Large Scale Clustering with Voltaire® InfiniBand HyperScale™ Technology

Contents

Scalable Interconnect Topology Tradeoffs

Since its inception, InfiniBand has been optimized for constructing clusters with very large compute power that are based on standard, commodity components. Over the past decade, these cluster-based supercomputers have been running the world's most complex calculations in industries such as government and academic research, financial services, energy, manufacturing, and bioscience. As microprocessors and server motherboards evolve, the ability to efficiently scale beyond the petaflop barrier is now a reality.

While extremely powerful in compute power, the overall solution cost of petascale clusters is high, and trading off between cost, complexity and bisectional bandwidth is a major challenge. To date, several topologies have been used to properly scale these clusters for achieving the required performance.

The most commonly used topology is Clos, also referred to as Fat Tree. Clos is the only topology that can scale to unlimited size, while maintaining full bisectional bandwidth and an equal latency between any two nodes in the network. The interconnect layout in Clos enables applications to scale while maintaining linear performance. The downside of such a topology is the high overall cost, as well as the amount of cabling required.

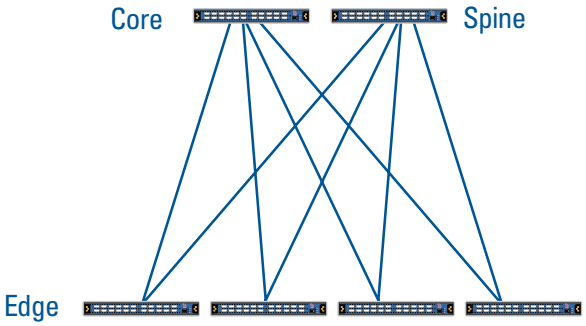

Figure 1: Standard Clos topology

Another topology practiced more rarely is Hypercube. In this topology, each node is directly connected to only a small set of neighbors. This allows users to build large clusters with fewer switching components at a lower cost. However, this topology has several significant downsides, including limited bisectional bandwidth and added latency caused by the many switch hops required to communicate between compute nodes in different parts of the fabric. In addition, there is often a higher degree of congestion spread and an increase in latency when many compute nodes correspond to fewer storage nodes and flood the fabric.

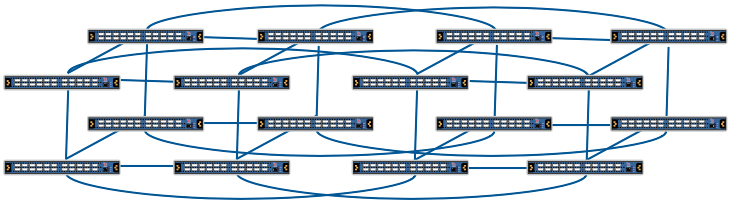

Figure 2: Hypercube topology

About HyperScale™

Drawing on its extensive experience in scale-out computing, Voltaire has introduced a new cluster topology called HyperScale as part of its 40Gb/s InfiniBand (QDR) switching solutions. HyperScale reduces the overall cost and complexity of medium to large clusters (thousands of cores), without impacting overall application performance. HyperScale is designed to make the most of the main benefits of each of the topologies mentioned above.

The basic building block of a HyperScale configuration is the Grid Director 4700 HyperScale fabric board. This unique fabric board carries double the switching capacity of a standard QDR fabric board, manifested in additional external 120 Gb/s 12x InfiniBand ports providing connectivity to other switches.

In traditional switches, the fabric boards define the spine of the Clos and all further expansion requires adding switches and cables in multiple tiers. With HyperScale, expansion starts directly from the spine and requires only a single tier of switching, thus significantly reducing the overall number of switch ports and cables required.

