USENIX Association

Proceedings of the FREENIX Track: 2004 USENIX Annual Technical Conference

Boston, MA, USA
June 27–July 2, 2004

© 2004 by The USENIX Association All Rights Reserved For more information about the USENIX Association:
Phone: 1 510 528 8649 FAX: 1 510 548 5738 Email: office@usenix.org WWW: http://www.usenix.org
Rights to individual papers remain with the author or the author's employer.
Permission is granted for noncommercial reproduction of the work for educational or research purposes.
This copyright notice must be included in the reproduced paper. USENIX acknowledges all trademarks herein.

Cluster Interconnect Overview

Brett M. Bode, Jason J. Hill, and Troy R. Benjegerdes
Scalable Computing Laboratory
Ames Laboratory
Ames, Iowa 50011

Abstract

Today cluster computers are more commonplace than ever and there are a variety of choices for the interconnect. The right choice for a particular installation will depend on a variety of factors including price, raw performance, scalability, etc. This paper will present an overview of the popular network technologies available today including Gigabit Ethernet, 10 Gigabit Ethernet, Myrinet, SCI, Quadrics, and InfiniBand. Where possible a comparison will be included for multiple vendors of a given technology. Included will be comparisons of cost and performance of each along with suggestions for when each might present the best choice for a cluster installation.

1. Introduction

Over the past few years the cost and performance of interconnects has progressed to the point where today most new clusters use a primary interconnect of 1 – 10 Gbps. There are several interconnection choices in this performance range that range in cost, latency and achievable bandwidth. Choosing the correct one for a particular application is an important and often expensive decision. This paper will present a direct comparison of Gigabit Ethernet, 10 Gigabit Ethernet, Myrinet, SCI, Quadrics and InfiniBand. For each of the network technologies we will examine issues of cost, performance (latency and bandwidth), and scalability.

Some of the network interconnects in this review have been around for quite some time such as Gigabit Ethernet and Myrinet. Others such as InfiniBand and 10 Gigabit Ethernet are quite new. In addition even well known technologies such as Myrinet are evolving in terms of both hardware and software implementations. For example Myricom now offers both single and dual port NICs and is in the process of finishing a substantial rewrite of their software stack.

Finally we will examine application behavior with each of the interconnect technologies. Even the highest performing network will be of little use if problems in the software stack cause scalability problems for real applications.

2. Point to Point Performance

The most important parameters for an interconnect are the latency and bandwidth that an application would experience. To measure these parameters we have used the NetPipe[1] program running between two Dell 2650's with dual 2.4Ghz Pentium 4 CPUs with the NICs installed in the PCIX 133Mhz slot. NetPipe functions as a normal user application using either MPI or TCP interfaces. NetPipe can also be used directly with low level interfaces, but since that is less relevant to applications we will not consider it in this work.

NetPipe measures performance versus message size over an exponentially increasing message size. This data is then plotted on a semi-log plot of bandwidth versus message size. There are several common traits to these graphs. First off the bandwidth achieved for small messages is very low for any type of interconnect since it is latency dominated. In addition most interconnects only achieve peak performance for messages over 128KB, some not until messages over 1MB. Finally message sizes in the range of 10KB - 1 MB tend to be the most relevant to applications.

2.1. Gigabit Ethernet

Ethernet of some sort has always been used in cluster computers and even today it is present in almost all clusters. While Fast Ethernet was often used as the high-speed interconnect in years past, it has been relegated to the service network today in most systems. The primary limiting factor with Ethernet in the cluster world has always been the switch. Since the underlying architecture of Ethernet requires smart switches which shoulder the full burden of routing packets Ethernet switches must maintain full routing tables and be capable of making route computations on the fly at wire speed. In addition wider market pressures have often led to switches including extra features such as layer 3 and

higher based routing that are unneeded in the cluster world.

