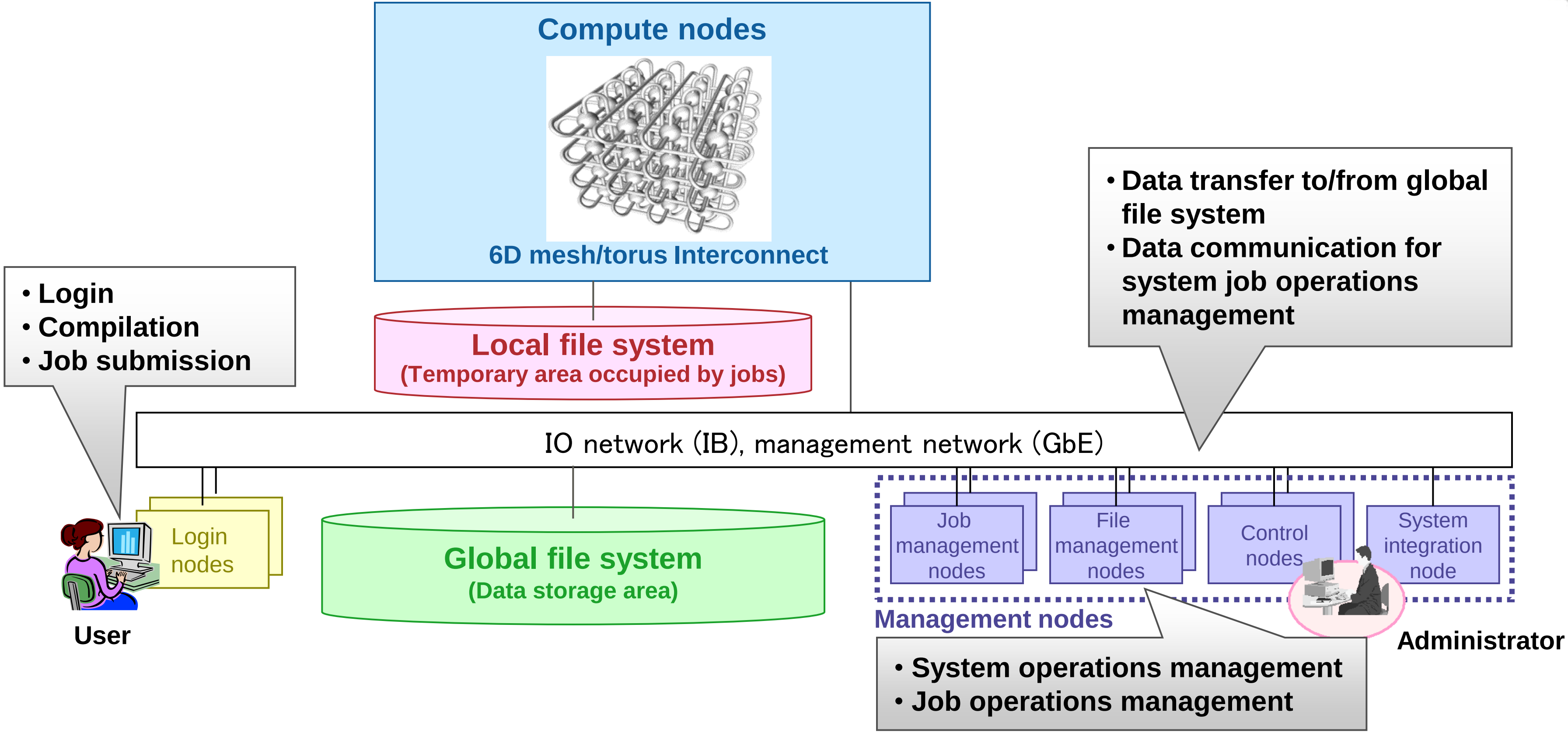
shaping tomorrow with you

Advanced Software for the Supercomputer PRIMEHPC FX10

System Configuration of PRIMEHPC FX10



Compute nodes

6D mesh/torus Interconnect

Local file system
(Temporary area occupied by jobs)

- Login
- Compilation
- Job submission

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

IO network (IB), management network (GbE)

Login nodes

User

Global file system
(Data storage area)

Job management nodes

File management nodes

Control nodes

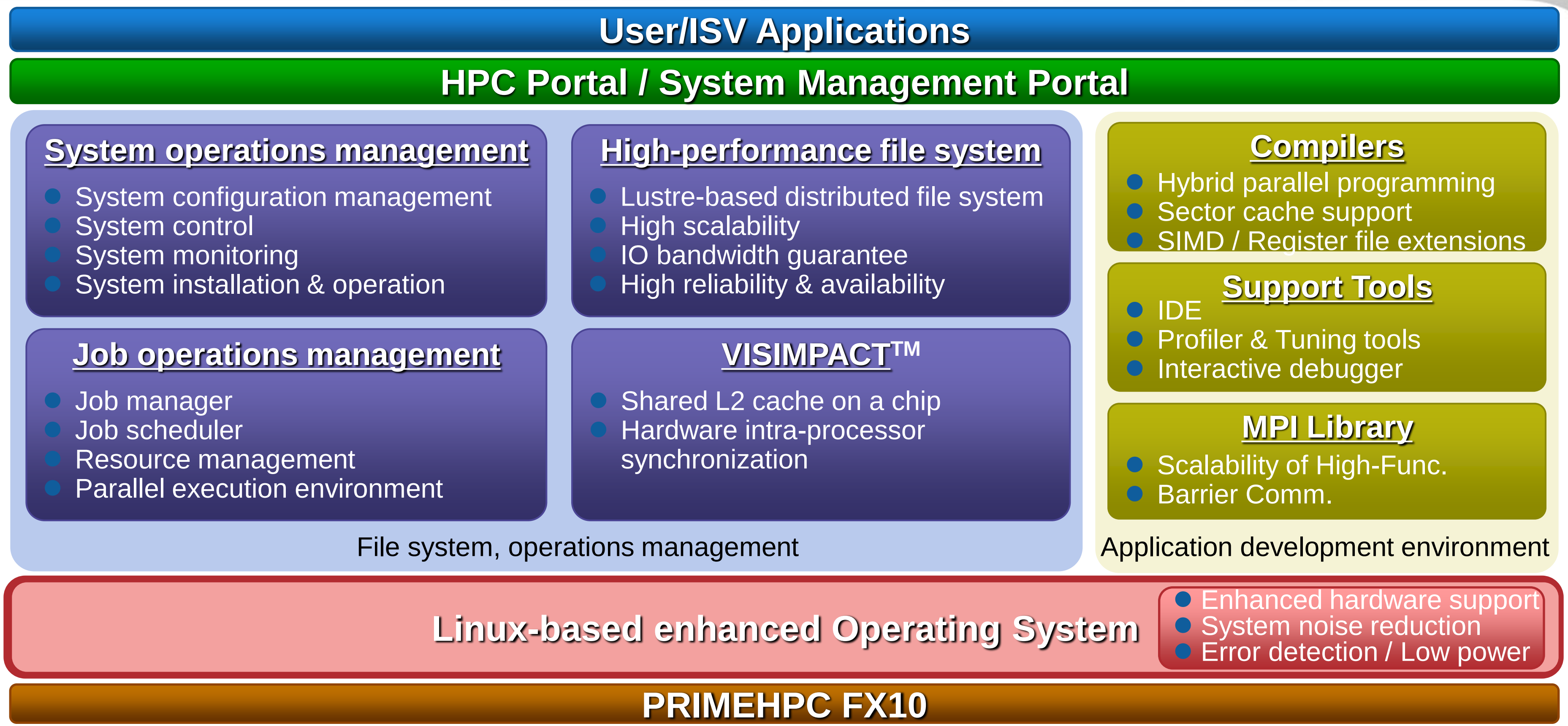
System integration node

Management nodes

Administrator

- System operations management
- Job operations management

System Software Stack



User/ISV Applications

HPC Portal / System Management Portal

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

Job operations management

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

VISIMPACT™

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

File system, operations management

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.

