FUJITSU

shaping tomorrow with you

Approach to Application Centric Petascale Computing

- When High Performance Computing Meets Energy efficiency -

16th Nov. 2010

Motoi Okuda
Fujitsu Ltd.

Agenda

FUJITSU

- Japanese Next-Generation Supercomputer, ***K computer***
  - Project Overview
  - System Overview
  - Development Status
- Fujitsu's Technologies for Application Centric Petascale Computing
  - Design Targets
  - CPU
  - VISIMPACT
  - Tofu Interconnect
- Conclusion

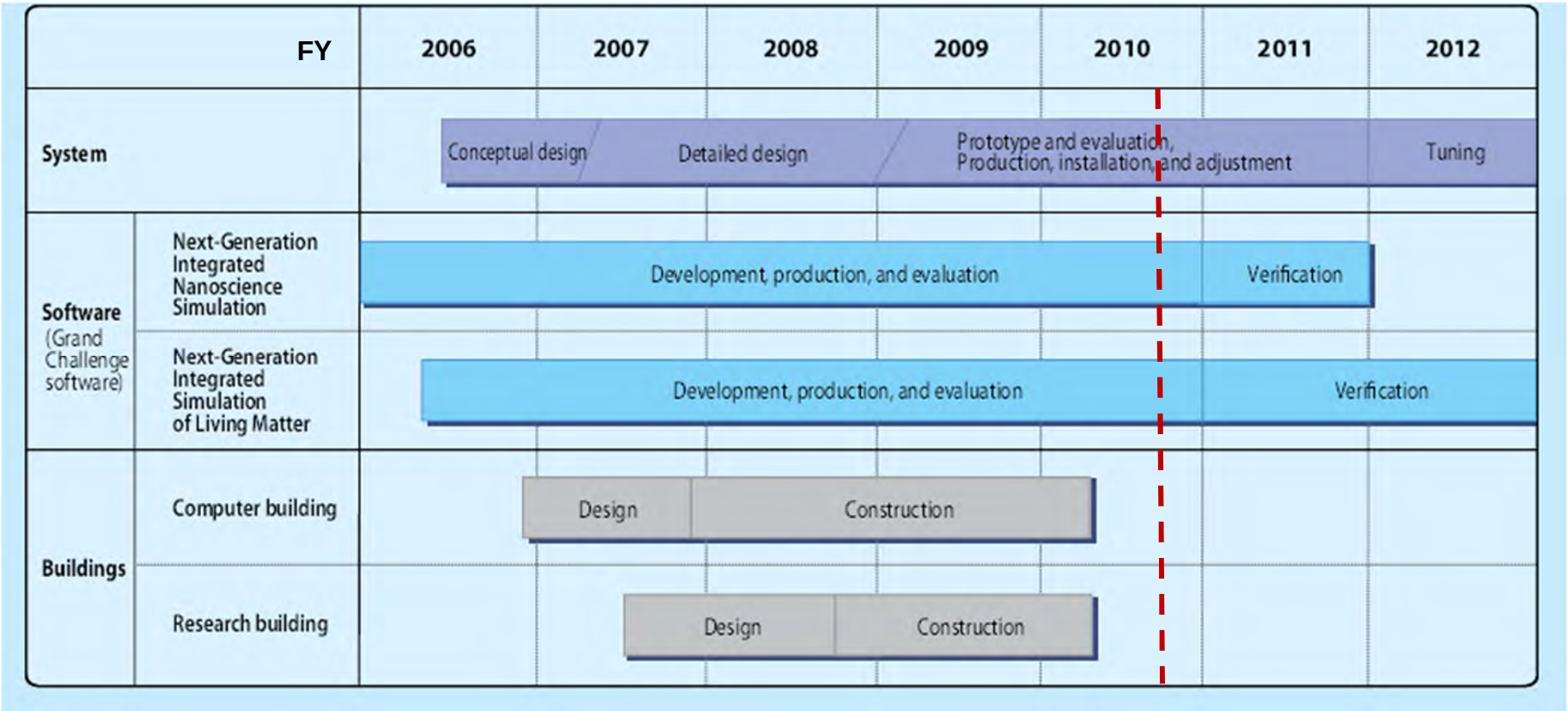
Japanese Next-Generation Supercomputer, *K computer*

- Project Overview
- System Overview
- Development Status

Project Schedule

- Facilities construction has finished in May 2010
- System installation was started in Oct. 2010
- Partial system will start test-operation in April 2011
- Full system installation will be completed in middle of 2012
- Official operation will start by the end of 2012

Courtesy of RIKEN



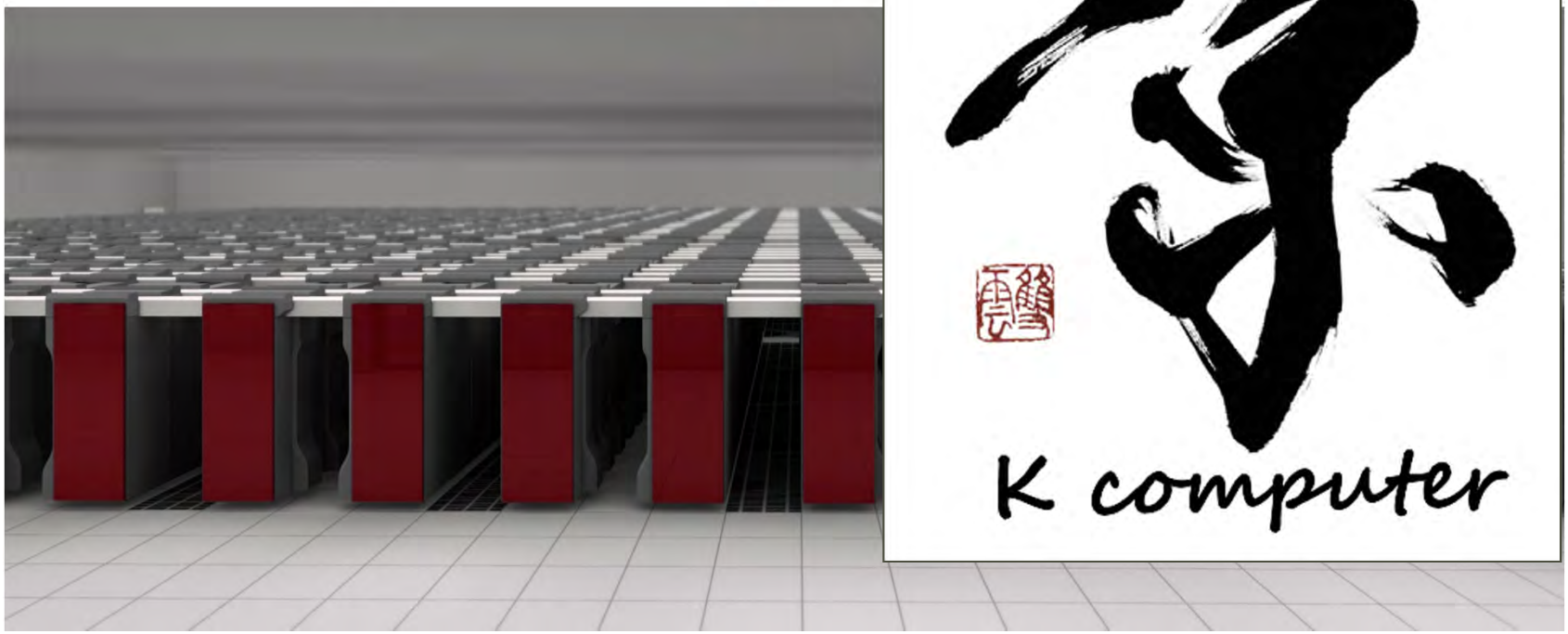
K Computer

FUJITSU

- Target Performance of Next-Generation Supercomputer
  - 10 PFlops = 10^{16} Flops = "京(Kei)" Flops,
  - "京" also means " the large gate".

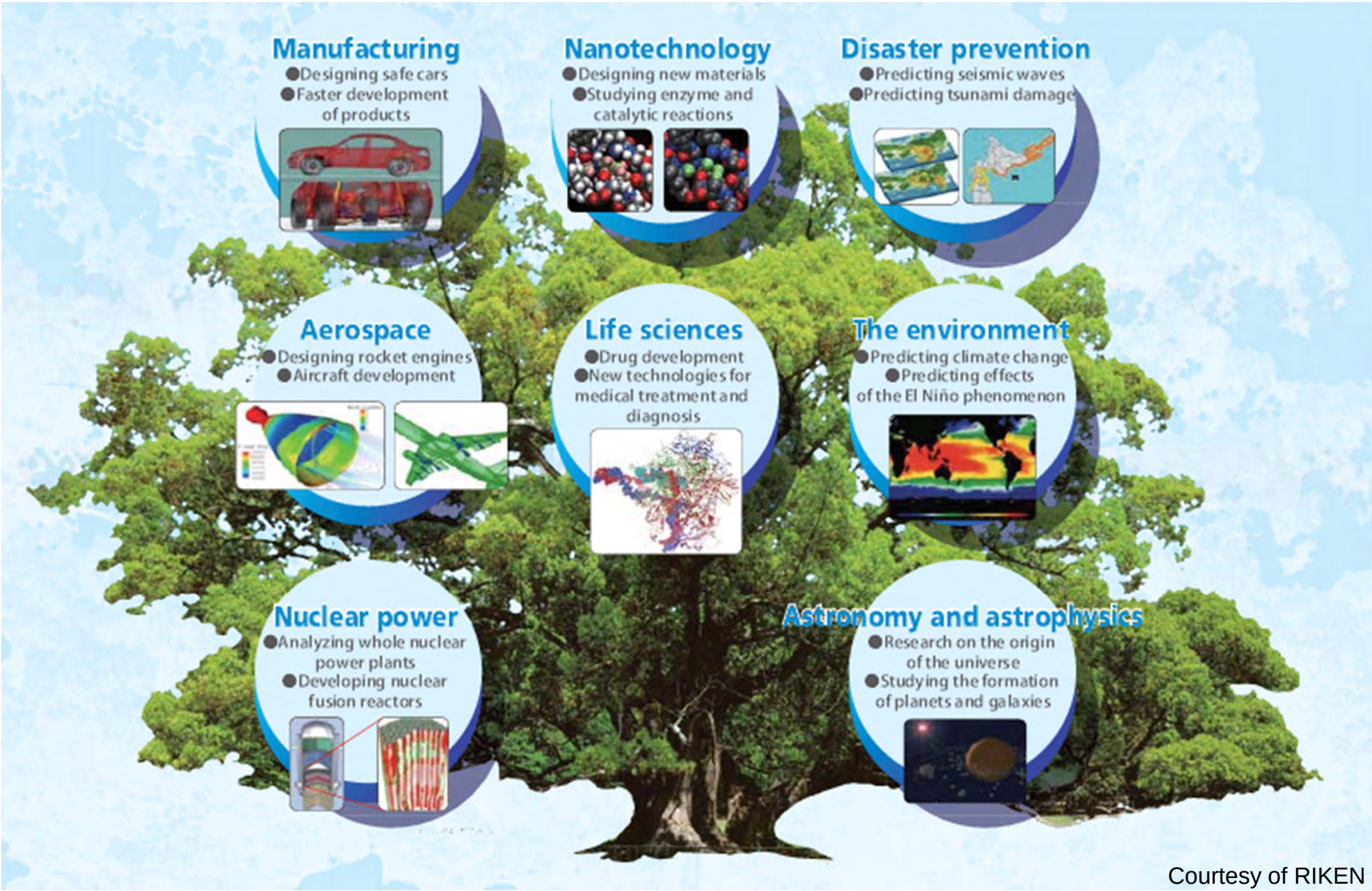
"京" (kei) computer

K computer



Full system installation (CG image)

