Fujitsu's challenge for Petascale Computing

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

System performance = Processor performance \times Number of processors

- **Many cores CPU or accelerator approach**
- **High-end general purpose CPU approach**
- **Our approach**
  - Give priority to application migration!
- **Low power consumption embedded processor approach**

\[
\begin{align*}
\text{Peak performance per processor (Gflops)} & \quad \text{Number of processors} \\
100Tflops & \quad 10Tflops & \quad 1Pflops & \quad 10Pflops
\end{align*}
\]
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

**Step 1: 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**

**Step 2: 2011 ~**

- **Petascale system with new high performance, high 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

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

**PRIMERGY**
- BX Series
- RX Series
- HX600
- IA/Linux

**SPARC Enterprise**
- SPARC Enterprise M9000
- SPARC64™ VII 64cpu

**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

**FX1**
- SPARC64™ VII
- sparc64

**SPARC/Solaris**
- IA/Linux

**RG1000**
- FPGA board
- NEW

**NEW**
Customers of large scale TC systems

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

<table>
<thead>
<tr>
<th>Customer</th>
<th>Type</th>
<th>No. of CPU</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan Aerospace Exploration Agency (JAXA)</td>
<td>Cluster Scalar SMP</td>
<td>&gt;3,500CPU</td>
<td>135TFlops</td>
</tr>
<tr>
<td></td>
<td>Manufacturer A</td>
<td>&gt;3,500CPU</td>
<td>&gt;80TFlops</td>
</tr>
<tr>
<td>KYOTO University Computing Center</td>
<td>Cluster Scalar SMP</td>
<td>&gt;2,000CPU</td>
<td>&gt;61.2TFlops</td>
</tr>
<tr>
<td>KYUSYU University Computing Center</td>
<td>Scalar SMP Cluster</td>
<td>1,824CPU</td>
<td>32TFlops</td>
</tr>
<tr>
<td>Manufacturer B</td>
<td>Cluster</td>
<td>&gt;1,200CPU</td>
<td>&gt;15TFlops</td>
</tr>
<tr>
<td>RIKEN</td>
<td>Cluster</td>
<td>3,088CPU</td>
<td>26.18TFlops</td>
</tr>
<tr>
<td>NAGOYA University Computing Center</td>
<td>Scalar SMP</td>
<td>1,600 CPU</td>
<td>13TFlops</td>
</tr>
<tr>
<td>TOKYO University KAMIOKA Observatory</td>
<td>Cluster</td>
<td>540CPU</td>
<td>12.9TFlops</td>
</tr>
<tr>
<td>National Institute of Genetic</td>
<td>Cluster Scalar SMP</td>
<td>324CPU</td>
<td>6.9TFlops</td>
</tr>
<tr>
<td>Institute for Molecular Science</td>
<td>Scalar SMP</td>
<td>320CPU</td>
<td>4TFlops</td>
</tr>
</tbody>
</table>
Latest case study

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

Peak Performance:
- 61.2 TFLOPS
- Memory: 13TB

Peak Performance:
- 8.96 TFLOPS
- Memory: 7TB

Peak throughput: 16GB/s
- Storage capacity: 883TB

Fiber channel switch

InfiniBand switch

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

I/O Server
- PRIMEQUEST580 [x 1] (32CPU)

LINPACK Performance: 50.5 TFLOPS (Efficiency 82.5%)

Cluster subsystem
- HX600 [x 416]
- QuadCore Opteron x 4/node
- InfiniBand DDR x 4/node

Disk Array
- ETERNUS2000 M200 [x 32]
FX1 Launch customer

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

**FX1 (3,392 nodes)**

135 TFlops

**THIN nodes**

**FAT node(SMP)**

1 TFlops

**SPARC Enterprise**

**SPARC Enterprise FX1**

3,392 nodes

First system will be installed at JAXA by the end of 2008
FX1 : New High-end TC Server - Outline -

- **High-performance CPU designed by Fujitsu**
  - SPARC64™ 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

![Diagram of CPU CHIP with L2 caches and cores]

**Advantage**
- Handles the multi-core CPU as one equivalent faster CPU
  - Reduces number of MPI processes to $1/n_{\text{core}}$ 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 of the number of cores

  ![Graph showing barrier overhead comparison between hardware and software]

- 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**
  - Applicability: wide
  - Overhead: frequent but low cost

- **Legacy parallelization on scalar machine**
  - Applicability: narrow (required wide range analysis)
  - Synchronization: occasional

- **Fine-grain parallelization on scalar machine**
  - Applicability: wide
  - Synchronization: frequent

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

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

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

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)**
  - \( n = 100 \quad \Rightarrow \quad 3.26 \text{ Gflops} \)
  - \( n = 40,000 \quad \Rightarrow \quad 37.8 \text{ Gflops (93.8%) } \)

- **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](image-url)
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
**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
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 nodes system

- Average
  - Software: 82.99 μ sec
  - Intelligent Switch: 11.86 μ sec

Graph showing performance over the number of reductions.
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

- High end TC server FX1 (>100TFlop/s)
- Petascale system (>10PFlop/s)

SMP System

- SPARC Enterprise
  - SPARC64™ VII
  - PRIMEQUEST
  - Itanium® 2

PC cluster

- PRIMERGY
  - Xeon
  - Opteron

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

ORAP Forum, 9th Oct. 2008
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

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

Source: RIKEN official report

RIKEN Next-Generation Supercomputer R&D Center

Development & Application of Next-Generation Supercomputer Project by MEXT

FY2006: 3,547 Million yen / FY2007: 7,736 Million yen

FY2006~FY2012 (total budget expected) about 110 billion 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).

