Fujitsu HPC and AI Processors

Takumi Maruyama
Senior Director
AI Platform Business Unit
Advanced System Research & Development Unit
Agenda

■ K computer

■ Fujitsu’s latest processors
  ■ HPC
  ■ UNIX

■ Future Fujitsu processors under development
  ■ Post K
  ■ AI processor: DLU

■ Summary
K Computer
WR#1
10.51 PFlops (Top500, 2011/11)
38,621 GTEPS (Graph500, 2016/11)
602.7 TFLOPS (HPCG, 2016/11)
Fujitsu Technologies in the K computer

- **High Performance Processor**: 8-core
- **Liquid Cooling**
- **4 Processors**
- **Torus Network**: 6D.
- **High Density Rack**
- **864 racks**
- **82,944 Compute nodes**
- **5,184 IO nodes**
- **24 boards**

©RIKEN
The Latest Fujitsu Processors
Perpetual Evolution > 60 years: Always Targeting No.1

Virtual Machine Architecture
Software on Chip
High-speed Interconnect

HPC-ACE
System on Chip
Hardware Barrier

Multi-core Multi-thread
L2$ on Die
Non-Blocking $
O-O-O Execution
Super-Scalar
Single-chip CPU
Store Ahead
Branch History
Prefetch

Mainframe/UNIX/HPC + AI incremental development

SPARC64™ XIfx Chip (HPC)

- **Architecture Features**
  - 32 computing cores + 2 assistant cores
  - HPC-ACE2 (256bit SIMD) Fujitsu’s ISA enhancements
  - Sector Cache: Cache with SW controllability
  - 24 MB L2 cache

- **20nm CMOS**
  - 3,750M transistors
  - 2.2GHz

- **Performance (peak)**
  - 1.1TFlops
  - HMC 240GB/s x 2 (in/out)
  - Tofu2 125GB/s x 2 (in/out)
SPARC64™ XII Chip (UNIX)

- **Architecture Features**
  - 12 cores x 8 threads
  - SWoC (“Software on Chip”)
    - Fujitsu’s ISA enhancements
  - 32MB L3 cache
  - Embedded MAC and IOC
- **20nm CMOS**
  - 25.8mm x 30.8mm
  - 5,450M transistors
  - 4.25GHz (up to 4.35GHz with “High Speed Mode” enabled)
- **Performance (peak)**
  - 417GIPS / 835GFlops
  - 153GB/s memory throughput

Multiple big cores, High CPU GHz
SPARC64™ XII (UNIX) Pipeline

Fetch | Decode | Issue | Reg-Read | Execute | Cache and Memory | Commit

- Branch Prediction
- Instruction Buffer
- L1 Instruction Cache 64KB
- Instruction Buffer
- Decode
- RSBR Reservation Station for Branch
- RSA Reservation Station for Address generation
- RSE Reservation Station for Execution
- RSF Reservation Station for Floating-point

Execute
- FPR x4
- FPR Update Buffer
- FUB
- GPR x4
- GUB
- EAGA
- EAGB
- EXA
- EXB
- FLA
- FLB
- FLD
- FLA
- FLB
- FLD

Cache and Memory
- L1 Data Cache 32KB
- L1 Data Cache 32KB
- dTLB
- L2 Cache
- L3 Cache
- CPU-CPU i/f
- MAC
- IOC

Commit
- Commit Stack Entry
- Program Counter x4
- Control Registers x4

Shared Micro-architecture

12 cores
Future Fujitsu Processors Under Development

- Post K
Project Overview

• RIKEN and Fujitsu are currently developing the post-K computer, which aims to be the most advanced general-purpose supercomputer in the world.

Goals of Japan’s Post-K Development Project

• Application performance
• Low power consumption
• User convenience
• Ability to produce ground-breaking results
## Post-K Processor and Interconnect Features

- Fujitsu Processor, adopting ARM ISA and enhanced Tofu interconnect
- Inherits and enhances the K computer’s innovative features

<table>
<thead>
<tr>
<th>Processor</th>
<th>Functions &amp; Architecture</th>
<th>Post-K</th>
<th>K computer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base ISA + SIMD Extensions</td>
<td>ARMv8-A+SVE</td>
<td>SPARCv9+HPC-ACE</td>
</tr>
<tr>
<td></td>
<td>SIMD width [bit]</td>
<td>512</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>FP16 (half precision) support</td>
<td>✔</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>FMA: Floating-point multiply and add</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Math. acceleration primitives</td>
<td>✔ Enhanced</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Inter-core barrier</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Sector cache</td>
<td>✔ Enhanced</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>Hardware “prefetch” assist</td>
<td>✔ Enhanced</td>
<td>✔</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Tofu</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copyright 2017 FUJITSU LIMITED
Post-K Processor Supports FP16

- Provides optimized precision for a wide range of applications
  - Superior performance
  - Reduces required bandwidth and power consumption

- Target applications:
  - Existing numerical applications
  - Brand-new applications, including Deep Learning

![Diagram showing Double Precision, Single Precision, Half Precision, and High Performance for More Applications]
Future Fujitsu Processor Development

- AI Processor (DLU™)
Processor Designed for Deep Learning

Utilizing technologies derived from the K computer

Features of DLU
- Architecture designed for Deep Learning
- Low power consumption design
- Optimized precision
  ➔ Goal: 10x Performance / Watt compared to competitors
- Scalable design with Tofu interconnect technology
  ➔ Ability to handle large-scale neural networks

