ICC: An Interconnect Controller
for the Tofu Interconnect Architecture

August 24, 2010

Takashi Toyoshima

Next Generation Technical Computing Unit
Fujitsu Limited

shaping tomorrow with you

Background

Requirements for Supercomputing Systems

Low latency

- Communication latency limits the scalability of applications

High bandwidth

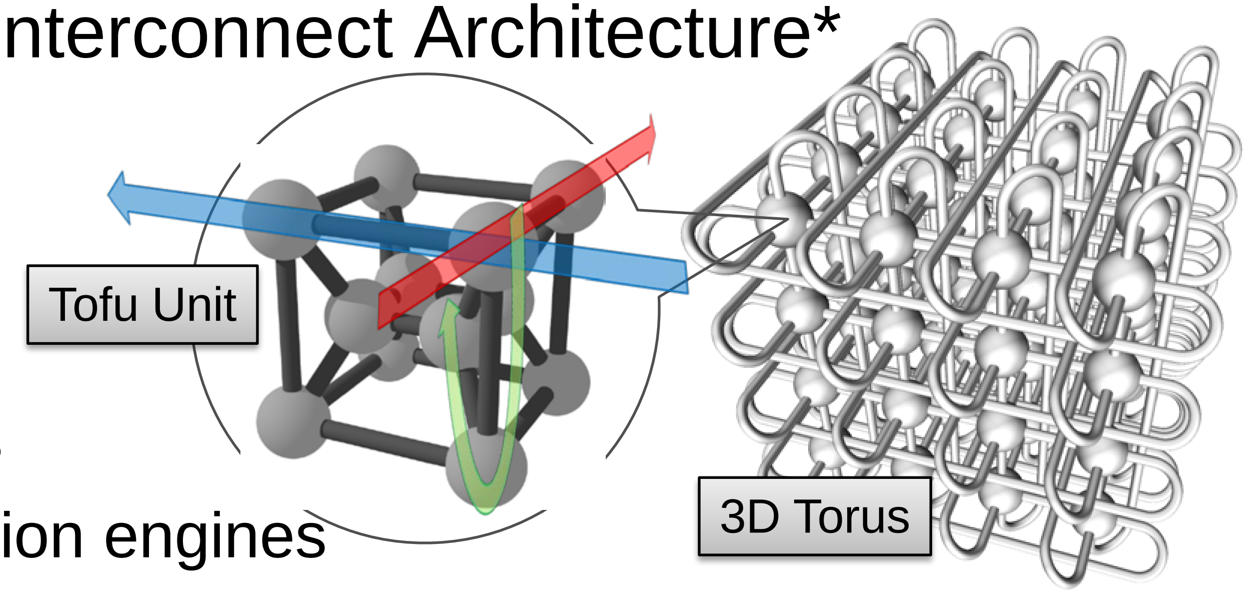
- Increasing calculation FLOPS requires higher network bandwidth be balanced with FLOPS

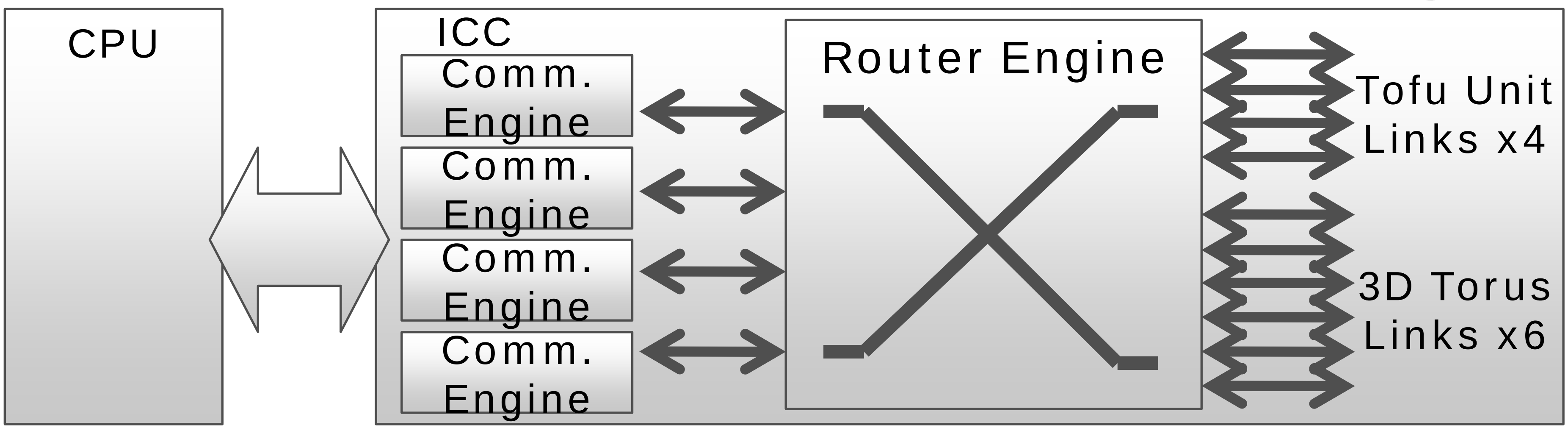
RAS – Reliability, Availability and Serviceability

- The risk of hardware faults in large systems increases along with the increased number of nodes

Fujitsu's New Interconnect Architecture

- 6D Mesh/Torus Interconnect Architecture*
  - Scalability
  - Fault-tolerance
- LSI Features
  - Ten network links
  - Four communication engines





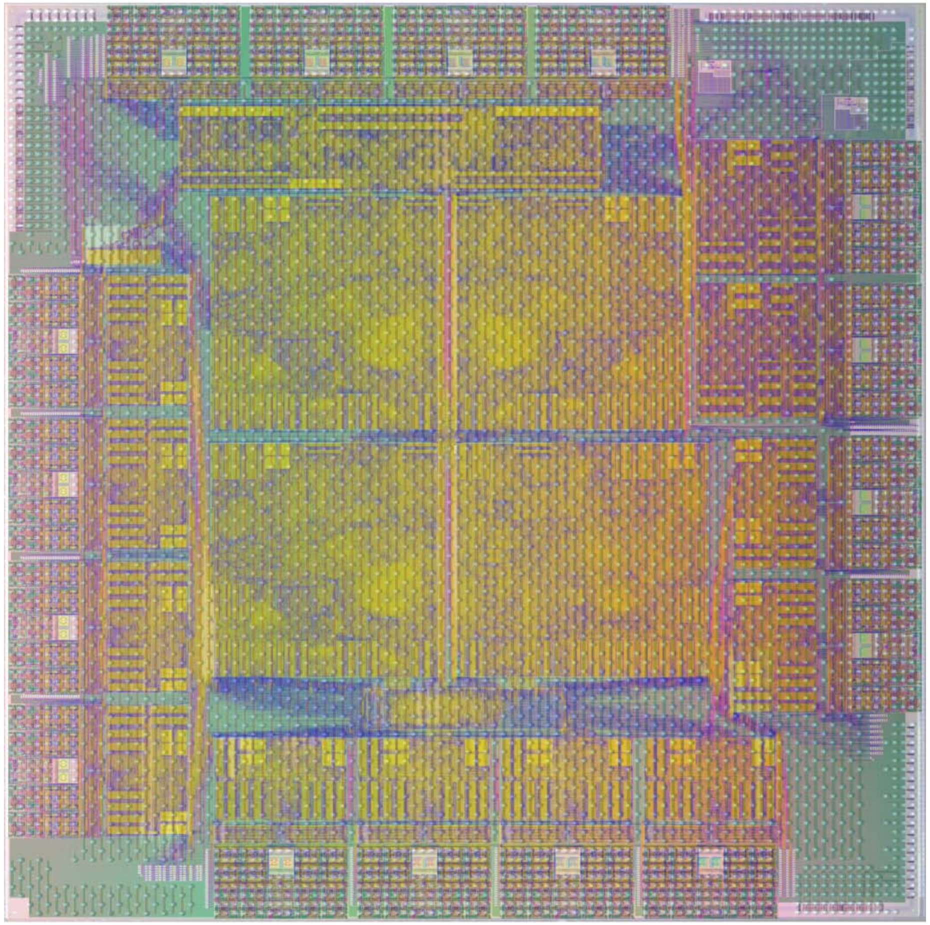
(*) "Tofu: A 6D Mesh/Torus Interconnect for Exascale Computers", IEEE Computer, vol.42, no.11, Yuichiro Ajima, Shinji Sumimoto, Toshiyuki Shimizu

Implementation

- Implementation
- Features
  - Overview
  - Interface features for latency and throughput
  - Network features for network utilization
- Conclusion

