ESE370: Circuit-Level Modeling, Design, and Optimization for Digital Systems

Lec 28: November 15, 2021

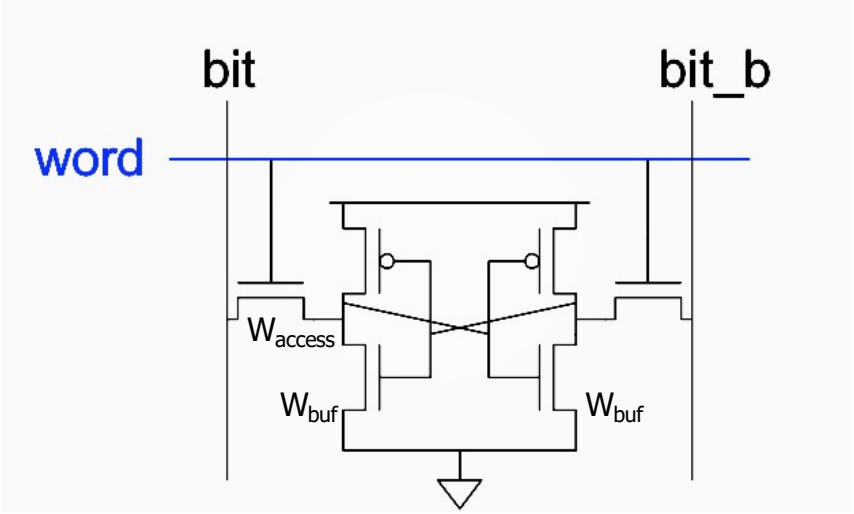
Memory Core Part 1

Today

- Memory Array examples
- Serial Access Memories
- Memory
  - Memory core
    - 6T SRAM
- Project 2 is on this

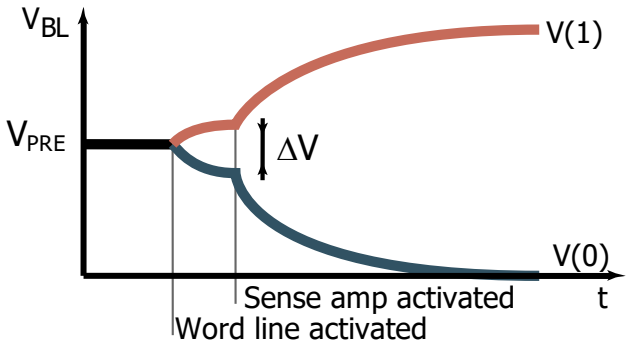
Column Capacitance Consequence

- Preclass1: What is capacitance of a bitline?
  - W_{access} (pass transistor size), d rows, $\gamma = C_{diff0}/C_0$
- Preclass2: What is the delay for the cell to drive the bitline during a read?
  - W_{buf} (inverter size in cell), R_0
- **Conclude:** Can't size up cell→driving bitline will be slow



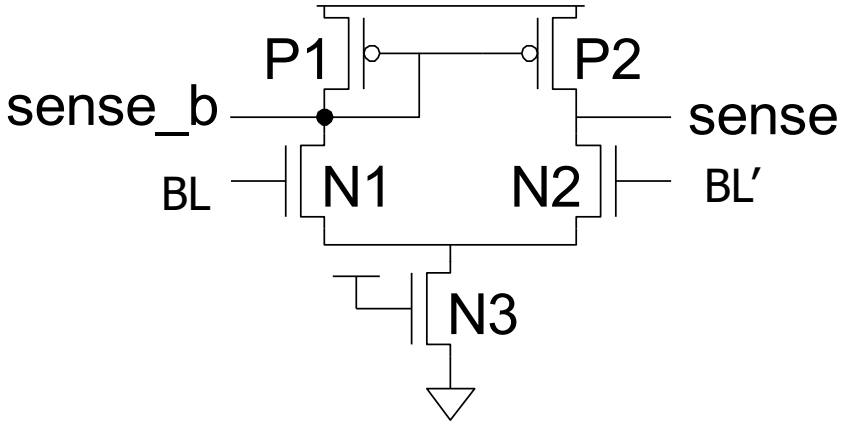
Sense Amplifiers

- Bitlines have many cells attached
  - Ex: 32-kbit SRAM has 128 rows x 256 cols
  - 128 cells on each bitline
- $t_{pd} \propto (C/I)\ \Delta V$
  - Even with shared diffusion contacts, 64C of diffusion capacitance (big C)
  - Discharged slowly through small transistors in each memory cell (small I)
- *Sense amplifiers* are triggered on small voltage swing (ΔV)



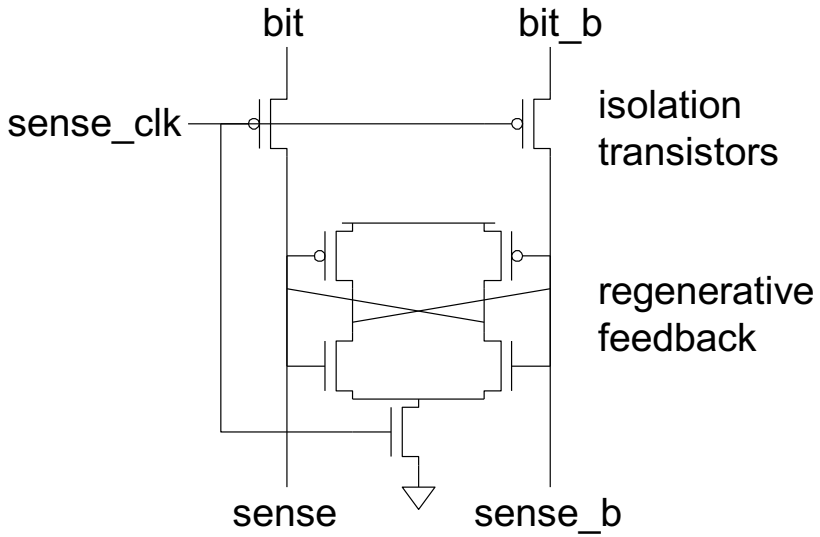
Differential Pair Amp

- Differential pair requires no clock
- But always dissipates static power



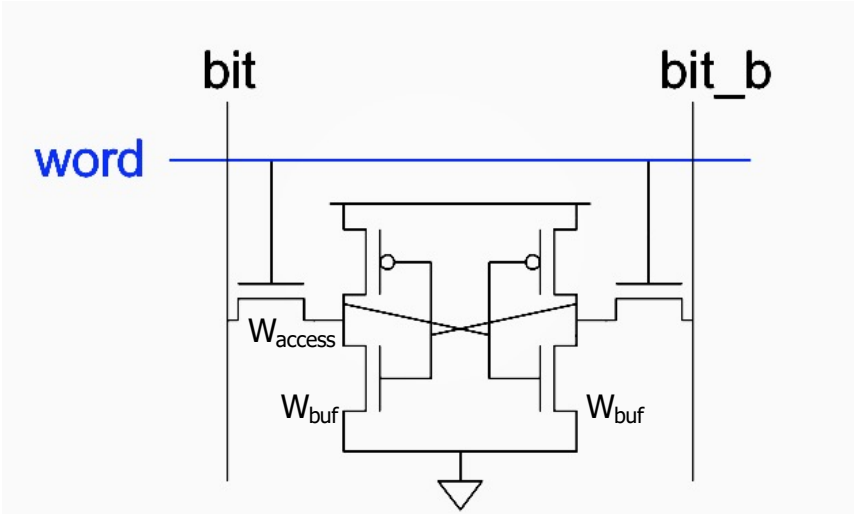
Clocked Sense Amp

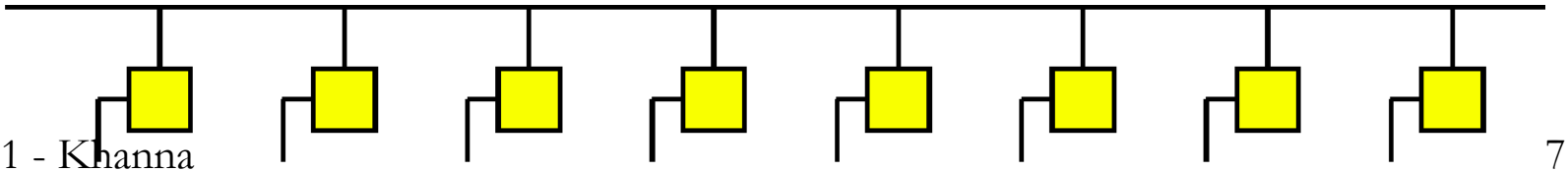
- Clocked sense amp saves power
- Requires sense_clk after enough bitline swing
- Isolation transistors cut off large bitline capacitance



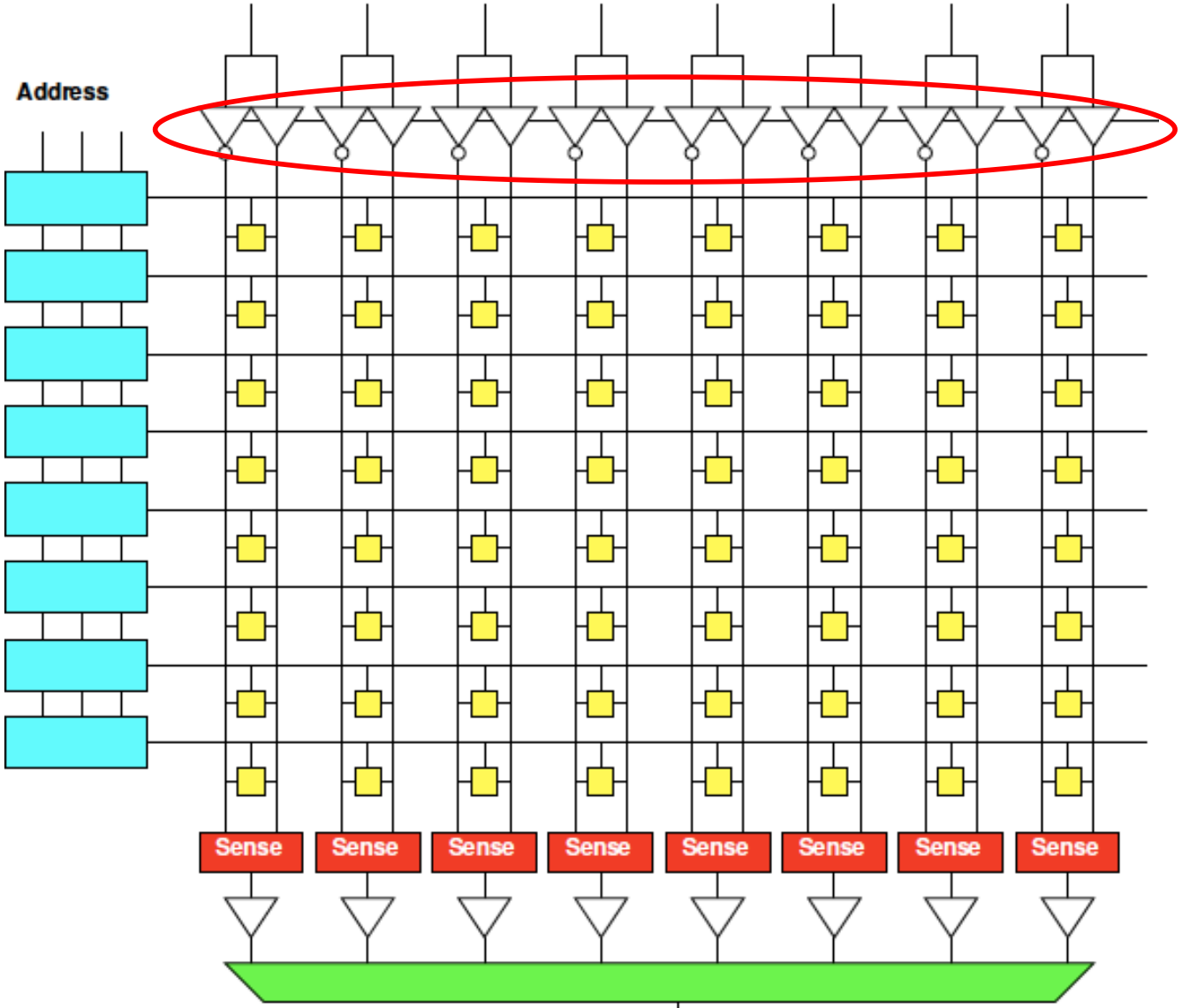
Word Line Capacitance

- Preclass4: What is capacitance of word line (row)?
  - W_{access}– transistor width of column device
  - w columns
  - $\gamma = C_{diff0}/C_0$
- Preclass5: Delay driving word line?
  - $W_{wldrive}$ Drive inverter



