FUJITSU

shaping tomorrow with you

Technical Computing Suite
Job Management Software

Toshiaki Mikamo

Fujitsu Limited

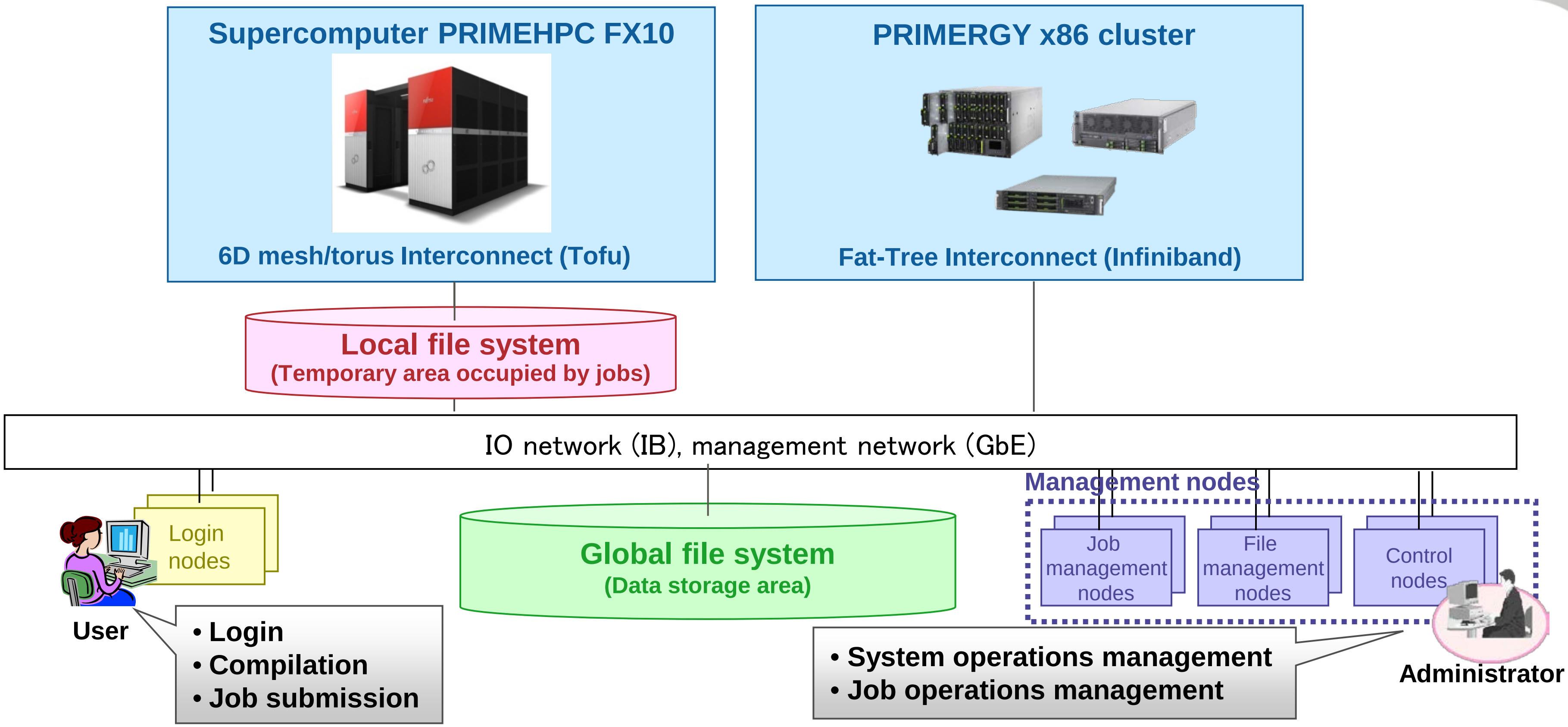
Supercomputer PRIMEHPC FX10

PRIMERGY x86 cluster

Outline

- System Configuration and Software Stack
- Features
- The major functions of job scheduler
  - Efficient Resource Usage
  - Fair Share Scheduling
  - System-optimal Resource Assignment
- Summary and Future

Hybrid System Configuration



Supercomputer PRIMEHPC FX10

6D mesh/torus Interconnect (Tofu)

