Networking Preview Session
OSDI/ATC 2021

Akshay Narayan, MIT CSAIL

Sessions

“Peeking over the Fence: RDMA”
Wednesday July 14, 8:45 PT (Track 1)

Naos: Serialization-free RDMA networking in Java

One-sided RDMA-Conscious Extendible Hashing for Disaggregated Memory

Characterizing and Optimizing Remote Persistent Memory with RDMA and NVM

MigrOS: Transparent Live-Migration Support for Containerised RDMA Applications

“I Can Smell That Fluffy Was Here: Networks”
Thursday July 15, 8:45 PT (Track 1)

Hashing Linearity Enables Relative Path Control in Data Centers

Live in the Express Lane

Understanding Precision Time Protocol in Today’s Wi-Fi Networks: A Measurement Study

AUTO: Adaptive Congestion Control Based on Multi-Objective Reinforcement Learning for the Satellite-Ground Integrated Network

Hey, Lumi! Using Natural Language for Intent-Based Network Management

Goals


Understand the technologies the papers use


Understand the types of problems the papers might want to solve


Learn what kind of questions you might want to ask the authors about their work


Paper details or motivations (go listen to the talks!)


Explanation for vaguely canine session names

Sessions

“Peeking over the Fence: RDMA”

Naos: Serialization-free RDMA networking in Java

One-sided RDMA-Conscious Extendible Hashing for Disaggregated Memory

Characterizing and Optimizing Remote Persistent Memory with RDMA and NVM

MigrOS: Transparent Live-Migration Support for Containerised RDMA Applications

“I Can Smell That Fluffy Was Here: Networks”

Hashing Linearity Enables Relative Path Control in Data Centers

Live in the Express Lane

Understanding Precision Time Protocol in Today’s Wi-Fi Networks: A Measurement Study

AUTO: Adaptive Congestion Control Based on Multi-Objective Reinforcement Learning for the Satellite-Ground Integrated Network

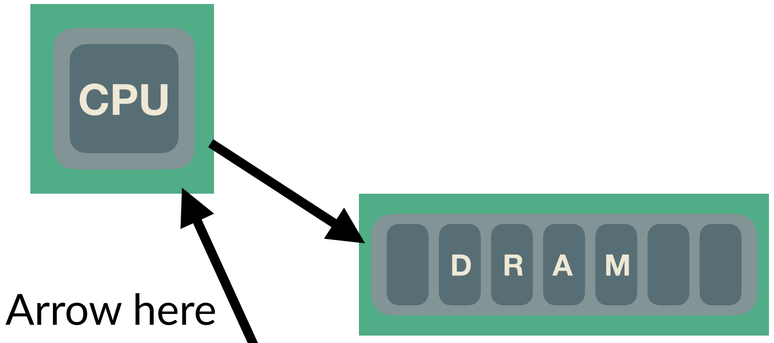
Hey, Lumi! Using Natural Language for Intent-Based Network Management

What is RDMA?

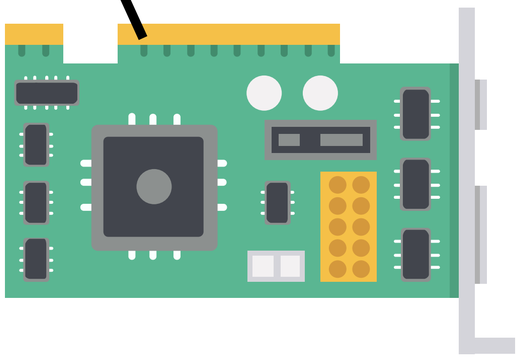
"Remote, Direct Memory Access"

Normal
(Sockets, etc)

CPU processes packets from the NIC (e.g., in the kernel)

