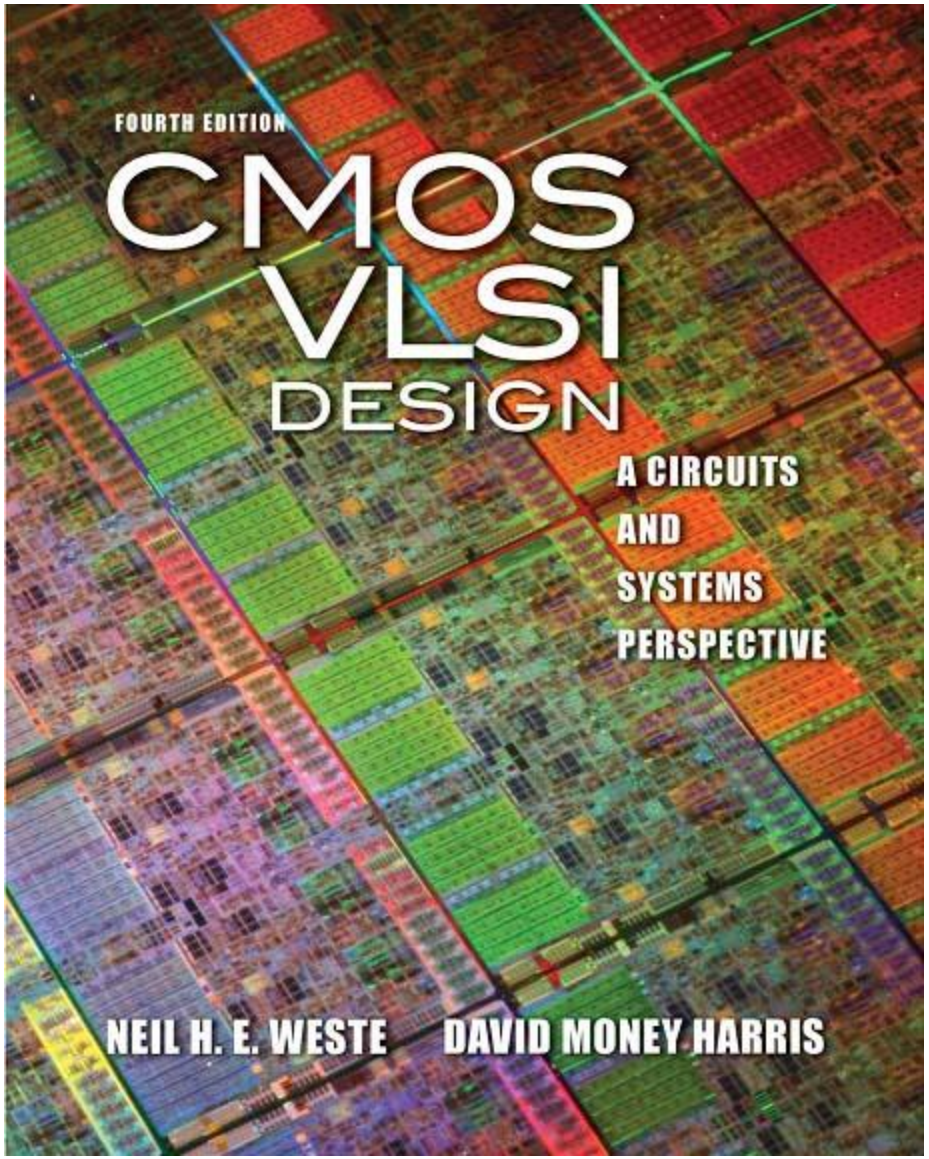
CPE 110408423
VLSI Design
Chapter 12: Array Subsystems

Bassam Jamil

[Computer Engineering Department,
Hashemite University]

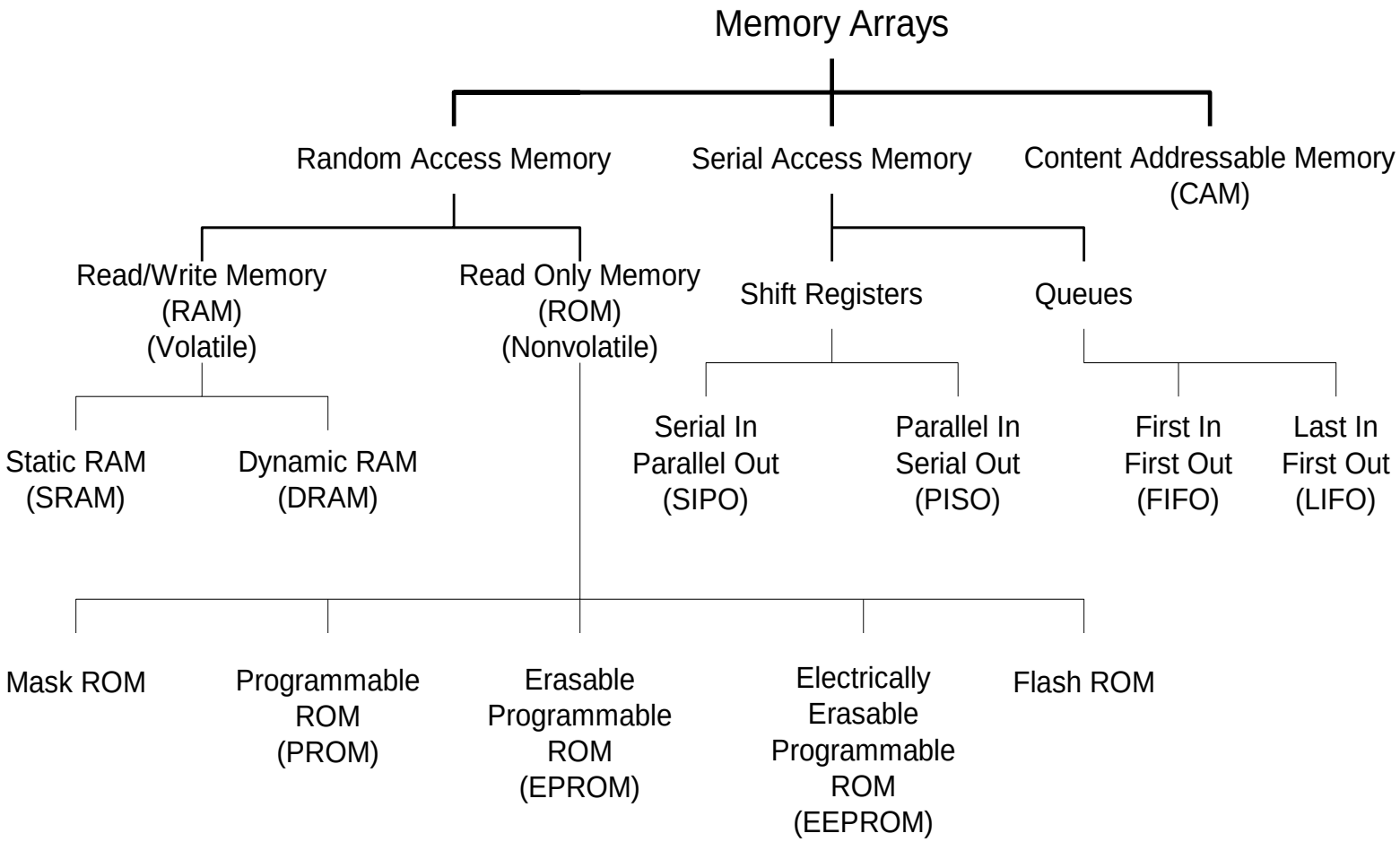


Lecture 3: SRAM

Outline

- Memory Arrays
- SRAM Architecture
  - SRAM Cell
  - Decoders
  - Column Circuitry
  - Multiple Ports

12.1 Memory Arrays



Memory Arrays

- *Random access memory (RAM)* is accessed with an address and has a latency independent of the address.
- *Serial access memories (SAM)* are accessed sequentially so no address is necessary.
- *Content addressable memories (CAM)* determine which address(es) contain data that matches a specified key.
- Volatile vs. nonvolatile memory.
  - Volatile memory retains its data as long as power is applied,
  - nonvolatile memory will hold data indefinitely.
  - RAM is synonymous with volatile memory, while ROM is synonymous with nonvolatile memory.

