CRAY

Infiniband At A Distance

Dave McMillen and Steve Woods

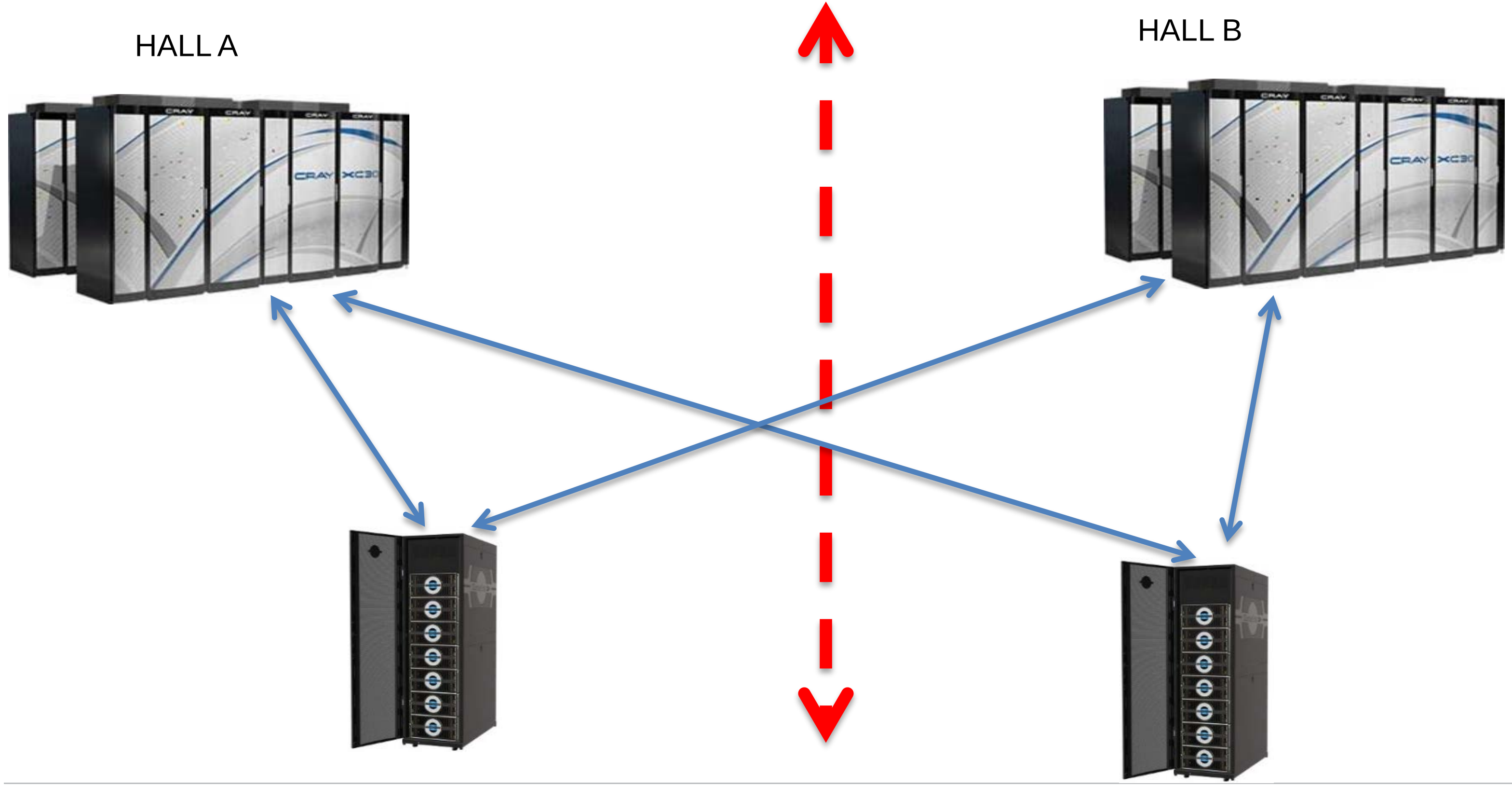
Lustre LNET Focus

- General long distance Infiniband has many potential use cases
- This presentation focuses exclusively on LNET and storage applications
- Mixed use of long distance Infiniband (i.e. compute-compute on same fabric as LNET) raises many non-trivial issues
- Typical Cray designs use LNET routers to isolate the Lustre server Infiniband fabric (SAN-like)

Why Infiniband At A Distance?

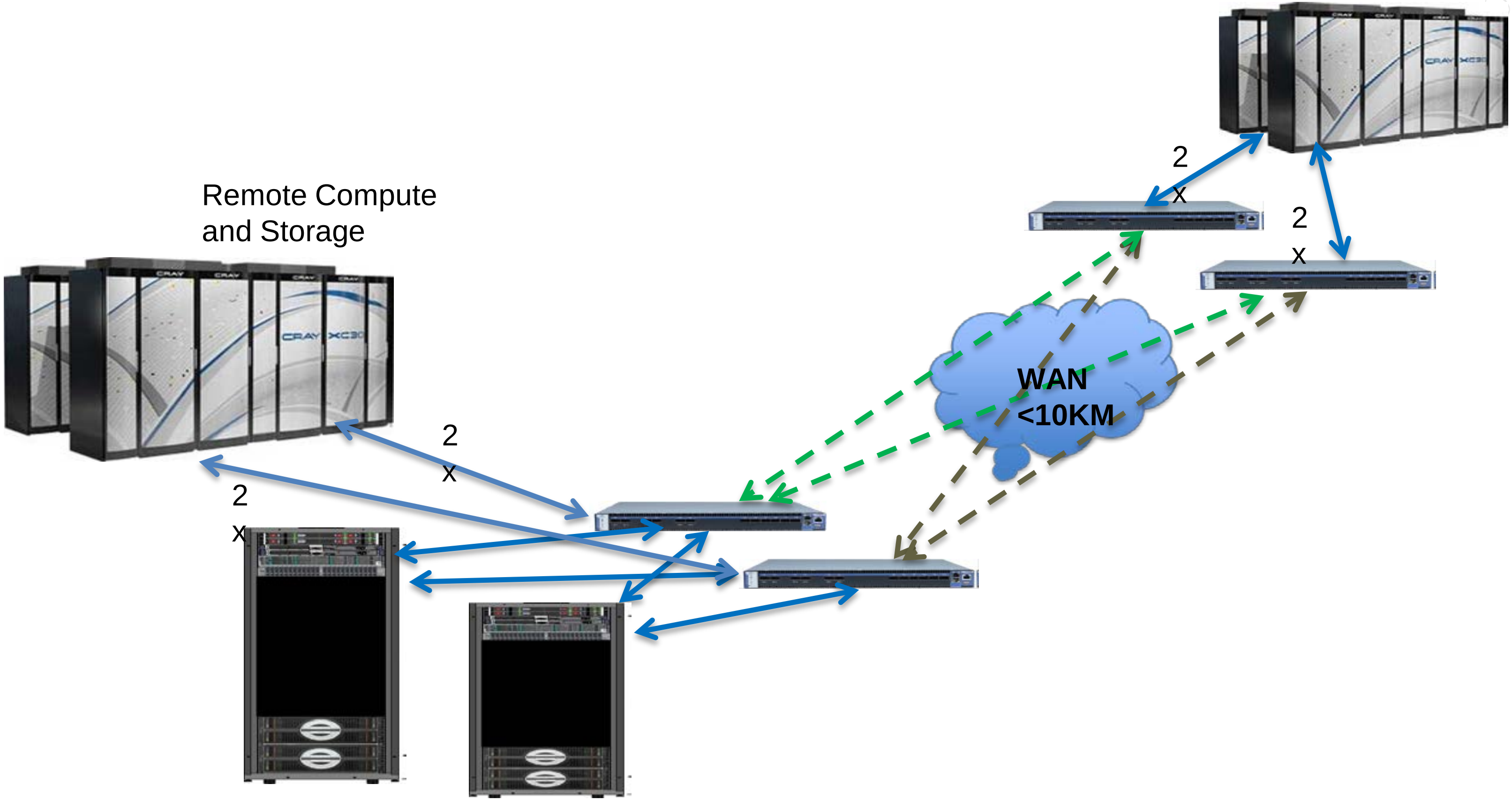
- Physical size of larger installations
- Isolation of selected components
- Desire to cross mount previously independent file systems
- Disaster resilience
- Retain simplicity of connectivity and management
- Avoid complexity of LNET routing, especially multiple hops

