



ISBC

Introduce Intel Enterprise Edition for Lustre(IEEL) Product

ISBC Korea, kimwh@isbc.co.kr

Date : 2013/12/18

1. ISBC 소개

Focus open technologies and High value solutions
Contribute and Collaboration with Global Community
Leading Solution and Business Integration



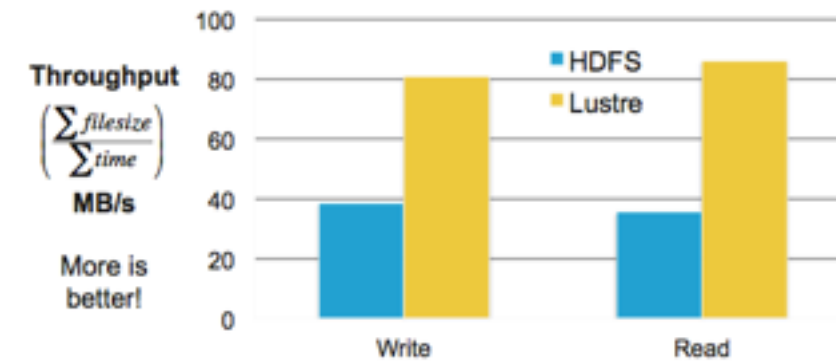
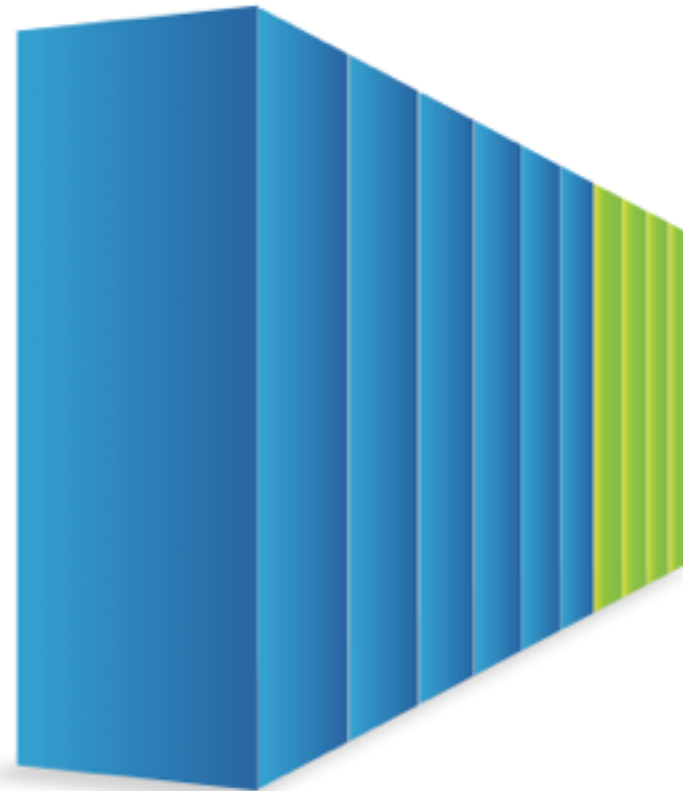
- Name(EN) ISBC (International Solution Business Consulting)
- Name(KR) 한국아이에스비씨주식회사
- CEO 김 완희
- Homepage <http://www.isbc.co.kr>



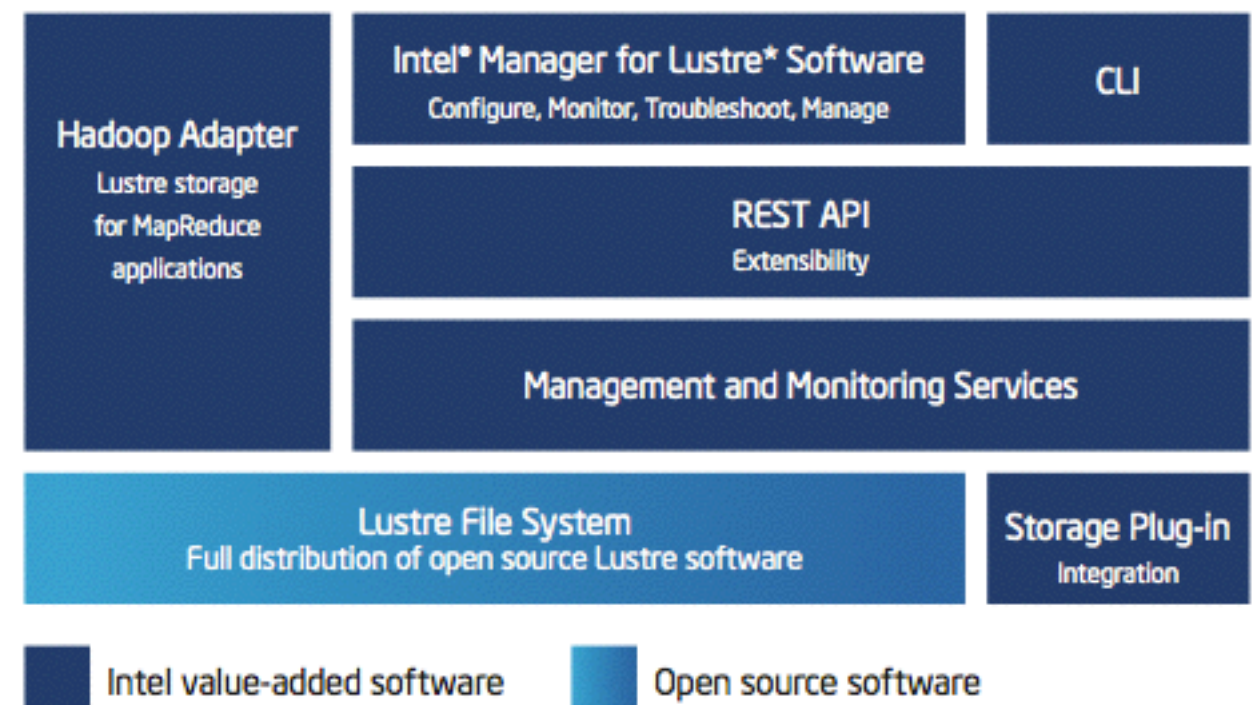
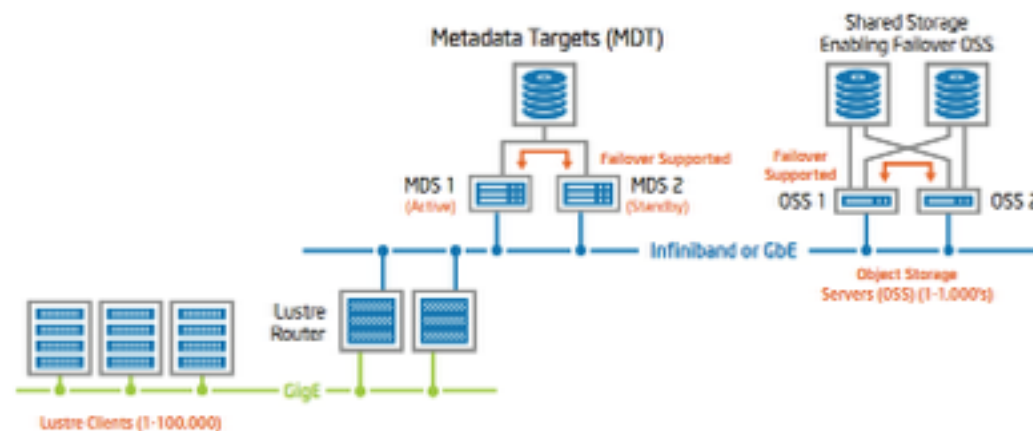
2. Intel Enterprise Edition for Lustre (IEEL)

IEEL은 웹기반으로 파일시스템 관리 및 운용 그리고 모니터링 기능을 포함 하고 있으며, IEEL의 핵심 러스터 파일 시스템은 전세계 Top100 슈퍼컴 시장에 60% 이상의 고객이 이미 사용 하고 있는 검증된 고성능 파일 시스템 입니다. 또한, HPC시장 뿐만아니라 최근 BigData환경에서도 IEEL을 사용할 수 있도록 Hadoop 제품들과 함께 연계 할수 있는 인터페이스도 제공 합니다.

Lustre powers
over
60%
of the top 100
supercomputers
worldwide¹



HDFS vs IEEL



2.1. IEEL 기능 소개



최대 100~100000 노드 까지 확장
 최대 2.5TB/Sec 의 성능
 평균 35000/s 메타데이터 처리
 최대 512PB 볼륨 확장
 최대 2.5PB의 단일 파일 크기
 1개의 디렉토리에 1천만개의 파일
 파일 시스템에 최대 40억개의 파일
 Hadoop Connector 제공
 웹 기반 관리 기능 제공
 고성능 파일 시스템 전용 라우팅

- 수많은 국가기관 연구소의 대규모 HPC 환경에서 Lustre 파일시스템은 수많은 핵심 기능을 바탕으로, 최적의 환경을 제공합니다. 또한 본 제품은 전세계 TOP 500의 고성능 컴퓨팅 환경에서 50% 이상의 고객이 사용하고 있으며, 고성능 IO가 필요로 하는 HPC시상에서는 필수적인 제품 으로 다음과 같은 환경을 제공 합니다:
 - ☒ 여러 노드에서 동시에 하나의 파일을 읽고 쓰기 작업을 할때에 Lustre는 병렬로 처리를 하며, 이를 기반으로 응용프로그램이 IO성능을 최대한 사용할수 있도록 만들어 줍니다.
 - ☒ Lustre는 시장의 표준 스토리지 여러대로 하나의 파일시스템을 구성할 수 있으며, 하나의 파일을 여러 스토리지에 분산/병렬로 저장 가능하며, 각각의 스토리지 성능을 하나로 합해서 운용가능 합니다. 이러한 방법으로 Lustre 와 함께 저렴한 스토리지로 대규모의 병렬 파일시스템을 제공 가능 합니다.
 - ☒ Lustre 파일 시스템의 I/O 는 1GbE, 10 GbE or Infiniband QDR 과 같은 다양한 종류의 네트워크와 함께 구성이 가능 합니다. 이것은 단일 클라이언드 기준에서 고성능 환경을 사용할 수 있도록 지원 가능하지만, 전통적인 HPC 응용프로그램 환경에서는 이러한 병렬 IO의 장점을 누릴수가 없었습니다.
 - ☒ Lustre 파일시스템은 여러개의 Disk를 하나의 파일시스템으로 구성하여 단일 마운트 포인트로 사용할수 있으며, 운용중에 Disk를 추가 하여 파일시스템 확장도 가능합니다. 이것은 대규모의 고성능을 요구하는 HPC환경에서 아주 유용한 기능 입니다.
 - ☒ Lustre 는 고가용성 및 복원력을 가지고 있으며, 이것은 대규모의 HPC환경을 위한 고가용성 및 장애 대응 구성을 지원 합니다.

