ALCF Site Update

Alex Kulyavtsev
Gordon McPheeters

4/1/2025

Topics To Be Covered

- ALCF filesystems and Clusters
- Aurora
- DAOS

ALCF Clusters mounting Lustre

Clusters:

- ***Theta*** retired. ***ThetaGPU*** cluster rebuilt as ***Sophia***.
- ***Sophia*** connects directly to core HDR network.
- ***Crux*** is CPU-only system for workloads not requiring GPU.
- ***Aurora, Polaris and Crux*** each have *separate* Slingshot 11 fabrics.

- Not shown in this report:
  - — ***few AI machines*** do not mount lustre servers
  - — TDS. **Gila** :: Sunspot ←→ Flare :: Aurora
    Tegu :: Sirius ←→ Eagle :: Polaris

	CPU	GPU	Computes	Gateways (LNet routers)
Sophia	AMD EPYC	NVIDIA DGX A100	24	Direct connection to HDR
Crux	AMD EPYC	N/A	256	26
Polaris	AMD EPYC	NVIDIA A100	560	50
Aurora	Intel Xeon Max	Intel GPU Max	10,624	100

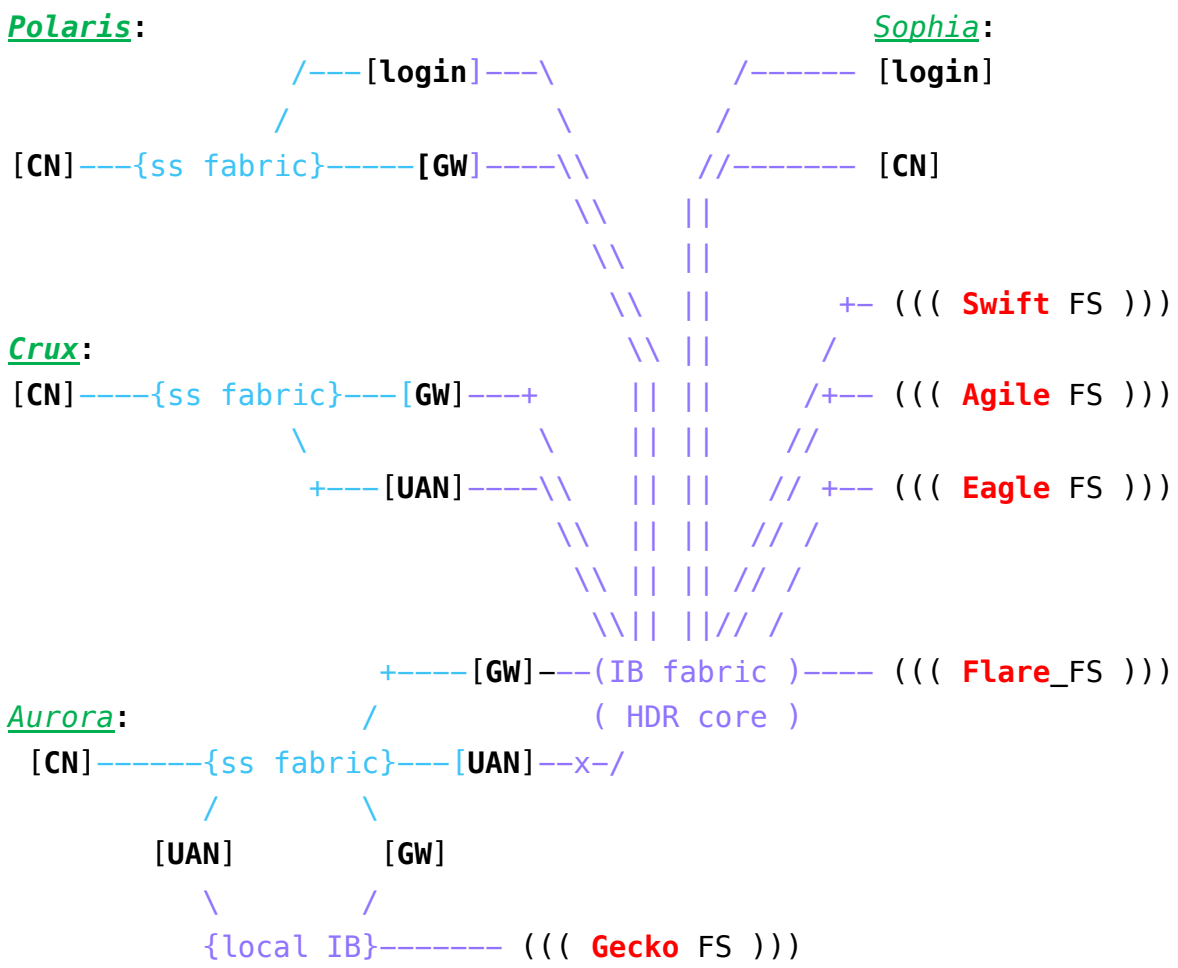
Cluster Components

- 40 TB/s HDR IB network is at the core of the network, common for all clusters.
- All Lustre storage systems reside on core IB fabric (see next slide)
  - except Aurora /home filesystem ***Gecko*** which is on small separate IB fabric
    - working to relocate Gecko to core IB fabric

- HPC clusters
  - *Compute Nodes* (CN) reside on Slingshot networks
    - Except Sophia which is fully on IB
  - *Gateways* (GW) :
    - Lnet on CN routed to storage through slingshot / IB gateways
    - No RDMA from CN to storage
  - user *Login* nodes or *User Access Nodes* (UAN) connected both to IB fabric for direct access to storage and also connect to slingshot
    - Aurora is different for now: it has small additional IB network used to isolate /home.