Application development environment

Linux-based enhanced Operating System

- Enhanced hardware support
- System noise reduction
- Error detection / Low power

PRIMEHPC FX10

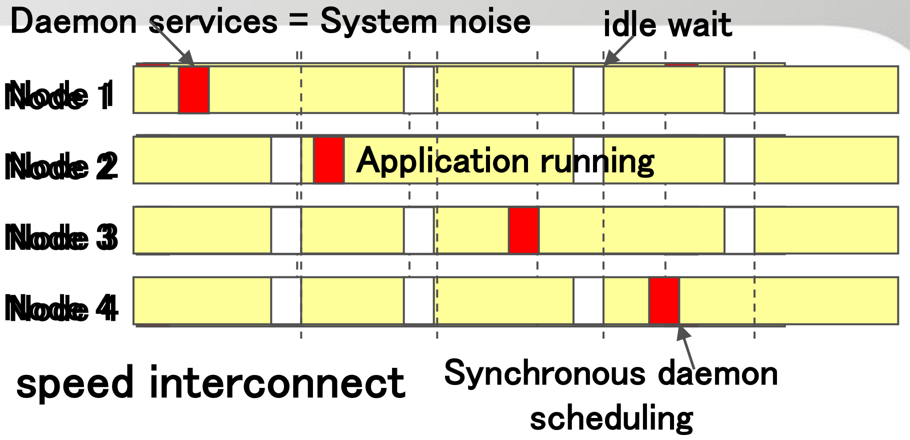
OS (Linux-based enhanced Operating System)

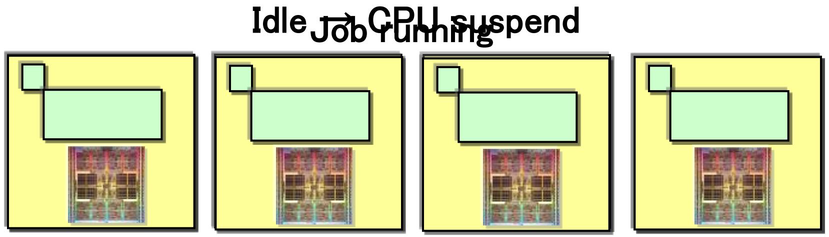
- Easy existing application porting
  - POSIX API: Linux kernel 2.6.x, glibc 2.x
- High performance / High scalability
  - Enhanced hardware support

    CPU registers, Large memory page, High speed interconnect
  - Reduce system noise in highly parallel program

    Inter−node OS scheduling
- High availability / low power consumption
  - Hardware error detection / isolation

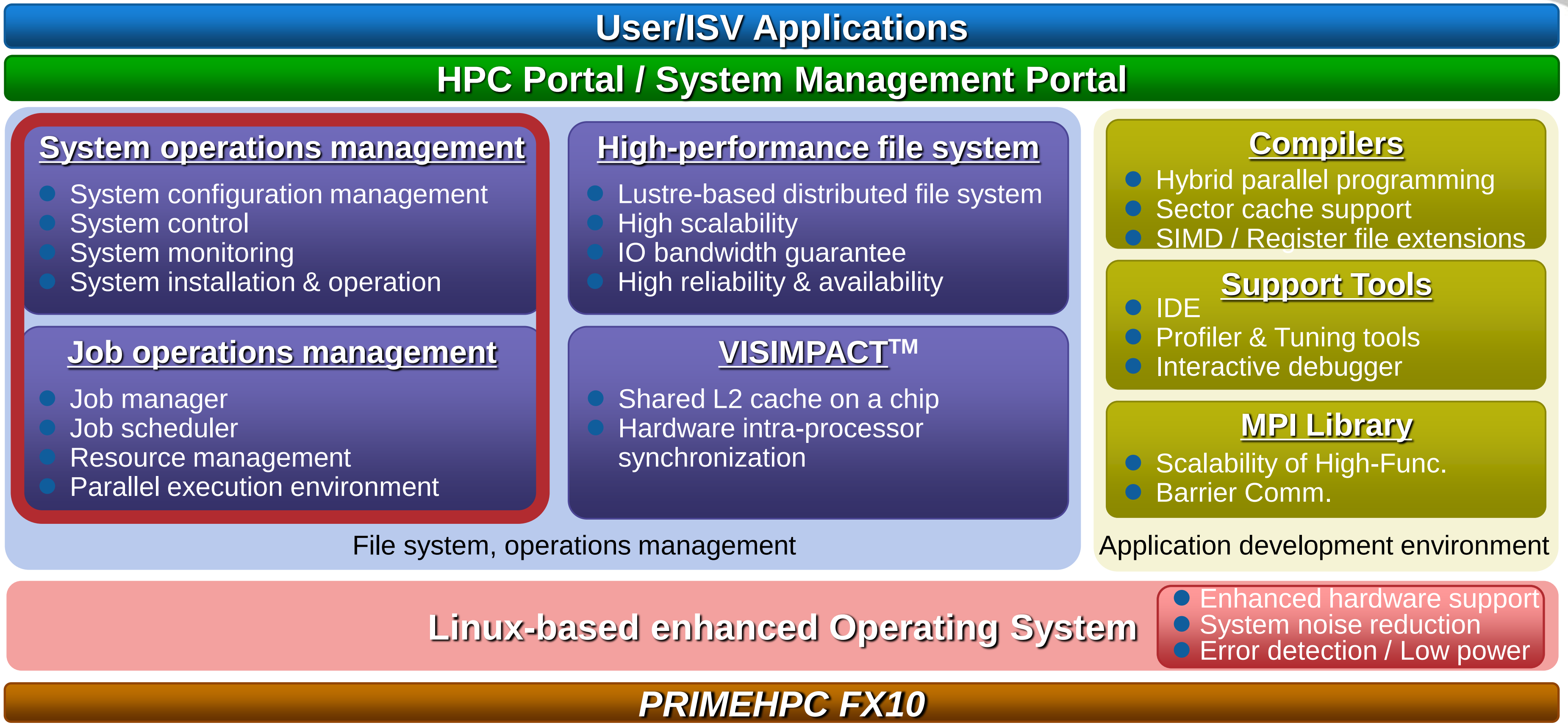
    memory patrol, io driver enhance.
  - CPU suspend during system idle state.





System Software Stack

FUJITSU



User/ISV Applications

HPC Portal / System Management Portal

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

File system, operations management

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.