PRIMERGY x86 cluster

Fat-Tree Interconnect (Infiniband)

Local file system

(Temporary area occupied by jobs)

IO network (IB), management network (GbE)

Login nodes

Global file system

(Data storage area)

Management nodes

Job management nodes

File management nodes

Control nodes

User

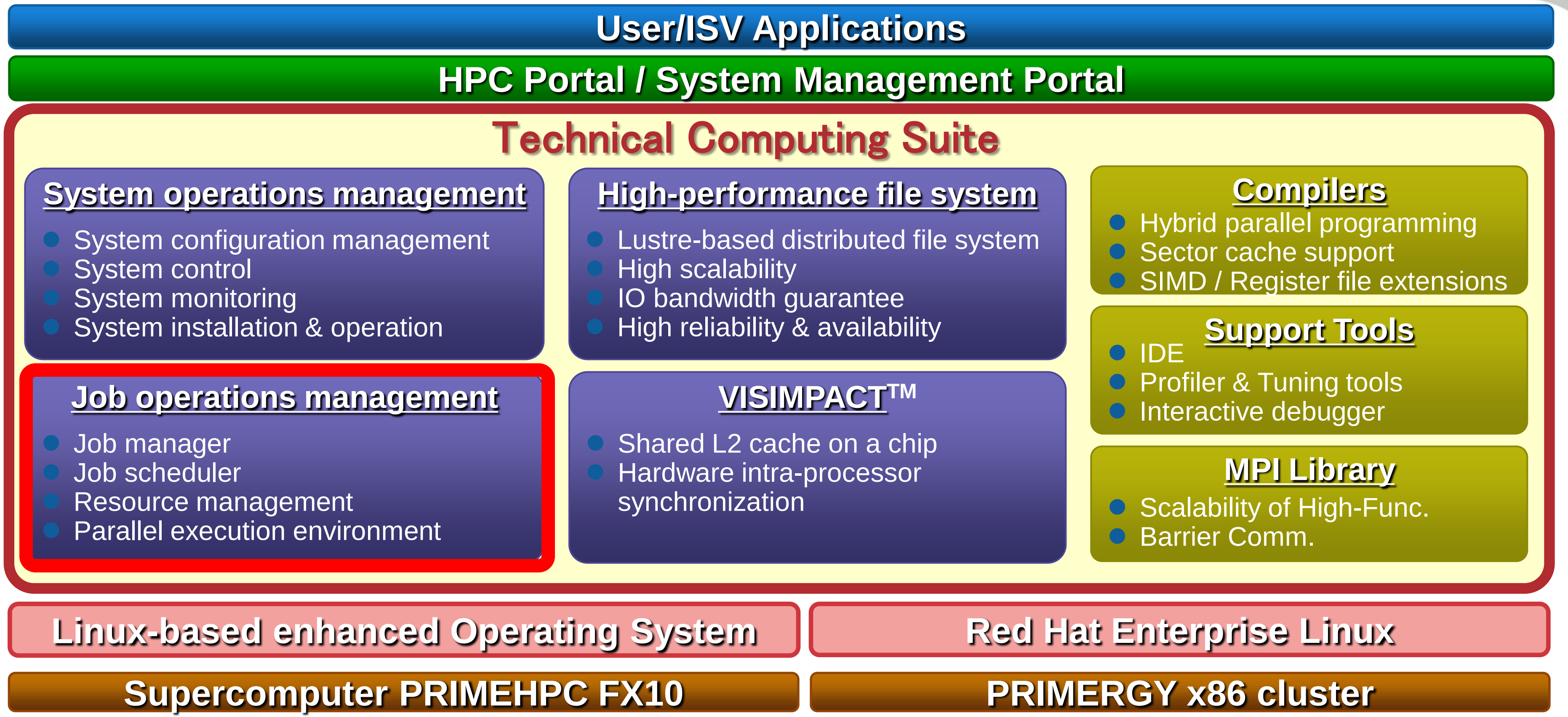
- Login
- Compilation
- Job submission

- System operations management
- Job operations management

Administrator

System Software Stack

FUJITSU



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

Job operations management

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

VISIMPACT^{TM}

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

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**

Supercomputer PRIMEHPC FX10 | **PRIMERGY x86 cluster**

Features

- Same job operations in FX10 and PRIMERGY
- Efficient, fair and system-optimal job scheduling
  - See slide below for details
- Resource / Access control
  - Elapsed time limit / CPU time limit / Physical memory limit
  - Enable / Disable execute permission of job operation commands
  - Reduce OS jitter / Power saving control
- Job statistical information
  - The amount of CPU time / Memory / IO
  - SIMD rate / MIPS / MFLOPS