Specifications

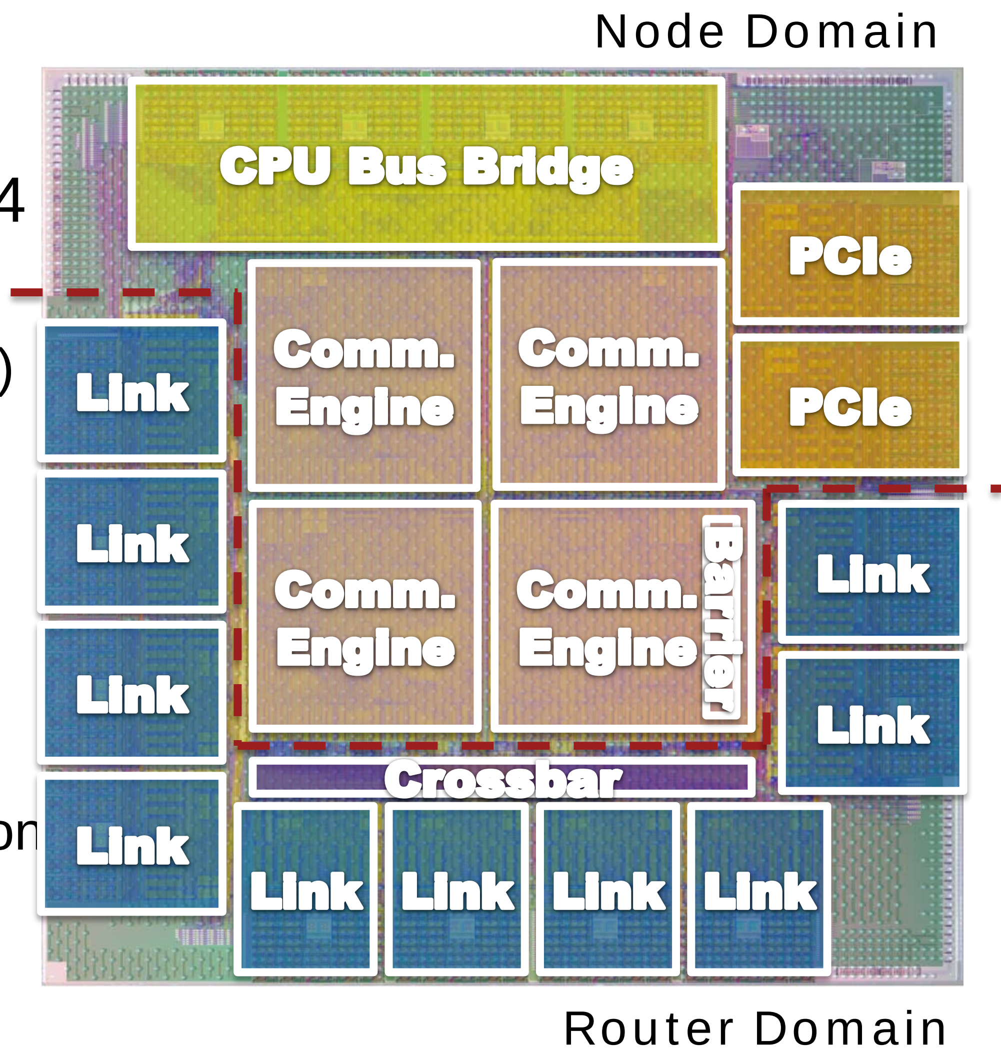
- Fujitsu's 65nm CMOS Technology
  - Die size
    - 18.2mm × 18.1mm
  - Transistors
    - 48M gates for logic
    - 12M-bit SRAM cells
  - I/O
    - 5GB/s Ports × 16
      - 6.25Gb/s × 8 links / port
  - Misc.
    - ASIC design flow
    - 312.5MHz/625.0MHz



Floor Plan

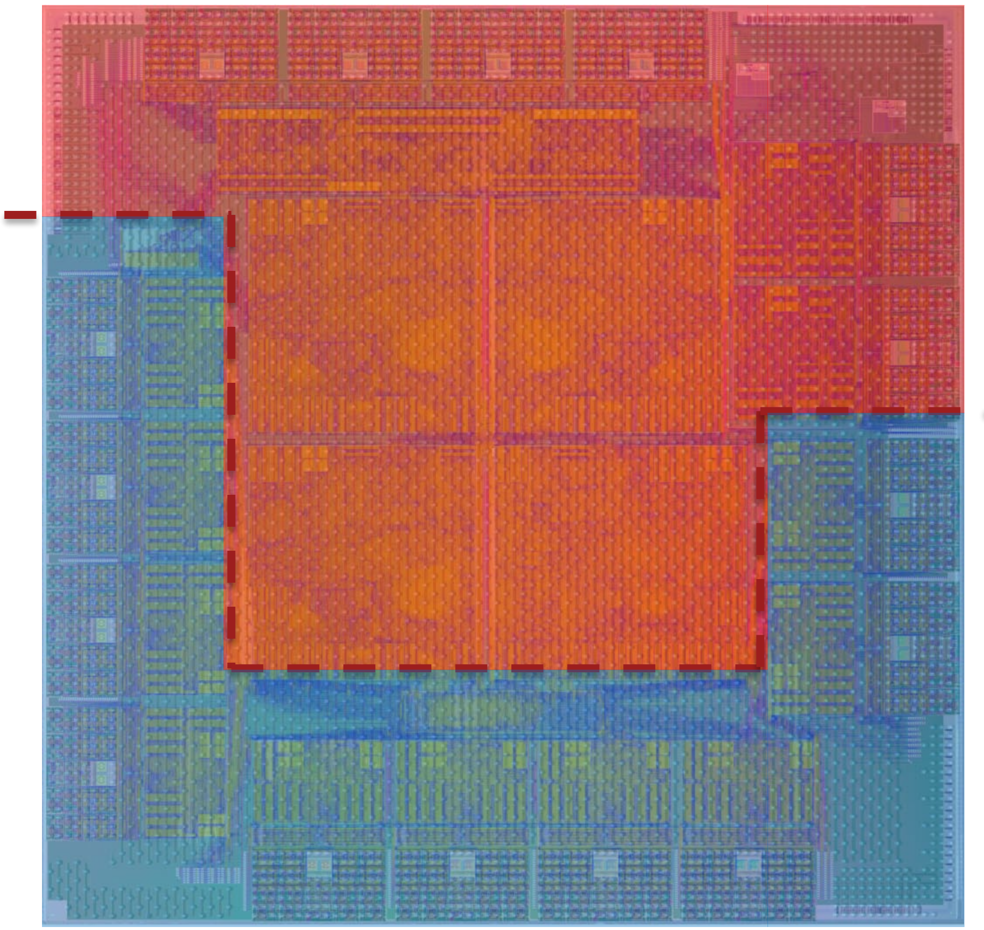
FUJITSU

- Node Domain
  - CPU bus bridge
    - 20GB/s in each direction
  - Communication engines × 4
    - 5GB/s in each direction
    - Barrier engine (Comm.#0 only)
  - PCIe 2.0 root complex × 2
    - Isolated power domain
- Router Domain
  - Crossbar
    - 14 ports 5GB/s in each direction
  - Link ports × 10
    - 5GB/s in each direction



RAS Features

- Fault Domain Isolation
  - Router continues to work on node faults
- Error Protection
  - Radiation-hardened FFs
  - ECC protection
    - RAM/Data path
  - Parity error detection
    - Control path
  - CRC protection
    - Data link/Transaction



Features
Overview

- Implementation
- Features
  - Overview
  - Interface features for latency and throughput
  - Network features for network utilization
- Conclusion

ICC Features

FUJITSU

	Latency	Throughput	RAS
System	✓Many Neighbors ✓Hop Reduction ✓3D Torus View ☆	✓Many Neighbors ✓Trunking ☆	✓Detour Path ☆ ✓Subnet Partitioning ☆
Network Interface	✓RDMA - Quick Start ☆ - Piggyback - Strong Order ☆ ✓Stream Offload ☆ ✓Barrier Engine ☆	✓GAP Control ☆ ✓Multi-Interfaces ☆ - User Thread x2 - Kernel Thread	✓Radiation-hardened FF ✓ECC ✓Parity ✓CRC
Router Engine	✓Cut-through ✓Grant Prediction ☆ ✓Straight Bypath ☆	✓Straight Bypath ☆ ✓New VC Scheduling ☆	✓Node Error Isolation ☆ ✓Radiation-hardened FF ✓ECC ✓Parity ✓CRC

