Argonne
NATIONAL LABORATORY

Lustre at Scale: ALCF Update

Alex Kulyavtsev
Gordon McPheeters

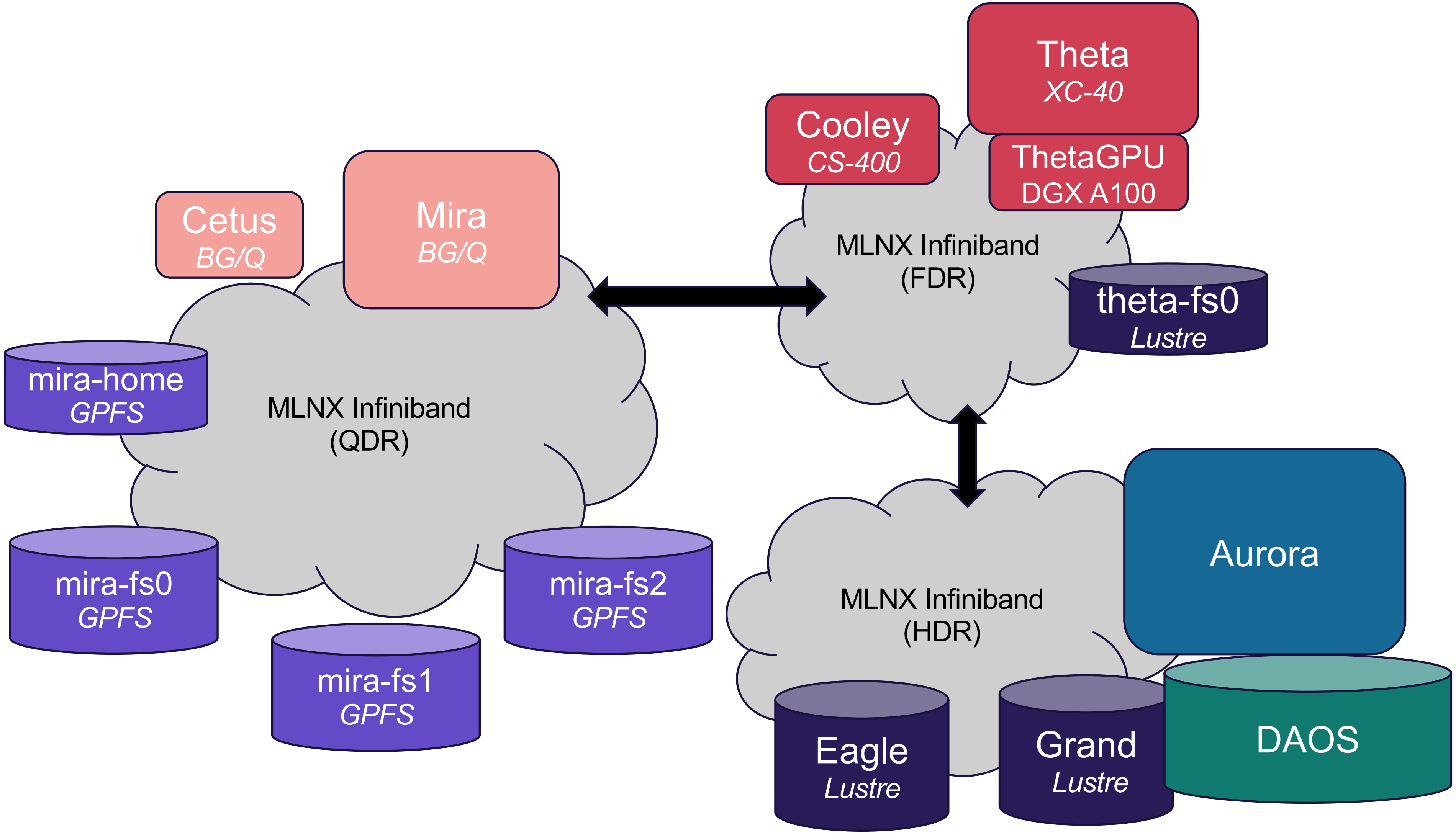
5/19/2021

Topics To Be Covered

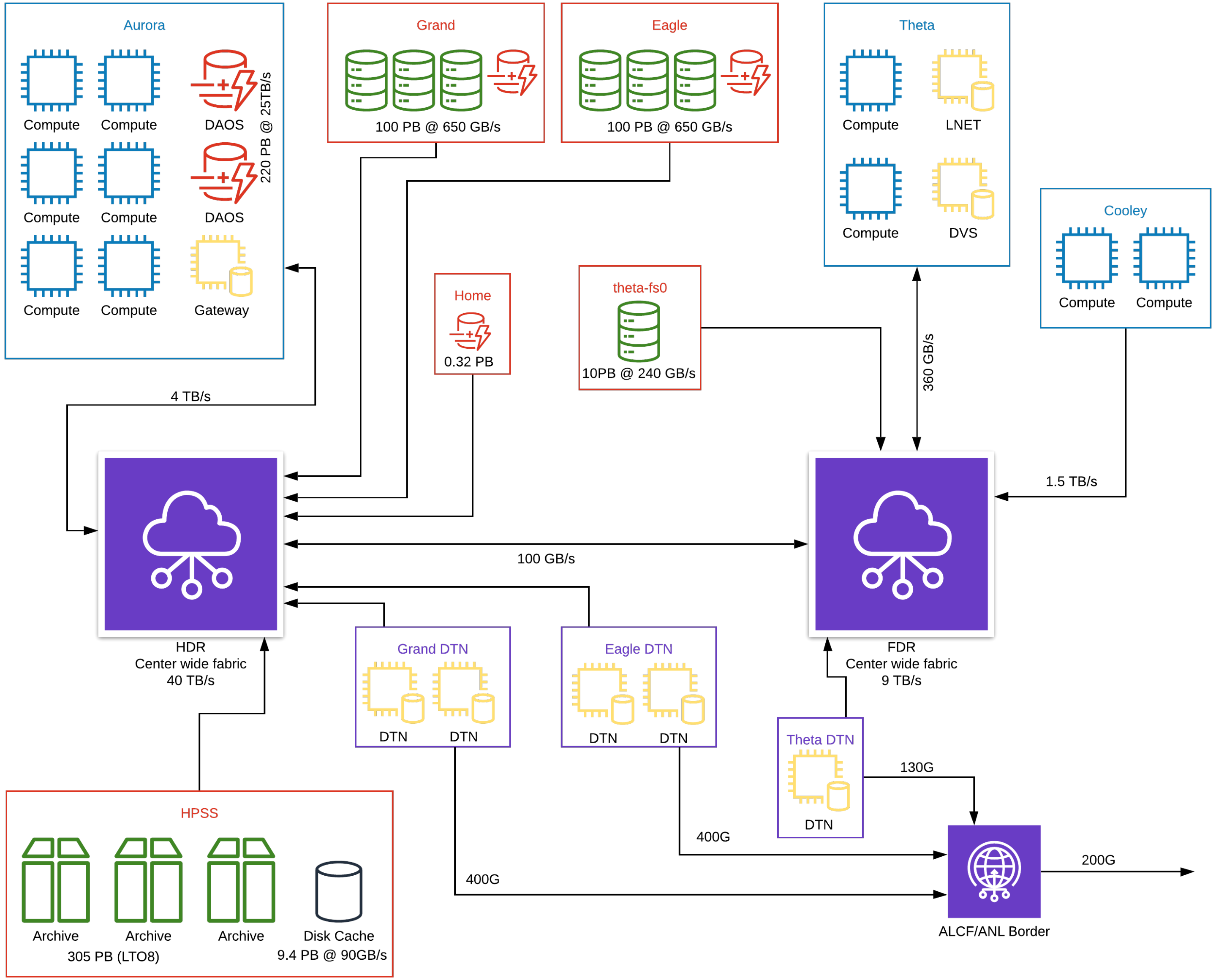
- review existing compute clusters
- review legacy file systems
- new HPE E1000 based file systems brought into production in December 2020
- installation and acceptance experiences

Update of ALCF's existing Compute Clusters

- since last year we have retired our IBM BG/Q machine in production since 2012 (GPFS based)
- continue to support Theta
  - Cray XC-40 4392 KNL nodes - 11.7 PF
- continue to support Cooley visualization cluster - Cray CS400
- installed 24 nodes of NVIDIA DGX A100 in Summer of 2020 (*ThetaGPU)*
- will be adding a new system based on AMD/GPU named Polaris - summer 2021 timeframe


