Numerical Simulation Research in NIFS and Fujitsu’s New Supercomputer PRIMEHPC FX100

Masanori Nunami\textsuperscript{a}, Toshiyuki Shimizu\textsuperscript{b}

\textsuperscript{a}National Institute for Fusion Science, Japan
\textsuperscript{b}Fujitsu, Japan
Outline

M. Nunami (NIFS)

- Overview of Numerical Simulation Reactor Research Project in NIFS
- Supercomputers in NIFS and Japan

T. Shimizu (Fujitsu)

- Fujitsu’s new supercomputer (FX100)
  - Features and evaluations
  - Application evaluation
- The next step in exascale capability
Research activities in NIFS

In NIFS, there exists 7 divisions and 3 projects for fusion research activities.

- **LHD project**
- **Numerical Simulation Research Project**
- **Fusion Engineering Research Project**

**Department of Helical Plasma Research**

- **High-Density Plasma Physics Research Division**
- **High-Temperature Plasma Physics Research Division**
- **Plasma Heating Physics Research Division**
- **Device Engineering and Advanced Physics Research Division**
- **Fusion Systems Research Division**
- **Fusion Theory and Simulation Research Division**
- **Fundamental Physics Simulation Research Division**
- **Rokkasho Research Center**
Numerical Simulation Reactor Research Project (NSRP)

NSRP

has been launched to continue the tasks in simulation group in NIFS, and evolve them on re-organization of NIFS in 2010.

Final Goal

Based on large-scale numerical simulation researches by using a super-computer,
- To understand and systemize physical mechanisms in fusion plasmas
- To realize ultimately the “Numerical Simulation Reactor” which will be an integrated predictive modeling for plasma behaviors in the reactor.

Realization of numerical test reactor requires

- understanding all element physics in each hierarchy,
- inclusion of all elemental physical processes into our numerical models,
- development of innovative simulation technology to interlock all hierarchies and all physical elements.

⇒ Multi-scale model, multi-hierarchy model, cross coupling model of core and periphery regions
Numerical Simulation Reactor Research Project

Numerical Reactor

Comparison with experiments

Integration

Core plasma

High energy particle

MHD

transport

kinetics

Core-periphery, multi-elements, multi-physics interlocked model

Plasma-wall interaction

Periphery plasma

Edge transport

Visualization

Multi-scale, compound physics

Hierarchy interlocked

Visualization

Core-periphery, multi-elements, multi-physics interlocked model

Integration

Comparison with experiments

Core plasma

High energy particle

MHD

transport

kinetics

Core-periphery, multi-elements, multi-physics interlocked model

Plasma-wall interaction

Periphery plasma

Edge transport

Visualization

Multi-scale, compound physics

Hierarchy interlocked
### Supercomputers in NIFS and Japan

<table>
<thead>
<tr>
<th></th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>2018</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NIFS</strong></td>
<td>Plasma</td>
<td>New</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simulator</td>
<td>Plasma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SR16000</td>
<td>Simulator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(315TF)</td>
<td>FX100(2.6 PF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RIKEN</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>K computer (10PF)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IFERC-CSC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Helios (1.2 PF)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Plasma Simulator** (NIFS, Fujitsu FX100) : 2.62 PFlops / 82 TB
- **K-computer** (RIKEN, Fujitsu SPARC 64 Vlllfx): 10.62 PFlops / 1.26 PB
- **Helios** (IFERC-CSC, Bull B510) : 1.52 PFlops / 256 TB
Fujitsu’s New Supercomputer PRIMEHPC FX100
Fujitsu’s new supercomputer PRIMEHPC FX100, delivering the next step in Exascale capability

Toshiyuki Shimizu

August 21st, 2015
## Past, PRIMEHPC FX100, and roadmap for Exascale

### K computer and PRIMEHPC FX10 in operation

Many applications running and being developed for science and industries.

### PRIMEHPC FX100 is ready

CPU and interconnect inherits K computer architectural concept.

