Data Center Fabric Convergence for Cloud Computing (the Debate of Ethernet vs. Fibre Channel is Over)

The evolution of the data center fabric has been well documented. The vision of a single fabric has been the goal for many years but the realistic challenges of interoperability, capabilities and the true total cost of ownership have prohibited the implementation of this vision for production networks. With the advent of host virtualization and cloud computing, a single fabric becomes a requirement, not just a luxury.
Ethernet for Synchronous Replication vs. Fibre Channel

The convergence challenges are real, especially for data center interconnectivity over WDM. The basic historical premise of having separate Ethernet fabrics for hosts and fibre channel fabrics for storage replication was to avoid disruptions in the fibre channel fabric that was extended over distance from impacting any host I/O. Fibre channel routing and fabric protocols join switches into a larger layer 2 network that maintained a view of all switches in the network. Any disruption to WDM links such as a failure would cause a ripple effect into the fabric and disrupt host I/O. Local application availability or host disruption due to non-network failure was unacceptable.

Fibre Channel over Ethernet (FCoE) was developed as a solution for the converged fabric. The idea is simple; encapsulate Fibre Channel Protocol (FCP) over Ethernet to run fibre channel over the existing Ethernet host network. Alas, the promises of simplicity are prohibited by vendor agendas and interoperability became a major factor of its non-success. Beyond the interoperability challenges were the true costs associated with FCoE, as no cost benefits were being recognized for convergence. Today FCoE is seen as a migration path or investment protection so customers can keep their existing fibre channel storage and let application requirements drive the timing to Ethernet-based storage fabrics.

As virtualization and cloud computing are driving the need for a single physical converged fabric, until now customers have been left with implementing the legacy architecture of separate fabrics.

Extreme Networks OpenFabric, featuring the BlackDiamond® X8, is leading the way to a true physical converged fabric: A single physical fabric for local and remote disk-to-disk connectivity as well as for host-to-disk connectivity. What is new with this architecture is the capability of Ethernet (iSCSI) providing disk-to-disk connectivity for synchronous replication over distance. To this day, the feature of synchronous replication over distance is the limiting factor for a true physical converged fabric for virtualization and cloud computing.

In cooperation with EMC and Superna, we have tested the capabilities of Extreme Networks OpenFabric. To better understand the solution, one must understand the challenge. It is proven that the most sensitive traffic in a data center is typically synchronous SRDF mirroring traffic. This traffic must be transported over the lowest latency transport solution possible to minimize the impact to host I/O latency. Synchronous I/O over SRDF will throttle the host I/O rate. This is why distances for synchronous replication are typically 100-200 km range to limit the impact on host I/O rates which are often transactional database applications.

![Figure 1: Standard Data Center Infrastructure](image1.png)

![Figure 2: All-Ethernet Data Center Architecture](image2.png)
The present mode of operation to transport SRDF for the vast number of deployments is fibre channel switches that use ISL connected to WDM equipment. This requires extra buffer credits on the fibre channel switch to ensure that flow control does not prevent the passing of I/O data due to lack of buffers on the port (buffers traverse the same latency over the fiber that the data travels, which requires a large number of buffers to be available to transmit while other buffers are in flight on the fiber). This is another reason switches are dedicated to replication network traffic to ensure frame buffers are available for high priority traffic. QOS is not typically used in fibre channel switches and interoperability between fibre channel switches is not well tested by vendors, and most SANs are single – vendor solutions. This makes widespread use of QOS low value with no potential for interoperability between vendors.

The key area for SAN extension testing focuses on link stability and link recovery over WDM networks. This testing involves pulling client and line (wide area) fibers and verifying that links recover without any impact to synchronous replication data. With prior testing of fibre channel it has been found that many vendor solutions switch from a working fiber to a backup fibre so that the SRDF and fibre channel switches bounce the link, causing synchronous replication traffic to stop. Several other scenarios were reboots of switches and devices in the path and do not always restore traffic to a working state. This is common in fibre channel if the device has not taken design steps to address these types of issues.

