PathScale™

AN INTRODUCTION TO THE PATHSCALE INFINIPATH™ HTX™ ADAPTER

LLOYD DICKMAN
Distinguished Architect, Office of the CTO
PathScale, Inc

Executive Summary

Cluster systems based on commodity processors and the Linux operating system have emerged as a powerful force for solving scientific problems. In addition, the Message Passing Interface (MPI) standard is a widely used paradigm for many current high performance computing applications.

Parallel high performance computing (HPC) applications require significant communication between computation nodes. The interprocess communications interconnect network, or fabric, is a critical component of application speedup and cluster efficiency. The better this communication performs (i.e., the lower the communications latency), the faster the time-to-solution.

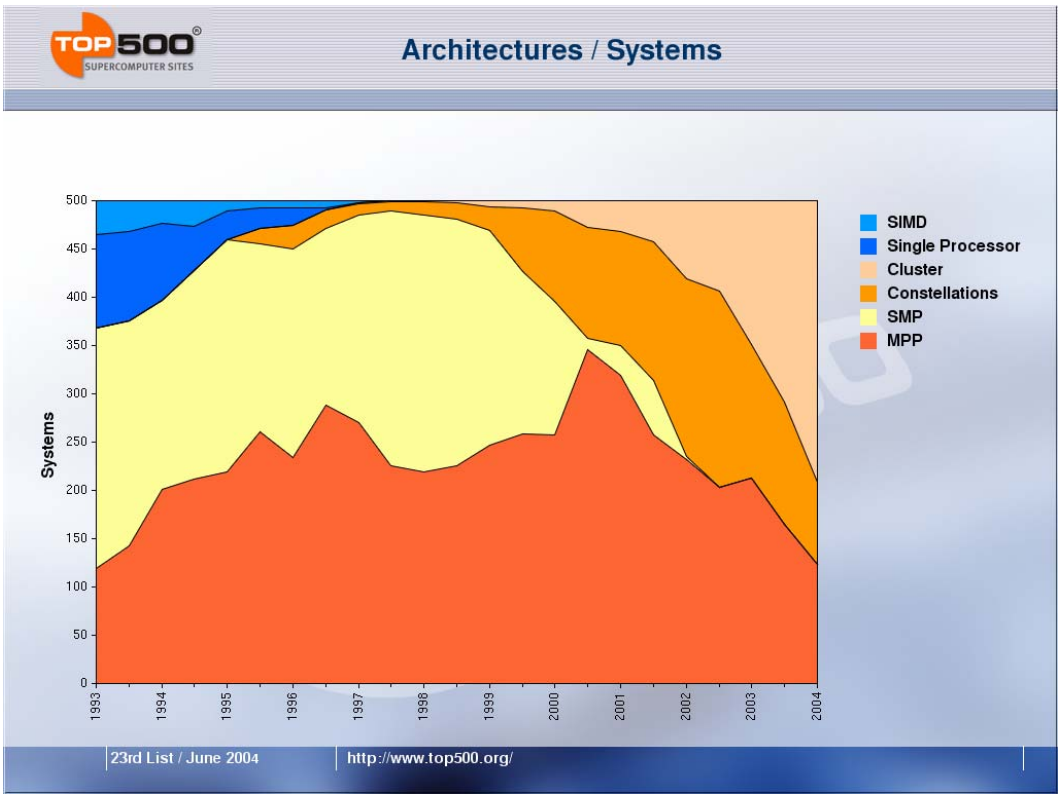
PathScale's InfiniPath™ is a single chip ASIC that directly connects processors supporting HyperTransport such as the AMD Opteron™ to the industry standard InfiniBand™ network fabric. InfiniPath, in combination with the InfiniBand switching fabric, is an industry leading interconnect for MPI codes. The design is optimized for communications patterns typically found in HPC applications. InfiniPath is optimized for both small message communications patterns requiring ultra-low latency, and medium to large messages requiring high-bandwidth. In addition, InfiniPath provides an outstanding $n_{½}$ message size and superior collectives performance to support high levels of application scaling. An IP transport is also provided to permit Sockets, TCP, and UDP applications as well as existing software stacks to coexist on this enhanced interconnect. Use of the InfiniBand switching fabric permits high bandwidth to be realized at a commodity fabric price point.

