

Technical Computing Suite supporting the hybrid system

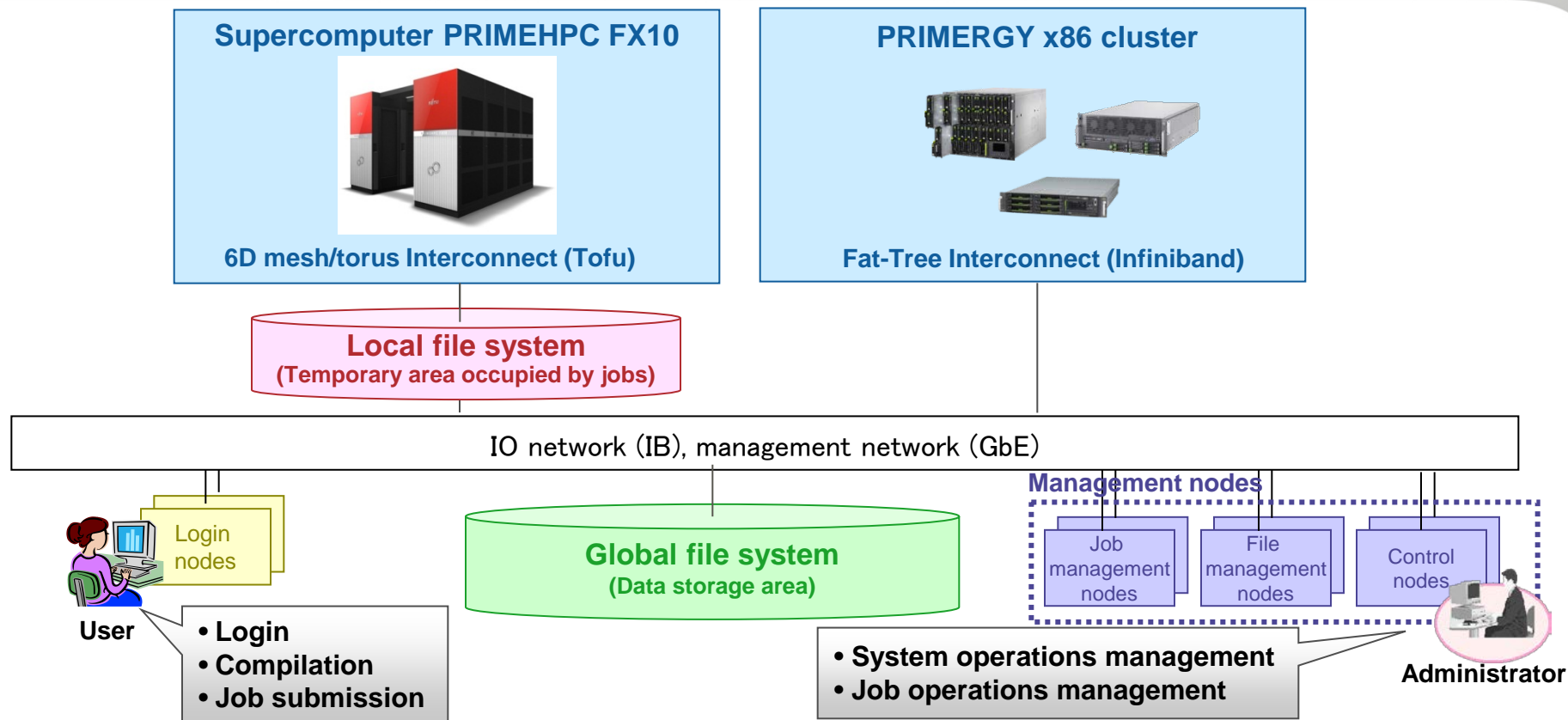


**Supercomputer
PRIMEHPC FX10**



**PRIMERGY
x86 cluster**

Hybrid System Configuration



System Software Stack



User/ISV Applications

HPC Portal / System Management Portal

Technical Computing Suite

System operations management

- System configuration management
- System control
- System monitoring
- System installation & operation

High-performance file system

- Lustre-based distributed file system
- High scalability
- IO bandwidth guarantee
- High reliability & availability

Compilers

- Hybrid parallel programming
- Sector cache support
- SIMD / Register file extensions

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

Job operations management

- Job manager
- Job scheduler
- Resource management
- Parallel execution environment

VISIMPACT™

- Shared L2 cache on a chip
- Hardware intra-processor synchronization

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Linux-based enhanced Operating System

