aiXcelerate 2024

CLAIX and Multi-GPU Setup

NHR4CES NHR for Computational Engineering Science | HPC.NRW | EDIH Rheinland | WEST AI KI-Servicezentrum | IT Center | RWTH AACHEN UNIVERSITY

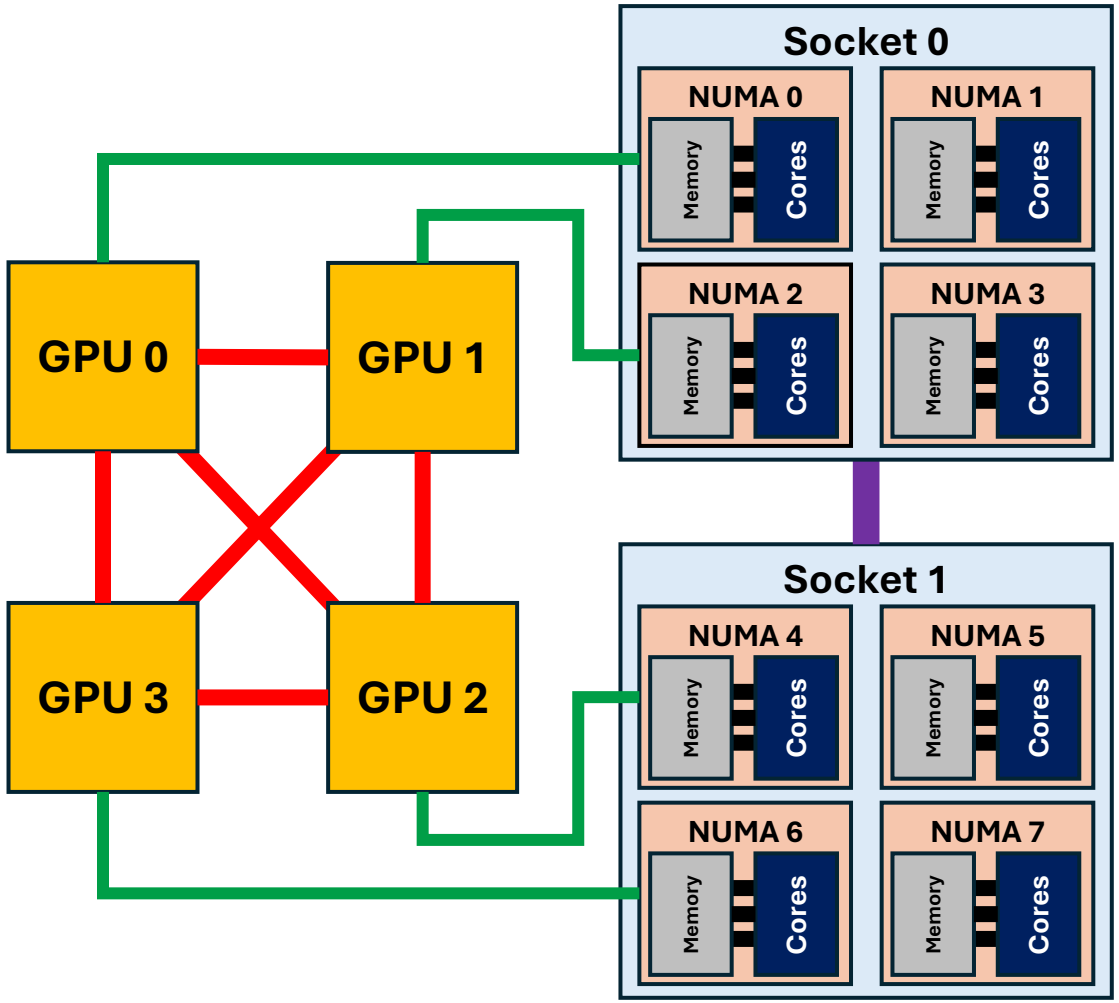
CLAIX-2023

CLAIX-2023 – Overview & Segments

CLAIX-2023 (Tier-2 + Tier-3)	
Segment: HPC	**632 HPC nodes (412 Tier-2 + 220 Tier-3)** 2-socket Intel Sapphire Rapids • 48 cores per socket @ 2.1 GHz • 1.5 TB local SSD • 470 nodes with 256 GB memory (c23ms) • 160 nodes with 512 GB memory (c23mm) • 2 nodes with 1024 GB memory (c23ml)
Segment: ML	**52 GPU nodes (32 Tier-2 + 5 Tier-3 + 15 WestAI)** 2-socket Intel Sapphire Rapids • 48 cores per socket @ 2.1 GHz • **4x Nvidia H100 94 GB HBM2e** • 695 GB local SSD • 512 GB memory
Segment: Interactive	Dedicated nodes with smaller GPUs (JupyterHub usage)
Fabric / Network	Infiniband NDR network
Storage	26 PiB Lustre (Parallel File System) GPFS ($HOME and $WORK) BeeOND: on-demand on local SSDs

