Hewlett Packard Enterprise

IXPUG INTEL EXTREME PERFORMANCE USERS GROUP **ISC26 Workshop** (26-Jun-2026)

Towards DAOS on Omni-Path: Mercury Performance on CN5000

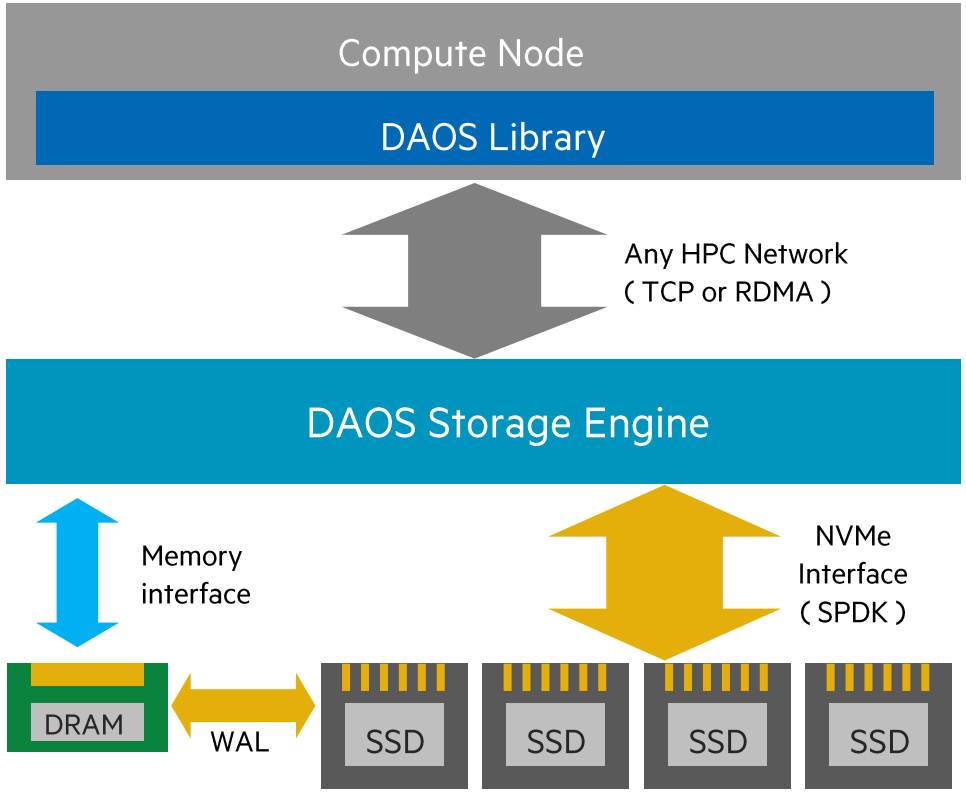
Michael Hennecke
Distinguished Technologist – HPE Cray Supercomputing Storage
Outreach Committee Chair – DAOS Foundation

DAOS

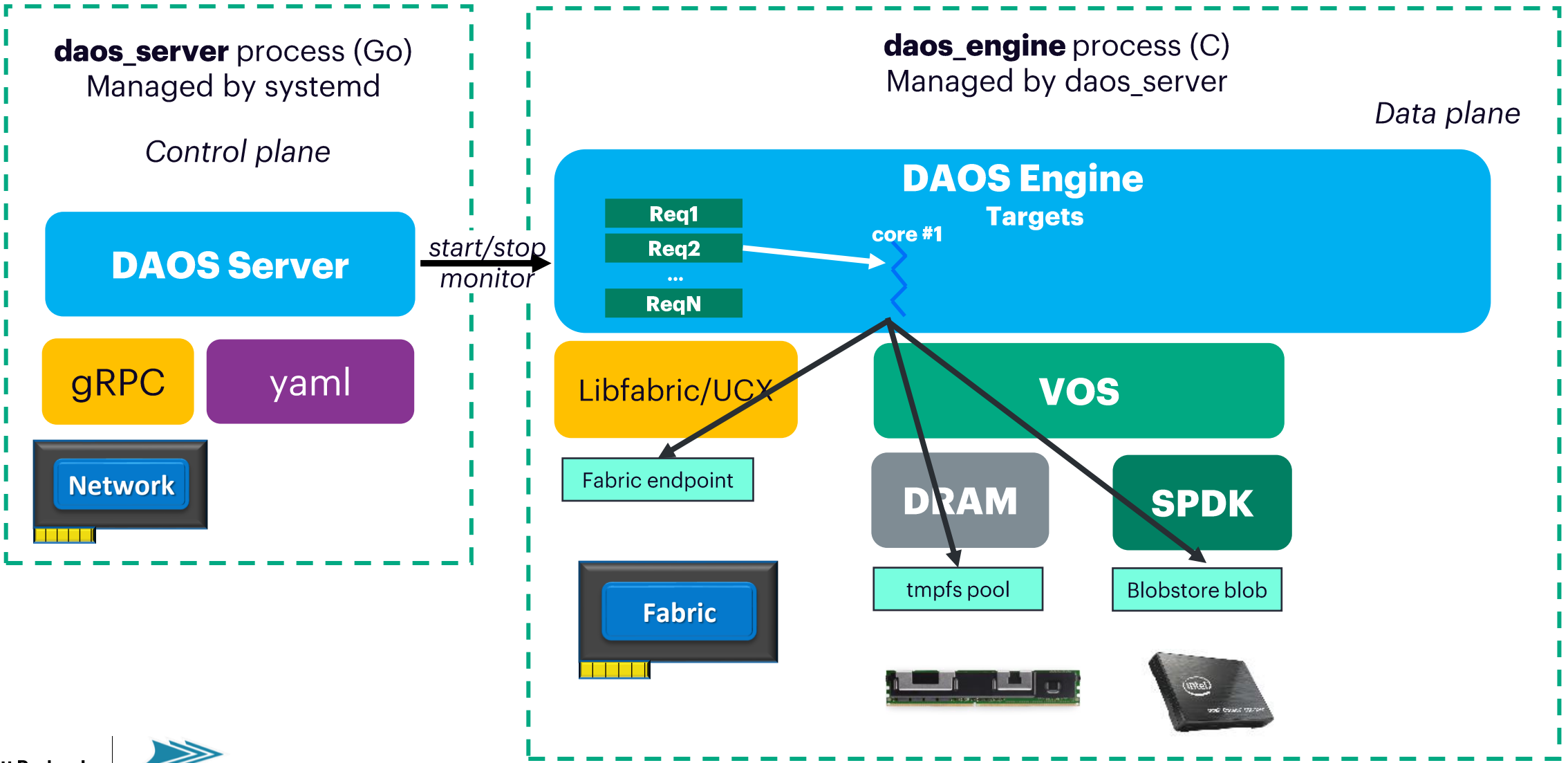
Distributed Asynchronous Object Storage (DAOS)

- DAOS is an **open source**, scale-out, all-flash system
  - Distributed key/value store; *versioned* byte-granular I/O
  - No centralized metadata servers; no locking; no kernel code
- Originally developed by **Intel** for PMem + NVMe
  - Prototype over Lustre 2012-2014; standalone since 2015
  - Optane PMem dependency removed in 2024
    - https://doi.org/10.1007/978-3-031-40843-4_26
- **DAOS Foundation** created 2023 (under THE LINUX FOUNDATION)

