Memory Design and Testing

Mahesh Awati/Dr Shashidhar

Department of Electronics and Communication Engg.

SEMICONDUCTOR MEMORY DESIGN AND TESTING

UNIT 2 – Static Random Access Memory

Mahesh Awati/Dr Shashidhar

Department of Electronics and Communication Engineering

UNIT 2 Outline

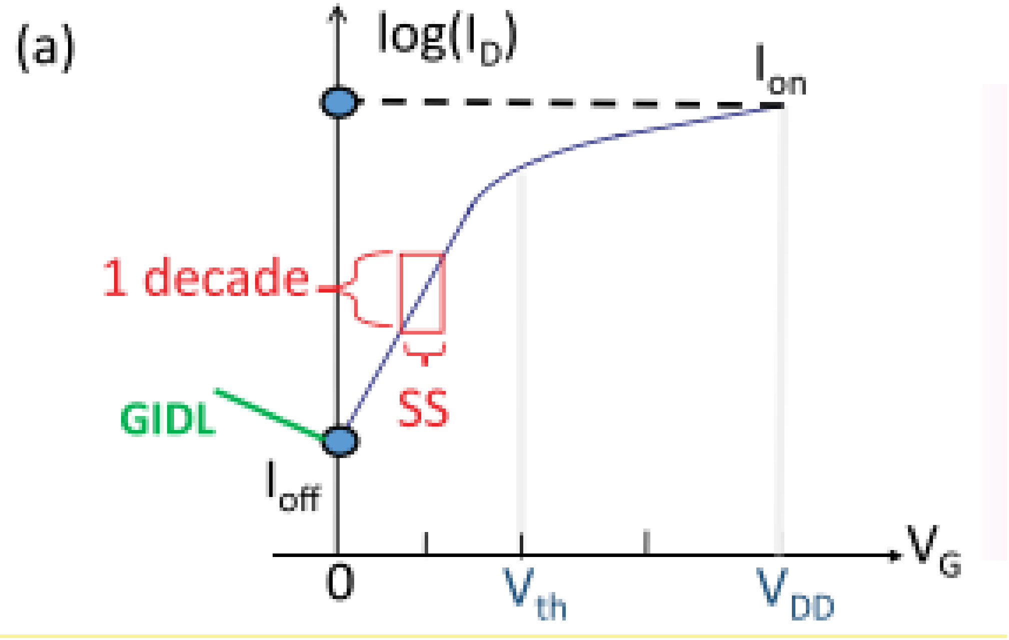
Unit 2: Static Random Access Memory

- SRAM’s leakage, variability and reliability
- SRAM layout and scaling trend
- FinFET based SRAM
- CAM topology, Binary CAM, Ternary CAM
- In memory computing, what is in memory computing
- Simple logic implementation using 6T SRAM

SRAM leakage

Transistor's sub threshold current

- It is due to carrier diffusion between source and channel
- Transistor's transfer characteristics is exponential decreasing log(I_D) v/s V_G
- Off state is defined as VG is zero VD as VDD
- On state is defined as VG is VDD and VD is also VDD

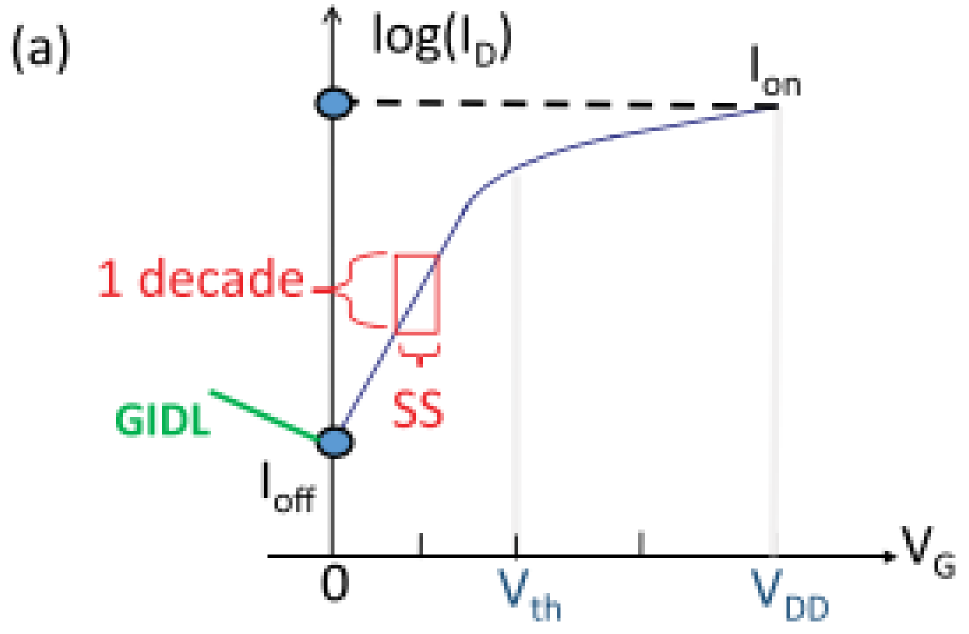


SRAM leakage

Transistor’s sub threshold current

- Subthreshold slope (SS) is defined in subthreshold region
- Where slope is given by,

$$S = \left(\frac{d \log I_D}{dV_G} \right)^{-1} = \frac{\partial V_G}{\partial \Psi_S} \frac{\partial \Psi_S}{\partial \log I_D} = \left(1 + \frac{C_{dm}}{C_{ox}} \right) \frac{kT}{q} \ln(10) = m \times 2.3 \frac{kT}{q}$$



SRAM leakage

Transistor’s sub threshold current

- Two factor determine subthreshold slope

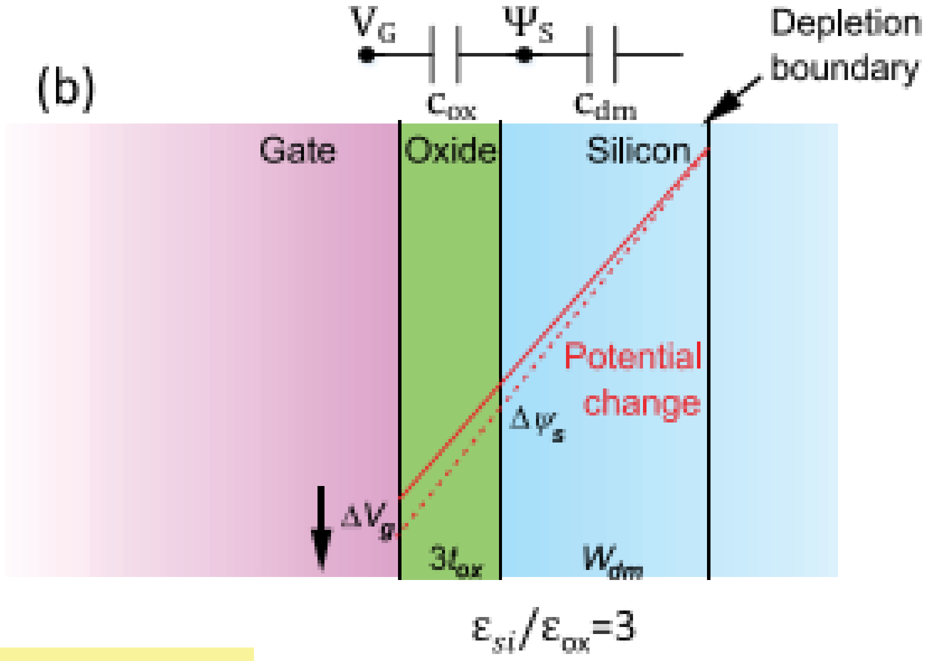
$$S = \left(\frac{d\log I_D}{dV_G}\right)^{-1} = \underbrace{\frac{\partial V_G}{\partial \Psi_S}}_{\text{Factor 1}} \underbrace{\frac{\partial \Psi_S}{\partial \log I_D}}_{\text{Factor 2}} = \left(1+\frac{C_{dm}}{C_{ox}}\right)\frac{kT}{q}\ln(10) = m\times 2.3\frac{kT}{q}$$

- Factor 1 reflects change of gate voltage with respect to surface potential
- Factor 2 reflects change of surface potential with respect to drain current

SRAM leakage

Transistor's sub threshold current

- Factor 1 modelling as voltage divider model



$$\frac{\Delta V_G}{\Delta \psi_s} = m = \left(\frac{C_{ox} + C_{dm}}{C_{ox}}\right) = \frac{W_{dm} + 3t_{ox}}{W_{dm}}$$

$$C_{ox} = \varepsilon_{ox}/t_{ox},$$

$$C_{dm} = \varepsilon_{si}/W_{dm},$$

SRAM leakage

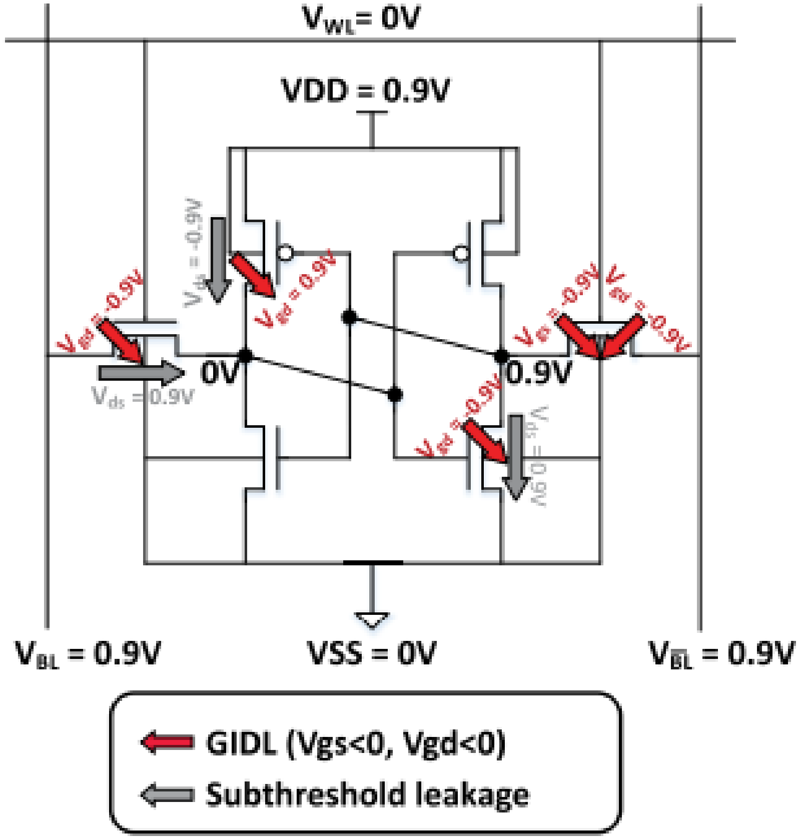
Transistor’s sub threshold current

- Factor 2 is dependent on physical constant.
- SS has lower limit or 60mv/dec means a change of 60mv for a decade of leakage current change
- Transistor may also suffer from Gate induced drain leakage(GIDL), occurs when gate potential is more negative than drain/source potential

SRAM leakage Reduction

Transistor’s leakage paths in SRAM

- Leakage paths when wordline is not active and bitline precharged
- Sum of these currents could lead to leakage power of nW for a cell and mW for MB level cache

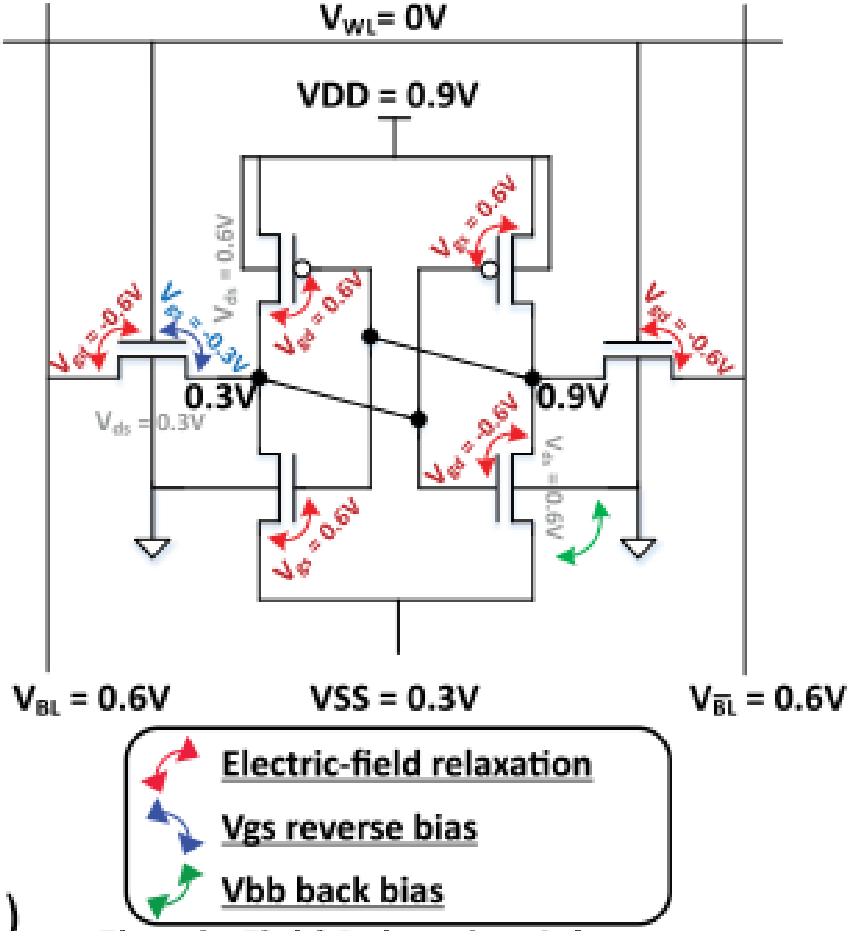


(a) Conventional SRAM cell

SRAM leakage Reduction

Transistor’s leakage paths in SRAM

- Leakage path currents can be reduced by high Vth devices
- Or we can raise lower voltage little bit

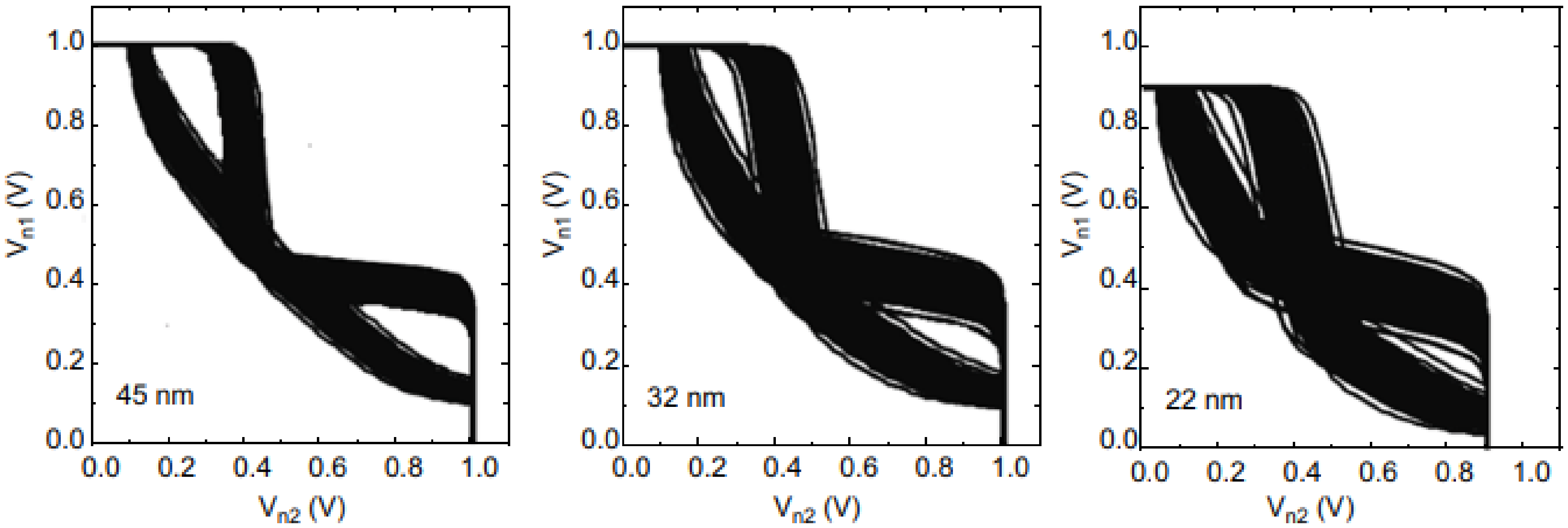


(b) Electric-Field Relaxation Scheme

Variability and Reliability

Transistor's intrinsic parameter fluctuations and impact on SRAM stability

- Variations in V_{th}, I_{on} and I_{off} of transistors
- Results in transistor mismatch and variations in butterfly curve
- Suppression in HSNM and RSNM curves

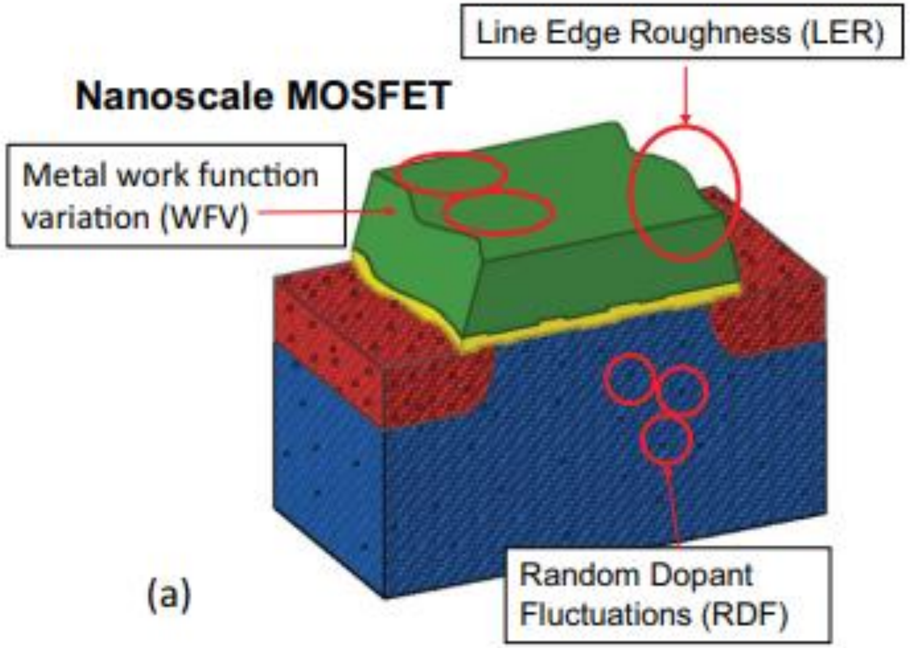


Butterfly curves

Variability and Reliability

Reason for variations

- Random Dopant Fluctuations(RDF)
- Line Edge Roughness(LER)
- Metal Work Function variation (MFV)



Line Edge Roughness (LER)
Nanoscale MOSFET
Metal work function variation (WFV)
Random Dopant Fluctuations (RDF)

