HPC.NRW

NHR4CES NHR for Computational Engineering Science

TECHNISCHE UNIVERSITÄT DARMSTADT | RWTH AACHEN UNIVERSITY

HPC Architecture Basics and RWTH Resources

What is a supercomputer?

Tim Cramer

GREAT COMPUTING COMES WITH GREAT SUPPORT.

Levels of parallelism in hardware

HPC.NRW

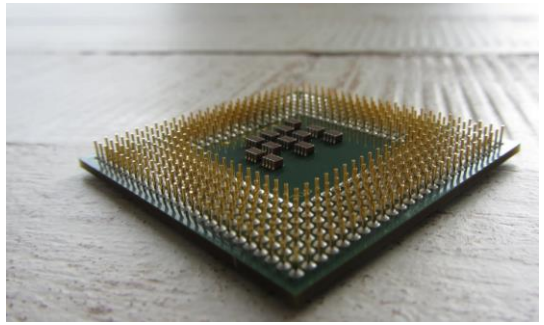
- A modern supercomputer may contain multiple levels of parallelism
  - Processor level parallelism: Superscalar, SIMD
  - Node/Chip level: Several cores/processors run in parallel with access to the same memory
  - System level: Several nodes run in parallel and are communicating over a network interconnect

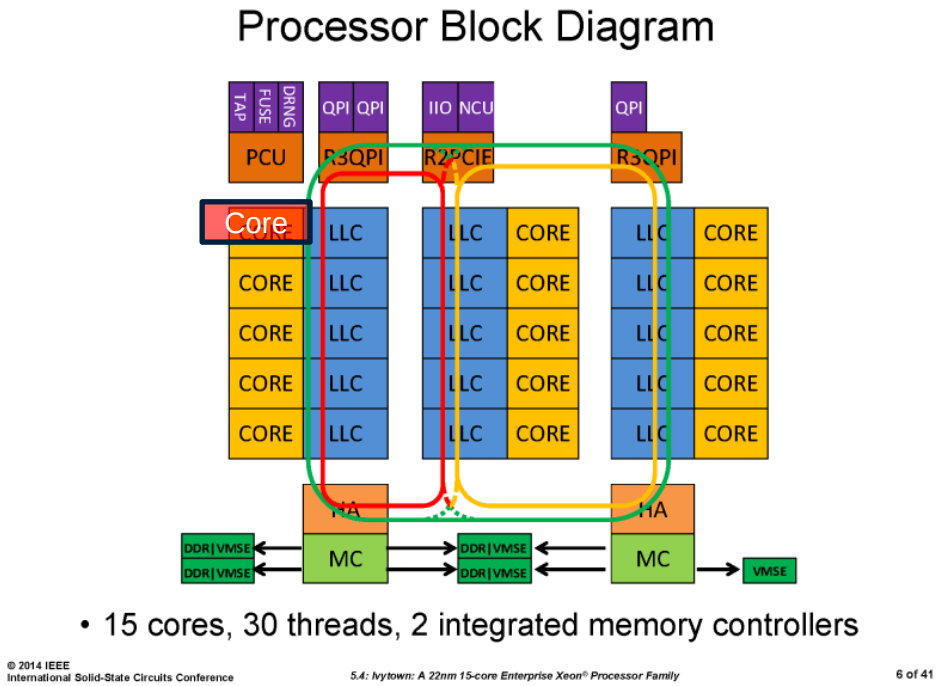
- Parallelism introduces overhead
  - Additional computational costs (cycles)
  - Implementation (hours of work)

- Overhead increases from processor to system level

Agenda

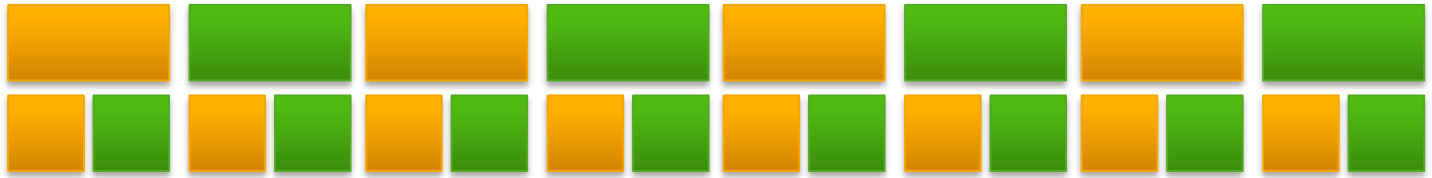
What is a core?





