Scalable Cache Coherent Shared Memory at Cluster Prices

White Paper

Redefining Scalable OpenMP and MPI Price-to-Performance with Numascale's NumaConnect

By: Douglas Eadline

About the Author: Douglas Eadline, Ph.D. has been writing and working with high performance computers and Linux for over twenty years. He is also the Editor of ClusterMonkey.net.

NW4V1

ABSTRACT

Using commodity hardware and the "plug-and-play" NumaConnect interconnect, Numascale delivers true shared memory programming and simpler administration at standard HPC cluster price points. One such system currently offers users over 1,700 cores with a 4.6 TB single memory image.

The NumaConnect cluster excels at both OpenMP and MPI computing within the same shared memory environment. No extra software or program modifications are needed to take advantage of the entire system. Results for the NASA Advanced Supercomputing (NAS) Parallel Benchmarks have set a new record for OpenMP core count and problem size. OpenMP results show good scalability, with best results coming from larger problem sizes.

In addition, NumaConnect shared memory MPI performance delivers better results than InfiniBand clusters, using standard tools without modification for the underlying hardware environment. A cost comparison with a small FDR InfiniBand cluster shows a comparable price point when ease of programming, high performance, and ease of administration are considered.

Several production systems are performing satisfactorily, including those in University of Oslo in Norway, Statoil, and Keele University.

Finally, we pose the question: ***If you can get scalable OpenMP and MPI performance, ease of programming, and ease of administration at commodity cluster price points, why limit yourself to an MPI cluster?***

The NumaConnect Breakthrough

Built to work as a "plug-and-play" hardware solution, the NumaConnect from Numascale uses industry-standard HyperTransport to seamlessly join commodity motherboards into a true shared memory system running a single Operating System (OS) image. Numascale's NumaConnect™ technology enables system vendors to build shared memory HPC clusters with servers at a fraction of the price of enterprise-level shared memory systems. At the heart of NumaConnect is the NumaChip™ that contains both cache coherency logic and switching capability to build large torus topologies. NumaConnect provides a record-breaking main memory capacity of up to 256TB in a single commodity-based system.

The following discussion will highlight some of the important advantages of NumaConnect systems and most importantly, provide performance and cost numbers that make NumaConnect-based HPC systems a compelling choice.

Shared Memory Advantages

Multi-processor shared memory processing has long been the preferred method for creating and running technical computing codes. Indeed, this computing model now extends from a user's dual core laptop to 16+ core servers. Programmers often add parallel OpenMP directives to their programs in order to take advantage of the extra cores on modern servers. This approach is flexible and often preserves the "sequential" nature of the program (pthreads can of course also be used, but OpenMP is much easier to use). To extend programs beyond a single server, however, users must use the Message Passing Interface (MPI) to allow the program to operate across a high-speed interconnect.

Interestingly, the advent of multi-core servers has created a parallel asymmetric computing model, where programs must map themselves to networks of shared memory SMP servers. This asymmetric model introduces two levels of communication, local within a node, and distant to other nodes. Programmers often create pure MPI programs that run across multiple cores on multiple nodes. While a pure MPI program does represent the greatest common denominator, better performance may be sacrificed by not utilizing the local nature of multi-core nodes. Hybrid (combined MPI/OpenMP) models have been able to pull more performance from cluster hardware, but often introduce programming complexity and may limit portability.

Clearly, users prefer writing software for large shared memory systems to MPI programming. This preference becomes more pronounced when large data sets are used. In a large SMP system the data are simply used in place, whereas in a distributed memory cluster the dataset must be partitioned across compute nodes.

In summary, shared memory systems have a number of highly desirable features that offer ease of use and cost reduction over traditional distributed memory systems:

- Any processor can access any data location through direct load and store operations, allowing easier programming (less time and training) for end users, with less code to write and debug.

- Compilers, such as those supporting OpenMP, can automatically exploit loop level parallelism and create more efficient codes, increasing system throughput and better resource utilization.