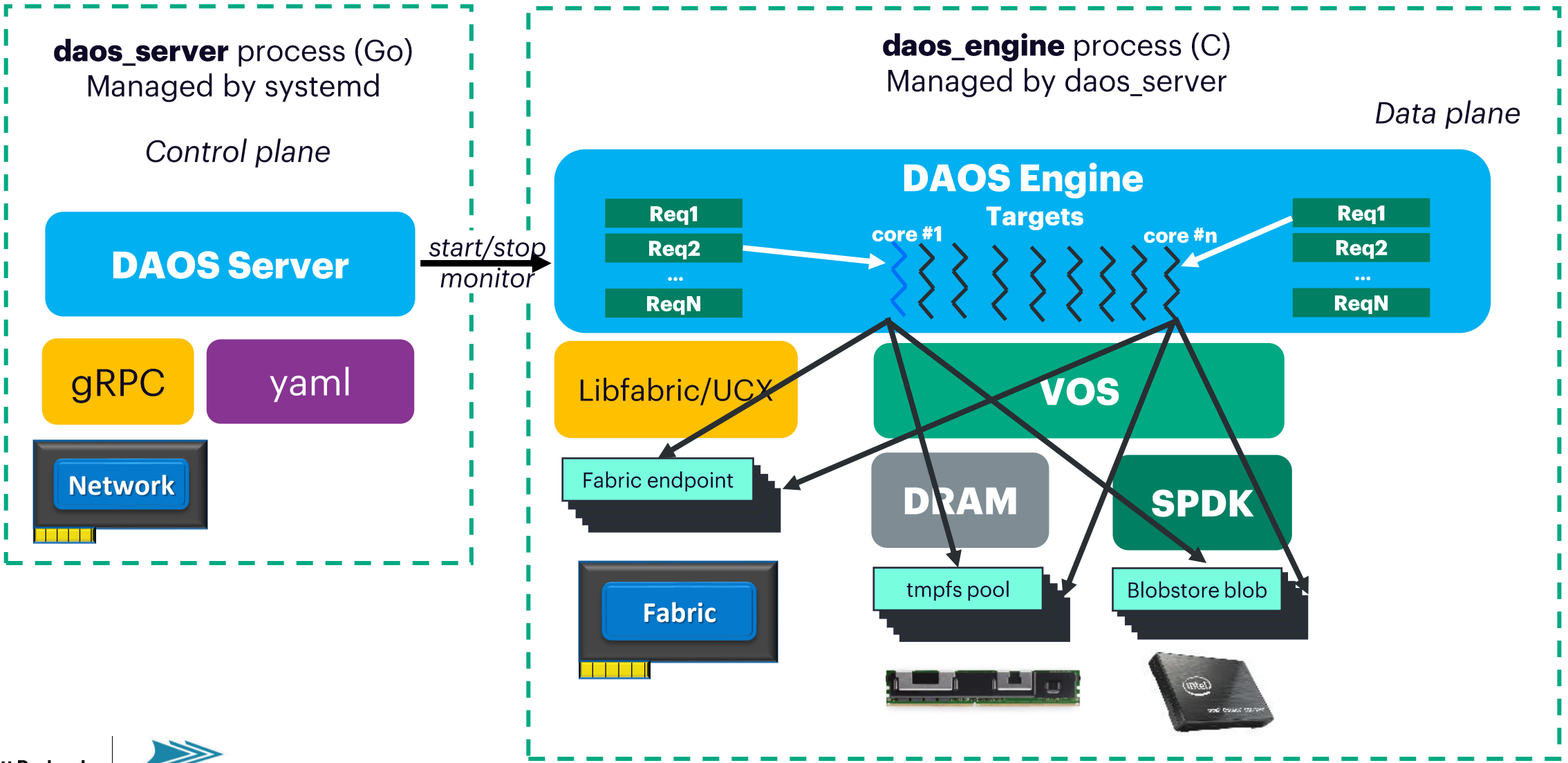
  intel · Argonne NATIONAL LABORATORY · Hewlett Packard Enterprise · Google · VDURA · MANGOBOOST
- DAOS development team moved to **HPE** in Dec-2024
  1. Continuing the open source development (→ daos.io)
  2. HPE Cray Supercomputing Storage „K3000“ product in 2026
     - LRZ „BlueLion“, OLCF-6 „Discovery“, …
- On **IDC**‘s Top Open Source Software Projects to Watch, 2026



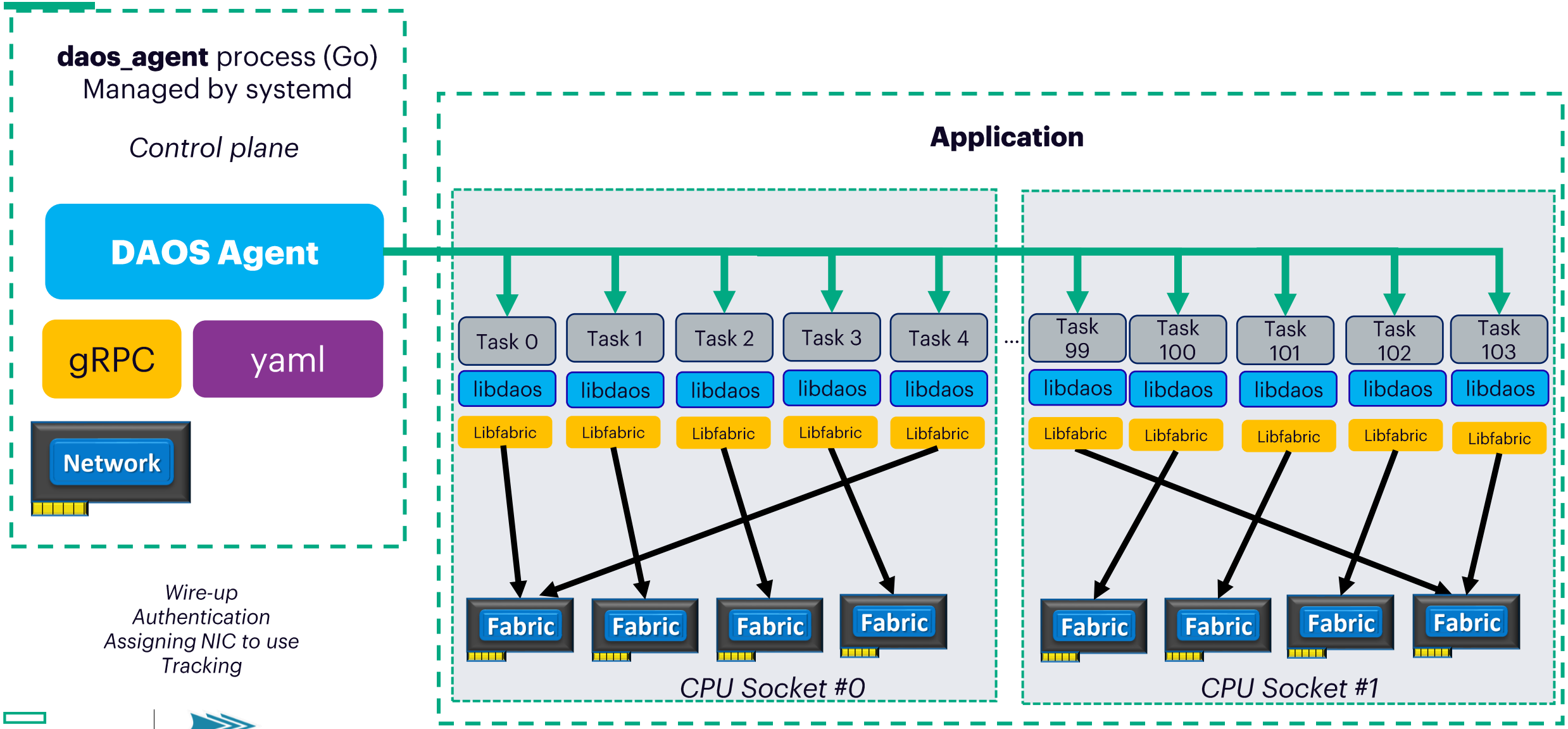
DAOS Server Service Structure



DAOS Server Service Structure – Using more targets increases performance



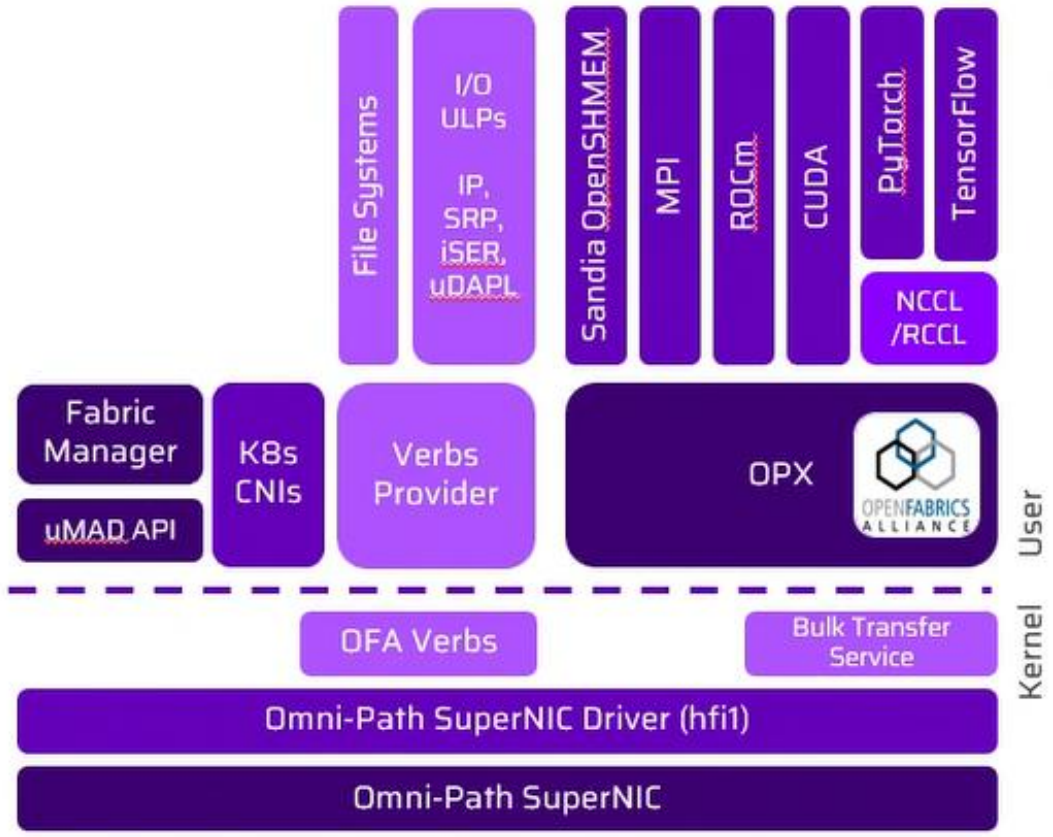
DAOS Client Structure – Every application process (e.g. MPI task) is a fabric endpoint