- Common services
  - Data Transfer Nodes (DTN) reside on core IB fabric

HPC clusters and Lustre File Systems



[CN] -- Compute Node
[UAN] -- User Access Node,
same as [login]
[GW] -- Gateway
{ss fabric} -- Slingshot-11 fabric

Production Lustre File Systems and Mounts

System	Vendor	Servers	Size, PB	OST	Clients	Mounted on	Purpose
Eagle	HPE	E1000	100	HDD	862	ALCF HPC clusters (*)	/projects
Agile	DDN	AI400X	0.242	SSD	864	ALCF HPC clusters (*)	/home
Flare	HPE	E1000	100	HDD	10,590	Aurora	/projects
Gecko	DDN	AI400X2	1 12	SSD HDD	10,590	Aurora	/home

(*) ALCF HPC Clusters refer:

Polaris, Crux, Sophia, Edith, Globus DTN, some service nodes

- In the works:
  - — mount all Lustre FS on all clusters
  - — working to move Gecko out of “Gecko’s HDR” to core HDR
  - — mount Gecko as global /home

HPE E1000: Flare(formerly Grand) and Eagle

Two identical HPE E1000 Lustre appliances

2 * (100 PB of available storage in 10 racks)

Each has

8,480 HDD in 80 HDD enclosures x 106 HDD
and SSDs for file bitmaps and Write Back Cache (WBC)
40 OSS/160 OST, GridRaid, ldiskfs
20 MDS/ 40 MDT, ldiskfs

- HDR (200 Gb/s) IB Network
  - two HDR interfaces per MDS (linux bond)
  - one HDR interface per OSS
- Software:
  - NEO 7.0 / HPE lustre 2.15.4
  - NEO 6.6 / HPE lustre 2.15.2

- E1000 Test and Development Systems (TDS)
- **Gila**: - mounted on Sunspot (128 Compute Node Aurora TDS)
- **Tegu**: Sirius (TDS for Polaris)

DDN EXAScalers and SFA18K

Lustre:

- *Agile* : AI400X
  - — /home FS for most clusters
  - — 242 TB usable (SSD) drive after capacity upgrade
  - — backed up to tape
- *Swift*: AI400X (older system)
  - — 120 TB usable (SSD)
  - — hot standby copy of /home on Agile
    - have higher space utilization
- *Gecko* : AI400X2
  - — /home for Aurora
  - — 1 PB SSD in two couplets
  - — 12 PB HDD connected to third couplet
  - — 8 HDR per couplet → presented as two HDR per VM.
  - — Exascaler 6.3.2 with DDN lustre 2.14.0_ddn191

HPSS Disk Cache

- 9 PB, SFA18K

Gecko : AI400X2

- used as a primary storage for Aurora during initial Aurora deployment
- served /home and /soft file systems to Compute Nodes as we scaled node count mounting lustre
- steep learning curve as we scaled up

- deployed on small HDR IB fabric (three TOR switches) to isolate Aurora during initial deployment from production clusters.
  - — few UAN and 8 GW connected directly to local Gecko IB.
  - — CN routed through slingshot.

- Initially deployed 12 PB HDD
  - — later added *separate* 1 PB SSD system to get more IOPS
  - — to minimize service interruption during deployment we moved data from HDD to new FS (rsync); reformatted HDD OST and added them to SSD system. Two short service interrupts: for final rsync and to add OSTs.

- Plan to relocate and connect through core IB fabric

Grand Migration, Fall 2023 / early 2024

- We considered too risky to mount main production file system Grand on a novel HPC system Aurora
- Got inspired by a report from Cameron Harr (LLNL) at LUG 2023:

  "Getting a Balanced, Full Production File System Migration to Work"

- Migrated all data (projects) from Grand to Eagle in 2023/H2 :
  - 14,852 TiB of data
  - 2,497,012,773 Files

  - data migrated to a separate subdirectory tree on Eagle
  - after the migration was completed, the migrated subtree was mounted as a Lustre Data Set on the same mount point /lus/grand as before to make it transparent for users

- **Migration process:**
- Early tests identified that Parallel File Utils at that time
  - — do not migrate named pipes, sockets
  - — create *separate* files for each hardlink

Grand Migration Process (cont.)

- Migrate data
  - copy, validate, delete original.

On a live system:

  - dwalk to identify special files and create lists
  - rsync *file lists* to create properly hardlinked files and create named pipes and sockets
  - dsync to replicate bulk of files to target
- Restrict access to files by end users for a set of projects (other continue to run) and execute
  - final dsync
  - validation (dcmp)
  - allow users to try to use migrated data
- Wait few weeks, delete original data on Grand
- It will be extremely helpful to have Parallel File Utils natively sync hardlinked files: it is *very* common pattern
- Migration requires running batch job with escalated privilege: executed on separate cluster.
- Upgraded Grand to the new major NEO/ ClusterStor version, mounted Grand on Aurora as a new FS Flare.

