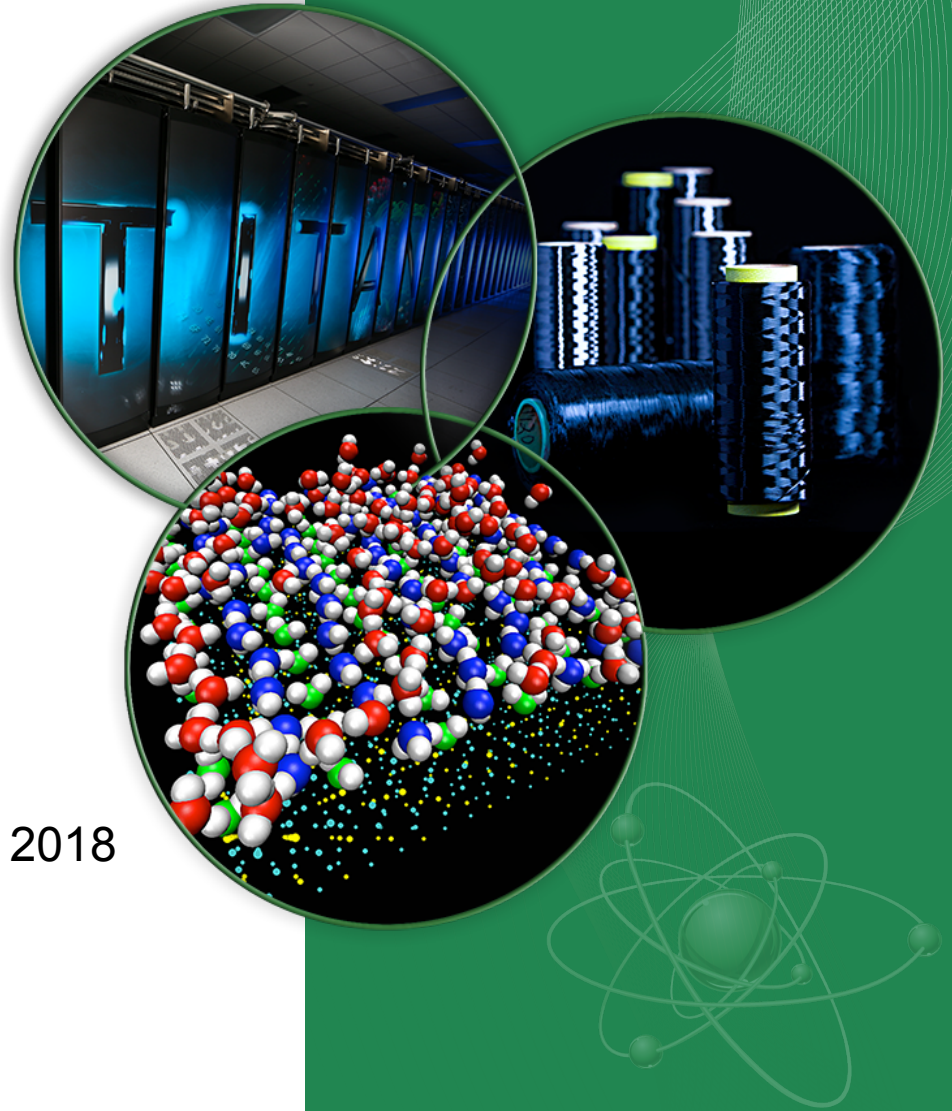


# Analytics of Wide-Area Lustre Throughput Using LNet Routers

Nagi Rao, Neena Imam,  
Jesse Hanley, Sarp Oral

Oak Ridge National Laboratory

Lustre User Group Conference – LUG 2018  
April 24-26, 2018  
Argonne National Laboratory  
Argonne, IL



# Outline

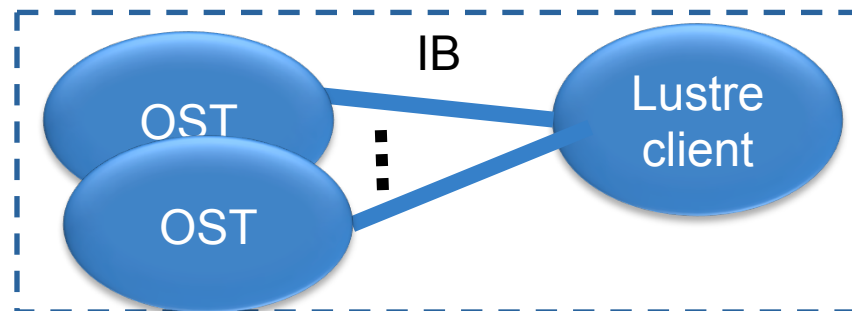
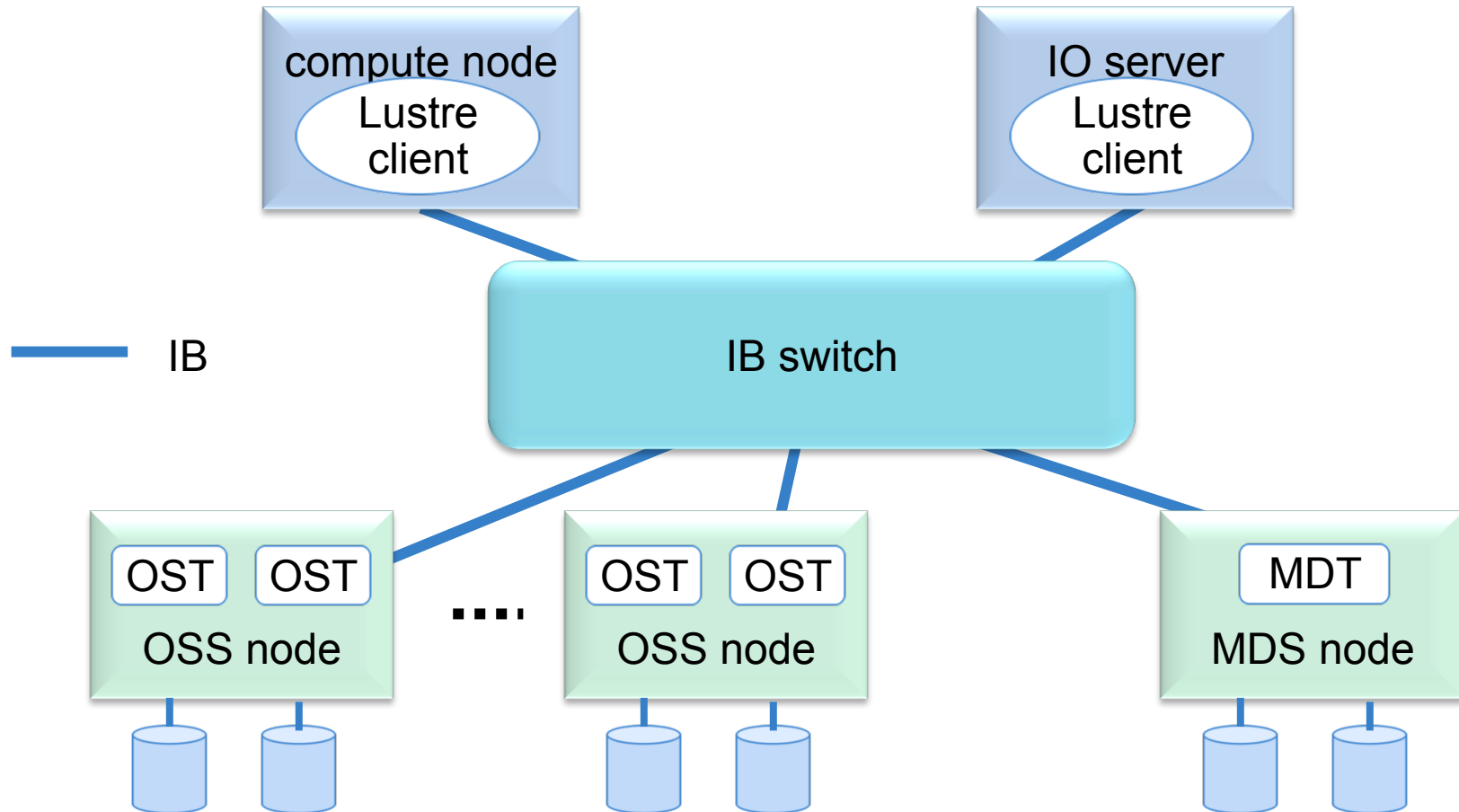
- Introduction
- Lustre Basic Configurations
- Luster Over Ethernet
- Test Configurations and Measurements
- LNet Buffers Impact
- Conclusions

# Lustre: Distributed File System

## Lustre file system

- One or more Meta Data Server (MDS)
  - supported by Meta Data Targets (MDT)
- One or more Object Storage Servers (OSS)
  - supported by one or more Object Storage Target (OST)
- Mounted by Lustre clients on hosts (IO and compute)
  - over IB, Ethernet, and other networks
  - compute host: typically nodes in a cluster – computations with files accesses
  - IO or Data Transfer Node (DTN) hosts: typically used for bulk and storage operations
- Lustre achieves high performance by effectively parallelizing I/O from multiple clients to multiple OSTs
  - files striped over multiple OSTs per a pre-defined pattern

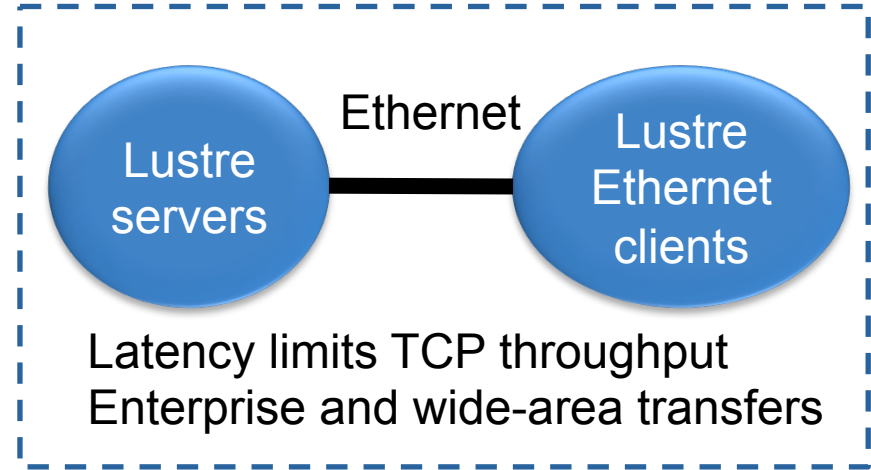
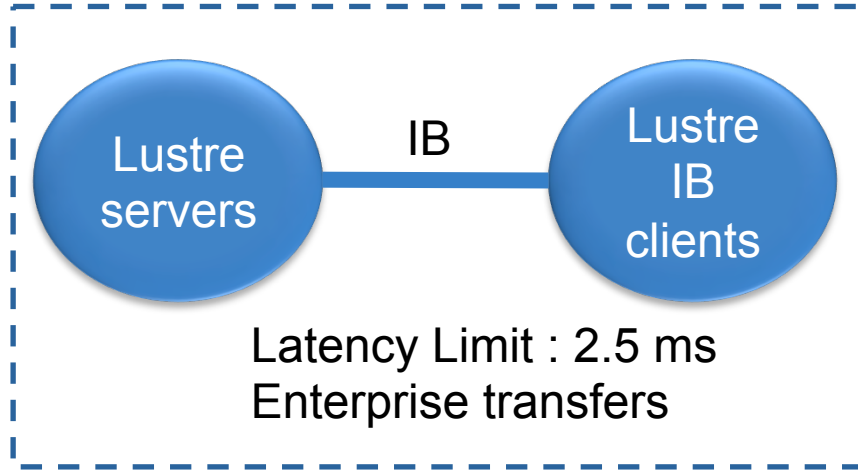
# Lustre over IB: Basic Configuration



# Lustre Over Wide-Area

- Desired features: Lustre mounted over wide-area
  - Obviates need for transfer services such as GridFTP, Aspera, XDD and others
  - Easier application integration with remote file operations
- Current Installations
  - Majority are supported over site IB networks
  - Time-out limitation: 2.5ms
  - IB WAN extenders: too expensive and not flexible
- Lustre over Ethernet (not as widely deployed)
  - TCP/IP implementation: uses existing networks
  - Very little infrastructure enhancements needed

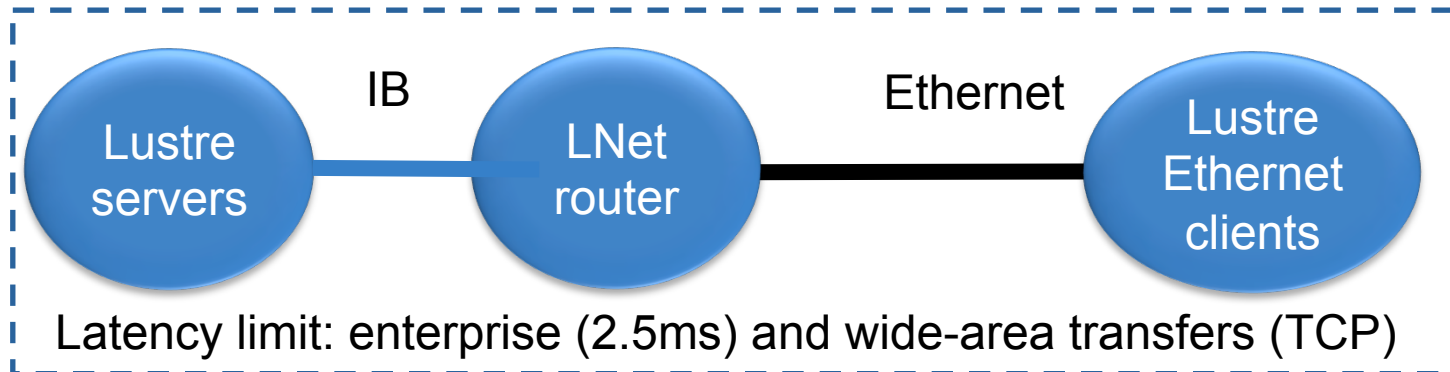
# Basic IB and Ethernet Lustre Configurations



