Title: Comparison of High Performance Networks: EDR InfiniBand vs. 100Gb RDMA Capable Ethernet

Author(s):
Van Wig, Faith Virginia
Kachelmeier, Luke Anthony
Erickson, Kari Natania

Intended for: HPC Mini Showcase

Issued: 2016-08-08

**Disclaimer:**
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

LA-UR-??-????

Los Alamos
NATIONAL LABORATORY
EST. 1943

# Comparison of High Performance Network Options:

## EDR InfiniBand VS. 100Gb RDMA Capable Ethernet

By Luke Kachelmeier, Faith Van Wig, and Kari Erickson

Mentors: Susan Coulter, Howard Pritchard

UNM

Missouri S&T

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH • UNIVERSITY

lkachelmeier@unm.edu

fvvnr5@mst.edu

kari.erickson@student.nmt.edu

# Outline

- Background
- Overview
- Experimental Setup
- Results
- Difficulties Faced
- Conclusions
- Future Work

# Background Information

- Ethernet

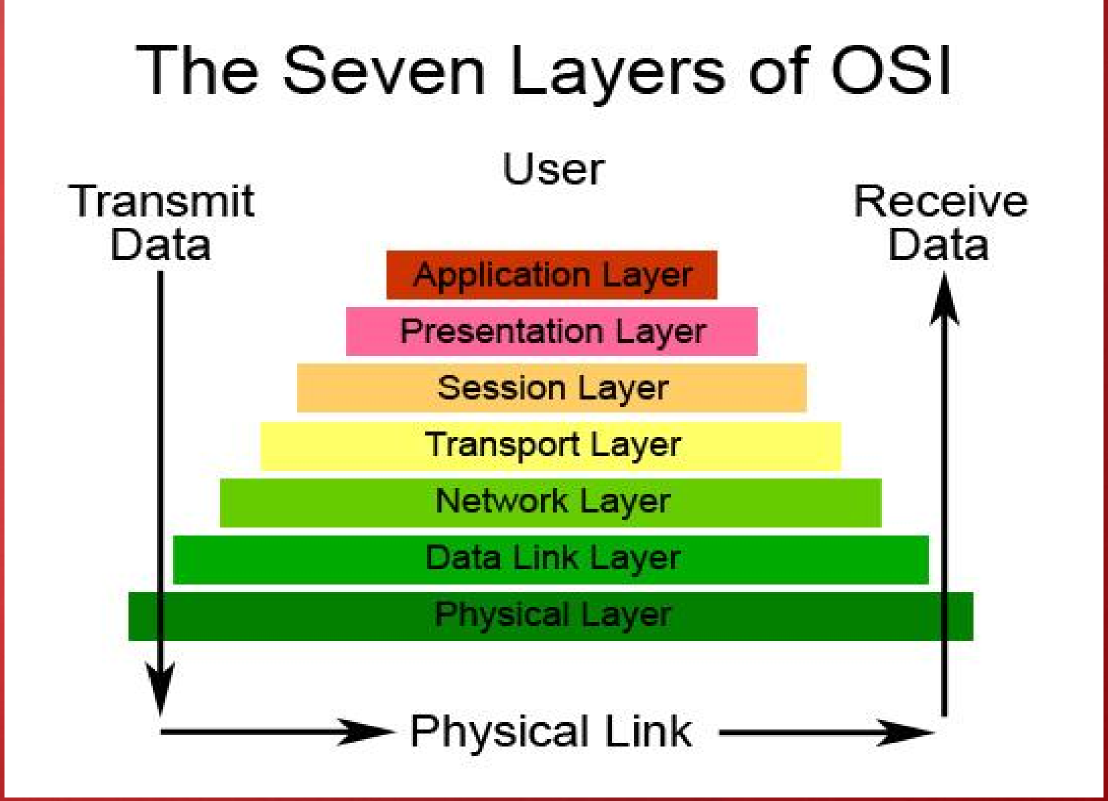
  TCP / IP Stack

- InfiniBand (IB)

  RDMA

  EDR - 4 lanes of 25Gb/s

- RoCE - RDMA over Converged Ethernet



The Seven Layers of OSI

User

Transmit Data

Receive Data

Application Layer

Presentation Layer

Session Layer

Transport Layer

Network Layer

Data Link Layer

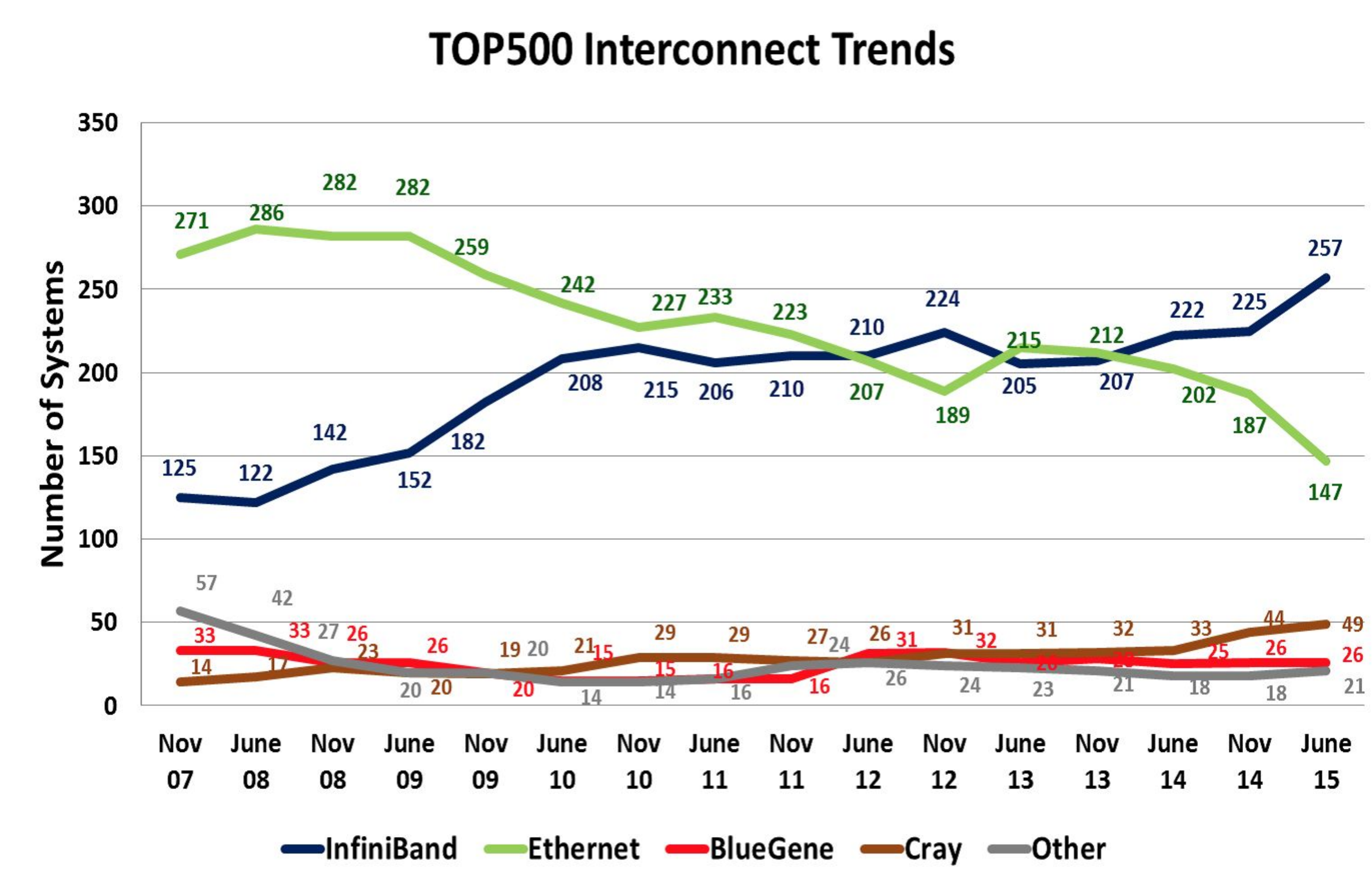
Physical Layer

Physical Link

Ethernet → Layer 3 (Network Layer)
InfiniBand → Layer 2 (Data Link Layer)

http://www.windowsnetworking.com/articles-tutorials/common/OSI-Reference-Model-Layer1-hardware.html

# Overview



http://blog.infinibandta.org/tag/top500/

# Project Description

Compare EDR InfiniBand, RoCE, and native Ethernet:

- Bandwidth Performance
- Latency Performance
- Ease of Configuration (or lack thereof)

OSU Benchmarks, Intel Micro Benchmarks

# Experimental Setup

https://www.attotech.com/products/cables-and-accessories/none/cables-and-accessories/CBL-0310-020

# Hardware

- Mellanox ConnectX-4 EDR adapter cards
- Mellanox SB7700 EDR InfiniBand Switch
- Juniper QFX5200 100Gb Ethernet Switch

http://www.mellanox.com/bttf/

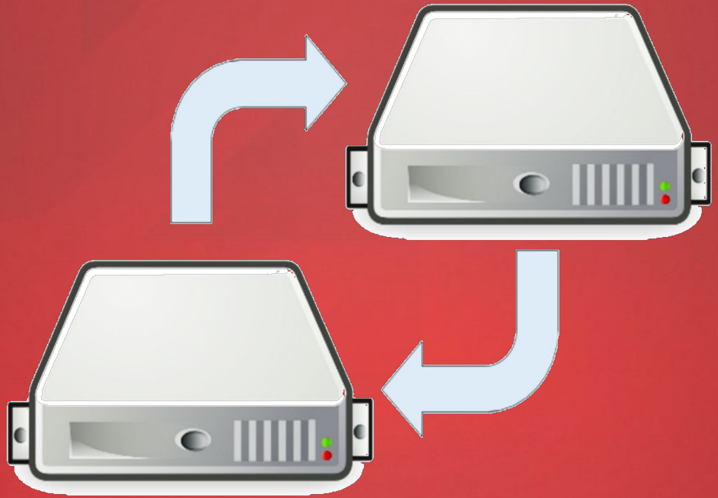
https://www.westconcomstor.com/us/en/westcon/westcon-vendors/juniper_networks.html

# Services

- CentOS 6 across the nodes
- OpenMPI 1.10 using MXM (Mellanox) libraries
- MLNX OFED drivers and related modules, *v. 3.3-1*

