Memory Design and Testing

Mahesh Awati
Department of Electronics and Communication Engg.

MEMORY DESIGN AND TESTING

UNIT 2 – Dynamic Random Access Memory

Mahesh Awati
Department of Electronics and Communication Engineering

UNIT 2 Outline

Unit 2: Dynamic Random Access Memory

- DRAM basics -
- Access and sense operations,
- Write operation,
- Opening a row,
- Open/Folded DRAM Array Architectures.
- The DRAM Array:
- The Mbit cell,
- The DRAM Capacitor - stacked and trench capacitors,
- The Sense Amplifier – Equilibration and bias circuits, Isolation Devices, Input/Output transistors, Nsense and Psense Amplifiers, Configurations, Operation;
- High Speed DRAMs: SDRAM, DDR

Introduction

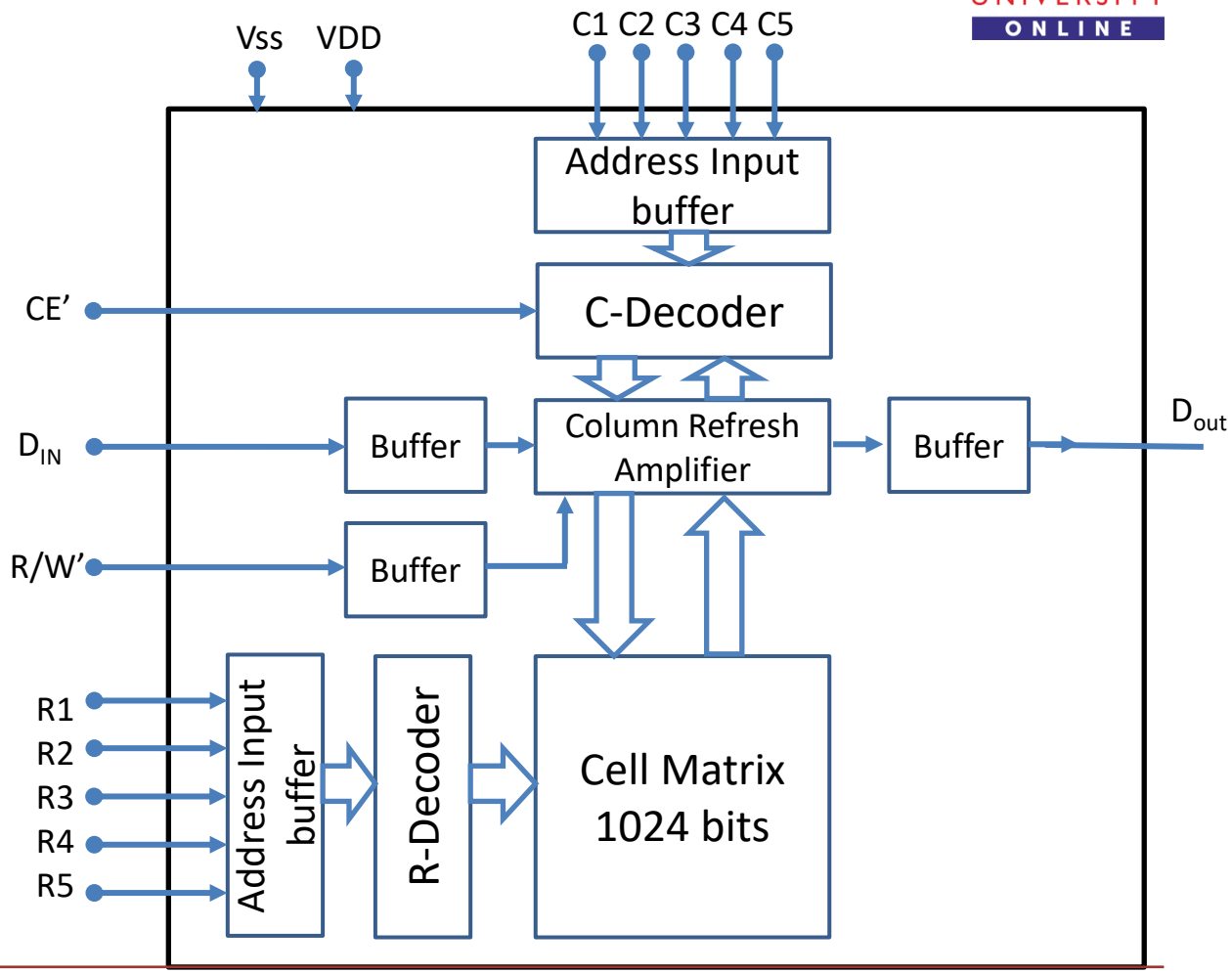
- **High Density Memory:** As the trend for high-density RAM arrays forces the memory cell size to shrink, alternative to 6T SRAM data storage **concepts must be considered to accommodate more cells in less silicon area.**
- **Charge on Capacitor:** In a **dynamic RAM cell**, binary data **is stored simply as charge in a capacitor**, where the **presence or absence of stored charge** determines the value of the stored bit.
- **Charge can't ne retained Indefinitely:** The data stored as **charge in a capacitor cannot be retained indefinitely**, because the **leakage currents** eventually remove or modify the stored charge.
- **Periodic Refreshing:** Thus, all dynamic memory cells require a **periodic refreshing of the stored data**, so that unwanted modifications due to leakage are prevented before they occur.
- **Access Devices are needed even though data is stored as charge in Capacitor:** Even though the binary data is stored as charge in a capacitor, **the DRAM cell must have access devices**, which can be activated externally for "read" and "write" operations. This requirement does not significantly affect the area advantage over the SRAM cell, since the cell access circuitry is usually very simple.
- **No static power is dissipated for storing charge on the capacitance:** Consequently, dynamic RAM arrays can achieve higher integration densities than SRAM arrays.

A Introduction to DRAM

First Generation (The 1k DRAM)

A (1024x1) bit DRAM

- A 1024 bit DRAM requires 10 address lines with 5 Row address lines (R1-R5) and Column address (C1-C5).
- Each address input is connected to a on-chip address input buffer. The input that drives the row (R) and Column (C) decoders.
- The 5 address inputs are connected through decoders to 1024 bit memory array in both row and column direction.
- The Memory array is having 32 rows[RA31:RA0] and 32 columns[CA31:CA0]. The DRAM Cell accessed depends on row address and column address.
- It may be Read from / Write to the DRAM Cell which is controlled by R/W' signal.
- The chip is enabled by CE' set as Low.



Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

A Introduction to DRAM

First Generation (The 1k DRAM)

The three basic operations involved are

1) Reading Data: 1 bit word from DRAM Array :

- First step set the chip in Read Mode by pulling R/W' = HIGH and then setting chip enable CE'=LOW.
- The address is placed on address lines which intern decodes the address of DRAM cell to be accessed and the data stored in the cell goes to Dout line through buffer.

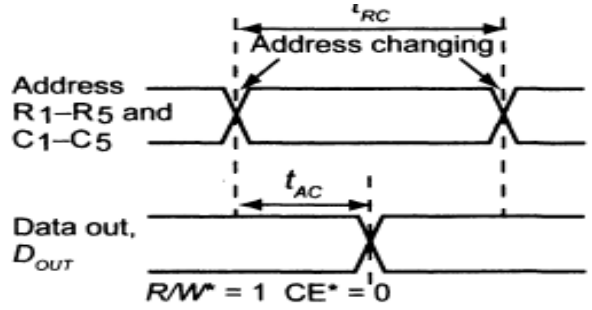
2) Writing Data: 1 bit word to DRAM Array

- Writing is accomplished by R/W'=0 with CE'=0 and valid data present on the D_{IN}
- The address is placed on address lines before the CE' and R/W' is made Low

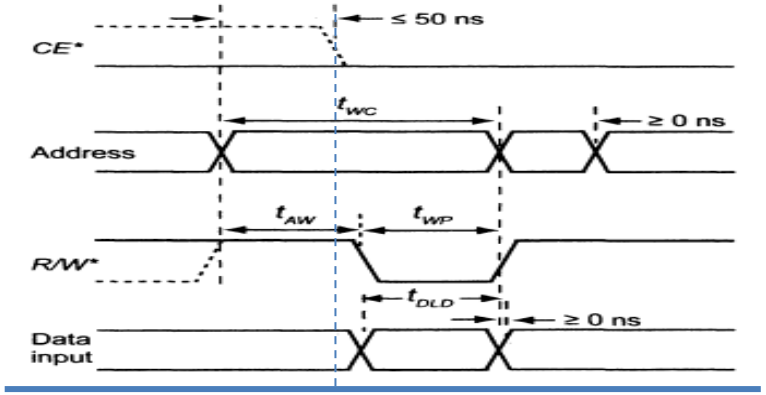
3) Refreshing the DRAM Array

- DRAM requires periodic refreshing so as to not loose the contents of the Memory Cells
- Refreshing is accomplished internally by reading data from selected cell and re-writing it back.
- R/W' itself is used as clock signal, R/W'=1 it reads data and when R/W'=0, It writes back the data

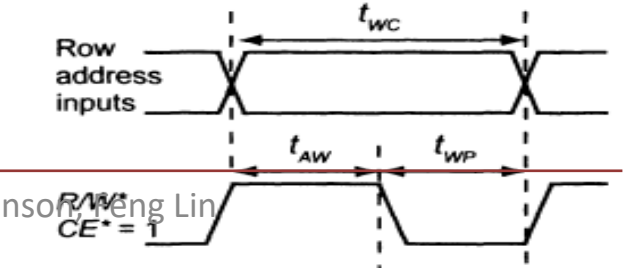
READ



WRITE



REFRESH

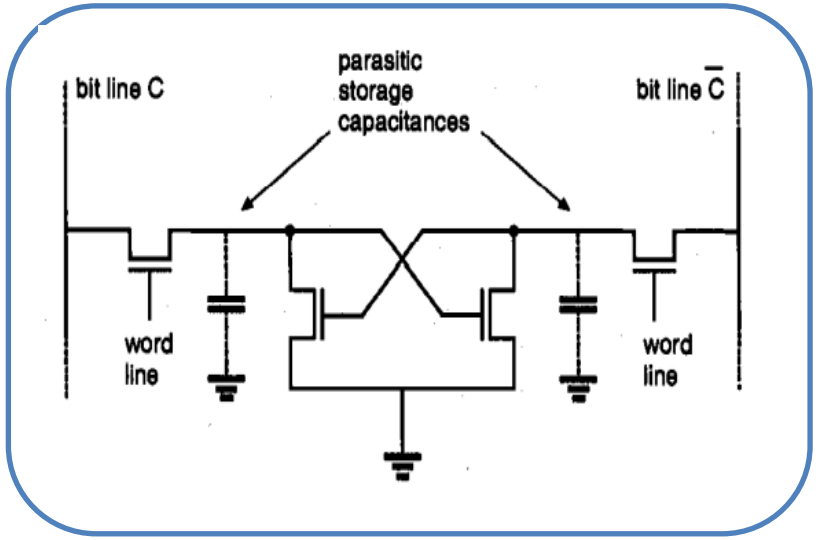


A Introduction to DRAM

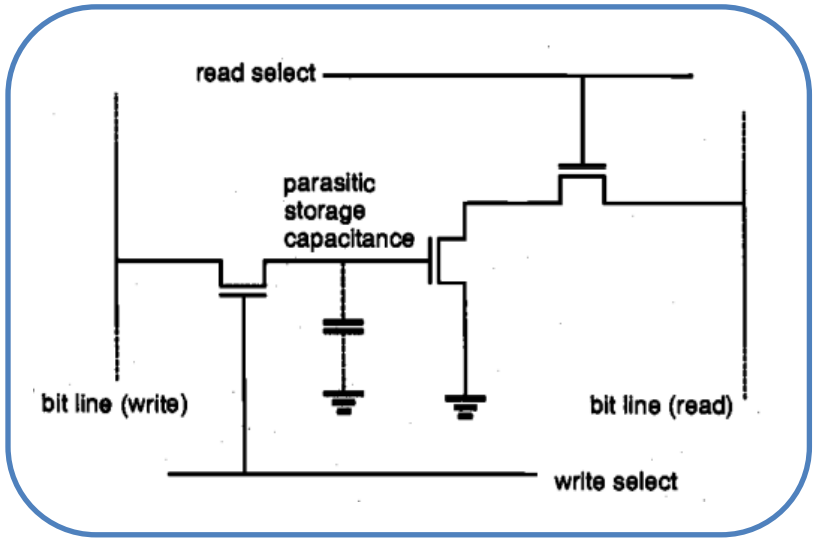
First Generation(The 1k DRAM)

- The historical evolution of the DRAM cell started with the four-transistor cell shown followed by Three transistor cell and then finally 1 transistor DRAM which has become a Industry standard for DRAM
- The first Generation of 1k DRAM used 3T DRAM Cells

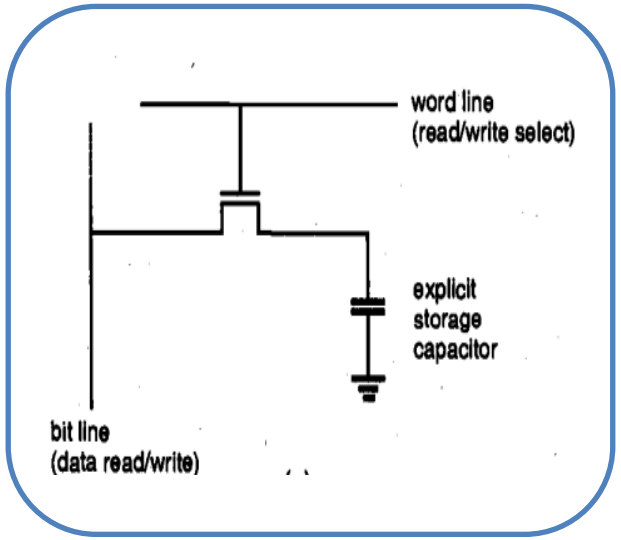
4T DRAM



3T DRAM



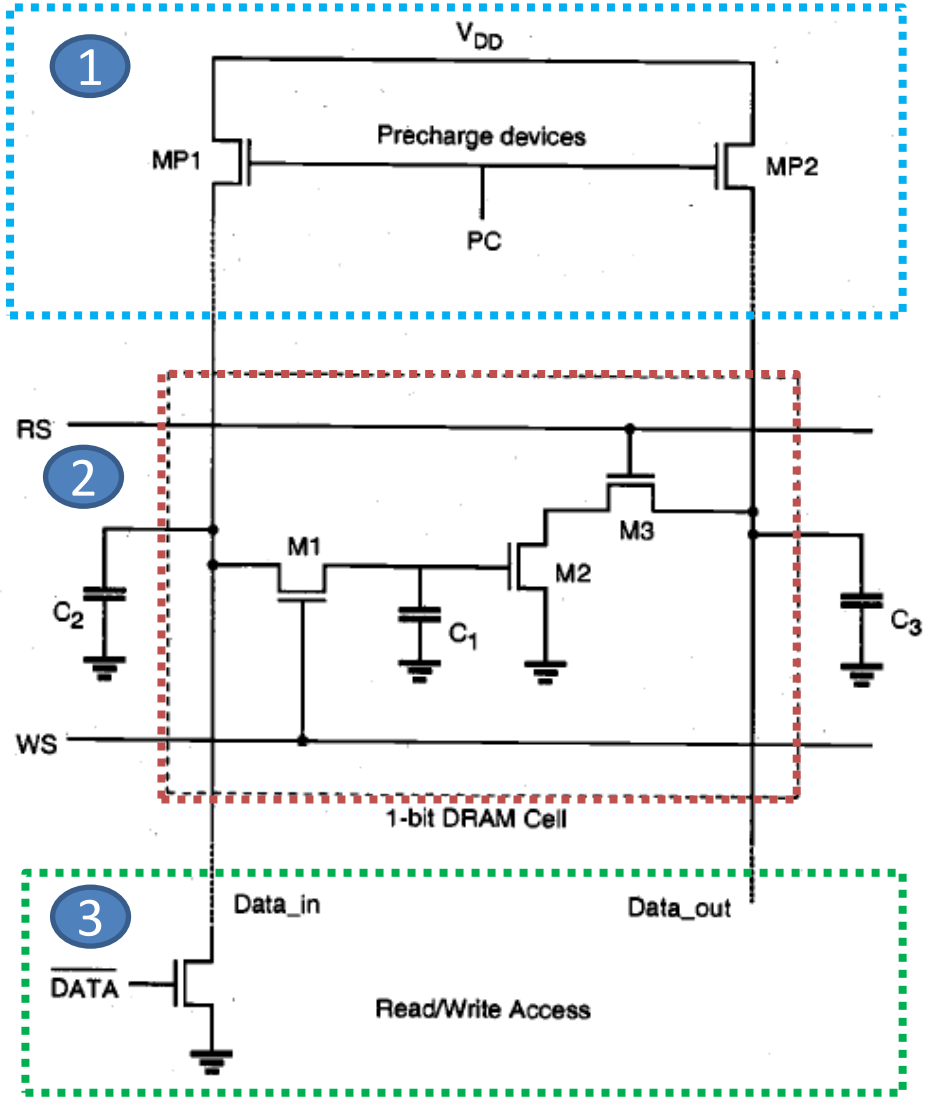
1T DRAM



First Generation

Second & Third Generation

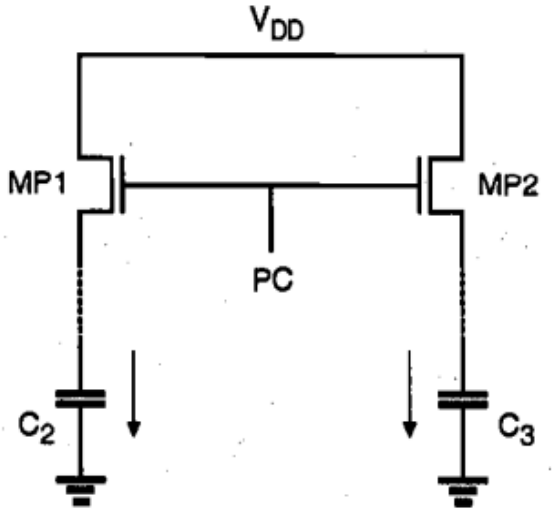
Three-Transistor DRAM Cell



- Three-transistor DRAM cell with the pull-up and read/write circuitry is as shown in figure. It basically consists of
  1. **Pre-charge Circuit**
  2. **3T DRAM Cell**
  3. **Read/Write Circuit**
- The cell has **two separate bit lines for "data read" and "data write"** as Data_in and Data_out.
- It also has **two separate word lines** as **WS and RS** to control the access transistors M1 & M3 respectively.
- Column capacitances C2 and C3 are being charged-up through MP1 and MP2 during the pre-charge cycle

Three-Transistor DRAM Cell

- The operation of the 3T DRAM and its peripheral circuitry is based on **a two-phase non-overlapping clock scheme (i.e., At any given Instant of Time both Ø1 and Ø2 will be simultaneously HIGH) .**
- The **pre-charge events are driven by Ø1 (PC. Ø1)**, whereas the **"read" and "write" events are driven by Ø2 (RS. Ø2 or WS. Ø2)**
- Every “Read" and “write" operation is preceded by a pre-charge cycle, which is initiated by making signal PC high.
- **During the pre-charge cycle,** the **column pull-up transistors MP1 and MP2 are activated,** and the corresponding column capacitances C2 and C3 are charged up to logic-high level.
- The pre-charge cycle is effectively completed when both capacitance voltages reach their steady-state values.



C2 and C3 are being charged up through MP1 and MP2 during the pre-charge cycle

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Three-Transistor DRAM Cell

Write Operation: Write ‘1’ i.e., Data=1

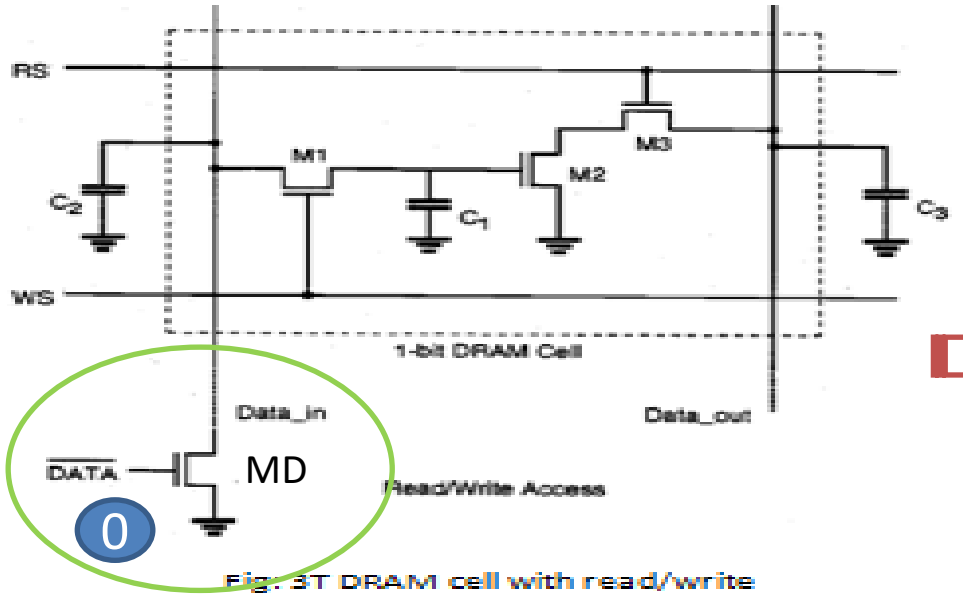
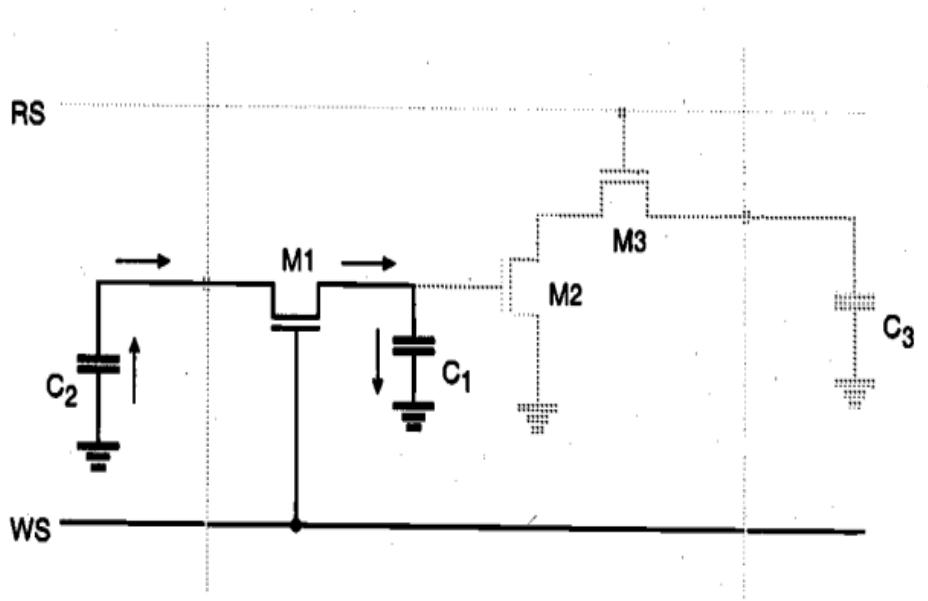


Fig. 3T DRAM cell with read/write circuitry



As Data=1, Data’=0 then Data’=0
Consequently, the "data write" transistor MD is turned off, and the voltage level on column Din remains high.

Now, the signal WS is pulled high during the active phase of **Ø2 (WS. Ø2=1)**.
Now, the write access transistor M1 is turned on. With M1 conducting, the charge on C2 is now shared with C1.

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Three-Transistor DRAM Cell

Write Operation: Write '1' i.e., Data=1

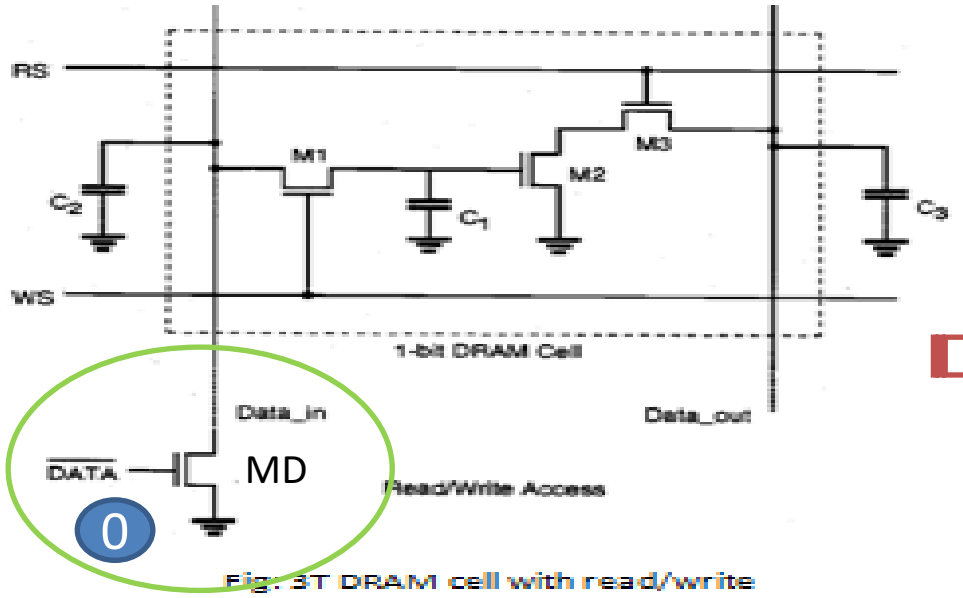
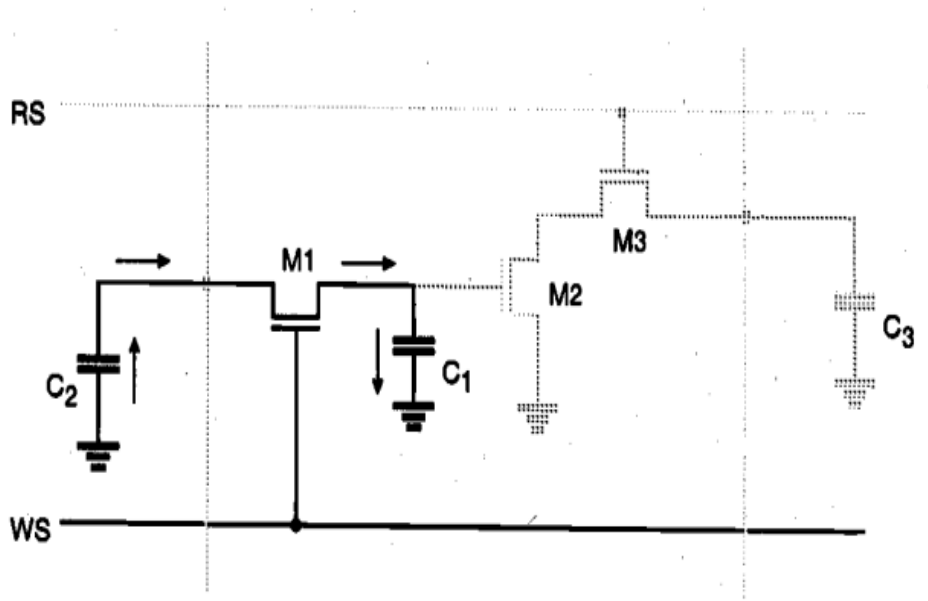


Fig. 3T DRAM cell with read/write circuitry



As Data=1, Data'=0 then Data'=0
Consequently, the "data write" transistor MD is turned off, and the voltage level on column Din remains high.

Since the C2 is very large compared to C1, the storage node capacitance C1 attains **approximately the same logic-high** level at the end of the charge-sharing process.

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Three-Transistor DRAM Cell

Write Operation: Write ‘1’ i.e., Data=1

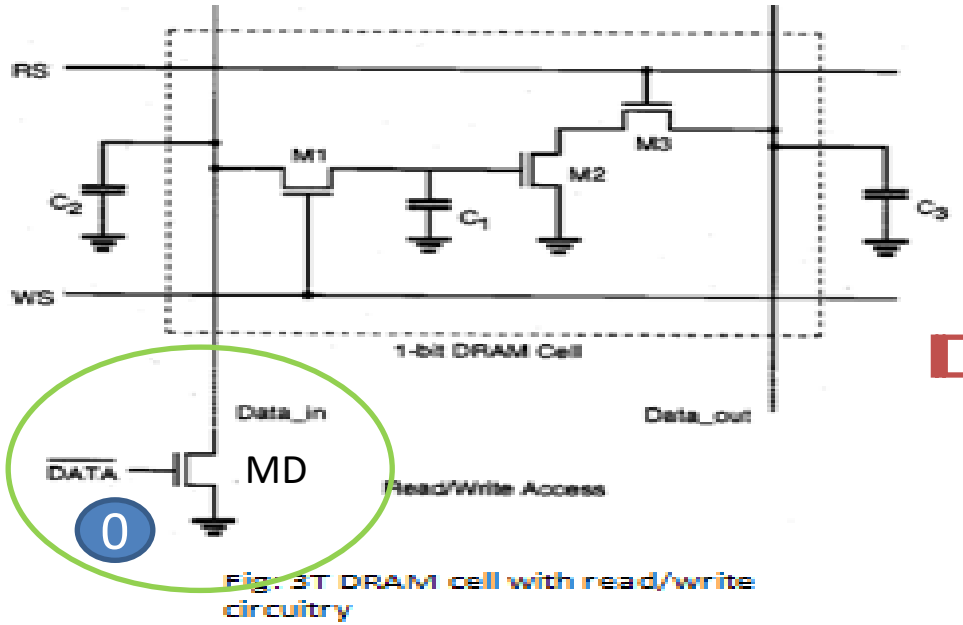
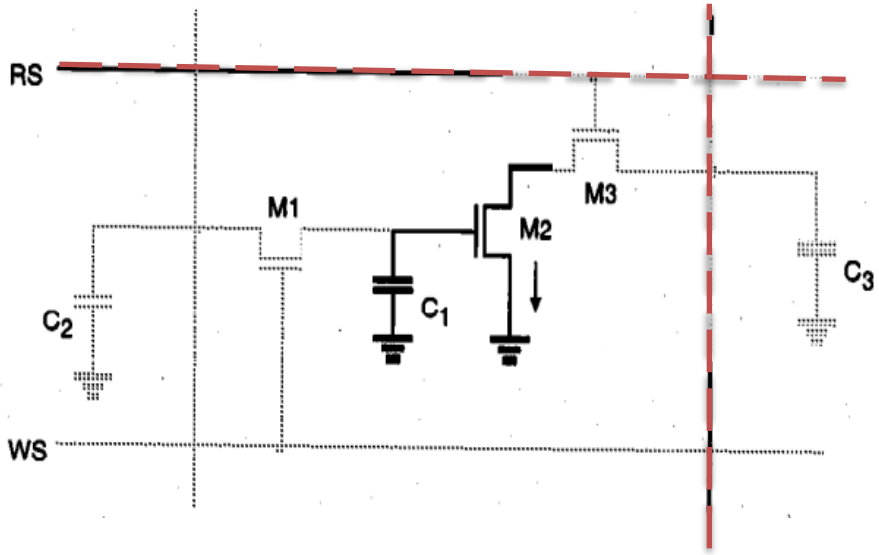


Fig. 3T DRAM cell with read/write circuitry



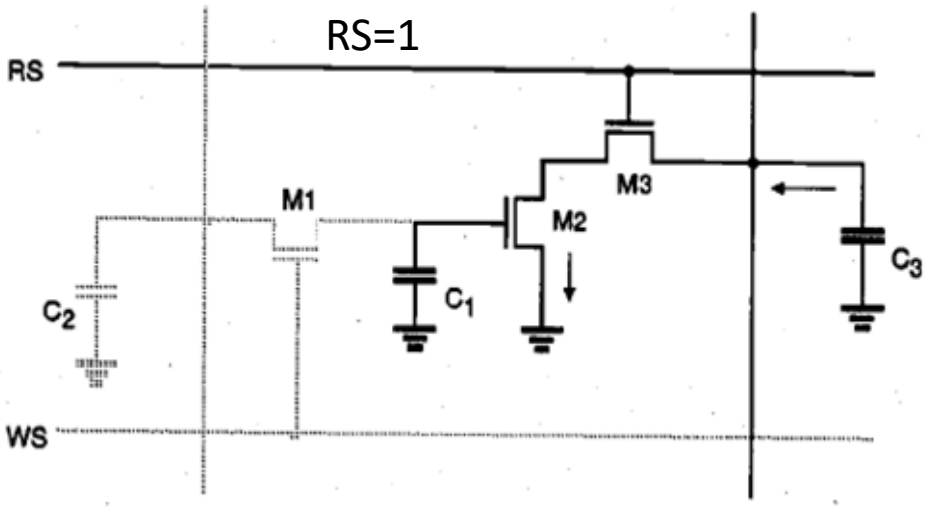
As Data=1, Data’=0 then Data’=0
Consequently, the "data write" transistor MD is turned off, and the voltage level on column Din remains high.

After the write "1" operation is completed, the **write access transistor M1 is turned off**. With the **storage capacitance C1 charged-up to a logic-high level**, transistor **M2 is now conducting.**

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

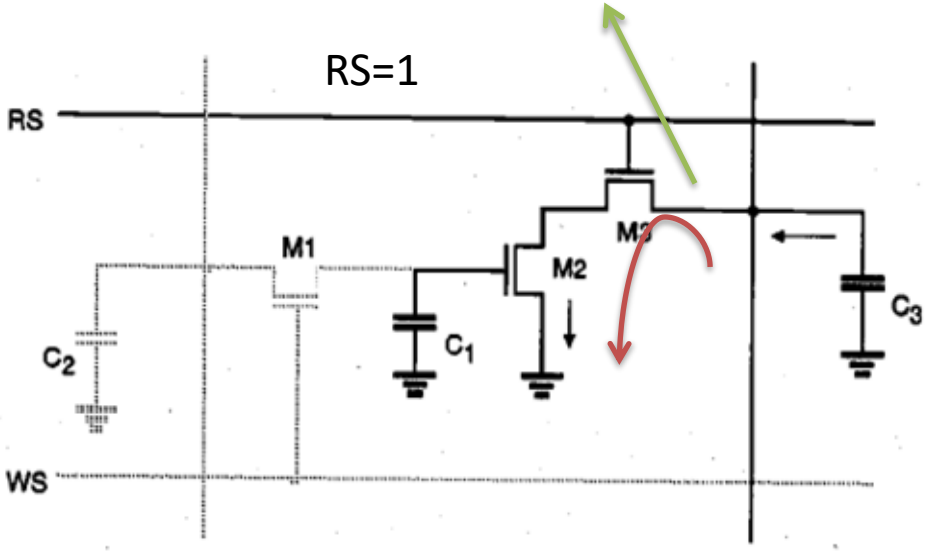
Three-Transistor DRAM Cell

Read Operation: Read ‘1’



The “read select” signal RS must be pulled high during the active phase of **Ø2 (WS. Ø2=1)**, following a pre-charge cycle.

The capacitance C3 discharges through M2 and M3, and the falling column voltage is interpreted by the "data read" circuitry as a stored logic "1."



As the read access transistor **M3 turns on, M2 and M3 create a conducting path between the "data read" column capacitance C3 and the ground.**

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Three-Transistor DRAM Cell

Write Operation: Write ‘0’ i.e., Data=0

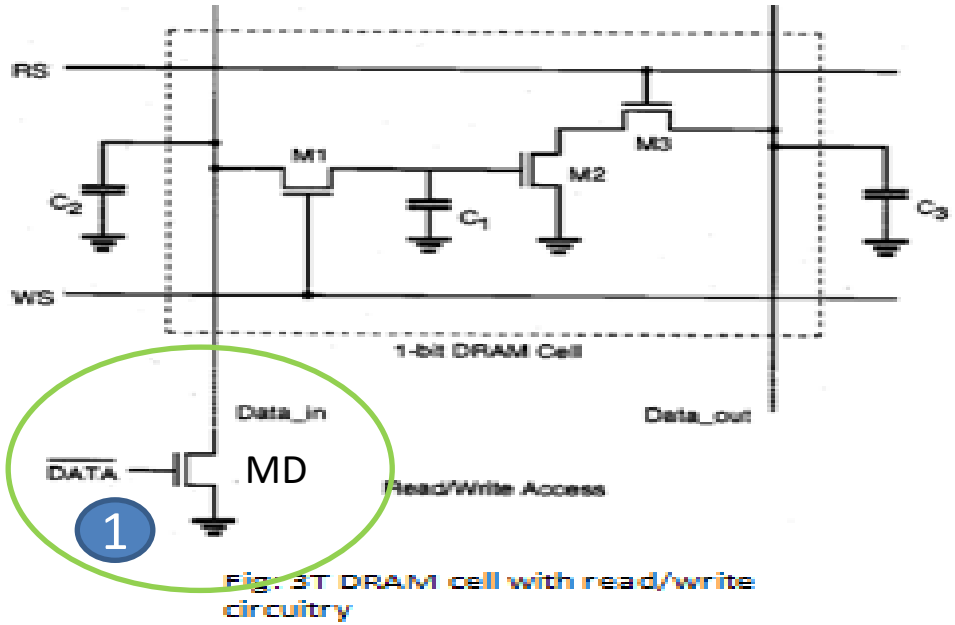
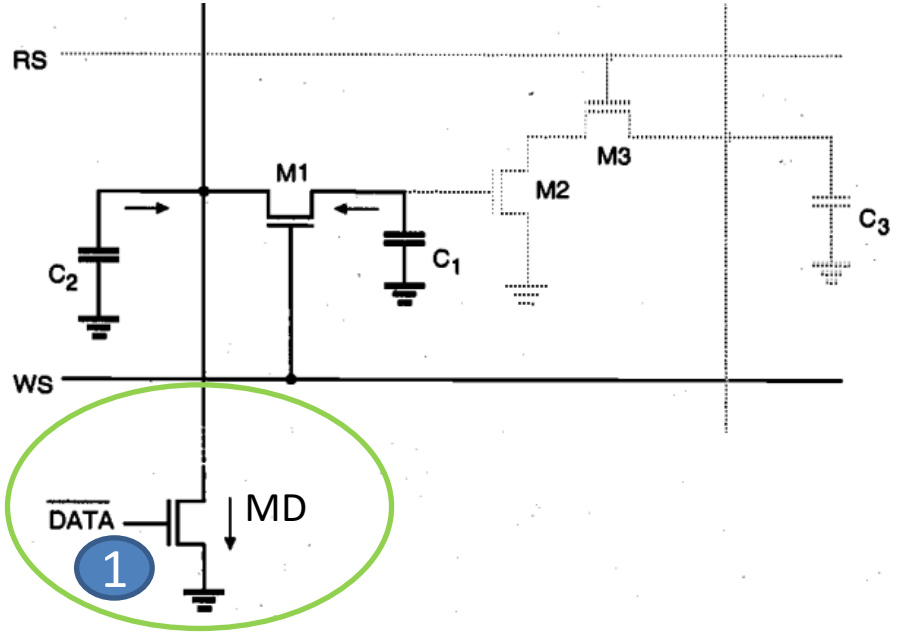


Fig. 3T DRAM cell with read/write circuitry



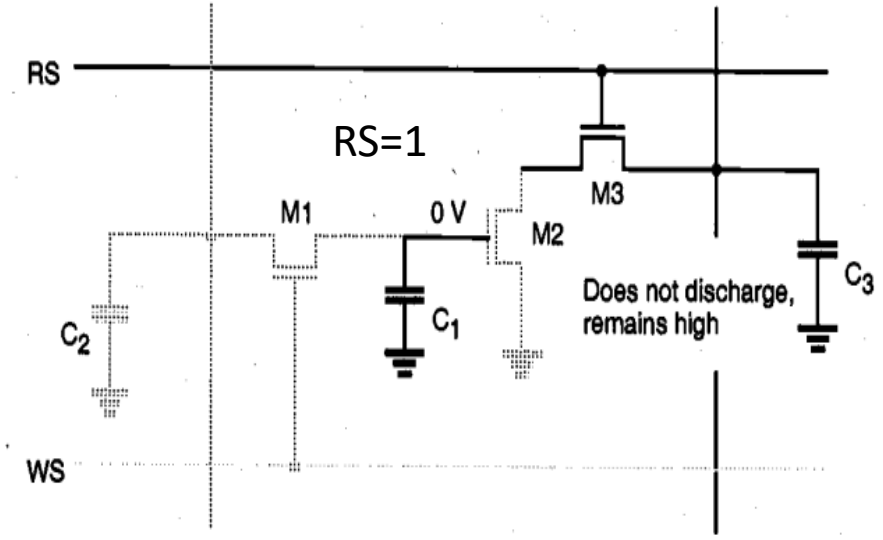
As Data=0, then Data’=0. This makes the "data write" transistor MD to turned ON and the voltage level on column Din is pulled to logic "0.