Application development environment

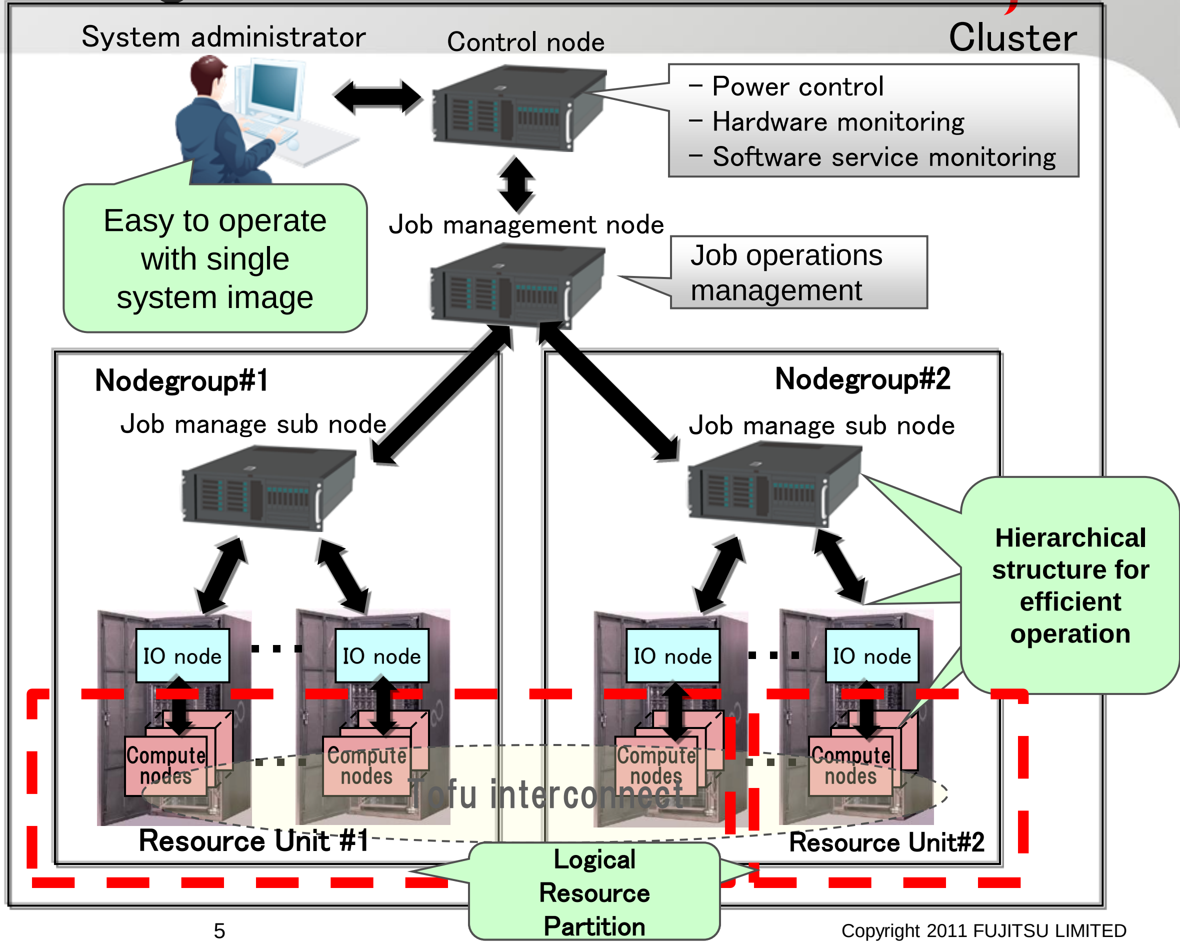
Linux-based enhanced Operating System

- Enhanced hardware support
- System noise reduction
- Error detection / Low power

PRIMEHPC FX10

System Operations Management

- Hierarchical structure for efficient system operation and adaptability to large-scale systems
    - The load is distributed by using the job management sub node.
- Easy to operate with a single system image
- The system is efficiently operated by dividing a logical resource partition named “Resource Unit”.

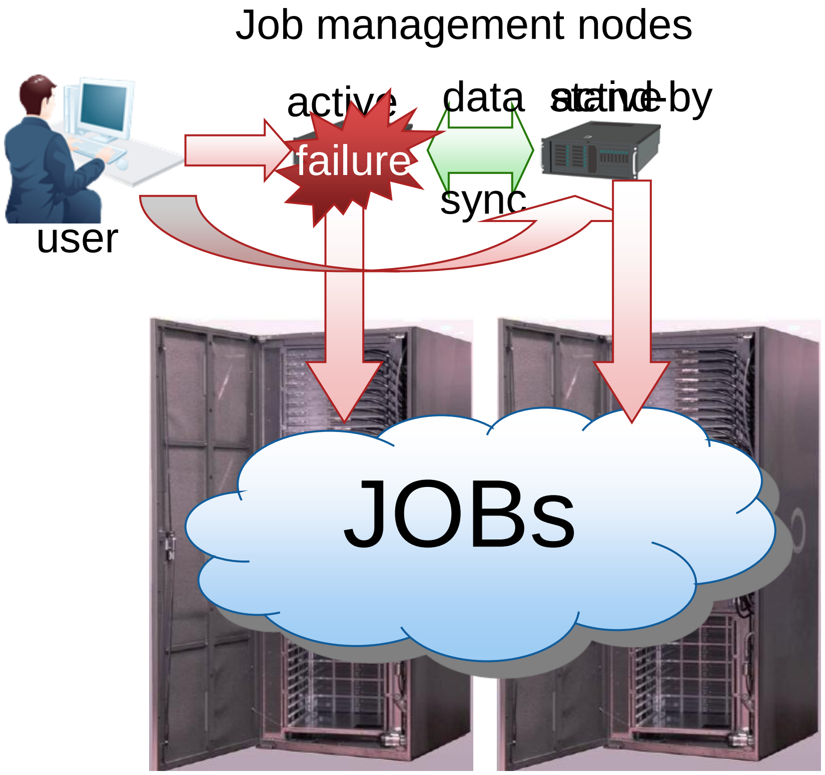


High Availability System

- The important nodes have redundancy
  - Control node
  - Job management node
  - Job management sub node
  - File servers

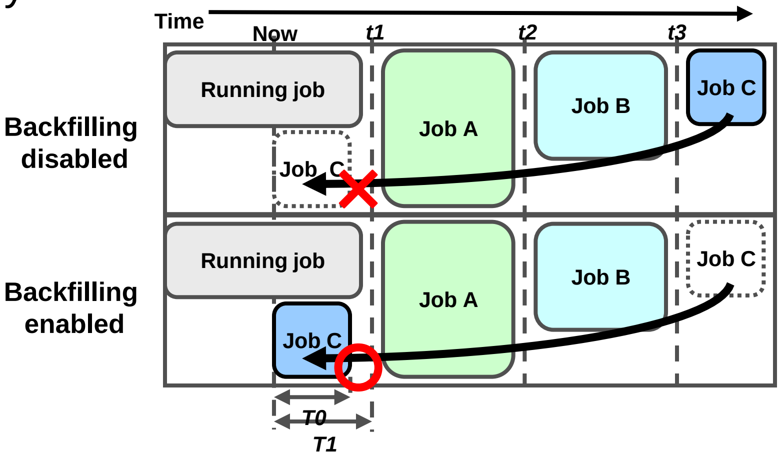
For example : right figure

- Continuing job execution even if the job management node is in failed status
  - The job data always synchronizes between active node and stand-by node.
  - Alternatively to stand-by node if active node is down.



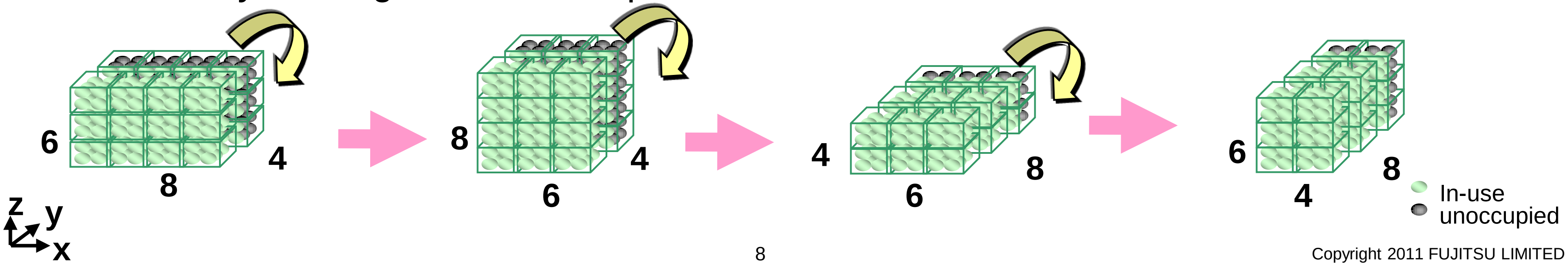
Job Operations Environment

- Efficient resource usage
  - Flexible job scheduling based on prioritized resource assignment
  - Interconnect topology-aware resource assignment
  - Backfill scheduling for keeping the resources busy
  - Asynchronous file staging
- High availability
  - Avoids assigning faulty resources to jobs
  - disconnects faulty nodes from job operations
  - Management nodes with failover support

