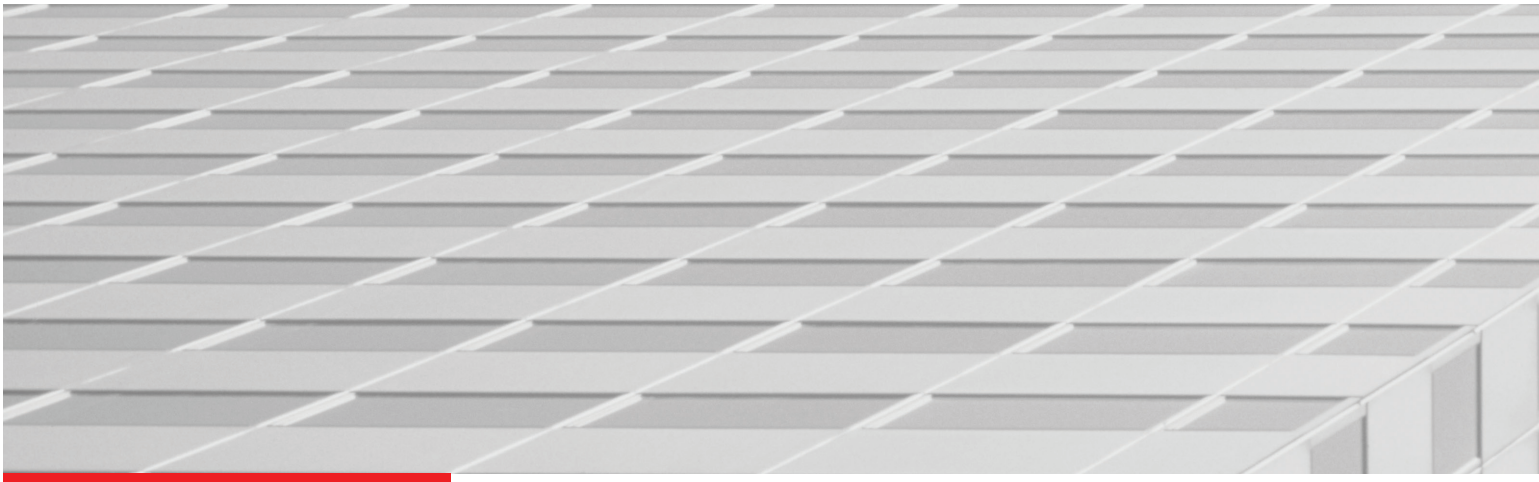

DATA CENTER

The Benefits of a Dedicated IP Network for Storage

Application, storage, and virtualization companies recommend a dedicated IP storage network to ensure deterministic and low latency. The importance of latency on application performance is demonstrated in this white paper.

BROCADE

The scale of IP storage, both block and file, has greatly expanded as more applications use files for managing storage than ever before while Ethernet bandwidth has increased to meet the demands of server virtualization hosting multiple applications per physical server. Today, many customers have as much, or more, NAS storage as they have block storage connected to their applications. In addition, a growing number of applications using IP storage are mission-critical tier 1 applications. This has changed old assumptions about how to design the network that connects application hosts to their NAS storage arrays.

STORAGE CHANNELS AND IP STORAGE NETWORKS

From the early days of computing, storage I/O used dedicated "channels" for moving data blocks between disk drives and the host operating system. In the mid-1980s, an open protocol for storage I/O channels, Small Computer System Interface (SCSI), was introduced. Today this protocol is ubiquitous. SCSI includes an initiator in the host and a target in the disk drive or tape device. I/O is controlled by the operating system kernel, so any application that needs to access data via a read or write calls the operating system SCSI device driver that completes the operation for the application.

Initially, SCSI used a parallel bus of copper wire to connect multiple devices to a single host adaptor. But, as storage volume grew, this dedicated architecture could not keep up with requirements. Two solutions were developed. One was the Fibre Channel Storage Area Network (FC SAN), which acts as a lossless, low latency network for SCSI traffic. Another solution is to use shared file systems such as Network File System (NFS) and (developed later) Common Internet File System (CIFS), which uses a "best effort" IP network with Transmission Control Protocol (TCP), ensuring delivery of the data.

