ETHERNET XXX
IWARP PERFORMANCE STUDY

Presenters

Michael Fenn

The Presenters

ethernet alliance

Michael Fenn is currently a Systems Administrator at D. E. Shaw Research, a research lab engaged in the field of computational biochemistry.

At the time of that the results mentioned in this talk were gathered and that the accompanying whitepaper was published[1], he was a Systems Administrator in the Research Computing and Cyberinfrastructure group at The Pennsylvania State University.

[1] https://www.openfabrics.org/ofa-documents/presentations/doc_download/514-iwarp-learnings-and-best-practices.html

ethernet alliance

ETHERNET ALLIANCE

The views WE ARE expressing in this presentation are our own personal views and should not be considered the views or positions of the Ethernet Alliance®, the Pennsylvania State University, or D. E. Shaw Research, LLC.

AGENDA

ethernet alliance

ETHERNET ALLIANCE

- Ethernet Alliance Overview
- What is RDMA?
- What is iWARP?
- Networking Considerations
- iWARP Software Setup
- MPI Considerations
- Test Description and Environment
- Performance Observations
- Multi-Fabric Hosts
- Testing Conclusions
- Acknowledgements
- Ethernet Alliance Membership Benefits

ETHERNET ALLIANCE MISSION

ethernet alliance

- To promote industry awareness, acceptance and advancement of technology and products based on, or dependent upon, both existing and emerging IEEE 802 Ethernet standards and their management.
- To accelerate industry adoption and remove barriers to market entry by providing a cohesive, market responsive, industry voice.
- Provide resources to establish and demonstrate multi-vendor interoperability.

Ethernet Alliance Strategic Vision

ethernet alliance

Expand Ethernet Ecosystem

- Facilitate interop testing
- Expand the market
- Go global

Support Ethernet Development

- Support consensus building
- Host Technology Exploration Forums (TEFs)
- Team with other orgs

Promote Ethernet

Marketing

Education

University of Ethernet Curriculum

ethernet alliance

- Completed and available online
- Planned
- Concept

Ethernet 101: Introduction to Ethernet

Physical Layer x00 Series	Protocols x10 Series	Applications x20 Series	Products x30 Series
Ethernet 102: The Physical Layer Of Ethernet	**Ethernet 111:** 802.1:Protocols Of Ethernet	**Ethernet 121:** The Applications Of Ethernet	**Ethernet 131:** Ethernet Products
Ethernet 202: 10GBASE-T Revamped	**Ethernet 211:** Data Center Convergence	**Ethernet 221:** Data Center Applications	**Ethernet 231:** Ethernet Switches
Ethernet 301: 40/100GbE Fiber Cabling and Migration Practices	**Ethernet 311:** Congestion Notification	**Ethernet 321:** Industrial Applications	**Ethernet 331:** Ethernet Server Adapters

Ethernet Alliance

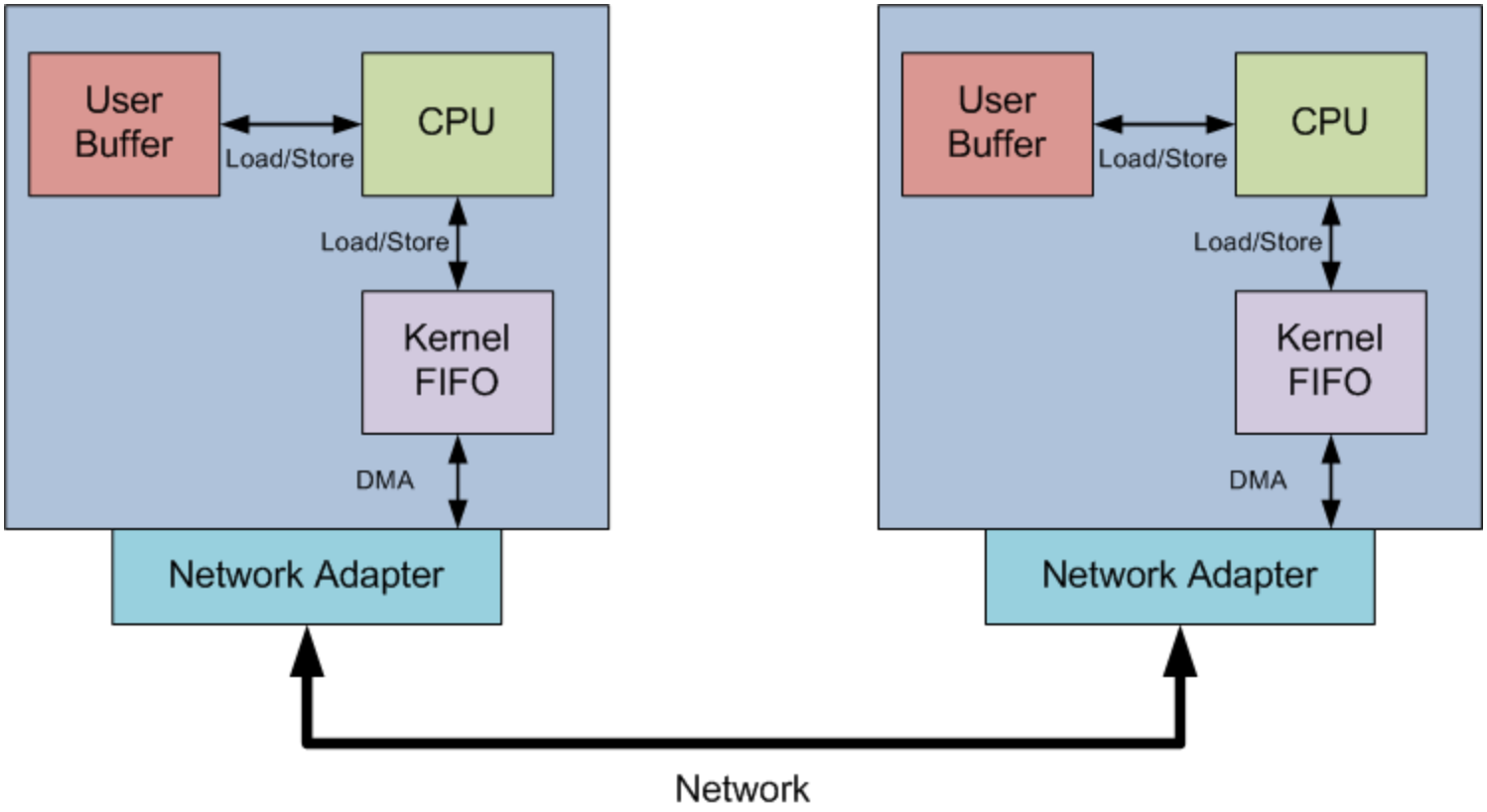
WHAT IS RDMA?

