



# Fujitsu's vision for High Performance Computing

October 31, 2006

Motoi Okuda

General Manager

Computational Science and Engineering Center

# Agenda

- Fujitsu's HPC Solution Offerings
  - Platform
  - Software
- Challenges of Petascale Computing
  - Petascale Interconnect Project
  - Contributions to Japan's Next Generation Supercomputing Project
  - HPC Product Roadmap
- Summary

# Fujitsu's HPC Solution Offerings

**Fujitsu provides an integrated HPC environment based on leading-edge technologies**

## User applications

## User programs / ISV software

## Infrastructure Software

- HPC portal solution
- High performance file system solution
- Management portal solution
- Language system solution



## Hardware Platforms

- Cluster solutions (Linux)
- Large-scale SMP system solution (Linux / Solaris)
- FPGA solution



## Semiconductor and System Design Technology

- Multi-core & Highly reliable SPARC CPU
- Leading-edge custom circuit design for chip set



# Fujitsu Hardware Platforms for HPC

## Cluster Solutions

- Optimal price/performance for MPI-based applications
- Highly scalable
- EM64T/Opteron-based (2 sockets)
- InfiniBand interconnect

## FPGA Solutions

- Ultra high performance for specific applications



**PRIMERGY**

BX Series

RX Series EM64T Opteron

RXI Series Itanium<sup>®</sup> 2

IA/Linux

## Large-scale SMP System Solutions

- Up to 1TB memory space for HPC applications
- High I/O bandwidth for I/O server
- High reliability based on main-frame technology
- High-end RISC MPU

**PRIMEQUEST**

PRIMEQUEST580  
Itanium<sup>®</sup> 2  
~32cpu

IA/Linux

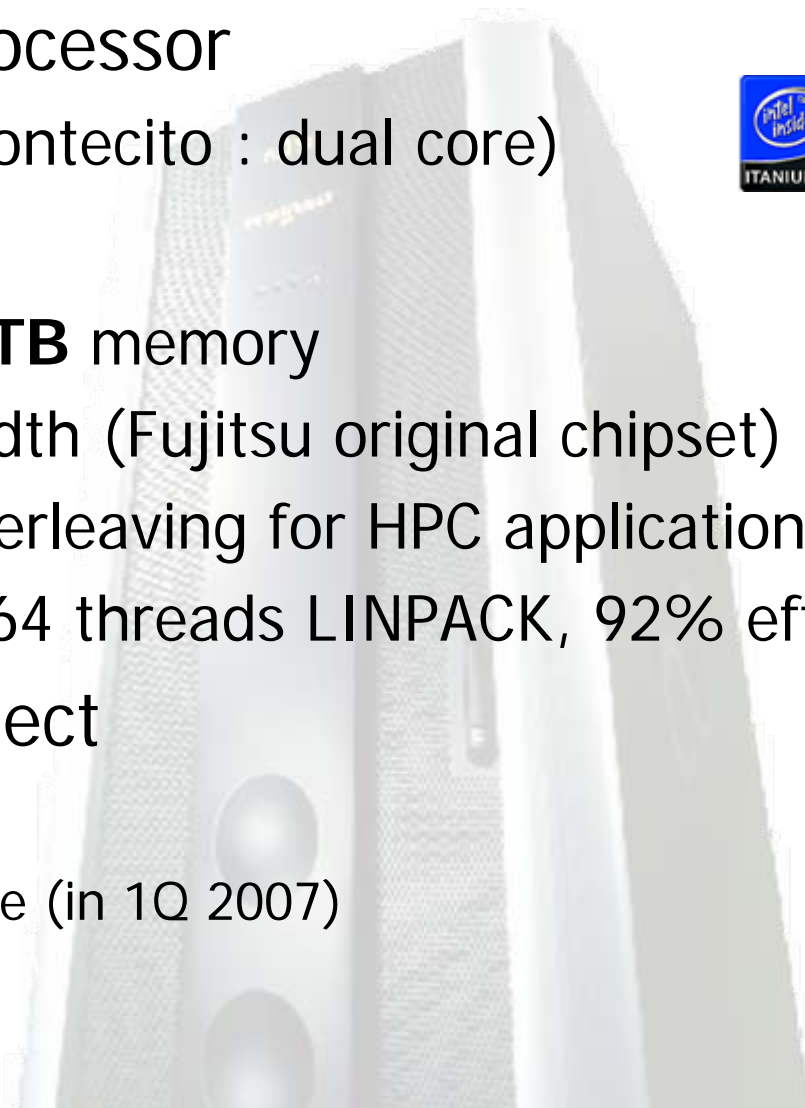
**PRIMEPOWER**

HPC2500  
SPARC64<sup>™</sup>  
~128cpu

SPARC/Solaris

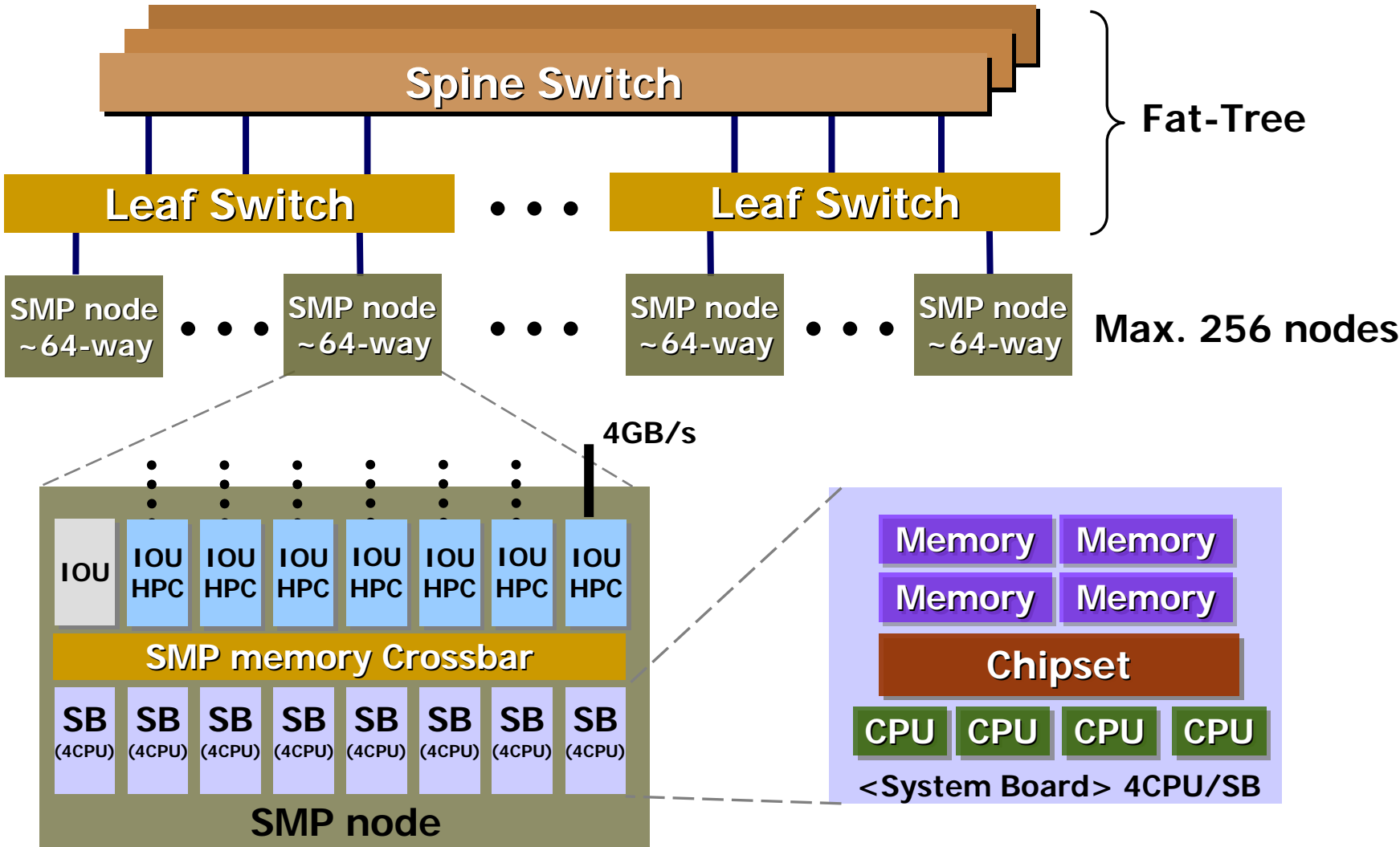
# Overview of PRIMEQUEST 580

- Latest Itanium<sup>®</sup> 2 processor
  - 1.6GHz/24MB L3\$ (Montecito : dual core)
- Large-scale true SMP
  - **64-way SMP** with **1TB** memory
  - High memory bandwidth (Fujitsu original chipset)
  - Extended memory interleaving for HPC applications
  - 375.5 GFlops /node (64 threads LINPACK, 92% efficiency)
- High speed interconnect
  - High bandwidth
    - ◆ Max. **28GB/s** per node (in 1Q 2007)
    - ◆ Fat-Tree topology
  - Scalability
    - ◆ Nodes can be connected up to **256 nodes** (in 1Q 2007)



