Empowered by Innovation NEC

Hot Interconnects 2014

End-to-End Adaptive Packet Aggregation for High-Throughput I/O Bus Network Using Ethernet

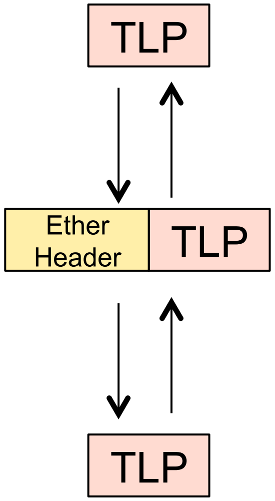
Green Platform Research Laboratories, NEC, Japan

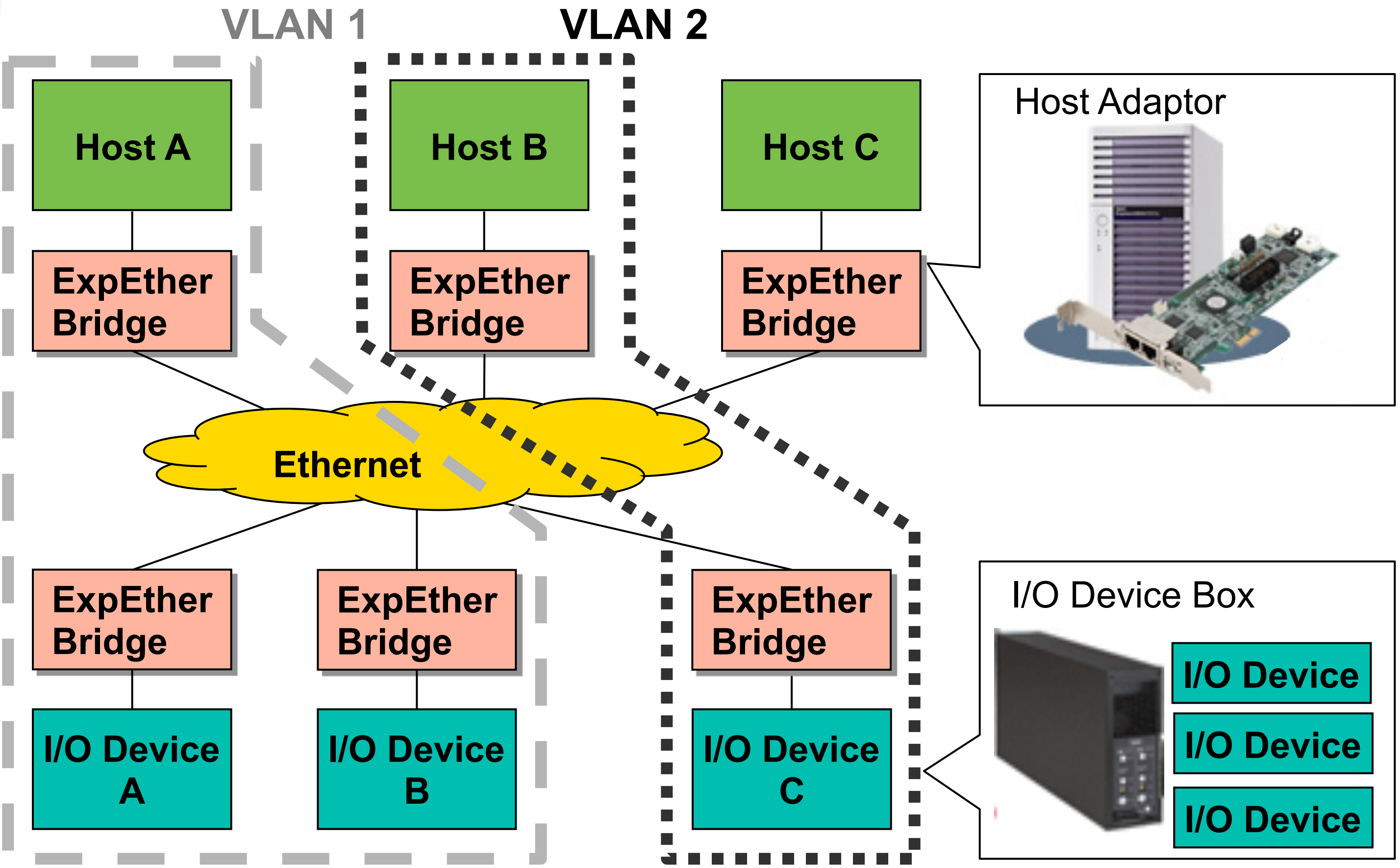
J. Suzuki, Y. Hayashi, M. Kan, S. Miyakawa, and T. Yoshikawa

Background: *ExpEther,* PCIe Interconnect Based on Ethernet

- Extension of PCI Express over Ethernet
- Encapsulation of PCIe packets (TLPs) into Ethernet frames
- Pros: Scalable connectivity, system reconfiguration, and separate implementation

