Unparalleled Parallelism? CPU & GPU Architecture Trends

…and Their Implications for HPC Software

March 11, 2021 | L. Morgenstern, I. Kabadshow, M. Werner | TU Chemnitz & Jülich Supercomputing Centre

JÜLICH Forschungszentrum

TECHNISCHE UNIVERSITÄT CHEMNITZ

The free lunch is over – again?

Reference to Herb Sutter's
The Free Lunch Is Over:
A Fundamental Turn Toward Concurrency in Software.

Research Question

How to Develop Sustainable HPC Software in a World of Massive Parallelism and Heterogeneity?

To answer this question, consider:

- Available architectures, their similarities and differences
- Type of available hardware parallelism (SIMD vs. SIMT vs. MIMD)
- **Trends in the design of CPU and GPU architectures**
- **Programming model efforts, especially by hardware vendors**

Uniform Architecture Model

Based on OpenCL Platform Model

- Platform consists of **host** CPU and multiple **devices**
- Device consists of multiple **compute units**
- Compute unit consists of multiple **processing elements**
- Mapping to actual hardware components not defined by OpenCL standard

Uniform Architecture Model

Mapping of OpenCL Terminology to CPU and GPU Architectures

OpenCL Platform Model	CPU	GPU (Nvidia/AMD)
Compute Unit	Core	Streaming Multiprocessor/Compute Unit
Processing Element	FP32 SIMD-lane	CUDA Core/Stream Processor

Data Selection

Hardware Architectures

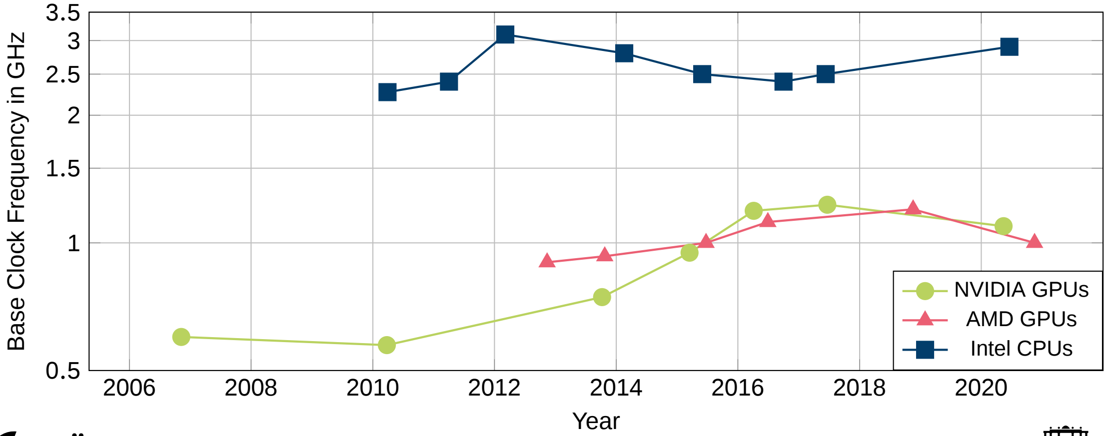
- ✓ Microarchitectures used in HPC systems only:
  - CPUs: Intel Xeon Processors and Intel Xeon Scalable Processors
  - GPUs: Nvidia Data Center (former Tesla) series and AMD Radeon Instinct (former FirePro S) series
- ✗ Excluding dual-GPU designs and multi-socket processors

Metrics

- ✓ Clock frequency: highest base clock frequency a chip/processor of the architecture was operated at
- ✓ Compute unit count: highest number of compute units a chip/processor was designed for
- ✓ Compute unit size: number of processing elements per compute unit