Now the WL is made HIGH during active phase of **Ø2 (WS. Ø2=1)**. As a result M1 is turned ON. The voltage level on C2, as well as that on the storage node C1, is pulled to logic "0“. C1 through M1 and the data write transistor, MD and C2 through MD.

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Three-Transistor DRAM Cell

Read Operation: Read ‘0’



The “read select” signal RS must be pulled high during the active phase of **Ø2 (WS. Ø2=1)**, following a pre-charge cycle.

As the read access transistor **M3 turns on,** but since M2 is off, **there is no conducting path between the column capacitance C3 and the ground**

Consequently, C3 does not discharge, and the logic-high level on the Dout column is interpreted by the data read circuitry as a stored "0" bit

Three-Transistor DRAM Cell

Refresh Operation

- The charge stored in C1 cannot be held indefinitely, even though the **"data read" operations do not significantly disturb the stored charge**.
- The **drain junction leakage current of the write access transistor M1** is the main reason for the gradual depletion of the stored charge on C1.
- In order to **refresh the data stored** in **the DRAM cells before they are altered due to leakage**,
  - the data must be **periodically read,**
  - **inverted** (since the data output level reflects the inverse of the stored data), and
  - **then written back** into the same cell location.
- This refresh operation is performed for **all storage cells in the DRAM array for every 2 to 4 ms.**
- Note that **all bits in one row can be refreshed at once,** which significantly simplifies the procedure.

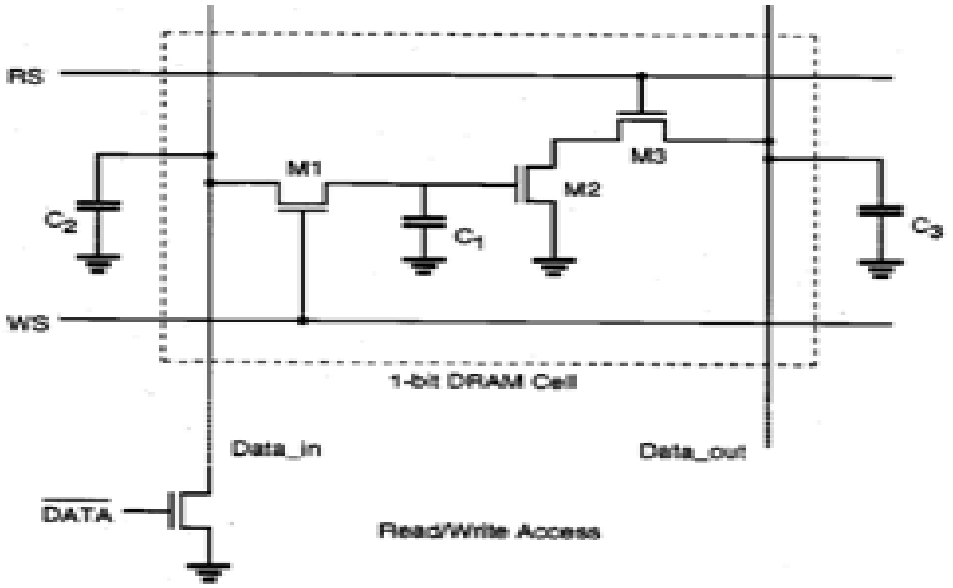


Fig: 3T DRAM cell with read/write circuitry

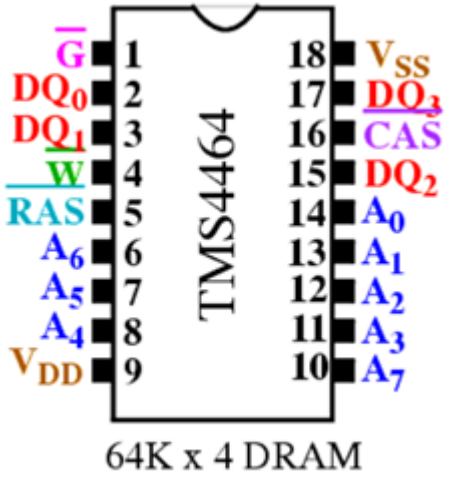
Dynamic RAM (DRAM)

Second Generation of DRAM

Density is very Important factor in DRAM –i.e., how many bits are packed in limited area is the important factor.

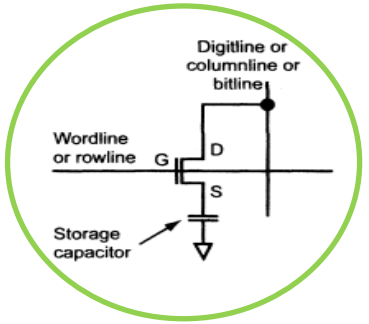
How High Density can be achieved?

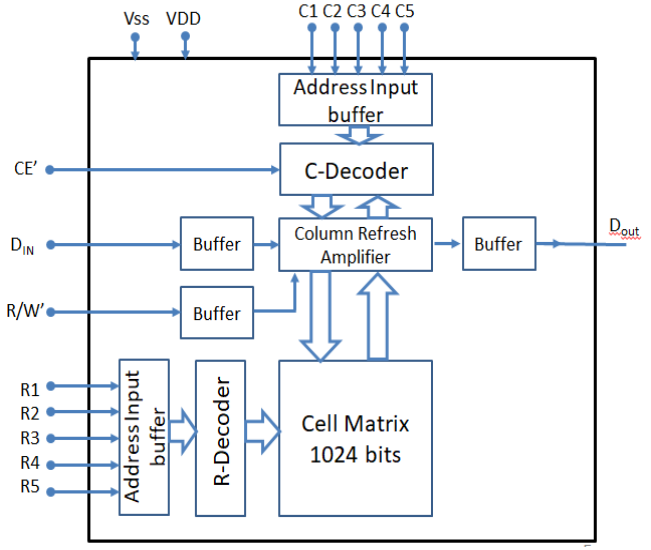
- By reducing the Size of the Cell – From 3T to 1T
- BY reducing the additional peripheral circuits required like PINs.



64K x 4 DRAM

Pin(s)	Function
A_0-A_7	Address
DQ_0-DQ_3	Data In/Data Out
$\overline{RAS}$	Row Address Strobe
$\overline{CAS}$	Column Address Strobe
$\overline{G}$	Output Enable
$\overline{W}$	Write Enable





First Generation

- 3T DRAM
- Separate PINs for Row and Column Decoding

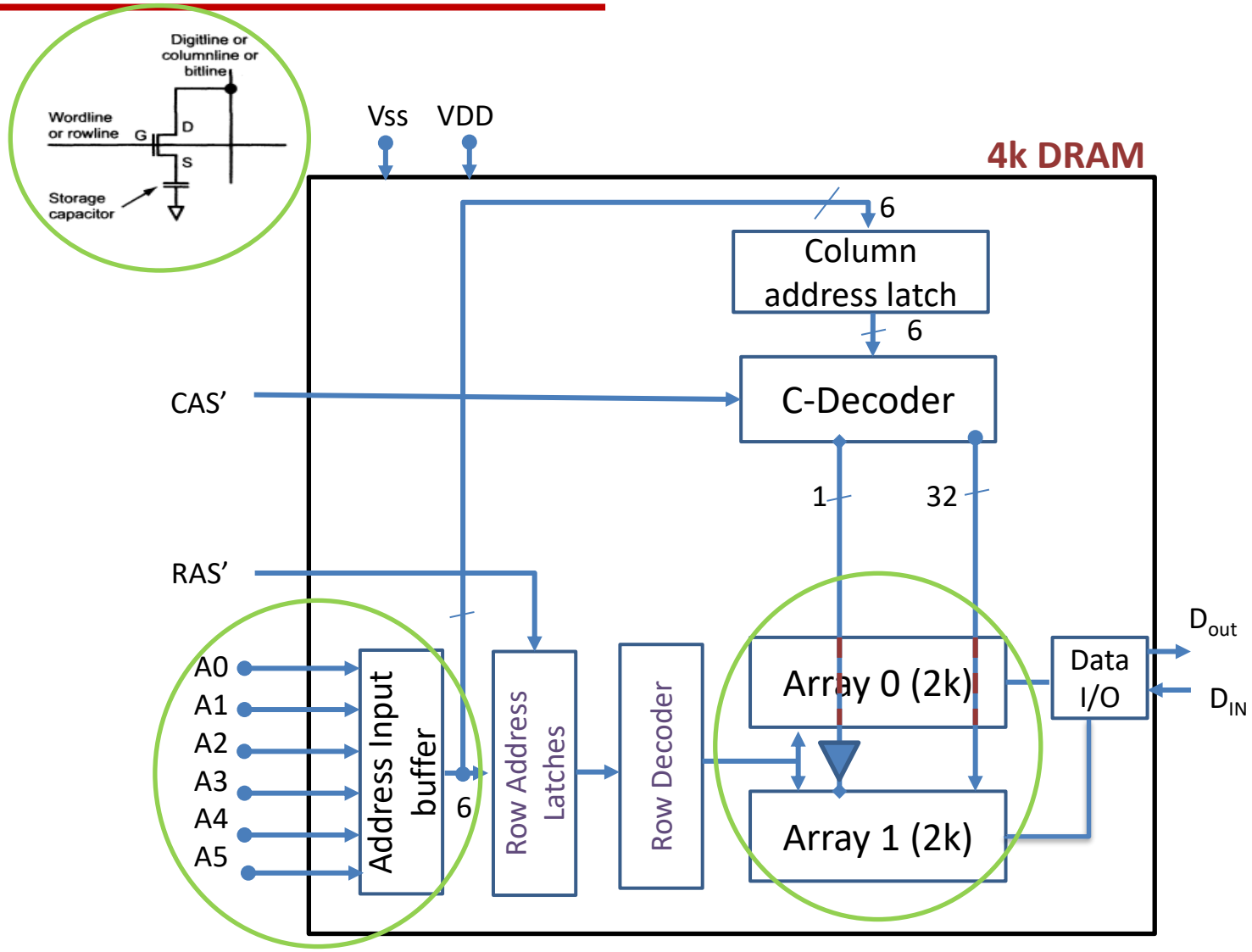
Dynamic RAM (DRAM)

Second Generation of DRAM

Key Features of 2nd generation are

- The Second generation of DRAM introduced
  - ✓ **1 Transistor / 1 Capacitor Memory Cell**
  - ✓ **Multiplexed address inputs**: Reduces the number of Address Input pins log2(4k)=12 but as Multiplexed address Inputs are used , it uses only 6 Address lines.
  - ✓ The Multiplexed addressing in which the same address inputs **are used for both the row and column addressing.**

0000 0000 0000 0000 0000 0000 0000 000 0
0000 0000 0000 0000 0000 0000 0000 001 0
0000 0000 0000 0000 0000 0000 0000 000 1
0000 0000 0000 0000 0000 0000 0000 001 1

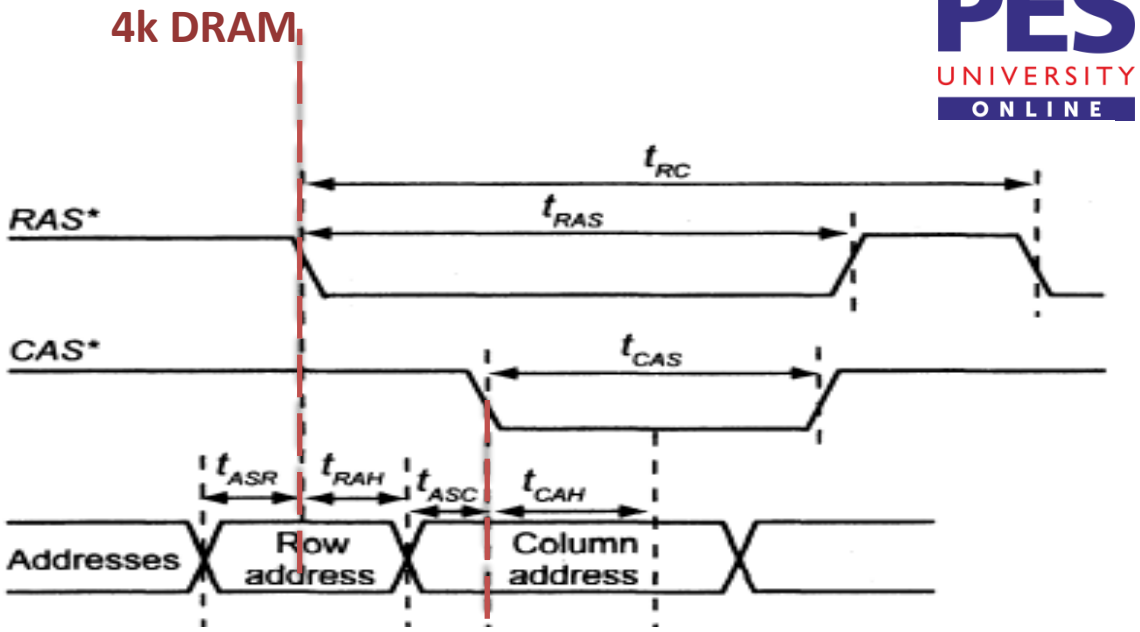


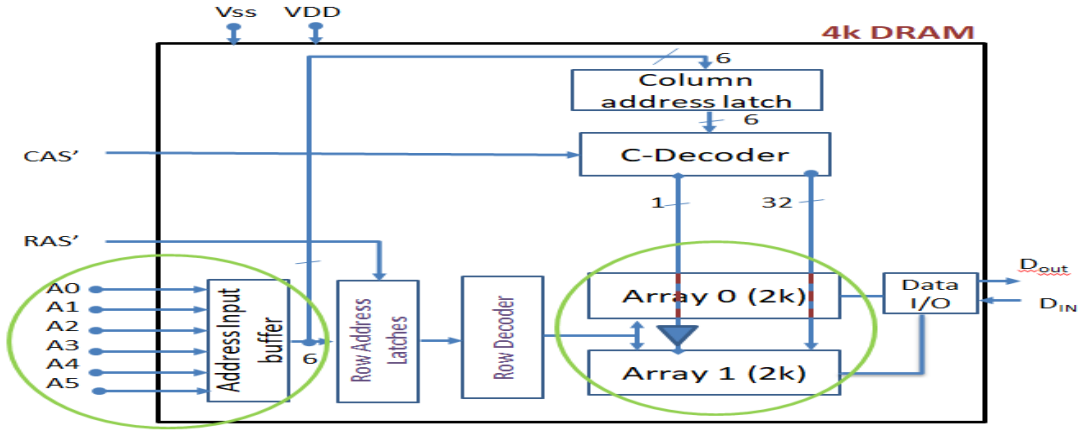
Dynamic RAM (DRAM)

Second Generation of DRAM

Key Features of 2nd generation are

- The Second generation of DRAM introduced
  - **1 Transistor / 1 Capacitor Memory Cell**
  - **Multiplexed address inputs**:
    - The **Row Address Strobe (RAS')** input clocks **the address on [A5:A0] into row address latch on it's falling edge.**
    - The **Column Address Strobe (CAS')** input clocks the **address on [A5:A0] into Column address latch on it's falling edge.**
    - The parameters t_{ASR}, t_{RAH}, t_{ASC}, and t_{CAH} indicate the **setup and hold times for the row and column addresses**, respectively.



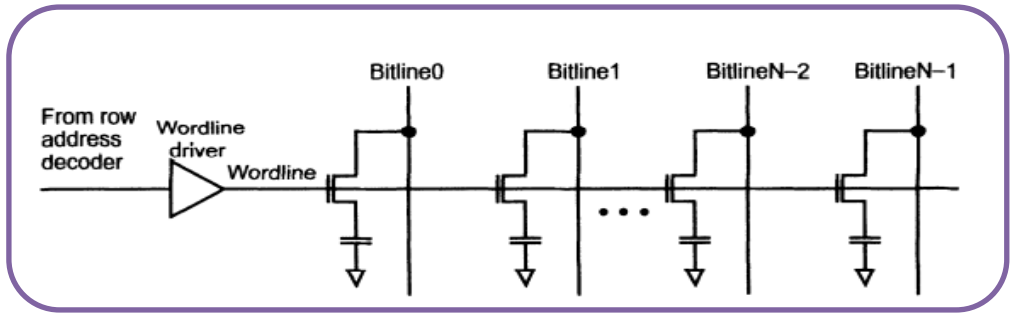


Dynamic RAM (DRAM)

Second Generation of DRAM

Key Features of 2nd generation are

- The Second generation of DRAM introduced **Multiple or Segmented Memory arrays**



Why Multiple Arrays ?

- ✓ Splitting up the memory into more than one array at the cost of a larger layout area is **to reduce the parasitic present in the dynamic memory circuit element.**
- ✓ A single Memory array may lead to large delay and limits the speed of DRAM
  - ✓ More bit-lines we use in an array, the longer the delay through the word-line
  - ✓ If we drive the **word-line HIGH on the left side, the signal will take a finite time to reach the end of the word-line on the right.**
  - ✓ This is due to the **distributed resistance/capacitance** structure formed by the resistance of the **poly-silicon word-line** and the **capacitance of the MOSFET gates**.
  - ✓ The delay limits the speed of DRAM operation i.e**, how quickly a row can be opened and closed.**

Solution to reduce this RC time along word-line

a) **A polycide word-line** which will be reducing the word-line resistance.
b) **A additional drivers can be placed** at different locations along the word-line.
c) The word-line can be **stitched at various locations with metal.**
d) **Increase the Number of word-line**

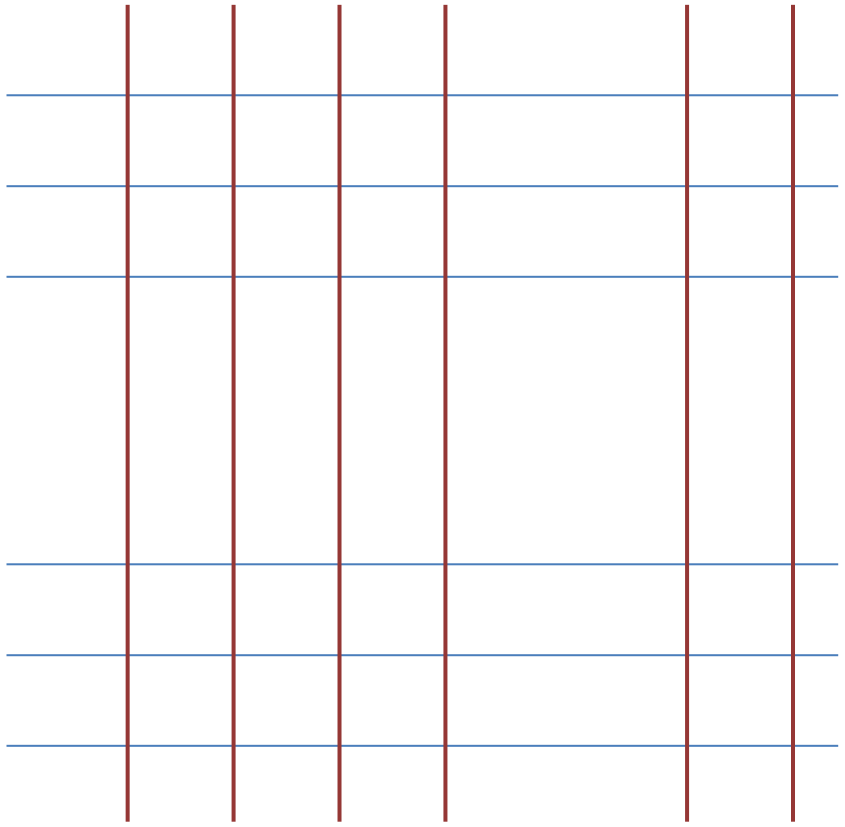
Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

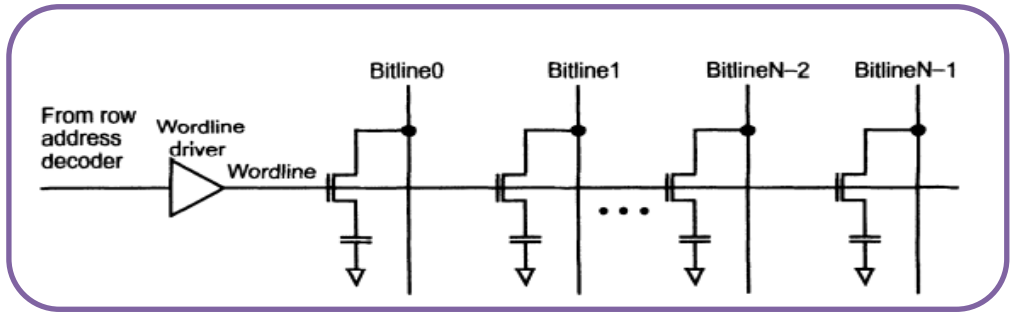
Dynamic RAM (DRAM)

Second Generation of DRAM

Key Features of 2nd generation are

- The Second generation of DRAM introduced **Multiple or Segmented Memory arrays**



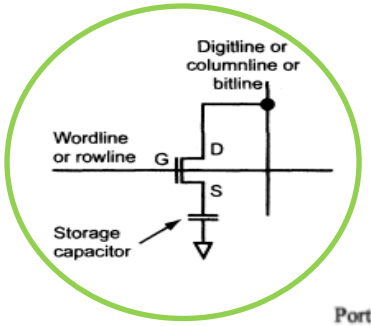


Solution to reduce this RC time along word-line

a) **A polycide word-line** which will be reducing the word-line resistance.
b) **A additional drivers can be placed** at different locations along the word-line.
c) The word-line can be **stitched at various locations with metal.**
d) **Increase the Number of word-line**

Dynamic RAM (DRAM)

Second Generation of DRAM

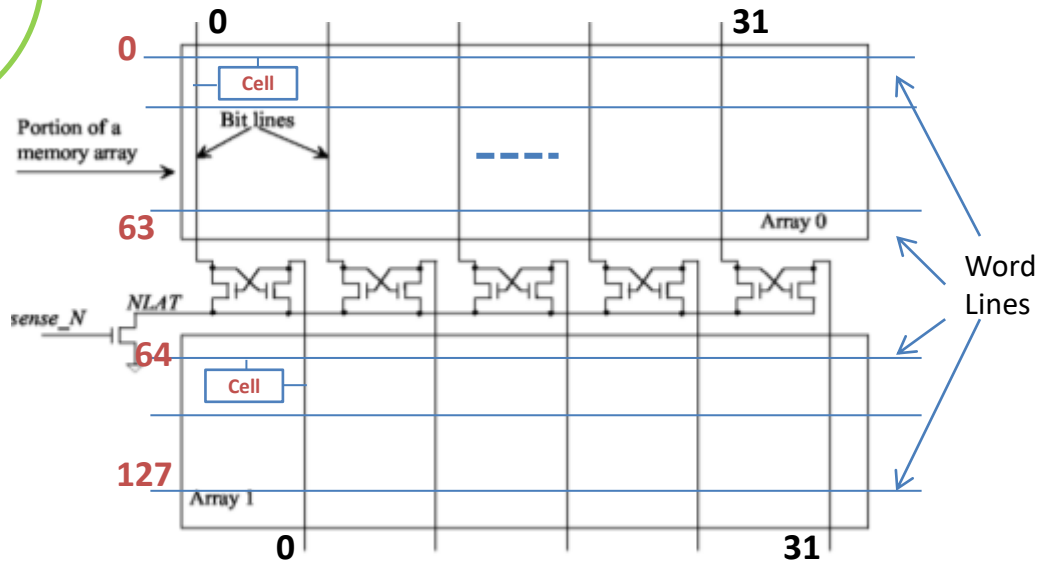


Key Features of 2nd generation are

- The Second generation of DRAM introduced **Multiple or Segmented Memory arrays**

Why there is a Limitation on word Lines ?

- **By adding more word-lines to the array,** more **parasitic capacitance is added to the bit-lines.**
- This parasitic capacitance becomes important when sensing the value of data charge stored in the memory element.
- Therefore, **to reduce the parasitic of bit-line the arrays are divided as Array 0 and Array1** instead of using it as Single Array.
- Ex: A DRAM of 4kbits used as Single array of 128 word lines and 32 bit-lines might have restricted word-lines length along the row but bit-line capacitance is HIGH.
- Instead divide 4kbit DRAM into two arrays of 2kbits. Each arrays having 64 word lines and 32 bit-lines (64x32=2048)



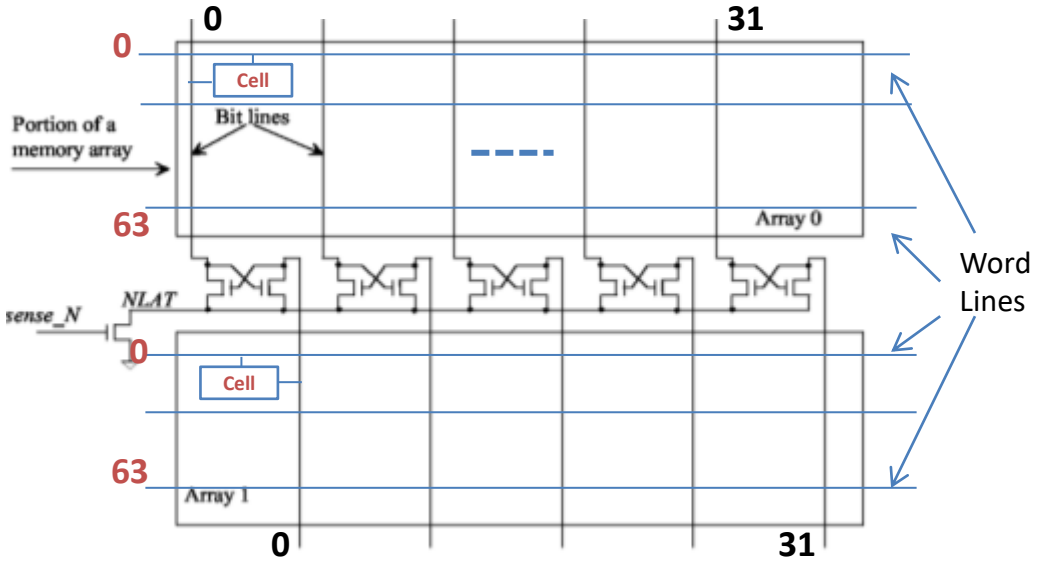
4k DRAM organized as 2 arrays of 2kbits

Dynamic RAM (DRAM)

Second Generation of DRAM

Memory Array Size:

- 4k DRAM is arranged as two 2k-DRAM memory arrays. Each 2k memory is composed of 64 wordlines and 32 bitlines for 2,048 memory elements per array.
- From our discussion earlier, we can open a row in Array0 while at the same time opening a row in Array1 by simply applying a row address to the input address pins and driving RAS* LOW.
- Once the rows are open, it is a simple matter of changing the column address to select different data associated with the same open row from either array.
- If our word size is 1 bit, we could define a page as being 64 bits in length (32 bits from each array).
- We could also define our page size as 32 bits with a 2-bit word for input/output.



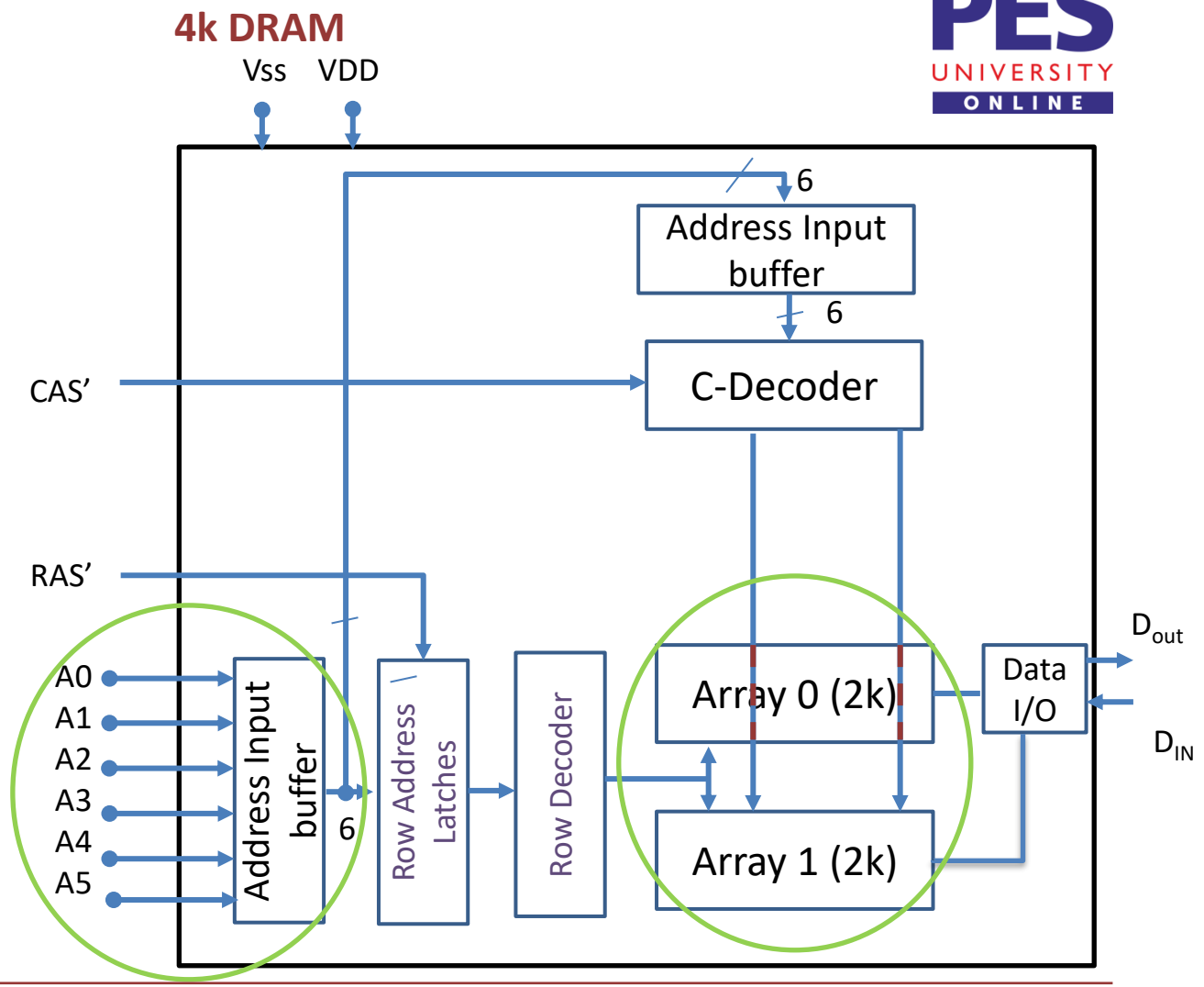
4k DRAM organized as 2 arrays of 2kbits

Dynamic RAM (DRAM)

Second Generation of DRAM

Key Features of 2nd generation are

- The Second generation of DRAM introduced - **They offer more modes of operations for grater flexibility and higher speed operations.**
  - ✓ Modes like
    - ✓ Page mode,
    - ✓ Nibble mode,
    - ✓ Static column mode,
    - ✓ Fast page mode and
    - ✓ Extended data output mode
- Second generation Memory size ranges from 4k to **64Mbits.**
- The technology advanced from NMOS to CMOS



Dynamic RAM (DRAM)

Second Generation of DRAM

Refreshing the DRAM.

- Refreshing the DRAM is **accomplished by sequentially opening each row** in the DRAM.
- In some DRAMs, an **internal row address counter is present to make the DRAM easier to refresh.**
- DRAM **rows need to be refreshed periodically**

  In a DRAM Memory, If every row has Refresh period is 2 milliseconds (Refresh to be done every 2 mSec). Assume that a DRAM chip has 2^{14} rows. The time taken for every refresh operation is 50nSec. What is the percentage of time available for DRAM for perform read/write operation.

  Answer: Refresh Period is 2mSec– It means all the Rows must be refreshed for every 2mSec.

  Total Refresh Time for a chip = No of Rows x 50n = 2^{14} x 50n=0.819 mSec

  Time for Read/write operation =2m-0.819m =1.181mSec

  % of time available for Read/Write operation is = (1.181m/2m)x 100 =59%.
- Refresh operation requirement makes the DRAM slower compared to SRAM.

Refresh period is decided by the Minimum Time taken by capacitor holding the data to get discharge through leakage paths which leads to change in information.

Dynamic RAM (DRAM)

Observation to be made

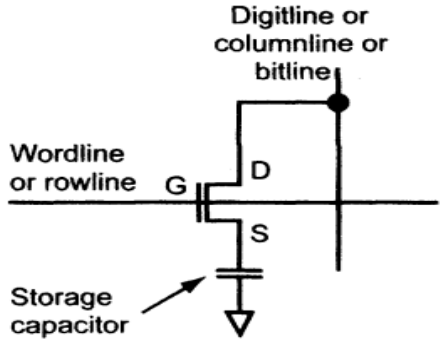


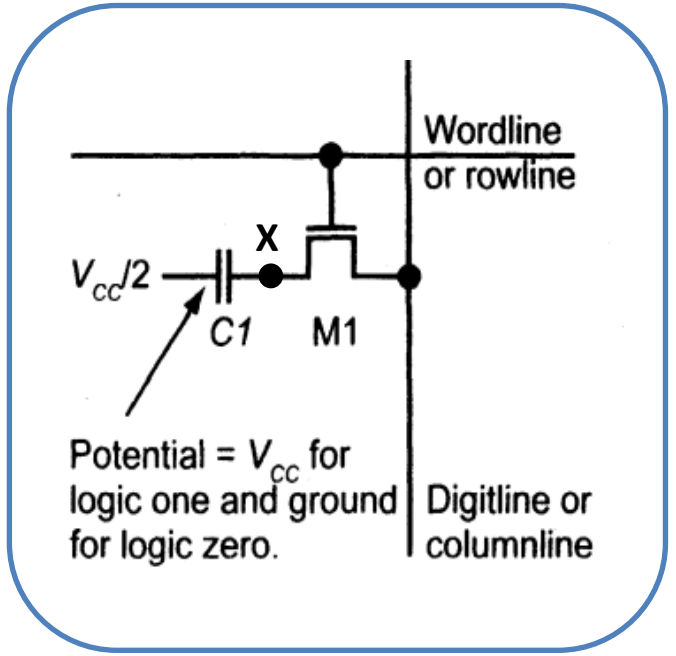
Figure 1.12 1-transistor, 1-capacitor (1T1C) memory cell.

1. The wordline (rowline) may be fabricated using polysilicon (poly). This allows the MOSFET to be formed by crossing the poly wordline over an n+ active area.
2. To write a full V_{CC} logic voltage (where V_{CC} is the maximum positive power supply voltage) to the storage capacitor, the rowline must be driven to a voltage greater than V_{CC} + the n-channel MOSFET threshold voltage (with body effect). This voltage, $> V_{CC} + V_{TH}$, is often labeled V_{CC} *pumped* (V_{CCP}).
3. The bitline (columnline) may be made using metal or polysilicon. The main concern, as we'll show in a moment, is to reduce the parasitic capacitance associated with the bitline.

