FUJITSU

shaping tomorrow with you

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

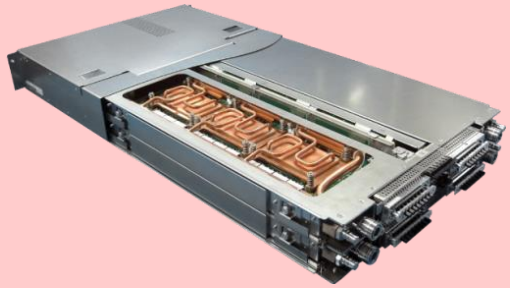
Tofu interconnect (Tofu1)

K computer

FX10

Tofu interconnect 2 (Tofu2)

Post-FX10

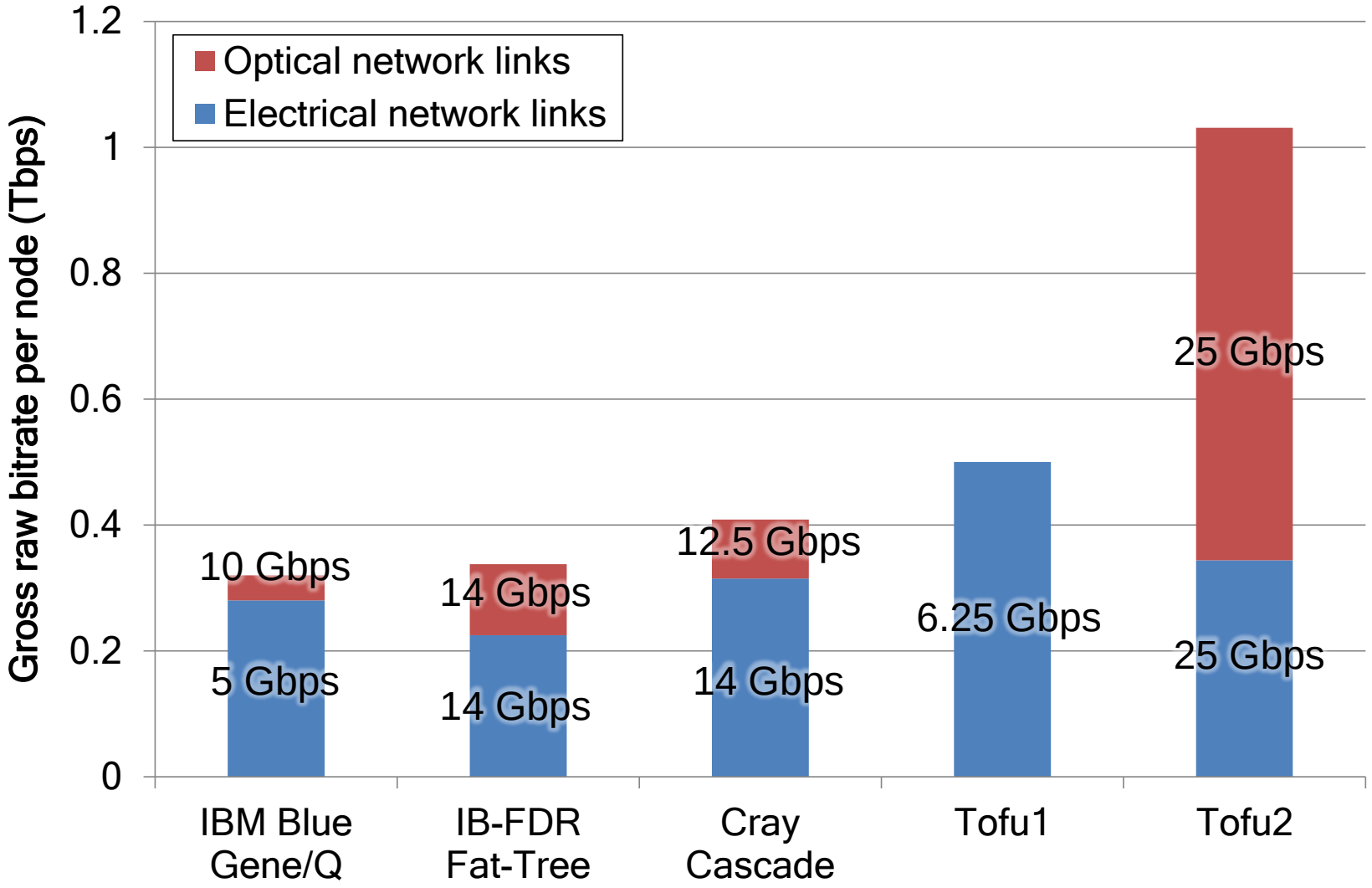


2010 → **2012** → **2015**

A Next Generation Interconnect

FUJITSU

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



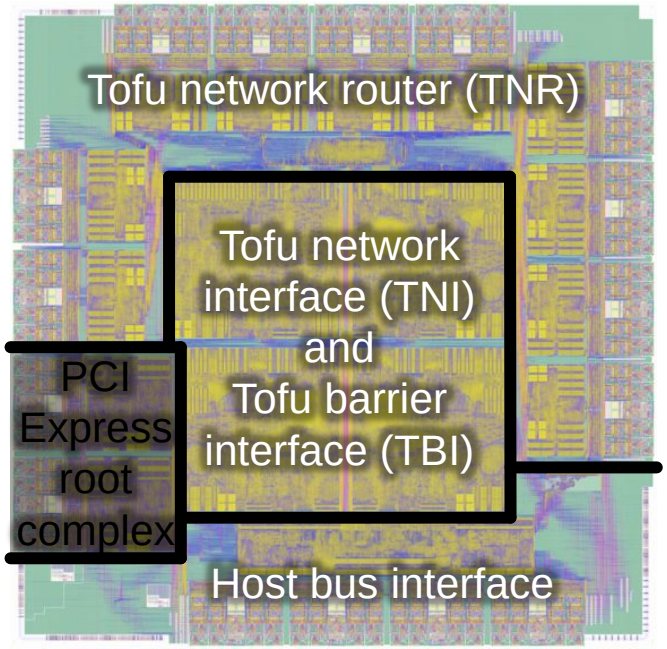
FUJITSU

Agenda

