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

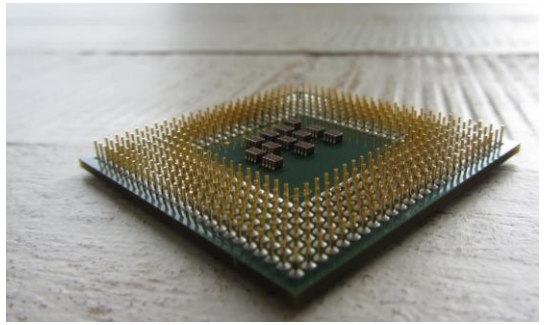
- A modern supercomputer may contain multiple levels of parallelism
  - Processor level parallelism: Superscalar, SIMD
  - Node/Chip level: Several cores/processors run in parallel with access to the same memory
  - System level: Several nodes run in parallel and are communicating over a network interconnect

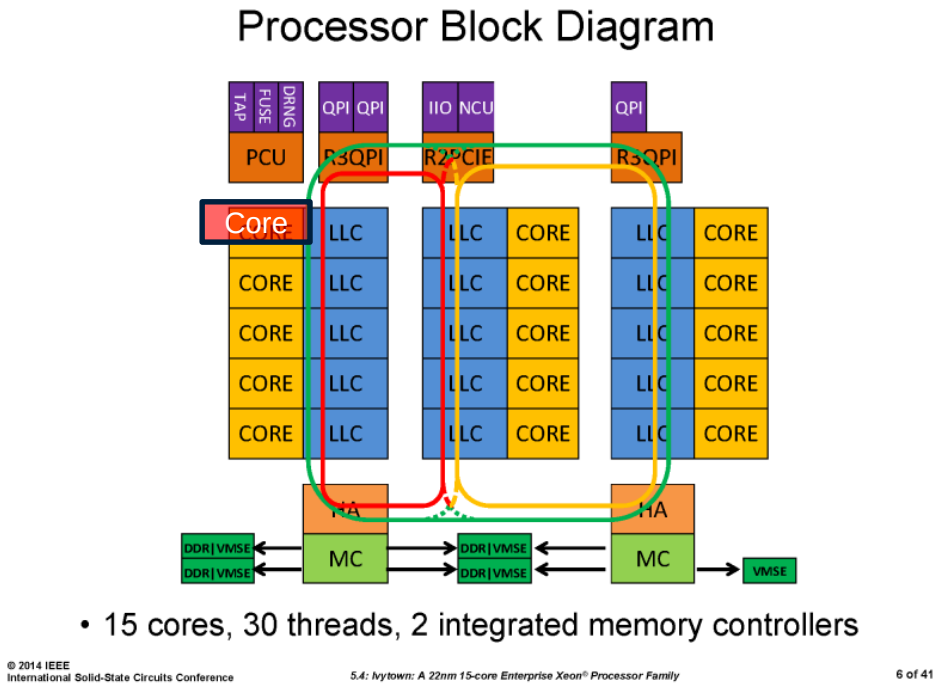
- Parallelism introduces overhead
  - Additional computational costs (cycles)
  - Implementation (hours of work)

- Overhead increases from processor to system level

Agenda

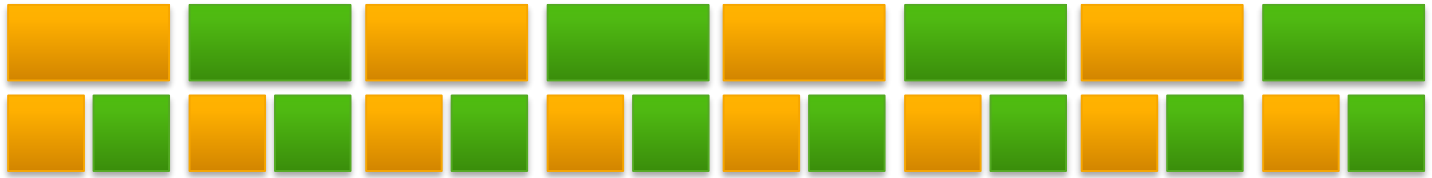
What is a core?





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

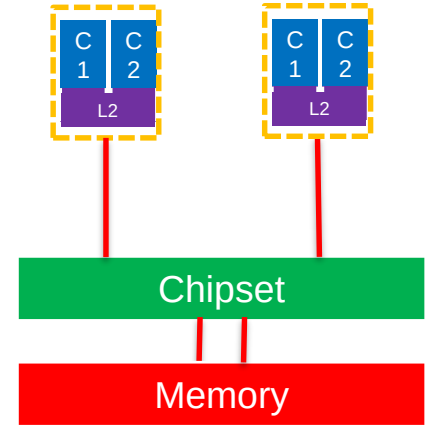
What is a Node in a Cluster?

