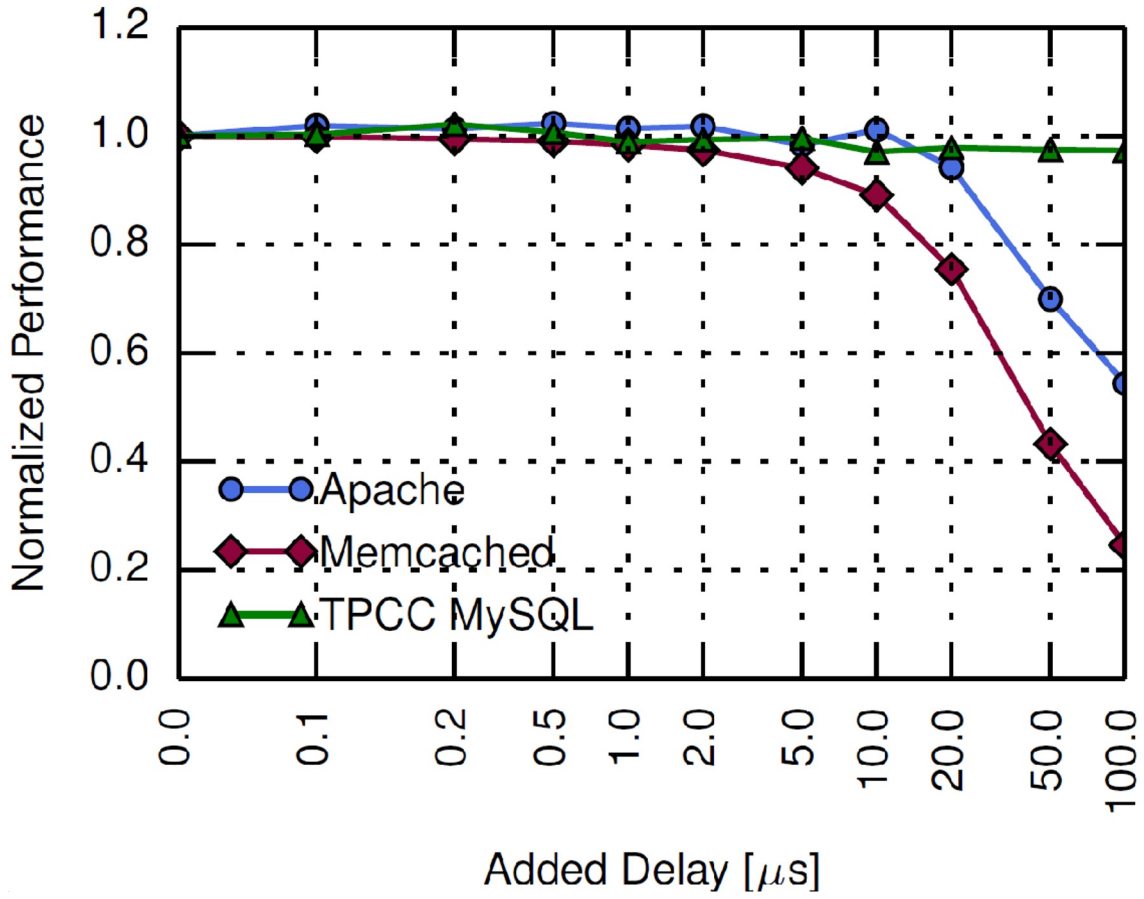
UNIVERSITY OF CAMBRIDGE

Where Has My Time Gone?

Noa Zilberman, Matthew Grosvenor, Diana Andreea Popescu, Neelakandan Manihatty-Bojan, Gianni Antichi, Marcin Wójcik, Andrew W. Moore

Latency matters



It's Time For Low Latency + Low Variance!

2011

It's Time for Low Latency

Stephen M. Rumble, Diego Ongaro, Ryan Stutsman,
Mendel Rosenblum, and John K. Ousterhout
Stanford University

ommunity has ignored network
the past, speed-of-light delays
and unoptimized network hard-
)μs round-trip times impossible.
w years datacenters will be de-
Ethernet. Without the burden
the datacenter campus and net-
rnet devices, it will be up to us
this benefit through to applica-
researchers must lead the charge
s to push the boundaries of low-
nunicati

	1983	2011	Improved
CPU Speed	1x10Mhz	4x3GHz	> 1,000x
Memory Size	≤ 2MB	8GB	≥ 4,000x
Disk Capacity	≤ 30MB	2TB	> 60,000x
Net Bwidth	3Mbps	10Gbps	> 3,000x
RTT	2.54ms	80μs	32x

Table 1: Network latency has improved far more slowly over the last three decades than other performance metrics for commodity computers. The V Distributed System [5] achieved round-trip RPC times of 2.54ms. Today, a pair of modern Linux servers require 80μs for 16-byte RPCs over TCP with 10Gb Ethernet.

