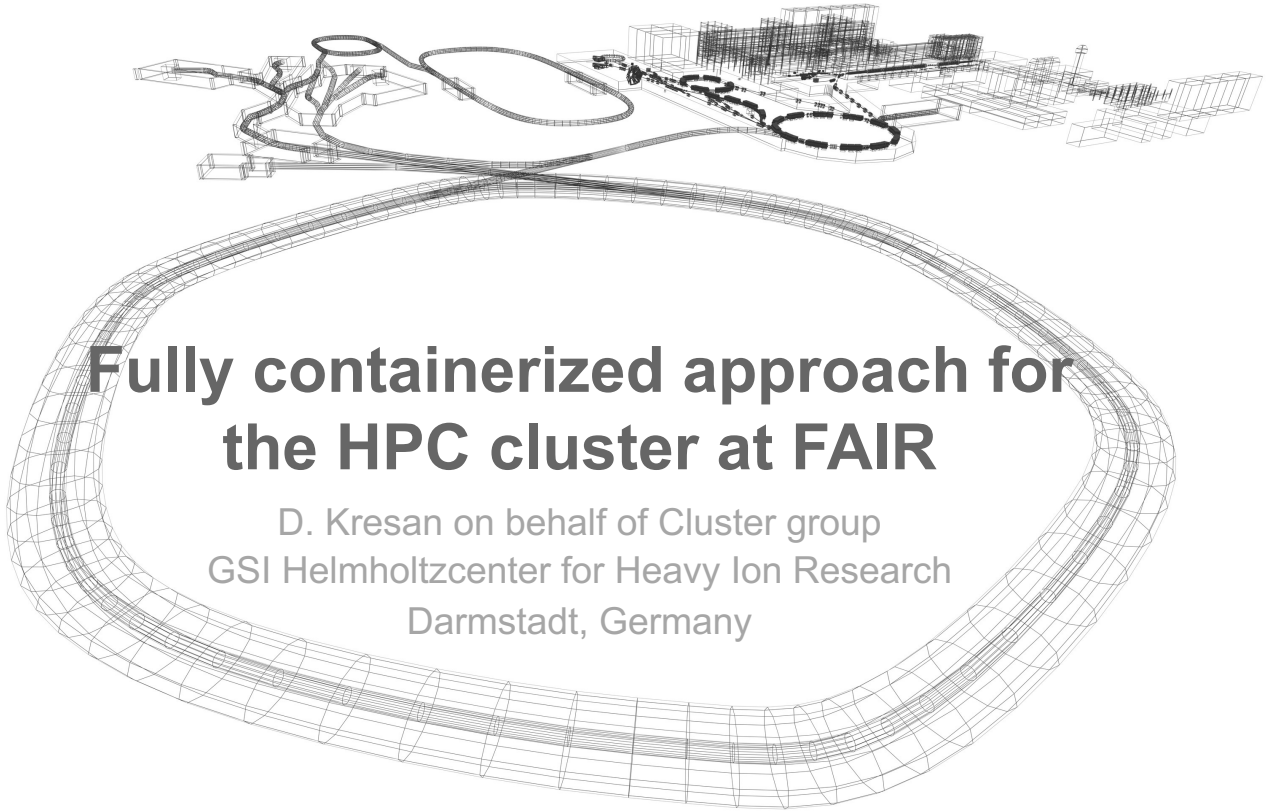


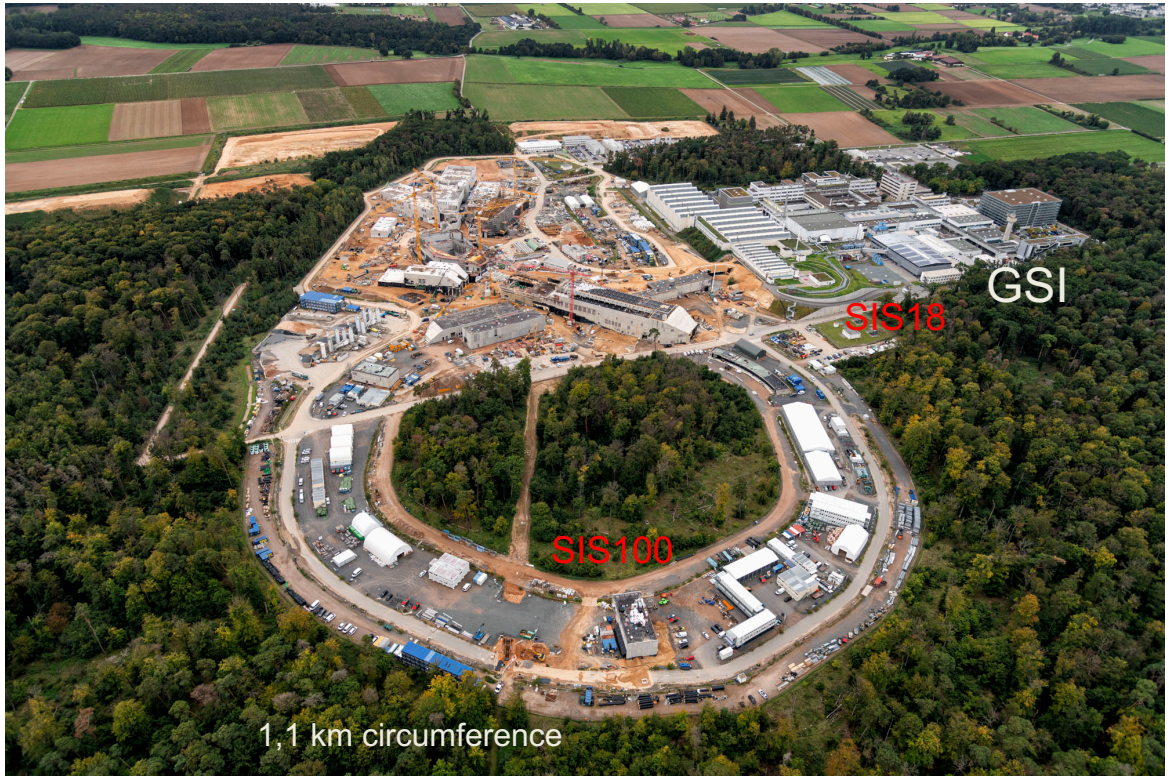
Fully containerized approach for the HPC cluster at FAIR

