High Performance Computing

Memory hierarchy

Martin Raum

Memory bus

The memory bus connects the CPU via the memory controller to system memory.

Memory bandwidth and latency are the most important features.

Memory bandwidth

The rate at which data can be loaded from and stored.

This results from a combination of

- Clock frequency of the memory controller / the bus.
- Transfers per clock cycle: There are types of memory that transfer twice (DDR) or four times (QDR) per base cycle.
- Memory bus width: The number of bits that can be transfered at once. Often 64 bits.
- Number of channels: There is multi-channel systems that multiply the effective width.

Memory bandwidth

Example data from AMD on the Epyc-CPU:

maximal memory bandwidth 170.7 GB/s

split over 8 channels, 21.3 GB/s per channel

on DDR4-2667, i.e., with 2,667 GHz base frequency.

Memory latency

The delay between a request for data in memory until receive on the CPU.

Burst mode allows to save hugely on latency when chunks of data are retrieved.

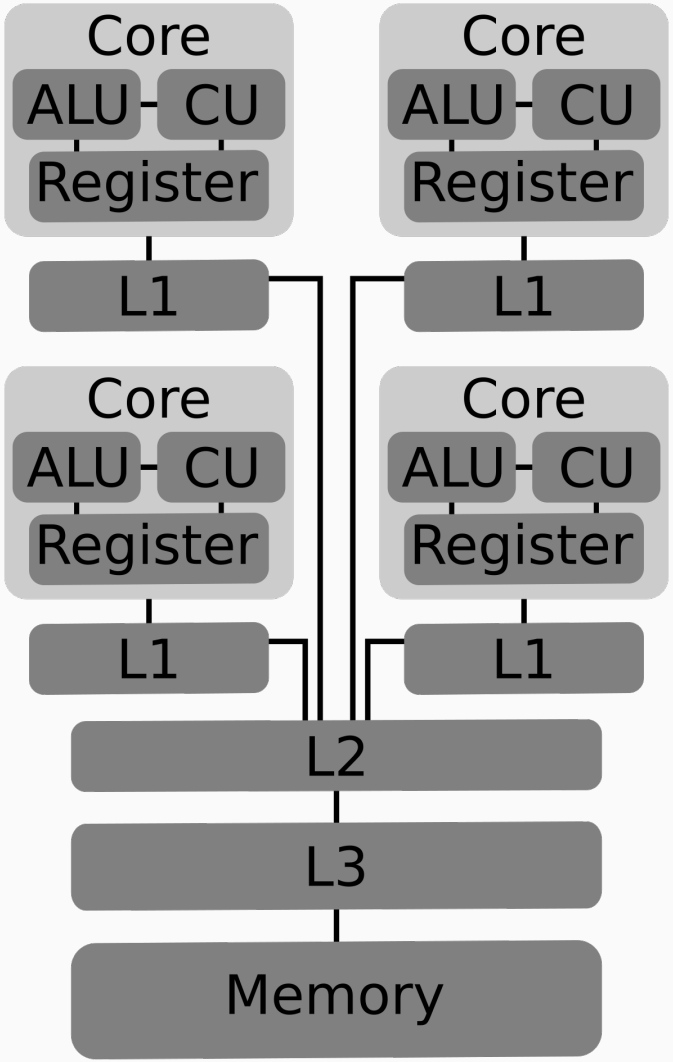
Latency applies to much more than just system memory, but also to cache and even registers.

Memory hierarchy

Memory can be organized by how fast it is connected to the CPU:

- Registers are directly accessible to the ALU, FPU, CU. They are considered part of the CPU.
- Cache comes in different levels: L1, L2, L3, possibly L4. The lower the number the faster and smaller it is.
- System memory is the largest and slowest of all memory that we consider.

Memory attached to the CPU



Cache size and latency

Example for one Epyc-CPU by AMD.

Level	Size (per core)	Latency
L0 Ops Cache	4,096 ops	
L1I Cache	64KiB	
L1D Cache	32KiB	ALU: 4 cycles (≈ 1.2ns)
		FPU: 8 cycles (≈ 2.4ns)
L2 Cache	512KiB	13 cycles (≈ 3.8ns)
L3 Cache	2MiB	≈ 34 cycles (≈ 10ns)
RAM	several GB	≈ 400 cycles (≈ 120ns)

L0, L1, L2 cache is attached to single cores, shared among two hardware threads.

L3 cache is attached to complexes of 4 cores each.

Cache size and latency visualization

Data sources

Technical data on chips can found at `en.wikichip.org`.

The micro-architecture for many CPUs can be found at `www.agner.org/optimize/microarchitecture.pdf`.