EECS 427
Lecture 17: Memory core and peripherals

Reading: 12.1-12.3

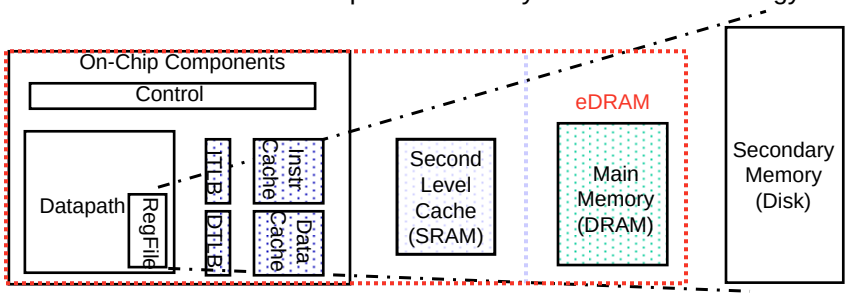
Overview

- SRAM and DRAM basics
- Decoders
- Sense amplifiers
- Our focus is mainly on SRAM

- HW 4 due now

A Typical Memory Hierarchy

- By taking advantage of the principle of locality:
  - Present the user with as much memory as is available in the cheapest technology
  - Provide access at the speed offered by the fastest technology

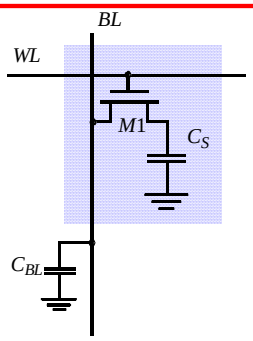


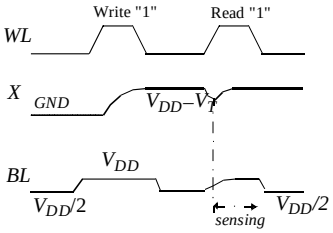
Speed (ns):	0.1's	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	T's
Cost:	highest				lowest

Read-write Memory Review

- SRAM
  - Data is stored as long as power is supplied
  - Relatively large cells, 6-transistors, lower density (vs. DRAM)
  - Fast – use closer to computation
  - Compatible with CMOS technology
- DRAM
  - Data must be periodically refreshed
  - Small cells, 1 transistor, VERY dense
  - Slower, use in larger main memories
  - Process not compatible with standard CMOS

1-Transistor DRAM cell review



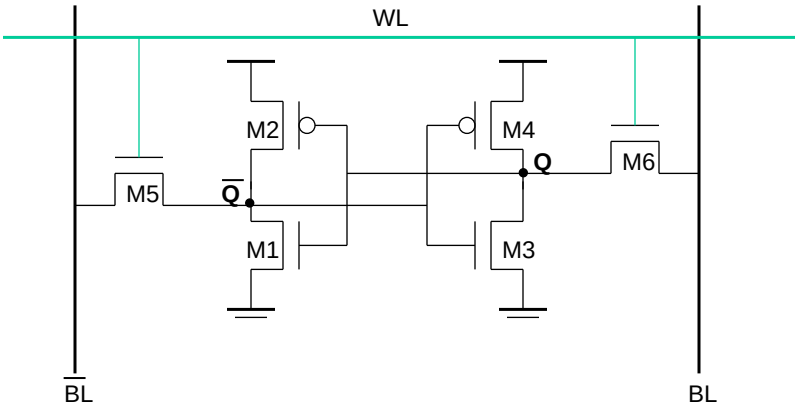


Write: C_S is charged or discharged by asserting WL and BL.
Read: Charge redistribution takes places between bit line and storage capacitance

$$\Delta V = V_{BL} - V_{PRE} = (V_{BIT} - V_{PRE})\frac{C_S}{C_S + C_{BL}}$$

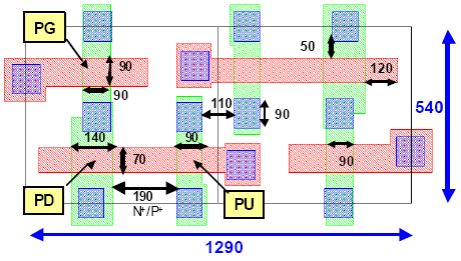
Voltage swing is small; typically around 250 mV.