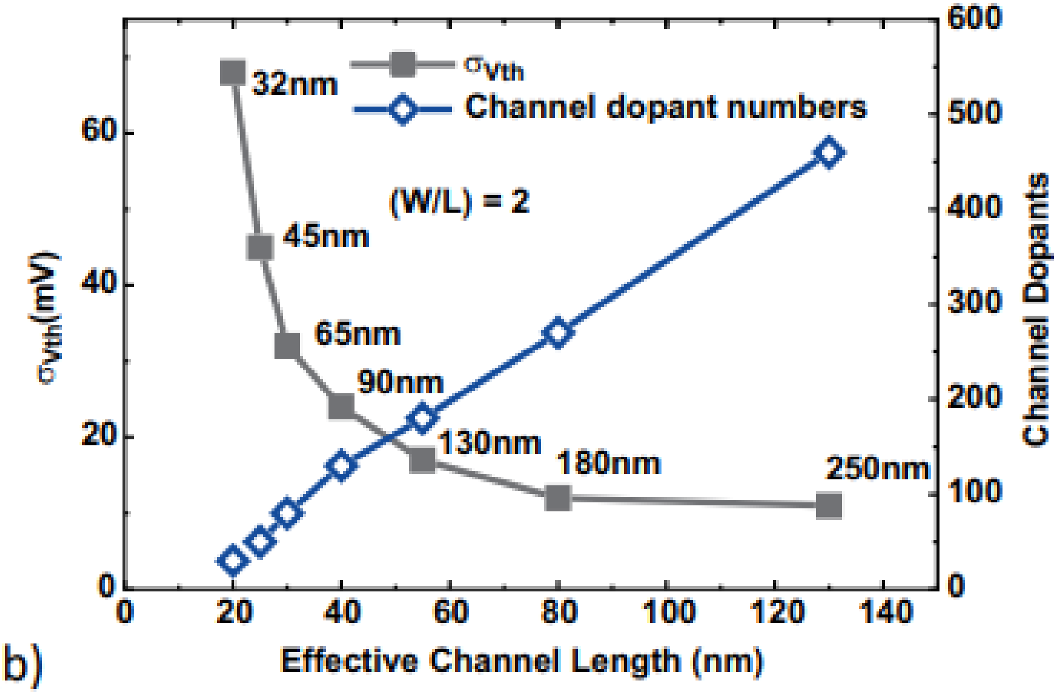
(a)

Variability and Reliability

Reason for variations

- Random Dopant Fluctuations(RDF)
- Line Edge Roughness(LER)
- Metal Work Function variation (MFV)

$$\sigma V_{th(RDF)} = \frac{q}{C_{ox}}\sqrt{\frac{N_d W_{dm}}{3LW}}$$



(b)

Variability and Reliability

Reason for variations

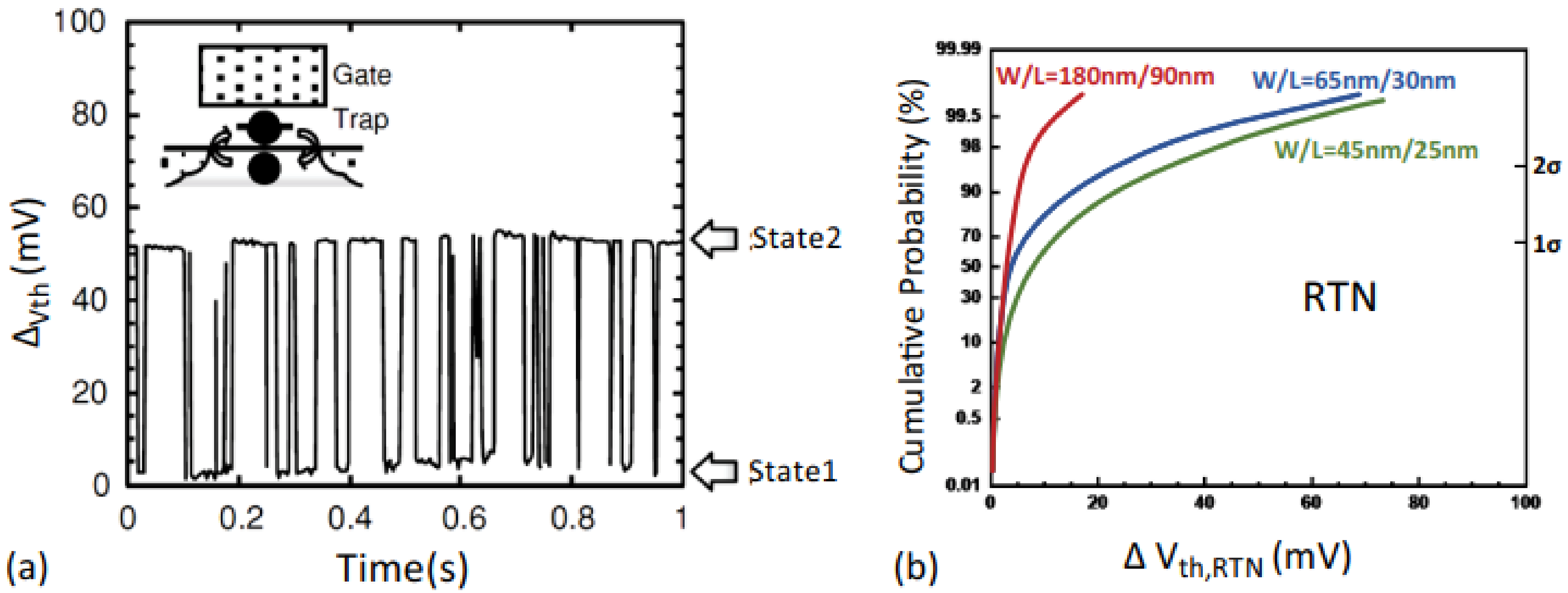
- LER is due to imperfect lithography, typical roughness of 1-2nm, a significant number is smaller dimension transistors
- WFV is due to metal alloys used for threshold voltage setting

$$\sigma V_{th(\mathrm{total})}{}^2 = \sigma V_{th(\mathrm{RDF})}{}^2 + \sigma V_{th(\mathrm{LER})}{}^2 + \sigma V_{th(\mathrm{WFV})}{}^2$$

Variability and Reliability

Temporal reliability issues

- These are random telegraphic noise and bias temperature instability

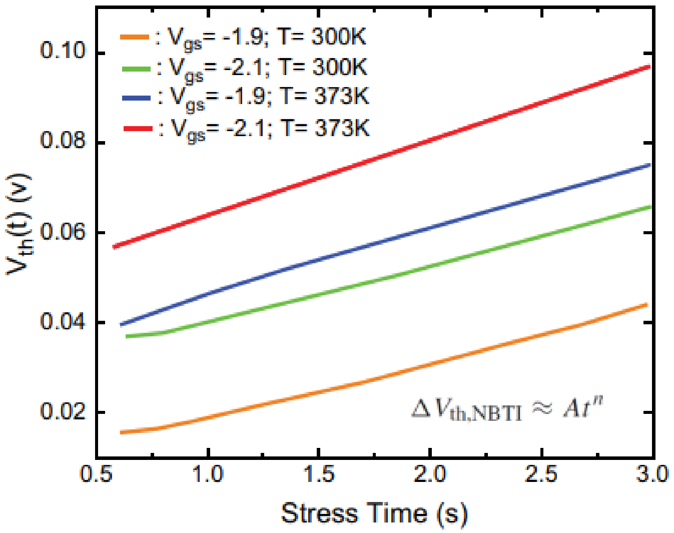


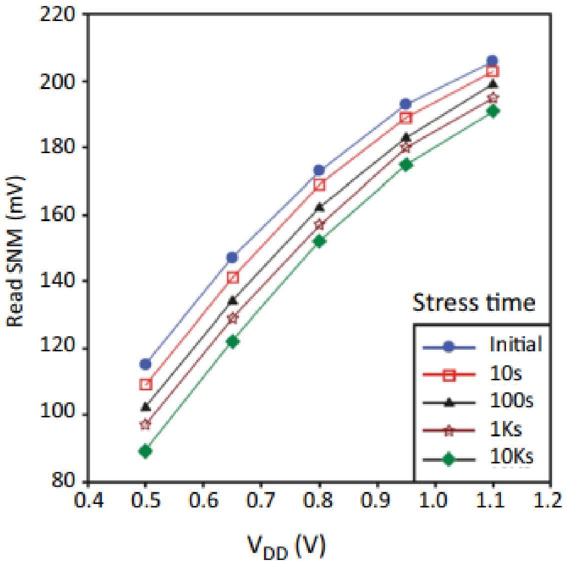
Effect of random telegraphic noise

Variability and Reliability

Temporal reliability issues

- Bias temperature instability is long term reliability issue, a drift in Vth happening due to stress
- More frequently seen in PMOS transistor





$$V_{\text{th,NBTI}}(t) = At^n$$

Where A and n are constants

Variability and Reliability

Soft errors due to radiation effect

- Single event upset (SEU) – Short term effect
- Total Ionizing Dose effect (TID) Long term effect

- These effects flip state of SRAM

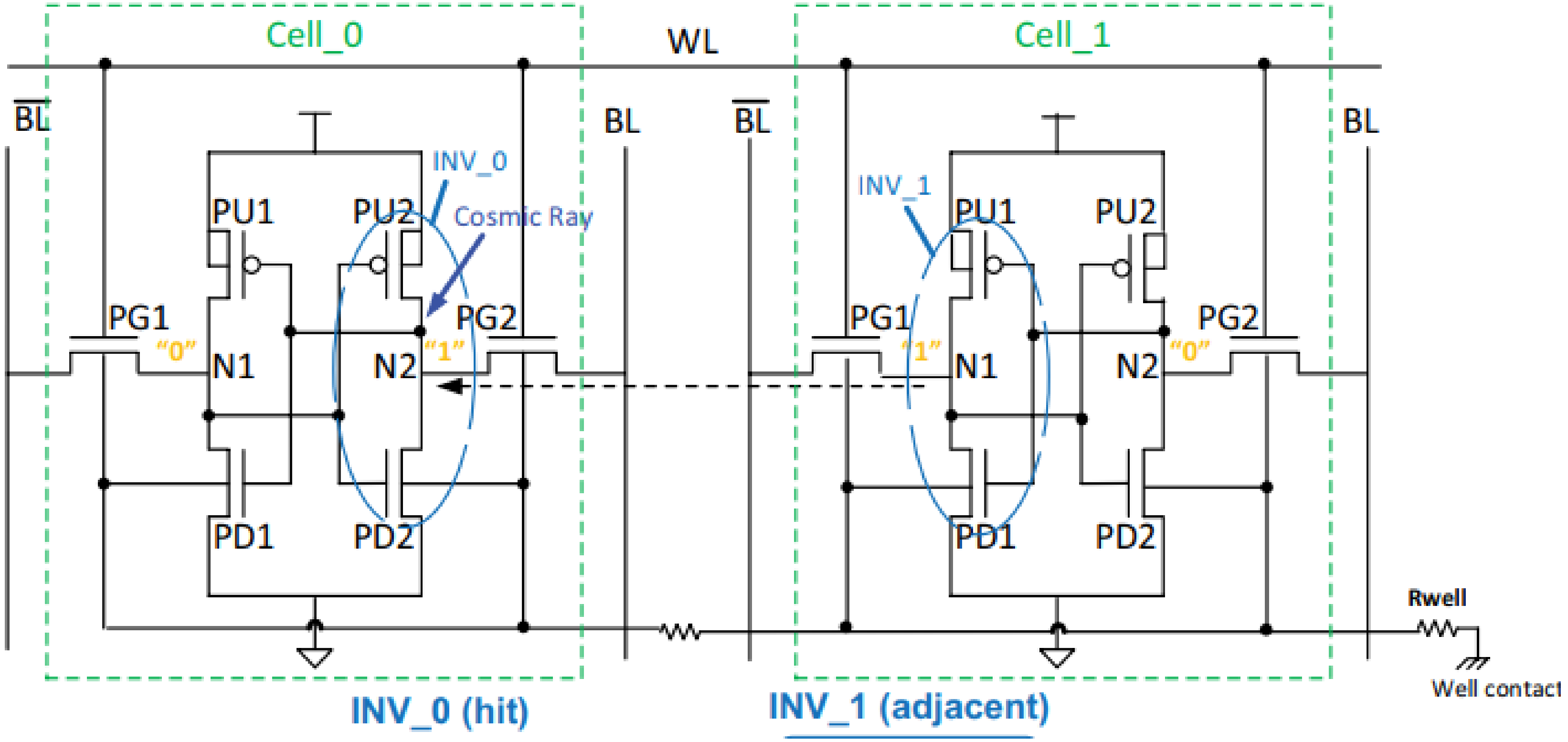
- Radiation effects are due to alpha particles and cosmic rays

