www.fujitsu.com

THE POSSIBILITIES ARE INFINITE FUJITSU

Fujitsu's challenge for Petascale Computing

Practical

October 9, 2008

Motoi Okuda

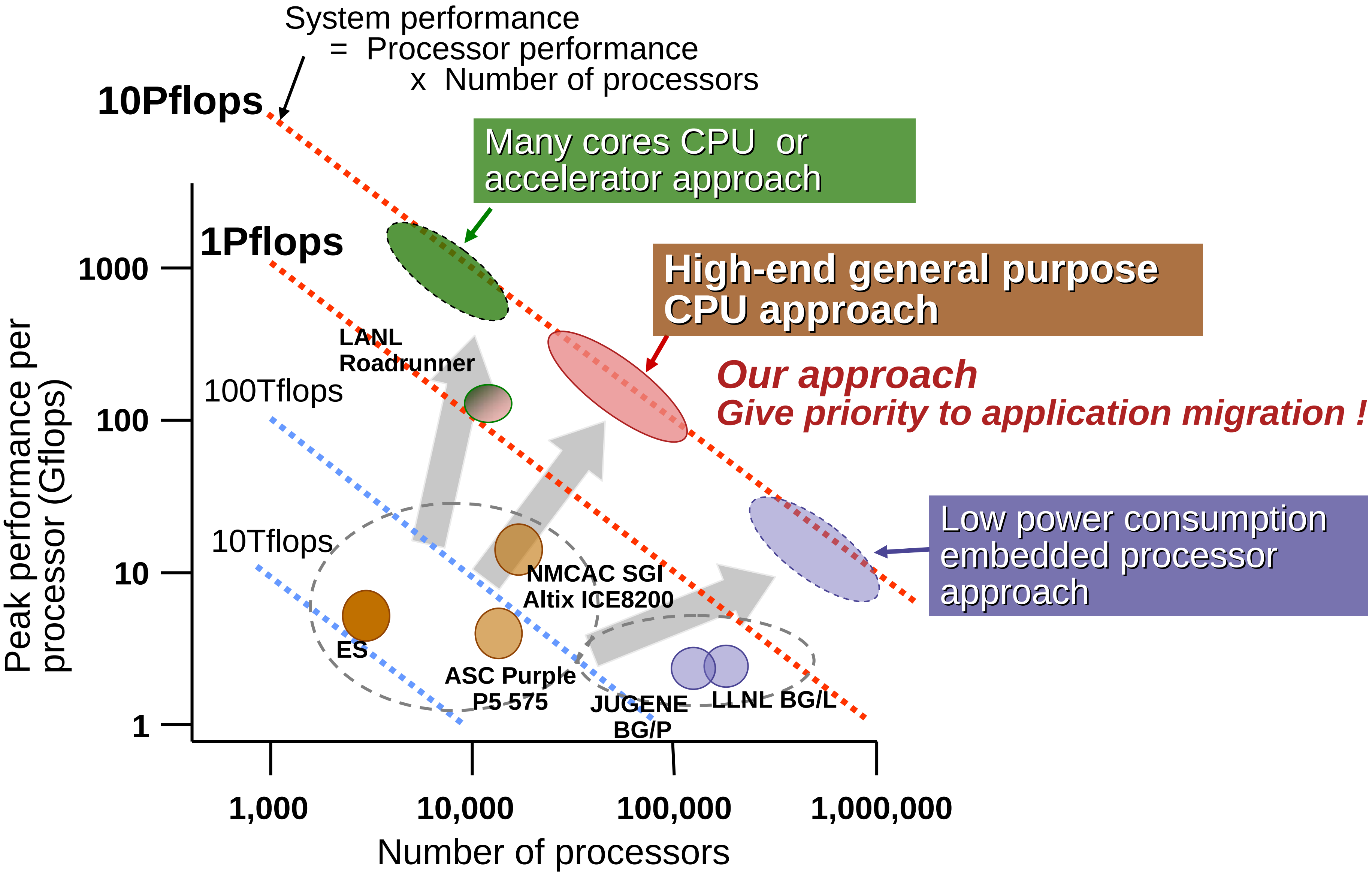
Technical Computing Solutions Group

Fujitsu Limited

Agenda

- Fujitsu's Approach for Petascale Computing and HPC Solution Offerings
- Japanese Next Generation Supercomputer Project and Fujitsu's Contributions
- Fujitsu's Challenges for Petascale Computing
- Conclusion

Fujitsu's approach for Scaling up to 10 Pflops



Key Issues for Approaching Petascale Computing

- How to utilize multi-core CPU?
- How to handle hundred thousand processes?
- How to realize high reliability, availability and data integrity of hundred thousand nodes system?
- How to decrease electric power and footprint?

- Fujitsu's stepwise approach to product release ensures that customers can be prepared for Petascale computing

Step1 : 2008 ~

- ***The new high end technical computing server FX1***
  - *New Integrated Multi-core Parallel ArChiTecture*
  - *Intelligent interconnect*
  - *Extremely reliable CPU design*
    - *Provides a highly efficient hybrid parallel programming environment*
- ***Design of Petascale system which inherits FX1 architecture***

Step2 : 2011 ~

- ***Petascale system with new high performance, high reliable and low power consumption CPU, innovative interconnect and high density packaging***

Current Technical Computing Platforms

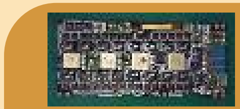
FUJITSU

Cluster Solutions

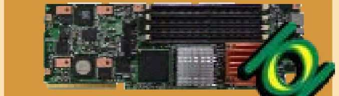
- Optimal price/performance for MPI-based applications
- Highly scalable
- InfiniBand interconnect

Solidware Solutions

- Ultra high performance for specific applications



FPGA board



RG1000

PRIMERGY

BX Series

RX Series

HX600

NEW

IA/Linux

High-end TC Solutions

- Scalability up to 100Tflops class
- Highly effective performance
- High-end RISC CPU

NEW

FX1

SPARC64™ VII

sparc64

SPARC/Solaris

Large-scale SMP System Solutions

- Up to 2TB memory space for TC applications
- High I/O bandwidth for I/O server
- High reliability based on main- frame technology
- High-end RISC CPU

PRIMEQUEST

PRIMEQUEST 580

Itanium® 2

~32cpu

IA/Linux

NEW

SPARC Enterprise

SPARC Enterprise M9000

SPARC64™ VII

~64cpu

sparc64

SPARC/Solaris

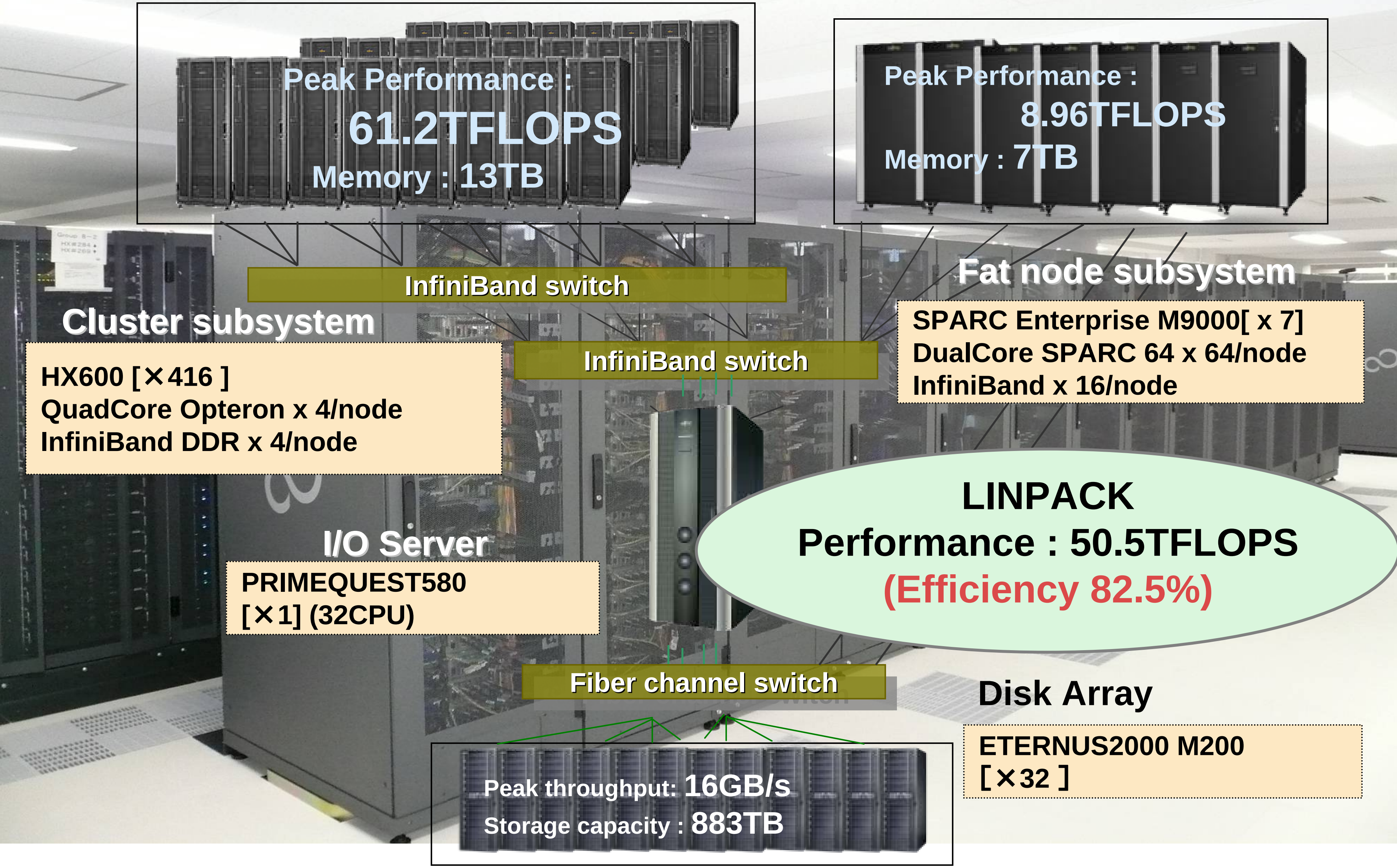
Customers of large scale TC systems

- Fujitsu has installed over 1200 TC systems for over 400 customers.

Customer	Type	No. of CPU	Performance
Japan Aerospace Exploration Agency (JAXA) *This system will be installed in end of 2008	Cluster Scalar SMP	>3,500CPU	135TFlops
Manufacturer A	Scalar SMP Cluster	>3,500CPU	>80TFlops
KYOTO University Computing Center	Cluster Scalar SMP	>2,000CPU	>61.2TFlops
KYUSYU University Computing Center	Scalar SMP Cluster	1,824CPU	32TFlops
Manufacturer B	Cluster	>1,200CPU	>15TFlops
RIKEN	Cluster	3,088CPU	26.18TFlops
NAGOYA University Computing Center	Scalar SMP	1,600 CPU	13TFlops
TOKYO University KAMIOKA Observatory	Cluster	540CPU	12.9TFlops
National Institute of Genetic	Cluster Scalar SMP	324CPU	6.9TFlops
Institute for Molecular Science	Scalar SMP	320CPU	4TFlops

Latest case study

- Kyoto University is one of the biggest computing centers in Japan.