☆: Unique Features

Today's topics are highlighted in red

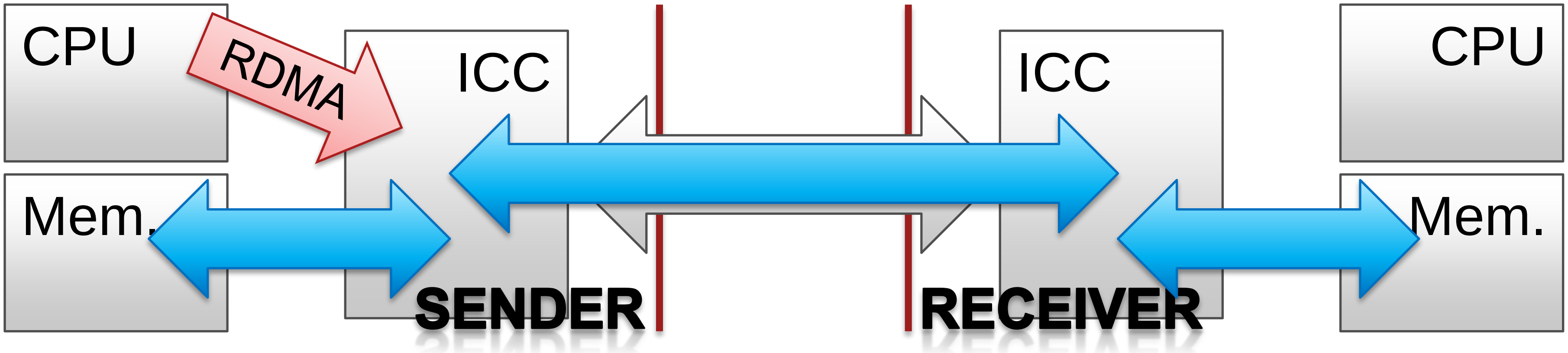
Features
Interface features

- Implementation
- Features
  - Overview
  - Interface features for latency and throughput
  - Network features for network utilization
- Conclusion

RDMA: Remote Direct Memory Access

FUJITSU

- Features



Operation	READ (Get) / WRITE (Put)
Length	~16MB
MTU	256B~1920B
Virtual Address	Support (64K set)

- Low Latency and High Throughput
  - Command supply throughput and latency
  - Out-of-ordered I/O memory bus

RDMA Optimizations

FUJITSU

- Sender Techniques
  - Direct descriptor
    - Quick command supply
  - Piggyback
    - Command embedded communication payload
    - Short message sending without any DMA

	Throughput	Latency
PIO		✓Good
DMA	✓Good	

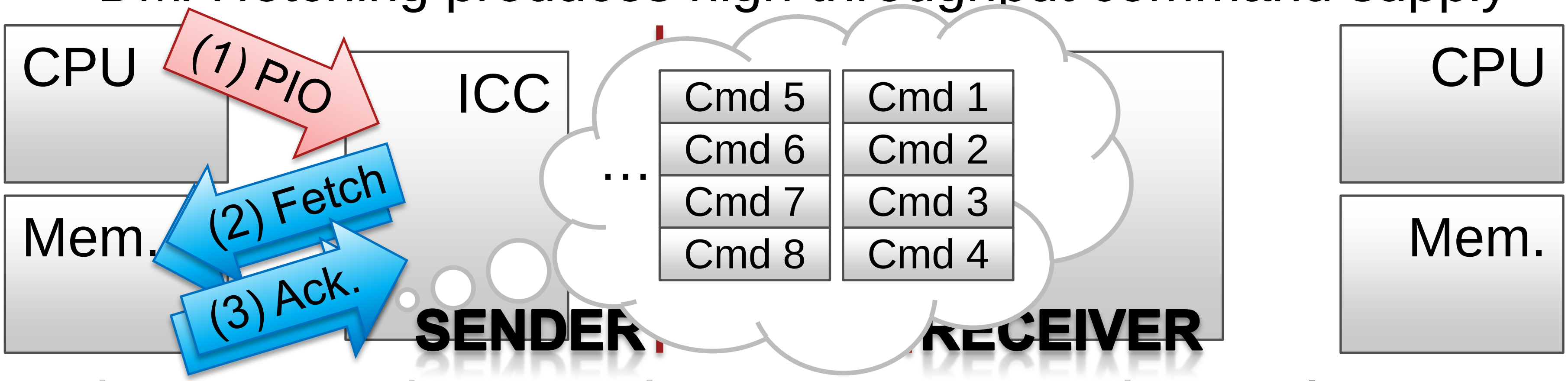
Command Supply Performance

- Receiver Techniques
  - Out-of-ordered I/O memory bus
    - High throughput bus transaction
  - Strong ordered store
    - In order completion of DMA transactions for buffer polling

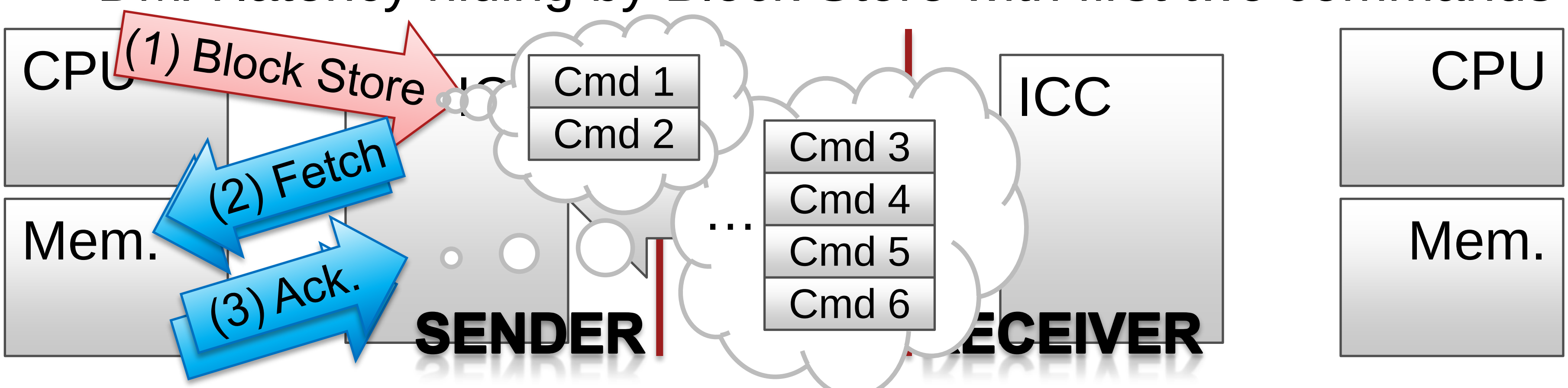
Direct Descriptor Feature

FUJITSU

- Normal Command Supply
  - DMA fetching produces high throughput command supply

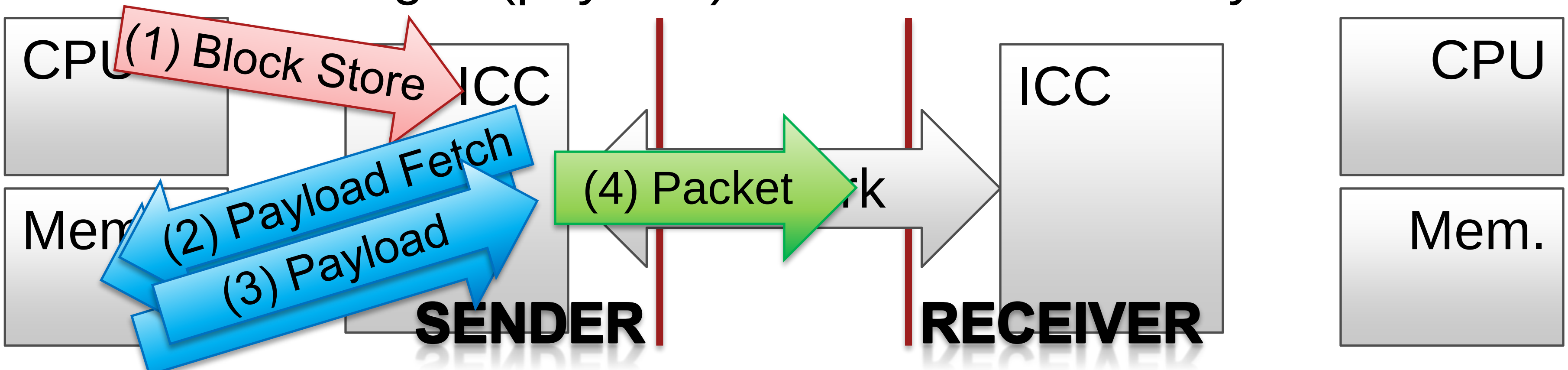


- Direct Descriptor and DMA Command Supply
  - DMA latency hiding by Block Store with first two commands

