

# An Evaluation of Ethernet Performance for Scientific Workloads

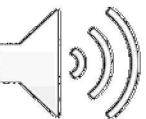


**Joseph P. Kenny**, Jeremiah J. Wilke, Craig D. Ulmer, Gavin M. Baker,  
Samuel Knight and Jerrold A. Friesen

*Scalable Modeling and Analysis*  
*Sandia National Laboratories, Livermore, CA, USA*

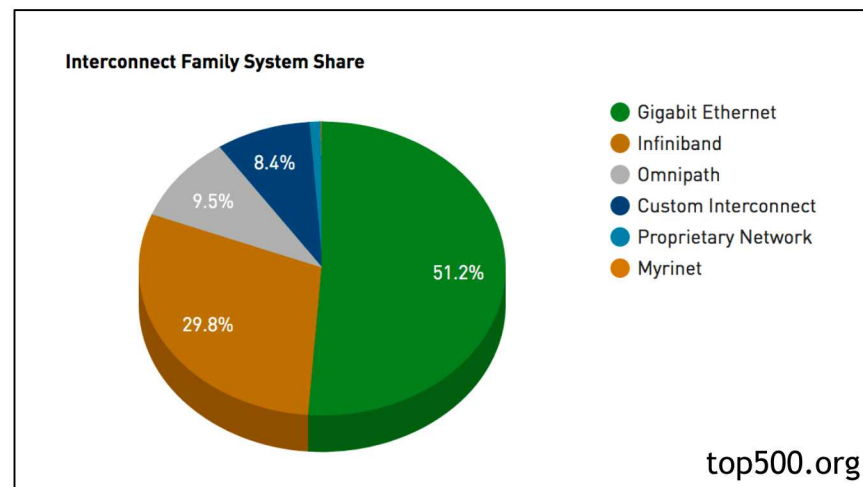


Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

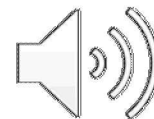


# Another Look at Ethernet for Scientific Workloads

- 51% of current TOP500 systems run on Ethernet
- Mellanox Ethernet revenues now exceed Infiniband (Mellanox Corporate Update, March 2020)
- HPE Cray Slingshot emphasizes Ethernet compatibility
- Storage, hyperscale and hyperconverged markets overwhelmingly Ethernet-focused
- Ethernet = risk mitigation?



- Sandia/CA unique procurement in 2017 to support network emulation
  - Required high performance Ethernet to support existing tools
  - See J. Floren et. al., “A reference architecture for emulytics clusters,” in Sandia Report, vol. SAND2009-5574, 2017
- Can future procurements support both network emulation alongside other scientific computing workloads with a single high speed network?



# Ethernet Performance Enhancements

- Data center bridging (DCB) features of potential interest for scientific computing were formally adopted to IEEE 802.1Q standard in 2011
- Priority Flow Control (PFC)
  - Improvement to global flow control, supports near lossless Ethernet for selected traffic priorities
  - Allows Fibre Channel over Ethernet, but also other lossless protocols
- Remote Direct Memory Access (RDMA) is the defining feature of high performance networks
  - Bypass OS kernel for high performance
  - Typically requires lossless network – PFC for Ethernet
  - RDMA over Converged Ethernet (RoCE) standard (IBTA) allows RDMA over Ethernet through the encapsulation of Infiniband packets.
  - RoCE v1 and v2 standards; v2 is routable; folklore of hardware with poor v1 performance
- Enhanced Transmission Selection (ETS)
  - Increased interest in Quality of Service (QoS) for optimizing performance in scientific computing installations
  - ETS: weighted round-robin algorithm for Ethernet QoS