- A noise current generated by these effects

- These are temporary failures hence called soft errors

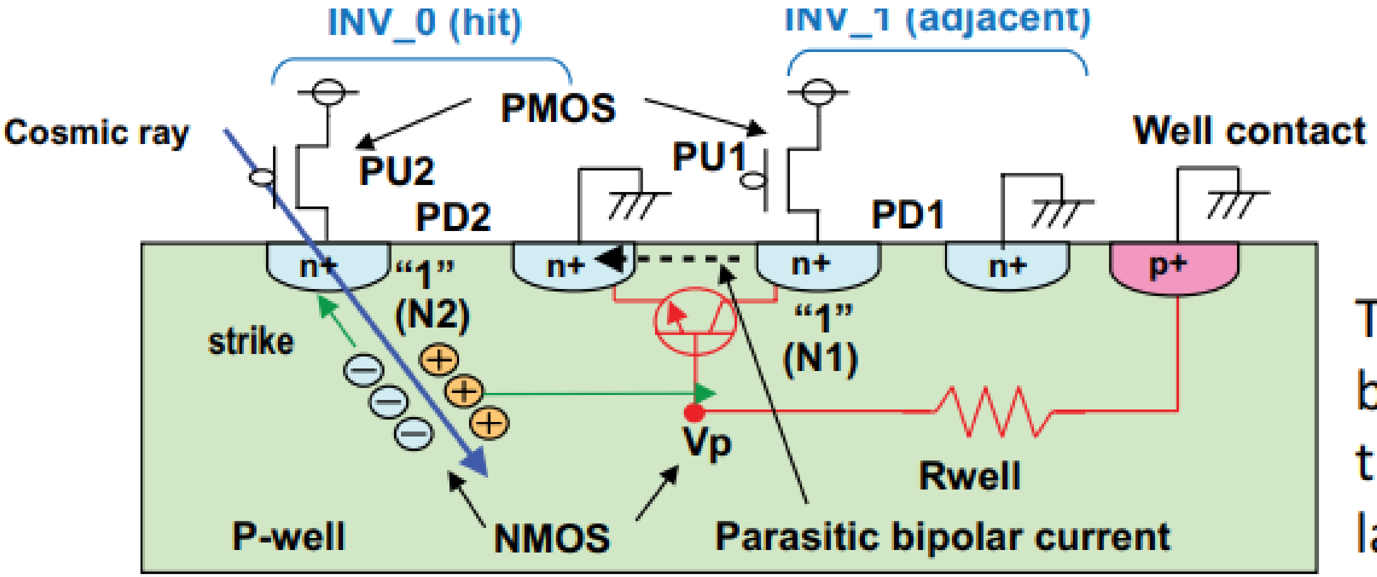
Variability and Reliability

Soft errors due to radiation effect



Variability and Reliability

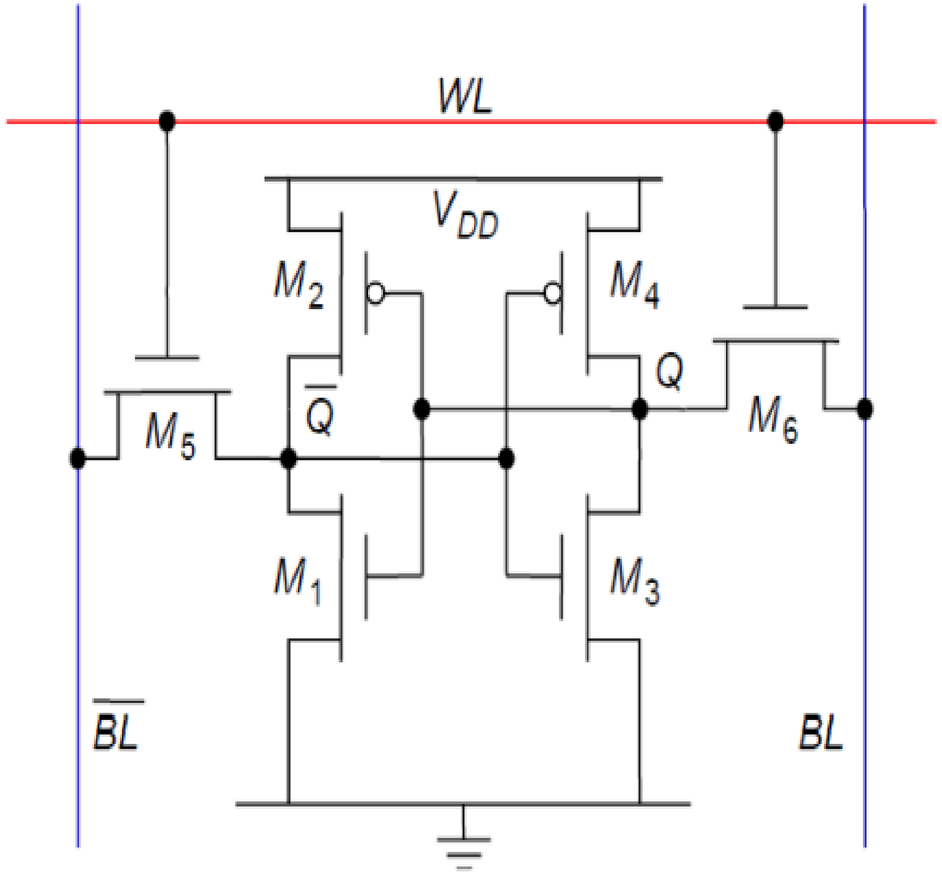
Soft errors due to radiation effect

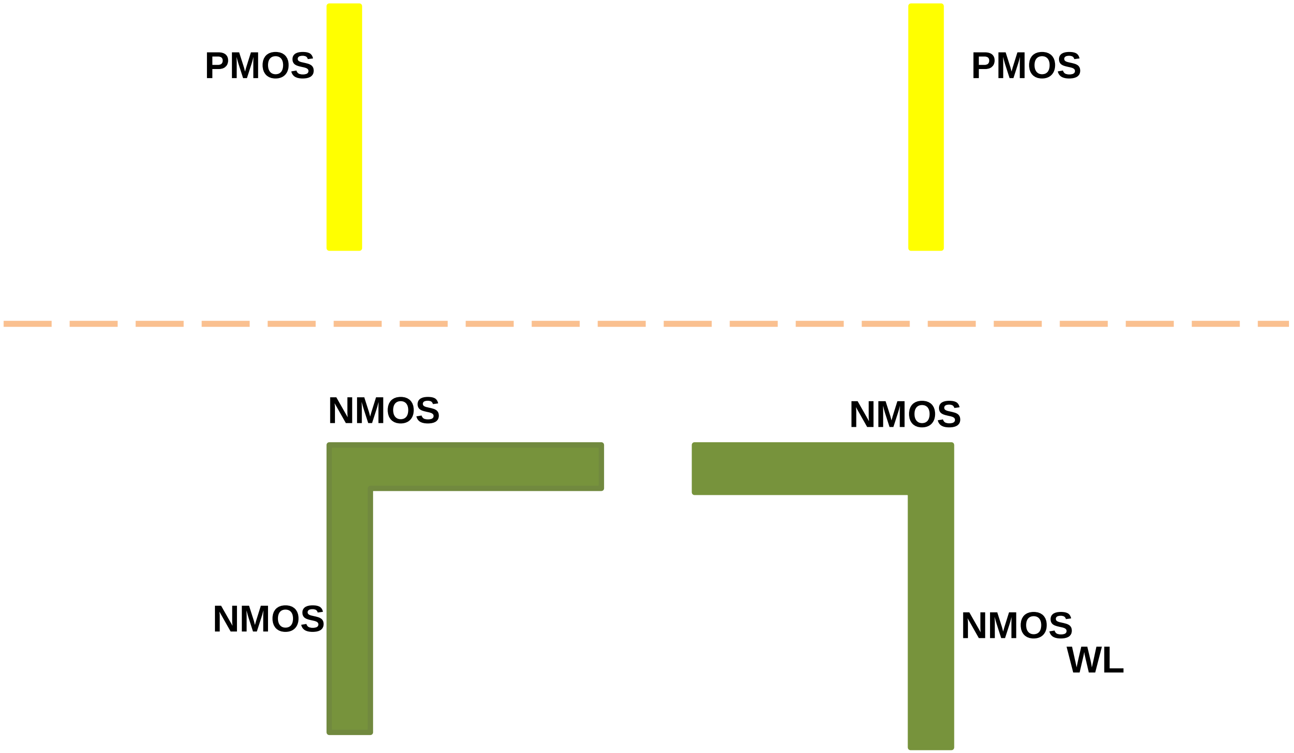


The parasitic n-p-n bipolar junction transistor leads to latch-up effect

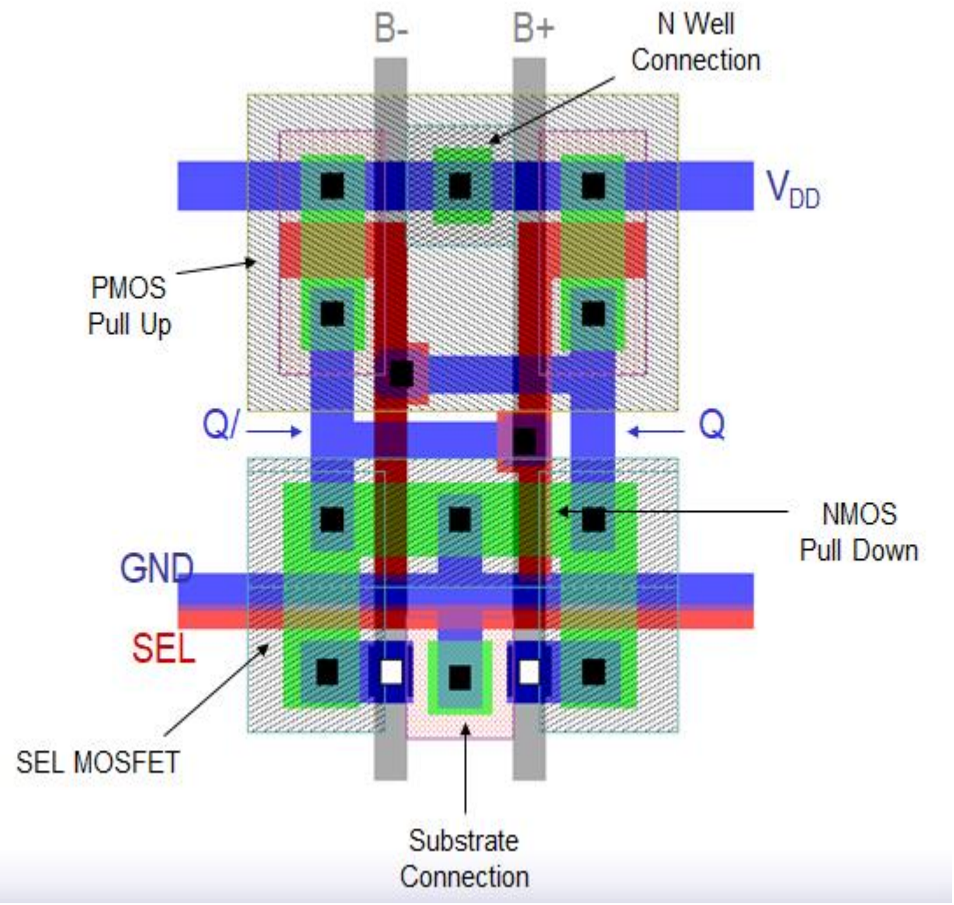
<u>**Silicon on Insulator can be used to avoid this effect**</u>

