Breaking the Rules: 
Bright Cluster Manager with Cisco UCS, a Complete HPC Solution

WHITE PAPER
Breaking the Rules: Bright Cluster Manager with Cisco UCS, a Complete HPC Solution

Executive Summary

Delivering a cost-effective HPC solution to Cisco Unified Computing System (UCS) customers has become easier thanks to Bright Computing. Bright Cluster Manager is a powerful software solution for managing HPC clusters cost-effectively. Cisco UCS is a highly successful integrated server solution that offers a wide range of capabilities from a single vendor. Used together, these products provide an HPC cluster solution that offers the following cost-saving features:

- A single, high-performance Ethernet network fabric that can be used for all management and communications within the cluster the cluster.
- Top500 HPL benchmark performance that provides 89.69% efficiency across 512 Intel Xeon cores using Cisco Ethernet. Compared to Dassault Systems Abaqus benchmarks for QDR InfiniBand, Cisco UCS performed as well as or better than InfiniBand.
- The Cisco User-Space NIC (usNIC) provides 2.2μs latency when OpenMPI and a Cisco switch are used. Using various usNIC features, OpenMPI can create dedicated virtual hardware queues for highly efficient message-passing performance.
- Cisco UCS is a vertically integrated platform optimized by a single vendor that provides a unified fabric supporting three network types—interconnection, storage, and management.
- With the introduction of Bright Cluster Manager 7, any UCS C-Series server model can be easily and seamlessly managed as an HPC resource.
- Bright Cluster Manager provides deep and comprehensive control and manageability for deploying and managing UCS C-Series servers.
- Bright Cluster Manager provides full configuration and management of the Cisco usNIC Ethernet and Fibre Channel virtual interface.
- There is no need for specialized expertise to manage high-performance interconnects. Managers can utilize their in-house Ethernet expertise to build and operate an HPC cluster.
- A single, standard converged Ethernet fabric can easily be deployed to fit almost any storage scenario, including NAS, SAN, and FC.

When used together, Bright Computing and Cisco Systems provide a complete, integrated HPC Ethernet solution that decreases the amount of time IT managers, administrators, and users spend managing a complex collection of vendors, networks, and hardware. With improved focus on results, productivity increases, and costs decrease.
Not Your Parents’ Ethernet

HPC clusters are often built piecemeal using the best components from various vendors. One vendor may supply the servers and racks, another may supply the InfiniBand interconnect and switches, and yet another may provide Ethernet switches for management. Countless software tools and libraries may be required to operate and manage the system, not to mention the user applications. A parallel storage component may also be added to the cluster to facilitate scaling for workloads that push the IOPS (Input/Output operations Per Second) barrier. In almost all cases, a highly specialized management solution such as Bright Cluster Manager is added to help manage, run, and sustain the HPC resource. Without this critical piece of software, administrators must manage, integrate, and orchestrate all aspects of the cluster themselves, which is a time-consuming and expensive responsibility.

For years, Bright Cluster Manager has managed highly sophisticated multi-vendor systems. With the introduction of Cisco’s Unified Computing System (UCS), the market can now procure optimized x86-64 servers, high-speed network interconnects, 10/40Gbps Ethernet switches, and centralized storage from a single source. Cisco manufactures and supports the entire hardware platform, not piecemeal clusters. When this hardware infrastructure is integrated with Bright Cluster Manager, a new level of solution is possible. The “single pane-of-glass” management provided by Bright Cluster Manager now extends throughout the entire system, and Bright Cluster Manager offers an unprecedented level of management to the global HPC market without sacrificing any of the performance offered by InfiniBand.

The astounding performance of the UCS compute fabric has changed the perception of Ethernet in the HPC market. Bright Computing offers Cisco UCS users a new level of management and performance over a single high-performance Ethernet fabric.
Shattering the Ethernet Performance Myth

The HPC market demands performance. One of the myths floating around the HPC community is that Ethernet does not provide the level of performance that InfiniBand fabrics do. The following section provides evidence that shatters the Ethernet performance myth.