# PRIMEQUEST - Multiple-node Configuration -

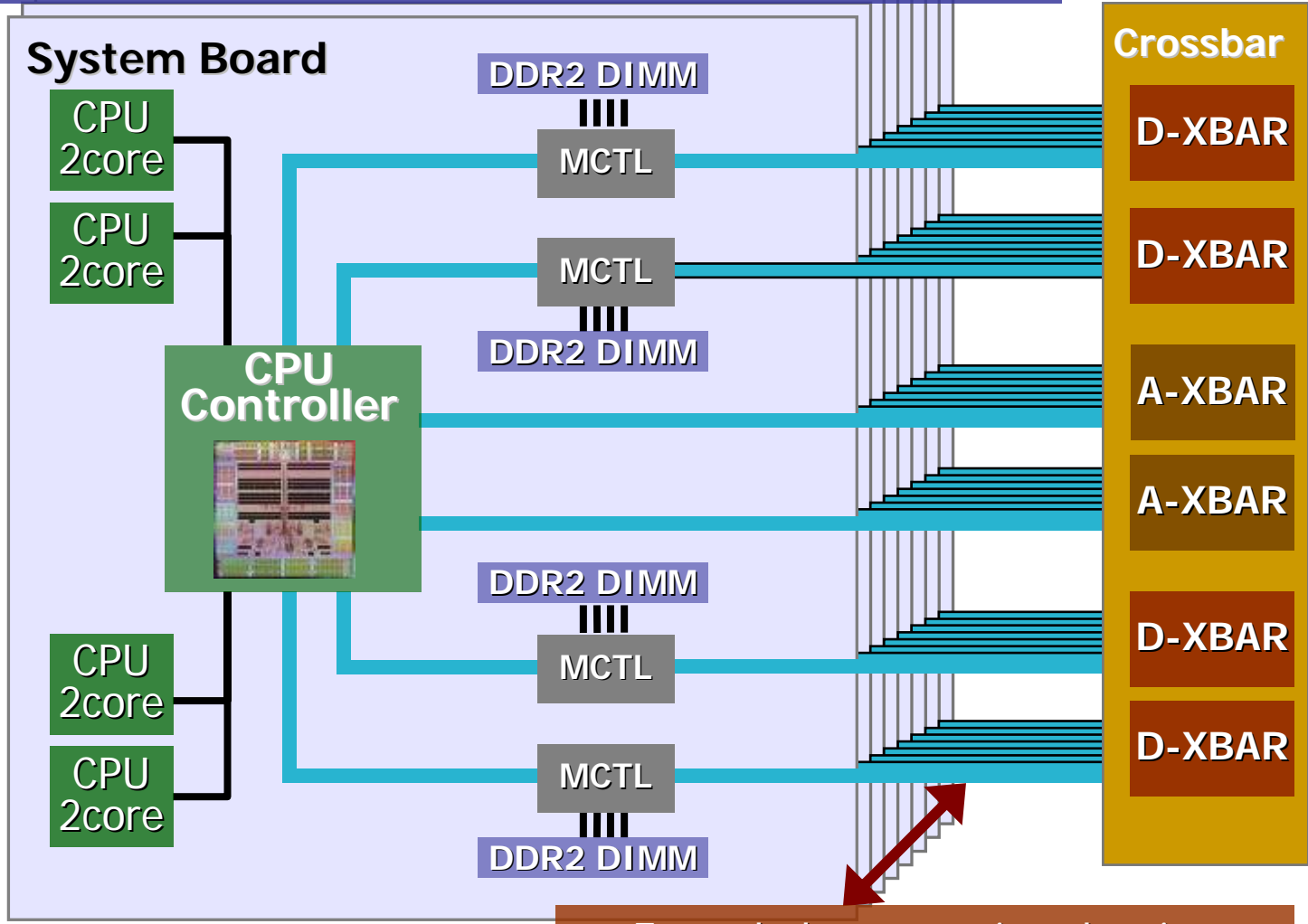
Up to 256 nodes through high speed interconnection



# PRIMEQEUST - System Block Diagram -

Fujitsu original chipset achieves high memory BW and low latency – a true SMP

- 4 CPU chips per System Board
- 32 DDR2 DIMMs per System Board
- High speed memory crossbar
- Extended memory interleaving
  - Assigning address extended to System Board
- Ultra-high speed synchronized bus

