The OrionX Constellation: Data Center
High Performance Interconnect

Evaluation Criteria

Stephen Perrenod, Ph.D.

Evaluation

The High Performance Interconnect (HPI) Market

With increasing adoption of scale-out architecture, cloud computing, artificial intelligence, and extreme-scale processing, High Performance Interconnect (HPI) technologies have become a more critical part of IT systems. Today, they represent their own market segment.

In this research report, we examine key evaluation criteria for HPI technologies. How do customers select a technology and a vendor? What are the key issues and tradeoffs? How does one match a given workload to an appropriate interconnect and topology?

Ethernet, InfiniBand from Mellanox, Intel's Omni-Path Architecture and proprietary technologies are covered in this study.

Distributed Nature of Modern Computation

Increasingly, both modern high-end enterprise applications and high performance computing (HPC) applications are best addressed with distributed systems, in order to satisfy overall performance, scalability, and price/performance considerations.

Decades ago, shared-memory symmetric multiprocessing (SMP) parallel systems replaced single CPU systems. This was followed by a move to distributed clusters and massively-parallel processing (MPP) systems in large part motivated by advances in microprocessor technologies and the advent of low-cost large volume systems that could be used as building blocks for large systems.

These have steadily displaced SMP systems for over 20 years in the HPC field and, especially during the last 10 years, in enterprise computing applications as well. This trend has been accelerated by the emergence of scale-out application architecture and the onslaught of important scale-out applications in social media, cloud architectures, and Big Data.

Both scale out and scale up are deployed in combination for commercial and HPC apps, but the architecture of choice tends to have large numbers of 'skinny' nodes, with typically 2 or perhaps 4 CPU sockets per node. Architectures of this type already support high degrees of intra-node

OrionX Constellation™ reports cover 5 Es: industry milestones (Events), historical view of a technology segment (Evolution), main vendors in a market segment (Environment), customer decision criteria (Evaluation), and how vendors in a segment score (Excellence) based on the OrionX methodology, which considers market presence and trends, customer needs and readiness, and product capabilities and roadmap.

parallelism with shared memory and dual CPUs each with 16 or more cores are now the state of the art. General-purpose many core architectures with over 50 cores are imminent.

As the node count increases, and as applications look to scale to more and more compute power, the connectivity between distributed memory nodes is of increasing importance. Communication patterns can vary greatly and can be highly arbitrary; there is no guarantee that the bulk of the communication for an application will be within its own node, or only with its nearest neighbors. As the inter-node communication pressure on the interconnect increases, the requirements for availability and performance increase concurrently.

Inter-node communication must be reliable, it must be of high bandwidth and it must have sufficiently low latency. Achievable performance should be high for both large message transmission and for small messages.

For enterprise applications, homogeneous, multi-tier, and heterogeneous workflow applications put significant demand on network performance and topology. Deployment of peer-to-peer and service-oriented/network-oriented applications in a production environment cannot be done without the ability to manage complex communication, security, availability, and quality-of-service (QoS) requirements.

For HPC applications (and increasingly for enterprise analytics applications), system throughput matters, but so does, especially, the ability to decompose and map single problems across a very large number of nodes. Problems are broken up into logical subunits, computation is performed, and results gathered. The load on the system interconnect varies with the phase of computation. For example, the computation of the so-called “solver” portion of some codes typically requires very heavy inter-node communication.

HPC simulations often have three-dimensional or even higher dimensional grids or algorithms iterated forward in time with very large numbers of elements. These can run into the millions and billions of cells or grid points or particles. As the grid size shrinks, so must the time step, so a factor of two reduction of the grid size can translate to 16 times as much work. These millions of elements must be mapped onto hundreds or thousands of nodes in the system.

A solution on, say, a 3-D grid, after decomposition, can require heavy amounts of communication across the boundaries of the subgrid contained within each given node and the subgrids in the neighboring node-s.

