MS Dissertation

RDMA as Low-latency Interconnect for Scaled-out Software Router

Sangwook Ma

Advisor: Sue Moon

School of Computing, KAIST

2015.12.15.

What Is Software Router?

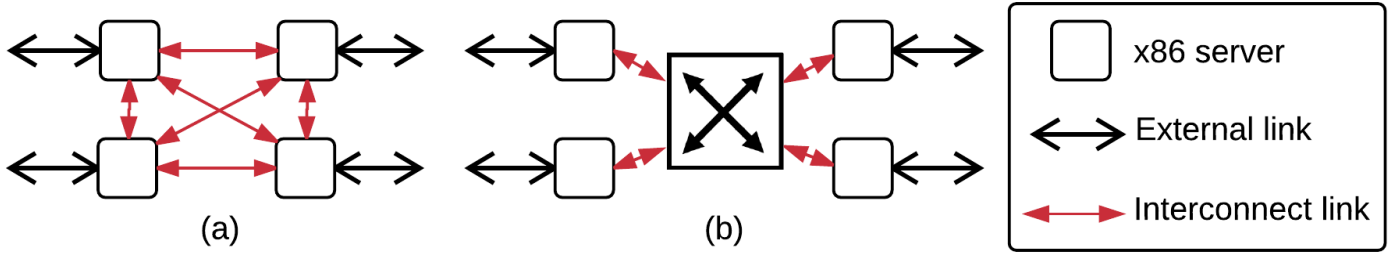
- Packet forwarding and processing on commodity x86 servers
- Low-cost alternatives for HW routers
  - Commercial products: Vyatta (Brocade), 1000V (Cisco)

- Pros: programmability & cost-effectiveness
  - Easy to add or modify packet processing functions
  - Cost-effective compared to proprietary routers
- Cons: lower throughput than HW routers

- Many SW router researches focus on enhancing throughput

Scaling Out Software Router

- Past works: RouteBricks [SOSP 09'], ScaleBricks [SIGCOMM 15']
  - Interconnect technology: Ethernet
  - Topology: Full mesh (RouteBricks) or via central switch (Scalebricks)



- Scale-out leads to inevitable increase in latency
  - Packets need to go through multiple nodes

I/O Batching in Software Router

- A technique sending and receiving packets in batches.
  - Amortizing per-packet overhead in packet I/O
- Essential in high-throughput packet processing engines
  - ex) Packet I/O engine [SIGCOMM 10'], netmap [ATC 12'], Intel DPDK
- The batching causes increase in I/O latency
  - Applications have to wait for NIC to fill packet buffer

Problem Statement

- Scaled-out router suffers from latency increase
- I/O batching trades-off latency for throughput in single node

- Goal of this work
  - Evaluate potential of RDMA (Remote Direct Memory Access) as an interconnect technology to reduce latency in scaled-out router

What is RDMA (Remote Direct Memory Access)?

- RDMA is a transfer mechanism allowing a server to access memory of remote server with minimal CPU involvement

- It provides low-latency and high-throughput by
  - Kernel bypassing: eliminating extra copying and buffering
  - CPU bypassing: offloading data transfer onto NIC

- In this work, we use RoCE (RDMA over Converged Ethernet) [1]
  - RoCE enables to run RDMA over data center Ethernet link

[1]: supplement to Infiniband Architecture Specification Volume 1, Release 1.2.1: Annex A16: RoCE

Why RDMA?

- Interconnect links are isolated from external network
  - No need for compatibility with traditional Ethernet
- RDMA provides low latency and high throughput
  - Mainly used in high-performance computing community
- We evaluate performance benefit of RDMA over Ethernet
  - Performance metrics: throughput and latency

