The System Design of the Next Generation Supercomputer: Post-K Supercomputer

Shinji Sumimoto, Ph.D.
Next Generation Technical Computing Unit
FUJITSU LIMITED

Oct. 25th, 2016
Outline

- Fujitsu in Technical Computing
  - PRIMEHPC FX100 overview
- The System Design of the Next Generation Supercomputer: Post-K
  - Background: FLAGSHIP2020 Project Overview
  - Requirements
  - System Design
  - Effectiveness for Meteorology
Fujitsu in Technical Computing
Past, PRIMEHPC FX100, and “Roadmap for Exascale”

PRIMEHPC FX10 in operation

Many applications are currently running and being developed for science and various industries.

PRIMEHPC FX100 in operation

The CPU and interconnect inherit the K computer architectural concept, featuring state-of-the-art technologies.

System software TCS supports the FX100 with newly developed technologies.

Towards Exascale

RIKEN and FUJITSU are working together for the Post-K computer.

Technical Computing Suite (TCS)

- Handles millions of parallel jobs
- FEFS: super scalable file system
- MPI: Ultra scalable collective communication libraries
- OS: Lower OS jitter w/ assistant core

Japan’s National Projects

Development

Operation of K computer

Post-K

HPCI strategic apps program

App. review

FS projects

Post-K computer development

Copyright 2016 FUJITSU LIMITED
Features of Fujitsu high-end supercomputer and Post-K

FUJITSU designed high performance CPU

Dedicated high performance interconnect Tofu

Application compatibility throughout generations

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
SMaC
Tofu interconnect 2
HMC & Optical connections
CY2015~
1TF~, 32-core/CPU

Post-K
PRIMEHPC FX100, design concept and approach

Provide steady progress for users

- Continue to keep performance portability among K computer, FX10 and FX100
- Facilitate the evolution of applications

Challenge to state-of-art technologies for future generation

- 20nm CMOS technology
- Hybrid Memory Cube (HMC)
- 25Gbps optical connection
**PRIMEHPC FX100 Overview**

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

**Chassis**
- 1 CPU/1 node
- 12 nodes/2U Chassis
- Water cooled
The System Design of Post-K

- Background: FLAGSHIP2020 Project Overview
- Requirements of Post-K
- Design for Application Performance Portability
- Effectiveness for meteorology
An Overview of Flagship 2020 project

- Developing the next Japanese flagship computer, so-called “post K”
- Developing a wide range of application codes, to run on the “post K”, to solve major social and science issues

The Japanese government selected 9 social & scientific priority issues and their R&D organizations.

Society with health and longevity
- Innovative Drug Discovery
  - RIKEN Quant. Biology Center
- Personalized and Preventive Medicine
  - Inst. Medical Science, U. Tokyo

Disaster prevention and global climate
- Hazard and Disaster induced by Earthquake and Tsunami
  - Earthquake Res. Inst., U. Tokyo
- Environmental Predictions with Observational Big Data
  - Center for Earth Info. JAMSTEC

Energy issues
- Innovative Clean Energy Systems
  - Grad. Sch. Engineering, U. Tokyo
- High-Efficiency Energy Creation, Conversion/Storage and Use
  - Inst. Molecular Science, NIMS

Industrial competitiveness
- Innovative Design and Production Processes for the Manufacturing Industry in the Near Future
  - Inst. of Industrial Science, U. Tokyo
- New Functional Devices and High-Performance
  - Inst. For Solid State Phys., U. Tokyo

Basic science
- Fundamental Laws and Evolution of the Universe
  - Cen. for Comp. Science, U. Tsukuba
# Target Applications’ Characteristics

