Tofu Interconnect 2: System-on-Chip Integration of High-Performance Interconnect

Yuichiro Ajima, Tomohiro Inoue, Shinya Hiramoto, Shunji Uno,
Shinji Sumimoto, Kenichi Miura, Naoyuki Shida, Takahiro Kawashima,
Takayuki Okamoto, Osamu Moriyama, Yoshiro Ikeda, Takekazu Tabata,
Takahide Yoshikawa, Ken Seki, and Toshiyuki Shimizu

Fujitsu Limited

25 June, 2014, ISC'14

Introduction

- **Tofu interconnect** (Tofu1) was developed for the K computer
  - '**To**rus **fu**sion' derives from its network topology
- Introducing **Tofu interconnect 2** (Tofu2)
  - Designed for Fujitsu's next generation machine Post-FX10
  - SoC integration, improved link speed and new efficient functions

K computer

8 core

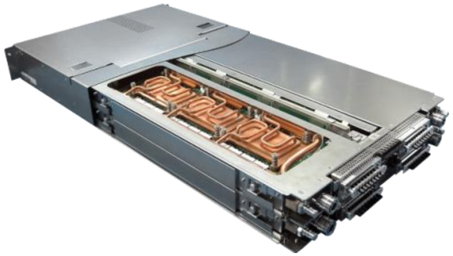
FX10

16 core

DDR3 SDRAM

Tofu interconnect

Post-FX10



32 core

Hybrid Memory Cube

Tofu interconnect 2

2010 **2012** **2015**

Index

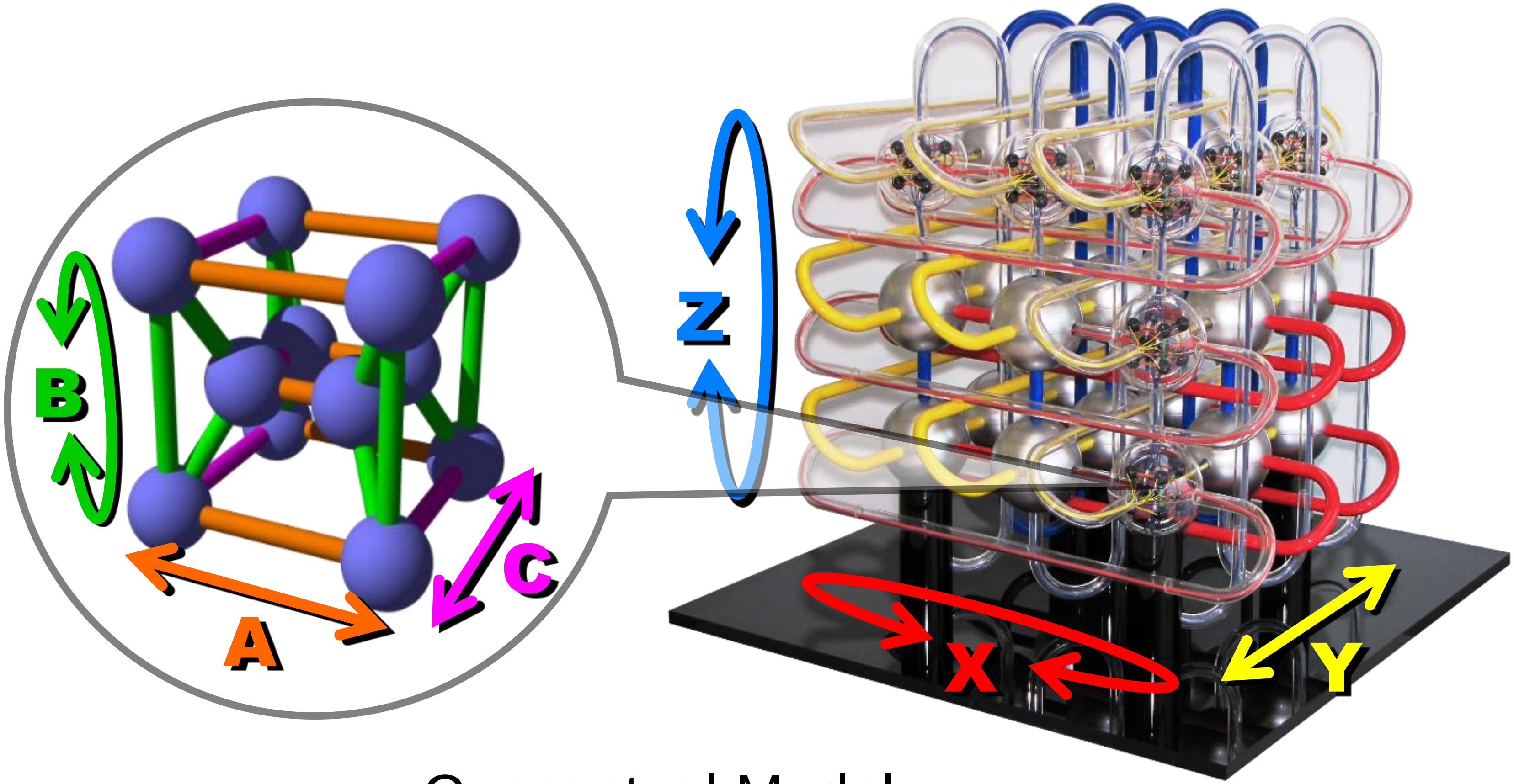
- Introduction
- Network topology
- SoC integration
- Improved link speed
- New efficient functions
- Summary