Lustre over IB: limited to enterprises – 2.5ms latency limit

Lustre over Ethernet for wide-area – TCP version limited by throughput

LNet routers: extent enterprise Lustre/IP to wide-area using Luster/Ethernet



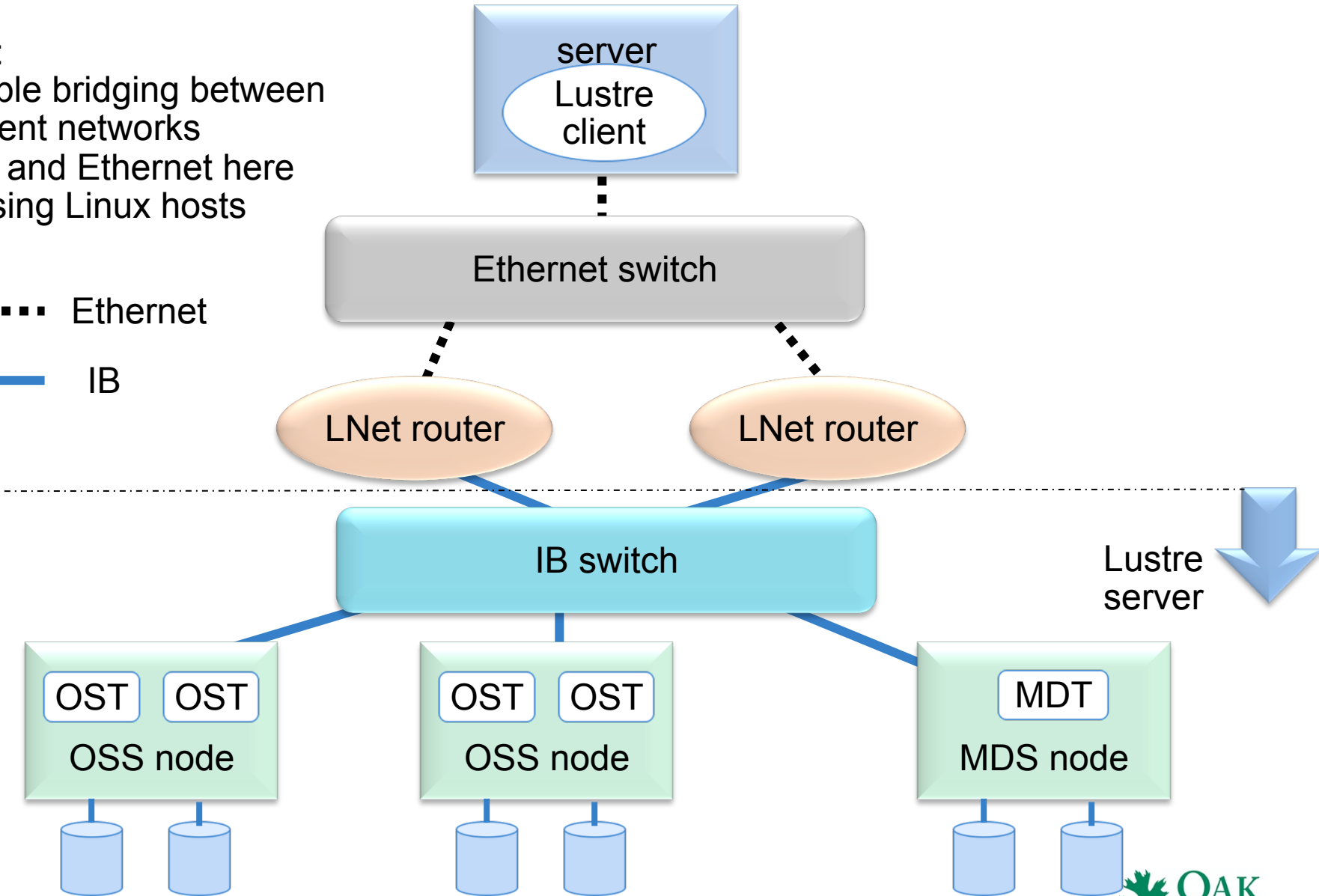
# Lustre over IB-Ethernet: LNet routers

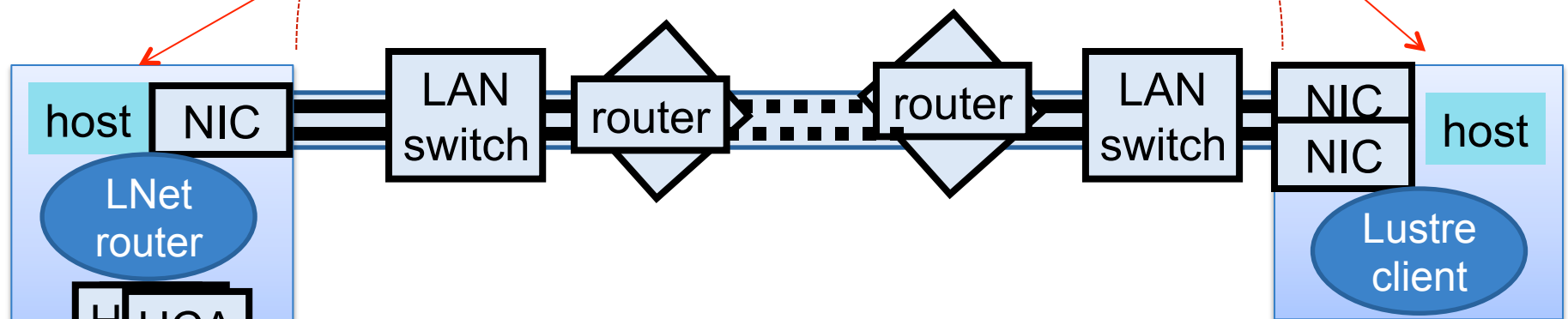
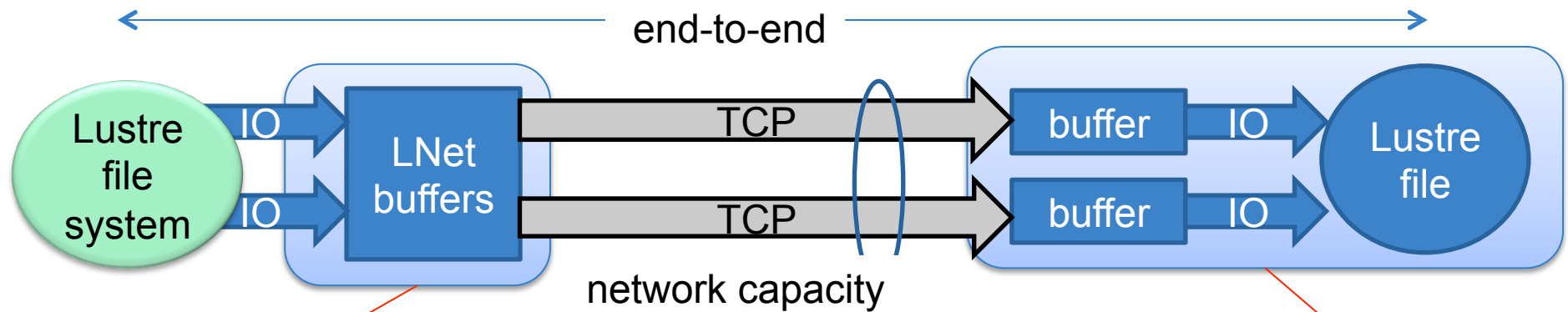
LNet:  
Flexible bridging between  
different networks

- IB and Ethernet here
- Using Linux hosts

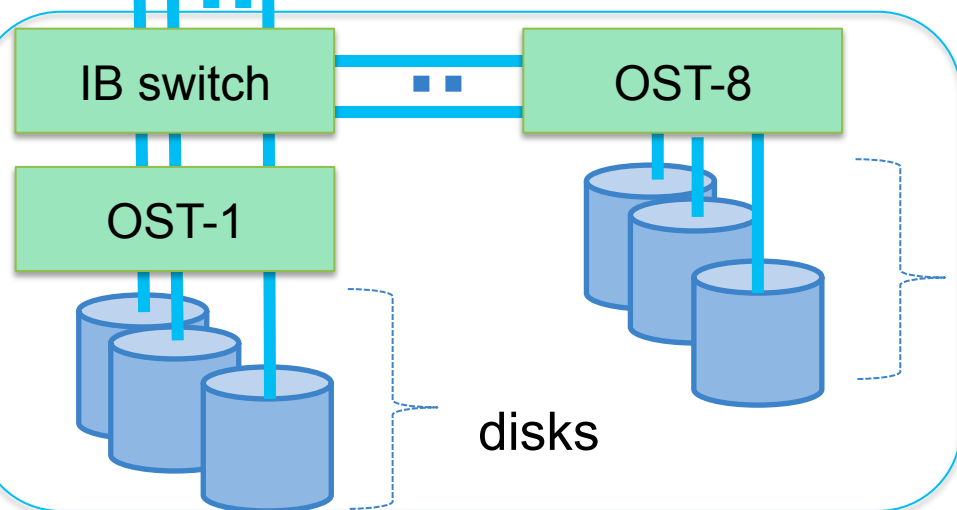
..... Ethernet

— IB





site IB Lustre



ORNL Testbed  
Configurations



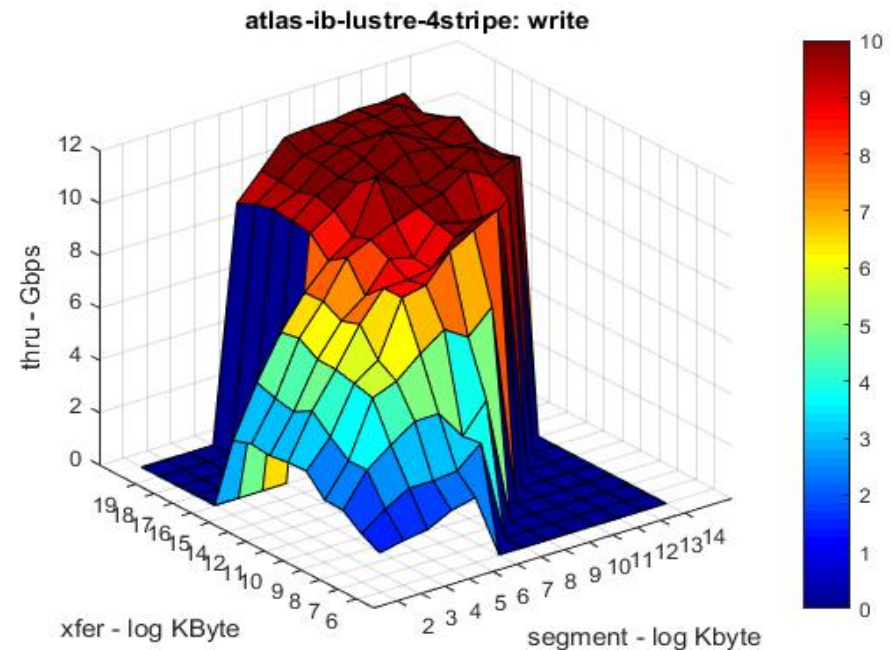
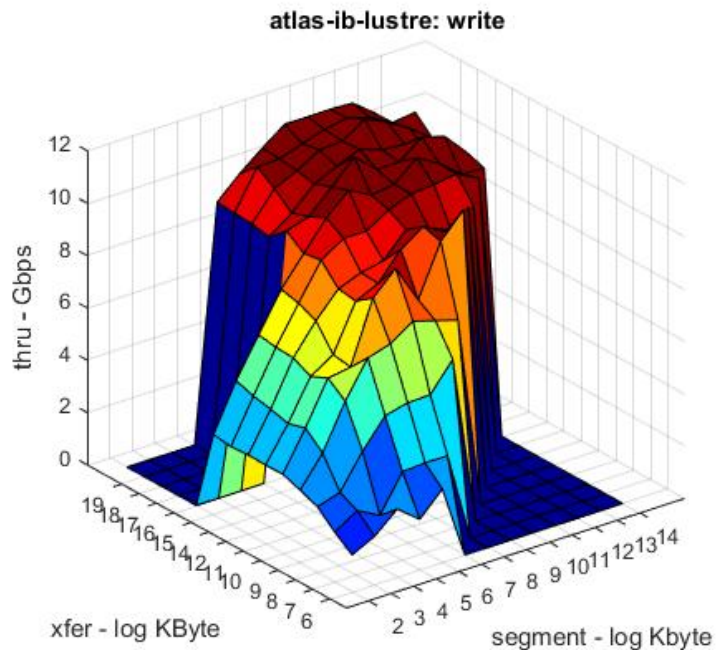
# Lustre and LNet Parameter Setup

- Lustre parameters:
  - Client-level: lustre\_ctrl
    - number of credits: default:8; current 256
  - File/Directory Level:
    - number of stripes: 2,8
    - stripe size: default
- LNet parameters
  - Base parameters: lnctl
  - ksocklnd.conf: LNet buffer-size range: 50K – 2G; default 65K
- TCP parameters
  - Congestion-control modules: CUBIC and Hamilton TCP
  - Buffer sizes: 200ms recommended values, largest allowed

# Lustre Throughput: IOZone (IB - Reference)

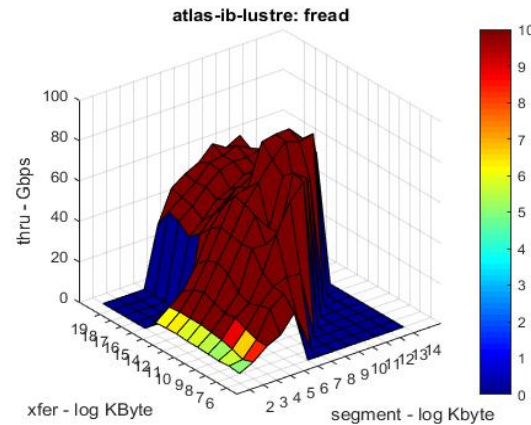
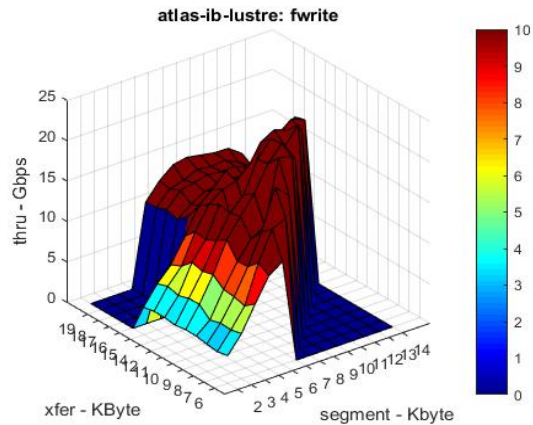
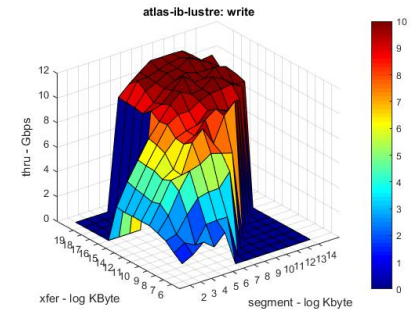
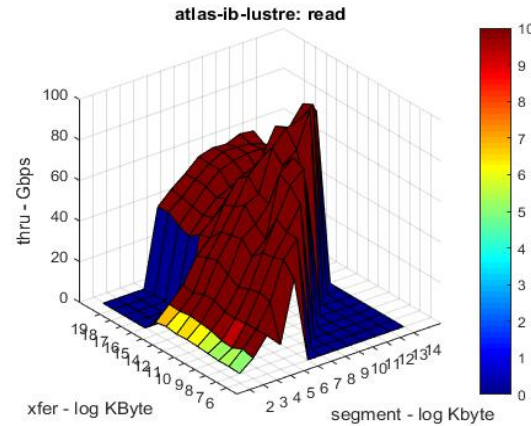
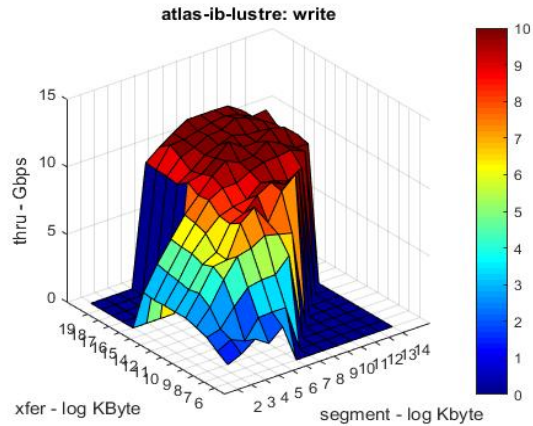
- ORNL OLCF

- Atlas/Spider Lustre system: 2K OSSs
- Data Transfer Node (DTN) server
- Write throughput: stable ~10Gigabits/sec (Gbps)  
= ~1.25GigaBytes/sec (GBps)



