Erlangen Regional Computing Center

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Node-Level Performance Engineering

Jan Eitzinger

Two-day Tutorial

GWDG, Göttingen

November 19-20, 2019

http://tiny.cc/NLPE-GWDG

High Performance Computing

Course schedule

	Day 1	
10:00	Welcome – Intro	JE
10:15	Computer architecture for software developers	JE
11:30	Performance engineering fundamentals	JE
12:00	Lunch	
13:00	Tools: Topology and affinity, frequency	JE
13:30	Exercise 1: Microarchitecture exploration	JE
14:30	Optimal use of parallel resources: SIMD	JE
16:00	Roofline Model: Basic	JE
17:00	End of day 1	
	Day 2	
9:00	Tools: Performance counters	JE
9:45	Optimal use of parallel resources: ccNUMA	JE
10:45	Exercise 2: Dense Matrix Vector Multiplication	JE
12:00	Lunch	
13:00	Performance Engineering: Basic skills	JE
14:00	Roofline case studies: Jacobi smoother + SpMVM	JE
16:00	Exercise 3: MiniMD Analysis	JE
17:00	End of day 2	

Quiz

http://tiny.cc/NLPE-GWDG

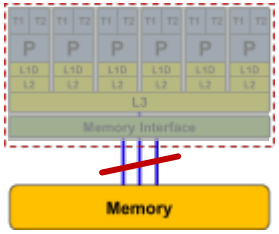
FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

- What does “clock frequency” mean in computers?

  The “heartbeat” of the CPU. A clock cycle is the smallest unit of time on a CPU chip. Typically < 1ns → $f \gtrsim 1\,\mathrm{GHz}$

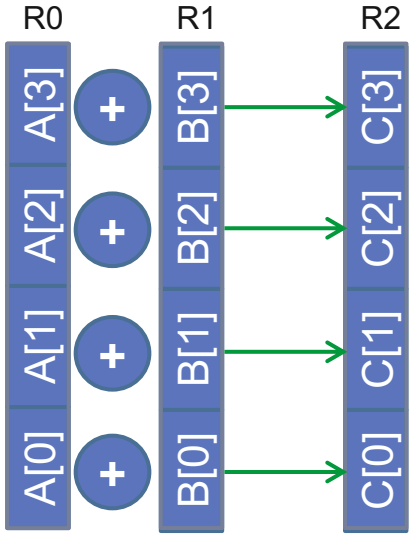
- What is “memory bandwidth”?

  Rate of data transfer between main memory (RAM) and CPU chip. Typical $b_S \approx 10 \ldots 100\,\mathrm{GB/s}$

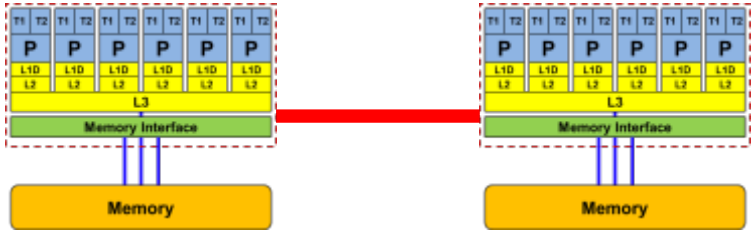


- What is SIMD vectorization?

  Single **I**nstruction **M**ultiple **D**ata.
  Data-parallel load/store and execution units.



- What is ccNUMA?



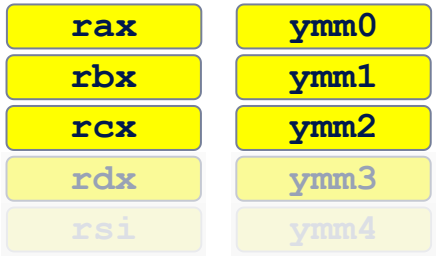
Quiz cont.

http://tiny.cc/NLPE-GWDG

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

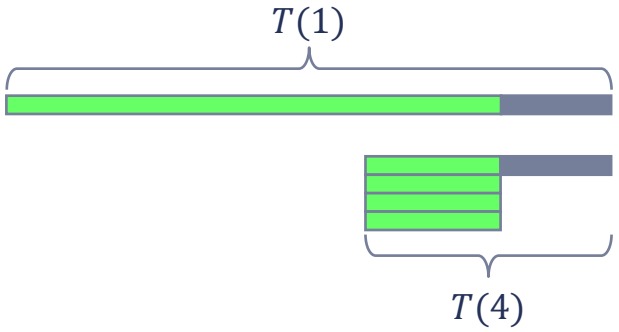
- What is a register?