Wire-up
Authentication
Assigning NIC to use
Tracking

Hewlett Packard Enterprise | DAOS

Optimized and Open Software



Optimized Performance
Tight Software-Hardware Integration and Co-Design

Seamless Integration
Adoption of Open Standards and Full Stack Support

Trusted Deployment
Open-Source Libfabric Provider and Upstreamed Kernel Driver

Supported Operating Systems
RHEL/Rocky 9.8, 10.1, 10.2
Ubuntu 24.04 HWE, 26.04 LTS
SLES 16.0

CORNELIS

Performance Optimizations for CN5000

Enable large MTU size for TCP (IP-over-IB)

```
# vim /etc/opa-fm/opafm.xml
…
<MulticastGroup>
  <Create>1</Create>
  <MTU>10240</MTU>
  <Rate>100g</Rate>
  <!-- <SL>0</SL> -->
  <QKey>0x0</QKey>
  <TClass>0x0</TClass>
</MulticastGroup>

# ip a show ibs3d1|head -1
9: ibs3d1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu
10236 qdisc mq state UP group default qlen 1000
```

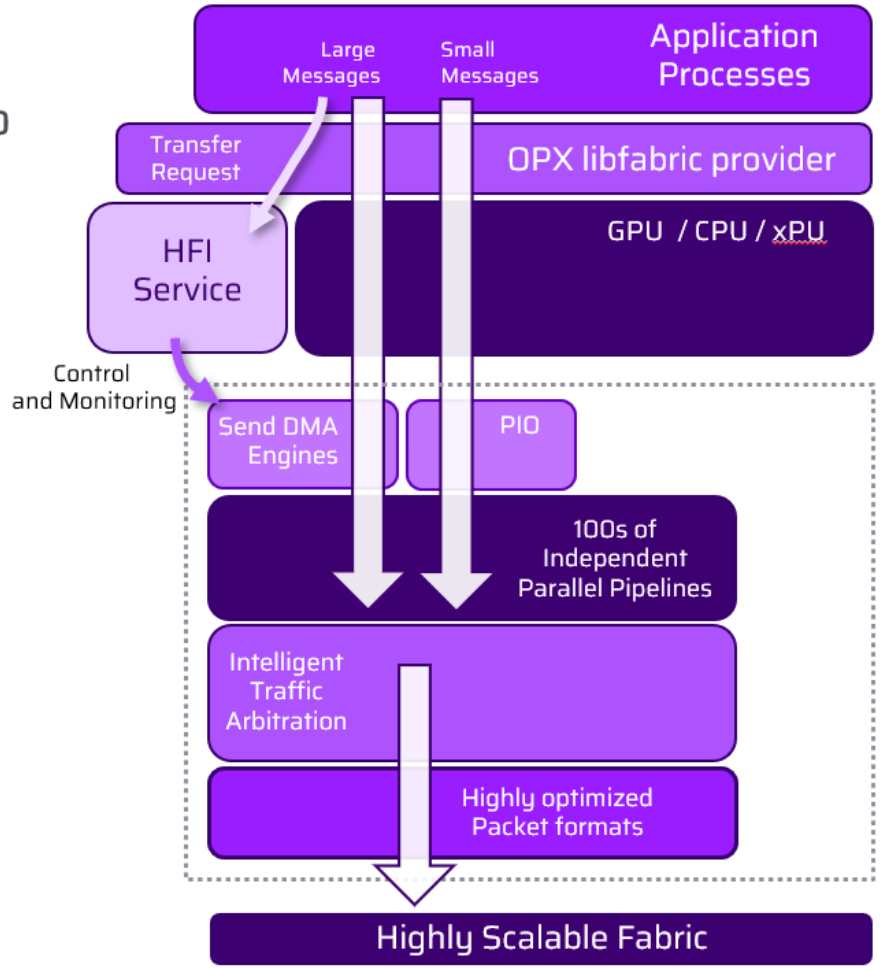
Enable Bulk Transfer Service (BTS) for VERBS

```
# ls -al /etc/modprobe.d/hfi1.conf #
should not exist after a fresh install

# echo "options hfi1 use_bulksvc=Y" \
    > /etc/modprobe.d/hfi1.conf

# dracut -f ; reboot

# cat /sys/module/hfi1/parameters/use_bulksvc
Y
```

Innovation: HFI Service

Previously called Bulk Transfer Service (BTS)…

- When sending large messages, it is most efficient to leverage the 16 Send DMA (SDMA) engines
- The OPXS software includes a kernel-level service called the HFI Service (HFISVC)
- HFISVC coordinates and optimizes the usage of the SDMA engines across all running processes
- Once a request has been made to the HFISVC, the process is free to continue computational progress
- HFISVC is currently used to optimize OPX and traditional RDMA verbs