### Towards Exascale

RIKEN selected Fujitsu as a partner for basic design of Post-K computer.

### Japan’s national projects

<table>
<thead>
<tr>
<th>Development</th>
<th>Operation of K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPCI strategic applications program</td>
<td></td>
</tr>
<tr>
<td>App. review</td>
<td>FS projects</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### PRIMEHPC FX10 and FX100

<table>
<thead>
<tr>
<th>Year</th>
<th>PRIMEHPC FX10</th>
<th>PRIMEHPC FX100</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 1.85 x CPU performance
- Easier installation
- Improved CPU & network performance
- High-density packaging & low power consumption

Copyright 2015 FUJITSU LIMITED
PRIMEHPC FX100, design concept and approach

Provide steady progress for users

- Natural extent of performance profile of K computer and FX10
- Facilitate the evolution of applications

Challenge to state-of-art technologies for future generation

- 20nm CMOS technology
- HMC
- 25G optical connection
Natural extent of perf. profile of K and FX10

Original high performance CPU for wide range of real applications

Highly scalable interconnect

<table>
<thead>
<tr>
<th></th>
<th>FX100</th>
<th>FX10</th>
<th>K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double Flops / CPU</td>
<td>Over 1 TF</td>
<td>235 GF</td>
<td>128 GF</td>
</tr>
<tr>
<td>Single Flops / CPU</td>
<td>Over 2 TF</td>
<td>235 GF</td>
<td>128 GF</td>
</tr>
<tr>
<td>Max. # of threads</td>
<td>32</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Memory / process</td>
<td>32 GB</td>
<td>32 GB</td>
<td>16 GB</td>
</tr>
<tr>
<td>SIMD width</td>
<td>256 bit</td>
<td>128 bit</td>
<td>128 bit</td>
</tr>
<tr>
<td>Byte per flop</td>
<td>0.4 ~ 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interconnect</td>
<td>Tofu 6D mesh/torus</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interconnect BW</td>
<td>12.5 GB/s</td>
<td>5 GB/s</td>
<td>5 GB/s</td>
</tr>
</tbody>
</table>

Compatibility with K computer and PRIMEHPC FX10

Binary compatible and make full use of performance by recompile
Compiler and libraries allow users to access new features
Feature and configuration of FX100

**Tofu interconnect 2**
- 12.5 GB/s × 2 (in/out)/link
- 10 links/node
- Optical technology

**CPU Memory Board**
- Three CPUs
- 3 × 8 Micron’s HMCs
- 8 opt modules, for inter-chassis connections

**Cabinet**
- Up to 216 nodes/cabinet
- High-density
- 100% water cooled with EXCU (option)

**Fujitsu designed SPARC64 XIIfx**
- 1TF~(DP)/2TF~(SP)
- 32 + 2 core CPU
- HPC-ACE2 support
- Tofu2 integrated

**Chassis**
- 1 CPU/1 node
- 12 nodes/2U Chassis
- Water cooled
System software for PRIMEHPC FX10 and FX100

Tuned Linux OS for HPC applications
• Supports large pages and OS jitter minimization

Combination of self-developed software and customized OSS
• System management software and languages are self-developed
• File system and MPI are developed based on OSS and the results were fed back to the communities

Single system images with x86 and hybrid configurations

System management portal and HPC portal

Technical Computing Suite

Management
- System management
  • Single system image
  • Single action IPL
  • Fail safe capability
- Job management
  • Highly efficient scheduler

File system (FEFS)
- Lustre based
- Higher scalability (thousands of I/O servers)
- Higher I/O performance (1.4 TB/s)

Programing environment
- Compiler
  • Fortran, XPF, C, C++
  • Automatic parallelization
  • SIMD support
- MPI: OpenMPI based
- Tools and math libraries
NIFS Plasma Simulator

- In HPCG (High Performance Conjugate Gradients), Plasma Simulator is ranked 2nd in Japan and 12th in the world as of June 2015.

