FUJITSU

shaping tomorrow with you

Fujitsu's Approach to Application Centric Petascale Computing

2nd Nov. 2010

Motoi Okuda
Fujitsu Ltd.

Agenda

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

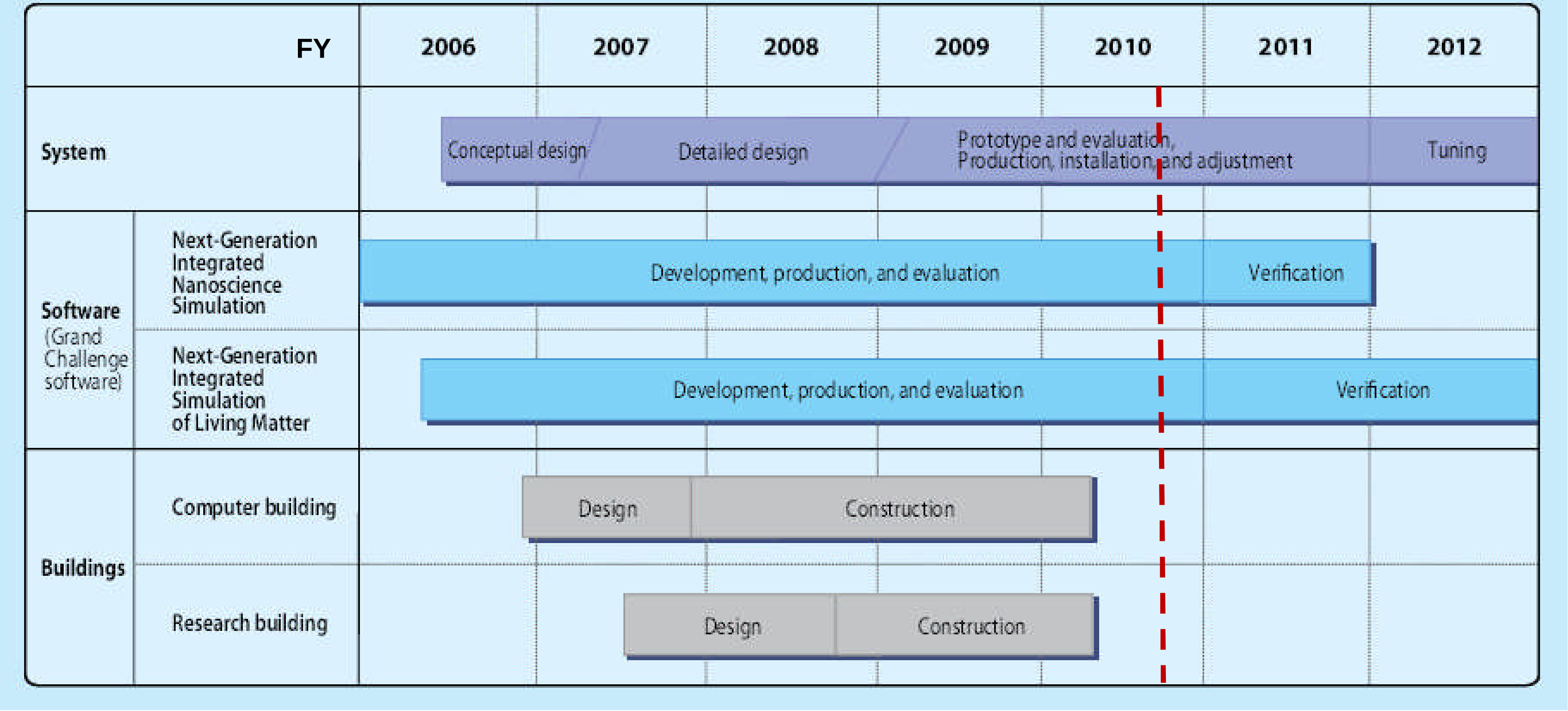
Japanese Next-Generation Supercomputer, *K Computer*

- Project Overview
- Design Targets
- 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



FY

2006

2007

2008

2009

2010

2011

2012

System

Conceptual design

Detailed design

Prototype and evaluation, Production, installation, and adjustment

Tuning

Software (Grand Challenge software)

Next-Generation Integrated Nanoscience Simulation

Development, production, and evaluation

Verification

Next-Generation Integrated Simulation of Living Matter

Development, production, and evaluation

Verification

Buildings

Computer building

Design

Construction

Research building

Design

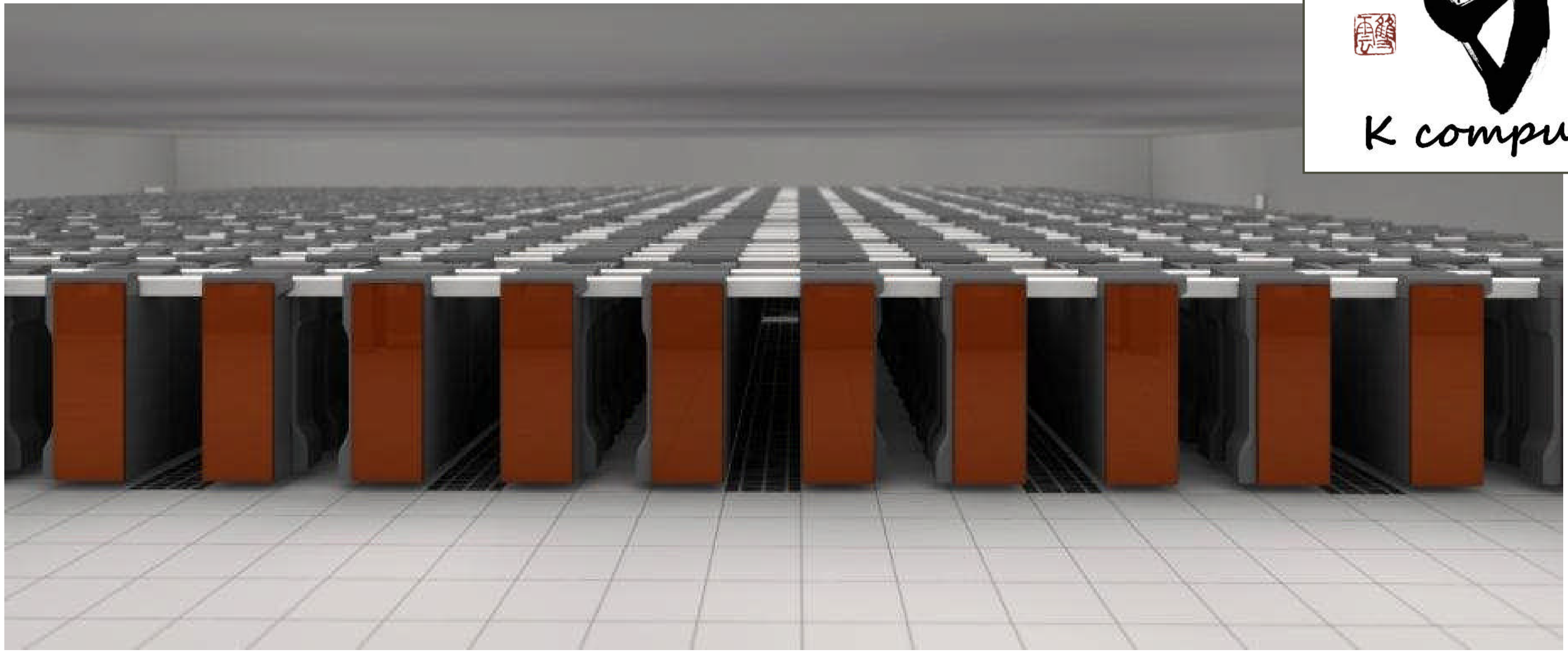
Construction

K Computer

- Target Performance of Next-Generation Supercomputer
  - 10 PFlops = 10^{16} Flops = “**京**(Kei)” Flops, “**京**” means the “Gate”.