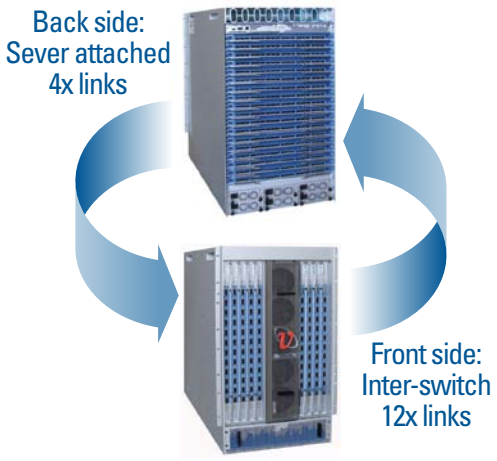

Figure 3: HyperScale Grid Director 4700 switch architecture fanning out an equivalent of 648 4x QDR ports from both sides

HyperScale comes in two different configurations: one designed to solve the challenges of medium-sized (300-600 nodes) clusters and the other to address those of large clusters (more than 650 nodes).

HyperScale Edge

The HyperScale Edge configuration expands the spine of the Clos through HyperScale fabric boards by adding edge switches connected as if they were additional line cards of the director switch. In other words, the fabric boards serves as a spine layer to connect the external edge swiches, which act as line boards.

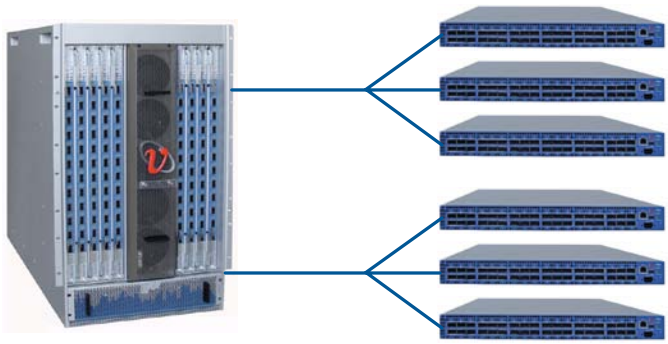

Figure 4: HyperScale-Edge

The benefit of such an approach is that users can create the right size switch and manage the cost of the solution while maintaining the option to grow the switching fabric in an incremental manner in the future. For instance, consider a case where a new cluster will require between 324 and 648 ports. Rather than utilizing one monolithic, partially populated 648 port chassis we can utilize a Voltaire Grid Director 4700 switch with expansion edge switches to obtain the capacity required. This approach has the following benefits:

- Decrease cable bulk and optimize cable routing
- Distribute weight by introducing flexible options for the placement of switch components in racks
- No need for forklift upgrades when node count exceeds 324
- Gradual scalability

Voltaire Grid Director™ Switches

Voltaire's Grid Director switches, in combination with Voltaire's advanced management software, vastly improve the performance of mission-critical applications.

Voltaire's fourth generation series of smart switches address the growing size and complexity of clusters with 40 Gb/s InfiniBand connectivity, advanced carrier-class management and a unique HyperScale™ stackable architecture.

Voltaire Grid Director 4700

With configurations of up to 324 ports or double-sided 648 ports of 40 Gb/s per port InfiniBand connectivity, the Voltaire Grid Director 4700 delivers an impressive 51.8 Tb/s of nonblocking bandwidth and the lowest latency in the industry. Its smart design provides unprecedented levels of performance and makes it easy to build clusters that can scale out to thousands of nodes. The Voltaire Grid Director 4700 is a high performance, ultra low latency and fully non-blocking InfiniBand switch for high performance clusters. The switch's HyperScale™ architecture provides a unique inter-switch link capability for stacking multiples of 324 ports to form highly scalable, cost effective, and low latency fabrics.

Voltaire Grid Director 4036

The Voltaire Grid Director 4036 is a high performance, low latency and fully non-blocking switch for high performance computing clusters. Perfect for use as an edge or leaf switch, it offers full bisectional bandwidth of 2.88 Tb/s. With built-in high availability and a design that is easy to maintain, the Grid Director 4036 is a cost-effective alternative to proprietary interconnect technologies. With 36 40 Gb/s ports in a 1U chassis, the Grid Director 4036 delivers high bandwidth and low latency at an affordable price.

The examples below compare a HyperScale Edge configuration for a 432-node cluster with a traditional configuration. The first example describes a cluster based on blade servers, and the second example is based on rack mount servers. The tables show the advantages of using HyperScale Edge vs. the traditional approach (a single 648 monolith switch) in each of these cases.