<table>
<thead>
<tr>
<th>Program</th>
<th>Brief description</th>
<th>Co-design</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GENESIS</td>
<td>MD for proteins</td>
<td>Collective comm. (all-to-all), Floating point perf (FPP)</td>
</tr>
<tr>
<td>2 Genomon</td>
<td>Genome processing (Genome alignment)</td>
<td>File I/O, Integer Perf.</td>
</tr>
<tr>
<td>3 GAMERA</td>
<td>Earthquake simulator (FEM in unstructured &amp; structured grid)</td>
<td>Comm., Memory bandwidth</td>
</tr>
<tr>
<td>4 NICAM+LETK</td>
<td>Weather prediction system using Big data (structured grid stencil &amp; ensemble Kalman filter)</td>
<td>Comm., Memory bandwidth, File I/O, SIMD</td>
</tr>
<tr>
<td>5 NTChem</td>
<td>molecular electronic (structure calculation)</td>
<td>Collective comm. (all-to-all, allreduce), FPP, SIMD</td>
</tr>
<tr>
<td>6 FFB</td>
<td>Large Eddy Simulation (unstructured grid)</td>
<td>Comm., Memory bandwidth</td>
</tr>
<tr>
<td>7 RSDFT</td>
<td>an ab-initio program (density functional theory)</td>
<td>Collective comm. (bcast), FPP</td>
</tr>
<tr>
<td>8 Adventure</td>
<td>Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)</td>
<td>Comm., Memory bandwidth, SIMD</td>
</tr>
<tr>
<td>9 CCS-QCD</td>
<td>Lattice QCD simulation (structured grid Monte Carlo)</td>
<td>Comm., Memory bandwidth, Collective comm. (allreduce)</td>
</tr>
</tbody>
</table>

*Copyright 2016 FUJITSU LIMITED*
An Overview of post K

**Hardware**
- Manycore architecture
- 6D mesh/torus Interconnect
- 3-level hierarchical storage system
  - Silicon Disk
  - Magnetic Disk
  - Storage for archive

**System Software**
- Multi-Kernel: Linux with Light-weight Kernel
- File I/O middleware for 3-level hierarchical storage system and application
- Application-oriented file I/O middleware
- MPI+OpenMP programming environment
- Highly productive programming language and libraries
Goal and Requirements of Post-K

**Goals:**
- World’s top class application performance with solving social and scientific priority issues
- Reliable system operation with limited power consumption

**Requirements:**
- World’s top class application performance with limited power consumption
- Keeping system reliability as much as possible as well as K computer.
- Easy migration of existing application from existing systems including K computer to expand system use

Especially, performance portability of existing application on K computer is important
Post-K Performance Portability Design

- Application Execution Model Backward Compatibility
  - Processor Architecture
  - System Architecture

- System Balance among Processor Core, Memory, Interconnect and Storage I/O Performance
  - Trying to keep system balance with limited power consumption and cost.

- Binary Compatibility Mechanism for Future Generations
  - Preferring to keep portability without re-compiling applications

- Execution Environment: Compiler, Runtime System, MPI, Batch Script etc.
  - Backward Compatibility of System Operation and Application Execution Environment
APPLICATION EXECUTION MODEL BACKWARD COMPATIBILITY
System Architecture for Performance Portability

- A scalable, many-core micro architecture concept: “SMaC,”
- Single Process Multi-Threaded Model in a Socket: CMG
- Scalable interconnect: “Tofu”
- High Performance Lustre Based Cluster File System: FEFS

Scalable System Software Architecture: Resource-saving, Flexible, Reliable

Single process multi-threading

Single chip CPU compute node single Linux on coherent memory

12-node Tofu unit

6D mesh/torus Tofu interconnect

SMaC Scalable Architecture

Tofu Scalable Architecture
Core Memory Group (CMG) Structure

- Cores in the group share the same L2 cache
- Dedicated memory and memory controller for the CMG provide high BW and low latency data access
- Loosely coupled CMGs using tagged coherent protocol share data with small silicon overhead
- Hierarchical configuration promises good core/performance scalability
SYSTEM PARAMETER BALANCE AMONG PROCESSOR CORE, MEMORY, INTERCONNECT AND STORAGE I/O
Post-K’s System Balance

FUJITSU original CPUs steadily increase their fundamental performance
Support a programming model, hybrid parallel execution
Uncompromised **system balance** for the best use of applications