Job Scheduler

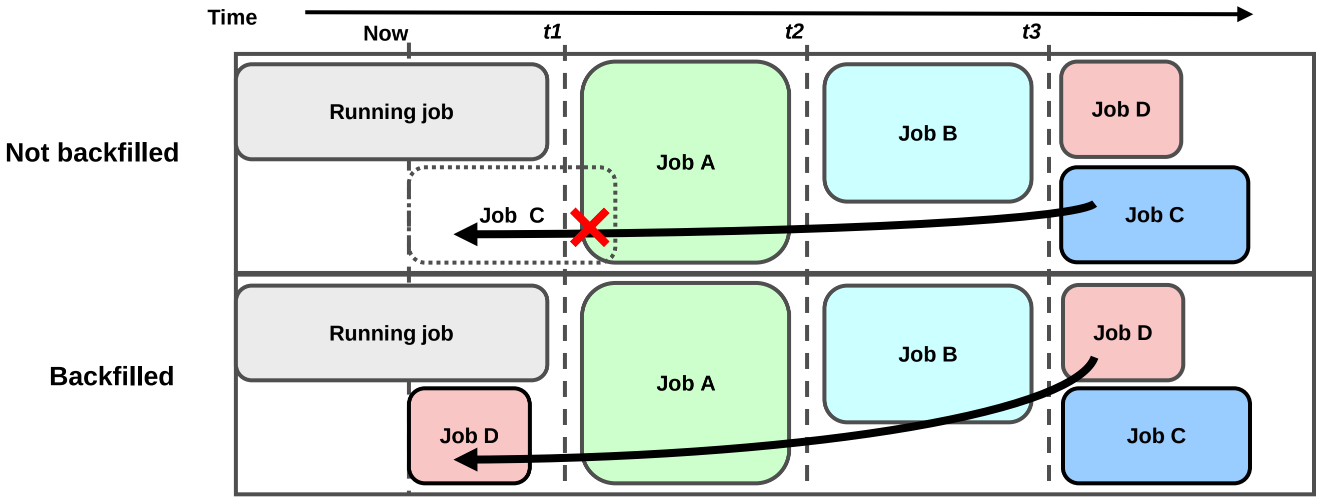
- Renew our job scheduler for large-scale system
- Our job scheduler features:
  - Multi-process

    enable to coexist multiple scheduler in a cluster.
  - Multi-thread

    enable to balance the load of scheduling.

Efficient Resource Usage

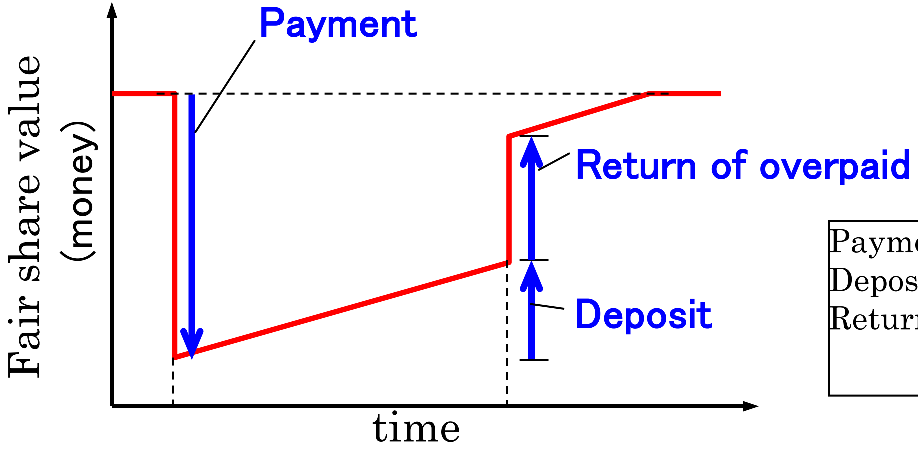
- Backfill scheduling for keeping the resources busy
  - Our scheduler manages space(compute nodes) and time.
  - It will backfill the low priority jobs so as not to prevent high priority jobs.