TLP: PCIe Transaction Layer Packet





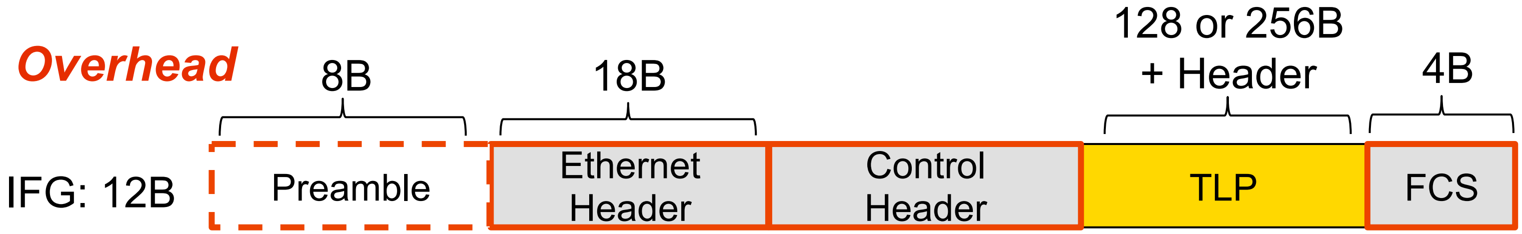
J. Suzuki *et al.*, 14th IEEE Symposium on High-Performance Interconnects, pp. 45-51, 2006.

Overhead of TLP Encapsulation

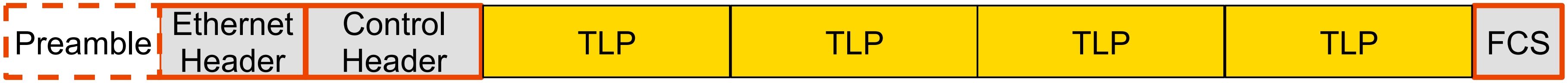
- Overhead of packet-by-packet encapsulation decreases PCIe throughput obtained through Ethernet connections

- Proposal: Aggregate multiple TLPs into single Ethernet frame

[Current System]



[Proposal]



Challenges of Packet Aggregation in PCIe over Ethernet

Low-latency is important and avoid additional delay

- PCIe traffic is sensitive to delay
- 1-us wait time (needs to aggregate TLPs by max Ethernet frame length) is large for short latency I/O device, e.g., PCIe-connected PCM

Needs to be End-to-End

- It is difficult to modify commercial Ethernet switches to aggregate TLPs

W/o modification of hosts' system stack

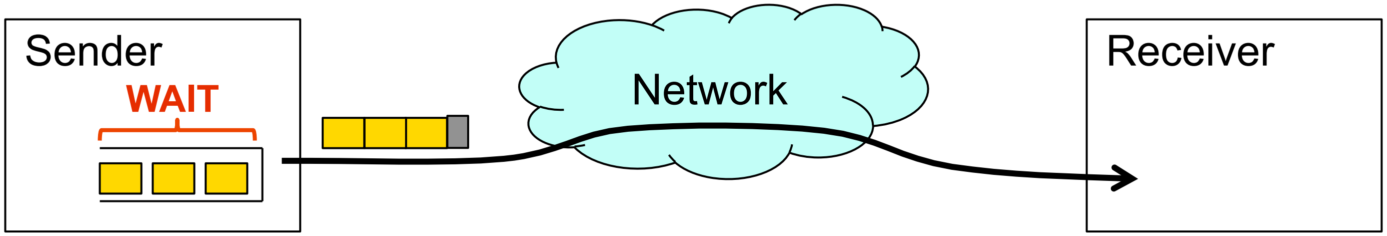
- Avoid modifying OS and device drivers

Related Work 1/2

- Previous work has been done in wireless and optical network
- They are categorized into two groups
  - *A. Jain, et al., PhD Thesis, Univ. of Colorado, 2003.*

[Method 1] Introduce wait time to aggregate packet with next one

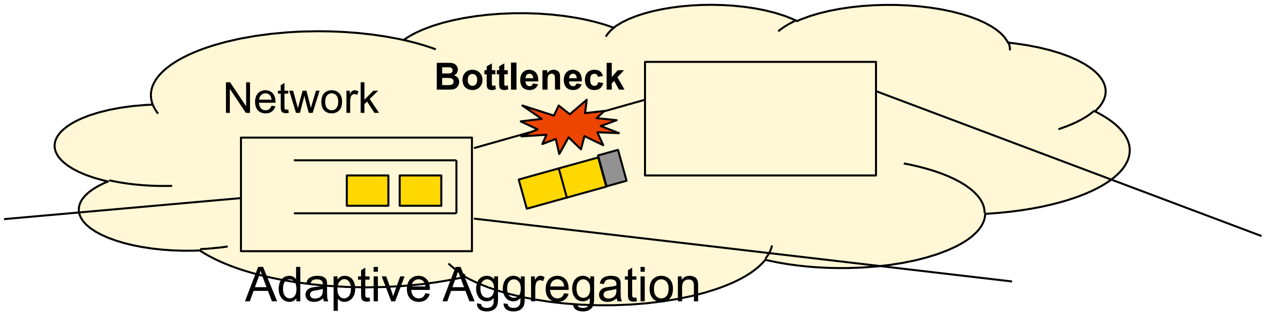
- End-to-end possible
- Increase transmission delay



Related Work 2/2

[Method 2] Adaptively aggregate packets if they are accumulated in queue (in network node adjacent to bottleneck link)

