Tanis Ahamed

16/01/24

DRAM

↳ When power is on, data can still decay.

① CPU → DRAM

NVM

③

② CPU → Cache → L2 → L3 → DRAM

LN → L3-N

distance means delay

L3 → DRAM

chip multiprocessor

④ CPU → Cache → Accelerator, Accelerator → DRAM

There accelators will produces high bandwidth

CIM is divided into three classes.

* # CIM → NMP (Near memory Processing)
→ PIM (Processing in memory)
→ ImC (In memory computing)

Requirements for CIM:

(i) design & compute capable memory & controller.

(ii) design & processor chip with in-memory unit.

(iii) design & system & hardware interfaces.

(iv) Fixing & system softwares & Languages.

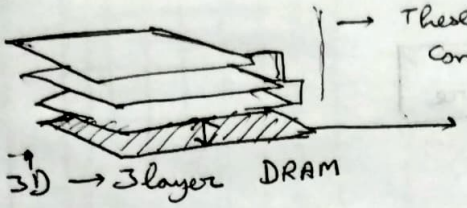
(v) Introduce memory within the CPU.

(vi) developing suitable algorith for CIM

(vii) Develop a new memory architecture.

▷ NMP :

2.5D / 3D	DRAM



→ These layer are connected via TVS

another layer which have a processing core

3D → 3layer DRAM

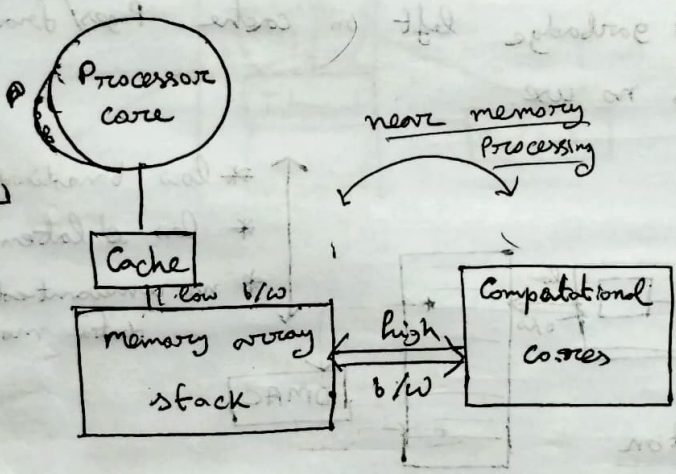
This is part of (PIM)

→ NVM : The computation will be lost as memory will be gone if it is a volatile memory.

so, IMC is has two sides → using DRAM
→ using storage

≠

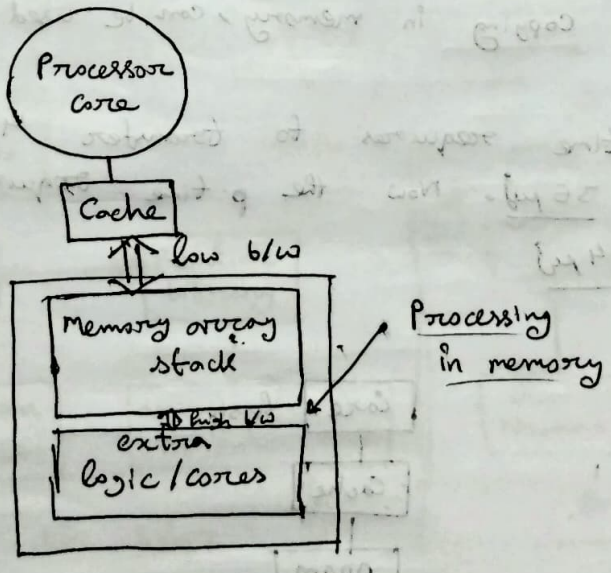
NMP



b/w → bandwidth

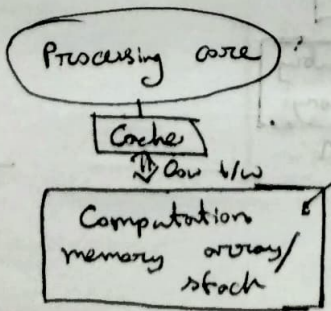
≠

PIM



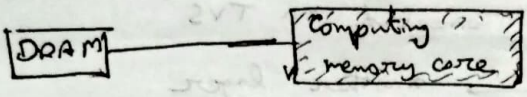
≠

IMC

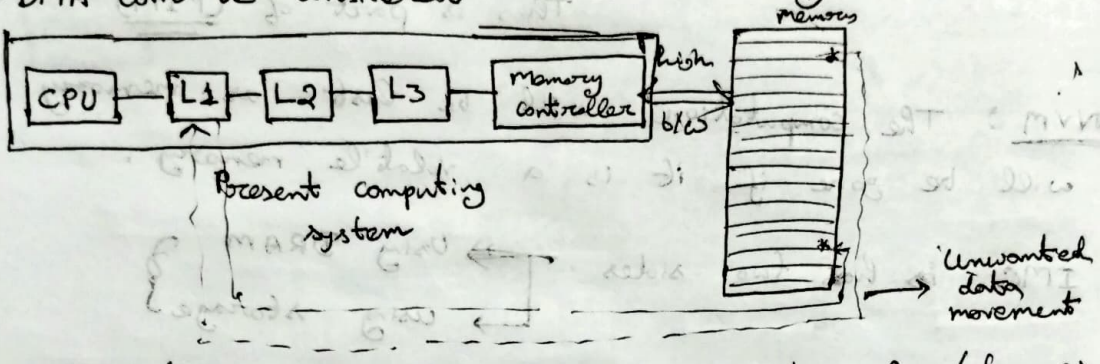


In memory computing

CIM

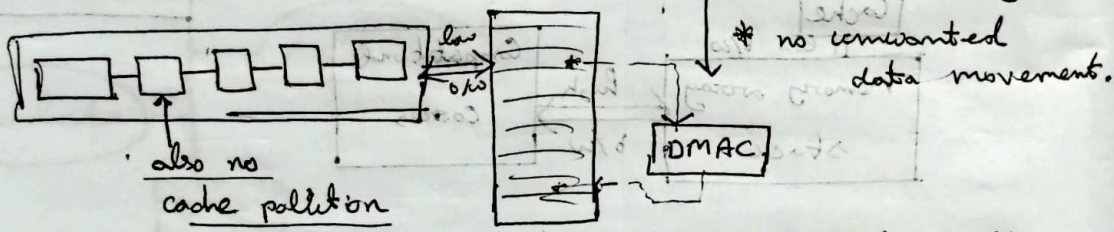


▷ DMA can be considered as near memory processing.