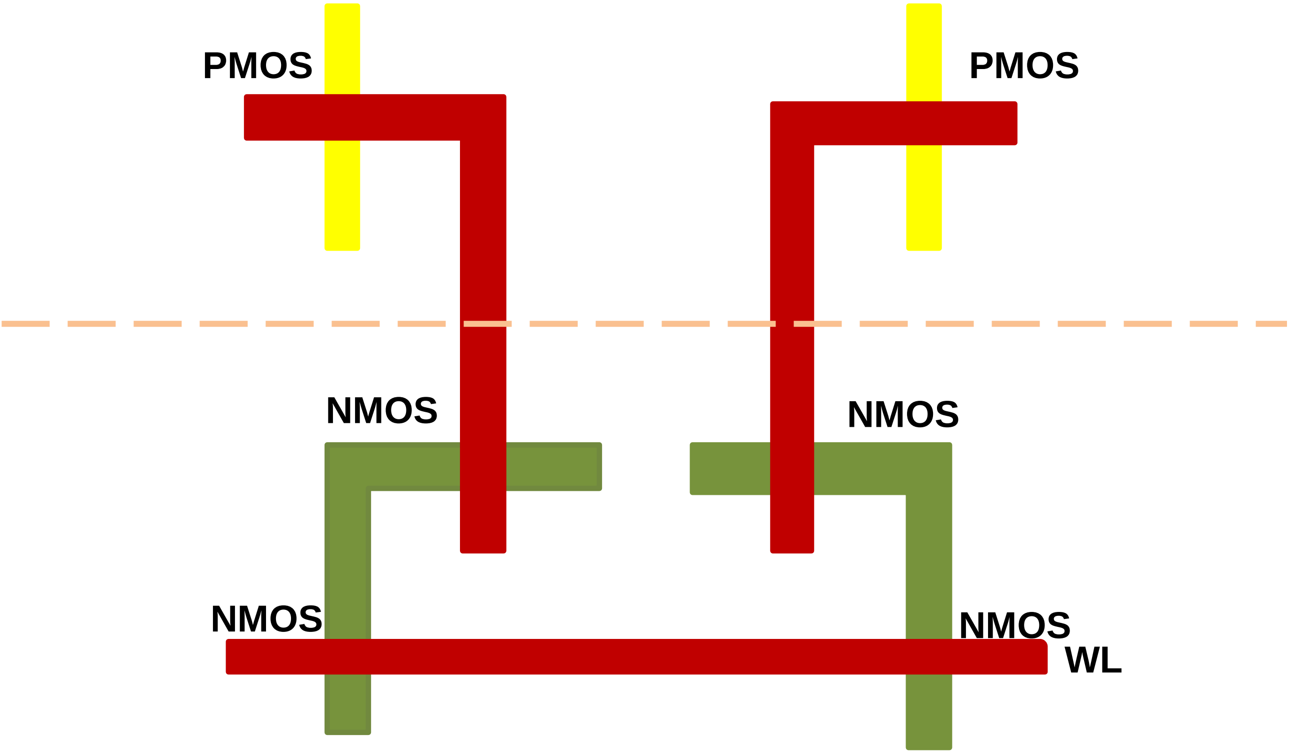
6T SRAM Cell – LAYOUT



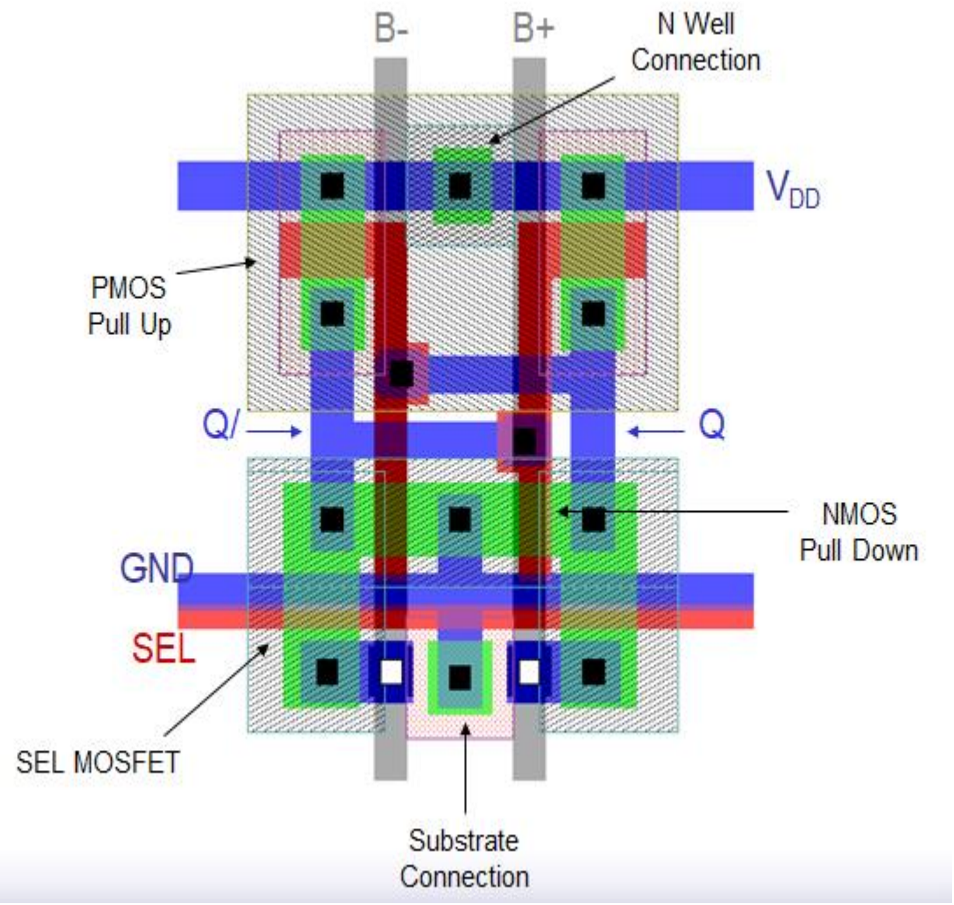


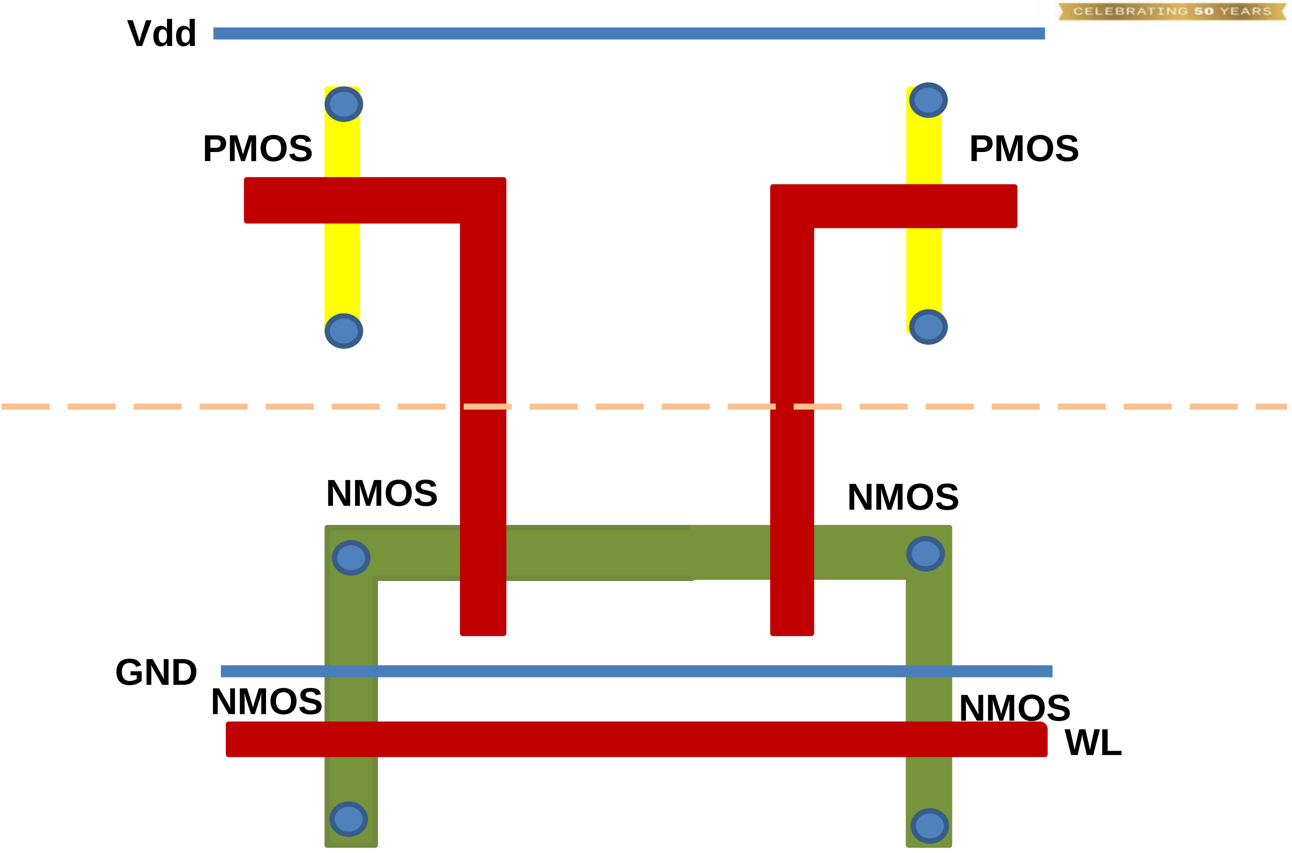
6T SRAM Cell – LAYOUT



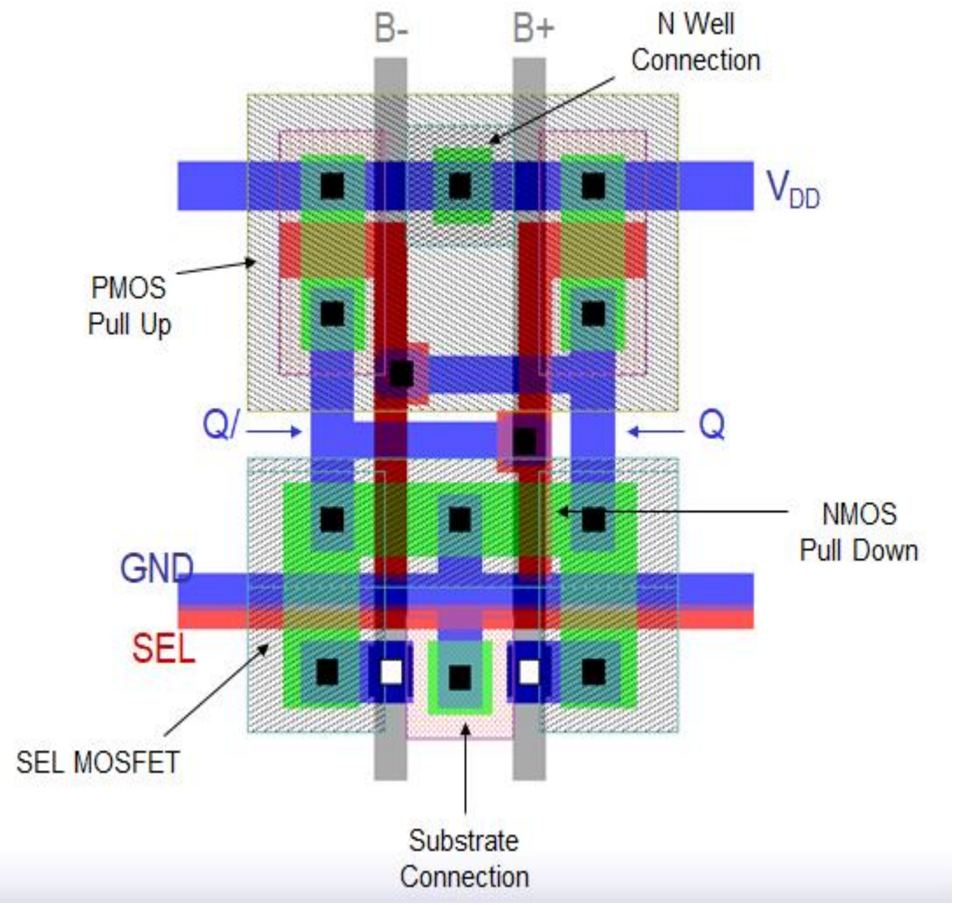


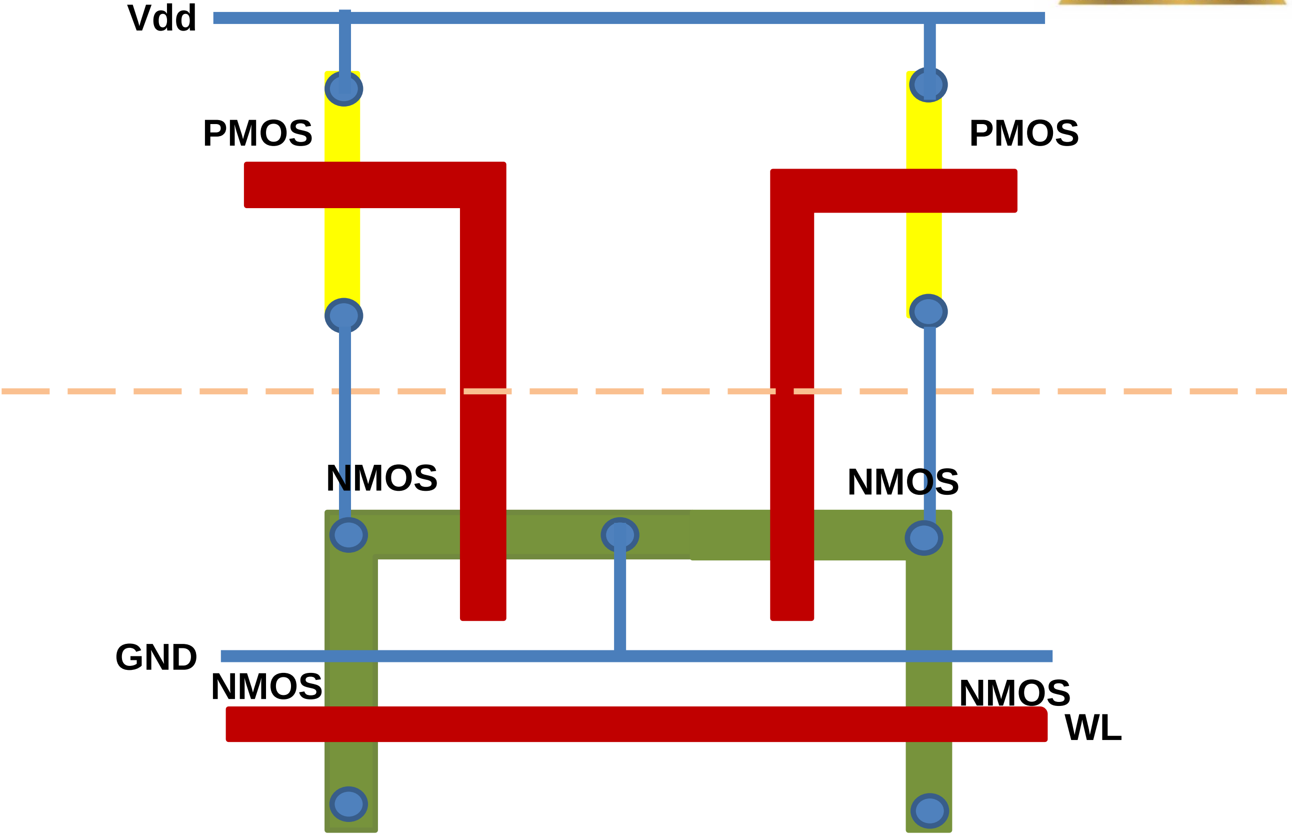
6T SRAM Cell – LAYOUT



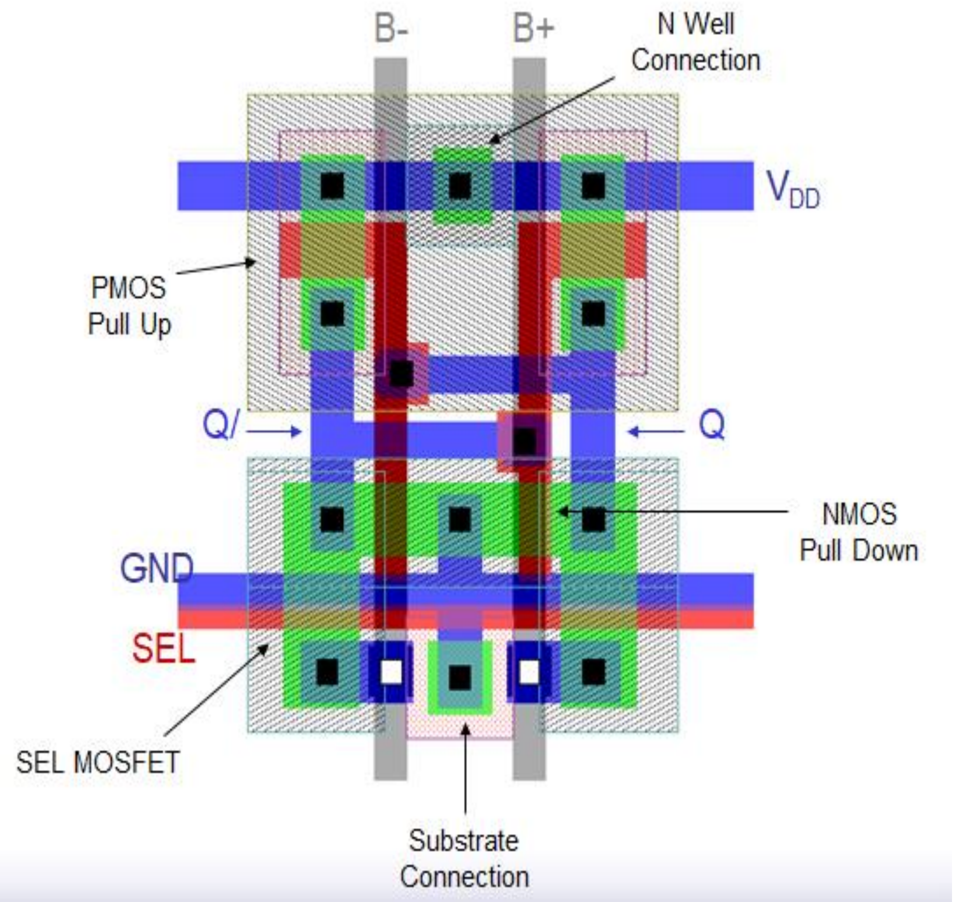


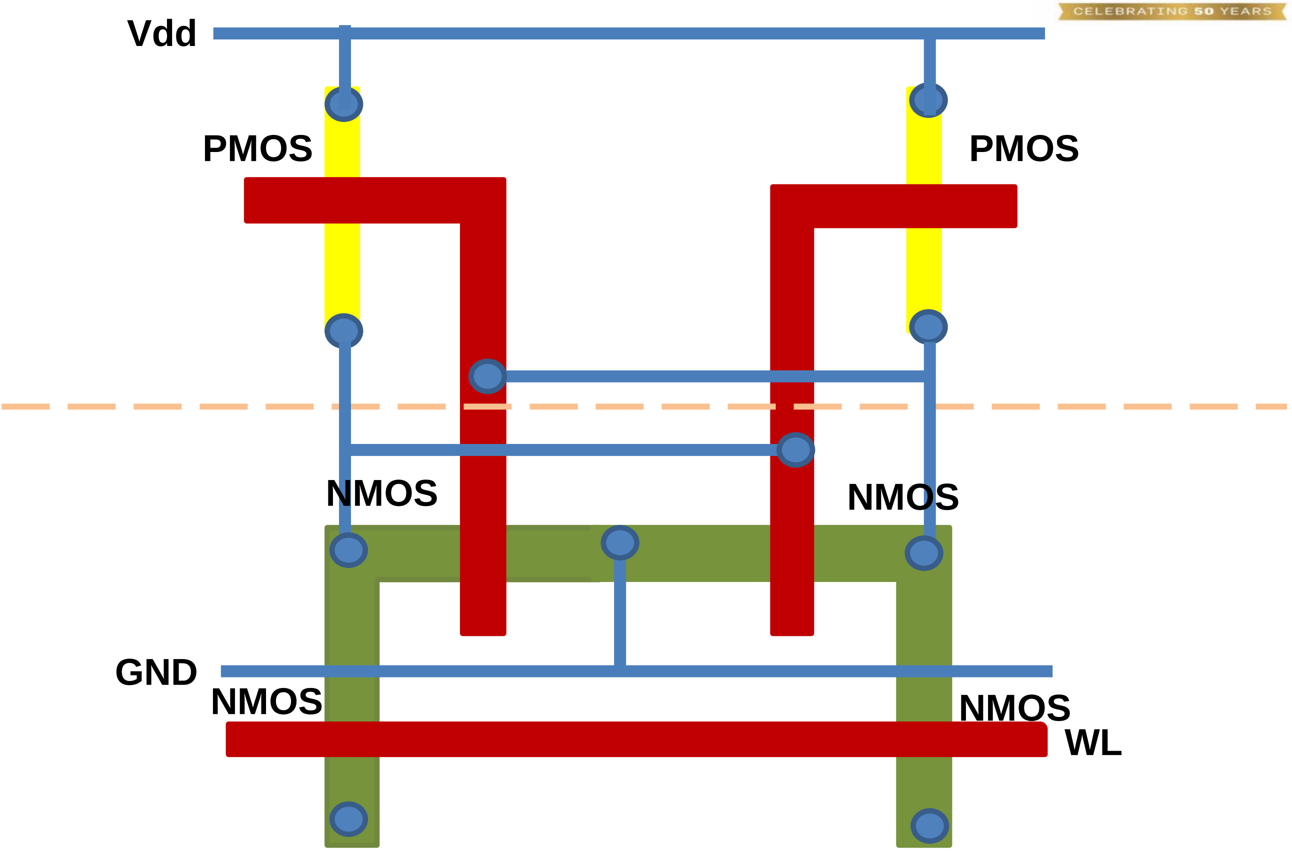
6T SRAM Cell – LAYOUT



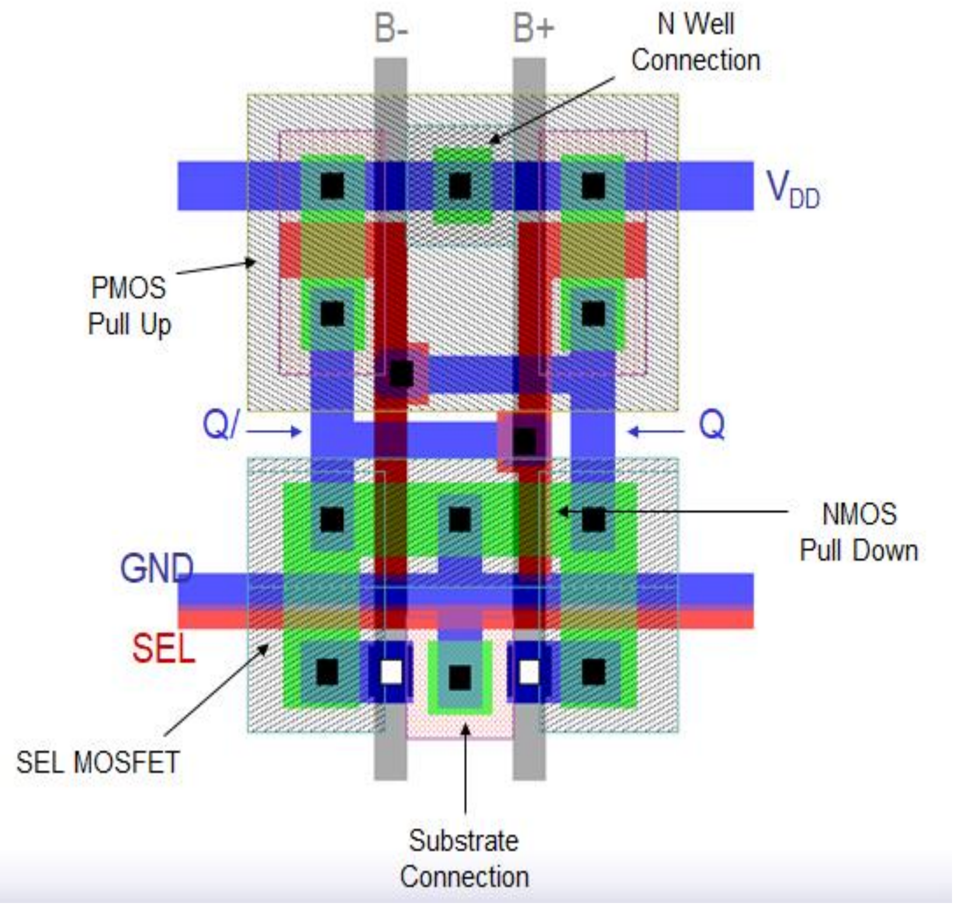


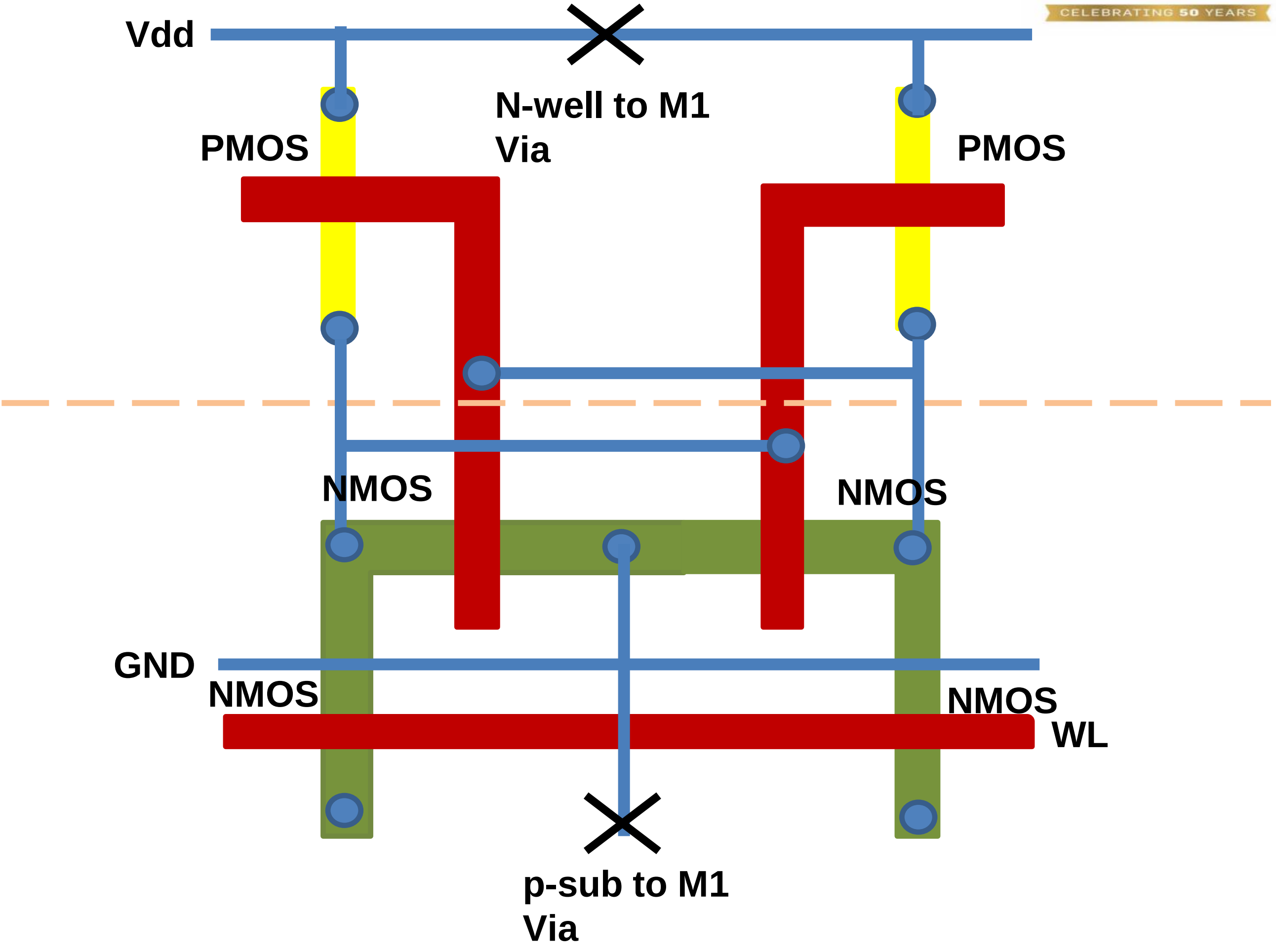
6T SRAM Cell – LAYOUT



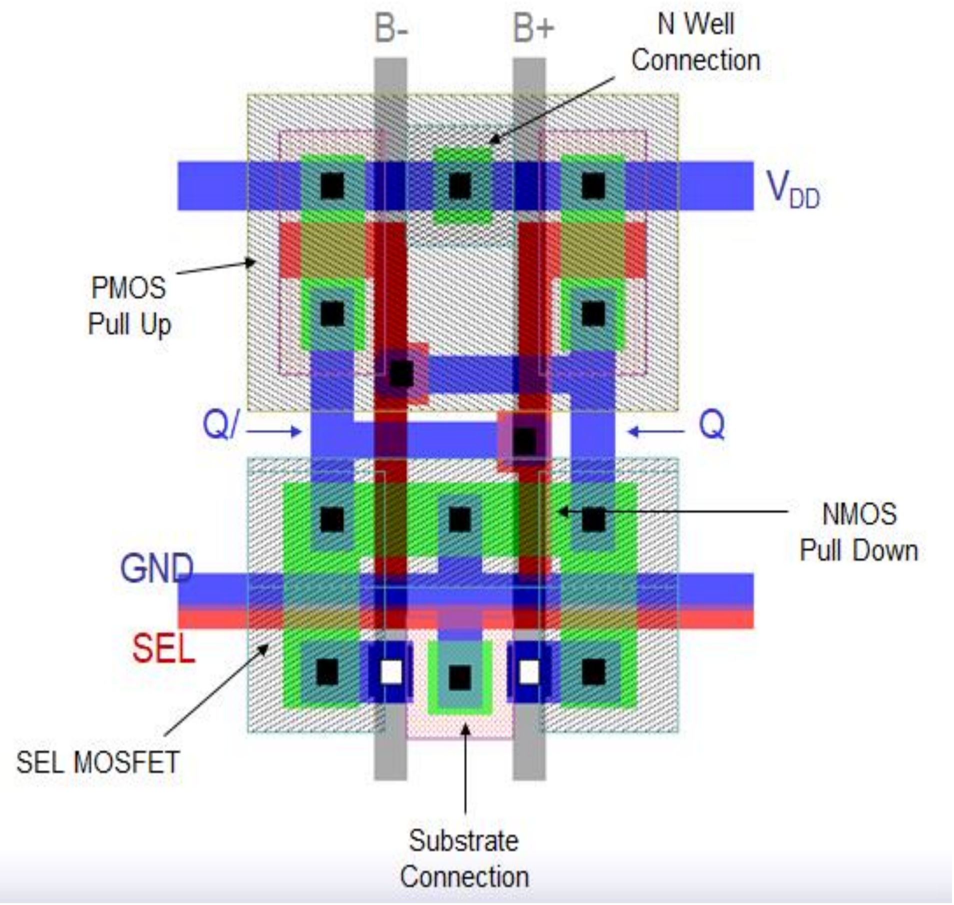


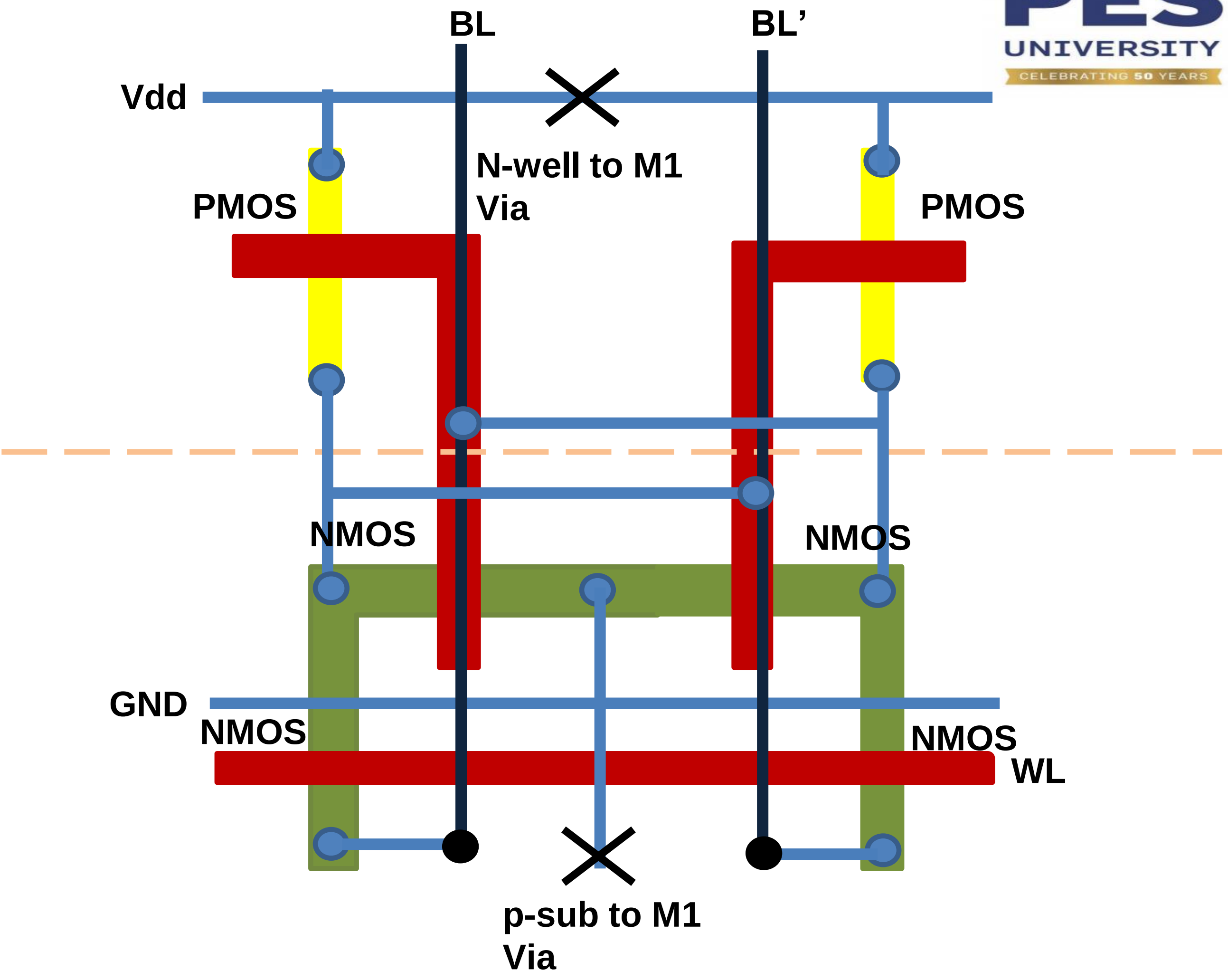
6T SRAM Cell – LAYOUT





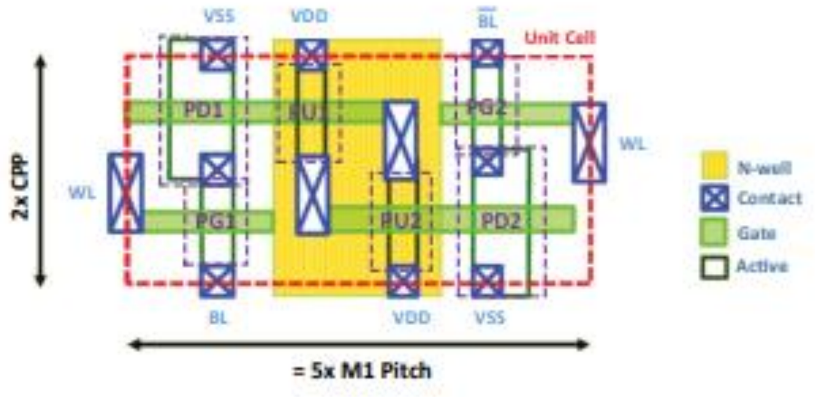
6T SRAM Cell – LAYOUT





6T SRAM Cell – Representative LAYOUT

- Representative layout diagram for technology node from 90nm to 32nm.
- Gate aligned horizontally and channel vertically
- PD:PG:PU aspect ratio is 2:1:1
- Cell area approximately 160F^2



