

Beyond Hero Numbers: Factors Affecting Interconnect Performance

By Lloyd Dickman, CTO, QLogic System Interconnect Group (formerly PathScale)

Executive Summary

As users tackle larger and more complex applications with ever-larger clusters, there is a marked increase in message rates and decrease in message sizes. The more processors that tackle a given task, the higher the frequency of small, bursty message communications. Interconnect performance has become a key factor in overall application and system performance and this will increasingly be the case as applications become more complex, more sophisticated, and are scaled onto larger and larger clusters to accelerate time-to-solution.

Latency and bandwidth are the key metrics that traditionally define interconnect performance. However, these two “hero numbers” alone are insufficient to determine effective interconnect performance for real applications.

Interconnect vendors show benchmarks that highlight interconnect performance for latency using zero-byte message sizes (no data payload), and peak bandwidth (large data payload). However, these “hero” numbers alone provide a limited perspective of real application performance since real applications tend to send messages in the tens to a few thousand bytes.

An additional measurement of interconnect performance is the half-power point ($n^{1/2}$ message size). This is the message size at which $1/2$ of the peak bandwidth is achieved. It provides a meaningful indication of interconnect performance for messages sizes used in a wide range of real applications.

Previously, the “sweet spot” for cluster nodes was two CPUs per node. With the availability of dual core processors, the “sweet spot” has moved up to four CPUs per node. In the near future, the number of processors per node is expected to increase further. The QLogic™ InfiniPath™ adapter was designed to support and capitalize on this industry trend, and as such, provides exceptional scaling with both node count and the number of processor sockets and cores per node.

2 | Beyond Hero Numbers – Factors Affecting Interconnect Performance

In summary:

- The two traditional “hero” numbers (zero-length message latency and peak bandwidth) do not adequately characterize real application performance on clusters.
- $n^{1/2}$ message size is an effective metric to add to the small set of interconnect comparative measures and an important indication of real application performance.
- SMP effects on interconnect performance, due to the proliferation of multi-socket / multi-core nodes, is a critical issue for cluster users to consider.

1. Traditional Hero Numbers Do Not Tell the Full Story

Traditional “hero” numbers characterize interconnect performance of point-to-point communications for two cases. One-way latency is measured for a message with a zero-length data payload in order to understand the basic communications latency. Then, peak bandwidth is measured with large sized data payloads to understand overall data transport capacity.

By studying industry marketing materials, users find that interconnect vendors (including QLogic) highlight results for these two basic metrics. However, real applications have communications messaging traffic that involve a variety of communication patterns different from those used by the micro-benchmarks. Assessing interconnects by only using these two basic metrics involves several implicit assumptions, including restricted communication patterns that are not representative of application performance. Here are some of those assumptions and why they are not representative:

3 Beyond Hero Numbers – Factors Affecting Interconnect Performance

Assumption

- *Communications is point-to-point between a single pair of nodes*
- *There is a single communicating process per node*
- *Latency for zero-length messages characterizes performance of short to medium messages.*
- *Peak bandwidth characterizes performance of medium to large messages.*
- *The receiver is always waiting for incoming messages.*

Why Not Reflective of Real Applications

- Applications consist of dozens to thousands of nodes simultaneously communicating
- Most compute nodes consist of 2 or more sockets. With dual-core processors, it will soon be typical that compute nodes will have 4 or more CPUs all contending for access to the interconnect fabric.
- Applications exchange messages with varying amounts of data, not empty messages.
- Application message sizes are significantly smaller than needed to achieve peak bandwidth on most commercially available interconnects.
- Hero numbers are determined by benchmarks where the receiver is always waiting for an incoming message. However, real applications use expected receives for only 10-50% of all messages. The performance of the 50-90% of messages that are not expected is not assessed by the hero numbers.

To better understand how a given interconnect solution will perform, it is necessary to have deeper insight into how cluster interconnect performance affects application performance:

- Understand how much bandwidth is usable by applications.
- Compare ideal latency characteristics versus actual latency curves.
- Understand how latency budgets are divided between host software and interconnection hardware devices.
- Include the effects of communication traffic with multi-socket and multi-core processors.

4 Beyond Hero Numbers – Factors Affecting Interconnect Performance

2. Performance Scorecard

The comparative MPI latency measurements from published benchmarks, vendor collateral, and QLogic's InfiniPath adapter measurements are shown in Figure 1. For MPI, both latency and bandwidth are traditionally quoted. This chart also shows the $n\frac{1}{2}$ message size since it plays such an important role in understanding real application performance. In addition, many applications and environments extensively use TCP for communications.