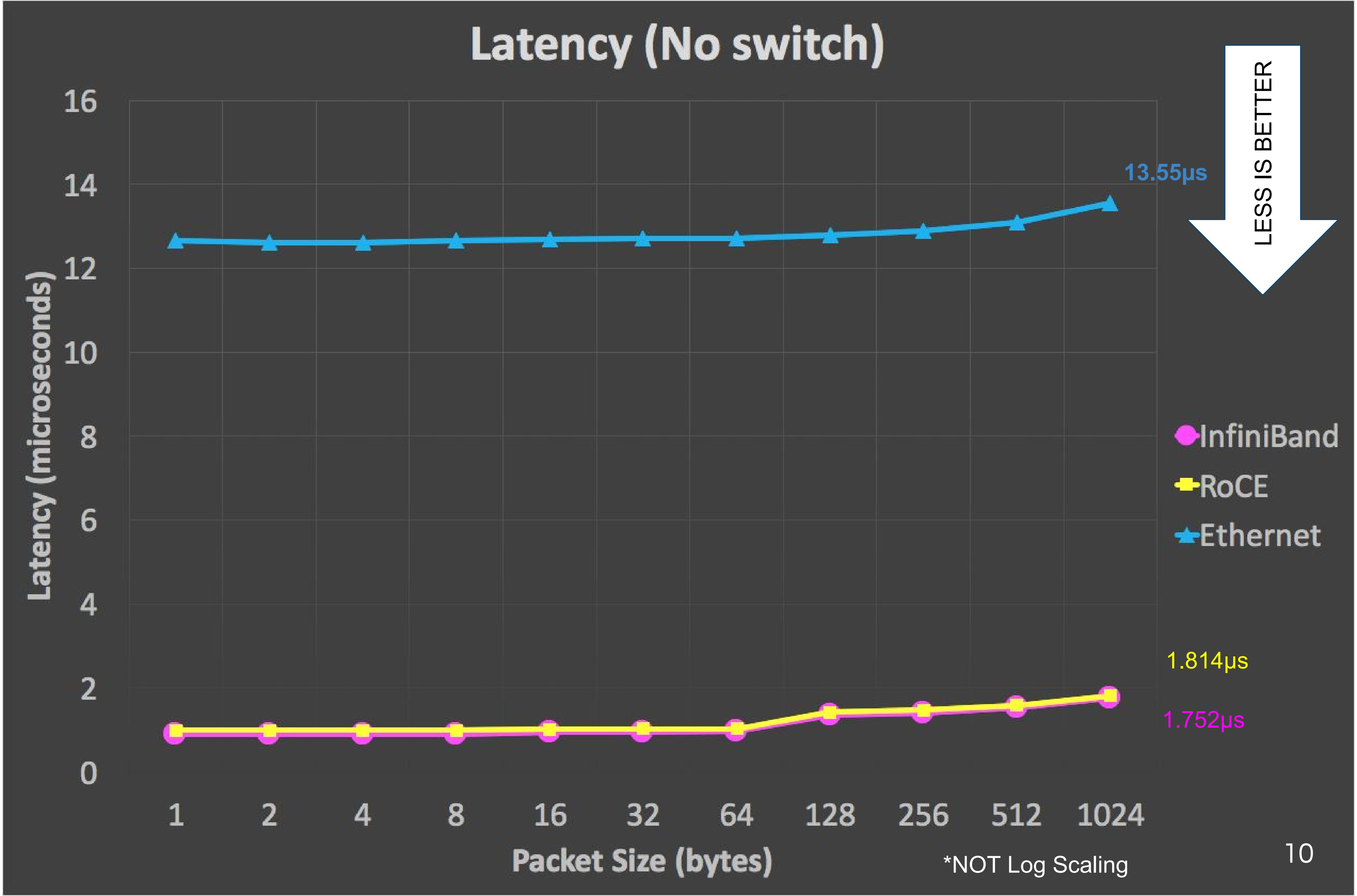
# Results

https://www.healthlottery.co.uk/results/



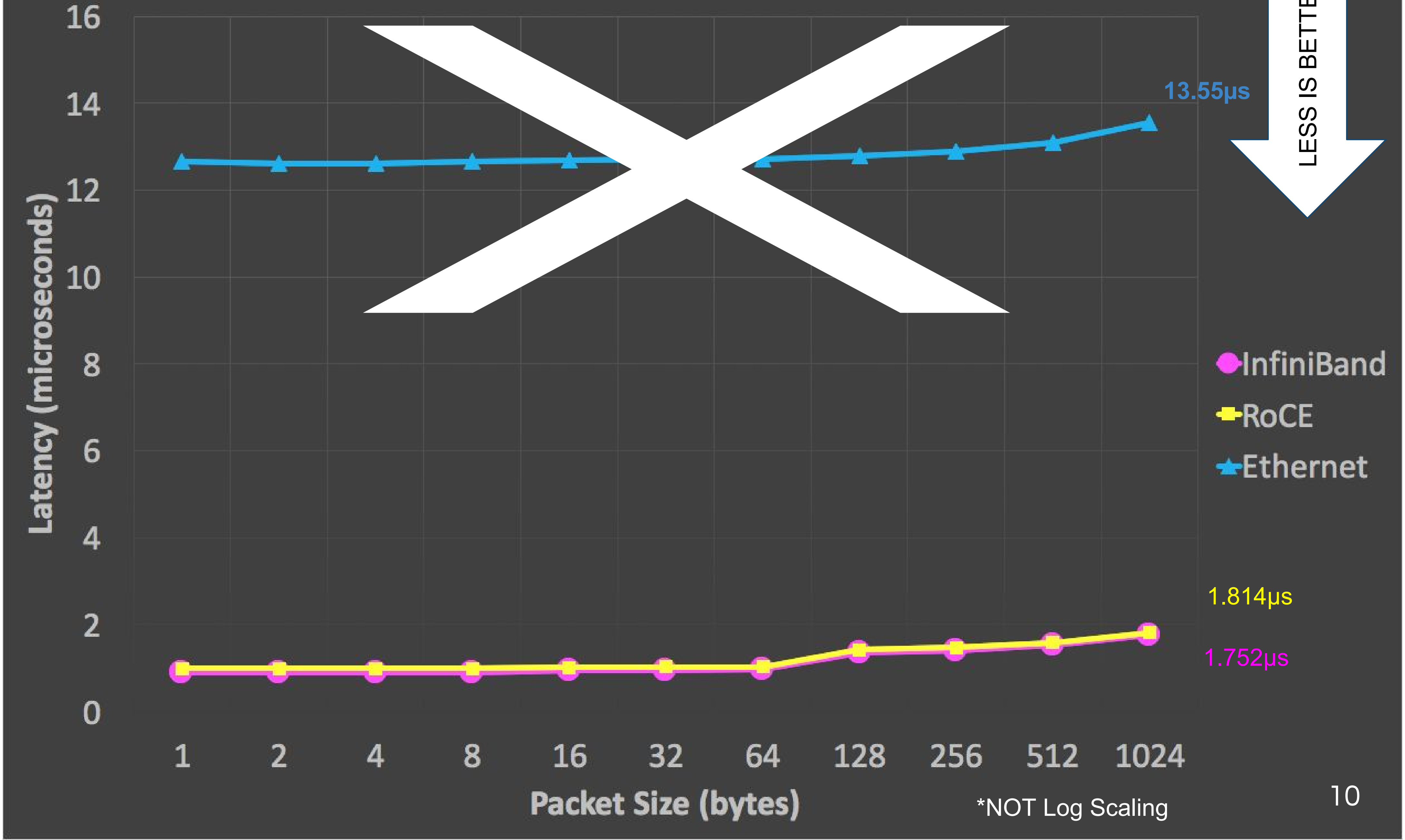
# Point-to-Point Connected Tests

OSU Benchmarks



*NOT Log Scaling

# Latency (No switch)



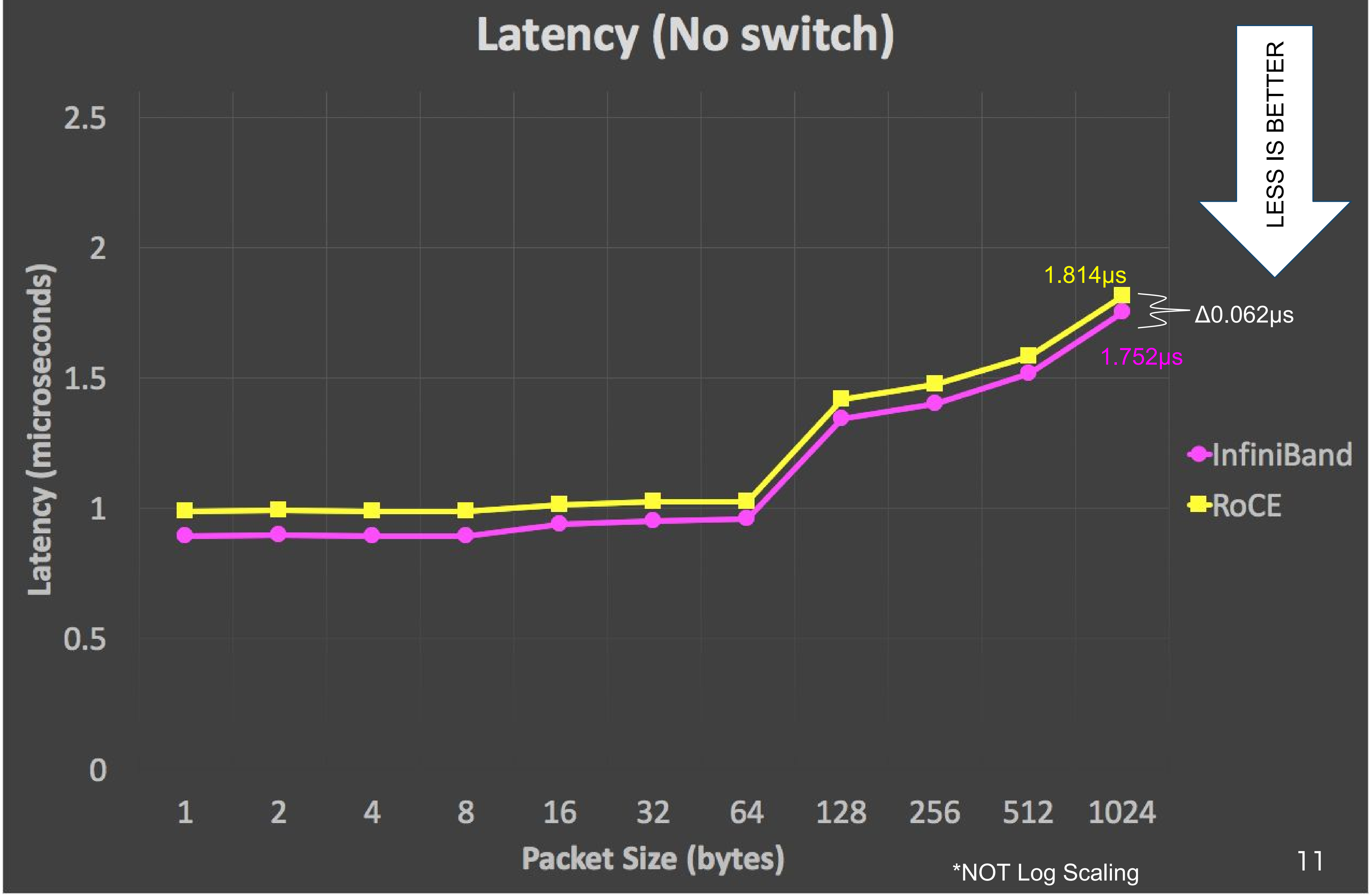
13.55µs

1.814µs

1.752µs

LESS IS BETTER

*NOT Log Scaling



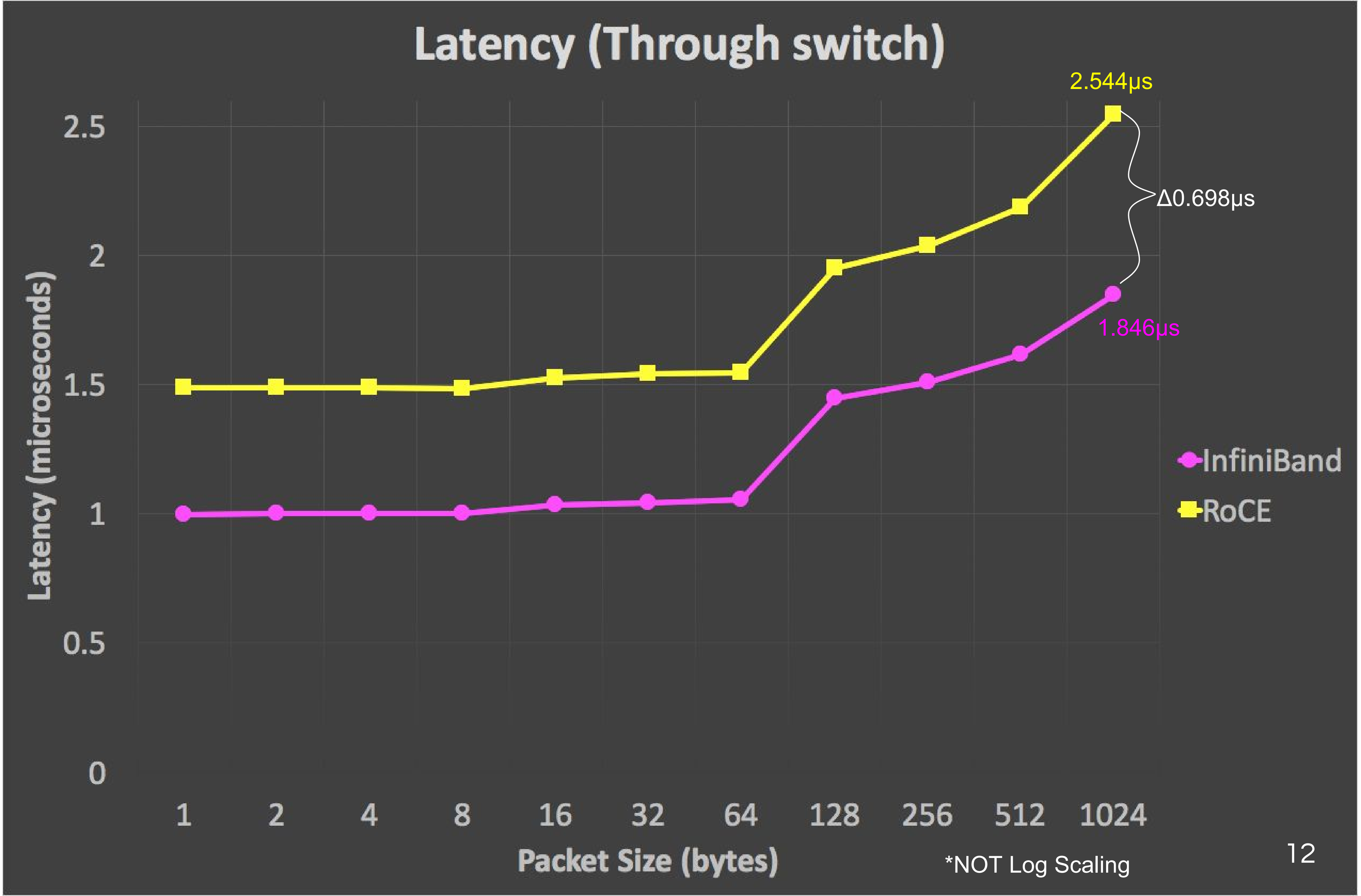
LESS IS BETTER