D. Kresan on behalf of Cluster group
GSI Helmholtzcenter for Heavy Ion Research
Darmstadt, Germany

Facility for Antiproton and Ion Research in Europe

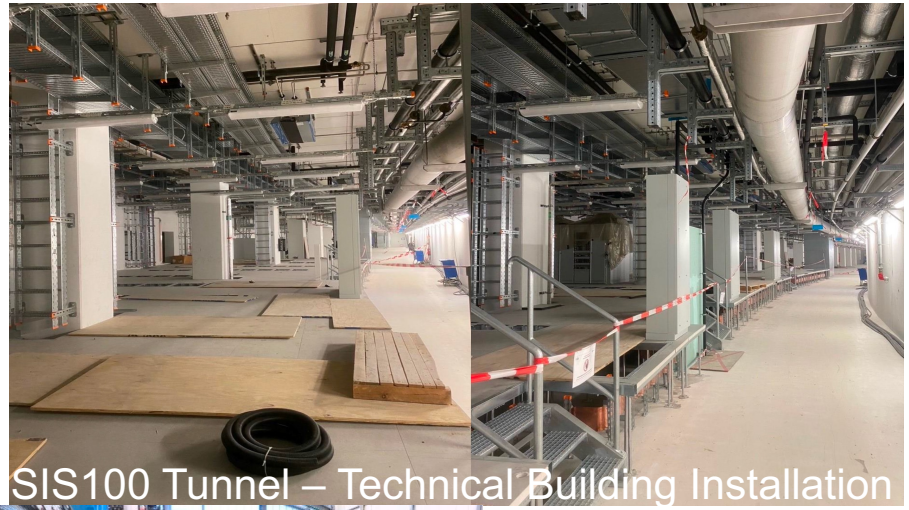
3000 scientists from
50 countries

First experiments
expected in 2028

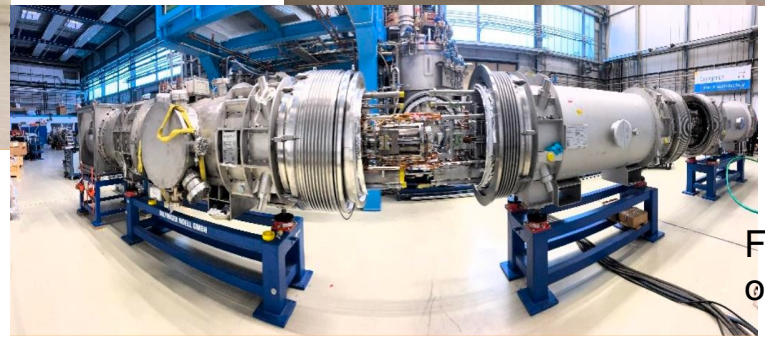


Highlights from FAIR Construction Site – installation started 2024

First power supply unit



SIS100 Tunnel – Technical Building Installation

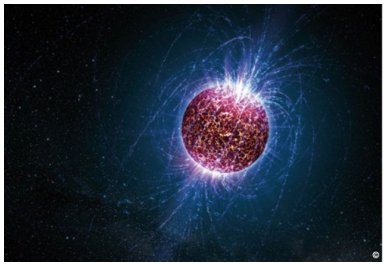


First thermal cycle
of the SIS100 string

4 Scientific Pillars of FAIR

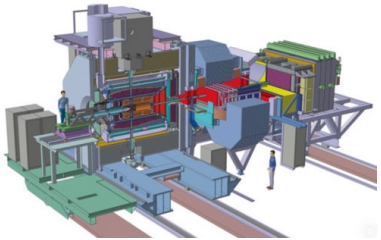
NUSTAR

Nuclear Structure Astrophysics and Reactions. Experiments with atomic nuclei are the key to understanding stars



CBM

Compressed Baryonic Matter. The collision of atomic nuclei at high speeds can simulate the conditions inside supermassive objects for a split second



PANDA

