

Topics from LUG 2019, ISC19, LAD & Lustre Developer Summit 2019

Shinji Sumimoto, Ph.D.
FUJITSU LIMITED Oct. 17th, 2019



- Lustre 20 year's Anniversary
 - ISC 19 workshop slides from Andreas

- Fugaku Update and FEFS
(Lustre based File System) Issues and Directions

- Topics from LUG19 and LAD19
 - LNET Multi-rail Improvement

Lustre 20 year's Anniversary

- 1999: Lustre file system 研究プロジェクト開始 by Peter J. Braam
- 2001: Cluster File Systems社設立
- 2007/9: Sun社がCluster File Systems社を買収
- 2009/4: Oracle社がSun社を買収、コミュニティに不安が走る
- 2010/4: Oracle社がサポートを自社ハードに制限、コミュニティ激震
 - 欧州EOFS、米国国研ベースOpenSFS、World WideユーザベースHPCFSの3つのコミュニティに分散
- 2010/9: Whamcloud社設立
 - Oracle社からWhamcloud社に技術者が集結
- 2010/12: Oracle社がLustre開発凍結
- 2011/4: LUG2011で3つのコミュニティがOpen SFS+EOFSとして再始動
- 2012/7: Intel社がWhamcloud社を買収
- 2018/6: DDN社がIntelのLustre事業を買収、新生Whamcloud誕生
- 2019: Luster 20 years anniversary

LUG2009: Lustre 10th Anniversary



LUG 2011: Single Community



Next steps

- Conduct weekly teleconference calls throughout the process
- Draft required changes to OpenSFS bylaws and contributor agreement with input from OpenSFS and HPCFS participants
 - Also require input from those in due-diligence process

changes to OpenSFS bylaws and contributor agreement


meeting mid-May in US with EOFS

intends to join OpenSFS at the promoter with board seat


OpenSFS.org 3

A Single Community with Global Participation

Guarantees the continued preeminence of the Lustre file system, now and in the future



EOFS
European Open File System



OpenSFS
Open Scalable File Systems, Inc.

April 2011 OpenSFS.org 4

Proposed changes to OpenSFS to facilitate merging HPCFS within OpenSFS

- 1) review contributor agreement - key goal is to make this simple. Can we do a GPL-V2 signoff only? Do we need a contributor agreement given the changes in the community?
- 2) review patent language - key goal is to constrain patent language as much as possible while meeting our agreed upon principal. Agreed upon principal is to protect end users of the OpenSFS codebase "stack" from patent litigation from any participant in OpenSFS
- 3) Governance model - establish a single board seat to represent adopter and supporter level participants
 - Additional board seats established as adopter and supporter participation grows (as currently covered in our bylaws)
- 4) Further our close relationship with EOFS via memorandum of understanding or other arrangement to facilitate improved collaboration and alignment

OpenSFS.org 2



2018/6: DDN社がIntelのLustre事業を買収





Lustre: The Next 20 Years

Andreas Dilger
Principal Lustre Architect
CTO Whamcloud



A Long Time Ago, In A Company Far, Far Away...

Stelias Computing



- Home
- About
- Projects
- News
- Press
- Relationships
- Contact

```
commit 139a736b8b98d38be9265b6e1fcf6b54ee4c0c71 (tag: refs/tags/0.0.0)
Author: pschwan <pschwan>
Date: Thu Jun 3 01:34:05 1999 +0000

    foo

commit 906f38ee0bc23a1b153fde2e9bf1063ccb0f40c9 (0.0.0-22-g906f38e)
Author: adilger <adilger>
Date: Wed Dec 1 18:45:16 1999 +0000

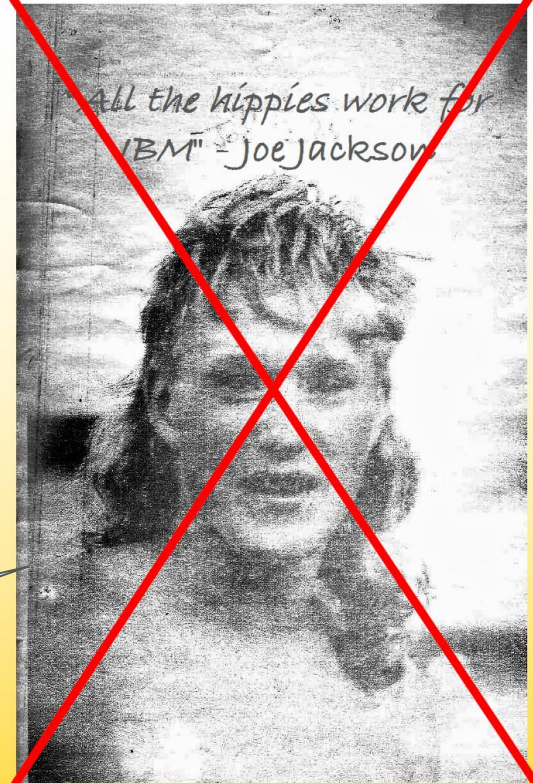
    Updated parameters for obdfs_writepage() to use struct *dentry instead of

commit 667e9cd3c1193c9e858512ced5ebccd26e0e6ab2 (0.0.0-108-g667e9cd)
Author: braam <braam>
AuthorDate: Sun Mar 11 00:53:49 2001 +0000

    a working file system!
```

2 Source: <http://www.myshared.ru/slide/137795/>

Someone Had A Modest Goal...



Young Andreas!?

Cluster File Systems, Inc



Someone Had A Modest Goal...

Lustre

The Inter-Galactic File System

Peter J. Braam

braam@clusterfs.com

<http://www.clusterfs.com>

Cluster File Systems, Inc

Source: [Lustre, The Inter-Galactic File System](#), Peter Braam, 2002



And Brought It All Together...



company information

- [search](#)
- [contact hp](#)
- [company information home](#)
- [about hp](#)
- [hp in the community](#)
- [hp labs](#)
- [investor relations newsroom](#)
- [executive team](#)

U.S. Department of Energy Agency Selects HP to Co-develop Linux Software for Clustered Computing

File System Software Targeted for Use on Clusters of up to 10,000 Systems

PALO ALTO, Calif., Aug. 8, 2002

HP (NYSE:HPQ) today announced it has been chosen by the U.S. Department of Energy's (DOE) National Nuclear Security Administration (NNSA) to develop and deploy file system software for Linux clusters.

The joint research and development effort between HP and NNSA to develop the software, code-named Lustre, is a three-year project. HP is supplying program management, development and test engineering, hardware, services and support to the initiative in a cost-sharing arrangement with NNSA and the DOE labs, including Lawrence Livermore National Laboratory, Los Alamos National Laboratory and Sandia National Laboratories. HP is working in conjunction with Cluster File Systems, Inc., which is serving as a subcontractor on the Lustre project.

Lustre is a high-performance, highly scalable, Linux-based file system designed to work on large compute clusters that provide more than 100 teraflops with high demand for storage and input/output performance. Lustre will be made available initially to each of the DOE labs, including the Pacific Northwest National Laboratory, on HP's Linux-based high-performance computing cluster solutions.