Processor Level Parallelism

- Parallelism at processor/ instruction level
  - Pipelining (overlap in execution: load, decode, execute)
  - Superscalar (redundant arithmetical units: Multiplication, Addition, …)
  - SIMD execution (e.g. 512 bit registers, AVX-512)



8 x double

16 x float

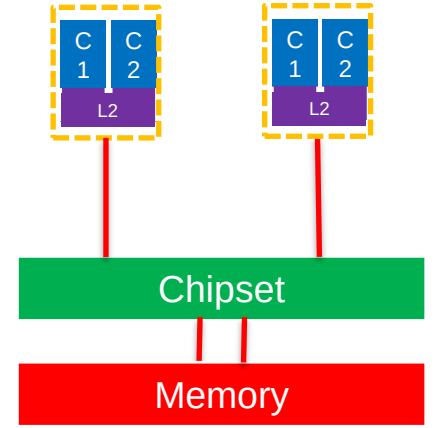
- Programming techniques
  - Code modifications: Unrolling, Cache reuse
  - Compiler optimizations

What is a Node in a Cluster?

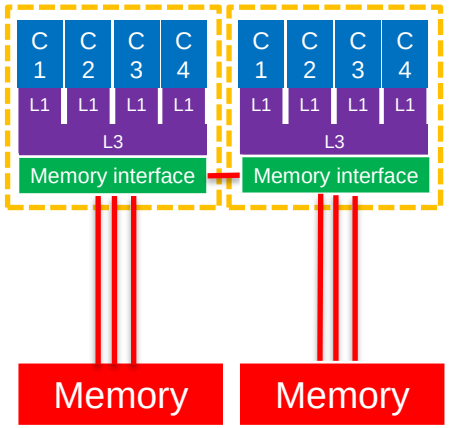
Node in a Cluster

HPC.NRW

- A node may contain
  - One or more (multi-core) processors
  - Memory hierarchy (caches, disks, etc.)
  - Interconnects, power supply, fans, …
  - Accelerators
- Multicore Designs
  - Early multicore design
    - Uniform Memory Architecture (UMA)
    - Flat Memory design
  - Recent multicore design
    - ccNUMA (Cache Coherent Non-Uniform Memory Architecture)
    - Memory Interface + HT/QPI provides inter-socket connectivity