Enterprise computing already uses many HPC-inspired technologies to achieve performance and scale. As Big Data analytics, machine learning, and business optimization become more common, High Performance Interconnect technologies are poised to become a standard part of scalable computing for the enterprise and for HPC alike.

Requirements of High Performance and Scalable Applications

The main categories of interconnect for distributed compute clusters are:

a) proprietary

b) Ethernet, and

c) InfiniBand and derivatives.

Evaluation of a proprietary interconnect can be difficult to disentangle from the overall proprietary system evaluation. Suffice it to say that the interconnect must scale gracefully, it must be reliable, and it must perform well for transmissions large and small. These are the same requirements that Ethernet or InfiniBand interconnects must meet, so it is natural to compare proprietary interconnects to those industry standard alternatives.

Ethernet and InfiniBand

Ethernet is ubiquitous, and its bandwidth has steadily increased over the years, but it suffers from high latency relative to proprietary alternatives and particularly in comparison to InfiniBand. Currently, Ethernet bandwidth peaks out at 100 Gb/s, with 400 Gb/s under development. These are quite respectable rates, and very large switches are available with capacities of the order of 100 Tb/s, supporting hundreds of wire speed ports. However, achievable latencies are another matter. Ethernet switch vendors are proud when their roundtrip TCP or UDP latencies are as low as 3 microseconds, while InfiniBand latencies can be significantly below 1 microsecond.

RDMA over Converged Ethernet (RoCE) is a protocol designed to provide efficient low-latency RDMA (Remote Direct Memory Access) transfers in Ethernet fabrics. The lowest achievable latency with RoCE is 1.3 microseconds, reported by Mellanox.

However, “try as it may, Ethernet cannot kill InfiniBand” said The Next Platform in April, 2015:

“With the ConnectX-4 InfiniBand adapters that Mellanox is sampling during the first quarter for deliveries later this year [2015], this EDR InfiniBand card has been tested to have a bi-directional throughput of 195 Gb/sec and an application latency of 610 nanoseconds.”

Application level latency is the name of the game, especially in HPC, but increasingly in commercial applications as well, including finance (real-time trading), Web 2.0 and real-time transaction, image, and voice processing, and for analytical processing.

Offload and Onload

The question of exactly where to process the networking protocol has led to distinct approaches to designing and building switches. If the processing is performed wholly or substantially on the network switch, it is referred to as an “offload” design since the processing is offloaded from server CPUs. By contrast, in an “onload” design, protocol processing is performed on server CPUs. At a high level, this leads to a trade-off between a more capable network switch where server CPUs can be allocated fully to server workloads (offload), vs. a less capable network switch and a reliance on server CPUs to handle not only server application workloads but also network processing (onload). In addition, onload can add latency since the network switch must wait for server CPUs to accomplish its tasks.

One of the major performance disadvantages of Ethernet is the connection-oriented protocol processing burden it places on the host CPU. Just 10 Gb/s of TCP network traffic can consume 4 cores of a 2.5 GHz processor. TCP offload engines (TOE) are a limited solution, more of a patch, for this problem is inherent in Ethernet’s design.

InfiniBand, which has been a standard since 1999, was expressly designed to offload network processing from the CPU as much as possible, and onto the Host Channel Adapter (HCA) hardware. InfiniBand achieves high bandwidth and low latency in part because of its ability to offload the server CPU and place protocol processing onto purpose-built ASICs located on switches and HCAs. RDMA is an intrinsic part of the InfiniBand design approach. With RDMA, data is moved between two distinct physical memories with little or no CPU involvement, reducing overhead by a factor of 10, approximately.

Omni-Path, a proprietary interconnect that is a derivative of InfiniBand, departs from the original InfiniBand philosophy by reinserting the CPU into the network processing pipeline.

It can become misleading to look at only the network performance when comparing an offload architecture with an onload architecture. The onload architecture might be pushing many messages down the pipes, but the tradeoff is that the CPUs are now much more consumed with networking functions, and less available to work on the actual user application. One must look at actual application performance and how CPU cores are being used – working on an app, or processing network functions.

