Leibniz Supercomputing Centre - DAOS Site Update

László Szűcs, Patrick Böhl

Leibniz Supercomputing Centre

of the Bavarian Academy of Sciences and Humanities

DATA STORAGE
MANAGE RESEARCH DATA PROFESSIONALLY

RESEARCH
DEVELOPING THE IT SERVICES FOR FUTURE SCIENCE

SUPERCOMPUTING
WORLD-CLASS MACHINES FOR EXCELLENT RESEARCH

IT SECURITY
PROTECT DATA AND STORE IT SECURELY

FUTURE COMPUTING WITH QUANTUM AND AI
ACCELERATING TIME TO WORLD-CLASS SCIENCE

VIRTUAL REALITY AND VISUALISATION
COMMUNICATING RESEARCH TO ALL SENSES

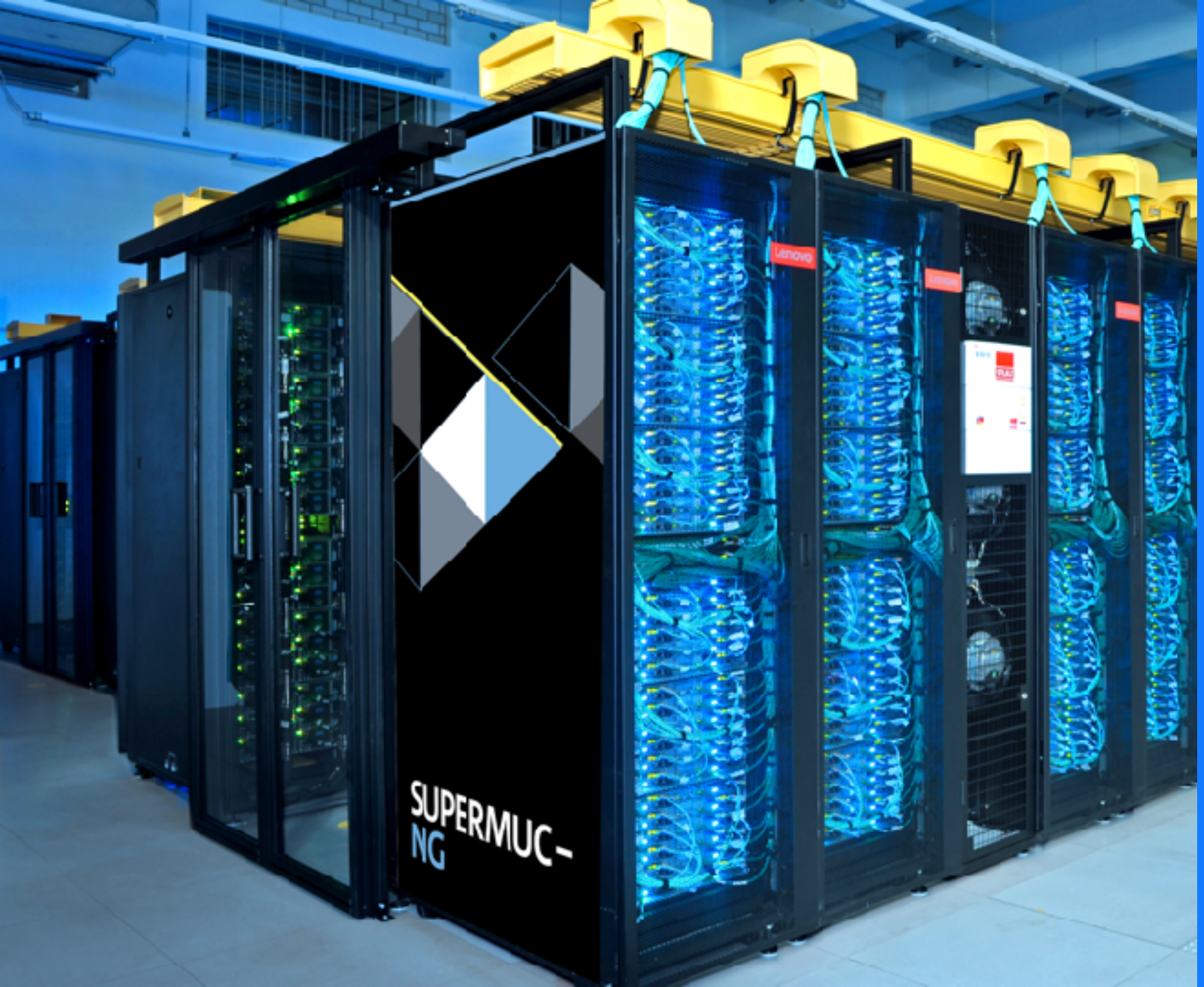
Context: HPC at LRZ

- Part of German GCS (Gauss Centre for Supercomputing)
  - Compute proposals to be granted
  - German / European
- Academics in Munich area

