



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

RWTHAACHEN  
UNIVERSITY

# HPC Architecture Basics and RWTH Resources

What is a supercomputer?

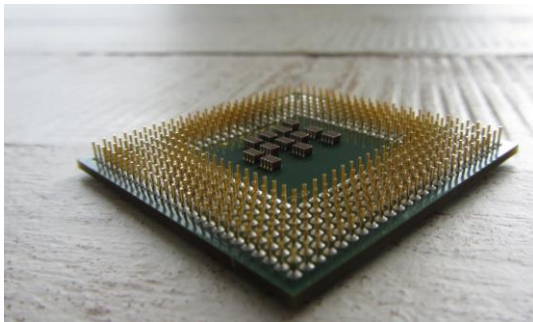
Tim Cramer

- 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

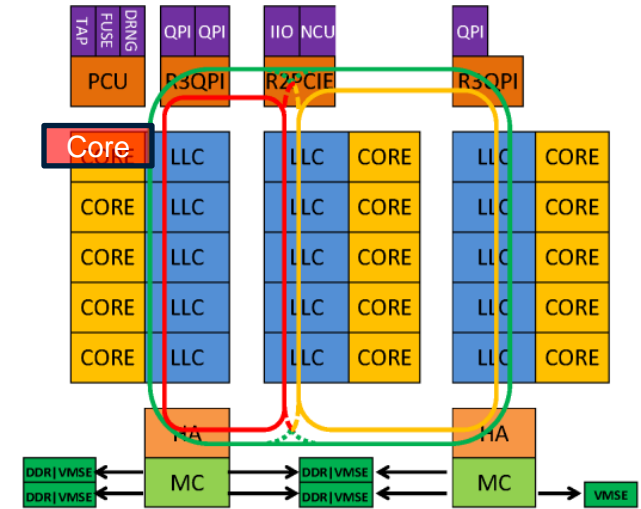




# What is a core?



Processor Block Diagram



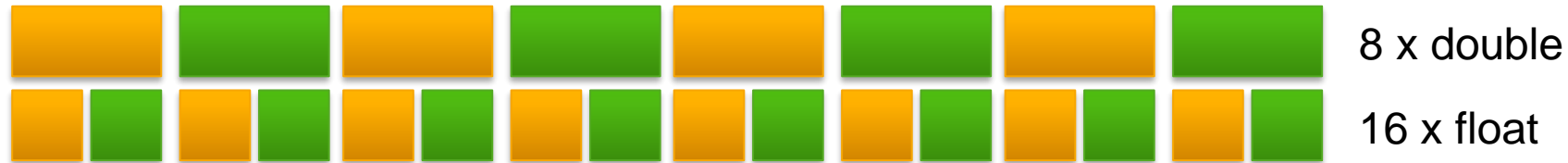
- 15 cores, 30 threads, 2 integrated memory controllers

© 2014 IEEE  
International Solid-State Circuits Conference

5.4: Ivytown: A 22nm 15-core Enterprise Xeon® Processor Family

6 of 41

- 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)



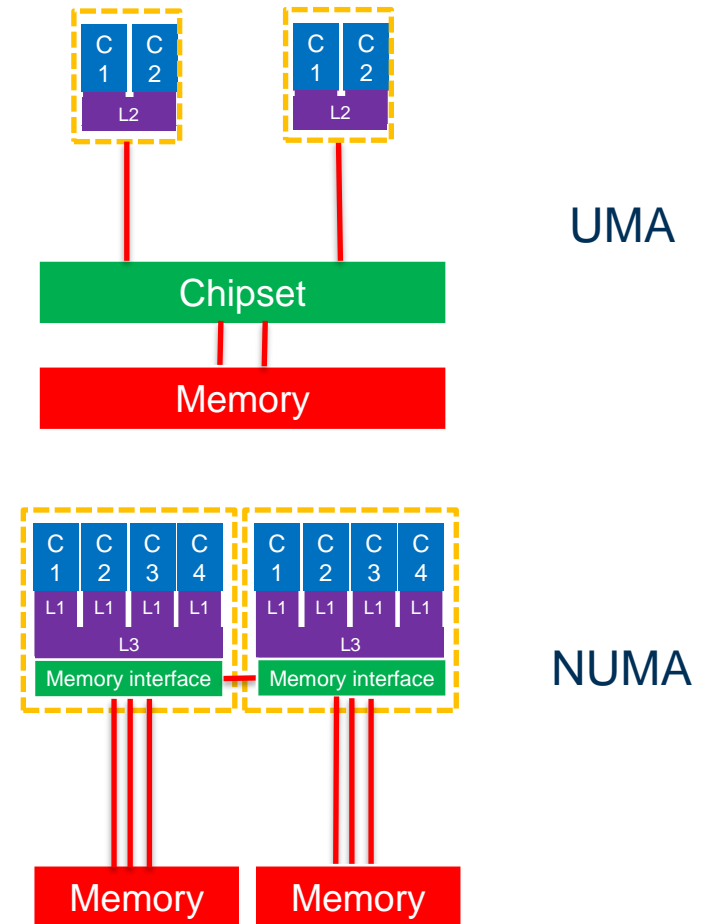
- Programming techniques
  - Code modifications: Unrolling, Cache reuse
  - Compiler optimizations