ethernet alliance

- Before talking about iWARP, we need to discuss RDMA
- What is RDMA? **R**emote **D**irect **M**emory **A**ccess
- The big performance inhibitor in data center networks is the number of times that data must be copied in order to get it from one application's buffers to another.
- RDMA allows for "zero-copy" data transfers from one host to another.

WHAT IS RDMA?

ethernet alliance

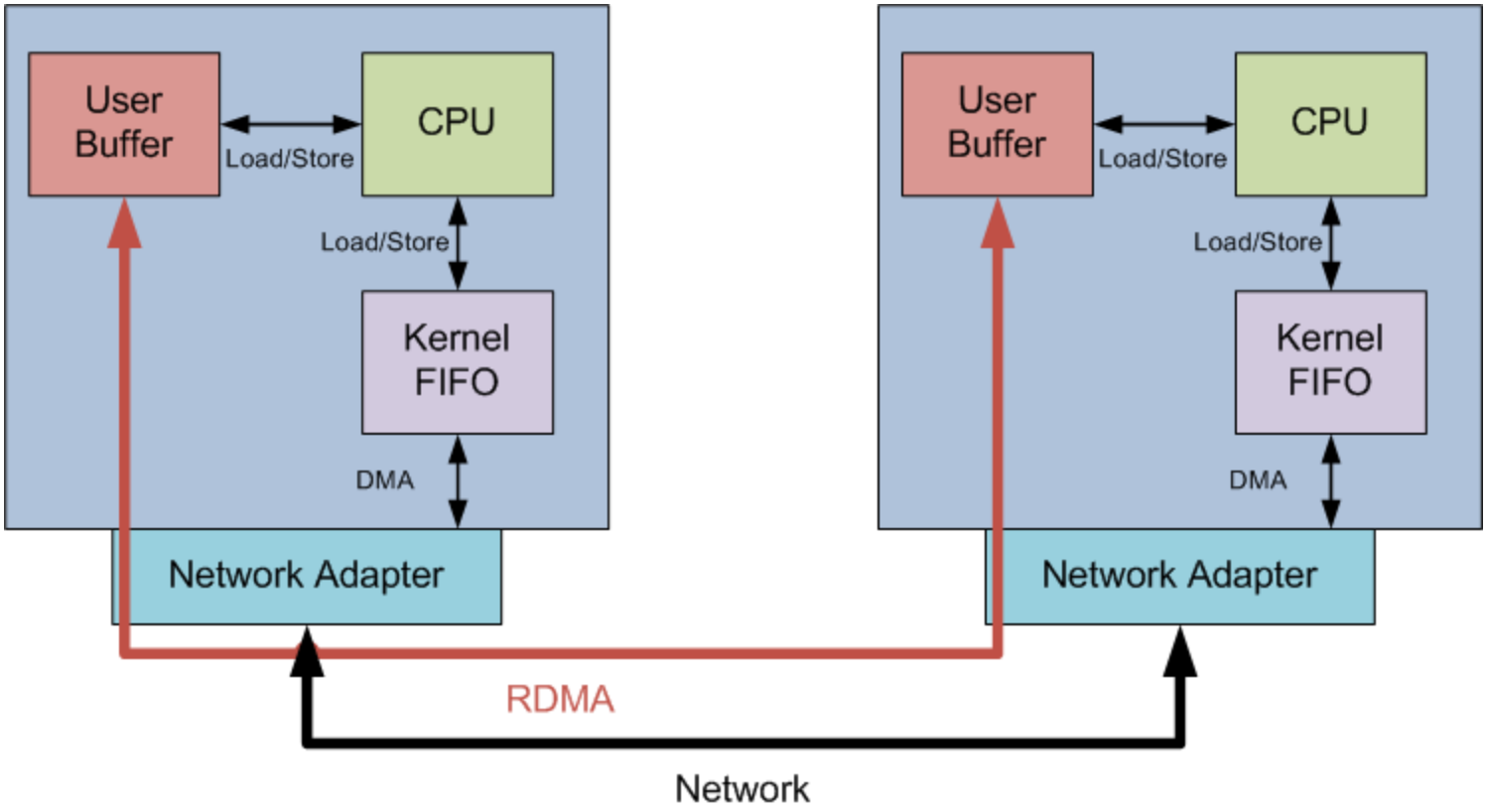


Traditional network architecture

ETHERNET ALLIANCE

WHAT IS RDMA?

ethernet alliance



RDMA-enabled network architecture

ETHERNET ALLIANCE

What is iWARP?

ethernet alliance

- Internet Wide Area RDMA Protocol
- Essentially, iWARP allows RDMA over TCP
- iWARP allows RDMA applications to work over an arbitrary TCP connection
- RDMA applications typically expect low latency communication
- Ways to lower latency:
  - Kernel bypass drivers
  - Acceleration of the transport protocols
  - A low-latency, well-provisioned fabric
- An Intel NetEffect card takes care of the first two, up to you to provide the third.

The Fabric

ethernet alliance

- For good performance in RDMA applications, you need a low-latency Ethernet switch
- Our 7148SX has 1.2µs latency
  - It's a couple years old, newer switches are well into the 100's of nanoseconds
- Dropped packets and retransmissions kill performance
  - Use flow control: 802.3x if you must, but 802.1Qbb if you can
  - Better yet, design a fully non-blocking network

Non-blocking Ethernet Fabrics

ethernet alliance

- Is the dream yet a reality?
- With a small number of hosts, this is easy:
  - < 48, use a fixed-port switch
  - < ~384, use a chassis switch with non oversubscribed line cards
- Beyond that?
  - That darn spanning tree gets in the way of designing a true fat tree fabric
  - Transparent Interconnect of Lots of Links (TRILL) seems to be the solution, but so far implementations are proprietary

JUMBO FRAMES

ethernet alliance

- If your network doesn't support Jumbo frames yet, don't worry
- VASP results:
  - iWARP without jumbo frames: 52.30 minutes
  - iWARP with jumbo frames: 51.26 minutes
  - Surprisingly (at least to me) 1500 byte frames are only about 2% slower than 9000 byte frames in message passing applications