Cetus
BG/Q
Mira
BG/Q
mira-home
GPFS
MLNX Infiniband
(QDR)
mira-fs0
GPFS
mira-fs1
GPFS
mira-fs2
GPFS
Cooley
CS-400
Theta
XC-40
ThetaGPU
DGX A100
MLNX Infiniband
(FDR)
theta-fs0
Lustre
MLNX Infiniband
(HDR)
Aurora
Eagle
Lustre
Grand
Lustre
DAOS

ALCF Storage



Legacy ACLF file systems

/lus/theta-fs0

- HPE Sonexion CS-3000. Running Neo 3.4 / Lustre 2.12
- 10PB based on 56 OSS/OST (41 x 6TB HDD dRAID)
- 6 billion inodes in 4 MDS/MDT based on 10K HDD

- small HPE L-300 for Test and Development (TDS) system

- retiring old GPFS filesystems on DDN SFA12KE and migrating data to Lustre
  - migrate 26 PB in two GPFS Mira project file systems to new Lustre (Grand)
  - evaluating the replacement of our current /home with Lustre DDN AI-400X (ExaScaler Appliance)
    - 4 MDT / 8 OST all NVMe flash

Global and Community File Systems: Grand and Eagle

- new capacity for ALCF
- accessible to all compute platforms in ALCF including Aurora
- each filesystem offers both *project/campaign* and *community* usages

- Grand - Global File System (GFS)
  - — combination of *campaign* and long term *project* storage

- Eagle - Community File System (CFS)
  - — designed to enable site wide and world wide sharing of ALCF project data within collaborations
  - — project space will be allocated on Eagle if project plans to share data

- upcoming: DAOS
  - – designed as the high performant primary campaign storage for Aurora
  - – novel object store based on Persistent Memory and NVMe
  - – name space integrated with Grand

Grand and Eagle file systems

two identical HPE E1000 Lustre appliances

100 PB of available storage in 10 racks each

- 40 OSS/160 OST
  - — 80 enclosures with 106 HDD drives each
  - — each OST is 53 drive dRAID 5*(8+2)+3 with SSD journal
  - — dual path Seagate 16TB HDD SAS-3
  - — 20 Viking enclosures for controllers and journal SSD
- 20 MDS/40 MDT
  - — 10 Viking enclosures
  - — 22 x 3.84 TB NVMe drives + 2 hot spares
  - — two RAID-10 arrays
  - — 160 billion inodes = 40 x 4 billion/MDT
- OST and MDT are ldiskfs based
- schedule precluded ZFS deployment, continue to watch ZFS dRAID developments
- HDR (200 Gb/s) IB Network - two interfaces per MDS, one interface per OSS
- meet the 650 GB/sec both read and write RFP requirement

- small 1 PB Test and Development system (TDS) named Gila

Community File System Sharing Technology

- there are dedicated Globus Data Transfer Nodes (DTN) which mount all lustre file systems.
- ALCF project Principal Investigators can freely share data with other Globus users without requiring them to have an ALCF account
- collaborators will authenticate to the Globus service using their institution's existing identity provider, or by creating their own Globus ID
- collaborators can transfer data from shares to other institutional Globus endpoints or to their own Globus Connect Personal endpoints.
- data is shared in-place from Eagle. There is no need to copy data to an intermediate location.

File System Client Zoo

- networks :
  - — Theta compute nodes on the Aries network
  - — all other clients are IB attached
- LNet routers to connect Aries to IB
  - — theta-fs0 : 30 routers with Fine Grained Routing
- various Mellanox OFED levels ranging from 4.9 to 5.1+
- range of OS including SLES, RHEL, Ubuntu
- Lustre clients:
  - — recently we ran a variety of HPE, upstream / community and DDN versions
  - — now standardized on client offered by HPE
  - — driven by first availability of patch for LU-12506 in 2.12: Single client mounting > 64 MDTs

Eagle and Grand Installation and Acceptance Experiences

- among the first customer ships of E1000 from HPE
- machines received and built during challenging (COVID-19) times
- global supply chain disruptions added to the challenge
- this was our first global file systems that was not explicitly tied to a specific compute platform
- no dedicated compute platform *yet* for benchmarking with sufficient bandwidth *between* networks
- to see full bandwidth we used Eagle's file system server nodes as clients to drive load on Grand, and then reversed
- still 1:1 ratio of clients / servers. To have 4:1 ratio we created pools representing of 1/4 file system and used all hosts from Eagle to drive 1/4 of Grand servers
- final tests with Theta using LNet routers was performed to ensure no problems were observed with a client count of 4392 and all filesystems mounted
- added an additional seperate set of 30 LNet nodes to route to Grand (18) and Eagle (12) to avoid interfering with production workloads
- separate LNet on same IB fabric per each Lustre FS
  - o2ib, o2ib22, o2ib23, … , more for FGR

