Fujitsu HPC Roadmap Beyond Petascale Computing

Toshiyuki Shimizu
Fujitsu Limited
Mission and HPC product portfolio

K computer*, Fujitsu PRIMEHPC, and the future
  - K computer and PRIMEHPC FX10
  - Post-FX10, successor to PRIMEHPC FX10

Summary

* K computer jointly developed by RIKEN and FUJITSU
Fujitsu’s mission and portfolio

HPC solutions for every aspect

Petascale supercomputer

Develop all key components: SPARC chips, Tofu interconnect, and software stacks

K computer
Developed with RIKEN

PRIMEHPC FX10

x86 clusters by PRIMERGY

Latest standard technologies: x86, Xeon Phi, and GPGPU

PRIMERGY CX400

BX900/BX400 RX200/RX900
Summary of K computer and Fujitsu PRIMEHPC

- Single CPU/node architecture for multicore
- Key technologies for massively parallel series
  - Tens of millions of cores (VISIMPACT*, Collective comm. HW,)
  - Original CPU and interconnect

* VISIMPACT: Virtual Single Processor by Integrated Multi-core Parallel Architecture

Copyright 2013 FUJITSU LIMITED
CPU chip of K computer: SPARC64 VIIIfx

- Eight core out-of-order super scalar CPU
- HPC-ACE instruction set extension
- VISIMPACT hybrid execution model support
- Low power consumption design
- Highly reliable design inherited from mainframe

<table>
<thead>
<tr>
<th>Technology</th>
<th>45 nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>128 GFLOPS</td>
</tr>
<tr>
<td>Memory bandwidth</td>
<td>64 GB/s</td>
</tr>
<tr>
<td>Power consumption</td>
<td>58 W</td>
</tr>
<tr>
<td># of transistors</td>
<td>760 M</td>
</tr>
</tbody>
</table>

Error detection by hardware and automatic recovery
Error detection by hardware
No affect on system operation
6D mesh/torus direct network
K: \((24 \times 18 \times 17) \times (2 \times 3 \times 2)\)
Low ave. hops and high bisectional BW
Virtual 3D torus topology for apps.
Hardware collective comm. support
Congestion control by inserting GAPs
System software for K computer and FX10

- Tuned Linux OS for HPC applications
  - Supports large pages and OS jitter minimization
- Combination of self-developed software and customized OSS (~14 Msteps)
  - 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 IO servers)
- Higher IO perf. (1.4 TB/s)

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

OS (Linux + HPC specific enhancement)
K computer application performance

- Stable operation of over 80k node system open for public use
- Emerging applications developed and run
- Leading highly scalable application performance provided by Dr. Minami of RIKEN

<table>
<thead>
<tr>
<th>Application</th>
<th>Brief explanation</th>
<th># of nodes</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINPACK</td>
<td>Solves dense linear system arithmetic</td>
<td>88,128</td>
<td>93%</td>
</tr>
<tr>
<td>NICAM</td>
<td>Non-hydrostatic icosahedral atmospheric model for global-cloud resolving simulation</td>
<td>81,920</td>
<td>8%</td>
</tr>
<tr>
<td>Seism3D</td>
<td>Simulation of seismic-wave propagation and strong ground motions</td>
<td>82,944</td>
<td>18%</td>
</tr>
<tr>
<td>PHASE</td>
<td>First-principle simulation based on the density functional theory using a plane wave expansion method</td>
<td>82,944</td>
<td>20%</td>
</tr>
<tr>
<td>RSDFT</td>
<td>First-principle simulation based on the density functional theory using a finite difference method</td>
<td>82,944</td>
<td>52%</td>
</tr>
<tr>
<td>FrontFlow/blue</td>
<td>Unsteady flow analysis based on large eddy simulation</td>
<td>80,000</td>
<td>3%</td>
</tr>
<tr>
<td>Lattice QCD</td>
<td>Simulation of elementary particle and nuclear physics based on the quantum chromodynamics theory</td>
<td>82,944</td>
<td>16%</td>
</tr>
<tr>
<td>ZZ-EFSI</td>
<td>Combination of CFD and structure analysis program simulating blood flow in vessel</td>
<td>82,944</td>
<td>46%</td>
</tr>
</tbody>
</table>
## Road to exascale computing

