Lecture 13: Memory technology

Monday, February 19, 2018 1:09 PM

Outline

- Memory technologies
- Memory hierarchy
  - Temporal and spatial locality
  - Blocks/lines
- Packaging technology

Perf. potential of modern CPUs

peak IPC ~ 4-6 inst.

freq ~ 4 GHz

per core: 16 billion inst/sec

↳ 4-8 per chip → 128 billion inst/sec

↳ floating point

call this gigaflops

floating point inst. per sec.

2 or 3 floating point #s

balance 1 flop

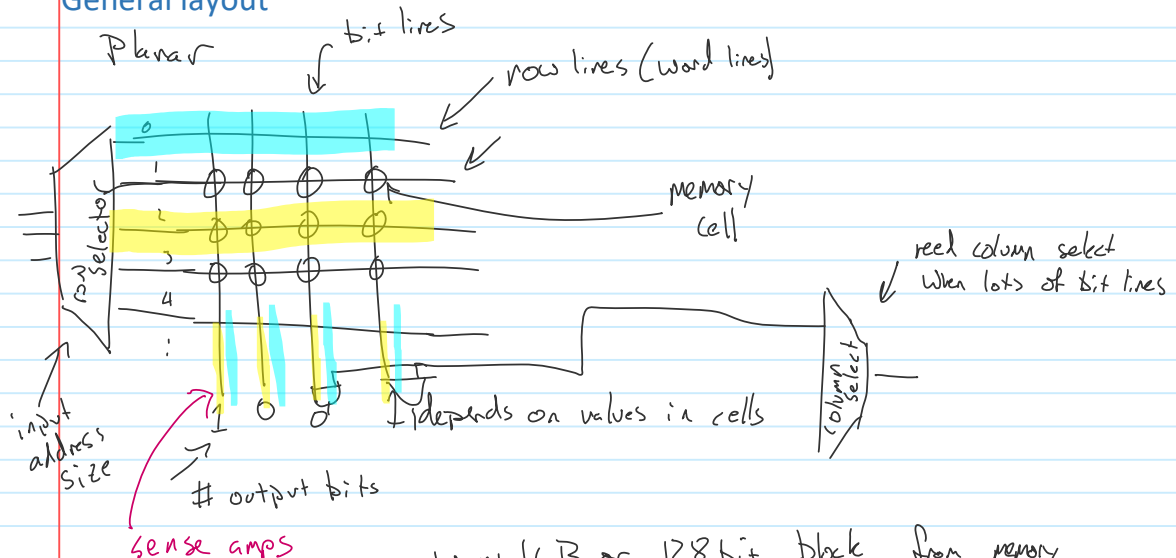
4 bytes ↳ 400 GB/s

Memory technologies

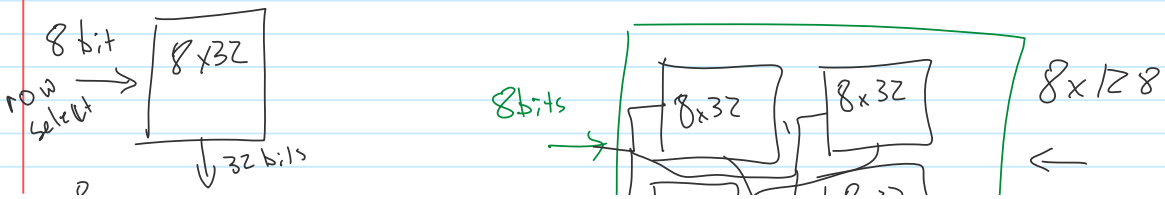
Key tradeoffs in memory technologies:

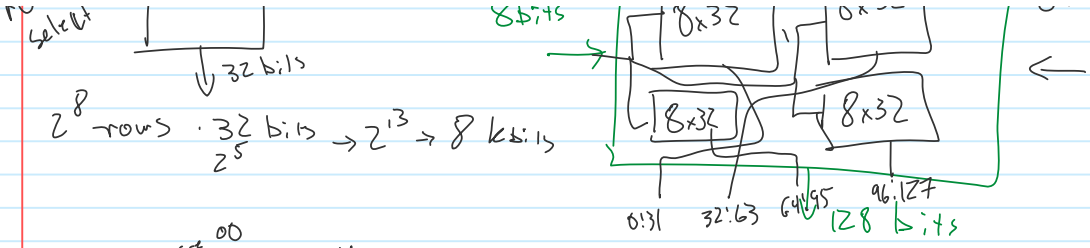
- size (density)
- speed / latency
- cost
- energy → power
  - read / write
  - ↳ static
  - ↳ dynamic
- compatibility → how to integrate
- reliability (ECC)
- volatility (keeps data w/o power)

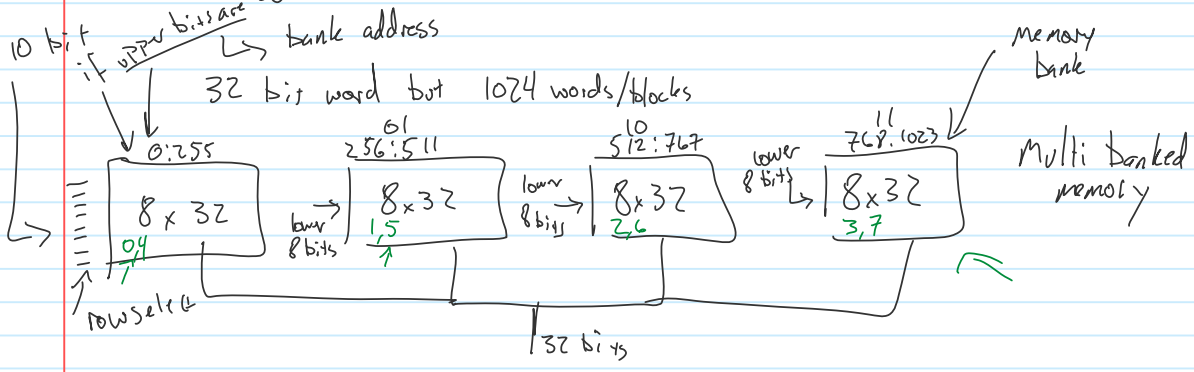
General layout



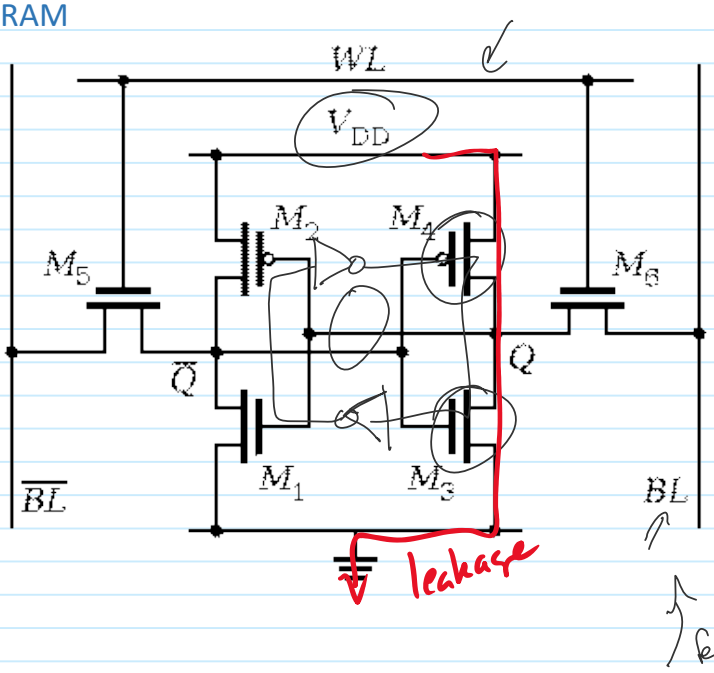
want 16B or 128 bit block from memory







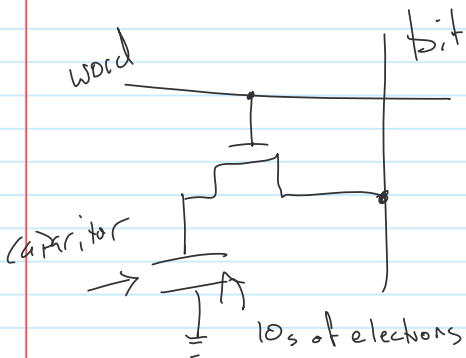
SRAM



Static RAM
↳ transistors in CMOS registers
← flip flop

latency: runs @ clock speed
0.5-10 ns
Volatile? yes
small ~ 6T
cost expensive
power? constant usage (static)