Gigabit Ethernet has been used as a high speed interconnect for a number of years. However, early on its performance in PC systems was limited to 300-500 Mbps. In addition Gigabit Ethernet switches were expensive and of limited port count. With the arrival of copper based NICs and inexpensive switches it became the low-cost solution for clusters up to 24 nodes. Beyond that size switches remained very expensive until very recently. Today, driven partly by the arrival of 10 Gigabit Ethernet, high-density Gigabit Ethernet switches are an affordable solution through around 480 nodes.

One significant issue with Ethernet has always been the relatively high CPU overhead of a full TCP/IP stack. This issue greatly limited performance on early systems. One common, but non-standard technique is to increase the Maximum Transmittable Unit (MTU). Standard Ethernet has always specified an MTU of 1500 bytes regardless of the speed. This creates a large overhead associated with packetization that can greatly impact performance. It has become fairly common to increase the MTU to 9000 bytes (aka Jumbo Frames), which has the effect of reduces the packetization overhead by a factor of six. However, all of the NICs and switches in the system must support the larger MTU size. Today CPUs and system busses have progressed to the point where most systems can achieve 90% or better utilization even using the standard MTU.

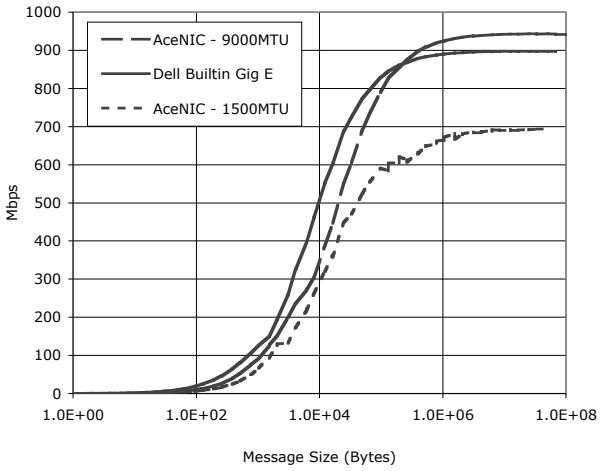

Figure 1: Gigabit Ethernet Performance

Figure 1 illustrates the performance of two Gigabit Ethernet NICs. The first and older NIC is based on the Alteon ACENIC chipset. The ACENIC provides a flexible CPU based architecture capable of jumbo frame support. Unfortunately the CPU based architecture has relatively high latency at around 90 μseconds. Partly due to the high latency the Alteon NIC requires the use of Jumbo Frames to achieve good results, but went Jumbo Frames are used performance around 940Mbps is seen. The second NIC is a built-in NIC on the Dell 2650 motherboard based on a Broadcom ASIC based chipset. The Broadcom chipset has become quite common and due to its low cost is the basis for many low cost as well as integrated Gigabit Ethernet solutions. As Figure 1 shows the Broadcom part performs quite well with a latency of 31 μseconds, which is excellent for any Ethernet NIC. In addition the lower latency translates into significantly better performance in the 10KB region. Finally the peak performance of 900 Mbps is quite good especially since it was achieved with the standard 1500 byte MTU.

A second issue is the inability to aggregate multiple switches in such a way as to provide full bi-section bandwidth. While most switches do provide the ability to trunk multiple ports together between switches this is both expensive and generally inadequate effectively limiting the size of an Ethernet based cluster to the size of the largest available switch, which is currently around 480 ports.

The Cost of Gigabit Ethernet has come down substantially over the last few years. While the Alteon, etc CPU based NICs are still relatively pricey at $300/NIC, the ASIC based NICs such as the Broadcom based products are widely available at less than $100/NIC, if they aren't integrated onto the motherboard. For small clusters there have been inexpensive (<$100/port) switches available for some time. However the price of the larger switches has now dropped substantially as well. For moderate size clusters (64-128 ports) per port pricing runs $200-$300/port while high-end switches run around $667/port. This puts the total cost of a Gigabit Ethernet solution at $150/port at the low end up to around $750/port at the high end which certainly makes it the lowest cost option considered in this paper.