Clock Frequency

Development of Clock Frequencies for CPUs and GPUs since 2007



Stagnating Clock Frequencies

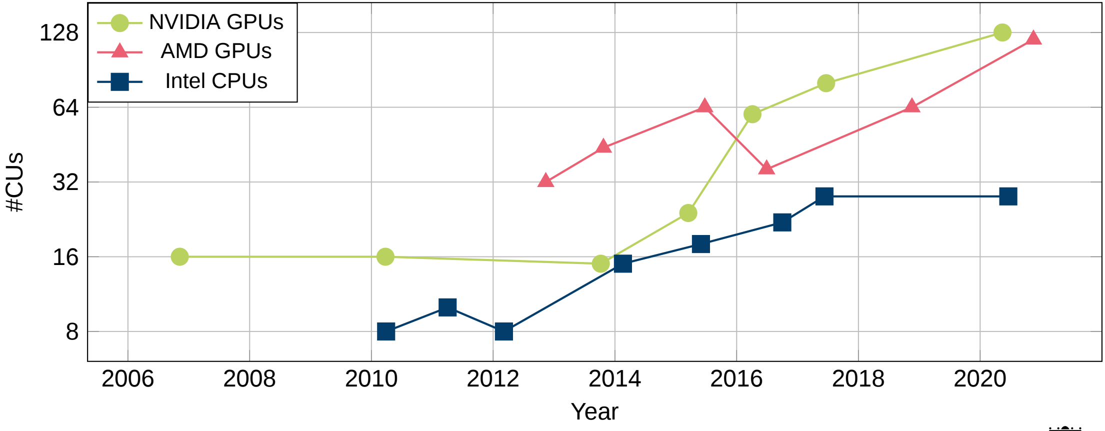
Implications for HPC Software

As soon as algorithmic data-parallelism is exhausted,

there is no more “automagical” speed-up at all!

Compute Unit Count

Development of Compute Unit Count for CPUs and GPUs since 2007



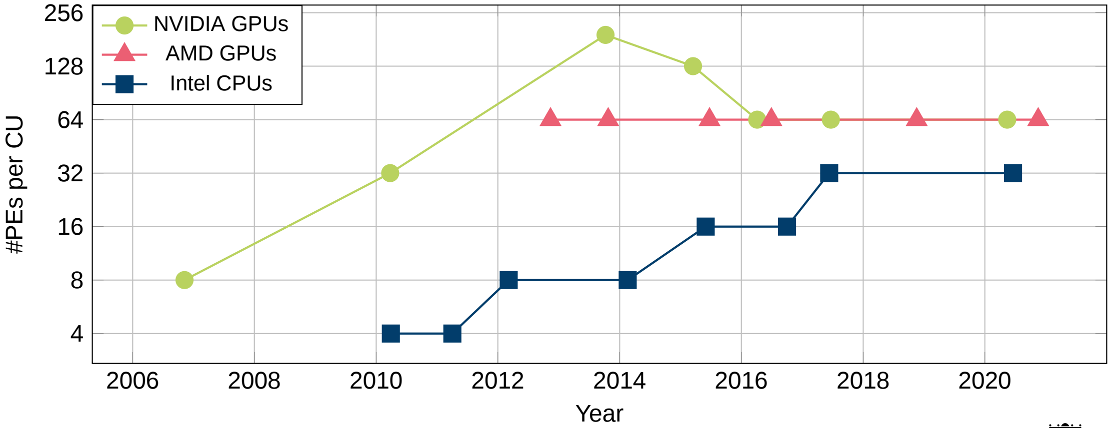
Increasing Compute Unit Count

Implications for HPC Software

- Increasing MIMD parallelism – on both architectures!
- Increases amount of kernel/threads that can be executed in parallel
- Increases programming flexibility → new possibilities to express task-parallelism:
  - CUDA asynchronous task graphs
  - ROCm's Asynchronous Task and Memory Interface
  - OpenMP tasks on GPUs

Compute Unit Size

Development of Compute Unit Size for CPUs and GPUs since 2007



Stagnating Compute Unit Size

Implications for HPC Software

- Reflects limited amount of data-parallelism in algorithms
- Compute units provide more flexibility:
  - Independent thread scheduling → synchronization, threading, tasking as on CPUs
  - Independent data paths for FP and INT operations → concurrent execution of compute and adressing operations

Programming Model Features…

…that Ease and Unify the Programmability of CPUs and GPUs

Memory Management

- CUDA's unified memory → uniform view on physically separated host and device memory
- OpenCL's shared virtual memory → seamlessly share pointers between host and device code

Unifying Programming Approaches

- Intel's SYCL and C++-based oneAPI
- Nvidia's libcu++
- AMD's ROCm

Conclusion

- CPUs and GPUs are developing in the same direction
- Stagnating SIMD, but increasing MIMD parallelism
- Unifying approaches that ease programmability
- Task-based and heterogeneous programming approaches developed in FGBS:
  - Whippletree
  - MxKernel
  - Eventify
  - PGASUS
  - CloudCL

The free lunch is over – again.

Unparalleled Parallelism? CPU & GPU Architecture Trends

…and Their Implications for HPC Software

March 11, 2021 | L. Morgenstern, I. Kabadshow, M. Werner | TU Chemnitz & Jülich Supercomputing Centre

JÜLICH Forschungszentrum

TECHNISCHE UNIVERSITÄT CHEMNITZ