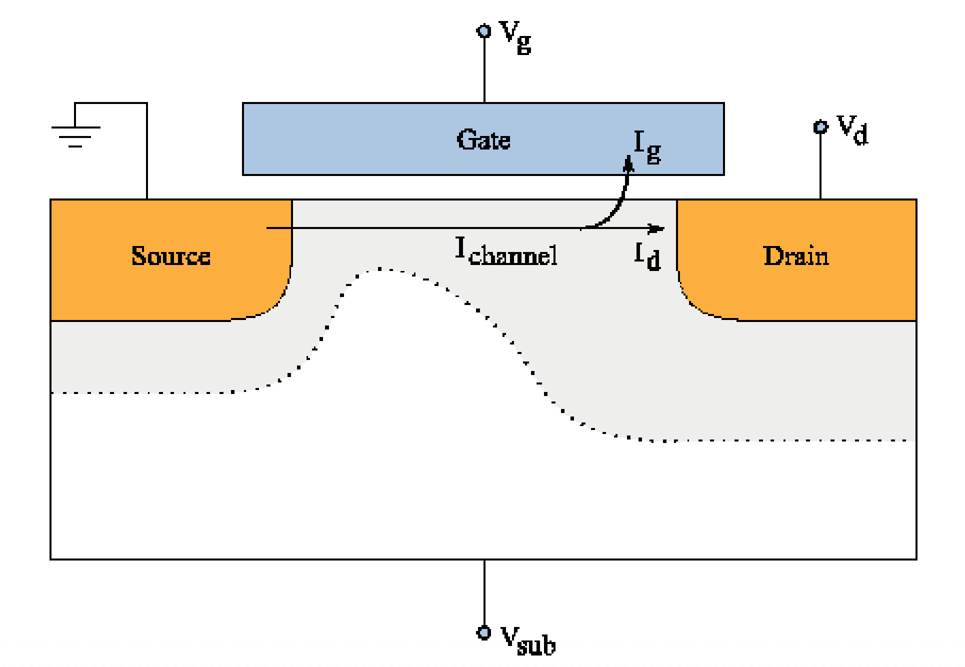


Schematic illustration of the channel hot carrier effect in (a) n-channel MOSFET, and (b) p-channel MOSFET.

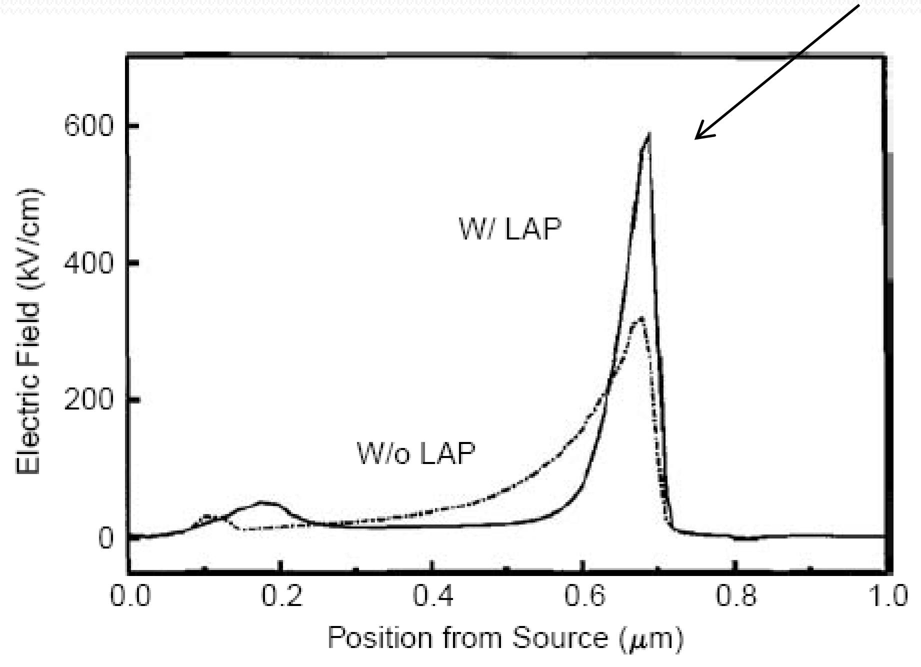
• HCI: Hot Carrier Injection

• either an electron or hole gains sufficient kinetic energy to overcome a potential barrier necessary to break an interface state

• can be injected into the gate dielectric, where they can get trapped



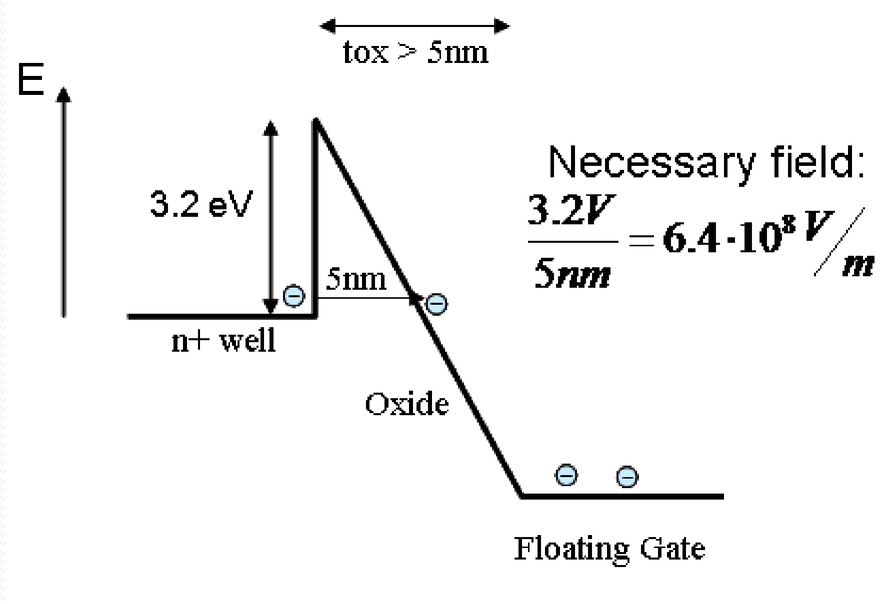
pinch-off region



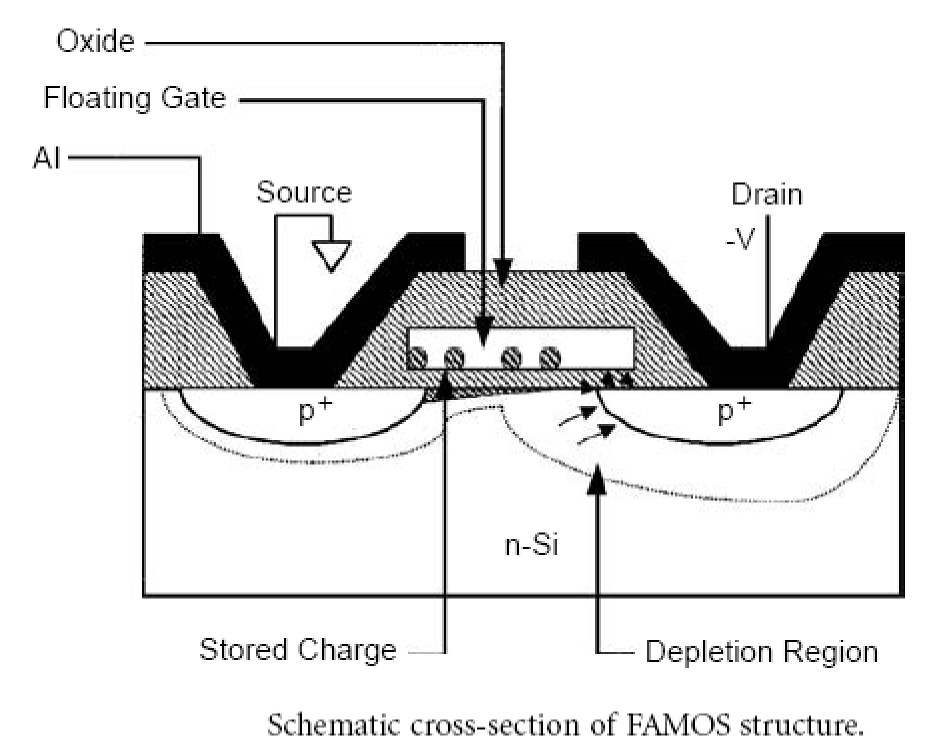
Simulated electric field along the channel in the n-channel MOSFET.

• Fowler-Nordheim tunneling

• current flowing across MOS structure at the high electric field in the oxide; electrons tunnel from semiconductor conduction band into oxide conduction band through part of the potential barrier at the semiconductor-oxide interface (oxide 5-10 nm thick)

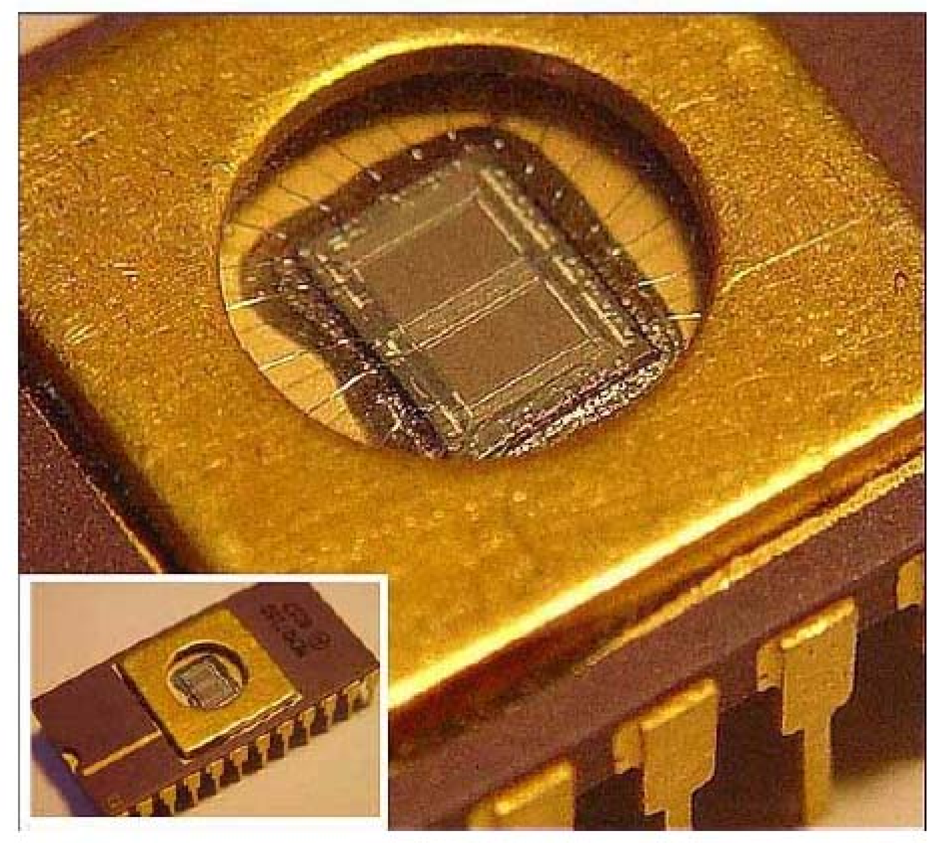


• **floating-gate MOS (FAMOS) transistor**



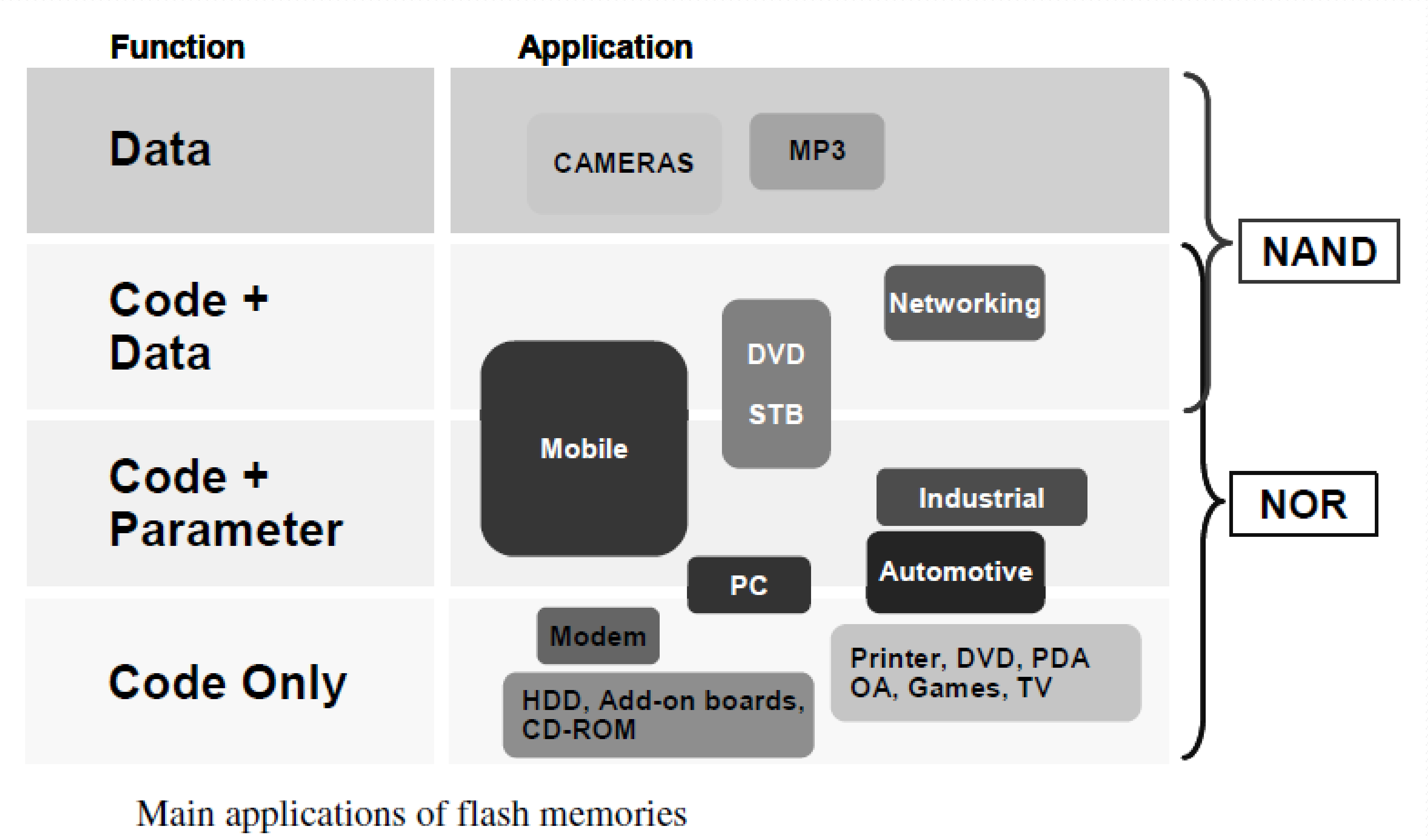
Schematic cross-section of FAMOS structure.

- not electrically erasable (only by time-consuming UV-irradiation)
- compatible with double-poly processes

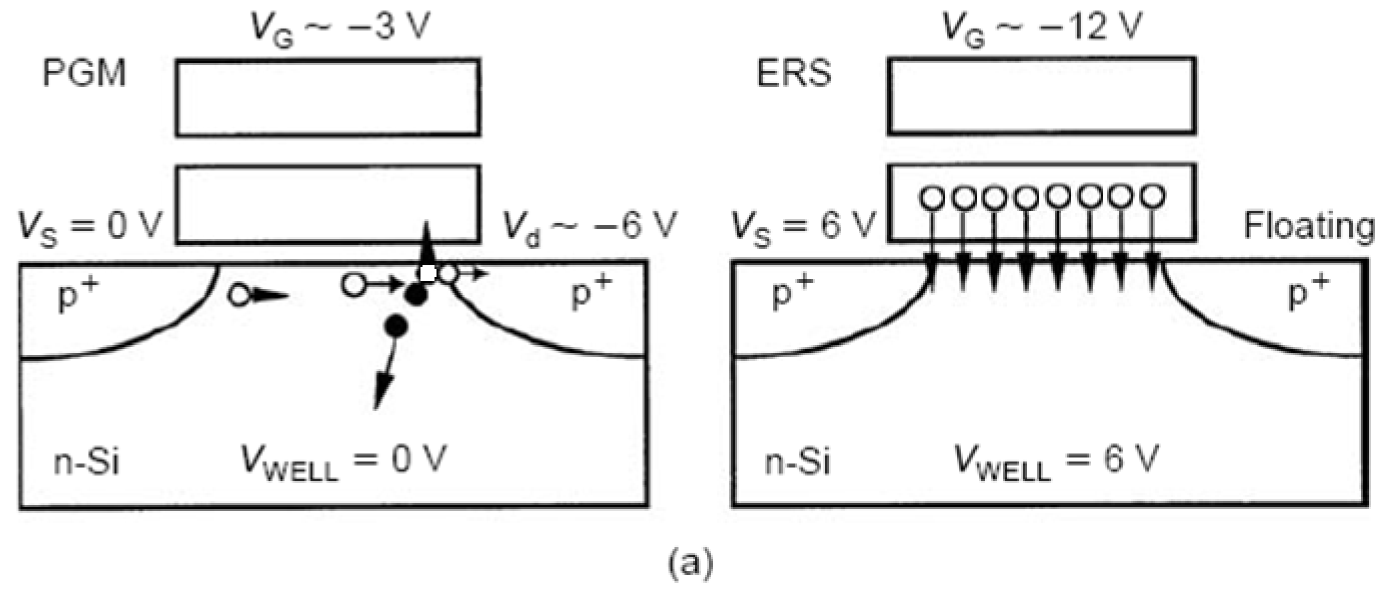


EPROM with UV-irradiation erasing

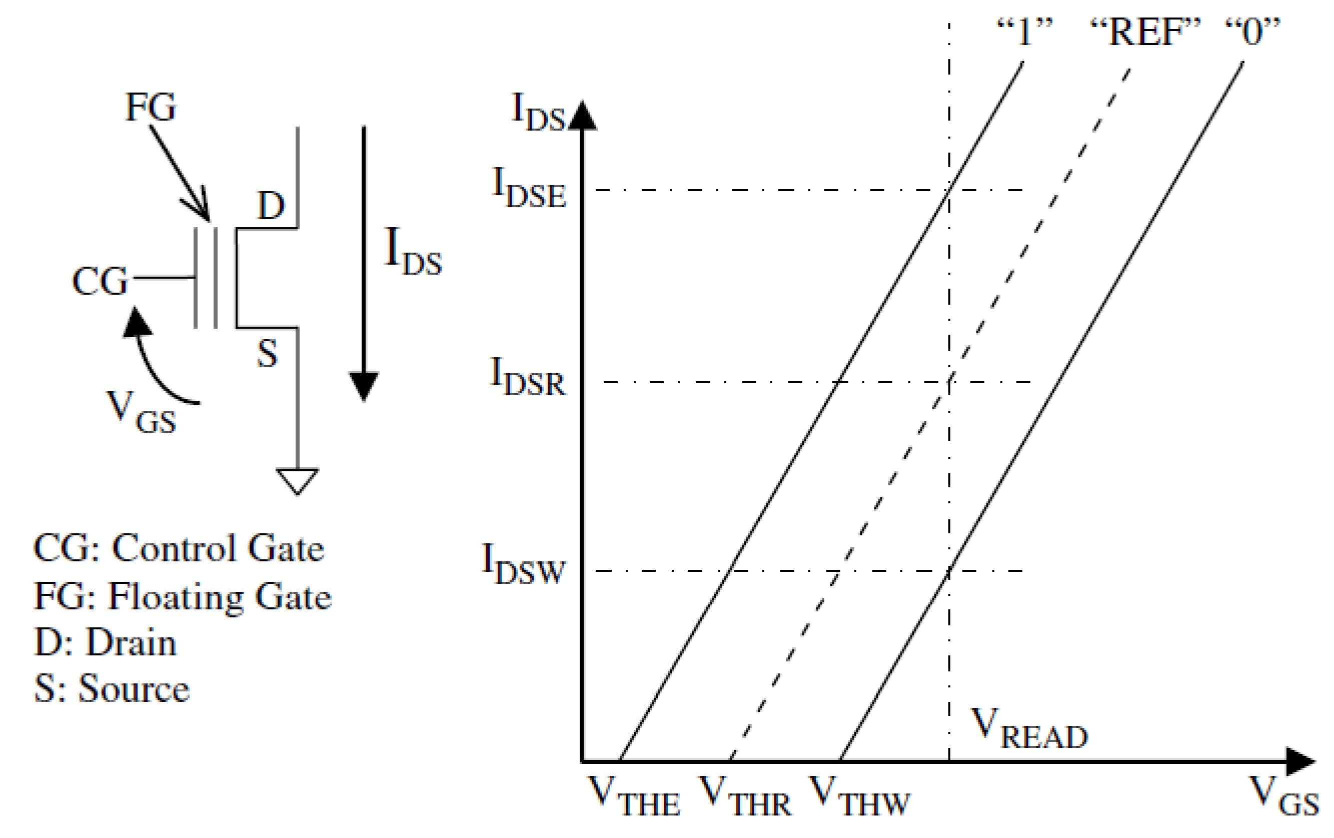
Flash Memories

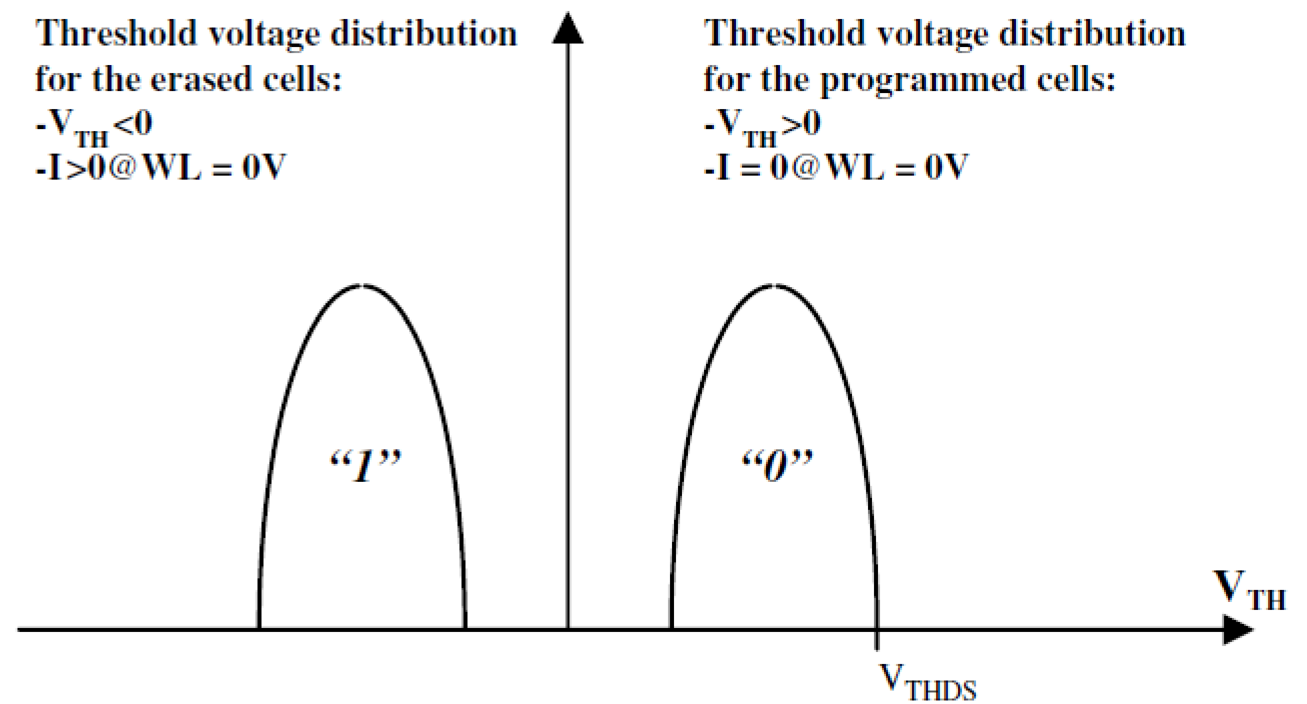


Main applications of flash memories

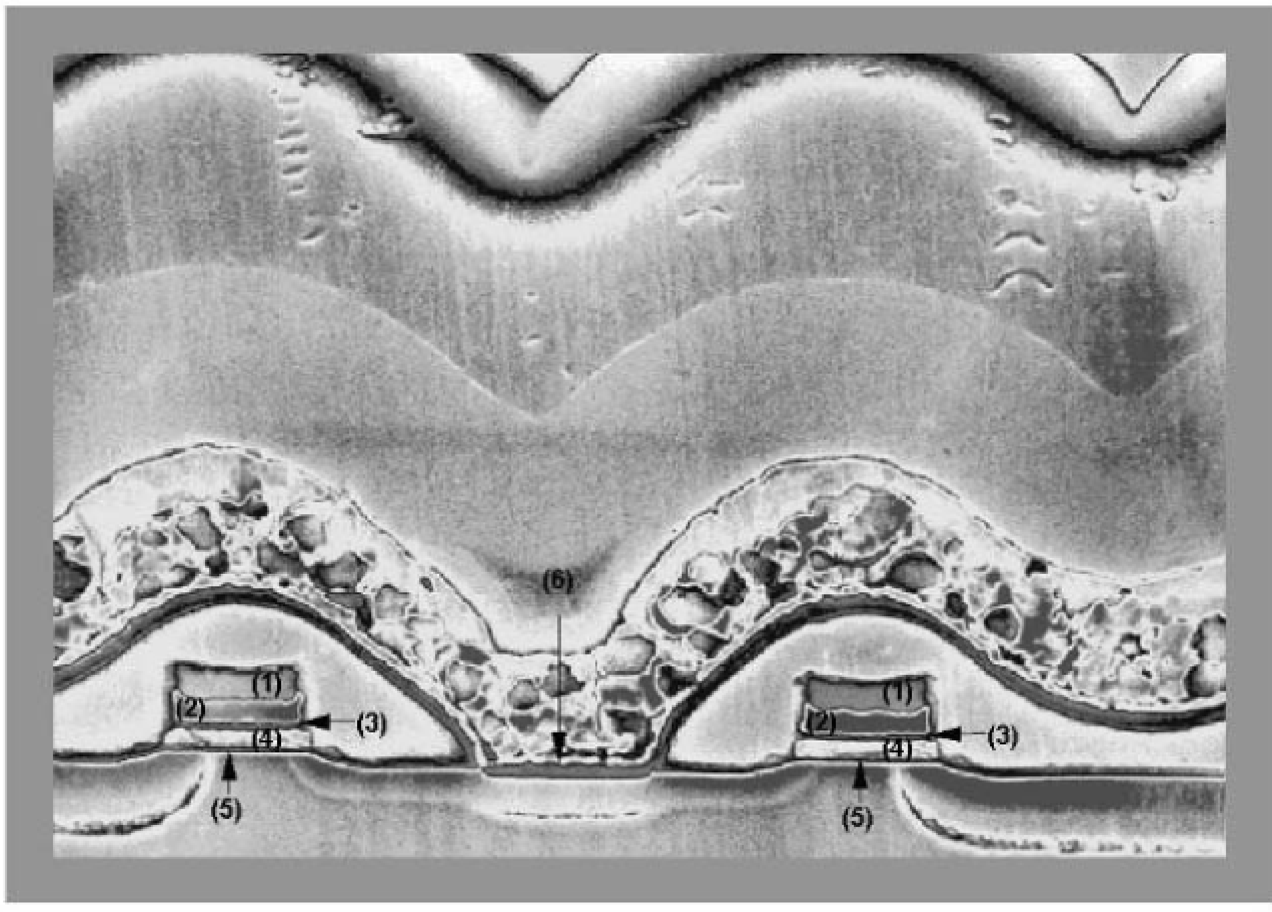


Different p-channel Flash write/erase operations (a) programming with DAHC and erase with FN tunneling action

