FUJITSU

shaping tomorrow with you

Fujitsu's Technologies Leading to Practical Petascale Computing: *K computer*, *PRIMEHPC FX10* and the Future

November 16th, 2011

Motoi Okuda

Technical Computing Solution Unit

Fujitsu Limited

Agenda

FUJITSU

- Achievements of ***K computer***
  - What did K computer achieve?
  - Technologies which realize the achievement
- Fujitsu's new Petascale supercomputer ***PRIMEHPC FX10***
  - Fujitsu supercomputer past and present
  - Second generation Petascale supercomputer PRIMEHPC FX10
- Conclusion

Latest World Supercomputer Ranking

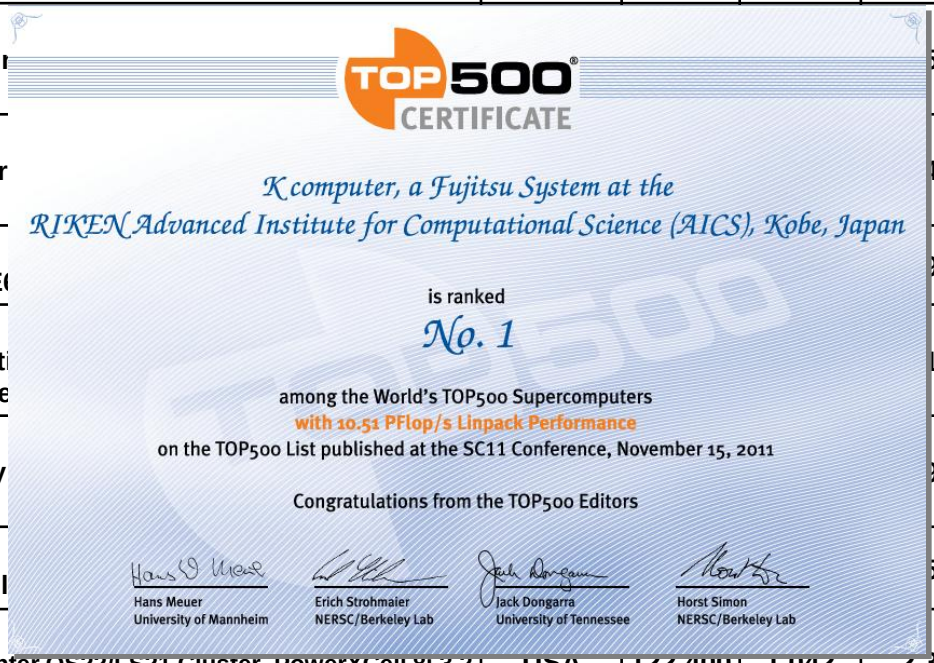
FUJITSU

- K computer takes consecutive No.1 on TOP500 List

38th List: TOP10

November 2011

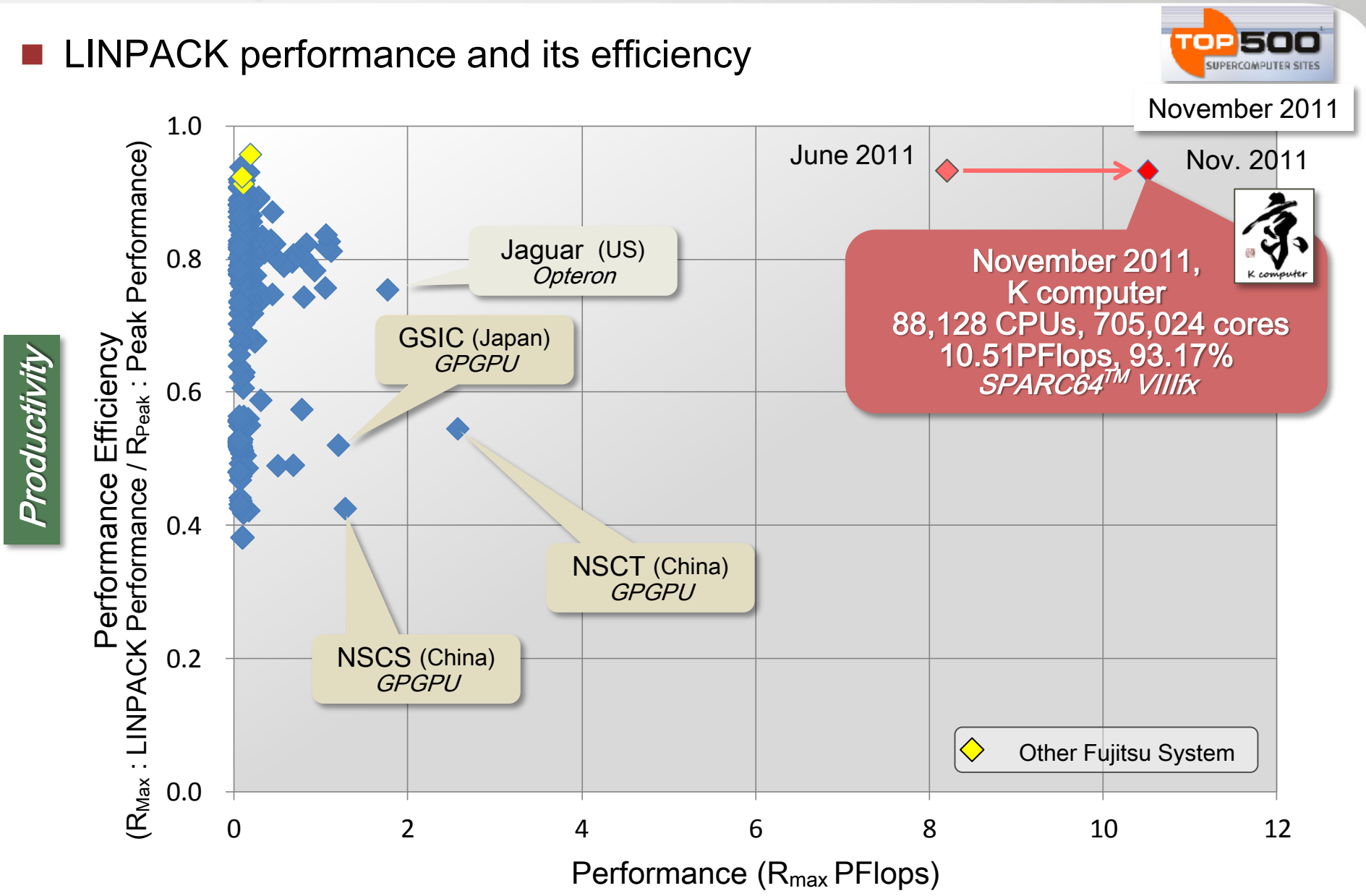
Rank	Site	Manufacture	Computer	Country	Cores	Rmax [Pflops]	Power [MW]
1	RIKEN Advanced Institute for Computational Science (AICS)	Fujitsu	K computer SPARC64 VIIIfx 2.0GHz,Tofu interconnect	Japan	705,024	10.510	12.66
2	National Supercomputing Center in Tianjin	NUDT	Tianhe-1A NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050	China	186,368	2.566	4.04
3	DOE/SC/Oak Ridge National Laboratory	Cray Inc.	Jaguar Cray XT5-HE Opteron 6-core 2.6 GHz	USA	224,162	1.759	6.95
4	National Supercomputing Centre in Shenzhen (NSCS)	Dawning	Dawni				58
5	GSIC Center, Tokyo Institute of Technology	NEC/HP	HP Pr				40
6	DOE/NNSA/LANL/SNL	Cray Inc	Cray XE				98
7	NASA/Ames Research Center/NAS	SGI	SGI Alt 3.0/Xe				10
8	DOE/SC/LBNL/NERSC	Cray Inc.	Cray				91
9	Commissariat a l'Energie Atomique (CEA)	Bull	Bul				59
10	DOE/NNSA/LANL	IBM	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband	USA	122,400	1.042	2.35