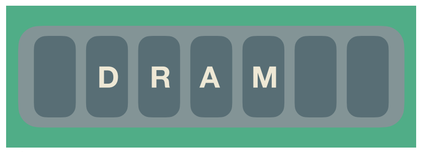


Arrow here makes things indirect

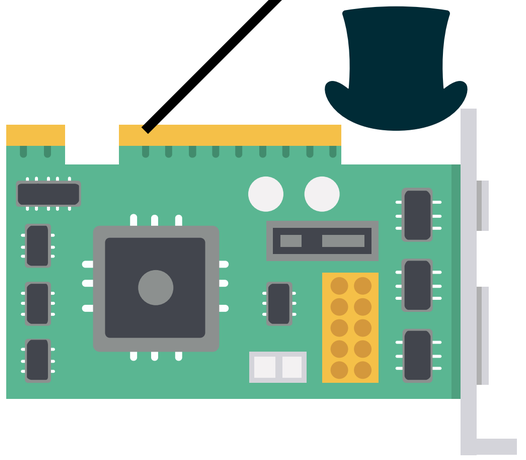


RDMA

CPU spends no cycles on processing packets

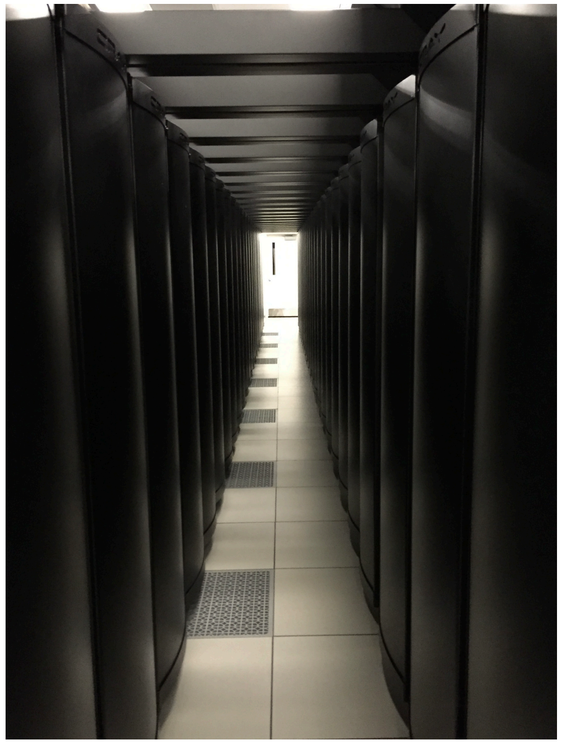


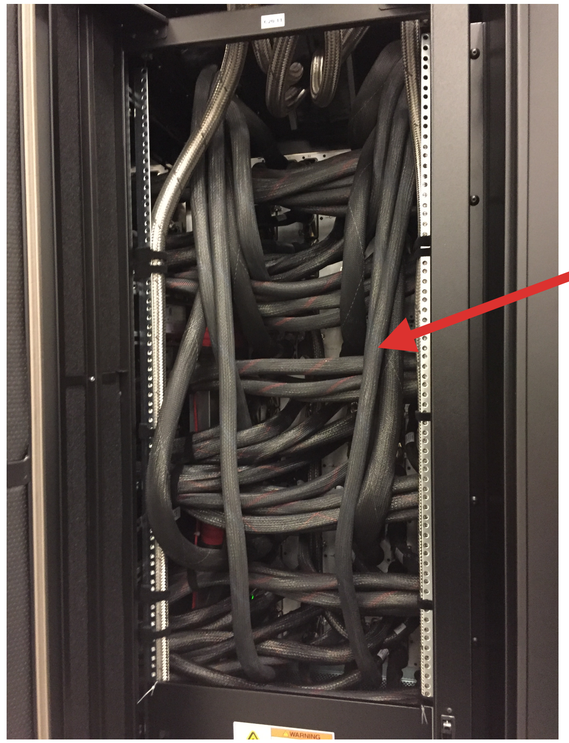
Fancy NIC (with hat)



NIC Image from @graziaprato via Justine Sherry

RDMA Origin Story





Special interconnect: Infiniband/RoCE

Special API: e.g. libibverbs

RDMA over Commodity Ethernet at Scale

Chuanxiong Guo, Haitao Wu, Zhong Deng, Gaurav Soni,
Jianxi Ye, Jitendra Padhye, Marina Lipshteyn
Microsoft
{chguo, hwu, zdeng, gasoni, jiye, padhye, malipsht}@microsoft.com