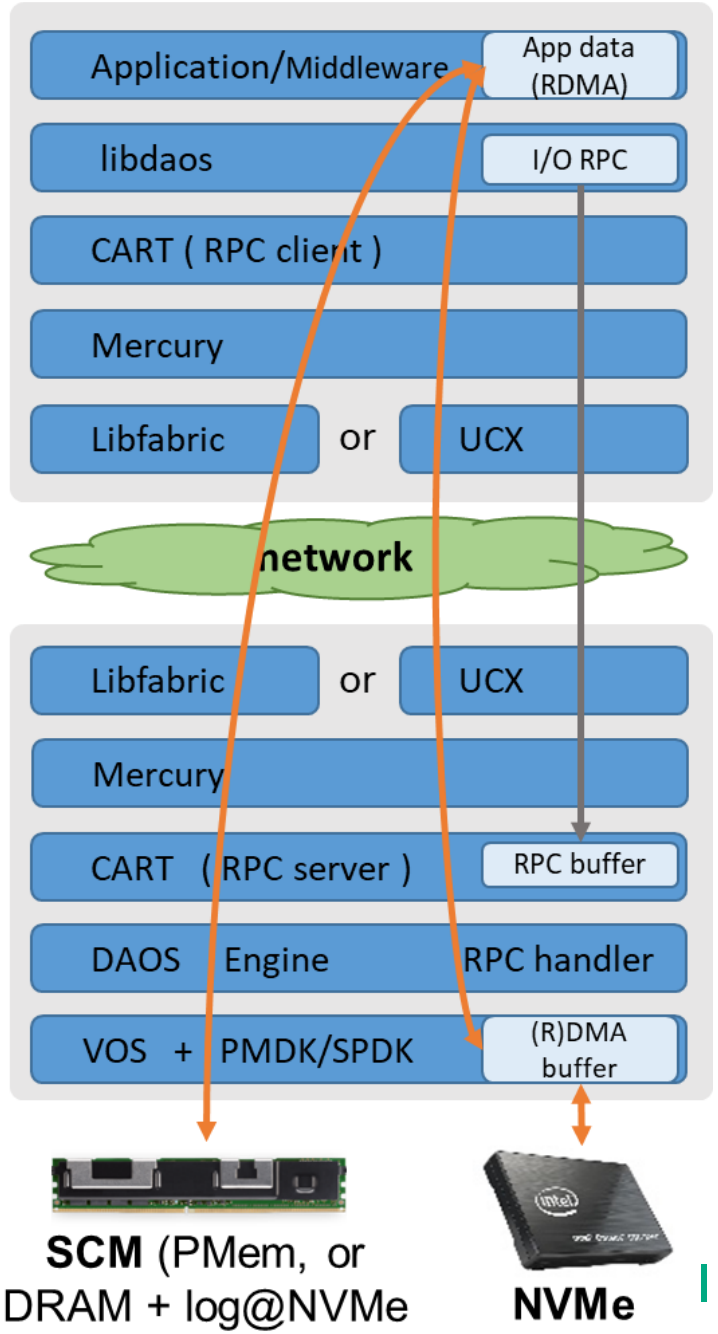
CORNELIS

Hewlett Packard Enterprise

DAOS

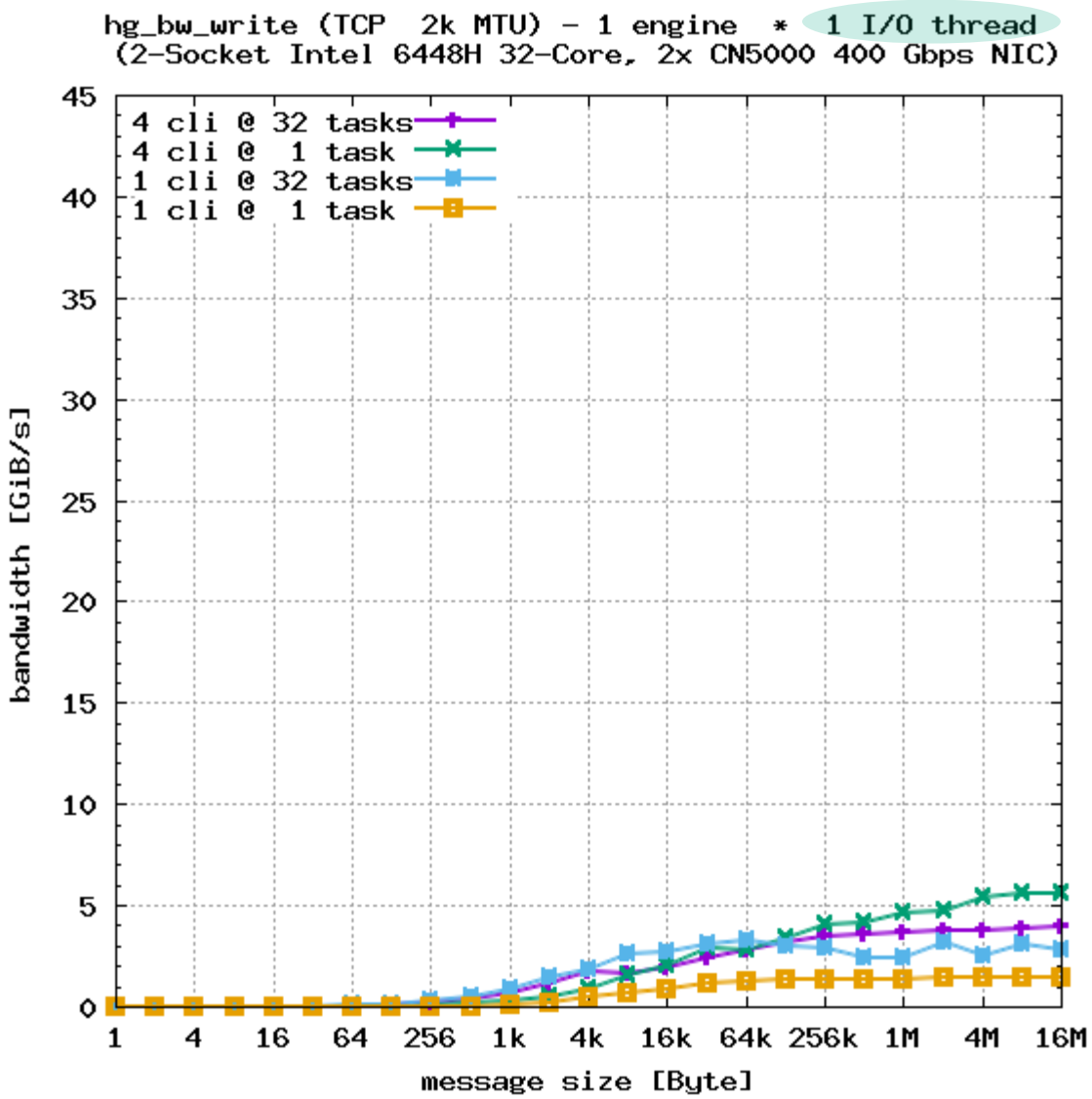
Benchmarking the DAOS Networking Stack

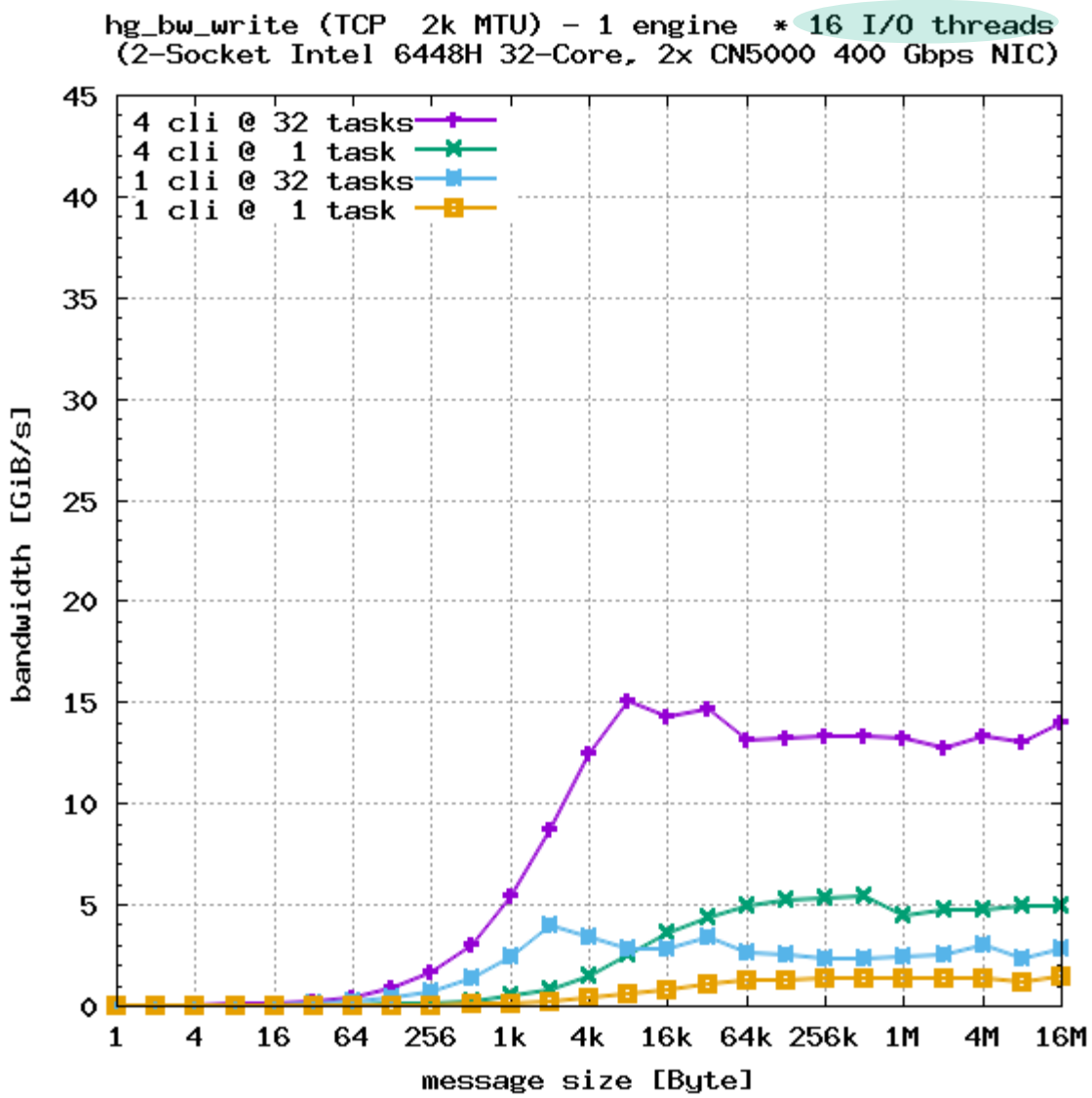
- HPC fabrics usually have **point-to-point** performance tools
  - `iperf` for generic TCP communication
  - `ib_write_bw -d hfi1_0 -i 2` for verbs
  - Libfabric has `fi_pingpong`
  - UCX has `ucx_perftest`
  - MPI benchmarks (IMB, OMB, …)
- But **storage traffic** is a **many-to-many, client/server** I/O pattern
- **Mercury** is the RPC layer used by DAOS
  - `na_write_bw` can test single client to many server-side endpoints
  - `hg_bw_write` is an **MPI-parallel** client application (to many server EPs)
- DAOS' CaRT (*C*ollectives *a*nd *R*PC *T*ransport) has `self_test`
- Actual I/O (e.g. `daos pool autotest`, running `IOR` or `mdtest`, …)