The K computer's Performance

FUJITSU

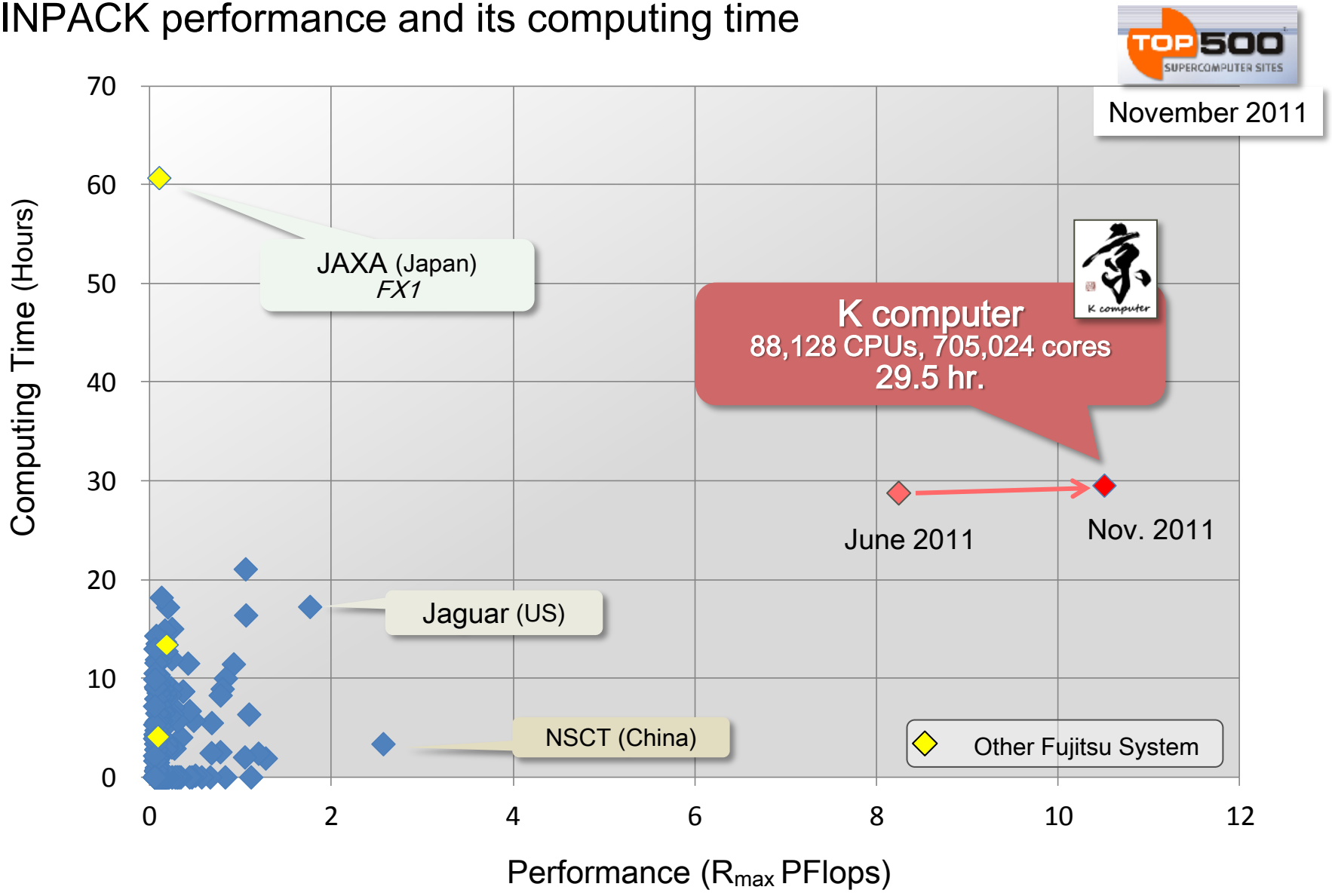
- LINPACK performance and its efficiency



The K computer's Performance (cont.)