~ 2.000
Researchers

100s of projects
- Astrophysics, Particle Physics
- Chemistry / Material Science
- Comp. Fluid Dynamics / Eng.
- Environmental / Life Sciences

LRZ user base is diverse
- **codes often developed by users**
- **porting/tuning is not contributing to science, so often not in focus**



SuperMUC-NG

Top500 - Nov 2018: #8
(June 2023: #31)

Lenovo Intel

311,040 cores
Intel Xeon Skylake

26.9 PetaFlops Peak

19.5 PetaFlops Linpack*

719 TeraByte Main Memory

70 PetaByte Disk

SuperMUC-NG

Top500 - Nov 2018: #8
(June 2023: #31)

Lenovo Intel

311,040 cores
Intel Xeon Skylake

26.9 PetaFlops Peak

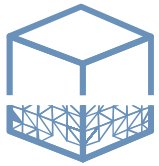
19.5 PetaFlops Linpack*

719 TeraByte Main Memory

70 PetaByte Disk

SuperMUC-NG Phase 2: Objectives

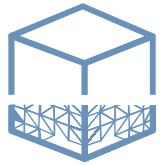
simulation and modelling



high performance artificial intelligence (HPAI)

high performance data analytics (HPDA)

integration of AI methods in HPC workflows



SuperMUC-NG Phase 2: Configuration

Accelerated node architecture

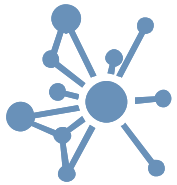
- 2x Intel® Xeon® Platinum 8480L, 56 cores
- 4x Intel® Data Center GPU Max 1550
- 512 GB DDR5 main memory
- Lenovo's SD650-I v3 platform

Distributed asynchronous object storage (DAOS)

- 1 PB NVMe capacity
- 78 TiB persistent memory
- > 750 GB/s write bandwidth

High speed interconnect

- 2x Mellanox HDR Infiniband HDR 200 Gb/s
- fat tree topology
- two uplinks per node
- separated from Phase 1



Integration

- Phase 1 accounts and HOME directories
- Phase 1 WORK and SCRATCH filesystems
- DSS volumes available
- direct warm water cooling

Joint Capacity

	Phase1 (general purpose)	Phase 2 (accelerated)
Number of compute nodes	6480	240
Main memory	719 TB DDR4	123 TB DDR5
Storage capacity	70 PB GPFS	1 PB DAOS
Throughput memory to disk	500 GB/s	>750 GB/s
Data Science Archive (DSA)	260 PB	
Throughput disk to tape	10 GB/s	
Network	Intel Omnipath	Mellanox HDR Infiniband
Power consumption	2,7 MW	0,5 MW

Phase 2 Storage Overview

- DAOS
  - 1 PiByte NVMe capacity, 78 TiB Persistent Memory
  - 750 GiByte/s IOR-Easy bandwidth (May 2023)
  - 9.6 million weighted metadata operations per second

  - 42x Lenovo SR630v2 DAOS Servers
    - 2x Icelake 8352Y (24core) CPUs
    - 16x PMem 128GB (BarlowPass)
    - 8x NVMe P5500 3.84TB
    - Erasure Coding 16+2P
    - 2x InfiniBand HDR200

  - 4x additional „data mover“ nodes

- Single node performance
  - Write: 19.6 GiB/s
  - Read: ~40 GiB/s
  - Capacity: 24.8 TiB usable

- Data Science Storage (DSS, HOME and PROJECT)
- GPFS (WORK, SCRATCH)

Experiences and challenges

- Limited first hand experience
  - System delivery in Autumn, until mid-November system fully occupied with bring up
  - First small scale LRZ tests on DAOS only in 2nd week of November

- PVC and DAOS are optimized with different fabrics (PSM3* vs MLX)

  * needed for RDMA

- Interception library is necessary to achieve good performance with Posix containers
  - Mixed GPFS + DAOS usage in job leads to error
  - DAOS Posix container within GPFS path (e.g. HOME) leads to unmounting and permission issues

- Figuring out how best to collect DAOS logs for debugging user jobs

Pilot phase (~2024 January)

- Selected users advised to utilize DAOS pools/containers
  - AI and machine learning training datasets
  - I/O limited workloads (e.g. Hanoi graphs)
  - Usage as fast scratch storage, especially astrophysics and computational life sciences (i.e. projects that can utilize HDF5 containers)
- Educate users: Scaling workshops will include porting to DAOS starting next spring

Summary

- System is delivered and acceptance testing is ongoing
- Performance metrics are promising
- Hands on experience is still lacking
- IO500 announcement later this week!

We have more to share on our PVC experience this afternoon at IXPUG PVC User Group (3 pm, same place)

Leibniz Supercomputing Centre

of the Bavarian Academy of Sciences and Humanities