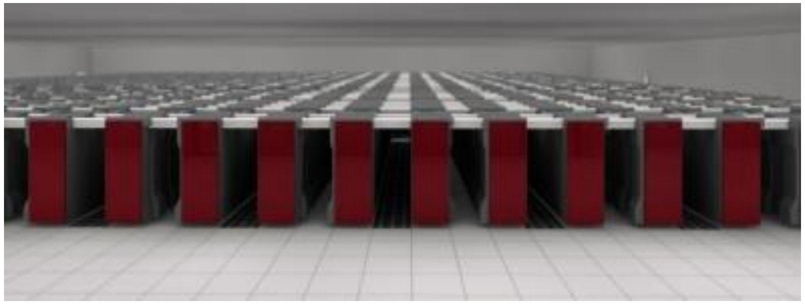
FUJITSU

shaping tomorrow with you

FEFS: Scalable Cluster File System



K Computer (RIKEN AICS)

"K computer" is the nickname RIKEN has been using for the supercomputer.

PRIMEHPC FX10

PRIMERGY

Copyright 2011 FUJITSU LIMITED

Outline

- Overview of FEFS
  - Major features
  - Target system
- I/O Architecture
  - Concept and system design
  - High Reliability
- Technical Issues
  - Lustre Extensions
  - I/O zoning
  - Fair-share QoS, Best-effort QoS
- Performance Evaluation
  - Throughput (IOR)
  - Response (mdtest)
- Summary
  - Contribution to the Lustre community.

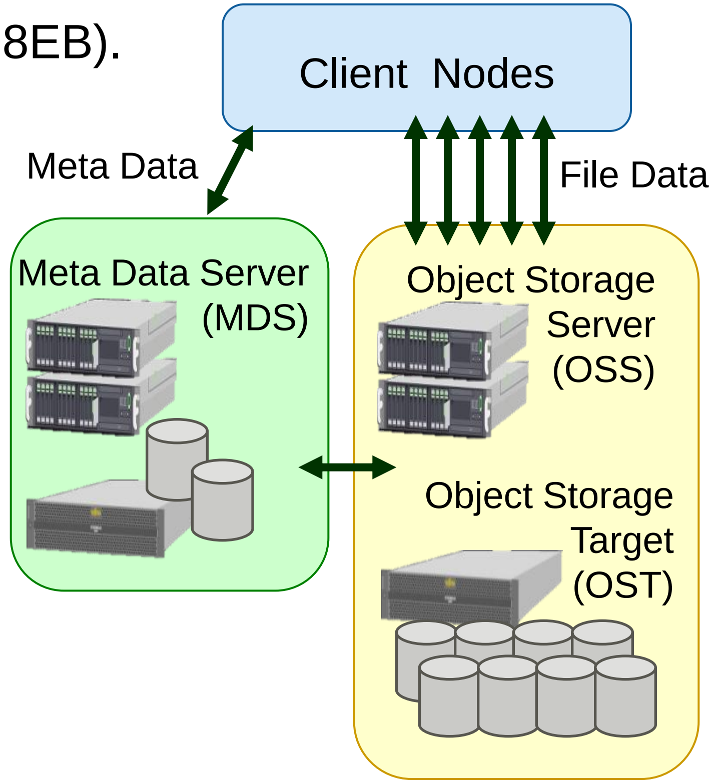
Copyright 2011 FUJITSU LIMITED

Features of FEFS

FEFS is a scalable cluster file system based on Lustre.

(FEFS: Fujitsu Exabyte File System)

- High Performance & High Scalability
  - Scalable I/O performance (~1TB/s) & capacity (~8EB).
- I/O Usage Management
  - Fair-share QoS
  - Best-effort QoS
- High Reliability & High Availability
  - Failover with redundant hardware and continuing file system service.

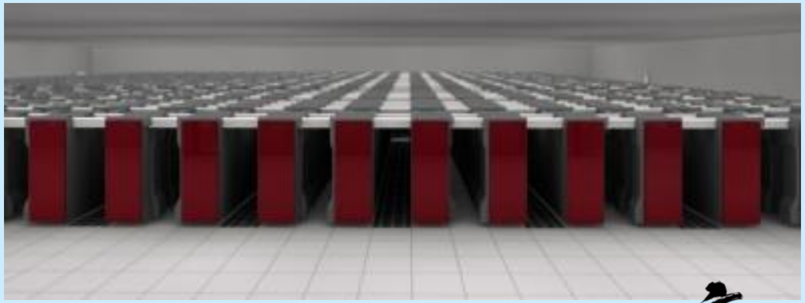


Copyright 2011 FUJITSU LIMITED

Target System

- K Computer
  - RIKEN and Fujitsu have been working together to develop the K computer.
    - To be installed at the RIKEN AICS, Kobe, by 2012
- PRIMEHPC FX10
  - Fujitsu’s brand-new supercomputer recently release.
- PC Cluster
  - PRIMERGY and third-party IA/Linux based servers.

Super Computer



K computer (RIKEN AICS)