- My theory is that these classes of application are more latency-sensitive than bandwidth-sensitive
- This was largely borne out in our larger comparison between IB and iWARP

iWARP Software Setup

ethernet alliance

- BIOS setup similar to other high-performance RDMA networks (IB, RoCE)
  - Disable C-states
  - Disable PCIe link power management
- Increase memlock ulimits
- Need to use a recent OFED
  - RHEL 5's bundled OFED is too old

More Software Setup

ethernet alliance

- The iw_nes driver needs some extra parameters in /etc/modprobe.conf:

```
options iw_nes nes_drv_opt=0x110
options rdma_cm unify_tcp_port_space=1
alias eth2 iw_nes
install iw_nes /sbin/sysctl -w
net.ipv4.tcp_sack=0 > /dev/null 2>&1;
/sbin/modprobe --ignore-install iw_nes
```

- Needing to keep track of which eth* device the NetEffect card is can somewhat complicate deployments on diverse hardware

Optional TCP Tuning

ethernet alliance

- These sysctl parameters control the behavior of the Linux TCP stack, but don't affect the hardware TCP engine in the NetEffect:

```
net.ipv4.tcp_timestamps=1
net.ipv4.tcp_sack=0
net.ipv4.tcp_rmem=4096 87380 4194304
net.ipv4.tcp_wmem=4096 16384 4194304
net.core.rmem_max=131071
net.core.wmem_max=131071
net.core.netdev_max_backlog=1000
net.ipv4.tcp_max_syn_backlog=1024
net.ipv4.tcp_window_scaling=1
net.core.rmem_default=126976
net.core.wmem_default=126976
net.core.optmem_max=20480
```

MPI Implementations

ethernet alliance

- iWARP is well-supported by popular Message Passing Interface (MPI) implementations
  - OpenMPI
  - MVAPICH2
  - Intel MPI
  - Platform MPI (neé HP-MPI)
- We used OpenMPI and HP-MPI in our testing

The Test

ethernet alliance

- Our testing goal was to evaluate the difference between InfiniBand and iWARP running over 10 Gb Ethernet.
- We already knew that either would be superior to a traditional 1 Gb Ethernet network.
- The test was to run various MPI applications and observe the relative scaling between IB and iWARP as we increased the number of cores.
- In other words, the problem stayed constant, so we expect to see faster run times as we increase the number of cores.