ABSTRACT

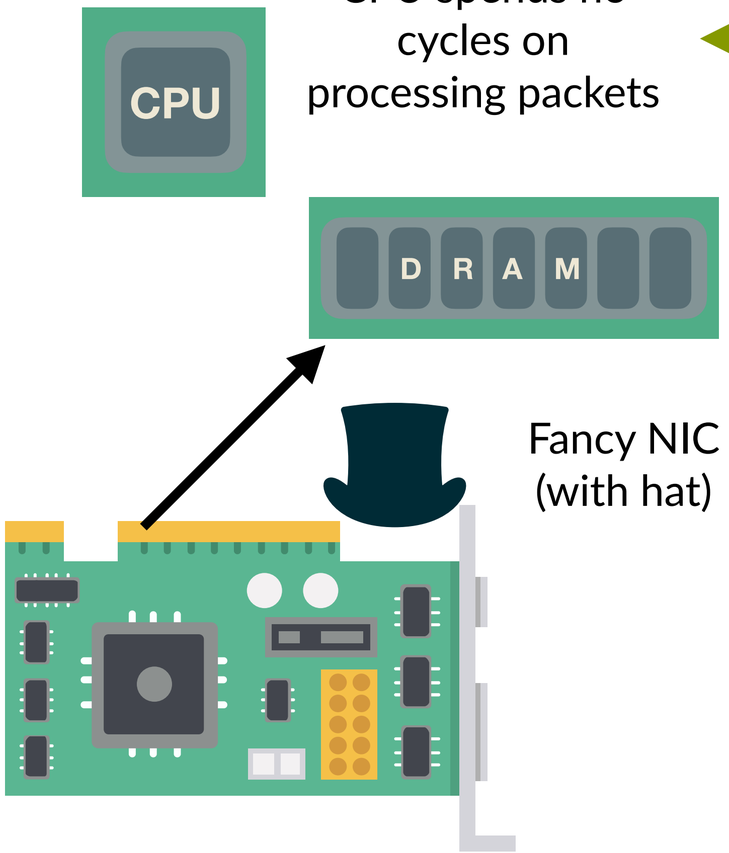
Over the past one and half years, we have been using RDMA over commodity Ethernet (RoCEv2) to support some of Microsoft's highly-reliable, latency-sensitive services. This paper describes the challenges we encountered during the process and the solutions we devised to address them. In order to scale RoCEv2 beyond VLAN, we have designed a DSCP-based priority flow-control (PFC) mechanism to ensure large-scale deployment. We

Ethernet switches and network interface cards (NICs). A state-of-the-art DCN must support several Gb/s or higher throughput between any two servers in a DC.

TCP/IP is still the dominant transport/network stack in today's data center networks. However, it is increasingly clear that the traditional TCP/IP stack cannot meet the demands of the new generation of DC workloads [4, 9, 16, 40], for two reasons.

First, the CPU overhead of handling packets in the

Why RDMA?



CPU spends no cycles on processing packets

We can sell these CPU cycles to someone else! $$$

Fancy NIC (with hat)

Fancy NIC can handle the packet processing for us

$$

RDMA over Commodity Ethernet at Scale

Chuanxiong Guo, Haitao Wu, Zhong Deng, Gaurav Soni,
Jianxi Ye, Jitendra Padhye, Marina Lipshteyn
Microsoft
{chguo, hwu, zdeng, gasoni, jiye, padhye, malipsht}@microsoft.com