An Example Layout: Two separate Halls, computes need access to both Halls' storage

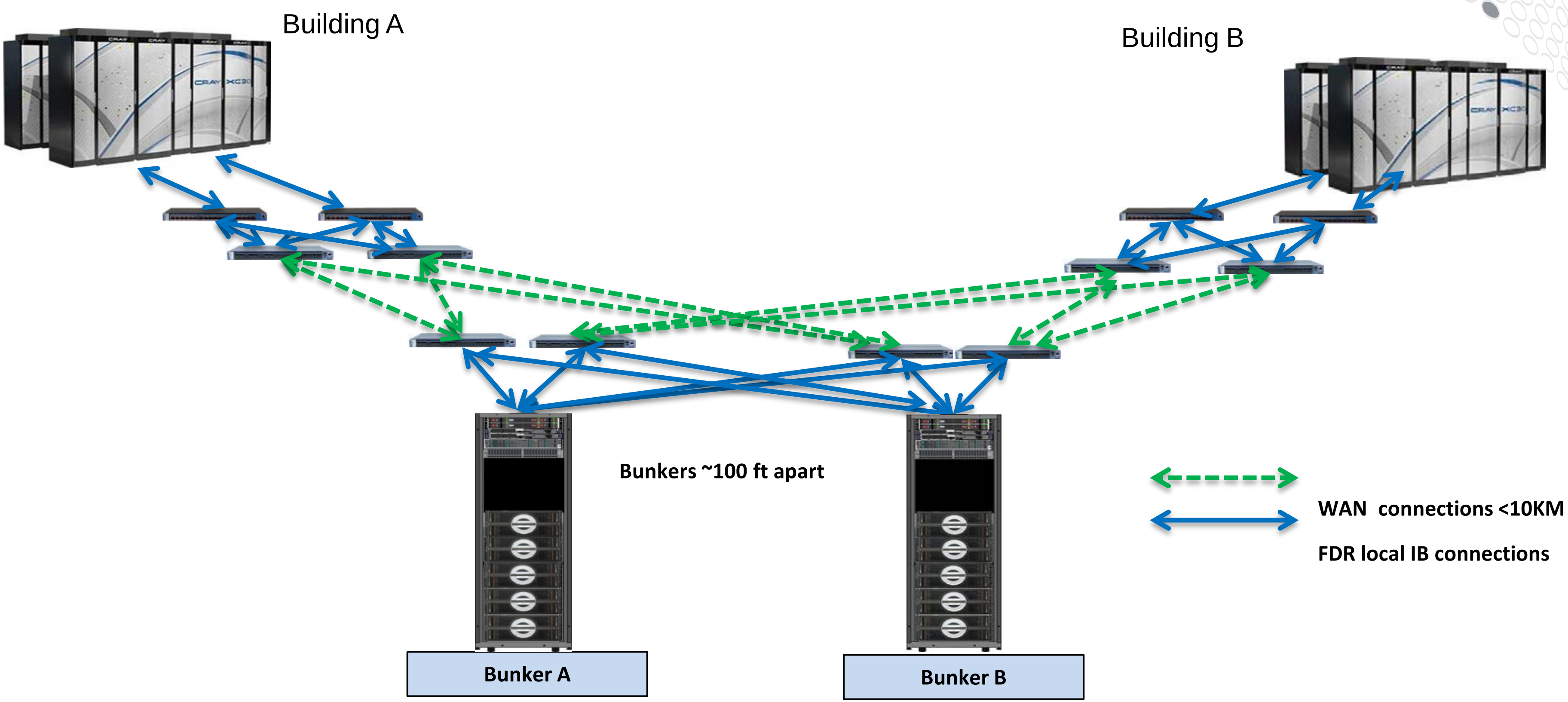


COMPUTE | STORE | ANALYZE

Remote and Local Access to Central Storage



Remote Access to Central Storage



Building A

Building B

Bunkers ~100 ft apart

WAN connections <10KM

FDR local IB connections

Bunker A

Bunker B

Alternatives to Infiniband At A Distance

- Build LNET Infrastructure using Ethernet
- Use IB-Ethernet LNET routers to transit longer distances
- Ethernet over WAN or over long physical links is well understood
- Lustre ksocklnd (@tcp NIDs) has very different characteristics than ko2iblnd (@o2ib NIDs)

How Far is “Distance”?