Antiproton Annihilation at Darmstadt. How can antimatter help us understand the mass of matter and the strong force?

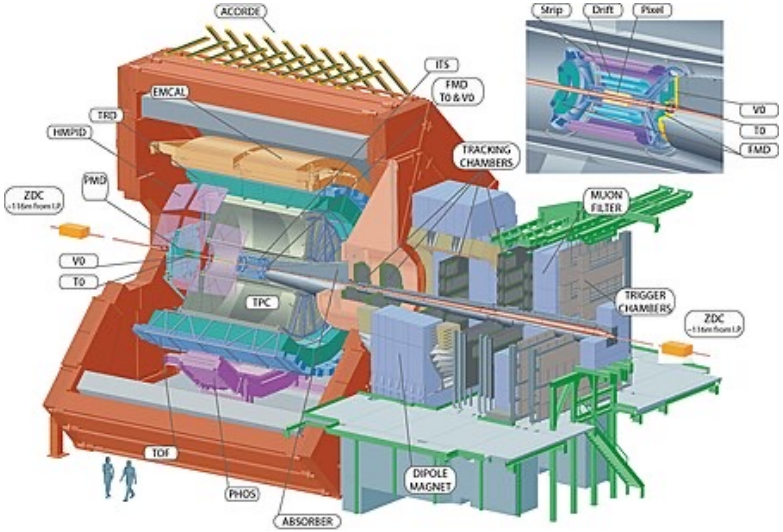
APPA

Atomic, Plasma Physics and Applications. From the investigation of atoms and macroscopic effects in materials or tissues all the way to engineering and medical applications

Computing for ALICE Experiment

ALICE

Analysis Facility



WLCG

Worldwide LHC Computing Grid

Computing Resource Requirements

CPU

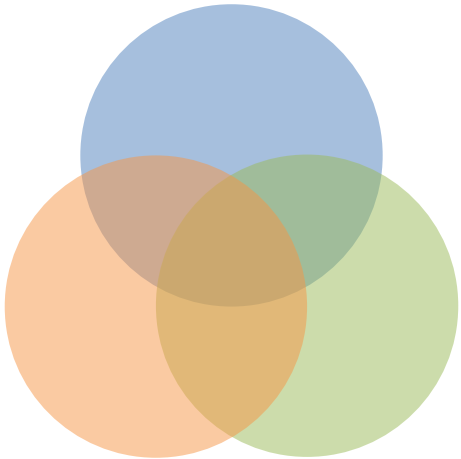
	NUSTAR	CBM	PANDA	APPA
Number of cores (a)	9 k	45 k	68 k	11 k
Number of cores (b)	7 k	45 k	34 k	-

(a) Resources for simulations
(b) Resources for online data reconstruction

Storage

	NUSTAR	CBM	PANDA	APPA
Disk total (TB)	34.250	103.000	60.680	7.037

Sharing of resources



- No dedicated / fixed hardware for an experiment
- Will not take beam all at the same time
- Computing resources will be shared dynamically