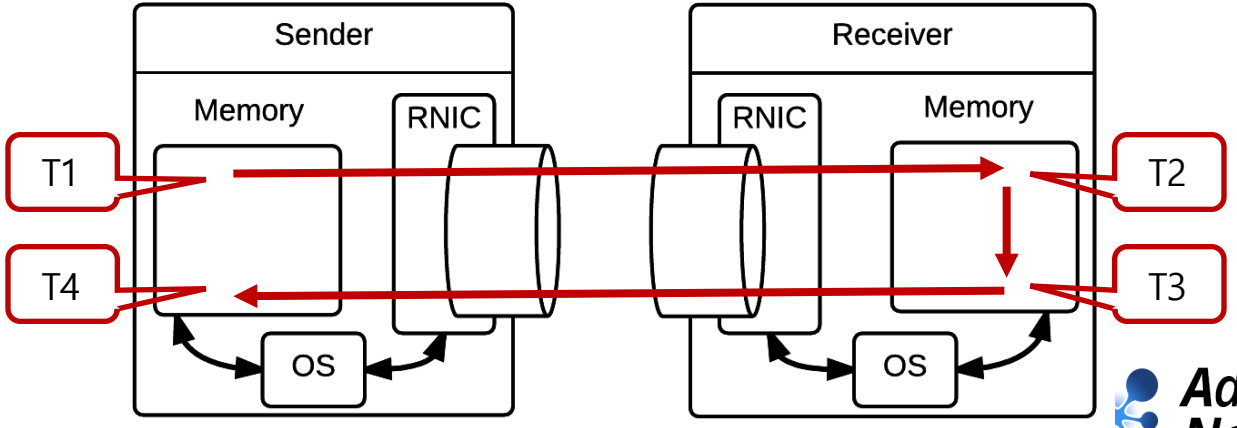
Experiment Setup

- 2 machines with Xeon E5-2670 (2.6 GHz), RAM 32GB
  - Connected using Mellanox ConnectX-3 40GbE NIC and SX1036 Switch
  - Used single CPU core in all measurements

- Software setup
  - OS: Ubuntu 14.04 with kernel v3.16
  - RoCE: Mellanox OFED (OpenFabrics Enterprise Distribution) v3.0.2
  - Ethernet: Intel DPDK (Data Plane Development Kit) v2.1.0
- Measurement program
  - RoCE: our own measurement program
  - Ethernet: own packet generator & DPDK Ethernet forwarding example

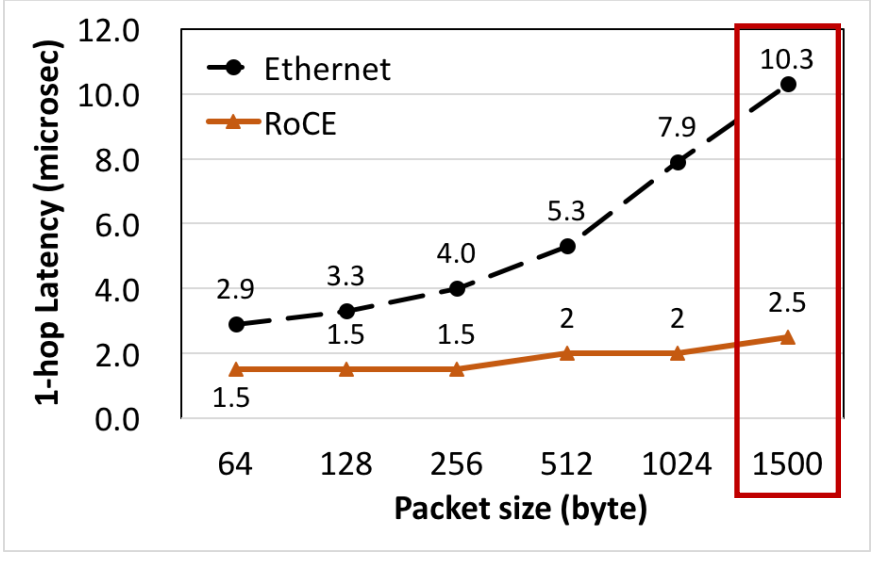
Latency Measurement

- Our definition of 1-hop latency
  - (the last time a sender can modify the packet)
    – (the time just after a receiver can modify the packet)
- 1-hop latency = ((T4 – T1) – (T3 – T2)) / 2
  - (T4 – T1): round-trip time
  - (T3 – T2): the time spent in receiver's OS



