

TrueScale Difference—Providing Maximum HPC Performance

TrueScale—Optimize for Performance and Scaling

Key Findings

- There are two types of InfiniBand host adapters, one based on an offload processor and the other based on an on-load architecture. Tests show that the on-load InfiniBand architecture with its lightweight InfiniBand protocol is better at scaling, message processing, and latency. The reason is that it can harness the processing power of today's faster and denser core count processors. The offload InfiniBand with its less capable microprocessor can become a bottleneck to today's processor, limiting the interconnect performance for each host on a cluster.
- The TrueScale™ on-load host architecture offers five times the message throughput of the other InfiniBand architecture. For many Message Passing Interface (MPI) applications, messages throughput is an important factor that contributes to overall performance.
- Cluster message rate, which is the ability of the InfiniBand host and switch technologies to process messages together, is another key factor in driving application(s) performance and scalability. QLogic's TrueScale-based cluster offers more than twice the bidirectional message rate throughput as a cluster based on the other InfiniBand provider's technology.
- End-to-end latency is the final determinate of an MPI application's performance and ability to scale. QLogic's TrueScale end-to-end latency is almost 50 percent better than the other major InfiniBand offering available in the market.
- Real application performance is the ultimate measure of an InfiniBand architecture's performance. QLogic tested a number of MPI applications and found that they performed up to 22 percent better on the cluster based on TrueScale.

Executive Summary

QLogic® conducted a performance study comparing the two major InfiniBand® architectures to determine their performance. The interconnect is a major factor in determining the performance of a High Performance Computing (HPC) cluster. InfiniBand has proven itself to be the interconnect of choice for HPC clusters requiring maximum performance while still using an open-standard interconnect. In fact, the list of the Top 500 supercomputers (Top500.org) shows that 88 percent of the systems are connected with either Gigabit Ethernet (GE) or InfiniBand. GE is utilized by 52 percent of the systems, but only accounts for 26 percent of the total Top 500 performance. InfiniBand is used by 36 percent of the systems, but

achieves a disproportionate 43 percent of the Top 500 performance. This means that InfiniBand has a 70 percent performance advantage over GE. The reasons InfiniBand is now the leading choice for major High Performance Computing (HPC) fabrics are that it offers the highest available bandwidth and the lowest available latency of the open/standards-based interconnect. But depending on the design of the InfiniBand architecture, its advantages can be squandered as the number of compute nodes scales up into the dozens, hundreds or thousands. One of the main challenges is achieving efficient cluster performance scaling, which can be impacted by the type of InfiniBand architecture that is used.

Background – Adapter-based vs. Host-based Processing

There are essentially two types of InfiniBand architectures available today in the marketplace. One was created in the early 2000s when InfiniBand was first being designed as a fabric for the enterprise data center. The other architecture is a more recent design, and was created when it became clear that HPC was the major market for InfiniBand. This later generation of InfiniBand was designed to run HPC/MPI applications, as well as accommodate the latest processor technology with its multi-core implementation. The traditional InfiniBand architecture had MPI retrofitted on top of its InfiniBand transport layer, which had been designed for the enterprise data center. In contrast, the latest generation of InfiniBand is designed for the HPC market and the current generation of faster, denser core count processors.

The two generations of InfiniBand handle protocol processing very differently. An organization’s choice of InfiniBand architecture can make a significant difference in overall fabric and application performance, particularly as the size of the cluster scales. Some vendors rely heavily on adapter-based (“on-load”) processing techniques, in which each InfiniBand adapter includes an embedded microprocessor that processes the communications protocols. Other vendors primarily use host-based processing, in which the server processes the communications protocols.

Adapter-Based/Offload Adapter Architecture

Just a few years ago, a typical server may have had just one or two single- or dual-core processors and relatively slow PCI or PCI-X buses. Since these processors had the ability to issue only one instruction per clock cycle at a relatively low clock rate, the servers benefitted from having communications processing offloaded to the adapter.