  A storage unit in the CPU core that can take one single value (a few values in case of SIMD). Operands for computations reside in registers.



- What is Amdahl's Law?

$$S_p = \frac{T(1)}{T(N)} = \frac{1}{s + \frac{1-s}{N}}$$



- What is a pipelined functional unit?

  An instruction execution unit on the core that executes a certain task in several simple sub-steps. The stages of the pipeline can act in parallel on several instructions at once.

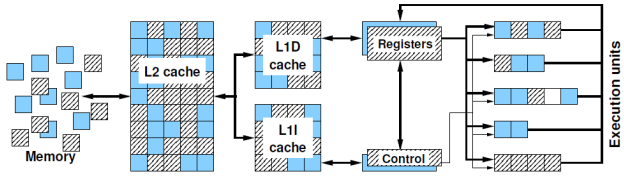
Quiz cont.

http://tiny.cc/NLPE-GWDG

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

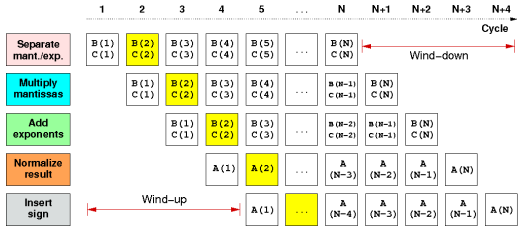
- What is SMT?

  Simultaneous **M**ulti-**T**hreading, a.k.a. hyper-threading.
  A CPU core can execute multiple threads concurrently.
  Such threads share all execution resources except the registers.



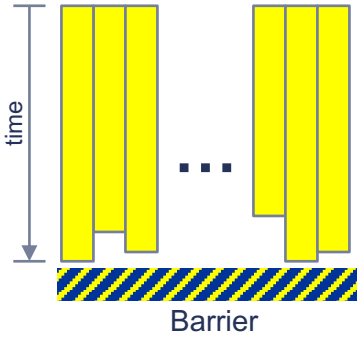
- How long does a single double-precision floating-point multiplication take?

  3-5 clock cycles, depending on the CPU



- How long does an OpenMP barrier take?

  Depending on the OpenMP runtime, the number of cores or threads, and where they are running, a barrier can take tens of thousands of cycles (typically a few 1000s on a full socket).



Cycle gymnastics

- 1 cycle = smallest unit of time on a CPU (“heartbeat”)
  - Clock speed of typical CPU: 2.4 Gcy/s (or GHz)
- Basic unit of work: Floating-point operation (Flop)
  - Typical peak performance of 20-core CPU: P_{peak} =768 Gflop/s
  - How many Flops per cycle per core is that? $\frac{768 \cdot 10^9 \frac{Flops}{s}}{20\ cores \cdot 2.4 \cdot 10^9 \frac{cy}{s}} = 16\ \frac{Flops}{cy \cdot core}$
  - Typical duration of a double precision multiply: 5 cycles
    - How much time is that? $\frac{5\ cy}{2.4 \cdot 10^9 \frac{cy}{s}} = 2.08 \cdot 10^{-9} s = 2.08\ ns$
- Basic unit of traffic: Byte
- Unit of bandwidth: Bytes/s
  - Typical memory bandwidth: 62 Gbytes/s = 6.2 · 10^{10} Bytes/s
  - How many bytes per cycle is that? $\frac{62 \cdot 10^9 \frac{Bytes}{s}}{2.4 \cdot 10^9 \frac{cy}{s}} = 25.8\ \frac{Bytes}{cy}$