CENG4480

Lecture 09: Memory 2

Bei Yu

byu@cse.cuhk.edu.hk

(Latest update: November 26, 2020)

Fall 2020

香港中文大學
The Chinese University of Hong Kong

CENG4480 v.s. CENG3420

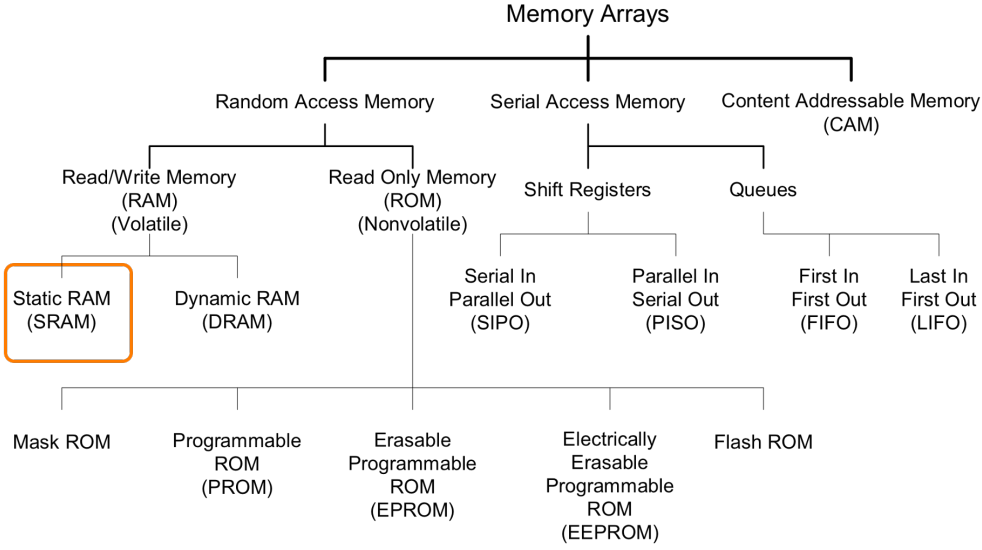
CENG3420:

- architecture perspective
- memory coherent
- data address

CENG4480:

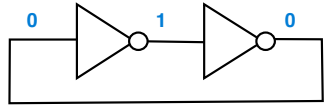
- more details on how data is stored

Memory Arrays

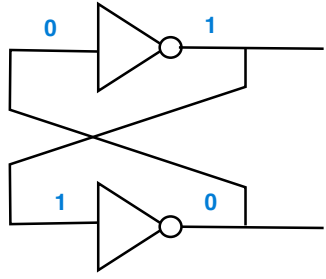


Memory Arrays

- What if we add feedback to a pair of inverters?

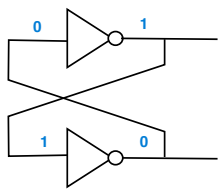


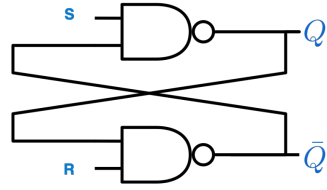
- Usually drawn as a ring of cross-coupled inverters
- Stable way to store one bit of information (w. power)



How to change the value stored?

- Replace inverter with NAND gate
- RS Latch

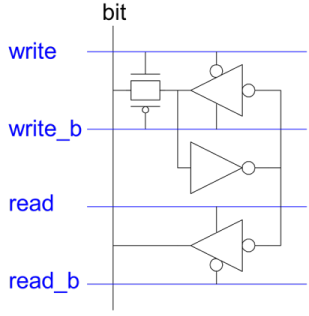


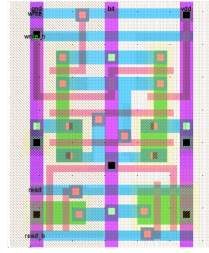


A	B	A nand B
0	0	1
0	1	1
1	0	1
1	1	0

12T SRAM Cell

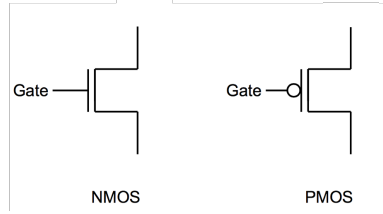
- Basic building block: SRAM Cell
  - Holds one bit of information, like a latch
  - Must be read and written
- 12-transistor (**12T**) SRAM cell
  - Use a simple latch connected to bitline
  - 46 $\times$ 75 λ unit cell