# IO Zone: Lustre Throughput

- ORNL OLCF
  - Write throughput: stable ~10Gbps



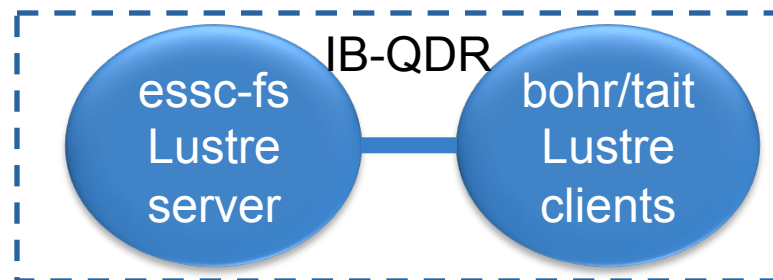
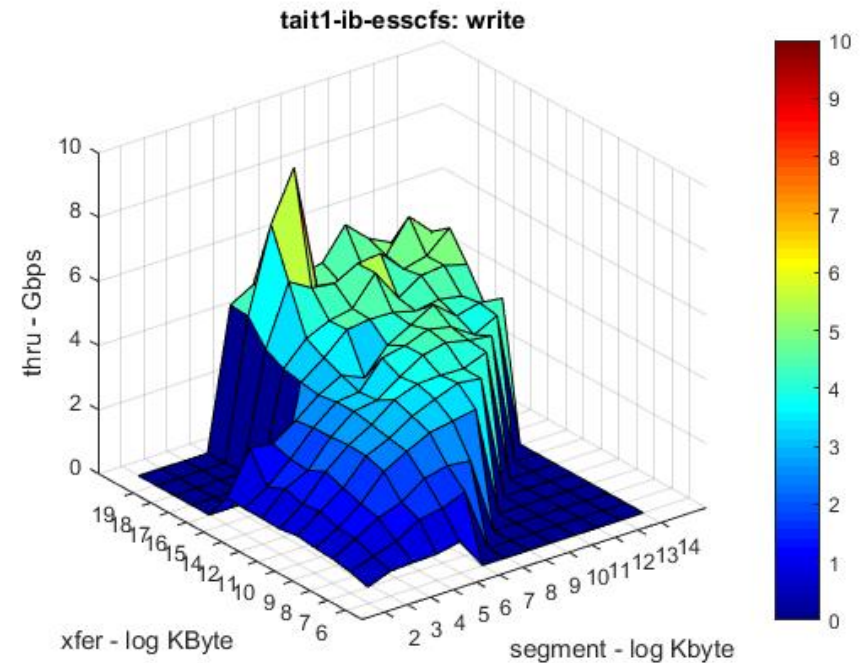
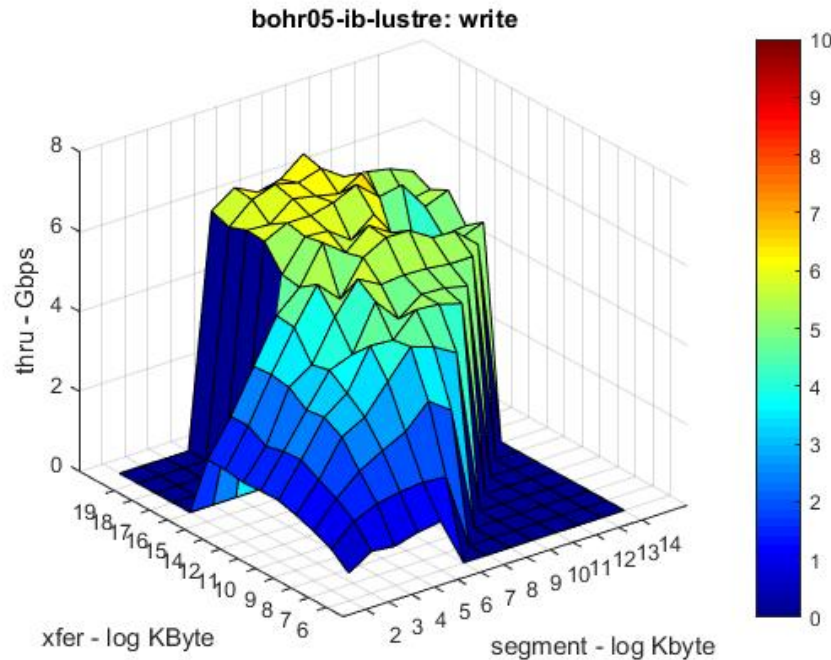
# Site Lustre over IB: ORNL Testbed

bohr IO servers - centos 6.8

- 48core, opteron, 2.2 GHz
- write peak throughput: ~6Gbps

tait compute servers – centos 6.8

- 24 core, xeon 2.6GHz
- Write peak throughput: ~5Gbps



# Site Lustre over IB: Test Hosts

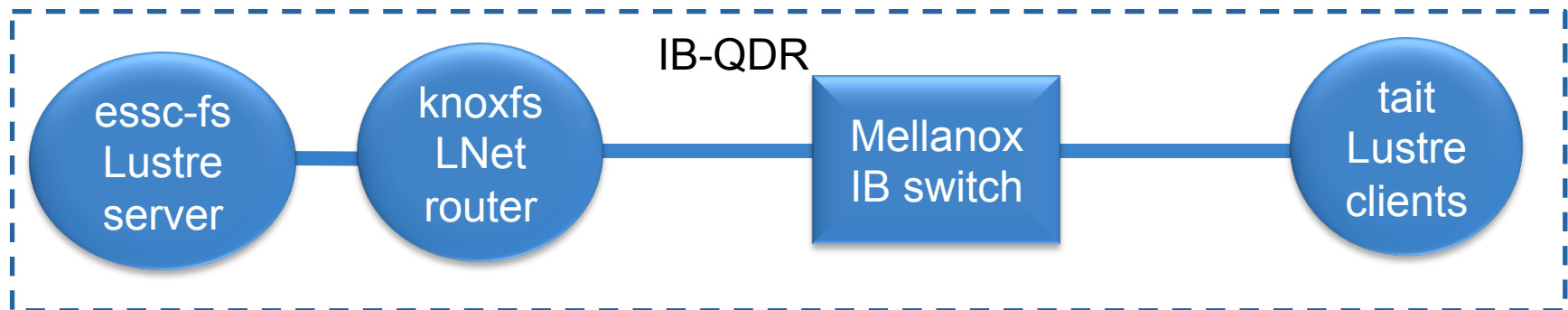
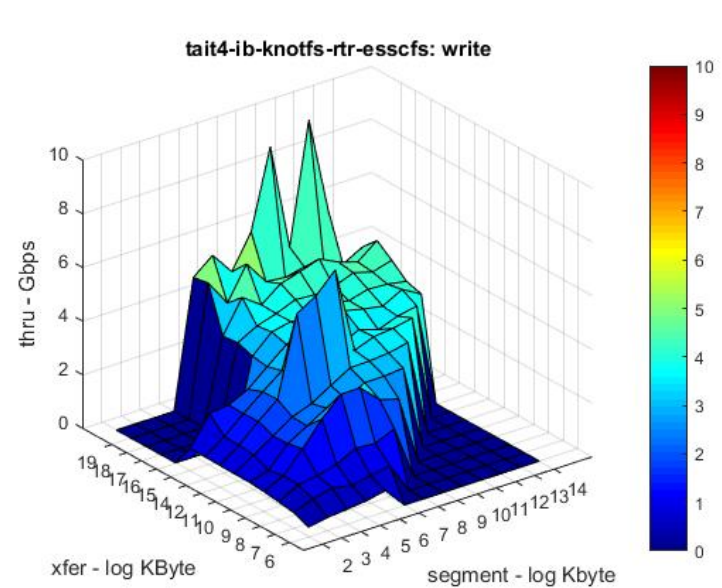
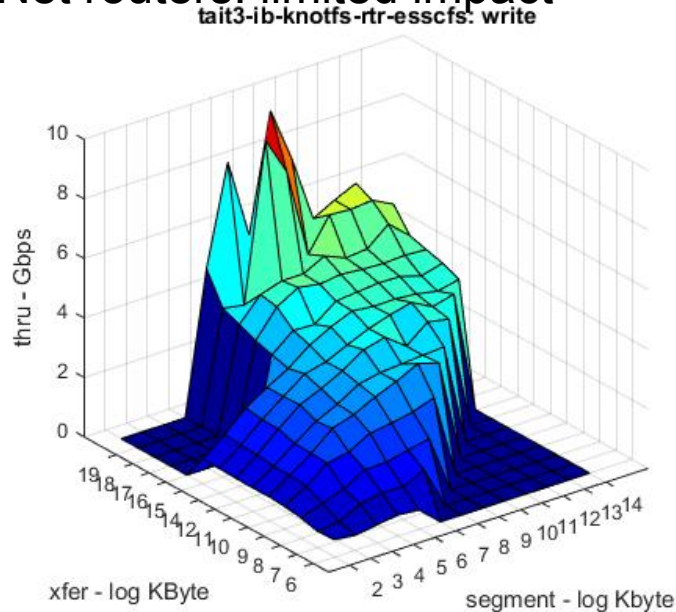
- bohr specs – primarily used for file transfers
  - HP ProLiant DL585 G7
  - AMD Opteron(tm) Processor 6176 SE: 48 cores and two sockets
  - 48x 8GB DDR3 1066MHz DIMMs
  - InfiniBand: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR / 10GigE] (rev b0): MT26428
  - Ethernet: Mellanox Technologies MT27700 Family [ConnectX-4]: MT4115
- tait specs – node on compute clusters
  - Dell PowerEdge R720
  - Dual Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz: 24 cores
  - 8x Micro Technology 16GB DDR3 1600MHz DIMMs
  - InfiniBand: Mellanox Technologies MT27500 Family [ConnectX-3]: MT4099
  - Ethernet: Mellanox Technologies MT27700 Family [ConnectX-4]: MT4115



# Site Lustre: LNet over IB Networks

tait compute servers – centos 6.8: write peak throughput: ~5Gbps

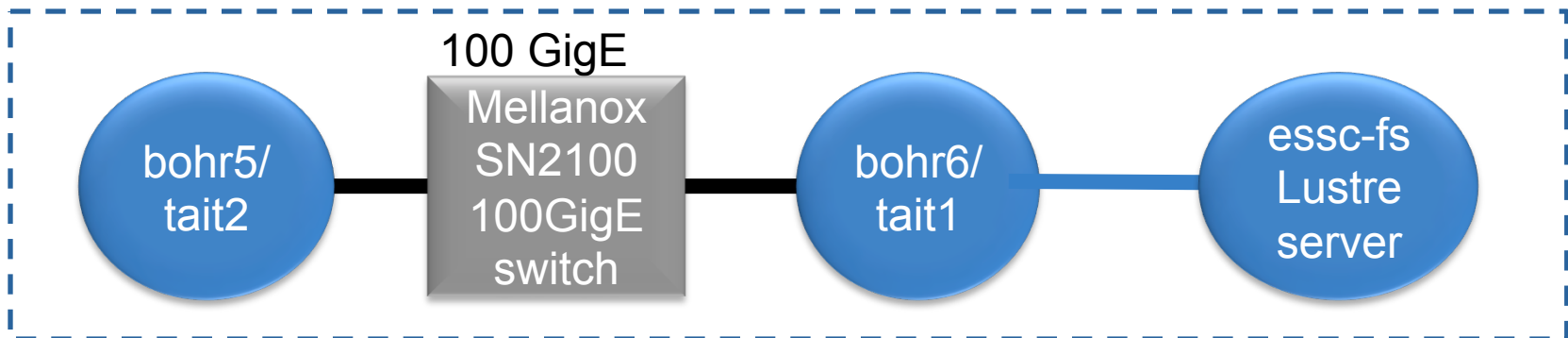
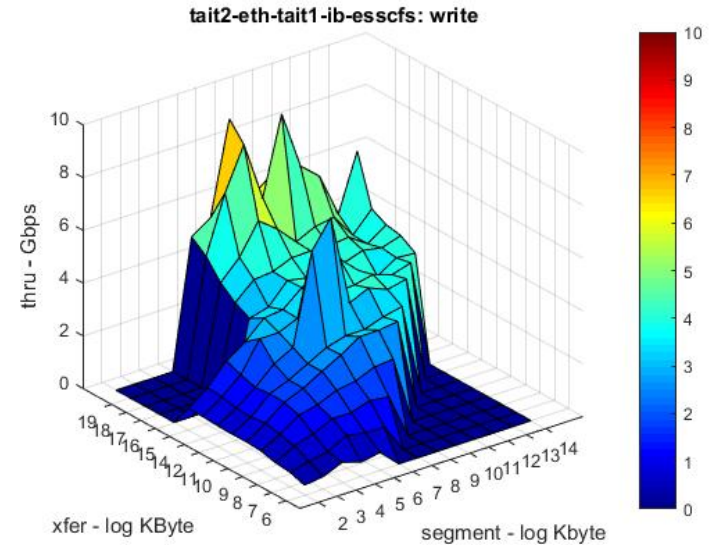
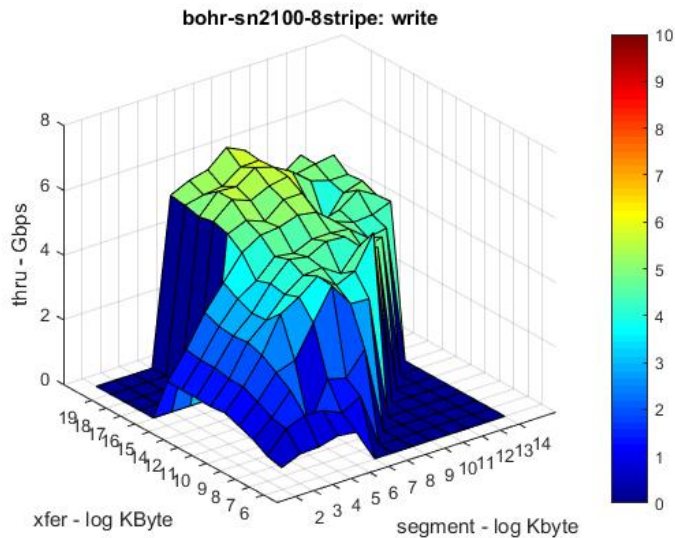
LNet routers: limited impact



