OrionX.net

Evolution

The OrionX Constellation: Data Center
High Performance Interconnect

The Emergence of High Performance Interconnects

Peter ffoulkes

Computer technology has been evolving to improve workload performance since its inception. Multi-processor architectures have been used to improve application performance since the 1960s which has significant impacts upon system and software design as well as the orchestration required to coordinate communication between processors, memory, storage, networks and any other component that affects workload performance.

The larger or more complex the workload the more important it is that the components of the system are well balanced and do not create a choke point that impairs expected performance. That balance and associated choke points have changed over time and will continue to do so.

System and Architecture Components

Server Processors

Server processor designs have changed radically over the years with the emphasis shifting from one performance attribute to others each time an issue was identified and addressed. Today, mainstream processor architectures emphasize floating point performance, multiple processor cores, multi-threaded performance, multi-layered cache, large memory support and energy efficiency. Future System on a Chip (SoC) designs promise to bring more capability to silicon level integration by incorporating Field-Programmable Gate Arrays (FPGAs), Digital Signal Processing (DSP), stacked memory and other novel technologies. Although advances at the processor level will always be required, the end of Moore's Law has caused much of the attention to shift to other aspects of system architecture for the near to mid-term future.

System Architectures

The use of multiprocessor architectures whether closely coupled or loosely coupled to enhance performance, requires coordination between system resources to support a single application or to coordinate the execution of multiple simultaneous applications (a workload) on a shared system.

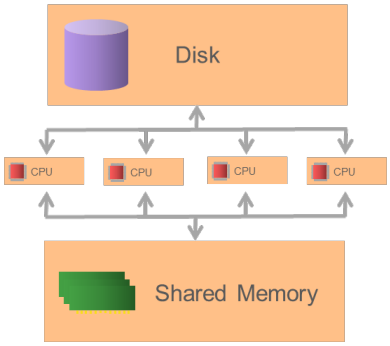
The essence of the problem and solution for scalable applications has always required a focus on the system interconnect and the transfer of information between the system components. The larger the system the

more difficult this becomes, and the greater the complexity of the workload the more critical it is to coordinate system resources efficiently. The interconnect technology becomes the focus and critical element of that optimization process in combination with the orchestration of available system resources.

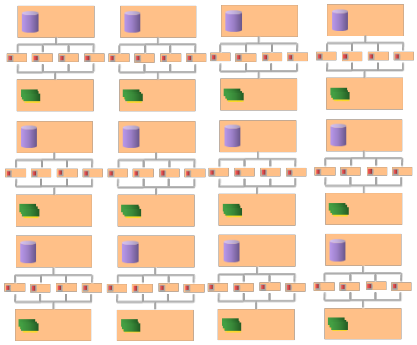
Tightly coupled multiprocessor systems allow two or more processors to share system resources, memory and storage under the control of a single operating system either symmetrically (or less often, asymmetrically, where control functions and applications functions are allocated to different processors.) The backplane of multi-processor systems provides the required interprocessor communication but leads to significantly higher system costs as the number of processors increases in a single system.

Computer Networks

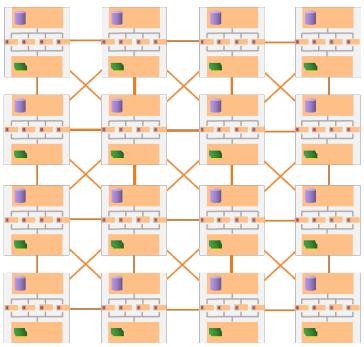
Commercial network technologies that connect multiple individual computer systems or nodes together have also evolved significantly since the 1960s, whether deployed as massively parallel processing (MPP) systems inside a single chassis or as distributed or clustered systems. These have the benefit of harnessing lower cost resources together than are typical of SMP systems but are heavily dependent on both the network performance and software orchestration to deliver efficient performance.



SMP System



MPP System



Distributed Cluster System

Distributed clusters are by far the most common format deployed for high performance systems, most typically based on x86 / x64 processor architectures. The individual systems are usually 1U, 2U or 4U servers with either two or four processors together with optional accelerator technologies to provide additional floating point performance. Intel-based systems currently use Intel's QuickPath Interconnect (QPI) designed for cache-coherent high bandwidth and low latency for single nodes.