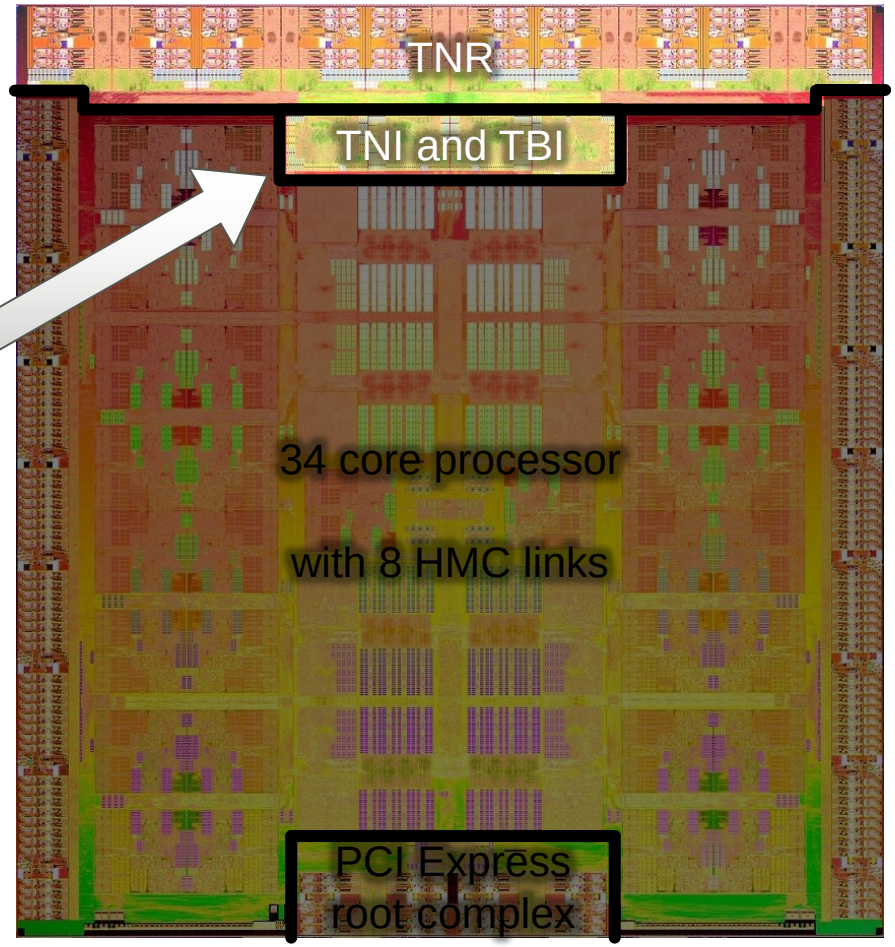
- Introduction
- Implementation
- Transmission Technology
- New Features
- Preliminary evaluations
- Summary

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



Tofu1 < 300mm^2
InterConnect Controller (65nm)



Tofu2 < 100mm^2 – SPARC64^{TM} XIfx (20nm)

Optical Modules

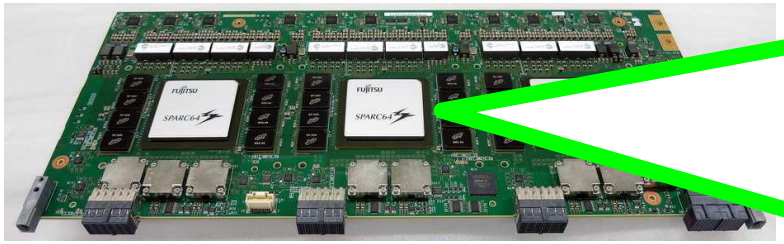
- Optical modules are placed next to the processor SoC
- 25 Gbps high-speed signals connect the optical modules