Today’s CPUs provide significantly more power than only a short time ago (Xeon® 5400 - Harpertown), and they are far faster than a microprocessor engine typically found in an offload device, such as those on a traditional Host Channel Adapter. The Intel® Xeon 5500 - Nehalem processor issues four instructions per clock cycle, and operates at a clock speed of 3GHz. As a result, each Nehalem processor has an *execution rate 24 times that of a generic microprocessor* engine operating at 500MHz (Figure 1). This difference in processing power can potentially overload the adapter- based/ offload microprocessor, thus making it a “bottleneck” to the host and HPC cluster’s performance.

	Clock Frequency	Instruction Issues per Clock	
Generic Micro-engine	500MHz	1	24 x
Intel Nehalem (core)	3GHz	4	
Advantage	6x	4x	

Figure 1. Nehalem vs. Host Channel Adapter offload processor instruction rates

With the recent releases of denser core count processors from Intel and AMD®, the potential for overloading the Host Channel Adapter’s microprocessor has increased. For example, Intel’s Xeon 5600 “Westmere” processor has six cores, which means that the processing overload on the adapter’s *microprocessor has now increased 36 times*. Therefore, with a *dual socket server the processing overload is approaching 72 times*. The adapter offload microprocessor can become a “bottleneck” for Nehalem-based servers and even more so with the Westmere processor.

Host-Based/On-Load Host Channel Adapter Architecture

InfiniBand that is designed with host-based processing is a much different approach. Host-based Host Channel Adapter architecture depends on the node or server to process the InfiniBand protocol. This allows the host-based protocol processing performance to scale in a much more linear fashion along with the number of available cores. More cores equal faster performance, enabling users to leverage Moore's Law; users can continually scale InfiniBand protocol processing provided they have an adapter that can take full advantage of the added power. Complementing this design is a transport protocol that was created for HPC/MPI market requirements. This transport protocol is a "lightweight" that is built around a tag that matches semantics similar in concept to high performance HPC interconnect pioneers: Myricom® and Quadrics®. By combining the host-based adapter's ability to leverage the greatly increased processing power of today's processors with an efficient InfiniBand protocol design, the host-based/on-load adapter can provide optimal HPC application performance and scaling.

Performance Study – Different InfiniBand Architectures

QLogic conducted a study of the two major InfiniBand architectures to determine their performance characteristics for HPC applications using MPI. This study was conducted at the NETTrack Developer Center located in Minnesota. The tests were completed using an Intel-based cluster consisting of 16 nodes and 128 cores. Each node had dual Intel Xeon 5570 – 2.93GHz processors and 24GB of memory. The InfiniBand interconnects tested were the QLogic TrueScale QDR hosts and switches and the Mellanox® QDR adapter with Mellanox/Voltaire® switches.

Performance Study Objectives

The goal of the study was to analyze the following performance characteristics of the two major InfiniBand architectures:

1. The host messaging rate performance of the InfiniBand interconnect architectures. A host's ability to process MPI messages is one of the key factors in determining how MPI applications will perform and scale.
2. The cluster-level message rate performance, both unidirectional and bidirectional, of both InfiniBand solutions to determine the optimal configuration of adapters and switches for a cluster. This test is an indication of how well an end-to-end InfiniBand architecture performs and how well a cluster will perform and scale.
3. End-to-end latency performance—another key factor in determining performance of most MPI applications.
4. The total communications capacity of the InfiniBand architectures. The study's test will show the ability of the underlying communications capabilities of the interconnect architectures.
5. The MPI application performance and scaling efficiency.

Host Message Rate Performance Test

The objective of this test is to determine the host messaging rate capabilities of the InfiniBand architectures. Host rate messaging is a key factor in the performance of an application, especially as the cluster scales. As users seek to solve more complex computational problems in less time, HPC clusters continue to grow both in terms of nodes per system and cores per node. Clusters that once used two to four cores per server now typically include eight, with 12-, 16-, and 24-core servers now available. To extract the most out of the additional compute power, the adapter (or adapters in some cases) needs to keep up with the exponential increase in communications throughput required by clusters based on the latest processor technologies.

The definitive test for measuring host rate message throughput is OSU's MPI Message Rate test. The message rate test evaluates the aggregate unidirectional message rate between multiple pairs of processes. Each of the sending processes sends a fixed number of messages back-to-back to the paired receiving process before waiting for a reply from the receiver. This process is repeated for several iterations. The objective of this benchmark is to determine the achieved message rate from one node to another node with a configurable number of processes running on each node.