Aurora

ALCF completed Aurora Acceptance Tests at the end of 2024.

In early 2025 Argonne National Laboratory has released its Aurora exascale supercomputer to researchers across the world.

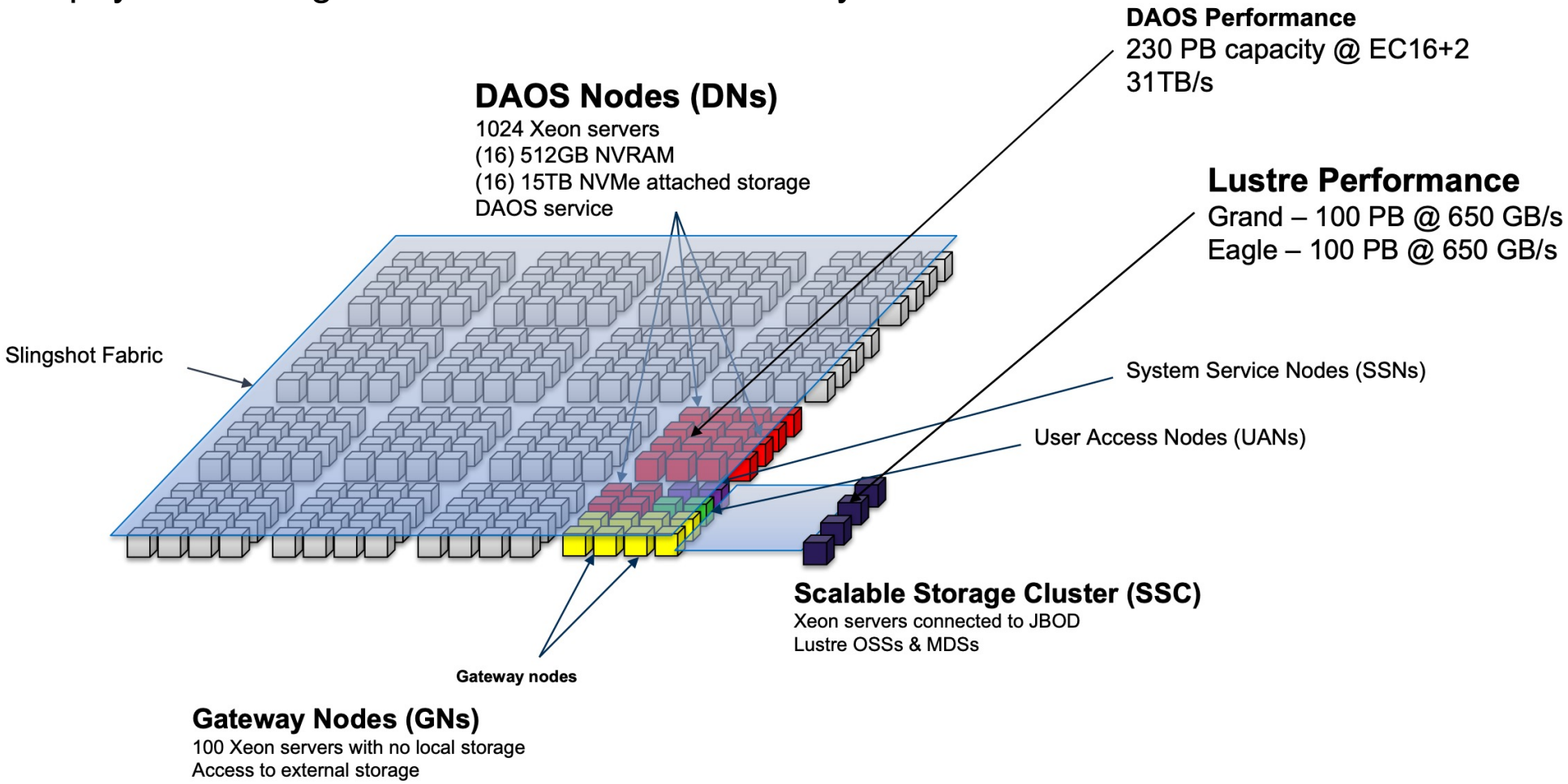
Compute nodes arranged

- 166 cabinets = 8 rows * 21 cabinets – 2 cabinets
  * 64 nodes/cabinet → 10,624 Compute Nodes
  32 slingshot L0 switches per water cooled cabinet

Compute Node	One Node	Total (* 10,624) Computes only
CPU	2x Intel® Xeon Max Series processors	21,248
GPU	6x Intel® Data Center GPU Max Series	63,744
Fabric Endpoints	8x Slingshot11	84,992