ABSTRACT

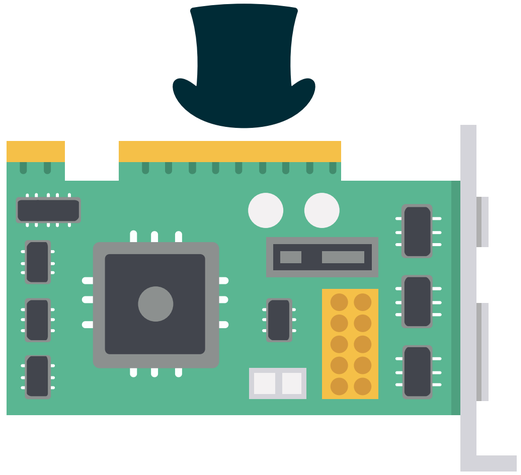
Over the past one and half years, we have been using RDMA over commodity Ethernet (RoCEv2) to support some of Microsoft's highly-reliable, latency-sensitive services. This paper describes the challenges we encountered during the process and the solutions we devised to address them. In order to scale RoCEv2 beyond VLAN, we have designed a DSCP-based priority flow-control (PFC) mechanism to ensure large-scale deployment. We

Ethernet switches and network interface cards (NICs). A state-of-the-art DCN must support several Gb/s or higher throughput between any two servers in a DC.

TCP/IP is still the dominant transport/network stack in today's data center networks. However, it is increasingly clear that the traditional TCP/IP stack cannot meet the demands of the new generation of DC workloads [4, 9, 16, 40], for two reasons.

First, the CPU overhead of handling packets in the

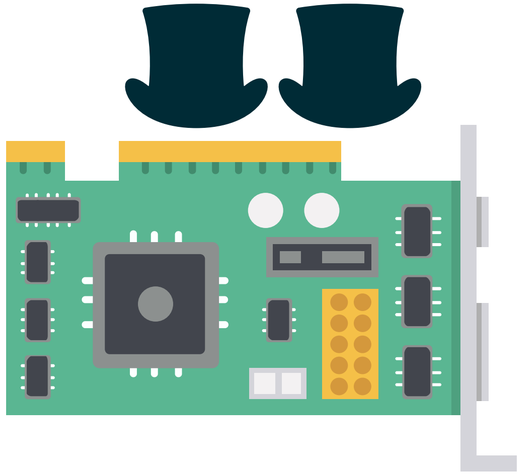
RDMA Choices



Mellanox TECHNOLOGIES

"[iWARP's] convoluted architecture is an ill-conceived attempt to fit RDMA into existing software transport frameworks."

https://www.mellanox.com/related-docs/whitepapers/WP_RoCE_vs_iWARP.pdf

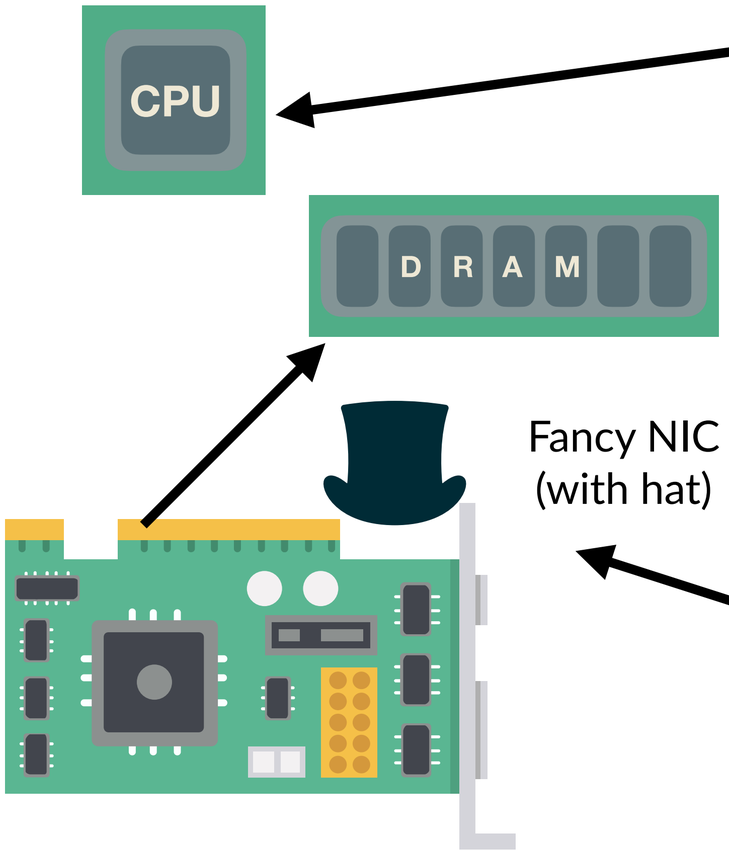


Chelsio Communications Accelerate

"RoCE doesn't scale. High performance iWARP implementations are available and compete directly with InfiniBand in real application benchmarks. iWARP allows use of existing hardware and lives alongside existing applications."

https://www.chelsio.com/roce/

RDMA Questions



Kernel functionality?
Application API issues?