Note: This is a test and the below results are based on non-coalesced message rate performance. As message rate has become more recognized as an important indication of HPC performance, technology providers want to portray their products in the best possible way. Coalescing artificially increases the overall message rate by up to 3200 percent, but it requires sending one stream of messages to only one other process, which is not typical of MPI interprocess messaging patterns. In addition, coalescing adds latency to the transaction because the sending process must wait and decide whether to send the packet of messages as is or wait for other messages to add to the packet.

This test was run between two Intel-based servers with dual Xeon 5570 (2.93GHz) processors. The first test used QLogic TrueScale QLE7340, the QDR host-based architecture, and adapters in each server connected to a 12300 TrueScale QDR switch. Then the test was iteratively repeated with generations of Mellanox adapter/offload-based QDR adapters (ConnectX® and ConnectX-2), which were connected to Mellanox MTS3600 QDR switch.

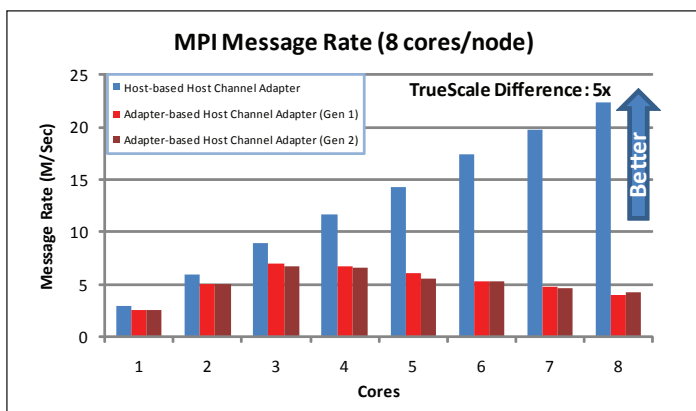


Figure 2. Message rate of host-based/on-loaded adapter vs. adapter-based/off-loaded adapter

Figure 2 illustrates that the adapter-based/on-board protocol processor adapter “tops out” at roughly seven-million messages per second. More significantly, the performance of this adapter actually declines as the number of processor cores moves beyond four. In contrast, the *host-based adapter offers more than five times more message throughput* at eight cores than the adapter/offload adapter. When you extrapolate this effect out over hundreds of nodes, it becomes clear that when adapter-based processing is the primary technique in use, the incremental benefit of adding nodes declines as the number of nodes increases because the adapters become the bottleneck.

Key Findings:

- Host-based adapters achieve five times more messages per second at scale
- Adapter-based/offload adapter performance peaked at four cores
- Near-linear scaling was achieved with the host-based adapter

Cluster Message Rate Performance Test

This test is designed to show the message rate performance for an entire cluster with its end-to-end InfiniBand fabric from the adapters to InfiniBand switches. Like the host messaging test, end-to-end InfiniBand fabric performance is a key factor in application performance and the cluster’s ability to scale. This test extends the OSU MPI Message Rate test to run across a cluster, thereby testing the throughput performance of adapters and switches on a cluster. The test was first run unidirectional, with one set of nodes on one side of the fabric sending messages to nodes on the other side of the cluster. The second run was bidirectional, with nodes on both sides of the fabric simultaneously sending messages to one another.

This test utilized a cluster consisting of fourteen Intel-based servers, each with dual Xeon 5570 processors and 24GB of memory. The following cluster configuration was used for the test, and it was designed to show the ability

of the InfiniBand switches to queue/buffer and process messages, along with the host adapter’s ability to generate and drive messages through the fabric.