FUJITSU

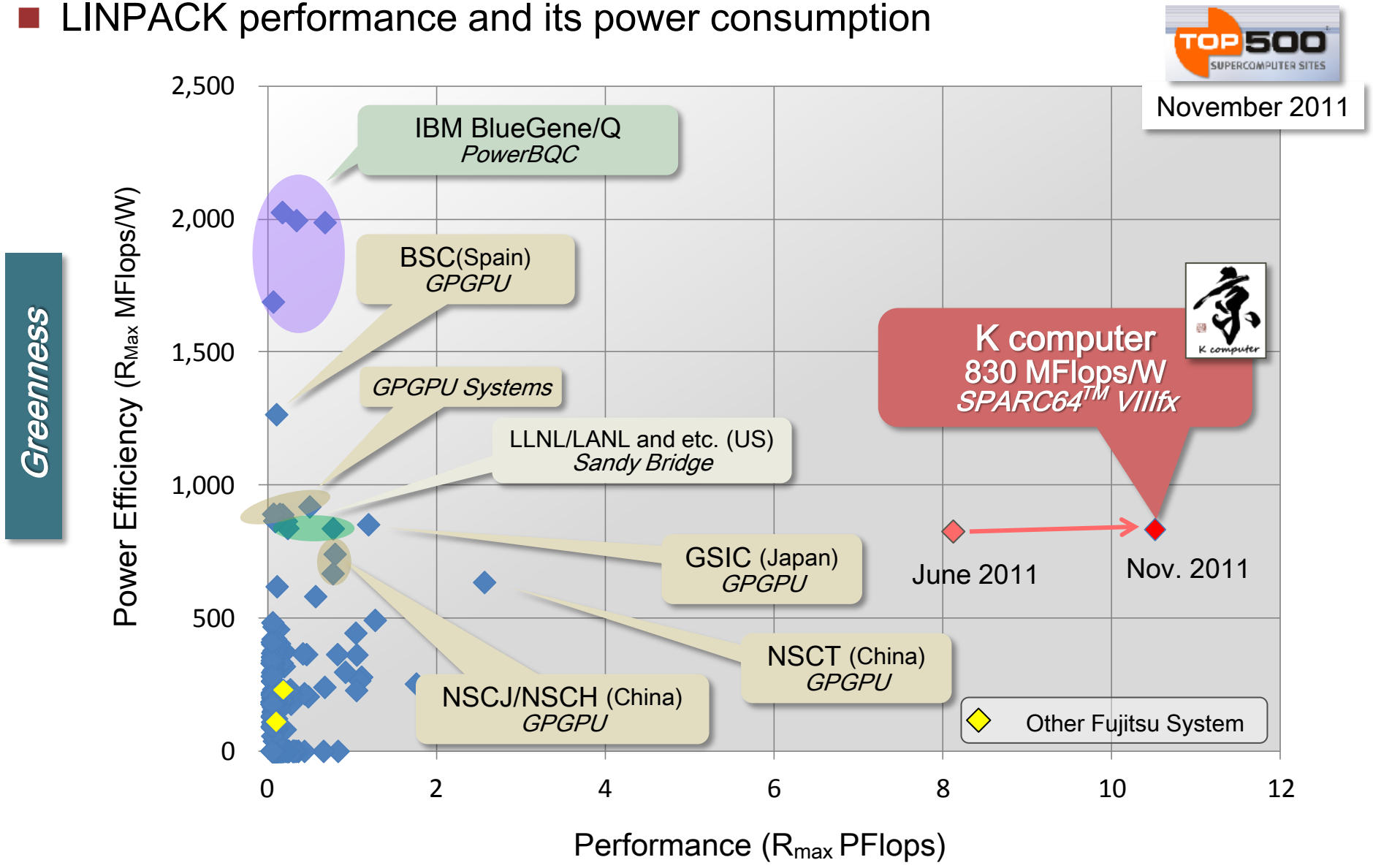
- LINPACK performance and its computing time



The K computer's Performance (cont.)

FUJITSU

- LINPACK performance and its power consumption



The K computer's Performance (cont.)

FUJITSU

- HPCC optimized run performance by 192 racks (2.36PFlops) K computer
- HPCC class 1 Awards

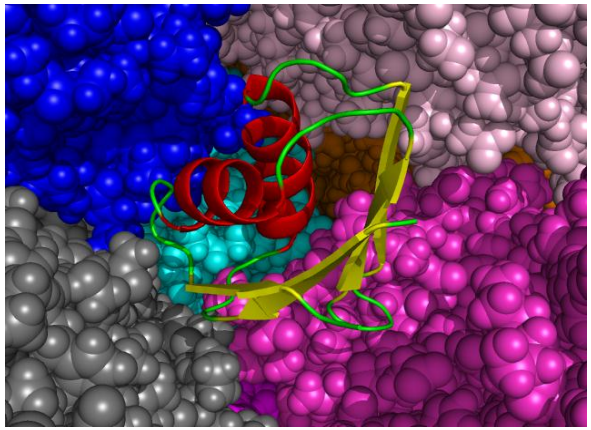
Program	Measured performance	Ranking
G-HPL(PFlops)	**2.118**	***#1***
G-Random Access (Gups)	**121.1**	***#1***
G-FFT (Tflops)	**34.72***	***#1***
EP-STREAM (PB/s)	**0.812**	***#1***

* : using FFTE5.0β developed by Prof . Takahashi of Tsukuba Univ.

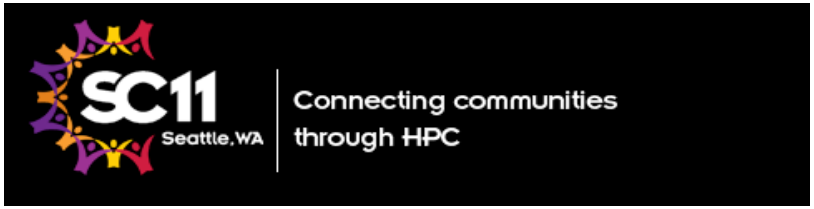
Applications Performance , not a paper tiger

FUJITSU

- March 2011
  - cppmd molecular dynamics
    - 181 million atom simulation
    - Running on 221,184 cores (3.54 PFlops)
    - Sustained performance of ***1.316 PFLOPS***
    - Efficiency of ***37%***



- 2011 Gordon Bell Award finalist
  - RIKEN, University of Tsukuba, University of Tokyo, Fujitsu
    - RSDFT program : first-principles calculations of electron states of a silicon nanowire
      Hasegawa, Iwata, Tsuji, Takahashi, Oshiyama, Minami, Boku, Shoji, Uno, Kurokawa, Inoue, Miyoshi, Yokokawa
    - 107,292-atom Si nanowire calculation
    - Running on 442,368 cores (7.08PFlops)
    - Sustained performance of ***3.08 PFLOPS***
    - Efficiency of ***43.6 %***

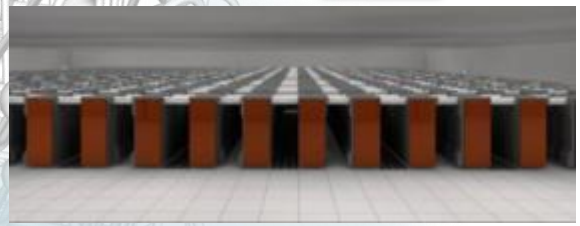


- Various application programs are in optimization and evaluation stage

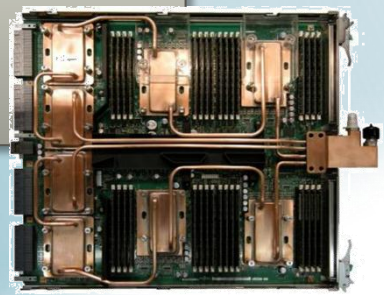
K computer outline and technologies

FUJITSU

- New Interconnect , *Tofu*
  - 6-dimensional Mesh/Torus topology
  - High speed, highly scalable, high operability and high availability interconnect for over 100,000 nodes system
  - Functional interconnect