Component	Delay	Round-Trip
Network Switch	10-30μs	100-300μs
Network Interface Card	2.5-32μs	10-128μs
OS Network Stack	15μs	60μs
Speed of Light (in Fiber)	5ns/m	0.6-1.2μs

Table 2: Factors that contribute to latency in TCP datacenter communication. "Delay" indicates the cost of a single traversal of the component, and "Round-Trip" indicates the total impact on round-trip time. Messages typically traverse 5 switches in each direction in a large datacenter network and must pass through the OS stack 4 times.

2016

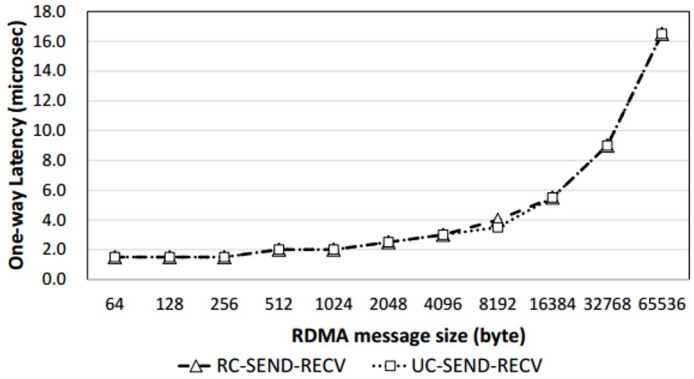


Fig. 4: Median one-way latency of RoCE RC and UC transport types.

"Exploring Low-latency Interconnect for Scaling Out Software Routers", Ma, Kim, and Moon

"Design Guidelines for High Performance RDMA Systems" Kalia, Kaminsky and Andersen

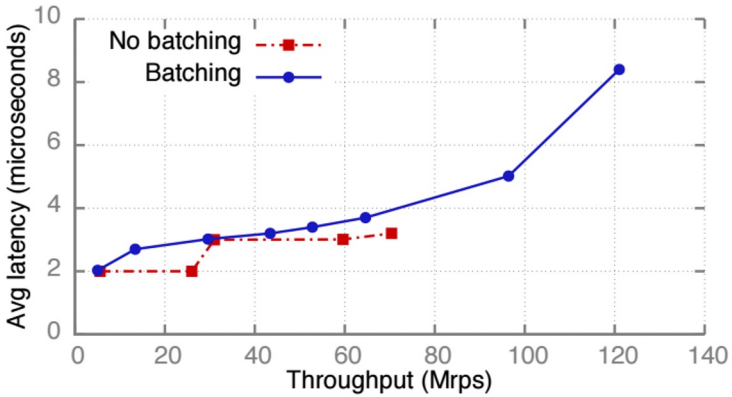


Figure 8: Impact of response batching on `Spec-S0` latency

It's Time For Low Latency + Low Variance!

2011

It's Time for Low Latency

Stephen M. Rumble, Diego Ongaro, Ryan Stutsman,
Mendel Rosenblum, and John K. Ousterhout
Stanford University

ommunity has ignored network
the past, speed-of-light delays
and unoptimized network hard-
)μs round-trip times impossible.
w years datacenters will be de-
Ethernet. Without the burden
the datacenter campus and net-
rnet devices, it will be up to us
this benefit through to applica-
researchers must lead the charge
s to push the boundaries of low-
nunicati

	1983	2011	Improved
CPU Speed	1x10Mhz	4x3GHz	> 1,000x
Memory Size	≤ 2MB	8GB	≥ 4,000x
Disk Capacity	≤ 30MB	2TB	> 60,000x
Net Bwidth	3Mbps	10Gbps	> 3,000x
RTT	2.54ms	80μs	32x

Table 1: Network latency has improved far more slowly over the last three decades than other performance metrics for commodity computers. The V Distributed System [5] achieved round-trip RPC times of 2.54ms. Today, a pair of modern Linux servers require 80μs for 16-byte RPCs over TCP with 10Gb Ethernet.

Component	Delay	Round-Trip
Network Switch	10-30μs	100-300μs
Network Interface Card	2.5-32μs	10-128μs
OS Network Stack	15μs	60μs
Speed of Light (in Fiber)	5ns/m	0.6-1.2μs

Table 2: Factors that contribute to latency in TCP datacenter

2016

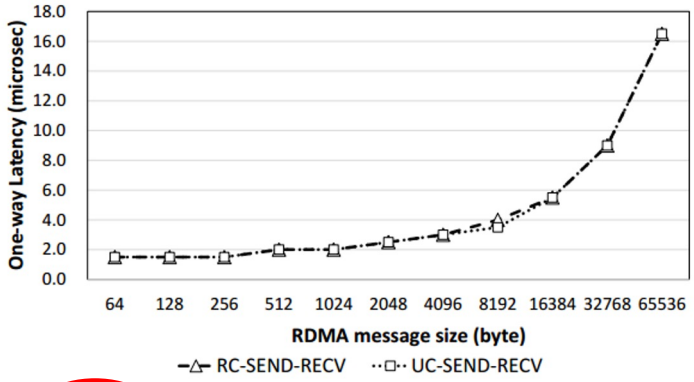


Fig. 4: Median one-way latency of RoCE RC and UC transport types

"Exploring Low-latency Interconnect for Scaling Out Software Routers", Ma, Kim, and Moon

