Lustre in Commercial Space Panel

Lustre User Group Meeting
April 17, 2013

Panelists

- Kent Blancett, *BP*
- Giuseppe Bruno, *Bank of Italy*
- Alan Wild, *ExxonMobil*
- Nathan Carman, *ITT Exelis*
- Josh Judd, *WARP Mechanics*

Lustre in Commercial Space

- What does your company/domain do with HPC that require a parallel file system?
- Why did your company/domain choose Lustre, instead of competitors?
- Can you provide configuration/architecture details for your Lustre file system?
- How do you use Lustre in your workload cycle?
- What do you want more from Lustre in future?

Lustre in Commercial Space Panel

Kent Blancett
BP

What/Why?

- What we do:
  - Provide computational resources for Seismic Imaging and Algorithm Research
  - All filesystems are for large project work – we have no scratch space – data loss is NOT cool!
- Why we chose Lustre over others – we didn't
  - We have Lustre, Panasas, and GPFS filesystems for different workloads and to foster competition
    - Now: Lustre = 7.5PB, Panasas = 4.5PB, GPFS = 6PB
    - Fall: Lustre = 10.5PB, Panasas = 5PB, GPFS = 3.5PB
  - Scalability and reliability are key

Compute Resources

- 2912 Sandybridge 2socket (16 core) – 128GB RAM, 10GbE
- 1920 Westmere 2socket (12 core) 48/96GB RAM, 10GbE
- 48 Westmere-EX 4socket (40 core) 512GB RAM, 10GbE
- 19 Dunnington 4socket (8 core) 256GB RAM, 10Gbe

- Expecting Ivybridge cluster in new building

Lustre Filesystems

- Now - 3 x 2.5PB production
  - Each has 16 OSS, 2 DDN SFA12K-40 couplets,1200 x 3TB SATA drives, each OSS has dual 10GbE
  - Failover, corosync/pacemaker 4 node groups
  - Lustre 2.1.2 servers, 1.8.8-wc1 clients due to client cache issue
- Fall – 3 x 2.5PB, 1 x 3PB production
  - Change from above all OSS move to dual 40GbE
  - 3PB filesystem will have 1600 total drives
- Bandwith: Now - 32GB each, Fall - ????

Implementation and Future

- Integration with DDN help, tweaks to corosync/ pacemaker for scale, tweaks to kernel tuning for failover/reconnect timeouts for scale
- L1 = BP, L2 = DDN, L3 = Intel
- Still looking for good Lustre log instead of debug log

Lustre in Commercial Space Panel

Giuseppe Bruno
Bank of Italy

BANCA D'ITALIA
EUROSISTEMA

Lustre at the Bank of Italy

- What does your company/domain do with HPC that require a parallel file system?
  - ➢ Seismic Processing, Reservoir Simulation
- Why did your company/domain chose Lustre, instead of competitors (e.g. PVFS, GPFS)?
  - ➢ We wanted a supported solution that worked well with our Cray.

BANCA D'ITALIA
EUROSISTEMA

Lustre at the Bank of Italy

- What does your company/domain do with HPC that require a parallel file system?
  - ➢ Seismic Processing, Reservoir Simulation
- Why did your company/domain chose Lustre, instead of competitors (e.g. PVFS, GPFS)?
  - ➢ We wanted a supported solution that worked well with our Cray.

BANCA D'ITALIA
EUROSISTEMA

Lustre at Bank of Italy

- Cluster Status for lustre-cluster1 @ Thu Apr 4 08:51:15 2013
- Member name — ID — Status
- Osiride-lp-030 — 1 — Online, Local, rgmanager
- Osiride-lp-031 — 2 — Online, rgmanager
- /dev/dm-45 — 0 — Online, Quorum Disk

Service name	Owner	State
ha_mdt	Osiride-lp-030	started
ha_mgs	Osiride-lp-030	started
ha_ost0	Osiride-lp-031	started
ha_ost1	Osiride-lp-031	started
ha_ost3	Osiride-lp-031	started
ha_ost4	Osiride-lp-031	started
ha_ost5	Osiride-lp-031	started

BANCA D'ITALIA
EUROSISTEMA

Lustre at the Bank of Italy

- Cluster Status for lustre-cluster1 @ Thu Apr 4 08:51:15 2013
- Member name — ID — Status
- Osiride-lp-030 — 1 — Online, Local, rgmanager
- Osiride-lp-031 — 2 — Online, rgmanager
- /dev/dm-45 — 0 — Online, Quorum Disk

Service name	Owner	State
ha_mdt	Osiride-lp-030	started
ha_mgs	Osiride-lp-030	started
ha_ost0	Osiride-lp-031	started
ha_ost1	Osiride-lp-031	started
ha_ost3	Osiride-lp-031	started
ha_ost4	Osiride-lp-031	started
ha_ost5	Osiride-lp-031	started

