EE671: VLSI Design
Spring 2024/25

Laxmeesha Somappa

Department of Electrical Engineering

IIT Bombay

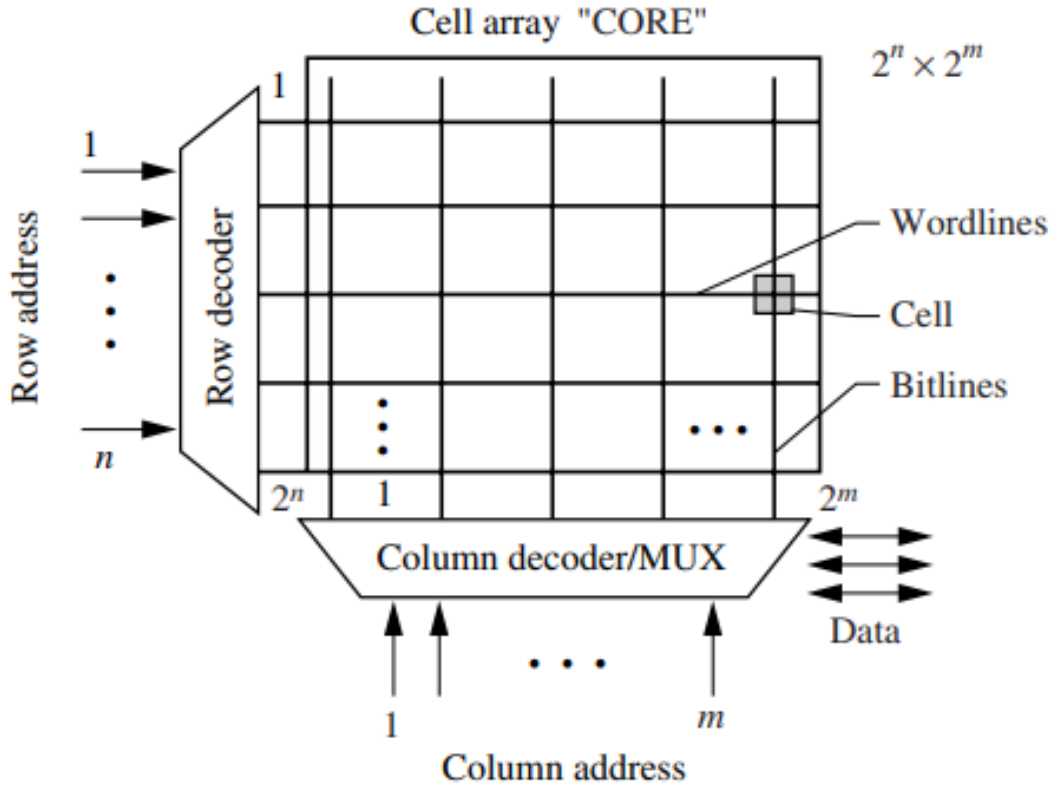
laxmeesha@ee.iitb.ac.in

LECTURE – 32
MEMORY: SRAM

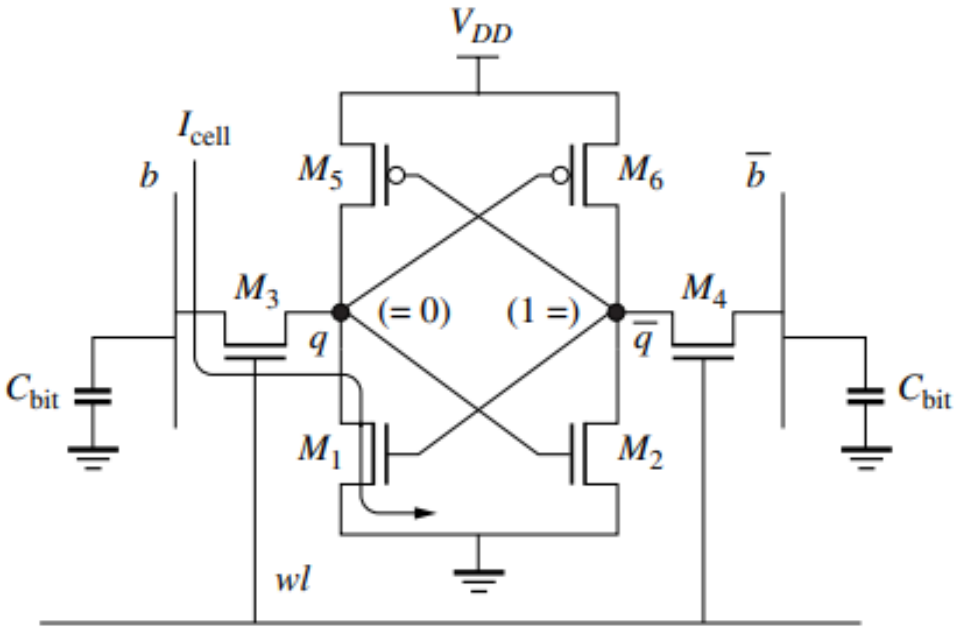
laxmeesha@ee.iitb.ac.in

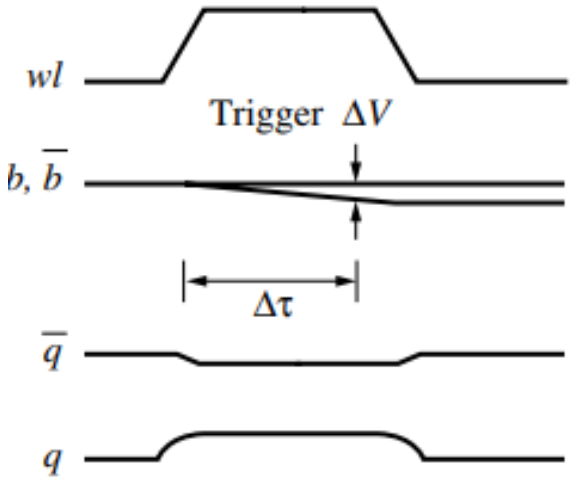
SRAM FULL PICTURE

- Implications:
  - If n is very large (tall memory) → the shared bitline per column is longer → this is a long metal in the layout → large bit-line capacitance → memory will be slower
  - If m is large (wide memory) → the wordline per row is longer → large wordline capacitance (because of routing