Dynamic RAM (DRAM)

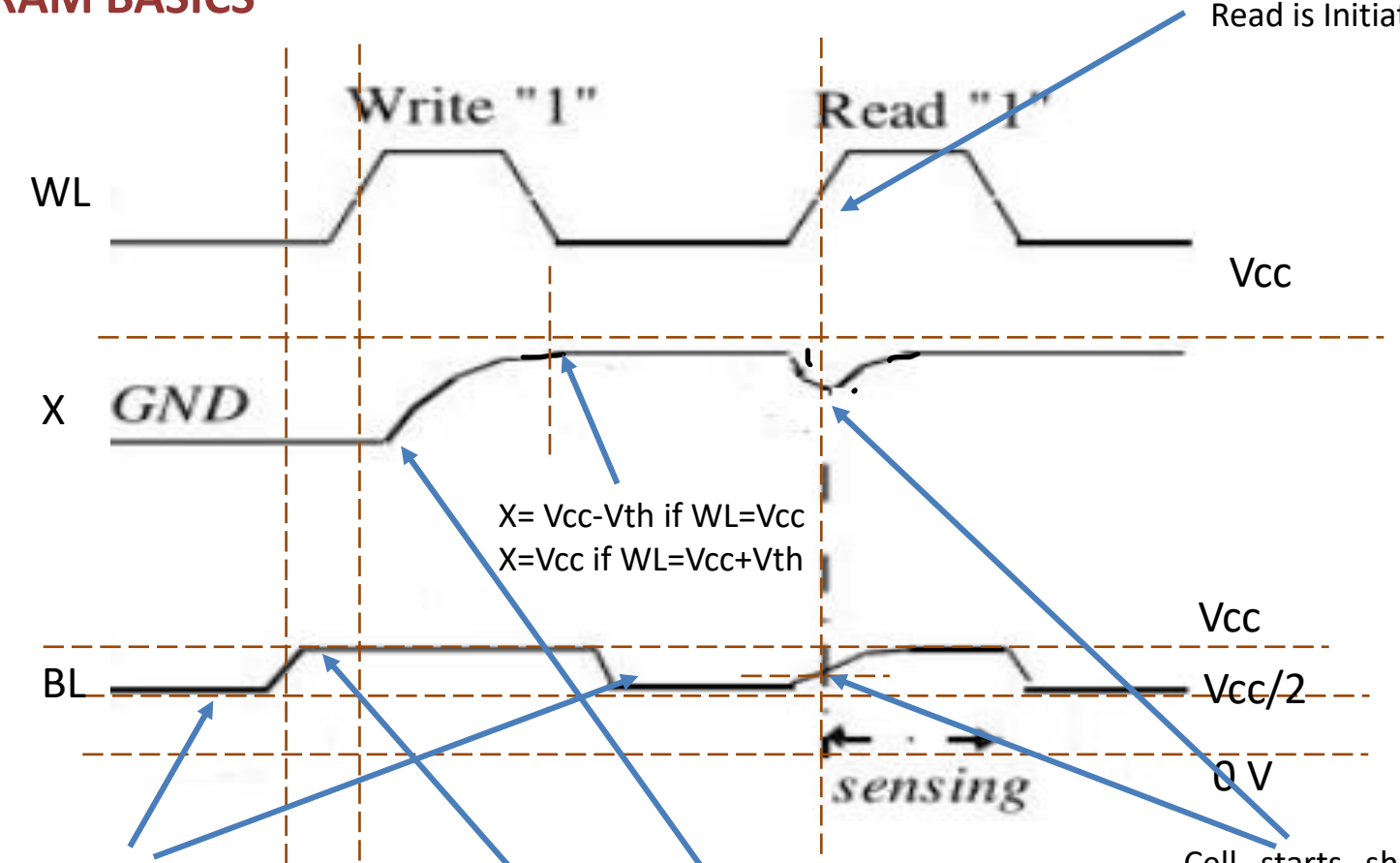
DRAM BASICS

- A modern DRAM memory cell or memory bit (mbit), is formed with one transistor and one capacitor, accordingly referred to as a **1T1C cell**.
- The mbit is capable of **holding binary information** in the form of **stored charge on the capacitor.**
- The **mbit transistor operates as access Transistor** through which mbit capacitor and the digit-line are connected for read or write operation.
- The capacitor's common node is biased at Vcc/2. This allows allows faster operation
- **Storing a logic one in the cell: It requires a capacitor with a voltage of +Vcc/2 across it.** i.e Node X should be charged to Vcc so that Voltage of +Vcc/2 across it.
- Therefore, the charge stored in the mbit capacitor is $Q = +C1.\left(\frac{Vcc}{2}\right)$
- **Storing a logic zero in the cell: It requires a capacitor with a voltage of -Vcc/2 across it.** i.e., Node X should be charged to GND potential, so that Voltage of -Vcc/2 across it.
- Note that the stored charge on the mbit capacitor for a logic zero is $Q = -C1.\left(\frac{Vcc}{2}\right)$



Dynamic RAM (DRAM)

DRAM BASICS



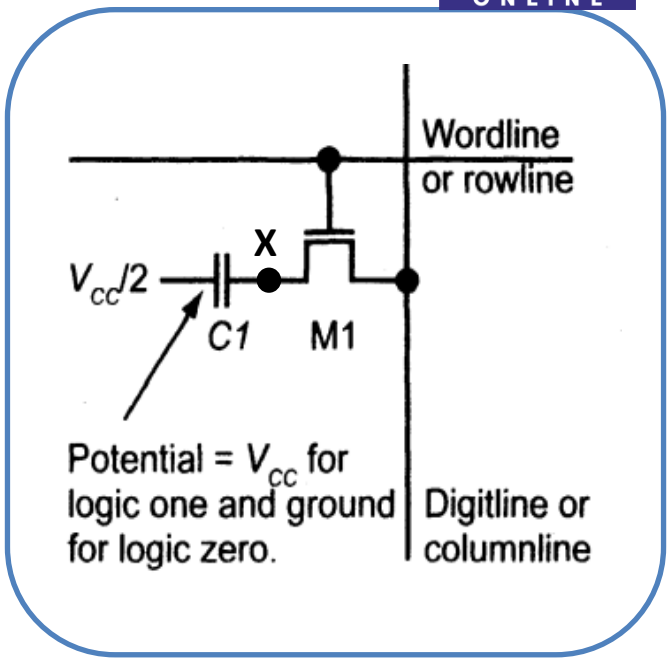
Read is Initiated – Reading '1'

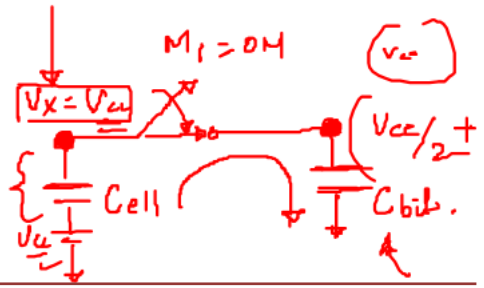
BL pre-charged to Vcc/2

Write '1' - BL driven to Vcc before WL =1

Write '1' is initiated & Cell starts charging increasing Node X potential as Access Transistor is ON

Cell starts sharing the charge with Cbit. Therefore Node X potential starts decreasing and BL starts increasing by delta





Dynamic RAM (DRAM)

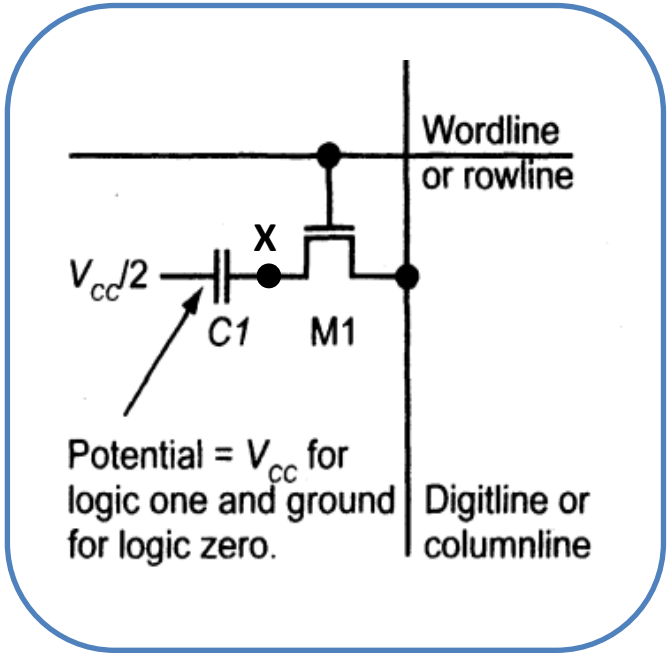
DRAM BASICS

Refresh Operation:

- The various leakage paths cause the stored capacitor charge to slowly deplete.
- To return the stored charge and thereby maintain the stored data state, the cell must be refreshed.
- The required refreshing operation is what makes DRAM memory dynamic rather than static.

Digitline:

- The digitline is a conductive line connected to a multiple of mbit transistors.
- The **conductive line is generally constructed** from either **metal or silicide/polycide polysilicon.**
- Because of the quantity of mbits connected to the digitline and its physical length and proximity to other features, the **digitline is highly capacitive.**
- A typical value for **digitline capacitance on a 50nm process might be 120fF**. Digitline capacitance is an important parameter because it dictates many other aspects of the design.

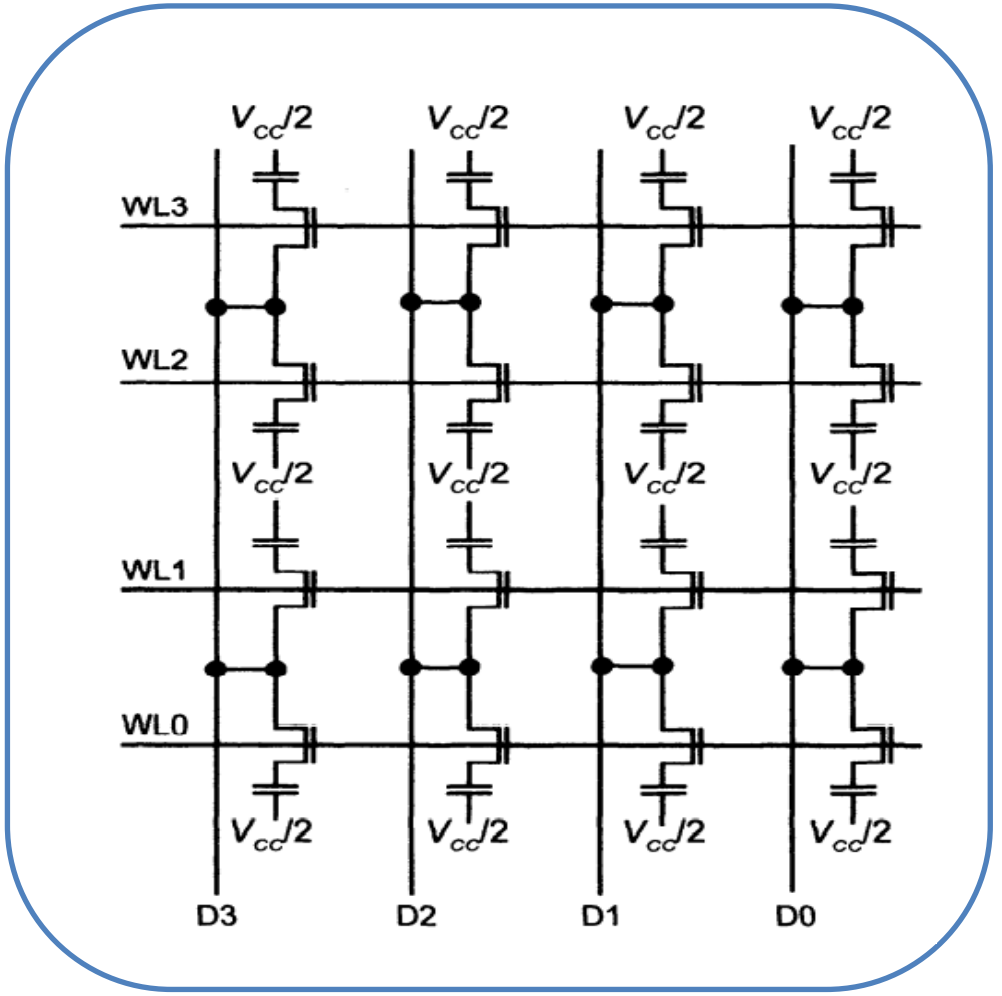


Dynamic RAM (DRAM)

DRAM BASICS

Wordline or Rowline:

- The **mbit transistor gate** terminal is connected to a wordline.
- **The wordline, which is connected to a multiple of mbits**, is actually formed of the same **polysilicon** as that of the transistor gate.
- The wordline is physically orthogonal to the digitline.
- A memory array is formed by tiling a selected quantity of mbits together **such that mbits along a given digitline do not share a common wordline** and mbits along a common wordline do not share a common digitline.

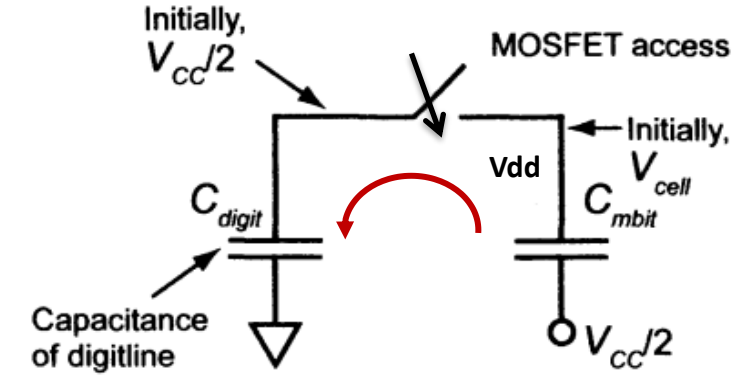


Dynamic RAM (DRAM)

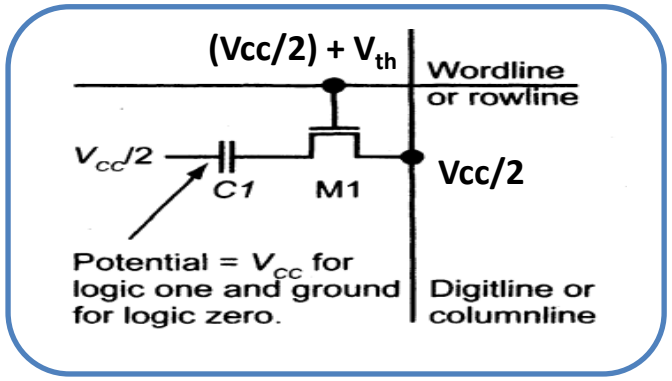
Access Operations

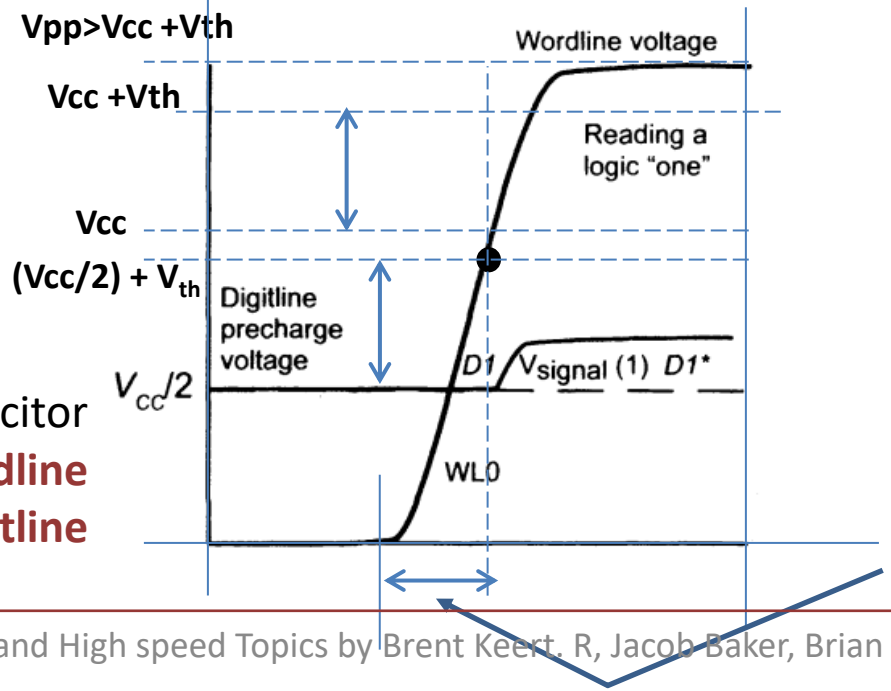
Read mbit1

- The wordline WL0 changes to $V_{pp}=V_{CC}+V_{th}$ to ensure that a full **logic one** value **can be written back into the mbit capacitor.**
- The mbit capacitor begins to discharge onto the digitline at two different voltage levels depending on the logic level stored in the cell.



Case1: **Assume , Stored Data = Logic 1.**
Vgs=Vth at WL=(Vcc/2) + V_{th} MOS Device turns on and start charging the Digiline





If a logic one was stored, the capacitor begins to **discharge when the wordline voltage exceeds the digitline PRECHARGE voltage by Vth.**

Because of the finite rise time of the wordline voltage, **this difference in turn-on voltage translates into a significant delay when reading ones.**

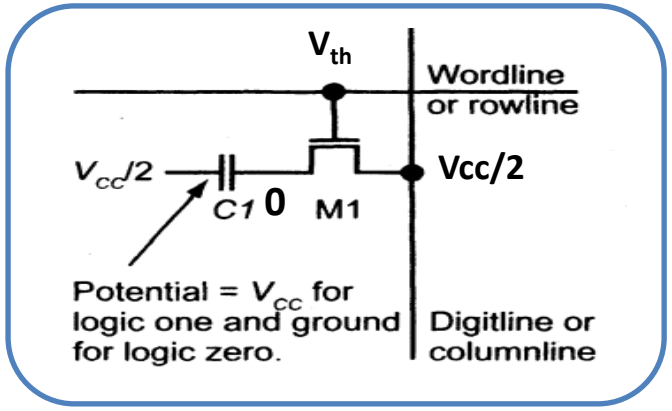
Dynamic RAM (DRAM)

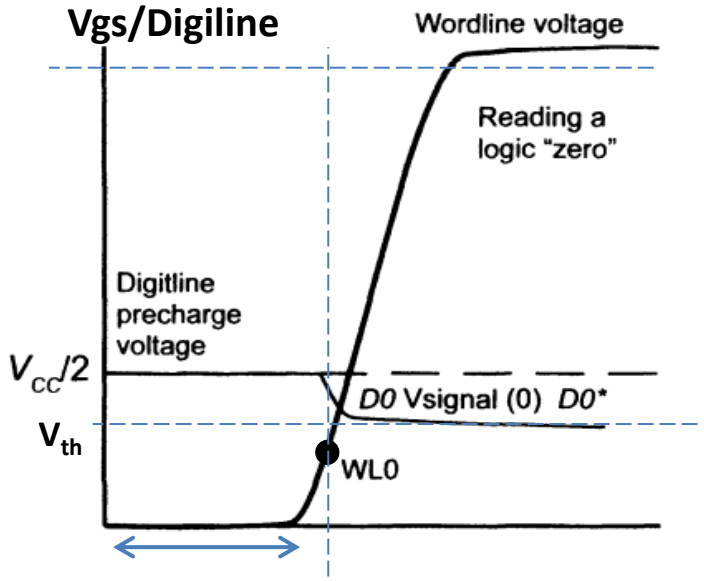
Access Operations : Reading mbit

- The wordline WL0 changes to $V_{pp}=V_{DD}+V_{th}$ to ensure that a full **logic one** value **can be written back into the mbit capacitor.**
- The mbit capacitor begins to discharge onto the digitline at two different voltage levels depending on the logic level stored in the cell.

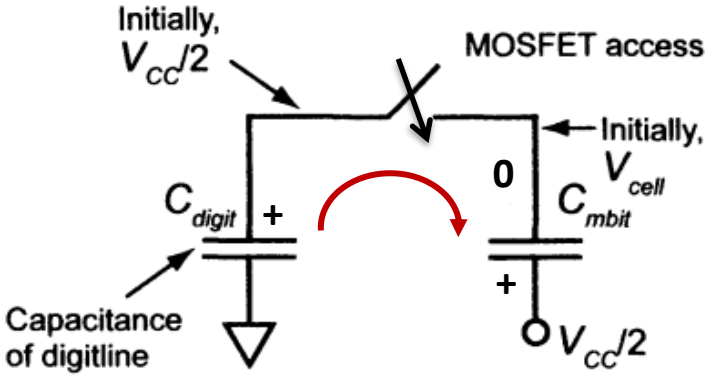
Case2: Assume mbit is holding Logic 0

Vgs=Vth at WL=V_{th} MOS Device turns on and start discharging the Digiline





For a logic zero, the capacitor begins to discharge when the word line voltage exceeds Vth·



Dynamic RAM (DRAM)

Access Operations : Reading mbit

Now it is clear that dataline (BL) is either getting charged or discharged by a factor ΔV.

The factor ΔV can be derived as follType equation here.ows

Dataline Voltage

$$\text{Vdigit} = \left(\frac{\text{Qdigit}}{\text{Cdigit}}\right) \quad ----------(1)$$

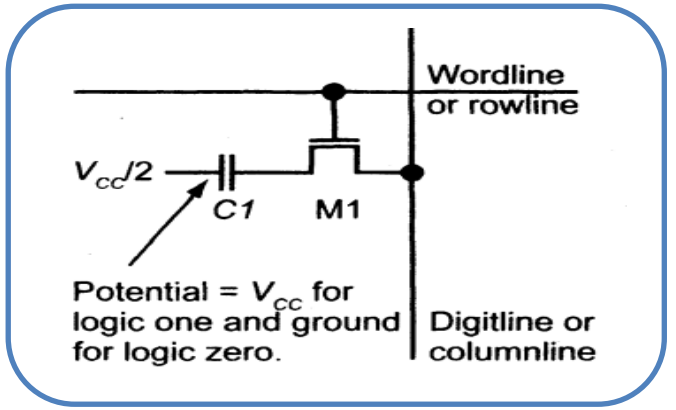
Initial Voltage across cell

$$\text{Vcell} = \left(\frac{\text{Qcell}}{\text{Ccell}}\right) \quad ----------(2)$$

Common Voltage after charge sharing

$$\text{Vf} = \left(\frac{\text{Total Charge}}{\text{Total Capacitance}}\right) = \left(\frac{\text{Qdigit} + \text{Qcell}}{\text{Cdigit} + \text{Cell}}\right) \quad ---(3)$$

$$\Delta\text{V} = \text{Vdigit} - \text{Vf} \quad ----------(4)$$



Dynamic RAM (DRAM)

Access Operations : Reading mbit

$$\Delta V = \left(\frac{Qdigit}{Cdigit}\right) - \left(\frac{Qdigit + Qcell}{Cdigit + Cell}\right)$$

By Simplification

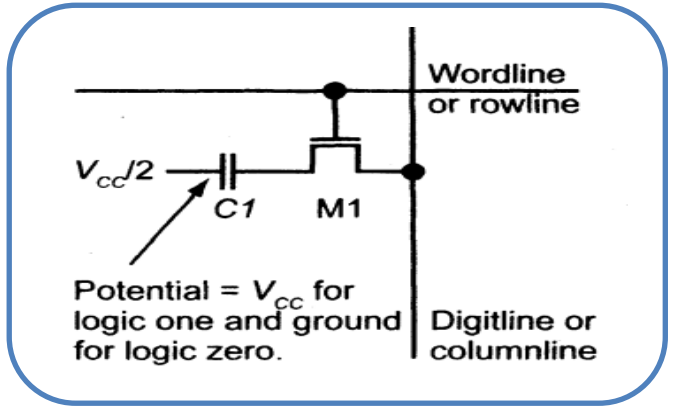
$$\Delta V = \left(\frac{Qdigit\,.\,(Cdigit + Cell) - (Qdigit + Qcell).\,Cdigit}{Cdigit\,.\,(Cdigit + Cell)}\right)$$

Dividing by Nr and Dr by Cdigit on RHS

$$\Delta V = \left(\frac{Qdigit\,.\,(1 + Cell/Cdigit) - (Qdigit + Qcell)}{(Cdigit + Cell)}\right)$$

$$\Delta V = \left(\frac{Qdigit + Qdigit.\frac{Cell}{Cdigit} - Qdigit - Qcell}{(Cdigit + Cell)}\right)$$

$$\Delta V = \left(\frac{Qdigit.\frac{Cell}{Cdigit} - Qcell}{(Cdigit + Cell)}\right) \quad ----------(4)$$



Dynamic RAM (DRAM)

Access Operations : Reading mbit

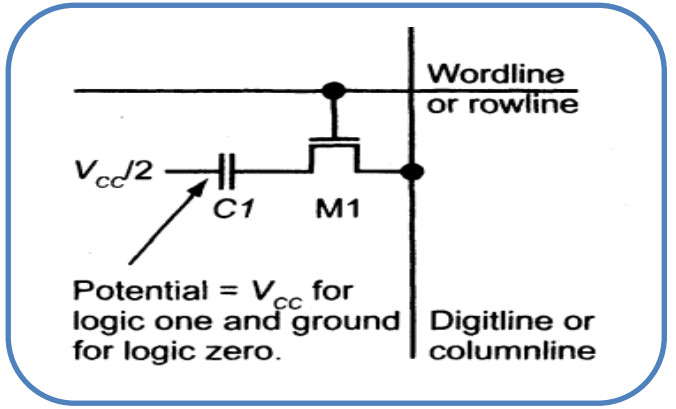
$$\Delta V = \left(\frac{\text{Qdigit}.\frac{\text{Cell}}{\text{Cdigit}} - \text{Qcell}}{(\text{Cdigit} + \text{Cell})} \right)$$

By taking Ccell common in Nr

$$\Delta V = \left(\frac{Ccell \ \left[\left(\frac{Qdigit}{Cdigit} \right) - \left(\frac{Qcell}{Ccell} \right) \right]}{(\text{Cdigit} + \text{Cell})} \right)$$

$$\Delta V = \left(\frac{Ccell \ [Vdigit - Vcell]}{(\text{Cdigit} + \text{Cell})} \right) \text{----------} -(5)$$

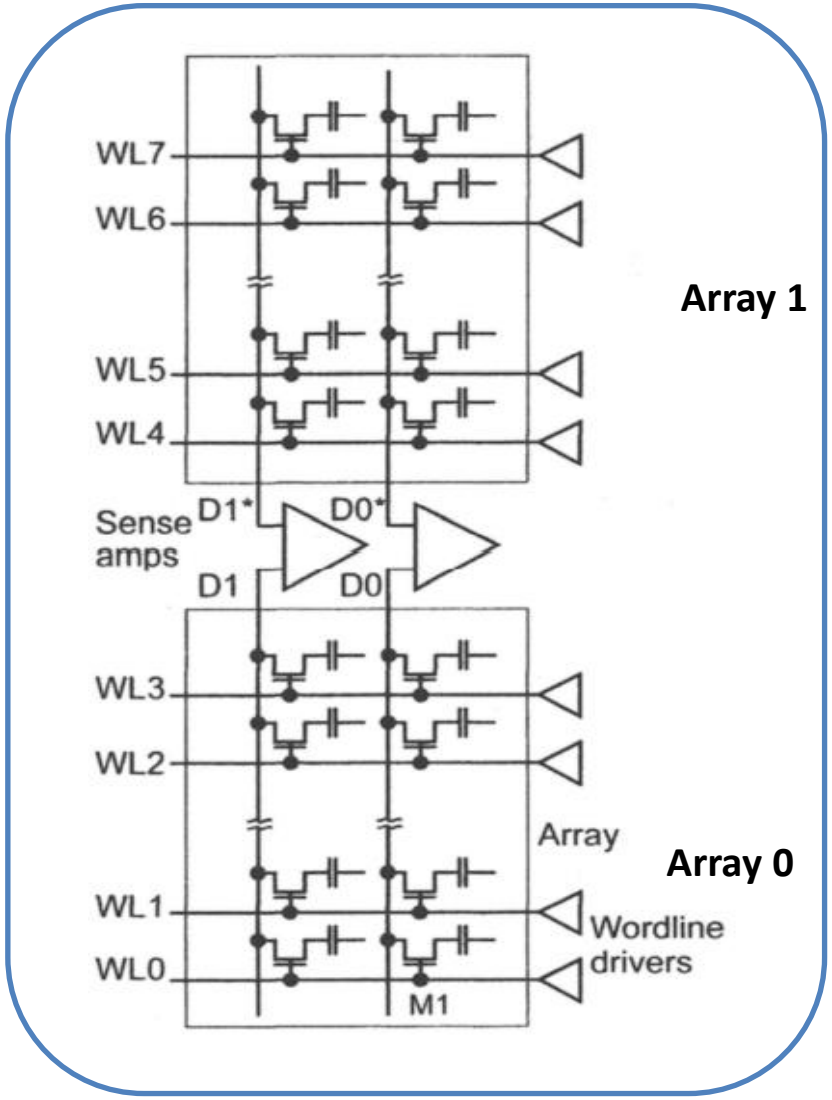
Where, $[Vdigit - Vcell]$ is the Initial difference.



Dynamic RAM (DRAM)

Access Operations

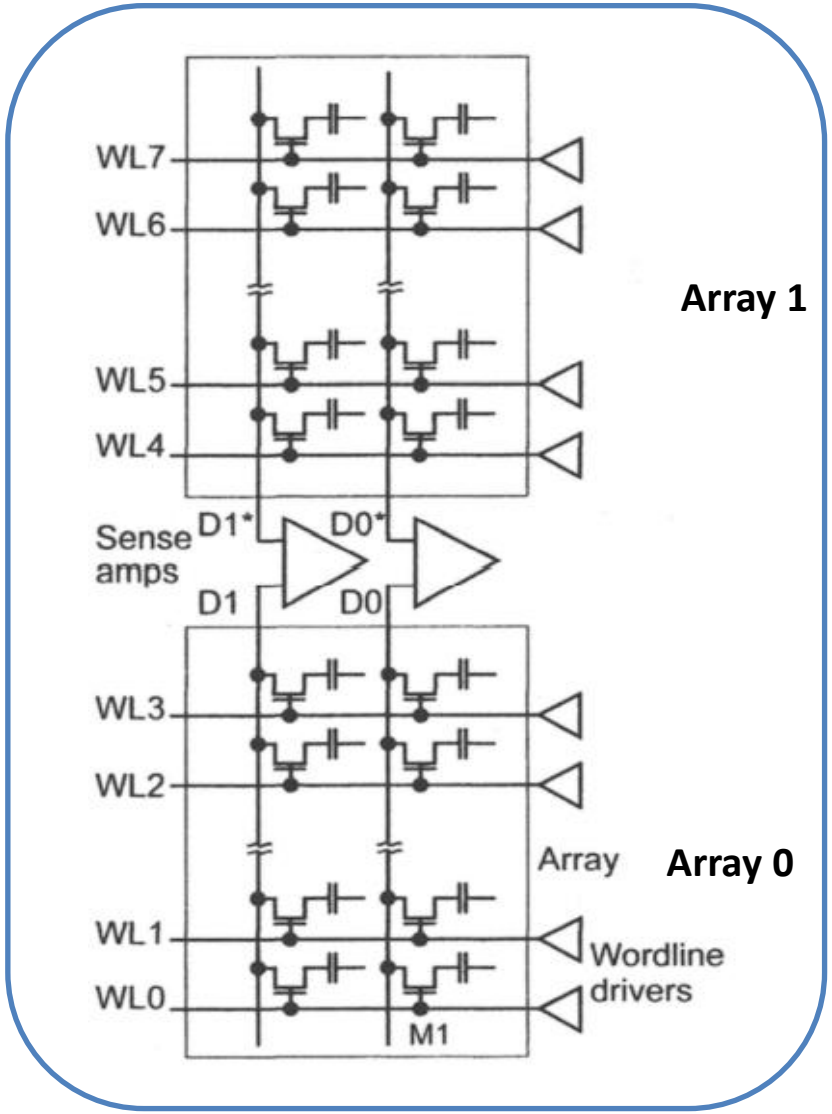
- Assume that the cells connected to D1, have logic one levels (+Vcc/2) stored on them and that the cells connected to D0 have logic zero levels (-Vcc/2) stored on them.
- Next, we form a digitline pair by considering two digitlines from adjacent arrays. The digitline pairs, labeled D0/D0* and D1/D1*, are initially equilibrated (precharged) to Vcc/2 V.
- All wordlines are initially at 0V, ensuring that the mbit transistors are OFF.
- Prior to a wordline firing, the digitlines are electrically disconnected from the Vcc/2 bias voltage and allowed to float.
- They remain at the Vcc/2 PRECHARGE voltage due to their capacitance.



Dynamic RAM (DRAM)

Access Operations

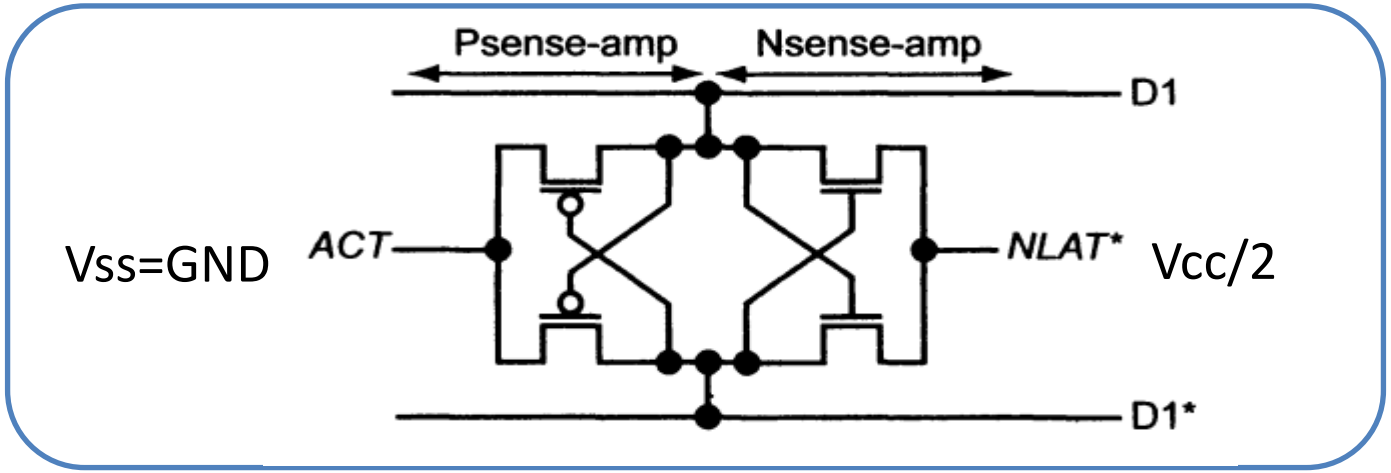
- Accessing a DRAM cell results in charge sharing between the mbit capacitor and the digitline capacitance.
- This charge sharing causes the **digitline voltage either to increase for a stored logic one** or **to decrease for a stored logic zero**.
- Ideally, only the digitline connected to the accessed mbit will change. In reality, the other digitline voltage also changes slightly, due to parasitic coupling between between the firing wordline and the other digitline.
- A differential voltage develops between the two digitlines.
- The magnitude of this voltage difference, is a function of the mbit capacitance (Cmbit), digitline capacitance (Cdigit), and voltage stored on the cell prior to access (Vcell)



Dynamic RAM (DRAM)

Sense Operations

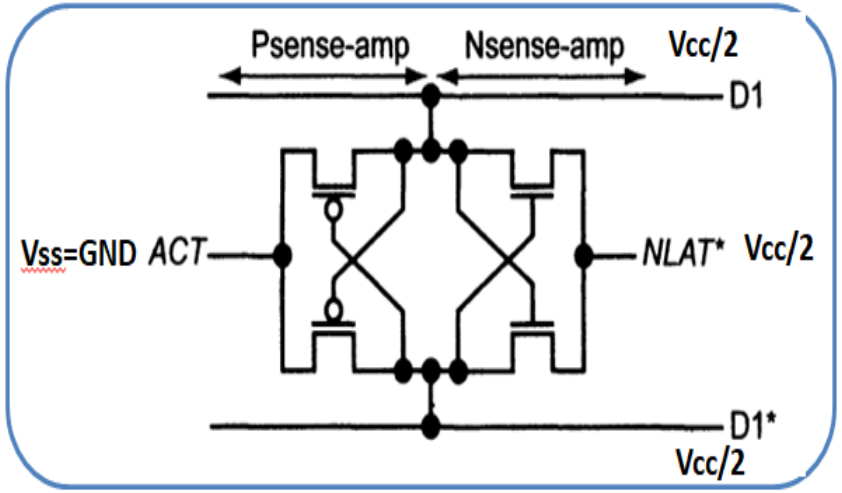
- Sensing is essentially the amplification of the digitline signal or the differential voltage between the digitlines. Sensing is necessary to properly read the cell data and refresh the mbit cells.
- A cross-coupled NMOS pair and a cross-coupled PMOS pair i.e., a pair of cross-coupled CMOS inverters in which ACT and NLAT* provide power and ground.
- The NMOS pair or Nsense-amp has a common node labeled NLAT* (for Nsense-amp latch).
- Similarly, the Psense-amp has a common node labeled ACT (for Active pull-up). **Initially, NLAT* is biased to Vcc/2, and ACT is biased to Vss or signal ground.**



Dynamic RAM (DRAM)

Sense Operations

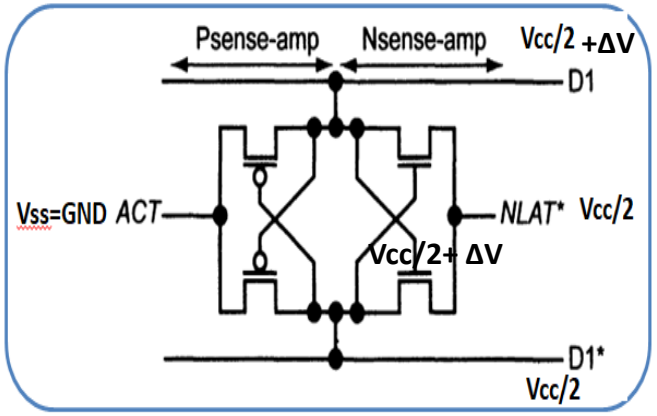
- **Initially, NLAT* is biased to Vcc/2, and ACT is biased to Vss or signal ground.**
- Because the **digitline pair DI and DI* are both initially at Vcc/2,** the Nsense-amp transistors are both OFF. Similarly, both Psenseamp transistors are OFF.
- Again, when the **mbit is accessed**, a signal develops across the digitline pair. While **one digitline contains charge from the cell access,** the other digitline does not but serves as a reference for the Sensing operation.
- The sense amplifiers are generally fired sequentially: the Nsense-amp first, then the Psense-amp.

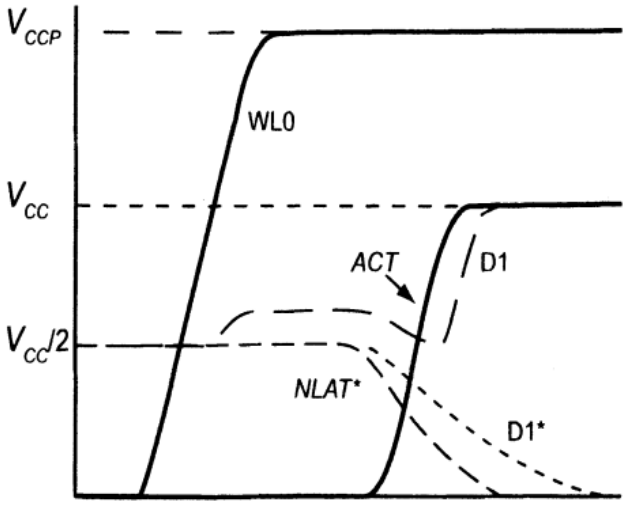


Dynamic RAM (DRAM)

Sense Operations

- The Nsense-amp is fired by bringing NLAT* (Nsense-amp latch) toward ground.
- As the voltage difference between NLAT* and the digitlines (D 1 and D 1 * in) approaches VTH, **the NMOS transistor, whose gate is connected to the higher voltage digitline, begins to conduct.**
- This conduction **occurs first in the sub-threshold and then in the saturation region as the gate-to-source voltage exceeds Vth** and causes the low-voltage digitline to discharge toward the NLAT* voltage.
- Ultimately, **NLAT* will reach ground and the digitline will be brought to ground potential.** Note that the other NMOS transistor will not conduct: its gate voltage is derived from the low-voltage digitline, which is being discharged toward ground.
- In reality, parasitic coupling between digitlines and limited subthreshold





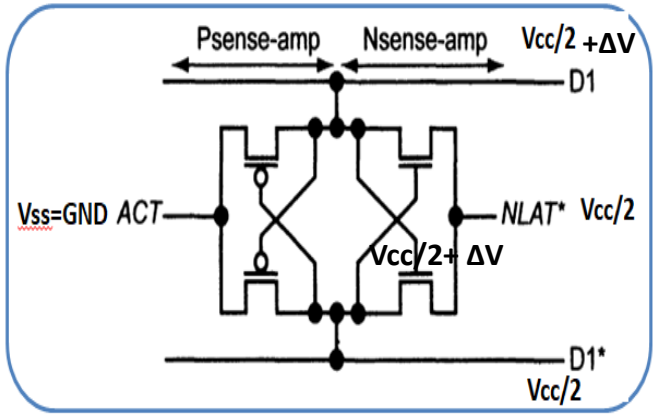
Dynamic RAM (DRAM)

Sense Operations

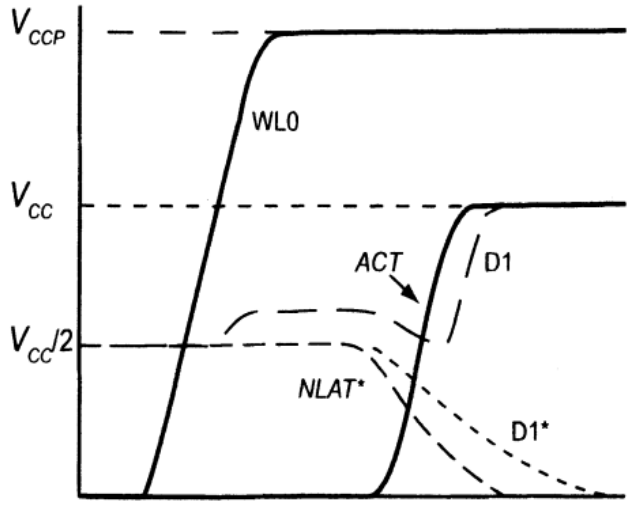
- In reality, parasitic coupling between digitlines and limited subthreshold conduction by the second transistor result in a temporary voltage drop on the high digitline

  Initially Vgs,pd = Vg –Vs
  = (Vcc/2+ ΔV) -NLAT
  = (Vcc/2+ΔV) - Vcc/2 ↓

  Initially Vgs,pu = Vg –Vs
  = (Vcc/2) -ACT
  = (Vcc/2) - GND ↑



- As NLAT is fired and going towards ground. This makes Vgs,pd >Vth and turns on NMOS and D1* goes towards zero. On the other side as ACT is increased towards Vdd , At a point when Vgs,pu>|Vtp| pmos conducts and D1 charges to Vcc



Dynamic RAM (DRAM)

Sense Operations

- Ultimately, **NLAT* will reach ground and the digitline will be brought to ground potential.** Note that the other NMOS transistor will not conduct: its gate voltage is derived from the low-voltage digitline, which is being discharged toward ground.

