FUJITSU

shaping tomorrow with you

Fujitsu's Requirements for OpenSFS Lustre Road Map

Fujitsu Limited

Feb.8 2011

Fujitsu's Lustre Products Development Policy

FUJITSU

- Fujitsu provides open standard Lustre based parallel file system to our HPC system customers.
- Fujitsu is currently engaged in the development of extension of Lustre for large scale HPC system customers. Fujitsu intends to contribute those extensions to community and/or OpenSFS if appropriate.
- Fujitsu provides Lustre based file system as one of component of systems to our customers.

Requirements for OpenSFS Lustre Extension(1/2)

FUJITSU

Features		Current Lustre	2012 Goals	
System Limits	Max file system size	64PB	100PB	
	Max file size	320TB	1PB	
	Max #files	4G	32G	
	Max #subdirs	65000	100k	
	Max OST size	8TB	100TB	
	Max stripe count	160	10k	
Node Scalability	Max #OSSs	1020	10k	
	Max #OSTs	8150	10k	
	Max #Clients	32k	100k	*
Block Size of *ldiskfs* (Backend File System)		4KB	512KB	
Patch-less Server		NA	Support	*

* in Sun's Lustre roadmap 2008.

Requirements for OpenSFS Lustre Extension(2/2)

- (Cont'd)

Features	Current Lustre	2012 Goals	
Big-endian support (MDS/OSS)	NA	Support	
Advisory Lock (fcntl/lockf support)	NA	Support	*
IPF Server & x86_64 Client	NA (restricted by page size)	Support	
Quota OST storage limit	<= 4TB	No limitation	*
Server Cache	NA	Support	*
InfiniBand bonding	NA	Support	*
Arbitrary OST assignment	NA	Support	
Cluster MDS	NA	Support	*
QOS	NA	Support	

＊ in Sun's Lustre roadmap 2008.

Requirements for OpenSFS Lustre Performance

FUJITSU

	2012 Goals
Capacity	~100PB
Performance	~1TB/s for file transfer & MPI-IO create 200k files/sec getattr 100k files/sec (ls –l) Client&server cache
Scalability	~100k Clients ~4K OSSs + 10k OSTs
Usability	ACL Quota Advisory file locking Adaptive OST addition/deletion
Reliability	Assured disk I/O OSS/MDS Failover Journaling + Disk check(fsck)

Requirements for OpenSFS -- Architectural Extensions for Exascale

FUJITSU

- Interface extensions/Architectural extensions :

	FY2018 Goals	FY2012 Goals
Maximum File System Size	8EB	100PB
Maximum File Size	8EB	1PB
Maximum Number of Files	8E	32G
Maximum Number of Subdirectories in a Single Directory	1M	100k
Maximum Number of Clients	1M	100k
Maximum Number of MDSs	1000	100
Maximum Number of OSSs	128k	10k
Maximum Number of OSTs	128k	10k
Maximum Size of OST	1PB	100TB
Maximum Block Size of *ldiskfs*	512KB	512KB
Maximum Stripe Count	128k	10k

Technical Issues to be Solved

FUJITSU

- Technical Issues
  - InfiniBand Bonding
  - QOS and QUOTA
  - MDS and OSS Scalability:
    SMP Scaling and server node scaling
- Other Issues

MDS and OSS Scalability

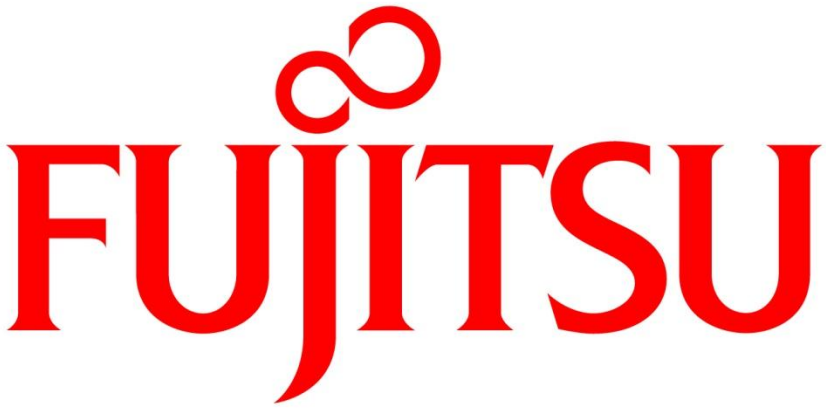
FUJITSU

- Locality is very important for increasing scalability
  - Locality : Storage, OSS, Interconnect, Server
- Utilize these locality as much as possible
  - MDS Scalability:
    - SMP Scaling up to over 64 core (8 core x 8 sockets)
    - Dividing file systems
    - Clustered MDS (Future)
  - OSS Scalability over 1000 of servers:
    - Minimizing OSS server response
    - Avoiding Interconnect congestion
    - Minimizing Storage(RAID) response

Other Issues

FUJITSU

- Kernel Independent Lustre Servers(MDS, OSS)
  - Current Implementation depends on kernel source codes, especially ext3 file system.
- LNET, OSS, MDS configuration for 100 thousands of Clients and servers
  - Automatic configuration is needed.
  - Checking connectivity among servers
- LNET Connections to 100 thousands of Client and OSS
  - Needs dynamic connection management
- Data Integrity
  - Fsck of whole data is impractical for over petabyte file system.
  - Backups and Snapshots
- Conversion Tools for Future Upcoming Lustre versions
  - FY2012 target file system should be upgraded to FY2018 target file system



shaping tomorrow with you