Peak Performance : 61.2TFLOPS

Memory : 13TB

Peak Performance : 8.96TFLOPS

Memory : 7TB

InfiniBand switch

Cluster subsystem

HX600 [×416]
QuadCore Opteron x 4/node
InfiniBand DDR x 4/node

InfiniBand switch

Fat node subsystem

SPARC Enterprise M9000[x 7]
DualCore SPARC 64 x 64/node
InfiniBand x 16/node

LINPACK
Performance : 50.5TFLOPS
(Efficiency 82.5%)

I/O Server

PRIMEQUEST580
[×1] (32CPU)

Fiber channel switch

Disk Array

ETERNUS2000 M200
[×32]

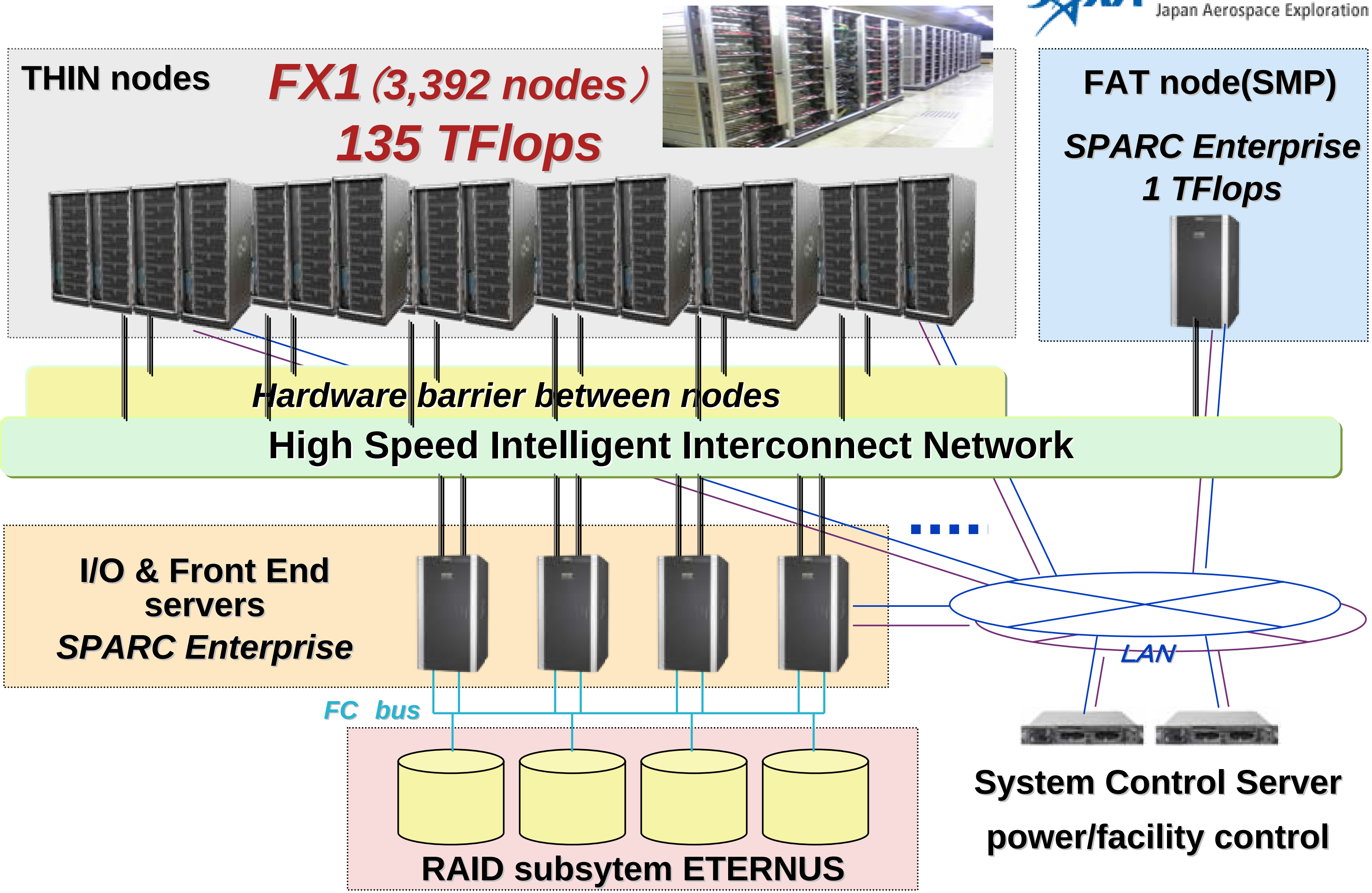
Peak throughput: 16GB/s

Storage capacity : 883TB

FX1 Launch customer

- First system will be installed at JAXA by the end of 2008

JAXA 宇宙航空研究開発機構
Japan Aerospace Exploration Agency



THIN nodes

FX1 (3,392 nodes)

135 TFlops

FAT node(SMP)

SPARC Enterprise
1 TFlops

Hardware barrier between nodes

High Speed Intelligent Interconnect Network

I/O & Front End servers
SPARC Enterprise

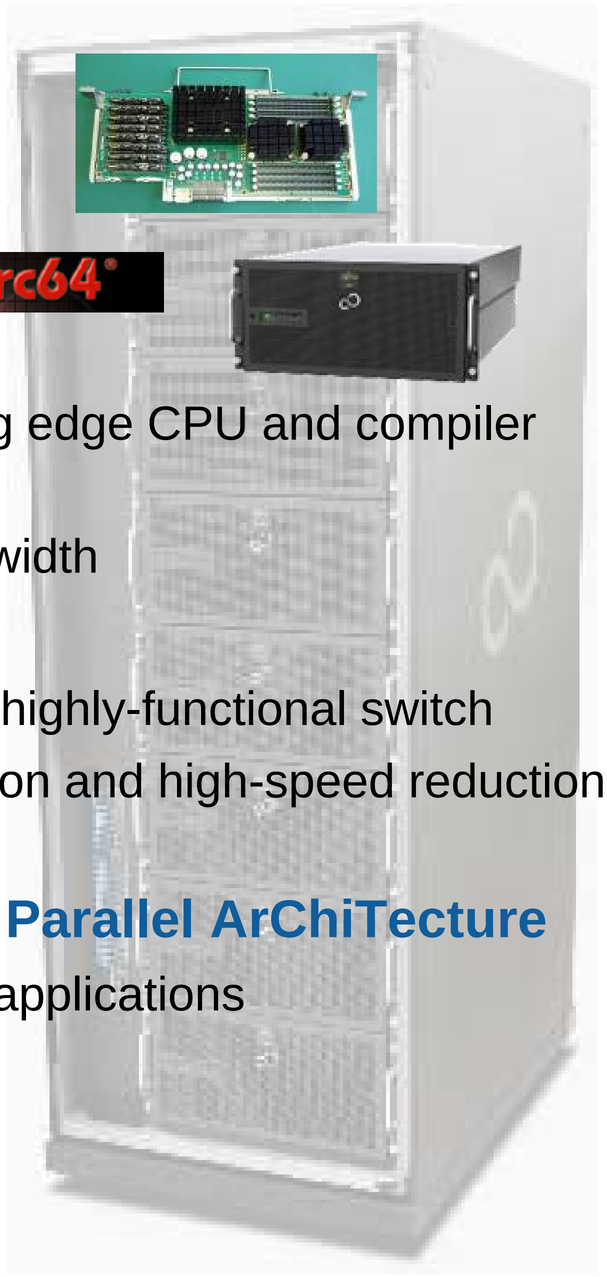
FC bus

LAN

RAID subsytem ETERNUS

System Control Server
power/facility control

FX1 : New High-end TC Server - Outline -



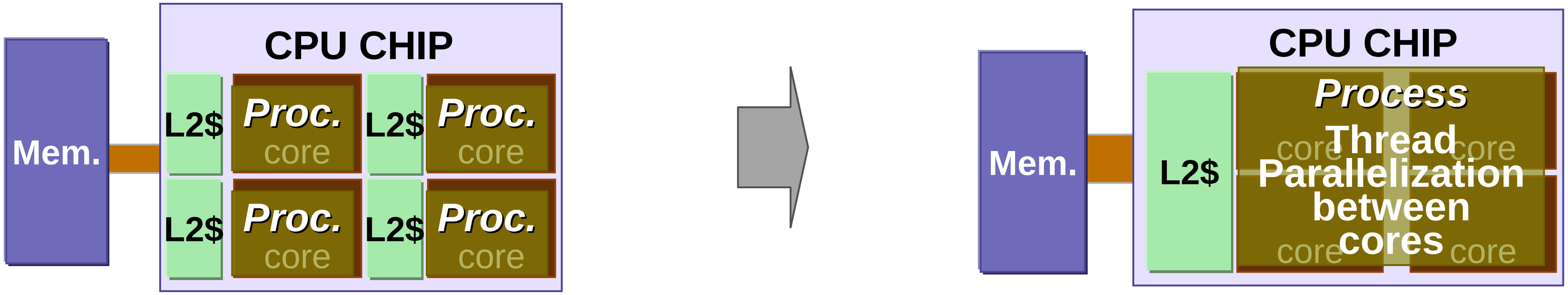
- **High-performance CPU designed by Fujitsu**
  - SPARC64$^{\text{TM}}$ VII : 4 cores by 65nm technology
  - Performance : 40 Gflops (2.5GHz)
- **New architecture for high-end TC server**
  - Integrated Multi-core Parallel ArChiTecture by leading edge CPU and compiler technologies
  - Blade type node configuration for high memory bandwidth
- **High-speed intelligent interconnect**
  - Combination of InfiniBand DDR interconnect and the highly-functional switch
  - Highly-functional switch realizes barrier synchronization and high-speed reduction between nodes by hardware
- **Petascale system inherits Integrated Multi-core Parallel ArChiTecture**
  - Suitable platform to develop and evaluate Petascale applications

Integrated Multi-core Parallel ArChiTecture

Introduction

Concept

- Highly efficient thread level parallel processing technology for multi-core chip



Advantage

- Handles the multi-core CPU as one equivalent faster CPU
  - ➔ Reduces number of MPI processes to $1/n_{core}$ and increases parallel efficiency
  - ➔ Reduces memory-wall problem

Challenge

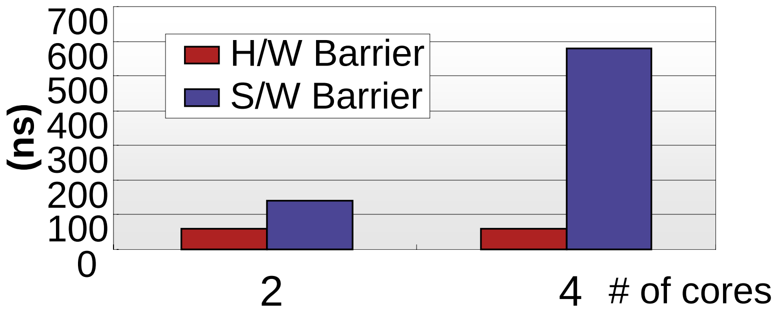
- How to decrease the thread level parallelization overhead?