HPC.NRW

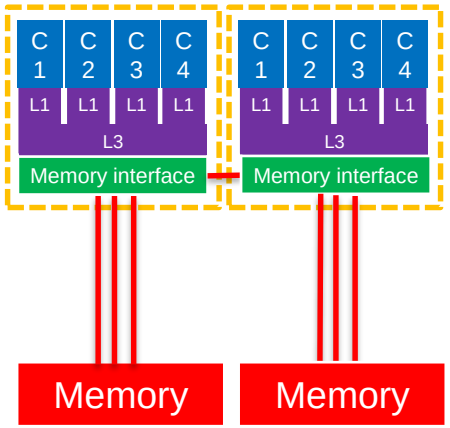
Node in a Cluster

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

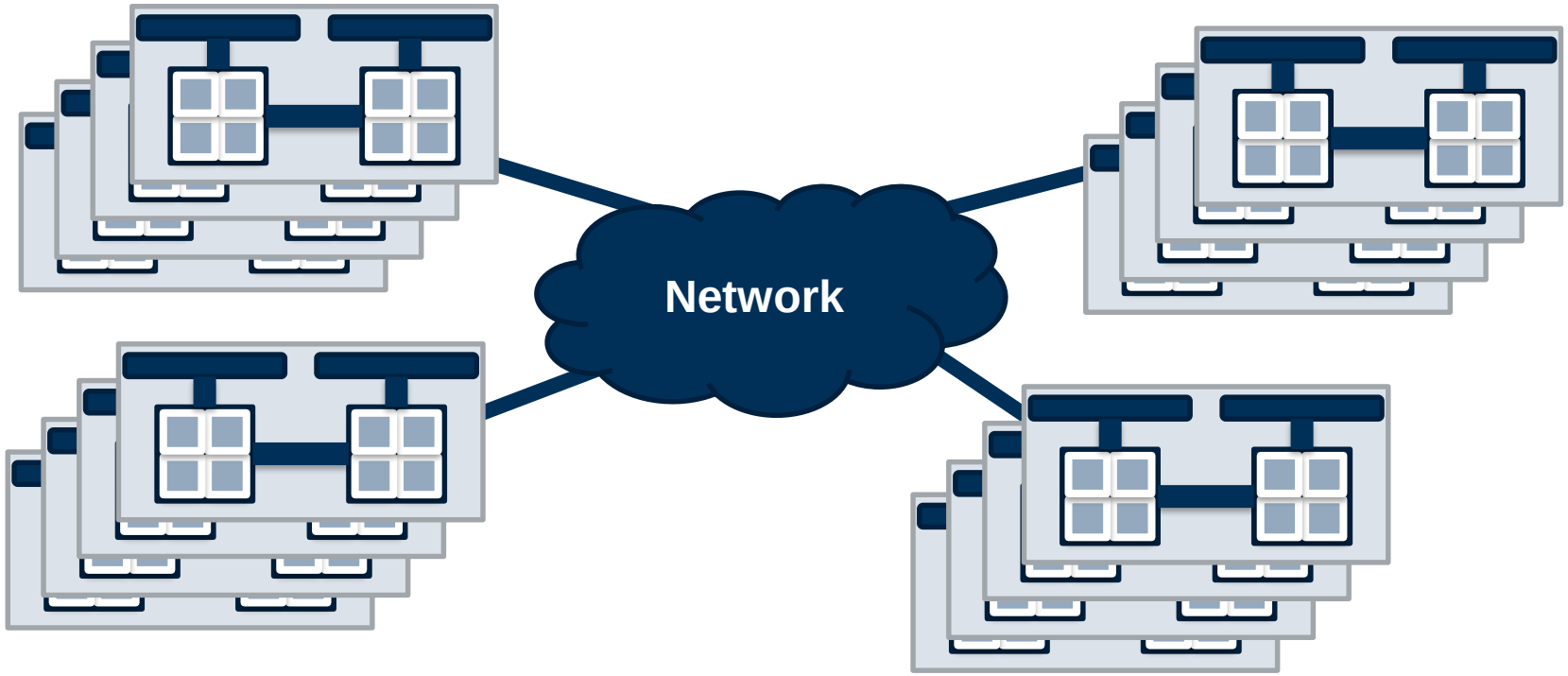
What is a Cluster?