History

The first Beowulf HPC clusters were built using Ethernet. This decision was primarily driven by the low cost and high availability of commodity Ethernet technology. The low cost of Ethernet was due to its ubiquitous use in areas of IT other than HPC. In the early days of cluster-based HPC, exploiting the economics of the larger market share was the most cost-efficient method for building systems.

As the HPC market grew, so did the need for faster cluster interconnects. Specialized technologies that included Myrinet, QSnets, SCI, and InfiniBand were developed and deployed. Each interconnect technology was designed to provide lower latency and higher bandwidth than commodity Ethernet. To achieve this goal, these interconnects eliminated reliance on the operating system for transporting data, and user applications communicate directly with the interconnect adapter hardware. The most popular low-latency/high-bandwidth HPC interconnect technology currently used throughout HPC is InfiniBand.

While high-performance interconnects were being deployed, Ethernet continued to improve in latency and throughput. Outside of the HPC world, Ethernet enjoyed great adoption and growth and became the standard interconnect for data centers everywhere. In particular, as each successive generation was deployed in large volumes, prices dropped, which further facilitated market uptake. Currently, 10 and 40 Gigabit Ethernet (GbE) technologies are being adopted at ever-increasing rates across all market verticals throughout the IT industry.

Ethernet Is Already There

Most HPC clusters in production today have two network fabrics connecting the compute nodes. One is an administrative network, most often referred to as the provisioning and management network, which is often 1 Gbps Ethernet. The other serves as the compute interconnect, and is almost always InfiniBand (IB). The compute interconnect network can be either 40 Gbps (QDR – Quad Data Rate) or 56 Gbps (FDR – Five Data Rate), and is used for high-speed/low-latency message-passing between the processing nodes.

Having separate networks increases the amount of work required for administration. InfiniBand is a good HPC interconnect but has some disadvantages:

- IB is not Ethernet – Ethernet is everywhere, and as a well-understood technology, it serves as reliable transport for all forms of network communication throughout the data center. IB is virtually a plug-and-play proposition for many deployments. There are TCP/IP emulation layers for IB, but using IB as a general-purpose network is not practical.
- IB needs a separate hardware infrastructure – Since IB is a separate network, it requires its own host channel adapters (HCA), cables, and switches. This adds cost and requires additional rack space.
- IB requires a separate tool chain – Although IB drivers are part of most kernels, tools for managing IB networks are often unique. For instance, all IB networks need a Subnet Manager. This manager may be part of the IB switch or run as software on an administrative node.
- IB has a short half-life – Once a new generation of IB technology is released, the previous generation tends to fade from the market. Even with backward compatibility, this situation can make it difficult to manage and maintain older systems. In contrast, previous generation Ethernet hardware gets cheaper and more ubiquitous before it disappears from the market.
- InfiniBand has no standard Application Programming Interface (API). The standard lists only a set of library functions that must exist called verbs. The syntax of these functions is left to the vendors.

A single, cluster-wide low-latency high-performance Ethernet network eliminates the challenges that face many IB clusters. For many workloads, moderately-low latency and good bandwidth are all that is required. For situations where the absolute lowest latency and the absolute highest bandwidth are needed, IB has traditionally been the go-to solution.
Myth Shattered: Ethernet Wins

A discussion of low-level network functionality might be considered an academic exercise, but a seasoned HPC veteran knows that it matters. The network must ensure that the high-level user applications perform well. Each application uses the network differently, and rigorous performance benchmarking is the only sure way to measure success. For instance, many HPC users are surprised to learn that Cisco’s recently developed high-speed, low-latency interconnect technology known as user-space NIC (usNIC) performed extremely well when coupled with a Cisco Nexus 3548 10Gb Ethernet switch. The out-of-the-box performance, without any software or hardware optimization, produced an 89.69% High-Performance Linpack (HPL) efficiency across 512 Intel Xeon Cores.