When using a storage network to transport SCSI, SCSI expects certain qualities from the network. In general, since SCSI is a channel protocol, not a client/server, it expects:

- Low, deterministic latency
- Guaranteed delivery

The evolution of dedicated block I/O channels from point-to-point busses into any-to-any networks has met the need for dramatic growth of data storage and larger bandwidth. In a similar way, any-to-any file systems, known as Network Attached Storage (NAS), are common. Instead of block transfer to and from storage, NFS uses a file system such as NFS or CIFS. Both rely on client/server architecture, in which clients make requests to file servers to access, read, and write entire files on the server on behalf of the client. The client operating system calls a TCP/IP network driver to create and manage a hierarchical file system of nested folders on the server and to read and write date from files in those folders and send them to the client.

THE RATIONALE FOR A DEDICATED IP STORAGE NETWORK

Here is a partial list of companies that recommend that IP storage traffic use a dedicated physical network. Note that IP storage includes both block (iSCSI) and file (NAS) storage.

- Apache CloudStack Recommendation:
  http://cloudstack-installation.readthedocs.org/en/latest/choosing_deployment_architecture.html#separate-storage-network
- VMware, Best Practices for Running VMware vSphere on iSCSI—Security Considerations, Private Network:
  http://www.vmware.com/files/pdf/iSCSI_design_deploy.pdf
- Citrix, Citrix Zen Server Design: Designing Zen Server Network Configurations—Chapter 6: Designing Your Storage Network Configuration:
  http://support.citrix.com/servlet/KbServlet/download/27046-102-666250/XS-design-network_advanced.pdf
- EMC, Clariion Best Practices for Performance and Availability; Release 30.0 Firmware Update, Applied Best Practices—Chapter 3: Network Best Practices:
  http://www.emc.com/collateral/hardware/white-papers/h5773-clariion-best-practices-performance-availability-wp.pdf
- IBM, Redbook, b-type Data Center Networking: Design and Best Practices Introduction—Chapter 9: IP Storage Area Networks, 9.4.4 Challenges in iSCSI SANs:
  http://books.google.com/books?id=QVbAAgAAQBAJ&pg=PA375&lpg=PA375&dq=IBM+separate+iscsi+network+best+practices&source=bl&ots=nX85tC_1z5&sig=f0Y9561MUr8ZOIVhKYukxVWWiOY&hl=en&sa=X&ei=DSLpU6PzL873yQSL5YDYCA&ved=0CDAQ6AEwAw#v=onepage&q=IBM%20separate%20iscsi%20network%20best%20practices&f=false
- NetApp, NetApp and VMware vSphere Storage Best Practices—Chapter 3: Storage Network Design and Setup, 3.3 Ethernet Storage Networking Basics, Separate Ethernet Storage Network:
  http://community.netapp.com/fukiw75442/attachments/fukiw75442/fas-and-v-series-storage-systems-discussions/2645/1/NetAppandVMwarevSphereStorageBestPracticesJUL10.pdf

The following are the primary reasons many vendors recommend a separate network for IP storage:

- Low, deterministic latency
- Guaranteed delivery
- Smaller administrative domain, which is easier to troubleshoot
- Fewer configuration compromises
- Better fault isolation
- Less complexity to upgrade and maintain

Storage I/O has to complete, again with low latency, so a network with guaranteed delivery is very important. In fact, this is a weakness when using Ethernet and IP as the transport for storage traffic, as they are "best effort" delivery transports, not guaranteed. This means that TCP has to handle the guaranteed delivery, which can be very difficult if congestion occurs. For example, compared to Fibre Channel block storage networks, IP networks rely on TCP for reliable delivery, but this can require multiple retries when congestion occurs. With more flows, any port congestion can trigger TCP slow start with considerable delay in completing storage I/O.

A single storage administrative domain simplifies management and troubleshooting. Consequently, a dedicated storage network simplifies storage administration, by eliminating "finger pointing" or split responsibilities when the storage administrator manages both the storage and its transport network.