and also the 2 access transistors gate cap per column) → Address decoder will drive larger load → delay limited by logical effort !!!!



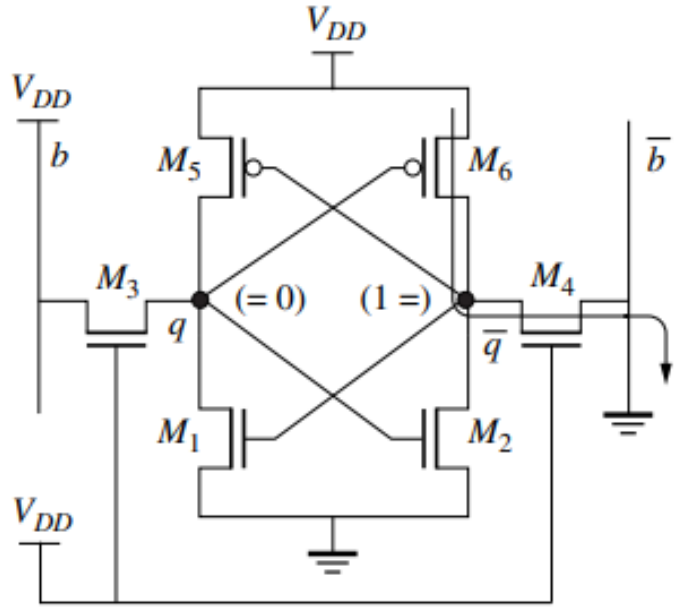
READ OPERATION

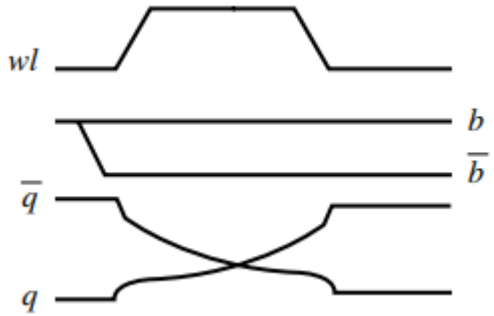




- M3 and M1 are actually fighting for node "q" → if M3 is strong it can turn on M2 !
- During read, we do not want to disturb the state of node "q"
  - i.e, change in node voltage "q" should not flip the previously stored data !
- To ensure the this, M3 and M1 will have to be sized accordingly
  - i.e., M1 should be stronger than M3 !! (M1 will have lower resistor than M3)