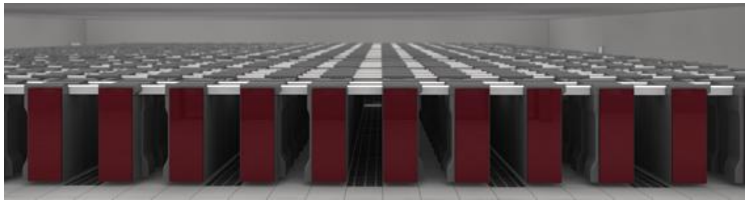
PRIMEHPC FX10

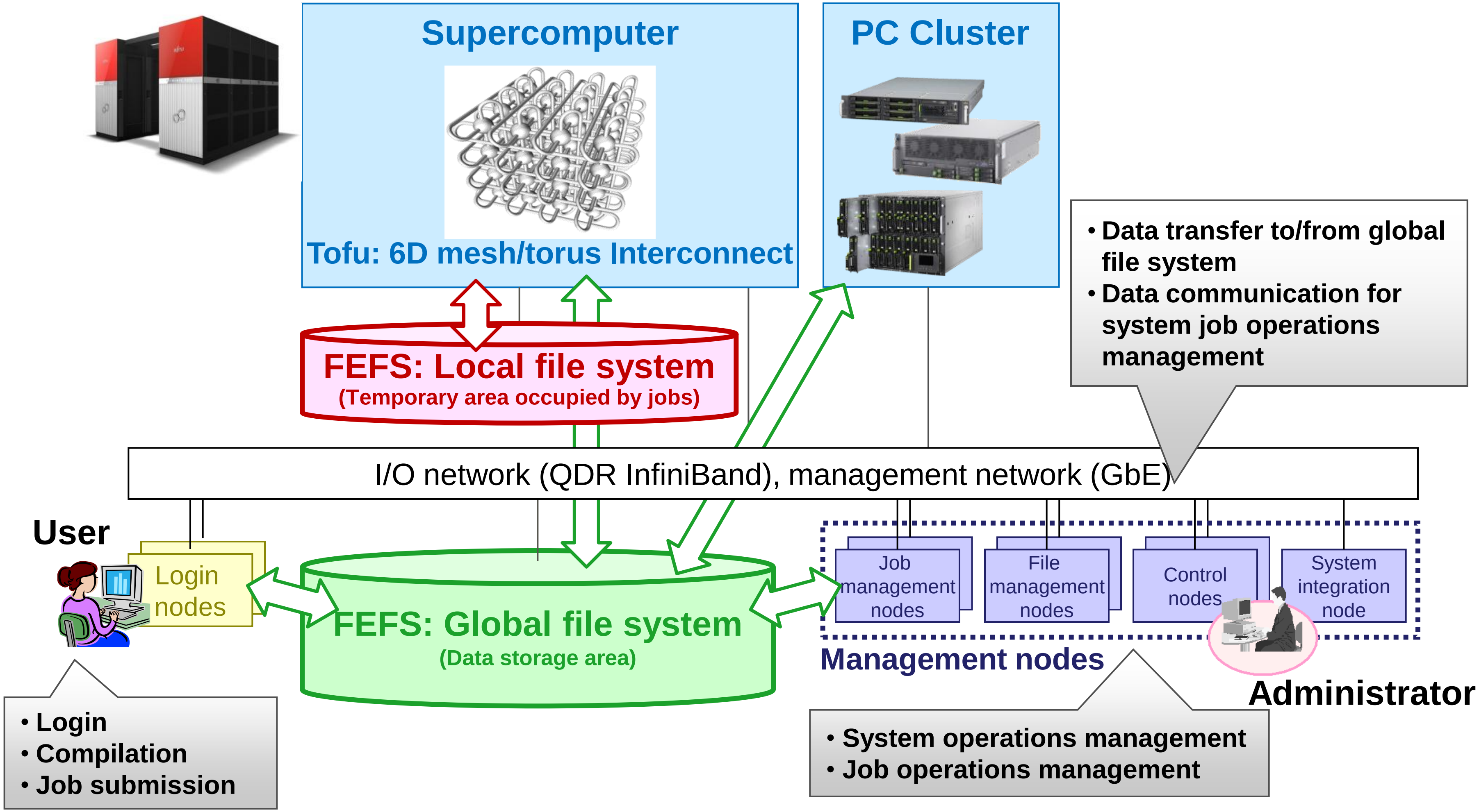
PC Cluster

PRIMERGY

Copyright 2011 FUJITSU LIMITED

System Configuration





Supercomputer

Tofu: 6D mesh/torus Interconnect

PC Cluster

- Data transfer to/from global file system
- Data communication for system job operations management

FEFS: Local file system

(Temporary area occupied by jobs)

I/O network (QDR InfiniBand), management network (GbE)

User

Login nodes

FEFS: Global file system

(Data storage area)