SPARC64™ XIfx integrating Tofu2

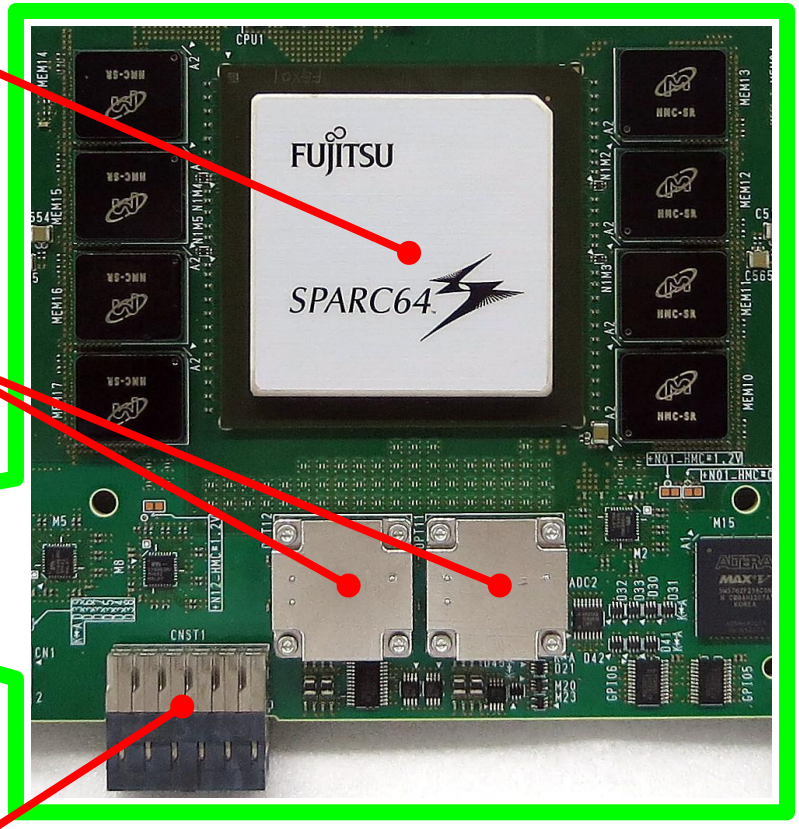
25 Gbps high-speed signals

Board-mounted optical assembly

(12 lanes for each)

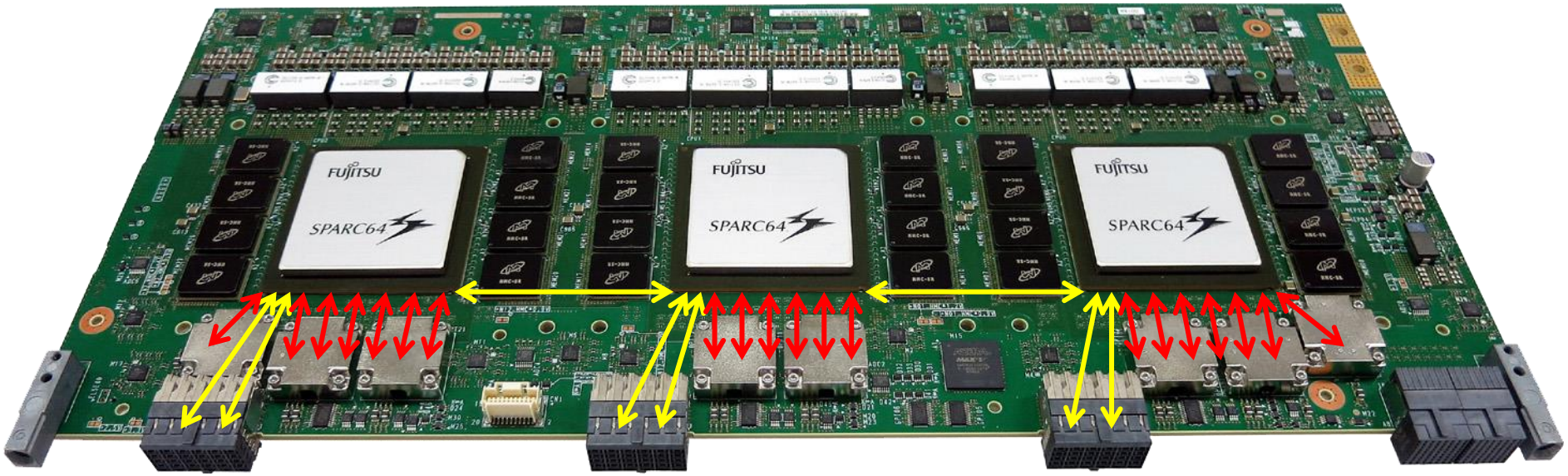


Electrical connector (8 lanes)



