Lecture 17: Cache performance and virtual memory

Tuesday, March 6, 2018 10:15 AM

Outline

- Cache performance
  - AMAT
- Multi-level caches
- Real examples
- Virtual memory
  - Why?
  - Multi-level page tables
- Improving VM performance

System performance with caching

of misses

Average memory access time

Miss rate/ratio = # misses / # of accesses
↳ lds + stores

Misses per instruction
↳ per 1000 instructions

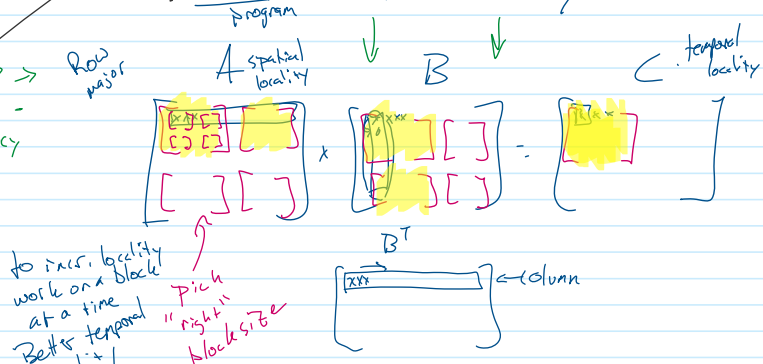
50 Mpki → .05 misses/inst.

Time for application to run / CPI

Time per app = (CPU exec. cycles + memory stall cycles) · cycle time
↳ assumes blocking cache

↳ # memory accesses / program · miss rate · miss penalty

how to reduce miss rate? →
↳ incr. cache capacity
↳ use diff replacement policy
↳ incr associativity



to incr. locality work on a block at a time
Better temporal locality
pick "right" blocksize

Average memory access time (AMAT)

AMAT = time for a hit + miss ratio · miss penalty

= 2 + 0.01 · 100 = 3 cycles

(1 − miss) · 2 + miss · (2 + 100)

→ hit time 2 cycles
→ miss ratio of 0.01
miss penalty of 100 cycles

Hit: lookup tags 2 cycles

Miss: lookup cache 2 cycles; send Mem req 100 cycles

Double freq

miss penalty → 200 cycles

AMAT = 2 + 0.01 200 = 4 cycles

Multi-level caches

to reduce → add a cache

AMAT = hit + miss ratio · miss penalty

= hit_{L1} + miss ratio$_{L1}$ · (hit_{L2} + miss ratio$_{L2}$ · miss penalty)

= 2 + 0.01 (10 + 0.05 · 200)

10 + 10

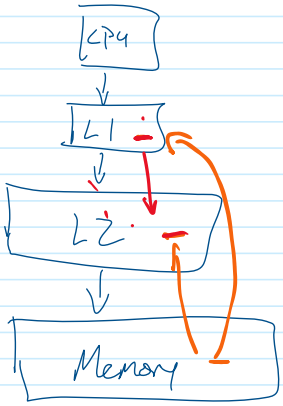
20

= 2 + .2

= 2.2 cycles

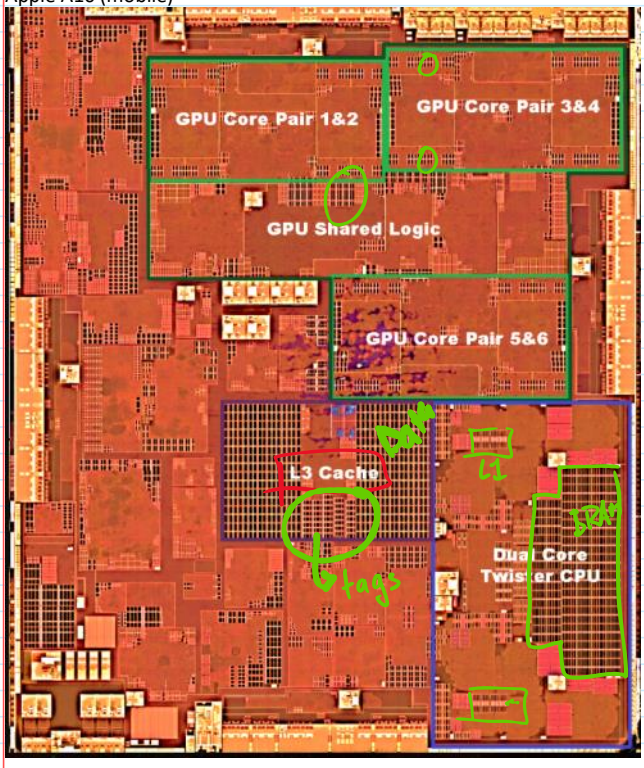
L2 miss ratio of 5%
hit time of 10 cycles