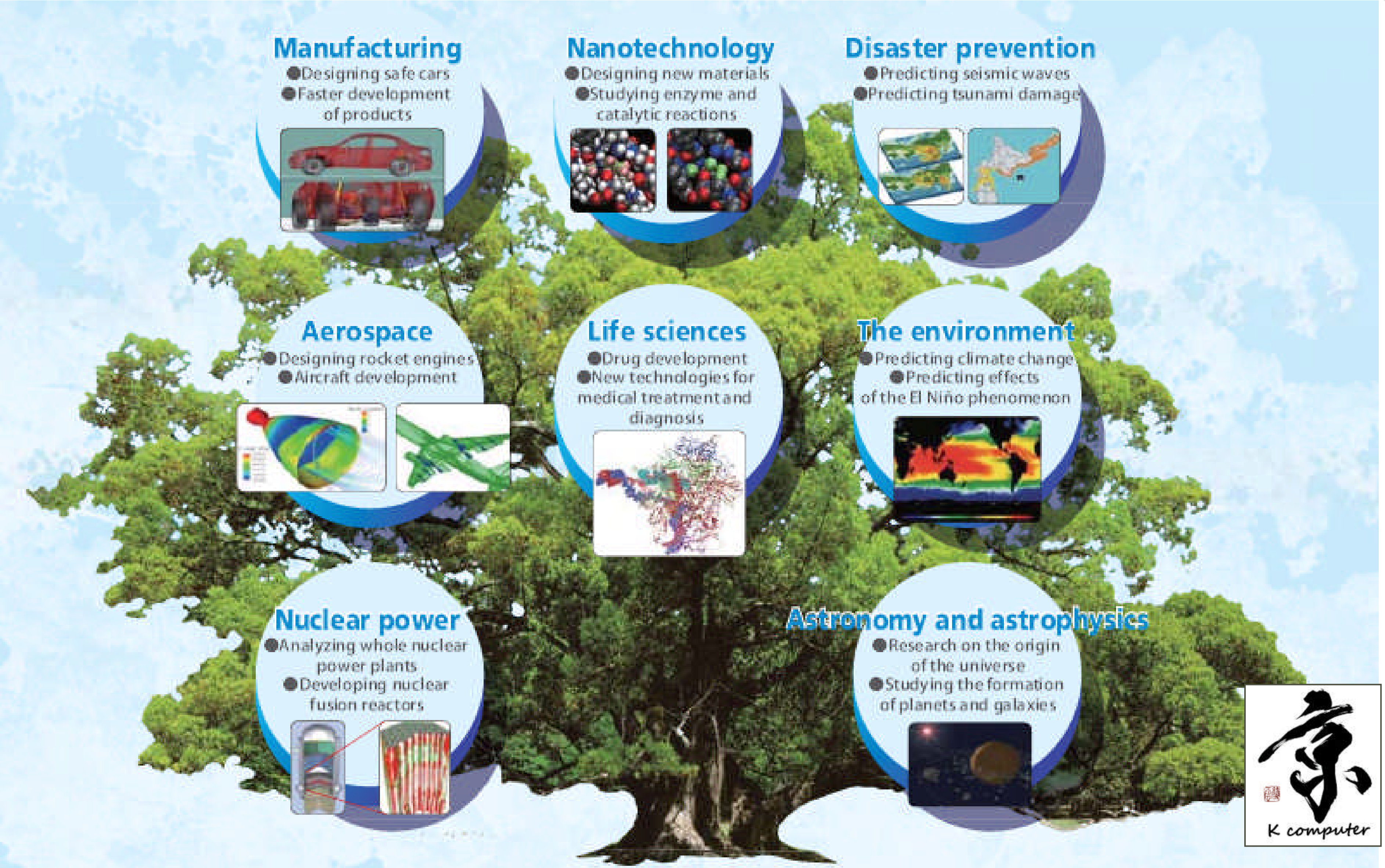
“京” (Kei) computer

K computer



Full system installation (CG image)

Applications of K computer



Manufacturing

- Designing safe cars
- Faster development of products

Nanotechnology

- Designing new materials
- Studying enzyme and catalytic reactions

Disaster prevention

- Predicting seismic waves
- Predicting tsunami damage

Aerospace

- Designing rocket engines
- Aircraft development

Life sciences

- Drug development
- New technologies for medical treatment and diagnosis

The environment

- Predicting climate change
- Predicting effects of the El Niño phenomenon

Nuclear power

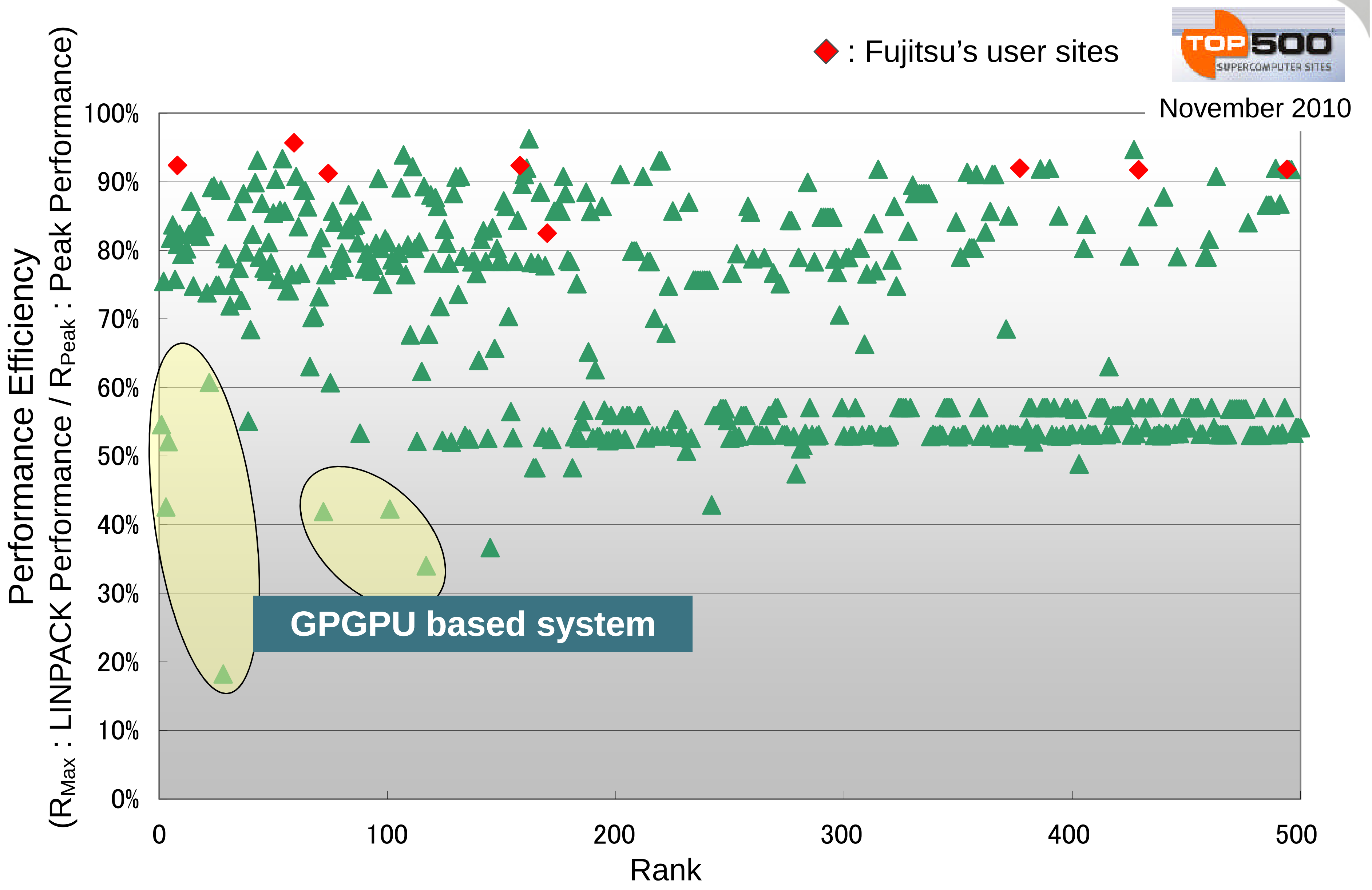
- Analyzing whole nuclear power plants
- Developing nuclear fusion reactors

Astronomy and astrophysics

- Research on the origin of the universe
- Studying the formation of planets and galaxies