Integrated Multi-core Parallel ArChiTecture

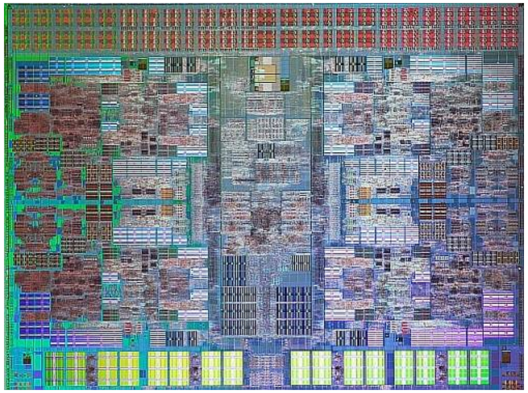
Key technologies

● CPU Technologies

- Hardware barrier synchronization between cores
  - ➔ Reduces overhead for parallel execution, 10 times faster than software emulation
  - ➔ Start up time is comparable to that of the vector unit
  - ➔ Barrier overhead remains constant regardless number of cores



Barrier Overhead



SPARC64™ VII

Real quad-core CPU for Technical Computing

(2.5GHz, 40Gflops/chip)

- Shared L2 cache memory(6MB)
  - ➔ Reduces the number of cache to cache data transfer
  - ➔ Efficient cache memory usage

● Compiler technologies

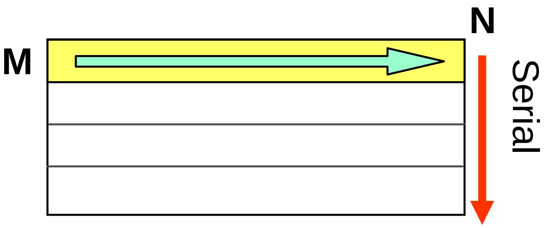
- Automatic parallelization or OpenMP on thread-based algorithm by vectorization technology

Integrated Multi-core Parallel ArChiTecture

Outline of parallelization methods

● Vectorization on vector machine

```
    DO J=1,N
V   DO I=1,M
V     A(I,J)=A(I,J+1)*B(I,J)
V   END
    END
```

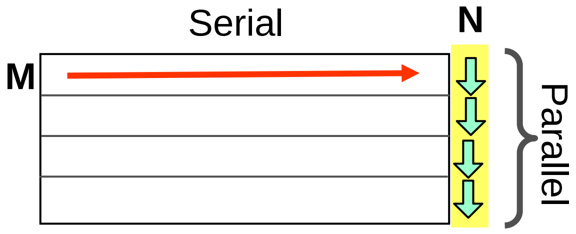


☺ Applicability : wide

☺ Overhead : frequent but low cost

● Legacy parallelization on scalar machine

```
P   DO J=1,N
P     DO I=1,M
P       A(I,J)=A(I,J)*B(I,J)
P     END
P   END
```

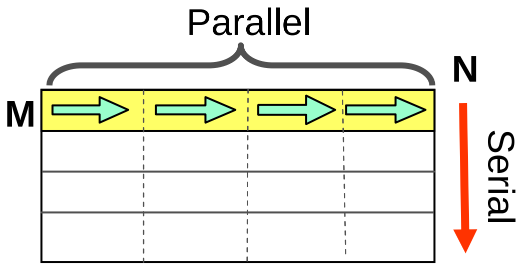


☹ Applicability : narrow (required wide range analysis)

☺ Synchronization : occasional

● Fine-grain parallelization on scalar machine

```
    DO J=1,N
P   DO I=1,M
P     A(I,J)=A(I,J+1)*B(I,J)
P   END
    END
```



☺ Applicability : wide

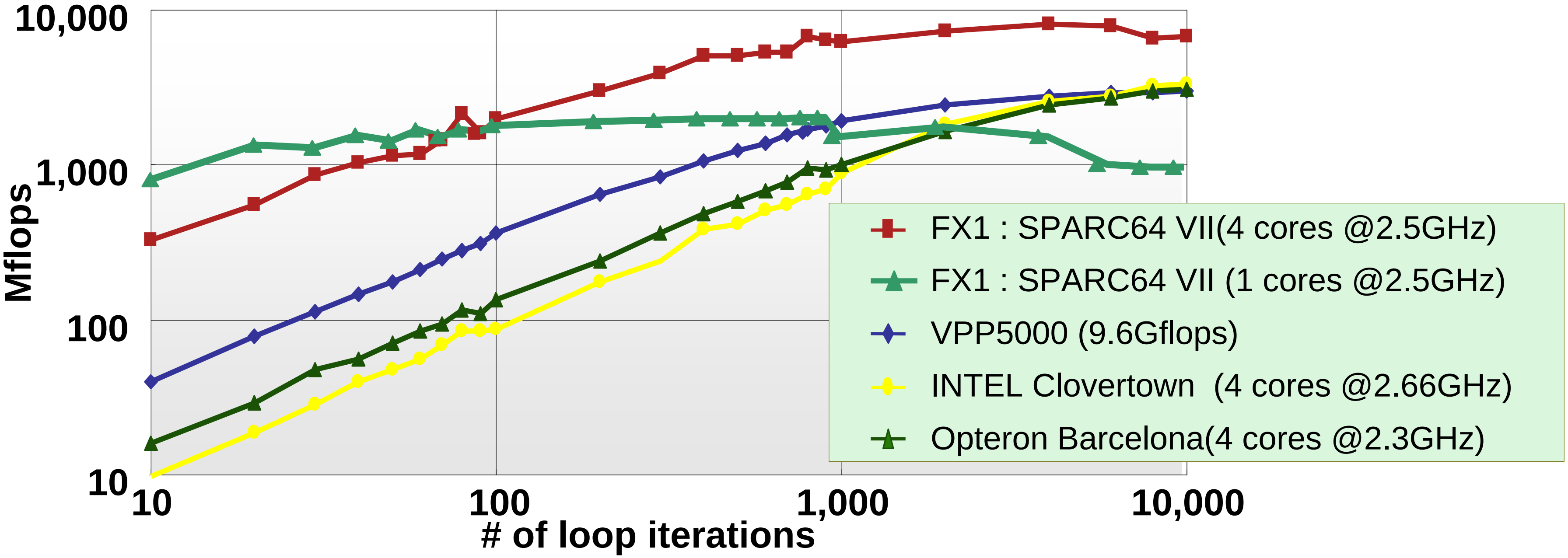
☹ Synchronization : frequent

Integrated Multi-core Parallel ArChiTecture takes cares of this weak point

Integrated Multi-core Parallel ArChiTecture, preliminary measured data

Performance measurement by automatic parallelization

- **LINPACK performance on 1 CPU(4 cores)**
  - n = 100 ➔ 3.26 Gflops
  - n = 40,000 ➔ 37.8 Gflops (93.8%)
- **Performance comparison of DAXPY (EuroBen Kernel 8) on 1 CPU**
  - 4core + IMPACT shows better performance than
    - ➔ 1core performance with small number of loop iterations
    - ➔ X86 servers

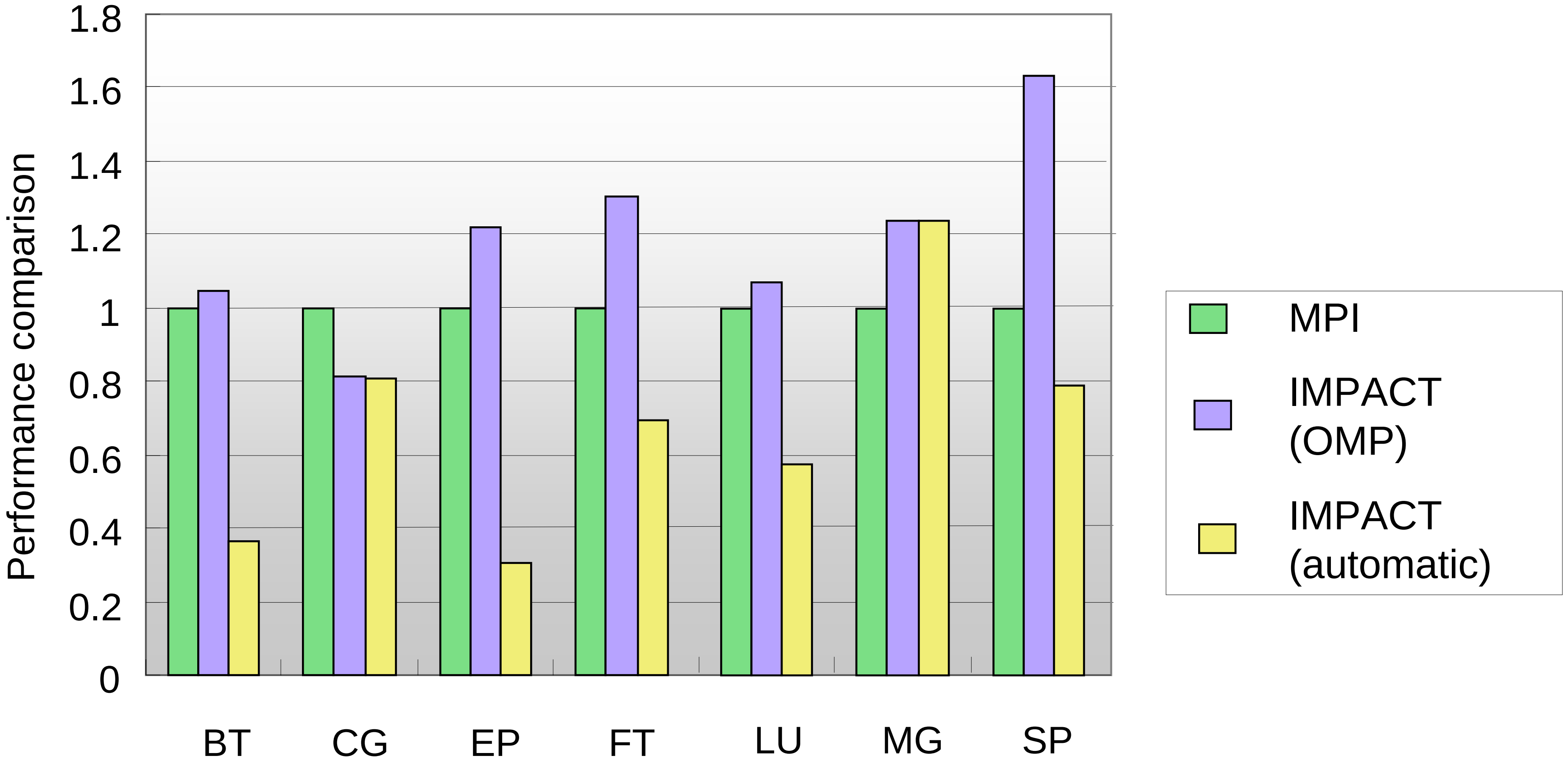


Performance of DAXPY

Integrated Multi-core Parallel ArChiTecture, preliminary measured data

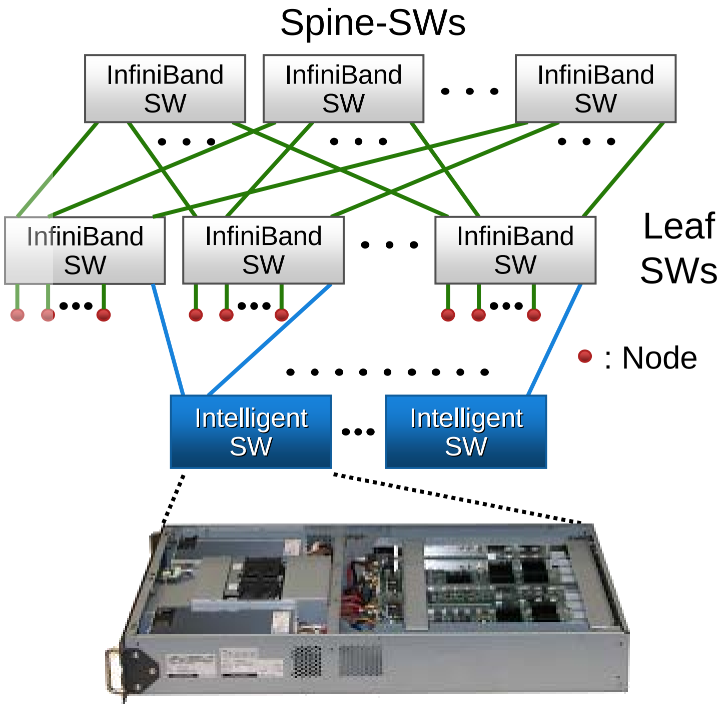
Performance measurement of NPB on 1 CPU

- **Performance comparison of NPB class C between pure MPI and Integrated Multi-core Parallel ArChiTecture on 1 CPU (4 cores)**
  - IMPACT(OMP) is better than pure MPI for 6/7 programs



