Fujitsu's Challenge for Petascale Computing

Sustained

October 16th, 2008

Motoi Okuda

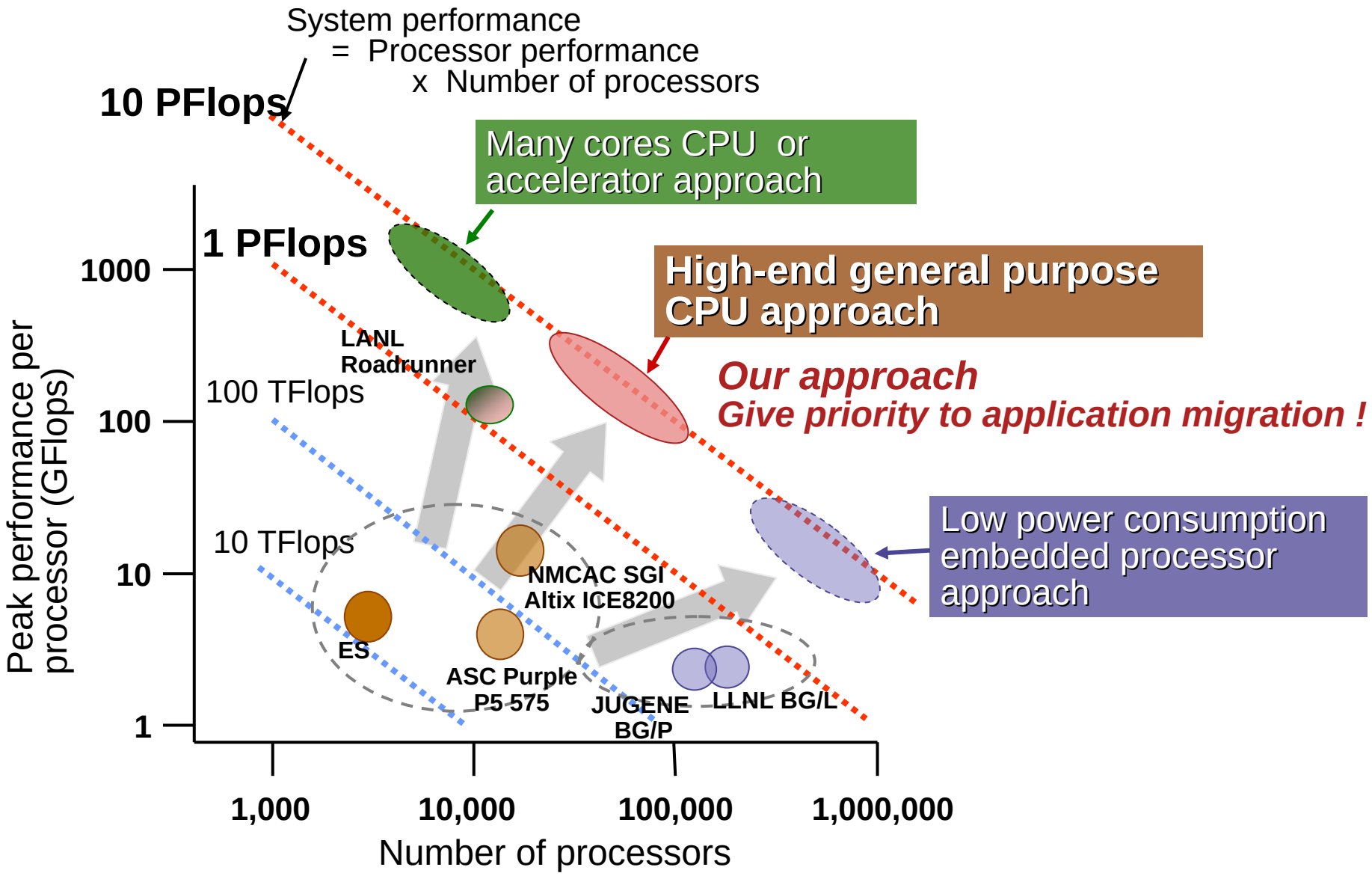
Technical Computing Solutions Unit

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 a hundred thousand processes?**
- **How to realize high reliability, availability and data integrity of a hundred thousand node 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, highly reliable and low power consumption CPU, innovative interconnect and high density packaging*

Current Technical Computing Platforms

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 100 TFlops class
- Highly effective performance
- High-end RISC CPU

NEW

FX1

SPARC64™ VII

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 mainframe technology
- High-end RISC CPU

PRIMEQUEST

PRIMEQUEST 580

Itanium® 2

~32cpu

IA/Linux

NEW

SPARC Enterprise

SPARC Enterprise M9000

SPARC64™ VII

~64cpu

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 (FX1) Scalar SMP（SPARC Enterprise)	>3,500	135 TFlops
Manufacturer A	Scalar SMP Cluster	>3,500	>80 TFlops
KYOTO Univ. Computing Center	Cluster(HX600) Scalar SMP（SPARC Enterprise)	>2,000	>61.2 TFlops
KYUSHU Univ. Computing Center	Scalar SMP（PRIMEQUEST) Cluster (PRIMERGY)	1,824	32 TFlops
Manufacturer B	Cluster	>1,200	>15 TFlops
RIKEN	Cluster (PRIMERGY)	3,088	26.18 TFlops
NAGOYA Univ. Computing Center	Scalar SMP (HPC2500)	1,600	13 TFlops
TOKYO Univ. KAMIOKA Observatory	Cluster (PRIMERGY)	540	12.9 TFlops
National Institute of Genetics	Cluster (PRIMERGY) Scalar SMP(SPARC Enterprise)	324	6.9 TFlops
Institute for Molecular Science	Scalar SMP（PRIMEQUEST)	320	4 TFlops