- System administration of a unified system (as opposed to a large number of separate images in a cluster) results in reduced effort and cost for system maintenance.

- Resources can be mapped and used by any processor in the system, with optimal use of resources in a single image operating system environment.

Shared Memory as a Universal Platform

Although the advantages of shared memory systems are clear, the actual implementation of such systems "at scale" has been difficult prior to the emergence of NumaConnect technology. There have traditionally been limits to the size and cost of shared memory SMP systems, and as a result the HPC community has moved to distributed memory clusters that now scale into the thousands of cores. Distributed memory programming occurs within the MPI library, where explicit communication pathways are established between processors (i.e., data is essentially copied from machine to machine). A large number of existing applications use MPI as a programming model.

Fortunately, MPI codes can run effectively on shared memory systems. Optimizations have been built into many MPI versions that recognize the availability of shared memory and avoid full message protocols when communicating between processes.

Shared memory programming using OpenMP has been useful on small-scale SMP systems such as commodity workstations and servers. Providing large-scale shared memory environments for these codes, however, opens up a whole new world of performance capabilities without the need for re-programming.

Using NumaConnect technology, scalable shared memory clusters are capable of efficiently running both large-scale OpenMP and MPI codes without modification.

Record-Setting OpenMP Performance

In the HPC community NAS Parallel Benchmarks (NPB) have been used to test the performance of parallel computers (http://www.nas.nasa.gov/publications/npb.html). The benchmarks are a small set of programs derived from Computational Fluid Dynamics (CFD) applications that were designed to help evaluate the performance of parallel supercomputers. Problem sizes in NPB) are predefined and indicated as different classes (currently A through F, with F being the largest).

Reference implementations of NPB are available in commonly-used programming models such as MPI and OpenMP, which make them ideal for measuring the performance of both distributed memory and SMP systems. These benchmarks were compiled with Intel ifort version 14.0.0. (Note: the current-generated code is slightly faster, but Numascale is working on NumaConnect optimizations for the GNU compilers and thus suggests using gcc and gfortran for OpenMP applications.)

For the following tests, the NumaConnect Shared Memory benchmark system has 1TB of memory and 256 cores. It utilizes eight servers, each equipped with two x AMD Opteron 2.5 GHz 6380 CPUs, each with 16 cores and 128GB of memory.

Figure One shows the results for running NPB-SP (Scalar Penta-diagonal solver) over a range of 16 to 121 cores using OpenMP for the Class D problem size.

