Semiconductor Memory Design and Testing

Prof. Mahesh Awati/Dr. Shashidhar/Prof. M S Sunita
Department of Electronics and Communication Engg.

SEMICONDUCTOR MEMORY DESIGN AND TESTING

UNIT 1 – Semiconductor Memory Technology overview

Mahesh Awati/Dr Shashidhar/Prof. M S Sunita

Department of Electronics and Communication Engineering

UNIT 1 Outline

Unit 1: Semiconductor Memory technology overview

- Introduction to memory hierarchy
- Data expslotion to zetta scale
- Memory hierarchy in memory sub systems
- Introduction to memory array architecture
- Generic memory array diagram
- Memory cell size and equivalent bit area
- Memory array’s area effeciency
- Peripheral circuits
  - Decoder
  - MUX and driver
  - Sense Amplifier,
- Industry technology scaling trend

SRAM:

- 6T SRAM cell operation
- SRAM stability analysis

Introduction

- Memory is the means of storing information in the form of 1s and 0s.

- Every digital device, today, has a significant amount of memory embedded inside which take up app 80% of the silicon real estate.

- Depending on Application the amount of Memory requirement varies.

- More than half of the transistors in high performance Microprocessors are devoted to cache memories and the ratio is expected to further increase.

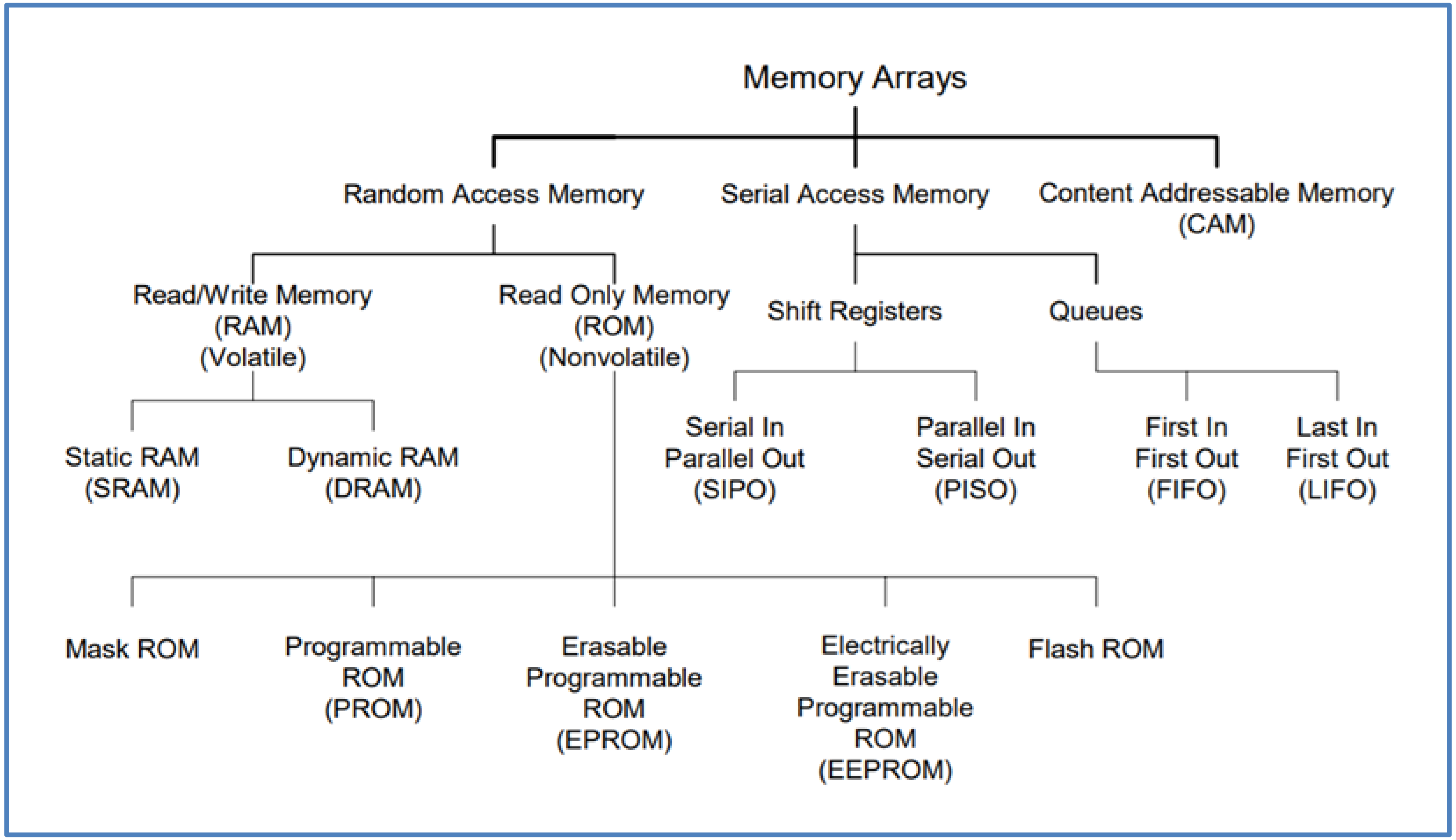
- Data creation to the level of zetta scale due to large number of devices connected to internet…estimated 75 billion

Introduction

- Memory holds short-term data with faster and more frequent read/write access, while storage holds long-term data with slower and less frequent read/ write access

- In general, the number of transistors for data storage is much larger than that for logic functions.

Classification of Semiconductor Memory



Memory Arrays

Random Access Memory

Serial Access Memory

Content Addressable Memory (CAM)

Read/Write Memory (RAM) (Volatile)

Read Only Memory (ROM) (Nonvolatile)

Shift Registers

Queues

Static RAM (SRAM)

Dynamic RAM (DRAM)

Serial In Parallel Out (SIPO)

Parallel In Serial Out (PISO)

First In First Out (FIFO)

Last In First Out (LIFO)

Mask ROM

Programmable ROM (PROM)

Erasable Programmable ROM (EPROM)

Electrically Erasable Programmable ROM (EEPROM)

Flash ROM

Factors to be considered while designing memory

1) Area efficiency of the memory array which is the number of stored data bits/unit area

   - Determines the overall storage capacity and hence the memory cost/bit

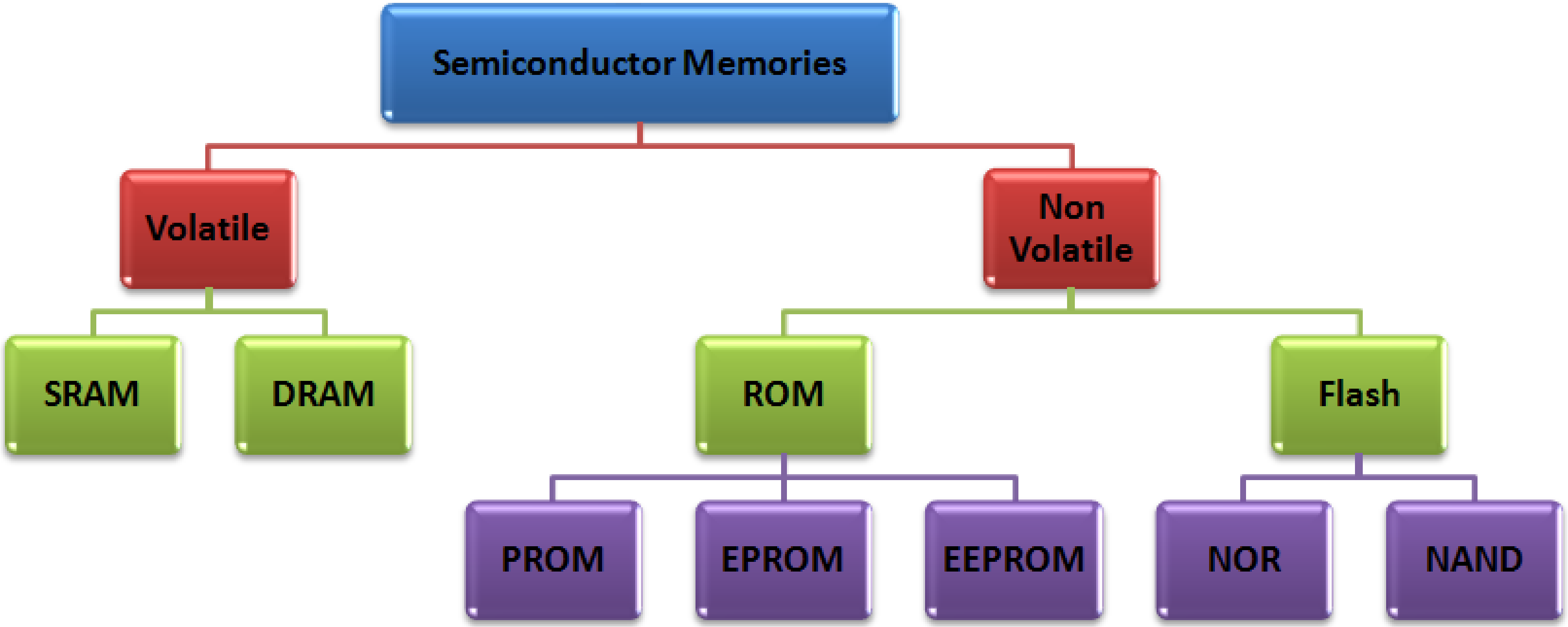
2) Memory access time which is the time required to store or retrieve data from the array

   - Determines the memory speed

3) Static and dynamic power consumption of memory

   - Due to the increasing importance of low – power applications

Overview of semiconductor memory types



Another type of Volatile Memory – CAM (Content Addressable Memory)

Other types of Non-Volatile Memories – FeRAM, MRAM, STTRAM, PCM

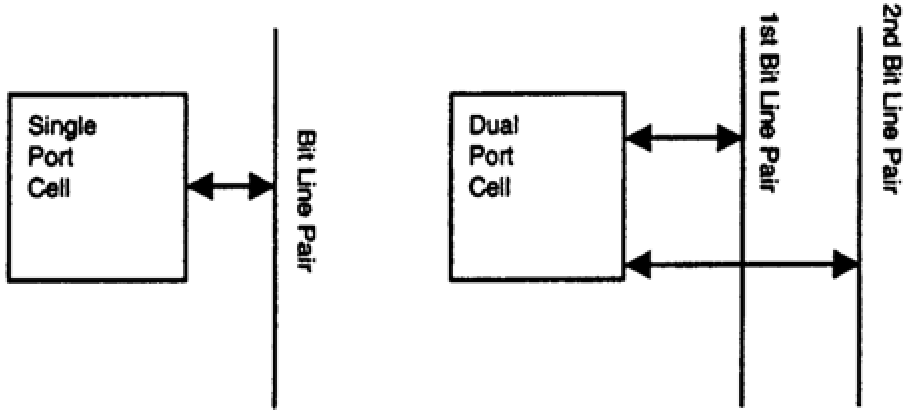
Classification of Semiconductor Memory

- Classification based on
  - ✓ **Input/output Architecture** – Single port /Multiple port

    1) **Single** – port memories: One port for both read and write operations

    2) **Multiport memories:** Separate ports for read and write operation. High performance but consumes high power.

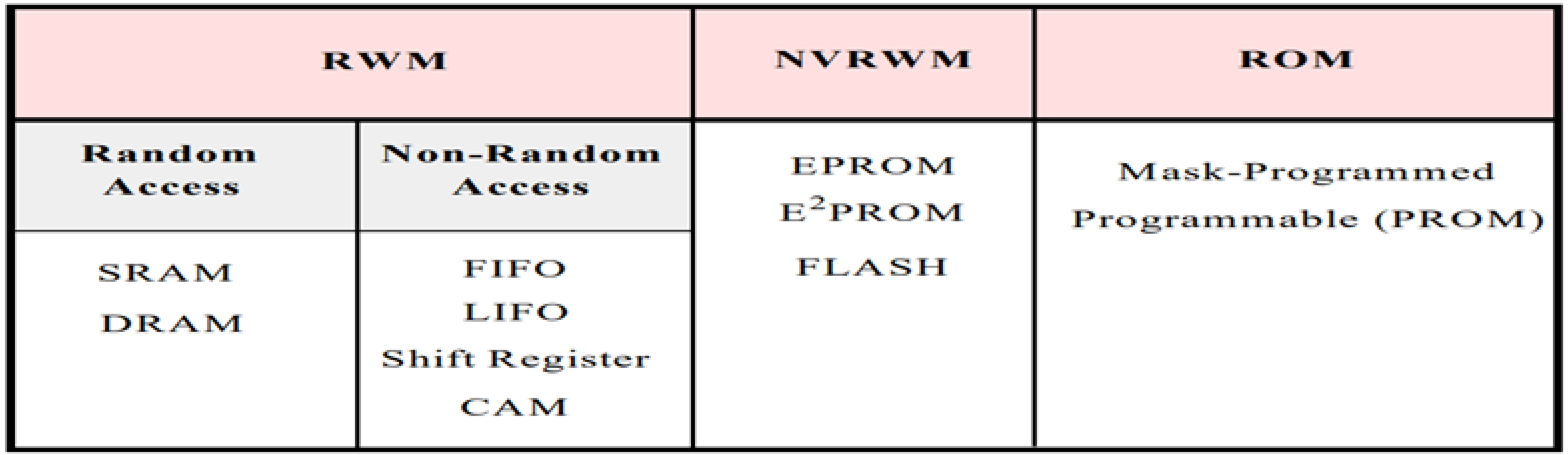
    Ex: 1 Read port + 1 Write port , 2 Read ports + 1 Write port
  - ✓ **Application** - Embedded / Secondary Memory.
    - **On-Chip Memories** i.e., Integrated on same die as logic functionality and is called embedded Memory.
    - **Off-Chip Memories** i.e., Not Integrated in the same die as logic functionality – Large Capacity Memories



Classification of Semiconductor Memory

- Semiconductor Memories are classified on the basis of
  - Functionality
  - Access Pattern required - Order in which data can be accessed.
  - Nature of the Storage Mechanism – Primary and Secondary storage
- Classification based on Functionality as Read Only Memory (ROM) and Read-Write Memory (RWM)

Volatile Memories	Non Volatile Memories

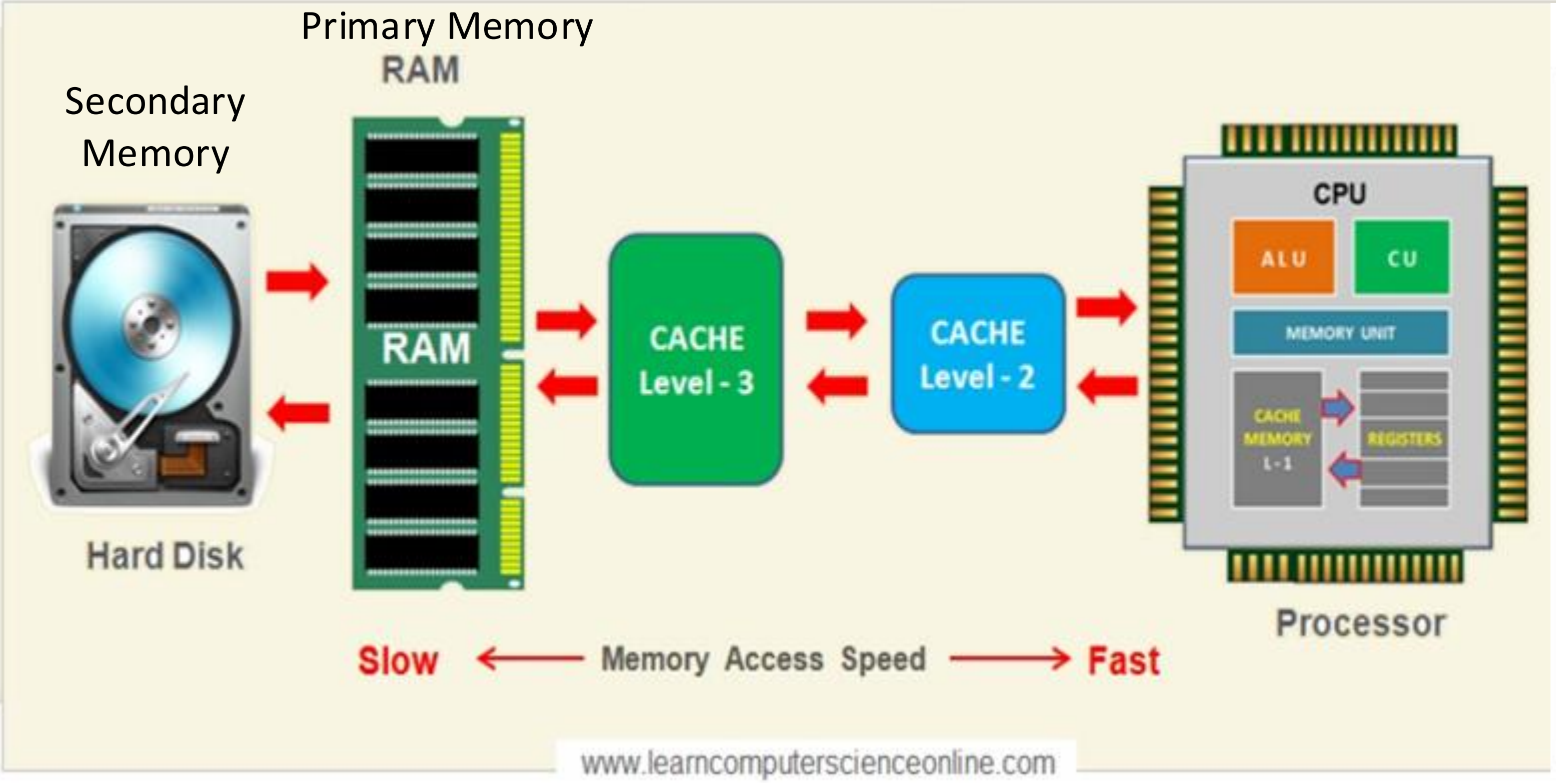


RWM		NVRWM	ROM
Random Access	Non-Random Access	EPROM E^2PROM FLASH	Mask-Programmed Programmable (PROM)
SRAM DRAM	FIFO LIFO Shift Register CAM		

Classification of Semiconductor Memory

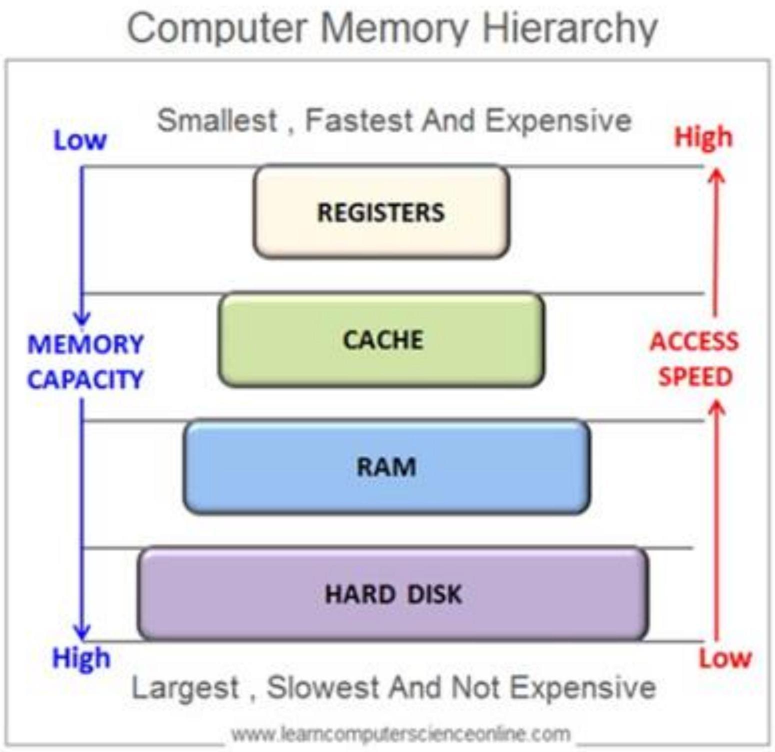
- Type of Memory preferred for the Application is function of the
  - ✓ Semiconductor Memories are classified on the basis of
    - Functionality
    - Access Pattern required - Order in which data can be accessed.
    - Nature of the Storage Mechanism
- Classification based on Access Pattern – Based on Order in which data can be accessed as
  - **Random Access Memory (RAM)** – It means Memory locations can be read or written in random order : Example – SRAM and DRAM
  - **Non Random Memory Access** – These type restrict the order of access which results in faster access times, smaller area Example: Serial Memories –FIFO, LIFO , Shift Registers and Video Memories and Content Addressable Memories (CAM)

Memory Hierarchy



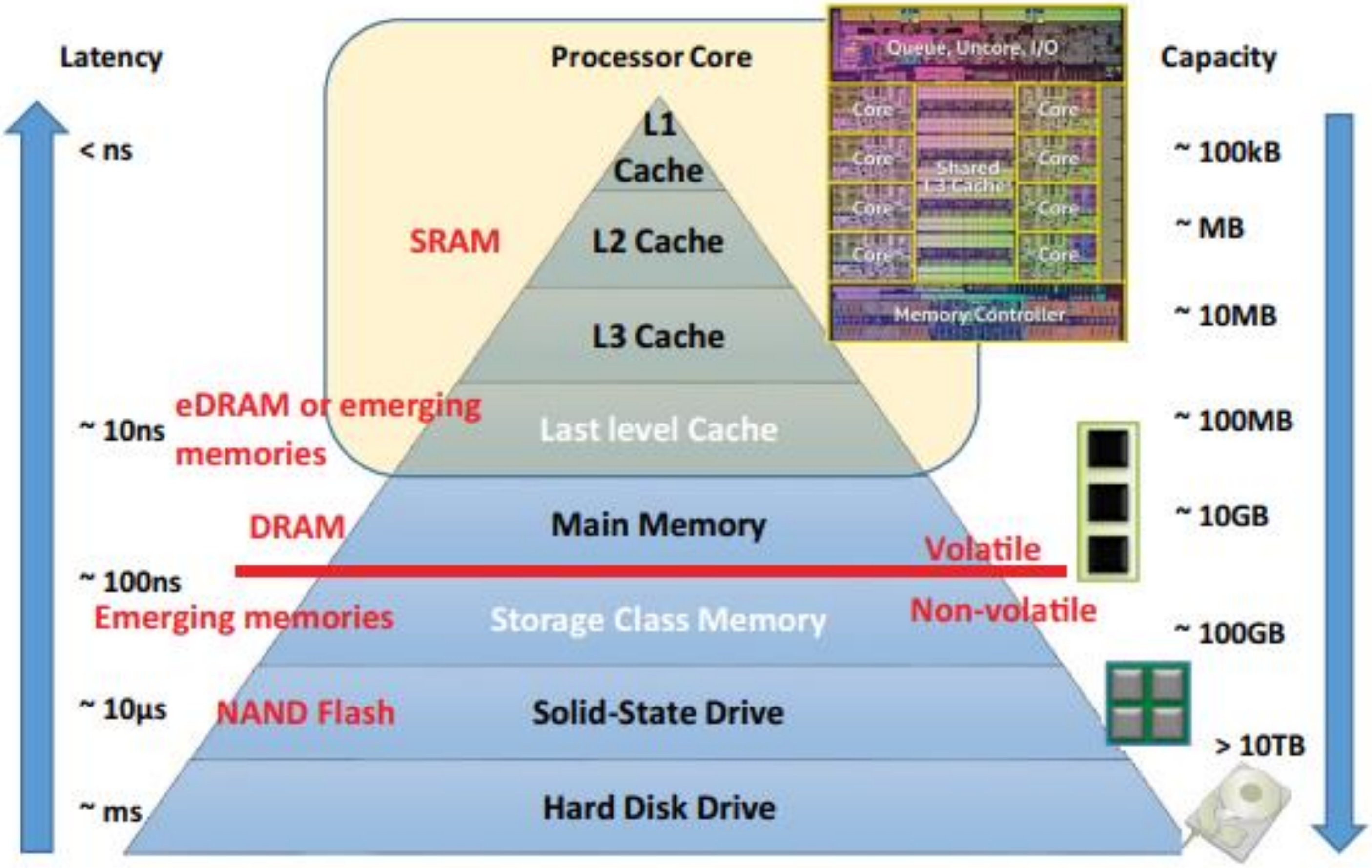
Primary Memory

Secondary Memory



Computer Memory Hierarchy
Smallest , Fastest And Expensive
Low
High
REGISTERS
CACHE
RAM
HARD DISK
MEMORY CAPACITY
ACCESS SPEED
High
Low
Largest , Slowest And Not Expensive
www.learncomputerscienceonline.com

