WHITE PAPER

WovenSystems

Woven Systems vSCALE Technology:

Using 10 Gigabit Ethernet to Create a Superior Data Center Fabric

June 2008

2455 Augustine Drive, Suite 211 • Santa Clara, CA 95054 • office +1.408.654.8100 • fax +1.408.654.8933 • www.wovensystems.com

Introduction

Ethernet is the most pervasive, most widely installed networking technology in use today. Because Ethernet is common, people rightly associate it with low cost solutions, and familiarity with the technology makes implementing it operationally simple. Ethernet's strengths are such that it is becoming a mainstream technology in realms where it traditionally has had limited use. As bandwidth requirements increase and networks are pushed to the limit, tremendous investment is being made in developing higher capability such as in 10 Gigabit Ethernet (GE) technologies, which promise high bandwidth at the low cost and easy implementation of Ethernet.

Woven Systems™ believes that the limits of traditional Ethernet switches in the data center can be addressed, and that Ethernet can serve as the basis for a superior data center solution by providing dynamic traffic and congestion management. Woven has developed its breakthrough vSCALE™ technology to do just that—deliver extraordinary throughput by employing Woven's innovative traffic management feature called Dynamic Congestion Avoidance.

This white paper is organized into three sections followed by a brief conclusion. The first section details the key requirements for Ethernet switching in the data center. The second section describes Woven's vSCALE technology and how it satisfies these requirements, making Ethernet fully suitable for use in large-scale data centers while maintaining full Ethernet standards compliance. The third section proves the point with conclusive test results comparing throughput performance of Ethernet with and without vSCALE.

Requirements for Ethernet Switching in the Data Center

The data center is a specialized environment with specific requirements. Traditionally, Ethernet switches have fallen far short of meeting those requirements.

The purpose of a networking fabric in the data center is to connect a set of computing and storage elements. Traffic patterns between the elements in a fabric can change – showing peaks and valleys according to the applications being run and as a result of virtualization and consolidation. To minimize peaks and maximize efficiency, the following seven requirements are essential for a functioning network fabric in a data center:

1. Ability to scale. Data centers are getting larger in terms of compute capacity, and each network connection is running closer to its capacity as CPUs increase in performance and as virtualization is adopted. The fabric, therefore, must have the capacity to connect increasingly large numbers of high-performance elements.

2. Operation at rated capacity without congestion. This requirement entails much more than providing non-blocking throughput. Ideally, given the capacity of a specific destination or target, the amount of available source or initiator bandwidth, respectively, should approach that capacity. This means that variable traffic flows should be dynamically and efficiently balanced across ports dedicated to a given destination or target, with the fabric itself never becoming a bottleneck.

3. Rapid failure recovery. The fabric should be able employ a resilient network topology, preferably a self-healing infrastructure, and must be able to recover rapidly from any path (link or switch) failure.

4. Ultra-low latency. High latency degrades throughput and application performance as computing elements are forced to wait on the networking fabric. Ultra-low latency is especially critical for inter-process communications in distributed applications, high-performance compute clusters, and for efficiently moving virtual machines from one physical server to another.

5. Flexible bandwidth-provisioning among endpoints. Data center applications today demand the ability to provide variable bandwidth between high capacity endpoints, particularly with server virtualization and storage consolidation. Traffic patterns change over time, and the mapping between endpoints must be adjustable—ideally in real time to accommodate dynamic application demands.

6. Comprehensive traffic management. The fabric itself must be an essential part of application performance management by clearly identifying egress ports that have oversubscribed destinations or targets, and then controlling those traffic flows at the source/initiator ingress ports to prevent packet drops and inefficient retransmissions.

7. Minimal power and space consumption. The fabric should provide the maximum amount of bandwidth for the least amount of power and space to enable data centers to grow "green" (both environmentally and financially).

Traditional Ethernet switches fail to satisfy these fabric requirements, rendering them undesirable for the data center. One of the most significant shortcomings, for example, is that latency is unacceptably high due to the use of store-and-forward switch architectures. Superior data center fabrics provide latencies of less than 10 microseconds, while traditional Ethernet switches experience latencies that are a factor of 100 or more times longer.