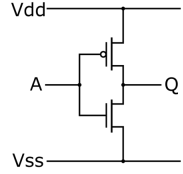




nMOS, pMOS, Inverter

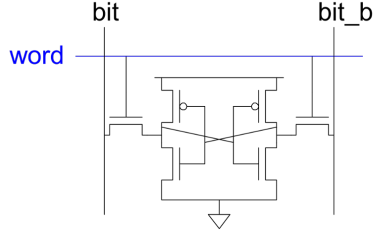
- nMOS:
  - Gate = 1, transistor is ON
  - Then electric current path
- pMOS:
  - Gate = 0, transistor is ON
  - Then electric current path
- Inverter:
  - Q = NOT (A)





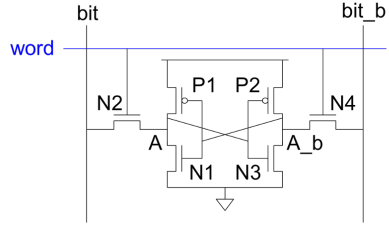
6T SRAM Cell

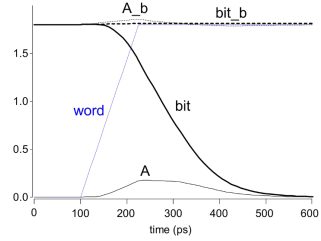
- Used in most commercial chips
- A pair of weak cross-coupled inverters
- Data stored in cross-coupled inverters
- Compared with 12T SRAM, 6T SRAM:
  - (+) reduce area
  - (-) much more complex control



6T SRAM Read

- Precharge both bitlines high
- Then turn on wordline
- One of the two bitlines will be pulled down by the cell
- Read stability
  - A must not flip
  - N1 >> N2

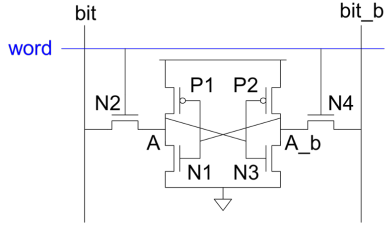




EX: 6T SRAM Read

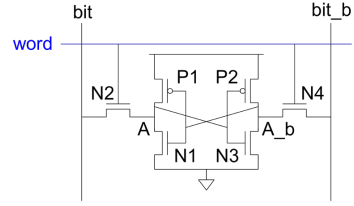
- Question 1: A = 0, A_b = 1, discuss the behavior:

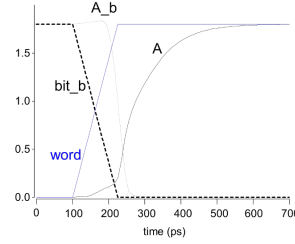
- Question 2: At least how many bit lines to finish read?



6T SRAM Write

- Drive one bitline high, the other low
- Then turn on wordline
- Bitlines overpower cell with new value
- Writability
  - Must overpower feedback inverter
  - N4 >> P2
  - N2 >> P1 (symmetry)

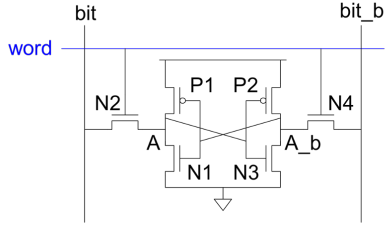




EX: 6T SRAM Write

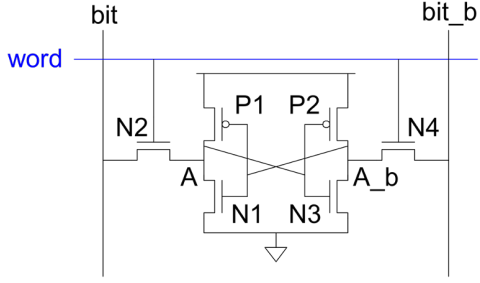
- Question 1: A = 0, A_b = 1, discuss the behavior:

- Question 2: At least how many bit lines to finish write?

