Post-K Supercomputer with Fujitsu's Original CPU, A64FX, Powered by Arm ISA

Toshiyuki Shimizu

Nov. 15th, 2018

Post-K is under development, information in these slides is subject to change without notice.
Agenda

- Background
  - K computer and Post-K
  - Post-K goals and approaches
- Fujitsu new Arm CPU A64FX
- Post-K system
  - System software
  - Configuration
- Performance discussion
- Summary and development status
Agenda

- Background
  - K computer and Post-K
  - Post-K goals and approaches
- Fujitsu new Arm CPU A64FX
- Post-K system
  - System software
  - Configuration
- Performance discussion
- Summary and development status
K computer, PRIMEHPC, and Post-K

- **K computer, PRIMEHPC FX**
  - Many applications are running and being developed for science and various industries
  - System software TCS supports hardware with newly developed technologies

- **Post-K**
  - RIKEN and Fujitsu are working together for Post-K
  - OS is running and design verification is proceeding as scheduled

---

**K computer, PRIMEHPC FX**

- 1.8x CPU perf.
- Easier installation
- 4x(DP) / 8x(SP) CPU & Tofu2
- High-density pkg & lower energy

**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**

- Post-K

**Development**

- Operation of K computer
- HPCI strategic apps program
- App. review
- FS projects
- Post-K computer development

---

Copyright 2018 FUJITSU LIMITED
Post-K goals and approaches

Post-K goals

- High application performance and good power efficiency
- Good usability and better accessibility for users
- Keeping application compatibility while advancing from predecessors

Our approaches

- Develops high performance and scalable, custom CPU cores
  - Performance: Wider SIMD & mathematical acc. primitives, high memory BW
  - Scalability: Scalable interconnect “Tofu”
  - Power efficiency: The best device tech, power control functions, optimal resources

- Adopts Arm ISA, binary compatibility
- Maintains performance balance and supports advanced features
Agenda

- Background
  - K computer and Post-K
  - Post-K goals and approaches
- Fujitsu new Arm CPU A64FX
- Post-K system
  - System software
  - Configuration
- Performance discussion
- Summary and development status
Fujitsu new Arm CPU: A64FX

- A64FX approach
- Specs and technologies
- Tofu interconnect D
- Power management
- A64FX summary
A64FX: Approach

- High performance Arm CPU
  - Develops original CPU core supporting Arm SVE (Scalable Vector Extension)
- Targeting high performance servers
  - Floating point calculations
  - High memory bandwidth
  - Low power consumption
  - Scalable performance and configuration
- Extension for emerging applications
  - Half precision (FP16) & INT16/8 partial dot product support

High BW/Calc. is the key for Real apps.
A64FX: Specs & technologies

CPU generations and key parameters

<table>
<thead>
<tr>
<th>CPU</th>
<th>SPARC64 VIIIfx</th>
<th>SPARC64 IXfx</th>
<th>SPARC64 XIfx</th>
<th>A64FX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st} system w/ CPU</td>
<td>K</td>
<td>FX10</td>
<td>FX100</td>
<td>Post-K</td>
</tr>
<tr>
<td>Si tech. (nm)</td>
<td>45</td>
<td>40</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>Core perf. (GFLOPS)</td>
<td>16</td>
<td>14.8</td>
<td>34</td>
<td>57~</td>
</tr>
<tr>
<td># of cores</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>48</td>
</tr>
<tr>
<td>Chip perf. (TFLOPS)</td>
<td>0.13</td>
<td>0.24</td>
<td>1.1</td>
<td>2.7~</td>
</tr>
<tr>
<td>Memory BW (GB/s)</td>
<td>64</td>
<td>85</td>
<td>480</td>
<td>1024</td>
</tr>
<tr>
<td>B/F (Bytes/FLOP)</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Combination of technologies

- SVE increases core performance
- CMG is scalable architecture to increase # of cores
- HBM enables high bandwidth
A64FX technologies: Core performance

- High calc. throughput of Fujitsu original CPU core w/ SVE
- 512-bit wide SIMD x 2 pipelines and new integer functions

**Core peak performance**

<table>
<thead>
<tr>
<th>Element size</th>
<th>Multiply and add</th>
<th>INT8 partial dot product</th>
<th>(GOPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64-bit</td>
<td>&lt;115</td>
<td>&gt;230</td>
<td>&gt;57</td>
</tr>
<tr>
<td>32-bit</td>
<td>&gt;115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16-bit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-bit</td>
<td></td>
<td></td>
<td>&gt;460</td>
</tr>
</tbody>
</table>

**INT8 partial dot product**

\[ C = \sum (A_i \times B_i) + C \]

- 8bit
- 8bit
- 8bit
- 8bit

32bit
A64FX technologies: Scalable architecture

- Core Memory Group (CMG)
  - 12 compute cores and an assistant core for OS daemon, I/O, etc.
  - Shared L2 cache
  - Dedicated memory controller