Memory Hierarchy



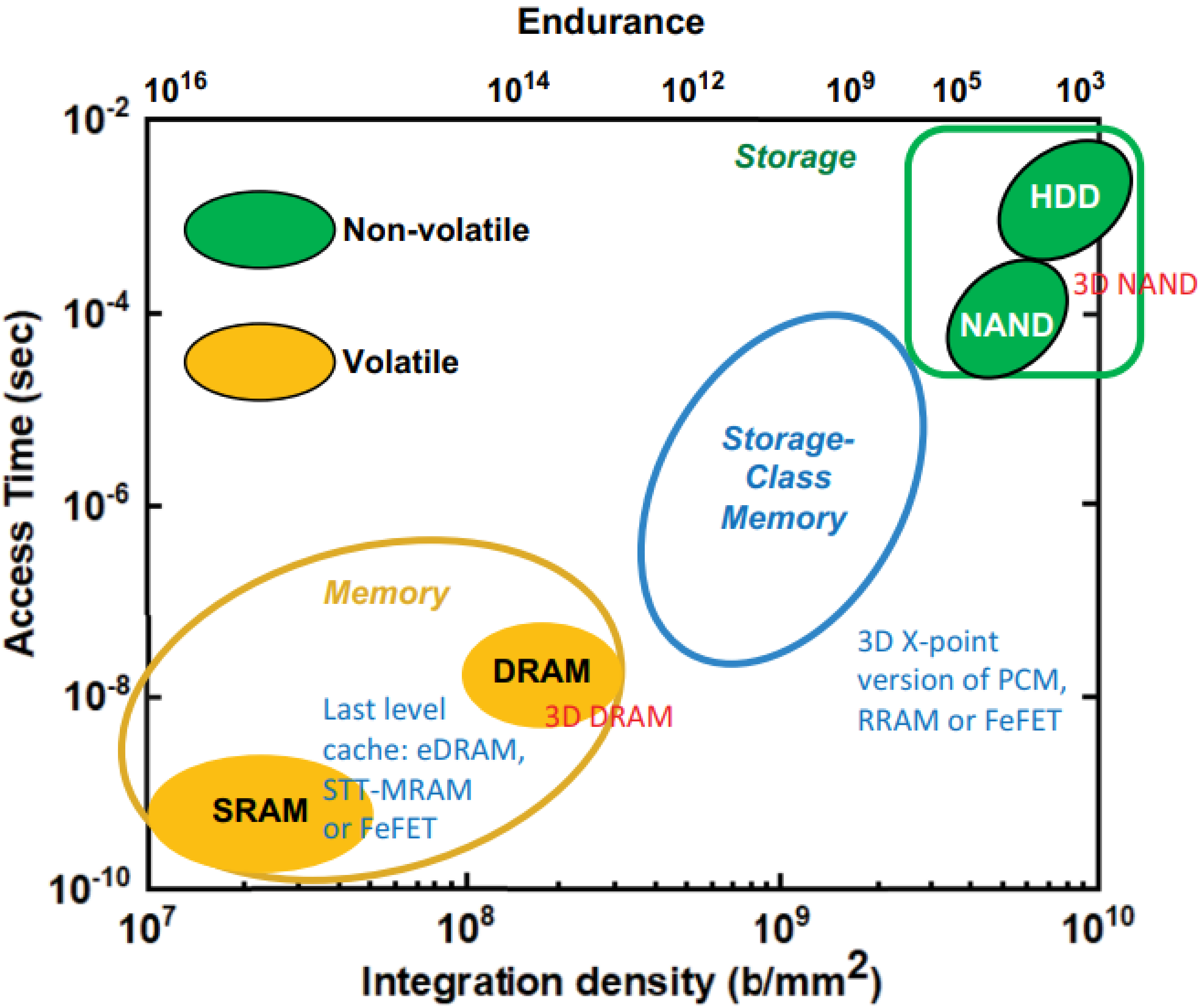
Introduction Terms to remember

- Volatile (Working) and Non-volatile (Storage) memories
- Cache and multilevel cache
- Off chip memory and on chip memory
- Storage class memory
- 3D memory or stacked memory

Performance parameters

- Access time: Refers to read/write access time.
- Cycle endurance: Specifies how many times the memory device could be written into before it fails
- Integration density: No. of bits stored/unit area

Performance parameters



Performance parameters

Access time vs capacity:

- L1 cache could be accessed in sub-ns and has a capacity of ~100 kB
- L2 cache could be accessed in 1–3 ns and has a capacity of ~1 MB
- L3 cache could be accessed in 5–10 ns and has a capacity of tens of MB.
- Main memory – DRAM which is accessible in 10s of ns capacity - GB
- NAND flash-based SSD accessed in 10s of µS and capacity of GB – TB
- Magnetic HDD accessed in ms and has a capacity of 10s of TBs

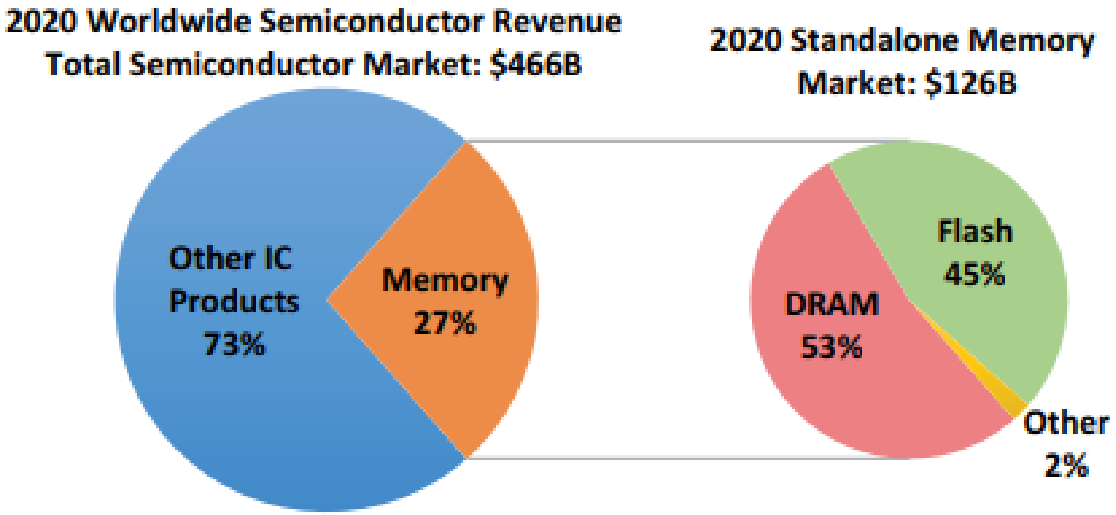
Performance parameters

Endurance:

- As working memories, SRAM or DRAM generally have >10^{16} endurance in a 10-year lifetime.
- NAND Flash is less often written with 10^3–10^5 endurance
- Storage-class memory is expected to have 10^9–10^{12} endurance.

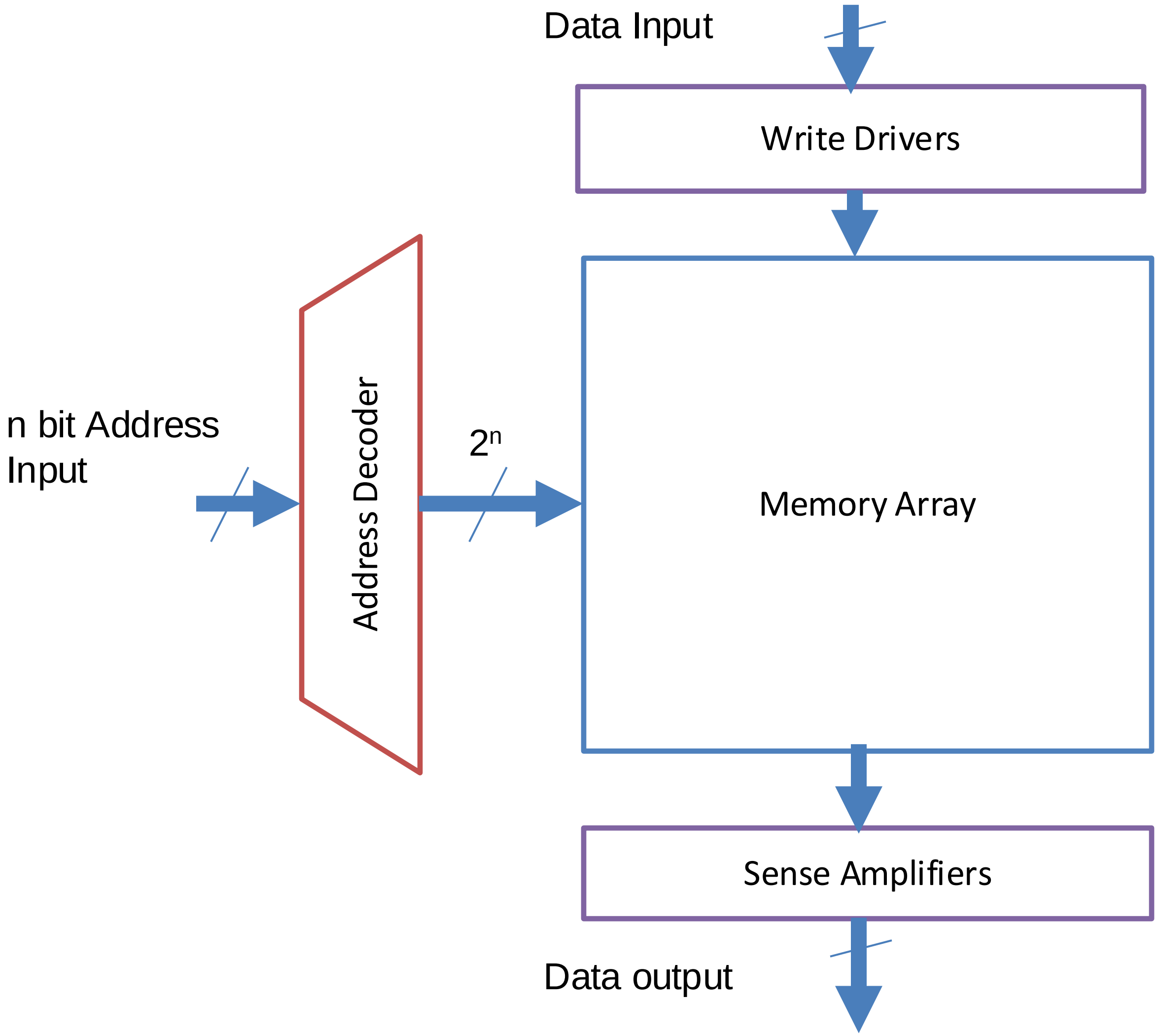
Semiconductor memory industry

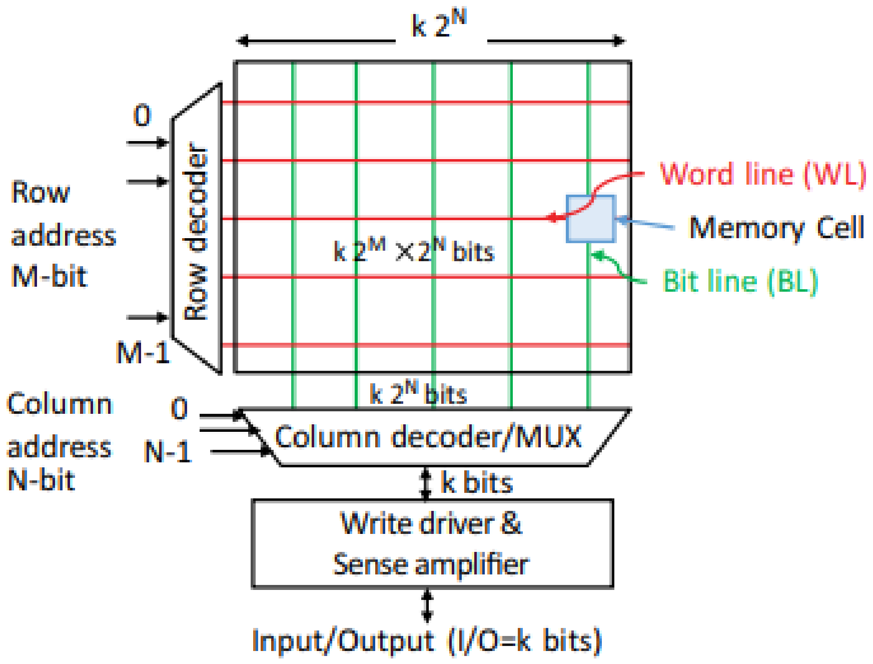
- Memory market 126 billion US dollar
  - From DRAM and NAND Flash
- Memory companies are
  - Samsung
  - Micron
  - SK Hynx
  - Intel
  - Western Digital



Memory Architecture and Building blocks

Basic Elements in a Memory System





Simple Block Diagram of Memory

Address decoders select the location into which either data is written into or read from.

The sense amplifiers speed up the read operation.

Due to the ever-increasing demand for larger data storage, the capacity of single-chip semiconductor memory approximately doubles every 2 years.

Memory Array Organisation

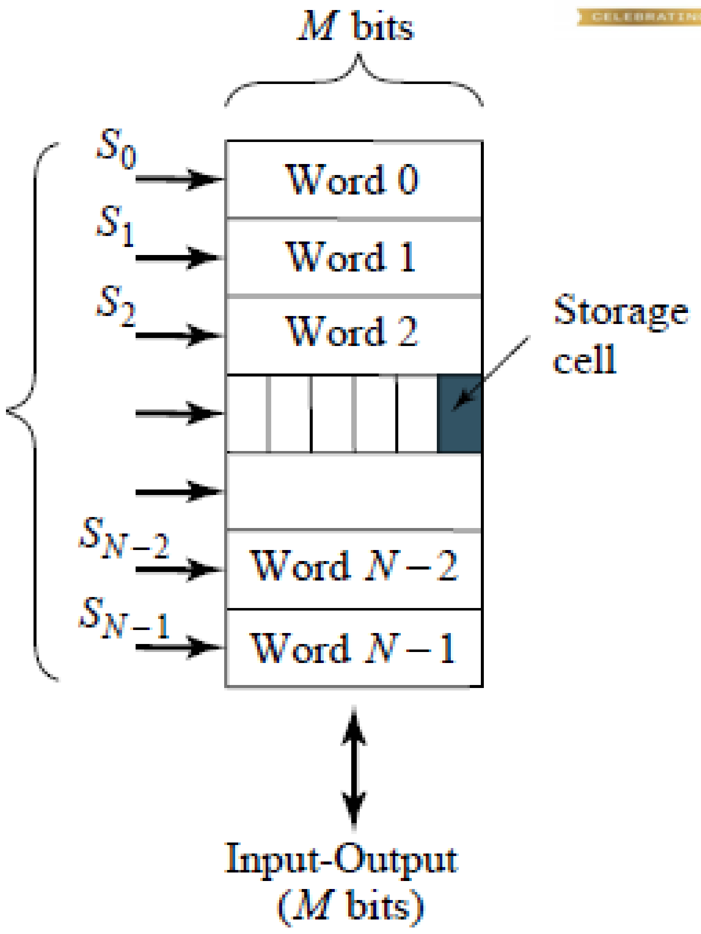
Simple Architecture: N x M memory

- N – words of size M – bits each
- Select line S is used to select each location for reading or writing.
- The number of select lines = number of locations

  = number of words
- Simple approach and ideal for small memories
- Not suitable for large memories.

  Ex: 1 MB memory would require ....

  2^{20} select lines

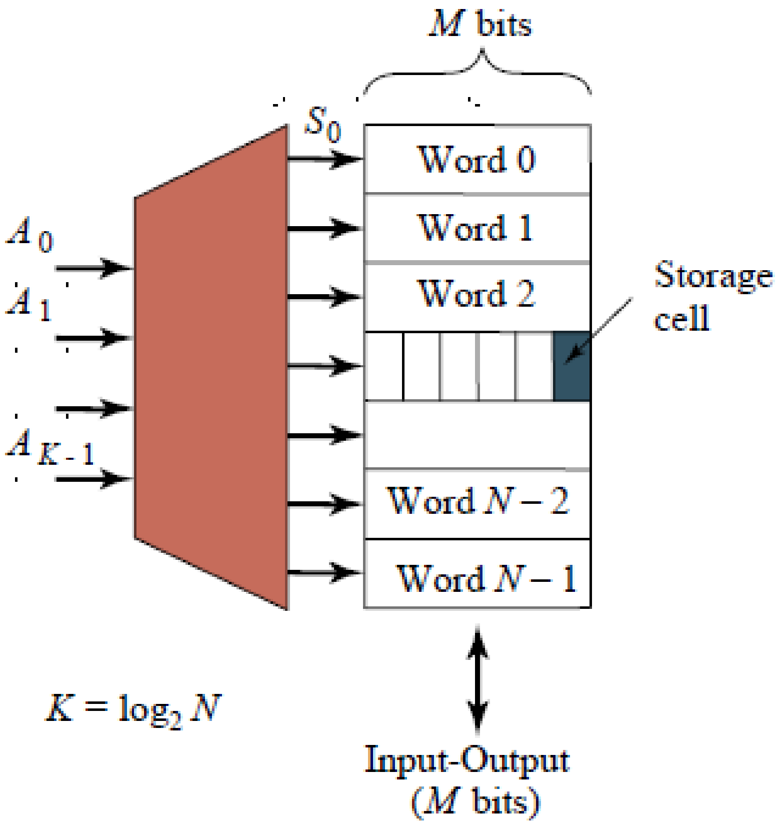


Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

Memory Array Organisation

Memory architecture with decoders: N x M memory

- Address decoder is used to reduce the number of select lines.
- Binary encoded address word of K – bits, $A_{K-1}A_{K-2}...A_1A_0$ is translated by the decoder into $N = 2^K$ select lines.
- Reduces the number of external address lines from 1 Million to 20 for a 1 MB memory which eliminates the wiring and packaging problems.