- **FDR Infiniband performance impact starts with 50 meter cables**
- **Many “distant” installations are less than 1 Kilometer maximum**
- **Mellanox MetroX simplifies connectivity up to 80 KM**
  - Increasing infrastructure costs with distance
- **Specialized products available for worldwide use**
  - Fairly expensive
  - Relatively small bandwidth increments per unit

What is the Problem with Distance?

- Individual Infiniband links (cables) are flow controlled using link credits (Infiniband is a lossless network)
- No transmission without credit, no new credit until remote side sends them
- Time * Bandwidth product tells you how many credits you need to fill a given single hop connection
- At FDR, 50 meters or more, you need more than the typical credits used for a “default” installation
- Using more credits means more data in flight

What is Time * Bandwidth?

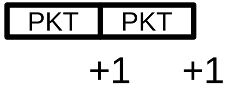
- **4x QDR = 0.25 ns/byte**
- **4x FDR10 = 0.206 ns/byte**
- **4x FDR = 0.146 ns/byte**
- **4x EDR = 0.082 ns/byte**

- **Optical Cable is typically 5 to 5.5 ns/meter**

- **Total delay is cable delay plus transceivers plus end point serialization/deserialization**
  - Transceiver models vary in delay time
  - End points vary in delay time

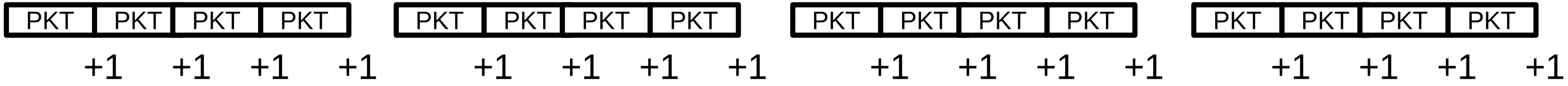
Infiniband Link Credits in Action

Short, normal cable


PKT PKT

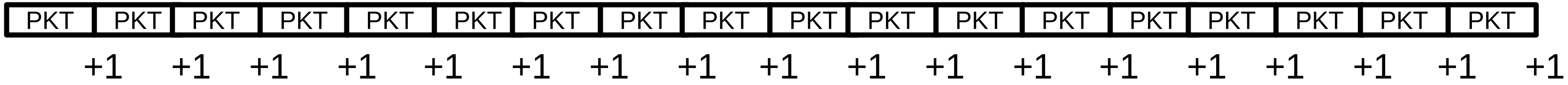
+1 +1

50 meter cable, not quite enough link credits


PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT

+1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1

Any cable, sufficient link credits


PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT PKT

+1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1 +1

How do you get more credits?

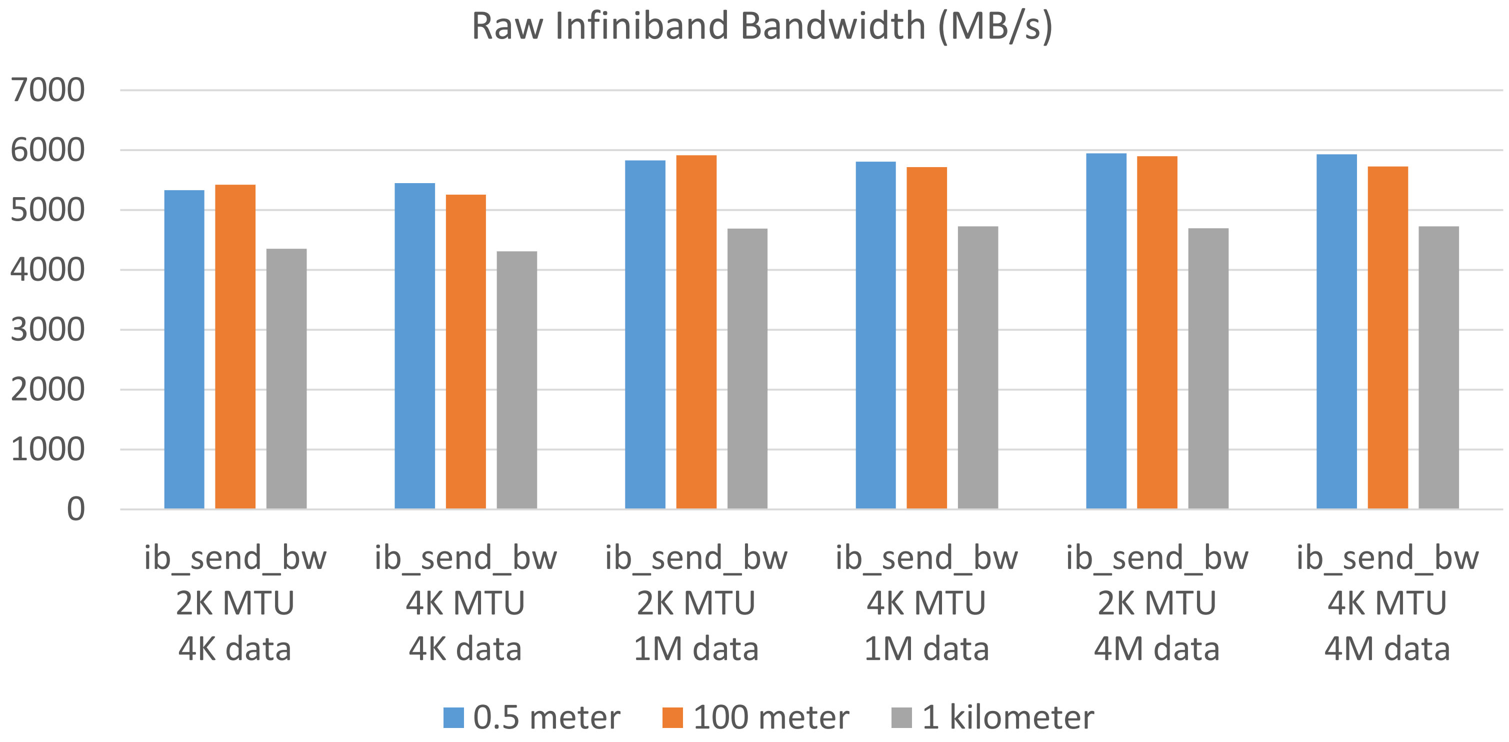
- **If the distance is short enough (< 100 meters) you can tune normal switches and end points to have more credits by reducing the number of Virtual Lanes (VLs)**
  - Credits = Total buffer space / Number of ports / Number of VLs
  - If you don’t know what VLs are, you don’t need them
- **As distance increases, specialized switch equivalents are used with appropriate credits for the long links**
  - More credits means more buffer space
  - Increasing costs for longer distances

Is it just about credits?