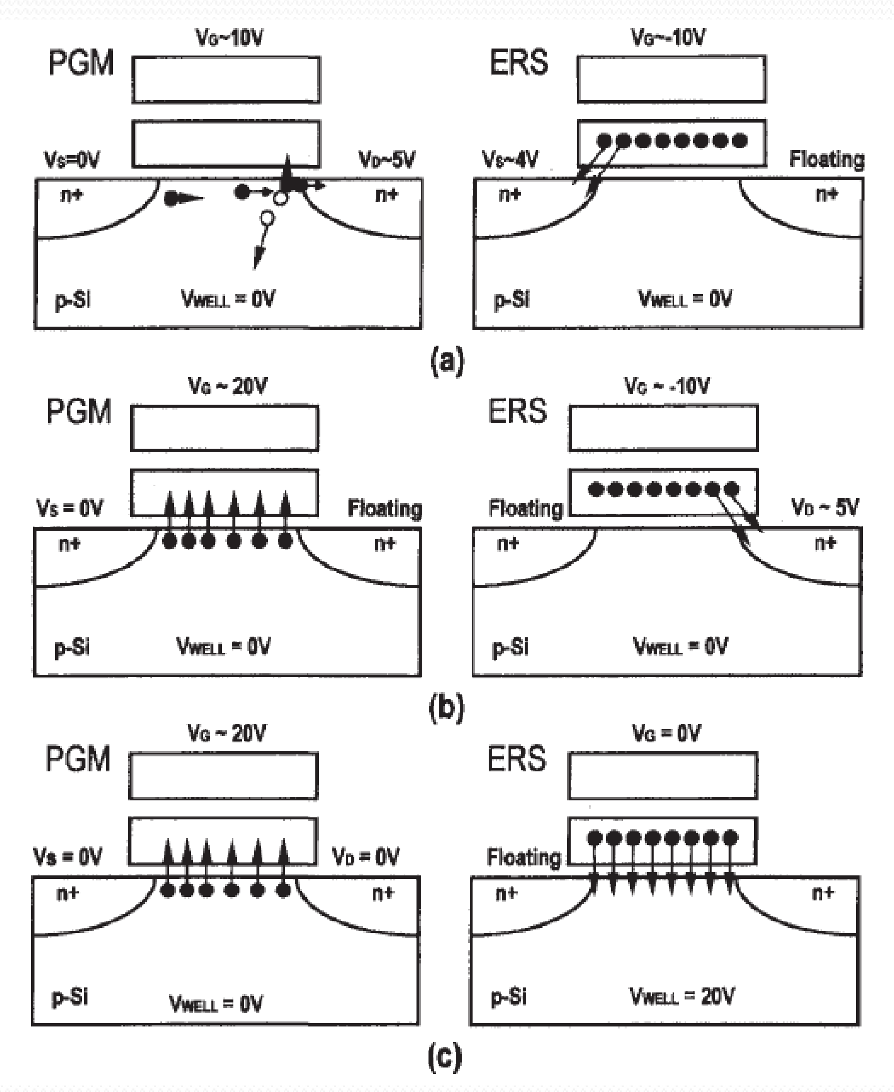

FG
D
CG
S
I_{DS}
V_{GS}
CG: Control Gate
FG: Floating Gate
D: Drain
S: Source
"1"
"REF"
"0"
I_{DS}
I_{DSE}
I_{DSR}
I_{DSW}
V_{READ}
V_{THE}
V_{THR}
V_{THW}
V_{GS}



Threshold voltage distributions for erased and programmed cells



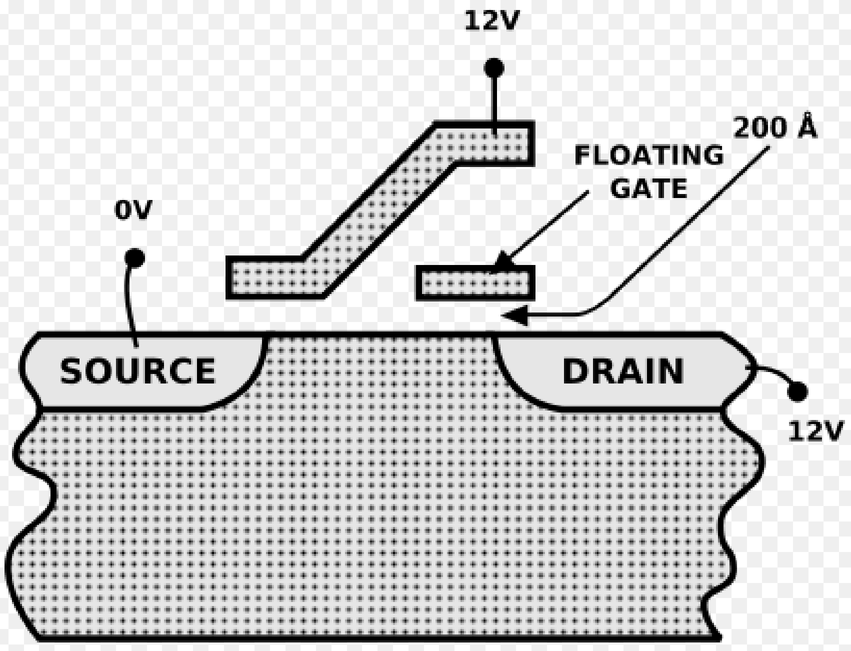
Cross-section of a Flash Memory device; silicide (1) and polysilicon (2) which constitute the control gate, the interpoly oxide layer (3), the floating gate (4), the thin oxide (5), drain contact, shared by the two cells and connected to Metal 1 (the granular wire)(6)) and source implant, deeper than the drain implant



(b) PGM: DAHC
ERS: FN tunneling (source)

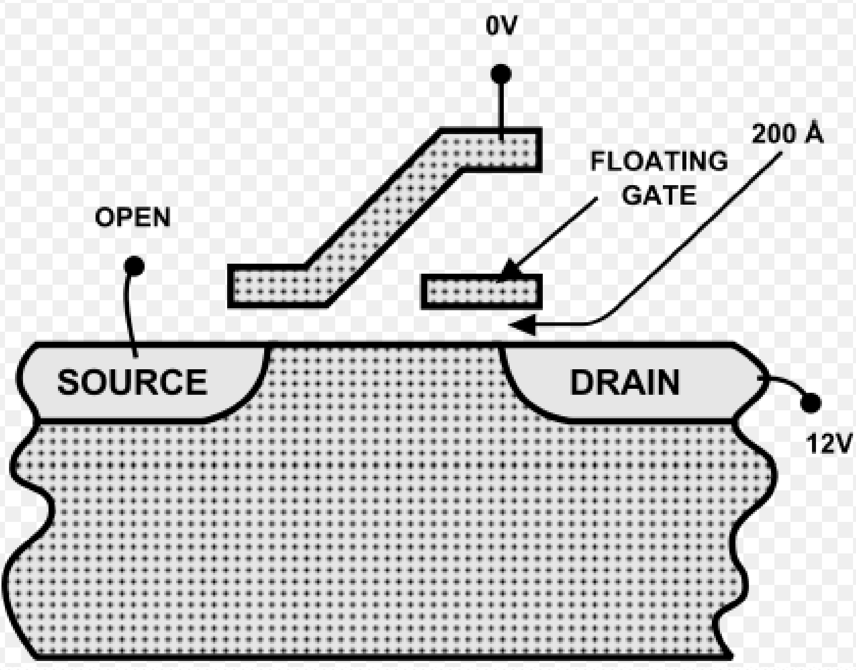
(c) PGM: HCI
ERS: FN tunneling (drain)

(d) PGM: HCI
ERS: FN tunneling

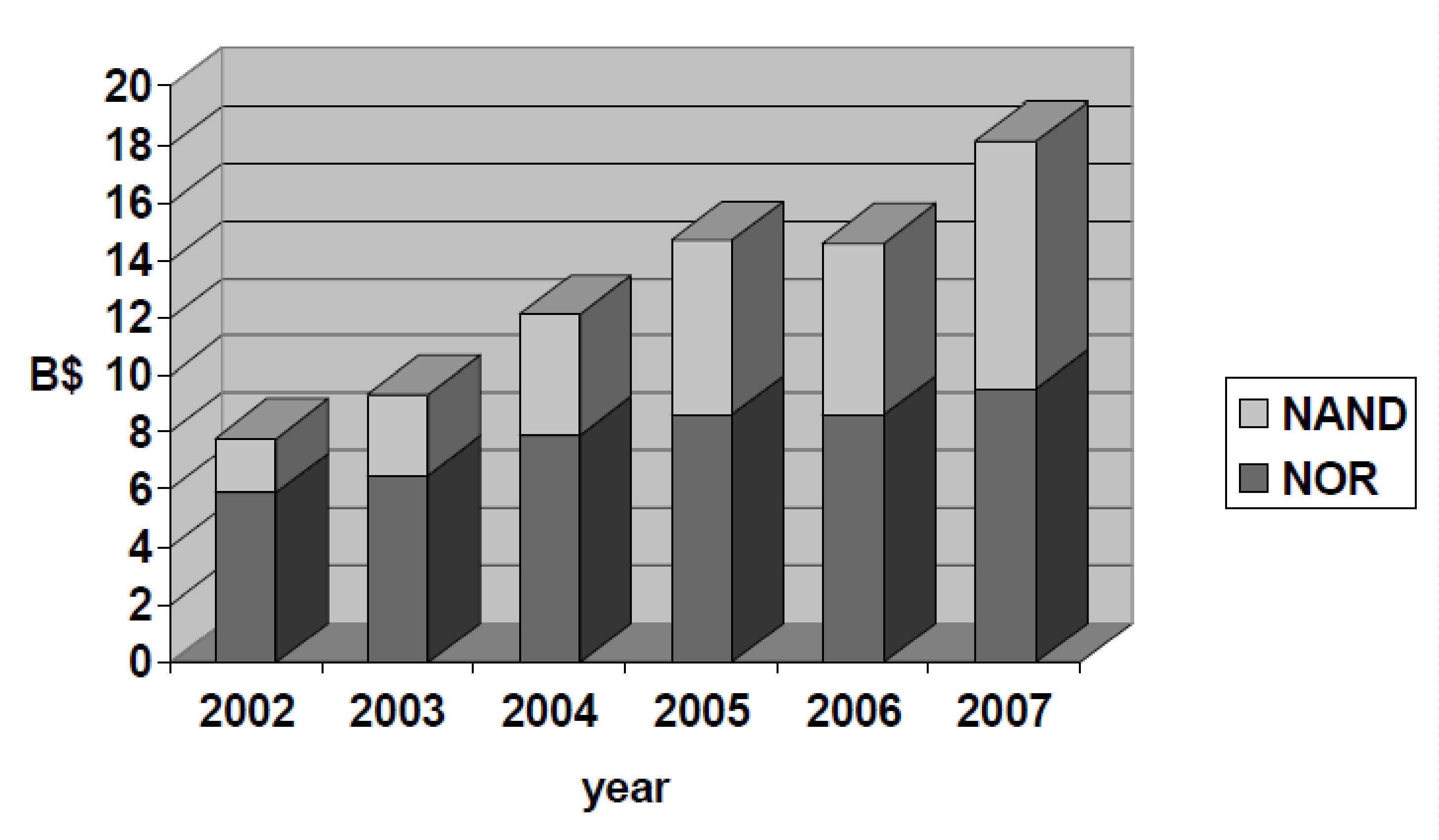


• Default state is “1”, as current will flow through the channel under application of an appropriate voltage to the control gate.

• Set to "0“ by applying > 5V (typically) to the CG. Channel is now turned on, so electrons can flow from the source to the drain. Current is then sufficiently large to cause some high energy (hot) electrons to jump through the insulating layer onto the FG.



• To erase a NOR flash cell (resetting it to the "1" state), a large voltage *of the opposite polarity* is applied between the CG and source, pulling the electrons off the FG through quantum tunneling



Flash memory market sharing by technology (source: Web Feet Inc.)

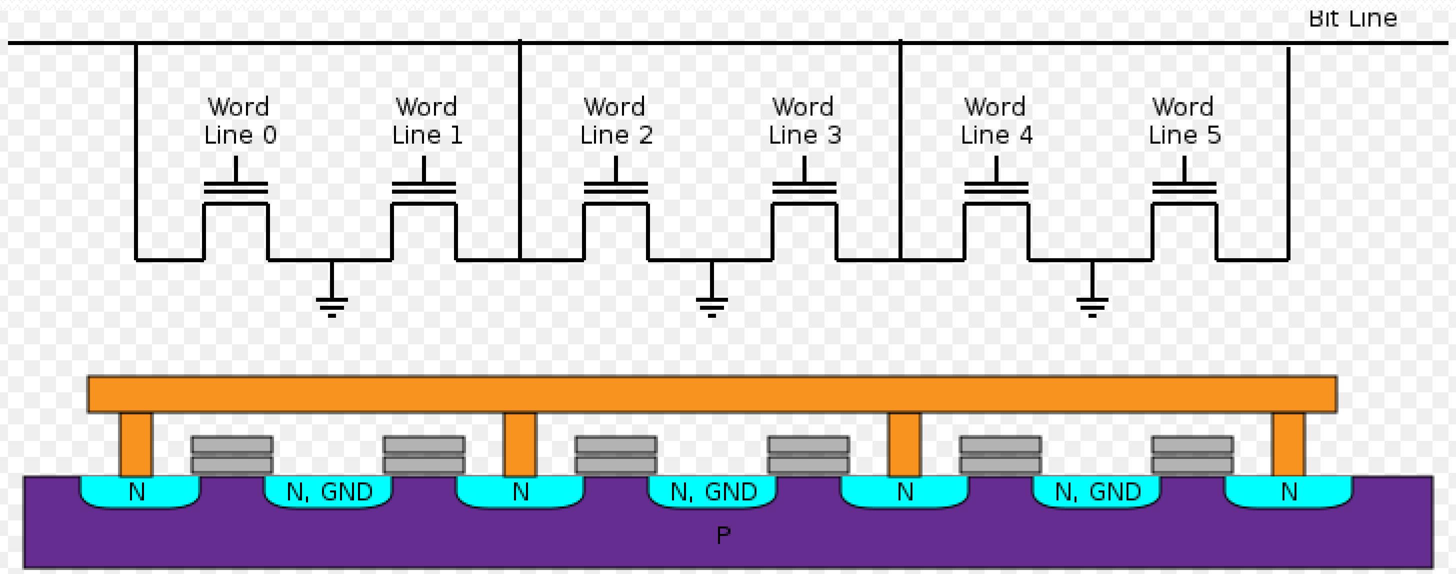
- NAND has significantly higher storage capacity than NOR

- NAND flash has found a market in devices to which large files are frequently uploaded and replaced: MP3 players, digital cameras and USB drives

- NOR flash is faster, but it's also more expensive. NOR is most often used in mobile phones

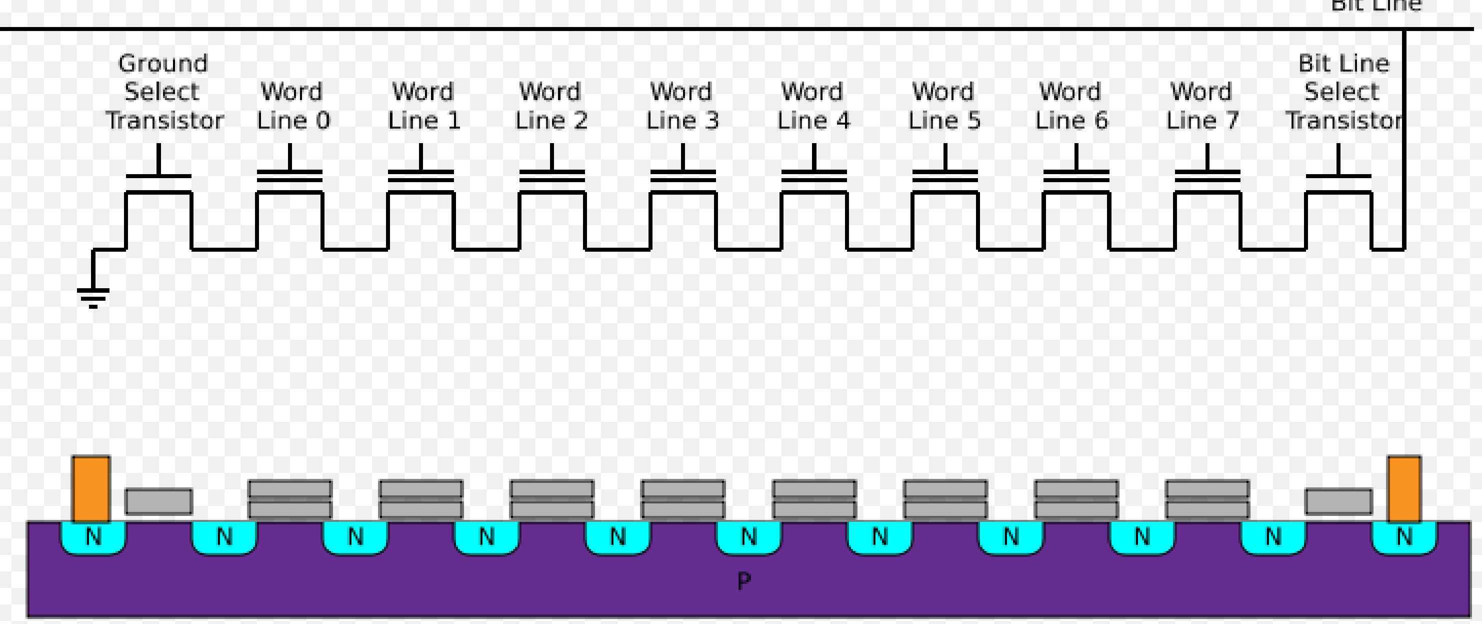
Flash memories

• specific type of EEPROM that is erased and programmed in large blocks



NOR-gate flash (Intel, 1988)

Nor-gate Flash: each cell has one end connected directly to ground, and the other end connected directly to a bit line.



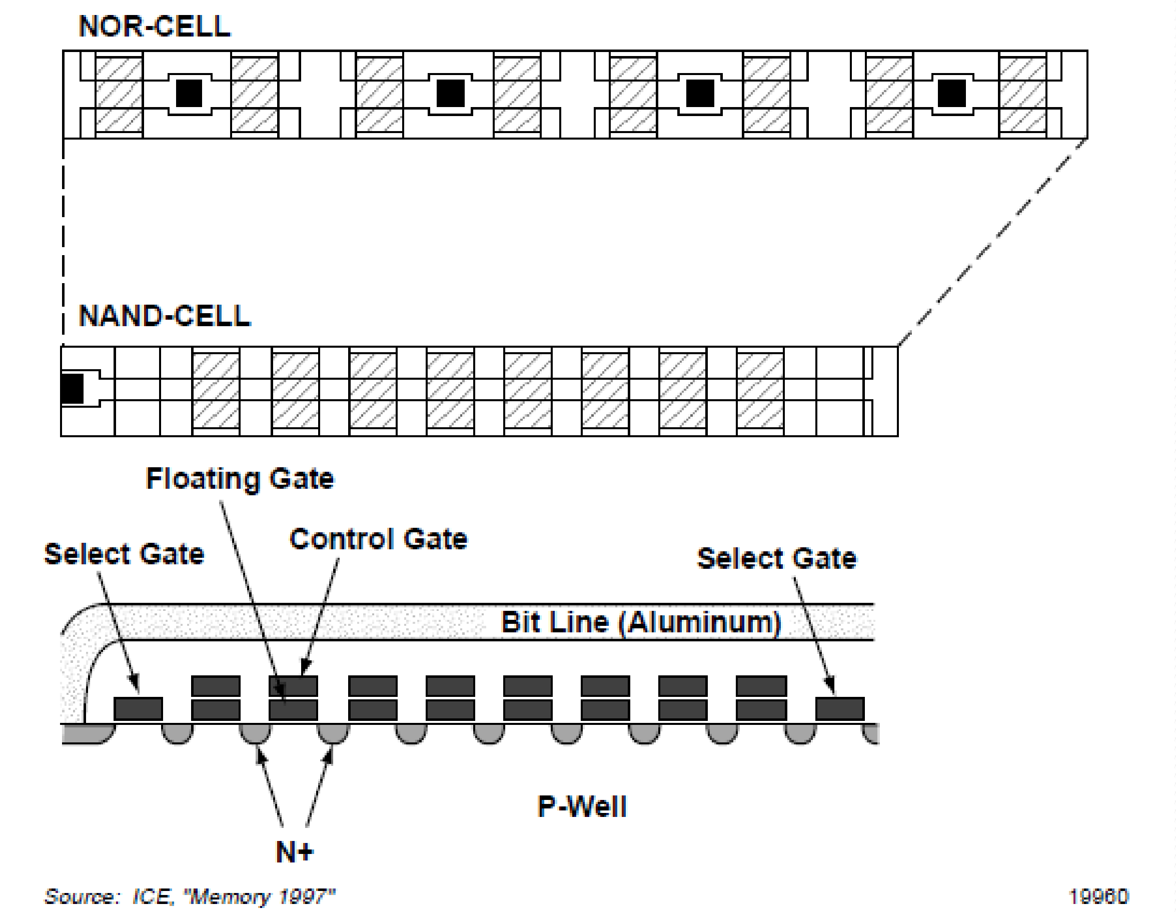
NAND-gate flash (Toshiba 1989)

- several transistors (8, 16, 32, …) are connected in series

When a cell is read, its gate is set to 0V, while the other gates of the stack are biased with a high voltage (typically 4–5 V, say)

⇨ the other gates work as pass-transistors, regardless of their threshold voltage.

