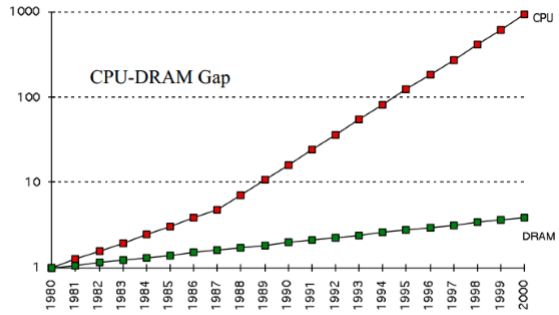
How L1 and L2 CPU caches work, and why they're an essential part of modern chips

Pedro Trancoso

By Joel Hruska (http://www.extremetech.com/extreme/188776-how-l1-and-l2-cpu-caches-work-and-why-theyre-an-essential-part-of-modern-chips)

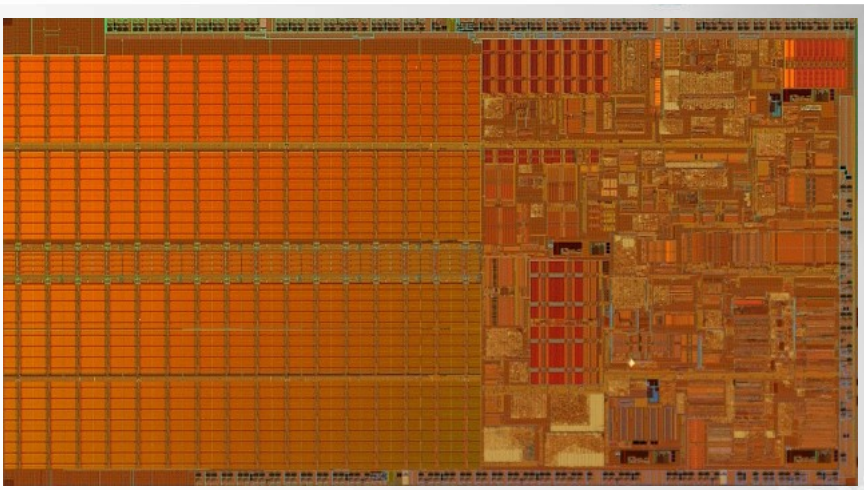
Memory-CPU Gap

- Processor vs Memory Performance



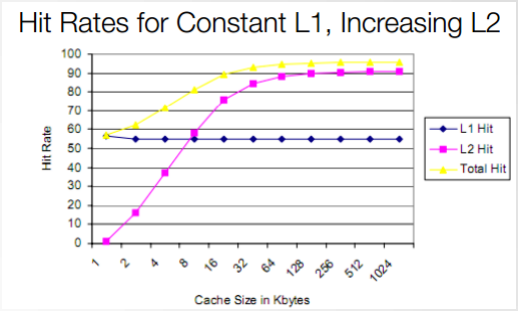
1980: no cache in microprocessor;
1995 2-level cache