UMA



NUMA

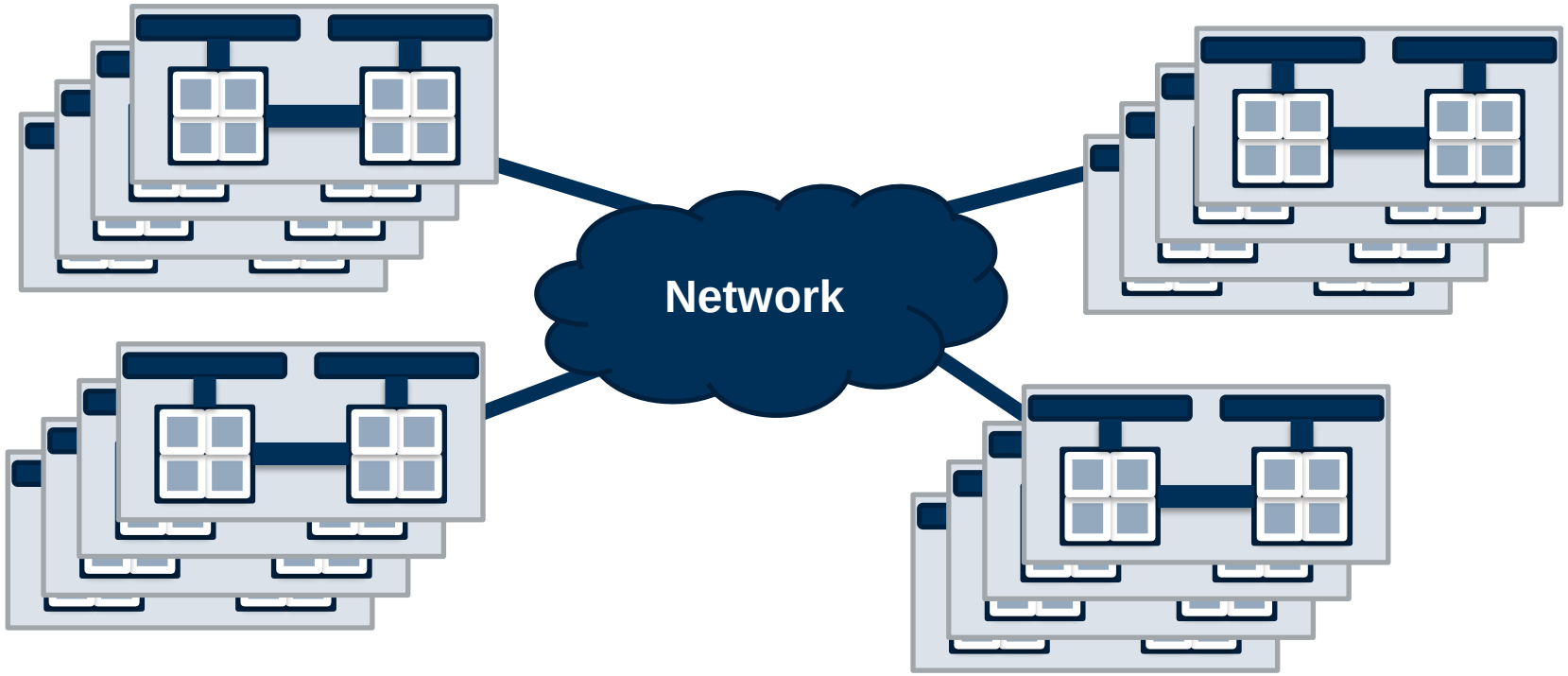
What is a Cluster?

What is a Cluster?