# Site Lustre with LNet over IB-Ethernet

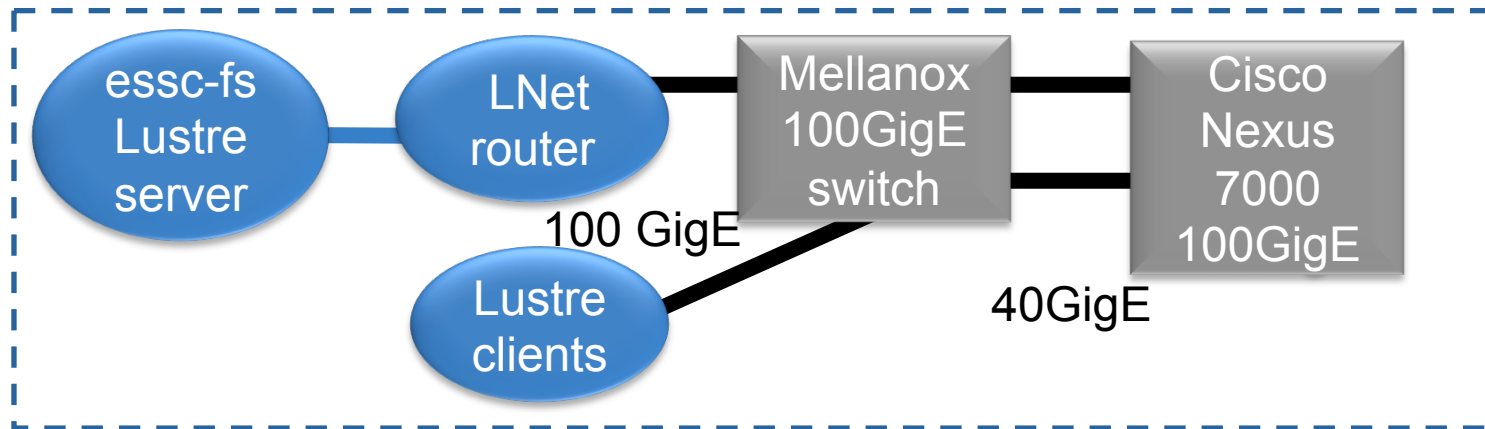
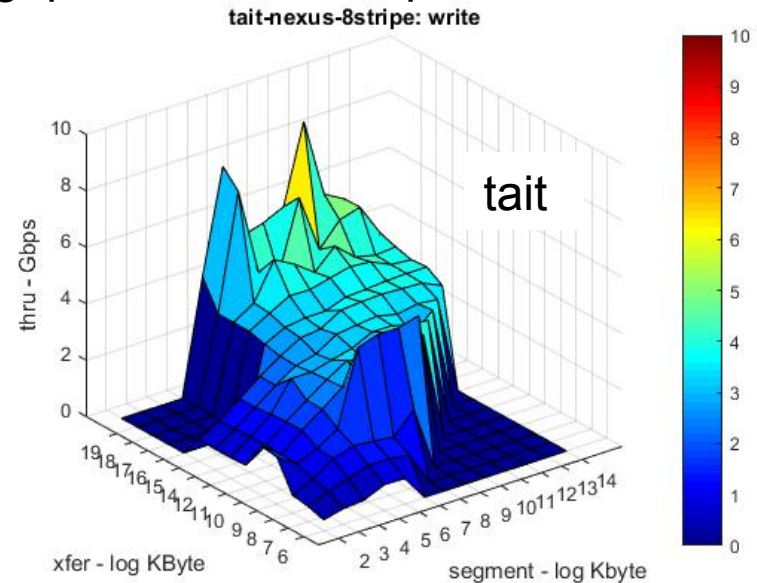
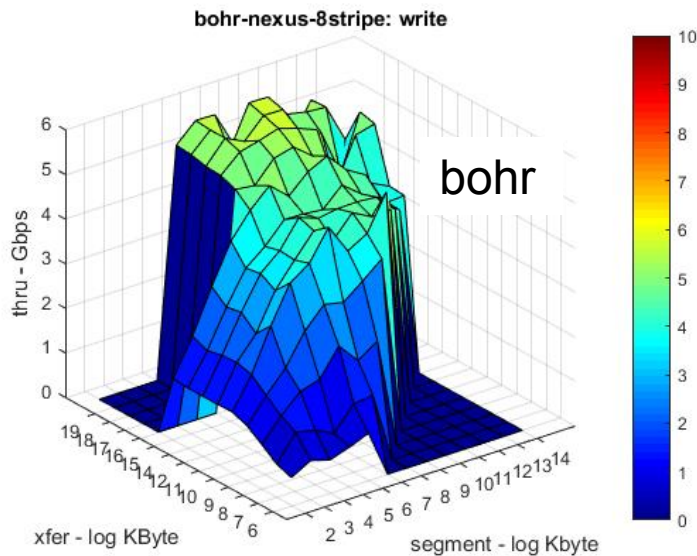
bohr and tait servers (6.8): write peak throughput: ~6 and 4 Gbps

LNet router and Ethernet switch: limited impact



# Site Lustre with LNet over IB-Ethernet

bohr and tait servers (6.8): write peak throughput: ~6 and 4 Gbps



Testbed configuration: complex with several “firsts” – IB, LNet, Lustre and n/w



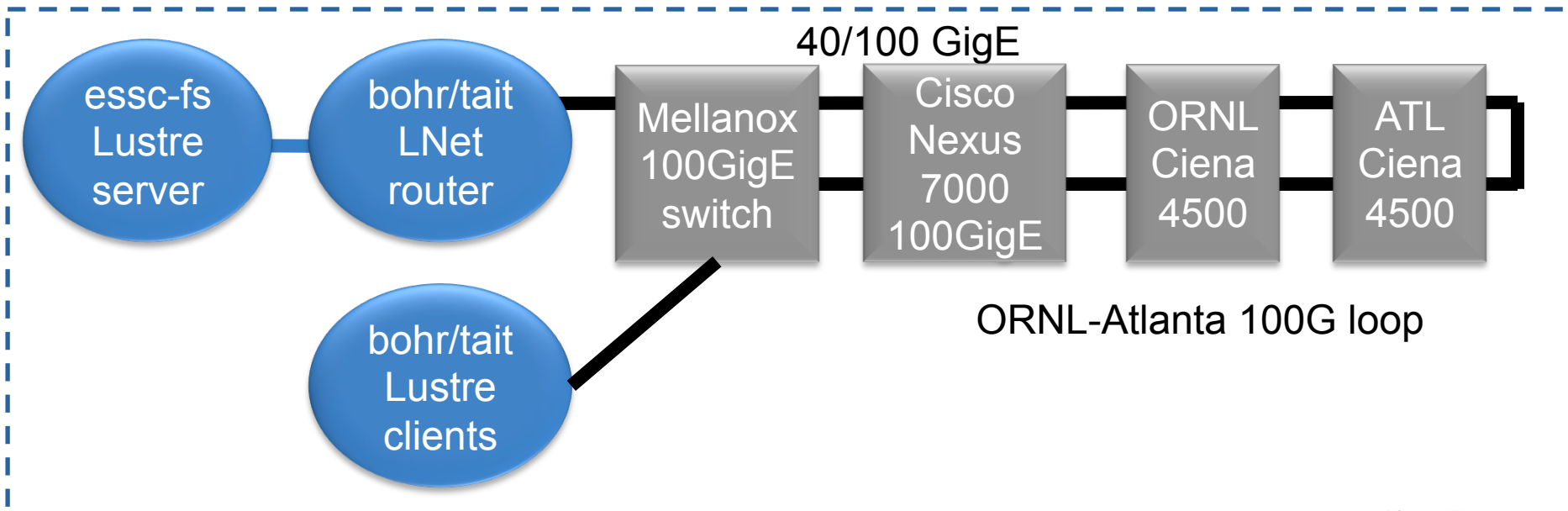
# Wide-Area Lustre: ORNL-Atlanta: 11.6ms



ORNL-Atlanta 100G loop  
11.6ms ~500miles

— 40G Ethernet

— IB QDR



# Wide-Area Lustre: ORNL-Atlanta: 11.6ms

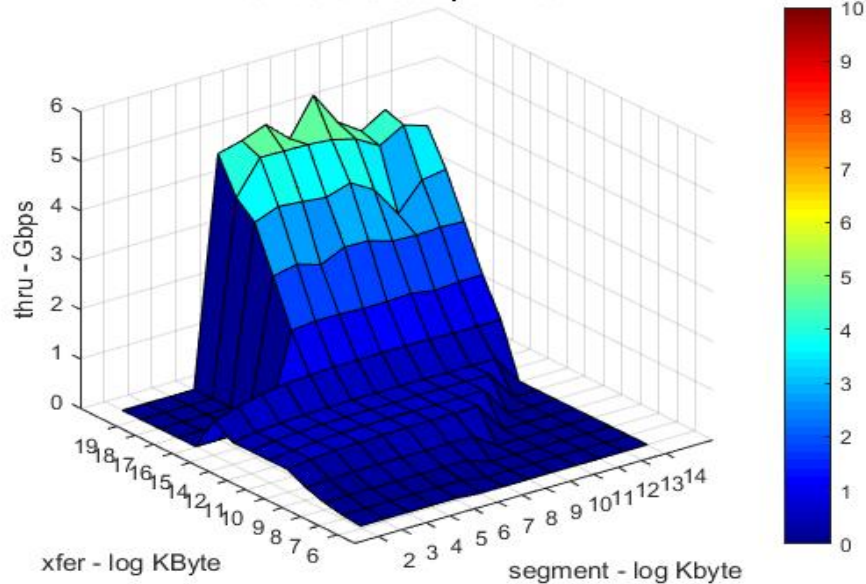
bohr IO servers: write peak throughput:

- wide-area: ~5Gbps
- local: ~6Gbps

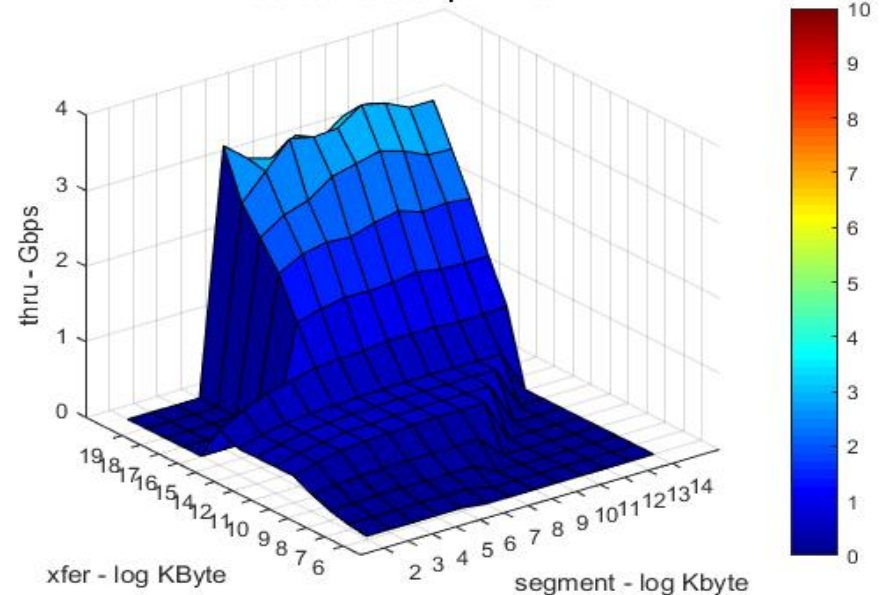
tait compute servers: write peak throughput:

- wide-area: ~3Gbps
- local: ~5Gbps

bohr-atlanta-8stripe: write



tait-atlanta-8stripe: write



Need deeper examination of networking: buffers and congestion control

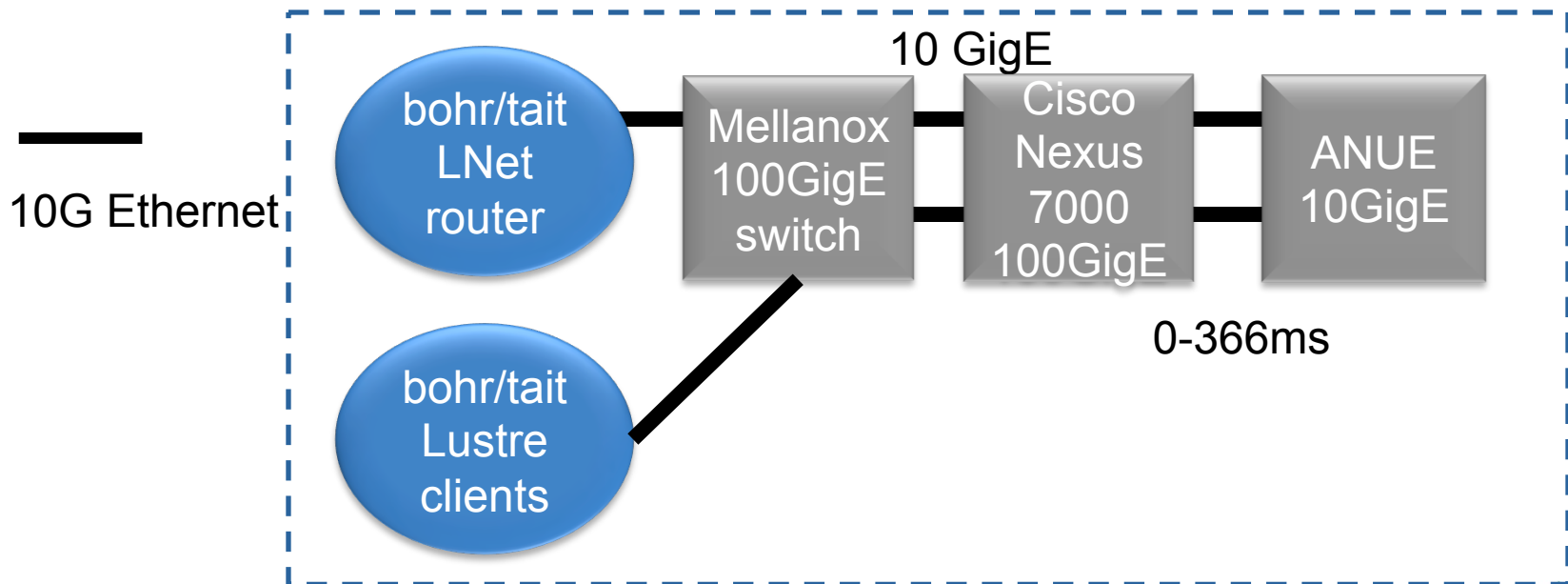
This default throughput is for centos 6.8

Default is lower ~10% when upgraded to centos 7.2

# Network Emulation: 0-366 ms

ANUE 10GigE hardware emulator:

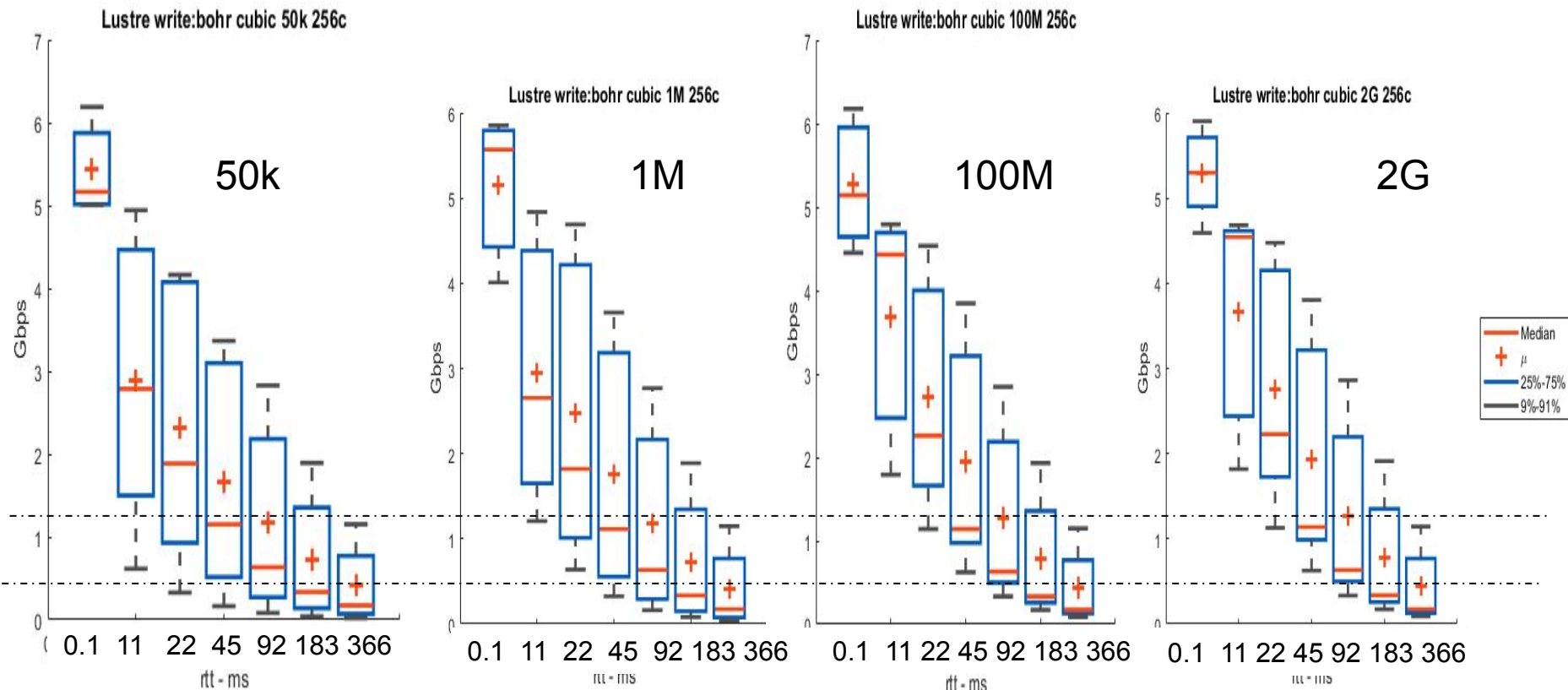
- sufficient since local throughput is below 10Gbps