<table>
<thead>
<tr>
<th>Year</th>
<th>PRIMEHPC FX10</th>
<th>Post-FX10 system</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>PRIMEHPC FX10</td>
<td>Post-FX10 system</td>
</tr>
<tr>
<td>2014</td>
<td>PRIMEHPC FX10</td>
<td>Post-FX10 system</td>
</tr>
<tr>
<td>2015</td>
<td>Post-FX10</td>
<td>Post-FX10 system</td>
</tr>
<tr>
<td>2016</td>
<td>Post-FX10</td>
<td>Post-FX10 system</td>
</tr>
<tr>
<td>2017</td>
<td>Post-FX10</td>
<td>Post-FX10 system</td>
</tr>
<tr>
<td>2018</td>
<td>Post-FX10</td>
<td>Post-FX10 system</td>
</tr>
<tr>
<td>2019</td>
<td>Post-FX10</td>
<td>Post-FX10 system</td>
</tr>
</tbody>
</table>

- **PRIMEHPC FX10**
  - 1.85 × CPU performance
  - Easier installation

- **Post-FX10 system**
  - Enhanced CPU & network performance
  - High-density packaging & low power consumption

## National projects

**Development**

- HPCI strategic applications program
  - App. review
  - FS projects
  - Exa-system development project

**Operation of K computer**

- Operation of K computer

---

Copyright 2013 FUJITSU LIMITED
Goals for Post-FX10 and exascale

- High performance per power and high density packaging

**Key technologies critical for exascale system**

- Latest Si technologies
- Power efficient microarchitecture
- Latest memory technology
- Dense packaging and SoC
- Latest optical technology
What is power efficient micro arch.?

Power efficiency = \[
\frac{\text{App. performance}}{\text{Power}}
\]

Example of design exploration of power efficient micro arch.

- SIMD
- Rich OOO
- Branch prediction
- Larger cache

<table>
<thead>
<tr>
<th>NoSIMD</th>
<th>SIMDx2</th>
<th>SIMDx4</th>
<th>SIMDx8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>Max power</td>
<td>Ideal FP perf.</td>
<td>IFP/Max power</td>
</tr>
<tr>
<td>Power</td>
<td>Perf</td>
<td>Perf/Power</td>
<td></td>
</tr>
</tbody>
</table>
**Post-FX10 power efficient micro arch.**

**Power efficiency =** \[
\frac{\text{App. performance}}{\text{Power}} \]

<table>
<thead>
<tr>
<th>Core config.</th>
<th>K</th>
<th>FX10</th>
<th>Post-FX10</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak FP performance</td>
<td>128 GF</td>
<td>236.5 GF</td>
<td>1-TF class</td>
<td>Maintain similar architecture for compatibility for applications</td>
</tr>
<tr>
<td>Execution unit</td>
<td>FMA × 2</td>
<td>FMA × 2</td>
<td>FMA × 2</td>
<td>Wider SIMD for better performance with small additional hardware</td>
</tr>
<tr>
<td>SIMD</td>
<td>2</td>
<td>2</td>
<td>&gt; 2</td>
<td>Double of DP</td>
</tr>
<tr>
<td>Dual SP mode</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td>Accelerates SP rich apps</td>
</tr>
<tr>
<td>Integer SIMD ope.</td>
<td>NA</td>
<td>NA</td>
<td>Support</td>
<td>Boost SIMD ratio, list vector</td>
</tr>
<tr>
<td>Single thread performance enhancement</td>
<td>-</td>
<td>-</td>
<td></td>
<td>Application performance often limited by single thread performance and no FP calc.</td>
</tr>
</tbody>
</table>

