The OrionX Constellation: Data Center
High Performance Interconnect

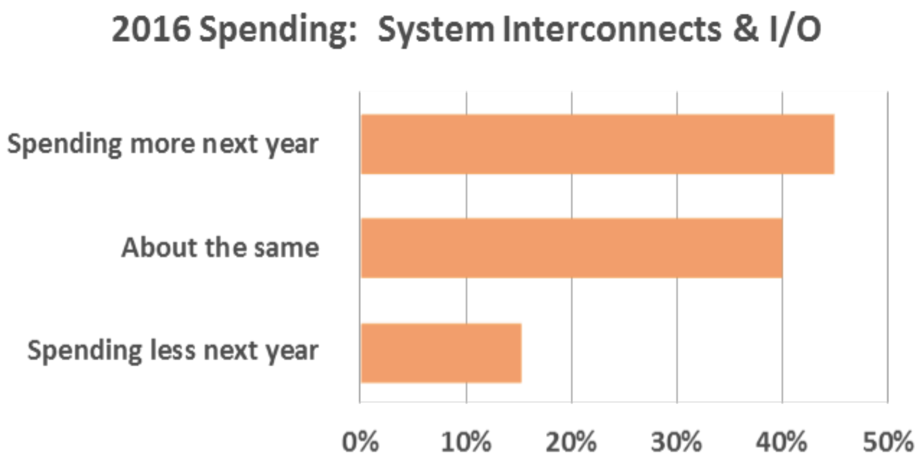
Market Landscape

Dan Olds

Environment

One of the most important components of a high performance computing solution is the interconnect that ties the resources together. In fact, interconnect speeds have advanced at a rate of 30% annually over the past four decades. This compares well with Moore's Law, which indicates 41% annual improvement. Thus the increase in interconnect performance has had, and will continue to have, a large impact on the performance of high performance systems.

In late 2015, we surveyed 175 high performance computing (HPC) and large enterprise data centers. In one question, we asked them about their spending priorities for 2016.



Given the role that interconnects play in system performance, it's not a surprise to see that the majority were planning to spend more money on interconnect technology in 2016.

There are two broad types of high performance computers today: MPP and clustered systems. Massively Parallel Processing (MPP) systems are often designed with a bias towards cases when a single application must run well over the entire set of nodes, which may number into the thousands. A clustered system is less constrained. It is typically used either in segments and to run many different applications at the same time, or runs well behaved embarrassingly parallel applications. Clustered systems can also utilize thousands of nodes.

The main difference is the interconnect. MPP systems tend to use the highest performing interconnects, including custom proprietary interconnects, while clustered systems pick from a wider set of choices.

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.

So where will they be spending their money? In this report, we run down the major players in the High Performance Interconnect (HPI) market.

The HPI market is the very high-end of the networking equipment market where high bandwidth and low latency are non-negotiable. It started out as a specialist proprietary segment but has blossomed into an indispensable, large, and growing area. Products in this category are used to build extreme-scale computing systems. They are typically not used for traditional telco, enterprise, or service provider networking needs.

In this report, we'll take a look at the technologies, market presence and focus, plus highlight some of their high-end technology. First up, let's take a look at high performance Ethernet players and technology.

Ethernet Providers

There are a number of companies that provide high performance interconnects based on Ethernet technology, In addition to large players such Cisco, HPE, Juniper, Brocade, and Hauwei, there is also Arista, Extreme Networks, Dell (Force10), IBM, Mellanox, and several others.

Ethernet is lower performance than other HPI alternatives so it occupies the low-end of the market. In addition, while Ethernet vendors have different levels of interest in the HPI segment, no clear leader has emerged in this segment. As a result, when it comes to assessing vendors, it is possible and reasonable to treat Ethernet vendors as a single virtual vendor. In this report, we take a closer look at the bigger players with the widest array of Ethernet products, as a proxy to all the Ethernet players in the HPI segment.

