First Experiences with CDI - PIO on DAOS

High Performance Computing in Meteorology

2021-09-24 | #HPCWS2021

Michael Hennecke (Lenovo)
Thomas Jahns (DKRZ)

Co-Workers: C. Pospiech (Lenovo), P. Adamidis (DKRZ), S. Eggerling (Lenovo)

ECMWF

Abstract

CDI-PIO is the parallel I/O component of the Climate Data Interface (CDI) that is developed and maintained by the Max-Planck-Institute for Meteorology and DKRZ. It is used by ICON, MPIOM, ECHAM, and the Climate Data Operator (CDO) toolkit. The two main I/O paths for output data are writing GRIB files using MPI-IO, and writing NetCDF4 files using HDF5 (which may then also use MPI-IO, or other VOL plugins).

The Distributed Asynchronous Object Storage (DAOS) is a new open source high performance object store for storage class memory and NVMe storage, which has been integrated into the ROMIO MPI-IO implementation. The HDF5 consortium is also developing a native HDF5 VOL plugin for DAOS.

This presentation will outline how CDI-PIO can be run on a DAOS storage system using the ROMIO DAOS backend. We will also report first performance results comparing Intel DAOS and IBM Spectrum Scale on similar NVMe storage hardware.

Climate Data I/O Interface (CDI)

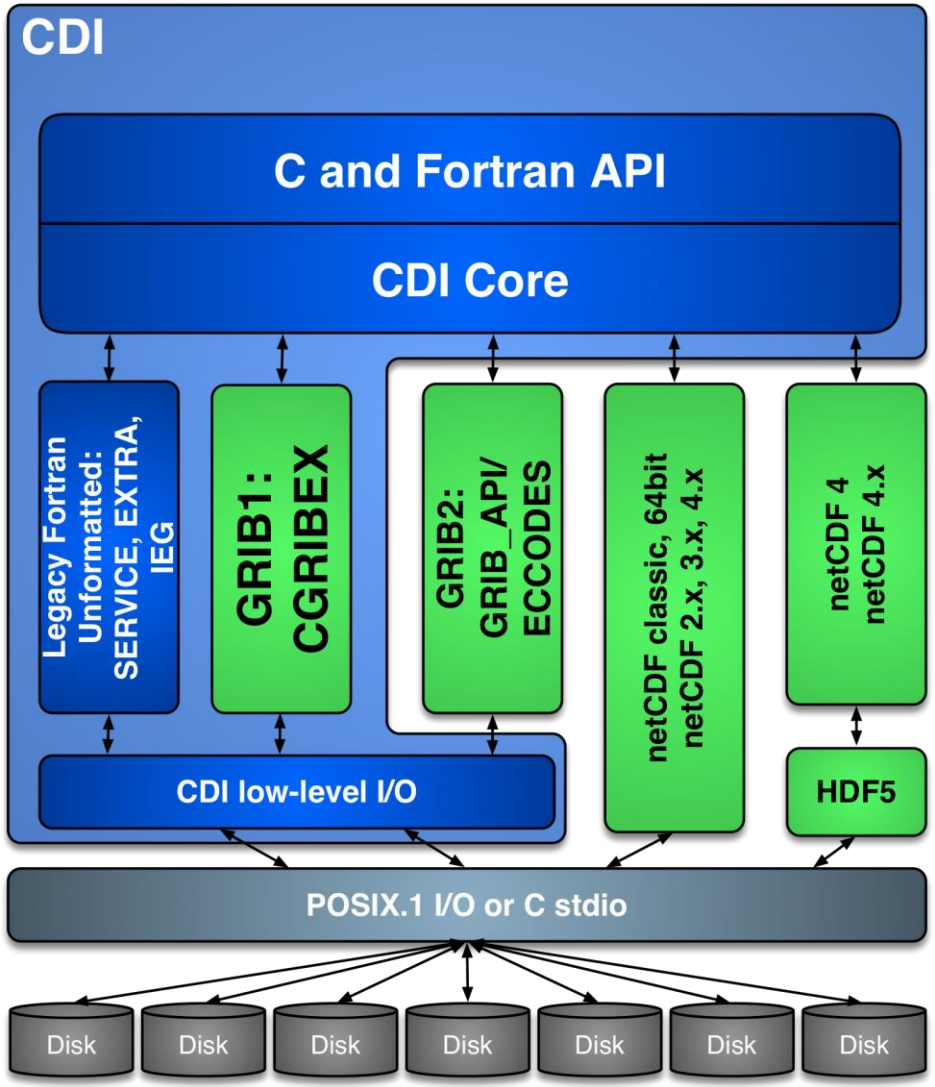
http://code.zmaw.de/projects/cdo

Climate Data Operators (CDO)

Part of CDO is the I/O interface CDI (Climate Data Interface), which it shares with all major MPI-M climate models (ICON,MPIOM,ECHAM)

- GRIB1 via CGRIBEX (MPI-M)
- GRIB2 via GRIB-API/ECCODES (ECMWF)
- NetCDF, CF-convention (UNIDATA)
- SERVICE, EXTRA, IEG (MPI-M binary formats)

GRIB support includes highly efficient, fast compression algorithms.



CDI with Parallel I/O

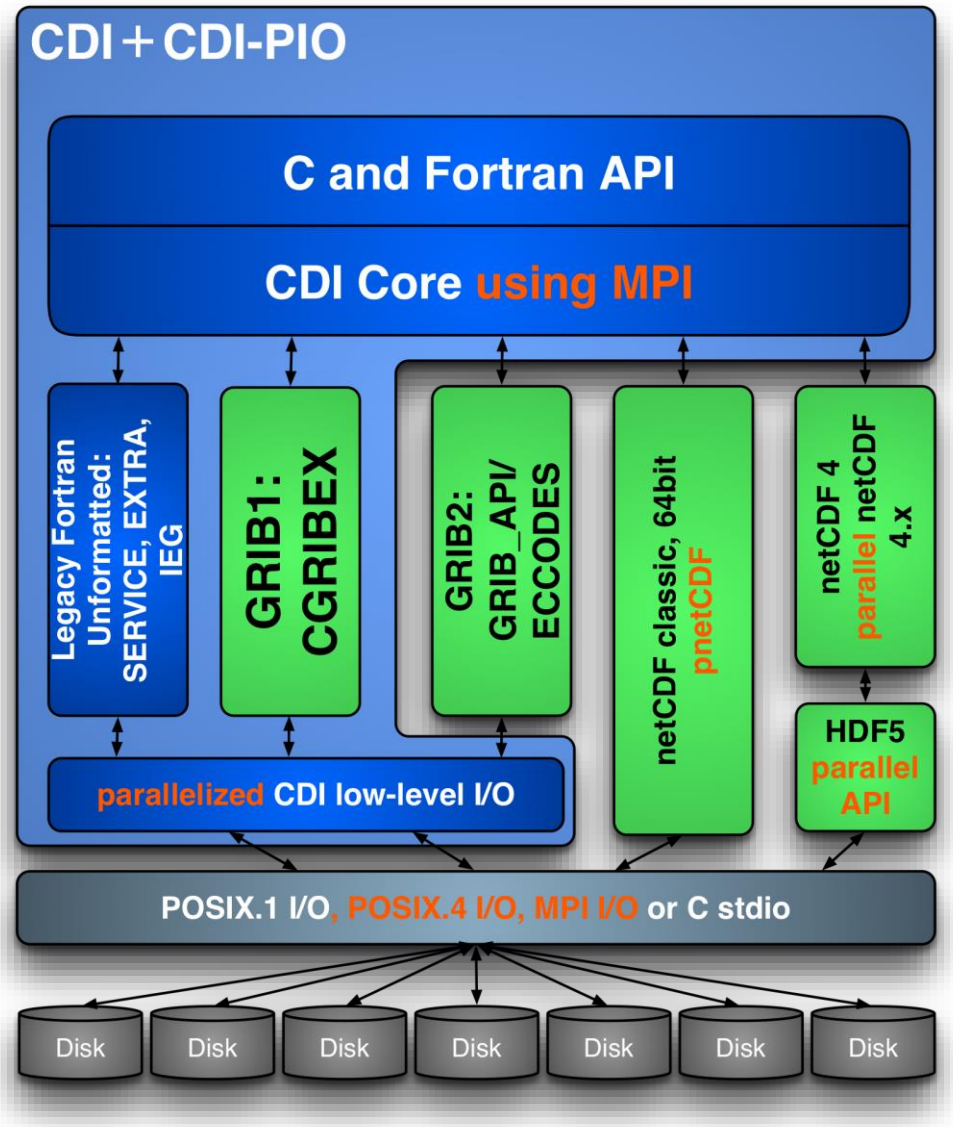
http://code.zmaw.de/projects/cdo

Climate Data Operators (CDO)