- In TOP500 Supercomputing Sites, Plasma Simulator is ranked 3rd in Japan and 27th in the world as of June 2015.

http://www-nsrp.nifs.ac.jp/news-e.html
Features and evaluations

- HMC
- SIMD
- Tofu interconnect 2
- Assistant core
- Real applications
Hybrid Memory Cube (HMC) support

HMC
- Higher density at BW
- Higher capacity and higher BW at package
- Lower power consumption at BW

Comparable capacity and bandwidth to those of K computer and FX10

<table>
<thead>
<tr>
<th>System</th>
<th>K computer</th>
<th>FX10</th>
<th>FX100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory device</td>
<td>DDR3 x 8</td>
<td>DDR3 x 8</td>
<td>HMC x 8</td>
</tr>
<tr>
<td>Capacity</td>
<td>16GB</td>
<td>32GB/64GB</td>
<td>32GB</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>64GB/s</td>
<td>85GB/s</td>
<td>480GB/s</td>
</tr>
<tr>
<td>Byte/Flop</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Improving memory throughput

By using HMC, node memory throughput increase $3x$, $4x$
SIMD extension of HPC-ACE2

256-bit wide SIMD with 128 FPRs

- Double precision x 4, single precision x 8, 8-byte integer x 4
- Stride Load/Store
- Indirect (list) Load/Store
- Permutation, Concatenate

---

SIMD

- reg S1
- reg S2
- reg D

Stride load

- Memory
- Specified stride
- reg D

Indirect load

- reg S
- Memory
- reg D

Permutation

- reg S
- Arbitrary shuffle
- reg D

Copyright 2015 FUJITSU LIMITED
Wider SIMD extensions

- DP 3x, SP 6x faster than FX10 in basic kernels
  - Improved L1 cache pipelines

### Basic kernels performance per core

<table>
<thead>
<tr>
<th></th>
<th>FX10</th>
<th>FX100 Double</th>
<th>FX100 Single</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcast</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector Copy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dot Product</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAXPY</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Normalized Performance
<table>
<thead>
<tr>
<th></th>
<th>Tofu</th>
<th>Tofu2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System</strong></td>
<td>K computer and FX10</td>
<td>FX100</td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>SPARC64 VIIIIfx/IXfx</td>
<td>SPARC64 XIIfx</td>
</tr>
<tr>
<td><strong>Integration</strong></td>
<td>No, dedicated LSI ICC is required</td>
<td>Yes, integrated into the CPU chip</td>
</tr>
<tr>
<td><strong>Topology</strong></td>
<td>6D mesh/torus topology</td>
<td>←</td>
</tr>
<tr>
<td><strong>Link bandwidth</strong></td>
<td>5 GB/s (6.25 Gbps x 8 lanes x 10 dirs)</td>
<td>12.5 GB/s (25 Gbps x 4 lanes x 10 dirs)</td>
</tr>
<tr>
<td><strong>Node bandwidth</strong></td>
<td>20 GB/s x in/out</td>
<td>50 GB/s x in/out</td>
</tr>
<tr>
<td><strong>Other features</strong></td>
<td>-</td>
<td>Cache injection, atomic operation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optical connection (2/3 of links are optical)</td>
</tr>
</tbody>
</table>
Communication performance

Throughput improves 2.4x higher than FX10

Good simultaneous multiple direction transfer

IMB Pingpong throughput

Two-direction simultaneous comm.

Throughput (GB/s)

Message length (KiB)

Throughput (GB/s)

InfiniBand FDR  FX10  FX100

InfiniBand FDR  FX10  FX100
Assistant core

- Two assistant cores available
  - For OS jitter reduction and overlap communication
Offloading of daemons, IO processing, MPI asynchronous communication to the assistant core reduces OS jitter

### Dispersion of calc. time caused by OS jitter

- fwq
- w 18  # Bits in work. # of loop iter. is $2^w$
- n 30000  # # of samples to take
- t 32  # # of threads to be executed

### Estimated slowdown caused by OS jitter (comm. interval=1ms)

- Slowdown rate vs. # of nodes
- x86 cluster, FX10, FX100

(OS optimization undergone)
Overlapping execution of non-blocking comm.

