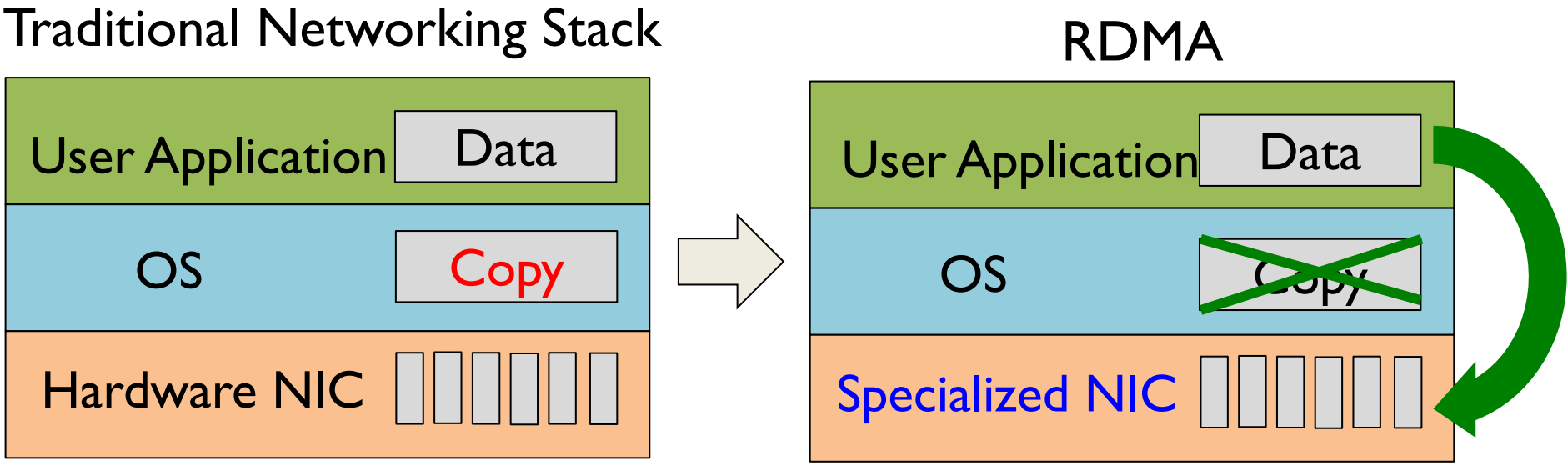
Revisiting Network Support for RDMA

Radhika Mittal[1],
Alex Shpiner[3], Aurojit Panda[1,4], Eitan Zahavi[3],
Arvind Krishnamurthy[2], Sylvia Ratnasamy[1], Scott Shenker[1]
(1: UC Berkeley, 2: Univ. of Washington, 3: Mellanox Inc., 4: NYU)

Rise of RDMA in datacenters



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.
  - Various known performance issues.

Current S

-

Practical DCB for Improved Data Center Networks

Brent Stephens, Alan L. Cox
Rice University

Ankit Singla
UIUC

John Carter, Colin Dixon, Wesley Felter
IBM Research, Austin

***Abstract*—Storage area networking is d**
center switches to support lossless Eth
nately, to enable DCB for all traffic
topologies, we must address several pr
lossless networks, e.g., large buffering d
line blocking, and deadlock. We propos
that not only addresses the first three
completion times by as much as 70%.
practical deadlock-free routing scher
while achieving aggregate network
ECMP routing. This small comp
capacity is well worth the gains in f
that our results on deadlock-free r
interest to the storage area netwo
our hardware testbed illustrates, t
without hardware changes to sw

incast [5], TCP congestion collapse that occurs in short, barrier
latency sensitive workloads.
standard is driven by the dema
(SAN) and LAN fabr
managemer

Unlocking Credit Loop Deadlocks

Alexander Shpiner, Eitan Zahavi, Vladimir Zdornov, Tal Anker and Matty Kadosh
Mellanox Technologies, Inc
Yokneam, Israel
{alexshp,eitan,vladimirz,ankertal,mattyk}@mellanox.com

Abstract
The recentl

CEE)
rder
as
en
e-
k

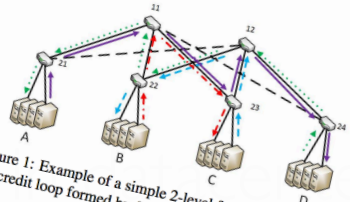


Figure 1: Example of a simple 2-level fat tree topology with the credit loop formed by failure of links 12-21 and 11-23.

RDMA over Commodity Ethernet at Scale

Chuanxiong Guo, Haitao Wu, Zhong D
Jianxi Ye, Jitendra Padhye, Mari
Microsoft
{chguo, hwu, zdeng, gasoni, jiye, padhye, r

Congestion Control for Large-Scale RDMA Deployments

Yibo Zhu[1,3] Haggai Eran[2] Daniel Firestone[1] Chuanxiong Guo[1] Marina Lipshteyn[1]
Yehonatan Liron[2] Jitendra Padhye[1] Shachar Raindel[2] Mohamad Haj Yahia[2] Ming Zhang[1]
[1]Microsoft [2]Mellanox [3]U. C. Santa Barbara