Physical Topology on a Board

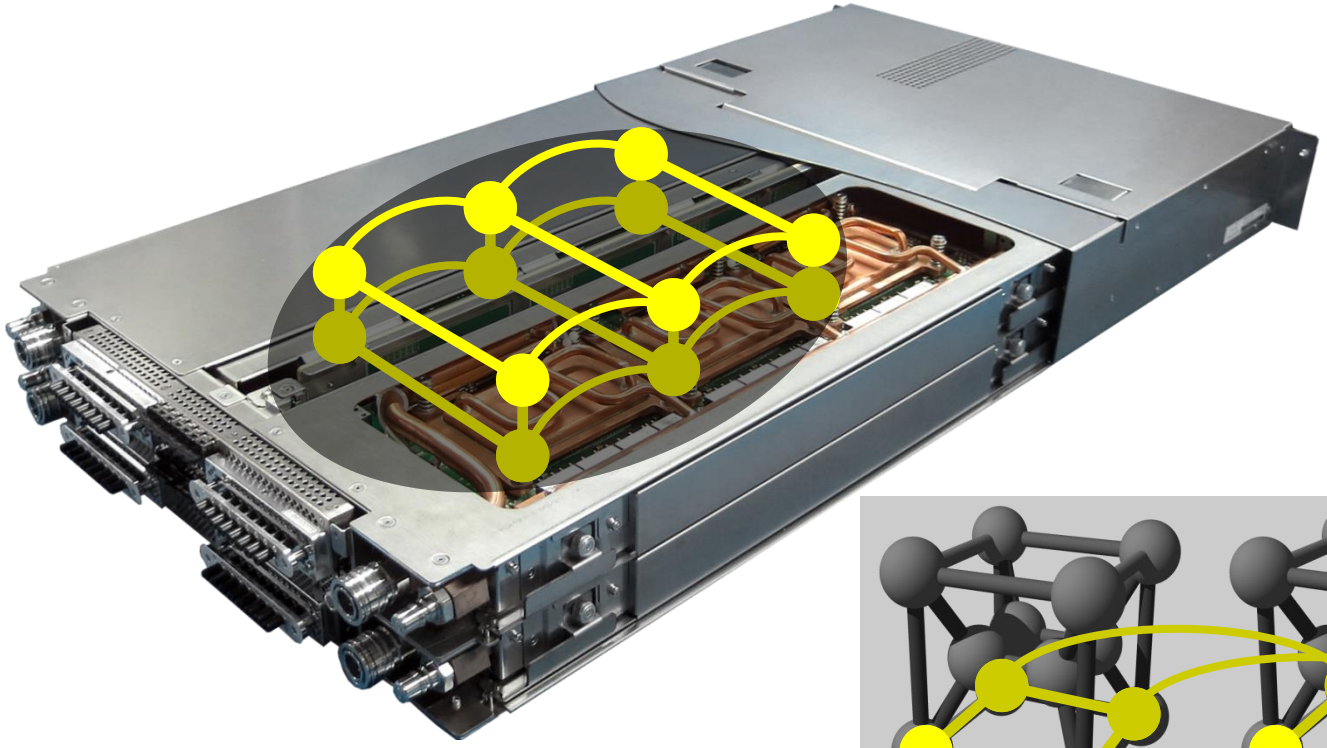
- 3 nodes on a CPU/memory board (CMB)
  - 10 ports for each node
- **20 ports for optical links**
- **10 ports for electrical links**
  - 4 ports for connection to the neighbor nodes on the CMB
  - 6 ports for connection to the other CMBs within the chassis