Job management nodes

File management nodes

Control nodes

System integration node

Management nodes

Administrator

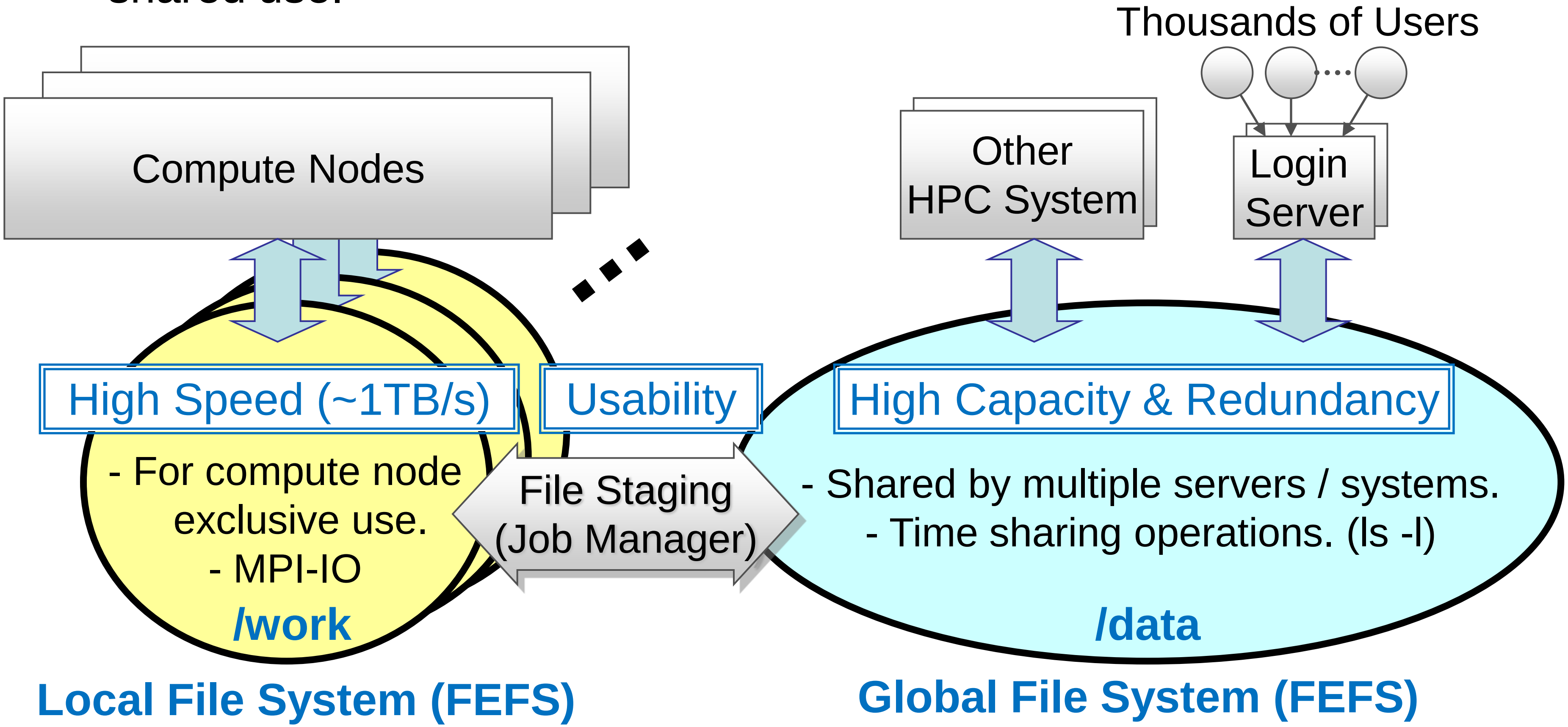
- Login
- Compilation
- Job submission

- System operations management
- Job operations management

Copyright 2011 FUJITSU LIMITED

I/O Architecture: Basic Concept

- Incompatible features is implemented by introducing Layered File System.
  - Local File System (/work): High Speed FS for dedicated use for jobs.
  - Global File System (/data): Large Capacity and Redundancy FS for shared use.


Thousands of Users

Compute Nodes

Other HPC System

Login Server

High Speed (~1TB/s)

Usability

High Capacity & Redundancy

- For compute node exclusive use.
- MPI-IO

/work

File Staging (Job Manager)

- Shared by multiple servers / systems.
- Time sharing operations. (ls -l)

/data

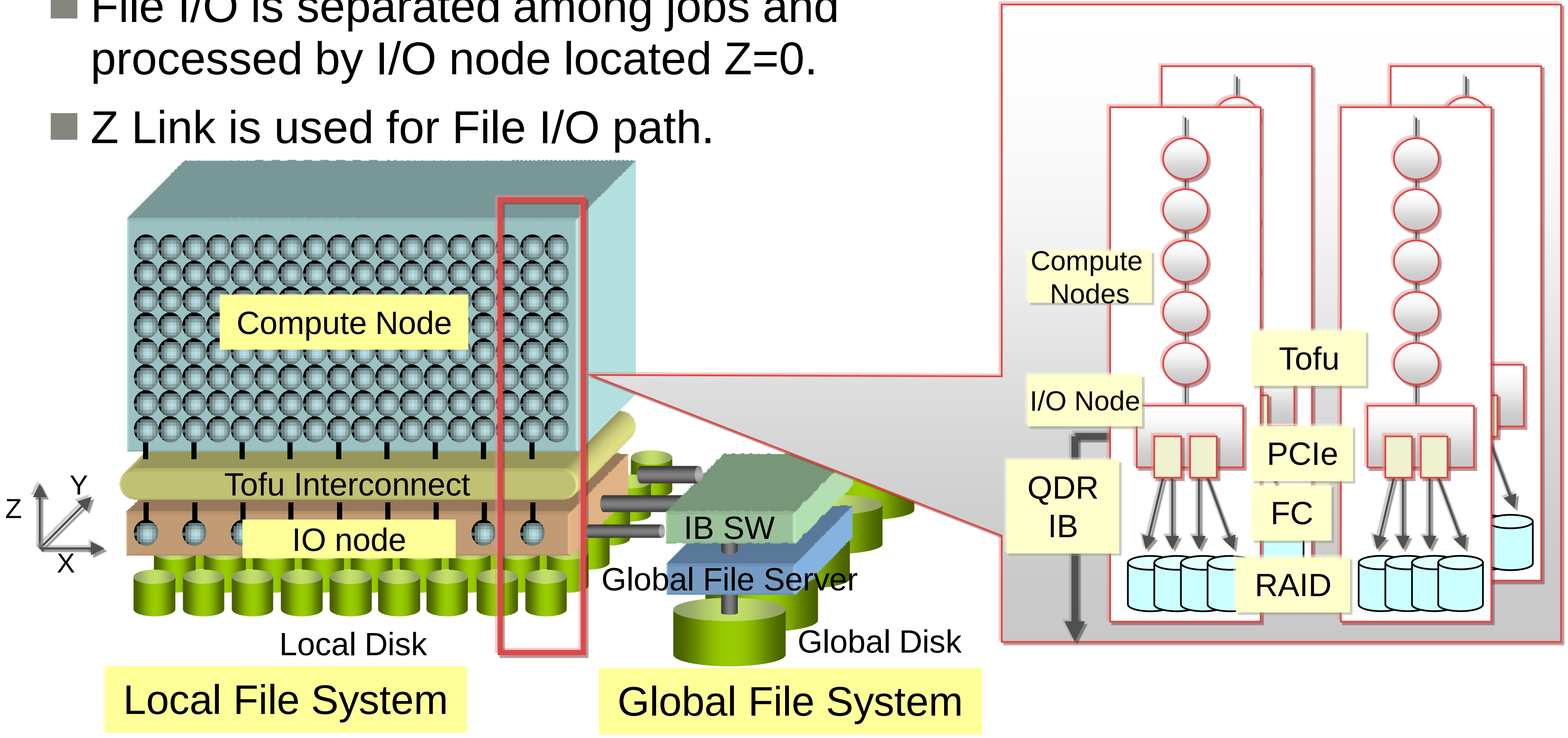
Local File System (FEFS)