1.814μs

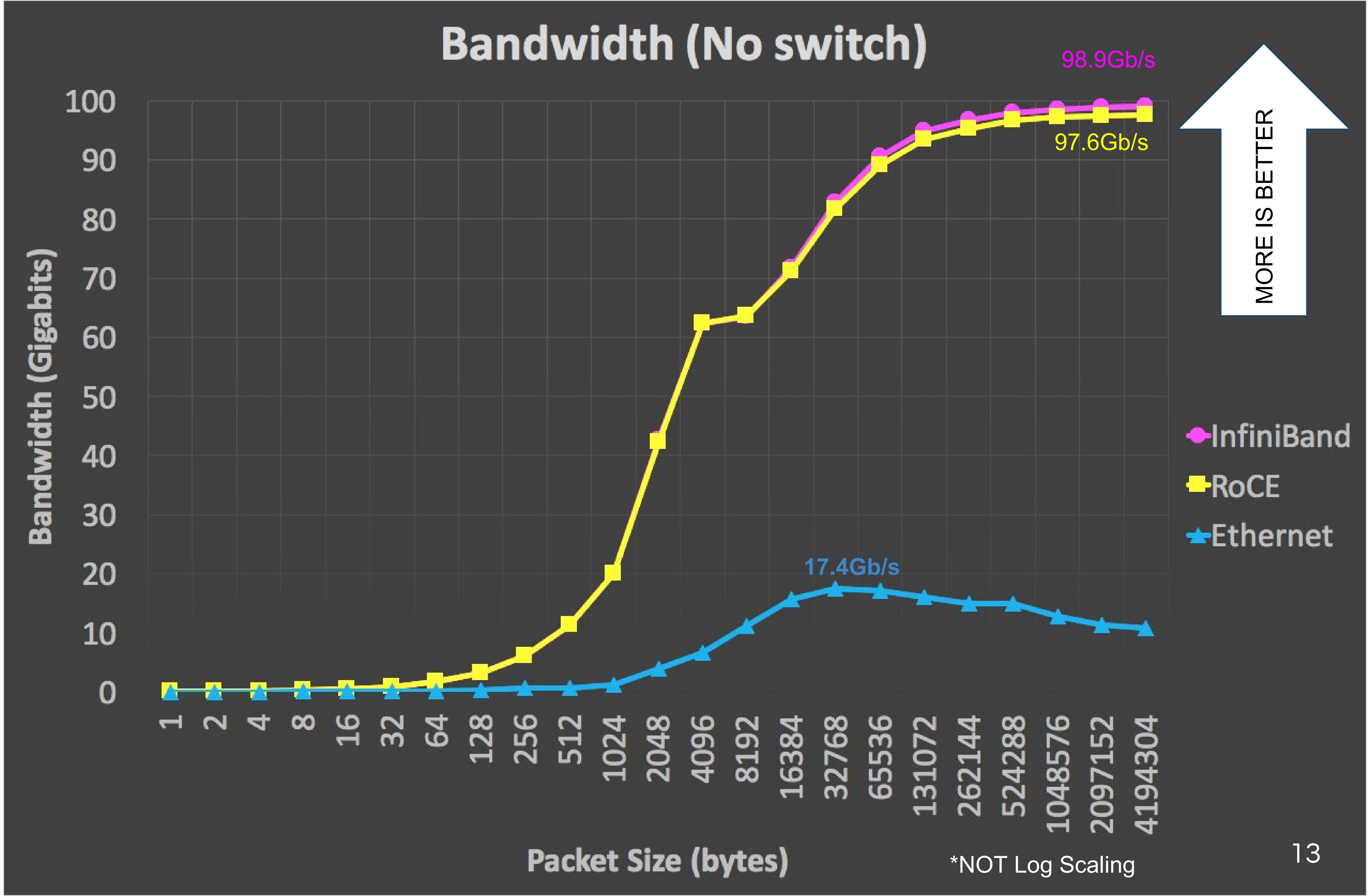
Δ0.062μs

1.752μs

*NOT Log Scaling

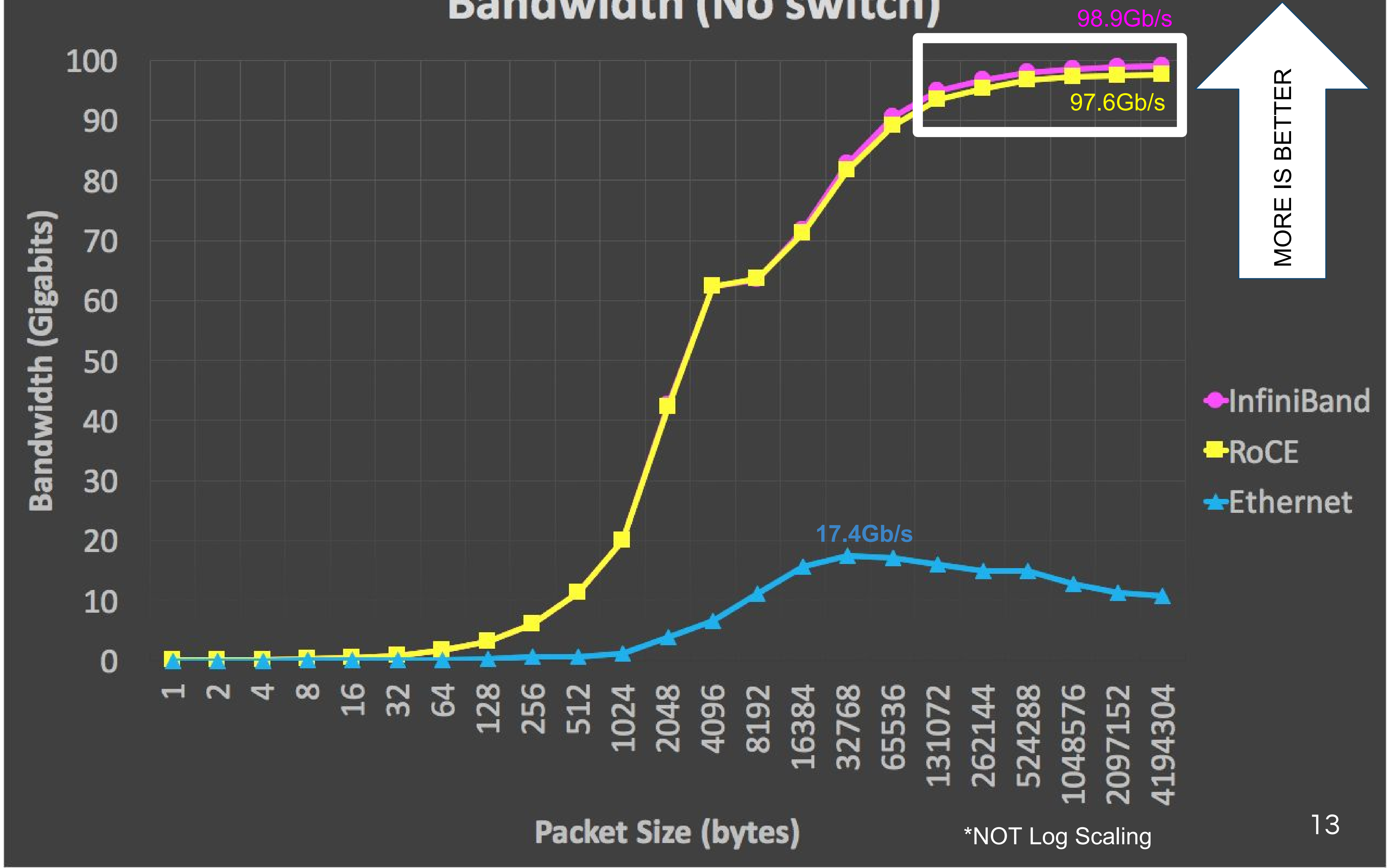


*NOT Log Scaling



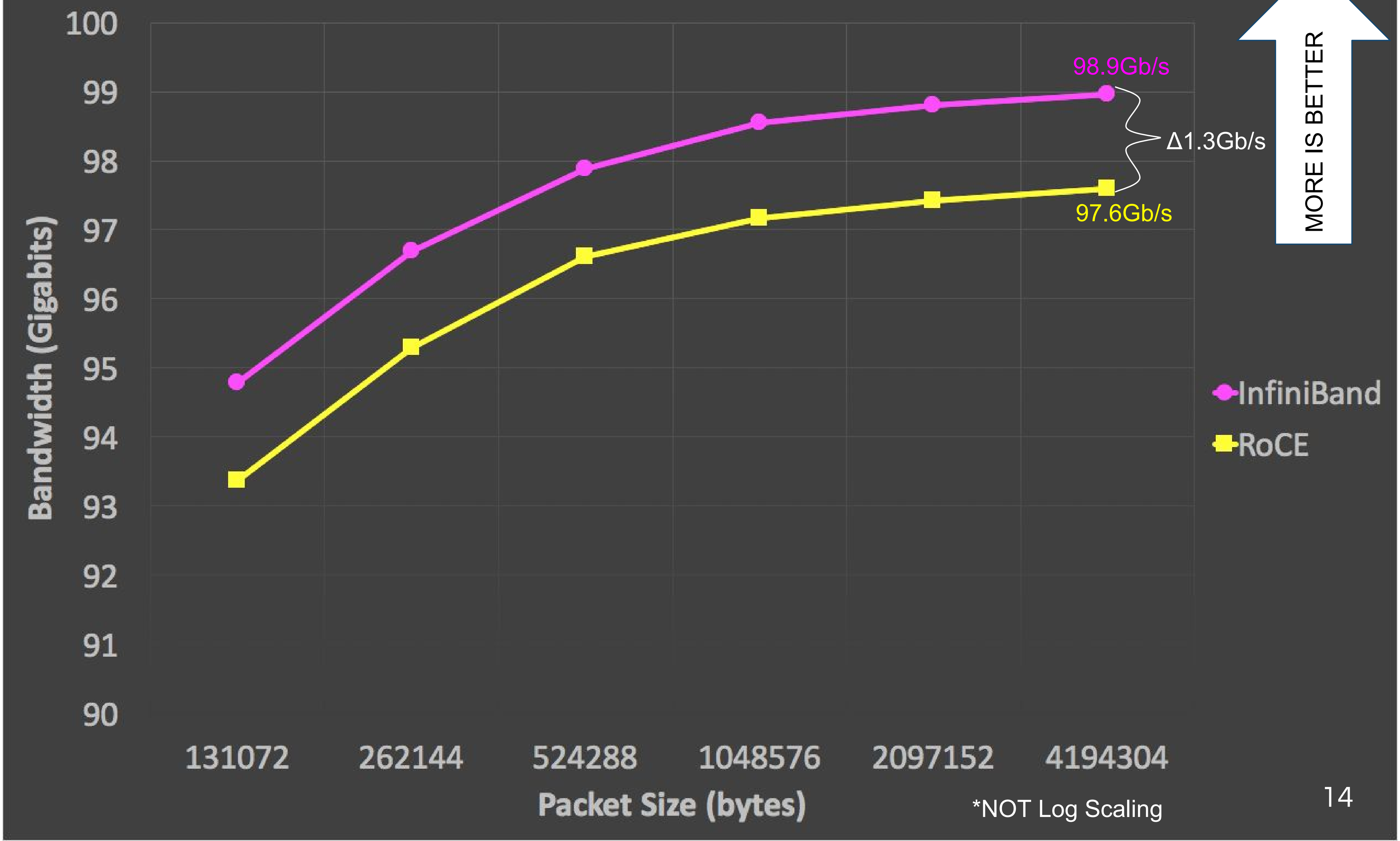
*NOT Log Scaling

# Bandwidth (No switch)

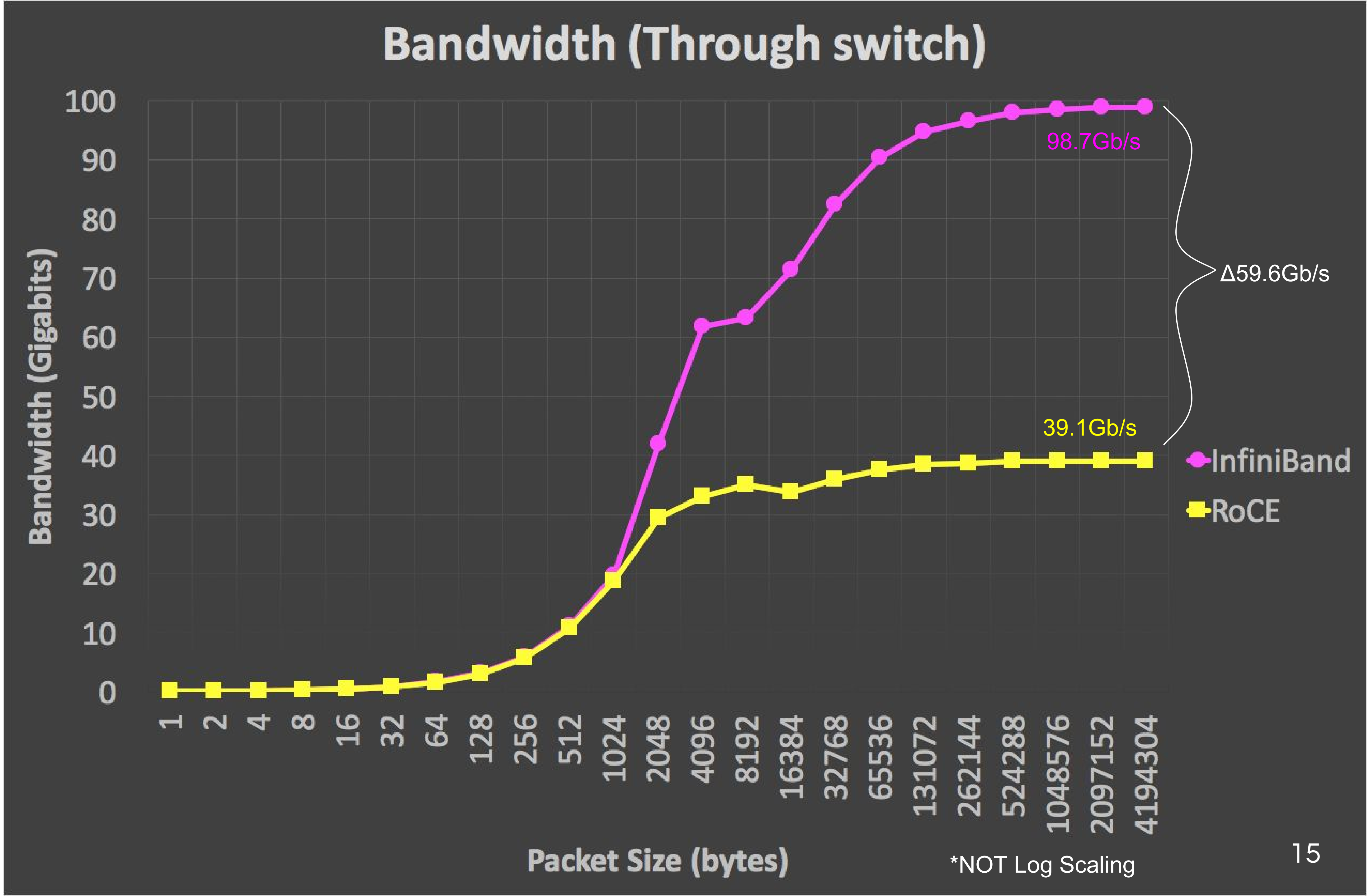


*NOT Log Scaling

# Bandwidth (No switch)

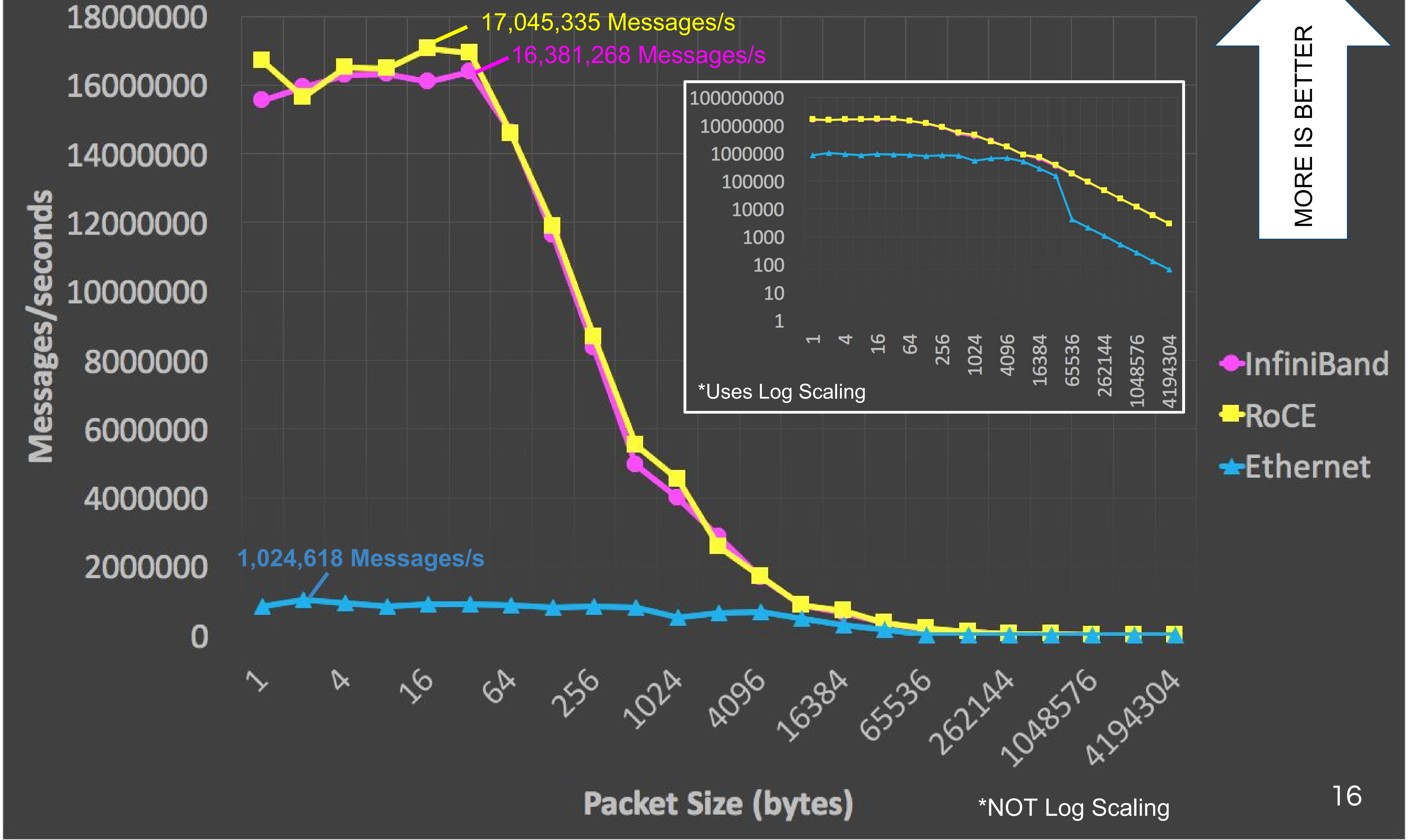


*NOT Log Scaling