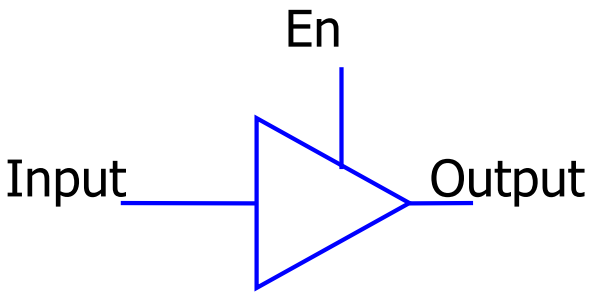


Column Drivers: Memory Bank



Tristate Buffer

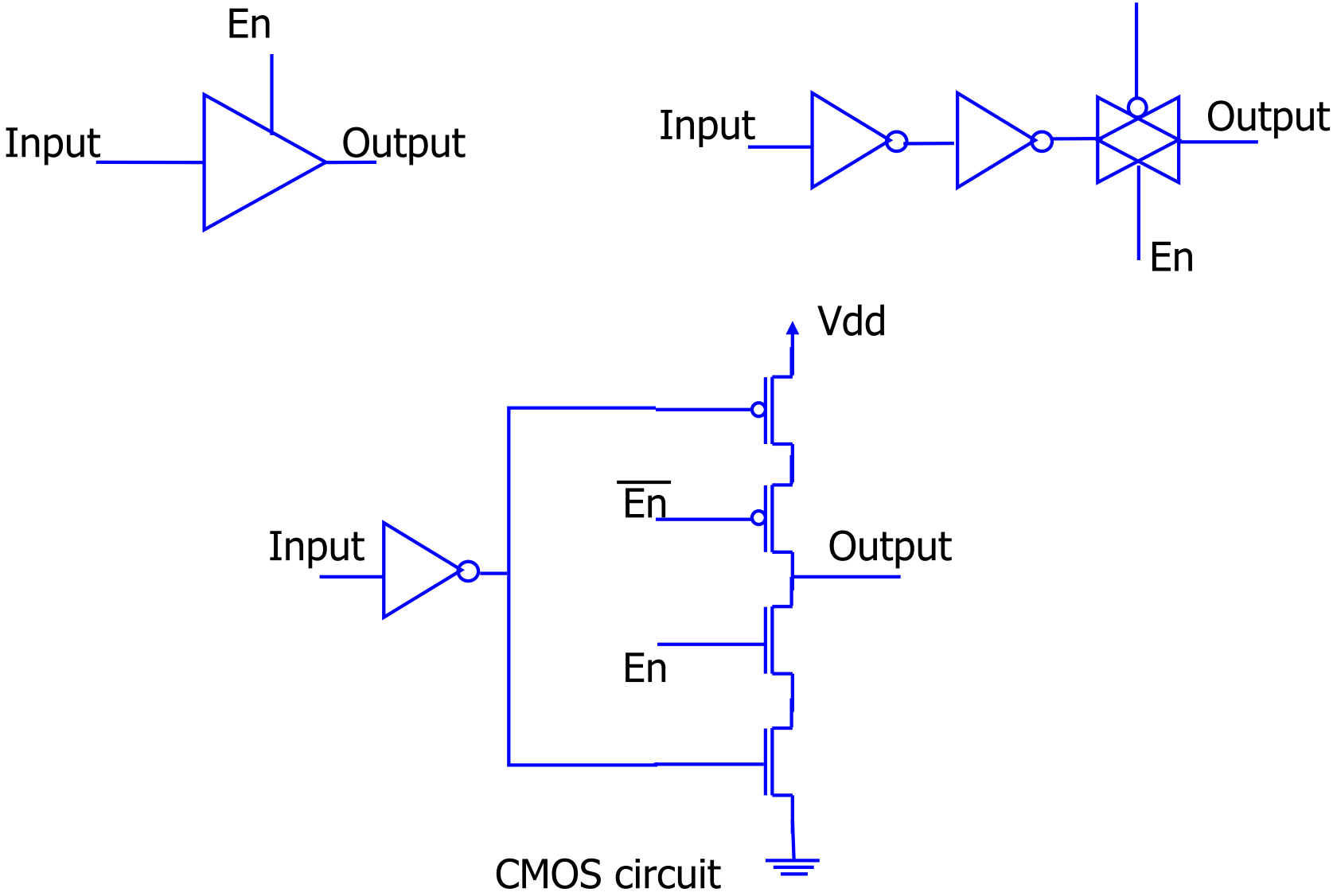
- Typically used for signal traveling, e.g. bus
- Ideally all devices connected to a bus should be disconnected except for active device reading or writing to bus
- Use high-impedance state to simulate disconnecting