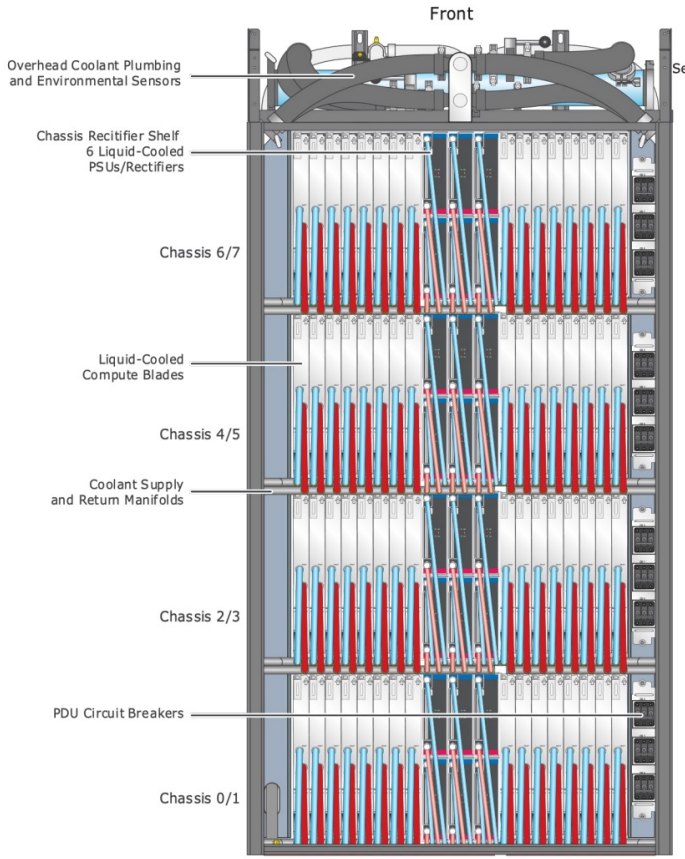
Aurora Storage Overview (ref [1])

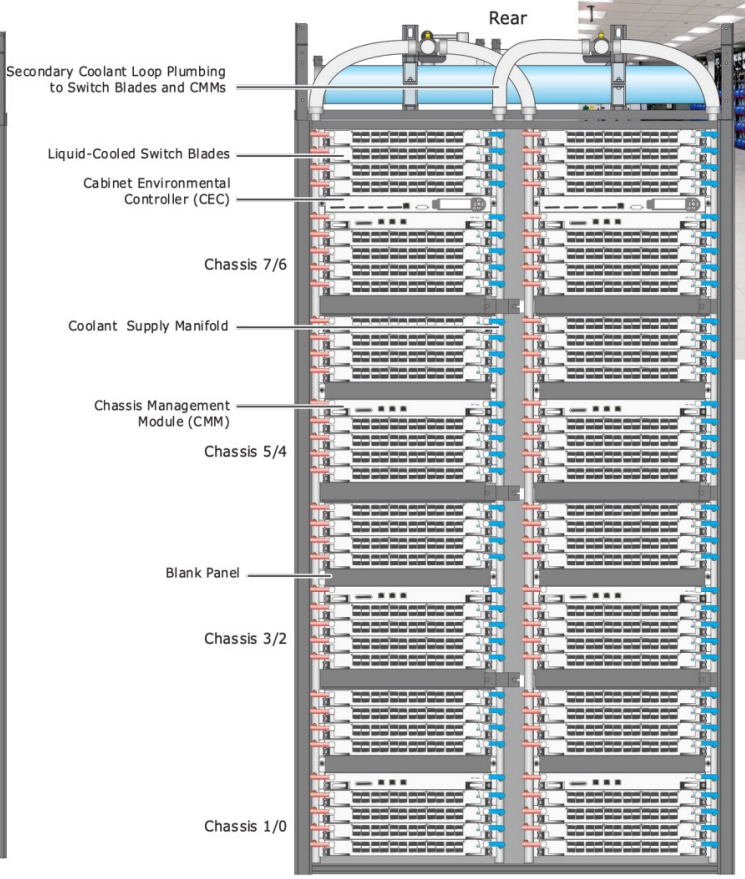
Note: the physical arrangement is different – 8 Rows by 21 Cabinets



Aurora Architecture Overview (ref. [1])