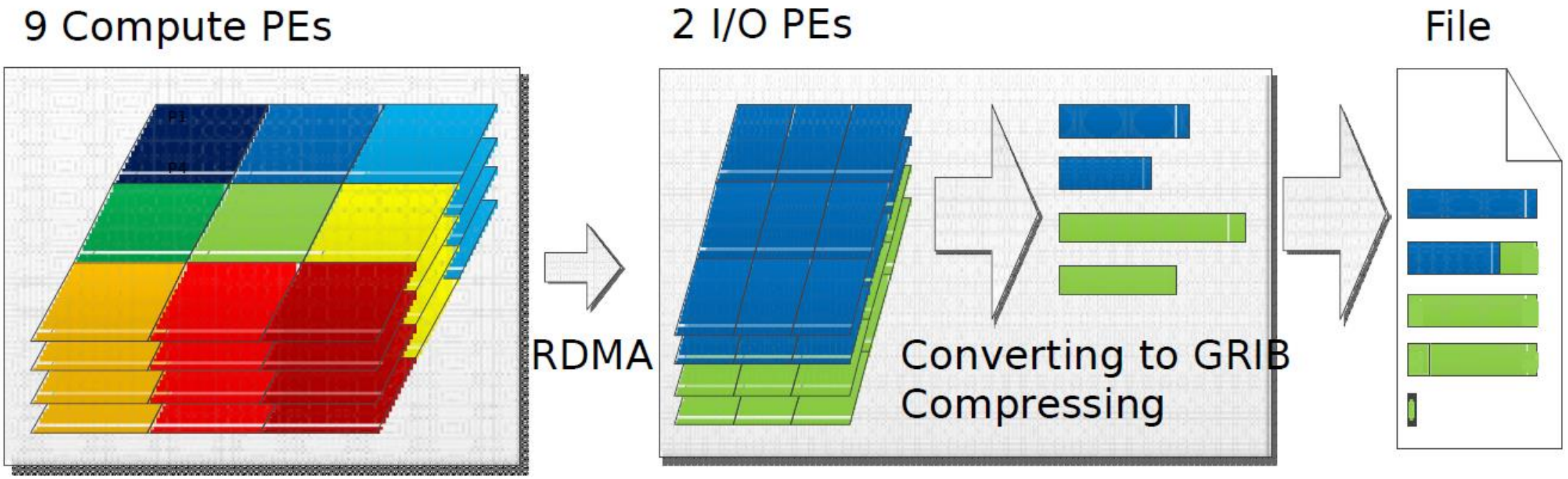
Part of CDO is the I/O interface CDI (Climate Data Interface), which it shares with all major MPI-M climate models (ICON,MPIOM,ECHAM)

- GRIB1 via CGRIBEX (MPI-M)
- GRIB2 via GRIB-API/ECCODES (ECMWF)
- NetCDF, CF-convention (UNIDATA)
- SERVICE, EXTRA, IEG (MPI-M binary formats)

GRIB support includes highly efficient, fast compression algorithms.



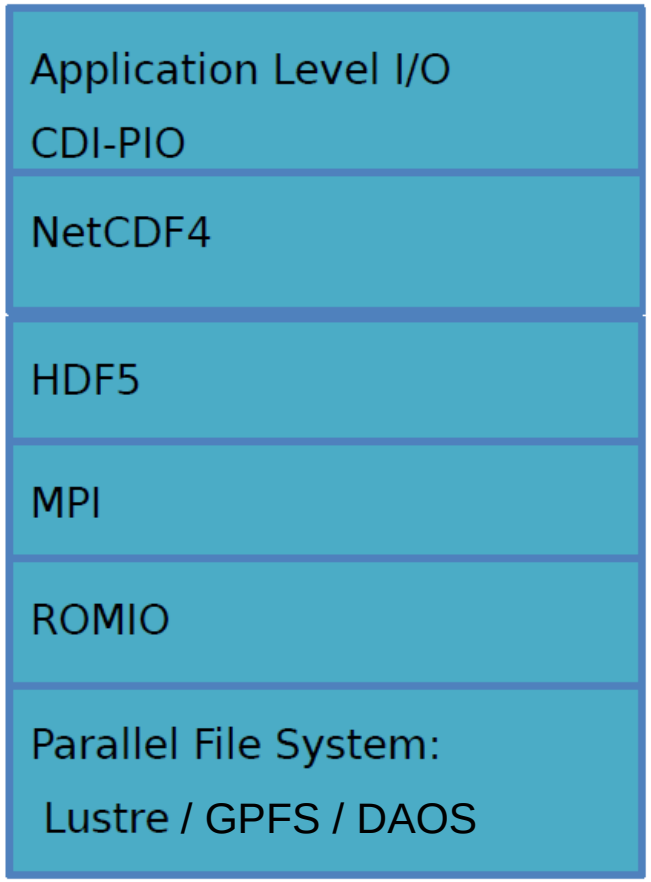
Data Flow Example with 9 Compute PEs and 2 I/O PEs