The scalability of traditional Ethernet switches is also relatively low, and data blocking occurs at the port, device, and network level. These shortcomings make data center network design excessively complex and expensive in the face of aggressive traffic growth and changing traffic patterns. In addition, power and space consumption become unsuitably high due to the need to construct hierarchies of switches to design around the capacity limitations of individual switches. Complex network designs also make isolating the impact of the network on application performance nearly impossible. These same limitations exacerbate the impact of network element failures causing significantly greater recovery times, while undermining the ability to assess available network bandwidth.

Besides requiring complex hierarchical network design, traditional Ethernet switches also depend on costly high capacity switches at the core of the hierarchy. While necessary to meet needs today, these switches become prohibitively expensive in the face of projected traffic growth, especially considering the short useful life due to their constrained backplane capacity.

Moreover, the available capacity of traditional Ethernet switches is often underutilized. Even if capacity exists across multiple paths between ingress and egress ports, that capacity isn't always exploited. While flows may be statically assigned across multiple paths initially, as those flows peak and/or as new flows are created, available capacity inevitably becomes underutilized—often substantially.

Woven Systems developed its innovative vSCALE technology to overcome these and other shortcomings of traditional Ethernet switches in the data center, and to fully satisfy all seven requirements for data center fabrics. The result is a new type of Ethernet switch—the Ethernet Fabric Switch—that enables full utilization of fabric bandwidth and maximizes visibility into the fabric's impact on application performance.

vSCALE Packet Processing Technology Enables Large-scale Ethernet Fabrics

Woven Systems determined it was necessary to address Ethernet's inherent limitations in silicon in order to scale Layer 2 Ethernet networks beyond the bandwidth capacity of a single switch and to provide Dynamic Congestion Avoidance capable of delivering non-blocking, lossless throughput at 10 Gbps across a resilient Ethernet switching fabric. The resulting vSCALE technology utilizes existing standards in novel ways to enhance the performance (both throughput and latency), resiliency, and scalability of Ethernet in the data center.

Significantly, vSCALE complies fully with all Ethernet standards to assure interoperability with any 10 GE NIC in any server, storage device, switch, or router. Specifically, vSCALE provides the following capabilities that satisfy all seven requirements for a networking fabric in the data center:

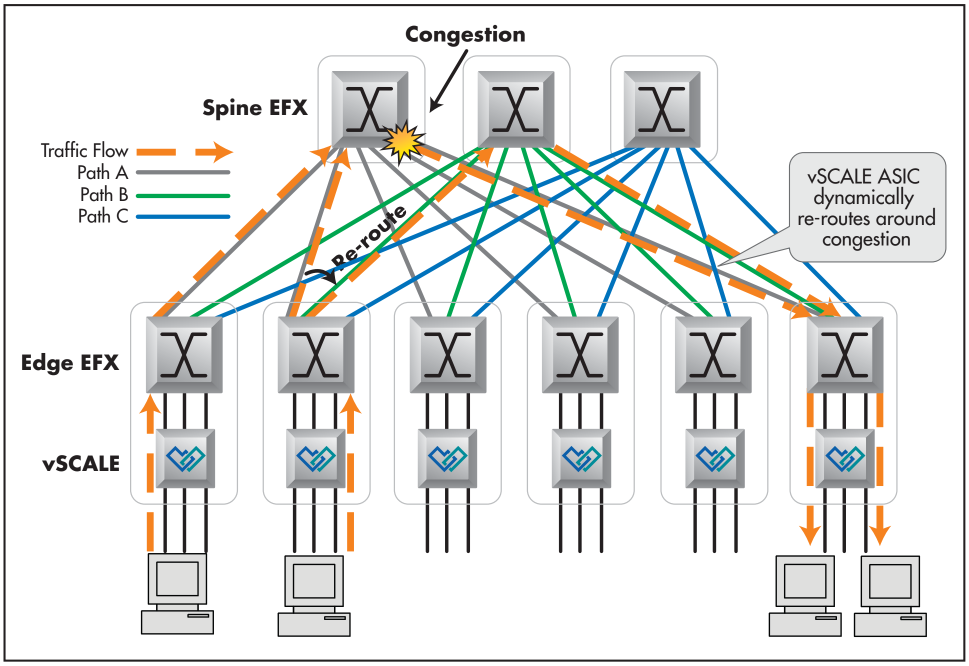
- Affords unprecedented scalability—well beyond the ports/bandwidth of a single switch chassis—supporting edge ports in a multi-path, multi-switch Ethernet fabric
- Distributes traffic flows continuously and automatically among all available paths in real time and rebalances traffic distribution to eliminate congestion within the Ethernet fabric and deliver non-blocking 10 GE throughput at all edge ports
- Delivers mission-critical reliability by utilizing a resilient, multi-path fabric topology that is able to detect and recover from path failures in under 10 milliseconds
- Employs cut-through switching to minimize latency to less than 4.6 microseconds in multi-switch fabrics
- Permits provisioning among any source/destination and initiator/target pairs by enabling multiple diverse active paths from any ingress port to any egress port
- Facilitates source rate control at the ingress port using the Ethernet Pause function when the destination or target at the egress port becomes congested
- Supports "greening" initiatives in the data center by consuming less than 17 Watts per 10 GE port