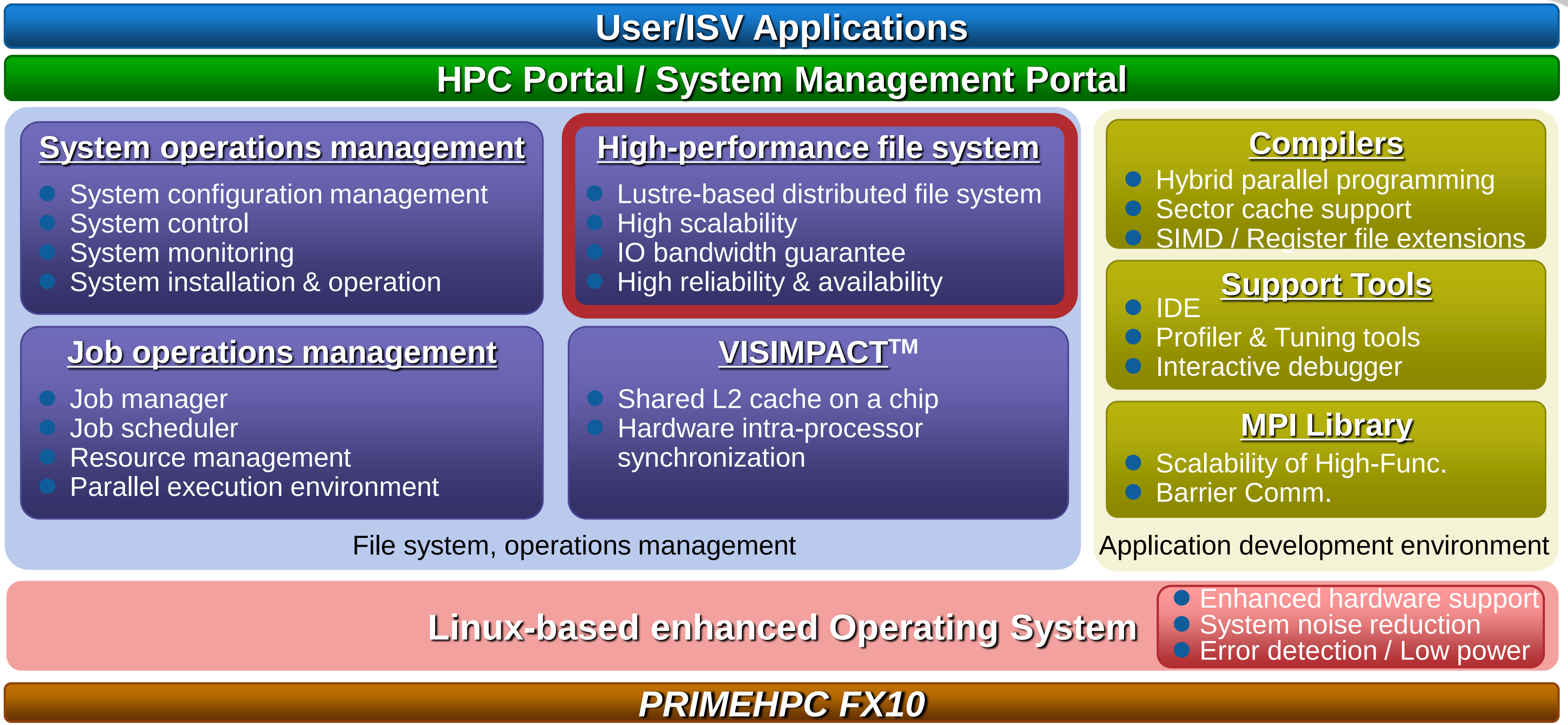


Resource Assignment

- Interconnect topology-aware resource assignment
  - Treats 12 compute nodes as one interconnect unit
  - Assigns cubic-shaped interconnect unit(s) to a job
  - → Using adjacent interconnect unit(s) is suitable for contiguous communication, and also avoids interfering with other jobs.
  - Optimizes the alignment of resources
  - → Rotating the cubic-shaped interconnect units This improves total system utilization by rotating the cubic shaped interconnect units.



System Software Stack



User/ISV Applications

HPC Portal / System Management Portal

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

Job operations management

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

VISIMPACT™

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

File system, operations management

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.

Application development environment

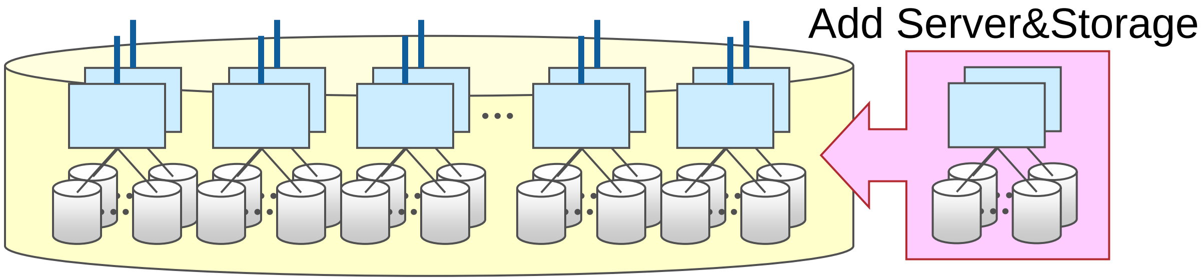
Linux-based enhanced Operating System

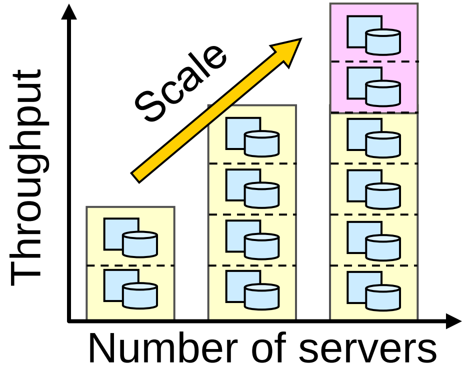
- Enhanced hardware support
- System noise reduction
- Error detection / Low power

PRIMEHPC FX10

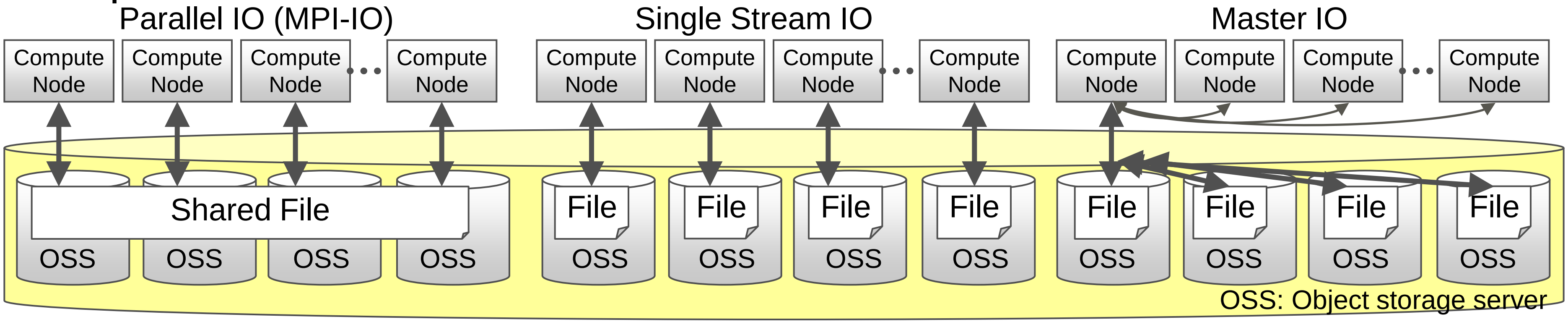
High Scalability

- Achieved high-scalable IO performance with multiple OSSes.