"Design Guidelines for High Performance RDMA Systems" Kalia, Kaminsky and Andersen

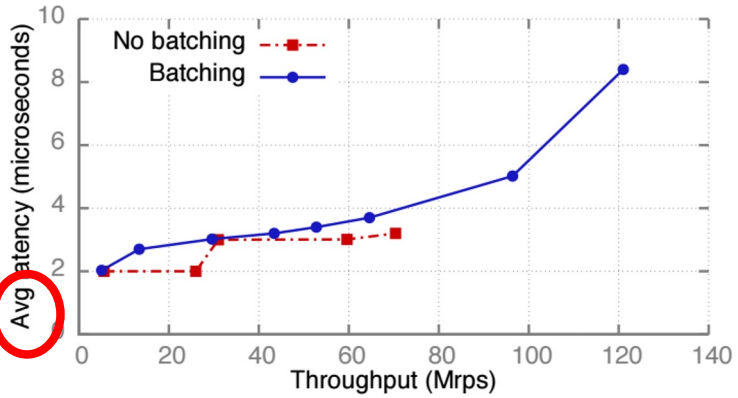


Figure 8: Impact of response batching on `Spec-S0` latency

It Usually Works

The Unavoidable Latency

I will talk about:

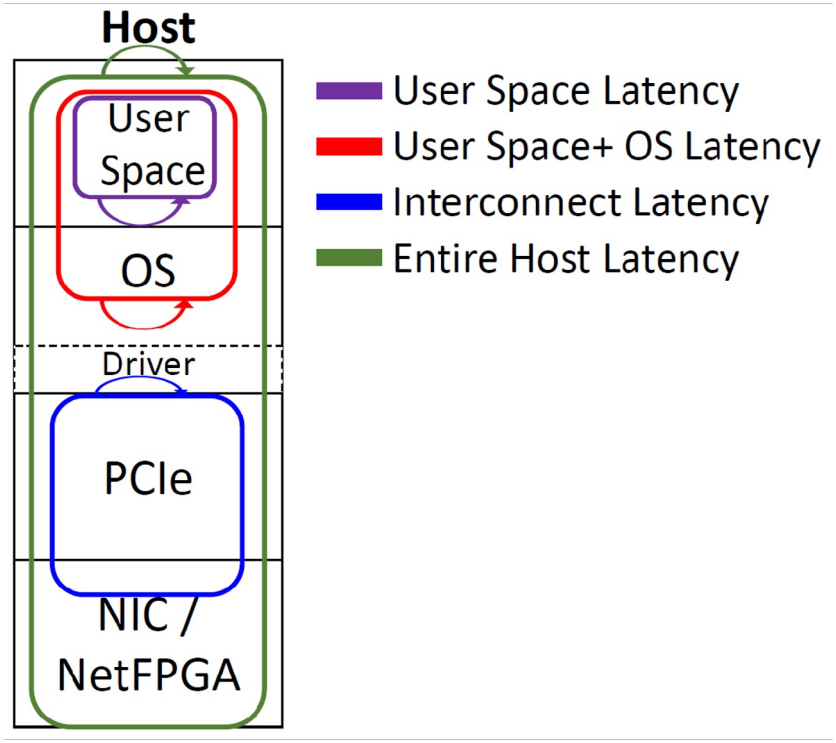
- The **essential** latency contributions
- **Commodity** hardware, standard coding
- Set a (reproducible) **baseline** for research

I will not talk about:

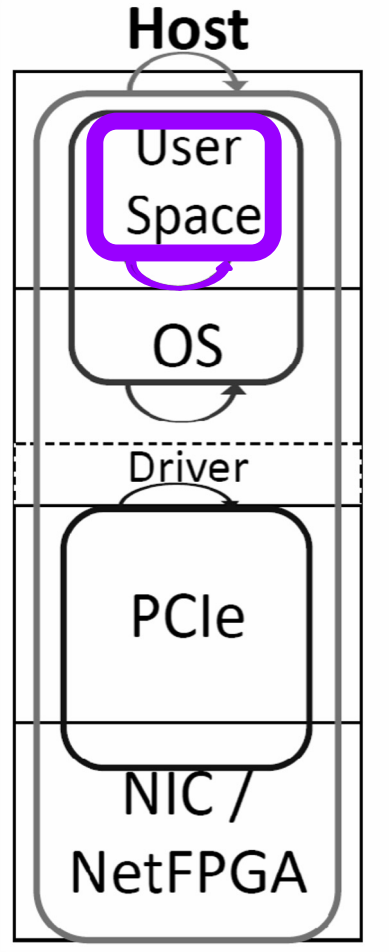
- TCP, DCTCP, MPTCP etc.
- Congestion, Buffer bloat, In-cast, etc.

