ELEC 402

Memory Design - I

Lecture 13 & 14

Reza Molavi
Dept. of ECE
University of British Columbia
reza@ece.ubc.ca

Slides Courtesy : Dr. Res Saleh (UBC), Dr. D. Sengupta (AMD), Dr. B. Razavi (UCLA)

Semiconductor Memory

- **30% of the worldwide semiconductor business**
- **technology advances are driven by memory design**
- **intense "device-level" circuit design process**
- **digital and analog design issues**
- **It serves as a good application of all the material that we have covered in this course!**

Random Access

- access time indep. of physical location
- dominant form of semiconductor memory

ROM = read-only memory

RAM = read/write memory

(static RAM, dynamic RAM)

Serial Access

- shift registers, CCDs
- access time depends on location in memory

Associative

- content addressable memory (CAM)
- compare input tag to contents of memory
- used for pattern recognition applications
- more complex and costly than sRAM

Random Access Chip Strcuture

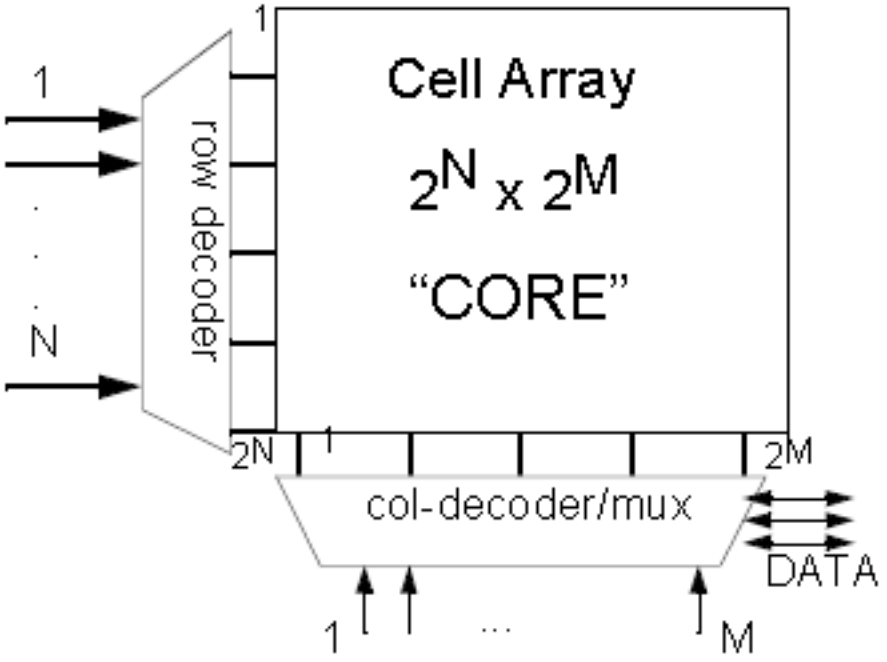
Address lines are split into row address and column address lines that feed "core".

CORE:
- keep square within a 2:1 ratio
- rows are **word lines**
- columns are **bit lines**
- data in and out on columns

DECODERS:
- needed to reduce total number of pins; N+M address lines for 2^{N+M} bits of storage

