EECS 427
Lecture 18/19: Memory reliability and power
Reading: 12.4, 12.5

Last Time

- Memory core
  - 6T SRAM and 1T DRAM cells are mainstays of the IC industry
  - Lots of details on sizing – but the KEY thing is:
    - Read and write place conflicting demands on cell sizing – sizing of access transistor vs. pull-down device is **crucial**
- Memory peripherals (SRAM-oriented)
  - Decoders: For large memories, they take up a large fraction of total access time (speed-critical)
  - Sense amplifiers: required to speed up read times, reduce voltage swing on bit lines

Lecture Overview

- Memories should be inherently robust/reliable by definition
- Memories comprise a growing portion of digital systems and hence their power consumption is appreciable

Reliability and Yield

- Semiconductor memories trade-off noise margin for density and performance

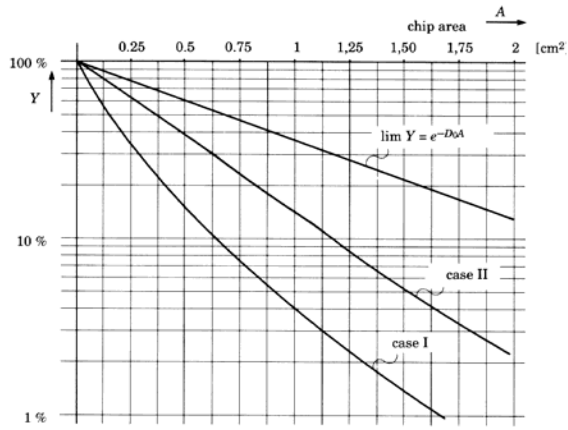
  Thus, they are highly sensitive to noise (cross talk, supply noise)
- High density and large die sizes cause yield problems (DRAM chips are the largest around)

- Yield = 100*(# of good chips/wafer)/(# of chips/wafer)

$$Y = [(1 - e^{-AD})/AD]^2$$

- A = Area, D = Defect Density
- Increase yield using error detection/correction and redundancy

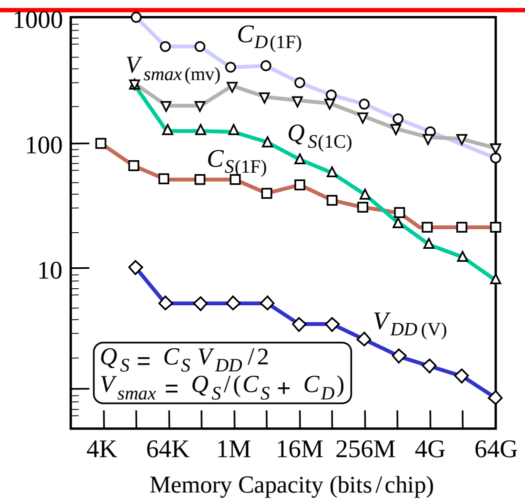
Yield



Yield curves at different stages of process maturity (from [Veendrick92])