Assistant core is used in MPI library
Boundary data transfer of stencil code

![Diagram showing overlapping execution and data transfer]

Normalized execution time

<table>
<thead>
<tr>
<th>Msg. length &amp; Assistant core usage</th>
<th>Ratio of execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not used 64KiB</td>
<td>Comm.: 0.6, Compute: 1.0</td>
</tr>
<tr>
<td>Used 64KiB</td>
<td>Comm.: 0.9, Compute: 1.0</td>
</tr>
<tr>
<td>Not used 256KiB</td>
<td>Comm.: 0.8, Compute: 1.0</td>
</tr>
<tr>
<td>Used 256KiB</td>
<td>Comm.: 0.9, Compute: 1.0</td>
</tr>
<tr>
<td>Not used 1MiB</td>
<td>Comm.: 0.7, Compute: 1.0</td>
</tr>
<tr>
<td>Used 1MiB</td>
<td>Comm.: 0.9, Compute: 1.0</td>
</tr>
</tbody>
</table>

Assistant core is used in MPI library
Boundary data transfer of stencil code
Application evaluation

- NAS Parallel Benchmarks FT Class C by OpenMP parallel
  - Time integration of a 3D partial differential equation using FFT (512^3)
- Quantum chromodynamics, CCS QCD Miniapp †
  - A linear equation solver with a large sparse coefficient matrix appearing in a lattice QCD problem (32x32x32x32)

† https://github.com/fiber-miniapp/ccs-qcd
NAS Parallel Benchmarks FT Class C (OpenMP)

Time integration of a 3D partial differential equation using FFT ($512^3$)

- 3.3x faster on FX100 with 32 threads
- Node performance is enhanced by higher cache/memory throughput, as well as increased CPU cores and SIMD width

Node performance (Gflops/node)

<table>
<thead>
<tr>
<th></th>
<th>FX10 (16 threads)</th>
<th>FX100 (32 threads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node performance</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

Breakdown of execution time

- 2-4 inst. committed
- 1 inst. committed
- wait (others)
- wait (instruction)
- wait (calculation)
- wait (cache)
- wait (memory)
A linear equation solver with a large sparse coefficient matrix appearing in a lattice QCD problem (32x32x32x32)

- 1.6x faster with 16 threads, 3.0x faster with 32 threads
- Enhanced memory bandwidth boosts the performance and Sector cache mechanism promotes data reuse on L2$

Node performance (Gflops/node)

<table>
<thead>
<tr>
<th>Sector cache enabled</th>
<th>Sector cache disabled</th>
<th>Sector cache enabled</th>
<th>Sector cache disabled</th>
<th>Sector cache enabled</th>
<th>Sector cache disabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>FX10 (1 proc. 16 threads)</td>
<td>FX100 (1 proc. X 16 threads)</td>
<td>FX100 (1 proc. X 32 threads)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† https://github.com/fiber-miniapp/ccs-qcd
**Summary, FX100**

**FX100 provides steady progress for users, natural extent of perf. profile**
- Single CPU/node architecture for multicore
- Good Byte/flop and scalability

**Leads apps toward highly scalable and introduces new technologies**
- Original CPU and interconnect
- Support for tens of millions of cores (VISIMPACT, Collective comm. HW)

**PRIMEHPC Series**

**K computer**
- VISIMPACT
- SIMD extension HPC-ACE
- Direct network Tofu
- CY2010~ 128GF, 8-core/CPU

**FX10**
- VISIMPACT
- HPC-ACE
- Direct network Tofu
- CY2012~ 236.5GF, 16-core/CPU

**FX100**
- VISIMPACT
- HPC-ACE2
- Tofu interconnect 2
- HMC & Optical connections
- CY2015~ 1TF~, 32-core/CPU
shaping tomorrow with you