New Storage System Solutions

Craig Prescott
Research Computing

May 2, 2013

Outline

- Existing storage systems
- Requirements and Solutions
- Lustre
- /scratch/lfs
- Questions?

Existing Storage Systems

File System	Purpose
/home	Source code, Makefiles, etc.
/scratch/hpc	Job I/O
/project	Longer-term projects, low/ moderate job I/O
/lts	Long-term storage
/rlts	Replicated Long-term storage

Requirements and Solutions

- Nodes must share data over the network
  - 1000+ compute nodes and 20k+ processors
    - Large scalability, throughput, IOP, and capacity requirements
- Resilience in the face of component failure
  - RAID
  - High-Availability
- POSIX Semantics
  - All nodes must maintain consistent views of a file's data and metadata
    - Locking subsystem

Requirements and Solutions

- Support for High-Performance RDMA networking
- Only two choices
  - Traditional Network File System (e.g. NFS) or Parallel File System
    - Parallel file systems offer scalability
- Solutions
  - Good proprietary parallel solutions are available
    - Expensive, vendor lock-in, reduced flexibility
  - Open Source – numerous possibilities
    - UF is a member of OpenSFS
    - Lustre – proven track record, in-house expertise

Lustre – What is it?

- Open-source parallel file system implementation
  - GPL
  - POSIX compliant(*)
- Horizontally scalable in capacity and throughput
  - Additional storage and server nodes can be added "online"
- Scales to very large numbers of clients
- Files and Directories can be "striped" over multiple servers and storage devices

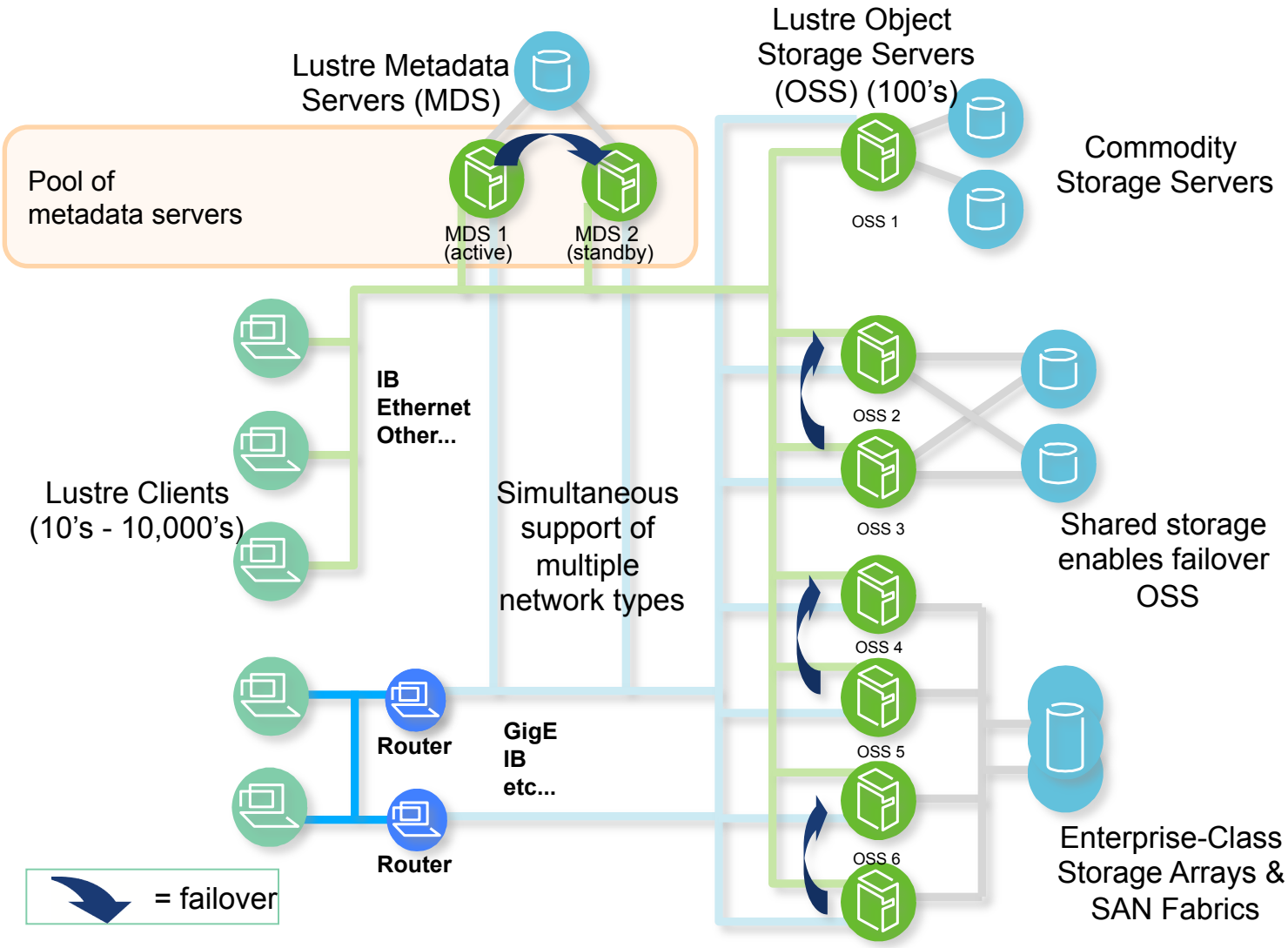
Lustre – What is it?