- Adapted various IO model

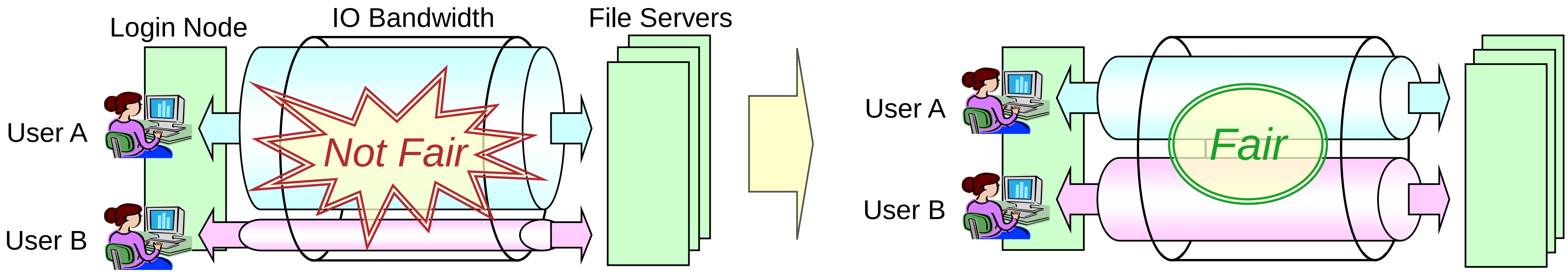


IO Bandwidth Guarantee

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

Without Fair Share QoS

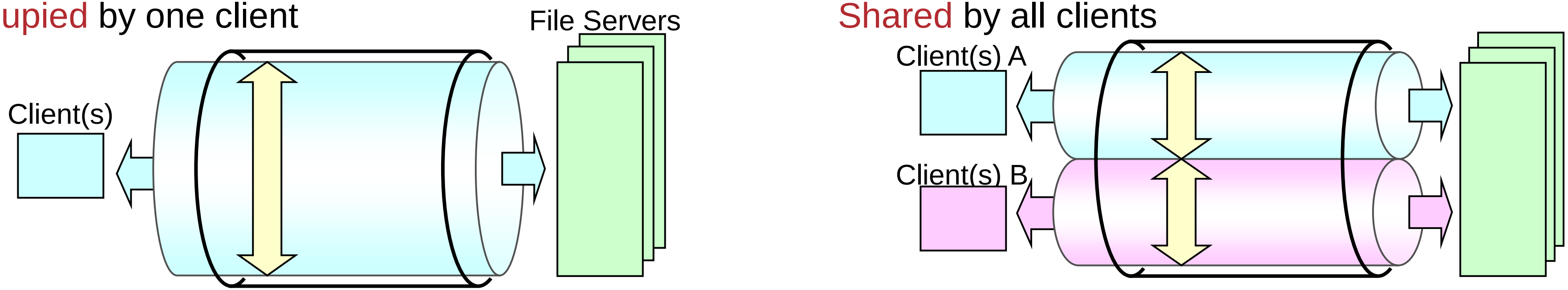
With Fair Share QoS



- Best Effort QoS: Utilize all IO bandwidth exhaustively.