- **SIMD**
- **Rich OOO**
- **Branch prediction**
- **Larger cache**
Wider SIMD
- Single instruction performs more operations simultaneously
- Enhanced to perform stride load/store, indirect (list) load/store, permutation, concatenate

Single precision and integer operations
- Dual single precision (double performance of DP), 8-byte integer arithmetic

**SIMD**
- Double precision
- Multiple simultaneous calculations

**Stride load**
- Specified stride
- Memory
- Reg S
- Reg D

**Indirect load**
- Reg S
- Memory
- Reg D

**Permutation**
- Arbitrary shuffle
- Reg S
- Reg D
Integer SIMD and indirect vector load

- Loops not accelerated by basic SIMD instructions only if they contain low true ratio if clauses

- List vector load SIMD inst. enables SIMDizing them with concatenate instructions

- Ex.: srtm_reftra of climate program, IFS
  - **If clauses** used to detect clouds
  - Indirect access accelerates 1.7 times

```
do i=1,N
  if (A(i)>0) then
    L(j)=I
    j=j+1
  endif
enddo
```

Transform to indirect vector load

```
do i=1,N
  if (A(i)>0) then
    L(j)=I
    j=j+1
  endif
enddo
```

Generate masks and compress flags

```
do i=1,j
  B(L(i))=C(L(i))+...
enddo
```

Body

```
do i=1,j
  B(L(i))=C(L(i))+...
enddo
```

Ratio of memory access ⇒ Low

if clauses
Latest memory technology, HMC

- Micron’s HMC as main memory
  - Higher density per BW
  - Higher capacity per package
  - Higher bandwidth per package
  - Lower power consumption per BW

HMC fulfills required BW for high performance multi-core CPUs

<table>
<thead>
<tr>
<th></th>
<th>Capacity</th>
<th>Bandwidth</th>
<th>Density</th>
<th>Other concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMC × 8</td>
<td>Good</td>
<td>Very good</td>
<td>Excellent</td>
<td></td>
</tr>
<tr>
<td>HBM × 8</td>
<td>Fair</td>
<td>Very good</td>
<td>Very good</td>
<td>Cost/SCM of 2.5 D</td>
</tr>
<tr>
<td>DDR4-DIMM × 8</td>
<td>Very good</td>
<td>Fair</td>
<td>Fair</td>
<td>No successor</td>
</tr>
<tr>
<td>GDDR5 × 8</td>
<td>Poor</td>
<td>Good</td>
<td>Poor</td>
<td></td>
</tr>
</tbody>
</table>
Configuration of Post-FX10 system

**Post-SPARC64 IXfx**
- Wide SIMD 1-TF class multicore CPU
- HPC-ACE2 support
- Tofu2 integrated

**Chassis (12 CPUs)**
- 1 CPU/1 node
- 12 nodes/2U Chassis
- Water cooled

**CPU Memory Board**
- Three CPUs
- Eight HMCs (hybrid memory cube)
- Optical modules

**Tofu2 Interconnect**
- Performance improvement
- Optical technology

**Cabinet**
- Over 200 nodes/cabinet
- High-density
- 100% water cooled with EXCU (option)
Summary

- Fujitsu continues to develop top end HPC machines
  - Run post petascale applications as well as K computer applications
  - Develop both CPU and interconnect for performance and availability

- Post-FX10 scheduled to be available around 2015
  - More calculation power with better power efficiency
  - Developing original HPC enhanced SPARC64 CPU chip
  - Tofu2 interconnect integrated and optical connection supported

- Considering exascale system design
  - Feasibility study project sponsored by MEXT and lead by the University of Tokyo
Fujitsu booth No. 2718

- Come and see our new CPU memory boards at the Fujitsu and Micron booths

CPU memory board of Post-FX10
HMCs used for main memory, delivering appropriate memory bandwidth for high performance CPU chips
shaping tomorrow with you