Stelias Computing in Peter Braam's basement Object-based storage project for Seagate

- Ethernet-connected HDDs with embedded Linux
- OBDfs developed from 1999/06 to 2000/03
- ext2 filesystem split in half with OBD API between
- Basic local-storage IO functionality demonstrated

Brief interlude at TurboLinux in Santa Fe, NM

CFS formed for ASCI Path Forward 2001/03

- One client+HDDs turned into distributed parallel fs
- US DOE required larger partners for project credibility
 - Enter HP+Intel for program mgmt., testing, etc. 2002/08

First production use on MCR (#3) at LLNL

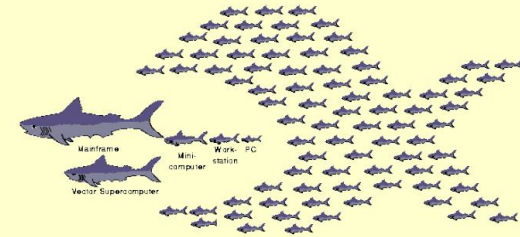
- First testing 2002/08, production 2002/11, 1.0 2003/12

Second install HPCS2 (#5) at PNNL in 2003/07

- Team lived & developed onsite for two weeks



Top500 Cluster



MCR LINUX CLUSTER
LLNL, LIVERMORE, CA
LINUX NETWORKX/QUADRICS
R_{max}: 5.69 TFlops

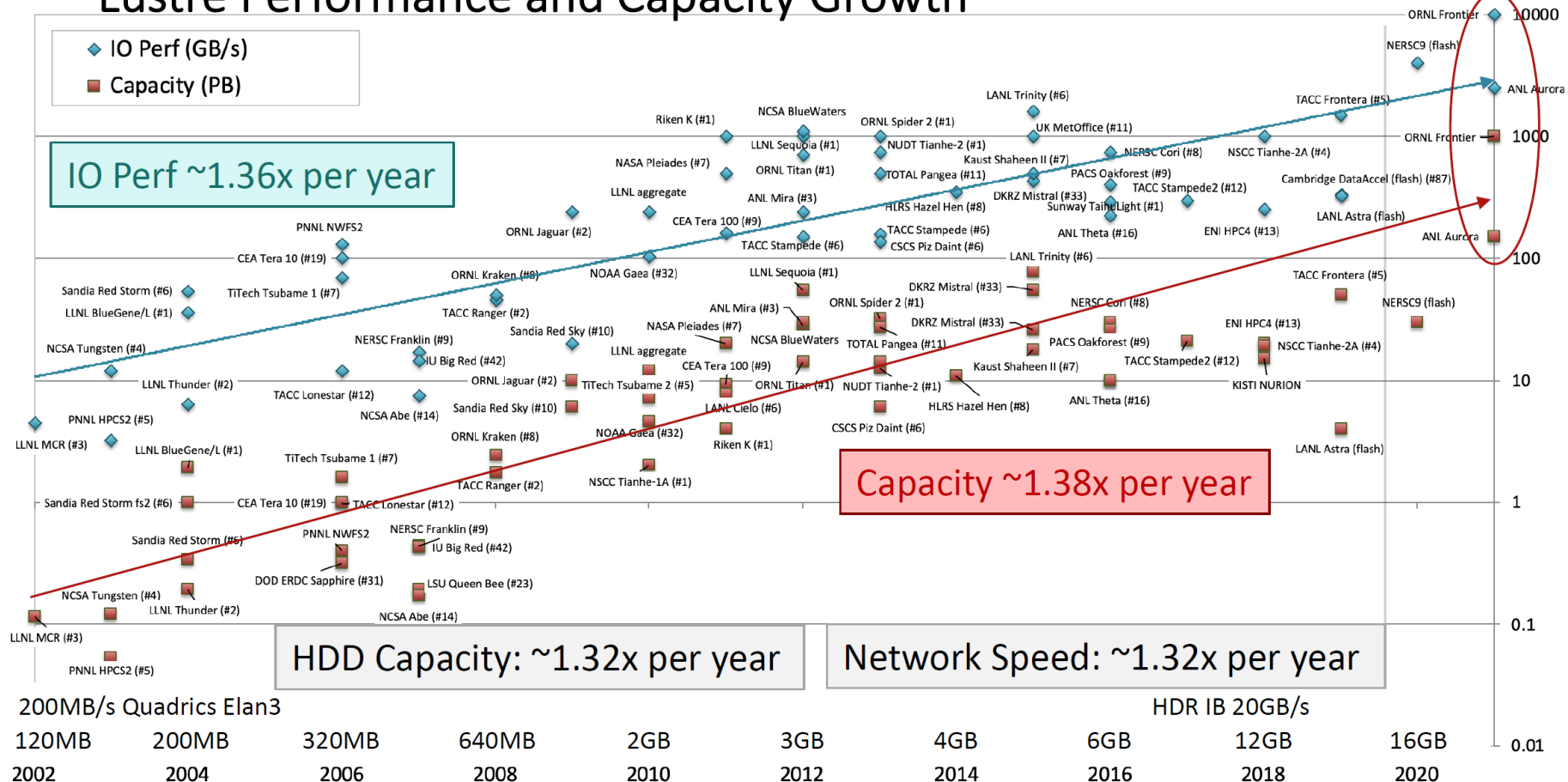


- 11.2 Tflops Linux cluster
- 4.6 TB of aggregate memory
- 138.2 TB of aggregate local disk space
- 1152 total nodes plus separate hot spare cluster and development cluster
 - 4 GB of DDR SDRAM memory and 120 GB Disk Space per node
- 2,304 Intel 2.4 GHz Xeon processors
- **Cluster File Systems, Inc. supplied the Lustre Open Source cluster wide file system**
 - 115TB capacity, 4.48GB/s peak I/O speed
- Cluster interconnect: QsNet ELAN3 by Quadrics,

A similar cluster with Myrinet connection announced for Los Alamos National Lab, planned for 2006

Source: [From the Earth Simulator to PC Clusters](#), Desy, SC'02

Lustre Performance and Capacity Growth



Lustre Feature Roadmap

Lustre (Lite) 1.0 (Linux 2.4 & 2.6)	Lustre 2.0 (Linux 2.6)	Lustre 3.0
2003	2004	2005
Failover MDS	Metadata cluster	Metadata cluster
Basic Unix security	Basic Unix security	Advanced Security
File I/O very fast (~100's OST's)	Collaborative read cache	Storage management
Intent based scalable metadata	Write back metadata	Load balanced MD
POSIX compliant	Parallel I/O	Global namespace

20 - NSC 2003 Source: [The Lustre Storage Architecture](#), Peter Braam, 2003

Cluster File Systems, Inc 

Lustre Release Plan

Lustre 1.6 Supported
until end April, 2010

April 2009

Lustre 1.8.0

- OST Pools
- OSS Read Cache
- Adaptive Timeouts
- Version based recovery (VBR)

Lustre 1.8.x

- Simplified Interoperation with 2.x

RHEL 5, SLES 10

RHEL 6 in 1.8.x after RHEL 6 GA

SLES 11 in 1.8.x

1.4 EOL is June 2009

1.6 EOL is 12 months after 1.8 GA

Q4 2009

Lustre 2.0.0

- Server and Client Restructure for CMD and ZFS
- Clustered MetaData Early Evaluation (No Recovery)
- Security GSS Early Evaluation
- Server Change Logs
- Commit on Share
- MDS Performance Enhancements