Volatile Memory

- The memory cells used in volatile memories can be divided into **static** structures and **dynamic** structures.
- **Static cells** use some form of feedback to maintain their state,
- **Dynamic cells** use charge stored on a floating capacitor through an access transistor.
  - Charge will leak away through the access transistor even while the transistor is OFF,
  - So dynamic cells must be periodically read and rewritten to refresh their state.
- Static RAMs (SRAMs) are **faster** and **less troublesome**, but require **more area per bit** than their dynamic counterparts (DRAMs).

Non-Volatile Memory

- **Volatility:**
  - Some nonvolatile memories are read-only (e.g. mask ROM), but many nonvolatile memories can be written, albeit more slowly than their volatile memories.
- **Types of non-volatile memories:**

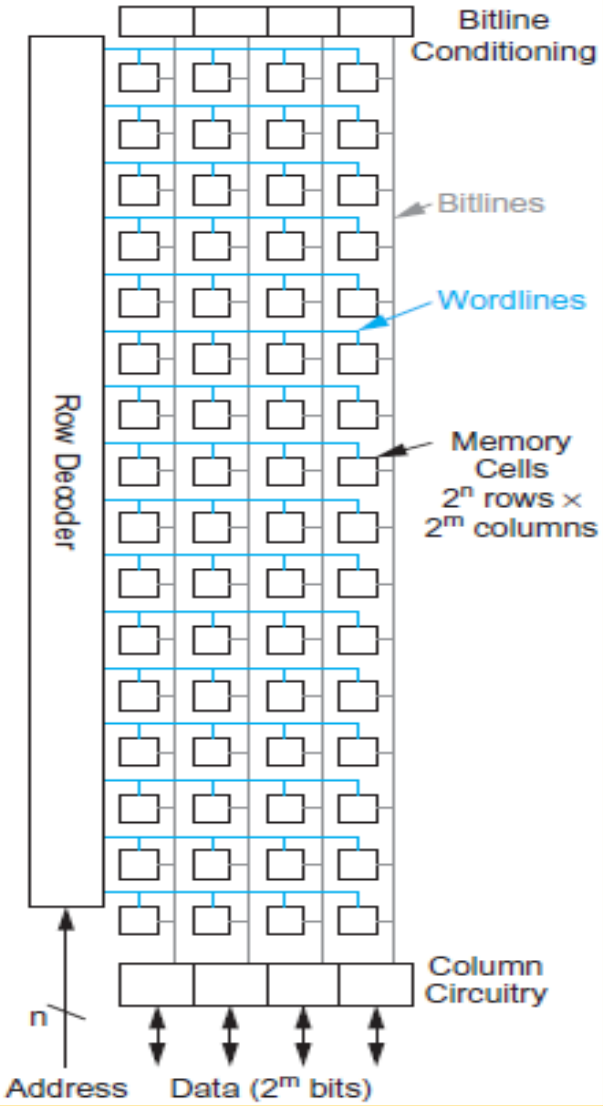
1. **Mask ROM**: the contents of a mask ROM are **hardwired** during fabrication and cannot be changed.
2. **Programmable ROM (PROM)** is **programmed** once after fabrication by blowing on-chip fuses with a special high programming voltage.
3. **Erasable programmable ROM (EPROM)** is programmed by storing charge on a floating gate.
   - It can be erased by exposure to ultraviolet (UV) light for several minutes to knock the charge off the gate. Then the EPROM can be reprogrammed.
4. **Electrically erasable programmable ROMs (EEPROMs)** are erased in microseconds with on-chip circuitry.
5. **Flash** memories are a variant of EEPROM that erases entire blocks rather than individual bits.
   - Sharing the erase circuitry across larger blocks reduces the area per bit.
   - Because of their good density and easy in-system reprogrammability, Flash memories have replaced other nonvolatile memories in most modern CMOS systems.

Memory Array Architecture

- Memory cells can have one or more ports for access.
- On a read/write memory, each port can be read-only, write-only, or both.
- A memory array contains 2^n words of 2^m bits each.
  - Total number of bits = $2^n \times 2^m$ bits
- memory array containing 16 4-bit words (n = 4, m = 2).
  - Total number of bits = 64 bits

Memory Array :One row per word and One column per bit

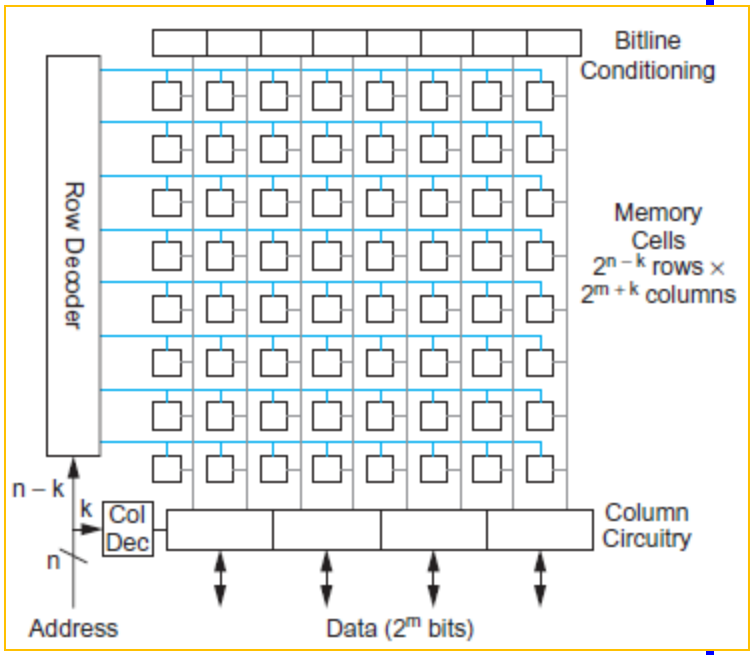
- The simplest design with one row per word and one column per bit.
- The **row decoder** uses the address to activate one of the rows by asserting the **wordline**.
- During a read operation, the cells on this wordline drive the **bitlines**.
- The **column** circuitry may contain amplifiers or buffers to sense the data.
- A typical memory array may have thousands or millions of words of only 8–64 bits each, which would lead to a tall, skinny layout that is hard to fit in the chip floorplan and slow because of the long vertical wires.
- Therefore, the array is often folded into fewer rows of more columns. After folding, each row of the memory contains 2^k words, so the array is physically organized as 2^{n-k} rows of 2^{m+k} columns or bits.



16 4-bit words:
(n = 4, m = 2, k=0)

Memory Array : Two-way Fold

- The array is physically organized as 2^{n-k} rows of 2^{m+k} columns or bits.
- The Figure shows a two-way fold (k = 1) with eight rows and eight columns.
- The column decoder controls a multiplexer in the column circuitry to select 2^m bits from the row as the data to access.
- Larger memories are generally built from multiple smaller sub-arrays so that the wordlines and bitlines remain reasonably short, fast, and low in power dissipation.



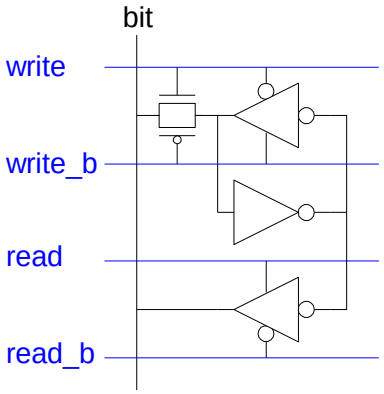
16 4-bit words:
(n = 4, m = 2, k=1)

12.2 SRAM

- Static RAMs use a memory cell with internal feedback that retains its value as long as power is applied. It has the following attractive properties:
  - Denser than flip-flops
  - Compatible with standard CMOS processes
  - Faster than DRAM
  - Easier to use than DRAM
- SRAMs are widely used i: caches, register files, tables and buffers.

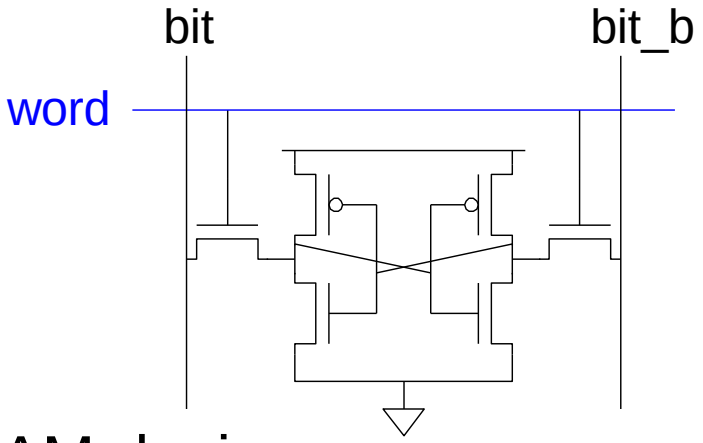
SRAM Cell: 12T SRAM Cell

- 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
  - Large area, so it is not used.
- Cell size accounts for most of array size
  - Reduce cell size at expense of complexity
  - The small cell size also offers shorter wires and hence lower dynamic power consumption.