Methodology

- Breakdown overall system latency into **component contributions**
- Use decompositional analysis to report component distributions e.g. min, median, 99.9th, max

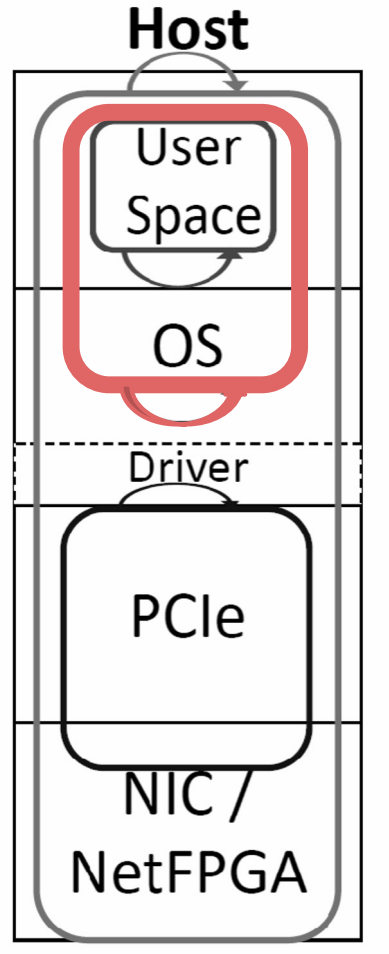


Time Stamp Counter (TSC) Measurement Accuracy



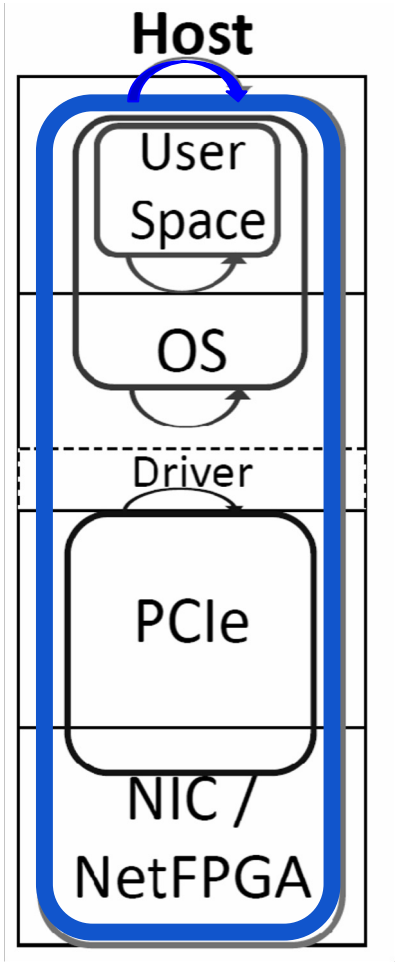
Experiment	Minimum	Median	99.9th	Tail
TSC - From User Space	9 ns	10 ns	11 ns	**49 µs**
TSC - Kernel	9 ns	9 ns	9 ns	**6.9 µs**
TSC - Kernel Early Boot	7 ns	7 ns	7 ns	**11 ns**
TSC - From VM User Space	12 ns	12 ns	13 ns	**64 ms**

A Breakdown of Basic Latency Components: User space + OS



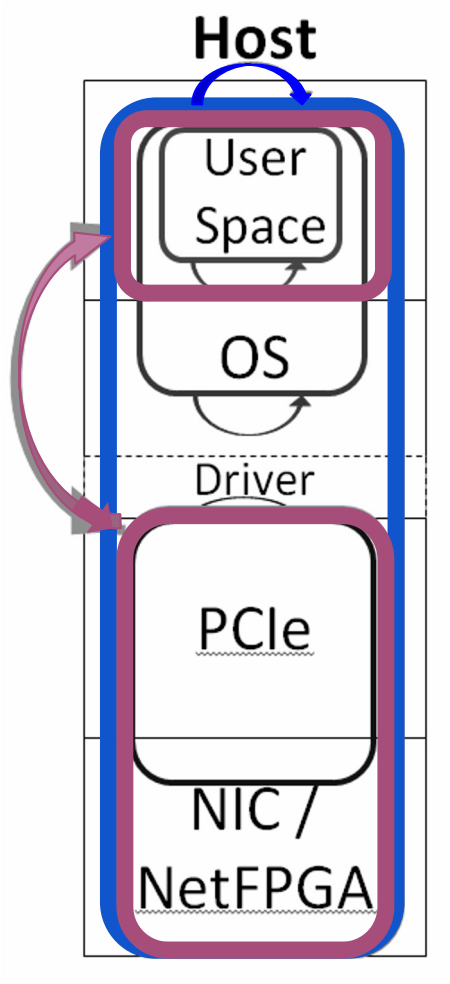
Experiment	Minimum	Median	99.9th	Tail
User space + OS (same core)	2 µs	2 µs	2 µs	68 µs
User space + OS (other core)	4 µs	5 µs	5 µs	31 µs