3.1. Lustre File System 기능

Ext4 FS기반의 메타데이터를 위한 성능 확장 : Lustre uses an improved version of the ext4 journaling file system to store data and metadata. This version, called ldiskfs, has been enhanced to improve performance and provide additional functionality needed by Lustre.

POSIX 호환 환경 제공 : The full POSIX test suite passes in an identical manner to a local ext4 filesystem, with limited exceptions on Lustre clients. In a cluster, most operations are atomic so that clients never see stale data or metadata. Lustre supports mmap() file I/O.

고성능의 이기종 네트워크 통합 가능 : Lustre supports a variety of high performance, low latency networks and permits Remote Direct Memory Access (RDMA) for Infiniband (OFED) and other advanced networks for fast and efficient network transport. Multiple RDMA networks can be bridged using Lustre routing for maximum performance. Lustre also provides integrated network diagnostics.

고 가용성 환경 : Lustre offers active/active failover using shared storage partitions for OSS targets (OSTs) and active/passive failover using a shared storage partition for the MDS target (MDT). This allows application transparent recovery. Lustre can work with a variety of high availability (HA) managers to allow automated failover and has no single point of failure (NSPF). Multiple mount protection (MMP) provides integrated protection from errors in highly-available systems that would otherwise cause file system corruption.

향상된 보안성 : By default TCP connections are only allowed from privileged ports. Unix group membership is verified on the MDS. (Client 기준으로 Password Less SSH 환경이 필요 하지 않음)

엑세스 제어 및 확장 속성 : the Lustre security model follows that of a UNIX file system, enhanced with POSIX ACLs. Noteworthy additional features include root squash.

객체 기반 시스템 설계 : Clients are isolated from the on-disk file structure enabling upgrading of the storage architecture without affecting the client.

다양한 Stripe사이즈 지원 : The layout of files across OSTs can be configured on a per file, per directory, or per file system basis. This allows file I/O to be tuned to specific application requirements within a single filesystem. Lustre uses RAID-0 striping and balances space usage across OSTs.

NFS 및 CIFS 연동 가능 : Lustre files can be re-exported using NFS or CIFS (via Samba) enabling them to be shared with non-Linux clients.

파일 시스템 복구 툴 제공 : Lustre provides a distributed file system check (lfsck) that can restore consistency between storage components in case of a major file system error. Lustre can operate even in the presence of file system inconsistencies, so lfsck is not required before returning the file system to production.

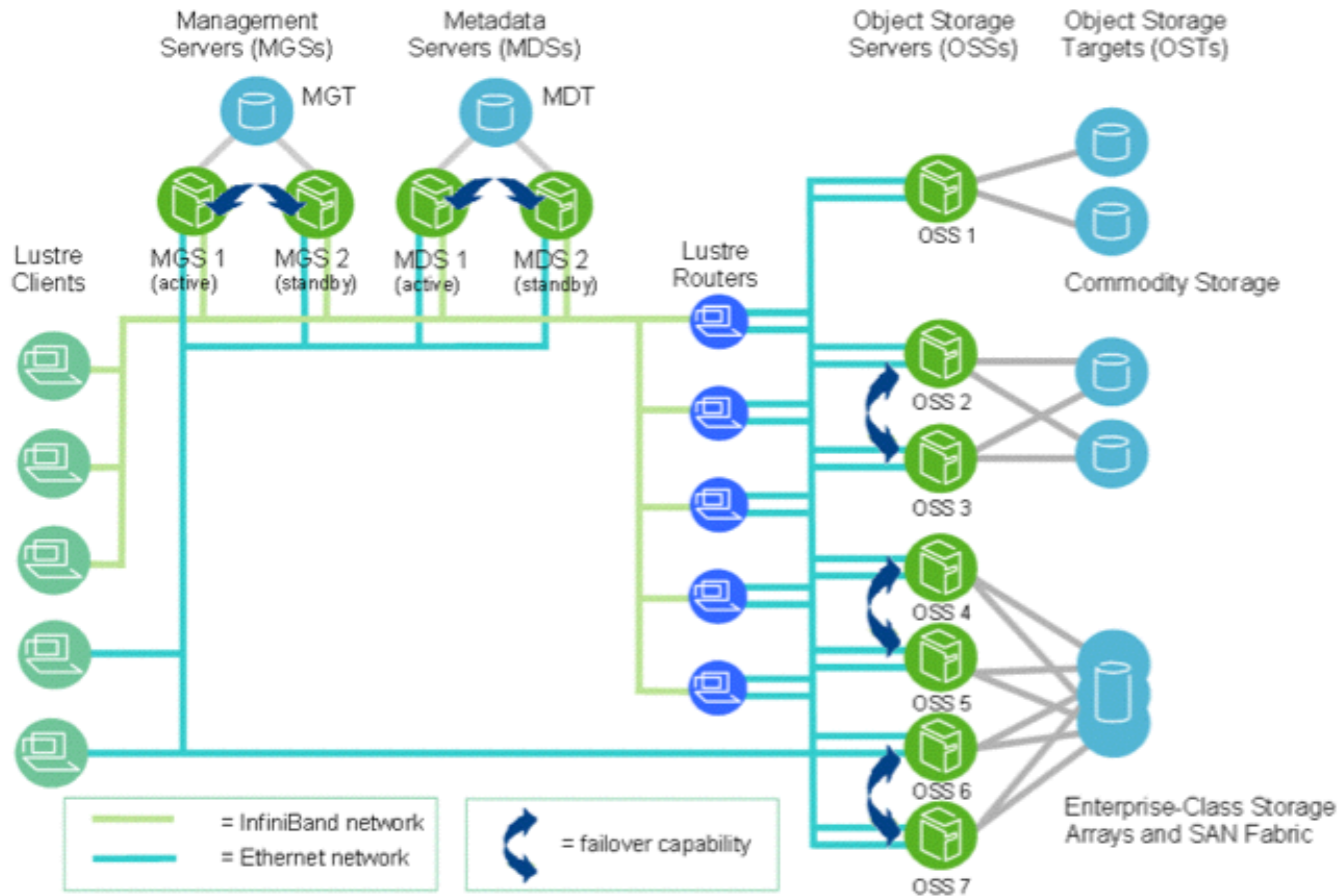
GUI기반의 성능 모니터링 : Lustre offers a variety of mechanisms to examine performance and tuning. (IEEL 은 IML을 포함하고 있습니다.)

3.2. Lustre 성능 및 운용 환경.

요구 사항	Lustre 제공 가능 환경
최대 파일 시스템 사이즈	하나의 파일시스템에 512PB까지 사용 가능
최대 단위 파일 크기	하나의 파일시스템에 최대 2.5PB 까지의 단위 파일 생성 가능하며, 작은 파일의 경우, OST단위별로 자동으로 로드밸런스 하여 운용됨
최고 성능	미국 LLNL에서 Sequoia시스템으로 1.5TB/Sec 성능 검증 DUG, Sangar 고객사에 300GB/Sec 환경으로 사용 검증
고성능의 메타데이터 환경	하나의 Directory에 천만개 화일 및 파일시스템에 40억개 이상의 파일 관리
수많은 클라이언트에서의 병렬 액세스 가능한 파일시스템	대규모 운용 환경에서 1000 - 20000 노드 까지 확장 운용 가능하며 중소 규모에서는 150 - 1000 노드의 범위로 운용 하면서 병렬 액세스 환경 제공.
고성능의 메타데이터 처리 속도	최대 25000/sec 번의 파일생성 및 100000/sec 번의 메타데이터 상태 변경 운용 성능

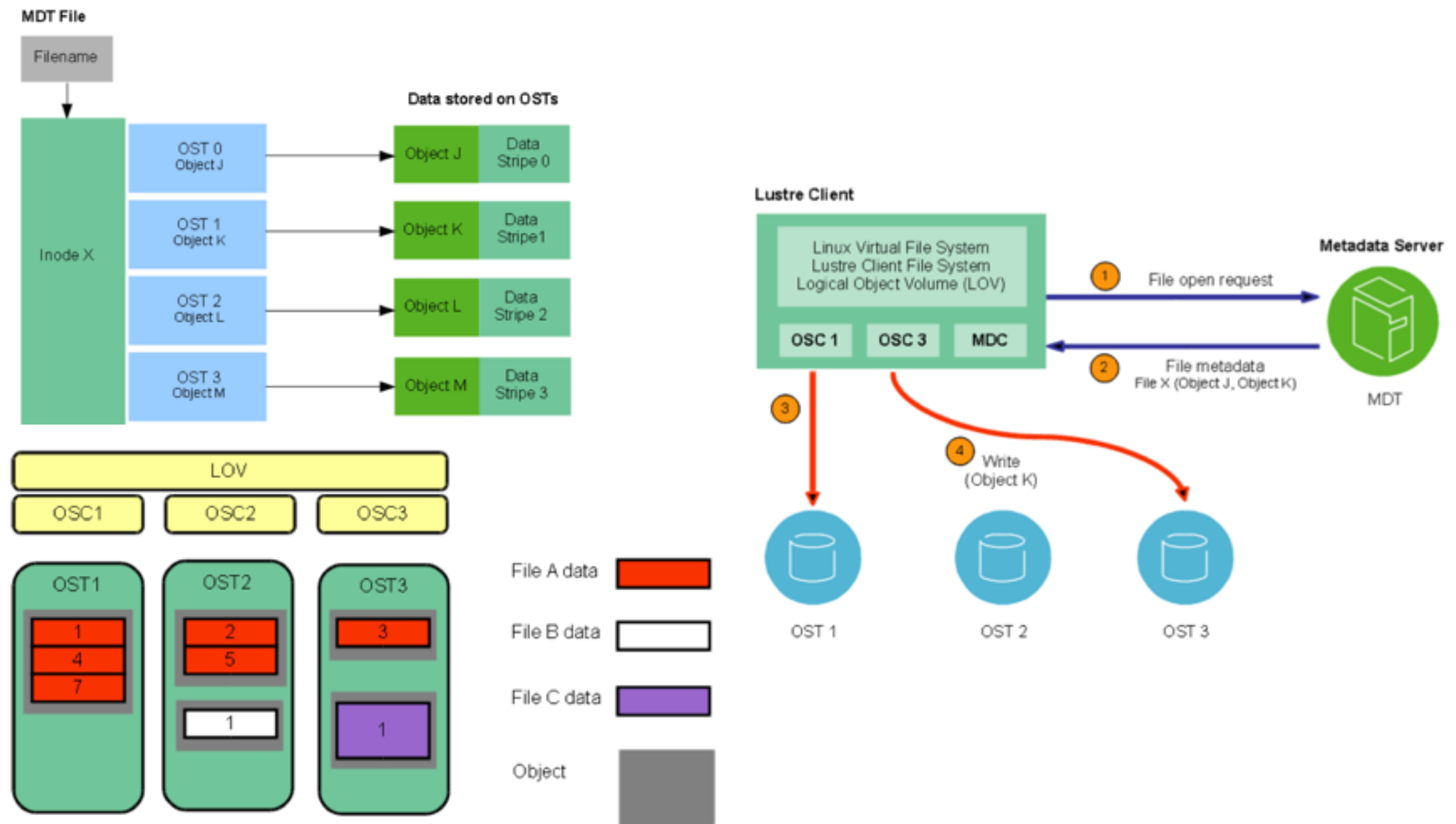
3.3. Lustre File System 기본 아키텍처

확장성 기준으로 기본적으로 수백 혹은 수천 노드를 포함할 수 있으며, 여러가지 네트워크를 혼용하여 운용 가능 하고, OSSs 사이에서고가용성 환경 구성을 기반으로 운용하도록 설계를 할수 있습니다.