SRAM Cell: 6T SRAM Cell

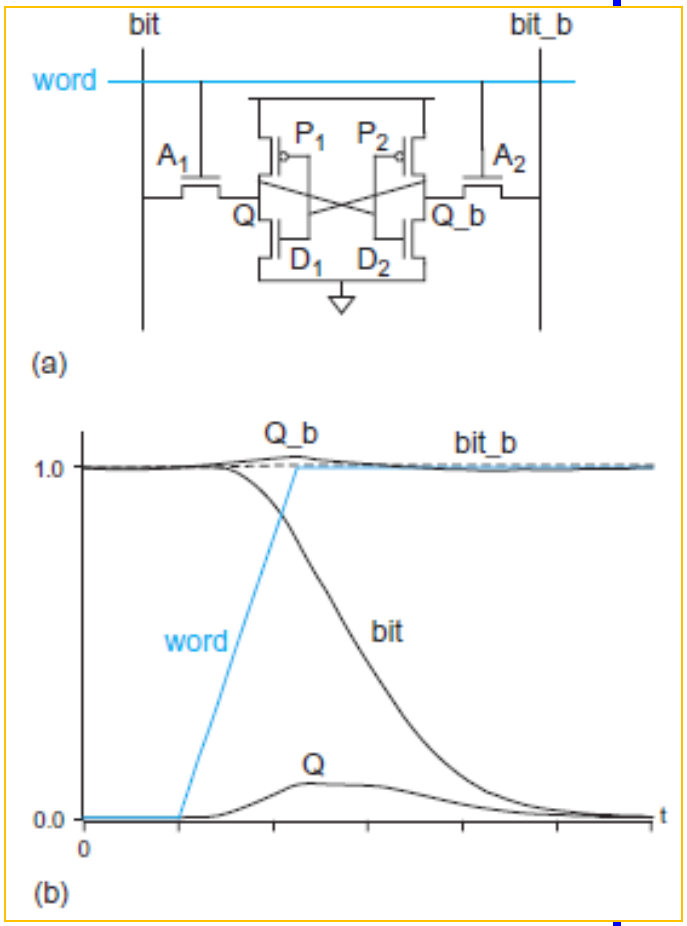
- 6T SRAM Cell
  - Used in most commercial chips
  - Data stored in cross-coupled inverters
- Read:
  - Precharge bit, bit_b
  - Raise wordline
- Write:
  - Drive data onto bit, bit_b
  - Raise wordline



- The central challenges in SRAM design are
  - minimizing its size
  - ensuring that the circuitry holding the state is weak enough to be overpowered during a write, yet strong enough not to be disturbed during a read.

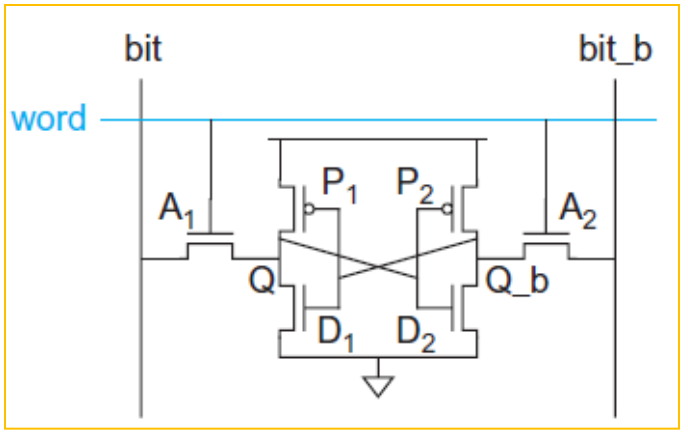
12.2.1.1 Read Operation

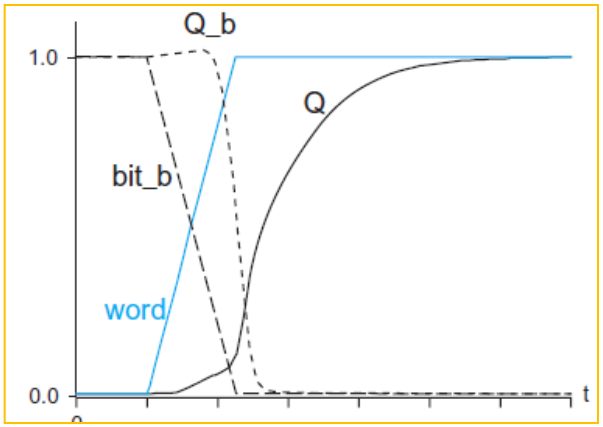
- Read operation:
  - Precharge both bitlines high
  - Then turn on wordline
  - One of the two bitlines will be pulled down by the cell
- Ex: Q= 0, Q_b = 1
  - bit discharges, bit_b stays high
  - But Q bumps up slightly
- *Read stability*
  - P2/D2 must not flip
  - D1 >> A1



12.2.1.2 Write Operation

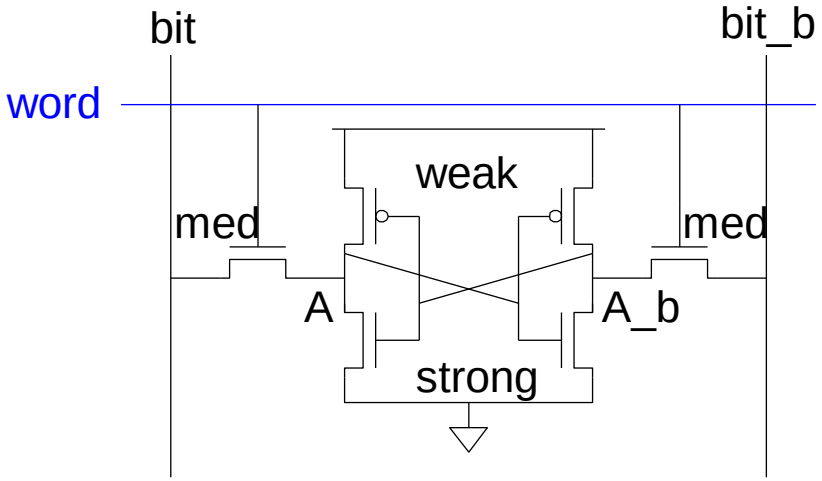
- Write operation:
  - Drive one bitline high, the other low
  - Then turn on wordline
  - Bitlines overpower cell with new value
- Ex: Q = 0, Q_b = 1, bit = 1, bit_b = 0
  - Force Q_b low, then Q rises high
- *Writability*
  - Must overpower feedback inverter
  - A2 >> P2





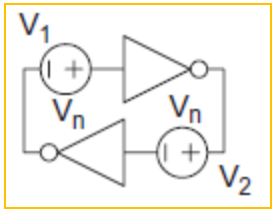
12.2.1.3 Cell Stability

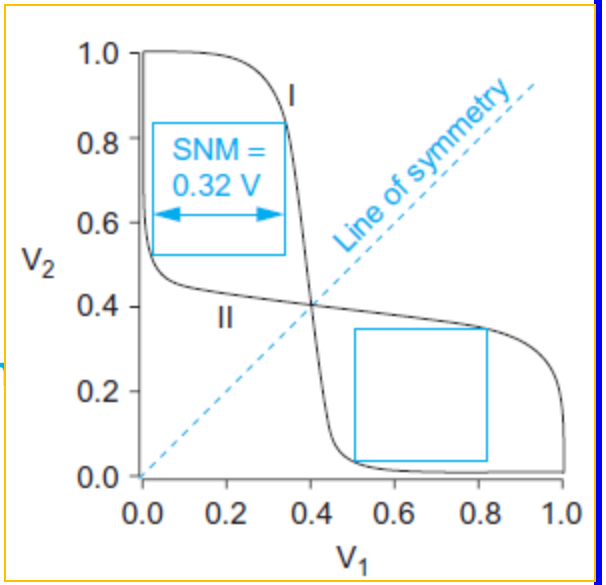
- **Read stability:**
  - The nMOS pulldown transistor in the cross-coupled inverters must be strongest.
- **Writibility:**
  - the access transistors are of intermediate strength, and the pMOS pullup transistors must be weak.