DLU™
(Deep Learning Unit)

FY2018 ~

Copyright 2017 FUJITSU LIMITED
DLU Design Target

- High Deep Learning performance / watt: 10x performance / watt

- However, high performance and low power is not easy to achieve at the same time

<table>
<thead>
<tr>
<th>High Performance</th>
<th>Conflicting Demands</th>
<th>Low Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>More transistors • state of the art O-O-O • many execution units/$ • Higher Frequency</td>
<td></td>
<td>Less Transistors • less control logic • fewer execution units/$ • Lower Frequency</td>
</tr>
</tbody>
</table>
Need for a New Architecture

- A new architecture is required for the DLU to achieve the target.
- The architecture is domain specific – Deep Learning
What’s the New Architecture for the DLU?

- Domain specific, Optimal precision, and Massively parallel.

<table>
<thead>
<tr>
<th>Conventional Architecture</th>
<th>The New Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Use</strong></td>
<td>1. Domain Specific</td>
</tr>
<tr>
<td>Complicated O-O-O cores</td>
<td>Domain specific cores</td>
</tr>
<tr>
<td><strong>High Precision</strong></td>
<td>2. Optimal Precision</td>
</tr>
<tr>
<td>Double/Single precision FP</td>
<td>Deep Learning Integer</td>
</tr>
<tr>
<td><strong>Sequential + Parallel</strong></td>
<td>3. Massively Parallel</td>
</tr>
<tr>
<td>Multiple strong cores</td>
<td>Many cores w/ on-chip network</td>
</tr>
</tbody>
</table>
DLU Architecture

1. Domain specific Domain specific Cores
   - Newly designed ISA
   - Simplified μ-architecture
   - Fully software visible and controllable
   - Heterogeneous cores ★
   - DPE and Large RF ★

2. Optimal Precision Deep Learning Integer ★

3. Massively Parallel Many DPUs with an On-chip Network

Large scale DLU interconnect through off-chip network
Heterogeneous Cores

- The combination of few large core (Master) and many small execution cores (DPU) results in more performance with less power consumption, compared to a conventional homogeneous structure.

Master Core: Memory Access and DPU control
- Push & Pull instructions and data for DLUs.
- Start/stop execution of DLUs.

DPU: Execution
- Execute DL operations based on master core’s control.

How to utilize many DPUs (convolution example)
- one CH-out / DPU
- multiple batch / DPU
DPE & Large RF (Register File)

- DPU consists of 16 DPEs connected with on-chip network
- DPE includes large RF and wide SIMD execution units to realize an efficient Deep Learning engine.
  - RF is fully SW controllable unlike cache to extract full HW potential

DPU: 128 SIMD* / 16DPE

DPE: 8SIMD* with large RF (~100x of typical CPU core) * For FP32

<table>
<thead>
<tr>
<th>Name</th>
<th>RF/$ structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNIX</td>
<td>SPARC64 XII</td>
</tr>
<tr>
<td>HPC</td>
<td>SPARC64 XIfx</td>
</tr>
<tr>
<td>AI</td>
<td>DLU</td>
</tr>
</tbody>
</table>

More SW controllability
Fujitsu’s “Deep Learning Integer” realizes necessary accuracy for Deep Learning with only a 16 or 8 bit data size (i.e. less power consumption compared with FP32)
Deep Learning Integer Accuracy

- Deep Learning Integer has shown similar accuracy with FP32 for Deep Learning

(*) ImageNet(subset): image size=96x96, #categories=25
DLU Roadmap

- Multiple generations of DLUs over time, as we currently do for HPC/UNIX/Mainframe processors

1st Generation
- Host CPU required
- Inter-DLU direct connection

2nd Generation
- Embedded host CPU

Other special processors
- Neuromorphic
- Combinatorial optimization

FY2018

Future

* Subject to change without notice
Summary
Fujitsu Processor Design Style

- Standard ISA with FJ enhancements / newly developed ISA
- Shared / Simple + SW visible micro-architecture
- The latest semiconductor technology
- Shared design infrastructure: Circuit, Methodology, People

<table>
<thead>
<tr>
<th></th>
<th>General Purpose</th>
<th>Domain Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instruction Set Architecture</strong></td>
<td>Mainframe</td>
<td>UNIX</td>
</tr>
<tr>
<td>(HW-SW I/F)</td>
<td>GS ISA</td>
<td>SPARC ISA</td>
</tr>
<tr>
<td><strong>Micro-architecture</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CPU internal structure)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>★ Performance/RAS</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Semiconductor Technology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Design Infrastructure</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Shared [FJ development] [Simple SW visible]
- The latest Circuit, Methodology, People

Copyright 2017 FUJITSU LIMITED
Fujitsu Processor Direction

- General purpose and Domain specific
- Wider variety of processors in the future to meet different requirements.
Summary

- Fujitsu has designed processors for a long time (> 60 years)
  - Perpetual evolution over generations

- SPARC64 IXfx (HPC), SPARC64 XII (UNIX), and Post-K
  - General purpose computing

- DLU
  - Domain specific
  - New Architecture
    - Heterogeneous, DPE and large RF, Deep Learning Integer

- Shared Design infrastructure: Circuit, Methodology, People

- Fujitsu will continue to develop cutting-edge processors to meet the needs of a new era.