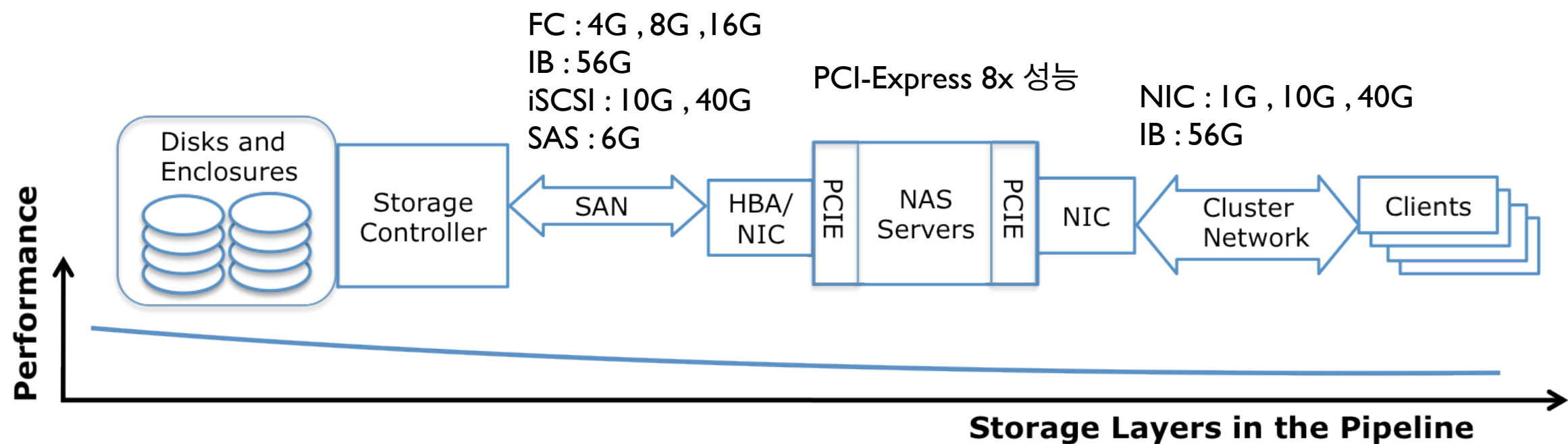
3.4. Lustre : 데이터 Striping

Client에서 데이터 IO가 발생 하게 되면, Client는 MDS로 파일 작업에 대한 메타데이터 처리를 요청 하게 되고, Directory 별 혹은 File system 별로 정해진 OSS 서버와 Strip사이즈에 따라서, 하나의 큰 단위 파일을 16K ~ 4GB 까지 다양한 사이즈로 각각의 OST에 저장할 수 있습니다.



3.5. Lustre 성능

- Lustre 는 Hardware의 최대한의 성능을 사용 할 수 있도록 설계 되어 있습니다.
 - ➡ 파일 시스템을 이루고 있는 모든 컴포넌트의 특징
 - ➡ 각각의 레이어에서 발생할수 있는 부하
 - ➡ Pipe Line 형태의 접근