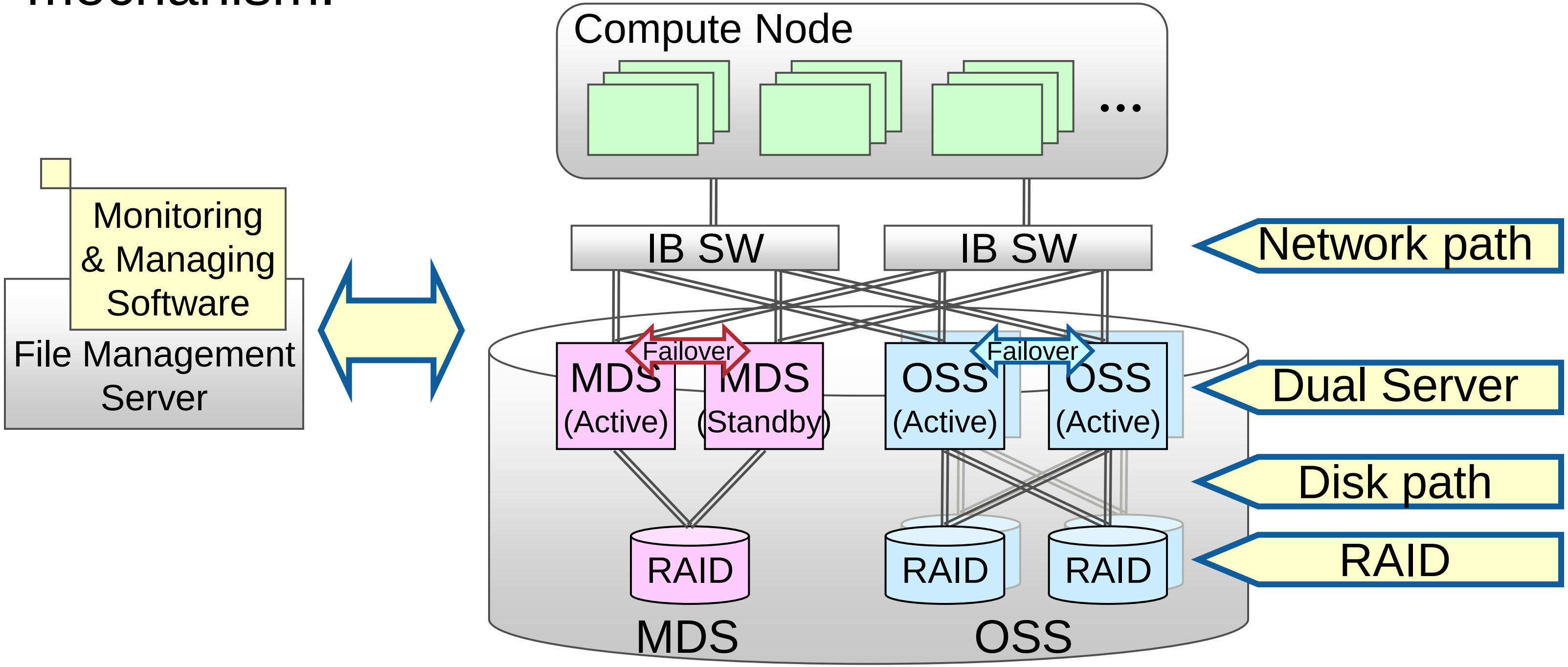
Occupied by one client

Shared by all clients

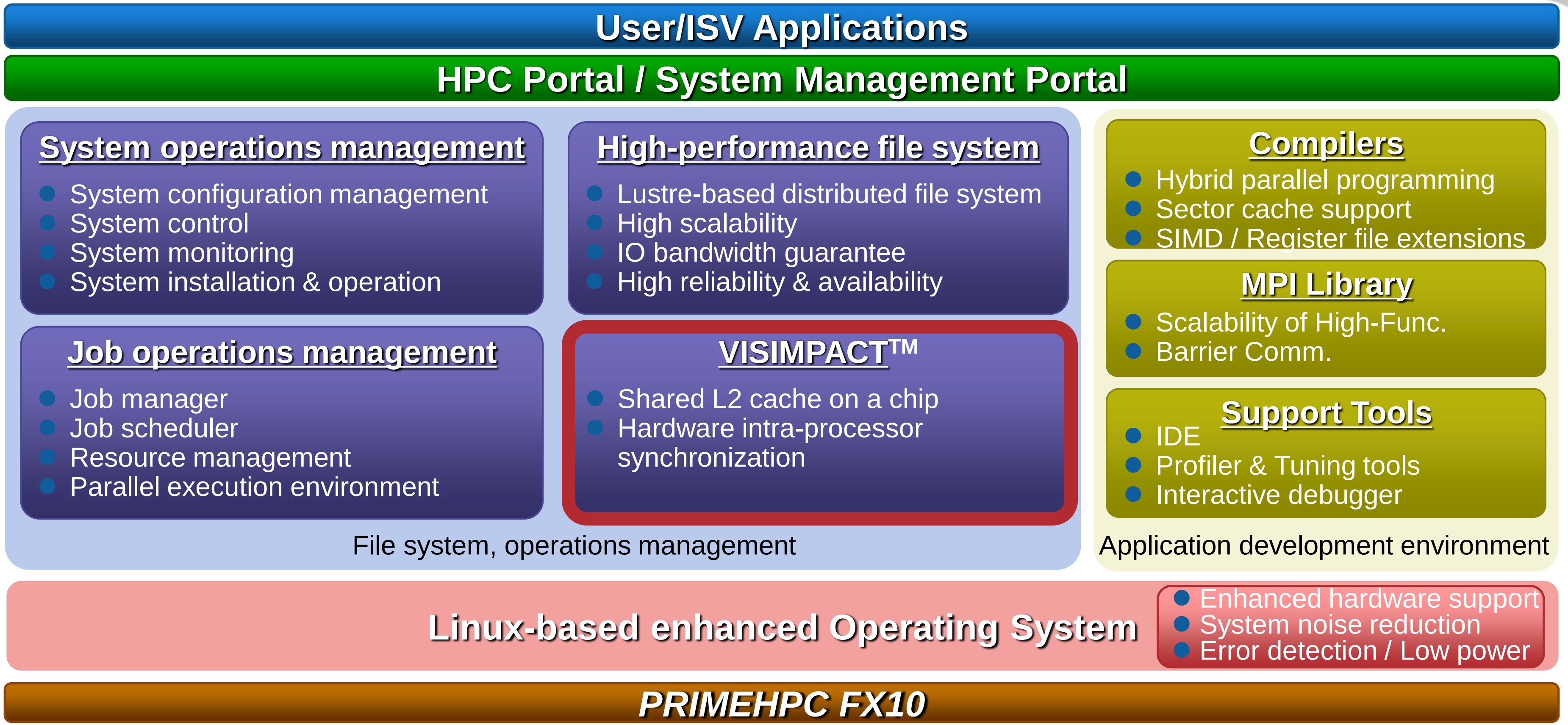


High Reliability and High Availability

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



System Software Stack



User/ISV Applications

HPC Portal / System Management Portal

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

Job operations management

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

VISIMPACT™

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

File system, operations management

Compilers

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

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

Application development environment

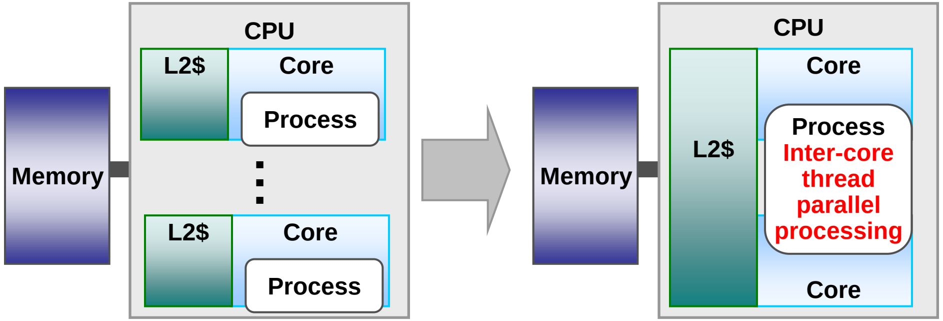
Linux-based enhanced Operating System

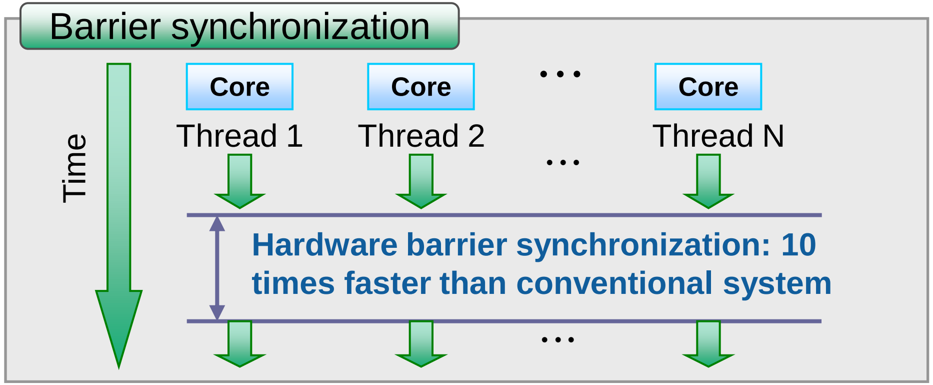
- Enhanced hardware support
- System noise reduction
- Error detection / Low power

PRIMEHPC FX10

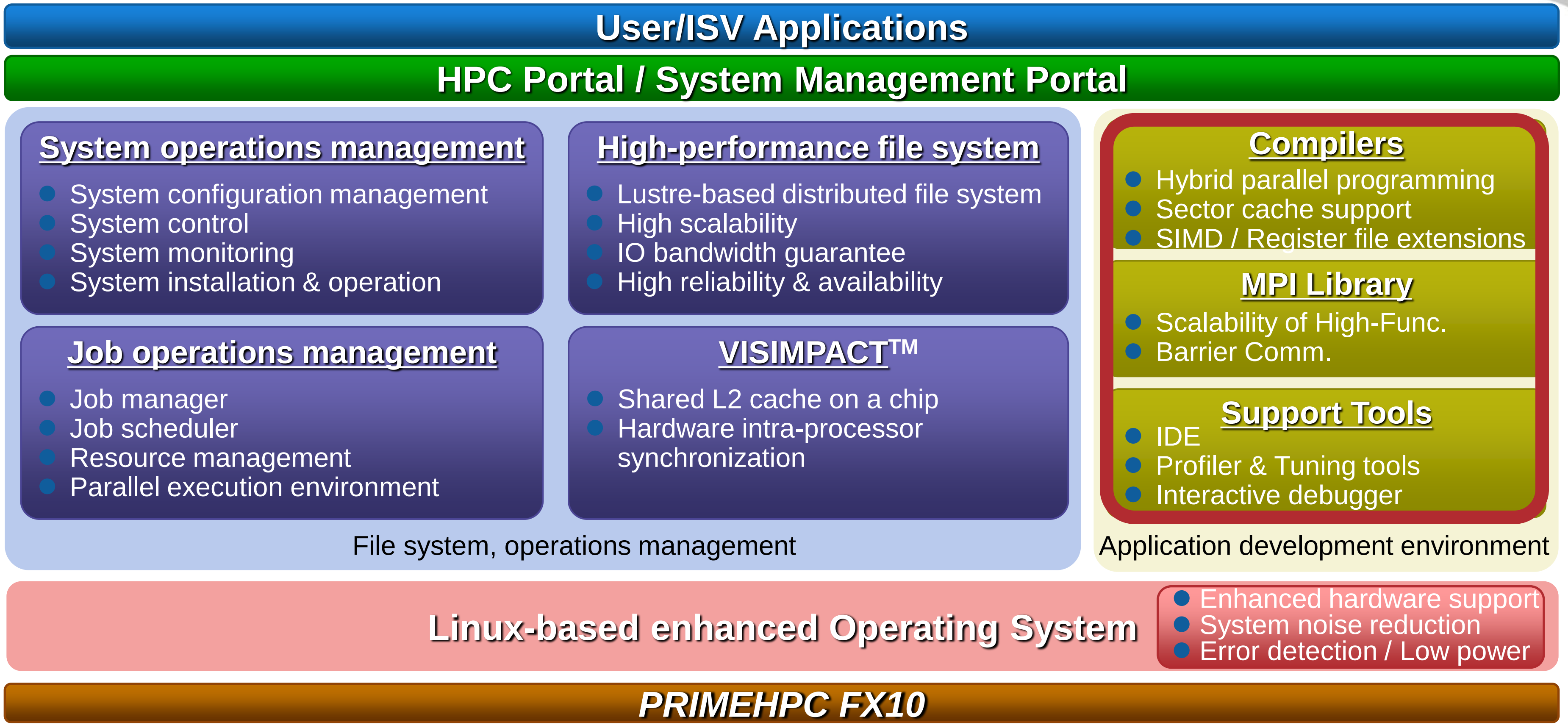
VISIMPACT™
(Virtual Single Processor by Integrated Multi-core Parallel Architecture)