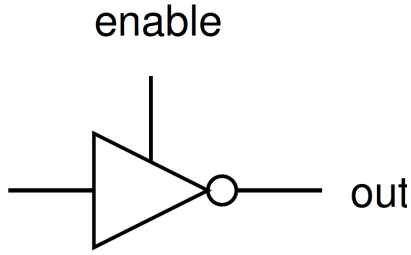
Active-high buffer

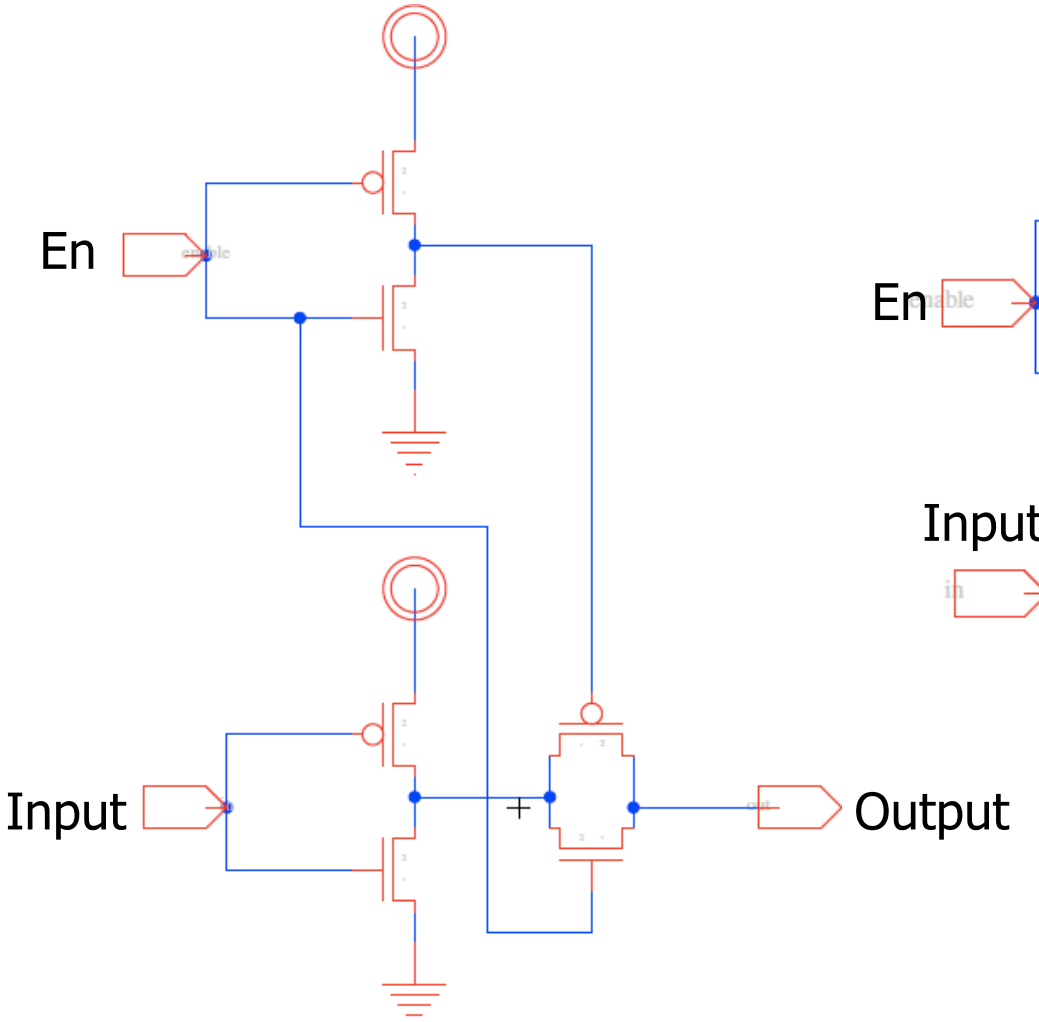
Input	En	Ouptut
0	0	Z
1	0	Z
0	1	0
1	1	1

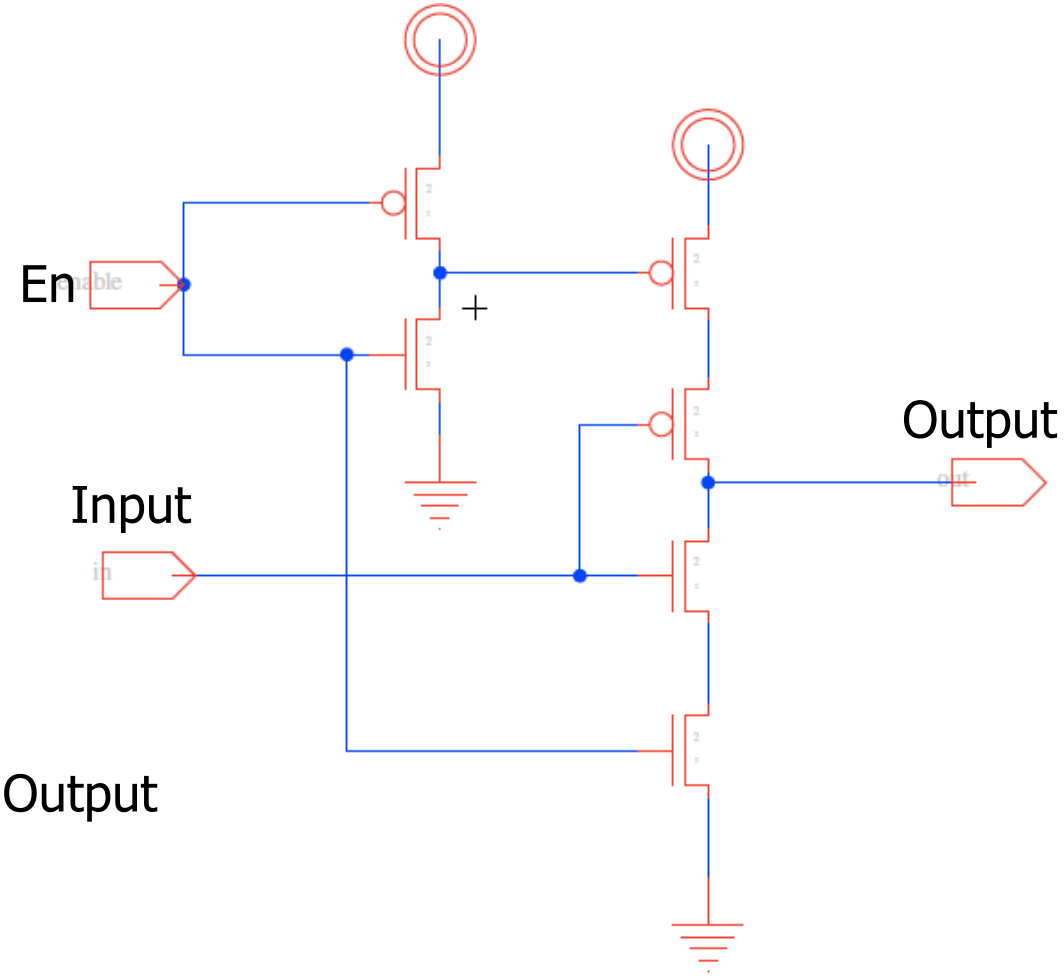
Tristate Buffer



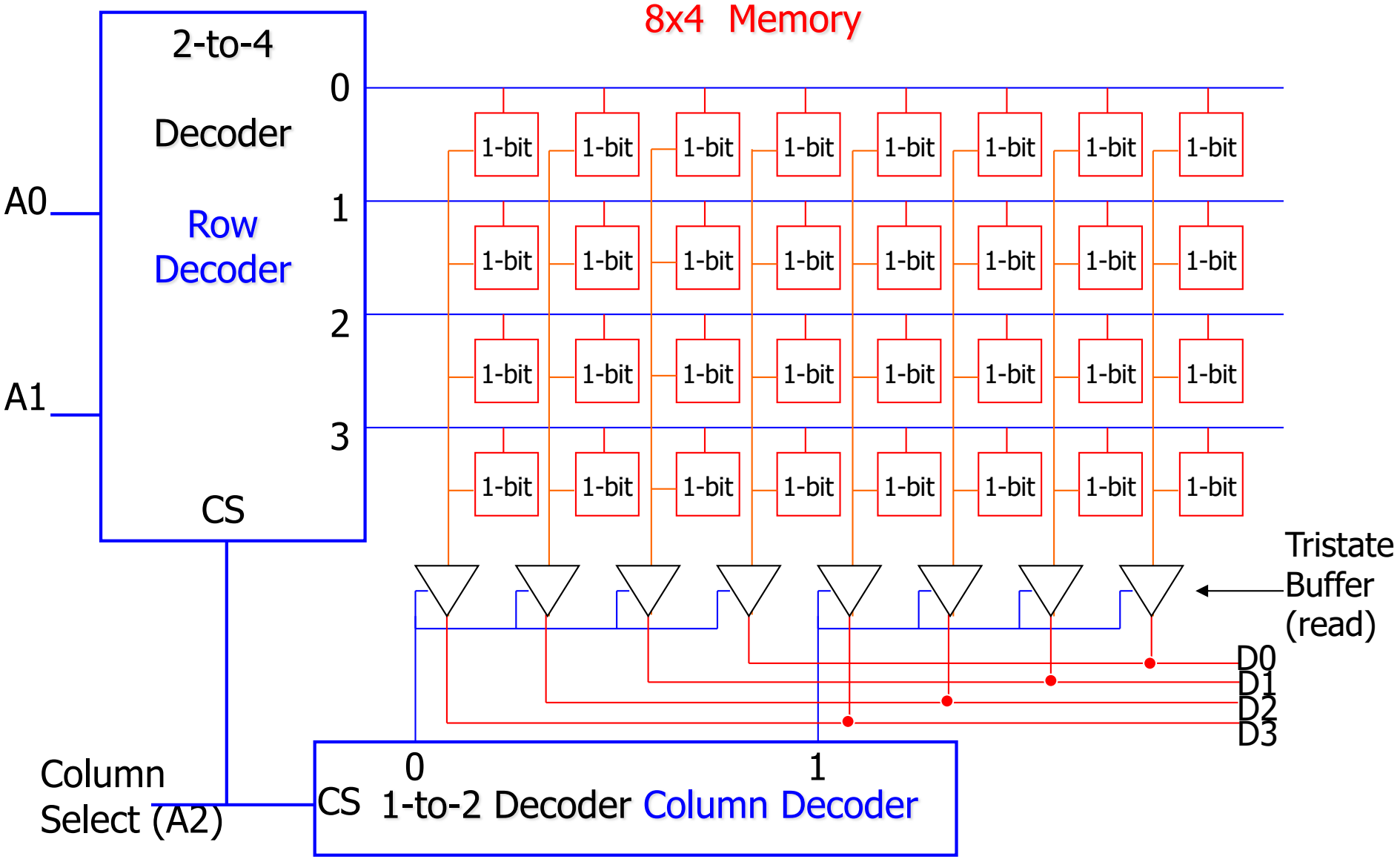
Tristate Inverters



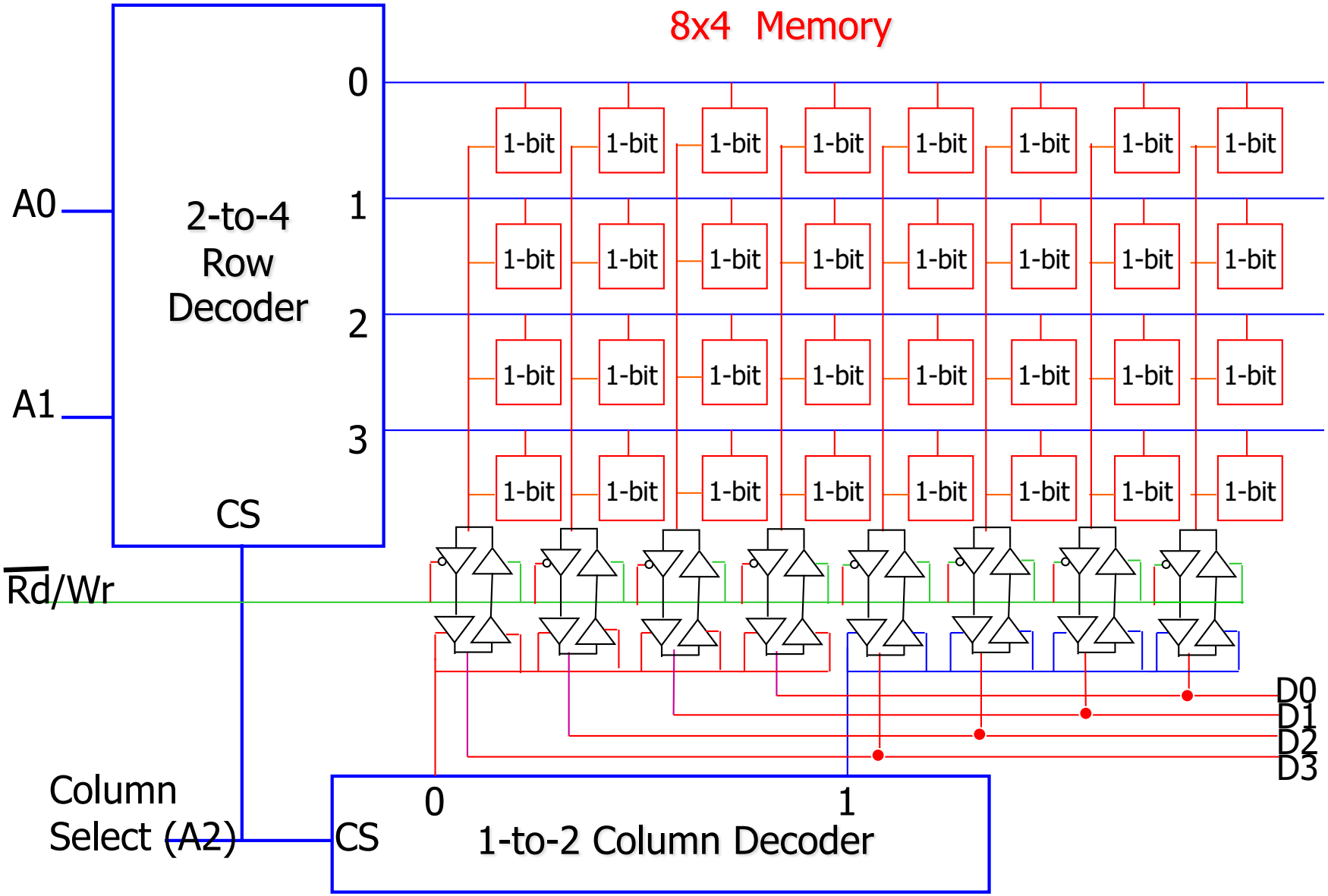




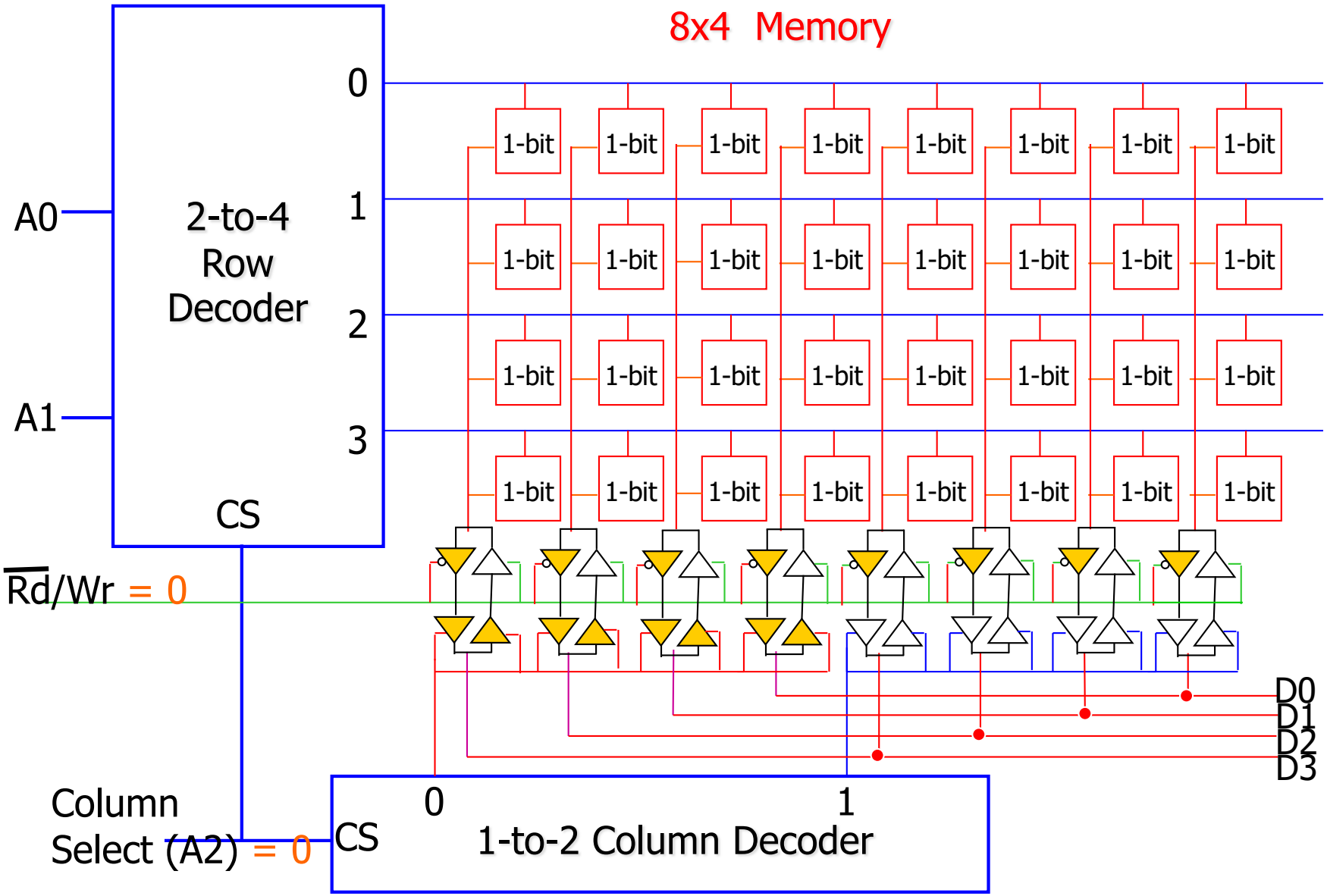
8x4 Memory with column decoder



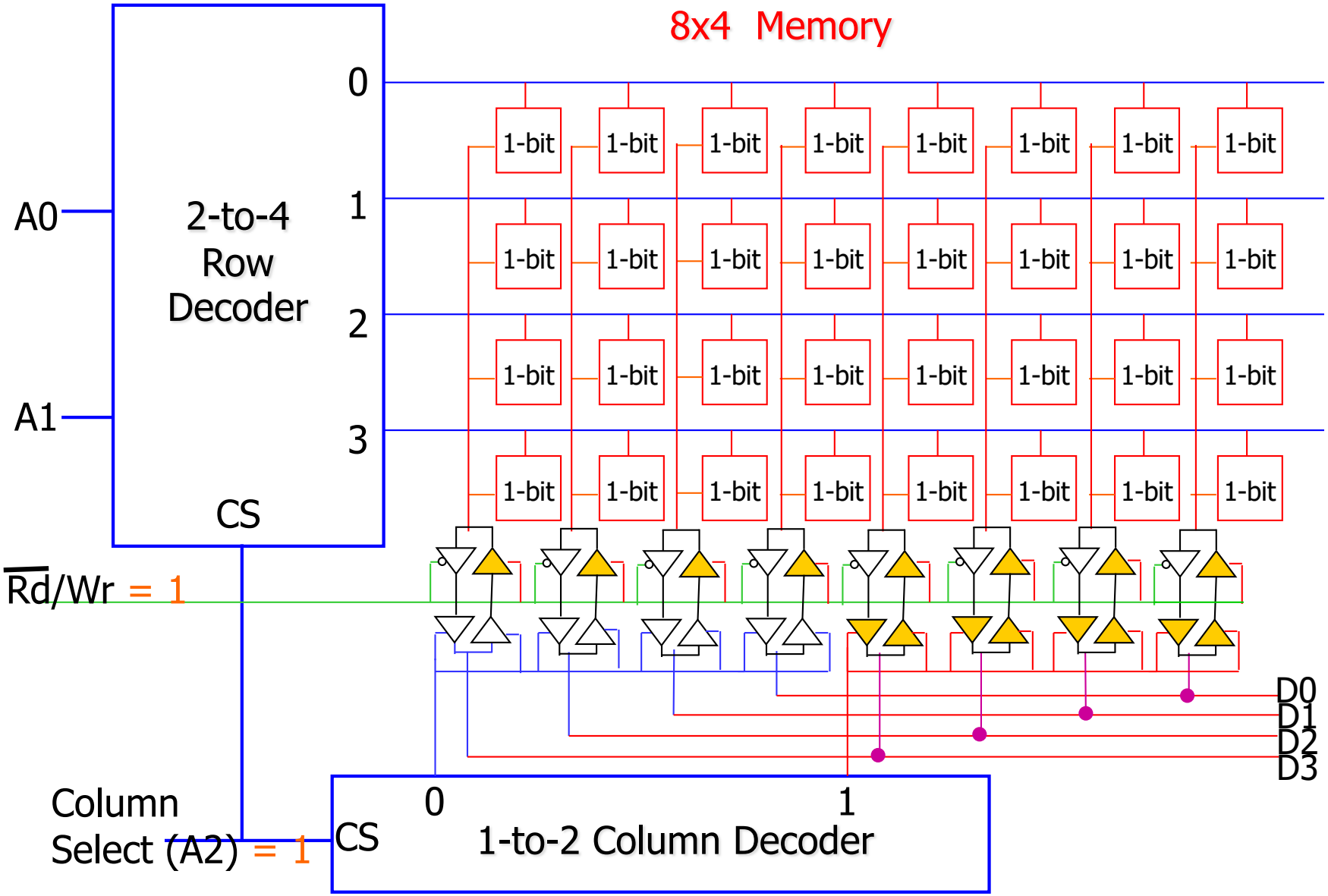
Read/Write Memory



Read/Write Memory



Read/Write Memory

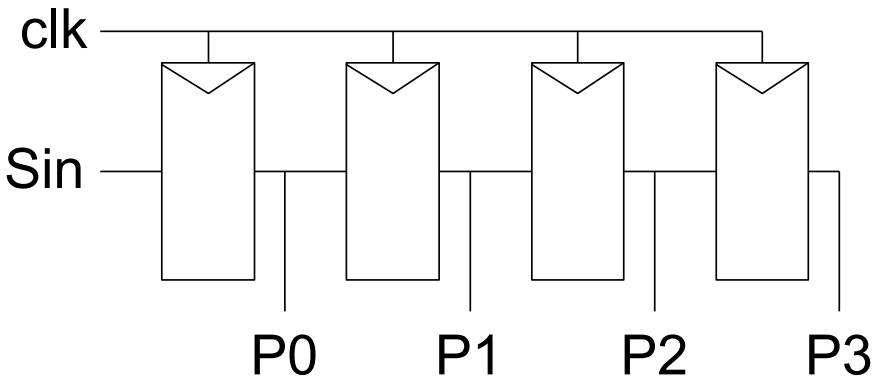


Serial Access Memories

- Serial access memories do not use an address
  - Serial In Parallel Out (SIPO)
  - Parallel In Serial Out (PISO)
  - Shift Registers
  - Queues (FIFO, LIFO)

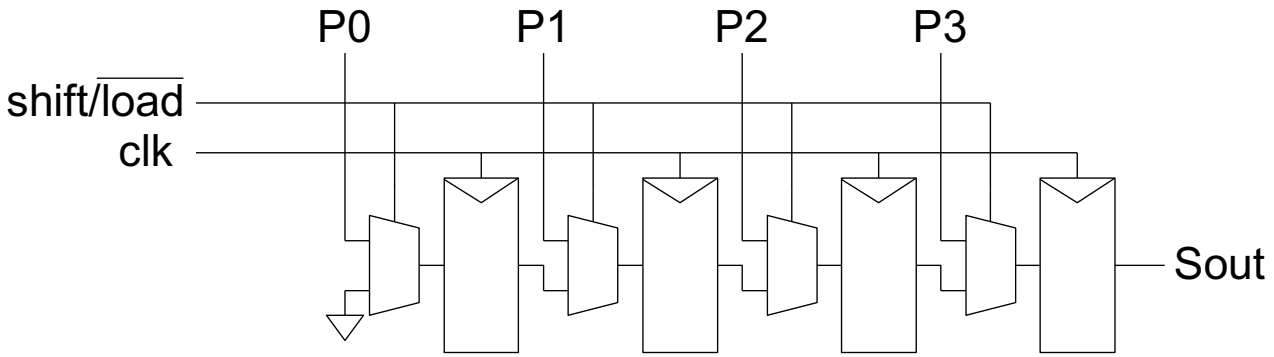
Serial In Parallel Out

- 1-bit shift register reads in serial data
  - After N steps, presents N-bit parallel output



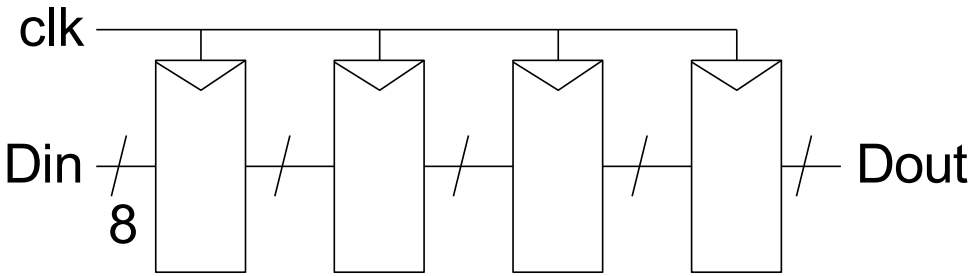
Parallel In Serial Out

- Load all N bits in parallel when shift = 0
  - Then shift one bit out per cycle