		InfiniBand		Proprietary		10 GbE
		PathScale InfiniPath	Mellanox Lx-Ex	Quadrics Elan 4	Myricom	Chelsio
MPI	Latency (μ s)	1.29 *	2.7 ~ 2.84* ~ 3.85*	1.29 ~ 2.0*	2 ~ 2.6 ~ 3.2 ~ 5.5*	10.2
	Bandwidth (MB/s) Uni-directional Bi-directional (streaming)	953 * 1,869 *	981* ~ 1,474* 1,670* ~ 2,645*	875 ~ 910* 901*	247 ~ 493* ~ 1,200 749*	862
	$n\frac{1}{2}$ Message Size (Bytes)	385	>1K	1K	2K	100K~300K
	Messaging Rate (Messages/s)	11.3M	450K~900K	125K	560K	150K
TCP	Latency (μ s)	6.7	23 ~ 37	24	18	8.9
	Bandwidth (MB/s)	583	316 ~ 425	482	247	950

MPI Sources/Notes:

- QLogic – QLogic measurements with one switch crossing, May 2005. MPI measurements by DK Panda, 8sep2005, 2.8 GHz Opteron.
- * Ohio State measurement results – DK Panda, Feb–Sep 2005
- Mellanox – 20jun2005 announcement
- Quadrics – IEEE Micro, to appear 2005, 22jun2005 announcement
- Myricom – Myricom website 4aug2005
- Benchmarks - OSU MPI Benchmarks 2.0 (streaming)

TCP Sources/Notes:

- QLogic – QLogic measurements with one switch crossing, May 2005
- Quadrics – Quadrics website
- Myricom – Myricom website, 19sep2005, F-card, 2G-MX
- Mellanox – QLogic measurements, 2sep2005
- Chelsio – Chelsio website and Wu Feng, Hot Interconnects 2005.
- Benchmark – netperf 2.3, one-way latency

Figure 1. Performance scorecard for several cluster interconnect solutions

3. Shape of the Bandwidth Curve – $n^{1/2}$ Message Size (Half-power Point)

A traditional way to understand the effectiveness of an interconnect on application performance is to measure performance with various message sizes and traffic patterns. One metric that captures interconnect performance at intermediate message sizes is the half-power point, or $n^{1/2}$ message size. $n^{1/2}$ is the message size on the bandwidth curve where half of the peak bandwidth is achieved. It is an excellent metric of interconnect performance for real applications that typically use small to medium sized messages.

A typical bandwidth curve is shown in Figure 2a. The half-power point is the message size for which half of the peak bandwidth is achieved.

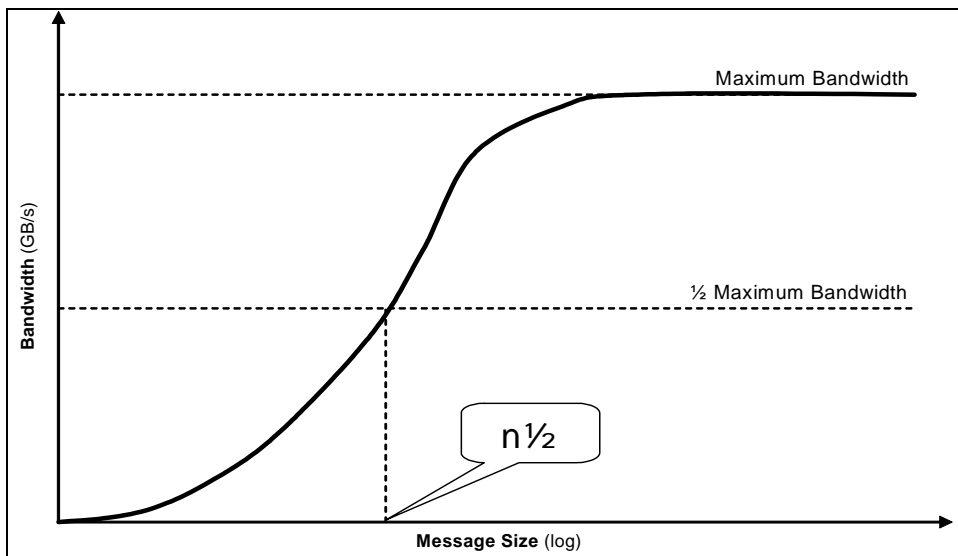


Figure 2a. The $n^{1/2}$ power point is the message size at which half of the maximum bandwidth is achieved. Smaller is better.

Figure 2b shows bandwidth curves for several interconnects, each of which exhibits identical zero-message latencies and peak bandwidths. The bandwidth curve for Interconnect #1 rises more rapidly than the others as message size increases. This is highly desirable since more of the potential link utilization is achieved for smaller message sizes.