WRITE OPERATION

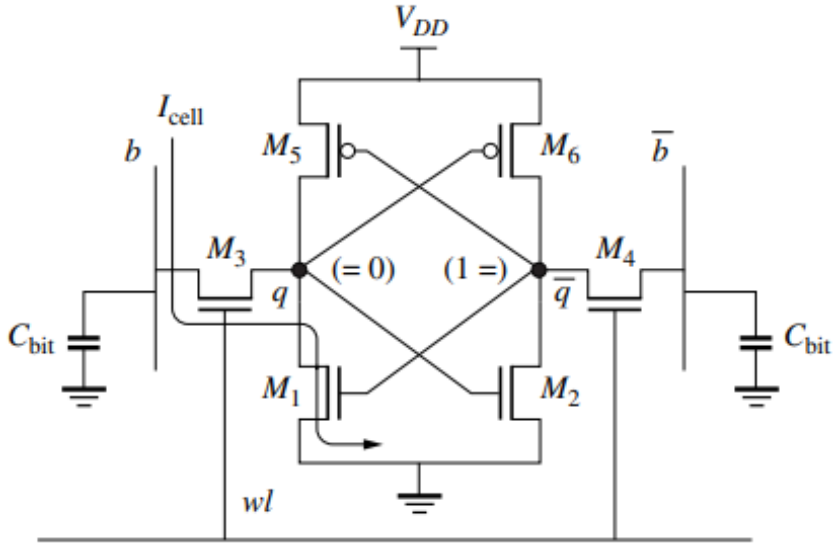


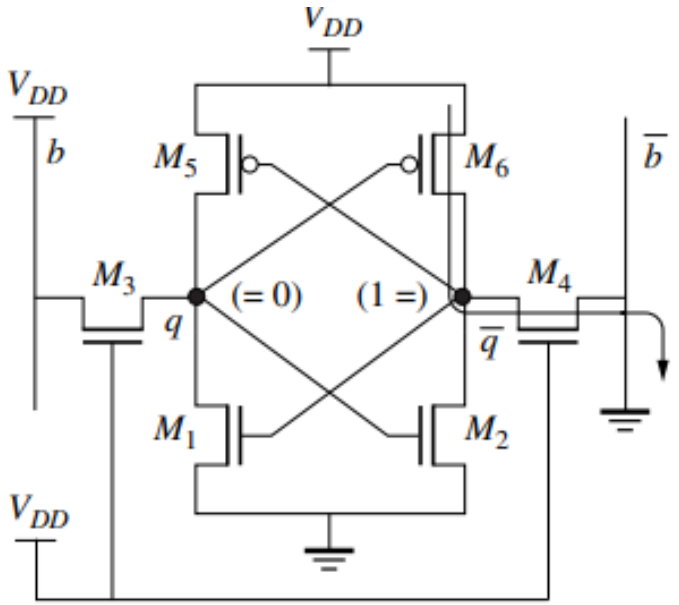


- To write new data → we need to overpower the existing cell data (if bits are different)
- During write → tie either BL or BLb to VDD/GND
- Say we want to write "1" into a cell that was earlier holding "0"
- In this case, M6 and M4 will fight to control node voltage at $\overline{q}$
- As a design criteria → M4 must be stronger than M6 !! (ie, M4 should have much smaller resistance compared to M6)

BIT-CELL SUMMARY

- Read operation dictates: M1,M2 stronger than M3, M4
- Write operation dictates: M3,M4 stronger than M5, M6
- Read:
  - Pre-charge bitlines (b and $\overline{b}$) to V_{DD} before a read operation (to remove history)
  - Fire the WL
  - Amplifier amplifies the change in difference voltage (between b and $\overline{b}$) to logic level
- Write:
  - b and $\overline{b}$ are tied to the write value (0,1 or 1,0)
  - Fire the WL
  - New data written to cell (pre-charge and amplifier not needed during write)