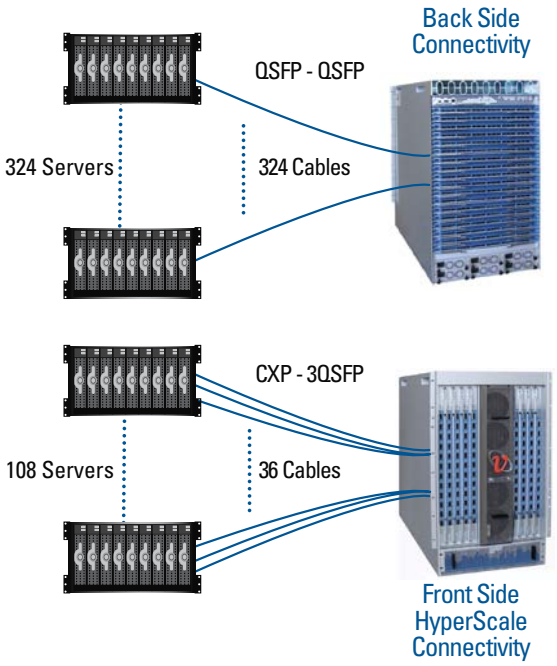

Figure 5: HyperScale Edge 432 cluster with blade servers

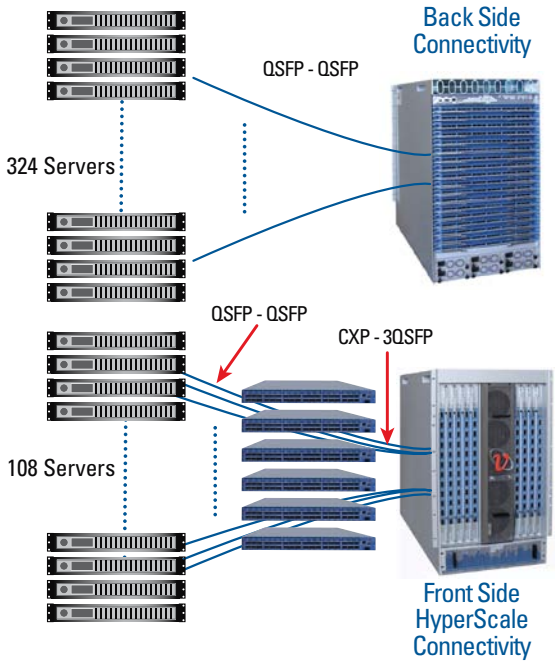

Figure 6: HyperScale Edge 432 cluster with rack mount servers

	Voltaire: HyperScale Edge	Traditional 648 monolith approach
Rack Space	19U	~30U
Inventory	Same components (chassis, line/fabric cards, PS, fans) for all node counts up to 648 nodes	648 port switch requires different chassis, line/fabric cards and fans from 324 port switch
Scalability	Increments of 18 ports all the way to 648 nodes	Forklift upgrade when passing 324 nodes
Latency	Maximum 4 hops	Maximum 5 hops

	Voltaire: HyperScale Edge	Traditional 648 monolith approach
Rack Space	19U entry point	~30U entry point
Flexible Racking	Optional split between 2 racks for avoiding power/cabling/weight challenges in certain racks	Single rack mandatory – could cause power/cabling/ weight challenges in certain racks
Inventory	Same components (chassis, line/fabric cards, PS, fans) for all node counts up to 648 nodes	648 port switch requires different chassis, line/fabric cards and fans from 324 port switch
Scalability	Increments of 18 ports all the way to 648 nodes	Forklift upgrade when passing 324 nodes

HyperScale Mesh

In addition to adding QDR ports to a Grid Director 4700, the HyperScale connectors can also be used to interconnect multiple Grid Director 4700 switches. One powerful way to do this is called HyperScale Mesh.

In a HyperScale Mesh configuration, a number of Grid Director 4700 switches are connected in an 'all to all' or 'fully connected' topology using 12X cables. Every Grid Director 4700 is connected to every other Grid Director 4700 via a bundle of 12X cables. Each 12X cable is equivalent to three 4X QDR cables and can carry 24 Gb/s of bi-directional bandwidth.