□ Cache polution → * garbadge left on cache. Pages/ frames & no use.

now for DMA,



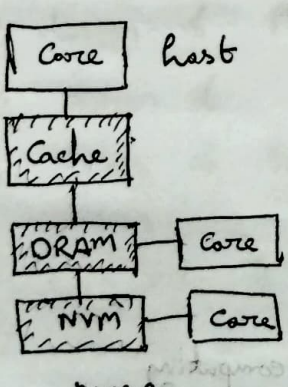
* low bandwidth
* low latency
* no unwanted data movement.

* Row clone → Copying in memory, can be used to bulk copy.

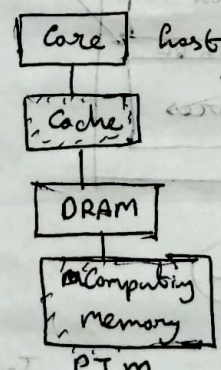
≠

→ 1046 ns time requires to transfer 4KB & power consumption was 36 μJ. Now the p. time required is 90 ns and power 0.04 μJ

17/01/24

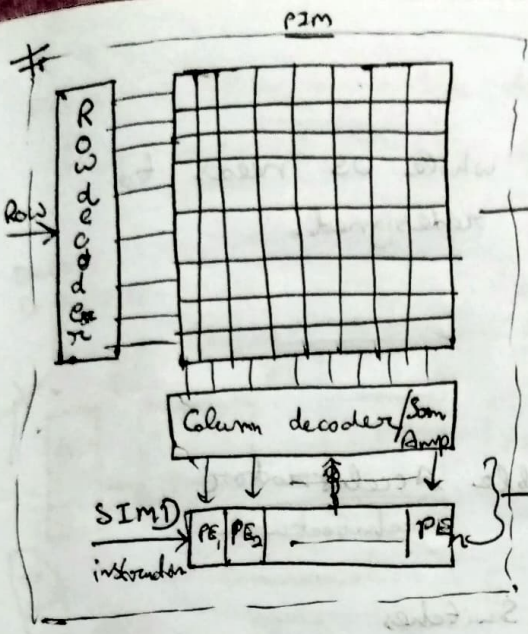


NMP



PIM

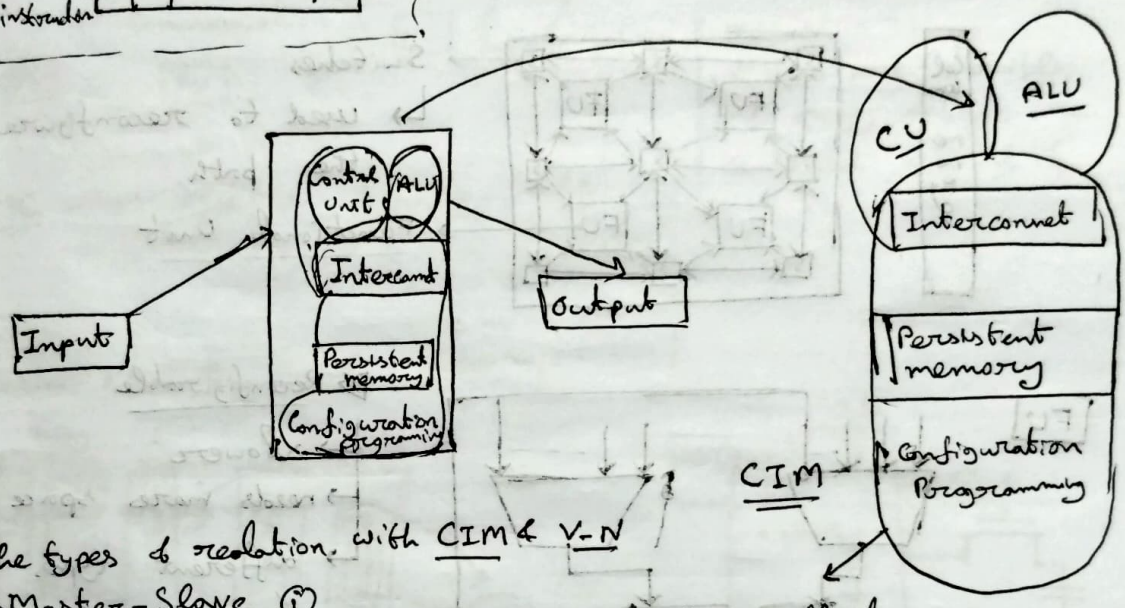
→



{ Error correction scheme

→ As this is with in a chip there's less possibility of error during data transfer.

→ Cell is not computing just this SIMD processor.

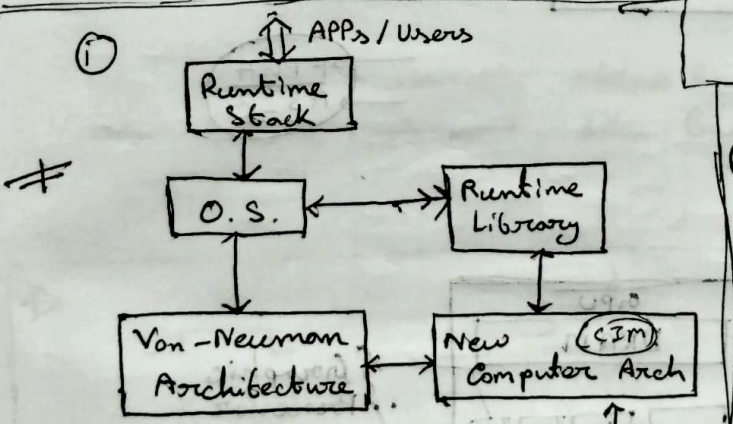


→ The types of realation with CIM & V-N

→ Master-Slave (i)
→ Co-operative (ii)
→ Independent (The picture above) (iii)

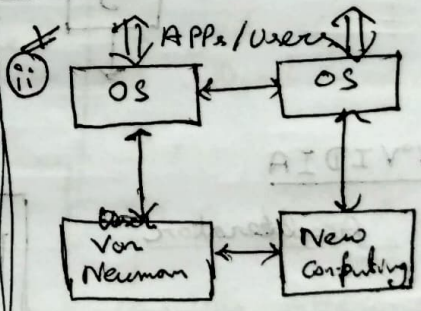
{ CIm without interaction with Von-Neuman Machine:
↳ not really feasible.