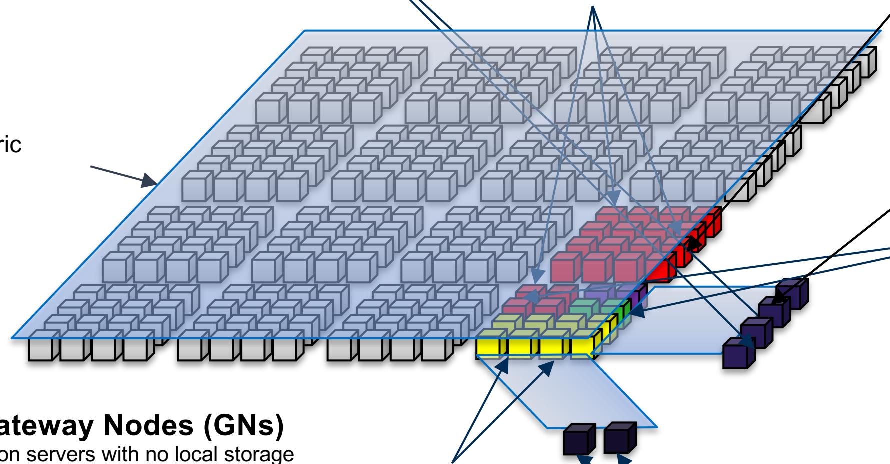
Acceptance: Lessons Learned

- expect the unexpected
- max number of files per directory is < 10 million in ldiskfs based Lustre formatted without ext4 `large_dir`, can vary
- maximum file size limited by ldiskfs max filesystem size:
  - — 2.5 PB = 160 OST * 16TB/OST
  - — looking forward to over-stripping support
  - — also observed for sparse file
- hit a limit with client mounting multiple file systems at the same time having a total of 80 MDTs (LU-12506)
- worked through some HA issues
- client/server Lustre version compatibility drives upgrade process
  - — the addition of Grand and Eagle at Lustre 2.12 as *global file systems* required client upgrades from older levels (2.7) to 2.12, which in turn drove upgrades to existing servers (also at 2.7) as well to support new clients.
  - — otherwise we could have left Theta clients / Sonexion server versions alone

Future Activity

- multiple new compute platforms over the next years that need to mount existing Lustre fs
- HPE's Lustre service stream is keeping us busy with upgrades which is good
- we will work on DAOS integration with Lustre
- investigate and deploy data management solution
- investigate and deploy monitoring tools to have top level view of entire Lustre infrastructure

ALCF DAOS Overview



DAOS Nodes (DNs)
Xeon servers
NVRAM and NVMe attached storage
DAOS service

DAOS Performance
220 PB capacity @ EC16+2
≥ 25 TB/s

Lustre Performance
Grand – 100 PB @ 650 GB/s
Eagle – 100PB @ 650 GB/s

Slingshot Fabric

System Service Nodes (SSNs)

User Access Nodes (UANs)

Scalable Storage Cluster (SSC)
Xeon servers connected to JBOD
Lustre OSSs & MDSs

Gateway Nodes (GNs)
Xeon servers with no local storage
Access to external storage

Existing storage systems

Ongoing Challenges

- There is little that can be prepared for in terms of new client code changes to address newly released security patches as the security patches are released as found and not on a particular cadence.
- However, hardware vendors pushing latest kernels and network driver levels to drive maximum performance leaves us with a difficult task building working LTS client codes when the patch may or may not yet exist in master and/or not backported to LTS. Proper testing at scale complicates this process even further.
- Example: IO slowdowns very recently related to grant issues LU-13972. Have mitigation in place, will need client updates.

- Metadata explosion. Machine Learning workloads (and not only) often create many files.

  e.g. project created 80M files on short time span - the last addition to 2 billion files

  we currently have on theta-fs0
- We are learning to use metadata restriping and facing issues with scalability of the process.

- Want to migrate 2 billions files / 10 PB of data from theta-fs0 to Grand.

- Performing software upgrades during continuous production.

Questions?