Coming soon!
CLAIX-2023 in
pilot phase

HPC Cluster

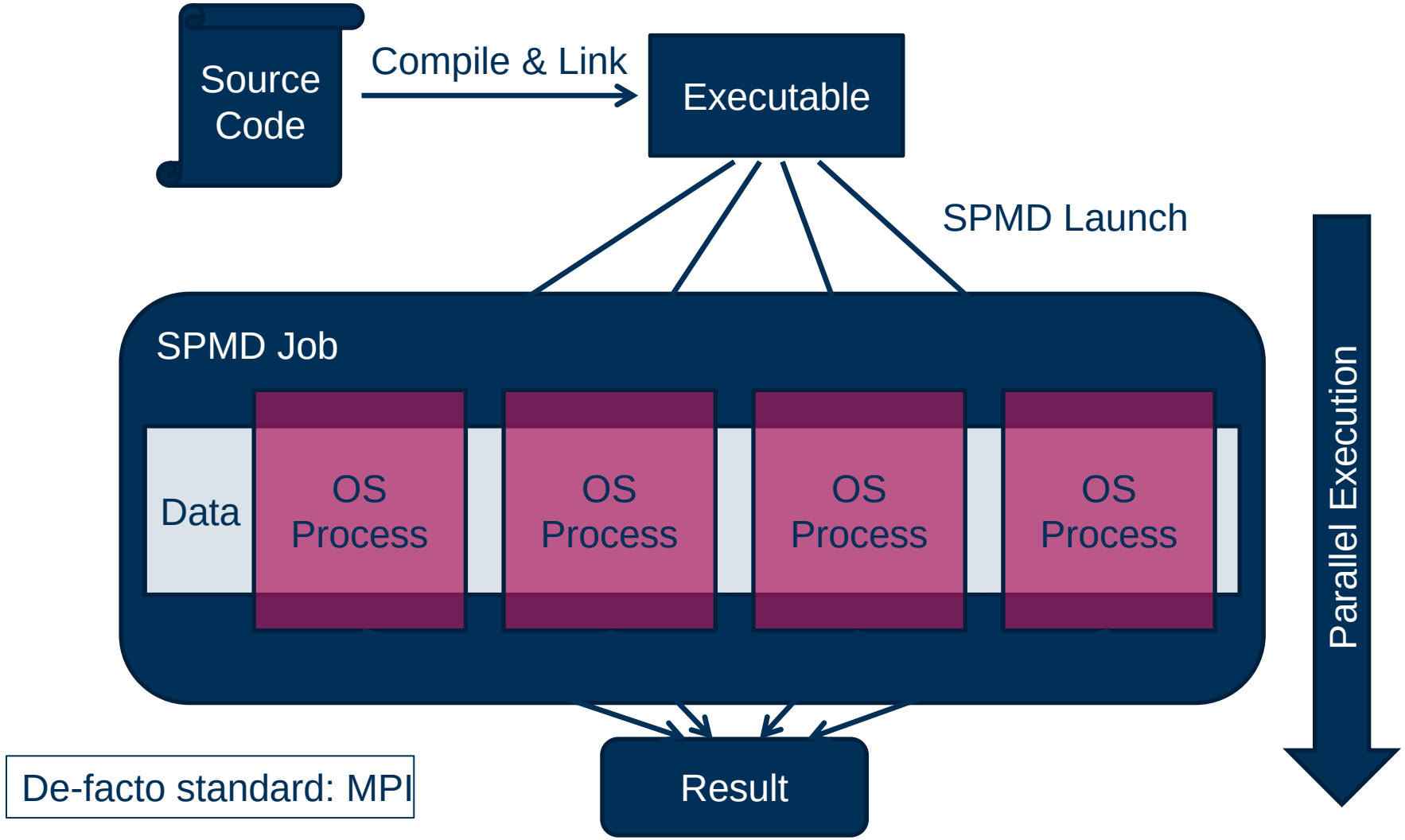
- HPC market is dominated by distributed memory multicomputers (clusters)
- Many nodes with no direct access to other nodes' memory



How to execute a program on the complete cluster?

SPMD:

Single Program
Multiple Data

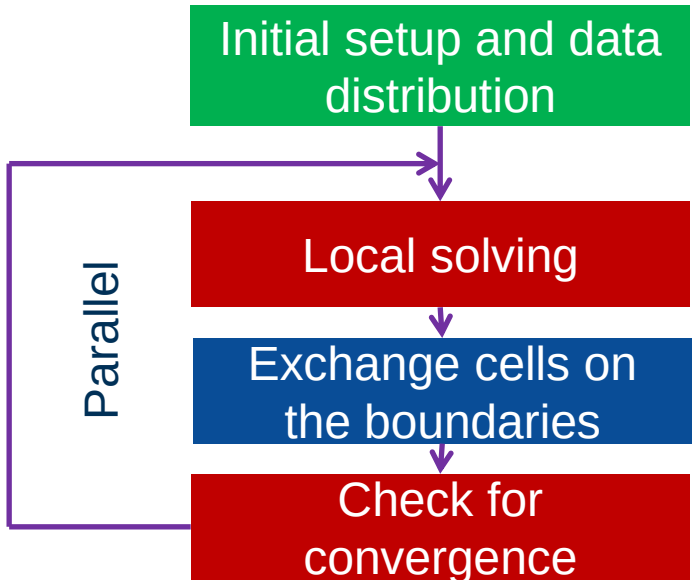


De-facto standard: MPI

System Level Parallelism

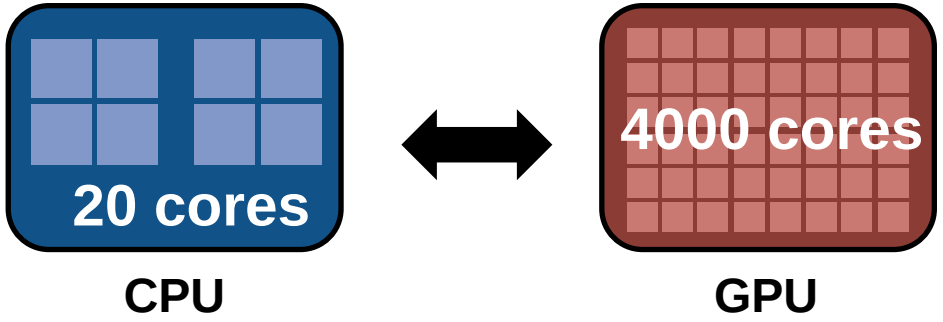
– Example: Domain decomposition in CFD: Mapping of 3D mesh to the processors