low dynamic power

DRAM



sensing charge in capacitor

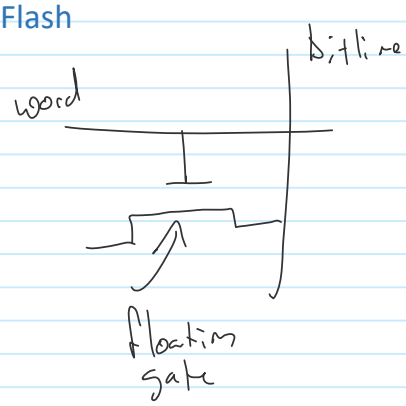
capacitor leaks ∴ you must refresh ← quite volatile
~ 64 ms
size → smaller than SRAM
cost → much lower than SRAM
latency → ~ 10+ ns w/ DIMM 10-50 ns
↳ Dual inline memory module

power → low, low leakage

Flash

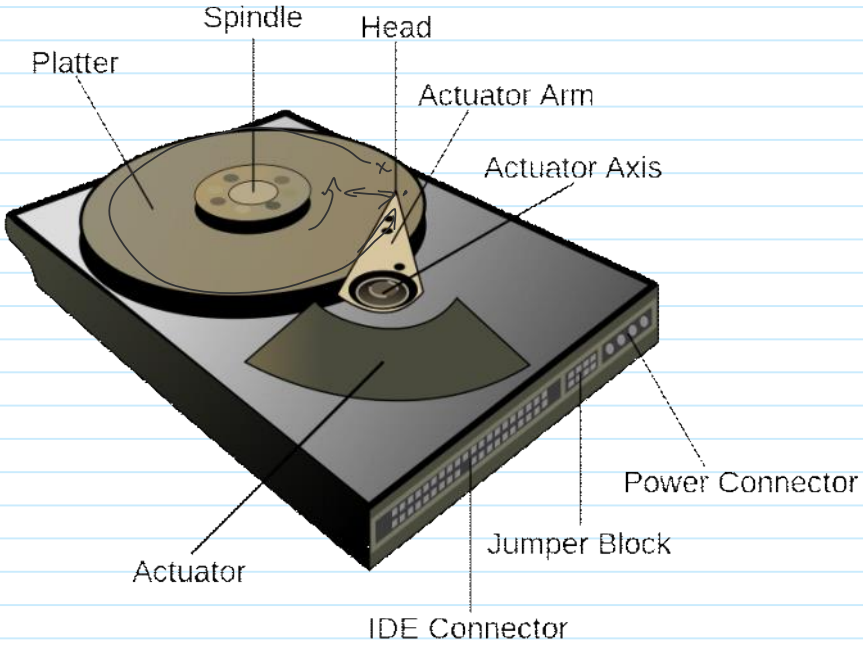
leakage

Flash



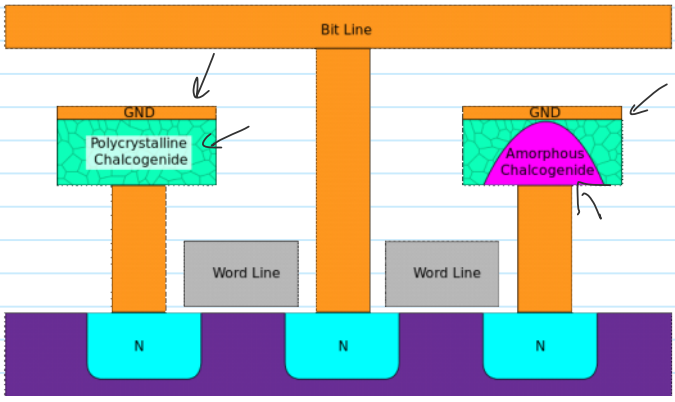
Smaller than DRAM → (no capacitor)
↳ store many bits per cell
(MLC or TLC)
↳ cheaper than DRAM
NON volatile
latency? ~ 10-100 ns but must operate on large pages 8-64 KB
↳ 10 µs +