ABSTRACT

Modern datacenter applications demand high throughput (40Gbps) and ultra-low latency (< 10 μs per hop) from the network, with low CPU overhead. Standard TCP/IP stacks cannot meet these requirements, but Remote Direct Memory Access (RDMA) can. On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol, which relies on Priority-based Flow Control (PFC) to enable a drop-free network. However, PFC can lead to poor application performance due to problems like head-of-line blocking and unfairness. To alleviates these problems, we introduce DCQCN, an end-to-end congestion control scheme for RoCEv2. To optimize DCQCN performance, we build a fluid model, and provide guidelines for tuning switch buffer thresholds, and other protocol parameters. Using a 3-tier Clos network testbed, we show that DCQCN dramatically improves throughput and fairness of RoCEv2 RDMA traffic. DCQCN is implemented in Mellanox NICs, and is being deployed in Microsoft's datacenters.

brutal economics of cloud services business dictates that CPU usage that cannot be monetized should be minimized: a core spent on supporting high TCP throughput is a core that cannot be sold as a VM. Other applications such as distributed memory caches [10, 30] and large-scale machine learning demand ultra-low latency (less than 10 μs per hop) message transfers. Traditional TCP/IP stacks have far higher latency [10].

We are deploying Remote Direct Memory Access (RDMA) technology in Microsoft's datacenters to provide ultra-low latency and high throughput to applications, with very low CPU overhead. With RDMA, network interface cards (NICs) transfer data in and out of pre-registered memory buffers at both end hosts. The networking protocol is implemented entirely on the NICs, bypassing the host networking stack. The bypass significantly reduces CPU overhead and overall latency. To simplify design and implementation, the protocol assumes a lossless networking fabric.

While the HPC community has long used RDMA in special-purpose clusters [11, 24, 26, 32, 38], deploying RDMA on a large scale in modern IP-routed datacenter networks presents

Deadlocks in Datacenter Networks: Why Do They Form, and How to Avoid Them

Shuihai Hu[1,2] Yibo Zhu[1] Peng Cheng[1] Chuanxiong Guo[1] Kun Tan[1]
Jitendra Padhye[1] Kai Chen[2]
[1]Microsoft [2]Hong Kong University of Science and Technology

TRACT

by the need for ultra-low latency, high throughput
CPU overhead, Remote Direct Memory Access (RDMA)
deployed by many cloud providers. To deploy RDMA
et networks, Priority-based Flow Control (PFC) must
PFC, however, makes Ethernet networks prone
ks. Prior work on deadlock avoidance has fo-
cessary condition for deadlock formation, which
r onerous and expensive solutions for deadlock
this paper, we investigate *sufficient* conditions
rmation, conjecturing that avoiding *sufficient*
t be less onerous.

UCTION

e discuss a problem that is quite (c)old,
re-emerged in a new context, and admit
a how to solve it completely. Our hope
unity's attention to this problem, and
is area.

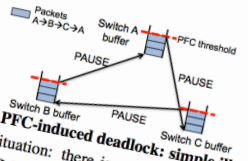


Figure 1: PFC-induced deadlock: simple illustration

a standstill situation: there is a *Cyclic Buffer Dependency* (CBD) among a set of switches. Each switch in the cycle holds all the buffer needed by its upstream switch, and meanwhile is waiting for its downstream switch to release some buffer and resume its packet transmission. A simple scenario is illustrated in Figure 1.

It is easy to see that when deadlock occurs, no switch in the cycle can proceed. Further, throughput of the whole network or part of the network will go to zero due to the back-pressure effect of PFC pause.

It is often believed that such deadlocks cannot occur in clos-structured datacenter networks, since a loop cannot form in such networks with valley-free routing [24]. However,

- Current S

Is a lossless network really needed?

No!

Incremental changes to RoCE NIC design can enable better performance without a lossless network.

Practical DCB for Improved Data Center Networks

Brent Stephens, Alan L. Cox — Rice University
Ankit Singla — UIUC
John Carter, Colin Dixon, Wesley Felter — IBM Research, Austin

Unlocking Credit Loop Deadlocks

Alexander Shpiner, Eitan Zahavi, Vladimir Zdornov, Tal Anker and Matty Kadosh
Mellanox Technologies, Inc
Yokneam, Israel
{alexshp,eitan,vladimirz,ankertal,mattyk}@mellanox.com

Congestion

Yibo Zhu[1,3] Haggai Era
Yehonatan Liron[2] Jitendra

ABSTRACT

Modern datacenter applications dema
(40Gbps) and ultra-low latency (< 10
network, with low CPU overhead. Standard TCP/IP stacks
cannot meet these requirements, but Remote Direct Mem-
ory Access (RDMA) can. On IP-routed datacenter networks,
RDMA is deployed using RoCEv2 protocol, which relies on
Priority-based Flow Control (PFC) to enable a drop-free net-
work. However, PFC can lead to poor application perfor-
mance due to problems like head-of-line blocking and un-
fairness. To alleviates these problems, we introduce DC-
QCN, an end-to-end congestion control scheme for RoCEv2.
To optimize DCQCN performance, we build a fluid model,
and provide guidelines for tuning switch buffer thresholds,
and other protocol parameters. Using a 3-tier Clos network
testbed, we show that DCQCN dramatically improves through-
put and fairness of RoCEv2 RDMA traffic. DCQCN is im-
plemented in Mellanox NICs, and is being deployed in Mi-
crosoft's datacenters.

Form,

Figure 1: PFC-induced deadlock: simple illustration

a standstill situation: there is a *Cyclic Buffer Dependency* (CBD) among a set of switches. Each switch in the cycle holds all the buffer needed by its upstream switch, and mean-while is waiting for its downstream switch to release some buffer and resume its packet transmission. A simple scenario is illustrated in Figure 1.