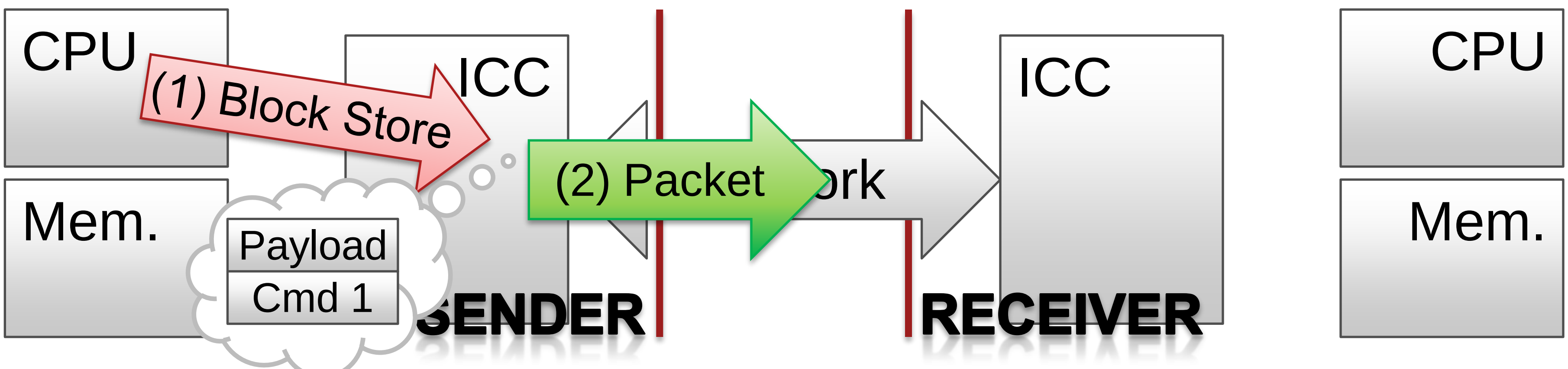


Piggyback Feature

- Normal Payload Supply
  - User messages (payload) should be fetched by DMA



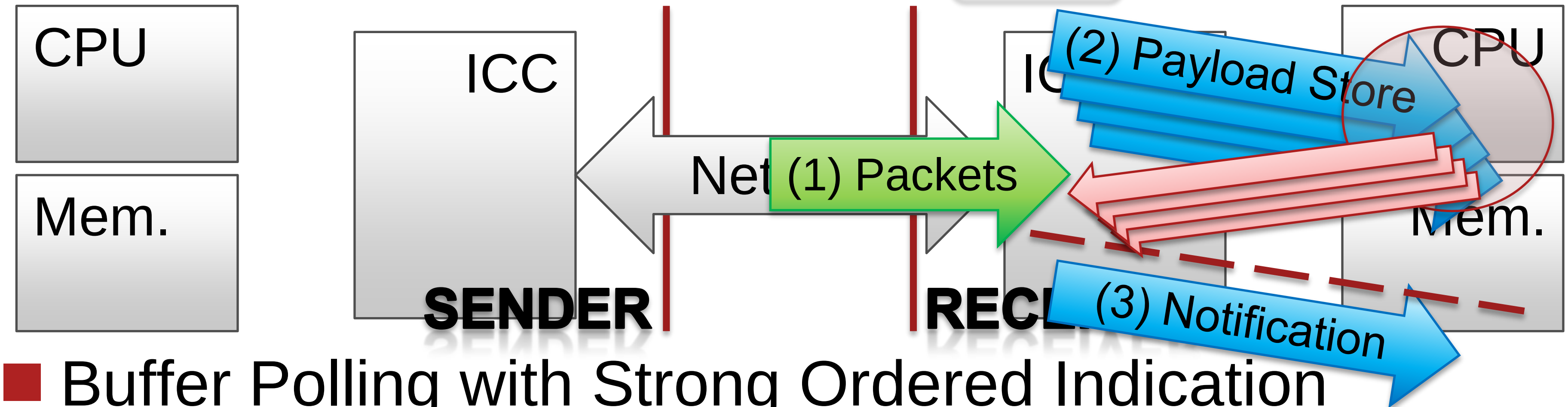
- Piggyback Payload Supply
  - User messages (payload) are embedded in commands



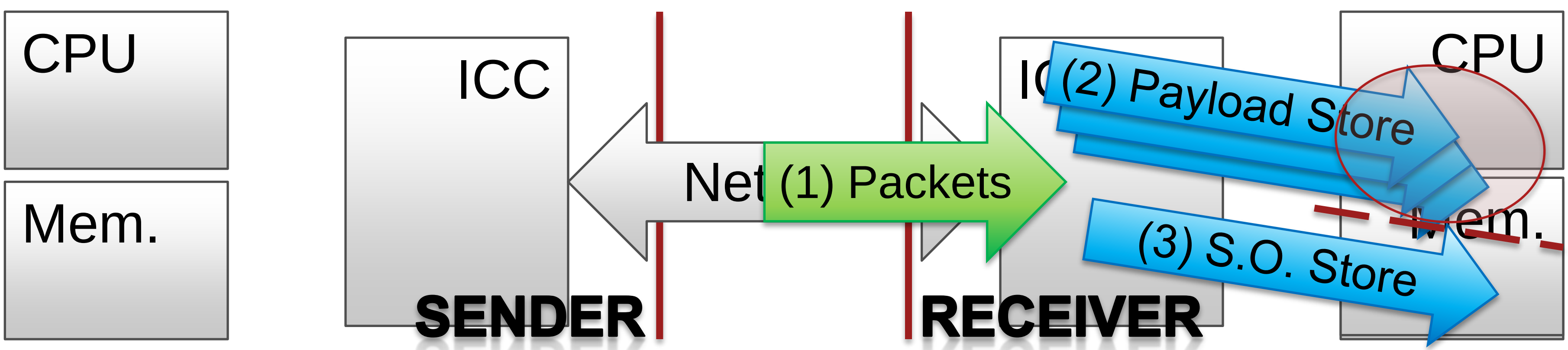
Out-of-Ordered I/O Memory Bus

FUJITSU

- Completion Notification Polling
  - ICC notifies after the completion of O-o-O DMA Stores

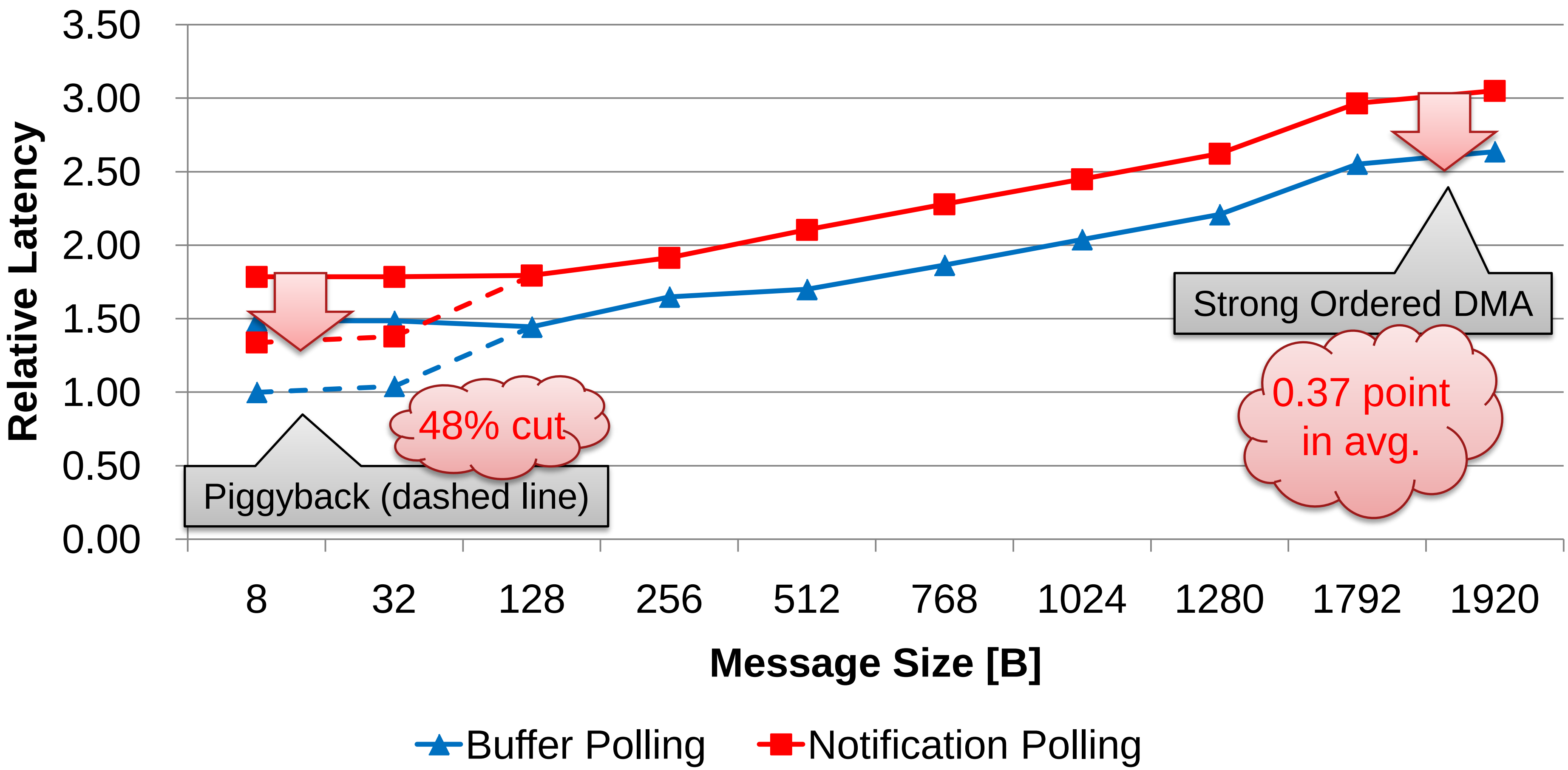


- Buffer Polling with Strong Ordered Indication
  - Memory controller guarantees specified DMA ordering