The Test Environment

ethernet alliance

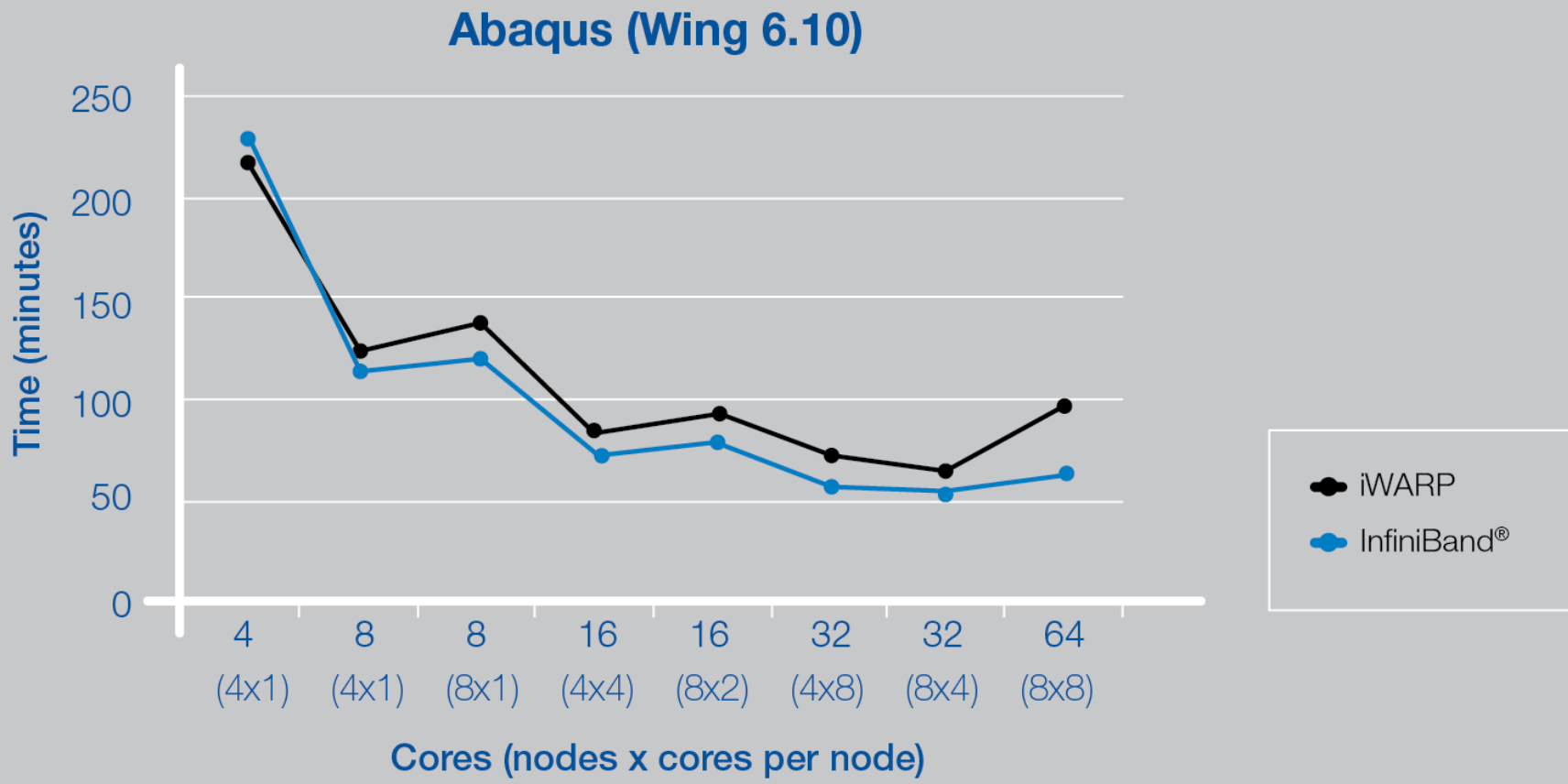
- The test results presented here were performed with the following hardware:
- Dell PowerEdge R710 servers
  - Two Xeon X5560 processors (2.80 GHz)
  - 48GB DDR3 1333 memory
  - Intel NetEffect 10Gb Ethernet Adapter
- Red Hat Enterprise Linux 5.6
- OFED 1.5.2
- Arista 7148SX 10Gb Ethernet Switch