(i)



Master - Slave

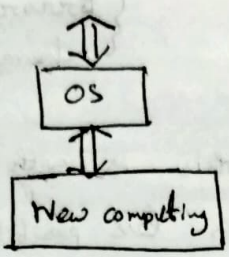
(ii)



Co-operative

As a part of Von-Neuman

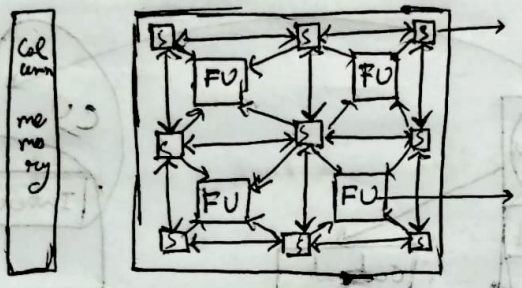
≠ 😐



→ The whole OS needs to be redesigned.

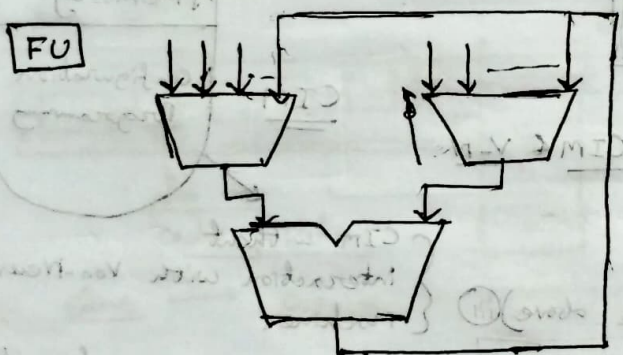
→ NMP:

CGRA → Coarse Grain Reconfigurable Accelerator Architecture

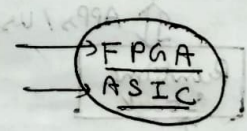


→ Switches
↳ used to reconfigure the path.

→ Functional Unit

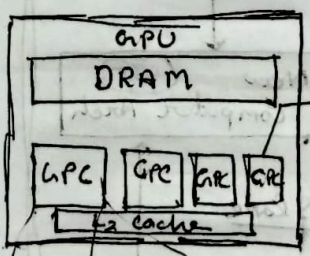


→ Reconfigurable
- → slower
- → needs more space
- → different types of tasks can be performed.

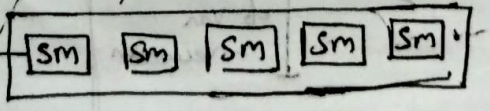


→ NVIDIA

↳ GPU Accelerator



{ Graphic Processor Cluster



{ series of multiprocessor

DM
23/01/24

FPGA → Reprogrammable designs

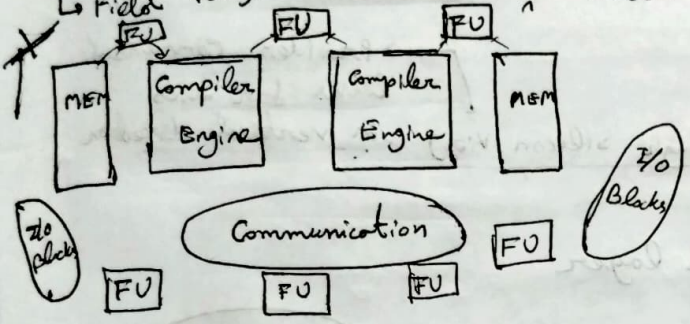
↳ Simulations are reun and re-programmed.

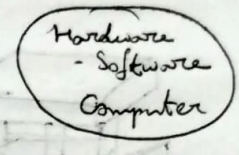
ANN → PPB me

▷ FPGA-based accelerator : Array.

Field Programmable Gate Array

Accelators need processors.

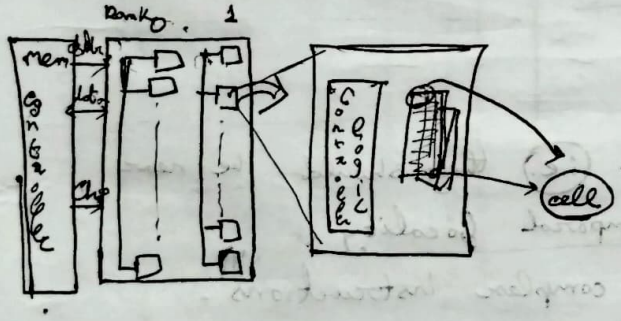


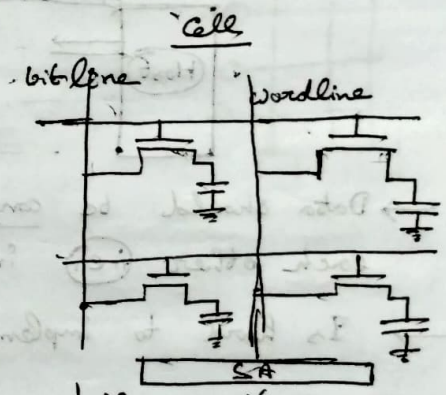


FPGA designs are flexible and many designs are centered around it.

HMC → Hybrid memory cube
(Designed from D-RAM)

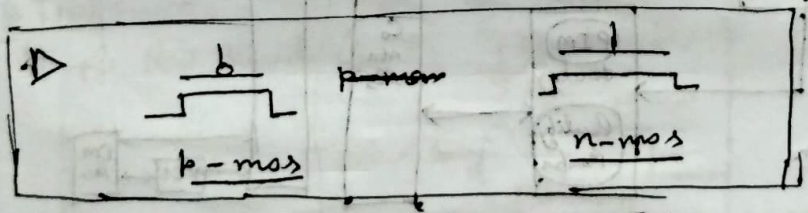
▷ Basic structure of DRAM:





The wordline is basically address line
The bitline " " " data line

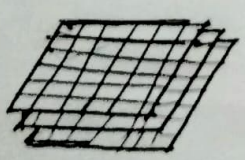
if capacitors are charged the it is '1'. Dyschanged means '0'.