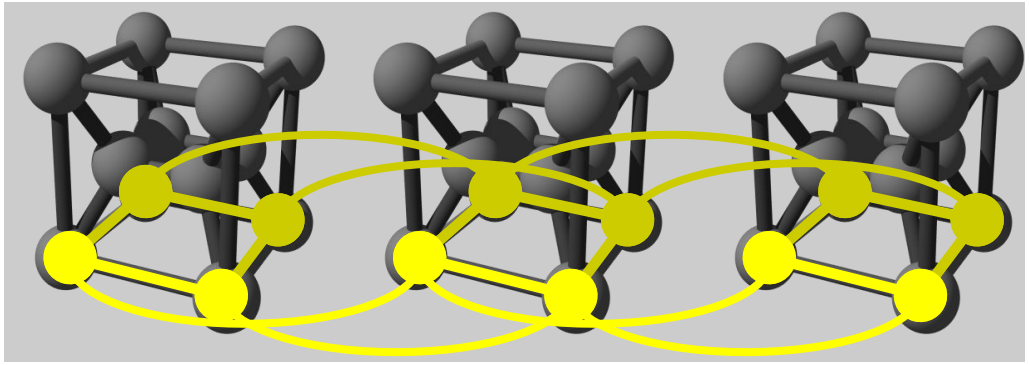


Logical Topology in a Chassis

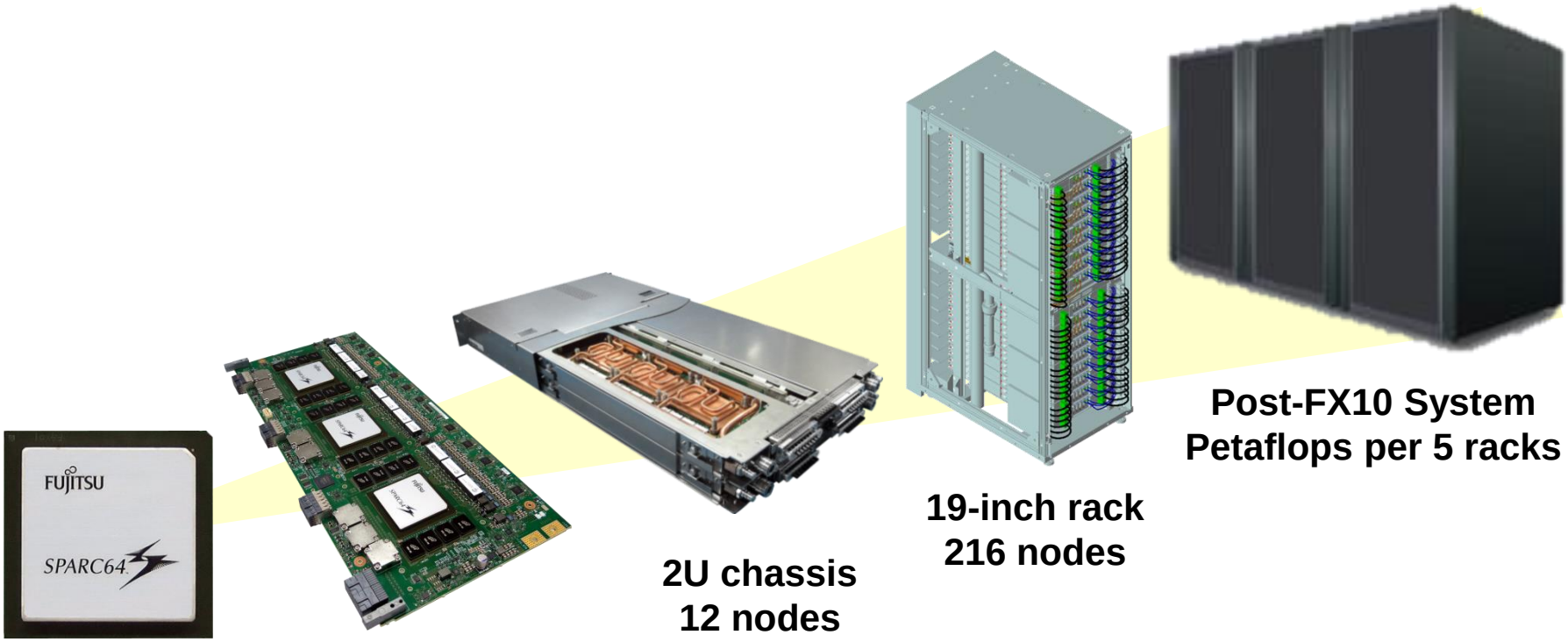
FUJITSU

- 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

FUJITSU

Agenda

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

Physical Layer

FUJITSU

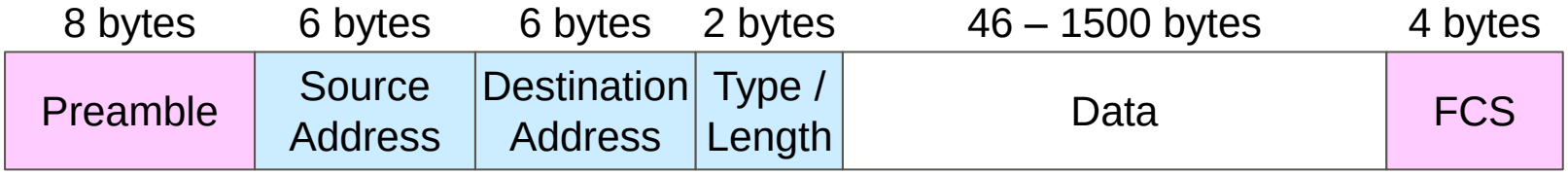
- The physical layer is based on 100GbE (100GBASE-SR4)
- The data transfer rate is 25.78125 Gbps
  - Increased more than fourfold from 6.25 Gbps
- The encoding scheme is 64b/66b
  - Enhanced from 8b/10b
- Each link has 4 lanes of signals
- Each link provides 12.5 GB/s peak throughput
  - Increased by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

	Tofu1	Tofu2
Data transfer rate	6.25 Gbps	25.78125 Gbps
Encoding	8b/10b	64b/66b
Signals per link	8 lanes	4 lanes
Link throughput	5.0 GB/s	12.5 GB/s