Cache-vs-CPU

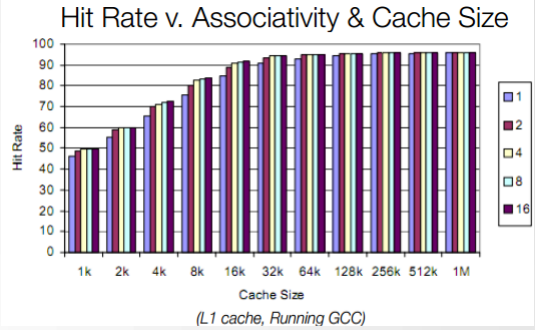


Intel Pentium M

Cache Size and Hierarchy

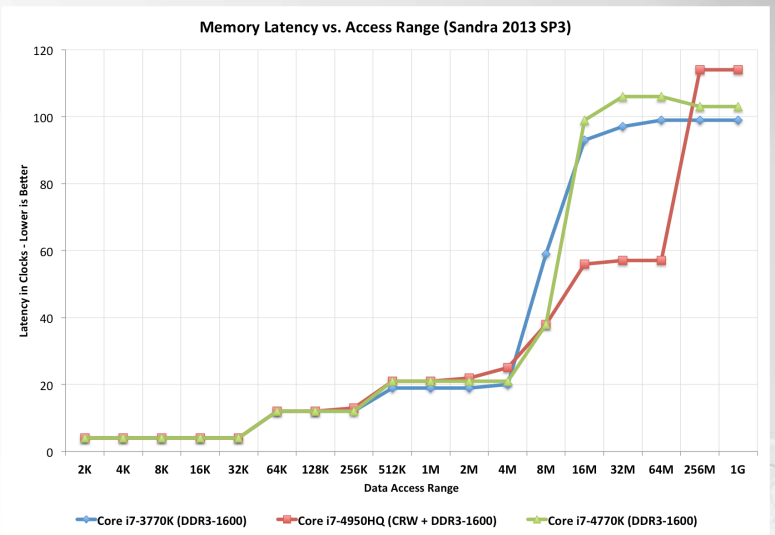


Cache Associativity



5

Memory Latency



6

Miss penalty and Performance

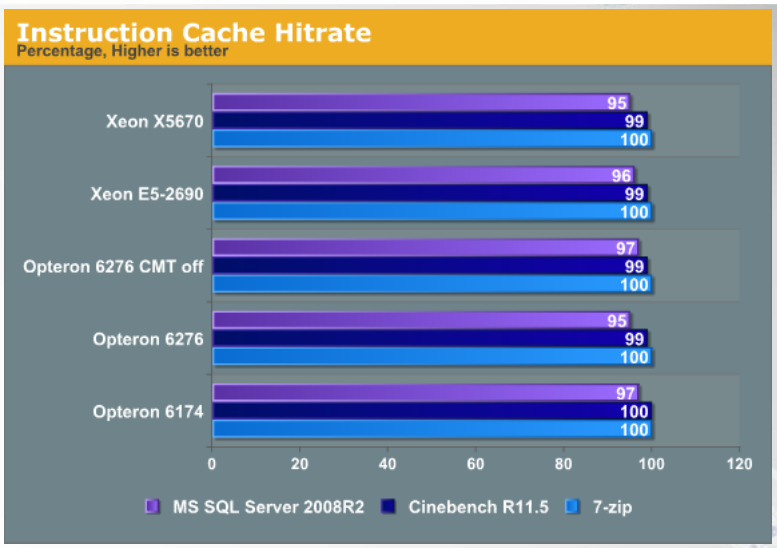
- With a cache with 99% hit rate, the CPU needs its 100th access is in L2, 10-cycle (10ns) access latency. That means: 99 nanoseconds for the first 99 reads and 10 nanoseconds for the 100th. A 1% reduction in hit rate has just slowed the CPU down by 10%.
- In real world: L1 cache typically has a hit rate 95-97%, but the *performance* impact is 14%. We're assuming the missed data is sitting in L2 cache. If the data is sitting in main memory, with an access latency of 80-120ns, the performance difference between a 95-97% hit rate could nearly double the total time needed to execute the code.

Caches in AMD Bulldozer/Piledriver/ Steamroller

Cache	Bulldozer	Piledriver	Steamroller
Level 1 code	64 kB, 2-way, 64 B line size, shared between two cores.	64 kB, 2-way, 64 B line size, shared between two cores.	96 kB, 3-way, 64 B line size, shared between two cores.
Level 1 data	16 kB, 4-way, 64 B line size, per core. Latency 3-4 clocks.	16 kB, 4-way, 64 B line size, per core. Latency 3-4 clocks.	16 kB, 4-way, 64 B line size, per core. Latency 3-4 clocks.
Level 2	1 - 2 MB, 16-way, 64 B line size, shared between two cores. Latency 21 clocks. Read throughput 1 per 4 clock. Write throughput 1 per 12 clock.	2 MB, 16-way, 64 B line size, shared between two cores. Latency 20 clocks. Read throughput 1 per 4 clock. Write throughput 1 per 12 clock.	2 MB, 16-way, 64 B line size, shared between two cores. Latency 19 clocks. Read throughput 1 per 4 clock. Write throughput 1 per 6 clock.
Level 3	0 - 8 MB, 64-way, 64 B line size, shared between all cores. Latency 87 clock. Read throughput 1 per 15 clock. Write throughput 1 per 21 clock.	0 - 8 MB, 64-way, 64 B line size, shared between all cores. Latency 87 clock. Read throughput 1 per 15 clock. Write throughput 1 per 21 clock.	None

Table 14.4. Cache sizes on AMD Bulldozer, Piledriver and Steamroller

Instruction Cache Performance



9

Conclusions

- Old rule of thumb that we add roughly one level of cache every 10 years
  - Intel's Haswell and Broadwell chips offer an enormous L4
- It's an open question at this point whether AMD will ever go down this path (AMD focuses on HAS- Heterogeneous Architecture)
- Cache design, power consumption, and performance will be critical to the performance of future processors

10