Evaluating DAOS Usage and Performance for a Classic HPC Application

10th International Parallel Data Systems Workshop, Supercomputing 2025

Steffen Christgau

Zuse Institute Berlin

ZIB
ZUSE INSTITUTE BERLIN

Zuse Institute Berlin (ZIB)

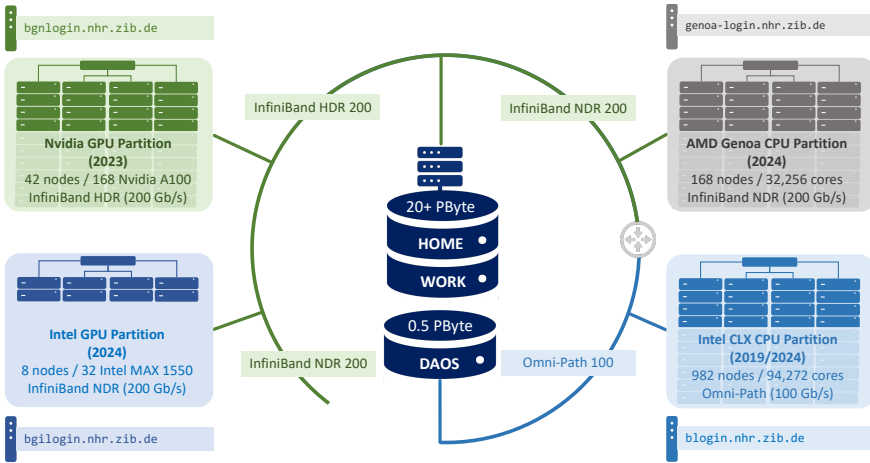
- Tier 2 HPC service provider for academic research (NHR Center)
- DAOS Installation currently ranked #4 in IO-500 10NP

Production | **10 Node Production** | Research | 10 Node Research | Full | Historical

Ranking of the research system submissions that used exactly ten client nodes. This is a subset of the Full List of submissions, showing only one highest-scoring result per storage system. Submitters who want a submission that is currently on the 10 client node Research List to be on the 10 client node Production List should contact the IO500 Steering Committee.

	INFORMATION						IO500			
# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	10	2,080	2,885.57	734.50	11,336.27	
2	ISC23	LRZ	SuperMUC-NG-Phase2-EC-10	Lenovo	DAOS	10	1,120	1,008.81	218.38	4,660.23	
3	ISC25	Hudson River Trading	HRT	DDN	EXAScaler	10	1,600	348.08	136.05	890.51	
4	ISC24	Zuse Institute Berlin	Lise	Megware	DAOS	10	960	324.54	65.01	1,620.13	

HPC System in a Picture



Application Study: PALM

ZIB
ZUSE INSTITUTE BERLIN

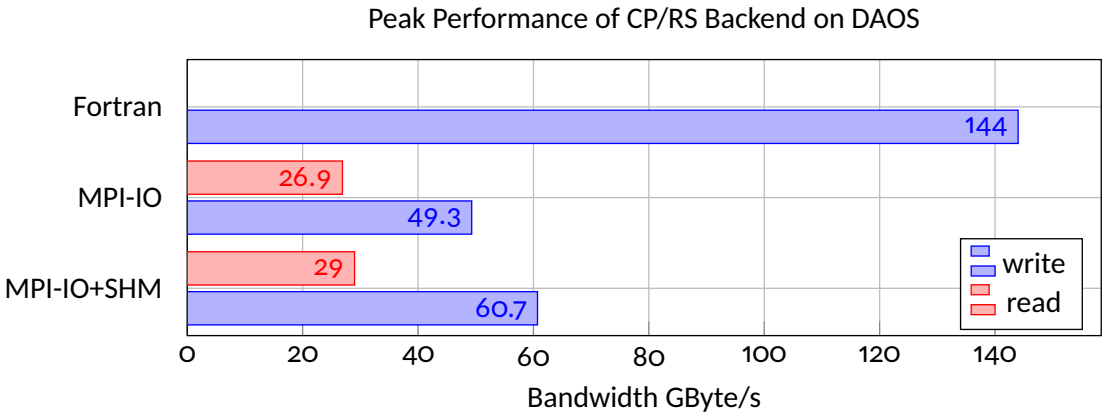
- Meteorological modeling system; Top 5 application codes running at NHR@ZIB
- Highly scalable MPI + OpenMP-parallelized Fortran 2008+ code
- Built-in **checkpoint/restart** (CP/RS) mechanisms/backends:
  1. **Fortran** unformatted I/O: one file per process, streaming of arrays
  2. **MPI-IO**: single large file using MPI datatypes, all process access file
  3. **MPI-IO + shared memory**: manual aggregation in leader process per node

All Backends work with DAOS out of the box

- Output dominated by 3D compute domain data, approx 8 doubles/grid point
- CP/RS is not time-critical but PALM's features enable application-focused testing
- Measurements with 96 × 96 = 9216 processes; about 5 TiB checkpoint size
- **Goals**:
  1. Evaluate Performance for CP/RS backend
  2. Compare DAOS with GPFS and Lustre production file systems

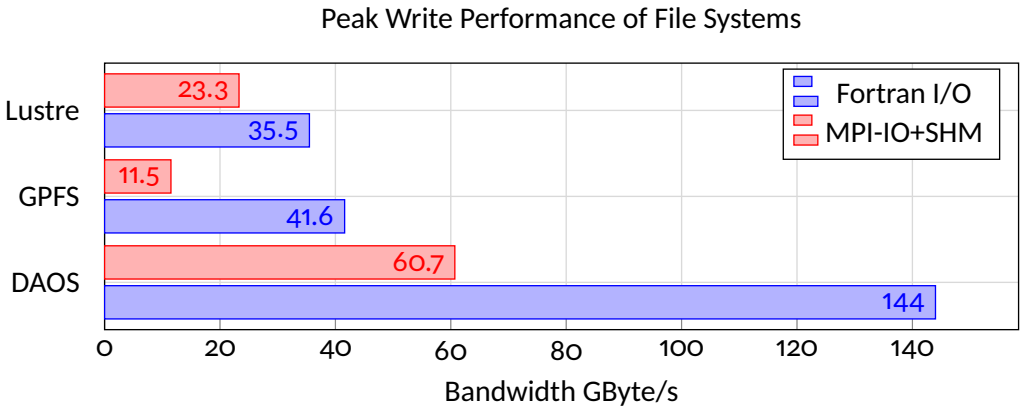
PALM's CP/RS backend performance on DAOS

- Fortran I/O benefits from simplicity (streaming per file), application issue for restore
- Slight benefit for MPI-IO+SHM over MPI-IO, but generally almost identical



Comparison with Production Filesystems

- 10 PB DDN Lustre – measured at storage's EOL, exclusive usage, 73% full
  - two pools: HDD and SSD, data shown for HDD → 35 OSTs HDD, 4 OST SSD, 8 MDTs
  - externally connected to CLX partition via OPA
- 20 PB IBM GPFS – natively connected to other partitions with 200 GBit/s IB, 26% full



Summary

- Good, **ready-to go application support** by DAOS
- **Superior performance of DAOS** compared to production Lustre and GPFS
- Future Work: Explore HDF5/netCDF support, dig into performance behaviors

Summary

- Good, **ready-to go application support** by DAOS
- **Superior performance of DAOS** compared to production Lustre and GPFS
- Future Work: Explore HDF5/netCDF support, dig into performance behaviors

For extended talk slides see DAOS User Group (DUG) from yesterday.

Thanks to Michael Hennecke.