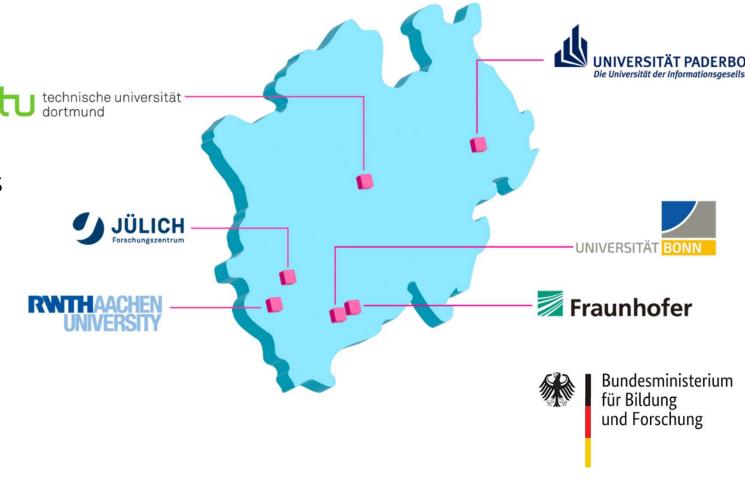


WestAl Service Center

John Arnold – RWTH Aachen University contact@westai.de westai.de

Goal and Partners of WestAl

Support researchers and industry partners in adapting AI/ML methods into their workflows by providing consulting services, trainings, and computing resources.





AI-Consultation



Entry Consultation

- Analysis on the feasibility of your AI idea
- Overview of current research trends
- Estimation of the required computational effort

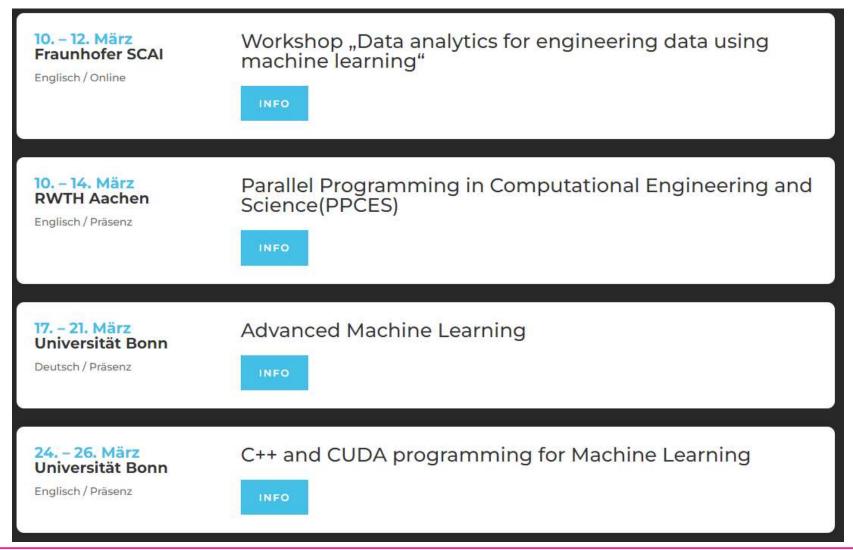


Research Cooperations

- Support in the development and optimization of AI prototypes
- Access to scientific evaluation
- Collaboration with renowned research institutions



Trainings





Al Research



Scalable data management and integration of heterogeneous data

- Generation of datasets for training large models
- Development and application of methods for handling sensitive data



Transfer of pre-trained models to different domains

- Model transfer to diverse domains and tasks (language, image, audio, video, etc.)
- Investigation of factors influencing transfer behavior



Large-scale model training

- Training of generic models transferable to different domains and tasks
- Deriving scaling laws



Transfer of models to applications with limited resources

- Model compression
- Transfer of models to resource-constrained use cases



Jureca (FZJ) / CLAIX-2023 (RWTH)





High Performance Computing

Memory

CPU

GPU







Hardware

RWTHAACHEN UNIVERSITY

15 Compute Nodes[†] 2 × Xeon 8468, 512 GB DDR5 4 × H100 GPUs*

Part of the 52 Nodes of CLAIX-2023



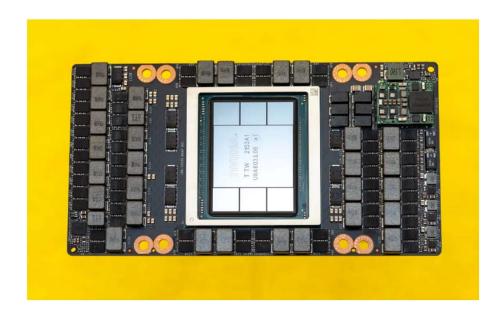
16 Compute Nodes[†] 2 × Xeon 8468, 512 GB DDR5 4 × H100 GPUs*