Cisco is the largest networking product firm in the world, with roughly 60% of the overall Ethernet switch market and overall revenues of $50 billion. While Cisco has products that are used as Ethernet HPI mechanisms, the company primarily aims their products at the enterprise market, which has different performance and application demands than the typical HPI user.

Hewlett Packard Enterprise (HPE) is also pursuing the enterprise Ethernet market, as well as the HPI segment. The company has a full complement of products along with skilled engineers (HPE acquired 3Com for $2.7B some years ago). HPE's networking unit now owns close to 10% of the Ethernet switch market and has been growing quickly.

Juniper Networks is another "short-list" competitor in the Ethernet market. The company, founded in 1996, is a $4.8 billon provider of enterprise, service provider, and HPI products.

Technology Strategy

Cisco is focusing on enterprise network products primarily, but is also concentrating on products that serve a wide variety of business needs beyond the data center. This includes consumer products, telepresence systems, call center equipment, and even television set top boxes. Cisco also builds their own x86 servers, the highly regarded Cisco Unified Computing System (UCS) converged servers that disaggregate networking and power/cooling from compute blades, simplify virtualization, and combine compute, storage, and network into a single highly efficient package.

HPE's network group is tightly focused on midrange to high-end enterprise and high performance networking products. This includes products that span the range from wireless LAN equipment up to service provider WANs.

Juniper Networks is also highly focused on providing a wide range of networking equipment, but concentrates a bit more on the very largest customers than either Cisco or HPE. Juniper also provides a set of network appliances and software designed to increase data and network security.

Technology Highlights

Cisco's largest offering in the cluster interconnect space is their Nexus 9516 switch. This product offers a maximum of 128 100GbE ports, and an overall maximum throughput of 60 Terabits/sec when configured with 100GbE ports.

HPE's FlexFAbric 12900 switch has a modular design that offers up to 57.6Tb/sec switching capacity when fully configured. The full system offers 64 100GbE ports.

Juniper's QFX100016 switch offers up to 96Tb/sec throughput and a total of 480 100GbE slots in a single system.

Cray

CRAY®
THE SUPERCOMPUTER COMPANY

Cray is synonymous with supercomputing and has steadily become an even larger player in the broader HPC market. In the last five years, Cray has expanded its product line into clusters, analytics, and storage; made acquisitions, and increased revenues from $236 million in 2011 to $725 million at the end of 2015. More importantly, the company achieved this revenue growth while remaining profitable, albeit with operating margins of 6% and net income of around 4%. Still, the company is healthy, growing, and has built a track record of solid results.

Cray's biggest market segment is supercomputing, the very high end of the HPC market, where it serves customers in areas such as research, academic, governmental, financial services, weather forecasting, energy. The company is also working to extend its reach into traditional enterprise clients who find that their computing needs require the kind of high-end design that Cray puts into its products.

Technology Strategy

Cray uses commodity computing components as building blocks for their supercomputers while offering customers a choice of interconnects. Customers can select standard Infiniband in a variety of speeds, or they can choose from two of Cray's custom interconnects – Gemini or Aries. Both of these interconnects exhibit scalability to the greatest heights available today.

Technology Highlights

Cray doesn’t disclose performance numbers or configurations for their proprietary interconnects. But we can draw some information by looking at Cray’s entries on the TOP500 (which lists the 500 largest computers in the world) list. Cray’s largest system (#3 on the June 2016 list) is called Titan. Titan is a MPP system (meaning it is designed to allow one application to efficiently use the entire machine) and it uses Cray’s Gemini interconnect to glue together 18,688 nodes.

At number seven on the list, Cray’s Trinity system uses their Aries interconnect to connect 9,436 individual nodes. This is another MPP system. When it comes to clustered supercomputers (systems which are designed to run multiple applications simultaneously), Cray tends to use InfiniBand interconnects.