4. Lustre 파일 시스템 Case Study

Sanger Institute

1주에 10~15TB 데이터 처리

1만개 이상의 Job 처리

Open Source 기반의 서비스

제한적인 비용으로 최대 성능

DownUnder GeoSolutions

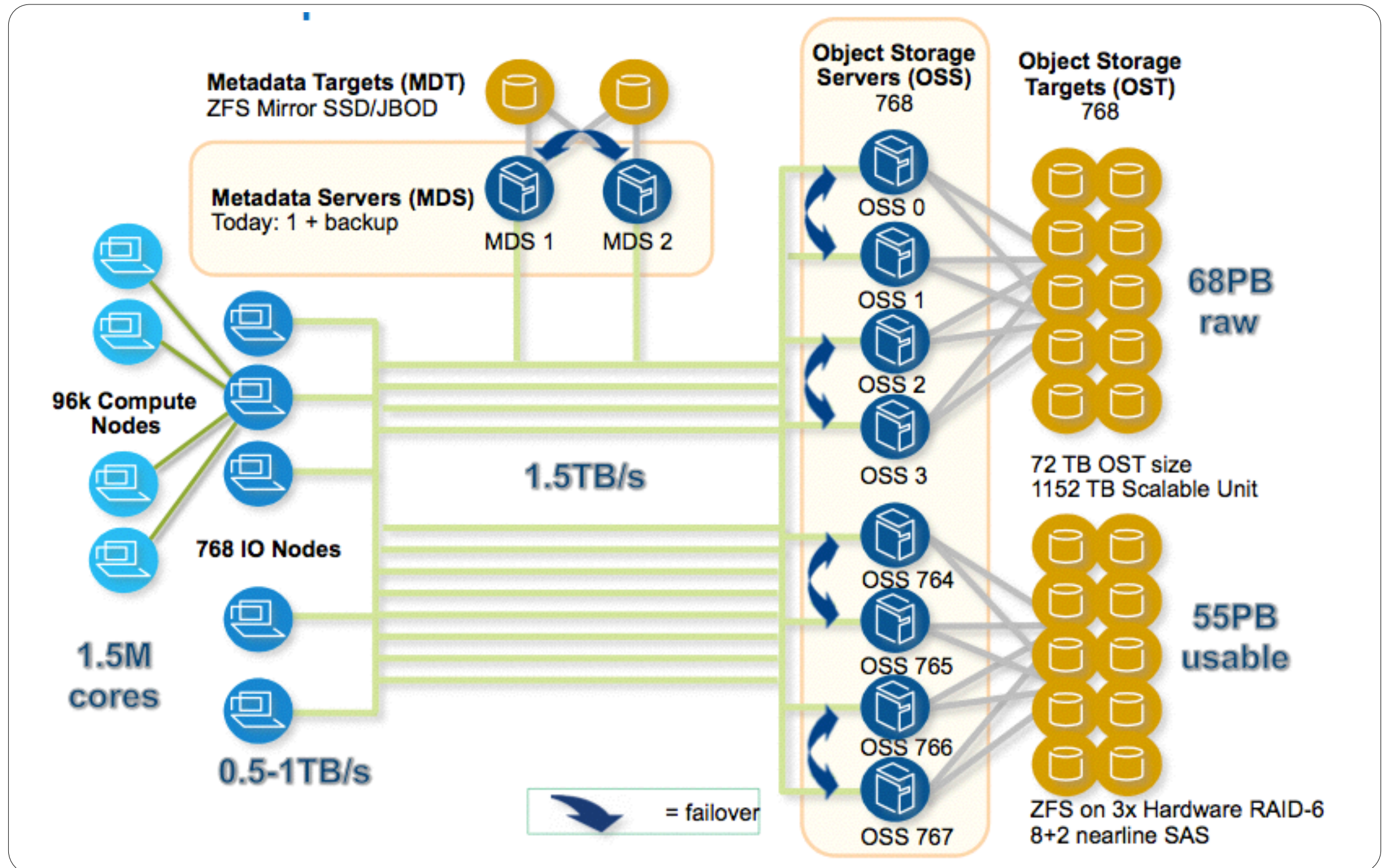
120 명의 임직원

1만2천 코어 클러스터

1.5페타 바이트

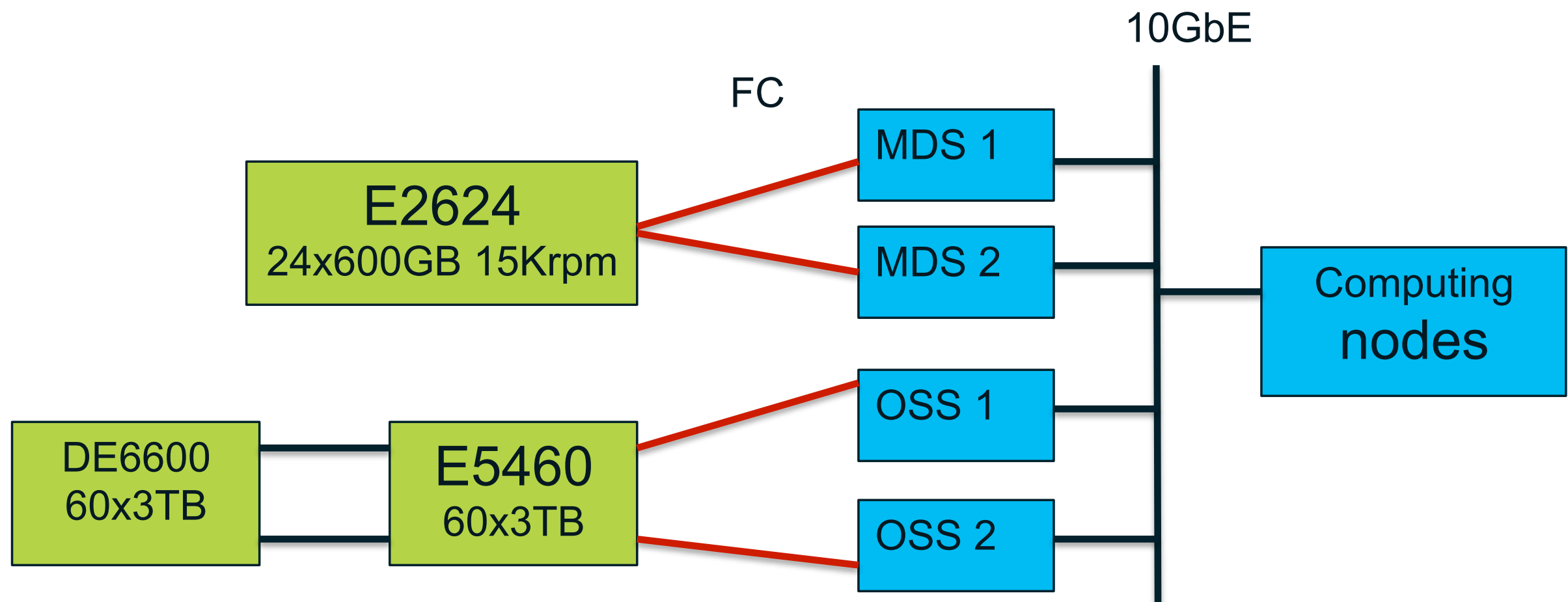
10GB/Sec 성능

4.1. LLNL Sequoia Lustre Architecture



4.2. Performance Sample Design #1

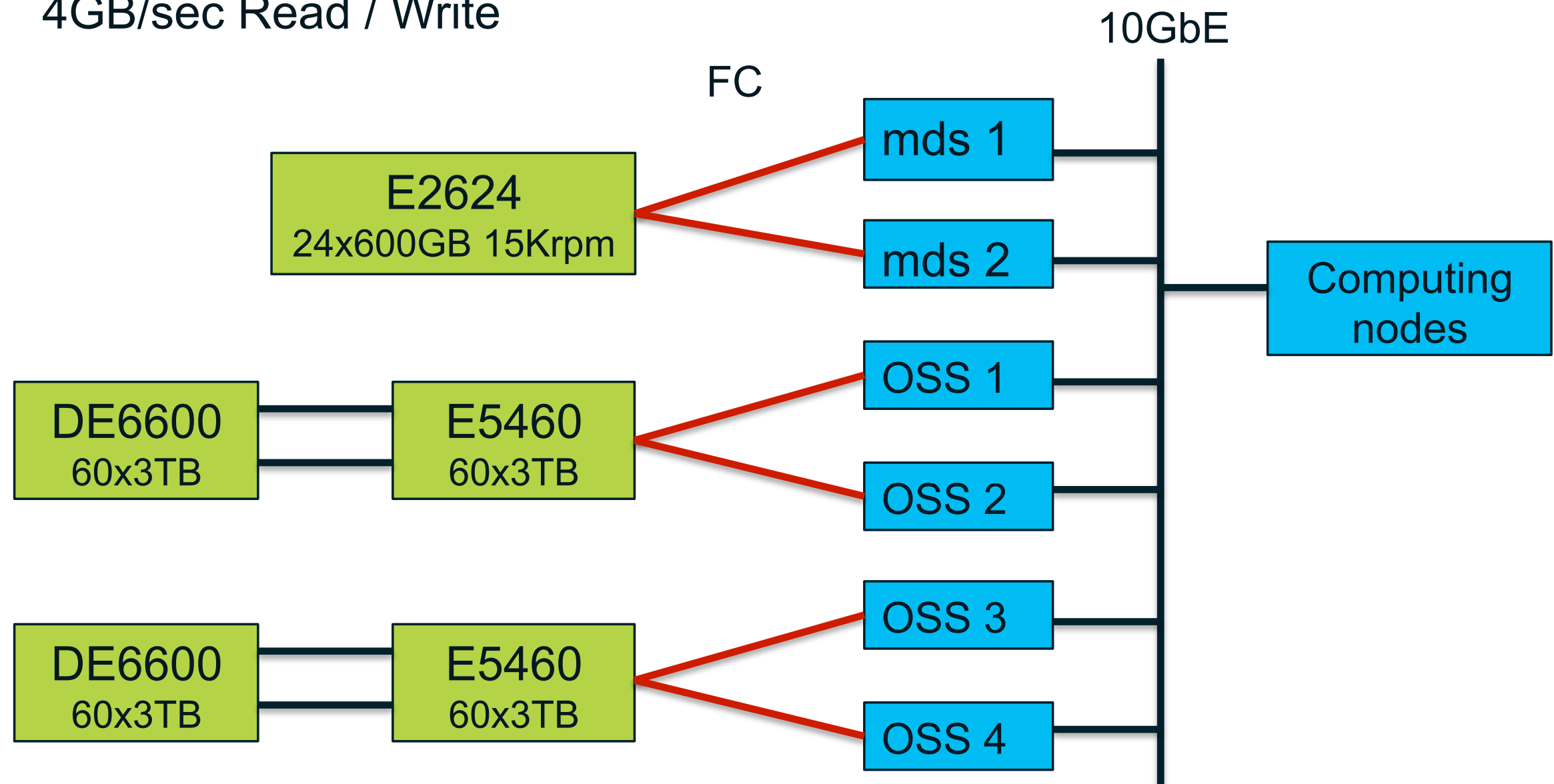
- 2GB/sec Read / Write
- Easier to fit into the existing infrastructure



4.3. Performance Sample Design #2

500TB High Availability Lustre storage

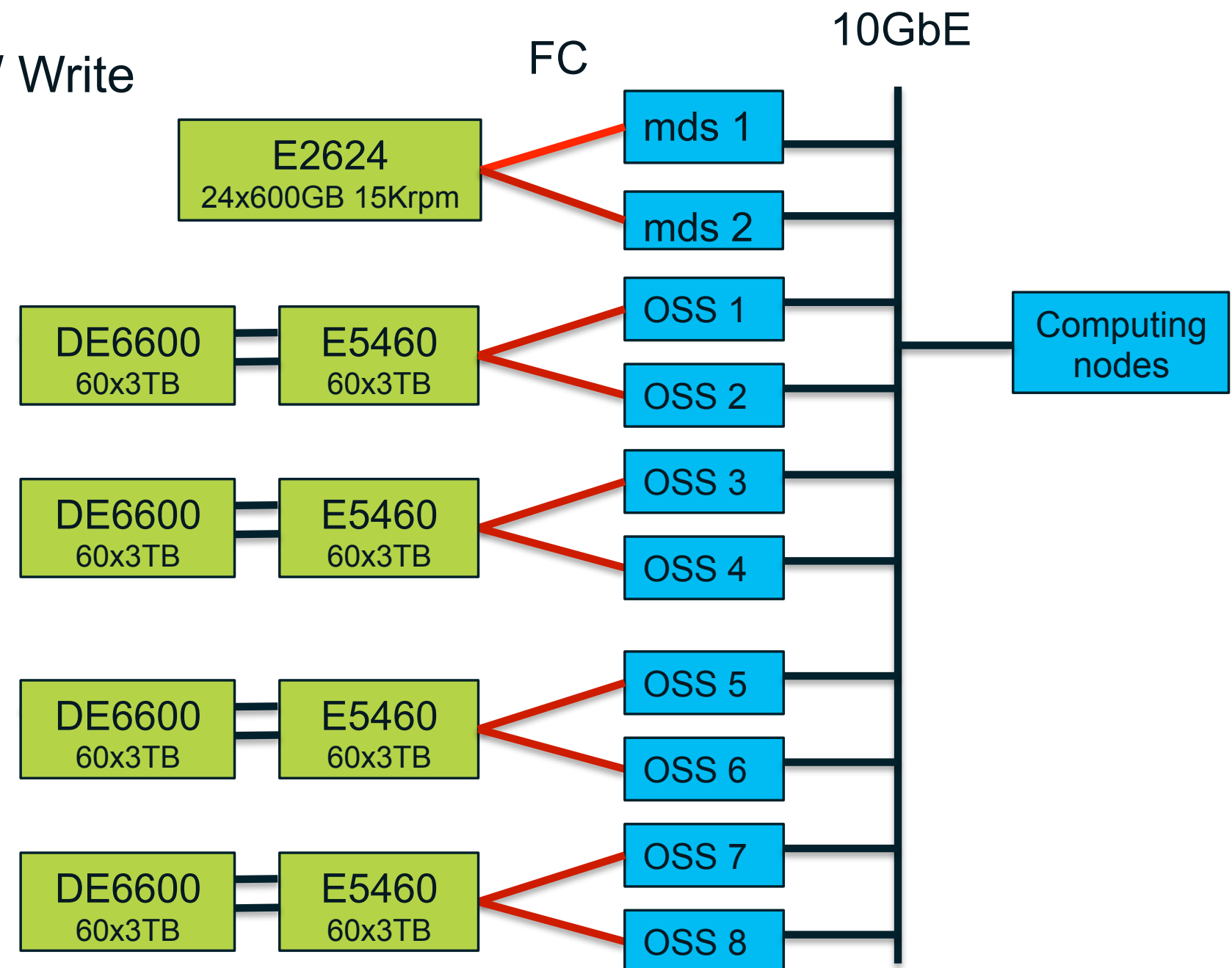
- 4GB/sec Read / Write



4.4. Performance Sample Design #3

1PB High Availability Lustre storage

- 8 GB/sec Read / Write



5. Intel Manager for Lustre (구, chroma)

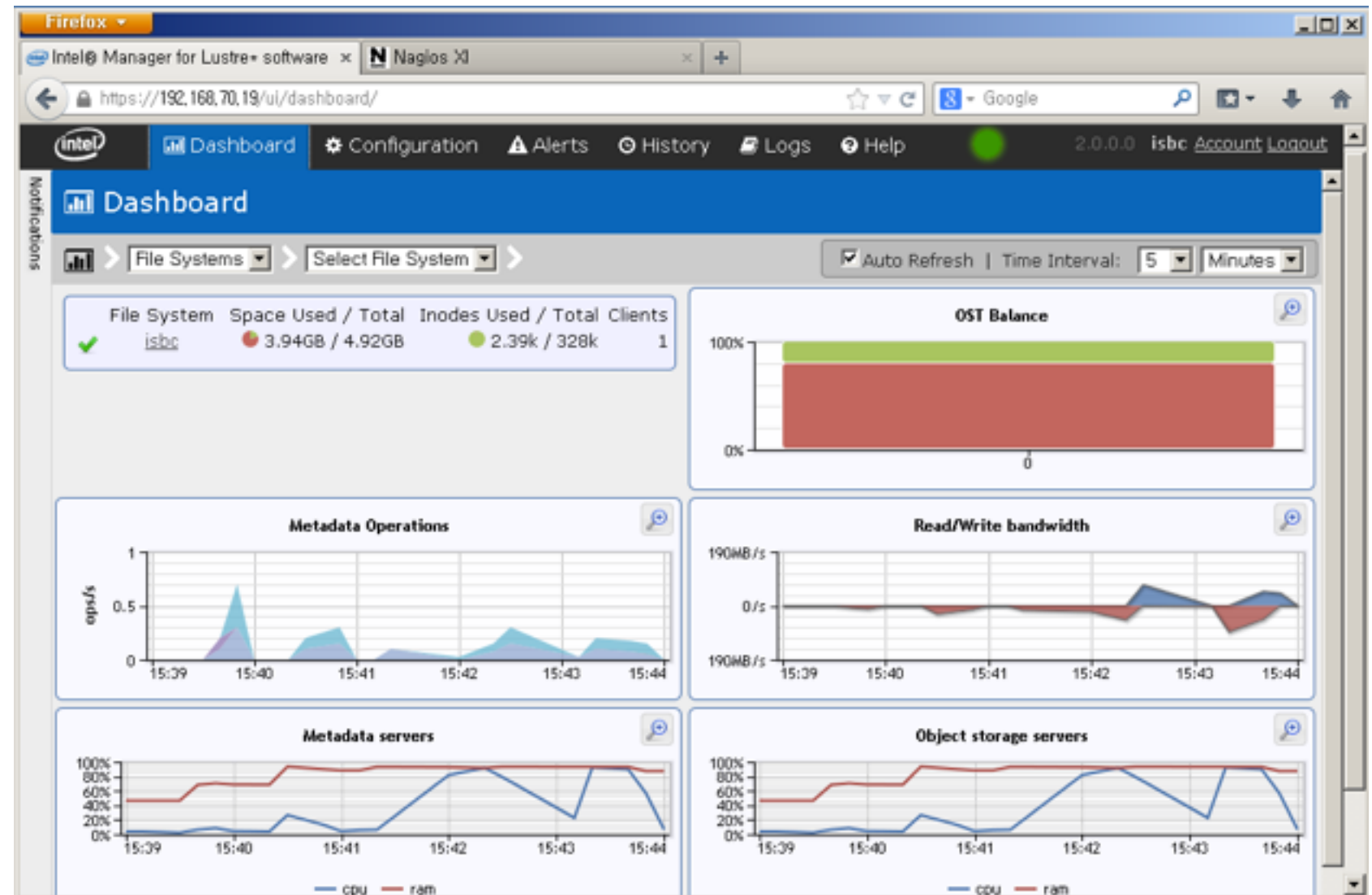
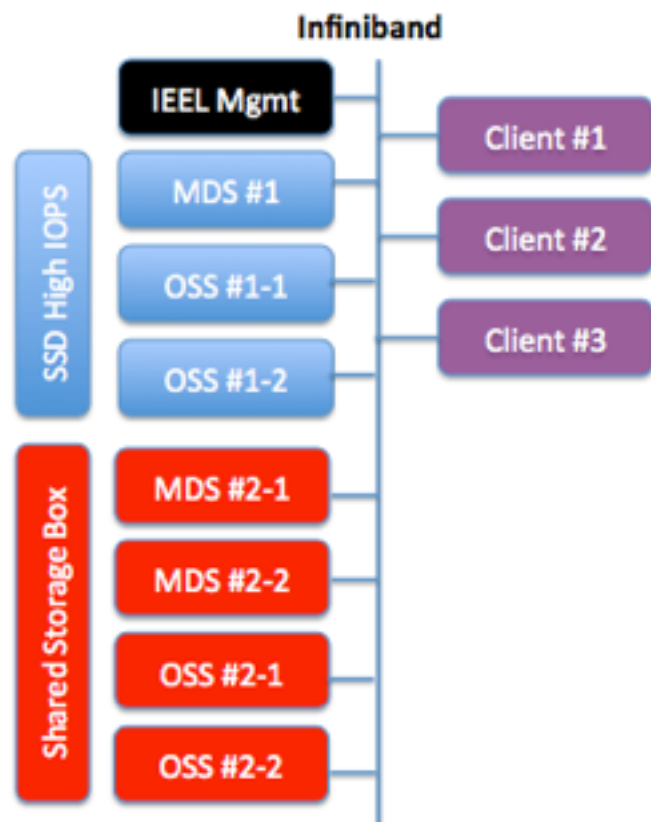
☑ Intel® Manager for Lustre는 손쉬운 관리와 생산적인 운용 환경을 제공합니다.