# What is a Node in a Cluster?



# Node in a Cluster

- 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



# What is a Cluster?



CLAIX-2018, not in production anymore





# What is a Cluster?



CLAIX-2018, not in production anymore



Only doors „survived“ as artwork by Tim Berresheim

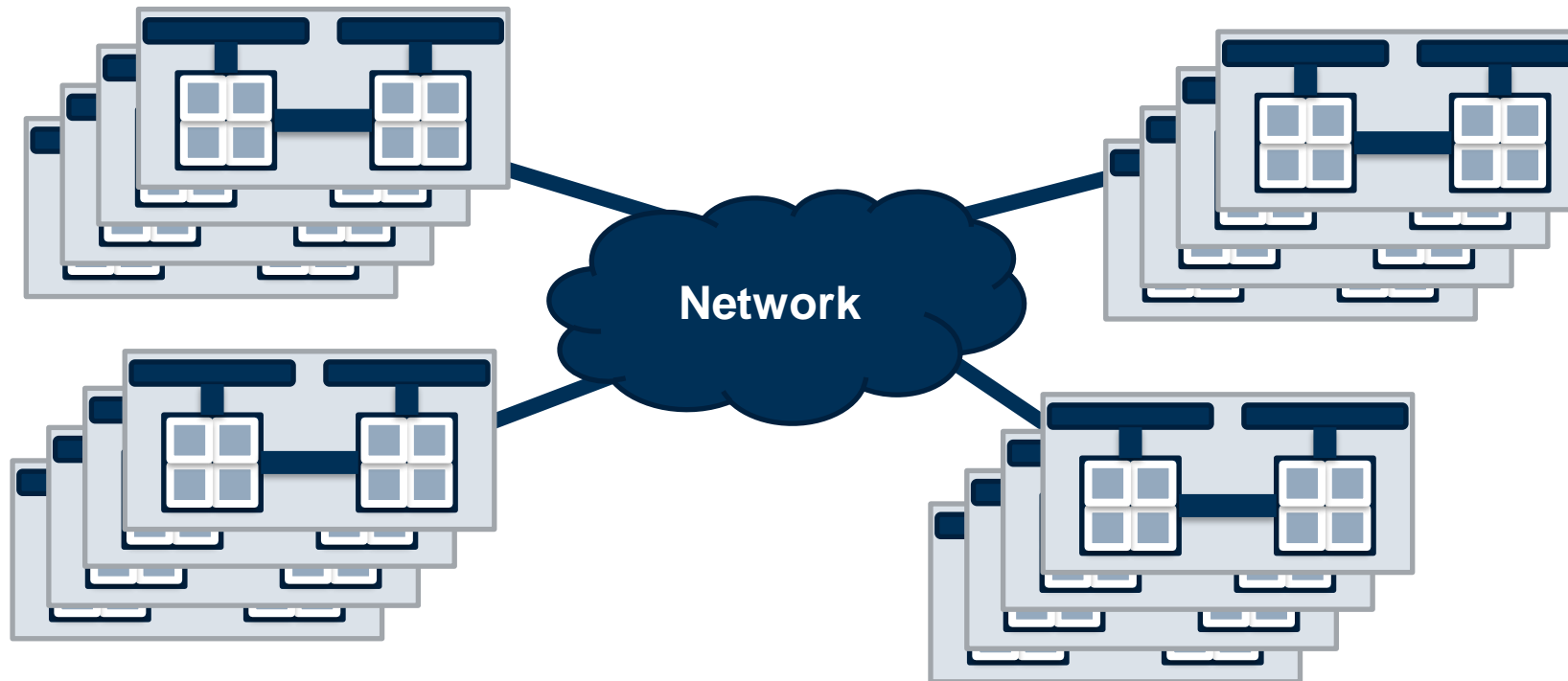
# What is a Cluster?