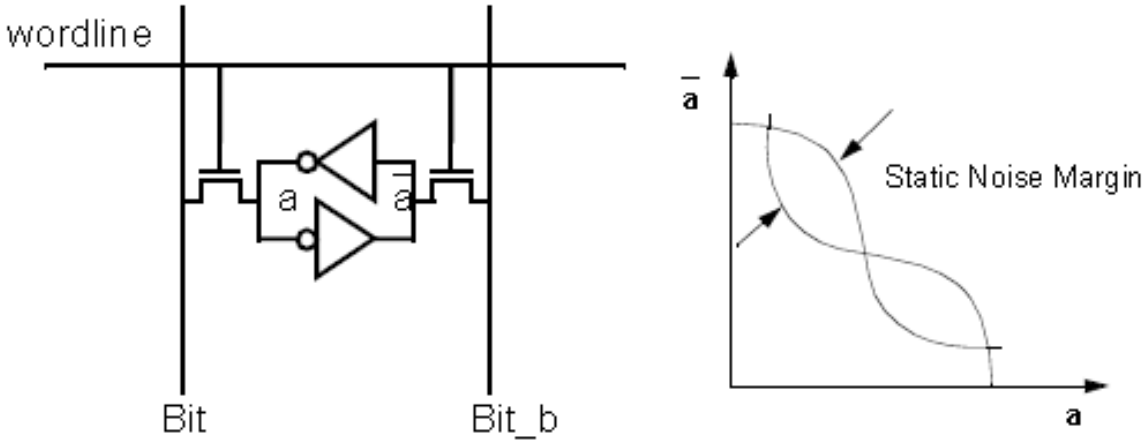
Ex: if N+M=20 → 2^{20}= 1Mb

MULTIPLEXING:
- used to select one or more columns for input or output of data



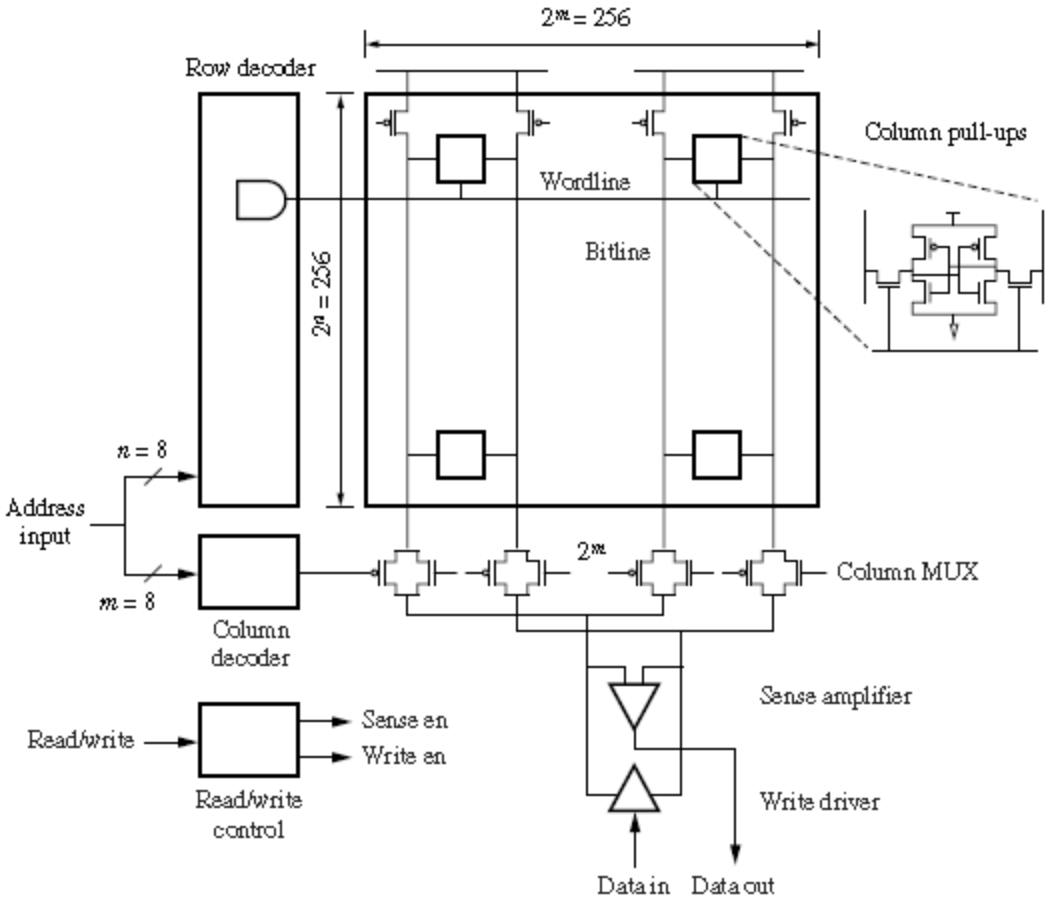
S-RAM Storage Cell

Uses only six transistors (called 6T cell):

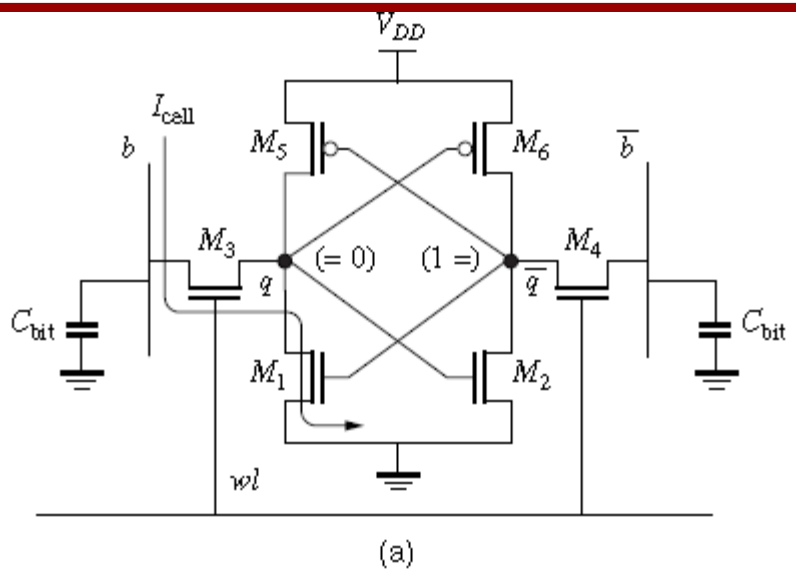


Read and write operations use the same port. There is one wordline and two bit lines. The bit lines carry complementary data. The cell layout is small since it has a small number of wires.

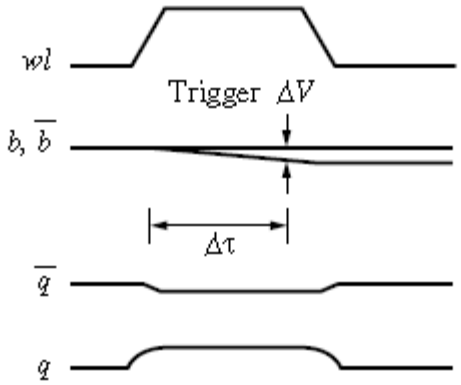
Memory Structure – Full Picture



Read Operation – Basic Steps



(a)



- The differential bitlines are precharged to vdd
- The wordline goes high high (select)
- The side connected to stored “0” (either b or b_bar) discharges a bit (note that M1 and M3 constitute saturated_load Inverter design)
- Small difference between b and b_bar is amplified to full rail

• Design must ensure the internal values are intact (they never get close to Vs of Transistors)

Read Operation- Example

Compute the widths of M_1 and M_3 in Figure 8.10 given that the circuit can only tolerate a rise in voltage of 0.1 V at node q during the read operation. Assume that $C_{bit} = 2$ pF and that the specification calls for a 200 mV transition of the bitline in 2 ns. Use 0.13 μm technology parameters.

When the wordline, *wl*, goes high, M_3 is a saturated enhancement load for the M_1 driver. The driver transistor is expected to be in the linear region of operation. Therefore, we can write the following equation:

$$\frac{W_1}{L_1}\frac{\mu_n C_{ox}}{\left(1+\frac{V_q}{E_{CN}L_1}\right)}\left[(V_{DD}-V_{T1})V_q - \frac{V_q^2}{2}\right] = \frac{W_3 v_{sat} C_{ox}(V_{DD}-V_q-V_{T3})^2}{(V_{DD}-V_q-V_{T3})+E_{CN}L_3}$$

Read Operation- Cont'd

$$\frac{W_1}{0.1\ \mu\text{m}} \frac{\left(270 \frac{\text{cm}^2}{V - \text{sec}}\right)}{\left(1 + \frac{0.1}{0.6}\right)} \left[(1.2 - 0.4)0.1 - \frac{0.1^2}{2}\right]$$

$$= \frac{W_3(8 \times 10^6\ \text{cm/s})(1.2 - 0.1 - 0.4)^2}{(1.2 - 0.1 - 0.4) + 0.6}$$

$$\therefore \frac{W_1}{W_3} \approx 1.7$$

This ratio would be smaller if body effect were taken into account. The actual values of the widths depend on the desired rate of change of the bitline voltage, the delay specification, and cell current. If we require a bitline transition of 200 mV in 2 ns, with a total bitline capacitance of 2 pF, then the cell current is

$$I_{\text{cell}} = C_{\text{bit}} \times \frac{\Delta V}{\Delta \tau} = 2\ \text{pF} \times \frac{200\ \text{mV}}{2\ \text{ns}} = 200\ \mu\text{A}$$

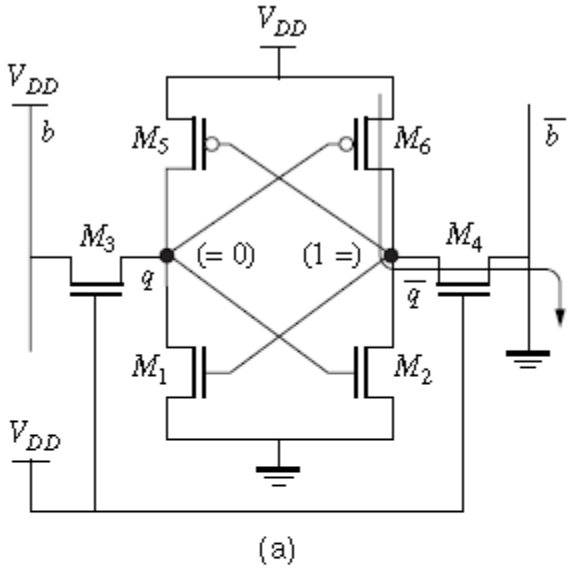
This is the average cell current through M_1 and M_3. As a rough estimate, we could simply use the current through the access transistor when it turns on:

$$I_{\text{cell}} \approx \frac{W_3(8 \times 10^6)(1.6\mu\text{F/cm}^2)(1.2 - 0.1 - 0.4)^2}{(1.2 - 0.1 - 0.4) + 0.6} = 200\ \mu\text{A}$$

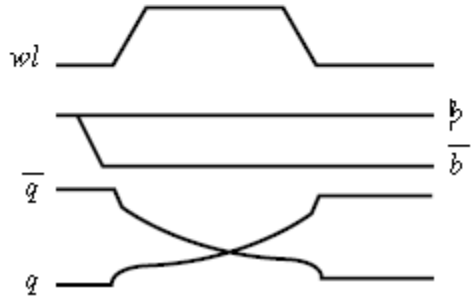
$$\therefore W_3 = 0.4\ \mu\text{m}$$

This implies that $W_1 = 0.7\ \mu$m. These two sizes are larger than we would desire if we were trying to create a 1 Mbit SRAM. However, this example is intended to show the steps in the design process.