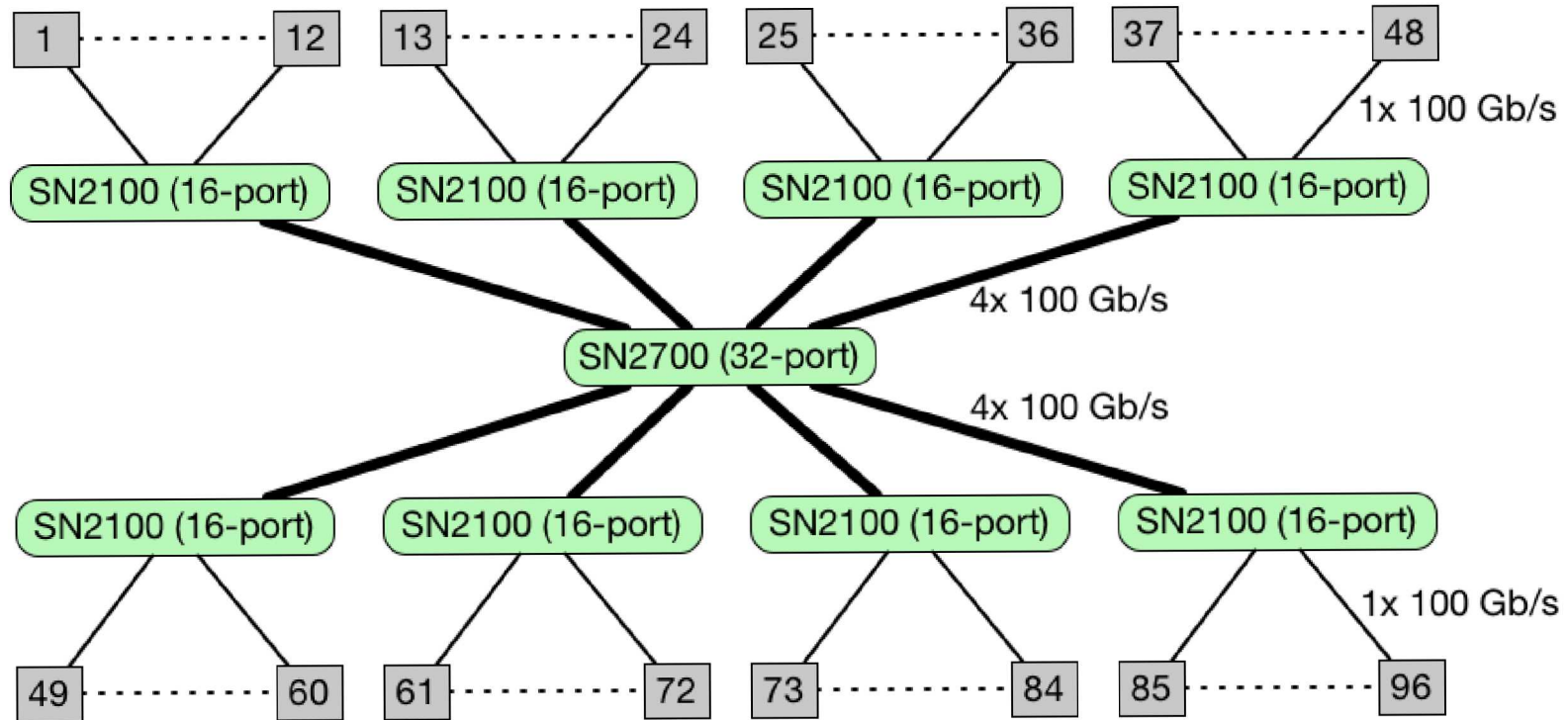


# Previous work

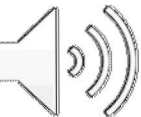
- Significant previous work in these areas is outlined in full paper
- Vienne et. al. -- comprehensive comparison of QDR/FDR Infiniband and 10/40 Gb/s RoCE, limited to single switch
  - J.Vienne et. al., “Performance Analysis and Evaluation of InfiniBand FDR and 40GigE RoCE on HPC and Cloud Computing Systems,” in 2012 IEEE 20th Annual Symposium on High-Performance Interconnects. IEEE, 2012, pp. 48–55.
- Mubarak et. al., Savoie et. al., and Wilke and Kenny -- simulations examining QoS for HPC workloads
  - L.Savoie et. al., “A Study of Network Quality of Service in Many-Core MPI Applications,” in 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2018, pp. 1313–1322.
  - M. Mubarak et al., “Evaluating Quality of Service Traffic Classes on the Megafly Network,” 2019.
  - J.J.Wilke and J.P. Kenny, “Opportunities and limitations of Quality-of-Service (QoS) in Message Passing (MPI) applications on adaptively routed Dragonfly and Fat Tree networks,” in 2020 IEEE International Conference on Cluster Computing (CLUSTER), 2020, in press.
- Balla et. al. used QoS to reduce RoCE latencies in the presence of interfering traffic, but did not consider application level benchmarks
  - D.Balla et. al., “Bounded latency with RoCE,” in Proceedings of the ACM SIGCOMM 2019 Conference Posters and Demos, 2019, pp. 134–135.
- Our work is distinguished by
  - 100G generation hardware
  - Size of testbed (9 switches, 96 nodes)



# Mellanox 100Gb/s Ethernet Testbed



- 3:1 tapering, should promote congestion
- Representative of typical of TOR leaf-spine designs (vs HPC)

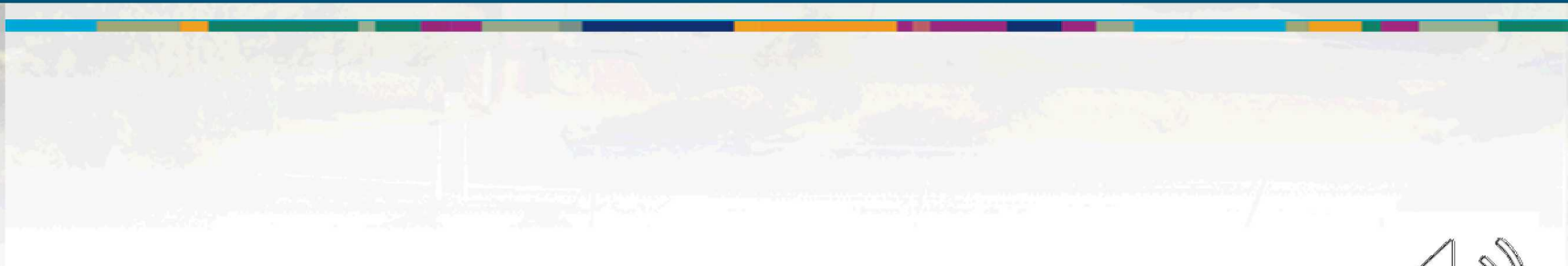


- Single Switch Bandwidth/Latency
  - MPI point-to-point bandwidth/latency [MVAPICH2]
  - Incast scanning up to 10 streams and up to 4 source nodes [custom script driving iperf3/ib\_write\_bw]
- Application Proxies
  - Latency-sensitive: fast Fourier transform (FFT) [subcom3d-a2a from LLNL/Chatterbug]
  - Bandwidth-sensitive: halo exchange (Halo3D) [halo3d-26 from SST/Ember]
  - MPI Parallel: High Performance Linpack benchmark (HPL) [UT-ICL/netlib.org]
- QoS Case Study
  - FFT running with interference from Halo3D background traffic
- MPI applications run with Open MPI 4.0.4
  - Easy to swap network transports and select RoCE service level
- Additional software/hardware details available in full paper and reproducibility artifact

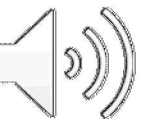
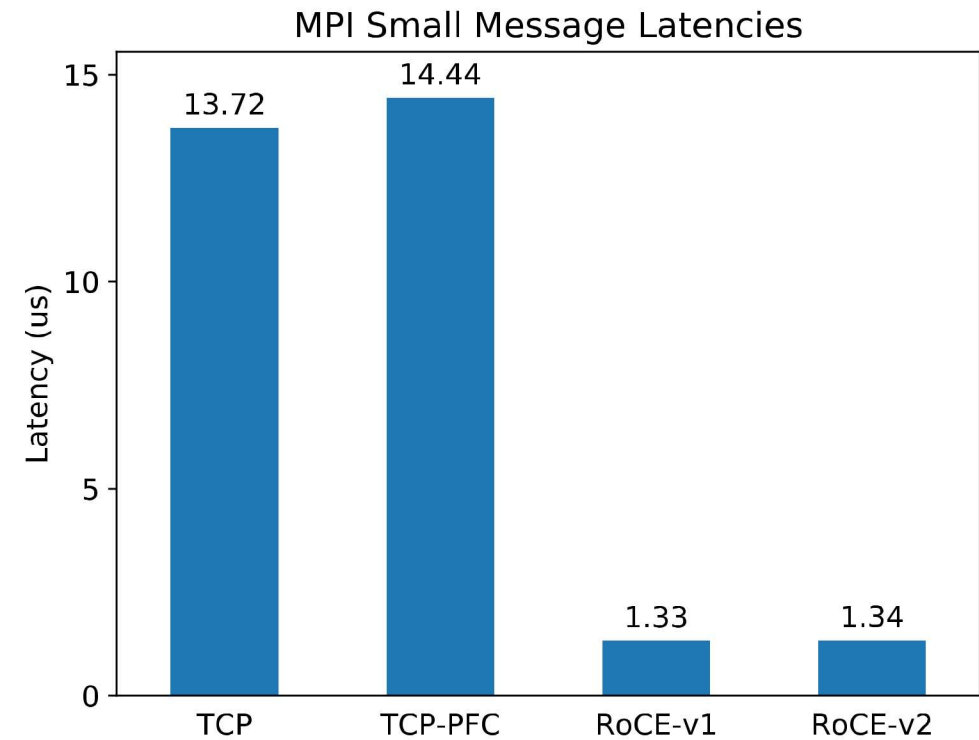
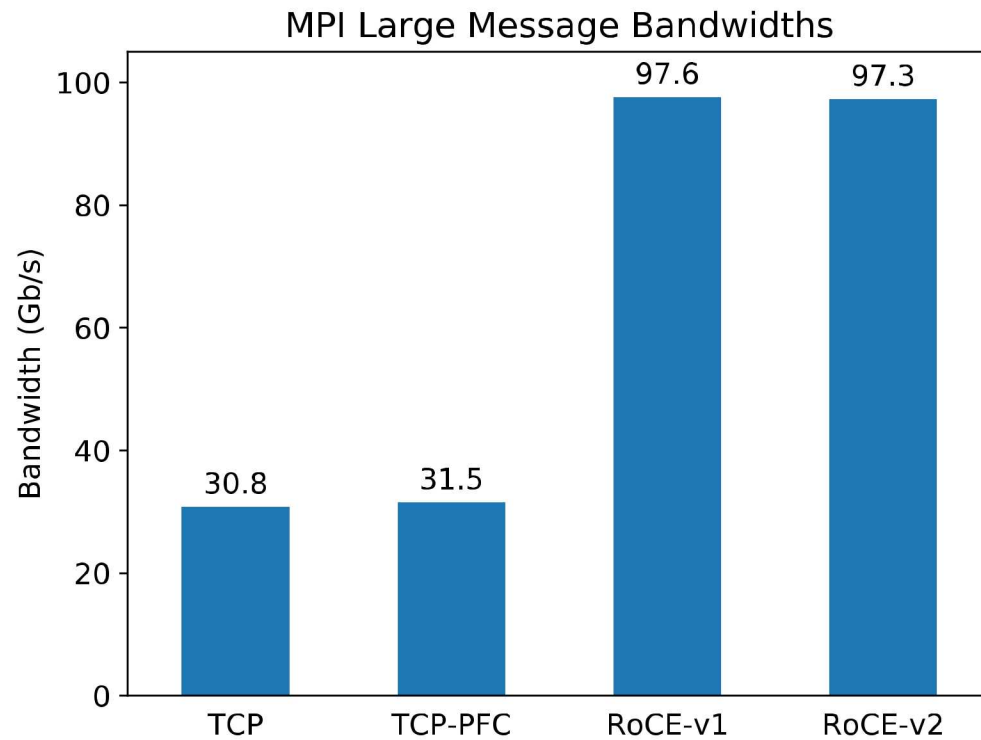