Introduction

- **Combination of Fat tree topology InfiniBand DDR interconnect and the highly-functional switch (Intelligent switch)**
- **Intelligent switch**
  - Result of the PSI (Petascale System Interconnect) national project
  - Functions
    - Hardware barrier function among nodes
    - Hardware assistance for MPI functions (synchronization and reduction)
    - Global ping for OS scheduling
  - Advantages
    - Faster HW Barrier speeds up OpenMP and data parallel FORTRAN (XPF)
    - Fast collective operations accelerate highly parallel applications
    - Reduces OS jitter effect

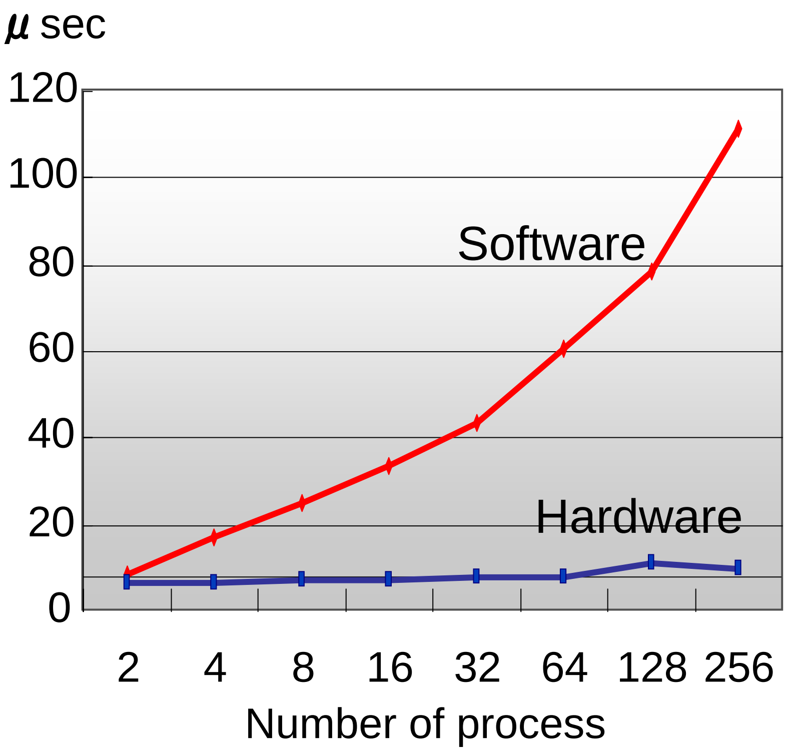


Intelligent Switch & its connection

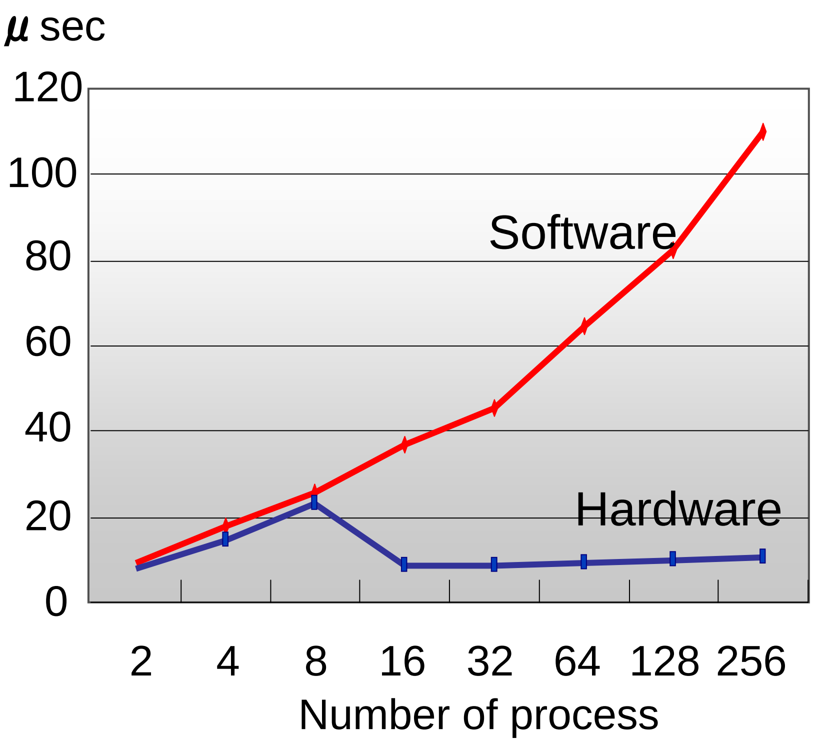
FX1 Intelligent Interconnect

High performance barrier & reduction hardware

- **Hardware barrier and reduction shows low latency and constant overhead in comparison with software barrier and reduction*.**



Barrier



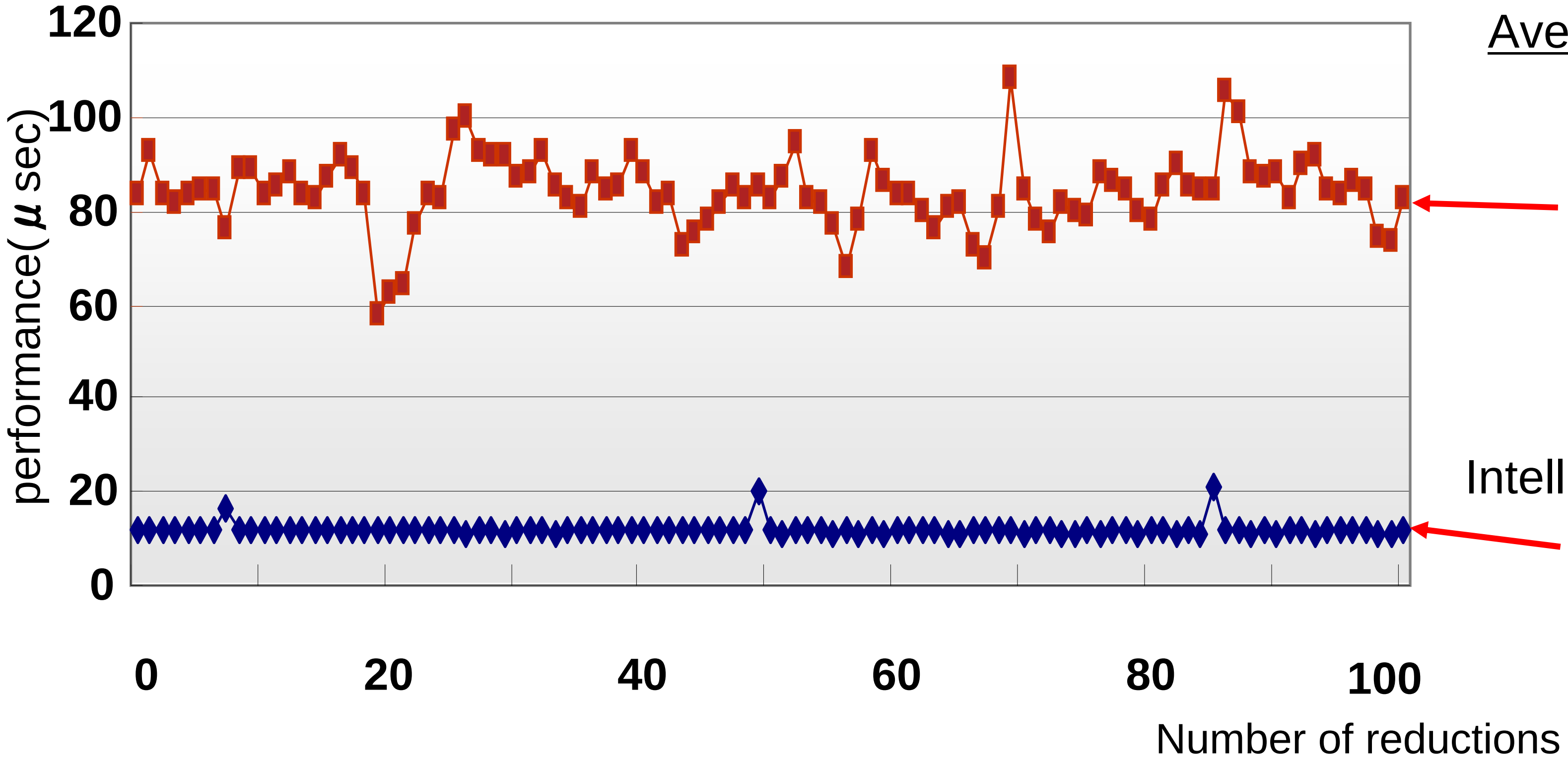
Reduction

*** :** Executed by host processor using butterfly network built by point to point communication.

FX1 Intelligent Interconnect

Stability of reduction function

- **Intelligent interconnect realizes stable reduction performance by global ping function**