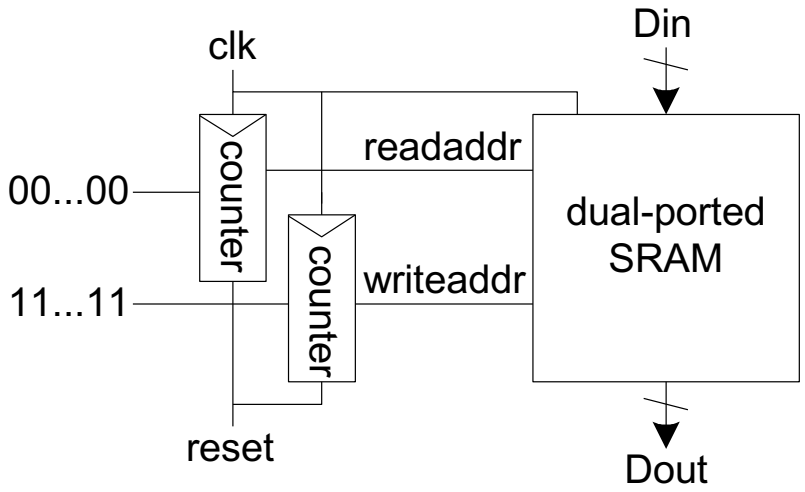
Shift Register

- *Shift registers* store and delay data
- Simple design: cascade of registers



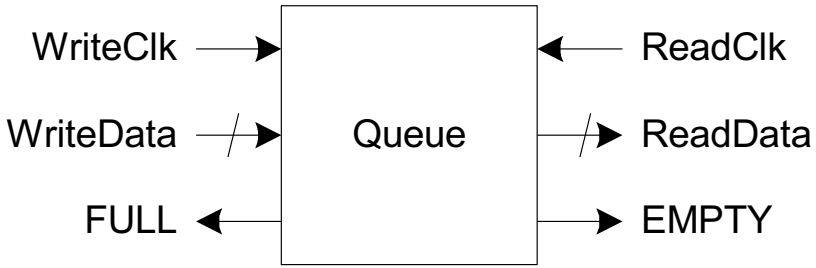
Denser Shift Registers

- Flip-flops aren't very area-efficient
- For large shift registers, keep data in SRAM instead
- Move read/write pointers to RAM rather than move data
  - Initialize read address to first entry, write to last
  - Increment address on each cycle



Queues

- *Queues* allow data to be read and written at different rates.
- Read and write each use their own clock, data
- Queue indicates whether it is full or empty
- Build with SRAM and read/write counters (pointers) storing read/write address

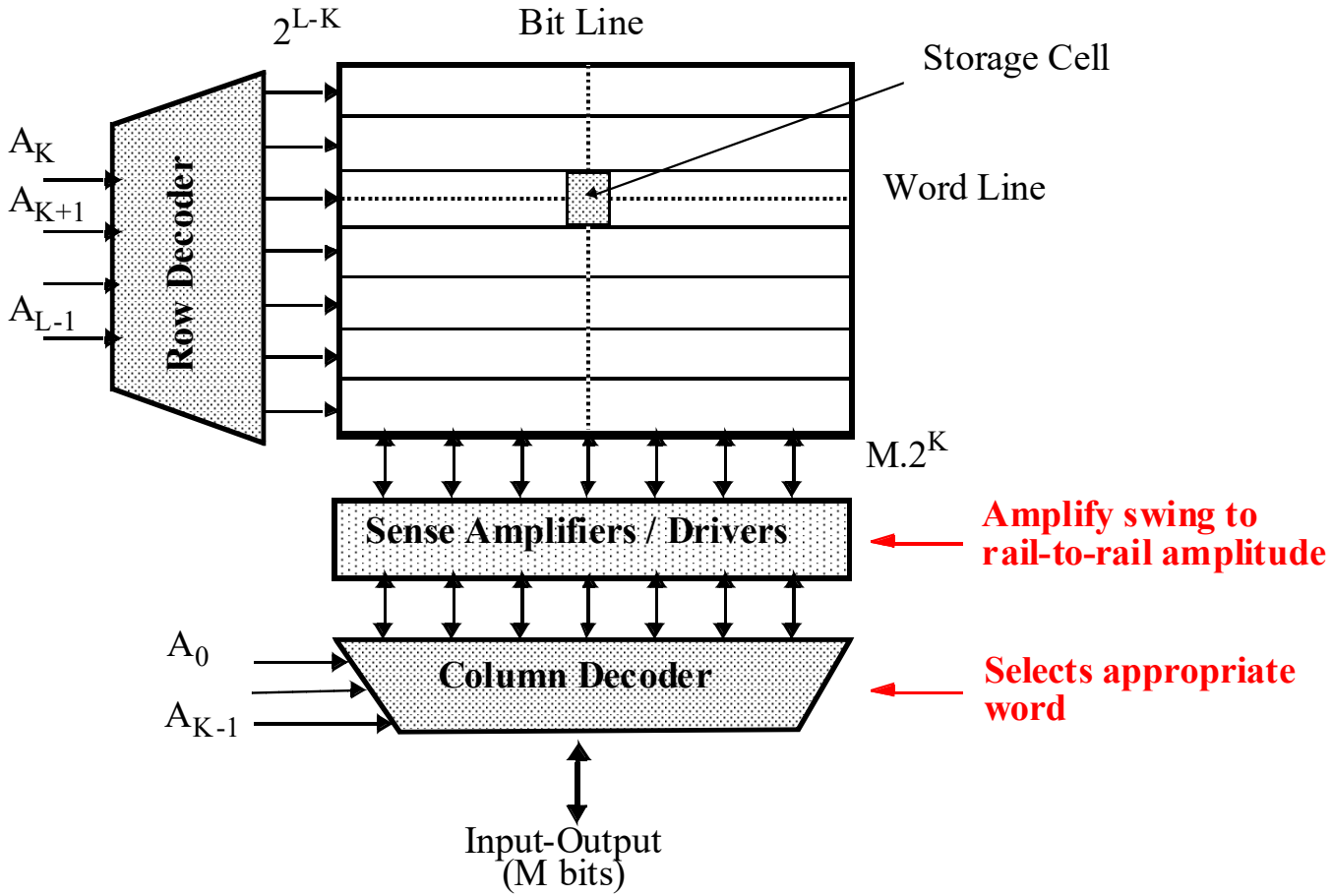


FIFO, LIFO Queues

- *First In First Out* (FIFO)
  - Initialize read and write pointers to first element
  - Queue is EMPTY
  - On write, increment write pointer
  - If write catches read, Queue is FULL
  - On read, increment read pointer
  - If read catches write, Queue is EMPTY
- *Last In First Out* (LIFO)
  - Also called a *stack*
  - Use a single *stack pointer* for read and write

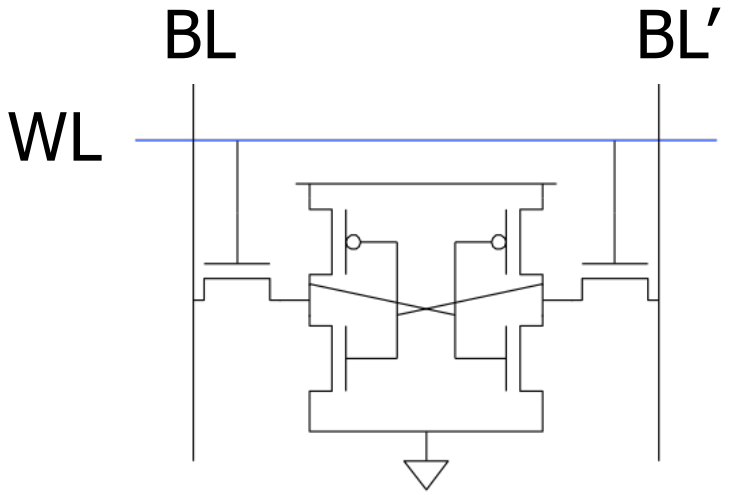
Array-Structured Memory Architecture

Problem: ASPECT RATIO or HEIGHT >> WIDTH

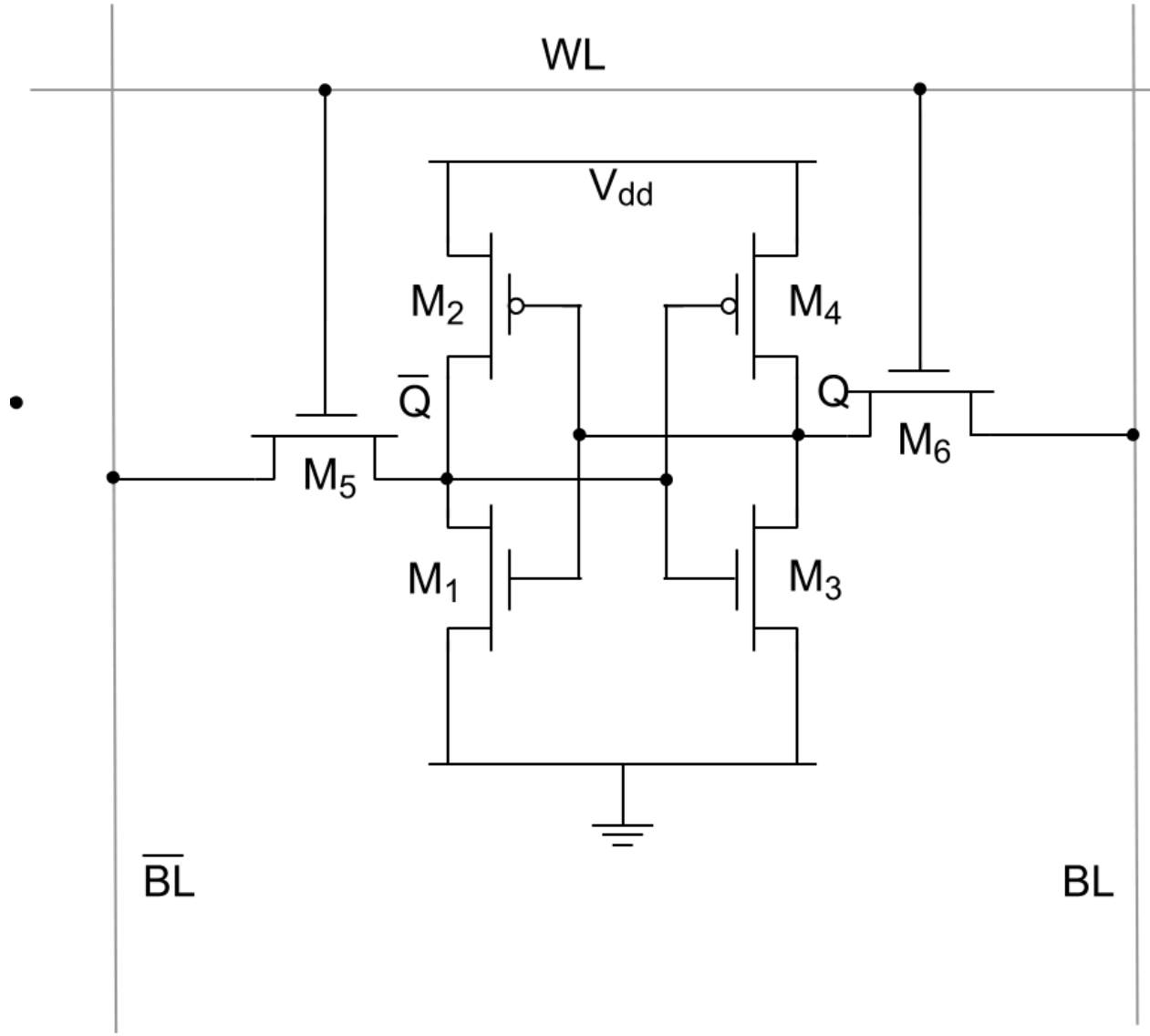


6T SRAM Cell

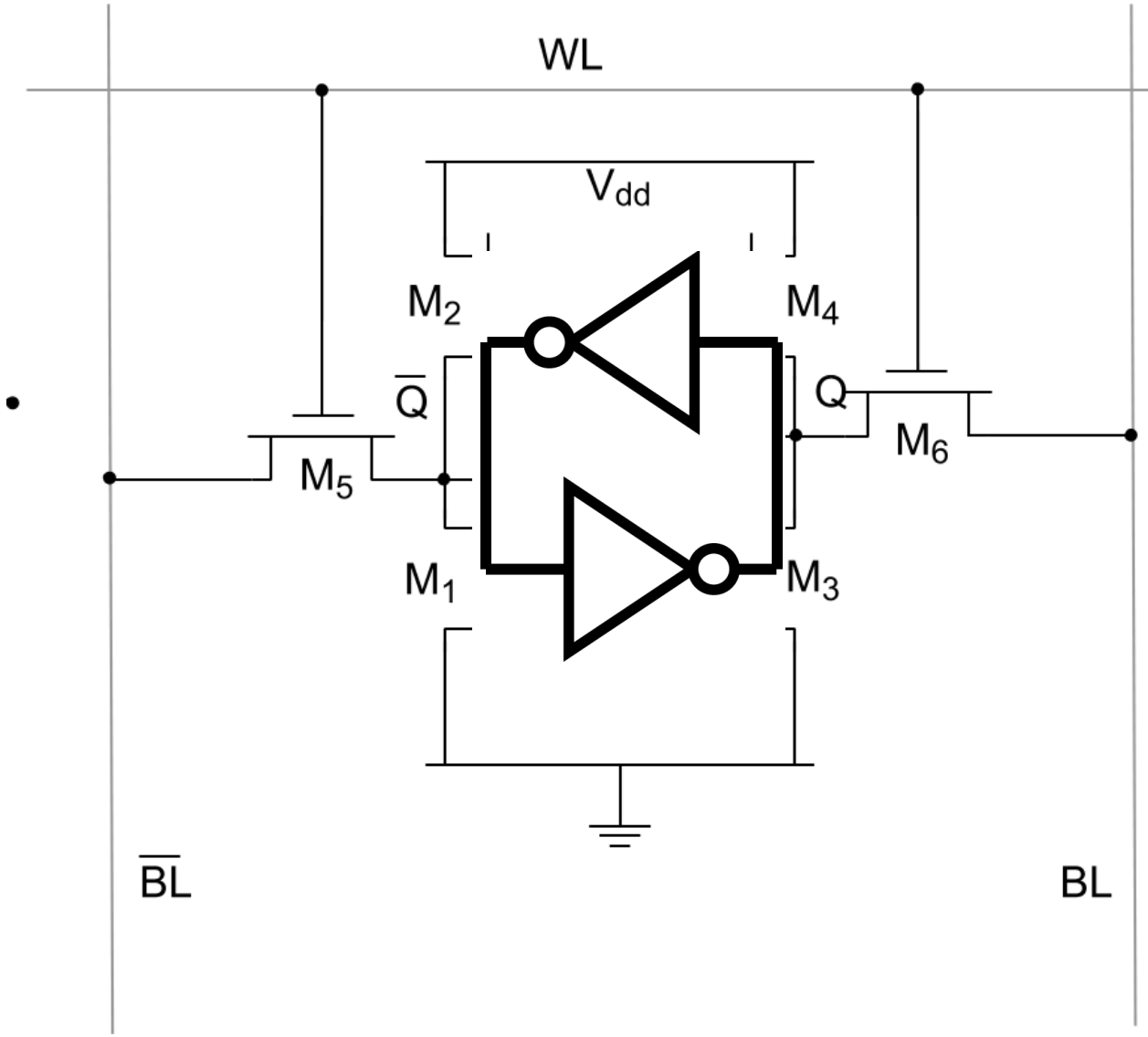
- Cell size accounts for most of array size
  - Reduce cell size at expense of complexity
- 6T SRAM Cell
  - Used in most commercial chips
  - Data stored in cross-coupled inverters
- Read:
  - Precharge BL, BL'
  - Raise WL
- Write:
  - Drive data onto BL, BL'
  - Raise WL