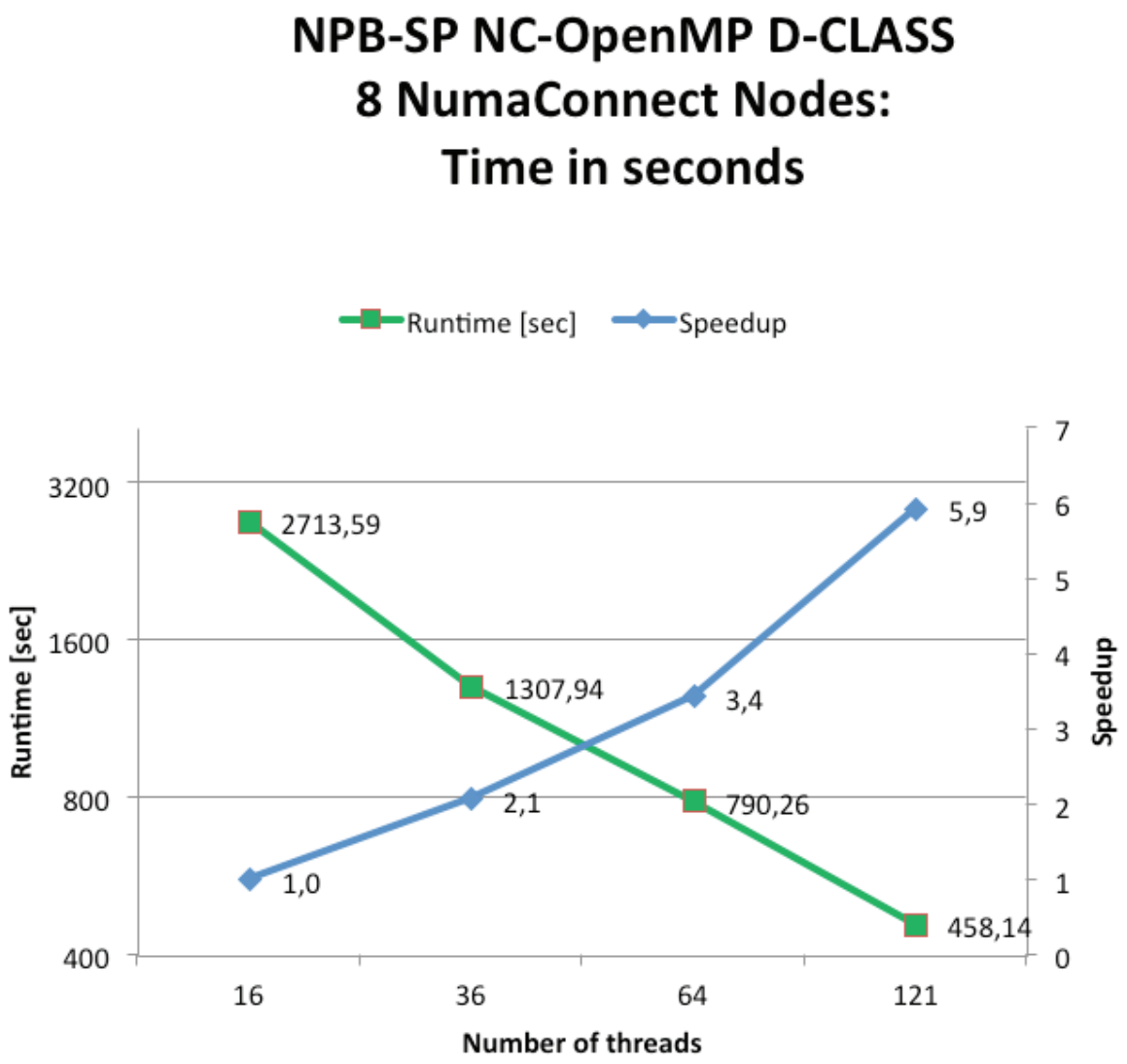

Figure One: OpenMP NAS Parallel results for NPB-SP (Class D)

Figure Two shows results for the NPB-LU benchmark (Lower-Upper Gauss-Seidel solver) over a range of 16 to 121 cores, using OpenMP for the Class D problem size.