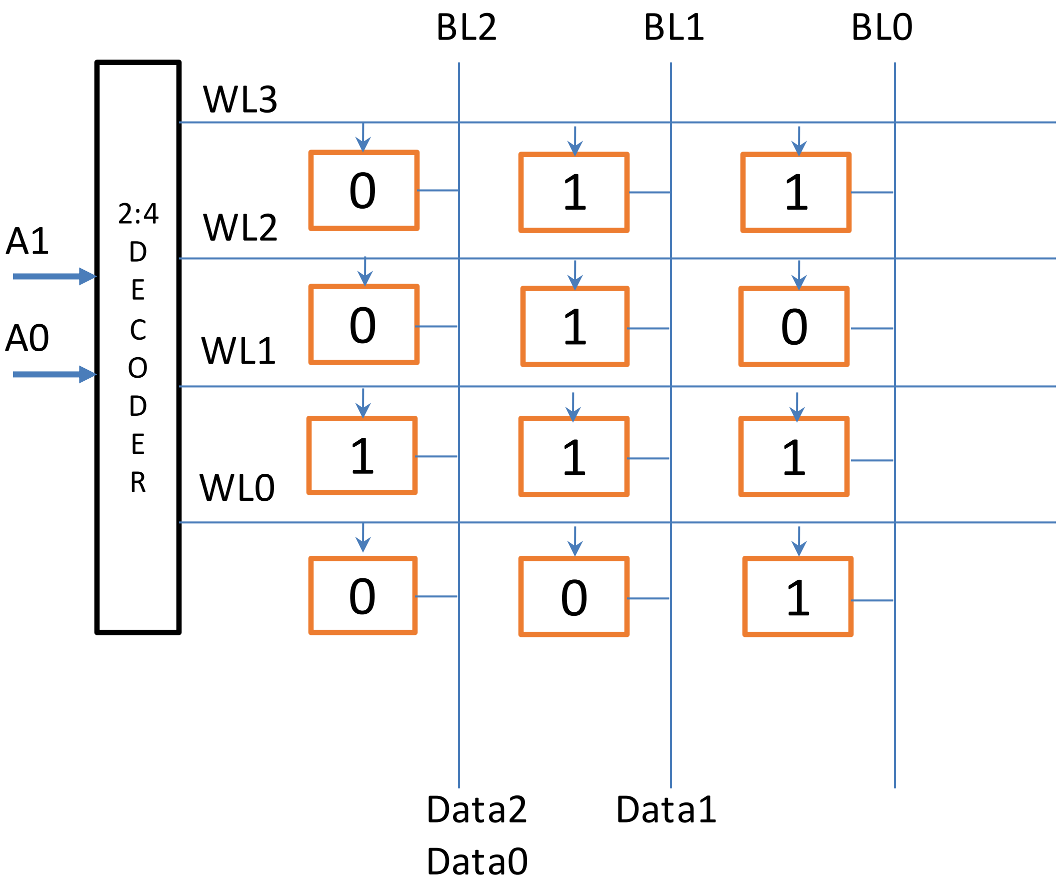


1MB memory has 1 Mega = 2^{20} locations, each storing a byte data

Diagrams courtesy "Digital Integrated Circuits: A Design Perspective" by Jan Rabaey

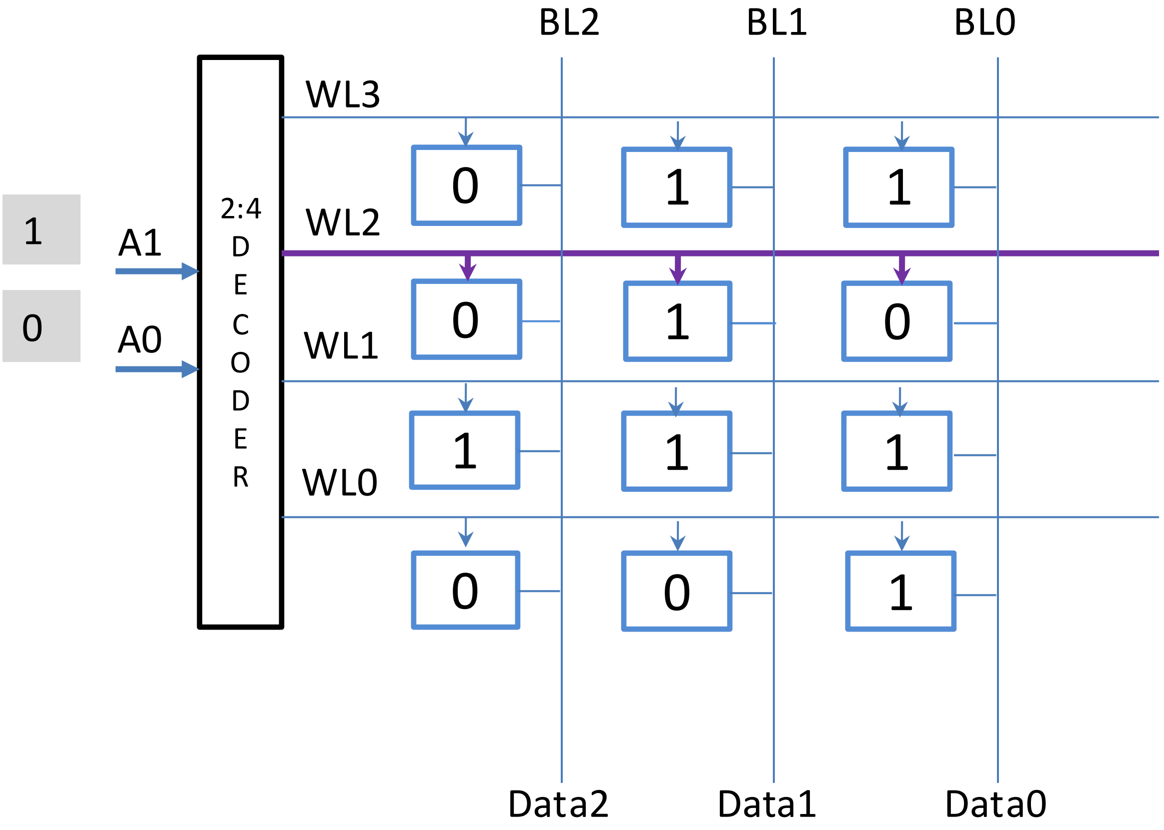
Memory Architecture and Building blocks

A simple N x M Memory Array with Address Decoder



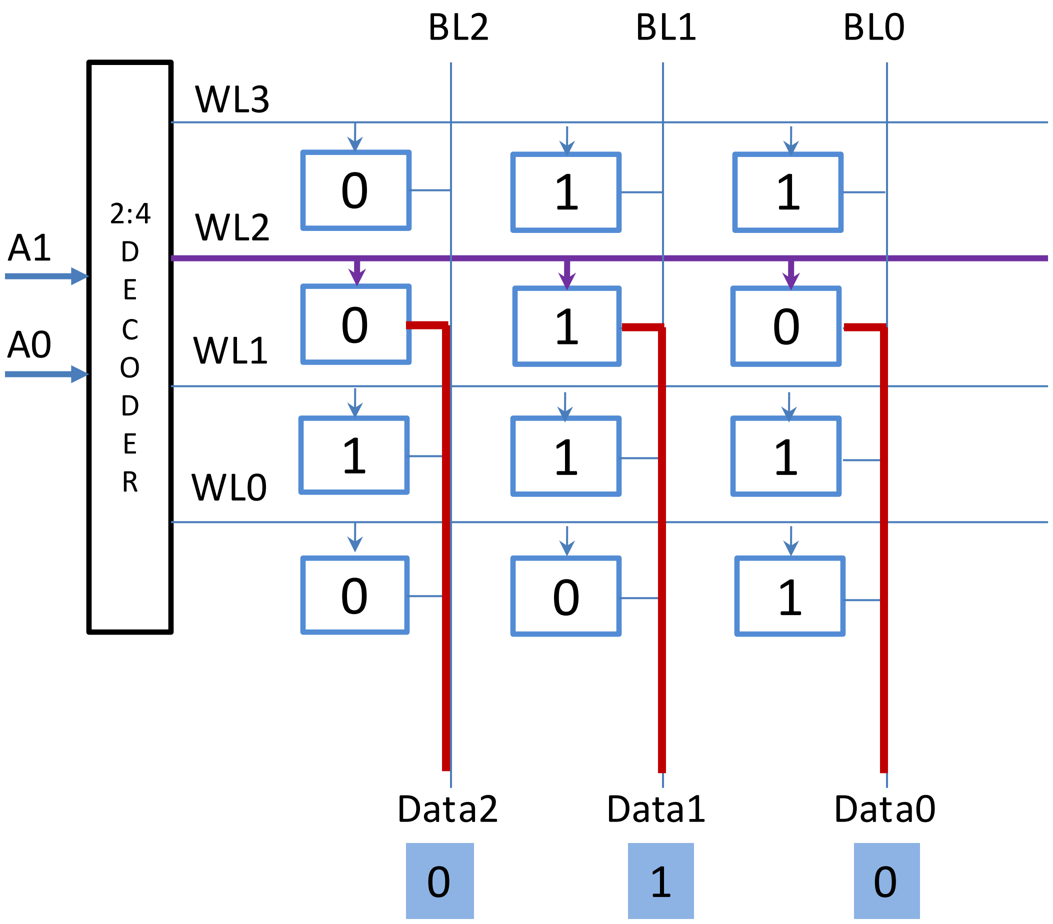
Memory Architecture and Building blocks

A simple N x M Memory Array with Address Decoder



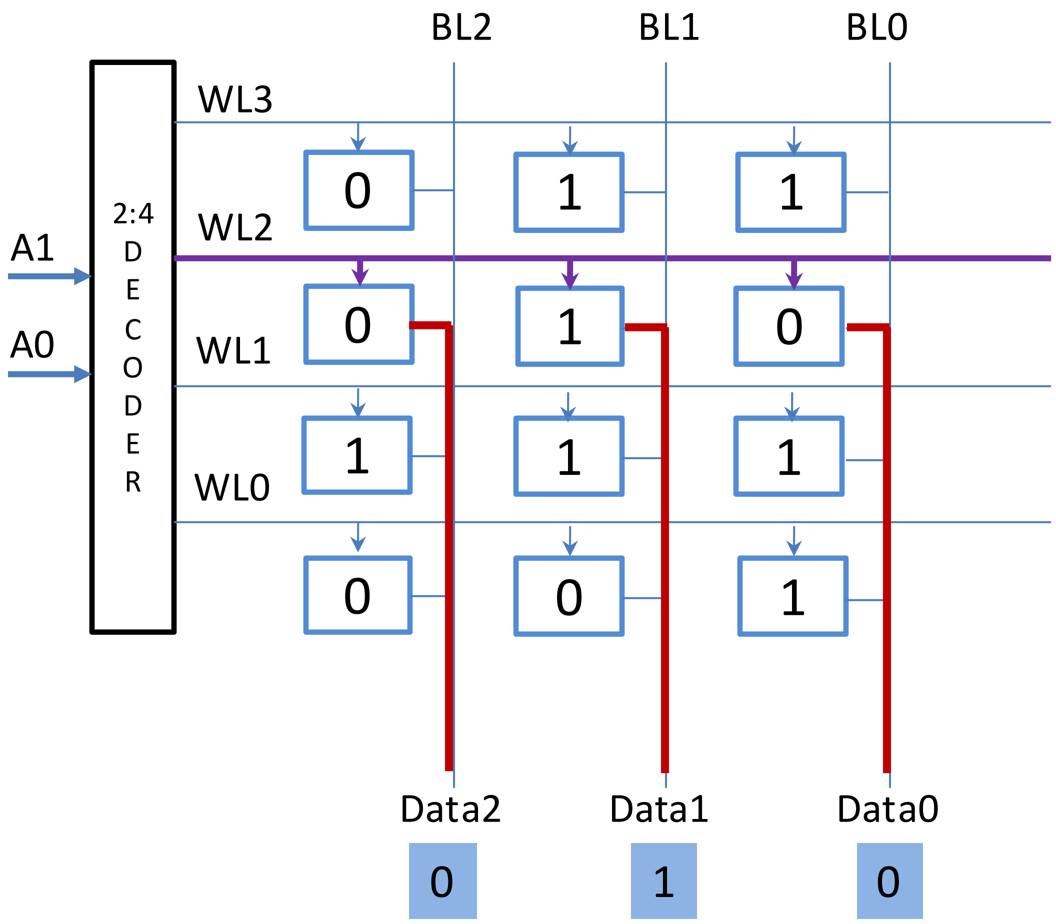
Memory Architecture and Building blocks

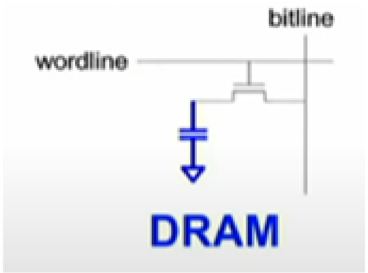
A simple N x M Memory Array with Address Decoder

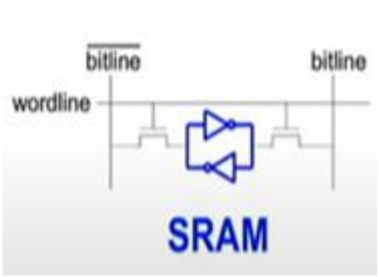


Memory Architecture and Building blocks

A simple N x M Memory Array with Address Decoder







Memory Array Organisation

Disadvantages:

Decoder resolves Select line problem but does not address the issue of memory aspect ratio.

The height of 1 MB memory is 2^{17} times larger than its width (2^{20} / 2^{3})

This results in a design that cannot be implemented.

Also, the design is extremely slow since the vertical wires connecting the storage cells to the input / outputs become excessively long increasing the read and write delays.

The delay of an interconnect line increases at least linearly with its length.

Memory Array Organisation

Array – Structured Memory Architecture

To address the problem of aspect ratio, memory arrays are organised so that vertical and horizontal dimensions are of the same order of magnitude and the aspect ratio approaches unity.

Multiple words are stored in a single row and are selected simultaneously.

To route the correct word to the I / O terminals, a column decoder is used.

Array – Structured Memory Architecture

The address word is partitioned into

Column address $A_0 – A_{K-1}$ and

Row address A_K to A_{L-1}

Number of rows = 2^{L-K}

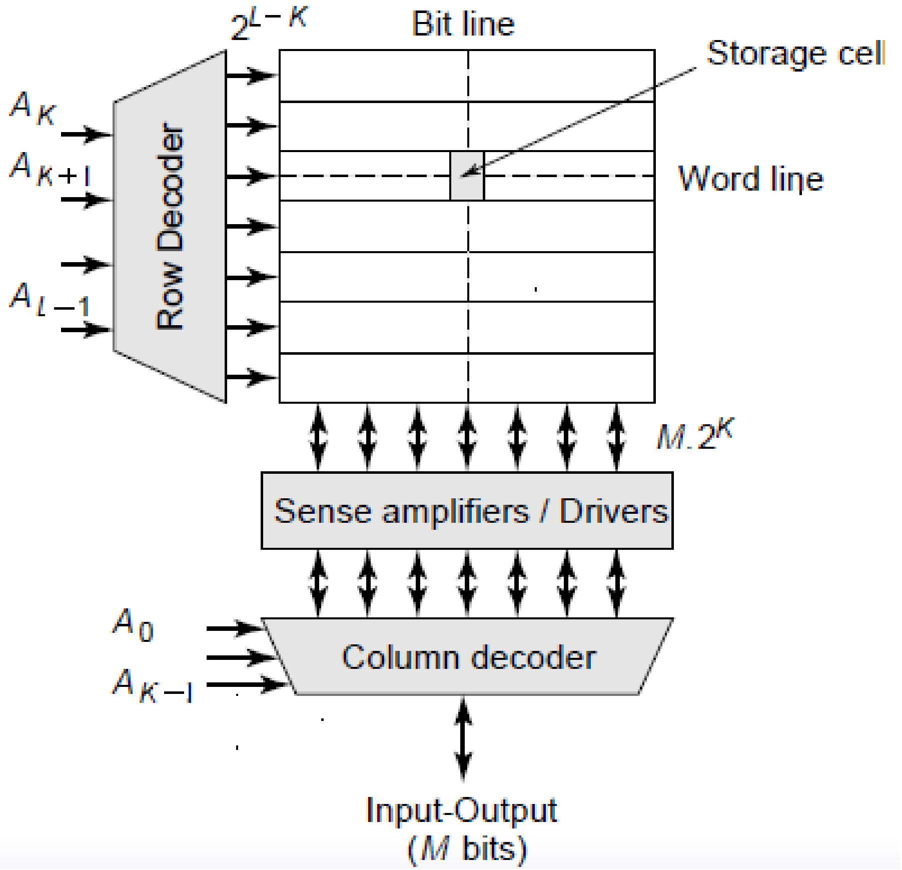
Number of columns = 2^K with each column

storing M – bit data

The row address enables one row of memory

for R / W, while the column address picks one

particular word from the selected row.

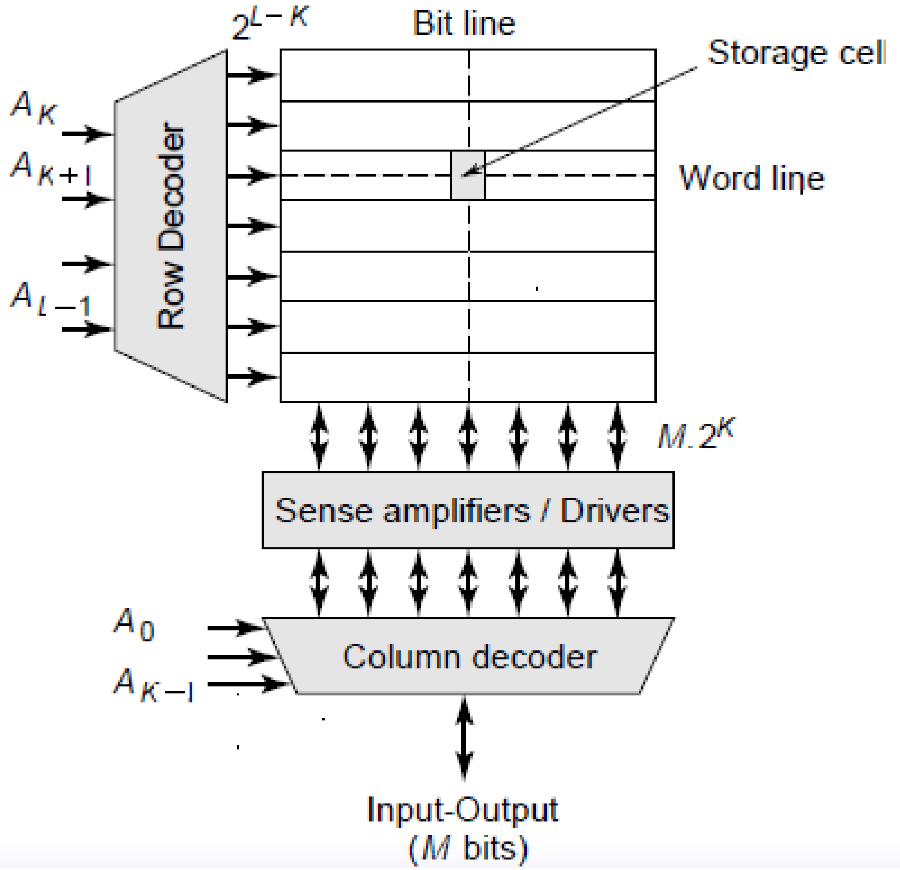


Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

Array – Structured Memory Architecture

The horizontal select line that enables a sir row of cells is called a Word Line

The vertical line that connects the cells in a single column to the I/O circuitry is called Bit Line



Array – Structured Memory Architecture

Ex: Organisation of 1 MB memory

Total data = 2^{20} x 2^{3} = 2^{23} bits

This is organised as an array of 4096 x 2048 cells = 2^{12} x 2^{11} cells

= 2^{23} cells

Each of the 4096 rows store 256 8-bit words.

The number of row address lines required to access 4096 (2^{12}) rows is 12.

The number of column address lines required to access 256 (2^{8})words in a column is 8.

Hence the total number of address lines required = 20

Memory Array Organization

For example if the Address generated is 0x 15F3D , Identify the Row and Column Address

A_{L-1} A_K	A_{K-1} A_0
A_{19} A_8	A_7 A_0

A_{19} A_{18} A_{17} A_{16}	$A_{15}A_{14}A_{13}$ A_{12}	A_{11} A_{10} A_9A_8	$A_7A_6A_5A_4$	$A_3A_2A_1A_0$
0 0 0 1	0 1 0 1	1 1 1 1	0 0 1 1	1 1 0 1

Row Number = 0x15F = 351

Column Number = 0x3D = 61

It means , Memory byte (word) being accessed is 61st byte of 351st Row

The array – structured architecture works well for memories up to a **range of 64K bits to 256K bits.**

Hierarchical Memory architecture

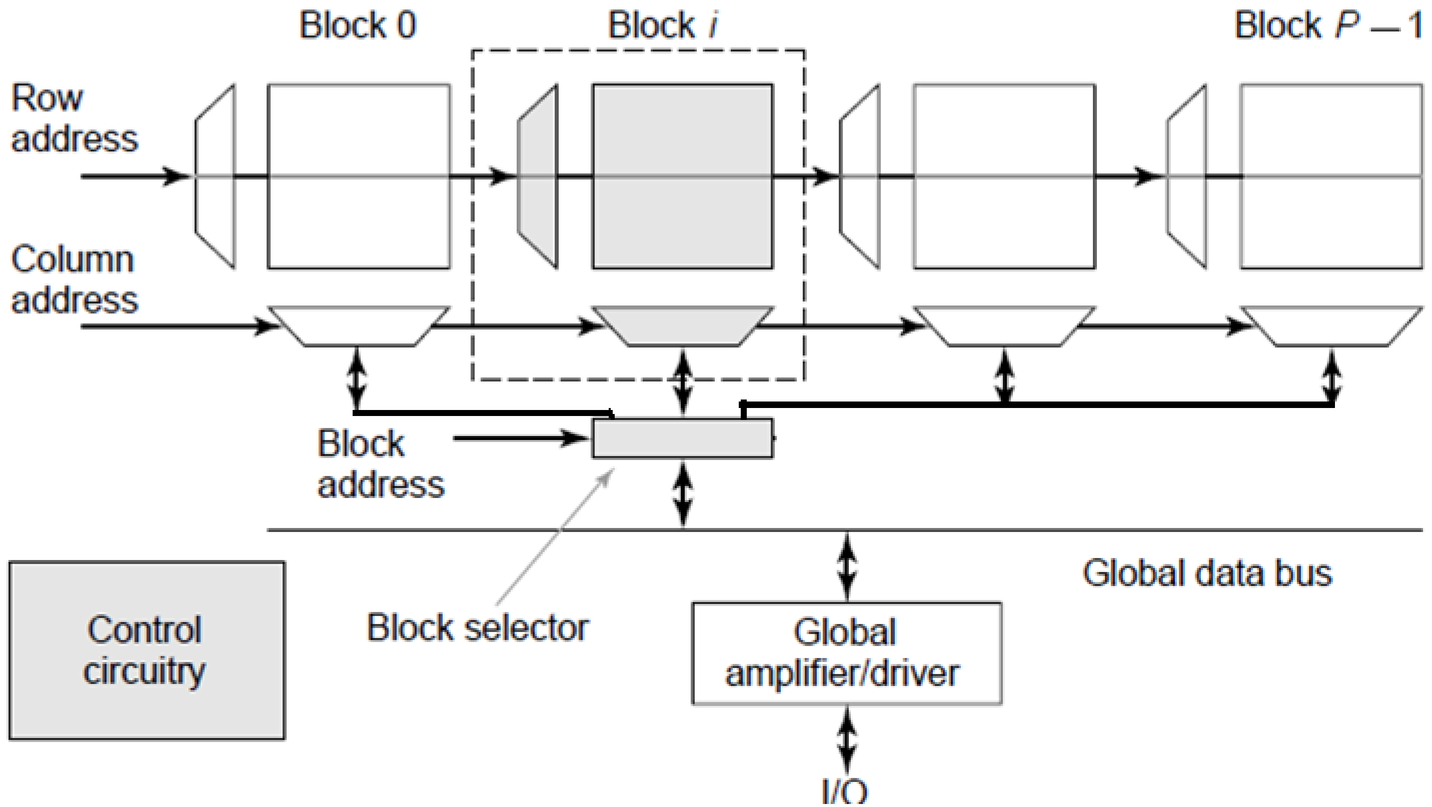
The array – structured architecture works well for memories up to a range of 64K bits to 256K bits.

Larger memories suffer from speed degradation as the length, capacitance and resistance of the word and bit lines become excessively large.

Hence large memories are partitioned into P – smaller blocks.

The composition of each of the blocks is identical to the array-structured memory.

Hierarchical Memory architecture



An extra address word called the block address, selects one of the P blocks to be read or written.

Diagrams courtesy "Digital Integrated Circuits: A Design Perspective" by Jan Rabaey

Hierarchical Memory architecture

A word is selected on the basis of row and column addresses that are broadcast to all the blocks.

This approach has dual advantage

1. The length of the local word and bit lines – that is, the length of the lines within the blocks – is kept within bounds, resulting in faster access time.
2. The block address can be used to activate only the addressed block. Non-active blocks are put in power-saving mode with sense amplifiers and decoders disabled. This results in a substantial power saving.

Hierarchical Memory architecture

Ex: Organisation of 4 Mbit SRAM into 32 blocks with a data size of 8 bits

4 Mbit = 2^{22} bits

Organised into 32 blocks, each of which contains 128 Kbits = 2^{17} bits.