– Programming techniques
  – Data parallel approach
    – Distribute data structures
  – Parallel algorithms
  – Explicit data exchange (MPI)



What is an Accelerator in a Node?

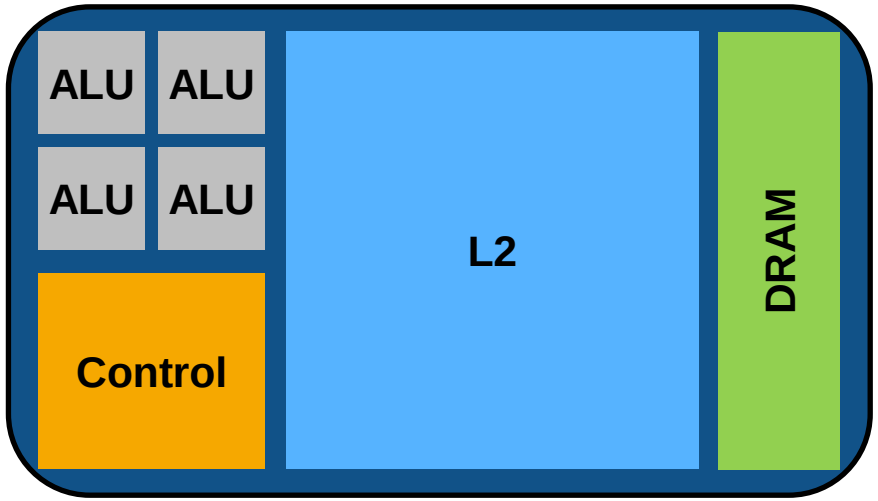
CPU vs. GPU



- GPU-Threads
  - Thousands (“few” on CPU)
  - Light-weight, little creation overhead
  - Fast switching
- Lots of parallelism needed on GPU to get good performance!

CPU vs. GPU

– Different design



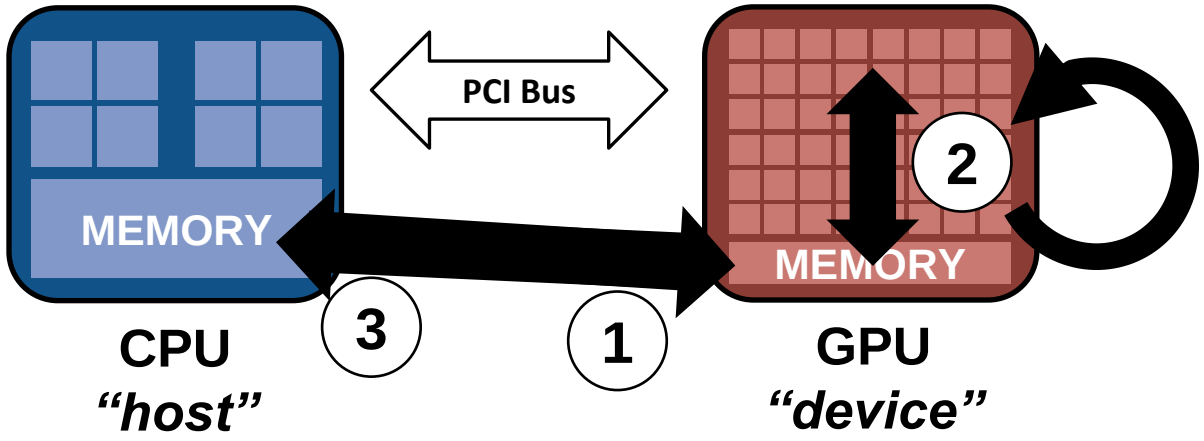
CPU

- → Optimized for low latencies
- → Huge caches
- → Control logic for out-of-order and speculative execution
- → **Targets on general-purpose applications**