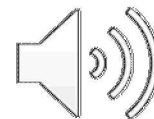
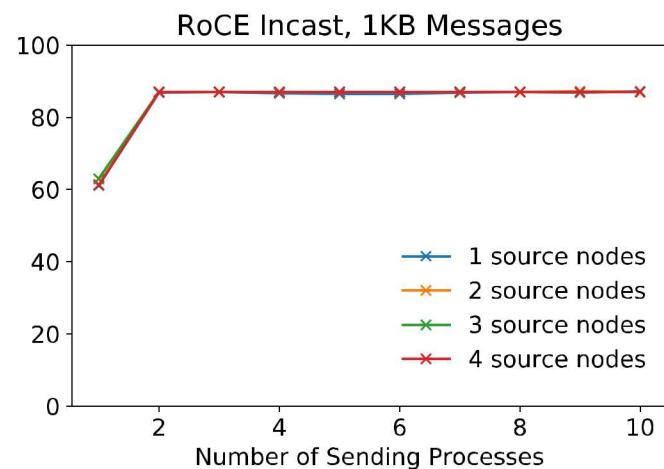
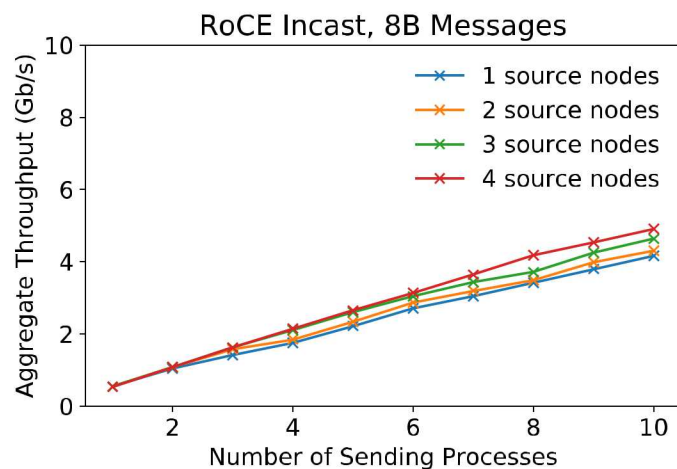
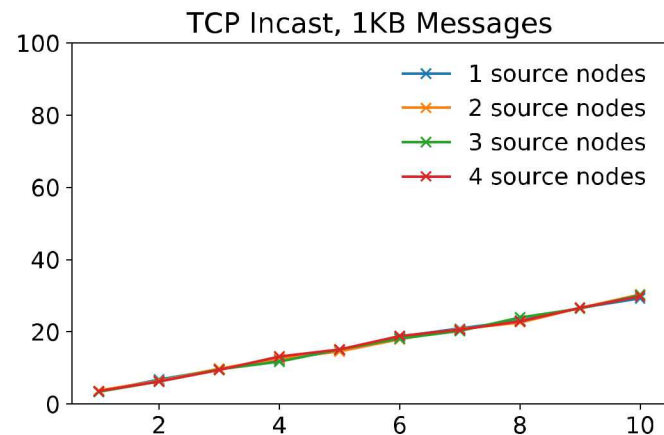
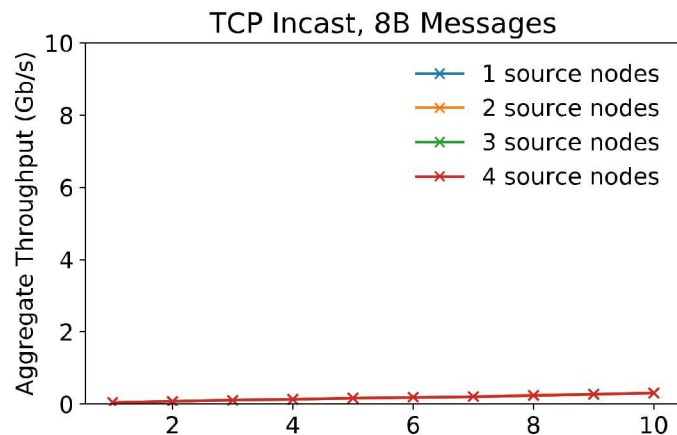
# Bandwidth and Latency Tests