6 Beyond Hero Numbers – Factors Affecting Interconnect Performance

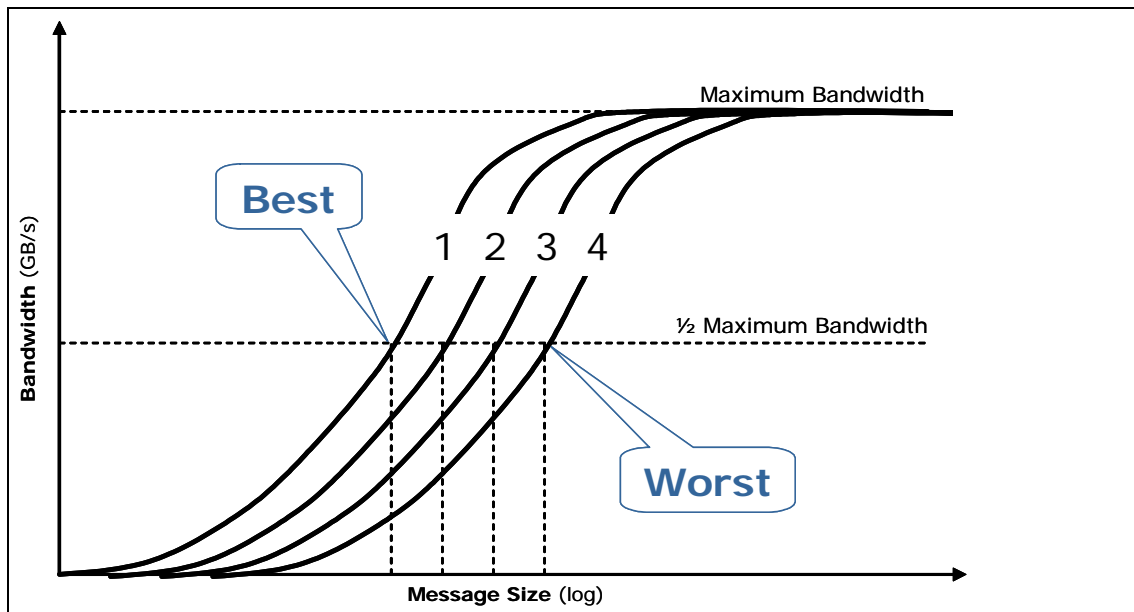


Figure 2b. Four interconnects with the same “hero” numbers of latency and peak bandwidth. Interconnect #1 rises fastest with message size, delivering superior bandwidth for small to medium messages.

The $n^{1/2}$ message size is also useful when you know the messaging patterns used by your applications. Figure 3 illustrates the bandwidth curves from Figure 2 overlaid with a typical application’s bi-modal messaging pattern – short-length messages for coordination, and medium-length messages for exchanging results among nodes.

Figure 3 illustrates that even modest changes in $n^{1/2}$ message size have significant implications on the ability of the application to benefit from the interconnect’s bandwidth. For Interconnect #1, a substantial portion of the interconnect’s bandwidth is realized by the application. If the application were to use Interconnects #2 ~ #4, less and less of the bandwidth potential is realized. Interconnects with lower $n^{1/2}$ message size provide greater effective bandwidth to the application and accelerate time-to-solution.

7 Beyond Hero Numbers – Factors Affecting Interconnect Performance

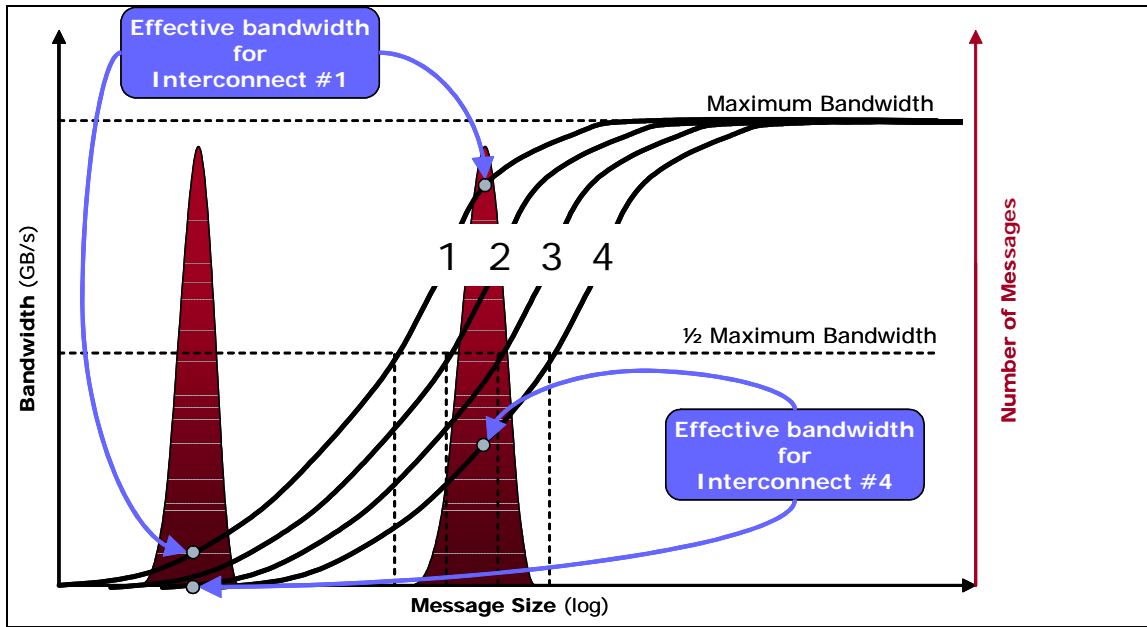


Figure 3. Typical bi-modal message pattern illustrates how $n^{1/2}$ message size is a valuable metric in predicting real application performance.

8 Beyond Hero Numbers – Factors Affecting Interconnect Performance

Figure 4 shows the streaming bandwidth micro-benchmarks results obtained at Ohio State University (OSU), courtesy of DK Panda, September 26, 2005.