Summary of Tofu Interconnect 2

	Tofu1	Tofu2
Network topology	6D-Mesh/Torus	←
# of network interfaces	4	←
# of network links	10	←
Implementation	Discrete chip	SoC integration
Link speed	40 Gbps	100 Gbps
# of optical links	0	6 - 7
RDMA Put	✓	✓
RDMA Get	✓	✓
RDMA Atomic RMW		✓
Barrier synchronization	✓	✓
Non-blocking collective		✓
Memory bypass (sender)	✓	✓
Memory bypass (receiver)		✓

6D-Mesh/Torus network topology

- Every 3-D Cartesian grid point embeds a 3-D structure
- Higher bisection bandwidth than 3D-Torus
- Virtual torus for topology-aware communication algorithms



Conceptual Model

Index

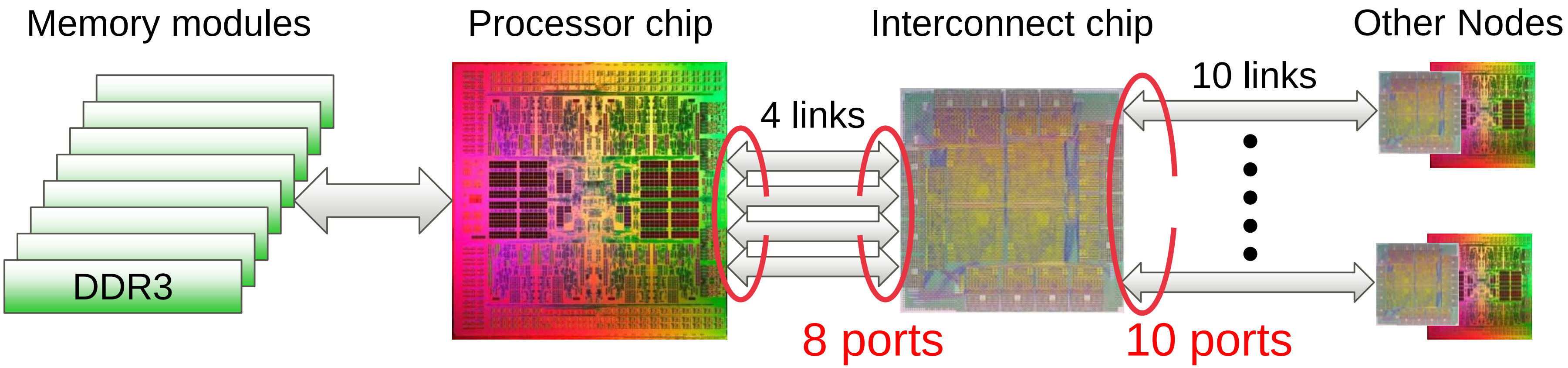
- Introduction
- Network topology
- SoC integration
- Improved link speed
- New efficient features
- Summary