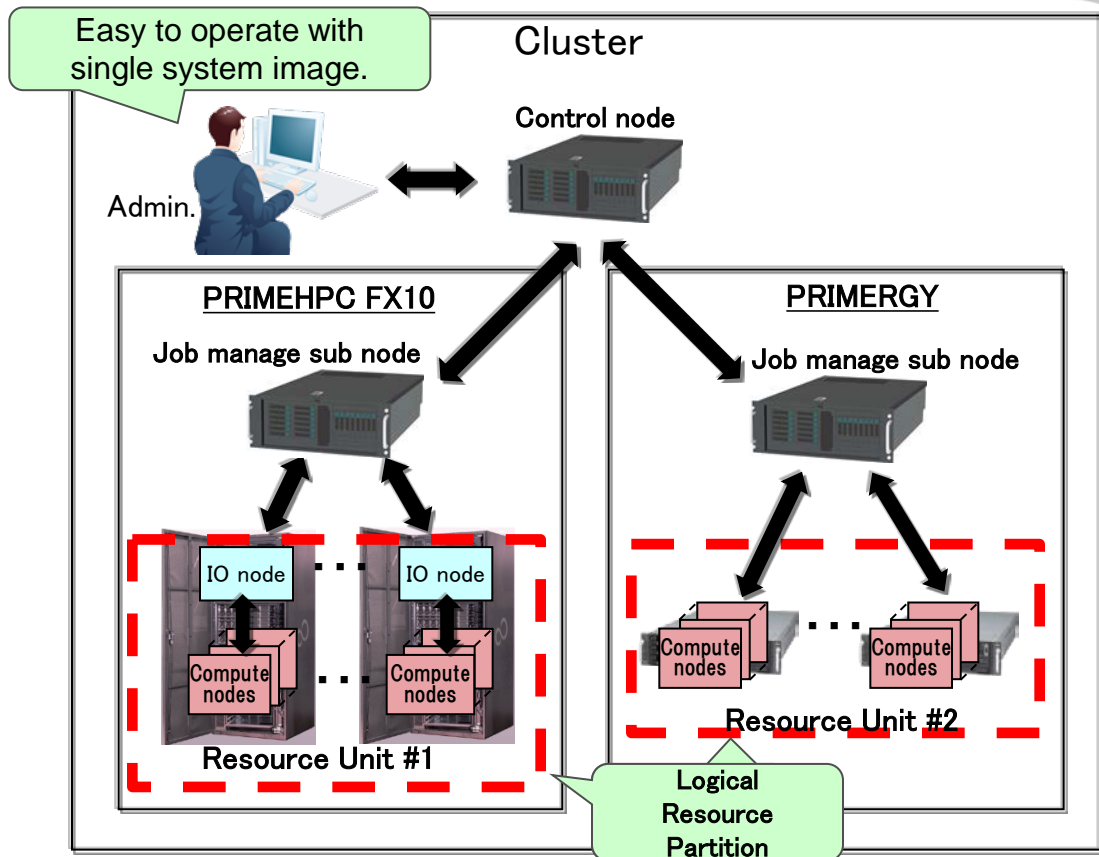
Red Hat Enterprise Linux

Supercomputer PRIMEHPC FX10

PRIMERGY x86 cluster

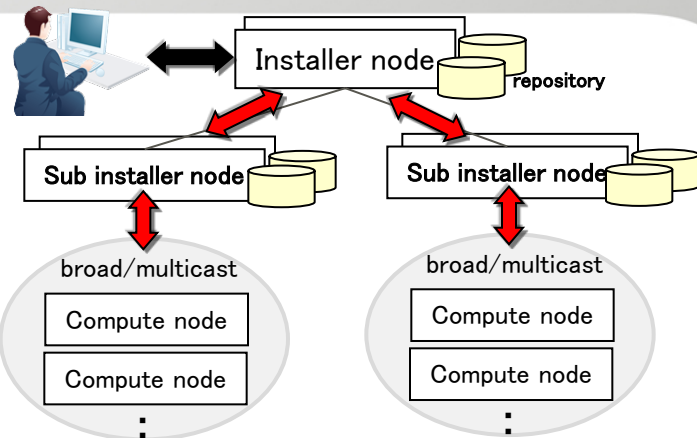
System Operations Management

- Single system image in FX10 and PRIMERGY
 - Installation / Update Packages
 - High Availability Control
 - Hardware/Software Monitoring
 - Power Control
- Hierarchical structure for large-scale systems
 - Load balance by using the job management sub node.
- Logical resource partition for efficient operations

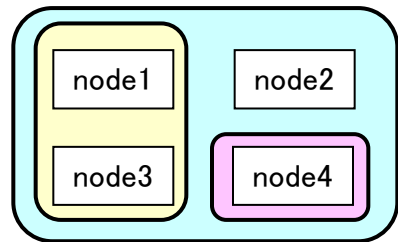


Installation / Update Packages

Installation
Update packages



* Support diskless node for FX10



2-tier (common/addition)
package management

Common package
PKG-A
PKG-B
PKG-C

Additional package-1
PKG-D
PKG-E

Additional package-2
PKG-F

■ Large-scale system support

- Hierarchical installer node structure

■ 2-tier package management

- Common packages: on all nodes.
- Additional packages: on some nodes.

High Availability Features

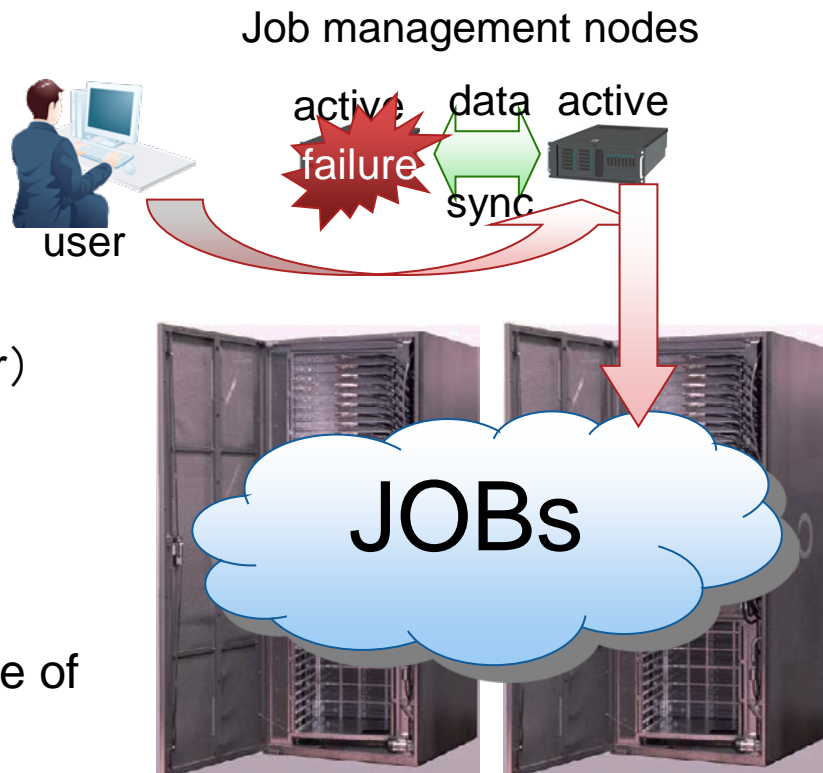
- The important nodes have **redundancy**.

- Control nodes (Installer nodes)
- Job management nodes
- Job management sub nodes
- File servers
(Meta Data Server / Object Storage Server)

- **Full automatic** failover

- Job management node/sub node is in **hot standby** mode.

- Continuing job execution even on the failure of the job management node
- Rapid failover without time lag



System Software Stack

User/ISV Applications

HPC Portal / System Management Portal

Technical Computing Suite

System operations management

- System configuration management
- System control
- System monitoring
- System installation & operation

High-performance file system

- Lustre-based distributed file system
- High scalability
- IO bandwidth guarantee
- High reliability & availability

Compilers

- Hybrid parallel programming
- Sector cache support
- SIMD / Register file extensions

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

Job operations management

- Job manager
- Job scheduler
- Resource management
- Parallel execution environment

VISIMPACT™

- Shared L2 cache on a chip
- Hardware intra-processor synchronization

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Linux-based enhanced Operating System