Courtesy of RIKEN

TOP500 Performance Efficiency



Design Targets

- Toward Application Centric Petascale Computing -

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

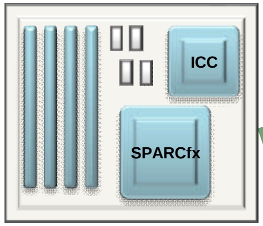
K computer Specifications

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

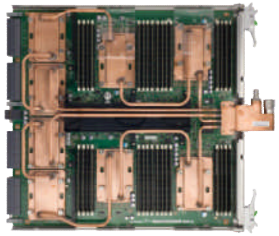
CPU

128GFlops
SPARC64™ VIIIfx
8 Cores@2.0GHz



Node

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

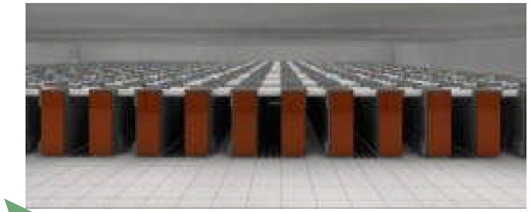


System Board

512 GFlops
64 GB memory

Rack

12.3 TFlops
15TB memory

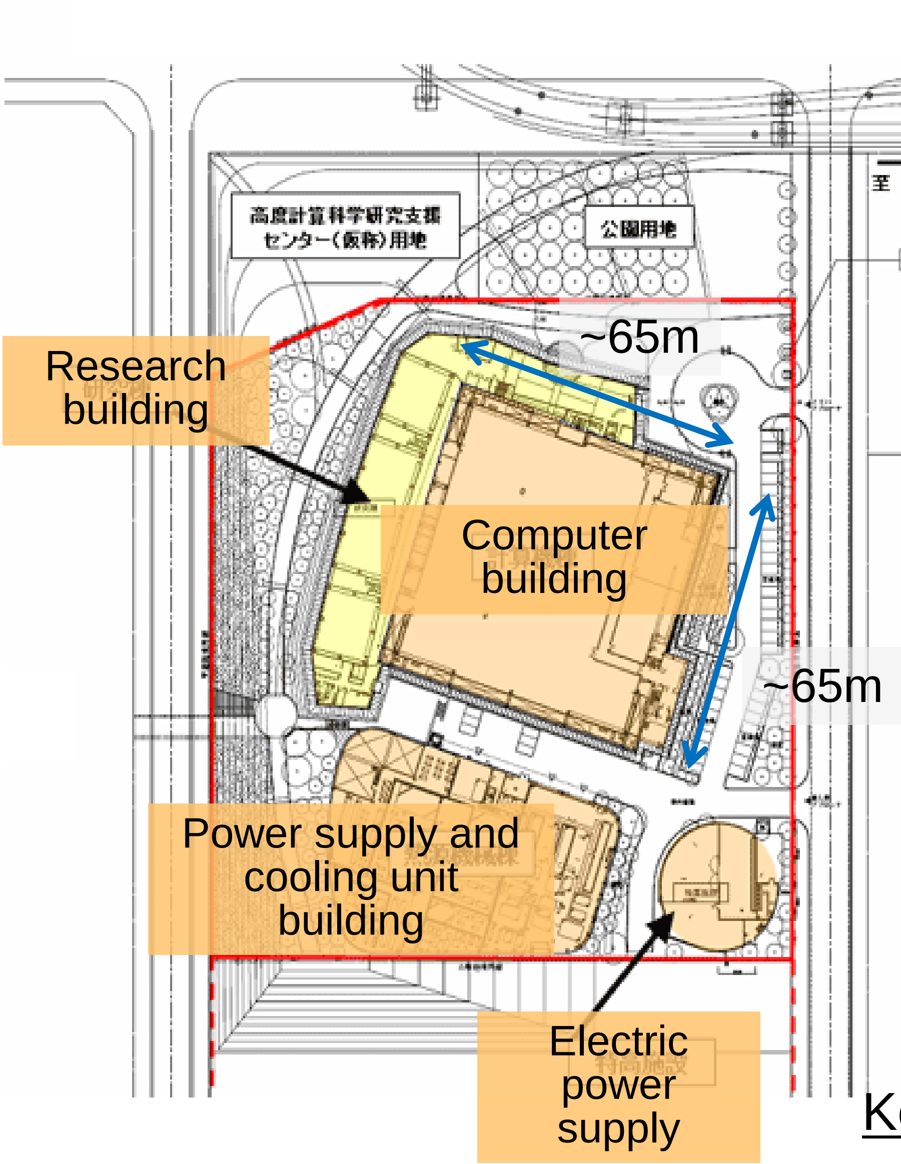


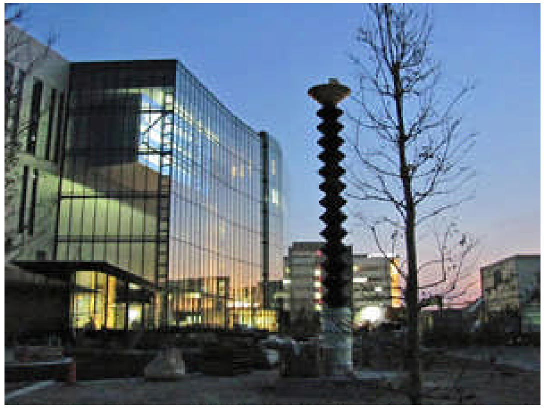
System

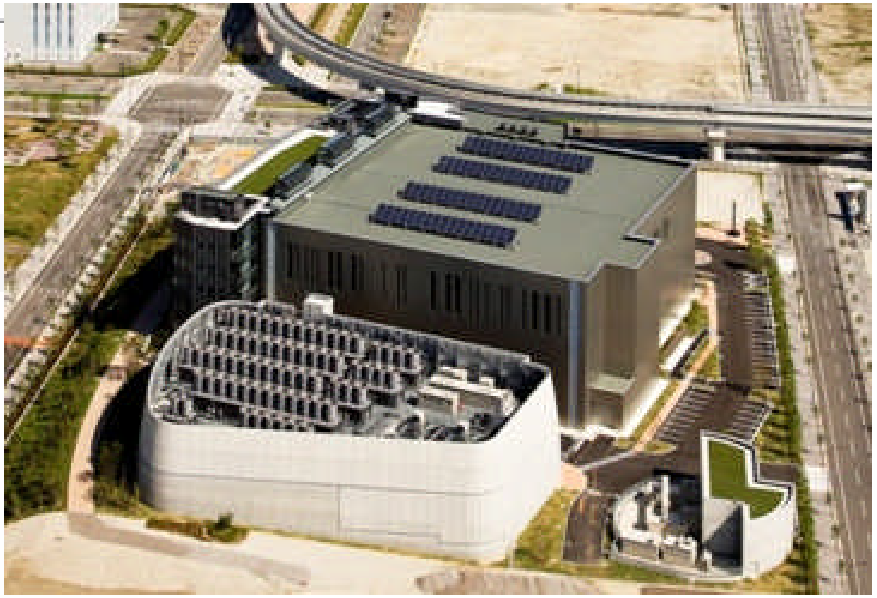
LINPACK 10 PFlops
over 1PB mem.
800 racks
80,000 CPUs
640,000 cores

* ICC : Interconnect Chip

Kobe Facilities



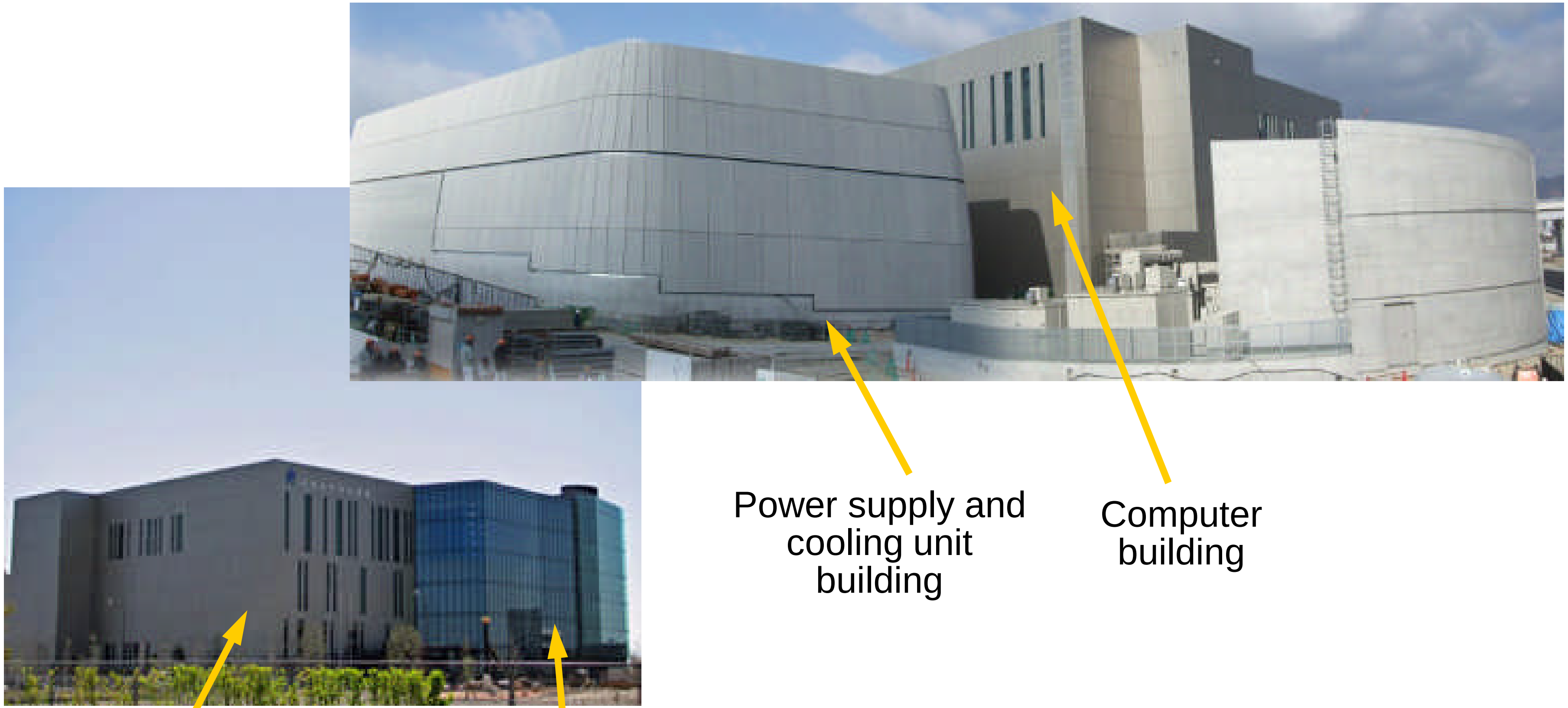




Kobe site ground plan and aerial photo

Courtesy of RIKEN

Kobe Facilities (cont.)