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

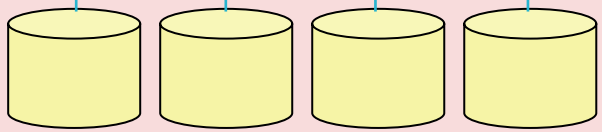
Hardware barrier between nodes

High Speed Intelligent Interconnect Network

I/O & front end servers
SPARC Enterprise

LAN

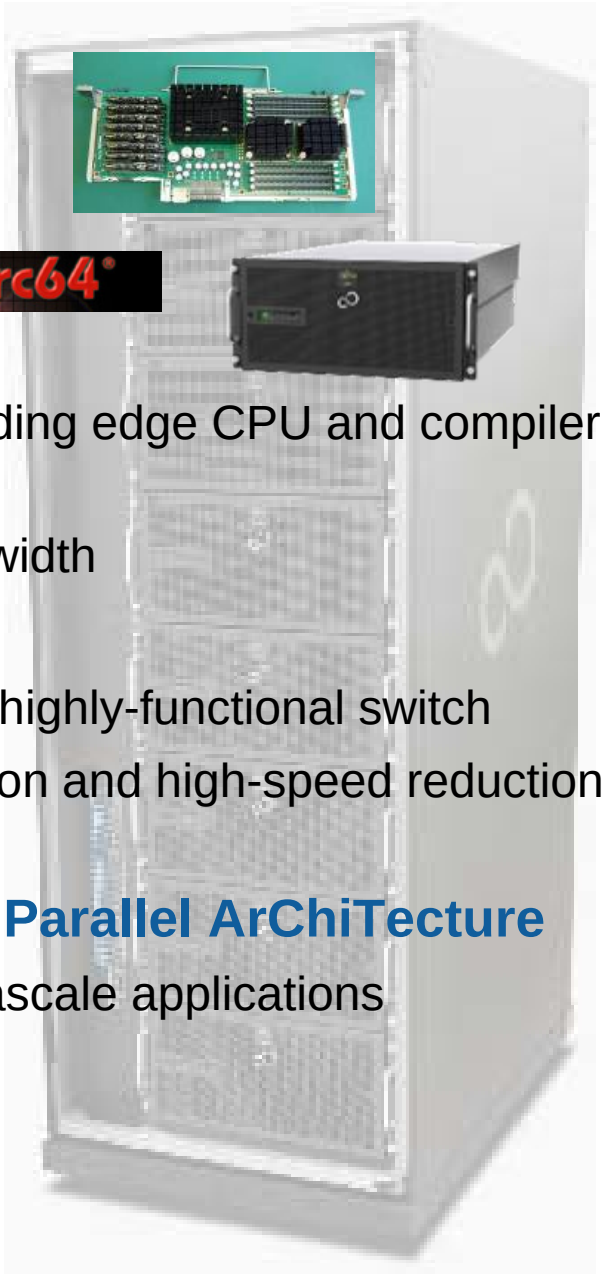
FC bus



RAID subsystem
ETERNUS

System Control Server
power/facility control

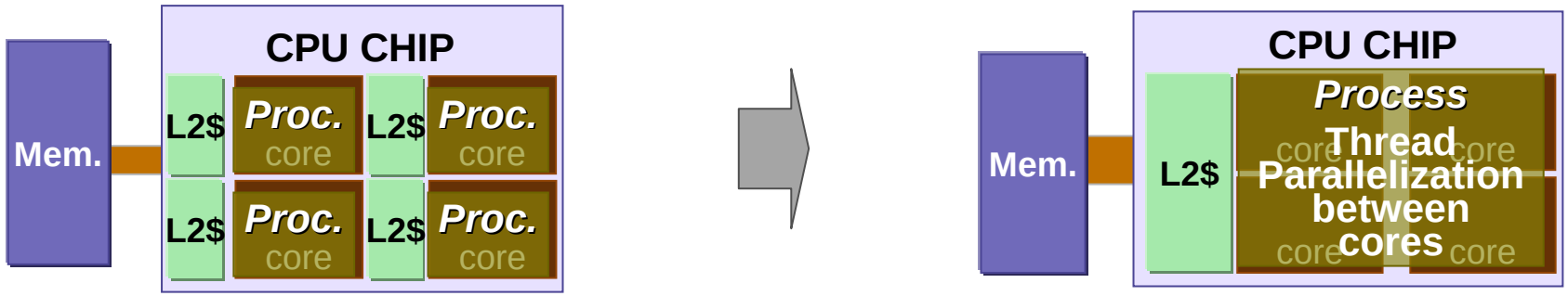
FX1 : New High-End TC Server - Outline -



- **High-performance CPU designed by Fujitsu**
  - SPARC64™ VII : 4 cores by 65 nm technology
  - Performance : 40 GFlops (2.5 GHz)
- **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**
  - FX1 is suitable platform to develop and evaluate Petascale applications

Introduction

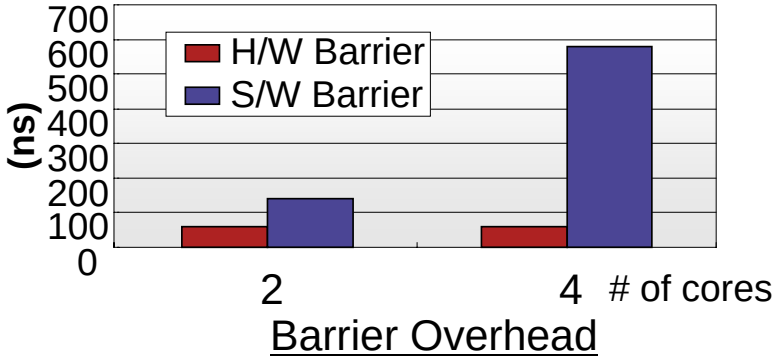
- **Concept**
  - Highly efficient thread level parallel processing technology for multi-core chip



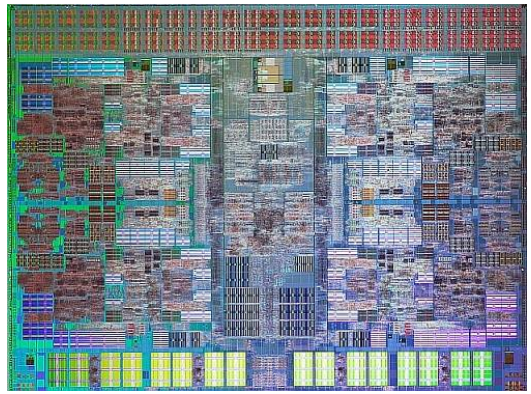
- **Advantage**
  - Handles the multi-core CPU as one equivalent faster CPU
    - ➔ Reduces number of MPI processes to $1/n_{core}$ and increases parallel efficiency
    - ➔ Reduces memory-wall problem
- **Challenge**
  - How to decrease the thread level parallelization overhead?