A Breakdown of Basic Latency Components: Host



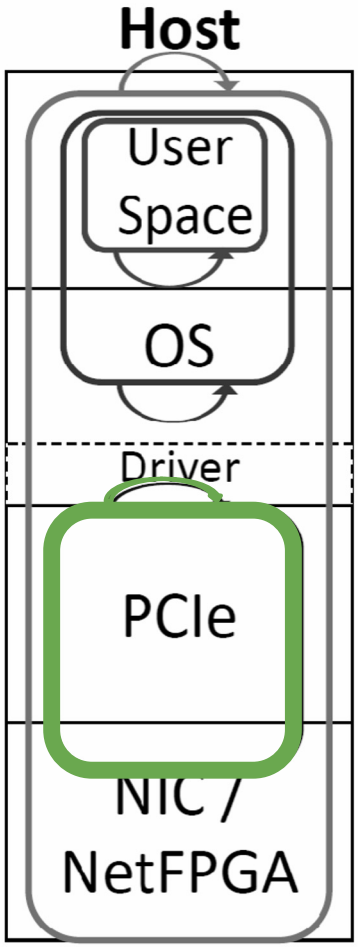
Experiment	Minimum	Median	99.9th	Tail
Host	3.9 µs	4.5 µs	21 µs	45 µs

A Breakdown of Basic Latency Components: Host



Experiment	Minimum	Median	99.9th	Tail
Host	3.9 µs	4.5 µs	21 µs	45 µs
Kernel Bypass	0.89 µs	0.94 µs	1.1 µs	5.4 µs

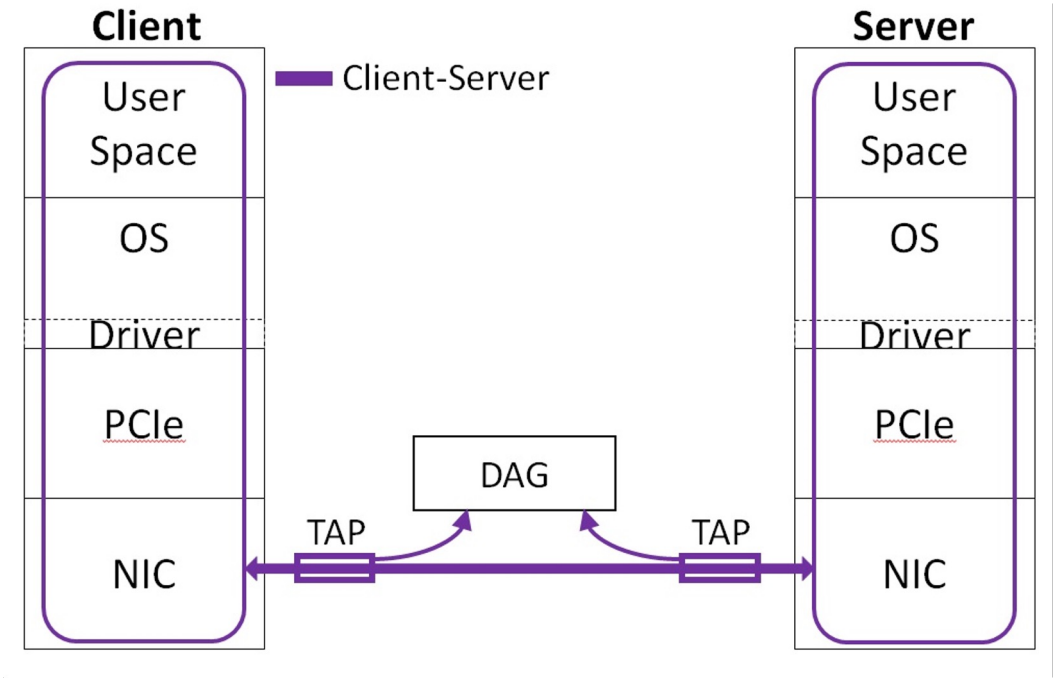
A Breakdown of Basic Latency Components: PCIe



Experiment	Minimum	Median	99.9th	Tail
Host	3.9 µs	4.5 µs	21 µs	45 µs
Kernel Bypass	0.89 µs	0.94 µs	1.1 µs	5.4 µs

Experiment	Minimum	Median	99.9th	Tail
PCIe Interconnect (64B)	**0.55 µs**	0.57 µs	0.59 µs	**0.6 µs**
PCIe Interconnect (1536B)	**0.97 µs**	0.98 µs	1.02 µs	**1.03 µs**