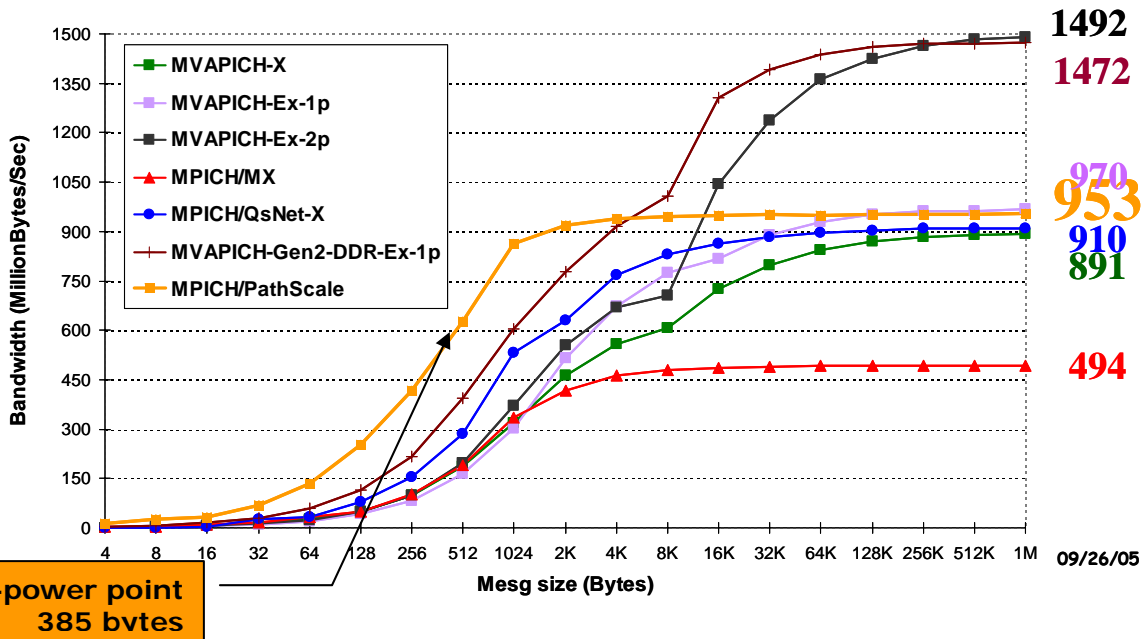


Figure 4. MPI unidirectional streaming bandwidth results. QLogic InfiniPath's bandwidth curve rises significantly more rapidly than other interconnects, delivering superior application performance.

The QLogic InfiniPath bandwidth curve rises faster than any other interconnect. The InfiniPath adapter's significantly lower $n\frac{1}{2}$ message size means that it can deliver higher bandwidth at smaller message sizes to real applications. Using the QLogic InfiniPath adapter, half of the maximum bandwidth is delivered for messages as small as 385 bytes; over 90% of the peak bandwidth is delivered with message sizes as small as 1 Kbyte.

4. Shape of the Latency Curve – Digging Deeper into MPI Latency

Figure 5 shows the latency micro-benchmarks results obtained at Ohio State University (OSU), courtesy of DK Panda, September 26, 2005.

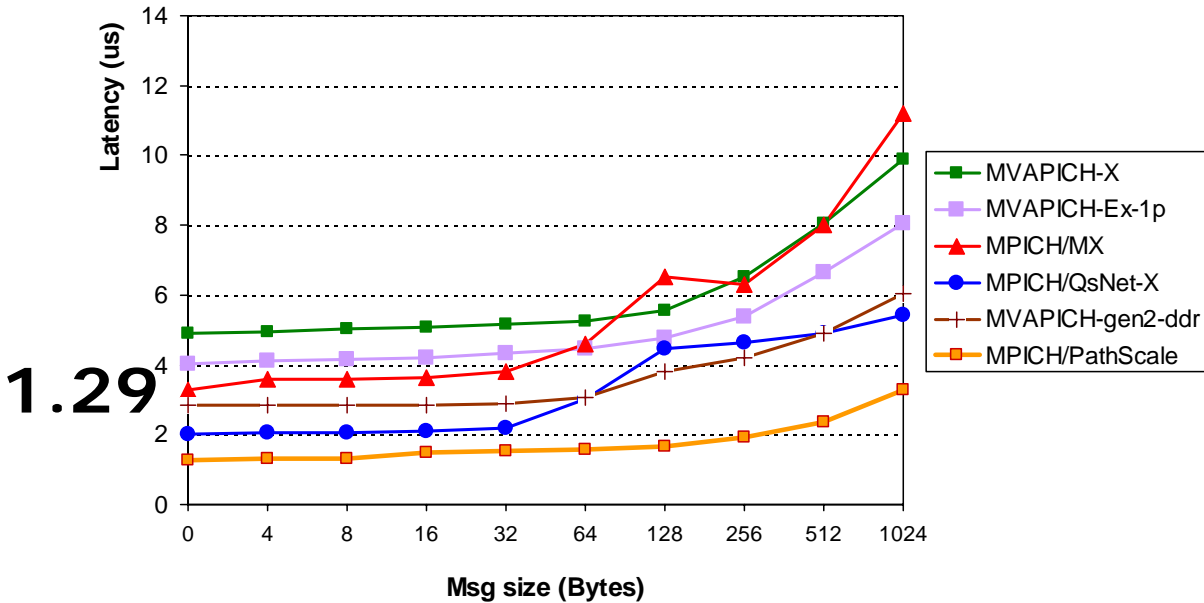


Figure 5. MPI small message latency for several interconnect solutions. QLogic delivers the lowest latency at all message sizes.