*NOT Log Scaling

# Message Rate (8 ranks) (No switch)



17,045,335 Messages/s

16,381,268 Messages/s

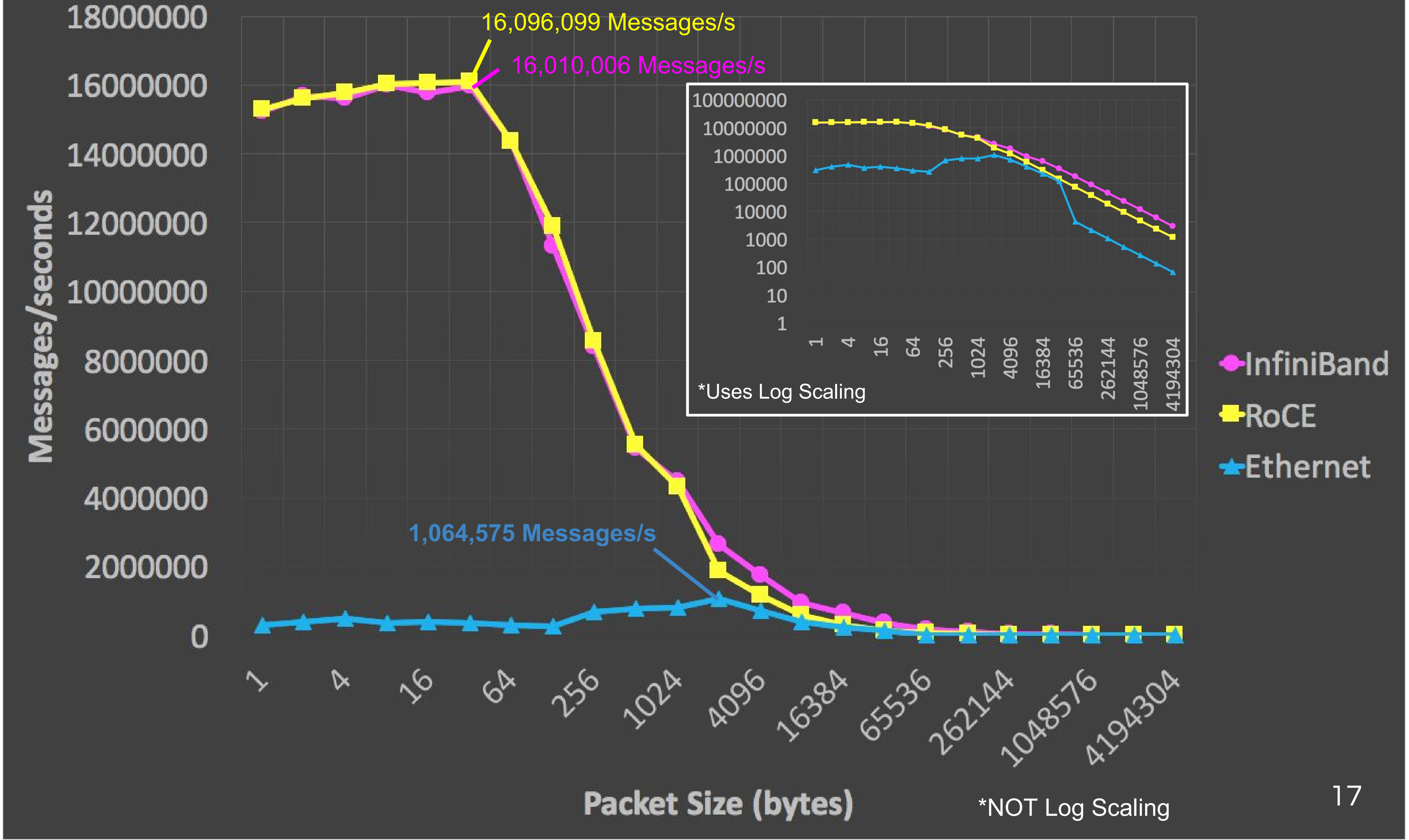
1,024,618 Messages/s

*Uses Log Scaling

*NOT Log Scaling

MORE IS BETTER

# Message Rate (8 ranks) (Through switch)



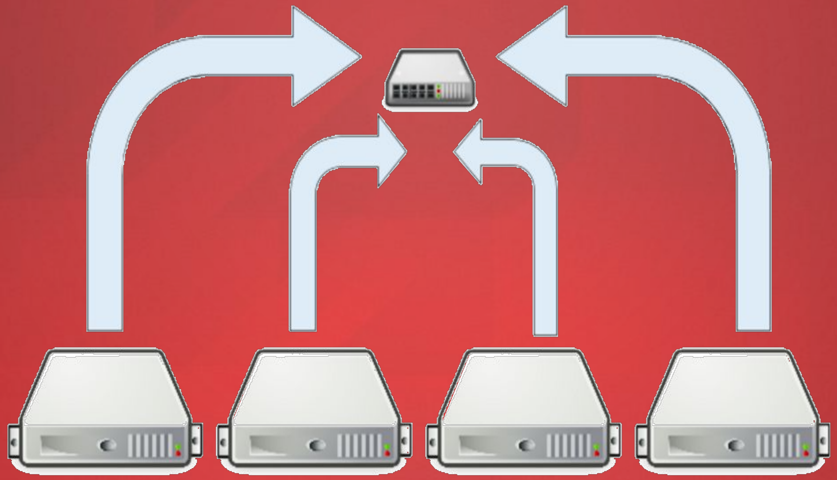
16,096,099 Messages/s

16,010,006 Messages/s

1,064,575 Messages/s

*Uses Log Scaling