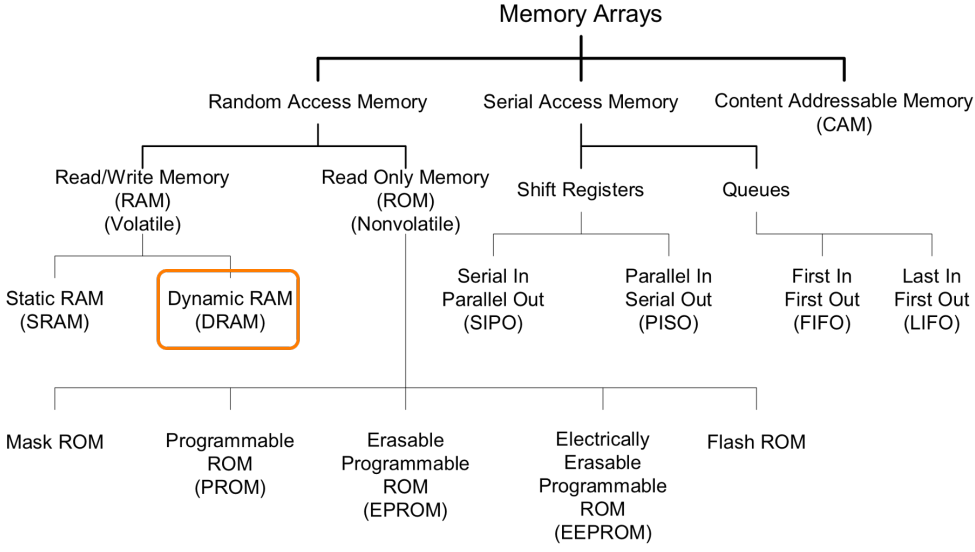


6T SRAM Sizing

- High bitlines must not overpower inverters during reads
- But low bitlines must write new value into cell



Memory Arrays

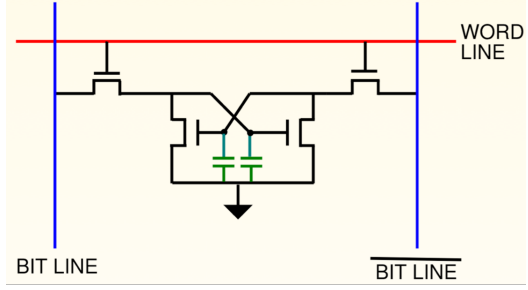


Dynamic RAM (DRAM)

- Basic Principle: Storage of information on capacitors
- Charge & discharge of capacitor to change stored value
- Use of transistor as "switch" to:
  - Store charge
  - Charge or discharge

4T DRAM Cell

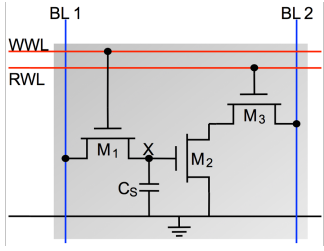
Remove the two p-MOS transistors from static RAM cell, to get a four-transistor dynamic RAM cell.

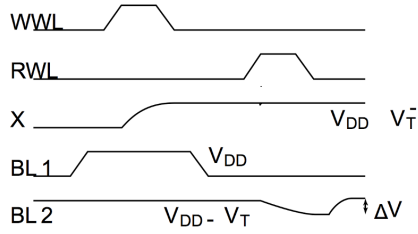


- Data must be refreshed regularly
- Dynamic cells must be designed very carefully
- Data stored as charge on gate capacitors (complementary nodes)

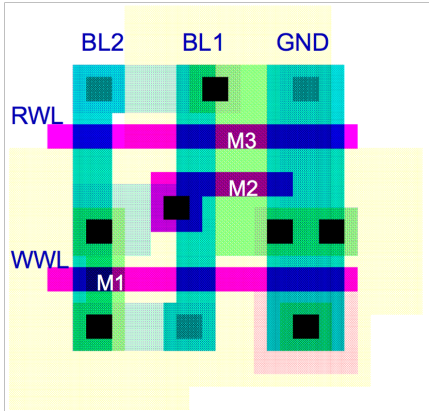
3T DRAM Cell

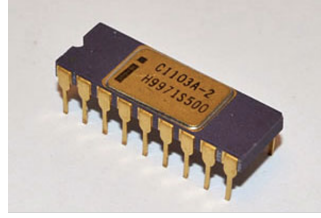
- No constraints on device ratios
- Reads are non-destructive
- Value stored at node X when writing a "1" = $V_{DD} - V_T$