CPP = Contact Width + 2 x Spacer + L_G
~ M1 Pitch ~ 4F

Here M1 pitch is not minimum, W/L=2 for PD

SRAM unit cell area
= 2 CPP x5 M1 Pitch
~ 10 CPP^2~160 F^2

6T SRAM Cell – LAYOUT Microscopic picture

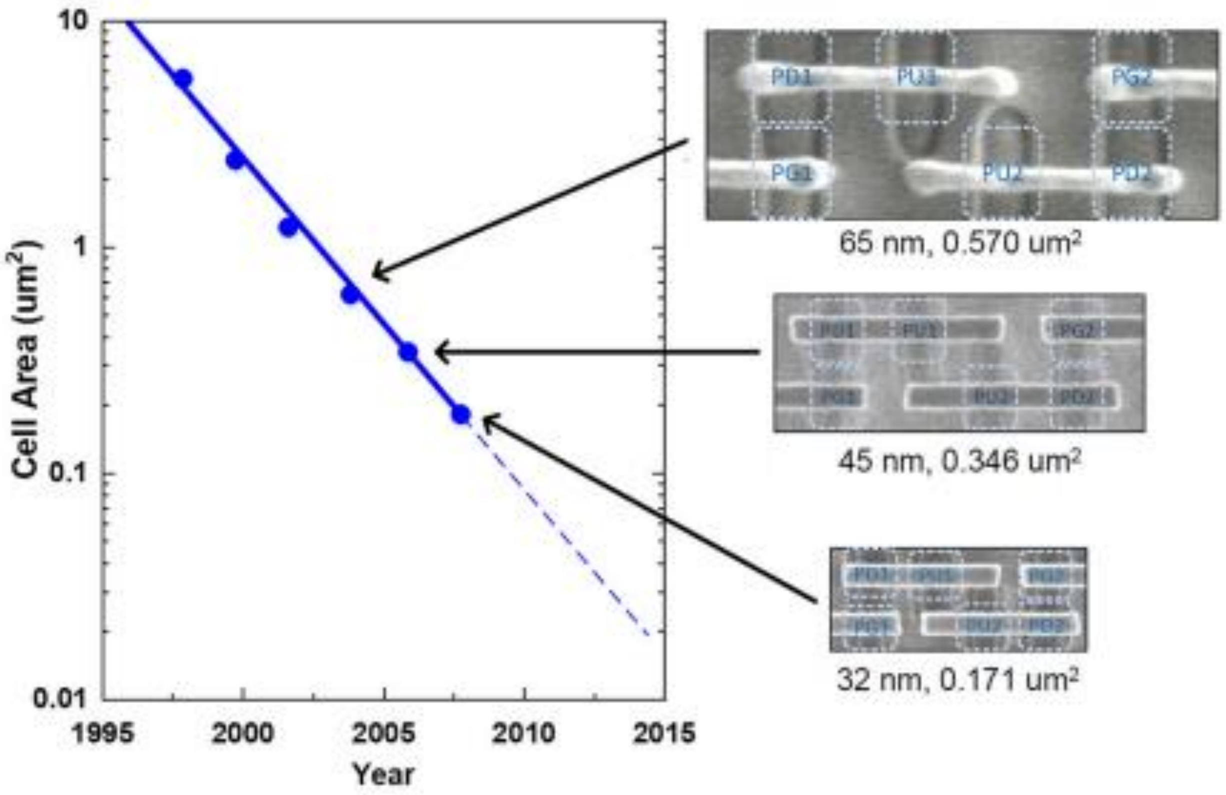


FIGURE 2.29 The SRAM layout scaling of the absolute cell area (in μm^2) and microscopic top-view images of the fabricated SRAM 6T cell in the planar transistor era.

CONTENT ADDRESSABLE MEMORY - CAM

Normally, RAMs are accessed by supplying an address.

The memory returns the data word stored in it.

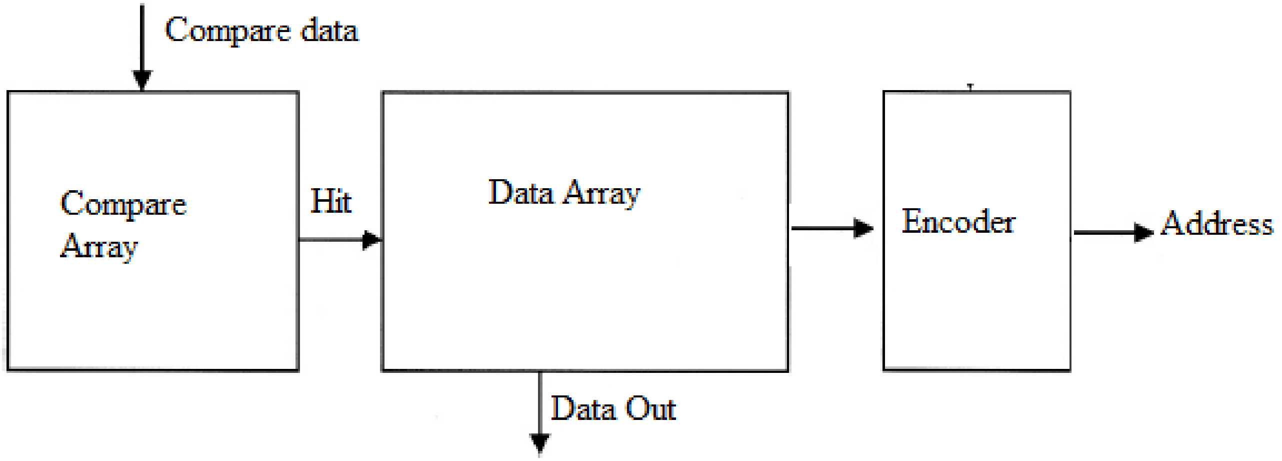
A CAM is designed such that the user supplies a data word and the CAM returns one or more addresses where the word was found.

CAM topology

A CAM contains 2 sections of memory – Compare Array and Data Array.

The Compare Array selects which section of the Data Array to read or write.

CONTENT ADDRESSABLE MEMORY



When a match occurs in the compare array with the data that is being applied at the inputs of the CAM, a “hit” occurs.

The “Hit” forces a Word Line in the data array to become active.

Data array is an SRAM array and accessing it is identical to SRAM Read operation.

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

CONTENT ADDRESSABLE MEMORY

The Word line outputs can be encoded to obtain the address of the location in which data is stored.

One or more entries can have a Hit.

Arbitration logic may be included to select which entry to send to the CAMs output.

A Priority Encoder is used to obtain the address, in such case.

To perform the function of Content based addressing requires the comparison of the input data with all of the possible locations of the compare array.

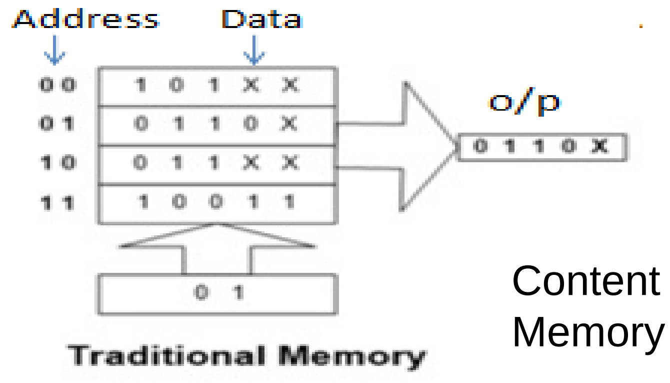
This comparison is done using XOR and NOR gates as shown

Content Addressable Memory (CAM)

Content Addressable Memories (CAM) has the ability to compare all stored data in parallel with incoming data in an efficient manner to identify the Memory address / addresses where the referred data is present.

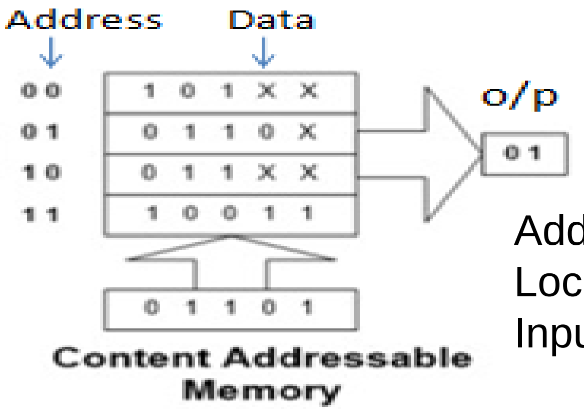
- CAM Architecture support 3 modes of Operations i.e., Read, Write and Match.
- Read and Write modes access and manipulate the data in CAM Array similar to ordinary memory.

- In MATCH Mode , It takes the DATA to be searched and finds the Memory locations in which the data is found.
- In case Multiple addresses holding the data searched then uses the Priority Encoder.



Address	Data
00	1 0 1 X X
01	0 1 1 0 X
10	0 1 1 X X
11	1 0 0 1 1

Content of addressed Memory location

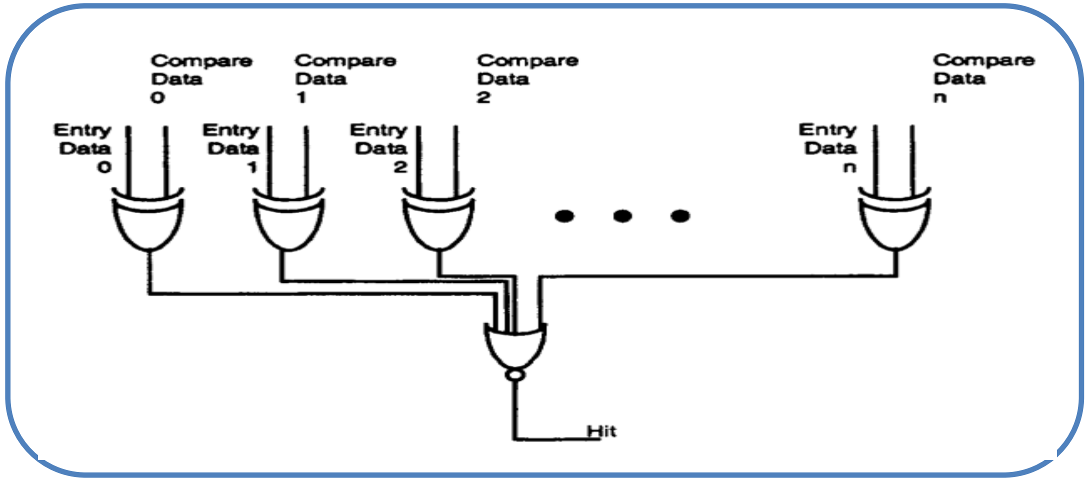


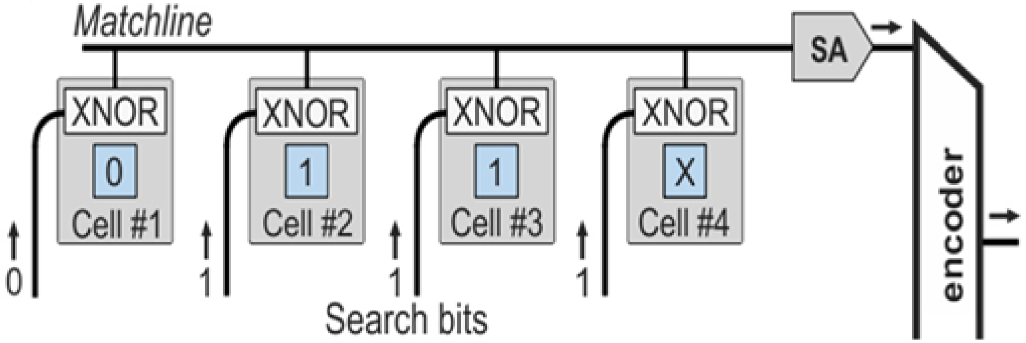
Address	Data
00	1 0 1 X X
01	0 1 1 0 X
10	0 1 1 X X
11	1 0 0 1 1

Address of Memory Location with which Input Data Matches

Content Addressable Memory (CAM)

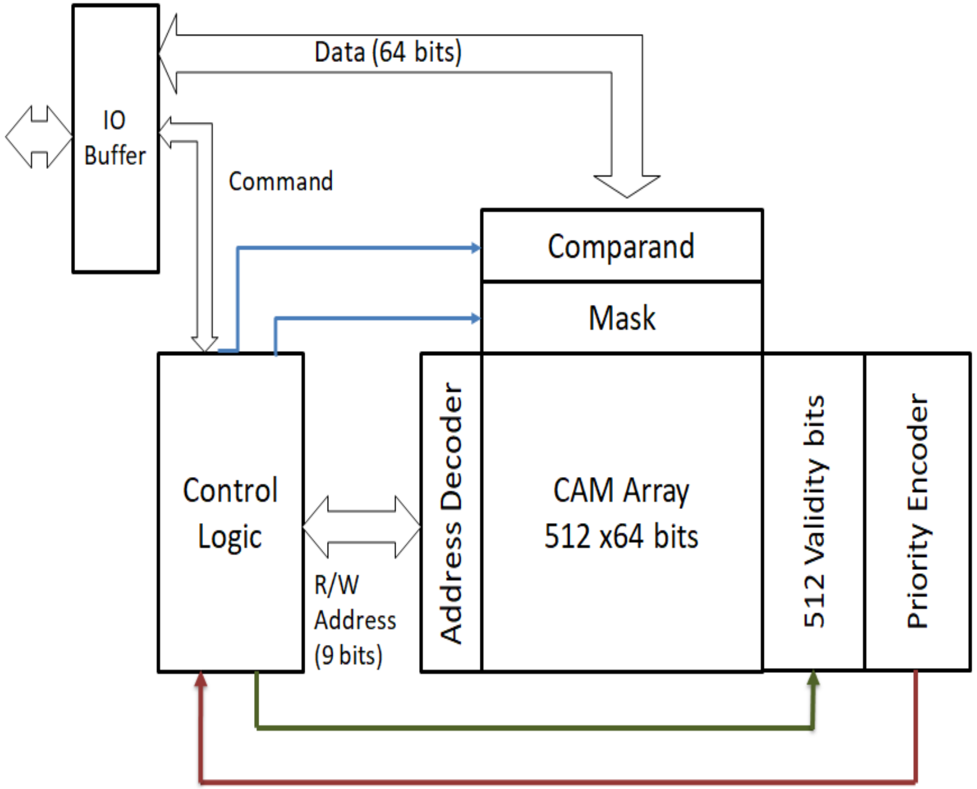
- One or more entries can have a hit in a given cycle.
- Depending on the CAM architecture, a multiple hit situation may be tolerable and **arbitration logic may be included to select which entry to send to the CAM's output.** A priority encoder is often used in this application. If multiple hits indicate an error condition, it should be identified as such at the CAM's output.
- Being able to perform a content based addressing **function requires a compare to be possible at all of the address locations in a memory**.
- A compare function is logically **done by XOR operation.**
- In compare mode, **each stored bit (compare data) is compared with searched bit (entry data) using an XOR** operation and fed into **a wide NOR gate**.
- If any bit mismatches then its XOR output will go to a "1" and the NOR output will go to a "0, indicating that there was a mismatch.
- This same compare function needs to be performed on each entry in the CAM's memory.





Content Addressable Memory (CAM)

Let us consider a 512 X 64 bit cam Architecture

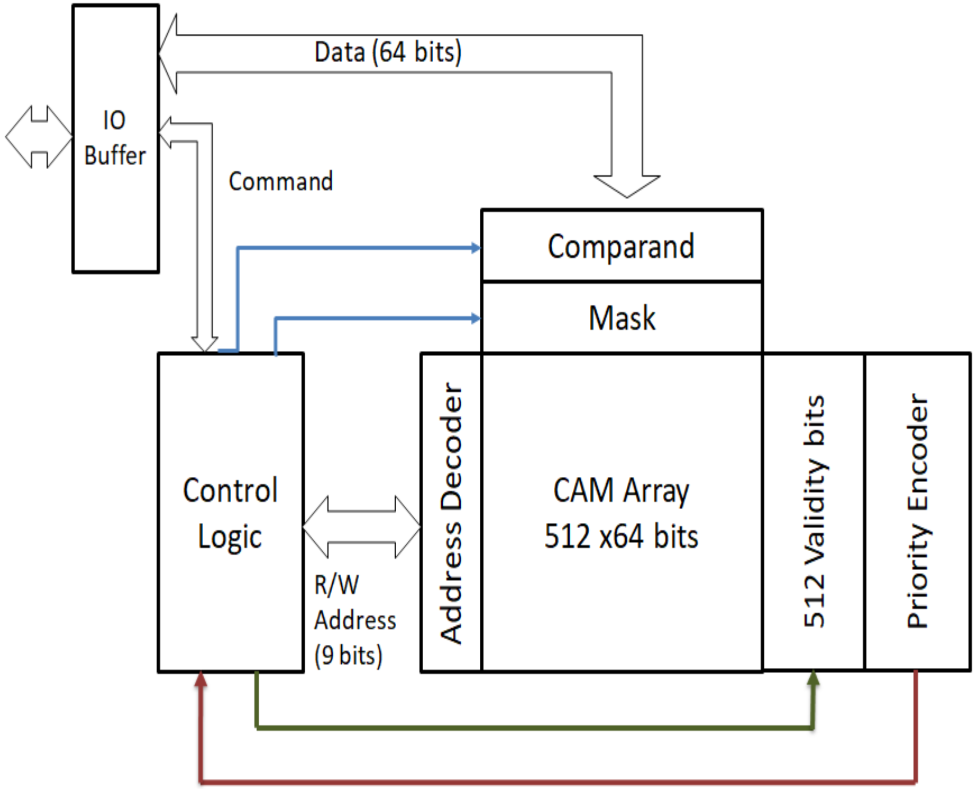


- Comparand block is filled with the data pattern to match.
- Mask word indicates which bits are significant in the word.

  Ex: To find all the words in the CAM Array that have pattern 0x123 in most significant bits, we should load Comparand with 0x1230 0000 and MASK with 0xFFF0 0000.
- All 512 rows of the CAM Array the simultaneously compares 12 MSBs of compared with the data contained in row.