Power supply and cooling unit building

Computer building

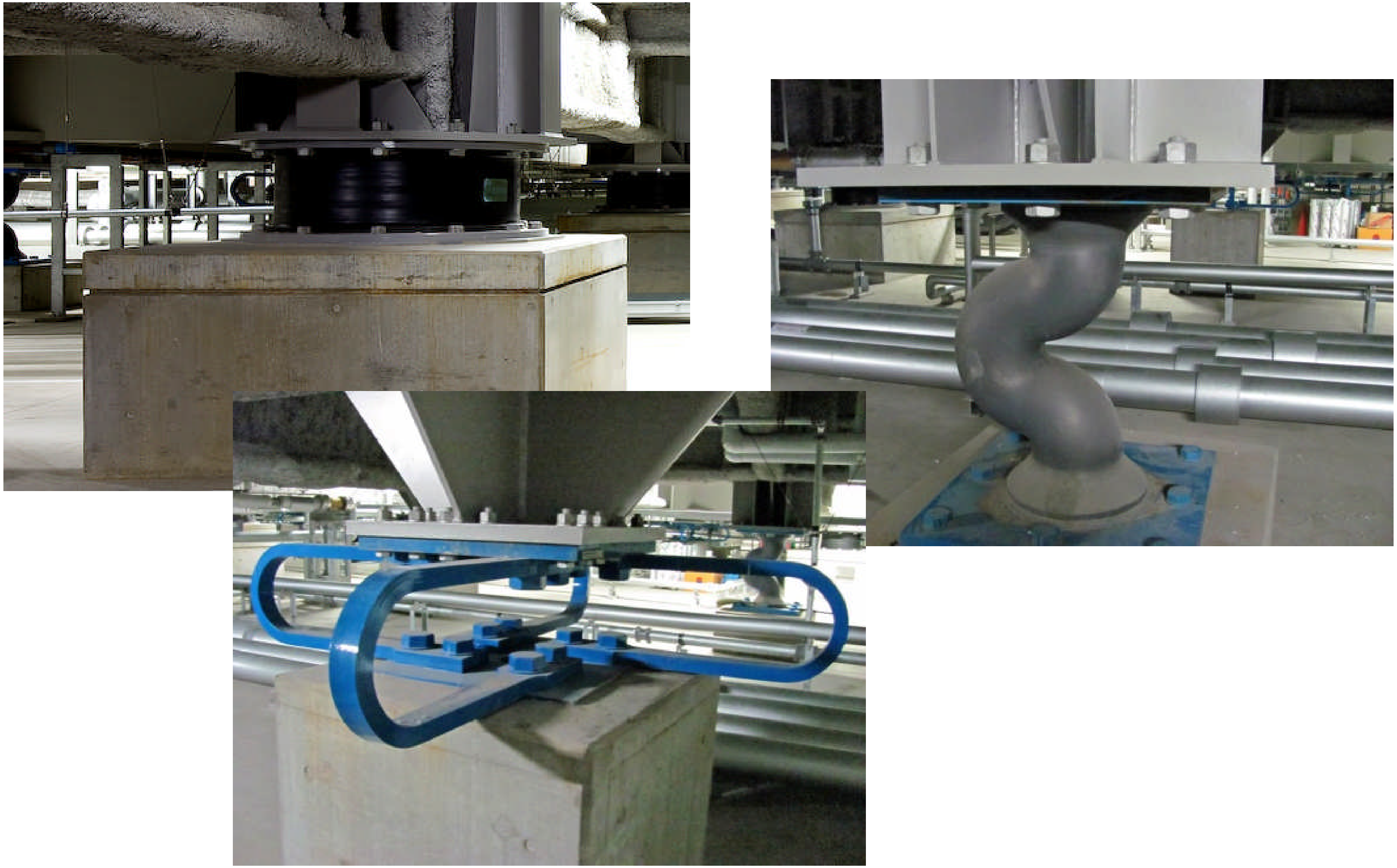
Computer building

Research building

Exterior of buildings

Courtesy of RIKEN

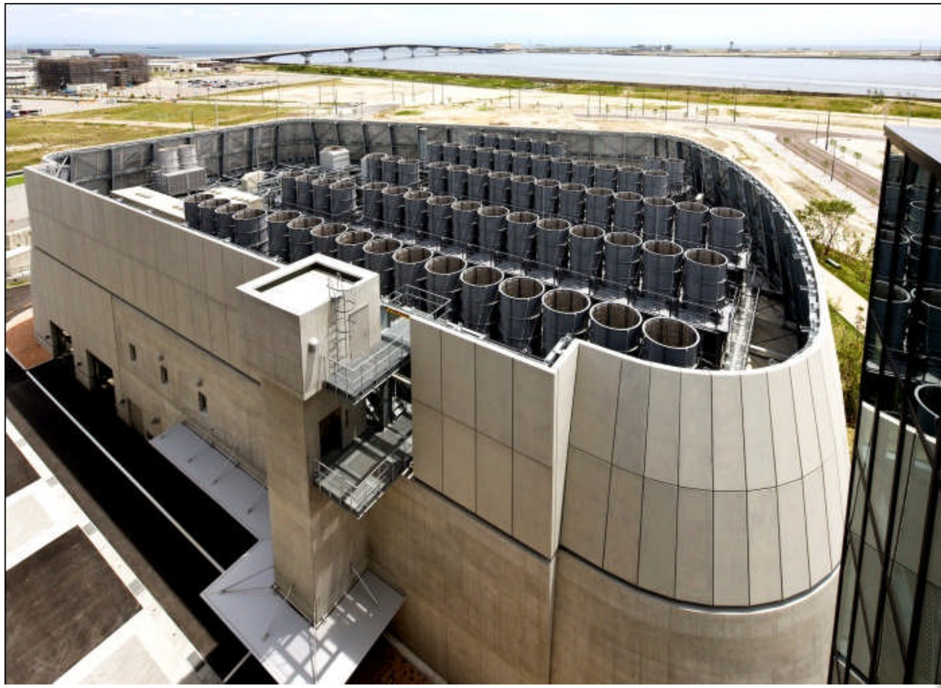
Kobe Facilities (cont.)



Seismic isolation structure

Courtesy of RIKEN

Kobe Facilities (cont.)



Cooling towers

Air Handling Units
(Computer building 2F)

Centrifugal chillers

Power Supply and Cooling Unit Building

Courtesy of RIKEN

Kobe Facilities (cont.)

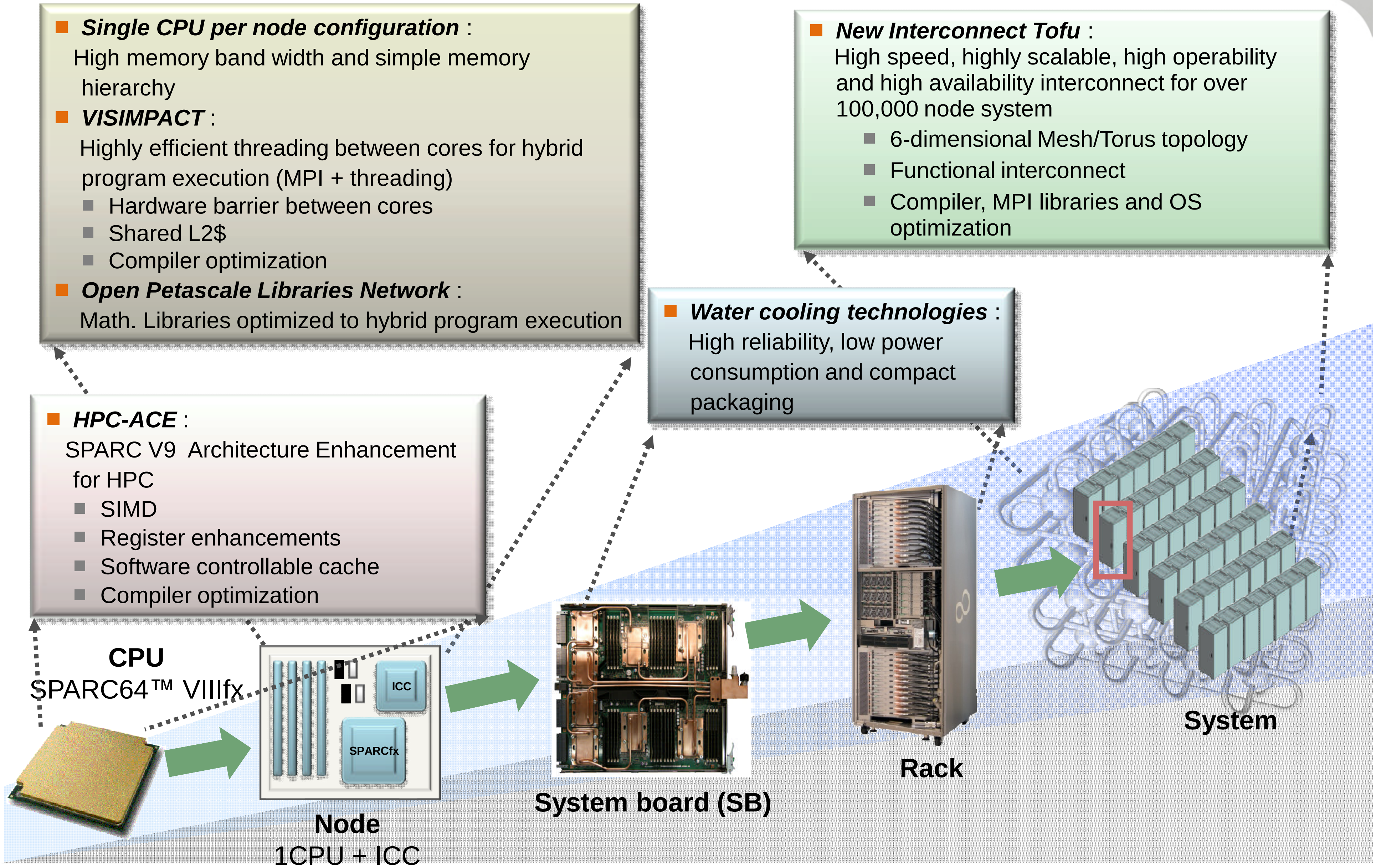
On Oct. 1st, First 8 racks were installed at Kobe site, RIKEN

Courtesy of RIKEN

Technologies
for
Application Centric Petascale Computing

- CPU
- VISIMPACT
- Interconnect

Technologies for Application Centric Petascale Computing



- ***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

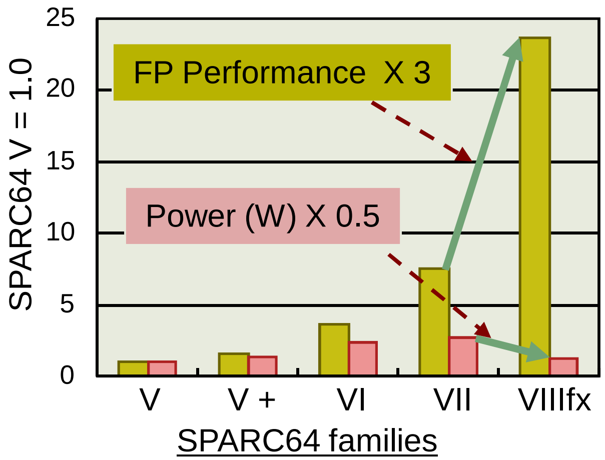
- ***New Interconnect Tofu*** :
  High speed, highly scalable, high operability and high availability interconnect for over 100,000 node system
  - 6-dimensional Mesh/Torus topology
  - Functional interconnect
  - Compiler, MPI libraries and OS optimization

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

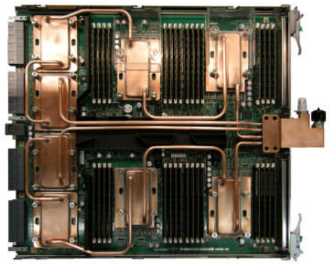
- ***HPC-ACE*** :
  SPARC V9 Architecture Enhancement for HPC
  - SIMD
  - Register enhancements
  - Software controllable cache
  - Compiler optimization

SPARC64^{TM} VIIIfx Processor

- Extended SPARC64^{TM} 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 per watt
  - 2 GHz clock, 128 GFlops
  - 58 Watts peak as design target
- Water cooling
  - Low current leakage of the CPU
  - Low power consumption and low failure rate of CPUs
- High reliable design
  - SPARC64^{TM} VIIIfx integrates specific logic circuits to detect and correct errors