Applications of K computer

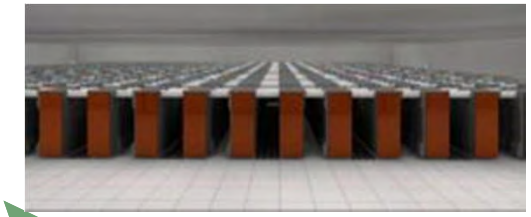


Courtesy of RIKEN

K computer Specifications

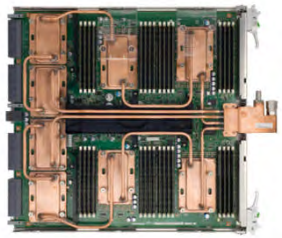
CPU *(SPARC64 VIIIfx)*	Cores/Node	8 cores (@2GHz)
	Performance	128GFlops
	Architecture	SPARC V9 + HPC extension
	Cache	L1(I/D) Cache : 32KB/32KB L2 Cache : 6MB
	Power	58W (typ. 30 C)
	Mem. bandwidth	64GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16GB (2GB/core)
System board(SB)	No. of nodes	4 nodes /SB
Rack	No. of SB	24 SBs/rack
System	Nodes/system	> 80,000

Inter-connect	Topology	6D Mesh/Torus
	Performance	5GB/s. for each link
	No. of link	10 links/ node
	Additional feature	H/W barrier, reduction
	Architecture	Routing chip structure (no outside switch box)
Cooling	CPU, ICC*	Direct water cooling
	Other parts	Air cooling

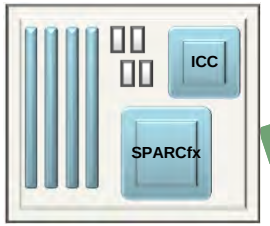


System
LINPACK 10 PFlops
over 1PB mem.
800 racks
80,000 CPUs
640,000 cores

Rack
12.3 TFlops
15TB memory

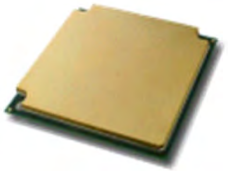


System Board
512 GFlops
64 GB memory



Node
128 GFlops
16GB Memory
64GB/s Memory band width

CPU
128GFlops
SPARC64™ VIIIfx
8 Cores@2.0GHz



* ICC : Interconnect Chip

Software Structure of K computer

FUJITSU

User / ISV Applications

HPC Portal / System Management Portal

Job/System Management

Job Scheduler
- Parallel Job execution
- Fair share schedule
- Job Accounting

HPC Cluster management
- System configuration Mgr.
- Power/IPL management
- Error monitoring

HPC enhancement
- CPU management
- Large page
- High speed interconnect

File System

FEFS
- Large scale File system (～100PB)
- Network File sharing
- High throughput File access

Language System

Automatic Parallelizing Compiler
- Fortran
- C/C++

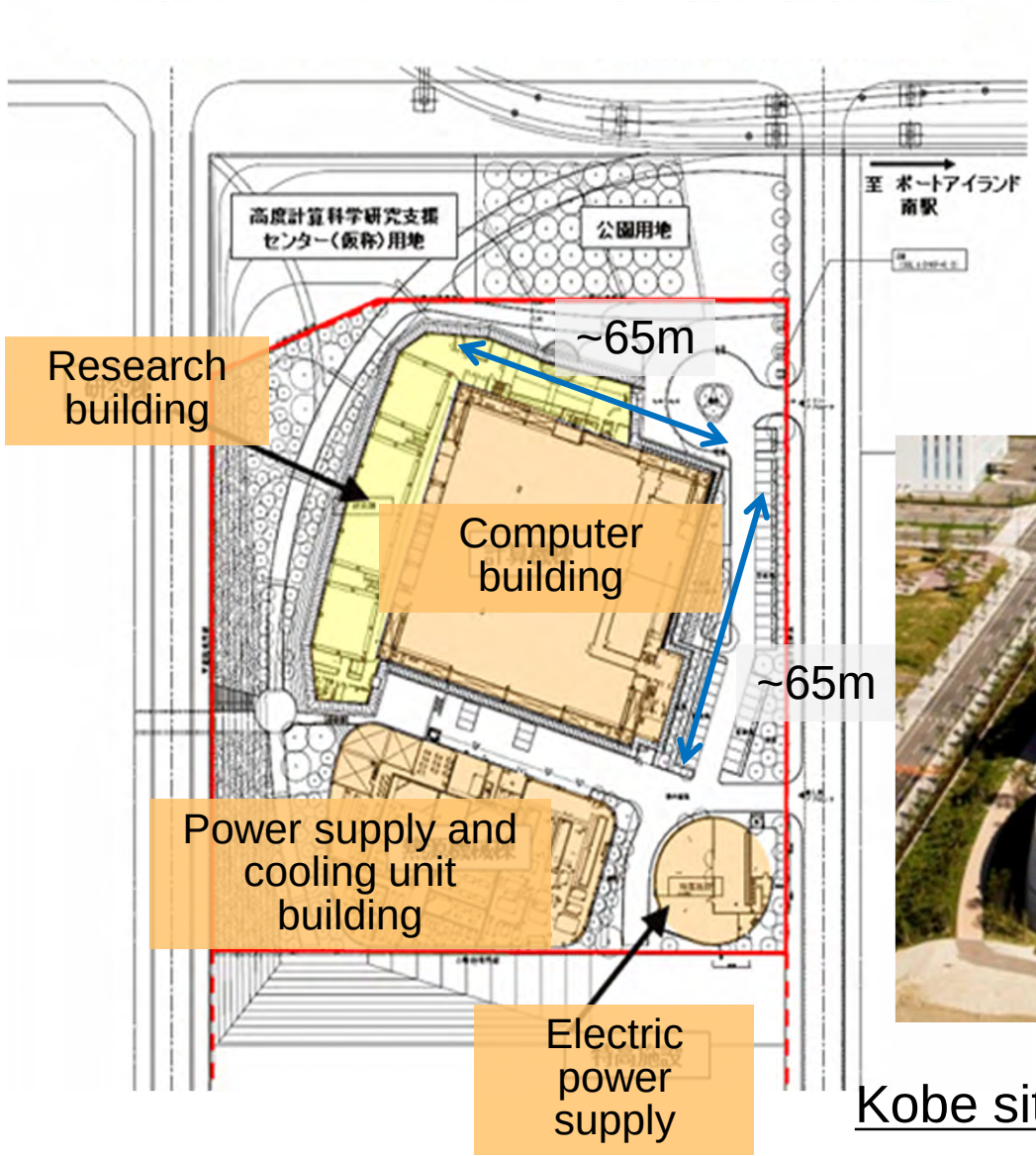
Parallel Programming
- OpenMP
- XPFortran
- MPI

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

Linux based enhanced OS

Hardware

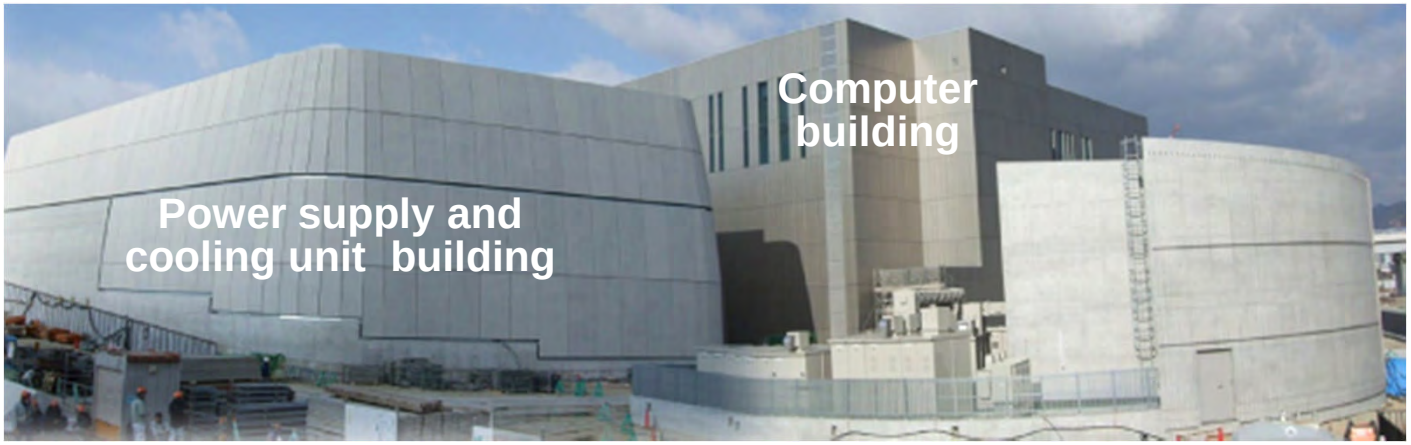
Kobe Facilities



Kobe site ground plan and aerial photo

Courtesy of RIKEN

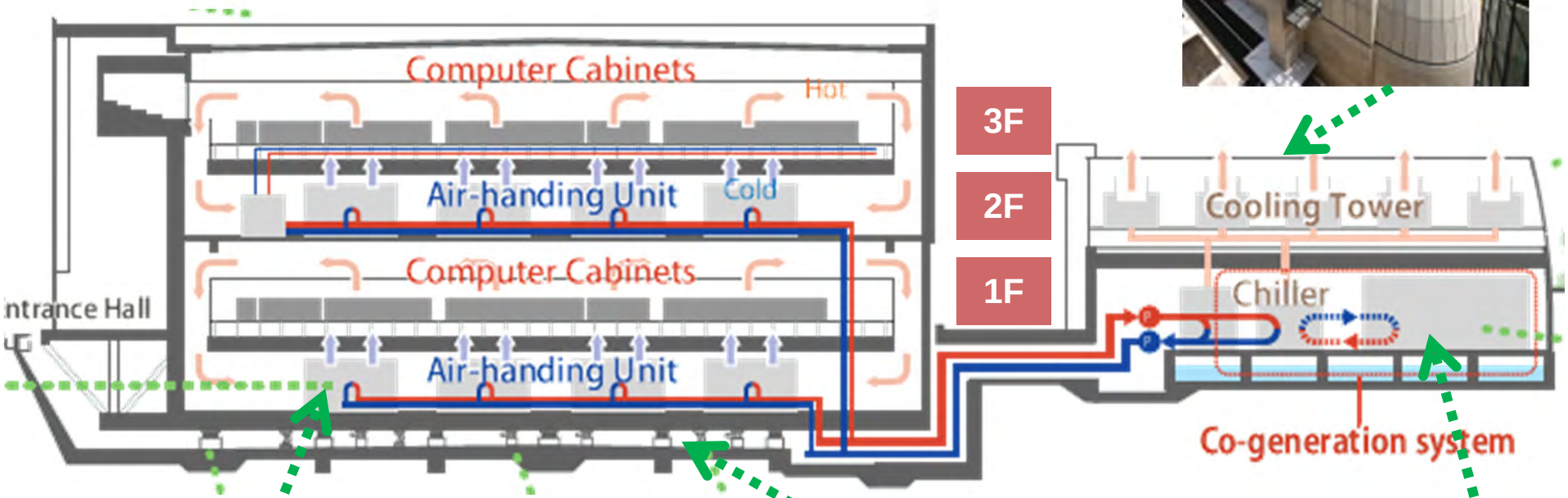
Kobe Facilities (cont.)