  Initially

  Vgs,pd = Vg –Vs

  = (Vcc/2+ ΔV) -NLAT ↓

  = (Vcc/2+ΔV) - Vcc/2

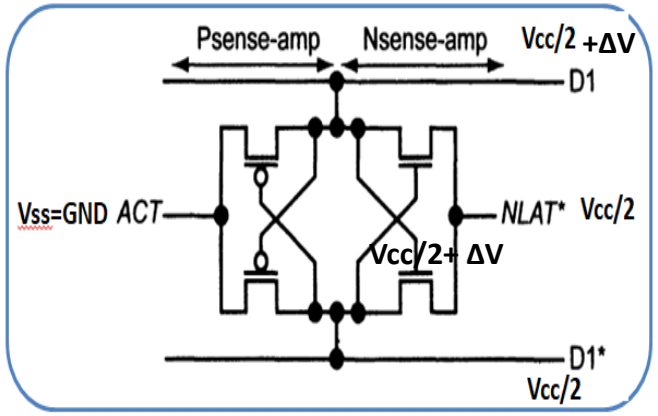
  Initially

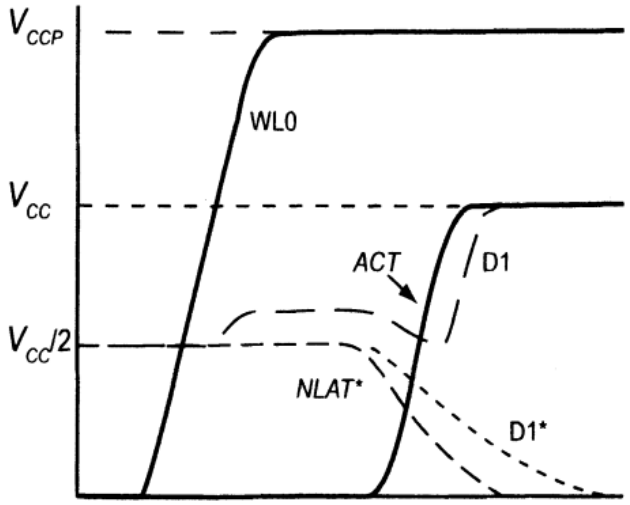
  Vgs,pu = Vg –Vs

  = (Vcc/2) -ACT ↑

  = (Vcc/2) - GND

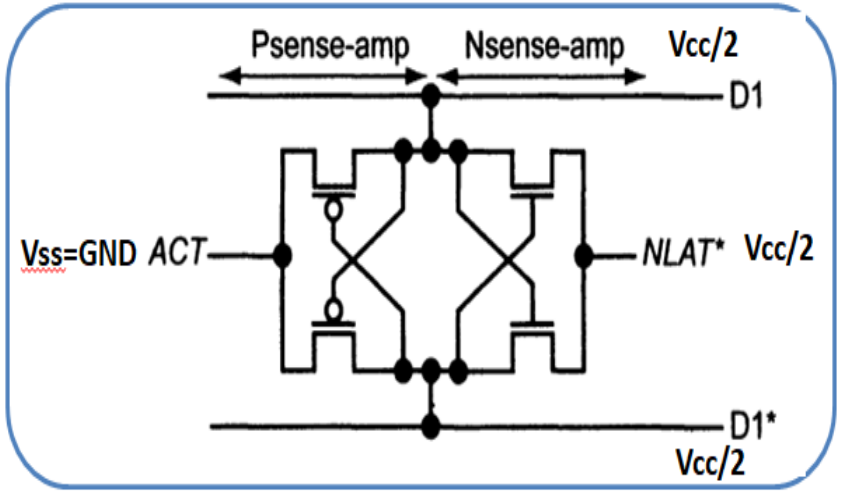
- As NLAT is fired and going towards ground. This makes Vgs,pd >Vth and turns on NMOS and D1* goes towards zero. On the other side as ACT is increased towards Vdd , At a point when Vgs,pu>|Vtp| pmos conducts and D1 charges to Vcc





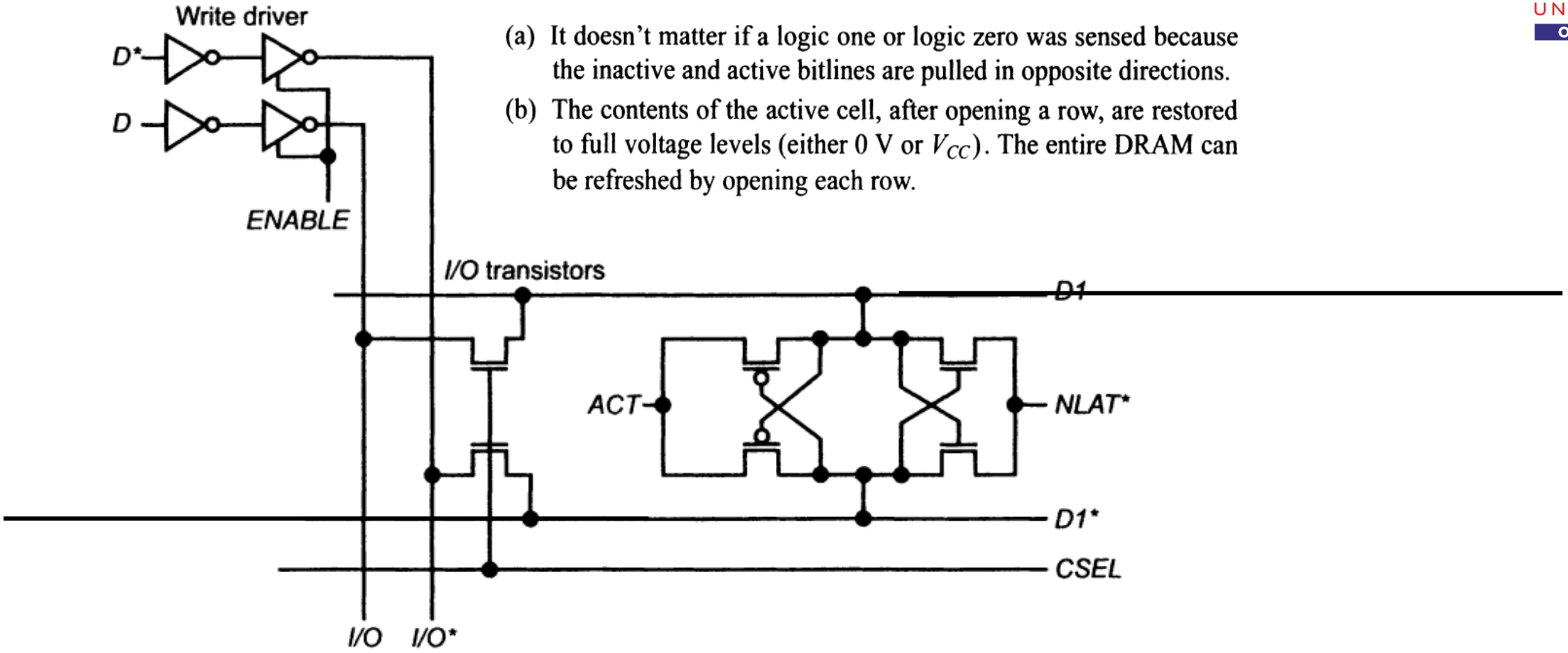
DRAM – Write Operation

- **Initially, NLAT* is biased to Vcc/2, and ACT is biased to Vss or signal ground.**
- Because the **digitline pair DI and DI* are both initially at Vcc/2,** the Nsense-amp transistors are both OFF. Similarly, both Psenseamp transistors are OFF.
- Again, when the **mbit is accessed**, a signal develops across the digitline pair. While **one digitline contains charge from the cell access,** the other digitline does not but serves as a reference for the Sensing operation.
- The sense amplifiers are generally fired sequentially: the Nsense-amp first, then the Psense-amp.



DRAM – Write Operation

Sense Amplifier with IO Devices

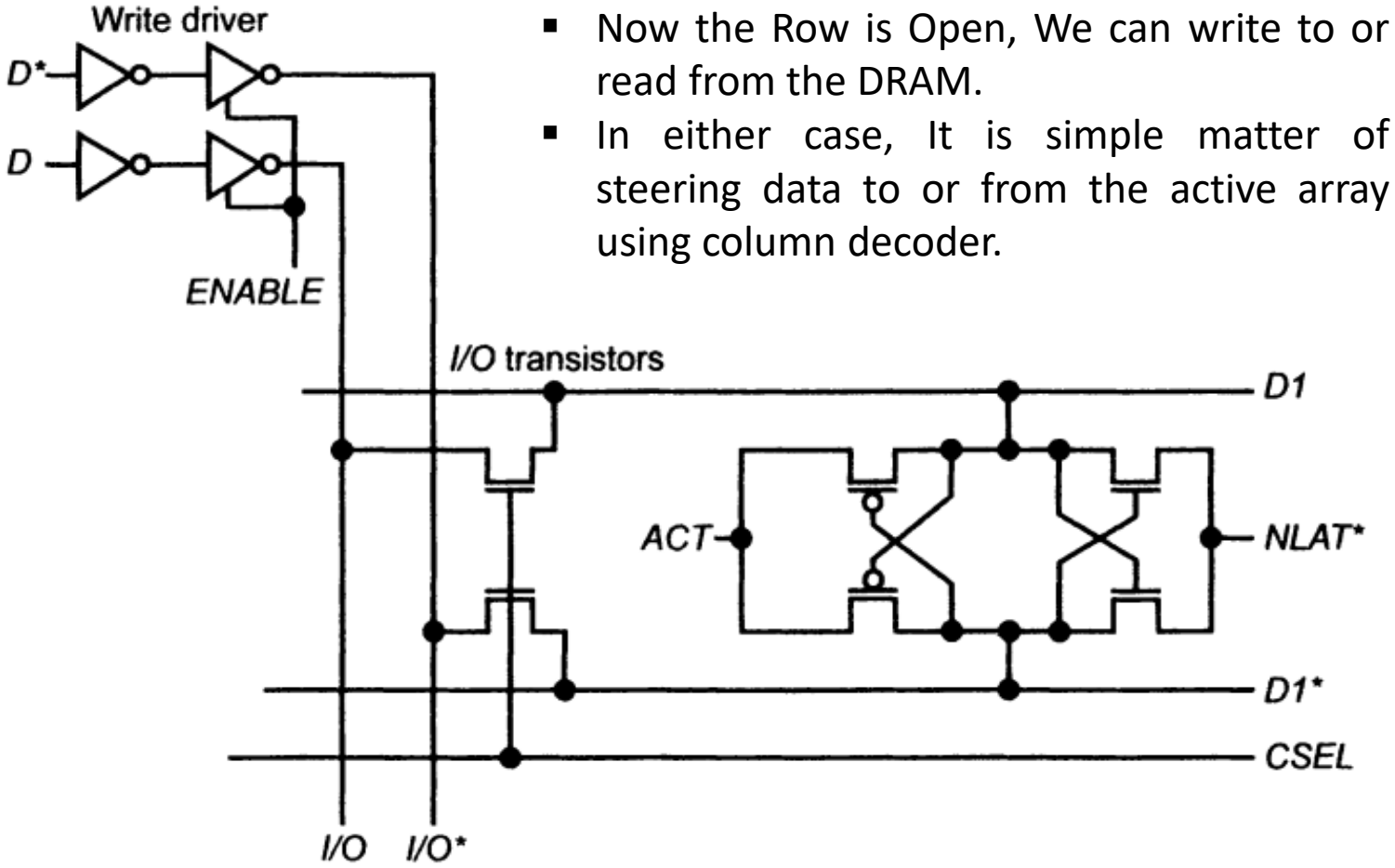


(a) It doesn't matter if a logic one or logic zero was sensed because the inactive and active bitlines are pulled in opposite directions.

(b) The contents of the active cell, after opening a row, are restored to full voltage levels (either 0 V or V_{CC}). The entire DRAM can be refreshed by opening each row.

DRAM – Write Operation

Sense Amplifier with IO Devices



- Now the Row is Open, We can write to or read from the DRAM.
- In either case, It is simple matter of steering data to or from the active array using column decoder.

- **If it is writing to the array** – the buffers (write driver circuit sets the new logic levels on bitlines through I/O transistors.
- The row is still open because, the wordline remains HIGH (WL=Vccp). The row remains open as long as RAS'=LOW

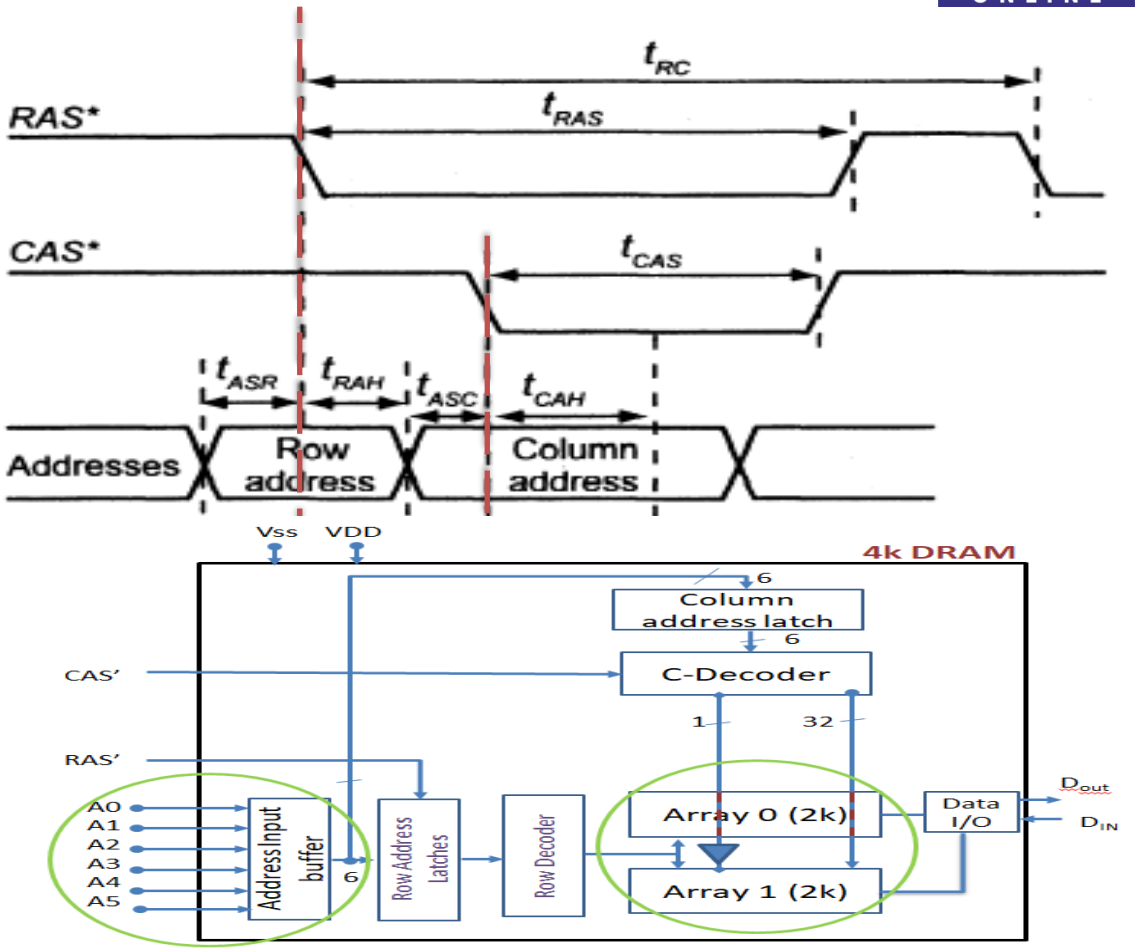
Dynamic RAM (DRAM)

Second Generation of DRAM

Key Features of 2nd generation are

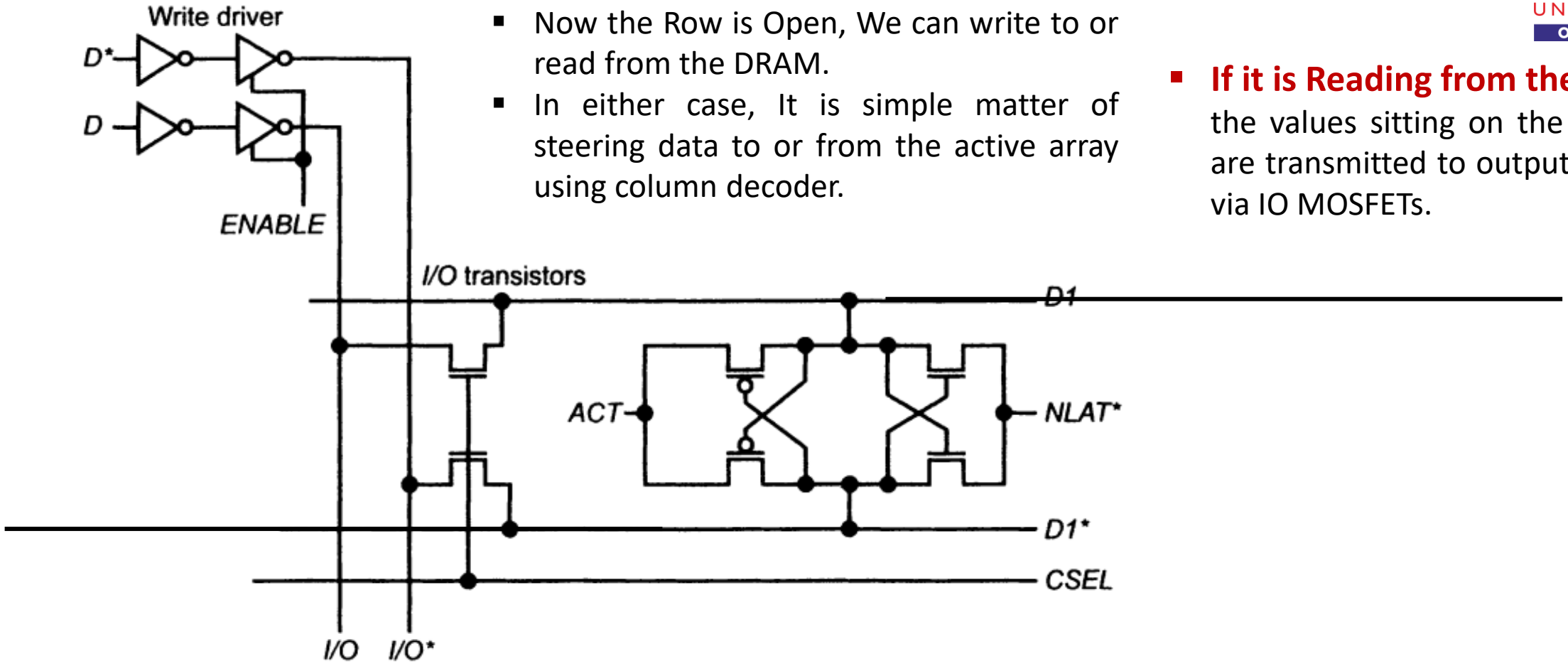
- The Second generation of DRAM introduced
  - **1 Transistor / 1 Capacitor Memory Cell**
  - **Multiplexed address inputs**:
    - The **Row Address Strobe (RAS')** input clocks **the address on [A5:A0] into row address latch on it's falling edge.**
    - The **Column Address Strobe (CAS')** input clocks the **address on [A5:A0] into Column address latch on it's falling edge.**
    - The parameters t_{ASR}, t_{RAH}, t_{ASC}, and t_{CAH} indicate the **setup and hold times for the row and column addresses**, respectively.

4k DRAM



DRAM – Write Operation

Sense Amplifier with IO Devices

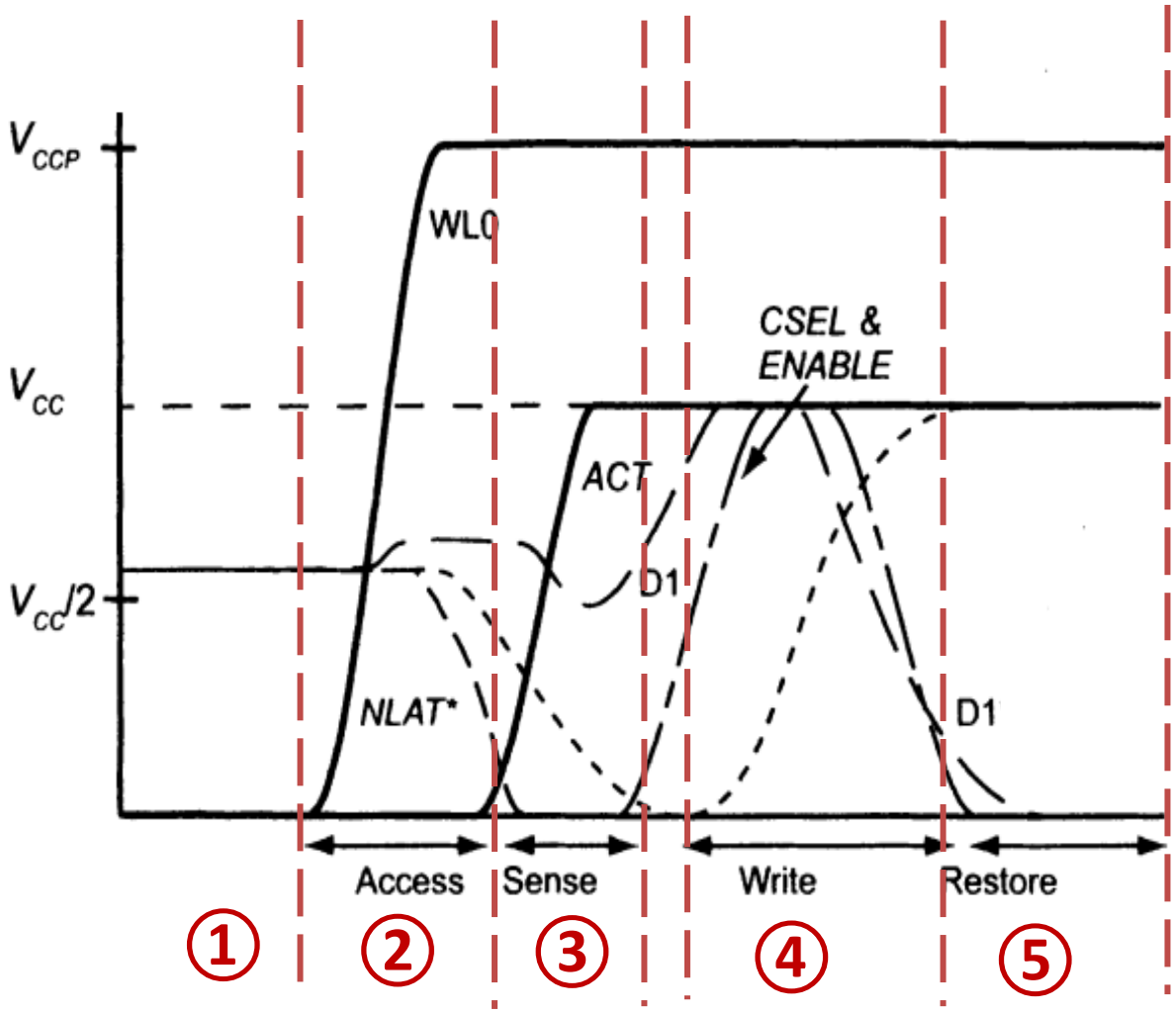


- Now the Row is Open, We can write to or read from the DRAM.
- In either case, It is simple matter of steering data to or from the active array using column decoder.

- **If it is Reading from the array:** the values sitting on the bitlines are transmitted to output buffers via IO MOSFETs.

DRAM – Write Operation

Sense Amplifier with IO Devices



Signals: WL0,D1,D1*,ACT,NLAT*, **Enable and CSEL**

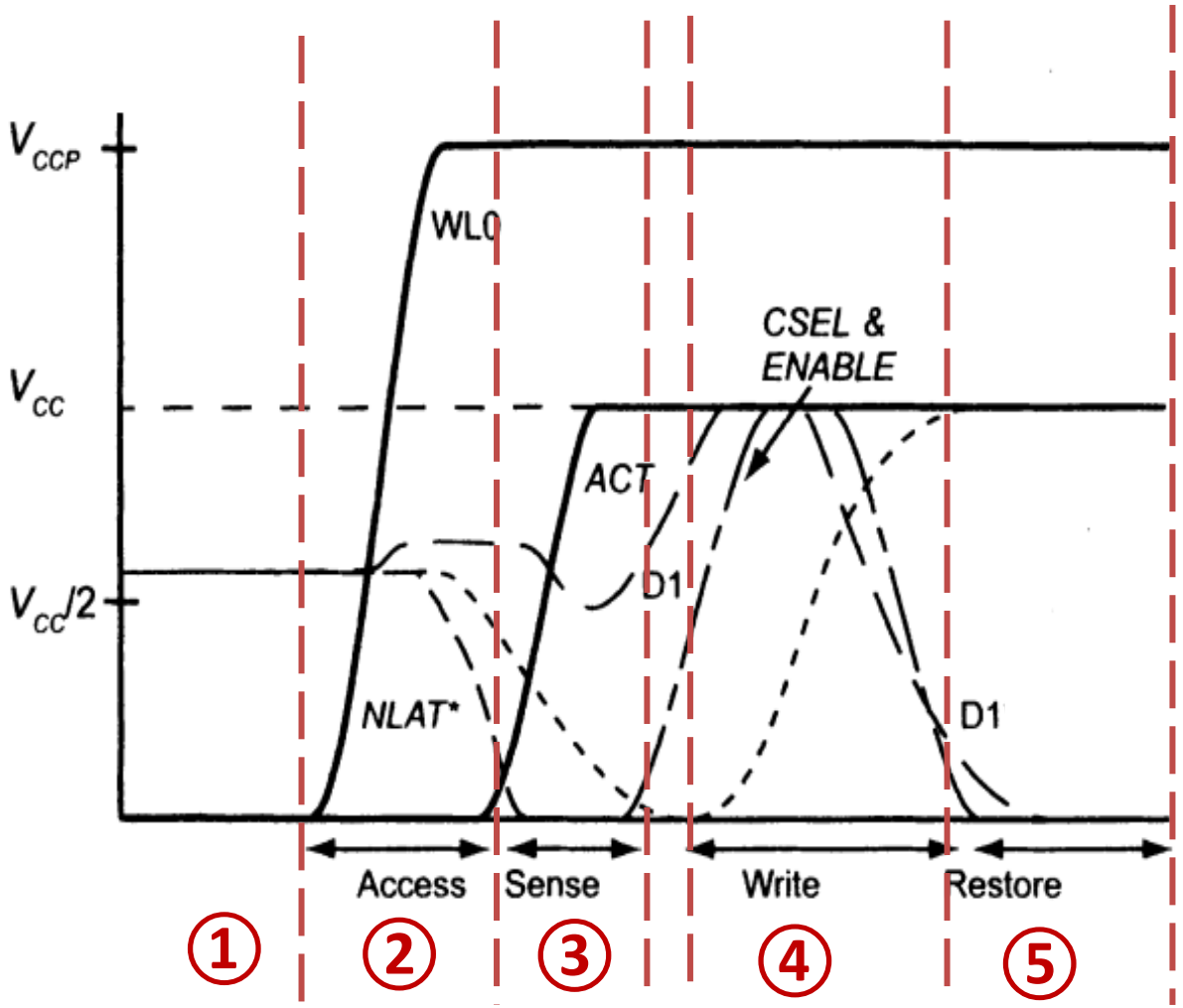
① - Row Not Selected : RAS* = HIGH

Under this state,

- Row is Inactive -WL =LOW,
- Sense Amplifier is disabled (ACT=GND and NLAT*= Vcc/2)
- Pre-charge is ON (D1 and D1* are pre-charged to Vcc/2)
- Write Driver and Column Select is Low (CSEL=0)

DRAM – Write Operation

Sense Amplifier with IO Devices



Signals: WL0,D1,D1*,ACT,NLAT*, **Enable and CSEL**

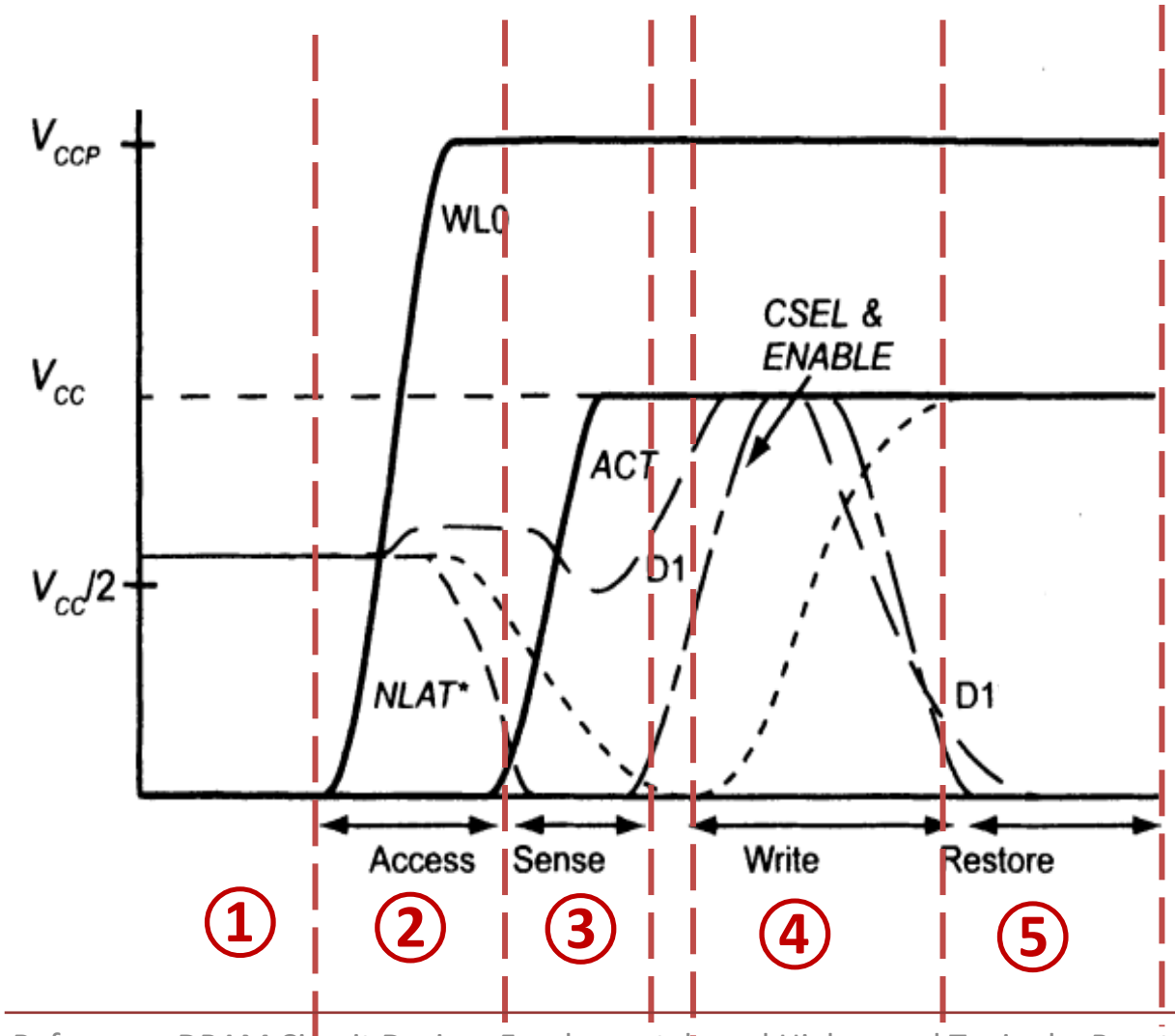
② **Access: Row is Selected : RAS* = HIGH – Low transition will enable the Row Decoder. It decodes the address and selects the Row by making WL=HIGH (Vccp)**

Under this state,

- Row is active -WL =HIGH,
- Pre-charge is OFF (D1 and D1* are already in pre-charged state)
- When WL=(Vcc/2 + Vtn) access transistor turns ON and charge sharing begins.
- Intially SA is OFF and Once sufficient charge is shared, SA is fired with Nsense first followed by Psense
- Write Driver and Column Select is Low.

DRAM – Write Operation

Sense Amplifier with IO Devices



Signals: WL0,D1,D1*,ACT,NLAT*, **Enable and CSEL**

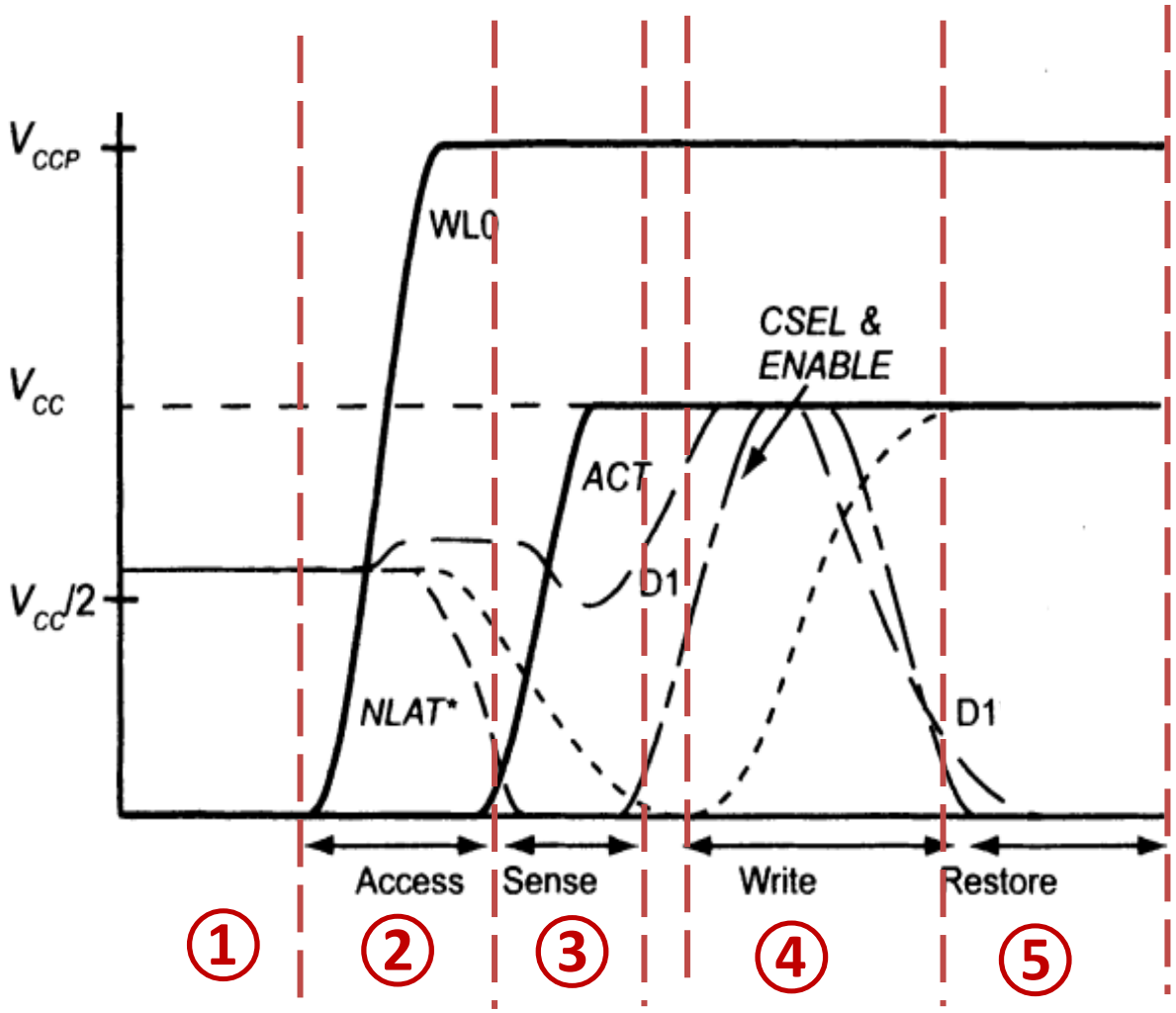
③ Sense: Row is Selected : RAS* = Low & CAS* =HIGH

Under this state,

- Row is active -WL =HIGH,
- Pre-charge is OFF
- When WL=(Vcc/2 + Vtn) access transistor turns ON and charge sharing begins.
- Initially SA is ON and Due to Firing Nsense first followed by Psense. The D1 is pulled down to GND potential and D1 is pulled up to Vdd.
- Write Driver and Column Select is Low (CSEL=0)

DRAM – Write Operation

Sense Amplifier with IO Devices



Signals: WL0,D1,D1*,ACT,NLAT*, **Enable and CSEL**

④ **Write: Row is Selected : RAS* = Low & CAS* =HIGH to Low.**

Under this state,

- Row is active -WL =HIGH, Pre-charge is OFF
- Now Enable =HIGH which enables the write Driver and CSEL (**output of Column Decoder) enables corresponding I/O MOSFETs. The new data placed by Write driver is taken as New data by Sense Amplifier and restores the bit lines rail to rail voltage.**
- **In waveform we can see that due to access D1=HIGH after Sensing, but due to write operation (Writing 0) , the D1 after Writing operation**

DRAM – Opening a Row (Summary)

Opening a Row is a fundamental operation for both reading and writing to DRAM array.