Write Operation- Basic Steps



(a)



Let's say we want to write 1 into cell (that means 0 into q_bar)

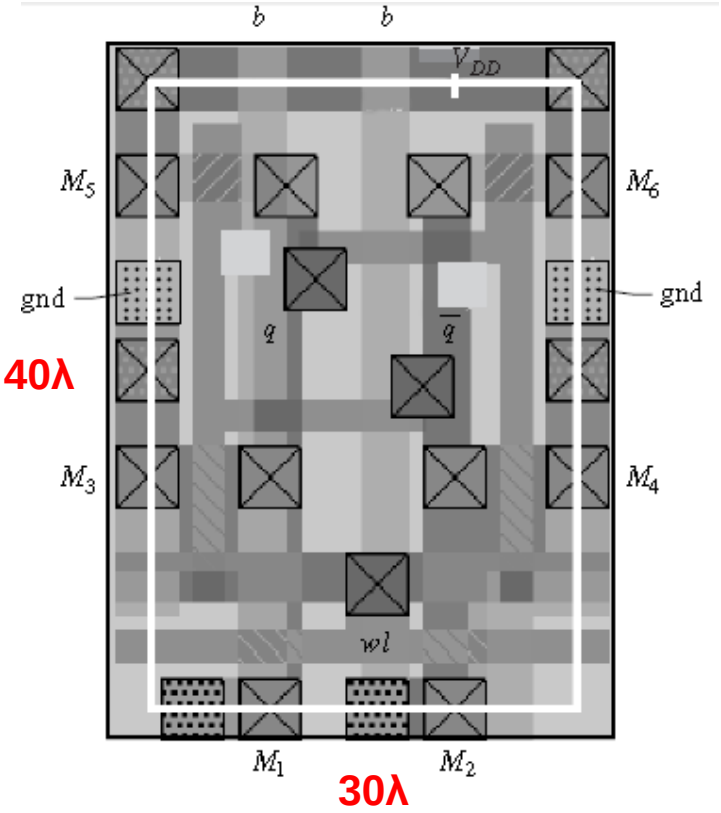
We set the b_bar to 0 and b to 1.

M6 and M4 constitute a pseudo-NMOS Structure and we should set the value of q_bar node to below Vs to make the write Work properly.

This translates to certain sizing between the devices:

$$\frac{W_4}{W_6} \approx 1.5$$ Typical size (not always)

SRAM Layout and I/O Circuitry



$$C_{word} = (2 \times \text{gate cap} + \text{wire cap}) \times \text{no. of cells in row}$$

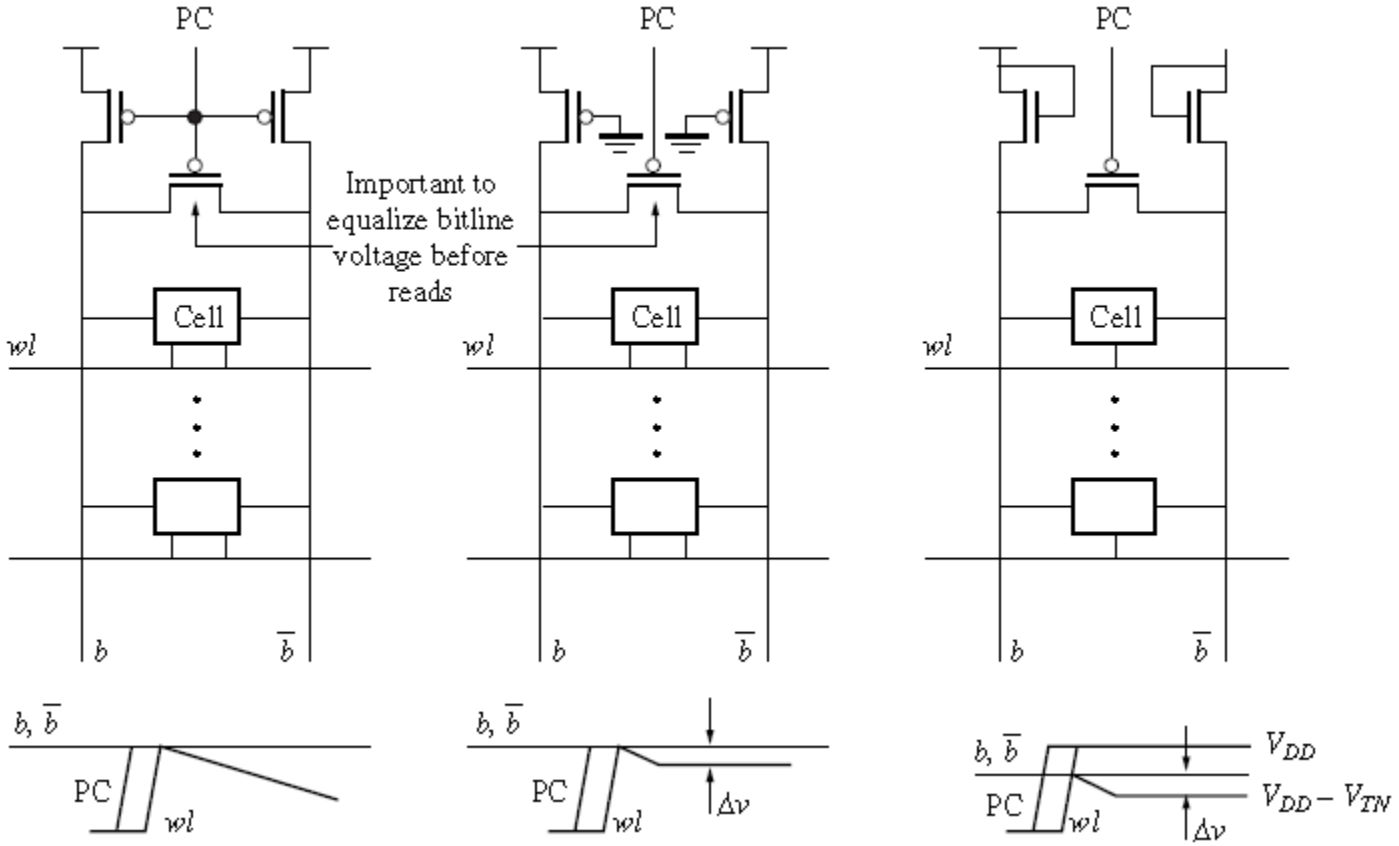
$$C_{bit} = (\text{source/drain cap} + \text{wire cap} + \text{contact cap}) \times \text{no. of cells in column}$$