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 of number of cores



Barrier Overhead



SPARC64™ VII

Real quad-core CPU for Technical Computing

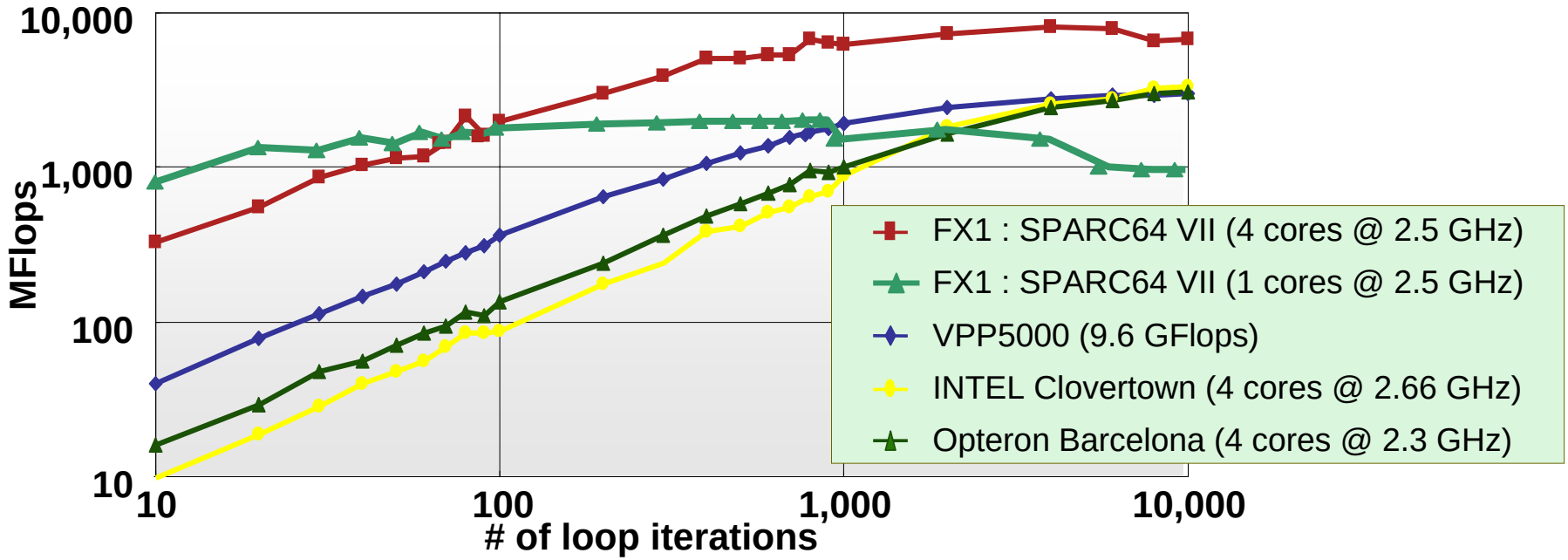
(2.5 GHz, 40 GFlops/chip)

  - Shared L2 cache memory (6 MB)
    - ➔ Reduces the number of cache to cache data transfers
    - ➔ Efficient cache memory usage
- **Compiler technologies**
  - Automatic parallelization or OpenMP on thread-based algorithm by vectorization technology

Integrated Multi-core Parallel ArChiTecture, preliminary measured data

Performance Measurement of Automatic Parallelization

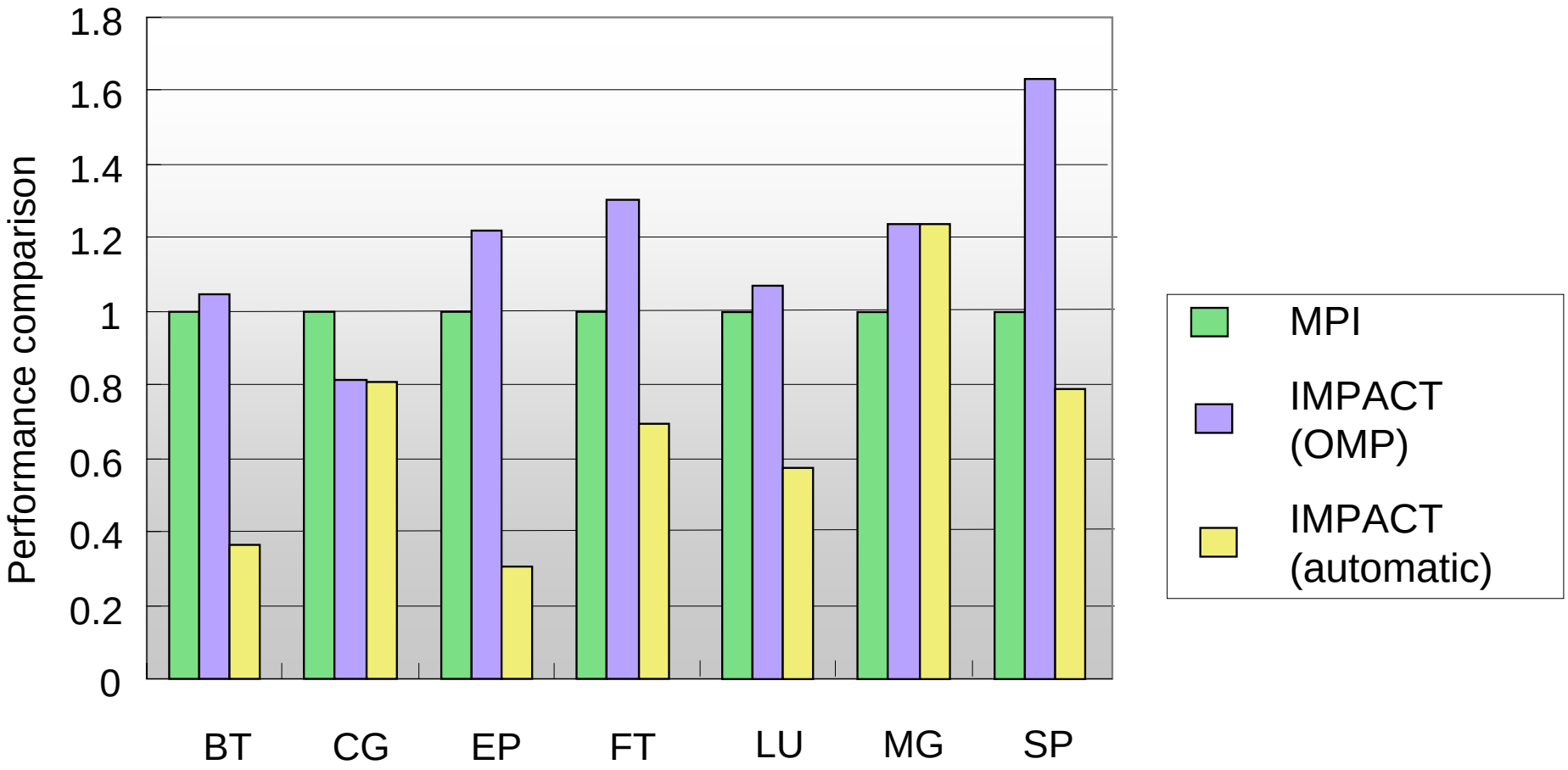
- **LINPACK performance on 1 CPU (4 cores)**
  - n = 100 ➔ 3.26 GFlops
  - n = 40,000 ➔ 37.8 GFlops (93.8%)
- **Performance comparison of DAXPY (EuroBen Kernel 8) on 1 CPU**
  - 4core + IMPACT shows better performance than
    - ➔ 1core performance with small number of loop iterations
    - ➔ X86 servers