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

<table>
<thead>
<tr>
<th>Year</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
</table>

- **System and Middleware**
  - NAREGI: Grid Project led by NII
  - Primary R&D projects for Next Generation Supercomputer
  - Major industry contributor

- **Next Generation Supercomputer Project**
  - Targeting LINPACK 10PFlops and led by RIKEN
  - Collaborative joint research of architecture
  - Grand design
  - Detailed design
  - Production

- **Application Software**
  - Life Science Application project led by RIKEN
  - Nano Science Application project led by IMS
  - CAE Application project led by IIS

**R&D and application optimization**
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
  - Drug design
  - Multi-sphere simulation of human body

- 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)

Through the courtesy of RIKEN
Japanese Next Generation Supercomputer Project
Basic Concept for Simulations in Life Sciences

Meso
- Micro-machine
- Tissue Structure
- Bio-MD
- Genome

Micro
- Gene Therapy

Organ
- Multi-physics
- HIFU
- DDS

Organism
- Blood Circulation
- Catheter

Vascular System
- Macro

Cell
- Tissue

Under total combination of simulation physics...

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

System performance = Processor performance x Number of processors

10Pflops

1 Pflops

100 Tflops

10 Tflops

1000

100

10

1

1,000

10,000

100,000

1,000,000

Peak performance per processor (Gflops)

Number of processors

Many cores CPU or accelerator approach

High-end general purpose CPU approach

Our approach
Give priority to application migration!

Low power consumption embedded processor approach

LANL Roadrunner

NMCAC SGI Altix ICE8200

ES

ASC Purple P5-575

JUGENE BG/P

LLNL BG/L

Our approach
Give priority to application migration!
Fujitsu’s Challenges for Petascale Supercomputer

**Fujitsu high-end CPU**

- 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

User

- Applications
- Will be ready for Petascale computing environment

Application developer

- Performance & environmental requirement
- Applications adapted for Petascale system
History of Fujitsu High–end Processor

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

### Mainframe Processor

- **GS8600**
  - Tr = 30M CMOS Al 250nm / 220nm

- **GS8800**
  - Tr = 45M CMOS Cu 180nm
  - Tr = 200M CMOS Cu 130nm

- **GS8800B**
  - Tr = 30M CMOS Cu 130nm

- **SPARC64 IV**
  - Tr = 10M CMOS Al 350nm

- **SPARC64 IV+**
  - Tr = 200M CMOS Cu + Low-k 90nm

- **GS21**
  - Tr = 500M CMOS Cu 90nm

- **SPARC64 V**
  - Tr = 45M CMOS Cu 130nm

- **SPARC64 V+**
  - Tr = 30M CMOS Cu 180nm / 150nm

- **GS8900**
  - Tr = 500M CMOS Cu + Low-k 90nm

- **GS21**
  - Tr = 540M CMOS Cu + Low-k 90nm

- **SPARC64 V+**
  - Tr = 200M CMOS Cu + Low-k 90nm

- **SPARC64 VI**
  - Tr = 200M CMOS Cu 130nm

- **SPARC64 VII**
  - Tr = 400M CMOS Cu + Low-k 90nm

**RAS Coverage of SPARC64 VII**
- Guaranteed Data Integrity

- 1bit error Correctable
- 1bit error Detectable
- 1bit error harmless
Interconnect for parallel computer system

- Interconnect type and its characteristic

<table>
<thead>
<tr>
<th>Interconnect type</th>
<th>Crossbar</th>
<th>Fat-Tree</th>
<th>Mesh / Torus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>◎ (Best)</td>
<td>○ (Good)</td>
<td>△ (Average)</td>
</tr>
<tr>
<td>Operability and usability</td>
<td>◎ (Best)</td>
<td>○ (Good)</td>
<td>× (Weak)</td>
</tr>
<tr>
<td>Cost, Packaging density and Power consumption</td>
<td>× (Weak)</td>
<td>△ (Average)</td>
<td>○ (Good)</td>
</tr>
<tr>
<td>Scalability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hundreds nodes</td>
<td></td>
<td>Thousands nodes</td>
<td>&gt;10,000 nodes</td>
</tr>
<tr>
<td>X (Weak)</td>
<td></td>
<td>Δ - ○ (Ave.-Good)</td>
<td>◎ (Best)</td>
</tr>
<tr>
<td>Representative</td>
<td>Vector Parallel</td>
<td>PC cluster</td>
<td>Scalar Massive parallel</td>
</tr>
</tbody>
</table>

- 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
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