CN5000 Measurements (release 12.1.1)

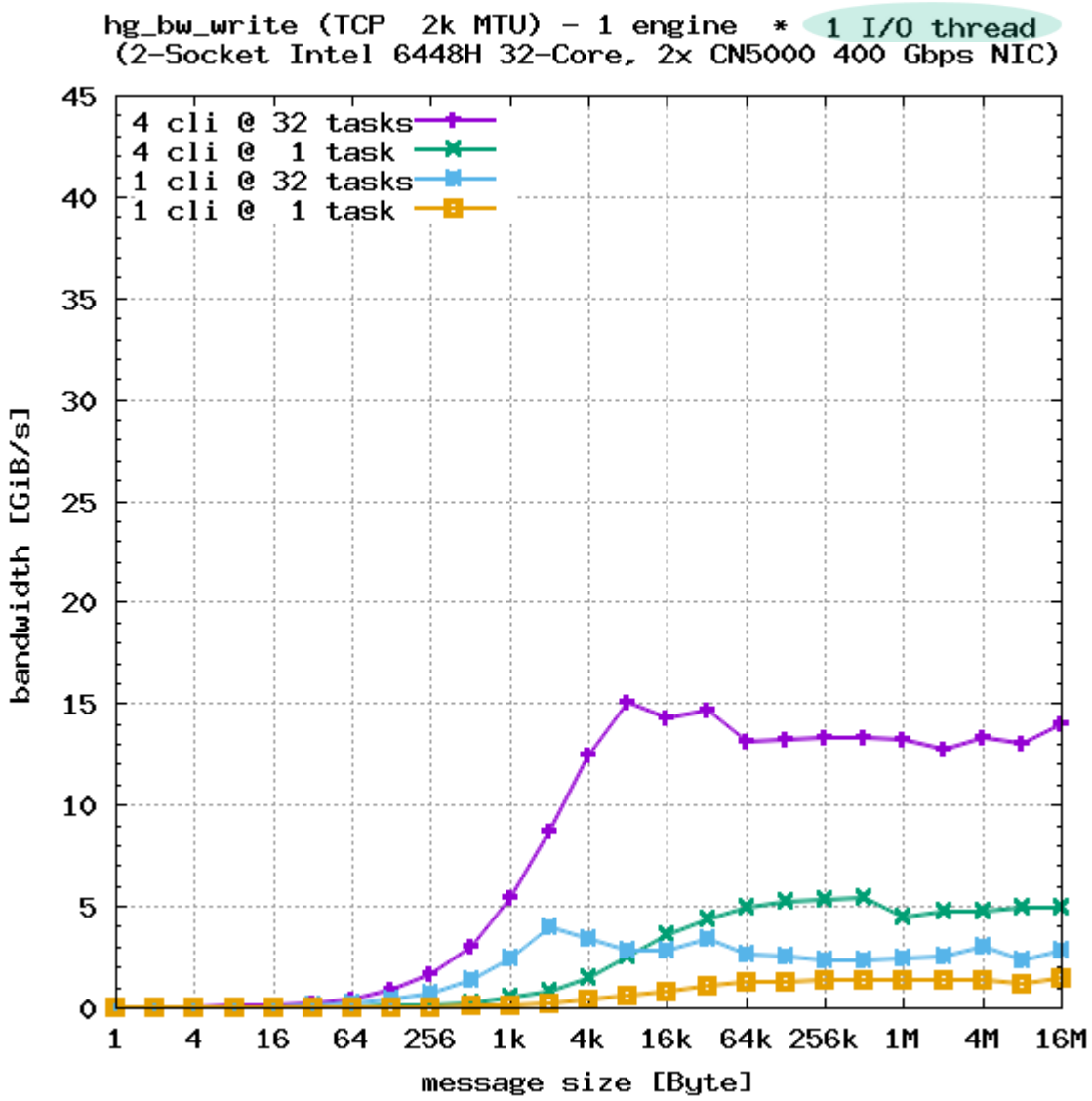
Single-engine hg_bw_write: libfabric TCP and MTU 2044