In many cases, comingling storage with other network traffic results in network configuration compromises that can complicate storage management. For example, TCP sliding window settings and round-trip timers affect the time required to recover a lost packet. Changing these to accommodate multiple traffic types can add complexity to the network configuration.

Anyone who has diagnosed technical problems has found that the smaller the domain of analysis, the easier it is to perform the analysis. This supports the idea of a dedicated network for storage traffic, if you assume that the capital cost of the network is comparable to the cost of expanding an existing network to handle the increased storage traffic.

Also, upgrading and maintaining a large network can become complicated. It is easier to maintain smaller scale networks that are dedicated to storage traffic than to maintain a large complex network that handles comingled storage and general-purpose network traffic.

Although these requirements seem simple, they place strict limits on how the network is designed and what the network protocols must be capable of. Consider an analogy: The human body contains a number of "networks," for instance, circulatory, nervous, and lymphatic. Each "network" connects to many of the same cells throughout the body, yet each is physically isolated from the other. One explanation for the three separate networks is that comingling them would place so many demands on the design that it would be difficult, if not impossible, to meet the collective requirements efficiently without compromising the health of a large, multicellular organism.

Similarly, comingling storage traffic on a single network that also carries application traffic, client/server traffic, and web traffic might place requirements on the design that could not be met efficiently and cost-effectively, due to the various constraints of each traffic type. For example, application performance is gated by many things, but slow response to a read or write of data in storage is devastating to the network. Therefore, low latency is essential. Similarly, deterministic latency that does not vary much is also highly desirable to ensure predictable application performance as the workload varies.

HOW SMALL EVENTS CAN CAUSE LARGE PROBLEMS FOR IP STORAGE

Since TCP is the protocol that is commonly used with NAS in a client/server design, it is important to appreciate how TCP reacts to momentary congestion events anywhere along the data path. In larger scale data centers, the network often uses a three-tier topology. The connections between switches and routers are called Inter-Switch Links (ISLs). These links multiplex all traffic flowing between the tiers. Consequently, ports on either side of an ISL link are more subject to congestion than ports connected to server Network Interface Cards (NICs) or storage ports.

When a diverse mix of application traffic is traversing the network tiers, sudden bursts in traffic from a few applications create a spike in frame forwarding at an ISL port. This can cause momentary congestion on a switch or router ISL port. When that happens, TCP attempts to recover from the congestion by dropping frames and throttling back the allowed frame forwarding rate on the port, to relieve the congestion. TCP then increases the flow over a period of time until the port is able to forward the presented traffic rate without congestion. This is called "slow start" and can take seconds before full flow is reestablished after a congestion event.

Therefore, for a comingled flow, where frames on an ISL link include application and IP storage traffic, the storage frames can be discarded when TCP discards frames, so they are not delivered. Then latency for all traffic on that port, including the storage traffic, can go up dramatically as TCP slow start recovery takes effect. The result is that while this process is taking place, delivery is held up, and latency gets very large and becomes quite variable.

Measuring the Impact of Momentary Congestion on IP Storage Latency

To see what can happen in a data center network, with multiple applications and NAS storage traffic sharing the same ISL links, Brocade conducted tests using the well-known VMware VMmark test suite. VMmark simulates a variety of typical applications running in Virtual Machines (VMs) and subjects them to varying workloads that are typically representative for each application. As the presented client workload varies, the storage I/O generated by the server to the backing store also varies.

Two measurements are taken, I/O latency and total I/O bandwidth. Latency is important when considering storage performance. It is correlated with transaction response time and is a proxy for application transactions per second. The business value of applications is directly related to the ability to process as many transactions as possible. Therefore, high and variable latency means lower transaction performance and less business value from the application. This is a well-known metric, particularly for web e-commerce transactions, where customer patience is low, and stalled application response can drive customers to a competitor's site.

The test network was configured in a two-tier, leaf-spine topology with Brocade® VDX® switches configured using Brocade VCS® Fabric technology mode. Brocade VCS Fabric technology is the Brocade implementation of an Ethernet fabric and is designed to optimize Layer 2 traffic while avoiding Spanning Tree Protocol (STP) limitations. Thus, Brocade VCS Fabric technology is an optimized network with high performance, resiliency, and simple management and configuration.