Global File System (FEFS)

Copyright 2011 FUJITSU LIMITED

I/O Architecture: System Design

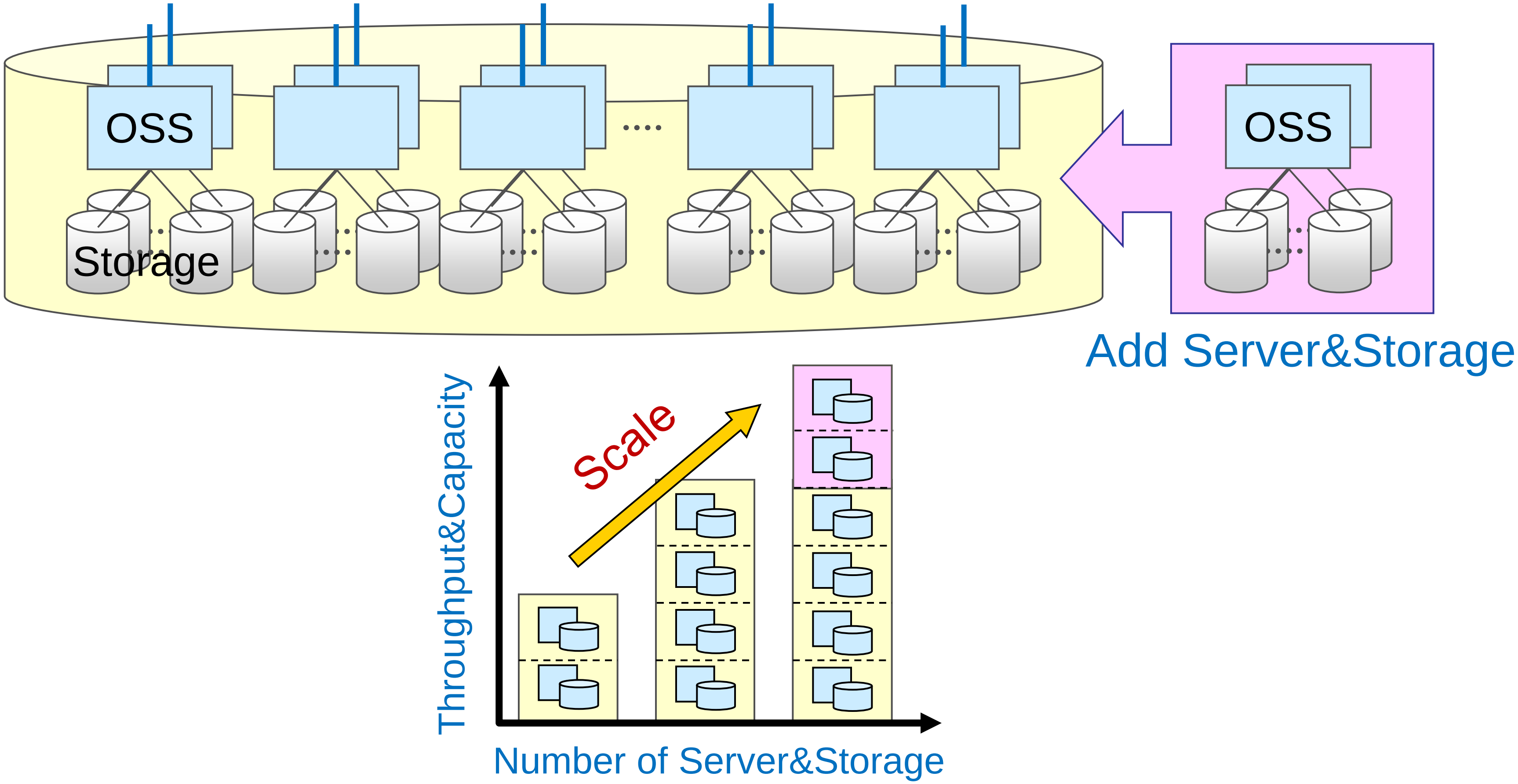
- Optimized for Scalable File I/O Operation
  - Achieving Scalable Storage Volume and Performance
  - Eliminating I/O Conflicts from Every Components
- I/O Zoning Technology for Local File System
  - File I/O is separated among jobs and processed by I/O node located Z=0.
  - Z Link is used for File I/O path.