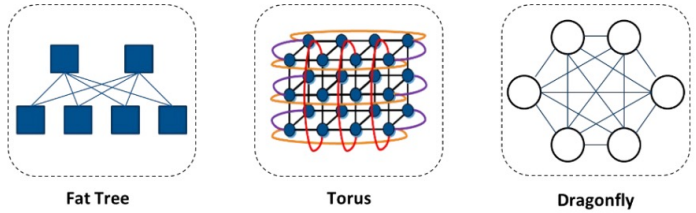
Aurora Cabinets at Argonne

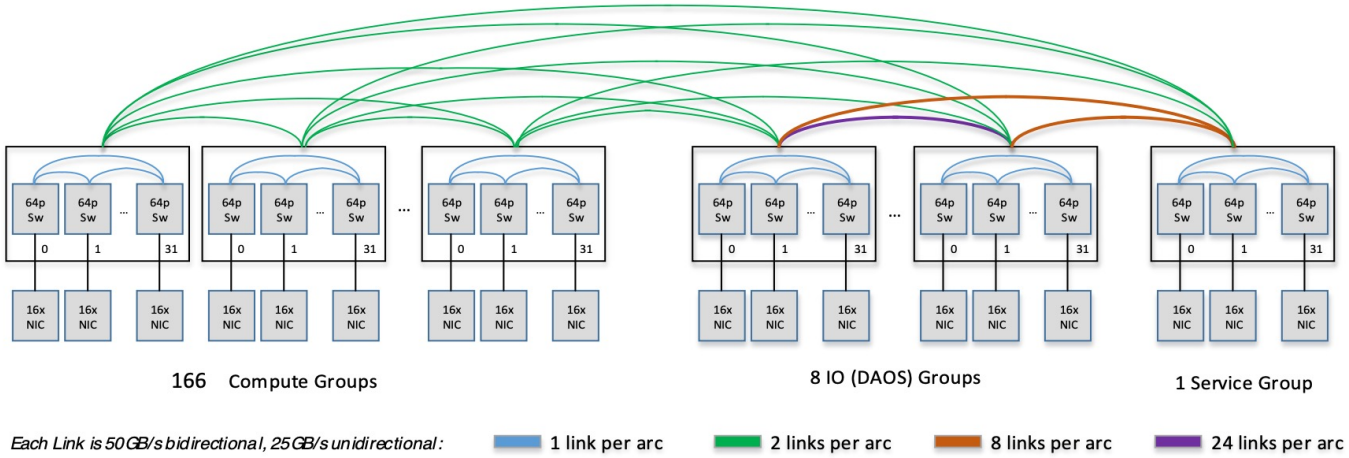


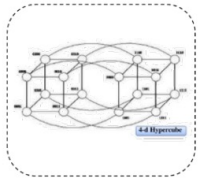


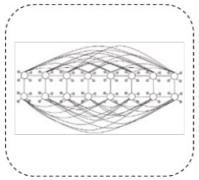
Aurora Architecture Overview (ref. [1])

Interconnect Topology






Hypercube

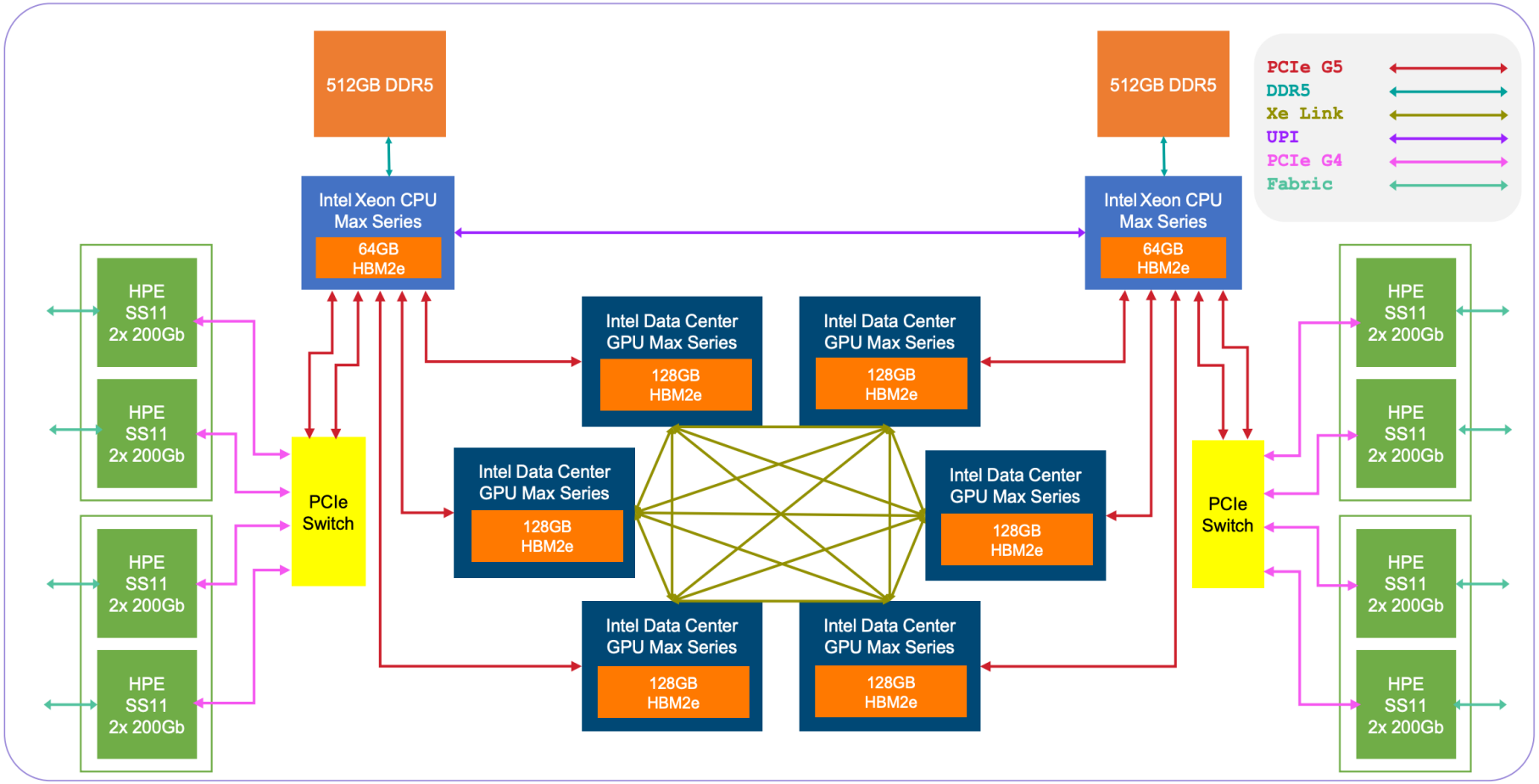

HyperX

- 1-D Dragonfly Topology - 175 total groups (166 compute + 8 IO + 1 Service),
- All the global links are optical, all the local links in compute groups are electrical
- 2 global links between any two compute groups
- 24 links between any two IO groups, 8 links between the Service group and each IO group
- Total injection bandwidth: 2.12PB/s
- Total bisection bandwidth: 0.69PB/s