5

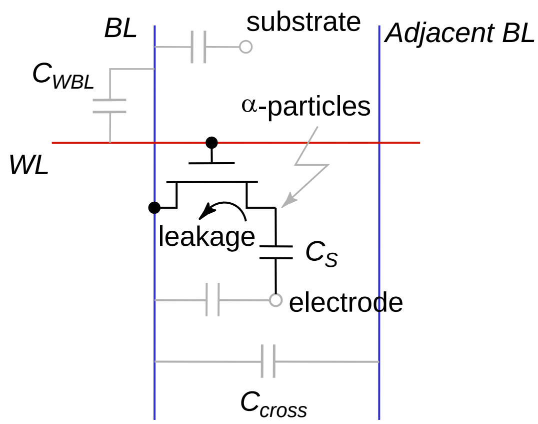
Sensing Parameters in DRAM



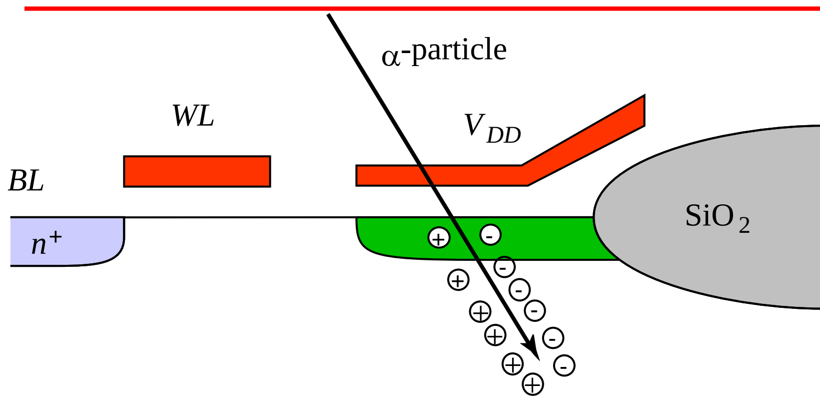
From [Itoh01]

6

Noise Sources in 1T DRAM



Alpha-particles (or Neutrons)



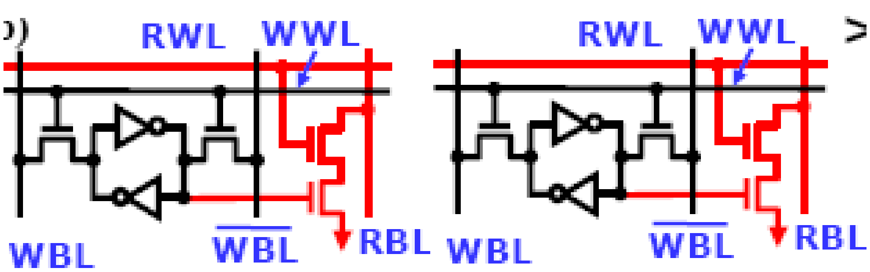
1 Particle ~ 1 Million Carriers

Implication: Interleave words → sometimes complicates matters

SRAM cell failure mode due to process variation

- Read failures are most common
  - '0' node to be read will spike up when access transistor turns ON; if spikes too high, could toggle the held state
    - "Destructive read" or read upset as discussed last time
    - Depends on relative strengths of various transistors in the 6T cell
      - **Random sources** of variation are the worst culprits in this regard (likely to have strong P, weak N, etc)
    - 4MB array: 5.44-sigma confidence to achieve only 1 failure