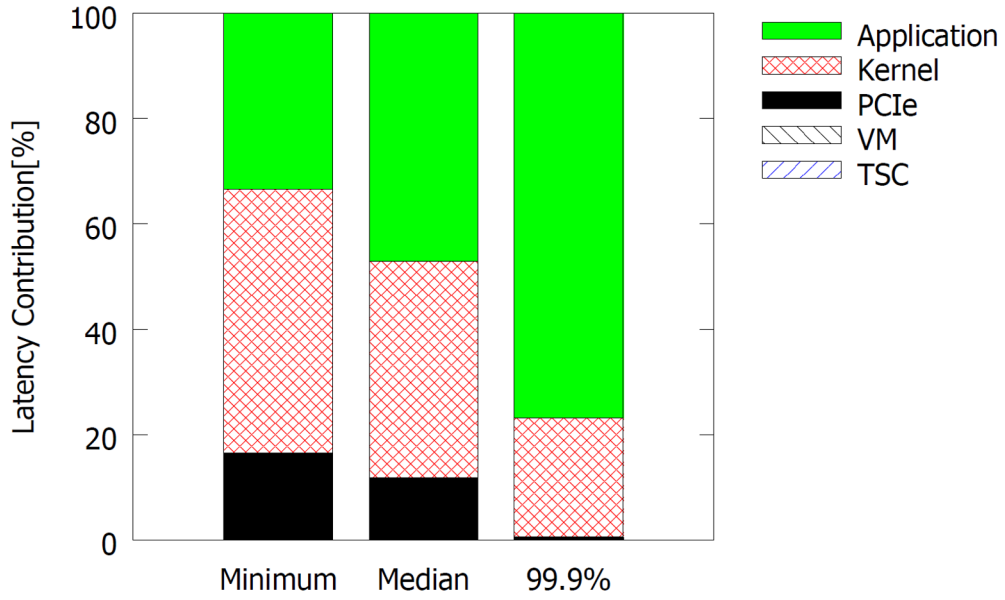
A Breakdown of Basic Latency Components: Client-Server



Experiment	Minimum	Median	99.9th	Tail
Client-Server (UDP)	7 µs	9 µs	107 µs	**203 µs**
Client-Server (Memcached)	10 µs	13 µs	240 µs	**20.3 ms**

Summary end host results

- All experiments bounded by application variance (TSC)
- PCIe latency variance is low