After test completion, it was found that the key test cases passed without any indication of issues that are common in fibre channel networks. For example, the link failure test is designed to run 2 links. We tested 2 protection modes for traffic, link aggregation and ring protection (using only 2 fibers). Both were tested to see the impact of synchronous replication when the active fiber was pulled. The results were that both protection schemes passed with synchronous replication traffic and volume synchronous state staying in sync. There are 2 factors for success: switch speed time to detect the loss of signal in conjunction with the time to switch, and that Ethernet has no convergence concept that pause data in the fabric and more advanced traffic control mechanisms to resolve path switching. In the testing, simple layer 2 mode was used with no MPLS or QOS features enabled. All settings were default with no VLANs used.

In a protected deployment, it is recommended to use VLANs to separate VMAX director ports and apply QOS to enable strict queuing of replication traffic. In our testing we noticed default queuing mechanisms worked well.

The second factor with regard to synchronous replication transport is congestion control. The testing only used QOS features and none of the advanced DCB features. Ethernet has larger frame sizes with jumbo frames (fibre channel MTU is 2148 bytes versus 9000 bytes for jumbo frames). Larger frame size means larger and deeper buffers to handle frame forwarding functions. The comparison to buffer credits with Ethernet is not possible with Ethernet using QOS as primary means to ensure traffic is not dropped. The recent addition of DCB will allow Ethernet networks to enable congestion avoidance features similar to buffer credit management used in Fibre channel.

**Summary of Ethernet Features for Synchronous Replication**

1. Ring protection with multi-node support for equipment and fiber redundancy
2. Link aggregation solution for N x link load sharing and link failure with no impact to synchronous replication traffic
3. 40Gb Ethernet provides faster links with fewer fibers (not tested with SRDF)
   - 120Gb LAG groups (not tested with SRDF)
4. QOS support with inter-vendor solutions and service provider as well as inter-working on WAN services and SLAs
5. Traffic shaping and rate limiting of storage traffic (TCP-based flows will flow control the rate)
6. Data Center Bridging – congestion avoidance plus QOS to enable lossless Ethernet fabric
7. Inter-working with MPLS, VLAN or Layer 2 circuits from service providers at N x 10Gb rates (fibre channel requires WDM solution to map 4G, 8G rates into a network transport solution)
8. DWDM optics available on Ethernet to build WDM networks in the metro with availability of carrier colored wavelength solutions
Ethernet for Host to Disk Connectivity vs. Fibre Channel

This test suite was designed to compare both 8Gb fibre channel and 10Gb iSCSI performance within a VMware ESX host. The test environment used an EMC VMAX target disk presented over a single hop Ethernet switch. The fibre channel path was also over a single hop switch in order to compare I/O latency and throughput in a similar configuration.

Fibre channel has been well deployed for host to disk in virtualized and non-virtualized environments. The fibre channel protocol and fabric switches are designed to provide a robust transport. iSCSI was designed to enable block storage over Ethernet IP networks (layer 3) based on TCP and allow both software initiator solutions and hardware offload solutions to accelerate performance in line with fibre channel. The solution tested was hardware assisted iSCSI. No consideration was made for testing software; only iSCSI initiators since it’s been shown this configuration has no ability to lower CPU utilization of the host adequately enough to ensure high throughput. FCoE is discussed briefly but was not the focus of the testing outlined below. As mentioned in the beginning of this paper, FCoE is considered a transition step to an all-Ethernet fabric and has many of the same advantages as iSCSI at a higher overall cost point. FCoE was not used during this testing.

The testing focused on running tests that compare iSCSI 10Gb and fibre channel 8Gb to comply with the following key attributes: 1) reliability and recovery to link failures, 2) performance with typical application I/O profiles (i.e. database), and 3) QoS meaning ability to ensure host to disk traffic is not impacted when other lower priority traffic (i.e. non-storage or other low priority storage traffic) is competing for link bandwidth between the host and the disk array.

iSCSI Network Design Considerations

1. The goal in host to desk network fabrics is low latency. The test bed used only a single switch between the host and disk for both fibre channel and iSCSI testing. Extreme Networks cut through switching increases performance by reducing latency per I/O transactions. When cascading switches, it’s best to reduce the switch hop count between the host and disk to no more than 5 switches to bound the built-up latency introduced when passing through transit switches.
   • Note: Extreme Networks high density platforms provide the best option to reduce the need to cascade more switches to increase port density.
2. Switching versus routing. Even though iSCSI can be routed, the test bed used Layer 2 switching between the host and disk array, this is comparable to fibre channel networks that do not route fibre channel frames.