Static Noise Margin

- The *static noise margin (SNM)* measures how much noise can be applied to the inputs of the two cross-coupled inverters before a stable state is lost (during hold or read) or a second stable state is created (during write).
- The static noise margin can be determined graphically from a butterfly diagram shown in Figure.
- The butterfly plot shows two stable states (with one output low and the other high) and one metastable state (with V1 = V2). A positive value of noise shifts curve I left and curve II up. Excessive noise eliminates the stable state of V1 = 0 and V2 = VDD, forcing the cell into the opposite state.
- **The static noise margin is determined by the length of the side of the largest square that can be inscribed between the curves.**

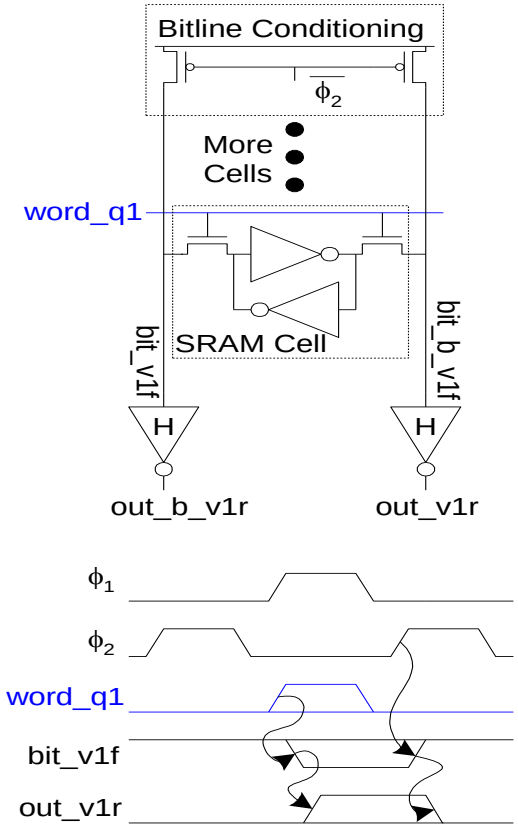




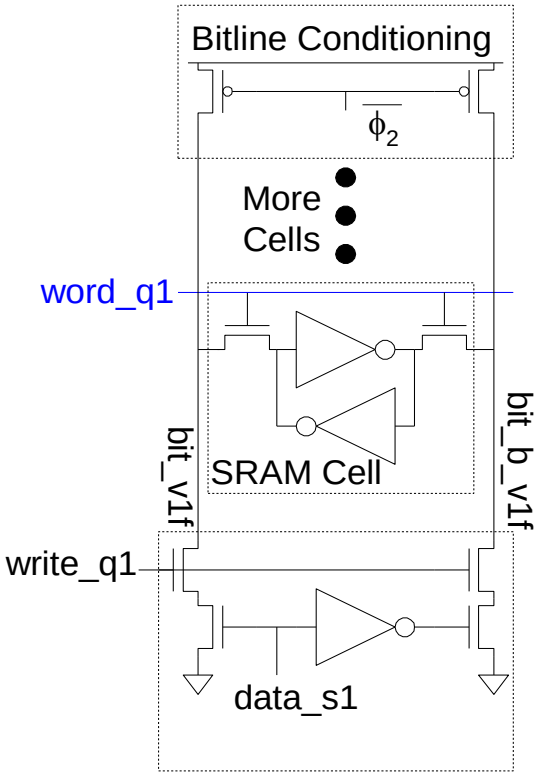
REST of the SLIDES Are reading material

SRAM Column Example

Read

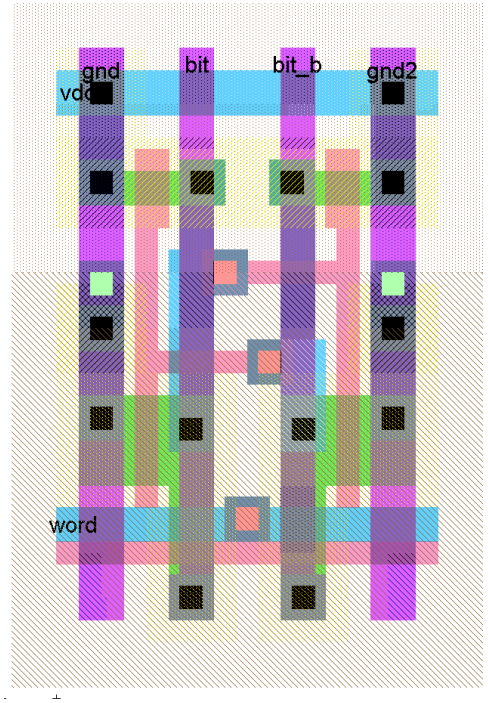


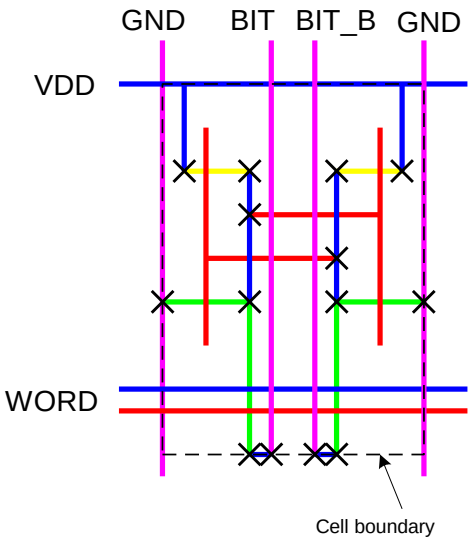
Write

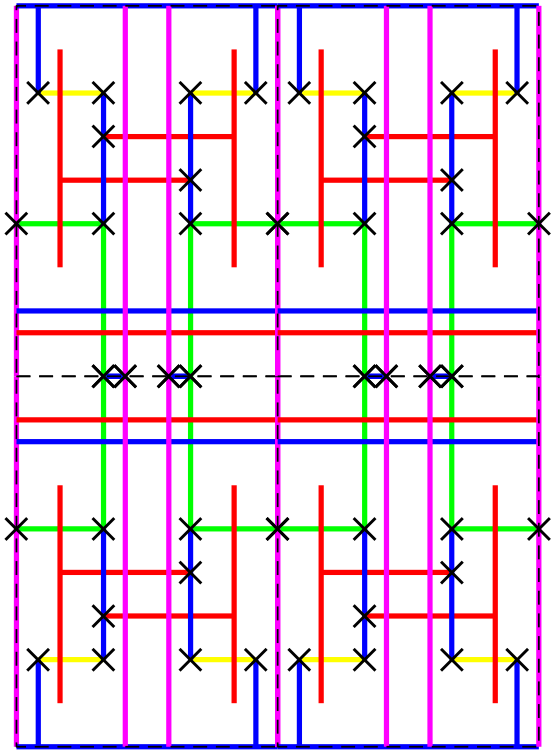


SRAM Layout: Physical Design

- Cell size is critical (left): 26 x 45 λ (smaller in industry)
- Tile cells sharing V_{DD}, GND, bitline contacts (6T:right).