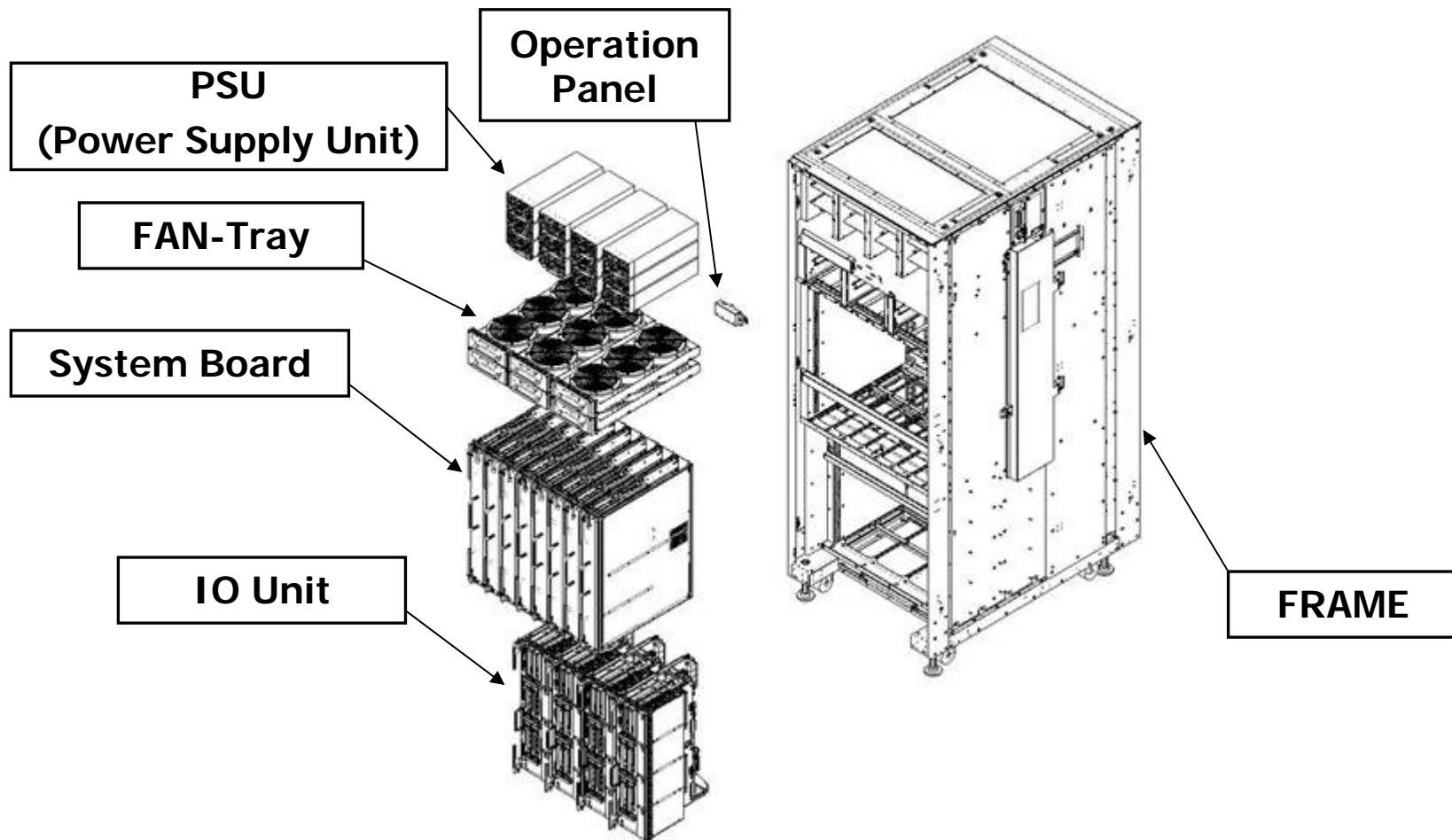


MCTL: Memory Controller  
 D-XBAR: Data Crossbar  
 A-XBAR: Address Crossbar

Extended memory interleaving

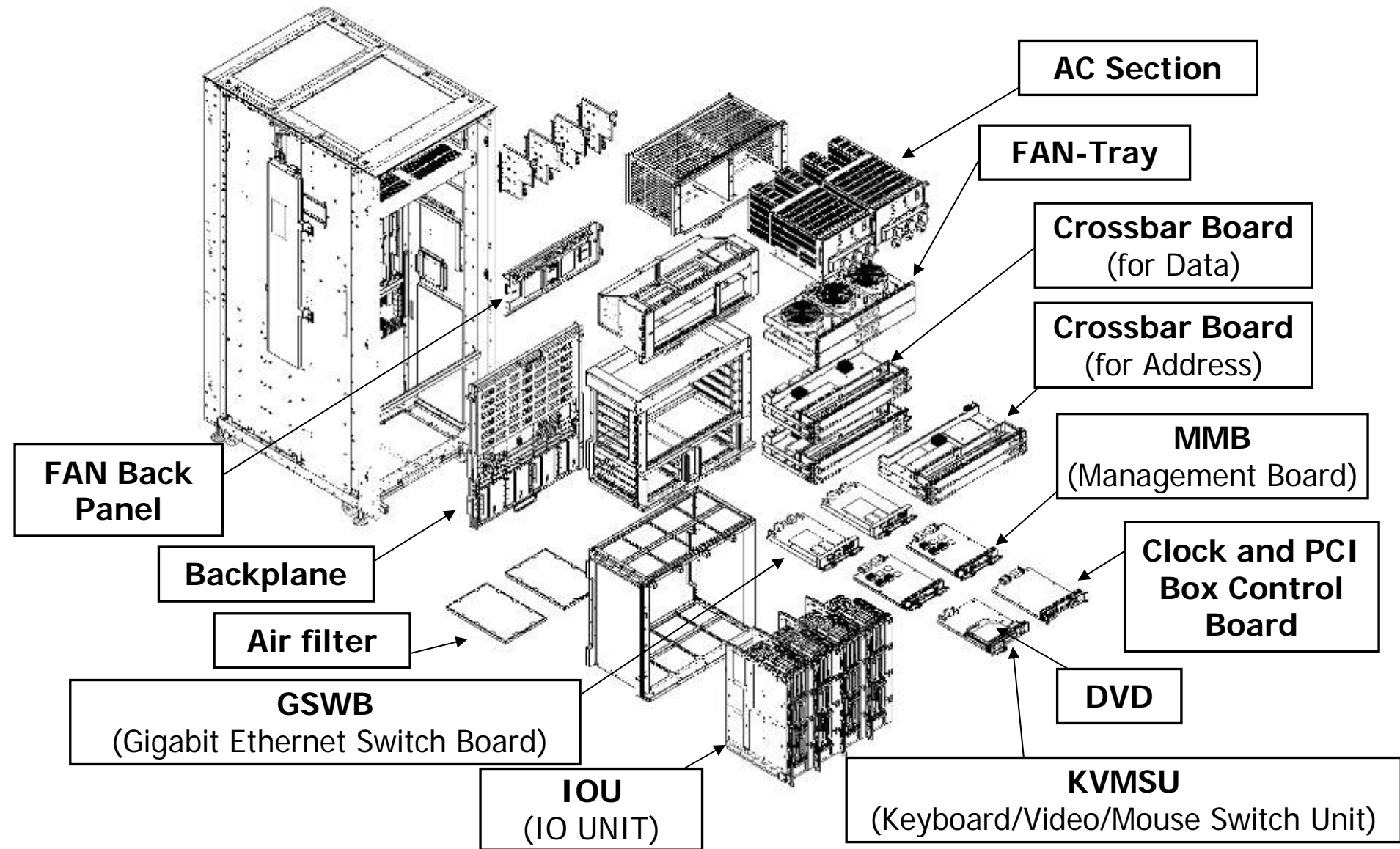
# PRIMEQUEST - Chassis (Front Side) -

Cable-less and Modular design  
for High Reliability and High Operability

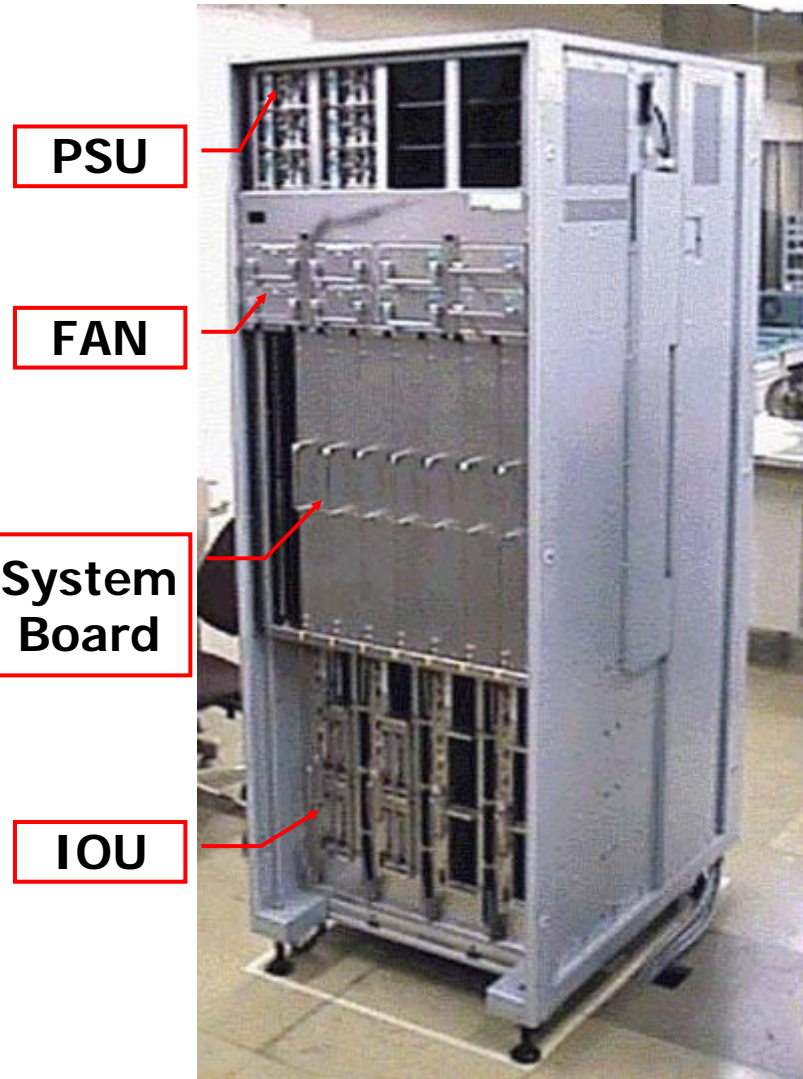




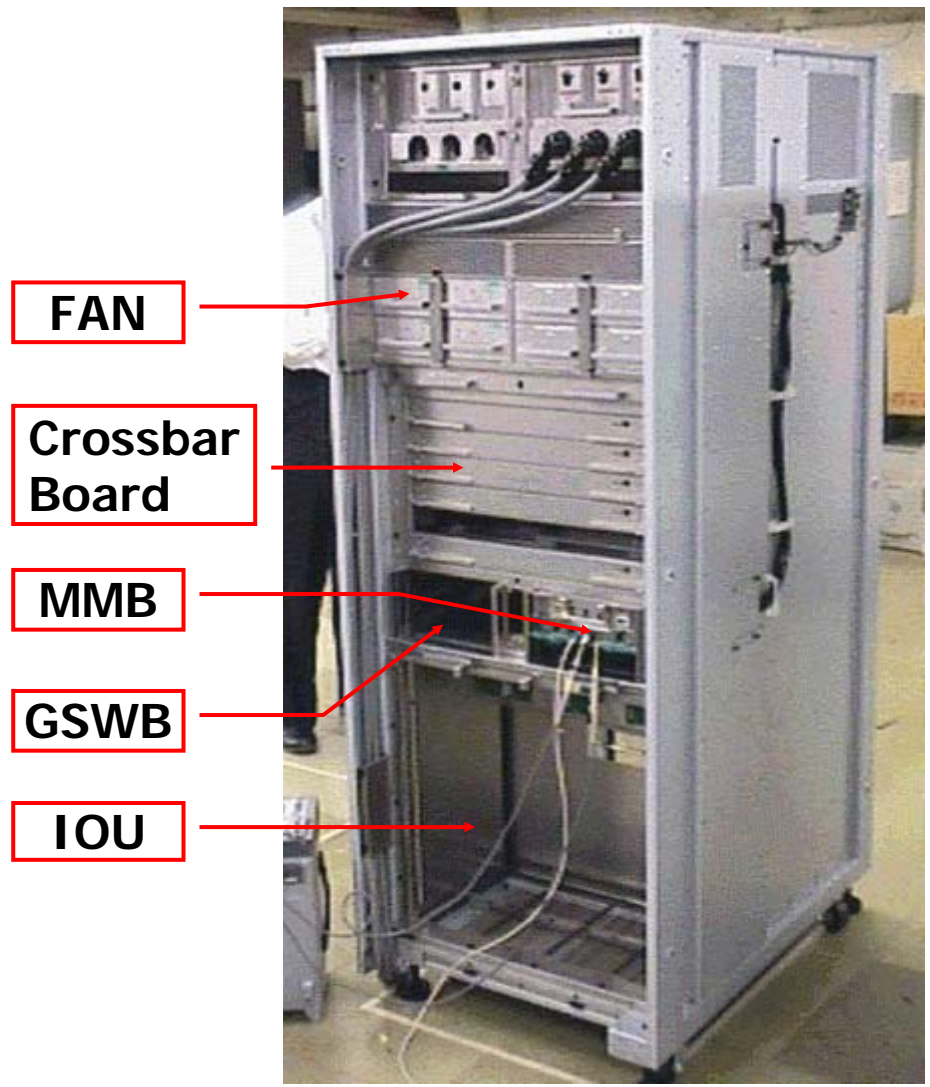
# PRIMEQUEST - Chassis (Rear Side) -



# PRIMEQUEST – Chassis photo -



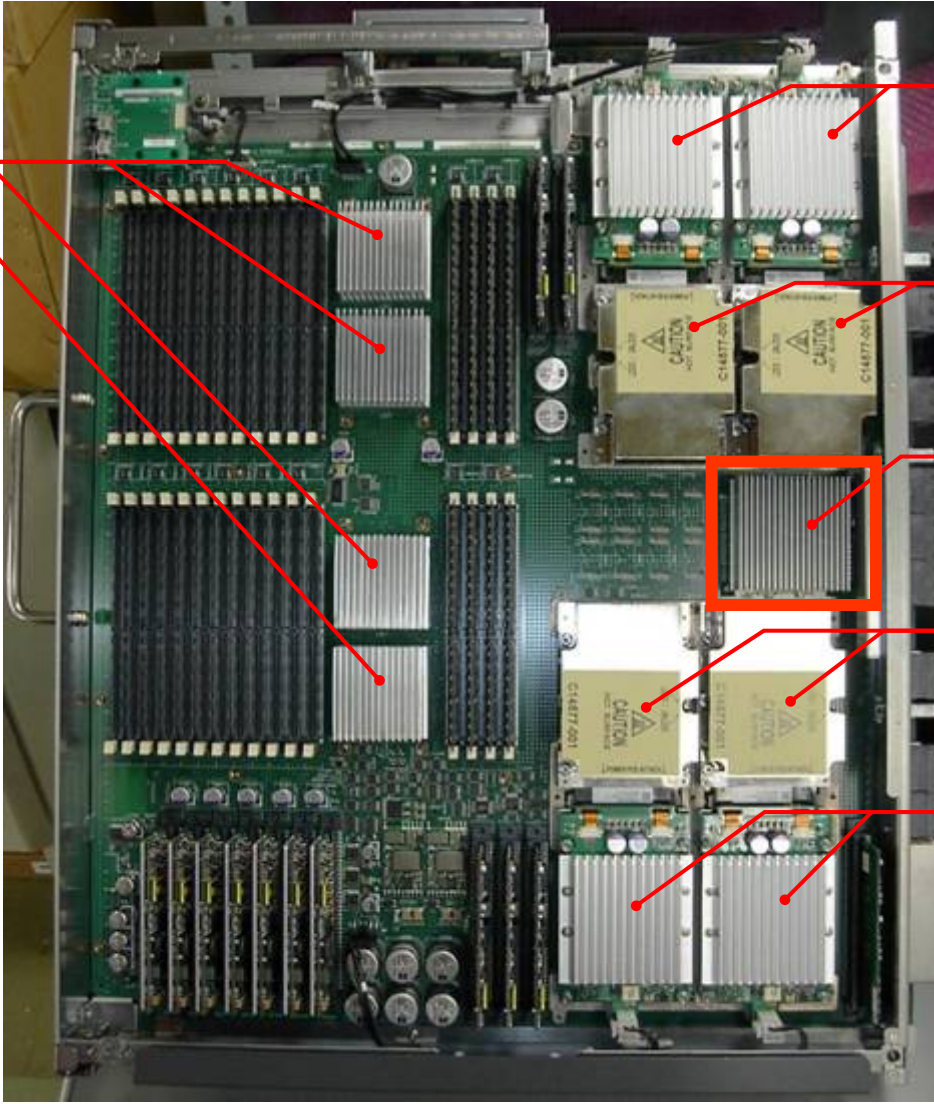
Front



Rear

# PRIMEQUEST - System Board Photo-

Memory Controller



PowerPod

CPU

CPU Controller

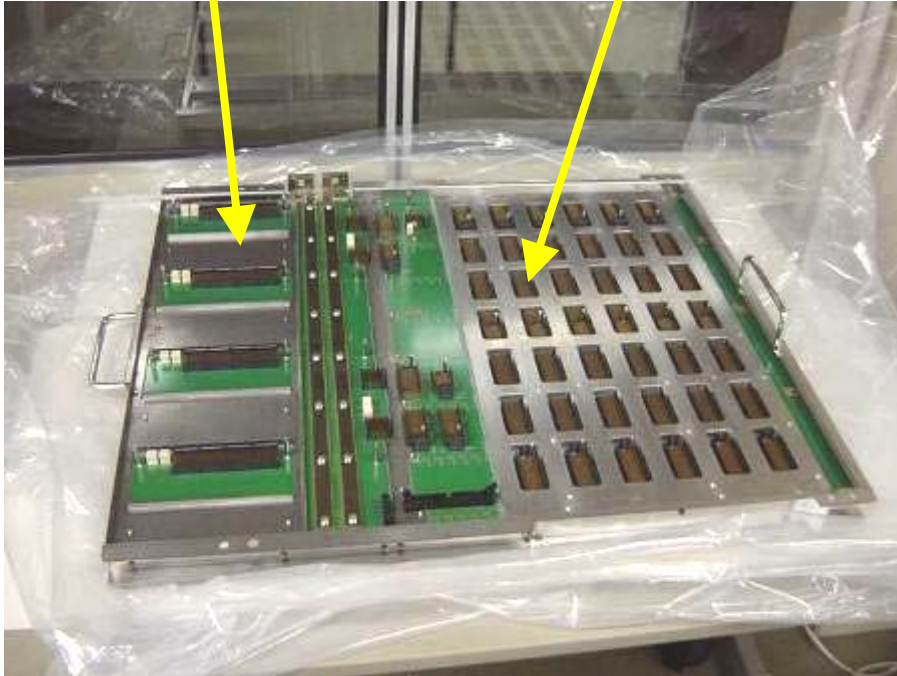
CPU

PowerPod

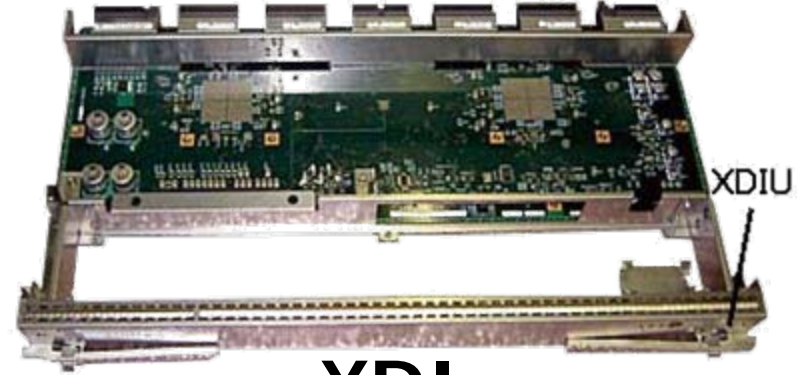
# PRIMEQUEST – Backplane & XDI/XAI Photo-

Connectors for 4 IOUs

Connectors for 4 XDIs and 2 XAIs

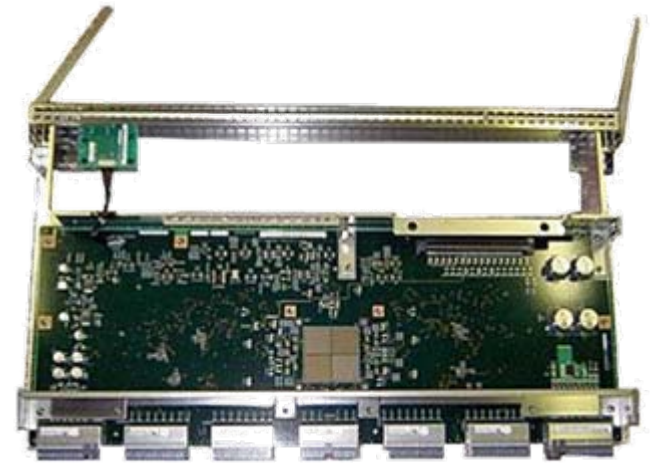


Backplane (rear side)



**XDI**

(Data Crossbar Interface)



**XAI**

(Address Crossbar Interface)

Top

# PRIMEQUEST latest installation

- Institute for Molecular Science -

**Login/TSS Server**  
PRIMEQUEST x 2

4 CPU cores  
16GB of mem.  
0.5TB of disk