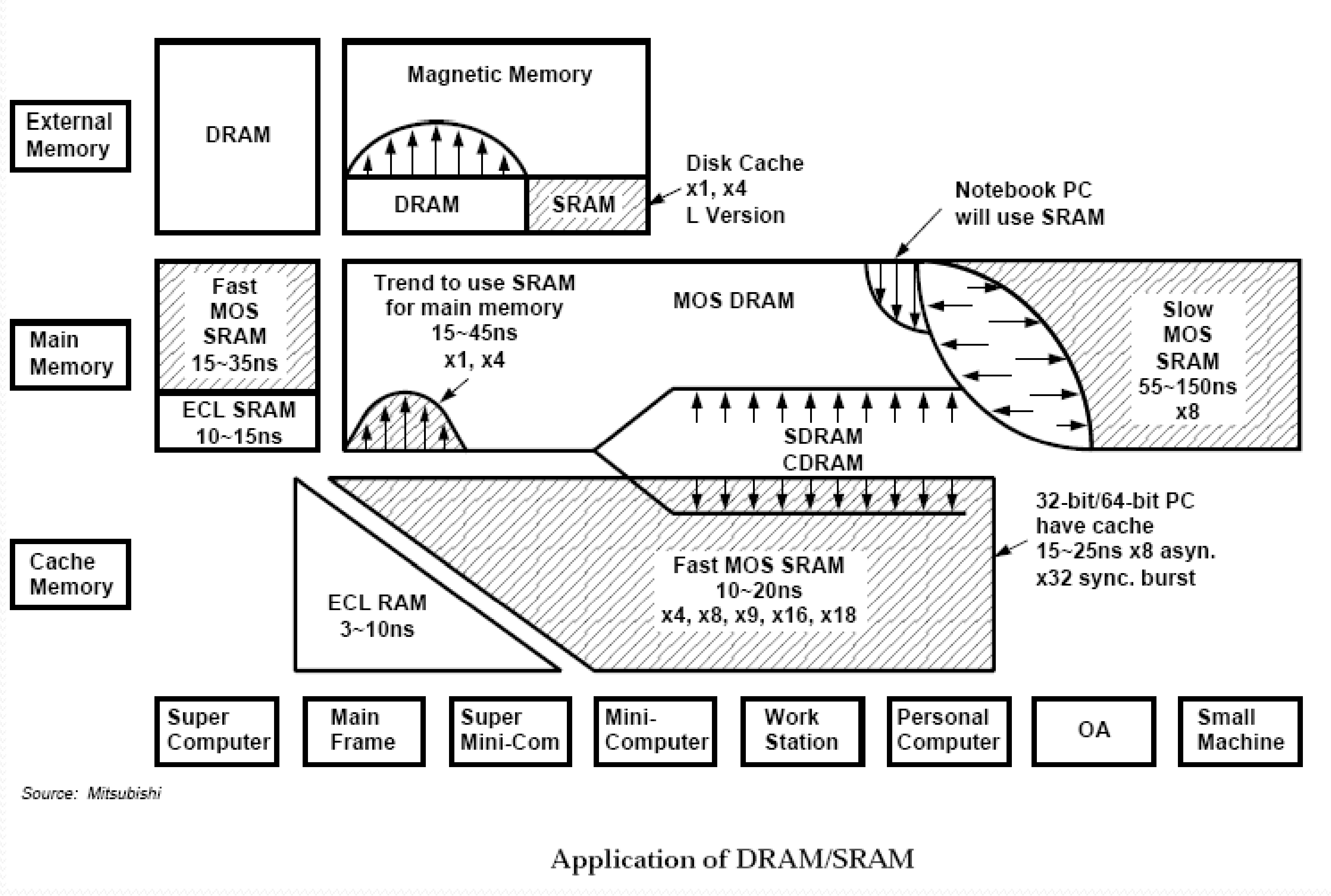
An erased NAND Flash cell has a negative threshold voltage; on the contrary, a programmed cell has a positive threshold voltage but, in any case, less than 4V.


NOR-CELL
NAND-CELL
Floating Gate
Select Gate
Control Gate
Select Gate
Bit Line (Aluminum)
P-Well
N+
Source: ICE, "Memory 1997"
19960

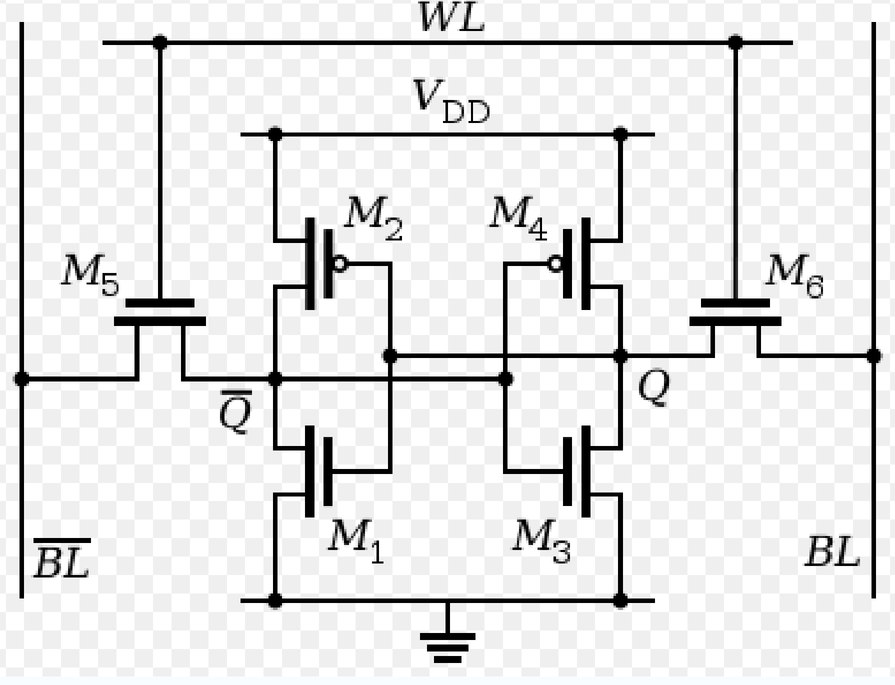
NOR Flash: is a random access device, appropriate for code storage application. It is designed for use in linear program storage for applications such as boot loaders and BIOS, with its key benefit being the ability to satisfy requirements that need to read code wherein each word of data is needed to carry out instructions.

NAND Flash: a sequential access device appropriate for mass storage applications. It is optimized for file structures where each word does not need to be read, but instead provides that sectors of data can be moved to and from media supporting a hard drive like repository structure for data storage to support file systems and allocation tables (FAT)

Static RAM (SRAM)

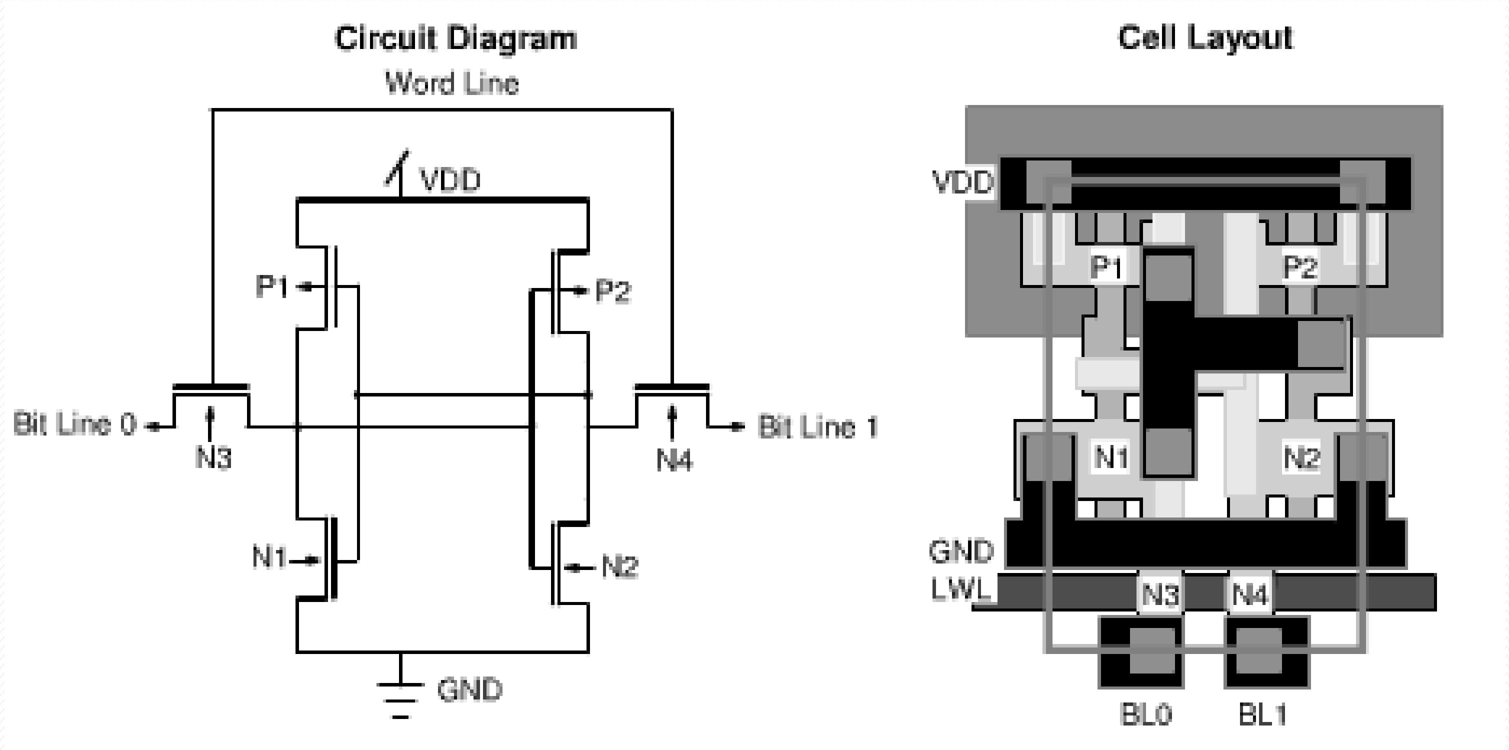


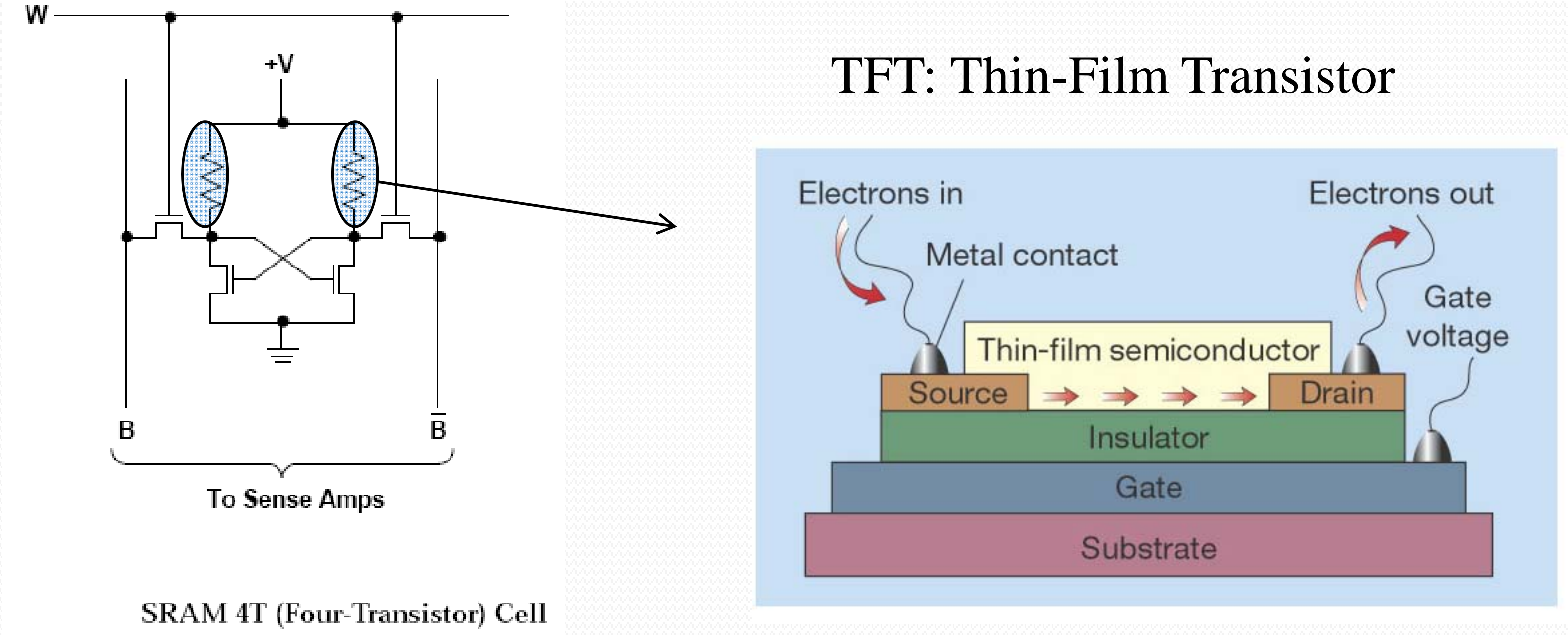
Application of DRAM/SRAM



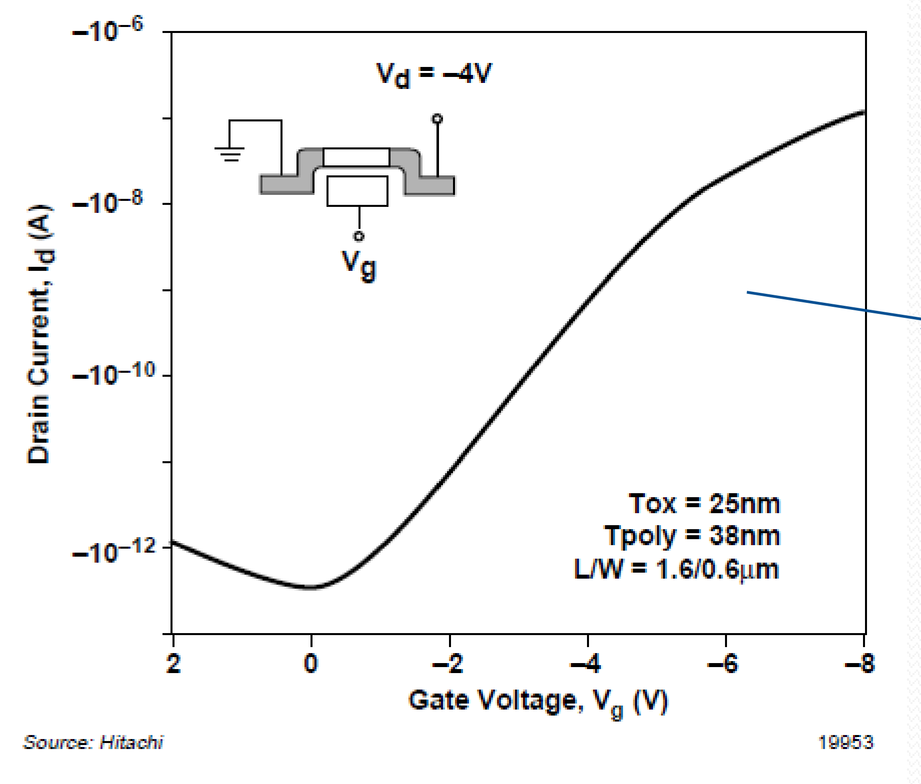
6-transistor SRAM cell

- high speed memory
- data stored as long as power is supplied to the circuit.
- low density ⇒ high cost!!


Circuit Diagram
Word Line
VDD
P1
P2
Bit Line 0
N3
N4
Bit Line 1
N1
N2
GND
Cell Layout
VDD
P1
P2
N1
N2
GND
LWL
N3
N4
BL0
BL1



- Load is formed by using polysilicon as a PMOS device. This PMOS transistor is called a Thin Film Transistor (TFT), and it is formed by depositing several layers of polysilicon above the silicon surface.



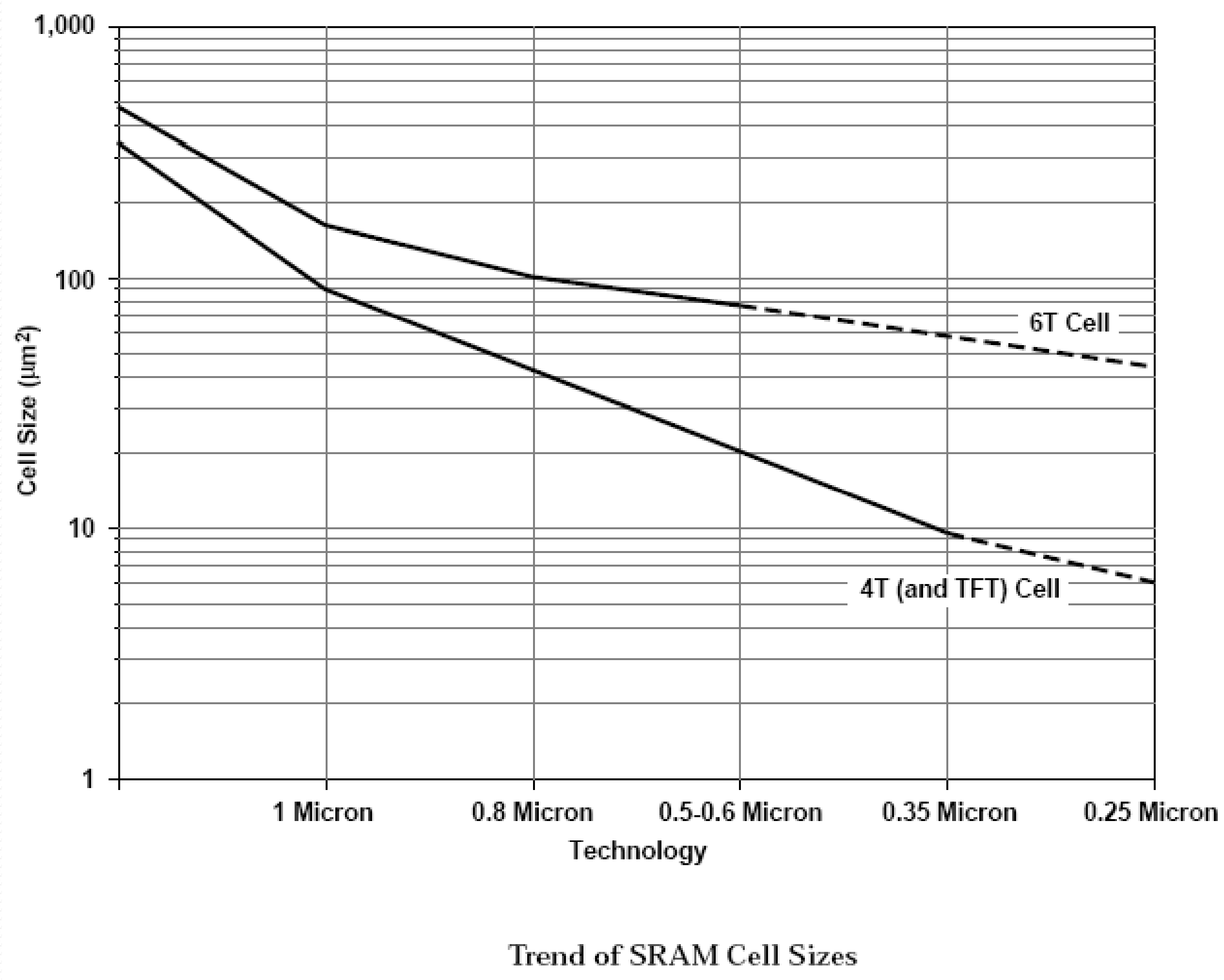
effective resistance ranges from ~ 11 x 10^{13} Ω to 5 x 10^{9} Ω

4T cells have several limitations:

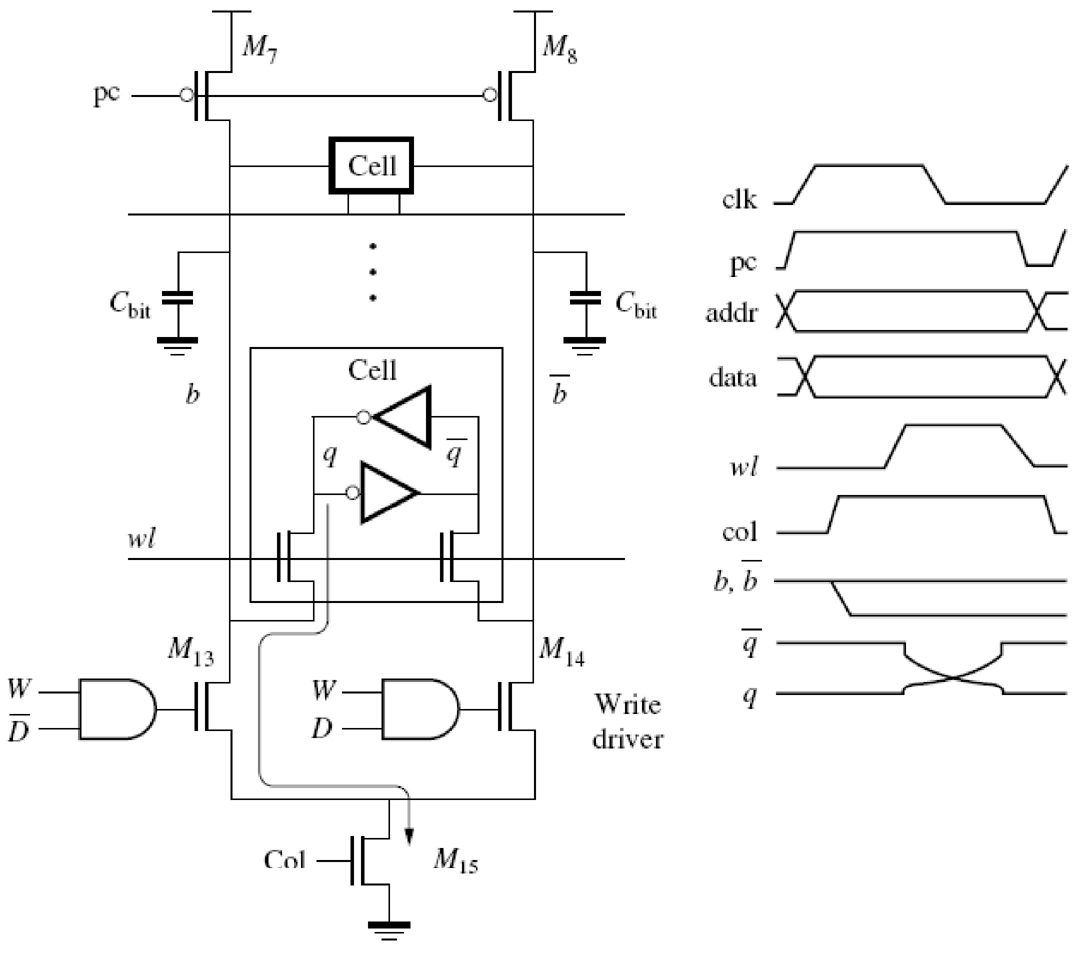
- Each cell has a current flowing in one resistor, i.e., high standby current.
- The cell is sensitive to noise and soft error rate because the resistors are so large.
- The cell is not the fastest cell design available.