Average

Software

82.99μsec

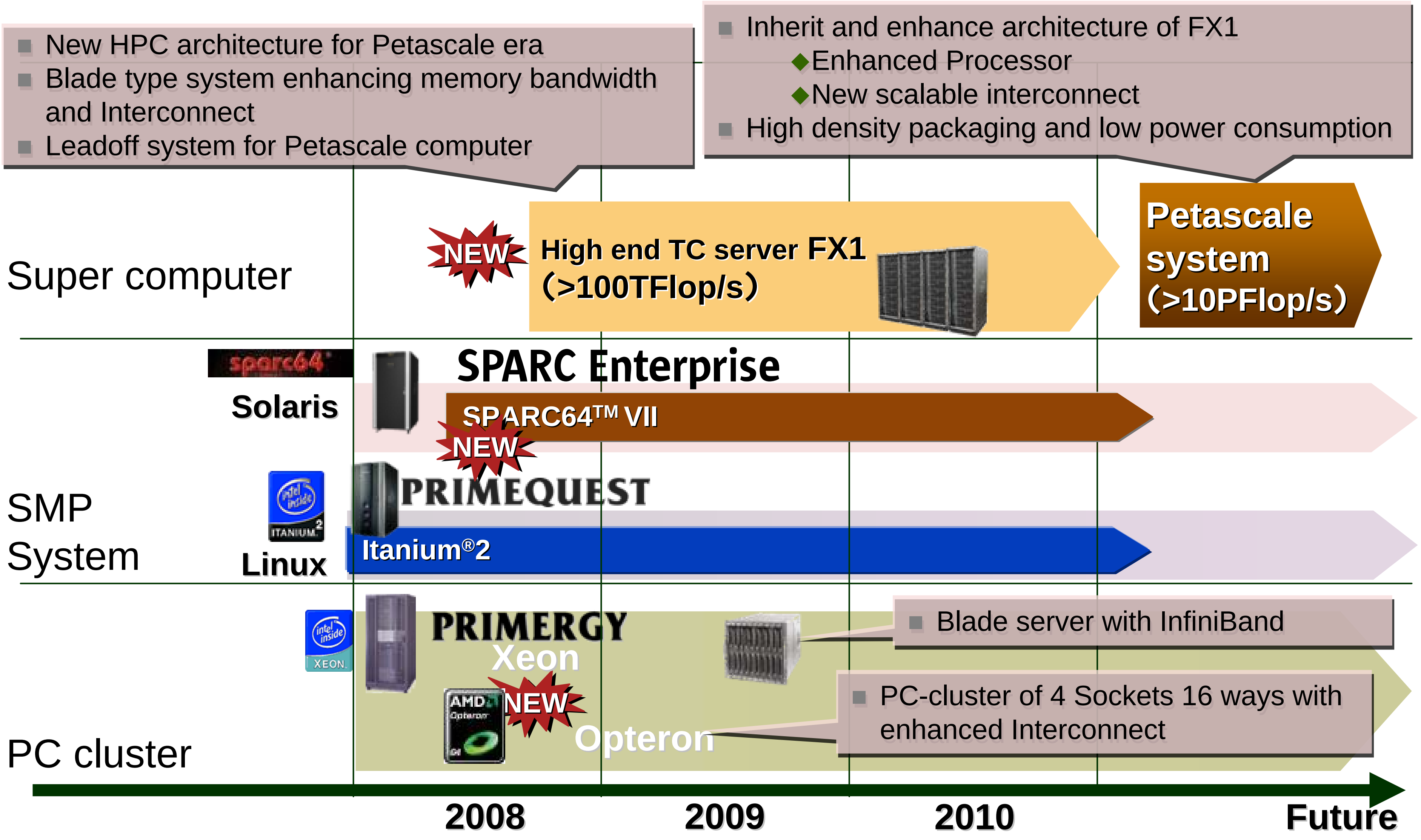
Intelligent Switch

11.86μsec

Reduction (All reduce) performance on 128 nodes system

Technical Computing server roadmap

- Development of the commodity based server and of the proprietary High End server for Technical Computing.



- New HPC architecture for Petascale era
- Blade type system enhancing memory bandwidth and Interconnect
- Leadoff system for Petascale computer

- Inherit and enhance architecture of FX1
  - ◆Enhanced Processor
  - ◆New scalable interconnect
- High density packaging and low power consumption

Super computer

NEW High end TC server FX1 (>100TFlop/s)

Petascale system (>10PFlop/s)

SMP System

Solaris

SPARC Enterprise

SPARC64™ VII NEW

Linux

PRIMEQUEST

Itanium®2

PC cluster

PRIMERGY

Xeon

NEW Opteron

- Blade server with InfiniBand
- PC-cluster of 4 Sockets 16 ways with enhanced Interconnect

2008 2009 2010 Future

Agenda

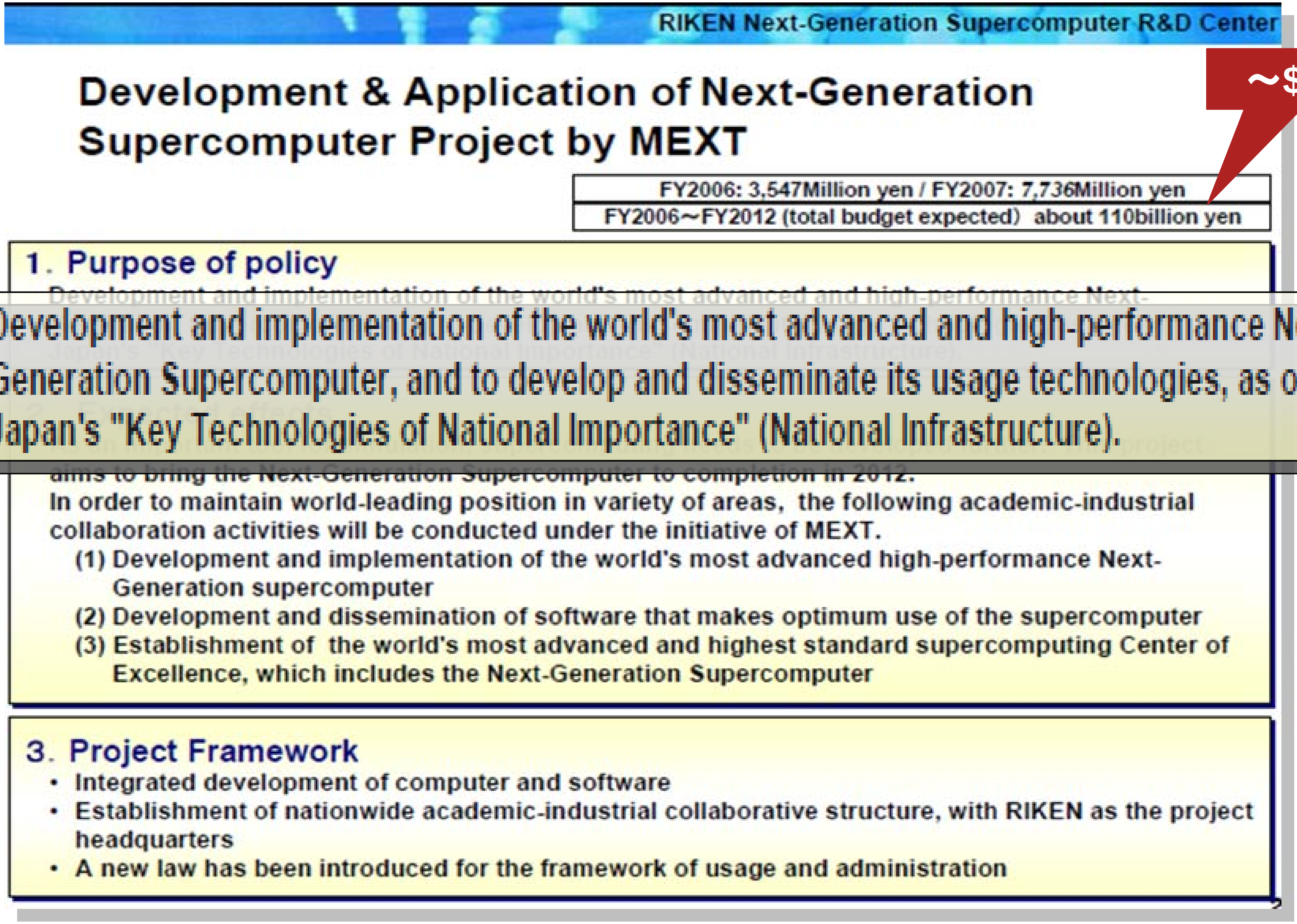
- Fujitsu’s Approach for Petascale Computing and HPC Solution Offerings
- **Japanese Next Generation Supercomputer Project and Fujitsu’s Contributions**
- Fujitsu’s Challenges for Petascale Computing
- Conclusion

Japanese Next Generation Supercomputer Project*

Project Target

Source: RIKEN official report

* : Sponsored by MEXT (Ministry of education, culture, sport, science and technology)



RIKEN Next-Generation Supercomputer R&D Center

Development & Application of Next-Generation Supercomputer Project by MEXT

FY2006: 3,547Million yen / FY2007: 7,736Million yen
FY2006～FY2012 (total budget expected) about 110billion yen

～$1.2 B

1. Purpose of policy

Development and implementation of the world's most advanced and high-performance Next-Generation Supercomputer, and to develop and disseminate its usage technologies, as one of Japan's "Key Technologies of National Importance" (National Infrastructure).

2. Expected effects

[illegible] project aims to bring the Next-Generation Supercomputer to completion in 2012.

In order to maintain world-leading position in variety of areas, the following academic-industrial collaboration activities will be conducted under the initiative of MEXT.

(1) Development and implementation of the world's most advanced high-performance Next-Generation supercomputer

(2) Development and dissemination of software that makes optimum use of the supercomputer

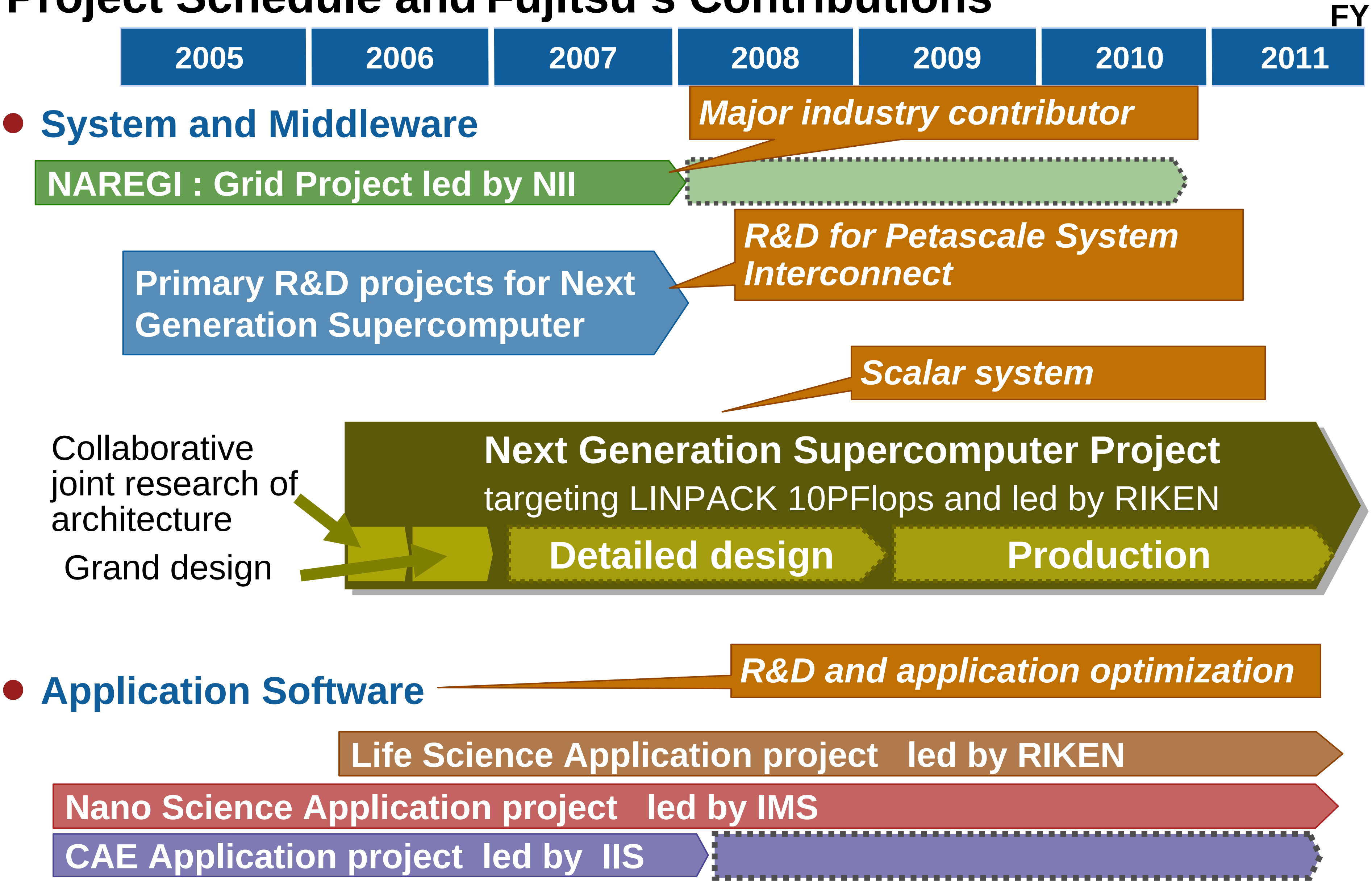
(3) Establishment of the world's most advanced and highest standard supercomputing Center of Excellence, which includes the Next-Generation Supercomputer