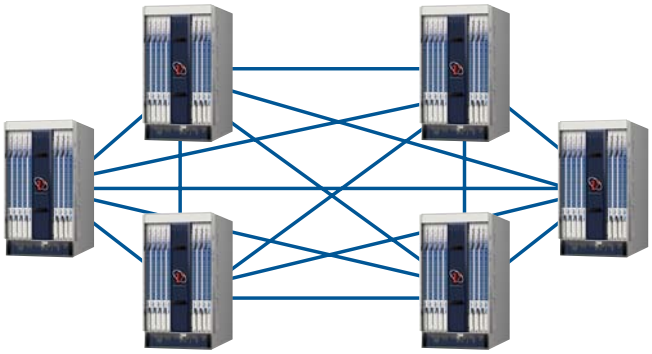

Figure 7: HyperScale Mesh

HyperScale Mesh can scale to several thousands of server nodes simply by adding more Grid Director 4700 switches. It differs from conventional Clos or Fat Tree topologies in several significant ways:

- Each Grid Director 4700 is a non-blocking QDR sub-cluster of 324 or more nodes.
- Worst-case latency between servers is only 4 hops, regardless of Mesh size. In a Clos or Fat Tree of similar size, the minimum worst case latency is 5 hops and can increase to 9 or more hops, within the range of comparison of 1000 to 6000 nodes, as cluster size increases.
- Each Grid Director 4700 sub-cluster communicates with the rest of the Mesh with as much as 2.6 Terabytes of bi-directional bandwidth, depending on the configuration. Unlike a Clos or Fat Tree topology, there is no need for leaf switches as an intermediate switching layer between the servers and the core switches. A Mesh is a single layer of Grid Director switches without the expense or additional cabling for another layer of switches.
- There are only 108 12X cables to each Grid Director 4700 chassis regardless of Mesh size.
- A non-blocking QDR sub-cluster of 324 servers using a modern CPU technology such as Intel Nehalem is a powerful Top500-class cluster in its own right.

A singular case of the HyperScale Mesh involves only two Grid Director 4700 switches. This configuration implements a fully non-blocking 648 port cluster. Unlike a monolithic switch it offers the flexibility of installing 648 ports of switching in one rack or splitting the switching across two racks to simplify cabling or reduce weight or power consumption per rack.

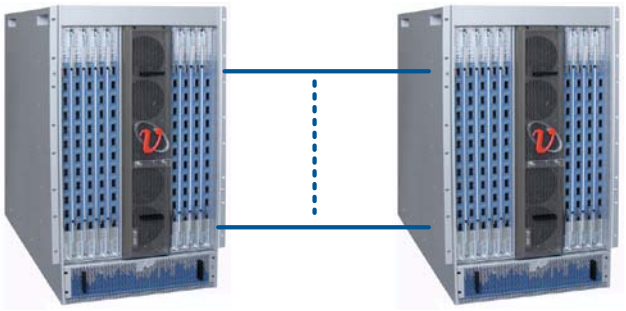

Figure 8: HyperScale - 648 ports

A HyperScale Mesh configuration can be viewed as a 'cluster of clusters'. Connecting non-blocking sub-clusters with Terabytes/second of intra-cluster bandwidth is a very cost-effective way to deploy a multi-application cluster that can run any mix of jobs from small to very large.

The example below compares a HyperScale Mesh configuration for a 1,296 node cluster, with the traditional, 2-tier Clos-based approach, which uses 36-port edge switches in the first tier, and 324/648 port switches in the second tier:

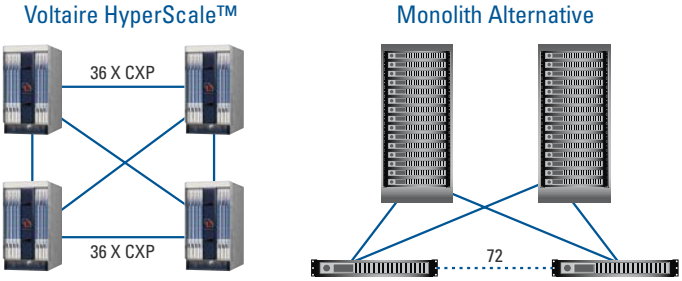

Figure 9: HyperScale Mesh 1,296-node example

	Voltaire: HyperScale Mesh	Competitor: Traditional Approach
Total number of switch ports	1296	2592
Total number of inter-switch cables	216	1296
Latency	Maximum 4 hops	Maximum 5 hops
Rack Space	2 full racks	4 full racks

HyperScale and UFM™: A Powerful Combination

In HyperScale Mesh, as in any InfiniBand topology, using a conventional subnet manager will often result in sub-optimal performance, since subnet managers have no knowledge of the actual traffic patterns among the compute nodes. Instead, they assume an 'any to any' communication pattern where each server sends an identical amount of traffic to every other server in the entire cluster. Even in a cluster of only a few servers this assumption is usually false, and can result in dramatic under-utilization of some links—even idle links—while other links are grossly overloaded. Due to this assumption, even non-blocking topologies typically deliver only a fraction of their theoretical bandwidth. The limitation of InfiniBand static routing is detailed in reference [1].

Voltaire Unified Fabric Manager™ (UFM) is aware of the fabric topology and the application traffic requirements and will optimize link utilization in conjunction with topology. For HyperScale Mesh configurations, UFM ensures that the 12X links are fully utilized, and will re-optimize as old jobs terminate and new jobs start. UFM also optimizes and balances inter-switch traffic with each Grid Director 4700 sub-cluster, which is equally important for overall fabric performance. UFM's set of traffic optimization algorithms are built on top of OpenSM routing algorithms and are referred to as Traffic Aware Routing Algorithms (TARA).

This approach takes the rearrangeable non-blocking configuration and rearranges it to the extent that there is an equal distribution of routes on each external and internal port in the fabric. This applies to both uniform and non-uniform distribution of traffic patterns. Further improvement is reached when the system administrator can provide information on the application pattern and traffic rates that are taken to further optimize the route distribution and performance results.

The following diagram shows a real life scenario where traffic was run with and without UFM's traffic optimization capability. UFM balanced the overall traffic over all the ports of the fabric by rearranging routing and, as evident from the results on the right, diffused the congestion spikes that existed originally. In all cases, Voltaire's Unified Fabric Manager will do a superior job of assigning HyperScale routes. Its knowledge of application connectivity enables it to optimize link utilization and deliver significantly higher performance to every application on the fabric.

[1] T. Hoefler. Multistage Switches are not Crossbars: Effects of Static Routing in High-Performance Networks. Oct. 2008.

Voltaire Unified Fabric Manager

Voltaire's Unified Fabric Manager™ (UFM™) software is a powerful platform for optimizing the performance of large server and storage scale-out fabrics. UFM enables data center operators to efficiently monitor and operate the entire fabric, boost application performance and maximize fabric resource utilization.

Unlike other management software platforms that are device-oriented and involve tedious manual processes, UFM software provides IT managers with a logical view of their infrastructure. This bridges the traditional gap between servers, applications and fabric elements, creating a more effective and business-oriented way to manage and scale out high-performance fabrics.

Advantages of the UFM approach include:

- Improved visibility into fabric performance and potential bottlenecks
- Improved performance due to application-centric optimizations
- Quicker troubleshooting time due to advanced event management
- Efficient management of dynamic and multi-tenant environments
- Higher utilization of fabric resources

UFM includes an advanced granular monitoring engine that provides real time access to switch and host data. The software also provides a unique congestion tracking feature that quickly identifies traffic bottlenecks and congestion events spreading over the fabric. This feature enables more accurate problem identification and quicker resolution.

UFM also optimizes routing algorithms by taking into consideration the fabric topology, the various services and active applications and their characteristics. UFM optimization features are built on top of the OpenSM industry standard routing engine, which provides fast fabric bring up with leading edge routing algorithms and maximizes the use of available fabric bandwidth. In addition, UFM enables segmentation of the fabric into isolated partitions, increasing traffic security and application performance.