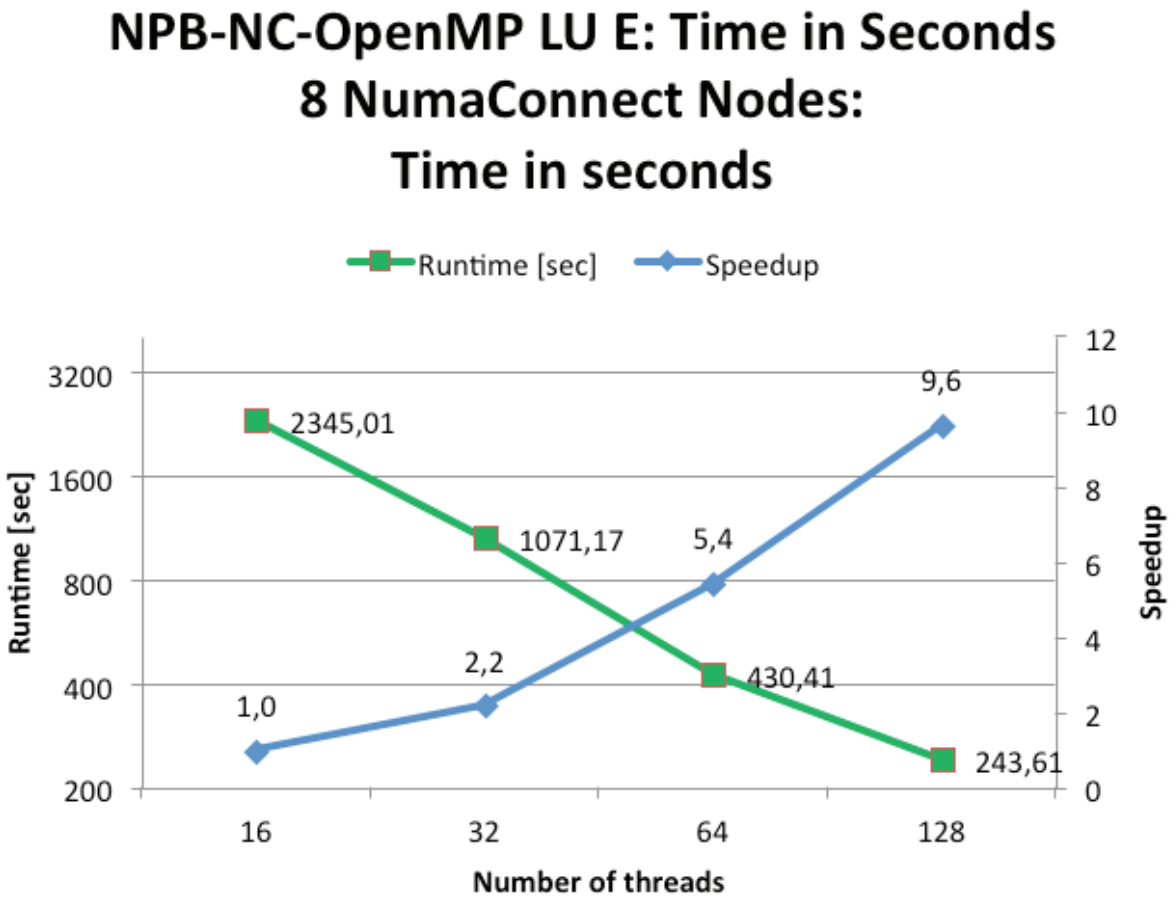

Figure Two: OpenMP NAS Parallel results for NPB-LU (Class E)

Figure Three shows he NAS-SP benchmark E-class scaling perfectly from 64 processes (using affinity 0-255:4) to 121 processes (using affinity 0-241:2). Results indicate that larger problems scale better on NumaConnect systems, and it was noted that NASA has never seen OpenMP E Class results with such a high number of cores.