“In-Situ Processing” and Co-processing

While onload/offload refers to some of the network protocol processing, other tasks at higher levels are also eligible for consideration. In such cases, the discussion and trade-offs are more in the realm of co-processing and the larger IT trend towards performing tasks where the data happens to be.

This is what we call “In-Situ Processing" where processing power is distributed across the data center to provide genuine data processing right where the data is. Often this capability simply augments the mechanism for control, which is already required, and thus not a completely new addition.

A prime example of higher-level tasks that can be performed by the network or by server CPUs in the HPI market is Message Passing Interface (MPI) system. Several MPI constructs not only direct data traffic, but also perform non-trivial tasks that are best handled as part of data communication. Performing such tasks within the network can provide significant latency advantages as explained below.

In general, the option to perform tasks in the network is an important factor in reducing latency today and increasingly so looking forward.

A recent study by Dosanjh et al. at the University of New Mexico and at the Sandia National Laboratories (presented at the 2015 IEEE International Conference on Cluster Computing) evaluated the performance of a Mellanox offloading 4X QDR adapter with a QLogic 4X QDR HCA using onloading. Their tests employed a 4-node system connected to a QLogic switch for all adapter tests. QLogic InfiniBand is a predecessor to Omni-Path that also pursued an onload design strategy.

The authors evaluated two applications, MILC and LULESH. MILC has almost 3000 MPI library calls per second, compared to just over 500 for the LULESH application. For LULESH, there was little difference in performance. However for the more intensive networking application MILC, the offload Mellanox HCA configuration provided from 7.7% to 10.6% better application performance than the alternative. They conclude that codes with large numbers of small communication calls will see more benefit from offloading.

The authors make the point that many core architectures may be less suited to onload networking since they tend to have simpler CPU core implementations and lower clock rates, and thus will have a harder time processing the networking functions. Furthermore they note that MPI (Message Passing Interface) code is highly serial and difficult to parallelize across cores. MPI is overwhelmingly the messaging library of choice in HPC, and can be considered as the uppermost layer of network processing required in support of highly parallel applications.

"Inter-node communication must be reliable, it must be of high bandwidth and it must have sufficiently low latency"

Switching

Networks that scale to moderately large node counts can be supported with one or two switches, but networks that scale to very high node counts require hierarchical architectures. Typically a leaf and core configuration will be used – one or two leaf switches (top-of-rack or ToR) will connect nodes within a single rack, a subcluster if you will. And above that first layer, some number of core switches will connect all of the subclusters together.

Oversubscription is often utilized to reduce the aggregate number of switches and system cost, by varying the ratio of upstream channels to downstream channels. Oversubscription is typically an important consideration in system design and costing. Two-to-one (2:1) and 4:1 oversubscriptions are common, with two or four times as many upstream channels as in a fully non-blocking configuration. These can be reasonable choices since workloads may be localized to various degrees, both for multi-job throughput work and for highly parallel single jobs.

Topology is another important consideration since it determines the degree of connectivity across the entire system and within subclusters. The topology dictates how many hops occur for end-to-end communication across the entire system.

InfiniBand and Omni-Path provide a switched fabric that can be used with multiple switches and hierarchical architectures to support various topologies. Commonly implemented topologies include toruses, typically 2-D or 3-D but also higher dimensions, and multilayer fat trees with various bristling ratios. Grid and hypercube topologies are some other possibilities, although less common these days. The appropriate choice will depend on the expected workload, the details of the switches deployed, node characteristics, and cost tradeoffs.

Interoperability: LANS and Storage

A compute cluster is not an isolated system; it must be part of a wider data center architecture that includes support for various forms of storage and for local area and wide area networks communicating with other shared resources in the organization. Thus support for other protocols is critical, especially Fibre Channel and iSCSI for storage and IP for networking.