- Proper calculations of caps are important for delay

The read and write operations mentioned so far require connection to outside world for Proper operation. Also all these operations should be done in timely manner and with proper Delay

SRAM I/O Circuitry

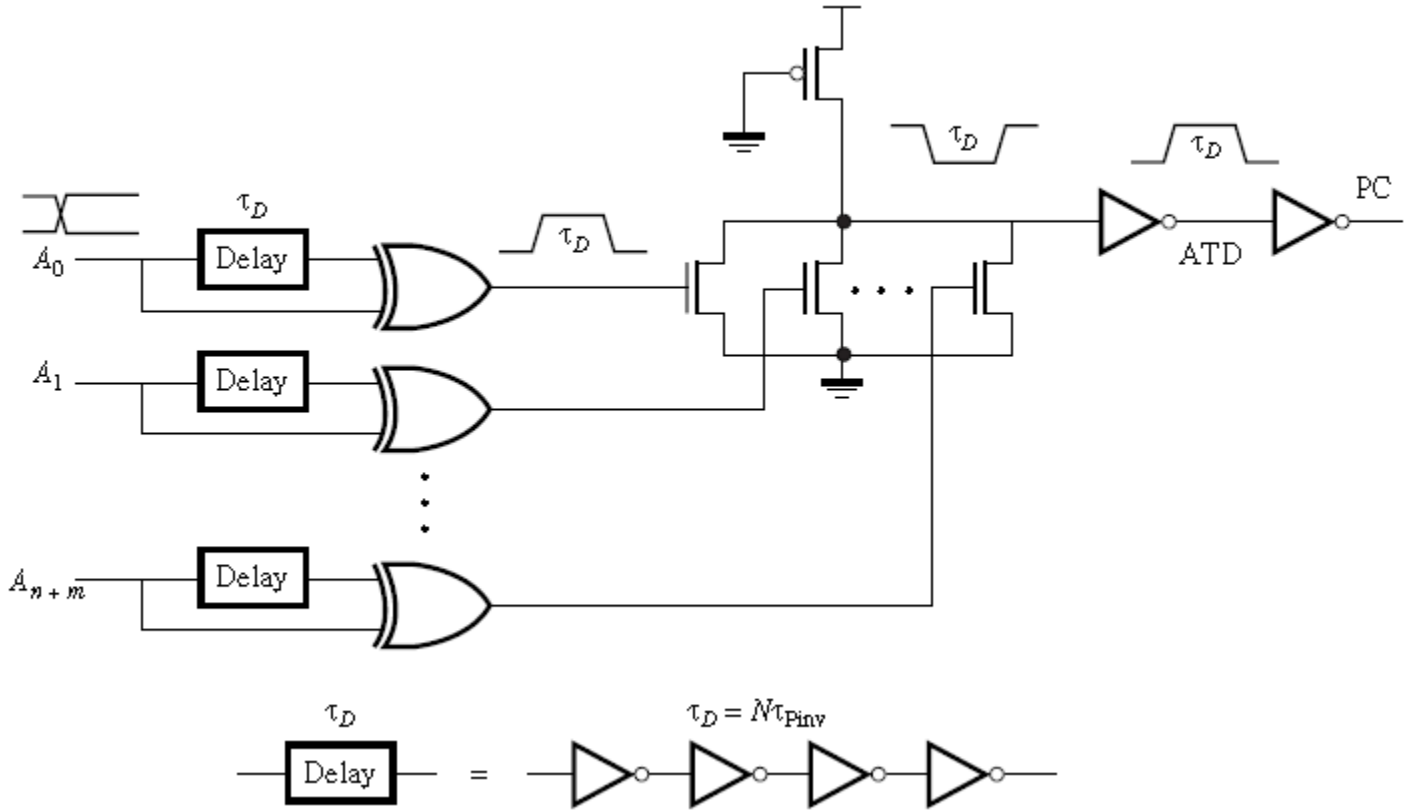
Different schemes of precharging bitlines



Suitable for latch-based sense amplifier

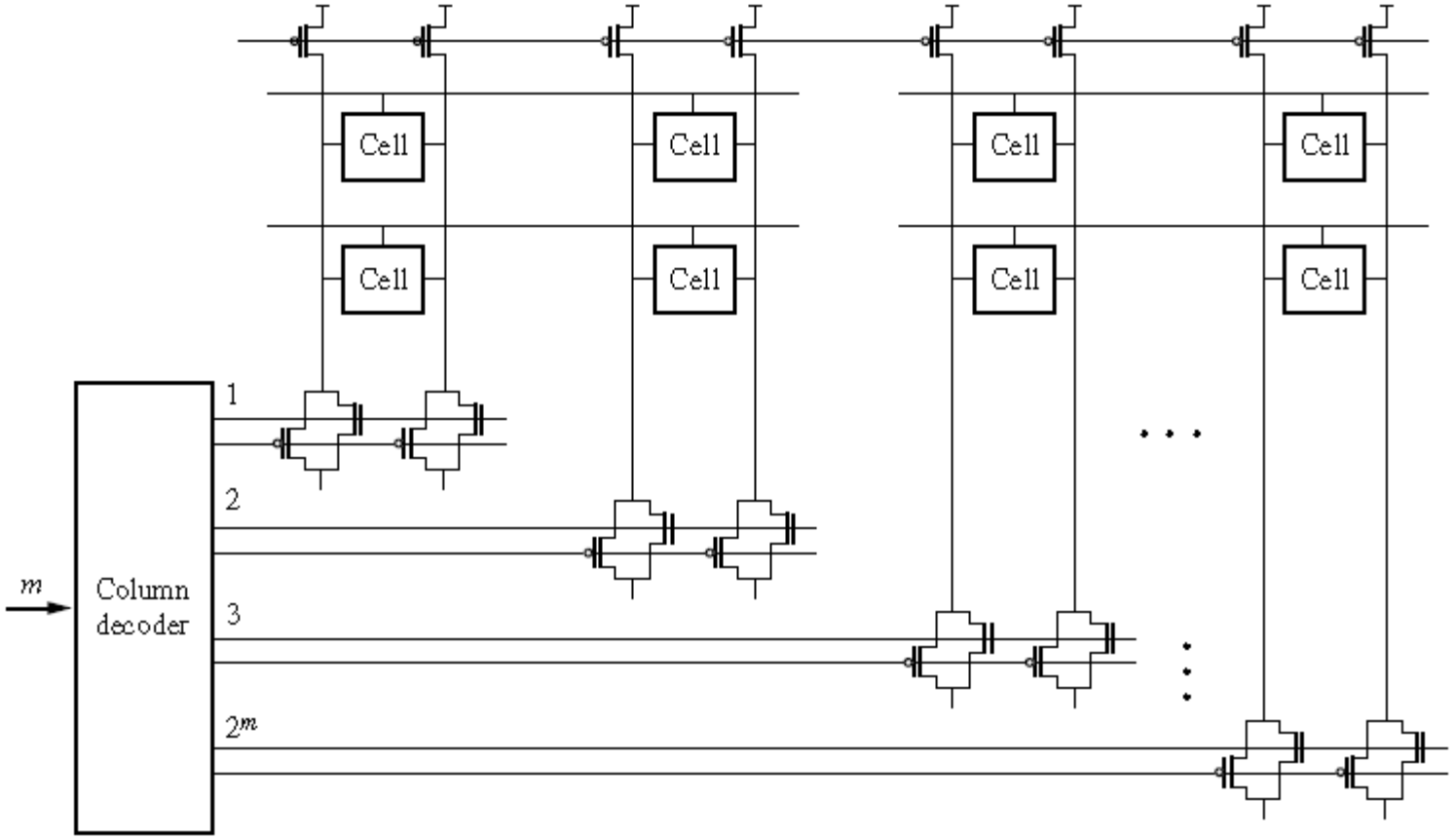
Suitable for differential-pair amplifier