Fair Share Scheduling

- Fairly share resources between users/groups based on past usage.
  ① Fair share value is issued in advance for each user/group.
  ② The value is changed by the result of resource usage.
  ③ The job execution priority is determined dynamically according to the value.

Fair share value is like money.

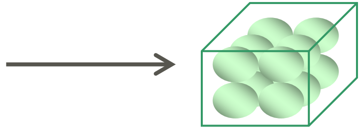


Payment[P] = (#Node allocated) x (Elapsed time limit of job)
Deposit[D] = (Elapsed time) x (Recovery rate)
Return of overpaid[R]
= P - ((#Node allocated) x (Actual elapsed time of job))

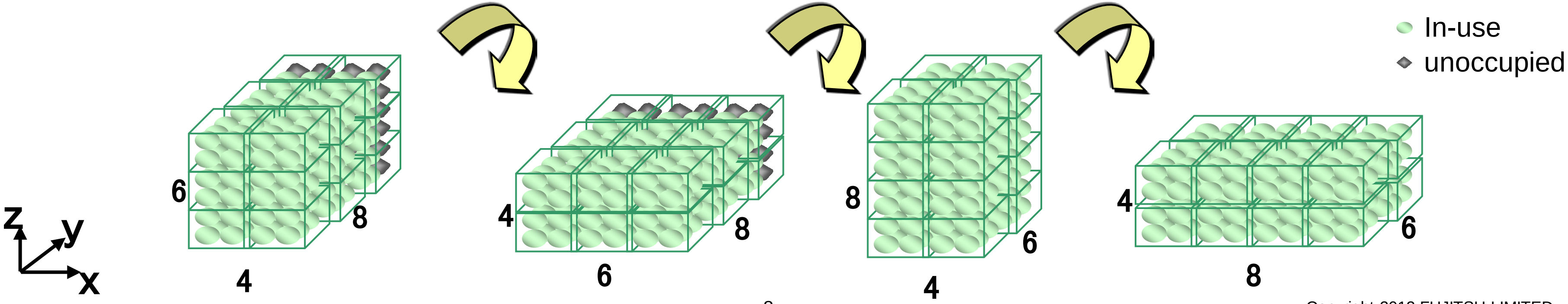
Optimal Job Scheduling for FX10

Interconnect topology-aware resource assignment

- One interconnect unit : 12 nodes (2 x 3 x 2)