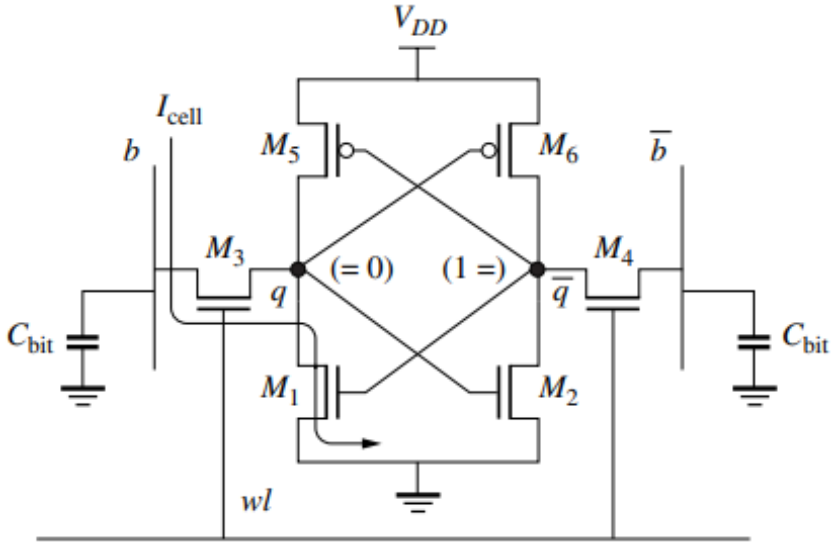


DESIGN EXAMPLE: READ OPERATION

- Need to design SRAM bit cell in a technology with:
  - L = 0.1 µm, V_{TN} = | V_{TP} | = 0.4 V, V_{DD} = 1.2 V, μ_n = 300 cm²/V-s, μ_p = 100 cm²/V-s, saturation velocity $v_{sat} = 8 \times 10^6$ cm/s, C_{ox} = 2 µF/cm²
- Design specifications of the SRAM bit cell are:
  - Total bit line capacitance = 1 pF
  - Amplifier needs minimum of ±200 mV difference to amplify output to logic levels within 1 ns
  - The bit-cell inverter nodes can only tolerate a change of 100 mV during read
- Design equations:
  - MOS triode current, $I_D \approx \mu Cox \frac{W}{L} [(V_{GS} - VT)V_{DS} - \frac{V_{DS}^2}{2}]$
  - MOS saturation current, $I_D \approx vs_{atCox} W(V_{GS} - VT)$
    - Under the assumption, for short channel MOSFETs, the velocity of the charges are saturated and hence the current will be lower than the current given by the quadratic equation

DESIGN EXAMPLE: READ OPERATION

- Let us first consider the read operation
- q = 0 is stored in the bit-cell
- The bitlines (b and $\bar{b}$) are pre-charged to V_{DD}
- What MOSFETs are ON?
  - M3 → is in saturation
  - M1 → is in triode
  - A current I_{cell} has to flow through these MOSFETs (discharging C_{bit})



- Equate the two currents:

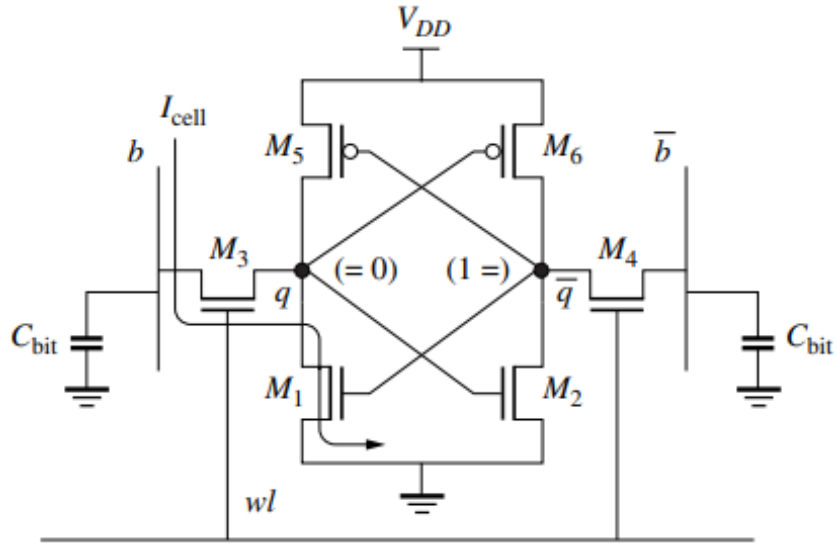
$$\mu Cox \frac{W_1}{L}\left[(V_{DD} - VT)V_q - \frac{V_q^2}{2}\right] = vs_{atCoxW3}(VD_D - Vq - VT)$$

- Substitute L = L_{min}, μ, v_{sat}, V_{DD}, V_T, V_q = 0.1 V (specs)

$$\frac{W_1}{W_3} \approx 2.5$$

Design Example: Read operation

- Let us first consider the read operation
- q = 0 is stored in the bit-cell
- The bitlines (b and $\overline{b}$) are pre-charged to V_{DD}
- Consider the next spec, amplifier needs at least 200 mV
- Node b should discharge from V_{DD} to (V_{DD}–0.2) in 1 ns
- Difference voltage between b and $\overline{b}$ will be 0.2 V
- We need to find out how much current is required to achieve this in 1 ns

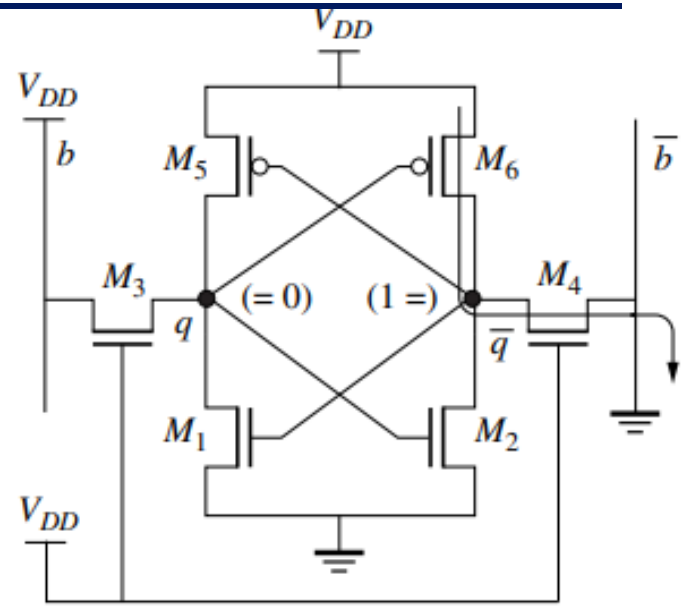


$$I_{cell} = \frac{\Delta V}{\Delta T} Cb_{it}$$ (what other insights from this equation? → tall memory?)