Performance Observations

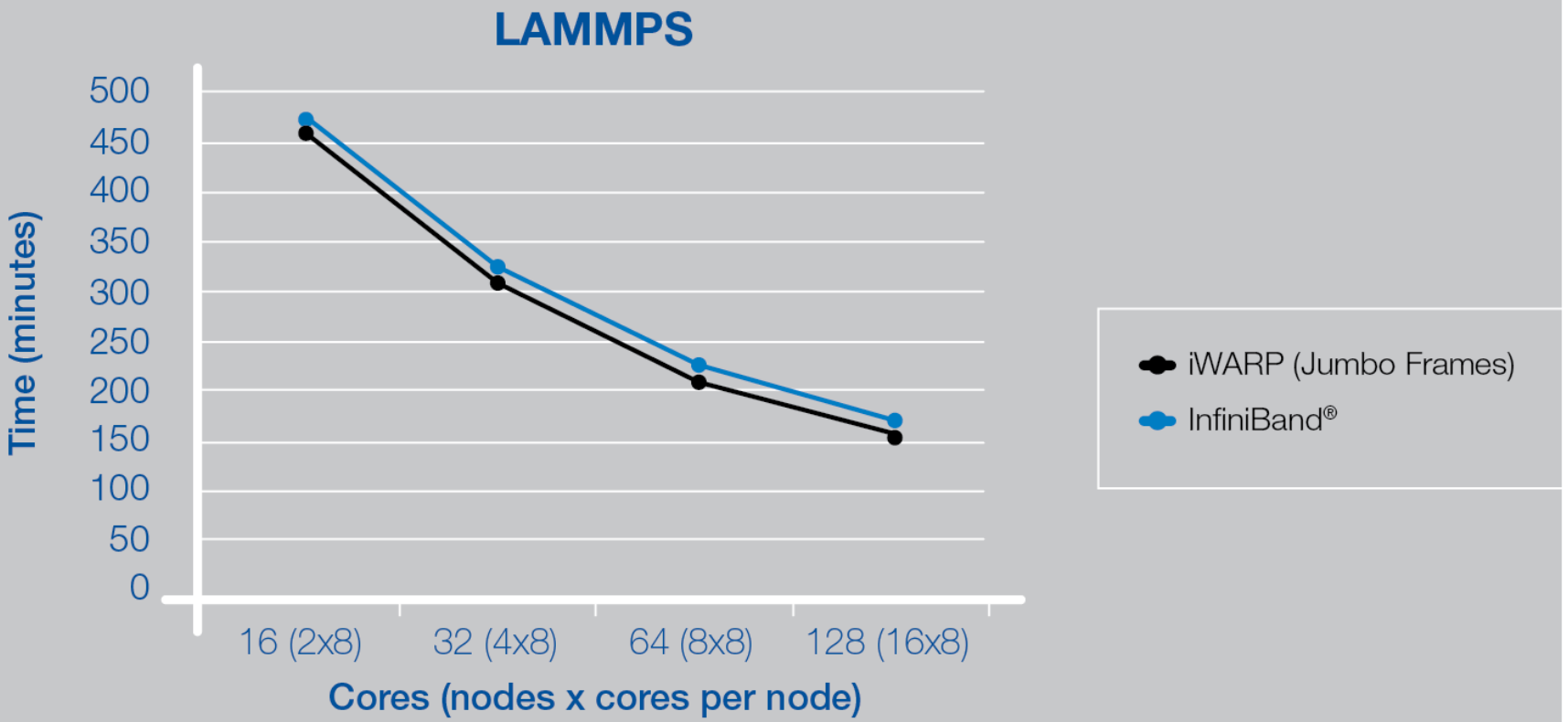
- With our 7148SX, 7.5µs IMB PingPong latency
  - 2.4µs is attributable to going through the 7148SX twice
- Application codes scaled well, (within the limits of our environment and benchmark)
  - Abaqus (HP-MPI)
  - LAMMPS (OpenMPI)
  - LS-DYNA with MPI i.e. MPPdyna (OpenMPI)
  - Quantum Espresso Plane Wave (OpenMPI)
  - VASP (OpenMPI)
  - WRF (OpenMPI)