RHEL 5 & 6,
SLES 10 & 11

2010

Lustre 2.x Release(s)

- ZFS Lustre GA
- Improved SMP Scaling
- Clustered MetaData Early Eval w/Recovery
- Size on MDS
- Imperative Recovery

Release in 2.x or 3.x Depending on Readiness

- HSM/HPSS
- HSM/SAM-QFS
- Windows Native Client
- Security GA
- Network Request Scheduler
- pNFS Exports
- Scalable health monitoring

2011+

Lustre 3.0.0

- Clustered MetaData GA
- Beginning of Other HPCS Enhancements

Lustre 3.x Release(s)

- Online Data Migration
- Write Back Cache
- Proxies

Lustre

The Inter-Galactic File System

Peter J. Braam

braam@clusterfs.com

<http://www.clusterfilesystems.com>

Cluster File Systems, Inc



Key requirements

- I/O throughput – 100's GB/sec
- Meta data scalability – 10,000's nodes, ops/sec, trillions of files
- Cluster recovery – simple & fast
- Storage management – snapshots, HSM
- Networking – heterogeneous networks
- Security – strong and global

The first 3 years...

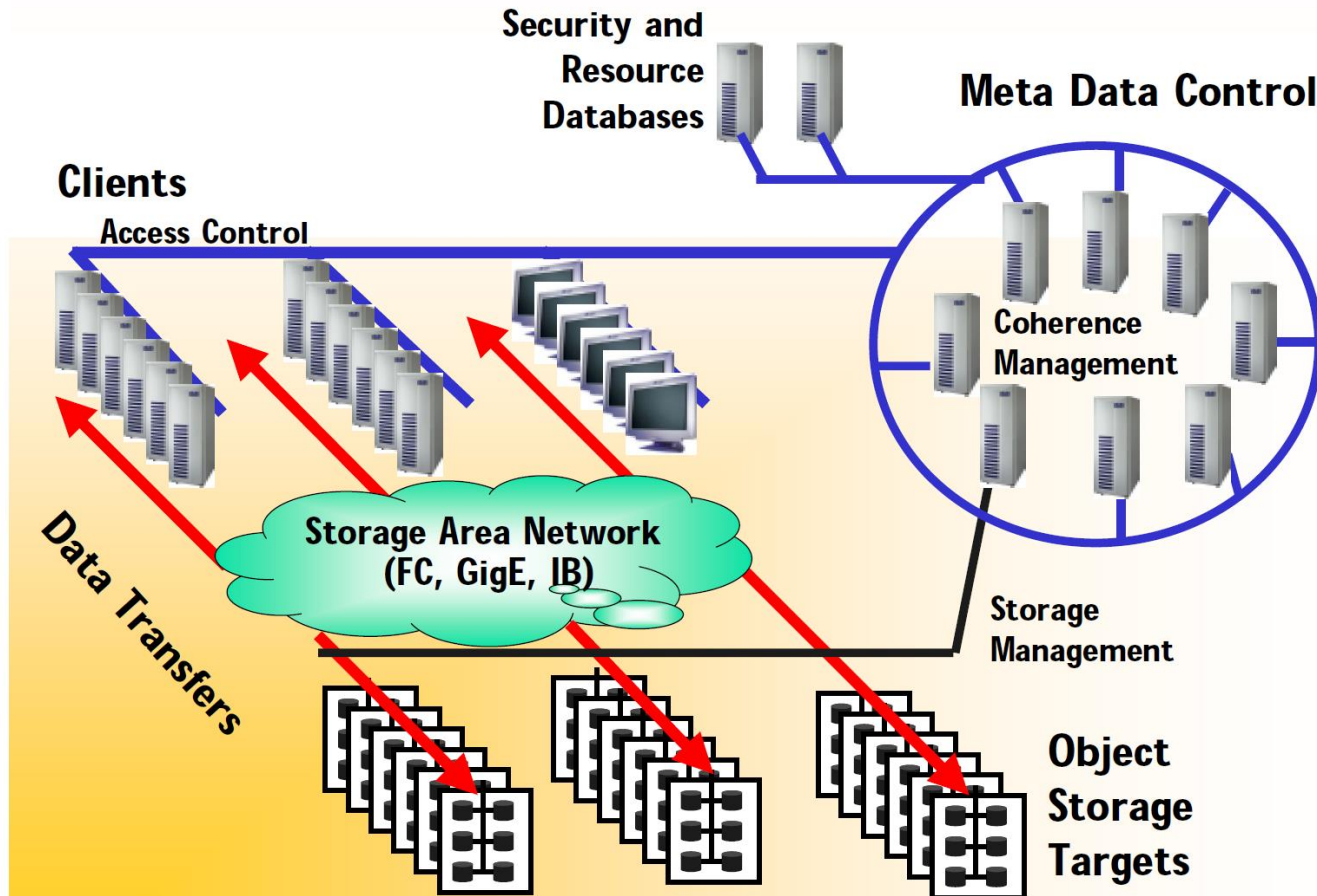
- 1999 CMU – Seagate – Stelias Computing
- 2000 Los Alamos, Sandia, Livermore:
 - need new File System
- 2001: Lustre design to meet the SGS-FS requirements?
- 2002: things moving faster
 - Lustre on MCR (1000 node Linux Cluster – bigger ones coming)
 - Lustre Hardware (BlueArc, others coming)
 - Very substantial ASCI pathforward contract (with HP & Intel)

Approach

- Initially Linux focused
- Was given blank sheet
- Learn from successes
 - GPFS on ASCI White
 - TUX web server, DAFS protocol
 - Sandia Portals Networking
 - Use existing disk file systems: ext3, XFS, JFS
- New protocols
 - InterMezzo, Coda