RDMA Performance

- Hardware Measured Results
  - Piggyback achieves low latency in short message
  - Strong ordered packet makes buffer polling possible

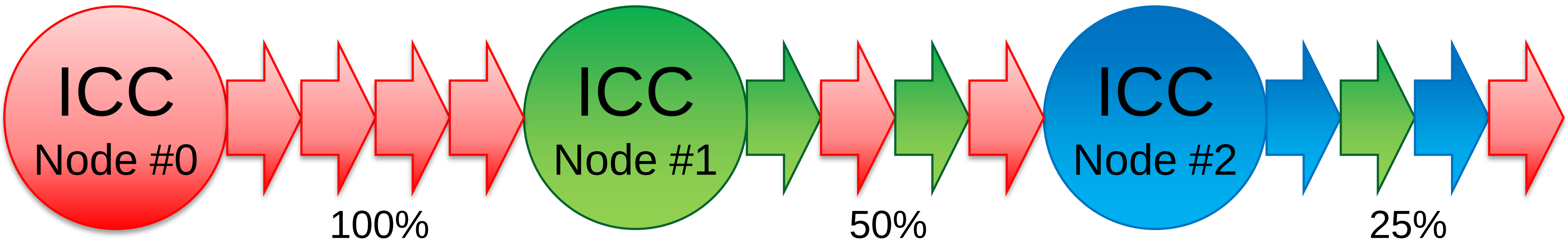


Features
Network Features

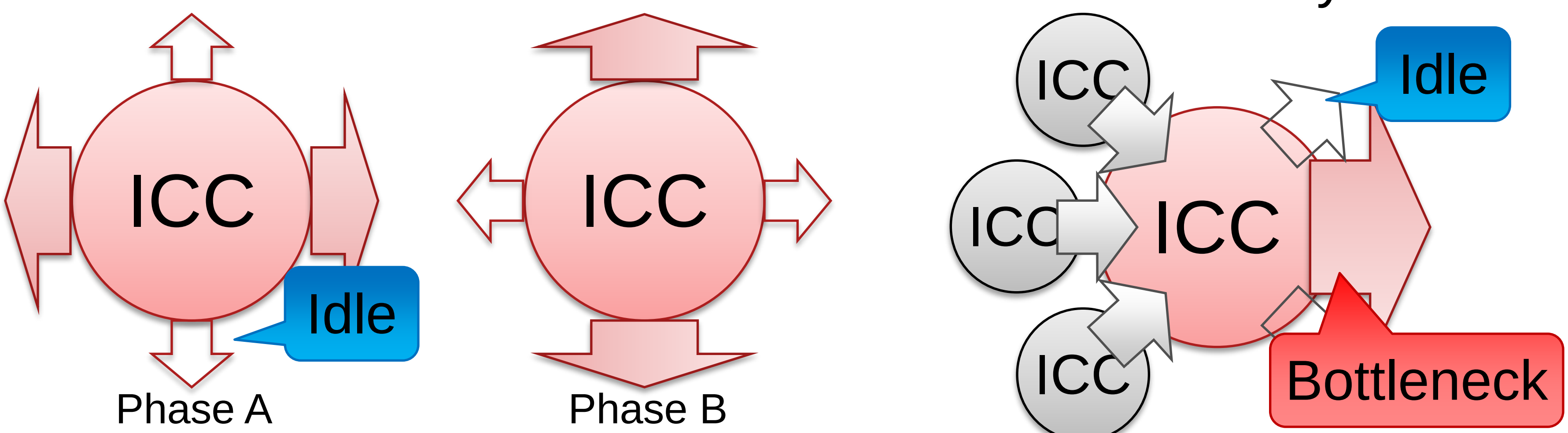
- Implementation
- Features
  - Overview
  - Interface features for latency and throughput
  - Network features for network utilization
- Conclusion

Network Utilization Problems

- Global Unfairness of Throughput
  - Arbitrations with local fairness cause global unfairness

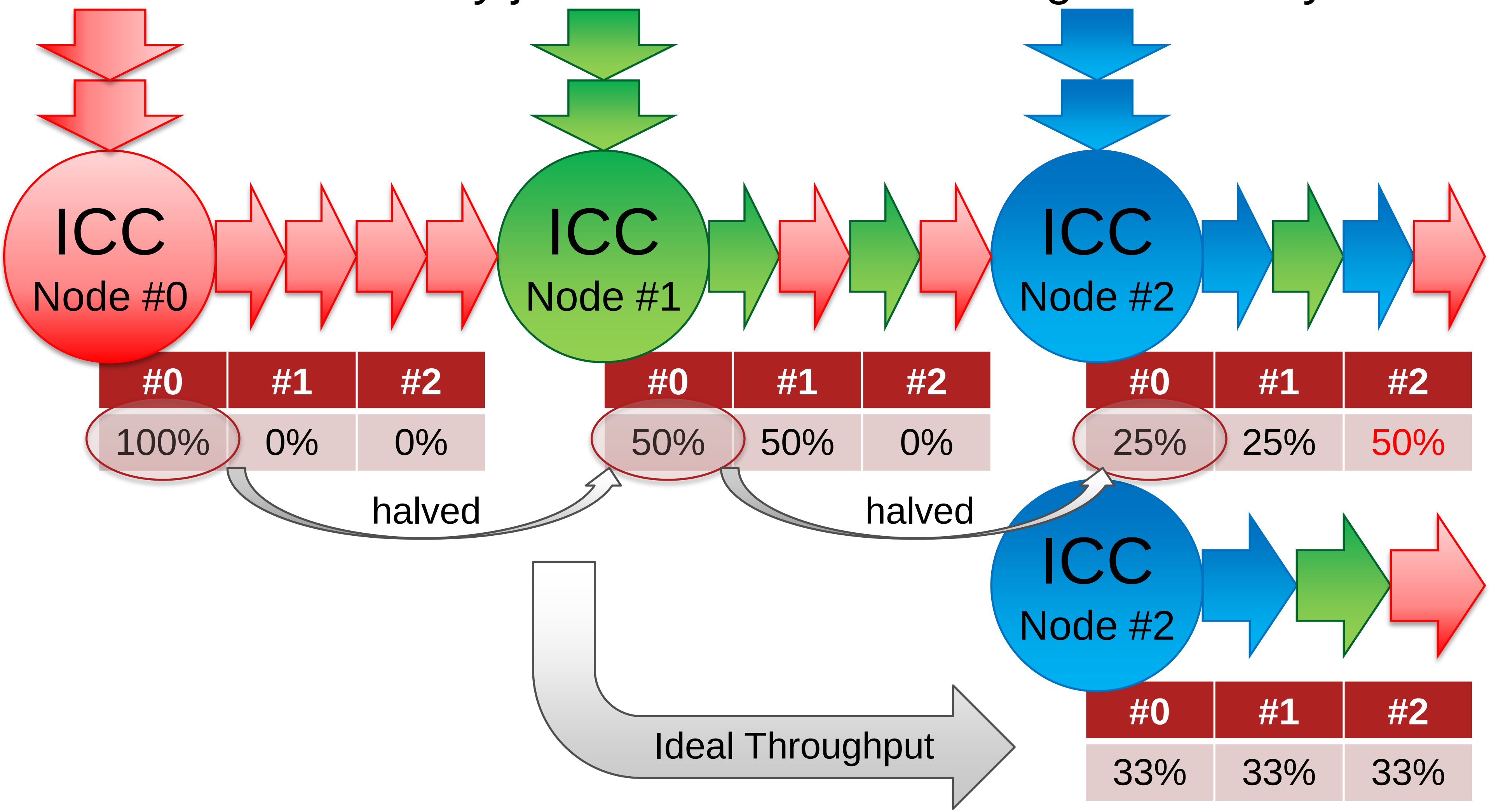


- Non-uniform Application Traffic in Time and Space
  - Bandwidth of idle links needs to be used effectively