Isolated: Exclusive for WestAI

†Infiniband Network *Full NVLINK, 96 GB HBM2e



High Performance Computing



H100: 67 TFLOPS (FP32) 96 GB HBM2e



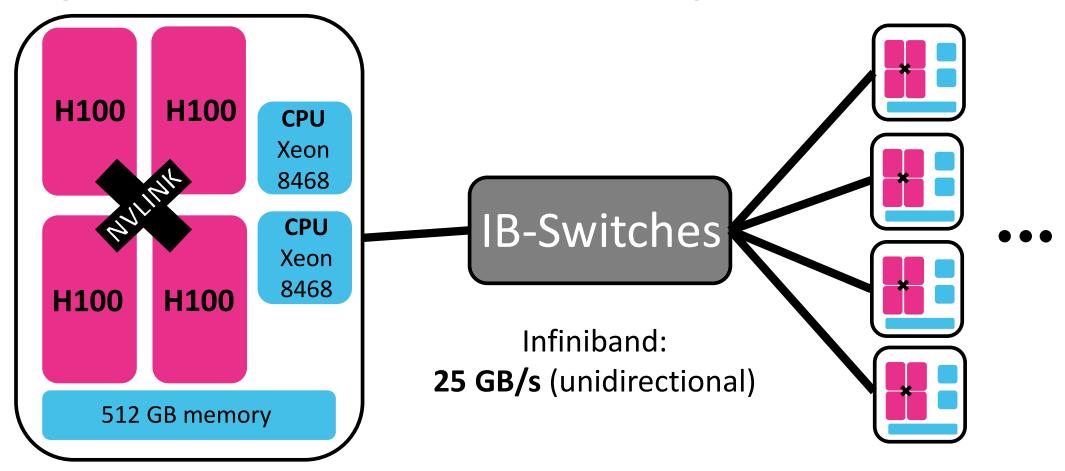
RTX 5090: 105 TFLOPS (FP32) 32 GB memory



https://www.techpowerup.com/gpu-specs/geforce-rtx-5090.c4216 https://www.cnet.com/pictures/see-nvidias-h100-hopper-chip-up-close/

High Performance Computing

Other GPU-Nodes



Performance results from parallelism



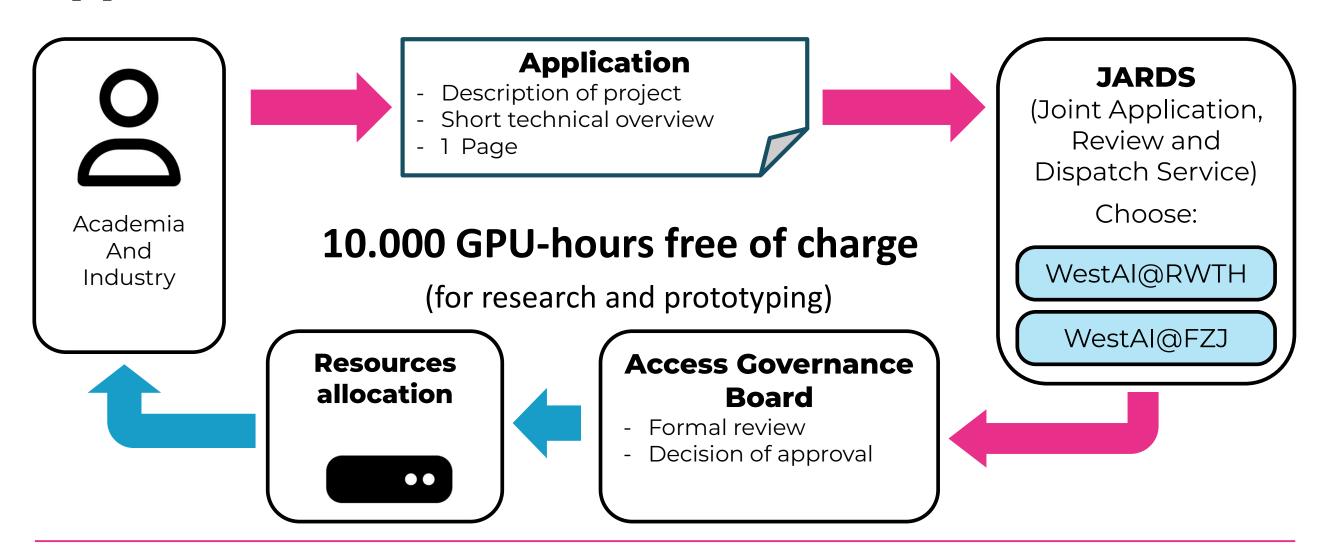
Hardware Offer

10.000 GPU-hours free of charge

(for research and prototyping)



Application Process





Discussion

- What work are you doing currently?
- What hardware resources are you using?
- What are limiting factors?



