Hewlett Packard
Enterprise

Virtual DAOS User Group (vDUG'26)

Fabrics Support

Michael Hennecke

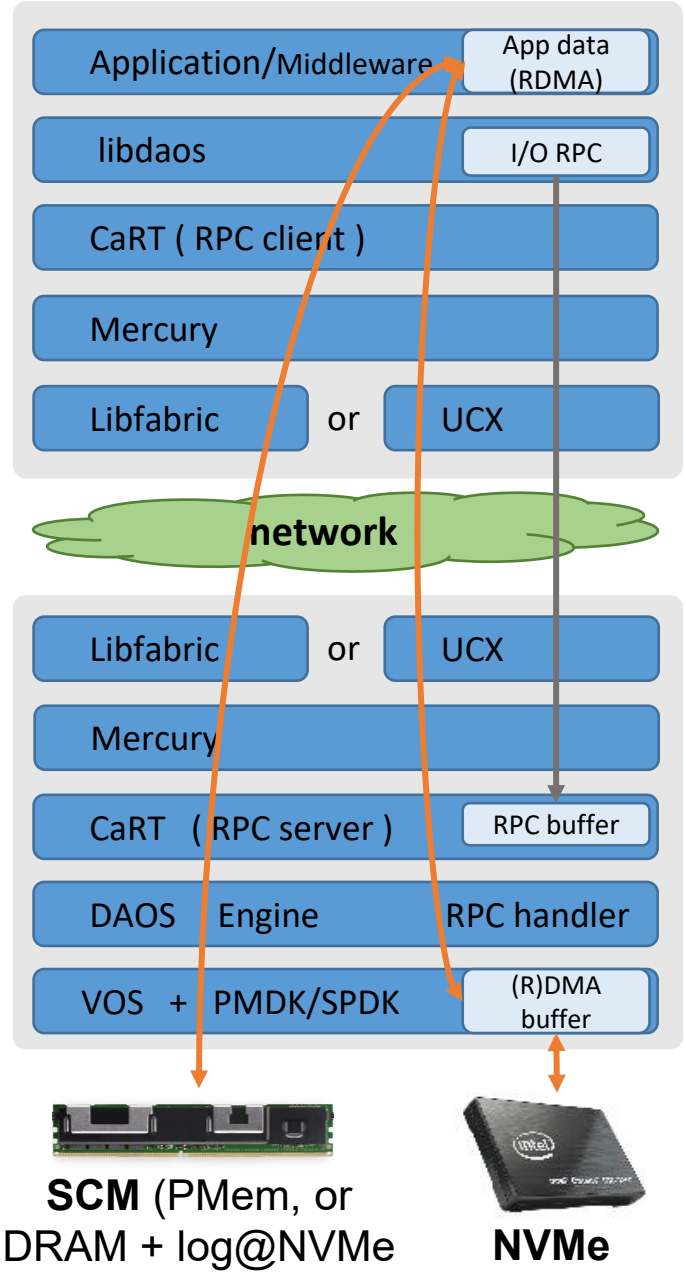
21-May-2026

Mercury Packaging Change

- DAOS uses Mercury as its point-to-point RPC layer
  - https://mercury-hpc.github.io/user/overview/
  - https://dx.doi.org/10.1109/CLUSTER.2013.6702617

- Until Mercury **2.4.0**, the `mercury` RPM shipped by DAOS **does** include the OFI/libfabric plugin
  - install `mercury-ucx` RPM to get the UCX plugin
- Since Mercury **2.4.1**, the `mercury` RPM shipped by DAOS **no longer** includes the OFI/libfabric plugin
  - install `mercury-libfabric` RPM to get the OFI/libfabric plugin
  - install `mercury-ucx` RPM to get the UCX plugin

- Plugin is loaded at runtime; failure if no plugin is found for the provider specified in daos_server.yml:
  - `na.c:570 na_plugin_scan_path()`
    `Could not open plugin (libna_plugin_ofi.so)`