From FX100, an **assistant core** and **CMG** is introduced

<table>
<thead>
<tr>
<th>Feature</th>
<th>Post-K</th>
<th>FX100</th>
<th>FX10</th>
<th>K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double Flops / CPU</td>
<td>TBD</td>
<td>1 TF</td>
<td>235 GF</td>
<td>128 GF</td>
</tr>
<tr>
<td>Single Flops / CPU</td>
<td></td>
<td>2 TF</td>
<td>235 GF</td>
<td>128 GF</td>
</tr>
<tr>
<td>SIMD width</td>
<td></td>
<td>256 bit</td>
<td>128 bit</td>
<td>128 bit</td>
</tr>
<tr>
<td># of CMG (# of cores / CMG)</td>
<td></td>
<td>2(17)</td>
<td>1(16)</td>
<td>1(8)</td>
</tr>
<tr>
<td># of cores / CPU</td>
<td></td>
<td>32 + 2xAC</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>Memory / CPU</td>
<td></td>
<td>32 GB</td>
<td>32 GB</td>
<td>16 GB</td>
</tr>
<tr>
<td>Memory BW</td>
<td></td>
<td>480 GB/s</td>
<td>83.5 GB/s</td>
<td>64 GB/s</td>
</tr>
</tbody>
</table>
Tofu Interconnect of K, PRIMEHPC and Post-K

Compatible highly scalable FUJITSU original interconnect

- Optimal implementation keeping high application scalability
- Flexible and efficient communication patterns for application performance
- Non blocking CPU off loadable DMA engines for calculation and communication overlapping

<table>
<thead>
<tr>
<th></th>
<th>Post-K</th>
<th>FX100</th>
<th>FX10</th>
<th>K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interconnect</td>
<td>Tofu 6D mesh/torus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interconnect BW</td>
<td>TBD</td>
<td>12.5 GB/s</td>
<td>5 GB/s</td>
<td>5 GB/s</td>
</tr>
<tr>
<td># of DMA engines</td>
<td>TBD</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Node injection BW</td>
<td>TBD</td>
<td>50 GB/s</td>
<td>20 GB/s</td>
<td>20 GB/s</td>
</tr>
<tr>
<td>Collective operations</td>
<td>TBD</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Next Generation File System and Storage Design

K computer File System Design
- How should we realize High Speed and Redundancy together?
- How do we avoid I/O conflicts between Jobs?
- These are not realized in single file system.
  - Therefore, we have introduced Integrated Layered File System.

K computer achieved 1 TB/s sustained file I/O performance

Next Generation File System/Storage Design
- Another trade off targets: Power, Capacity, Footprint
  - Difficult to realize enough capacity and performance file system in limited power consumption and footprint.

Third Storage layer for Capacity is needed: Three Layered File System
- Local File System for Performance
- Global File System for Shared Use
- Archive File System for Capacity
BINARY COMPATIBILITY MECHANISM FOR FUTURE GENERATIONS
FUJITSU HPC CPU Transition to the ARM V8 + SVE

Post-K fully utilizes FUJITSU proven supercomputer microarchitecture

FUJITSU, as a lead partner of ARM SVE development, is contributing to complement ARM SVE (Scalable Vector Extension), for application performance efficiency

ARM V8+SVE brings out the real strength of FUJITSU’s microarchitecture

<table>
<thead>
<tr>
<th>ISA</th>
<th>Functions for Perf.</th>
<th>Post-K</th>
<th>FX100</th>
<th>FX10</th>
<th>K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVE w/ FJ contribution</td>
<td>FMA</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Math. acc. prim.*</td>
<td>✔ Enhanced</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Fujitsu Extension</td>
<td>Inter core barrier</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Sector cache</td>
<td>✔ Enhanced</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Prefetch</td>
<td>✔ Enhanced</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

*Mathematical acceleration primitives include trigonometric functions, sine & cosines, and exponential...
Overview of SVE (Scalable Vector Extension)

- **HPC Extension Instruction Set of AArch64**
  - SVE is not an extension of NEON, independent of NEON for HPC processing
  - SVE instruction and NEON instruction are able to execute independently
  - SVE includes various amount instruction set to support higher SIMD execution