PC Generation

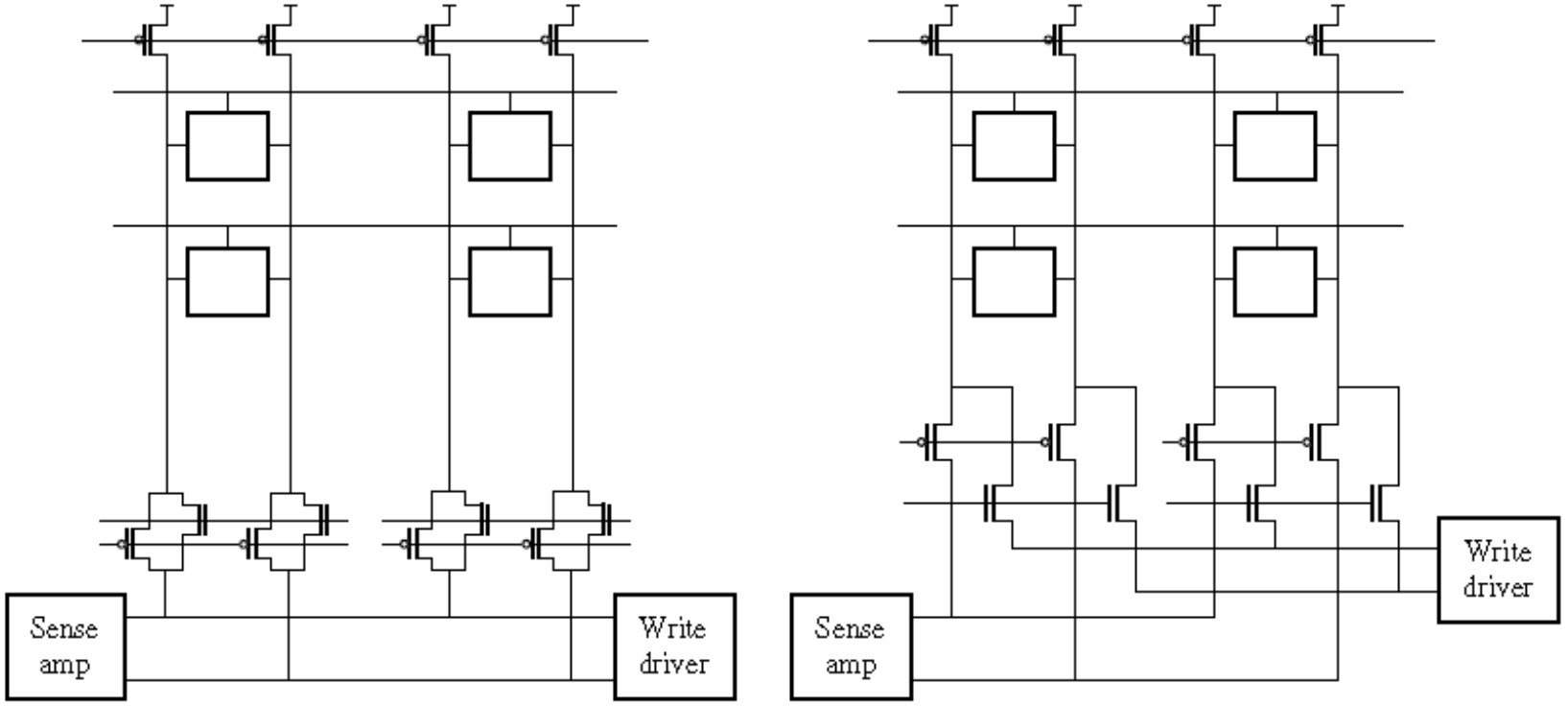


This is called monostable circuitry

Column Selection – Overall view

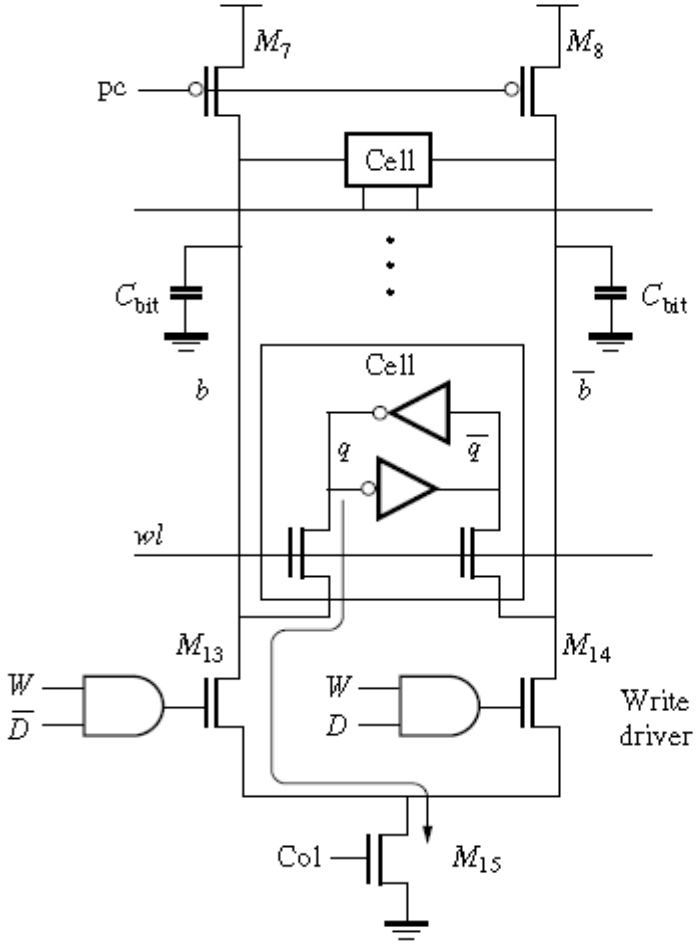


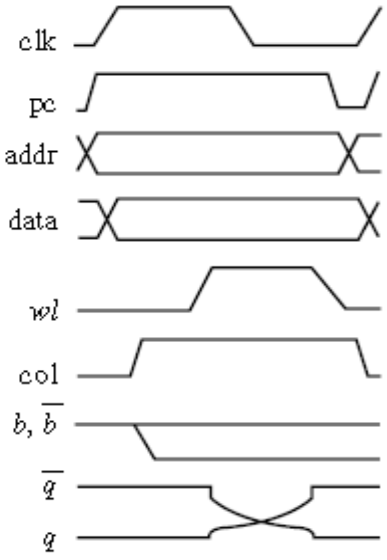
Column Selection – Optimization



If we consider that for read operation, bitlines are close to Vdd and PMOS is a better pass Transistor, connection to sense amplifier can heavily be simplified

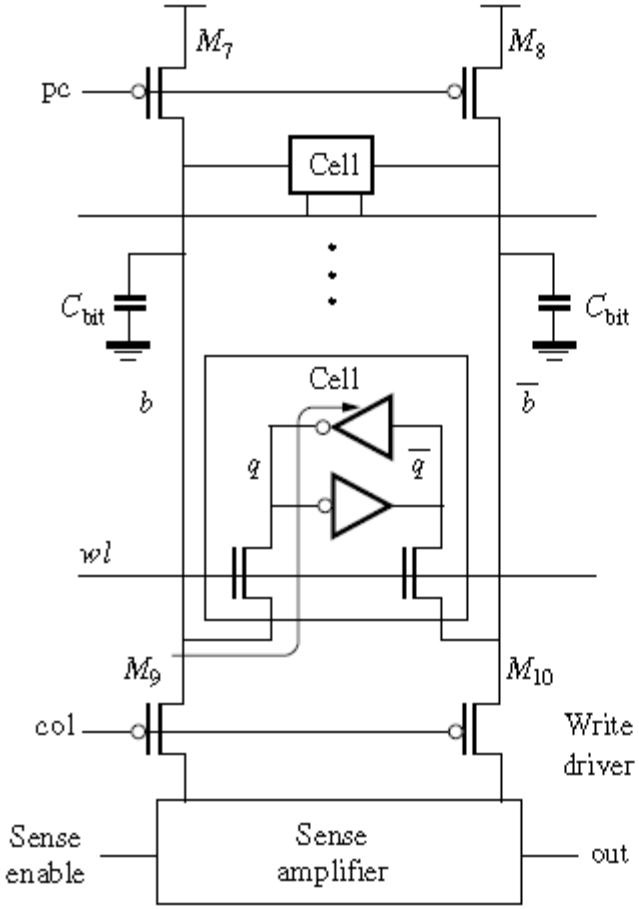
Write Complete Circuitry

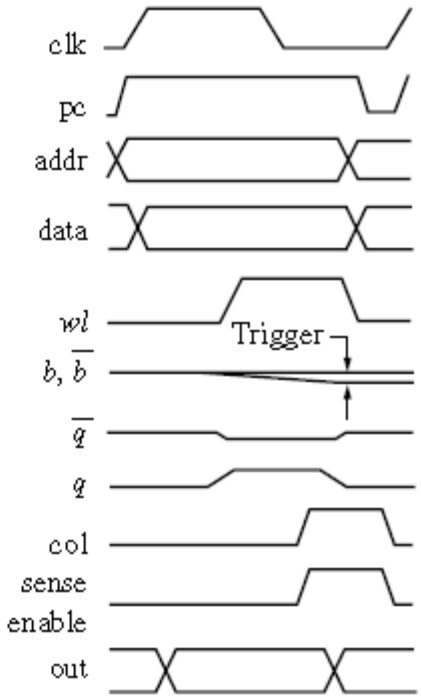




1. First column precharge Using M7 and M8
2. Address and data are stable