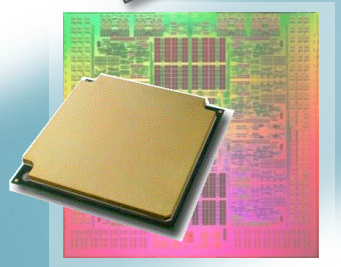


- ***SPARC64™ VIIIfx CPU***
  - Fujitsu designed and fabricated CPU
  - 8 cores, 128GFlops@2GHz
  - HPC-ACE (SPARC V9 Architecture Enhancement for HPC) :128GFlpos
  - Main frame CPU level of high reliable design
  - Low power consumption : ~58W



LINPACK 10.51 PFlops
17 .6 PB
864 racks
88,128 CPUs
705,024 cores

RIKEN FUJITSU



- Single CPU per node configuration
  - High memory bandwidth and simple memory hierarchy
- CPU/ICC direct water cooling
  - High reliability, low power consumption and compact packaging

Agenda

FUJITSU

- Achievements of ***K computer***
  - What did K computer achieve?
  - Technologies which realize the achievement
- Fujitsu's new Petascale supercomputer ***PRIMEHPC FX10***
  - Fujitsu supercomputer past and present
  - Second generation Petascale supercomputer PRIMEHPC FX10
- Conclusion

Fujitsu HPC Servers

FUJITSU

Petascale Supercomputer Series

FX10

No.1 in Top500
(June and Nov., 2011)

K computer

Vector Supercomputer Series

FX1

Most Efficient Performance in Top500 (Nov. 2008)

World's Fastest Vector Processor (1999)

VPP5000

SPARC Enterprise

NWT*
Developed with NAL

©JAXA

No.1 in Top500
(Nov. 1993)
Gordon Bell Prize
(1994, 95, 96)

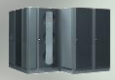
VPP300/700

PRIMEQUEST

Scalar Supercomputer Series

PRIMERGY CX400
Skinless server
(coming soon)

PRIMEPOWER HPC2500

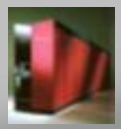


World's Most Scalable Supercomputer (2005)

PRIMERGY BX900
Cluster node

VPP500

VP Series



Scalar MPP Series

AP3000

HX600
Cluster node

F230-75APU

AP1000

PRIMERGY RX200
Cluster node

Japan's Largest Cluster in Top500
(July 2004)

x86 HPC Cluster Series

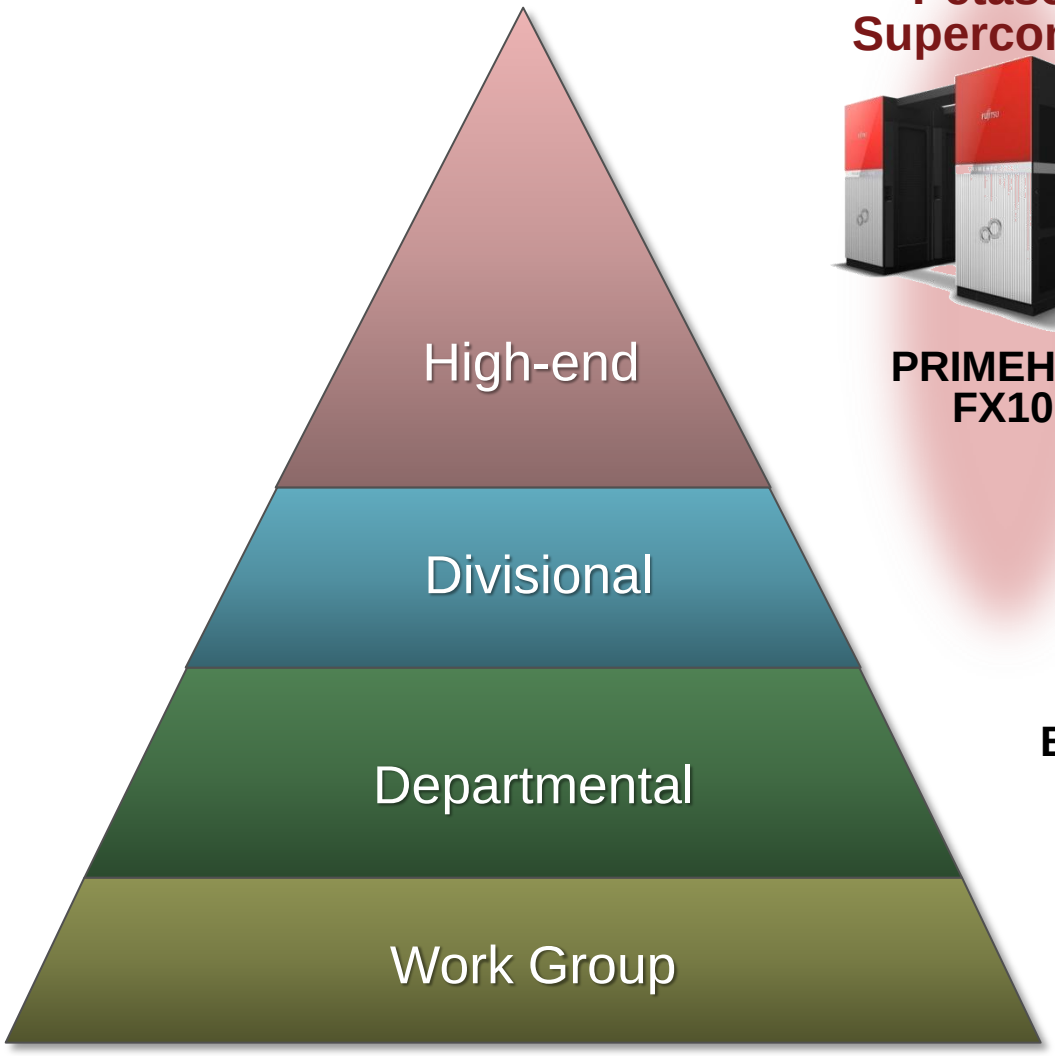
Japan's First Vector (Array) Supercomputer (1977)

*NWT:
Numerical Wind Tunnel

~1985 1990 1995 2000 2005 2010 2015

FUJITSU

- Full range coverage with choice of HPC platform



Petascale Supercomputer

PRIMEHPC FX10

x86 HPC Cluster

CX400
Skinless server (coming soon)