Fancy NIC
(with hat)

In what scenarios is the
hassle worth it?

Or, can we make using
RDMA less painful?

How fancy should the NIC be?
How does RDMA interact with other HW trends?

Sessions

“Peeking over the Fence: RDMA”

Naos: Serialization-free RDMA networking in Java

One-sided RDMA-Conscious Extendible Hashing for Disaggregated Memory

Characterizing and Optimizing Remote Persistent Memory with RDMA and NVM

MigrOS: Transparent Live-Migration Support for Containerised RDMA Applications

“I Can Smell That Fluffy Was Here: Networks”

Hashing Linearity Enables Relative Path Control in Data Centers

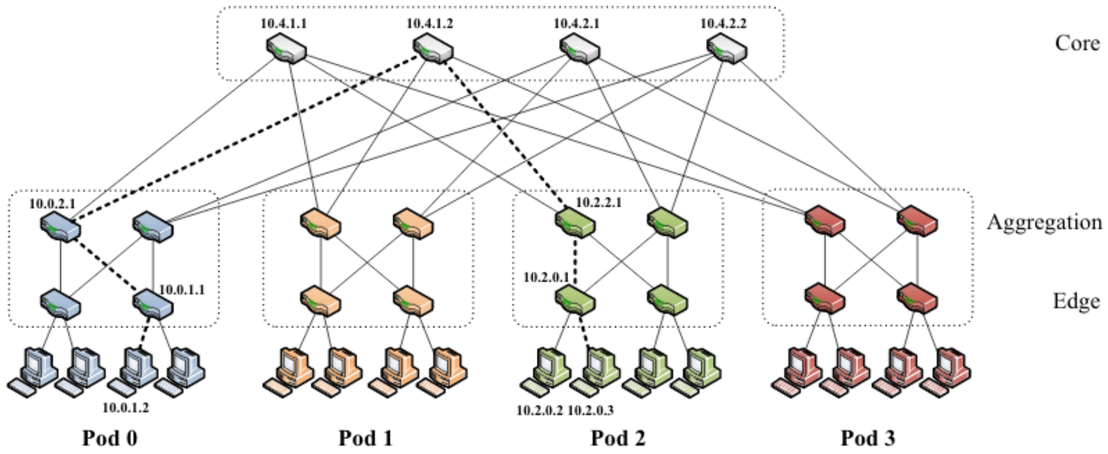
Live in the Express Lane

Understanding Precision Time Protocol in Today’s Wi-Fi Networks: A Measurement Study

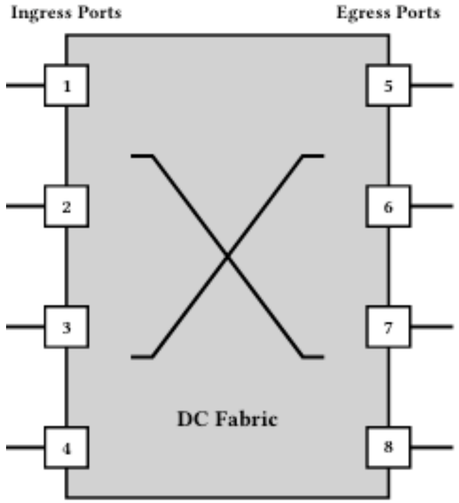
AUTO: Adaptive Congestion Control Based on Multi-Objective Reinforcement Learning for the Satellite-Ground Integrated Network