PathScale InfiniPath is available as a chip for system manufacturers to design onto motherboards, or as an add-on adapter card (PathScale InfiniPath™ HTX™ Adapter) using the new HyperTransport HTX connector standard. Wide availability of industry standard InfiniBand switches, cables and management solutions provide a robust and cost-effective interconnect fabric.

1. The need for a low latency interconnect

Industry Trends

Cluster computing has come of age. The amazing improvements in microprocessor technology along with supporting improvements in high performance interconnects enabled clusters to displace SMPs and become the dominant computing model of the Top 500 computing list in the past 5 years:



Source: top500.org

This powerful trend is driven by the following factors:

- Availability of powerful, cost-effective commodity microprocessors with
    - 64-bit virtual addressing
    - Large physical memories
    - Excellent multi-processing performance for compute nodes
- The emergence of low-latency, high-bandwidth interconnects
- Applications needing to scale aggressively to handle higher resolution problems and to accelerate time-to-solution

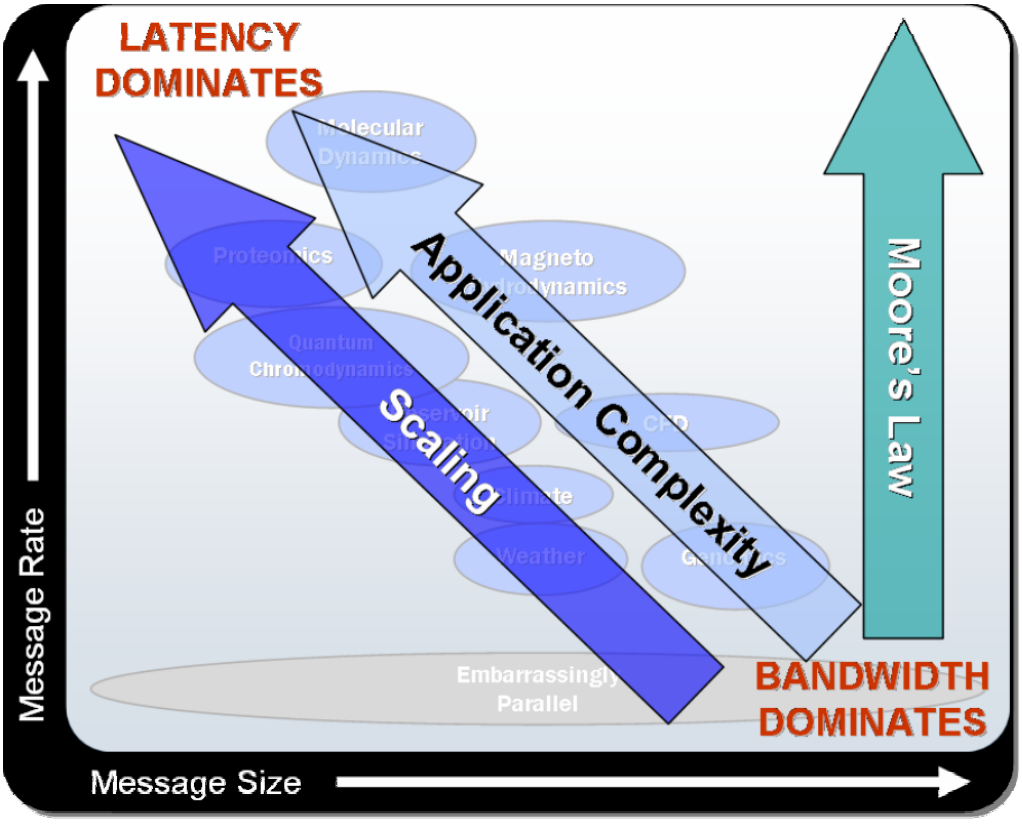
People who make cluster buying decisions have every right to expect a high degree of efficiency. Adding more nodes should result in cost-effective performance improvement.