BX Series
BX900

CX1000

Large-Scale SMP System

RX900

RX Series
RX200

PRIMERGY series

Design targets and features of FX10

FUJITSU

- High Performance
  - High peak performance and high effective performance

- Highly parallel application productivity
  - Easy to extract high performance from the highly paralleled programs without inordinate burden to programmers

User requirement of over PFlops computing

➡ *FX10 design targets*

- High operability
  - Low power consumption
  - High reliability and easy to operate

- K computer compatibility
  - Binary compatibility
  - Same programing environment

Design targets and features of FX10

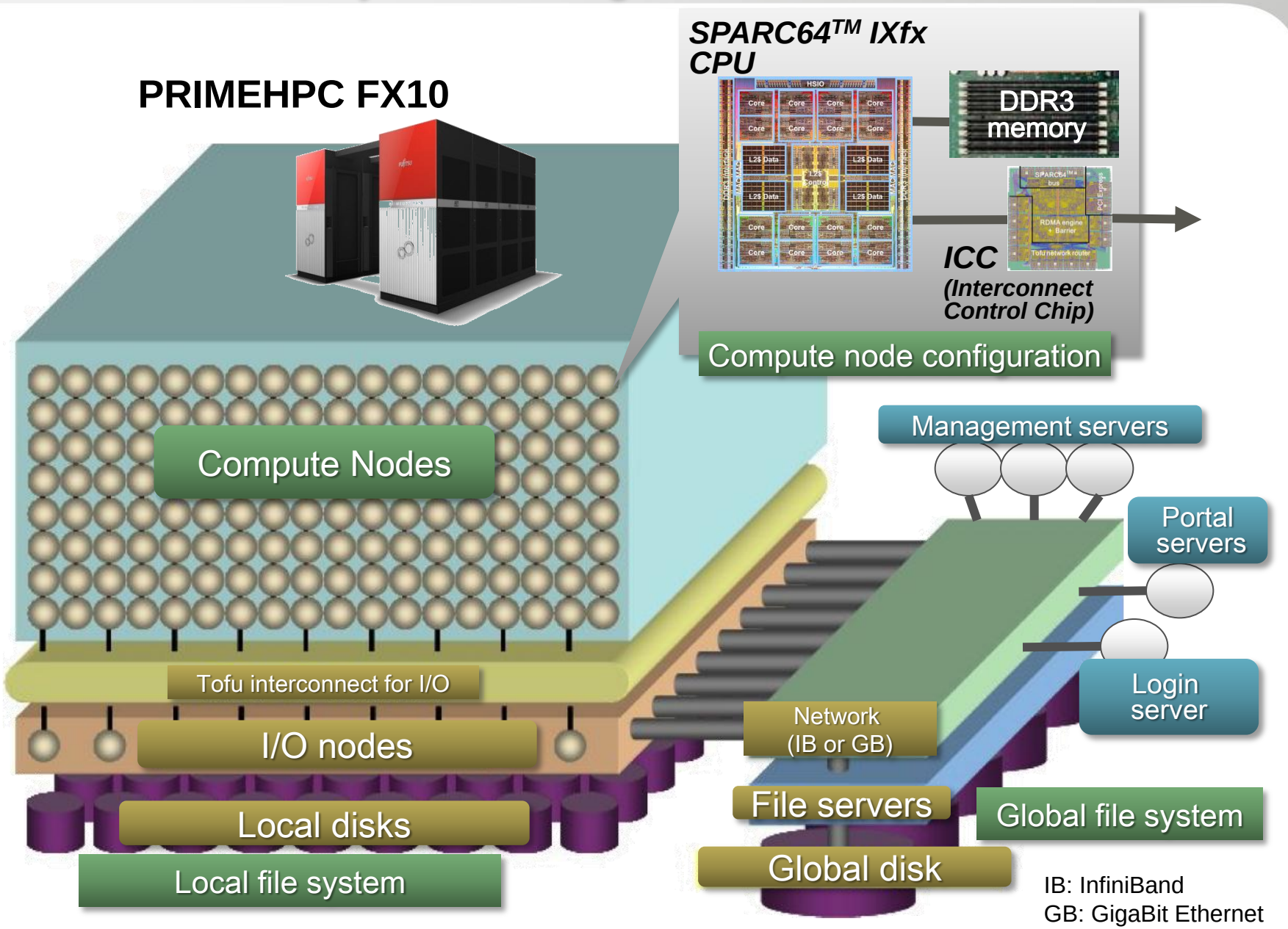
FUJITSU

- High Performance
  - High-performance CPU "*SPARC64 IXfx*" with SPARC V9 + HPC-ACE architecture
  - High performance, highly reliable and fault tolerant 6D mesh/torus interconnect "Tofu*1"
  - Water cooling system
- High parallel application productivity
  - Easy to extract high
  - "*VISIMPACT*"*2 supports efficient hybrid parallel execution
  - hordinate burden to
  - ammers
  - Parallel Language, programing tools and Petascale HPC middleware for high reliability and operability
- High operability
  - Low power consumption
  - High reliability and easy to operate
- K computer compatibility
  - Binary compatibility
  - Same programing
  - onment
- High reliability components & functions based on mainframe development experience

*1) Tofu: Torus Fusion
*2) VISIMPACT: Virtual Single Processor by Integrated Multicore Parallel Architecture. (The feature uses multiple cores as a single high-speed CPU)

PRIMEHPC FX10 System Configuration

FUJITSU



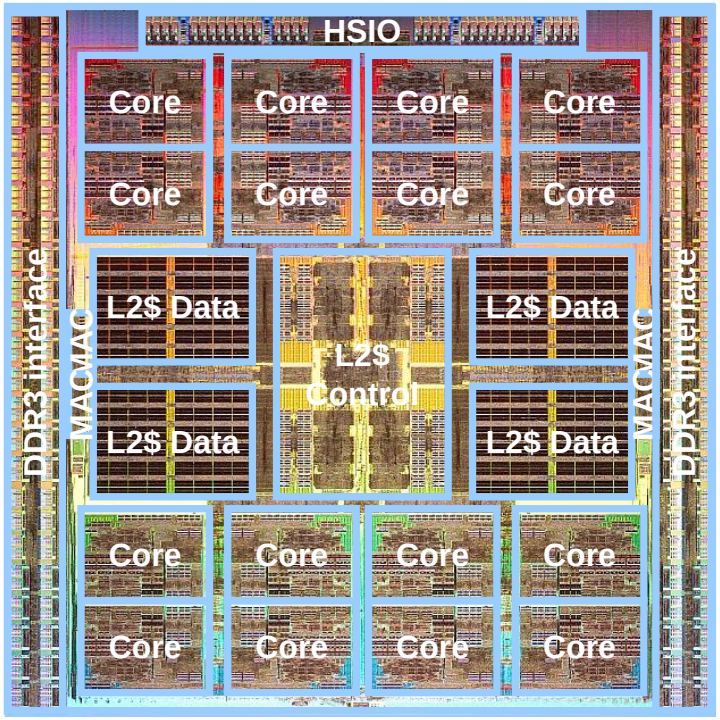
SPARC64™ IXfx

FUJITSU

- High-performance and low-power multi-core CPU
  - High performance core by HPC-ACE
    - Register # extension, SIMD operation, software controllable cache, ・・・
  - VISIMPACT : Support highly efficient hybrid execution model (thread + process)
    - Shared 2nd cache, hardware barriers among cores and compiler

