The Tofu Interconnect 2

Yuichiro Ajima, Tomohiro Inoue, Shinya Hiramoto, Shun Ando, Masahiro Maeda, Takahide Yoshikawa, Koji Hosoe, and Toshiyuki Shimizu

Fujitsu Limited
Introduction

- **Tofu interconnect**
  - Highly-scalable six-dimensional network topology

- **Tofu interconnect 2**
  - Link speed has been upgraded from 40 Gbps to 100 Gbps
  - System-on-Chip integration
  - New features

---

**Tofu interconnect (Tofu1)**

- K computer
- FX10

**Tofu interconnect 2 (Tofu2)**

- Post-FX10

---

2010 2012 2015
A Next Generation Interconnect

- Optical-dominant: 2/3 of network links are optical

![Gross raw bitrate per node (Tbps)]

- IBM Blue Gene/Q: 10 Gbps
- IB-FDR Fat-Tree: 14 Gbps (5 Gbps, 14 Gbps)
- Cray Cascade: 12.5 Gbps (14 Gbps)
- Tofu1: 6.25 Gbps (25 Gbps)
- Tofu2: 25 Gbps (25 Gbps)
Agenda

- Introduction
- Implementation
- Transmission Technology
- New Features
- Preliminary evaluations
- Summary
Reduction in the Chip Area Size

- The process technology shrinks from 65 to 20 nm
- System-on-chip integration eliminates the host bus interface
- Chip area shrinks to 1/3 size
Optical Modules

- Optical modules are placed next to the processor SoC
- 25 Gbps high-speed signals connect the optical modules

SPARC64™ XI fx integrating Tofu2

25 Gbps high-speed signals

Board-mounted optical assembly

(12 lanes for each)

Electrical connector (8 lanes)
Physical Topology on a Board

- 3 nodes on a CPU/memory board (CMB)
  - 10 ports for each node
- 20 ports for optical links
- 10 ports for electrical links
  - 4 ports for connection to the neighbor nodes on the CMB
  - 6 ports for connection to the other CMBs within the chassis
Logical Topology in a Chassis

- Six-dimensional network: \{X, Y, Z, A, B, C\}
- The size of a chassis = \{1, 1, 3, 2, 1, 2\}
  - 3 nodes on a CMB connects along the Z-axis
  - 4 CMBs in a chassis connect by the A- and C-axes
Packaging Hierarchy

- A chassis is 2U sized
- A 19-inch rack mounts a maximum of 18 chassis
- Racks are interconnected by links to the X- and Y-axes

CPU/memory board
3 nodes

2U chassis
12 nodes

19-inch rack
216 nodes

Post-FX10 System
Petaflops per 5 racks

SPARC64™ XI fx
1 TFlops

Copyright 2014 FUJITSU LIMITED
Agenda

- Introduction
- Implementation
- Transmission Technology
- New Features
- Preliminary evaluations
- Summary
Physical Layer

- The physical layer is based on 100GbE (100GBASE-SR4)
- The data transfer rate is 25.78125 Gbps
  - Increased more than fourfold from 6.25 Gbps
- The encoding scheme is 64b/66b
  - Enhanced from 8b/10b
- Each link has 4 lanes of signals
- Each link provides 12.5 GB/s peak throughput
  - Increased by 2.5 times from 5.0 GB/s

<table>
<thead>
<tr>
<th></th>
<th>Tofu1</th>
<th>Tofu2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data transfer rate</td>
<td>6.25 Gbps</td>
<td>25.78125 Gbps</td>
</tr>
<tr>
<td>Encoding</td>
<td>8b/10b</td>
<td>64b/66b</td>
</tr>
<tr>
<td>Signals per link</td>
<td>8 lanes</td>
<td>4 lanes</td>
</tr>
<tr>
<td>Link throughput</td>
<td>5.0 GB/s</td>
<td>12.5 GB/s</td>
</tr>
</tbody>
</table>
Frame Format

- Tofu2 harnesses the Ethernet frame format
  - Preamble, frame check sequence and inter-frame gap are the same
- The frame size is a multiple of 32 bytes
- The maximum frame size is 2016 bytes
- Single frame encapsulates transport and data-link packets
  - Transport and data-link layer packets of Tofu1 had their own frame

### Ethernet frame format

<table>
<thead>
<tr>
<th>8 bytes</th>
<th>6 bytes</th>
<th>6 bytes</th>
<th>2 bytes</th>
<th>46 – 1500 bytes</th>
<th>4 bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preamble</td>
<td>Source Address</td>
<td>Destination Address</td>
<td>Type / Length</td>
<td>Data</td>
<td>FCS</td>
</tr>
</tbody>
</table>

### Tofu2 frame format

<table>
<thead>
<tr>
<th>8 bytes</th>
<th>8 bytes</th>
<th>32 – 1984 bytes</th>
<th>16 bytes</th>
<th>4 bytes</th>
<th>4 bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preamble</td>
<td>Routing Header</td>
<td>Transport Layer Packet</td>
<td>Data-Link Layer Packet</td>
<td>Flags</td>
<td>FCS</td>
</tr>
</tbody>
</table>
Agenda