3. Project Framework

- Integrated development of computer and software
- Establishment of nationwide academic-industrial collaborative structure, with RIKEN as the project headquarters
- A new law has been introduced for the framework of usage and administration

Japanese Next Generation Supercomputer Project

Project Schedule and Fujitsu’s Contributions



- System and Middleware

Major industry contributor

NAREGI : Grid Project led by NII

R&D for Petascale System Interconnect

Primary R&D projects for Next Generation Supercomputer

Scalar system

Collaborative joint research of architecture

Grand design

Next Generation Supercomputer Project

targeting LINPACK 10PFlops and led by RIKEN

Detailed design

Production

- Application Software

R&D and application optimization

Life Science Application project led by RIKEN

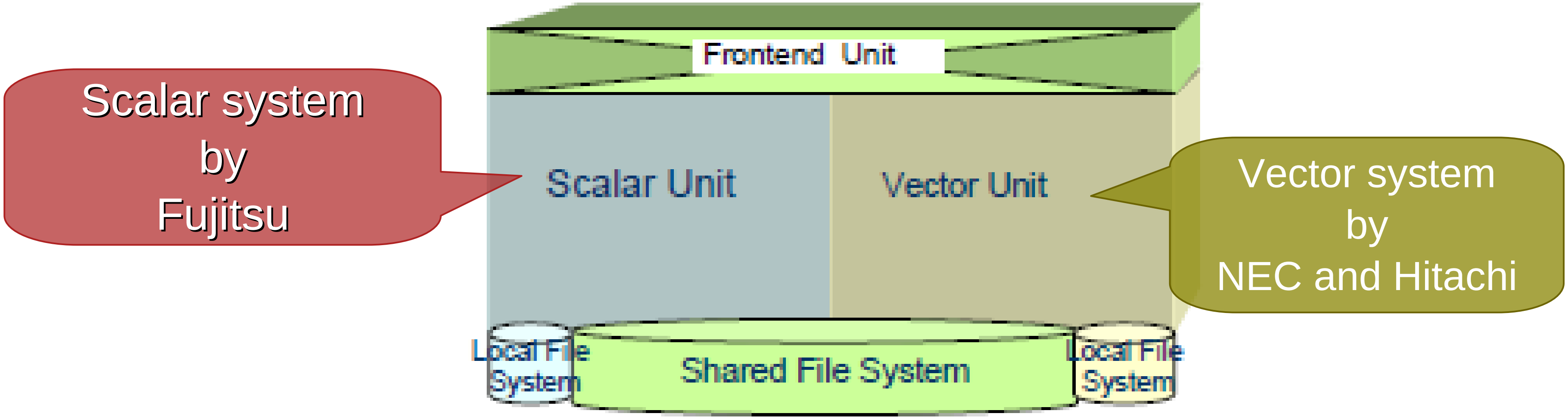
Nano Science Application project led by IMS

CAE Application project led by IIS

Japanese Next Generation Supercomputer Project

Project Outline

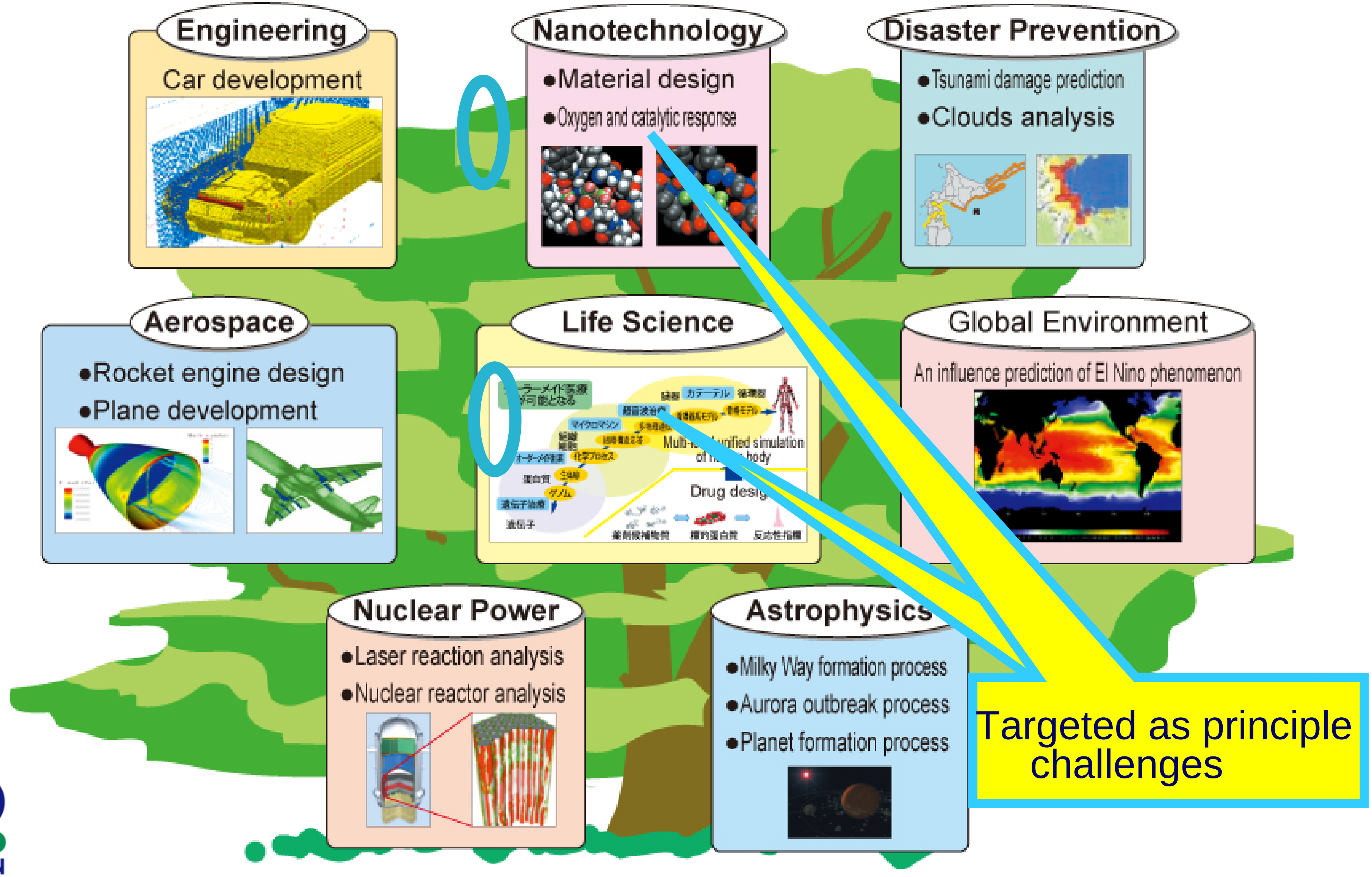
- **System configuration**
  - The hardware system consists of scalar and vector processor units.
- **The target performance**
  - 10PFlops on LINPACK BMT
- **Contributor**
  - Fujitsu, Hitachi and NEC join the project as the system developers.
- **Schedule**
  - Prototype system will be available for operation from the end of FY2010 and full system will be available from the end of FY2011.



Source: CSTP evaluation working group report

Japanese Next Generation Supercomputer Project

Major Applications of Next Generation Supercomputer



Engineering
Car development

Nanotechnology
- Material design
- Oxygen and catalytic response

Disaster Prevention
- Tsunami damage prediction
- Clouds analysis

Aerospace
- Rocket engine design
- Plane development

Life Science

Global Environment
An influence prediction of El Nino phenomenon

Nuclear Power
- Laser reaction analysis
- Nuclear reactor analysis

Astrophysics
- Milky Way formation process
- Aurora outbreak process
- Planet formation process

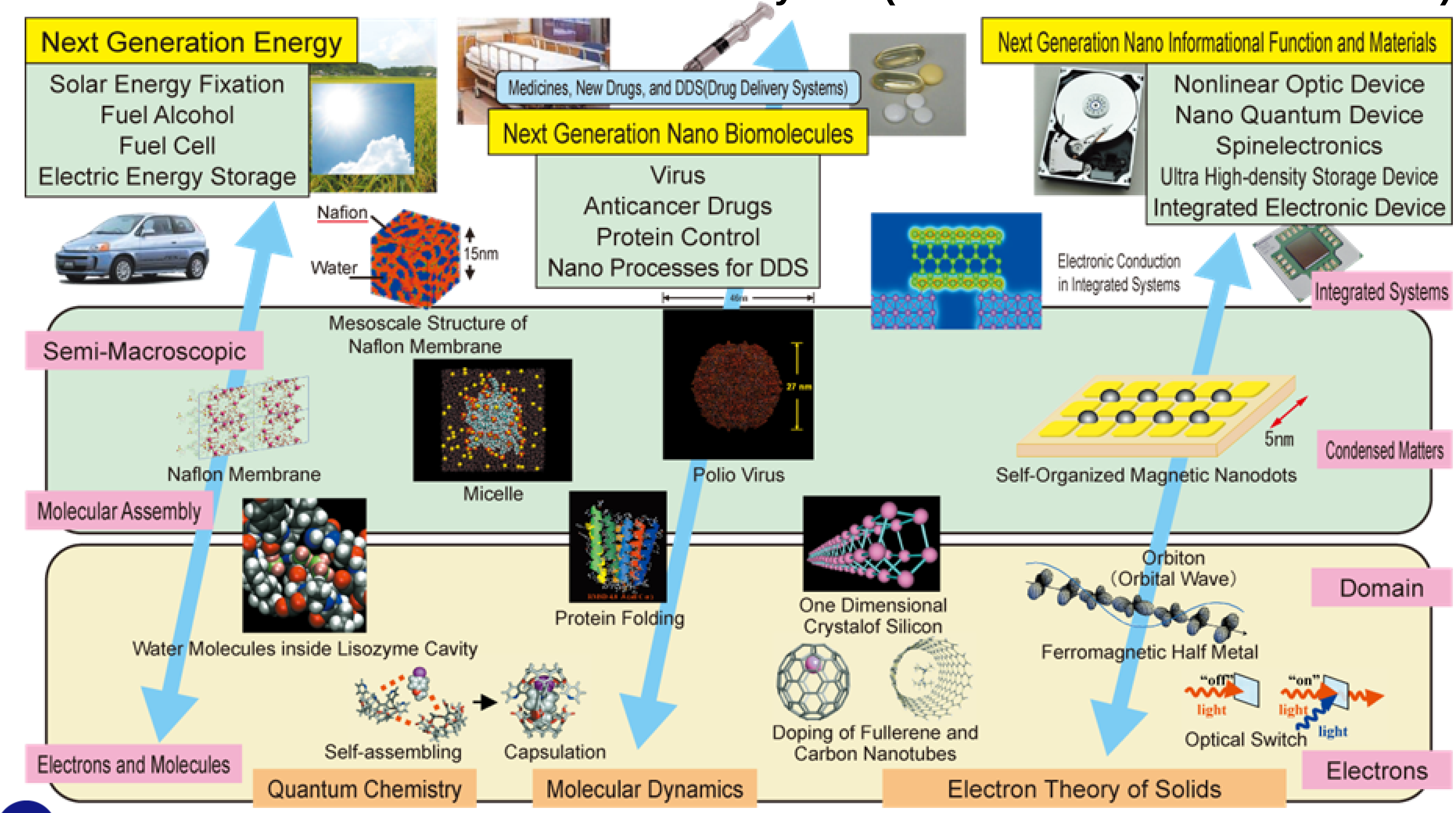
Targeted as principle challenges

Through the courtesy of RIKEN

Japanese Next Generation Supercomputer Project

Basic Concept for Simulations in Nano-Science

Led by IMS (Institute for Molecular Science)