CLAIX-2023



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

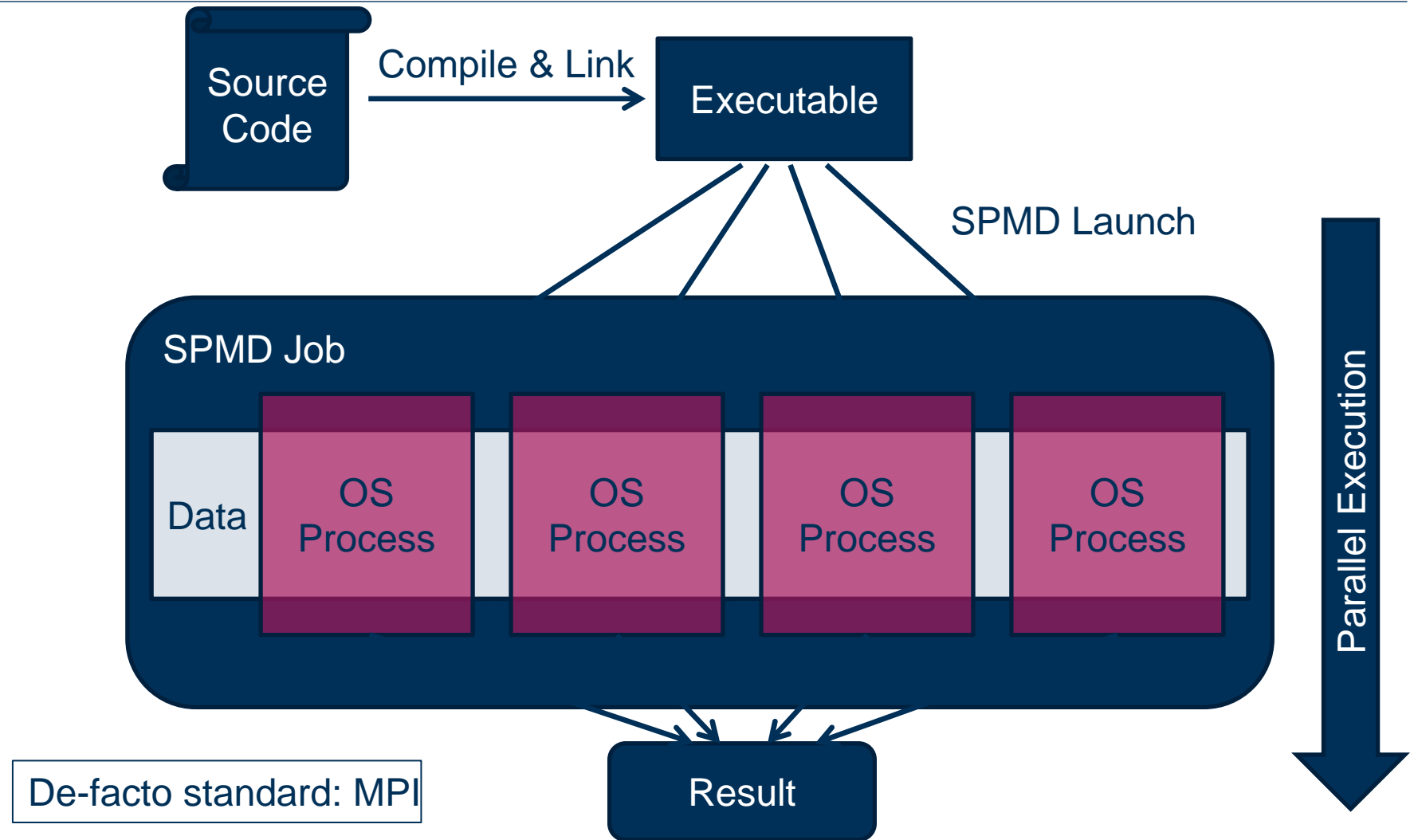


# How to execute a program on the complete cluster?

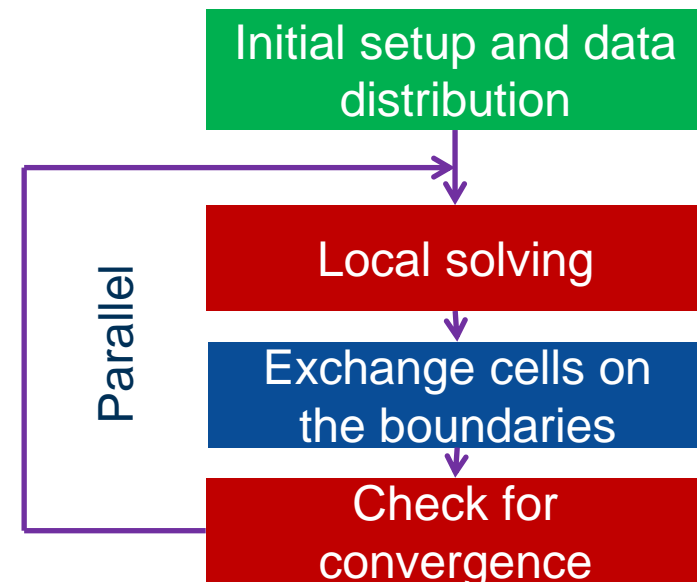


SPMD:

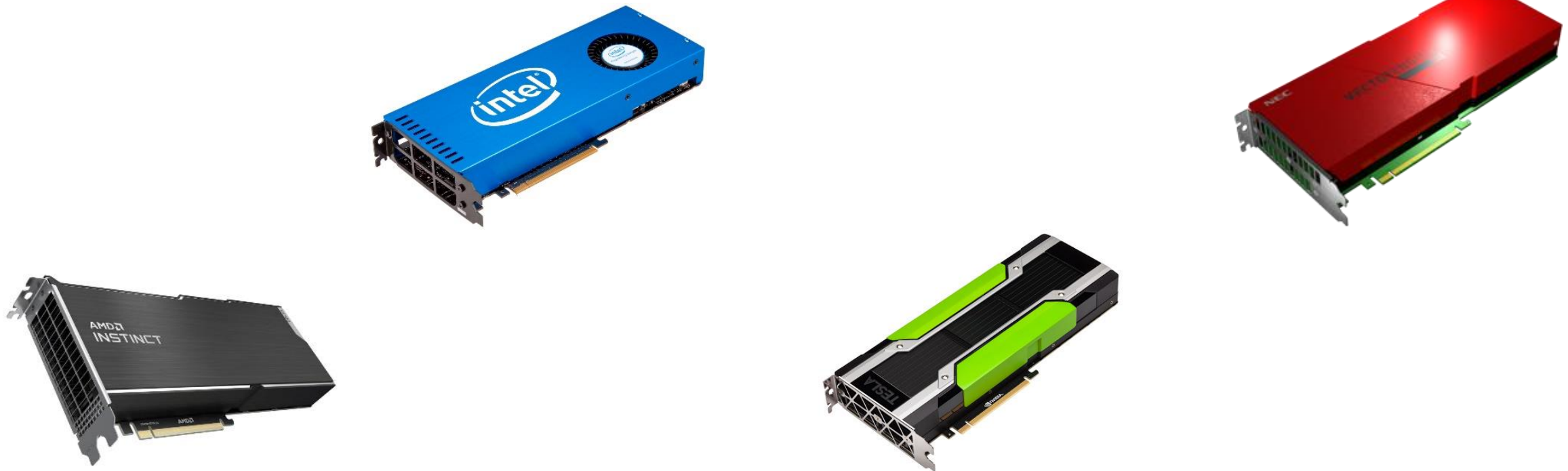
Single Program  
Multiple Data

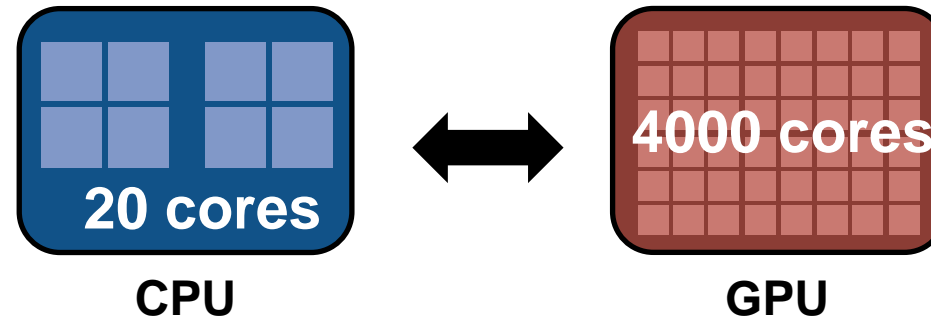


- Example: Domain decomposition in CFD: Mapping of 3D mesh to the processors
- Programming techniques
  - Data parallel approach
    - Distribute data structures
  - Parallel algorithms
  - Explicit data exchange (MPI)



# What is an Accelerator in a Node?



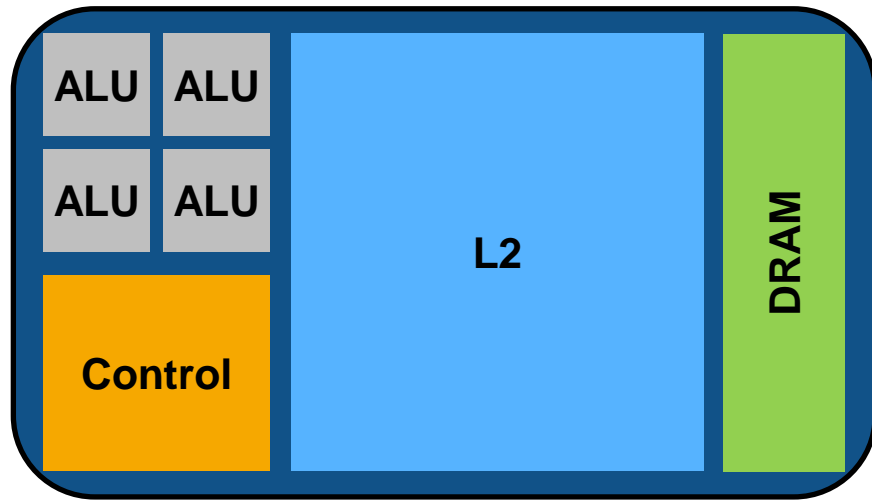


- GPU-Threads
  - Thousands (“few” on CPU)
  - Light-weight, little creation overhead
  - Fast switching
- Lots of parallelism needed on GPU to get good performance!