Intel

Intel is best known for their processors and associated system technology. Founded in 1968, Intel has grown into a $55 billion company with approximately 107,000 employees worldwide. Revenue growth in the last five years has slowed to a crawl, growing on average about .5% since 2011. In recent news, Intel has re-planned its mobile processor strategy and is shedding more than 10,000 employees. However, the company is highly profitable, with gross margins of 62% and net margins of 21%.

Intel’s biggest market for processors has traditionally been the consumer and business PC market while seeing significant growth in server markets (small/large enterprise plus HPC). System interconnects are a new market for Intel, coming about after it purchased InfiniBand assets from QLogic. Intel’s interconnect product set is marketed as Omni-Path Architecture (OPA). Intel also acquired relevant intellectual property from Cray. The impact of that acquisition is slated to become visible in the future. At this point in time, Intel is really just a well-funded startup when it comes to the HPI market and must continue to prove itself.

While Intel has the smallest market share in high performance interconnects, it has successfully parlayed its strong presence in CPUs into securing a few high-end installations for OPA (among them Texas Advanced Computing Center and Pittsburgh Supercomputing Center) that should make future TOP500 lists.

Technology Strategy

Over the years, Intel has added more and more functions to its chipset – eventually incorporating everything from I/O ports to RAID controllers to audio. It is now looking to do the same with servers. Intel arguably controls all of the crucial technology inside the server. By adding high performance interconnects to their portfolio, it is expanding its reach to a rack of servers. Indeed, Intel’s impressive rack scale architecture promises to disaggregate compute, storage, and network resources and re-aggregate these resources for more efficient utilization.

Omni-Path Architecture is similar to QLogic’s product set and continues the different architectural approach that QLogic provided compared to other HPI products, namely, utilizing an “on-load” mechanism that uses the main server CPU to process networking operations.

There is historical and current controversy around where exactly to process what task. For example, if processing is not latency-constrained and/or is amenable to parallelism, it might be worthwhile to move the processing to wherever CPU or GPU cycles are available and bring the results back. On the other hand, if processing is possible where the data is already residing, it would be efficient to perform the tasks as close to the data as possible. Arguments are made in favor of each approach, but there is a growing trend across the IT industry towards performing tasks where the data happens to be.

While onload/offload refers to some of the network protocol processing, other tasks at higher levels are also eligible for consideration. When these higher-level tasks are considered for processing, the discussion and trade-offs are more in the realm of co-processing and the larger IT trend towards “in-situ processing”.

A prime example in the HPI market is Message Passing Interface (MPI) message-passing system some of whose constructs require non-trivial serial processing. Performing such tasks within the network can provide significant latency advantages.

Technology Highlights

Intel’s premiere Omni-Path high performance interconnect product is their Fabric Director Switch 100 Series. It supports up to 768 ports in a 20U design, with an advertised aggregate bandwidth of 153.6 terabits/sec, which breaks down to 100Gb/s per port.

The Director also comes in a 7U, 192 port version, plus a customizable switch for OEM needs that can support up to 48 ports.

Mellanox

Since its founding in 1999, the company has been focused exclusively on network connectivity products, including InfiniBand and Ethernet. The company is headquartered in both Yokneam, Israel, and Sunnyvale, California, and has 2,600 employees.

Over the last five years, Mellanox revenues have grown at a 34% CAGR, with 2015 revenue coming in at $658 million. The company is profitable with gross margin of 72% and net income of just over 14% for their 2015 fiscal year. A little over 68% of the company’s revenue was derived from InfiniBand products, while 24% came from sales of Ethernet based products.

Mellanox has a large presence in the high performance interconnect market. Almost half of the TOP500 list use Mellanox InfiniBand interconnects, making Mellanox the interconnect vendor with the largest presence on the list.

Today, Mellanox primarily competes in the supercomputing market, serving customer in energy, financial services, research, academia, and pharma. The company is making inroads into enterprise markets including Big Data, Analytics, and Webscale markets as well.

Technology Strategy

Mellanox sees system interconnects and networking hardware as perhaps the most vital component to spurring future system performance and scalability. The company is dedicated entirely to high performance networking.