RoCE vs. Ethernet: 1-hop Latency

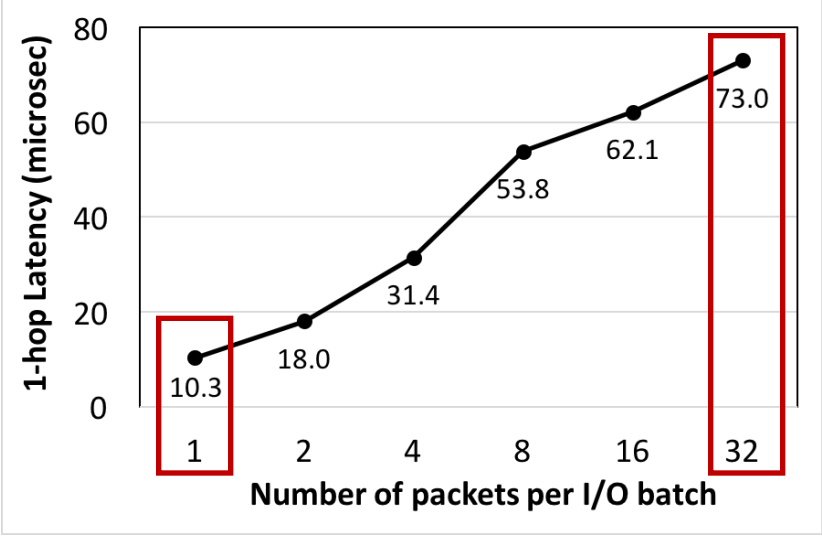
- RoCE latency < 3 usec in all cases
- Max Ethernet latency: 10.3 usec
- We should consider the effect of I/O batching on Ethernet latency.



Median 1-hop latency of RoCE and **Ethernet without I/O batching**

Effect of I/O Batching on Ethernet Latency

- Max 1-hop latency of Ethernet with I/O batching: 73 usec
  - 30x higher than RoCE case
- It is significant overhead in scaled-out router
  - HW router with 40Gbps port maintains latency below 27 usec [2]



1-hop median latency of Ethernet by I/O batch size. Packet size is 1500B.

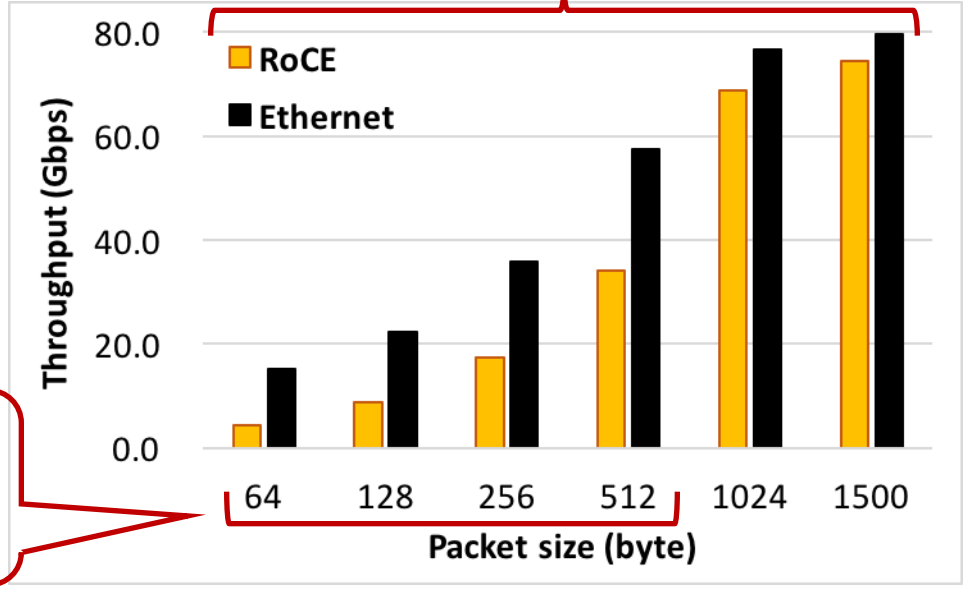
[2]: Performance-Comparison Testing of IPv4 and IPv6 Throughput and Latency on Key Cisco Router Platforms
http://www.cisco.com/c/dam/en/us/products/collateral/ios-nx-os-software/enterprise-ipv6-solution/IPv6perf_wp1f.pdf