Cluster File Systems, Inc 

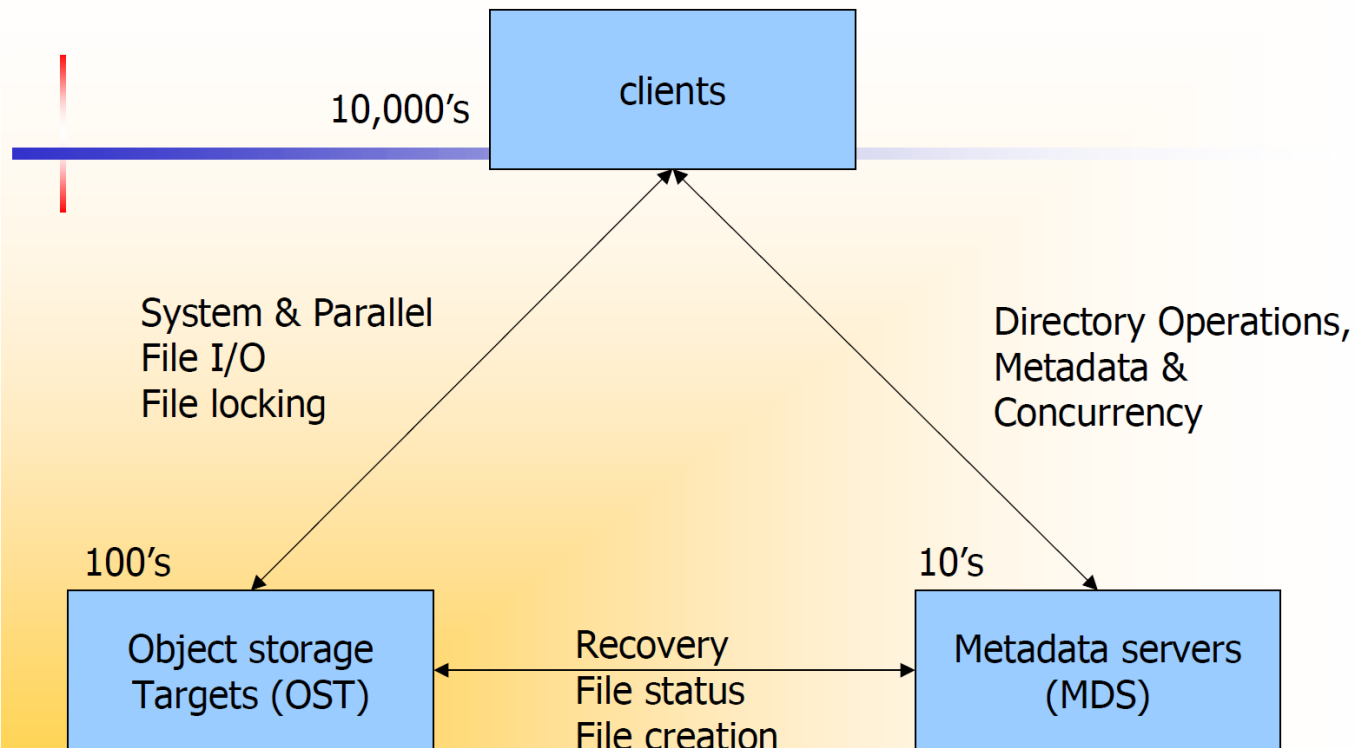
From "BlueGene/L Applications, Algorithms and Architectures" Workshop: Lustre The Inter-Galactic File System



6 6/6/2002

Cluster File Systems, Inc 

From "BlueGene/L Applications, Algorithms and Architectures" Workshop: Lustre The Inter-Galactic File System



Lustre System

7 6/6/2002

Cluster File Systems, Inc 

I/O bandwidth requirements

- Required: 100's GB/sec
- Consequences:
 - Saturate 100's – 1000's of storage controllers
 - Block allocation must be spread over cluster
 - Lock management must be spread over cluster
- This almost forces object storage controller approach

Fugaku Update and FEFS (Lustre based File System) Issues and Directions

Update on Fugaku 富岳 Development

Yutaka Ishikawa
Leader, Flagship2020 Project
RIKEN Center for Computational Science
ISC2019, 2019/06/18, 13:45-14:07



Update on Fugaku Development(1)

Fugaku



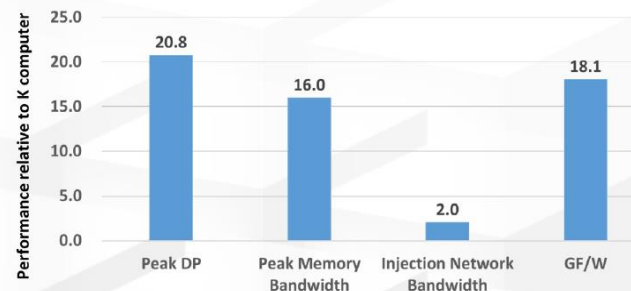
- ❑ A Fugaku prototype machine was built in Summer 2018. Since then, Fujitsu has been testing and evaluating the machine.
- ❑ Ten racks of Fugaku achieve almost the same performance of K computer (864 racks)



X 10 =



		Fugaku	K
CPU Architecture		A64FX (Armv8.2-A SVE +Fujitsu Extension)	SPARC64 VIIIfx
Node	Cores	48	8
	Peak DP performance	2.7+ TF	0.128 TF
	Main Memory	32 GiB	16 GiB
	Peak Memory Bandwidth	1024 GB/s	64 GB/s
	Peak Network Performance	40.8 GB/s	20 GB/s
Rack	Nodes	384	102
	Peak DP performance	1+ PF	< 0.013PF
Process Technology		7 nm FinFET	45 nm



An Overview of Fugaku Hardware



- **150k+ node**

2.7 TF x 150k+ = 405+ PF

- **Two types of nodes**

- Compute Node and Compute & I/O Node connected by Fujitsu TofuD, 6D mesh/torus Interconnect

- **3-level hierarchical storage system**

- 1st Layer

- One of 16 compute nodes, called Compute & Storage I/O Node, has SSD about 1.6 TB

- Services

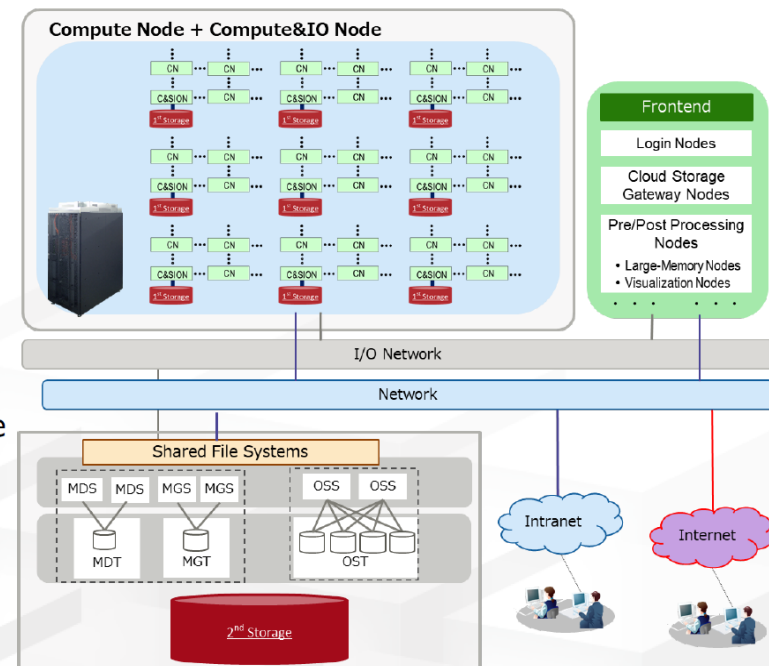
- ~ Cache for global file system
- ~ Temporary file systems
 - Local file system for compute node
 - Shared file system for a job

- 2nd Layer

- Fujitsu FEFS: Lustre-based global file system

- 3rd Layer

- Cloud storage services



An Overview of System Software Stack



✓ **Fujitsu proprietary batch job system with RIKEN power-aware scheduler**

✓ **FEFS (Fujitsu Exabyte File System)**
Lustre-based parallel file system

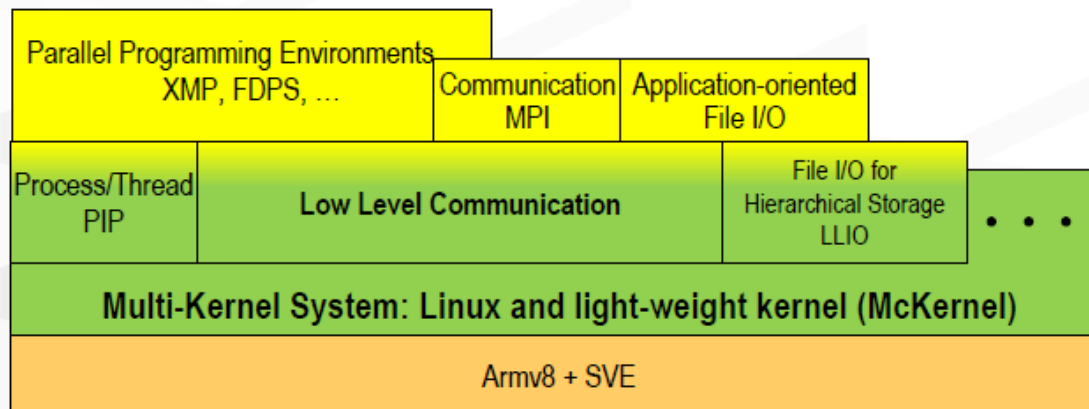
✓ **LLIO (Lightweight Layered IO-Accelerator)**
NVMe-based file IO accelerator