To make the aspect ratio close to unity, number of rows can be 2^{9} and number of columns = 2^{8}

Hence, No. of row address lines = 9

Data size = 8 = 2^{3} bits. No. of column address lines = 5

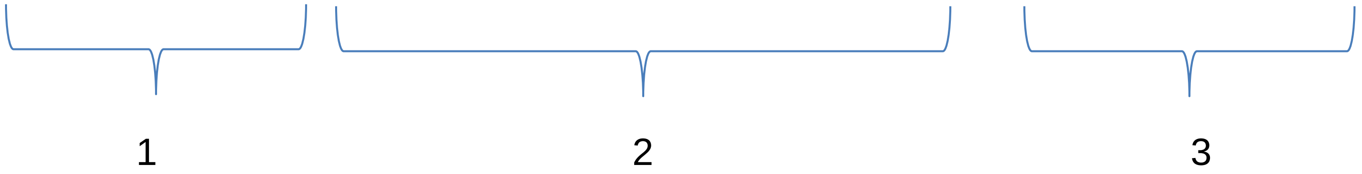
No. of block address lines = 5

Total number of address lines = 19

Memory Array Organization

If the address is 0x51EC7

A_{18} A_{14}	A_{13} A_{5}	A_{4} A_{0}
Block Address	Row Address	Column Address
1 0 1 0 0	0 1 1 1 1 0 1 1 0	0 0 1 1 1



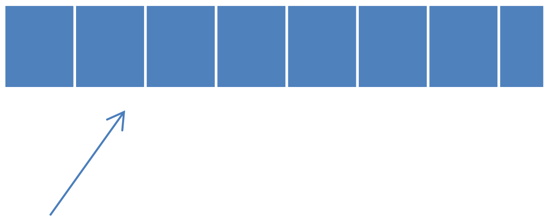
1 2 3

1. Block Address = 0x14 = 20
2. Row Address = 0x0F6 = 246
3. Column Address = 0x07 = 07 and

It means In block Number 20, Row Number 246, 7th Column (byte) is accessed

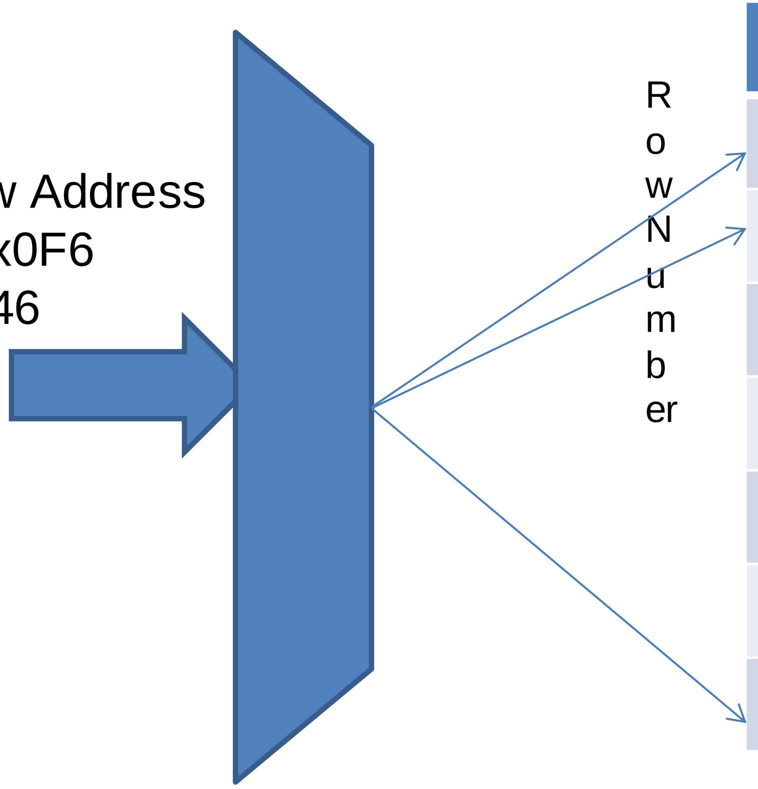
Memory Array Organization

Block of 128Kbits = 16384 bytes



Column Number

Row Address
= 0x0F6
= 246



R\C	31	..	07	..	2	1	0
0							
1							
2							
..							
246			9D				
..							
511							

Column Address = 0x07 = 07

Address=0x51EC7

1. Block Address = 0x14 = 20
2. Row Address = 0x0F6 = 246
3. Column Address = 0x07 = 07

Assignment

1. Organise a 16M Byte Memory into 64 blocks. The data has a word size of 8 bits.

2. Organise a 64 KByte Memory into an array . The data has a word size of 8 bits.

Reference material: Digital Integrated circuits: A Design Perspective, Rabaey, Chandrakasan and Nikolic, 2nd edition.

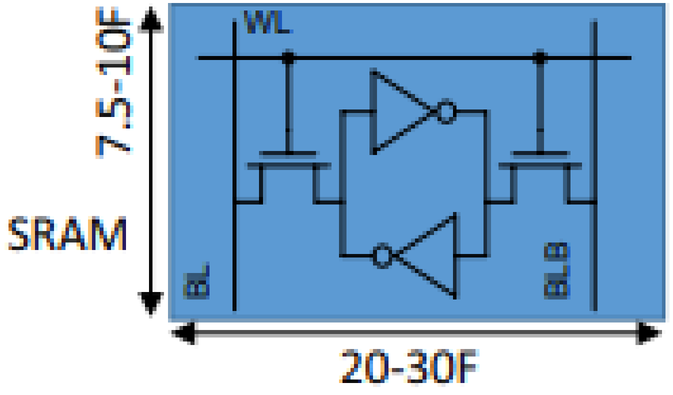
Memory cell size and equivalent bit area

- Each type of memory requires specific number of devices to represent one bit
  Ex SRAM 6 transistors, DRAM one transistor and one capacitor
- Integration of devices depends on technology and integration defines area required
- Also higher integration means lower area to represent a bit
- Lower cost per bit area is essential for large capacity memories

- Metrics used is F^2, which is memory cell size in a particular technology

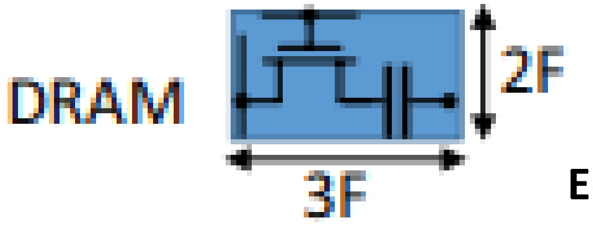
Memory cell size and equivalent bit area

- Equivalent bit area



150 F^2 to 300 F^2

Equivalent bit area



6 F^2

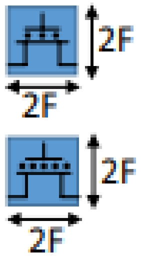
Equivalent bit area

- DRAM has lesser equivalent bit area hence can be used for larger density requirements

Memory cell size and equivalent bit area

- Equivalent bit area

2D NAND Flash (SLC)

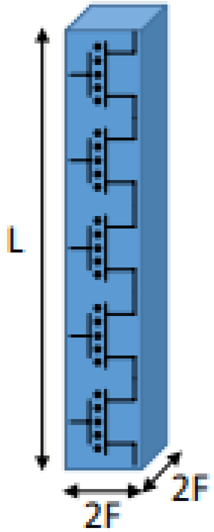


$4\ F^2$

2D NAND Flash (MLC=n-bit)

$4\ F^2/n$

Equivalent bit area



3D NAND Flash (MLC=n-bit and L layers)

$4\ F^2/n/L$

Equivalent bit area

- SLC – Single layer cell
- MLC – Multi layer cell
- N bit – number of bits in multi layer cell

Memory array’s area efficiency

- It defines area occupied by memory array in a given memory chip area
- Higher the area efficiency lower the cost per bit
- NAND flash memory has highest area efficiency about 70 to 80 %
- DRAM has about 60 to 70 % area efficiency
- SRAM has about 50%

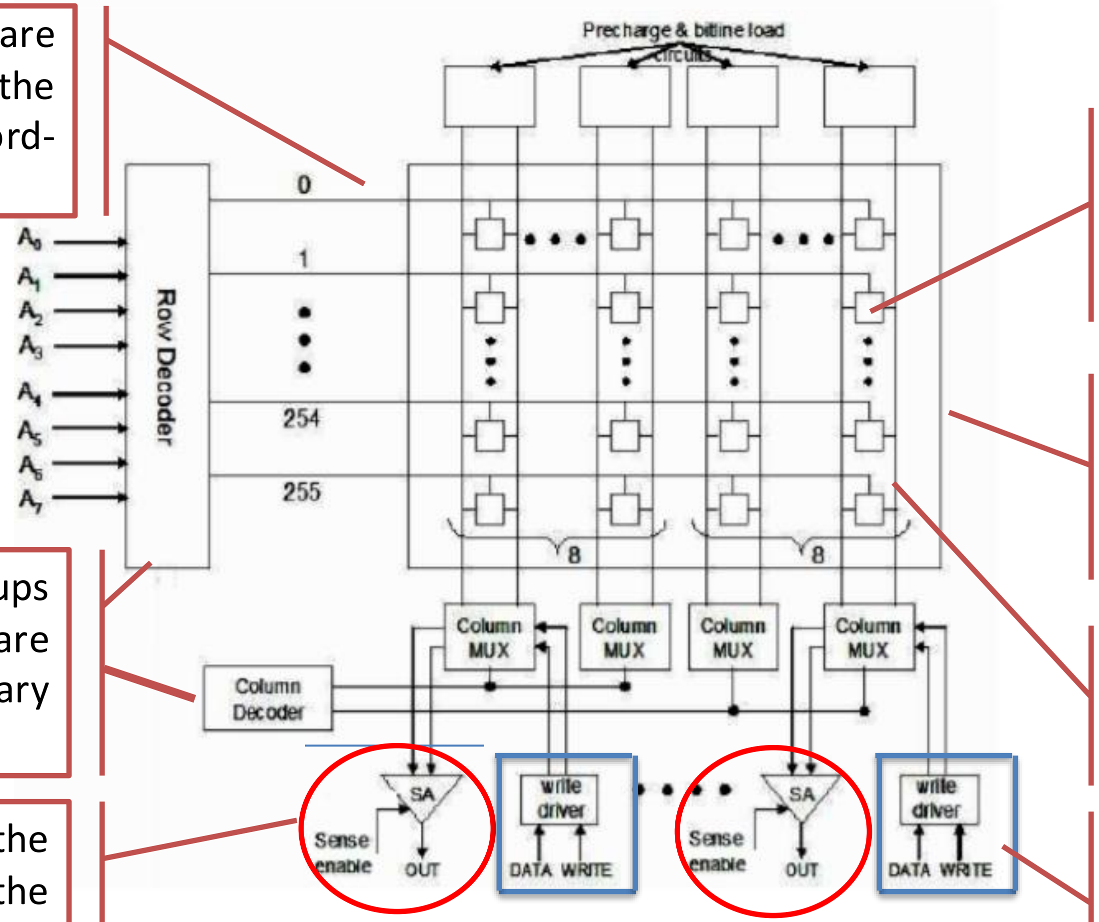
SRAM Memory Array Organisation

SRAM Memory Architecture

The horizontal lines, which are driven only from outside the storage array, are called word-lines(WLs)

The row and column (or groups of columns) to be selected are determined by decoding binary address information.

A sense amplifier is part of the read circuitry. It amplify the small voltage swing on bit lines to recognizable logic levels



SRAM Cell holding bit of data as 1 or 0. A cell is accessed for reading or writing by selecting its row and column.

The storage array of Memory Cells circuits are arranged to share **connections in horizontal rows** and **vertical columns.**

The vertical lines, along which data flow into and out of cells, are called bit lines.

Write Driver Circuits forces one of the **bit-lines to zero**, while maintaining the pre-charged value on the other bit-line.

Random Access Memories

SRAM and DRAM are 2 types of Volatile Memories.

Data is stored as long as power is applied

SRAM

- Earliest memories produced
- Large in size (6 – 10 transistors / cell)
- Capacity is 1/4th of DRAM
- Fast
- Differential output
- Used in high – speed applications such as L1 caches of microprocessors.

DRAM

- Small in size (1- 3 transistors / cell)

- Cost / bit is 1/4th that of SRAM

- Slower

- Single – ended output

- Periodic refresh required

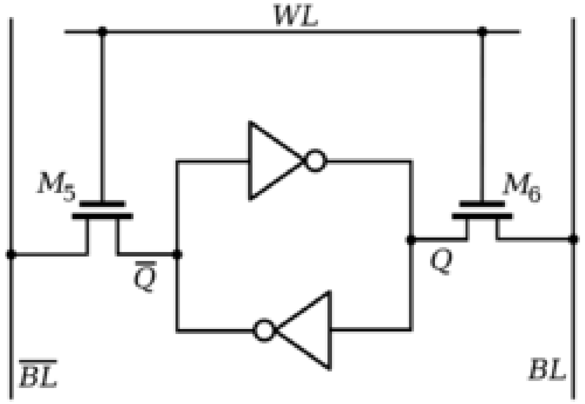
- Consume more power than SRAM

- Used in high – density applications

SRAM Cell

6 – transistor SRAM Cell

Made of 2 inverters connected back – to – back with 2 access transist

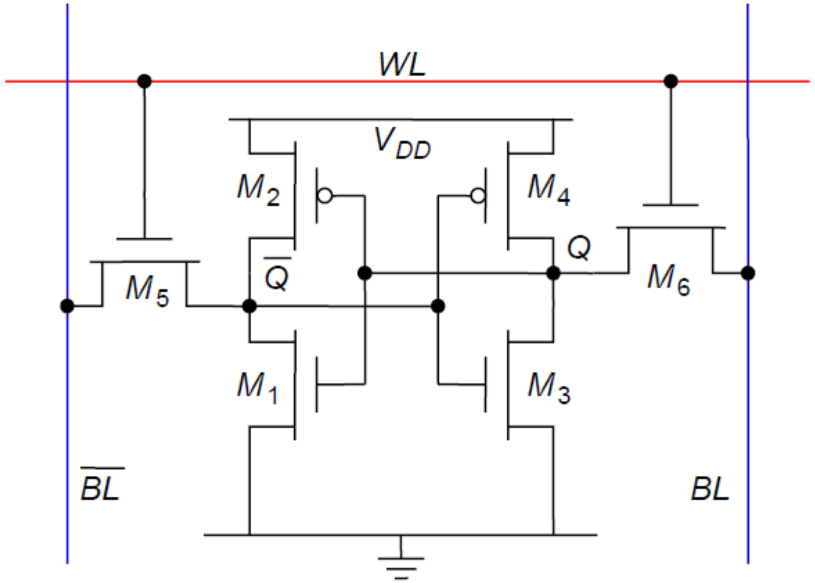


It is similar to the static SR latch. It requires 6 transistors per bit.

Q is the data stored in the cell.

Access to the cell is enabled by the word line, WL, which controls the 2 pass transistors, M_5 and M_6

6 – transistor SRAM Cell



Transistor pairs M_1 – M_2 and M_3 – M_4 form the 2 inverters.

2 Bit Lines are used to transfer data and its complement to and from the cell.

Providing both polarities improves the noise margins during read and write operations

Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

Other configurations of SRAM Cell

6T SRAM, while simple and reliable, consumes a substantial area.

Placing the 2 PMOS transistors in the N-well significantly contributes to the area.

Large memories have other cell structures with special devices.

Some of the other SRAM configurations

1) 4 – Transistor SRAM Cell

2) 7T SRAM

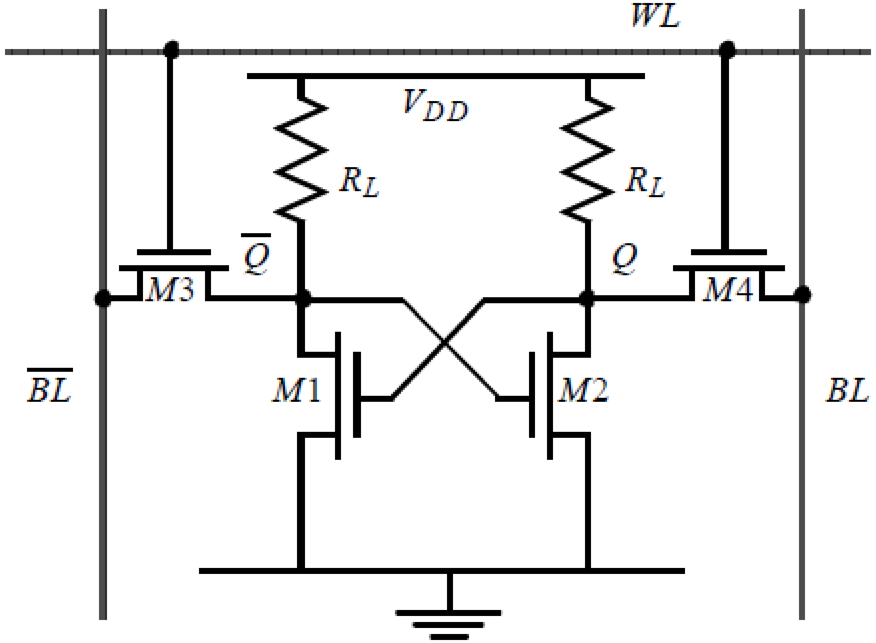
3) 8T SRAM

4) 9T SRAM

Other configurations of SRAM Cell

4 – Transistor SRAM Cell

Also called the Resistive load SRAM



The cross-coupled CMOS inverter pair is replaced by a pair of resistive-load NMOS inverters.

Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

Other configurations of SRAM Cell

4 – Transistor SRAM Cell

Advantage:

Cell size is reduced by approximately 1/3rd for a 1 Mbit SRAM

Disadvantage:

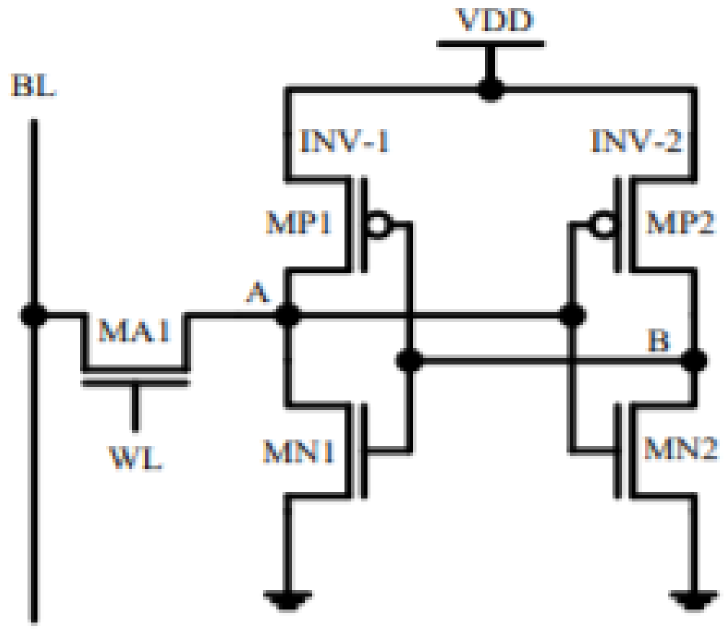
Static power dissipation when the pull – down transistors M_1 and M_2 turn ON.

To reduce power dissipation, special resistors are used that are made as large as possible.

Since the bit lines are externally precharged, the cell is not involved in the pull-up process and there is no penalty for having large pull-ups.

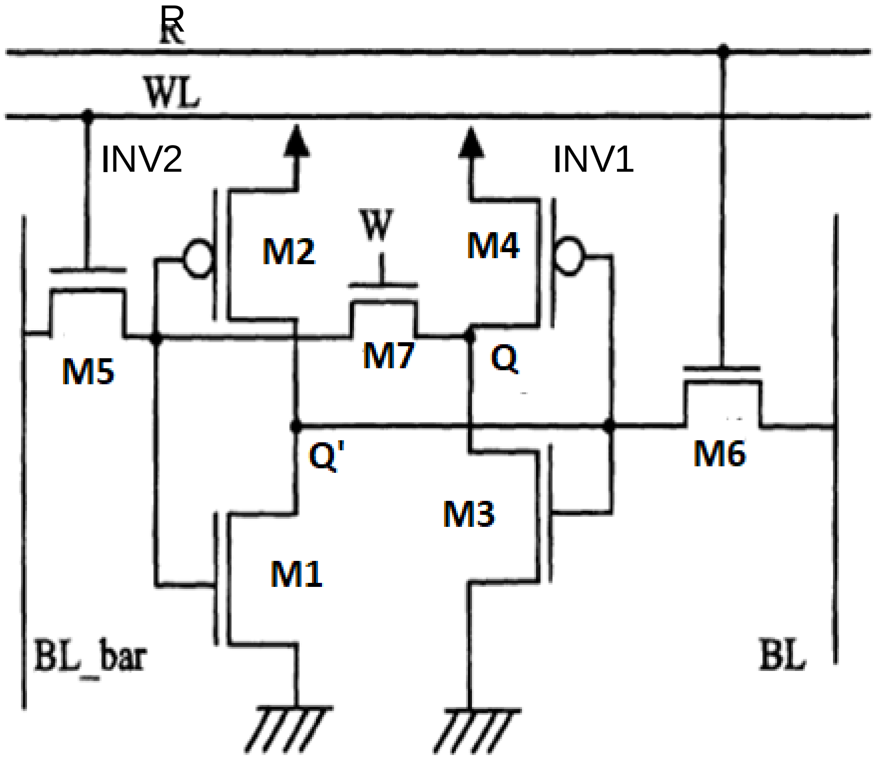
5T SRAM Cell

It is Single ended SRAM Cell consisting only True bit-line BL which is connected to Memory Cell through access transistor M5.