Copyright 2011 FUJITSU LIMITED

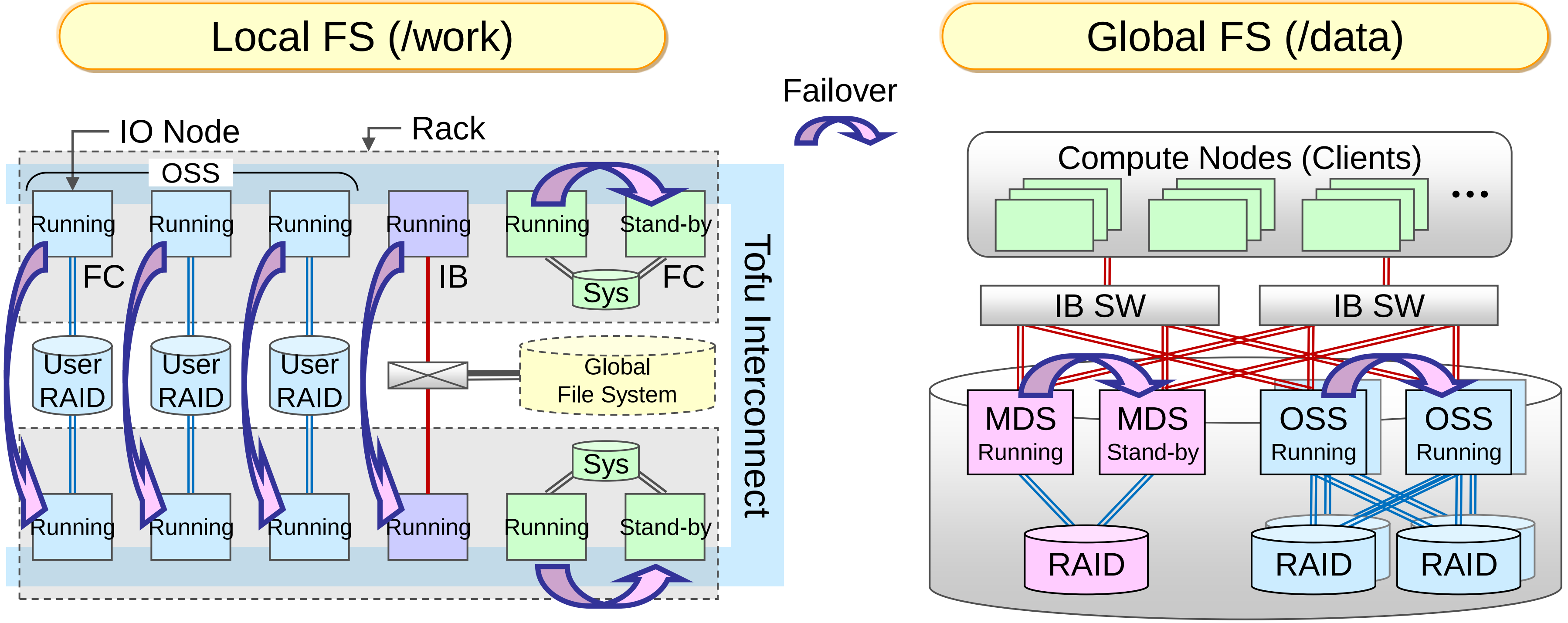
Scalable Performance & Capacity

- High speed throughput and large capacity have achieved by multiple OSSs.
  - Scale out throughput & capacity by adding servers and storages.



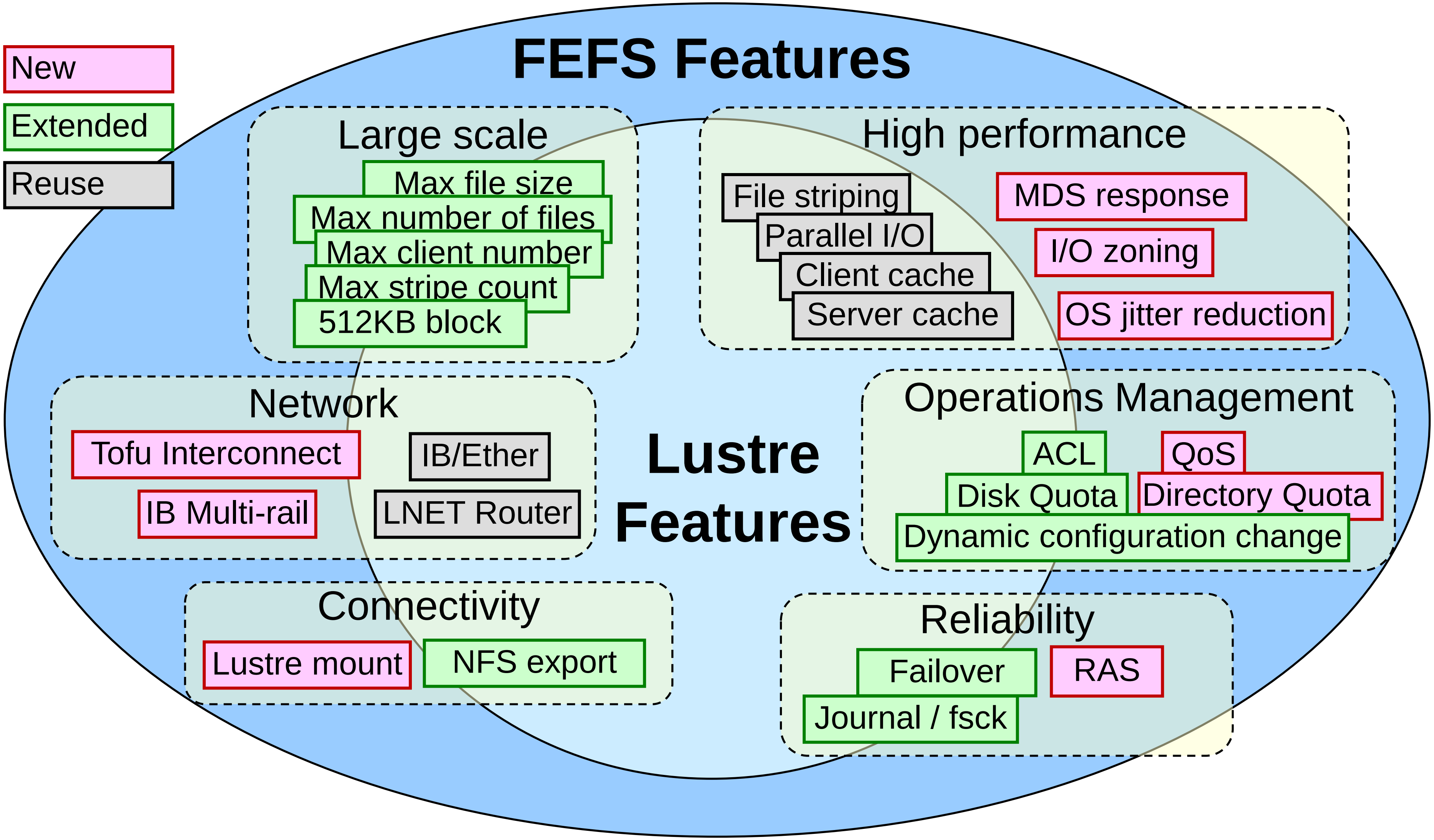
High Reliability and High Availability

- Keeping file system service against failures.
  - Redundant hardware
    - Duplex paths of InfiniBand, Fibre Channel, I/O Server
    - RAID disks (MDS：RAID10, OSS: RAID5/6)
  - System Management software
    - Detect failure and switch to alternate path or server automatically