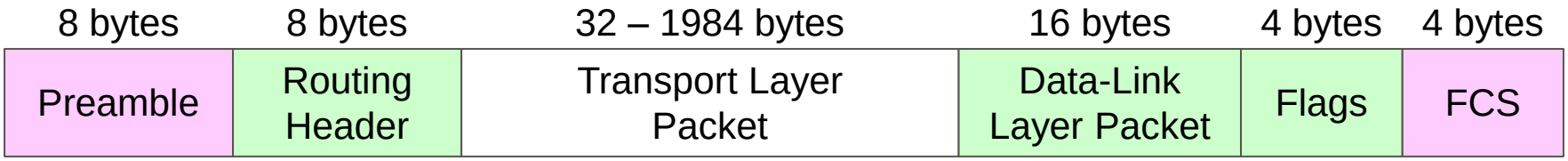
Frame Format

- 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



Tofu2 frame format



FUJITSU

Agenda

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

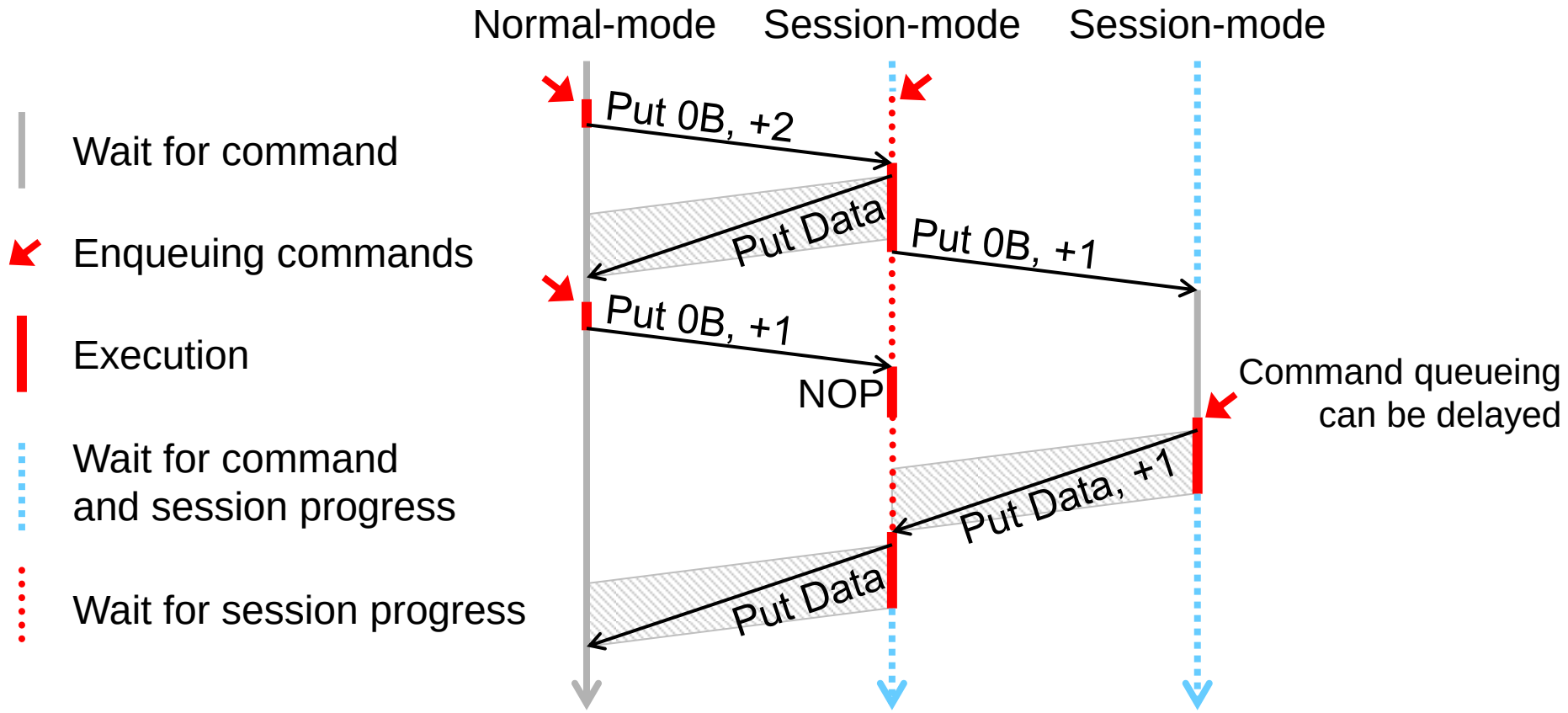
Cache Injection

FUJITSU

- 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

RDMA Atomic Read-Modify-Write

FUJITSU