GPU

- → Optimized for data-parallel throughput
- → Architecture tolerant of memory latency
- → More transistors dedicated to computation
- → **Suited for special kind of apps**

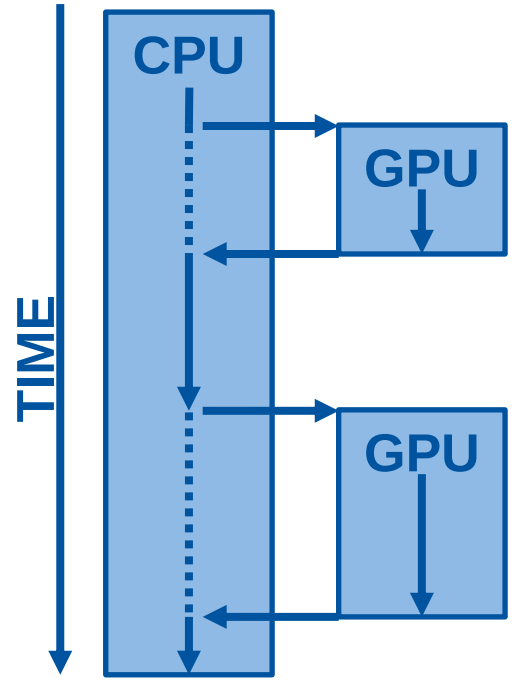
Offloading



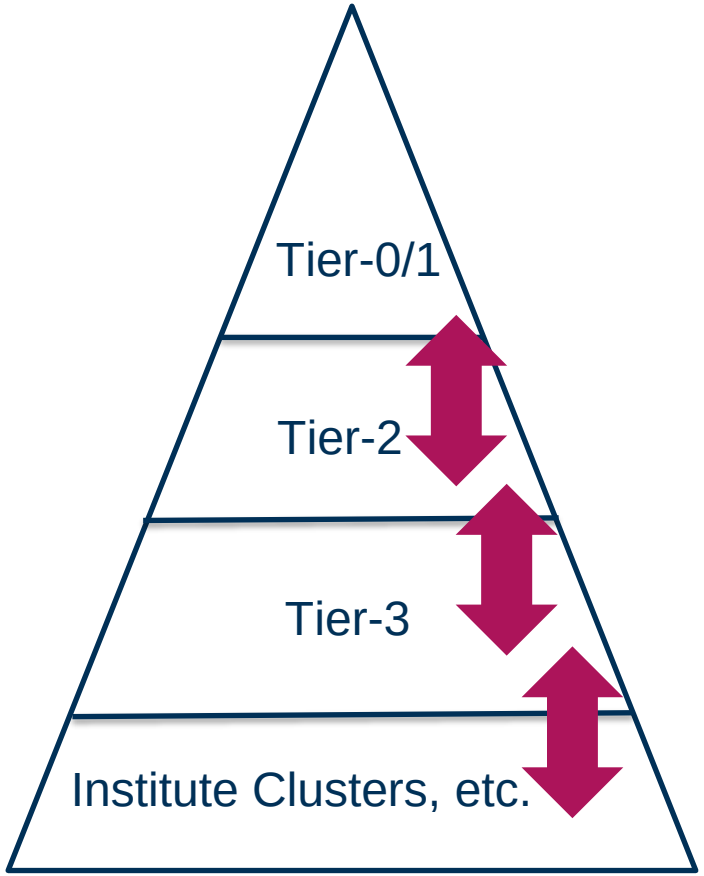
We refer to "discrete GPUs" here.