**Compute Server (4TFlops)**  
PRIMEQUEST580 x 10 nodes with High Speed Interconnect

64 CPU cores  
256MB of mem.  
3.5TB of disk



**640 cores of Itanium2 (Montecito)**

InfiniBand™  
Switch

...x16 SWs

1GB/s x 16

GbE x2



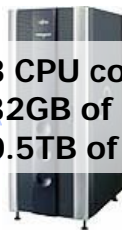
**File Server**

PRIMEQUEST with  
two ETERNUS3000 m700

1  
of s



8 CPU cores  
32GB of mem.  
0.5TB of disk



# HPC Software Structure

## User/ISV Applications

### HPC Portal / System Mgmt. Portal

#### Job/System Management (Parallelnavi)

##### Job Scheduler

- Parallel Job execution
- Fair share schedule
- Job Accounting

##### HPC enhancement

- CPU management
- Large page
- High speed interconnect

##### HPC Cluster management

- System configuration Mgr.
- Power/IPL management
- High speed interconnect

#### File System

##### SRFS

- High speed network file system

##### GFS/GDS

- SAN global file system
- Large file system

#### Language System (Parallelnavi)

##### Compiler

- Fortran
- C/C++
- XPFortran

##### Parallel Programming

- Auto-Parallelization
- OpenMP
- MPI

##### Tools/Libraries

- Programming Tools
- Scientific Library (SSL II/BLAS etc.)