- 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

FUJITSU

Agenda

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

Evaluation Environment

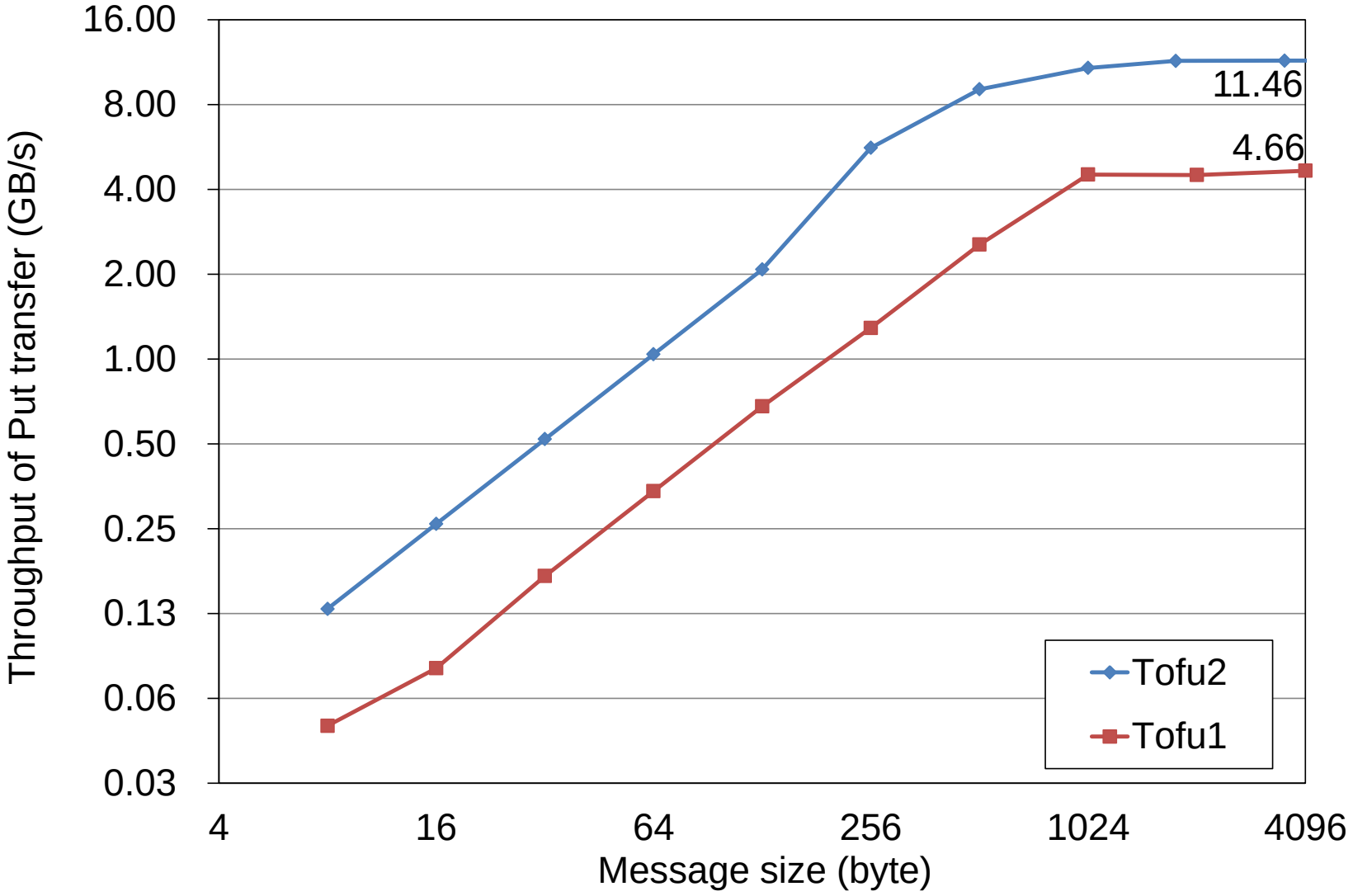
FUJITSU

- 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

FUJITSU

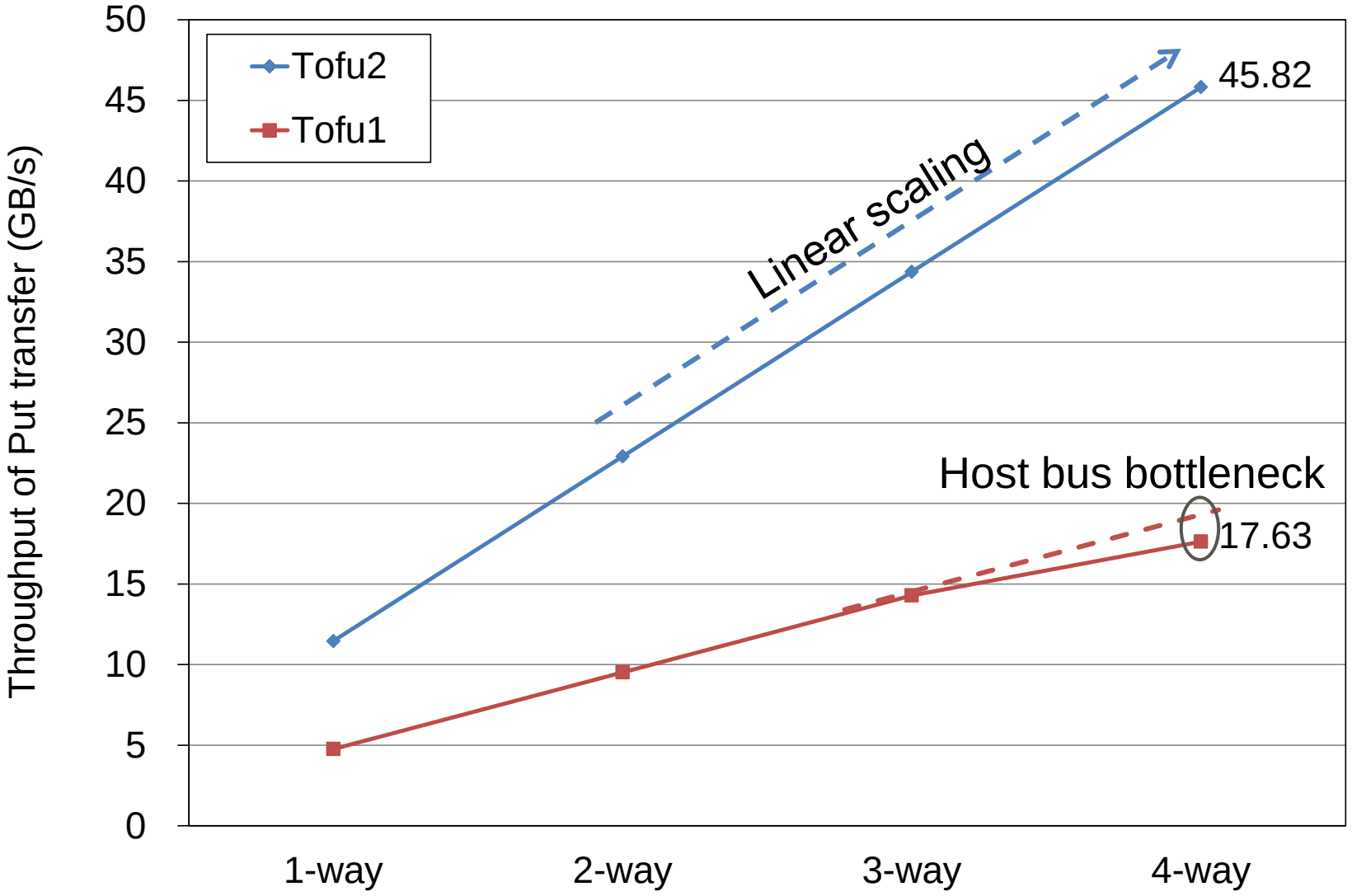
- Achieved 11.46 GB/s of throughput which is 92% efficiency