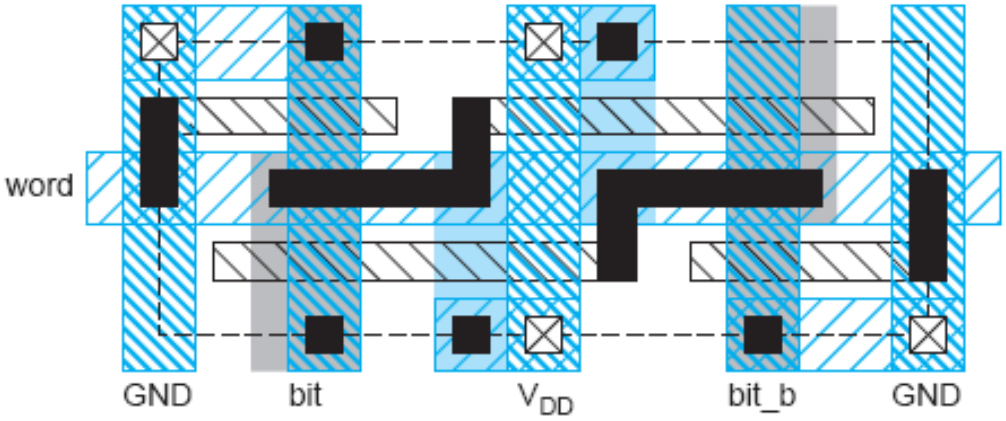






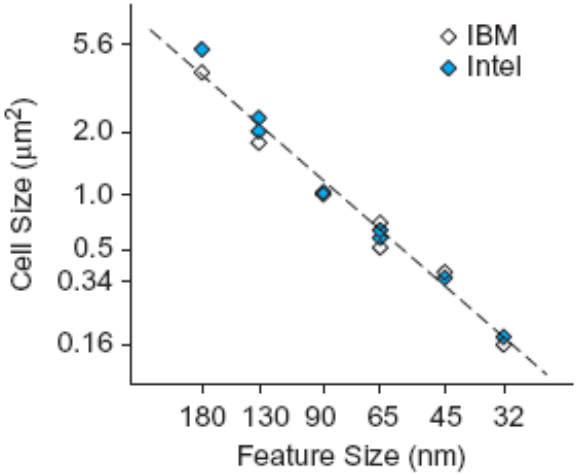
Thin Cell

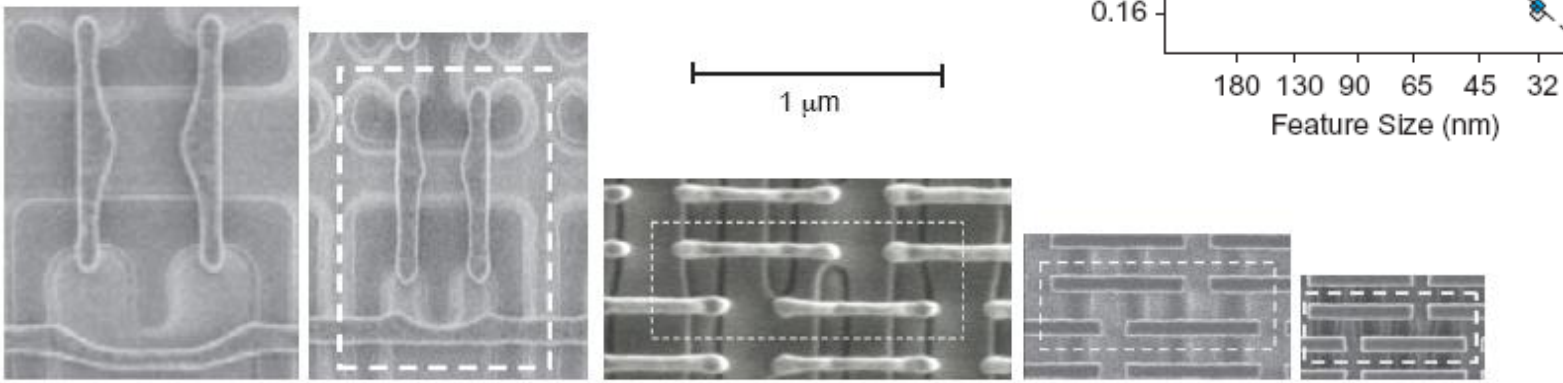
- In nanometer CMOS
  - Avoid bends in polysilicon and diffusion
  - Orient all transistors in one direction
- *Lithographically friendly* or *thin cell* layout fixes this
  - Also reduces length and capacitance of bitlines



Commercial SRAMs

- Five generations of Intel SRAM cell micrographs.
  - Transition to thin cell at 65 nm
  - Steady scaling of cell area



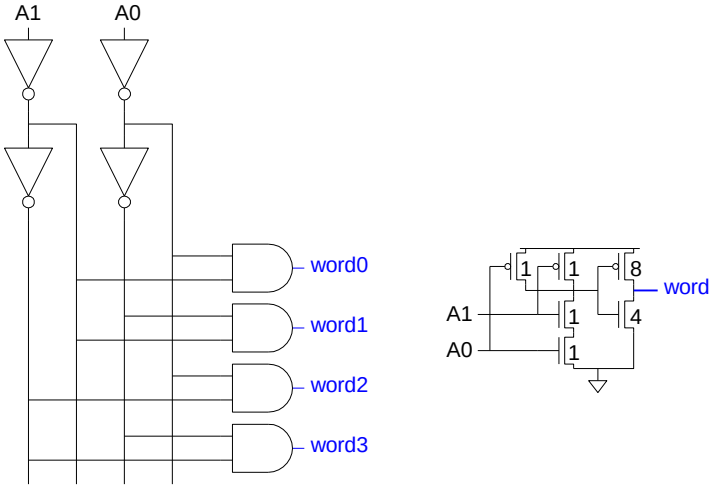


130 nm [Tyagi00] 90 nm [Thompson02] 65 nm [Bai04] 45 nm [Mistry07] 32 nm [Natarajan08]

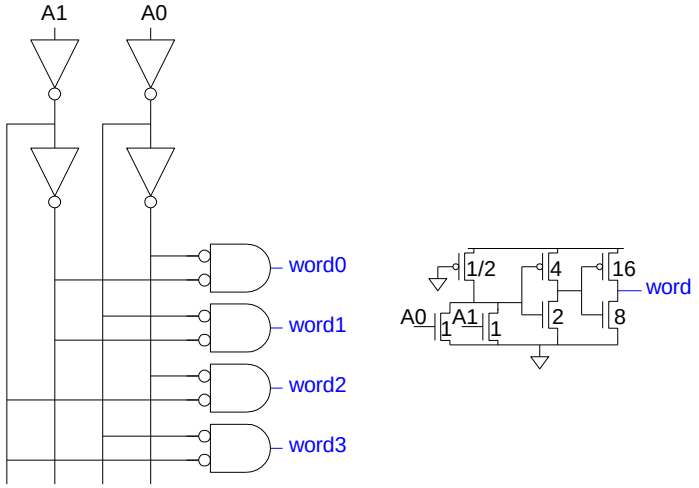
Row Circuitry: Decoders

- Row circuitry consists of decoder & worldline drivers
- n:2^n decoder consists of 2^n n-input AND gates
  - One needed for each row of memory
  - Build AND from NAND or NOR gates