# Lustre wide-area: bohr – cubic – LNet

Overall Summary: TCP and Lustre tuned

Increasing Lnet router buffers: 50k-2G: improves throughput within 1 %



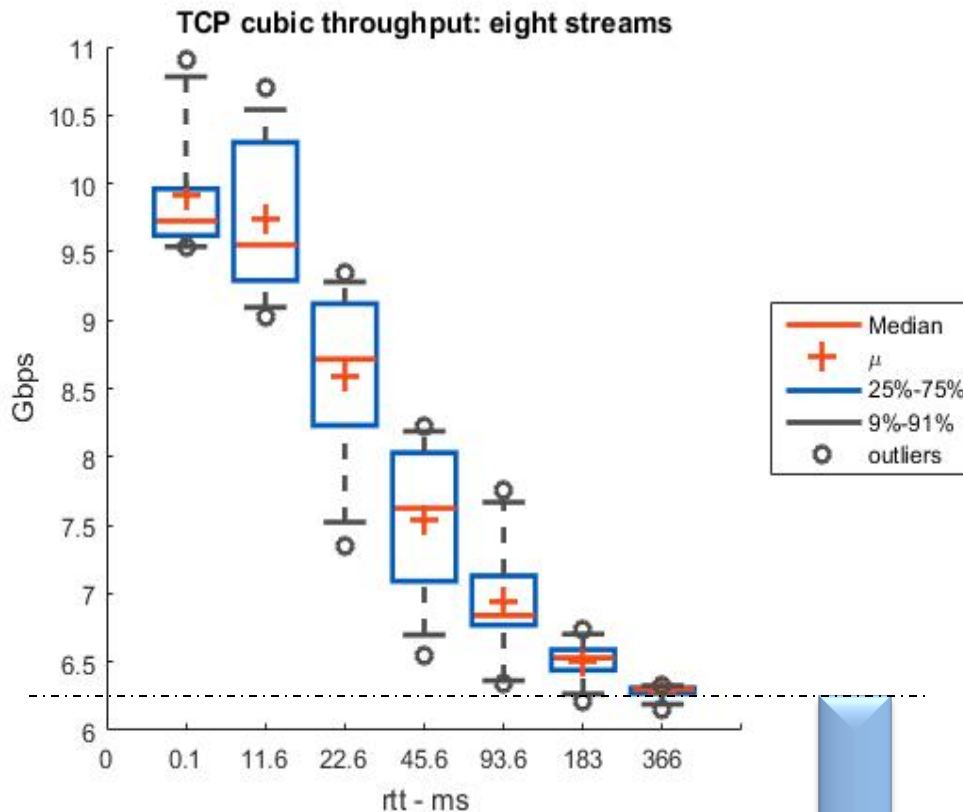
bohr IO servers – TCP tuned

- 48core, opteron, 2.2 GHz, Centos 7.2
- write peak throughput: ~6Gbps

# Network Throughput - CUBIC

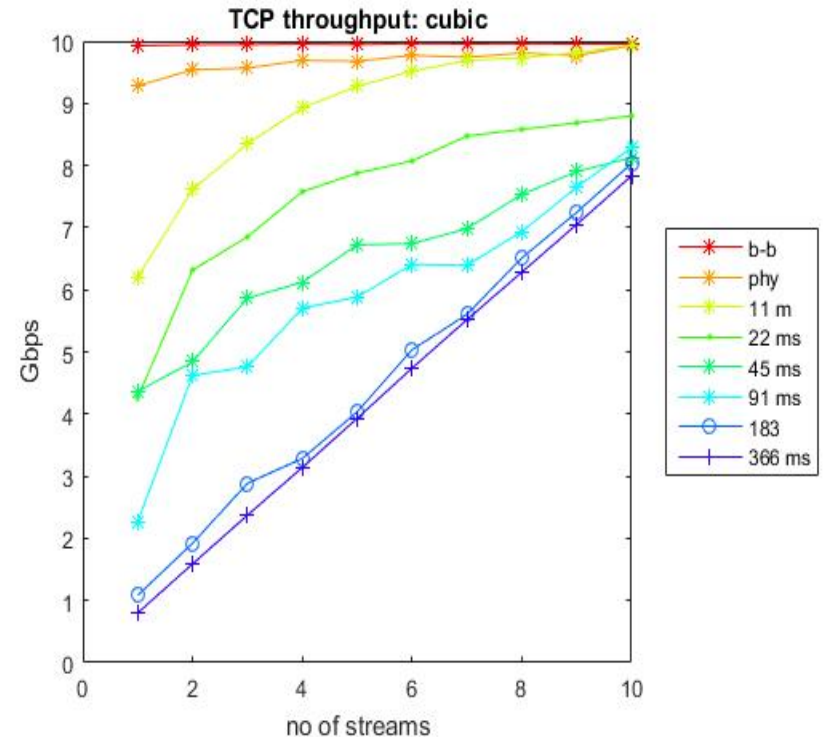
bohr IO servers – TCP tuned

- 48core, opteron, 2.2 GHz
- write peak throughput: ~6Gbps



country

globe



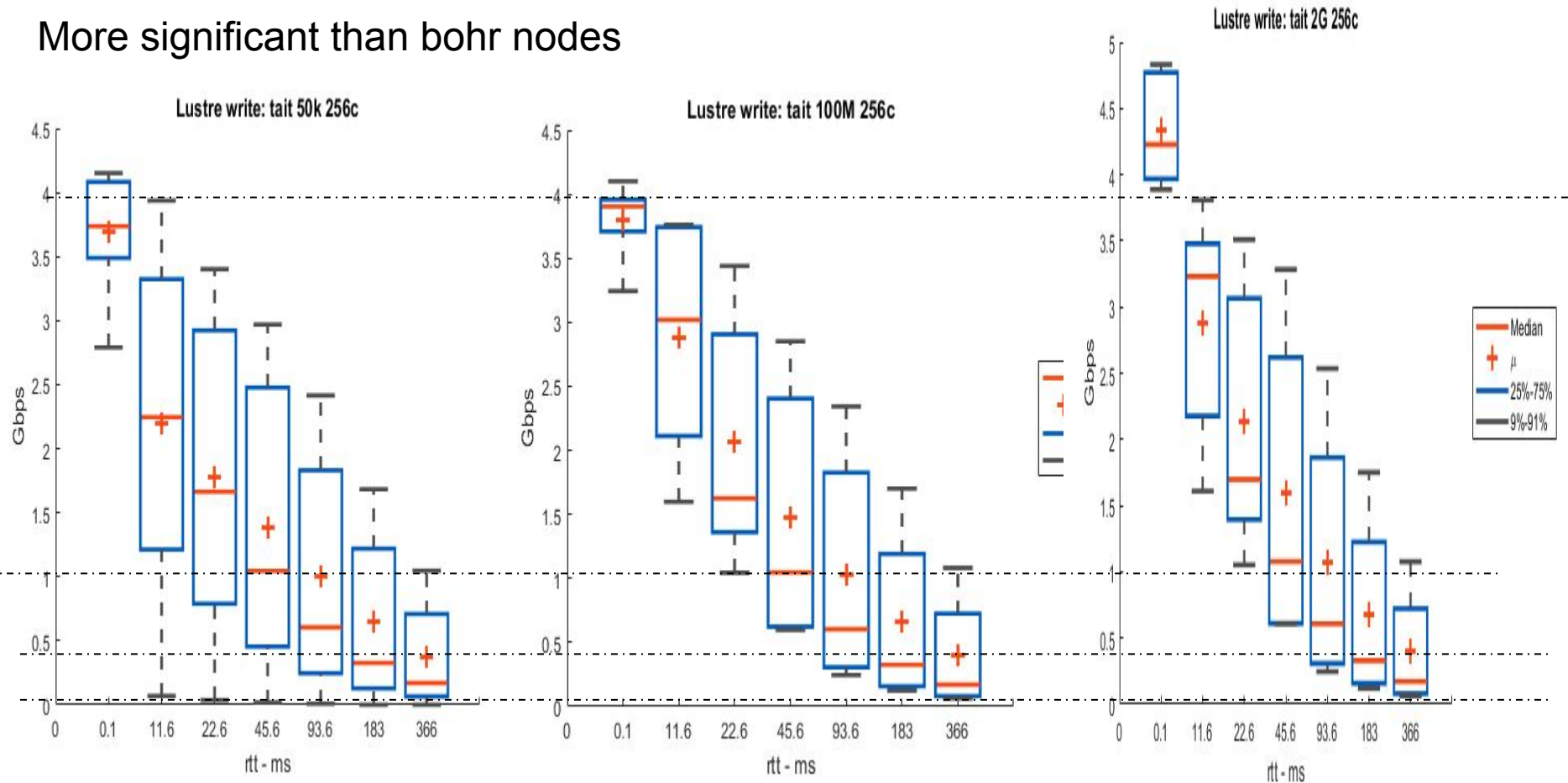
Lustre throughput peak ~6Gbps

# Lustre wide-area: tait – cubic – LNet

Overall Summary: TCP and Lustre tuned

Increasing Lnet router buffers: improves throughput 10% with 2G buffer

More significant than bohr nodes



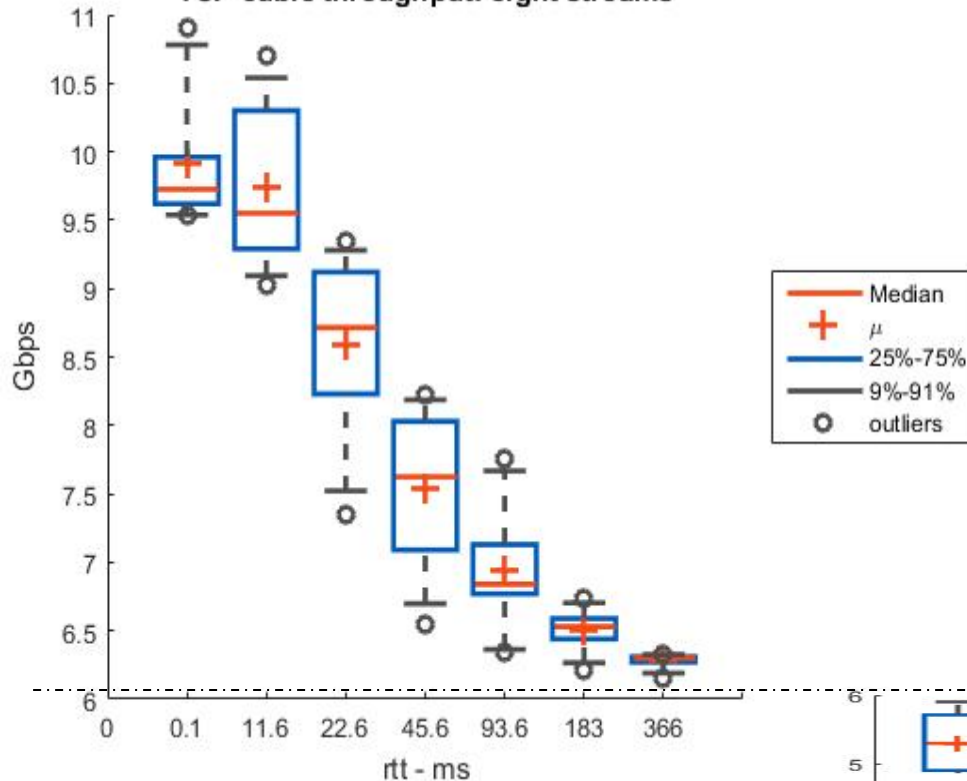
taic compute-cluster servers – centos 7.2

- 24 core, xeon 2.6GHz



# Lustre wide-area: bohr - cubic

TCP cubic throughput: eight streams



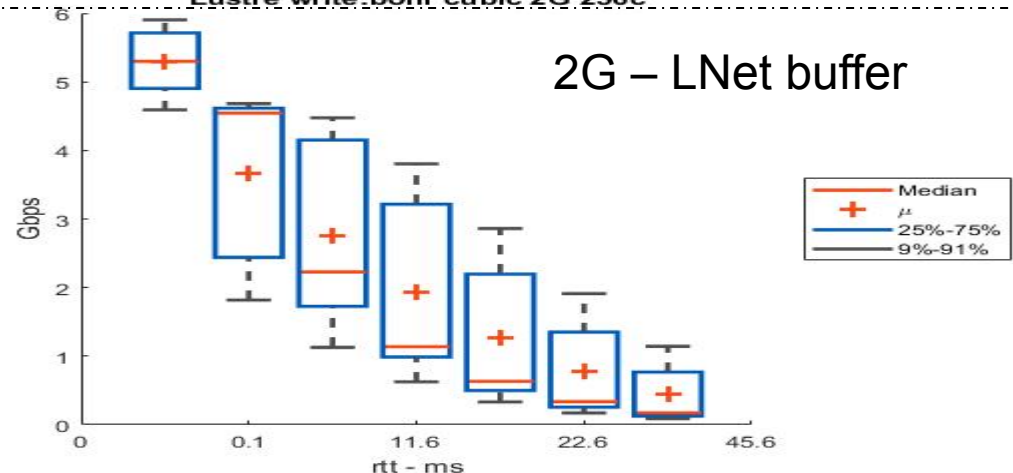
bohr IO servers – centos 7.2

- 48core, opteron, 2.2 GHz
- write peak throughput: ~6Gbps

Lnet router 2G buffers:

Lustre throughput is much below TCP memory throughput

Lustre write:bohr cubic 2G 256c



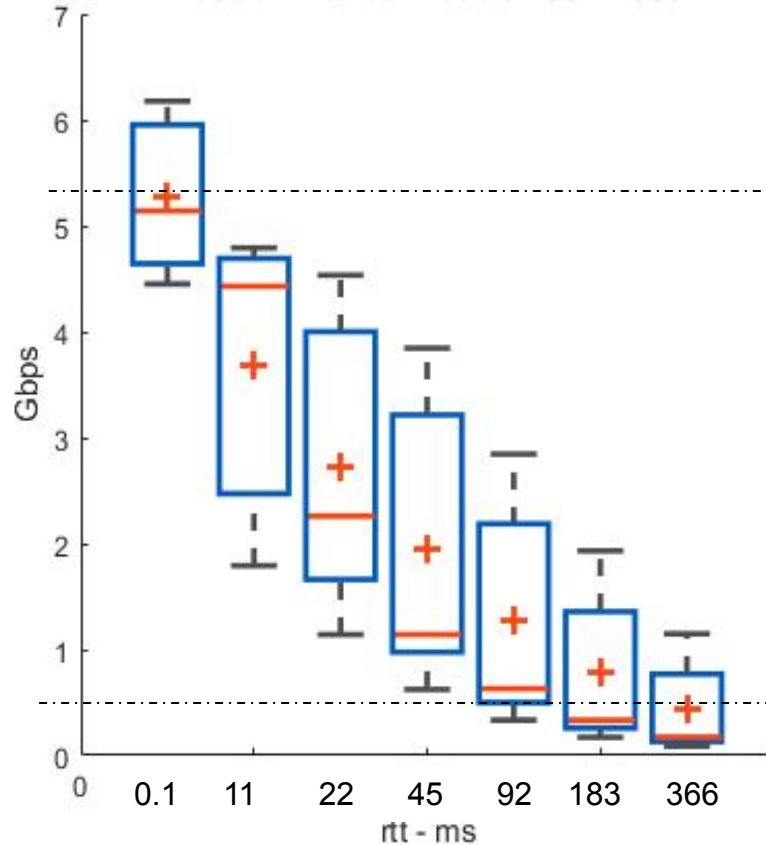
# Lustre wide-area: bohr – cubic and htcp