Performance of DAXPY

Performance Measurement of NPB on 1 CPU

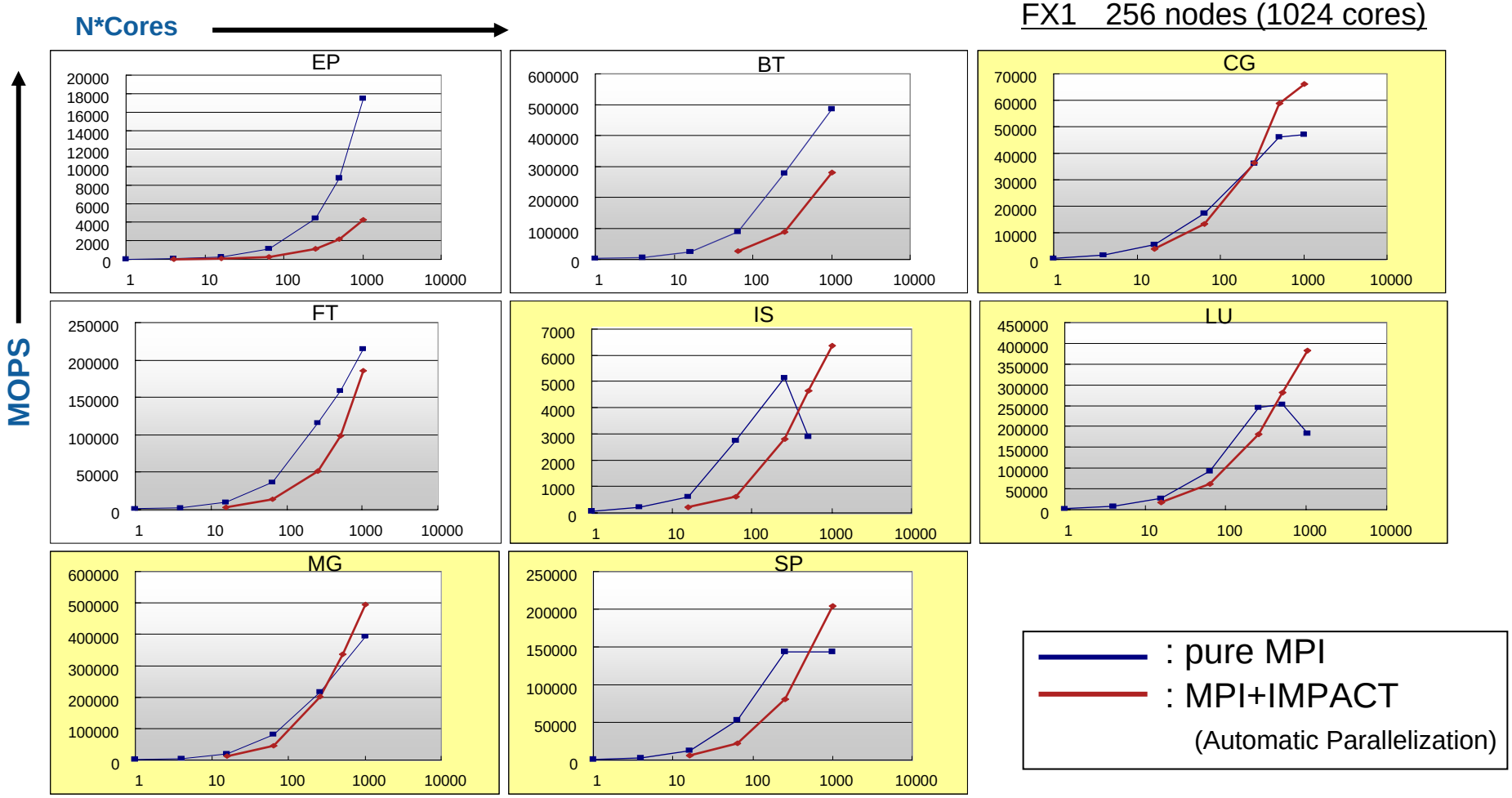
- **Performance comparison of NPB class C between pure MPI and Integrated Multi-core Parallel ArChiTecture on 1 CPU (4 cores)**
  - IMPACT(OMP) is better than pure MPI for 6/7 programs



Performance measurement of NPB on 256 CPUs (1)

- **Performance comparison of NPB class C between pure MPI and MPI + Integrated Multi-core Parallel ArChiTecture**
  - MPI + IMPACT (Automatic Parallelization) is better than pure MPI with 5/8 programs

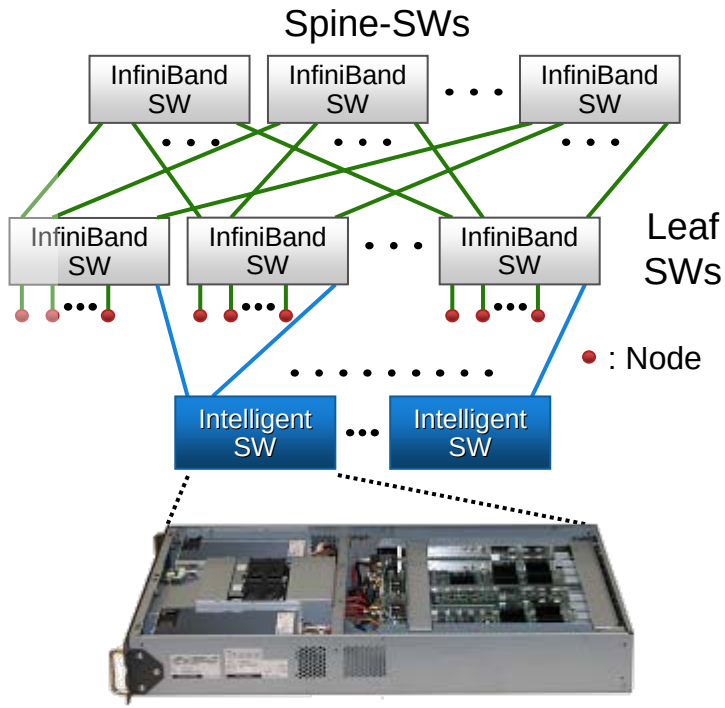
FX1　256 nodes (1024 cores)



FX1 Intelligent Interconnect

Introduction

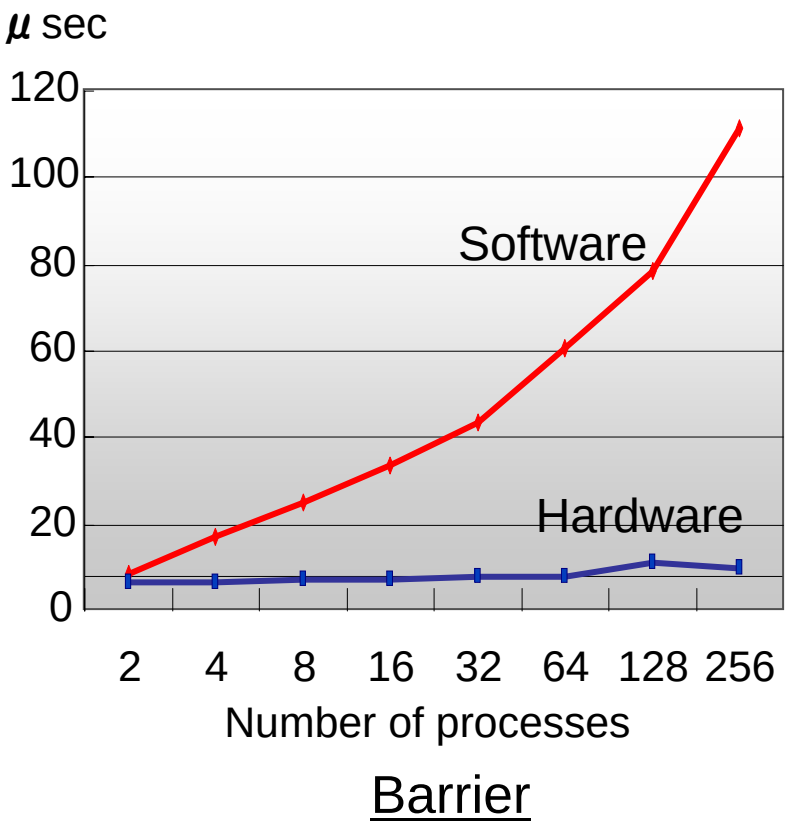
- **Combination of fat tree topology InfiniBand DDR interconnect and the highly-functional switch (Intelligent switch)**
- **Intelligent switch**
  - Result of the PSI (Petascale System Interconnect) national project
  - Functions
    - Hardware barrier function among nodes
    - Hardware assistance 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