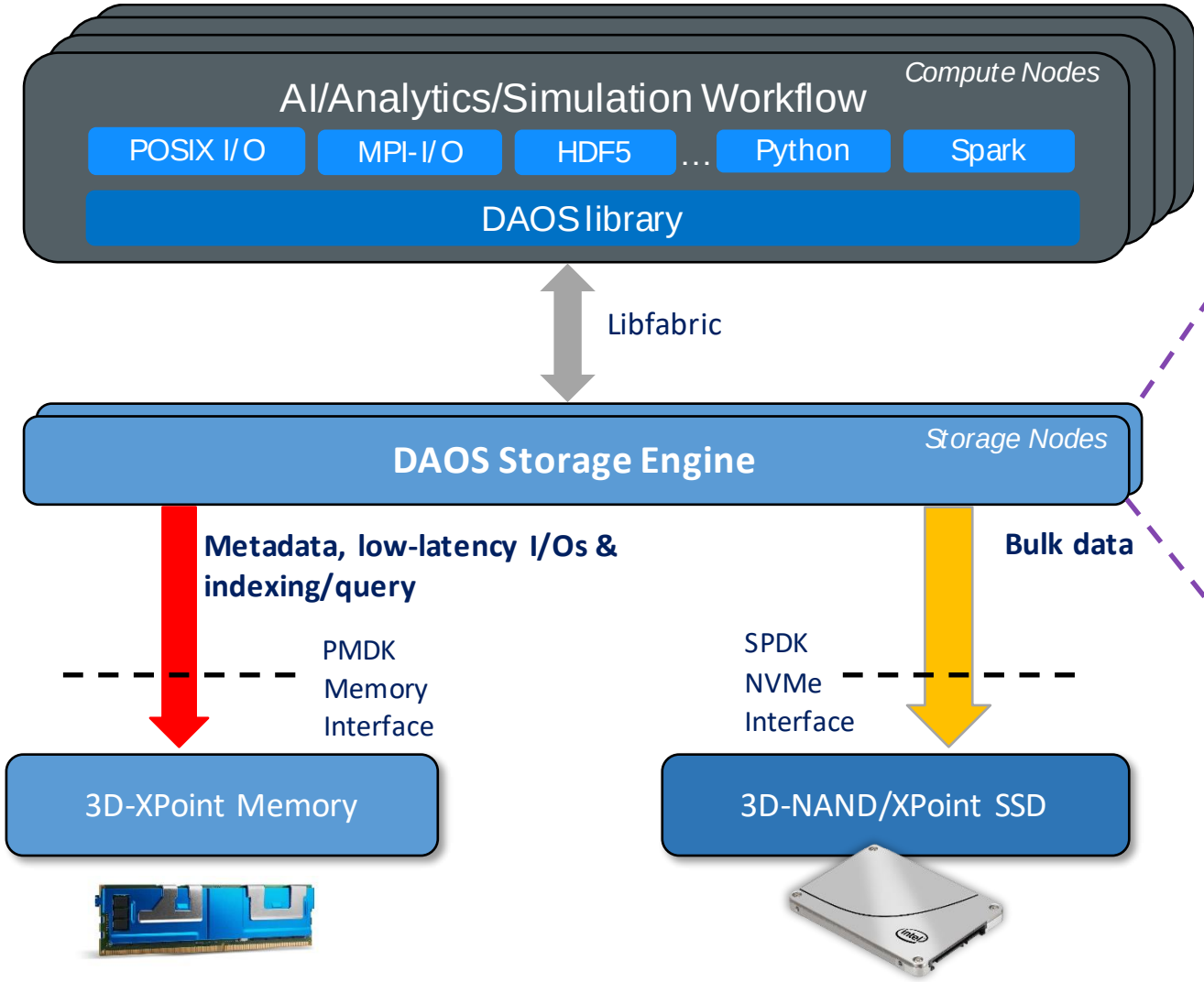
The I/O Software Stack

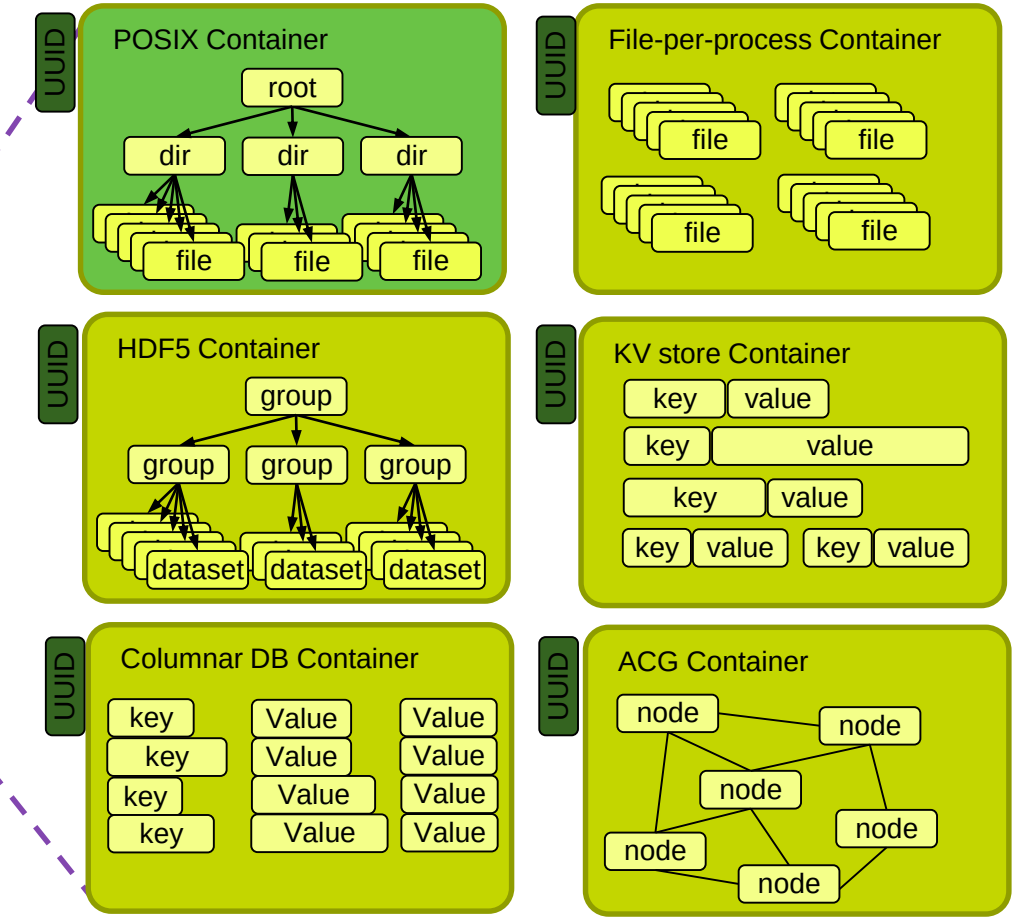
GRIB output

NetCDF4 output

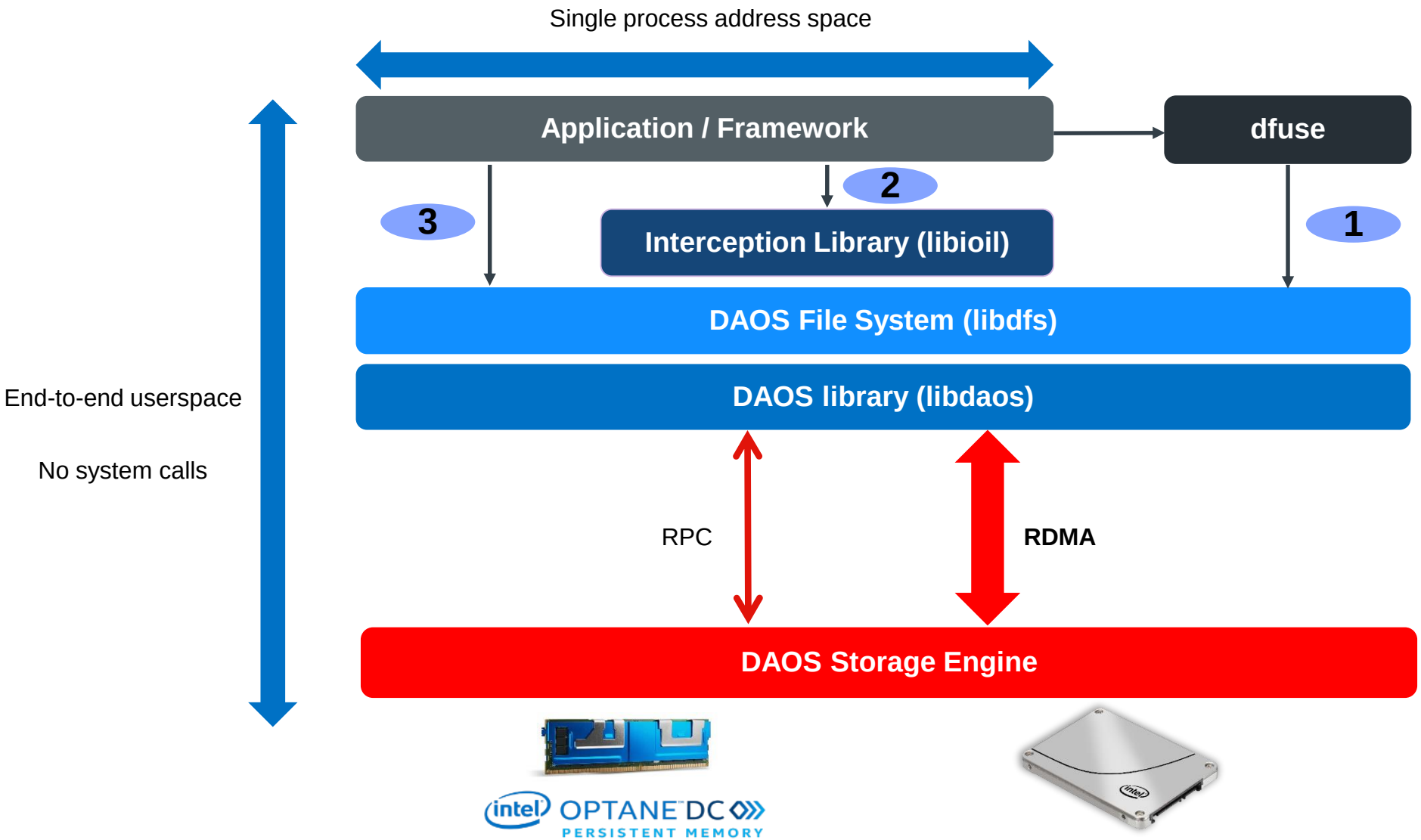


The DAOS Exascale Storage Stack – Software Architecture