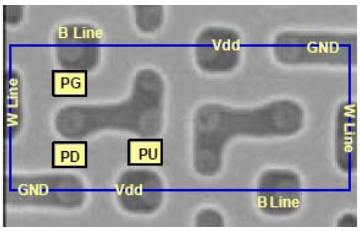
CMOS 6T SRAM cell review



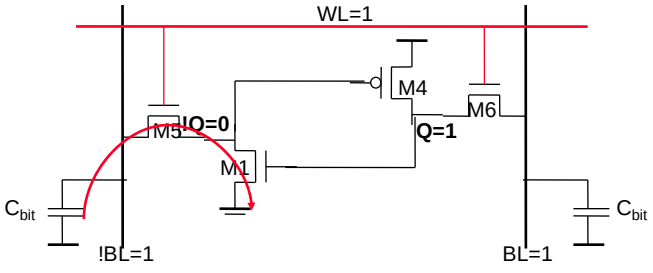
Example layout

- Freescale 65nm cell, 0.69μm^2





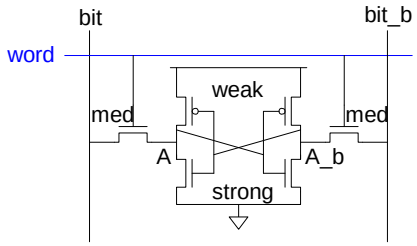
Stability: Read Upset



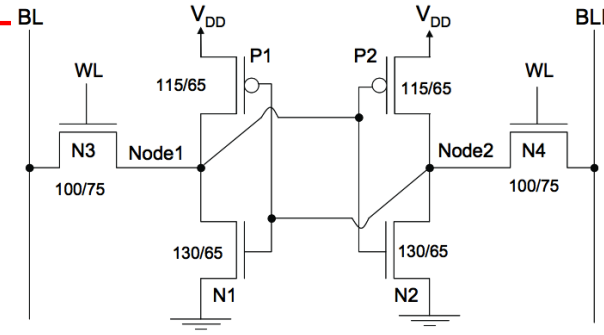
Read-disturb (read-upset): must carefully limit the allowed voltage rise on !Q to a value that prevents the read-upset condition from occurring while simultaneously maintaining acceptable circuit speed and area constraints

