

Open SFS Roadmap

Presented by David Dillow
TWG Co-Chair

TWG Mission

- Work with the Lustre community to ensure that Lustre continues to support the stability, performance, and management requirements of the OpenSFS members as HPC compute platforms continue to scale
- Responsible for creating and managing the roadmap for the OpenSFS community
 - Gather requirements from the Lustre HPC community,
 - Prioritize and recommend development projects to the Board,
 - Initiate RFPs for important features, and
 - Work with contractors to meet these requirements

Who is the TWG?

- ▶ The following have attended TWG meetings and/or contributed content to our requirements:

Bull/EOFS	LLNL
Cray	NRL
DDN	ORNL
Fujitsu	RAID, Inc.
Indiana University	Whamcloud
LBL	Xyratex

Importance of Community

- ▶ Community contribution is crucial
 - Fujitsu
 - EOFS
- ▶ Broadening the scope of requirements
- ▶ No monopoly on good ideas
- ▶ Avoiding duplicate effort

Process History

- ▶ Developed feature-based proposals
 - Presented to Board in January 2011
 - Rejected in favor of requirement-based approach
- ▶ Gathered and prioritized requirements
 - Reached out to broader community members for requirements
 - Prioritized by consensus

Prioritized Requirements

- ▶ Near-term requirements
 - Metadata server performance
 - Metadata server scaling
- ▶ Long-term requirements
 - Support alternate storage backends
 - Scalable fault management
 - Start investigations of alternate storage backends
- ▶ Improve the code base
 - Reduce maintenance effort
 - Reduce cost of new features

Process History

- ▶ Presented requirements and recommendations to the Board March 2011
 - <http://goo.gl/cZSWG+>
 - Board accepted our recommendations
- ▶ Developed RFPs for top two priorities
 - RFPs open to the public April 7, 2011
 - Metadata Performance and Scalability
 - Space Quota Accounting and Enforcement
 - http://www.opensfs.org/?page_id=149

Roadmap Caveats

- ▶ OpenSFS doesn't have direct control
 - Development performed by contractors or members
 - Clearinghouse for requirements
 - Host community architecture discussions
- ▶ RFPs open
 - Expect some traditional ideas to be proposed
 - Encourage novel ideas
 - Tough to predict exact roadmap!

Scaling Requirements

Metric	Lustre 2.0/2.1	Q2 2012	Q1 2014
maximum number of files in file system	4 billion	100 billion	1 trillion
maximum number of files in directory	10 million	50 million	10 billion
maximum number of subdirectories	10 million	1 million	10 million
maximum number of clients	131072	64 thousand	128 thousand
maximum number of OSS nodes	-	1 thousand	4 thousand
maximum number of OSTs	8150	2 thousand	8 thousand
maximum OST size	16 TB	32 TB	128 TB
maximum file system size	64 PB	100 PB	256 PB
maximum file size	320 TB	1 PB	-
maximum object size	2 TB	16 TB	64 TB
peak aggregate file creates/s	-	200 thousand	400 thousand
peak directory listings/s (ls -l)	-	-	100 thousand
maximum single client open files	~3 thousand	100 thousand	-
peak single client file creates/s	-	30 thousand	-

Metadata Server Performance

- ▶ Vertical Scaling
 - LNET scaling
 - RPC/MDS operation scaling
 - Size-on-MDS
 - Other novel ideas proposed by respondents
- ▶ Horizontal Scaling
 - Phase 1 - distributed namespace
 - Phase 2 - striped directories
- ▶ Long-term
 - Rework service model
 - Network Request Scheduler

Alternate Storage Backends

- ▶ Ldiskfs is nearing the end of its useful life
 - Requires external assistance for redundancy
 - Increasing disk capacities require larger LUNs for efficiency
 - No checksum of data
 - No online filesystem consistency check

Alternate Storage Backends

- ▶ Refactor obdfilter to allow new backends
 - Object Storage Device interface work
 - Partially funded by LLNL
- ▶ Requires work on quota system
 - Currently intimate with details of Idiskfs quotas
 - Lustre quotas need to be independent of backend
 - RFP out for this work

Storage Backends

- ▶ Top contenders
 - Ldiskfs
 - ZFS
 - BTRFS
 - Another upstart?
- ▶ OSD work to support all of these

Other Requirements

- ▶ Reliability
 - Fault detection, recovery, reporting
- ▶ Layout improvements
 - Allow layouts to adapt as the file grows and/or ages
 - Dynamic storage balancing
 - Snapshots/replication
- ▶ Environmental
 - Patchless server/support for newer distros
 - Support mixed endianness and page sizes
 - Improved configuration
 - Ipv6 support

Summary

- ▶ Full link to requirements document:

http://www.opensfs.org/wp-content/uploads/2011/03/OpenSFSTWGRequirements_2011-03-22.pdf

- ▶ TWG Archives

<http://lists.opensfs.org/pipermail/twg-opensfs.org/>

- ▶ Join us at discuss@lists.opensfs.org

dillowda@ornl.gov

carrier@cray.com