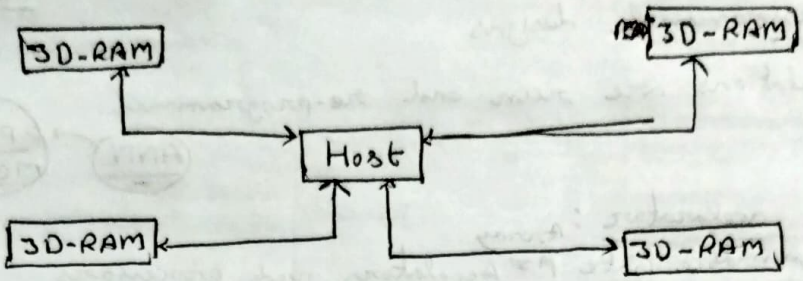
SA → Sense Amplifier
↓
considers difference b/w bitline & bitline

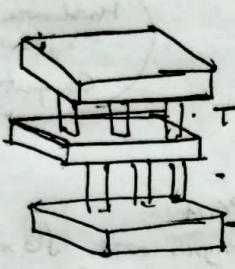
This basic structure of DRAM will not be used in our subject, we will use HMC'

▷



→





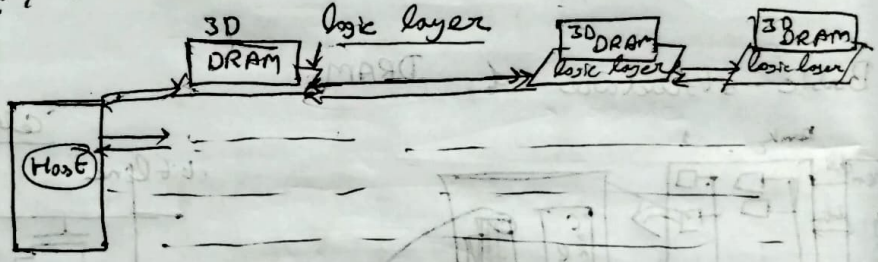
TSV (Through Silicon Via) → Provides communication links for dies in vertical direction

Logic layer

Map Reduced Workload

▷ TOP-PIM :

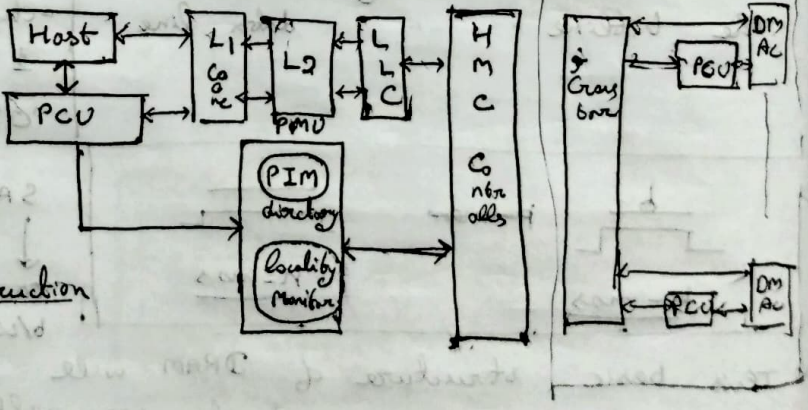
→ Daisy-chained : → Daisy chained (series)



→ Data should be congwood (i.e.) they should be near each other (i.e.) in temporal locality.

→ Is hard to implement complex instructions.

→ ISA :

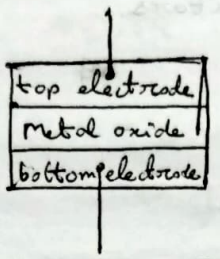


PEI Computer unit

PIM Enabled Instruction

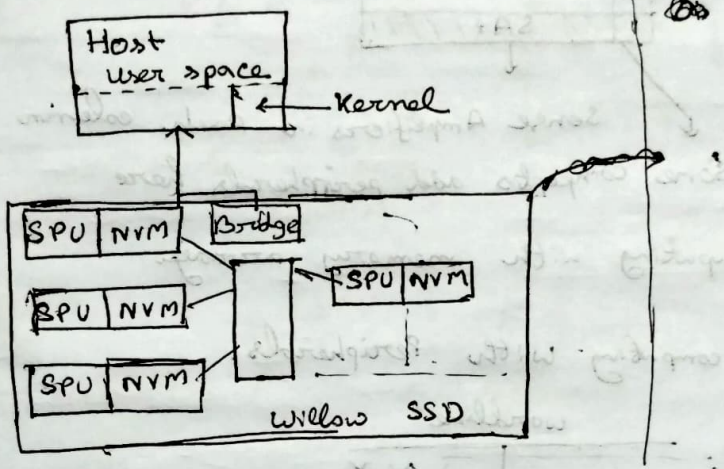
→ ~~Vault~~ :

Re-RAM → Resistive RAM



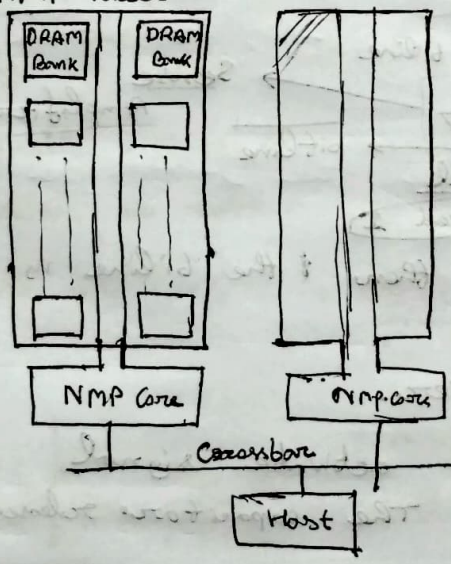
Memristor → key component in IMC

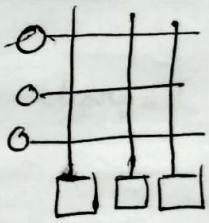
& implements



SMART MEMORY

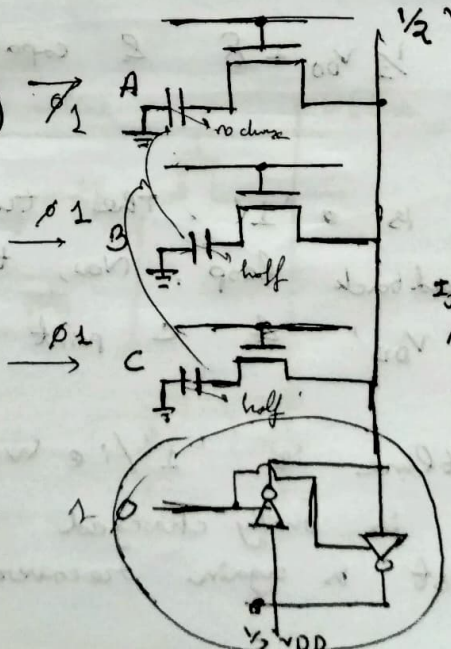
NMP Vault





Ambit

TRA :
(Triple Row Activation)