	TOSHIBA TC554161FTL-70L 4Mb (x16) 9509	SAMSUNG KM732V588 1Mb (x32) 1995	GALVANTECH GVT7132C32Q7 1Mb (x32) 9524	HITACHI HM67W1664JP-12 1Mb (x16) 9539	NEC D461018LG5-A12 1Mb (x18) 9436	MOTOROLA MCM67C618FN7 1Mb (x18) 9443
Technology	CMOS	BiCMOS	CMOS	BiCMOS	BiCMOS	BiCMOS
Die Size	7.7 x 18.7mm ($144mm^2$)	5 x 6.6mm ($33mm^2$)	4.5 x 6.8mm ($31mm^2$)	6.4 x 10.1mm ($64mm^2$)	5.7 x 11.7mm ($67mm^2$)	9.2 x 11.8mm ($108mm^2$)
Min Gate - (N)	0.65μm	0.5μm	0.4μm	0.45μm	0.6μm	0.6μm
Cell Pitch	3.7 x 6μm	3.0 x 4.75μm	3.4 x 4.9μm	3.3 x 5.7μm	3.5 x 5.5μm	4.9 x 8.2μm
Cell Area	$22\mu m^2$	$14.25\mu m^2$	$16.5\mu m^2$	$19\mu m^2$	$19\mu m^2$	$40\mu m^2$
Cell Type	4T	4T	4T	4T	4T	4T
V_{CC}	5V	3.3V	3.3V	3.3V	3.3V	5V
Access Time	70ns	—	7ns	12ns	12ns	7ns

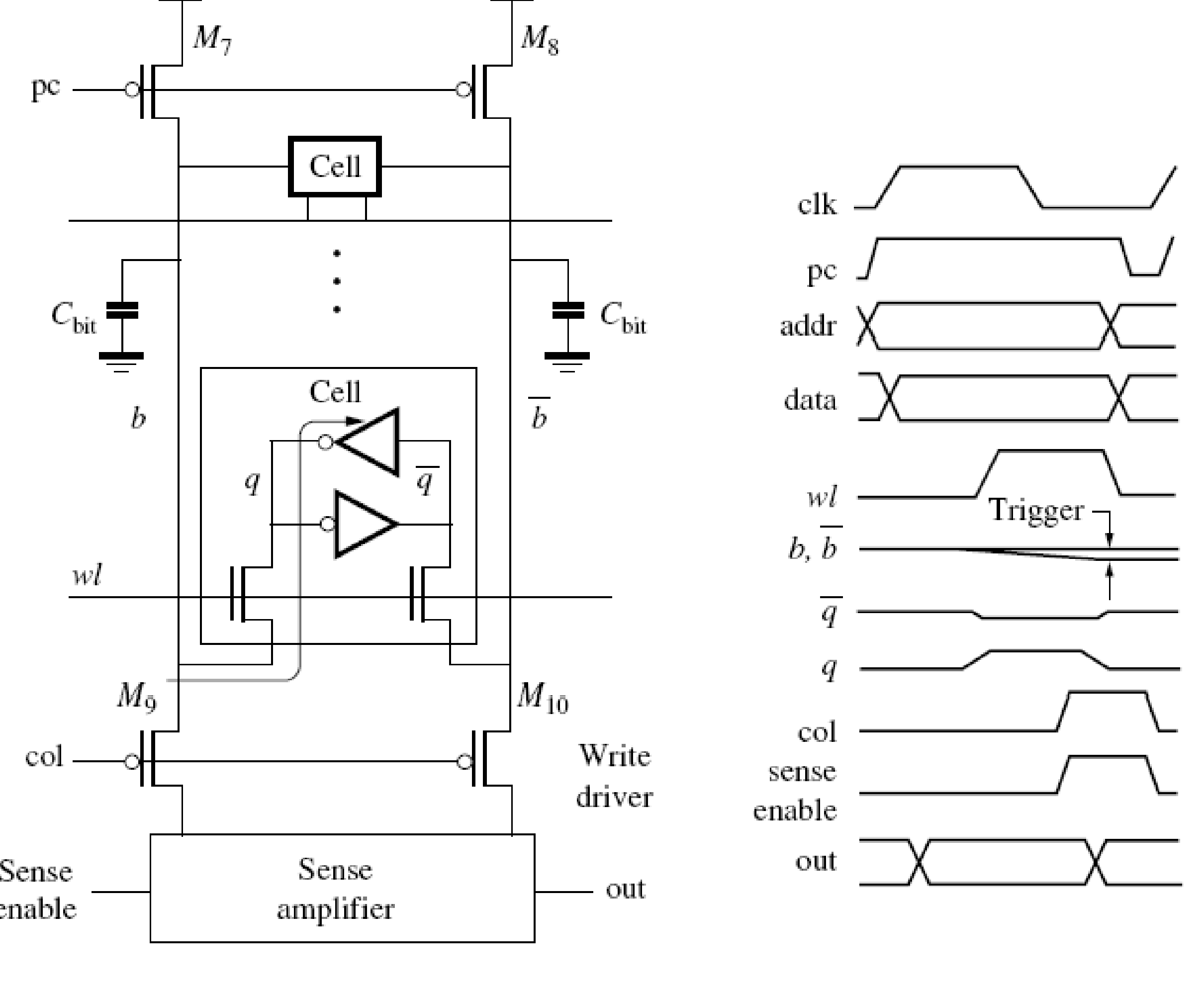
Physical Geometries of SRAMs



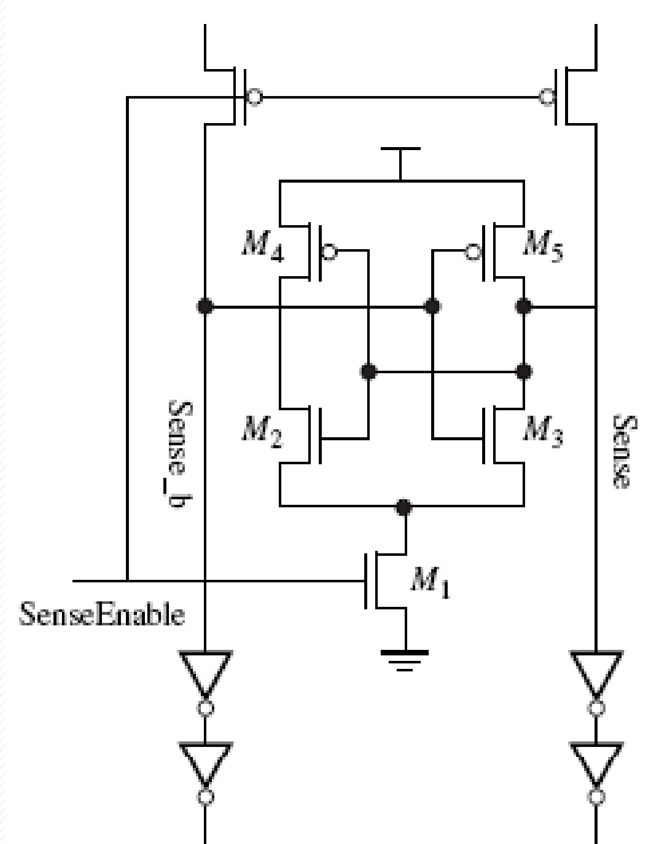
Trend of SRAM Cell Sizes

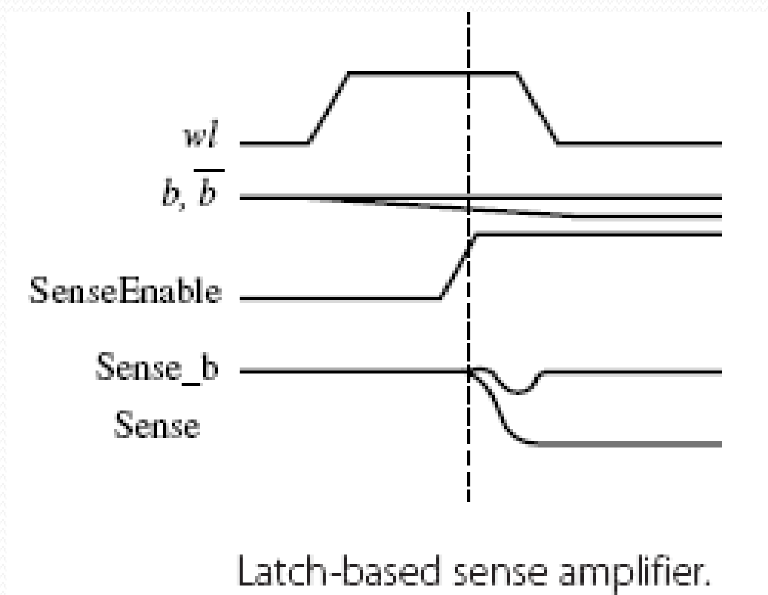


Write driver circuit.



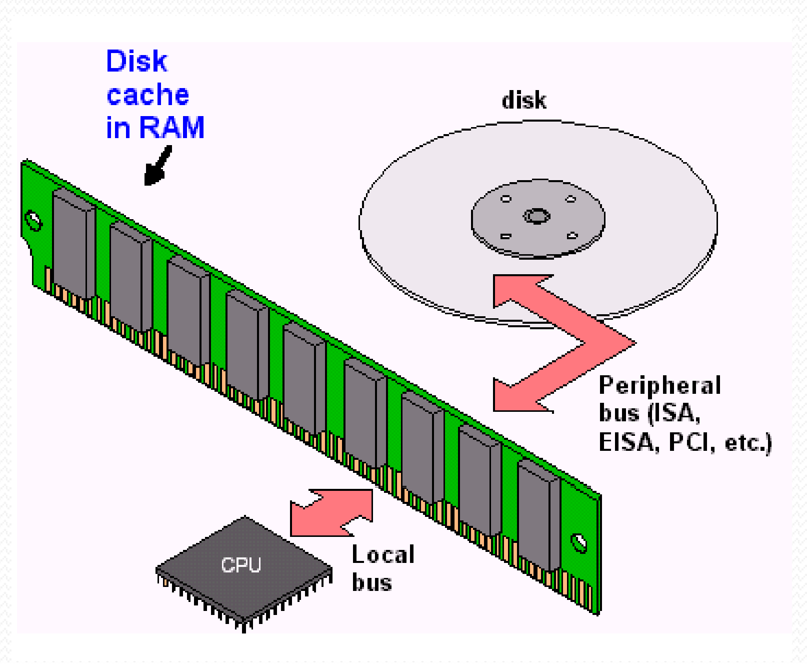
Basic read circuitry.

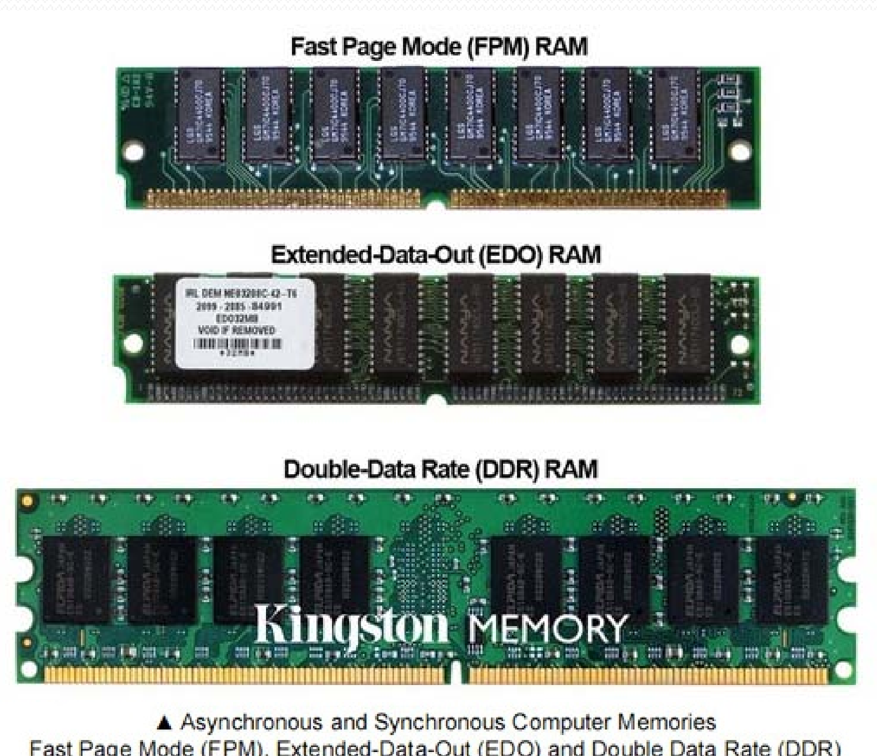




Latch-based sense amplifier.

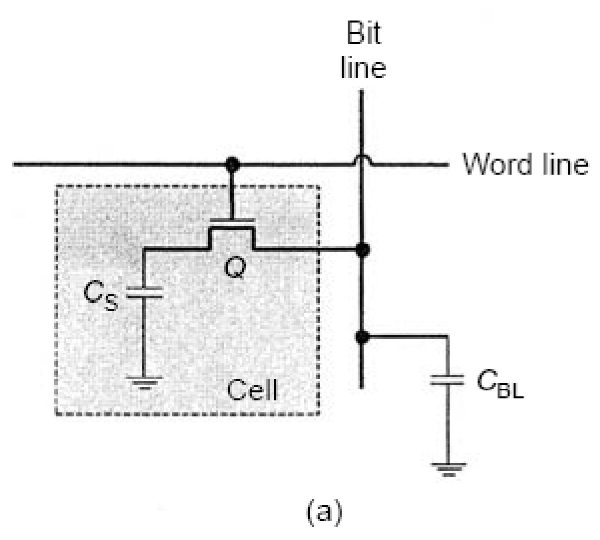
Dynamic RAM (DRAM)



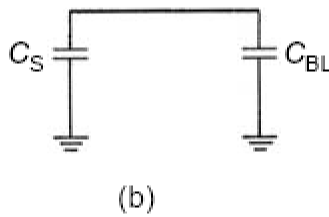


▲ Asynchronous and Synchronous Computer Memories
Fast Page Mode (FPM), Extended-Data-Out (EDO) and Double Data Rate (DDR)

DYNAMIC RAM (DRAM) MEMORY TECHNOLOGIES					
Type	Year of Intro.	Maximum Clock Rate	Bus Width	Peak Bandwidth	Volts
FPM	1990	25MHz	64 bits	200 MBps	5v
EDO	1994	40MHz	64 bits	320 MBps	5v
SDRAM	1996	133MHz	64 bits	1.1 GBps	3.3v
RDRAM	1998	400MHz (x2)	16 bits	800 MBps	2.5v
DDR SDRAM	2000	266MHz (x2)	64 bits	4.2 GBps	2.5v
DDR2 SDRAM	2003	533MHz (x2)	64 bits	8.5 GBps	1.8v
DDR3 SDRAM	2007	800MHz (x2)	64 bits	12.8 GBps	1.5v

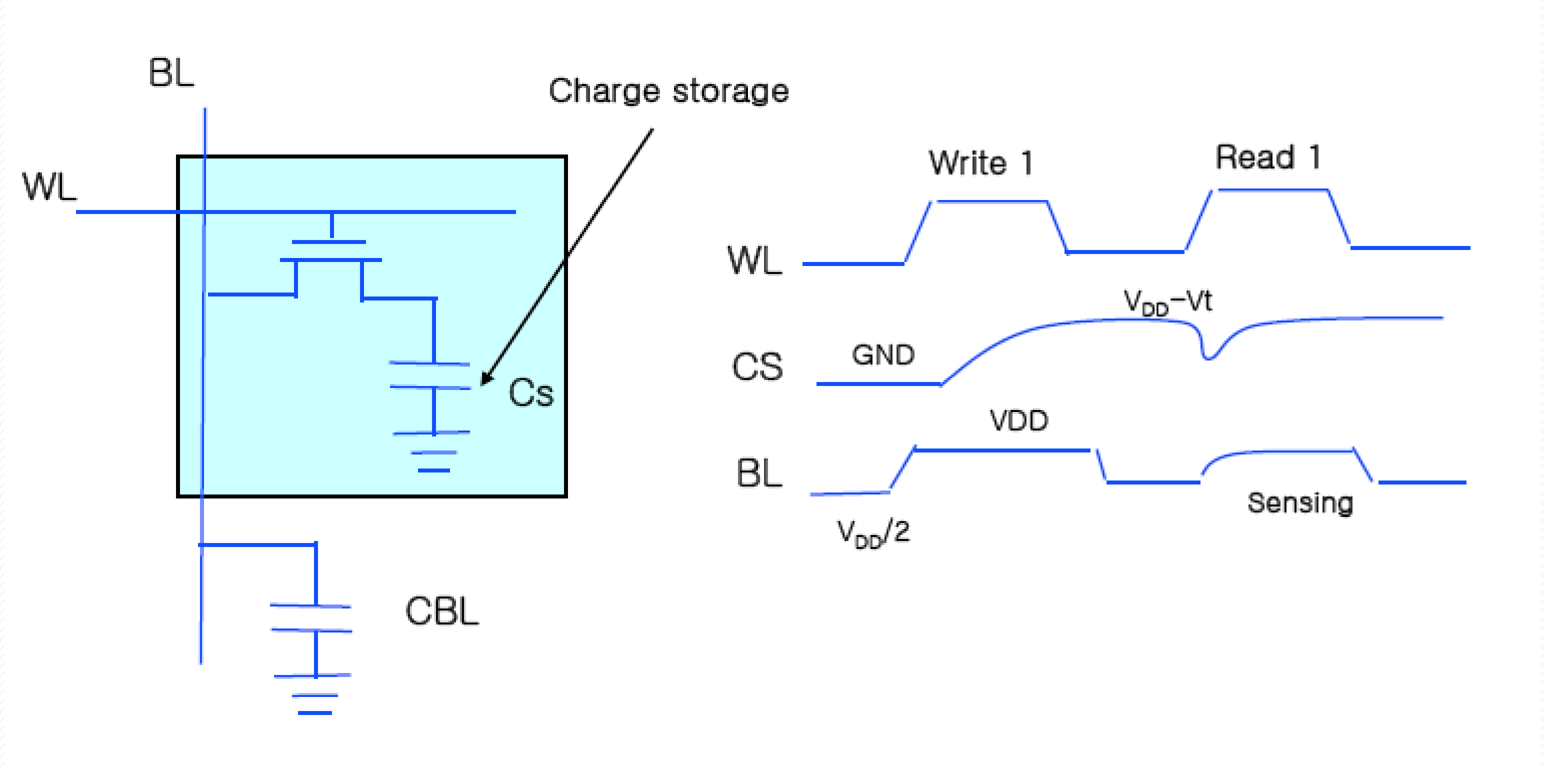


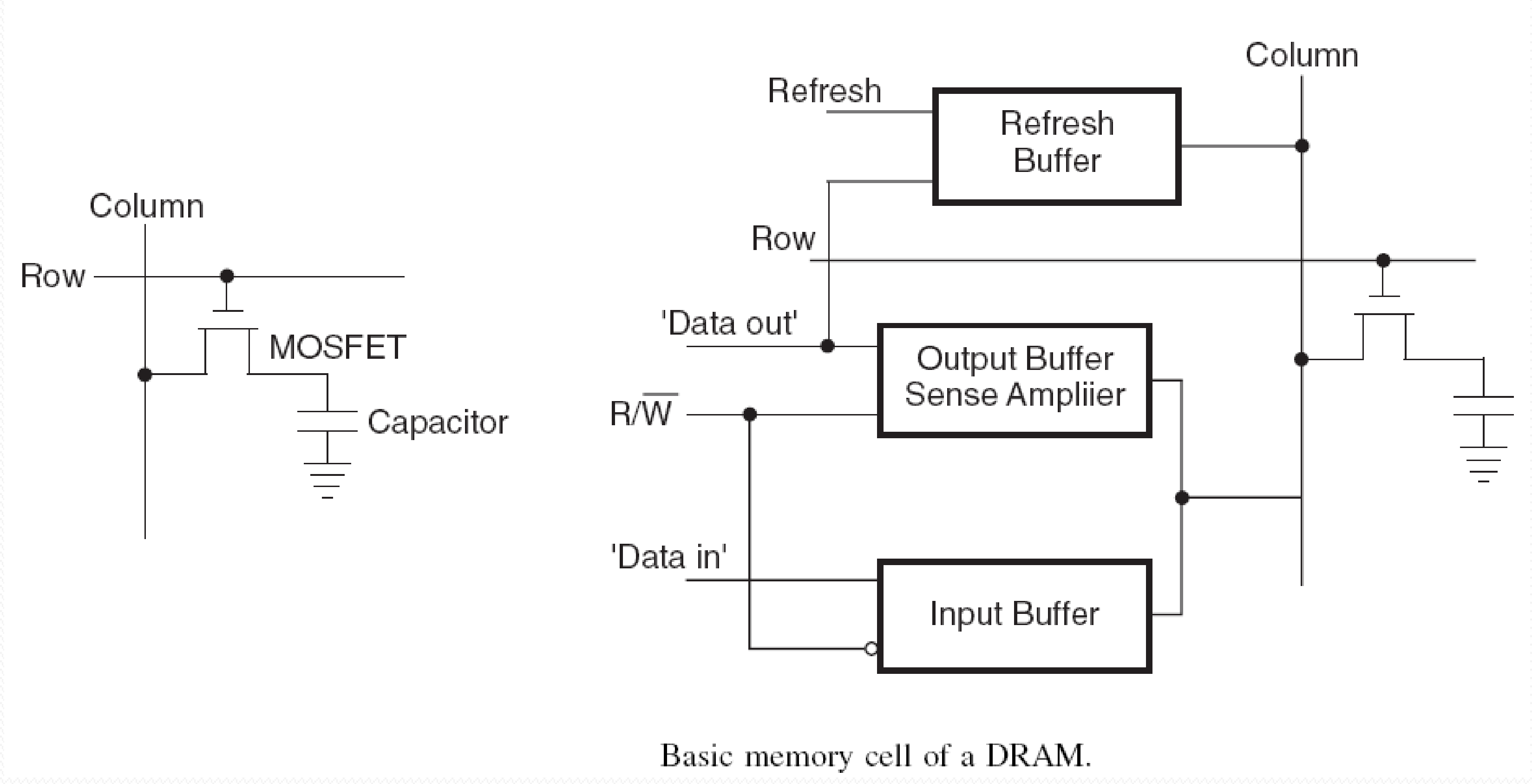
CS: typically, 30-50fF
"1": VCS = VDD – VTH
"0": VCS = 0V



(a) The one-transistor DRAM cell; and (b) during the READ operation, the voltage of the selected word-line is high, thus connecting the storage capacitor C_s to the bit-line capacitance C_{BL}.

READ: the voltage of the selected word-line is high;
readout process is destructive !!!
write process should occur after reading.


BL
Charge storage
WL
Cs
CBL
Write 1
Read 1
WL
VDD-Vt
CS
GND
VDD
BL
Sensing
VDD/2



Basic memory cell of a DRAM.

• bit-line voltage variation

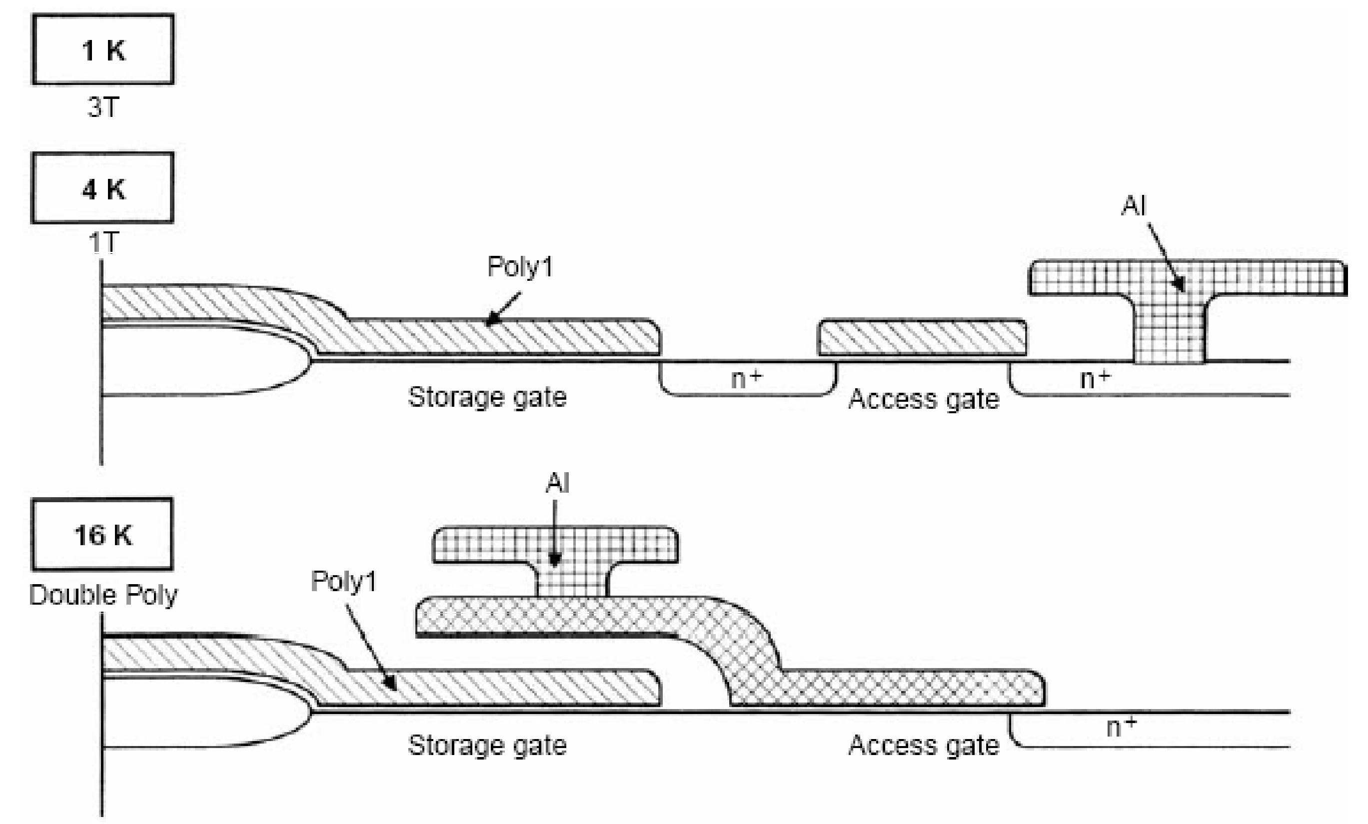
$$V_s = \Delta V_{BL} = \frac{C_S}{C_{BL} + C_S}\left(V_{cs} - \frac{V_{DD}}{2}\right)$$

$R = C_{BL}/C_s$

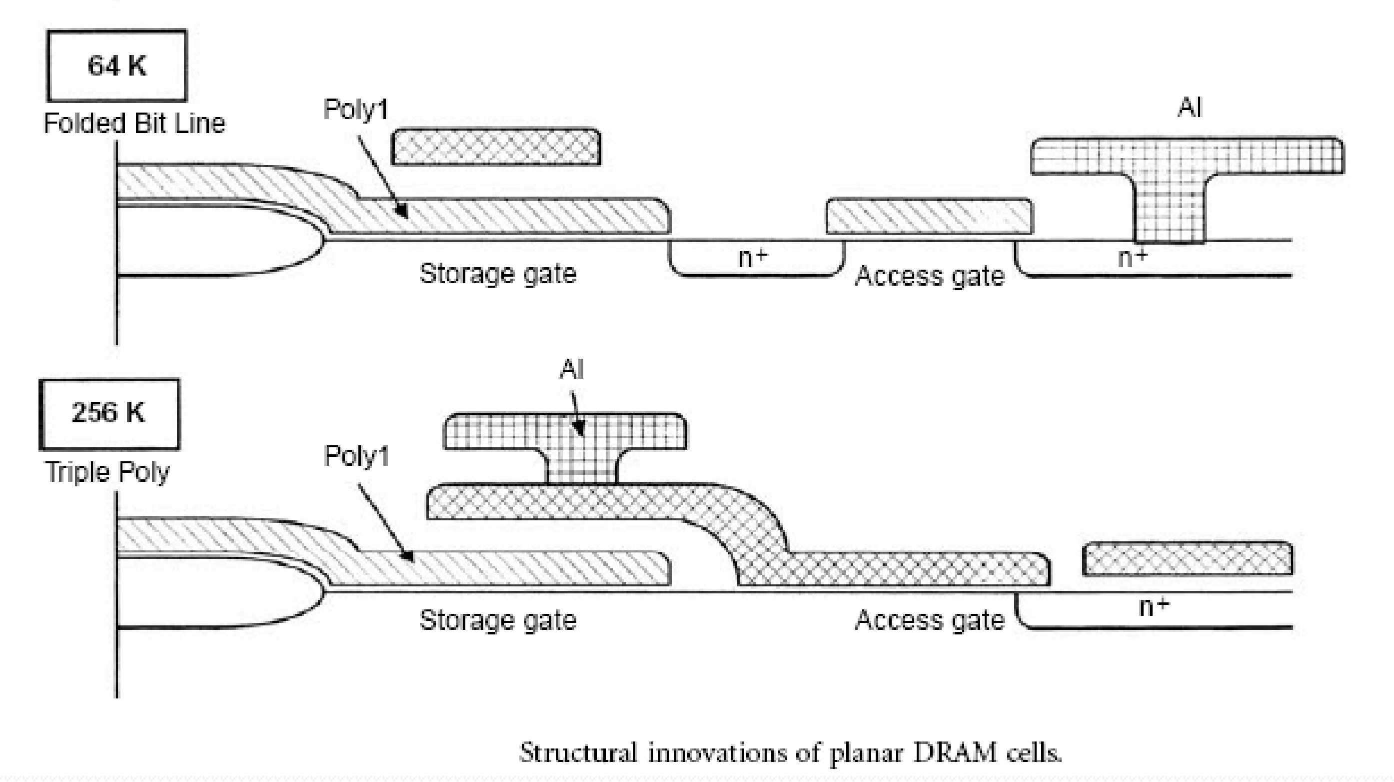
$$\Delta V\,(1) = \frac{1}{1+R}\left(\frac{V_{DD}}{2} - V_t\right)$$

$$\Delta V\,(0) = \frac{1}{1+R}\left(\frac{V_{DD}}{2}\right)$$

Since ratio $R = C_{BL}/C_s$ is large, these readout bit-line sense signals $\Delta V(1)$ and $\Delta V(0)$ are very small. Typical values for the sense signal are about 100 mv.