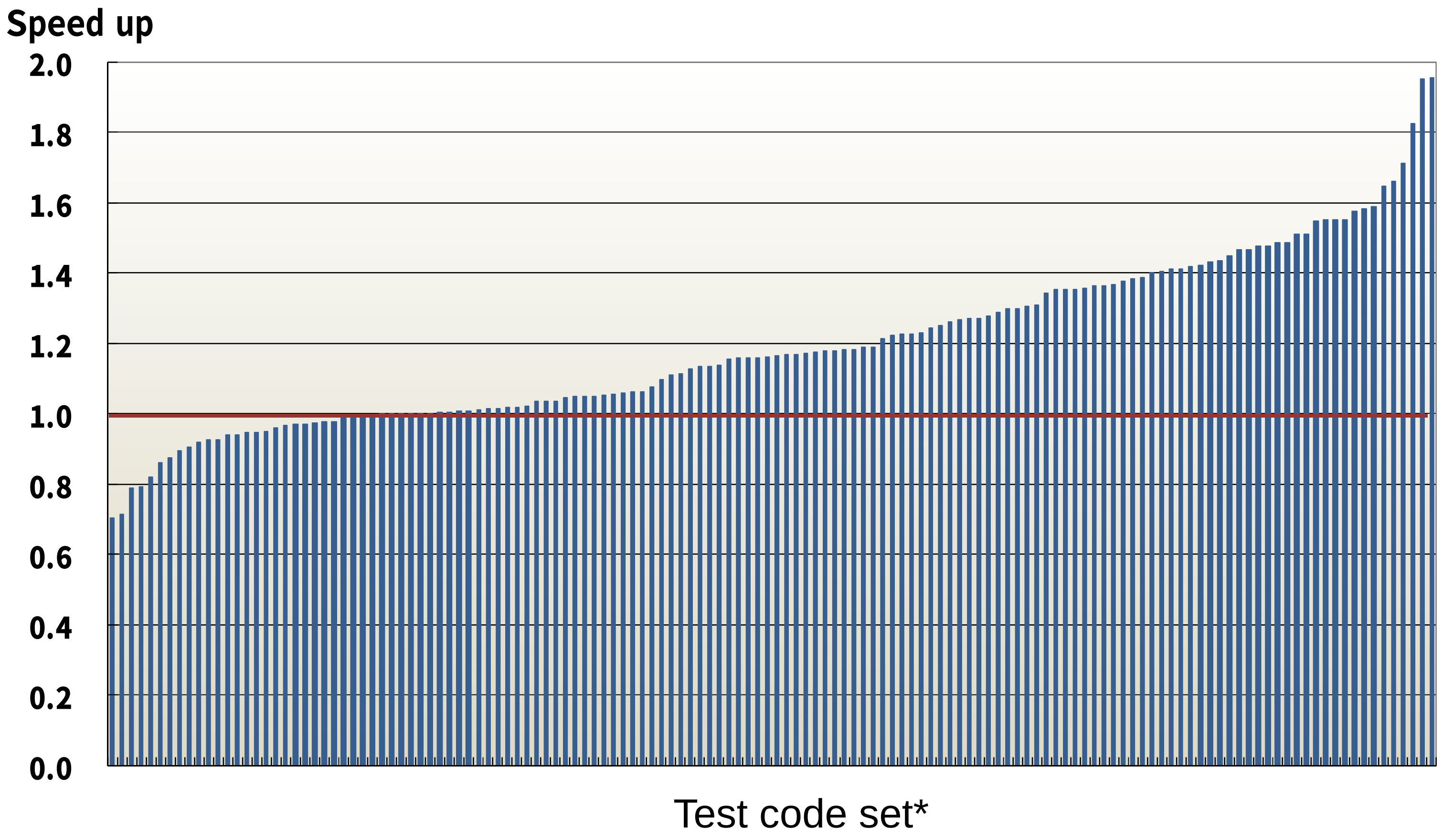
History of Peak Performance & Power



Direct water cooling System Board

SIMD Extension (1)

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

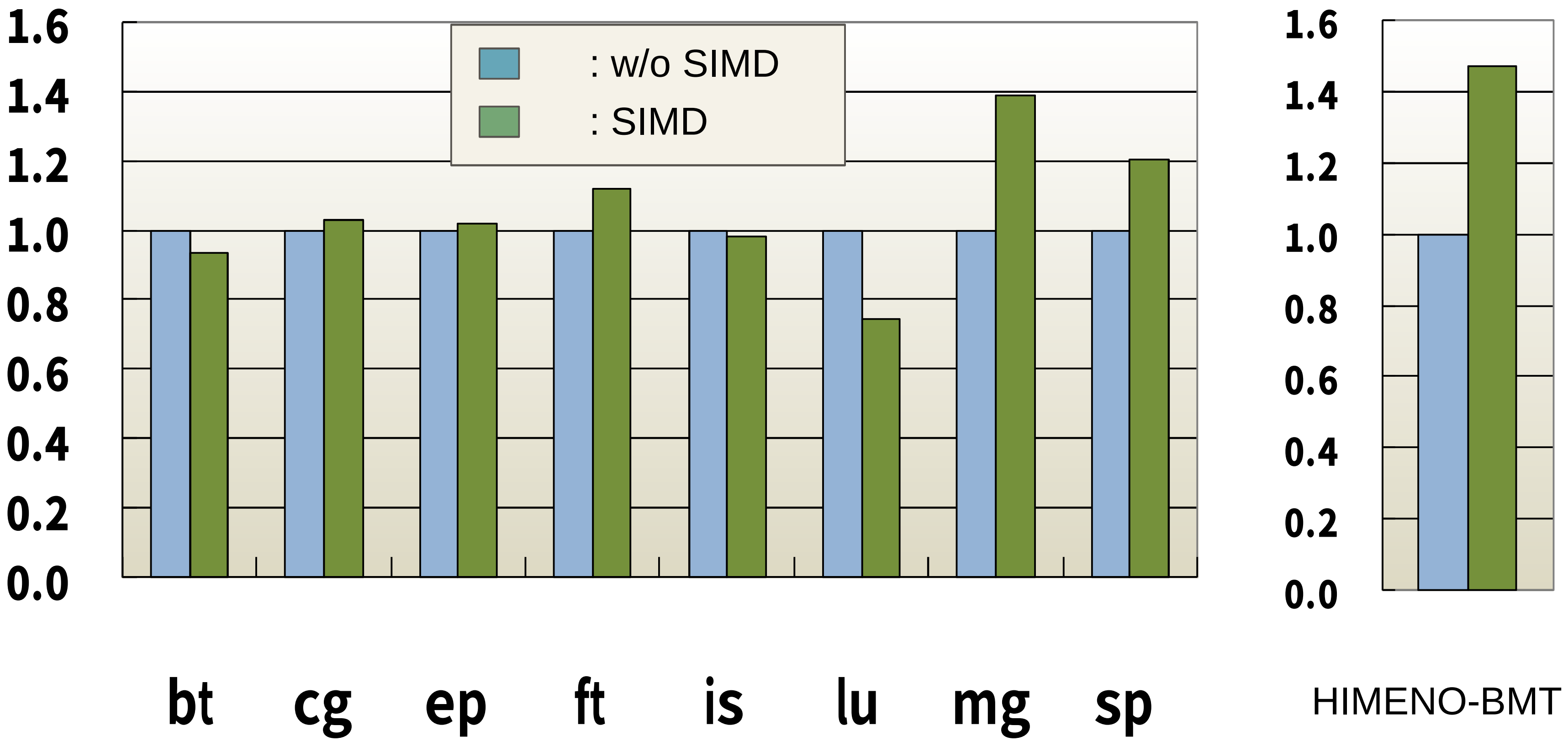


Effect of SIMD extension on one core of SPARC64^{TM} VIIIfx

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

SIMD Extension (2)