- Mechanism that treats multiple cores as one high-speed CPU
  - Easy and efficient execution of inter-core thread parallel processing with a multi-core CPU
  - Supports the realization of a highly-efficient Hybrid model (Automatic parallelization + MPI)
- CPU technologies
  - Large-capacity shared L2 cache memory decrease in the influence of false sharing
  - Inter-core hardware barrier facilities 6-10 times faster than conventional software barrier




Barrier synchronization
Time
Core
Thread 1
Core
Thread 2
Core
Thread N
…
Hardware barrier synchronization: 10 times faster than conventional system

System Software Stack



User/ISV Applications

HPC Portal / System Management Portal

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

Job operations management

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

VISIMPACT™

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

File system, operations management

Compilers

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

MPI Library

- Scalability of High-Func.
- Barrier Comm.

Support Tools

- IDE
- Profiler & Tuning tools
- Interactive debugger

Application development environment

Linux-based enhanced Operating System

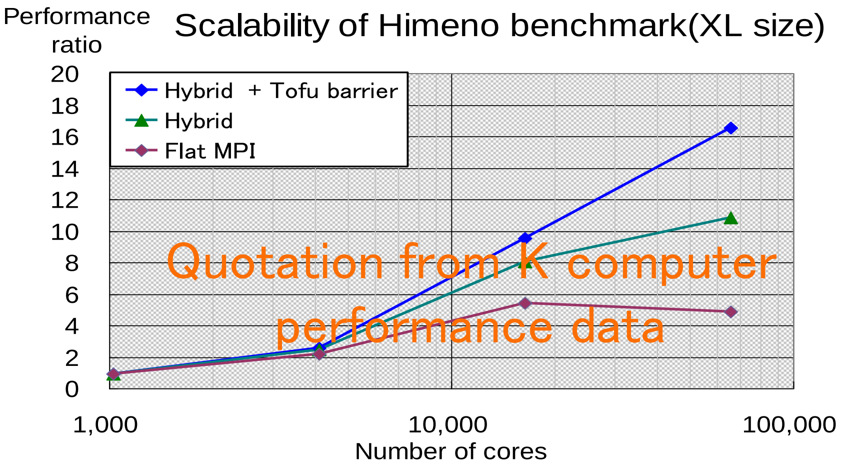
- Enhanced hardware support
- System noise reduction
- Error detection / Low power

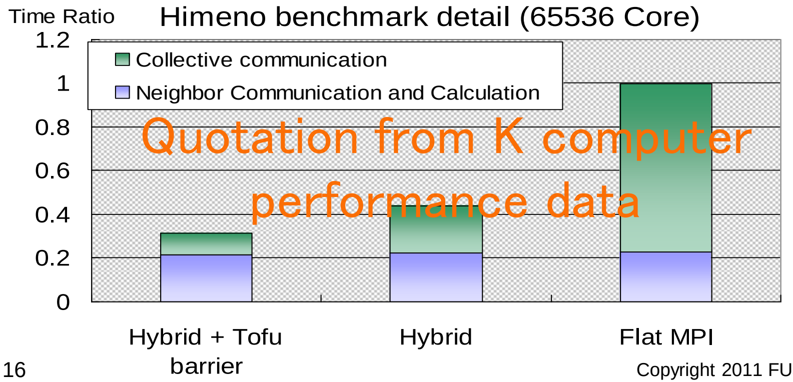
PRIMEHPC FX10

Programming Model for High Scalability

Hybrid parallelism by VISIMPACT and MPI library

- VISIMPACT
  - Automated multi-thread parallelization
  - High performance thread barrier used Inter-core hardware barrier facility
- MPI library
  - High performance collective communications used Tofu barrier facility





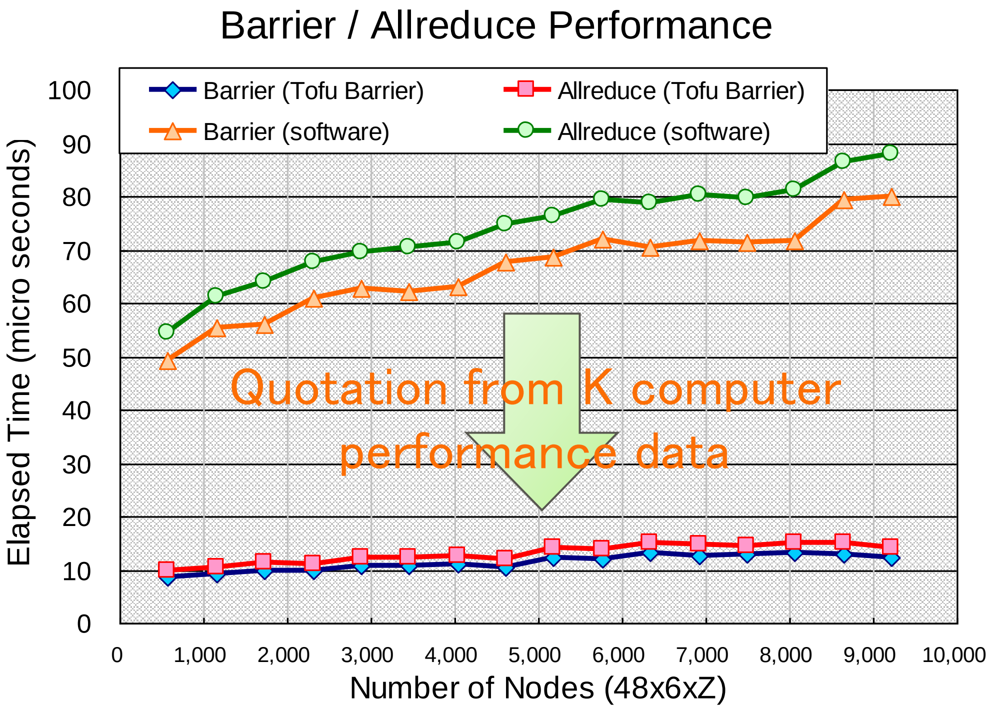
Customized MPI Library for High Scalability

Point-to-Point communication

- Use a special type of low-latency path that bypasses the software layer
- The transfer method optimization according to the data length, process location and number of hops