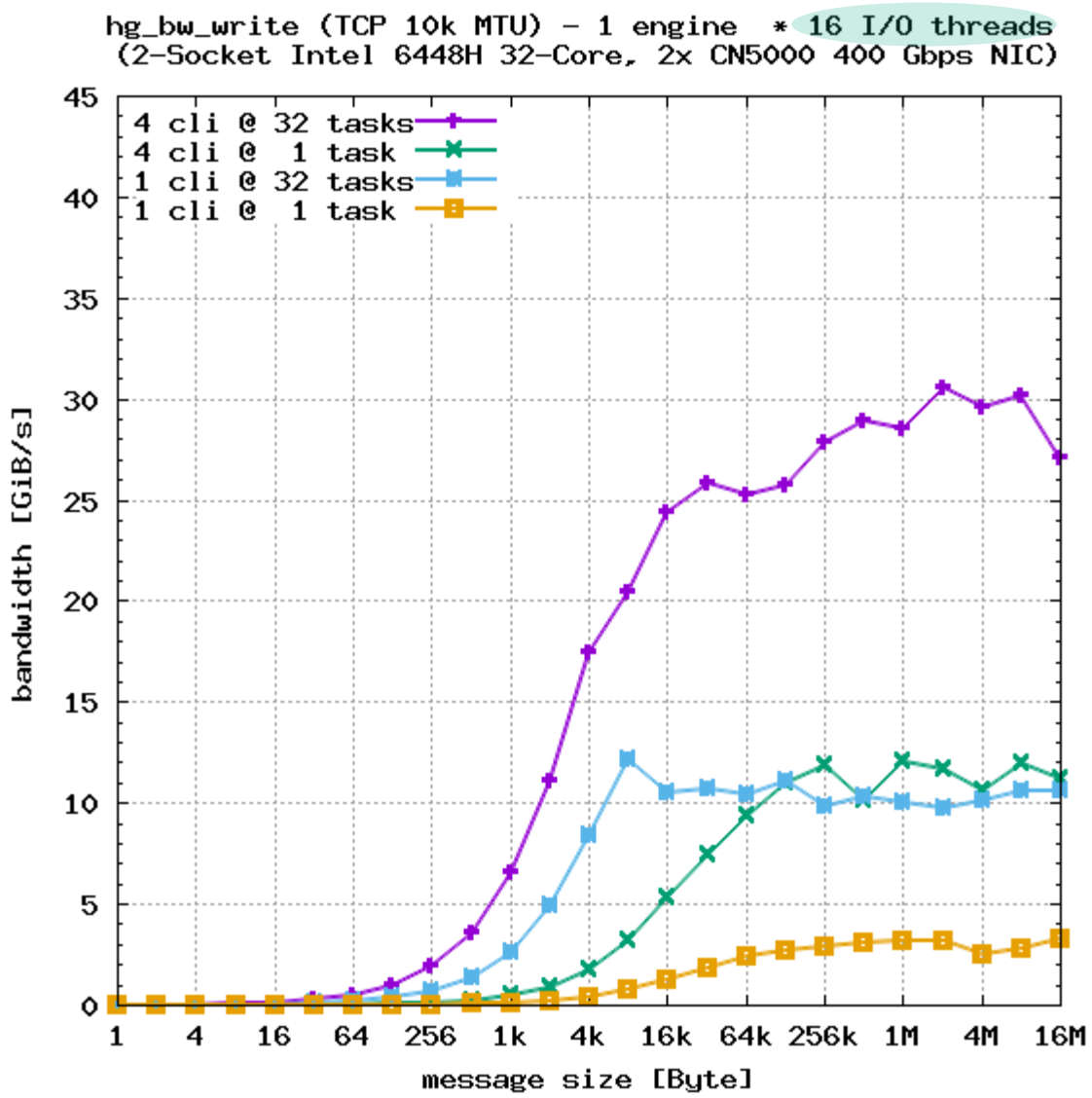




Hewlett Packard Enterprise | DAOS

Single-engine hg_bw_write: libfabric TCP and MTU 10236





Hewlett Packard Enterprise | DAOS

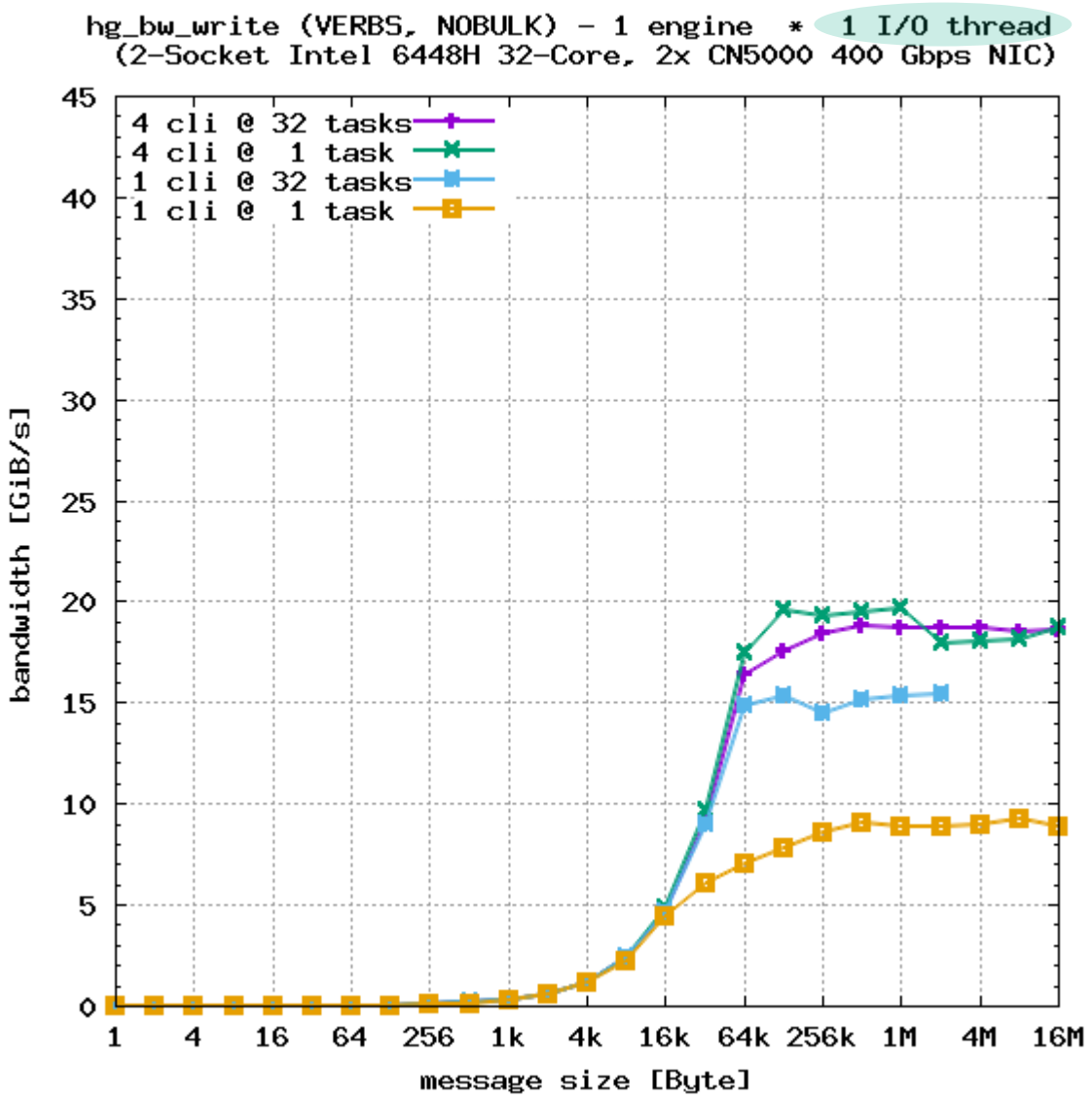
IO500-ISC24 : ZIB-Lise (19 Servers @ 2x OPA-100 ; 10 clients @ 1x OPA-100, 96 tasks)

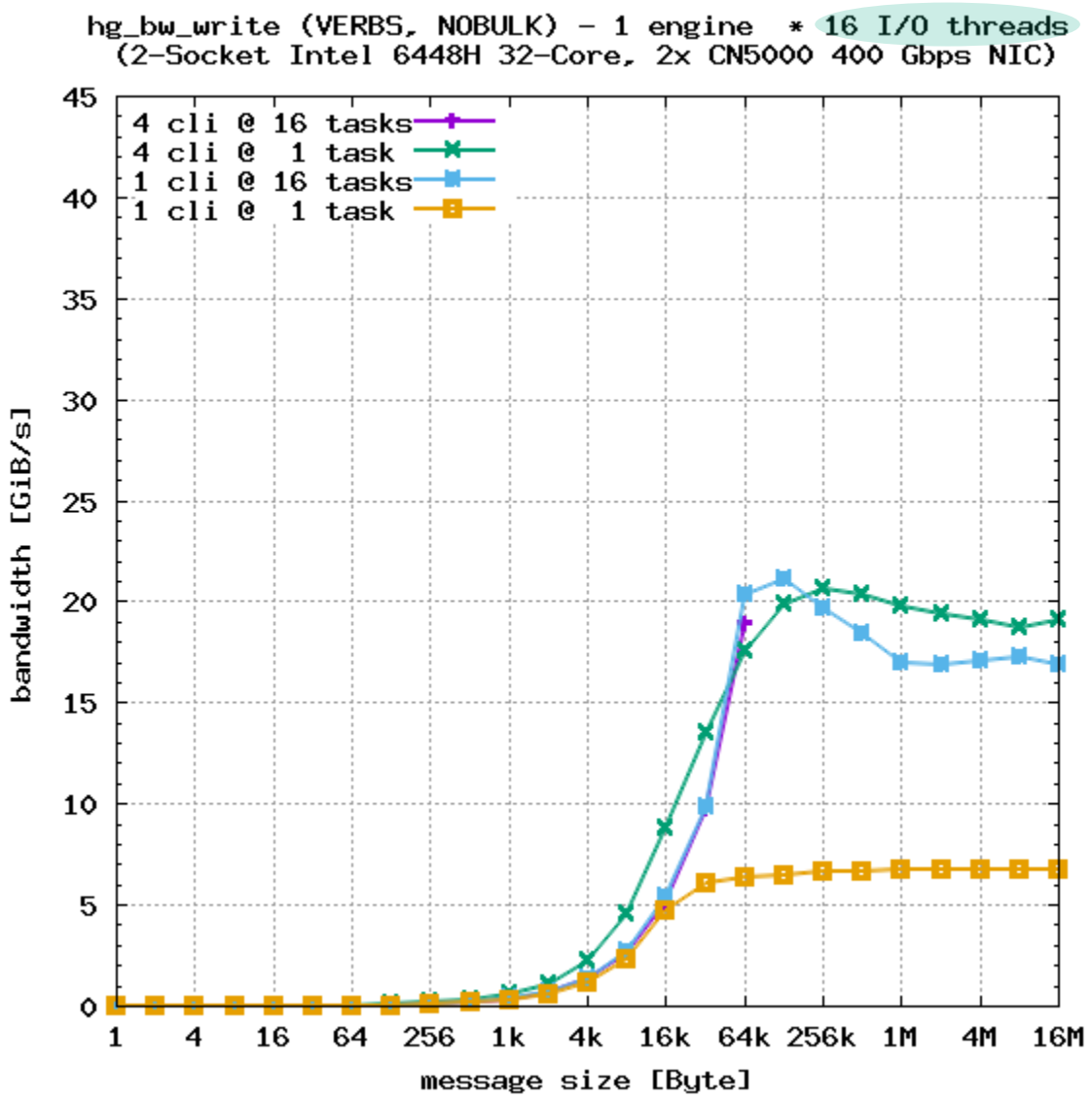
```
IO500 version io500-isc24_v3 (standard)
[RESULT]       ior-easy-write       90.518305 GiB/s : time 322.489 seconds
[RESULT]    mdtest-easy-write     1283.701416 kIOPS : time 306.948 seconds
[      ]            timestamp        0.000000 kIOPS : time 0.001 seconds
[RESULT]       ior-hard-write       48.087384 GiB/s : time 314.447 seconds
[RESULT]    mdtest-hard-write      450.497469 kIOPS : time 310.202 seconds
[RESULT]                 find     2881.264049 kIOPS : time 184.683 seconds
[RESULT]        ior-easy-read       71.368873 GiB/s : time 408.933 seconds
[RESULT]     mdtest-easy-stat     3894.997663 kIOPS : time 101.815 seconds
[RESULT]        ior-hard-read       57.505494 GiB/s : time 262.942 seconds
[RESULT]     mdtest-hard-stat     3787.388936 kIOPS : time 37.793 seconds
[RESULT]   mdtest-easy-delete      736.179489 kIOPS : time 536.722 seconds
[RESULT]     mdtest-hard-read     3190.894658 kIOPS : time 44.668 seconds
[RESULT]   mdtest-hard-delete      822.070343 kIOPS : time 209.762 seconds
[SCORE ] Bandwidth 65.012435 GiB/s : IOPS 1620.126639 kiops : TOTAL 324.543338
```