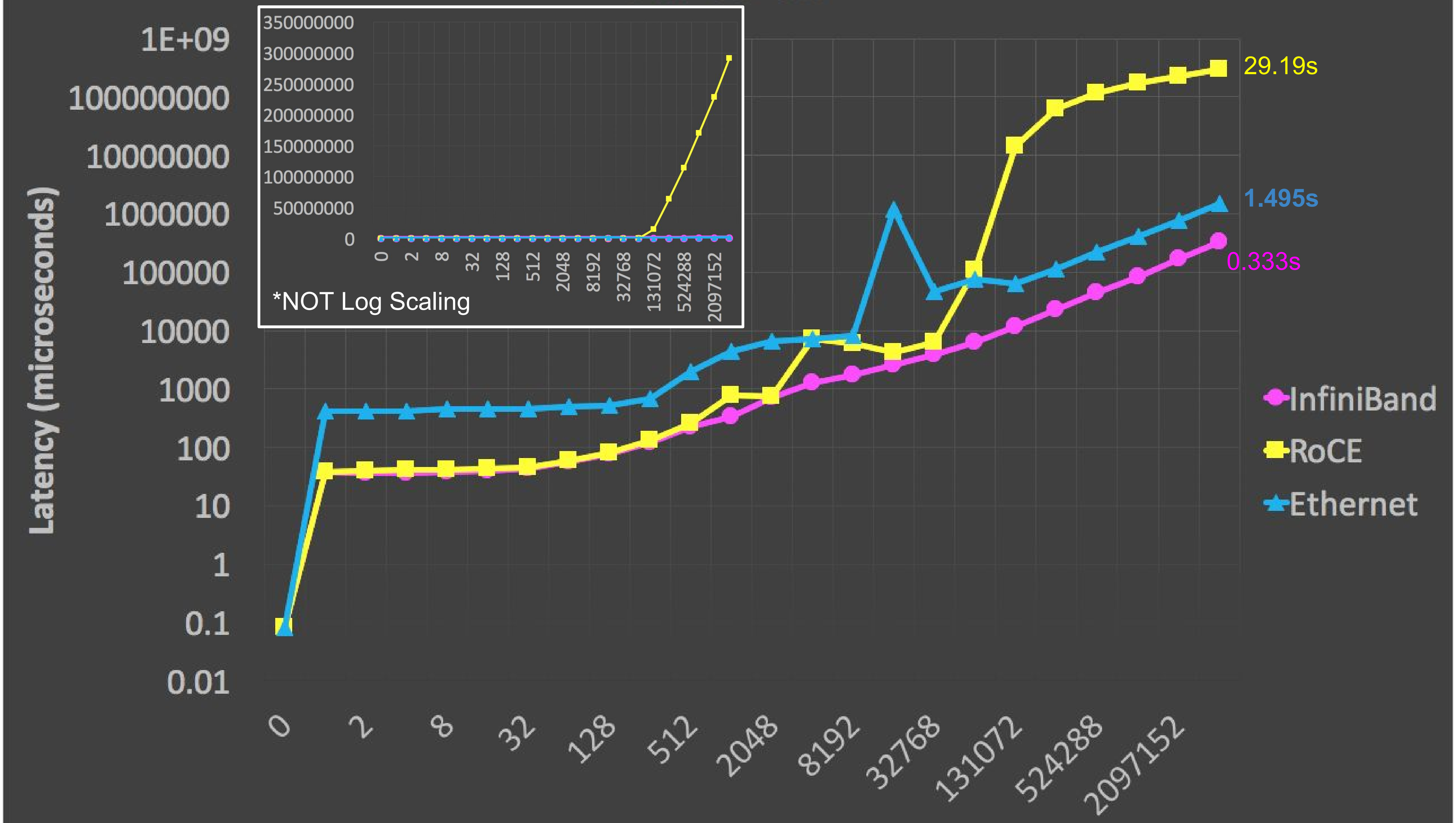
*NOT Log Scaling



# Collective Tests Through Switches

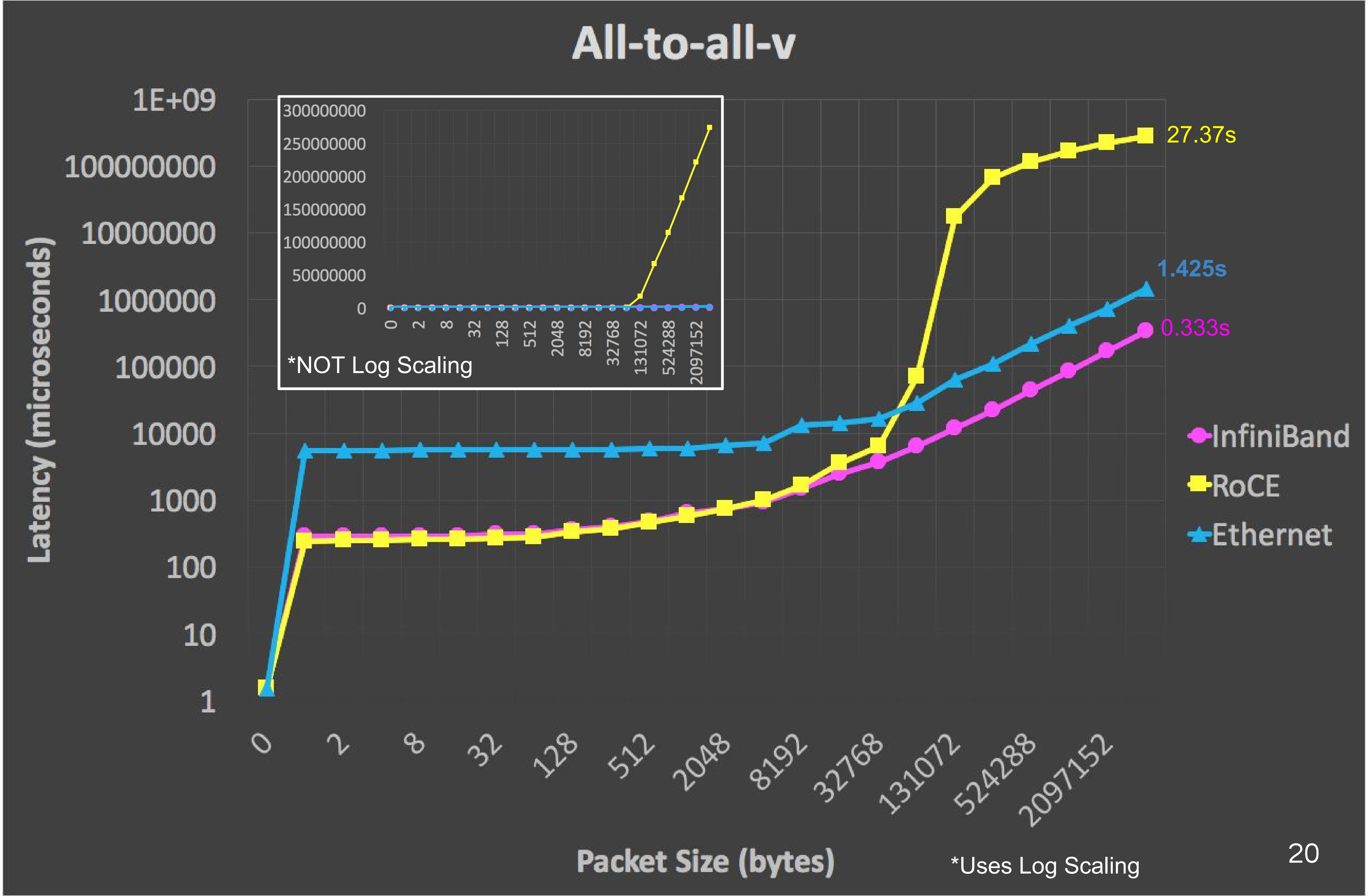
Intel Micro Benchmarks

# All-to-all

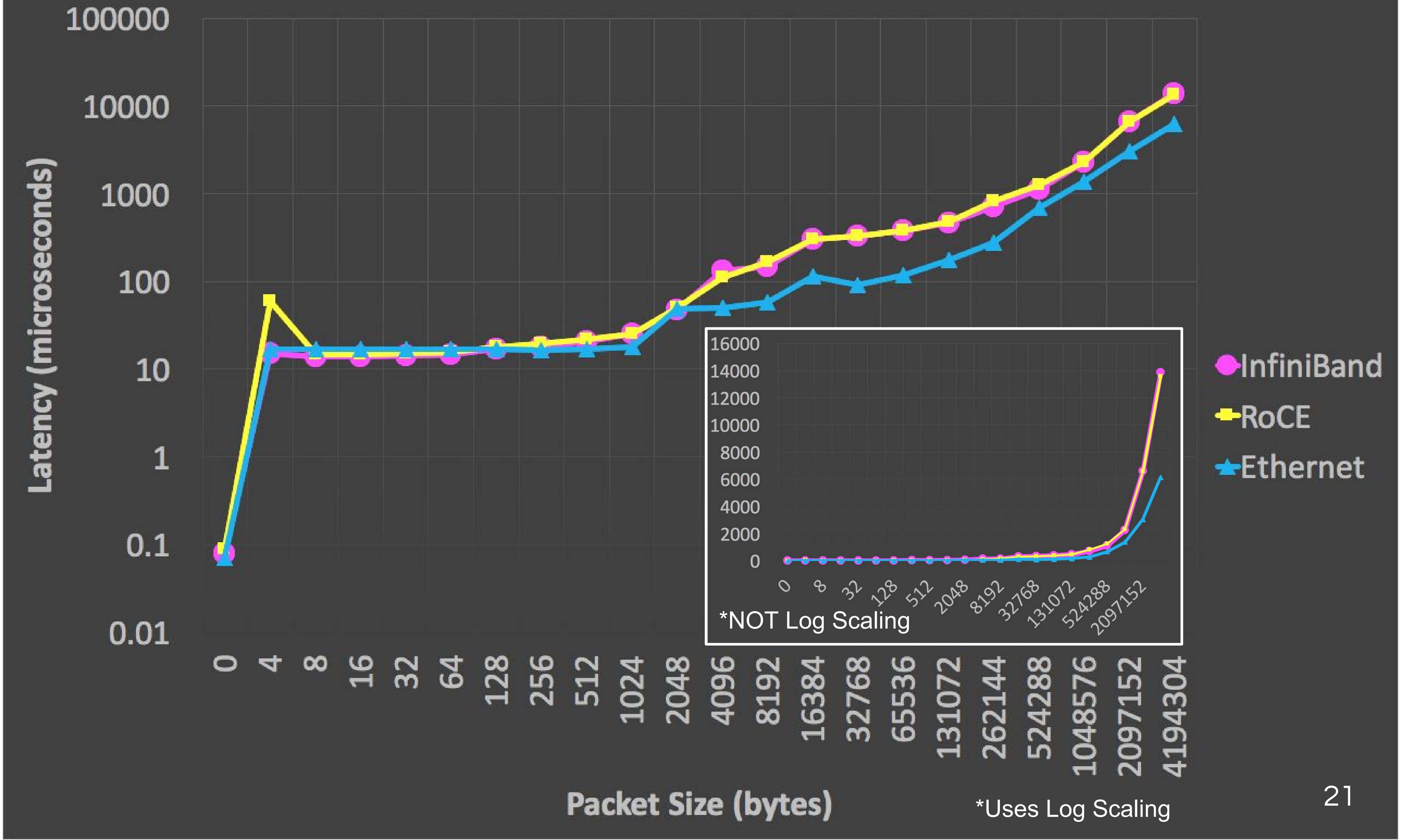


*NOT Log Scaling

*Uses Log Scaling


All-to-all-v
Latency (microseconds)
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1
Packet Size (bytes)
0
2
8
32
128