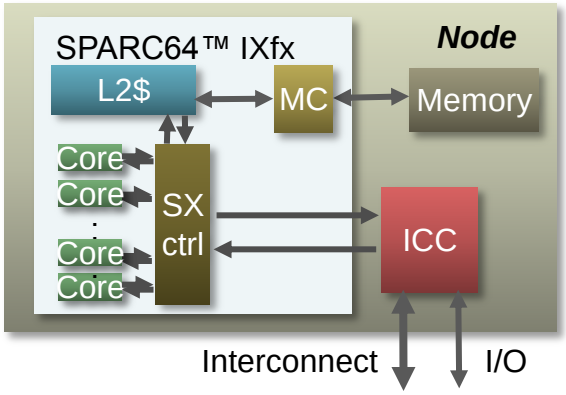
SPARC64™ IXfx specifications	
Architecture	SPARC64™ V9 + HPC-ACE
# of FP operations /clock/core	8 (= 4 Multiply and Add)
No. of cores	16
Peak performance and clock	236.5 Gflops@1.848GHz
Memory bandwidth	85 GB/s
Power consumption	110 W (typical)

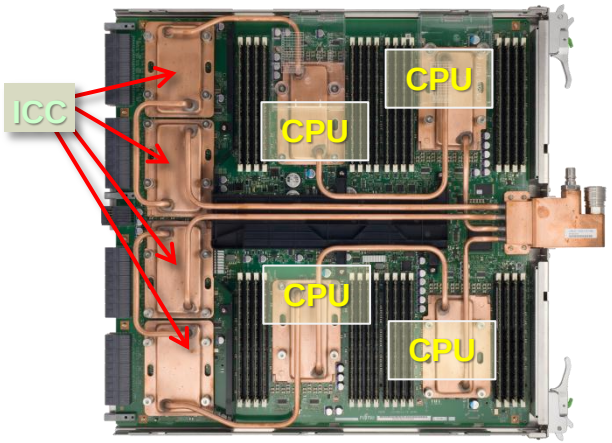
  - High performance-per-power ratio and High reliability
    - Water cooling system has lowered the CPU temperature and leak current
    - Wide-ranging error detection/self-recovery functions, instruction retry function



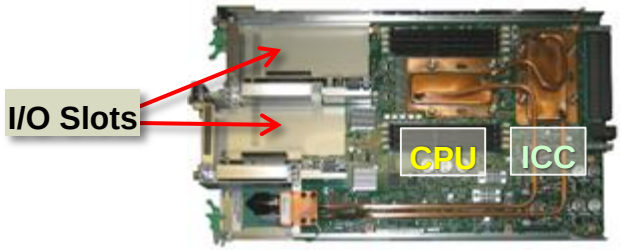
Node Configuration

- Single CPU as a node design
  - SPARC64™ IXfx based
  - 32 or 64 GB memory capacity
  - Single CPU per node to maximize memory bandwidth
  - High memory bandwidth of 85 GB/s
- On board InterConnect Controller (ICC)
  - Direct RDMA and global synchronization operations
  - No external switch
- Node type
  - Compute node
    - Consist of CPU, ICC and memory
    - Without I/O capability
    - Four nodes are mounted on a SB (system board)
  - I/O node
    - Same CPU as compute node
    - Includes four PCI Express Gen2 x8 slots
    - 8 GB/s I/O bandwidth per I/O node
    - One node is mounted on an I/O SB (system board)





System Board

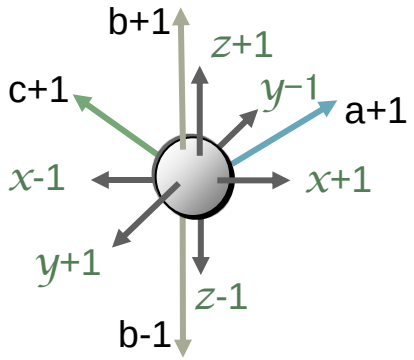


I/O SB

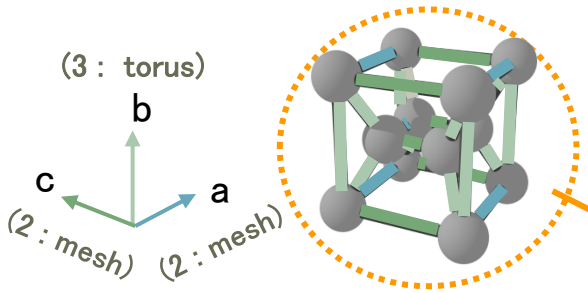
New Interconnect : Tofu

FUJITSU

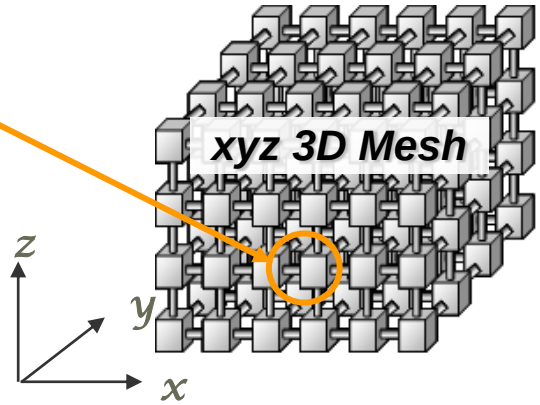
- Design targets
  - Scalabilities toward 100K nodes
  - High operability and usability
  - High performance