Summary of Tofu Interconnect 2

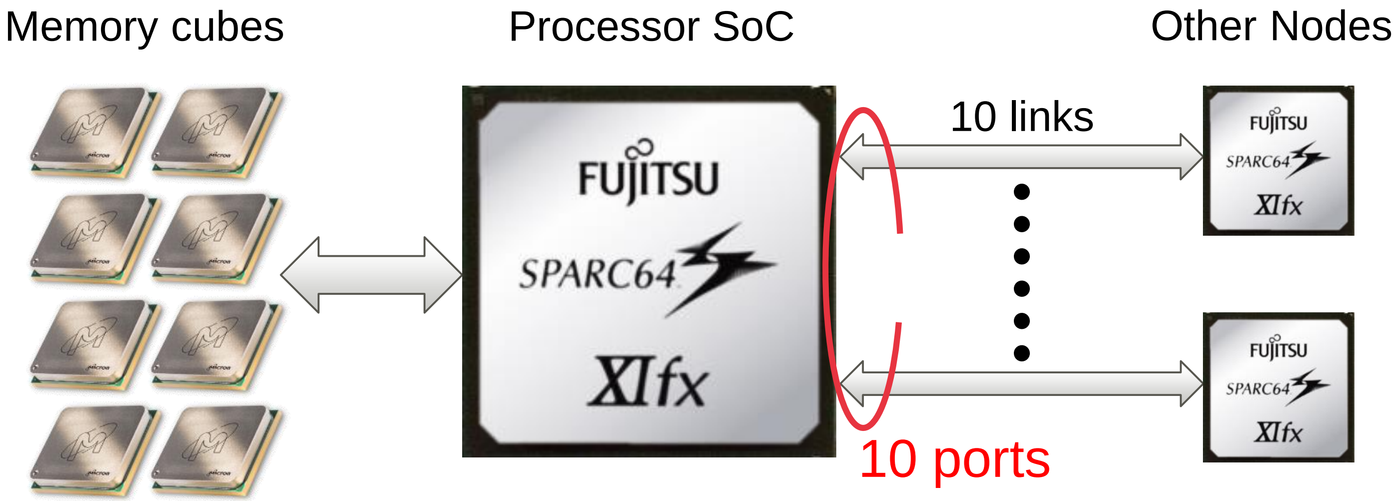
	Tofu1	Tofu2
Network topology	6D-Mesh/Torus	←
# of network interfaces	4	←
# of network links	10	←
Implementation	Discrete chip	SoC integration
Link speed	40 Gbps	100 Gbps
# of optical links	0	6 - 7
RDMA Put	✓	✓
RDMA Get	✓	✓
RDMA Atomic RMW		✓
Memory bypass (sender)	✓	✓
Memory bypass (receiver)		✓
Barrier synchronization	✓	✓
Non-blocking collective		✓