- Low-latency
- Hop-by-hop only and switch needs to be adapted



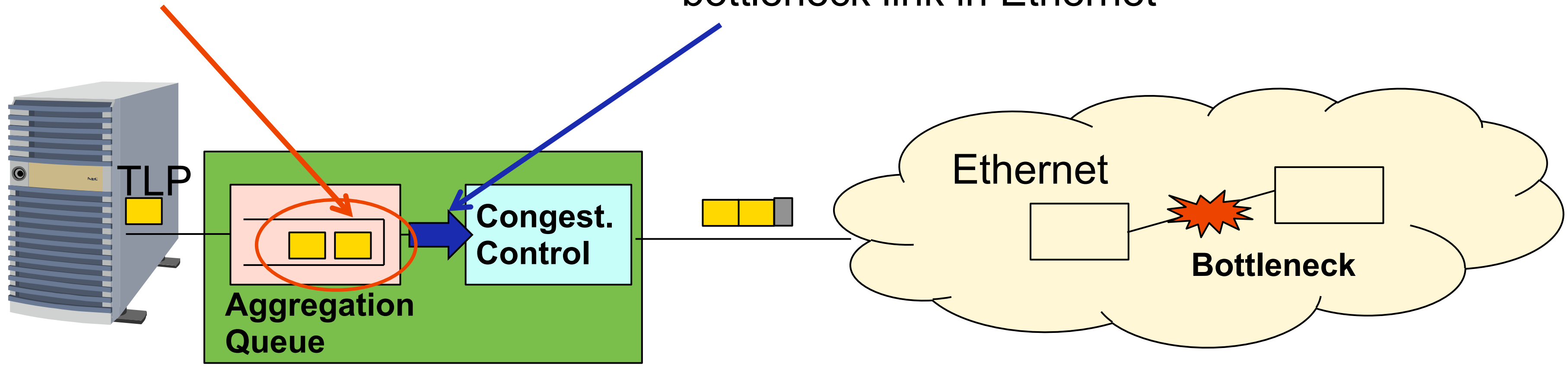
Proposed Method

Adaptive aggregation in *End-to-End*

- Inside PCIe-to-Ethernet bridge, perform adaptive aggregation *behind congestion control unit*
  - Our congestion control was proposed in another work, *H. Shimonishi et al., CQR Workshop, 2008.*
- TLPs are extracted from aggregation queue at the rate of bottleneck link and aggregated if multiple packets are accumulated then

Adaptive aggregation if TLPs are accumulated

Extraction at the rate of bottleneck link in Ethernet



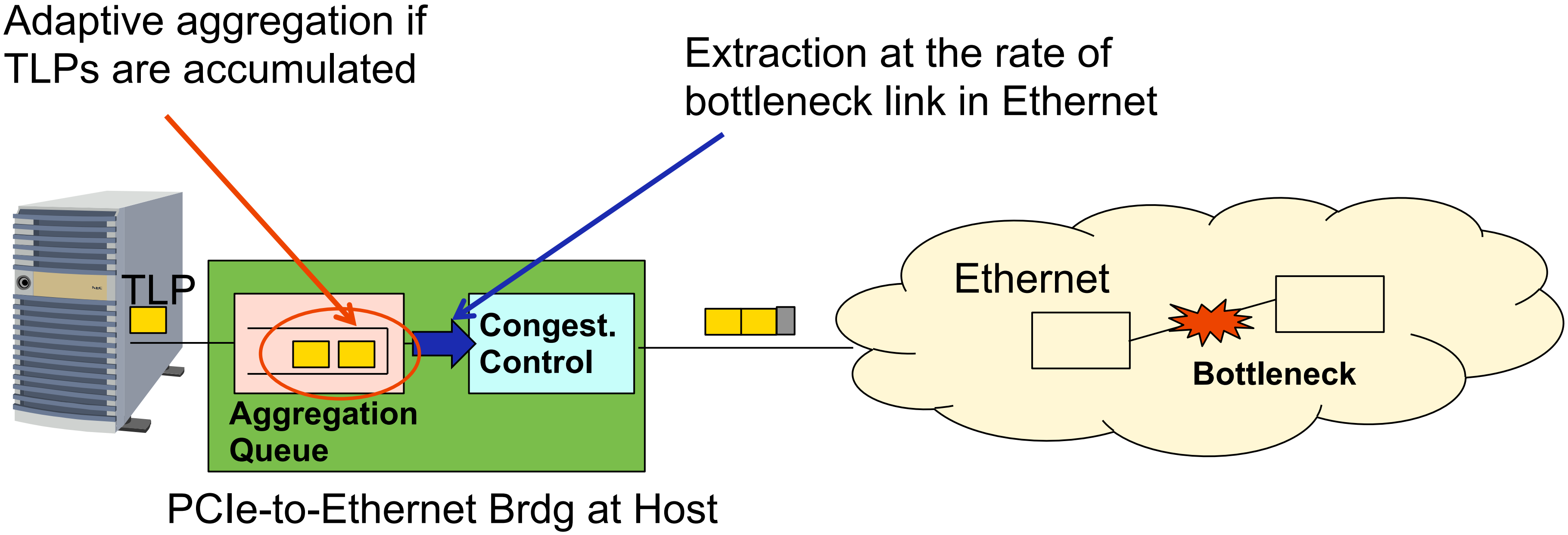
PCIe-to-Ethernet Brdg at Host

Proposed Method

Adaptive aggregation in *End-to-End*

- Inside PCIe-to-Ethernet bridge, perform adaptive aggregation *behind congestion control unit*
  - Our congestion control was proposed in another work, *H. Shimonishi et al., CQR Workshop, 2008.*