CLAIX-2023 – GPU System Overview

- **2x Intel Xeon 8468 Sapphire Rapid sockets**
  - 48 cores per socket → 96 cores total
  - Sub-NUMA clustering enabled
- **Memory: 512 GB memory**
- **4x NVIDIA H100 GPUs per node**
  - Connected to NUMA domains 0, 2, 4, 6
  - 2 GPUs per socket
- **Billing for GPU jobs**
  - Case: 1 GPU
    - You get 25% resources (cores and memory)
    - 1 GPU-h = 24 core-h
  - Case: 2 GPUs
    - You get 50% resources (cores and memory)
    - 2 GPU-h = 48 core-h

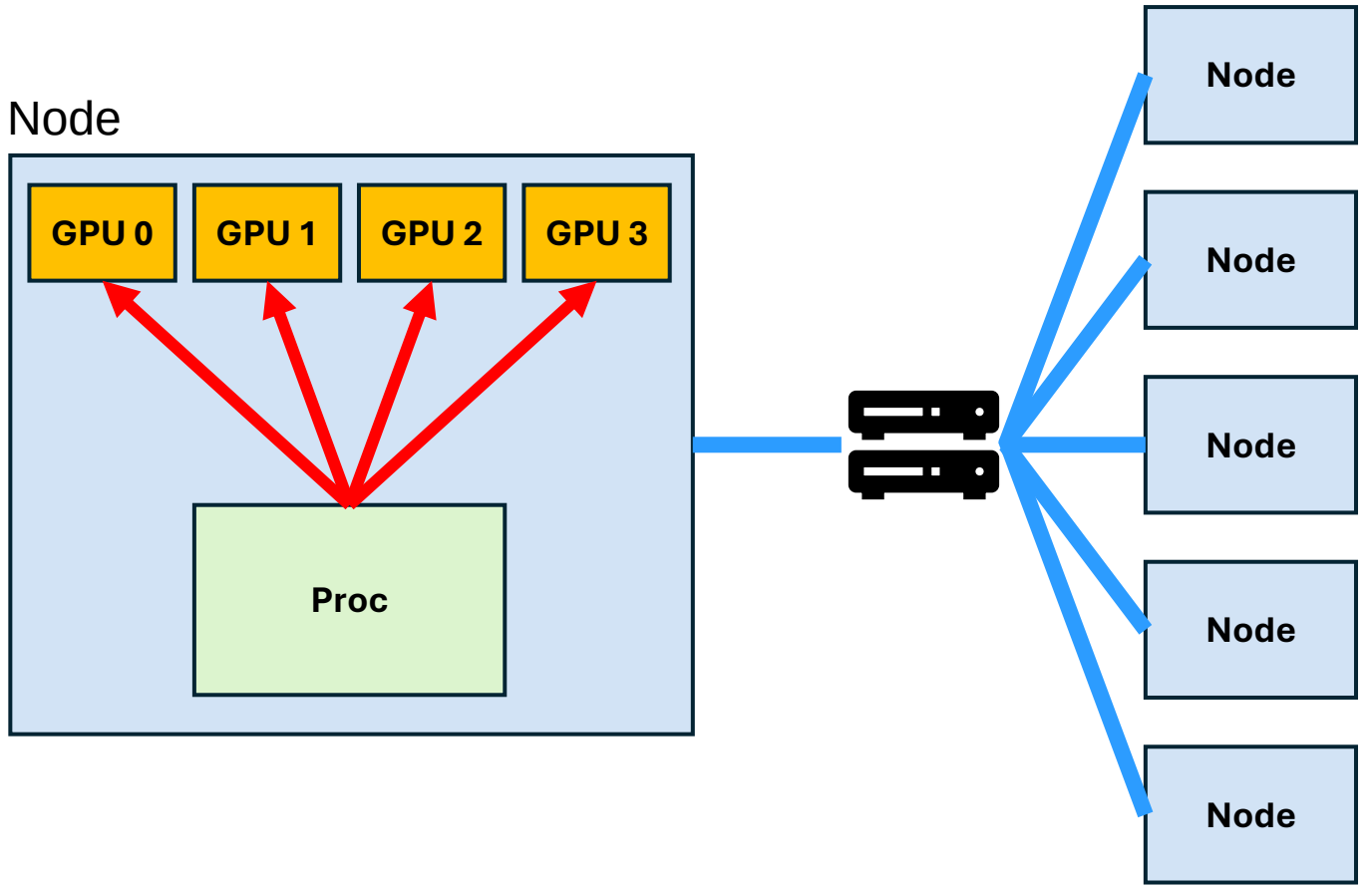


CLAIX-2023 – Use cases with single or multiple GPUs?