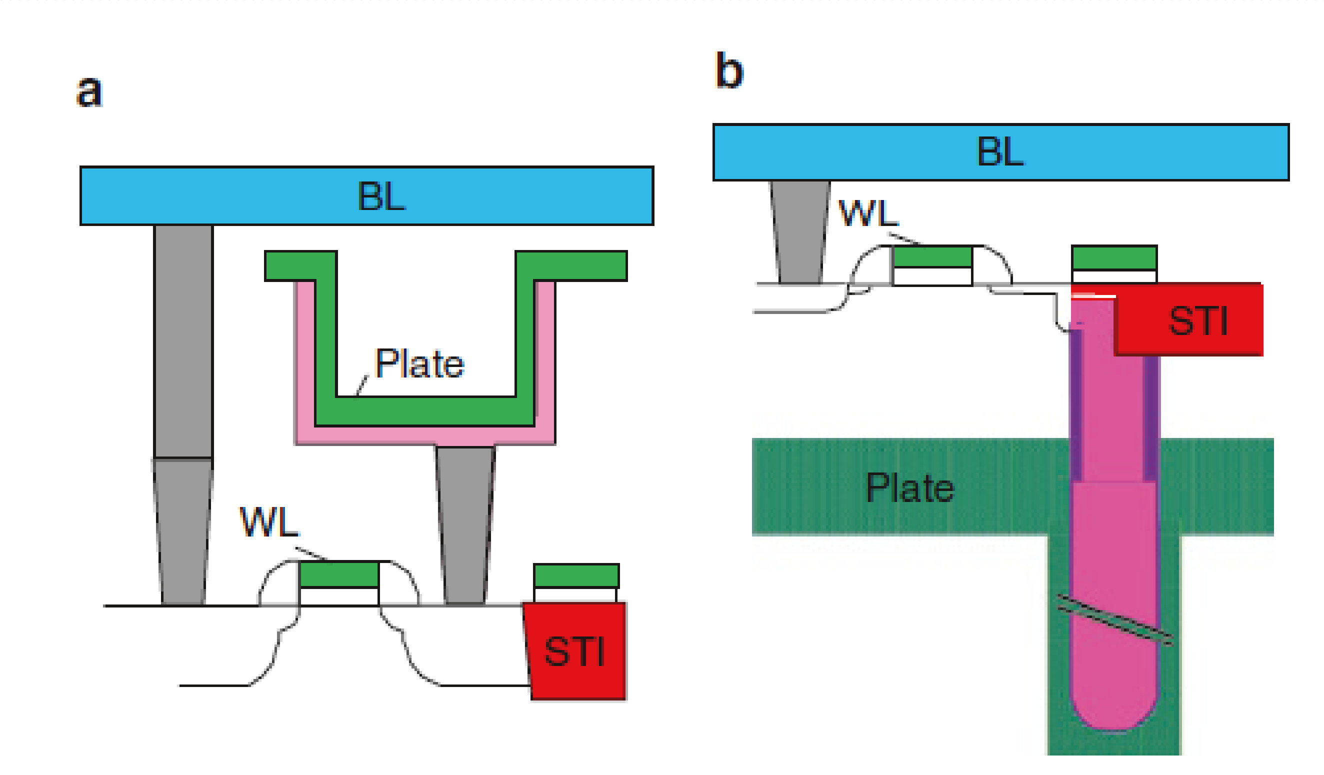


Structural innovations of planar DRAM cells.

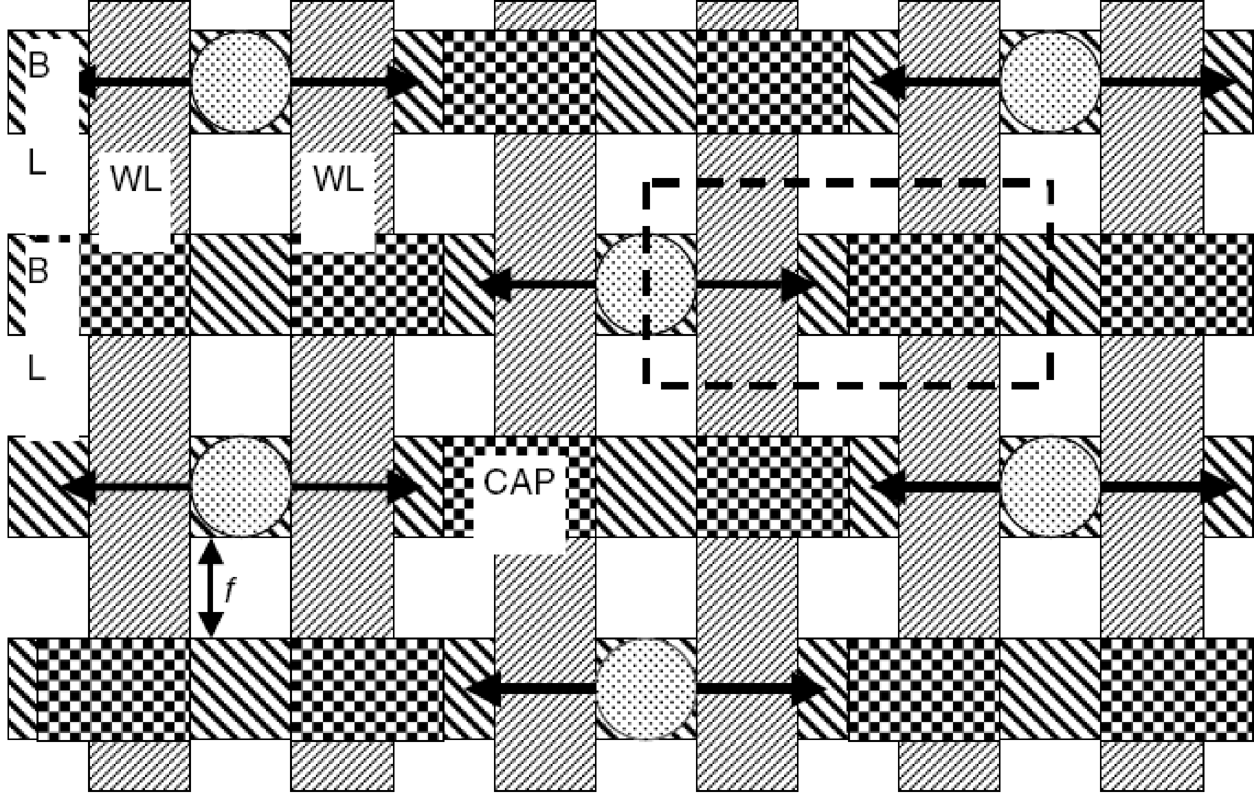


Structural innovations of planar DRAM cells.

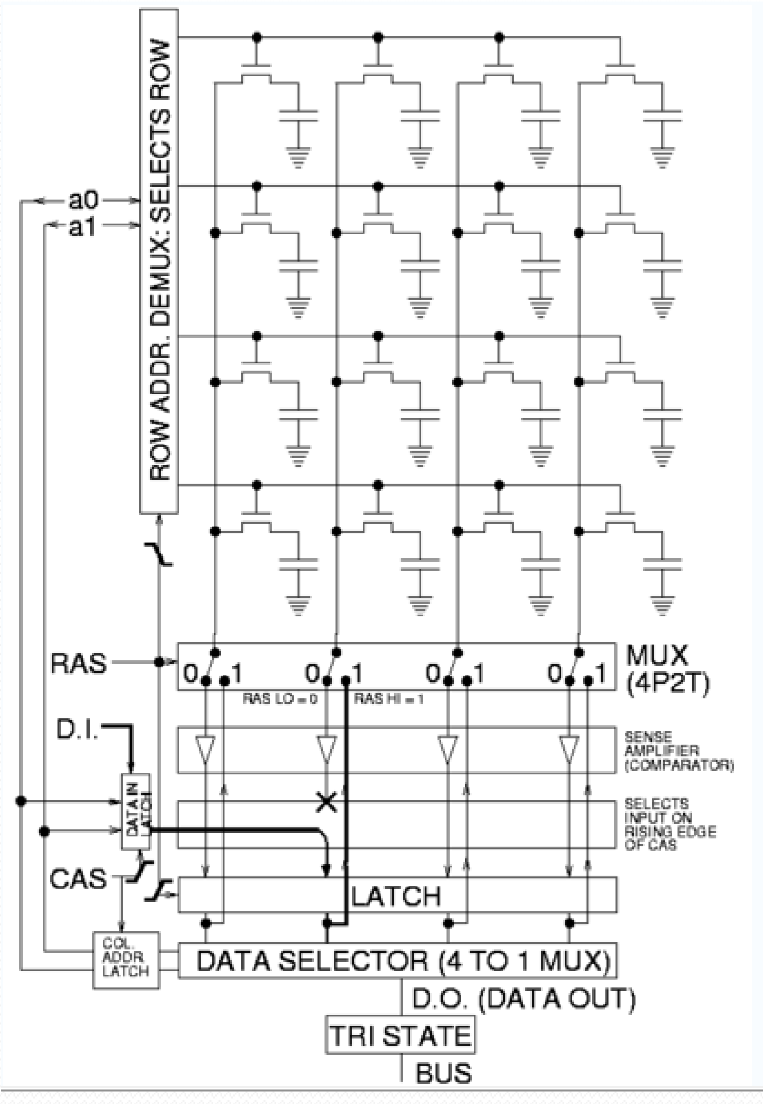
1-Mb DRAM.



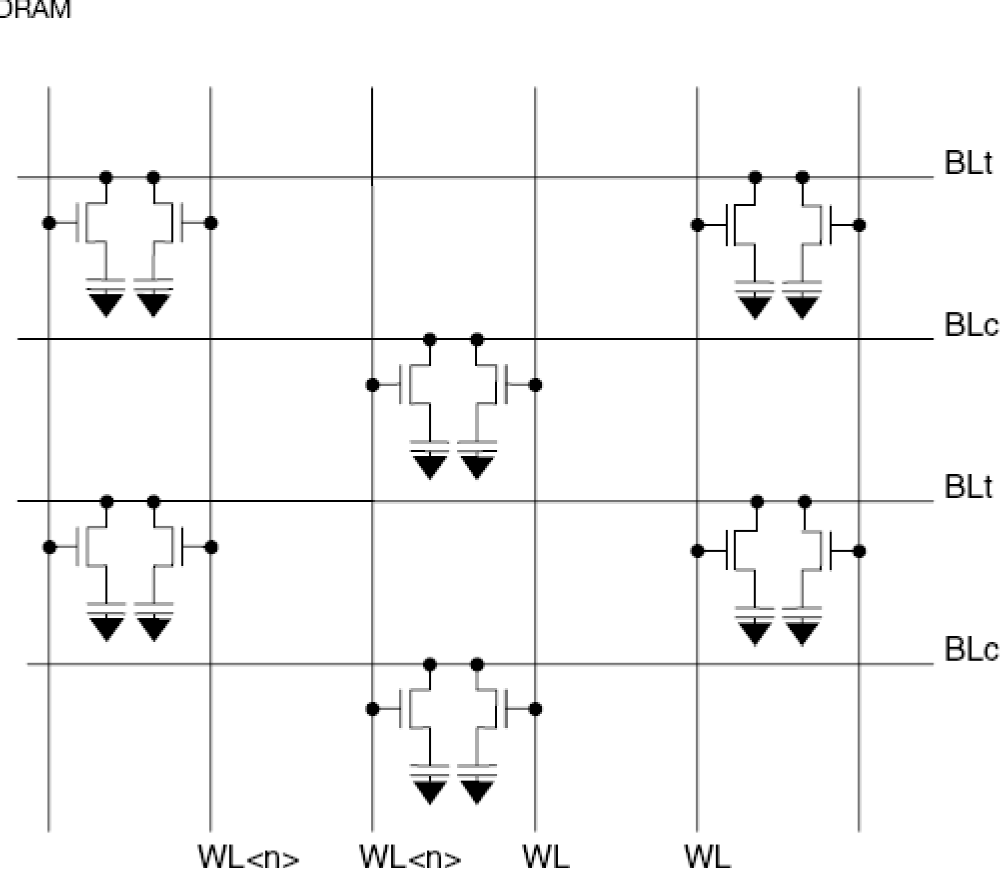
Cells: (a) stacked capacitor and (b) trench capacitor



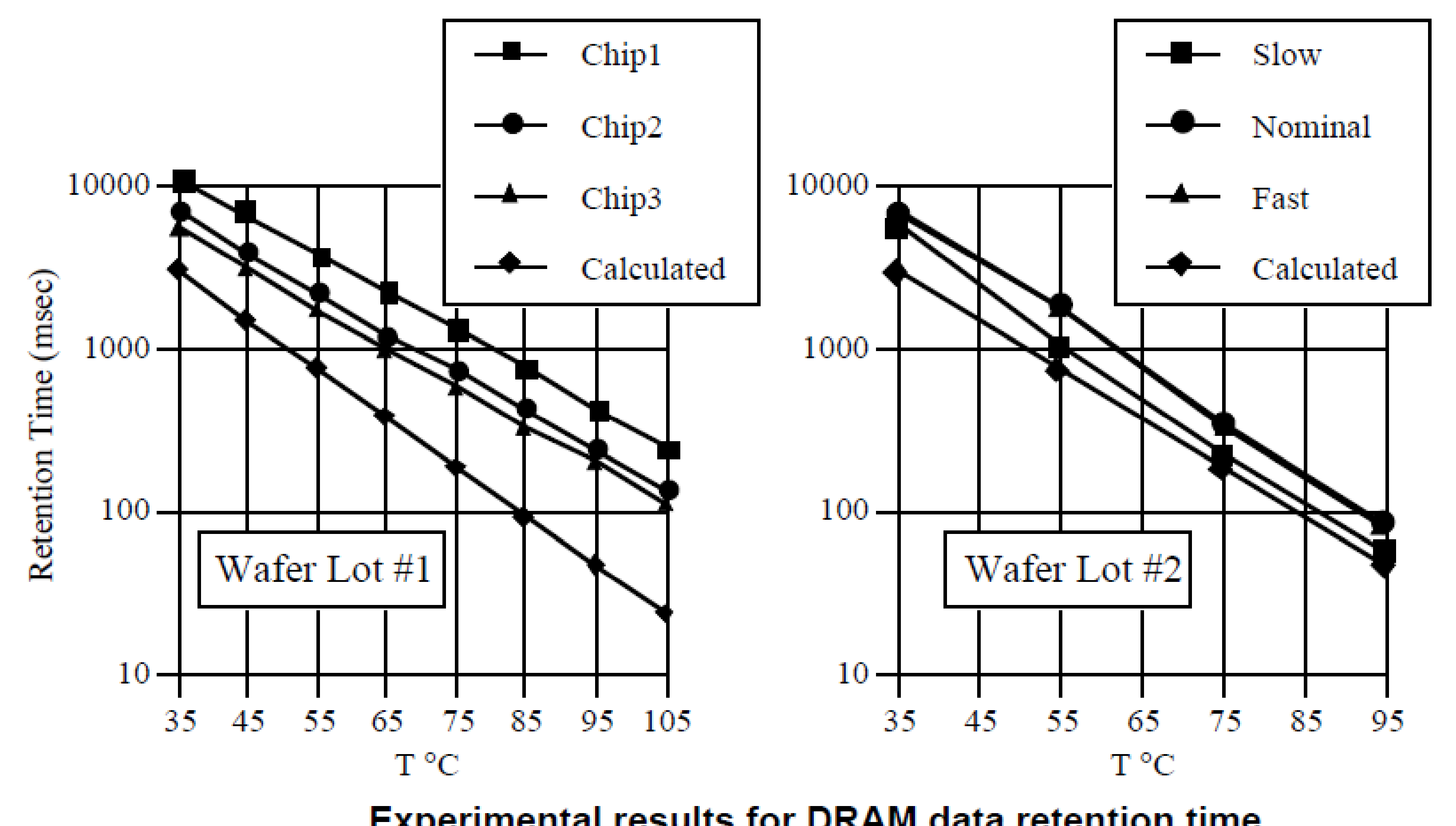
Idealized layout of an $8f^2$ DRAM array. The memory cell outline is given by dashed lines. Arrows indicate the direction of current flow through the transfer device. Circles indicate contact from bit line to diffusion of the transfer device. Word lines (WLs) run vertically; bit lines (BLs) are oriented perpendicularly. A storage capacitor (CAP; checkerboard pattern) may occupy more than $1f^2$ area.


ROW ADDR. DEMUX: SELECTS ROW
a0
a1
RAS
MUX
(4P2T)
RAS LO = 0
RAS HI = 1
SENSE
AMPLIFIER
(COMPARATOR)
D.I.
DATA IN
LATCH
SELECTS
INPUT ON
RISING EDGE
OF CAS
CAS
LATCH
COL.
ADDR.
LATCH
DATA SELECTOR (4 TO 1 MUX)
D.O. (DATA OUT)
TRI STATE
BUS

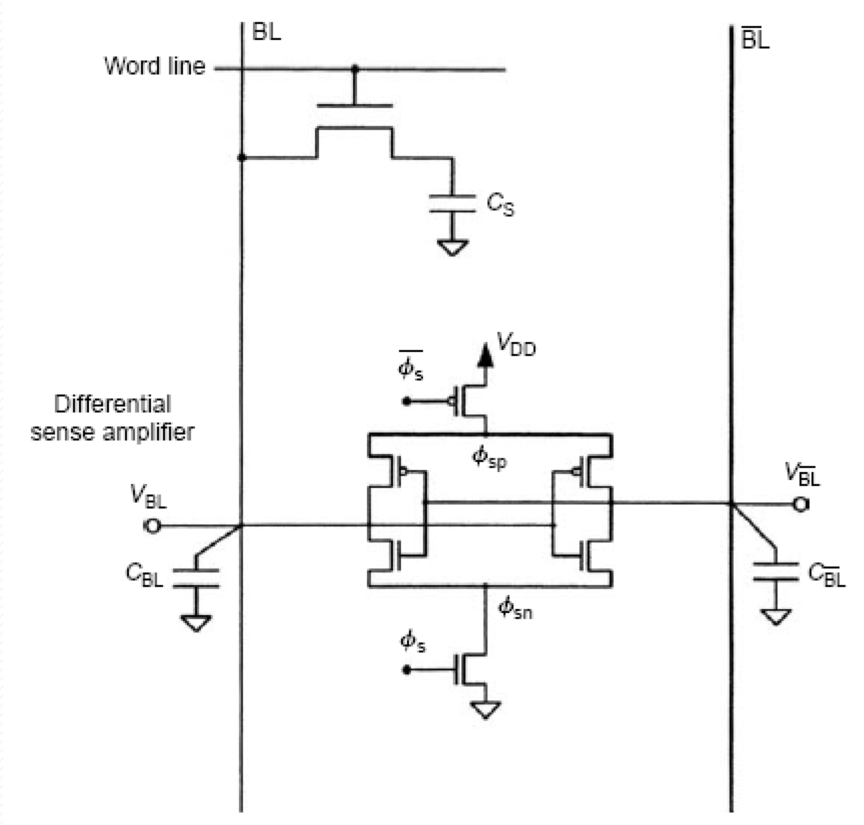
DRAM



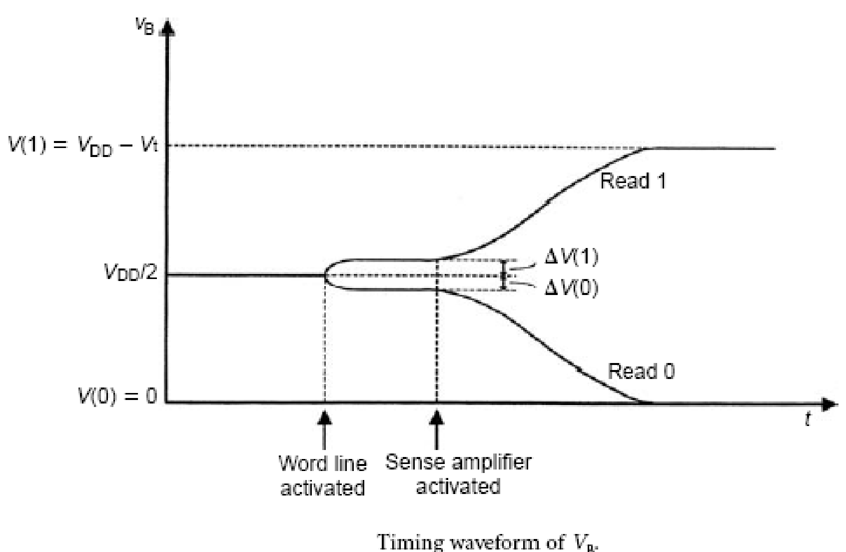
Schematic representation of a DRAM array. Word lines (WLs) are running vertically, bit lines (BLs) horizontally. Logically, bit lines are organized into true/complement pairs.



Experimental results for DRAM data retention time.

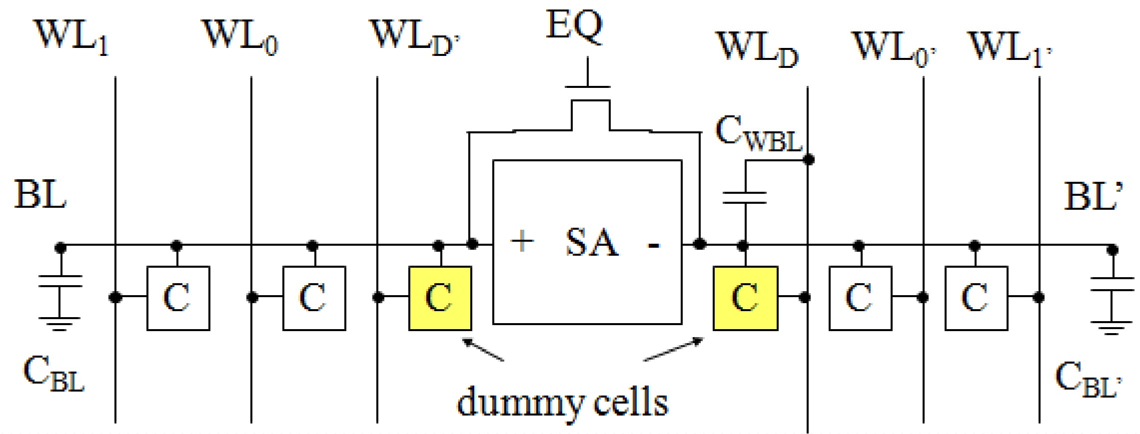


A differential sense amplifier connected to the bit-line.