Hey, Lumi! Using Natural Language for Intent-Based Network Management

The Dream: One Big Switch



Conventional design: fat-tree network

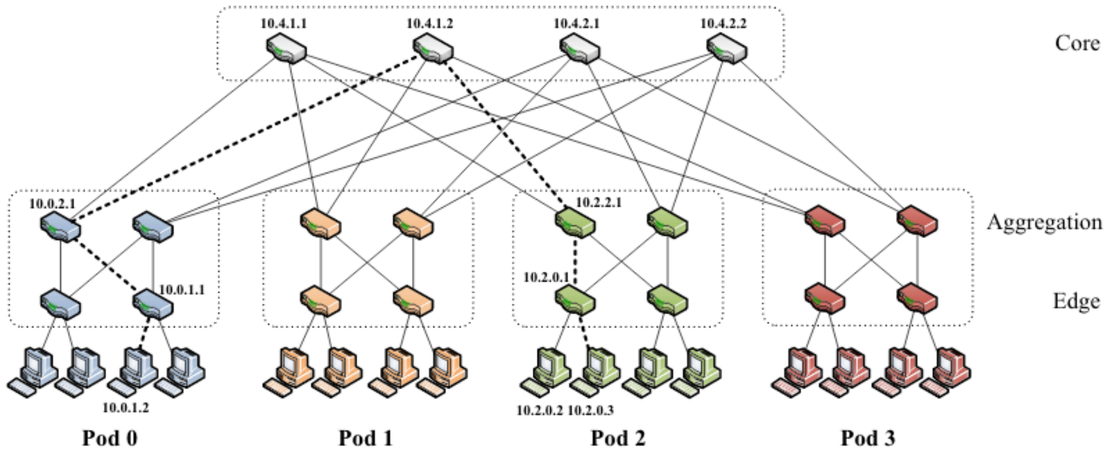


What users want: “One Big Switch”

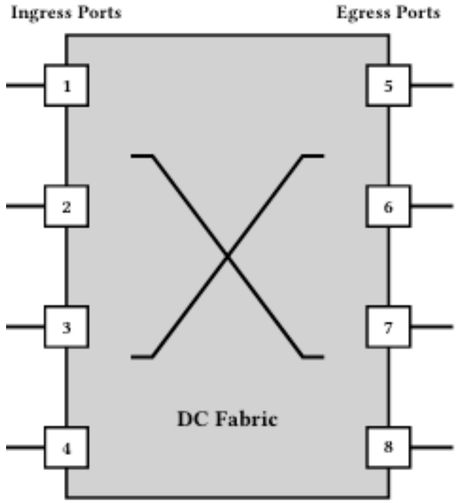
Image from M. Al-Fares et al., SIGCOMM 2008