Exterior of buildings

Courtesy of RIKEN

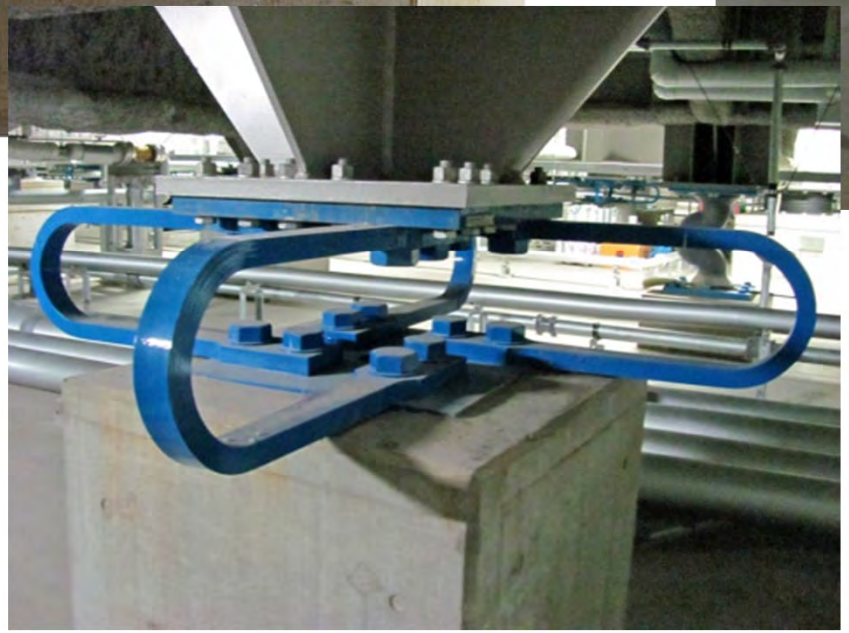
Kobe Facilities (cont.)



Cooling System

Courtesy of RIKEN

Kobe Facilities (cont.)



Seismic isolation structure

Courtesy of RIKEN

Kobe Facilities (cont.)

Cooling towers

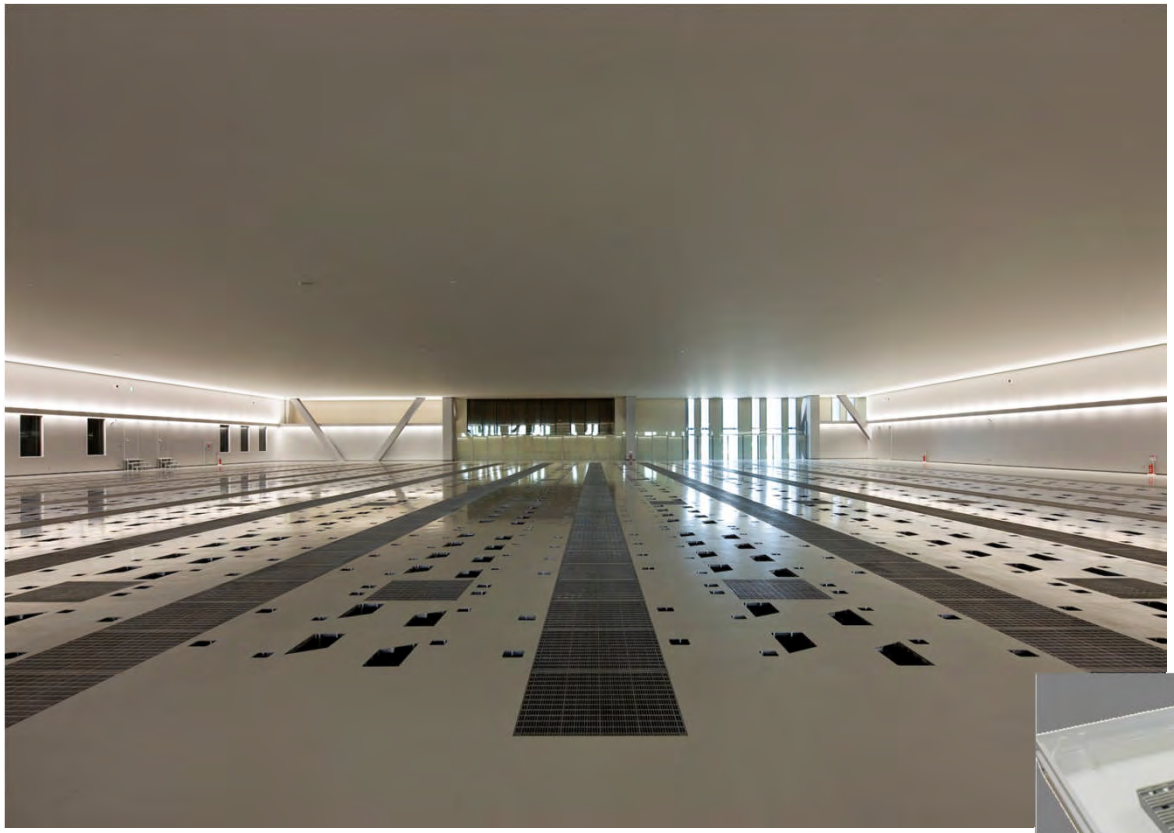
Air Handling Units
(Computer building 2F)

Power Supply and Cooling Unit Building

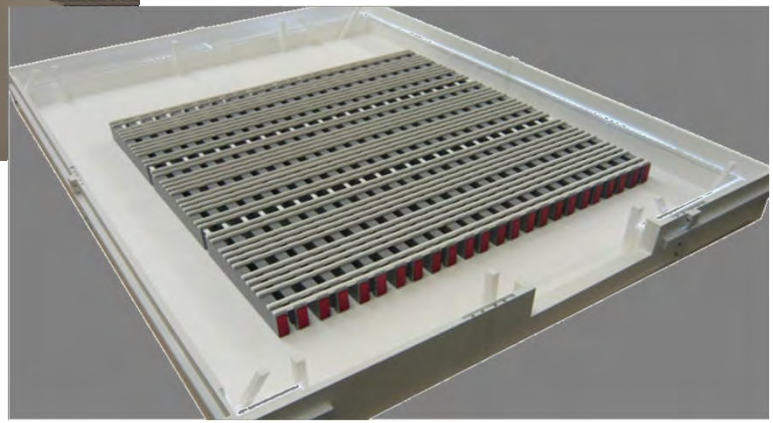
Centrifugal chillers

Courtesy of RIKEN

Kobe Facilities (cont.)



3rd Computer Floor



Full system installation (CG image)

Courtesy of RIKEN

Kobe Facilities (cont.)

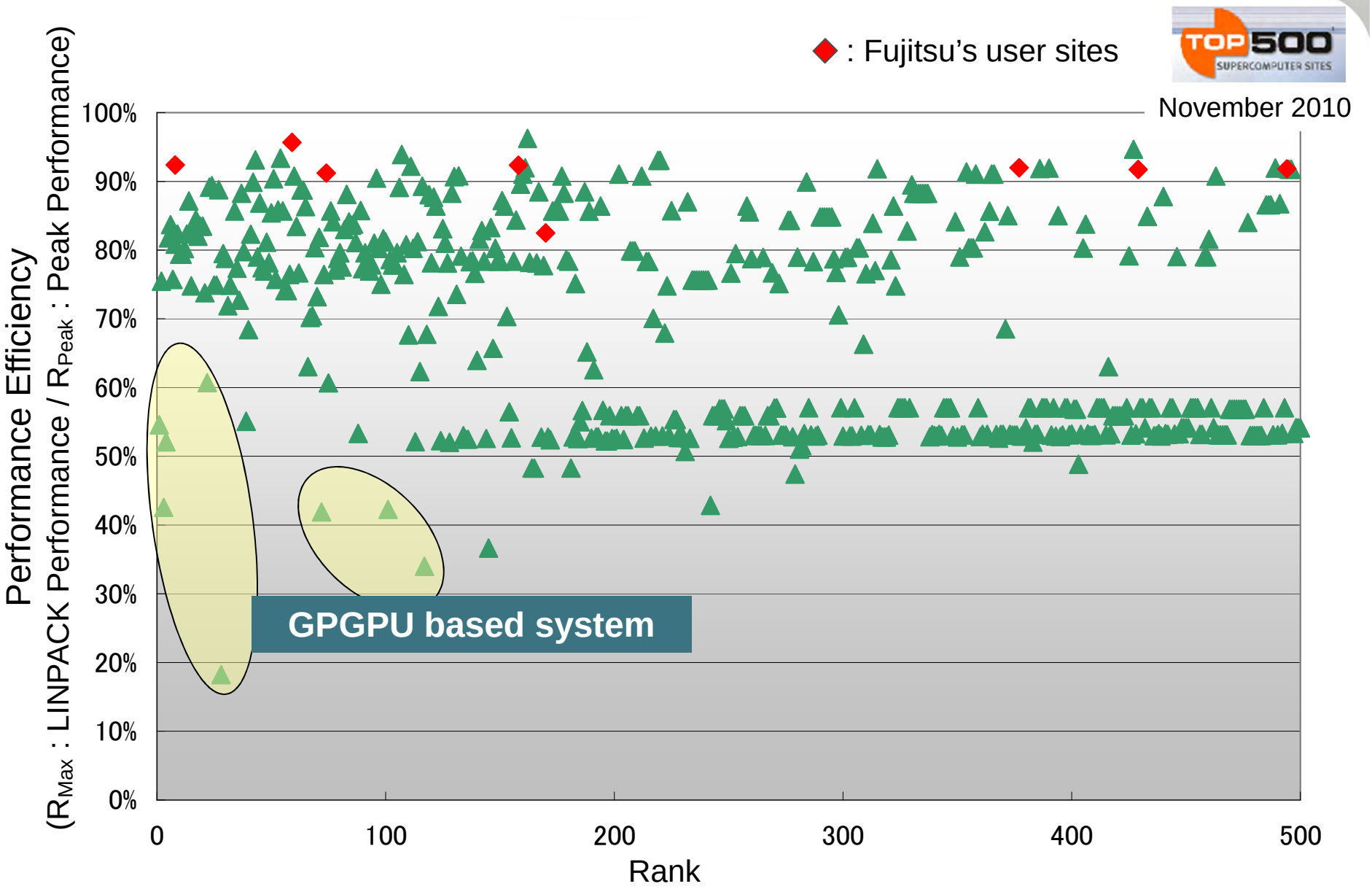
On Oct. 1st, First 8 racks were installed at Kobe site, RIKEN