Copyright 2011 FUJITSU LIMITED

Lustre Extension of FEFS: Features



- 🟥 New
- 🟩 Extended
- ⬛ Reuse

FEFS Features

Lustre Features

Large scale
- 🟩 Max file size
- 🟩 Max number of files
- 🟩 Max client number
- 🟩 Max stripe count
- 🟩 512KB block

High performance
- ⬛ File striping
- ⬛ Parallel I/O
- ⬛ Client cache
- ⬛ Server cache
- 🟥 MDS response
- 🟥 I/O zoning
- 🟥 OS jitter reduction

Network
- 🟥 Tofu Interconnect
- ⬛ IB/Ether
- 🟥 IB Multi-rail
- ⬛ LNET Router

Operations Management
- 🟩 ACL
- 🟥 QoS
- 🟩 Disk Quota
- 🟥 Directory Quota
- 🟩 Dynamic configuration change

Connectivity
- 🟥 Lustre mount
- 🟩 NFS export

Reliability
- 🟩 Failover
- 🟥 RAS
- 🟩 Journal / fsck

Copyright 2011 FUJITSU LIMITED

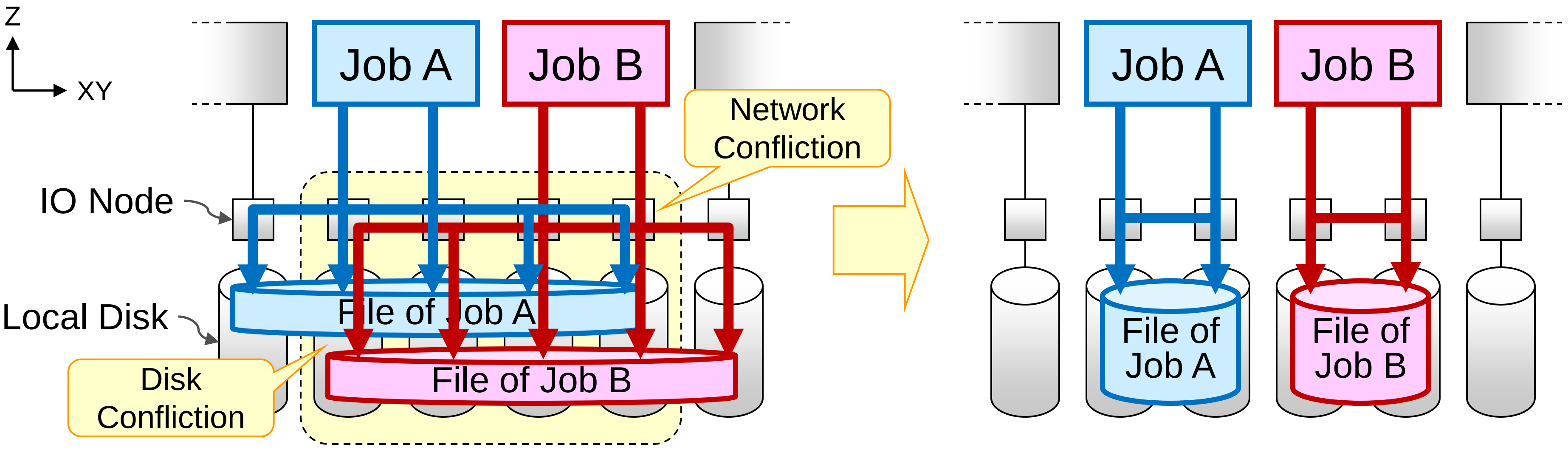
Lustre Extension of FEFS: Specification

Features		FEFS	Current Lustre
System Limits	Max file system size	100PB (8EB)	64PB
	Max file size	1PB (8EB)	320TB
	Max #files	32G (8E)	4G
	Max OST size	100TB (1PB)	16TB
	Max stripe count	20k	160
	Max ACL entries	8191	32
Node Scalability	Max #OSTs	20k	8150
	Max #clients	1M	128K
Usability	QoS	Yes	No
	Directory Quota	Yes	No
InfiniBand Multi-rail		Yes	No
Block Size (Backend File System)		~512KB	4KB

Copyright 2011 FUJITSU LIMITED

I/O Zoning: I/O Separation among Jobs

- Issue: Job’s I/O conflicts on hardware.
  - Sharing disk volumes, network links among jobs cause I/O performance degradation because of their confliction.
- Our Approach: Separate hardware among jobs.
  - Separating of disk volumes, network links among jobs as much as possible.