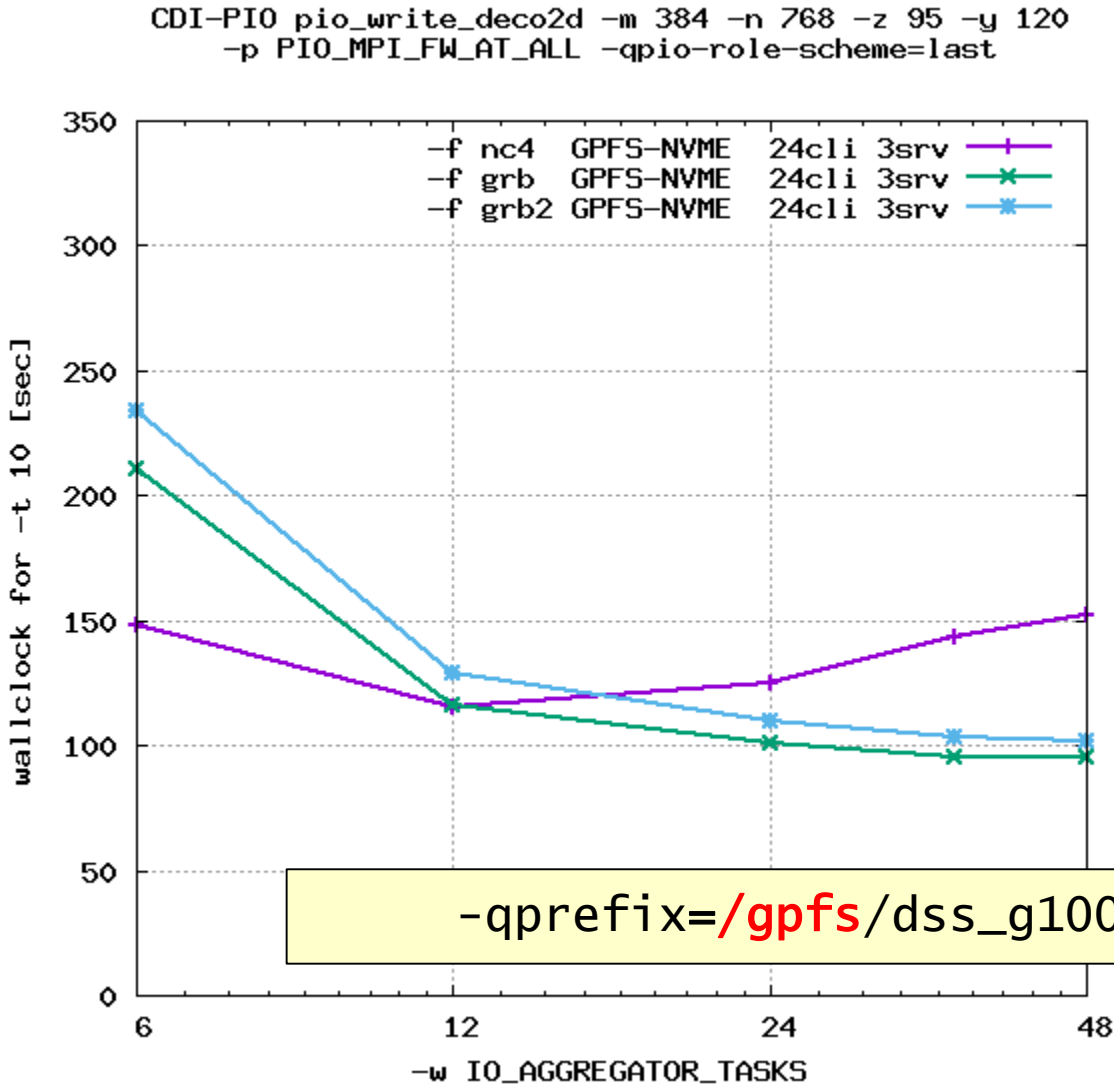




Three Ways of POSIX Filesystem Support in DAOS

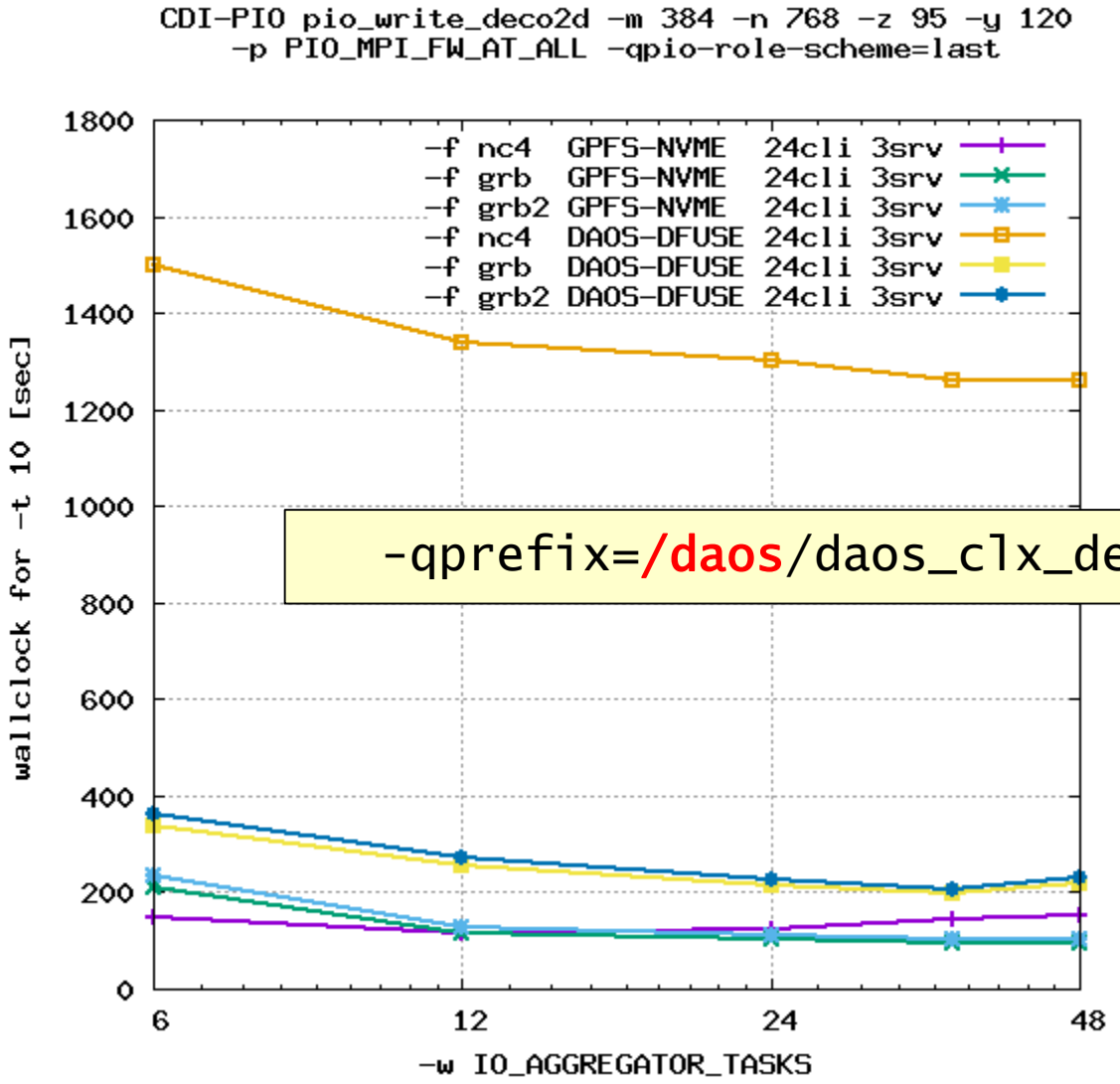


Porting CDI-PIO to DAOS – Baseline on GPFS-NVMe



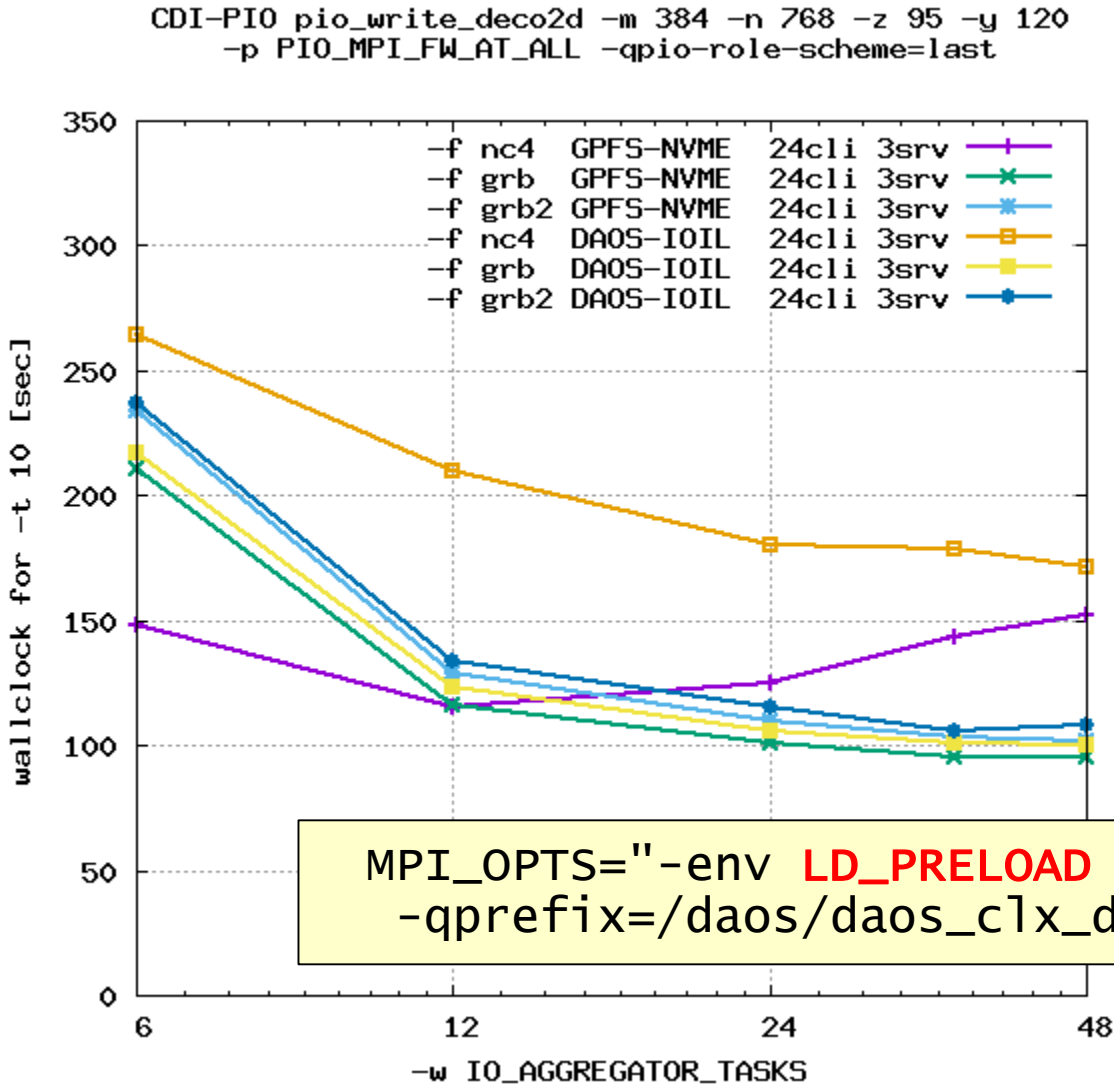

```
-qprefix=/gpfs/dss_g100/mhennecke/cdi-pio
```