2.2. Myrinet[2]

Myrinet was one of the first interconnect technologies designed specifically with clustering in mind. Because of this design criterion it makes some tradeoffs not possible in more general-purpose technologies such as Ethernet. The main feature of the design is that packets are source routed over a fat-tree based network made up of relatively small switch elements (16 port switches). This obviously requires that each node know the full network topology or map and that it be fairly static.

The big payoff is that the switch elements can be very simple since they do not perform any routing calculations. In addition the software design is based on an OS-bypass like interface to provide low-latency and low-CPU overhead.

Current Myrinet hardware utilizes a 2 Gbps link speed with PCI-X based NICs providing one or two optical links. The dual-port NIC virtualizes the two ports to provide an effective 4 Gbps channel. The downside to the two port NIC is cost. Both the cost of the NIC and the extra switch port it requires. Myrinet is designed to be scalable. Until recently the limit has been 1024 nodes, but that limit has been removed in the latest software. However, there are reports that the network mapping is not yet working reliably with more than 1024 nodes. The cost of Myrinet runs around $850/node up to 128 nodes, beyond that a second layer of switches must be added increasing the cost to $1737/node for 1024 nodes. The dual port NICs effectively double the infrastructure and thus add at least $800 per node to the cost.

Myricom provides a full open-source based software stack with support for a variety of OS's and architectures. Though some architectures, such as PPC64, are not tuned as well as others. One of the significant plusses for Myrinet is its strong support for TCP/IP. Generally TCP/IP performance has nearly matched the MPI performance, albeit with higher latency.

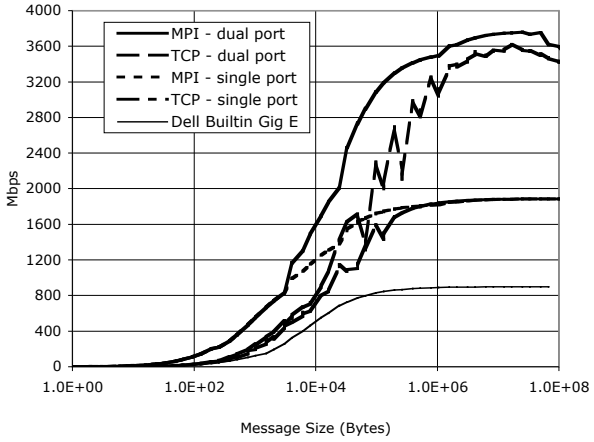

Figure 2: Myrinet Performance

Figure 2 illustrates the performance of both the single and dual port NICs with the Broadcom based Gigabit Ethernet data included from Figure 1 for reference. In both cases the MPI latencies are very good at 6-7 μseconds and TCP latencies of 27-30 μseconds. Also both NICs achieve around 90% of their link bandwidth with over 3750 Mbps on the dual NIC and 1880 Mbps on the single port NIC. TCP/IP performance is also excellent with peak performance of 1880 Mbps on the single port NIC and over 3500Mbps on the dual port NIC.

2.3. Scalable Coherent Interface[3]

The Scalable Coherent Interface (SCI) from Dolphin solutions is the most unique interconnect in this study in that it is the only interconnect that is not switched. Instead the nodes are connected in either a 2D wrapped mesh or a 3D torus depending on the total size of the cluster. The NIC includes intelligent ASICs that handle all aspects of pass through routing, thus pass through is very efficient and does not impact the host CPU at all. However, one downside is that when a node goes down (powered off) its links must be routed around thus impacting messaging in the remaining system.

Since the links between nodes are effectively shared the system size is limited by how many nodes you can effectively put in a loop before it is saturated. Currently that number is in the range of 8-10 leading to total scalability in the range of 64-100 nodes for a 2D configuration and 640-1000 nodes for the 3D torus. Because there are no switches the cost of the systems scales linearly with the number of nodes, $1095 for the 2D NIC and $1595 for the 3D NIC including cables. Unfortunately cable length is a significant limitation with a preferred length of 1m or less, though 3-5m cables can be used if necessary. This poses quite a challenge in cabling up a systems since each node must connected to 4 or 6 other nodes.