Courtesy of RIKEN

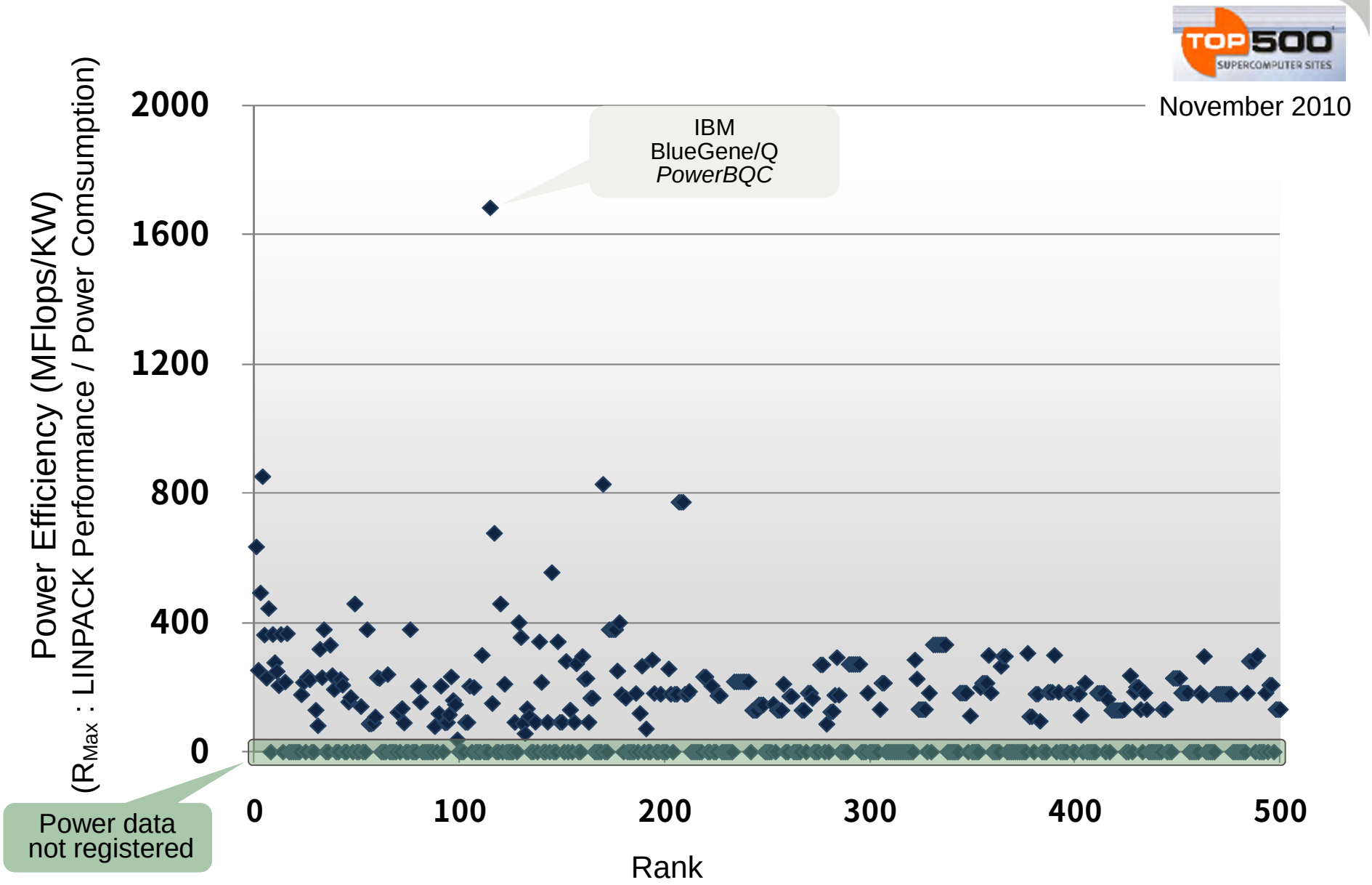
Fujitsu's Technologies for Application Centric Petascale Computing

- Design Targets
- CPU
- VISIMPACT
- Interconnect

TOP500 Performance Efficiency

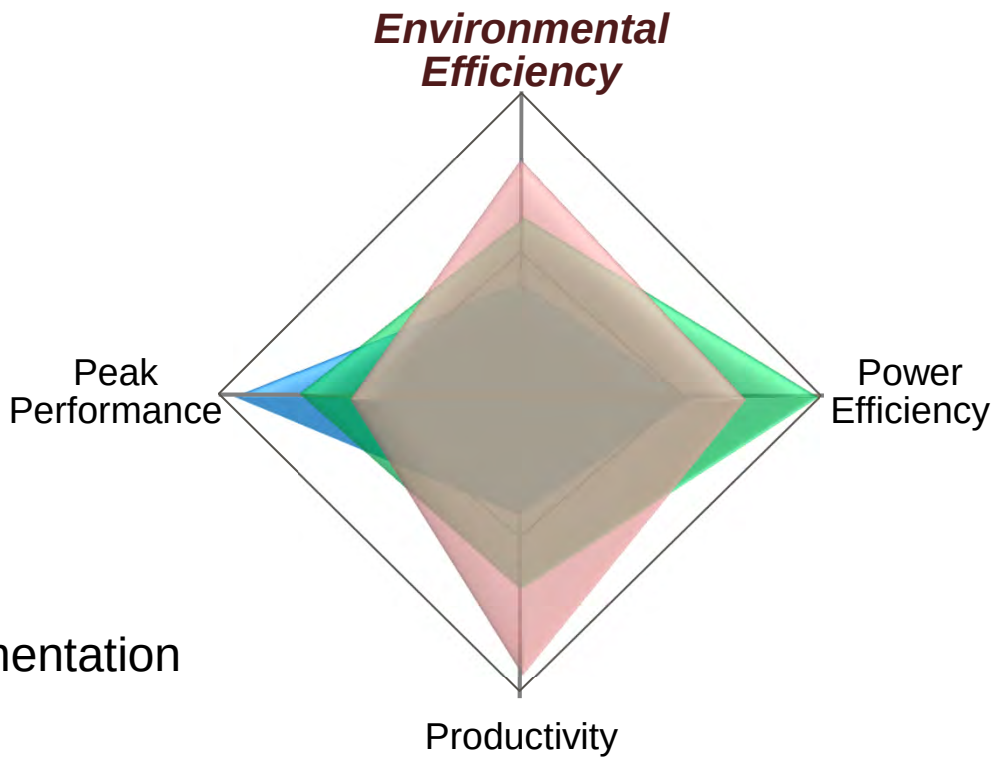


TOP500 Power Efficiency



Design Targets

- High performance
  - High peak performance
- Power efficiency and installation
  - Low power consumption
  - Small footprint
- High productivity
  - High performance efficiency
    / High sustained performance
  - High scalability
  - Less burden to application implementation
  - High reliability and availability
  - Flexible and easy operation



Environmental Efficiency =

***f* (Performance, Power efficiency, Productivity)**

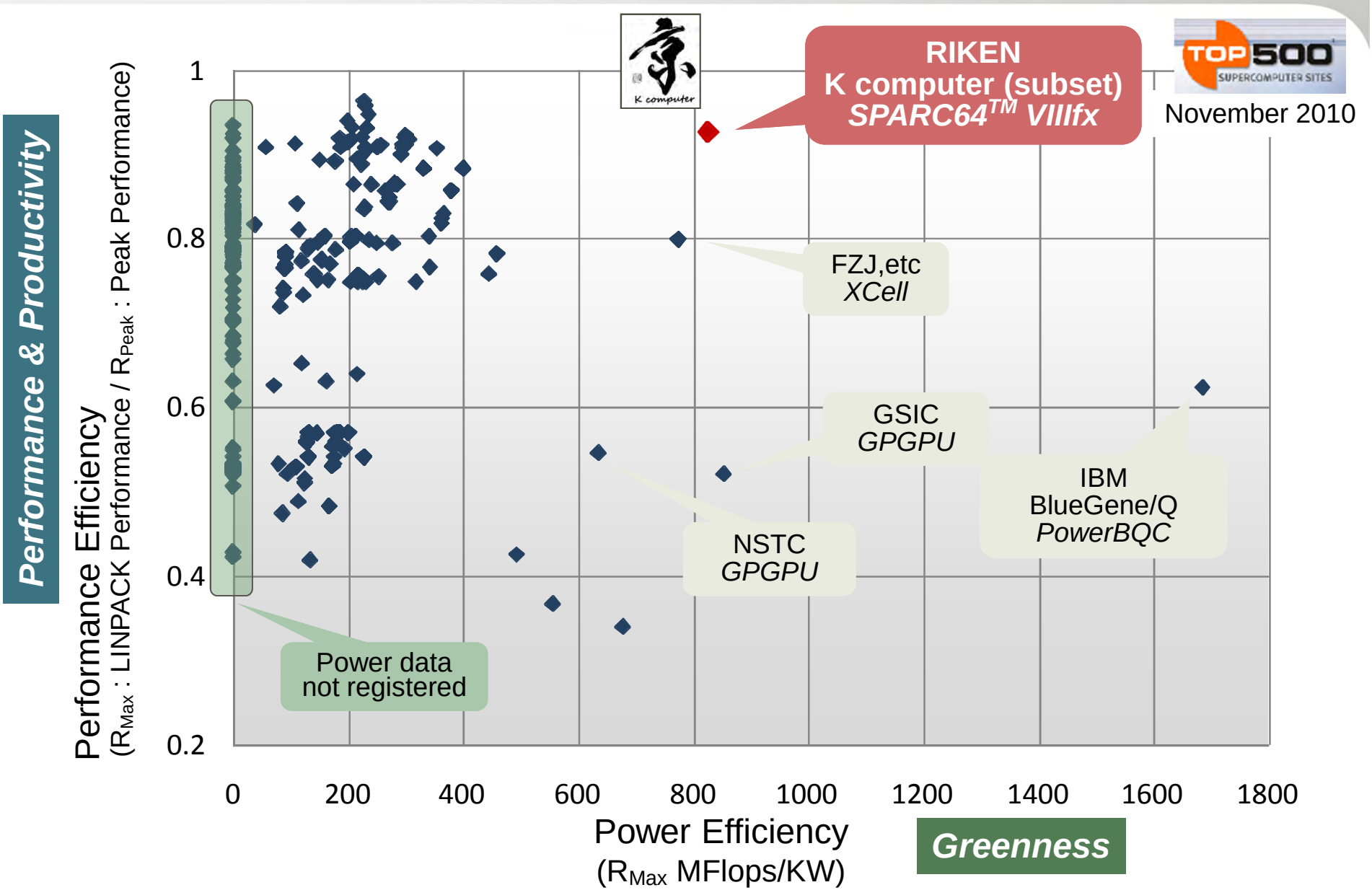
- Toward Application Centric Petascale Computing -

K computer(subset) TOP500 Score , Nov. 2010

- Information
  - Name of the site : RIKEN Advanced Institute for Computational Science, Japan
  - Machine / Year / Vendor :

    K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2010 / Fujitsu
  - No. of cores : ***3,264 cores (408 CPUs)***
- Measured result
  - R_{Max} (LINPACK) : ***48.03 TFlops***
  - R_{Peak} : 52.22 TFlops
    - ➜ LINPACK Efficiency : ***92.0%***
  - Power consumption : 57.96 KW
    - ➜ Greenness : ***828.7 MFlops/W***
- Ranking
  - TOP500 : 170th
  - Green500 : 4 th