3. Before activation of wordline one column is discharged through write drivers (sizing matters?)
4. Wordline goes high for Write operation
5. Select bits go to standby

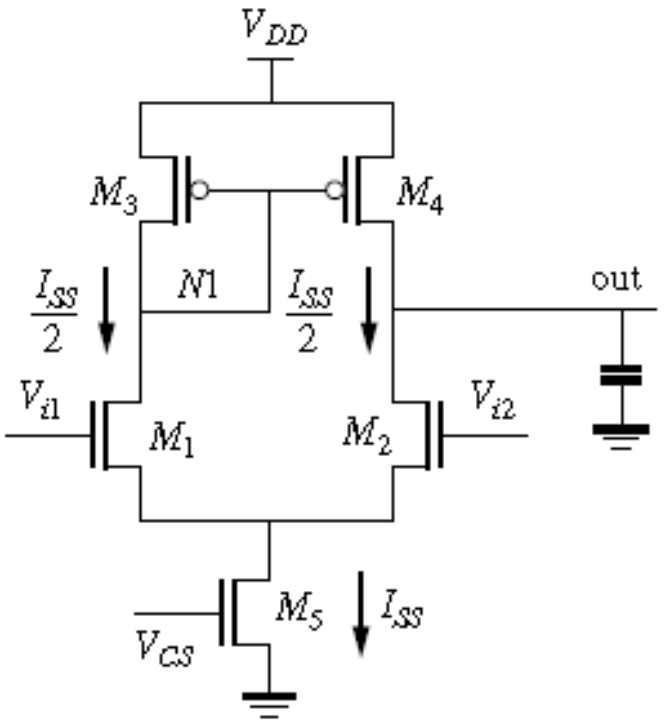
Read Complete Circuitry





1. First column precharge Using M7 and M8
2. Address and data are stable (not used for read)
3. Address is translated into
4. Column select and wordline for rows (usually same time)
4. One side starts to drop (usually 150mV – 200mV)
5. Sense amplifiers takes this difference and regenerates it to full-rail outputs

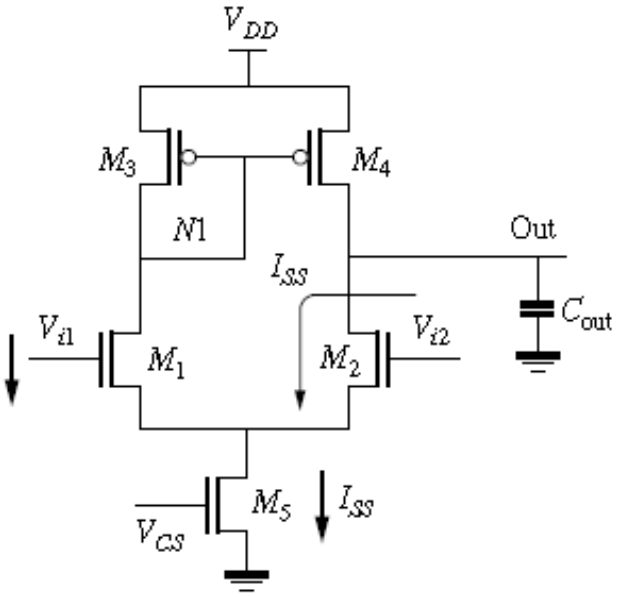
Sense Amplifier I – Differential Pair Amplifier