Batch Job and Management System

Hierarchical File System

Open Source Management Tool Spack

Red Hat Enterprise Linux 8



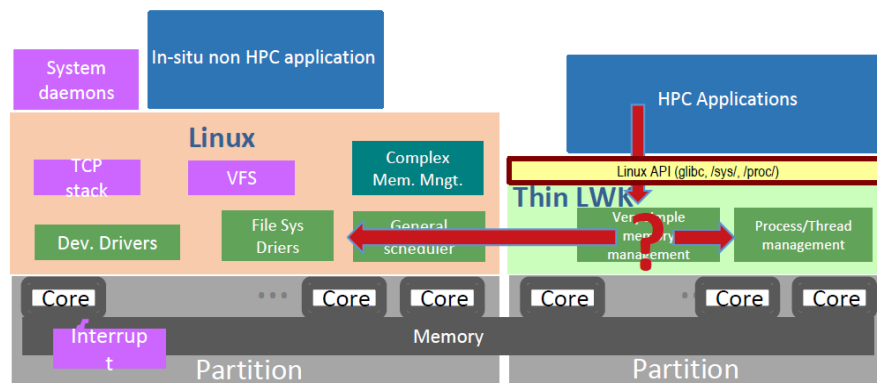
Operating System on Compute Node



- **Linux kernel on 2 or 4 cores**
 - System daemons and in-situ non HPC applications
 - Device drivers
- **Light-weight kernel(LWK), McKernel on other cores**
 - HPC applications

• McKernel

- Executes the same binary of Linux without any recompilation
- One of advantages is that McKernel provides much larger page sizes
 - Applications, accessing a huge memory area randomly, may benefit
- User may select one of McKernel configurations without rebooting



	McKernel (4K)	McKernel (64K)	Linux
.text	4K	64K	64K
.data	64K, 2M, 32M, 1G	2M, 512M	2M
.bss	64K, 2M, 32M, 1G	2M, 512M	2M
Stack	64K, 2M, 32M, 1G	2M, 512M	2M
malloc	64K, 2M, 32M, 1G	2M, 512M	2M
thread stack	64K, 2M, 32M, 1G	2M, 512M	2M
System V IPC	64K, 2M, 32M, 1G	2M, 512M	64K
POSIX shm	4K	64K	64K
XPMMEM	64K, 2M, 32M, 1G	2M, 512M	64K

1 Peta FLOPS System: K computer vs. Post-K

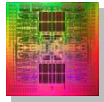
■ K computer

■ 80x compute racks & 20x disk racks

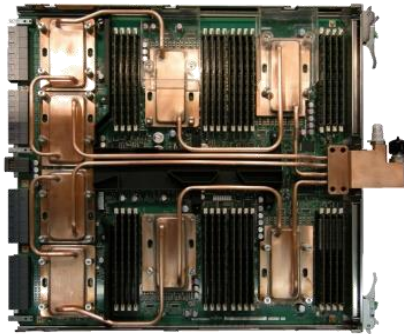
■ Post-K (Now Fugaku)

■ 1x rack w/ SSDs

SPARC64 VIIIfx

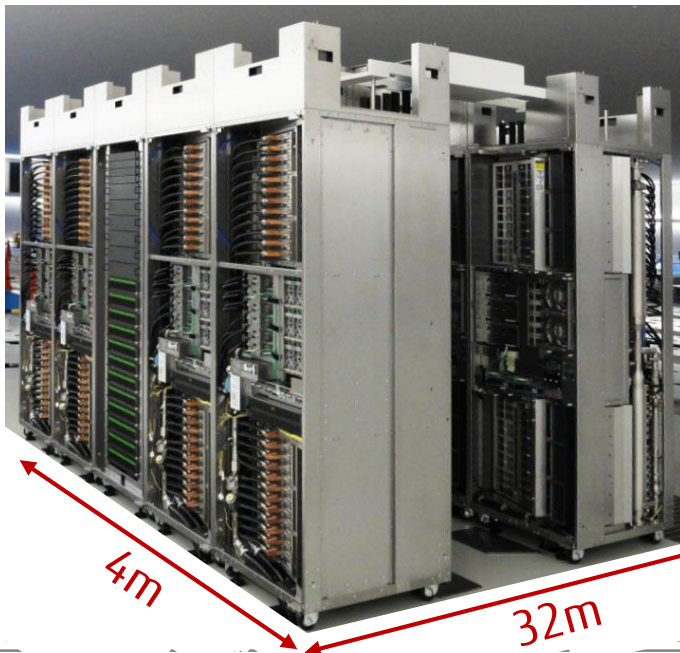


+ ICC

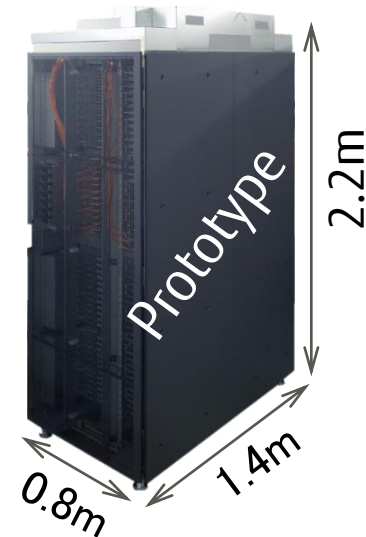


	K computer	Post-K
Compute nodes	7,680(=96x80)	384
IO nodes	4,80(=6x80)	
Footprint (m ²)	128(=4x32)	1.1
	SPARC Linux	Arm Linux

A64FX



More applications as well as system software will come in collaboration with
Open Source Community