TOP500 Performance & Power Efficiency



Technologies for Application Centric Petascale Computing

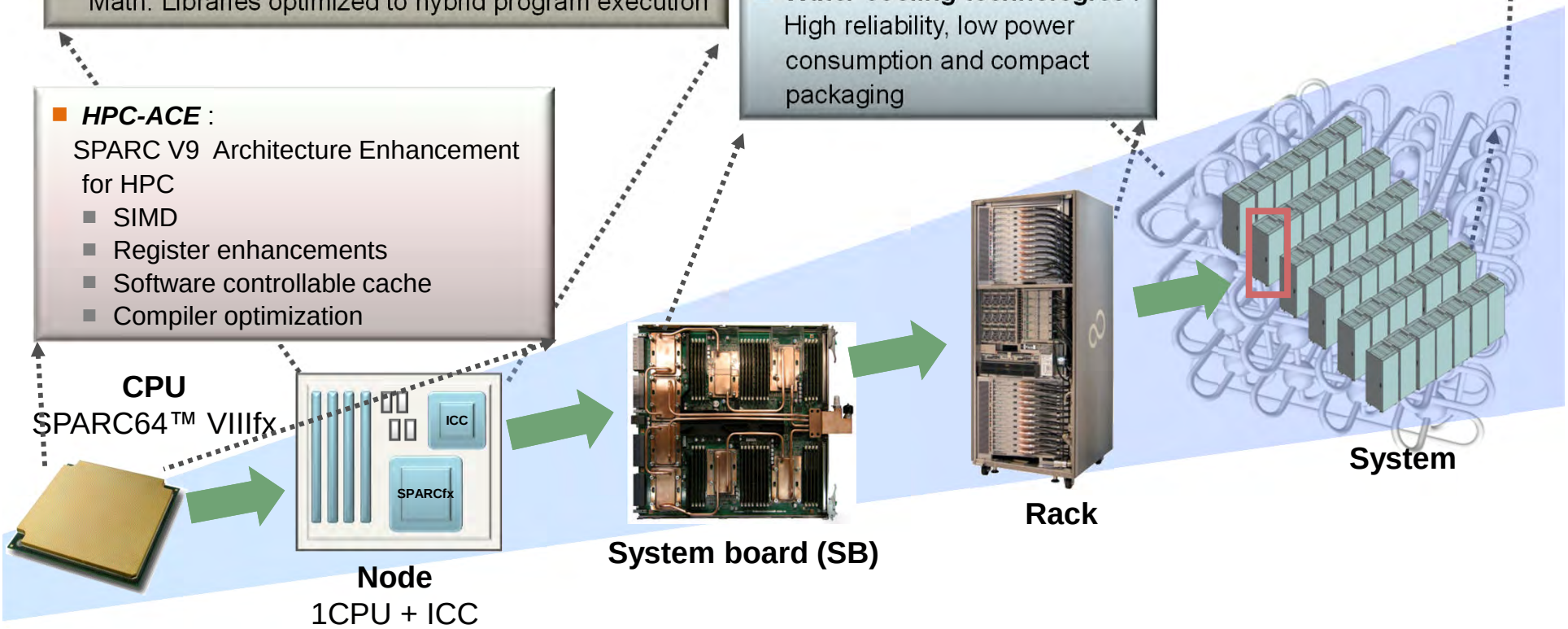
FUJITSU

- ***Single CPU per node configuration*** :
  High memory band width and simple memory hierarchy
- ***VISIMPACT*** :
  Highly efficient threading between cores for hybrid program execution (MPI + threading)
  - Hardware barrier between cores
  - Shared L2$
  - Compiler optimization
- ***Open Petascale Libraries Network*** :
  Math. Libraries optimized to hybrid program execution

- ***New Interconnect Tofu*** :
  High speed, highly scalable, high operability and high availability interconnect for over 100,000 node system
  - 6-dimensional Mesh/Torus topology
  - Functional interconnect
  - Compiler, MPI libraries and OS optimization

- ***Water cooling technologies*** :
  High reliability, low power consumption and compact packaging

- ***HPC-ACE*** :
  SPARC V9 Architecture Enhancement for HPC
  - SIMD
  - Register enhancements
  - Software controllable cache
  - Compiler optimization

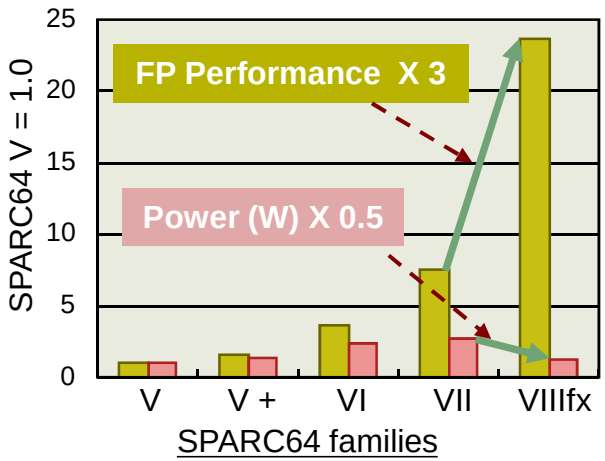


SPARC64™ VIIIfx Processor

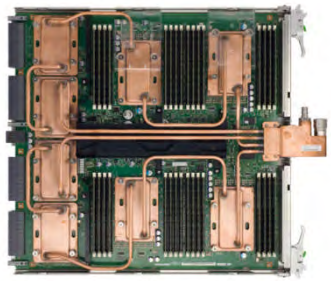
FUJITSU

- Extended SPARC64™ VII architecture for HPC
  - HPC extension for HPC : ***HPC-ACE***
    - 8 cores with 6MB Shared L2 cache
    - SIMD extension
    - 256 Floating point registers per core
    - Application access to cache management

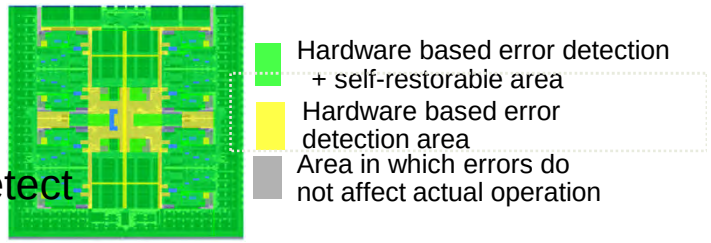
    :
  - Inter-core hardware synchronisation (barrier) for high efficient threading between core
- High performance and low power consumption
  - 2 GHz clock, 128 GFlops
  - 58 Watts/CPU as design target
- Water cooling
  - Low current leakage of the CPU
  - Low power consumption and low failure rate of CPUs
- High reliable design
  - SPARC64™ VIIIfx integrates specific logic circuits to detect and correct errors



History of Peak Performance & Power



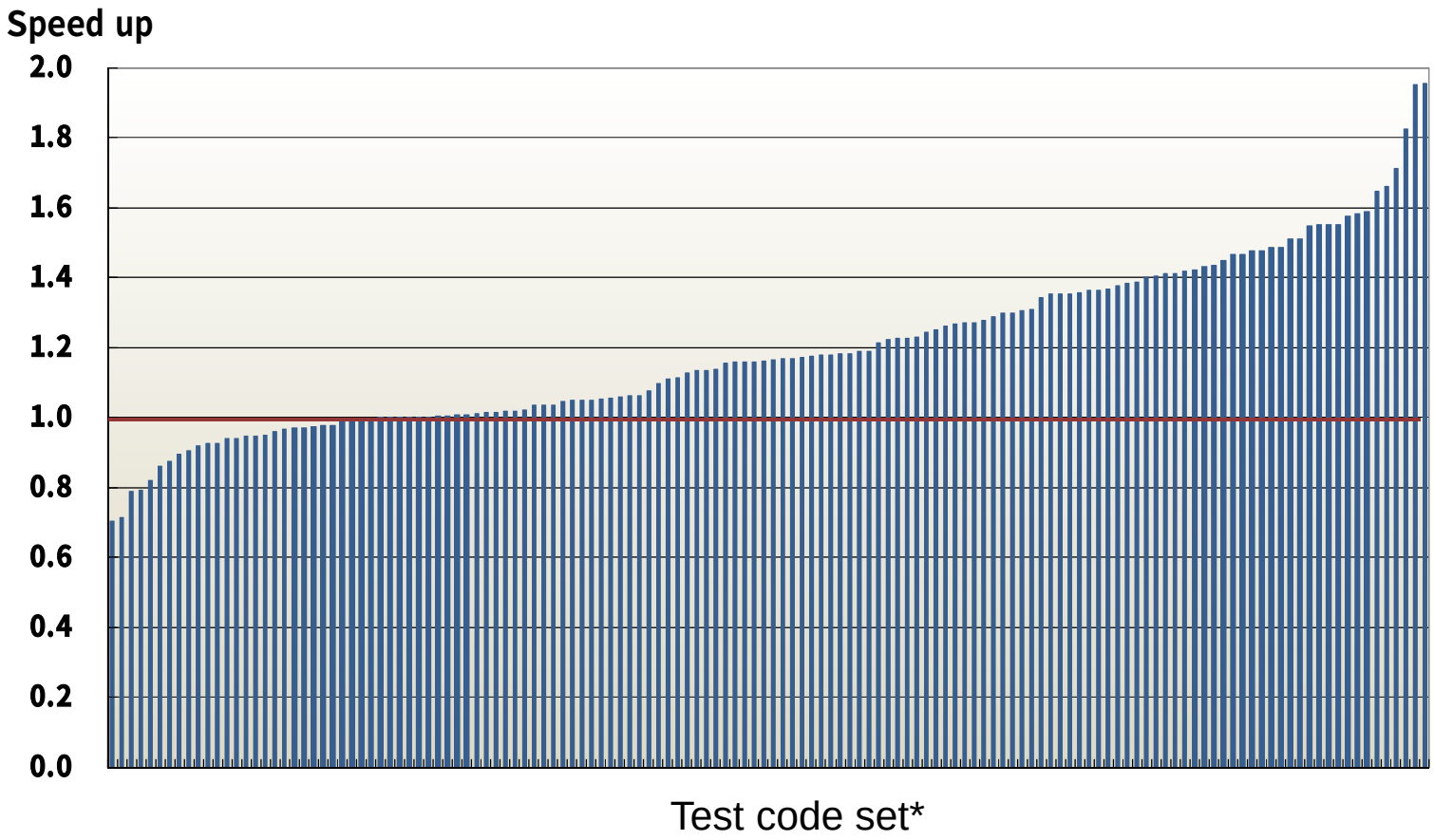
Direct water cooling System Board



SPARC64™ VIIIfx RAS coverage

Performance of SIMD Extension

- Performance improvement on Fujitsu test code set*
- *We expect further performance improvement by compiler optimization*