Content Addressable Memory (CAM)

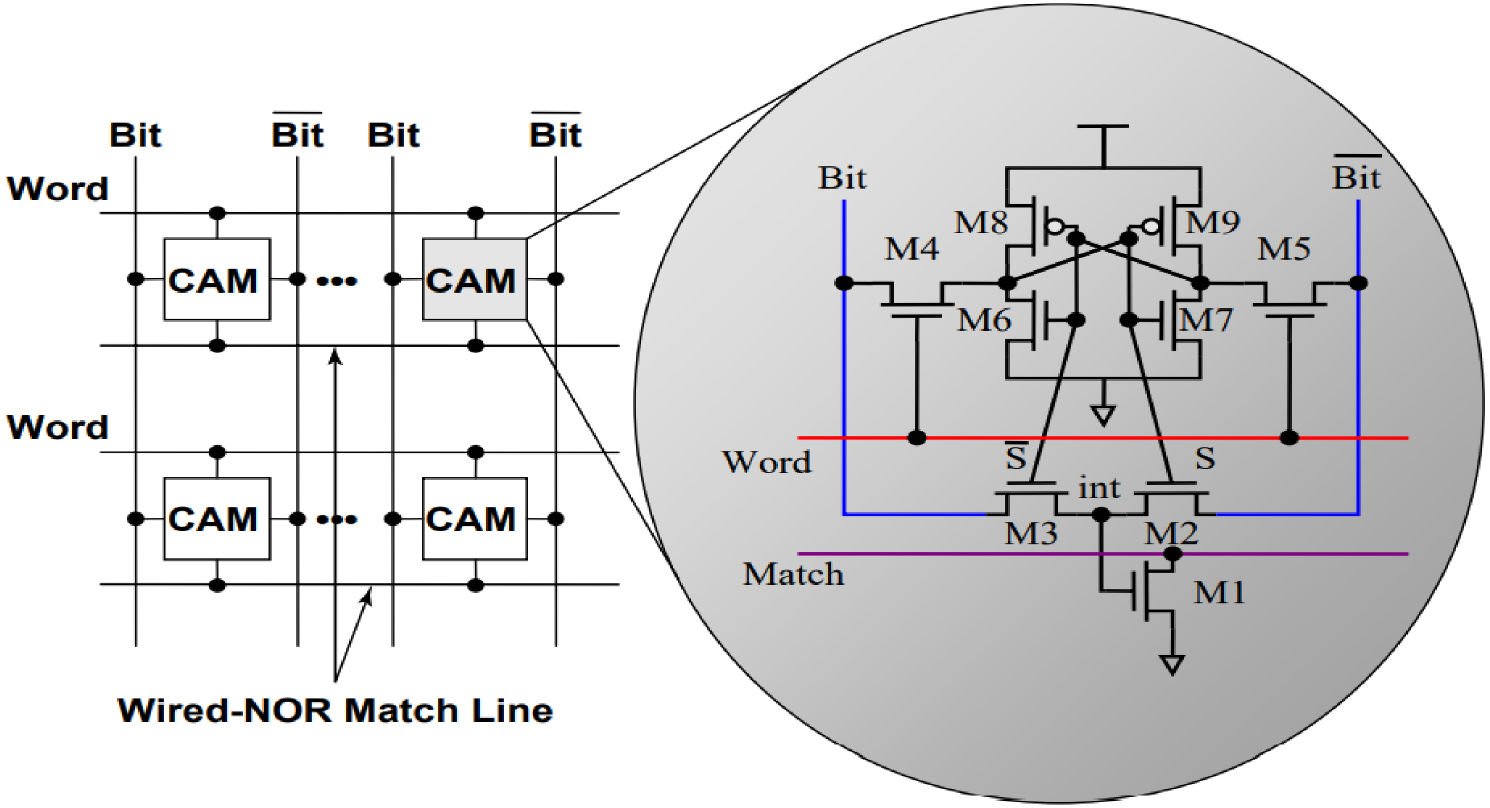
Let us consider a 512 X 64 bit cam Architecture



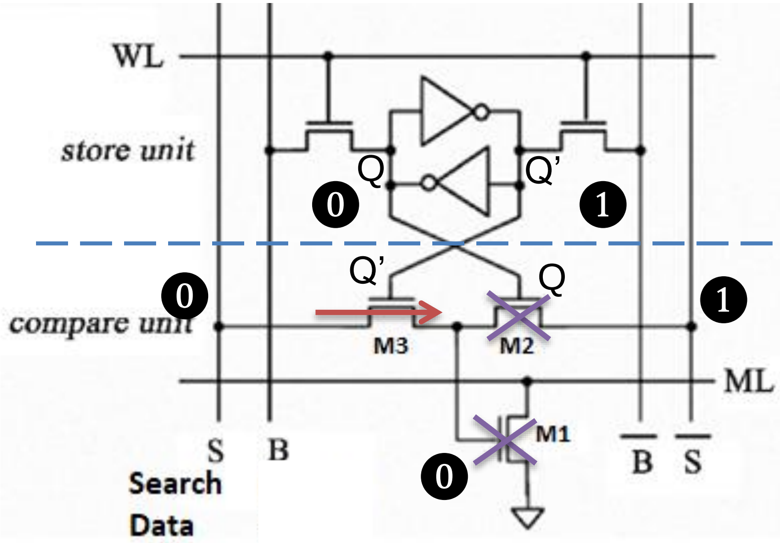
- Every ROW that matches the pattern is passed to VALIDITY block.
- The Valid rows that MATCH are passed to Priority Encoder to select one of the MATCH i.e., one with the Highest Address and encodes it in binary.
- Since 512 ROWs are present in the Array, 9 bits are required to indicate the row that matched.
- One additional Match found bit provided to indicate NO Match (If None of the ROWs matches the Pattern)

Binary Content Addressable Memory (BCAM)

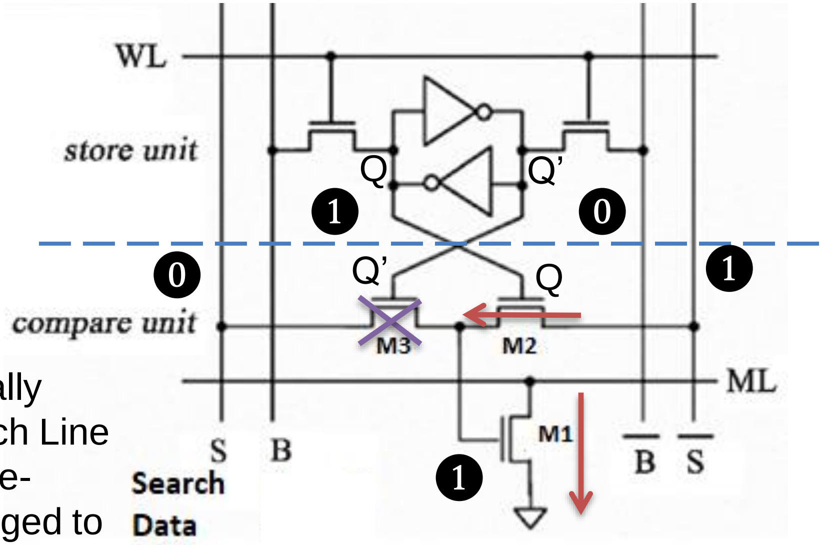
Content Addressable Memories (CAM) search for the Memory address / addresses where the referred data is present.
The Cell combines traditional 6T SRAM Cell with additional circuitry to perform 1 bit comparison using (M1-M3)



Binary Content Addressable Memory (BCAM) – 9T BCAM



Initially Match Line is pre-charged to Vdd



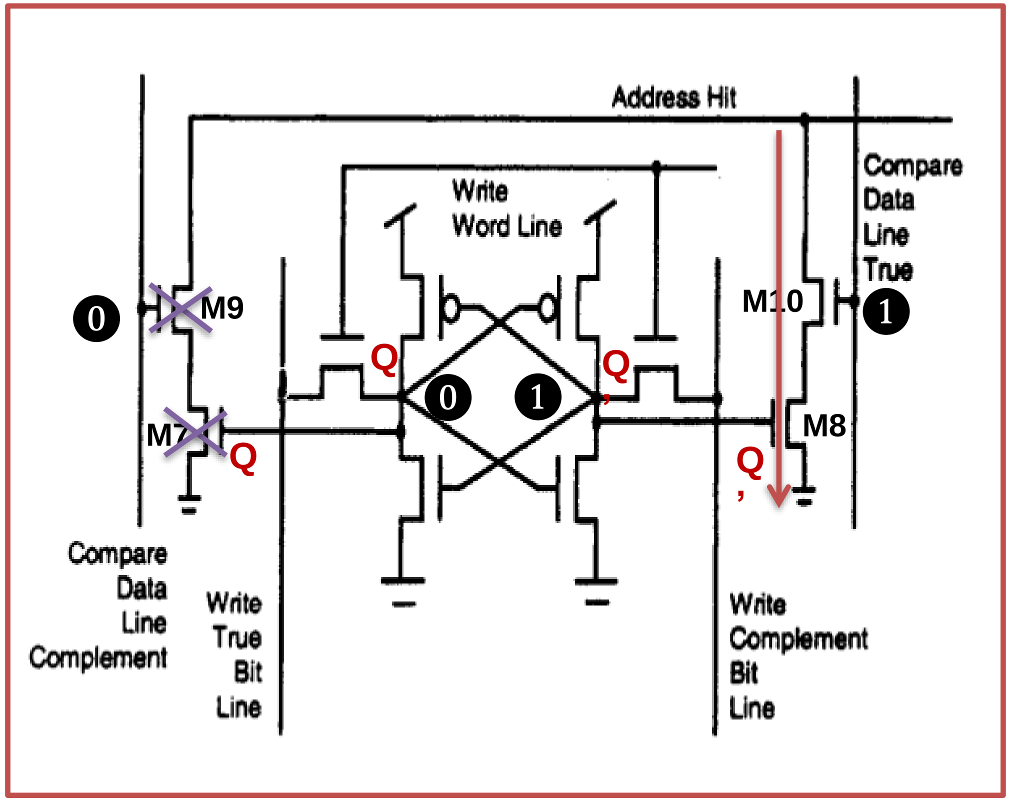
Match : Search Data =0 and Stored data(Q) =0

Miss-Match : Search Data =0 and Stored data(Q) =1

Search Data	Stored Data (Q)	M3 Pass Transistor	M2 Pass Transistor	Resultant Transistor M1	Match Line (ML)	Comparison Result XNOR
0	0	ON Passing '0'	OFF	OFF	Vdd	Match
0	1	OFF	ON Passing '1'	ON	Vss	Mismatch
1	0	ON Passing '1'	OFF	ON	Vss	Mismatch
1	1	OFF	ON Passing '0'	OFF	Vdd	Match

Binary Content Addressable Memory (BCAM)-10T BCAM

CAM cell with a built in XOR function. It is a 10 Transistor Cell in which 6 transistor make simple cross coupled latch with access transistors and other 4 does the function of XOR gate.

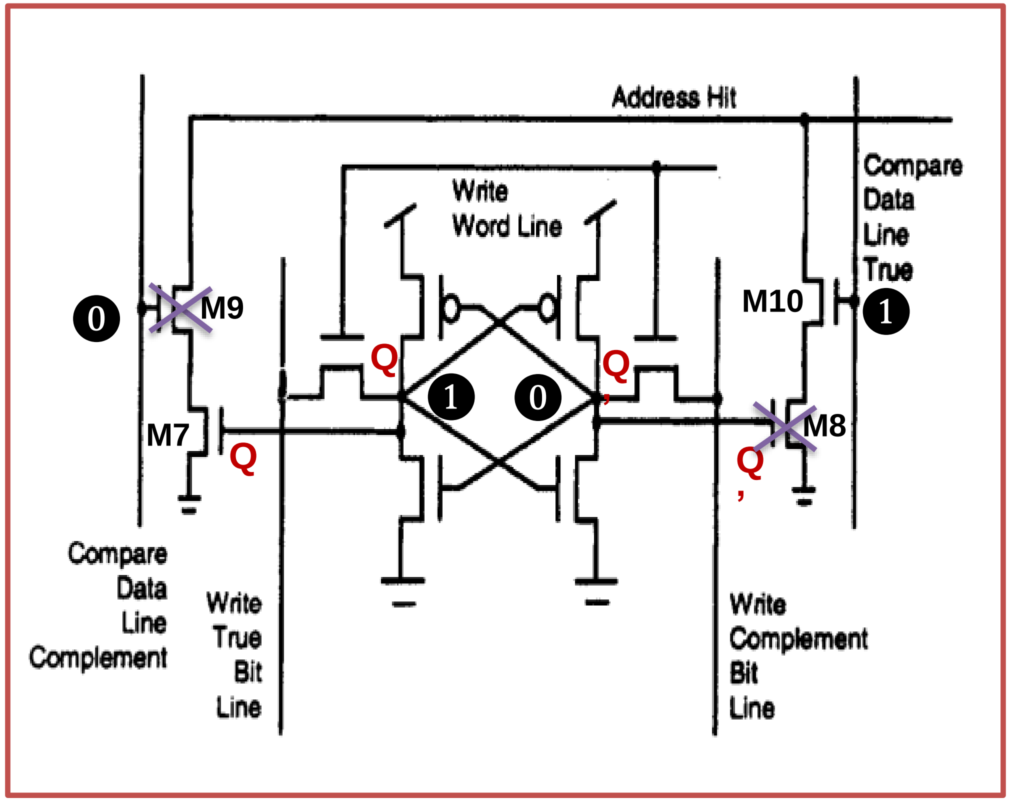


- The other four provide the XOR function.
- If there is a **mismatch** then either the true side or the complement side causes the **hit signal to be discharged.**
- If a match occurs on a cell, then **neither the true nor the complement** NFET stacks discharges the hit signal

Miss-Match : Search Data =1 and Stored data(Q) =0
M7,M9 are OFF and M8,M10 are ON providing discharging path for Hit(Match) Line i.e., (Hit=LOW)

Binary Content Addressable Memory (BCAM)- 10T BCAM

CAM cell with a built in XOR function. It is a 10 Transistor Cell in which 6 transistor make simple cross coupled latch with access transistors and other 4 does the function of XOR gate.



- The other four provide the XOR function.
- If there is a **mismatch** then either the true side or the complement side causes the **hit signal to be discharged.**
- If a match occurs on a cell, then **neither the true nor the complement** NFET stacks discharges the hit signal

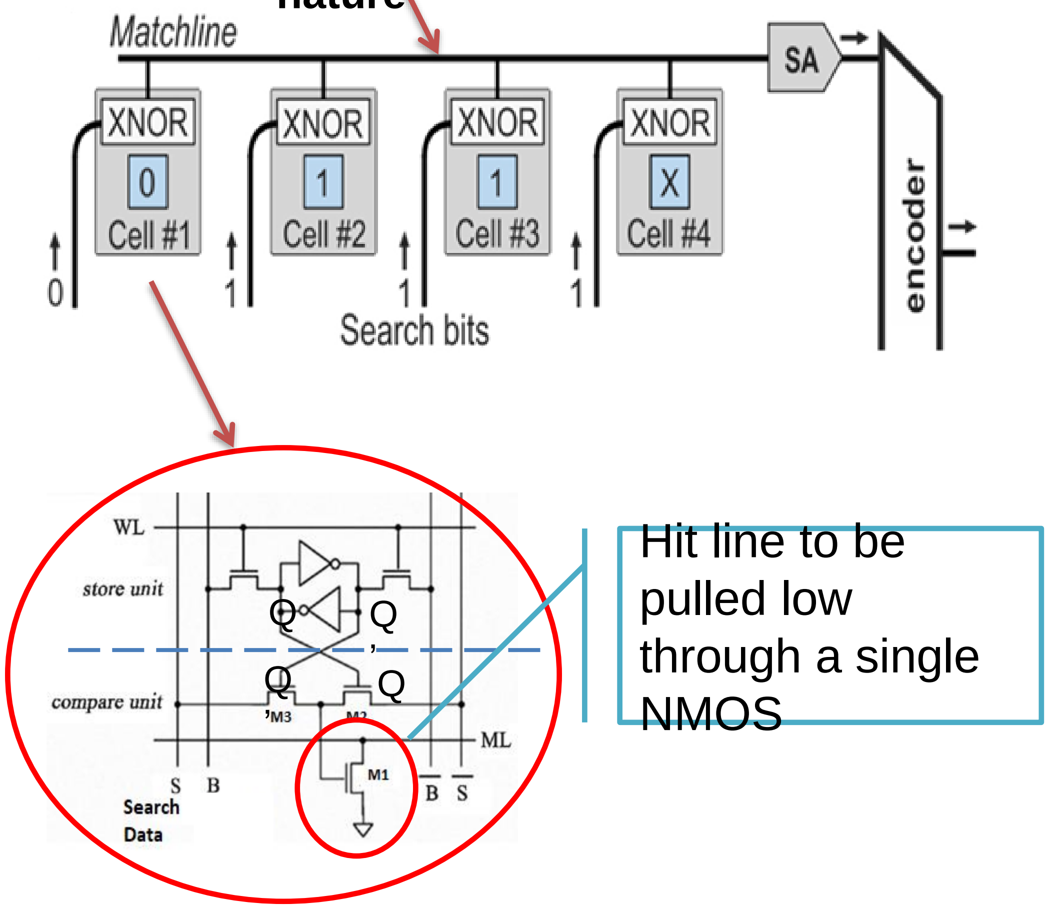
Match : Search Data =1 and Stored data(Q) =1
M7 is ON ,M9 is OFF and M8 is OFF , M10 ON . No path for discharging of Hit/MATCH line i.e.,(HIT=HIGH)

Binary Content Addressable Memory (BCAM)

Design Issue to be addressed

- At the beginning of each cycle the CAM entry address hit signal is pre-charged high.
- Any mismatch causes the hit signal to be pulled low .
- The most **difficult mismatch to detect is a single bit mismatch**. The **hit signal** spreads horizontally across the compare array and **is heavily capacitive.**
- A single bit mismatch requires the hit line to be pulled low through a single NMOS stack in 9T BCAM and stack (pair) of NMOS in 10T BCAM. Therefore careful timing analysis and simulation must be performed to ensure that **worst case conditions still allow a single bit mismatch to discharge the hit signal with sufficient margin.**
- A timing chain exists in the CAM with this topology, **which clocks the hit signal into the word line of the data array.**
- This function can be accomplished simply with an AND gate. **The clock is ANDed with the hit signal to drive the appropriate word line high** in order to read the data array.

Heavily Capacitive in nature



Hit line to be pulled low through a single NMOS

A Ternary Content Addressable Memory (TCAM)

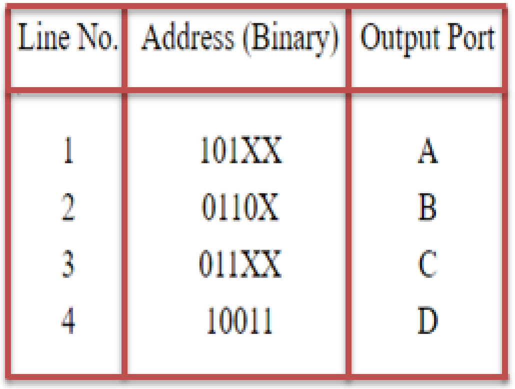
- In certain applications there are bits which are “don’t cares” in either the compare data that is stored in the array or that is applied for compare at the CAM’s inputs.
- These compare inputs can be masked with a masking bit on a per bit basis.
- In this case, each time a mask bit is set the number of bits that must match is reduced by one.
- If all of the mask bits are set then all of the entries will indicate a match, which is obviously a useless condition but does illustrate the circuit operation.

CAM Router Address Lookup

- Internet routers forward **data packets from an incoming port using an address lookup function**.
- The address lookup function examines the packet's destination address and chooses an output port associated with that address.