- For shorter distances (you can easily walk everywhere) all you really need are the credits
- For longer distances there are complexities coordinating activities at the different locations
- If resiliency is desired, consideration must be given to possible isolation (split fabric)
  - Power failure is the common culprit
  - Cables run in common space can be simultaneously lost
  - Need Infiniband subnet management distributed

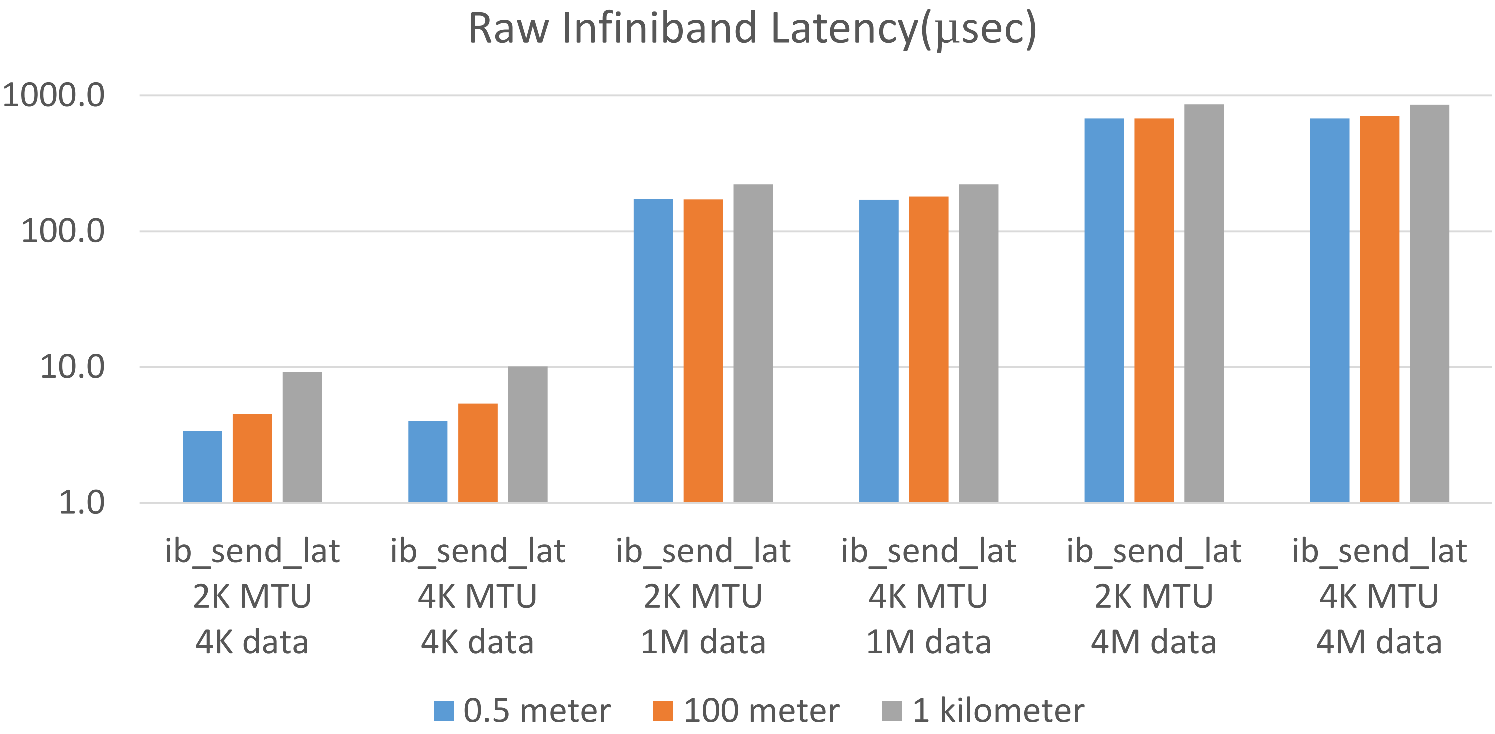
Test Results

Bandwidth at Three Distances



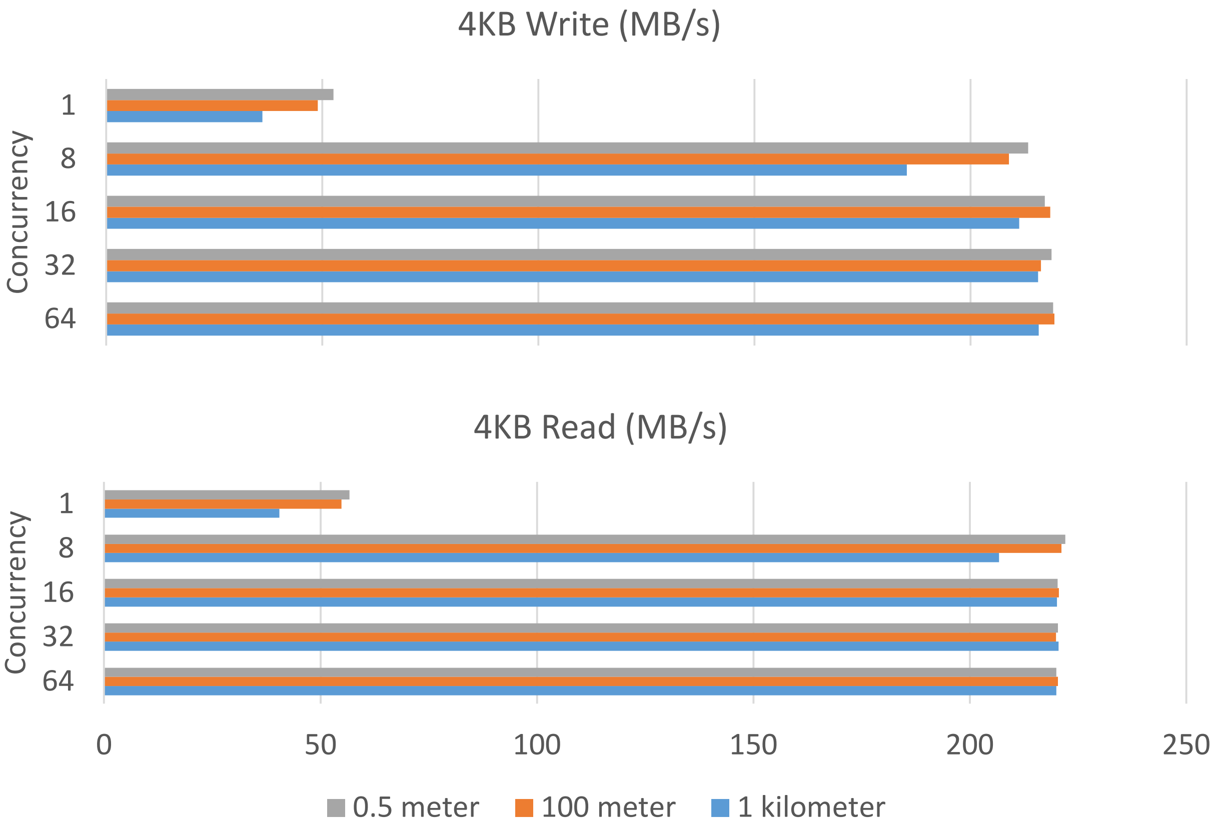
FDR Infiniband for 0.5 and 100 meters, FDR10 for 1 kilometer