7 Hosts -> Switch <- Single ISL -> Switch <- 7 Hosts

The following InfiniBand equipment was used for the test:

- QLogic: 12300 QDR switches and QLE7340 QDR adapters
- Mellanox: MTS3600 QDR switches and ConnectX-2 QDR adapters

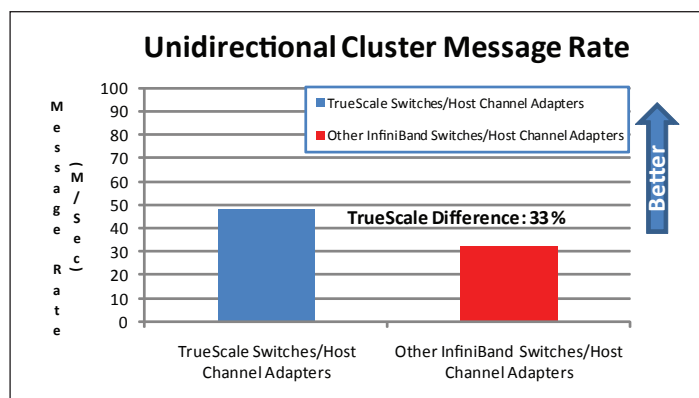


Figure 3. Unidirectional cluster message rate

The unidirectional cluster message rate test (Figure 3) shows that the end-to-end TrueScale fabric achieves almost 48M messages/second, while the Mellanox fabric provides only 32.1M messages. This means that the *TrueScale InfiniBand fabric provides 33 percent better message throughput* than the same end-to-end Mellanox-based fabric. The TrueScale adapter and switches will drive significantly more throughput, which provides better cluster performance, scalability, and application performance.

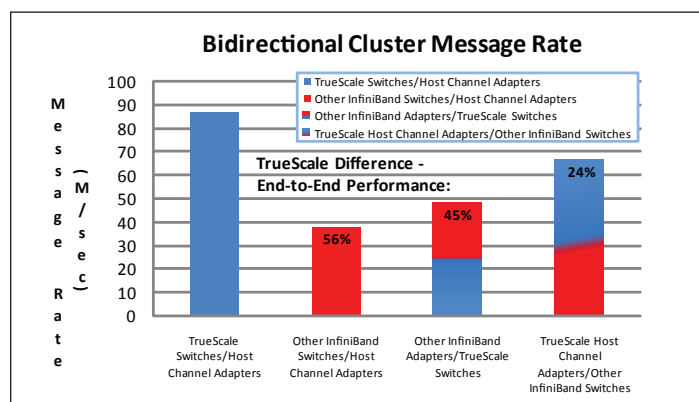


Figure 4. Bidirectional cluster message rate

The bidirectional cluster messaging test is more representative of how MPI applications generate traffic on an HPC cluster. In this test, the end-to-end TrueScale fabric was able to generate and process almost 87 million messages per second, which is an *81 percent increase over the unidirectional*

test. Mellanox generated almost 38 million messages per second, which is *only 18 percent better than the unidirectional test*. The end-to-end TrueScale fabric processed 56 percent more bidirectional messages/second than the Mellanox-based fabric. In fact, TrueScale adapters with a Mellanox switch is the second-best performing combination, and in this combination it performs 75 percent better than the all-Mellanox fabric. The reason that the TrueScale combination performs so much better than the all-Mellanox configuration is the TrueScale adapter’s superior ability to drive message throughput.

Key Findings:

- End-to-end TrueScale fabric provides more than 56 percent of additional throughput to the Mellanox-based fabric.
- The TrueScale-based fabric showed an 81 percent increase in message throughput going from unidirectional to bidirectional, whereas Mellanox only showed an 18 percent improvement.
- The second-best performing configuration is the TrueScale adapters with Mellanox switches. This combination provides 75 percent better performance than an all-Mellanox fabric. The reason for this result is the superior message handling capabilities of the TrueScale adapter architecture.

End-to-End Latency Performance Test

The HPC Challenge (HPCC) was used to test and determine the latency for the fabrics based on the two InfiniBand solutions. HPCC was designed to examine the performance of HPC architectures, including the interconnect. It is a more challenging test than other standard benchmarks, such as the High Performance Linpack (HPL). HPCC provides benchmarks that bound the performance of many real applications. Latency is one of the most important factors that will impact an MPI application’s performance and its ability to scale. The latency test used in this study determines end-to-end latency, which is a function of the InfiniBand adapter and the host InfiniBand stack and switch. The following tests were used to determine and analyze the performance of the InfiniBand architectures:

- Maximum Ping-Pong Latency – reports the maximum latency for a number of non-simultaneous ping-pong tests. The ping-pongs are performed between as many distinct pairs of processors as possible.
- Randomly Ordered Ring Latency – reports latency in the ring communication pattern. The communication processes are ordered randomly in the ring.
- Average Ping-Pong Latency – reports the average latency for a number of non-simultaneous ping-pong tests. The ping-pongs are performed between as many distinct pairs of processors as possible.
- Naturally Ordered Ring Latency – reports latency achieved in the ring communication pattern.