In terms of application performance, Figure 1 plots the scaling results from Dassault Systems Abaqus running a Dodge Neon crash simulation using IB (QDR) and Cisco (usNIC). The lines indicate the run time for the simulation. In this test, Ethernet performed better than QDR IB. A similar behavior was observed for additional Abaqus jobs where usNIC and Ethernet either matched or surpassed the same number of cores compared to QDR IB.

These results show that Ethernet technology from Cisco can go head-to-head with IB in many application areas. Similar results were obtained for Ansys Fluent and StarCCM+ from CD-Adapco.

IB is no longer the only option for HPC clusters. With usNIC interconnect technology from Cisco, UCS servers and Nexus switches managed via Bright Cluster Manager offer all the performance of InfiniBand without the additional cost, complexity, or time investment traditionally associated with a multi-fabric cluster administrator and manager.

---

**Figure 1:** Benchmark results for Ethernet (usNIC) and IB (QDR). The lines indicate run times for the Dodge Neon crash simulation. The bars indicate the number of cores used.
Cisco usNIC Changes
the Ethernet MPI Equation

Cisco has introduced a low-latency Ethernet technology as part of their Unified Computing System (UCS). Cisco UCS is a data center server platform composed of computing hardware, virtualization support, switching fabric, and management API. A key technology is the Cisco usNIC featured on the VIC 1200/1300 series Converged Network Adapters (CNAs) that supports multiple payloads (e.g., Ethernet and Fibre Channel over Ethernet, FCoE) at 10 and 40 Gigabit per second speeds over a unified Ethernet network fabric.

By using usNIC as a firmware enhancement of Cisco 1200/1300 series VIC adapters, the host using open source software libraries can bypass the kernel and deliver extremely low-latency communications for HPC applications that leverage OpenMPI.

Figure 2 illustrates how Cisco usNIC differs from traditional “through kernel” TCP/IP communication methods. As shown on the left side of Figure 2, when data is sent via TCP/IP, the data must first traverse the user buffers, then the kernel buffers, and then the Ethernet driver layer before finally getting to the physical network adapter hardware. The same is true for data received by the physical hardware. The data must travel back through the kernel TCP/IP stack before the user-mode application can use the data being transmitted and received. Although TCP/IP communication through the kernel has many advantages, a large amount of overhead latency is added in this mode. In HPC, this overhead translates to latency.
Cisco uSNIC was designed to circumvent the need for kernel interaction and work directly with the user’s application. As shown on the right side of Figure 2, after the connection is set up, user applications can communicate directly with the physical hardware using the OS-bypass Linux Verbs API stack, which is also used in user-space IB applications. Once installed, the uSNIC software provides all the normal Verbs support commands.

The results are extremely low latency for user communications. The following are preliminary Half Round Trip (HRT) ping-pong latencies reported by Cisco for VIC 1200/1300 series adapters. In these tests, numerous small data packets are sent and returned between two systems; then the total time is measured and averaged over the number of number of packets sent resulting in the round-trip time.

HRT is half of this value or one-way latency.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Latency (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network cards back to back</td>
<td>1.7μs</td>
</tr>
<tr>
<td>Through a Cisco N3548 switch</td>
<td>1.9μs</td>
</tr>
<tr>
<td>Using Open MPI N3548 switch</td>
<td>2.2μs</td>
</tr>
</tbody>
</table>

The VIC series is a converged adapter, which means user-space Verbs communication, TCP/IP, and Fibre Channel (FCoE) communication can take place at the same time over the same hardware. There is no need for a second high-performance interconnect in HPC clusters as Ethernet can now address all cluster communication needs. Additionally, all user-space software, the OpenMPI stack, the VIC driver, and the Verbs libraries are open source.

A more complete communication profile for MPI can be found in Figure 3. In this figure, the standard Net-Pipe benchmark was run over a range of packet sizes. The results indicate an “HPC ready” profile for Ethernet.