Latency at Three Distances

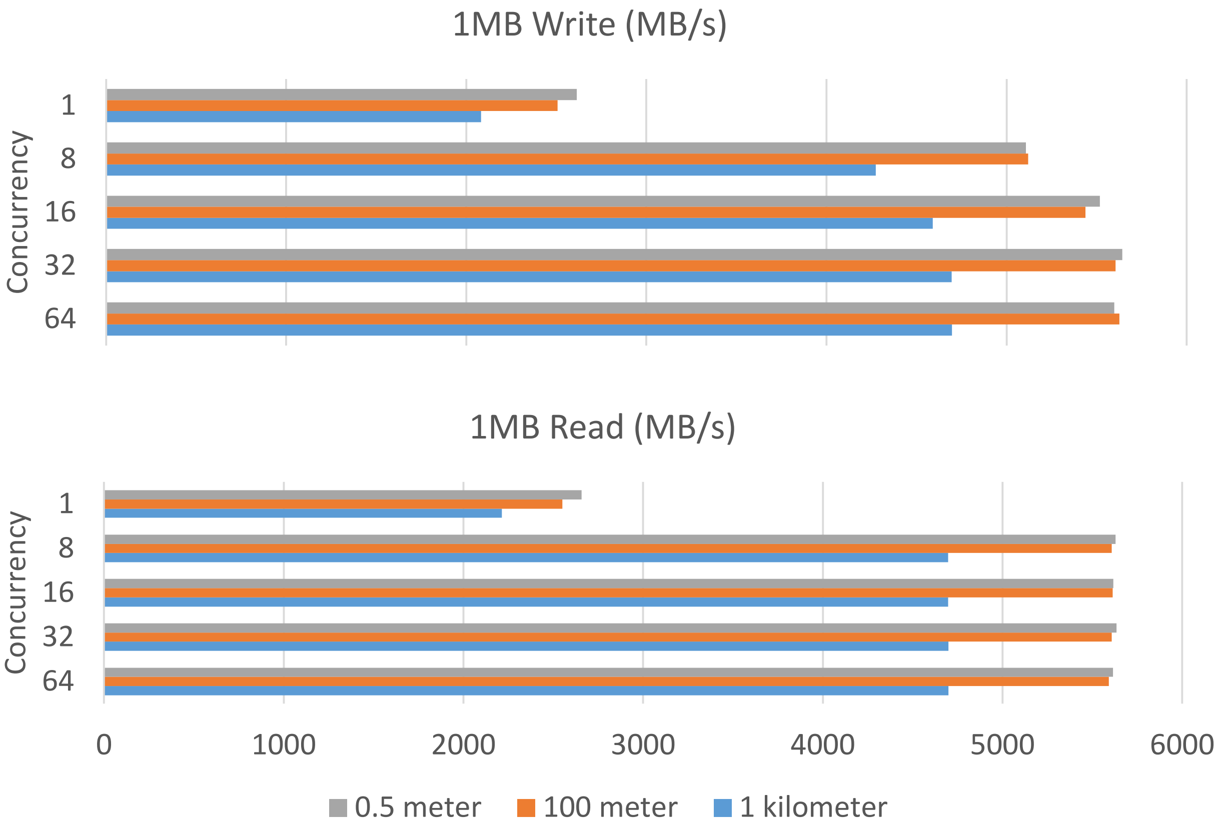


FDR Infiniband for 0.5 and 100 meters, FDR10 for 1 kilometer