Example:

- There are four entries in the table represented by 5-bit words, with the *don't care* bit, *X*, matching both a 0 and a 1 in that position.
- Because of the *X* bits, the first three entries in Table represent a range ofinputaddresses,
  i.e. Entry in Line 1(101xx) indicates that **all addresses in the range of 10100_2—10111_2 are forwarded to port *A*.**

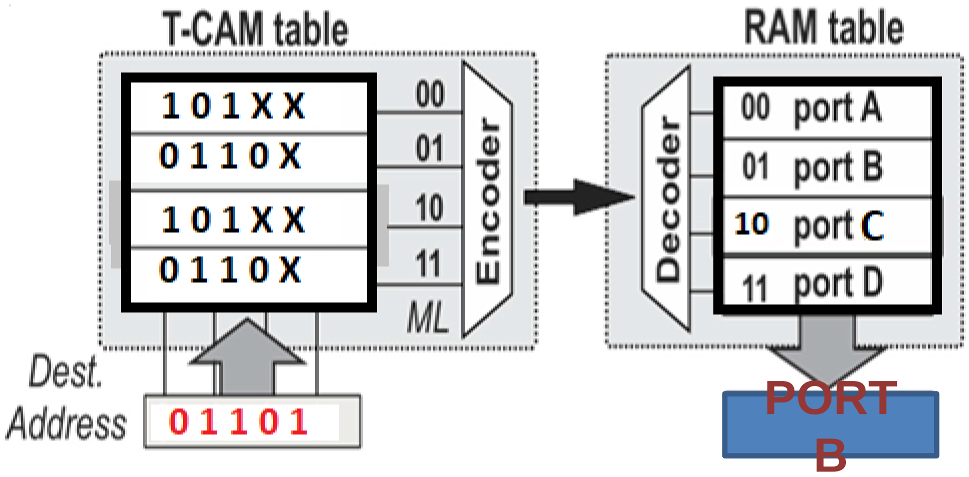


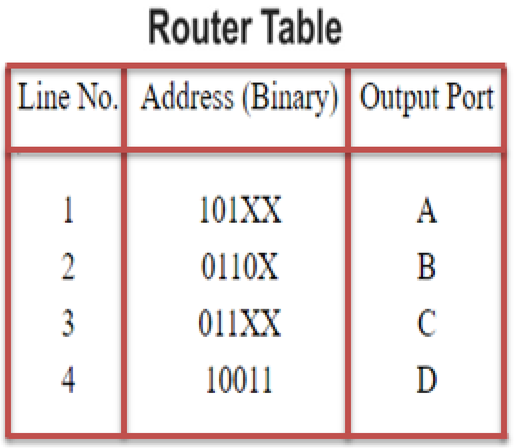
Line No.	Address (Binary)	Output Port
1	101XX	A
2	0110X	B
3	011XX	C
4	10011	D

How it works ?

- The router searches for the destination address of each incoming packet in the address lookup table to find the appropriate output port.
- For example, if the router receives a packet **with the incoming address 01101, the address lookup matches both Line 2 and Line 3 in the table**.
- **Line 2 is selected since it has the most defined bits,** indicating it is the most direct route to the destination.

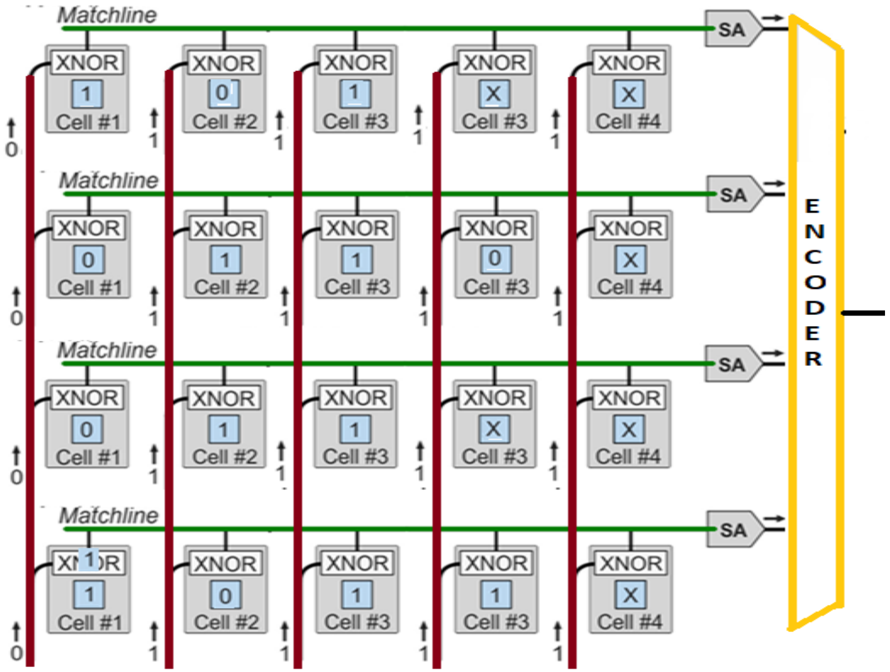
Content Addressable Memory (CAM)



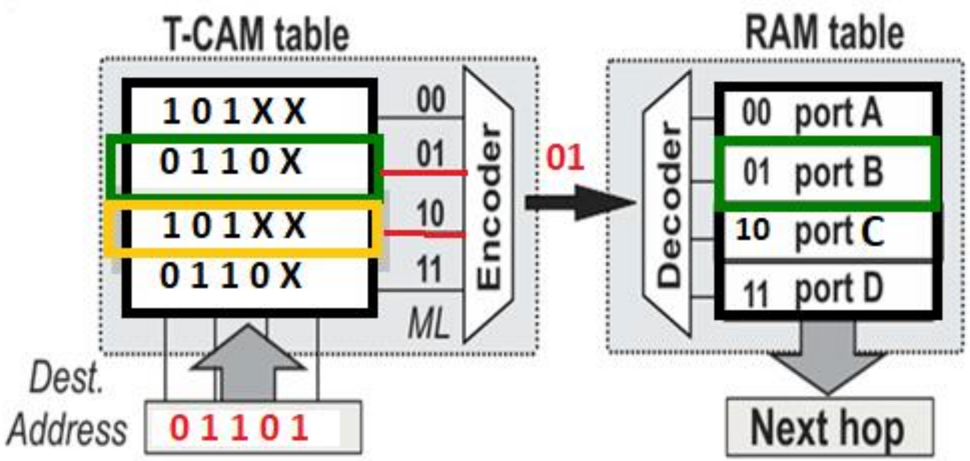


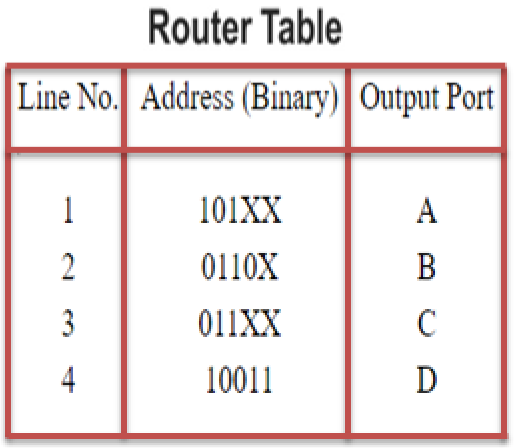
Router Table

Line No.	Address (Binary)	Output Port
1	101XX	A
2	0110X	B
3	011XX	C
4	10011	D



Content Addressable Memory (CAM)



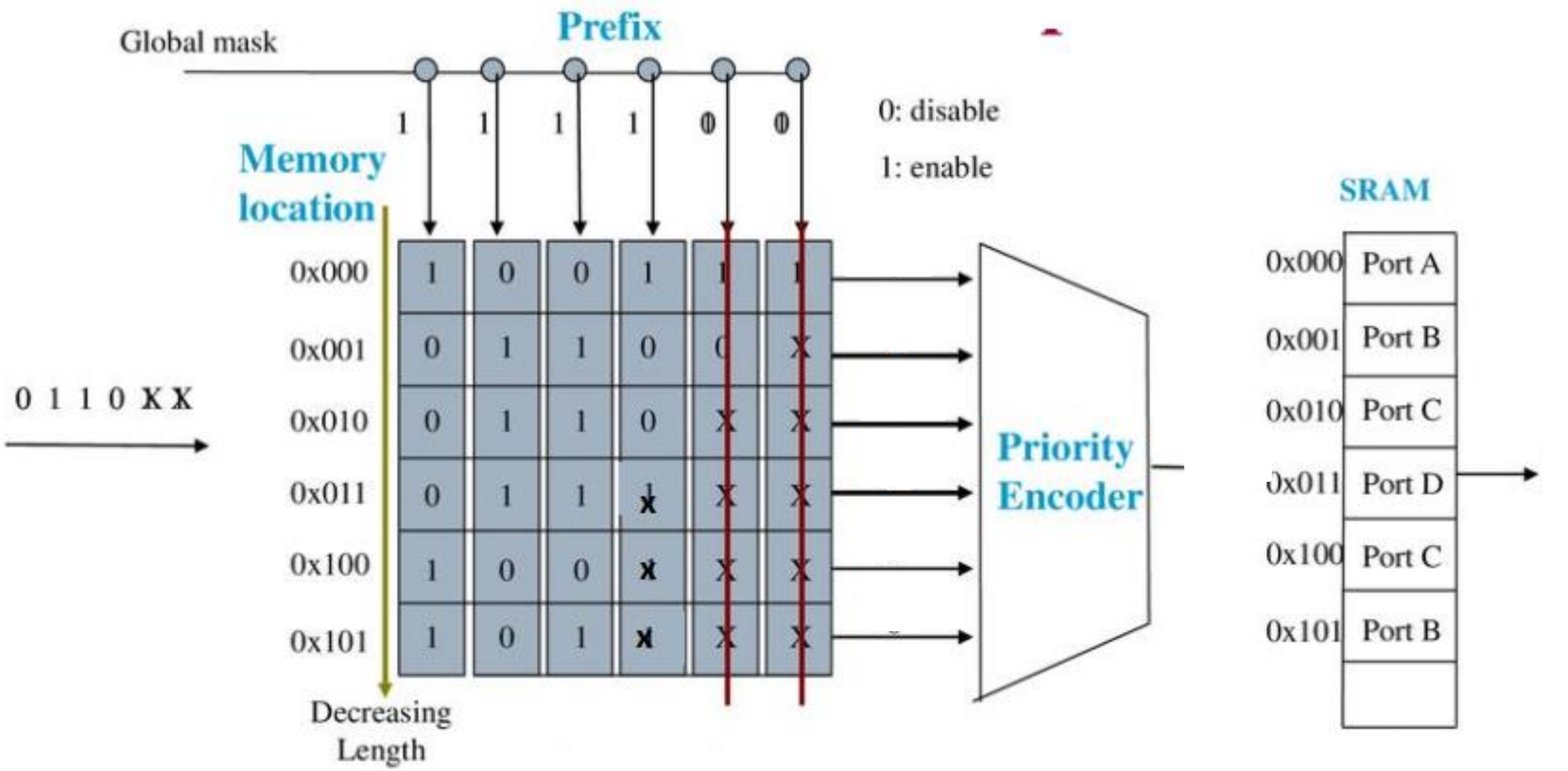


Router Table

Line No.	Address (Binary)	Output Port
1	101XX	A
2	0110X	B
3	011XX	C
4	10011	D

Content Addressable Memory (CAM)

Identify the destination address searched by the router searches for each incoming packet given a) 0110xx b) 100xxx

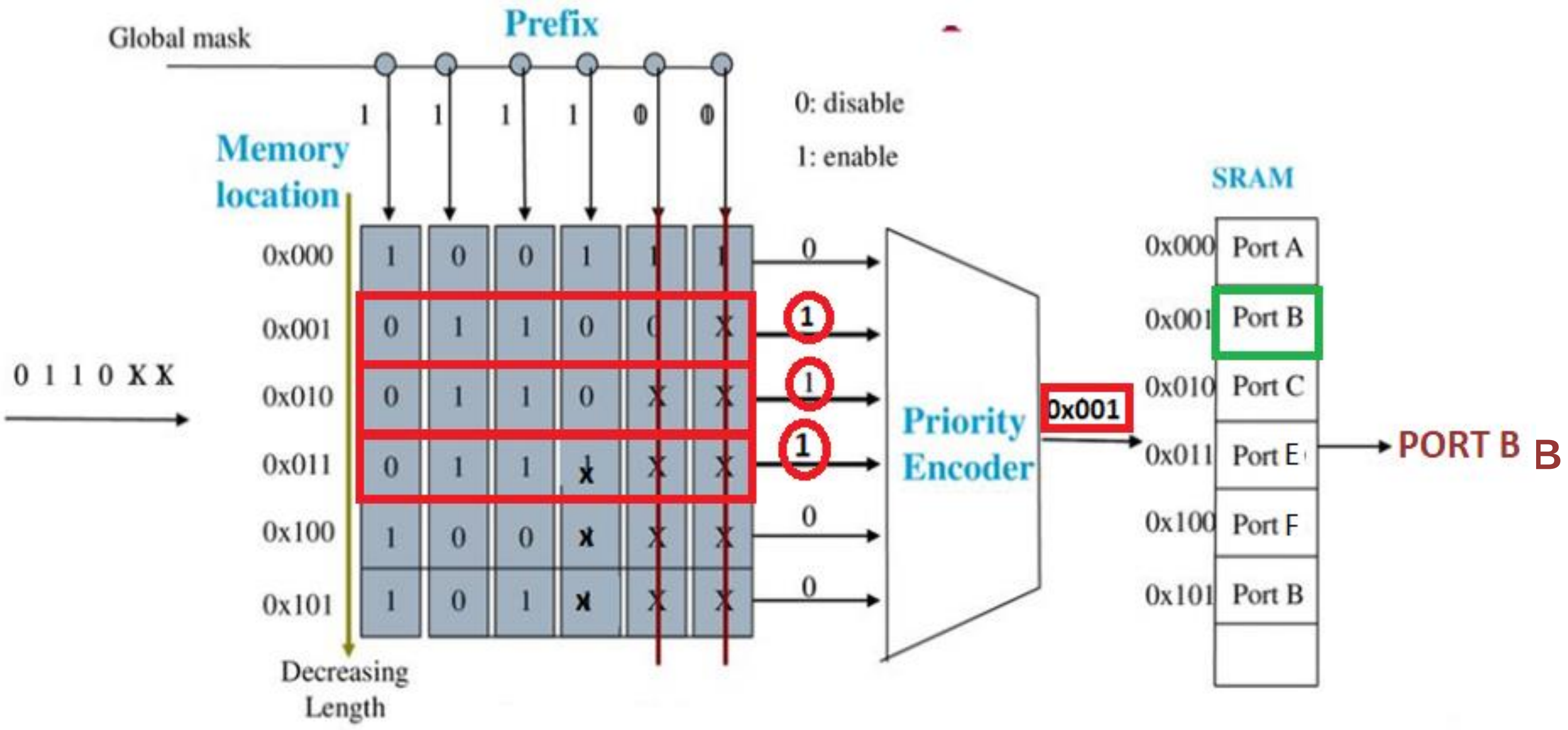


Memory location						
0x000	1	0	0	1	1	1
0x001	0	1	1	0	0	X
0x010	0	1	1	0	X	X
0x011	0	1	1	X	X	X
0x100	1	0	0	X	X	X
0x101	1	0	1	X	X	X

SRAM	
0x000	Port A
0x001	Port B
0x010	Port C
0x011	Port D
0x100	Port C
0x101	Port B

Content Addressable Memory (CAM)

Solution: a) 0110xx

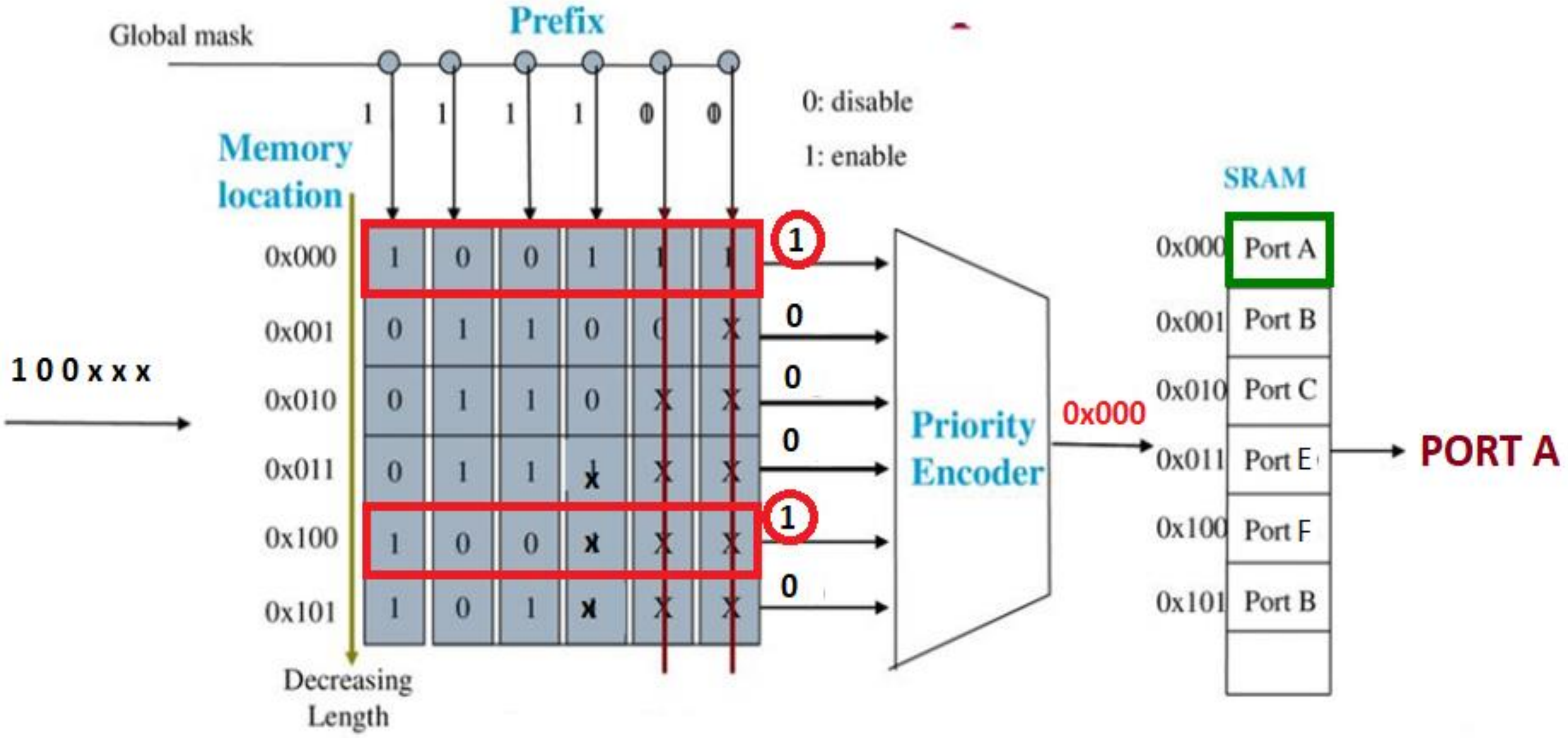


Memory location						
0x000	1	0	0	1	1	1
0x001	0	1	1	0	0	X
0x010	0	1	1	0	X	X
0x011	0	1	1	X	X	X
0x100	1	0	0	X	X	X
0x101	1	0	1	X	X	X

SRAM	
0x000	Port A
0x001	Port B
0x010	Port C
0x011	Port E
0x100	Port F
0x101	Port B

Content Addressable Memory (CAM)