System-on-Chip Integration

- Tofu1 was implemented as a discrete chip



- Tofu2 is integrated into a processor SoC



- Number of ports per node decreased from 18 to 10

Index

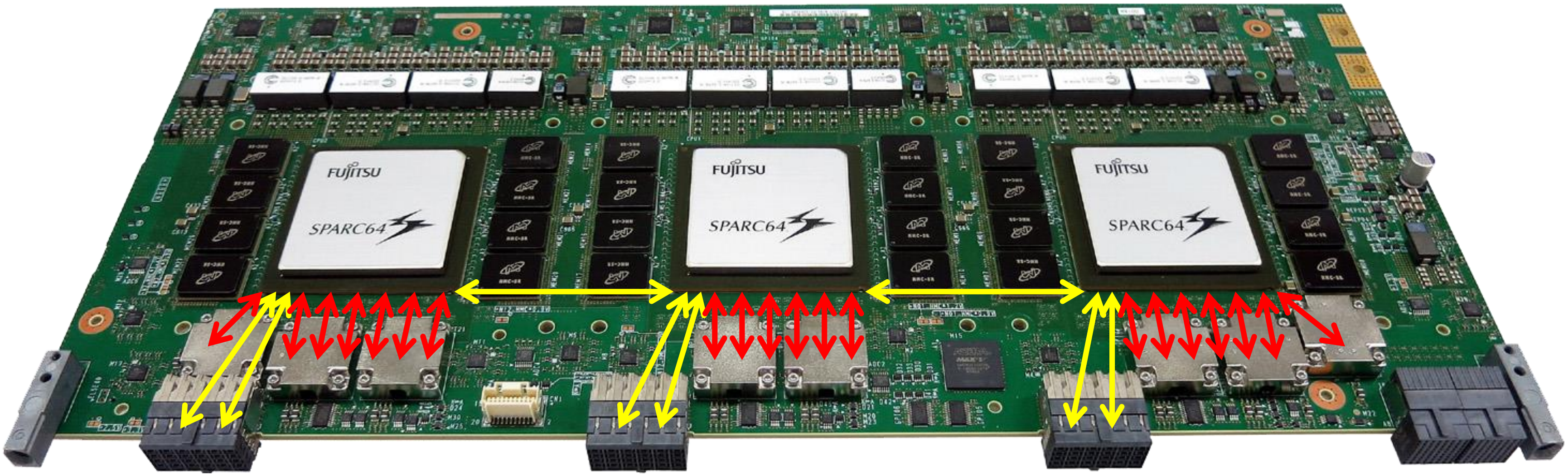
- Introduction
- Network topology
- SoC integration
- Improved link speed
- New efficient functions
- Summary

Summary of Tofu Interconnect 2

	Tofu1	Tofu2
Network topology	6D-Mesh/Torus	←
# of network interfaces	4	←
# of network links	10	←
Implementation	Discrete chip	SoC integration
Link speed	40 Gbps	100 Gbps
# of optical links	0	6 - 7
RDMA Put	✓	✓
RDMA Get	✓	✓
RDMA Atomic RMW		✓
Memory bypass (sender)	✓	✓
Memory bypass (receiver)		✓
Barrier synchronization	✓	✓
Non-blocking collective		✓

Transmission Technology

- Number of signals per link decreases from 8 to 4 lanes
  - Tofu1: 8 lanes × 18 ports = 144 lanes
  - Tofu2: 4 lanes × 10 ports = 40 lanes (limited pin-count)
- Link speed increases from 40 Gbps to 100 Gbps
  - By increasing data transfer rate fourfold from 6.25 to 25.78125 Gbps
- 2/3 of the links are optical
  - 1 out of 3 nodes uses **6 optical links** + **4 electrical links**
  - 2 out of 3 nodes use **7 optical links** + **3 electrical links**



Index

- Introduction
- Network topology
- SoC integration
- Improved link speed
- New efficient functions
- Summary

Summary of Tofu Interconnect 2

FUJITSU

	Tofu1	Tofu2
Network topology	6D-Mesh/Torus	←
# of network interfaces	4	←
# of network links	10	←
Implementation	Discrete chip	SoC integration
Link speed	40 Gbps	100 Gbps
# of optical links	0	6 - 7
RDMA Put	✓	✓
RDMA Get	✓	✓
RDMA Atomic RMW		✓
Memory bypass (sender)	✓	✓
Memory bypass (receiver)		✓
Barrier synchronization	✓	✓
Non-blocking collective		✓

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 each network interface
  - Strong atomicity: Any memory accesses cannot break atomicity
  - Mutual atomicity: Atomic operations of processor and Tofu2 mutually guarantee their atomicity

- The mutual atomicity enables an efficient implementation of unified multi-process and multi-thread runtime

Summary of Tofu Interconnect 2

	Tofu1	Tofu2
Network topology	6D-Mesh/Torus	←
# of network interfaces	4	←
# of network links	10	←
Implementation	Discrete chip	SoC integration
Link speed	40 Gbps	100 Gbps
# of optical links	0	6 - 7
RDMA Put	✓	✓
RDMA Get	✓	✓
RDMA Atomic RMW		✓
Memory bypass (sender)	✓	✓
Memory bypass (receiver)		✓
Barrier synchronization	✓	✓
Non-blocking collective		✓

Cache Injection

- Injecting received data into L2 cache directly
  - Bypassing main memory
- Injection flag On/Off is indicated by the sender
- Reduction of communication latency
  - The evaluations used the Verilog RTL codes for the production
  - Communication pattern: Ping-Pong of Put transfer