Ethernet over InfiniBand allows encapsulated Ethernet packets to move across the InfiniBand fabric. This enables transparent communication between a compute cluster and LANs and also enables InfiniBand to play an important role in converging infrastructure. Convergence improves efficiency and economy of operations by reducing the management burden and by allowing for fabric and cable consolidation in the data center. This can be very helpful in virtualized environments, since some of the processing for guest OSes can be offloaded to the fabric, freeing more compute resources for the VMs.

Big data, cloud, software-defined storage, database clusters, converged storage, software-defined storage, and the use of flash SSDs are major trends driving modern scale-out storage architectures. Traditional Fibre Channel SANs are being replaced with scale-out storage running on Ethernet or InfiniBand networks. Cluster-oriented file systems including Lustre, IBM Spectrum Scale (GPFS), VMware SAN, EMC ScaleIO, Hadoop and other solutions can provide the bandwidth necessary to move data in and out of large compute clusters at lower cost. Hyper-converged infrastructure can combine compute and storage in the same networking layer for simplified management.

InfiniBand supports RDMA connections to SCSI and Fibre Channel based storage systems. Fibre Channel storage can be supported with the RoCE extensions and FCoE (Fibre Channel over Ethernet). iSER provides RDMA extensions for iSCSI, in order to provide the benefits of block transfers with low latency and high bandwidth to iSCSI storage targets. OpenStack (Cinder) can be used in conjuction with iSER for cloud environments.

The Intel Omni-Path Architecture (OPA) fabric uses OFA (Open Fabrics Alliance) software. A single OFA software environment can support Mellanox InfiniBand HCAs and Intel OPA HFIs. Storage connections can be routed over IB or Ethernet to storage servers running NFS or parallel file systems including Lustre and GPFS.

Implementation

Primary components include switches, gateways, adapters, cables, and related software.

If one chooses a proprietary vendor interconnect (e.g. Cray Aries or SGI NUMALink with shared memory support), then the implementation of the full stack will typically be from a single vendor, including software components, the proprietary switching integral to the system, and gateways, adapters and cables. Oracle integrates InfiniBand into its engineered systems. Cray and SGI also offer InfiniBand interconnects for their cluster systems.

For InfiniBand and Omni-Path, the primary vendors for these components are Mellanox and Intel, respectively, although there are many cable vendors in the IB space. The cluster node provider (e.g. Cisco, Dell, HPE, Lenovo, Oracle, major ODMs) will often provide the Mellanox InfiniBand or Intel Omni-Path components directly and integrate those for the buyer.

Ecosystem

Two important standards bodies for high-bandwidth low-latency fabrics are IBTA (the InfiniBand Trade Association) and OFA (the Open Fabrics Alliance).

The IBTA was founded in 1999 to promote the InfiniBand architecture and RDMA implementations including over Ethernet with RoCE. The steering committee includes Broadcom, Cray, HPE, IBM, Intel, Mellanox, Microsoft, Oracle and QLogic. The IBTA sponsors interoperability testing of commercial products and publishes an Integrators' List of compliant products covering HCAs, SRP storage targets, switches and cables, and details of interoperability tests.

The Open Fabrics Alliance was founded in 2004 (as the OpenIB Alliance). Its vision is "to deliver a unified, cross-platform, transport-independent software stack for RDMA and kernel bypass" so that users can transparently run apps over InfiniBand, RoCE, iWARP and other fabrics. Alliance membership includes a large portion of the major players in enterprise computing and high performance computing. The Open Fabrics Software (OFS) is open-source code for RDMA and kernel bypass applications. Fabrics/networks that are supported by

OFS include Ethernet, iWARP for Ethernet, RoCE, and InfiniBand. OFS is available for many Linux and Windows distributions.

Maturity

The proprietary interconnects such as those from Cray and SGI are highly mature with long histories as technically robust solutions. Cisco is an example of an Ethernet-oriented cluster provider that is the leader in the networking space, and has also provided integrated compute cluster solutions for seven years. Cray and SGI have been around since 1972 and 1982, respectively, and have been leaders in high performance computing systems and highly scalable commercial systems requiring high bandwidth, low latency, highly scalable interconnects for decades.

