Very Early Experiences with a 0.5 PByte DAOS Testbed

Steffen Christgau, Tobias Watermann, Thomas Steinke

Supercomputing Department
Zuse Institute Berlin

DAOS User Group Meeting 2020

Computing + Data Storing @ Zuse Institute Berlin

- ZIB operates HLRN-IV "Lise" for German science community
- **Motivation for DAOS:**
  - Our current Lustre installation w/o Burst Buffer ⇒ poor IOPS performance
  - Complement Lustre? — Burst Buffer, BeeGFS/BeeOND, DAOS
  - Workloads that can benefit from DAOS: turbulence simulation, astrophysics, small/many files I/O, ··· AI/DL
  - Evaluation of new storage concepts vs. "traditional" concepts
- DAOS as research and later production platform

HLRN-IV "Lise" @ ZIB

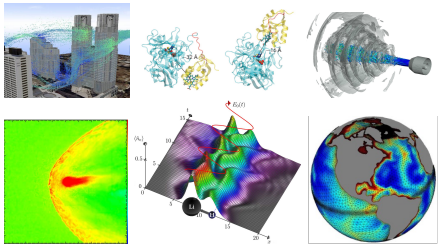
8 PFlop/s peak

1,270 nodes, Intel CLX AP

121,920 cores

3,000+ users — 200+ projects

8 PB Lustre



- **Chemistry** incl. **Material Science**
- **Earth Science** incl. **Climate Research**
- **Engineering**
- **Life Science** (biology, medicine)
- **Physics** incl. **Astro** and **High-Energy Physics**

DAOS @ ZIB: Exploration Testbed

- DAOS user since July 2019
- version 0.6… and git commits before that
- manual compilation process

- **Exploration Testbed**: used for DCPM and DAOS exploration
  - isolated from HLRN
  - 2 Inspur dual-socket nodes (CLX-SP Platinum 8260L)
  - 3 + 6 TB Optane DCPMM and 384 GB + 768 GB DRAM
  - 8 + 16 TB Optane SSD
  - single 100 Gb/s OmniPath back-to-back
  - CentOS 7

DAOS @ HLRN: Integration Testbed

larger testbed **to be integrated in HLRN** infrastructure

- 20 dual-socket server nodes (CLX Gold 6240R)
- 192 GB DRAM
- 1.5 TB DCPM
- 25.6 TB NVMe NAND SSD
- 2 x 100 Gbit/s OmniPath

total capacity 512 TB SSD + 30 TB DCPM

Installation Experiences with DAOS 1.0 (I)

compared to earlier versions

- Prebuilt DAOS packages are a good thing! We use these, no in-house build.
- online **package repositories** would be even better (no login please, see oneAPI)
- better OS integeration, support for installation and deployment

- **documentation** improved a lot
- documentation sometimes more promising than reality

- **configuration defaults** and comments from examples do not apply
- *# scm_class default: dcpm → scm config validation failed: scm_class not set*

- ***immutable after reformat*** hints are good

Installation Experiences with DAOS 1.0 (II)

compared to earlier versions

- **error messages** not always helpful: *failed to connect to pool: -1026*
- logs are more useful (sometimes)
- content of stderr vs. log files vs. system logging (journal)

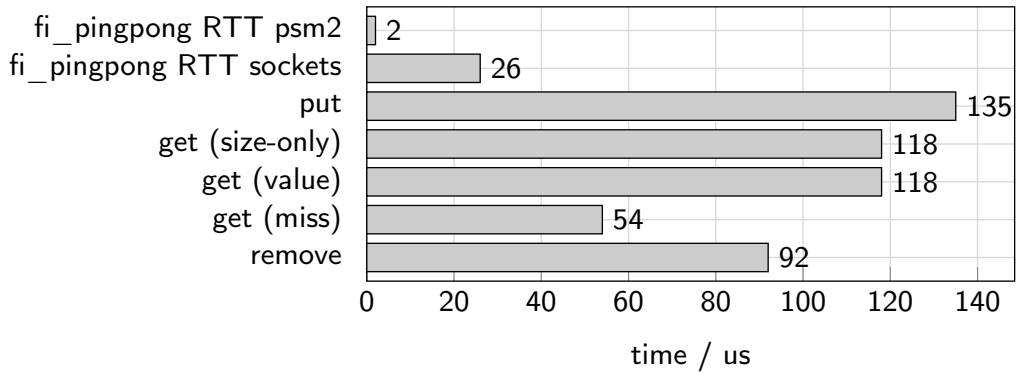
- **MPI**(CH) integration appreciated!
- **middleware** in general: unclear version management $\rightarrow$ packages?!

- user interaction with fusefs/POSIX container: orphaned/forgotten mounts
- dfuse daemon might be a good thing

- **PSM2 issue**: multi-tennant usage with OmniPath not supported
- intent to use sockets or **verbs** instead

Simple DAOS Key Value API Benchmark

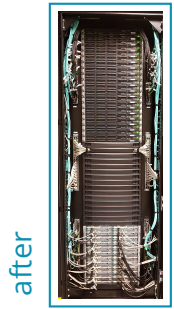
- use DAOS key value (kv) library, DAOS 1.0.1 for very simple test
- perform operation 1000× in blocking fashion, median reported
- key = 4 Bytes, value = 32 Bytes



Current Status

Phase 1: Installation / integration

- hardware shipped mid September, ready to boot OS end of September
- DAOS software integration in HLRN cluster near completion
- early tests on exploration testbed with critical HLRN workloads
  - MPI IO middleware works seemless with MPI-ready application, see DUG'19
  - netCDF/HDF5 testing planned (waiting for compatible versions)



Next Planned Step

Phase 2: Research with integration testbed

- Usability & user interface: application integration for a few test cases
- Administration: experiences with capabilities of the management of pools, containers, . . ., monitoring & performance

Phase 3: Access for selected user projects

- intent to use per-project pool
- provide DAOS as optional and **additional offer for heavy IO worklads** besides existing Lustre (work), NFS (home), and SSD drive (scratch)
- support power users in migration $\rightarrow$ easy when MPI/HDF/netCDF is used for IO
- disseminiation planned for 2021 ff.

Summary

- DAOS improved significantly since our first contact. . . and is still improving
- integration into production system in progress
- mature for early access of (power) users

Thanks for your attention!
Questions?