SRAM Sizing

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

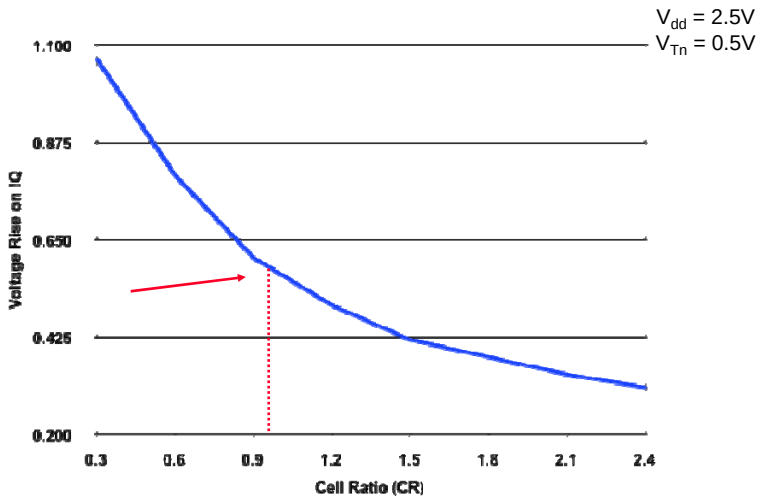


Example 65 nm Sizing

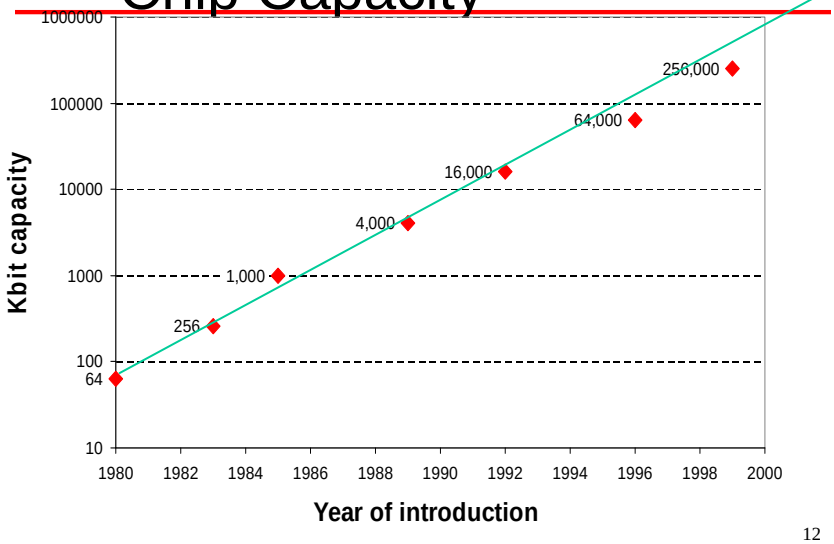


High Read Stability and Low Leakage Cache Memory Cell
Zhiyu Liu; Kursun, V.;
Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on
27-30 May 2007 Page(s):2774 - 2777

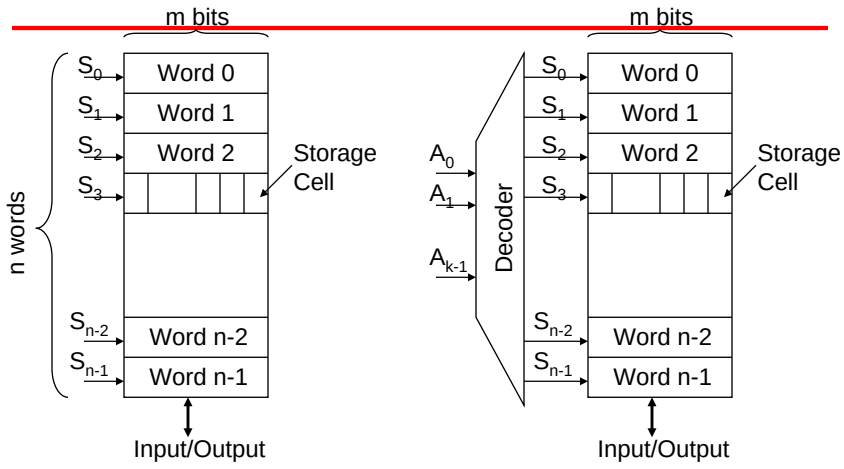
Stability: Read Upset



Growth in DRAM Chip Capacity



1D Memory Architecture

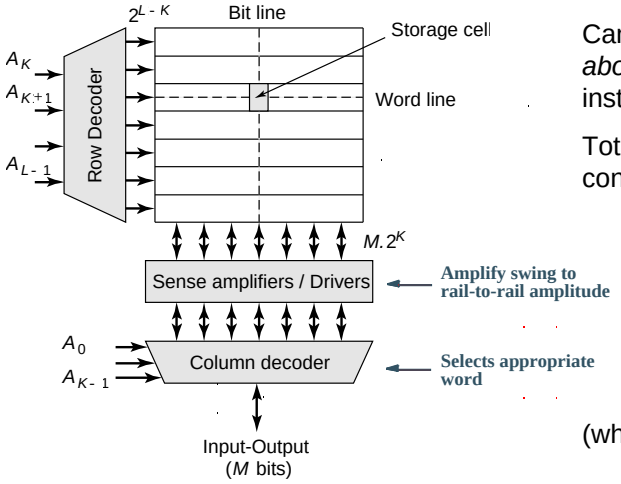


n words → n select signals
too many signals

Decoder reduces # of inputs
$k = \log_2 n$

2D (array) Memory Architecture

Problem: ASPECT RATIO or HEIGHT >> WIDTH



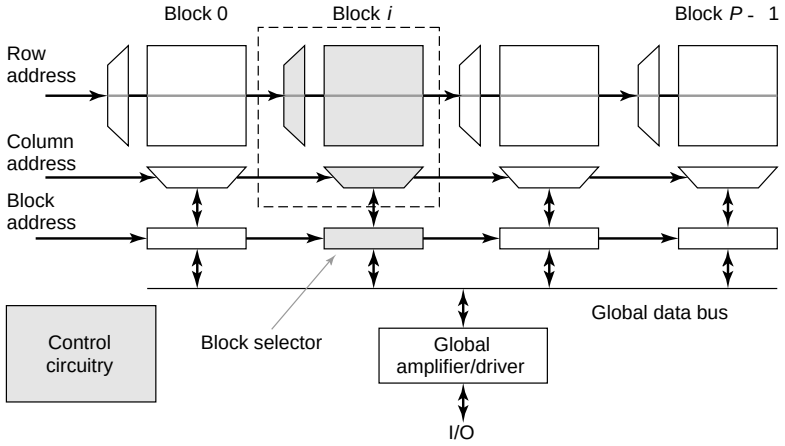
Can put column decode *above/before* sense amps instead

Total memory access time consists of many components

- Row decode
- Word line drivers
- Cell driving bit line cap
- Sense amp delay
- Column decode
- Output driving circuitry

(which dominate?)