Porting CDI-PIO to DAOS – Using dfuse mount 1



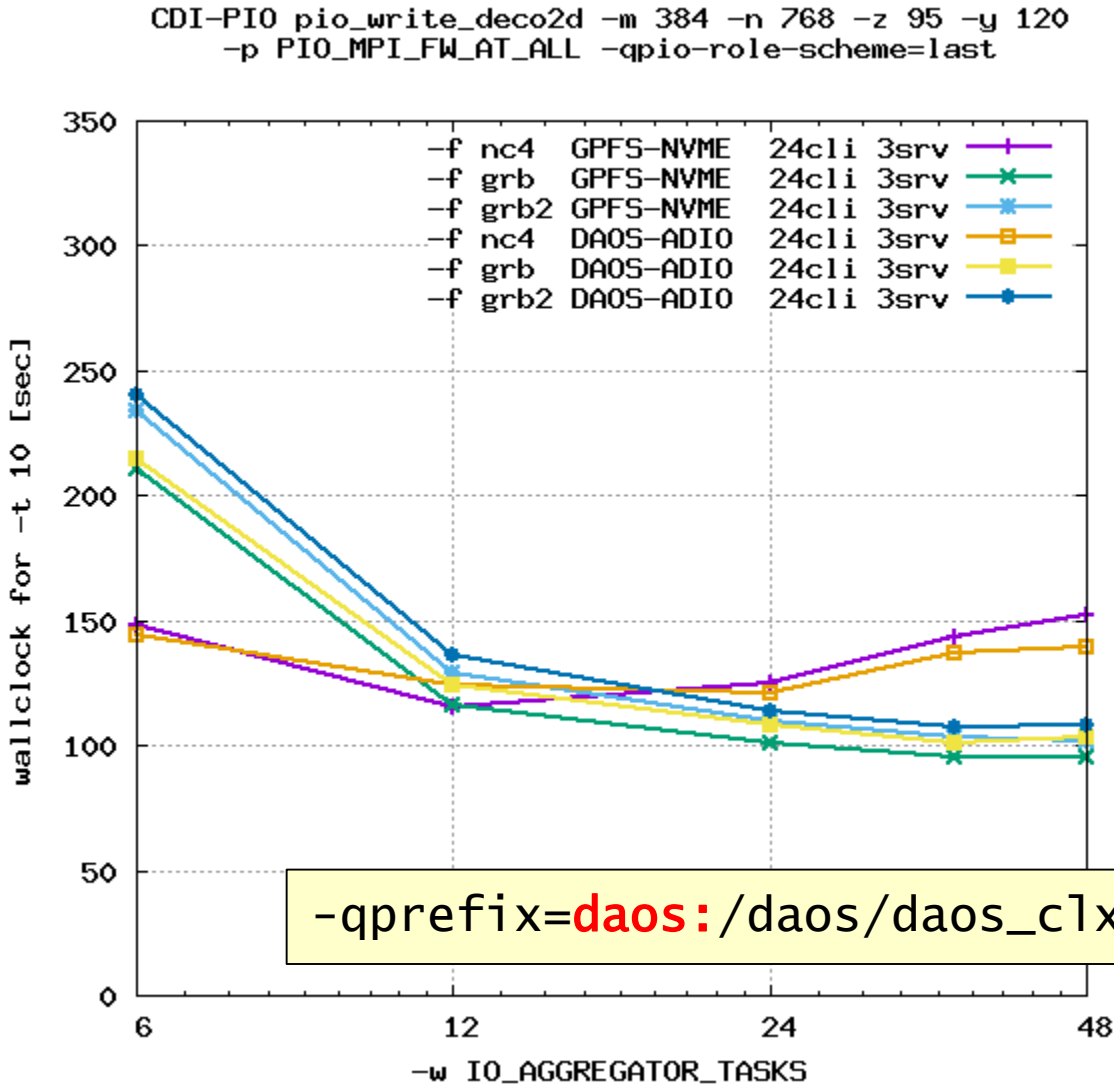
-qprefix=/daos/daos_clx_dev1/mhennecke/cdi-pio

Porting CDI-PIO to DAOS – Using `dfuse` with `IOIL` library 2



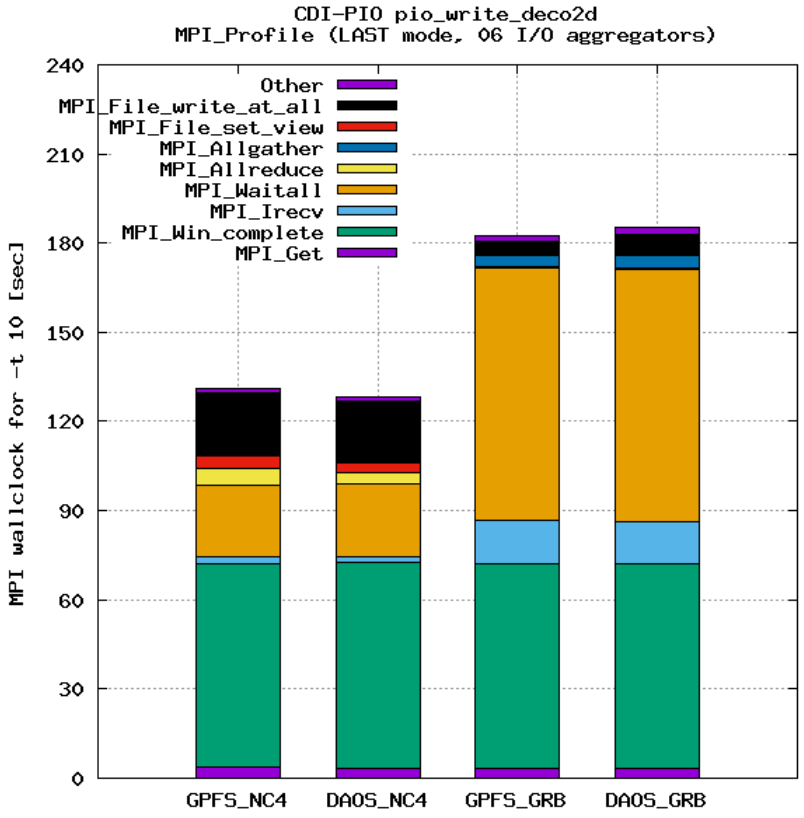

```
MPI_OPTS="-env LD_PRELOAD /usr/lib64/libioil.so“
 -qprefix=/daos/daos_clx_dev1/mhennecke/cdi-pio
```