Dolphin provides an open source driver and a 3rd party MPICH based MPI is under development. However,

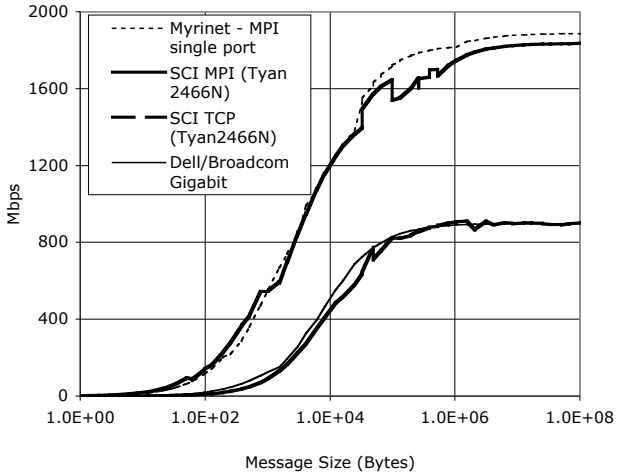

Figure 3: SCI Performance

currently we have not gotten the MPI to function correctly in some cases. An alternative to the open source stack is a package from Scali[4]. This adds $70 per node, but does provide good performance. Unfortunately the Scali packages are quite tied to the Red Hat and Suse distributions that they support. Indeed it is difficult to get the included driver to work on kernels other than the default distribution kernels.

Figure 3 illustrates the performance of SCI on a Tyan 2466N (dual AMD Athlon) based node. Since Dolphin does not currently offer a PCI-X based NIC it is unlikely that the performance would be substantially different in the Dell 2650 nodes used for the other tests. The Scali MPI and driver were used for these tests. The MPI performance is quite good with a latency of 4 μseconds and peak performance of 1830Mbps nearly matching that of the single port Myrinet NIC, which is a PCI-X, based adapter. However, the TCP/IP performance is much less impressive as it barely gets above 900 Mbps and more importantly proved unreliable in our tests.

2.4. Quadrics[5]

The Quadrics QSNET network has been known mostly as a premium interconnect choice on high-end systems such as the Compaq AlphaServer SC. On systems such as the SC the nodes tend to be larger SMPs with a higher per node cost than the typical cluster. Thus the premium cost of Quadrics has posed less of a problem. Indeed some systems are configured with dual Quadrics NICs to further increase performance and to get around the performance limitation of a single PCI slot.

The QSNET system basically consists of intelligent NICs with an on-board IO processor connected via copper cables (up to 10m) to 8 port switch chips arranged in a fat tree. Quadrics has recently released an updated version of their interconnect called QSNet II based on ELAN4 ASICs. Along with the new NICs Quadrics has introduced new smaller switches, which has brought down the entry point substantially. In addition the limit on the total port count has been increased to 4096 from the original 1024. Still it remains a premium option with per port costs starting at $2400 for a small system, $2800 per port for a 64 way system, up to $4078 for a 1024 node system.

On the software side quadrics provides an open source software stack including the driver, userland and MPI. The DMA engine offloads most of the communications onto the IO processor on the NIC. This includes the ability to perform DMA on paged virtual memory addresses eliminating the need to register and pin memory regions. Unfortunately their supported software configuration also requires a licensed, non-open source resource manager (RMS). In our experience the RMS system was by far the hardest part to get working.