Dashboard 를 통한 전체 성능 관리

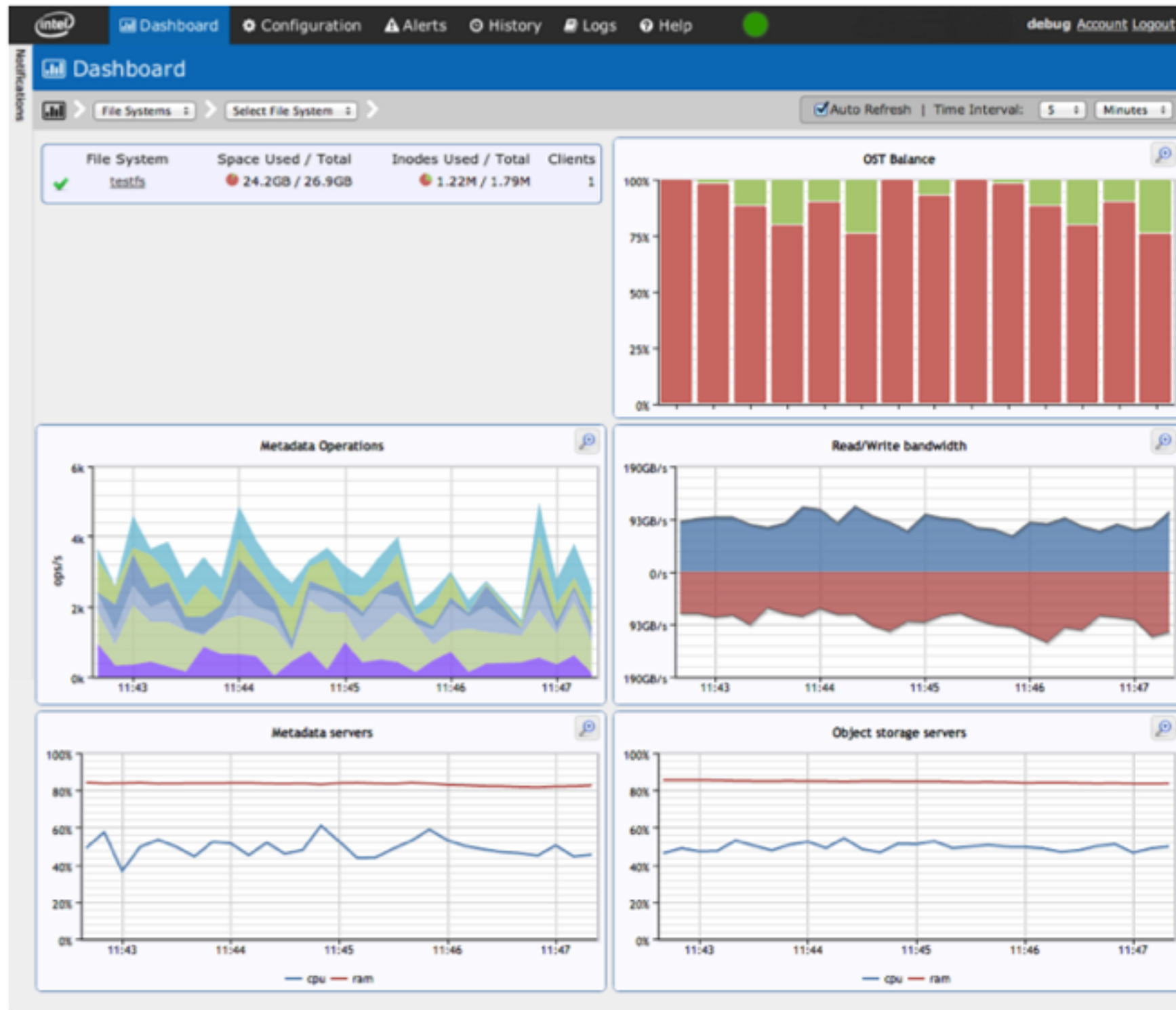
파일 시스템 관리를 위한 웹 인터페이스 및 콘솔 명령어 기반의 인터페이스 제공
중앙에서 파일시스템 구성, 상태, 사용량등을 모니터링

파일 시스템 구성 관리

단일화된 소프트웨어 패키지
설치 및 구성 배포의 용이성



5.1. IML – Dashboard 로그인 초기 화면



5.2. IML – File System 관리, HA 구성

File System tests

Overview

Management Server: [client-28vm2.lab.whamcloud.com](#)
 Metadata Server: [client-28vm2.lab.whamcloud.com](#)
 OSTs: 2
 Alerts: ✓ No alerts
 Actions: [Actions](#)

104MB/1.97GB 96/524k inodes

[Update Advanced Settings](#) [View Client Mount Information](#)

Management Target

Name	Volume	Primary server	Failover server	Started on	
MGS	11ET 00010001	client-28vm2.lab.whamcloud.com	client-28vm3.lab.whamcloud.com	client-28vm2.lab.whamcloud.com	Actions ✓

Metadata Target

Name	Volume	Primary server	Failover server	Started on	
testfs-MDT0000	11ET 00010002	client-28vm2.lab.whamcloud.com	client-28vm3.lab.whamcloud.com	client-28vm2.lab.whamcloud.com	Actions ✓

Object Storage Targets

5.3. IML – Server 구성, 추가 및 삭제

The screenshot displays the Intel Management Library (IML) web interface. The top navigation bar includes links for Dashboard, Configuration (selected), Alerts, History, Logs, Help, and Account. The main content area is titled 'Servers' and shows a list of server entries. The list has columns for Hostname, LNet State, and Actions. The current view shows 4 entries, all with LNet up status.

Hostname	LNet State	Actions
client-28vm2.lab.whamcloud.com	🌿 LNet up	Actions ▾
client-28vm3.lab.whamcloud.com	🌿 LNet up	Actions ▾
client-28vm5.lab.whamcloud.com	🌿 LNet up	Actions ▾
client-28vm6.lab.whamcloud.com	🌿 LNet up	Actions ▾

Showing 1 to 4 of 4 entries

Server Actions

Detect File Systems Re-scan NIDs Re-write Target Configuration Install Updates

5.4. IML – Event History

Notifications

Intel Dashboard Configuration Alerts **History** Logs Help ● debug Account Logout

Build 3033

History

Event History

Host: All Severity: Any Event type: Any Filter

Show 25 entries

	Time	Host	Message
i	Yesterday 21:23	client-28vm6.lab.whamcloud.com	testfs-OST0001 started
i	Yesterday 21:23	client-28vm5.lab.whamcloud.com	testfs-OST0000 started
i	Yesterday 21:23	client-28vm2.lab.whamcloud.com	testfs-MDT0000 started
i	Yesterday 21:15	client-28vm2.lab.whamcloud.com	MGS started
i	Yesterday 20:51	client-28vm5.lab.whamcloud.com	LNet started on server 'client-28vm5.lab.whamcloud.com'
i	Yesterday 20:51	client-28vm6.lab.whamcloud.com	LNet started on server 'client-28vm6.lab.whamcloud.com'
i	Yesterday 20:51	client-28vm6.lab.whamcloud.com	Host is back online client-28vm6.lab.whamcloud.com
i	Yesterday 20:51	client-28vm5.lab.whamcloud.com	Host is back online client-28vm5.lab.whamcloud.com
i	Yesterday 20:51	client-28vm2.lab.whamcloud.com	LNet started on server 'client-28vm2.lab.whamcloud.com'
i	Yesterday 20:50	client-28vm3.lab.whamcloud.com	LNet started on server 'client-28vm3.lab.whamcloud.com'

5.5. IML – Alerts, 파일 시스템의 경고 모니터링

The screenshot displays the Intel IML Alerts interface. The top navigation bar includes links for Dashboard, Configuration, Alerts (selected), History, Logs, Help, Build 3033, debug, Account, and Logout. The main content area is titled 'Alerts' and features a sidebar with 'Notifications'. The 'Active Alerts' section shows a dropdown for 'Show 25 entries' and a table with columns 'Started', 'Entity', and 'Message'. Below this, it states 'No data available in table' and 'Showing 0 to 0 of 0 entries'. The 'Alert History' section also has a 'Show 25 entries' dropdown and a table with columns 'Started', 'Finished', 'Entity', and 'Message'. The table contains 10 rows of alert history data.