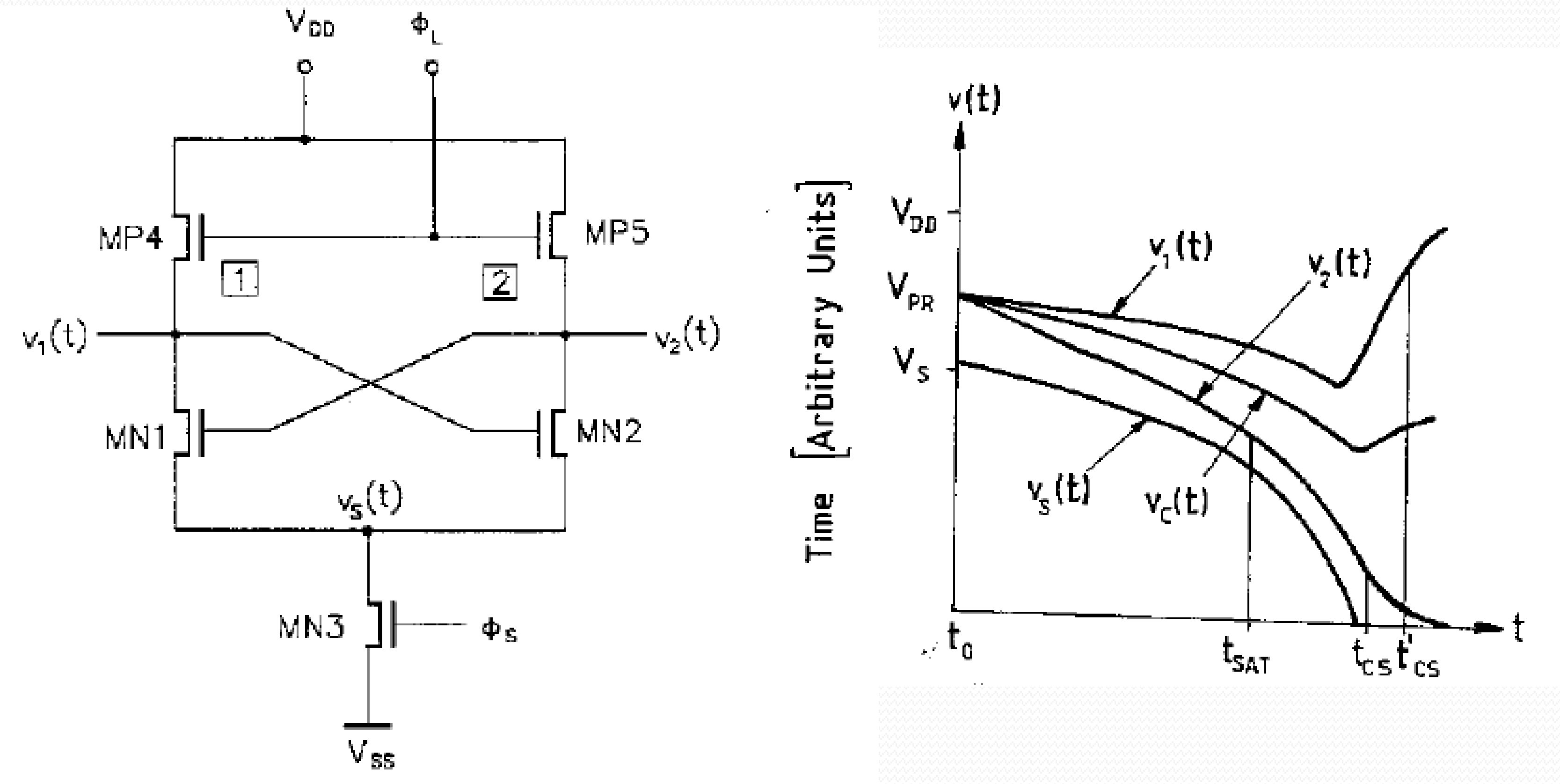
Timing waveform of V_B.

Sense amplifier with dummy-cell sensing structure



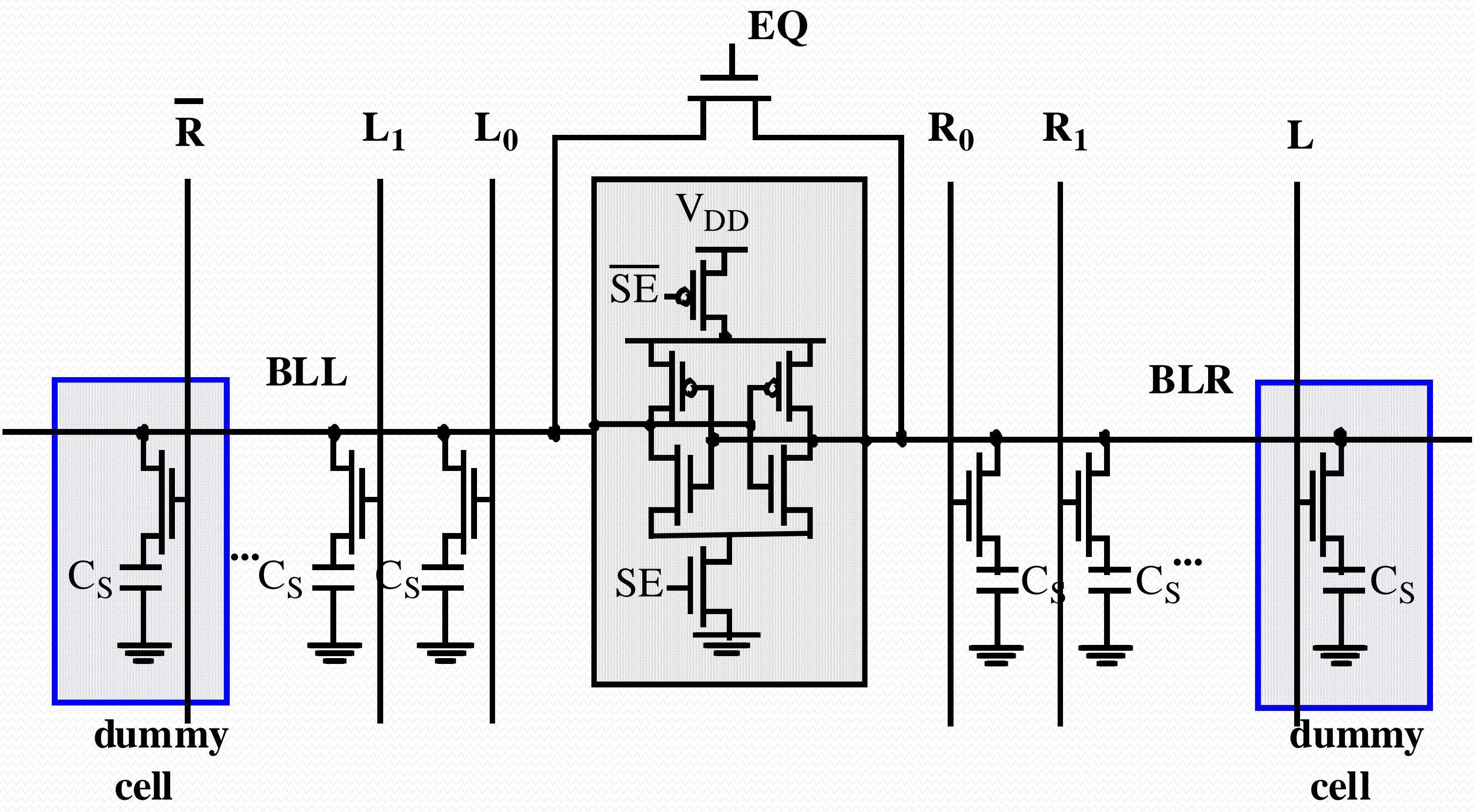
• During reading, if LHS cell is read, the dummy word line on RHS is acessed (and vice-versa).

• **sense amplifier**

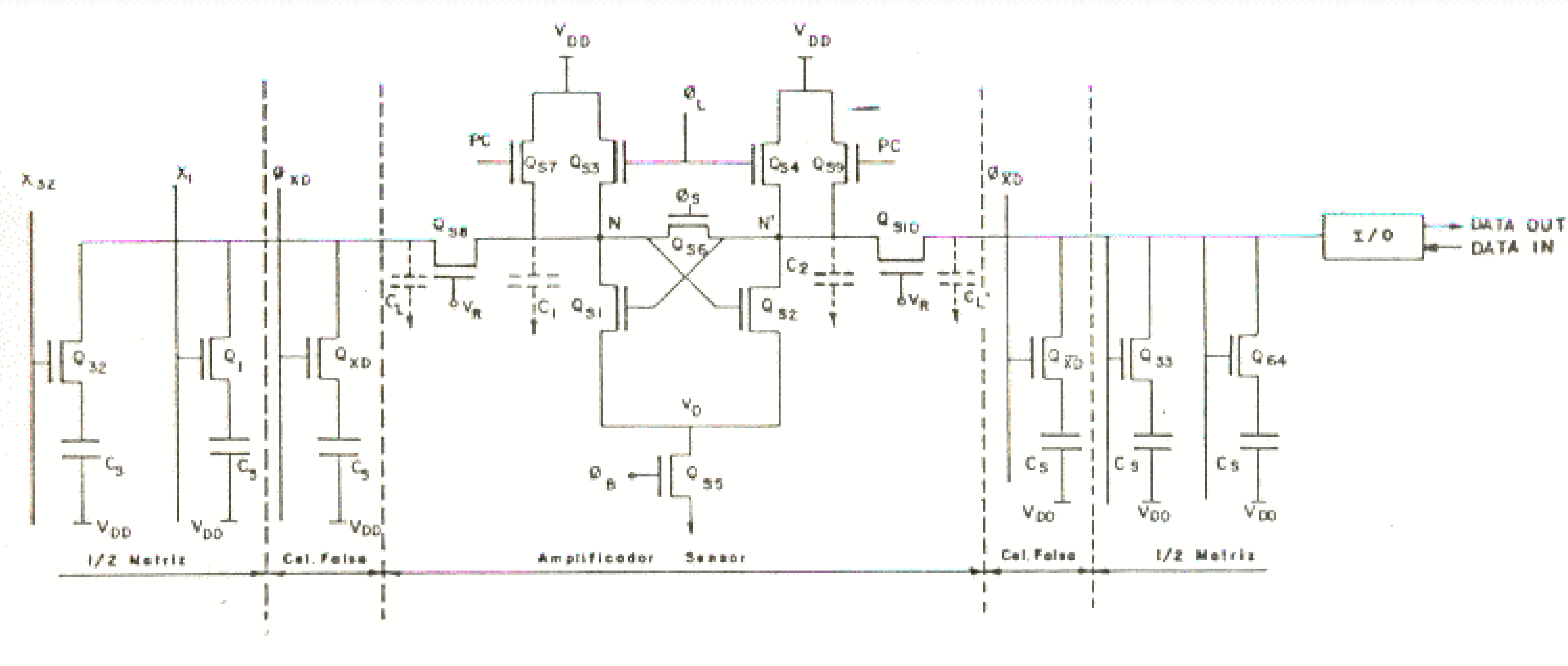


Positive feedback differential voltage sense amplifier circuit.

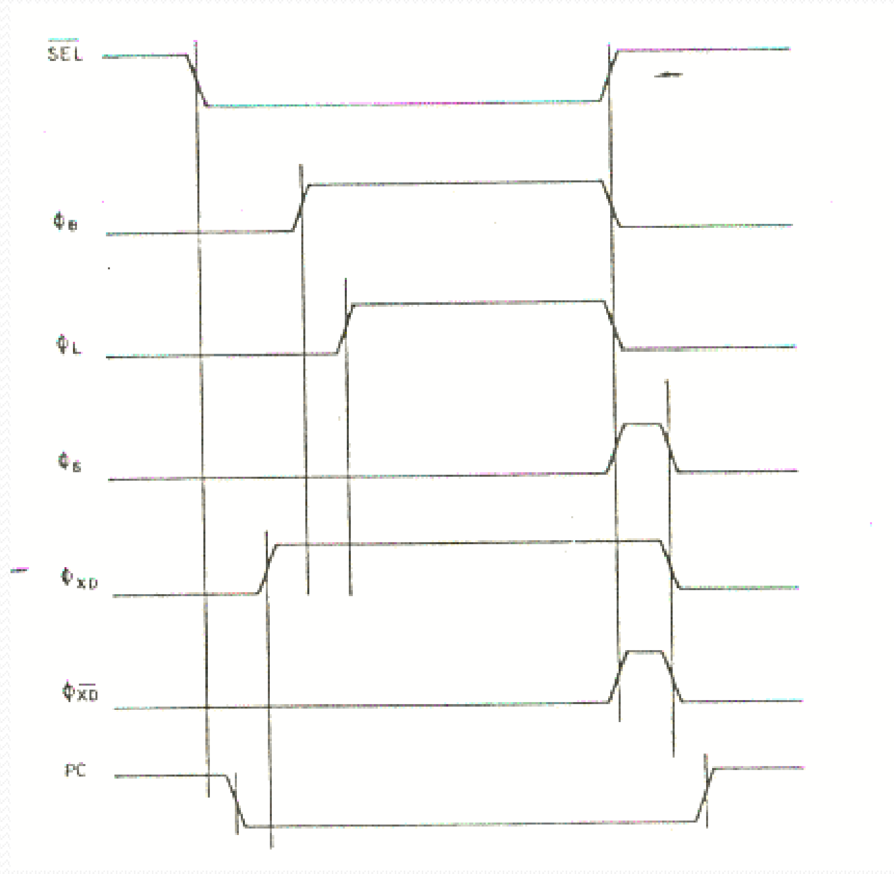
Sense amplifier with dummy-cell sensing structure



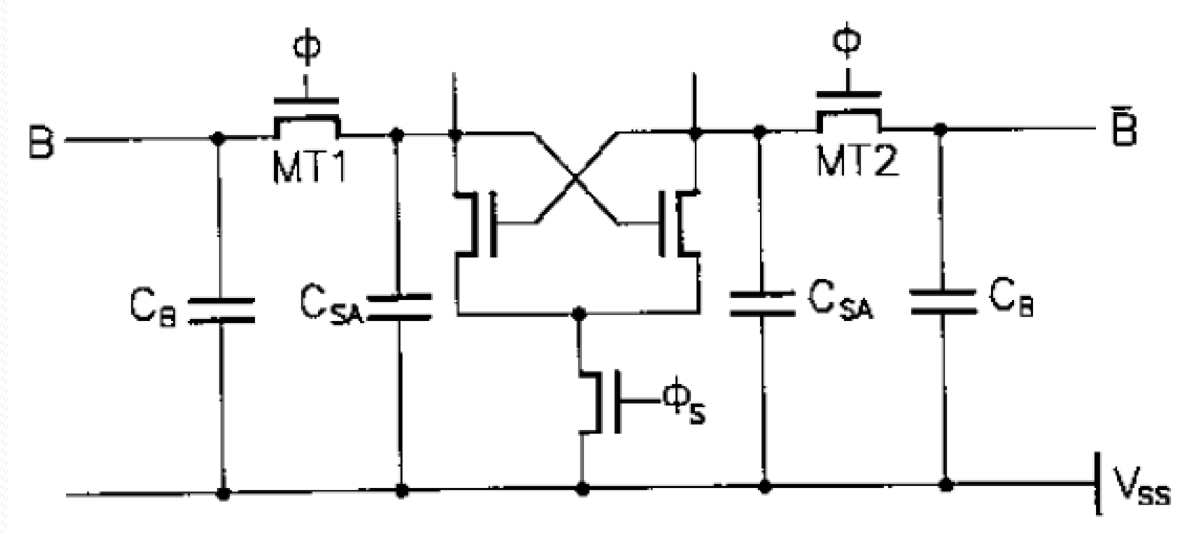
Sense amplifier with dummy-cell sensing structure



Sense-amplifier timing (dummy-cell approach)



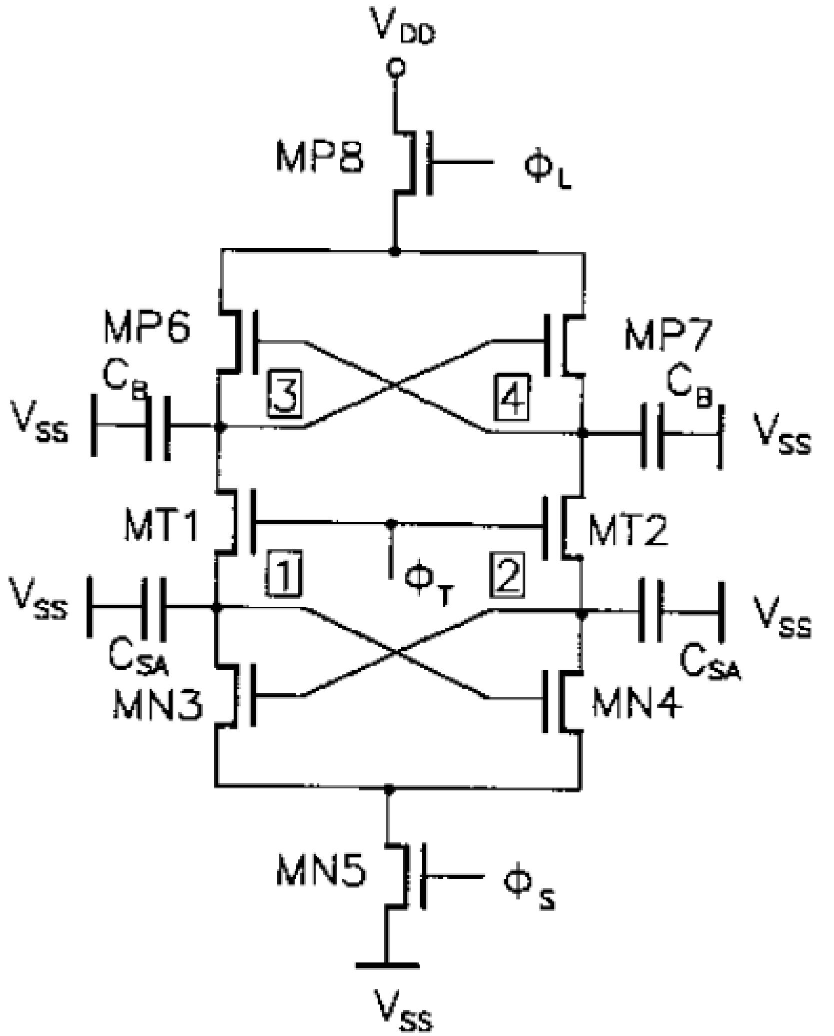
• Decoupling of bit-line capacitance



• decoupling devices MT1 and MT2 are initially turned on. SA is not active and load at its input is CBL + CSA, with CB >> CSA.

• AS is activated when differential input-voltage Vin is large enough. At this moment, MT1 and MT2 are turned off and load at AS input is only CSA. ⇒Faster arbitration.

• during re-writing, MT1 and MT2 are turned on again.



Sense amplifier incorporating decoupler devices.

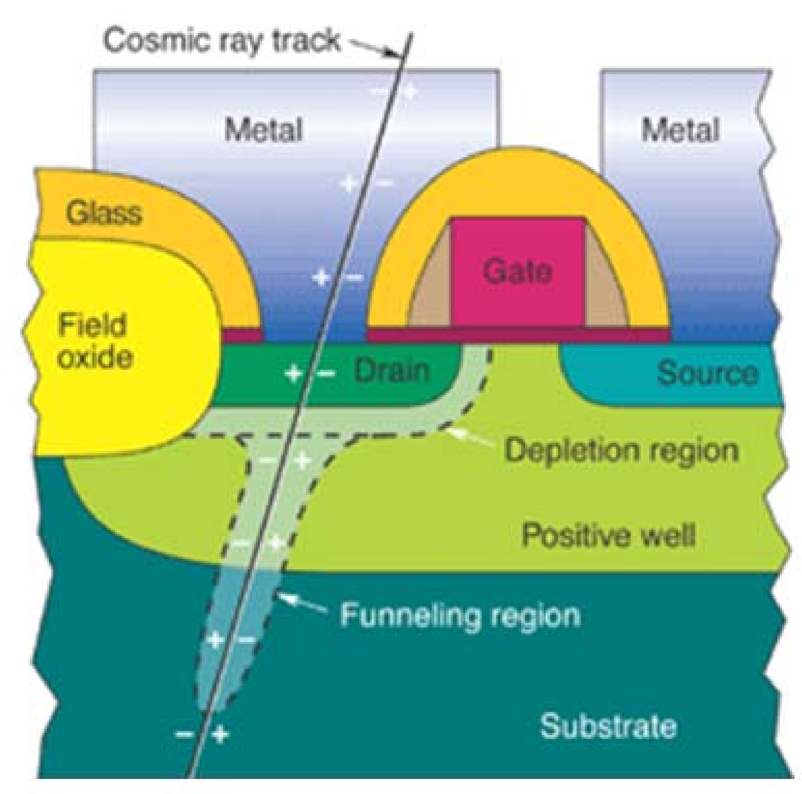
- initially, Φs turns on MN5.

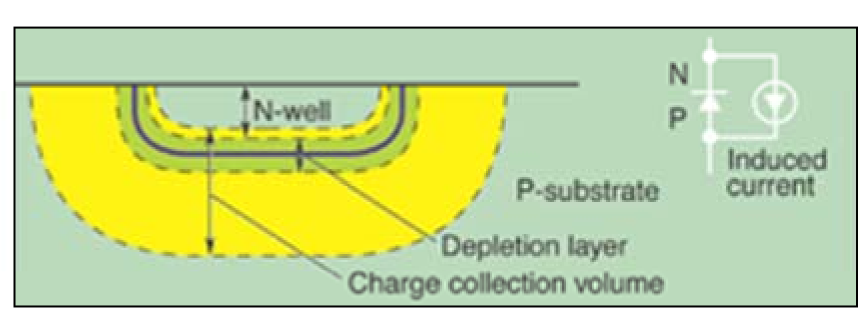
- Φt turns on decoupling devices MT1 and MT2. Both are turned off with differencial Vin is large enough.

- Vin is then quickly amplified by MN3, MN4 and MN5, as bit-line capacitances are de-coupled.

- when Vin reaches a given value, Φt goes high, turning on MT1 and MT2, which allows full Vin swing.

Alpha-particles

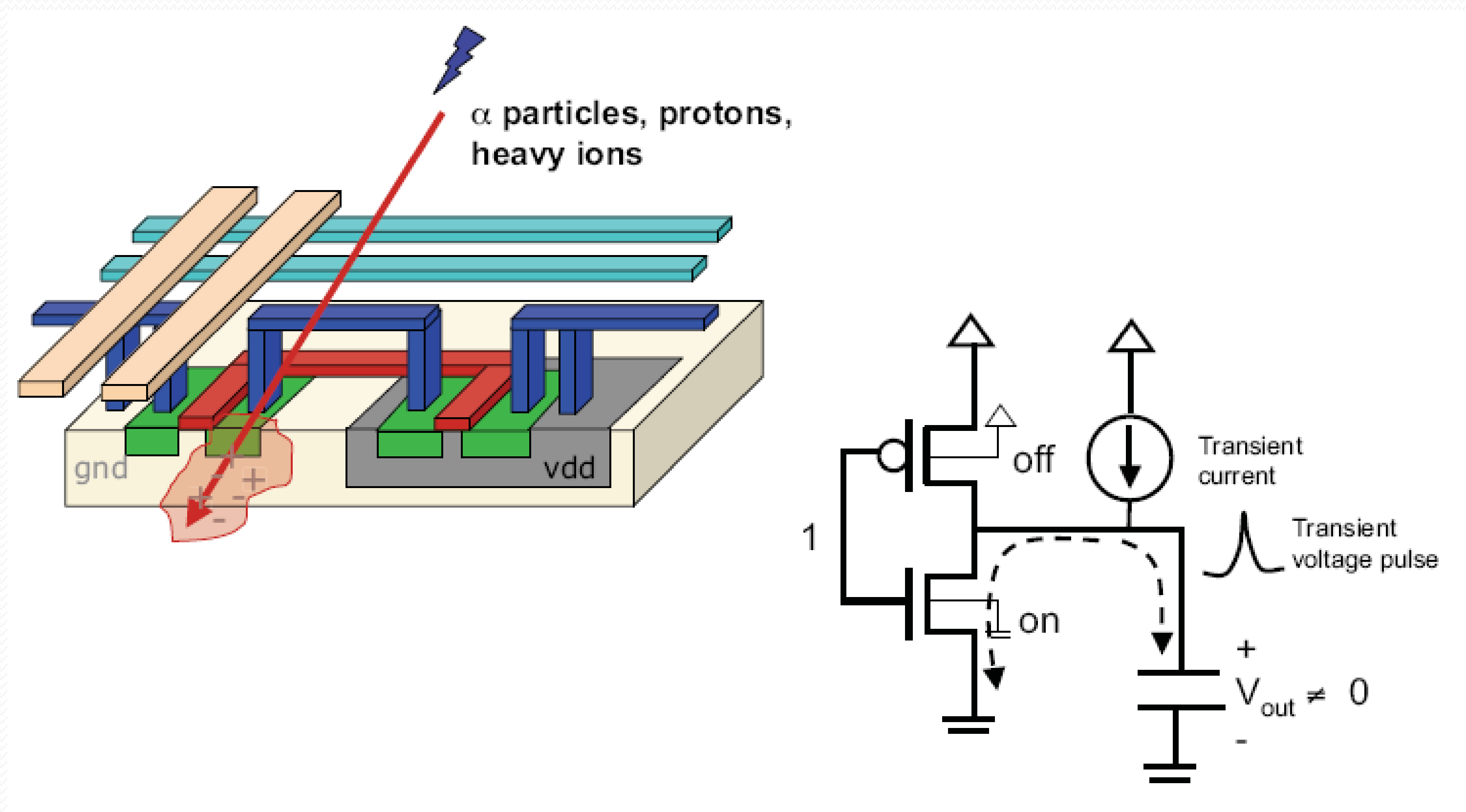




Current is generated in a p-n junction exposed to transient radiation. It can be modeled as a transient current source in parallel with a diode.