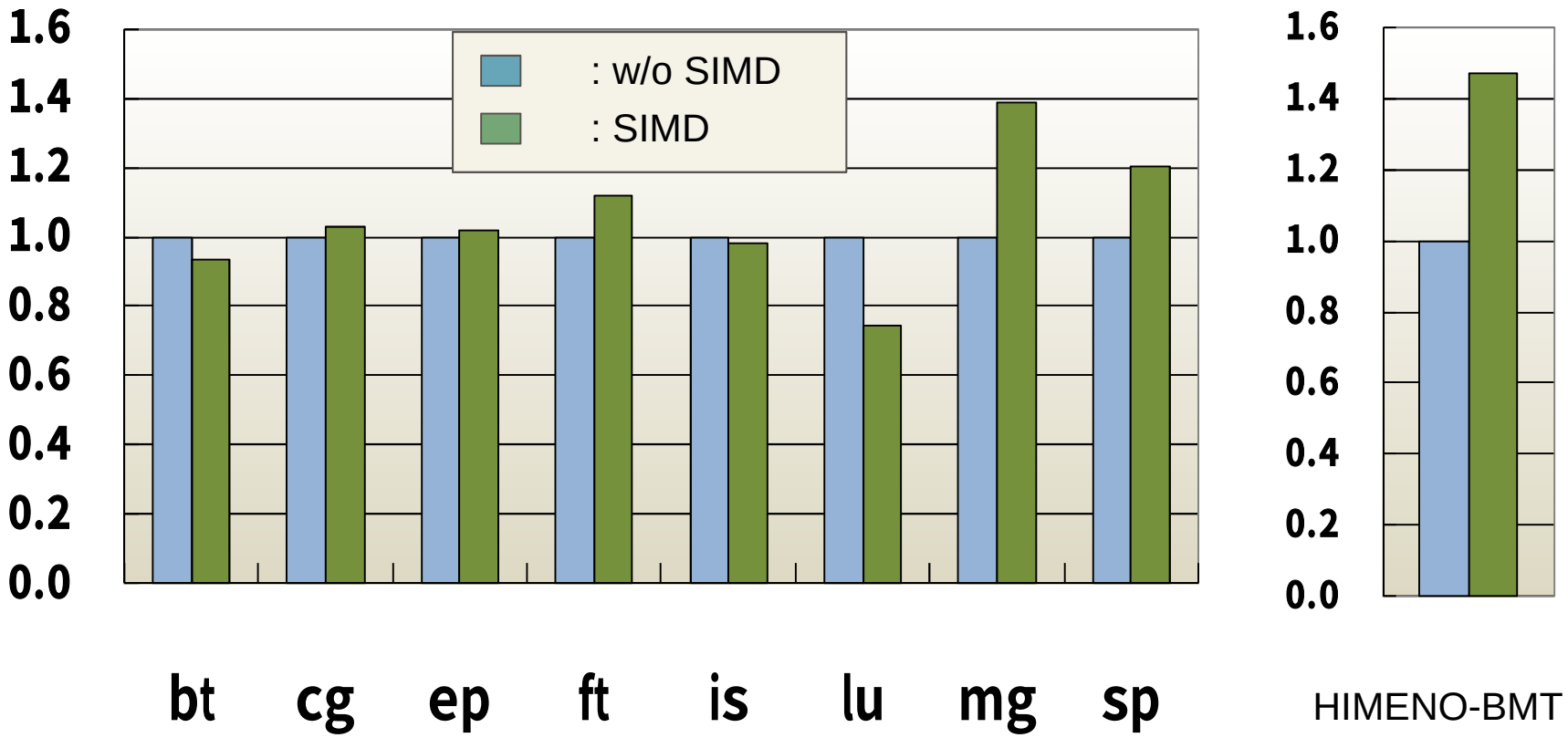
Effect of SIMD extension on one core of SPARC64^{TM} VIIIfx

* : Fujitsu internal BMT set consist of 138 real application kernels

Performance of SIMD Extension (cont.)

FUJITSU

- Performance improvement on NPB (class C) and HIMENO-BMT*
- *We expect further NPB performance improvement by compiler optimization*

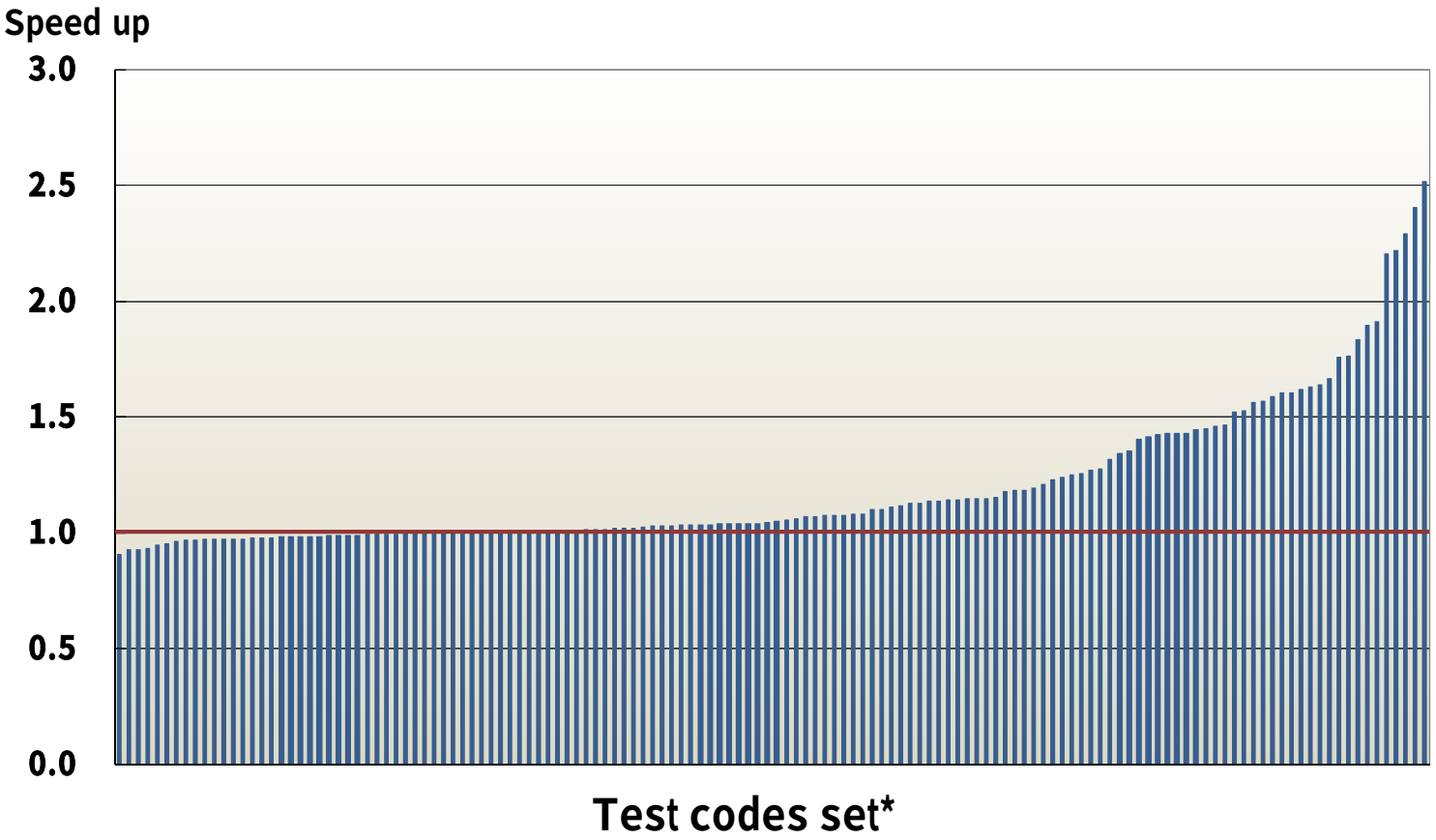


Effect of SIMD extension on one core of SPARC64^{TM} VIIIfx

* : HIMENO-BMT, Benchmark program which measures the speed of major loops to solve Poisson's equation solution using Jacobi iteration method. In this measurement, Grid-size M was used.

Performance of FP Registers Extension

- Performance improvement on Fujitsu test code set*
- No. of floating point registers : 32 ➔ 256 /core



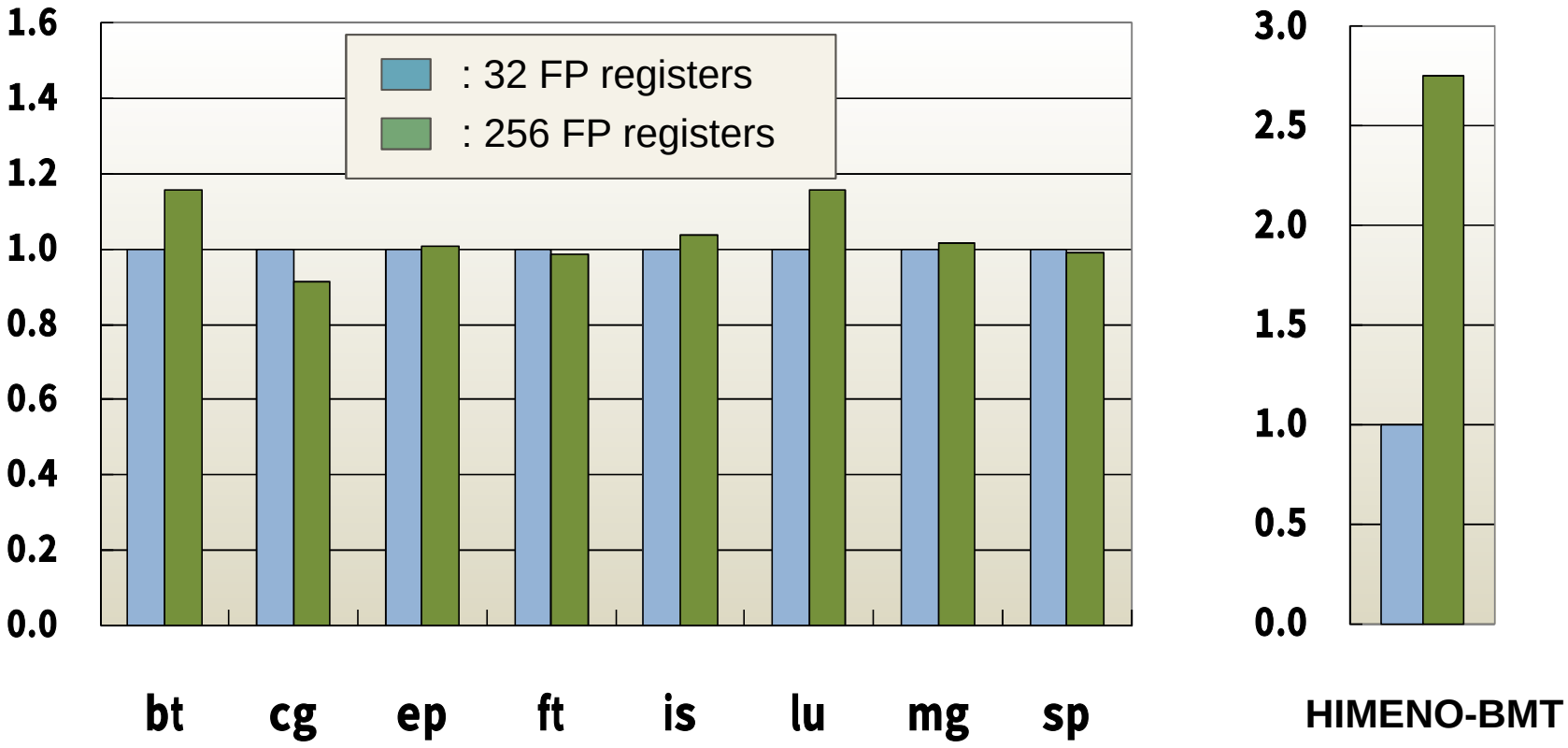
Effect of register size extension (32 to 256) on one core of SPARC64^{TM} VIIIfx

* : Fujitsu internal BMT set consist of 138 real application kernels

Performance of FP Registers Extension (cont.)