[1] Overview on Aurora Exascale Compute Blade, Servesh Muralidharan

Aurora Architecture Overview (ref. [2])

Aurora Exascale Compute Blade – Data Flow



Aurora LNet

- Split lnet configuration into several files in /etc/lustre/ per network, router group, and global lnet setting.
- Compute Node in `lnet.service` :

```
/sbin/modprobe    lnet
/usr/sbin/lnetctl lnet configure
/usr/sbin/lnetctl import       /etc/lustre/lnet.global.conf
/usr/sbin/lnetctl import --add /etc/lustre/flare.net.conf
/usr/sbin/lnetctl import --add /etc/lustre/flare.route.conf
/usr/sbin/lnetctl import --add /etc/lustre/gecko.net.o2ib25.conf
```

- Separate hardware and Lnet networks used to route to different servers to simplify troubleshooting

```
tcpXX00(hsn0,hsn4)             → gecko GW  [8] → gecko (/home )
tcpYY00(hsn2,hsn3, hsn6,hsn7) → flare GW [48] → flare (/project )
```

- Gateway, network facing computes:

```
- net type: tcpXX00
  local NI(s):
    - interfaces:
        0: hsn0
      tunables:
        peer_timeout: 0
        peer_credits: 32
        peer_buffer_credits: 0
        credits: 65536            ←---- talk to 10 K computes
      lnd tunables:
        conns_per_peer: 1         ← set to 4 increase performance; dropped back to :1 to get stability
```

Scaling up

- Aurora successfully mount Flare and Gecko file systems on 10 K compute nodes.
- Most issues we faced were due to *COMBINATION* and/or *INTERPLAY* of *MULTIPLE* factors related to multirail, routing through GW, lnet discovery, corner cases during error processing and the scale of the system.
- Example
  - RPC failed because client panicked. Server considers options :
    - `[A] server ---→ GW ---→ Client [X]`
    - `[B] server ---→ GW --X→ Client`
    - `[C] server --X→ GW ---→ Client`
  - and server mistakenly decreases GW health [C]. We had to permanently set the router health percentage to 0.
- Example:
  - Server powered down. Some computes extremely slow to mount peer as they do lnet discovery and wait for RPC to timeout before trying peer.
  - patches with asynchronous lnet discovery and trying peer faster addressed the issue.
- Changes in upstream kernel may result in new scaling restrictions: OBD count got limit to 8192 nodes till lustre got a patch. Got surprised as other HPC clusters scaled to 20 K mounts while ago.
- Lnet / lustre has many interrelated parameters (multiple timeouts; peer credits on different networks). We have to resort to consult Subject Matter Expert to verify changes (Thank you, Chris !!!).

Aurora DAOS

- Distributed Asynchronous Object Store based on Persistent Memory and NVMe
- Designed as the high performant primary campaign storage for Aurora [5]
- Currently Number One on *io500.org*
  - “Production SC24 List” and
  - “10 Node Production SC24 List”

- 64 racks hosting 1024 DAOS server nodes (NCN, Air cooled)
- 230 PB, 25 TB/sec

Aurora DAOS

- Three DAOS Object Stores created on Aurora
  - 128 node daos_user, production cluster. DAOS 2.6.3 rc3
  - 128 node daos_perf pre-production cluster dedicated to prove user applications
  - ~700 node cluster still with HPE to complete DAOS testing with high server count
  - DAOS is targeting final production cluster approximately 800 nodes
  - as HPE testing completes nodes will be used to grow daos_user to its final size
- name space integration with Lustre [4]
  - DAOS integration with Lustre uses the Lustre foreign file/dir feature
    - Server: LU-11376 “Special file/dir to represent DAOS Containers.”
    - Client: LU-12682 adds DAOS specific support to the Lustre foreign file/dir feature.
- eight Aurora Gateway node allocated as DAOS data movers
  - move DAOS containers between DAOS ← → Lustre
  - data mover node has DAOS and Lustre mounted
  - data movement will be accomplished by user PBS jobs executing mpifileutils to copy data

DAOS Performance at Large Scale

- Intel testing at large server counts and large client counts prior to acceptance testing

- IOR
  - 793 daos servers (2 engines each)
  - 8,192 compute nodes (65,536 endpoints)
  - 1 MiB EC cell size
  - EC_16P2GX

  - IO-500 ”easy” configuration using DFS
    - 16 MiB block size
    - 22,489 GiB/s write
    - 22,618 GiB/s read

  - IO-500 “hard” configuration using DFS
    - 21115360 chunk size
    - 1,223 GiB/s write
    - 8,729 GiB/s read

Data graciously provided by Dalton Bohning, Intel DAOS

- mdtest
  - 775 daos servers (2 engines each)
  - 8,192 compute nodes (64,536 endpoints)