- Four CMGs maintain cache coherence with on-chip directory
  - Threads binding within a CMG allows linear speed of up to 48 compute cores
A64FX: L1D cache uncompromised BW

- 128B/cycle sustained BW even for unaligned SIMD load

- “Combined Gather” doubles gather (indirect) load’s data throughput, when target elements are within a “128-byte aligned block” for a pair of two regs, even & odd

Maximizes BW to 32 bytes/cyc.
A64FX: Tofu interconnect D

- Integrated w/ rich resources
  - Increased TNIs achieves higher injection BW & flexible comm. patterns
  - Increased barrier resources allow flexible collective comm. algorithms
- Memory bypassing achieves low latency
  - Direct descriptor & cache injection

<table>
<thead>
<tr>
<th></th>
<th>TofuD spec</th>
<th>Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rate</td>
<td>28.05 Gbps</td>
<td></td>
</tr>
<tr>
<td>Link bandwidth</td>
<td>6.8 GB/s</td>
<td></td>
</tr>
<tr>
<td>Injection bandwidth</td>
<td>40.8 GB/s</td>
<td></td>
</tr>
<tr>
<td>Put throughput</td>
<td>6.35 GB/s</td>
<td></td>
</tr>
<tr>
<td>PingPong latency</td>
<td>0.49~0.54 µs</td>
<td></td>
</tr>
</tbody>
</table>
A64FX: Power monitor and analyzer

- **Energy monitor (per chip)**
  - Node power via Power API*1 (~msec)
  - Averaged power of a node, CMG (cores, an L2 cache, a memory) etc.

- **Energy analyzer (per core)**
  - Power profiler via PAPI*2 (~nsec)
  - Fine grained power analysis of a core, an L2 cache and a memory

*1: Sandia National Laboratory

*2: Performance Application Programming Interface
A64FX: Power Knobs to reduce power consumption

- "Power knob" limits units’ activity via user APIs
- Performance/W would be optimized by utilizing Power knobs, Energy monitor & analyzer

- Frequency reduction
- EXA pipeline only
- FLA pipeline only
- HBM2 B/W adjust (units of 10%)
A64FX: Summary

- Arm SVE, high performance and efficiency
- DP performance >2.7 TFLOPS, >90%@DGEMM
- Memory BW 1024 GB/s, >80%@STREAM Triad

**A64FX**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISA (Base, extension)</td>
<td>Armv8.2-A, SVE</td>
</tr>
<tr>
<td>Process technology</td>
<td>7 nm</td>
</tr>
<tr>
<td>Peak DP performance</td>
<td>&gt;2.7 TFLOPS</td>
</tr>
<tr>
<td>SIMD width</td>
<td>512-bit</td>
</tr>
<tr>
<td># of compute cores</td>
<td>48</td>
</tr>
<tr>
<td>Memory capacity</td>
<td>32 GiB (HBM2 x4)</td>
</tr>
<tr>
<td>Memory peak bandwidth</td>
<td>1024 GB/s</td>
</tr>
<tr>
<td>PCIe</td>
<td>Gen3 16 lanes</td>
</tr>
<tr>
<td>High speed interconnect</td>
<td>TofuD integrated</td>
</tr>
</tbody>
</table>

CMG: Core Memory Group  NOC: Network on Chip
Agenda

- Background
  - K computer and Post-K
  - Post-K goals and approaches
- Fujitsu new Arm CPU A64FX
- Post-K system
  - System software
  - Configuration
- Performance discussion
- Summary and development status
Software development: System software

RIKEN and Fujitsu are developing software stacks for Post-K

<table>
<thead>
<tr>
<th>Management software</th>
<th>File system</th>
<th>Programming environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>System management for high availability &amp; power saving operation</td>
<td>FEFS Lustre-based distributed file system</td>
<td>XcalableMP</td>
</tr>
<tr>
<td>Job management for higher system utilization &amp; power efficiency</td>
<td>LLIO NVM-based file I/O accelerator</td>
<td>MPI (Open MPI, MPICH)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OpenMP, COARRAY, Math.libs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compilers (C, C++, Fortran)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Debugging and tuning tools</td>
</tr>
</tbody>
</table>

Linux OS / McKernel (Lightweight kernel)

Post-K system hardware

Post-K Under development w/ RIKEN

RIKEN and Fujitsu are developing software stacks for Post-K.
### FEFS vs Lustre, LLIO vs Burst buffer