Started	Finished	Entity	Message
Yesterday 21:23	Yesterday 21:23	testfs-OST0001	Target testfs-OST0001 offline
Yesterday 21:23	Yesterday 21:23	testfs-OST0000	Target testfs-OST0000 offline
Yesterday 21:23	Yesterday 21:23	testfs-MDT0000	Target testfs-MDT0000 offline
Yesterday 21:15	Yesterday 21:15	MGS	Target MGS offline
Yesterday 20:51	Yesterday 21:14	client-28vm5.lab.whamcloud.com	Updates are ready for server client-28vm5.lab.whamcloud.com
Yesterday 20:51	Yesterday 21:14	client-28vm6.lab.whamcloud.com	Updates are ready for server client-28vm6.lab.whamcloud.com
Yesterday 20:51	Yesterday 20:51	client-28vm6.lab.whamcloud.com	Host is offline client-28vm6.lab.whamcloud.com
Yesterday 20:51	Yesterday 20:51	client-28vm5.lab.whamcloud.com	Host is offline client-28vm5.lab.whamcloud.com
Yesterday 20:51	Yesterday 21:14	client-28vm2.lab.whamcloud.com	Updates are ready for server client-28vm2.lab.whamcloud.com
Yesterday 20:50	Yesterday 21:14	client-28vm3.lab.whamcloud.com	Updates are ready for server client-28vm3.lab.whamcloud.com

5.6. IML – System Logs, Kernel 레벨의 상세 로그

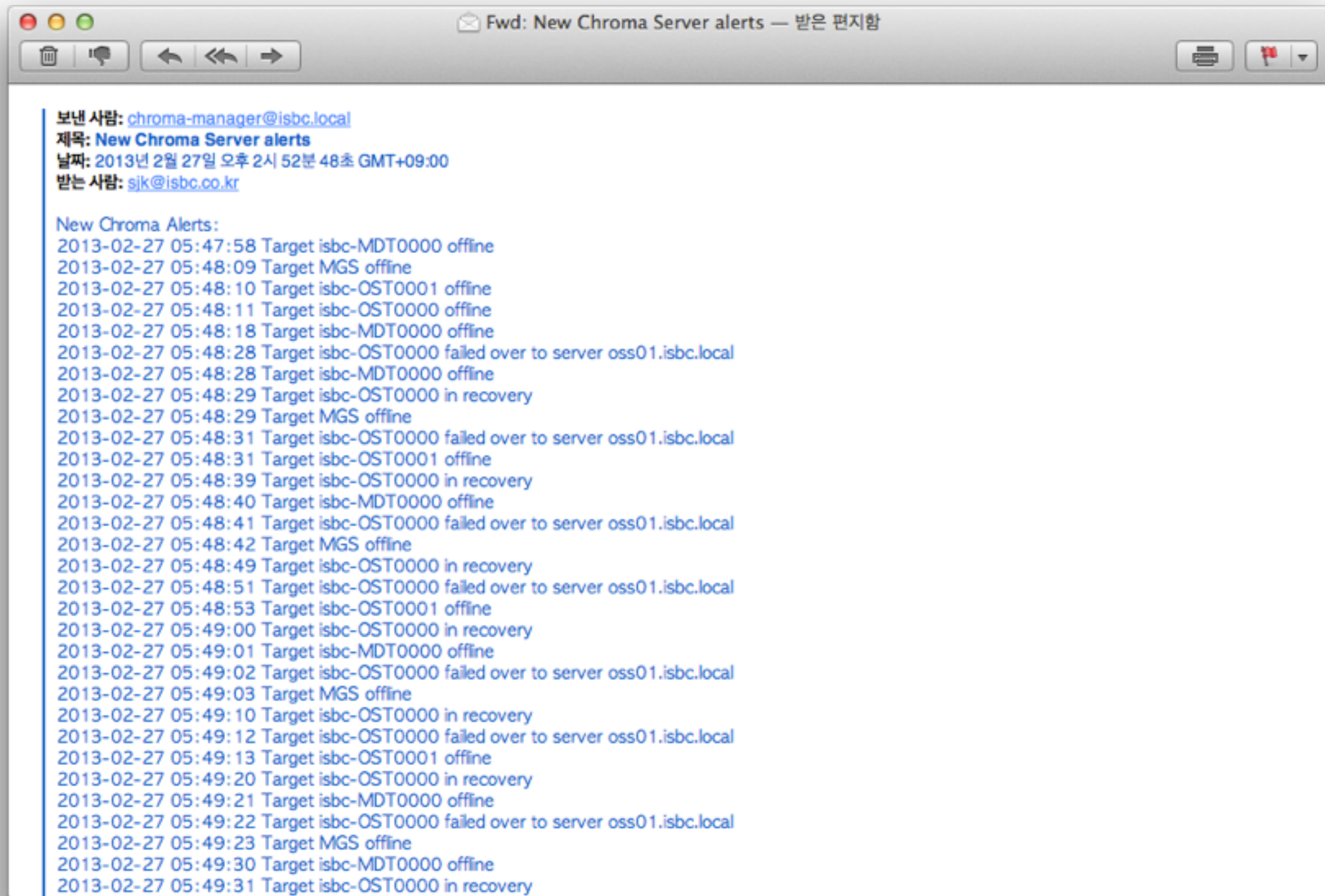
System Logs

Host: From: To: Only Lustre Messages: ☒

Show entries Search:

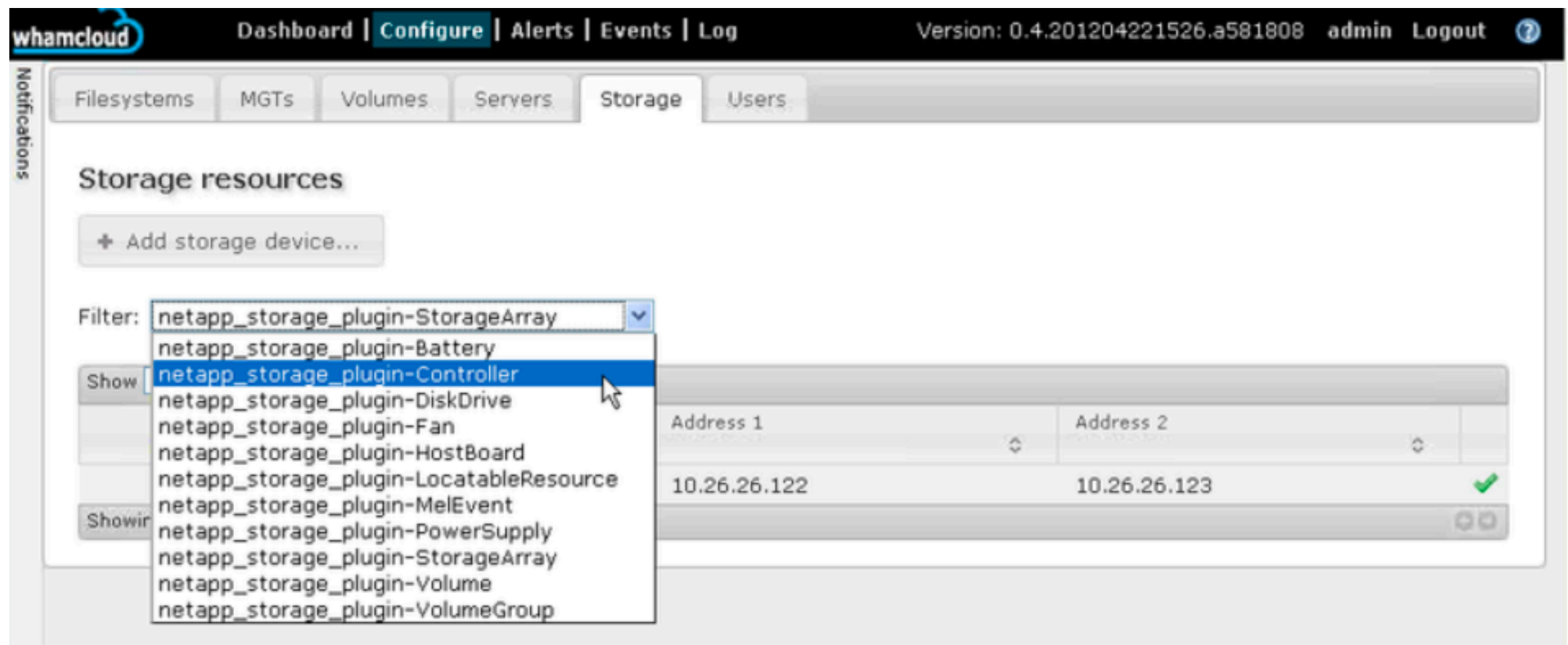
Date	Host	Service	Message
Yesterday 21:23	client-28vm2.lab.whamcloud.com	kernel	Lustre: MDS mdd_obd-testfs-MDT0000: testfs-OST0001_UUID now active, resetting orphans
Yesterday 21:23	client-28vm2.lab.whamcloud.com	kernel	Lustre: MDS mdd_obd-testfs-MDT0000: testfs-OST0000_UUID now active, resetting orphans
Yesterday 21:23	client-28vm6.lab.whamcloud.com	kernel	Lustre: testfs-OST0001: received MDS connection from client-28vm2.lab.whamcloud.com
Yesterday 21:23	client-28vm5.lab.whamcloud.com	kernel	Lustre: testfs-OST0000: received MDS connection from client-28vm2.lab.whamcloud.com
Yesterday 21:23	client-28vm6.lab.whamcloud.com	kernel	Lustre: testfs-OST0001: Imperative Recovery enabled, recovery window shrunk from 300-900 down to 150-450
Yesterday 21:23	client-28vm6.lab.whamcloud.com	kernel	Lustre: testfs-OST0001: Now serving testfs-OST0001 on /dev/sde with recovery enabled
Yesterday 21:23	client-28vm6.lab.whamcloud.com	kernel	Lustre: MGCclient-28vm2.lab.whamcloud.com: Reactivating import
Yesterday 21:23	client-28vm5.lab.whamcloud.com	kernel	Lustre: testfs-OST0000: Imperative Recovery enabled, recovery window shrunk from 300-900 down to 150-450
Yesterday 21:23	client-28vm5.lab.whamcloud.com	kernel	Lustre: testfs-OST0000: Now serving testfs-OST0000 on /dev/sdd with recovery enabled
Yesterday 21:23	client-...	kernel	Lustre: MGCclient-28vm2.lab.whamcloud.com: Reactivating import

5.7. IML - Email 통보기능



5.8. IML – Storage Plug-in Module

IML은 스토리지 벤더에서 제공하는 plug-in을 통합하여 파일시스템과 함께, Volume의 물리적인 상태 까지 확인할 수 있도록 되어 있습니다.



5.9. Lustre Server Monitoring – Hardware and OS Part (Nagios XI)

Nagios XI를 사용하여, Lustre Server 의 Hardware 및 OS를 모니터링 가능 합니다.