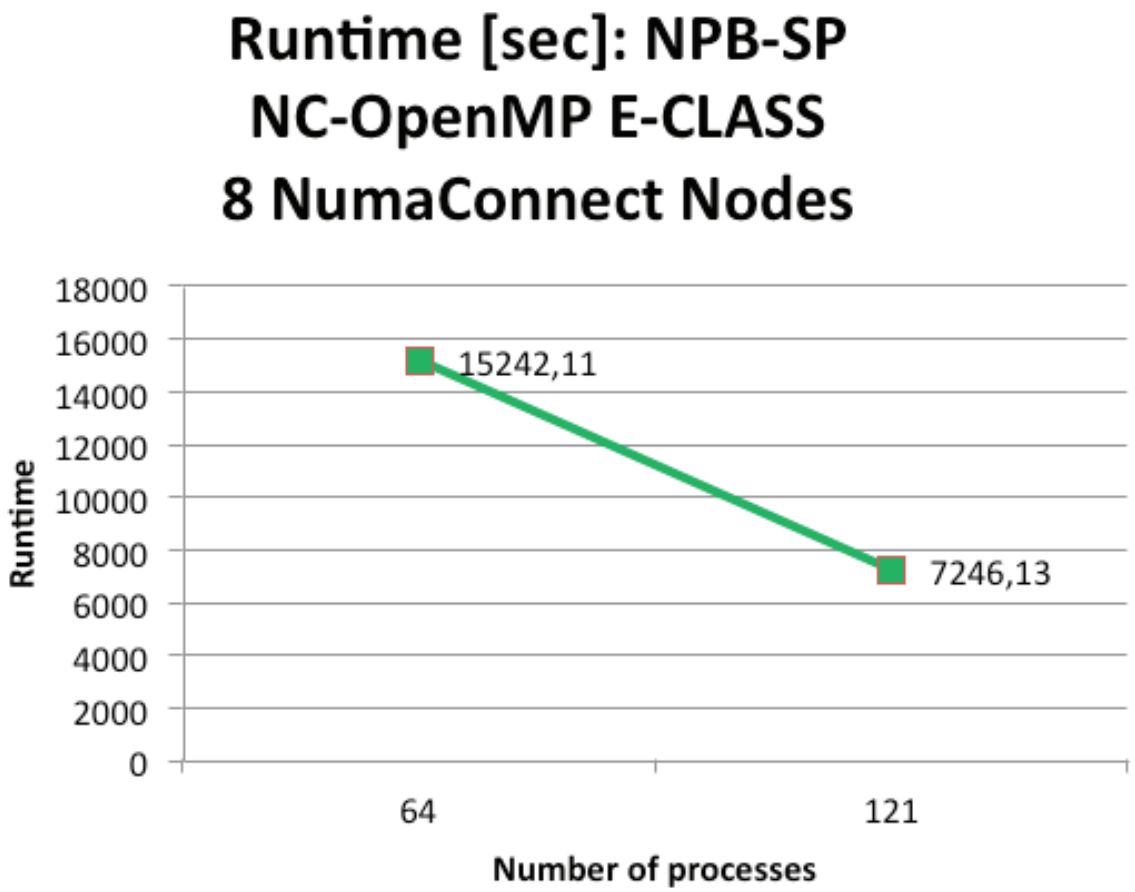

Figure Three: OpenMP NAS results for NPB-SP (Class E)

OpenMP applications cannot run on InfinBand clusters without additional software layers and kernel modifications. The NumaConnect cluster runs a standard Linux kernel image

Surprisingly Good MPI Performance

Despite the excellent OpenMP shared memory performance that NumaConnect can deliver, applications have historically been written using MPI. The performance of these applications is presented below. As mentioned, the NumaConnect system can easily run MPI applications. Figure Four is a comparison of NumaConnect and FDR InfiniBand NPB-SP (Class D). The results indicate that NumaConnect performance is superior to that of a traditional distributed InfiniBand memory cluster. MPI tests were run with OpenMPI and gfortran 4.8.1 using the same hardware mentioned above.

Both industry-standard OpenMPI and MPICH2 work in shared memory mode. Numascale has implemented their own version of the OpenMPI BTL (Byte Transfer Layer) to optimize the communication by utilizing non-polluting store instructions. MPI messages require data to be moved, and in a shared memory environment there is no reason to use standard instructions that implicitly result in cache pollution and reduced performance. This results in very efficient message passing and excellent MPI performance.

Similar results are shown in Figure Five for the NAS-LU (Class D). NumaConnect's performance over InfiniBand may be one of the more startling results for the NAS benchmarks. Recall again that OpenMP applications cannot run on InfiniBand clusters without additional software layers and kernel modifications.