The QLogic InfiniPath adapter delivers an ultra-low latency of 1.29 μs. For 256 byte messages, its latency is below that of the next nearest interconnect’s latency for zero-byte messages. This has enormous performance implications for applications and middleware that exhibit significant chatty small message exchanges, both for scientific / technical computing, and in the data center. Reduced message latencies improve response time, shorten wait queues, and improve system resource utilization.

While the latency for zero-length messages is an important metric, users need to understand how latency changes across a spectrum of message sizes. We look next at the difference between a measured latency curve such as shown in Figure 5, and the predicted ideal latency curve represented by the following equation:

$$ideal\ latency_{size} = t_{size=0} + \frac{size}{\beta_{size=\infty}}$$

10 Beyond Hero Numbers – Factors Affecting Interconnect Performance

This ideal latency equation suggests that latency for an arbitrary-sized data payload is a function consisting of the two “hero” metrics — basic zero-payload latency, and the peak bandwidth achieved by transferring a maximum-sized data payload. It is not surprising to find that the measured latencies differ from the “ideal” latencies as suggested by the equation above. The measured and ideal latency curves for QLogic InfiniPath are shown in Figure 6.

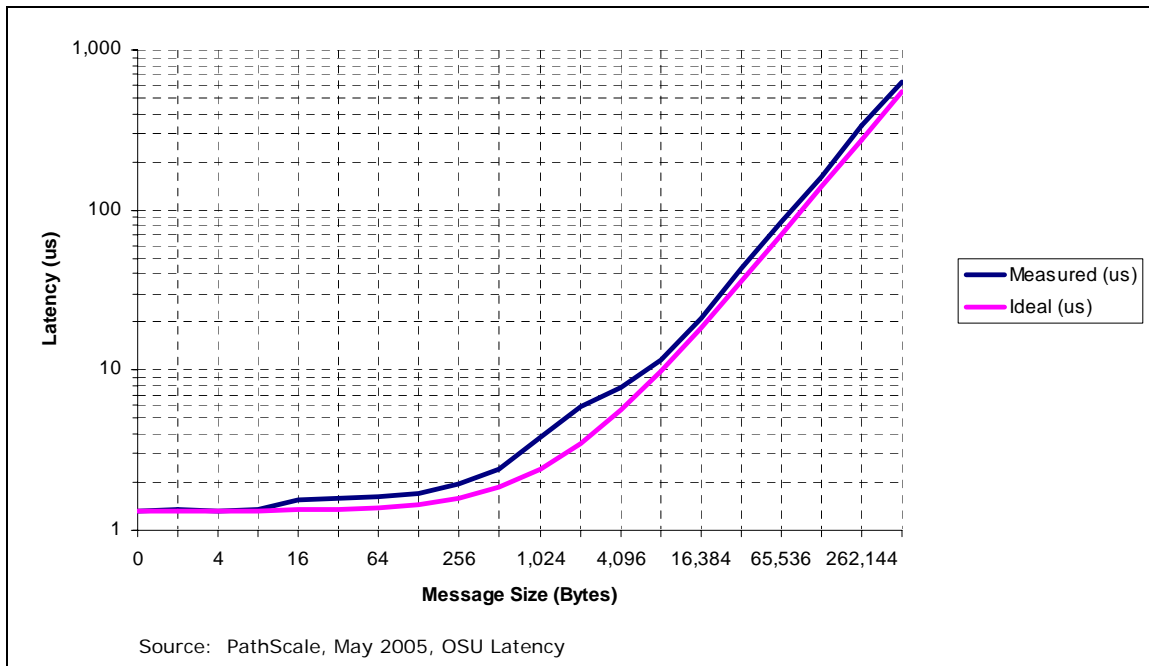


Figure 6. Latency versus message size

11 Beyond Hero Numbers – Factors Affecting Interconnect Performance

Figure 7 shows the latency “surprises” as the ratio of the measured to ideal latency curves for the interconnect solutions Figure 5 across a range of message sizes. Data is courtesy of DK Panda at Ohio State, September 26, 2005.