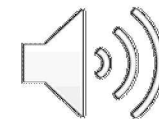
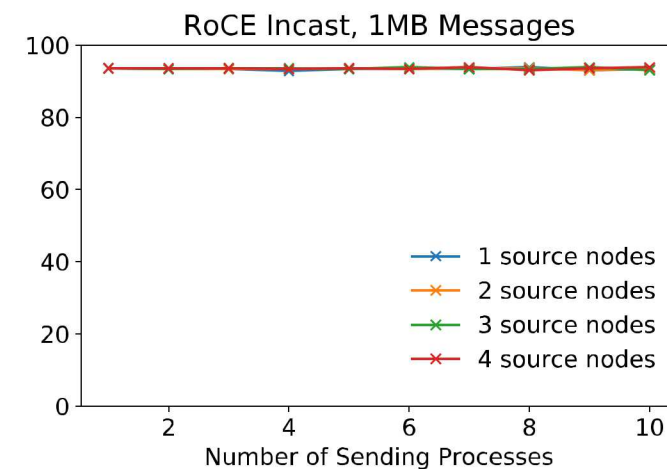
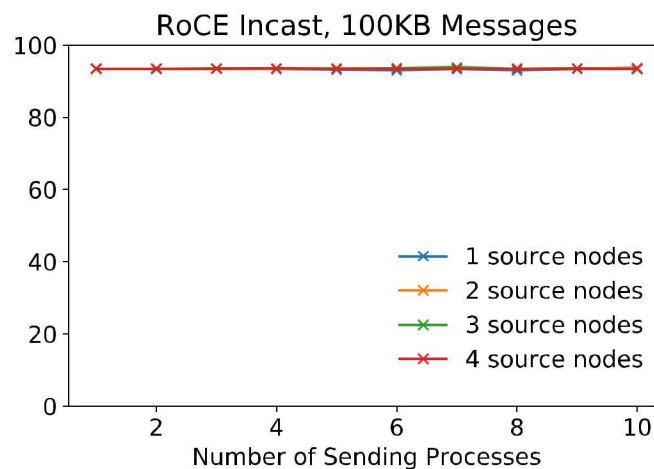
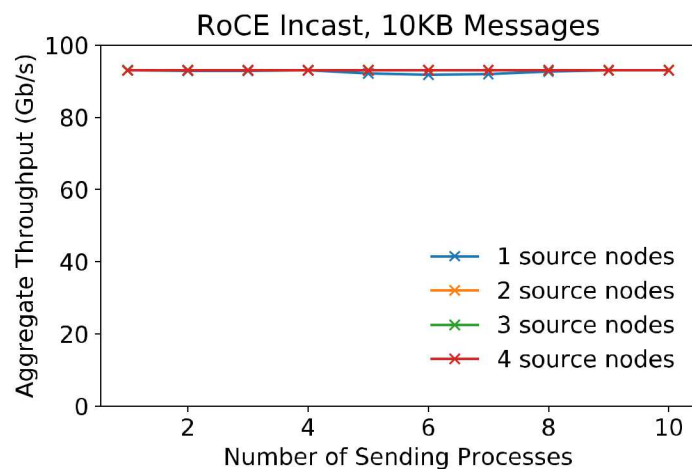
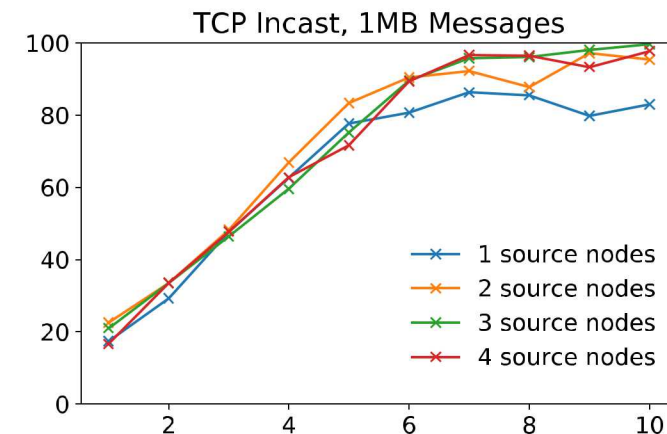
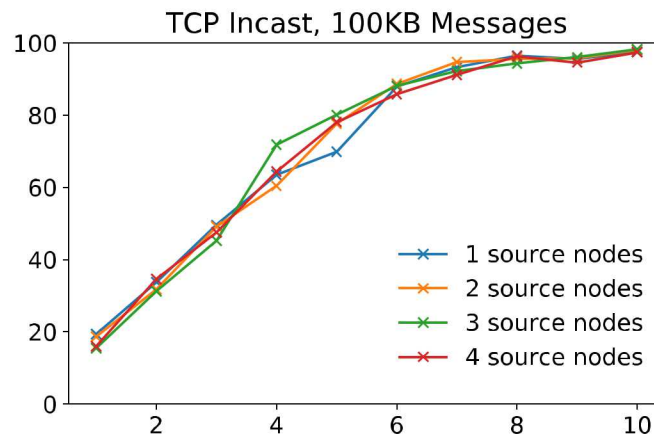
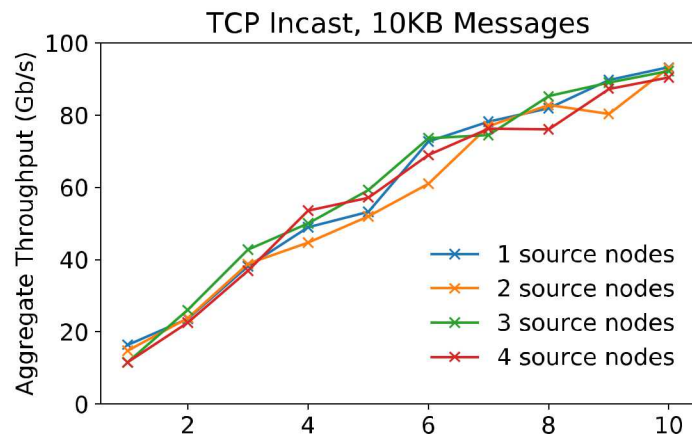
# MPI Point-to-Point Bandwidth/Latency