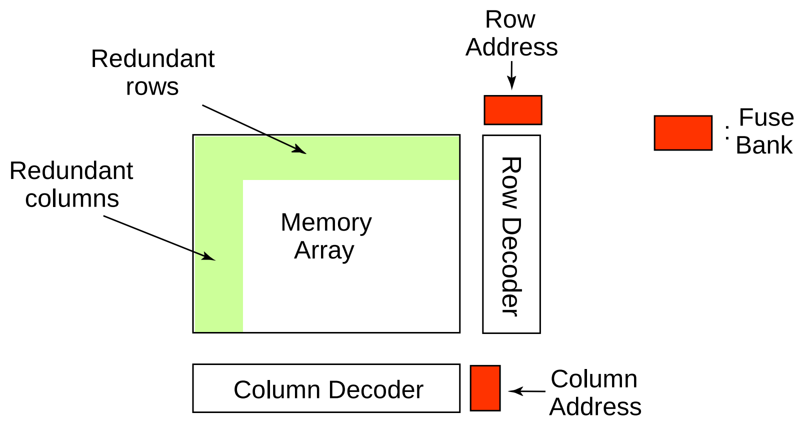
8T cell



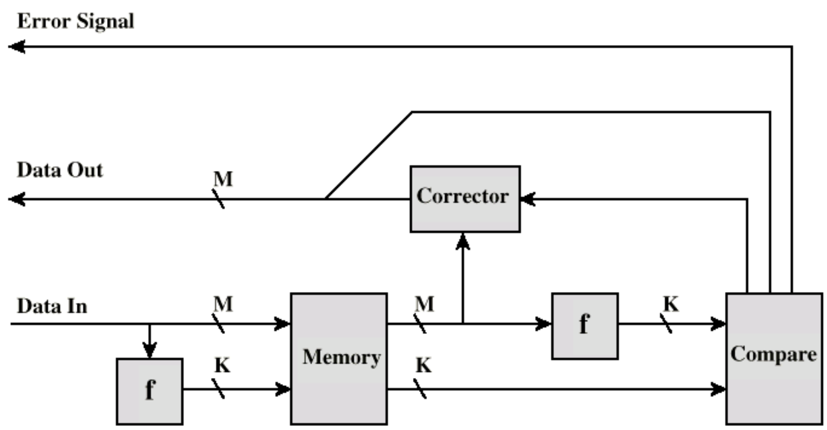
Pseudo-read: Impacts cells on WWL that are not being written to (column muxing)

Gaining significant traction in industry despite area

Redundancy



Error Correcting Code Function



Hamming based codes

- M data bits
- K parity bits
- Must transmit M+K bits instead of just M (code overhead)
- Ex: 8 bit data (M=8) → 11000100

Bit positions:

1 2 3 4 5 6 7 8 9 10 11 12

P_1 P_2 1 P_4 1 0 0 P_8 0 1 0 0

Compute parity bits as XORs of various bit combinations

P_1 = XOR of bits 3,5,7,9,11 = 0

P_2 = XOR of bits 3,6,7,10,11 = 0

P_4 = XOR of 5,6,7,12 = 1

P_8 = XOR of 9,10,11,12 = 1

Checking of data

- 4 check bits (XOR parity bits with their original data):

C_1 = XOR of 1,3,5,7,9,11

C_2 = XOR of 2,3,6,7,10,11

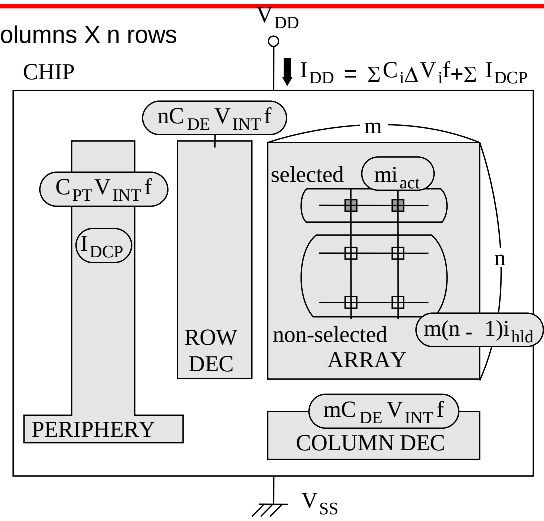
C_4 = XOR of 4,5,6,7,12

C_8 = XOR of 8,9,10,11,12

- If $C_1=C_2=C_4=C_8=0$, no errors
- Otherwise, bit C is in error ($C_8C_4C_2C_1$ encoded)
- Ex: if bit 5 is flipped erroneously in the memory array, C_4 and C_1 would be computed as high:
  - $C_8C_4C_2C_1$ = 0101 → bit location 5
  - Corrector circuit would then flip the state of bit 5 back to correct value

Sources of Power Dissipation in Memories

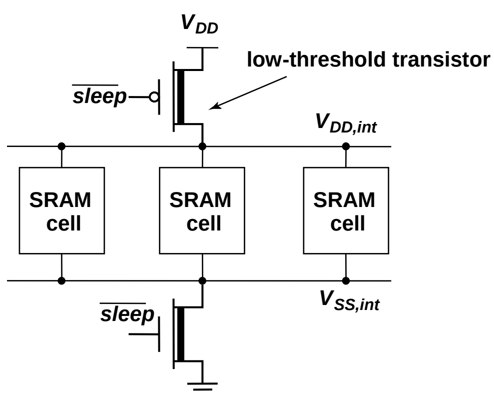
m columns X n rows

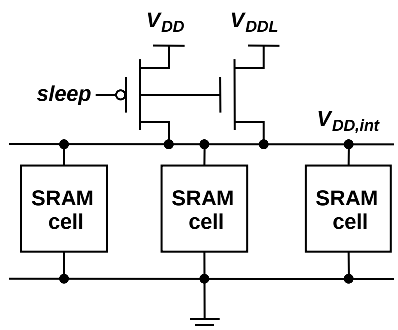


- Leakage is a *major* issue in memories
- Lots of transistor width, most of it isn't switching in a given access
- Memory partitioning very important – can turn off unused blocks after decoding address