PERFORMANCE OBSERVATIONS

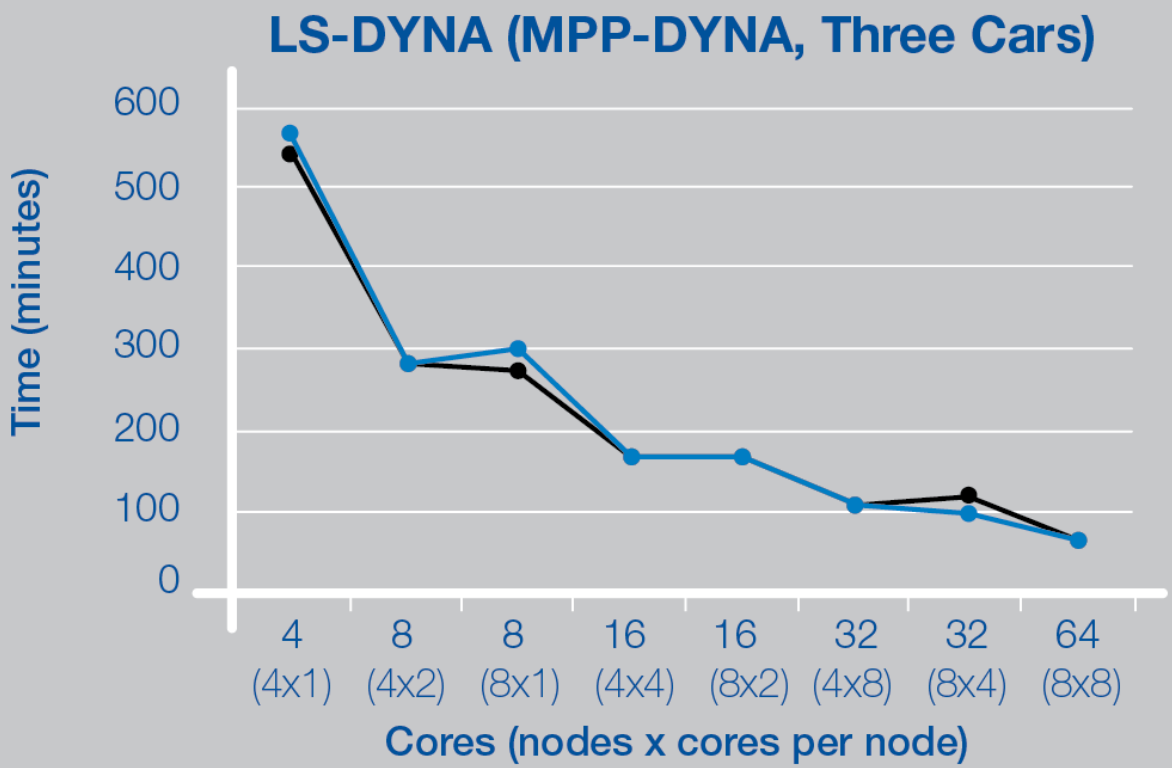
ethernet alliance



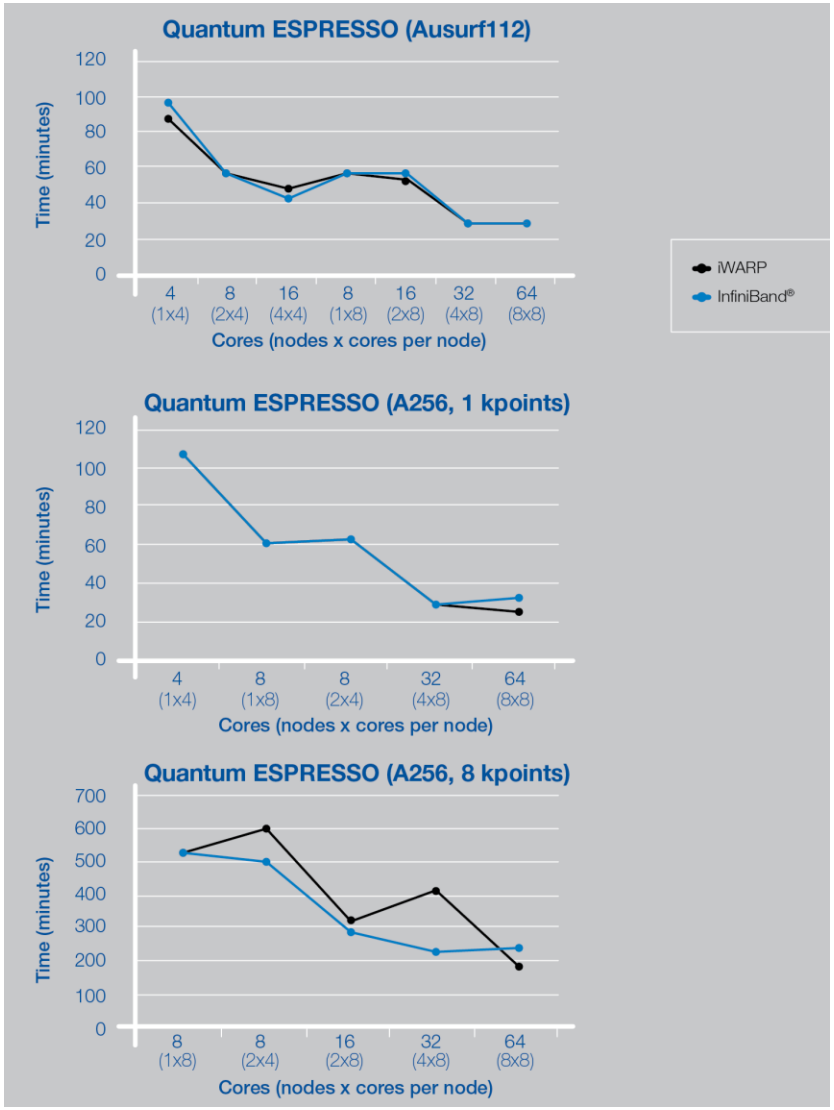
Performance Observations



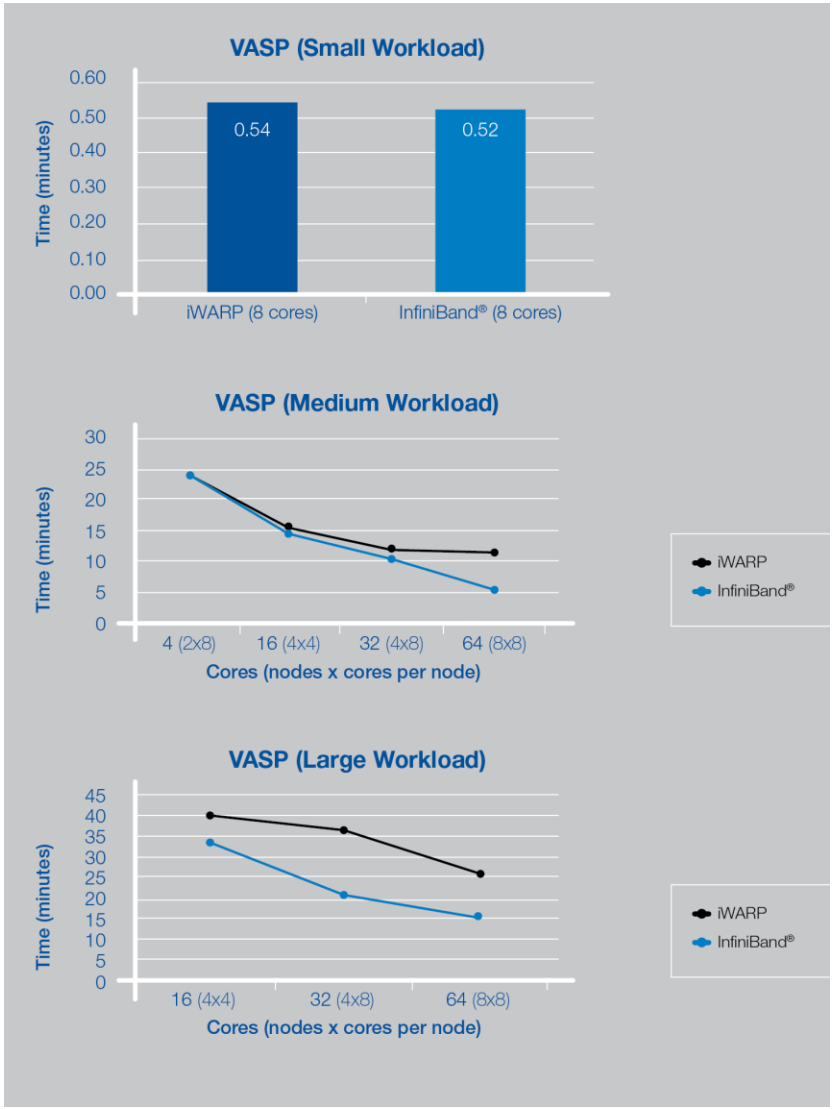
Performance Observations



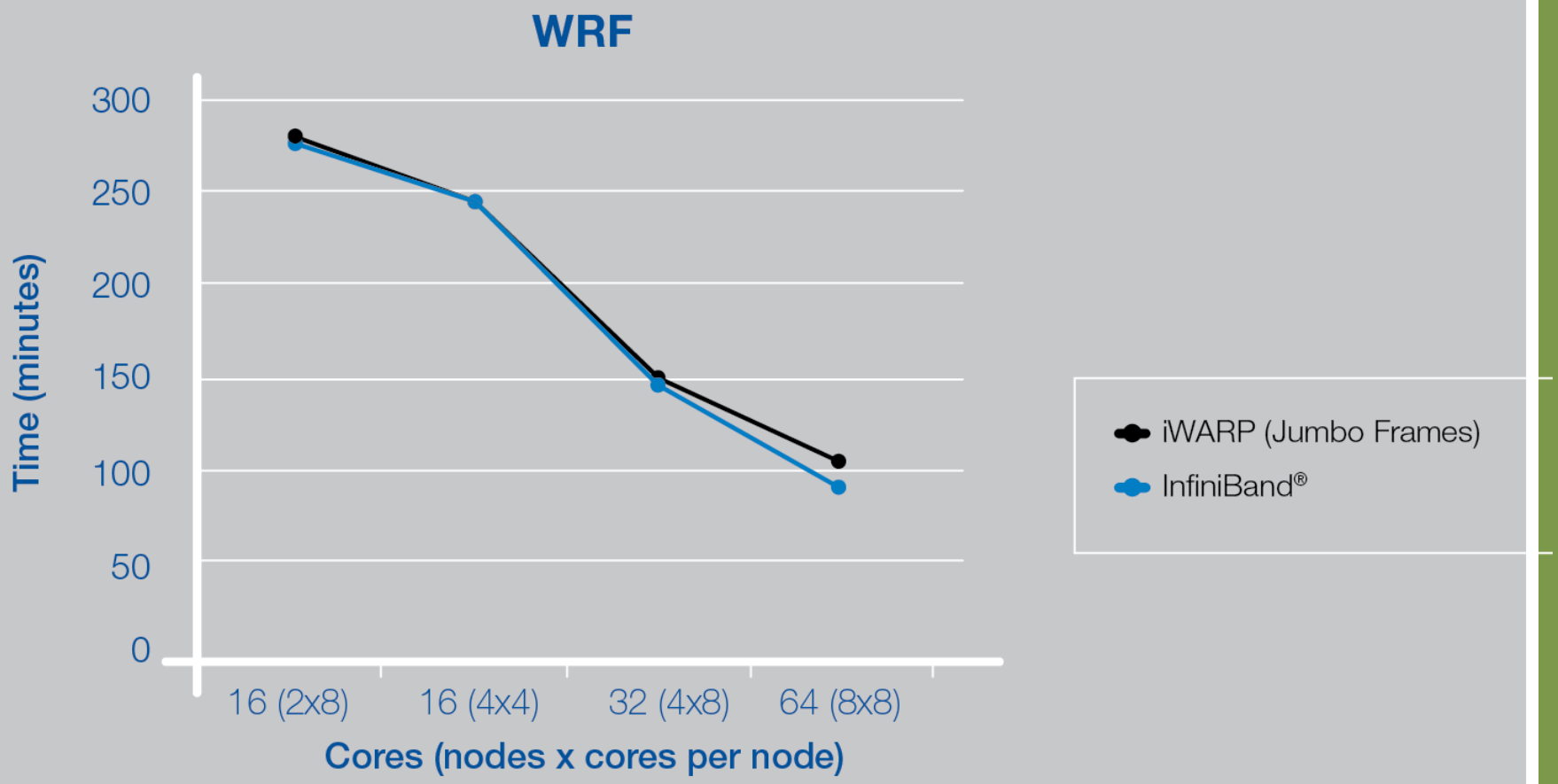
PERFORMANCE OBSERVATIONS



PERFORMANCE OBSERVATIONS



Performance Observations



Performance Observations

ethernet alliance

- Sometimes we did notice performance degradation
  - High amount of time spent in systems calls
  - No apparent extra load on the network
  - Some more driver tuning would be useful (probably addressed in newer OFED)
- Read the paper for the full results

Multi-Fabric Hosts

ethernet alliance

- What if you want to have iWARP and Infiniband interconnects on the same machine?
  - Could imagine a situation where RDMA over TCP is used for (say) storage, but Infiniband is used for MPI interconnect