# Small/Medium Message Incast

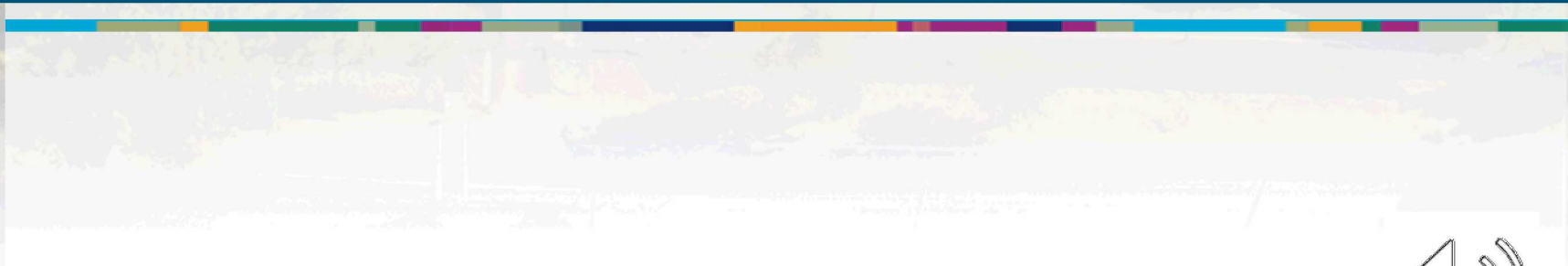


# Large Message Incast

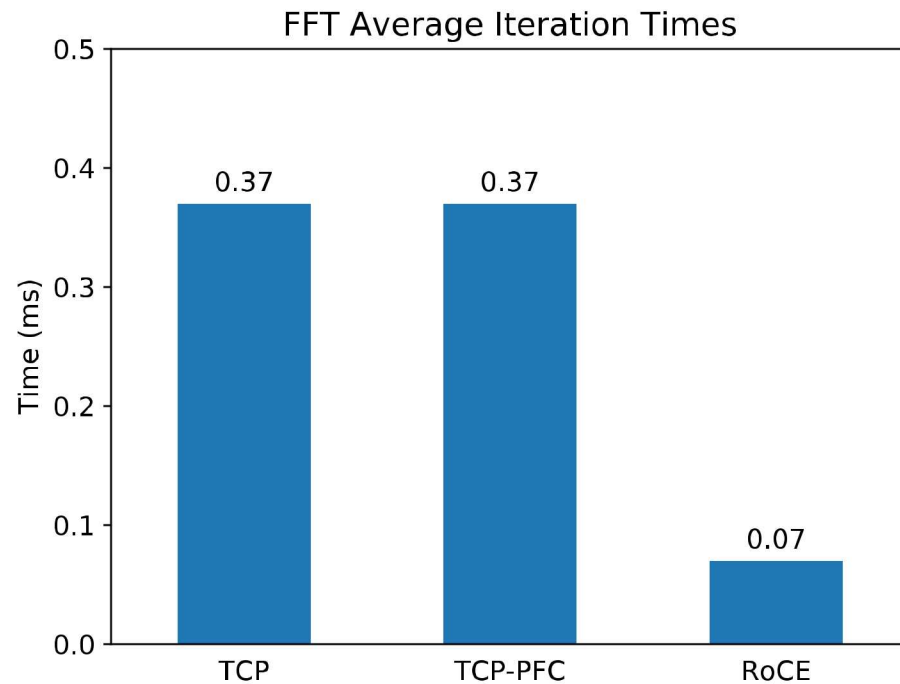




# Application Proxy Performance



# Latency Sensitive: FFT

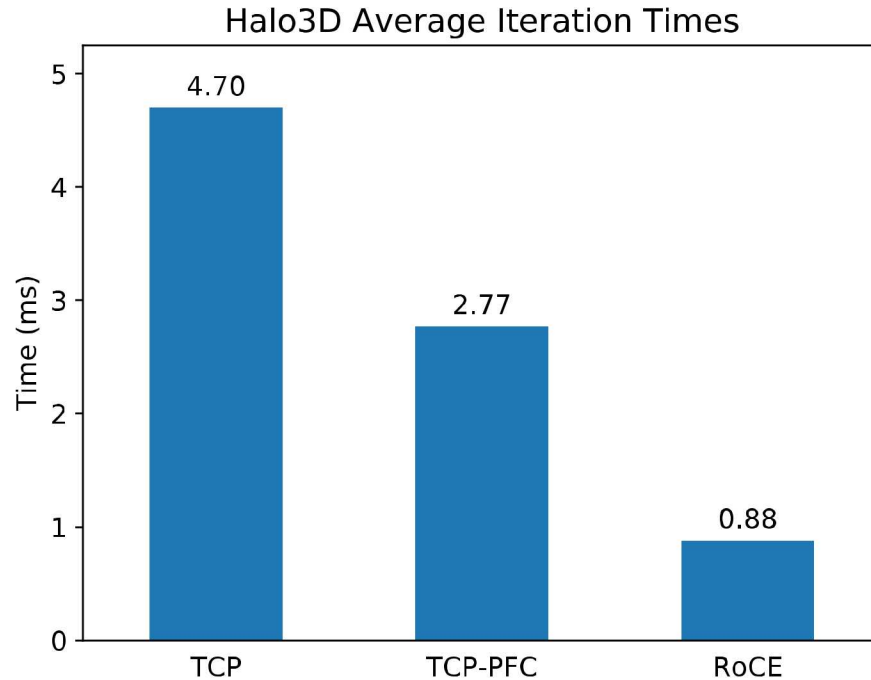


	Rx Pause Duration	Tx Pause Duration
TCP-PFC	0	0
RoCE	0	0

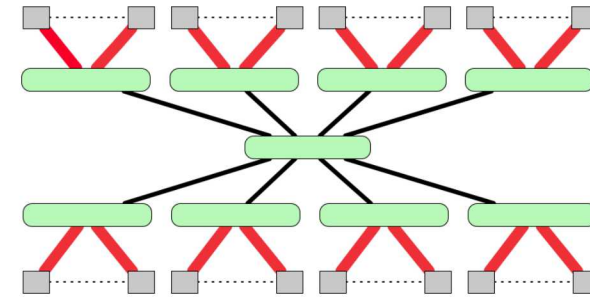
- No congestion, RoCE latency is a big win



# Bandwidth Sensitive: Halo Exchange

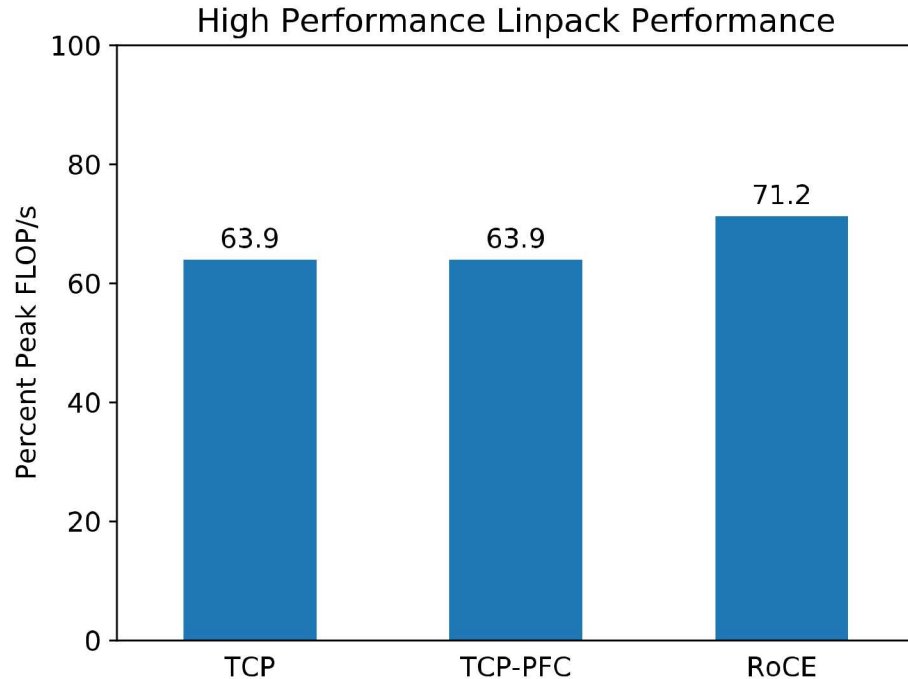


	Rx Pause Duration	Tx Pause Duration
TCP-PFC	6602760	0
RoCE	120121	0

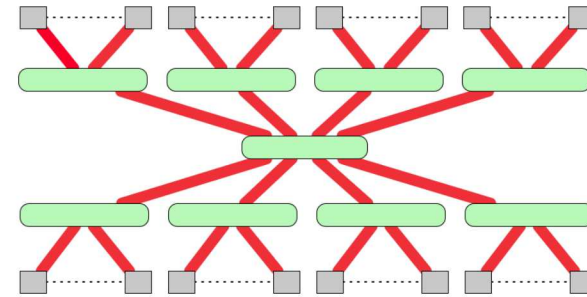


- Congestion limited to ejection link (leaf to node)
- RoCE kernel bypass improves message handling
- PFC improves TCP performance