Access and Software @ RWTH



Batch Mode

Login via SSH SLURM

Apptainer (Singularity) Module System



Jupyter Hub / Lab Web Interface

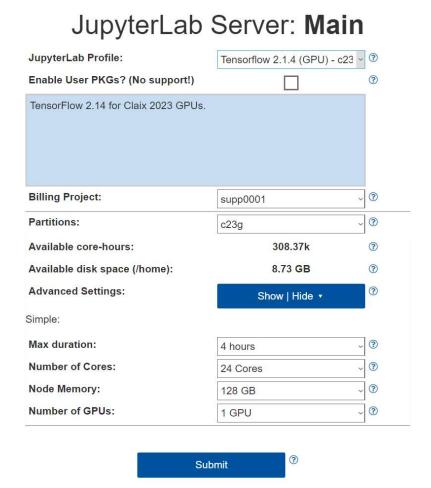
Predefined Containers





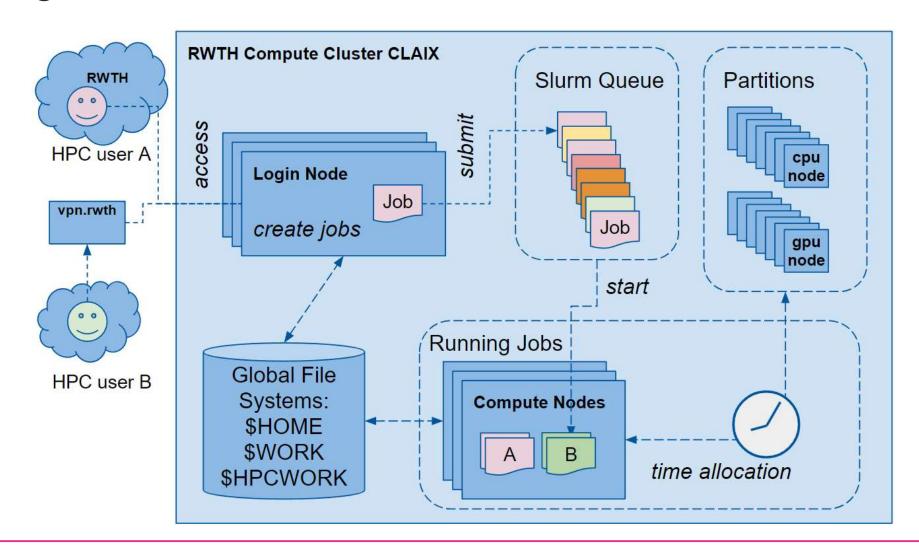
JupyterHub @ RWTH

- Easy access to HPC-Resources via the web browser
- Bring your own Jupyter Notebooks and start working right away
- All usual ML-Frameworks are available (Pytorch, Tensorflow, Keras, Scikit-Learn)





Batch-System





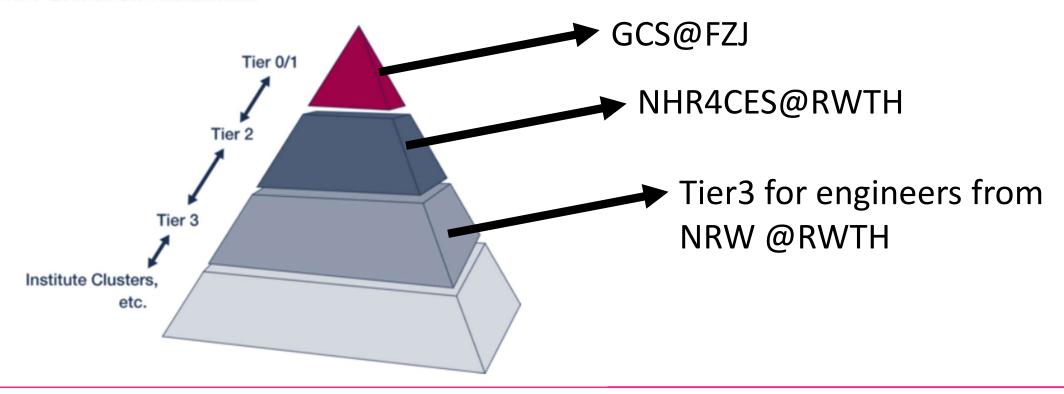
Al Service Centers





Other Initiatives

- Tier-0 PRACE und EuroHPC
- Tier-1 Gauss Center for Supercomputing (GCS, JSC, HLRS, LRZ)
- Tier-2 Verbund für Nationales Hochleitungsrechnen (NHR) und weitere Tier-2 Zentren
- Tier-3 lokale HPC-Zentren der Hochschulen





https://hpc.dh.nrw/en/hpc-access/hpc-performance-pyramid

Additional Resources

- https://westai.de
- https://www.hpc.itc.rwth-aachen.de/
- https://help.itc.rwth-aachen.de
- https://hpc-wiki.info
- https://www.nhr4ces.de/hpc-access/
- https://www.fz-juelich.de/en/ias/jsc/systems/supercomputers/applyfor-computing-time





Thanks for listening!

If you have any questions:
contact@westai.de
westai.de