FUJITSU

- Performance improvement on NPB (class C) and HIMENO-BMT*
- *We expect further NPB performance improvement by compiler optimization*



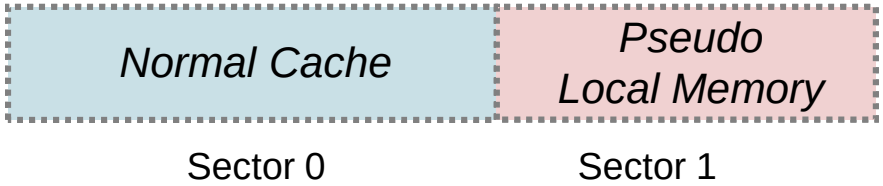
Effect of register size extension on one core of SPARC64^{TM} VIIIfx

* : HIMENO-BMT, Benchmark program which measures the speed of major loops to solve Poisson's equation solution using Jacobi iteration method. In this measurement, Grid-size M was used.

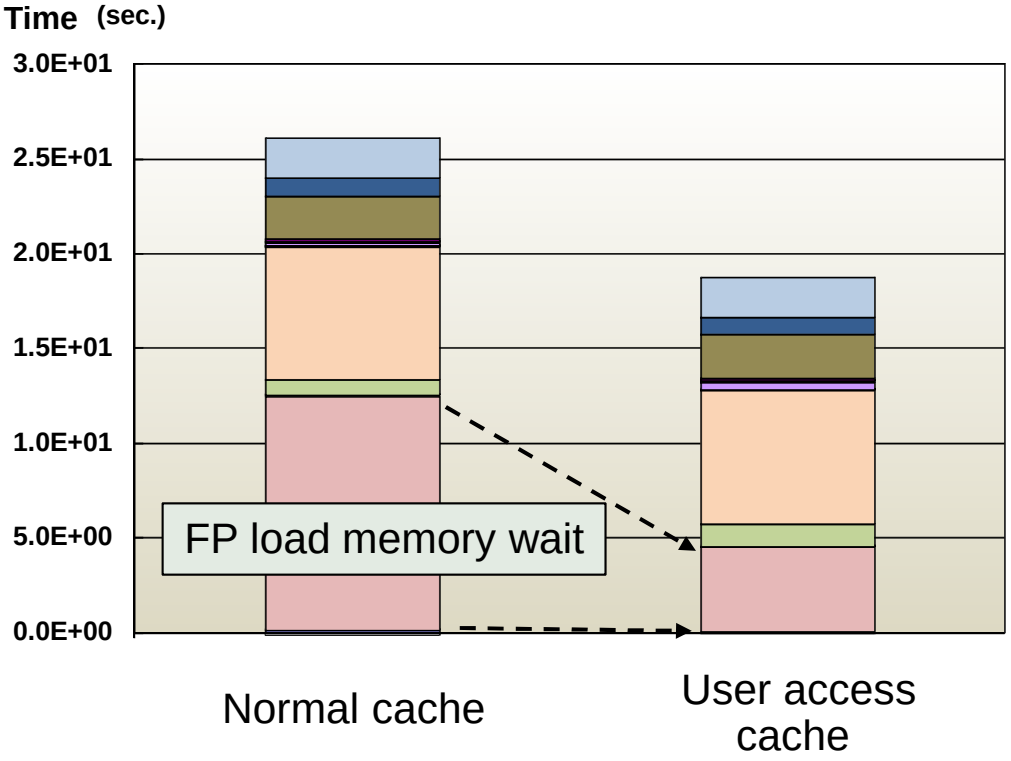
Performance of Application Accessible Cache