All of the tests use MPI standard send and receive routines.

The following table summarizes the configuration used for all the HPCC tests.

	Mellanox	QLogic TrueScale
System	Intel-based servers	Intel-based servers
CPU	Intel X5570	Intel X5570
Memory	24GB – DDR3	24GB – DDR3
Switch	Voltaire QDR 1:1	QLogic QDR 1:1
Host Channel Adapter	MLX ConnectX-2 QDR	QLogic TrueScale QDR
Compiler	Intel 11.1	Intel 11.1
Math Library	GotoBLAS 1.26	GotoBLAS 1.26
MPI	MVAPICH 1.2	MVAPICH 1.2
CPU settings	TURBO	TURBO
# ports used per node	Single rail	Single rail
	Intel X5570, 24GB – DDR3, MLX ConnectX-2 QDR, TURBO	Intel X5570, 24GB – DDR3, QLogic TrueScale QDR, TURBO

Figure 5 summarizes the results of the four HPCC latency tests. The fifth set of bars is an average of the four tests. In each of the tests, the QLogic TrueScale InfiniBand architecture achieved significantly better latency than its counterpart. The Randomly Ordered Ring Latency test showed the most performance difference; TrueScale-based fabric had a three-times latency advantage. The TrueScale average latency is almost two times faster than Mellanox.

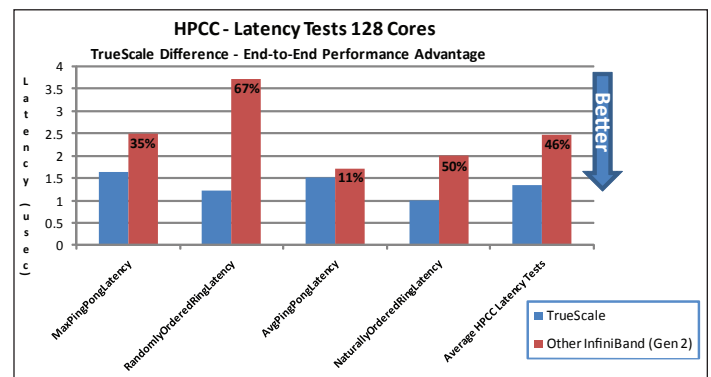


Figure 5. HPCC latency results at 128 cores

Key Findings:

- Latency is a key factor impacting the performance of most MPI applications
- QLogic provides almost two times better latency than the other major InfiniBand offering on the market

- TrueScale has 10 to 67 percent better latency depending on the test
- Average latency advantage for TrueScale is 46 percent

Application Test – Fast Fourier Transform

The HPCC benchmark has an MPI application-level test. This test is a good example to show how InfiniBand host message rate, cluster message rate, and latency can impact the performance of an MPI application. HPCC’s MPIFFT test measures the floating point rate of execution of the double precision complex one-dimensional Discrete Fourier Transform (DFT). The rating is in gigaflops or billions of floating point operations per second.

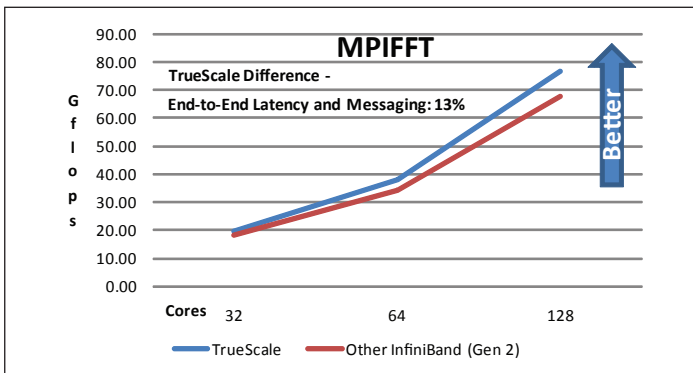


Figure 6. HPCC MPI Fast Fourier Transform results

The MPIFFT test results show that *TrueScale has a nine percent performance advantage at 32 cores and a 13 percent advantage at 128 cores.* TrueScale achieves almost nine Gflops more performance using the same Intel-based servers as when Mellanox InfiniBand is used. This test shows advantages in message rate, cluster message rate, and latency. Even though the TrueScale adapter uses the host CPU and memory, it still achieves better performance than the Mellanox offload.