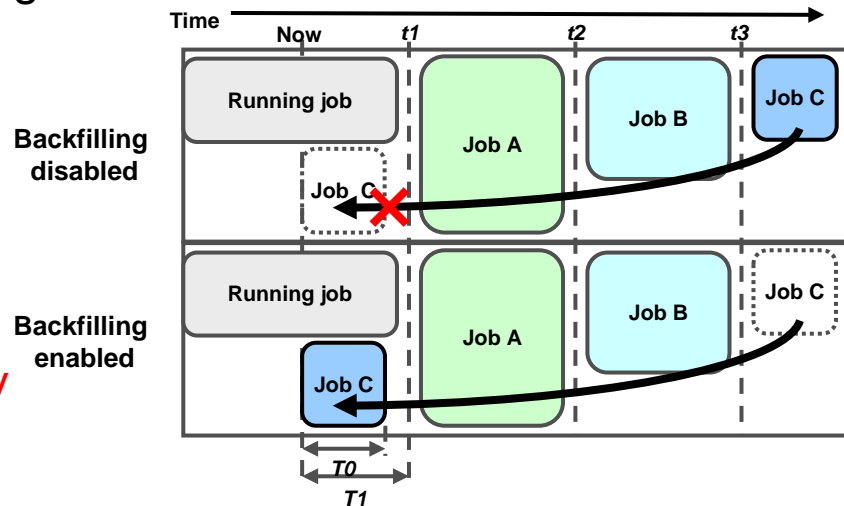
Red Hat Enterprise Linux

Super Computer: PRIMEHPC FX10

PC cluster: PRIMERGY

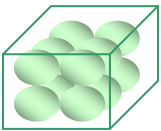
Job Operations Management

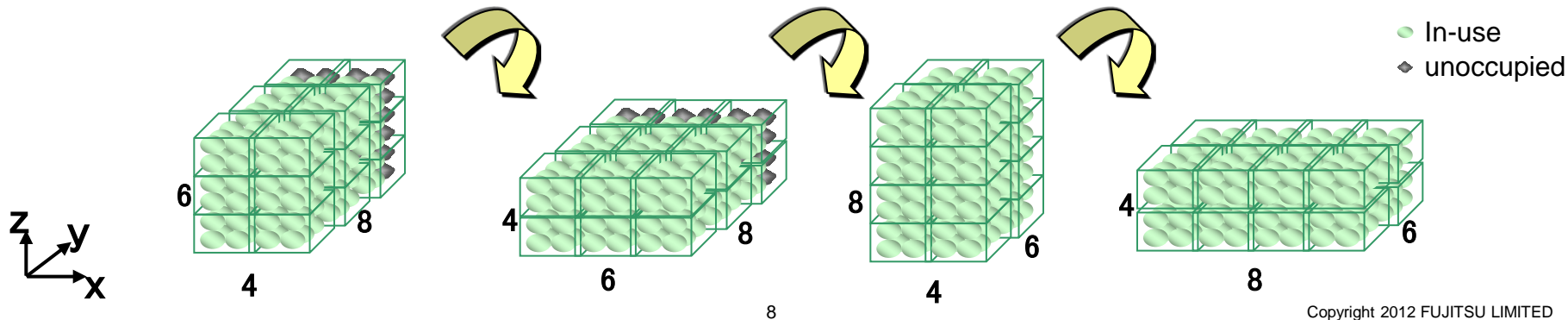
- Same job operations in FX10 and PRIMERGY
- Efficient, fair and system-optimal resource usage
 - Backfill scheduling
 - Fair share scheduling
 - System-optimal job scheduling
- Resource / Access control
 - Elapsed time / CPU time / Physical memory
 - Permission of job operation commands
 - Reduce OS Jitter / Power saving control



Optimal Job Scheduling for FX10

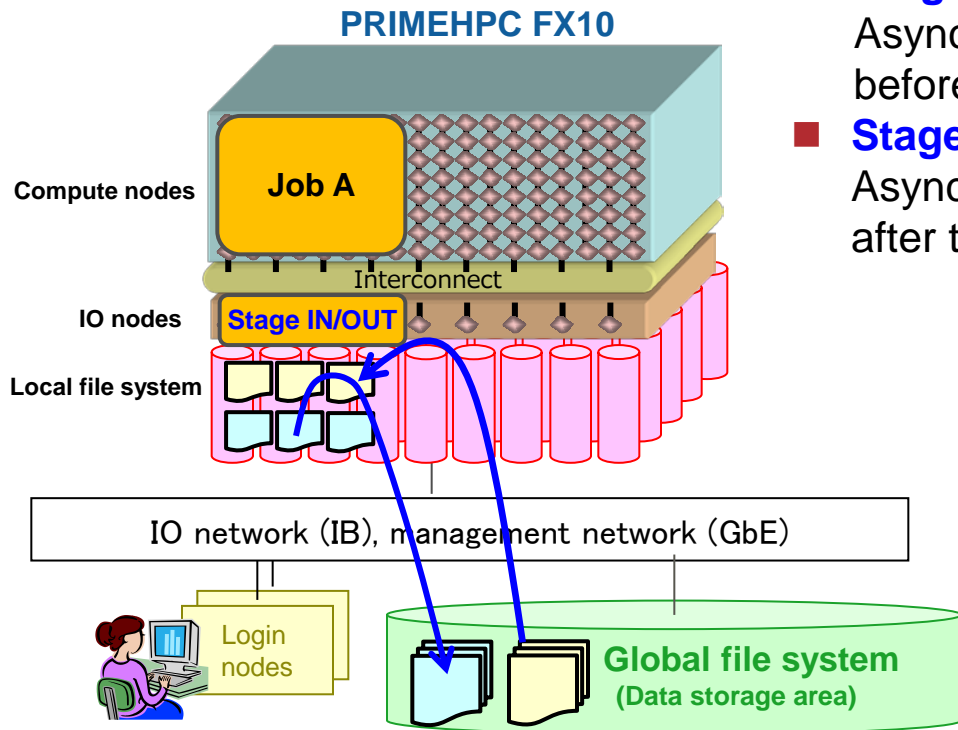
■ Interconnect topology-aware resource assignment

- One interconnect unit : 12 nodes (2 x 3 x 2) → 
- Job assignment rule: rectangular solid shape
 - ➔ Guaranteeing neighbor communication
 - ➔ Avoiding interfering with other jobs
- Rotates nodes to reduce fragmentation



Optimal Job Scheduling for FX10

■ Asynchronous file staging

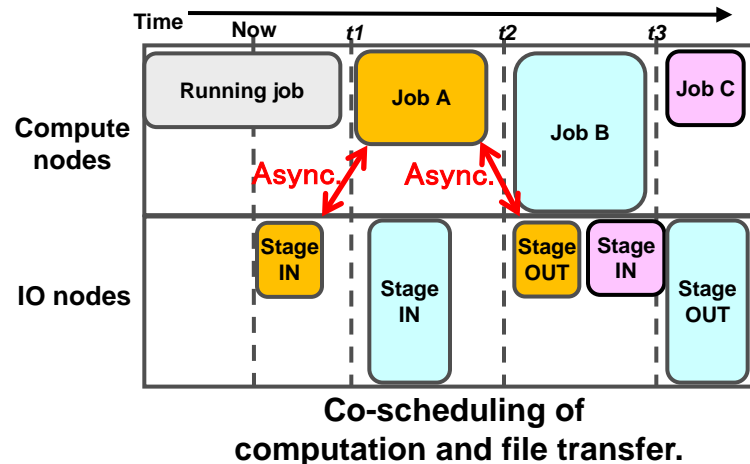


■ Stage IN

Asynchronously transfer files from Global to Local FS before the job starts.