Core
Accelerator
Node
Cluster

HPC.NRW

HPC Cluster

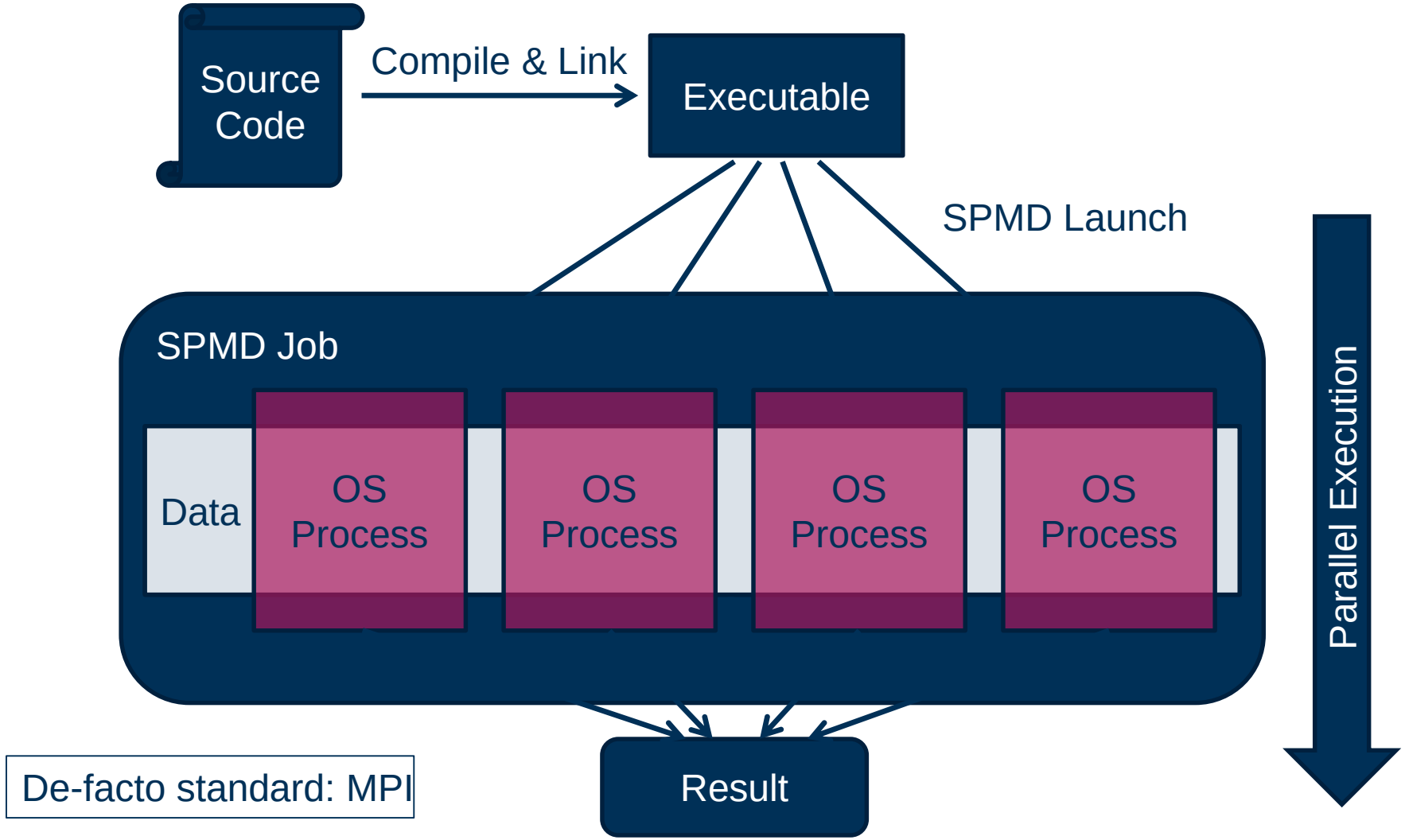
- HPC market is dominated by distributed memory multicomputers (clusters)
- Many nodes with no direct access to other nodes' memory



How to execute a program on the complete cluster?