7T SRAM Cell

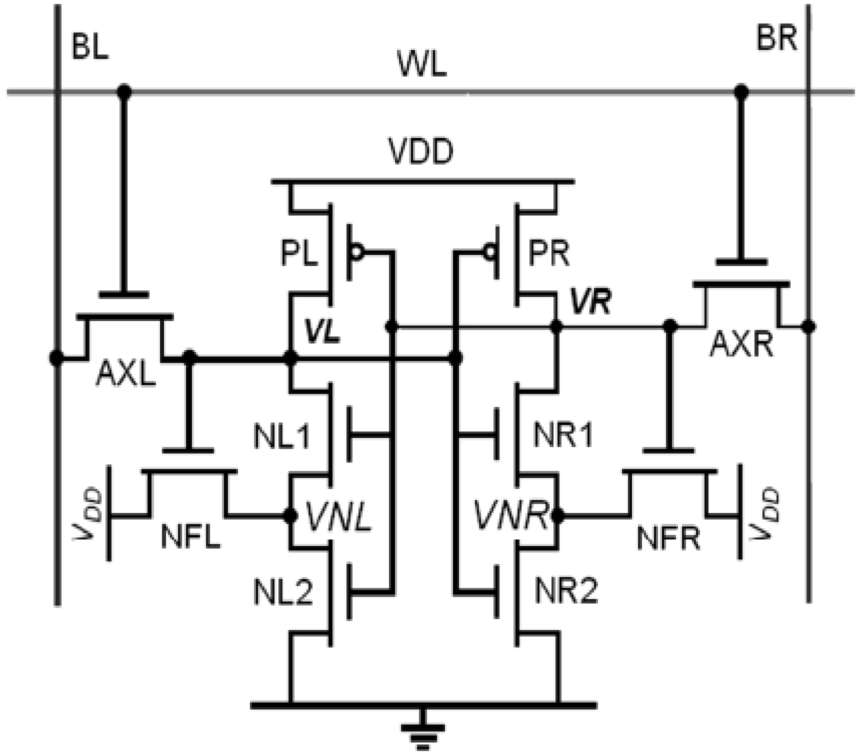
7T SRAM has additional NMOS M7 connected between output of INV1 with Input INV2.



R- Read Signal
W- Write Signal

- Write Operation: W=0, M7 is turned OFF and data is written from BL_bar through M5 into the cell.
- Since data is written only through one line, power consumption is reduced. Read operation is similar to that of a 6T SRAM Cell.
- Read Operation: R=1 and W=1 which means it acts like a conventional 6T SRAM Cell

10T SRAM Cell

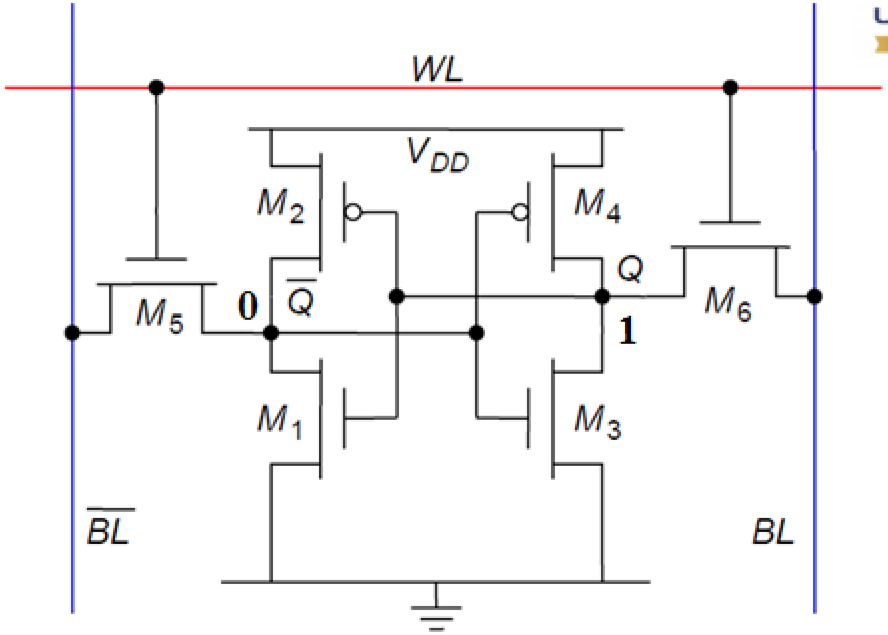


SRAM Cell – Read Operation

Before the start of *Read* operation, both bit lines are precharged to V_{DD} .

Assume that a ‘1’ is stored at Q.

Read cycle is started by asserting WL, turnir ON M_5 and M_6.



For the correct read operation, the values stored in Q and $\overline{Q}$ are transferred to the Bit lines.

Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

MEMORY DESIGN AND TESTING

SRAM Cell – Read Operation

The simplified model of SRAM cell just before the read begins is shown

Since $\overline{Q}$ Q = 1, M_1 is ON and M_2 is OFF.

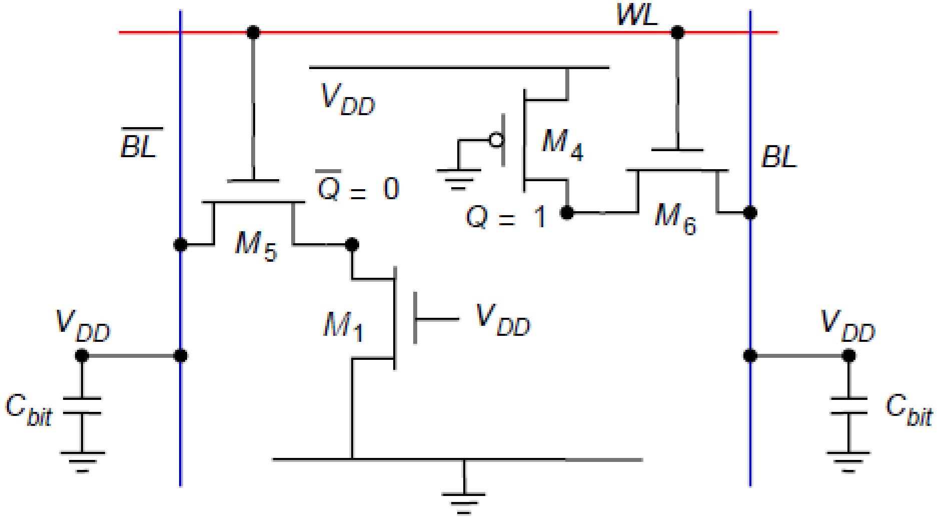
Also = 0. Hence M_3 - OFF and M_4 - O

C_{bit} refers to the bit-line capacitance whic

is of the order of pF for large memories.

This capacitance holds the Bit Lines at the

precharge value, before the read begins.



Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

MEMORY DESIGN AND TESTING

SRAM Cell – Read Operation

Once read operation begins, ie., WL = 1, then, $\overline{\mathrm{BL}}$ is discharged to ground through the series transistors M_5 – M_1. BL remains at its precharge value.

For a small-sized cell, W/L is small. Therefore, the discharging current is less. Also, the Bit Line capacitance is relatively high. This results in a slow discharge of $\overline{\mathrm{BL}}$.

As the difference between the two bit lines builds up, the sense amplifier is activated to accelerate the reading process. It quickly discharges one of the bit lines.

SRAM Cell – Read Operation

Initially, upon the rise of WL, $\overline{Q}$ is pulled up towards the precharge value of $\overline{BL}$.

This voltage rise, ΔV, must be low enough not to cause substantial current through M_3 – M_4 inverter, which in the worst case could flip the cell.

This type of malfunction is called a read upset.

To prevent this, resistance of M_5 should be made larger than M_1

Node $\overline{Q}$ is held low by M_1 but raised by M_5.

Hence M_1 must be made stronger than M_5

$(W/L)_1 >> (W/L)_5 \longrightarrow$ Resistance of M_1 << Resistance of M_5

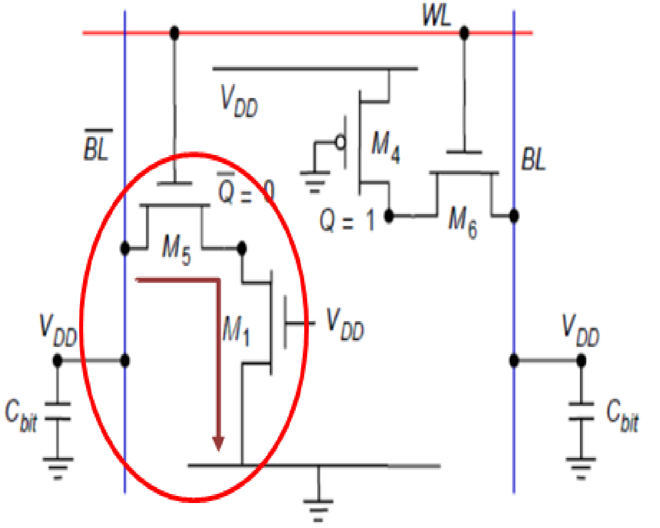
6T SRAM Cell – Read Operation

- **Resistance of M1 << Resistance of M5 so that ΔV at Q’ does not exceed Vtn of M3.** If it exceeds it may lead to unintended change of the stored state.
- The Key design issue for the data-read operation is then to guarantee that the ***voltage*** ΔV ***does not exceed the threshold of M3***, so that the transistor M3 remains turned OFF during the read phase.
- The boundary constraints on device sizes can be determined by solving the current equation at the maximum allowed value of the voltage ripple ΔV .

$$\Delta V \leq V_{tn3} \text{ -- } (1)$$

- We can assume that after the access transistors are turned on, the column voltage Vc remains approximately equal to VDD. Hence , **M5 operates in saturation while M1 operates in the linear region.**

$$\frac{k_{n\,5}}{2}(V_{DD} - \Delta V - Vtn)^2 = \frac{k_{n\,1}}{2}(2(VDD - Vtn)\Delta V - \Delta V^2) \quad \text{--------}(2)$$



$V_{GS} \geq V_{T0}$

Saturation

$V_{DS} > (V_{GS} - V_{T0})$

$$I_D = \frac{Kn}{2}(V_{GS} - V_{T0})^2$$

Linear

$V_{DS} \leq (V_{GS} - V_{T0})$

$$I_D = \frac{Kn}{2}[2(V_{GS} - V_{T0})V_{DS} - V_{DS}^2$$

6T SRAM Cell – Read Operation

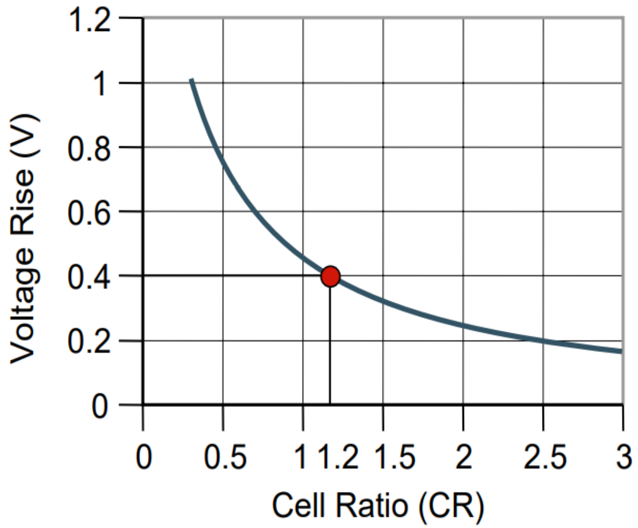
Substitute ΔV as V_{tn} in equation 2 and by simplifying the equation

$$\frac{k_{n,5}}{k_{n,1}} = \frac{\left(\frac{W}{L}\right)_5}{\left(\frac{W}{L}\right)_1} < \left[\frac{2(V_{DD} - 1.5Vtn)V_{tn}}{(V_{DD} - 2Vtn)^2}\right] \text{--------------------}(3)$$

By, Inverse of en (3) we get

$$\text{Cell Ratio(CR)= } \beta = \frac{\left({}^{W}/_{L}\right)_1}{\left({}^{W}/_{L}\right)_5} \text{-----------------------}(4)$$

The variation of ΔV as a function of CR, for 250nm technology is shown below



To keep the node voltage (ΔV) from rising above the transistor’s (M3) **threshold voltage (0.4V)**, **the cell ratio must be greater than 1.2.**
A **β value of 1.5 to 2.0** is typical in the industry. A β value less than 1 implies that, each time the cell is read, it is disturbed as well

To achieve the desired value of CR, W1 is increased while L1 = L5.

Stability of SRAM Cell

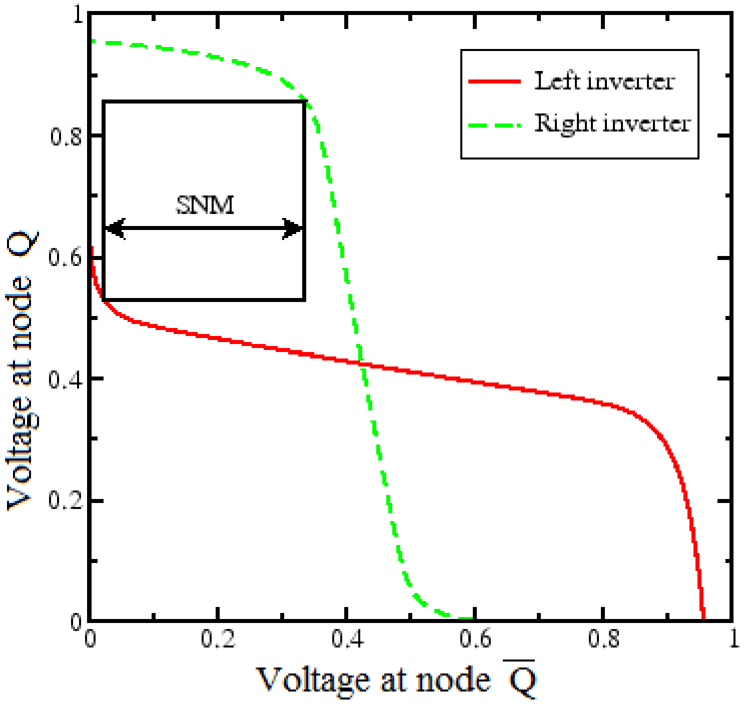
An SRAM cell's stability can also be described by the "Butterfly curve"

The Butterfly curve is obtained by super-imposing the Voltage Transfer curves of the 2 inverters.

The stability of the cell is given by the size of the square box that can fit inside the "Butterfly wing"

The larger the box, higher the β value and hence greater the stability.

The curve flattens during a read operation, reducing the box size thus, implying reduced stability.



SRAM Cell – Read Operation

The square box also gives the SNM – Static Noise Margin, which is the tolerance of an SRAM cell to noise before it risks losing the stored bit.

SNM is defined as the length, in volts, of the largest possible square in the Butterfly curve.

If the external DC noise is larger than SNM, SRAM can lose its data.

Bit Line, BL, clamps Q to V_{DD} which makes the inadvertent toggling of the cross-coupled inverter pair difficult. This is one of the major advantages of the dual bit line architecture.

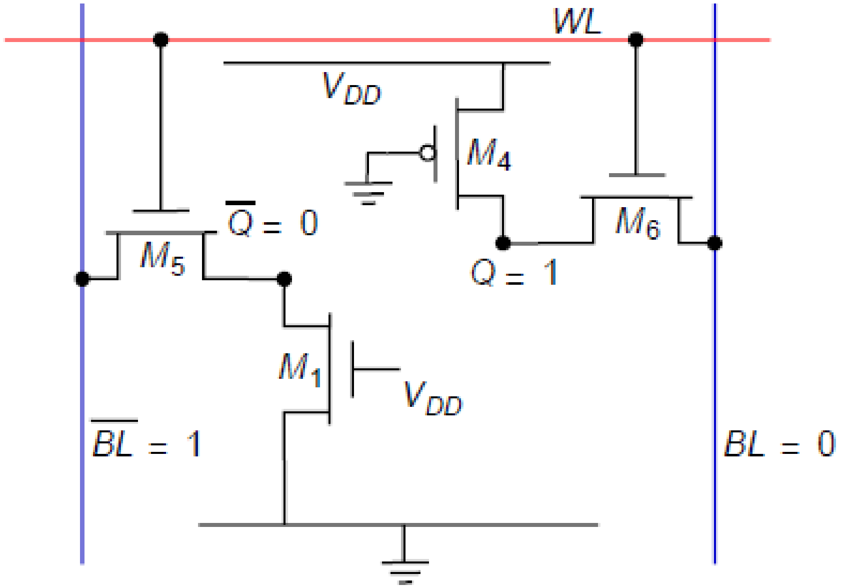
MEMORY DESIGN AND TESTING

SRAM Cell – Write Operation

Assume that a ‘1’ is stored in the cell, ie., Q = 1.

A ‘0’ is written into the cell by setting $\overline{\mathrm{BL}}$ to ‘1’ and BL to ‘0’.

The simplified model of SRAM cell just before write begins is shown



Since Q = 1, M_1 is ON and M_2 is OFF.

Similarly, since $\overline{\mathrm{Q}}$ = 0, M_3 is OFF and M_4 is ON.

SRAM Cell – Write Operation

$\overline{Q}$ side of the cell cannot be pulled high enough to ensure the writing of ‘1’

The sizing constraint, imposed by the read stability, ensures that this voltage is kept below 0.4V.

Hence the new value of the cell has to be written through transistor M_6.

Data ‘1’ will be written into the cell if node Q is pulled down, below the threshold voltage of M_1 to turn it OFF.

6T SRAM Cell - Write Operation

Writing ‘0’

Transistor M6 has to push the data ‘0’ from BL to Q and On the other hand M4 , tries to pull-up Q to 1.

Hence for reliable writing, M6 must be stronger than M4 so that Node Voltage at Q reduces to less than Vtn, M1 to turn it OFF.

Under this condition (Q=Vtn) , the transistor **M6 operates in linear region** and **M4 operates in the saturation region**

$$\frac{k_{n,4}}{2}(0 - VDD\ - V_{tp})^2 = \frac{k_{n,6}}{2}(2(VDD - Vtn)V_{tn} - V_{tn}{}^2) \quad \text{----------(1)}$$

by simplifying the equation

$$\frac{\mathrm{k}_{\mathrm{n},4}}{\mathrm{k}_{\mathrm{n},6}} < \left[\frac{2(\mathrm{V}_{\mathrm{DD}} - 1.5\mathrm{Vtn})\mathrm{V}_{\mathrm{tn}}}{(\mathrm{V}_{\mathrm{DD}_{+}}\mathrm{Vtp})^2}\right]$$

$$\frac{\left(\frac{\mathrm{W}}{\mathrm{L}}\right)_4}{\left(\frac{\mathrm{W}}{\mathrm{L}}\right)_6} < \frac{\mu_n}{\mu_p} \cdot \left[\frac{2(\mathrm{V}_{\mathrm{DD}} - 1.5\mathrm{Vtn})\mathrm{V}_{\mathrm{tn}}}{(\mathrm{V}_{\mathrm{DD}} + \mathrm{V}_{\mathrm{tp}})^2}\right] \quad \text{-----------------------(2)}$$

To summarize, the transistor M1 will be forced into cut-off mode during the write”0” operation if the above conditions are satisfied. It means M3 turns ON.

6T SRAM Cell - Write Operation

Writing ‘0’

Thus **the pull –up ratio PR** for write operation is given by

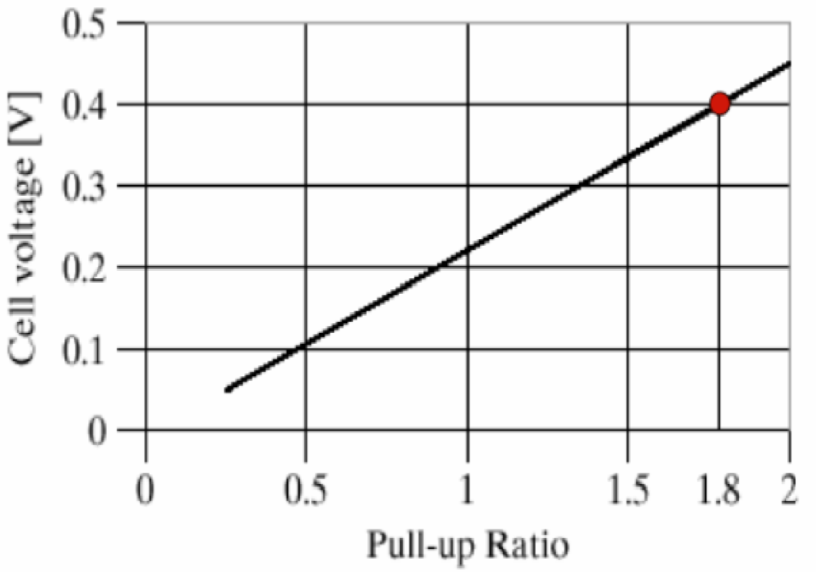
$$PR=\frac{(W/L)_4}{(W/L)_6} \quad \text{------------------------(3)}$$

The dependence of Voltage at node Q, V_Q , on PR for a 0.25 um process is shown below

To pull – down node Q below the threshold voltage, 0.4V, of M1 , PR should be less than 1.8.

Therefore, the transistor sizing for efficient read and write operations:

Pull – down transistors M1 , M3 – strongest,
Pull- up transistors M2 , M4 – weakest,
Transfer transistors - Medium



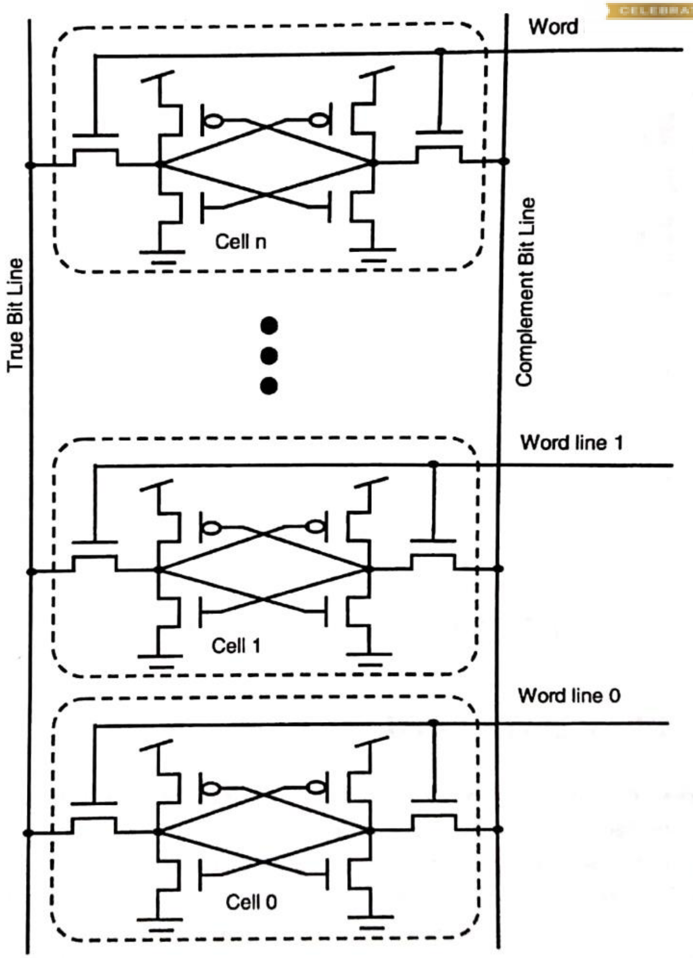
MEMORY DESIGN AND TESTING
SRAM Peripheral Circuitry

Bit – Line pair

The cells of a memory are joined along a column by a bit line pair.

The number of cells along a differential bit line column is normally large, often being 256, 512 or 1024 in number.