- Weak memory model
  - Host + device memory = separate entities
  - No coherence between host + device
    - **Data transfers** needed
- Host-directed execution model
  - Copy input data from CPU mem. to device mem.
  - Execute the device program
  - Copy results from device mem. to CPU mem.

processing flow *(simplified)*



HPC Pyramid



Tier 0: European Level

- Partnership for Advanced Computing in Europe (PRACE)
- https://prace-ri.eu/hpc-access/calls-for-proposals

Tier 1: National Level (large scale)

- Gauss Centre for Supercomputing (GCS)
- Jülich (JSC), Munich (LRZ), Stuttgart (HLRS)
- https://www.gauss-centre.eu/for-users/hpc-access

Tier 2: Regional-National Level

- Nationales Hochleistungsrechnen (NHR): https://www.nhr-verein.de
- Gauss Allianz (GA): https://gauss-allianz.de
- Aachen, Cologne, Paderborn (and others outside NRW)

Tier 3: Regional Level

- E.g. local universities

HPC Resources @ RWTH Aachen University

HPC.NRW

	CLAIX-2018 (Tier-2 + Tier-3)
Peak Perform. Hosts and GPUs (avail. resources)	2.3 + 0.7 PFlops (400 Mio Coreh + 80 Mio Coreh)
MPI Nodes	**1032 + 216 MPI nodes** 2-socket Intel Skylake processors (Platinum 8160, 2x24 cores, 192 GB, 2.1 GHz) additional dialog and service nodes
GPU Nodes	**48 + 6 GPU nodes** 2-socket Intel Skylake processors (Platinum 8160, 2x24 cores, 192 GB, 2.1 GHz) 2 NVIDIA Tesla V100 GPUs (16 GB HBM2)
Fabric	Intel OmniPath network (OPA) 1:2 blocking, 16X PCIgen3
Storage	10 PB Lustre Storage BEEOND on SSDs (480 GB)

Reaches EOL soon!
CLAIX-2023 in pilot phase

HPC Resources @ RWTH Aachen University

HPC.NRW

	CLAIX-2023 (Tier-2 + Tier-3)
Theoretical Peak Performance CPUs	2.6 + 1.4 PFlops
Theoretical Peak Performance GPUs	4.4 + 0.7 PFlops
Available resources CPUs	346 + 185 Mio Coreh
Available resources GPUs	27 + 4 Mio Coreh (1 GPU-h == 24 Core-h)
HPC Segment	**412 + 220 HPC nodes** 2-socket Intel Sapphire Rapids (Xeon 8468, 2x48 cores, 2.1 GHz) • 470 nodes with 256 GB • 160 nodes with 512 GB • 2 nodes with 1024 GB
ML Segment	**32 + 5 ML nodes** 2-socket Intel Sapphire Rapids (Xeon 8468, 2x48 cores, 2.1 GHz, 256 GB) **4x NVIDIA H100** 96 GB HBM2e per node
Interactive Segment	Additional nodes with smaller GPUs (e.g., for JupyterHub usage)
Fabric	Infiniband NDR network (OPA) 2:1 blocking
Storage	26 PiB Lustre Storage BEEOND on SSDs (1.4 TB per node)

Coming soon!
Pilot phase started!

HPC Resources @ RWTH Aachen University

HPC.NRW

	CLAIX-2023 (Tier-2 + Tier-3)
Theoretical Peak Performance CPUs	2.6 + 1.4 PFlops
Theoretical Peak Performance GPUs	4.4 + 0.7 PFlops
Available resources CPUs	185 Mio Coreh
Available resources GPUs	Coreh (1 GPU-h == 24 Core-h)
HPC Segment	**220 HPC nodes** socket Intel Sapphire Rapids (Xeon 8468, 2x48 cores, 2.1 GHz) • 470 nodes with 256 GB • 160 nodes with 512 GB • 2 nodes with 1024 GB
ML Segment	**32 + 5 ML nodes** 2-socket Intel Sapphire Rapids (Xeon 8468, 2x48 cores, 2.1 GHz, 256 GB) **4x NVIDIA H100** 96 GB HBM2e per node
Interactive ent	Additional nodes with smaller GPUs (e.g., for JupyterHub usage)
Fabric	Infiniband NDR network (OPA) 2:1 blocking
Storage	26 PiB Lustre Storage BEEOND on SSDs (1.4 TB per node)

Coming soon: Additional WestAI resources: 15 ML nodes

Coming soon! Pilot phase started!

Conclusion

HPC.NRW

Questions?