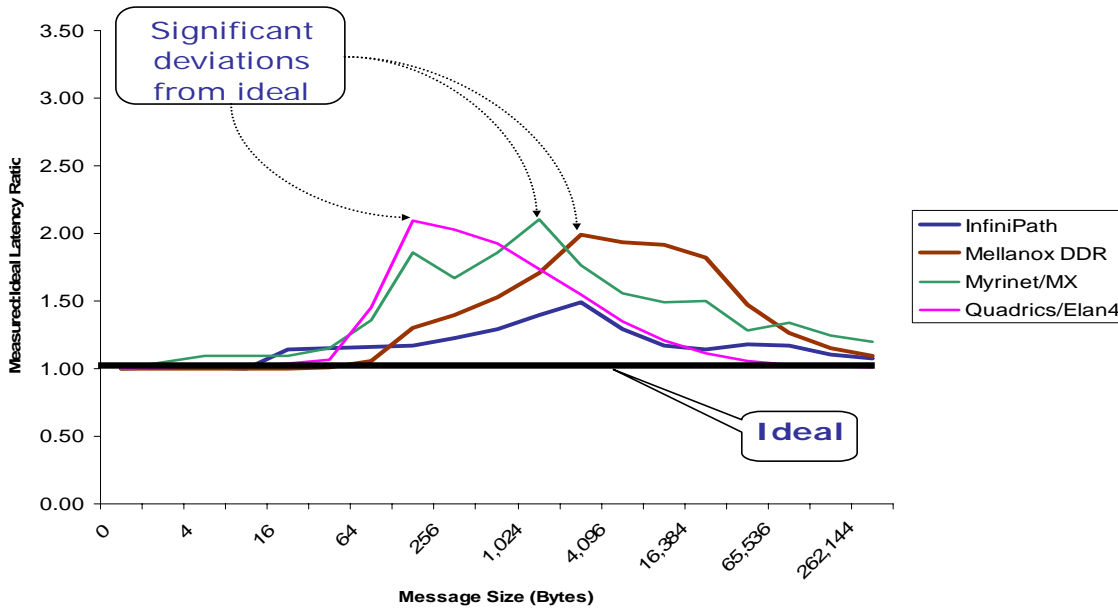


Figure 7. Latency “surprises”. Ratio of Measured to Ideal latency versus message size for several interconnect solutions. Implementation factors in the interconnect cause the measured latency to deviate from the ideal model. Flattest is best. Minimize area under the curve.

Note that the ratio of measured to ideal latency can be substantial. While all interconnect solutions in Figure 7 match their ideal latency curves for extremely short message sizes (represented by the zero-length message latency), and for very large messages (which approach the peak bandwidth), message sizes in the area of interest to most applications deviate, sometimes substantially, from the ideal curve. This is due to constraints and artifacts that reflect specific design decisions.

It is desirable that the latency ratio be as flat (close to 1) as possible, meaning that the “hero” metrics accurately represent measured performance across a wide range of message sizes. This would result in few latency surprises for message sizes of interest to real applications. The QLogic InfiniPath adapter is superior to the other available interconnects in meeting this objective. QLogic delivers the most consistent performance

12 Beyond Hero Numbers – Factors Affecting Interconnect Performance

across the full range of message sizes, including the small to medium message sizes of interest to real applications.

5. Composition of the Latency Budget Affects Performance

Interconnect vendors make design decisions affecting all aspects of performance. One of the key design decisions is how to partition functions across host software and the interconnect adapter.

The comparative MPI latency budgets of several commercially available interconnect solutions is shown in Figure 8. For zero-byte messages, it shows the time spent in host software (message send and receive), in adapter related operations (message send and receive), and the network switching fabric (single level of switch).