Note: A Brocade VCS fabric has unique capabilities, such as ISL trunks, that frame stripe all traffic over all links in the trunk, as well as the ability to use all equal-cost paths between end devices. These features were not used in this test.

Figure 1 shows the test configuration.

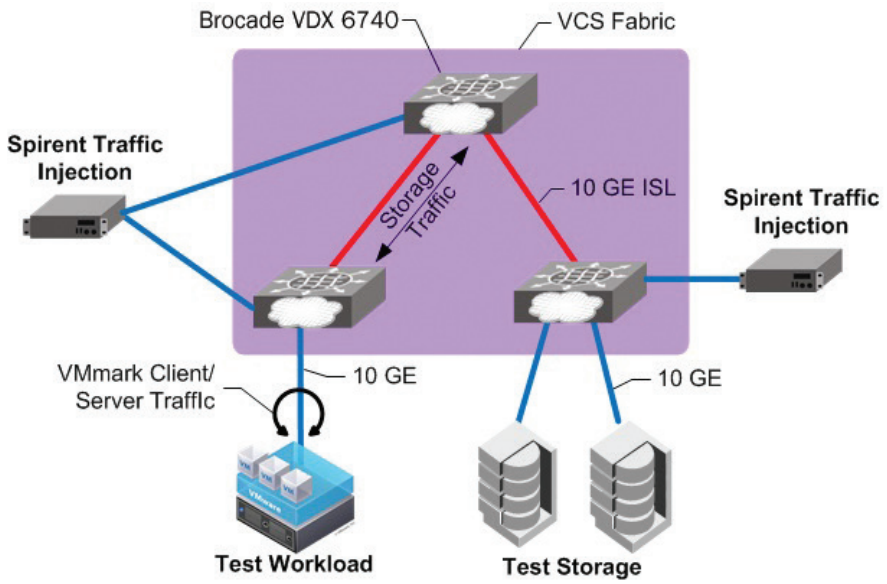

Figure 1.
Test configuration.

Although this configuration does not include multiple ISLs combined into a Link Aggregation Group (LAG) or port channel for resiliency, the behavior of IP storage traffic under congestion is well represented. Layer 2 networks with STP allow only a single path between switches. A LAG allocates a single flow to a single link that is shared with other flows on that link. Thus, congestion created by traffic bursts on a single link affects any storage traffic using that link, when a link is part of a LAG or port channel. Therefore, the test configuration is a good representation of most existing networks supporting IP storage traffic without the complexity of setting up a LAG.

Base Test Case

The base test case used the VMmark load generator and collected disk latency at the ESXi NIC for each application. The configuration isolated the application's client/server traffic to the hypervisor on a single ESXi server, as both the clients and servers were running in VMs on a single ESXi server. However, storage traffic between the server and the backing store traversed the VCS fabric. Therefore, the VCS fabric is carrying only NAS traffic. In the base test, no traffic was inserted into the VCS fabric by the Spirent traffic-generation tool.

Base Test Case Latency and Bandwidth

Figure 2 and Figure 3 show the latency measured during VMmark benchmark testing for the DVD Store (e-commerce store simulation) and the Olio (Web 2.0 simulation) applications in the VMmark test suite.