In this situation the efficiency of the network interconnect and orchestration software between server systems becomes even more important to overall workload performance. Physical network connections have evolved beyond copper-based cables to optical fiber which benefits both performance and energy usage. Looking to the future, Silicon photonics technologies promise to deliver even more advantages in the coming years.

Network Technology and Protocols

Early network protocols evolved to connect distributed users to centralized computer systems. As with most computing technology, proprietary protocols evolved to enable communication between computers. IBM's Systems Network Architecture (SNA) and Digital Equipment's DECnet in the mid-1970s are common examples.

Leading edge developments typically began with proprietary technologies and become standardized in various ways as the industry evolved. The mid-seventies also saw the birth of Ethernet at Xerox PARC, today's ubiquitous standard for local area and commercial office networking. Along the way, a large body of knowledge has been gathered about protocol processing, network topologies, data traffic patterns, network services, and communication requirements for distributed applications.

The need for high performance interconnects emerged with the advent of massively parallel systems and scale-out computing. The significant gap between memory access vs. network bandwidth and latency required the highest performance interconnects. While networking technologies have progressed extremely well, that requirement remains. Today, high performance interconnects can be divided into three categories:

Ethernet

Despite competition from significant proprietary offerings including IBM's Token Ring and others, Ethernet emerged as the dominant low level interconnect standard for mainstream commercial computing requirements. Above the physical level, the software layers to coordinate communication remained proprietary until the OSI stack and TCP/IP filled the rest of requirements in a non-proprietary manner. Its layered design as well as its adoption in the client/server wave of computing that started in early 80s helped create a formidable ecosystem, resulting in TCP/IP becoming widely as the primary commercial networking protocol.

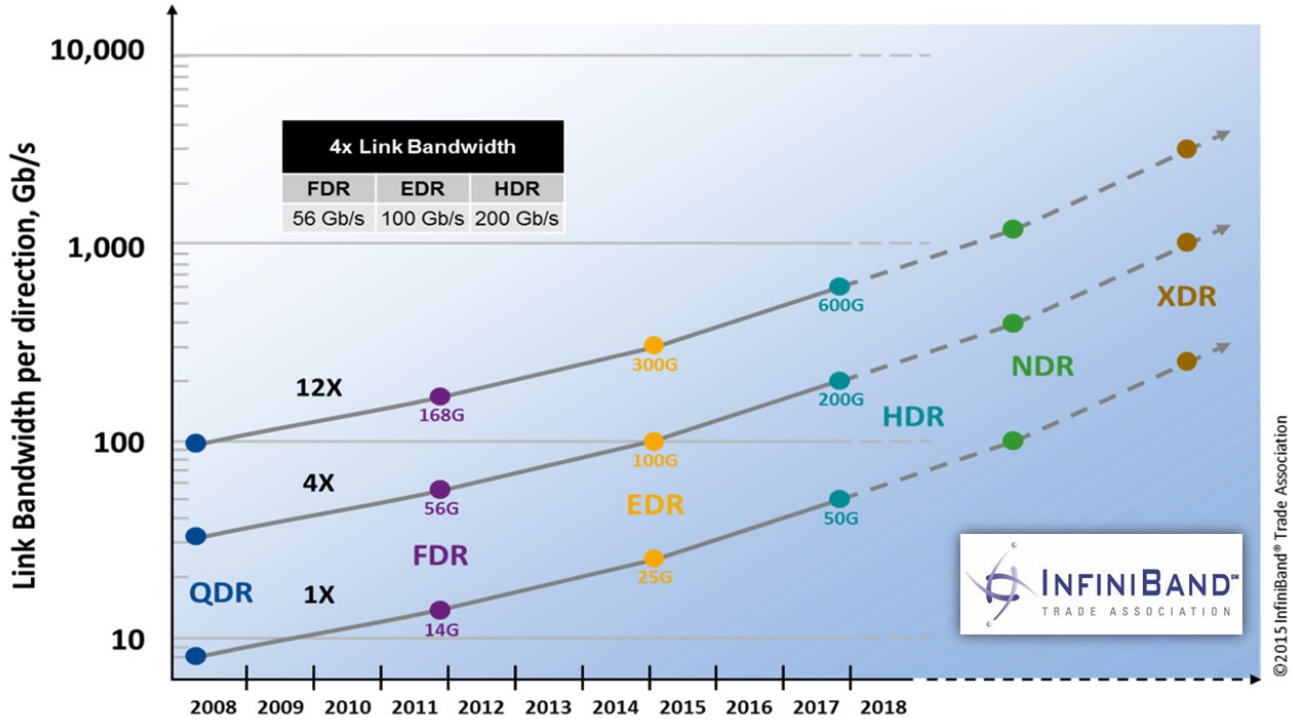
Ethernet has continued to evolve since then, driving specifications to ever better performance levels from the initial 3 Mbps to 100Gbps currently, with 400Gbps expected in 2017. Based upon its ubiquity and continuing development, Ethernet is clearly the dominant network for mainstream computing needs where a physical connection is required. When it fits, it is often the best option, but for high bandwidth and low latency deployments, better options have emerged.