RAS* Timing chain

1. Initially, both *RAS** and *CAS** are HIGH. All bitlines in the DRAM are driven to $V_{CC}/2$, while all wordlines are at 0 V. This ensures that all of the mbit's access transistors in the DRAM are OFF.
2. A valid row address is applied to the DRAM and *RAS** goes LOW. While the row address is being latched, on the falling edge of *RAS**, and decoded, the bitlines are disconnected from the $V_{CC}/2$ bias and allowed to float. The bitlines at this point are charged to $V_{CC}/2$, and they can be thought of as capacitors.
3. The row address is decoded and applied to the wordline drivers. This forces only one rowline in at least one memory array to V_{CCP}. Driving the wordline to V_{CCP} turns ON the mbits attached to this rowline and causes charge sharing between the mbit capacitance and the capacitance of the corresponding bitline. The result is a small perturbation (upwards for a logic one and downwards for a logic zero) in the bitline voltages.

DRAM – Opening a Row (Summary)

Opening a Row is a fundamental operation for both reading and writing to DRAM array.

RAS* Timing chain

4. The next operation is Sensing, which has two purposes: a) to determine if a logic one or zero was written to the cell and b) to refresh the contents of the cell by restoring a full logic zero (0 V) or one (V_{CC}) to the capacitor. Following the wordlines going HIGH, the Nsense-amp is fired by driving, via an n-channel MOSFET, *NLAT** to ground. The inputs to the sense amplifier are two bitlines: the bitline we are sensing and the bitline that is not active (a bitline that is still charged to $V_{CC}/2$—an inactive bitline). Pulling *NLAT** to ground results in one of the bitlines going to ground. Next, the *ACT* signal is pulled up to V_{CC}, driving the other bitline to V_{CC}. Some important notes:

   (a) It doesn't matter if a logic one or logic zero was sensed because the inactive and active bitlines are pulled in opposite directions.

   (b) The contents of the active cell, after opening a row, are restored to full voltage levels (either 0 V or V_{CC}). The entire DRAM can be refreshed by opening each row.

DRAM – Opening a Row (Summary)

Opening a Row is a fundamental operation for both reading and writing to DRAM array.

RAS* Timing chain

Now that the row is open, we can write to or read from the DRAM. In either case, it is a simple matter of steering data to or from the active array(s) using the column decoder. When writing to the array, buffers set the new logic voltage levels on the bitlines. The row is still open because the wordline remains HIGH. (The row stays open as long as *RAS** is LOW.)

DRAM Array Architecture

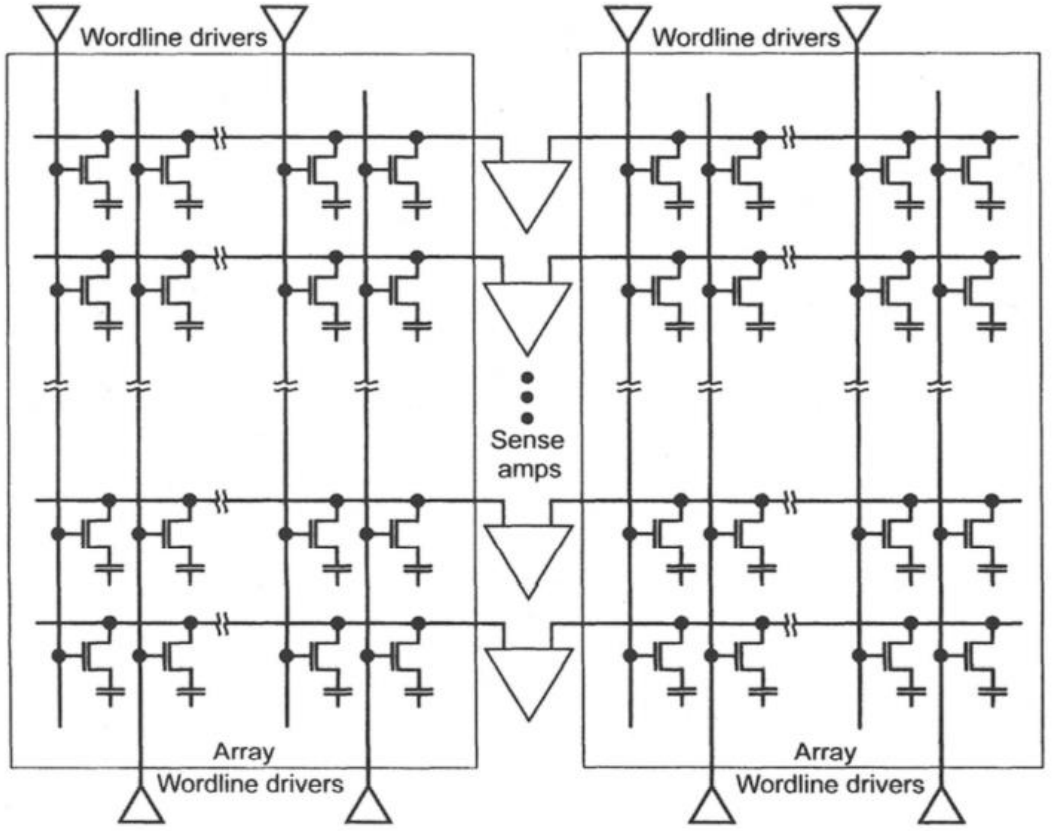
There are two basic types of DRAM Array

1. Open digitline Array Architecture
2. Folded Array Architecture

1. Open digitline Array Architecture

Features:

- **It was used in DRAM early generation where the size was limited to 64K bits**
- the sense amplifier circuits between two sets of arrays.
- The **true and complement digitlines (D and D*) connected to each sense amplifier pair** come from separate arrays (adjacent pair of arrays).
- This arrangement precludes **(prevents) using digitline twisting to improve signal-to-noise** performance, which is the prevalent reason why the industry switched to folded arrays.
- Each wordline in an open digitline architecture connects to mbit transistors on every digitline, creating crosspoint-style arrays.



DRAM Array Architecture -Open digitline

- In these layouts, mbits are paired to share a common contact to the digitline, which reduces the array size by eliminating duplication.

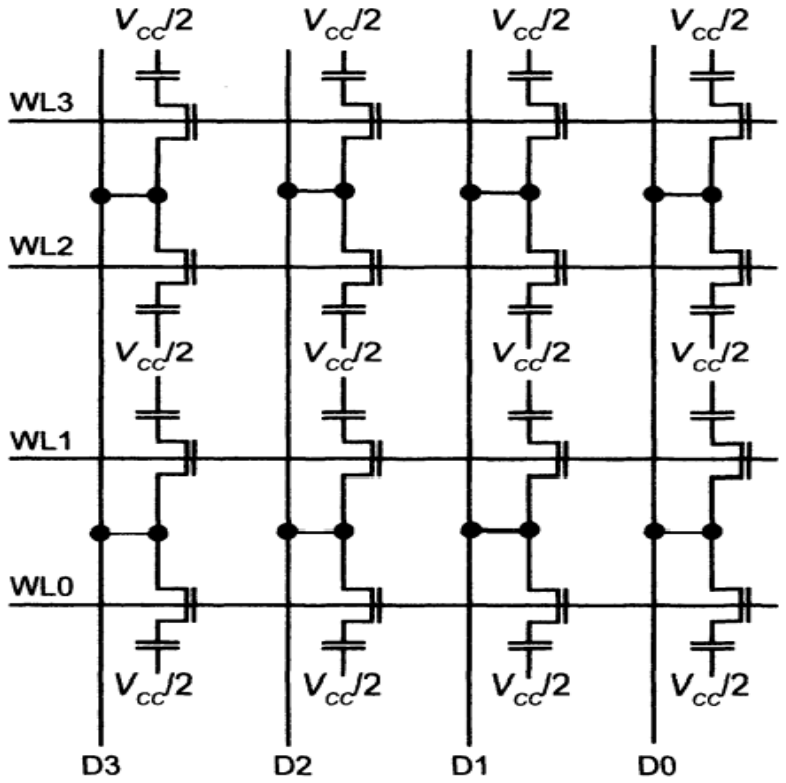
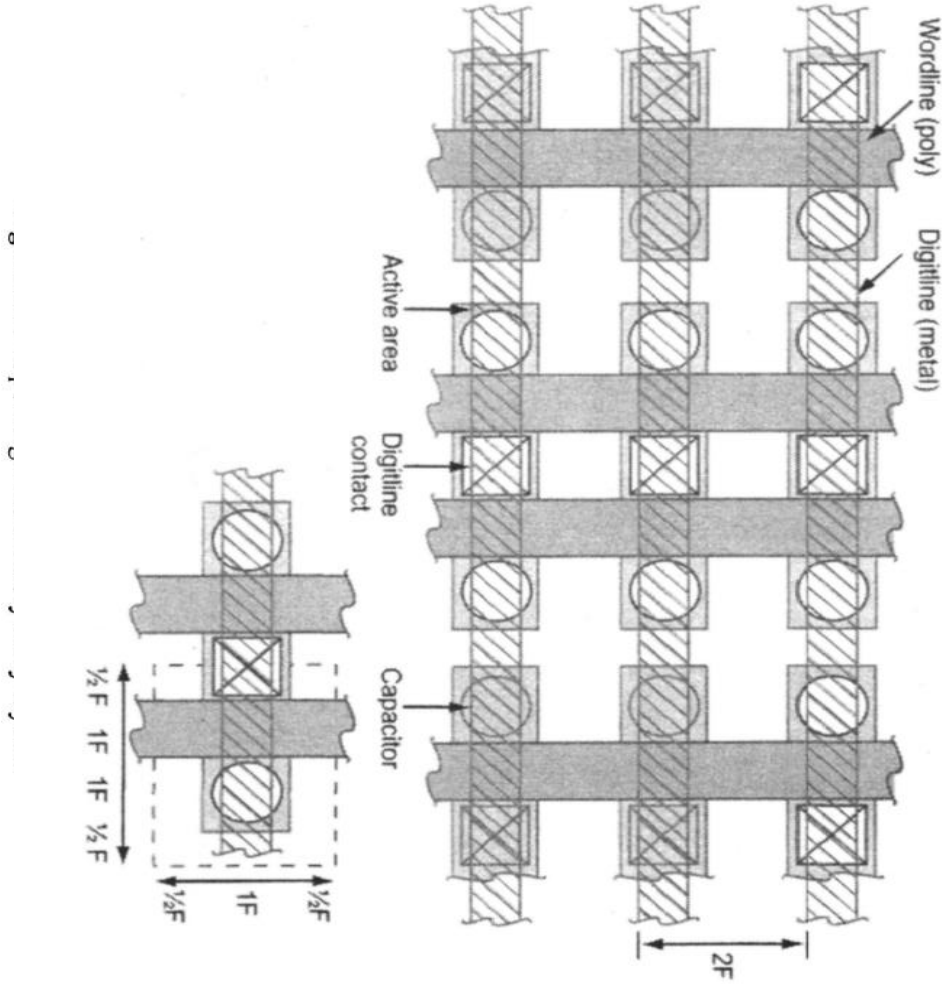


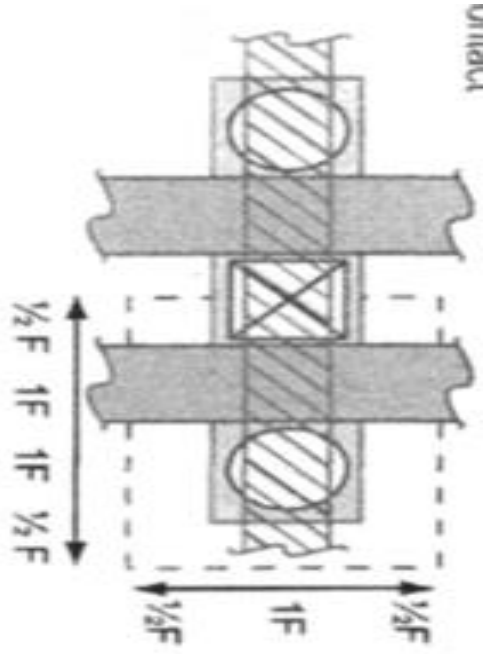
Figure 1.22 Open digitline memory array schematic.



DRAM Array Architecture :Open digitline

Because sharing a contact significantly reduces overall cell size, DRAM mbits are constructed in pairs. In this way, a digitline contact can be shared.
The mbits comprise

- An active area rectangle (in this case, an n+ active area),
- A polysilicon wordlines,
- A single digitline contact, a metal or polysilicon digitline, and
- A cell capacitors formed with an oxide-nitride-oxide dielectric between two layers of polysilicon.
- For most processes, the wordline polysilicon is silicided to reduce sheet resistance, permitting longer wordline segments without reducing speed.



DRAM Array Architecture :Open digitline

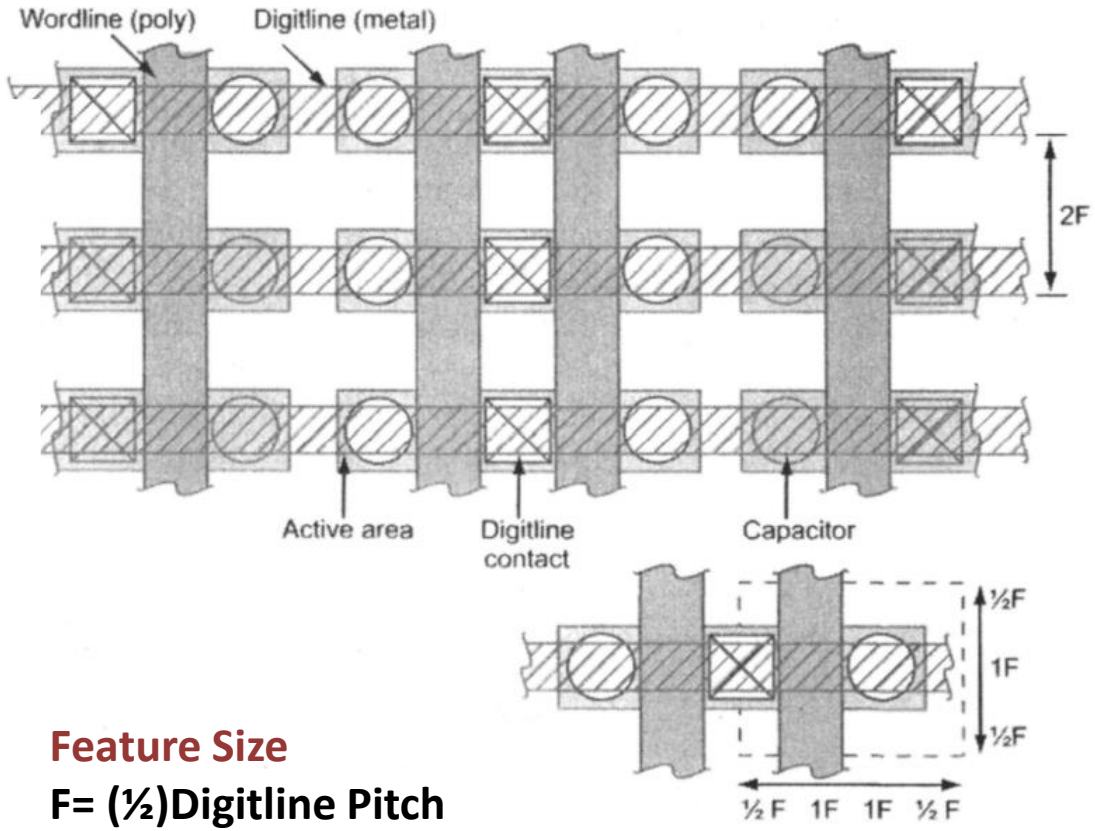
First

- Digitline pitch= Width of Digitline + Spacing
- It dictates the active area pitch and the capacitor pitch.
- **The digitline Pitch = 2F**
- The Feature (**F) = (½)Digitline Pitch**

Second

- The wordline pitch (width plus space) dictates the space available for
  - The digitline contact,
  - transistor length,
  - active area,
  - field poly width, and capacitor length.

- This Array architecture uses mbit size of $6F^2$
- A box is drawn around one of the mbits to show the $6F^2$ cell boundary.
- Again, two mbits share a common digitline contact to improve layout efficiency.



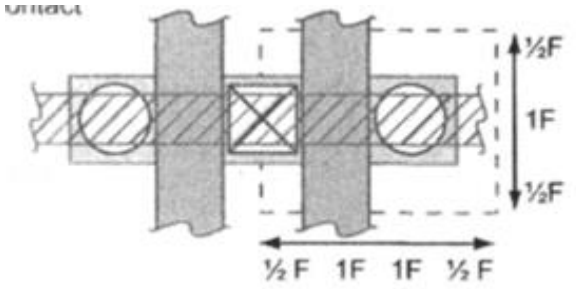
Feature Size

F= (½)Digitline Pitch

Figure 1.23 Open digitline memory array layout.

DRAM Array Architecture :Open digitline

- **Digitline capacitive components, contributed by each mbit, include**
- **The following are the capacitive components contributed by each mbit**
  - Junction capacitance,
  - Digitline to cellplate (poly3),
  - Digitline to wordline,
  - Digitline to digitline,
  - Digitline to substrate, and
  - Digitline to storage cell (poly2) capacitance.
- Therefore, each mbit connected to the digitline adds a specific amount of capacitance to the digitline.
- Most modern DRAM designs have no more than 512 mbits connected to a digitline segment.



Digitline Pitch

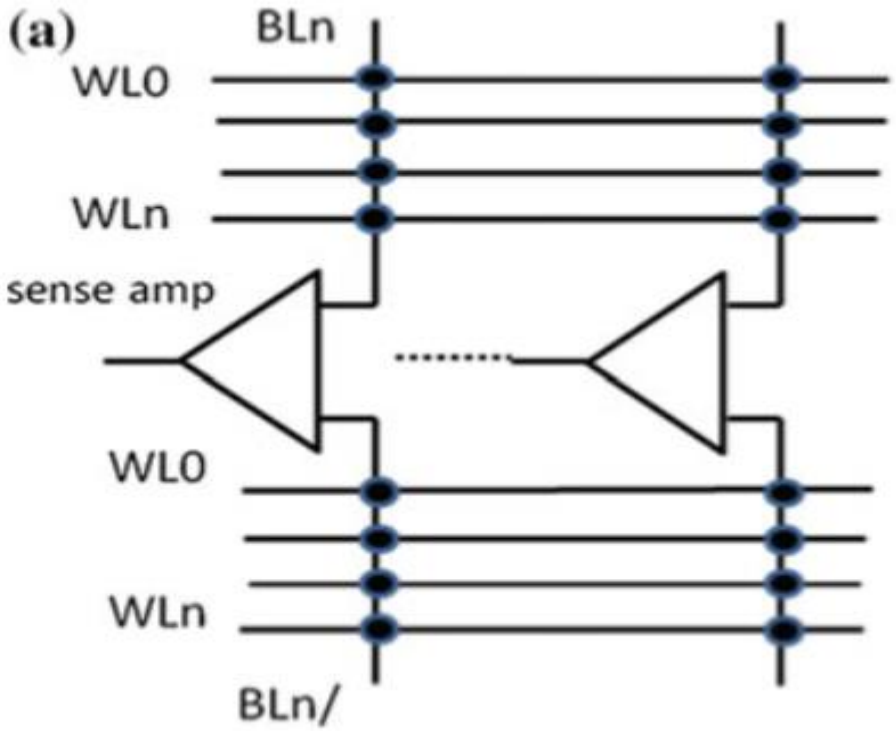
Feature Size

F= (½)Digitline Pitch

DRAM Array Architecture :Folded Array

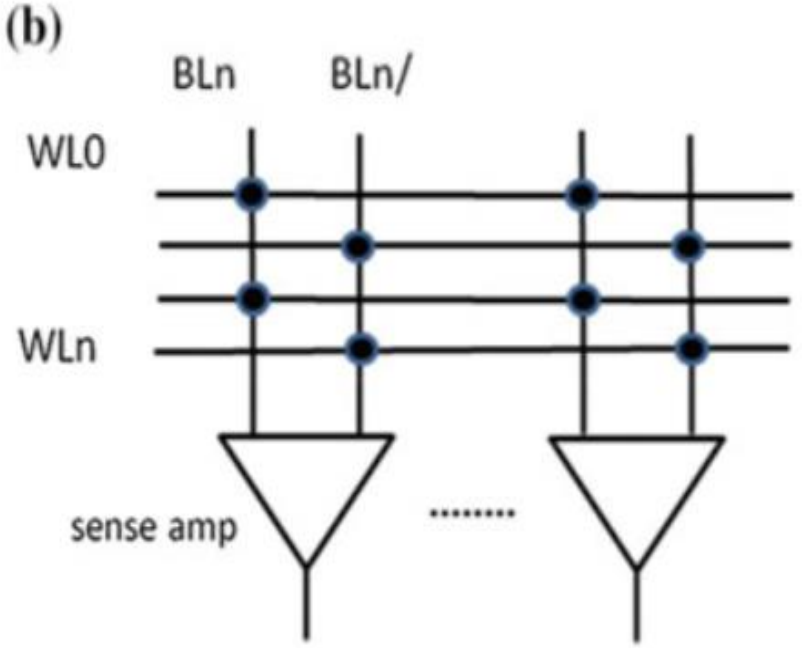
2. Folded Array Architecture

Open digitline Array Architecture



This precludes **(prevents) using digitline twisting to improve Signal to Noise performance as D0 and D0* are from two different arrays**

Folded Array Architecture



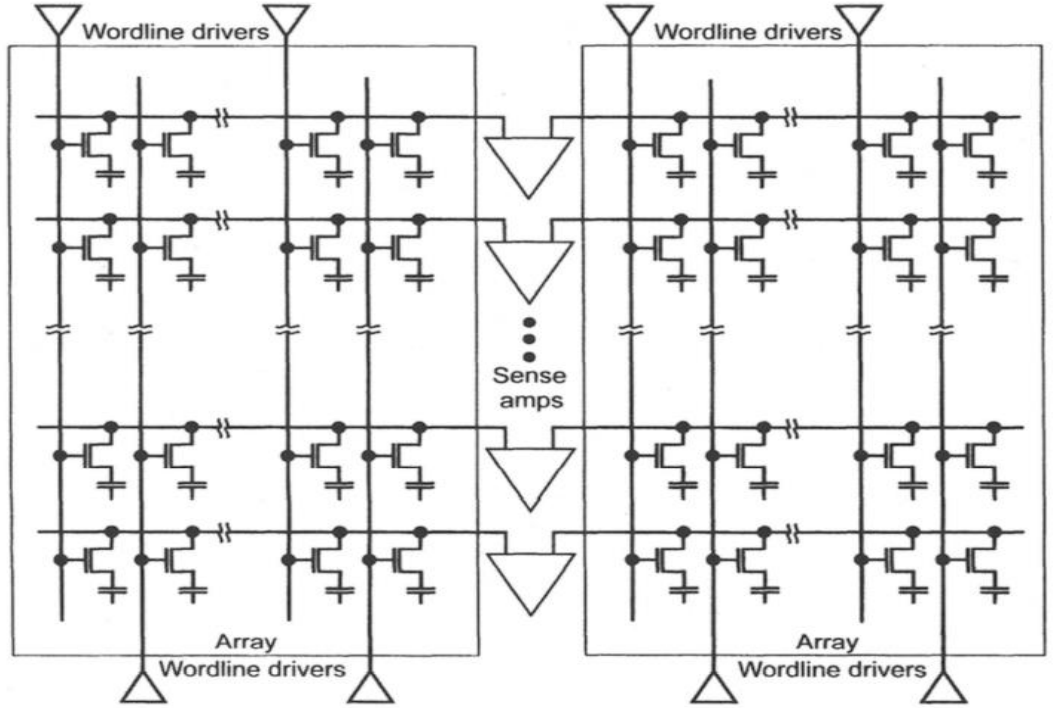
This allows **digitline twisting to improve Signal to Noise performance as D0 and D0* are from same arrays but at the cost of more area**

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

DRAM Array Architecture :Folded Array

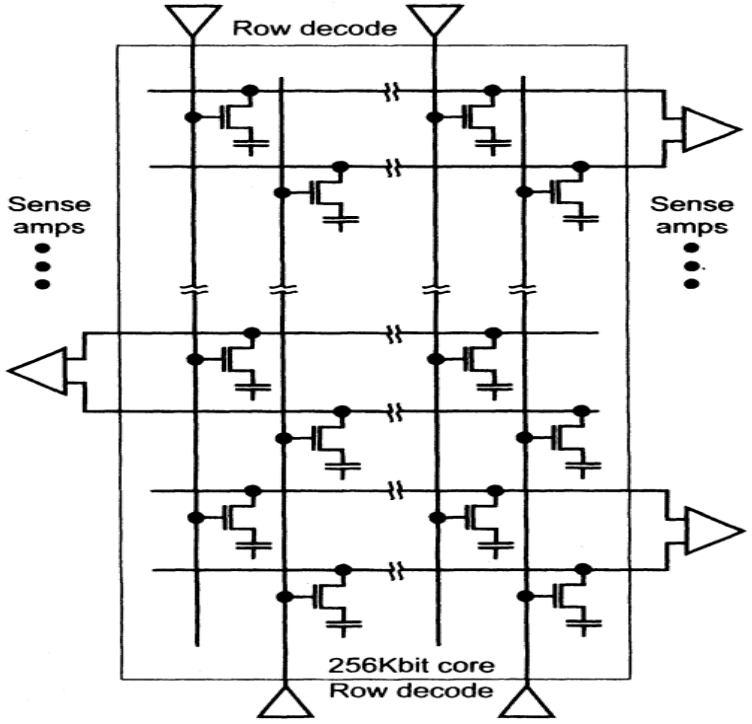
2. Folded Array Architecture

Open digitline Array Architecture



This precludes **(prevents) using digitline twisting to improve Signal to Noise performance as D0 and D0* are from two different arrays**

Folded Array Architecture



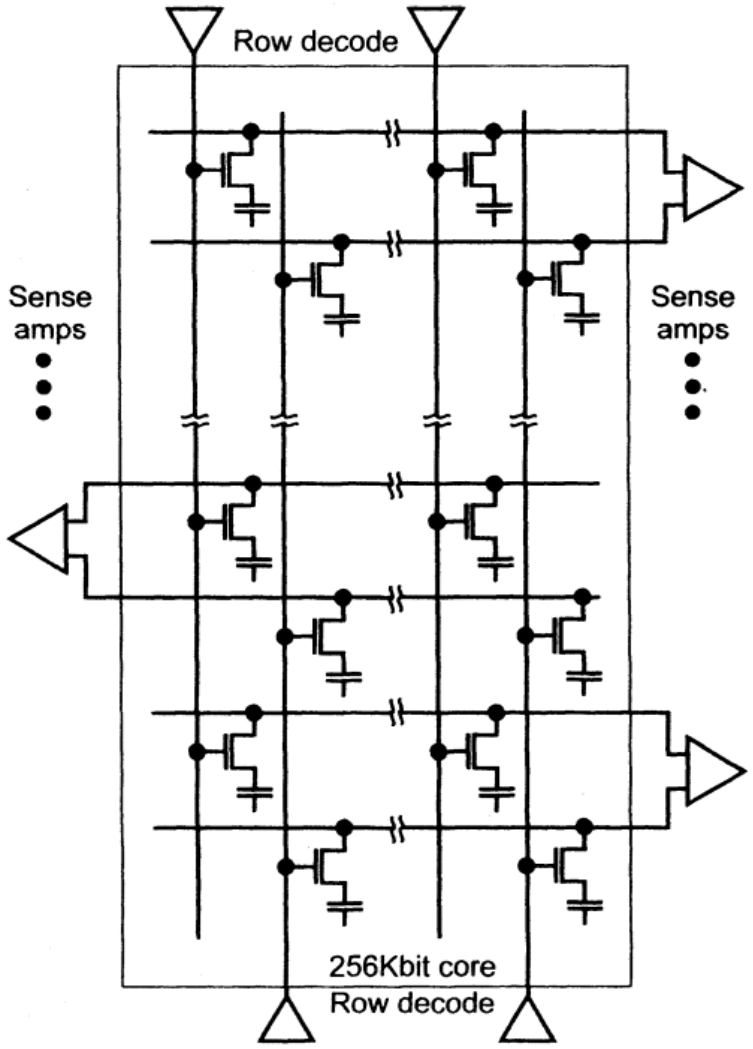
This allows **digitline twisting to improve Signal to Noise performance as D0 and D0* are from same arrays**

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

DRAM Array Architecture :Folded Array

Features:

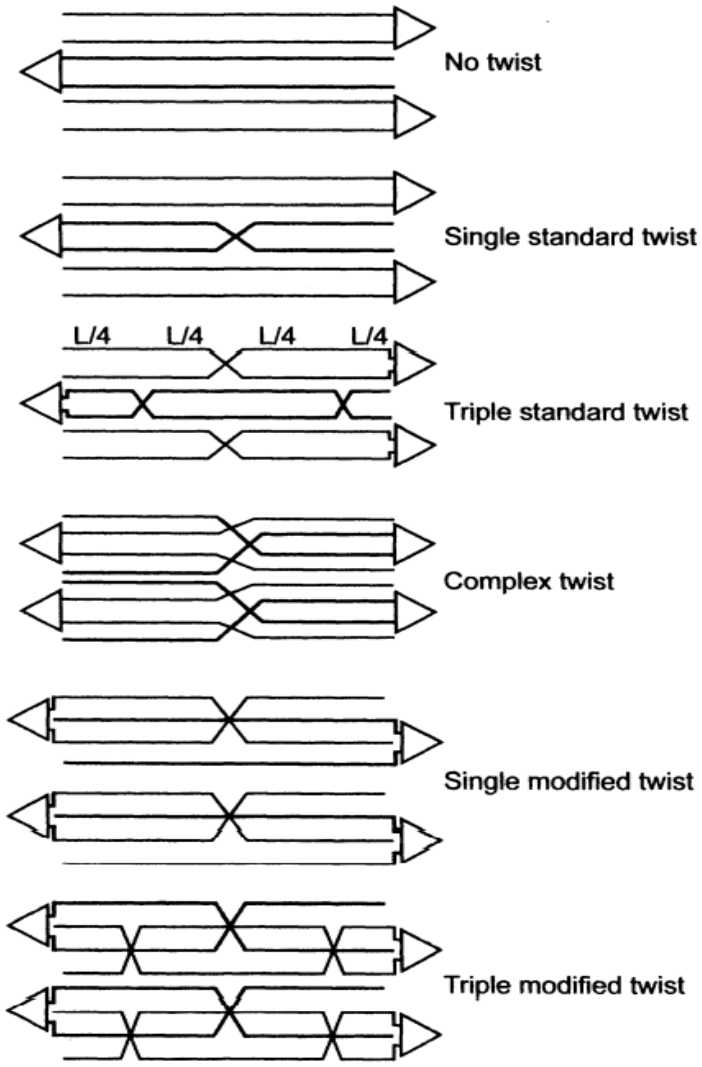
- In folded array architecture **Sense amplifier circuits placed at the edge of each array** connect to both true and complement digitlines (D and D*) **coming from a single array**.
- Optional **digitline pair twisting in one or more positions** reduces and balances the coupling to adjacent digitline pairs and **improves overall signal-to-noise characteristics .**



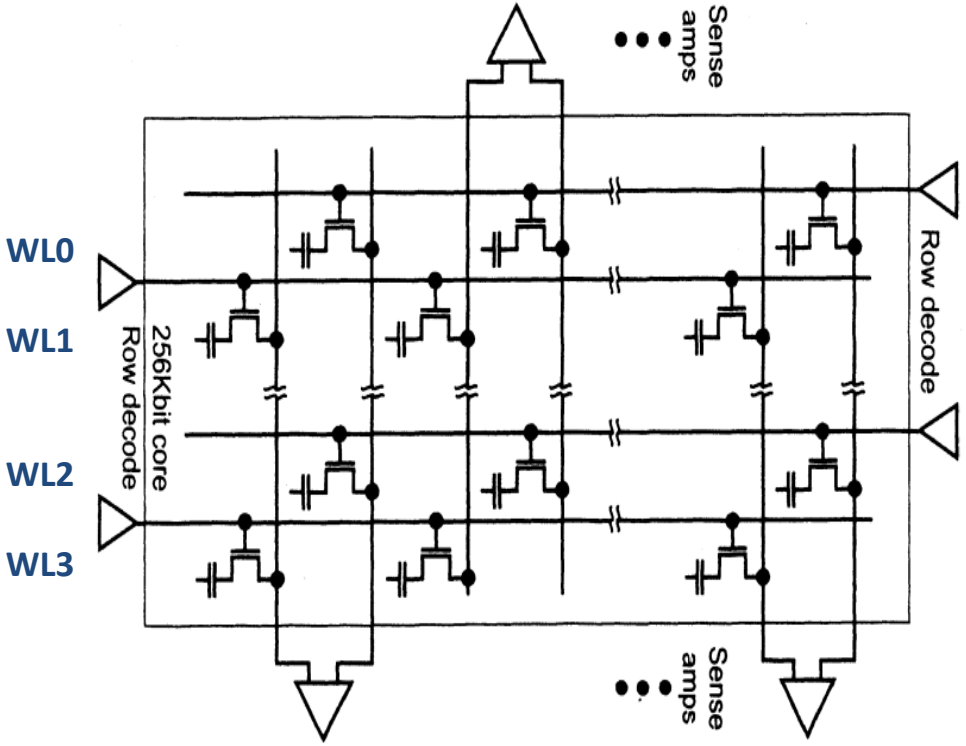
DRAM Array Architecture :Folded Array

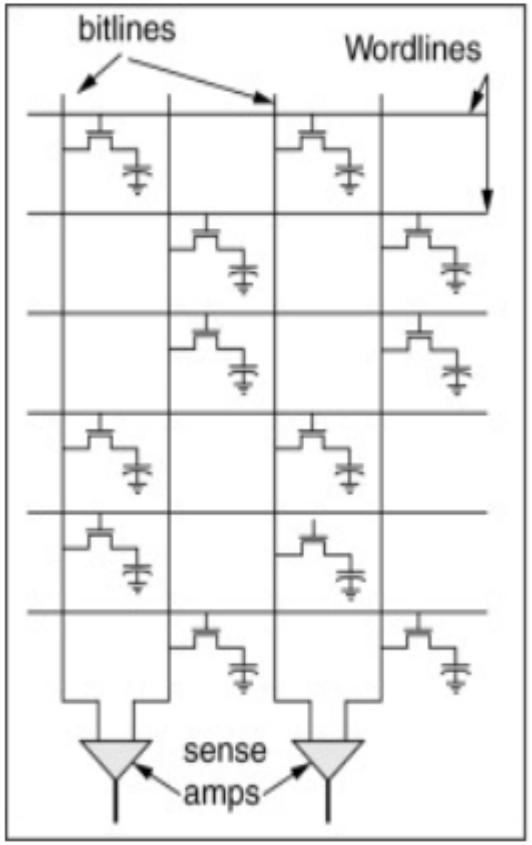
Features:

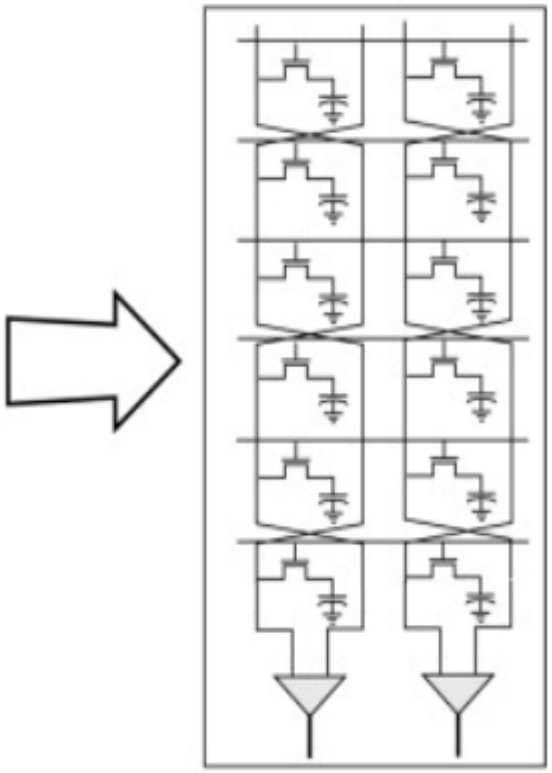
- There are variety of twisting schemes used throughout the DRAM industry.
- Ideally, a **twisting scheme equalizes the coupling terms from each digitline to all other digitlines**, both true and complement.
- If implemented properly, **the noise terms cancel or at least produce only common-mode noise to which the differential sense amplifier is more immune.**
- Each digitline twist region consumes valuable silicon area. Thus, design engineers resort to the simplest and most efficient twisting scheme to get the job done.
- Because the coupling between adjacent metal lines is inversely proportional to the line spacing, the signal-to-noise problem gets increasingly worse as DRAMs scale to smaller and smaller dimensions.
- Hence, the industry trend is toward use of more complex twisting schemes on succeeding generations .