Lustre throughput: cubic and htcp almost same

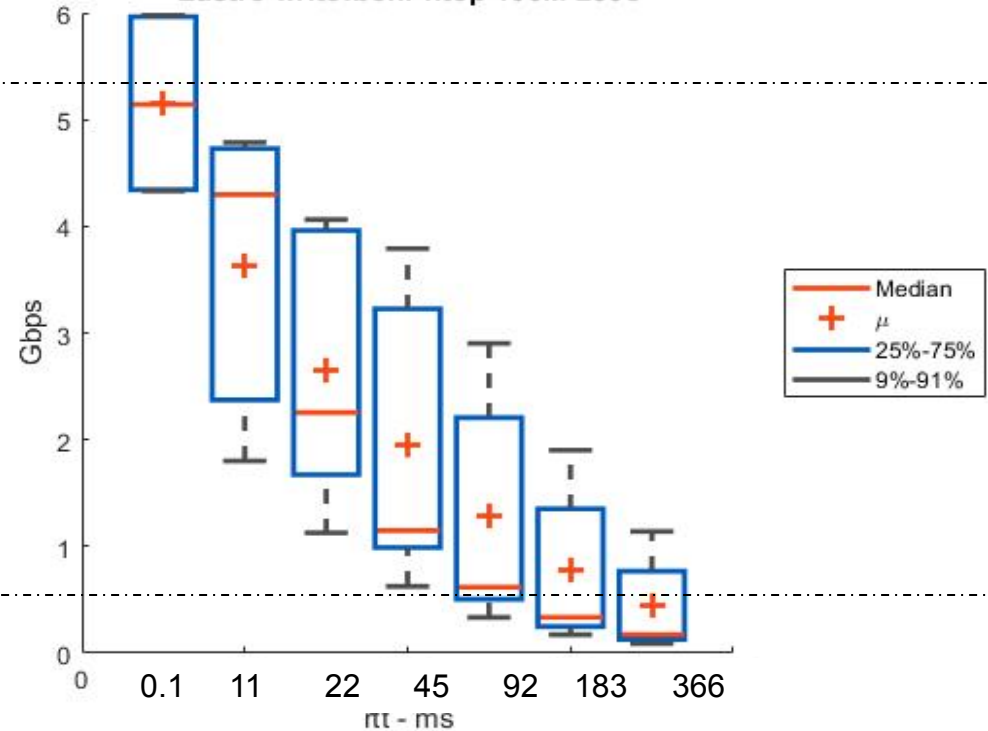
bohr IO servers – centos 7.2

- 48core, opteron, 2.2 GHz
- write peak throughput: ~6Gbps

Lustre write:bohr cubic 100M 256c

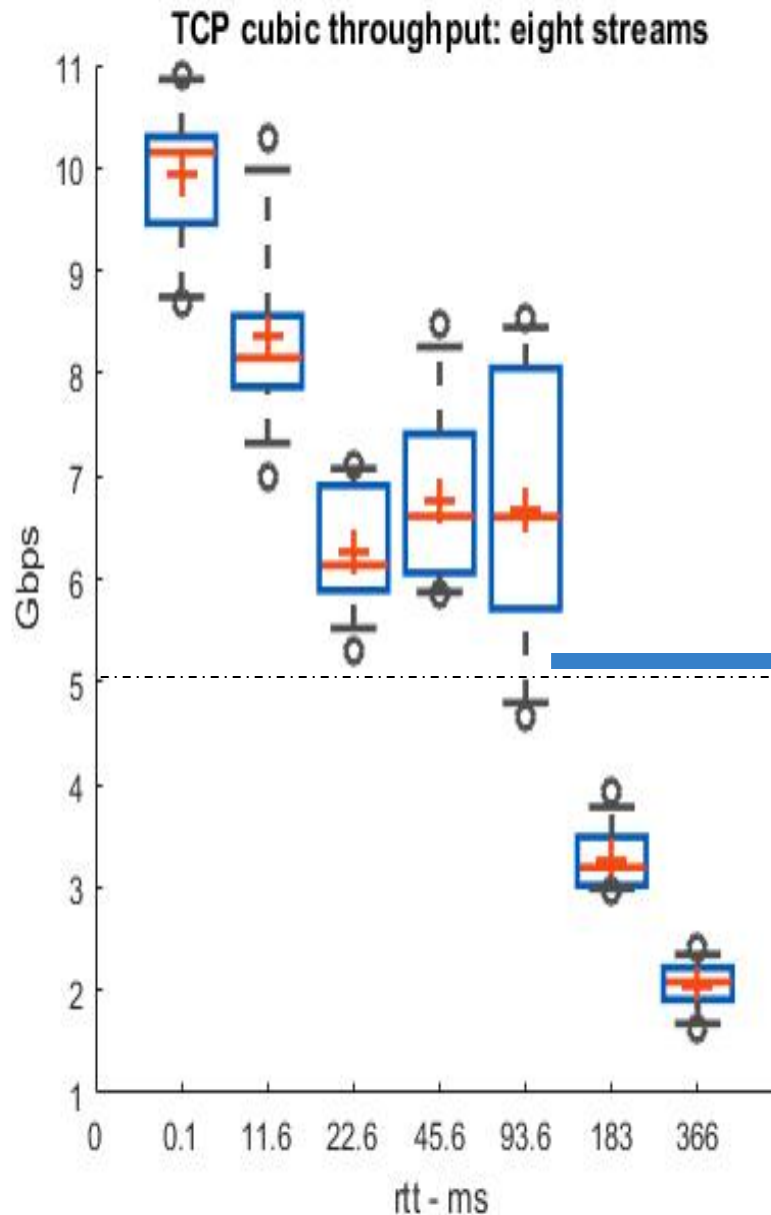


Lustre write:bohr htcp 100M 256c



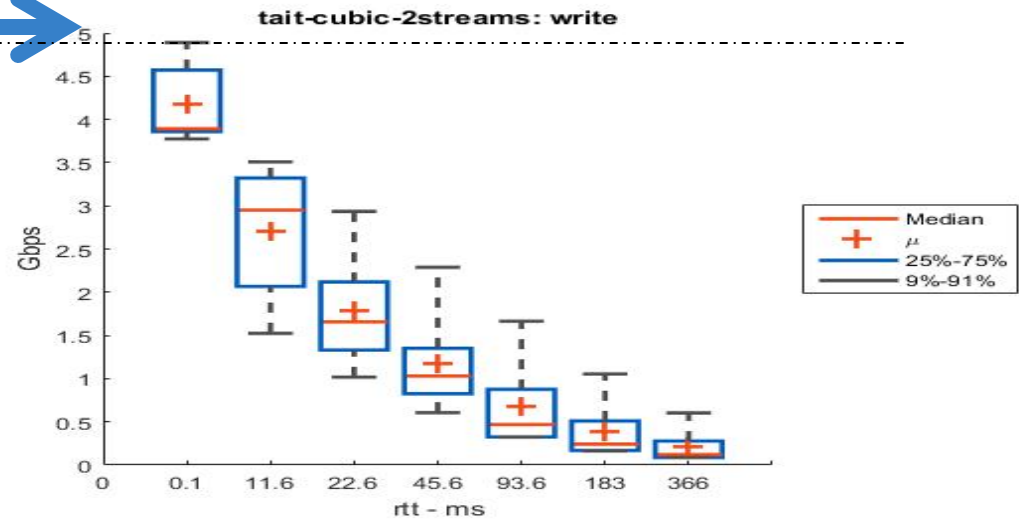


# Lustre wide-area: tait - cubic

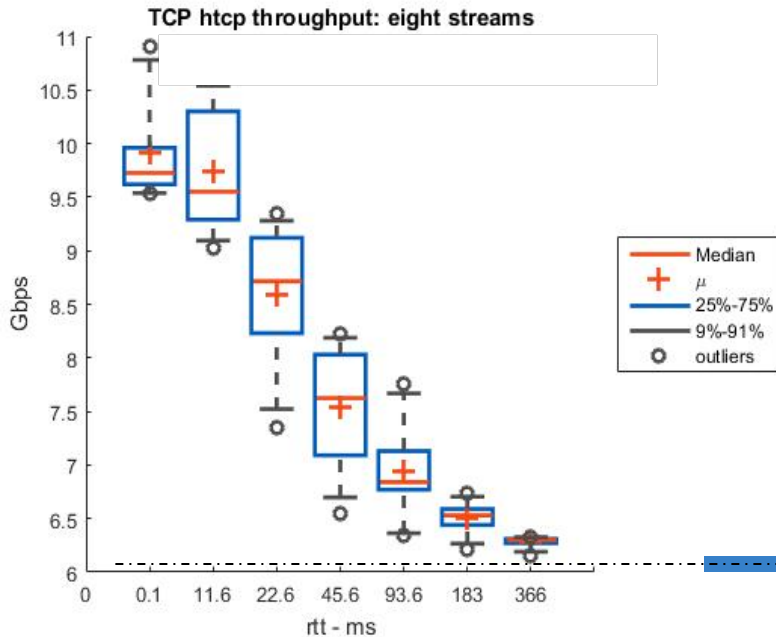


taic compute servers – centos 6.8

- 24 core, xeon 2.6GHz
- write peak throughput: ~5Gbps
- lower than corresponding iperf throughput



# Lustre wide-area: bohr – Hamilton TCP



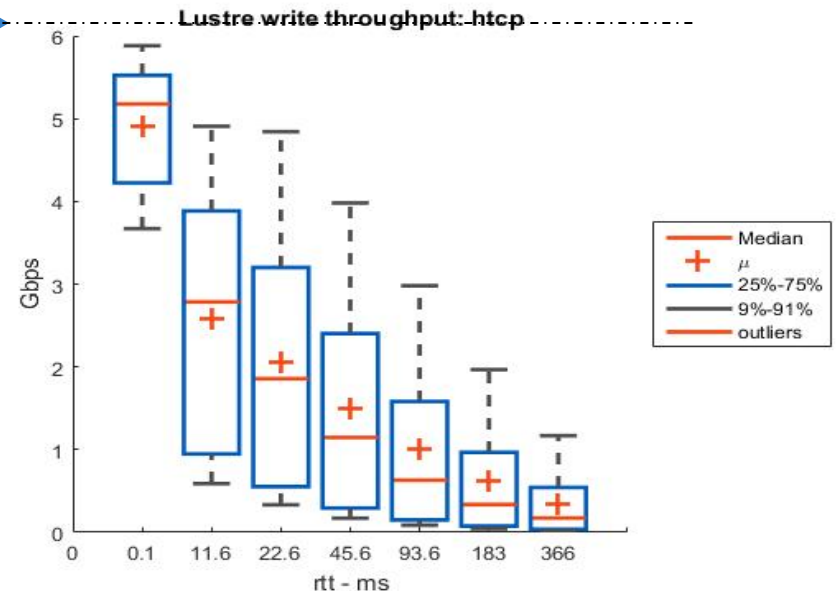
bohr IO servers – TCP tuned

- 48core, opteron, 2.2 GHz
- write peak throughput: ~6Gbps
- lower than lowest iperf throughput
- Centos 6.8

Hamilton TCP: Not much difference

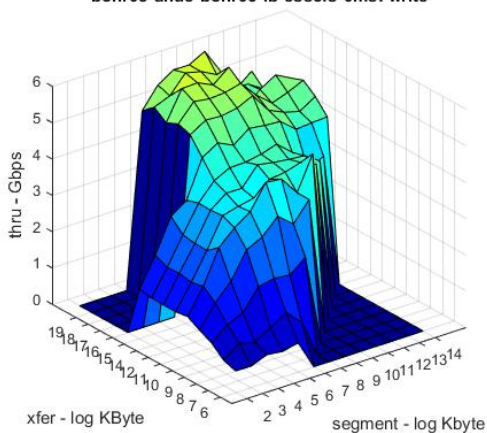
Recommended for large transfers over long (cross-country and inter-continental) distances

- Used in DOE Data Transfer Nodes (DTNs)
- CUBIC is default in Linux

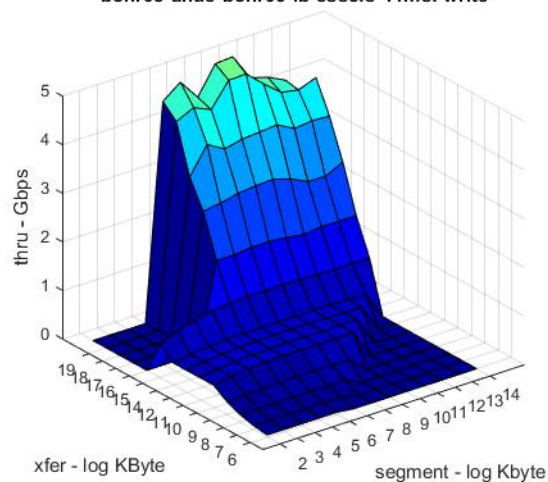


# Lustre wide-area: bohr - htcp

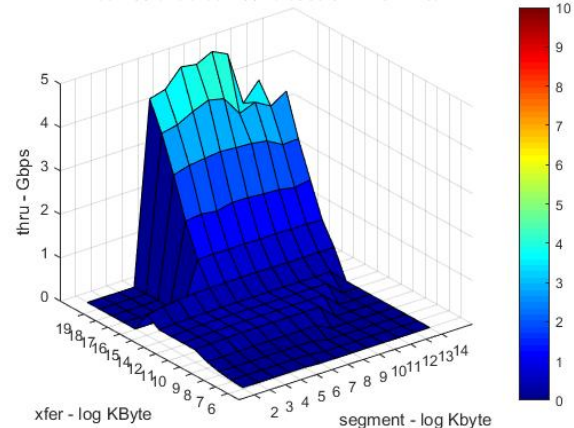
bohr05-anue-bohr06-ib-esscfs-0ms: write



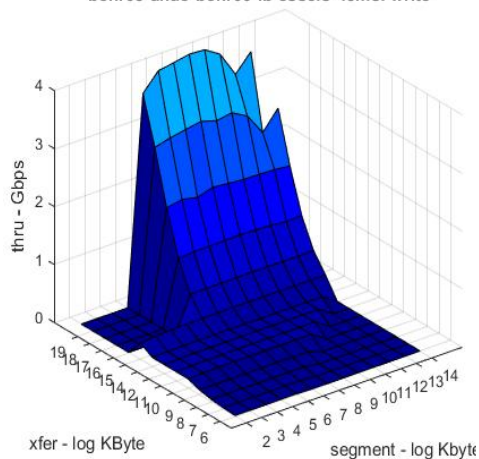
bohr05-anue-bohr06-ib-esscfs-11ms: write



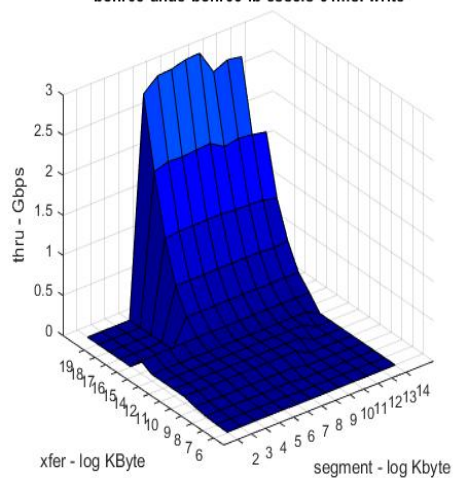
bohr05-anue-bohr06-ib-esscfs-22ms: write