512
2048
8192
32768
131072
524288
2097152
27.37s
1.425s
0.333s
InfiniBand
RoCE
Ethernet
300000000
250000000
200000000
150000000
100000000
50000000
0
*NOT Log Scaling
*Uses Log Scaling

# All-reduce



*NOT Log Scaling

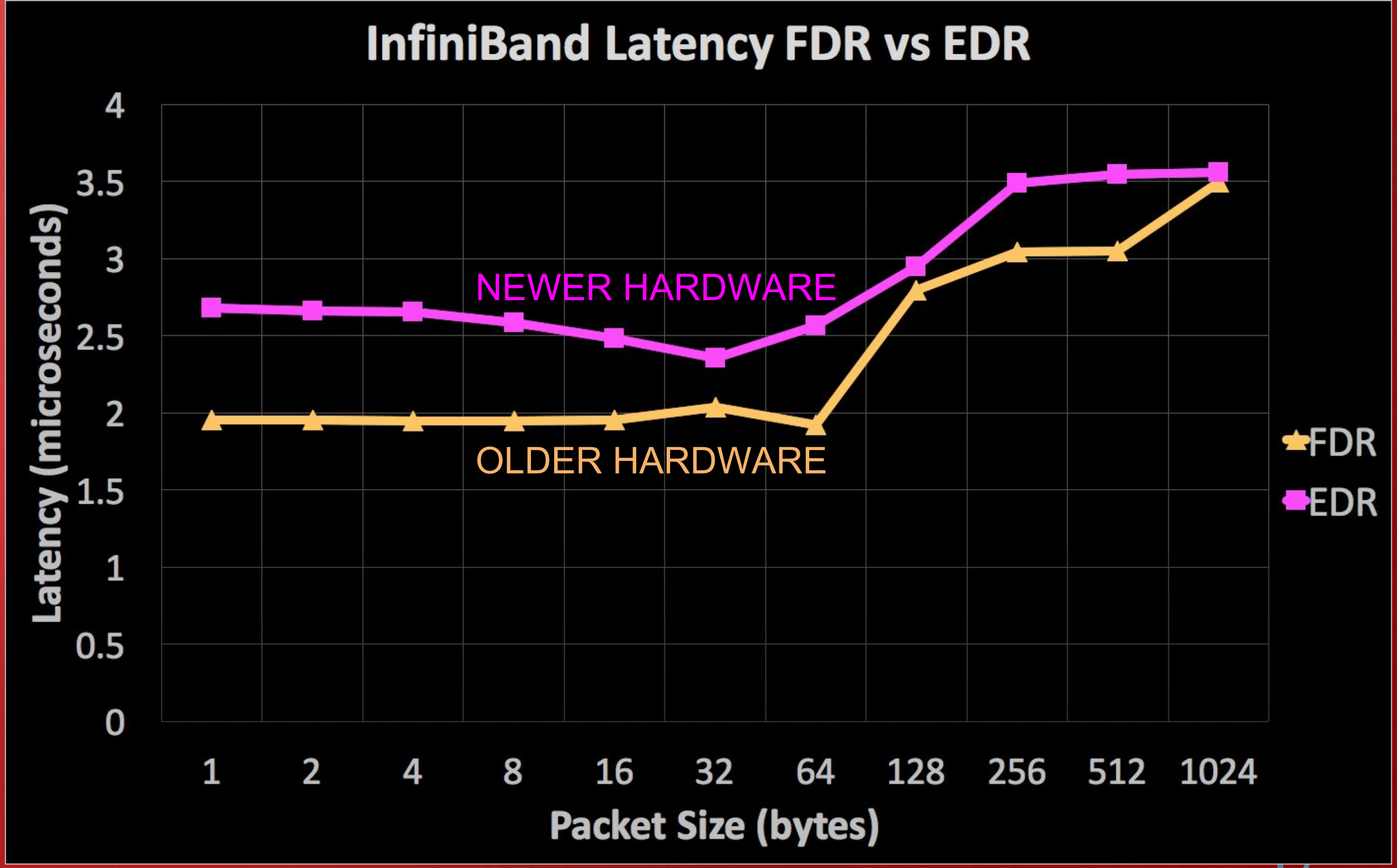
*Uses Log Scaling

# Difficulties Faced

http://fidelisspies.blogspot.com/2014/03/day-212-can-i-live-without-frustration.html