■ ファイルアクセス性能向上

- メタデータアクセス性能向上
- 単体I/O処理高速化

■ 耐故障性向上

- フェイルオーバ時のI/O エラー防止
- ファイルサーバ異常発生からファイルI/O 再開までの時間短縮
- 両系故障時のアクセスハング抑止

■ 保守性向上

- ファイルシステム復旧時間短縮

■ 負荷耐性向上：メタアクセス、データアクセス

- MDS高負荷防止：大量ファイルアクセス時のメタアクセスレスポンス低下防止
- インタコネクト高負荷時のノード異常の誤検出回避

■ アクセス公平性確保

- 特定プロセスのファイルアクセスによりほかのプロセスが大きく影響を受ける

■ 省メモリ性向上



Computer simulations create the future

Operation of the K computer and the facility

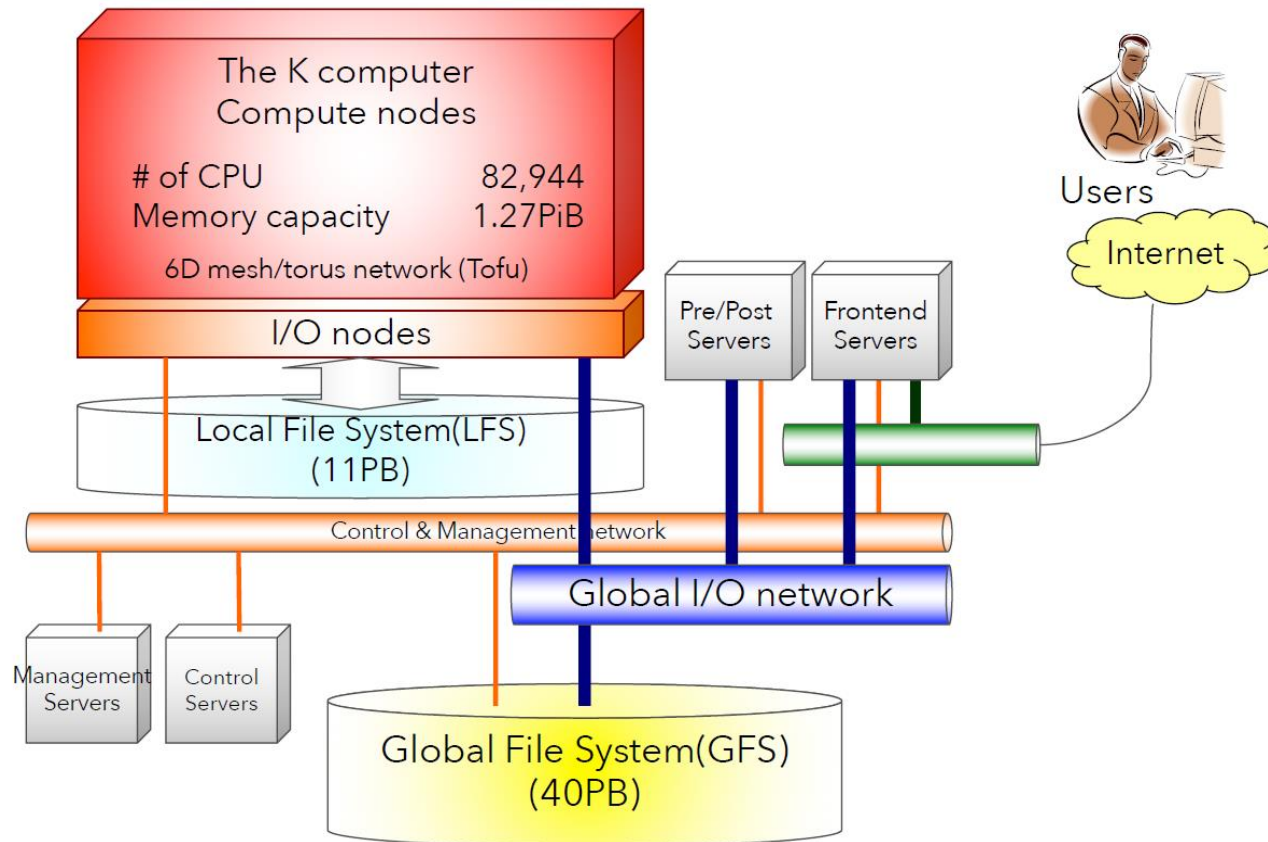
Fumiyoshi Shoji (Division Director)
Operations and Computer Technologies Div.
RIKEN Center for Computational Science



RIKEN Center for Computational Science

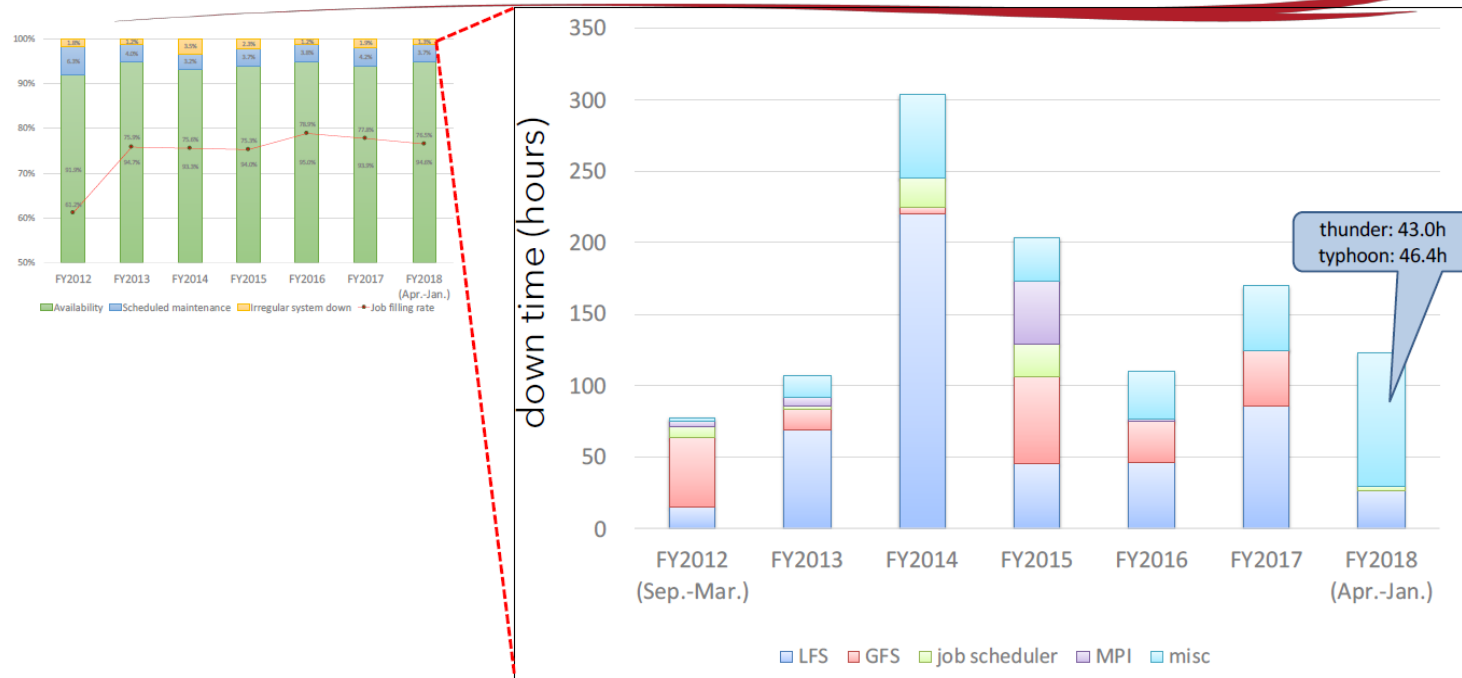
Slide from "Operation of the K computer and the facility" by Dr. Shoji

System overview



Slide from "Operation of the K computer and the facility" by Dr. Shoji

Irregular system down



- File system failures (GFS & LFS) are dominant irregular system down
- We changed our mind to give priority to resuming service earlier than investigating the cause of failures since FY2015.
- Misc. in FY2018 includes failure of power supply facility due to terrible rain and wind by typhoon (8/20) and power outage by thunder (6/8).

