Fujitsu's Challenge for Petascale Computing

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

Our approach
Give priority to application migration!

System performance
= Processor performance
  x Number of processors

Many cores CPU or accelerator approach

High-end general purpose CPU approach

Low power consumption embedded processor approach

Fujitsu's Approach for Scaling up to 10 PFlops

Peak performance per processor (GFlops)

Number of processors

10 PFlops
1 PFlops
1000
100
10
1
1,000,000
1,000,000
100,000
10,000
1,000
100
10
1

LANL Roadrunner
JUGENE
ES
NMCAC SGI Altix ICE8200
LLNL BG/L
LANL
P5 575
ASC Purple
Longhorn

10 TFlops
100 TFlops
1 PFlops
1,000,000
10,000
100,000
10,000
100
1

10 TFlops
100 TFlops
1 PFlops
1,000,000
10,000
100,000
10,000
100
1

Many cores CPU or accelerator approach

High-end general purpose CPU approach

Low power consumption embedded processor approach
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

**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, 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

#### High-end TC Solutions
- Scalability up to 100 TFlops class
- Highly effective performance
- High-end RISC CPU

#### 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

---

**Solidware Solutions**
- Ultra high performance for specific applications

**PRIMERGY**
- BX Series
- TX Series
- HX600

**FX1**
- SPARC64™ VII

**PRIMEQUEST**
- 580 Itanium® 2
- ~32cpu

**SPARC Enterprise**
- SPARC Enterprise M9000
- SPARC64™ VII
- ~64cpu
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 (FX1)</td>
<td>&gt;3,500</td>
<td>135 TFlops</td>
</tr>
<tr>
<td></td>
<td>Scalar SMP (SPARC Enterprise)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>*This system will be installed in end of 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manufacturer A</td>
<td>Scalar SMP Cluster</td>
<td>&gt;3,500</td>
<td>&gt;80 TFlops</td>
</tr>
<tr>
<td>KYOTO Univ. Computing Center</td>
<td>Cluster (HX600)</td>
<td>&gt;2,000</td>
<td>&gt;61.2 TFlops</td>
</tr>
<tr>
<td></td>
<td>Scalar SMP (SPARC Enterprise)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KYUSHU Univ. Computing Center</td>
<td>Scalar SMP (PRIMEQUEST)</td>
<td>1,824</td>
<td>32 TFlops</td>
</tr>
<tr>
<td></td>
<td>Cluster (PRIMERGY)</td>
<td>&gt;1,200</td>
<td>&gt;15 TFlops</td>
</tr>
<tr>
<td>Manufacturer B</td>
<td>Cluster (PRIMERGY)</td>
<td>3,088</td>
<td>26.18 TFlops</td>
</tr>
<tr>
<td>RIKEN</td>
<td>Cluster (PRIMERGY)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAGOYA Univ. Computing Center</td>
<td>Scalar SMP (HPC2500)</td>
<td>1,600</td>
<td>13 TFlops</td>
</tr>
<tr>
<td>TOKYO Univ. KAMIOKA Observatory</td>
<td>Cluster (PRIMERGY)</td>
<td>540</td>
<td>12.9 TFlops</td>
</tr>
<tr>
<td>National Institute of Genetics</td>
<td>Cluster (PRIMERGY)</td>
<td>324</td>
<td>6.9 TFlops</td>
</tr>
<tr>
<td></td>
<td>Scalar SMP (SPARC Enterprise)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institute for Molecular Science</td>
<td>Scalar SMP (PRIMEQUEST)</td>
<td>320</td>
<td>4 TFlops</td>
</tr>
</tbody>
</table>
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) SPARC Enterprise  1 TFlops

Hardware barrier between nodes

High Speed Intelligent Interconnect Network

I/O & front end servers SPARC Enterprise

RAID subsystem ETERNUS

FC bus

System Control Server

power/facility control

LAN
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
Integrated Multi-core Parallel Architecture

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 $1/n_{\text{core}}$ 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 number of cores
  - 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

**SPARC64™ VII**
Real quad-core CPU for Technical Computing
(2.5 GHz, 40 GFlops/chip)
Integrated Multi-core Parallel ArChitect, 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

![Graph showing 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

![Performance comparison chart](chart.png)
Integrated Multi-core Parallel ArChiTecture, preliminary measured data

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

![Graphs showing performance comparison of NPB class C between pure MPI and MPI + IMPACT.](image)
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 node system

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

---

Supercomputer

- High end TC server FX1 (>100 TFlops)
- Petascale system (>10 PFlops)

SMP System

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

PC cluster

- PRIMERGY
  - Xeon
  - Blade server with InfiniBand
  - PC cluster of 4 sockets 16 ways with enhanced interconnect

- NEW Sparc64
- Solaris
- Linux
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

Development & Application of Next-Generation Supercomputer Project by MEXT

~$1.2 B

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>
<tbody>
<tr>
<td><strong>System and Middleware</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAREGI: Grid Project led by NII</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary R&amp;D projects for Next Generation Supercomputer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Major industry contributor&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grid operation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;R&amp;D for Petascale System Interconnect&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;Scalar system development&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Collaborative joint research of architecture |
| Grand design |

| Next Generation Supercomputer Project |
| targeting LINPACK 10 PFlops and led by RIKEN |
| 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 code optimization"
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

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
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

**User**
- Will be ready for Petascale computing environment

**Application developer**
- Applications adapted for Petascale system
- Applications
- Performance & environmental requirement
- Applications

---

IDC HPC User Forum, Oct. 16th, 2008

All Rights Reserved, Copyright FUJITSU LIMITED 2008
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

**RAS Coverage of SPARC64VII**
Guaranteed Data Integrity
- : 1bit error correctable
- : 1bit error detectable
- : 1bit error harmless

<table>
<thead>
<tr>
<th>Year</th>
<th>Model</th>
<th>Speed</th>
<th>Technology</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>GS8600</td>
<td>10M</td>
<td>CMOS AI 350 nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GS8800</td>
<td>30M</td>
<td>CMOS AI 250 nm / 220 nm</td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>SPARC64II</td>
<td>45M</td>
<td>CMOS Cu 180 nm</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>SPARC64I</td>
<td>500M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>SPARC64V</td>
<td>200M</td>
<td>CMOS Cu 130 nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GS8900B</td>
<td>200M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SPARC64GP</td>
<td>45M</td>
<td>CMOS Cu 180 nm / 220 nm</td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>GS21</td>
<td>540M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SPARC64VI</td>
<td>300M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SPARC64V+</td>
<td>400M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>GS21</td>
<td>500M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SPARC64VI</td>
<td>450M</td>
<td>CMOS Cu + Low-k 90 nm</td>
<td></td>
</tr>
</tbody>
</table>
Interconnect for Parallel Computer System

- **Interconnect type and characteristics**

<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>X Weak</td>
</tr>
<tr>
<td>Cost, packaging density and power consumption</td>
<td>X Weak</td>
<td>△ Average</td>
<td>○ Good</td>
</tr>
<tr>
<td>Scalability</td>
<td>Hundreds nodes</td>
<td>Thousands nodes</td>
<td>&gt;10,000 nodes</td>
</tr>
<tr>
<td>Representative</td>
<td>Vector Parallel</td>
<td>PC cluster</td>
<td>Scalar Massive parallel</td>
</tr>
</tbody>
</table>

- **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 PFIops performance, Fujitsu has embarked on the Petascale Computing challenge