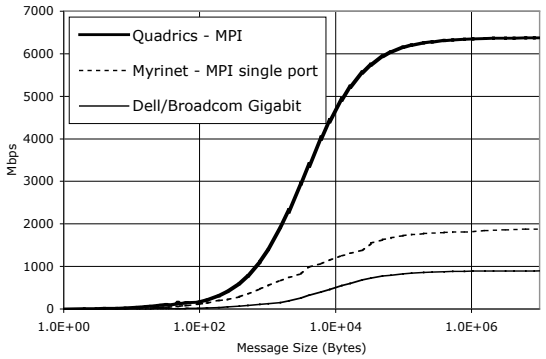

Figure 4: Quadrics Performance

Figure 4 illustrates the performance of Quadrics using the MPI interface. TCP/IP is also provided, but we were unable to get it to build on our system in time for this paper due to incompatibilities with our compiler version. The MPI performance is extremely good with latencies of 2-3 μseconds and peak performance of 6370 Mbps. Indeed this is the lowest latency we have seen on these nodes.

2.5. Infiniband

Infiniband has received a great deal of attention over the past few years even though actual products are just beginning to appear. Infiniband was designed by an industry consortium to provide high bandwidth communications for a wide range of purposes. These purposes range from a low-level system bus to a general purpose interconnect to be used for storage as well as inter-node communications. This range of purposes leads to the hope that Infiniband will enjoy commodity-like pricing due to its use by many segments of the computer market. Another promising development is the use of Infiniband as the core system bus. Such a system could provide external Infiniband ports that would connect directly to the core system bridge chip bypassing the PCI bus altogether. This would not only lower the cost, but also provide a significantly lower latency. Another significant advantage for Infiniband is that is designed with scalable performance. The basic Infiniband link speed is 2.5Gbps (known as a 1X link). However the links are designed to bonded into higher bandwidth connections with 4 link channels (aka a 4X links) providing 10Gbps and 12 link channels (aka 12X links) providing 30 Gbps.

Current Infiniband implementations are available as PCI-X based NICs and 8 or 24 port switch chips utilizing 4X (10Gbps) links. The switch chips can be configured for other link speeds as well including 12X links. This makes switch aggregation somewhat easier since you can configure a switch with 12 4X links to connect to nodes and 4 12X links to connect to other switches. Current Infiniband pricing is enjoying the benefits of aggressive venture capital funding while the various vendors attempt to define their market. Thus there are a variety of vendors competing with slightly different NICs and switches even though the core silicon in most current implementations is from Mellanox[6]. Current pricing ranges from $1200-$1600/per depending on the vendor and cluster size (pricing would be higher for very large clusters).

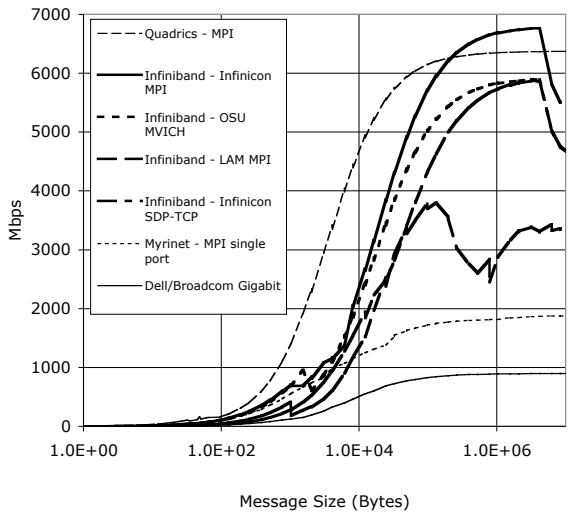

Figure 5: Infiniband Performance

One of the primary ways the vendors are attempting to differentiate their products is through their software stack. This has created a general reluctance to release the software stack as full open-source. In addition many of the software elements are still very much in development. For our tests we attempted to us a couple of MPI implementations, but found the latest MPICH2 based code unstable. Instead we used an older MVICH based implementation. Also included is a non-open source stack from Infinicon Corporation[7].

Figure 5 plots the performance of Infiniband versus a couple of other technologies. Clearly there are substantial differences between the different MPI implementations. The Infinicon MPI shows quite good peak performance over 6750 Mbps, the other two peak out around 5900 Mbps. All three show a large drop off above 4MB in message size when the message size exceeds the on NIC cache size. In addition current latencies of 6-7 μseconds are a bit higher than the other OS-bypass technologies.