Image from S. Agarwal et al., SIGCOMM 2018

The Dream: One Big Switch



Conventional design: fat-tree network



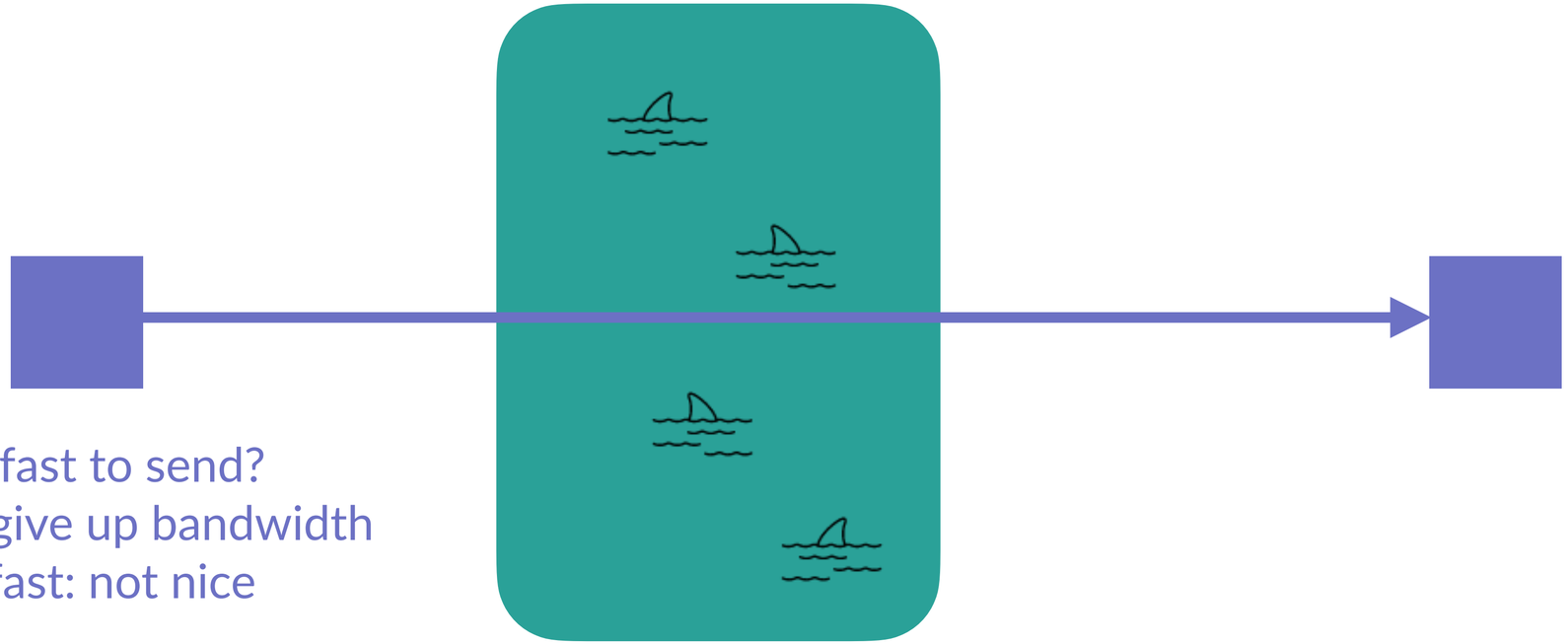
What users want: “One Big Switch”

Complexity? Efficiency?
Headaches for Apps?
Impacts on latency?