- TLPs are extracted from aggregation queue at the rate of bottleneck link and aggregated if multiple packets are accumulated then


Adaptive aggregation if TLPs are accumulated
Extraction at the rate of bottleneck link in Ethernet
TLP
Congest. Control
Aggregation Queue
Ethernet
Bottleneck
PCIe-to-Ethernet Brdg at Host

Feature of the method

Low-latency

- No additional wait time is introduced for aggregation

No manual parameter settings

- #aggregated TLPs are automatically decided

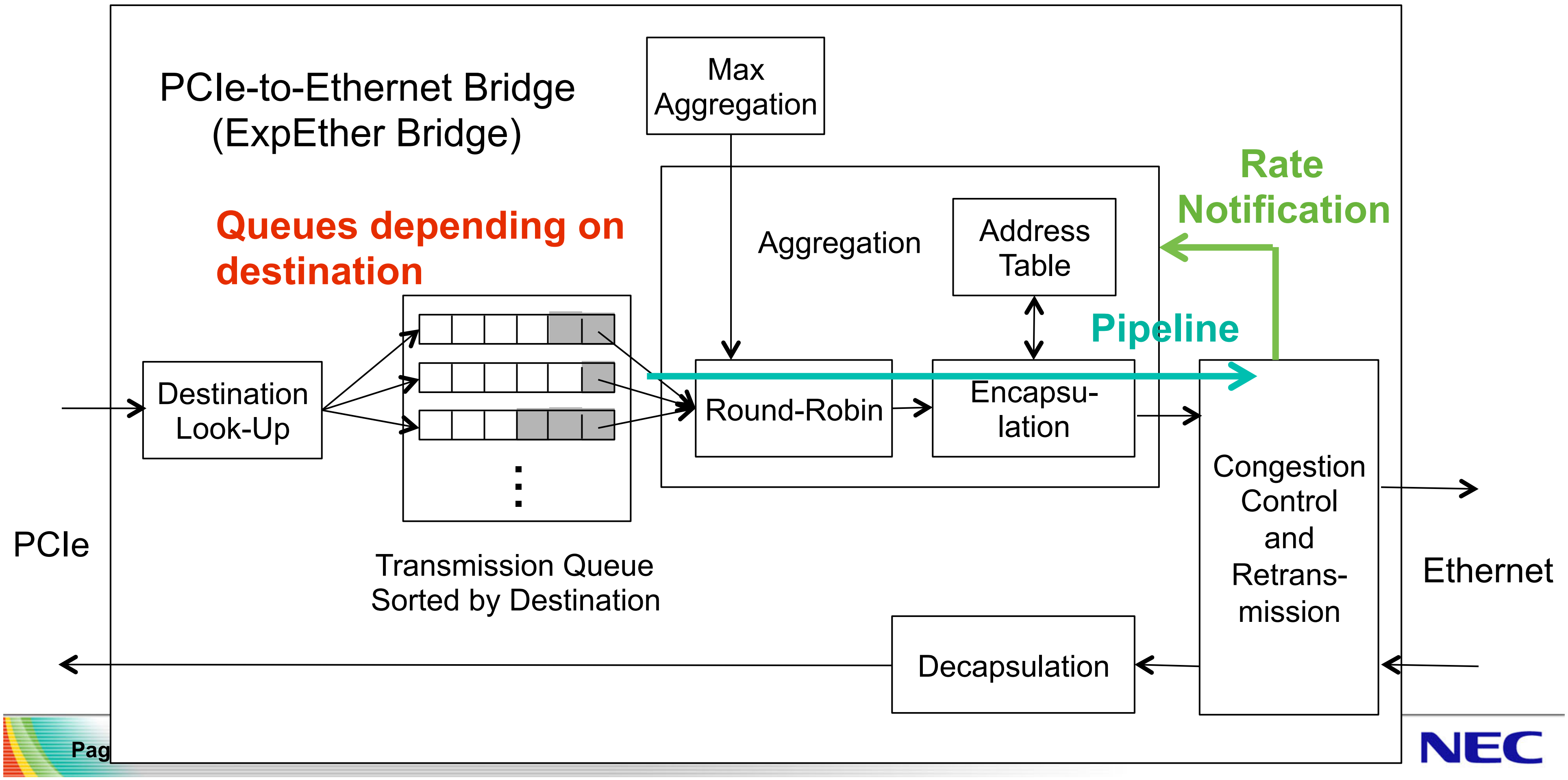
Off-the-shelf OS, device drivers, I/O devices, and Ethernet

Reduced hardware footprints

- Implementing aggregation function before congestion control reduces internal bus width compared to that implemented inside it