6-transistor CMOS SRAM Cell



6-transistor CMOS SRAM Cell

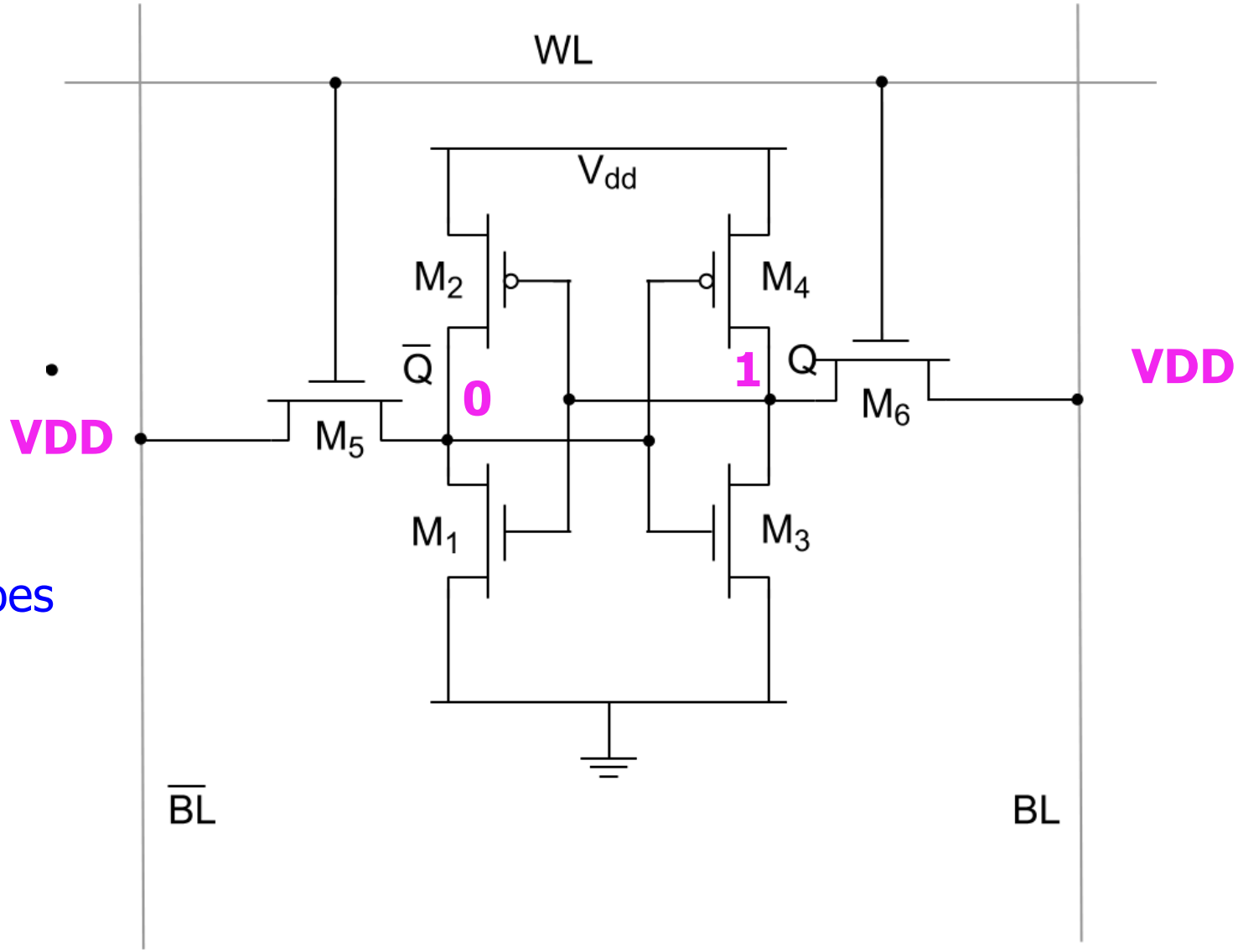


6-transistor CMOS SRAM Cell (preclass 1)

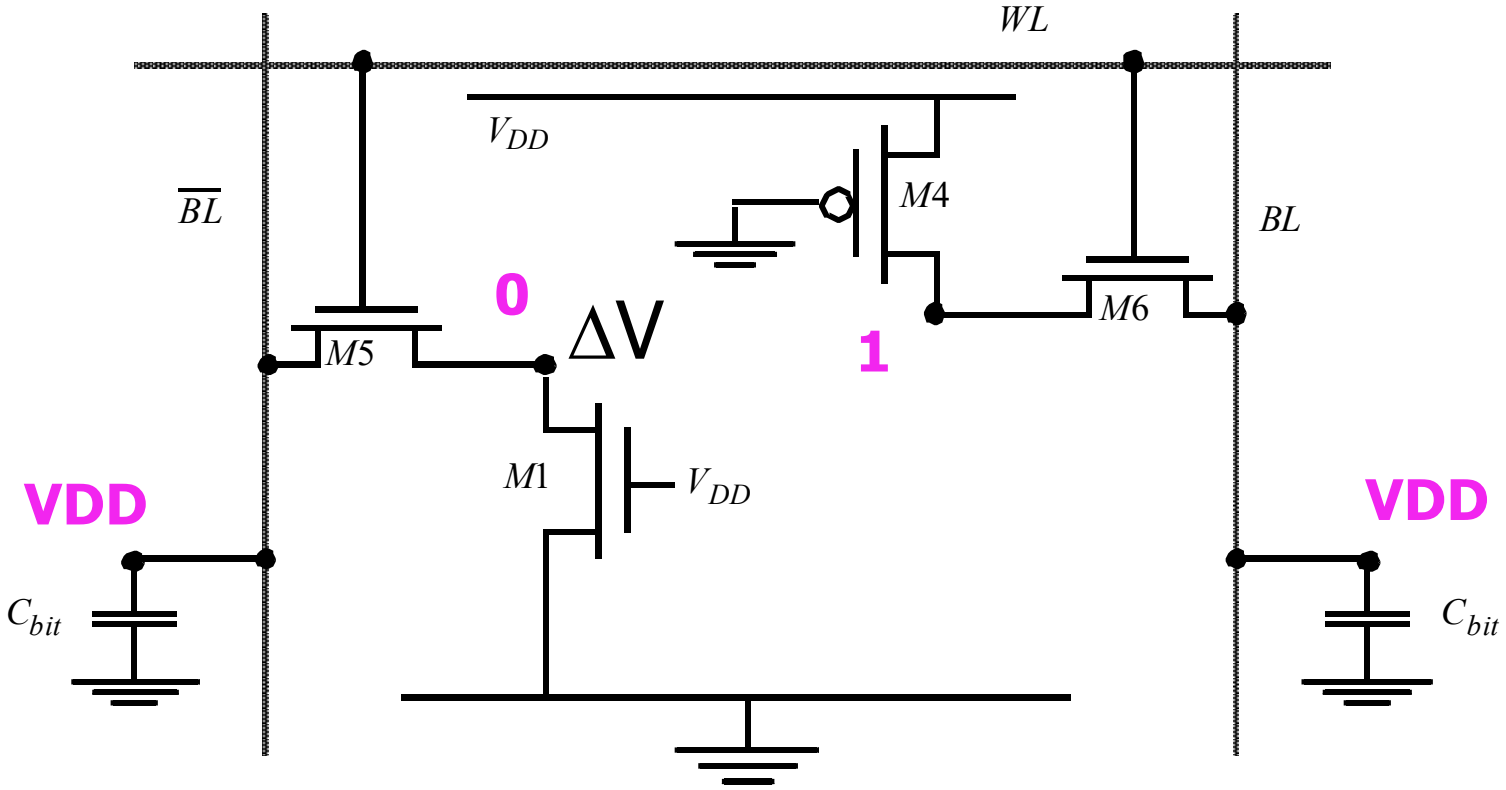
Assume 1 is stored (Q=1)

Read Operation:

- First bitlines get precharged high (Vdd)
- Then wordline goes high (Vdd)
  - Precharge disconnected

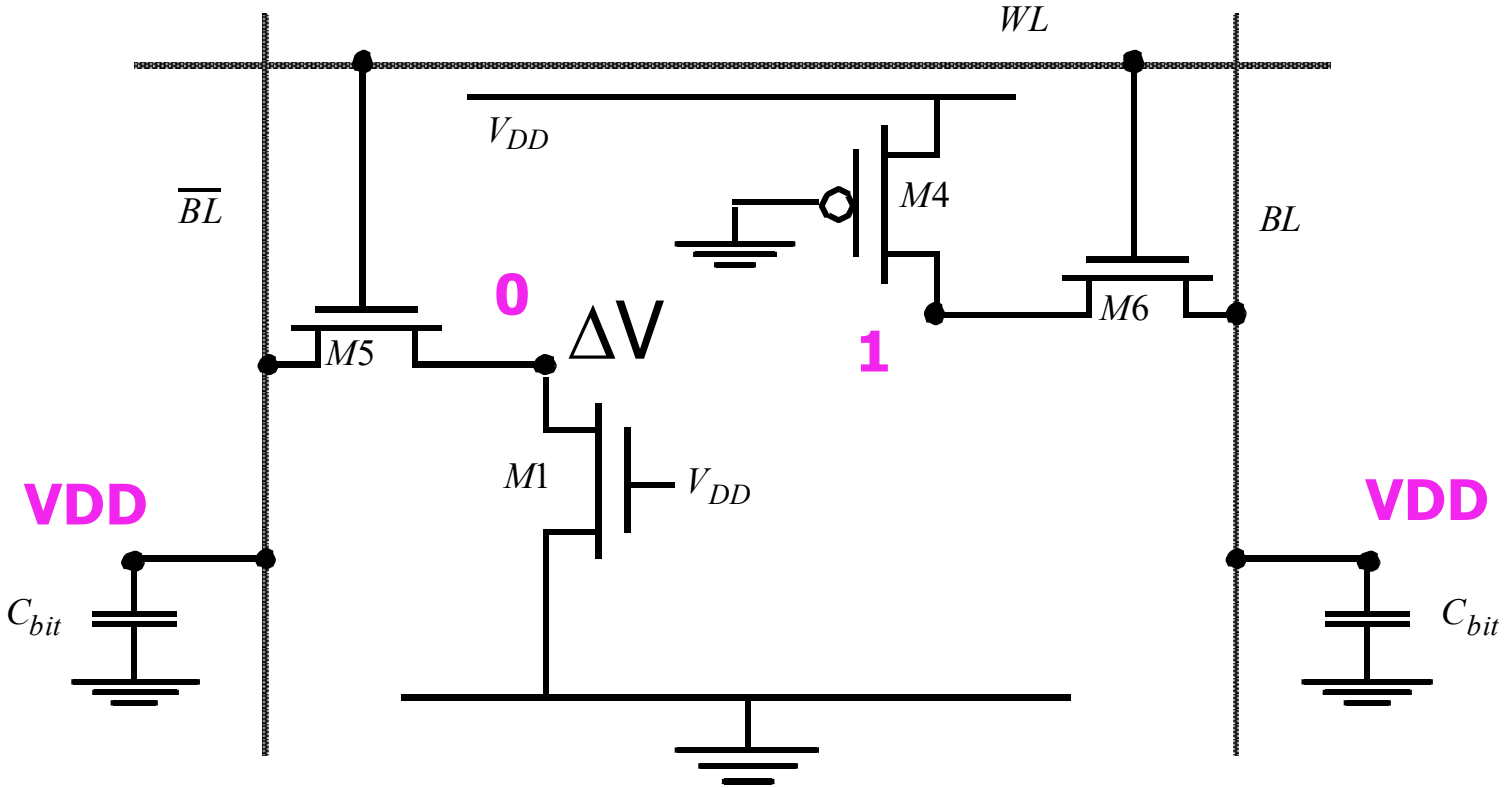


CMOS SRAM Analysis (Read) (preclass 1)