■ Stage OUT

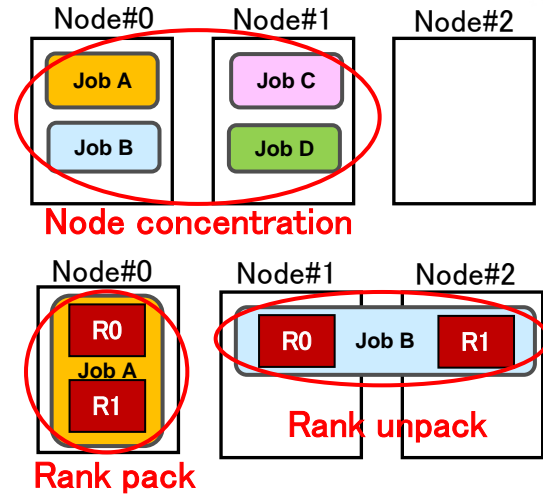
Asynchronously transfer files from Local to Global FS after the job ends.



Optimal Job Scheduling for PRIMERGY

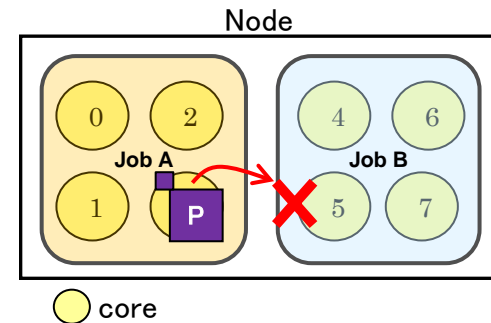
■ Fine-grained node assignment

- Node selection method : balancing / concentration
- Rank placement policy : pack / unpack
- Priority control of allocated nodes
- Execution mode : node is occupied or not by a job.



■ Strict core assignment

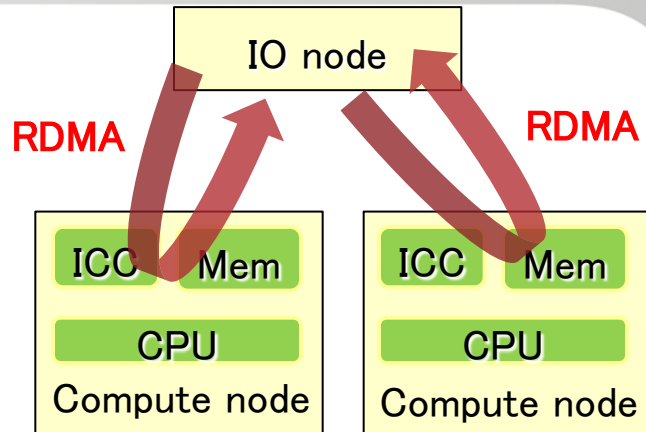
- Processes are bound to cores in the job territory.
- No process can move to cores in other job territory.



Reduce OS Jitter

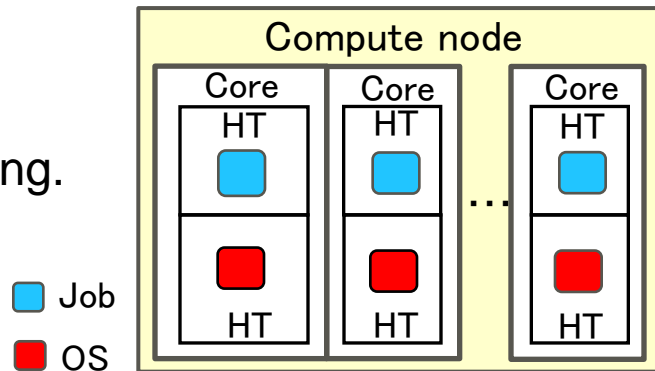
■ PRIMEHPC FX10

- Stripped-down system processing
- Minimizes OS Jitter by using RDMA of Tofu.
 - a. Node / service health check
 - b. System information monitor (remote sadc)
 - c. Job information monitor (CPU time/used memory)



■ PRIMERGY

- Isolates OS Jitter from jobs by using Hyper-Threading.
- ➡ Avoiding the conflict between job and OS Jitter.



System Software Stack

User/ISV Applications

HPC Portal / System Management Portal

Technical Computing Suite

System operations management

- System configuration management
- System control
- System monitoring
- System installation & operation

Job operations management

- Job manager
- Job scheduler
- Resource management
- Parallel execution environment

High-performance file system

- Lustre-based distributed file system
- High scalability
- IO bandwidth guarantee
- High reliability & availability

VISIMPACT™

- Shared L2 cache on a chip
- Hardware intra-processor synchronization

Compilers

- Hybrid parallel programming
- Sector cache support
- SIMD / Register file extensions

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Linux-based enhanced Operating System

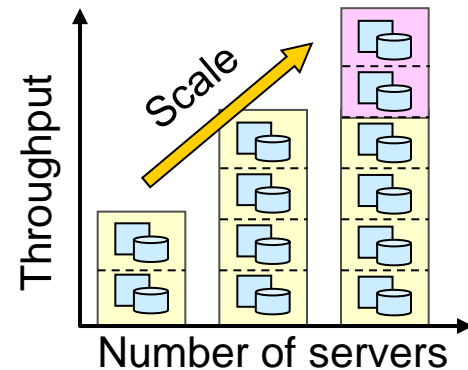
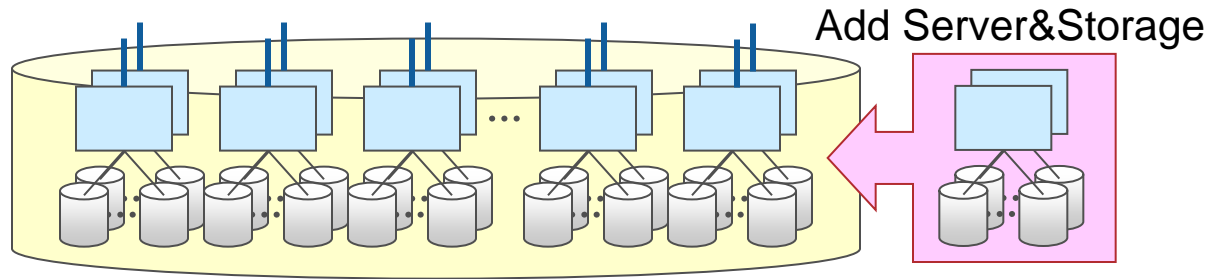
Red Hat Enterprise Linux