BC → ~~half~~ ¼

A → ½

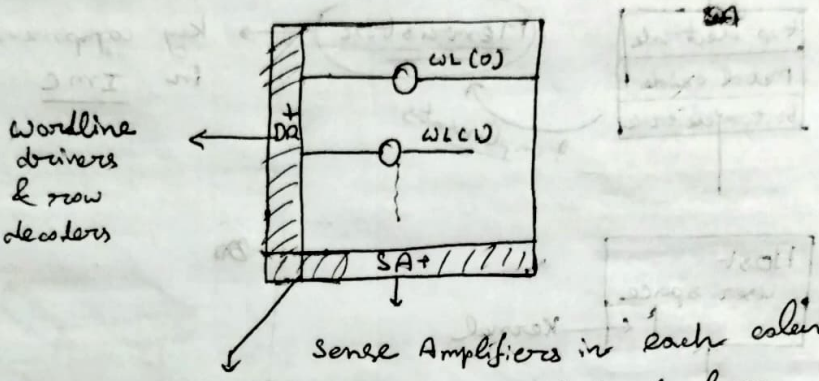
Charge ~~due~~ sharing

If we control ABC & the below input we can use different logics.

e.g. AB + BC + CA

24/01/24

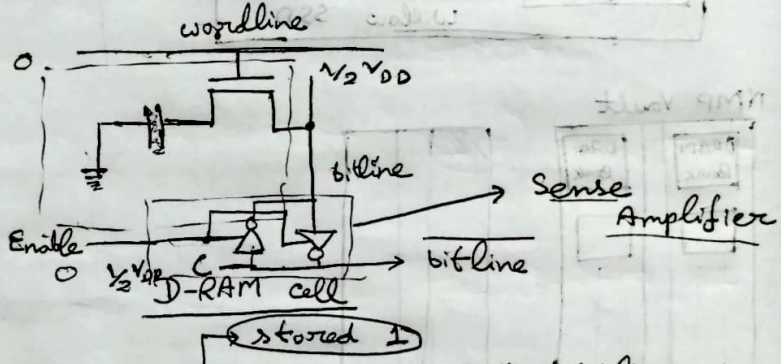
D-RAM is a series of transistors & capacitors.



Some compapts add peripherals here.

IM-A
↳ In memory computing with memory arrays.

IM-P
↳ In memory computing with Peripherals.



→ If the capacitor is charged then the bitline is ½ V_{DD}.

→ Enable controls sense amplifiers.

→ If we want to read, send activate signal, wordline will become '1'. The capacitor releases charge.

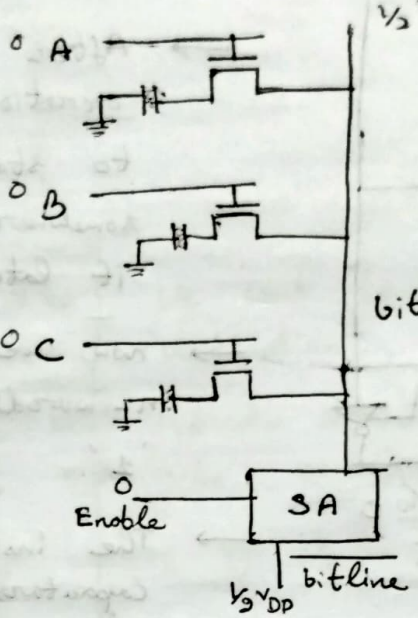
bitline charge is ½ V_{DD} + δ & capacitor is ½ V_{DD} + δ.

→ Now, the enable is '1'. The twin inverse will be in a feedback loop. Now, the bitline goes to 'V_{DD}' & C point will go to '0'.

→ Hence, the bitline is '1'/i.e. 'V_{DD}'. Also the capacitor is fully charged again. Hence the data we lost is again recovered.

TRA :

A simple AND/OR logic was implemented using three cells.



SA
Is same as last picture.

$$\delta = \frac{(2K-3)C_c}{6C_c + 2C_b} V_{DD}$$

$C_c \rightarrow$ cell capacitance

$C_b \rightarrow$ bitline capacitance

$K \rightarrow$ no of cells in a fully charged state.

→ so, if $K = 2$ or 3 then the $\delta \rightarrow$ +ve from expression then the SA will make bitline '1'

→ Enable = 0, A = B = C = 0, capacitors are charged this data stored state.

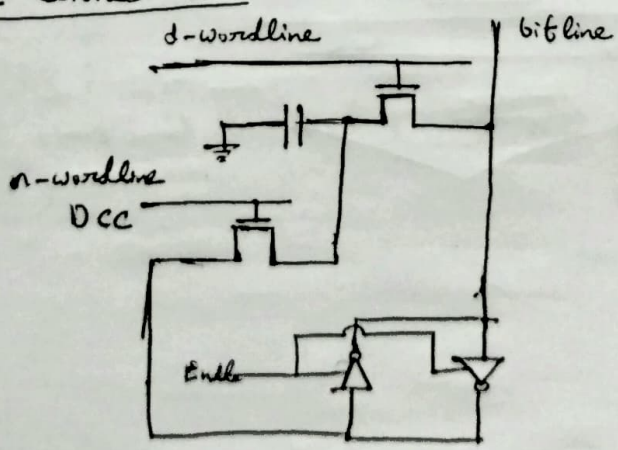
→ so, bitline = $AB + CA + BC$

so, if majority is one.

∴ if, $C = 0$, AND → AB

if, $C = 1$ OR → $A + B$

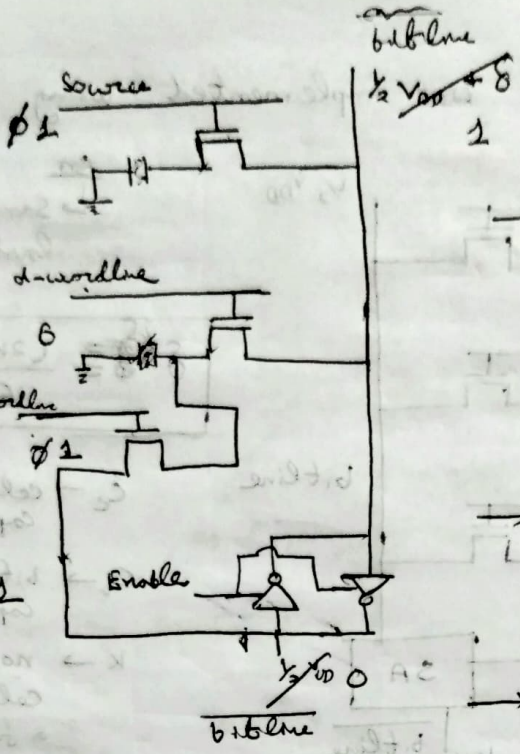
Dual-Connect Cell :



d-wordline
n-wordline

d → data
n → negation

now,



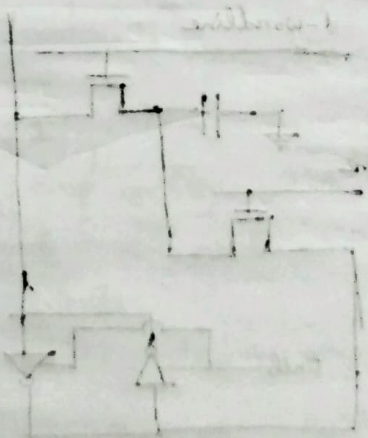
→ After normal operation, we need to store bitline somewhere to know, if later.