- **Post-K processor ISA will be ARM with SVE**

<table>
<thead>
<tr>
<th>SVE</th>
<th>ARMv8-A</th>
<th>v8.1-A</th>
<th>v8.2-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEON</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# SIMD ISA Extension Comparison

<table>
<thead>
<tr>
<th></th>
<th>SVE</th>
<th>HPC-ACE2</th>
<th>AVX-512</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base ISA</strong></td>
<td>ARMv8-A</td>
<td>SPARC V9</td>
<td>Intel 64</td>
</tr>
<tr>
<td><strong>SIMD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bit width</td>
<td>128~2048</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>SP Elements</td>
<td>4~64</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>DP Elements</td>
<td>2~32</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td><strong>GP Registers (#)</strong></td>
<td>31+SP (Same as v8-A)</td>
<td>32+32</td>
<td>16</td>
</tr>
<tr>
<td><strong>Vector Registers (#)</strong></td>
<td>32</td>
<td>128</td>
<td>32</td>
</tr>
<tr>
<td><strong>Predicate Registers (#)</strong></td>
<td>16</td>
<td>(Included in Vector Regs.)</td>
<td>8</td>
</tr>
</tbody>
</table>
Scalable Vector Length for ABI Compatibility

- ISA does not fix Vector Length
  - SVE supports VL from 128 to 2048 bit with multiples of 128 bit
  - VL is set by processor before executing a binary dynamically

- Single execution binary can be executed on processors with multiple VLs
  - Vector-Length Agnostic (VLA) programming enables ABI Compatibility

**Execution Binary Portability**

- Execution Binary does not depend on processor’s VL
  - Increasing dynamic instruction steps to double
  - Reducing dynamic instruction steps to half

512bit SIMD  

256bit SIMD
EFFECTIVENESS FOR METEOROLOGY ON POST-K
Effectiveness for Meteorology of Post-K

Achievements with K computer

- NICAM performance on K computer is good scalability up to 81920node x 8 threads with 0.9 PFLOPS
  - “Recent performance of NICAM on the K-computer and activities towards post-petascale computing”, Hisashi Yashiro (Riken/AICS), Workshop on Scalability (ECMWF, 14-15 April, 2014)

Effectiveness for meteorology on Post-K

- Wider SIMD with good system balance using meteorology application: IFS
Weak scaling test

- Same problem size per node, same steps
- Full configuration / full components
- Realistic boundary / initial data set
- Good scalability up to 81920node x 8threads with 0.9PFLOPS

![Bar chart showing performance efficiency and elapsed time for different numbers of cores]
Performance Portability from K to FX100 and Post-K

- SIMD width and memory bandwidth are enhanced from K to FX100
  - 4-way SIMD is supported on FX100 and Hybrid Memory Cube (HMC) provides higher memory bandwidth

- Speedup of IFS on FX100 is realized by wider SIMD and good system balance
  - 2-way SIMD can benefit from high memory bandwidth
  - 4-way SIMD accelerates calculations and drives memory bandwidth more

- Trying to keep system balance will be expected to provide higher performance on the next-generation machine: Post-K with more wider SIMD width

<table>
<thead>
<tr>
<th></th>
<th>K computer</th>
<th>FX100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flops / CPU</td>
<td>128 Gflops</td>
<td>1 Tflops</td>
</tr>
<tr>
<td>SIMD width</td>
<td>128 bit</td>
<td>256 bit</td>
</tr>
<tr>
<td>Memory BW</td>
<td>64 GB/s</td>
<td>480 GB/s</td>
</tr>
<tr>
<td>Byte per flop</td>
<td>0.4 ~ 0.5</td>
<td></td>
</tr>
</tbody>
</table>

![Speedup of IFS (TL159, 32 cores)](image)
Summary

Fujitsu in Technical Computing

PRIMEHPC FX100 Overview

The System Design of the Next Generation Supercomputer: Post-K

- Performance Portability: Trying to keep system balance with limited power consumption and cost will be expected to provide higher performance on the next-generation machine: Post-K with more wider SIMD width.

- Application Binary Compatibility: Scalable Vector Length will help to keep binary compatibility for the future systems without re-compilation of programs.
shaping tomorrow with you