Static CMOS

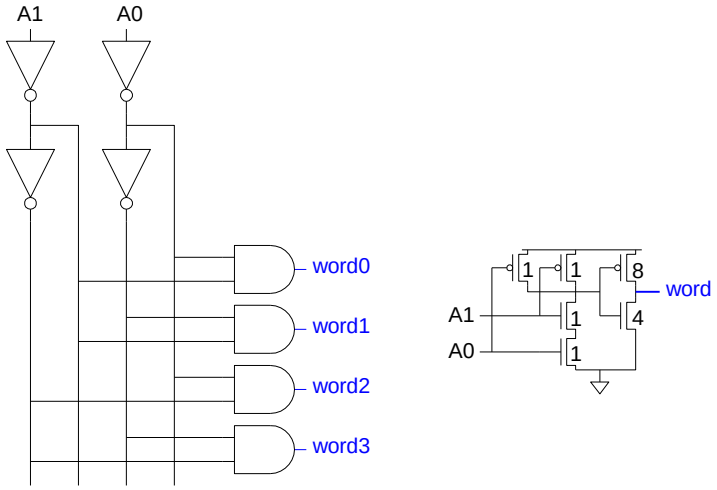


Pseudo-nMOS

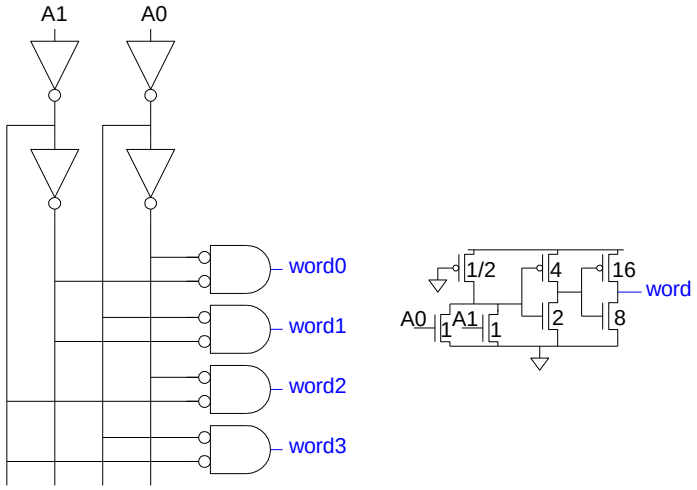


Row Circuitry: Decoders

Static CMOS



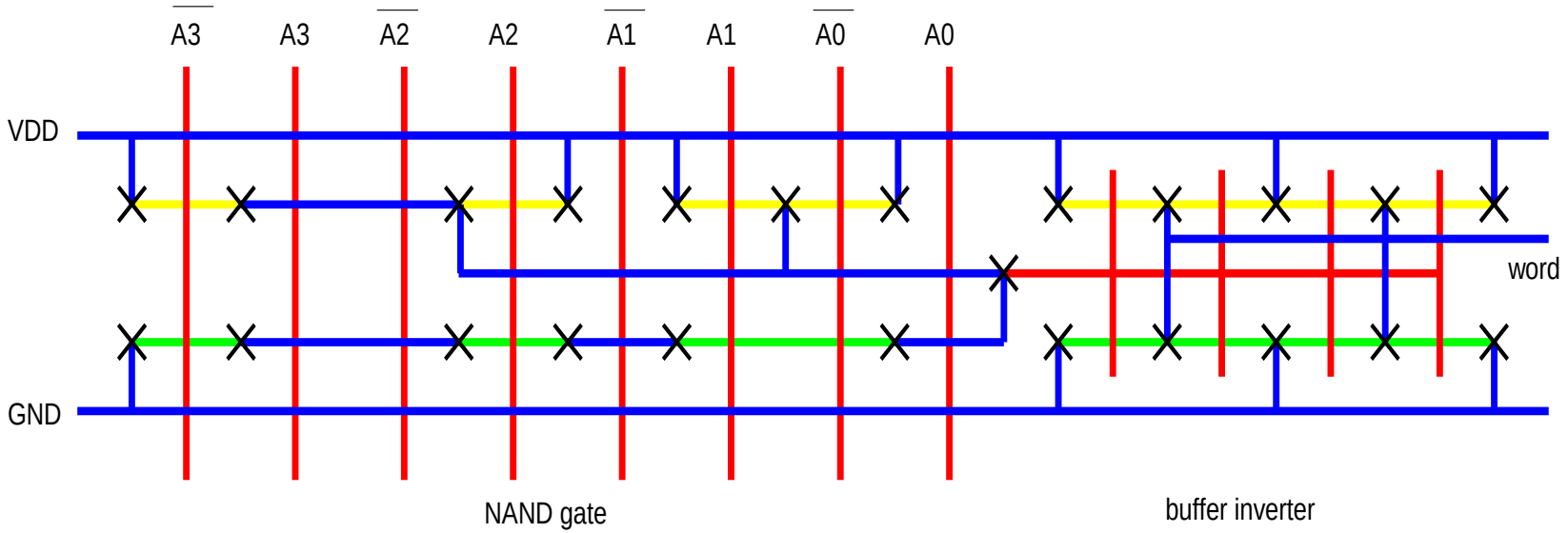
Pseudo-nMOS



- Wordline must be qualified with clock (shared clock and transistors).

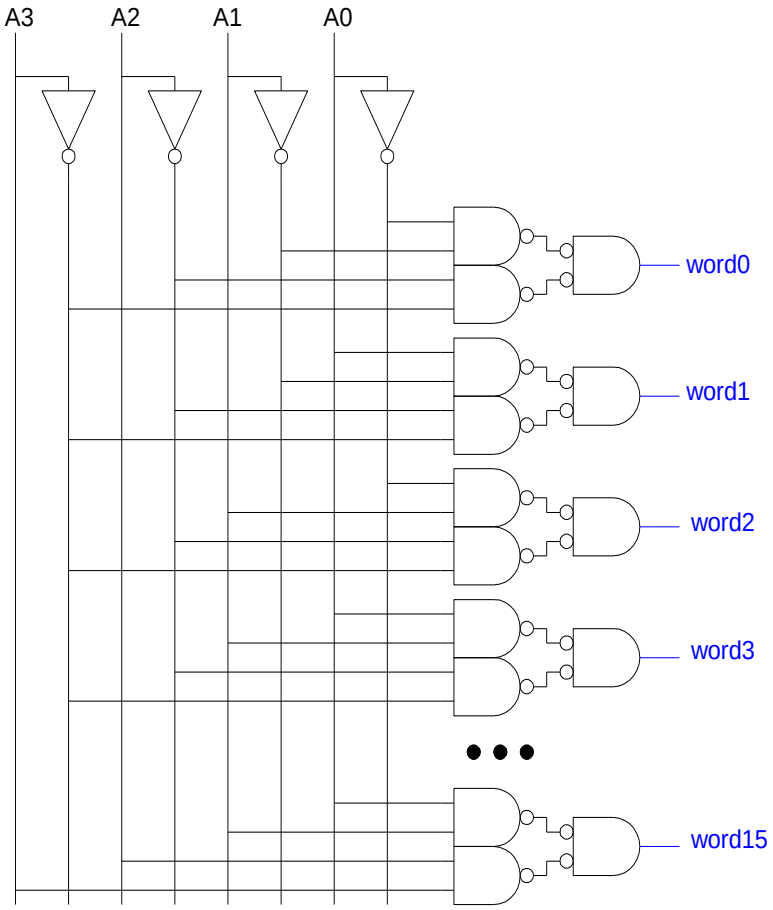
Decoder Layout

- Decoders must be pitch-matched to SRAM cell: height of decoder gate must match height of the row it drives.
  - Requires very skinny gates



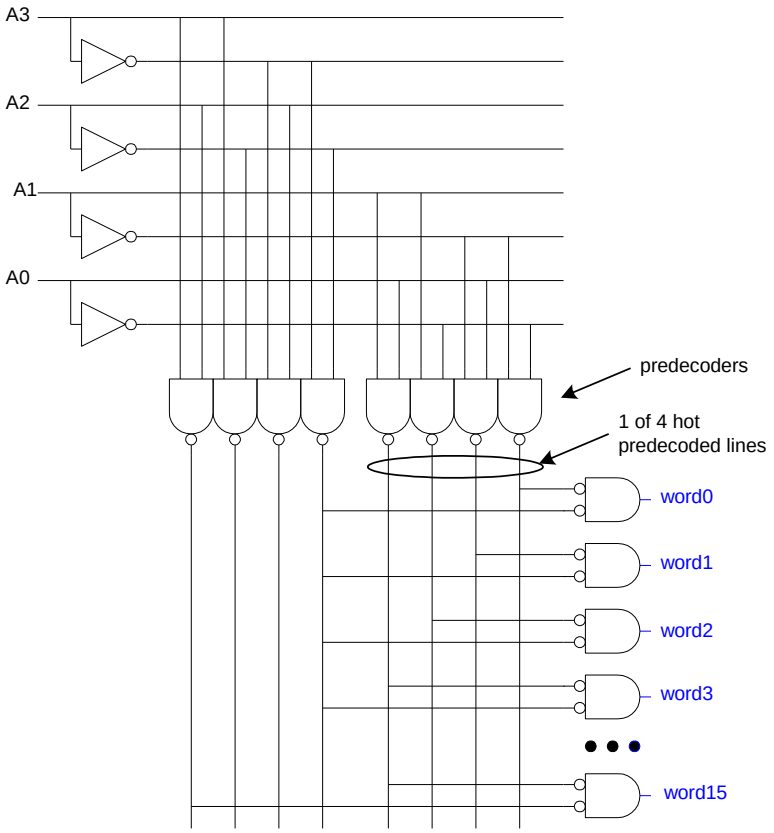
Large Decoders

- 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 into predecoder
  - Saves area
  - Same path effort

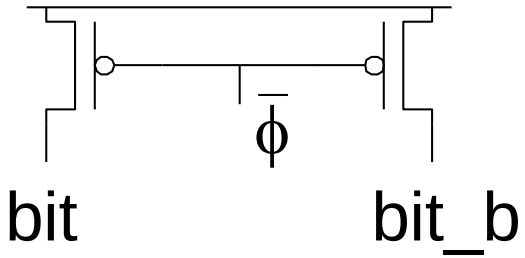


Column Circuitry

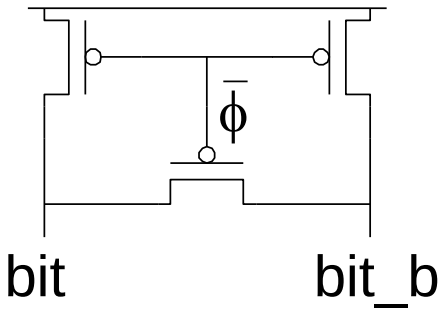
- Some circuitry is required for each column
  - Bitline conditioning
  - Write driver
  - Bitline sense amplifiers
  - Column multiplexing