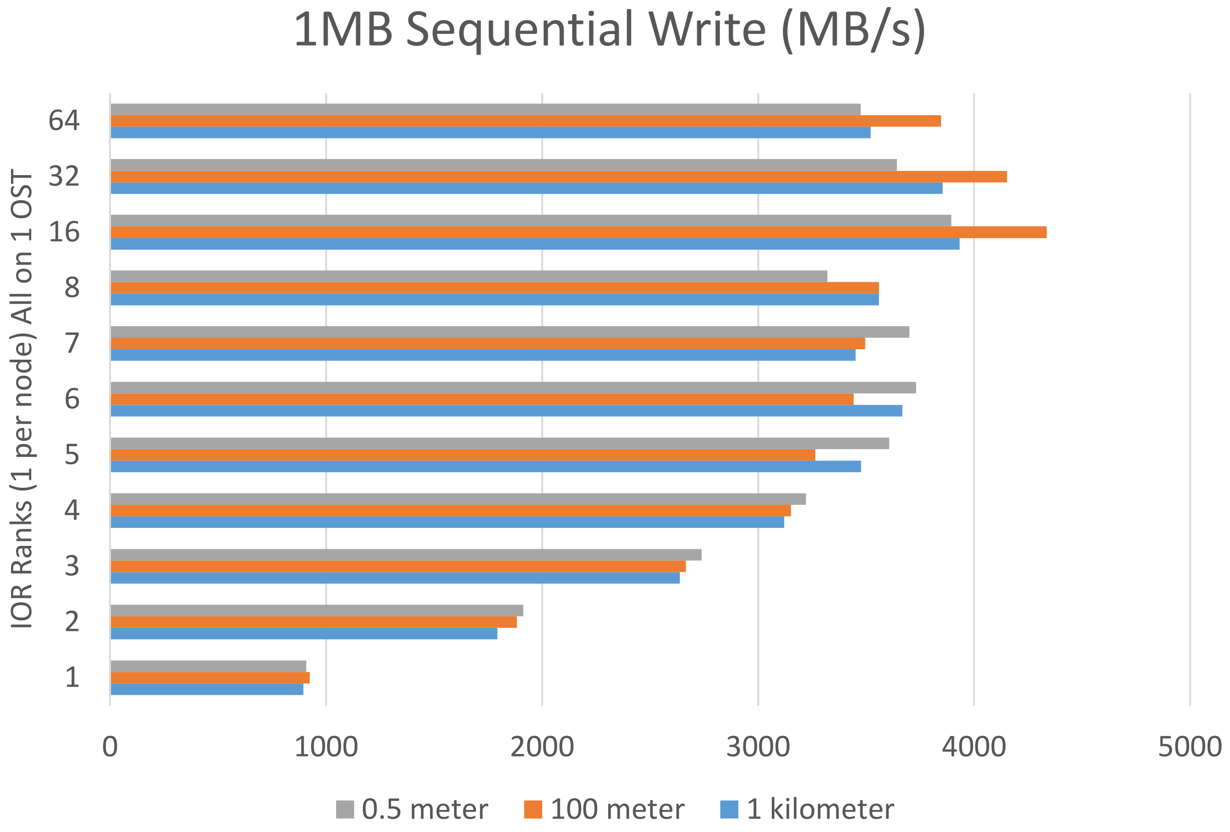
LNET Self Test



LNET Self Test

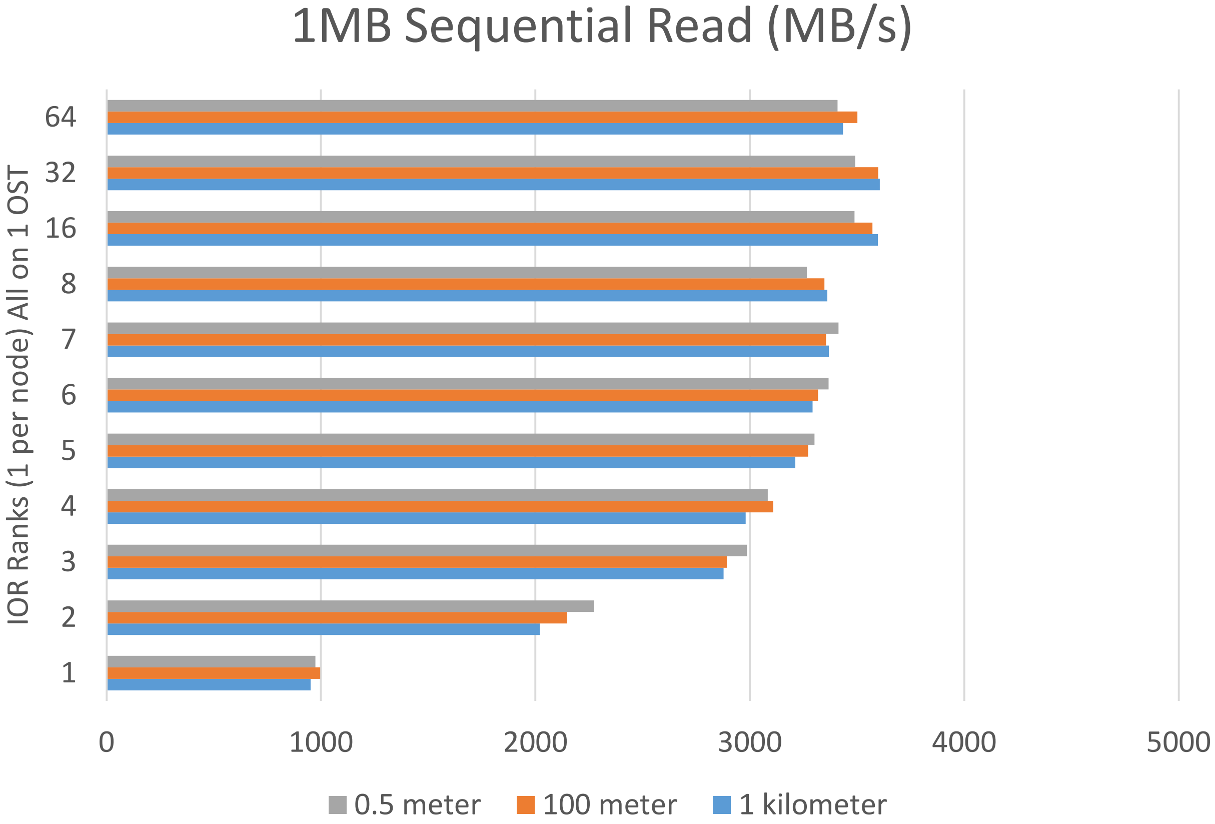


IOR (Actual File I/O)