It is easy to see that when deadlock occurs, no switch in the cycle can proceed. Further, throughput of the whole net-work or part of the network will go to zero due to the back-pressure effect of PFC pause.

It is often believed that such deadlocks cannot occur in clos-structured datacenter networks, since a loop cannot form in such networks with valley-free routing [24]. However,

History of RDMA

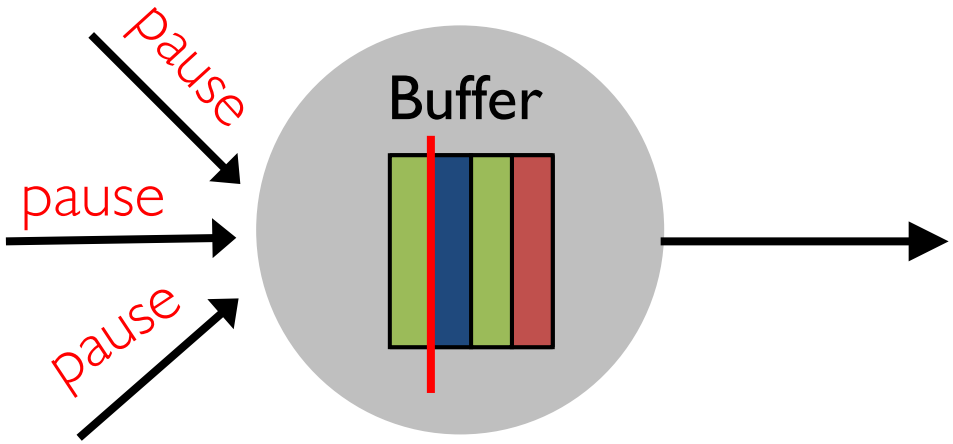
- RDMA traditionally used in Infiniband clusters.
  - Losses are rare (credit-based flow control).
- Transport layer in RDMA NICs not designed to deal with losses efficiently.
  - Receiver discards out-of-order packets.
  - Sender does *go-back-N* on detecting packet loss.

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 a lossless network for good performance.

Network made lossless by enabling PFC

- PFC: Priority Flow Control



- Complicates network management.
- Performance issues:
  - head-of-the-line-blocking, unfairness, congestion spreading, deadlocks.

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
- Tagger: Practical PFC deadlock prevention in datacenter networks, CoNext 2017

Can we alter the RoCE NIC design such that a lossless network is not required?

Why not iWARP?

- Designed to support RDMA over a fully general network.
  - Implements entire TCP stack in hardware.
  - Needs translation between RDMA and TCP semantics.

- General consensus:
  - iWARP is more complex, more expensive, and has worse performance.

iWARP vs RoCE

NIC	Cost in Dec 2016	Throughput	Latency
iWARP: *Chelsio T-580-CR*	$760	3.24Mpps	2.89us
ROCE: *Mellanox MCX 416A-BCAT*	$420	14.7Mpps	0.94us

**Could be due to a number of reasons besides transport design: different profit margin, engineering effort, supported features etc.*

Our work shows that

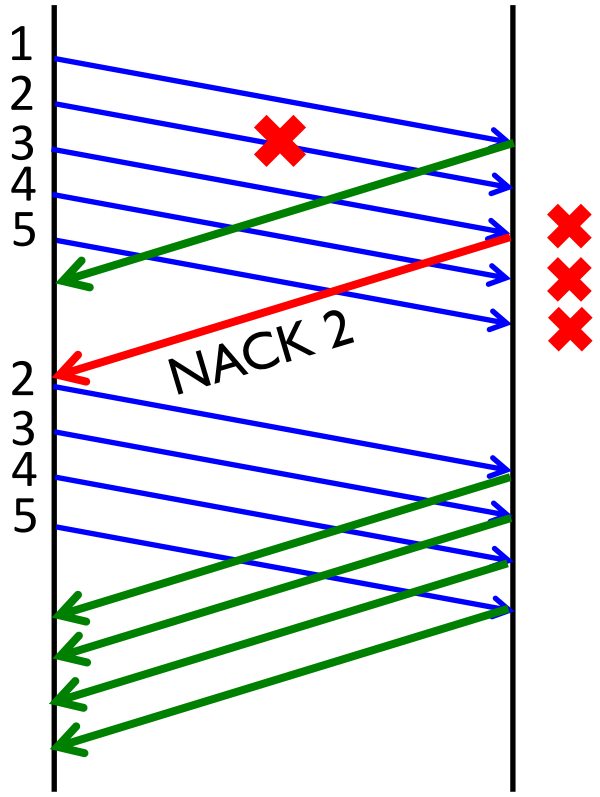
- iWARP had the right philosophy.
  - NICs should efficiently deal with packet losses.
  - Performs better than having a lossless network.

- But we can have a design much closer RoCE.
  - No need to support the entire TCP stack.
  - Identify incremental changes for better loss recovery.
  - Less complex and more performant than iWARP.