Throughput Measurement

- Bidirectional throughput through 40GbE
  - Up to 80Gbps
- Protocol overhead included
  - Ethernet = 14 bytes / packet
  - RDMA over Ethernet = 72 bytes / message
- Ethernet packet sizes of 64 to 1500 bytes used
- Throughput averaged over 10 seconds

RoCE vs. Ethernet: Throughput

Ethernet performs better than RoCE in all packet sizes.



Throughput gap is worse in smaller packet sizes.

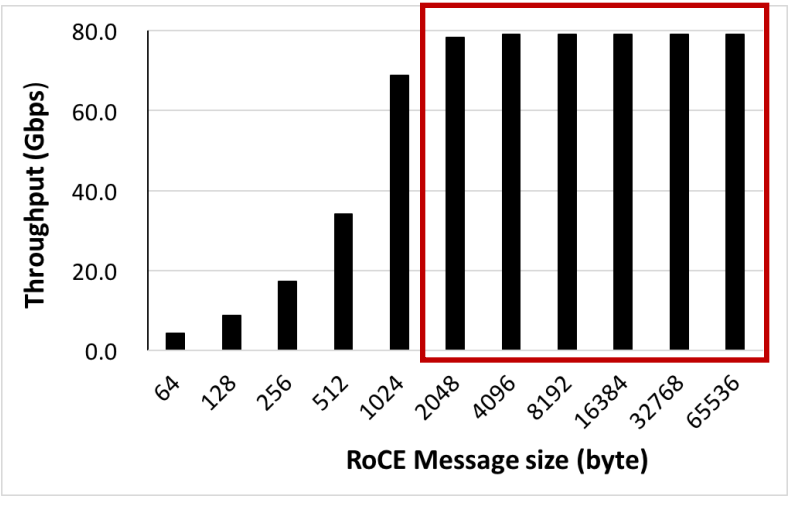
Average throughput of Ethernet with 32-packet I/O batching and RoCE

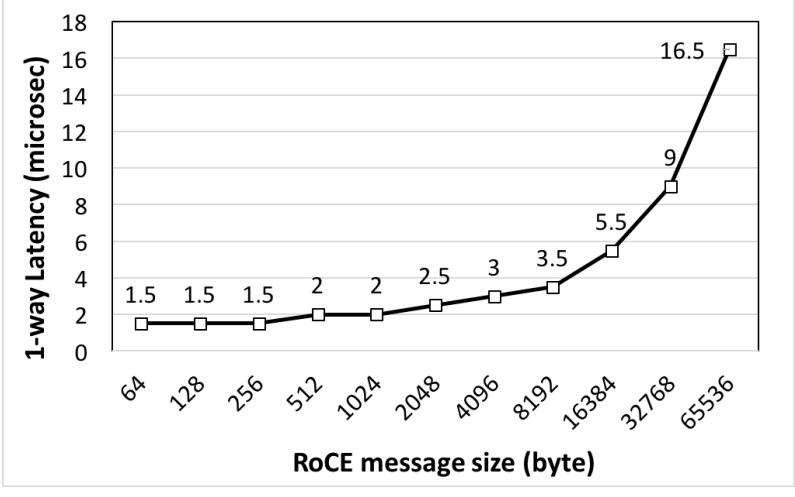
Advantage and Challenge of Using RDMA

- RDMA maintains latency under 3 usec in all packet sizes
  - Up to 30x lower than Ethernet in the same conditions
- RDMA throughput < Ethernet throughput when (packet size ≤ 1500B)
- Solution to low RDMA throughput: NIC-assisted packet batching

Motivation to Batch Packets Using RDMA

- Line-rate throughput in message sizes ≥ 1500B
- Latency < 20 usec even in messages ≥ 1500B
- RDMA scatter/gather mechanism for discontinuous memory segments