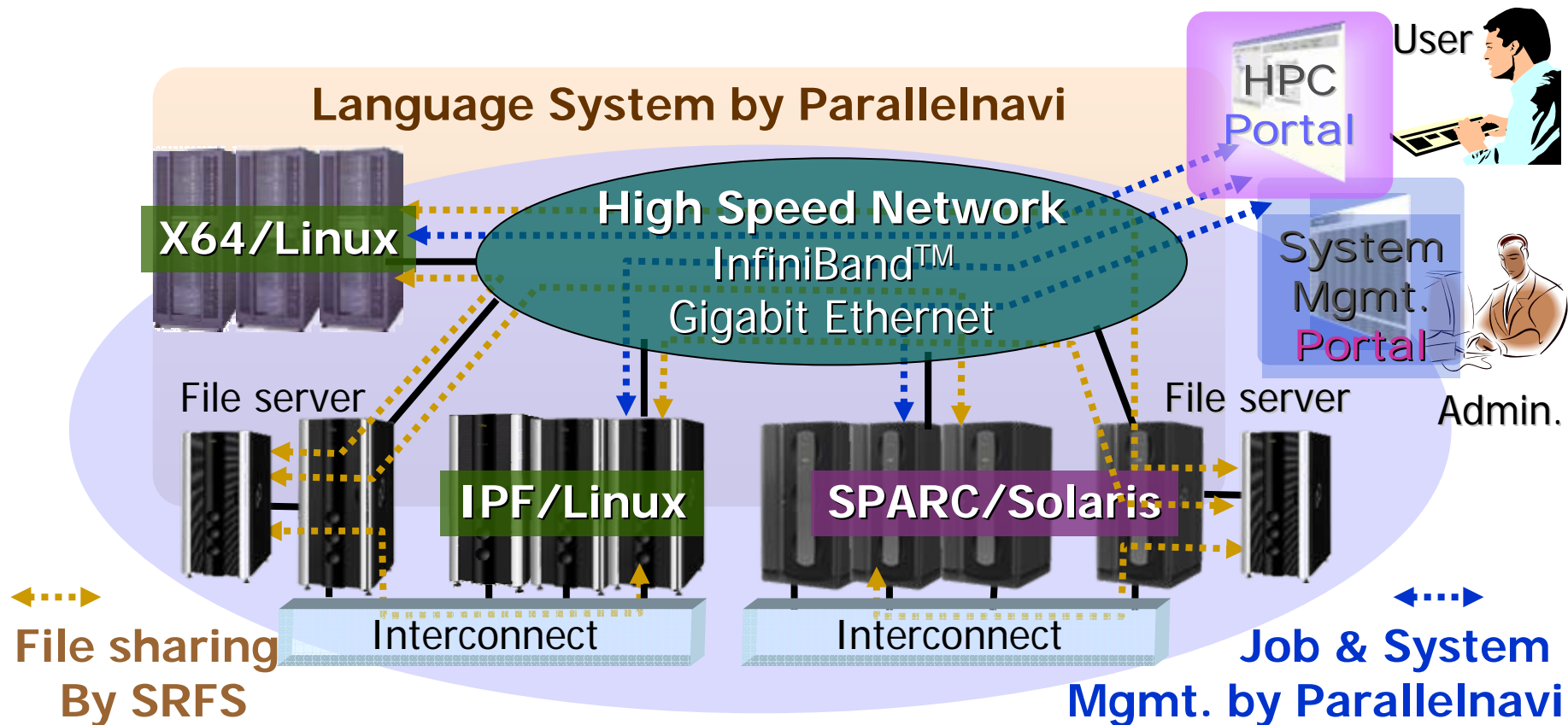
Solaris / Red Hat Enterprise Linux v.4



PRIMEPOWER / PRIMEQUEST / PRIMERGY

# Integrated HPC Environment

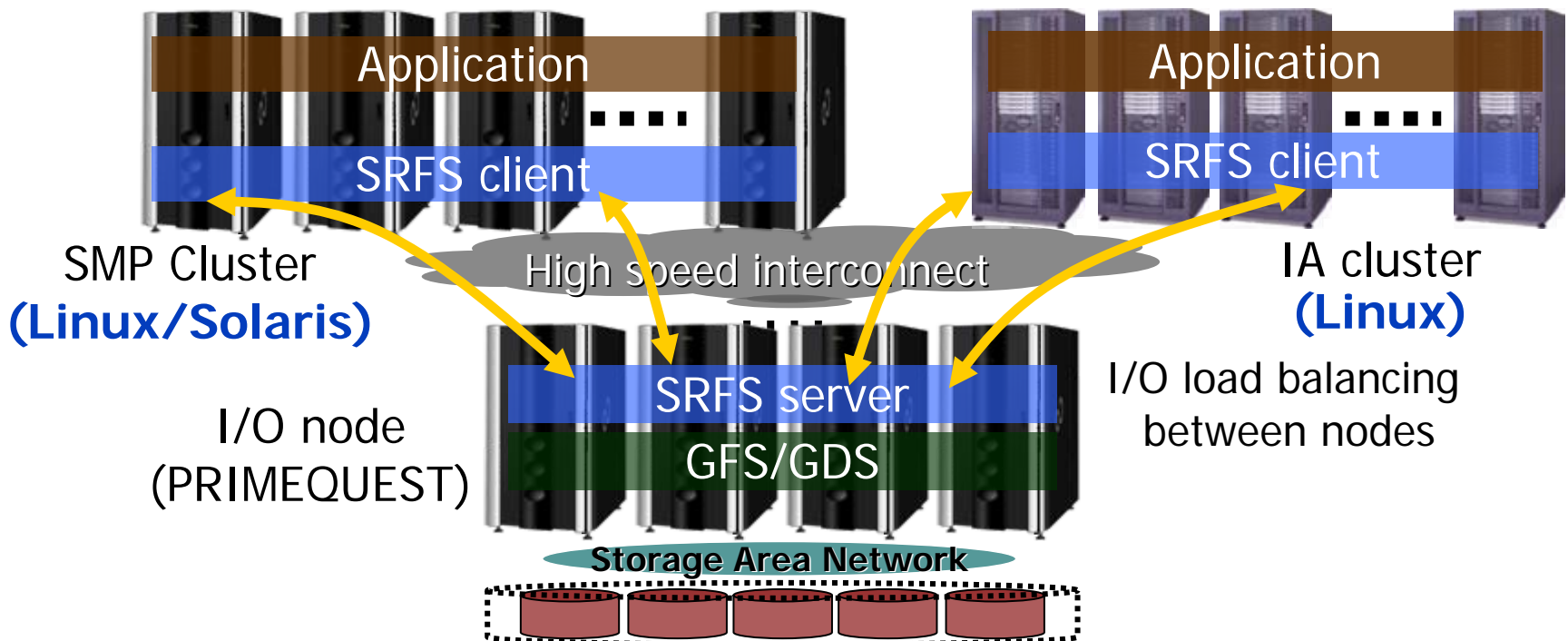
- Use heterogeneous platform resources via HPC Portal
- Monitor and manage resources via System Management Portal
- High performance network file sharing via high speed network
- High performance language system for SPARC, X64, and IPF



# Shared Rapid File System (SRFS)

Huge file sharing among huge heterogeneous clusters

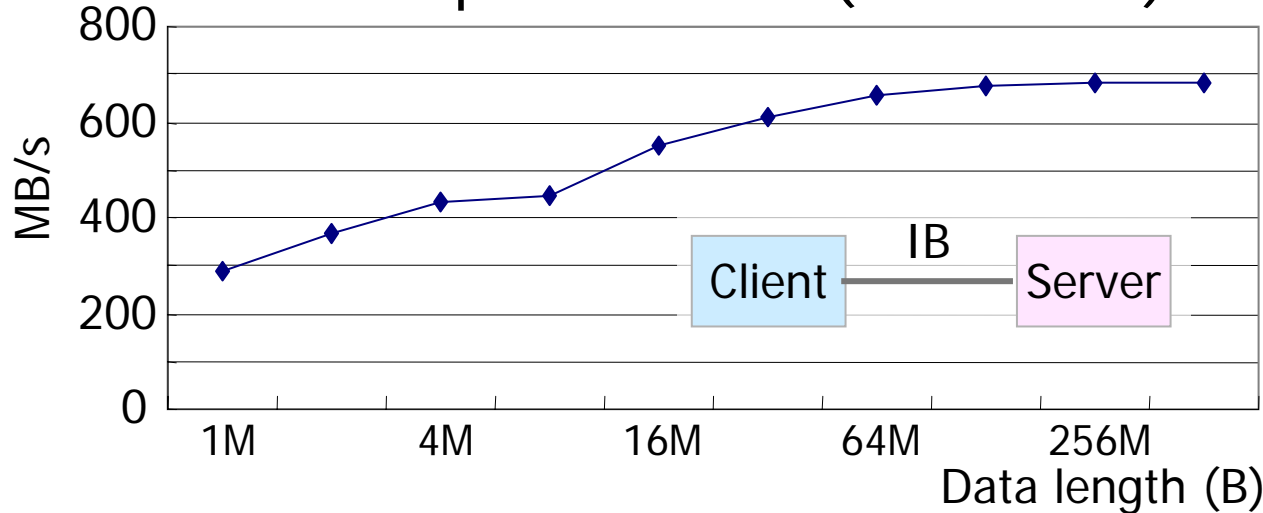
- Available from existing applications (POSIX standard file access API)
  - High performance: High speed interconnect and huge file cache
  - High Availability: Redundant interconnect/IO nodes
  - File Consistency: Keeps file consistency between clients
  - Huge Capacity: Extend client nodes by multiple IO nodes sharing SAN disk using GFS





