Revisiting Network Support for RDMA

Radhika Mittal¹,
Alex Shpiner³, Aurojit Panda¹, Eitan Zahavi³,
Arvind Krishnamurthy², Sylvia Ratnasamy¹, Scott Shenker¹

(¹: UC Berkeley, ²: Univ. of Washington, ³: Mellanox Inc.)
Rise of RDMA in datacenters

 RDMA enables low CPU utilization, low latency, high throughput
Current Status

• RoCE (RDMA over Converged Ethernet)
  – canonical approach for deploying RDMA in datacenters.
  – needs lossless network to get good performance.

• Network made lossless using Priority Flow Control (PFC)
  – Complicates network management, congestion spreading, deadlocks
RoCE (RDMA over Converged Ethernet) – canonical approach for deploying RDMA in datacenters. Needs lossless network to get good performance.

Network made lossless using Priority Flow Control (PFC) – Complicates network management, congestion spreading, deadlocks.

Current Status
is losslessness really needed? No!

Simple changes to NIC design enable similar and often better performance without losslessness.
Background
Evolution of RDMA

• Traditionally used in Infiniband clusters.
  – Losses are rare (credit-based flow control).

• NICs were not designed to deal with losses efficiently.
  – Receiver discards out-of-order packets.
  – Sender does go-back-N on detecting packet loss (via timeouts/negative acks).
Go-Back-N Loss Recovery

Retransmit all packets sent after the last acknowledged packet.
RDMA over Converged Ethernet

- RoCE: RDMA over Ethernet fabric.
- RoCEv2: RDMA over IP-routed networks.
- Infiniband transport was adopted as it is.
  - Go-back-N loss recovery
  - Needs lossless for good performance.
- Network was made lossless using PFC.
Why not iWARP?

- Designed to support RDMA over generic (non-lossless) networks.
- Implements entire TCP stack in hardware along with multiple other layers.
- Not as popular as RoCE
  - Less mature, more power, more expensive.

RoCE + PFC emerged as popular choice.
Priority Flow Control (PFC)

- XOFF frame sent when queuing exceeds a certain threshold to pause transmission.
Drawbacks of PFC

• Adds complexity to network management.
  – Need enough headroom to absorb all packets in flight.
Drawbacks of PFC

- Adds complexity to network management.
  – Need enough headroom to absorb all packets in flight.
Drawbacks of PFC

• Performance Issues
  – Unfairness, HoL blocking
Drawbacks of PFC

- Performance Issues
  - Unfairness, HoL blocking
Drawbacks of PFC

- Performance Issues
  - Unfairness, HoL blocking
Drawbacks of PFC

- Performance Issues
  - Unfairness, HoL blocking
Drawbacks of PFC

Congestion Spreading
Drawbacks of PFC

Congestion Spreading

Switch A

Switch B
Drawbacks of PFC

Deadlocks caused by cyclic buffer dependency
**Advanced Congestion control**

**RoCE**

**DCQCN**: Zhu et al, SIGCOMM 2015 (Microsoft)
- ECN-based congestion control
- Implemented on NIC hardware (Mellanox ConnectX4)

**Timely**: Mittal et al, SIGCOMM 2015 (Google)
- Delay-based congestion control
Recent Works highlighting PFC Issues

• RDMA over commodity Ethernet at scale, SIGCOMM 2016
• Deadlocks in datacenter: why do they form and how to avoid them, HotNets 2016
• Unlocking credit loop deadlock, HotNets 2016
Our approach

• Based on the principle of iWARP
  – NIC efficiently deals with packet losses

• But on the implementation of RoCE
  – Incorporate only necessary bare-minimal features
Is losslessness needed for RDMA?
Experimental Setup

- Mellanox simulator modeling MLX CX4 NICs, built on Omnet/Inet.
- Three layered fat-tree topology.
- Links with capacity 10Gbps and delay 2us.
- Heavy-tailed distribution at 70% utilization.
Metrics

• Average Flow Completion Time

• 99%ile Flow Completion Time

• Average Slowdown
Current RoCE NICs

PFC helps a lot.
Current RoCE NICs

PFC helps a lot.