InfiniBand is a mature technology dating back into the past century and with market support from a broad range of system and component vendors. Mellanox (the original provider of InfiniBand networking silicon) is a mature technology provider for InfiniBand chips, adapters, gateways and highly scalable switches as well as the software components required for these. Mellanox InfiniBand is used in 45 percent of the world's petascale computing systems found in the upper reaches of the TOP500. Almost half of the TOP500 systems currently use InfiniBand.

Intel acquired QLogic's InfiniBand assets in 2012. The same year, Intel also acquired Cray's Aries interconnect IP and development team. Omni-Path is a new offering with a small, but high profile, customer base in the U.S. and Europe.

Evaluation Criteria

We divide customer evaluation criteria into three broad categories:

- Market
- Product (technology)
- Customer (adoption)

Market

It's important to evaluate how long a technology has been around and how widespread its utilization is. One also should look at how long a particular company has been providing that technology suite. An evaluation will also take a look at the installed base for a particular company's products in the interconnect space.

Market maturity is greatest for the proprietary interconnects from Cray and SGI because of their long histories and proven track records, and for Ethernet from Cisco, due to its long history and very large installed base as the major network standard for LANs.

After that, InfiniBand as provided by Mellanox, has the longest pedigree for standards-based high-end interconnect technology. It also has a very respectable installed base in HPC and commercial applications.

Intel's Omni-Path is the new kid on the block. It has some pedigree by inheritance from the QLogic InfiniBand technology acquisitions. But it has moved away from the InfiniBand standards and philosophy (e.g. by returning networking functions to the CPU), and has currently only a handful of customers, so the jury is still out.

Product and Technology

Ethernet switches with up to 100 Gb/s bandwidth and support for 25 Gb/s and 50 Gb/s channel rates are now available, allowing for significant enhancements in bandwidth relative to traditional Ethernet cluster environments, but latencies remain relatively high. TCP latencies usually exceed five microseconds.

The latest EDR technology for InfiniBand and Omni-Path supports bandwidths with 4X links aggregating to 100 Gb/s. The IBTA roadmap indicates 200 Gb/s 4X bandwidths in the 2017/2018 timeframe. Latencies are of order one microsecond or less. Message rates for both technologies are of the order 100 million messages per second.

One can evaluate interconnect options based on the specifications and benchmarks for bandwidth, latency, message/packet rates, and also packet loss/retransmission rates. One can also look at CPU utilization rates – how many cores on average of a multicore or many core processor are required to process network functions, and are therefore lost from the pool of cores used to actually run applications.

One also must look at the switches available, in terms of size and configurability and take into consideration the appropriate topology for a given application suite and environment. Fewer hops across a hierarchical switching fabric is desirable in order to cut end-to-end latency. Using fewer, but larger, switches for the same node count leads to fewer hops, but can be more expensive.

Ultimately, what matters are the application requirements. Some applications are more latency sensitive, some more bandwidth sensitive. In addition the efficiency will vary as the node count is increased. The real question is how do the major applications perform on a given interconnect configuration? The evaluation should consider all of the I/O functions for the applications and account for OS and network processing overhead and overall software performance considerations. Are the applications compute bound, I/O bound, or interconnect bound? How do applications actually scale as node count increases? Where is the sweet spot in node count beyond which the utilization of additional resources has limited benefit?

Customer Adoption

There are a few considerations here. One is how widespread is customer adoption for the technology and vendor under evaluation? Another is does the vendor have presence in your industry? Are customers with similar workloads to yours adopting technology from the vendor in question? Have they moved away from Ethernet, or not? Are there barriers to adoption, and how have others overcome those? How important is incumbency?

Customers' ability and readiness to adopt a given technology is ultimately the most important evaluation criteria. While features, benefits, maturity, and economics of a solution have a large influence on customers' decision, so do costs to retire existing technology and implementing the new one, finding or building skilled staff, and current and anticipated needs.