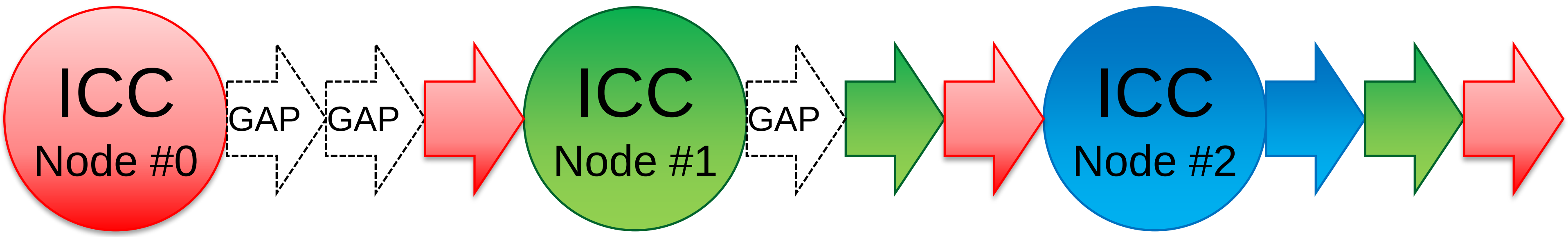
Global Unfairness of Throughput

- Local Fairness of arbitration cause global unfairness
  - Arbiters on every junction treat all incoming traffic fairly



Injection Rate Control

- Software Specify the Inter-Packet GAP Parameter

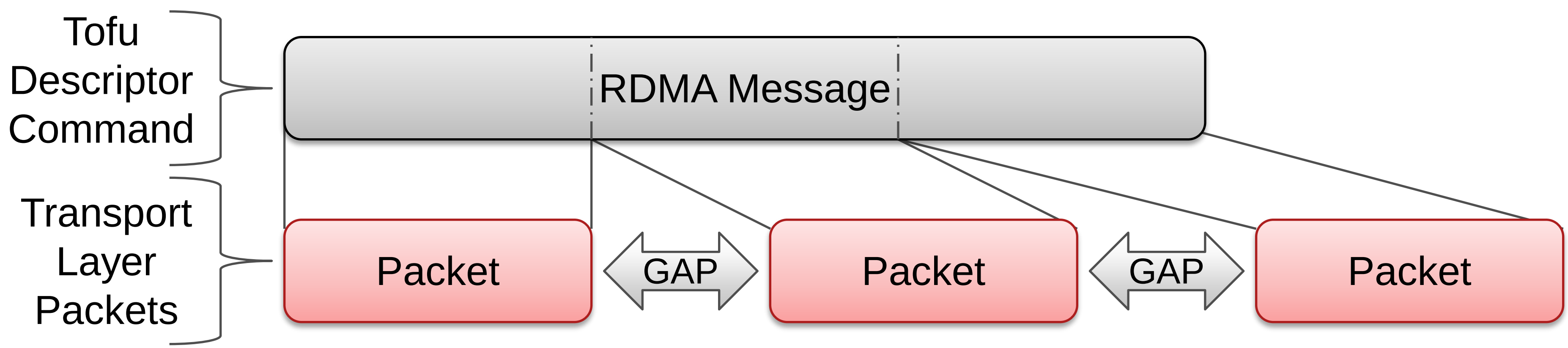


#0	#1	#2
33%	0%	0%

#0	#1	#2
33%	33%	0%

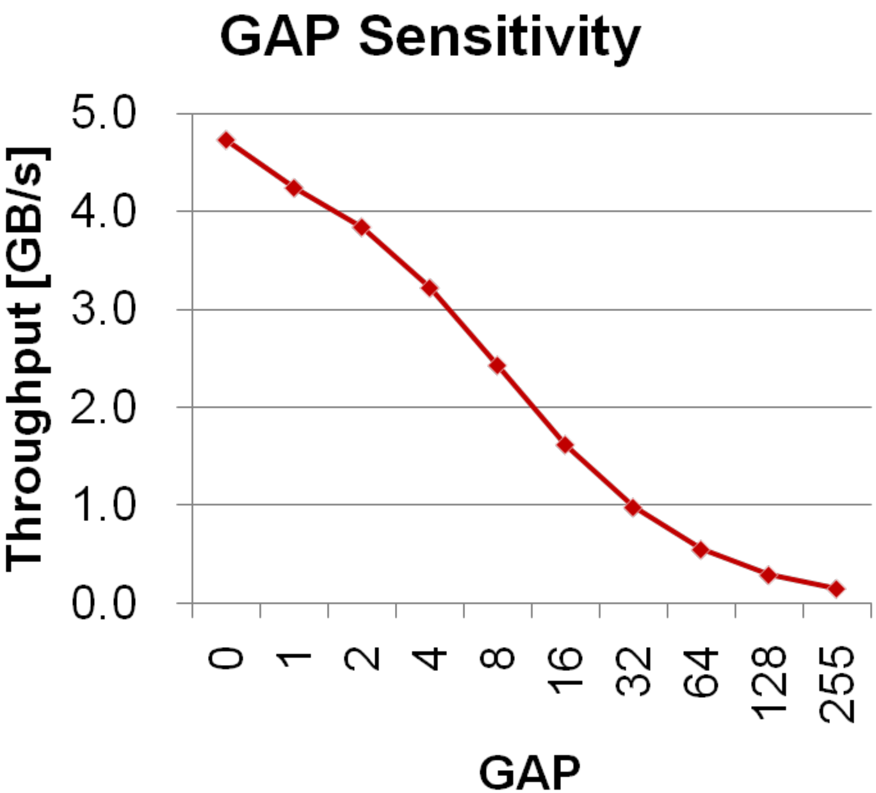
#0	#1	#2
33%	33%	33%

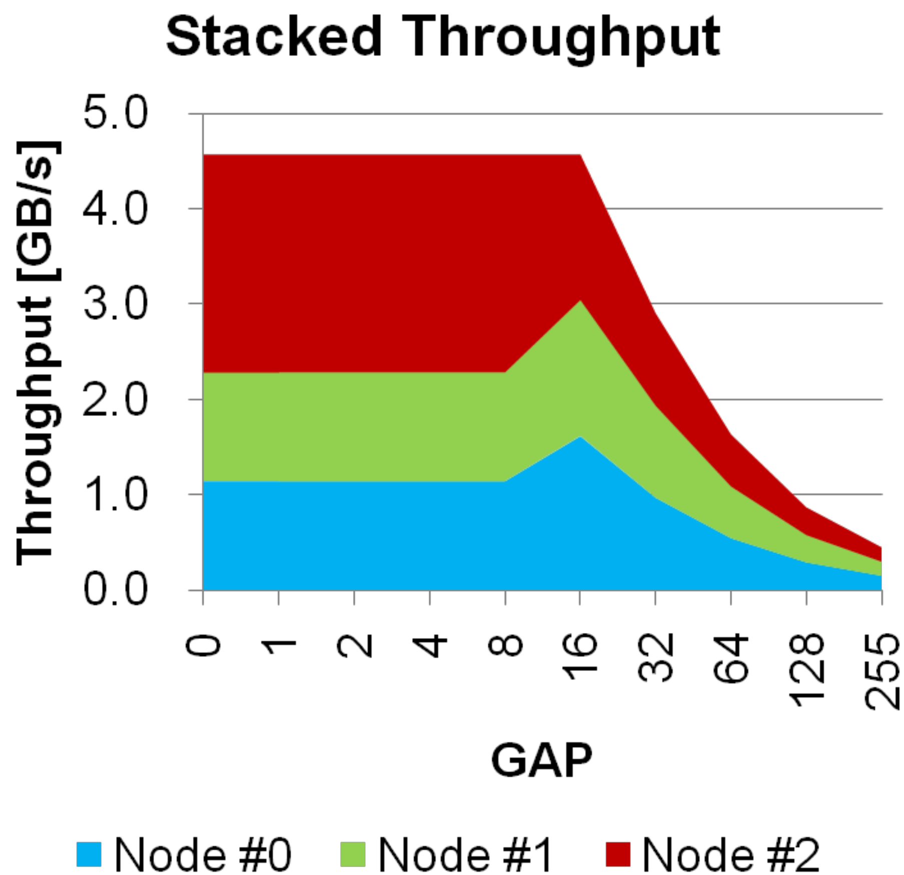
- Communication engine works to control injection rate
  - Insert temporal gaps between transmitting packets
  - Interval can be specified by the user