9

■ 階層ファイルシステムの導入

- SSD採用の第1階層LLIO、第2階層のLustreベースファイルシステム
- 16ノード毎にLLIOサーバSIOを配置、第2階層はSIO毎にマウント

■ SSD採用の第1階層LLIO導入効果

- アプリケーションファイルアクセスの高速化
 - ノードローカル&ジョブ内共有&第2階層キャッシュアクセス(SSD base)
- 第2階層のLustreベースファイルシステムへの負荷削減と制御性向上
 - クライアント数、京：10万規模→Fugaku：1万規模
 - 第2階層サーバのメモリ使用量の削減、ネットワーク負荷の削減
- SSDはジョブ実行中の一時データのみ格納： SSDは非冗長構成
 - 高速性確保と機器コスト削減
 - SSD, SIO故障：
 - 関連ジョブ異常終了で処置： 故障部品交換でリカバリ処理不要
 - 影響は小数Jobに限定： 負荷分散とシステム全体の故障耐性向上

- Lustre 2.x ベースで開発・試験開始： 状況に応じてUpdate
 - コミュニティコードベース開発が基本： 将来のUpdate障壁緩和
- 耐故障性向上
 - サーバ異常発生からファイルI/O 再開までの時間短縮： Lnet MultiRail改善
- 保守性向上
 - ファイルシステム復旧時間短縮： ストレージのデバイスリカバリ高速化等
- 負荷耐性向上
 - MDS高負荷防止： DNE, DNE2採用によるメタデータ処理分散
- アクセス公平性確保
 - QoS機能の導入

- 2019/8 京運用停止、2020年Fugakuインストール本格化
 - 徐々に規模を拡大してシステム試験を実施
- 性能安定性の鍵は、ファイルシステムの安定性に大きく依存
 - 特に第2階層のLustreベースファイルシステムの安定性が重要
- Fugakuのファイルシステムの安定化：
Lustre開発コミュニティと密に連携
- 引き続き、Bug fix、機能フィードバックに貢献していきます

LAD19講演より抜粋

LNET Multi-rail Improvement

■ Tatsushi Takamura and Shinji Sumimoto, Ph.D.

Next Generation Technical Computing Unit

FUJITSU LIMITED Sept. 24th, 2019



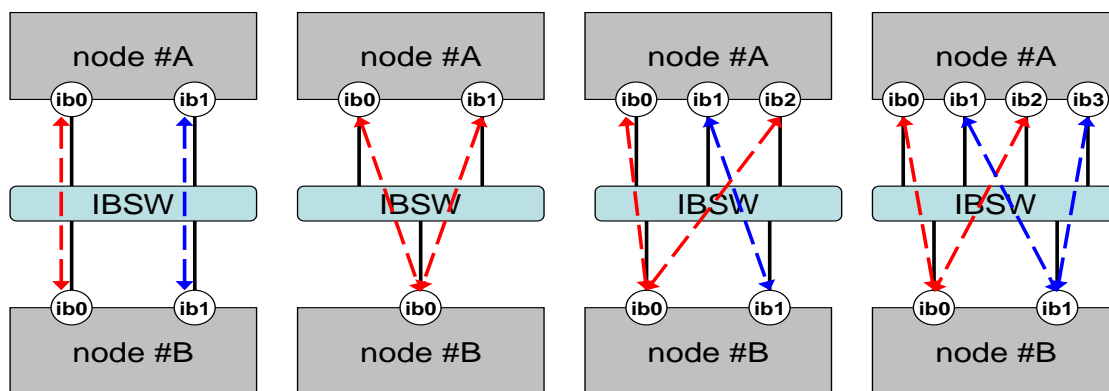
Backgrounds (FEFS IB Multi-rail)

- Fujitsu developed FEFS IB Multi-rail and operated on K computer and other HPC systems for over 7 years



- IB Multi-rail Features:

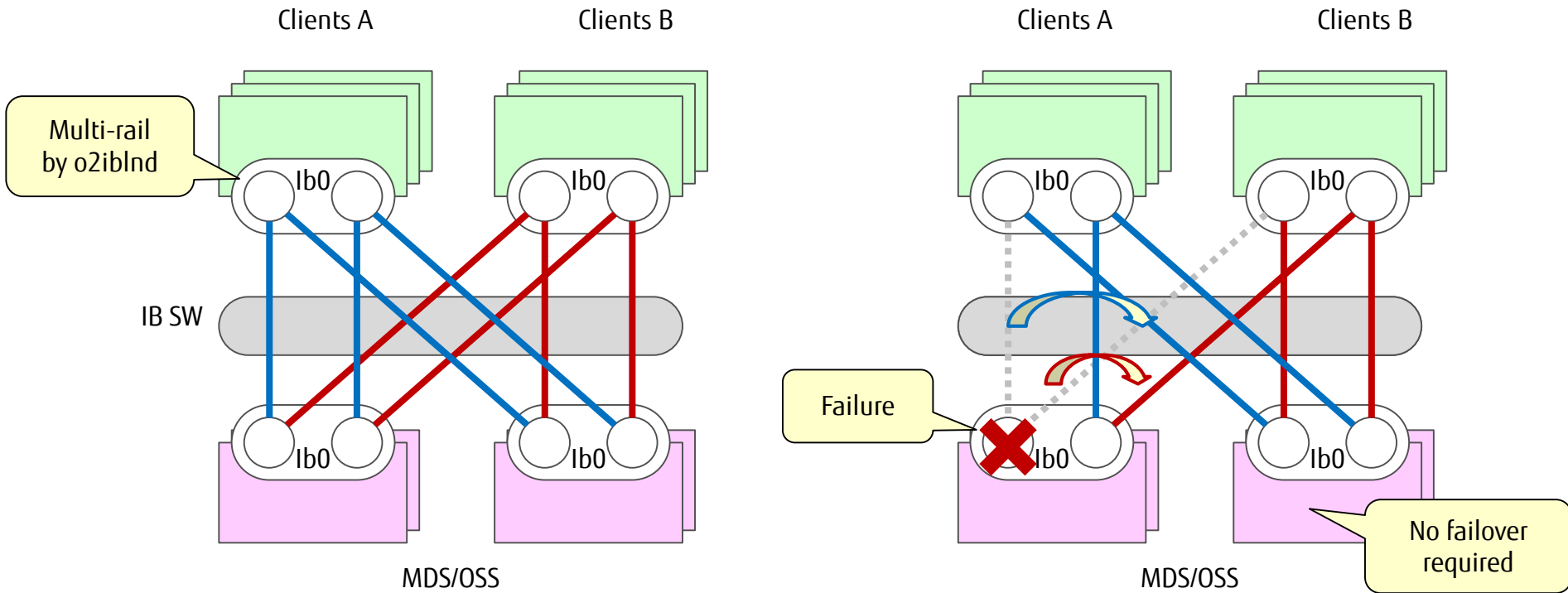
- High availability even if a single point of IB failure occurs
- High throughput by using multiple IB interfaces
- Various configurations
 - Not only Symmetric connections but also Asymmetric connections