NIC-assisted Packet Batching

- Transfers multiple Ethernet packets in a single, combined message
- Utilizes scatter/gather mechanism of RDMA

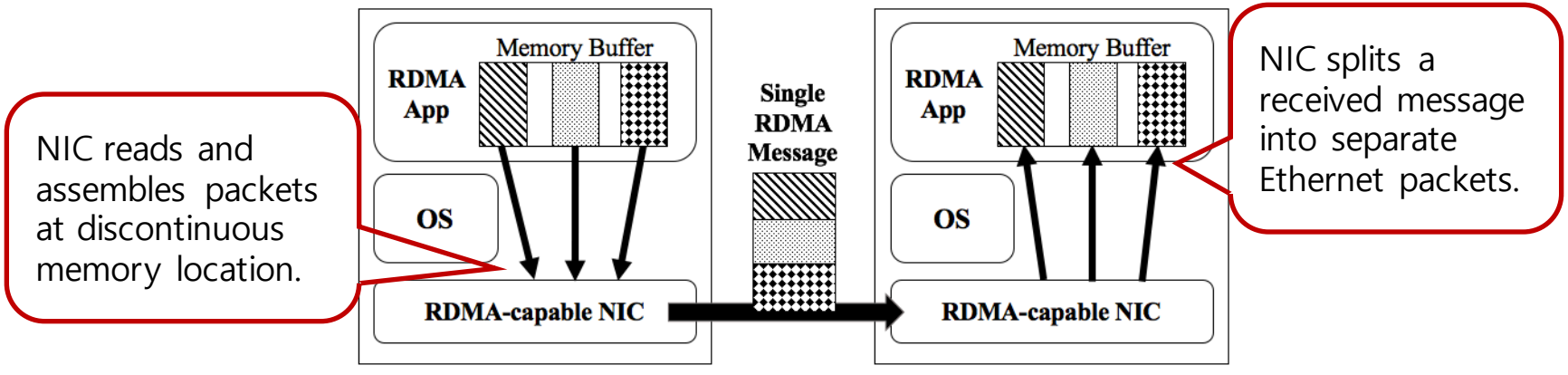
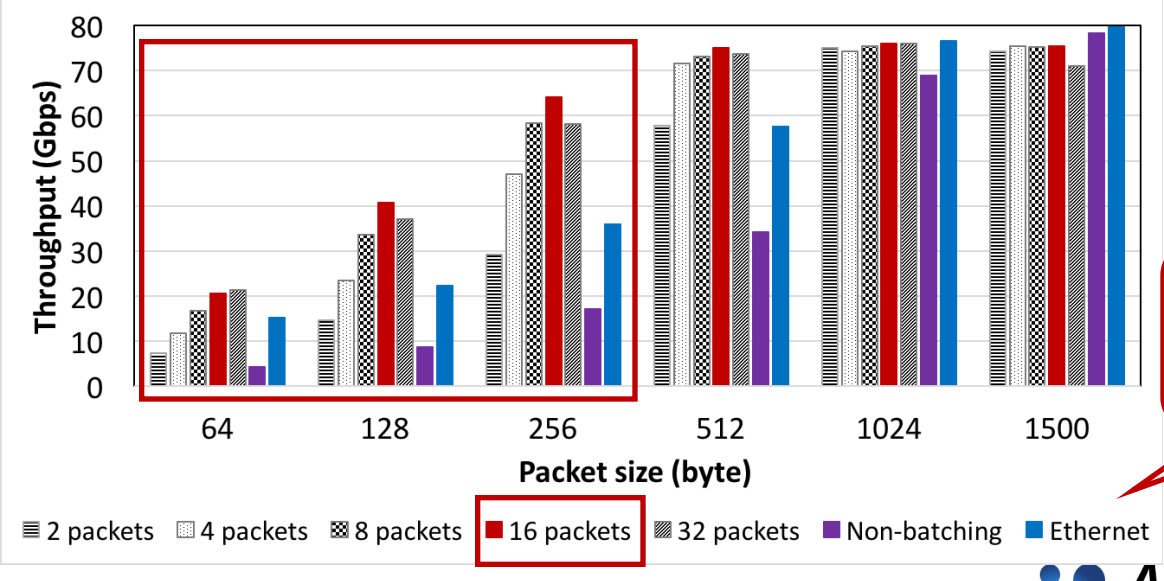