→ now, we make n-wordline is make '1'.

→ The inside cell capacitors will store charge.

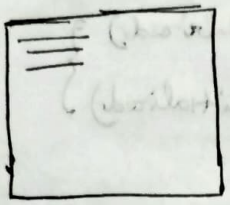
So, '0' represents no charge } (1/0) in bitline
'1' " charged }

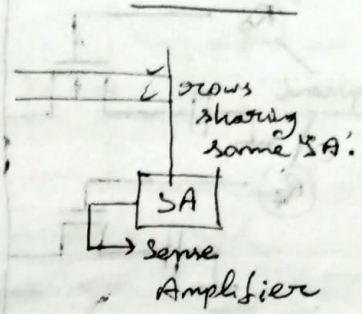
Through this, we can make the D-RAM cell compute simple logic operation.

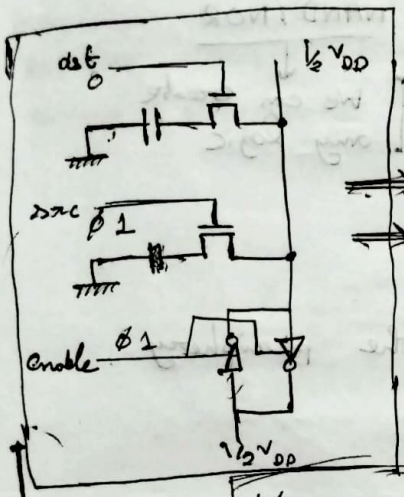


31/01/24

Bulk copy / Initialization :





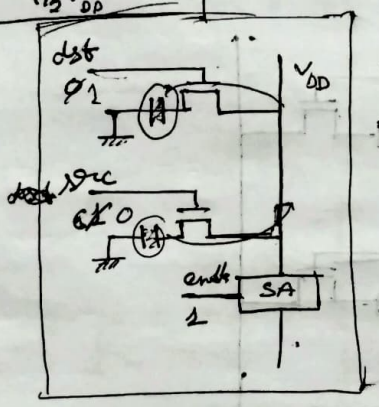


We want to transfer 'source' to 'destination'.

⇒ Activate source (0 → 1)

⇒ make enable → 1

now by '½ V_{DD} + δ' & 'SA'



⇒ deactivating 'source' (1 → 0)

⇒ activating 'destination' (0 → 1)

now, charge is shared to destination.

Hence, source is copied to destination

⇒ In reality, charge sharing is not 100% error proof.

⇒ If we activate one row, the whole subarray will be copied. We cannot copy a part of subarray.

DRISA :

3T1C → 3 transistor 1 cell | 1T1C → IN-P → adding logics to the D-RAM

IN-A

→

Adder

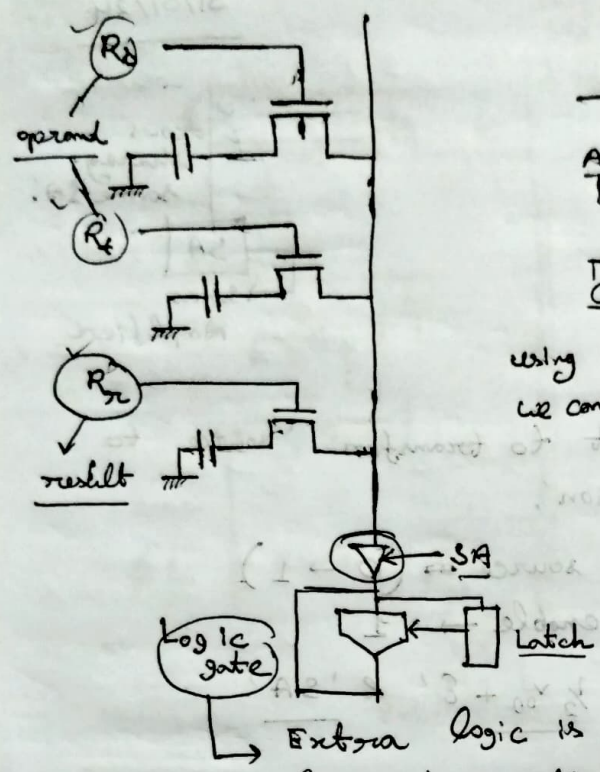
$\longrightarrow R_3R_t + R_tR_r + R_rR_3$

AND
- R_{3r} = 0 (initialized)
- R_{tr} = 1 (initialized)