Injection flag	Estimated half round-trip latency
Off	0.87 usec
On	0.71 usec

0.16 usec reduction

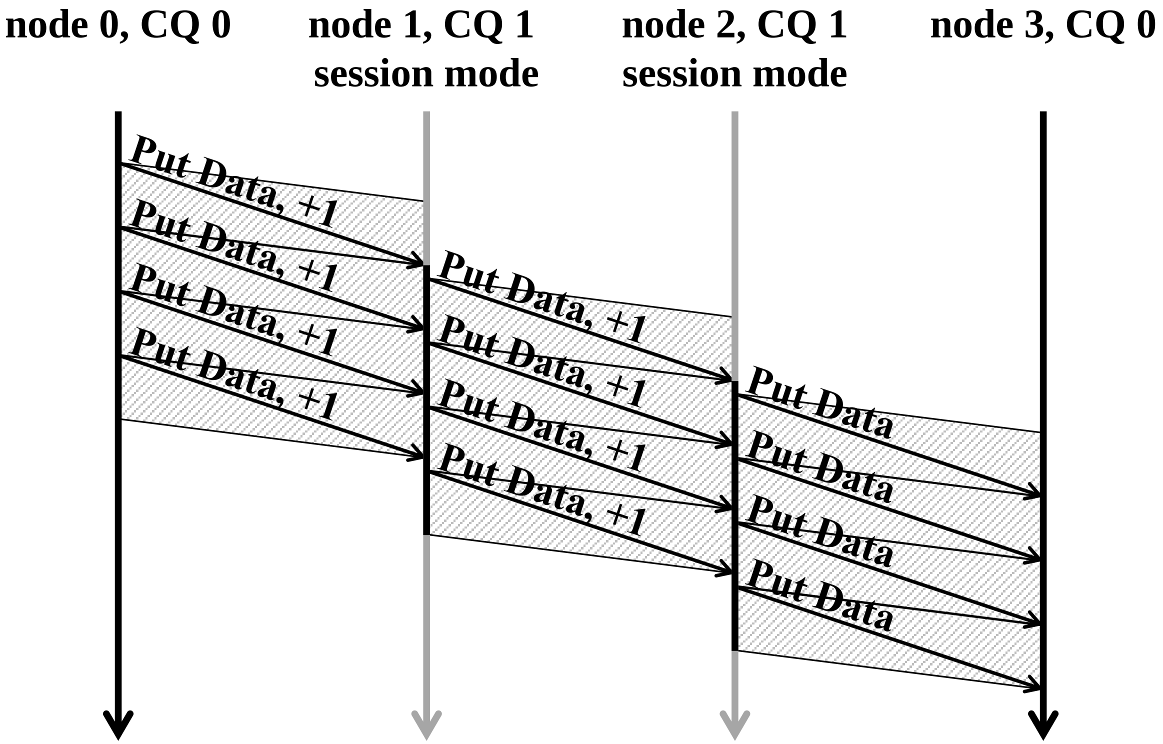
- Harmless injection
  - Cache injection is only performed when cache hits and the line is in exclusive state
  - A cache line in exclusive state is highly likely to be polled by a corresponding processor core

Summary of Tofu Interconnect 2

	Tofu1	Tofu2
Network topology	6D-Mesh/Torus	←
# of network interfaces	4	←
# of network links	10	←
Implementation	Discrete chip	SoC integration
Link speed	40 Gbps	100 Gbps
# of optical links	0	6 - 7
RDMA Put	✓	✓
RDMA Get	✓	✓
RDMA Atomic RMW		✓
Memory bypass (sender)	✓	✓
Memory bypass (receiver)		✓
Barrier synchronization	✓	✓
Non-blocking collective		✓

Session-mode Control Queue (CQ)