**Figure 3: Net-pipe MPI profile using OpenMPI**
OpenMPI and the usNIC

Cisco usNIC has low-latency communication due to the VIC series of Ethernet NIC, which exploits a combination of Intel VT-d (Intel Virtualization Technology for Directed I/O) and PCI-Express SR-IOV (Single Root I/O Virtualization Specification) technologies. With PCI-Express SR-IOV, multiple virtual functions (VFs) or hardware queues can be created for use with user-space applications. Each VF may be considered an independent hardware queue resource. Multiple VFs can reside under a single network card interface or Physical Function (PF).

Because the VF shares a common address space with user space software using Intel VT-d, it is ideal for kernel bypass of network traffic. OpenMPI creates one VF per MPI process and uses the Unreliable Datagram (UD) Verbs interface to provide the low-latency message passing shown in Figure 3. As shown in Figure 4, each MPI process can create its own send-and-receive queue pairs that exist in their own (isolated) PCI-E virtual function (using VT-d). Combined with the Verbs UD library, each MPI process now has its own exclusive user-space communications pipeline that sits (virtually) directly on top of the physical hardware.

The Big Picture: Cisco UCS Clusters

Cisco Unified Computing System (UCS) combines industry-standard x86-architecture blade and rack servers, networking, and enterprise-class management into a single system. The system’s configuration is entirely programmable and provides a unified I/O infrastructure.

This system uses a high-bandwidth, low-latency unified fabric to support networking, storage I/O, and management traffic. Cisco UCS covers the entire technology stack from server to switch.

The physical compute chassis used by Cisco UCS is available in two form factors:

1. B-Series blade chassis with 8 blade servers per unified chassis
2. C-Series rack mount servers

In terms of HPC, Cisco UCS offers several advantages. In addition to the very low-latency Ethernet, UCS offers a converged network, extended memory technology, and scalability.

Cisco UCS is designed with a unified fabric that condenses three network types—interconnection, storage, and management—into one. This “wire-once” philosophy means that the system is wired only once, when it is installed, with bandwidth allocations and I/O configurations managed dynamically through the system’s embedded management features. Since all servers are physically wired with the same 10/40 Gigabit Ethernet and Fibre Channel over Ethernet (FCoE) network, all nodes within the system can host the same workloads simply by changing configurations through software.
Cisco Extended Memory Technology provides more than twice as much memory (384 GB) as traditional two-socket servers, increasing performance and capacity for demanding virtualization and large-data-set workloads.

As shown in Figure 4, highly scalable clusters can be designed with Nexus 3400 and 6000 switches to achieve low latency and high port counts (e.g., 9,216 10GbE ports with a switching latency of about 1.4 µsecs).

Support the Whole Stack
A key advantage of Cisco UCS is its integrated support. Because Cisco designed, integrated, and tested UCS systems from servers to switches to storage, UCS-specific issues are resolved by a single vendor, regardless of the problem. There is no need to manage multiple hardware vendors (server, interconnect, switching, storage, etc.) when an issue occurs.
Efficient Management Unlocks Performance

Managing a complex computing resource is expensive in terms of time and cost. Bright Cluster Manager provides a single, system-wide management solution that transforms hardware into a coherent and efficient HPC resource. Bright Cluster Manager is designed to fully manage HPC clusters and offers many features that are specific to HPC administration and user environments (i.e., these features are not available with Cisco UCS tools).

Bright Cluster Manager provides a unified and integrated approach to managing HPC clusters. It is designed and written from the ground up using a single-cluster management agent that provides all functionality. A single, central database configures, stores, and monitors all cluster data. Bright Cluster Manager provides a “single-pane-of-glass” management GUI for all cluster management tasks.

In terms of HPC systems, Bright Cluster Manager is easy to use, extremely scalable, and secure.

Bright Cluster Manager — Elements

![Diagram of Bright Cluster Manager elements]

Bright elements are shown in Figure 6. An abbreviated list of important HPC capabilities is as follows:

- Flexible Node Provisioning
- Cluster Management Shell
- GPU/Accelerator and ScaleMP Management
- Integrated Workload Managers
- Parallel Shell
- Deep System Monitoring
- HPC Tools/Library Management
Bright Cluster Manager and Cisco UCS Integration

With the introduction of Bright Cluster Manager 7, any Cisco UCS C-Series server model can be easily and seamlessly managed via API calls. It can configure, monitor, and manage C-Series rack servers directly in standalone mode or via the UCS Manager. Cisco Integrated Management Controllers (CIMCs) can be configured, controlled, and provisioned directly from Bright Cluster Manager without the 6200 Series embedded software UCS Manager.