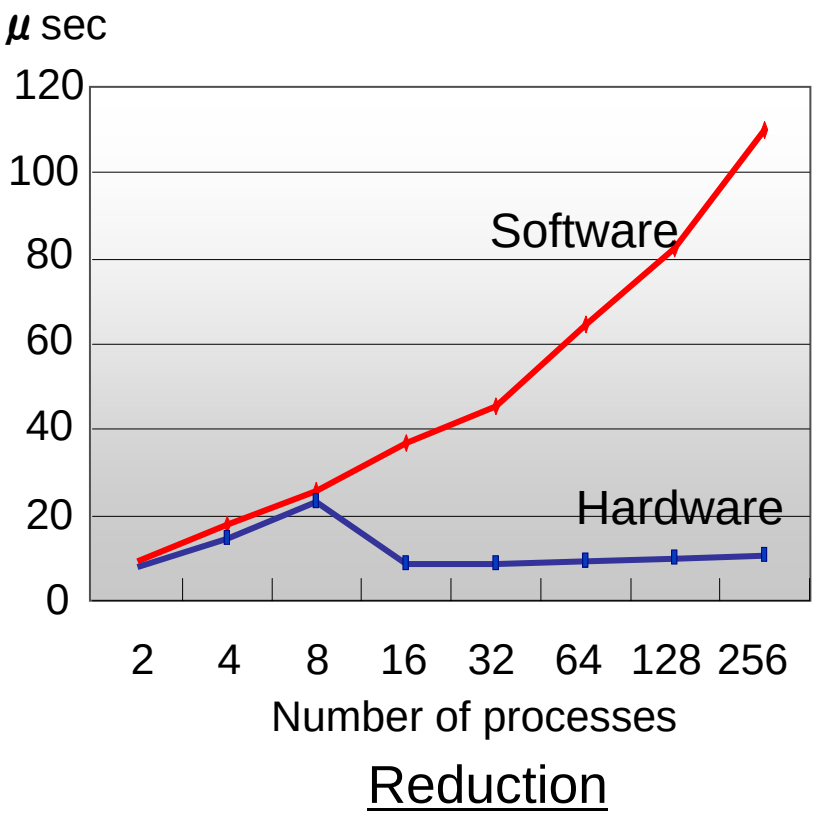


Intelligent Switch & its connection

High Performance Barrier & Reduction Hardware

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



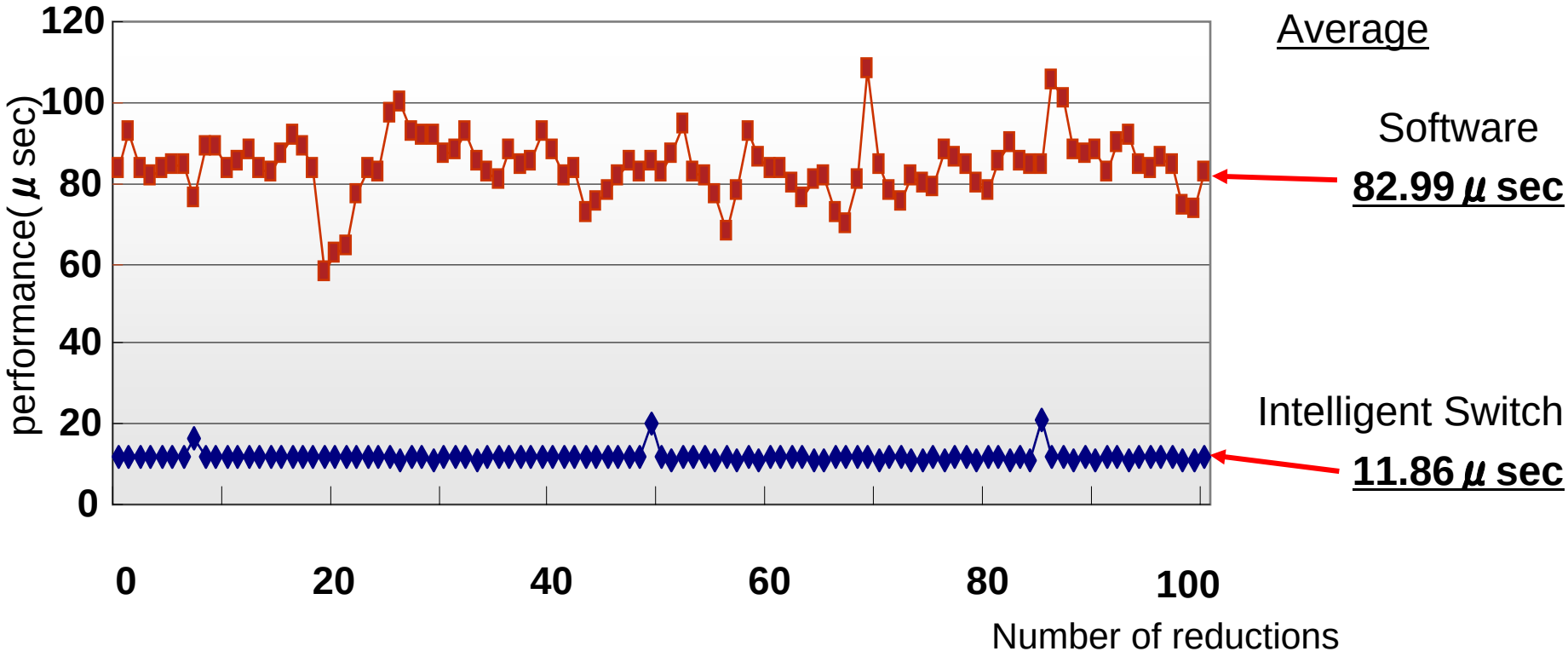


*** :** Executed by host processor using butterfly network built by point to point communication.

FX1 Intelligent Interconnect

Stability of Reduction Function

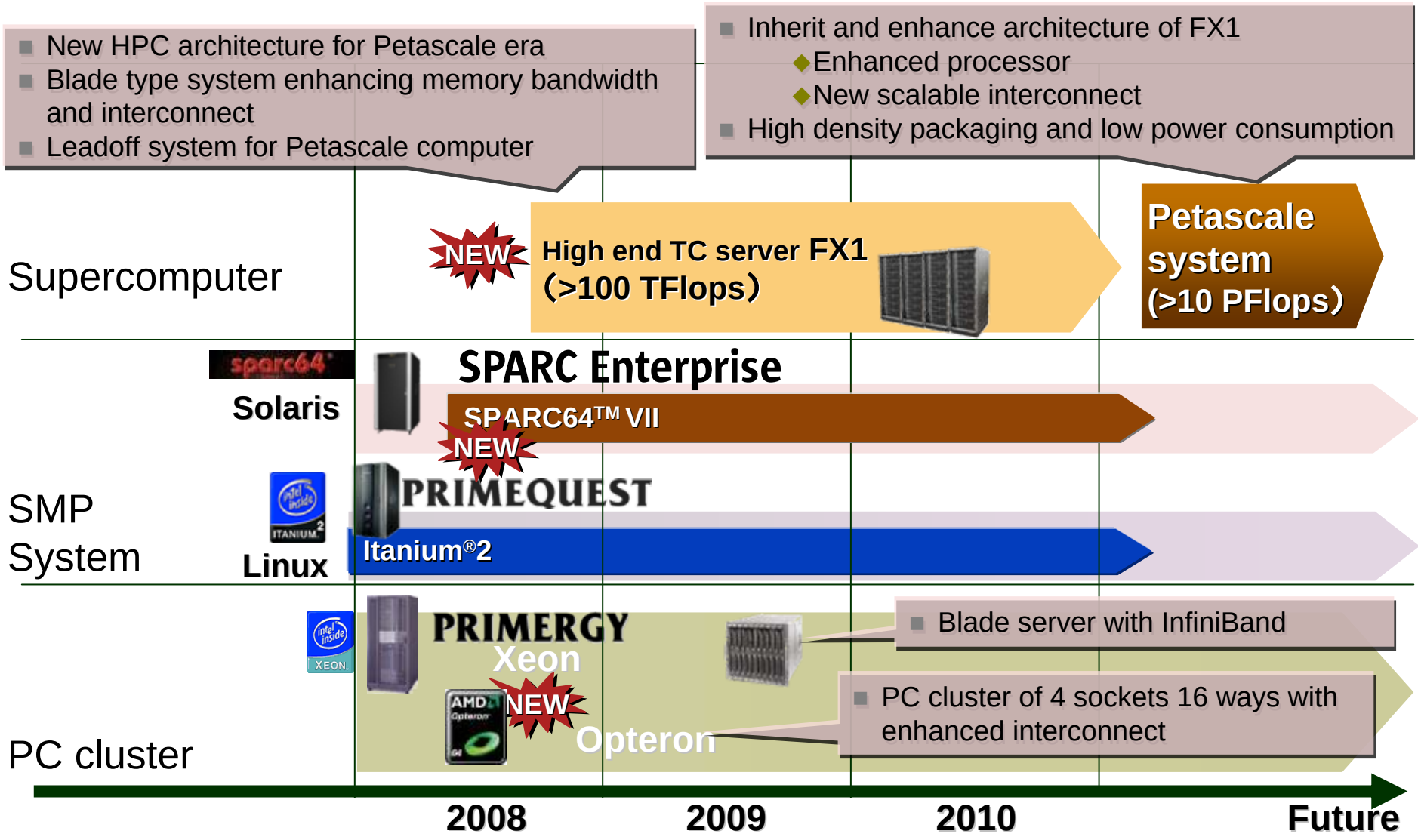
- **Intelligent interconnect realizes stable reduction performance by global ping function**



Reduction (All reduce) performance on 128 node system

Technical Computing Server Roadmap

- **Development of the commodity based server and of the proprietary high end server for Technical Computing**