2.6. 10 Gigabit Ethernet

10 Gigabit Ethernet is the next step in the Ethernet series. Because of its heritage its design inherits all the advantages and disadvantages of previous Ethernet implementations. The primary advantages include interoperability with previous Ethernet generations, wide portability and a ubiquitous interface (TCP/IP). The primary disadvantages have always been relatively high latency and CPU load, as well as expensive switches. In the present case a portion of the cost is being driven by the cost of the optics which currently run around $3000 by themselves. Some Gigabit Ethernet switches are now offering 10 Gigabit uplink ports essentially at the cost of installing the optics that is making it somewhat more practical to trunk multiple Gigabit Ethernet switches together. However, full 10 Gigabit Ethernet switches are still fairly expensive at around $10000 per switch port even though that figure has dropped by a factor of 3-5 in the last year. In addition 10 Gigabit Ethernet switches are currently fairly limited in port count with large switches offering only around 48 ports. The cost and low-port density clearly make cluster based on 10 Gigabit Ethernet unlikely in the near term. In the longer term these issues will likely mitigate as the technology commoditizes.

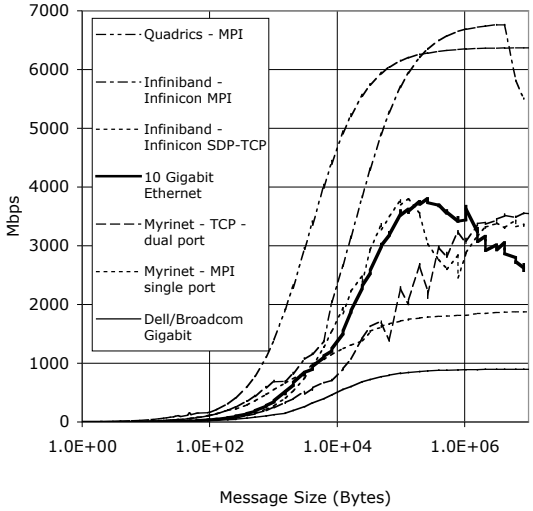

Figure 6: 10 Gigabit Ethernet Performance

Figure 6 plots the performance of two Intel 10Gigabit Ethernet ports against samples of the other technologies. Clearly the performance of 10 Gig. E. is disappointing with performance topping out around

3700Mbps. However this does appear a common limit of TCP/IP among the other technologies and thus may be in part limited by the overhead of TCP/IP itself.

3. Application Performance

Beyond raw network performance there are other issues that can dramatically effect the overall performance of the interconnect. Chief among the issues is how the software and hardware deal with running more processes than there are processors. Especially in the case of applications that use auxiliary processes as data servers. This type of processes typically must block in a receive call. Under many MPI implementations this results in a polling behavior that can take substantial CPU time away from the compute processes. In our local environment the dominant application is the quantum chemistry package GAMESS[8]. Thus we have ran a sample, communication intensive, calculation on each of the interconnect choices to see just how much impact the network has on the timings.

#node, 2Procs per	1	2	4
Gigabit Ethernet	5026	1961	1529
Myrinet MPI	3277	1984	1559
Myrinet TCP	3477	1924	1279
Infiniband MPI	4868	2785	1319
Quadrics MPI		2593	1427
# node, 1 proc per	**1**	**2**	**4**
Gigabit Ethernet	5823	3062	1852
Myrinet MPI	5651	3030	1743
Myrinet TCP	5843	3059	1801
Infiniband MPI	5743	3006	1727
Quadrics MPI		3052	1775

Table 1: GAMESS results

Table 1 lists some timings for a calculation that utilizes a dual process model. One process functions as the compute process and the second process acts as a data server for a pseudo global shared memory segment. The first block of timings overload the CPUs with 2 compute and 2 data servers on each dual CPU node. The second block run only 1 compute and 1 data server per node.