High speed SRAMs always use a pair of bit lines whereas smaller, low performance SRAMs can have a single bit line for reading and another for writing.

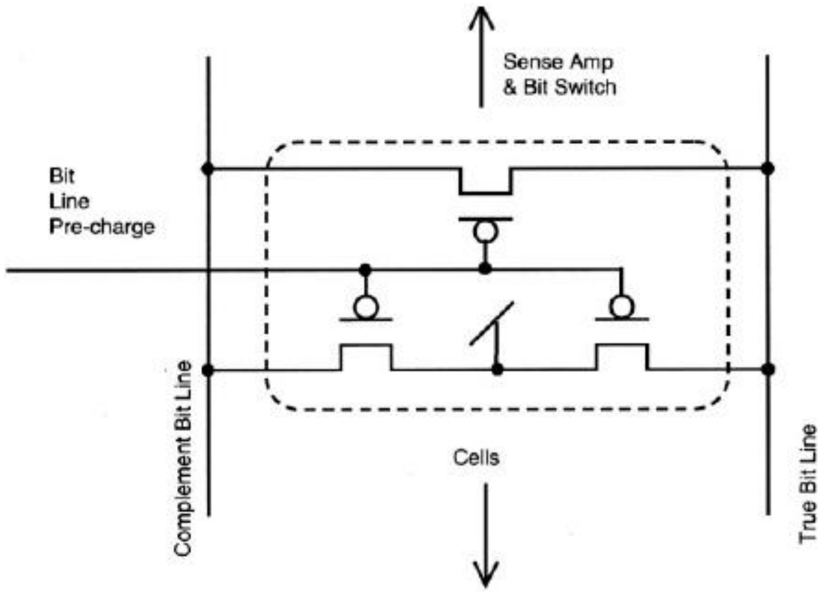


Diagrams courtesy “High Performance Memory Testing” by R Dean Adams

Pre-charge circuitry

Before a *Read* operation, both bit lines are pre-charged to V_{DD}

The precharge circuitry consists of 3-transistors as shown



This circuit is sometimes referred to as a “crow – bar”

The bit line precharge signal is active low.

Diagrams courtesy “High Performance Memory Testing” by R Dean Adams

Pre - charge Circuitry

There is a PFET pulling each bit line to V_{DD} and a third PFET connecting the 2 bit lines together to equalize their potentials.

Need for equalizer transistor:

- Sense amplifiers are highly susceptible to differential noise on the bit lines, since they are designed to detect small voltage differences.

- If the bit lines are not charged long enough, residual voltages on the lines from the previous ‘read’ may cause pattern-dependent failure.

- The equalizer transistor reduces the required precharge time by ensuring that both bit lines are at nearly equal voltages even if they are not precharged to V_{DD}.

Pre - charge Circuitry

Half-select state:

During a *Read*, the precharge circuit is turned off for a column that is being read.

For the columns that are not being read, the precharge is left ON.

The word line which goes high corresponding to the input address selects a row of cells.

With the precharge circuit ON for the unselected columns, the cell fights against the precharge circuit.

Pre - charge Circuitry

Such cells are said to be in “half-select” states since the WL has gone high but the column is not yet selected for a *read.*

The cells can still retain data in the half-select state since NFET transfer devices have poor pull-up characteristics (Weak 1).

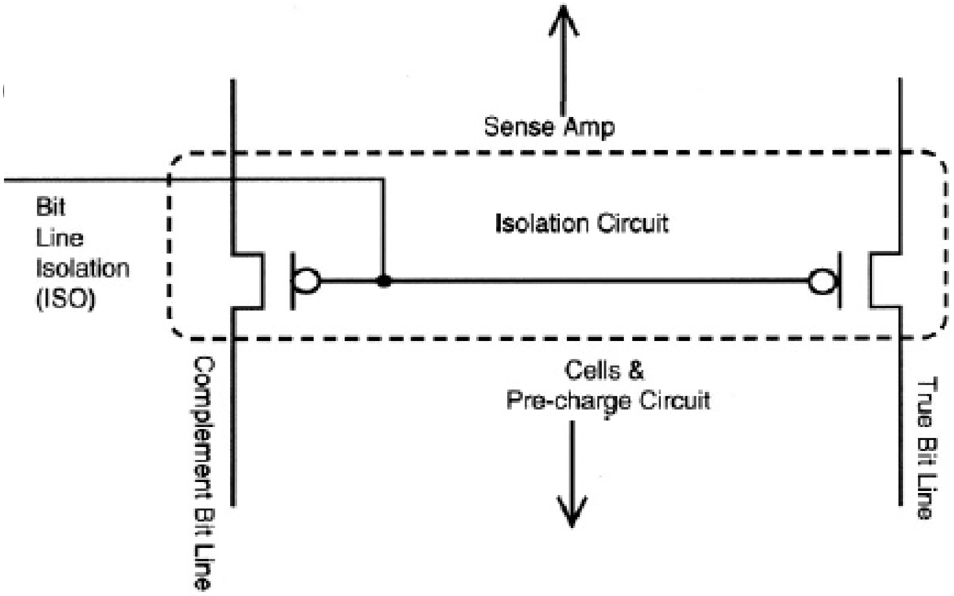
The half-select state can be utilized as a feature to weed out defective or weak cells, since defect free SRAM cells can retain data in a half-select state.

MEMORY DESIGN AND TESTING

Isolation Circuit

Bit lines are connected to Sense Amplifier through Isolation circuitry.

Once sufficient signal is developed across the bit lines, the Bit lines are isolat



Bit line isolation signal (ISO) goes high, turning OFF the PFETs

Diagrams courtesy “High Performance Memory Testing” by R Dean Adams

Isolation Circuit

Need for isolation:

A single Sense amplifier exists for each bit line pair.

Bit lines are long with many cells attached to them.

Also, the long metallization of the bit lines makes it highly capacitive.

Hence the bit line that is connected to data 0 discharges very slowly, from its precharge value.

The sense amplifier completes the discharge process, once it gets isolated from the cells.

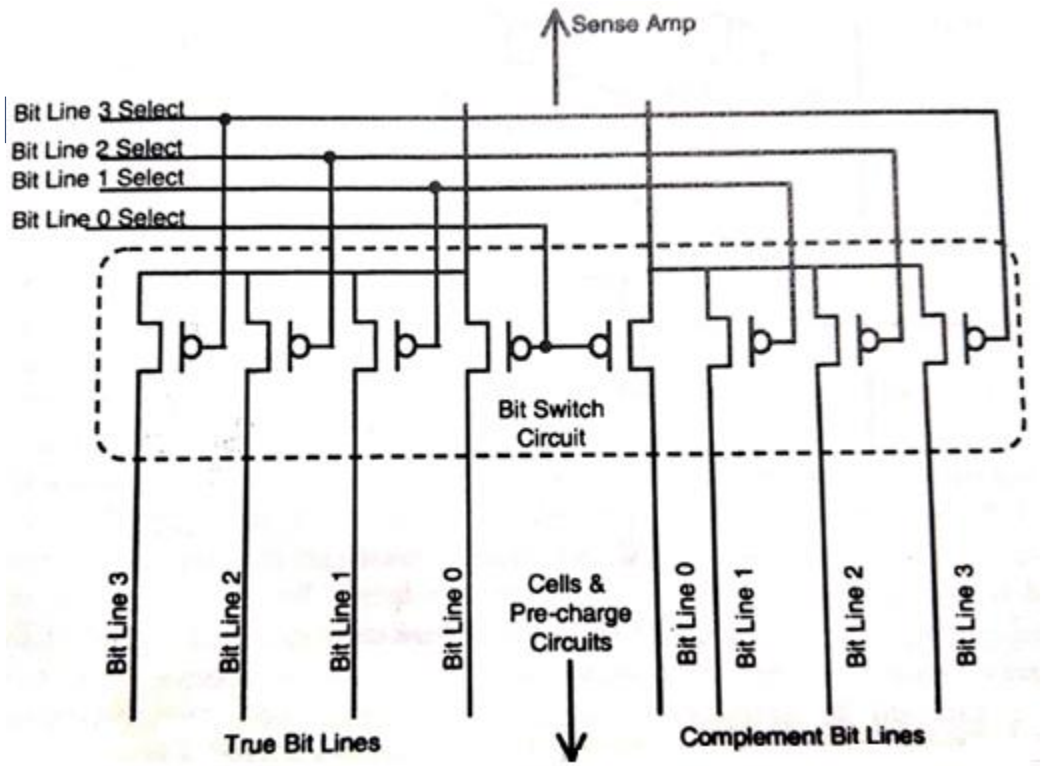
Isolation Circuit

Bit switch isolation circuit:

Large memories have a single isolation circuitry for multiple columns.

In such cases, the isolation circuit is rep

The output of the column decoder is connected to Bit line select signal which connects the appropriate pair of Bit lines to the sense amplifier.



Diagrams courtesy “Digital Integrated Circuits: A Design Perspective” by Jan Rabaey

Sense Amplifier

The *Read* operation requires the discharging of large bit line capacitance through the stack of 2 small transistors.

To accelerate the read process, SRAMs use sense amplifiers.

A sense amplifier is a differential voltage amplifier that amplifies small differential input voltage (bit line voltage) into larger output voltage.

Sense Amplifier

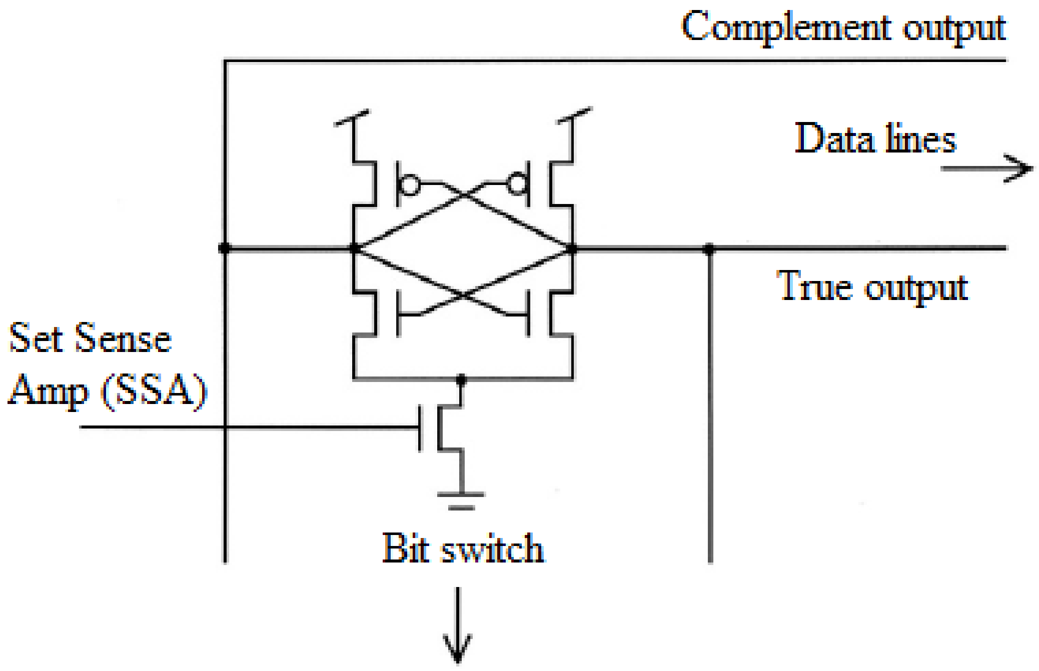
A differential approach is used due to its advantage of common – mode rejection, which is its ability to reject noise that is equally present at both the inputs.

The impact of these noise signals can be substantial since the amplitude of the signal to be sensed is generally small.

The effectiveness of a differential amplifier is characterized by its ability to reject the common noise and amplify the true difference between the signals

Latch – type Sense Amplifier

Latch – type sense amplifier:



The sense amplifier is activated by the SSA signal going high.

This will turn on the 2 inverters.

One of the output lines will be quickly discharged to ground potential.

Diagrams courtesy “High Performance Memory Testing” by R Dean Adams

Latch – type Sense Amplifier

Functions of a sense amplifier:

1) Delay reduction: The differential bit lines have enormous capacitive loads. Due to this they swing slowly. The sense amplifier can detect a small swing on the bit lines and bring it up to normal logic levels.

In other words, it reduces the delay by avoiding full swing on the bit lines themselves.

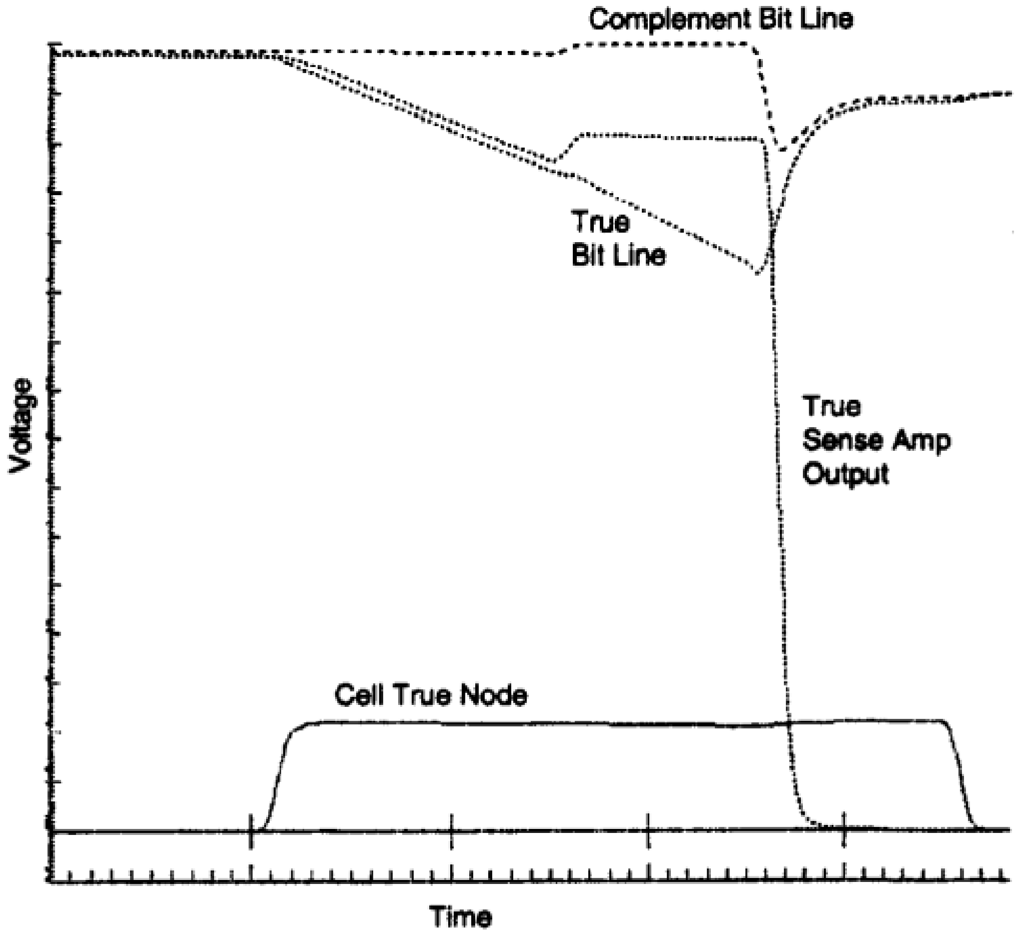
2) Power reduction: Reducing the signal swing on the bit lines can eliminate a substantial part of power dissipation related to charging and discharging of the bit lines.

3) Signal restoration: In DRAM cells, the read and refresh functions are intrinsically linked. Hence the sense amplifiers are involved in driving the bit line to full signal range after sensing.

Reference Material “High Performance Memory Testing” by R Dean Adams

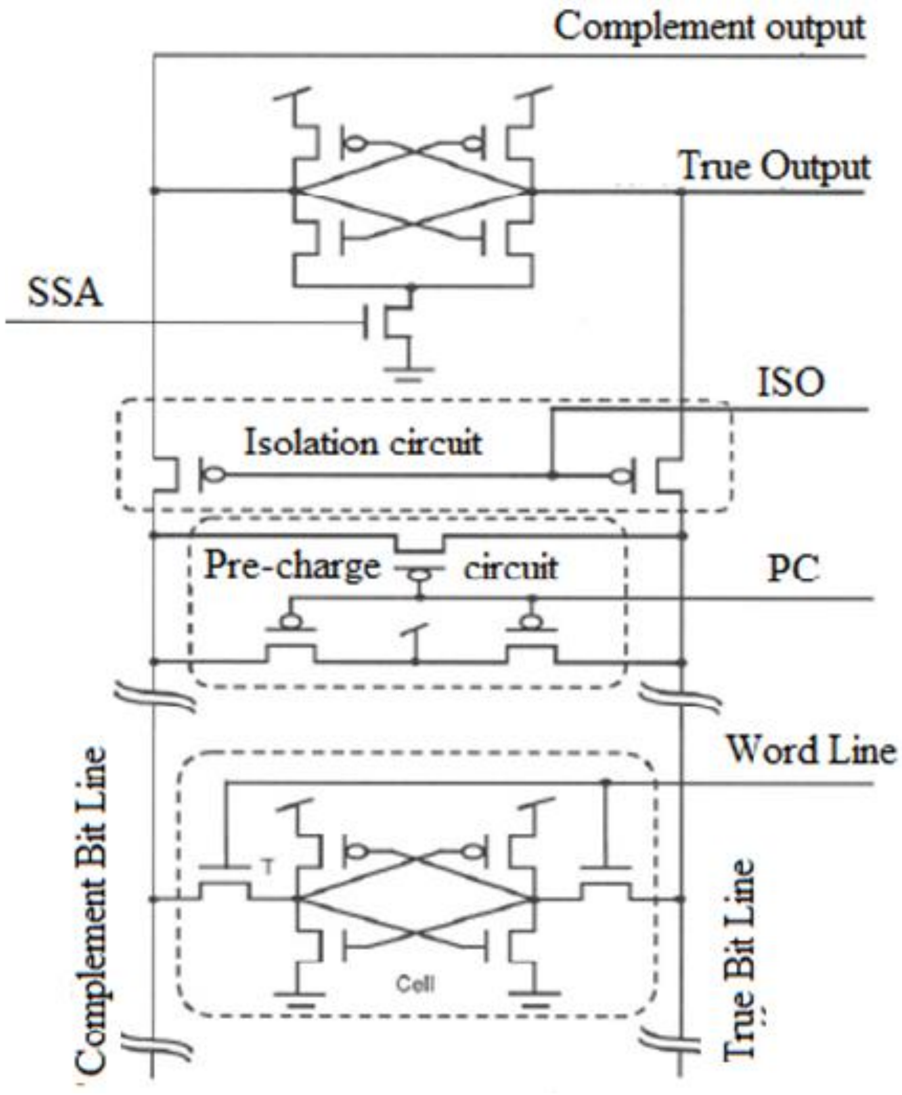
Sense Amplifiers

Cross Coupled CMOS Latch used as Sense Amplifier



- The cell's true node pops up indicating that the word line has gone high.
- The signal starts to develop on the true bit line. The complement bit line remains high.
- Signal stops developing on the sense amplifier true node when the IS0 signal goes high.
- The SSA signal goes active causing the true sense amplifier output to go low.
- In Figure the IS0 and SSA signals have been purposely separated in time to illustrate their respective coupling effects on the bit-line pair.

SRAM Read data path



Typical column in an SRAM array

Diagrams courtesy “High Performance Memory Testing” by R Dean Adams

SRAM Read data path

A read cycle proceeds as follows:

1. Bit lines are precharged to V_{DD} by making Pre- Charge signal low. This turns ON the equalising transistor ensuring that the initial voltage on both the bit lines is identical. Equalisation is critical when the bit lines are precharged through NMOS pull-ups since there is a variation in the device threshold.

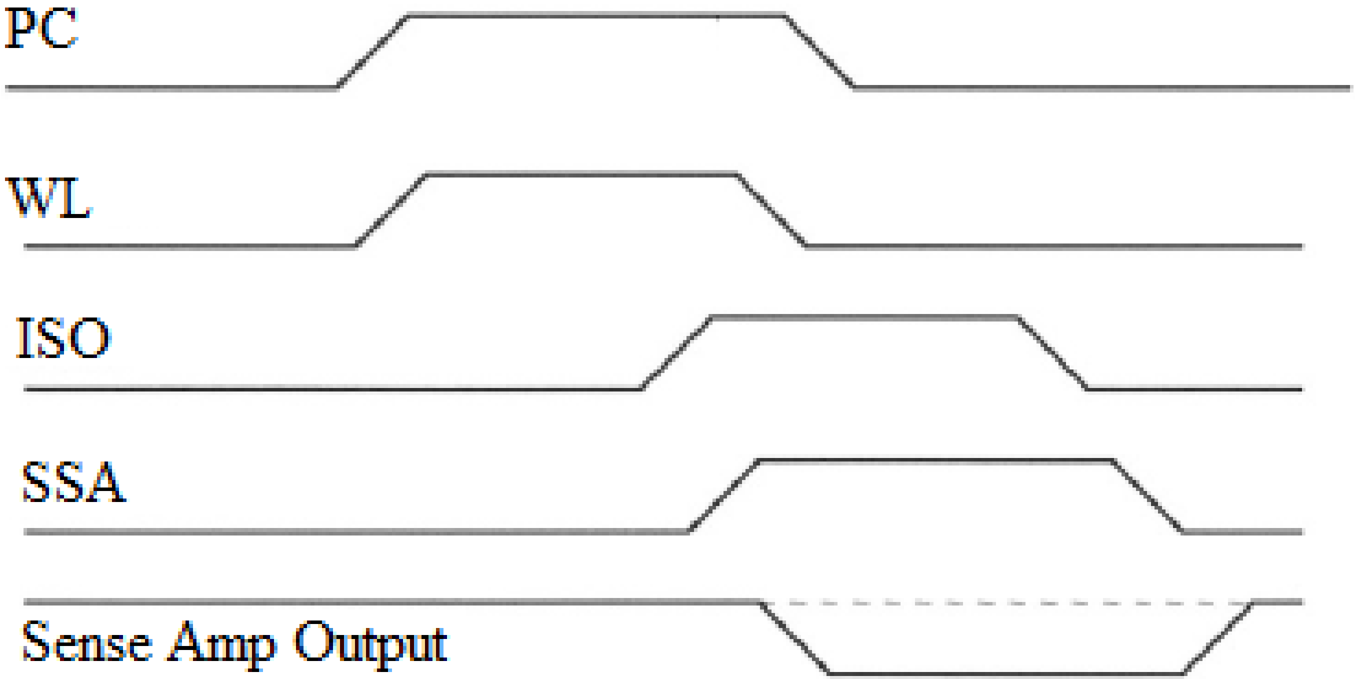
2. Read operation is started by disabling the PC signal and enabling one of the word lines. One of the bit lines is pulled low by the selected memory cell in that column

SRAM Read data path

3. Once a sufficient signal is built-up, ISO is made high followed by the SSA signal.
4. If there is a single bit line pair per sense amplifier, the ISO and SSA signals can be combined into a single signal.
5. The differential input signal is amplified by the sense amplifier and eventually a rail – to – rail output is produced
6. Once ISO goes high to isolate the cells from the sense amplifier, PC is made low to precharge the bit lines.