Architectural Diagram of PCIe-to-Ethernet Bridge

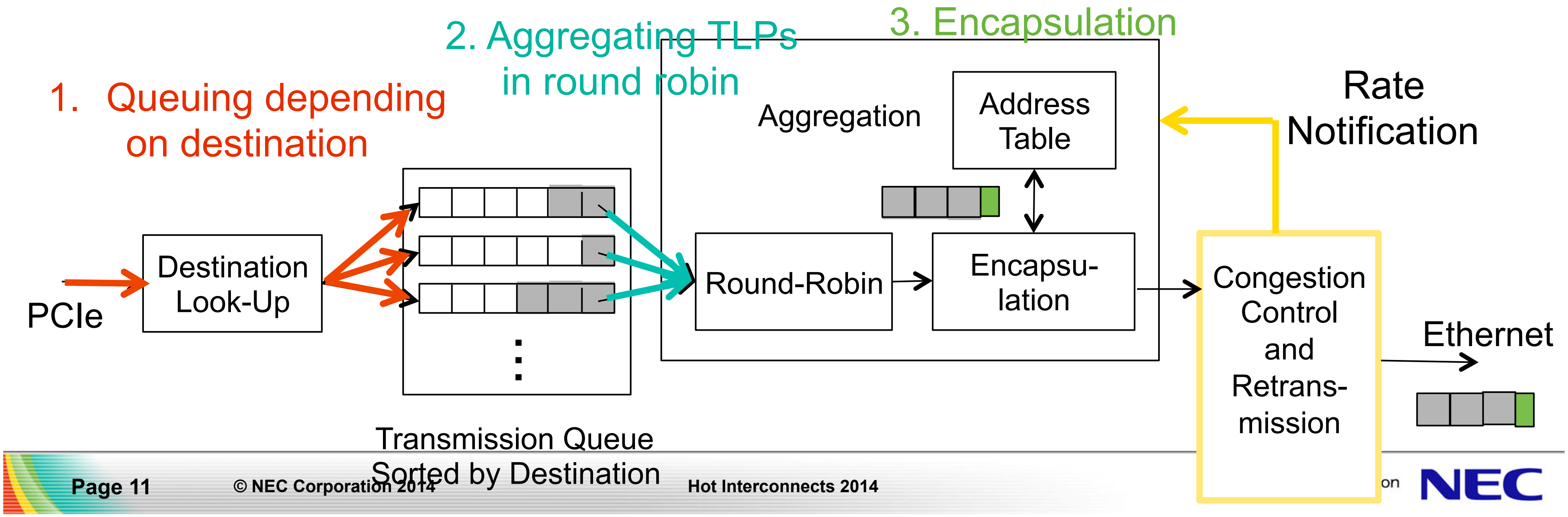
- Multiple aggregation queues depending on destination node
- TLPs are aggregated in pipeline at the rate notified by congestion control function to achieve high throughput transmission



Sending TLPs

1. TLPs received from PCIe are sorted and stored in queues depending on their destination
2. TLPs are extracted from queue in round-robin. All TLPs in each queue up to the number limited by max Ethernet frame are extracted at one time
3. Aggregated TLPs are encapsulated and sent to Ethernet

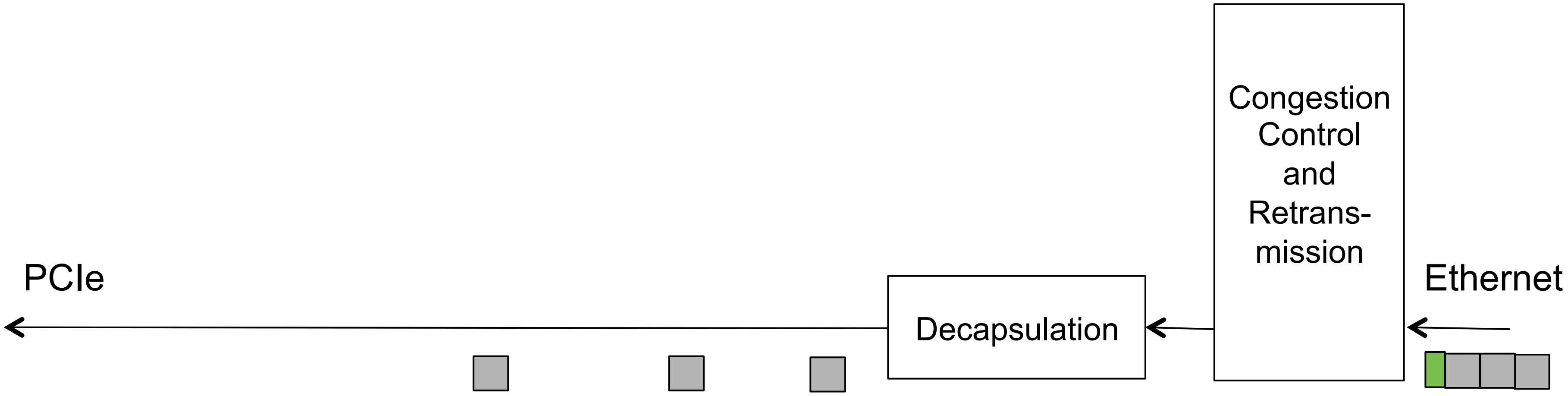
Rate of round-robin TLP extraction is set to the transmission rate notified by congestion control function