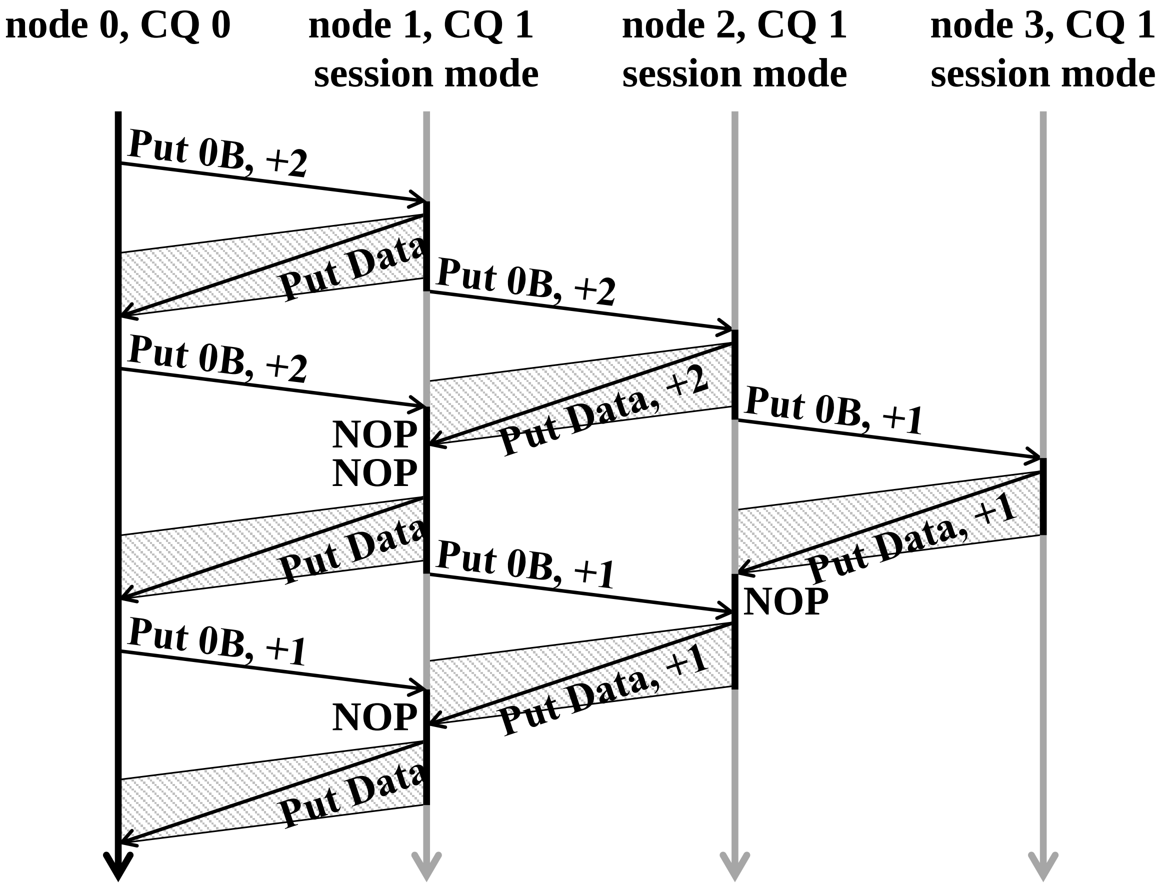
- Offloading a collective communication of long messages
- Command execution in a session-mode CQ is only advanced on a successful reception of Put transfer



Example of offloading pipelined Broadcast communication

Flexibility of Session-mode CQ

- Control flow can be branched or joined
  - Branch by advancing multiple commands
  - Join by enqueuing no operation commands



Example of offloading handshaking Gather communication

Index

- Introduction
- Network topology
- SoC integration
- Improved link speed
- New efficient functions
- Summary

Summary

- Introduced Tofu interconnect 2
  - Designed for the next generation Post-FX10 machine
- System-on-chip integration
  - The number of link ports per node decreases from 18 to 10
- Link speed increases from 40 Gbps to 100 Gbps
  - 2/3 of the links are optical
- New efficient functions
  - Tofu2’s atomic RMW operations guarantee the mutual atomicity that enables to unify multi-thread and multi-process shared variables
  - The cache Injection function bypasses main memory on a receiver side and reduces communication latency by 0.16 usec without cache pollution
  - The flexibility of session-mode CQ enables offloading various non-blocking collective communication algorithms

FUJITSU

shaping tomorrow with you