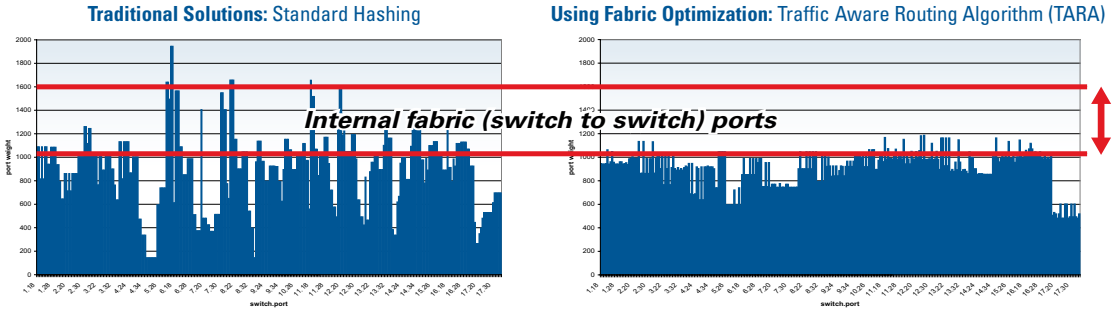

Figure 10: UFM Traffic optimization results

Racking & Cabling

Another advantage of the HyperScale configuration compared to a monolith 648-port switch is the flexibility in terms of racking. Mounting a 648-port switch in a single rack introduces several challenges that might exceed the per-rack limits in certain data centers:

- Number of cables per rack
- Maximum weight per rack
- Maximum power consumption per rack

The modular approach of HyperScale allows spreading the load among multiple racks, in a way that can fit the specific requirements of any data center. While a huge 648 port switch in a single rack requires extra space on both sides to manage horizontal cable guide apparatuses, any two Grid Director 4700 switches can be placed in adjacent racks and function as a distributed switch with true standard racking.

HyperScale Mesh Single Rack | HyperScale Edge Single Rack | HyperScale Mesh Dual Rack | HyperScale Edge Dual Rack

Figure 11: Flexible Racking

Summary

Voltaire HyperScale technology and products provide a more modular and less costly way to scale-out clusters to thousands of nodes while maintaining and exceeding the performance characteristics of the classic Clos topology. HyperScale is based on a more modular and flexible building block approach than alternative switches, and provide cluster architects and managers with more deployment options, less cabling and better usability.

Many customers have already embraced this innovation and benefit daily from HyperScale-based clusters in HPC production environments.

HyperScale Mesh Benefits

- Scale to 10,000 nodes with linear performance
- Scale best price-per-port-count configurations (pay as you grow)
- 4-hops maximum latency to any port
- Expandable in increments of 324 ports with port counts while maintaining pure 324 non-blocking islands
- Flexible component placement in racks to distribute weight and power consumption
- Optimize cable management, decrease the overall number of cables and ensure best practices for cabling the entire cluster

HyperScale Edge Benefits

- Up to 648 QDR ports, non-blocking Fat Tree topology, full bisection bandwidth
- Expandable in increments of eighteen ports with port counts ranging from 18 to 648 total ports
- 4-hops maximum latency to any port
- Uses same chassis & parts for configurations both above and below 324 nodes for simpler inventory management

About Voltaire

Voltaire (NASDAQ: VOLT) designs and develops server and storage switching and software solutions that enable high-performance grid computing within the data center. Voltaire refers to its server and storage switching and software solutions as the Voltaire Grid Backbone™. Voltaire's products leverage InfiniBand technology and include director-class switches, multi-service switches, fixed-port configuration switches, Ethernet and Fibre Channel routers and standards-based driver and management software. Voltaire's solutions have been sold to a wide range of end customers including governmental, research and educational organizations, as well as market-leading enterprises in the manufacturing, oil and gas, entertainment, life sciences and financial services industries. More information about Voltaire is available at www.voltaire.com or by calling 1-800-865-8247.

Contact Voltaire to Learn More

1.800.865.8247
info@voltaire.com
www.voltaire.com