Key Findings:

- The TrueScale-based cluster performed better across the range of tested cluster sizes
- At scale, TrueScale has a nine Gflop or 13 percent advantage (16 node/128 core)

Application Test – ANSYS FLUENT

The next application tested was ANSYS® FLUENT®, which is one of the most popular computational fluid dynamics applications in the market. FLUENT is an example of a commercial MPI application used in many industries, including aerospace, automotive, oil and gas, and medical, just to name a few. ANSYS has made strong strides in designing FLUENT to scale and perform well on HPC clusters.

The size of the CFD model has a direct relationship to the performance of the interconnect used. A CFD simulation is composed of a number of cells.

A “relatively” small CFD simulation will be broken up into a small number of cells per core. The cells per core are solved quickly for a specific step of the simulation. This then requires each node to send its information from the completed step to the other nodes. This means there is frequent MPI communications (that is, messages) between nodes. A smaller model best shows the performance of the interconnect and is a good indication of how a much larger model would run and scale on a larger cluster.

The FLUENT test was run on the Intel-based server cluster, which consists of 16 servers and one NFS server node. Each server has dual quad-core Intel Xeon 5570 “Nehalem” 2.93GHz processors and 24GB of memory. The Intel-based server cluster has a total of 128 cores and 384GB of memory. The following InfiniBand was tested:

- QLogic: 12300 QDR switches, QLE7340 QDR Host Channel Adapters
- Mellanox: MTS3600 QDR switches, ConnectX-2 QDR Host Channel Adapters

Eddy 417K Model	
Number of cells	417,000
Cell type	Hexahedral
Models	k-eps turbulence
Solver	Segregated implicit

The FLUENT Eddy 417K model was selected for testing, since this model stresses the CPU and communications of the cluster. This is especially true on a cluster size of 16 nodes. It produces significant volumes of small messages that grow exponentially as the cluster scales. The FLUENT Eddy model is also highly latency sensitive. Finally, FLUENT is a processor-intensive application. This means that if an InfiniBand architecture, such as a host based, on-load adapter, requires too many processing cycles, then the performance of FLUENT would be affected. The results from the test show just the opposite.

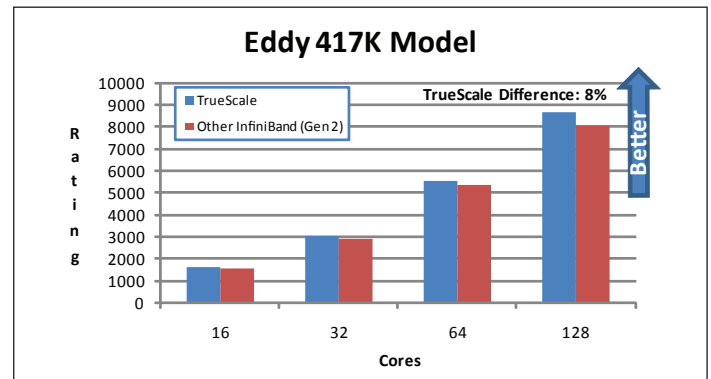


Figure 7. ANSYS FLUENT 12.1 – Eddy 417K Results

The TrueScale host-based/on-load architecture scales well with the FLUENT applications, especially on models, such as Eddy 417K, that tax the cluster fabric. In summary, *TrueScale shows a 3.6 percent advantage at 16 cores and grows to an 8 percent difference at 128 cores.*

Turbo 500K Model	
Number of cells	500,000
Cell type	Mixed
Models	Spallart-Allmaras turbulence
Solver	Coupled implicit