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

~$1.2 B

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

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).

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.

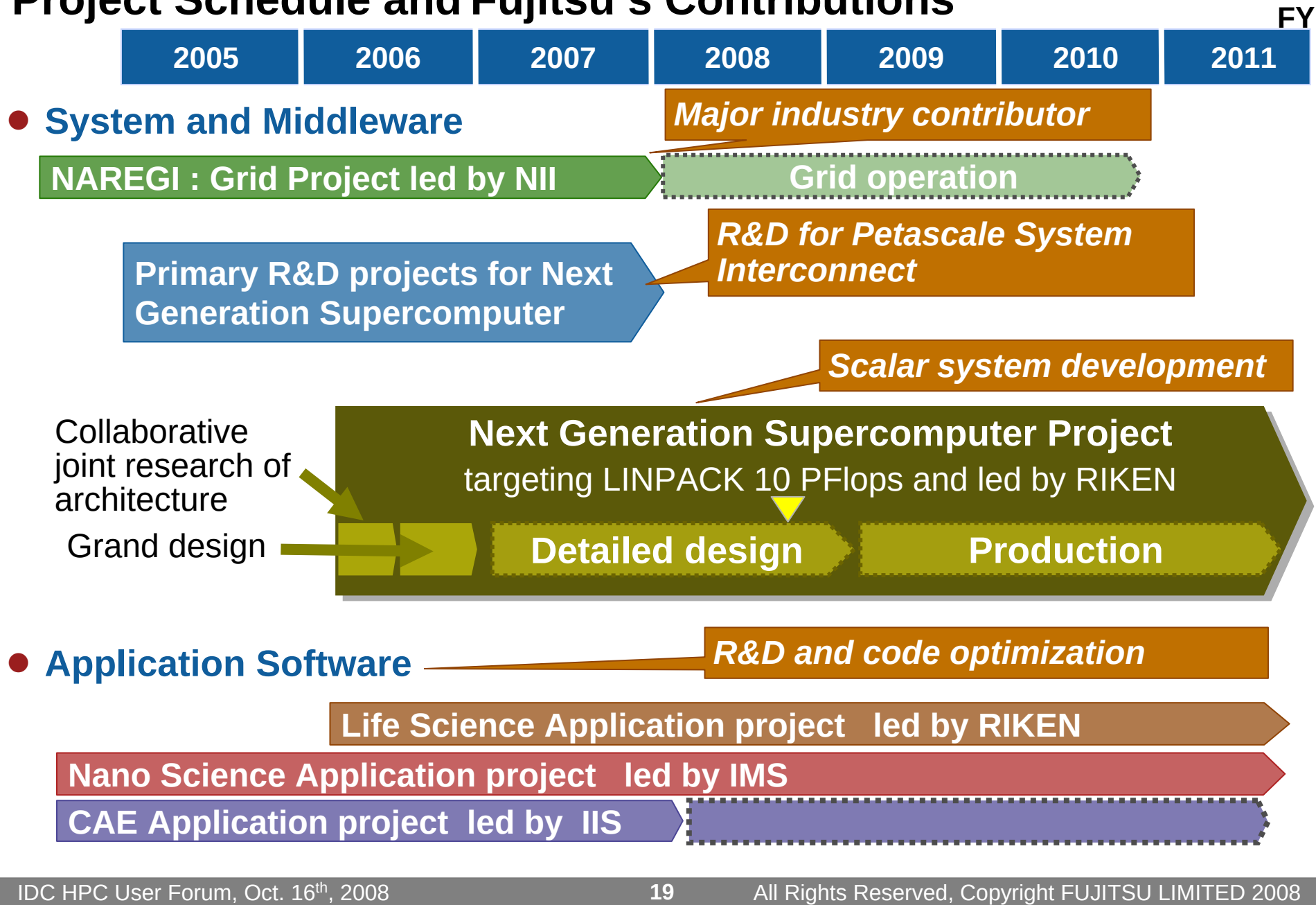
(1) Development and implementation of the world's most advanced high-performance Next-Generation supercomputer
(2) Development and dissemination of software that makes optimum use of the supercomputer
(3) Establishment of the world's most advanced and highest standard supercomputing Center of Excellence, 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