HPI technologies are complex and their behavior can vary under different conditions and for different use cases, making the selection process more involved.

Economic considerations

What matters most is the achievable application performance and system throughput for a given total cost of ownership (capital cost plus operating expense), plus considerations of ease of upgradeability, ease of operation, and reliability.

The onloading vs. offloading debate comes into play here. With an offload-based architectural design as found with InfiniBand, a somewhat smaller number of CPUs may be needed, since the network processing load on CPUs is reduced. Fewer CPUs can also mean fewer switches.

One must look at the cost tradeoff for the total system in aggregate with the two approaches (onload with OPA vs. offload with IB) including the system nodes plus network cards and switches for two different architectures that ultimately would have the same application performance.

In addition, one needs to look at the reasonable topology options for a given node count and then look at how many and which size of core and leaf switches would be required. Then one adds in the cost of the necessary adapters and cables to complete the configuration pricing, along with the aggregate cost of the compute nodes and storage associated with the system. There may also be some routers and gateways that are desired for communication with storage and LAN resources. Software for the interconnect will generally be bundled or open source so may not be an additional capital cost item.

For all of the various options the annual service costs, including software, must be examined and 3-year or 5-year TCO costing done.

This is an optimization problem. One trades off the number of nodes and switches and the topology in order to maximize system throughput and application performance for a given budget. What percent of the CPU cores will be doing real application work rather than simply network processing?

It may even be that the fat tree configuration favors one vendor, and the 3-D torus favors another, for example. So the optimization is across node types, switches types and topologies, and compares different suppliers (e.g. Intel with OPA, Mellanox with IB, and Cisco with Ethernet). There can also be a capital vs. operating cost tradeoff as service and other operating costs including power, cooling, floor space and system administration costs are incorporated.

This is a multi-dimensional space and an iterative process is necessary to converge on the best reasonable or available solution within a given budget framework. Interconnect technology can represent over one quarter of the total capital cost of a large-scale cluster.

Summary, Conclusions

InfiniBand (IB) offers substantial advantages over Ethernet as a proven standards-based high bandwidth, low latency interconnect with RDMA inherent in the stack. InfiniBand is fully scalable, to tens of thousands of nodes.

Omni-Path Architecture (OPA) has its roots in InfiniBand, and while it has moved in a proprietary direction, can be an interesting alternative to evaluate. One of the major issues is the offloading philosophy (as for IB) versus the onloading approach (used by OPA). One needs to look at full application performance, not just networking kernels, in order to properly evaluate the two alternatives.

Both IB and OPA are suitable for the implementation of converged architectures incorporating Ethernet LAN functions and communication with storage systems. They both support Ethernet encapsulation and a variety of storage protocols and parallel file system interfaces. Both support the Open Fabric Software stack.

Both offerings support a range of switch sizes, including director class switches, allowing for robust hierarchical topologies such as fat trees or 2-D and 3-D toroidal configurations. A full economic analysis would examine different topologies and component switch options, with different degrees of oversubscription. This analysis should be done in light of the actual application requirements since different applications will put stress on the interconnect infrastructure in different ways and to varying degrees.

The cost of adapters, cables, and required gateways and annual service and software licenses should also be included.

InfiniBand has the established pedigree and substantial installed base for High Performance Interconnects, including many of the largest computer systems in the world, thus it is the first option to consider if Ethernet is deemed to be insufficient to the cluster workload requirements.

This is the third paper in a four-part series examining the HPI market. The fourth paper in this series is "Excellence" and discusses our evaluation of the industry players and our recommendation for customers looking at HPI interconnects today.
Please visit OrionX.net/research for additional information and related reports.

The information contained in this publication has been obtained from sources believed to be reliable. OrionX does not warrant the completeness, accuracy, or adequacy of this report and bears no liability for errors, omissions, inadequacies, or interpretations of the information contained herein. Opinions reflect the judgment of OrionX at the time of publication and are subject to change without notice.