Throughput Performance

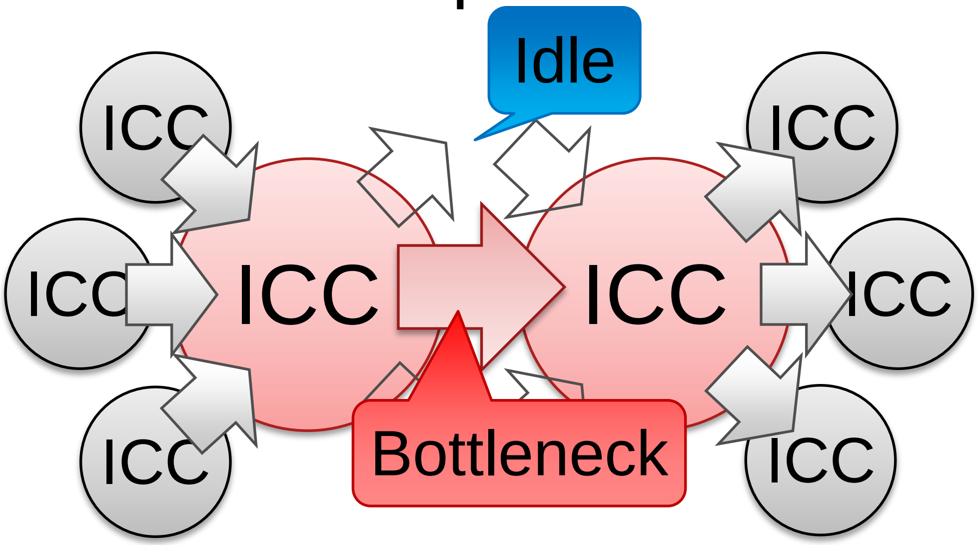
- Hardware Measured Results
  - Software can specify fine grained GAP parameters: 0-255
  - GAP works to control throughput effectively