# InfiniBand Setup Issues

- Older version of InfiniBand hardware better initial performance
- Collective tests hung indefinitely



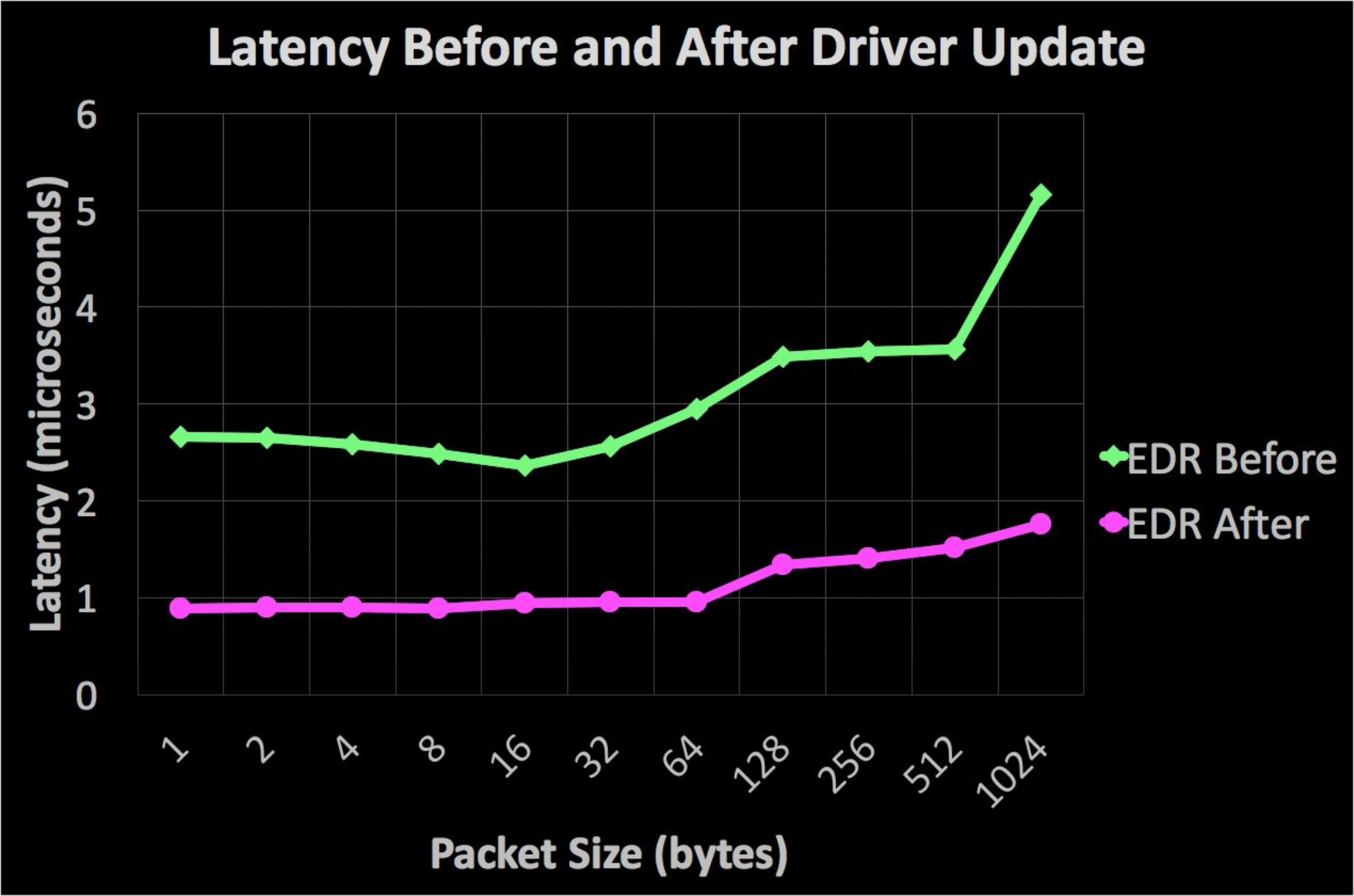
# Impact of InfiniBand Optimization

One-time driver update:

2.2-1 → 3.3-1

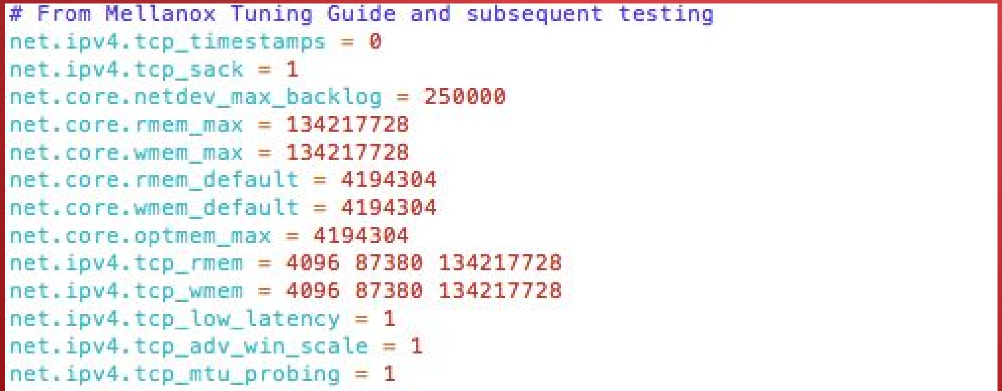
Significant latency improvement!

Drivers will be updated throughout HPC division

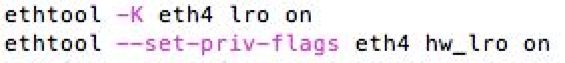


# Ethernet Setup Issues

- Initial iperf tests (with 6 sockets) were only 60Gb/s
  - Eventually reached 87Gb/s through *lots* of configuration



```
# From Mellanox Tuning Guide and subsequent testing
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 1
net.core.netdev_max_backlog = 250000
net.core.rmem_max = 134217728
net.core.wmem_max = 134217728
net.core.rmem_default = 4194304
net.core.wmem_default = 4194304
net.core.optmem_max = 4194304
net.ipv4.tcp_rmem = 4096 87380 134217728
net.ipv4.tcp_wmem = 4096 87380 134217728
net.ipv4.tcp_low_latency = 1
net.ipv4.tcp_adv_win_scale = 1
net.ipv4.tcp_mtu_probing = 1
```

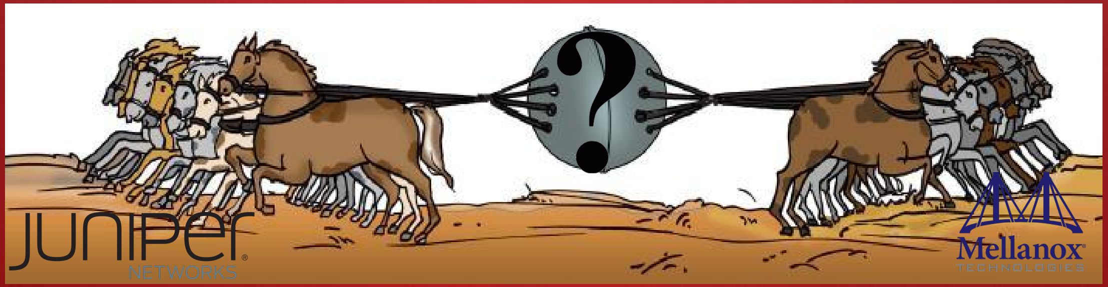


```
ethtool -K eth4 lro on
ethtool --set-priv-flags eth4 hw_lro on
```

- OpenMPI gives <20Gb/s for native Ethernet
  - Potentially lack of socket optimization

# Juniper Setup Issues

- Juniper switch reported Mellanox passive 100Gb copper cables as 40Gb/s
- EDR cards didn’t acknowledge Juniper passive optics whatsoever
- Juniper switch didn’t acknowledge Mellanox active optics



http://www.ekshiksha.org.in/eContent-Show.do?documentId=59

# Mellanox Card Compatibilities

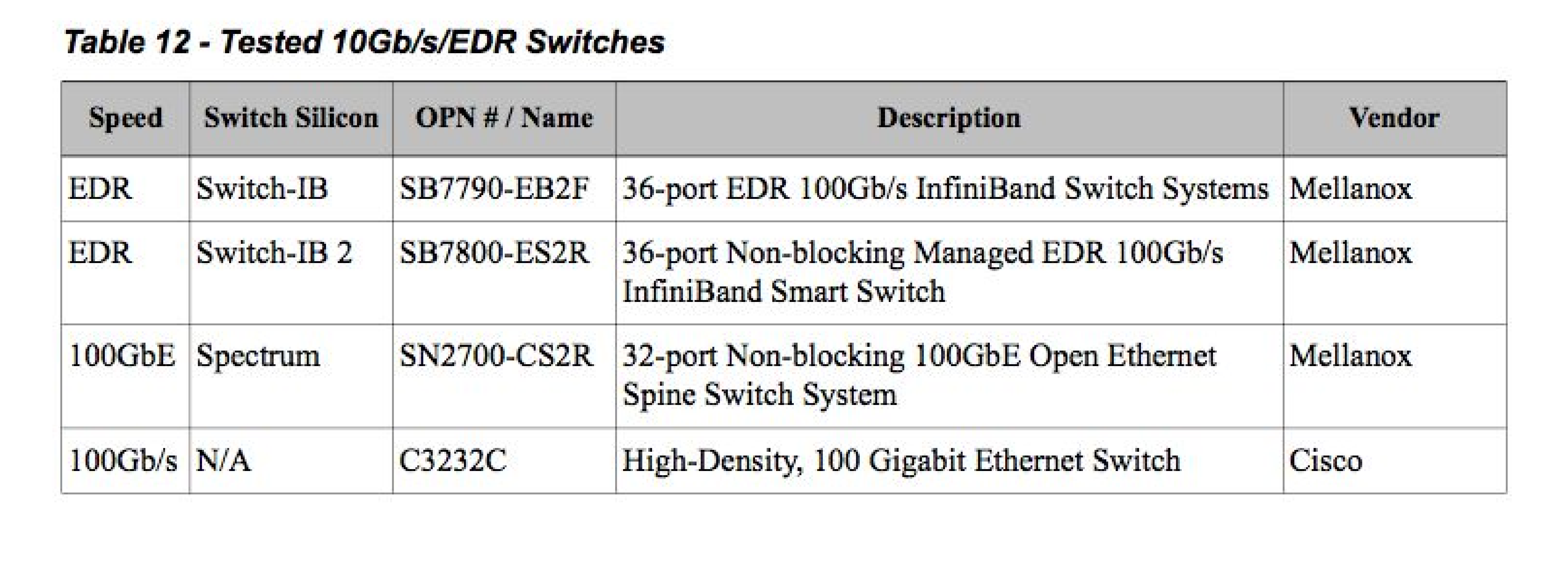
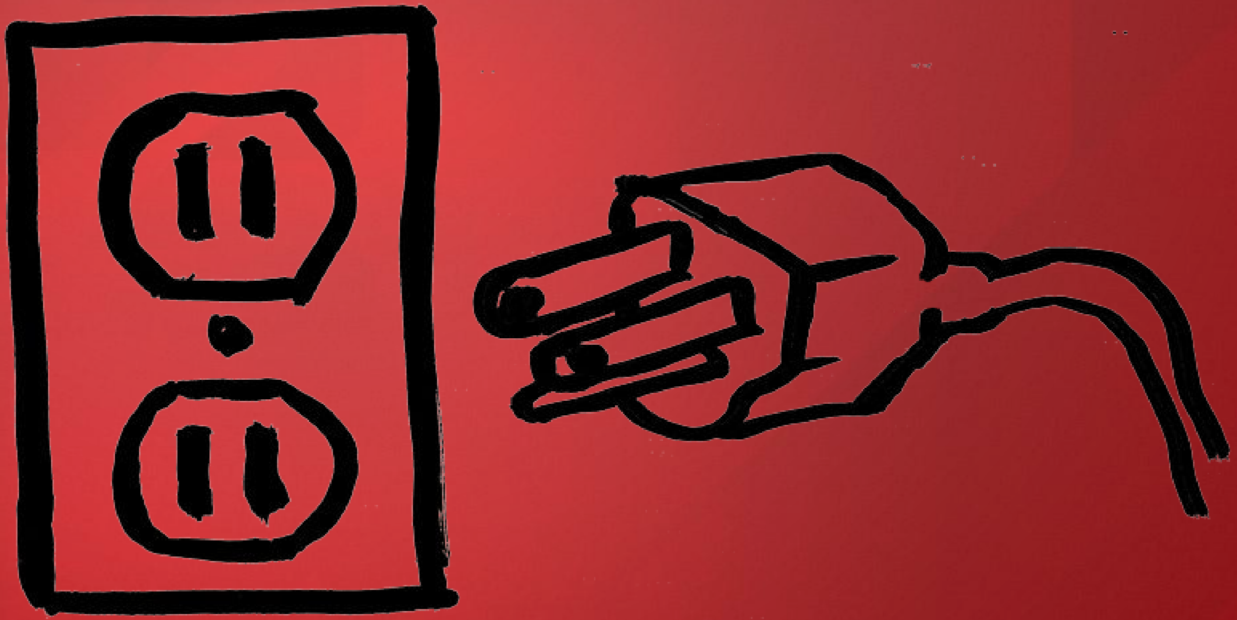