Magnetic disk



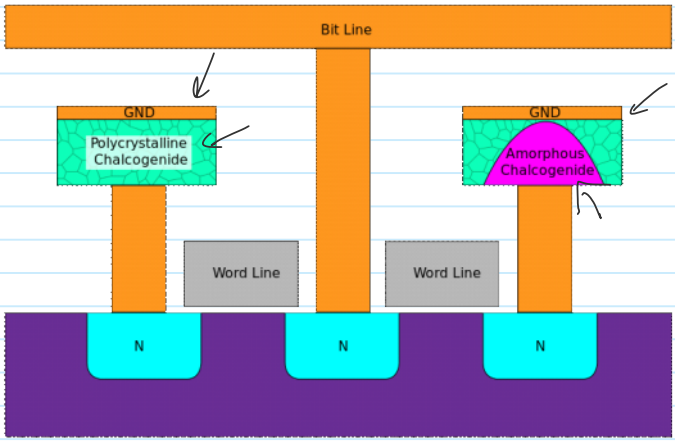
Size → much smaller than flash
non volatile
latency → much higher
10+ ms
power → high

Emerging technologies



↙ PCM phase change memory
↳ Intel's 3D-xpoint memory

non volatile
theoretically very dense
latency → slow to write ~ 5-10 µs
reads are fast → 1-10 ns
power → reading is low
writing is high



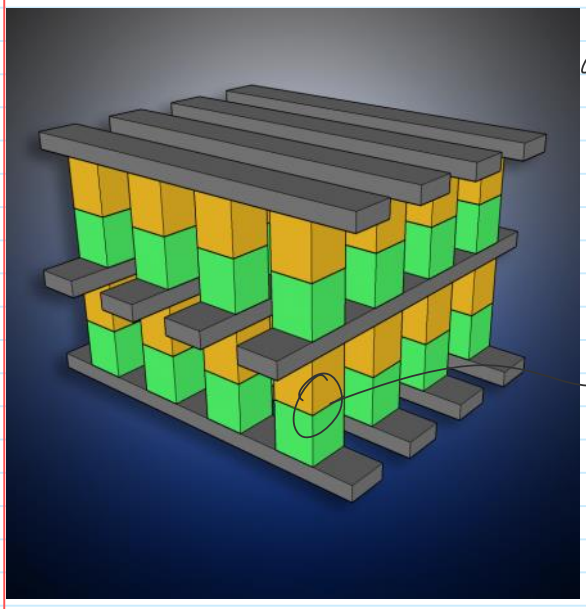
↳ Intel's 3D-Xpoint Memory

Non volatile

theoretically very dense

latency → slow to write ~ 5-10 μs
reads are fast → 1-10 ns

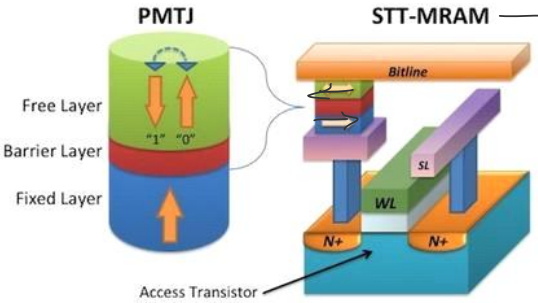
power → reading is low
writing is high



Intel's 3D Xpoint

can be really dense by stacking in 3D

PCM



non volatile

→ Spin-transfer torque memristive RAM

low power and low latency