Improved RoCE NIC (IRN)

1. Better loss recovery.

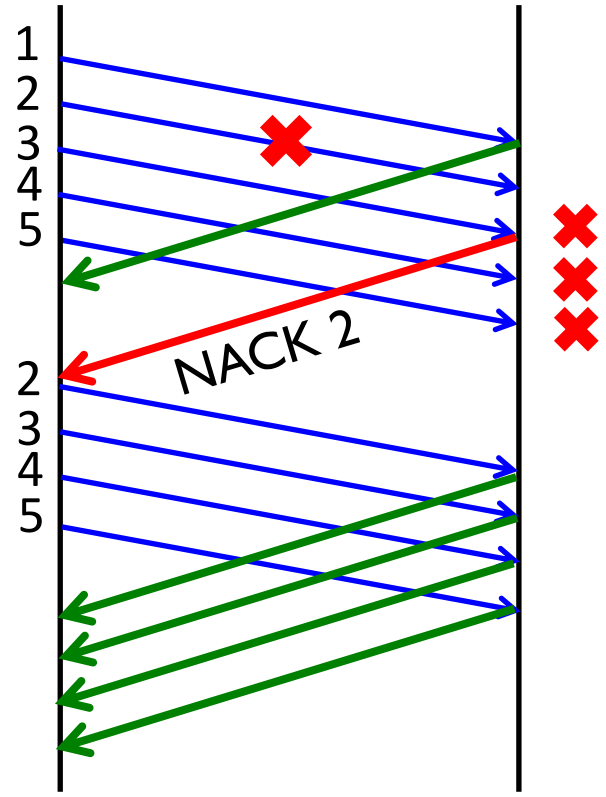
RoCE uses go-back-N loss recovery



Receiver discards all out-of-order packets.

Sender retransmits all packets sent after the last acked packet.

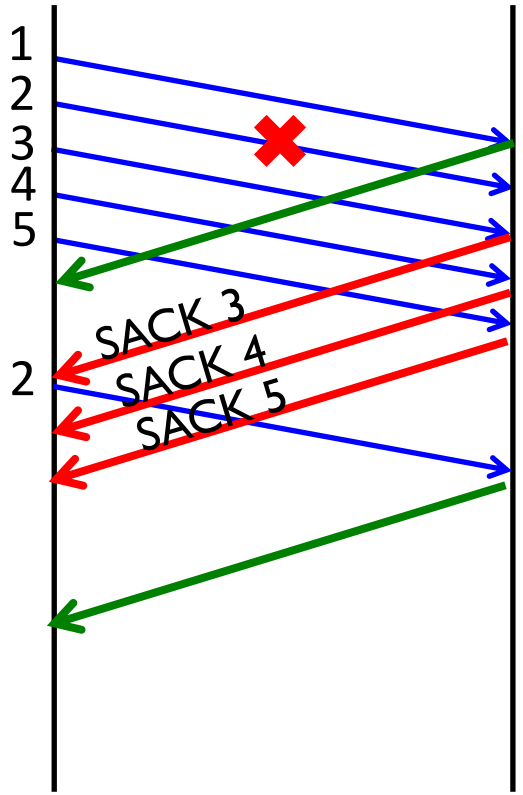
Instead of go-back-N loss recovery…



Receiver discards all out-of-order packets.

Sender retransmits all packets sent after the last acked packet.

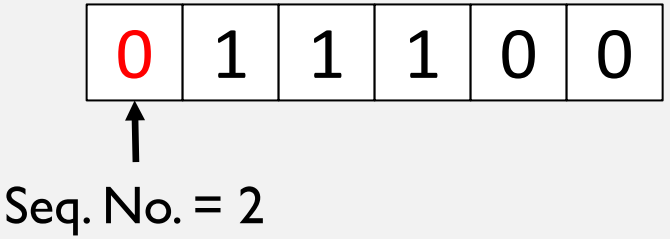
…use selective retransmission



Receiver does not discard out-of-order packets and *selectively acknowledges* them.

Sender retransmits only the lost packets.

Use bitmaps to track lost packets.



Handling timeouts

- Very small timeout value
  - Spurious retransmissions.
- Very large timeout value
  - High tail latency for short messages.

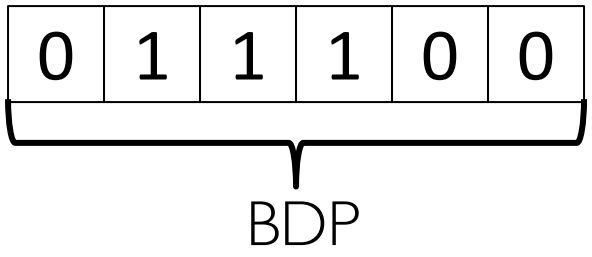
- IRN uses two timeout values
  - RTO_{low}: Less than N packets in flight.
  - RTO_{high}: Otherwise.

Improved RoCE NIC (IRN)

1. Better loss recovery.
   - Selective retransmission instead of go-back-N.
     - Inspired from traditional TCP, but simpler.
   - Two timeout values instead of one.
2. BDP-FC: BDP based flow control.

BDP-FC

- Bound the number of in-flight packets by the bandwidth-delay product (BDP) of the network.
- Reduces unnecessary queuing.
- Strictly upper-bounds the amount of required state.



Improved RoCE NIC (IRN)

1. Better loss recovery.
   - Selective retransmission instead of go-back-N.
     - Inspired from traditional TCP, but simpler.
   - Two timeout values instead of one.
2. BDP-FC: BDP based flow control.
   - Bound the number of in-flight packets by the bandwidth-delay product (BDP) of the network.