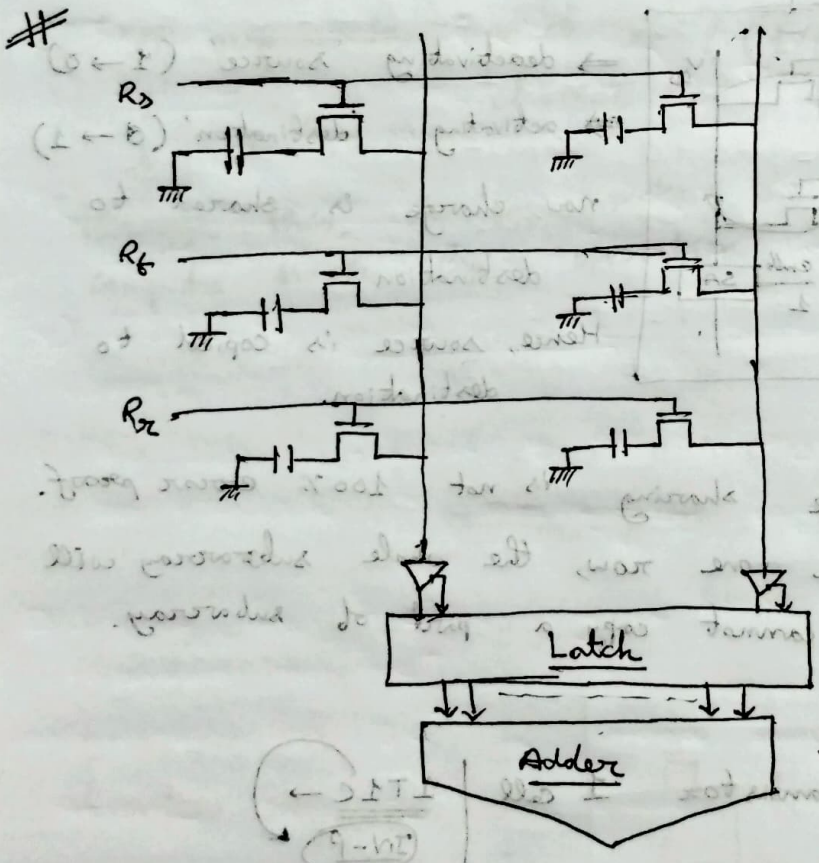
OR

using extra logic, we can realize NAND / NOR

[We can create any logic

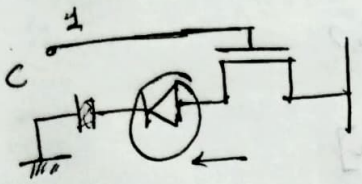


Extra logic is added in the periphery hence, it is called IM-P



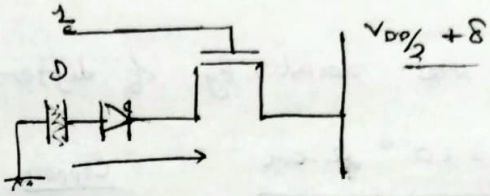
← Additional logic to perform arithmetic operation

ROC :



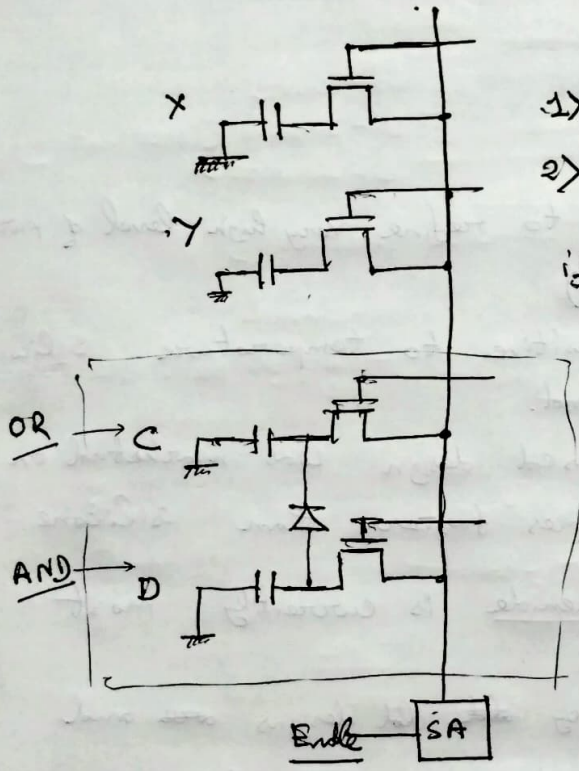
if this is charged nothing happens.

only when discharged ~~current~~ charge moves to the showed direction



If this is discharged nothing happens.

only when ~~dis~~charged, charge moves to the showed direction.



X + Y

1) Copy 'X' to 'C'

2) Copy 'Y' to 'D'

if D ~~d~~ has charge, then charge can flow to D → C but not vice versa, and ~~by~~ by this we can make sure if only 'D' is charged, charge is moved to 'C'.

by this AND/OR can be realized.

06/02/2024

What are resistivity of different materials?

→ $\rho_{Cu} = 10^{-6}\ \Omega.cm$ — Copper [Conductors]

→ $\rho_{Ge} = 50\ \Omega.cm$ — Germanium } Semiconductors

$\rho_{Si} = 10^{3}\ \Omega.cm$ — Silicon }

→ $\rho_{mica} = 10^{12}\ \Omega.cm$ [Insulator]

~~Compound~~ →

~~Germanium~~ Germanium:

i) Easily extractable, easy to refine, very high level of purity.

ii) Fabrication process. (=)

iii) Silicone is more sensitive to temperature. Silicone is more ~~abde~~ abdeentdand.

iv) Gralium arsenide based design was marketed on 1970 and it was 5 times faster than silicone.

v) Currently galium arsenide is currently most used.

vi) Silicone is favoured by ~~dd~~ dld designs ~~are~~ and ~~are~~ research is ~~beig beig beig~~ being done.

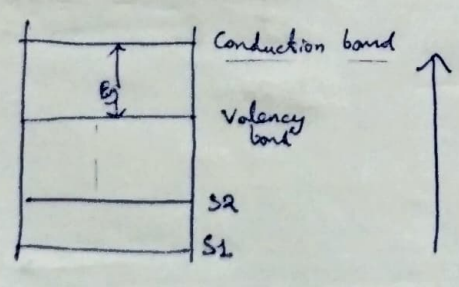
Silicone has $4e^-$ in valence band $(14e^-)$

Germanium has $4e^-$ " " " $(32e^-)$

Gralium has $3e^-$ in valence band $(31e^-)$

Arsenic has $5e^-$ in valence band $(32e^-)$

→



E_g → > 5eV (Insulator)

E_g → ≥ 0.67eV (semi)

E_g → ~0 (insulator)

↓

• they overlap.

▶ Let's assume

Si, $1e^-$ in conduction band among 10^{12} atom.

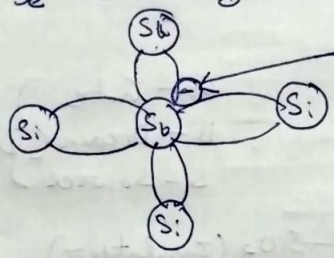
now, 1 impurity in 10^7 atoms.

↓

10^5 times increase in conduction.

▶ n-type \ P-type :

Sb → Antimony → $5e^-$ in valency band.



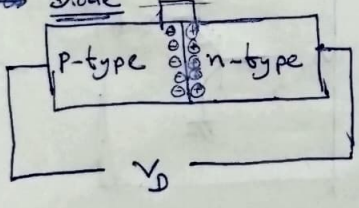
additional electron moves to conduction band.

n-type → majority carriers e^-

P-type → " " holes (h^+)

I'll be writing like this

Diode



Depletion region

→ can be biased using V_D.

→ no-bias → ($V_D = 0$)
forward-bias
reverse-bias

→ In reverse bias → The depletion region is extended.