  - Another case is in a benchmarking environment, it is very useful to be able to run tests back to back with no hardware reconfiguration required
- This is possible, at least for some subset of cases

More Multi-Fabric Hosts

ethernet alliance

- OpenMPI is easy, it is an mca parameter:
  - NetEffect: `--mca btl_openib_if_include nes0`
  - Mellanox: `--mca btl_openib_if_include mlx4_0`
  - Others are possible
- HP-MPI should be easy, just change MPI_HASIC_UDAPL
  - However, since /etc/dat.conf parsing is broken, changing fabrics ends up requiring an system config file change

CONCLUSIONS

ethernet alliance

- iWARP and RDMA over Ethernet networks in general require a change in mindset
  - A 48-port 10GbE switch with a few uplinks is not sufficient
  - Need a fully non-blocking network, either in a chassis switch or with TRILL
- NetEffect hardware “looks” similar enough to IB to be supported by MPI with minor alterations (MPI applications themselves don’t care)

More Conclusions

ethernet alliance

- However, the rest of the ecosystem is still catching up
  - ISV codes bundled with old MPI versions are the biggest offender
- Impact depends on your environment
  - Could be a non-issue for environments with heavy open-source or community code usage
  - If you heavily rely on ISV codes, it could be a big impediment to an iWARP deployment

ACKNOWLEDGEMENTS

- Julie Cummings of Intel for providing expert technical assistance
- Tom Stachura and William Meigs of Intel for coordinating the testing process
- David Fair of Intel for coordinating the Ethernet Alliance webinar

Benefits of Membership

- Be part of the Voice of Ethernet!
  - Network with Ethernet Thought Leaders
  - Participate in the Debate of Ethernet Futures
  - Contribute to Ethernet Alliance Social Media
- Visibility Through Participation
  - Global Exposure
  - Broad Market Exposure
- Prove Your Interoperability
  - Plugfest
  - Live Demonstrations
- Education

Discussion/Q&A

ethernet alliance

THANK YOU