Solution: b) 100xxx

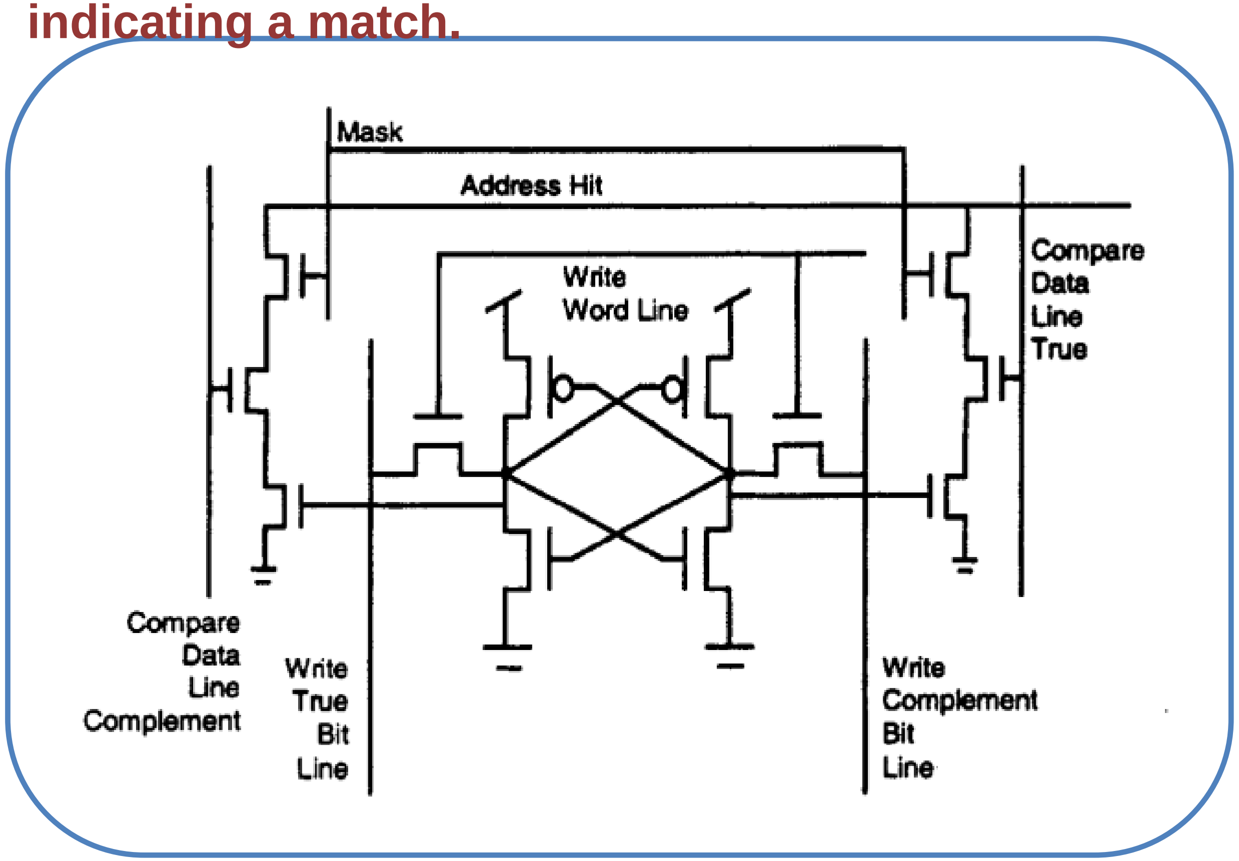


Memory location						
0x000	1	0	0	1	1	1
0x001	0	1	1	0	0	X
0x010	0	1	1	0	X	X
0x011	0	1	1	X	X	X
0x100	1	0	0	X	X	X
0x101	1	0	1	X	X	X

SRAM	
0x000	Port A
0x001	Port B
0x010	Port C
0x011	Port E
0x100	Port F
0x101	Port B

A Ternary Content Addressable Memory (TCAM)

In this case an additional NMOS is added to the stack as shown with control signal as Mask . **If a masking bit is set, the inputs to the corresponding NFETs are both driven low**. i.e., turns OFF these NMOS in stack, This prevents that cell from discharging the hit line and thus **indicating a match.**



- A ternary CAM, or TCAM, allows individual bits in entries to be treated as "don't care" bits.
- A ternary CAM has three states per bit: they are "0, "l", and X.
- If an X is encountered then either a "1" or a "0” will be considered to be a match on that bit

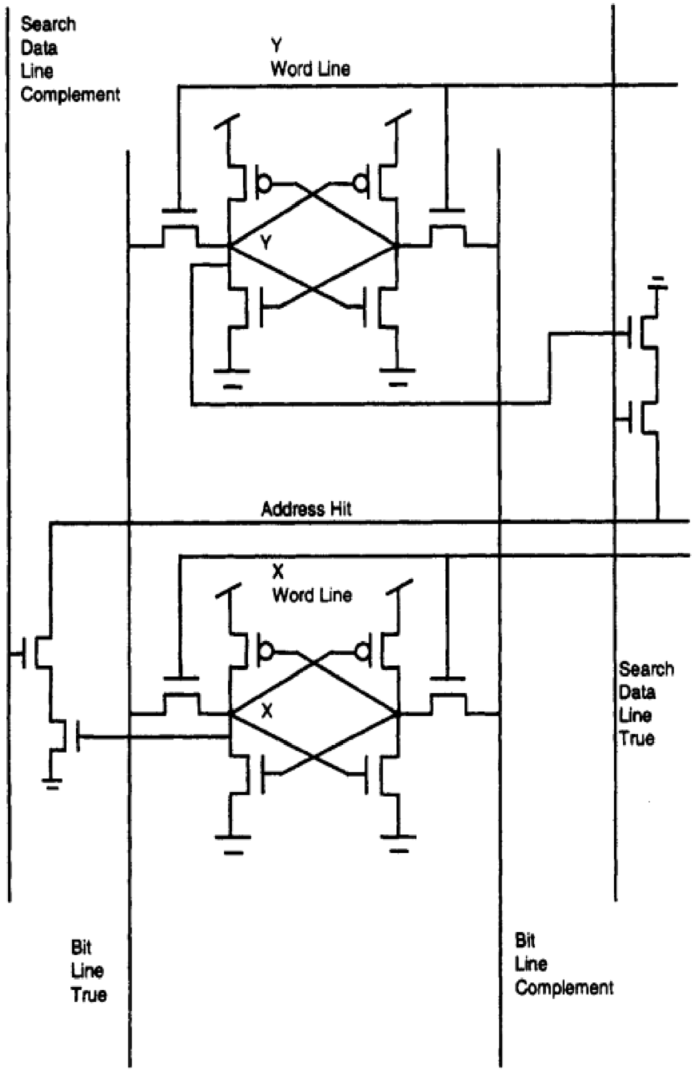
A Ternary Content Addressable Memory (TCAM)

Since an SRAM cell, the core of a CAM, stores only two states, i.e., "1" and "0", more than one cell is required to store the three states.

A ternary CAM implemented with two static cells. Two cells normally have four valid possible states but in the case of a TCAM, only three of the states are considered valid.

For a ternary CAM, the two bit cells are referred to as cell X and cell Y. The valid encoded states of these two cells are listed below

X	Y	Data	Hit	Comment
0	0	0	1	MATCH, Stored Data is X
0	0	1	1	
0	1	0	1	MATCH, Stored Data is 0
0	1	1	0	MISMATCH
1	0	0	0	MISMATCH
1	0	1	1	MATCH, Stored Data is 1
1	1			Invalid



FinFET Basics

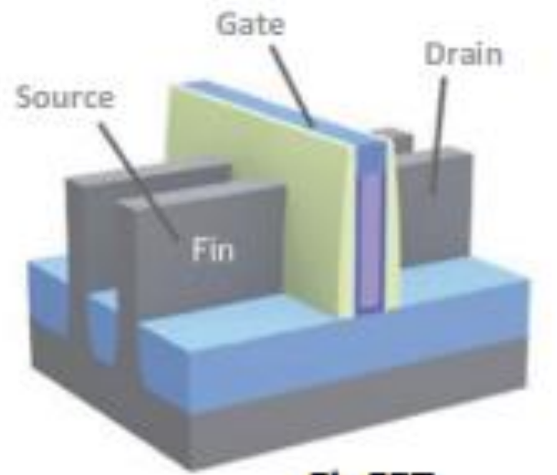
FinFETs are known as non planar technology

FinFETs usually too have multiple fins

Fins conduct current

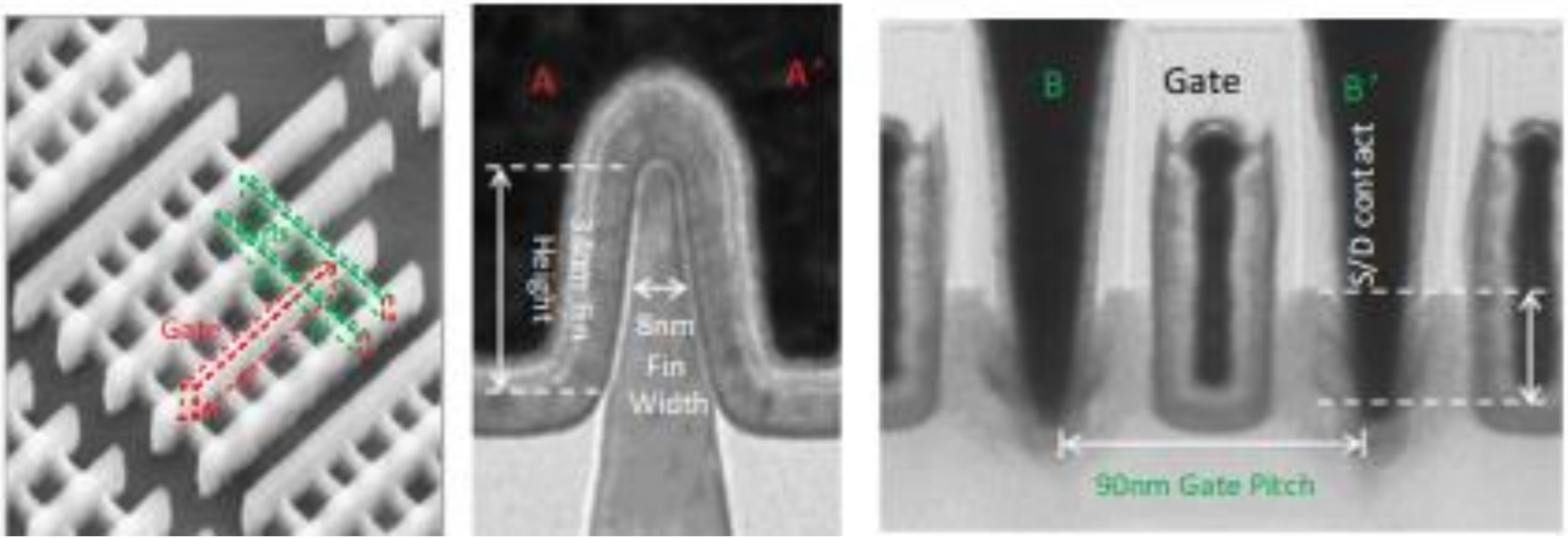
Fins connected to bulk of the transistor

It is also called as tri gate structure as gate covers three sides of fins



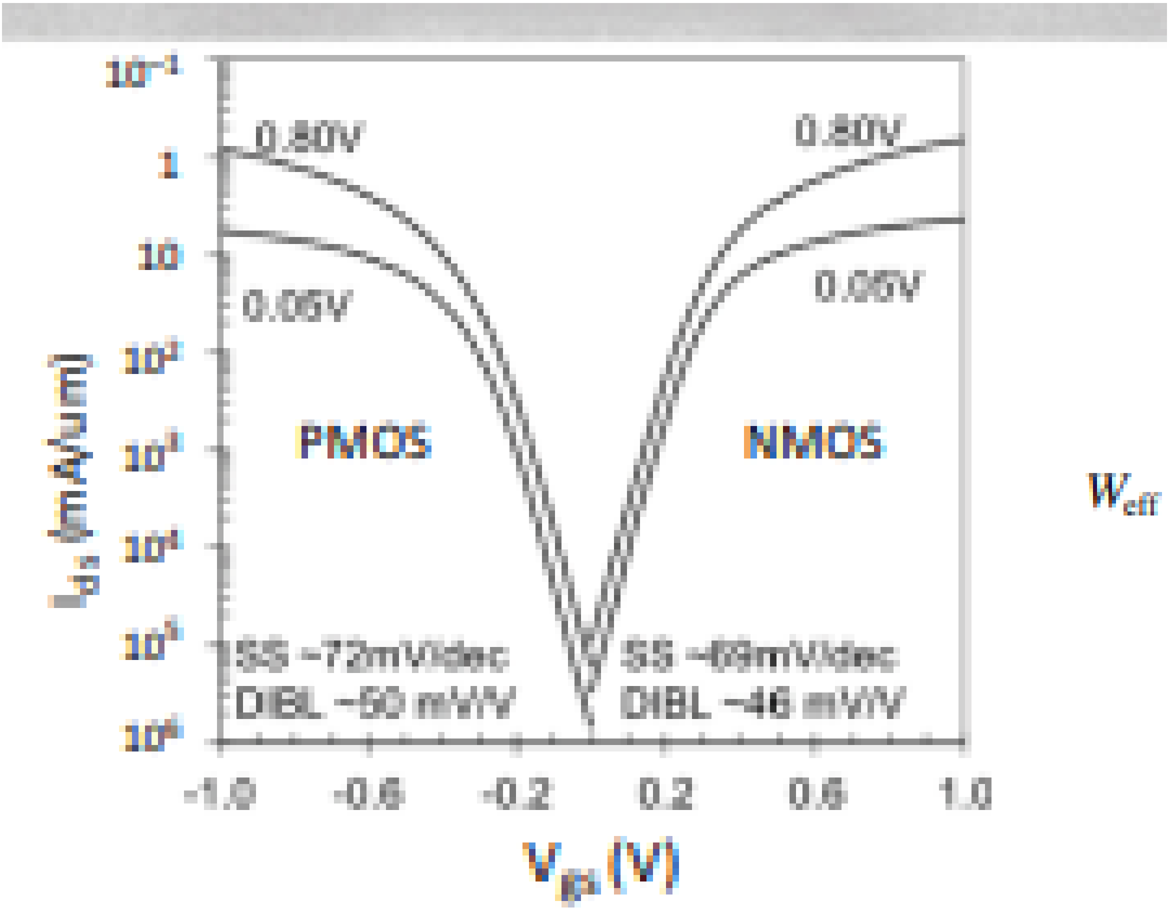
FinFET

FinFET Basics



Microscopic image of Intel’s 22nm FinFET

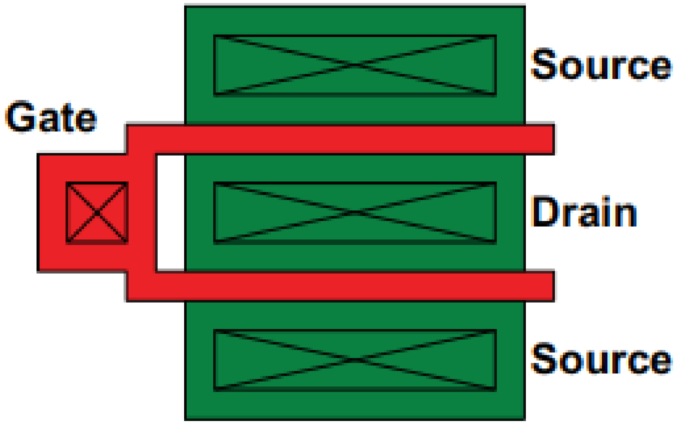
FinFET Basics



$$W_{eff} = 2 \times Fin_{\_height} + Fin_{\_width}$$

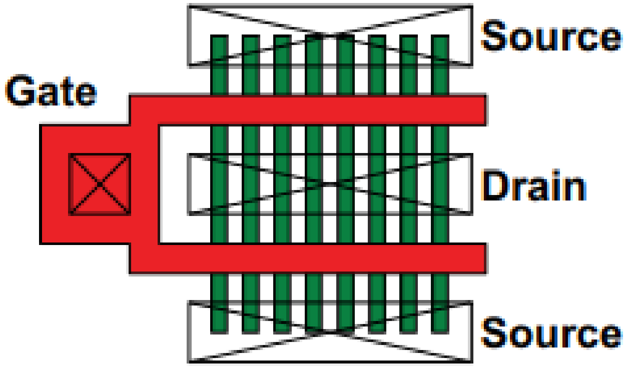
V-I characteristics

FinFET Basics

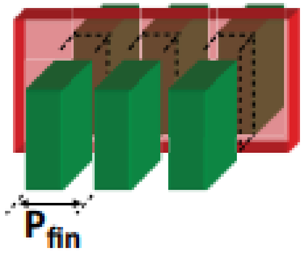


Planar MOSFET

$W_{eff} = (2 * H_{fin} + T_{fin}) * N_{Fins}$



FinFET



$W_{Phy} = N * P_{fin}$

FinFET Basics

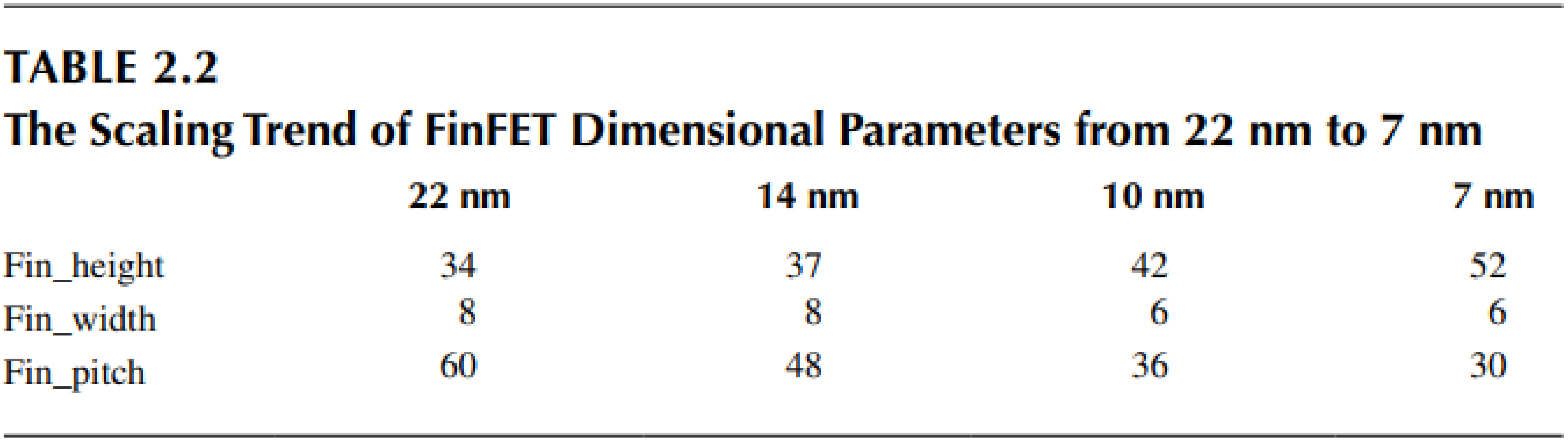


TABLE 2.2
The Scaling Trend of FinFET Dimensional Parameters from 22 nm to 7 nm

	22 nm	14 nm	10 nm	7 nm
Fin_height	34	37	42	52
Fin_width	8	8	6	6
Fin_pitch	60	48	36	30

THANK YOU

Mahesh Awati/Dr Shashidhar

Department of Electronics and Communication

stantry@pes.edu

+91 9845695028