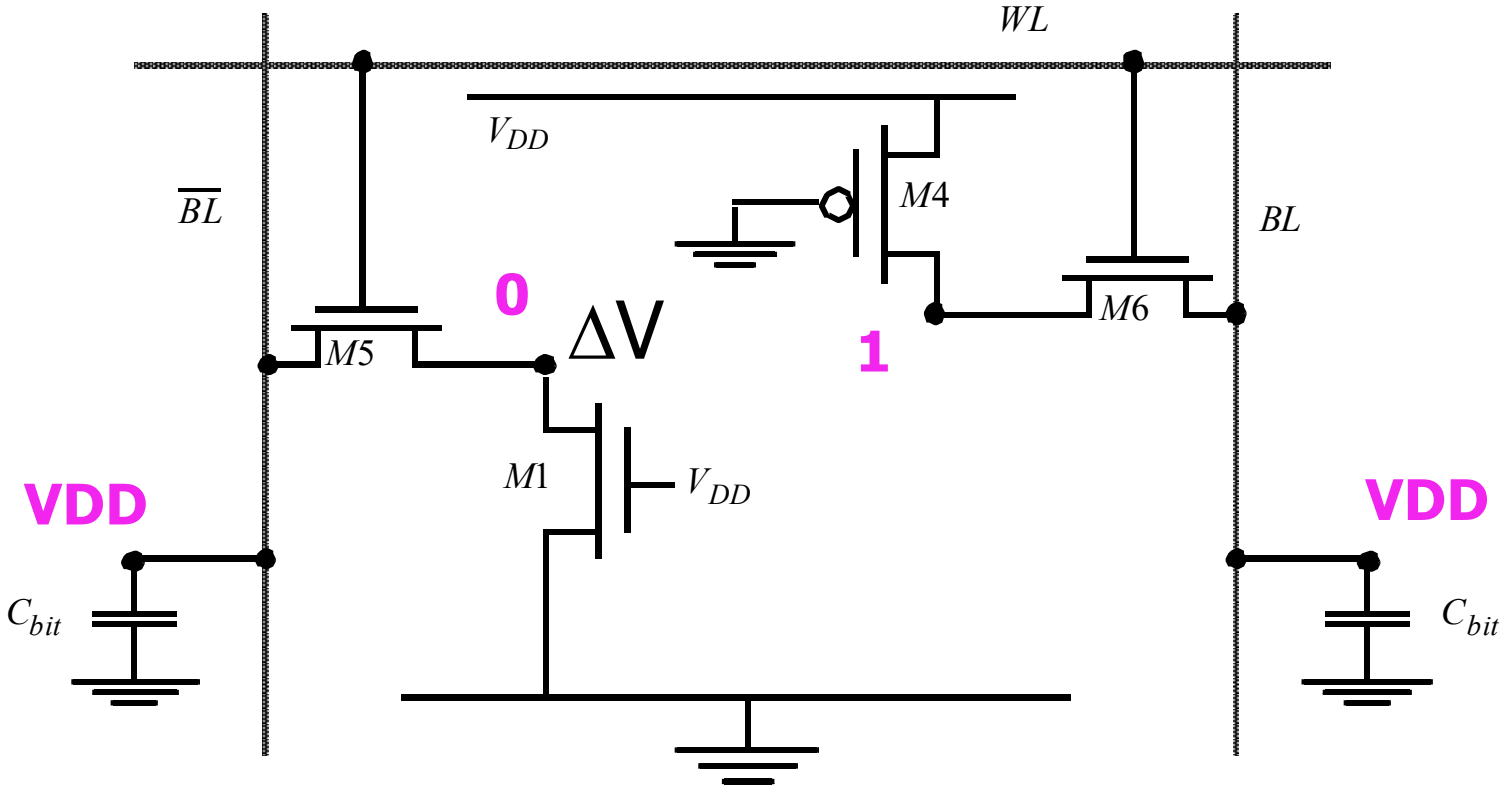
Which two transistors are discharging \BL to Gnd?
What regions of operation are the transistors in?
Write the KCL equation for the two transistors

CMOS SRAM Analysis (Read)



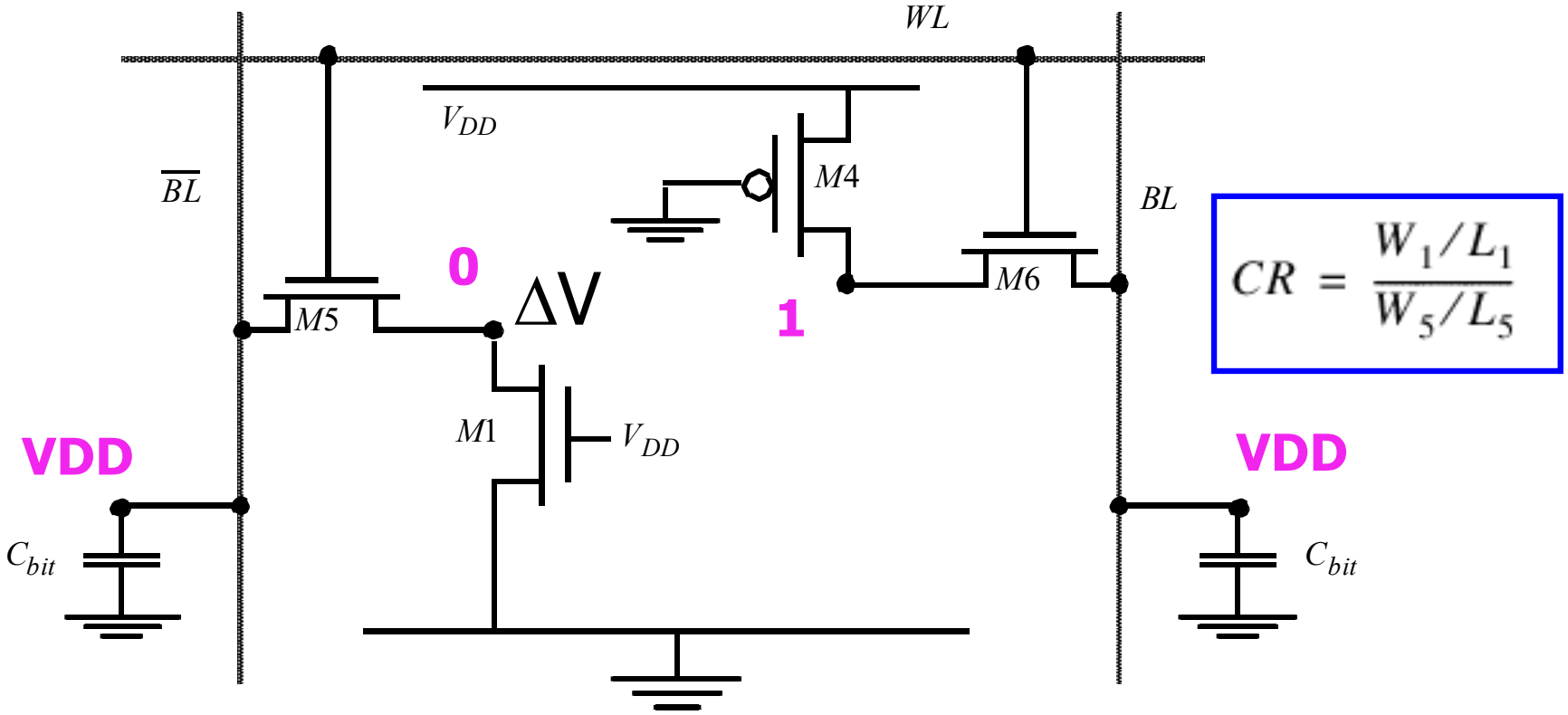
$$k_{n,M5}\left(V_{DD}-\Delta V-V_{Tn}\right)^2 = k_{n,M1}\left((V_{DD}-V_{Tn})\Delta V-\frac{\Delta V^2}{2}\right)$$

CMOS SRAM Analysis (Read)



$$\frac{k_{n,M1}}{k_{n,M5}} = \frac{\left(\frac{W}{L}\right)_1}{\left(\frac{W}{L}\right)_5} = \frac{\left(V_{DD} - \Delta V - V_{Tn}\right)^2}{(V_{DD} - V_{Tn})\Delta V - \frac{\Delta V^2}{2}}$$

CMOS SRAM Analysis (Read)

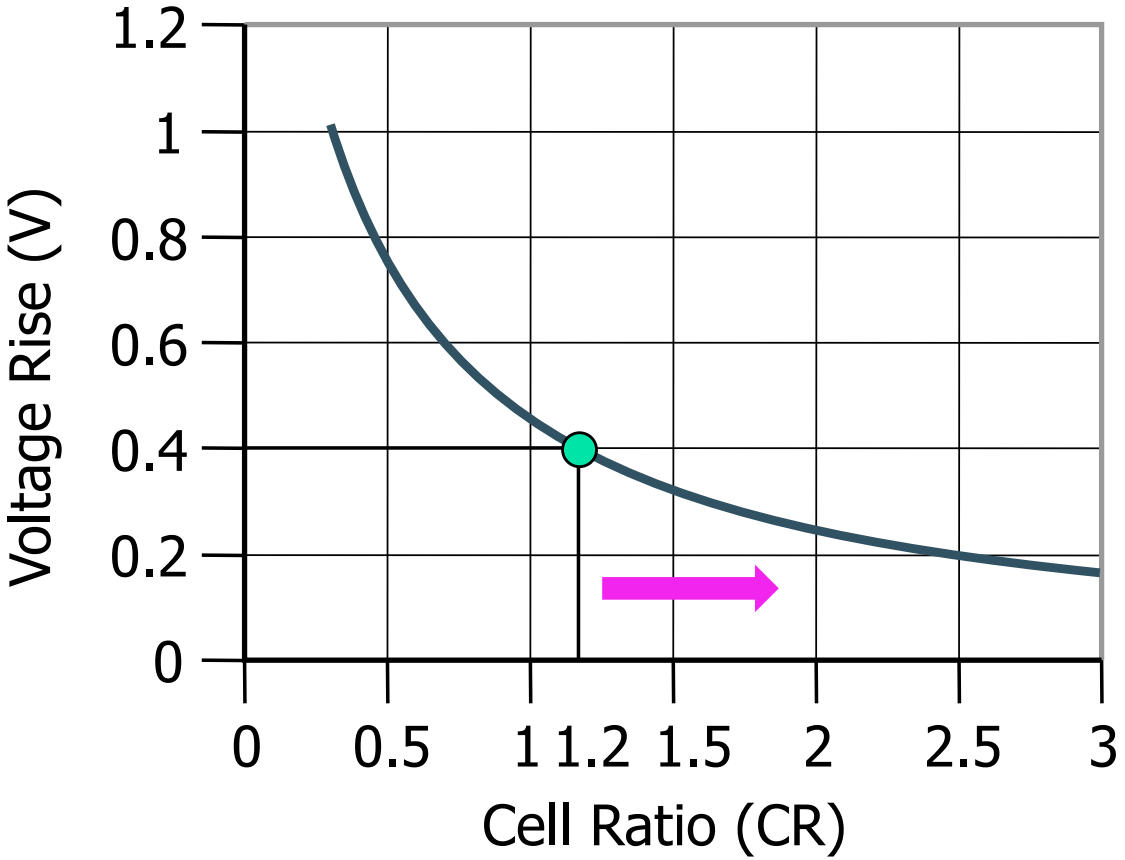


$$CR = \frac{W_1/L_1}{W_5/L_5}$$

$$\frac{k_{n,M1}}{k_{n,M5}} = \frac{\left(\frac{W}{L}\right)_1}{\left(\frac{W}{L}\right)_5} = \frac{\left(V_{DD} - \Delta V - V_{Tn}\right)^2}{(V_{DD} - V_{Tn})\Delta V - \frac{\Delta V^2}{2}}$$

$$\Delta V = V_{Tn} \Rightarrow \frac{\left(\frac{W}{L}\right)_1}{\left(\frac{W}{L}\right)_5} = \frac{\left(V_{DD} - 2V_{Tn}\right)^2}{(V_{DD} - 1.5V_{Tn})V_{Tn}}$$

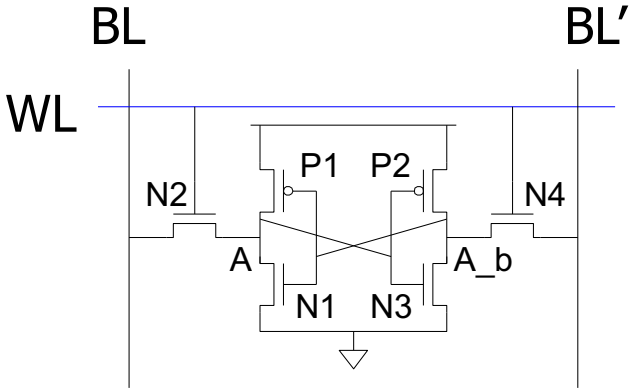
CMOS SRAM Analysis (Read)

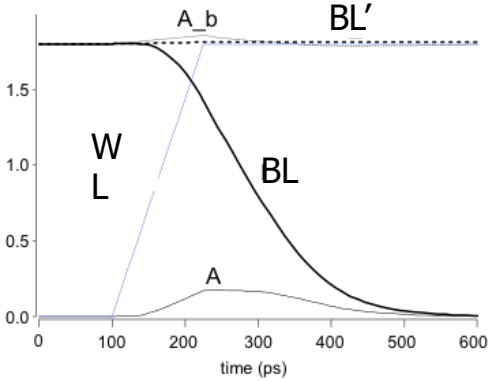


$$CR = \frac{W_1/L_1}{W_5/L_5}$$

SRAM Read

- Precharge both bitlines high
- Then turn on wordline, WL
- One of the two bitlines will be pulled down by the cell
- Ex: A = 0, A_b = 1
  - BL discharges, BL' stays high
  - But A bumps up slightly
- *Read stability*
  - A must not flip
  - N1 > N2