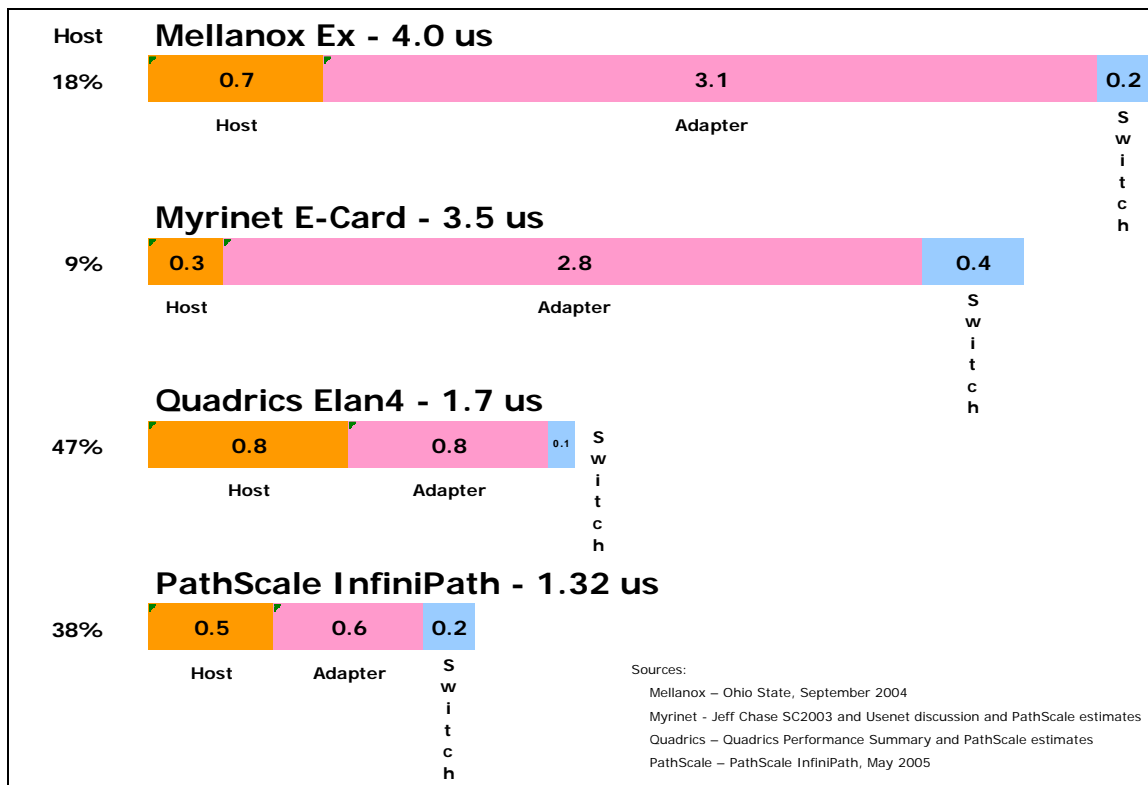


Figure 8. Competitive MPI latency budgets. It is desirable to have a significant portion of the latency budget in host software.

The QLogic InfiniPath adapter exhibits low host processor utilization because there is a highly efficient interface between host software and the adapter. The 38% of the QLogic latency budget that runs in the sending and receiving hosts will be further reduced as host processors increase in speed. This is not true for interconnects where little of the latency budget is in host software.

6. Performance with Multi-Core Processors

The architecture of the interconnect and how it partitions its latency budget affects performance as the number of processing cores per node increases. If the bulk of the protocol processing is in the interconnect link hardware, then node traffic suffers as more processors are added. Recalling Amdahl's Law, system performance is limited by the sequential component. Interconnects with a significant portion of their latency in the adapter are expected to scale poorly on current and future multi-socket / multi-core nodes.

Because the QLogic InfiniPath adapter places more of the latency budget in the host processor, parallelism is achieved and performance scales as more processors are added. This approach places some protocol processing in the host processor. However, the host processor is the fastest and most cost-effective computing resource. Since typical communication patterns often leave the host processor idle, this is a wise choice.

Because the QLogic InfiniPath adapter chose to place a significant portion of the latency budget in the host processor, its performance improves naturally as

- processors get faster, and
- the number of processing cores per node increases.

This will not be true for adapters that partition most of the adapter function into the adapter. Such devices may bottleneck as processor speeds and processor cores increase.

Figure 9 shows how the QLogic InfiniPath interconnect scales with multiple processor cores. The shape of the bandwidth curve improves with additional processor cores because the QLogic InfiniPath adapter exploits multi-core parallelism. An important implication of this is a drastic reduction in $n^{1/2}$ message size from 385 bytes with one processor core per adapter, to a remarkable 95 bytes using 4 processor cores in a node.

Measurements were made with a modified version of OSU Bandwidth 2.0 that used 1, 2 and 4 MPI process pairs communicating across two dual-socket 2.2 GHz Opteron nodes. Note that the single processor half-power point is 415 bytes @ 2.2 GHz, rather than 385 bytes @ 2.6 GHz. This is another example of how the QLogic InfiniPath adapter benefits from improvements in host processor performance.