The interconnect and network fabric is often the single greatest determinant of parallel application speed and cluster efficiency. The interconnect must have low latency and high bandwidth to match the demanding interprocess communication needs of the applications. Many HPC applications exhibit a bimodal communications pattern. There are frequent periods of chatty communications exchanges where latency is paramount. In addition there are times when long messages requiring high-bandwidth are sent. As microprocessors march up the curve of Moore's law, improved interconnect latencies are needed to maintain system balance.

With scalability and efficiency as buying criteria, the interconnect fabric choice becomes critical. It gates the scalability of applications, and consequently the efficiency of the cluster. The PathScale InfiniPath interconnect represents a departure from previous choices, offering ultra-low latency using a commodity fabric, while providing industry-leading application scaling.

Evaluating Application Communication Needs

The communication patterns of parallel applications span a wide range of messaging rates and message sizes as shown below.



Applications have wide variation in communication patterns. For example, genome matching programs, seismic processing, and EDA Verilog simulation are typical “embarrassingly parallel” applications. Their workloads are partitioned such that each parallel node can work with limited interaction with other nodes. Applications with limited communication needs work well over low-cost networks such as 100 Mbps (10 MB/s) or 1 Gpbs (100 MB/s) Ethernet.

Applications that communicate with larger message sizes are sensitive to the network bandwidth since they must communicate large amounts of data quickly. Applications with high messaging rates are latency sensitive since they exchange small to modest amounts of data rapidly among a set of cooperating nodes in order to move the computation forward towards a solution. Applications that are typically latency and bandwidth sensitive include reservoir simulation, proteomics, weather / climate modeling, computational fluid dynamics, computational chemistry, molecular modeling, and weapons simulations.

The straightforward solution of many scientific problems can be programmed using a “fine grained” representation of the problem that directly reflects the underlying physical phenomena. Ultra-low latency interconnects are an excellent match to such approaches. Previously, many communications networks have had poor communication latencies relative to their bandwidth, requiring in significant computer science and programming effort to convert important codes into a “course grained” representation that communicated using fewer, but larger messages.

Even with such heroic efforts, the need to scale programs to provide greater application fidelity, or to increase the number of nodes to accelerate time-to-solution, again drives the communication patterns towards a requirement for improved latency.

Lastly, as microprocessors march up the curve of Moore’s law, lower and lower latencies are needed to maintain the system balance as the improved speed of computation pushes towards higher messaging rates.

Bottom line: *Communications latency is a significant, perhaps the primary, determinant of applications performance, and will become more so over time.*

2. Evolution of Interconnect Networks

At the lower end of the cost and performance spectrum are clusters connected by Ethernet. Although the emerging 10 Gbps (1 GB/s) Ethernet solutions offer competitive bandwidth, they are currently extremely expensive and are projected to remain so for the foreseeable future. Additionally, they offer unacceptably poor latencies when compared with alternative fabrics.

A recent entry into the interconnect arena is standard InfiniBand. InfiniBand interconnect fabrics are now available with 1 GB/s bandwidth (in each direction) with low-latency, high port-count, cut-through switches available from several vendors including Infinicon, Mellanox, TopSpin, and Voltaire.

Leaving the realm of standard interconnect architectures, we find proprietary solutions from Myricom and Quadrics. Quadrics has traditionally been the latency leader, although at a significantly higher price point than other alternatives. Myricom has traditionally offered lower pricing than Quadrics, although at a reduced performance level for both latency and bandwidth.

With scalability, efficiency, and cost-effectiveness as buying criteria, the choice of interconnect fabric becomes critical. It gates the scalability of applications, and consequently the efficiency of the cluster and time-to-solution of the applications. Although many HPC cluster users would also enjoy moving away from proprietary interconnect fabrics and use InfiniBand-based solutions, generic InfiniBand host adapters lack adequate latency when compared with proprietary solutions.

The PathScale InfiniPath interconnect represents a dramatic departure from the continuum above, providing ultra-low latency based on the commodity high-bandwidth InfiniBand industry standard network fabric.