InfiniBand

InfiniBand originated in 1999 to specifically address workload requirements that were not adequately addressed by Ethernet, and interoperability requirements that the then-current proprietary technologies were unable to meet. The initial specification released in 2000 by the InfiniBand Trade Association (IBTA) led to today's InfiniBand standard that leads in high bandwidth and low-latency and co-exists with Ethernet.

InfiniBand is designed for scalability, using a switched fabric network topology together with remote direct memory access (RDMA) to reduce CPU overhead. Efforts to implement RDMA over Converged Ethernet (RoCE) and other initiatives promise to continue the quest for the highest performing interconnect technologies and standards.

The InfiniBand protocol stack is less burdensome than that required for Ethernet. This enables InfiniBand to maintain a performance and latency edge in many high performance workloads. The IBTA roadmap shows bandwidth for HDR InfiniBand reaching 600 Gbps by 2017.



Proprietary Interconnects

Proprietary technologies frequently have a time to market (and therefore performance) advantage over standardized offerings if for no other reason than not having to deal with the overhead of the standardization process. For example, the fastest TOP500 systems usually include a healthy proportion of systems built with proprietary interconnects. Currently, proprietary interconnects are concentrated in the TOP50 and dominate the TOP10.

In recent years, the most common of the proprietary interconnects have been the IBM Blue Gene and Cray Aries interconnects deployed in combination with InfiniBand, Ethernet or Fibre Channel (FC) for connection to storage systems.

There is some significant change occurring in the proprietary interconnect landscape, with older companies being acquired and other technologies being superseded by InfiniBand. The most significant of these have been by Intel which acquired QLogic's InfiniBand assets as well as Cray's Gemini and Aries interconnect technologies. These acquisitions accelerated, and formed the foundation of, Intel's strategy to enter the high performance interconnect market.

Introduced in 2015, Intel's end-to-end Omni-Path Architecture (OPA) targets the InfiniBand market, claiming higher messaging rates and lower latency in addition to advanced features such as traffic flow optimization, packet integrity protection and dynamic lane scaling.

Onload vs. Offload and CPU-centric vs. In-Situ Processing

There is more to this scenario than mere vendor positioning. While OPA is derived from InfiniBand, the technology approaches are fundamentally different. The Intel philosophy quite naturally places the message processing predominantly on the server CPUs (the onload approach), with less dependency on the network hardware. The Mellanox approach (the offload approach) places less work on servers, using dedicated engines in the network switch and host card adapters (HCAs).

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 discussed below. 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.

The efficacy of each approach remains a matter of significant debate and contention. The precise distribution of functionality is not clear cut in an era where multi- or many-core CPUs are becoming the norm, energy efficiency is of paramount importance, and network function virtualization is in vogue. On the other hand, extreme scale is increasingly common, bandwidth

and latency more important than ever, and the parallelizability of networking tasks is a case-by-case affair.

Whichever philosophy is chosen, any impediment to message processing limits performance with significant consequences to the entire system's workload throughput and value to its host organization. Perhaps the biggest issue is that not all workloads are created equal and vary in processor and message passing requirements. In such a scenario the offload approach appears to be more workload agnostic.

Software Orchestration

There are two primary application programming interfaces for deploying workloads in parallel across multiple CPU resources and systems,

- Open Multi-Processing (OpenMP) supports multi-platform shared memory multiprocessing programming in C, C++, and Fortran on most platforms. It consists of a set of compiler directives, library routines, and environment variables that influence run-time behavior using a portable, scalable model that gives programmers a flexible interface for developing parallel applications for platforms ranging from the standard desktop computer to the supercomputer.
- Message Passing Interface (MPI) is a standardized and portable message-passing system designed to function on a wide variety of parallel computers and distributed clusters. The standard defines the syntax and semantics of a core of library routines useful to a wide range of users writing portable message-passing programs in C, C++, and Fortran. There are several well-tested and efficient implementations of MPI, many of which are open-source or in the public domain.