14 Beyond Hero Numbers – Factors Affecting Interconnect Performance

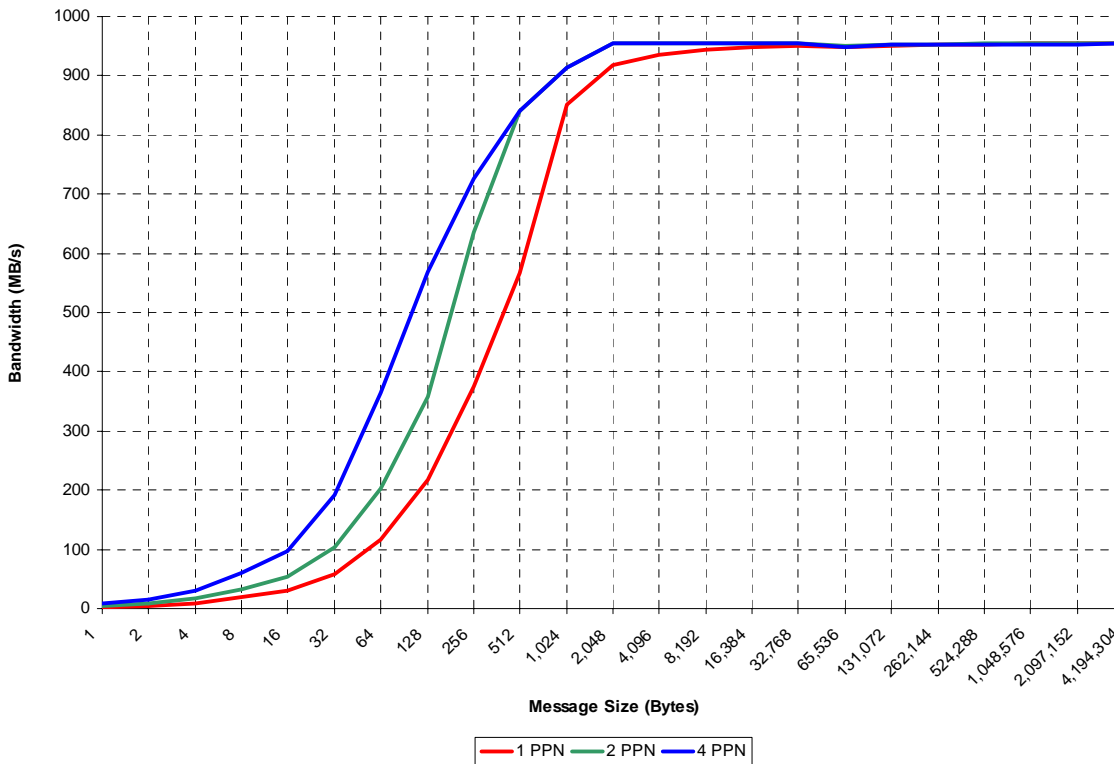


Figure 9. Effect of exploiting multi-core parallelism dramatically improves InfiniPath’s bandwidth and dramatically reduces the $n^{1/2}$ message size. $n^{1/2}$ moves down to 95 bytes on a four core compute node.

How these factors influence applications is shown in Figure 10. One of the HPC Challenge Benchmarks measures traditional MPI latency as a one-way latency between processors. In addition, it measures the inter-processor latency in a dynamically changing logical “ring” which connects all processors in the cluster.

Most tested interconnects fall below the black line; that is, their random ring latency across many processors does not keep up with the “hero” latency as measured between a single pair of processors. Some interconnects miss by a significant amount. The QLogic InfiniPath adapter is not only lowest in “hero” latency, its ultra-low latency is maintained in a random ring when all processors are communicating.

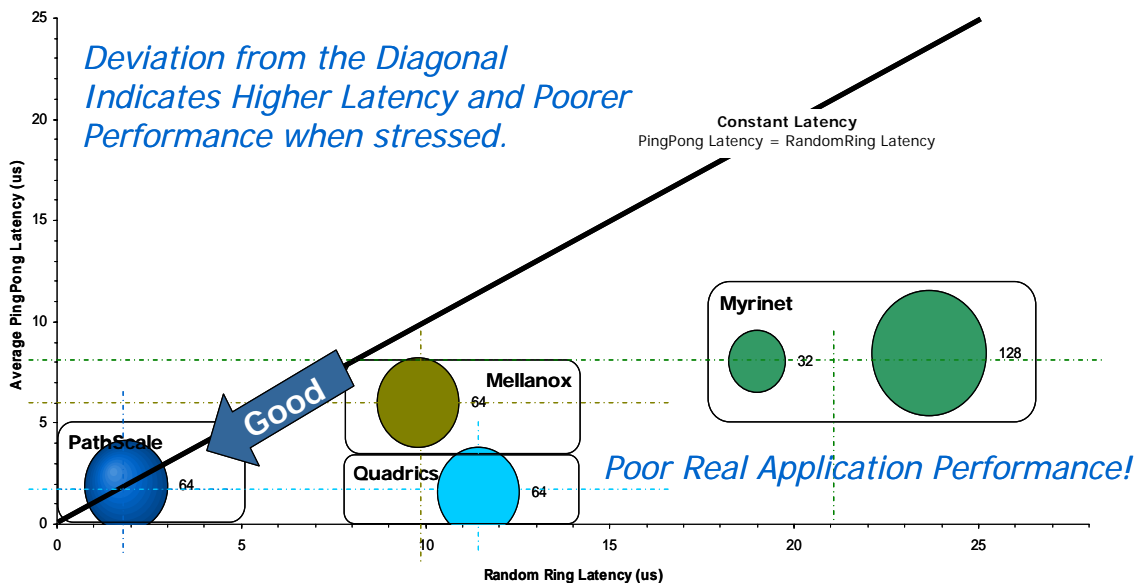


Figure 10. HPC Benchmark results for point-to-point and random ring latency. Good hero numbers do not always yield strong application performance when all CPUs are in use. Size of each bubble and the number adjacent to it are the number of CPUs in the cluster. Best position is close to the origin, on the black line!

7. Conclusion

When estimating interconnect performance for your application, it is necessary to go beyond the hero numbers of best-case latency and peak bandwidth. Delving deeper into interconnect metrics such as the $n/2$ message size (half-power point), latency across a spectrum of message sizes, latency budget partitioning, and message parallelization gives insight into real application performance. The QLogic InfiniPath adapter exhibits ultra-low messaging latency, an industry leading low $n/2$ message size, high bandwidth, and scalability with increasing processor / multi-core performance – all on a standard and cost-effective 10 Gbps switching fabric.

For More Information

Additional information on QLogic InfiniBand products can be obtained by visiting <http://www.qlogic.com/QLogic> on the World Wide Web, or by contacting:



QLogic, inc.
System Interconnect Group Phone: (650)-934-8100
2071 Stierlin Court, Suite 200 Fax: (650) 428-1969
Mountain View, CA. 94043 USA www.qlogic.com/QLogic

Copyright 2005 QLogic, Incorporated.

QLogic, QLogic and InfiniPath are trademarks of QLogic, Incorporated.
AMD and AMD Opteron are trademarks of Advanced Micro Devices, Incorporated.
All other trademarks and registered trademarks are the property of their respective owners.