	Rx Pause Duration	Tx Pause Duration
TCP-PFC	241264	174764
RoCE	6929284	9404312

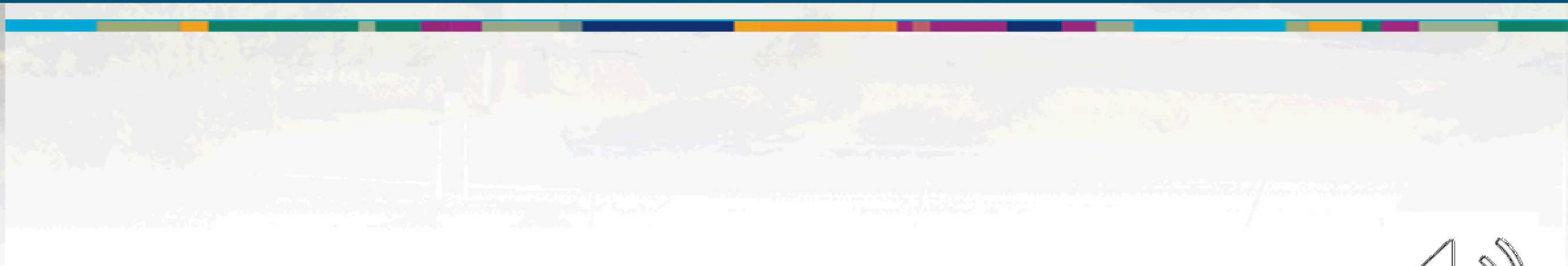


- Congestion spread throughout network
- RoCE increases congestion (unlike Halo3D)
- Many TCP streams effectively use available bandwidth

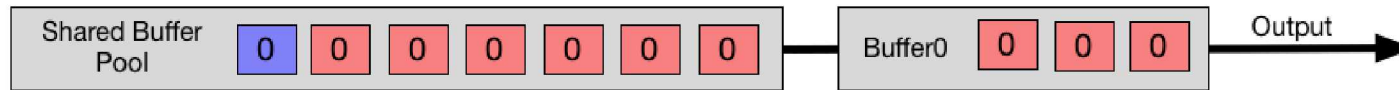




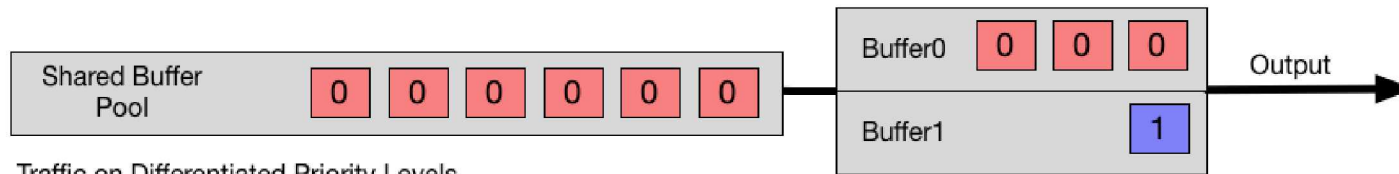
# Managing Interference with ETS



# Enhanced Transmission Selection



All Traffic Shares Priority Level



Traffic on Differentiated Priority Levels

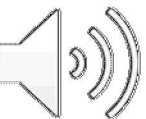


Bandwidth Consumer Packet

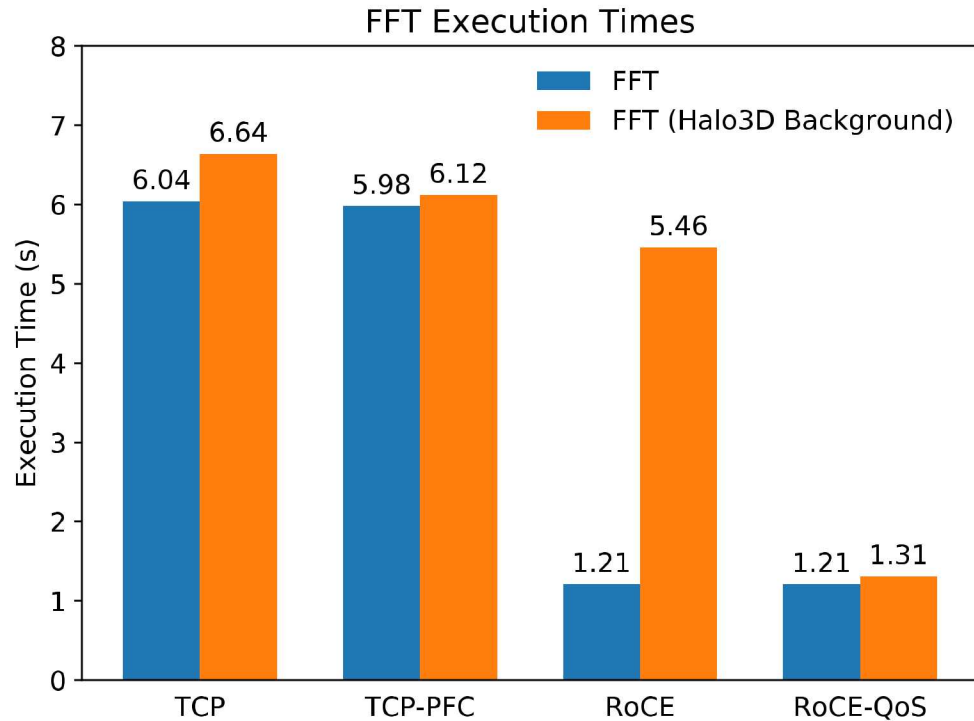


Latency-Sensitive Packet

- QoS provides dedicated buffer resources and differentiated service
- Bandwidth shaping/guarantees appropriate for relatively static workloads (commercial datacenters – storage, streaming multimedia, etc.)
- ETS provides weighted round-robin arbitration, better for dynamic scientific applications (no hard limits, maximal bandwidth utilization)