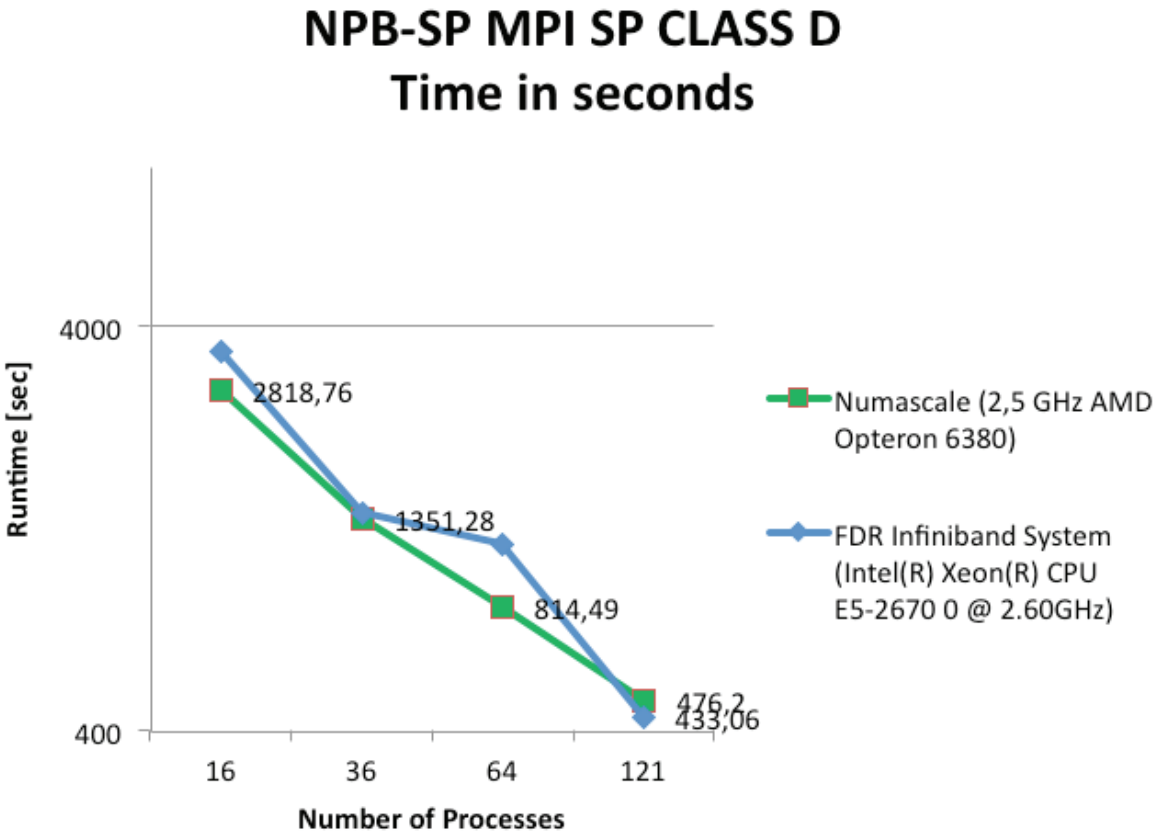

Figure Four: NPB-SP comparison of NumaConnect to FDI InfiniBand

Comparable Cost and Better Value

The cost of a NumaConnect cluster is very close to that of an InfiniBand cluster. As an example, consider a 36-node cluster. The cost of the nodes is identical, so it is interesting to examine the