Hewlett Packard Enterprise | DAOS

For reference only: ZIB's "Lise" has an Omni-Path 100 fabric, not CN5000…

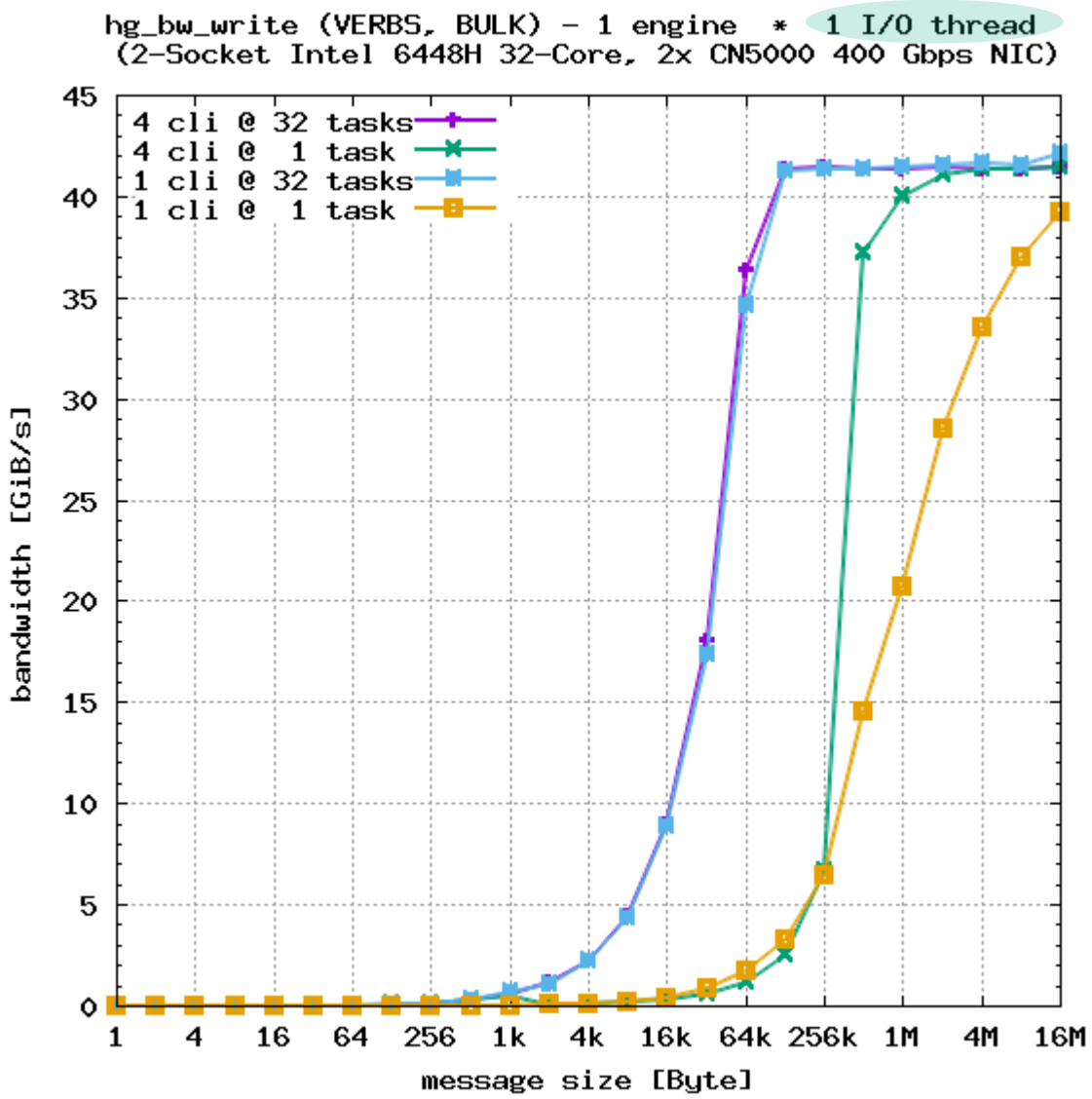
Single-engine hg_bw_write: libfabric VERBS, no bulk transfer service