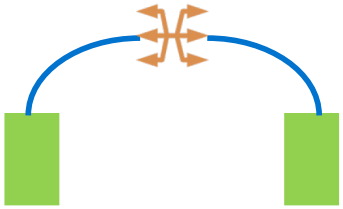


Basic Latency Components of the Network

Single Rack

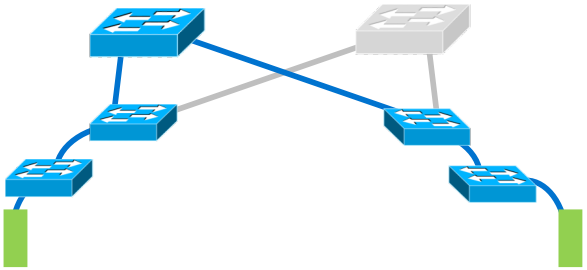
Cut Through ToR, 2m fibres, 10GE

HFT



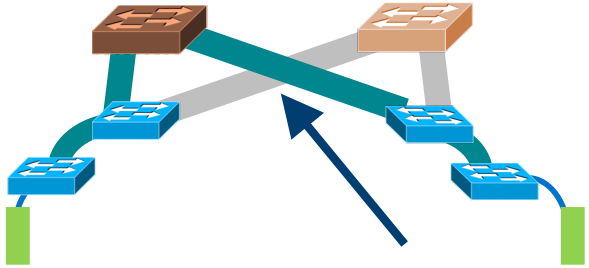
L1 *Switching*, 1.5m copper, 10GE

Fat tree

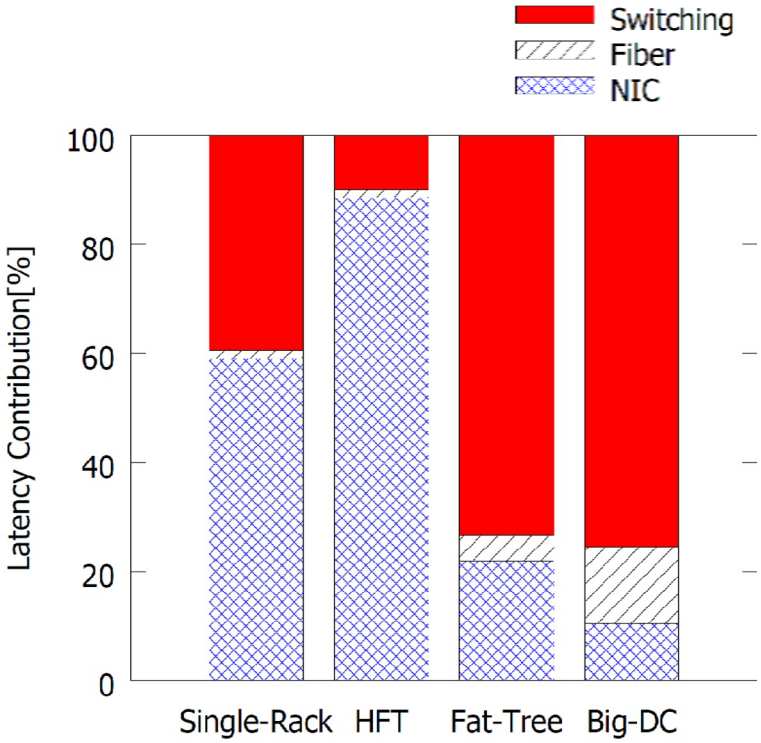


Cut Through, 2m, 5m & 10m fibres, 10GE

***Big DC* Fat tree**



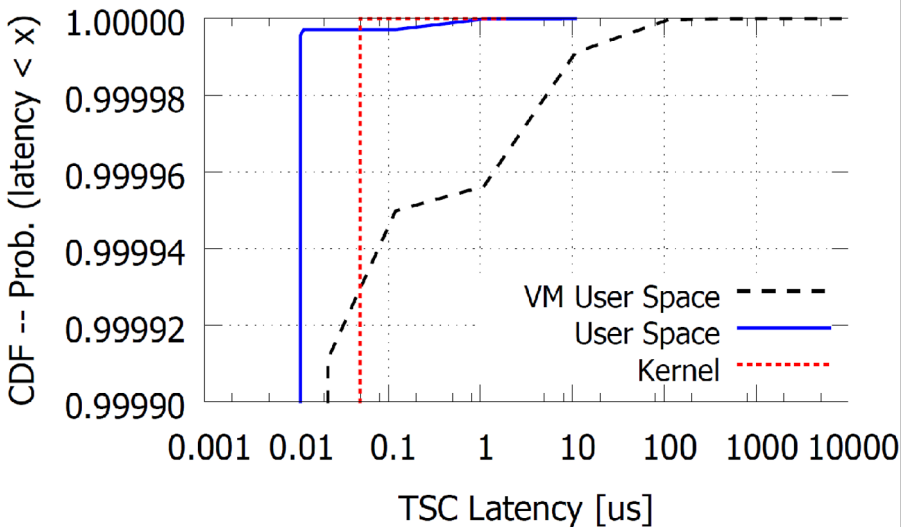
Store-forward spine, 100m /100GE fibres

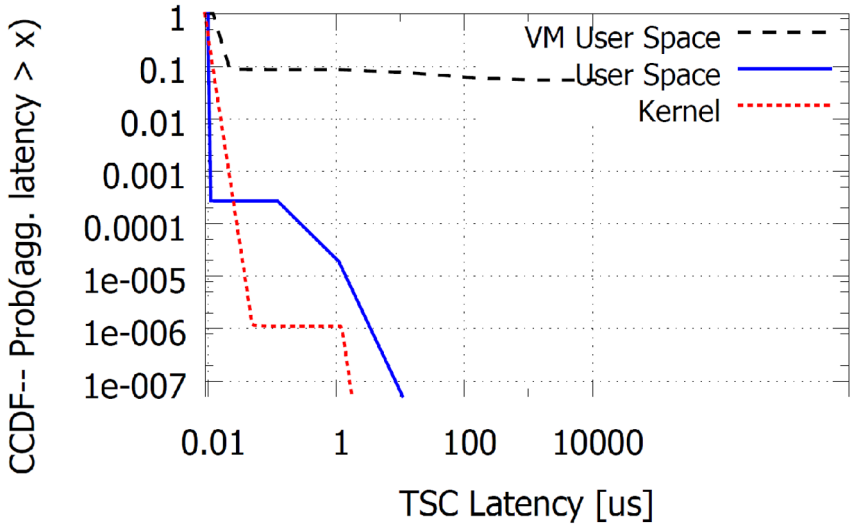


Based on median results

Tail Latency

- The long latency tail can take much more than you think
- For VM, events of 1ms or longer take almost 5% of the time.





The Good, The Bad and The Ugly

Good (≤1μs)

- Simple kernel and user space operations
- PCIe
- Single through-switch latency (no queueing)

Bad (1μs-100 μs)

- Sending packets over user space+OS
- Host latency and client-server latency
- Multi-stage network topology
- RTT over 100m fibers

Ugly (>100μs)

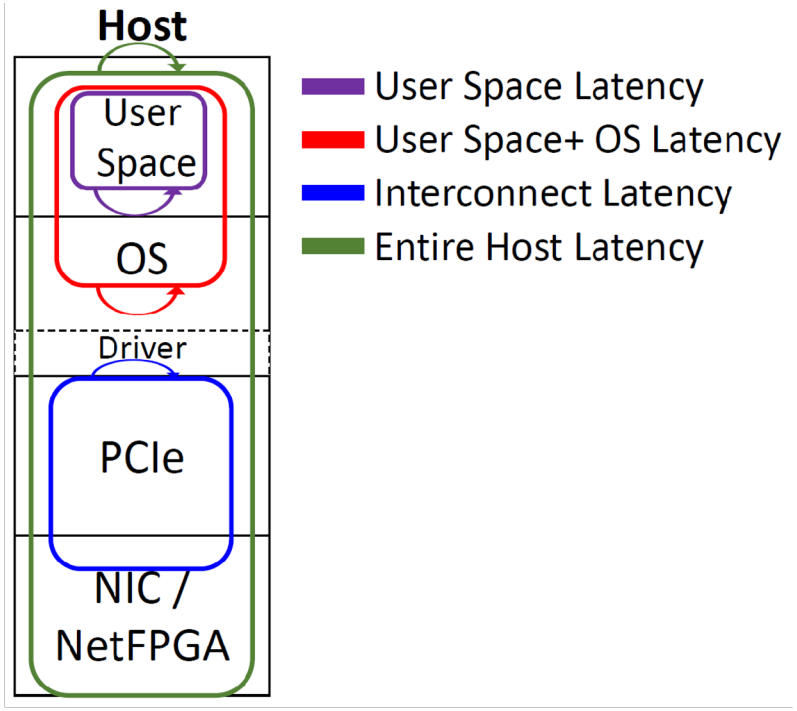
- The far end of the tail, i.e. the “variance”
- Mostly in user space and within a VM.