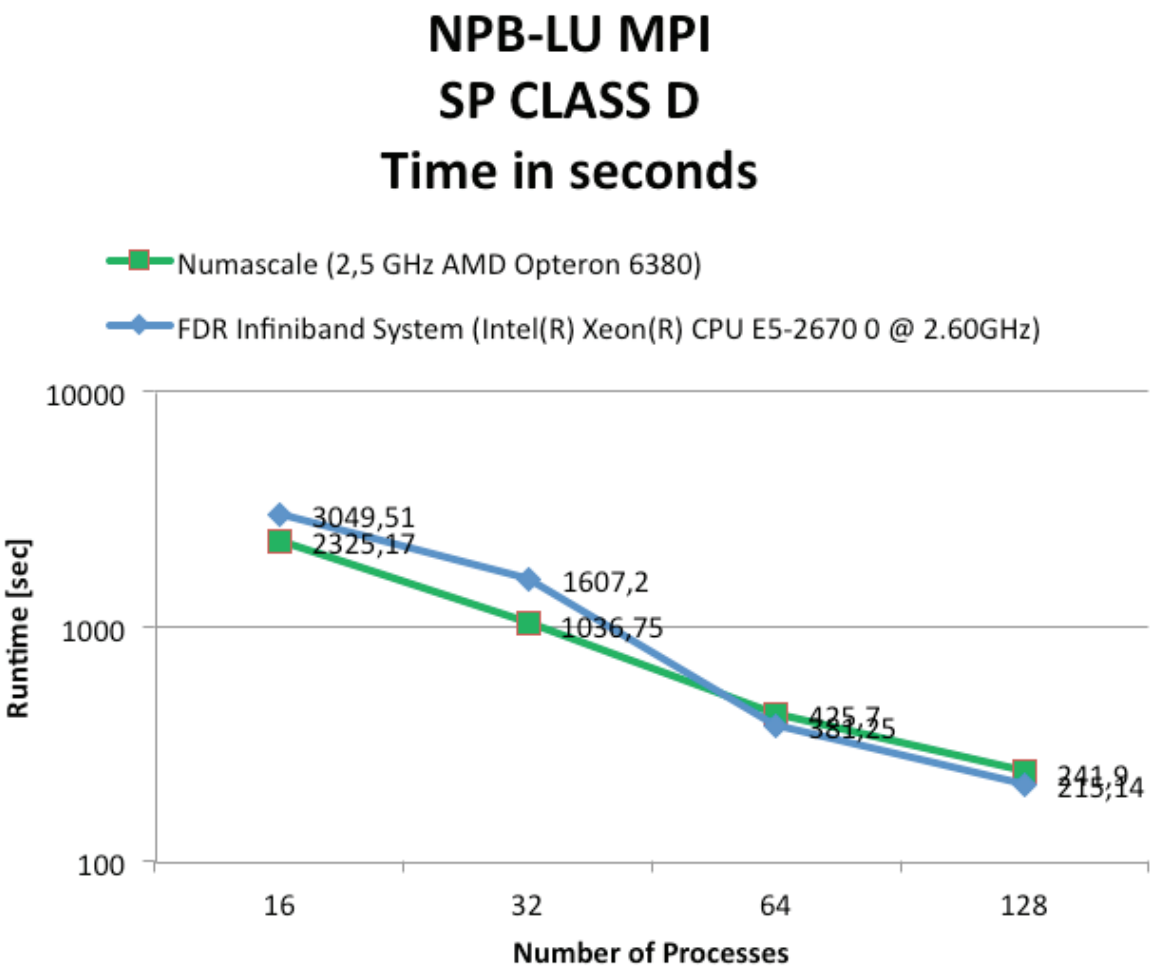

Figure Five: NPB-LU comparison of NumaConnect to FDI InfiniBand

cost of the interconnect itself. The following costs are intended to be "industry average" costs and may change depending on overall system design and requirements.

InfiniBand

Using pricing found online (Colfax Direct) the following minimal equipment is required for InfiniBand connection. The cost per port is calculated as follows:

1	Mellanox MSX6036F-1SFR FDR 36-Port InfiniBand Switch	$10,400
36	Mellanox Active Fiber Cable, VPI, InfiniBand FDR, 3 meters	$13,860
36	ConnectX-3 VPI InfiniBand Adapter Card 1X	$23,544
	Total	$47,804
	Cost per port	$1,328

NumaConnect

Using the standard list price for NumaConnect (a switchless interconnect) the following cost per port can be computed (note: NumaConnect is "switchless" and requires three cables to build a full 3D torus connecting all nodes):

1	NumaConnect Adapter Card N313	$1,750
3	NumaConnect Cables (3D torus)	$360
	Cost per port	$2,110

In summary, for a difference of less than $800 per port, a traditional MPI cluster can become a much more powerful and easier to maintain single image SMP cluster using NumaConnect. A good analogy is the difference between Gigabit Ethernet and InfiniBand. The Ethernet cost per port is much less than that of InfiniBand and all applications will run over both interconnects, but InfiniBand makes the cluster much more efficient. As can be seen from the above benchmarks, NumaConnect offers a similar proposition for cluster users. In addition, NumaConnect's per-port cost is scalable, while InfinBand costs will increase when a larger number of switch ports are added to the cluster.