- ΔV = 0.2 V, ΔT = 1 ns, C_{bit} = 1 pF (all from the specs)
- I_{cell} = 200 μA
- Substitute current in sat equation, $W_3 \approx 0.2\ \mu m$, $W_1 \approx 0.5\ \mu m$
- $W_4 = W_3$ and $W_2 = W_1$ and all $L = 0.1\ \mu m$

DESIGN EXAMPLE: READ OPERATION

- To size M_6 and M_5, lets look at the write operation
- Recall that to write a new data,
  - we need to flip the existing data in the bit cell
  - Remember M1 is stronger than M3 and will fight for node q
  - Node $\overline{q}$ must start discharging to logic 0 to write new data
  - For this to happen, M2 must start turning ON
    - M2 will decisively turn ON when M1 is OFF
    - M1 will turn OFF when node $\overline{q}$ is just below V_T (M_1 $V_{GS} = V_T$)

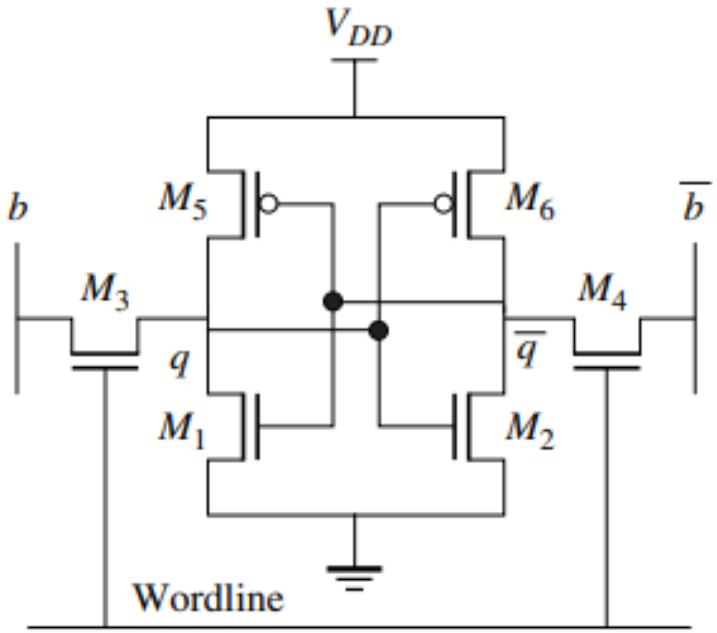


- What MOSFETs are ON?
  - M4 is in saturation
  - M6 is in triode
- Equate the currents and set $V_q = V_T = 0.4$ V

$$\frac{W_4}{W_6} \approx 0.5$$

Design Example

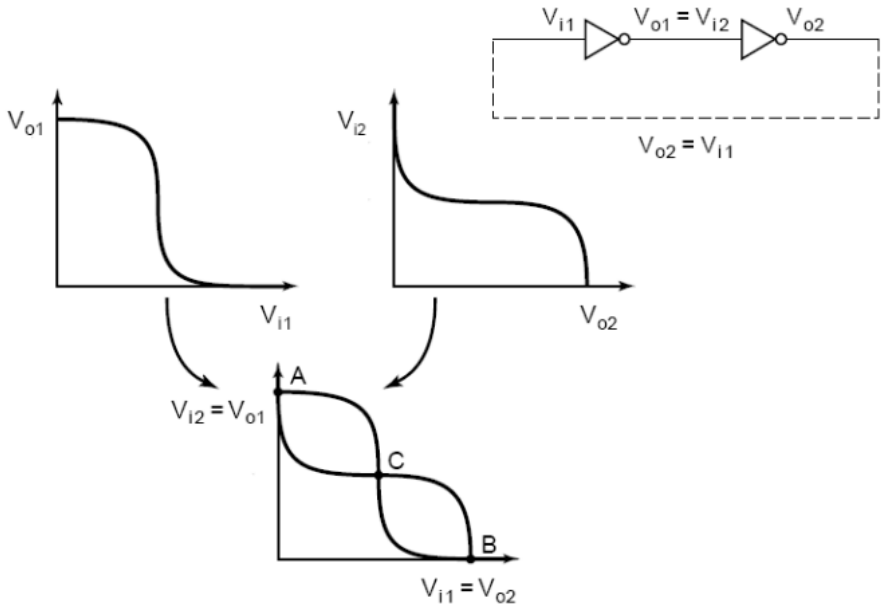
- $\frac{W_4}{W_6} \approx 0.5$
- $\frac{W_1}{W_3} \approx 2.5$
- $W_4 = W_3 = 0.2\ \mu m$
- $W_2 = W_1 = 0.5\ \mu m$
- $W_5 = W_6 = 0.4\ \mu m$
- L = $0.1\ \mu m$