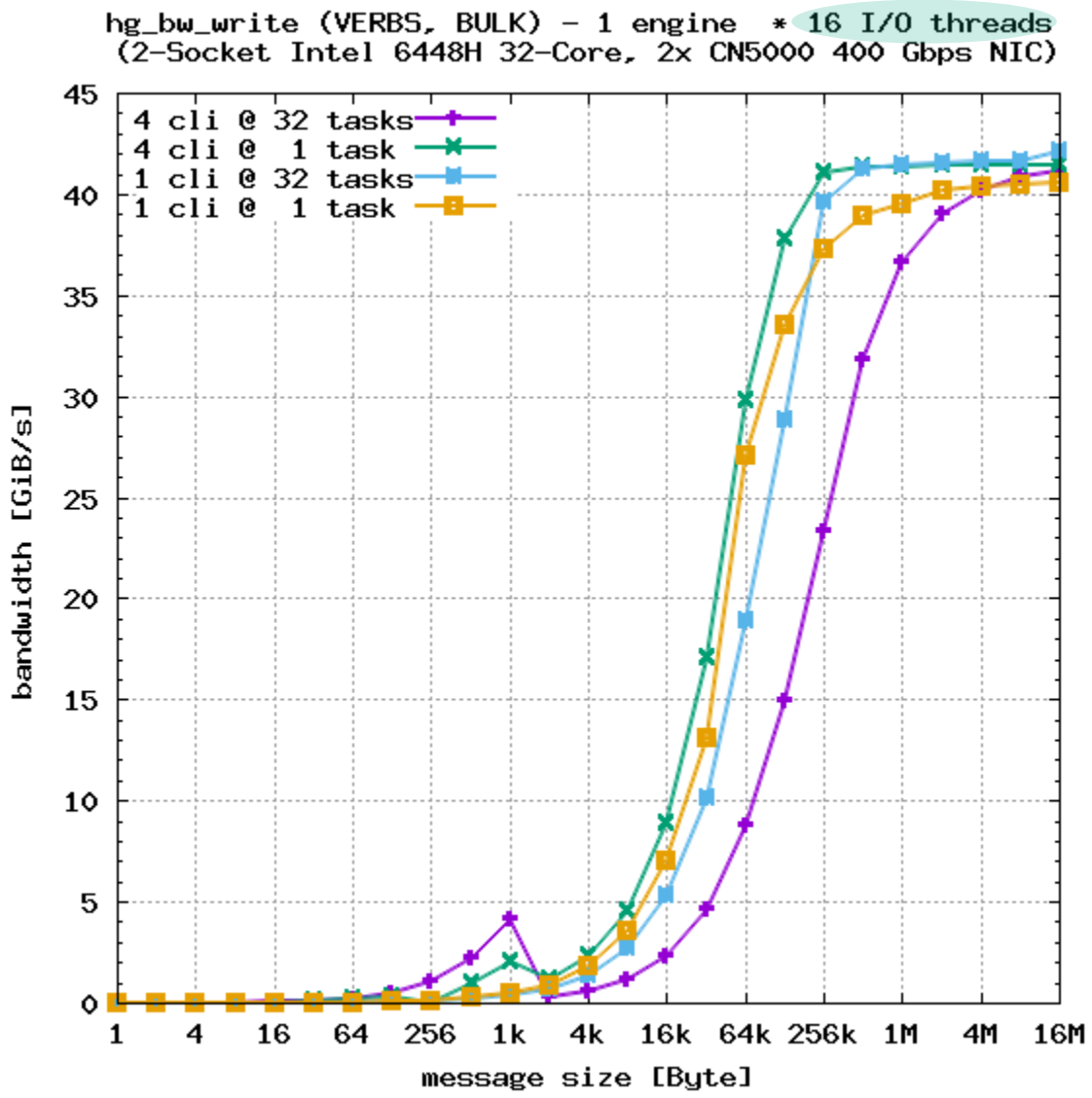




Hewlett Packard Enterprise | DAOS

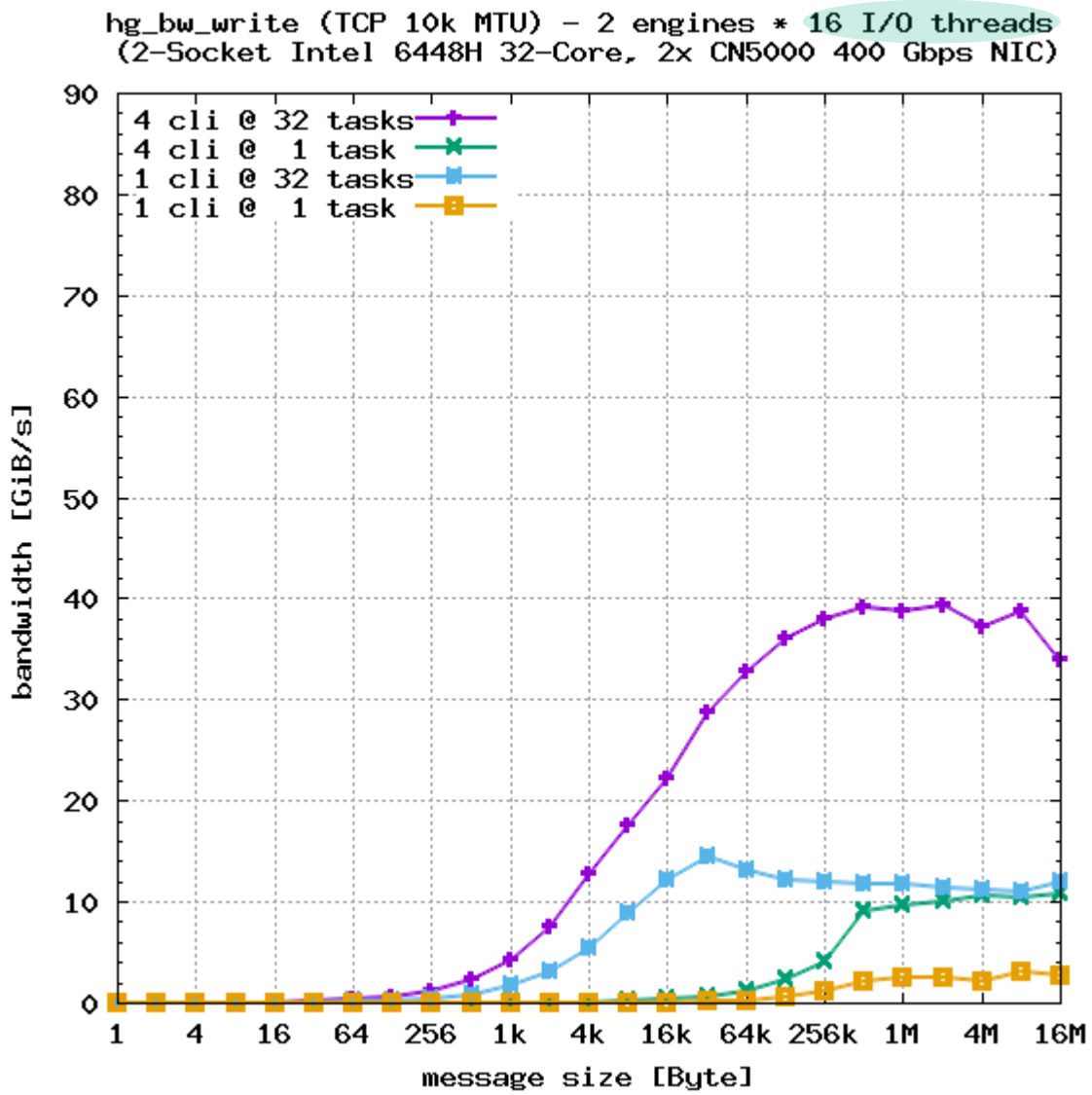
Single-engine hg_bw_write: libfabric VERBS, bulk transfer service enabled

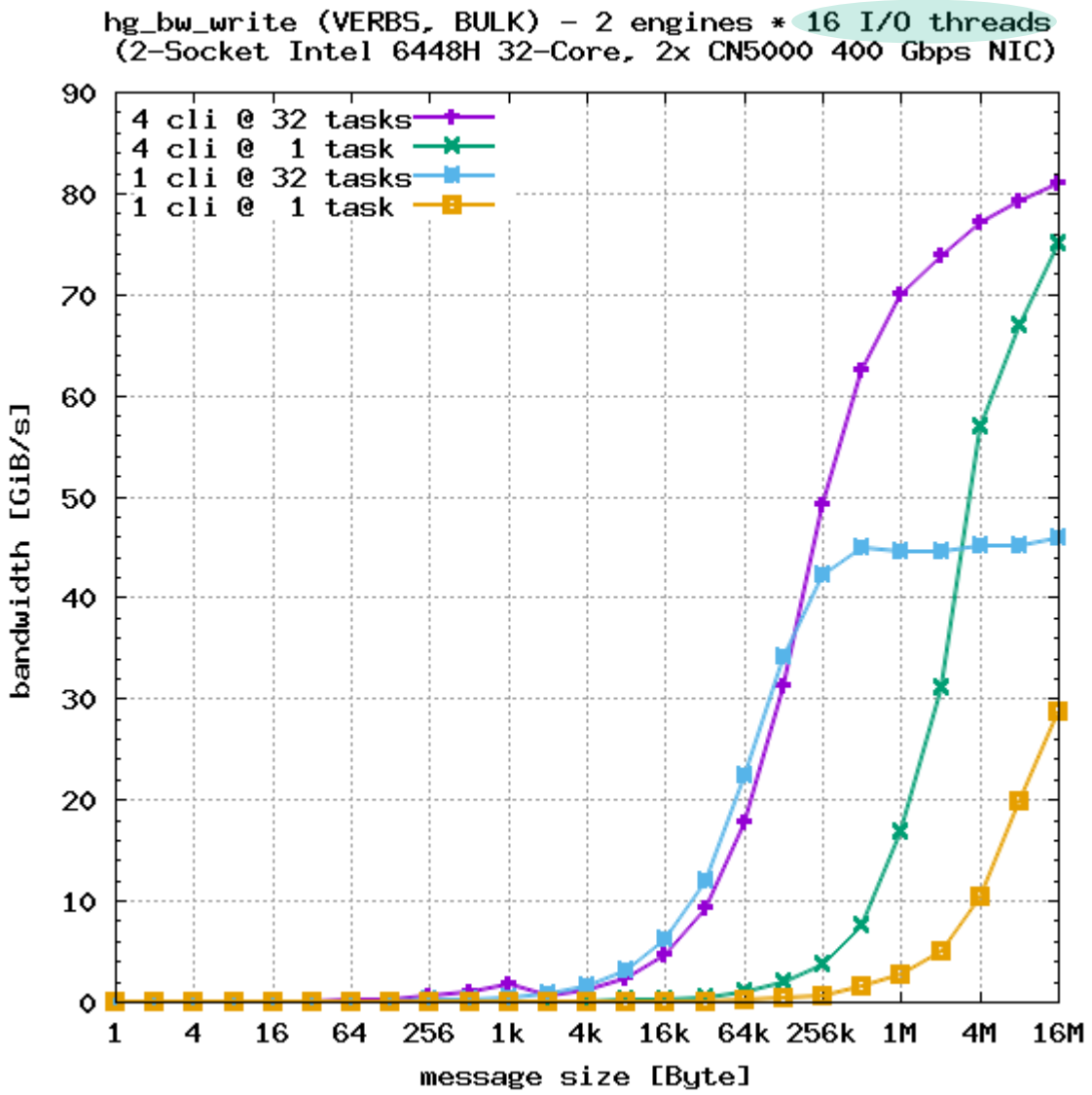




Hewlett Packard Enterprise | DAOS

Dual-engine hg_bw_write: TCP @ MTU 10236 (left) ; VERBS @ BTS (right)





Summary and Next Steps

- TCP provider (IP-over-IB) with 10k MTU was sufficient for OPA-100, but cannot saturate OPA-400

- VERBS provider now works well with Bulk Transfer Services (recently renamed to HFI Services)
  - Performance testing on Mercury level was reported here (with CN5000 release 12.1.1)
  - Need relatively large transfer sizes to saturate 400Gb link (further improvements exected)
  - DAOS-level testing showed some issues ➔ planning to re-run with CN5000 12.1.2 bugfix release

- Ultimate goal for DAOS is to use the OPX provider
  - Unlike other storage systems, DAOS is in user space (not a separate system process)
    - Cannot use different libfabric providers for MPI and for DAOS in the same user process
  - Working with Cornelis Networks to identify and close functional gaps in OPX for DAOS usage

Thank you