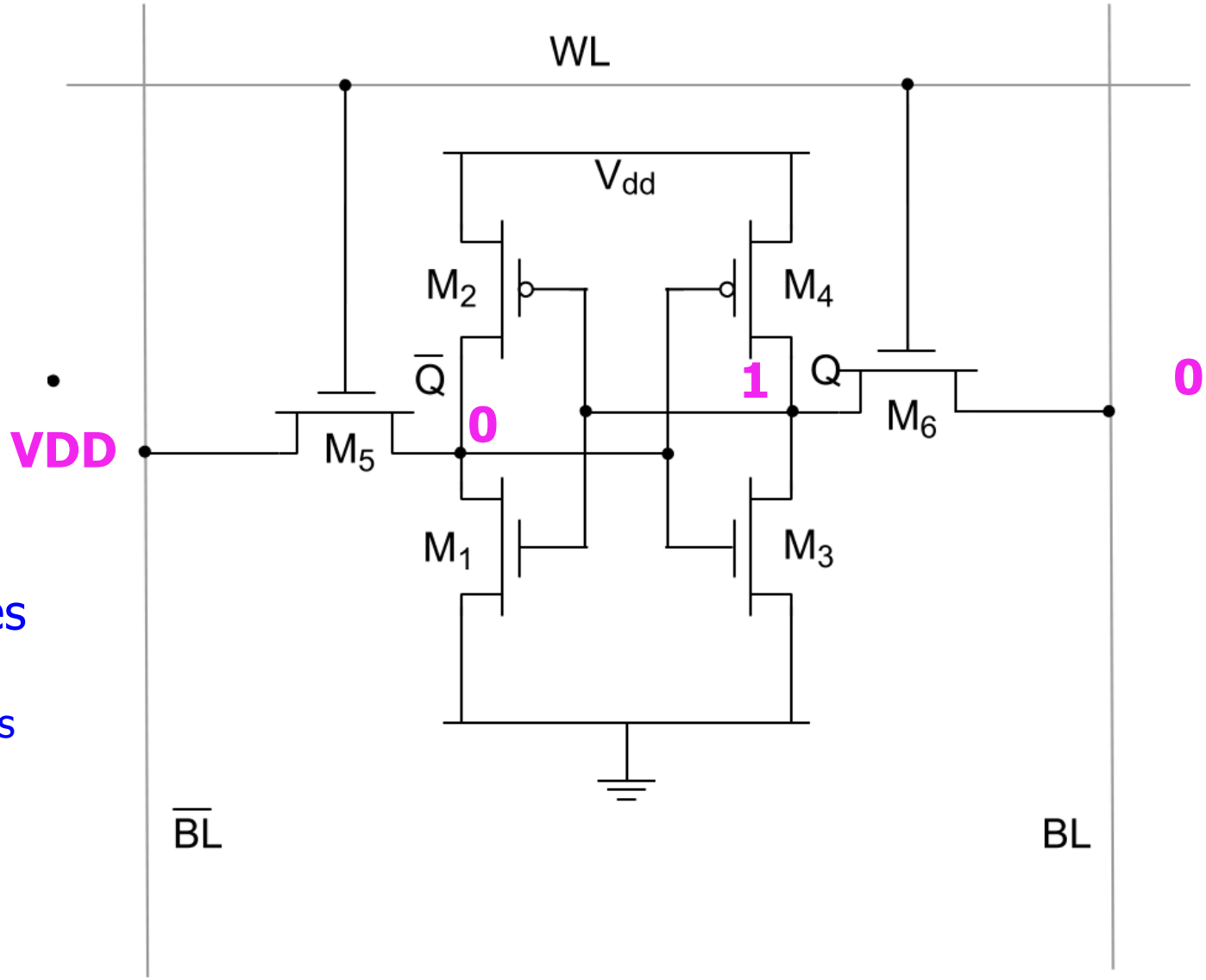


6-transistor CMOS SRAM Cell (preclass 2)

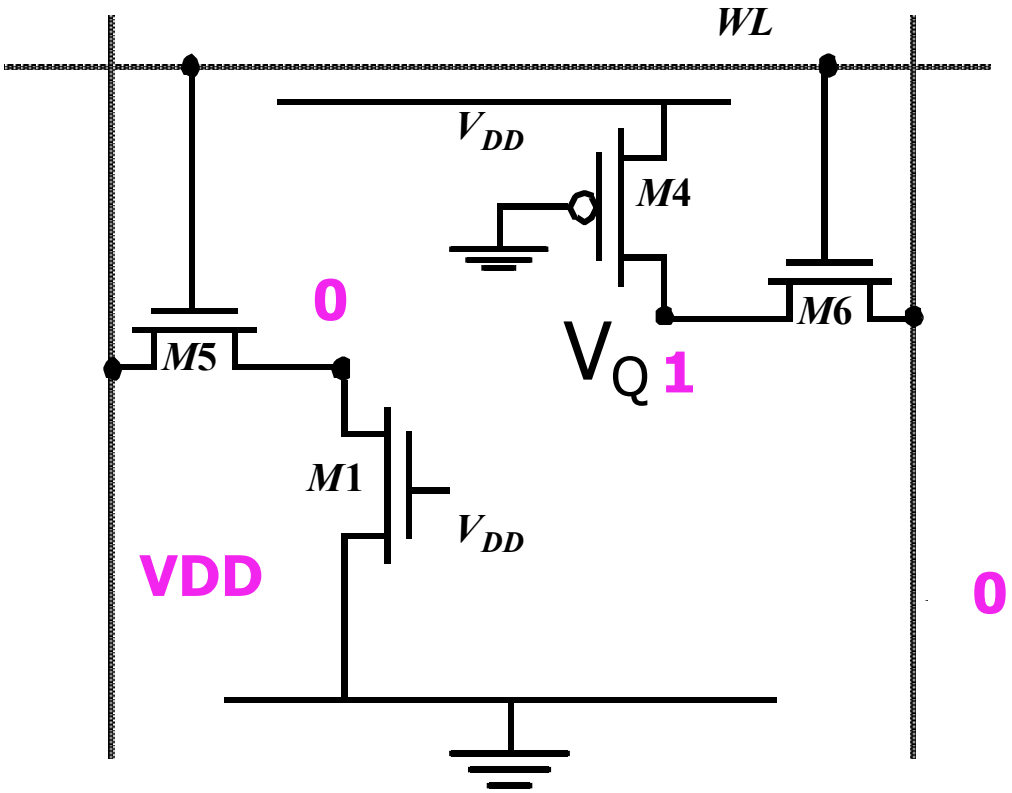
Assume 1 is stored (Q=1)

Write Operation:

- Want to write a 0
- First drive bitlines with input data
- Then wordline goes high (Vdd)
  - Still driving bitlines

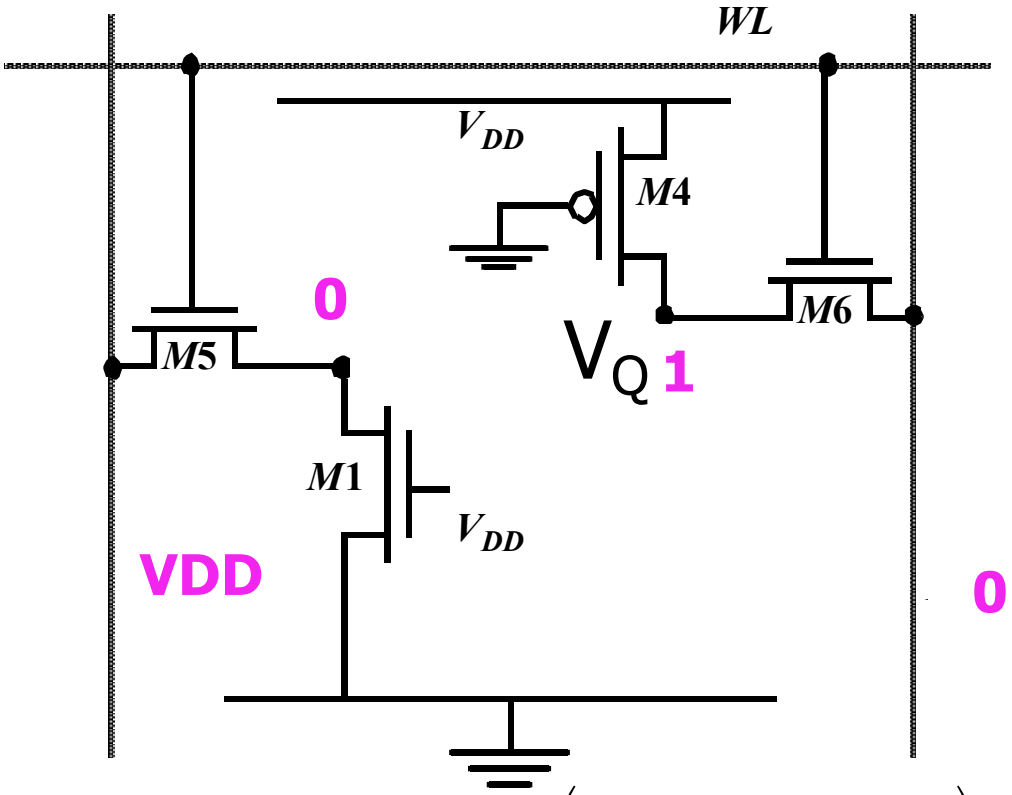


CMOS SRAM Analysis (Write) (preclass 2)



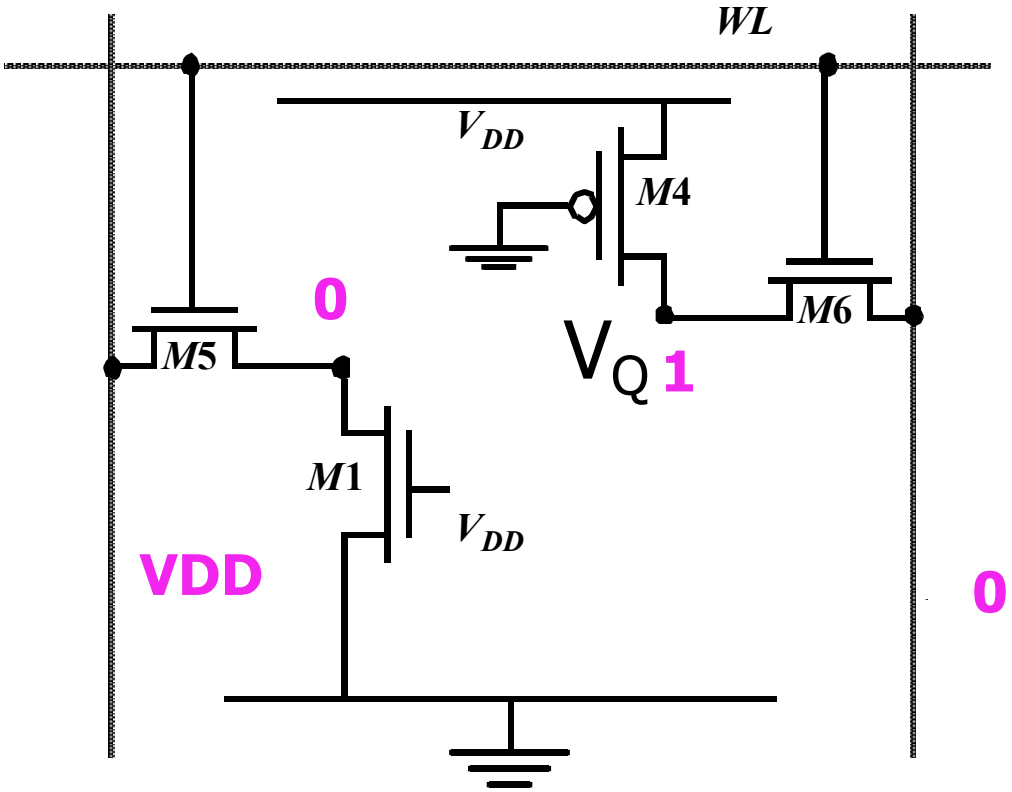
Which transistor discharges Q to Gnd?
Which transistor must it overpower?
What regions of operation are the transistors in?
Write the KCL equation for the two transistors

CMOS SRAM Analysis (Write)



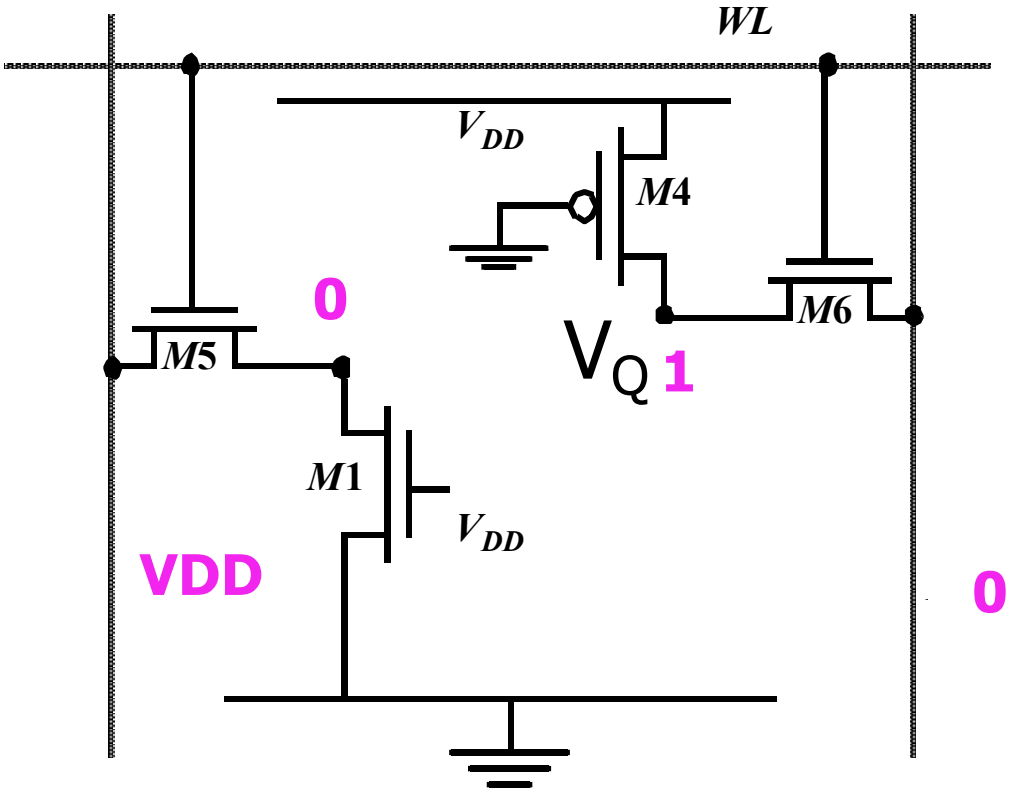
$$k_{p,M4}\left(V_{DD}-\left|V_{Tp}\right|\right)^2 = k_{n,M6}\left((V_{DD}-V_{Tn})V_Q-\frac{V_Q^2}{2}\right)$$

CMOS SRAM Analysis (Write)



$$\frac{k_{p,M4}}{k_{n,M6}} = \frac{(V_{DD} - V_{Tn})V_Q - \frac{V_Q^2}{2}}{\left(V_{DD} - |V_{Tp}|\right)^2}$$

CMOS SRAM Analysis (Write)