Porting CDI-PIO to DAOS – Using ROMIO DAOS driver 3

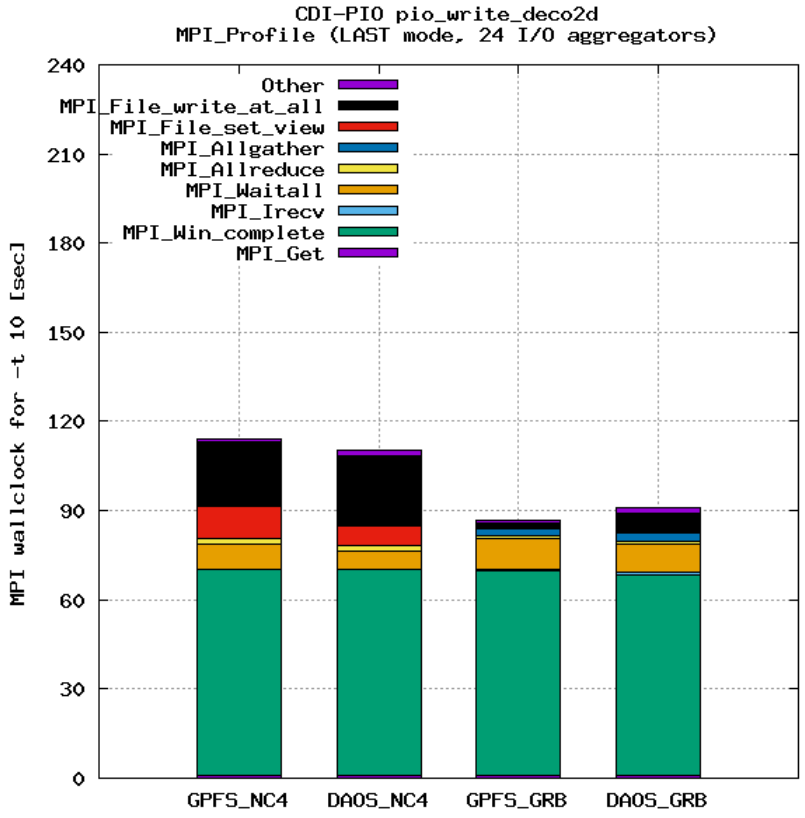



```
-qprefix=daos:/daos/daos_clx_dev1/mhennecke/cdi-pio
```

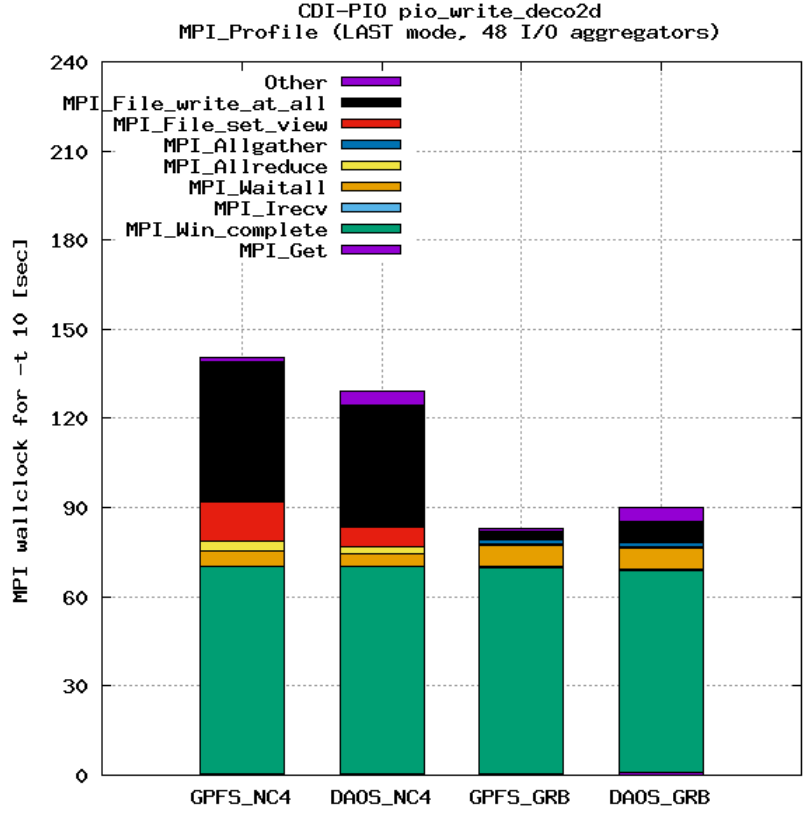
A Closer Look at the MPI Profiles („LAST“ Mode)...



6 I/O aggregators



24 I/O aggregators

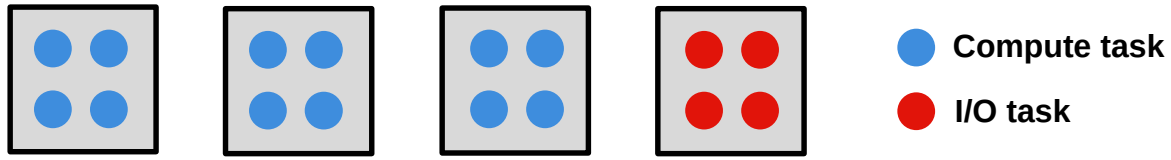


48 I/O aggregators

„BALANCED“ Mode: One I/O Aggregator Task per Node

CDI-PIO „**LAST**“ Mode:

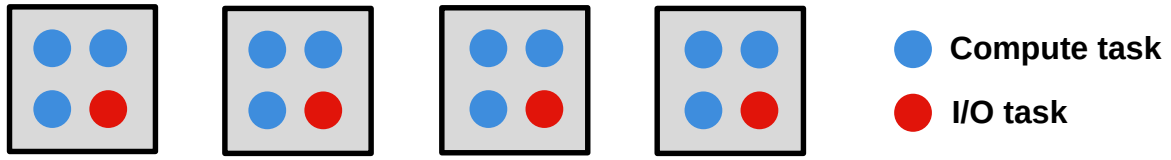
- I/O aggregator tasks are the last MPI ranks in the job → get allocated on the last node:



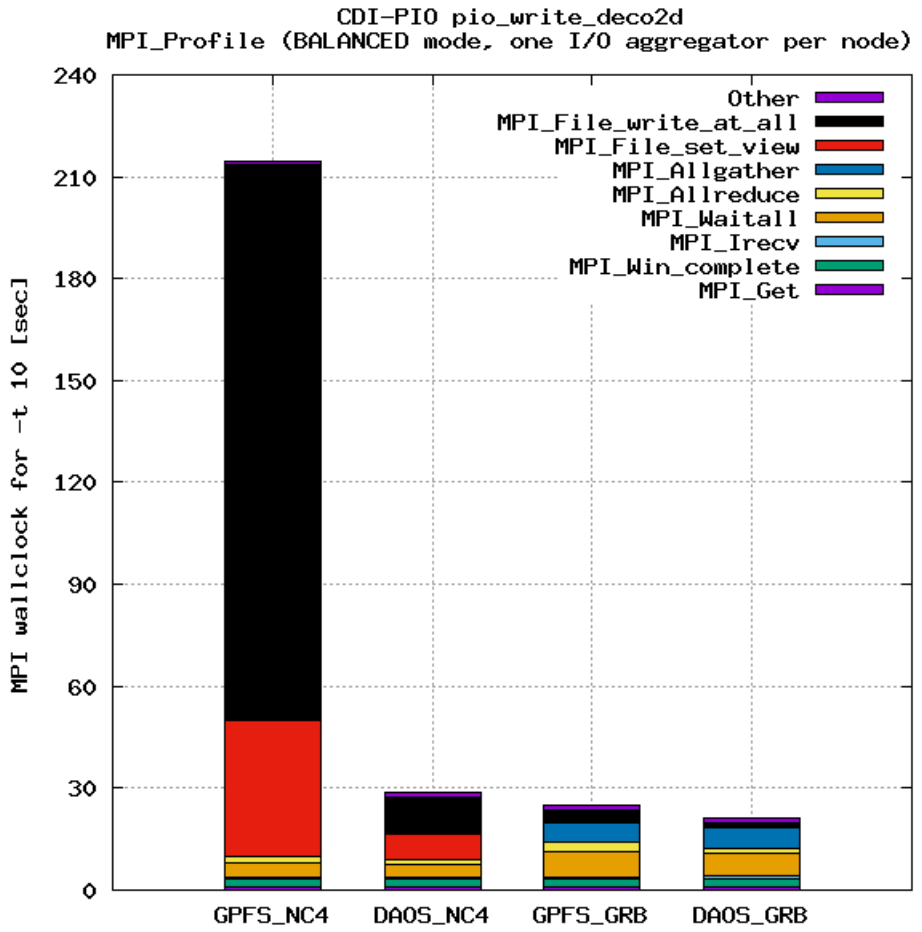
- Used in production, works well with task allocation of simulation codes like ICON

CDI-PIO „**BALANCED**“ Mode:

- One I/O aggregator task per node:



- Not yet used in production, but promising...