3T DRAM Layout

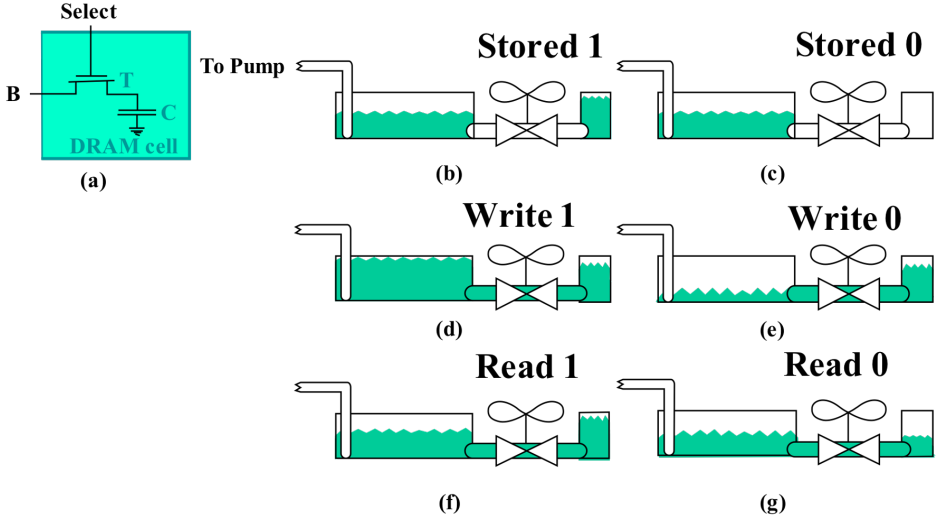




[1970: Intel 1003]

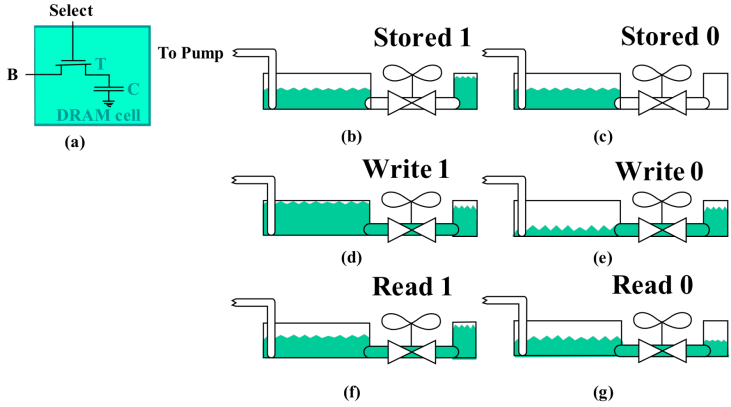
- 576 λ 3T DRAM v.s. 1092 λ 6T SRAM
- Further simplified

1T DRAM Cell



- Need sense amp helping reading

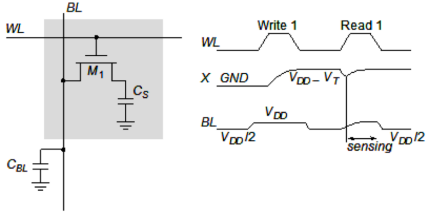
1T DRAM Cell

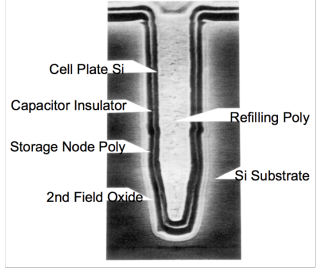


- Read
  - Pre-charge large tank to VDD2
  - If Ts = 0, for large tank: VDD2 - V1
  - If Ts = 1, for large tank: VDD2 + V1
  - V1 is very insignificant

1T DRAM Cell

- **Write:** Cs is charged or discharged by asserting WL and BL
- **Read:** Charge redistribution takes place between bit line and storage capacitance
- Voltage swing is small; typically around 250 mV





Trench-capacitor cell [Mano87]