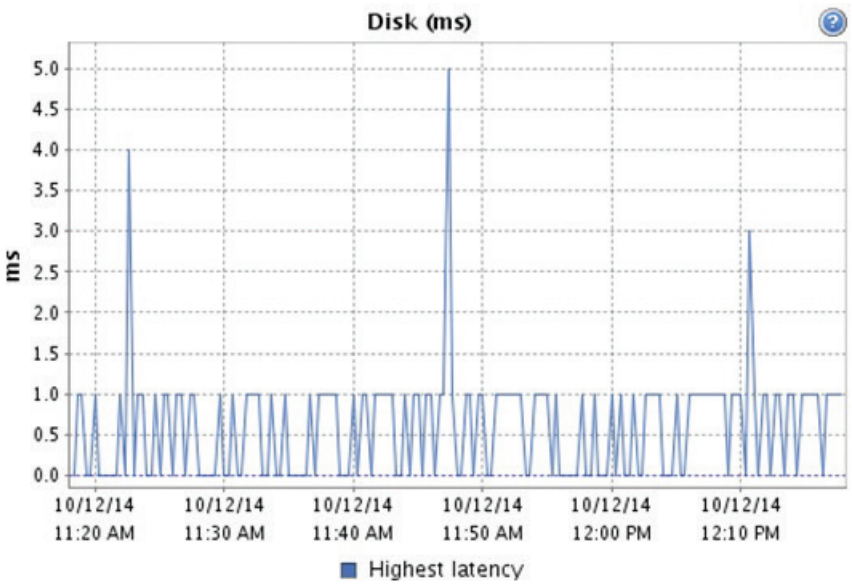

Figure 2.
DVD Store ESXi disk I/O latency, base case.

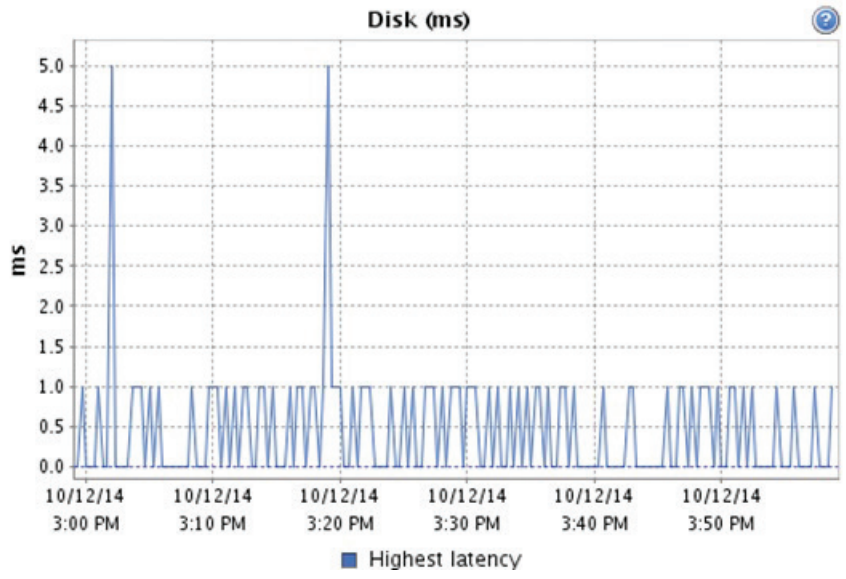

Figure 3.
Olio ESXi disk latency, base case.

As these measurements show, disk latency is excellent. It is consistently low, with only a few minor spikes occurring when the VMmark work load driver occasionally ramps up the application work load.

The corresponding disk bandwidth for the base case is shown in Figure 4 and Figure 5, for both workloads. Note that the storage I/O is quite low, only occasionally hitting 1 MB/s based on the VMmark load generation for the DVD Store and Olio applications.

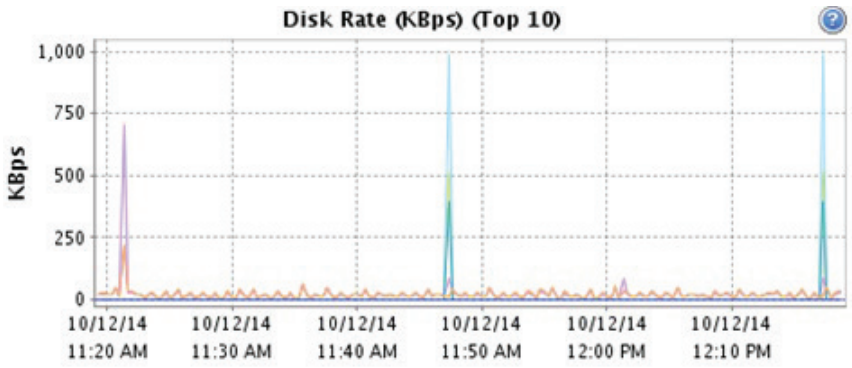

Figure 4.
DVD Store ESXi disk bandwidth, base case.