SPMD:

Single Program
Multiple Data

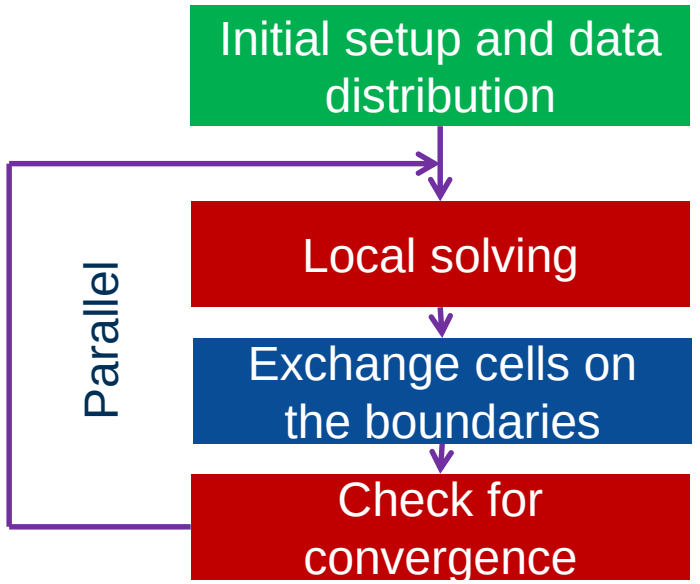


De-facto standard: MPI

System Level Parallelism

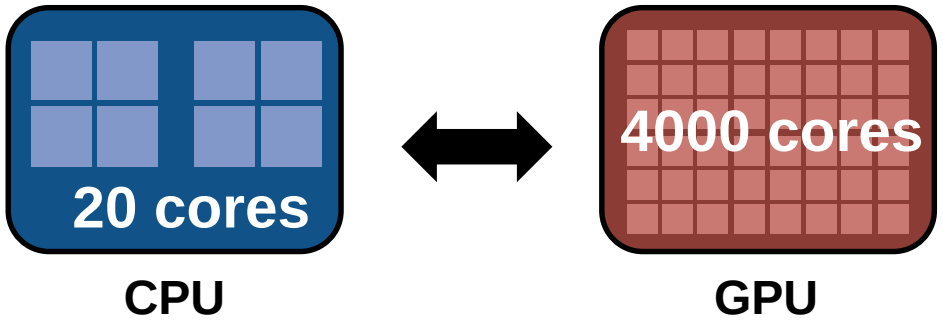
– Example: Domain decomposition in CFD: Mapping of 3D mesh to the processors

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



What is an Accelerator in a Node?