EX. 1T DRAM Cell

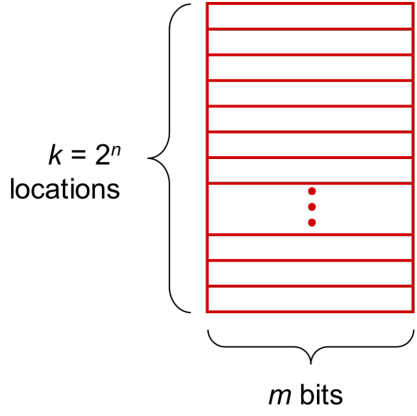
- Question: V_{DD}=4V, C_S=100pF, C_{BL}=1000pF. What's the voltage swing value?
- Note: $\Delta V = \frac{V_{DD}}{2} \cdot \frac{C_S}{C_S + CBL}$

SRAM v.s. DRAM

- **Static (SRAM)**
  - Data stored as long as supply is applied
  - Large (6 transistorscell)
  - Fast
  - Compatible with current CMOS manufacturing
- **Dynamic (DRAM)**
  - Periodic refresh required
  - Small (1-3 transistors/cell)
  - Slower
  - Require additional process for trench capacitance

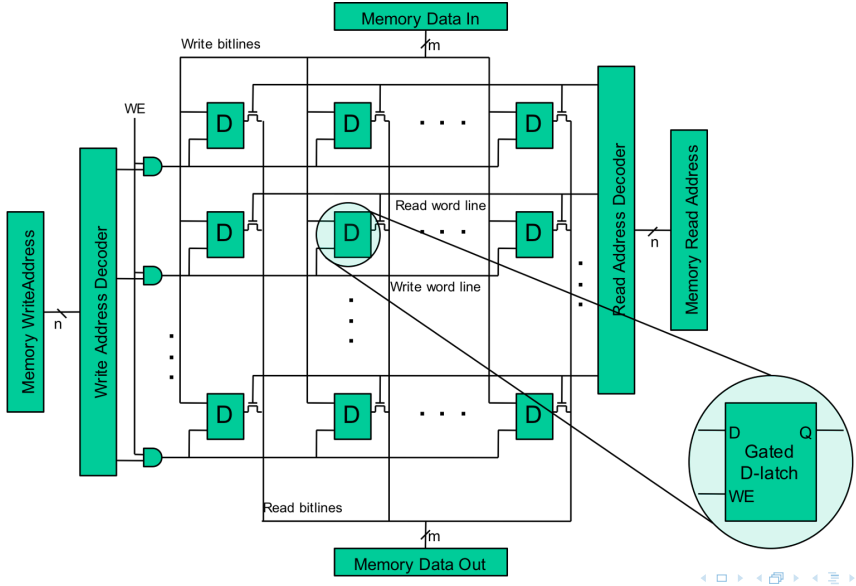
Array Architecture

- 2ˆn words of 2ˆm bits each
- Good regularity - easy to design



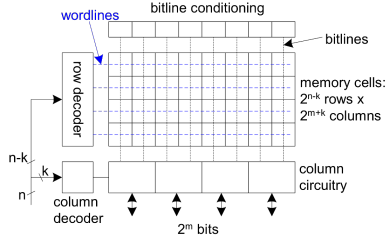
SRAM Memory Structure

- Latch based memory



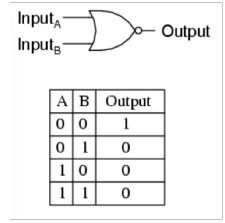
Array Architecture

- 2^n words of 2^m bits each
- How to design if n $>>$ m?
- Fold by 2k into fewer rows of more columns



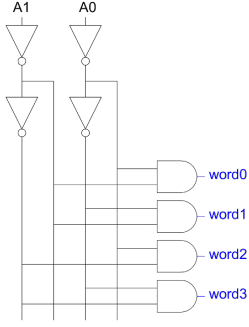
Decoders

- n:2^n decoder consists of 2^n n-input AND gates
  - One needed for each row of memory
  - Build AND with NAND or NOR gates