Woven's vSCALE technology adds a layer of intelligence within the Ethernet switching fabric, while maintaining full compatibility with existing and emerging Ethernet standards. The vSCALE application-specific integrated circuit (ASIC) uses a combination of latency and latency jitter measurements to serve as a proxy for detecting congestion along any path. These one-way forward measurements are made at regular intervals using special time-stamped packets for the active and alternate paths between pairs of edge ports. The mathematical algorithms Woven employs are able to calculate Layer 4 traffic flow latencies with sub-microsecond accuracy. These measurements also enable operators to monitor throughput and latency constantly, providing visibility into the impact of the Ethernet fabric on application performance, and collecting statistics for reports of current and historical performance to assist in capacity planning.

The constant awareness of congestion along all paths across the fabric is what enables Woven's switch, the EFX 1000 Ethernet Fabric Switch, to provide Woven's signature Dynamic Congestion Avoidance in a resilient, multi-path Ethernet fabric topology. To assure peak application performance, vSCALE selects uncongested, alternate paths to achieve minimum aggregate latency. This automatic process of rebalancing traffic flows is both fast, with a response time of less than a millisecond, and stable, allowing for lossless throughput with no reordering of packets. By combining this real-time rebalancing with cut-through switching, the EFX 1000 is able to deliver 10 Gbps throughput for all edge ports with the lowest possible edge-to-edge latency.

When multiple EFX 1000 Ethernet Fabric Switch units are interconnected in a large, multi-stage 10 GE fabric, vSCALE enables lossless, non-blocking throughput with an industry-leading port-to-port latency of less than 4.6 microseconds. When configured in smaller deployments on a single EFX 1000 chassis with 144 ports, edge-to-edge latency is reduced to 1.6 microseconds.

Because vSCALE functions edge-to-edge and not link-by-link across the Ethernet fabric, it has no role to play in intermediate Spine switches or at the internal-facing ports on Leaf switches. This "bypass" mode is, therefore, configured automatically for all non-edge ports. The vSCALE ASIC can also be bypassed intentionally, effectively turning the EFX 1000 into an ordinary Ethernet switch. Such a configuration is utilized in the tests reported in the next section.