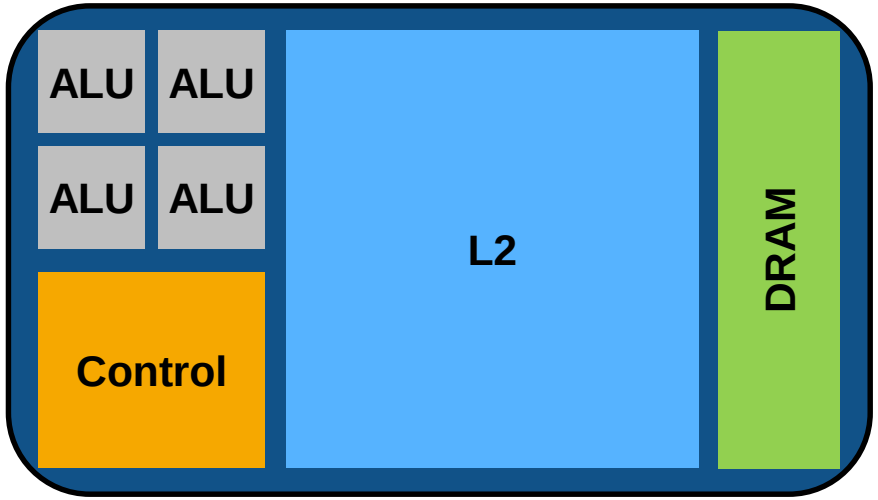
CPU vs. GPU



- 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



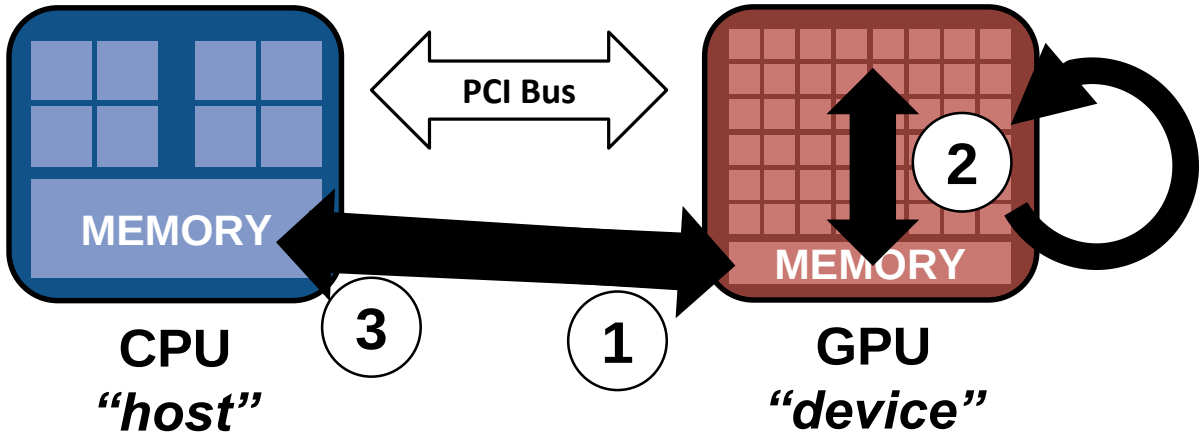
CPU

- → 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**

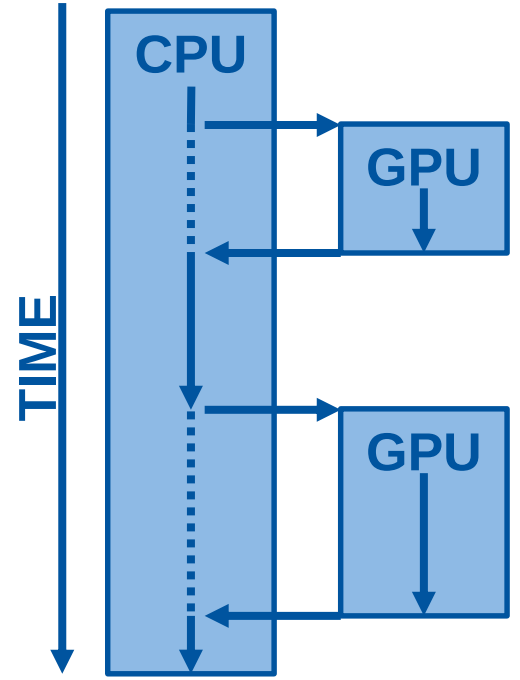
Offloading



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 Resources @ RWTH Aachen University

HPC.NRW

	CLAIX-2016	CLAIX-2018 (Tier-2 + Tier-3)
Peak Perform. Hosts + GPUs (avail. resources)	0.56 + 0.13 PFlops (125 Mio Coreh)	2.3 + 0.7 PFlops (400 Mio Coreh + 80 Mio Coreh)
MPI Nodes	**609 MPI nodes** 2-socket Intel Broadwell processors (E5-2695v4, 2x12 cores, 128 GB, 2.2 GHz) 8 additional nodes with 2 TB NVMe additional dialog and service nodes	**1032 + 216 MPI nodes** 2-socket Intel Skylake processors (Platinum 8160, 2x24 cores, 192 GB, 2.1 GHz) additional dialog and service nodes
SMP Nodes	**8 SMP nodes** 8-socket Intel Broadwell processors (E7-8860v4, 8x18 cores, 1024 GB, 2.2 GHz)	-
GPU Nodes	**10 GPU nodes** 2-socket Intel Broadwell processors (E5-2695v4, 2x12 cores, 128 GB, 2.2 GHz) 2 NVIDIA P100 GPUs (16 GB)	**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	Intel OmniPath network (OPA) 1:2 blocking, 16X PCIgen3
Storage	3 PB Lustre Storage 1.3 PB Home Storage (EMC2 Isilon)	10 PB Lustre Storage BEEOND on SSDs (480 GB)

HPC Resources @ RWTH Aachen University

HPC.NRW

	CLAIX-2016	CLAIX-2018 (Tier-2 + Tier-3)
Peak Perform. Hosts + GPUs (avail. resources)	0.56 + 0.13 PFlops (125 Mio Coreh)	2.3 + 0.7 PFlops (400 Mio Coreh + 80 Mio Coreh)
MPI Nodes	**609 MPI nodes** 2-socket Intel Broadwell (E5-2695v4, 2x12 cores, 12 8 additional nodes additional dialog e nodes	**1032 + 216 MPI nodes** 2-socket Intel Skylake processors (Platinum 8160, 2x24 cores, 192 GB, 2.1 GHz) /Me additional dialog and service nodes
SMP Nodes	**8 SMP nodes** 8-socket In ell processors (E7-8860v s, 1024 GB, 2.2 GHz)	-
GPU Nodes	**10 GP** 2-so Broadwell processors (E 12 cores, 128 GB, 2.2 GHz) P100 GPUs (16 GB)	**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	mniPath network (OPA) 1:2 blocking, 16X PCIgen3	Intel OmniPath network (OPA) 1:2 blocking, 16X PCIgen3
Storage	3 PB Lustre Storage 1.3 PB Home Storage (EMC2 Isilon)	10 PB Lustre Storage BEEOND on SSDs (480 GB)

Out of maintenance, will be taken out of production in 2022

Conclusion

HPC.NRW

Questions?