  - DFS with files configured RP_3G1 and directories RP_3GX

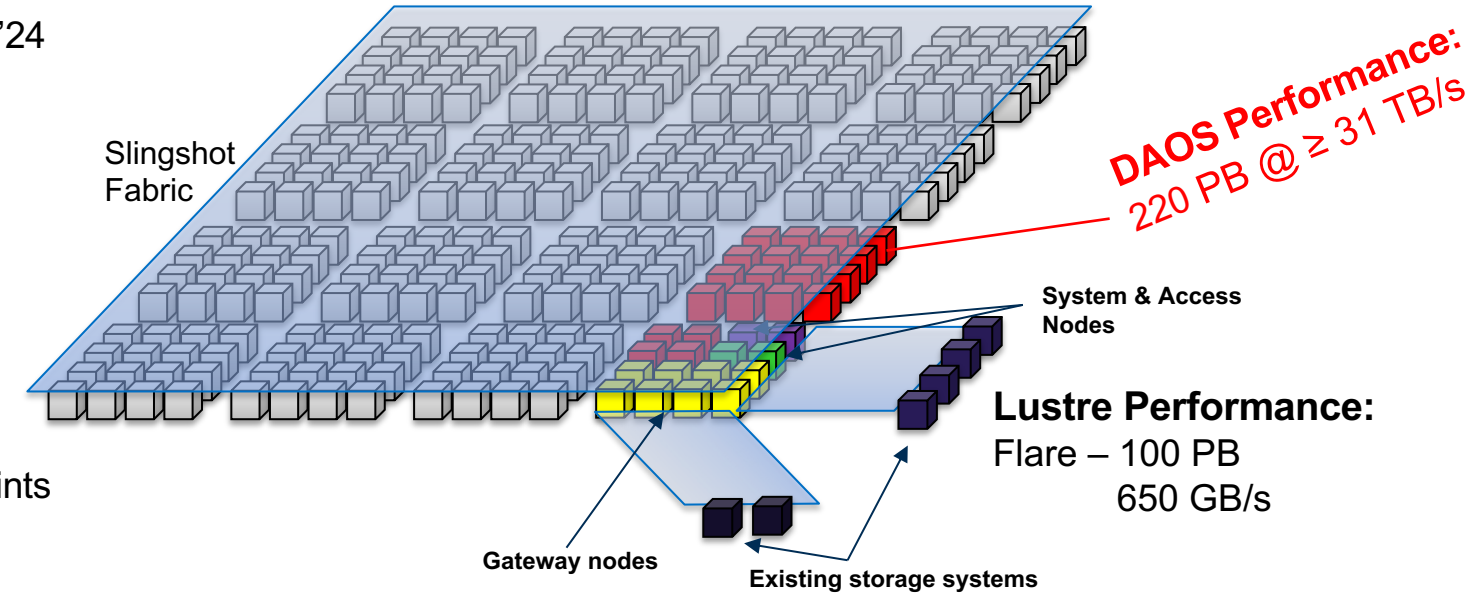
  - IO-500 “easy”
    - Create: 48,438 k/ops
    - Stat: 225,635 k/ops
    - Remove: 31,407 k/ops

  - IO-500 “hard”
    - Create: 35,626 k/ops
    - Stat: 237,749 k/ops
    - Remove: 35,312 k/ops

Aurora DAOS Performance

- System configuration during Aurora Acceptance Oct'24
  - daos_perf
    - 800 daos server nodes
    - Used for acceptance testing
  - daos_user
    - 128 daos server nodes
    - General access for Intel and Argonne users

- Both systems running v2.6.2-tb2 and SHS 11.0.0
- Testing up to 8000 compute nodes or 64,000 endpoints
- Core focus is on hardening and bug fixing



num CN nodes	ppn	num DAOS nodes	interception lib	daos backend mode	access	type	transfer size	block size	agg file size	Write TB/s	Read TB/s
256	16	128	libpil4dfs.so	posix	single-shared-file	independent	16MiB	2 GiB	8TiB	5.23	4.47
256	16	128	libpil4dfs.so	mpiio-daos:/	single-shared-file	independent	16MiB	2 GiB	8TiB	4.64	4.52
256	16	128	libpil4dfs.so	dfs	file-per-process	independent	16MiB	2 GiB	8TiB	5.51	4.12
256	16	128	libpil4dfs.so	dfs	file-per-process	collective	16MiB	2 GiB	8TiB	5.54	4.11
256	16	128	libpil4dfs.so	dfs	single-shared-file	independent	16MiB	2 GiB	8TiB	5.73	4.44
256	16	128	libpil4dfs.so	dfs	single-shared-file	collective	16MiB	2 GiB	8TiB	5.73	4.37

SX – no redundancy, roofline performance

DAOS Performance

Good Overview [3] of per-client performance and tuning,

results on small cluster in PDSW'24:

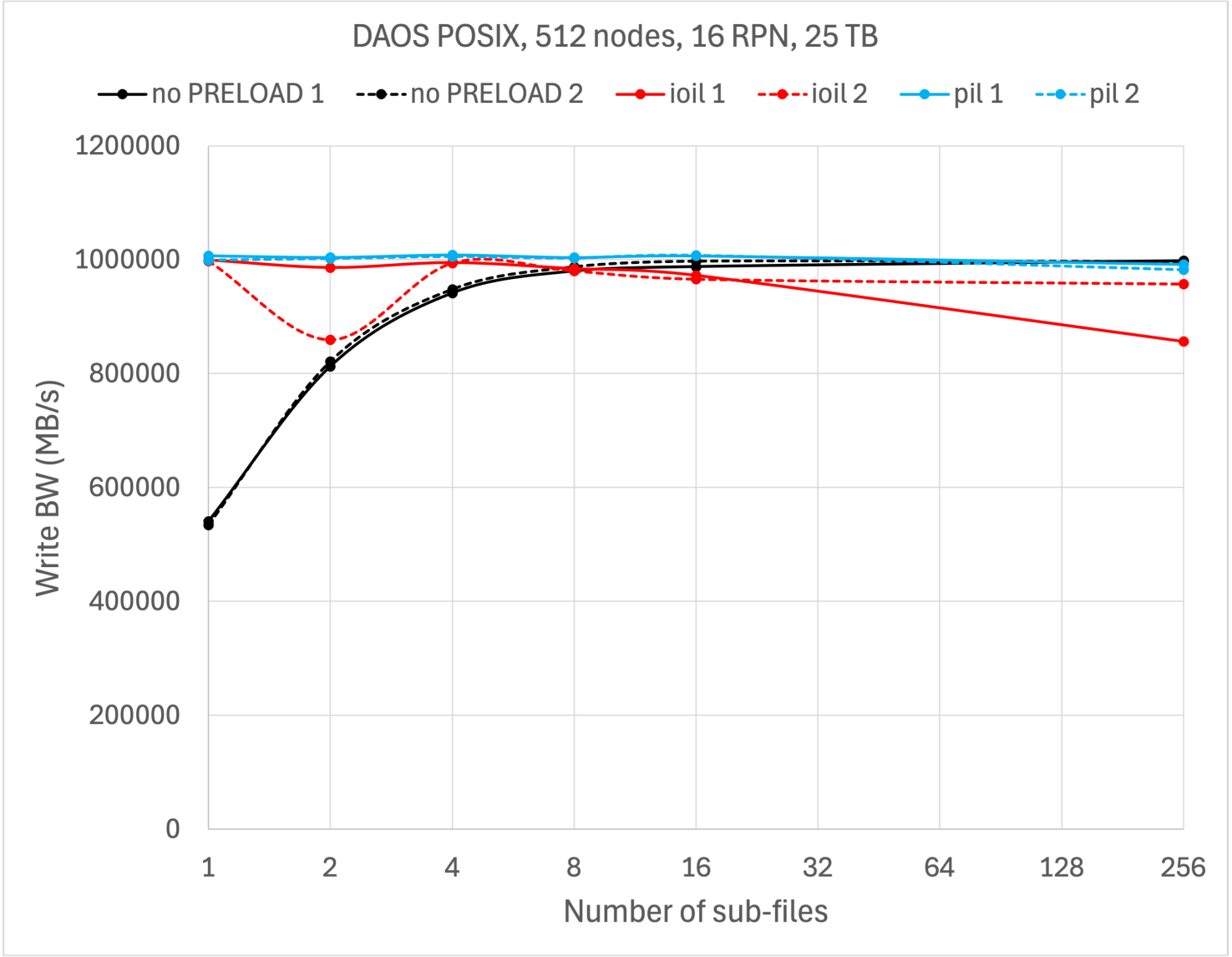
R. Latham *et al*.,

"Initial Experiences with DAOS Object Storage on Aurora,"

SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2024, pp. 1304-1310, doi: 10.1109/SCW63240.2024.00171.

DAOS:: HACC

- Running HACC Generic IO
  - HACC is Cosmology code which simulates galaxy formation
  - Generic IO is a benchmark using their application I/O code
- Running at 512 compute nodes
  - Writing 25TB checkpoint
  - Using POSIX API
- Using daos_user (20 servers)
- Code has been write optimized
- 1 TB/s with SX



Ongoing

- Lustre DAOS integration
- *kfi* lnet driver
- Move gecko to production HDR network from Aurora local HDR fabric

References

ATESPEC. Argonne Training Program On Extreme Scale Computing

[1] Muralidharan, Servesh. “Aurora Exascale Architecture”, ATESPEC 2023,

- Aurora Exascale Architecture – Intel Data Center GPU Max Series, DAOS and HPE Slingshot

https://extremecomputingtraining.anl.gov/wp-content/uploads/sites/96/2023/08/ATPESC-2023-Track-1-Talk-3-Servesh-Mulalidharan-Aurora.pdf . Accessed 3/24/25

[2] Muralidharan, Servesh, “Overview on Aurora Exascale Compute Blade”,

https://www.alcf.anl.gov/sites/default/files/2024-07/Aurora-DataFlow-28Feb24.pdf . Accessed 3/24/25

[3] R. Latham *et al*., "Initial Experiences with DAOS Object Storage on Aurora," *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Atlanta, GA, USA, 2024, pp. 1304-1310, doi: 10.1109/SCW63240.2024.00171.

[4] DAOS Documentation, “Tiering and Unified Namespace”

https://docs.daos.io/v2.6/admin/tiering_uns/?h=lustre+daos+tiering

[5] Distributed Asynchronious Object Storage https://docs.daos.io/v2.6/

Questions?

Extra Slides

Extra slides

- File: ALCF_Site_Update-v02