Image from M. Al-Fares et al., SIGCOMM 2008

Image from S. Agarwal et al., SIGCOMM 2018

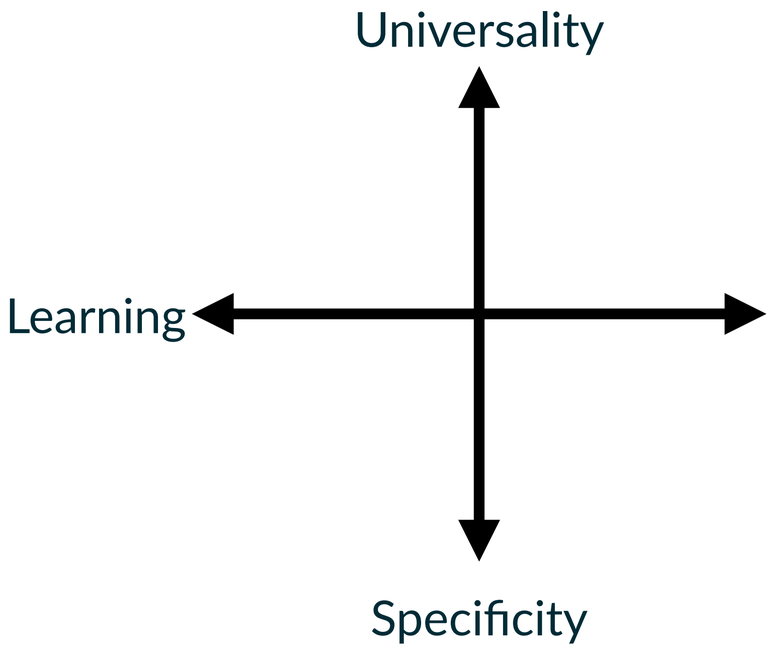
Congestion Control



How fast to send?
Too slow: give up bandwidth
Too fast: not nice

Internet: A Scary Place

Congestion Control



Congestion Control

Congestion-Control Throwdown

Michael Schapira
Hebrew University of Jerusalem
schapiram@huji.ac.il

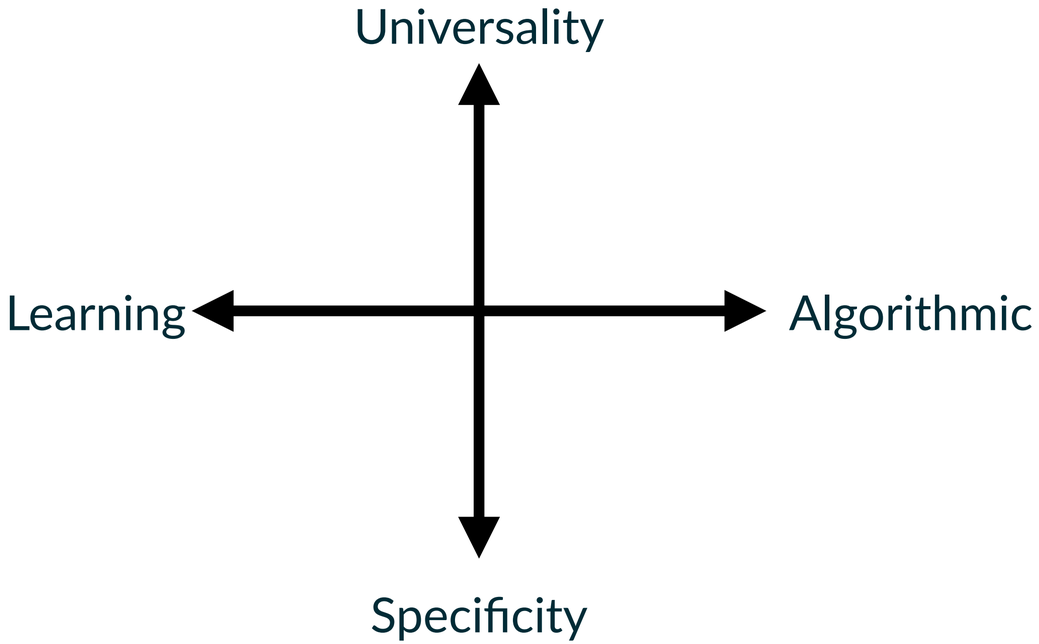
Keith Winstein
Stanford University
keithw@cs.stanford.edu

ABSTRACT

Congestion control is a perennial topic of networking research. In making decisions about *who* sends data *when*, congestion-control schemes prevent collapses and ultimately determine the allocation of scarce communications resources among contending users and applications.

The field has seen considerable recent activity. Even after three decades of research, basic principles and techniques remain up for debate. In this throwdown-as-paper, the authors find themselves at loggerheads over the fundamental tenets of congestion control.

bottleneck links, and also the designer's *global optimization* objective, say, proportional fairness. Remy then generates a *model of the network* and seeks a "good" mapping from observed network state (average of packet ACKs inter-arrival times, ratio of current RTT and minRTT, etc.) to control actions (such as a multiplier/increment to the congestion window). BBR's design philosophy is different; BBR models the network pipe as a single link, repeatedly probes the bandwidth and RTT, and paces the rate so as to track the bottleneck link's bandwidth. Lastly, PCC continuously associates the sending rate with a numerical *utility* value that reflects a *local* performance objective (say, "high throughput and low



Modelling assumptions?

Sessions

“Peeking over the Fence: RDMA”
Wednesday July 14, 8:45 PT (Track 1)

Naos: Serialization-free RDMA networking in Java

One-sided RDMA-Conscious Extendible Hashing for Disaggregated Memory

Characterizing and Optimizing Remote Persistent Memory with RDMA and NVM

MigrOS: Transparent Live-Migration Support for Containerised RDMA Applications

“I Can Smell That Fluffy Was Here: Networks”
Thursday July 15, 8:45 PT (Track 1)

Hashing Linearity Enables Relative Path Control in Data Centers

Live in the Express Lane

Understanding Precision Time Protocol in Today’s Wi-Fi Networks: A Measurement Study

AUTO: Adaptive Congestion Control Based on Multi-Objective Reinforcement Learning for the Satellite-Ground Integrated Network

Hey, Lumi! Using Natural Language for Intent-Based Network Management

akshayn@csail.mit.edu