Super Computer: PRIMEHPC FX10

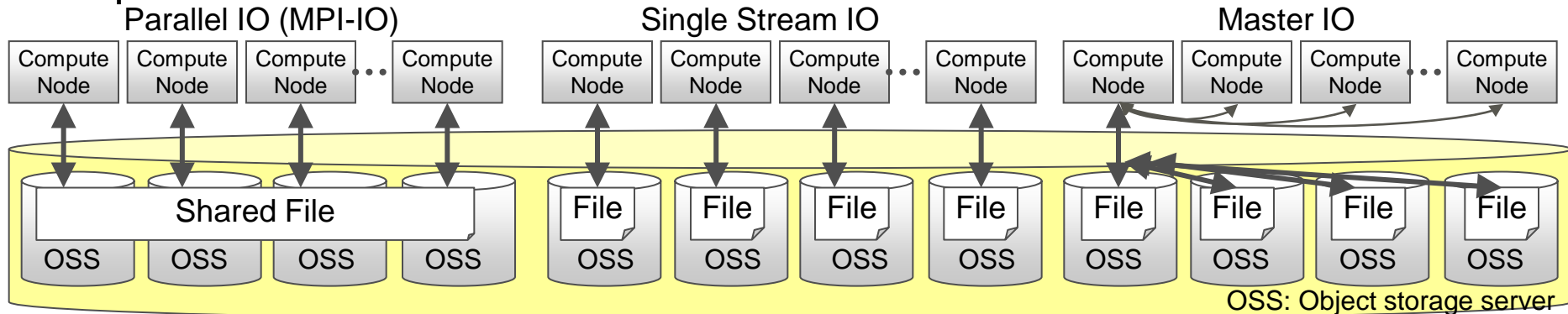
PC cluster: PRIMERGY

High Scalability

- Achieved high-scalable IO performance with multiple OSSes.



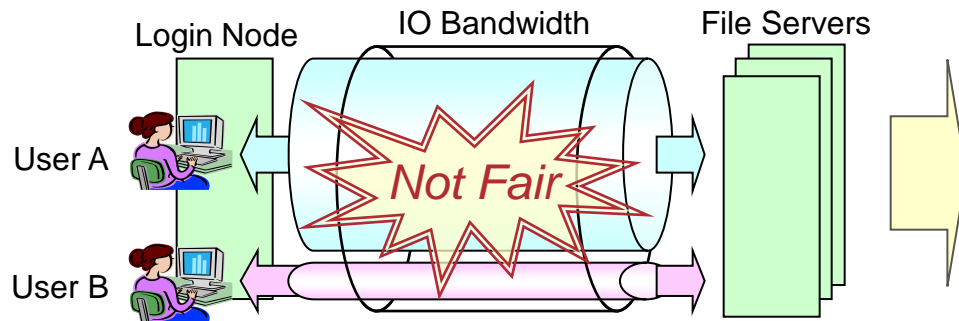
- Adapted various IO model



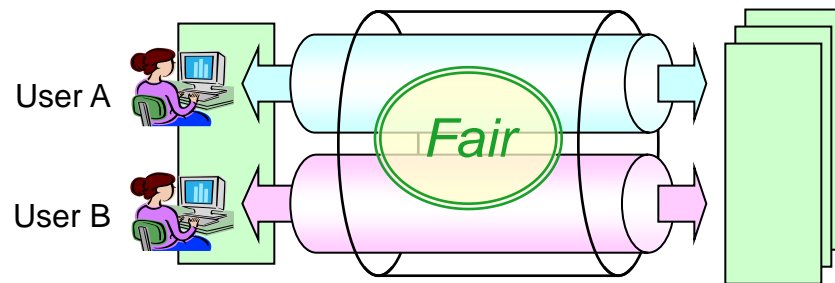
IO Bandwidth Guarantee

■ Fair Share QoS: Sharing IO bandwidth with all users.

Without Fair Share QoS

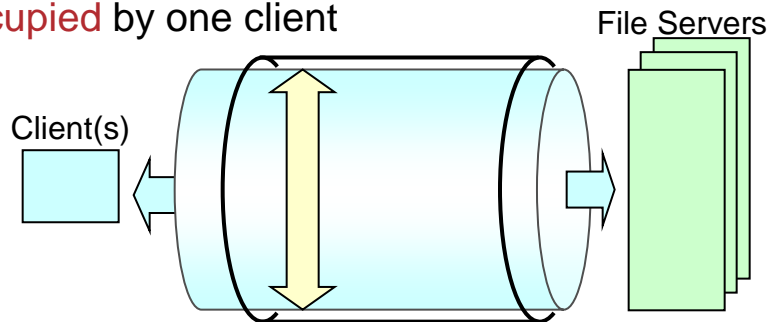


With Fair Share QoS

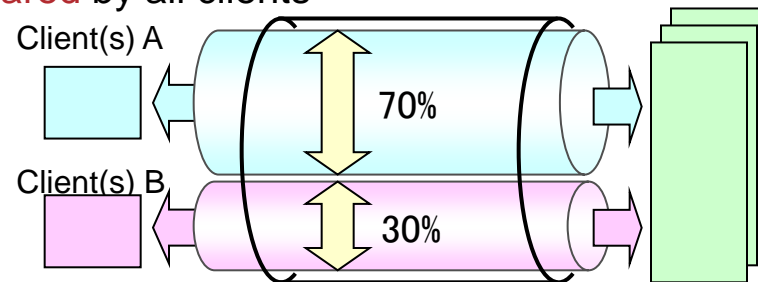


■ Best Effort QoS: Utilize all IO bandwidth exhaustively.