Next Generation Energy

Solar Energy Fixation
Fuel Alcohol
Fuel Cell
Electric Energy Storage

Medicines, New Drugs, and DDS(Drug Delivery Systems)

Next Generation Nano Biomolecules

Virus
Anticancer Drugs
Protein Control
Nano Processes for DDS

Next Generation Nano Informational Function and Materials

Nonlinear Optic Device
Nano Quantum Device
Spinelectronics
Ultra High-density Storage Device
Integrated Electronic Device

Semi-Macroscopic

Mesoscale Structure of Nafion Membrane

Electronic Conduction in Integrated Systems

Integrated Systems

Naflon Membrane

Micelle

Polio Virus

Self-Organized Magnetic Nanodots

Condensed Matters

Molecular Assembly

Protein Folding

One Dimensional Crystalof Silicon

Orbiton (Orbital Wave)

Domain

Water Molecules inside Lisozyme Cavity

Ferromagnetic Half Metal

Self-assembling

Capsulation

Doping of Fullerene and Carbon Nanotubes

Optical Switch

Electrons and Molecules

Quantum Chemistry

Molecular Dynamics

Electron Theory of Solids

Electrons

Through the courtesy of RIKEN

Japanese Next Generation Supercomputer Project

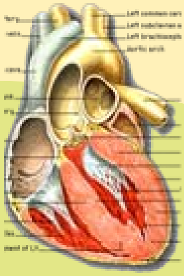
Basic Concept for Simulations in Life Sciences

Led by RIKEN

Meso

Micro-machine

Organ



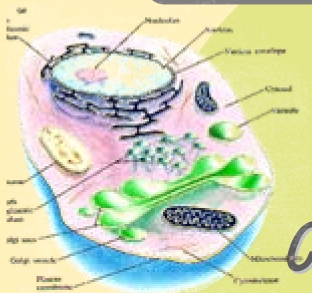
http://info.med.vale.edu/

Organism

Blood Circulation

Vascular System

Macro



http://ridge.icu.ac.jp

Tissue Structure

Multi-physics

HIFU

Cell

Tissue

Catheter

RIKEN

Chemical Process

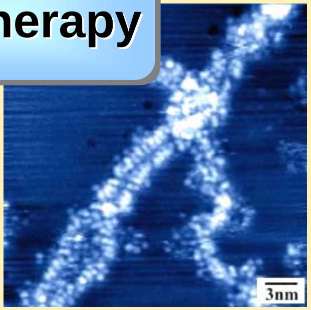
Bio-MD

DDS

Protein

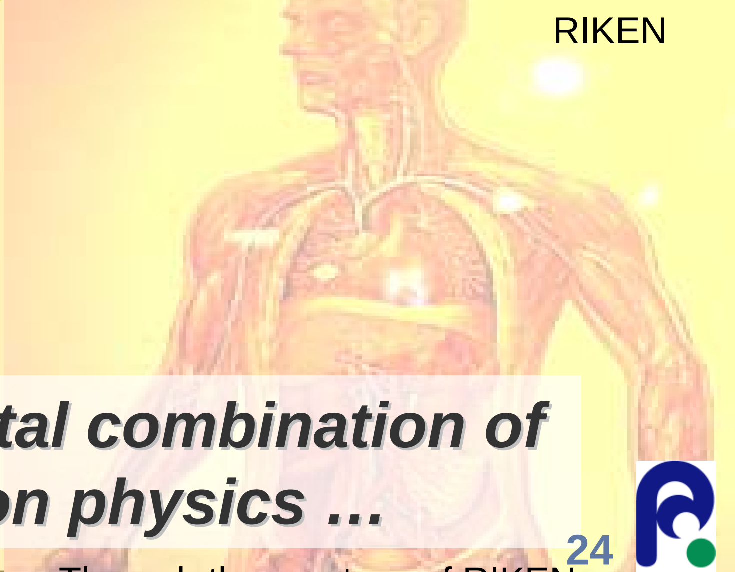
Genome

Gene Therapy



Genes

Micro



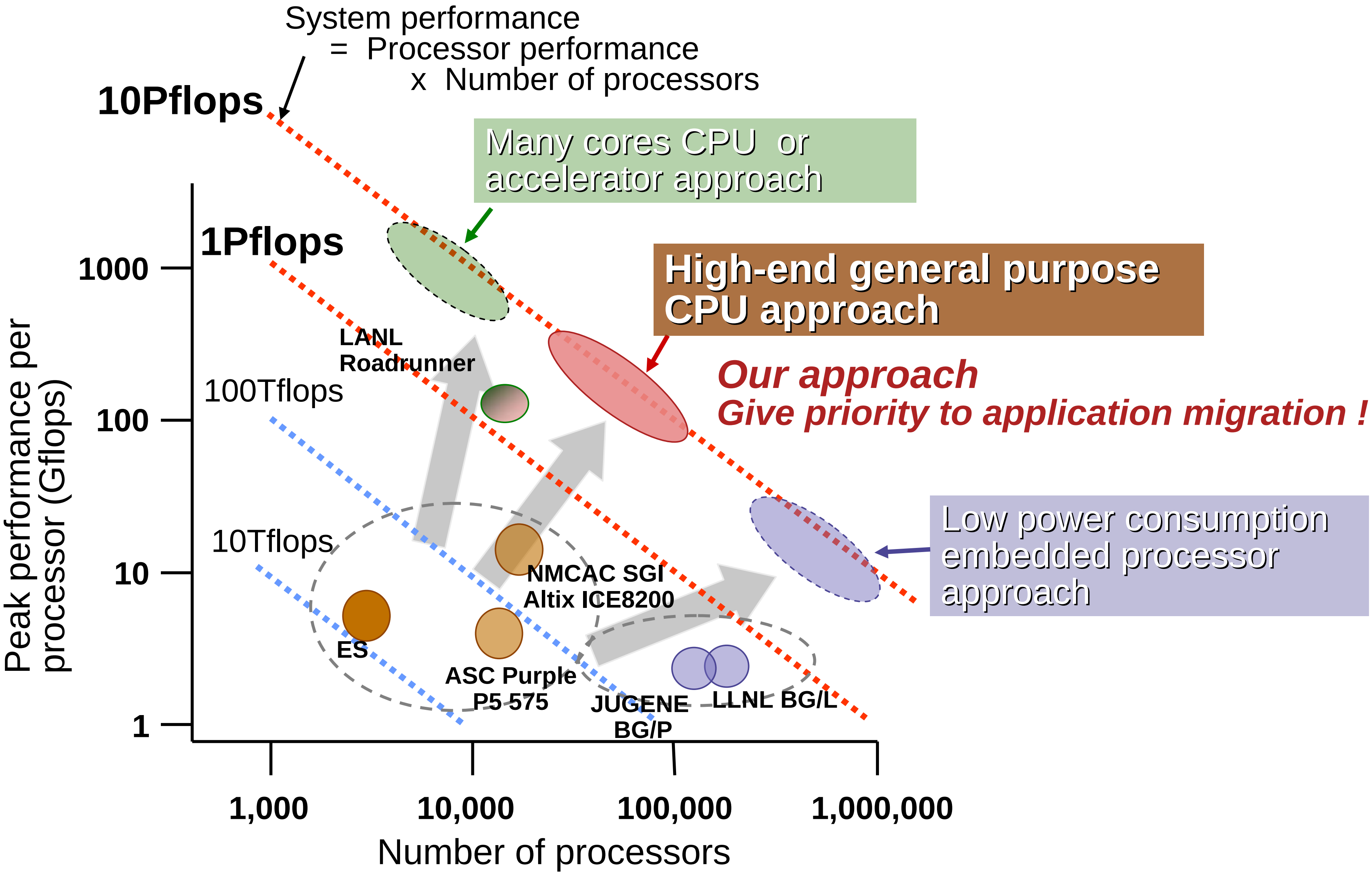
Under total combination of simulation physics …

Through the courtesy of RIKEN

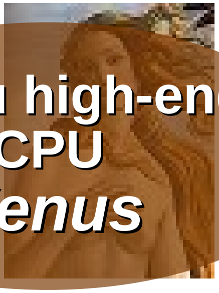
Agenda

- Fujitsu’s Approach for Petascale Computing and HPC Solution Offerings
- Japanese Next Generation Supercomputer Project and Fujitsu’s Contributions
- Fujitsu’s Challenges for Petascale Computing
- Conclusion

Fujitsu's approach for Scaling up to 10 Pflops



Fujitsu's Challenges for Petascale Supercomputer



Fujitsu high-end CPU *Venus*

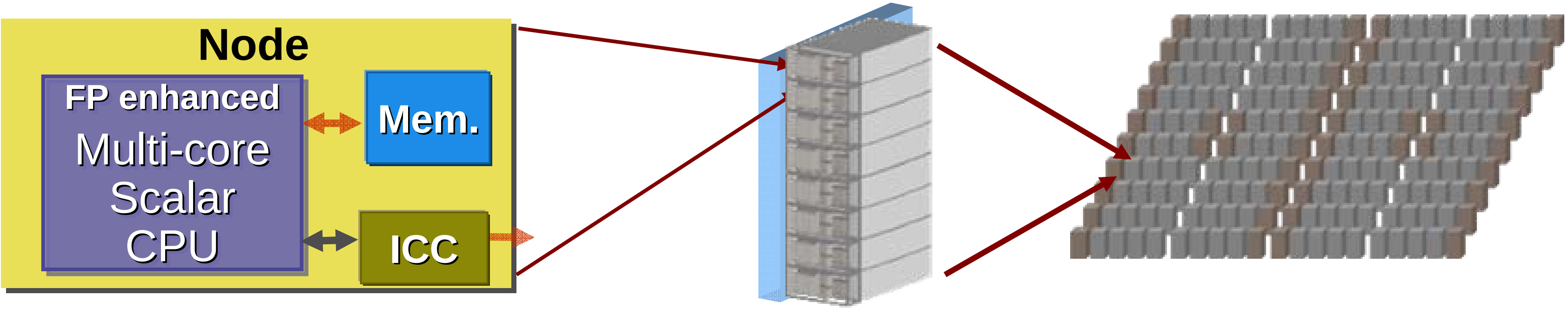
- FP enhanced multi-core scalar CPU (over 100Gflops/cpu) with main-frame level reliabilities
- Inherit Integrated Multi-core Parallel ArChiTecture of the FX1
- Low power consumption, targeting ~1/10 power consumption per flop