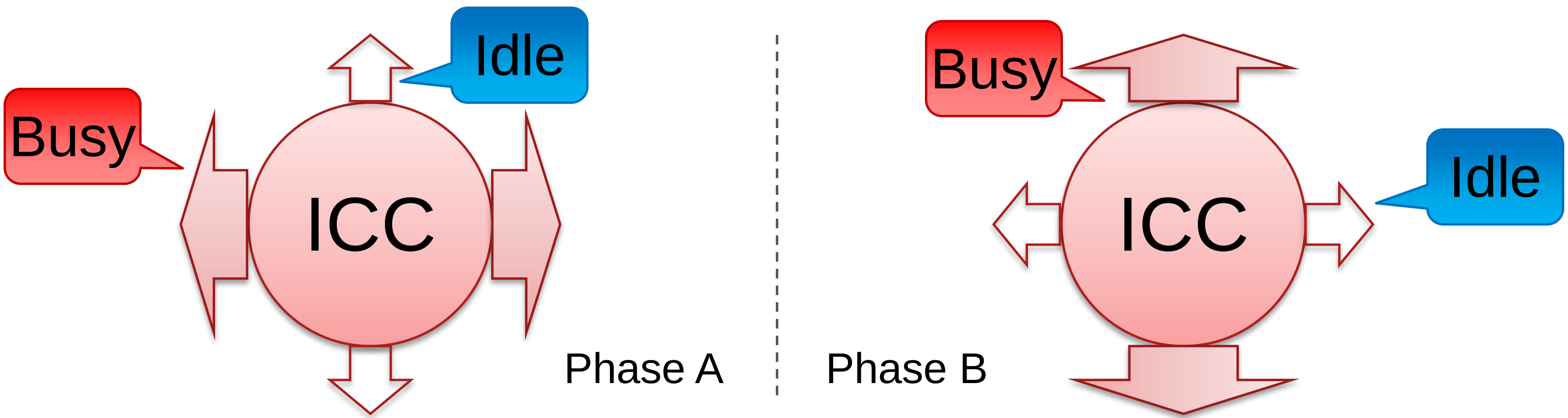


Non-uniform Application Traffic

- Non-uniform Traffic in Space

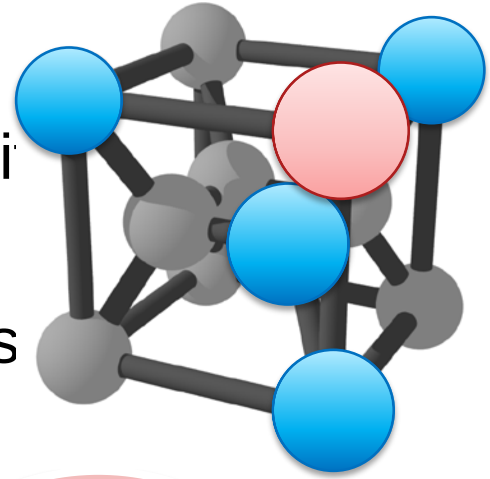


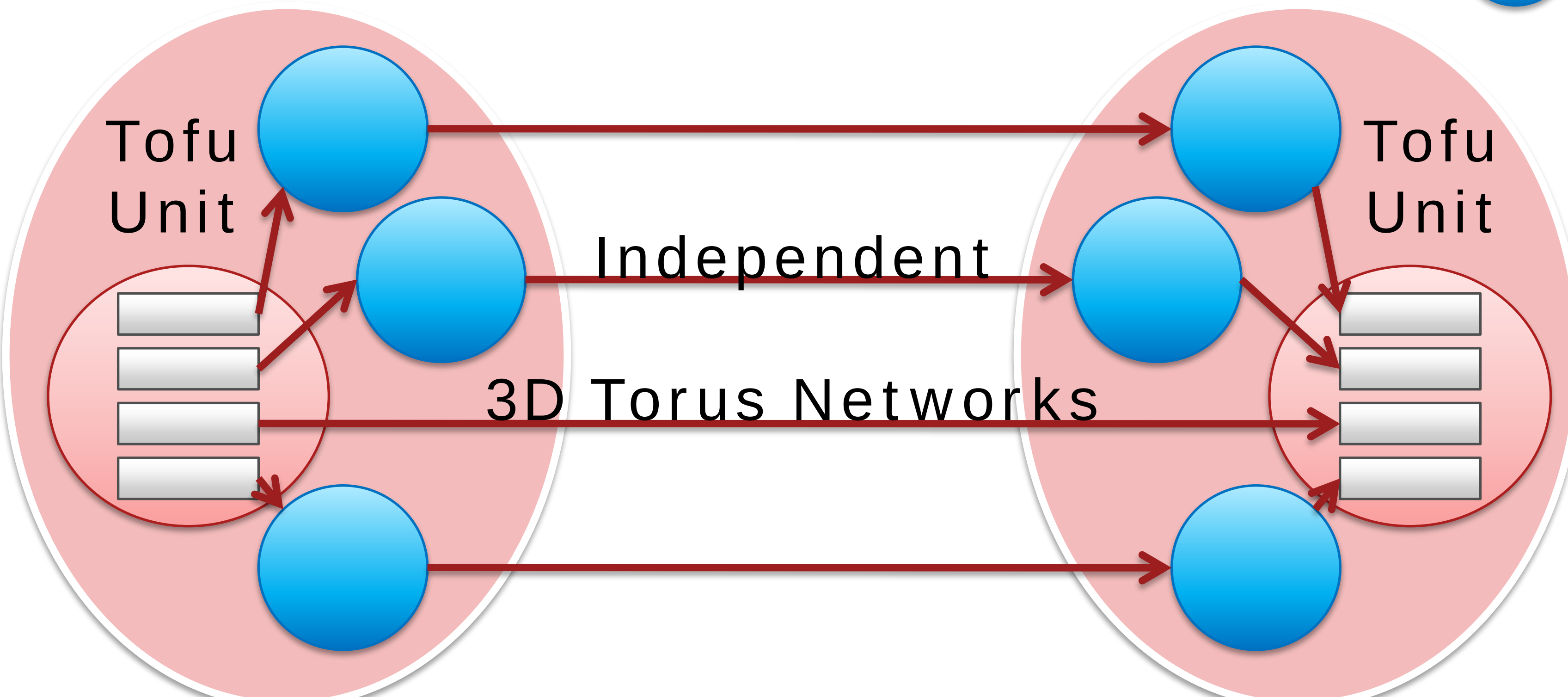
- Non-uniform Traffic in Time



Trunking Communication

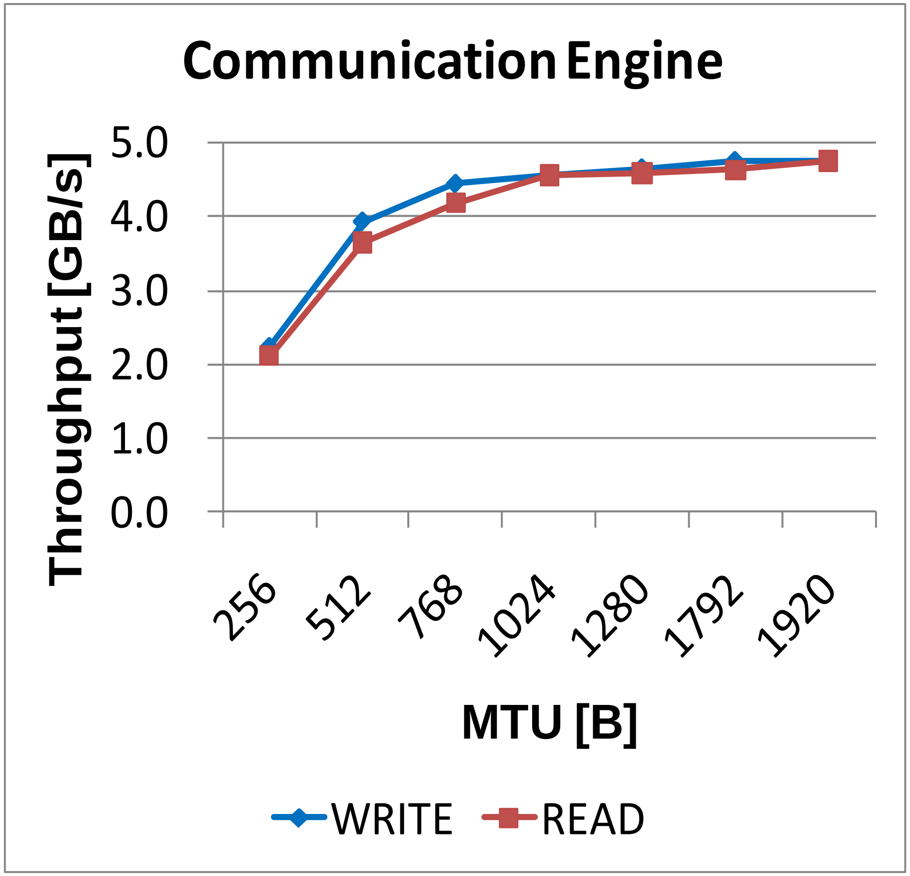
- Trunking Independent Idle Paths
  - Nodes have four neighborhoods in Tofu Unit
    - Independent links and 3D-Torus networks
  - Each node has four communication engines
    - Up to × 4 throughput

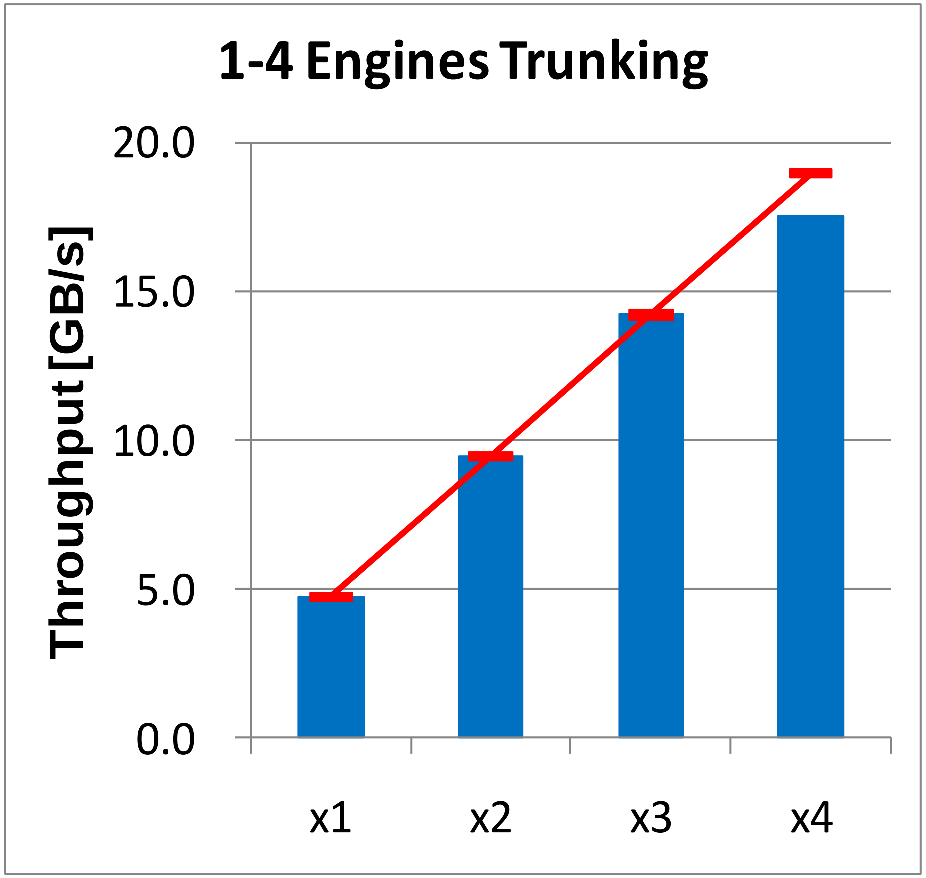




Trunking Performance

- Results
  - Communication engines achieve good performance
  - Trunking mechanisms scale up to four engines



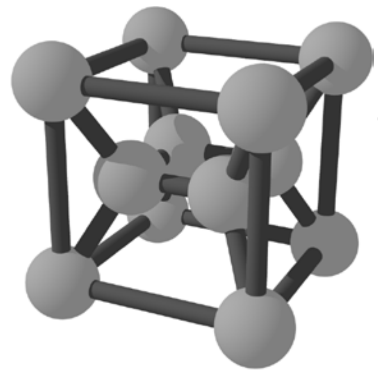


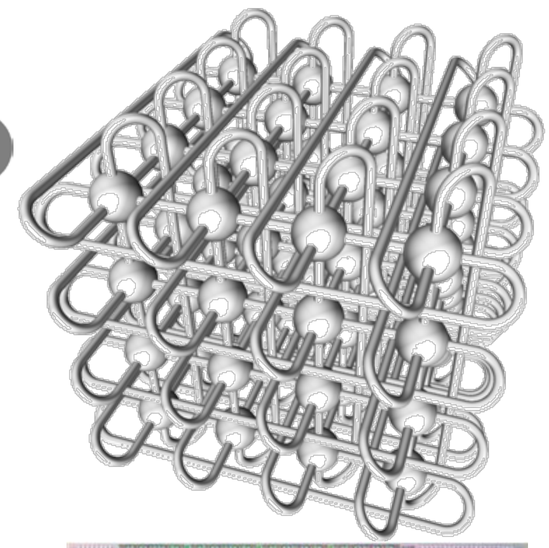
Conclusion

- Implementation
- Features
  - Overview
  - Interface features for latency and throughput
  - Network features for network utilization
- Conclusion

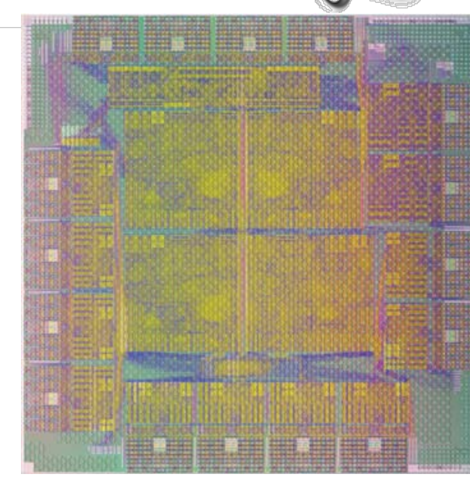
Concluding Remarks

- Tofu: A 6D mesh/torus interconnect architecture
  - Interconnect for Fujitsu's Peta/Exascale computing systems
  - Low latency, High bandwidth and RAS





- Features
  - High-throughput and low-latency RDMA
    - Direct Descriptor and Piggyback
    - Out of Ordered I/O Memory Bus
  - Network features for network utilization
    - Network injection rate control
    - Trunking up to four times throughput



Thank You for Listening

Thanks to…

Next Generation Technical Computing Unit

Aiichiro Inoue, Yuji Oinaga

Tofu Architecture Team

Toshiyuki Shimizu, Yuichiro Ajima, Tomohiro Inoue, Shinya Hiramoto

ICC Design Team

Takeo Asakawa, Akira Asato, Takumi Maruyama,
Koichiro Takayama, Koichi Yoshimi, Osamu Moriyama,
Masao Yoshikawa, Shinichi Iwasaki, Takekazu Tabata,
Yoshiro Ikeda, Yuzo Takagi,
Yoshihito Matsushita, Toshihiko Kodama, Satoshi Nakagawa,
Masato Inokai, Shigekatsu Sagi, Ikuto Hosokawa,
Yaroku Sugiyama, Takahide Yoshikawa

FUJITSU

shaping tomorrow with you