Occupied by one client

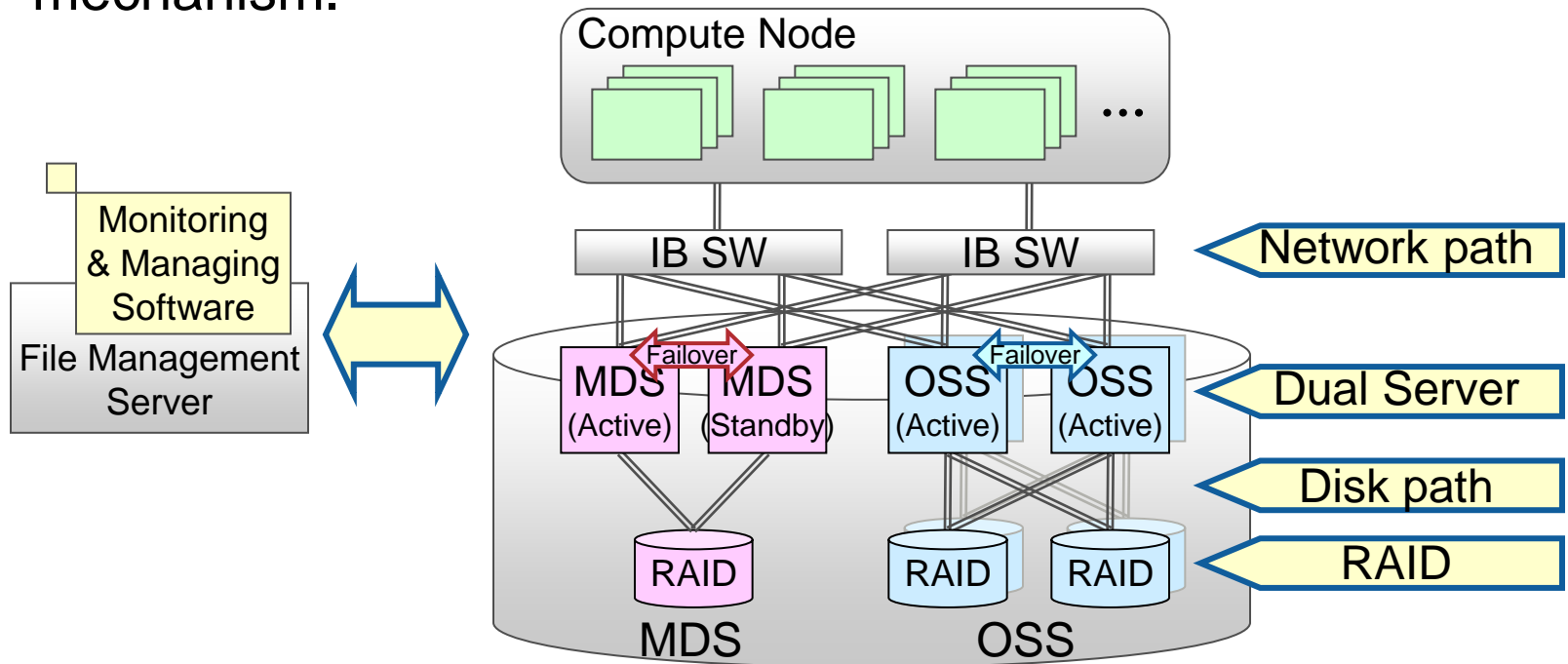


Shared by all clients



High Reliability and High Availability

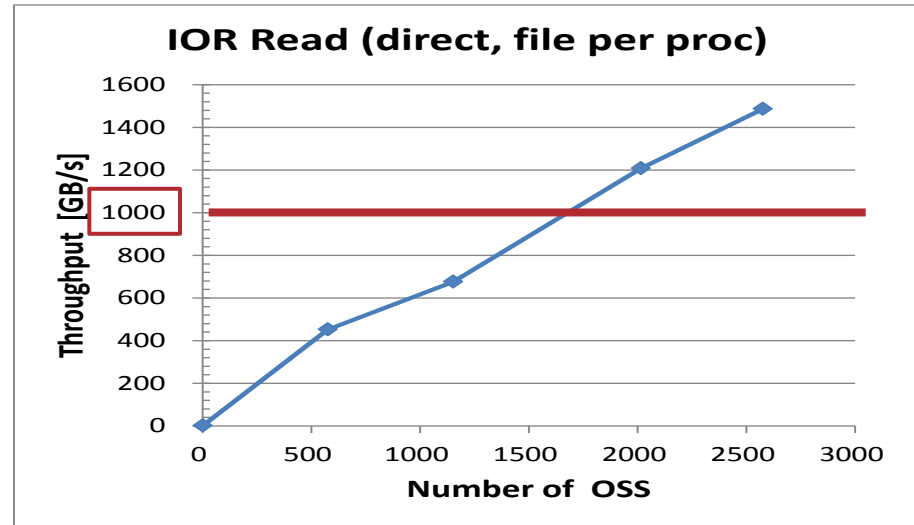
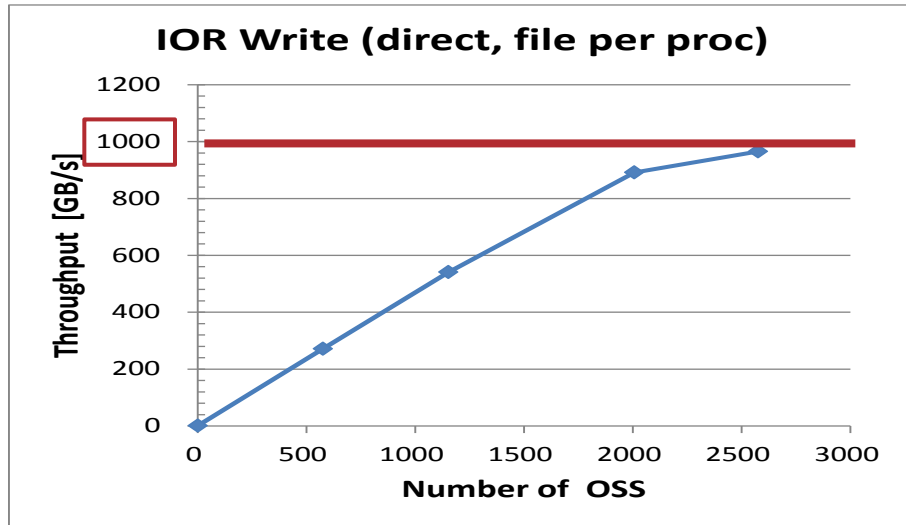
- Avoiding single point of failure by redundant hardware and failover mechanism.