# 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**

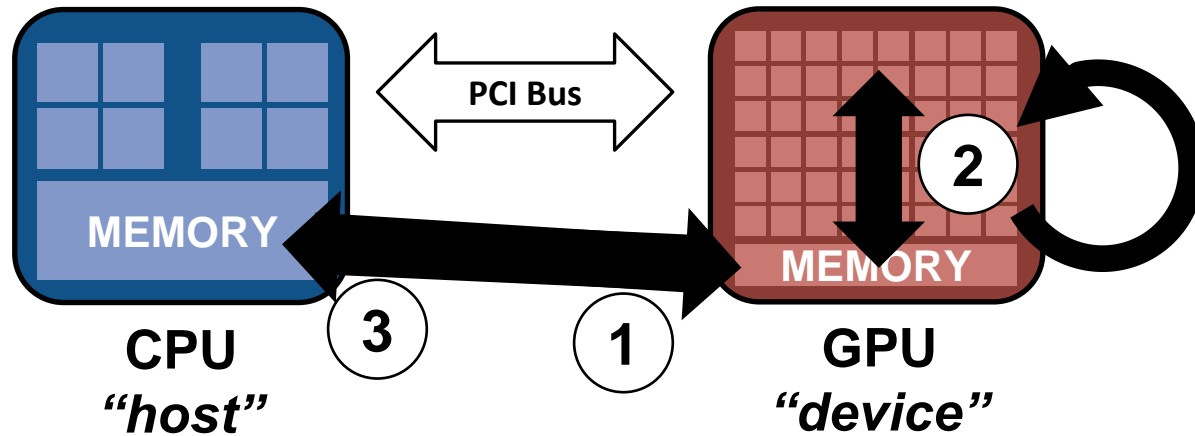


### GPU

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



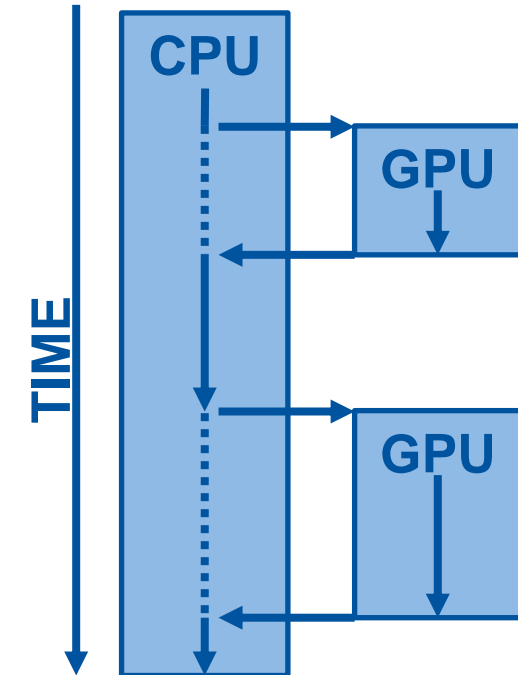
# Offloading

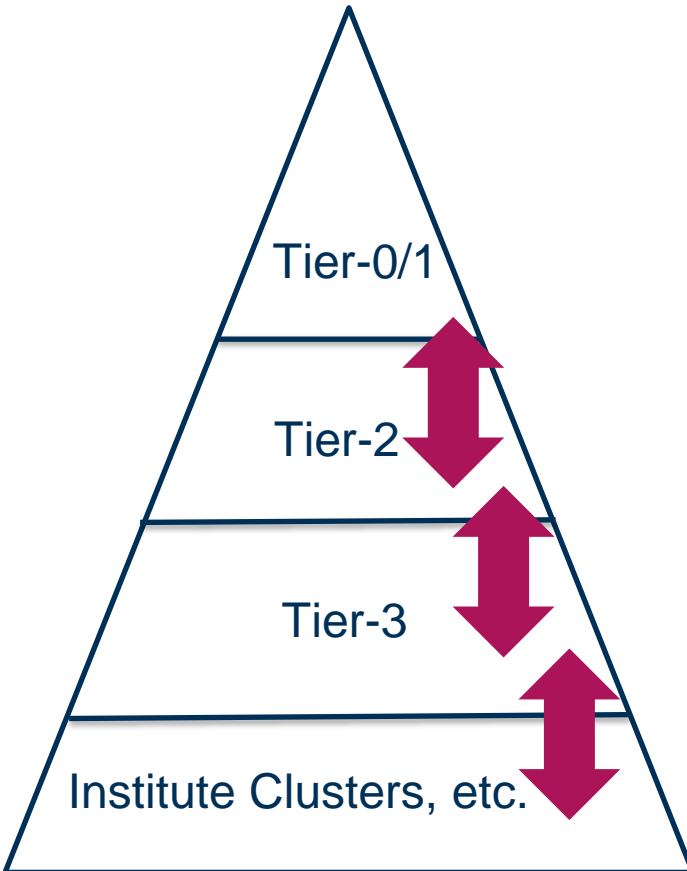


We refer to “discrete GPUs” here.

- 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)*





## 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

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	<b>412 + 220 HPC nodes</b> 2-socket Intel Sapphire Rapids (Xeon 8468, 2x48 cores, 2.1 GHz) <ul style="list-style-type: none"> <li>• 470 nodes with 256 GB</li> <li>• 160 nodes with 512 GB</li> <li>• 2 nodes with 1024 GB</li> </ul>
ML Segment	<b>32 + 5 ML nodes</b> 2-socket Intel Sapphire Rapids (Xeon 8468, 2x48 cores, 2.1 GHz, 256 GB) <b>4x NVIDIA H100 96 GB HBM2e per node</b>
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)

Additional WestAI resources:  
 15 ML nodes

# Questions?