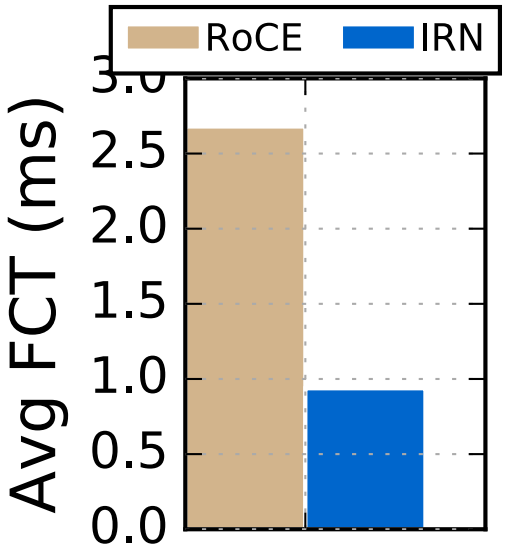
Can IRN eliminate the need for a lossless network?

Default evaluation setup

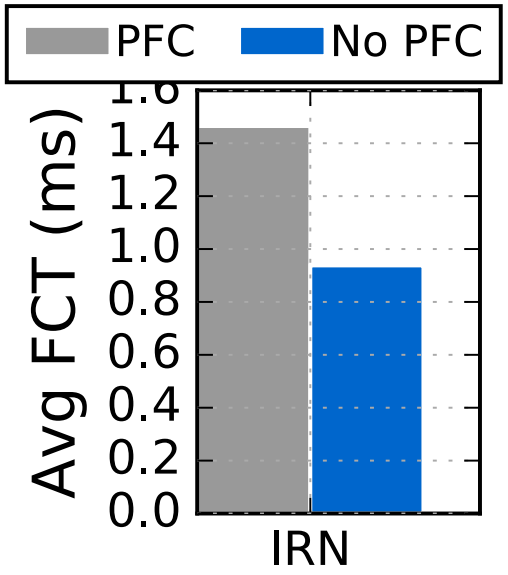
- Mellanox simulator modeling ConnectX4 NICs.
  - Extended from Omnet/Inet.
- Three layered fat-tree topology.
- Links with capacity 40Gbps and delay 2us.
- Heavy-tailed distribution at 70% utilization.
- Per-port buffer of 2 x (bandwidth-delay product).

Key results

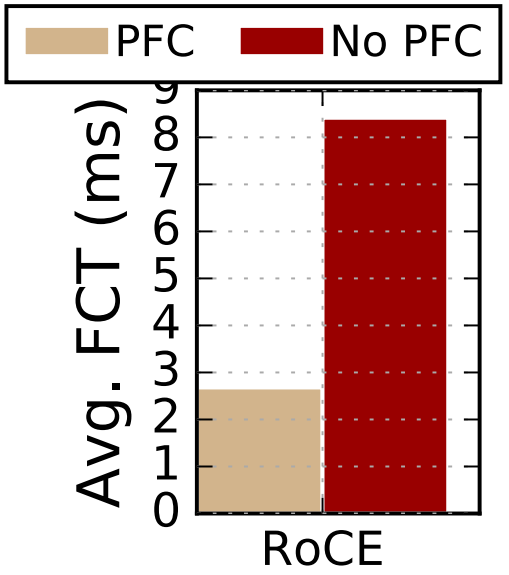
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.

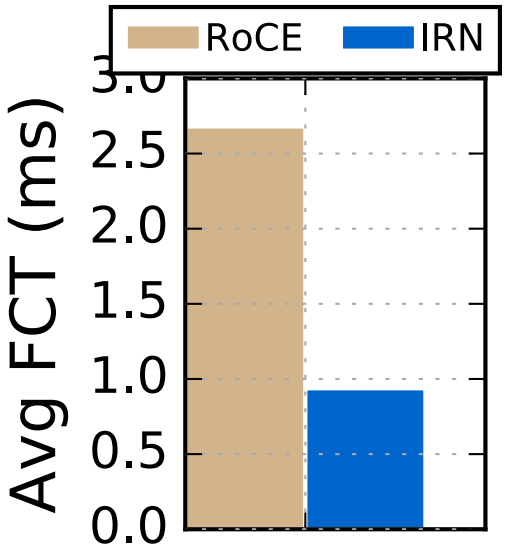


RoCE requires PFC.

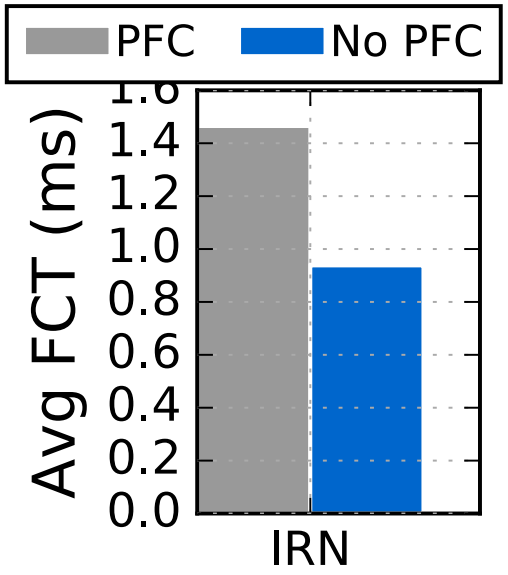


Average flow completion times

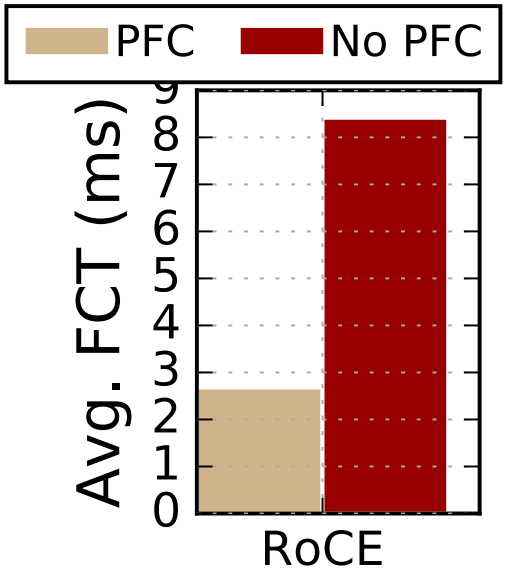
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.