Performance: I/O Throughput of FEFS

- Achieved the world's top-level throughput on K computer using over 2,500 OSS.
- We were encountered with serious problems: Memory shortage & System noise issues.
- Write : **965GB/s**
- Read : **1486GB/s**

Collaborative work with RIKEN



System Software Stack

User/ISV Applications

HPC Portal / System Management Portal

Technical Computing Suite

System operations management

- System configuration management
- System control
- System monitoring
- System installation & operation

Job operations management

- Job manager
- Job scheduler
- Resource management
- Parallel execution environment

High-performance file system

- Lustre-based distributed file system
- High scalability
- IO bandwidth guarantee
- High reliability & availability

VISIMPACT™

- Shared L2 cache on a chip
- Hardware intra-processor synchronization

Compilers

- Hybrid parallel programming
- Sector cache support
- SIMD / Register file extensions

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Linux-based enhanced Operating System

Red Hat Enterprise Linux

Super Computer: PRIMEHPC FX10

PC cluster: PRIMERGY

VISIMPACT

thread & process Hybrid-Parallel Programming

■ Auto-parallel + MPI

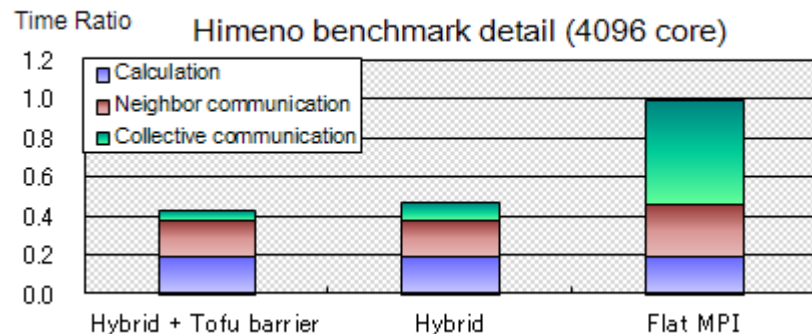
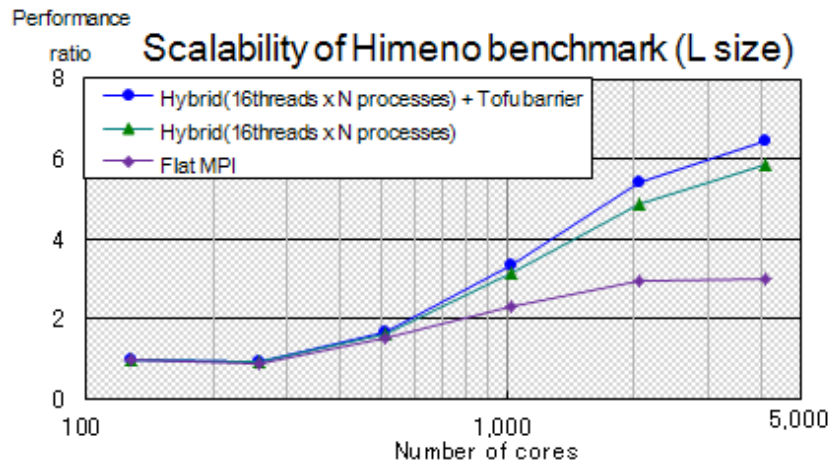
■ “Auto-thread parallel” in a chip

■ VISIMPACT: CPU Architecture for low overhead parallelism among cores

- Inter-core hardware barrier
- Shared L2 Cache
- Automatic parallelization

■ Process parallel by MPI

■ Tofu barrier facility for collective communication



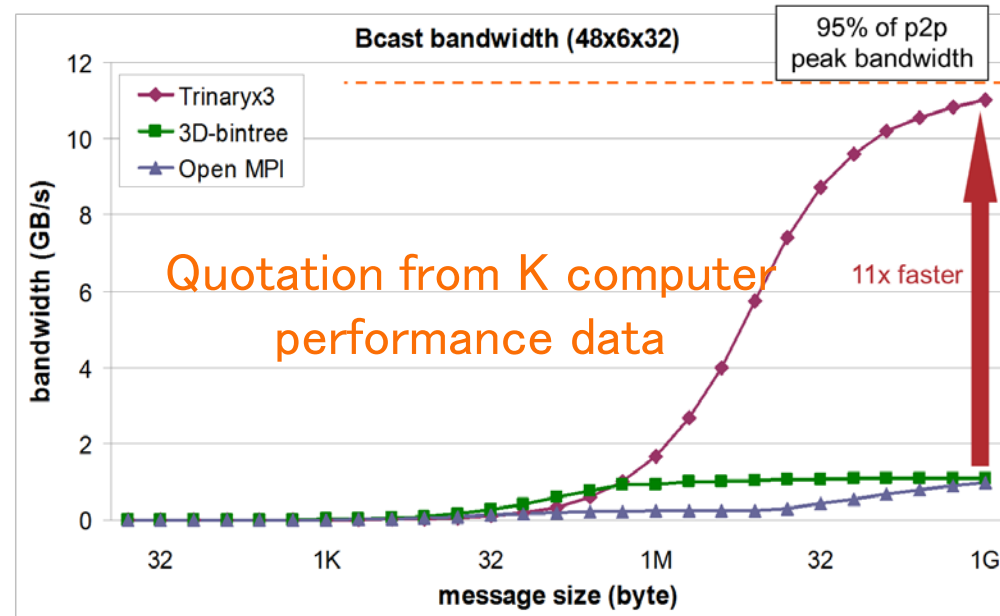
Customized MPI Library for High Scalability