2. Aggregating TLPs in round robin
3. Encapsulation
1. Queuing depending on destination
Aggregation
Address Table
Rate Notification
PCIe
Destination Look-Up
Round-Robin
Encapsu-lation
Congestion Control and Retrans-mission
Ethernet
Transmission Queue Sorted by Destination

Receiving TLPs

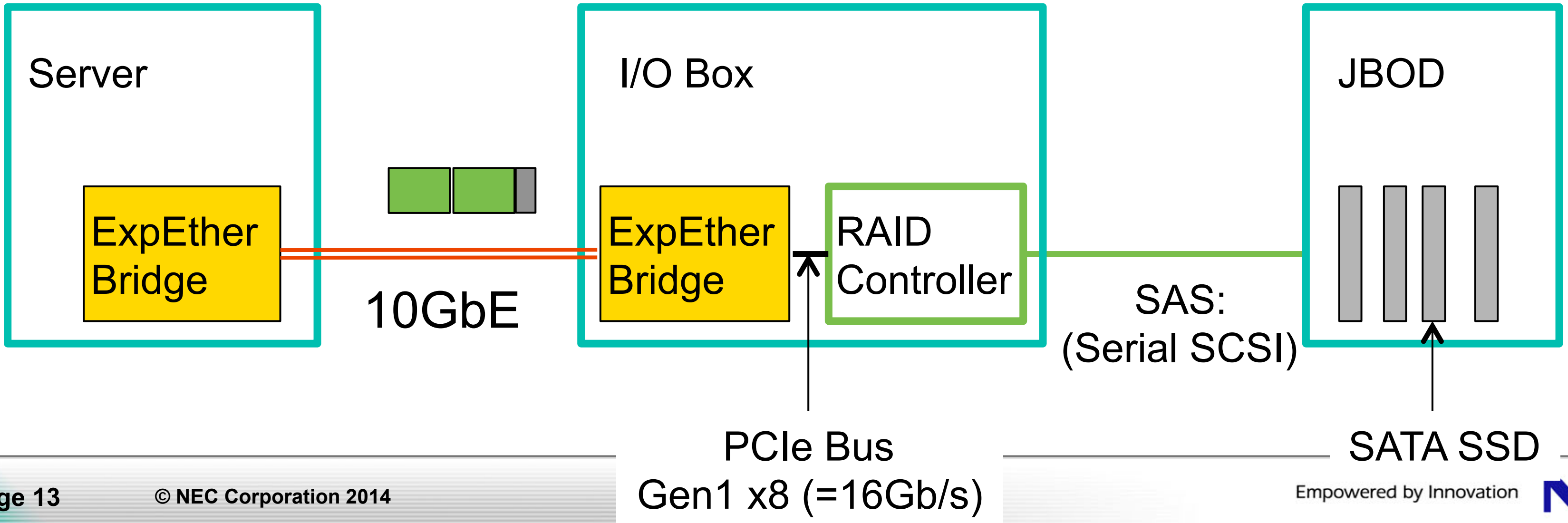
- Aggregated TLPs are decapsulated and sent to PCIe bus



Evaluation using prototype

- Aggregation method was implemented into FPGA-based ExpEther bridge

- I/O performance was evaluated with 1:1 host and I/O device connection
  - Size of TLP payload: 128B
  - RAID0 was configured using SATA SSDs accommodated in JBOD
  - Implemented congestion control unit for evaluation: simple rate limit function to virtualize network congestion



Performed Evaluation

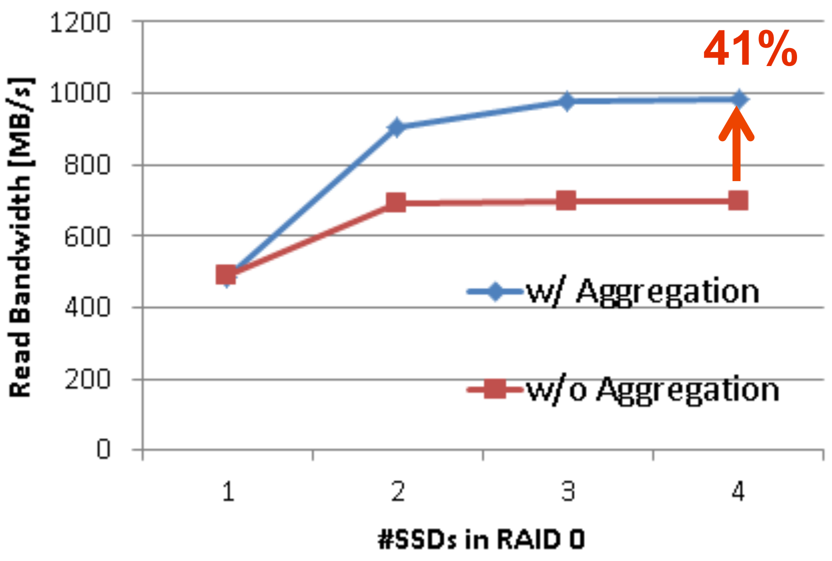
1. Whether TLPs are adaptively aggregated depending on the performance of connected I/O device
   - Even when full Ethernet bandwidth is available, it is bottleneck for some devices and not for others
   - Vary #SSDs configuring RAID0