Applications requiring the highest levels of interconnect performance have traditionally run on SMPs because of the relatively low latency and high bandwidth possible using shared memory systems. The ultra-low latency of PathScale's InfiniPath makes it possible to consider lower-cost cluster computing platforms for such applications.

3. PathScale InfiniPath Solution

In keeping with the customer trends and needs above, PathScale recently introduced a new interconnect for High Performance Computing, the PathScale InfiniPath Interconnect. InfiniPath represents an ultra-low latency, highly-scalable interconnect attaching directly to HyperTransport systems such as the AMD Opteron and using the industry standard InfiniBand switching fabric.

InfiniPath provides an aggressive solution for the principal interconnect metrics:

- Latency
- Bandwidth
- Scaling: Collectives and $n_{½}$ half-power point (message size for which half the bandwidth is achieved)

With an innovative internal architecture and form-factors appropriate for system vendors and end-user deployment.

Ultra-Low Latency

InfiniPath directly connects to HyperTransport links. HyperTransport is available on microprocessors such as the AMD Opteron, IBM PowerPC 970, and others. This gives it a distinct advantage over connections made through PCI-X or even PCI Express, because of fewer chip crossings and an inherently low-latency processor communications protocol. HyperTransport provides the absolute lowest latency method of communicating with processors.

With PathScale InfiniPath, the host CPU provides the MPI protocol processing cycles. In latency sensitive environments dominated by communication overheads, it is preferable to assign critical path operations to the fastest processing component, rather than off-load the critical path to a processor that is 5~10X slower and endure additional protocol crossing overhead. The PathScale approach is consistent with microprocessor industry directions for multi-thread processors, multi-core

processors, and increasingly aggressive processing rates. PathScale is riding with Moore's Law rather than fighting against it!

Host-based MPI protocol support provides additional advantages for achieving ultra-low latency by eliminating extraneous memory traffic. Typically, the data payload for short MPI messages resides in the processor's cache. By moving this payload directly to the InfiniPath interconnect, InfiniPath avoids the additional memory operations incurred by other interconnects that move the data payload back into memory prior to fetching it again for launch over the interconnect fabric.

InfiniPath messaging avoids kernel intervention in all but exceptional error recovery cases. This further reduces the end-to-end latency, and reduces additional sources of OS noise by eliminating system calls and process scheduling events.

Bottom line: *An industry leading 1.5 us for end-to-end MPI messaging, inclusive of switching fabric.*

High Bandwidth

PathScale InfiniPath directly connects HyperTransport links to the InfiniBand switching fabric. HyperTransport is a high-bandwidth, low-latency, new generation host system interconnect. InfiniPath simultaneously moves 3.2+3.2 GB/s (bi-directional) to and from the AMD Opteron host.

Each InfiniPath Interconnect chip provides an InfiniBand 4X link to the switching fabric, and can simultaneously move 1+1 GB/s (bi-directional) to and from the switching fabric.

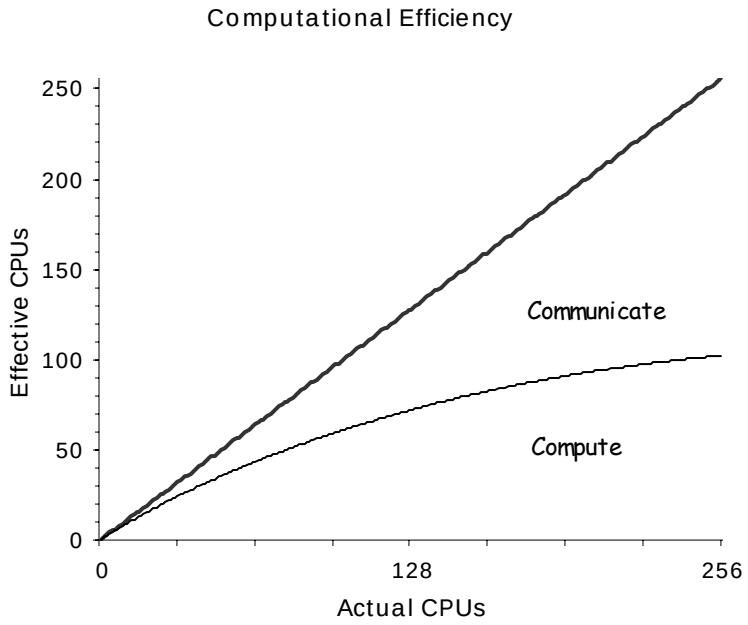
The InfiniBand ecosystem has evolved to provide robust, cost-effective, and manageable switching fabrics available from several vendors. State-of-the-art InfiniBand switch chips provide 48 GB/s of non-blocking aggregate bandwidth over 24-ports with 200 ns latency. These switch chips are used by InfiniBand vendors to provide up to 288-port full bisection bandwidth switches with 576 GB/s of non-blocking aggregate bandwidth at under 420 ns latency. Robust, full bisection bandwidth switching fabrics can be constructed to support thousands of nodes.