$AMAT_{L2}$

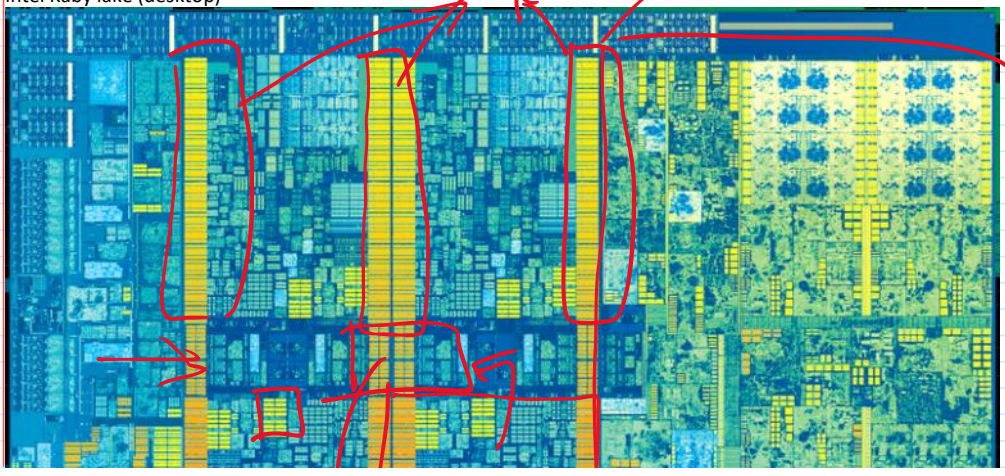


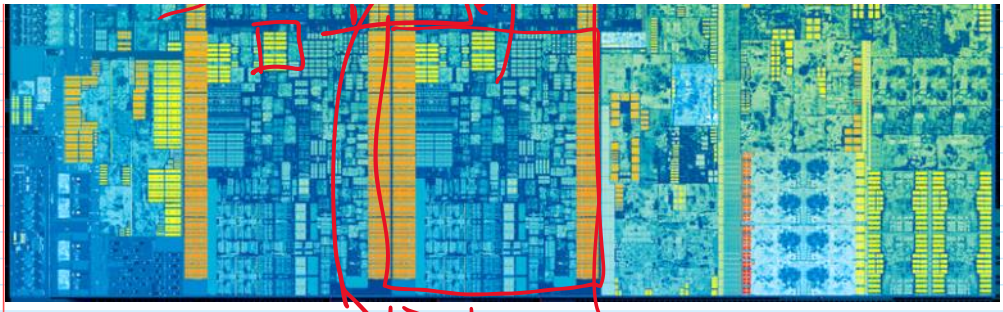
Cache examples

Apple A10 (mobile)