2. Whether TLPs are adaptively aggregated depending on the bandwidth of Ethernet bottleneck link
   - By using rate limiting function implemented into FPGA

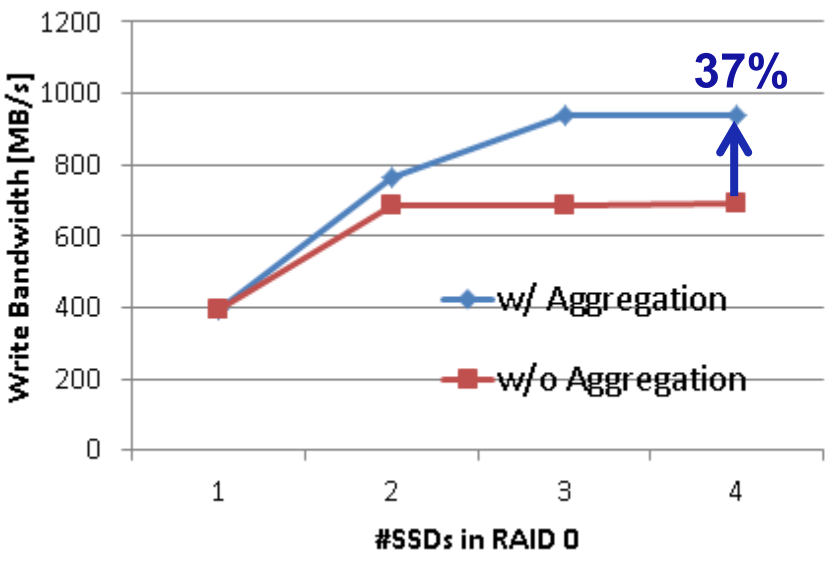
3. Whether TLP aggregation increases transmission delay

[Eval. 1] #SSDs in RAID0 were varied

- Performance of I/O device is increased with #SSDs
- Read and write throughput were increased up to 41% and 37%, respectively
  - ●TLPs were started to be aggregated when #SSDs was two
  - ●Throughput was saturated at 982MB/s (=7.9Gb/s). Further improvement seemed difficult because of TLP and Ethernet header



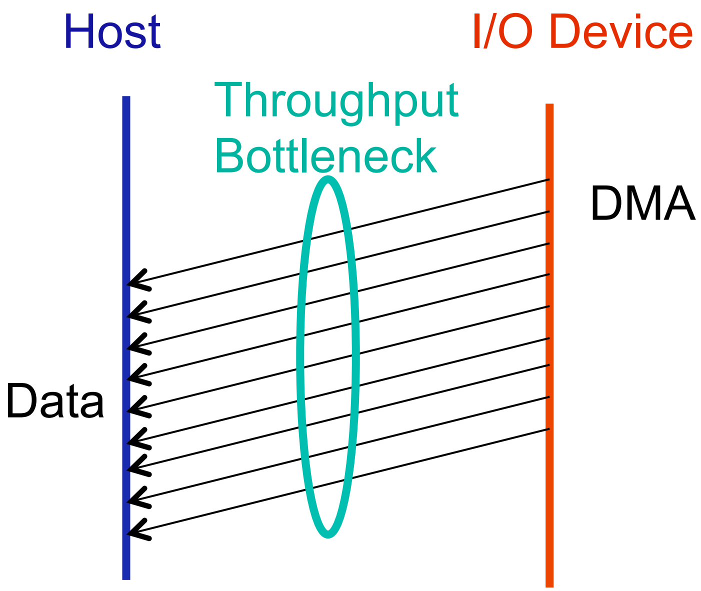
[Read Throughput]



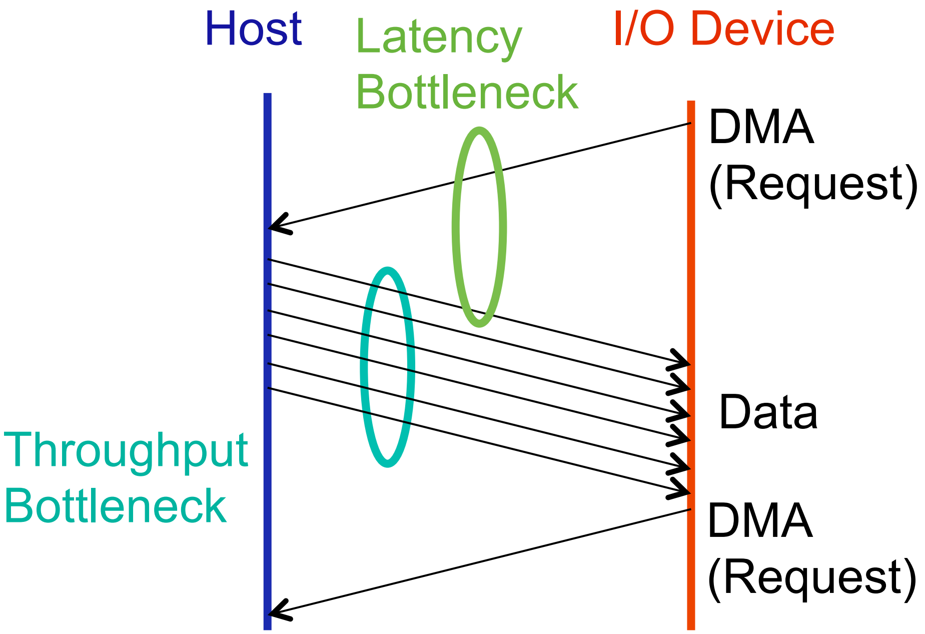
[Write Throughput]

Why the improvement better for read?