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

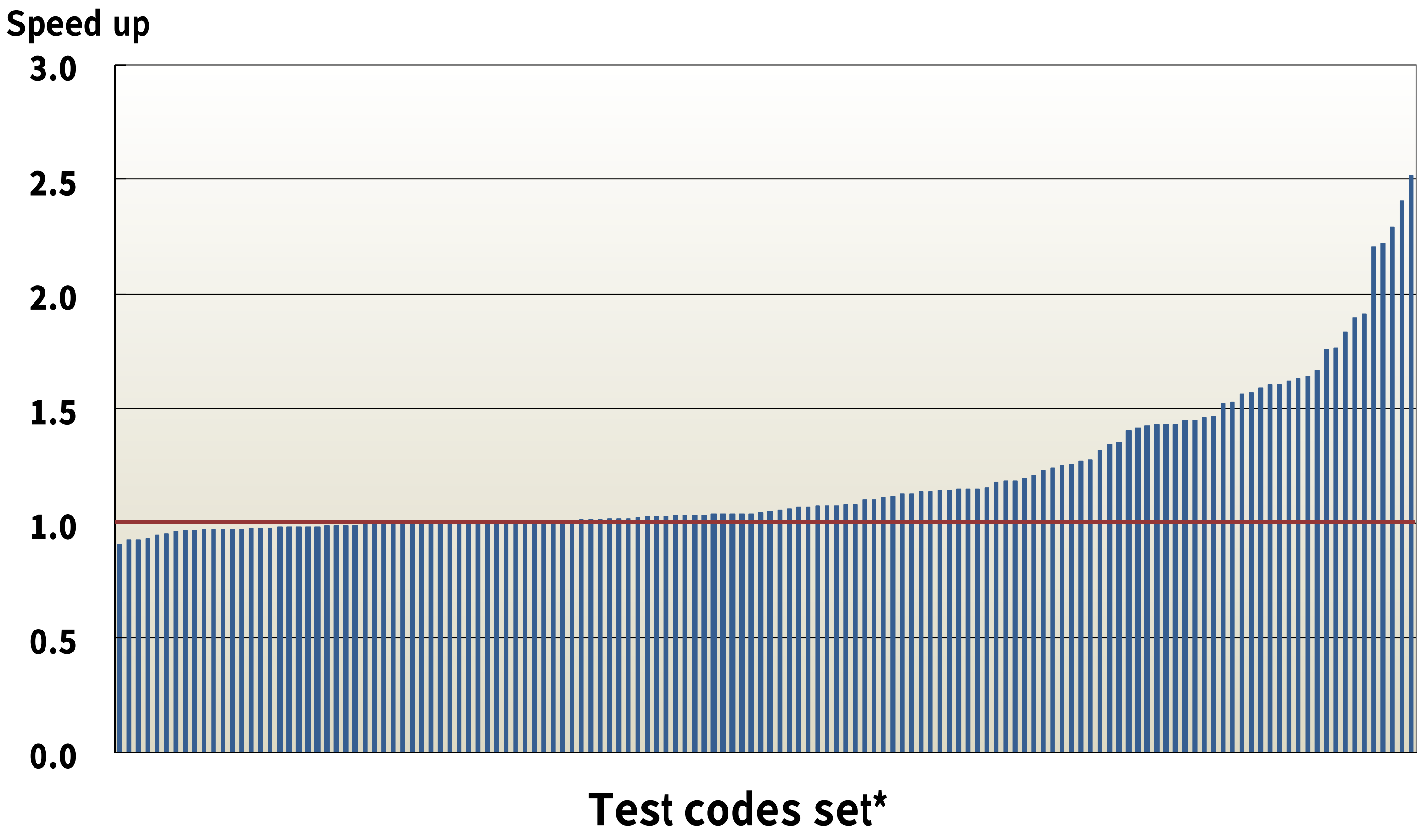


Effect of SIMD extension on one core of SPARC64^{TM} VIIIfx

* : 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.

Floating Point Registers Extension (1)

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

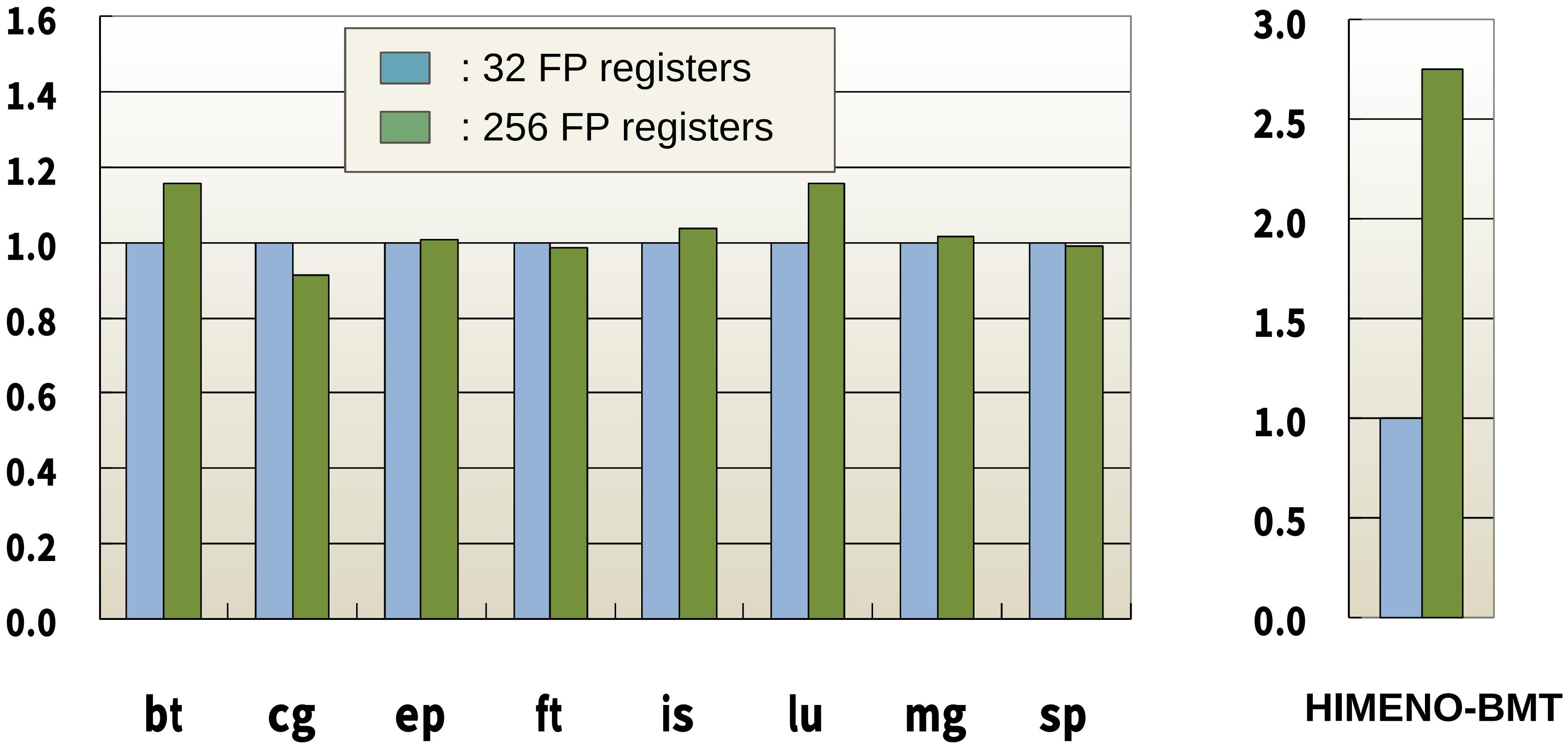


Effect of register size extension (32 to 256) on one core of SPARC64^{TM} VIIIfx

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

Floating Point Registers Extension (2)

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

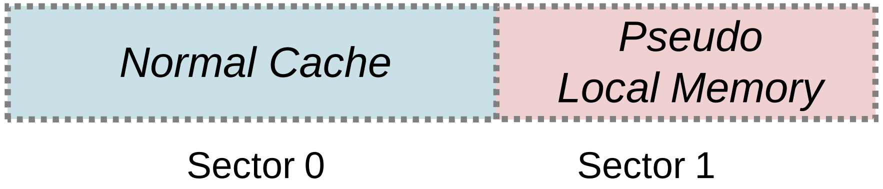


Effect of register size extension on one core of SPARC64^{TM} VIIIfx

* : 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.

Application Access to Cache Management

- 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



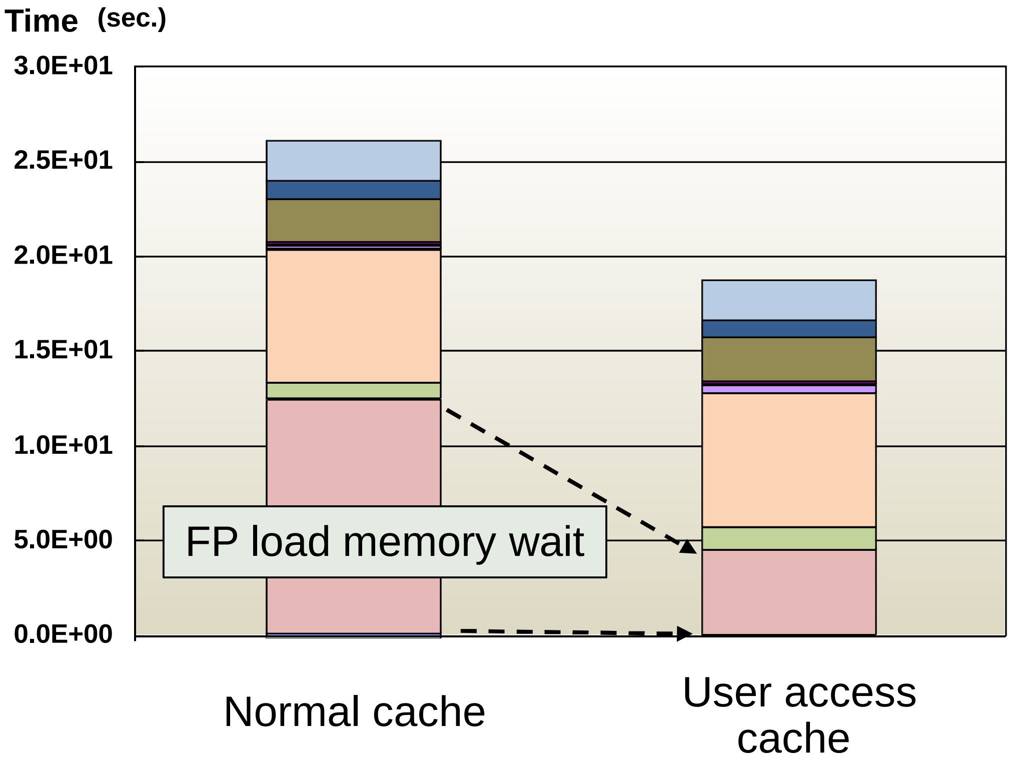
Normal Cache

Pseudo Local Memory

Sector 0

Sector 1

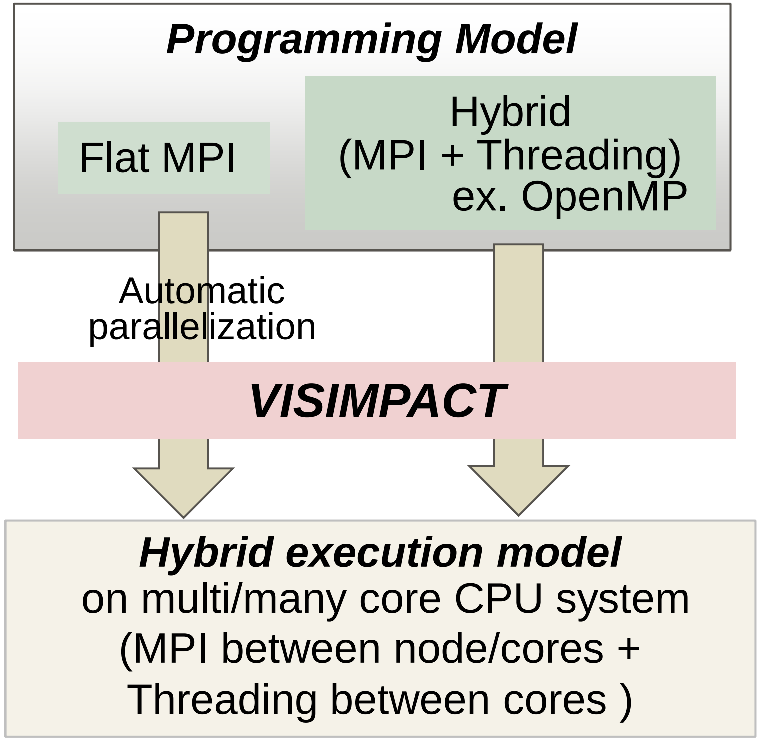
```
39                c--------------------------
40                !ocl cache_sector_size (3, 9)
41    1   s   s          do iter=1, itmax
42    1   s   s            call sub(a, b, c, s, n, m)
43    1   s   s          enddo
44                c--------------------------
:
~~~~~~~~~~~
52                     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