3D Memory Architecture



Advantages:

1. Shorter wires within blocks

2. Block address activates only 1 block => power savings

Peripheral Components of Memories

- **Decoders (both row & column)**
- **Sense Amplifiers**

Not discussing:

- **Input/Output Buffers**
- **Control / Timing Circuitry**

Row Decoders (M to 2^M)

Simplest visualization: Collection of 2^M complex logic gates Organized in regular and dense fashion

(N)AND Decoder

$$WL_0 = \bar{A}_0\bar{A}_1\bar{A}_2\bar{A}_3\bar{A}_4\bar{A}_5\bar{A}_6\bar{A}_7\bar{A}_8\bar{A}_9$$

$$WL_{511} = \bar{A}_0A_1A_2A_3A_4A_5A_6A_7A_8A_9$$

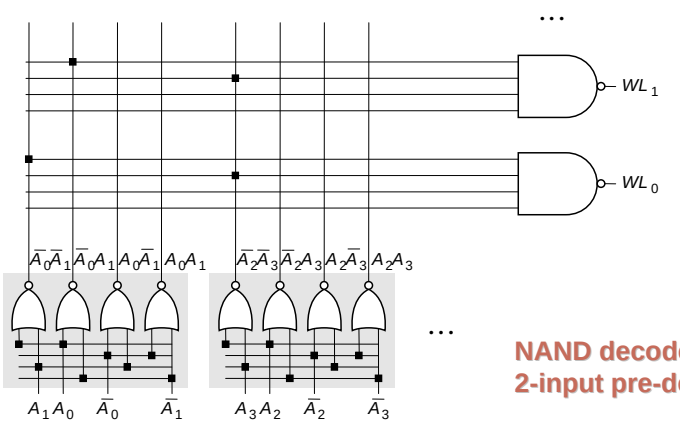
NOR Decoder

$$WL_0 = \overline{A_0 + A_1 + A_2 + A_3 + A_4 + A_5 + A_6 + A_7 + A_8 + A_9}$$

$$WL_{511} = \overline{A_0 + \bar{A}_1 + \bar{A}_2 + \bar{A}_3 + \bar{A}_4 + \bar{A}_5 + \bar{A}_6 + \bar{A}_7 + \bar{A}_8 + \bar{A}_9}$$

Hierarchical Decoders

Multi-stage implementation improves performance

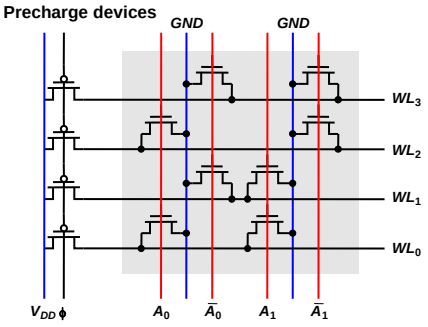


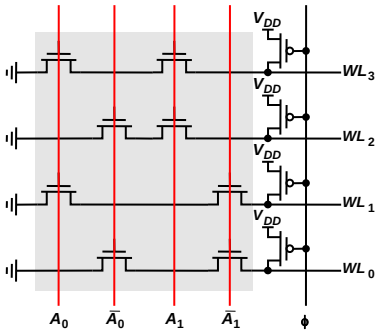
NAND decoder using 2-input pre-decoders

Decoder design goals

- Only 1 critical transition
  - 1 signal must go HIGH, going low again is not as crucial
- Can skew the gate sizing
- Can use dynamic logic with precharge