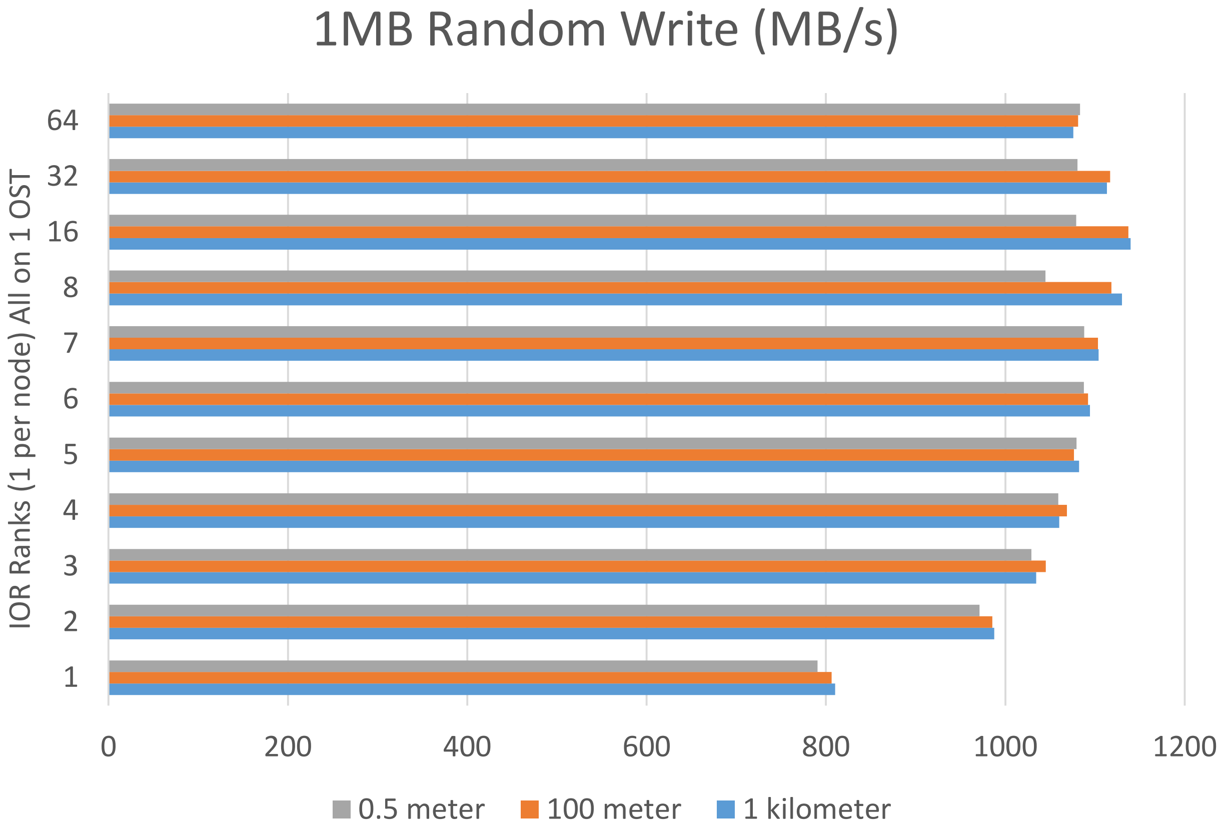


COMPUTE | STORE | ANALYZE

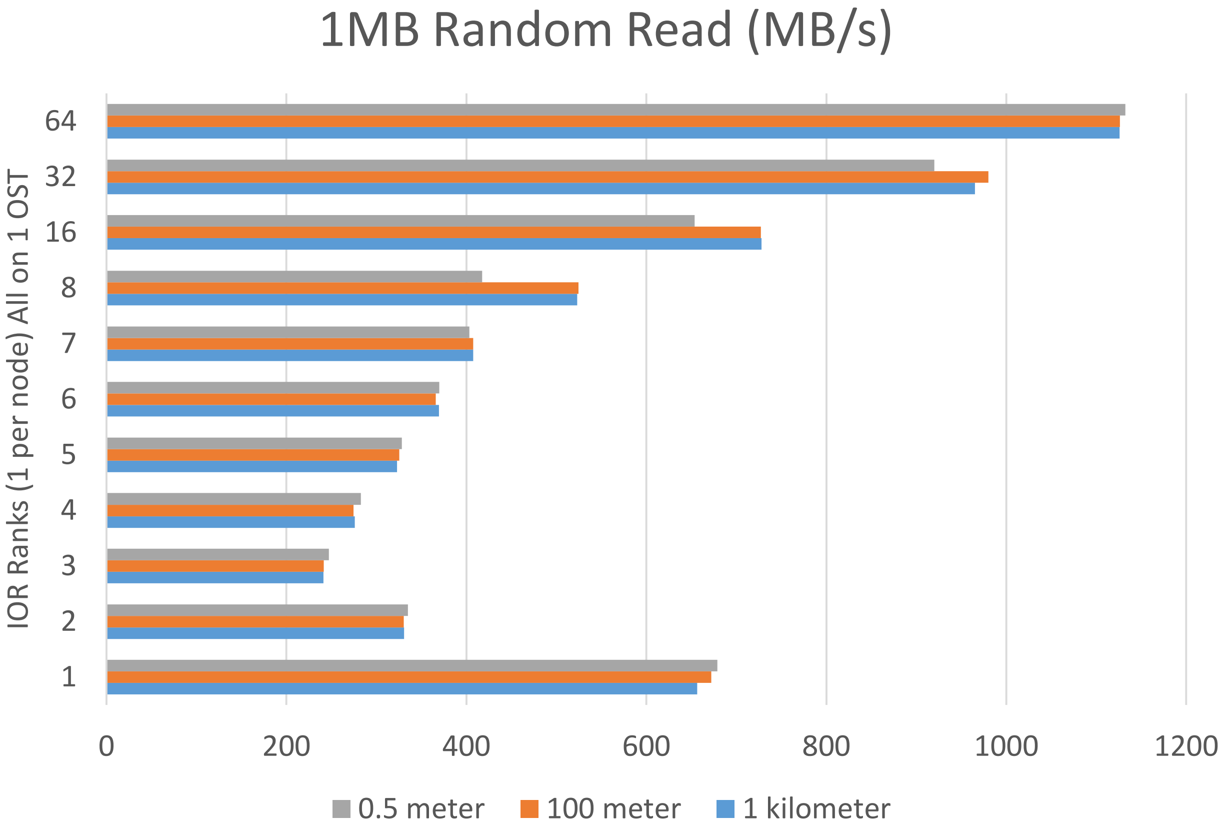
IOR (Actual File I/O)



IOR (Actual File I/O)