- Bottleneck of read performance is Ethernet throughput
  - DMA requests are sent sequentially by I/O device
- Bottleneck of write performance is both Ethernet latency and throughput
  - Response of DMA requests are waited to send next ones



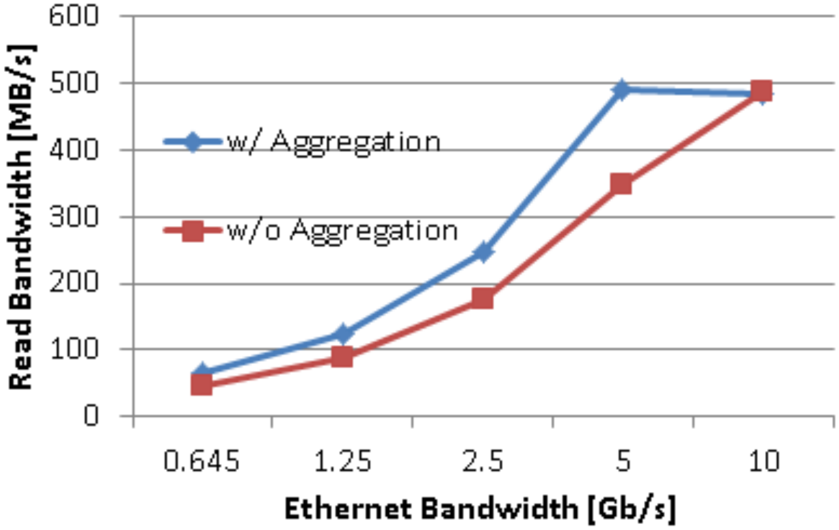
Read from I/O Device



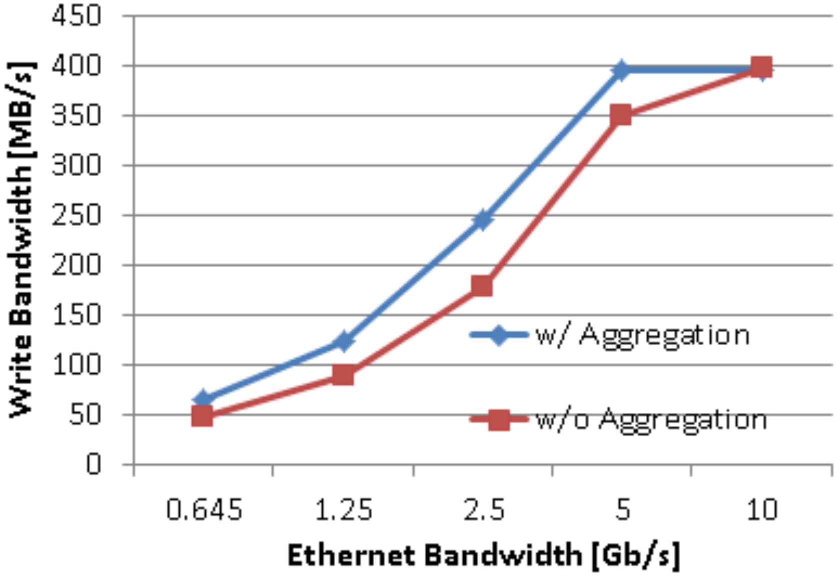
Write to I/O Device

[Eval. 2] Ethernet bandwidth were varied

- No improvement when Ethernet bandwidth was 10 Gb/s
  - Ethernet was not the bottleneck
- Limiting Ethernet bandwidth below 5 Gb/s had TLPs be aggregated
  - Maximum improvement in throughput: 41% in read, 39% in write



[Read Throughput]



[Write Throughput]

#SSD in RAID0 = 1

[Eval. 3] Increase of TLP transmission delay

- Measured degradation of I/O performance using file I/O benchmark "fio"
- 4KB read and write (when Ethernet throughput was not bottleneck)
  - No degradation
- 64MB read and write (when Ethernet throughput was bottleneck)
  - Latency improved because I/O performance was improved

	Read [us]	Write [us]
4KB w/ Aggregation	70.68	98.18
4KB w/o Aggregation	69.5	98.21
64MB w/ Aggregation	65180	65038
64MB w/o Aggregation	91622	89993

↑ 102% ↑ 100%

↑ 71% ↑ 72%

No degrade in short I/O

Conclusion

End-to-end adaptive I/O packet (TLP) aggregation

- Aggregation behind congestion control inside PCIe-to-Ethernet bridge
- Low-latency
- Off-the-shelf OS, device drivers, I/O devices, and Ethernet switches

Evaluation of prototype implemented using FPGA

- I/O performance improved by up to 41%
- No degradation of application performance due to increase of latency

Future work

- Full Implementation of congestion control function
- Evaluation using multiple hosts and I/O devices

Empowered by Innovation

NEC