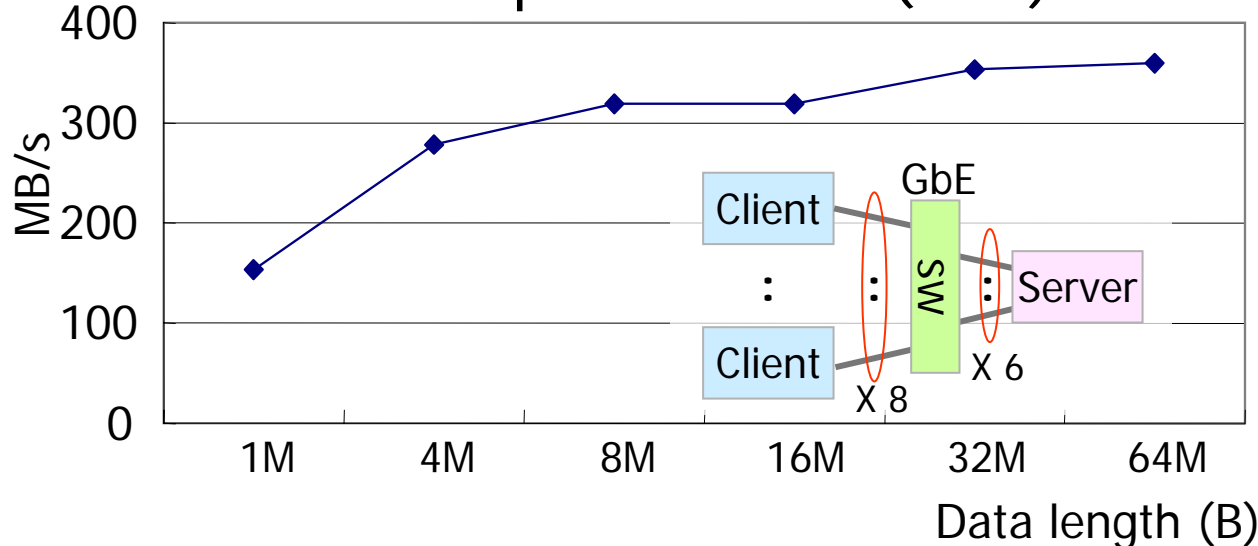
# Performance Example of SRFS + GFS

## Read performance (InfiniBand)



- **Client node**
  - PG RX200S3 x 1
- **Server node**
  - PQ580
- **Interconnect**
  - InfiniBand X 1
- **Disk**
  - ETERNUS8000M-90:4GbFC X 8

## Read performance (GbE)

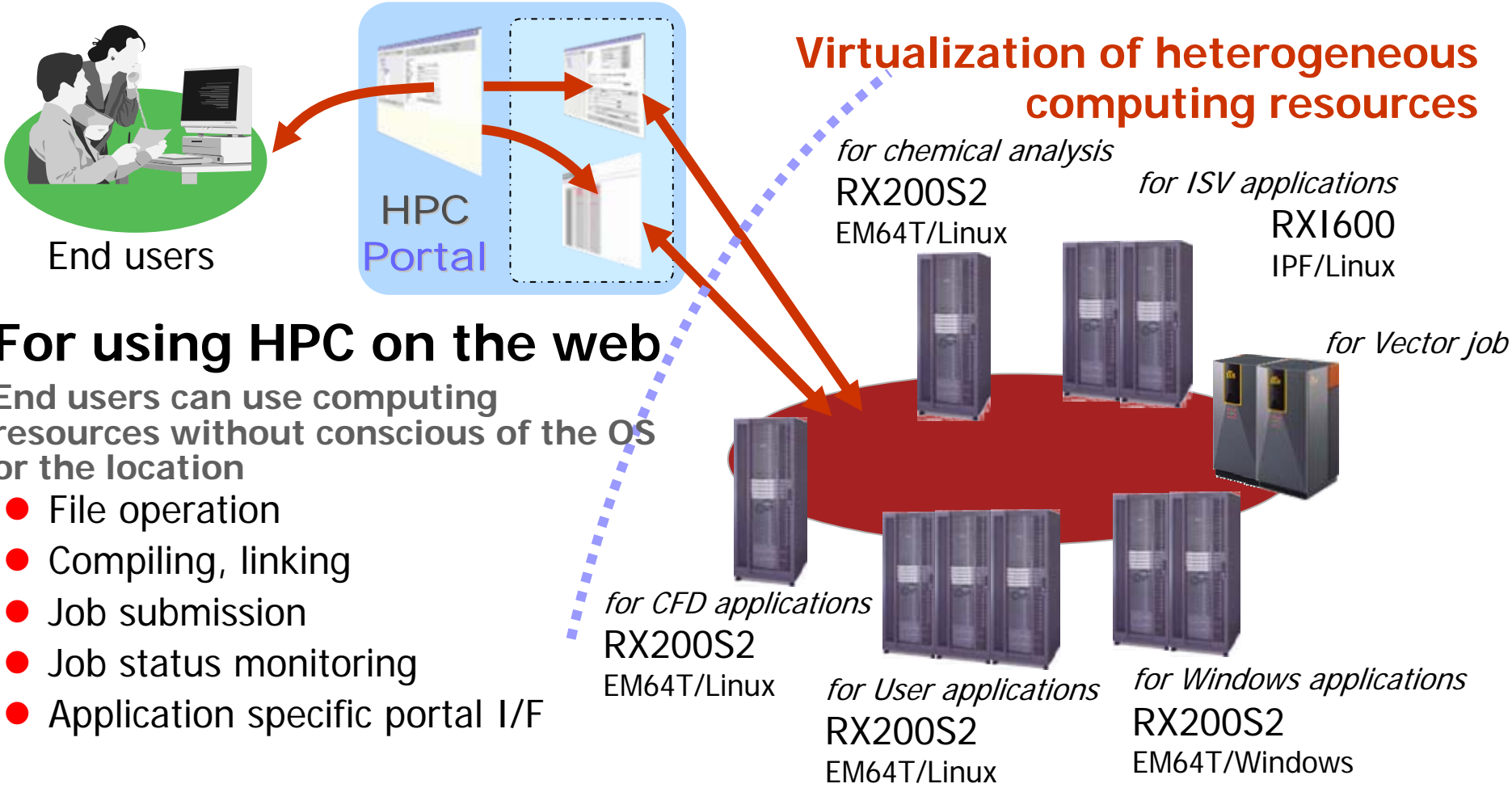


- **Client node**
  - PG RX200S2 X 8
- **Server node**
  - PQ480
- **Interconnect**
  - Gigabit-E X 6
- **Disk**
  - ETERNUS4000 : 2GbFC X 4

# HPC Portal

## Unified operation of heterogeneous resources

An Electrical Equipment Company: *"Simulation Cockpit"*



### For using HPC on the web

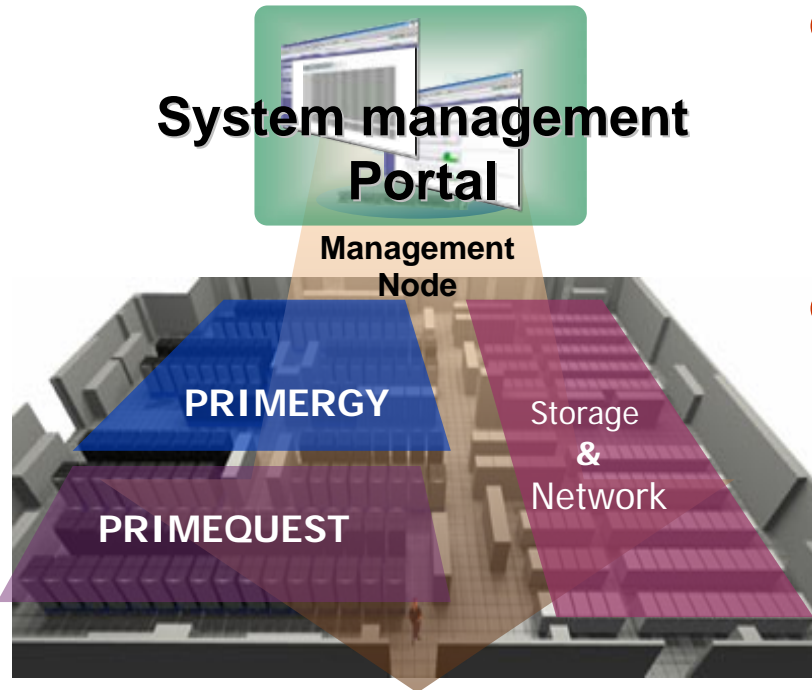
End users can use computing resources without conscious of the OS or the location

- File operation
- Compiling, linking
- Job submission
- Job status monitoring
- Application specific portal I/F

# System Management Portal

Single system image management for HPC platform enables simplified system operation for huge clusters ⇒ **Reduces TCO**

A management node controls the whole system



System management portal provides easy web interface

- **Centralized cluster management for heterogeneous cluster**
- **Power, IPL control**
  - Simultaneous, grouping, hierarchy
  - Collect hardware/OS log from nodes
  - Dump instruction from manager
- **Assist auto-pilot**
  - Hardware failure monitoring
  - Software (daemon) status monitoring
  - Call centre routine on event
  - Cooperate with job scheduler (failure isolation)
- **Easy installation**
  - Centralized install

# Overview of Language System

- Highly scalable thread parallel performance based on Fujitsu's vectorization technology
- Hybrid parallel programming environment (combination of thread and message passing )

	Compiler/MPL	Tool	Math. Library
<b>Serial</b>	Fortran	<b>Parallelnavi Programming Tools *2</b>	SSL II, BLAS, LAPACK
	C		C-SSL II
	C++		
<b>Data Parallel</b>	Auto-parallel		SSL II BLAS, LAPACK
	OpenMP		
	XPFortran *1		SSL II/XPF *3
<b>MPL</b>	MPI		ScaLAPACK