DRAM Array Architecture :Folded Array



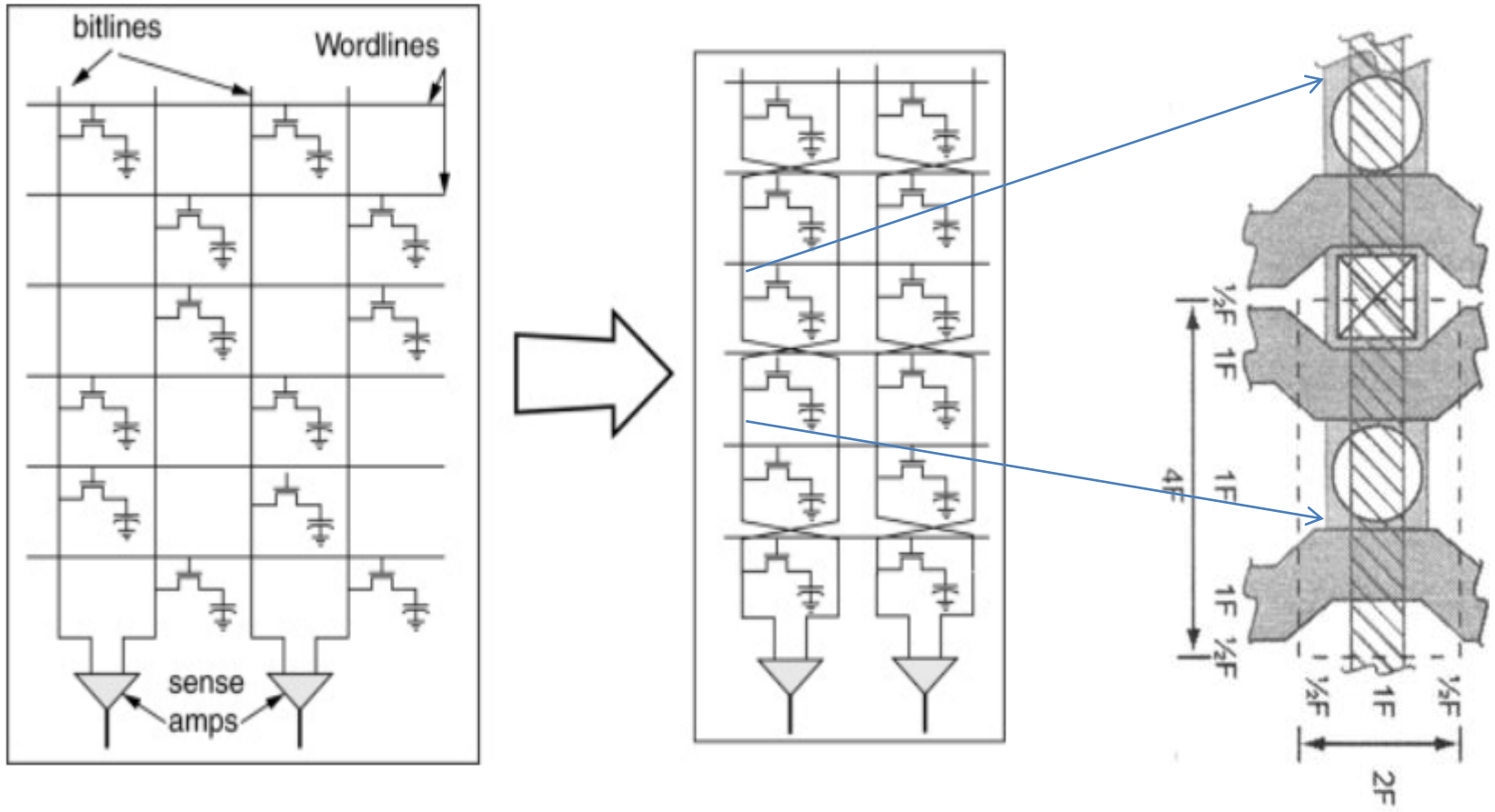




DRAM Array Architecture:Folded Array

Features:

Buried capacitor DRAM mbit pair.



The layout of Folded Array architecture:

The mbits comprise

- An active area rectangle (in this case, an n+ active area),
- A pair of polysilicon wordlines,
- a single digitline contact, a metal or polysilicon digitline, and
- a pair of cell capacitors.

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

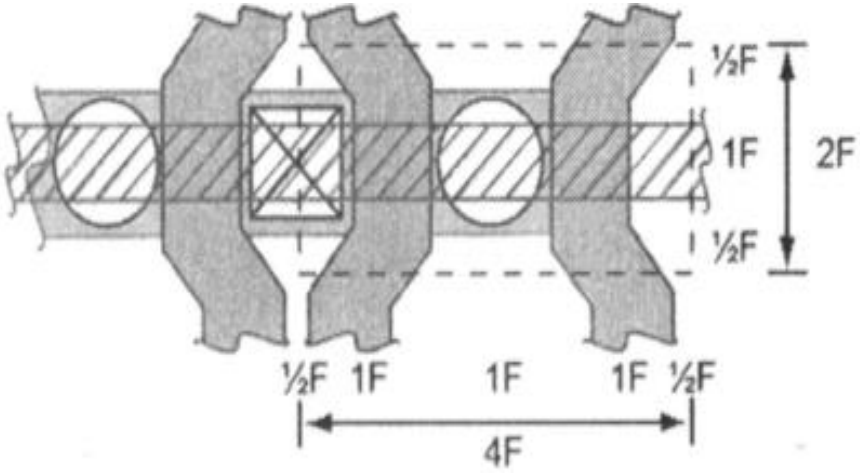
DRAM Array Architecture :Folded Array

Features:

- Folded Array Architecture has a area of $8F^2$
- An imaginary box drawn around the mbit defines the cell's outer boundary.
- Along the x-axis, this box includes
  - one-half digitline contact feature,
  - one wordline feature,
  - one capacitor feature,
  - one field poly feature, and
  - one-half poly space feature
- Along the y-axis, this box contains
  - two one-half field oxide features and
  - one active area feature.

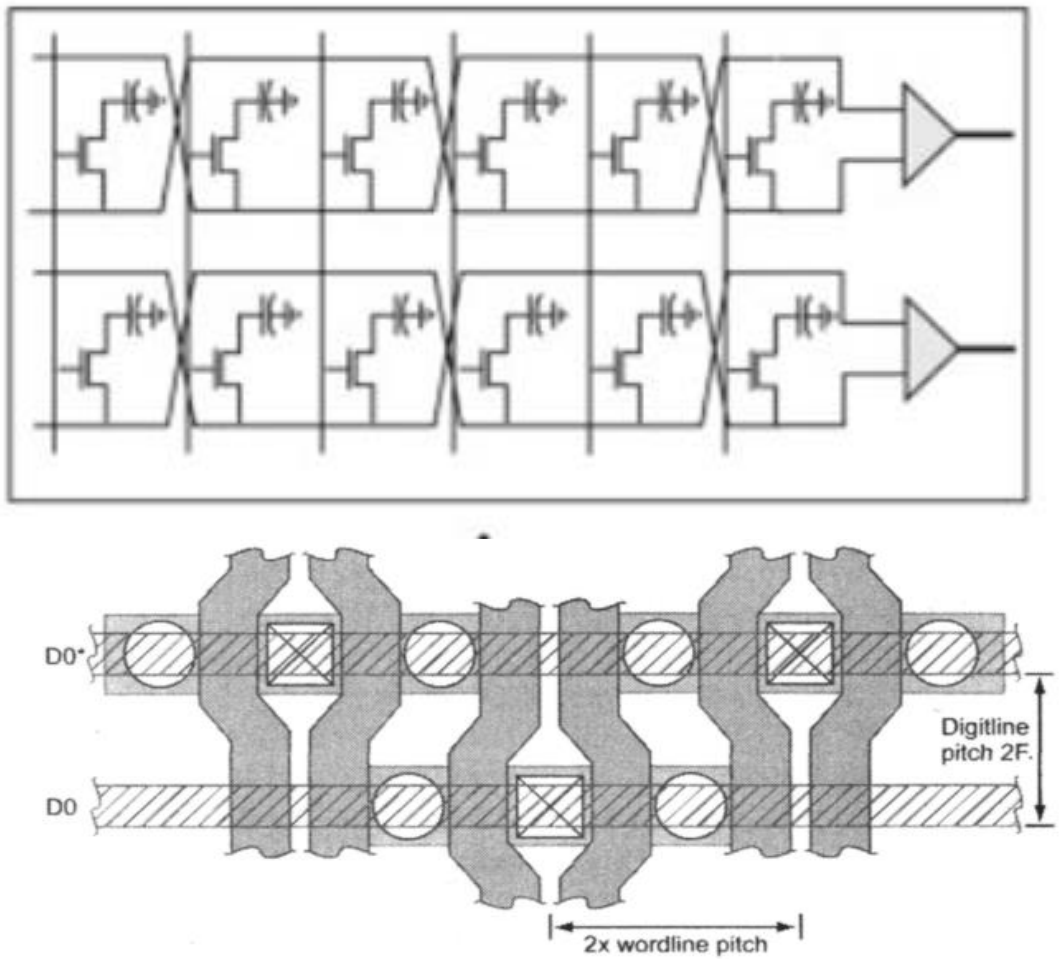
The area of the mbit is therefore

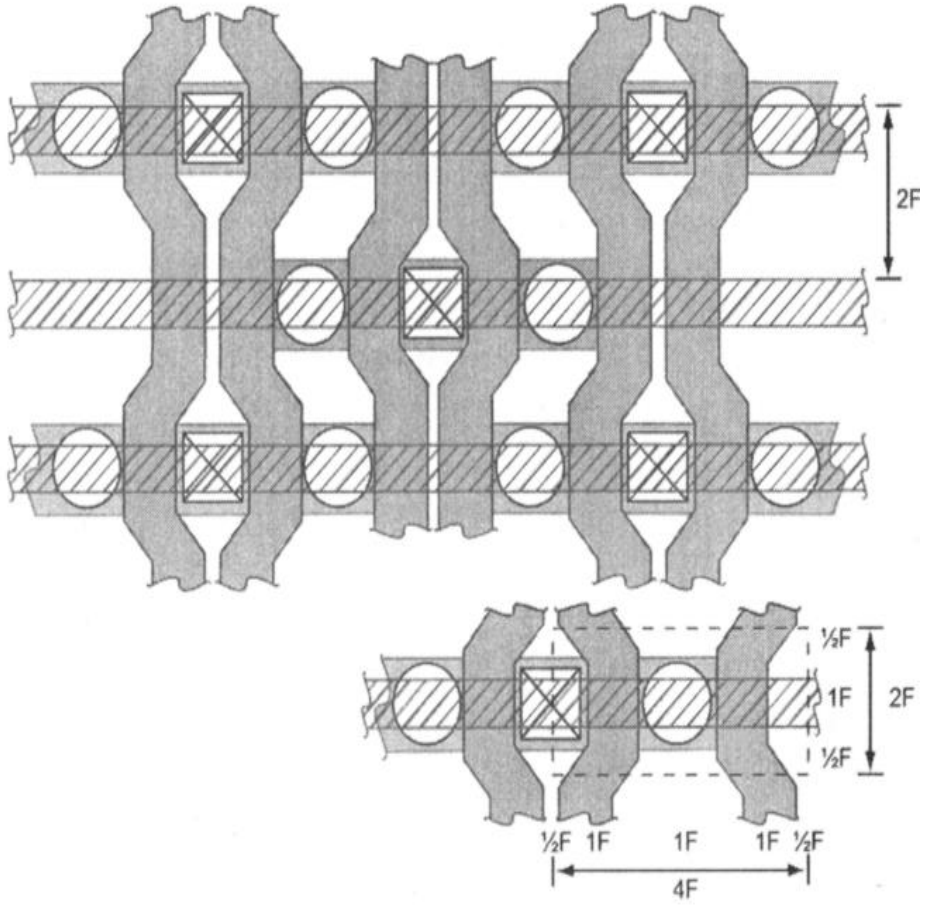
$$4F \cdot 2F = 8F^2$$



DRAM Array Architecture : :Folded Array

Features:





Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Memory Design and Testing

Mahesh Awati
Department of Electronics and Communication Engg.

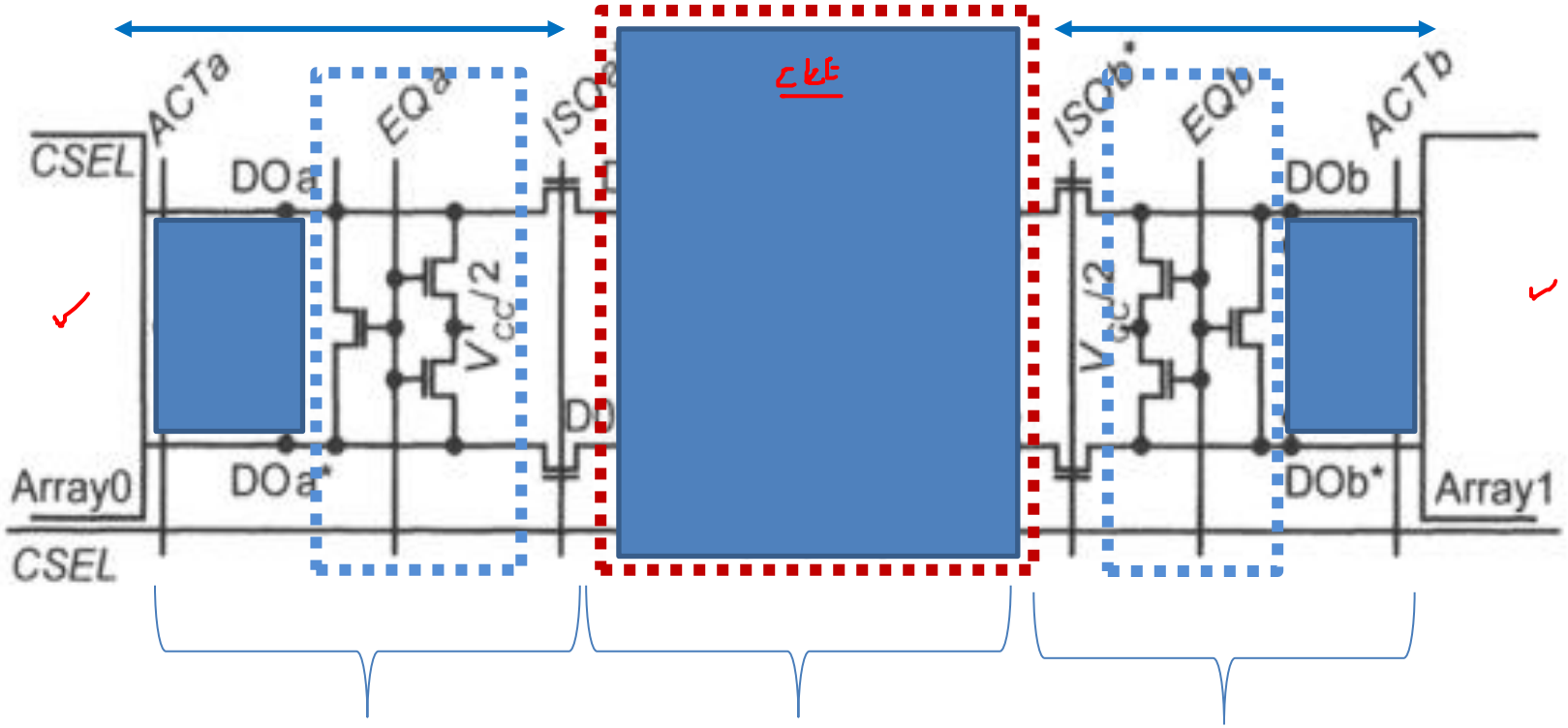
MEMORY DESIGN AND TESTING

UNIT 2 – Dynamic Random Access Memory

Mahesh Awati
Department of Electronics and Communication Engineering

Standard sense amplifier block

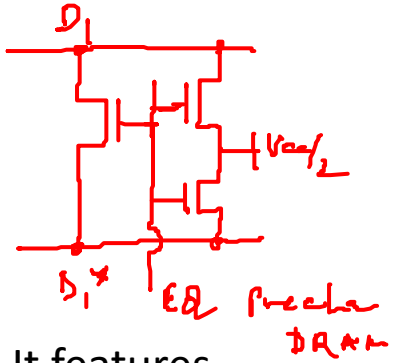
Equilibration and Bias Circuits (Pre-charge)



Peripheral Circuit of Array0

Common Peripheral Circuits

Peripheral Circuit of Array1

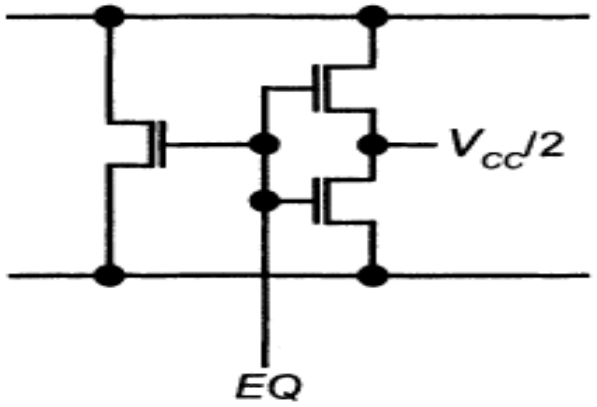


It features

a) Two Psense amplifiers outside the isolation transistors,
b) A pair of EQ/bias (EQa and EQb) devices,
c) A single Nsense amplifier, and
d) A single I/O transistor for each digitline.

DRAM Array Architecture

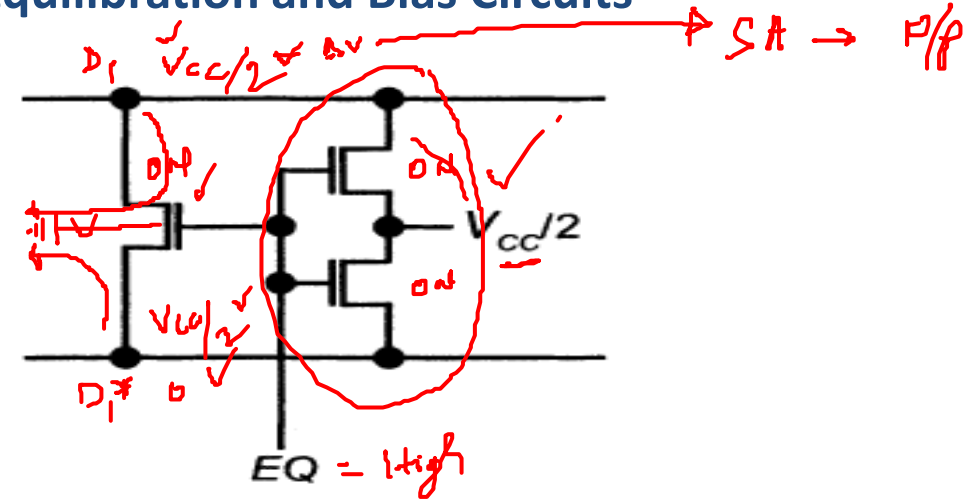
Equilibration and Bias Circuits



- Digital lines are pre-charged to Vcc/2 prior to cell accessing and Sensing.
- For Sensing operation, It is important that digitlines are at same potentially i.e., Vcc/2 as the digitline pair acts as differential signal for SA.
- Any offset between pair of digitines, reduces the effective signal produced during the access operation.
- The pre-charge circuit is made up NMOS transistors – It consists of
  - Equilibration Transistor and
  - Bias Transistors
- The Gate of all these NMOS devices are connected to EQ signal. EQ is held HIGH when RAS* is HIGH or else LOW , It means Precharge is active when access transistors of Cell are Inactive or vica versa.

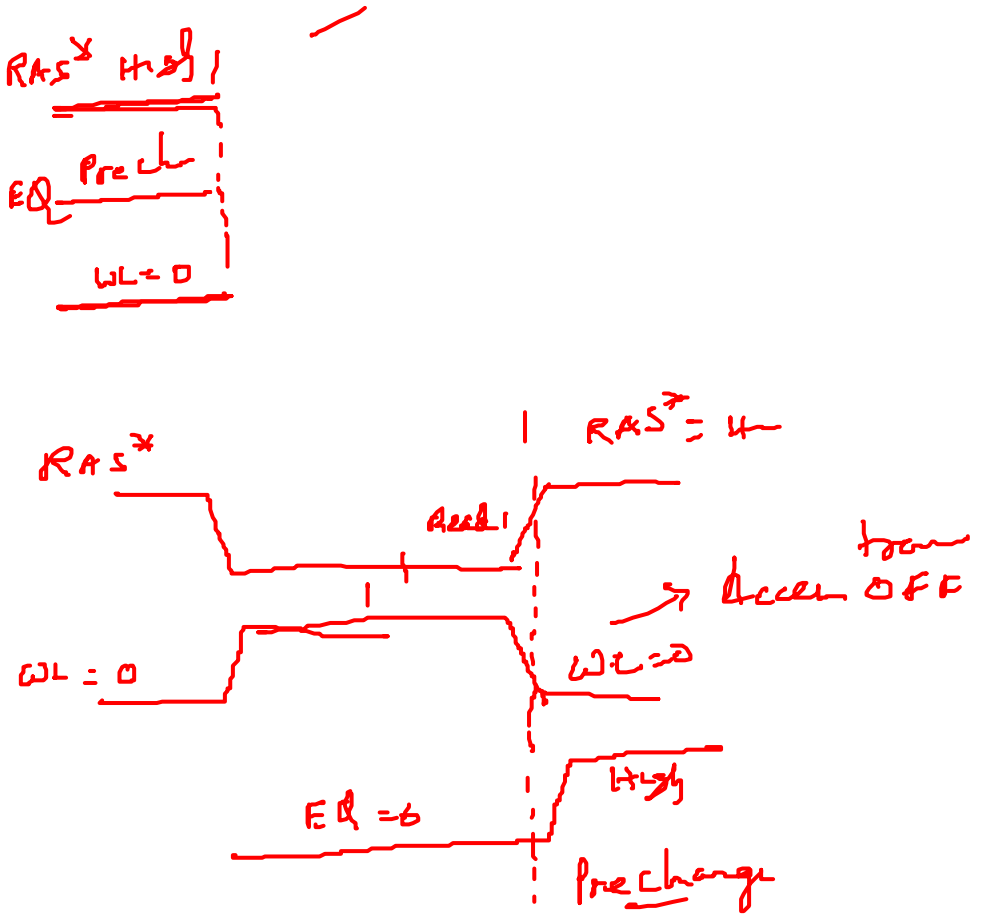
DRAM Array Architecture

Equilibration and Bias Circuits



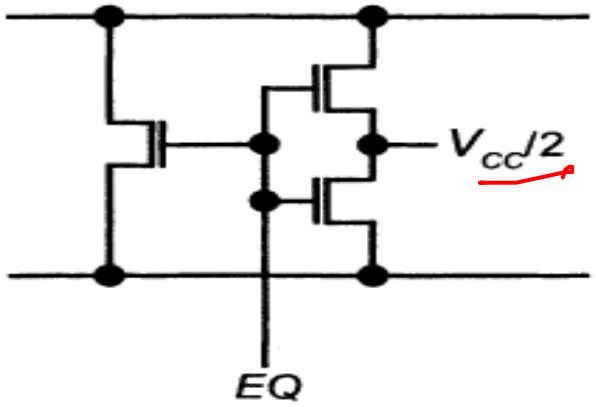
Working Sequence

- **EQ=HIGH , when RAS* = HIGH** , this indicates precharge or Inactive state of DRAM.
- **After Falling edge of RAS*, (High to Low Transition) , the EQ=LOW** – which turns of equilibration transistors just prior to WL goes HIGH.
- **EQ=HIGH at the end of RAS* cycle to** force equilibration of digit lines.



DRAM Array Architecture

Equilibration and Bias Circuits

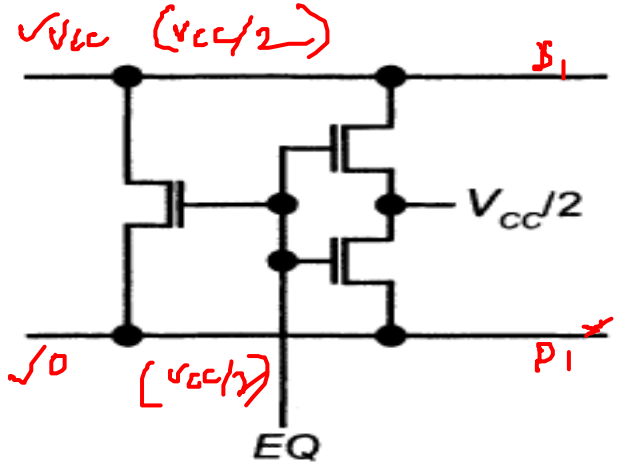


Why NMOS ?

- NMOS have **High driving capacity which results in faster equilibrium**.
- After the RAS* cycle, **Normally the digitlines are at Vcc and gnd which will get equilibrate to Vcc/2 by charge sharing.**
- The other two NMOS provide bias voltage Vcc/2. These operate in conjunction with equilibration NMOS to ensure **bitlines are at Vcc/2 despite the leakage path that would otherwise discharged them.**
- **This ensures that the bitlines are at prescribed voltage level for sensing**

DRAM Array Architecture

Equilibration and Bias Circuits

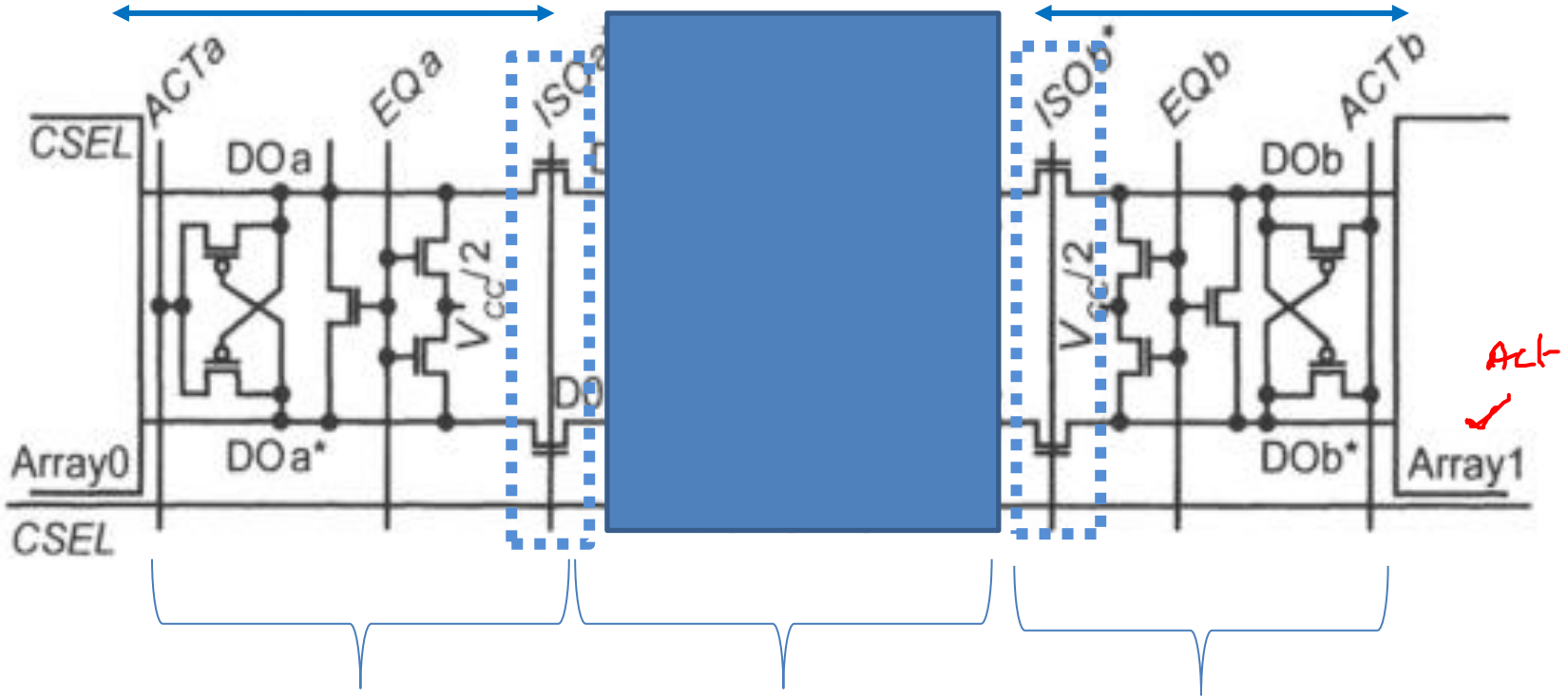


Why Vcc/2 ?

- Vcc/2 pre-charge is used as it
  - **Reduces the Power Consumption** – As the pre-charge voltage can be obtained by equilibrating the digitlines (which are at Vcc and GND) at the end of each cycle.
  - **The charge sharing between the digitlines produces the Vcc/2 without additional Icc current.**
  - **Reduces the Read – Write Time as Pre-charge time required is less**: Every Read or Write involves pre-charging and as it is pre-charged to Vcc/2 only , it will be faster.
  - **Improves Sensing operation.**

Standard sense amplifier block

Isolation Device

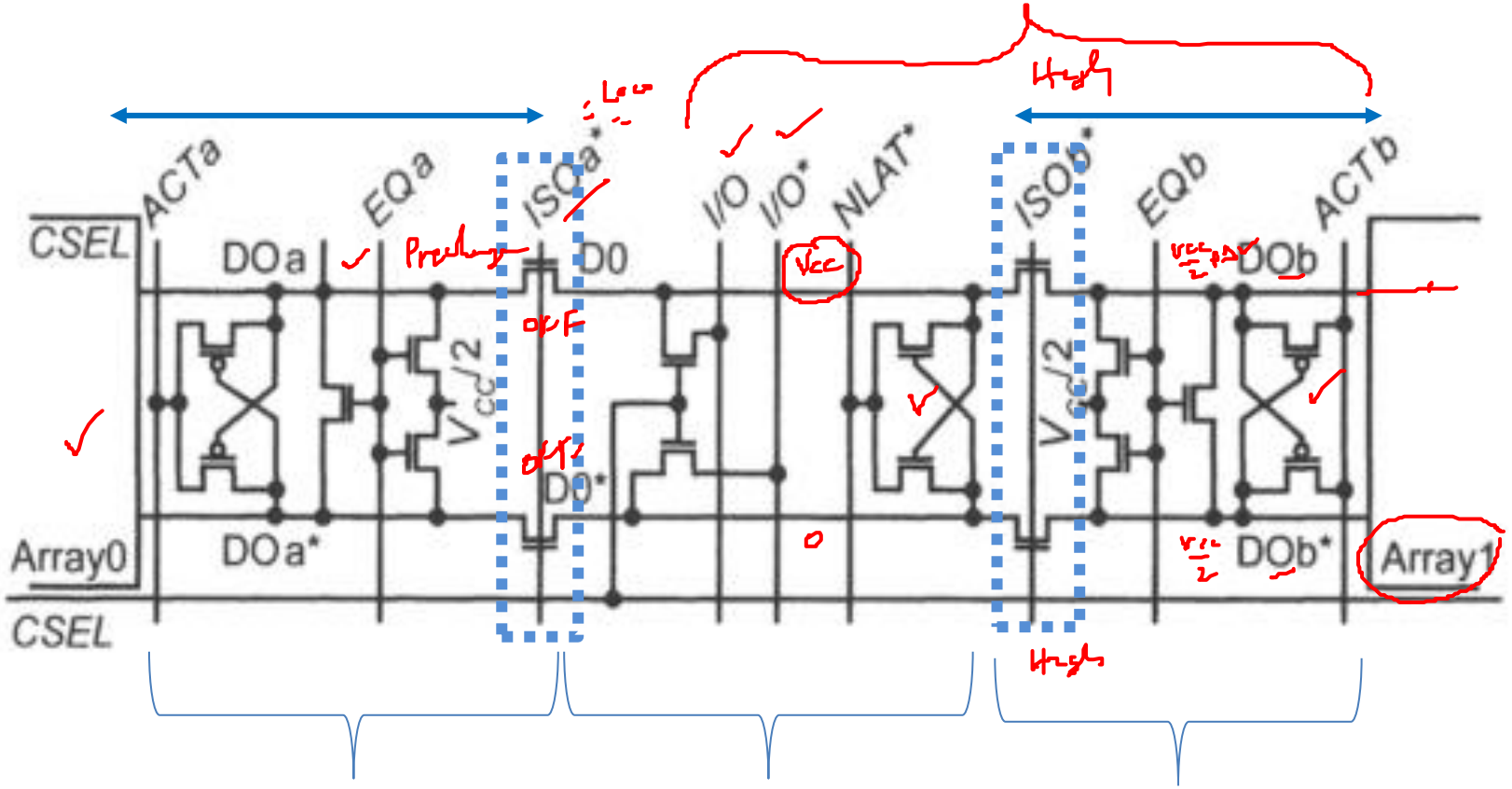


It features

a) Two Psense amplifiers outside the isolation transistors,
b) A pair of EQ/bias (EQa and EQb) devices,
c) A single Nsense amplifier, and
d) A single I/O transistor for each digitline.

Standard sense amplifier block

Isolation Device



It features

a) Two Psense amplifiers outside the isolation transistors,
b) A pair of EQ/bias (EQa and EQb) devices,
c) A single Nsense amplifier, and
d) A single I/O transistor for each digitline.

DRAM Array Architecture

Isolation Devices

Isolation devices are important to the sense amplifier circuits.
These devices are NMOS transistors placed between the array digitlines and the sense amplifiers.
The isolation devices provide two functions.

1. if the **sense amps are positioned between and connected to two arrays, they electrically isolate one of the two arrays.**
   - This is necessary whenever a wordline fires in one array because of the isolation of the second array – **The digitline capacitance driven by sense Amplifier reduces.**
   - This leads to
     - **Speeding Read-Write times,**
     - **Reducing power consumption, and**
     - **Extending Refresh for the isolated array.**

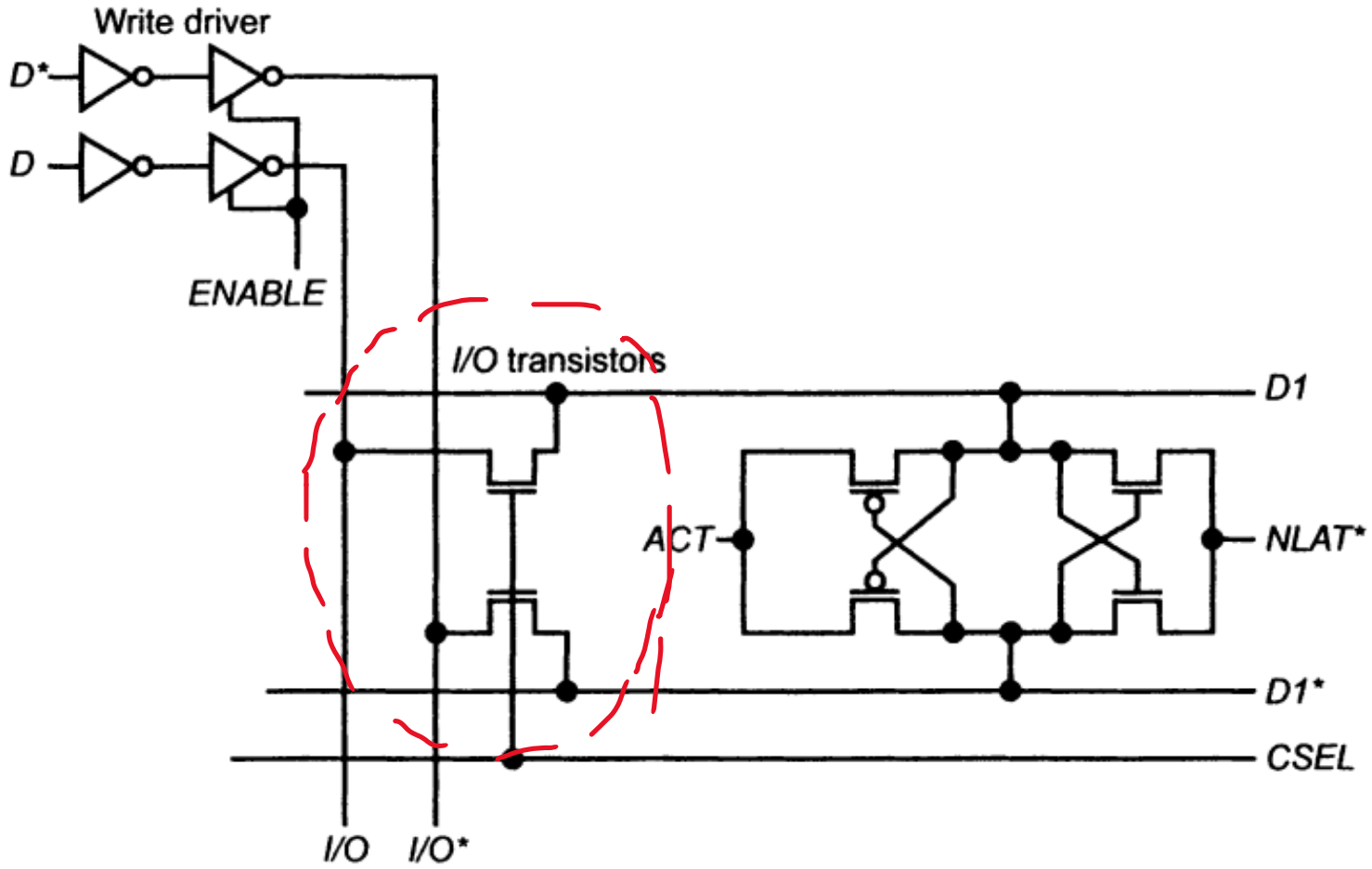
DRAM Array Architecture

Isolation Devices

2. The isolation devices provide **resistance between the sense amplifier and the digitlines.**
   - This resistance stabilizes the sense amplifiers and **speeds up the Sensing operation by somewhat isolating the highly capacitive digitlines from the low-capacitance sense nodes** .
   - Capacitance of the **sense nodes between isolation transistors is generally less than 15fF**, permitting the sense amplifier to **latch much faster** than if it were solidly connected to the digitlines.
   - The isolation transistors slow Write-Back to the mbits, but this is far less of a problem than initial Sensing

DRAM Array Architecture

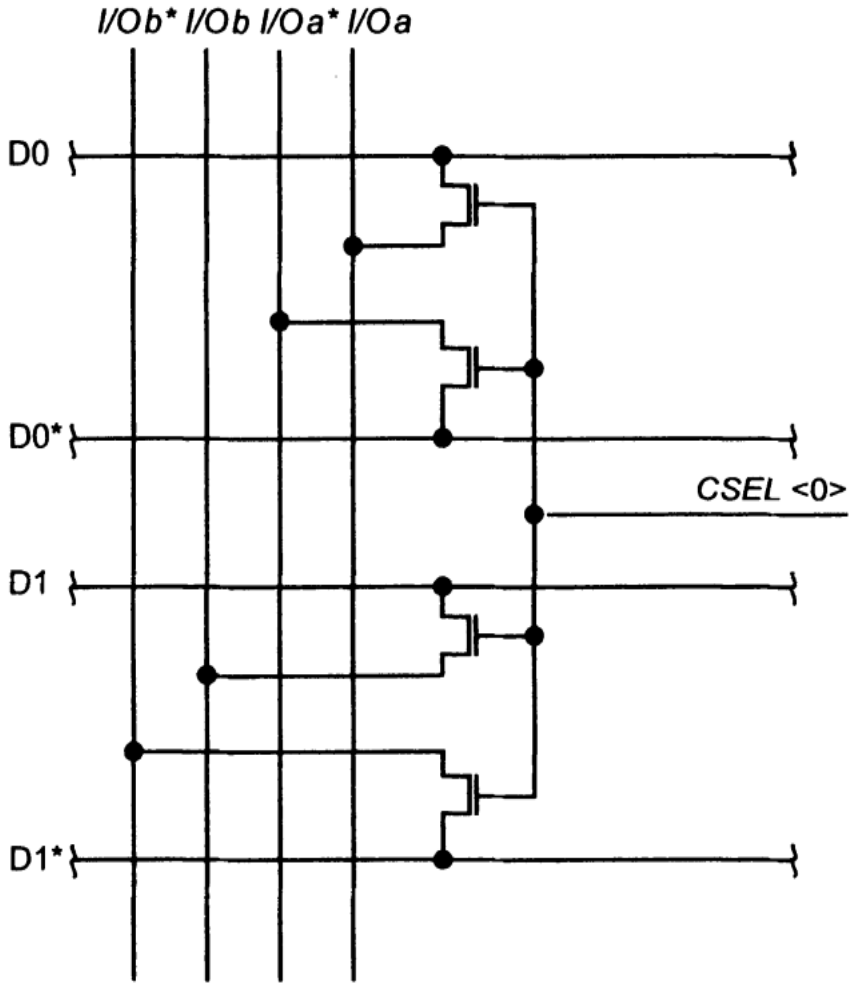
Input-Output Transistors



DRAM Array Architecture

Input-Output Transistors

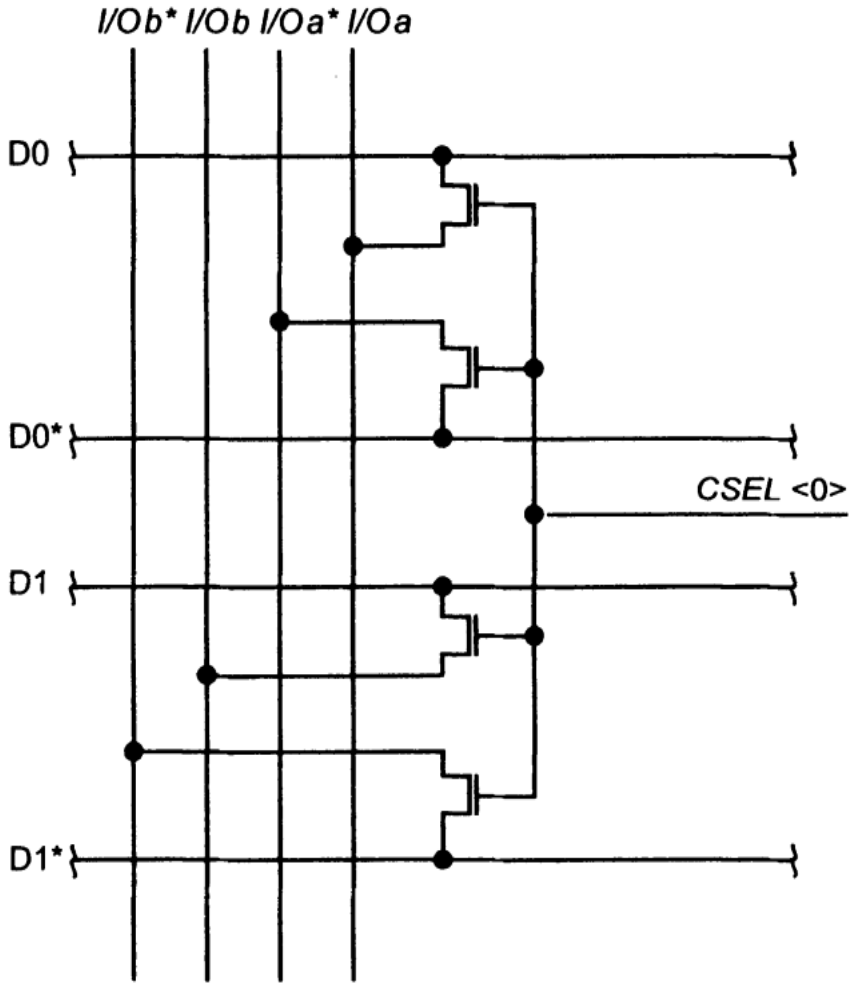
- The Input – Output Transistors allows the data to be read from or written to specific digitline pair.
- The output of IO transistor pair is connected to IO signal pair (I/Ob*,I/Ob..)
- Commonly, there are two pairs of I/O signal lines, which permit four I/O transistors to share a single column select (CSEL) control signal.
- **The I/0 transistors must be sized carefully** to ensure that **instability is not introduced into the sense amplifiers by the I/O bias voltage or remnant voltages** on the I/0 lines.