- M1, M2 strongest
- M3, M4 stronger than M5, M6 and weaker than M1, M2
- M5, M6 weakest (account for the mobility of devices when talking about strength) – in this example, $\mu_n = 3\,\mu_p$
- A rule of thumb in lower nodes: strength ratio is kept at 1.5

TRANSFER CURVE

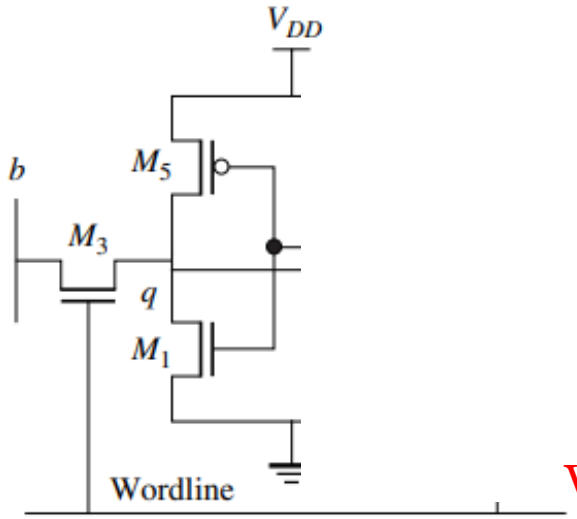
- For an SRAM bit cell, the transfer curve is also called as "Butterfly diagram"



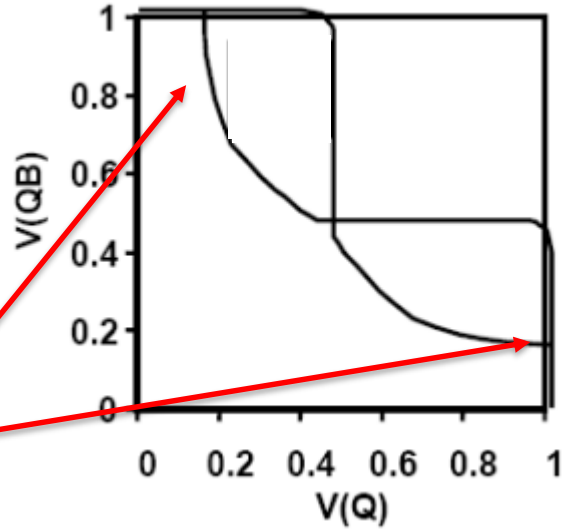
- Now that we have sized the MOS, if we plot the butterfly of the back-to-back inverter, we get a plot similar to the one above
- However, we now have a series access transistor → what is the impact of that?
- Let us look at the noise margin during read

Static Read Noise Margin (SNM)

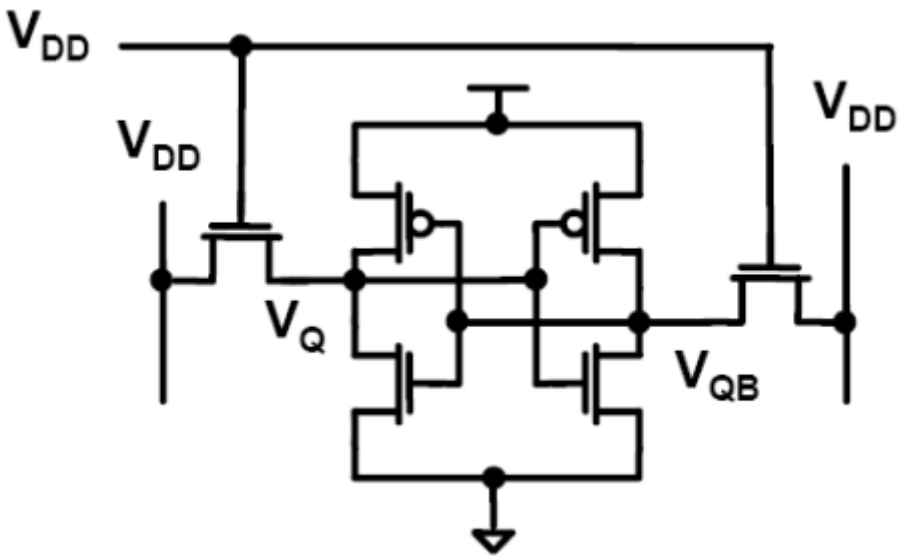
- Before a read, the bitlines are pre-charged to V_{DD}
- Similar to an inverter, we sweep the input of the inverter from 0 to V_{DD} (to plot the transfer curve)
  - But, with bit line at V_{DD} and WL at V_{DD}
- When input to the inverter is 0, output will be at V_{DD} → no issues
- When input to the inverter is V_{DD}, both M1 and M3 are ON
  - Even though M1 is stronger than M3, the node *q* will never really reach 0 V !!! → the transfer curve of this combination (during read) will be different than that of only inverter