From [Itoh00]

Suppressing Leakage in SRAM



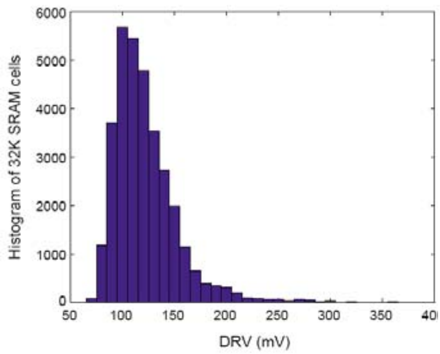


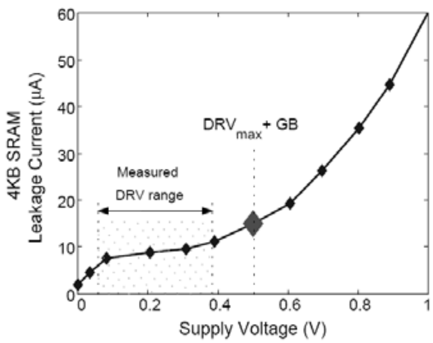
Inserting Extra Resistance

Reducing the supply voltage AKA "drowsy cache"

Data retention in SRAM

- To retain data, some minimum Vdd must be maintained





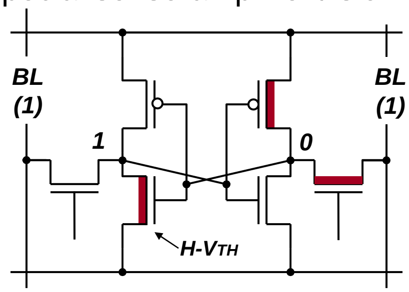
H. Qin et al.

Additional Techniques

- Turn off unused blocks in an array
- Apply body bias to sections of the array not being used to increase Vth and reduce leakage
- Just use a higher Vth in the memory cells (speed penalty amortized over entire access time)
  - Ex: 100mV higher Vth in 6T cell leads to speed penalty of ~20%
  - If 6T access time is only 1/3 of total delay, the total memory access delay penalty is 7% and leakage will be reduced by nearly 10X since array leakage dominates

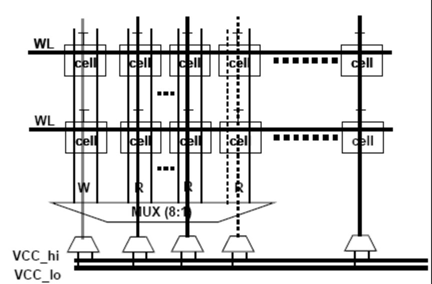
Dual V_{TH}-Based Techniques

- Asymmetric Dual-V_{TH} Cache
  - Optimizing leakage power of SRAM cell for storing "0" using high-V_{TH} devices in SRAM cells
    - exploiting highly biased memory bits to "0" in many apps
    - requiring special sense-amplifier / slower access time



Column-Based Dynamic Power Supply (Intel)

- Dynamic Vdd concept
  - Columns are tied to 2 different possible Vdd's through a MUX
  - During a read operation, higher Vdd is used to enhance read stability
  - During write the active column is run at low Vdd to allow for easier overwriting of data
  - When a given bank is not being accessed, it is tied to low Vdd to reduce leakage



Conclusions

- Redundancy and ECCs used to improve memory reliability
  - Biggest problem today is in variability and corresponding read upset in small # of cells
- Leakage power is even more significant in memories than in logic
  - Due to low switching activity in most of the array, huge amount of transistor width