SRAM Read data path

The typical read timing waveforms are shown



Write driver circuit

The circuit that writes data into the cell is called a Write Driver

The writing of a cell is accomplished by writing a ‘0’ into either the true or the complement side of the cell and the cell latch causes the opposite side to go to a ‘1’ state.

This is because, the cells have NFET transfer devices which drive a strong ‘0’ and do not drive a ‘1’ very effectively.

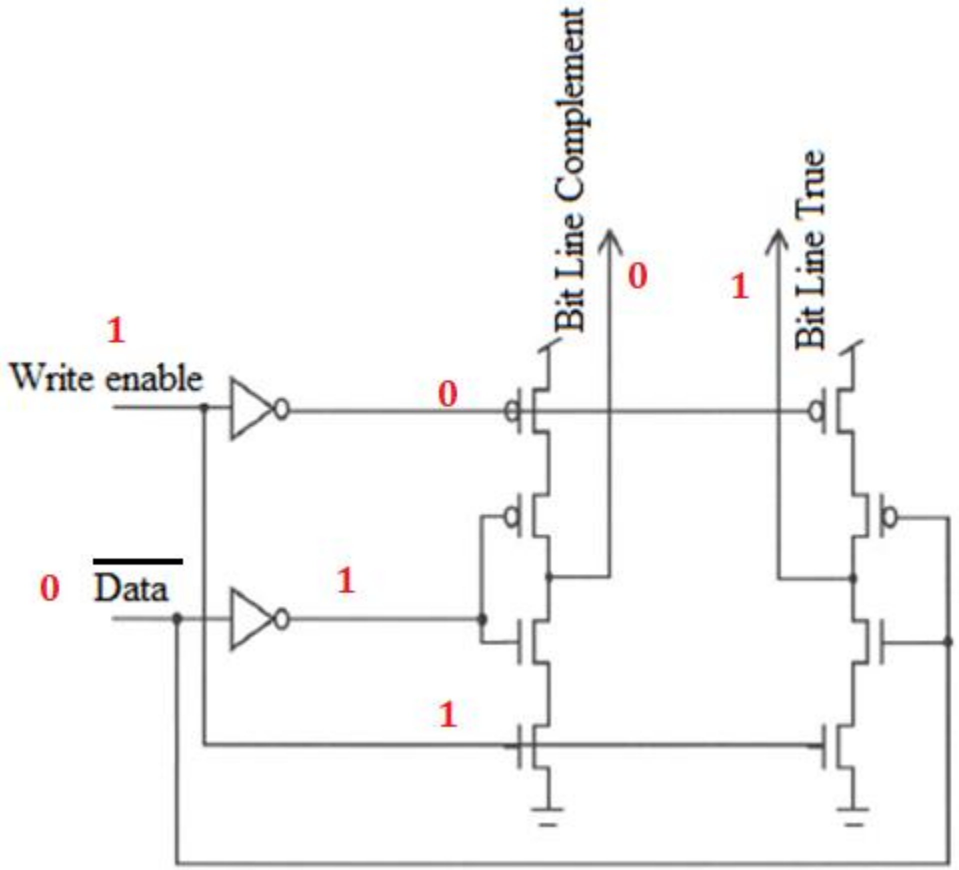
A simple write driver is shown

Write driver circuit

The 4 vertical devices in series are often referred to as a ‘Gated Inverter’.

When *Write Enable* is made high, the 2 inverters turn ON, driving the appropriate data values onto the true and complement Bit lines.

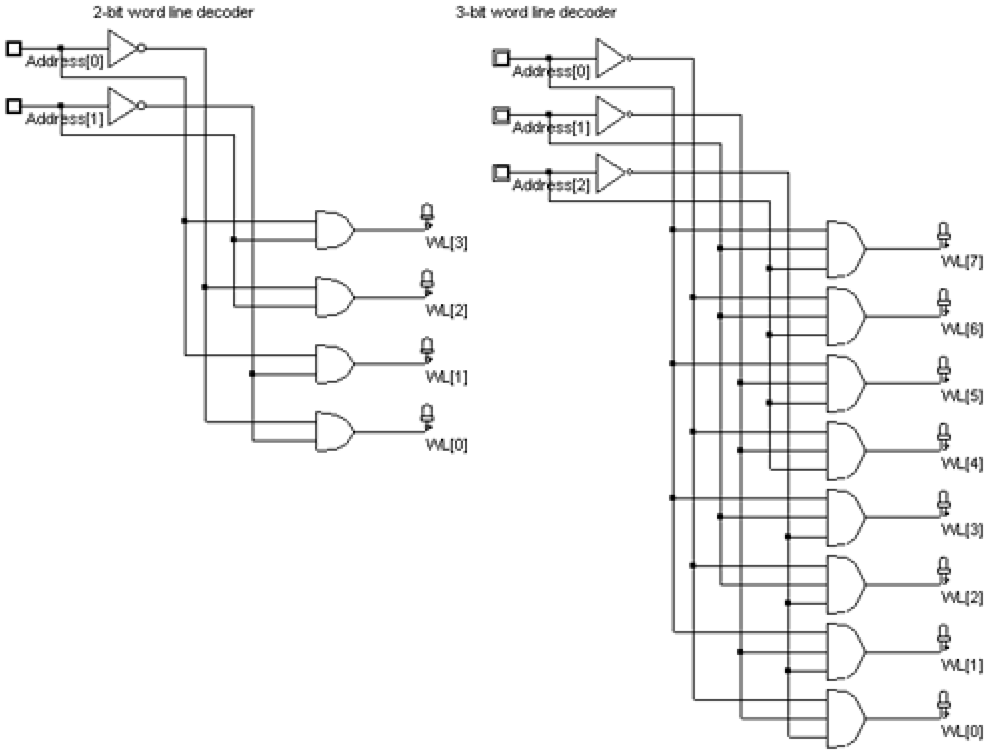
Since the primary objective is to drive a ‘0’ the NFETs and PFETs are similarly sized rather than the typical 2:1 ratio used for other logic.



Diagrams courtesy “High Performance Memory Testing” by R Dean Adams

Address Decoder

- Address Decoder are used to decode the Address of Memory Location to be accessed.
- Based on Memory size, there might be only Row Decoders or Both Row and Column Decoders. Smaller size memories will have only Row Decoders.
- Classification of Address Decode
  1) Static Decoders
     a) Row Decoders
     b) Column Decoders
  2) Dynamic Decoders
     a) Row Decoders
     b) Column Decoders

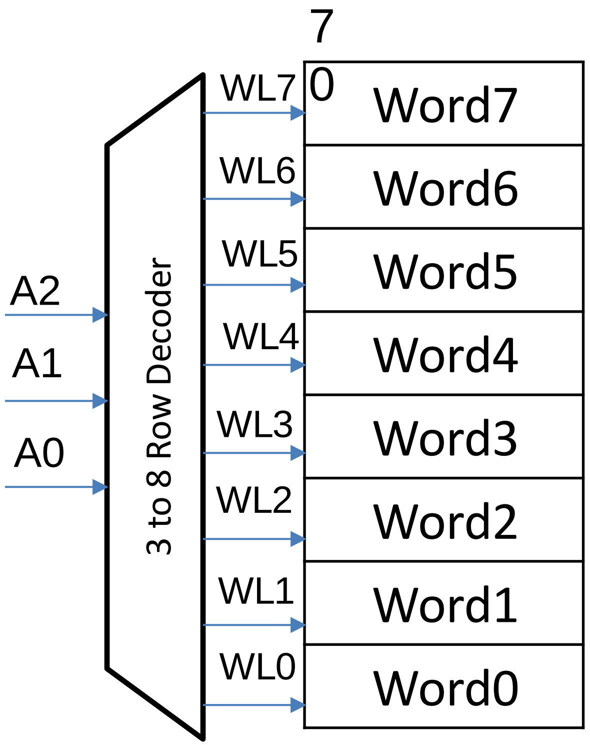


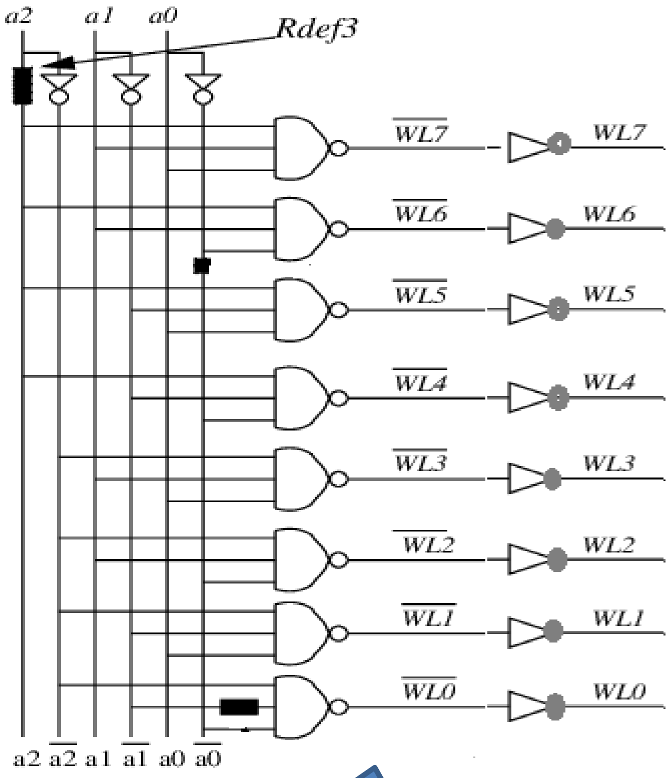
Address Decoder – Static Row Decoders

A 3 bit Address Decoder (Row)

Let us say, A Memory Size of 8 words of 8 bit Size to be addressed.

A2	A1	A0	WL
0	0	0	WL0
0	0	1	WL1
0	1	0	WL2
0	1	1	WL3
1	0	0	WL4
1	0	1	WL5
1	1	0	WL6
1	1	1	WL7

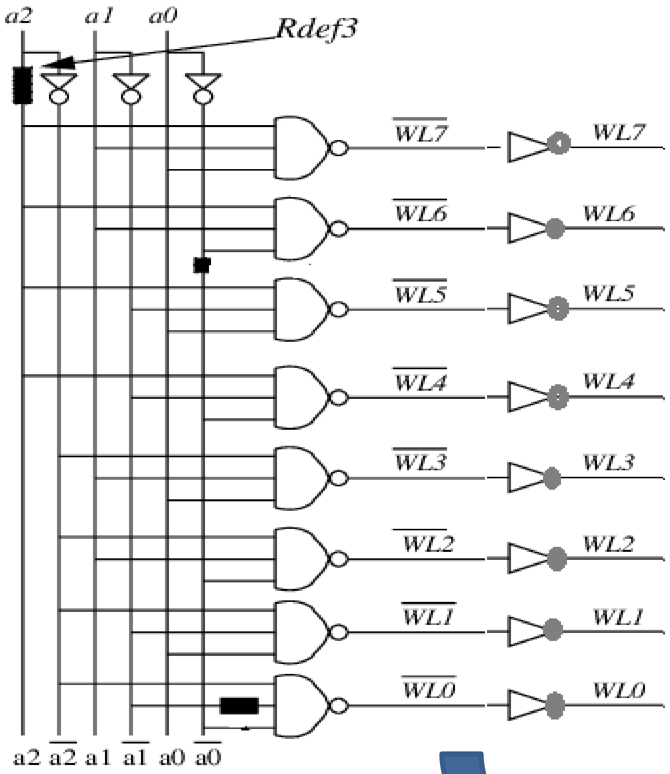




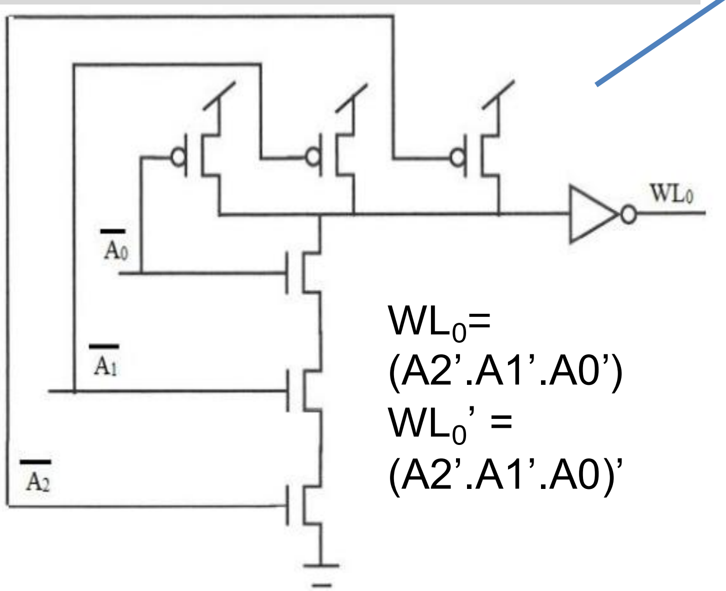
Address Decoder Static Row Decoders

A 3 bit Address Decoder (Row)

Let us say, A Memory Size of 8 words of 8 bit Size to be addressed.



Two Stage Implementation Using a 3 input NAND and INV



WL_0= (A2’.A1’.A0’)
WL_0’ = (A2’.A1’.A0)’

8 such logics will give a 3 to 8 Row Decoders

This logic can also be Implemented using NOR gate
We have
WL_0= (A2’.A1’.A0’)
which can be written as
WL_0= (A2+A1+A0)’

Single Stage Implementation using NOR gate reduces the Number of MOS devices required as No Inverter required.

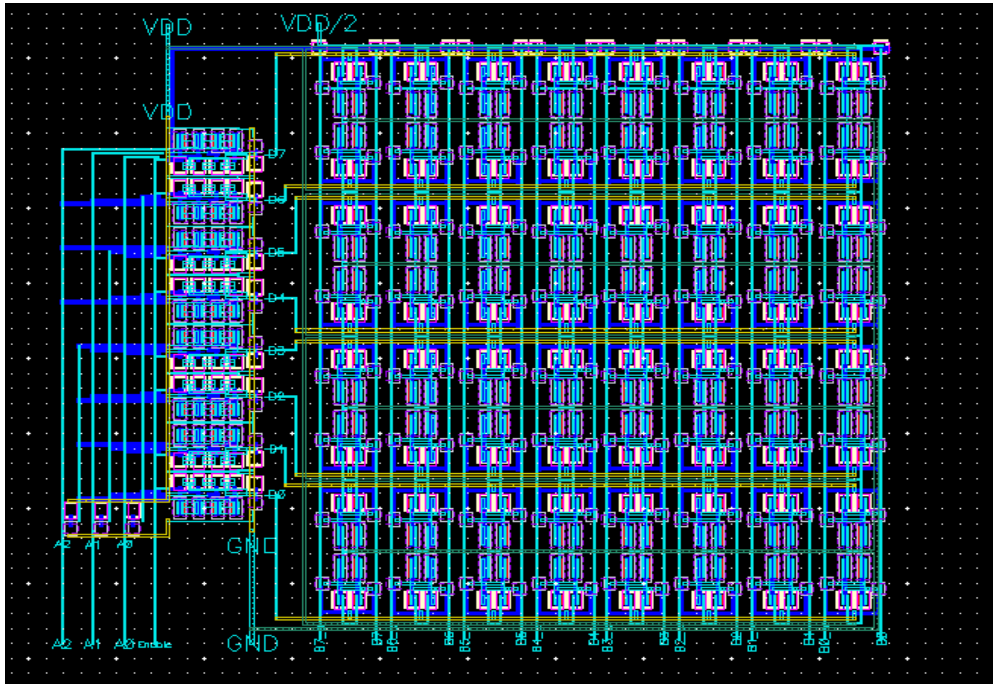
Address Decoder -Static Row Decoders

A 3 bit Address Decoder (Row)

Let us say, A Memory Size of 8 words of 8 bit Size to be addressed.

Challenges of Decoders using NOR gates

- The Layout of Decoder using NOR will have challenge of Fitting within the word line Pitch
  - ✓ Height of the Decoder logic for enabling a Row must match with Height of the Row Cells.
  - ✓ Ex: If the Memory Size is increased to 256x8 then, 256 rows need to be enables. It means a 8:256 decoder is required, which means a 8 input NOR gate for each WL.
  - ✓ In a Memory, Row Pitch will almost remain same even though size increases, but the Size of NOR gate used for Decoder increases.
- The Large FAN IN of the gate has Negative Impact on the Performance. The propagation delay Increases which effects both READ and WRITE access times.



- NOR gates should Drive Large Capacitive Word Lines

Implementation of such Decoders is Impractical

Address Decoder -Static Row Decoders

Hierarchical Decoders

- Now, Splitting of Complex Logic gates into two or more logic Layers makes it Faster and Economical.
- This decomposition of CMOS Decoder makes them fast and area efficient. In this, Decoder is split in to Two Layers as
  - ✓ Pre-Decoder and
  - ✓ Final Decoder

Address Decoder -Static Row Decoders

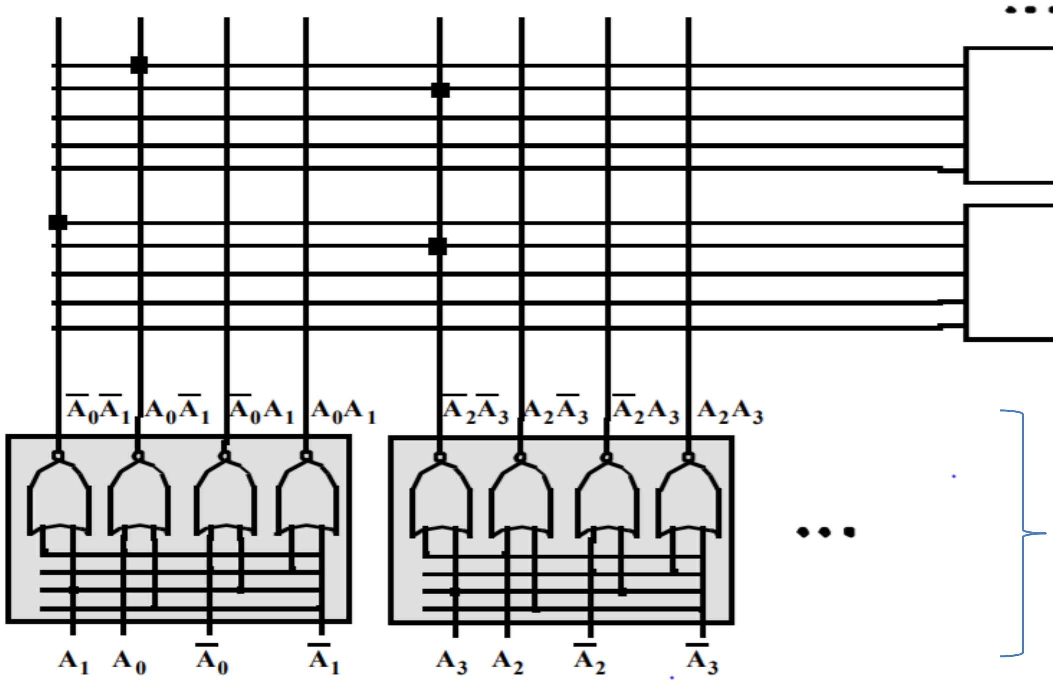
Hierarchical Decoders

- Let us consider 8:256 Decoder to be Implemented. The expression for Implementation of WL_0 using regrouping is as below

$$WL_0 = \overline{\overline{A_0}.\overline{A_1}.\overline{A_2}.\overline{A_3}.\overline{A_4}.\overline{A_5}.\overline{A_6}.\overline{A_7}}$$

$$= \overline{\overline{(A_0 + A_1)}\,\overline{(A_2 + A_3)}\,\overline{(A_4 + A_5)}\,\overline{(A_6 + A_7)}}$$

- We can see that Address in Stage 1 is partitioned into section of two bits that are decoded in advance.
- The resulting signals are combined using 4 input NAND gates to produce corresponding WL signals.



Pre Decoder Stage

Address Decoder -Static Row Decoders

Advantages Hierarchical Decoders

1) Reduces Number of Transistors required.
2) As the Number of Inputs to NAND gates is Halved (4), the propagation Delay is reduced approximately by factor of 4.

   Gates with Fan IN greater than 4 should be avoided in order to reduce the

   propagation delay.

1) Adding a Select Signal each of the Predecoders (NOR gate) makes it possible to disable the decoder when Memory block is not selected. This results into Power Saving

Address Decoder -Static Row Decoders

Advantages Hierarchical Decoders

1) Reduces Number of Transistors required.

Calculate Number of Transistors required to design 8:256 Decoder using

a) Single stage NAND Decoder

b) NAND gate Decoder using predecoder

Write a comment based on Number of Transistors used

Total Number of Rows to be decoded is 256, (WL_{255} to WL_0)

a) Single stage NAND Decoder

In this,

Total Number of Transistors Required is = 256, 8 INPUT CMOS NAND gate

= 256 x 16 = 4096

a) NAND gate Decoder using predecoder

In this,

Total Number of Transistors Required is

= 256, 4 INPUT CMOS NAND in Final stage + 4 Sets of 4 , 2INPUT NAND

= (256 x 8) + (4 X 4 X 4) = 2112

Comment: The NAND using predecoder requires 52% Transistors of Single stage Decoders

Address Decoder -Static Row Decoders

Drawback Hierarchical Decoders

Still these designs require 4 input NAND gate are require to drive the large capacitive word lines.

The best Driver which can be used for driving large capacitive loads is an Inverter. Therefore output NAND should be buffered.