MPI is the primary API for distributed cluster systems in HPC environments. MPI workloads can also be run on shared memory systems as well as in conjunction with OpenMP on clusters of shared-memory systems. Both approaches are important and are optimized for different system and workload types. Since inter-node communication is frequently important, the interconnect and protocol processing take the dominant role for communication between systems rather than the API choice.

Open Fabric Alliance

The Open Fabrics Alliance (OFA) will be increasingly important in the coming years as a forum to bring together the leading high performance interconnect vendors and technologies to deliver a unified, cross-platform, transport-independent software stack. Founded in 2004 as the OpenIB Alliance, the Alliance was originally focused on developing a vendor-independent, Linux-based InfiniBand software stack.

Since then the organization has expanded its charter to include support for iWARP, RoCE and the OpenFabrics Interfaces working group to investigate and incorporate support for other high performance networks. Today, the vision of the OpenFabrics Alliance is to deliver a unified, cross-platform, transport-independent software stack for RDMA and kernel bypass to enable users to run their applications agnostically over InfiniBand, iWARP, RoCE, or other fabrics, including Intel's OPA though the OpenFabrics Software (OFS) open source offering.

The Once, Now, and Future landscape for High Performance Interconnects

OnceScape

Beginning in 1993 the TOP500 list has ranked the majority of the world's fastest computers together with many details of the system architectures. These systems run the most demanding and complex parallel workloads, and the larger the system the more critical high-bandwidth, low-latency interconnects become in overall system performance. A brief analysis of the TOP500 data highlights the roles that Ethernet, InfiniBand and proprietary interconnect technologies have played and how their distribution among the TOP500 ecosystem has changed.

NowScape

Today, Gigabit and 10G Ethernet-based systems collectively represent 44% of the TOP500 systems, but none are in the TOP50, and only 5 in the TOP100.

Since the first Petascale system was introduced in 2008, InfiniBand systems have grown from just under to 25% to over 40% of the TOP500 by June 2016.

Proprietary system interconnects account for just 15% of the TOP500, but completely dominate the highest performing (and cost) systems with 100% of the TOP10 on the June 2016 list. However, the race for superiority is far from over. When looking at the TOP11-30 systems it's 50/50.

FutureScape: The OrionX Perspective

The future is rooted firmly in the past. Concurrent with the evolution of computer interconnects, which mostly dates back to the sixties, is the question of the criteria that should be used to make a decision about strategic choices.

There are multiple aspects to consider:

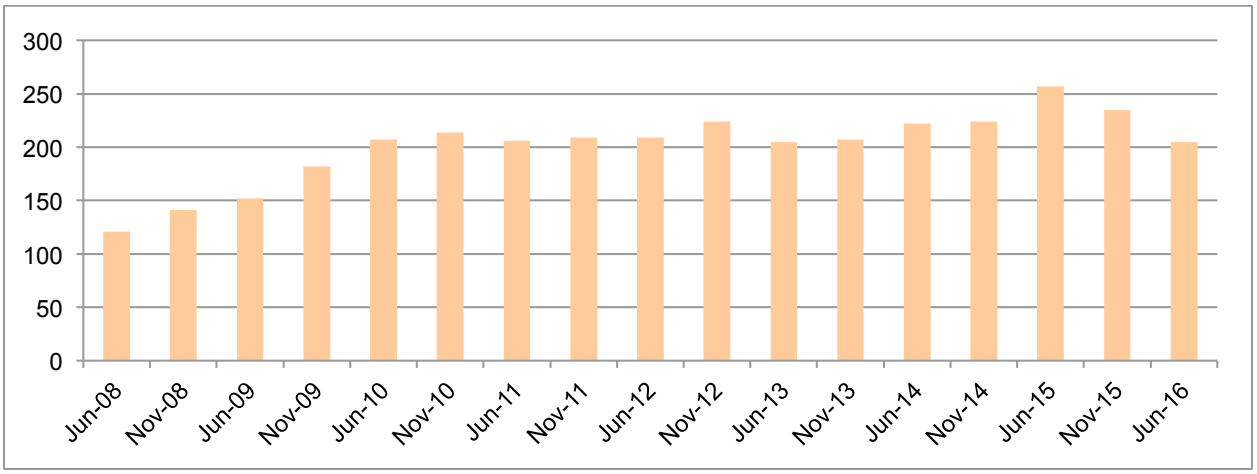
Technology considerations:

- In-Situ Processing: going beyond traditional offload vs onload methodologies to do more and more processing where the data happens to be or must pass through.
- Protocol processing: while some network functions are amenable to parallelism, many are not. This is the case also for MPI libraries.

Market presence and resilience:

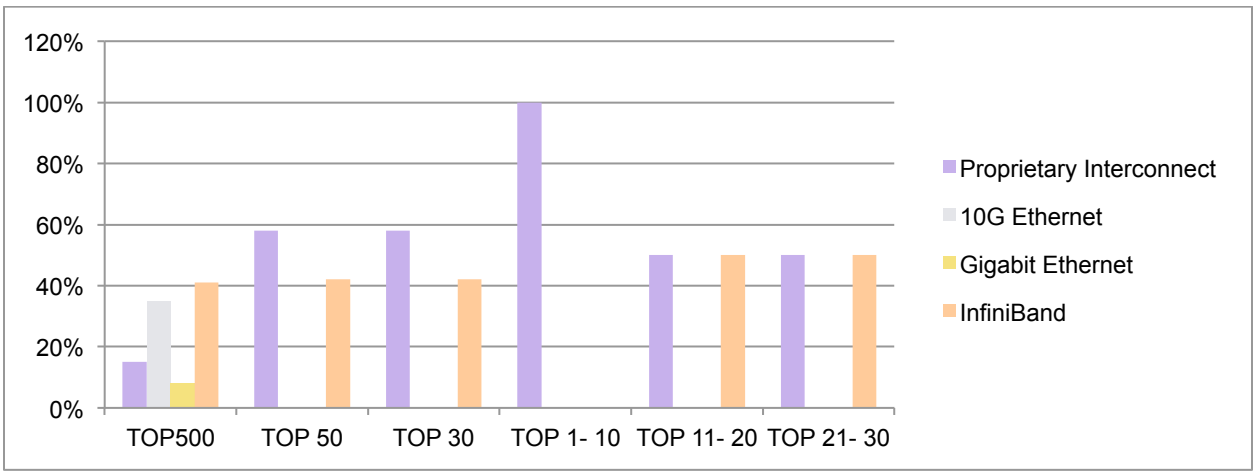
- All vendors of high performance interconnects enjoy strong market presence. Mellanox is a nimble innovative company with annual revenues approximating $700M. Intel has been able to dominate the microprocessor industry for decades while demonstrating considerable ability to adapt and persevere against onslaughts on multiple fronts. But it is new to the interconnect business. Major Ethernet vendors such as Cisco or Juniper are quite well-established. Proprietary vendors are well known in the HPC market and likely participate in all major bids.

TOP500 InfiniBand Systems



OrionX analysis of TOP500 data from June 2008 to June 2016

TOP500 HPI Distribution



OrionX analysis of TOP500 data from June 2016

Customer adoption preferences:

- At the high-end, InfiniBand and Mellanox have the dominant position and Intel is the new kid on the block. From both a technology and proven performance perspective InfiniBand appears to be the safe bet, but the HPC arena is where big bets are made and sometimes lost.

For the next TOP500 DOE Leadership computing systems expected in 2017 from the CORAL initiative, two have been awarded to systems based upon Mellanox InfiniBand technology, one has been awarded to the Intel-based Omni-Path and Cray-based technology. Only time will tell with Intel and Mellanox being quite evenly matched as the game commences.

For now, InfiniBand and its vendor community, notably Mellanox appear to have the upper hand from a performance and market presence perspective, but with Intel entering the HPI market, and new server architectures based on ARM and Power making a new claim on high performance servers, it is clear that a new industry phase is beginning. A healthy war chest combined with a well-executed strategy can certainly influence a successful outcome.

References:

1. Re-evaluating Network Onload vs. Offload for the Many-Core Era - 2015 IEEE International Conference on Cluster Computing.
2. The OpenFabrics Alliance (OFA) Overview

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