After deployment, Bright Cluster Manager interacts with the rack servers through the CIMC XML API. This information is stored as the properties of a node or a category of nodes in Bright Cluster Manager. These properties can then be modified, monitored, and managed using standard Bright Cluster Manager capabilities.

The integration of Bright Cluster Manager with UCS minimizes the number of manual steps that must be performed by the HPC administrator. With Bright Cluster Manager, the administrator can:

- Assign and modify usNIC profiles for Cisco VIC interfaces
- Modify Fibre Channel Adapter and Interface properties
- Launch the remote KVM console to nodes
- Modify the network-related settings of a CIMC, such as NTP and syslog
- Monitor metrics related to power supply, fans, temperature, voltage, and storage
- Automatically configure the CIMC user and network

Bright Cluster Manager works with the UCS cluster to provide a high-level, HPC-focused management interface. Bright provides a single point of interaction for the HPC administrator, eliminating the need for the administrator to move back and forth between the UCS Manager and Bright Cluster Manager.

As an example, Figure 7 illustrates the Bright configuration window for managing the usNIC virtual interface. Bright Cluster Manager also provides a similar window for managing the virtual Fibre Channel interface of the UCS Ethernet fabric.
Summary: The Complete Ethernet HPC Solution

The combination of Bright Cluster Manager and Cisco UCS servers provides an entirely new level of integration, management, and performance previously unavailable in the HPC market. The highly desirable all-Ethernet HPC cluster has become a viable solution for technical computing with the following capabilities:

- True HPC performance over 10 and 40 Gigabit Ethernet that is on par with InfiniBand QDR/FDR interconnects
- Decreased cost and complexity through the ability to consolidate multiple network fabrics (IB, FC) onto a single high-performance Ethernet network
- Comprehensive and cost-effective HPC administration from Bright Cluster Manager
- Deep management through the integration of UCS API’s in Bright Cluster Manager 7 allowing customers to fully manage C-Series servers. The UCS Manager supports only deep management of UCS blades
- Bright Cluster Manager is the only HPC Cluster Management solution validated for use with Cisco UCS servers and Nexus switches
- Using Bright Cluster Manager, Cisco VIC adapters can be easily provisioned without moving cables in the data center providing cluster flexibility and a means to quickly increase compute capacity.

With the single fabric solution for HPC, IT managers can now manage all of their resources in the same way. IT managers do not have to develop specialized expertise to provide high-performance interconnects. Engineering can focus on designing and testing products instead of managing HPC clusters.

A single-standard converged Ethernet fabric can easily be deployed to fit any scenario, regardless of the backend technology. Storage networks that use NAS, SAN, and FC, and direct-attach methods are greatly simplified because they no longer require a separate physical layer or the need to propagate over an InfiniBand fabric.

A system built on Cisco UCS using Bright Cluster Manager is a complete HPC Ethernet solution that frees the IT manager, administrator, and users to focus on HPC results instead of managing a complex collection of hardware and software.
About the Author
Douglas Eadline, PhD, began his career as a practitioner and a chronicler of the Linux cluster HPC revolution and now documents Big Data analytics. Before starting and editing the popular ClusterMonkey.net website in 2005, he served as editor-in-chief of ClusterWorld Magazine and was senior HPC editor for Linux Magazine. Currently, he is a writer and consultant to the HPC industry. He is a coauthor of “Apache Hadoop YARN: Moving Beyond MapReduce and Batch Processing with Apache Hadoop 2” (Addison Wesley), author of “HPC for Dummies” (Wiley), and presenter/author of several popular Hadoop videos from Addison Wesley. He can be contacted at deadline@clustermonkey.net