Computing Use Cases at GSI / FAIR

- Heavy-Ion Collisions and Hadron Physics:
  - Online event processing, HTC, MC Simulations
- Nuclear Physics:
  - mostly offline batch processing
- Bio-physics, Med-physics:
  - MC simulations, AI with GPUs
- Plasma physics:
  - large scale MPI
- Theory:
  - large scale MPI, memory-bound

- All 6 communities need different software tools

HPC System for FAIR needs to …

- … scale with number of software tools
- … scale with number of workflows
- … scale with number of nodes
- … be easy to operate and upgrade

Fully Containerized Approach

- Separate user application space from host system
- Jobs are executed in a container
- **Minimal host system**: HW drivers + Slurm and Apptainer

Advantages

- Users are free to choose Linux flavor and install any required software
  - Flexibility in supporting different use cases and workflows
- Admins are free to upgrade host OS and/or Slurm at any time
  - Makes Virgo cluster more scalable

Bare Metal Submitter Node

- `ssh virgo.hpc`
- Submit job in container

Virtual Application Environment (VAE)

- Ready-to-use solution provided by GSI-IT

- `ssh vae23.hpc`
- Login into container - interactive session on submitter node
  - Edit, compile, test, debug

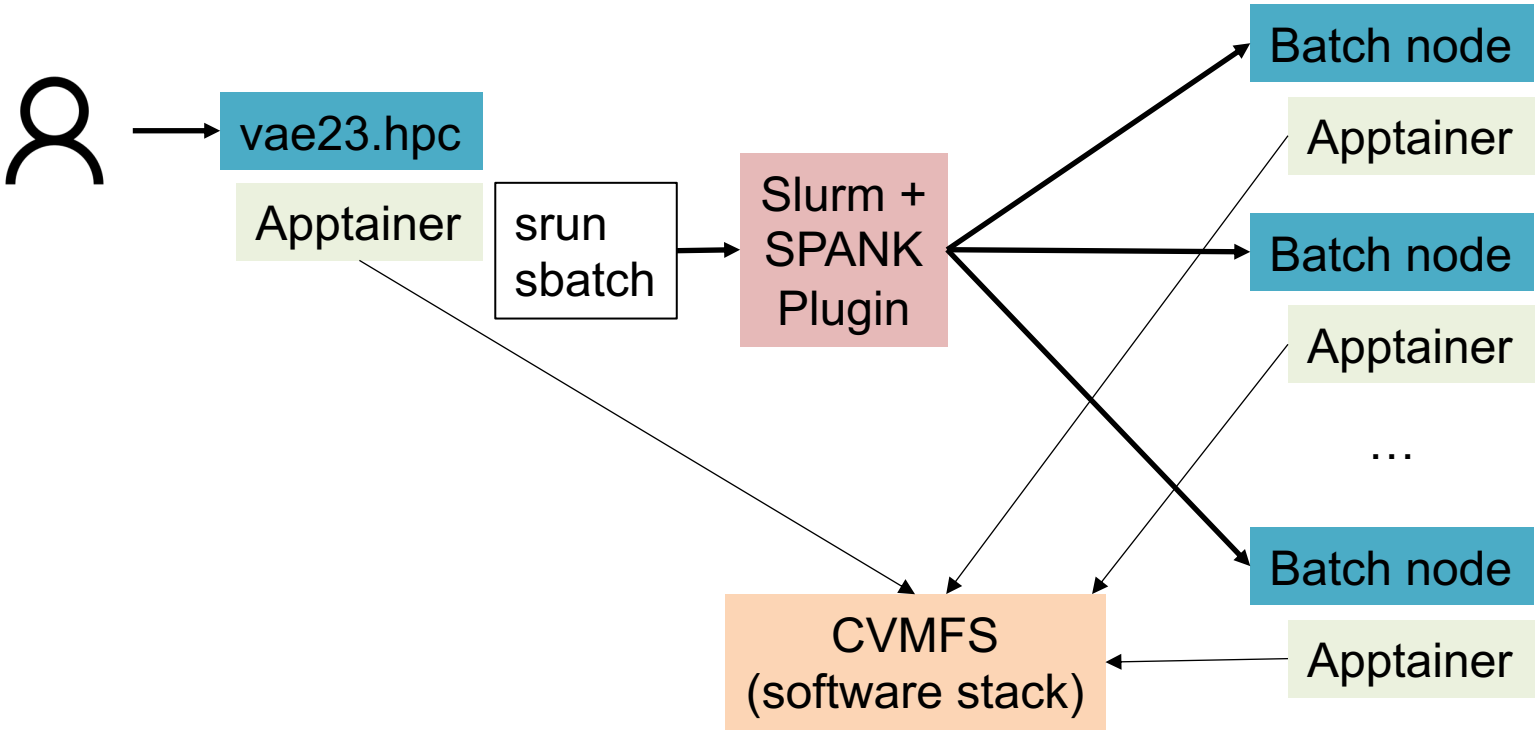
- Submit job which will run in the same VAE