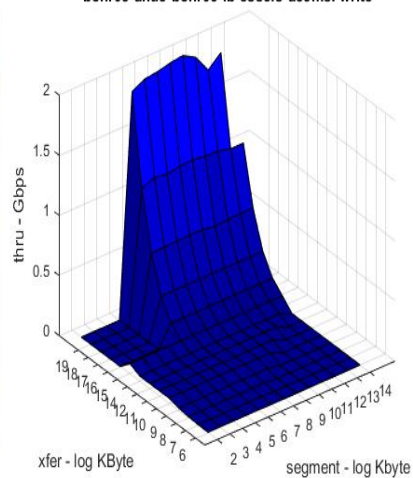
bohr05-anue-bohr06-ib-esscfs-45ms: write



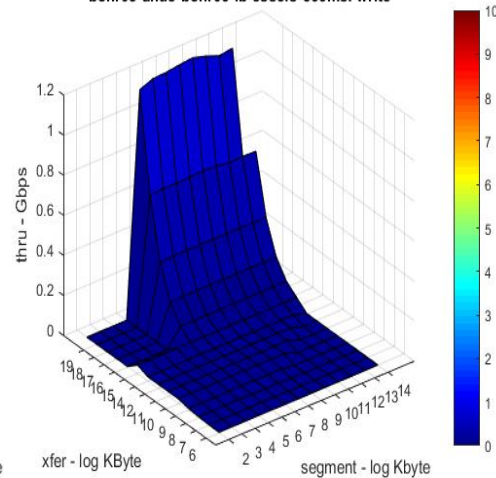
bohr05-anue-bohr06-ib-esscfs-91ms: write



bohr05-anue-bohr06-ib-esscfs-a83ms: write



bohr05-anue-bohr06-ib-esscfs-366ms: write

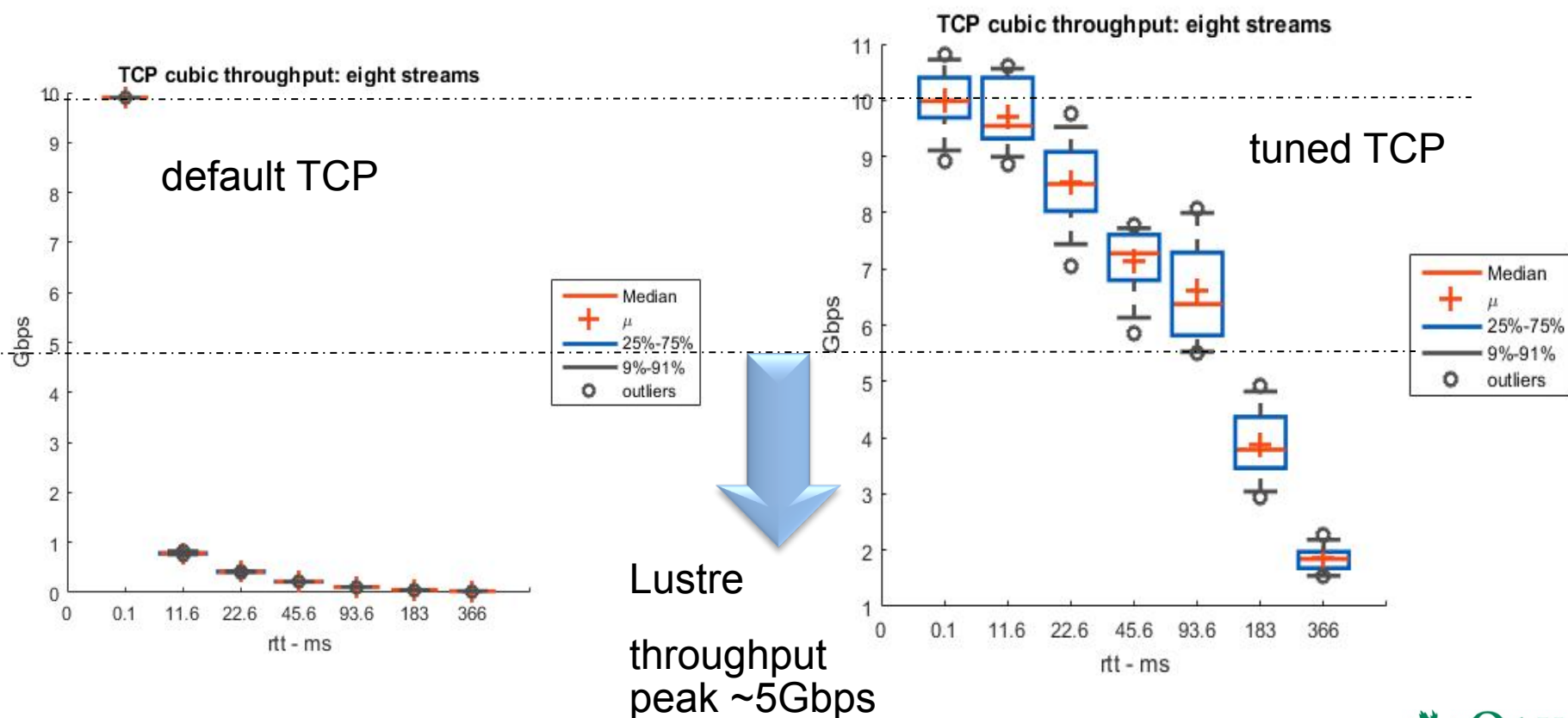


# Network Throughput: Cluster Nodes

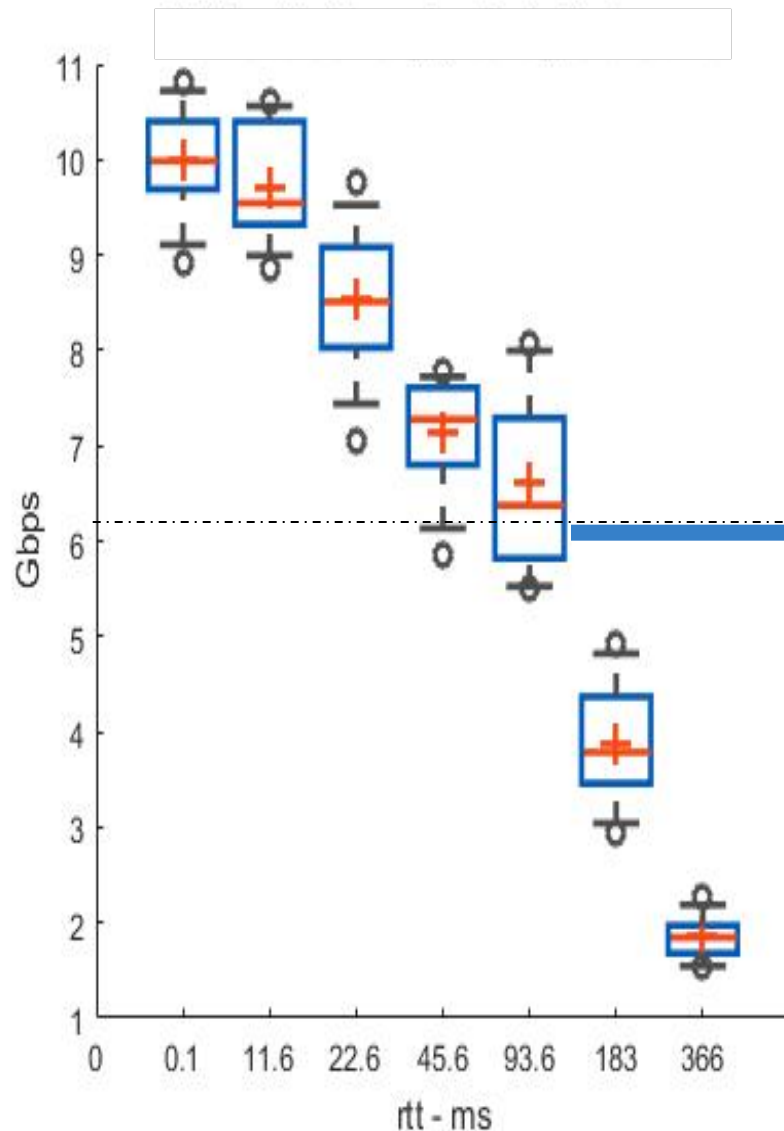
tail compute servers – centos 6.8

- 24 core, xeon 2.6GHz
- write peak throughput: ~5Gbps

Compute nodes are not tuned well for wide-area data transfers

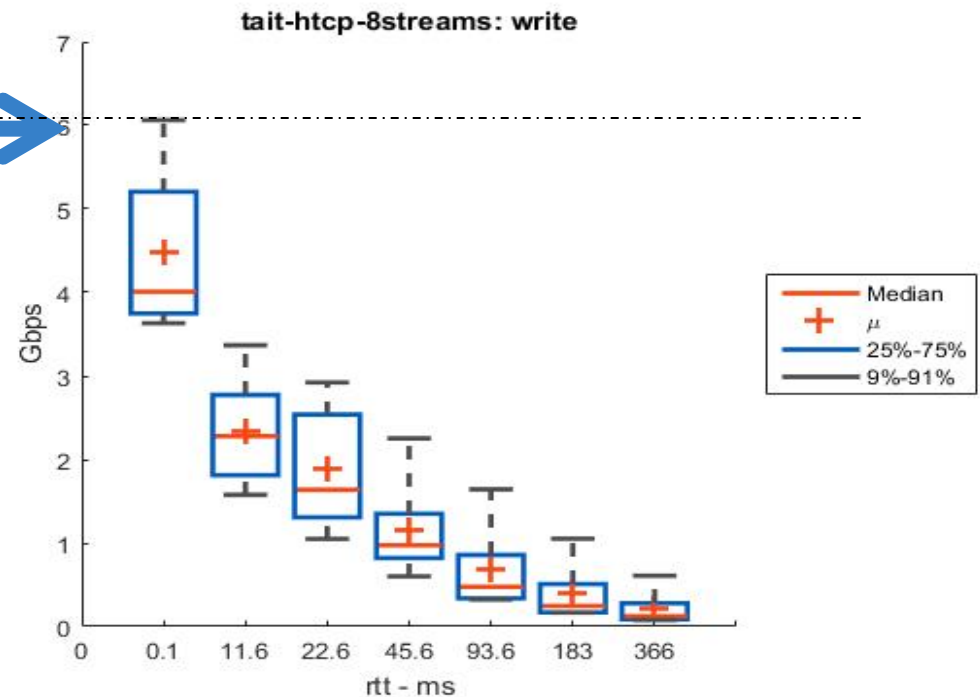


# Lustre wide-area: tait - htcp



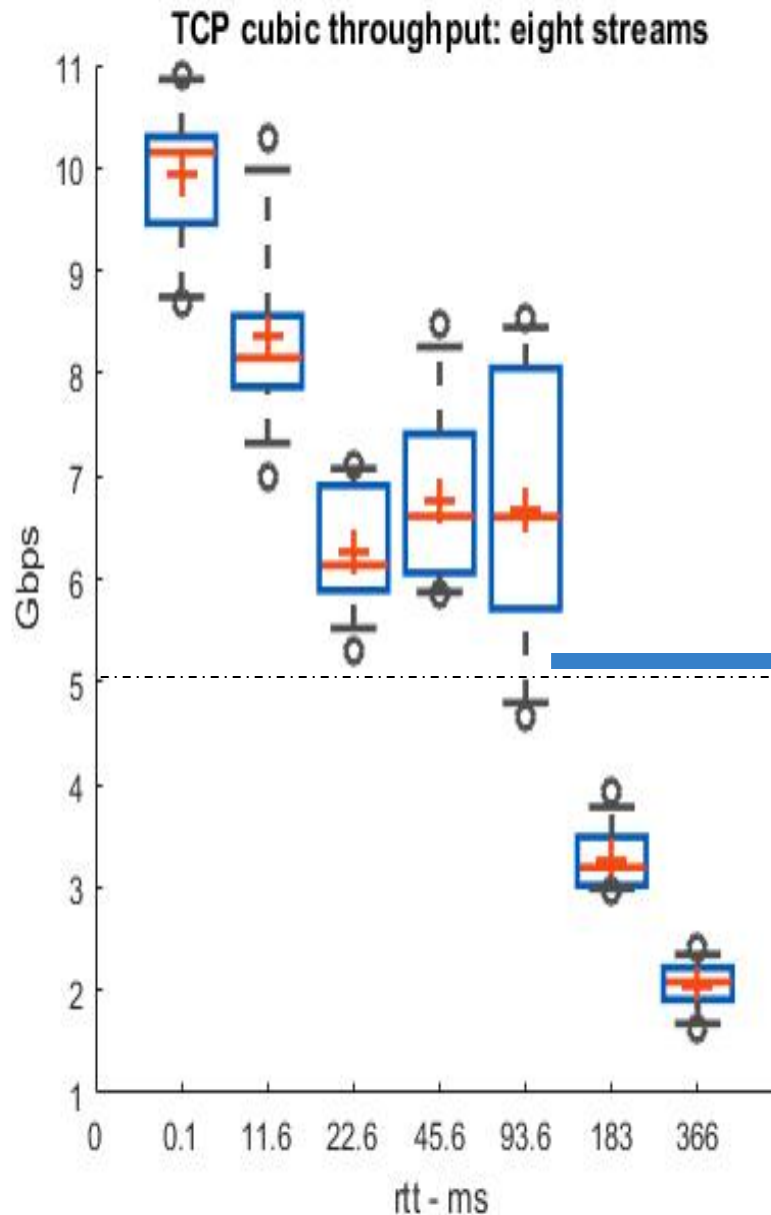
tait compute servers - centos 6.8

- 24 core, xeon 2.6GHz
- write peak throughput: ~6Gbps
- lower than corresponding iperf throughput