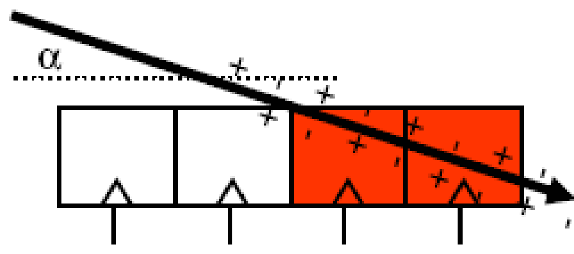
1 particle ~1million carriers

Alpha-particles

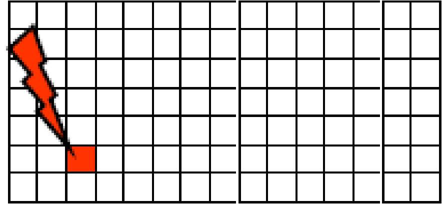


Challenges in Sequential Logic

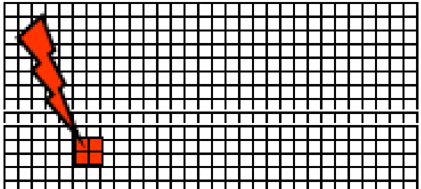
MULTIPLE BIT UPSETS



- Particle incidence angle
- Transistor Dimensions ↓
- Voltage Supply ↓
- Memory Array Density ↑



Single memory cell

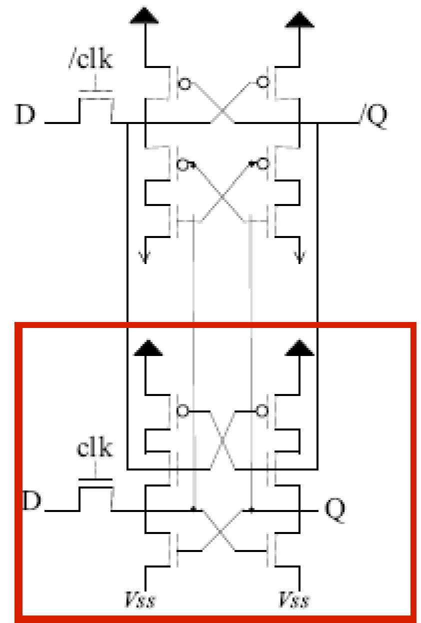


Multiple memory cells

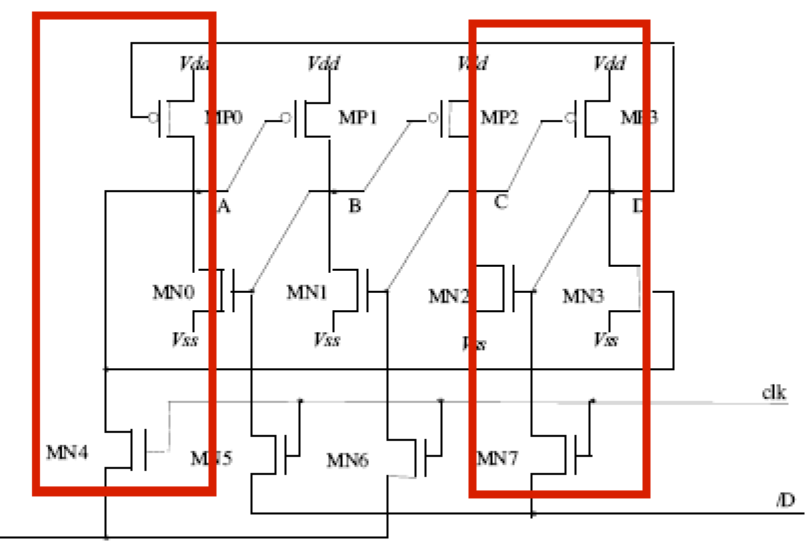
Hardened Memories

- The principle is to store the data in two different locations within the cell in such way that the corrupted part can be restored.

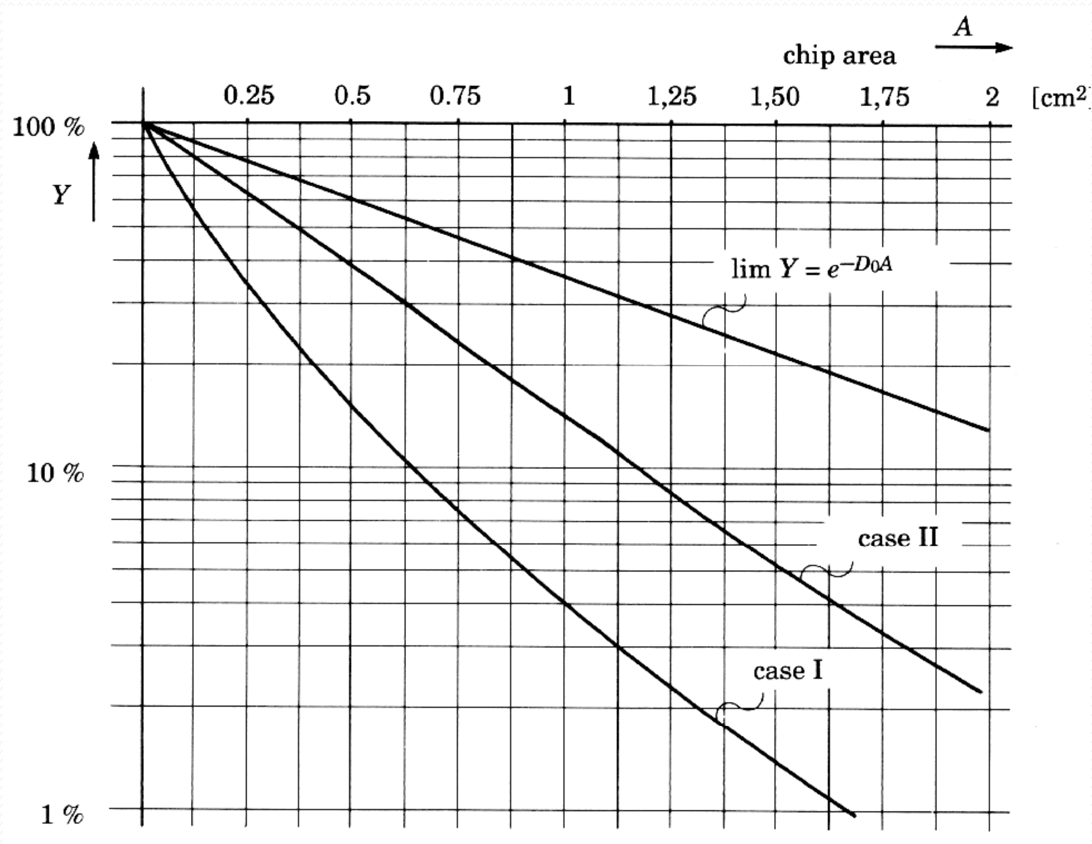
Whitaker/Liu Memory Cell [Liu, 92]



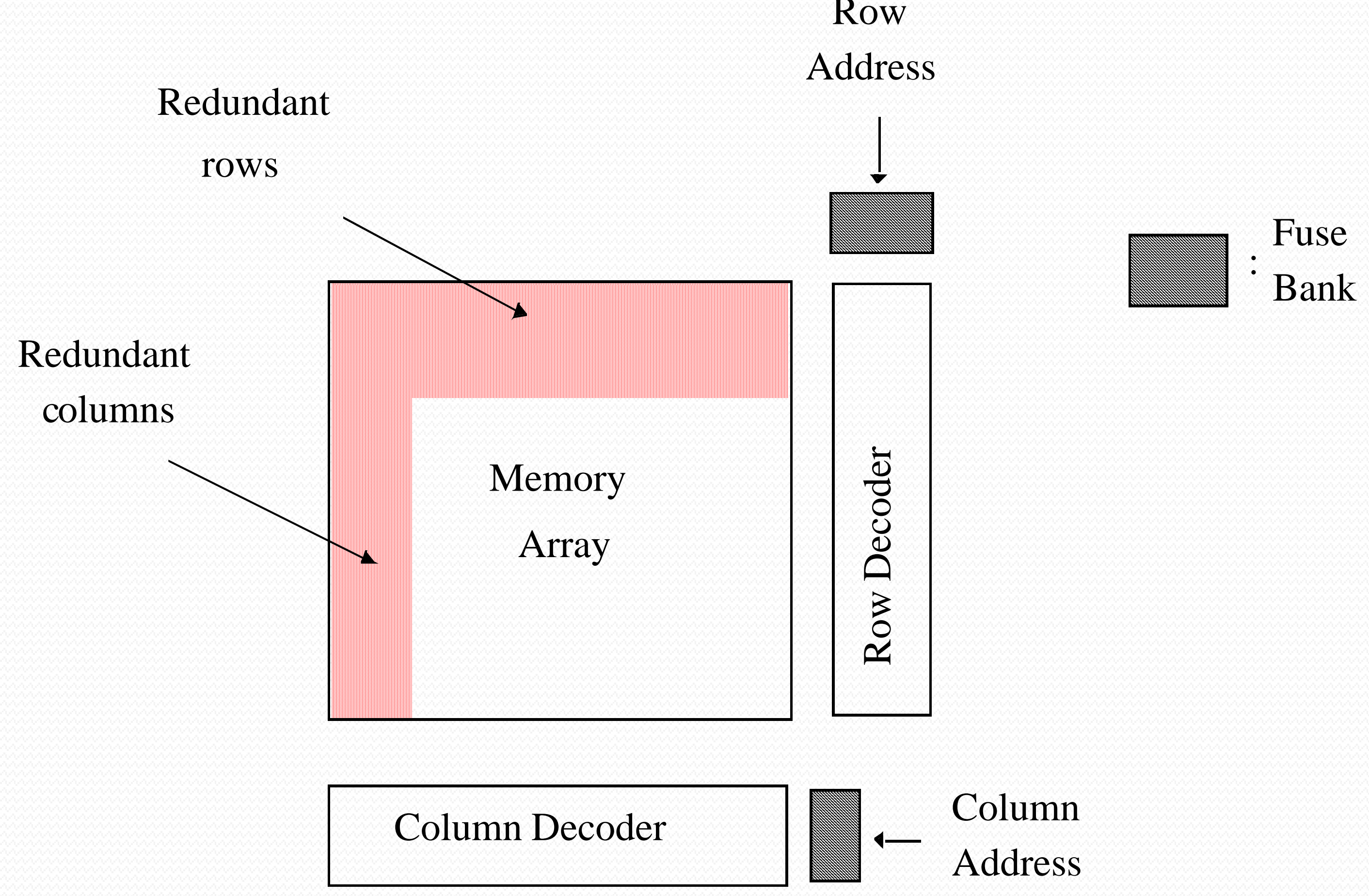
DICE Memory Cell [Calin, 96]



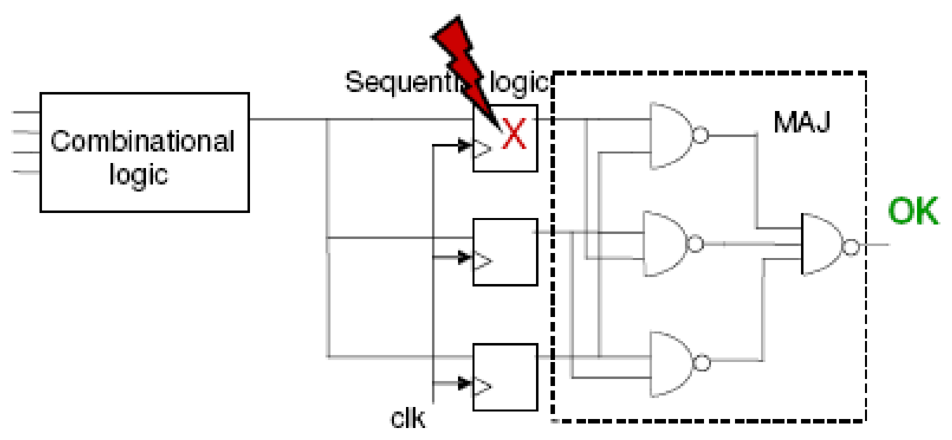
Yield

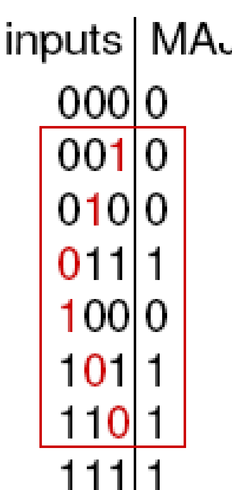


Redundancy



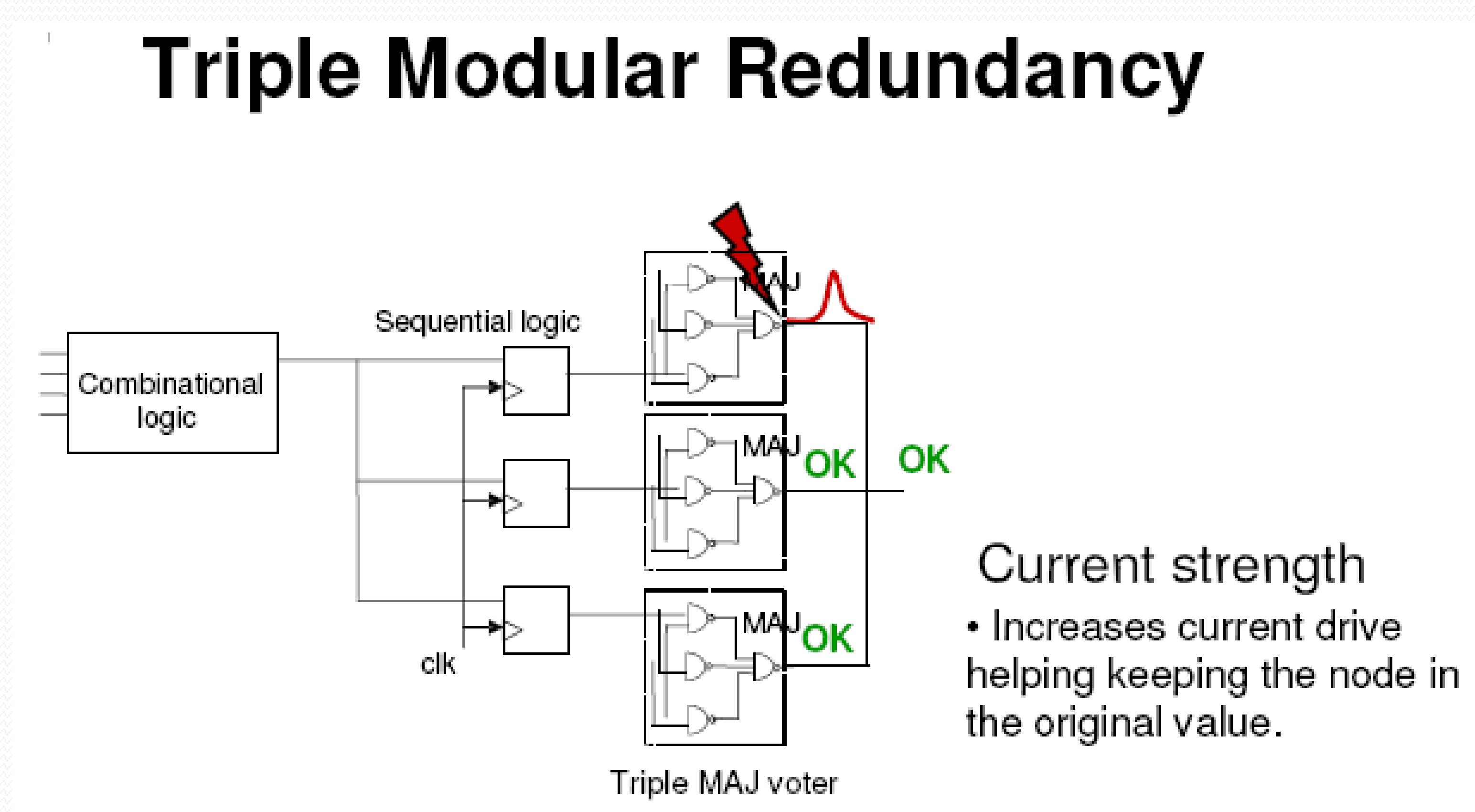
Triple Modular Redundancy



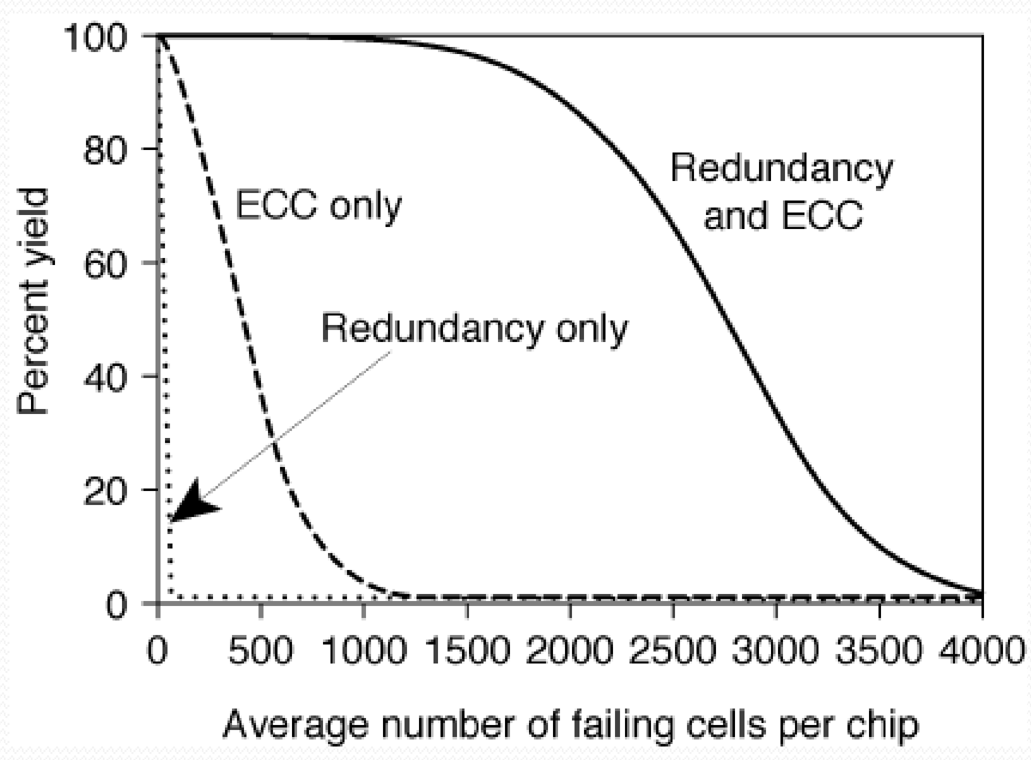


inputs	MAJ
000	0
001	0
010	0
011	1
100	0
101	1
110	1
111	1

Triple Modular Redundancy

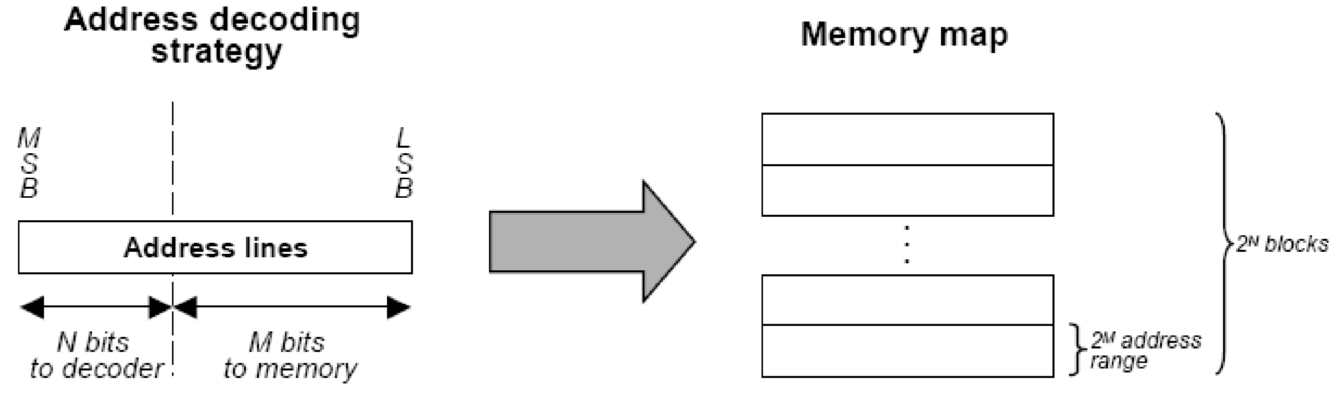


Redundancy and Error Correction



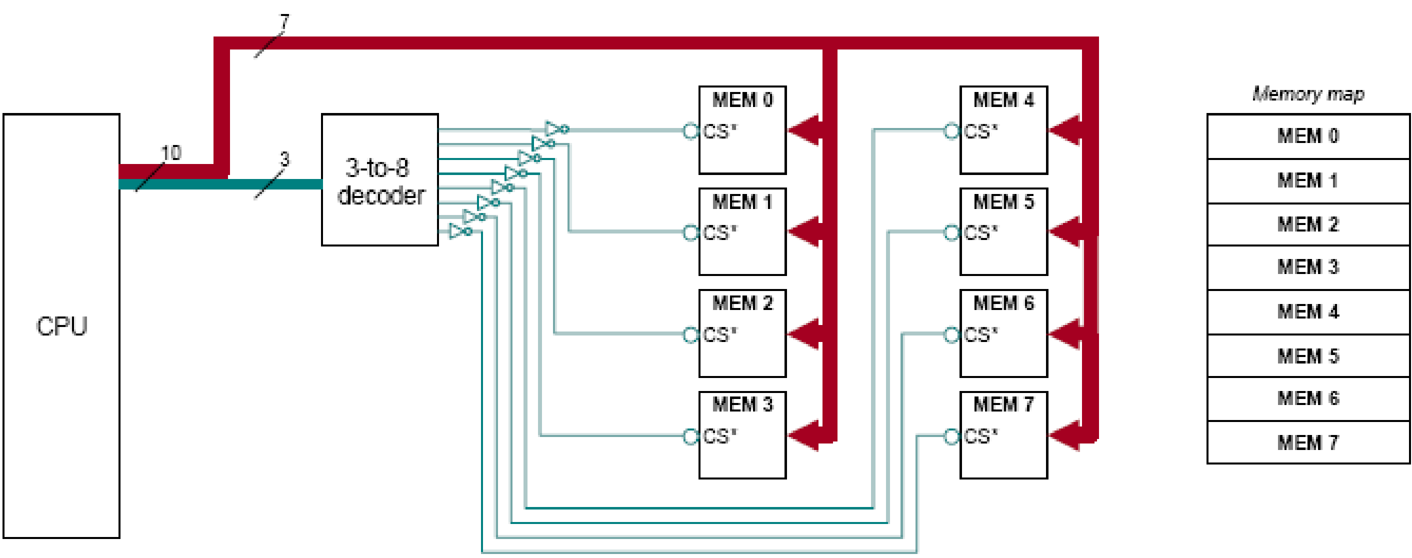
Address Decoding

- **Address decoding is the process of generating chip select (CS*) signals from the address bus for each device in the system**
- **The address bus lines are split into two sections**
  - the N most significant bits are used to generate the CS* signals for the different devices
  - the M least significant signals are passed to the devices as addresses to the different memory cells or internal registers



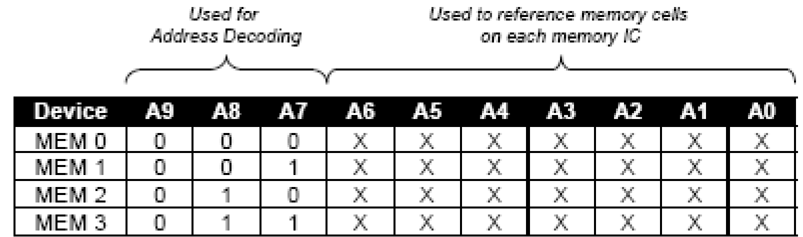
Example:

- **Let's assume a very simple microprocessor with 10 address lines (1KB memory)**
- **Let's assume we wish to implement all its memory space and we use 128x8 memory chips**
- **SOLUTION**
  - We will need 8 memory chips (8x128=1024)
  - We will need 3 address lines to select each one of the 8 chips
  - Each chip will need 7 address lines to address its internal memory cells



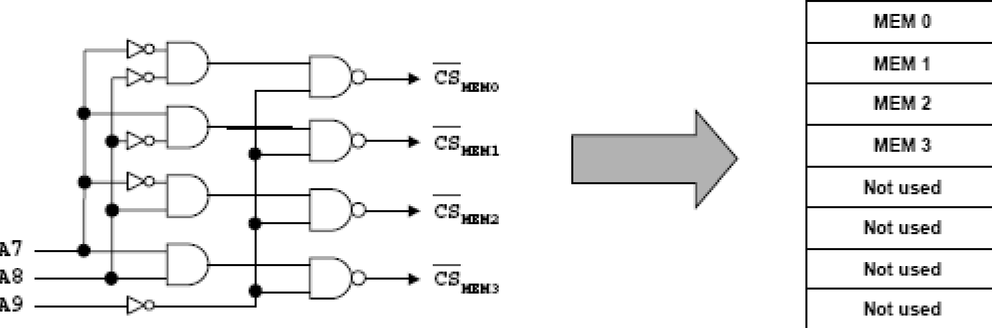
Full address decoding

- **Let's assume the same microprocessor with 10 address lines (1KB memory)**
  - However, this time we wish to implement only 512 bytes of memory
  - We still must use 128-byte memory chips
  - Physical memory must be placed on the upper half of the memory map
- **SOLUTION**