HPE Slingshot (provider: ofi+cxi)

- Aurora optimisations: see CUG 2025 paper (https://doi.org/10.1145/3757348.3757350)

- Aurora @ Slingshot 200 (1024x servers) is #1 on the IO500 Production list since SC23 (run used 642 servers)
- First Slingshot 400 DAOS cluster (HPE-DL320-SS400) in IO500-SC25 list (8 servers @ 1-socket, 1 NIC)
  - See SC'25 DUG presentation for performance details
  - Validated and supported for HPE K3000 (with 400 Gbps and 200 Gbps NICs)

- Slingshot 400 needs at least Slingshot Host Software (SHS) **13.1.0** (02-Feb-2026)
  - supports EL 9.6, SLES 15 SP7
  - uses libfabric-2.2
  - includes VNI improvements for DAOS
- Testing with Slingshot Host Software (SHS) **14.0.0** (23-Mar-2026) is in progress
  - supports EL 9.7, SLES 15 SP7
  - uses libfabric-2.3.1-SHS14.0.0

- Slingshot Host Software installs libfabric in a non-default location, need to set the library search path:
  `LD_LIBRARY_PATH=/opt/cray/libfabric/2.3.1/lib64`
  - Needed in engine env in daos_server.yml, and in user environment on clients

NVIDIA InfiniBand and Ethernet (provider: ucx+dc_x)

- IB verbs has known scalability limitations (# of QP)
- For DAOS on InfiniBand, recommendation is to use `provider: ucx+dc_x`
  - Dynamically connected transport, shares QPs
  - Single-server IO500-SC25 entry: HPE-DL360-NDR
  - Validated and supported with NDR for HPE K3000
- RoCE with CX-7 adapters and SN5000 switches:
  - Validation of `provider: ucx+dc_x` is in progress for HPE K3000
- RoCE on non-NVIDIA hardware:
  - Juniper QFX5130 switches with CX-7 adapters planned after initial HPE K3000 product release
  - Non-NVIDIA adapters TBD

IO500 List Submission		oclass	DAOS Version	Fabric Provider	Servers	Clients [nodes * tpn = tasks]	IO500 Score	ior-rnd4K-read [GiB/s]	[M IOPS]
Research	SC23	SX	**2.4.0-2**	**ofi+verbs**	42	**90 * 72 =6480**	**4585**	--	--
Research	SC25	SX	**2.6.4-2**	**ucx+dc_x**	42	**192 * 112 =21504**	**6308 (up 38%)**	**189,1**	**49,6**
Production	SC23	EC_16P1	**2.4.0-2**	**ofi+verbs**	42	**90 * 72 =6480**	**2508**	--	--
Production	SC25	EC_16P1	**2.6.4-2**	**ucx+dc_x**	42	**192 * 112 =21504**	**3470 (up 38%)**	**218,3**	**57,2**

Cornelis Omni-Path Evaluation Status

The “native” Omni-Path `psm2` and `opx` providers cannot be used with DAOS (functional gaps)

- Talk to us, and to your Cornelis contacts, if you want to run DAOS in Omni-Path environments…

Omni-Path 100:

- `provider: ofi+tcp` works fine
- As with all high-speed TCP fabrics, MTU size matters…
  - Default is 2 kiB, increase it in /etc/opa-fm/opafm.xml
- ZIB’s “Lise” supercomputer was #3 on the IO500-ISC24 Ten-Node production list (#6 at SC25)

Omni-Path 400 (CN5000):

- `provider: ofi+tcp` works, but cannot saturate 400 Gbps (even with 10 kiB MTU)
- Cornelis has implemented a verbs API over CN5000, enabling `provider: ofi+verbs`
  - Together with the new CN5000 Bulk Transfer Service (BTS), this provides good performance
  - Functional DAOS validation is ongoing
- Detailed presentation on DAOS@CN5000 at IXPUG-ISC26 workshop

Thank you

michael.hennecke@hpe.com

Hewlett Packard Enterprise | DAOS