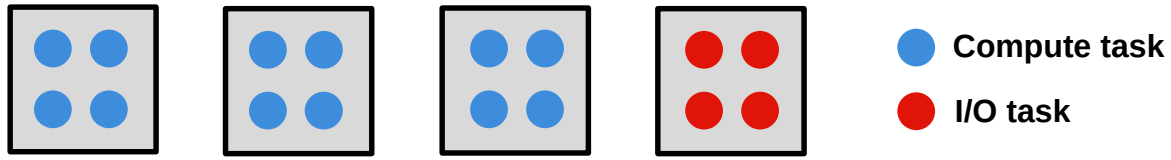


24 I/O aggregators on 24 nodes

„BALANCED“ Mode: One I/O Aggregator Task per Node

CDI-PIO „**LAST**“ Mode:

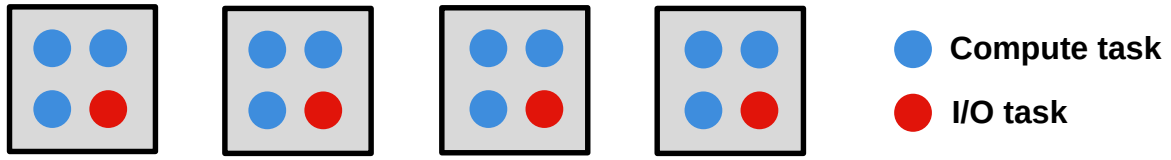
- I/O aggregator tasks are the last MPI ranks in the job → get allocated on the last node:



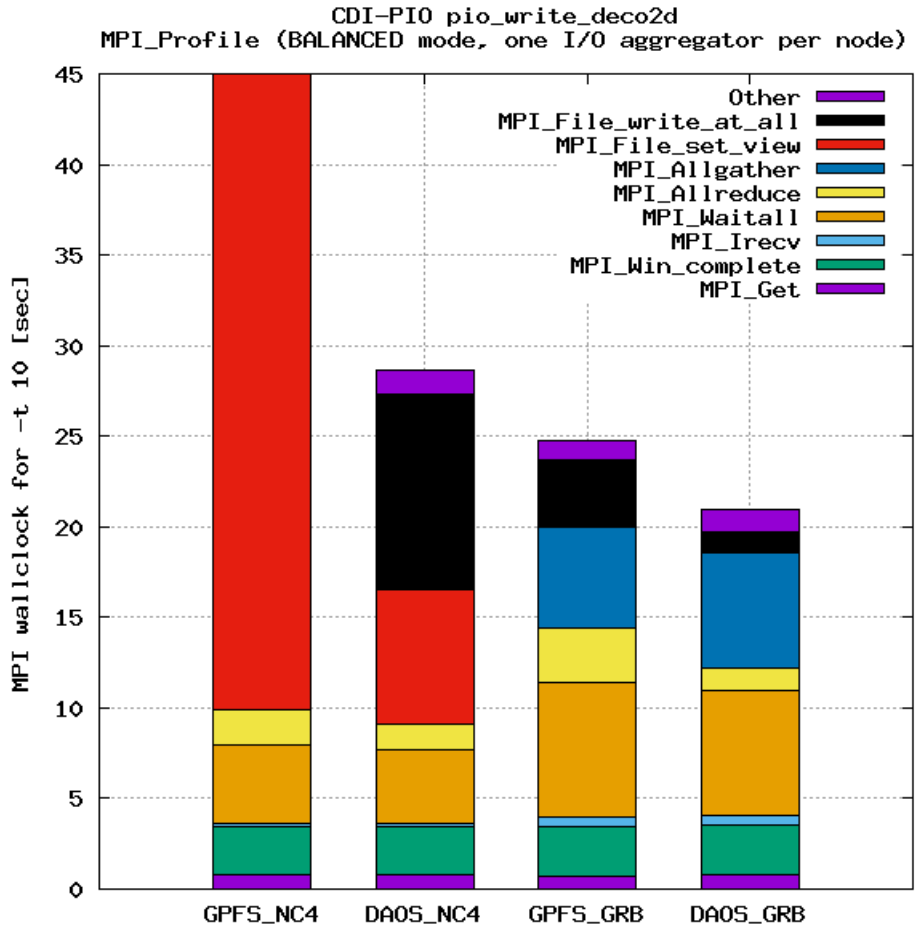
- Used in production, works well with task allocation of simulation codes like ICON

CDI-PIO „**BALANCED**“ Mode:

- One I/O aggregator task per node:



- Not yet used in production, but promising...



zoomed in from previous chart...

Next Step: HDF5 VOL Plugin for DAOS (no MPI-IO, no POSIX)

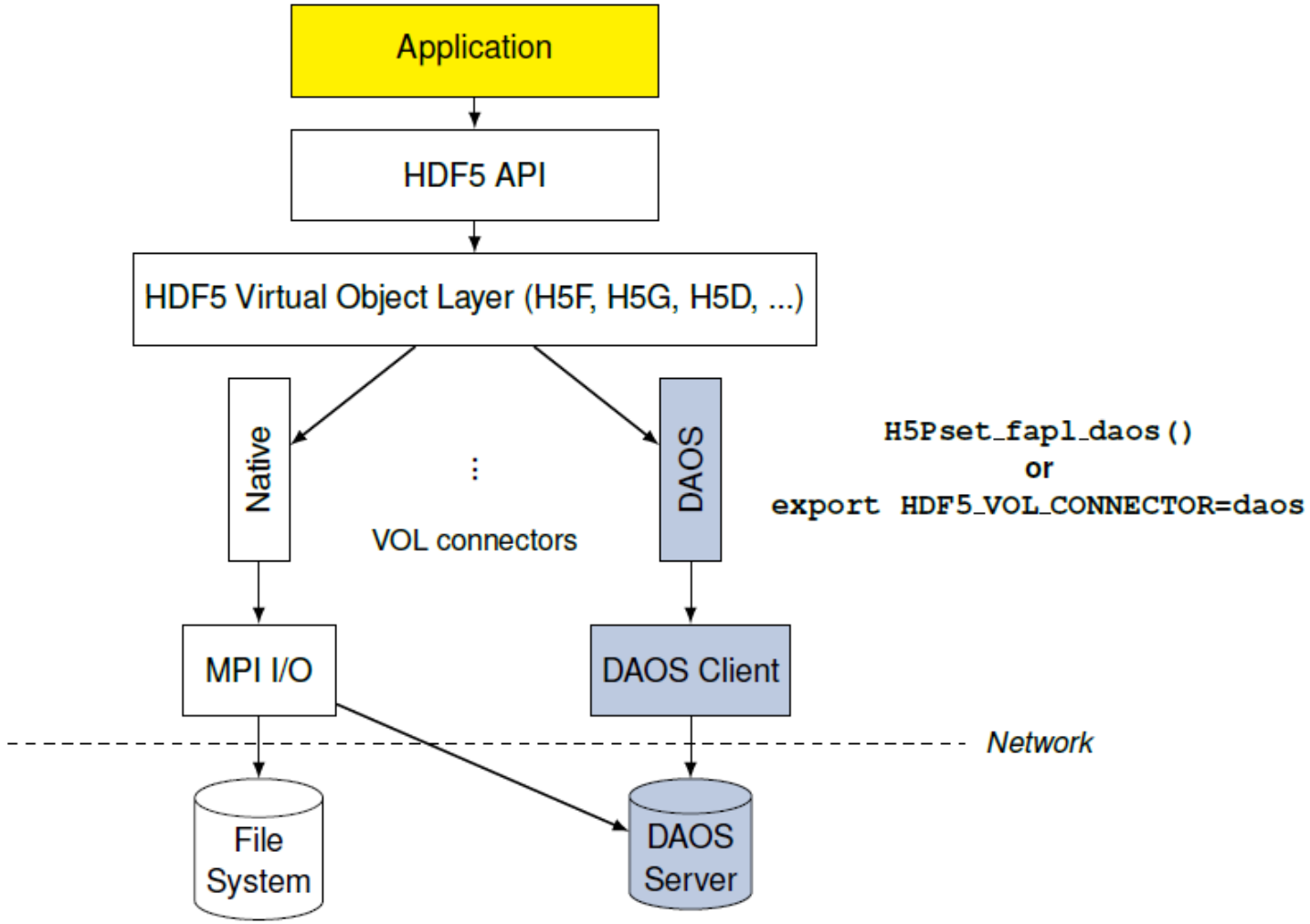
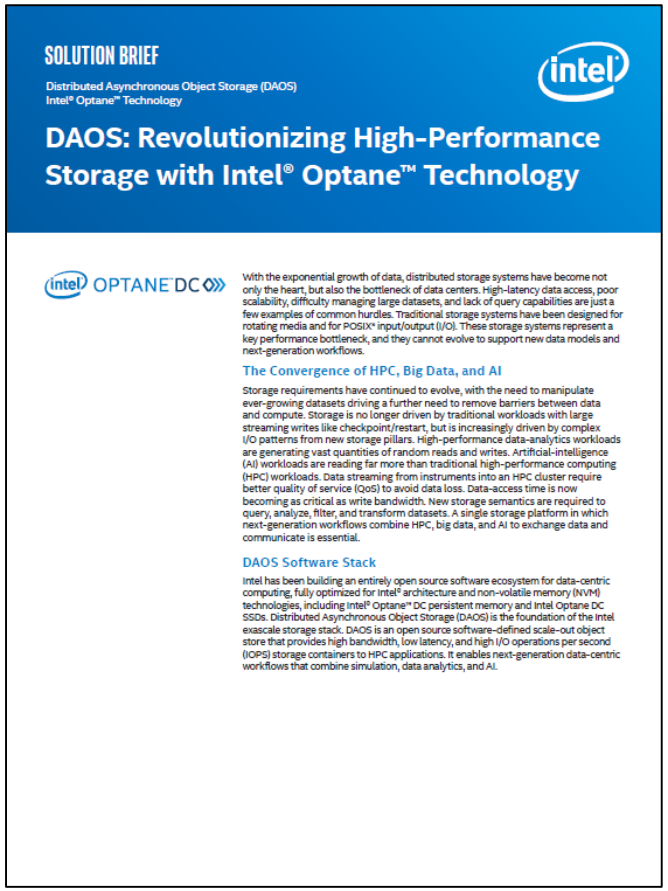
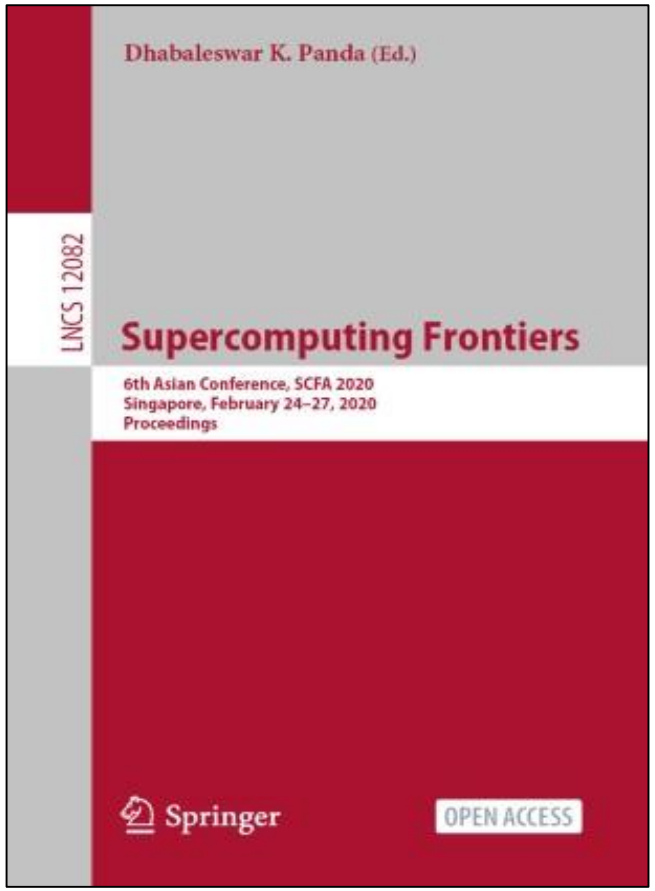