Bitline Conditioning

- Precharge bitlines high before reads or writes



- Equalize bitlines to minimize voltage difference when using sense amplifiers



Bitline Sense Amplifiers

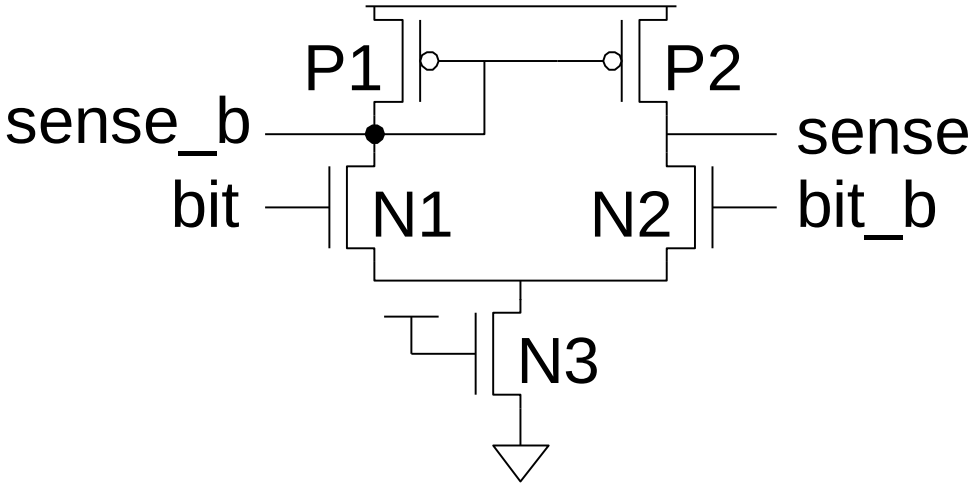
- Bitlines is classified: large-signal (single-ended sensing) or small-signal. (differential sensing).
- Large-signal: bitline swings between Vdd and GND.
- Small-signal: one of the two bitlines changes by a small amount to save delay and reduce energy consumption.

Bitline 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 (small I)
- *Sense amplifiers* are triggered on small voltage swing (reduce ΔV)

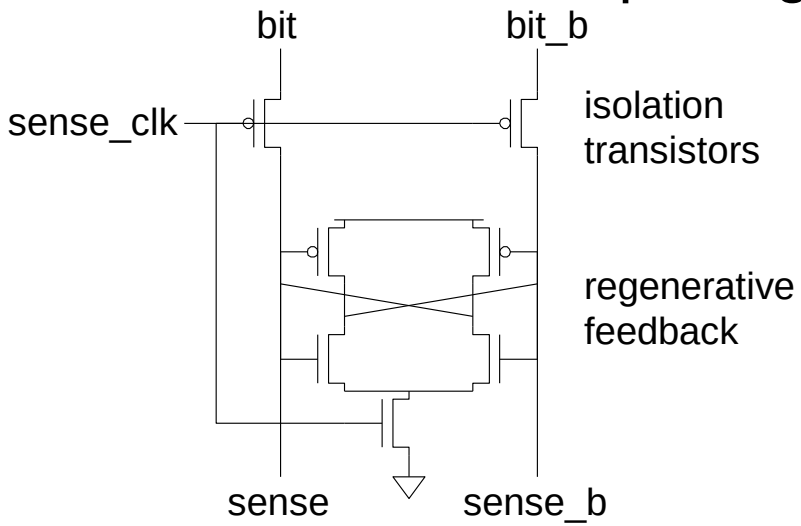
Differential Pair Amp

- Differential pair requires no clock
- But always dissipates static power
- Consumes a significant amount of DC power.



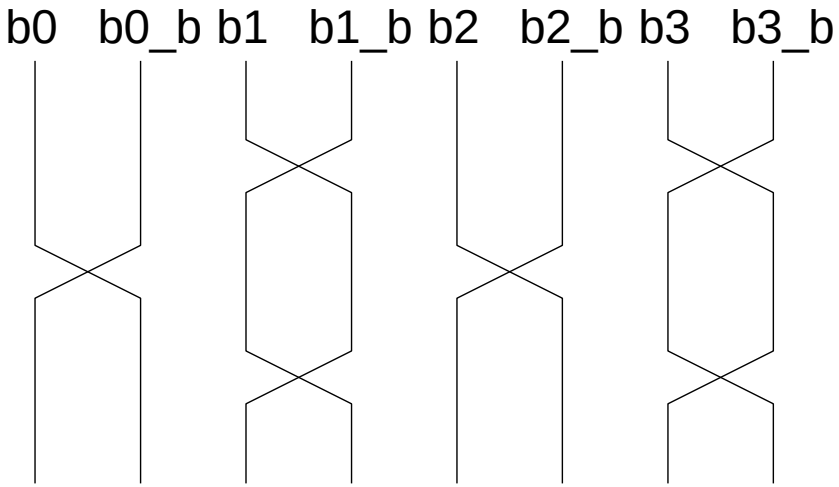
Clocked Sense Amp

- Clocked sense amp saves power (consumes power only while activated).
- Requires sense_clk after enough bitline swing
- Isolation transistors cut off large bitline capacitance, regenerative feedback to make one output high and the other low.



Twisted Bitlines

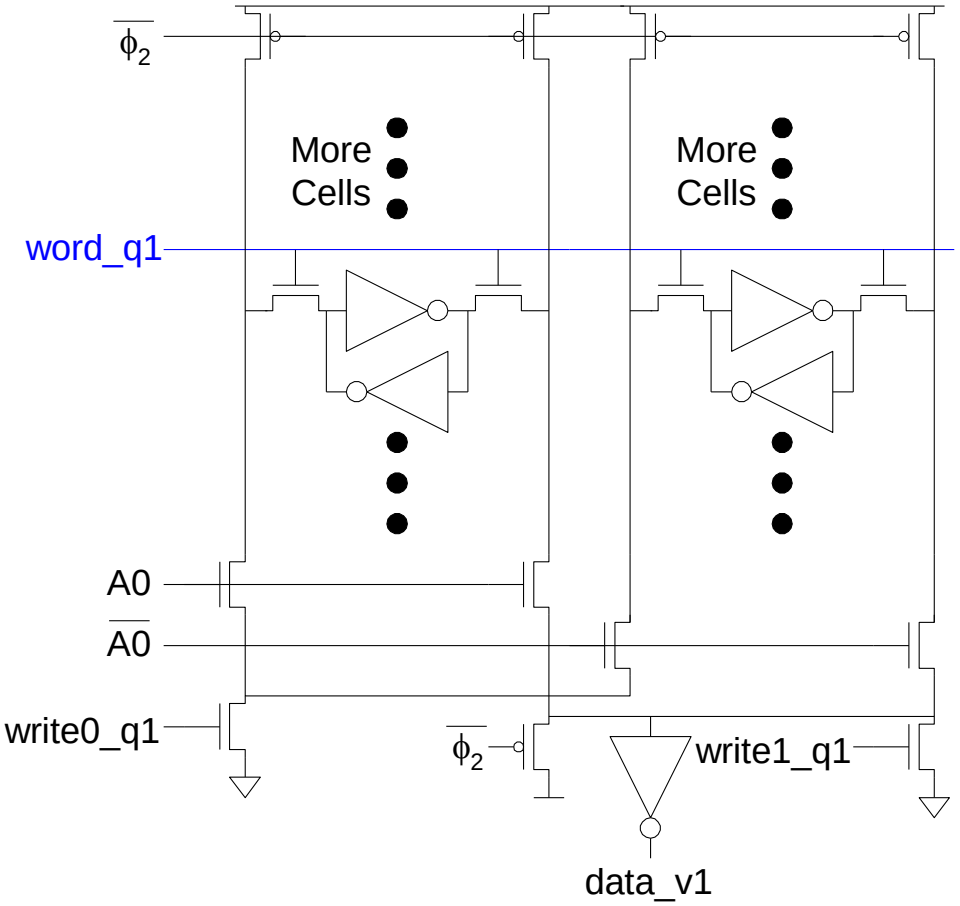
- Sense amplifiers also amplify noise
  - Coupling noise is severe in modern processes
  - Try to couple equally onto bit and bit_b
  - Done by *twisting or transposing* bitlines