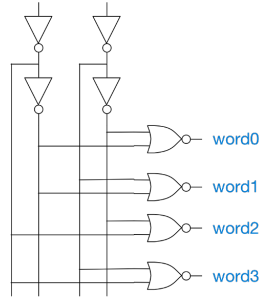


A	B	Output
0	0	1
0	1	0
1	0	0
1	1	0

Static CMOS

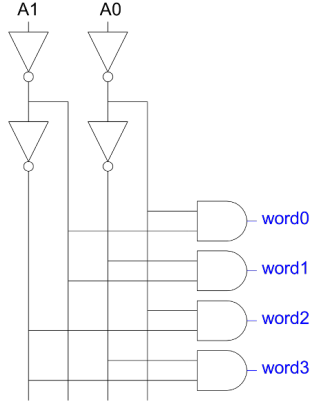


Using NOR gates



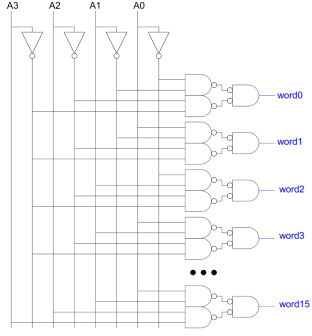
EX. Decoder

- Question: AND gates => NAND gate structure



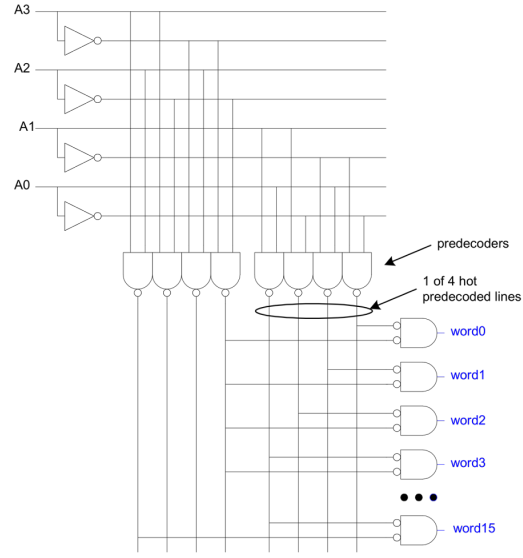
Larger Decoder

- For n > 4, NAND gates become slow
  - Break large gates into multiple smaller gates



Predecoding

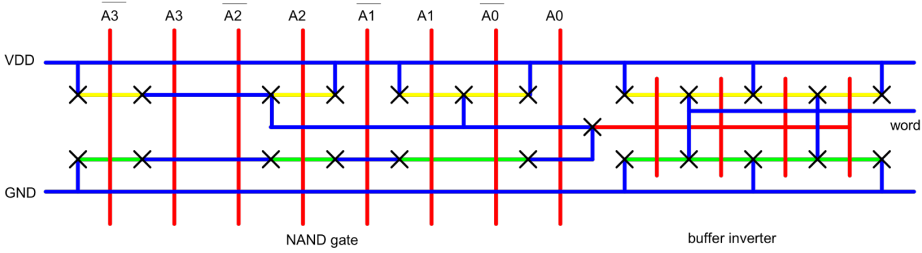
- Many of these gates are redundant
  - Factor out common gates
  - => Predecoder
  - Saves area
  - Same path effort



- Question: How many NANDs can be saved?

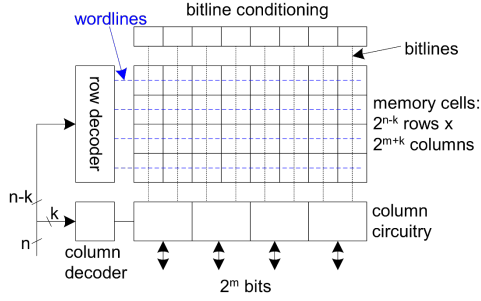
*Decoder Layout

- Decoders must be pitch-matched to SRAM cell
  - Requires very skinny gates



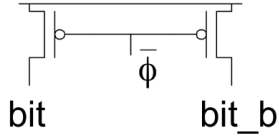
*Column Circuitry

- Some circuitry is required for each column
  - Bitline conditioning
  - Column multiplexing
  - Sense amplifiers (DRAM)