DRAM Array Architecture

Input-Output Transistors

- **I/O transistors are designed to be two to eight times smaller than the Nsense amplifier transistors.**
- This is sometimes referred to as beta ratio. A beta ratio between **five and eight is considered standard;**



THANK YOU

Mahesh Awati
Department of Electronics and Communication

mahesha@pes.edu

+91 9741172822

Memory Design and Testing

Mahesh Awati
Department of Electronics and Communication Engg.

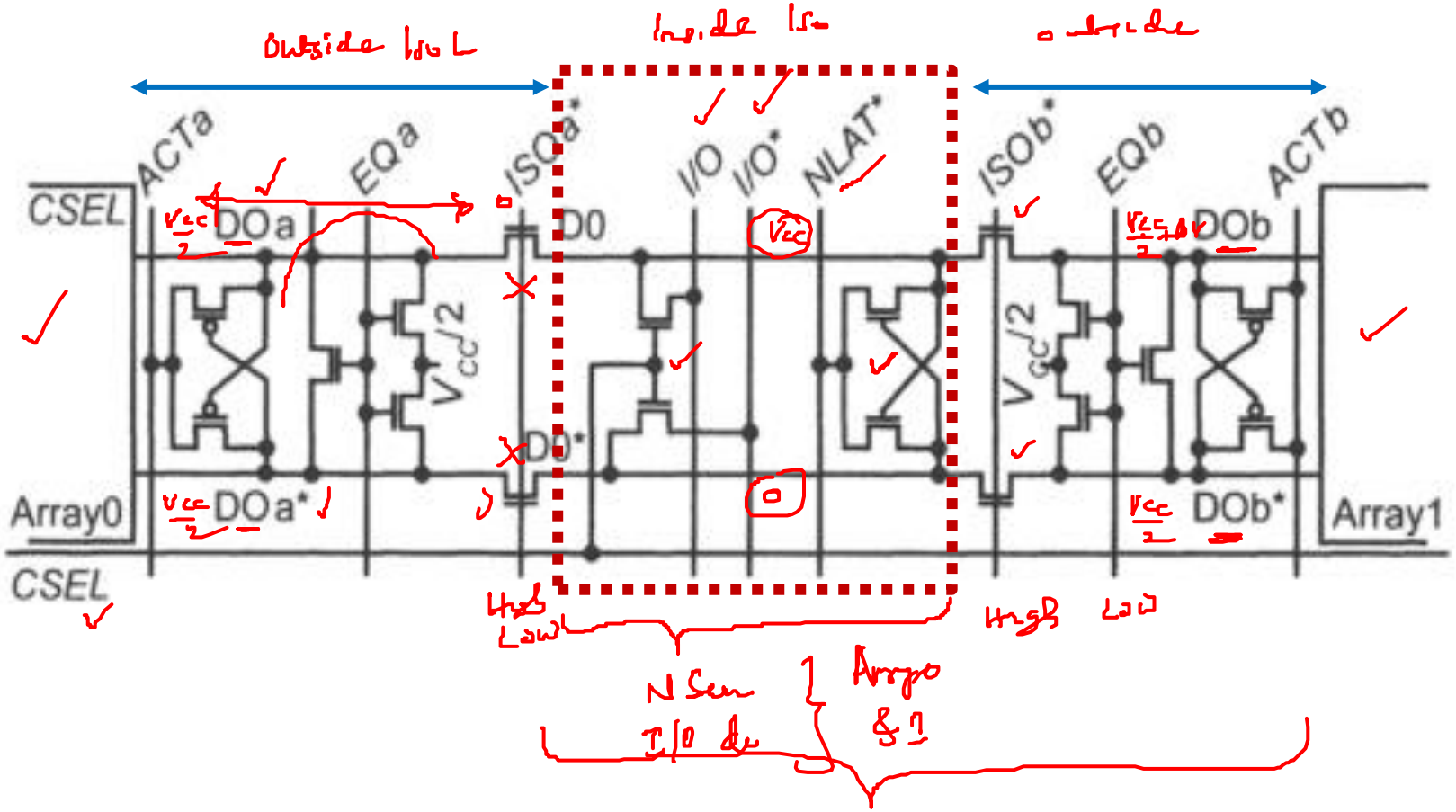
MEMORY DESIGN AND TESTING

UNIT 2 – Dynamic Random Access Memory

Mahesh Awati
Department of Electronics and Communication Engineering

Standard sense amplifier block

Standard sense amplifier block (Folded Array Architecture)



It features

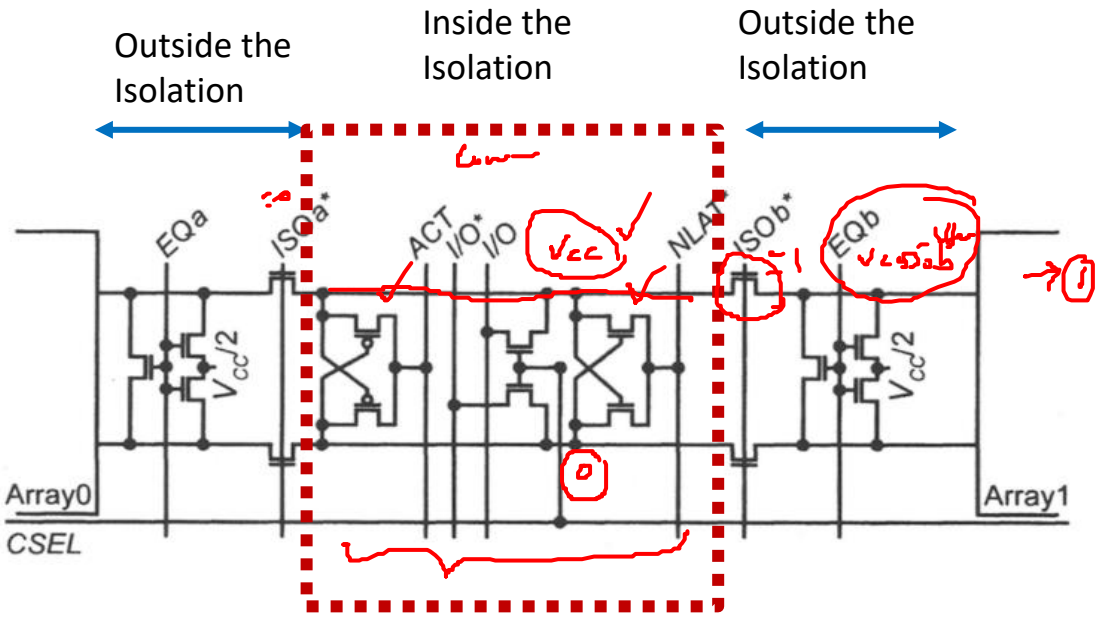
a) Two Psense amplifiers outside the isolation transistors,
b) A pair of EQ/bias (EQa and EQb) devices,
c) A single Nsense amplifier, and
d) A single I/O transistor for each digitline.

Standard sense amplifier block

Why Psense amplifiers are outside the Isolation ?

- To understand why Psense amplifiers are placed otside the Isolation in standard Sense Amplifier block, Let us consider **the Reduced Sense Amplifier Block with Psense Amplifier Inside the Isolation**
- Consider Writing '1' into mbit of Array1
  - ISOb*=HIGH and ISOa*=Low means Isolation Circuit is ON in Array1 side.
  - CSEL = High and I/O and I/O* are holding '1' and '0' ;
  - This data is made available on D0b and D0b* by SA which are already in Fired state.
  - As Psense is inside Isolation, Psense drives D0b= Vcc but it has to pass through Isolation Transistor to be written into mbit.
  - As Isolation transistors are nmos , they cant pass strong '1'. In order to do so , either the Isolation transistors must be Depletion type nmos or gate must be boosted to at least Vcc+Vth

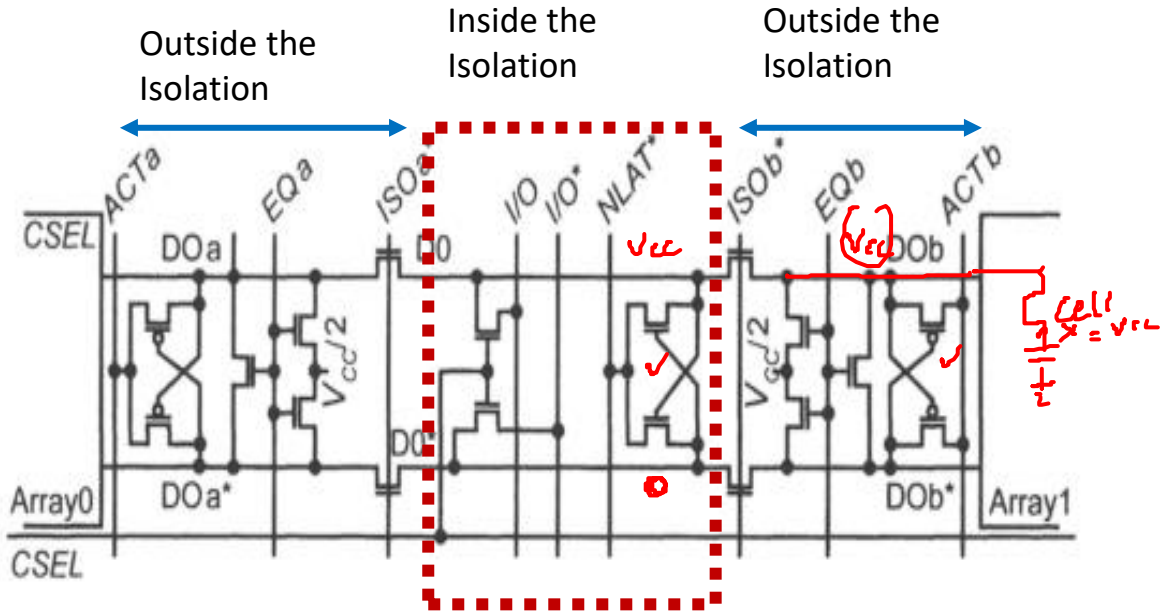
Reduced Sense Amplifier Block



Standard sense amplifier block

Standard sense amplifier block

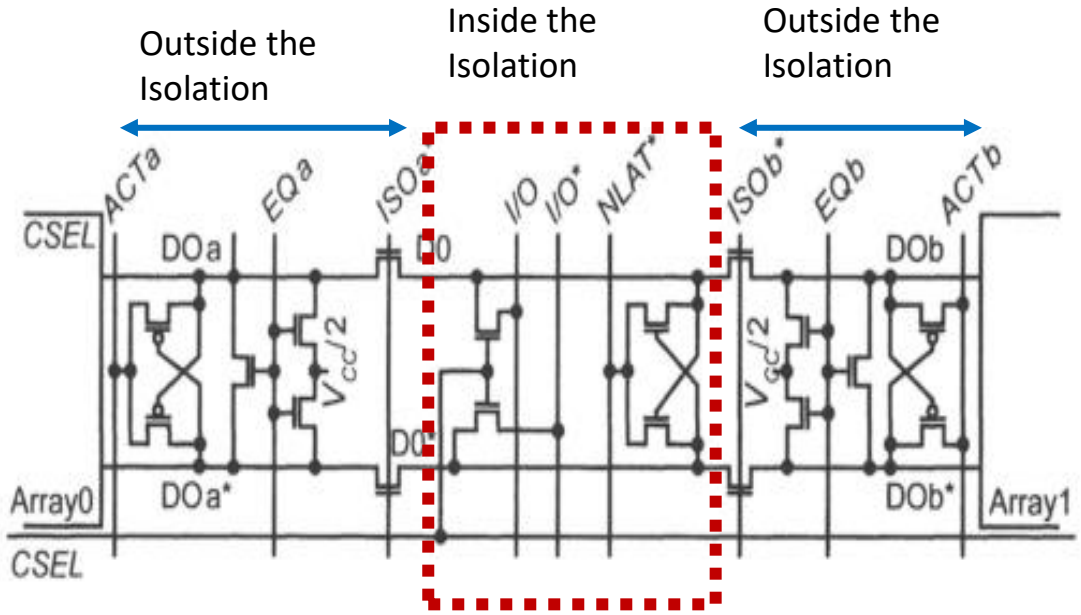
- Therefore, in standard Sense Amplifier block Psense Amplifiers are placed outside the Isolation device.
- Placement of the **Psense amplifiers outside the isolation devices provides a full one level (Vcc) directly available at the mbit (No need for Isolation device to pass the data to be written)**
- **EQ/bias transistors** are placed **outside of the ISO devices to permit continued equilibration** of digitlines in arrays that are isolated.
- The **I/0 transistor gate terminals** are connected to a **common CSEL signal for four adjacent digitlines.**
- **Each of the four I/0 transistors is** tied to a separate I/0 bus.
- This sense amplifier, though simple to implement, is somewhat larger than other designs due to the presence of two Psense amplifiers.



Standard sense amplifier block

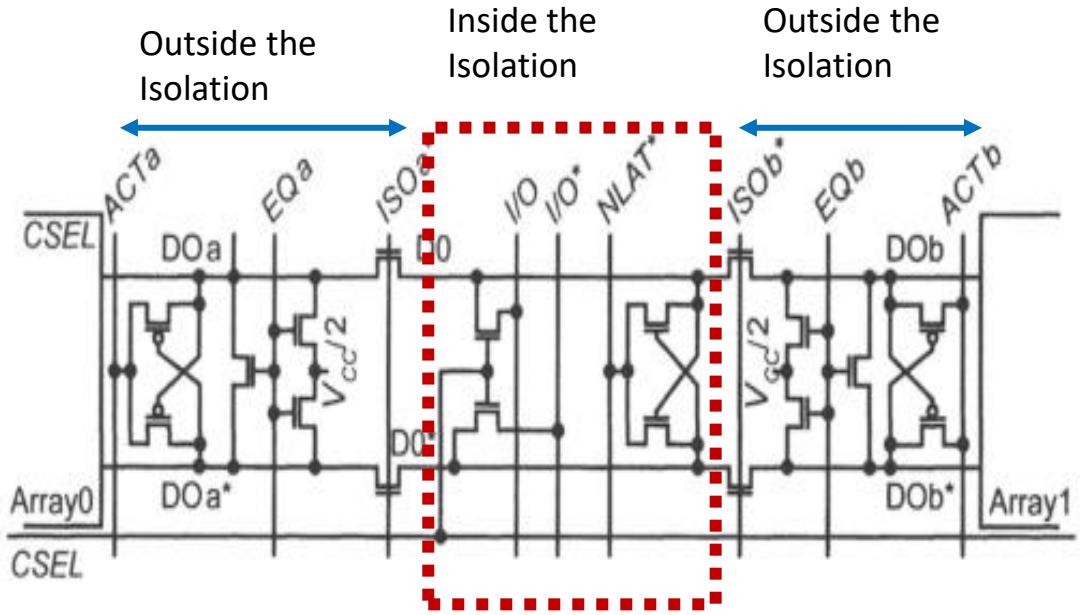
Why EQ/bias transistors are placed outside of ISO devices?

- **EQ/bias transistors** are placed **outside of the ISO devices to permit continued equilibration** of digitlines in arrays that are isolated.
- At a given point of time one side of the Active and performing memory accessing (ISOb*=1) but at the same time on the other side , the other array is isolated (ISOa*=0).
- Therefore the **EQ/bias transistors** placed outside Isolation can be turned ON so that, it continues **equilibration** of digitlines in arrays that are isolated.



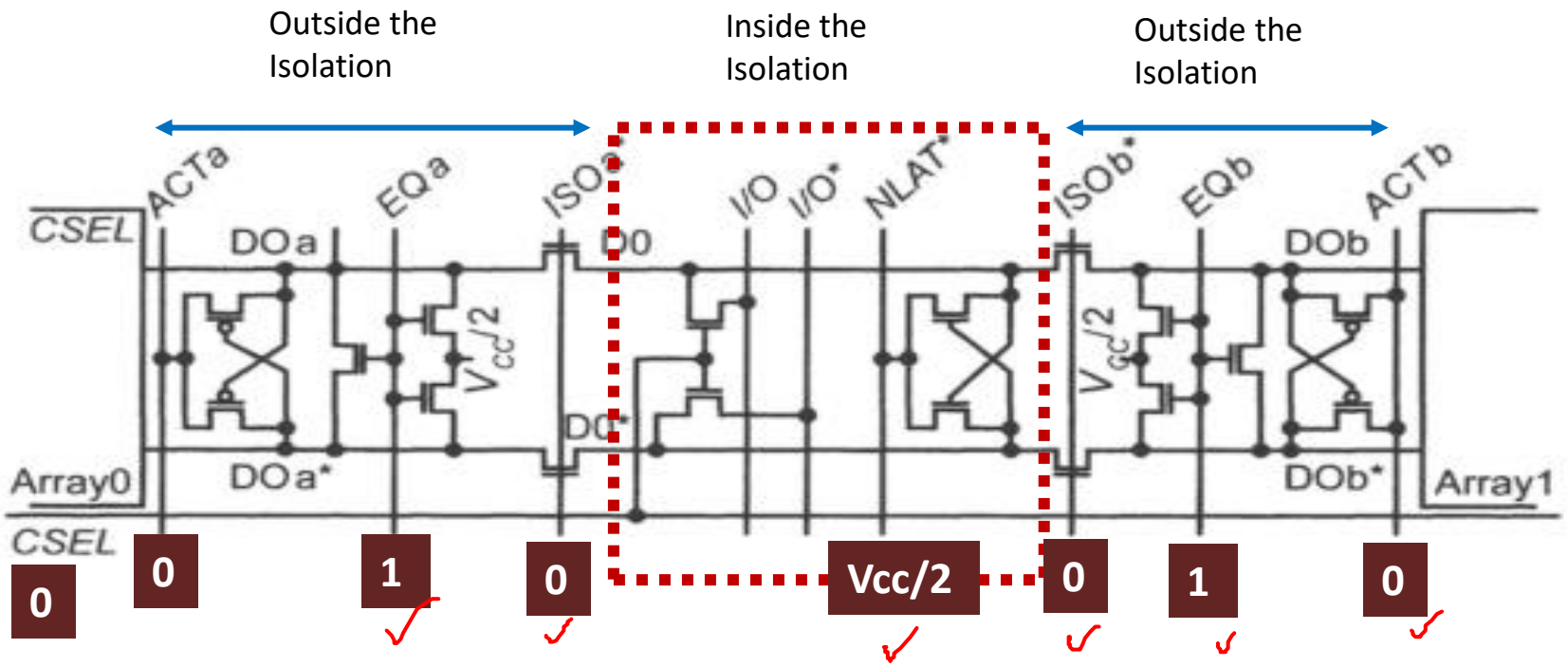
Standard sense amplifier block

- The **I/0 transistor gate terminals** are connected to a **common CSEL signal for four adjacent digitlines.**
- **Each of the four I/0 transistors is** tied to a separate I/0 bus.
- This sense amplifier, though simple to implement, is somewhat larger than other designs due to the presence of two Psense amplifiers.



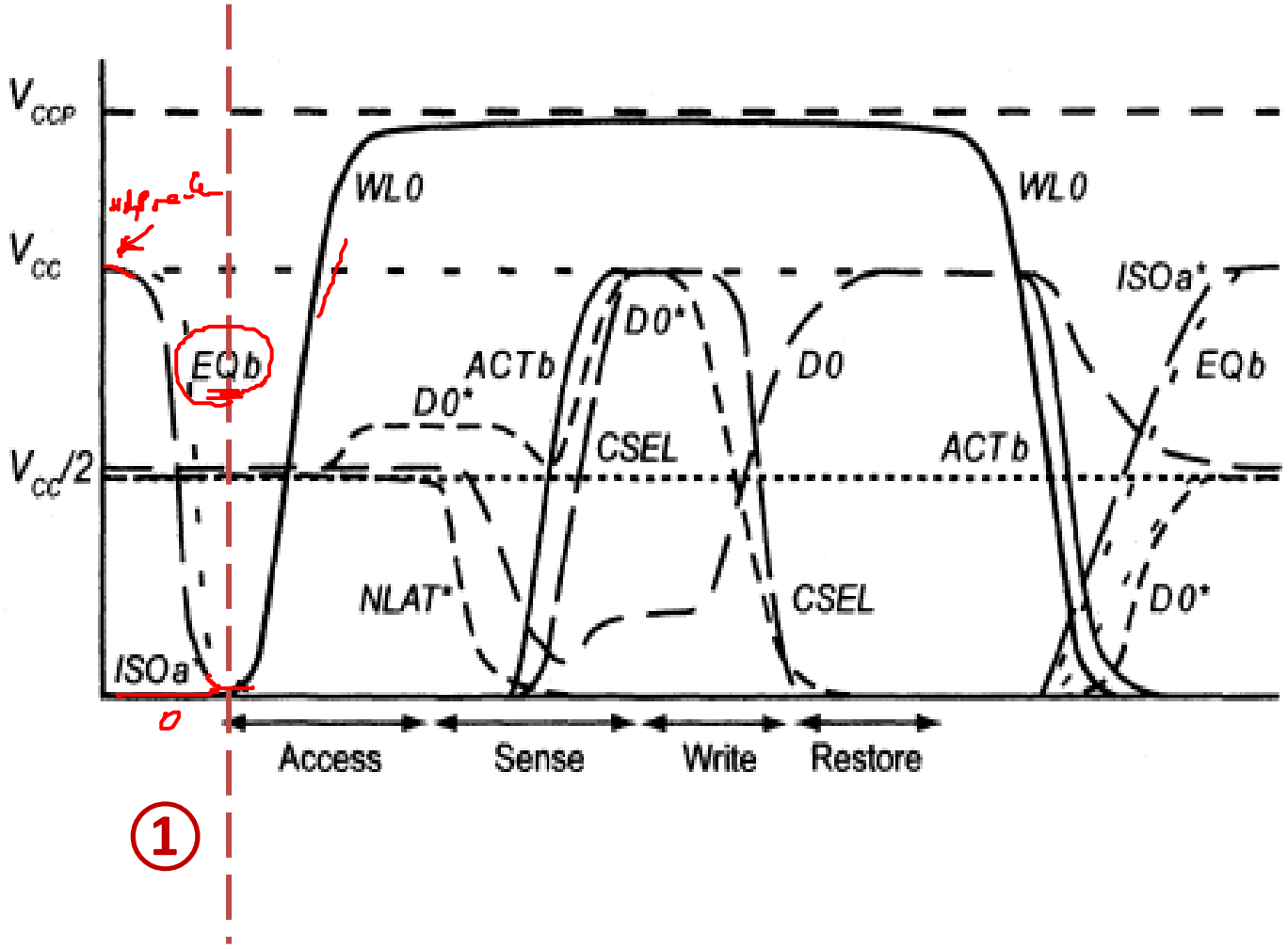
Standard sense amplifier block

Operation of Standard Sense Amplifier block



Standard sense amplifier block –Timing Diagram

Read-Modify-Write cycle – Reading From Array 1 and Writing New data



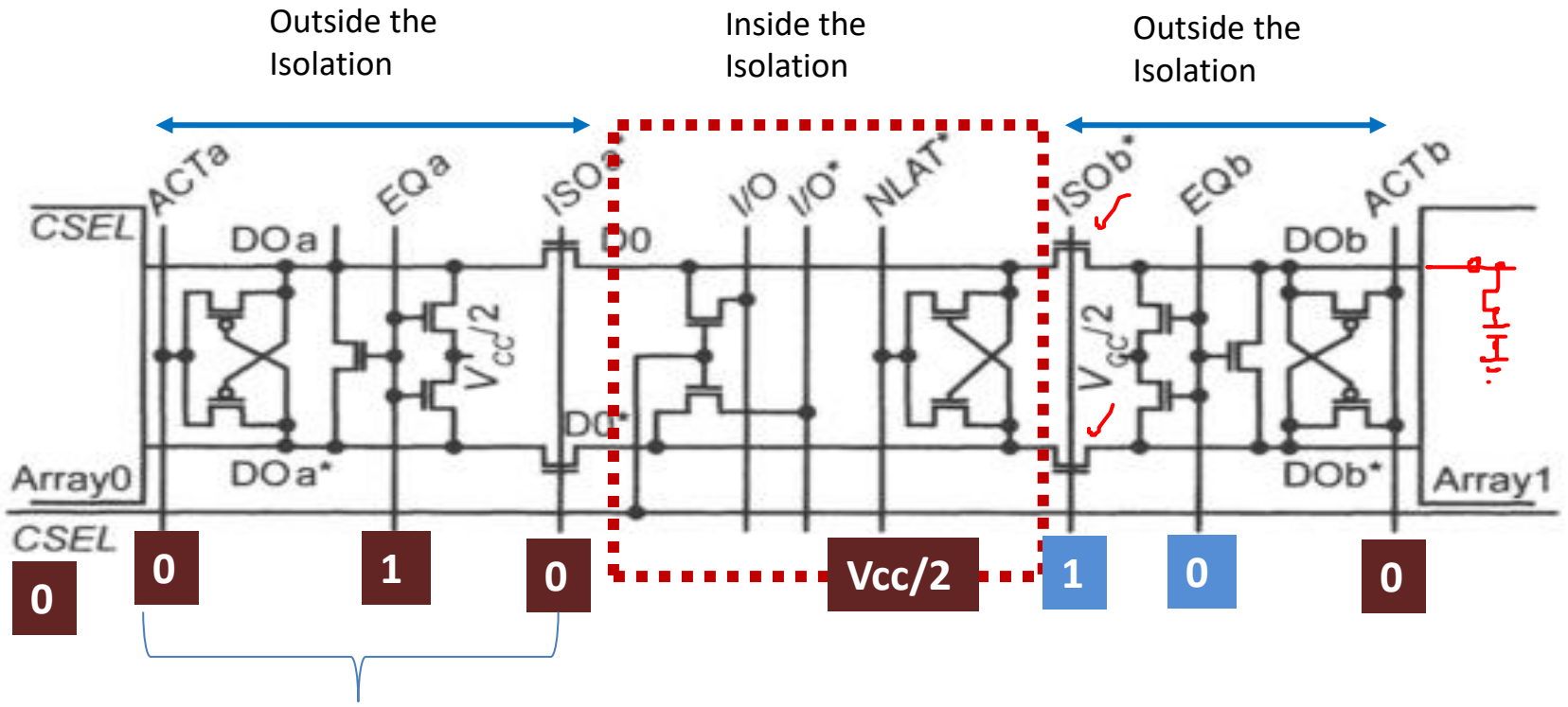
Region ① - Array 0 to be disabled

- WL0=Low; Row is Inactive
- EQa=EQb=HIGH – Precharge is ON and Digitlines of Array1 and Array0 are pre-charged to Vcc/2 (D0 and D0*)
- Sense Amplifier is disabled (ACT=GND and NLAT*= Vcc/2)
- **CSEL and Enable are at GND** potential
- Prior to firing of wordline in Array 1, the **ISOa is made LOW** to disable Array0 and EQb is made LOW to disable pre-charge circuit of Array 1.

Standard sense amplifier block

Operation of Standard Sense Amplifier block

Firing the wordline(WL=High) in Array1 to Read mbit



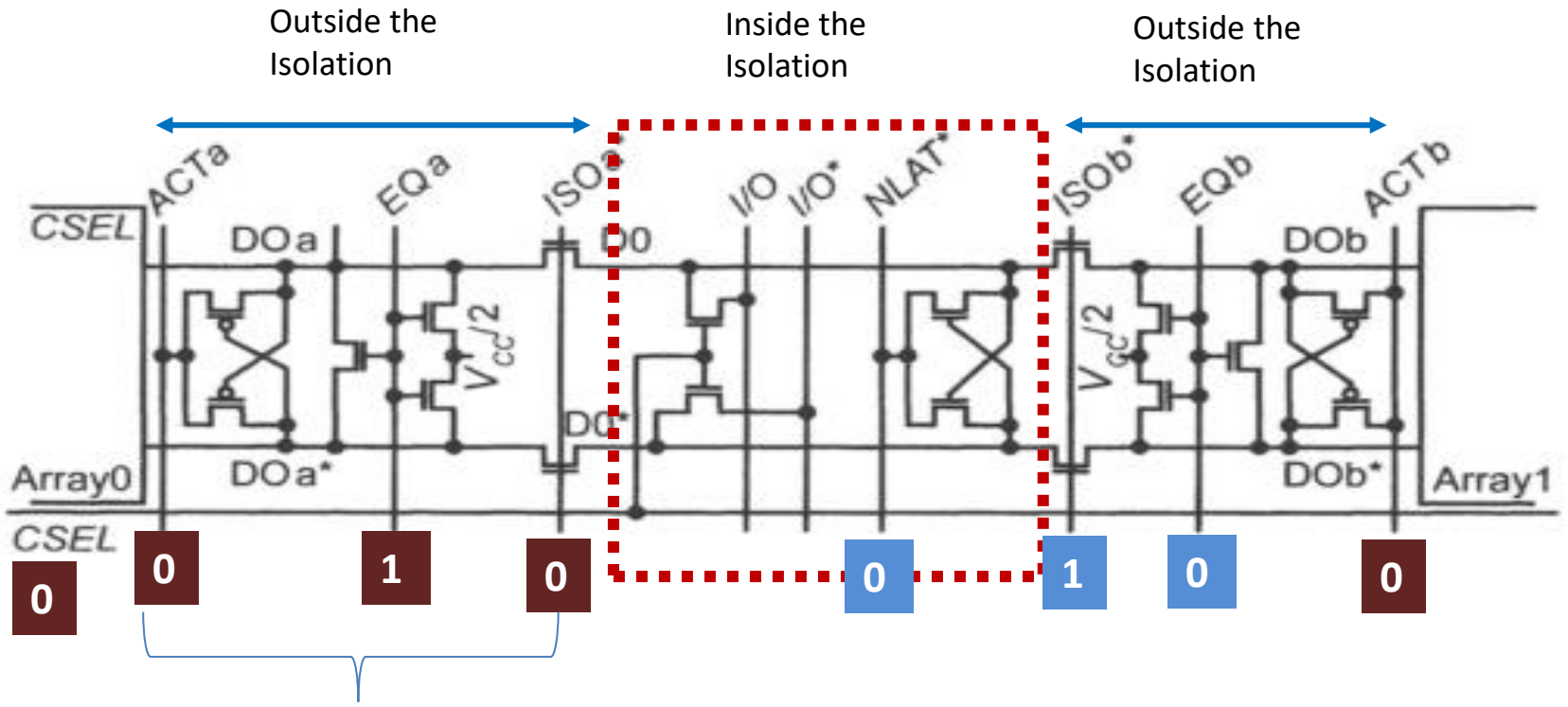
Isolated Array0 and EQ/bias is on to pre-charge D0a and Doa*

Access followed by Firing Sense Amplifier with Nsense first followed by Psense

Standard sense amplifier block

Operation of Standard Sense Amplifier block

Firing the wordline(WL=High) in Array1 to Read mbit



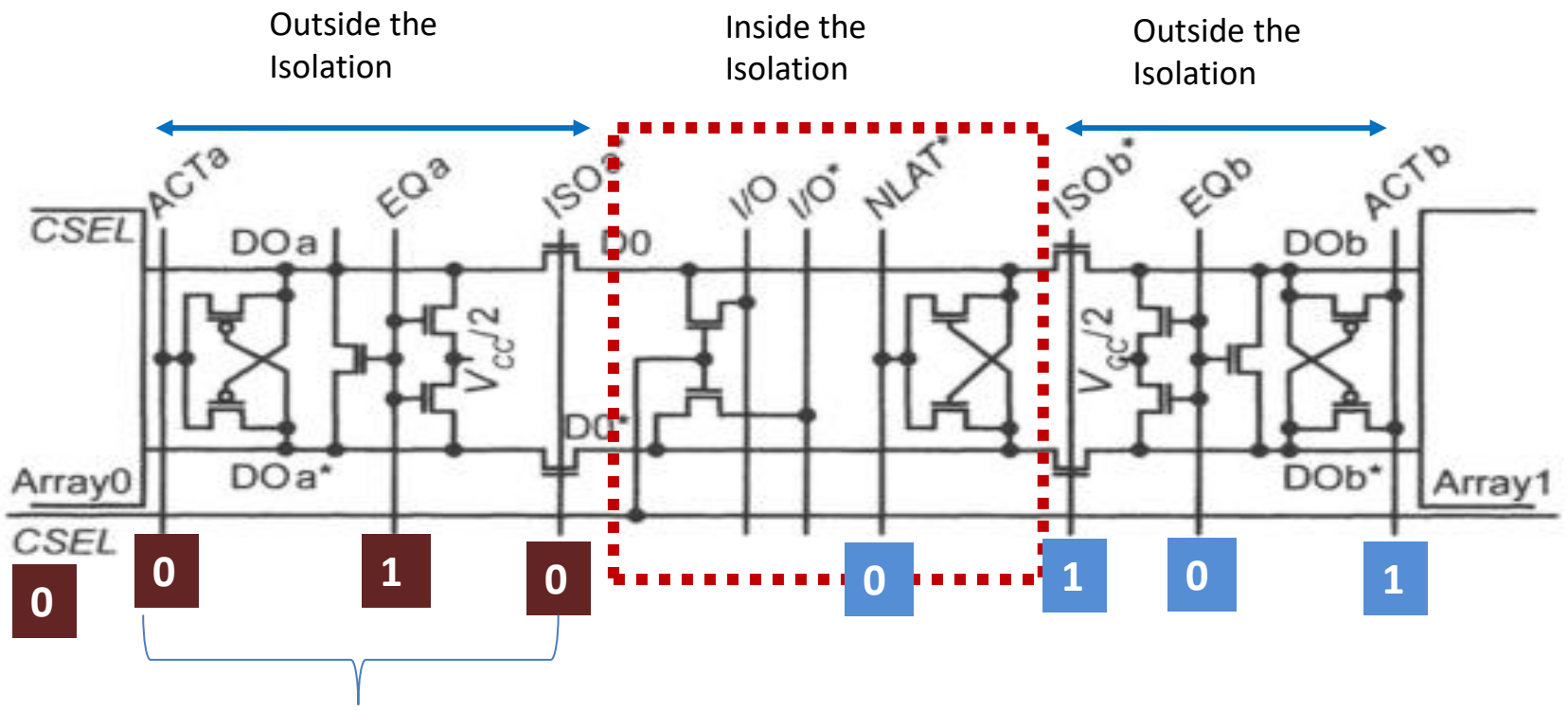
Isolated Array0 and EQ/bias is on to pre-charge D0a and Doa*

NLAT* is made 0; therefore D0 starts discharging towards zero.