Bottom line: *Over 1.9 GB/s of bi-directional user data per PathScale InfiniPath link across a full bisection bandwidth InfiniBand switching fabric.*

It's really about scaling

While important to understand, latency and bandwidth metrics are often surrogates for assessing how scalable an interconnect solution can be. The desire to solve problems with finer fidelity and resolution places increasing pressure on reducing the time-to-solution. Improved time-to-solution has enormous economic leverage for accelerating research and engineering development cycles.

For clusters, available time is split between node-based application computation, and inter-node communications. The challenge is to exploit as much computation capability as possible, and reduce the time spent time in inter-node communications. This becomes more difficult with increasing cluster node counts because of the higher levels of communications traffic. The ideal is to use all of the actual computational capability, as shown by the perfect scaling line in the next example. At even modest cluster sizes, the ratio of effective to actual computational capability is compelling.



The half-power point, or $n_{½}$ message size is a key metric for assessing scaling. This is the message size at which half of the maximum interconnect bandwidth is achieved. While latency is a key metric for small messages, and maximum bandwidth is a key metric for long messages, communication patterns for real applications often exchange moderate amounts of data. $n_{½}$ message size is a measure of how effectively these moderate message lengths utilize the potential streaming bandwidth, and by inference, what messaging rates can be supported. For $n_{½}$ message size, smaller is better.

Bottom line: *InfiniPath achieves an industry leading $n_{½}$ message size of under 600 bytes, a 2X to 3X improvement over alternative high-speed interconnects.*

Collectives Performance – As applications scale to larger clusters, the performance of MPI collectives operations becomes a major determinant of time-to-solution and scaling efficiency. InfiniPath provides several performance optimizations:

- Optimized for short collective size. Typical collective operations use small messages. The ultra-low latency of the PathScale InfiniPath interconnect is an exceptionally good match.
- Collective algorithms automatically adapt to cluster size and physical topology. This is especially important in mid to large size clusters.
- Collective processing in host nodes permit all operations, including floating point, to be efficiently handled.
- Responsive aggregation of collective results. Current industry and academic research demonstrates application slow-down due to operating system induced process variation, highlighting the performance vulnerabilities of existing solutions.
- Multicast communication enhances efficiency in large clusters.

Bottom line: *InfiniPath delivers outstanding performance for collective operations.*

Architecture

The PathScale InfiniPath HTX Adapter for AMD Opteron based clusters provides extremely high performance over a commodity interconnect fabric. InfiniPath is a single chip ASIC solution. It does not have, or need, an embedded processor. There is no external memory. It provides proven architecture features including OS-bypass and zero-copy, while incorporating robust data-integrity features to support large, mission critical cluster applications.

InfiniPath provides a connection directly from the HyperTransport link of an AMD Opteron cluster node to a standard InfiniBand 4X switch fabric. By interfacing at the physical and link layers, it exploits the inherent low-latency of the InfiniBand network fabric.

Host CPU cycles are used to drive InfiniPath, since host cycles are 5~10X faster than solutions dependent on embedded microprocessors. Protocol and error handling are done on the host with low processor overhead.