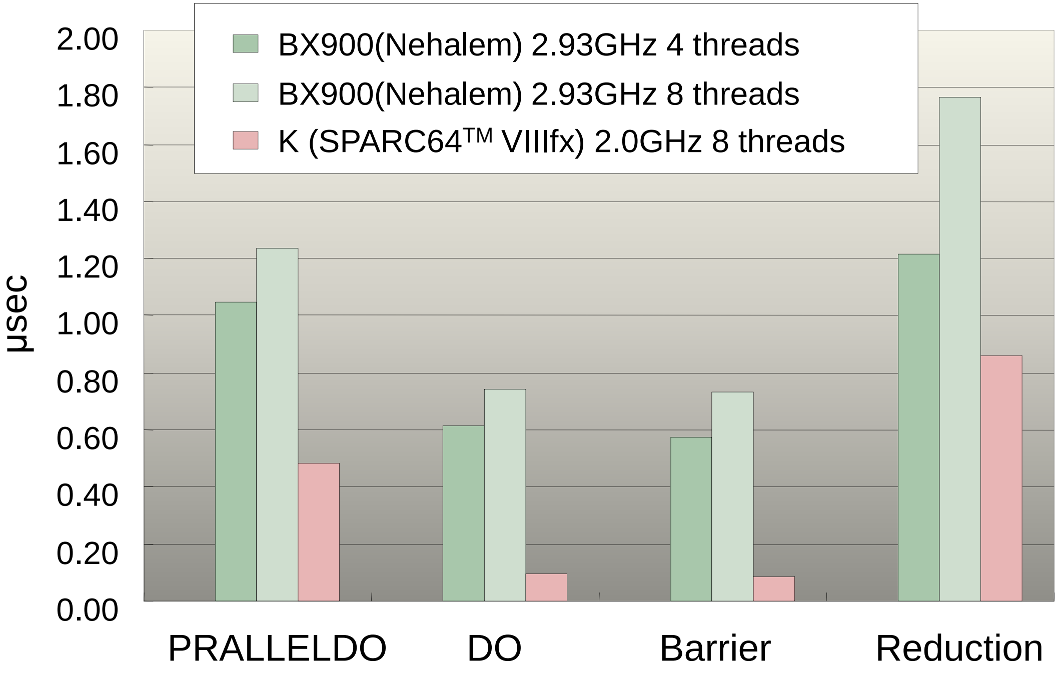
Performance of VISIMPACT (Integrated Multi-core Parallel ArChiTecture)

- 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)



Programming Model
Flat MPI
Hybrid (MPI + Threading) ex. OpenMP
Automatic parallelization
VISIMPACT

Hybrid execution model
on multi/many core CPU system
(MPI between node/cores +
Threading between cores)



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

Interconnect for Petascale Computing

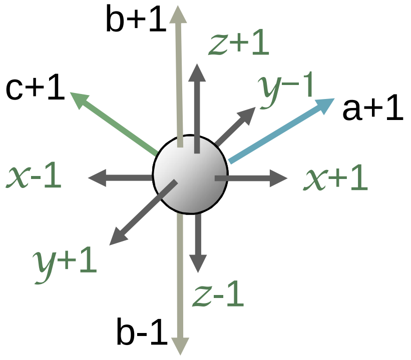
topology Characteristics	Cross bar	Fat-Tree/ Multi stage	Mesh / Torus
Performance	Best	Good	Average
Operability and Availability	Best	Good	Weak
Cost and Power consumption	Weak	Average	Good
Topology uniformity	Best	Average	Good
Scalability	Hundreds nodes Weak	Thousands nodes Ave.-Good	>10,000 nodes Best
Example	Vector Parallel	x86 Cluster	Scalar Massive parallel

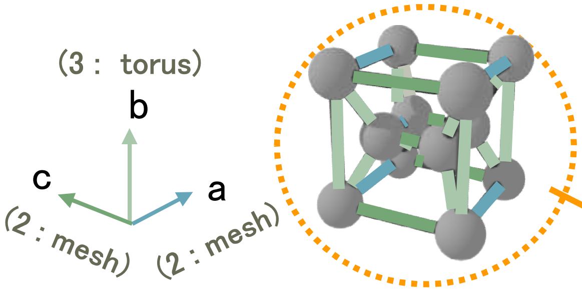
- Which type of the topology can scale up over 100,000 node?

➡ Improvement of the performance, operability and availability of mesh/torus topology are our challenge

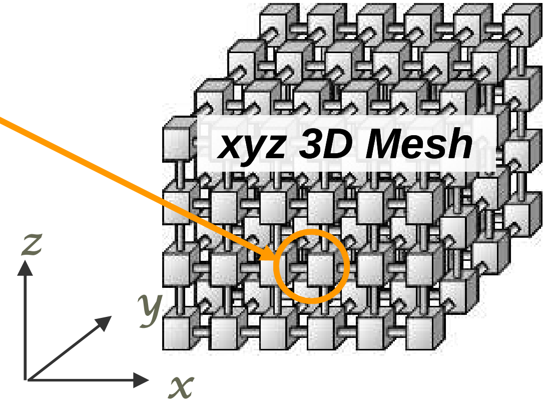
New Interconnect (1) : *Tofu Interconnect*

- 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

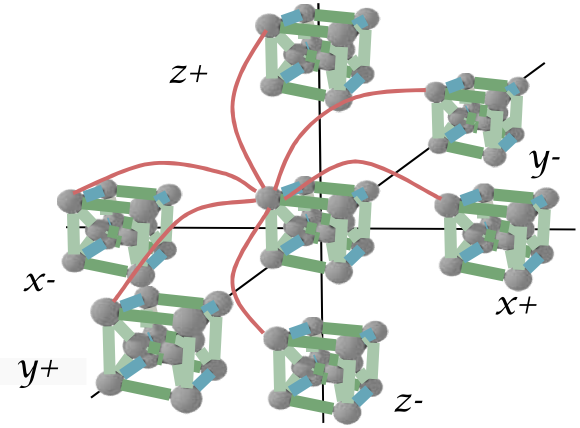




Node Group Unit
(12 nodes group, 2 x 3 x 2)



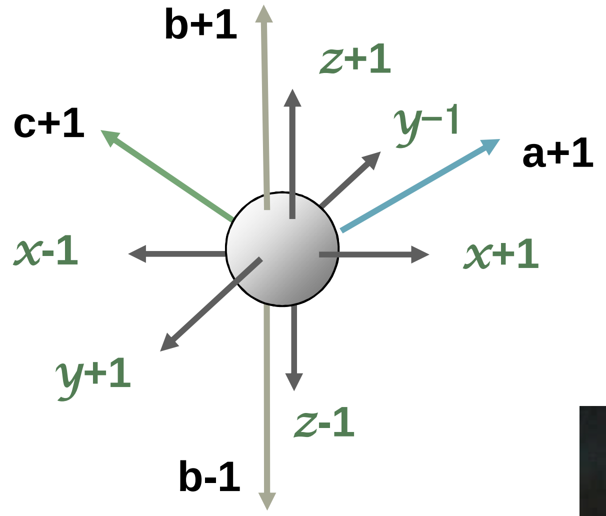
Node Group 3D connection



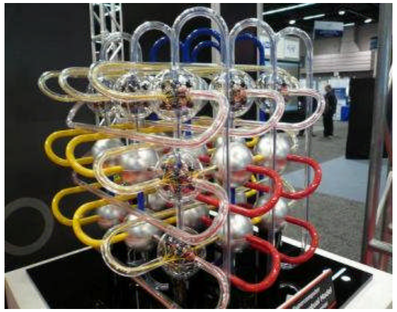
3D connection of each node

New Interconnect (2) : *Tofu Interconnect*

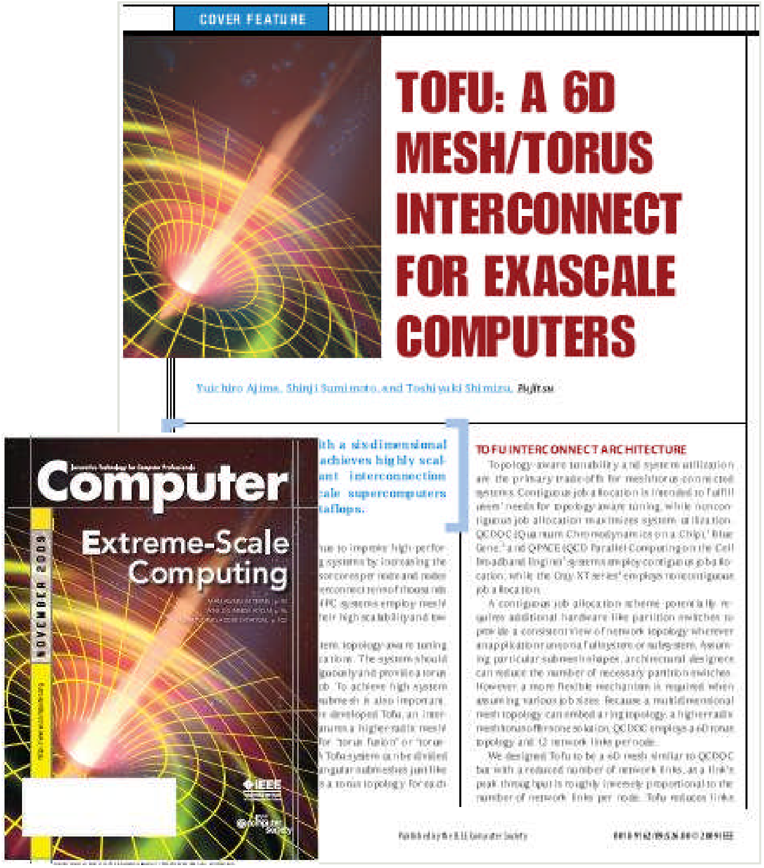
- Technology
  - Fast node to node communication : *5 GB/s x 2 (bi-directional) /link, 100GB/s. throughput /node*
  - Integrated MPI support for collective operations and global hardware barrier
  - Switch less implementation



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



Conceptual Model



COVER FEATURE

TOFU: A 6D MESH/TORUS INTERCONNECT FOR EXASCALE COMPUTERS

Yuichiro Ajima, Shinji Sumimoto, and Toshiyuki Shimizu, Fujitsu

ith a six-dimensional achieves highly scal- ant interconnection cale supercomputers taflops.