***Table 12 - Tested 10Gb/s/EDR Switches***

| Speed | Switch Silicon | OPN # / Name | Description | Vendor |
|---|---|---|---|---|
| EDR | Switch-IB | SB7790-EB2F | 36-port EDR 100Gb/s InfiniBand Switch Systems | Mellanox |
| EDR | Switch-IB 2 | SB7800-ES2R | 36-port Non-blocking Managed EDR 100Gb/s InfiniBand Smart Switch | Mellanox |
| 100GbE | Spectrum | SN2700-CS2R | 32-port Non-blocking 100GbE Open Ethernet Spine Switch System | Mellanox |
| 100Gb/s | N/A | C3232C | High-Density, 100 Gigabit Ethernet Switch | Cisco |

# Conclusion



http://blog.friendfriend.net/2014/05/10/54-incompatible-plug/

# Conclusion

- Good potential, as shown with direct results.
- 100Gb technology is too new, not standardized
  - Deployment Effort is complex for both options
- Different companies not necessarily compatible
- If you want 100Gb/s, get it all from one place.

# Greater Impact

- Mellanox OFED driver updates throughout HPC Division
- Juniper and Mellanox will begin to collaborate with hardware compatibilities
  - Because of us, they have a place to start

# Future Work

- Comparison with Intel’s OmniPath
- Collective tests after technologies are compatible
- Work with OpenMPI community to optimize Ethernet performance
- NFS over RDMA
- TCP/IP over InfiniBand

# ACKNOWLEDGEMENTS

Mellanox TECHNOLOGIES

Aleksey Senin

Martijn Van Breugel

JUNIPER NETWORKS

Erik DeHaas

Matthew Herzog

Los Alamos NATIONAL LABORATORY EST.1943

Our mentors:

Susan Coulter

Howard Pritchard

Our instructors:

Jarrett Crews

Eric Holm

http://shottec.com/wp-content/uploads/2014/09/Shot-Tec-Thank-You.png

# Questions?

Faith Van Wig fvvnr5@mst.edu

Luke Kachelmeier lkachelmeier@unm.edu

Kari Erickson kari.erickson@student.nmt.edu

```
    o O___  _________  _________  _________  _________  _________  _________  _________
  _] [__|o | |O O O O| |O O O O| |O O O O| |O O O O| |O O O O| |O O O O| |O O O O|
 <________| |________| |________| |________| |________| |________| |________| |________|
  /O -O -O     o    o     o    o     o    o     o    o     o    o     o    o     o    o
```

# Backup Slides

## ifcfg-eth4

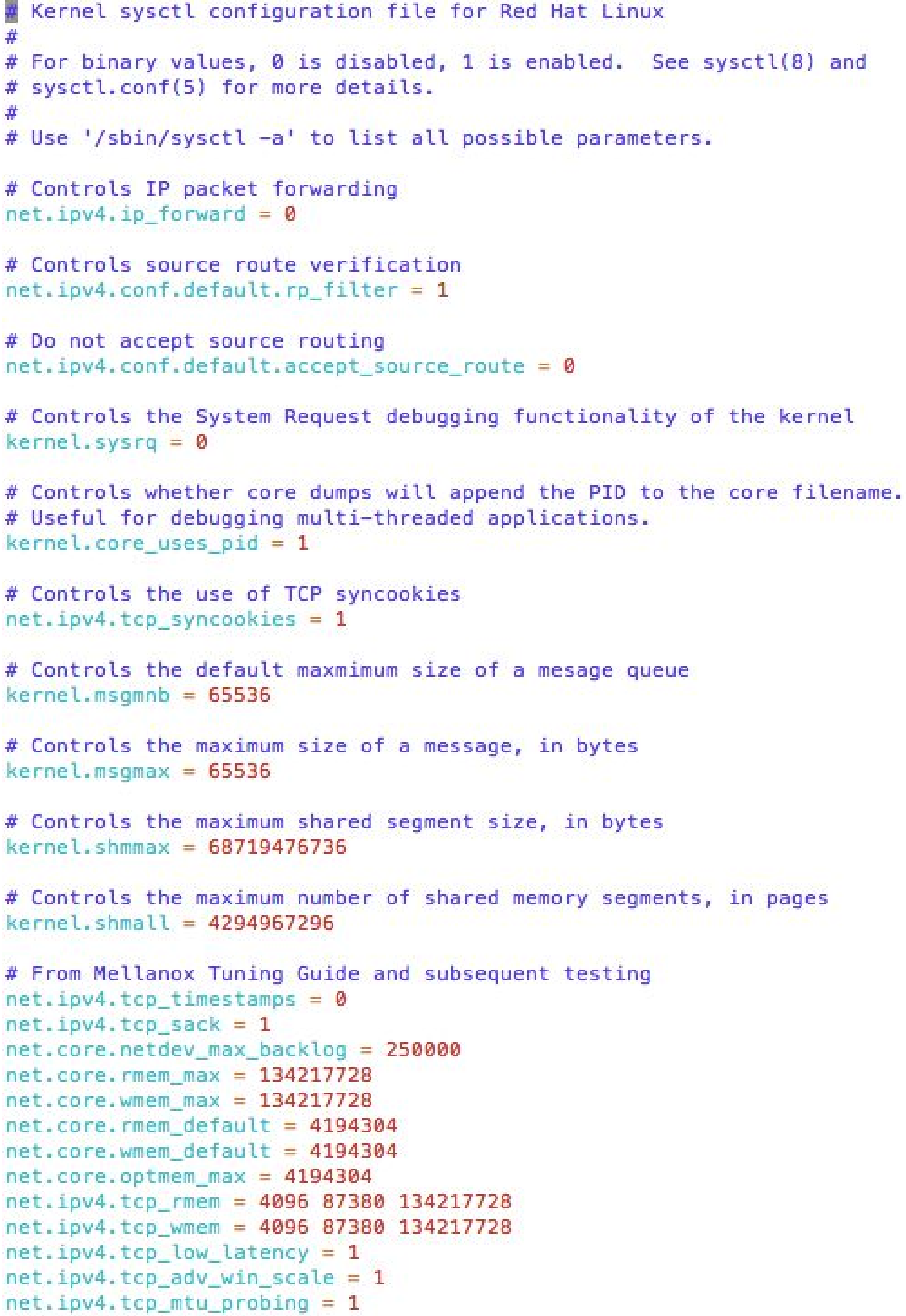
```
DEVICE=eth4
BOOTPROTO=dhcp
PEERDNS=no
ONBOOT=yes
GATEWAY=192.168.1.1
MTU=9000
```

## rc.local (Onboot)

```
ethtool -K eth4 lro on
ethtool --set-priv-flags eth4 hw_lro on
```

## whenever you run OMPI:

```
export OMPI_MCA_tcp_max_send_size=2097152
export OMPI_MCA_btl_tcp_if_include=eth4
export OMPI_MCA_btl_tcp_eager_limit=8388608
export OMPI_MCA_btl_tcp_max_send_size=2097152
export OMPI_MCA_btl_tcp_links=1
export OMPI_MCA_btl_tcp_sendbuf=4194304
export OMPI_MCA_btl_tcp_recvbuf=4194304
```



```
# Kernel sysctl configuration file for Red Hat Linux
#
# For binary values, 0 is disabled, 1 is enabled.  See sysctl(8) and
# sysctl.conf(5) for more details.
#
# Use '/sbin/sysctl -a' to list all possible parameters.

# Controls IP packet forwarding
net.ipv4.ip_forward = 0

# Controls source route verification
net.ipv4.conf.default.rp_filter = 1

# Do not accept source routing
net.ipv4.conf.default.accept_source_route = 0

# Controls the System Request debugging functionality of the kernel
kernel.sysrq = 0

# Controls whether core dumps will append the PID to the core filename.
# Useful for debugging multi-threaded applications.
kernel.core_uses_pid = 1

# Controls the use of TCP syncookies
net.ipv4.tcp_syncookies = 1

# Controls the default maxmimum size of a mesage queue
kernel.msgmnb = 65536

# Controls the maximum size of a message, in bytes
kernel.msgmax = 65536

# Controls the maximum shared segment size, in bytes
kernel.shmmax = 68719476736

# Controls the maximum number of shared memory segments, in pages
kernel.shmall = 4294967296

# From Mellanox Tuning Guide and subsequent testing
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 1
net.core.netdev_max_backlog = 250000
net.core.rmem_max = 134217728
net.core.wmem_max = 134217728
net.core.rmem_default = 4194304
net.core.wmem_default = 4194304
net.core.optmem_max = 4194304
net.ipv4.tcp_rmem = 4096 87380 134217728
net.ipv4.tcp_wmem = 4096 87380 134217728
net.ipv4.tcp_low_latency = 1
net.ipv4.tcp_adv_win_scale = 1
net.ipv4.tcp_mtu_probing = 1
```