End-to-end data integrity is supported by an extensive parity and redundancy checking on all memories and data buses, as well as internal hardware protocol validation.

System Form Factors

PathScale InfiniPath is a single chip interconnect solution that directly attaches to an AMD Opteron HyperTransport link on the host side and an industry standard InfiniBand 4X link on the fabric side. HyperTransport and InfiniBand bi-directional paths each run simultaneously without interference. No external memory is used.

The InfiniPath chip is offered separately to system or motherboard vendors who require a “Landed on Motherboard” interconnect solution. The InfiniPath chip uses the HyperTransport tunnel configuration to co-exist with other HyperTransport devices in a single HyperTransport chain, if required. This is extremely beneficial in larger Opteron multiprocessors such as 4-way and 8-way configurations.

In addition, PathScale offers the InfiniPath chip on a HyperTransport add-in card using the HyperTransport HTX standard. The PathScale InfiniPath HTX Adapter is designed for customers and vendors who want the flexibility of an add-in card solution. The HyperTransport Technology Consortium recently announced the HTX slot and connector standard. Iwill Corporation, a leading AMD Opteron motherboard manufacturer is the 1st motherboard supplier to offer a standard EATX form-factor dual processor Opteron board with an HTX slot standard on the motherboard. This Iwill board, model DK8S2-HTX, can be incorporated into any server that accommodates EATX form factor boards, commonly found in 1U servers.

4. InfiniPath is Standards-based

The PathScale InfiniPath interconnect solution is standards based. The standards embraced include: InfiniBand, Message Passing Interface (MPI), HyperTransport, and Linux.

- InfiniBand is the network fabric for the InfiniPath solution. It is an industry standard managed by the InfiniBand Trade Association (www.infinibandta.org). Users can acquire InfiniBand switches and cable solutions from several vendors. InfiniPath is interoperable with existing InfiniBand fabrics, including switches, cables, and systems management. InfiniPath utilizes existing fabric management solutions.

- MPI is a widely used standard for parallel programming (www-unix.mcs.anl.gov/mpi/). InfiniPath is MPICH compliant.

- The HyperTransport host connection is used to connect to processors such as the AMD Opteron. It is an industry standard, managed by the HyperTransport Technology Consortium (see www.hypertransport.org). It is used by several computer system vendors to support a range of microprocessor architectures. HyperTransport members include AMD, Apple, IBM, Sun, Cisco, Broadcom, Iwill, nVidia, and Network Appliance.

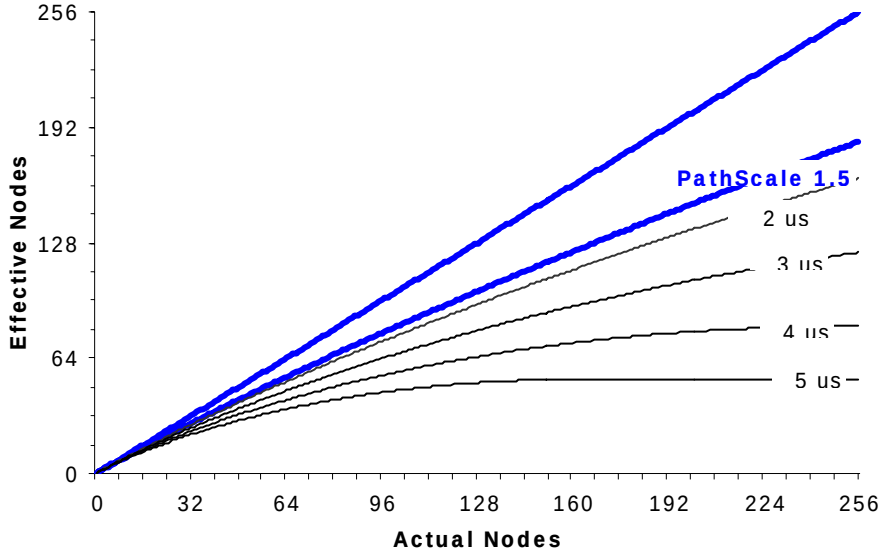
  HyperTransport provides a direct point-to-point processor connection, eliminating traditional NorthBridge structures. Its dual unidirectional links provide the fastest interconnect available for on-board components, including processors and I/O devices.