...ue to improve high-perfor- g systems by increasing the sor cores per node and nodes erconnect terms of thousands HPC systems employ mesh/ heir high scalability and low

tem, topology-aware tuning cations. The system should guously and provide a torus ob. To achieve high system submesh is also important. e developed Tofu, an inter- atures a higher-radix mesh/ for "torus fusion" or "torus A Tofu system can be divided angular submeshes just like s a torus topology for each

TOFU INTERCONNECT ARCHITECTURE

Topology-aware tunability and system utilization are the primary trade-offs for mesh/torus-connected systems. Contiguous job allocation is intended to fulfill users' needs for topology-aware tuning, while noncontiguous job allocation maximizes system utilization. QCDOC (Quantum Chromodynamics on a Chip),[1] Blue Gene,[2] and QPACE (QCD Parallel Computing on the Cell Broadband Engine)[3] systems employ contiguous job allocation, while the Cray XT series[4] employs noncontiguous job allocation.

A contiguous job allocation scheme potentially requires additional hardware like partition switches to provide a consistent view of network topology wherever an application runs on a full system or subsystem. Assuming particular submesh shapes, architectural designers can reduce the number of necessary partition switches. However, a more flexible mechanism is required when assuming various job sizes. Because a multidimensional mesh topology can embed a ring topology, a higher-radix mesh/torus offers one solution; QCDOC employs a 6D torus topology and 12 network links per node.

We designed Tofu to be a 6D mesh similar to QCDOC but with a reduced number of network links, as a link's peak throughput is roughly inversely proportional to the number of network links per node. Tofu reduces links

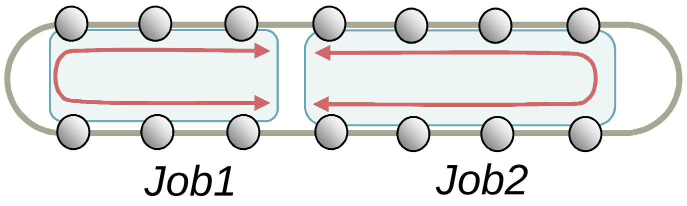
Published by the IEEE Computer Society

IEEE Computer Nov. 2009

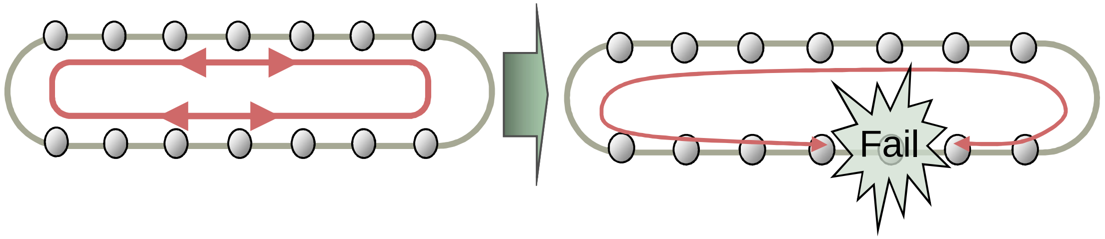
Why 6 dimensions?

- 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__

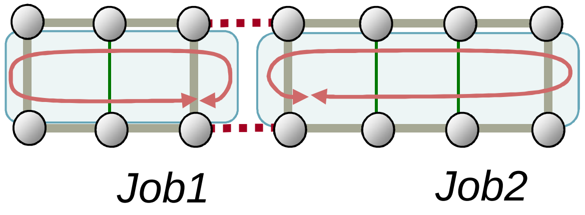


High hop-count & mesh

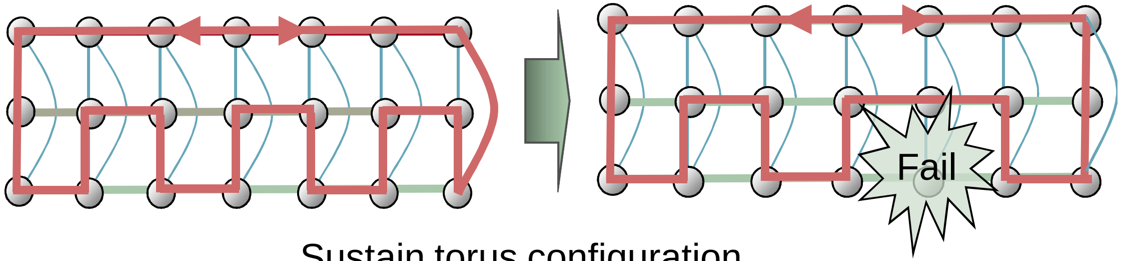


Torus topology can't be configured

__Tofu interconnect__



Low hop-count & keep torus
with job isolation



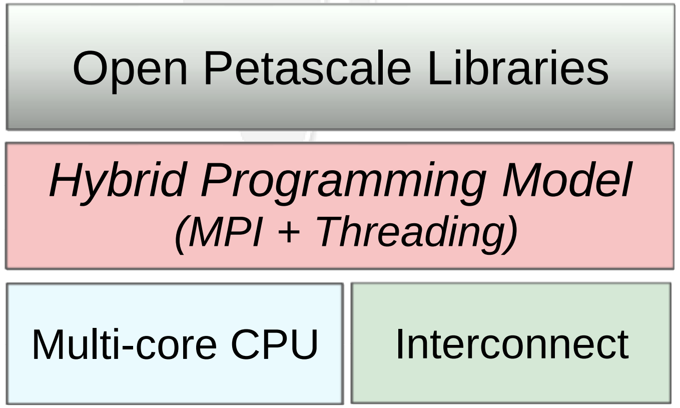
Sustain torus configuration

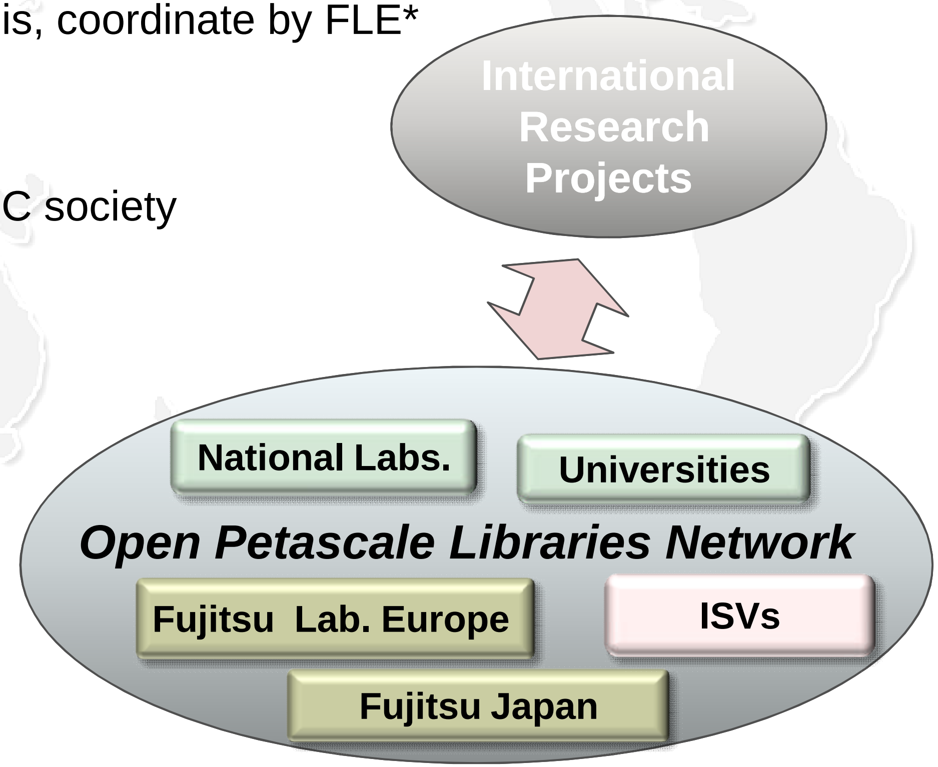
Open Petascale Libraries Network

FUJITSU

- 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

openpetascale LIBRARIES




International Research Projects
National Labs.
Universities
Open Petascale Libraries Network
Fujitsu Lab. Europe
ISVs
Fujitsu Japan

*: Fujitsu Labs Europe, located in London

Conclusion

Toward Application Centric Petascale Computing

- Installation of RIKEN’s K Computer has started and the system is targeting
  - High performance
  - Environmental efficiency
  - High productivity
- Leading edge technologies are applied to K computer
  - New CPU
  - Innovative interconnect
  - Advanced packaging
  - Open Petascale Libraries
- Those technologies shall be enhanced and applied to Fujitsu’s future commercial supercomputer

京

K computer

FUJITSU

shaping tomorrow with you