While the company sells Ethernet products, Mellanox selected InfiniBand as their premiere high performance interconnect technology. Mellanox pursues an “offload” and “In-Situ Processing” architecture, where network adapters and switches handle the network processing and higher level tasks that are best performed right on the network instead of on server processors. This gives Mellanox the high ground in the race to provide higher bandwidth, lower latency, and more intelligent interconnect products.

The company was first to market with their 100Gb/s EDR InfiniBand products and is currently leading in that space. They were also first to market with 25/50/100 GbE products. Additionally, they’ve been a pioneer in smart multi-mode switches that can run both InfiniBand and Ethernet protocols in the same box.

Mellanox is also a leader when it comes to their roadmap. They are promising 200Gb/s InfiniBand in 2017 and 400Gb/s technology a few years after that.

Technology Highlights

Mellanox switches scale from 8 ports to 648 ports at up to 100Gb/s per port. These switches can support thousands of nodes, and can be configured to ensure granular quality of service for clusters, LANs, and SANs.

SGI

SGI is another company that is closely identified with the supercomputing market. SGI has a long and distinguished history in several market segments including the HPC and In-Memory Computing markets.

Over the past five years, revenue for SGI has dropped from a peak high of $767 million in 2013 to $521 million in 2015. The company has been generally unprofitable during this period despite a respectable product portfolio and overtures into the commercial enterprise market which continue to look promising.

As a well-established but smaller competitor in the HPC market, SGI has 26 systems on the June 2016 TOP500 list. SGI’s unique large-memory x86 products are also used as part of the SAP HANNA ERP analytics solution, which are resold by HPE and Dell.

Technology Strategy

SGI is striving to be a full-service HPC vendor in a ruthlessly competitive market. For the most part, they use commodity products to build out their supercomputing solutions. SGI has their own proprietary SGI UV advanced NUMA shared memory interconnect, but it is integrated with and offered only with the purchase of a SGI UV system. It is not available as a stand-alone product. The company doesn’t disclose performance statistics for this interconnect.

However, SGI has used InfiniBand technology exclusively in their computers that landed on the June 2016 TOP500 list.

Today's Choices in High Performance Interconnects

There is really only one broad choice in HPI technology today: Mellanox's InfiniBand. Intel's OPA cannot be ignored, and its early success in some very high-end systems makes it, at a minimum, a technology to watch closely and likely evaluate.

Given that most every system vendor in the market today is partnered with both Intel and Mellanox, and has ready access to Ethernet switches, customers are going to be able to choose between multiple interconnect architectures.

In our opinion, the right choice for customers in the HPI market today is InfiniBand. While both InfiniBand and OPA technologies boast roughly the same performance numbers, there are some real differences in the implementation and maturity of their architectures that we believe will make a significant difference when used at scale. For example:

- Offload vs. Onload, and In-Situ Processing: Because the requirements for low latency is paramount in the HPI space, we believe that the demanding and serial nature of some of the protocols and parallelism constructs (data reduction, eg) currently favor an offload/In-Situ processing strategy – particularly at scale.
- Customer Experience: There are thousands of Mellanox installations running today in both small and extremely large clusters. Conversely, there are very few Intel OPA installations up and running. Most organizations don't have the time or money to spend pioneering a completely new technology.
- Market Maturity: Hand in hand with the above point, there are very large numbers of service organizations, vendors, and skilled workers who have deep experience with InfiniBand. The same cannot be said for Intel's OPA at this time.

Summary

InfiniBand continues to be the de-facto choice for a high performance interconnect today. It offers the highest performance standards based technology, has proven that it can perform at scale, and has a track record with some of the largest installations in the world. It also offers an architecture that can scale better and offer predictable performance in any situation.

This is the second paper in a four-part series examining the HPI market. The third paper is "Evaluation" and discusses customer evaluation criteria for high performance interconnects.
Please visit OrionX.net/research for additional information and related reports.