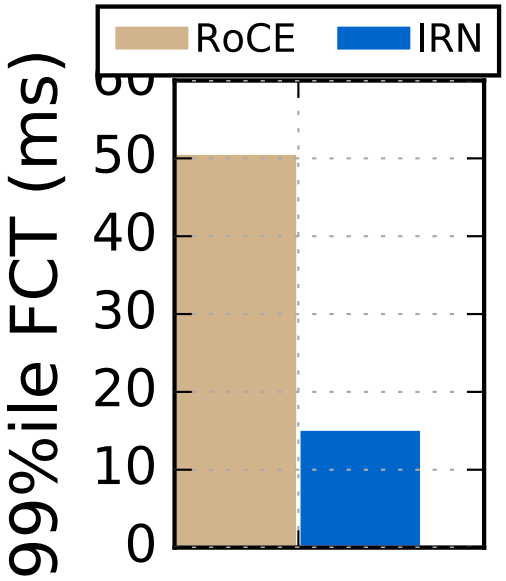


RoCE requires PFC.

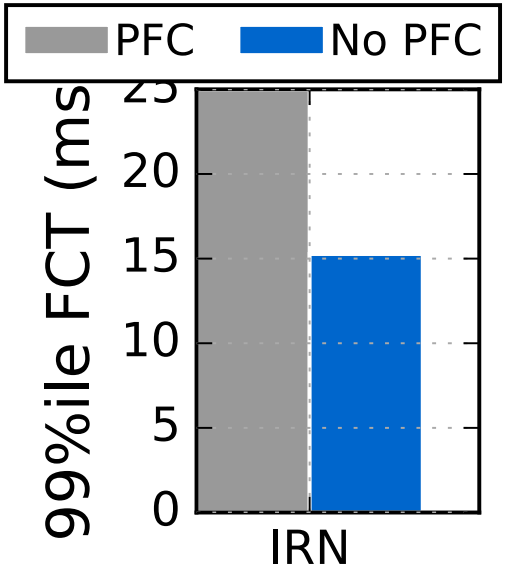


Tail flow completion times

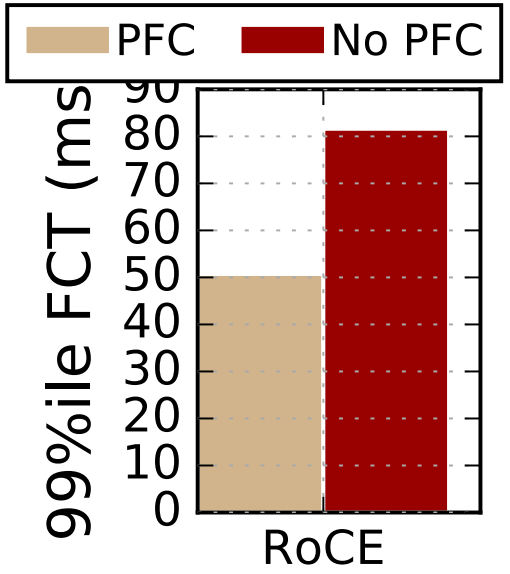
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.

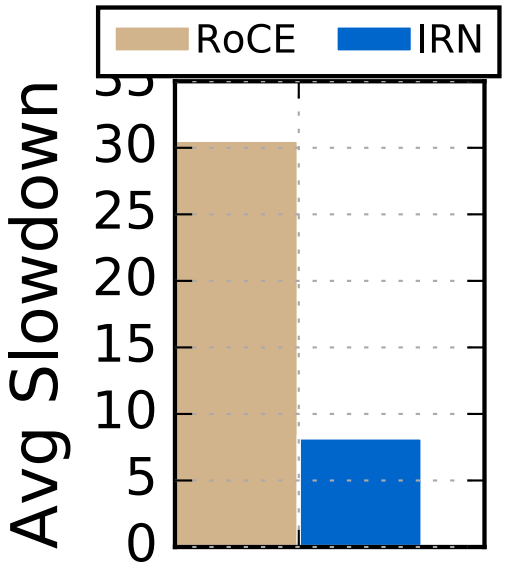


RoCE requires PFC.