- Supports High-Performance RDMA networking devices
  - RDMA – Remote Direct Memory Access
    - NIC transfers data directly to/from application memory
    - Implemented in InfiniBand interconnect, widely used in HPC environments
- Built-in High-Availability Support
  - If a server goes down, another server takes over

Lustre – Who uses it?

- Largest market share in HPC systems
  - Primary file system of the "petascale club"
    - 50-70% of top 10, top 100, and top 500 entries at top500.org in recent years
    - Some very large deployments – many 1000s or tens of 1000s of compute nodes
      - Titan (ORNL), Blue Waters (NCSA), Sequoia (LLNL), Stampede (TACC), are examples of big places
- Top-tier HPC vendors use Lustre
  - Dell, DDN, Cray, Xyratex, Fujistu, HP, Bull, ...
- Minimal requirements - commodity compute and I/O hardware with Linux (any block storage device)

Lustre – Basic Architecture



Lustre Components - Object Storage Servers

- Object Storage Server (OSS)
  - Server node that hosts storage device(s) and NIC(s)
  - Manages object data and attributes
  - Manages RDMA/zero-copy from clients/ routers
  - Can be configured in active-active mode for failover

Lustre Components - Object Storage Targets

- Object Storage Target (OST)
  - Formatted storage device hosted by OSS
    - Ldiskfs ("ext4+") file system, zfs option soon – on-disk format is hidden from clients
    - Each OSS can have multiple OSTs – RAID6 arrays are common
  - OSTs are independent, thus enabling linear scaling
  - Optimized for large I/O (1MiB client RPCs)
    - Client writeback cache allows I/O aggregation

Lustre Components - Metadata Server & Target

- Metadata Server (MDS)
  - Server node with storage and NIC
  - Manages namespace only – filenames, attributes, and ACLs, no data
  - Determines file layout on OSTs at creation time
    - Default policies balance I/O and space usage over OSSs/OSTs
- Metadata Target (MDT)
  - Storage device hosted by MDS
  - Ldiskfs backing file system
  - Generally a RAID1+0 device

/scratch/lfs

- 24 OSS nodes
  - Active-Active
  - 128 GB RAM per OSS (read cache)
  - Six RAID6 ("8+2") OSTs per OSS
  - 1440 2TB drives total
  - 2.1 Petabytes usable
- ~50 GB/s streaming reads/writes observed
- ~100k random IOPS for small I/Os
- Highly-Available
- FDR InfiniBand Interconnect (56Gb/s)

Historical Perspective

- /scratch file systems through the years...

Name	Era	# Servers	# Disks	Interconnect	HA?
/scratch/ufhpc	2005-2011	8	144	SDR IB (8Gbps)	N
/scratch/crn	2007-2011	2	144	SDR IB (8Gbps)	N
/scratch/hpc	2011-	8	384(192)	QDR IB (32Gbps)	Y
/scratch/lfs	2013-	24	1440	FDR IB (54Gbps)	Y

Questions?

- Thank you!