- Fully transparent to a user and easy to use
  - SPANK plugin for Slurm starts container in the background

Software for VAE

- Large and diverse community
  - Many software packages required in VAE
- Issues with container size

- Solution: software stack for VAE is installed on CVMFS and mounted during run-time
- Container size: ~ 1 GB


vae23.hpc
Apptainer
srun
sbatch
Slurm +
SPANK
Plugin
Batch node
Apptainer
Batch node
Apptainer
…
Batch node
Apptainer
CVMFS
(software stack)

Package Manager

- ~ 700 software packages available in the current version of VAE
- Use Spack to build and install them: https://github.com/spack/spack

- Package manager developed for HPC systems
- Handling of dependencies
- Support of
  - Multi architectures
  - Multi compilers
  - Multi versions
  - Mixed toolchains

CVMFS

- Key service for the cluster
- Redundant setup (HW based) for stable operation

- (nearly) Every group at GSI/FAIR has own CVMFS repo

- GSI-IT provides solution for ready-to-use publisher

OpenMPI in Apptainer

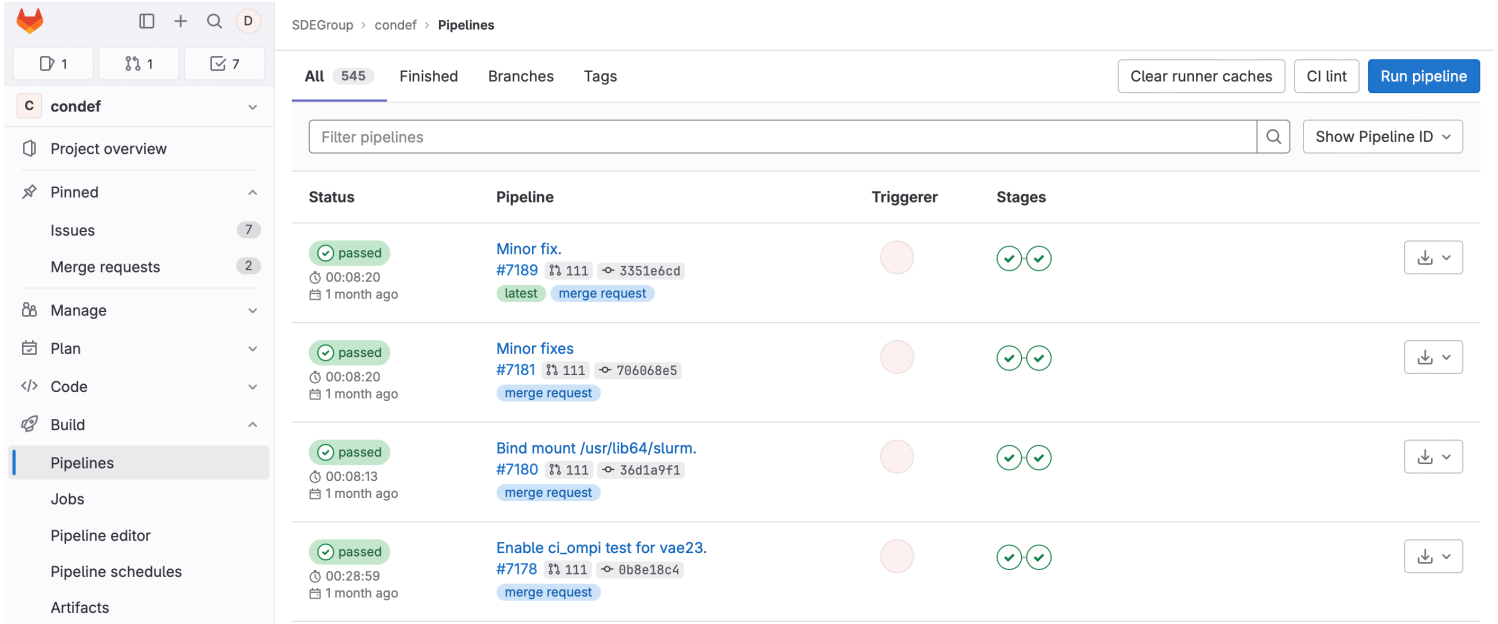
- New way to use OpenMPI
- Alternative to standard bind or hybrid mode