Summary

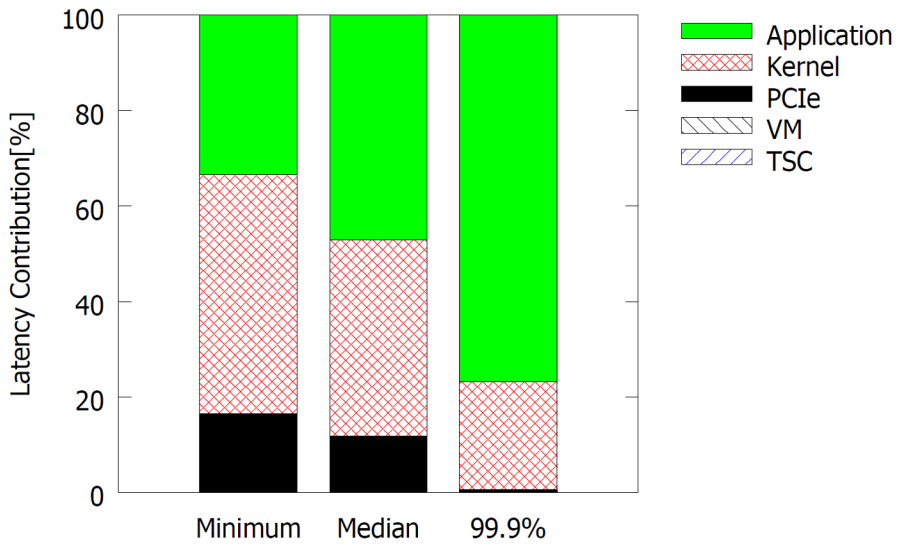
- End host latency has significantly improved over half a decade, but
  - Apps and VMs are the source of significant variance
- Fibre length matters
- “Variance” is no longer negligible
  - Fewer opportunities to improve end-host networking & interconnect

- The toolkit and reproduction environment are available!

  www.lowlatencylab.org/data/pam2017


Host
User Space
OS
Driver
PCIe
NIC / NetFPGA
User Space Latency
User Space+ OS Latency
Interconnect Latency
Entire Host Latency

Questions?


Latency Contribution[%]
100
80
60
40
20
0
Minimum
Median
99.9%
Application
Kernel
PCIe
VM
TSC

Acknowledgements

UNIVERSITY OF CAMBRIDGE

The Leverhulme Trust

EPSRC

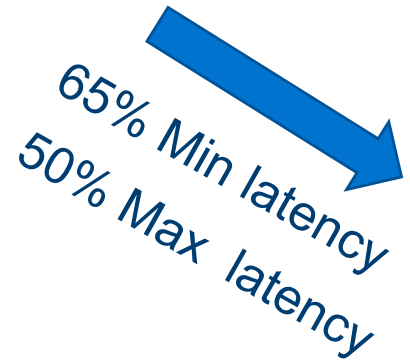
Pioneering research and skills

SOLARFLARE®

EXABLAZE

Improving Tail Latency

```
1   while (!done)
2   {
3       //Read TSC once
4       do_rdtscp(tsc, cpu);
5       //If the gap between the current and the previous
            TSC value is above a certain threshold, save
            it
6       if ((tsc - last > threshold) && (cpu == lastcpu))
7          buffer[samples++] = tsc-last;
8       last = tsc;
9       lastcpu = cpu;
10  }
```




```
1   while (!done)
2   {
3       //Read TSC twice, one immedately after the other
4       do_rdtscp(tsc, cpu);
5       do_rdtscp(tsc2,cpu2);
6       //If the gap between the two reads is above a
            certain threshold, save it
7       if ((tsc2 - tsc > threshold) && (cpu == cpu2))
8          buffer[samples++] = tsc2-tsc;
9   }
```

- Running in real time
- Pinning to a core
- Inhibiting interrupts
- Coding practices

Application Only

- What we do: Read TSC
- Min: 9ns
- Median: 10ns
- 99.9%: 11ns
- Max: 10's to 100's of µs

- 50-100 events/second > 1 µs