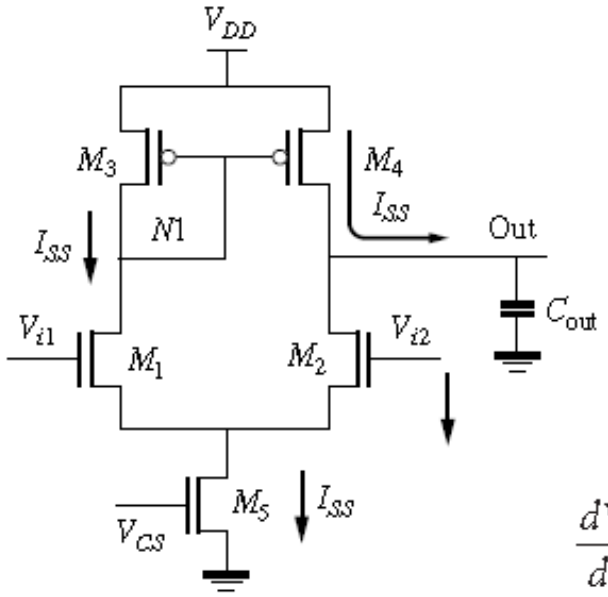
- M1 and M2 sense a small voltage difference between their gates and amplify it to a large voltag
  Vout

M3 and M4 constitute a current mirror structure

Sense Amplifier I – Differential Pair



(a) Discharging output



(b) Charging output

$$\frac{dV}{dt} = \frac{I_{SS}}{C_{out}}$$

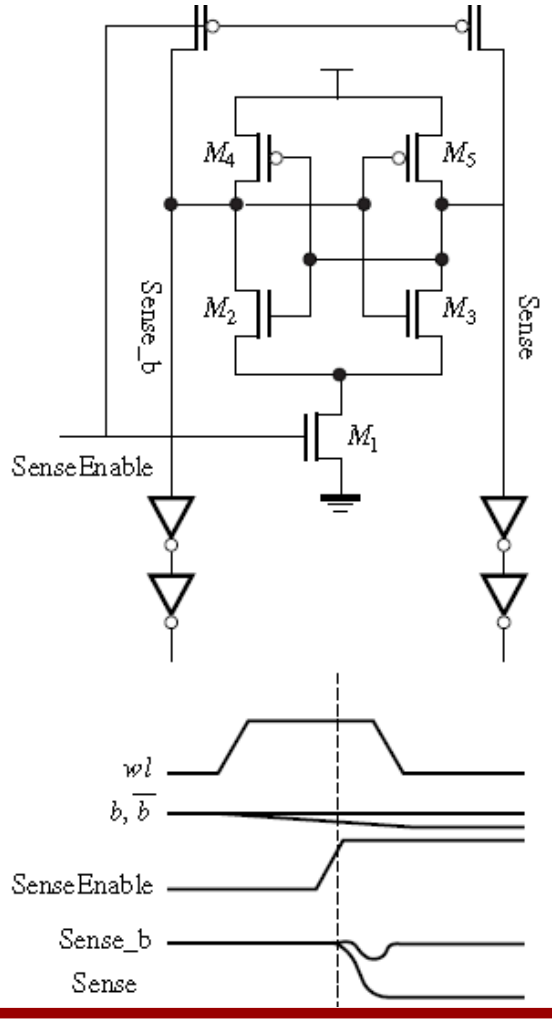
$$\Delta\tau = \frac{C_{out}\,\Delta V_{out}}{I_{SS}}$$

Not good when Vi1 and Vi2 are too close to Vdd (why?)

$$P = I_{SS}V_{DD}$$

Slew-rate at the output determines the speed, but a subtle trade-off with power

Sense Amplifier II – Latch-based Amplifier



If the precharge value of bitlines are close to vdd
This scheme proves more useful
-Inherently slower (relies on regeneration effect)
- the output and input are connected together so there needs to be a means of isolating them when the regerenation is happening (that is the job of SenseEnable signal)
- The timing of this signal is very important
what happens if activated to early (or too late)?

Also consider PVT variations! What should we do?

Dynamic-RAM

-In SRAM design we employ the stable state of flip-flops, i.e. the two inverters, to store the data and data is retained as long as power is available.

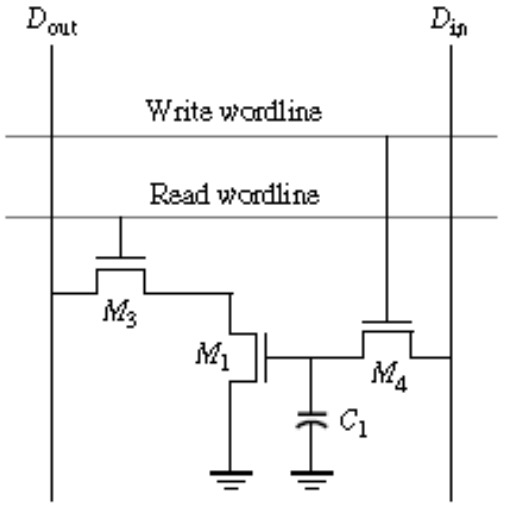
-In DRAM data is stored as charge on capacitance (therefore does not need constant power) and can potentially be a smaller cell.

However, due to leakage of capacitance charge, it requires constant *refreshing*

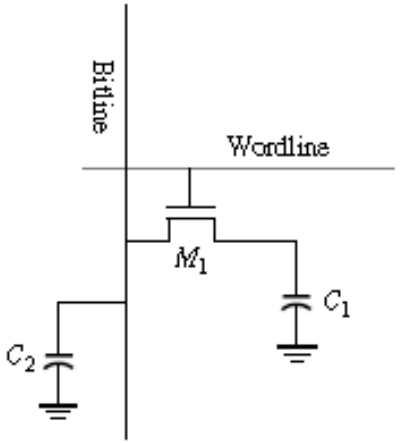
3T DRAM one of the first widely used structures

-It does not depend on the ratio of devices (unlike SRAM)

- Data is stored on parasitic C1
- Data read in D_{out} is the inverse of stored data (an inverter can be added)
- Current sensing amplifier on D_{out} also detects the value of cell