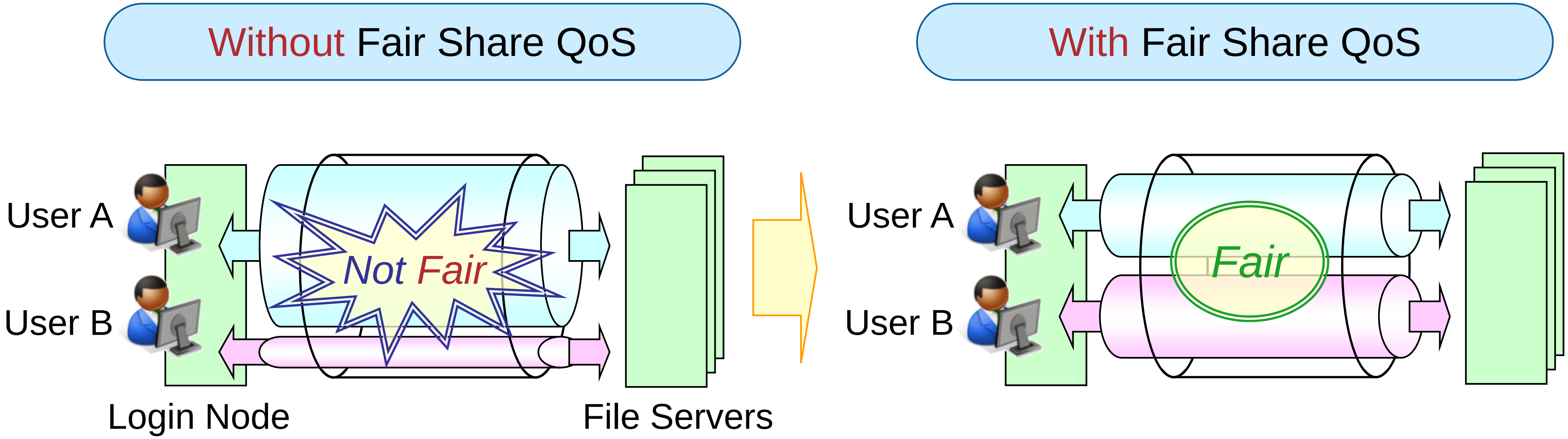
No-good: w/ I/O Confliction

Good: w/o I/O Confliction

Copyright 2011 FUJITSU LIMITED

QoS: Fair-share QoS

- Issue
  - Avoiding from some one’s occupying file I/O resources.
- Our approach
  - Limit the number of I/O requests each user can execute simultaneously on the client node.

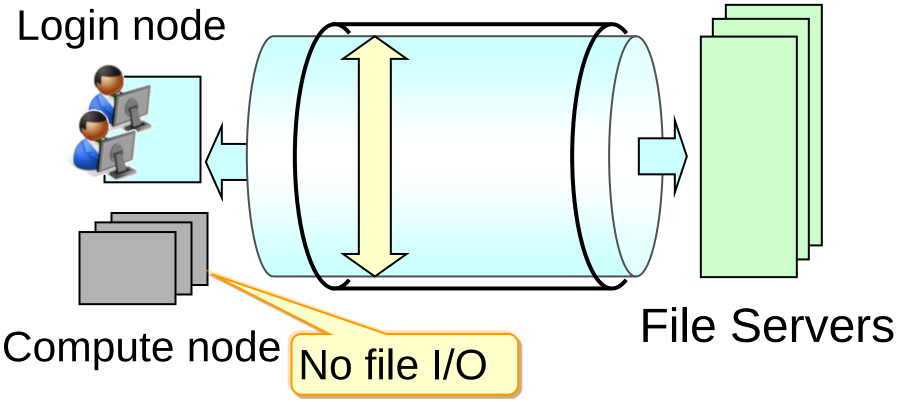


Copyright 2011 FUJITSU LIMITED

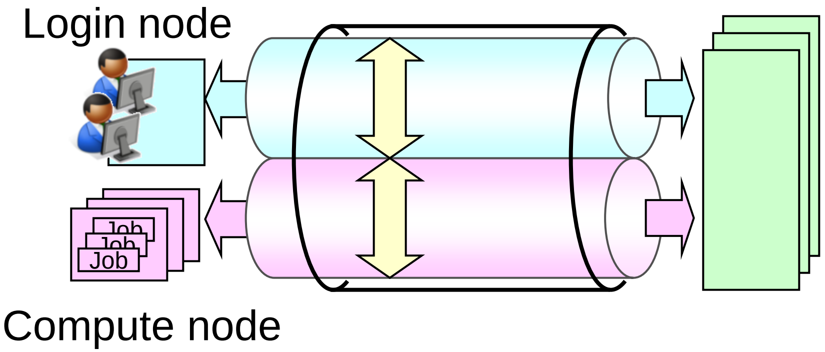
QoS: Best-effort QoS

- **Issue**
  - Utilize all I/O resources effectively.
- **Our Approach**
  - Assign all server resources to clients that execute file I/O.

Occupied by 1 node



Shared by multiple nodes



Copyright 2011 FUJITSU LIMITED

Evaluation of FEFS: QoS

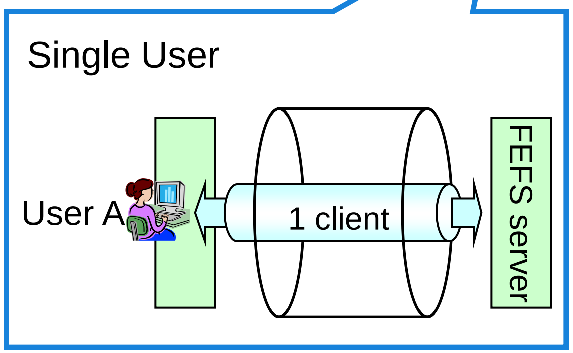
QoS efficiency on PC Cluster

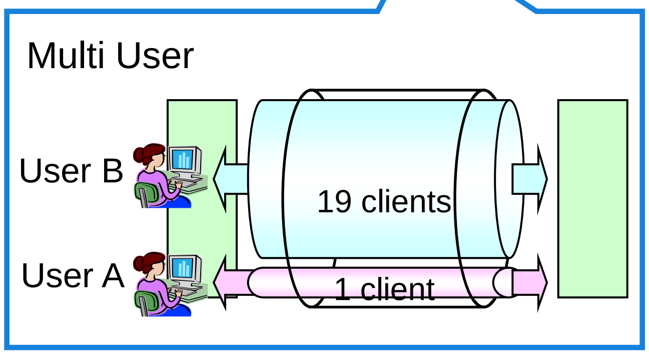
- User A: 1 node Job ⇒ Measure creation/removal time of 10,000 files.
- User B: 19 node Job

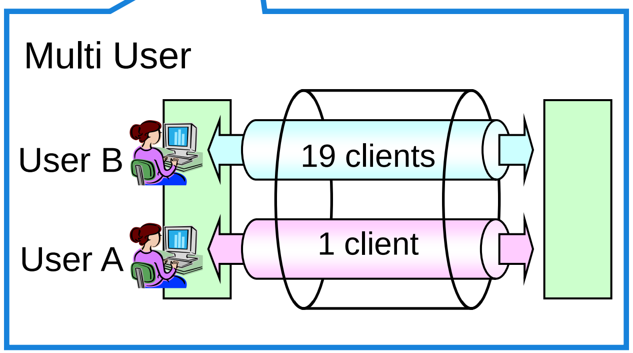
Influence of User B's file operation is suppressed.