- No OpenMPI installed on the host, only in Apptainer

- Slurm + pmix is used to launch the calculation

Building of Container Images

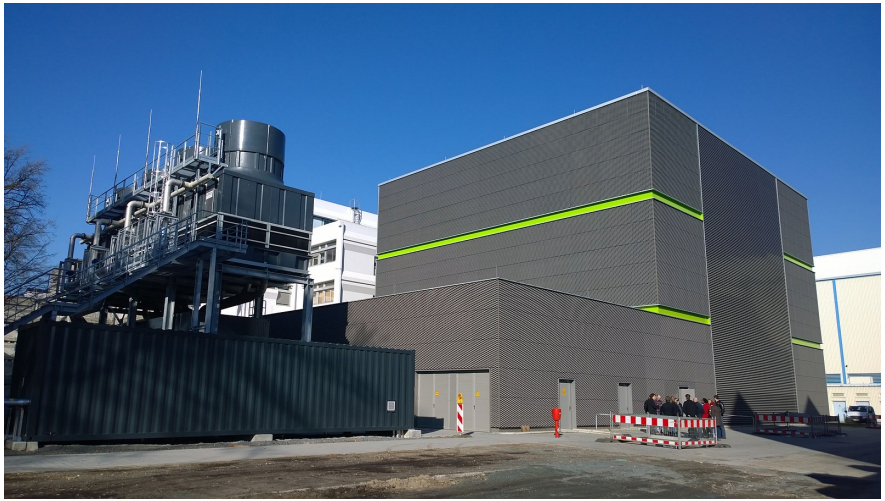
- Definition files are in a git repository
- CI workflow for building, testing and deployment



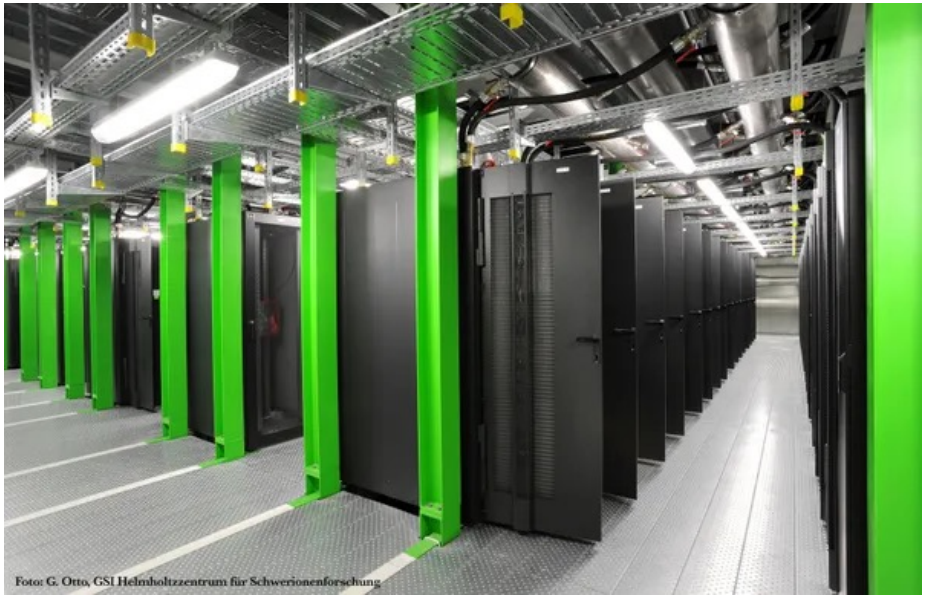
Testing

- Tool for benchmarking of network and I/O performance as MPI job
- Helped to discover and fix network issues in the past
- Integrated into CI / CD

Green IT Cube



In operation since 2016



PUE < 1,07
4 MW cooling
Capacity for 768 racks on 6 floors

Summary

- 3 years of stable operation with containers
- ~700 users on ~70.000 CPUs and 400 GPUs
- Multiple host OS upgrades performed
- Slurm upgrades performed. Current version: 21-08. Now rolling out 23-11!
- HPC Cluster for future FAIR, based on fully containerized approach, is scalable

Thank you!

d.kresan@gsi.de