To drive Large Capacitive load Multi stage of logics should be used. Therefore eqn

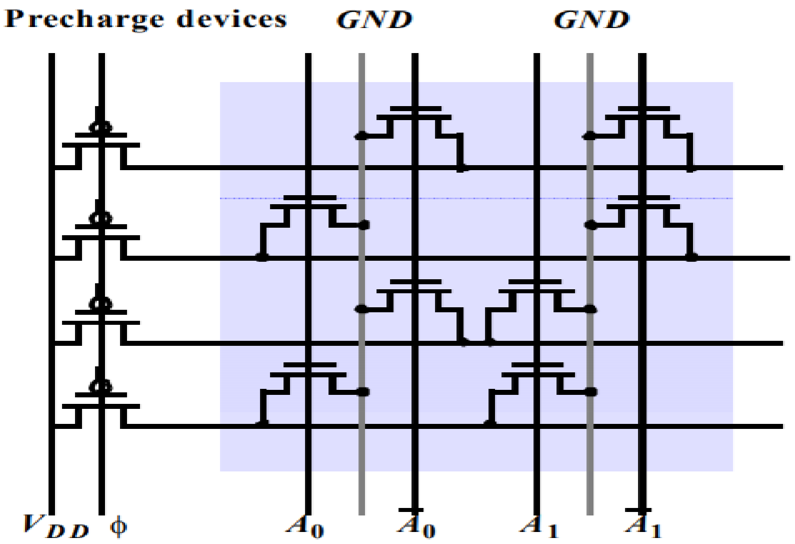
$$WL_0 = \overline{\overline{A_0}.\overline{A_1}.\overline{A_2}.\overline{A_3}.\overline{A_4}.\overline{A_5}.\overline{A_6}.\overline{A_7}}$$

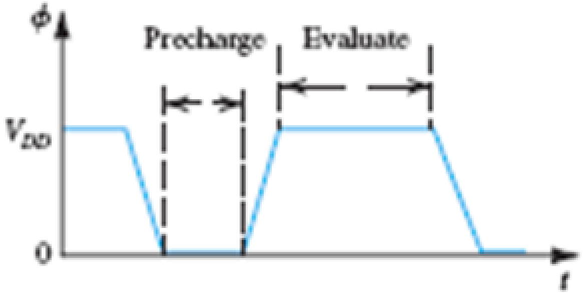
Can be broken into additional levels of logic, each of which consists of 2 input NANDs, 2 input NORs or Inverters

Address Decoder -Dynamic Decoders

Dynamic 2 to 4 NOR Decoder

- Dynamic Logic 2 input NOR gate is used in this topology.
- When Clock Ø=0; The outputs are in Pre-charged state (WL3-WL0) are HIGH.
- When Clock Ø=1; the circuit is Evaluation state where in the o/p depends on address lines A1,A0



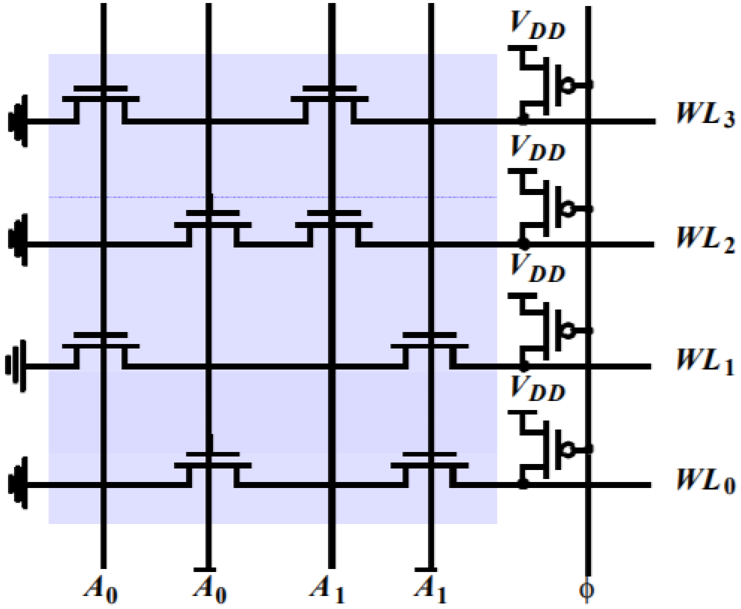


Ø	A1	A0	WL3	WL2	WL1	WL0	State
0	X	X	1	1	1	1	Pre-charged State
1	0	0	0	0	0	1	Evaluate State (WL0 is Selected)
1	0	1	0	0	1	0	Evaluate State (WL1 is Selected)
1	1	0	0	1	0	0	Evaluate State (WL2 is Selected)
1	1	1	1	0	0	0	Evaluate State (WL3 is Selected)

Address Decoder -Dynamic Decoders

Dynamic 2 to 4 NAND Decoder : It gives Active Low output.

- Dynamic Logic 2 input NAND gate is used in this topology.
- When Clock Ø=0; The outputs are in Pre-charged state (WL3-WL0) are HIGH.
- When Clock Ø=1; the circuit is Evaluation state where in the o/p depends on address lines A1,A0.
- In this, All the outputs of the Array are HIGH by default except the Selected row based on A1A0.



Ø	A1	A0	WL3	WL2	WL1	WL0	State
0	X	X	1	1	1	1	Pre-charged State
1	0	0	1	1	1	0	Evaluate State (WL0 is Selected)
1	0	1	1	1	0	1	Evaluate State (WL1 is Selected)
1	1	0	1	0	1	1	Evaluate State (WL2 is Selected)
1	1	1	0	1	1	1	Evaluate State (WL3 is Selected)

Comparison of NOR and NAND Decoders

1) Delay:

NOR decoders are faster than NAND decoders.

In NOR decoders, one line remains in the precharge state and 3 lines are discharged during the evaluation phase.

Each of the lines is pulled-down to ground simultaneously by 2 NMOS transistors.

On the other hand, in NAND decoders, only one line is discharged during the evaluation phase, by a pair of 2 series transistors.

Thus NAND decoders discharge more slowly than NOR

Comparison of NOR and NAND Decoders

2) Power:

NOR decoders consume more power than NAND decoders.

In NOR decoders, 3 lines discharge during the evaluation phase and precharge once again during the precharge phase.

On the other hand, in NAND decoders, only one line discharges and recharges.

Thus the charging – discharging current of 3 Word lines causes higher power dissipation in NOR decoders.

Comparison of NOR and NAND Decoders

3) Area:

NOR decoders consume more area than NAND decoders due to the presence of additional ground lines.

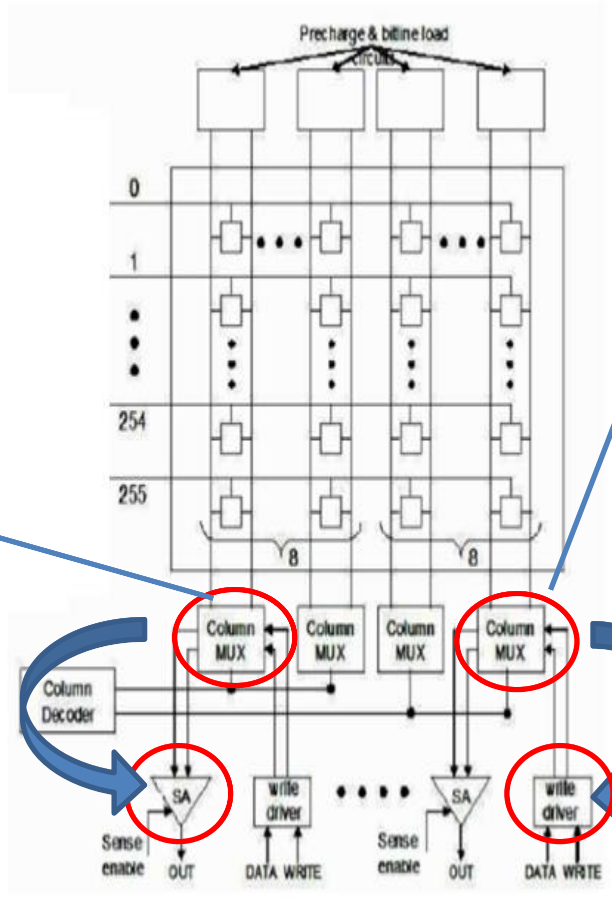
Similar to the static decoder design, large decoders are built using a multilayer approach

Reference material: Digital Integrated circuits: A Design Perspective, Rabaey, Chandrakasan and Nikolic, 2nd edition.

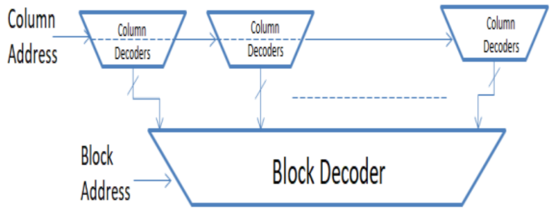
Column and Block Decoder

- Column Decoders should match with bit line pitch of the Memory Array.
- Column and block Decoder does the function of 2^k Input Multiplexer, **where k is size of the address word.**
- **For Read-Write Operations** these Multiplexers may be shared or separate.

During Read operation they have to provide the discharge path from pre-charged bit lines to sense Amplifiers (SA)

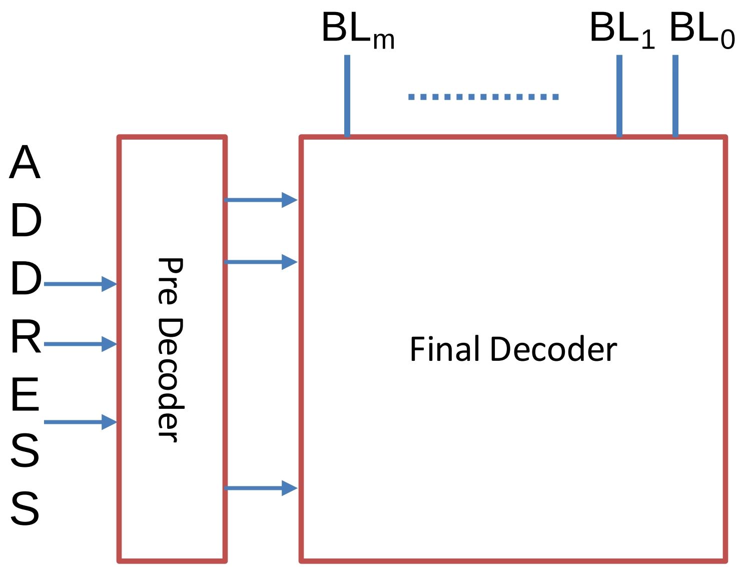


During Write operation they have to be able to drive the bit line low to write 0 in memory cell.

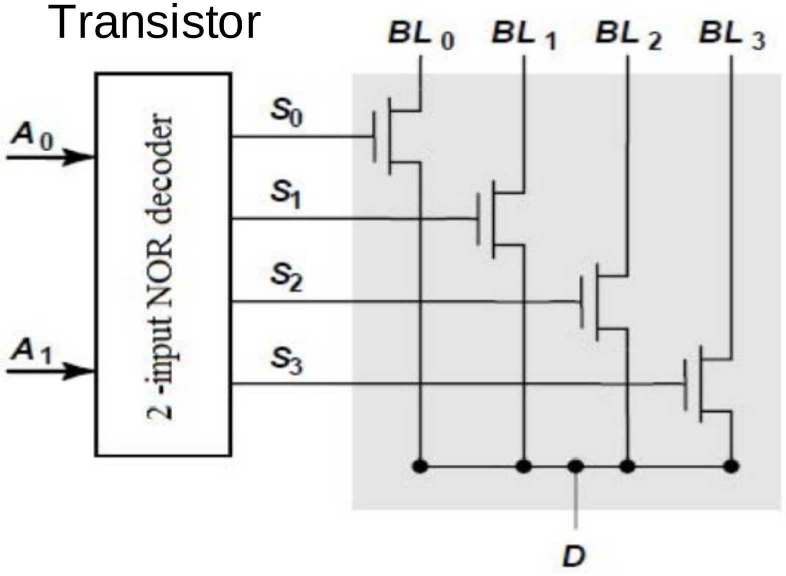


Column and Block Decoder

Column Decoder using NOR predecoder



1) MUX using NMOS Pass Transistor



2) MUX using CMOS Transmission Gate can also be used.

A1	A0	S3	S2	S1	S0	Out put D
0	0	0	0	0	1	BL0
0	1	0	0	1	0	BL1
1	0	0	1	0	0	BL2
1	1	1	0	0	0	BL3

- The Control Signal for the pass Transistors is generated by using k to 2^k predecoder.
- A 4:1 Column Decoder using nmos logic is as shown in the circuit.
- The speed of this approach is HIGH as Single pass Transistors in present in each path, which introduces minimal resistance in path.
- These use large number of Transistor i.e., **Predecoder+ Final Decoder**

$= (k+1) \times 2^k + 2^k$

Column and Block Decoder

Column Decoder using NOR predecoder

Calculate the Number of Transistor required for Column decoder using pass transistor logic for k=2

Soln:

Number of Transistors required(T) = **Predecoder + Final Decoder**

= $(K+1) \times 2^k + 2^k$

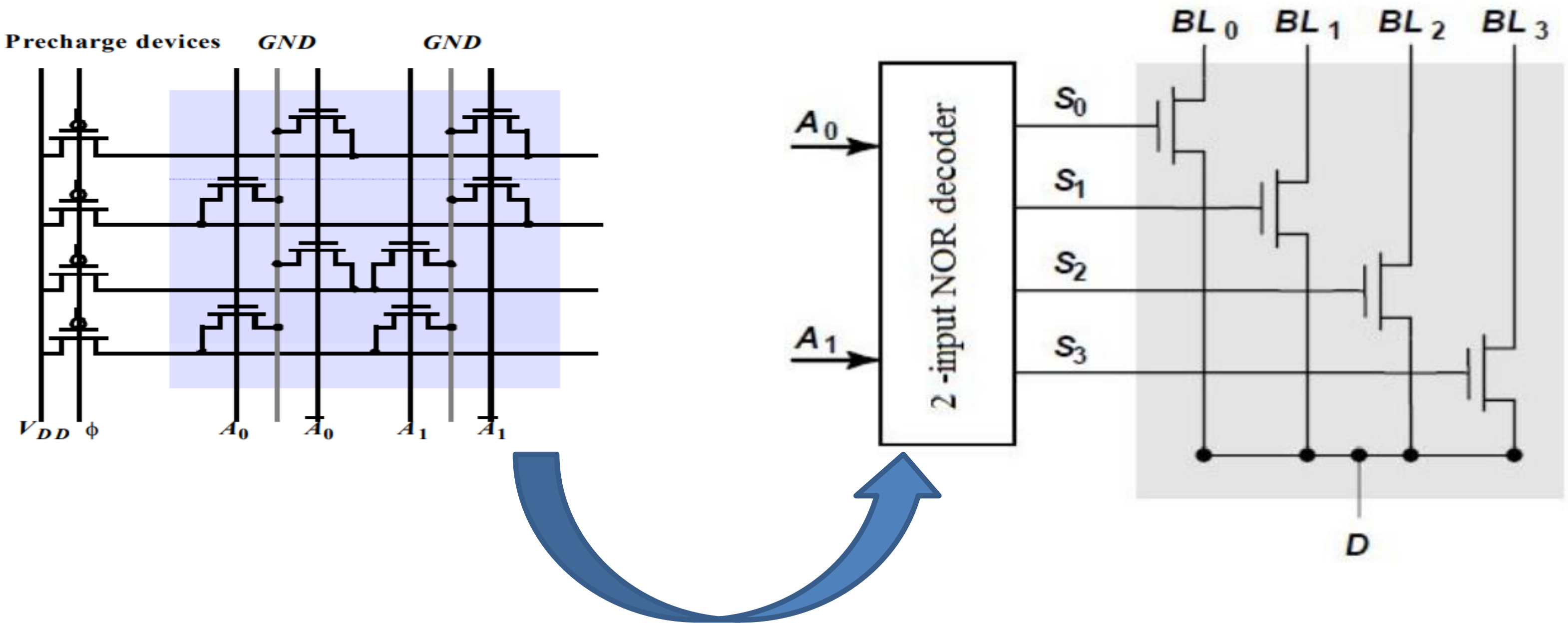
= $(2+1) \times 2^2 + 2^2$ =12 + 4 = 16

If 1024:1 Column Decoder is required, then k=10

No of Transistors

= $(10+1) \times 2^{10} + 2^{10}$

= **12,288**



Column and Block Decoder

Tree based Column Decoder

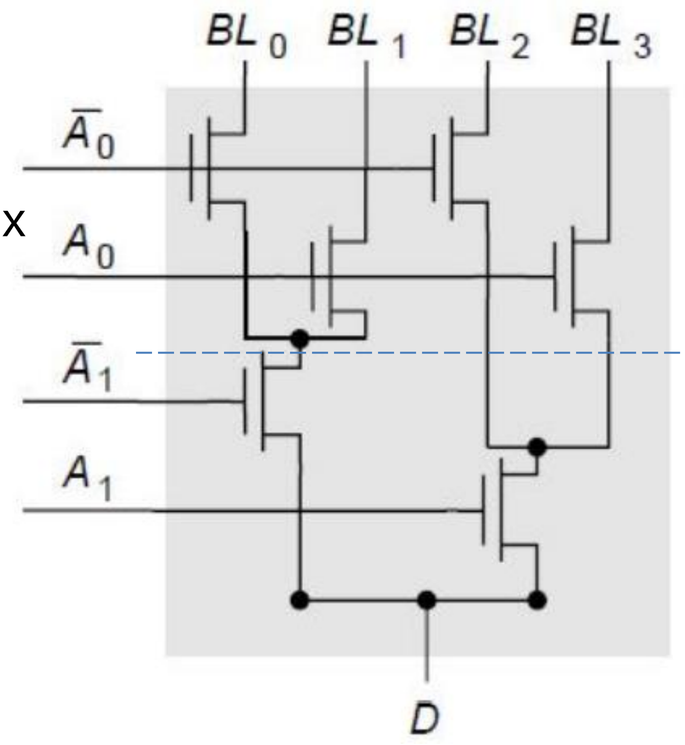
- It uses a binary reduction scheme.
- No need of predecoder
- No of Transistors required

$2^k + 2^{k-1} + 2^{k-2} + \dots\dots\dots\; 2^2 + 2^1 = 2x(2^k - 1)$

If 1024:1 (then k=10)Column Decoder is designed using Tree Decoder, then Number of Transistors required,

No of Transistors $= 2x(2^k - 1)$

$= 2x(2^{10}-1) =$ **2046**



$2^k = 2^2 = 4$

$2^{k-1} = 2^1 = 2$

Number of Transistors is reduced by a factor 6!

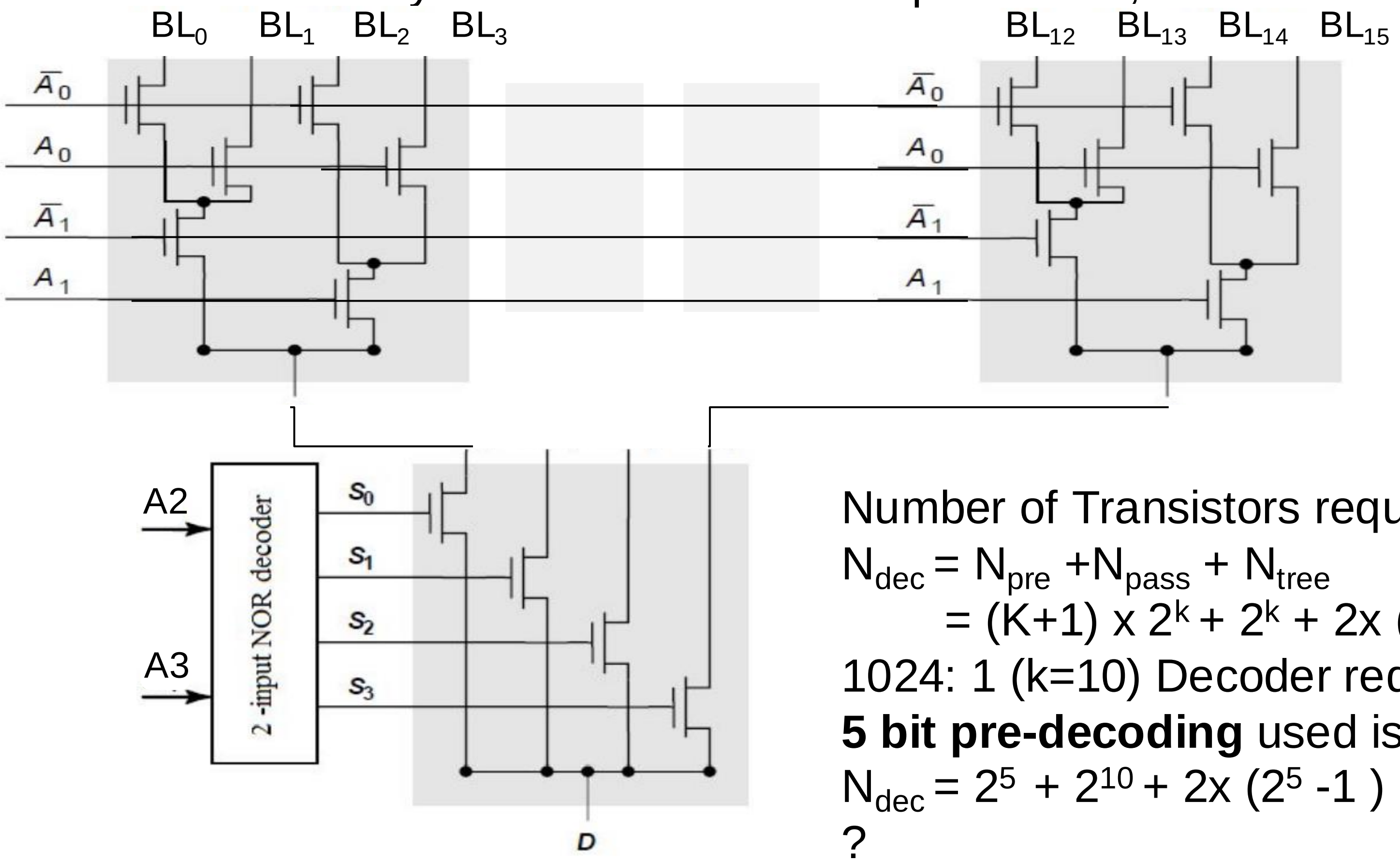
Drawback : A chain of k series connected pass Transistors is inserted in the signal path. It makes the tree approach slow which can be addressed by using buffers or progressive sizing.

Column and Block Decoder

Hybrid Column Decoders

In this, both predecoder and Tree decoder is combined to use the advantages of both approaches.

Let us see how a 4:16 Hybrid Decoder can be implemented,



Fraction of the address word is predecoded (MSB bits) and the remaining bits are tree decoded.

Ex: 4:16 , 2 bits are pre-decoded and 2 remaining bits are Tree decoded.

Number of Transistors required

$N_{dec} = N_{pre} + N_{pass} + N_{tree}$

$= (K+1) \text{ x } 2^k + 2^k + 2\text{x } (2^k - 1)$

1024: 1 (k=10) Decoder requires if **5 bit pre-decoding** used is

$N_{dec} = 2^5 + 2^{10} + 2\text{x } (2^5 - 1) = 1278$

?

THANK YOU

Mahesh Awati/Dr Shashidhar/M S Sunita

Department of Electronics and Communication

stantry@pes.edu

+91 9845695028