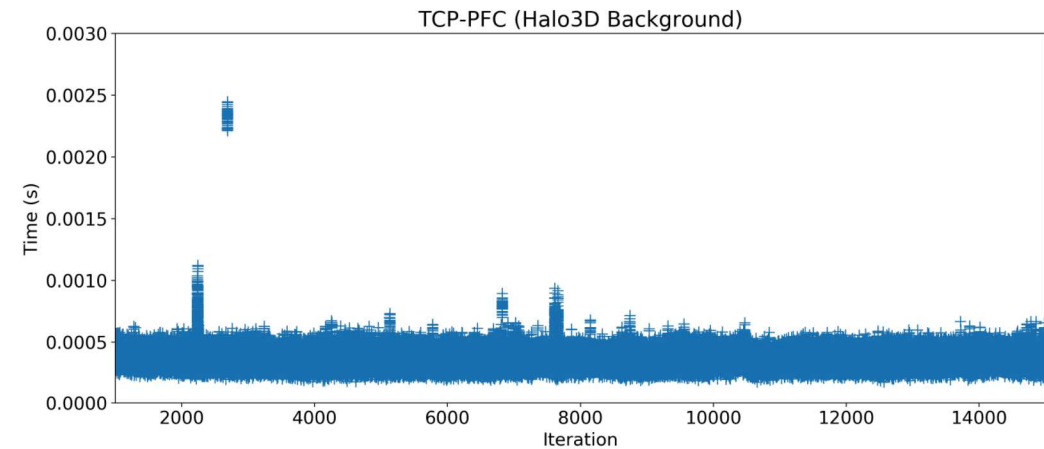
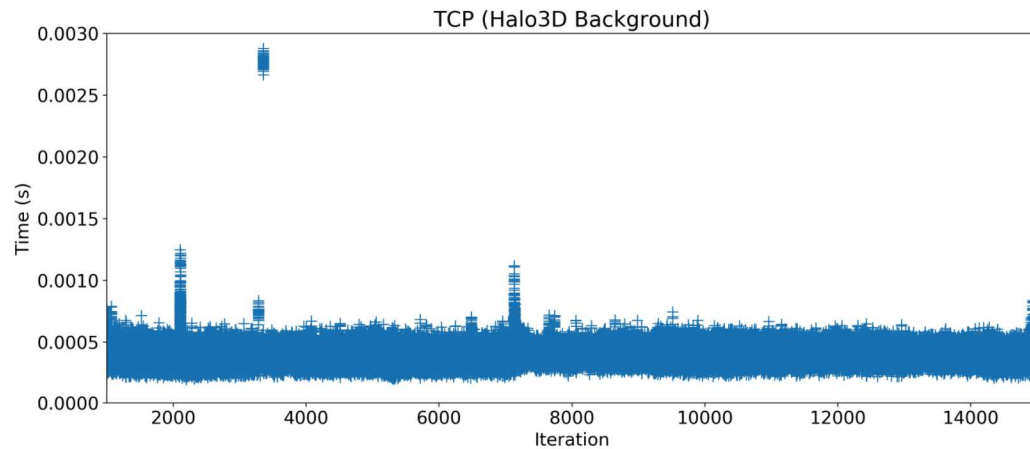
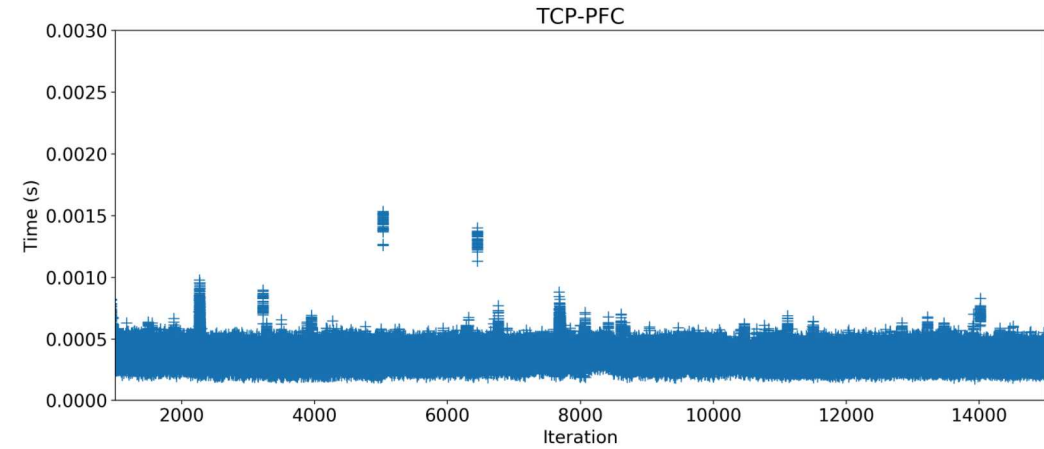
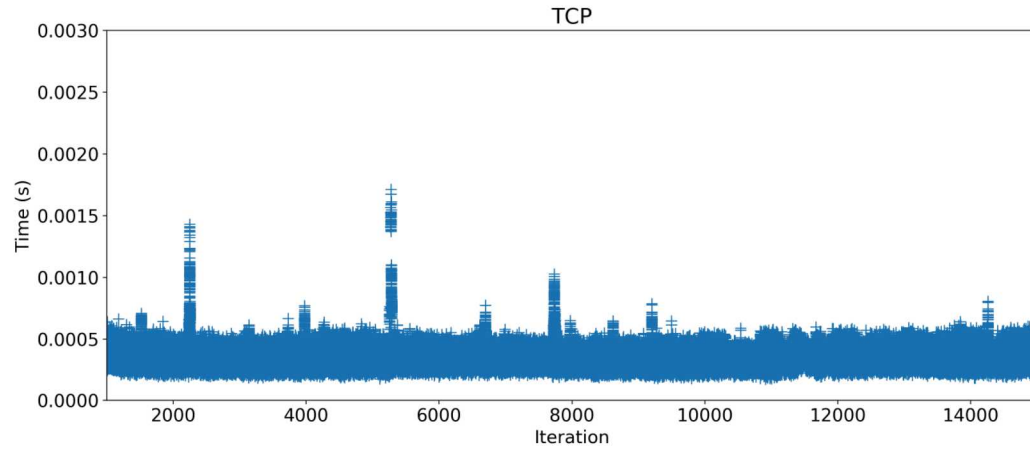
# Bandwidth Consumers vs Latency-Sensitive Traffic



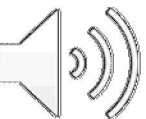
- Halo3D increases FFT network delay
- Latency bottleneck shifts to switches
- RoCE kernel bypass benefit much reduced
- ETS moves FFT traffic to “front of the line”



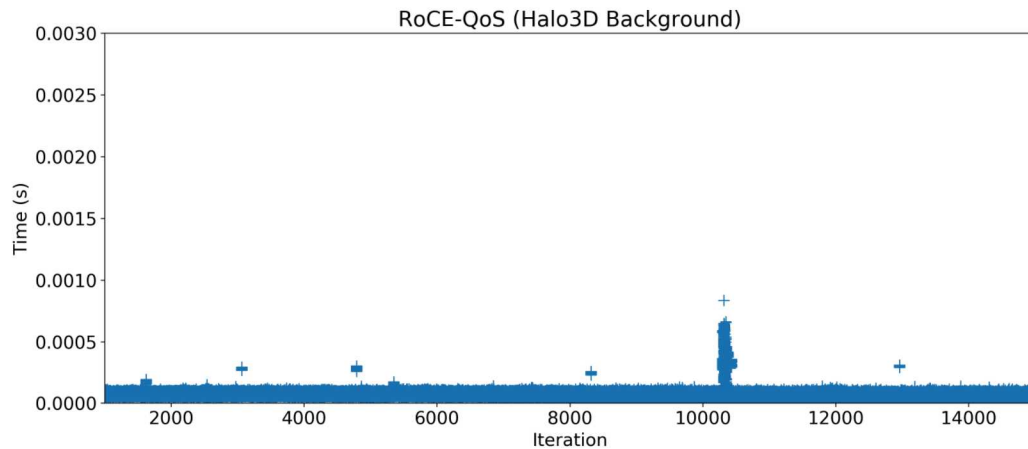
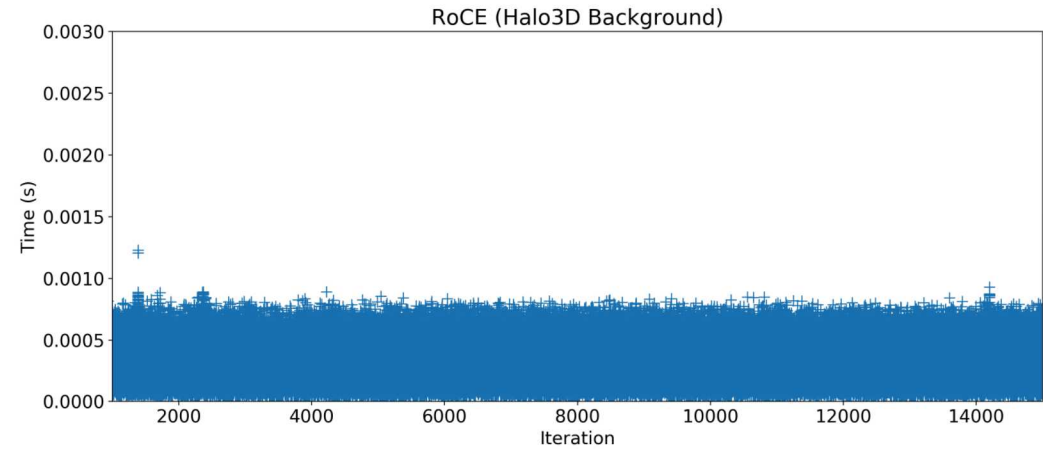
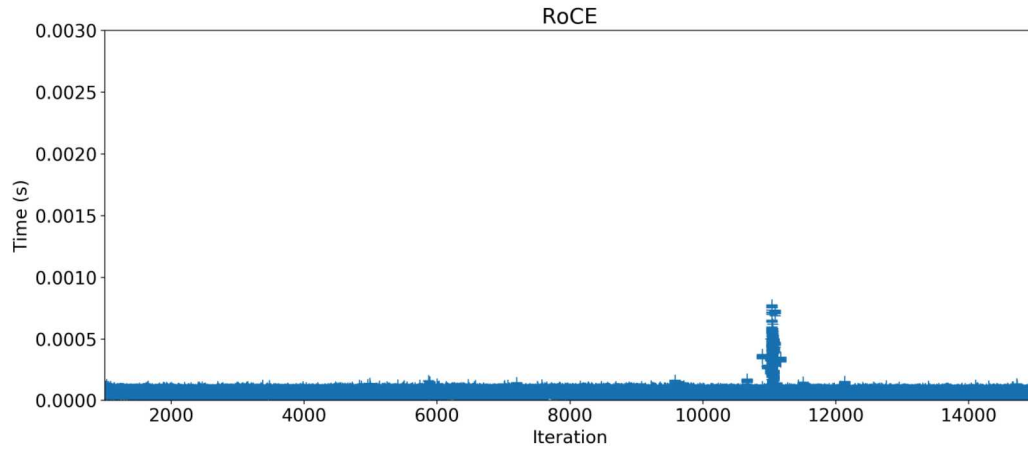
# FFT Per-Node Iteration Times (TCP)



