Hardware optimizations with CUDA

Robin Dorstijn

Universität Heidelberg

Febuary 2021

In this presentation

- Processors/hardware architecture
- Optimization
- GPU vs CPU
- Writing code

Remark

Project with real code: traffic simulation!
(Over 30h invested)

The project - Requirements

- Cars driving on a highway
- Physical objects
- Aware of each other
- Infinite highway $\rightarrow$ donut

Background - Cache miss

- Intel i7 clock speed: $\approx$ 5Ghz
- $\Rightarrow$ min 0.2ns per operation
- Loading memory from RAM: 60-100ns

Background - Memory types in a processor

- Registers (orders of bits, 0.2ns)
- L1 cache (256KB, 1-2ns)
- L2 cache (256KB - 8MB, 3-5ns)
- L3 cache (4MB - 50MB, 12-40ns)

Background - Memory types in a processor

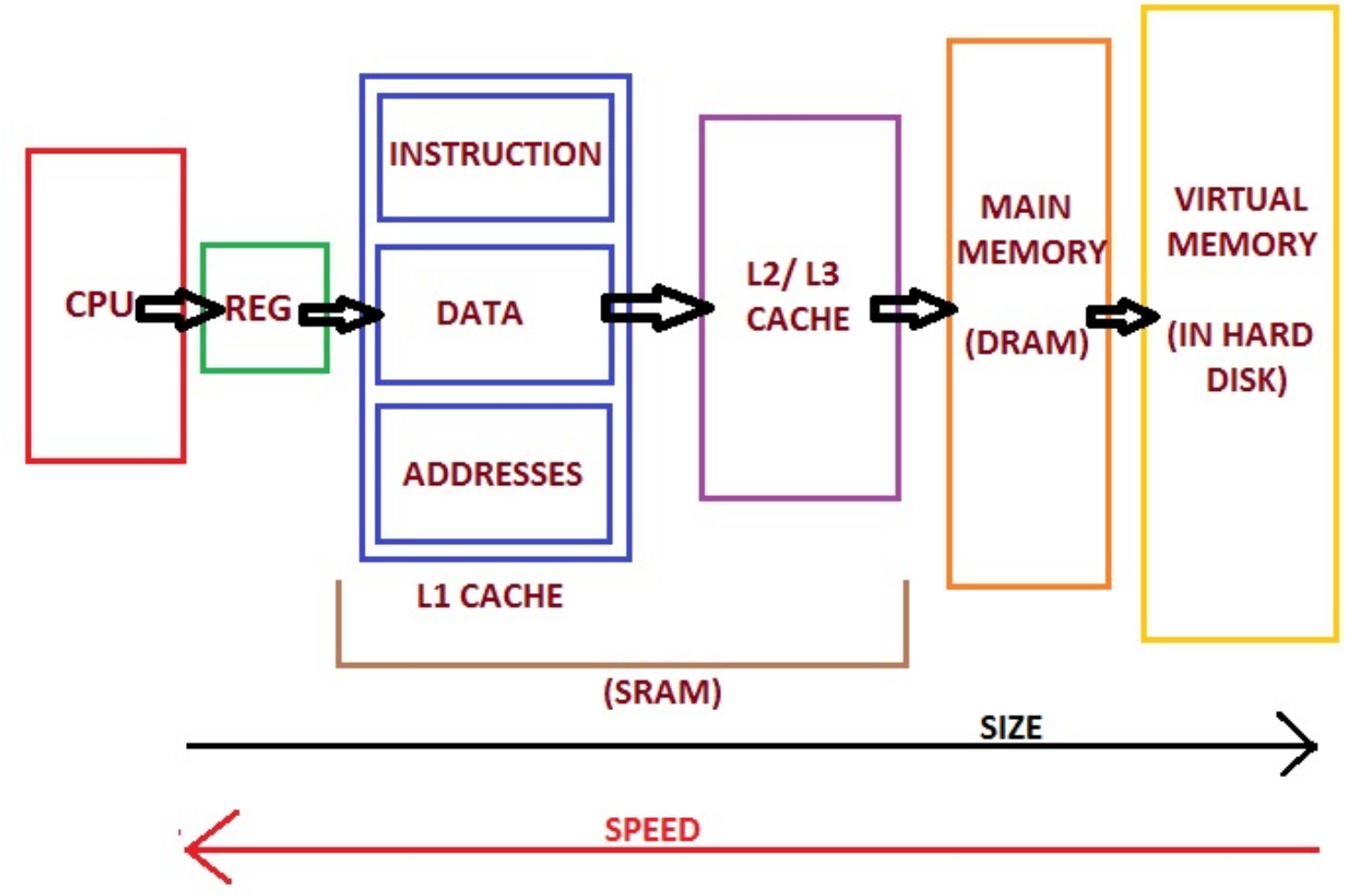

Figure: Memory layout in the CPU

Background - Data oriented design

- Difference struct and class: ability to hide data at compile time.
- Many repetitions of same pointer to functions: much wasted space.
- Therefore: Linear array of structs.

Remark

*Array in the sense of CS, not C++