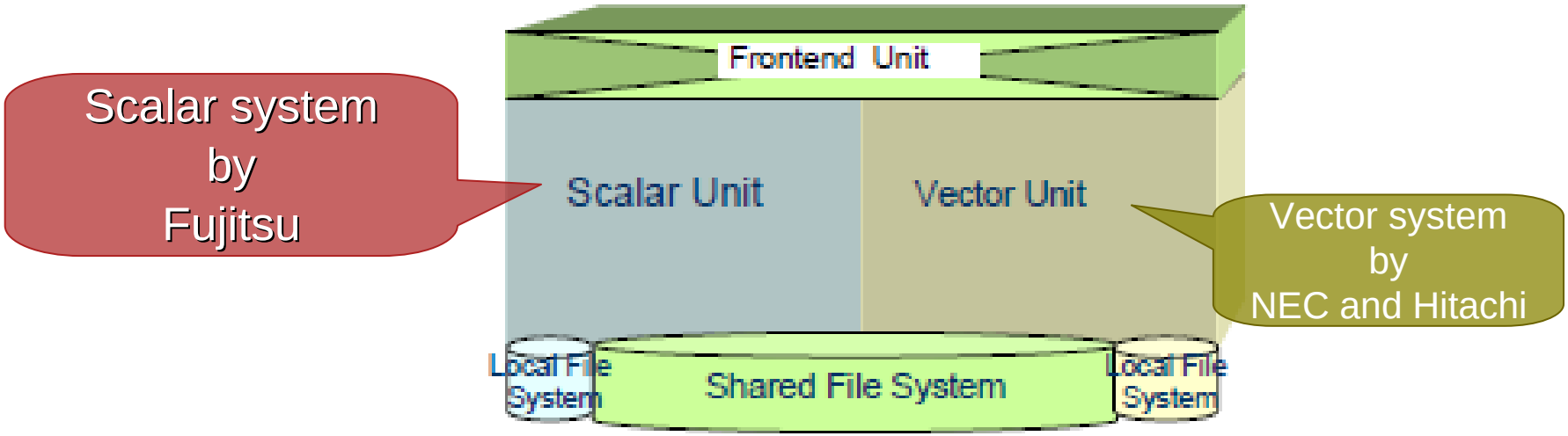
Project Schedule and Fujitsu's Contributions



Japanese Next Generation Supercomputer Project

Project Outline

- **System configuration**
  - The hardware system consists of scalar and vector processor units
- **The target performance**
  - 10 PFlops 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