Dynamic Decoders

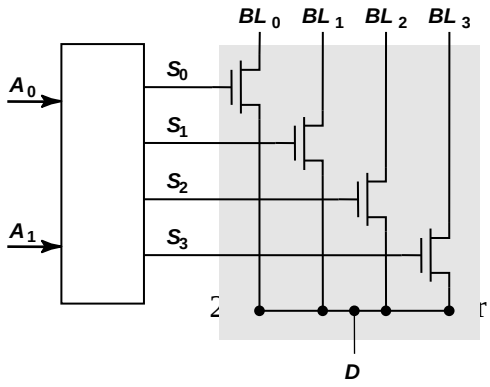




2-input NOR decoder **2-input NAND decoder**

NOR is faster but consumes more power since all but one word line pulls down each cycle

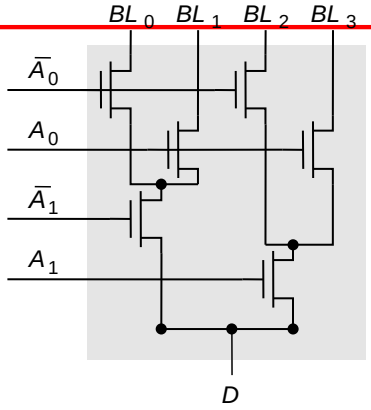
4-input pass-transistor based **column** decoder



Often use PMOS pass transistors instead since BLs precharge high (speed penalty)

Advantages: speed (Decode portion doesn't add to overall memory access time)
Only one extra transistor in signal path
Disadvantage: Large transistor count

4-to-1 tree based **column** decoder



Number of devices drastically reduced
Delay increases quadratically with # of sections; prohibitive for large decoders
Solutions: buffers
progressive sizing
combination of tree and pass transistor approaches

Sense Amplifiers

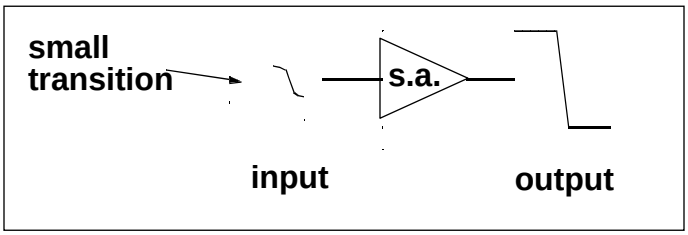
$$t_p = \frac{C \times \Delta V}{I_{av}}$$

make ΔV as small as possible

(b/c of large arrays) large

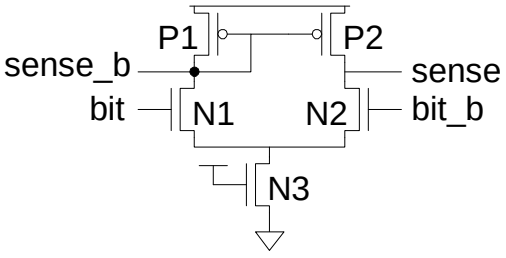
small devices must be small for density

Idea: Use Sense Amplifier

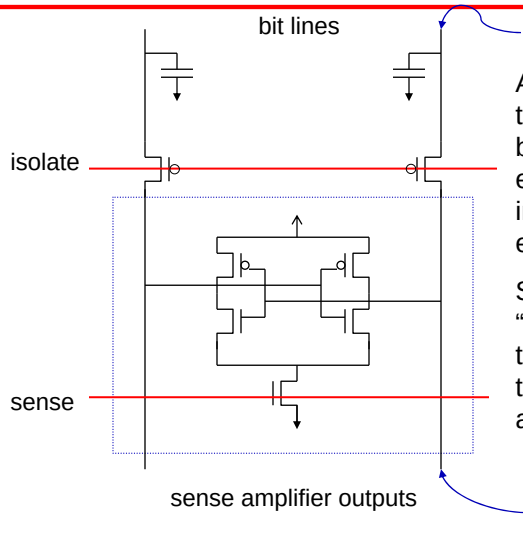


Differential Pair Amp

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



SRAM sense amp design



$\Delta V \approx 0.1V_{dd}$

An equalization pass-transistor often placed b/w BL and $\overline{BL}$ to ensure proper initialization before evaluation

Sense signal must "fire" at appropriate time: Too fast = errors, too slow = larger access time

$\Delta V = V_{dd}$

Summary

- Memory performance is critical to overall system performance
  - Memory hierarchy develops based on speed and size requirements
- On-chip SRAM very common today
  - 6T SRAM cells have became very compact
  - Complete memory architecture involves arrays + row/column decoders, sense amps, output drivers