Graph showing 99th percentile FCT (s) with and without PFC for RoCE, RoCE + Timely, and RoCE + DCQCN.
Current RoCE NICs

PFC helps a lot.
Key results

• PFC is needed with current RoCE NICs.

• PFC is not needed with IRN.

• IRN performs better than current RoCE NICs.
Improved RoCE NIC (IRN)

• Three key changes:
  – Selective retransmit instead of go-back-N
  – Back-off on losses
  – Cap the number of outstanding bytes by BDP
IRN: no advanced congestion control

Disabling PFC gives much better performance.
Disabling PFC gives much better performance.
IRN: no advanced congestion control

Disabling PFC gives much better performance.
IRN with advanced congestion control

PFC is not needed.
IRN with advanced congestion control

PFC is not needed.
IRN with advanced congestion control

PFC is not needed.
Key results

- PFC is needed with current RoCE NICs.
- PFC is not needed with IRN.
- IRN performs better than current RoCE NICs.
IRN vs RoCE

IRN performs better than RoCE.
IRN vs RoCE

IRN performs better than RoCE.
IRN vs RoCE

IRN performs better than RoCE.
Key results

• PFC is needed with current RoCE NICs.

• PFC is not needed with IRN.

• IRN performs better than current RoCE NICs.
Robustness of results

- Tested a wide range of experimental scenarios:
  - Higher link bandwidth of 40Gbps and 100Gbps
  - More uniform workload distribution
  - Varying link delay
  - Varying link utilization

- Our key take-away hold across all of these scenarios.
Efficient RDMA transport does not require a lossless network.

With simple NIC changes, generic out-of-the-box Ethernet network outperforms a lossless network.
Deep Dive
Improved RoCE NIC (IRN)

• Three key changes:
  – Selective retransmit instead of go-back-N
  – Back-off on losses
  – Cap the number of outstanding bytes by BDP
Improved RoCE NIC (IRN)

- Three key changes:
  - Selective retransmit instead of go-back-N
  - Back-off on losses
  - Cap the number of outstanding bytes by BDP
Need for Selective Retransmit

Disabling selective retransmit for IRN+DCQCN increased avg FCT by 28% for our default scenario.
Improved RoCE NIC (IRN)

• Three key changes:
  – Selective retransmit instead of go-back-N
  – Back-off on losses
  – Cap the number of outstanding bytes by BDP
Benefit of Backing Off on Losses

• Exploit losses as congestion signal

• Found to be particularly useful
  o When no advanced congestion control is being used.
    - Disabling this feature for IRN increased avg FCT by 158% for our default scenario.
  o For scenarios where advanced congestion control does not react fast enough.
    - Disabling this feature for IRN+DCTCP increased avg FCT by 140% when link bandwidth is 100Gbps.
Improved RoCE NIC (IRN)

• Three key changes:
  – Selective retransmit instead of go-back-N
  – Back-off on losses
  – Cap the number of outstanding bytes by BDP
Benefit of capping by BDP

• Upper bound on number of out-of-order packets.

• Improves performance irrespective of whether PFC is enabled or disabled.
  – Disabling this feature for IRN+DCQCN increased the avg FCT by 64% for our default scenario.
Design Details

• IRN supports both window mode or rate mode.
• Start at line rate (or with cwnd = BDP)
• Losses detected via:
  o Three dupacks
    – selective retransmit + reduce rate or cwnd by half
  o Partial ack
    – selective retransmit
  o Fixed timeout
    – go-back-N + reduce rate or cwnd by half
• Option for additive increase on new acks.
• For RDMA reads: requester generates read acks.
Implementation Feasibility

- **Support for out of order packet delivery at the receiver**
  - New feature in Mellanox CX5 NICs for adaptive routing.
  - Straight forward to implement selective retransmit.

- **Other IRN changes**: few additional instructions.

- **Increase in per-flow state**: less than 3%.
Summary

• IRN performs better than RoCE and does not need PFC.

• Holds across various congestion control algorithms and experimental scenarios.

• The NIC changes required are simple and feasible.
Thank you!