BANCA D'ITALIA
EUROSISTEMA

Lustre at the Bank of Italy

- Cluster Status for lustre-cluster1 @ Thu Apr 4 08:51:15 2013
- Member name — ID — Status
- Osiride-lp-020 — 1 — Online, Local, rgmanager
- Osiride-lp-021 — 2 — Online, rgmanager
- /dev/dm-50 — 0 — Online, Quorum Disk

Service name	Owner	State
ha_ost06	Osiride-lp-020	started
ha_ost07	Osiride-lp-020	started
ha_ost08	Osiride-lp-020	started
ha_ost09	Osiride-lp-020	started
ha_ost10	Osiride-lp-020	started
ha_ost11	Osiride-lp-020	started
ha_ost12	Osiride-lp-021	started
ha_ost13	Osiride-lp-021	started
ha_ost14	Osiride-lp-021	started
ha_ost15	Osiride-lp-021	started
ha_ost16	Osiride-lp-021	started
ha_ost17	Osiride-lp-021	started

Lustre in Commercial Space Panel

Alan Wild
ExxonMobil

ExxonMobil
Taking on the world's toughest energy challenges.™

Lustre in Commercial Space

- What does your company/domain do with HPC that require a parallel file system?
  - ➢ Seismic Processing, Reservoir Simulation
- Why did your company/domain chose Lustre, instead of competitors (e.g. PVFS, GPFS)?
  - ➢ We wanted a supported solution that worked well with our Cray.

ExxonMobil
Taking on the world's toughest energy challenges.™

Lustre in Commercial Space

- Can you provide brief configuration/architecture details for your Lustre file system?
  - Cray Sonexion 1600 storage. 4 live filesystems with about 8PB of capacity online today.
- How do you use Lustre in your workload cycle?
  - "scratch", but we define scratch differently then most
- What do you want more from Lustre in future?
  - Stability, distributed name space, HSM

Lustre in Commercial Space Panel

J. Nathan Carman
Lead Systems Engineer
Innovations - AIS

Agility and Ingenuity for the 21st Century

C4ISR Electronics & Systems (C4ISR)

Electronic Systems

Providing Customers With:

The ability to sense and deny threats to manned and unmanned aircraft, ships, submarines, ground vehicles and personnel and to provide warfighters with enhanced situational awareness.

Geospatial Systems

Providing Customers With:

Next-generation imaging that integrates space and airborne sensors into broader, coordinated systems.

Night Vision & Tactical Communications Systems

Aerostructures

Providing Customers With:

Aerospace Composite Tanks, Fiberglass Piping Systems, Marine Piping Systems, Multi-angle Piping Systems, Primary Airframe Structures and Space Structures,

Information & Technical Services (I&TS)

Information Systems

Providing Customers With:

Data fusion, network integration and critical decision support services.

Mission Systems

Providing Customers With:

A broad range of critical service, support and logistics solutions that enable efficient operations in the most demanding environments.

Why does Exelis need a HPC parallel file system?

- Developing, Deploying, and Operating Integrated Distributed Data Centers

Integrated Distributed Data Center

Key Architectural Attributes

- High Bandwidth Circuits (where possible)
- Efficiently Use Circuits
  - Extend Data Center Protocols Across the WAN

Initial and Life Cycle Cost Savings

- Allows for strategically choosing where equipment is installed
- Changes Data Duplication CONOP
- Improved Timely Analytics

	Ingest Site	Access Site	Processing Site	Storage & Server Site	System Capability
Network Layer	Yes	Yes	Yes	Yes	Yes
Server Farm			Yes	Yes	Yes
Storage Farm				Yes	Yes
Services, Ingest	Yes				Yes
Services, Access		Yes			Yes
Services, Transport				Yes	Yes
Services, Processing			Yes		Yes

Sample System Total Capability

Why Lustre?

- Performance (Block I/O, local and across WAN)
- Open Source
  - License costs
  - Compliance with government objectives
    - "Adapt the current technology acquisitions process to default to Open Technology Development implementations."
- Vendor Independence

Exelis need a HPC parallel file system?

- Installed on High Performance (private) WANs, distributed resource across the WAN
- DIY Custom built systems
- Provide our own Level 1/2/3(some) support
- Compliant with required Information Assurance and Cyber Security processes and procedures

How is Lustre Used?

- Fundamental Component Of Distributed High Performance System
- Used To Provide Low Latency / High Bandwidth User Access To Large Format Data

What is Needed in the Future?

- Installation/Management/Monitoring Layer
  - To include an API
- Failover/High Availability Features
  - More robust and easier to manage
- Windows clients
  - As Lustre Grows Beyond HPC World Into Enterprise World The Issue Will Grow

Thank you!