- Halo3D traffic throttled by protocol
- Network not stressed enough to adversely affect FFT



# FFT Per-Node Iteration Times (RoCE)



- Halo3D traffic increases spread in FFT iteration times
- ETS largely recovers FFT performance
- Intermittent slow down of small node subset



- PFC standard clearly “*allows link flow control to be performed on a per-priority basis*”

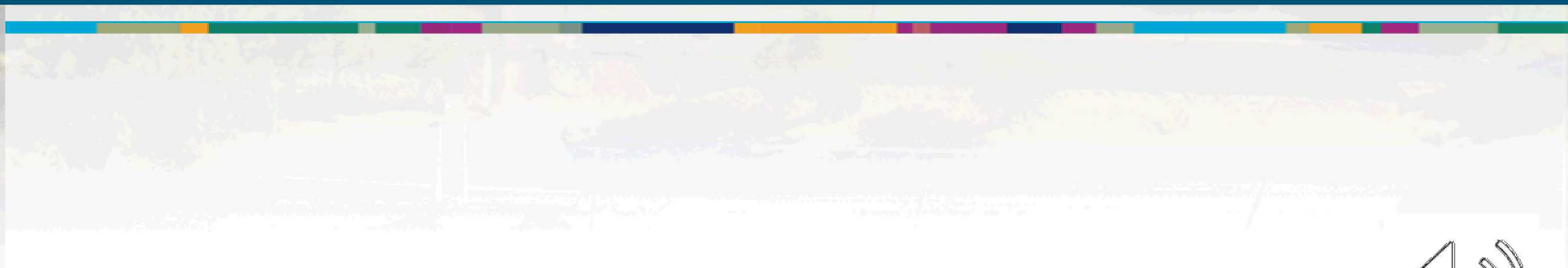
	Rx0 Pause Packets	Rx0 Pause Duration	Rx1 Pause Packets	Rx1 Pause Duration
TCP-PFC	1580102	11477936	1581330	11488489
RoCE	14312	64272	14312	64270
RoCE-QoS	23750	126279	23750	126279

- Priority 1 reports pauses even without QoS enabled
- Priority 1 and 2 pauses are nearly identical
- Attribute QoS performance to arbitration/forwarding priority, not differentiated pause behavior





# In Conclusion



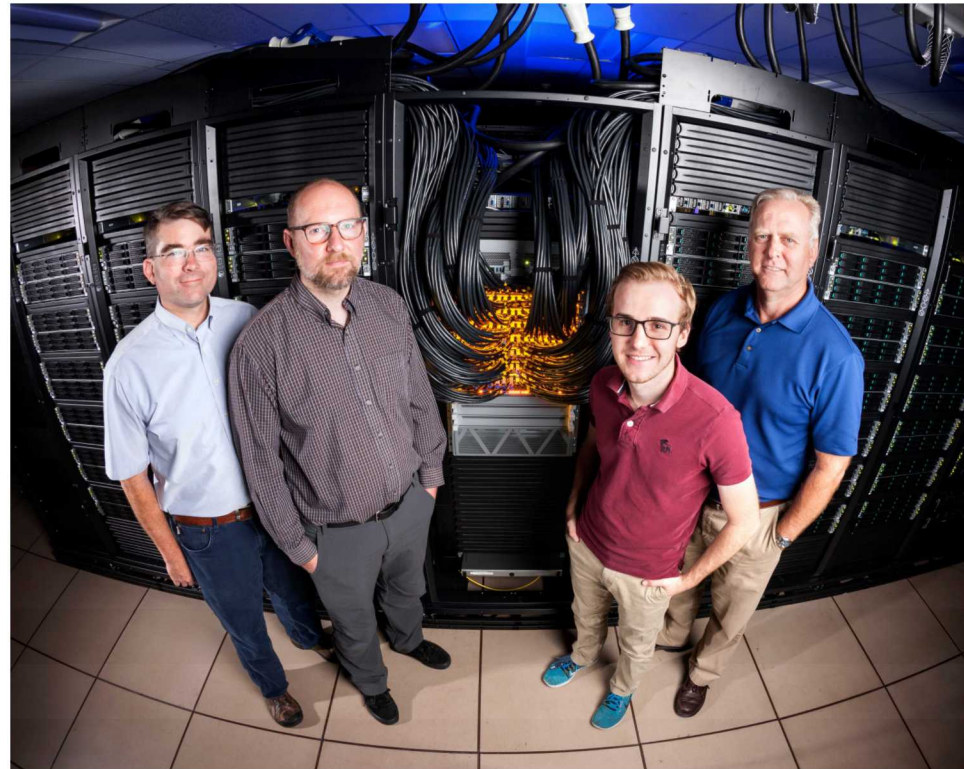
# Conclusions

- RoCE bandwidth and latency can be competitive with modern high performance networks
- For some workloads performance benefits vs TCP are substantial
- QoS is getting more attention in scientific computing for good reason... Ethernet can do that
- RoCE is more challenging to configure than HPC networks (but not as hard to tune as TCP!)
- Is the ecosystem mature enough?
- High-end Ethernet hardware is probably not a cost savings

*Where particular device support or user demands shift requirements,  
Ethernet seems viable for new general purpose scientific computing clusters.*



*Thank you to the organizers, my co-authors and the audience.*



*Craig, Joe, Gavin and Jerry*