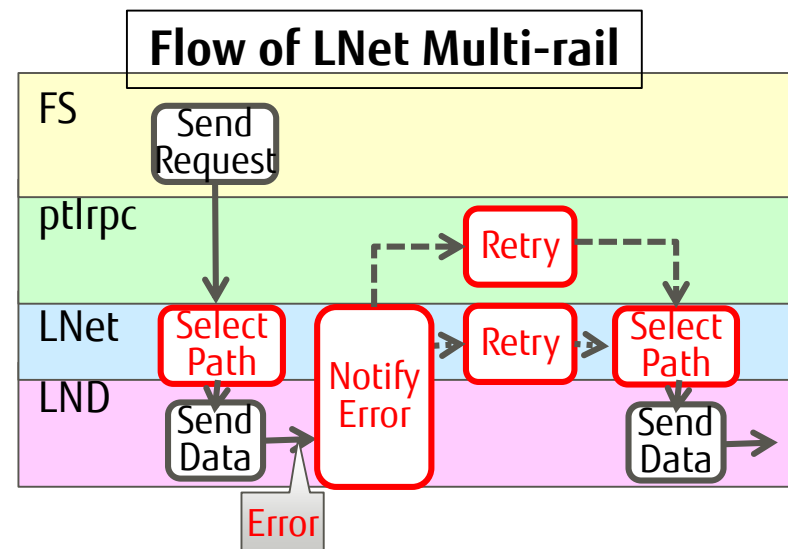
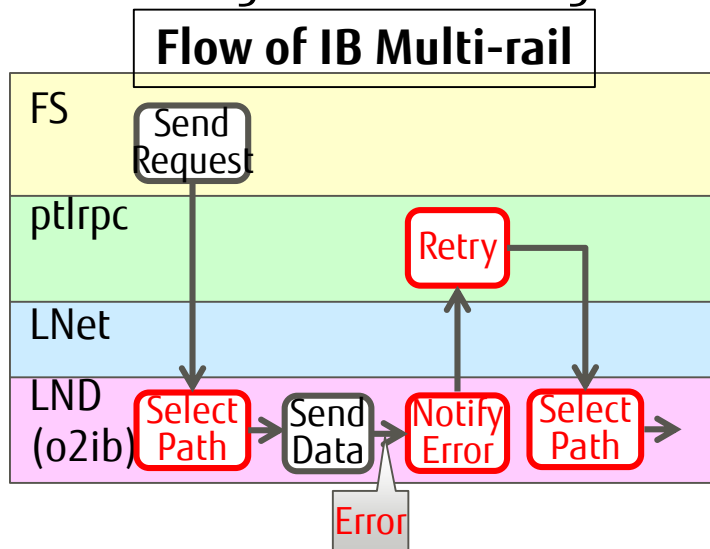
- Lustre community is now developing similar Multi-rail features on LNet level
 - LNet Multi-rail, LNet Network Health, Etc...
- However the development is still going on...
- Therefore, we have evaluated the Lustre LNet Multi-rail assuming the same features of FEFS IB Multi-rail
 - In order to give feedback to current LNet Multi-rail implementation

FEFS IB Multi-rail (presented at LAD14)

- FEFs Approach: Add IB Multi-rail function into Lustre network driver (o2ibln).
- All IB I/F on the client can be used to communicate with a server.
- All IB connections are used by round-robin order.
- Continue communication when single point of IB failure occurs.
- All IB connections are used by round-robin order by each requests.



- LNet Multi-rail: Introduced in Lustre 2.10(LU-7734)
 - Using multiple interfaces including Ethernet and InfiniBand
- LNet Network Health: Introduced in Lustre 2.12(LU-9120)
 - Detecting network failures of local interface, remote interface, network timeouts and etc.
 - Switching and resending among different interfaces



The basic idea is the same as FEFS IB Multi-rail
(The difference is LND level or LNet level)

■ Detecting device status

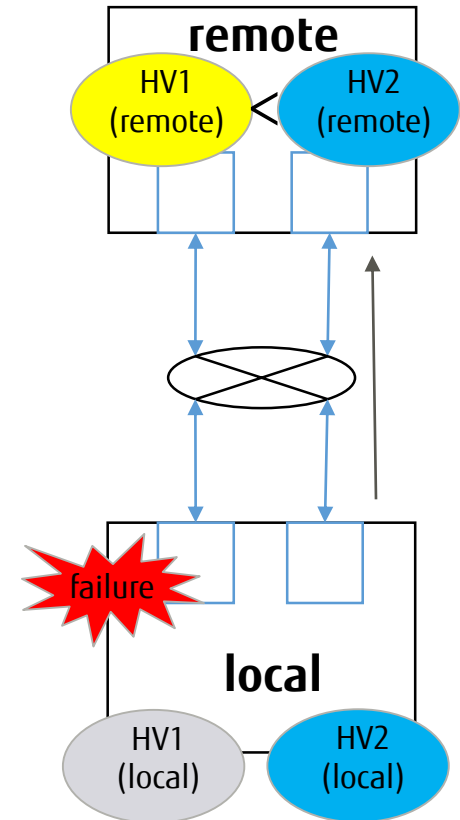
- A local Network Interface (NI) is marked fatal, if the device has gone into a fatal
 - ex. IB_EVENT_DEVICE_FATAL, IB_EVENT_PORT_ERR

■ Maintaining health value

- Each NI (both local, remote) has a health value (HV)
- HV is decremented when communications fail and incremented when succeeds

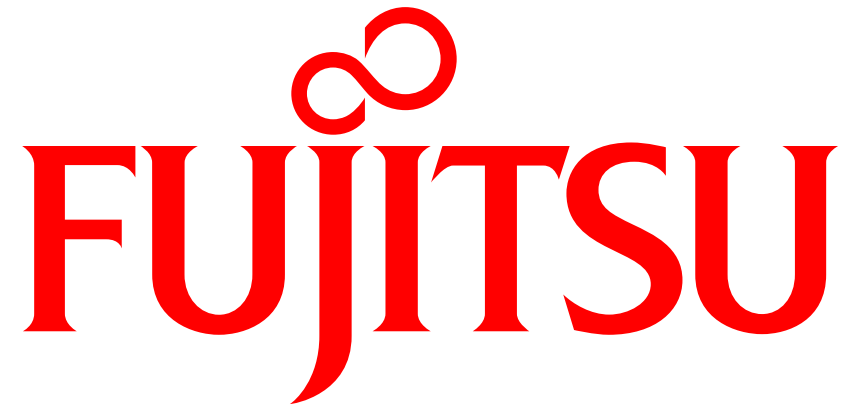
■ Controlling path selection

- Selecting the healthiest local NI by HV
 - Fatal NI is removed from the candidates
- Selecting the healthiest remote NI which belong to the same network which the local NI connected
- Communicating using the selected NIs



Summary of Issues and Modifications

Issue	Description	Modification
No.1	Unable to detect IB hardware failure (NI is not marked fatal).	We handled IB hardware failure and a path is selected without waiting for a HV decremented
No.2	Decrementing a health (HV) of normal NI	We set HV appropriately and reduced extra resending
No.3	Unable to use Multi-rail on asymmetric NIs	We switched switch another normal NI to avoid for the system to become unusable
No.4	After recovery of NI failure, the NI is not used for a while (1000sec)	We stopped HV decrement at recovery processing to use the NI in a few seconds after device recovery



shaping tomorrow with you