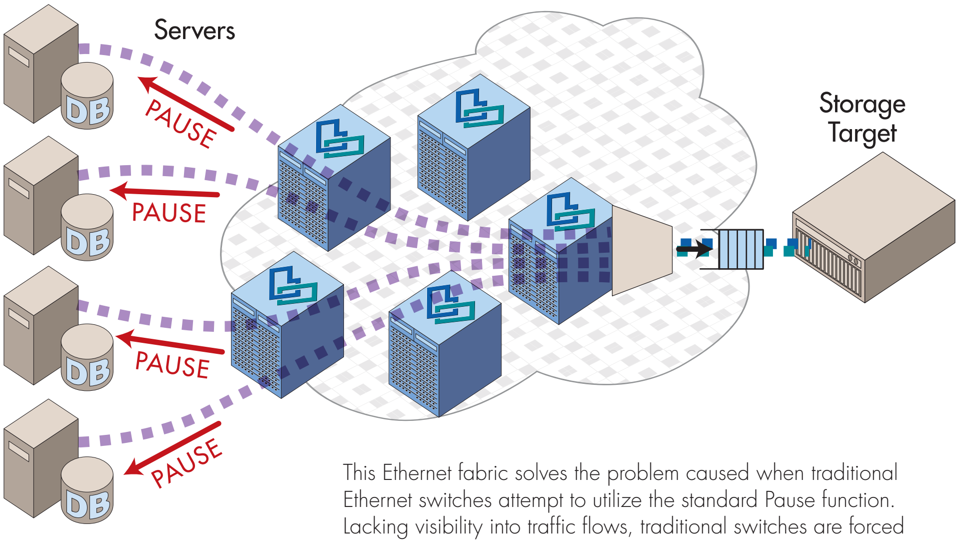


The sample Ethernet fabric shown (with six 6-port Leaf switches and three 6-port Spine switches) shows how traffic is balanced throughout the fabric by rerouting flows.

Flow is rerouted from congested Path A6 (Spine switch A to Leaf switch #6) to an available path, in this case Path B6 (Spine switch B to Leaf switch #6).

Note how the vSCALE ASIC in the Leaf switches provides edge-to-edge congestion management, and that vSCALE is disabled within the fabric.

Although the EFX 1000 is capable of eliminating congestion within the fabric, congestion can still occur for two reasons. The first is lack of sufficient alternate paths within the switching fabric—a situation resulting from the classic price/performance tradeoff made when intentionally oversubscribing a network. The second involves congestion at the target or destination external to the switching fabric. In both cases, the EFX 1000 permits a more intelligent use of Ethernet Pause from the ingress port rather than from the egress port, which eliminates the problem of congestion spreading.



This Ethernet fabric solves the problem caused when traditional Ethernet switches attempt to utilize the standard Pause function. Lacking visibility into traffic flows, traditional switches are forced to issue an Ethernet Pause from the egress port, which can cause congestion spreading back to all ingress ports.

With its edge-to-edge visibility into all traffic flows, the Woven Ethernet Fabric is able to issue an Ethernet Pause from the ingress port only for those resources affected.

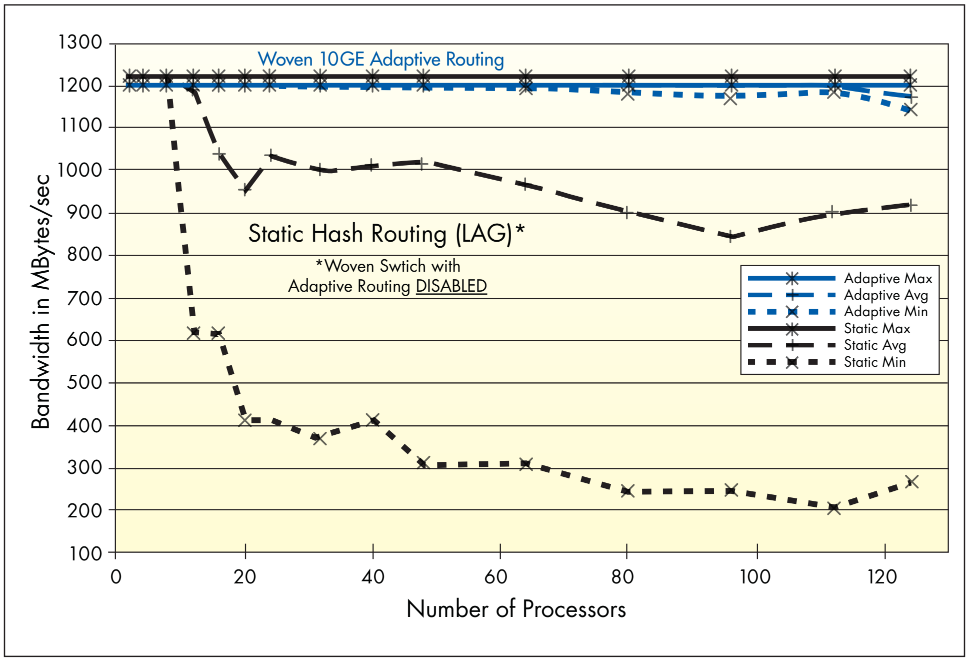
Benchmark Test Results with and without vSCALE