- **Classical use cases**
  - MPI + CUDA (not in focus)
  - MPI + OpenMP target (not in focus)
- **AI / ML / DL use cases**

PyTorch

TensorFlow

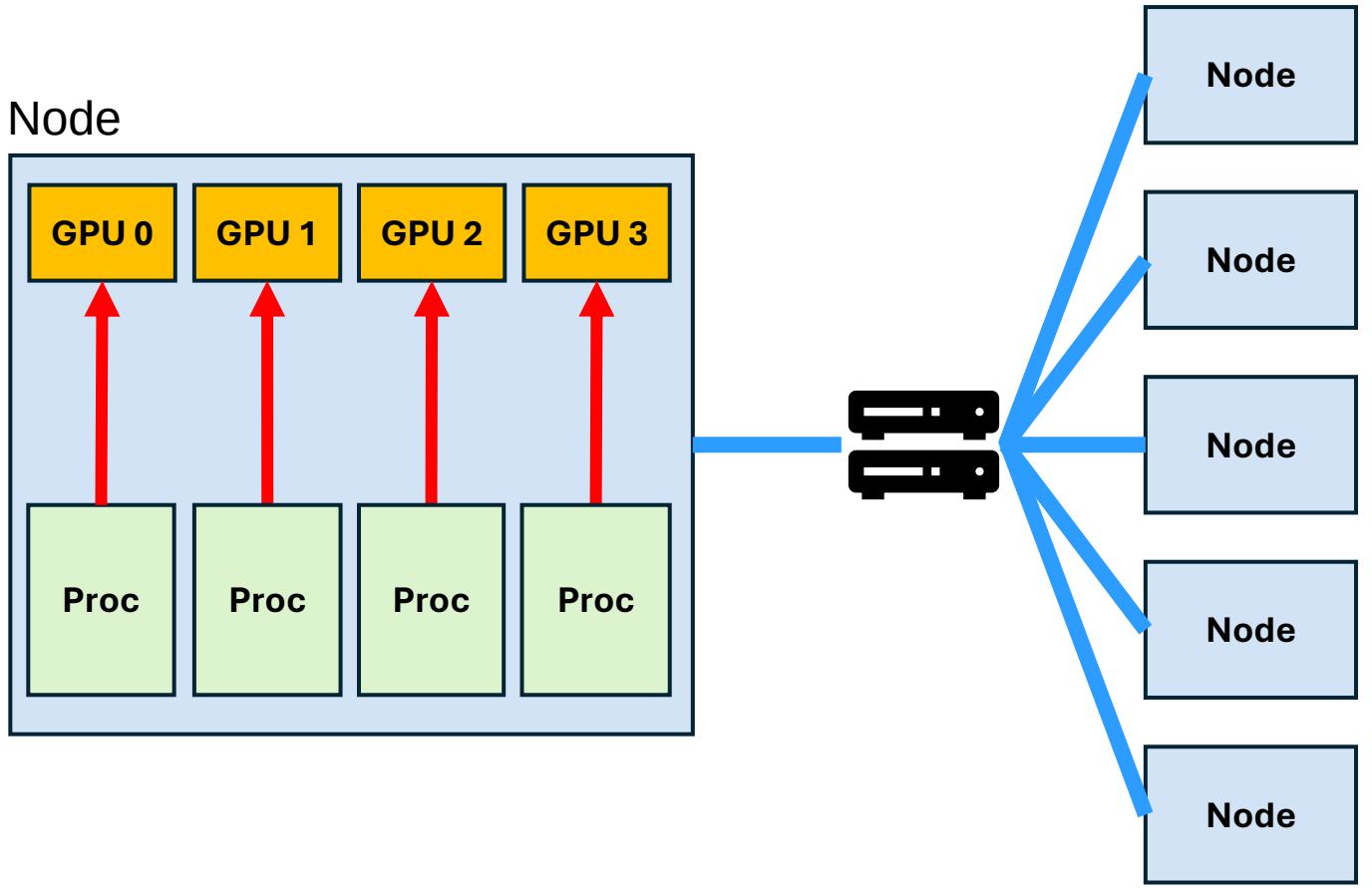


CLAIX-2023 – Use cases with single or multiple GPUs?