Standard sense amplifier block

Operation of Standard Sense Amplifier block

Firing the wordline(WL=High) in Array1 to Read mbit



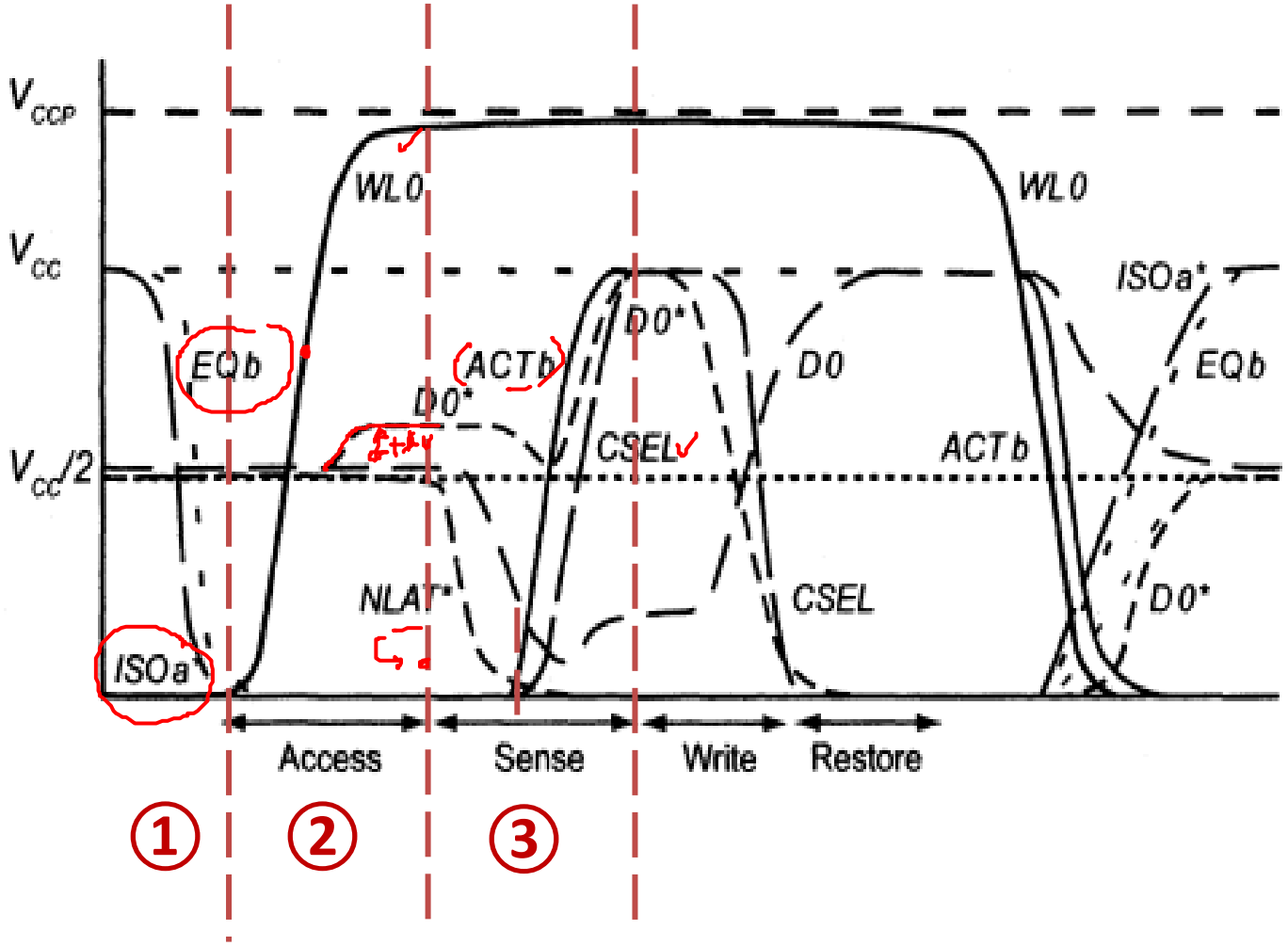
Isolated Array0 and EQ/bias is on to pre-charge D0a and Doa*

Psense is Fired by making ACT = Vcc; therefore D0* starts charging towards Vcc

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Standard sense amplifier block –Timing Diagram

Read-Modify-Write cycle – Reading From Array 1 and Writing New data



Region ② - Reading '0' from Array1

- **Read is initiated by firing** wordline of Array 1 (WL0=HIGH).
- Accessing an mbit, which **dumps charge onto D0* .**

Region ③ - Sensing and Reading

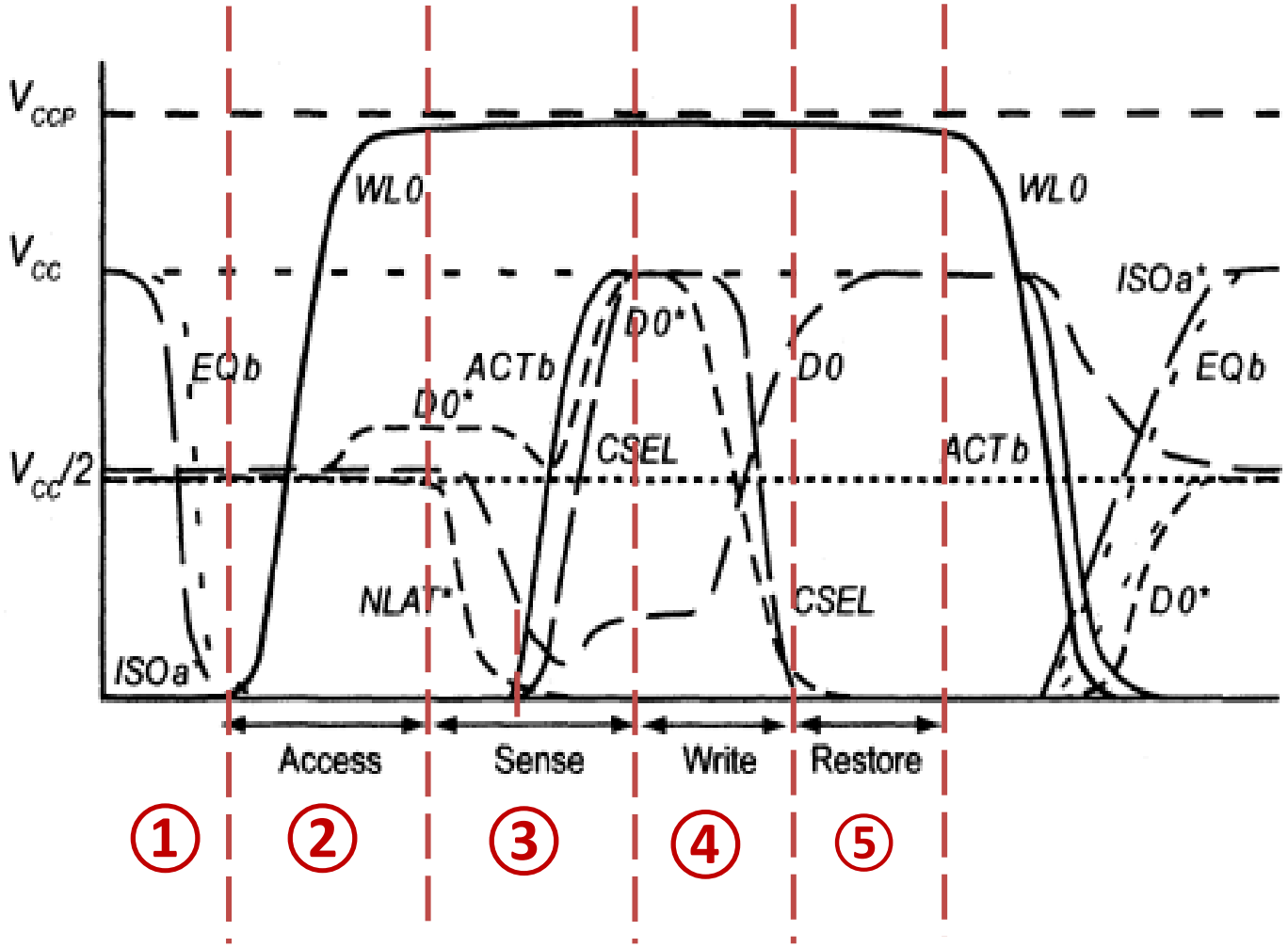
NLAT*, which is initially at Vcc/2, drives LOW to begin the Sensing operation and pulls DO toward ground. Then, ACT fires, moving from ground to Vcc, activating the Psense amplifier and driving DO* toward Vcc.

After separation has commenced, **CSEL0 rises to Vcc, turning ON the I/0 transistors** so that the cell data can be read by peripheral circuits.

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Standard sense amplifier block –Timing Diagram

Read-Modify-Write cycle – Reading From Array 1 and Writing New data



Region ④

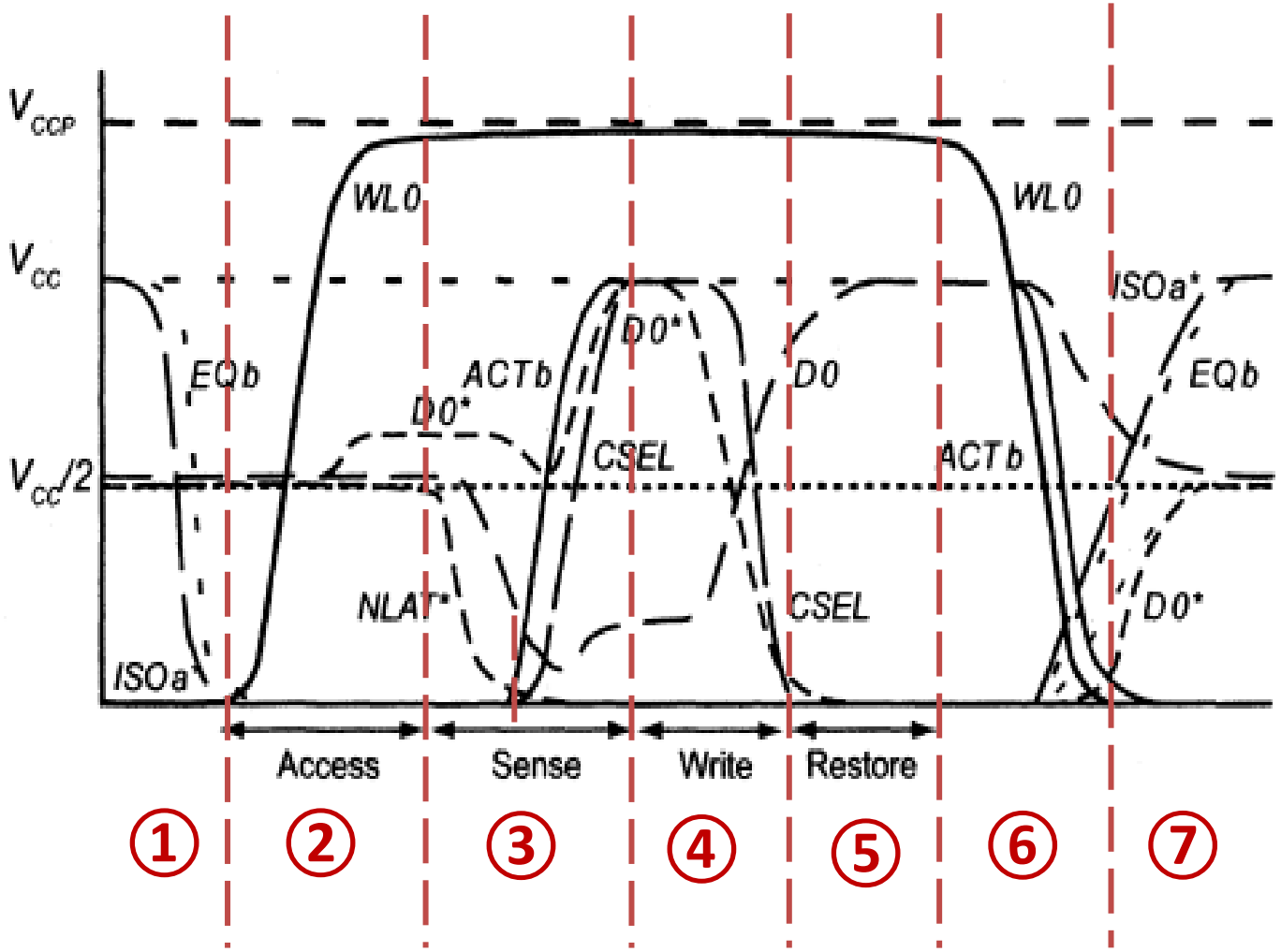
- After the Read is complete, **Write drivers in the periphery tum ON** (Enable =HIGH) and **drive the I/0 lines to opposite data states and writes over the existing data stored in the SA (D0*=0 and D0=1),** Data on D0* goes to mbit which is ON.
- SA latches the new data and It can be noted the when **CSEL and Enable is made LOW (Write Driver and I/O devices are shut down), This allows the SA to finish the restoring of digitlines to full level**

Region ⑤ - It stored data on D0* into mbit active in Array1

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Standard sense amplifier block –Timing Diagram

Read-Modify-Write cycle – Reading From Array 1 and Writing New data



Region ⑥ and ⑦

- The wordline (WL0) transitions LOW to shut OFF the mbit transistor.
- Finally, EQb= Vcc and ISOa*=HIGH to equilibrate the digitlines back to Vcc/2 and reconnect Array0 to the sense amplifiers respectively.
- **ACTb is made Zero (Psense is Turned OFF in Array 1) and NLAT* is made Vcc/2 (Nsense is turn OFF).**

Reference :DRAM Circuit Design -Fundamentals and High speed Topics by Brent Keert. R, Jacob Baker, Brian Johnson, Feng Lin

Standard sense amplifier block –Timing Diagram

- The timing for each event of Figure depends on
  - Circuit design,
  - transistor sizes,
  - layout,
  - device performance,
  - parasitics, and temperature.
- While timing for each event must be minimized to achieve optimum DRAM performance, it cannot be pushed so far as to eliminate all timing margins.
- Margins are necessary to ensure proper device operation over the expected range of process variations and the wide range of operating conditions.

THANK YOU

Mahesh Awati
Department of Electronics and Communication

mahesha@pes.edu

+91 9741172822

Memory Design and Testing

Mahesh Awati
Department of Electronics and Communication Engg.

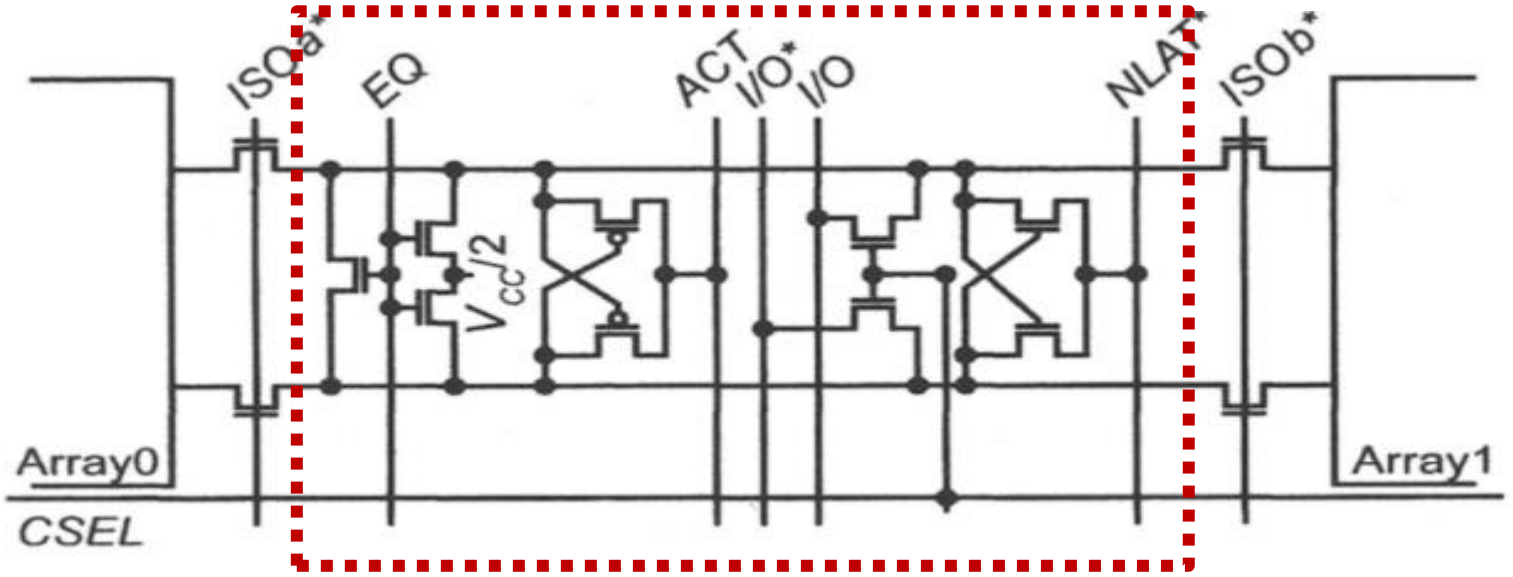
MEMORY DESIGN AND TESTING

UNIT 2 – Dynamic Random Access Memory

Mahesh Awati
Department of Electronics and Communication Engineering

Minimum Sense Amplifier block

- This block has Sense Amplifier (both Psense and Nsense) , I/O circuit and EQ/bias circuit Inside the Isolation. This reduces the Number of Psense to one, EQ/bias to one and Psense to one.
- Only a handful of designs operates with a single EQ/bias circuit inside the isolation devices.
- Historically, DRAM engineers tended to shy away from designs that permitted digitlines to float for extended periods of time.

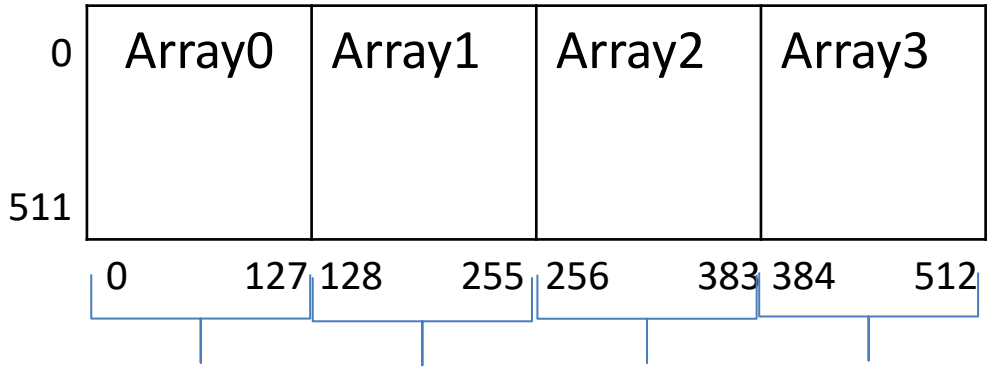


High Speed DRAM

- Suppose we have a 64-Meg DRAM organized as 16 Meg x 4 (4 bits input/output) using 4k row address locations and 4k column address locations (12 bits or pins are needed for each 4k of addressing).

B0	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**
B8							**B15**
B57							**B63**

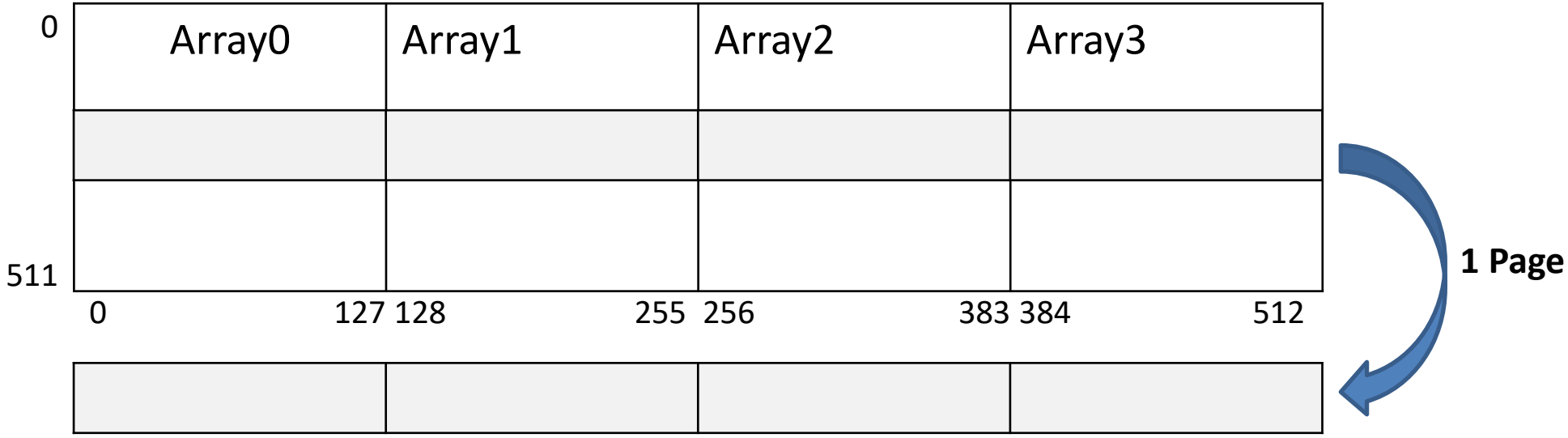
- $Log_2(4K)$ =12 Row address and 12 Column address.
- In 12 bits Row and column address, 3 bits are used for block addressing. Therefore, 9 bit Row address and 9 bit column address will be available.



128x4bit=512 bits

High Speed DRAM

- Each Array arranged as 512 Rows and 512 columns. We'll assume that there are 512 wordlines and 512 bitlines (digitlines), so that the memory array is logically square. (However, physically, as we shall see, the array is not square.)



- We **can define a page of data** in the DRAM by realizing that when we open a row in each of the four memory arrays, we are accessing 2k of data (4 arrays x 128columns x 4 bits =512 bits/array= 2Kbits).

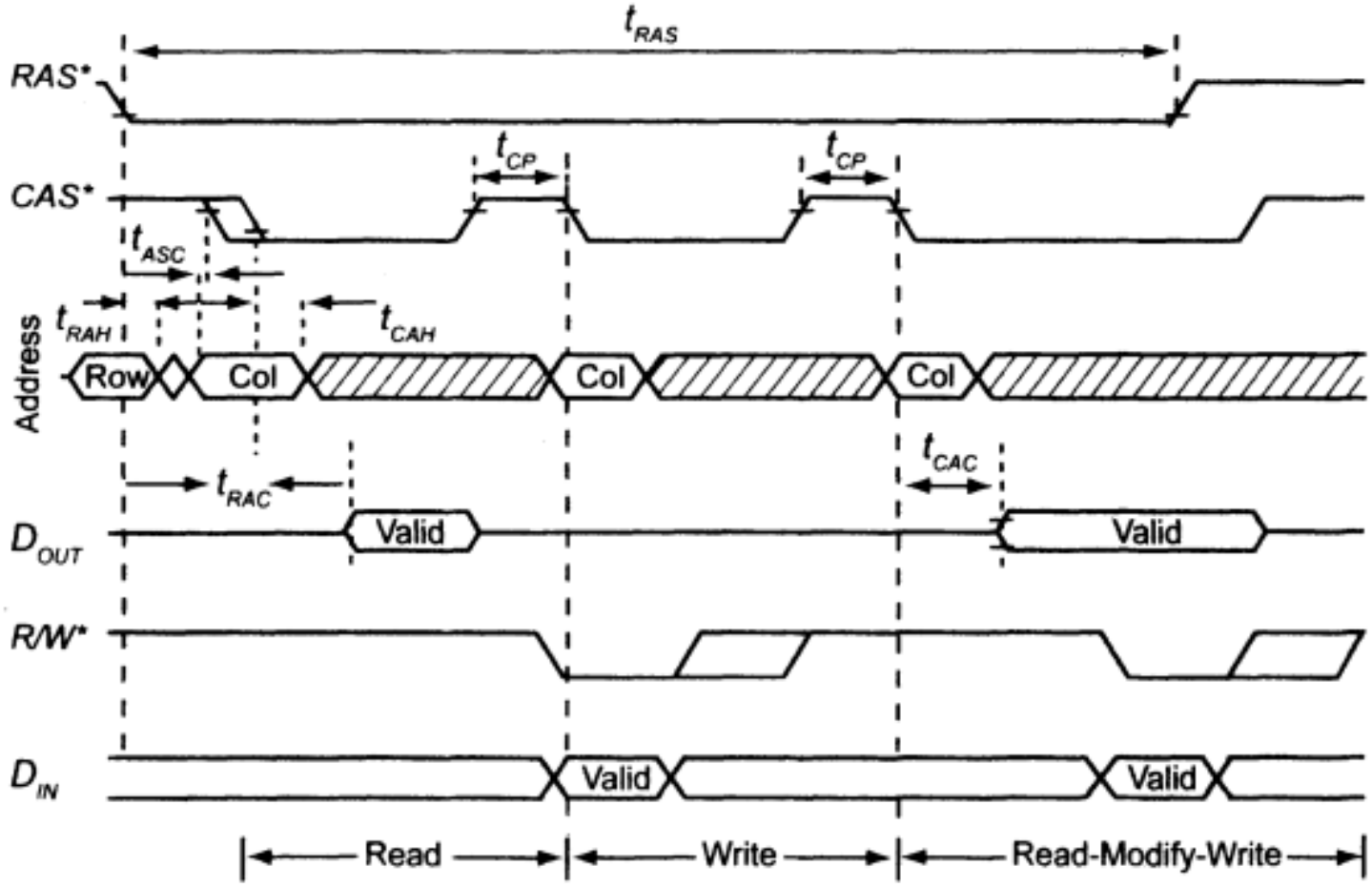
High Speed DRAM

- Now By **simply changing the column address without changing the row address and thus opening another group-of-four wordlines, we can access the 2k "page" of data.**
- With a little imagination, we can see different possibilities for the addressing.
- For example**, we could open 8 group-of-four memory arrays with a row address (8x2K=16K page) and thus increase the page size to 16k**, or we could use more than one bit at a time from an array to increase word size.

Group0	Group1	Group2	Group3	Group4	Group5	Group6	Group7
2K	**2K**	**2K**	**2K**	**2K**	**2K**	**2K**	**2K**

High Speed DRAM

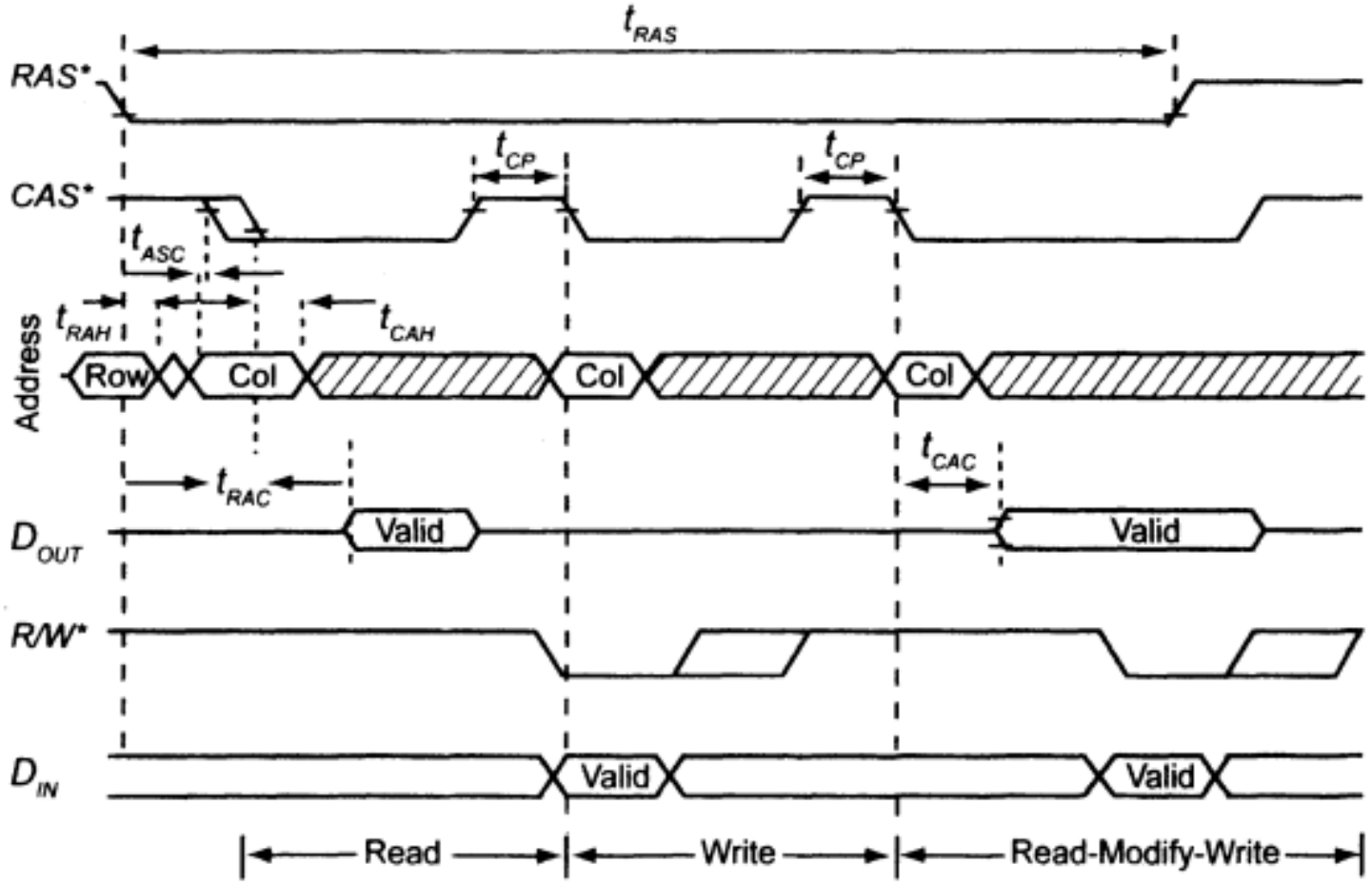
Page Mode : Read-Write- Reda_Modify_Write



- **First RAS* goes LOW,** we clock in a row address, decode the row address, and then drive a wordline in one or more memory arrays to Vccp·
- One row can be opened in any single array at a time.
- Also notice at this time that data out, Dout, is in a Hi-Z state;

High Speed DRAM

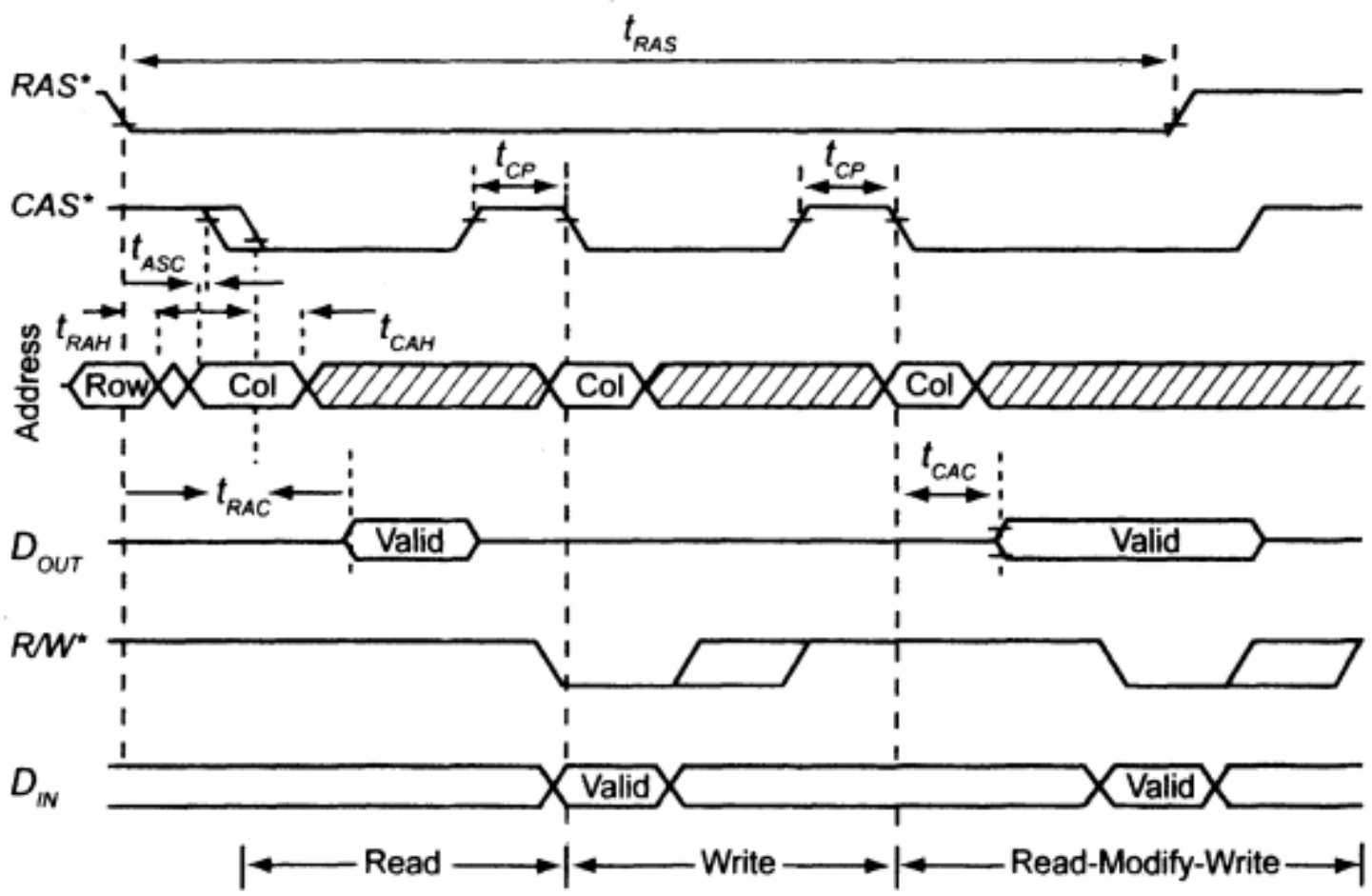
Page Mode : Read-Write- Reda_Modify_Write



- **Next CAS* goes LOW** and the column address is clocked into the DRAM.
- The column address is decoded, and, assuming that the data from the open row is sitting on the digitlines, it is steered using the column address decoder to Dout.
- It might have an opened row of 512 bits, but we are steering only one bit to Dout-
- Notice that when CAS* goes HIGH, Dour goes back to the Hiz state
- By strobing in another column address with the same open row, we can select another bit of data to steer to the Dour pin.

High Speed DRAM

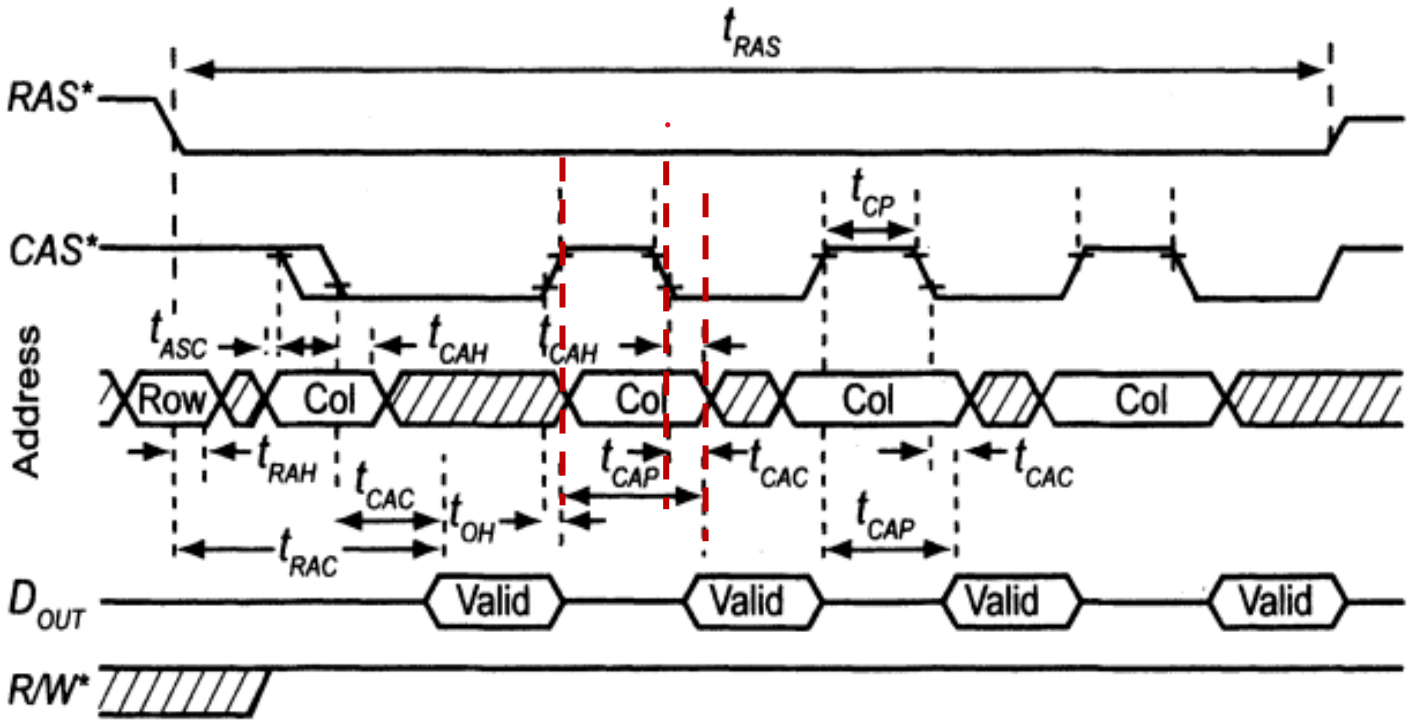
Page Mode : Read-Write- Reda_Modify_Write



- In the timing diagram, the DRAM is changed to the Write mode. This allows write operation with the same row open via the DIN pin to any column address on the open row.
- The final set of timing signals **read data out of the DRAM with RIW* HIGH, change RIW* to a LOW, and then write to the same location**. Again, when CAS* goes HIGH, Dout goes back to the Hi-Z state.

High Speed DRAM

Fast Page Mode : In this Column column address to propagate into the column circuits while CAS* is HIGH



FPM allows the column address to propagate into the column circuits while CAS* is HIGH. The speed of the DRAM thus improves by reducing the delay between CAS* going LOW and valid data present, or accessed, on Dout (t_{CAC})

SDRAM

- SDRAMs are potentially faster than 2 nd generation DRAMs such as FPM and EDO.
- SDRAMs are divided into banks and use pipelining in their data paths. Hence it is possible to
  - Activate a row in one bank and then,
  - while the row is opening, perform an operation in some other bank (such **as reading or writing**).
- In addition, one of the banks can be in Precharge Mode, while accessing one of the other banks, in effect, hiding PRECHARGE and allowing data to be continuously written to or read from SDRAM

DDR-RAM

- DDR DRAMs are faster than SDRAMs since data is transferred over **2 clock edges.**
- SDRAMs are also referred to as SDR SDRAM. Thus the transfer rate of **DDR DRAM is double of SDR** DRAM without increasing the clock frequency.
- For Ex: A DDR2 device with 8 data pins (x8) prefetches 32 bits of data during each Read operation. Each data pin (DQ pin) drives out 4 bits of data serially over 2 full clock cycles – one data bit per clock edge. Prefetch buffer size is 4 bits. Prefetch buffer size in DDR DRAM is 2 bits whereas in DDR3 DRAM, it is 8 bits
- Thus if a clock of frequency 133 MHz is used, then the transfer rate for each of the DRAMs is
    1. SDRAM – 133 MT/sec M(Megha)T – 10^6words/sec
    2. DDR SDRAM – 266 MT /sec
    3. DDR2 SDRAM – 533 MT /sec
    4. DDR3 SDRAM – 1066 MT /sec

THANK YOU

Mahesh Awati
Department of Electronics and Communication

mahesha@pes.edu

+91 9741172822