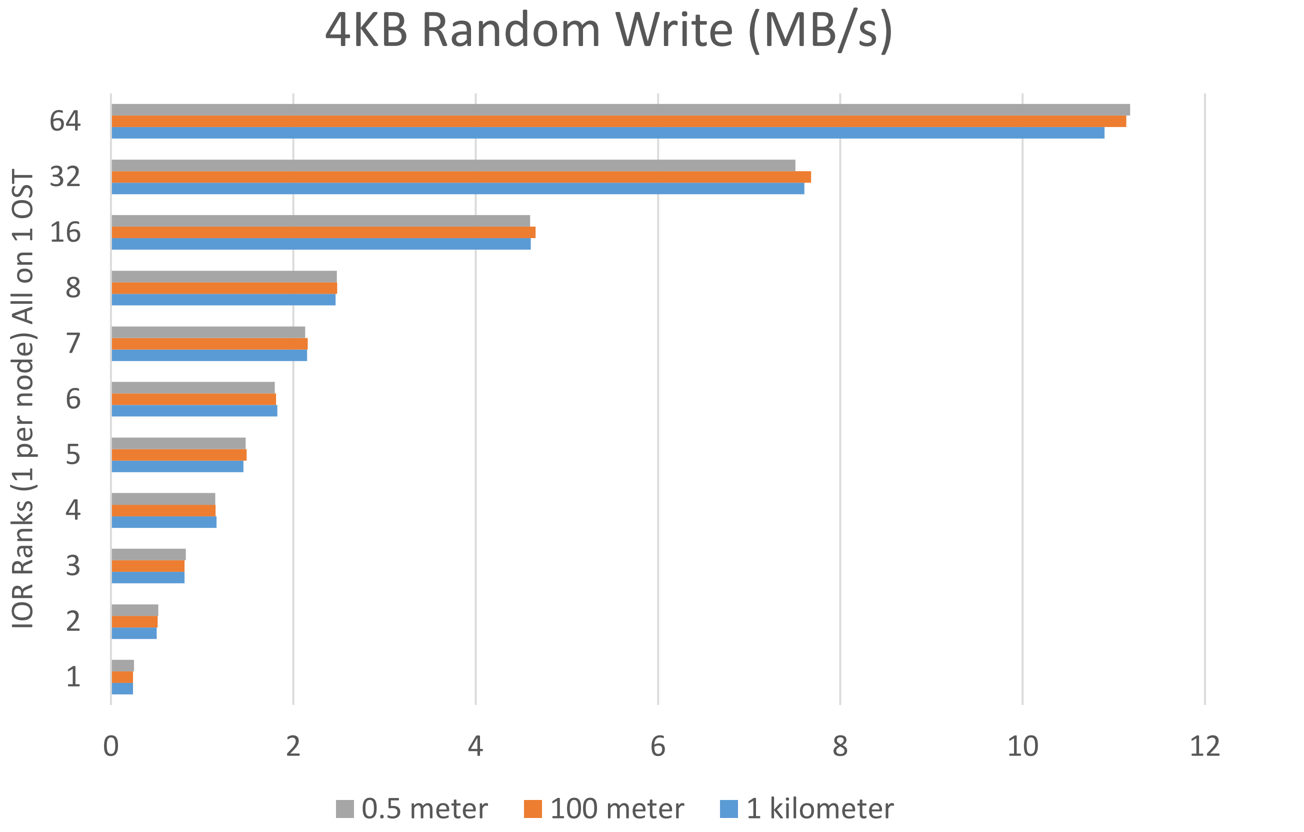


IOR (Actual File I/O)



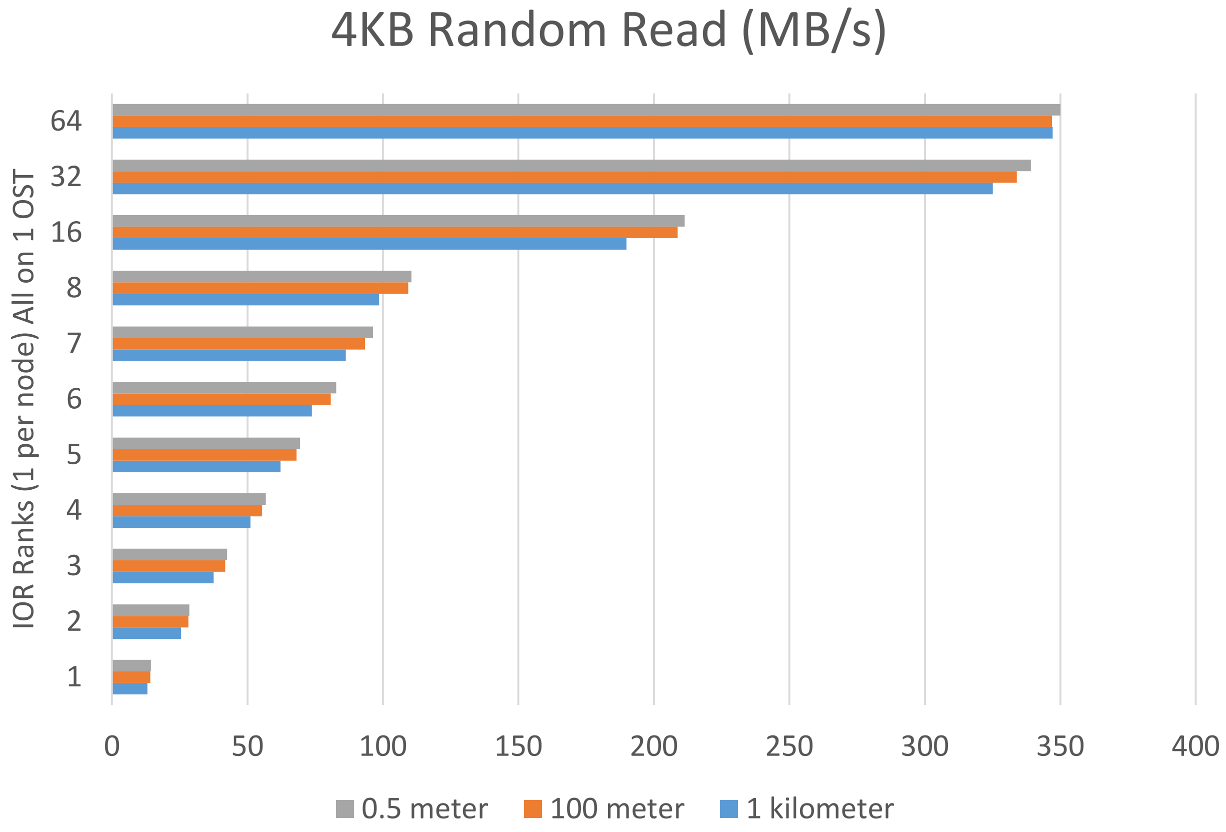
COMPUTE | STORE | ANALYZE

IOR (Actual File I/O)



COMPUTE | STORE | ANALYZE

IOR (Actual File I/O)



What Workloads Operate Over Distance?

- **Single client is fully exposed to the round trip latency**
  - Simple single-buffered I/O performance will suffer
  - Can use multiple buffers to overcome problem
  - Striped files can help, especially for writing
  - Not usually easy to fix
- **Multiple clients already interleave I/O requests**
  - Shared use of Lustre servers means network sees multiple buffers
  - The more clients simultaneously sharing, the lower the impact
- **File systems with 100+ clients active will see almost no performance difference with distances under 10 Kilometers**
- **OST traffic has larger buffers and works better than MDT traffic**

Conclusion

- Increasing your Infiniband fabric diameter up to 10 KM can be straightforward
- Beyond 10 KM, or for smaller numbers of clients, consideration must be given as to how the pipelines will be filled
- These solutions are in production today

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, and URIKA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.