  The HyperTransport Technology Consortium recently published a new standard for a HyperTransport motherboard slot called HTX. The specifications on the new HTX slot standard can be found at www.hypertransport.org.

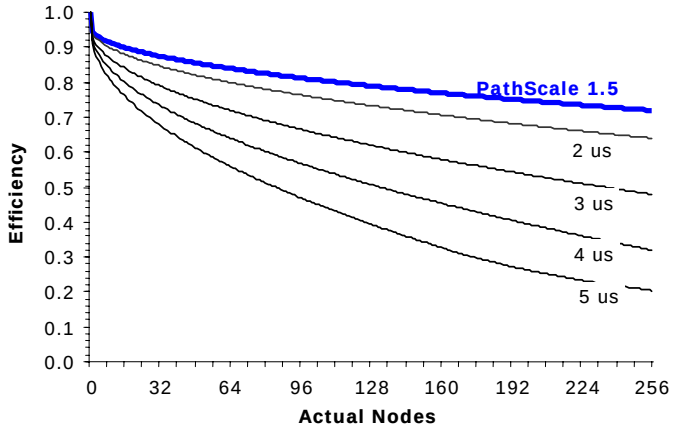
- Linux has gained significant acceptance in high performance cluster computing community (see www.kernel.org). InfiniPath software stacks are provided for widely used Linux distributions. InfiniPath software is loaded as a user library and kernel driver. It does not alter the kernel.

5. Value of a high speed interconnect, and do you need one?

As mentioned earlier, there are several choices one can make for the interconnect fabric. The following chart shows how basic interconnect latency affects the time-to-solution for an example molecular dynamics computation. “Effective nodes” is directly proportional to the relative time-to-solution. Compared with existing high-speed interconnect alternatives, PathScale InfiniPath enables the use of more effective nodes that can be applied to the application. This provides greater cluster efficiency and a superior return on investment.



An alternate way of looking at the same results is to examine how well cluster efficiency and time-to-solution hold up with increasing node counts. The importance of an ultra-low latency InfiniPath interconnect solution is apparent:



Up to this point, we have been talking about a single application on a cluster. While clusters are often justified by a single application, many must support a range of applications. This is particularly true at large HPC centers that support hundreds of users and dozens of applications. Since all of the application requirements aren't necessarily understood at procurement time, it is prudent to acquire the fastest interconnect affordable. For example, from the efficiency chart above, we see that there is a doubling of efficiency at 128 nodes when using the InfiniPath interconnect solution rather than a 4 us alternative. The economic value of halving the time-to-solution must be determined by the user, but is likely significant.

6. Conclusions

Linux clusters represent the fastest growing platform for high performance computing. Use of clusters for high performance computing demands a low latency interconnect fabric.

The industry has shown a clear preference for standards-based approaches as standards make systems more cost-effective.

InfiniPath, from PathScale, provides an elegant solution to the interconnect challenge; ultra-low latency and high bandwidth with outstanding scaling characteristics based on a cost-effective industry standard network fabric. Because it is based on standards for InfiniBand, MPI, HyperTransport, and Linux, PathScale InfiniPath offers an immediately deployable interconnect solution for clusters of all sizes.

When coupled with the PathScale EKOPath compilers, the world's highest performance 64-bit compilers, and the PathScale OptiPath MPI Acceleration Tools, the industry's 1st easy to use root cause analysis toolkit for MPI applications, cluster efficiency can be substantially enhanced.

Give your domain experts the tools that deliver time-to-solution, from PathScale!

For More Information

Additional information on PathScale, Inc. and its products can be obtained by visiting http://www.pathscale.com on the World Wide Web, or contacting:

PathScale, Inc.
477 N. Mathilda Avenue
Sunnyvale, CA. 94085 USA

Phone: (408) 746-9100
Fax: (408) 746-9150
www.pathscale.com