In many cases the need for users to convert or write code using MPI represents an opportunity cost barrier. The NumaConnect cluster eliminates this "lost opportunity cost" at a price point comparable to that of traditional MPI clusters. Furthermore, the ease of administration — a single Linux OS system image vs. a multitude of system images — removes much of the administrative time and cost.

Quickly Establishing a Track Record

IBM and Numascale installed a system at University of Oslo in 2012. The cluster consists of 72 IBM x3755 2U servers connected in a 3D torus with NumaConnect, using four cabinets with 18 servers apiece in a 3x6x4 topology. Each server has 24 cores and 64GB of memory, providing a single system image of 4.6TB of memory to all 1,728 cores. As part of the PRACE (Partnership for Advanced Computing in Europe) initiative the system is set to begin meeting user demand for "very large" Linux-based applications running under a single large

memory space on commodity x86 based servers.

After a successful testing of Numascale's innovative technology for large-scale real shared memory systems under Statoil's LOOP program, Statoil Technology Invest chose to partner with Numascale. Statoil's applications have extremely large memory requirements and a corresponding interest in the capability to directly address and access very large amounts of data. A NumaConnect cluster was tested with some of Statoil's codes in the LOOP program and has shown impressive results, offering Statoil a very cost-effective and energy-saving processing capability.

In November of 2012 Integrex HPC, a UK-based HPC integrator, was successful in a competitive bid to supply a 1,000-core cluster to Keele University. The system is unique in that 256 cores of the cluster are connected via the latest Numascale fabric. The remaining cores are connected via QDR InfiniBand. The total solution is unified under a single cluster management solution adopting IBM Platform HPC Suite 3.2. Under this environment jobs are submitted via Load Share Facility (LSF) to either the InfiniBand nodes or to the Numascale machine, which has 576GB RAM and 256 cores.

System reliability has been quite good. Thus far, systems with close to 2,000 cores have uptimes measured in months between reboots, which are mostly due to some planned form of maintenance such as replacing ECC memory modules that are reporting errors.

Finally, NumaConnect systems have an inherent ability to support large numbers of mass storage devices connected to every node in the system. These systems can be managed using Linux software RAID or other applications as parallel file systems. In addition, all storage devices are directly accessible to all applications, as is the case in any SMP system. With NumaConnect, however, the amount of total storage can be much larger than just a single server. Since all applications see all memory and all storage, NumaConnect offers a very powerful I/O capability that has not been available to date.

Conclusion

NumaConnect demonstrates price, performance, and capability advantages that have been previously unavailable. From a cost perspective, those seeking to maximize their HPC invest-

ment should consider NumaConnect as way to deliver all existing compute capabilities while enjoying the ease of shared memory computing. The following conclusions should also be noted:

- NumaConnect provides "plug and play" shared memory computing. There is no extra software or configuration required; user applications "see" all cores and memory in the entire cluster.

- Shared memory systems offer easier programming (OpenMP) and maintenance than distributed memory clusters.

- Large OpenMP programs scale quite well on NumaConnect and have set a new record for benchmark size and number of cores.

- Standard MPI libraries and applications often run faster on NumaConnect than on comparable FDR InfiniBand clusters.

- A cost comparison to a small minimal FDR InfiniBand cluster shows comparable price points.

- One live system currently offers users over 1,700 cores with a 4.6TB single memory image. Several production systems are running, including those in University of Oslo in Norway, Statoil, and Keele University.

Ultimately we pose the question to HPC users who wish to focus on technical computing rather than on the details of their machines: ***If you can get scalable OpenMP and MPI performance, ease of programming, and ease of administration at commodity cluster price points, why limit yourself to an MPI cluster?***