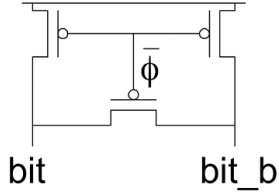


*Bitline Conditioning

- Precharge bitlines high before reads

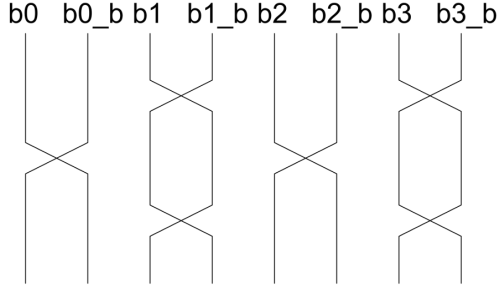


- Equalize bitlines to minimize voltage difference when using sense amplifiers



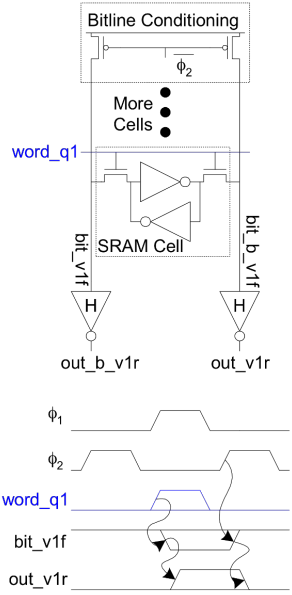
*Twisted Bitlines

- Sense amplifiers also amplify noise
  - Coupling noise is severe in modern processes
  - Try to couple equally onto bit and bit_b
  - Done by twisting bitlines

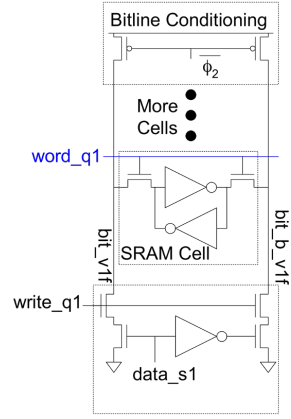


*SRAM Column Example

read



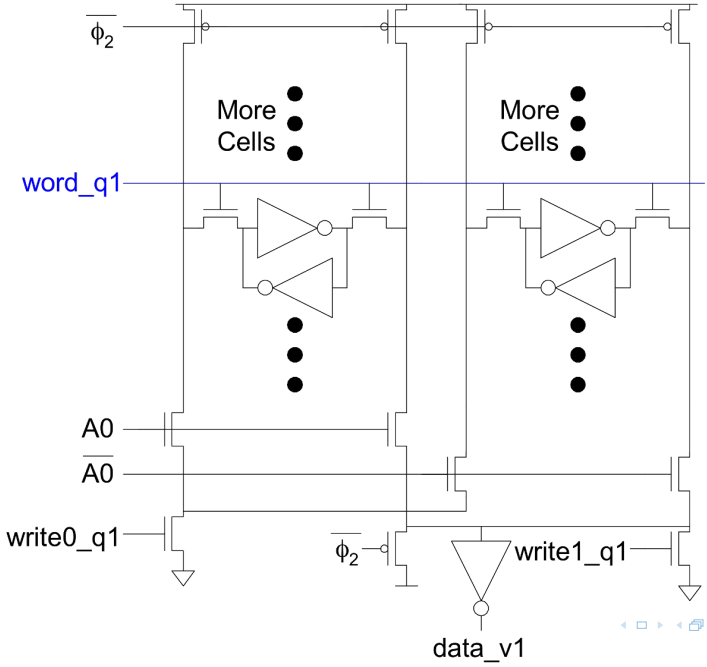
write



*Column Multiplexing

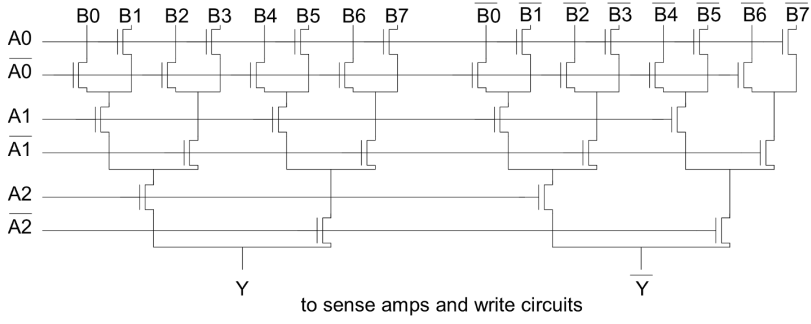
- Recall that array may be folded for good aspect ratio
- Ex: 2 kword x 16 folded into 256 rows x 128 columns
  - Must select 16 output bits from the 128 columns
  - Requires 16 8:1 column multiplexers