Figure 1 – DAOS within Virtual Object Layer. All of the HDF5 I/O related calls are routed to the DAOS VOL connector.

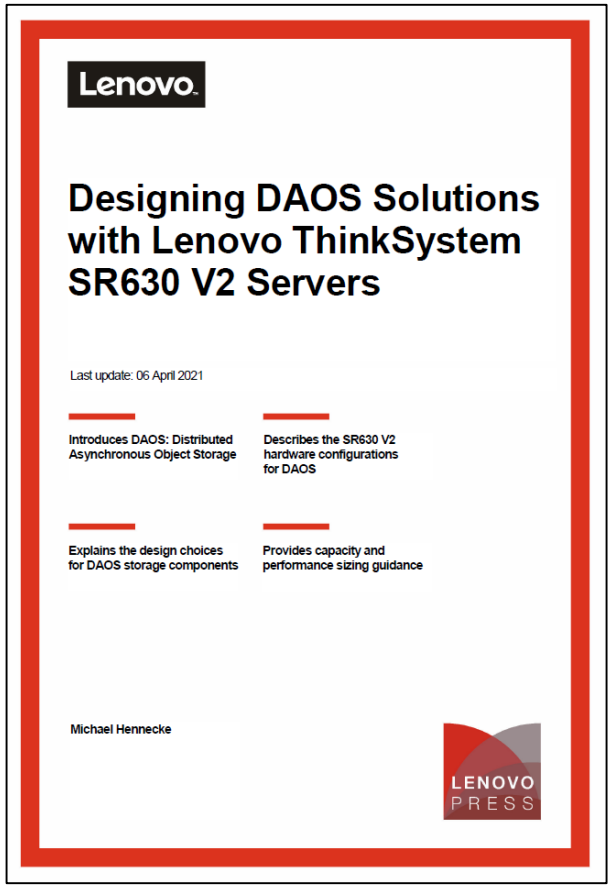
DAOS Documentation



Intel's HPC **Solution Brief:**
DAOS with Optane Technology



Intel / Lenovo **DAOS Article**
(SC-Asia, Springer LNCS 12082)



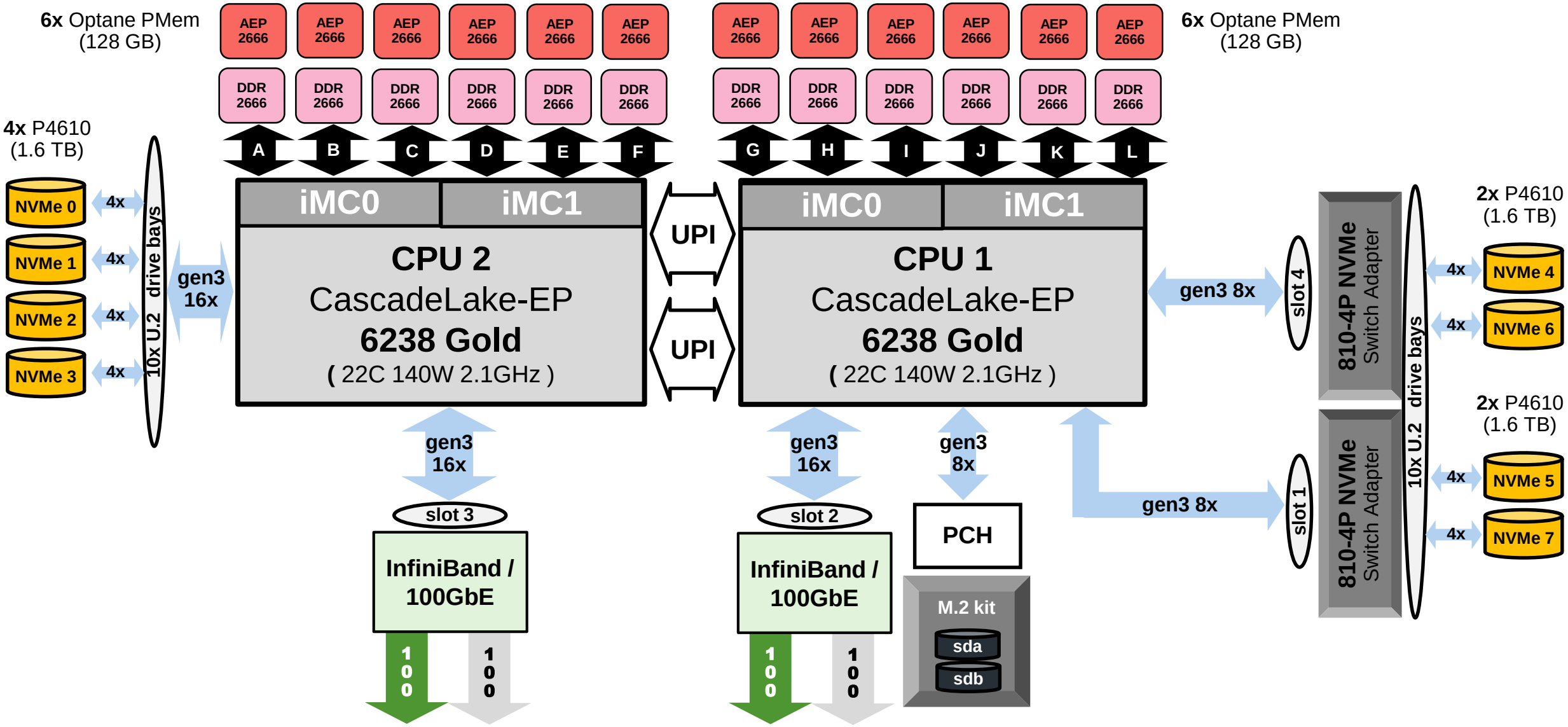
DAOS on Lenovo SR630, **SR630v2**
(LenovoPress LP1398, LP1421)

thanks.

High Performance
Computing
in Meteorology

ECMWF

DAOS Server Architecture: Lenovo ThinkSystem SR630



GPFS Server Architecture: Lenovo ThinkSystem SR630