- Intel IEEL v1.0 IML 은 Lustre 파일 시스템 구성 및 운용을 담당
- Nagios XI 는 Lustre Server가 설치된, HW/OS 등을 담당 (HP ILO4, IBM IMM, Dell DRAC 및 Onboard RAID 부문 담당)

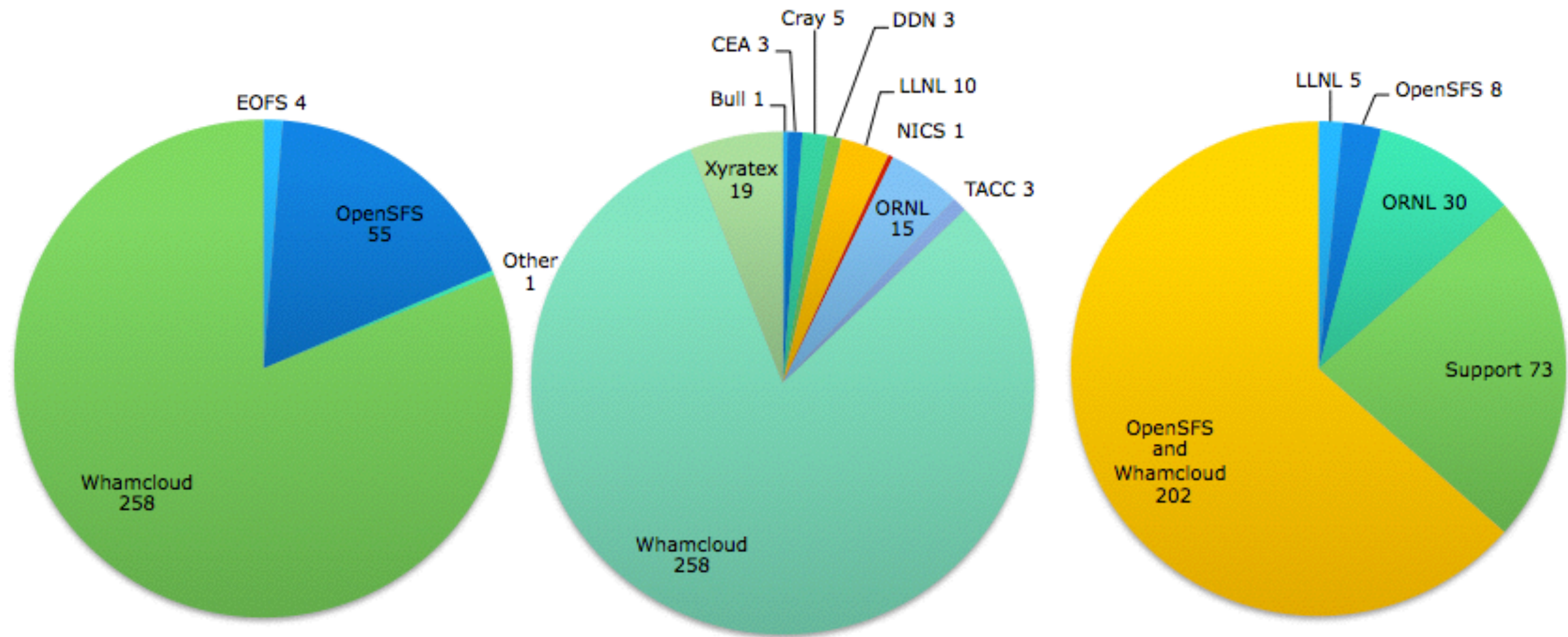
** Nagios XI는 별도 추가 구매 필요

The screenshot displays the Nagios XI web interface. The top navigation bar includes links for Home, Views, Dashboards, Reports, Configure, Tools, Help, and Admin. The user is logged in as 'nagiosadmin'. The main content area shows a search for 'lustre' with 12 matches. The table below lists the monitored services and their status.

Host	Service	Status	Duration	Attempt	Last Check	Status Information
lustre	/Disk Usage	Ok	35m 8s	1/5	2013-09-05 15:44:03	DISK OK - free space: / 15205 MB (90% inode=96%):
	CPU Stats	Ok	34m 38s	1/5	2013-09-05 15:44:33	CPU STATISTICS OK: user=3.14% system=0.81% iowait=0.71% idle=95.34%
	Cron Scheduling Daemon	Ok	3m 8s	1/5	2013-09-05 15:46:03	crond (pid 1554) is running...
	Load	Ok	33m 38s	1/5	2013-09-05 15:45:38	OK - load average: 11.16, 6.05, 2.62
	Memory Usage	Ok	33m 8s	1/5	2013-09-05 15:46:03	OK - 696 / 992 MB (70%) Free Memory, Used: 851 MB, Shared: 0 MB, Buffers: 4 MB, Cached: 555 MB
	Open Files	Ok	32m 38s	1/5	2013-09-05 15:46:33	OK: 1312 open files (1% of max 98092)
	Ping	Ok	32m 8s	1/5	2013-09-05 15:47:04	OK - 192.168.70.21: rta 0.301ms, lost 0%
	SSH Server	Ok	31m 38s	1/5	2013-09-05 15:47:33	openssh-daemon (pid 1449) is running...
	Swap Usage	Ok	31m 8s	1/5	2013-09-05 15:48:03	SWAP OK - 98% free (1962 MB out of 2015 MB)
	Total Processes	Ok	30m 38s	1/5	2013-09-05 15:48:33	PROCS OK: 203 processes
	Users	Ok	34m 53s	1/5	2013-09-05 15:44:18	USERS OK - 2 users currently logged in
	Yum Updates	Ok	3m 23s	1/5	2013-09-05 15:45:48	YUM OK: O/S is up to date.

The interface also includes a sidebar with navigation options like Home Dashboard, Tactical Overview, and various problem views. The footer shows the Nagios XI version (2012R2.3) and copyright information.

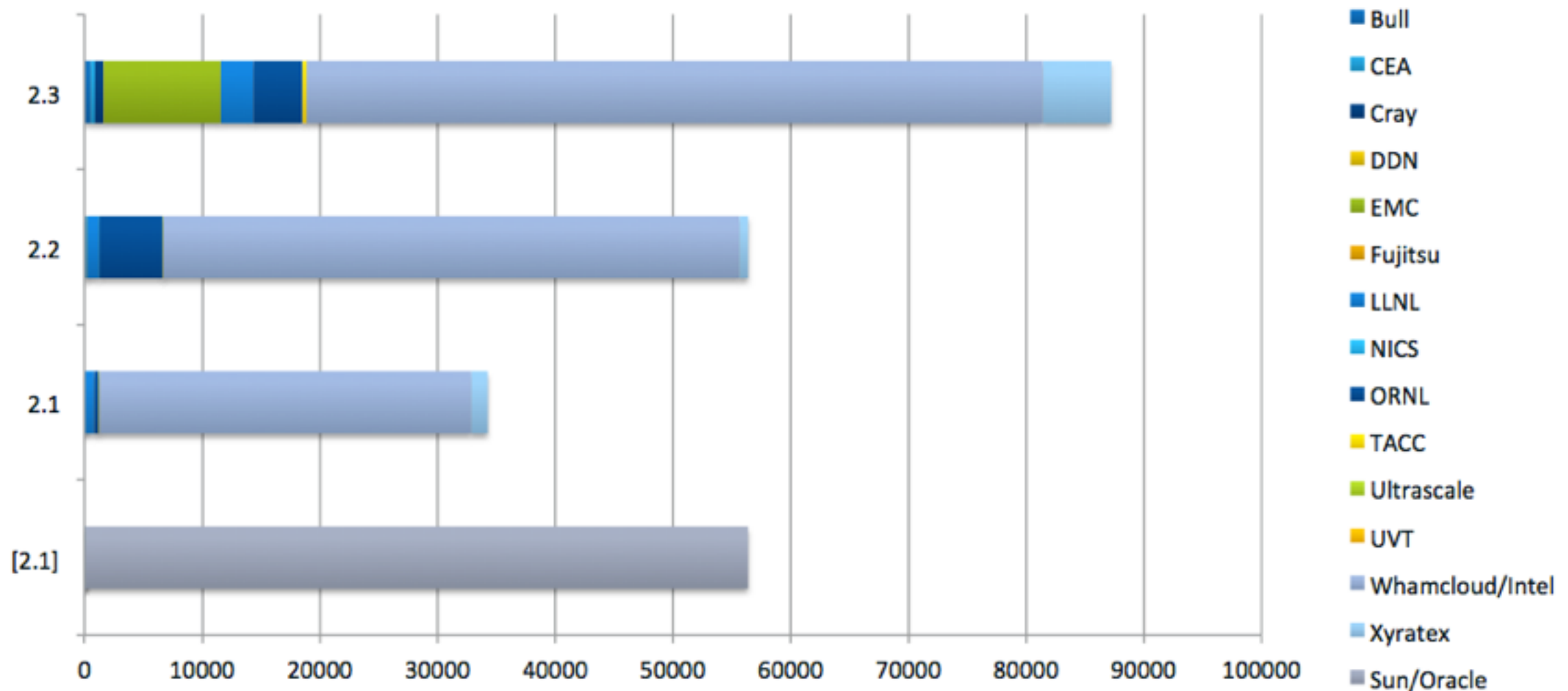
6. Lustre 2.2 커뮤니티



- http://git.whamcloud.com/?p=fs/lustre-release.git;a=shortlog;h=refs/heads/b2_2
- <http://wiki.whamcloud.com/display/PUB/Lustre+Community+Development+in+Progress>
- <http://wiki.whamcloud.com/display/PUB/LUG+2012>

6.1. Evolving Lustre Community

An increasing number of organizations contribute



<http://git.whamcloud.com/?p=fs/lustre-release.git;a=summary>

7. Lustre 로드맵.



IEEL

1.1

Basies on Lustre 2.5 Release

- Include Previous IEEL Features
- HSM
- Dynamic LNET Configuration
- Enhanced Alerts & Logging
- Intel Xeon Phi Support

2.6

LFSCCK MDT-OST Consistency
UDI Mapping & Shared Key

2014.2Q

2.5

LFCK MDT-OST Consistency
HSM

1.0

Base on Lustre 2.3 Release and 2.4 Client

- Lustre file system plus Selected Stability patch
- Intel Manager for Lustre Software
- Hadoop Adapter for Lustre Software
- Enterprise Support enhancements
- Maintenance Releases
- RHEL6.4 Support Only

2.4

OSD restructuring / Support ZFS(No GPL)
Distributed namespace (Remote Directory)
Online Check/Scrub
Network Request Scheduler 4MB I/O RPC

2013.3Q

2.3

Imperative recovery
Metadata Performance
Server Stack SMP Scaling
Enhanced CheckSum
Quota Accounting