- **Classical use cases**
  - MPI + CUDA (not in focus)
  - MPI + OpenMP target (not in focus)
- **AI / ML / DL use cases**

PyTorch

TensorFlow



CLAIX-2023 – Allocation Examples using Slurm

- **Single GPU Job (1 Node; 1 Process)**

```
#!/usr/bin/zsh
#################################################################
### Slurm flags
#################################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1       → (MPI) processes
#SBATCH --cpus-per-task=24        → cores per process
#SBATCH --gres=gpu:1              → GPUs per node

#################################################################
### Modules, Execution
#################################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 1 GPU → 25 %
  - 24 cores → 25 %
  - ~128 GB memory → 25 %
- Billing: 1 hour = 24 core-h
  - Even if cores not fully used or less allocated

CLAIX-2023 – Allocation Examples using Slurm

- **Single GPU Job (1 Node; 1 Process)**

```
#!/usr/bin/zsh
#############################################################
### Slurm flags
#############################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1       → (MPI) processes
#SBATCH --cpus-per-task=24        → cores per process
#SBATCH --gpus-per-node=1         → GPUs per node

#############################################################
### Modules, Execution
#############################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 1 GPU → 25 %
  - 24 cores → 25 %
  - ~128 GB memory → 25 %
- Billing: 1 hour = 24 core-h
  - Even if cores not fully used or less allocated

CLAIX-2023 – Allocation Examples using Slurm

- **Multi-GPU Job (1 Node; 1 Process)**

```
#!/usr/bin/zsh
#############################################################
### Slurm flags
#############################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1       → (MPI) processes
#SBATCH --cpus-per-task=48        → cores per process
#SBATCH --gpus-per-node=2         → GPUs per node

#############################################################
### Modules, Execution
#############################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 2 GPU → 50 %
  - 48 cores → 50 %
  - ~256 GB memory → 50 %
- Billing: 1 hour = 48 core-h
  - Even if cores not fully used or less allocated

CLAIX-2023 – Allocation Examples using Slurm

- **Multi-GPU Job (1 Node; 2 Processes)**

```
#!/usr/bin/zsh
#################################################################
### Slurm flags
#################################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=2       → (MPI) processes
#SBATCH --cpus-per-task=24        → cores per process
#SBATCH --gpus-per-node=2         → GPUs per node

#################################################################
### Modules, Execution
#################################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 2 GPU → 50 %
  - 48 cores → 50 %
  - ~256 GB memory → 50 %
- Billing: 1 hour = 48 core-h
  - Even if cores not fully used or less allocated

CLAIX-2023 – Allocation Examples using Slurm

- **Multi-GPU Job (2 Nodes, 2 Processes)**

```
#!/usr/bin/zsh
#############################################################
### Slurm flags
#############################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=1       → (MPI) processes
#SBATCH --cpus-per-task=96        → cores per process
#SBATCH --gpus-per-node=4         → GPUs per node

#############################################################
### Modules, Execution
#############################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 8 GPU (2x4) → 100 %
  - 192 cores → 100 %
  - ~1024 GB memory → 100 %
- Billing: 1 hour = 192 core-h
  - Even if cores not fully used or less allocated

CLAIX-2023 – Allocation Examples using Slurm

- **Multi-GPU Job (2 Nodes, 8 Processes)**

```
#!/usr/bin/zsh
#################################################################
### Slurm flags
#################################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=4       → (MPI) processes
#SBATCH --cpus-per-task=24        → cores per process
#SBATCH --gpus-per-node=4         → GPUs per node

#################################################################
### Modules, Execution
#################################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 8 GPU (2x4) → 100 %
  - 192 cores → 100 %
  - ~1024 GB memory → 100 %
- Billing: 1 hour = 192 core-h
  - Even if cores not fully used or less allocated

CLAIX-2023 – Allocation Examples using Slurm

- **Special Case: Single GPU but more memory required**

```
#!/usr/bin/zsh
#################################################################
### Slurm flags
#################################################################
#SBATCH --time=00:15:00
#SBATCH --partition=c23g
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1      → (MPI) processes
#SBATCH --cpus-per-task=24       → cores per process
#SBATCH --gpus-per-node=1        → GPUs per node
#SBATCH --mem=256G               → Memory per node
#################################################################
### Modules, Execution
#################################################################
...
```

- Submit the job to Slurm with

```
$ sbatch myjob.sh
```

- Assigned resources
  - 1 GPU → 25 %
  - 24 cores → 25 %
  - ~256 GB memory → 50 %
- Billing: 1 hour = 48 core-h
  - Even if cores not fully used or less allocated
  - You now exceed the 25% resources

Thank you!