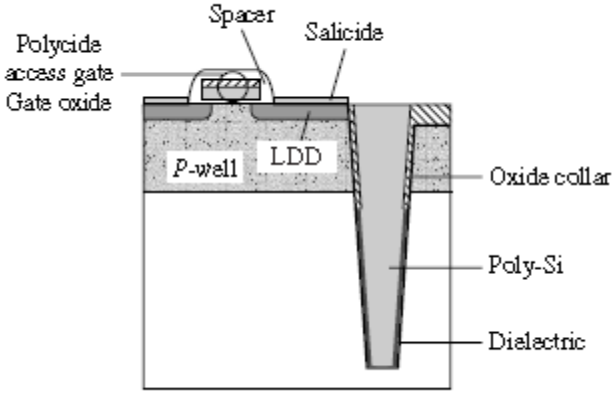


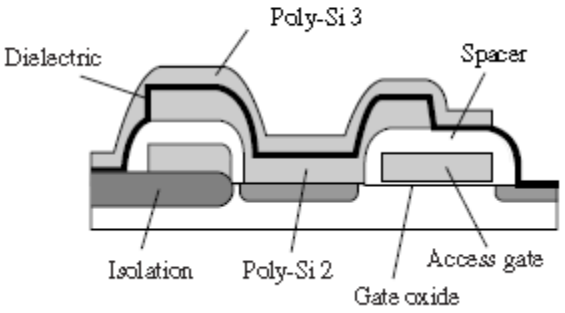
1T DRAM



An explicit cap of C1 is added (it is desired to have A large capacitance (to store more charge) in a very Small area

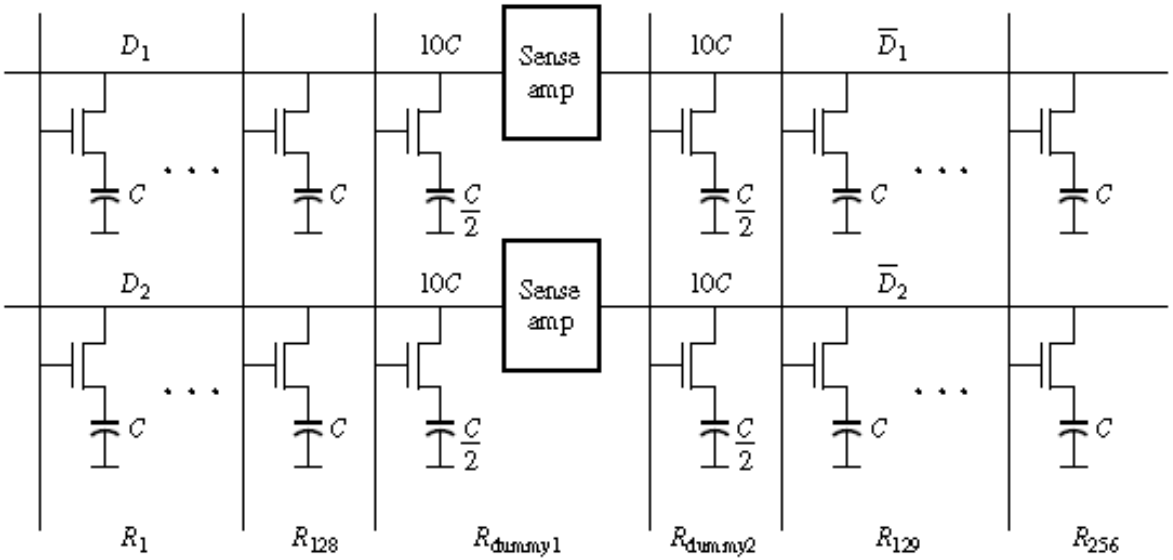
-Also a larger voltage might be used on wordline to increase the stored charge (note that Q=CV so increasing either C or V helps better storage)





During read C1 and C2 (which is ~10C1) share Their charges. If C1 too Small, the detection of voltage Change in C2 is hard

1T DRAM Refreshing



Since reading the value exposes its content, it is a destructive process

-The array is cut in half placed two Sides with an additional dummy

-We rely on regenerative switching of sense amplifier (like latch) to Detect small change and restore it.

- Columns are precharged to vdd/2

-Dummy columns (both sides) are turned on to write vdd/2 into its cap

- the desired worline (say R1) goes high, at the same time the dummy cell on opposite side is chosen

- the side connected to dummy does not see any change while the opposite side experience slight shift (voltage goes slightly above or below vdd/2 based on the stored value).

1T DRAM Example

In Figure 9.16, assume that the column lines are precharged to $V_{DD}/2$ and the internal dummy cell voltages are also set to $V_{DD}/2$. Compute the voltages on the bitlines when reading a "1" and reading a "0." Assume that boosted wordlines are used so that full V_{DD} levels can be stored in each cell when writing a "1." The column capacitance is $10C$ and the cell capacitance is C. Also compute the reference voltage level and then explain why it is required.

The output voltage is due to charge sharing. For a stored "1," we can write

$$\frac{V_{DD}}{2} C_{column} + V_{DD} C_{cell} = V^* (C_{column} + C_{cell})$$

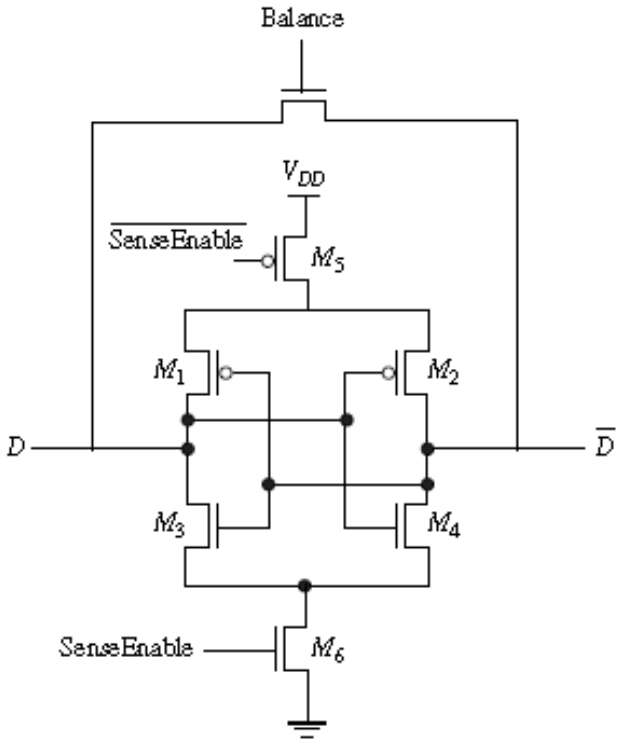
$$\therefore V^* = \frac{\frac{V_{DD}}{2} C_{column} + V_{DD} C_{cell}}{C_{column} + C_{cell}} = \frac{\frac{V_{DD}}{2} 10C + V_{DD} C}{10C + C} = 0.55\, V_{DD}$$

On the other hand, if a "0" is stored, the output on D is

$$\frac{V_{DD}}{2} C_{column} + 0 \times C_{cell} = V^* (C_{column} + C_{cell})$$

$$\therefore V^* = \frac{\frac{V_{DD}}{2} C_{column}}{C_{column} + C_{cell}} = \frac{\frac{V_{DD}}{2} 10C}{10C + C} = 0.45\, V_{DD}$$

Sense Refresh Circuitry



Operates similar to latched based sense amplifier in SRAM (take small Difference between D and Dbar and amplifies it to full rail values)