#### Post-K supports superior performance file system, FEFS + LLIO

<table>
<thead>
<tr>
<th></th>
<th>FEFS</th>
<th>Lustre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage network</td>
<td>Tofu, InfiniBand, Ether, OPA</td>
<td>InfiniBand, Ether, OPA</td>
</tr>
<tr>
<td>Injection BW per node</td>
<td>40.8 GB/s</td>
<td>12.5 GB/s (IB EDR)</td>
</tr>
<tr>
<td>MDS scalability</td>
<td>Yes</td>
<td>Yes (ver. 2.4~)</td>
</tr>
<tr>
<td>Interoperability</td>
<td>x86, Arm (A64FX)</td>
<td>x86, Power, Arm?</td>
</tr>
<tr>
<td></td>
<td>FEFS + LLIO*</td>
<td>Lustre + Burst buffer**</td>
</tr>
<tr>
<td>Local temporary file acc.</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Cache file acc.</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Shared file within a job.</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Explicit function call</td>
<td>Not required (by mount point)</td>
<td>Required (Source modification)</td>
</tr>
</tbody>
</table>

*Lightweight Layered IO-Accelerator  
**DDN IME
LLIO (Lightweight Layered IO-accelerator) overview

- Maximizing application file I/O performance
- Easy access to User Data: File Cache of Global File System
- Higher Data Access Performance: Temporary Local FS (in a process)
- Higher Data Sharing Performance: Temporary Shared FS (among processes)
## Post-K system configuration

### Scalable design

<table>
<thead>
<tr>
<th>Unit</th>
<th># of nodes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>1</td>
<td>Single socket node with HBM2 &amp; Tofu interconnect D</td>
</tr>
<tr>
<td>CMU</td>
<td>2</td>
<td>CPU Memory Unit: 2x CPU</td>
</tr>
<tr>
<td>BoB</td>
<td>16</td>
<td>Bunch of Blades: 8x CMU</td>
</tr>
<tr>
<td>Shelf</td>
<td>48</td>
<td>3x BoB</td>
</tr>
<tr>
<td>Rack</td>
<td>384</td>
<td>8x Shelf</td>
</tr>
</tbody>
</table>
CMU: CPU Memory Unit

- A64FX CPU x2
- QSFP28 x3 for Active Optical Cables
- Single-side blind mate of signals & water
- ~100% direct water cooling
Agenda

- Background
  - K computer and Post-K
  - Post-K goals and approaches
- Fujitsu new Arm CPU A64FX
- Post-K system
  - System software
  - Configuration
- Performance discussion
- Summary and development status
Preliminary performance evaluation results

- Over 2.5x faster in HPC & AI benchmarks than SPARC64 XIfx
- Arm SVE, Fujitsu’s microarchitecture, and high bandwidth

### A64FX chip performance measurements & architectural contributions

<table>
<thead>
<tr>
<th>Throughput tests</th>
<th>Application kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory BW</td>
<td>512-bit SIMD</td>
</tr>
<tr>
<td>2.5 TF</td>
<td>840 GB/s</td>
</tr>
<tr>
<td>3.0x</td>
<td>Combined gather</td>
</tr>
<tr>
<td>2.8x</td>
<td>L1$ BW</td>
</tr>
<tr>
<td>3.4x</td>
<td>L2$ BW</td>
</tr>
<tr>
<td>2.5x</td>
<td>INT8 partial dot product</td>
</tr>
<tr>
<td>DGEMM</td>
<td>STREAM</td>
</tr>
<tr>
<td>Fluid dynamics</td>
<td>Atmosphere</td>
</tr>
<tr>
<td>Seismic wave propagation</td>
<td>Convolution FP32</td>
</tr>
<tr>
<td>Convolution low precision</td>
<td>9.4x</td>
</tr>
</tbody>
</table>

Baseline: SPARC64 XIfx
Agenda

- Background
  - K computer and Post-K
  - Post-K goals and approaches

- Fujitsu new Arm CPU A64FX

- Post-K system
  - System software
  - Configuration

- Performance discussion

- Summary and development status
Summary of Post-K

- High application performance and good power efficiency
  - High memory bandwidth & wider SIMD
- Arm SVE support
  - Scalable vector extension, state of art Arm instruction set architecture
  - Binary compatibility advances and expands the power of ecosystem by incorporating HPC technologies and applications
- Optimizing for new supercomputer standards
  - Low power consumption in many aspects
  - Being focusing on the existing applications and emerging applications

Post-K inheriting the strong and proven microarchitecture with HPC system software will be more powerful with Arm and its ecosystem
Fujitsu is currently developing Post-K, successor to K computer, with RIKEN OS running and design verification is underway as scheduled. RIKEN announced Post-K early access program to begin around Q2/CY2020.

System development status: Post-K

- **Fujitsu** is currently developing Post-K, successor to K computer, with RIKEN OS running and design verification is underway as scheduled.
- RIKEN announced Post-K early access program to begin around Q2/CY2020.

![Post-K prototype rack](https://postk-web.r-ccs.riken.jp/sched.html)
FUJITSU

shaping tomorrow with you