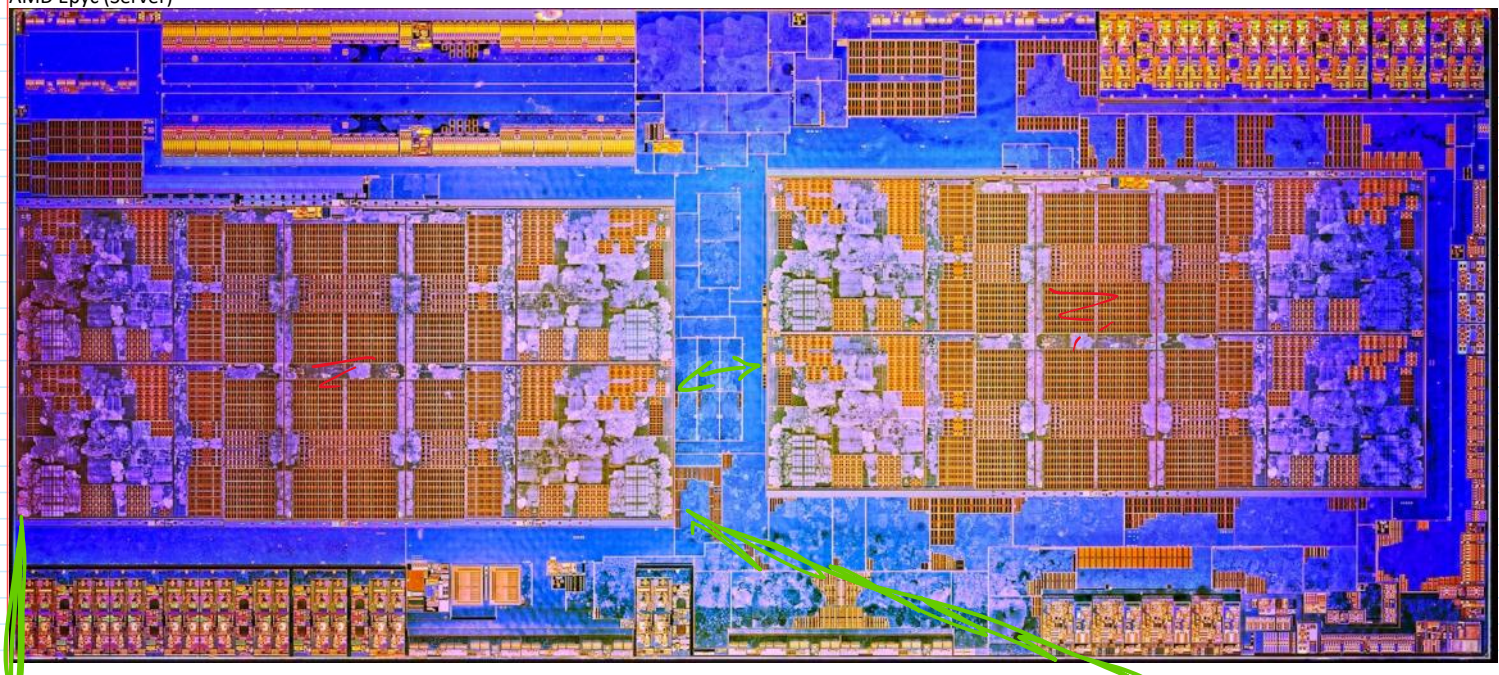


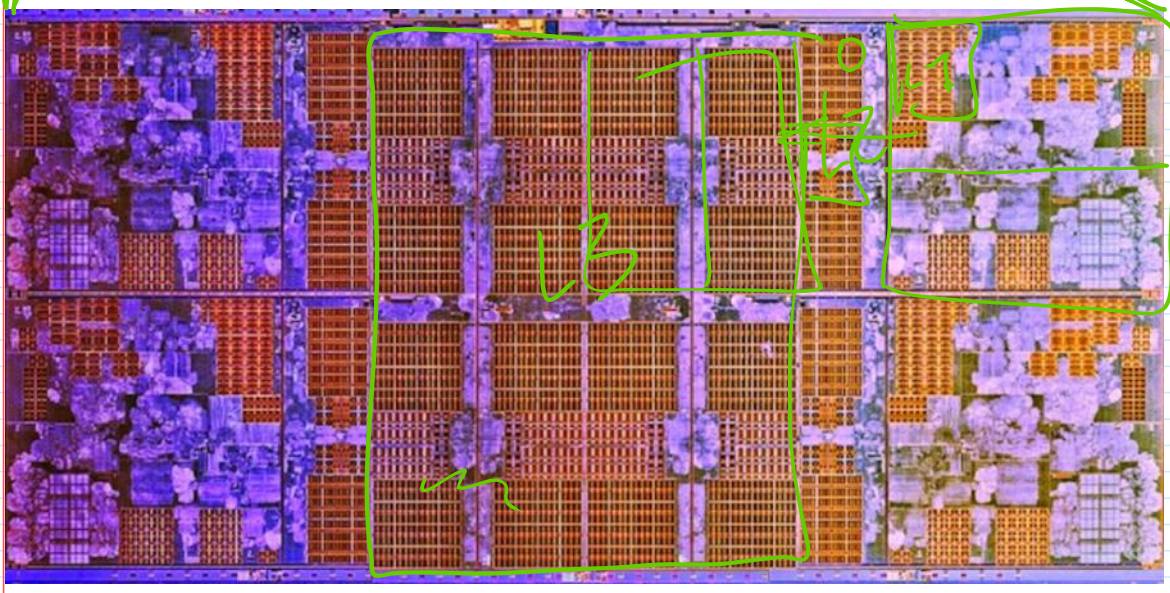
Intel Kaby lake (desktop)





AMD Epyc (Server)





Virtual Memory

What is VM?

Why?

→ illusion to programmer that she has more memory

→ security & flexibility in assigning physical memory

→ combat fragmentation

→ programs appear to be running on their own machine

→ part of ISA s/w - h/w promise

↳ 32 bit ISA

64 bit

→ originally looked like more memory by being a cache of hard drive

↳ swapping

Like a cache

Block size? → 4KB or 2MB

↳ page size

associativity? fully associative b/c → any page (or physical frame) can be mapped anywhere in virtual address space

Capacity? size of main/physical memory