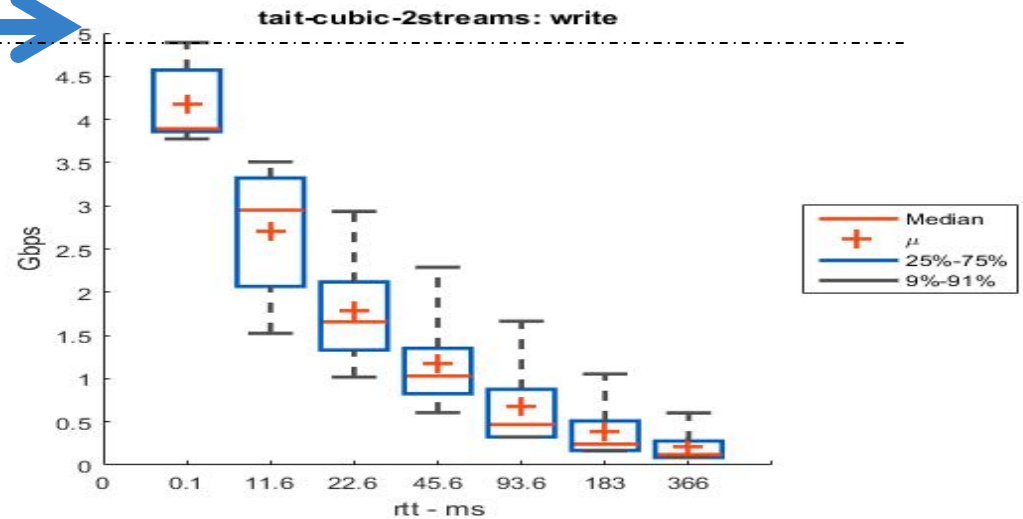


# Lustre wide-area: tait - cubic



taic compute servers –centos 6.8

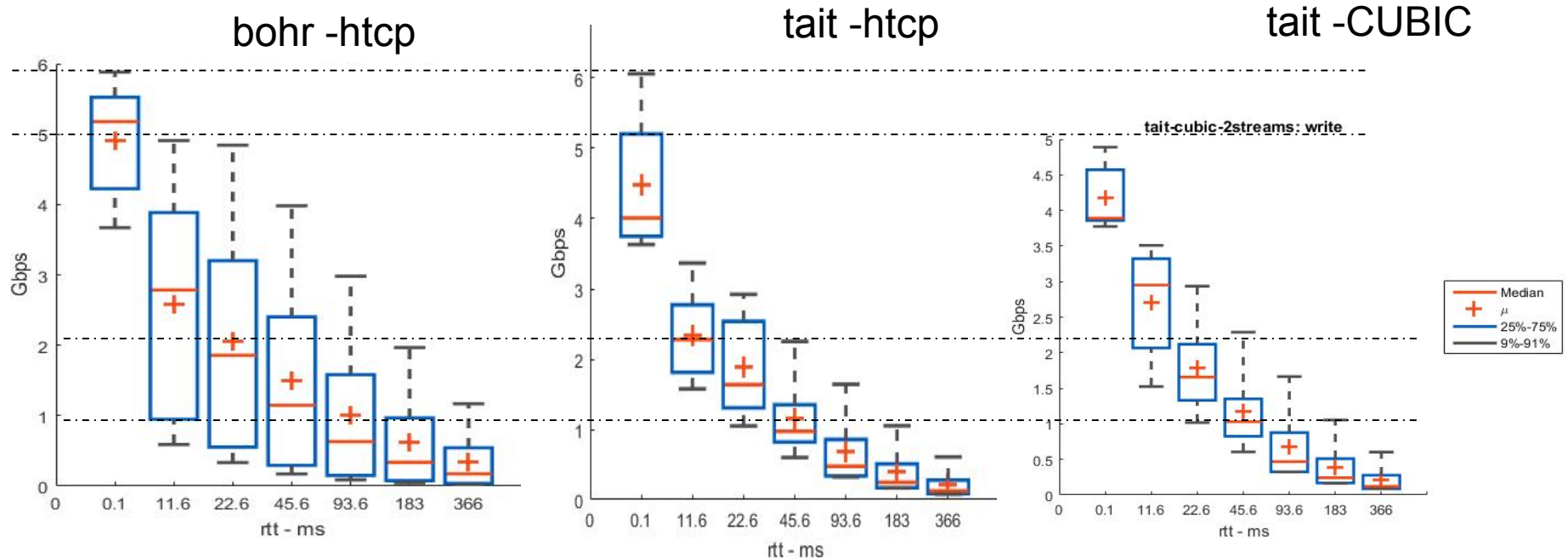
- 24 core, xeon 2.6GHz
- write peak throughput: ~5Gbps
- lower than corresponding iperf throughput



# Lustre wide-area: bohr - tait default LNet

Throughput is somewhat higher for bohr servers – centos 6.8

- their network and local Lustre throughputs are higher

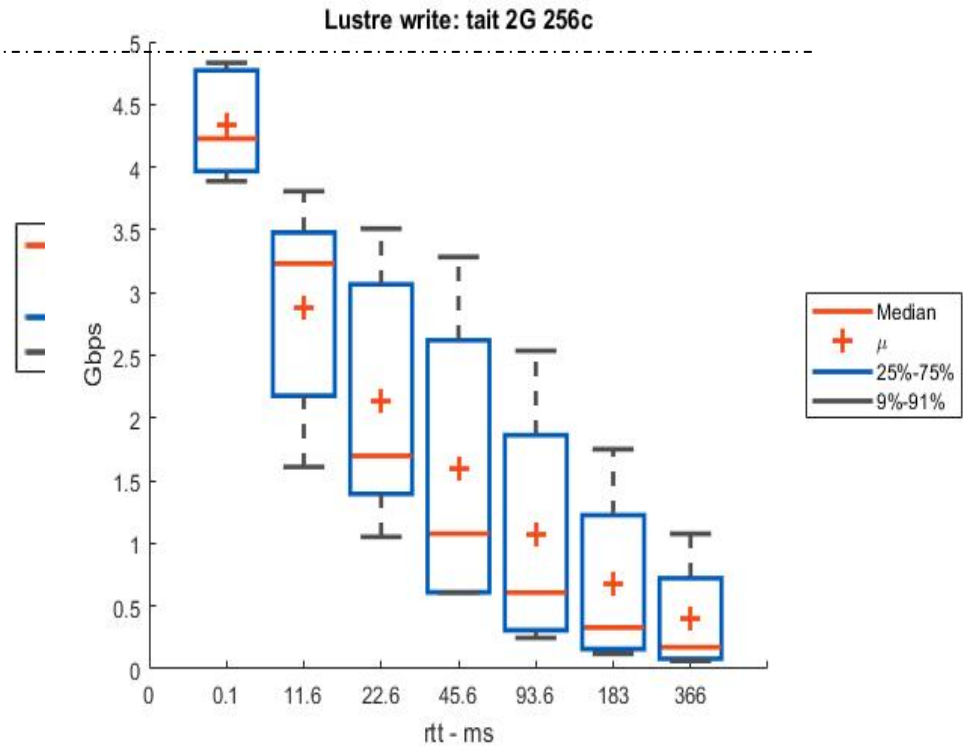
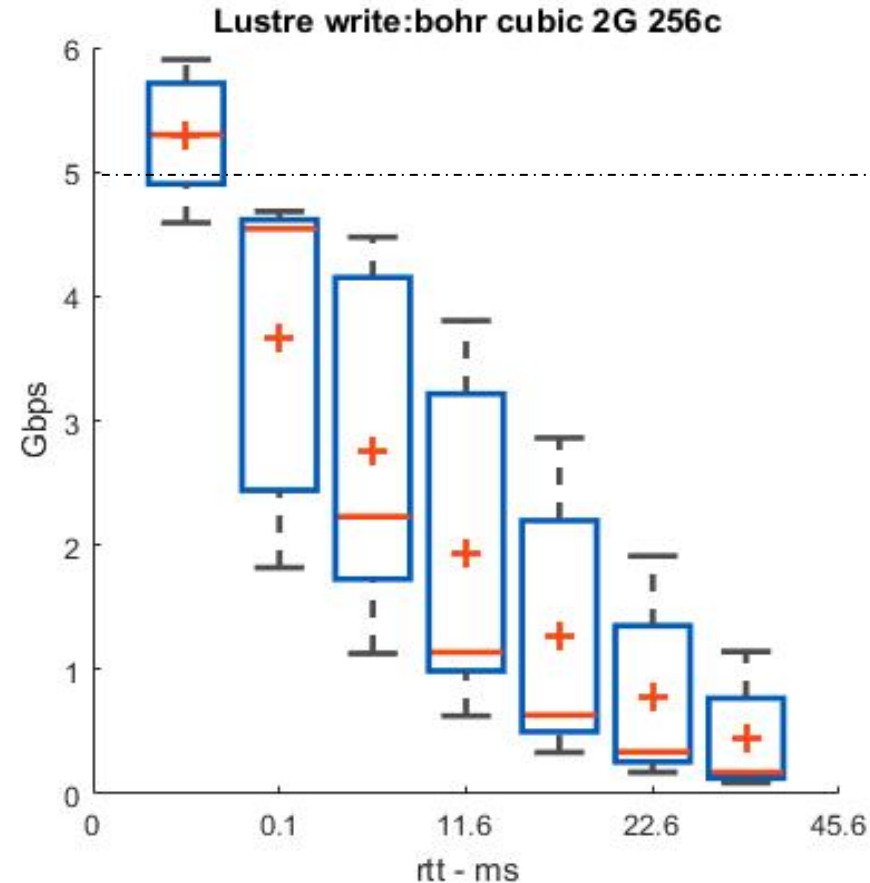


Throughput is similar for Hamilton TCP and CUBIC for tait

- network throughput is higher for Hamilton TCP but does not make much difference for Lustre

# Lustre wide-area: bohr - tait 2G LNet

Throughput is somewhat higher for bohr servers (7.2) – similar to default case





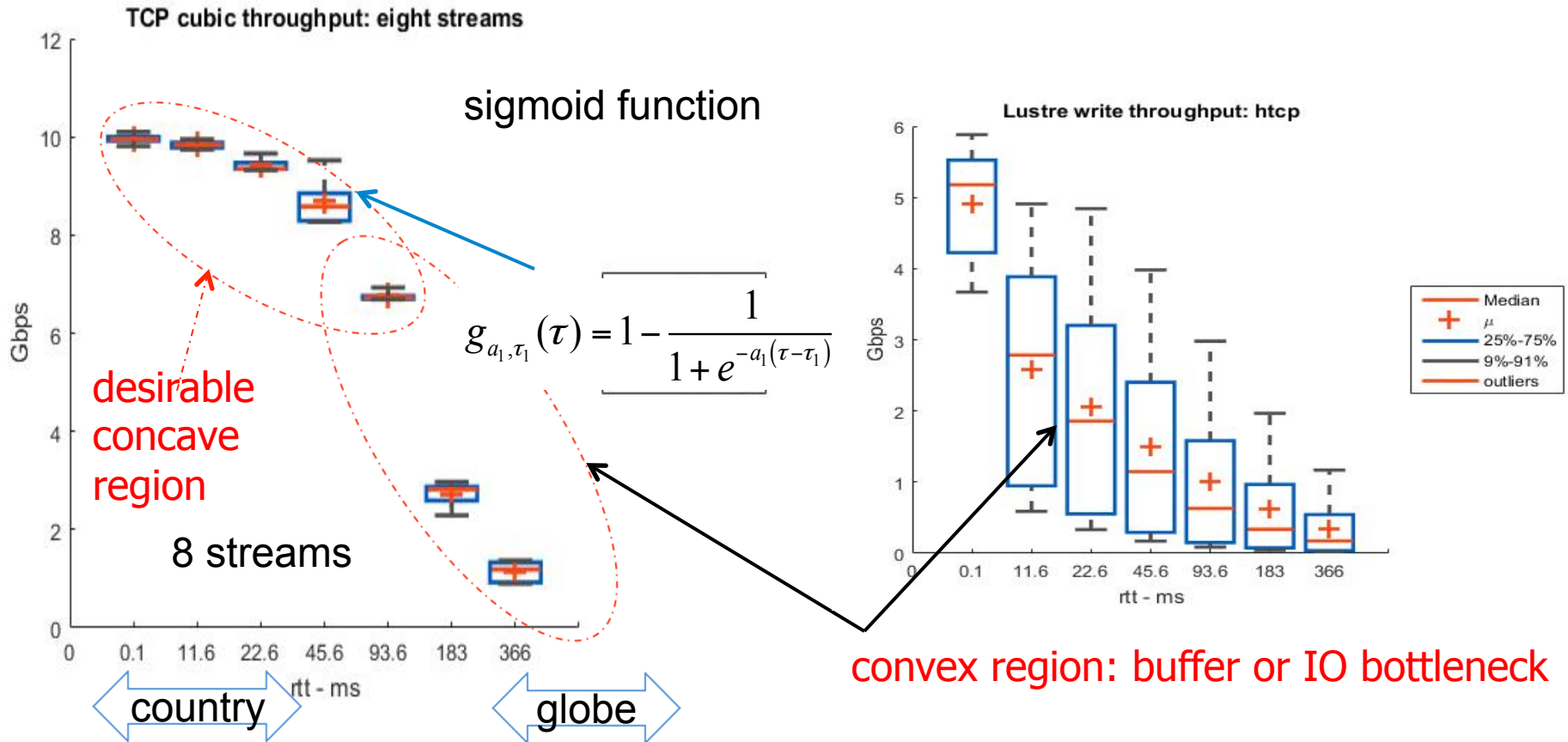
# IO or Network Bottleneck?

TCP memory transfers: concave-convex regions

10Gbps: CUBIC TCP buffers tuned for 200ms rtt

Concave region: indicates buffer, IO bottleneck

Our Lustre configuration indicates IO limit



RTT: cross-country (0-100ms), cross-continentals (100-200ms), across globe (366ms)

# Generic Model for Data, Disk and File Transfers

Buffer size, IO throughput or available processing power limit data in transit:  
connection capacity (bps):  $C$

RTT:  $\tau$

data unacknowledged within a slot of period:  $\tau$

no IO or processor limit:  $C\tau$

under IO or processor limit:  $B < C\tau$

example: limited buffer size

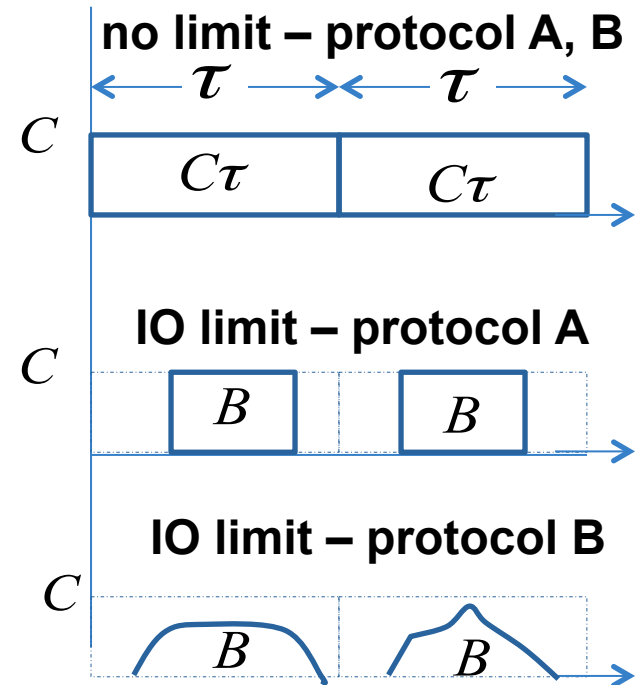
Throughput averaged over each slot of width  $\tau$ :

$$\theta^\tau(t) = \frac{B}{\tau}$$

Throughput profile:  $\Theta_o(\tau) = \frac{1}{T_o} \int_0^{T_o} \theta^\tau(t) dt = \frac{B}{\tau}$

Throughput derivative:  $\underbrace{\frac{d\Theta_o}{d\tau}} = -\frac{B}{\tau^2}$

increasing function of  $\tau$  implies convexity of  $\Theta_o(\tau)$



Transport methods may have different shapes of  $B$  – but subject to convexity

- convex profile indicates disk or file throughput limit
- due to peer credits on IB and Ethernet sides of LNet

# Conclusions

## Summary

- Demonstrated Luster mounted over long-haul connections
  - file transfers do not require specialized solutions, e.g. XDD, Aspera, GrifFTP
  - distributed applications with file accesses are supported transparently
- LNet-based solution
  - extends local IB-based Lustre (no changes to file servers)
- Measurements over 0-366ms connection suites
  - provided insights for performance: configurations constrained by IO throughput

## Future Work

- Comprehensive tests and analysis
  - other TCP congestion control methods – Scalable TCP, BBR, ...
- Continued analysis and performance tuning
  - host and Lustre performance parameters and tuning
- Comparative Analysis and Analytics:
  - TCP, ROCE(RDMA over Converged Ethernet) and IB Lustre solutions

# Acknowledgements



This work was supported by the United States Department of Defense (DoD) and also in part by RAMSES project of Department of Energy (DoE), and used resources of the Computational Research and Development Programs at Oak Ridge National Laboratory.



# Questions?



Computational Research &  
Development Programs