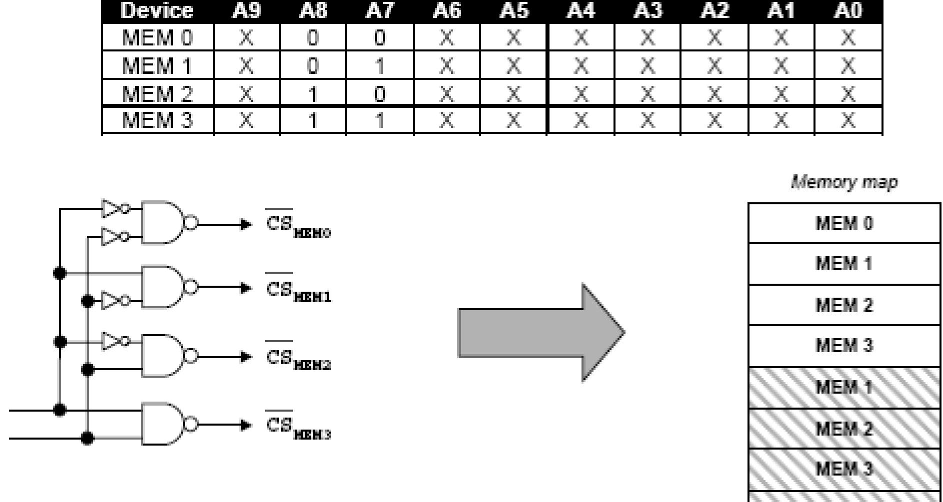
Used for Address Decoding (A9–A7) — *Used to reference memory cells on each memory IC* (A6–A0)

Device	A9	A8	A7	A6	A5	A4	A3	A2	A1	A0
MEM 0	0	0	0	X	X	X	X	X	X	X
MEM 1	0	0	1	X	X	X	X	X	X	X
MEM 2	0	1	0	X	X	X	X	X	X	X
MEM 3	0	1	1	X	X	X	X	X	X	X



Partial address decoding

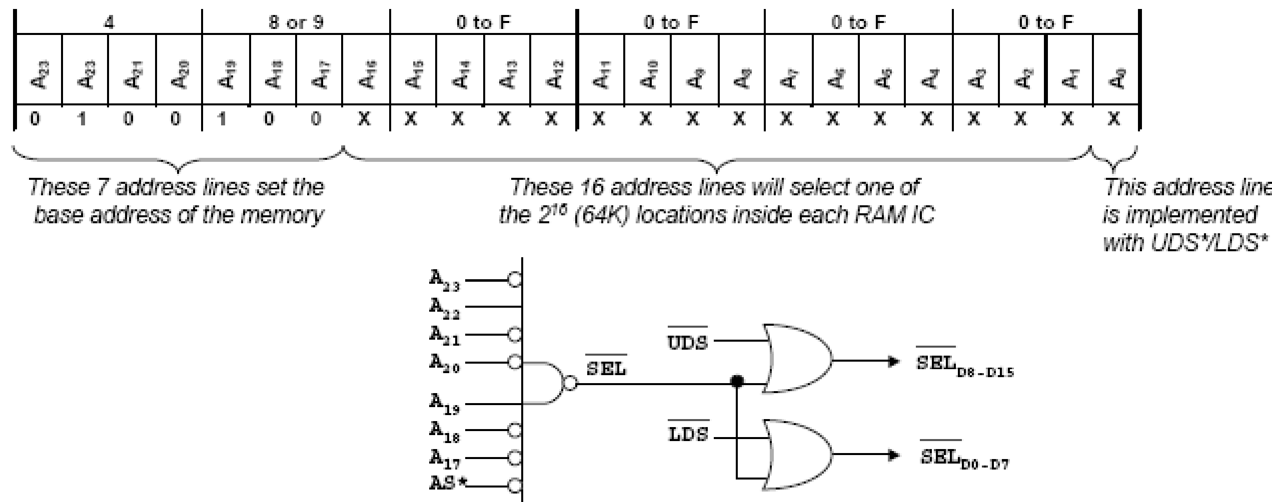
- **Let's assume the same microprocessor with 10 address lines (1KB memory)**
  - However, this time we wish to implement only 512 bytes of memory
  - We still must use 128-byte memory chips
  - Physical memory must be placed on the upper half of the memory map
- **SOLUTION**



Device	A9	A8	A7	A6	A5	A4	A3	A2	A1	A0
MEM 0	X	0	0	X	X	X	X	X	X	X
MEM 1	X	0	1	X	X	X	X	X	X	X
MEM 2	X	1	0	X	X	X	X	X	X	X
MEM 3	X	1	1	X	X	X	X	X	X	X

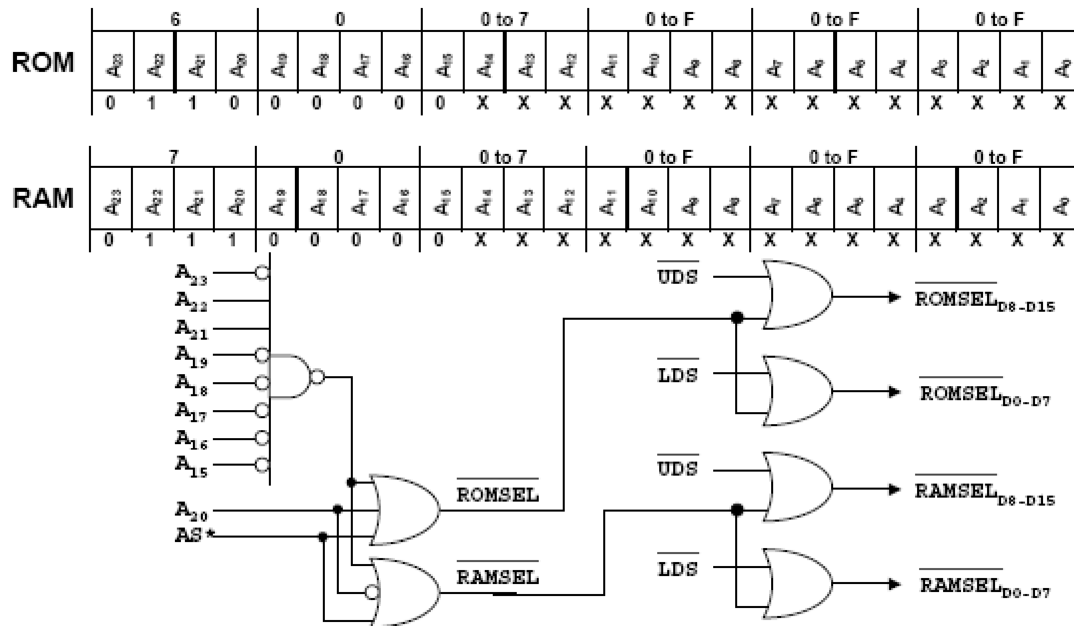
Example 1

- **A circuit containing 64K words of RAM is to be interfaced to a 68000-based system, so that the first address of RAM (the *base* address) is at $480000**
  - What is the entire range of RAM addresses?
  - Design a FULL address decoder using two 64K×8 RAM ICs
- **Solution**
  - The address range for the RAM is from $480000 to $480000+(128K=$20000)=$4A0000-1=$49FFFF
  - The two ICs must be differentiated through UDS*/LDS* (since the 68000 DOES NOT have A_0)



Example 2

- **A 68000-based system is to be built with these memory requirements**
  - a 16K word EPROM with a starting address of $60 0000
  - a 16K word RAM with a starting address of $70 0000
- **Design a FULL address decoder for this application using 16K×8 chips for both EPROM and RAM**
  - $60 0000 + (16KWord=32KB=$8000)-1=$60 7FFF
  - $70 0000 + (16KWord=32KB=$8000)-1=$70 7FFF

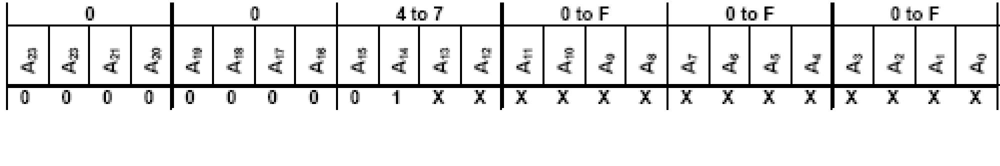


ROM	6				0				0 to 7				0 to F				0 to F				0 to F			
	A_{23}	A_{22}	A_{21}	A_{20}	A_{19}	A_{18}	A_{17}	A_{16}	A_{15}	A_{14}	A_{13}	A_{12}	A_{11}	A_{10}	A_9	A_8	A_7	A_6	A_5	A_4	A_3	A_2	A_1	A_0
	0	1	1	0	0	0	0	0	0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

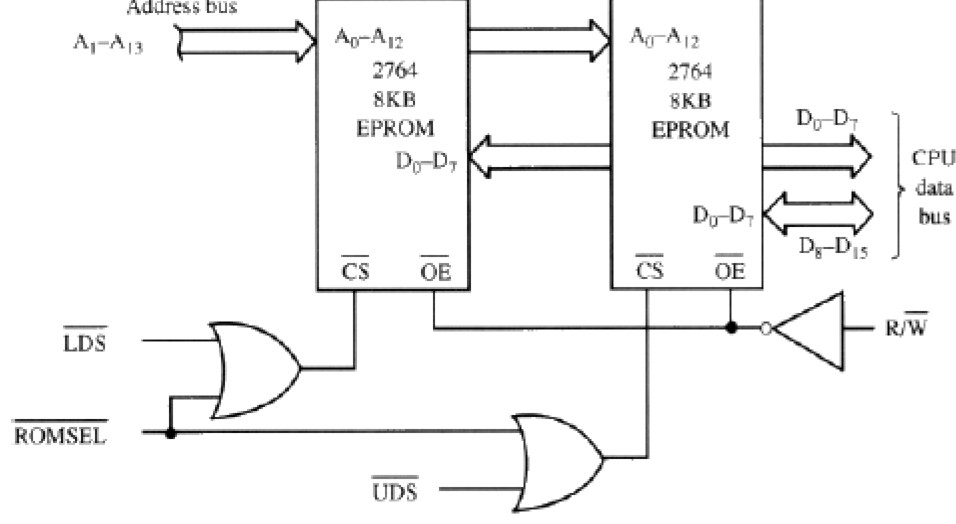
RAM	7				0				0 to 7				0 to F				0 to F				0 to F			
	A_{23}	A_{22}	A_{21}	A_{20}	A_{19}	A_{18}	A_{17}	A_{16}	A_{15}	A_{14}	A_{13}	A_{12}	A_{11}	A_{10}	A_9	A_8	A_7	A_6	A_5	A_4	A_3	A_2	A_1	A_0
	0	1	1	1	0	0	0	0	0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Example 3

- **Design a PARTIAL address decoder for a 68000-based system with only 8K words of EPROM space, and a base address at $4000, using 8Kx8 memory chips**

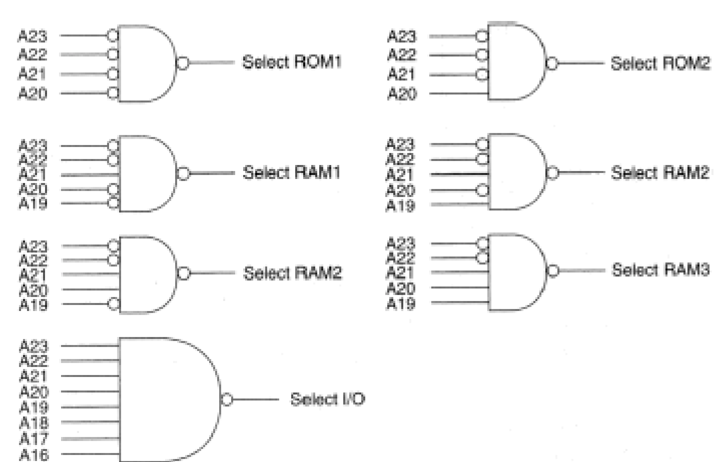


0				0				4 to 7				0 to F				0 to F				0 to F			
A_{23}	A_{22}	A_{21}	A_{20}	A_{19}	A_{18}	A_{17}	A_{16}	A_{15}	A_{14}	A_{13}	A_{12}	A_{11}	A_{10}	A_9	A_8	A_7	A_6	A_5	A_4	A_3	A_2	A_1	A_0
0	0	0	0	0	0	0	0	0	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X



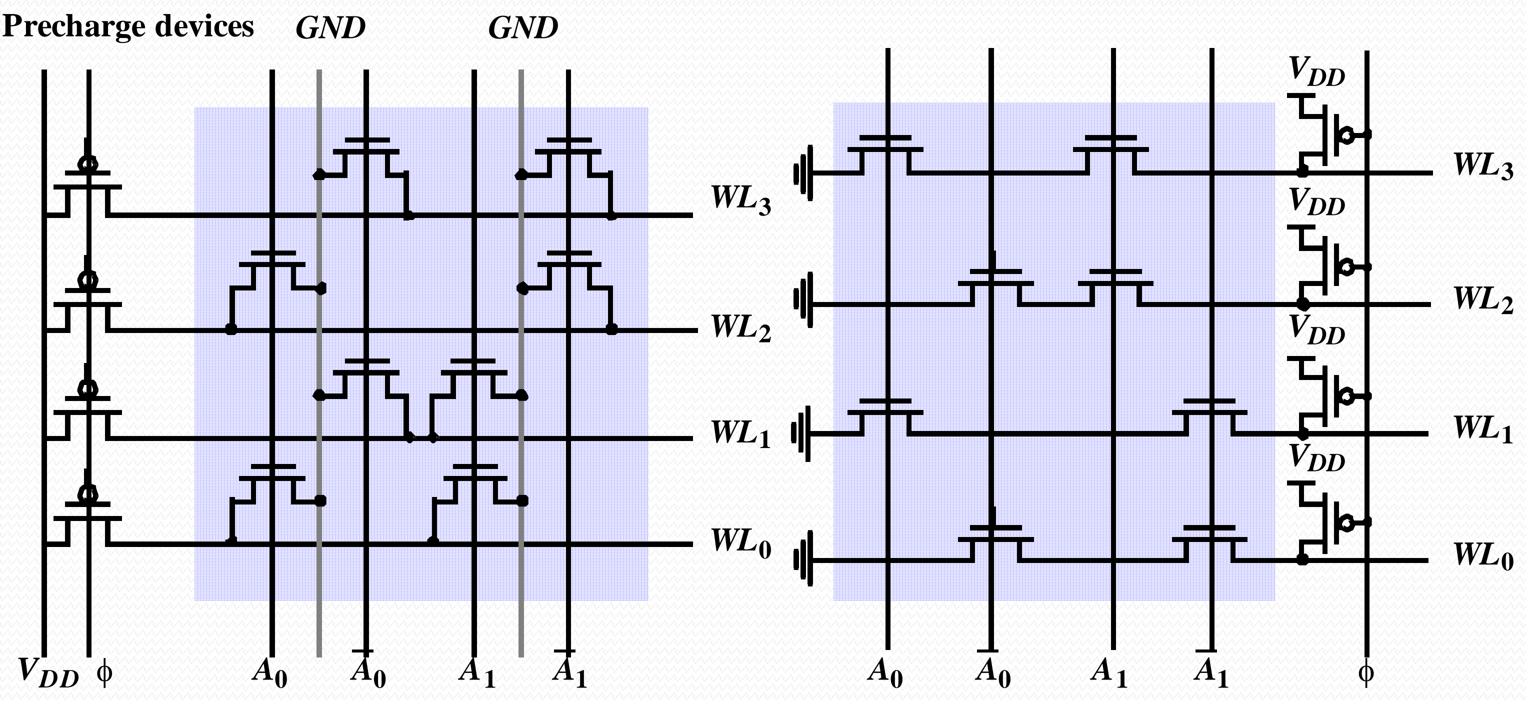
Example 4

- **Design a partial address decoder for a 68000-based system that contains**
  - 2MB of EPROM at a starting address $00 0000 using 512Kx8 chips
  - 2MB of RAM at a starting address $10 0000 using 256Kx8 chips
  - 64KB I/O space starting at $FF0000
- **SOLUTION**
  - For the EPROM we will need 4 512Kx8 chips, organized as 2 pairs of 512x8 chips (in order to use UDS*/LDS*). We will call these pairs ROM1 and ROM2
  - For the RAM we will need 8 256Kx8 chips, organized as 4 pairs of 256Kx8: RAM1 to RAM4



	A_{23}	A_{22}	A_{21}	A_{20}	A_{19}	A_{18}	A_{17}	A_{16}	A_{15}	A_{14}	A_{13}	A_{12}	A_{11}	A_{10}	A_{09}	A_{08}	A_{07}	A_{06}	A_{05}	A_{04}	A_{03}	A_{02}	A_{01}	A_{00}
ROM1	0	0	0	0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ROM2	0	0	0	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
RAM1	0	0	1	0	0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
RAM2	0	0	1	0	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
RAM3	0	0	1	1	0	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
RAM4	0	0	1	1	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
I/O	1	1	1	1	1	1	1	1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

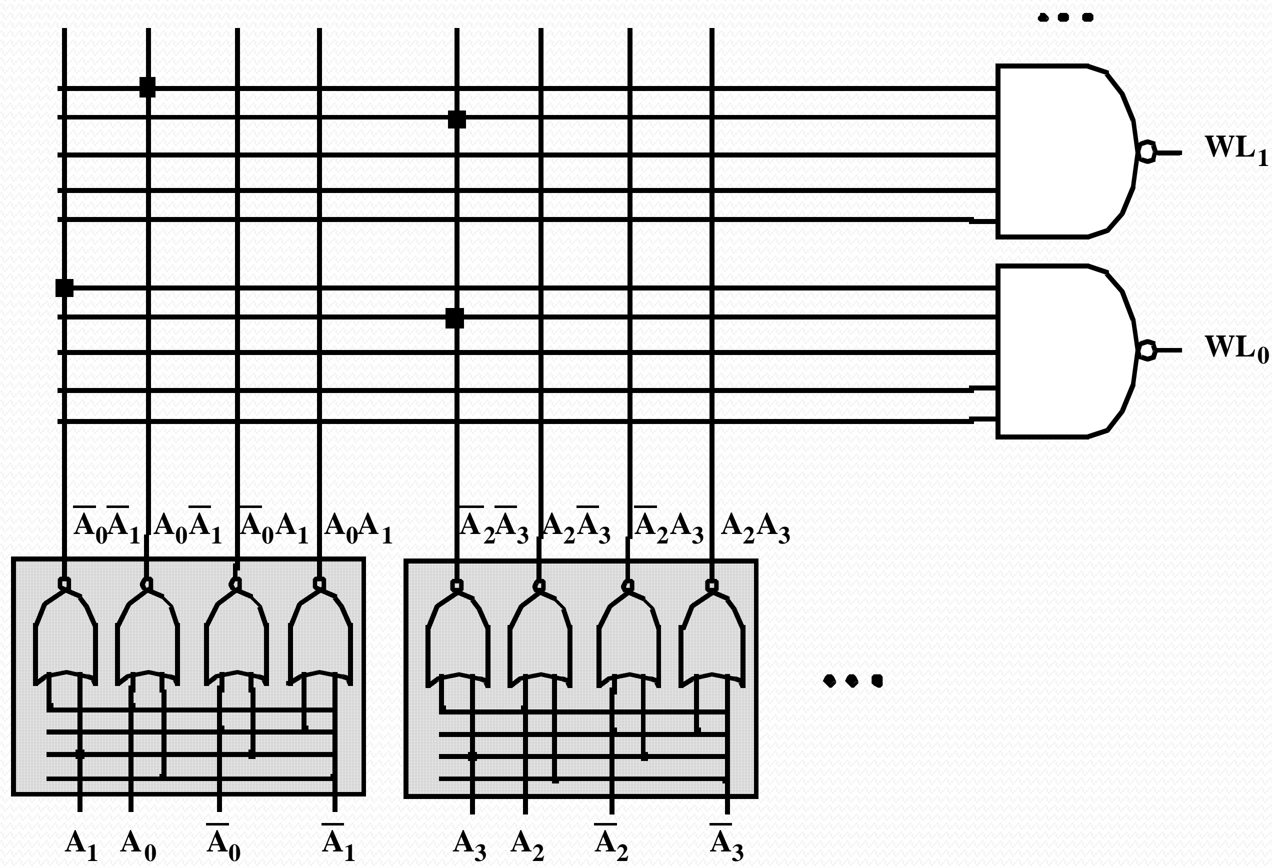
Dynamic Decoding



Dynamic 2-to-4 NOR decoder | **2-to-4 MOS dynamic NAND Decoder**

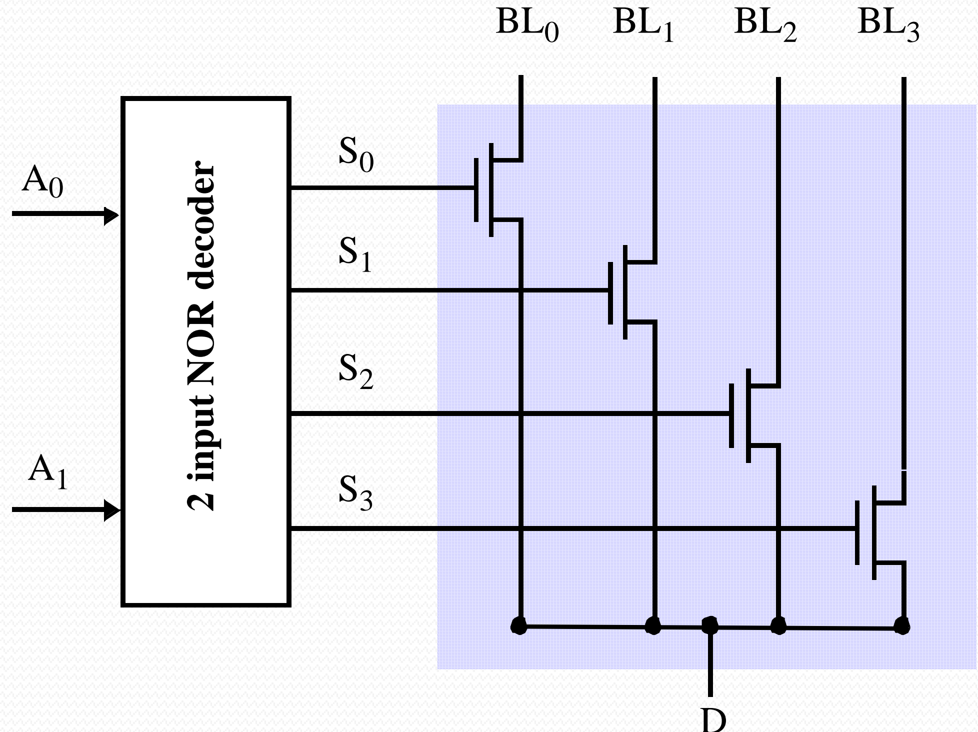
Propagation delay is primary concern

A NAND decoder using 2-input pre-decoders



Splitting decoder into two or more logic layers produces a faster and cheaper implementation

4 input pass-transistor based column decoder

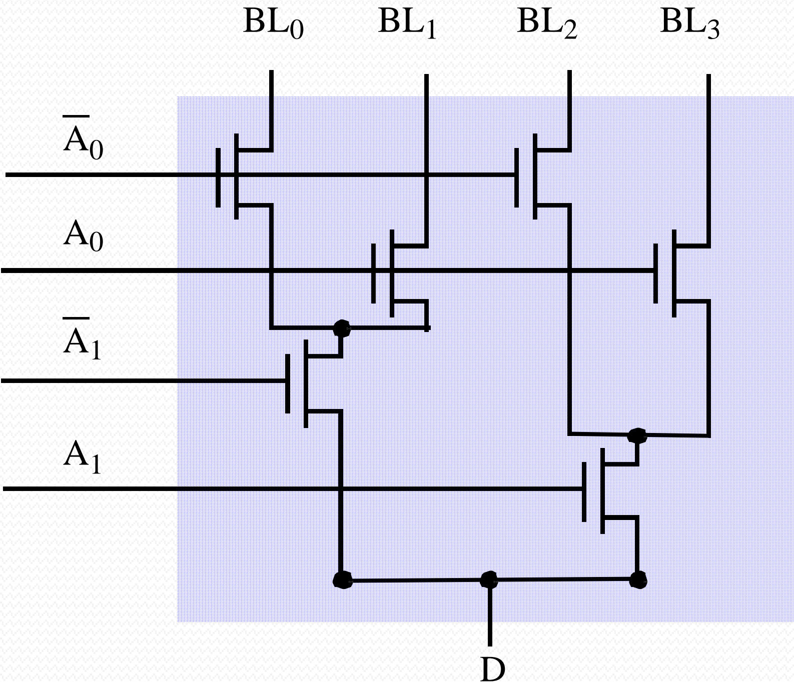


lvantage: speed (t_{pd} does not add to overall memory access time)

only 1 extra transistor in signal path

sadvantage: large transistor count

4-to-1 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