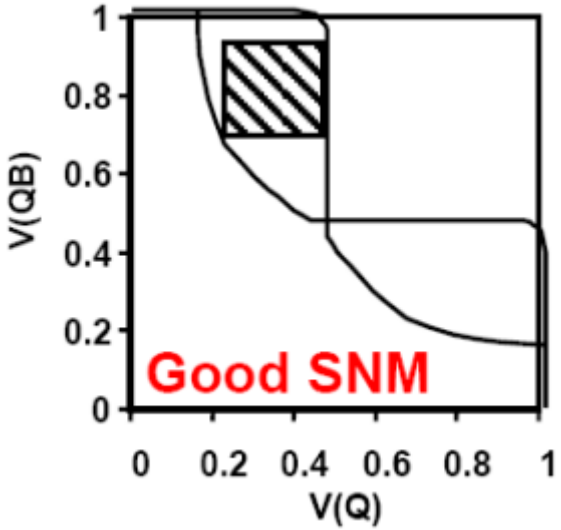


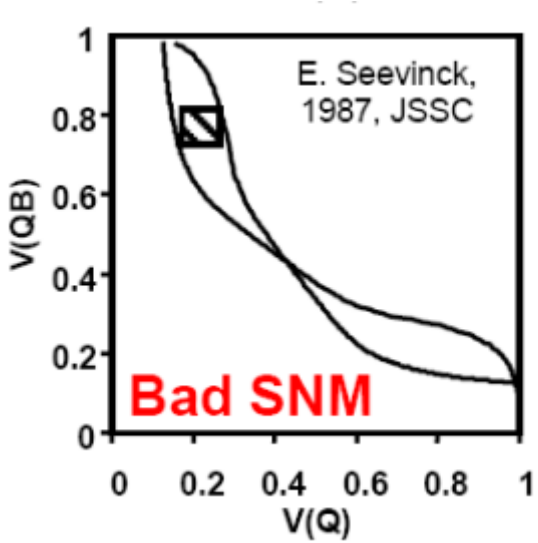
Voltages never go to zero



READ SNM

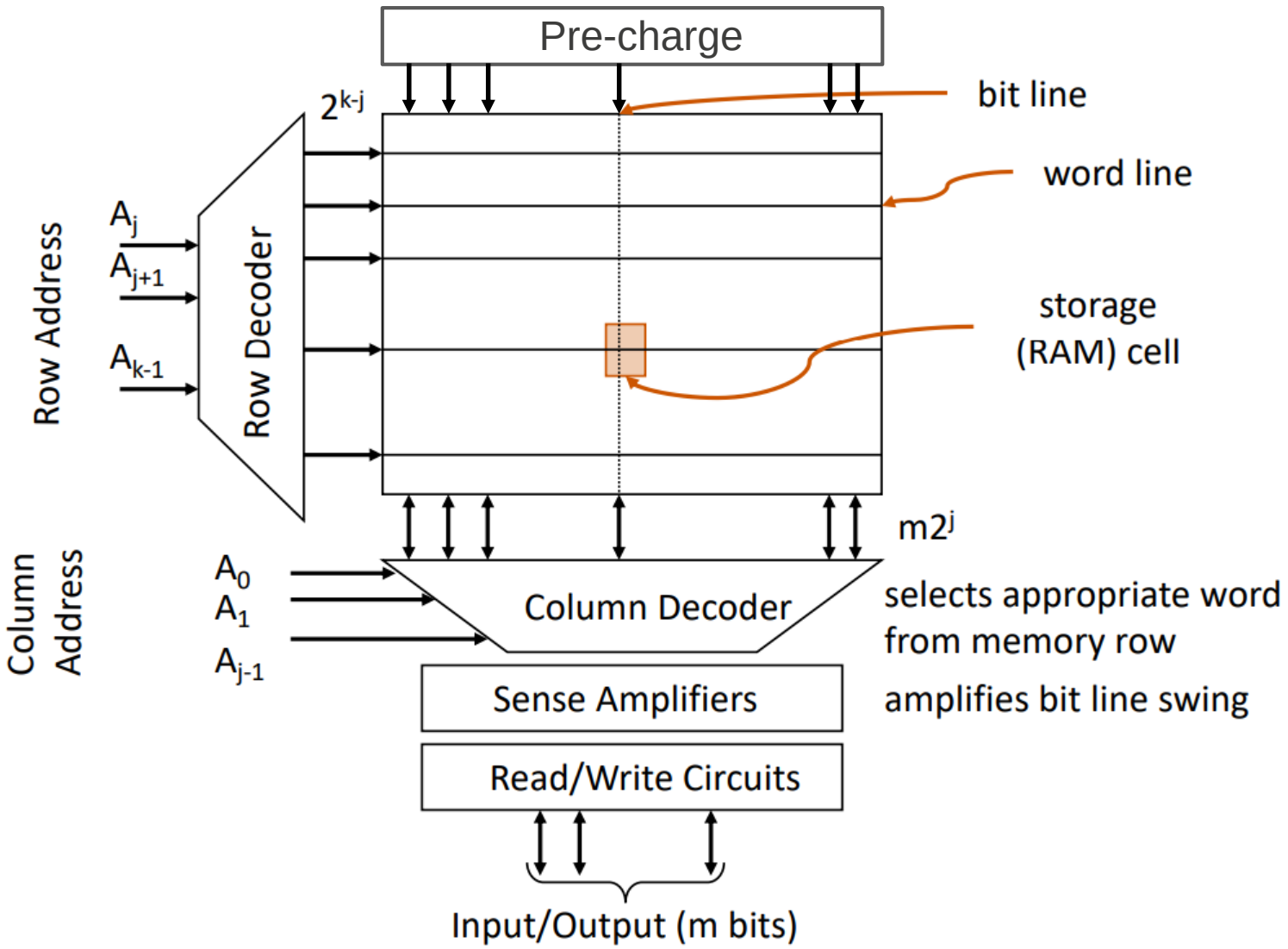






- Addition of access transistors → reduced the noise margin (compared to inverter)
- To avoid read destruction→ Static Noise Margin (SNM) must be increased
- Bigger the box you can fit inside the two transfer curves → better SNM !
- Design: start with the initial W values → keep optimizing to increase SNM
- But remember → entire idea behind SRAM is to reduce the area → we cannot size MOSFETs beyond a certain limit!

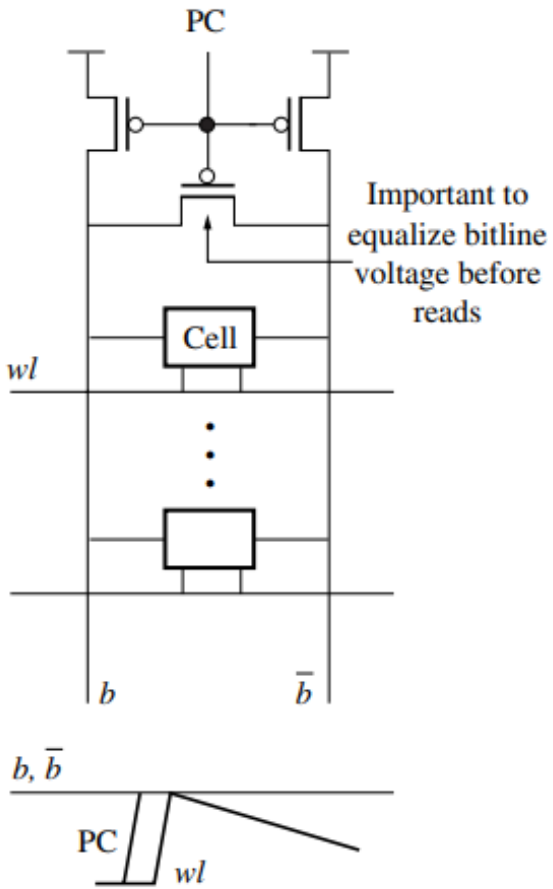
SRAM: Full Picture



- We have looked at the Row decoder (logical effort design)
- Looked at design of individual bit-cell
- Next, we will look at the pre-charge circuit, column decoders, sense amplifiers

COLUMN PULL-UPS: PRE-CHARGE

- To remove all history → perform pre-charge before every read and write
- The nature of pre-charge (PC) circuit depends on the amplification (also called sense amplifier (SA)) topology



Suitable for voltage based amplifiers

- Sequence:
  - PC is low by default (Pre-charging bit lines)
  - PC goes high
  - After some finite time → WL goes high
  - Read operation starts
  - Bitlines start developing delta voltage
  - Sense amplifier (SA) enabled after finite time to amplify this delta voltage
  - Once logic levels are obtained at the output, disable WL, SA and enable PC