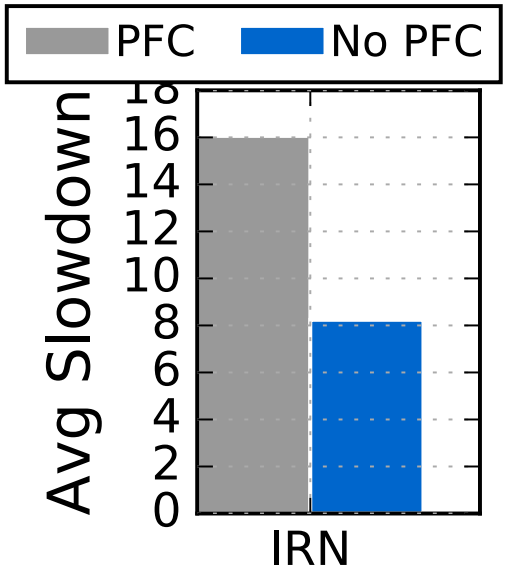


Average slowdown

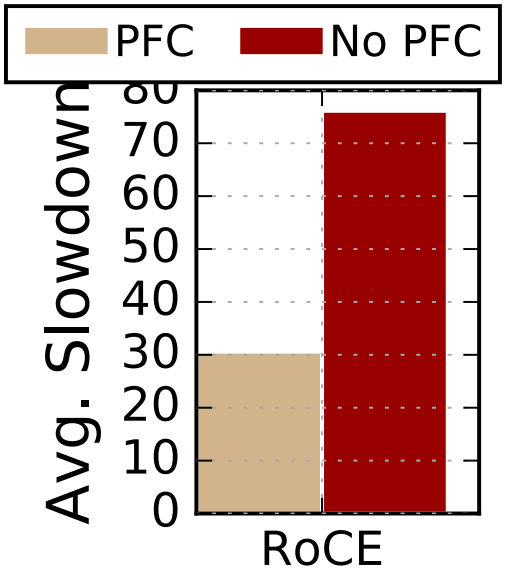
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.

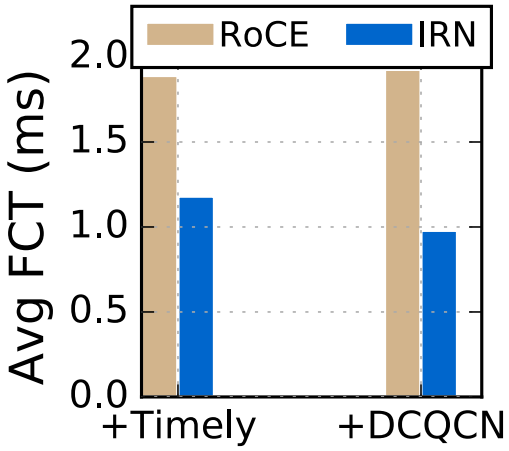


RoCE requires PFC.

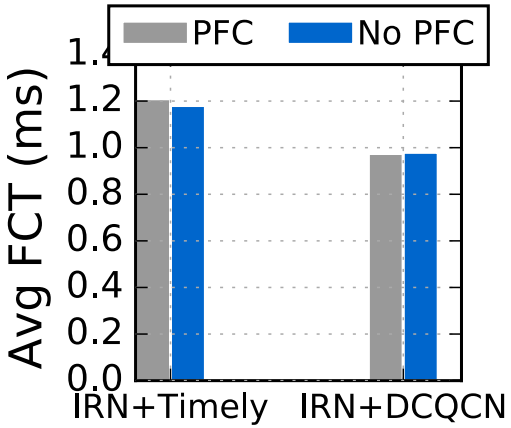


With explicit congestion control

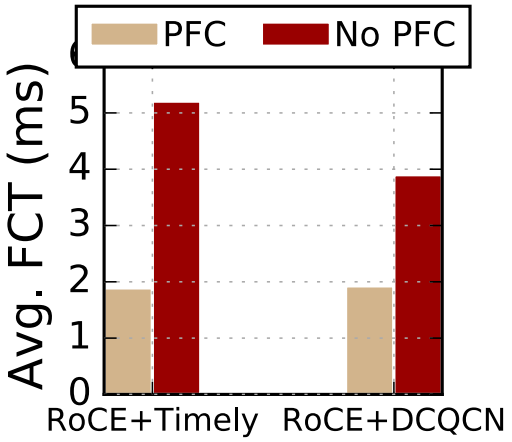
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.

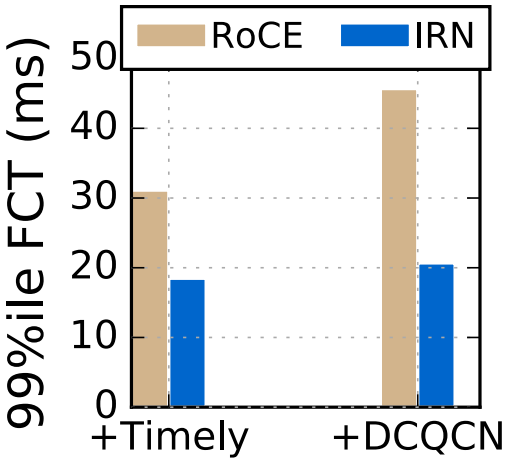


RoCE requires PFC.