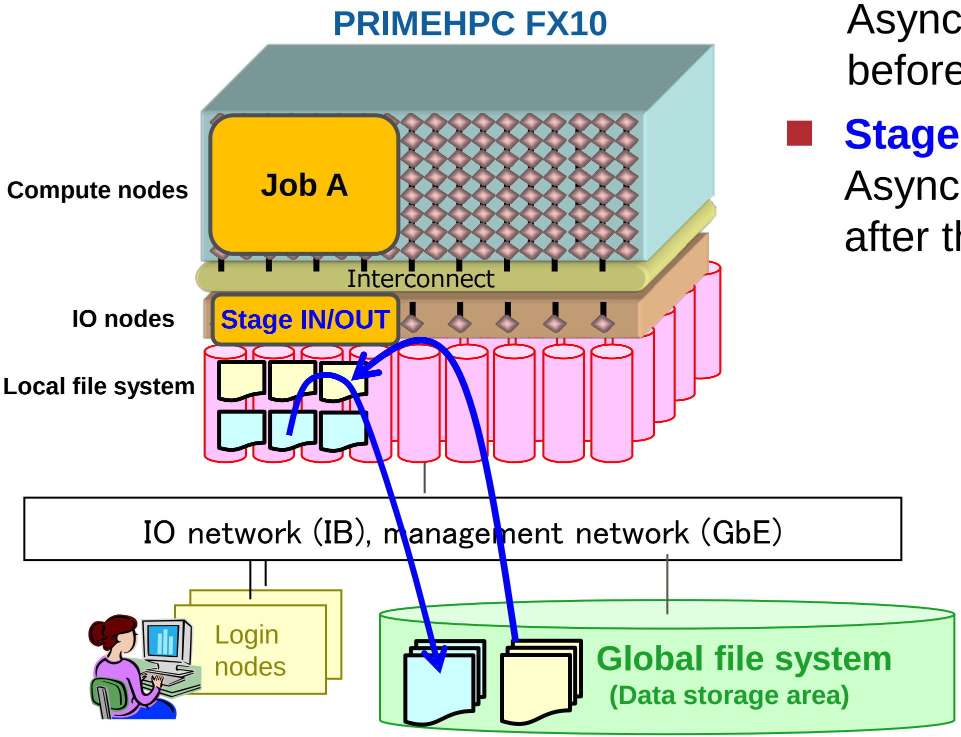


- Job assignment rule: rectangular solid shape
  - ➔ Guaranteeing neighbor communication
  - ➔ Avoiding interfering with other jobs
- Rotates rectangular solid of interconnect unit to reduce fragmentation

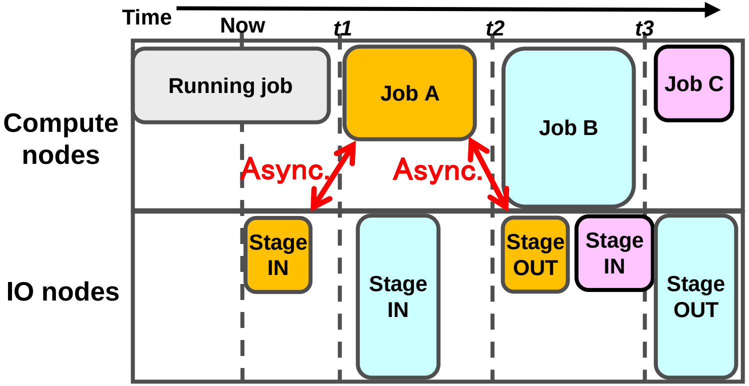


Optimal Job Scheduling for FX10

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.

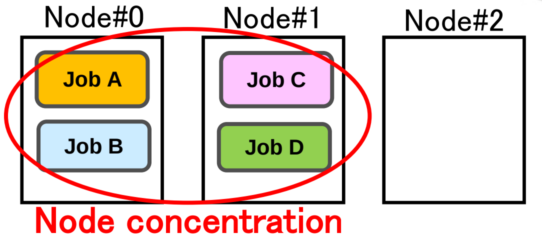


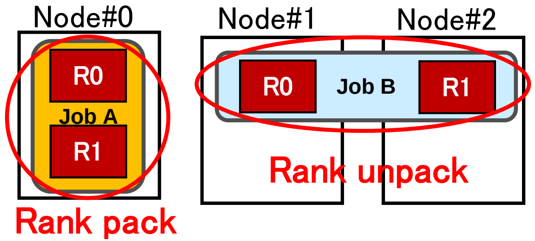
Co-scheduling of computation and file transfer.

Optimal Job Scheduling for PRIMERGY

Fine-grained node assignment

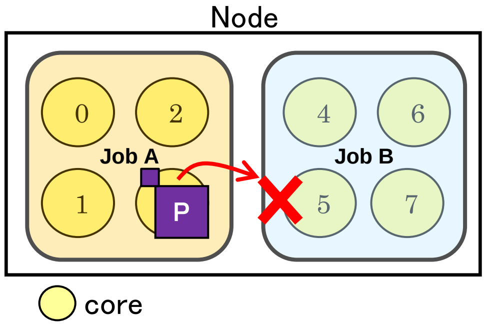
- Node selection method : balancing / concentration
- Rank placement policy : pack / unpack
- Priority control of allocated nodes
- Execution mode : node is occupied or not by a job.





Strict core assignment

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



Summary and Future

- We developed the job management software.
  - Unified operability on PRIMEHPC FX10 and PRIMERGY
  - New job scheduler : Efficiency, Fairness and System-optimization
  - Practical resource control and job statistical information
- Future Work
  - Operation simulator
    Administrator will be able to simulate the operation situation subsequent to operation parameter changes.

FUJITSU

shaping tomorrow with you