\*1: eXtended Parallel Fortran (Distributed parallel language: includes VPP Fortran)

\*2: Integrated programming environment

\*3: Same functions as SSL II/VPP

# Agenda

- Fujitsu's HPC Solution Offerings
  - Platform
  - Software
- Challenges of Petascale Computing
  - Petascale Interconnect Project
  - Contributions to Japan's Next Generation Supercomputing Project
  - HPC Product Roadmap
- Summary

# Next Generation Supercomputer Project of MEXT\* Japan

## *Fujitsu's contributions*

2005

2006

2007

2008

2009

2010

2011

Fundamental R&D projects  
for Next Generation  
Supercomputer  
2005-2007(FY)

Fujitsu collaborates with Kyushu University on R&D work for the *PSI project*

- Opto-Electric hybrid interconnect
- Low cost MPI communication by intelligent switch and run time optimization
- Petascale system evaluation methodology

**Next Generation Supercomputer Project  
led by RIKEN**  
(expected total budget is about \$US 1 billion)

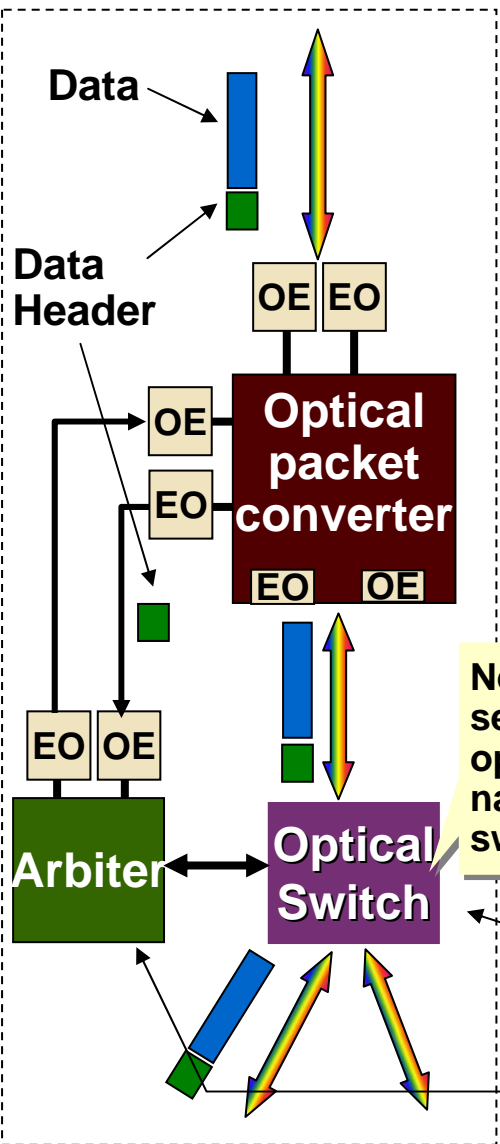
Fujitsu is a contributor in the  
*conceptual design of the  
Next Generation Supercomputer*

\*MEXT : Ministry of Education, Culture, Sport, Science and Technology

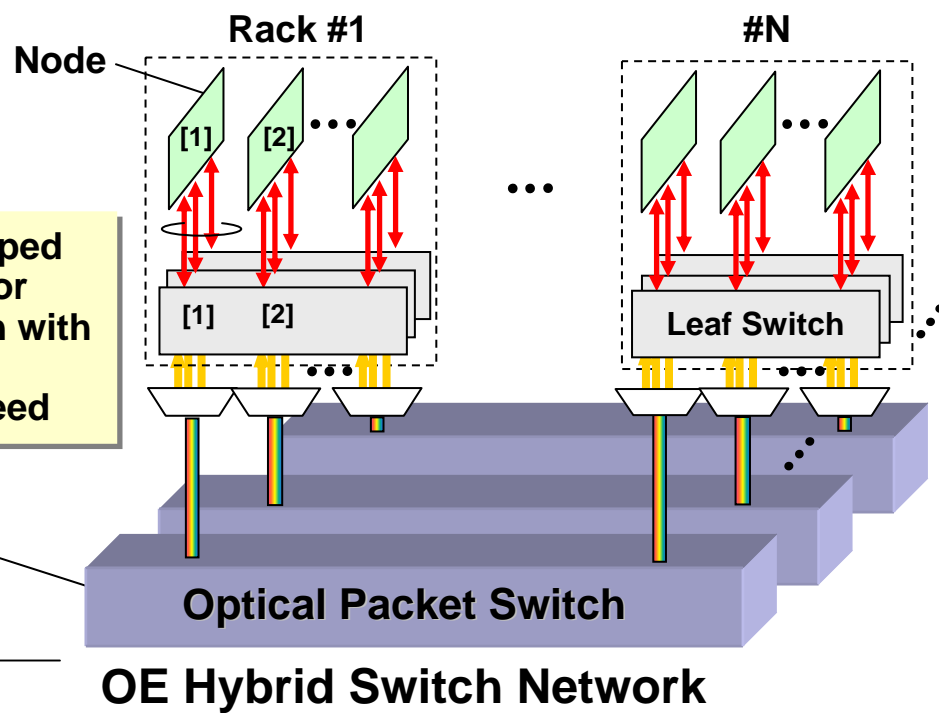
# PSI Project

## Ultra High Speed Optical-Electric Hybrid Interconnect

- Development of high capacity, high speed and compact interconnect based on an optical packet switch technology

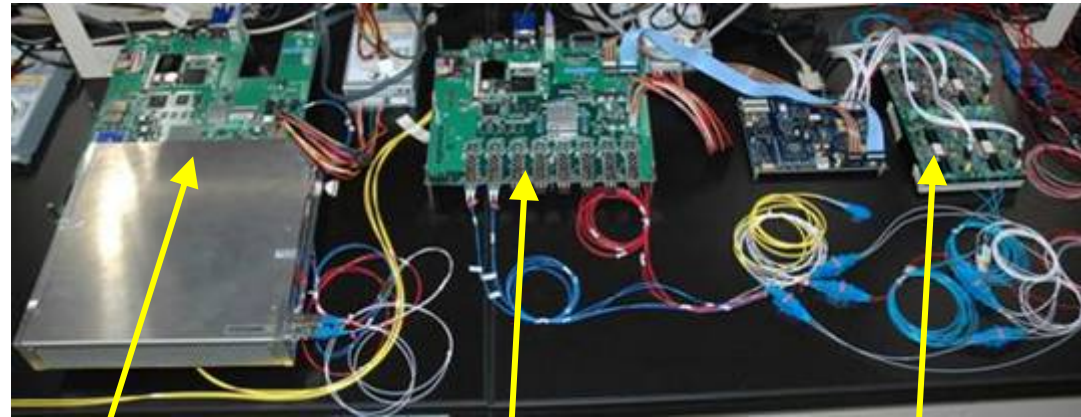
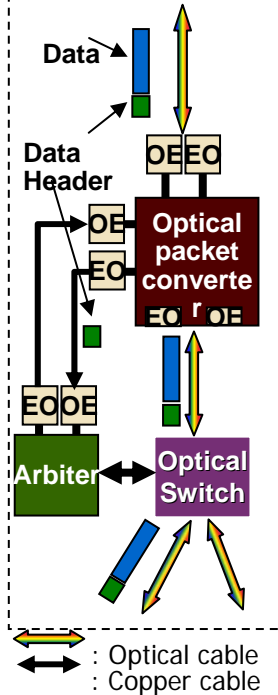