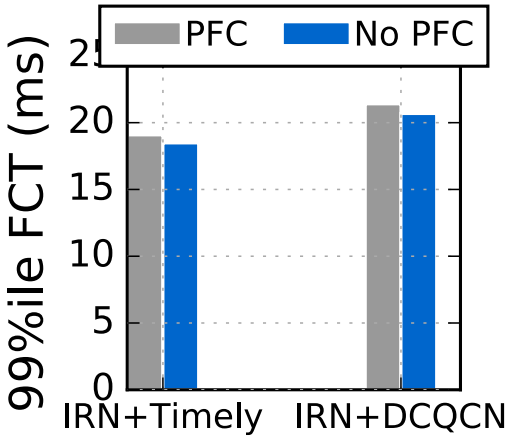


With explicit congestion control

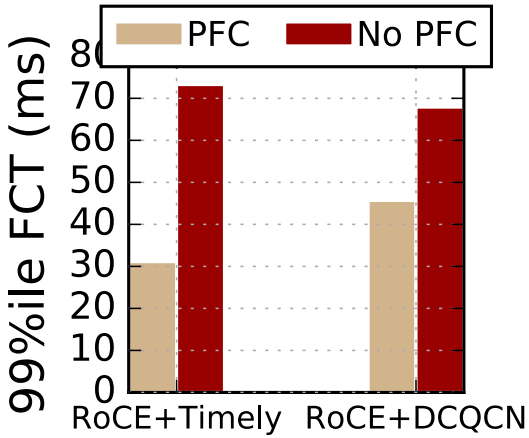
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.

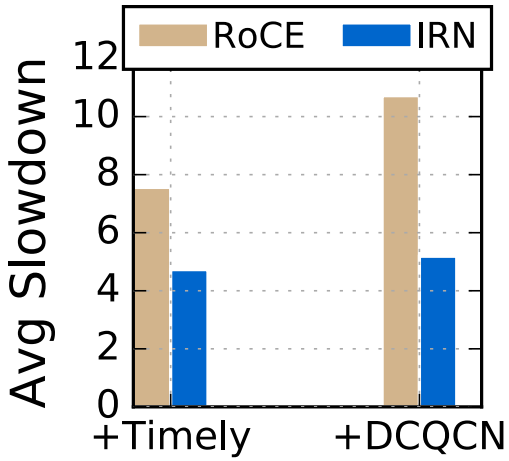


RoCE requires PFC.

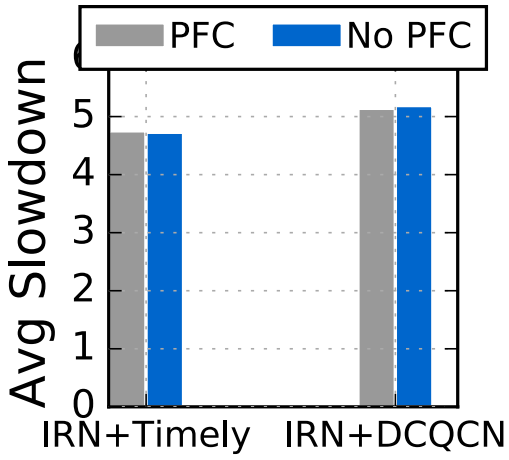


With explicit congestion control

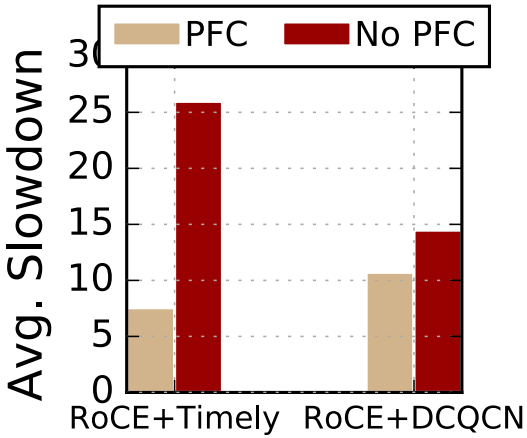
IRN *without PFC* performs better than RoCE *with PFC.*



IRN does not require PFC.



RoCE requires PFC.



Robustness of results

- Tested a wide range of experimental scenarios:
  - Varying link bandwidth.
  - Varying workload.
  - Varying scale of the topology.
  - Varying link utilization.
  - Varying buffer size.
  - …

- Our key takeaways hold across all of these scenarios.

Can IRN eliminate the need for a lossless network? Yes.

Can IRN be implemented easily?

Implementation challenges

- Need to deal with out-of-order packet arrivals.
  - Crucial information in first packet of the message.
    - Replicate in other packets.

Implementation challenges

- Need to deal with out-of-order packet arrivals.
  - Crucial information in first packet of the message.
    - Replicate in other packets.
  - Crucial information in last packet of the message.
    - Store it at the end-points.

Implementation challenges

- Need to deal with out-of-order packet arrivals.
  - Crucial information in first packet of the message.
    - Replicate in other packets.
  - Crucial information in last packet of the message.
    - Store it at the end-points.
  - Implicit matching between packet and work queue element (WQE).
    - Explicitly carry WQE sequence in packets.

Implementation challenges

- Need to deal with out-of-order packet arrivals.
  - Crucial information in first packet of the message.
    - Replicate in other packets.
  - Crucial information in last packet of the message.
    - Store it at the end-points.
  - Implicit matching between packet and work queue element (WQE).
    - Explicitly carry WQE sequence in packets.

- Need to explicitly send Read Acks.

Implementation overheads

- New packet types and header extensions.
  - Upto 16 bytes.
- Total memory overhead of 3-10%.
- FPGA synthesis targeting the device on an RDMA NIC.
  - Less than 4% resource usage.
  - 45.45Mpps throughput (without pipelining).

Can IRN eliminate the need for a lossless network? Yes.

Can IRN be implemented easily? Yes.

Summary

- IRN makes incremental updates to the RoCE NIC design to handle packet losses better.
- IRN performs better than RoCE without requiring a lossless network.
- The changes required by IRN introduce minor overheads.

Contact: radhika@eecs.berkeley.edu

Code: http://netsys.github.io/irn-vivado-hls/

Thank You!