Figure 4: An example of NIC-assisted batching with 3 Ethernet packets

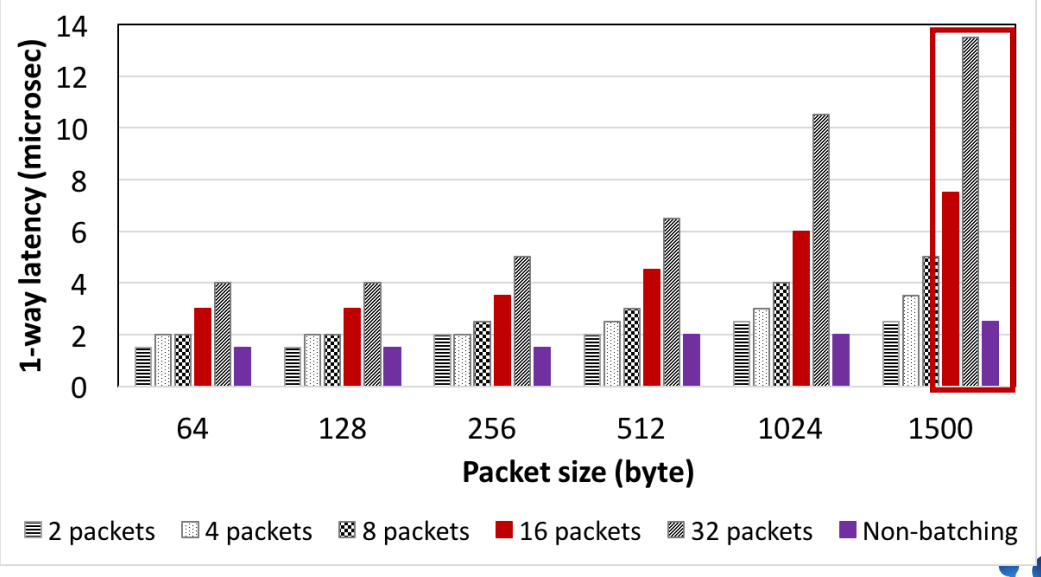
Effect of NIC-assisted Batching: Throughput

- Throughput higher or close to throughput of Ethernet with I/O batching
  - Line-rate throughput when packet size ≥ 512B
  - 3.7~4.8x increase when packet size ≦ 256B



Effect of NIC-assisted Batching: Latency

- Linear increase according to the size of batch and packet
- Max latency: 13.5 usec when packet size is 1500B
  - Still 5.4 times lower than Ethernet's 73 usec



Conclusion

- RDMA is a valid alternative as an interconnect of scaled-out SW router
  - It can reduce I/O latency up to 30x compared to Ethernet
- Challenge is its low throughput in packet sizes ≤ 1500 bytes

- We propose NIC-assisted batching as a solution to enhance throughput
  - It batches multiple Ethernet packets in a single RDMA message.

- The batching can achieve throughput higher or close to Ethernet while still maintains 1-hop latency under 14 usec.

Future Works

- Examine the effect of the number of RDMA connections on performance
- Measure throughput and latency using real traffic traces
- Implement scaled-out SW router prototype using RDMA interconnect
  - Cluster composed of Ethernet ports for external interface and RoCE ports for interconnect

Related Works on RDMA

- Implementation of distributed key-value stores
  - Pilaf [ATC 13'], HERD [SIGCOMM 14'], FaRM [NSDI 14']
- Acceleration of existing applications
  - MPI [ICS 03'], Hbase [IPDPS 12'], HDFS [SC 12'], Memcached [ICPP 11']
  - They replace socket interface with RDMA transfer operations
- RDMA-like interconnects for rack-scale computing
  - Scale-out NUMA [ASPLOS 14'] , R2C2 [SIGCOMM 15'], Marlin [ANCS 14']

Q & A