Column Multiplexing

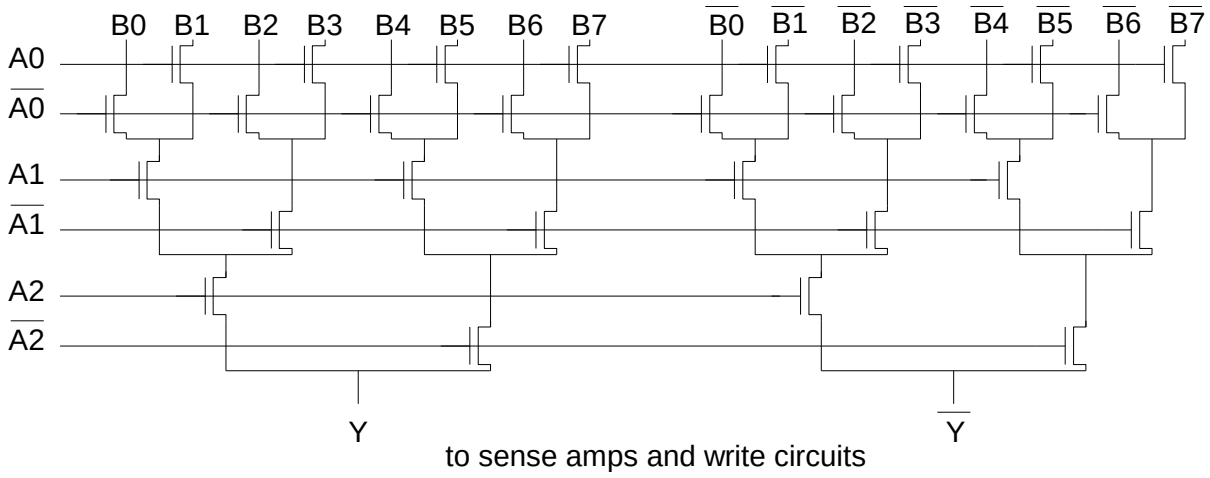
- Recall that array may be folded for good aspect ratio
- Ex: 2 kword x 16 folded into 256 rows x 128 columns
  - Must select 16 output bits from the 128 columns
  - Requires 16 8:1 column multiplexers

Ex: 2-way Muxed SRAM



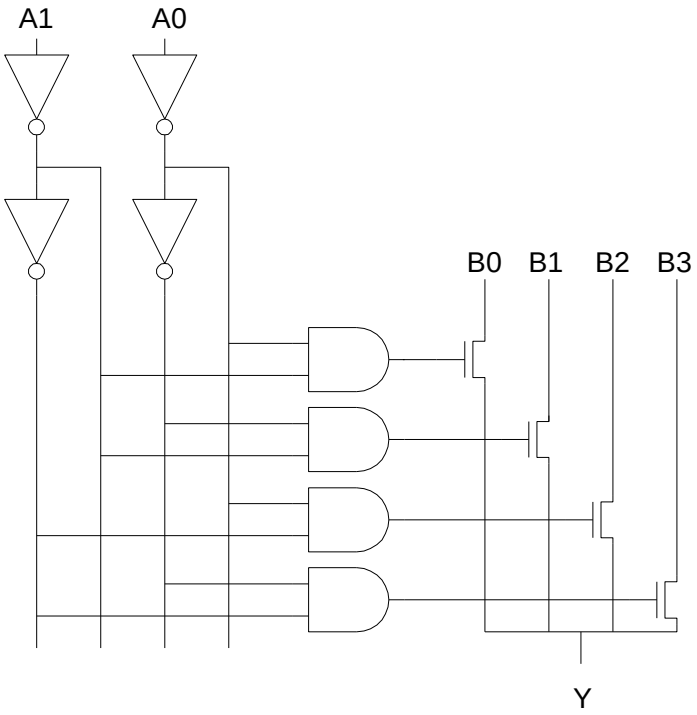
Tree Decoder Mux

- 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



Single Pass-Gate Mux

- Or eliminate series transistors with separate decoder

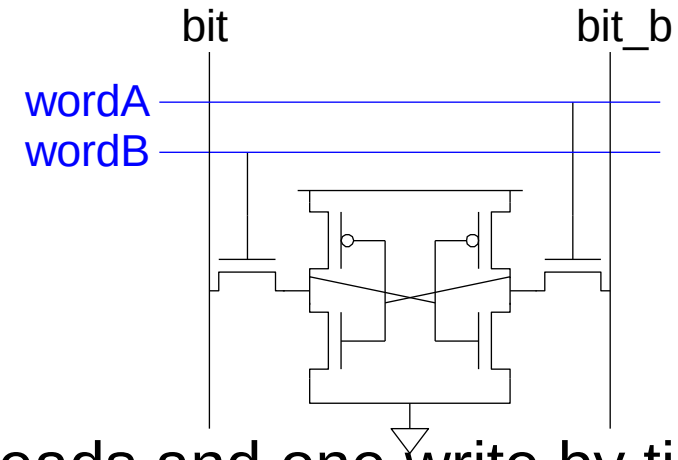


Multiple Ports

- We have considered single-ported SRAM
  - One read or one write on each cycle
- *Multiported* SRAM are needed for register files
- Examples:
  - Multicycle MIPS must read two sources or write a result on some cycles
  - Pipelined MIPS must read two sources and write a third result each cycle
  - Superscalar MIPS must read and write many sources and results each cycle

Dual-Ported SRAM

- Simple dual-ported SRAM
  - Two independent single-ended reads
  - Or one differential write



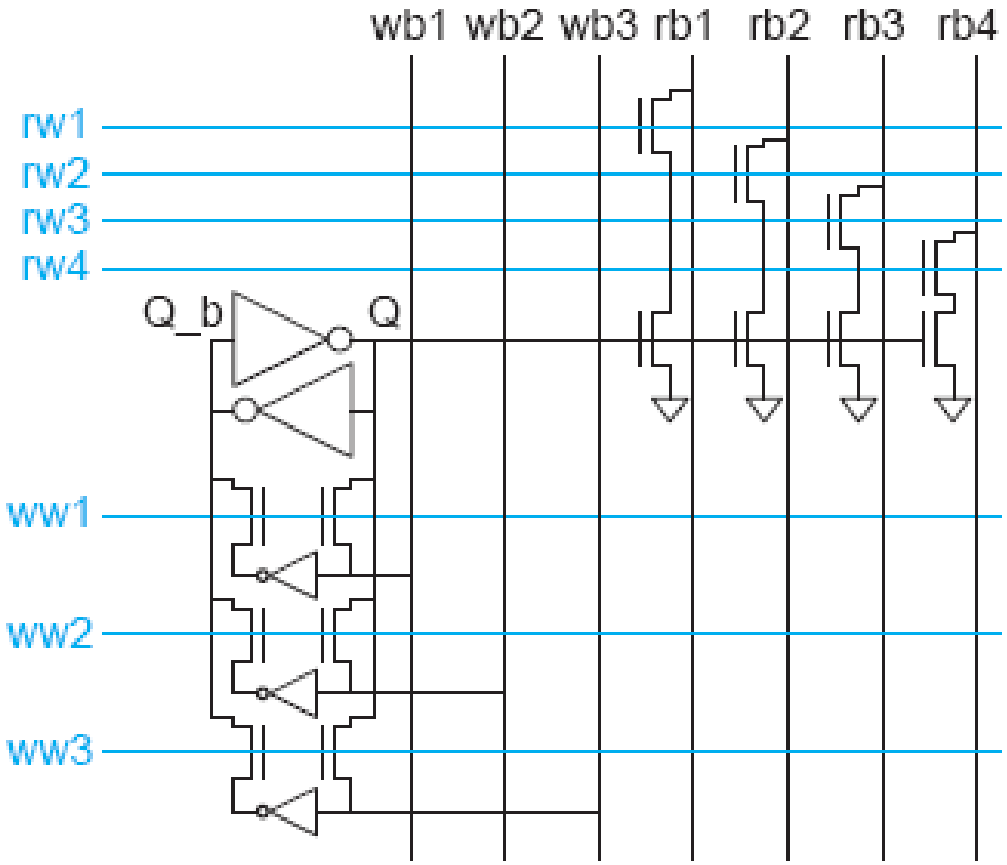
- Do two reads and one write by time multiplexing
  - Read during ph1, write during ph2

Multi-Ported SRAM

- Adding more access transistors hurts read stability
- Multiported SRAM isolates reads from state node
- Differential read ports double the number of read bitlines and transistors.
- Single-ended bitlines save area
- Multiple write ports simply attach the ports to the state node.

Multi-Ported SRAM

- 3 write ports and 4 read ports:



Large SRAMs

- Large SRAMs are split into banks or subarrays for speed and reduce power
- Ex: UltraSparc 512KB cache
  - 4 128 KB subarrays
  - Each have 16 8KB banks
  - 256 rows x 256 cols / bank
  - 60% subarray area efficiency
  - Also space for tags & control