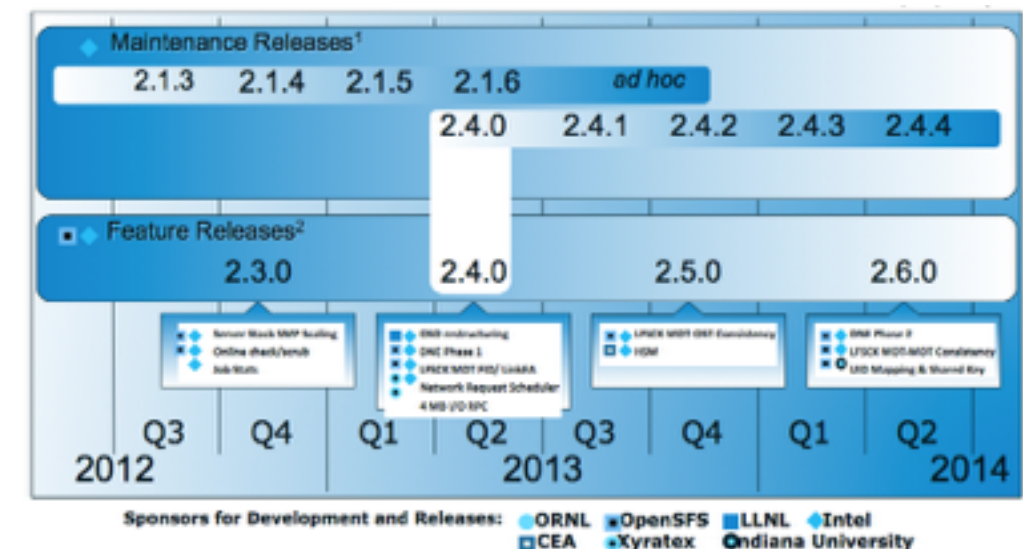
2.2

LNEXT SMP Scaling
Client Parallel CheckSums
Parallel Directory Operation

2.1

Full RHEL6 Support
Asyc Journal Commits on By default
ext4 by default
128TB Luns
Stability enhancements

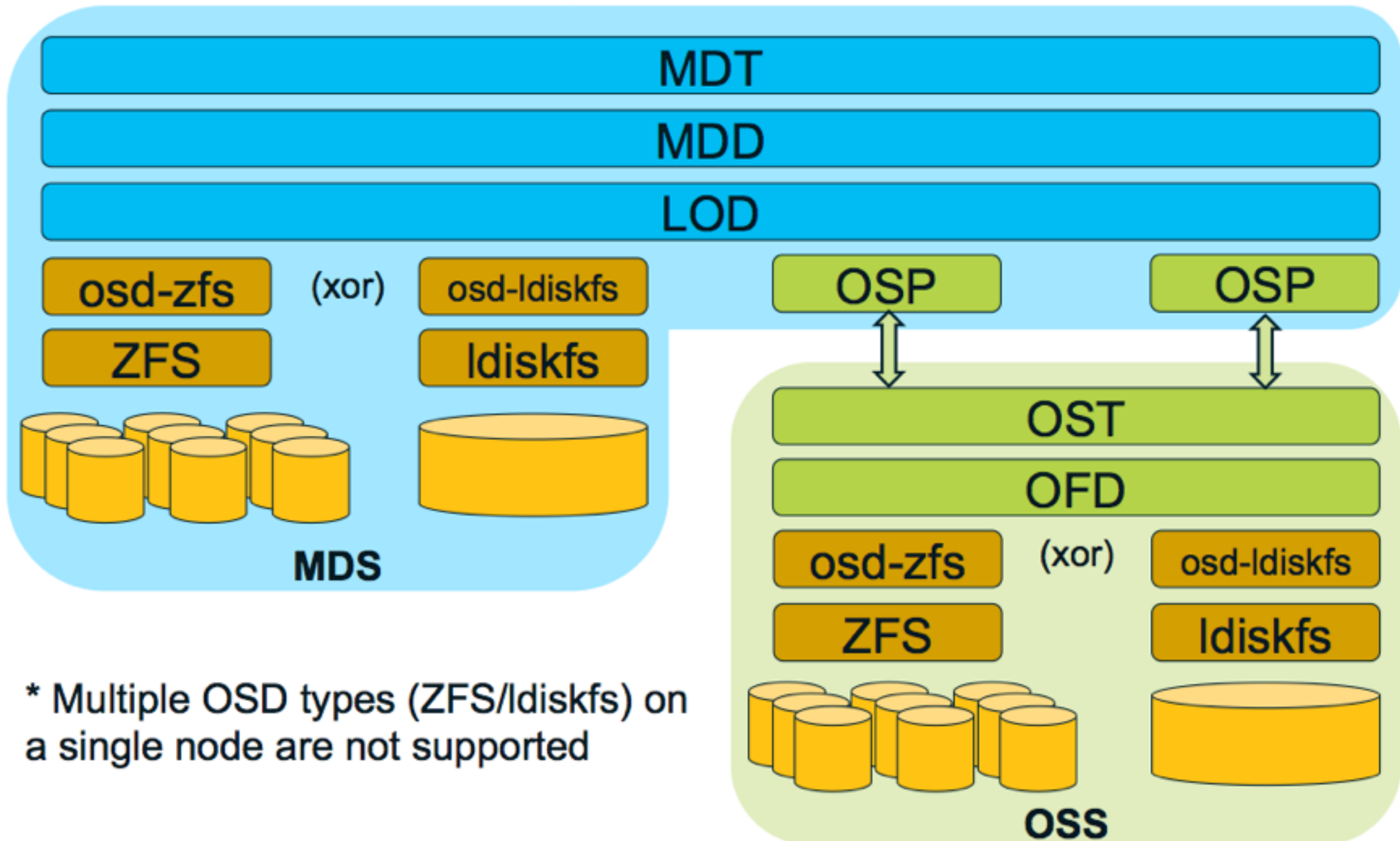
Community



7.1. Lustre 2.4 – ZFS OST/MDT Backing Storage

- Can Leverage many ZFS feature in Lustre 2.4
 - ▶ Robust code with 10+ years maturity
 - ▶ Data checksums on disk + Lustre checksums on network
 - ▶ Online filesystem check/scrib/repair – no more e2fsck!
 - ▶ Scales beyond current filesystem limits (Object count/size, file system size)
 - ▶ Easier management of many disks, commodity JBODs without RAID hardware
 - ▶ Integrated with flash storage cache (L2ARC read Cache)
 - ▶ Data compression improves real-world IO performance and space utilization
- Other features will need extra effort to work with Lustre
 - ▶ <http://en.wikipedia.org/wiki/Zfs>
 - ▶ <http://zfsonlinux.org/lustre.html>

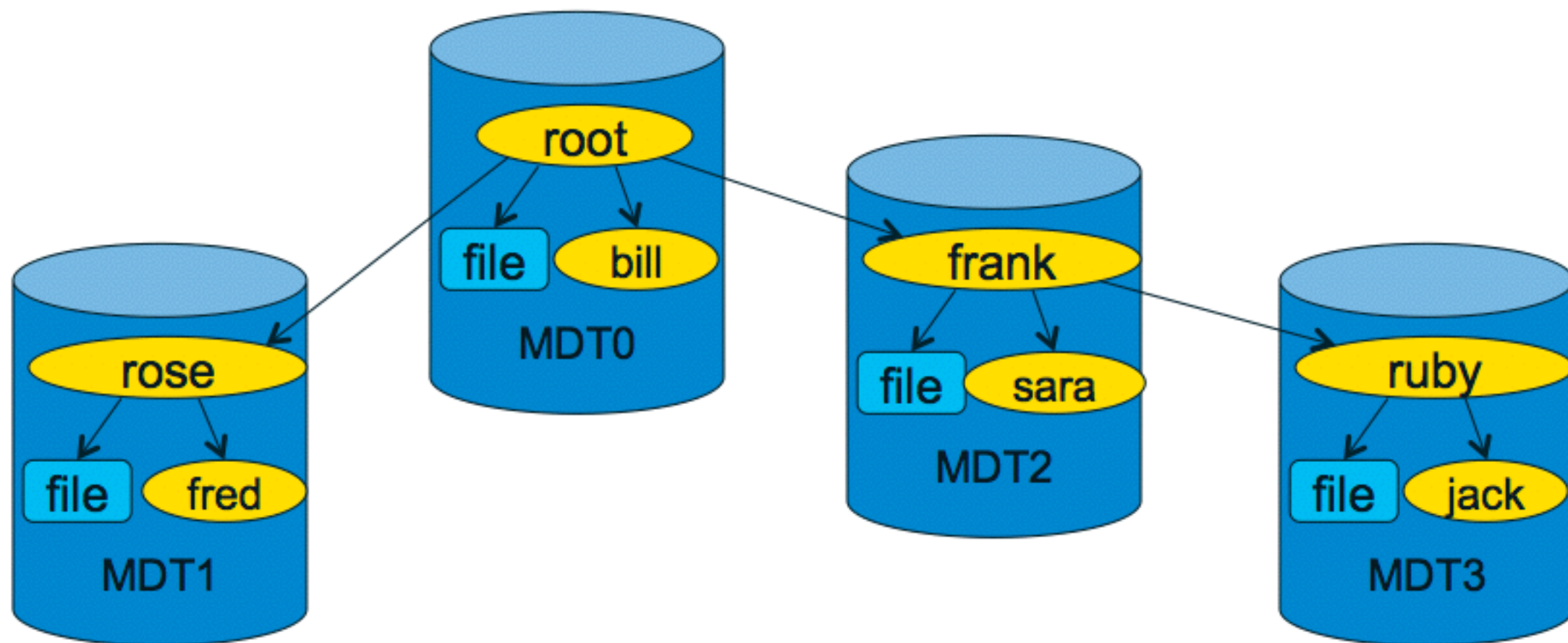
7.2. Lustre 2.4 – Updated MDS/OSS Module Layering



MDD - Metadata Device Driver / LOD - Logical Object Device / OSP - Object Storage Proxy
 OSD - Object Storage Device / Object Filter Device

7.3. Lustre 2.4 –Remote Directory (2.4)

- Subdirectories on remote MDT created by administrator
- Scales namespace in similar manner to data servers
- Allows isolated metadata performance for users/jobs
- Shared OST IO bandwidth among all files on all MDTs



7.4. Lustre 2.4 Other Updates

- LFSCK phase 1.5
 - LFSCK is online Lustre File System Check Tool
 - Verify Lustre File Identifier(FID) stored in each directory entry
 - Verify inode -> parent back-pointer in link xattr
 - Piggy-backs on OI Scrub inode iteration
 - Verifies MDT/OST FID attr / OST object
- Network Request Scheduler (NRS)
 - Recoder RPC request on the Server
 - Currently implements multiple optimization policies
 - Allows other policies to be developed in the future(2.5)
- Linux Kernel Client Updates (2.4/2.5)
 - Desire to include Lustre client in upstream Linux kernel
 - Need to clean up ten years of legacy code to Linux style
 - EMC is doing most of this work (Change Landing Incrementally in 2.3/2.4/2.5)



THANK YOU

MERCI
DANK
GRAZIE
BEDANKEN
IGRACIAS!