User A's processing time

User A 10,000 Files	w/o QoS Single User	w/o QoS Multi User	w/ QoS Multi User
Create Files	4.1 sec	10.1 sec	3.9 sec
Remove Files	4.2 sec	14.0 sec	5.5 sec







Copyright 2011 FUJITSU LIMITED

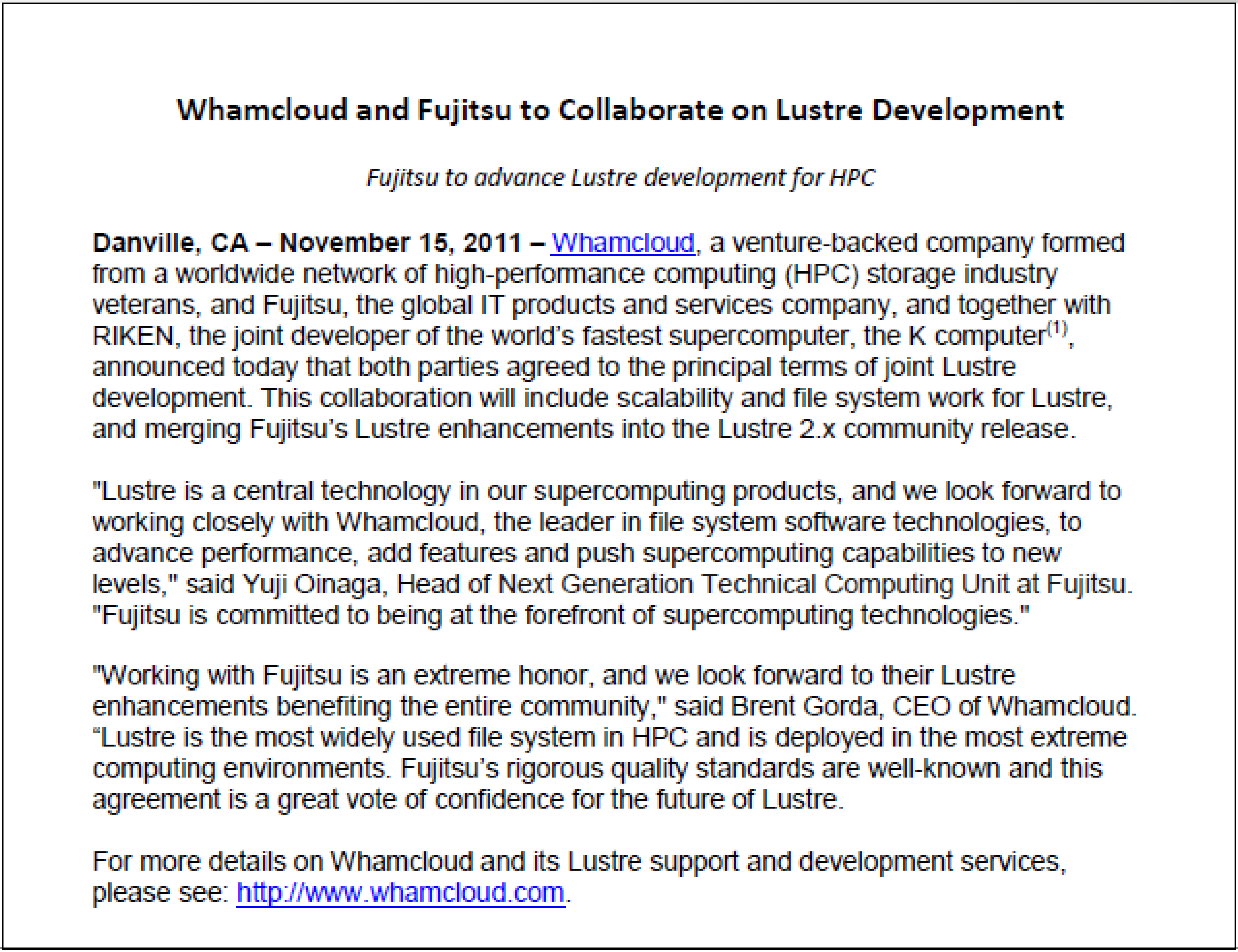
Summary and Future Works

- Fujitsu developed Lustre based cluster file system FEFS.
  - High-speed file I/O (~1TB/s), Huge capacity (~8EB)
  - High-reliability and High-availability
  - Luster enhancements: QoS, IB multi-rail, directory Quota.

- Future Works
  - Contribute our efforts to the Lustre community.
  - Merge our enhancements into future release of Lustre.

Copyright 2011 FUJITSU LIMITED

Press Release



Whamcloud and Fujitsu to Collaborate on Lustre Development

Fujitsu to advance Lustre development for HPC

Danville, CA – November 15, 2011 – Whamcloud, a venture-backed company formed from a worldwide network of high-performance computing (HPC) storage industry veterans, and Fujitsu, the global IT products and services company, and together with RIKEN, the joint developer of the world's fastest supercomputer, the K computer[1], announced today that both parties agreed to the principal terms of joint Lustre development. This collaboration will include scalability and file system work for Lustre, and merging Fujitsu's Lustre enhancements into the Lustre 2.x community release.

"Lustre is a central technology in our supercomputing products, and we look forward to working closely with Whamcloud, the leader in file system software technologies, to advance performance, add features and push supercomputing capabilities to new levels," said Yuji Oinaga, Head of Next Generation Technical Computing Unit at Fujitsu. "Fujitsu is committed to being at the forefront of supercomputing technologies."

"Working with Fujitsu is an extreme honor, and we look forward to their Lustre enhancements benefiting the entire community," said Brent Gorda, CEO of Whamcloud. "Lustre is the most widely used file system in HPC and is deployed in the most extreme computing environments. Fujitsu's rigorous quality standards are well-known and this agreement is a great vote of confidence for the future of Lustre.

For more details on Whamcloud and its Lustre support and development services, please see: http://www.whamcloud.com.

Copyright 2011 FUJITSU LIMITED

FUJITSU

shaping tomorrow with you