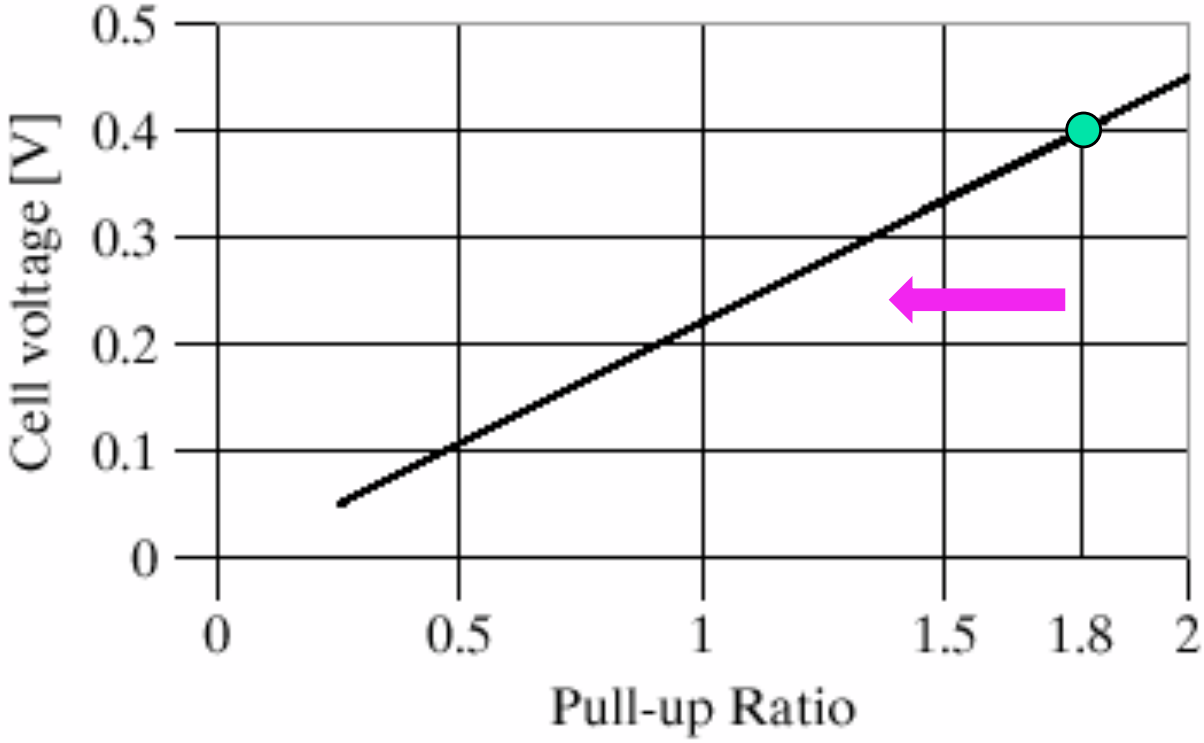
$$PR = \frac{W_4/L_4}{W_6/L_6}$$

$$\frac{k_{p,M4}}{k_{n,M6}} = \frac{(V_{DD} - V_{Tn})V_Q - \frac{V_Q^2}{2}}{\left(V_{DD} - |V_{Tp}|\right)^2}$$

$V_Q = V_{Tn}$

$$\frac{k_{p,M4}}{k_{n,M6}} = \frac{(V_{DD} - V_{Tn})V_{Tn} - \frac{V_{Tn}^2}{2}}{\left(V_{DD} - |V_{Tp}|\right)^2}$$

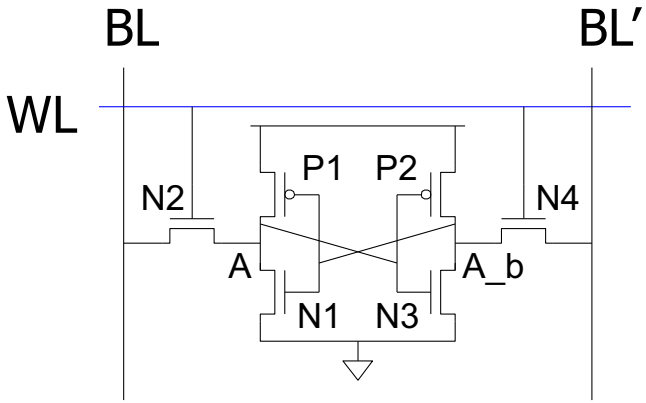
CMOS SRAM Analysis (Write)

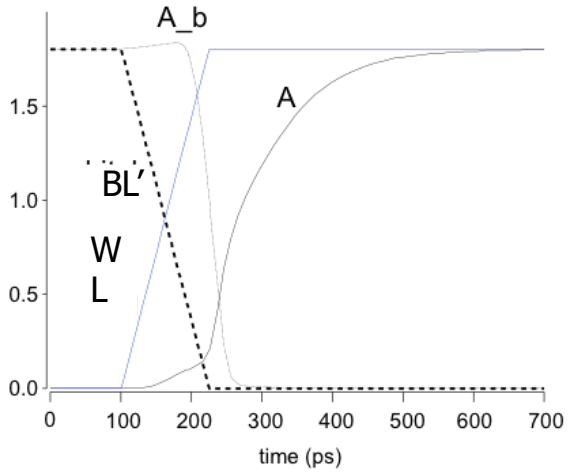


$$PR = \frac{W_4/L_4}{W_6/L_6}$$

SRAM Write

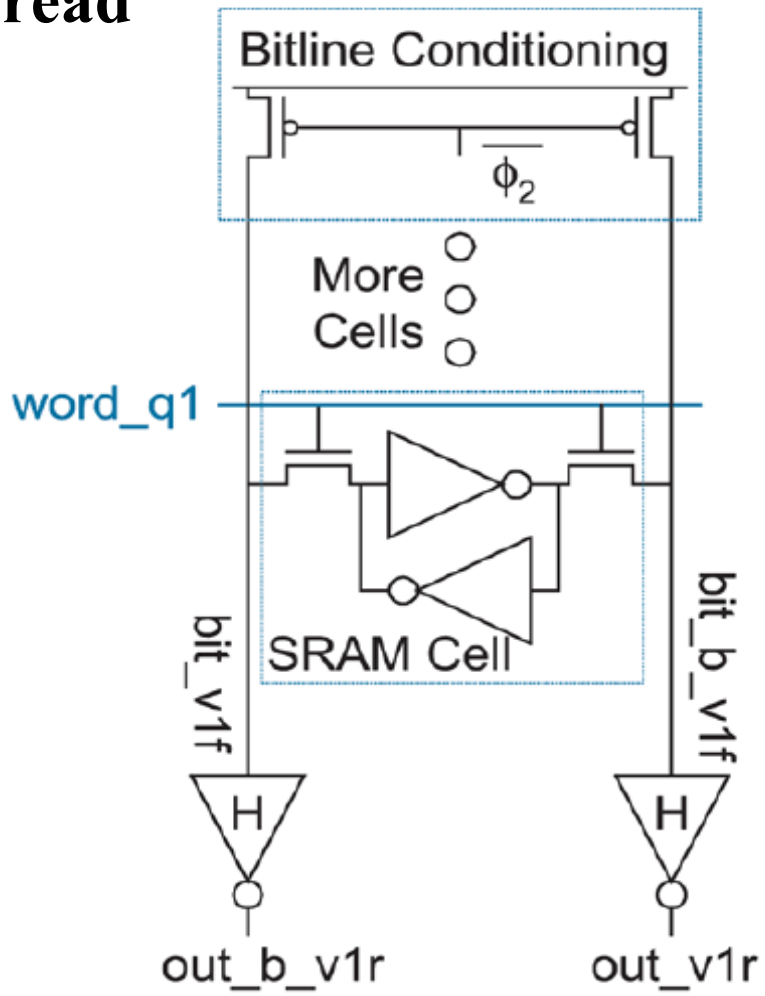
- Drive one bitline high, the other low
  - Depending on write data
- Then turn on wordline, WL
- Bitlines overpower cell with new value
- Ex: A = 0, A_b = 1, BL = 1, BL' = 0
  - Force A_b low, then A charges high
- *Writability*
  - Must overpower feedback inverter
  - N4 >> P2



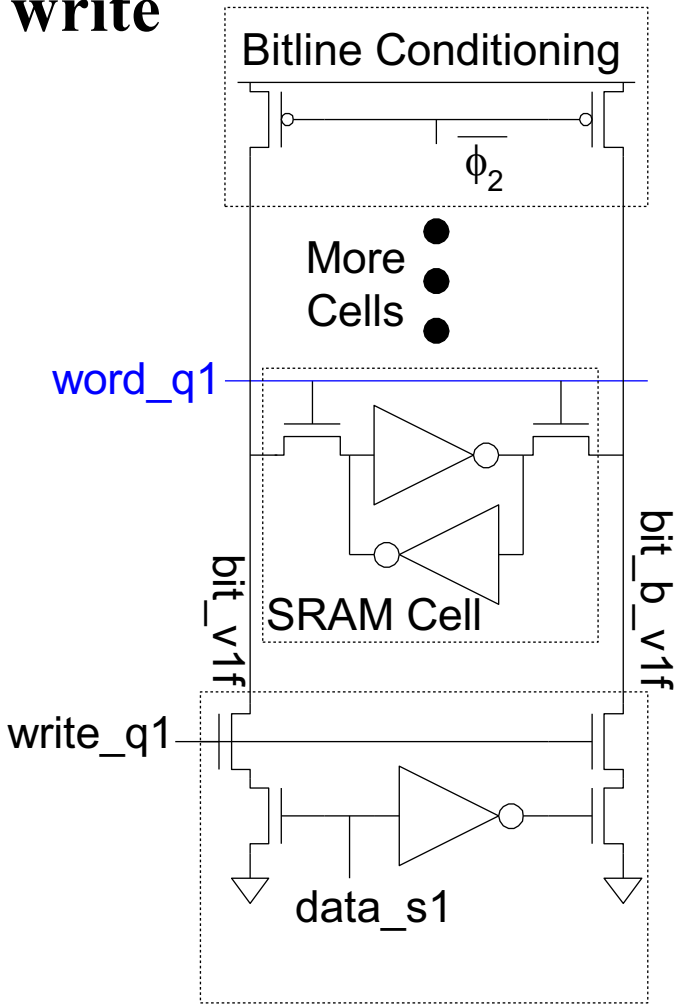


SRAM Column Example

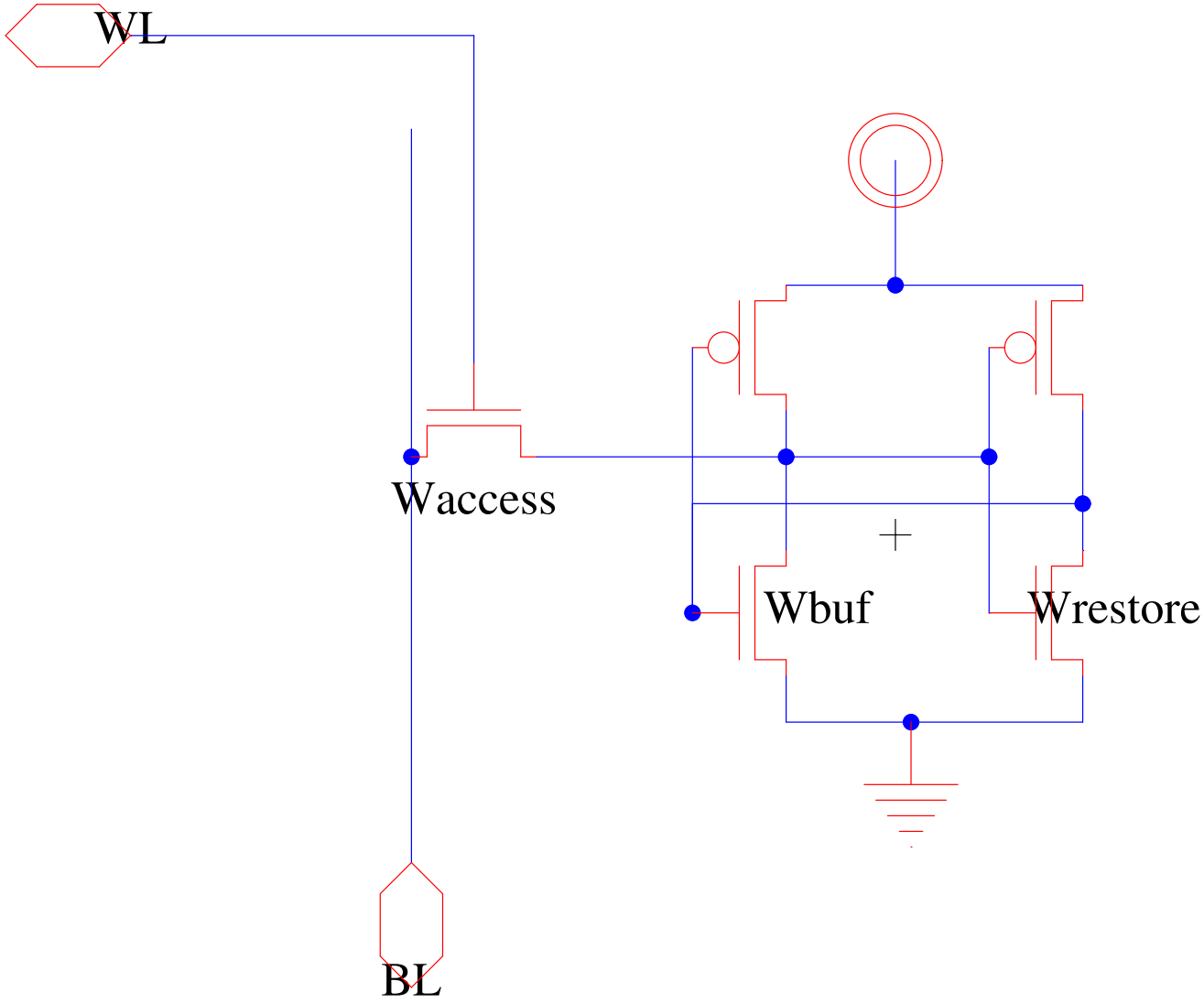
read



write



5T SRAM



Idea

- 6T SRAM
  - Robust cell when sized carefully
- 5T SRAM
  - More sensitive to sizing than 6T SRAM
  - More next time…

Admin

- Project 2 out - START NOW!
  - Work in teams up to 2
  - Milestone due Monday 11/22
    - I will give feedback by Wed (night) 11/24
  - Final report due 12/3