Collective communication

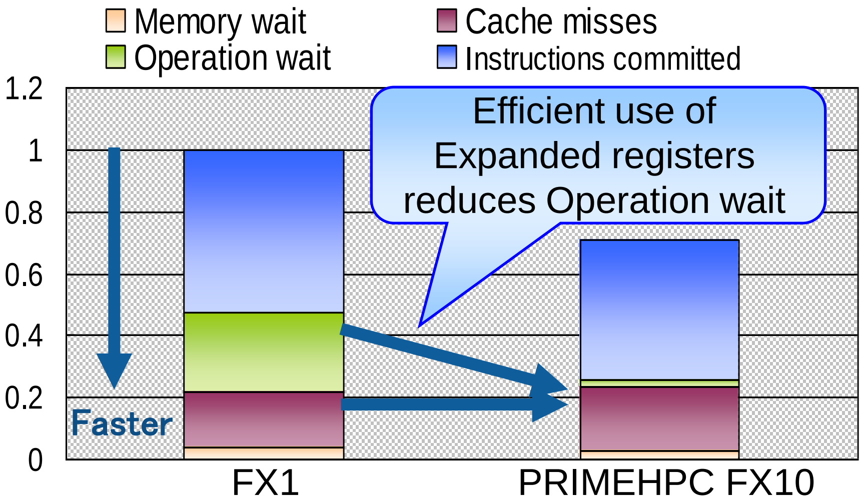
- High performance Barrier, Allreduce, Bcast and Reduce used Tofu barrier facility
- Scalable Bcast, Allgather, Allgatherv, Allreduce and Alltoall algorithm optimized for Tofu network



Compiler Optimization for High Performance

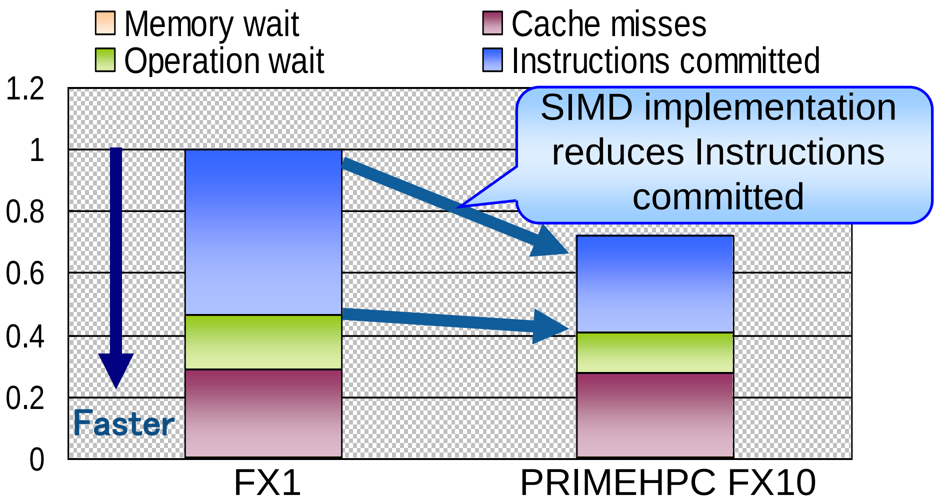
- Instruction-level parallelism with SIMD instructions
- Improvement of computing efficiency used Expanded registers
- Improvement of cache efficiency used Sector cache

NPB3.3 LU
Execution time comparison (relative values)



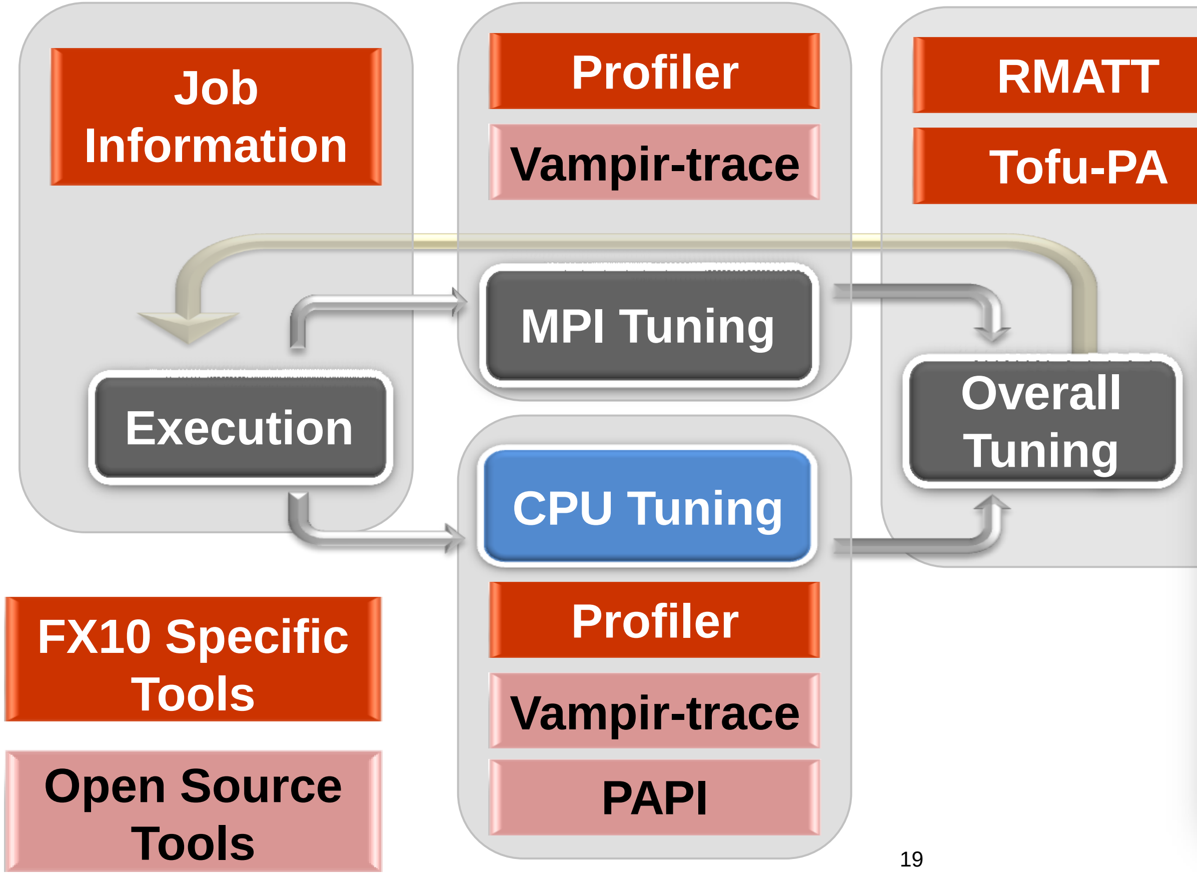
Efficient use of Expanded registers reduces Operation wait

NPB3.3 MG
Execution time comparison (relative values)

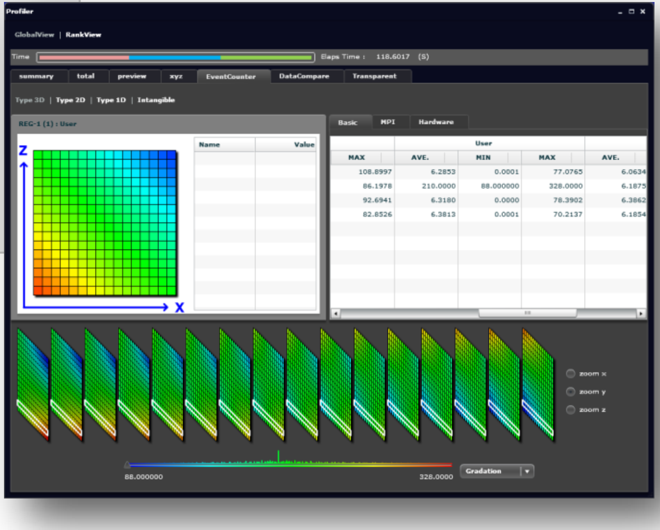


SIMD implementation reduces Instructions committed

Application Tuning Cycle and Tools



Profiler snapshot



Elaps Time : 118.6017 (S)

Name	Value

	User			
MAX	AVE.	MIN	MAX	AVE.
108.8997	6.2853	0.0001	77.0765	6.0634
86.1978	210.0000	88.000000	328.0000	6.1875
92.6941	6.3180	0.0000	78.3902	6.3862
82.8526	6.3813	0.0001	70.2137	6.1854

FUJITSU

shaping tomorrow with you