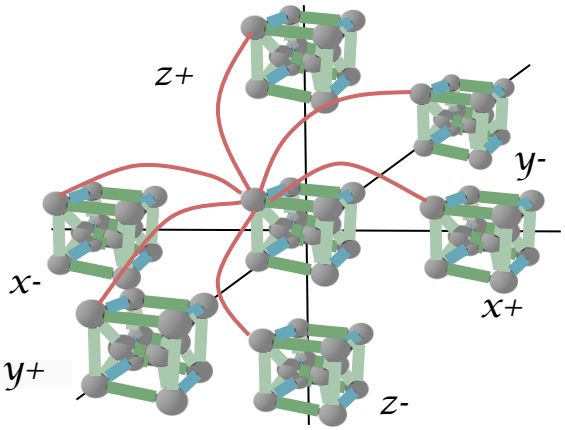
- Topology and performance
  - User view/Application view : Logical 3D Torus (X, Y, Z)
  - Physical topology : 6D Torus / Mesh addressed by (x, y, z, a, b, c)
    - 10 links / node, 6 links for 3D torus and 4 redundant links
  - Performance : 5GB/sec. /link x 2 (bi-directional)



Node Group Unit
(12 nodes group, 2 x 3 x 2)



Node Group 3D connection



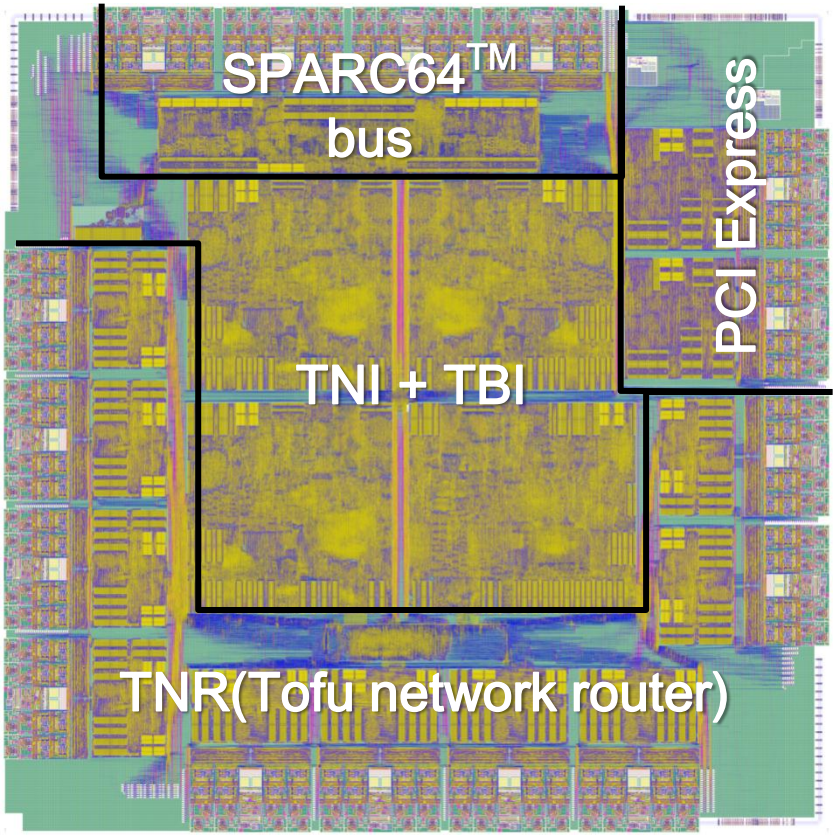
3D connection of each node

ICC : Tofu Interconnect Controller

FUJITSU

- ICC integrates Tofu interconnect and PCI Express I/F
- Tofu interconnect
  - Tofu network router (TNR) : 10 x TNRs transfer packets among ICCs,
  - Tofu network interface (TNI) : 4 x RMDA communication engines
  - Tofu barrier interface (TBI) : Collective communication capability(Barrier and Allreduce)
- High reliable design

ICC specifications	
# of concurrent connections	4 transmission + 4 reception
Switching capacity	100 GB/s
Link speed x number of ports	5 GB/s x Bidirectional x 10 ports
Operating frequency	312.5 MHz
PCI Express	5 Gbps x 16 lanes



FX10 Software stack

FUJITSU

Applications

HPC Portal / System Management Portal

Technical Computing Suite

System Management

- System management
- System control
- System monitoring
- System operation support

Job Management

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

High Performance File System *FEFS*

- Lustre based high performance distributed file system
- High scalability, high reliability and availability

Automatic parallelization compiler

- Fortran
- C
- C++

Tools and math. libraries

- Programming support tools
- Mathematical libraries (SSL II/BLAS etc.)