One obvious result is that the fastest network does not necessarily produce the fastest timing for this application. When the CPU's are not overloaded (ie one compute and one data server per node) the timings are fairly similar for all interconnect types. However when the number of processes per node is doubled the results are more interesting. First off Myrinet turns in substantially faster timing for a single node. Most likely this indicates a good intra-node message passing implementation. However when running on 4 nodes the Myrinet TCP run is faster than the others. The reason the MPI's are likely slower is due to the way many MPI's handle blocking in a receive call. Many implementations assume that each MPI process essentially owns a CPU and thus implement a polling mechanism that eats CPU time. For applications such as GAMESS this results in a substantial loss of performance. In addition many of the interconnects seem to perform worse when there are lots of processes utilizing the interconnect at the same time. By this we mean there seems to be substantial overhead in multiplexing and demultiplexing the messages when multiple processes are simultaneously active.

4. Conclusions

The good news is that today there are several good choices for a high-speed interconnect on a cluster at a range of price. At the low-end Gigabit Ethernet has emerged as a solid option with a cost of $750/node or below up to several hundred nodes. Of course Gigabit Ethernet is also the lowest performing network in the survey, but its solid, ubiquitous, software support make it a good choice for applications that do not require cutting edge communications performance.

In the middle of the pack in terms of cost and performance are SCI and Myrinet. Both products offer good performance at a moderate cost. SCI has some interest due to its flat cost per node as the cluster size is scaled up. However, the software stacks available for SCI have several issues, some of which significantly impact the performance of the system. Myrinet on the other hand offers, a complete open source software package that is the most solid software stack surveyed apart from TCP/IP based Ethernet. The current Myrinet hardware provides good performance at a reasonable cost and is thus a quite solid choice for most applications.

Infiniband stands out as a technology with a great deal of promise, but quite a few rough edges. Chief among the rough edges are the problems with the varied software stacks. Currently it is possible to setup an Infiniband based network and find a usable software stack. However, such a network will require a bit more maintenance as the software is revised.

In the long term 10 Gigabit Ethernet will likely play a role in clusters, but it will likely take at least a couple more years before the price becomes competitive. During that time CPUs and busses will also increase in

performance to the point where 10 Gigabit Ethernet will also be likely to deliver a more acceptable percentage of its theoretical bandwidth.

At the very high-end of the spectrum, in terms of both performance and cost, lies the Quadrics interconnect. It provides very low latency and very good peak performance, but its cost comes in at nearly twice its competitors. This will likely continue to leave Quadrics as a niche product used only on very high-end systems with either big SMP nodes (ex 4 way Alpha or IA64 systems) or those requiring the ultimate in scalability. Even in these cases it would be nice to see the dependence on the licensed RMS software removed.

5. Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy under contract W-7405-Eng-82 at Ames Laboratory operated by the Iowa State University of Science and Technology. Funding was provided by the Mathematical, Information and Computational Science division of the Office of Advanced Scientific Computing Research.

---

[1] Quinn, Snell O.; Mikler, Armin R.; Gustafson, John L; Helmer, Guy. "NetPIPE: a Network Protocol Independent Performance Evaluator". Scalable Computing Laboratory, Ames Laboratory. http://www.scl.ameslab.gov/Projects/NetPIPE/index.html.

[2] http://www.myricom.com/

[3] http://www.dolphinics.com/

[4] http://www.scali.com/

[5] http://www.quadrics.com/

[6] http://www.mellanox.com/

[7] http://www.infinicon.com/

[8] M.W.Schmidt, K.K.Baldridge, J.A.Boatz, S.T.Elbert, M.S.Gordon, J.H.Jensen, S.Koseki, N.Matsunaga, K.A.Nguyen, S.J.Su, T.L.Windus, M.Dupuis, J.A.Montgomery *J. Comput. Chem.* **14**, 1347-1363 (1993).