*Ex: 2-way Muxed SRAM

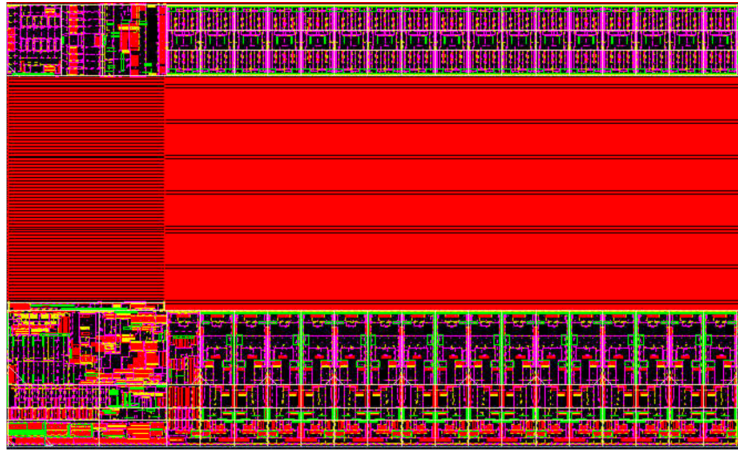


*Tree Decoder Mux

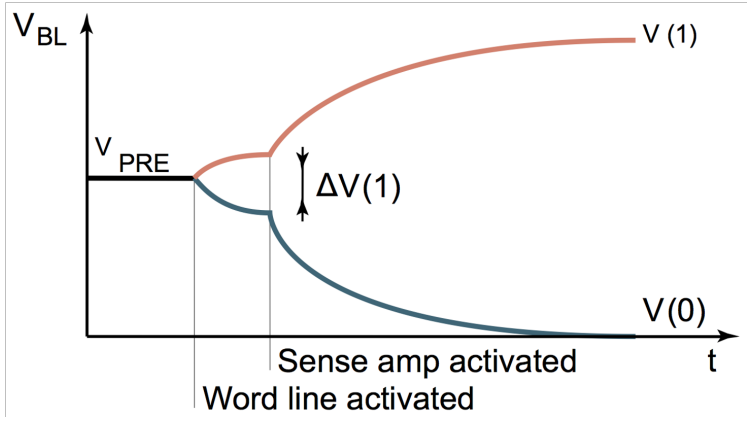
- Column mux can use pass transistors
  - Use nMOS only, precharge outputs
- One design is to use k series transistors for 2^k:1 mux
  - No external decoder logic needed



*SRAM from ARM



Sense Amp Operation for 1T DRAM



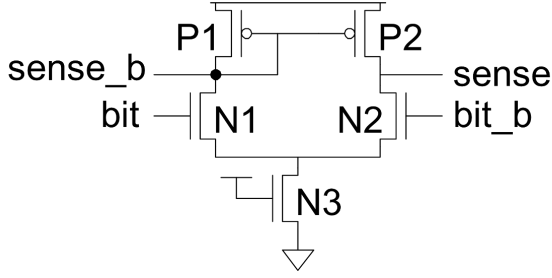
- 1T DRAM read is destructive
- Read and refresh for 1T DRAM

*Sense Amplifiers (DRAM)

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

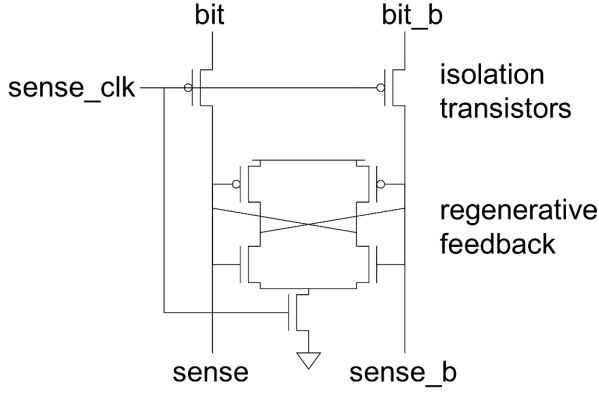
*Differential Pair Amp

- 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



Thank You :)