Parallel languages and libraries

- OpenMP
- MPI
- XPFortran

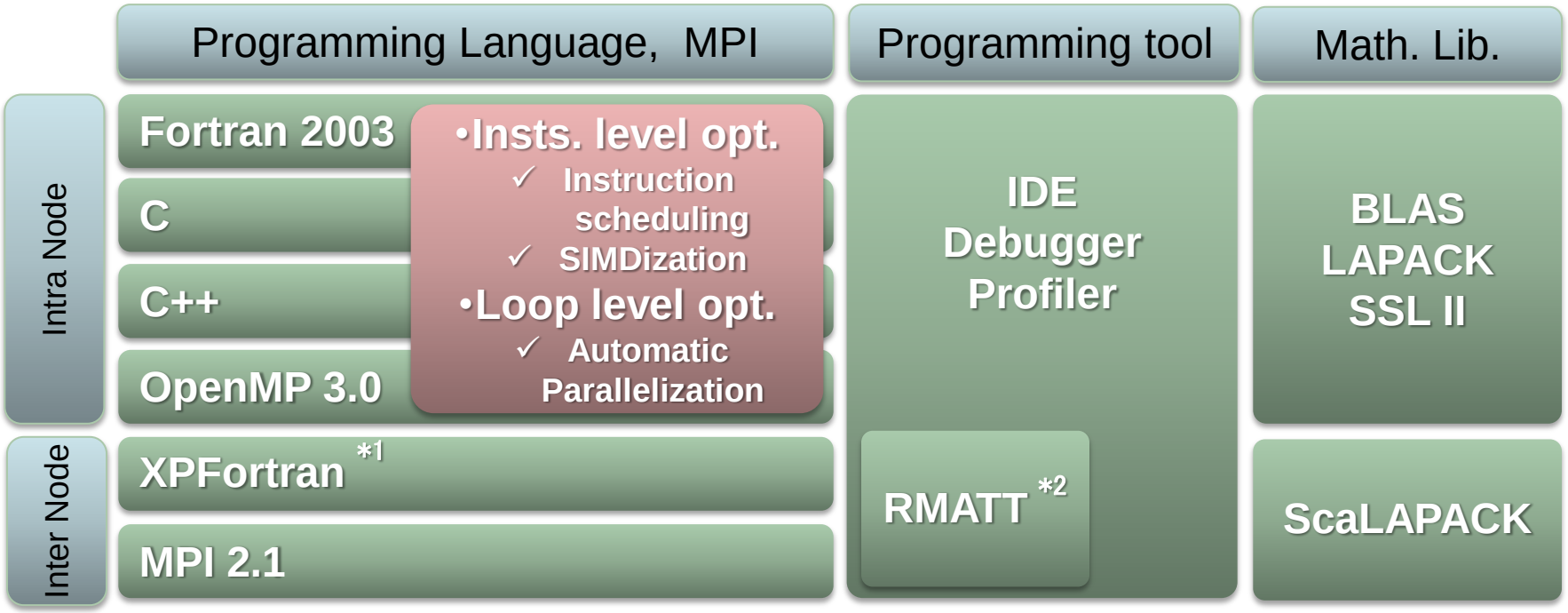
Linux based OS enhanced for FX10

PRIMEHPC FX10

Language System overview

FUJITSU

- Fortran C/C++/Fortran Compiler
- Programming model (OpenMP, MPI, XPFortran)
- Instruction level /Loop level optimization using HPC-ACE
- Debugging and Tuning tools for highly parallel computer



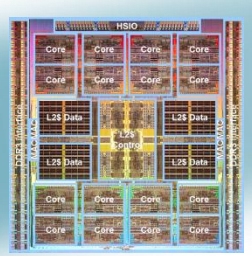
*1: eXtended Parallel Fortran (Distributed Parallel Fortran)
*2: Rank Map Automatic Tuning Tool

FX10 System H/W Specifications

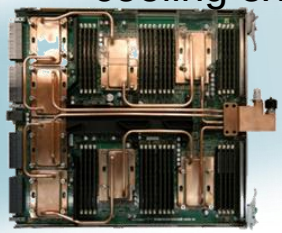
FUJITSU

FX10 H/W Specifications		
CPU	Name	*SPARC64™ IXfx*
	Performance	236.5GFlops@1.848GHz
Node	Configuration	1 CPU / Node
	Memory capacity	32, 64 GB
Rack	Performance/rack	22.7 TFlops
System	No. of compute node	384 to 98,304
	No. of racks	4 to 1,024
	Performance	90.8TFlops to 23.2PFlops
	Memory	12 TB to 6 PB

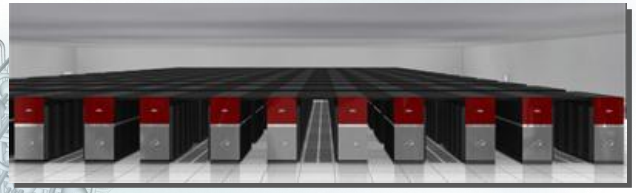
- ***SPARC64™ IXfx CPU***
  - 16 cores/socket
  - 236.5 GFlops



- ***System rack***
  - 96 compute nodes
  - 6 I/O nodes
  - With optional water cooling exhaust unit



- ***System board***
  - 4 nodes (4 CPUs)



- ***System***
  - Max. 23.2 PFlops
  - Max. 1,024 racks
  - Max. 98,304 CPUs

K computer and FX10
Comparison of System H/W Specifications

FUJITSU

		K computer	FX10
CPU	Name	*SPARC64™ VIIIfx*	*SPARC64™ IXfx*
	Performance	128GFlops@2GHz	236.5GFlops@1.848GHz
	Architecture	SPARC V9 + HPC-ACE extension	←
	Cache configuration	L1(I) Cache:32KB, L1(D) Cache:32KB	←
		L2 Cache: 6MB	L2 Cache: 12MB
	No. of cores/socket	8	16
	Memory band width	64 GB/s.	85 GB/s.
Node	Configuration	1 CPU / Node	←
	Memory capacity	16 GB	32, 64 GB
System board	Node/system board	4 Nodes	←
Rack	System board/rack	24 System boards	←
	Performance/rack	12.3 TFlops	22.7 TFlops

Comparison of System H/W Specifications (cont.)

FUJITSU

		K computer	FX10
Interconnect	Topology	6D Mesh/Torus	←
	Performance	5GB/s x2 (bi directional)	←
	No. of link per node	10	←
	Additional feature	H/W barrier, reduction	←
	Implementation	No external switch box	←
Cooling	CPU, ICC(interconnect chip), DDCON	Direct water cooling	←
	Other parts	Air cooling	Air cooling + Exhaust air water cooling unit (Optional)

Agenda

FUJITSU

- Achievements of ***K computer***
  - What did K computer achieve?
  - Technologies which realize the achievement
- Fujitsu's new Petascale supercomputer ***PRIMEHPC FX10***
  - Fujitsu supercomputer past and present
  - Second generation Petascale supercomputer PRIMEHPC FX10
- Conclusion

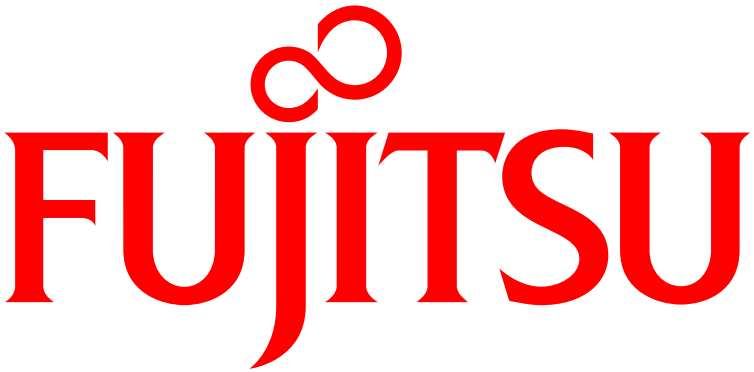
“Computing” ideal future

FUJITSU

- ***Achievements of K computer***
  - Archived LINPACK 10PFlops as scheduled
  - No.1 in 37th and 38th TOP500 list, #1 in all HPCC class 1 Awards
  - *Over PFlops performance in real practical applications*
- ***PRIMEHPC FX10*** a new Petascale supercomputer for10PFlops era
  - Enhanced technologies applied to K computer
  - Application expertise through K computer project
  - Practical Petascale computing supported by HPC software stack

- ***Challenges to ExaFlops computing***
  - 100PFlops class system is *already* in sight
  - Our challenges continues to achieve ExaFlops computing
  - ExaFlpos computing will more quickly and objectively enable us to discover ways to deliver the future world, that as yet we have been unable to perceive.

Fujitsu continues to work on realizing an ideal world through supercomputer development and its applications with you



shaping tomorrow with you