- Introduction
- Implementation
- Transmission Technology
- New Features
- Preliminary evaluations
- Summary
Cache Injection

- The major issue of latency: cache flush
  - Ordinary DMA from a peripheral causes flushing of CPU cache

- Tofu2 can inject received data into L2 cache directly
  - Bypassing main memory
  - Data to be injected is indicated with a flag bit that is set by the sender

- Tofu2’s cache injection feature is harmless
  - Cache injection is only performed when cache hits and the line is in exclusive state
  - A cache line in exclusive state is highly likely to be polled by a corresponding processor core
Session-mode Control Queue

- Offloading a collective communication of long messages
- Command execution may be delayed until a reception of Put
- Control flow can be branched or joined

An example of handshaking pipelined gather in a ring logical-topology

Normal-mode  Session-mode  Session-mode

Wait for command
Enqueuing commands
Execution
Wait for command and session progress
Wait for session progress

Command queueing can be delayed

Put 0B, +2
Put Data
Put 0B, +1
Put Data, +1
Put Data
NOP

Wait for session progress
RDMA Atomic Read-Modify-Write

- Atomically read, modify and write back remote data
  - Typical operations: compare-and-swap and fetch-and-add
  - Usage: software-based synchronization and lock-free algorithms

- Atomicity
  - Guaranteed by extending the coherency protocol of processor
    - not by extending each network interface
  - Strong atomicity: No memory accesses can break atomicity
  - Mutual atomicity: Atomic operations of processor and Tofu2 mutually guarantee their atomicity

- Mutual atomicity enables an efficient implementation of unified multi-process and multi-thread runtime
Agenda

- Introduction
- Implementation
- Transmission Technology
- New Features
- Preliminary evaluations
- Summary
Evaluation Environment

- System-level logic simulations
  - Using Cadence Palladium XP hardware emulator

- Simulation model
  - Verilog RTL codes for the production
  - Multiple nodes of Post-FX10

- Test programs
  - Executed on the simulated processor cores
  - Used Tofu2 hardware directly

- Simulation waveform
  - Provides detail information of signal propagation delay in logic circuits
  - Latency results can be obtained directly from simulation waveforms
  - Throughput results were calculated from measured latency values
Throughput of Single Put Transfer

- Achieved 11.46 GB/s of throughput which is 92% efficiency.
Throughput of Concurrent Put Transfers

- Linear increase in throughput without the host bus bottleneck

![Graph showing throughput of put transfers for Tofu2 and Tofu1 with linear scaling and host bus bottleneck](image-url)

- Tofu2
- Tofu1

- 1-way
- 2-way
- 3-way
- 4-way

- Throughput of Put transfer (GB/s)

- Linear scaling

- Host bus bottleneck

- Throughput values:
  - Tofu2: 45.82 GB/s (4-way)
  - Tofu1: 17.63 GB/s (4-way)
Communication Latencies

- Cache injection reduced latency by 0.16 μsec

- Overheads of the other two features are low
  - Session-mode introduced no additional latency
  - The overhead of an atomic operation was about 0.1 μsec

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Method</th>
<th>Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-way</td>
<td>Put 8-byte to memory</td>
<td>0.87 μsec</td>
</tr>
<tr>
<td></td>
<td>Put 8-byte to cache</td>
<td>0.71 μsec</td>
</tr>
<tr>
<td>Round-trip</td>
<td>Put 8-byte ping-pong by CPU</td>
<td>1.42 μsec</td>
</tr>
<tr>
<td></td>
<td>Put 8-byte ping-pong by session</td>
<td>1.41 μsec</td>
</tr>
<tr>
<td></td>
<td>Atomic RMW 8-byte</td>
<td>1.53 μsec</td>
</tr>
</tbody>
</table>
Breakdown of Latency

- More than 0.2 μsec less latency than Tofu1

Estimated cumulative latency (nsec)

- Cache injection (~0.16 μsec)
- Host bus elimination (~0.06 μsec)

Tofu1 vs Tofu2

- Rx CPU
- Rx Host bus
- Rx TNI
- Packet Transfer
- Tx TNI
- Tx Host bus
- Tx CPU
Agenda

- Introduction
- Implementation
- Transmission Technology
- New Features
- Preliminary evaluations
- Summary
Summary

- **Tofu interconnect 2: the next generation interconnect**
  - Optical-dominant: 2/3 of network links are optical
  - System-on-Chip integrated controller

- **New features**
  - Cache injection reduces communication latency without cache pollution
  - Session-mode offloads various collective communication algorithms
  - Atomic RMW functions guarantee mutual atomicity

- **Preliminary evaluation results**
  - Throughput 11.46 GB/s
  - Latency 0.71 μsec
  - Cache injection reduced latency by 0.16 μsec
  - Elimination of the host bus reduced latency by about 0.06 μsec
  - Session-mode introduced no additional latency
  - Overhead of an atomic operation was about 0.1 μsec
shaping tomorrow with you