FUJITSU

- Application access to cache management
  - 6MB L2$ can be divided by two sectors, normal cache and Pseudo Local Memory
  - Programmer can specify reuse data by compiler directive




```
39                c----------------------------
40                !ocl cache_sector_size (3, 9)
41    1   s   s        do iter=1, itmax
42    1   s   s          call sub(a, b, c, s, n, m)
43    1   s   s        enddo
44                c----------------------------
:
~~~~~~~~~~~
52                    subroutine sub(a, b, c, s, n, m)
53                    real*8    a(n), b(m), s
54                    integer*4 c(n)
55
56                !ocl cache_subsector_assign (b)
                <<< Loop-information Start >>>
                <<<  [PARALLELIZATION]
                <<<    Standard iteration count: 728
                <<<  [OPTIMIZATION]
                <<<    SIMD
                <<<    SOFTWARE PIPELINING
                <<< Loop-information  End >>>
57    1  pp  4v        do i=1,n
58  1 p   4v             a(i) = a(i) + s * b(c(i))
59    1   p   4v       enddo
60
61                    end
```

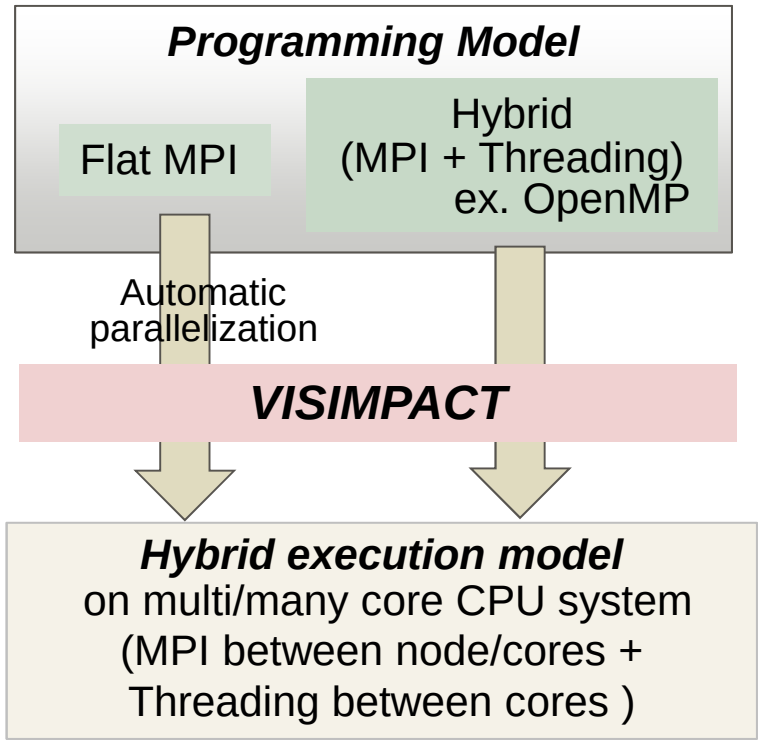


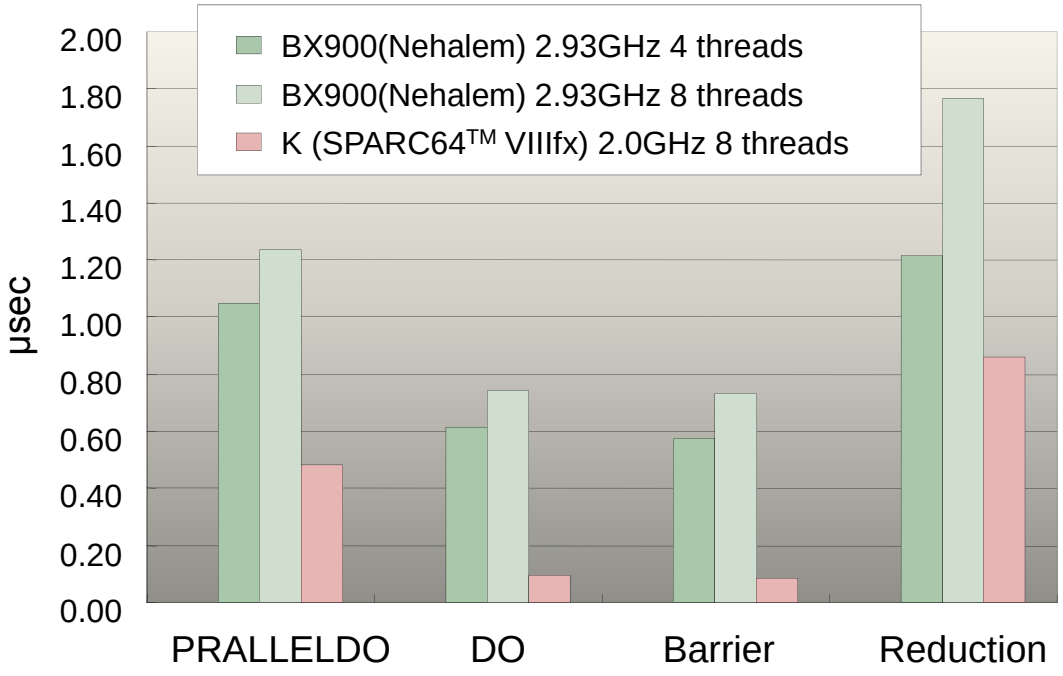
Effect of application accessible cache on one chip of SPARC64^{TM} VIIIfx

Performance of VISIMPACT (Integrated Multi-core Parallel ArChiTecture)

FUJITSU

- Concept
  - Hybrid execution model (MPI + Threading between core)
    - ➜Can improve parallel efficiency and reduce memory impact
    - ➜Can reduce the burden of program implementation over multi and many core CPU
- Technologies
  - Hardware barriers between cores, shared L2$ and automatic parallel compiler
    - ➜High efficient threading ***: VISIMPCT*** (Integrated Multi-core Parallel ArChiTecture)

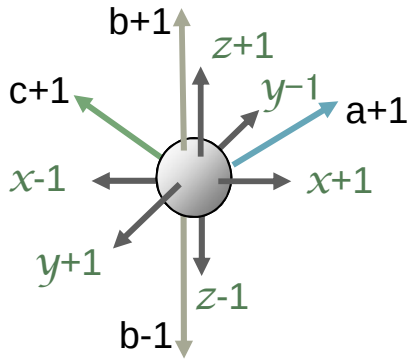


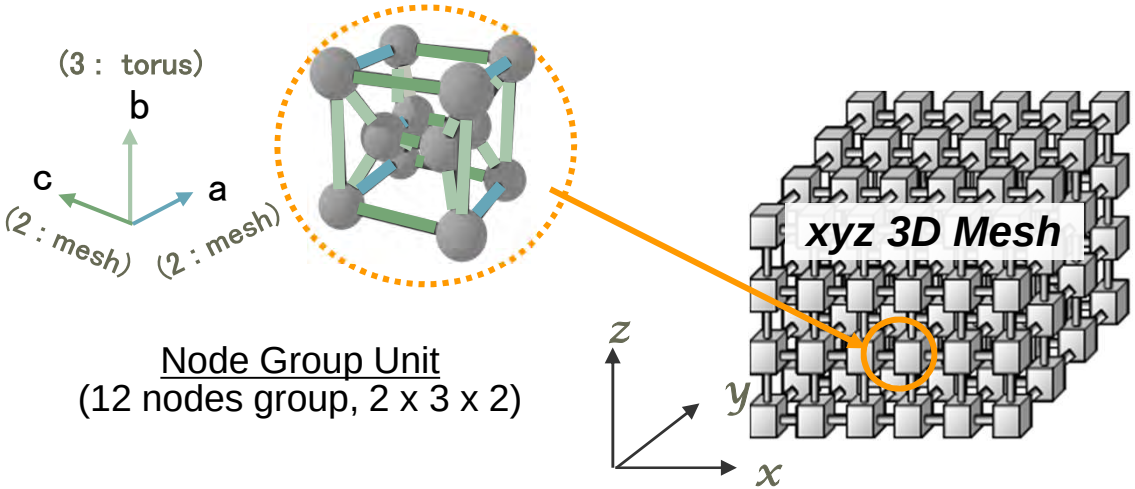


Comparison of OpenMP micro BMT performance between SPARC64^{TM} VIIIfx and Nehalem

New Interconnect : *Tofu*

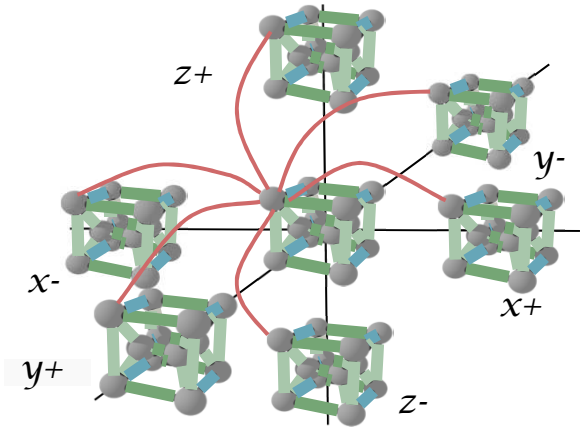
- Design targets
  - Scalabilities toward 100K nodes
  - High operability and usability
  - High performance
- Topology
  - User view/Application view : Logical 3D Torus (X, Y, Z)
  - Physical topology : 6D Torus / Mesh addressed by (x, y, z, a, b, c)
    - 10 links / node, 6 links for 3D torus and 4 redundant links





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

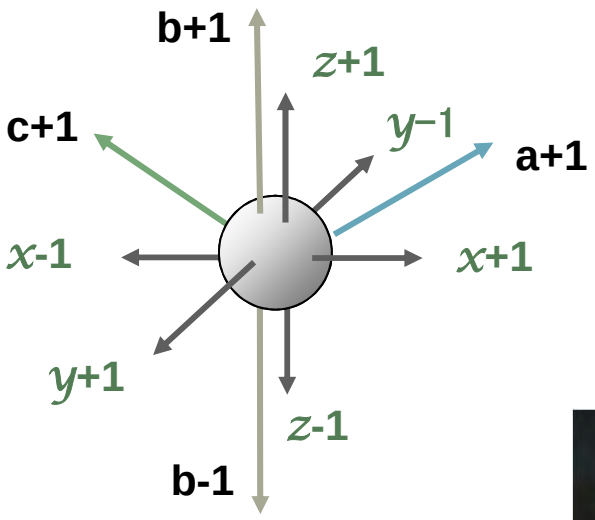
Node Group 3D connection



3D connection of each node

New Interconnect (cont.)

- Technology
  - Fast node to node communication : *5 GB/s x 2 (bi-directional) /link, 100GB/s. throughput /node*
  - Integrated MPI support for collective operations and global hardware barrier
  - Switch less implementation



Each link : 5GB/s X 2
Throughput : 100GB/s/node

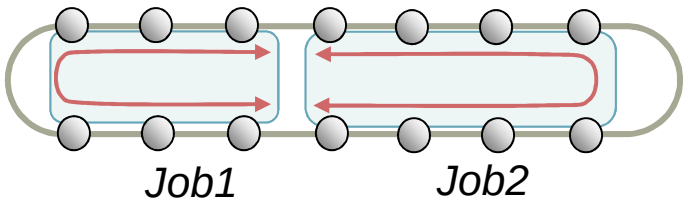
Conceptual Model

IEEE Computer Nov. 2009

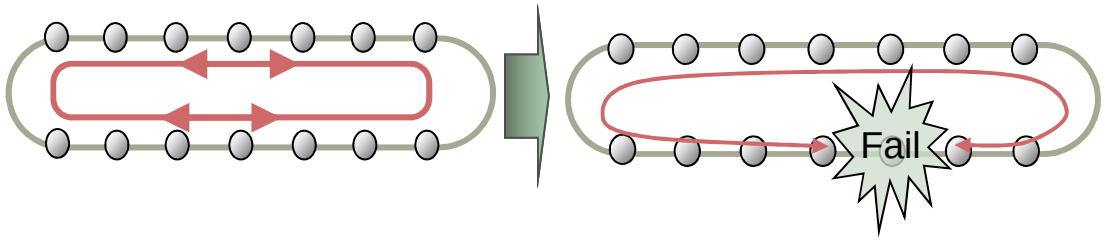
Why 6 dimensions?

- High Performance and Operability
  - Low hop-count (average hop count is about ½ of conventional 3D torus)
  - The 3D Torus/Mesh view is always provided to an application even when meshes are divided into arbitrary sizes
  - No interference between jobs
- Fault tolerance
  - 12 possible alternate paths are used to bypass faulty nodes
  - Redundant node can be assigned preserving the torus topology

Conventional Torus

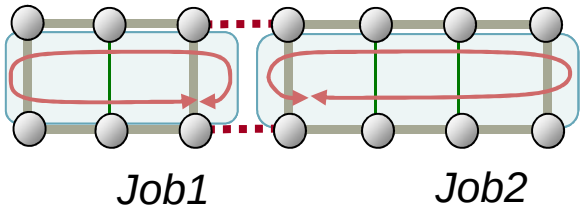


High hop-count & mesh

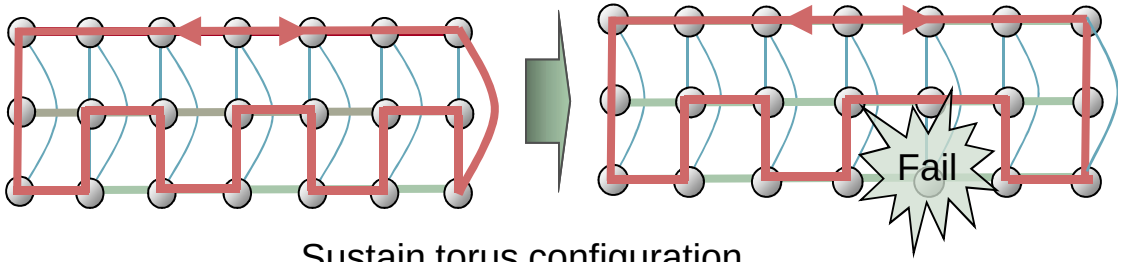


Torus topology can't be configured

Tofu interconnect



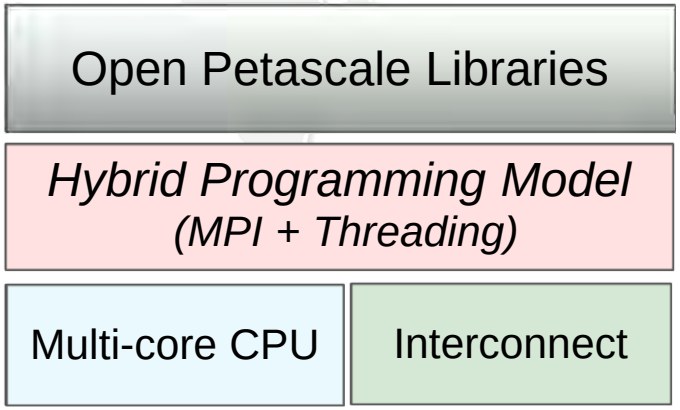
Low hop-count & keep torus with job isolation

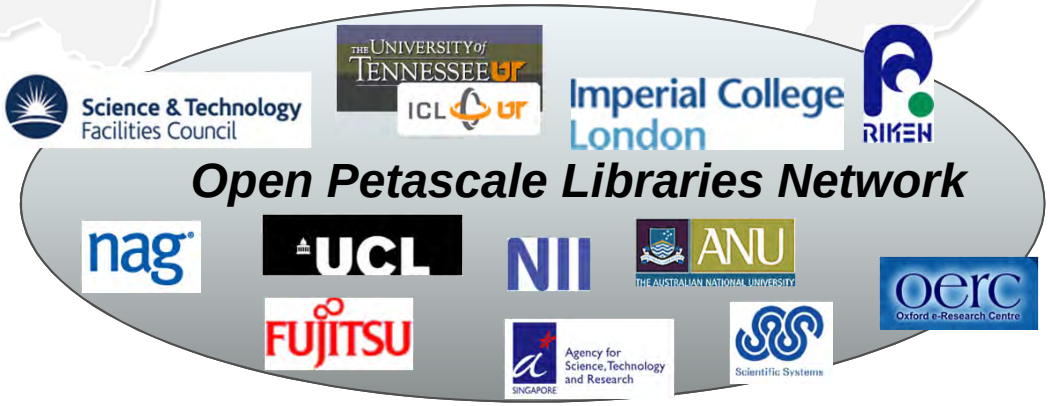


Sustain torus configuration

Open Petascale Libraries Network

- How to reduce the burden to application implementation over multi/many core system, i.e. How to reduce the burden of the two stage parallelization?
- Collaborative R&D project for Mathematical Libraries just started
  - Target system
    - Multi-core CPU based MPP type system
    - Hybrid execution model (MPI + threading by OpenMP/automatic parallelization)
  - Cooperation and collaboration with computer science, application and computational engineering communities on a global basis, coordinate by FLE*
- ***Open-source implementation***
  - Sharing information and software
  - Results of this activity will be open to HPC society





*: Fujitsu Labs Europe, located in London

Open Petascale Libraries Network (cont.)

FUJITSU

Global | Search

Products | Services | Solutions | Support | Corporate Information

Home > News > Press Releases > Archives > By Month > 2010 > Fujitsu Launches Global Initiative to Develop Mathematical Library for Petascale

News
- Press Releases
  - Recent
  - Archives
    - By Month
      - 2010
      - 2009
      - 2008
      - 2007
      - 2006
      - 2005
      - 2004
      - 2003

Fujitsu Limited
Fujitsu Laboratories of Europe Limited

Fujitsu Launches Global Initiative to Develop Mathematical Library for Petascale Computing Applications

To be employed in maximising the performance of the Next-Generation Supercomputer (the "K computer")

Tokyo, November 9, 2010 — Fujitsu Limited and Fujitsu Laboratories of Europe Limited today announced the launch of the Open Petascale Libraries (OPL) project, a global collaboration initiative to develop a mathematical library(1) that will serve as a development platform for applications running on petascale-class supercomputers. Initially involving ten partners, including universities and research institutions, the project will make the developed code publicly available in open-source form, thereby contributing to the computational science community as a whole. In addition, the output from the OPL project will be applied to help accelerate the application development for the Next-Generation Supercomputer (the "K computer")(2), which is scheduled to begin operation in fiscal 2012. As a result, this project is expected to make an important contribution to a range of fields, such as the life sciences, development of new materials and sources of energy, disaster prevention and mitigation, manufacturing technologies and basic research into the origins of matter and the universe.

The launch of the OPL project is scheduled to coincide with SC10, a conference bringing together supercomputer professionals from around the world, with the project's inaugural workshop to be held on November 14 in New Orleans, LA.

Nov. 9th, 2010

Conclusion

Toward Application Centric Petascale Computing

FUJITSU

- Installation of RIKEN's K computer has started

- Leading edge Fujitsu's technologies are applied to K computer
  - ***High environmental efficiency***
    - High performance-efficiency
    - Low power consumption
    - High productivity
  - ***Technologies***
    - New CPU
    - Innovative interconnect
    - Advanced packaging
    - Open Petascale Libraries Network

- Those technologies will be enhanced and applied to Fujitsu's future commercial supercomputer

FUJITSU

shaping tomorrow with you