In order to claim industry-leading performance and scalability, Woven Systems conducted extensive testing of the EFX 1000 Ethernet Fabric Switch. Tests conducted both by Woven and by independent laboratories, produced nearly identical results. The Cbench Rotate test results reported here compare the performance of the EFX 1000 switch with the vSCALE ASICs enabled to its performance with vSCALE disabled. The test evaluates the effectiveness of the vSCALE technology itself, and its ability to provide Dynamic Congestion Avoidance that delivers full bandwidth, non-blocking throughput, regardless of traffic patterns.

The Cbench Rotate test was developed at Lawrence Livermore National Laboratory as a way of testing bandwidth and latency between dynamically changing pairs of nodes across a high-performance computing (HPC) interconnect, thereby forcing traffic to take different paths through the switch or fabric topology. This approach takes into account the effects and overhead of dynamic or multi-path routing, which are common in HPC clusters utilizing the Message Passing Interface (MPI). The test, therefore, provides an accurate means for estimating the performance of MPI-based applications that leverage collective communications—from all-to-all to various reductions. The all-inclusive source/destination pairing nature of the test also makes it suitable for assessing performance in storage area network (SAN) and network-attached storage (NAS) applications.

It is important to note that even though disabling vSCALE turns the EFX 1000 into the equivalent of an Ethernet switch built using the latest generation of commercially available Ethernet switching silicon, other switches might not be able to perform as well as this "vanilla" version of the EFX 1000. The reason is Woven's use of a 5-tuple hash that results in greater randomization when calculating the static switching paths. This Layer 4 approach is superior to the usual Layer 2 source/destination hash, which is known to increase congestion by creating common paths even with a relatively modest number of source/destination pairs.

With this more-than-fair bias, Woven's testing provides a valid comparison between the EFX 1000 and other Ethernet switches employing only commercially available Ethernet switching silicon with Link Aggregation Group (LAG), Equal-Cost Multi-Path (ECMP) routing or any other statically-routed protocol.



The test above demonstrates vSCALE's value by measuring throughput in high-performance computing clusters with up to 128 processors. The results compare throughput with the vSCALE technology in the EFX 1000 switch enabled (in blue) to provide Dynamic Congestion Avoidance and disabled (in black).

Disabling vSCALE causes the EFX 1000 to behave as an ordinary Ethernet switch with statically routed LAG paths. Note how the average performance with static paths degrades to less than 70% of the maximum throughput with a cluster size of only around 20 nodes, while vSCALE's Dynamic Congestion Avoidance feature maintains close to line-rate throughput independently of the cluster size and traffic patterns.

Two trends in large-scale data centers make the need to deliver lossless, non-blocking performance across any-to-any connections even more profound: server virtualization and storage consolidation. Virtualization and consolidation, with their constantly changing and unpredictable traffic patterns, significantly exacerbate the problems caused by Ethernet's inherent limitations. Woven's resilient, Ethernet fabric solution, enabled by the vSCALE packet processing ASIC technology, delivers the scale and performance required in data centers supporting large numbers of high performance servers.

Conclusion

A technology like vSCALE is both necessary and sufficient to satisfy all seven fabric requirements for deploying Ethernet in large-scale data centers. It is necessary to overcome the many shortcomings of traditional Ethernet switches. And it is sufficient because Woven Systems has engineered the vSCALE technology to deliver lossless, non-blocking throughput with industry-leading low latency across a multi-path, multi-chassis Ethernet fabric. No other Ethernet switch available comes even close to matching this level of scalable resilience and performance.

Woven's Dynamic Congestion Avoidance, based on vSCALE technology, makes the EFX 1000 Ethernet Fabric Switch suitable for both HPC clusters and storage area networks. As such, vSCALE constitutes nothing short of the breakthrough advance needed for unifying data center networking on a familiar and affordable protocol: Ethernet.

To learn more about how your data center can benefit from deploying the EFX 1000 Ethernet Fabric Switch with breakthrough vSCALE technology, visit Woven Systems on the Web at *www.wovensystems.com* or call +1.408.654.8100.