COMP1203 - RAM and Cache

Dominik Tarnowski (tdom.dev)

18/11/19

RAM and Cache

- **DRAM** - main RAM (sticks)
- **SRAM** - Static Random Access Memory - inside of CPUS and chips

Memory Hierarchy

1. Registers - in the CPU
2. Internal or main memory
3. Other RAM elsewhere in the computer - disk cache, GPU cache, controllers, etc...

Static RAM

- Bits stored in one-bit latch
- More complex structure than DRAM - larger and more expensive
- Fast ($\approx 1 - 10$ ns)
- Chips can have 64M bit

SRAM only systems

- Possible, especially on a small microcontroller
- Fast but expensive per byte
- RPI uses DRAM as main memory and SRAM as cache

Measuring Memory Performance

- Access time - delay between requesting the address and getting the data
- Memory Cycle time
  - Sometimes time is required for memory to "recover" since last access

- Transfer rate
  - Rate at which data is moved

Dynamic RAM

- Bits stored in capacitors
- Chargers leak so need *refreshing* periodically even when powered
- Simpler construction
- Cheaper & Smaller / bit
- Slower (6-60ns)

DRAM Refresh

- Each bit discharges over time and is boosted back by the refresh
- Chip on the RAM circuit
- Slightly slows down performance, as memory cannot be accessed during refreshes

Types of ROM & Flash

- Can be written during manufacturing
- Useful for BIOS and embedded software
- PROM - programmable ROM

Error Correction - ECC RAM

- DRAM can sometimes loose data
- Hard Failure - permanent defect
- Soft Error - no permanent damage to memory
- Detected and fixed using error correcting algorithms and usage of **extra bits**.

Caches

DRAM is too slow (6-60ns) to access data in its memory. This is why we use cache and cache sequential memory locations each time memory fetching is performed.

- Small SRAM on the CPU chip
- Acts as a middle man between main memory and CPU

Cache as a middle man

- CPU requests contents of memory location from cache
- If memory location in cache, value is returned immediately (fast)
- Otherwise, request the location's whole block from main RAM (slow)
- Copy data to cache and return the location

Latency

DRAM with a CL of 5 takes **at least** 5 clock cycles to return the data, whereas static RAM only takes 1 cycle.

- Value of **CL** represents the latency of memory.
  - $CL = 15$ means 15 clocks per transaction.

Multi-level caches

- we usually use more than one cache level, such as lower latency smaller L1 and bigger, slower L2.
- L3 cache is often used to "shield" ram from the CPU, acting like a middle-man
- L1 and L2 caches are core specific, whereas L3 is shared between all