Leading edge interconnect

- 3D torus interconnect with scalability up to over 10Pflops, high bandwidth, high reliability and low latency

Latest packaging & cooling technology

- Targeting X ~10 packing density per flop by liquid cooling technology



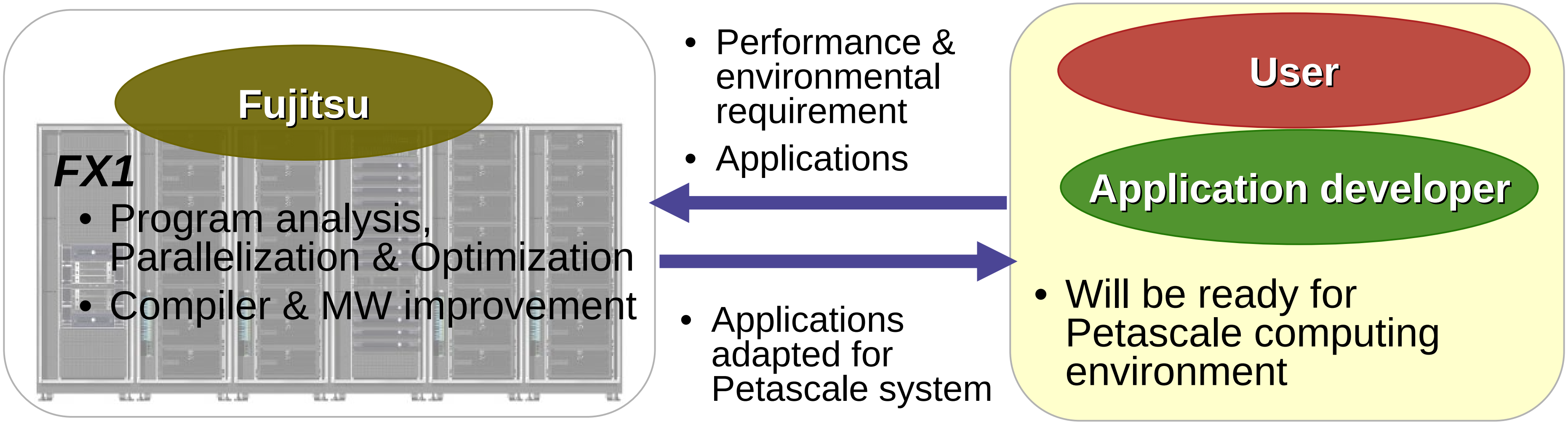
Fujitsu’s Challenges for Petascale Supercomputer

Middleware for Highly-parallel system

- Sophisticated compiler for program with 100,000 processes on multi-core CPU
- System management software for system with 100,000 nodes

Highly parallel Application S/W

- Optimization of highly parallel applications
- Collaboration with users and ISVs to optimize their software for Petascale system



Fujitsu

FX1

- Program analysis, Parallelization & Optimization
- Compiler & MW improvement

- Performance & environmental requirement
- Applications

- Applications adapted for Petascale system

User

Application developer

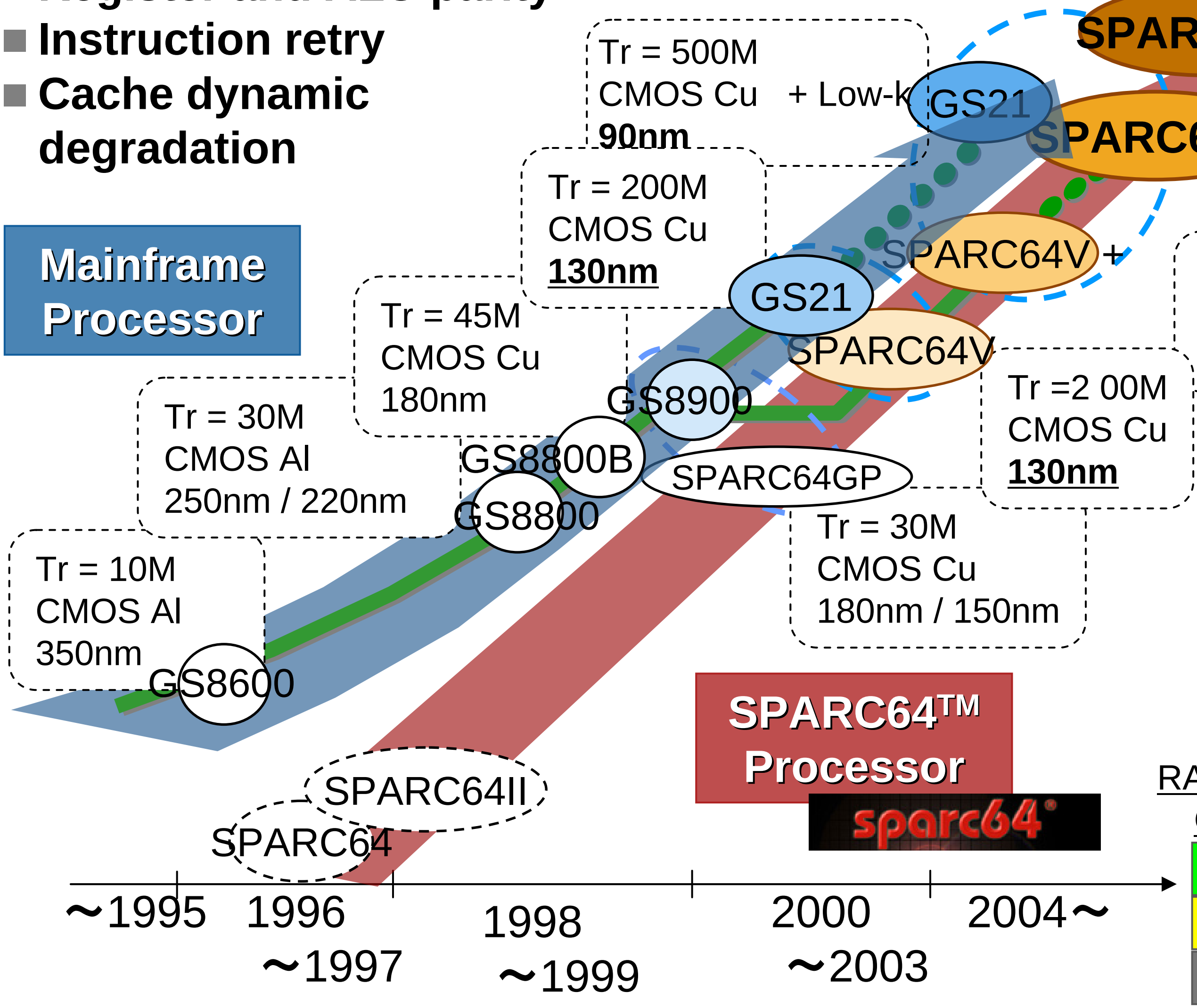
- Will be ready for Petascale computing environment

History of Fujitsu High–end Processor

- **High reliability and data integrity**
  - **Cache ECC**
  - **Register and ALU parity**
  - **Instruction retry**
  - **Cache dynamic degradation**

Venus

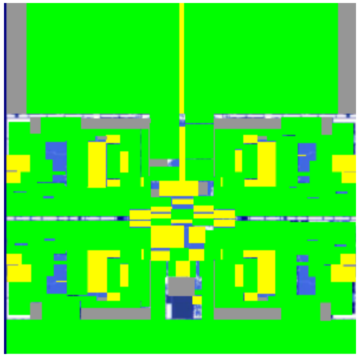
CPU for Petascale supercomputer



CMOS Cu+Low-k
65nm

Tr = 540M
CMOS Cu+Low-k
90nm

Tr = 400M
CMOS CU + Low-k
90nm



RAS Coverage of SPARC64VII
Guaranteed Data Integrity

🟩	: 1bit error Correctable
🟨	: 1bit error Detectable
⬜	: 1bit error harmless

Interconnect for parallel computer system

- **Interconnect type and its characteristic**

Interconnect type	Crossbar	Fat-Tree	Mesh / Torus
Performance	**◎(Best)**	**○(Good)**	**△(Average)**
Operability and usability	**◎(Best)**	**○(Good)**	**×(Weak)**
Cost, Packaging density and Power consumption	**×(Weak)**	**△(Average)**	**○(Good)**
Scalability	**Hundreds nodes** **×(Weak)**	**Thousands nodes** **△ - ○(Ave.-Good)**	**>10,000 nodes** **◎(Best)**
Representative	**Vector Parallel**	**PC cluster**	**Scalar Massive parallel**

- **Targeting over 10,000 nodes parallel system**
  - Cost, packaging density and power consumption are essential issues
  - Too much number of hops are needed for Mesh interconnect.

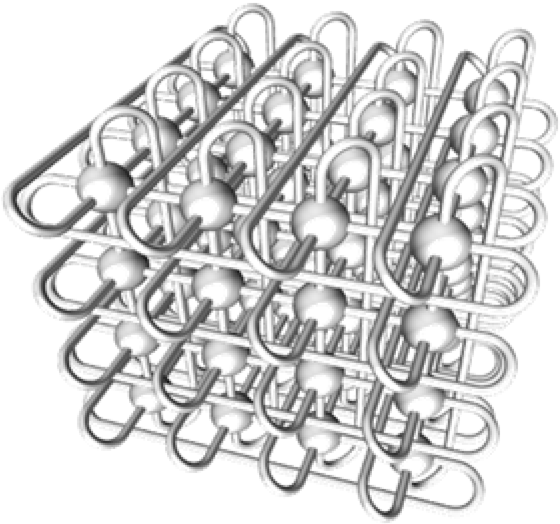
    ➔ Torus interconnect is a strong candidate

    ➔ The greatest challenge of Torus interconnect is operability and usability

- **Fujitsu challenges to develop an innovative Torus interconnect**

Fujitsu's Interconnect for Petascale computer system

- **Architecture**
  - Improved 3D Torus
  - Switchless
- **Advantages**
  - Low latency and low power consumption
  - Scalability over 100,000 nodes
  - High reliabilities and availabilities
  - High density packaging
  - Reduce wiring cost
  - Simple 3D torus logical (application) view



Improved 3D torus Architecture

Agenda

- Fujitsu’s Approach for Petascale Computing and HPC Solution Offerings
- Japanese Next Generation Supercomputer Project and Fujitsu’s Contributions
- Fujitsu’s Challenges for Petascale Computing
- Conclusion

Conclusion

- Fujitsu continues to invest in HPC technology to provide solutions to meet the broadest user requirements at the highest levels of performance
- Targeting sustained Pflops performance, Fujitsu has embarked on the Petascale Computing challenge

FUJITSU

THE POSSIBILITIES ARE INFINITE