A second FLUENT test was performed using the Turbo 500K model. This test showed an even more pronounced performance difference at scale. The Turbo 500K model, like the Eddy test, stresses the processing and communications infrastructure of the cluster. The test results showed that the *TrueScale-based cluster had a 22 percent performance advantage over the other cluster.*

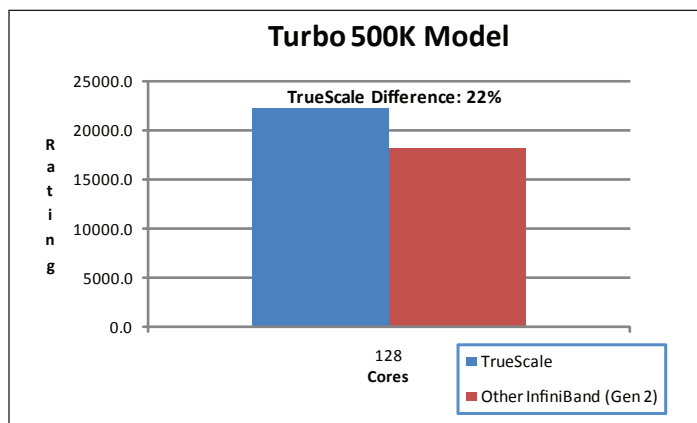


Figure 8. ANSYS FLUENT 12.1 – Turbo 500K Results

Key Findings:

- The smaller FLUENT models best show the performance capability of the interconnect. The reason is that these types of models are broken into smaller numbers of cells per core, which requires extensive MPI communication between cores and nodes to complete the simulation.
- FLUENT is both a latency-sensitive and CPU/processor-intensive application, which places stress on an interconnect, especially at scale.
- TrueScale has a performance advantage ranging from 3.6 to 22 percent, depending on the FLUENT model.

Conclusion

The InfiniBand architecture does make a difference in the performance of the cluster and in applications. QLogic TrueScale InfiniBand host and switch technologies provide the interconnect infrastructure to maximize an HPC cluster’s overall performance. The TrueScale host-based (on-load) protocol processing engine, with its lightweight, semantic-based interface, provides an efficient architecture that maximizes MPI application performance. Combine the TrueScale host performance with the advanced messaging handling of the TrueScale switch, and you have the ultimate combination for optimizing a cluster’s performance. With the use and size of HPC clusters expanding at a rapid pace, TrueScale InfiniBand architecture and technology extracts the most out of your investment in compute resources by eliminating adapter and switch bottlenecks.

Disclaimer

Reasonable efforts have been made to ensure the validity and accuracy of these performance tests. QLogic Corporation is not liable for any error in this published white paper or the results thereof. Variation in results may be a result of change in configuration or in the environment. QLogic specifically disclaims any warranty, expressed or implied, relating to the test results and their accuracy, analysis, completeness or quality.



Corporate Headquarters QLogic Corporation 26650 Aliso Viejo Parkway Aliso Viejo, CA 92656 949-389-6000

www.qlogic.com

International Offices UK | Ireland | Germany | France | India | Japan | China | Hong Kong | Singapore | Taiwan

© 2010–2011 QLogic Corporation. Specifications are subject to change without notice. All rights reserved worldwide. QLogic and the QLogic logo are registered trademarks of QLogic Corporation. TrueScale is a trademark of QLogic Corporation. AMD is a registered trademark of Apple, Inc. ANSYS is a registered trademark of ANSYS, Inc. ConnectX is a registered trademark of Mellanox Corporation. FLUENT is a registered trademark of ANSYS, Inc. InfiniBand is a registered trademark and service mark of the InfiniBand Trade Association. Intel is a registered trademark of Intel Corporation. Mellanox is a registered trademark of Mellanox Technologies. Myricom is a registered trademark of Myricom, Inc. Quadrics is a registered trademark of Quadrics Ltd. Voltaire is a registered trademark of Voltaire, Inc. Xeon is a registered trademark of Intel Corporation. All other brand and product names are trademarks or registered trademarks of their respective owners. Information supplied by QLogic Corporation is believed to be accurate and reliable. QLogic Corporation assumes no responsibility for any errors in this brochure. QLogic Corporation reserves the right, without notice, to make changes in product design or specifications.