   Note: The option to route is very useful in storage VMotion configurations as VMware will leverage routing to move datastores between ESX hosts. This requires a flat layer 2 fibre channel fabric or VSAN to achieve the same thing and this is typically more complicated and expensive to achieve than simply adding routes or VLANs to an existing network.

3. Link redundancy was provided by Ethernet ring protection scheme. LAG could have also been used, but the advantage of a ring-based fabric is the that it requires less fiber connections and takes advantage of spatial ring bandwidth usage that can send traffic over either side of the ring and add or delete nodes from a ring to attach more hosts to the iSCSI network.

4. Jumbo frames (9000 bytes) provide a key advantage on large data reads and are recommended to be enabled on all devices. This is typically not the default setting on most devices and requires specific settings to enable. MTU path discover was not automatic and any device between the OS and the disk that does not support 9000 byte MTU frames caused LUN discovery to fail with no reason that indicated why.

   Note: fibre channels' largest frame size is 2148 bytes.

5. The testing showed resiliency to recovery from link failures and fast LUN rediscovery – this behavior was the same between iSCSI and fibre channel host attached disks. When connected to a VMware host, VMware performs multi-pathing and LUN discover and isolates the host VMware from link errors. It’s important that temporary cable disconnects or path failures do not result in corrupted OS file systems. It was seen that no disk corruption resulted when using VMware to isolate the host OS from the storage layer.

6. QOS tests showed the Extreme Networks default queuing was more than adequate to ensure traffic was not dropped under heavy load. The test specifically attempted leave QOS settings in default settings and run a high rate of UDP traffic parallel to iSCSI host to disk tests and compare the results with and without QOS enabled settings. The setup used only VLAN-based prioritization of data and did not use any other filters to tag flows of data that could be explored to identify iSCSI flows in a more dynamic way.

7. It is recommended to use port-based VLAN QLogic adaptors into a storage VLAN since it is not considered a network interface in VMware. The software VMware Ethernet switch, or vSwitch, cannot be used for tagging externally from the ESX host.

8. 3 iSCSI TOE vendors were tested and only QLogic 8242 was considered adequate to replace a fibre channel interface. This is based on the following reasons:
   - The percentage of iSCSI offload can be TCP fragment and reassemble in hardware plus checksum calculations.
   - SCSI PDU payload assembles and reassembles in hardware plus TCP fragment functionality.
   - Vendors that rely on VMware software initiator drivers care only about performing TCP offload from VMware and we found this was not sufficient to provide the same performance as fibre channel. The reason is fibre channel HBA’s offload all SCSI and fibre channel functions into hardware, versus some vendors’ implementations that depend on the SCSI layer being provided by the OS.

### Summary of Ethernet Features for Host to Disk Connectivity

Testing shows that MTU has a big impact on performance for larger I/O’s and that smaller MTU is beneficial for smaller I/O sizes.

The database pattern I/O testing results show that queuing more I/O’s with the same workload resulted in a small improvement in fibre channel throughput.

Overall, iSCSI outperformed fibre channel for reads and MTU flexibility was a key advantage to allow tuning an OS for the applications needs (i.e. small or big read I/O). Fibre channel has an advantage for large I/O write transactions. Typical applications read the vast majority of the time compared to writes, so in the end the advantage may be a moot point.
The real advantage of iSCSI over fibre channel comes down to many factors outlined below:

1. Single fabric with Ethernet versus dual fabrics (lower cost ports, more tools to manage, existing investment can be leveraged).
2. Routed or layer 2 versus only layer 2 with fibre channel (the requirement to route can assist with islands of storage to be more easily connected and migrated between physical locations).
3. SAN extension is simplified with Ethernet versus fibre channel (WDM or SONET mapping devices).
4. Cost of the HBA, CNA is about the same.
5. VLANs are simpler to implement for all applications versus VSAN and zoning for fibre channel which only address storage applications.
6. Reduced cabling in high density blades.
7. Ethernet tools and IT skills are more readily available, especially open source tools versus fibre channel networks which lack basic things (for example, a ping from host to target capability).
8. Cost performance benefit is an important consideration as the TCO for the vast majority of applications will operate fine with iSCSI.
9. iSCSI can buffer Ethernet off-the-shelf QOS settings, but it was very straightforward to enable strict queuing on storage VLAN and see real-time storage traffic return to normal. Fibre channel on the other hand has no standard-based QOS settings that are customer settable. It’s a one-size-fits-all solution for fibre channel and QOS.