Newly developed semiconductor optical switch with nanoseconds switching speed



# PSI Project Switch

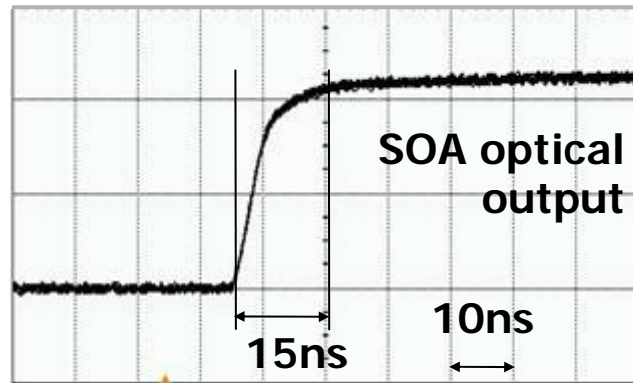
# 2x2 Optical Packet Prototype



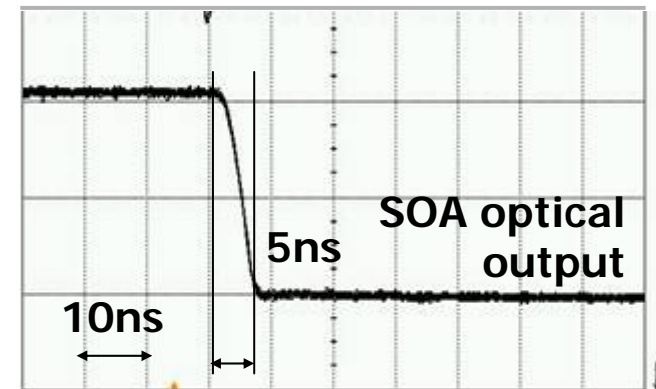
**Optical Packet Converter**

**Switch Arbitrer**

**2x2 SOA\* Optical Switch**



(a) Rise time



(b) Fall time

## Switching Speed

\*Semiconductor Optical Amplifier

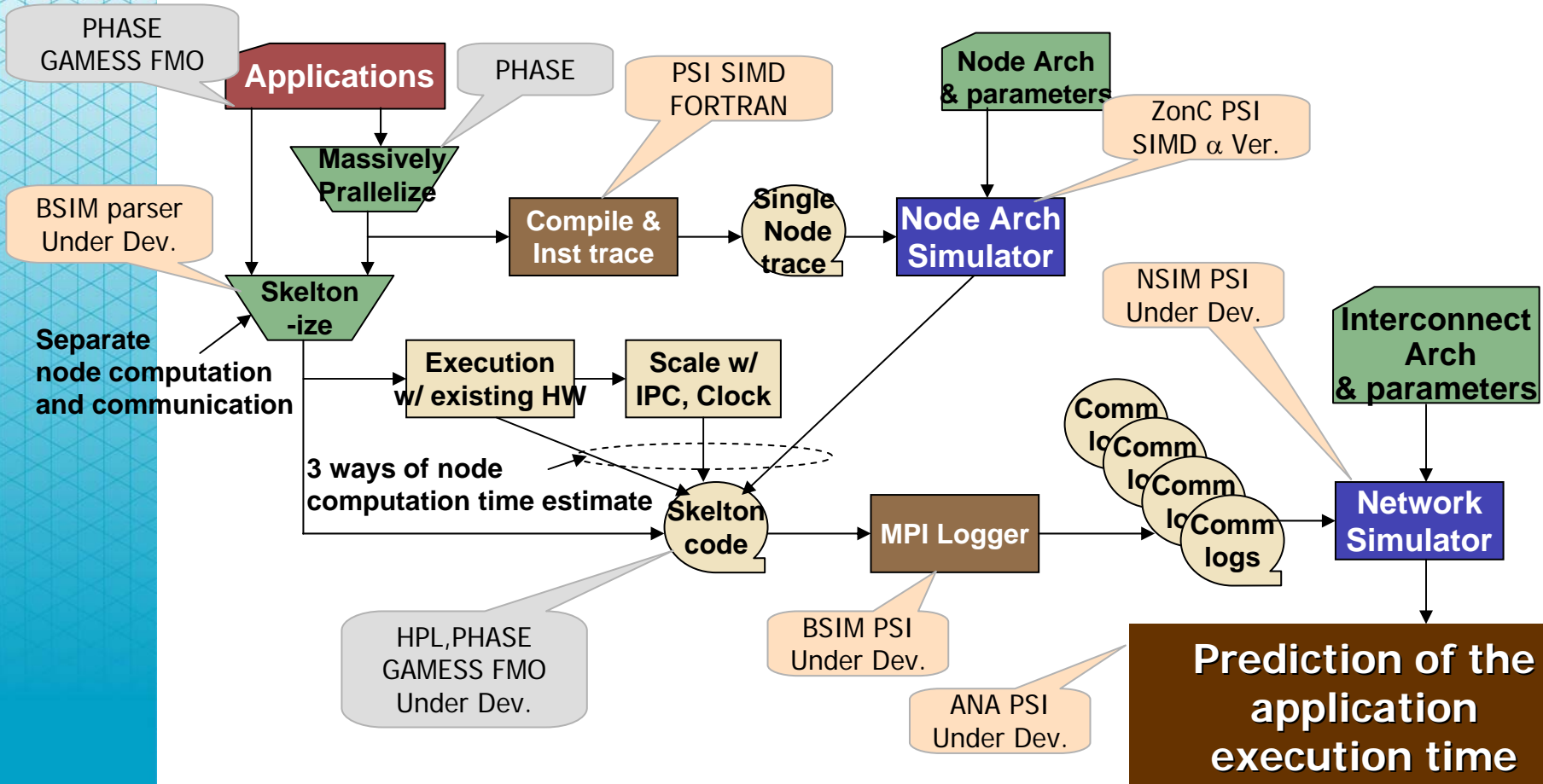
The development of high speed optical switch fabric is partly supported by the National Institute of Information and Communications Technology (NICT) of Japan.



# PSI project

# Petascale System Evaluation Methodology

Investigate the peta-scale system architecture by analyzing applications and creating the system evaluation model



# Towards Petascale Computing



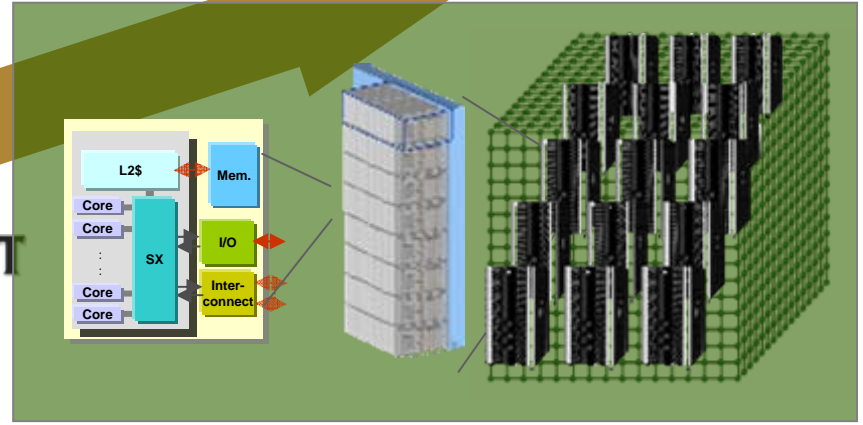
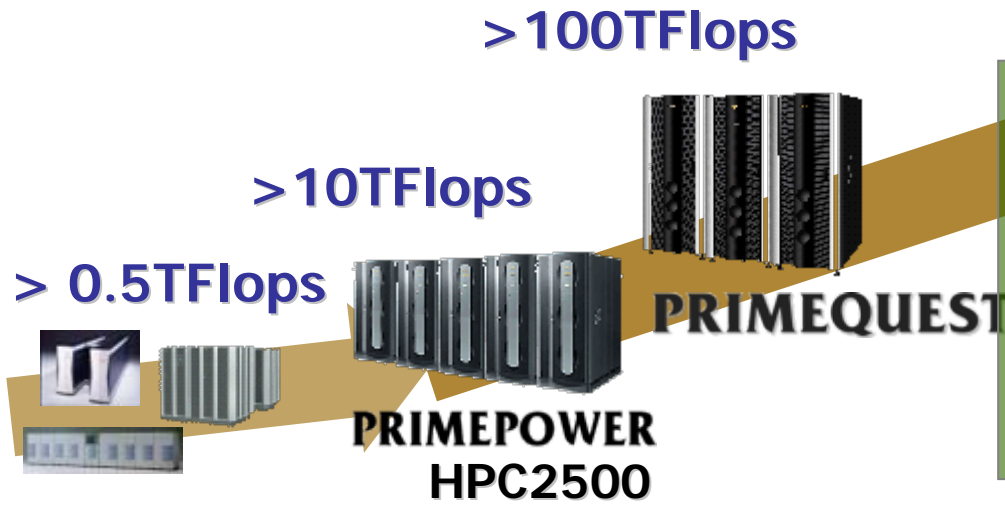
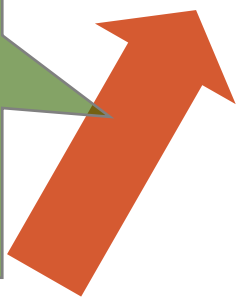
2005  
 Fundamental R&D projects with Kyushu Univ.

2008  
 Next Generation Supercomputer Project of Japan

2011  
*Petascale Computer*

**Challenges in achieving Petaflops**

- High performance & low power consumption CPU
- Highly reliable system
- Highly scalable and reliable interconnect

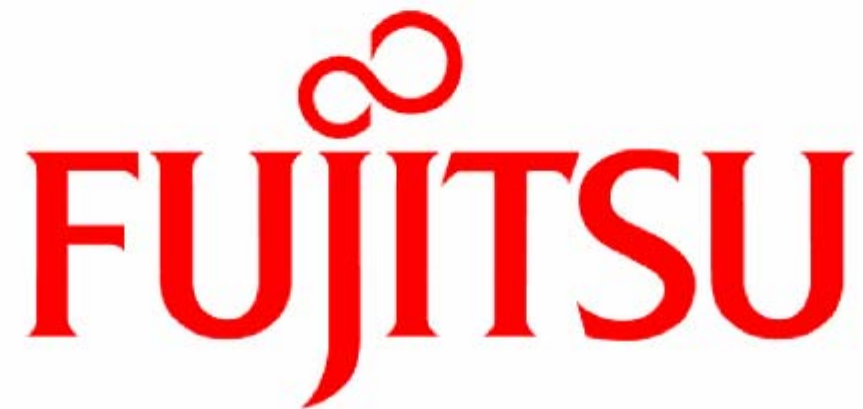


# Agenda

- Fujitsu's HPC Solution Offerings
  - Platform
  - Software
- Challenges of Petascale Computing
  - Petascale Interconnect Project
  - Contributions to Japan's Next Generation Supercomputing Project
  - HPC Product Roadmap
- Summary

# Summary

- Fujitsu continues to invest in HPC technology to provide solutions to meet the broadest user requirements at the highest levels of performance
- Fujitsu has embarked on the **Petascale Computing** challenge for its future HPC Product Line



**FUJITSU**

**THE POSSIBILITIES ARE INFINITE**