Field effect Transistor :

→ JFET
→ MOSFET → metal oxide Field
→ MESFET

▷ Comparison b/w FET & BJT :

i) FET is unipolar.

ii) In BJT, current control.

FET, Voltage control.

↓

Output current, I_D is a function of gate voltage.

→

(iii) FET are more temperorily stable

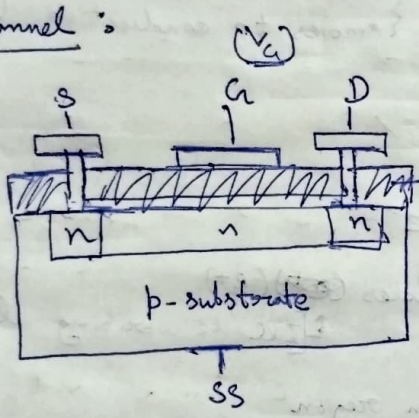
(iv) FET are smaller.

▶ MOSFET:

→ Depletion MOSFET
→ Enhancement MOSFET } .

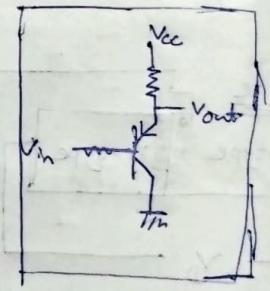
n-channel p-channel → Depletion MOSFET } .

▶ n-channel:



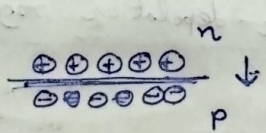
G - Gate
D - Drain (Destination)
S - Source

n-channel
↳ carrier is electron



V_{GS} → Gate to Source voltage
↳ controles the current flow.

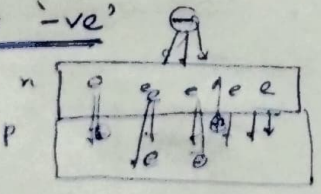
When, $V_G = 0$ (High impedence)
The, n-p junction will have



V_G → Input, I_{DS} → output (Drain to source current)

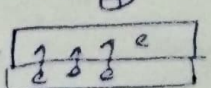
→ $V_G = 0$
$V_{DS} = +V_{DD}$ → I_D current

→ V_G = '-ve'

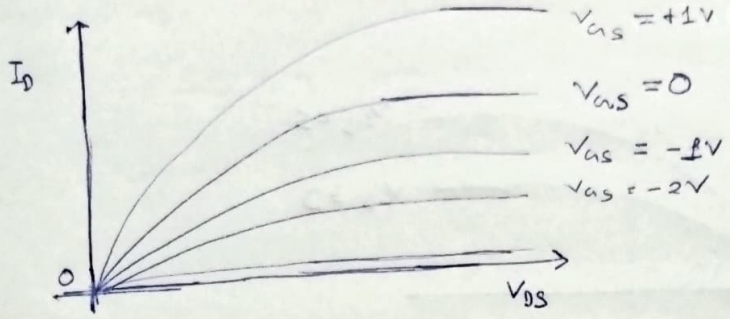


→ no. of free electron is reduced.
hence, I_D is less.

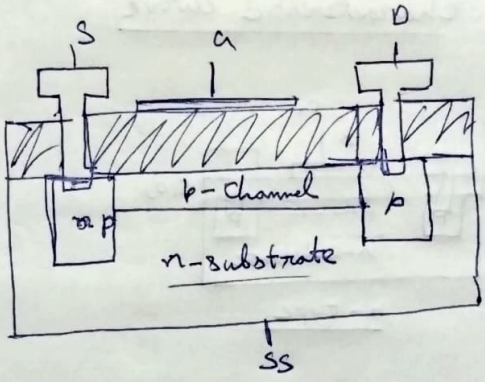
→ V_G = '+ve' ⊕



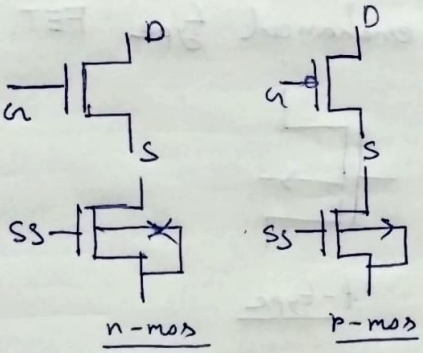
→ no. of free electrons is increased
hence, I_D is more.



Characteristic curve



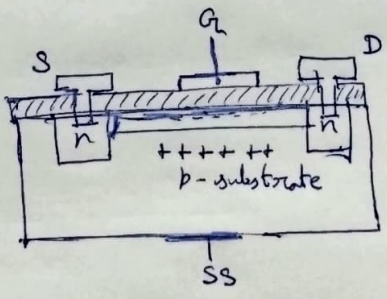
The logic is same as, n-channel.



Depletion type

07/02/24

Enhanced Type:



Threshold voltage

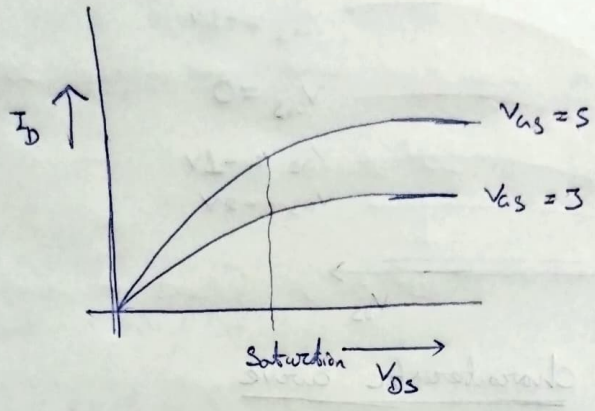
$V_{GS} \geq V_T$

p-type

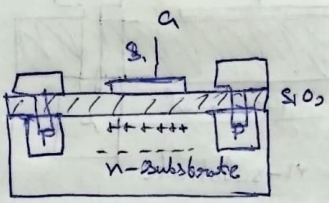
$V_{GS} > 0$

$V_{DS} > 0$

If we keep increasing V_{DS}, at some point current will be saturated as no more minority carrier will be available.



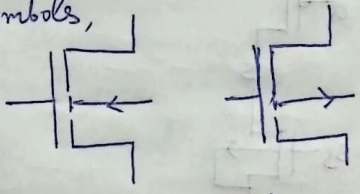
characteristic curve



n-type

* $V_G < 0$
  $V_{DS} < 0$

* There is no channel in enhanced type FET.
  hence, the symbols,



n-type p-type