■ Point-to-Point communication

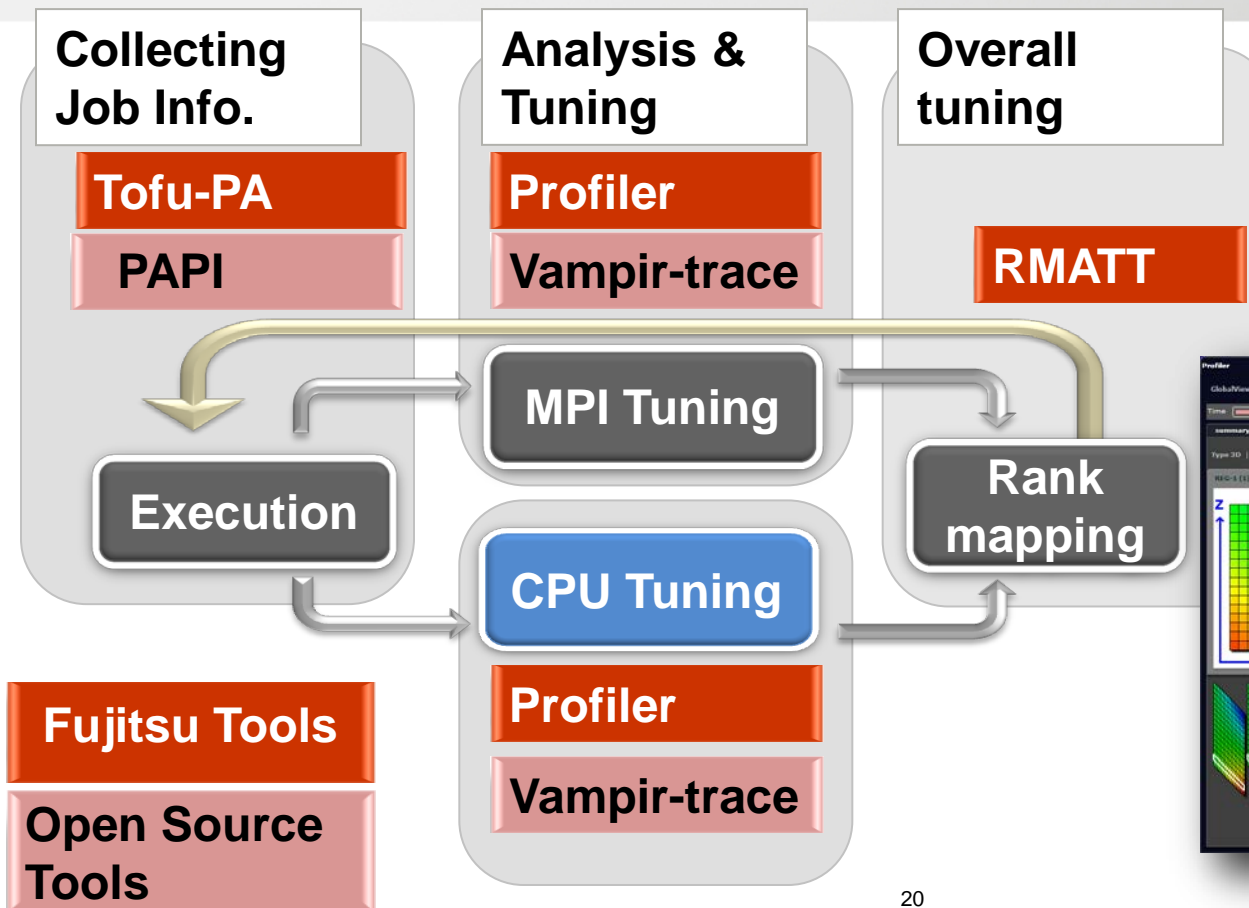
- The transfer method selection according to the data length, process location and number of hops

■ Collective communication

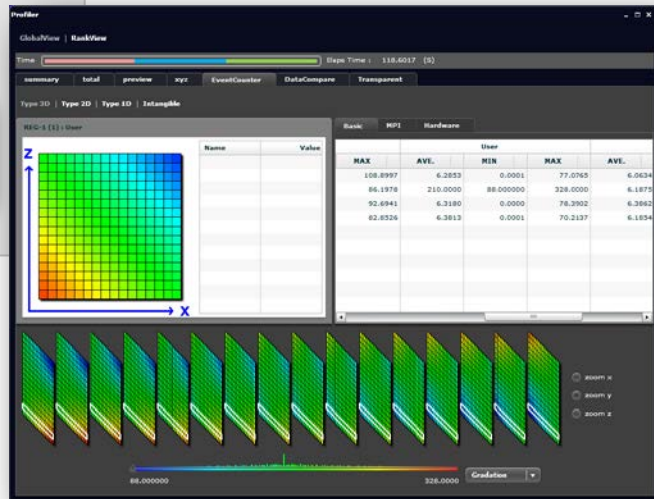
- Barrier, Allreduce, Bcast and Reduce use Tofu-barrier & Reduction facility
- Bcast, Allgather, Allgatherv, Allreduce and Alltoall use Tofu-optimized algorithm



Application Tuning Cycle and Tools

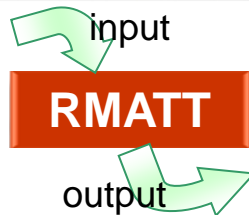


Profiler snapshot



Rank Mapping Optimization (RMATT)

Network Construction
Communication Pattern (Communication
processing contents between Rank)



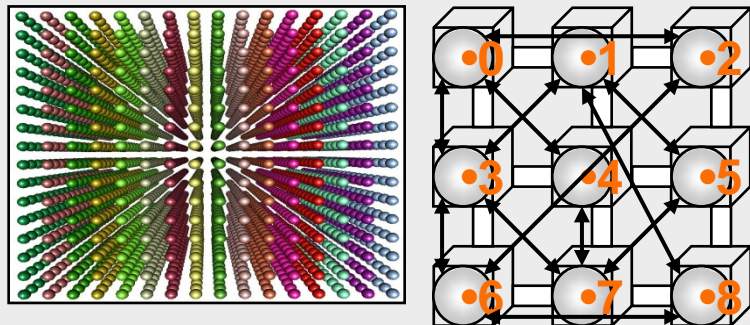
Optimized Rank Map
Reduce number of hop and congestion



Apply MPI_Allgather Communication Processing Performance

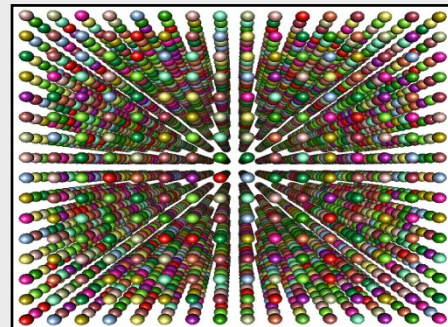
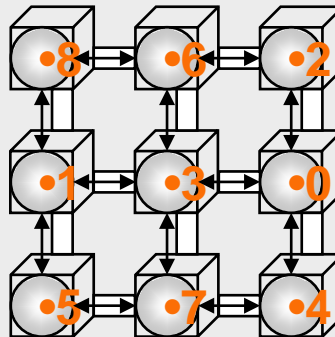
- Rank number : 4096 rank
- Network construction : 16x16x16 node (4096)

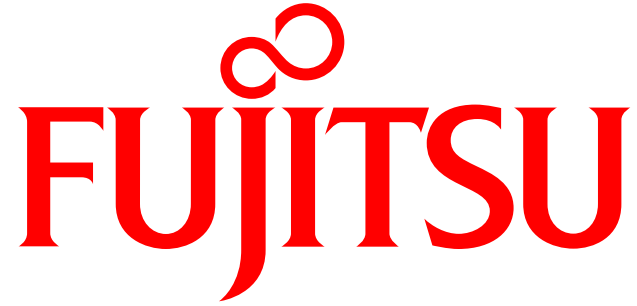
x,y,z order mapping 22.3ms



Remapping used RMATT 5.5ms

4 times performance Up





shaping tomorrow with you