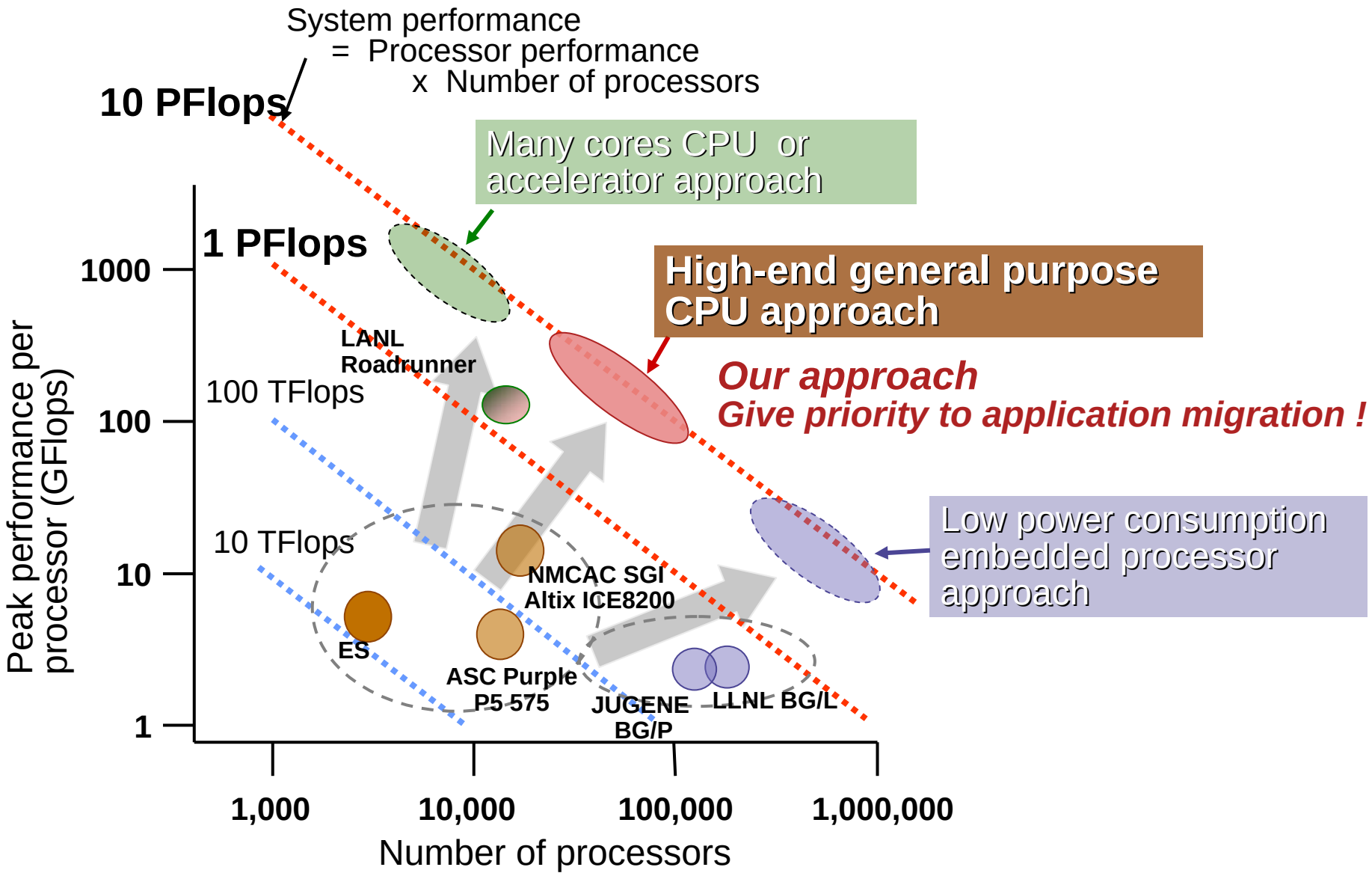


Source: CSTP evaluation working group report

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*

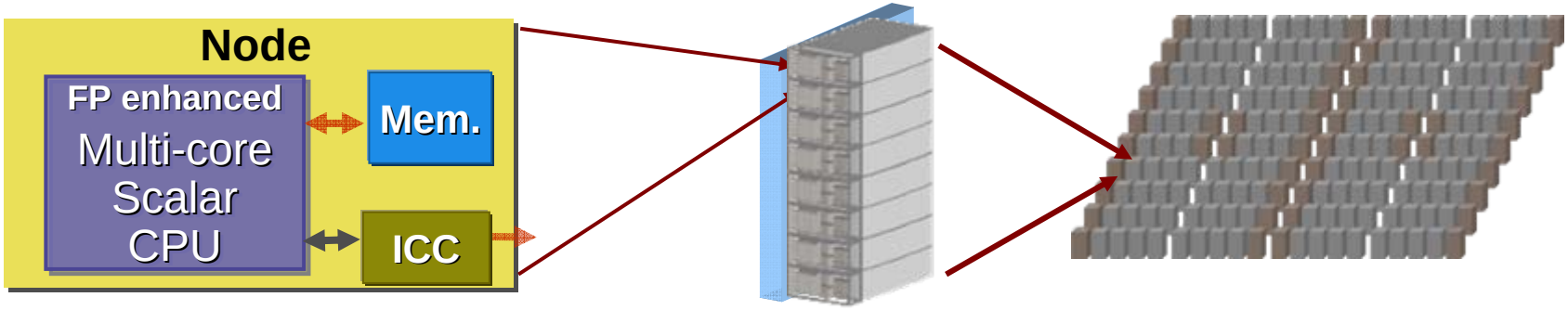
- FP enhanced multi-core scalar CPU (over 100 GFlops/CPU) with mainframe level reliability
- Inherit Integrated Multi-core Parallel ArChiTecture of the FX1
- Low power consumption, targeting ~1/10 power consumption per flop

Leading edge interconnect

- 3D torus interconnect with scalability up to over 10 PFlops, 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

Mainframe Processor

SPARC64™ Processor

SPARC64VII — CMOS Cu+Low-k 65 nm

SPARC64VI — Tr = 540M CMOS Cu+Low-k 90 nm

GS21 — Tr = 500M CMOS Cu + Low-k **90 nm**

SPARC64V+ — Tr = 400M CMOS CU + Low-k **90 nm**

GS21 — Tr = 200M CMOS Cu **130 nm**

SPARC64V — Tr =2 00M CMOS Cu **130 nm**

GS8900 — Tr = 45M CMOS Cu 180 nm

GS8800B — Tr = 30M CMOS Al 250 nm / 220 nm

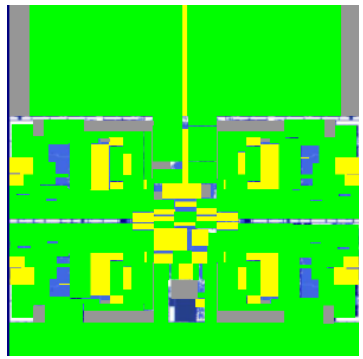
SPARC64GP — Tr = 30M CMOS Cu 180 nm / 150 nm

GS8800

GS8600 — Tr = 10M CMOS Al 350 nm

SPARC64II

SPARC64



RAS Coverage of SPARC64VII Guaranteed Data Integrity

: 1bit error correctable

: 1bit error detectable

: 1bit error harmless

~1995 | 1996 ~1997 | 1998 ~1999 | 2000 ~2003 | 2004~

Interconnect for Parallel Computer System

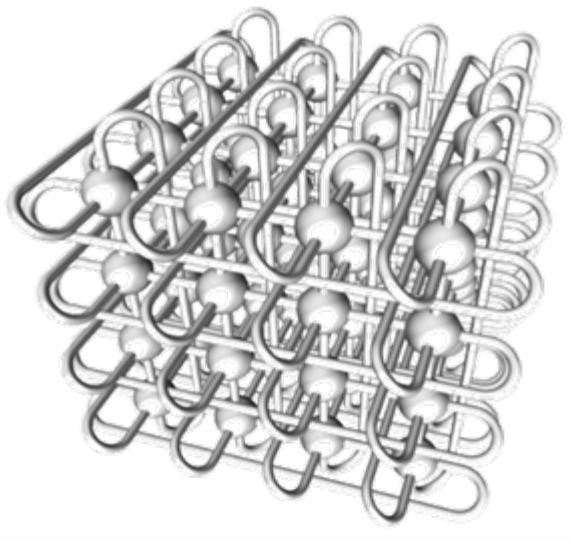
- **Interconnect type and characteristics**

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

- **Targeting over 100,000 nodes parallel system**
  - Cost, packaging density and power consumption are essential issues
  - Too many hops are needed for mesh interconnect
    - ➔ Torus interconnect is a strong candidate
    - ➔ The greatest challenge of torus interconnect is operability and usability
- **Fujitsu's challenge is 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 to over 100,000 nodes
  - High reliability and availability
  - High density packaging
  - Reduced wiring cost
  - Simple 3D torus logical (application) view

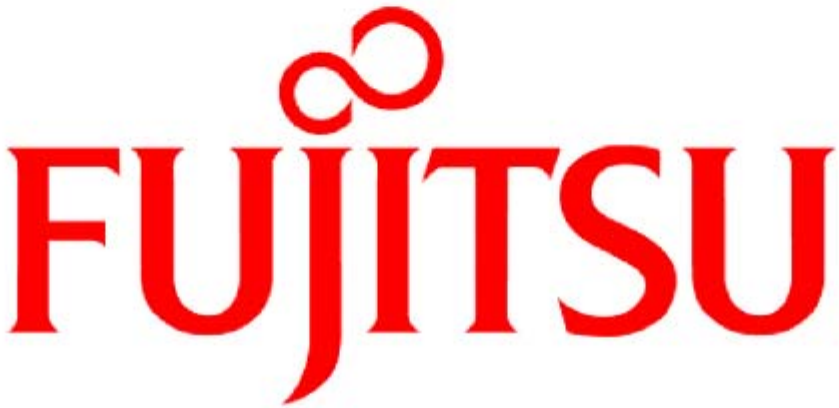


Agenda

- 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**



THE POSSIBILITIES ARE INFINITE