Throughput of Concurrent Put Transfers

FUJITSU

- Linear increase in throughput without the host bus bottleneck



Communication Latencies

FUJITSU

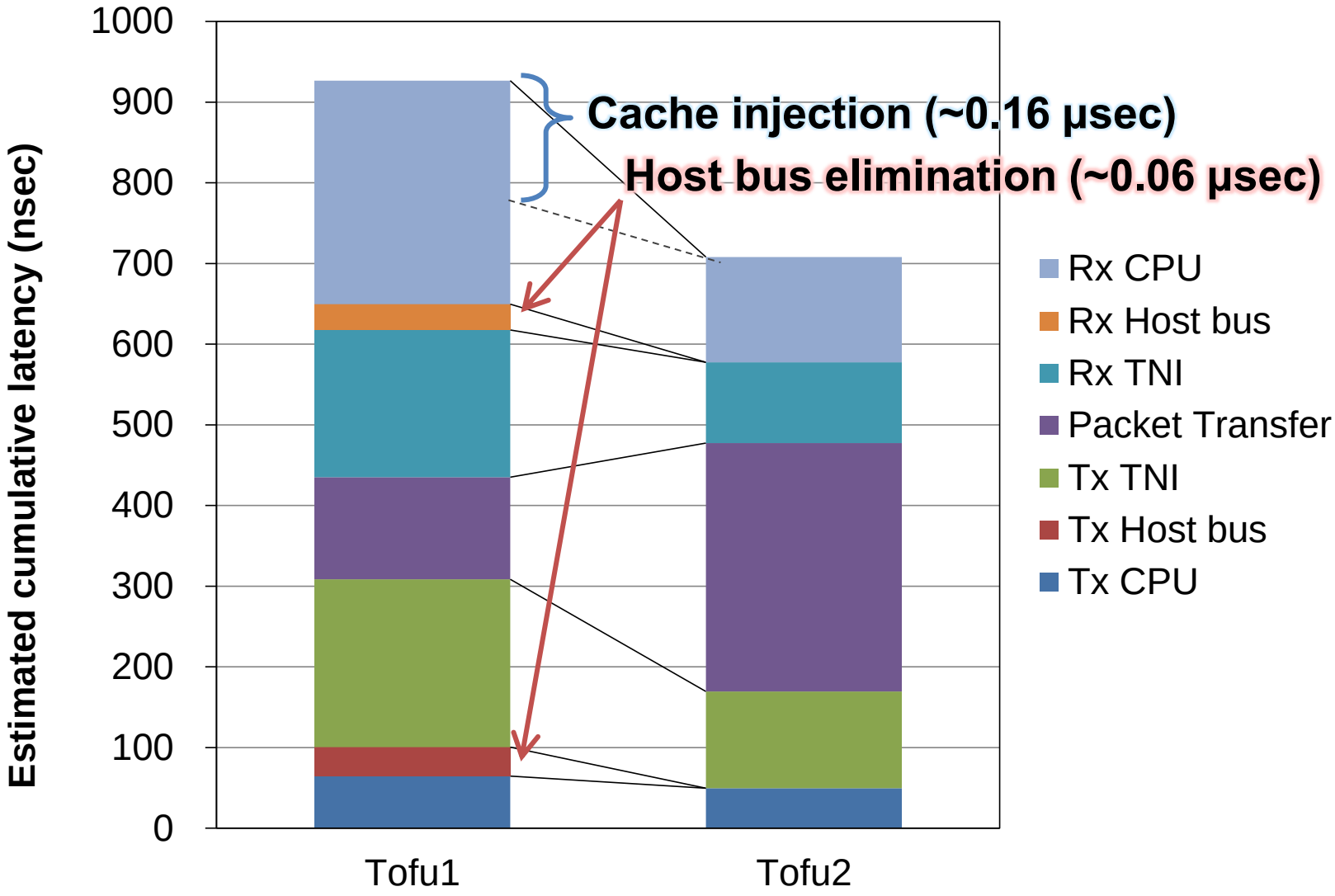
- Cache injection reduced latency by 0.16 µsec
- Overheads of the other two features are low
  - Session-mode introduced no additional latency
  - The overhead of an atomic operation was about 0.1 µsec

Pattern	Method	Latency
One-way	Put 8-byte to memory	0.87 µsec
	Put 8-byte to cache	0.71 µsec
Round-trip	Put 8-byte ping-pong by CPU	1.42 µsec
	Put 8-byte ping-pong by session	1.41 µsec
	Atomic RMW 8-byte	1.53 µsec

Breakdown of Latency

FUJITSU

- More than 0.2 μsec less latency than Tofu1



FUJITSU

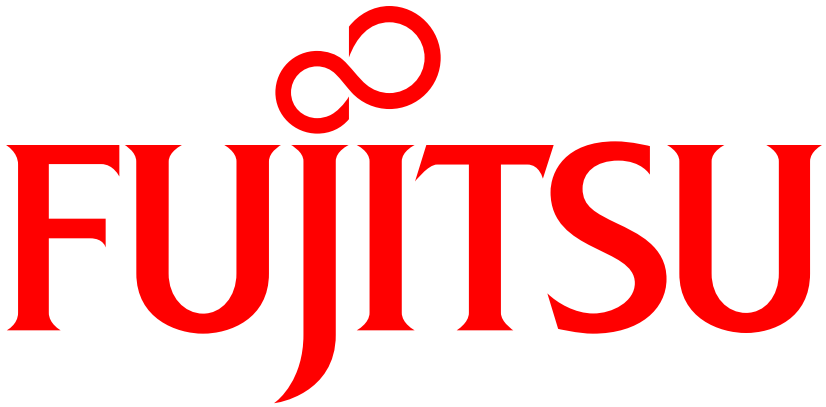
Agenda

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

Summary

FUJITSU

- 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