lustre

Phil Schwan
phil@clusterfs.com
http://www.clusterfs.com

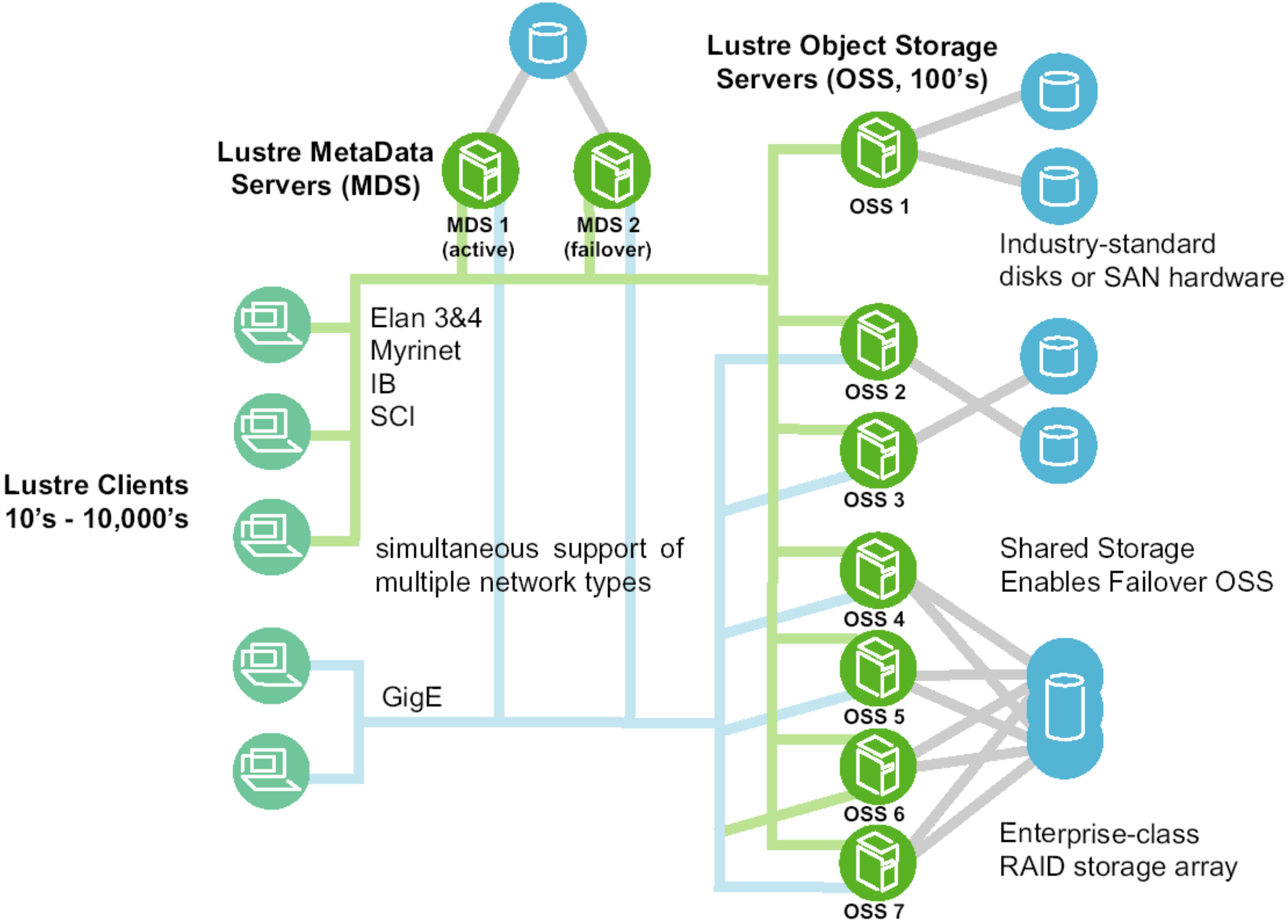
Agenda

- Whirlwind introduction
- Lustre success stories
- Thoughts for new clusters
- Cluster File Systems, Inc.

0-60 in 5 Minutes

- Very cross-platform (on Linux)
- Maximum scalability
- Maximum availability

Architecture



Lustre MetaData Servers (MDS)

Lustre Object Storage Servers (OSS, 100's)

Lustre Clients 10's - 10,000's

Industry-standard disks or SAN hardware

Shared Storage Enables Failover OSS

Enterprise-class RAID storage array

simultaneous support of multiple network types

Cross-Platform

- Stable: Linux on i386, ia64, x86-64
- Testing: Linux on PPC
- Development: OS/X client
- liblustre available for others

Maximum Scalability

- Today's production code supports
  - 1,000s of clients
  - 100s of servers
  - Linear I/O scalability
  - High metadata concurrency
  - Heterogeneous networking

Common HPC Workloads

- High metadata concurrency. Consider:
  - 2,048 tasks start
  - each creates 10 files in same directory
  - 5,000 creates/sec
- I/O parallelism
  - n to n
  - n to 1
  - HDF5
  - MPI-I/O

Maximum Availability

- Zero single points of failure
- Automatic recovery (ie, server reboot)
- Automatic failover (ie, node on fire)
- Application-transparent
- Fast – no voting, quorum, group membership, etc.
- Starts quickly
  - "Thunder" mounts in 7 seconds

Success Stories

PNNL: HPCS2

- 1,024 dual-cpu Itanium clients
- Elan3/Elan4
- 600 MB/s from a single client
- 3.2 GB/s aggregate (disk bottleneck)
- 53 TB
- In production for more than a year
- *nwchem* über alles

LLNL: Thunder

- 1,024 quad-cpu Itanium clients
- ia32 servers
- Elan4 clients, GigE servers
- 150 TB
- pre-production testing

LLNL: MCR/PVC, ALC, Lilac

- MCR/PVC:
  - 1150 dual-cpu Xeon compute clients
  - 64 visualization clients
  - heterogeneous Elan3/GigE
  - 100 TB / 20M files
  - In production since September
- ALC: (same configuration)
  - Different user community
  - Half production / half testing
- Lilac: (same configuration)
  - Classified

NCSA: Tungsten

- 1,280 dual-cpu Xeon clients
- 104 server nodes
- All GigE
- 11.1 GB/s aggregate (disk bottleneck)
- 150 TB
- User jobs started in January

Cross-industry, cross-continent

- earth science: weather/seismic/chem
- oil/gas industry applications
- digital movie production
- pharmaceuticals
- classified weapons research
- homeland security
- US, Europe, Australia, China

New Clusters

- Completely commodity (almost)
  - Some shared storage required for failover
- GigE, Elan3/Elan4, Myrinet today
- InfiniBand coming
- Tiny files work fine, but larger files perform best
- Clients on servers not optimal

Support Option One

- $5,000 per year, per cluster, includes:
  - 10 hours of support
  - Access to the latest versions
  - License for the management tools
  - Documentation and training materials
  - Next business day email response
  - Additional hours available
  - Custom development, flat-rate deliverables available

Support Option Two

- For the mission-critical deployment:
  - 24/7, 4-hour response time available
  - Telephone or email support
  - On-site training and support available
  - Cost varies by deployment size, difficulty, and needs
  - Generally flat-rate

Fin

lustre

http://www.clusterfs.com
http://www.lustre.org