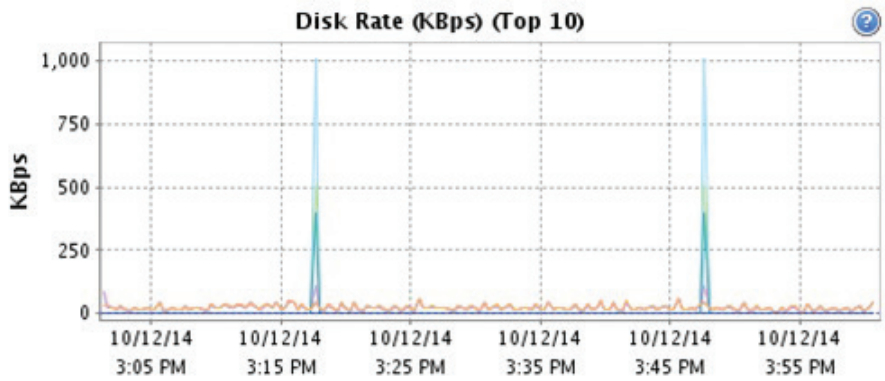

Figure 5.
Olio ESXi disk bandwidth, base case.

Note that the maximum ISL bandwidth available is 10 gigabits per second (Gbps) but the actual storage bandwidth for these two applications is quite low: from 20 kilobytes per second (KBps) (or 160 kilobits per second, which is abbreviated as kbps) with occasional 30-second spikes up to 700 to 1,000 KBps (5,600 kbps to 8,000 kbps). Although not shown, the cumulative storage bandwidth from all the applications in the VMmark benchmark used no more than 10 percent to 20 percent of the available bandwidth of a 10 Gigabit Ethernet (GbE) NIC installed in the ESXi server.

Momentary Congestion Test Case

A series of tests were conducted using the Spirent test tool to inject traffic into the Brocade VDX switches, so that congestion occurs on the ISL paths during the VMmark test. A square wave cycle of traffic was injected that ranged from 1 second on, 1 second off up to 20 seconds on, 20 seconds off. During each 30-minute VMmark test, only one square wave duration was used, for example, 1 second on, 1 second off during a VMmark test, and then another VMmark test that used a duration of 5 seconds on, 5 seconds off, and so forth. Examples of traffic that might cause these spikes include desktop computer backups, YouTube video traffic, sending large file attachments, uploading large documents to cloud storage services such as Dropbox, and so on.

In the congestion tests, the long-term average aggregate bandwidth (storage plus Spirent traffic) on the ISL paths did not exceed 60 percent of the available bandwidth of a 10 GbE ISL port when measured over several minutes. Figure 6 shows the VMmark test traffic (blue) with the Spirent traffic generator traffic spikes (grey bars) superimposed. You can see that the long-term average Spirent traffic (red line) is 50 percent, while the long-term average aggregate traffic does not exceed 60 percent of the available link capacity.

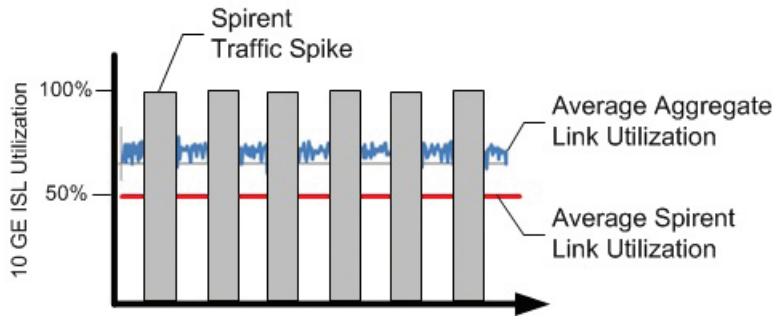

Figure 6.
Congestion test: Average ISL link utilization.

On a longer term average of a few minutes, the ISL links are not congested. However, during the short duration of the Spirent traffic injection, they are congested. This simulates the effect of spikes in traffic from other applications that share the network with the NAS traffic.

Congestion Test Case Latency and Bandwidth

Figure 7 and Figure 8 show the latency measured for both the DVD Store and Olio ESXi servers for a typical congestion test run.

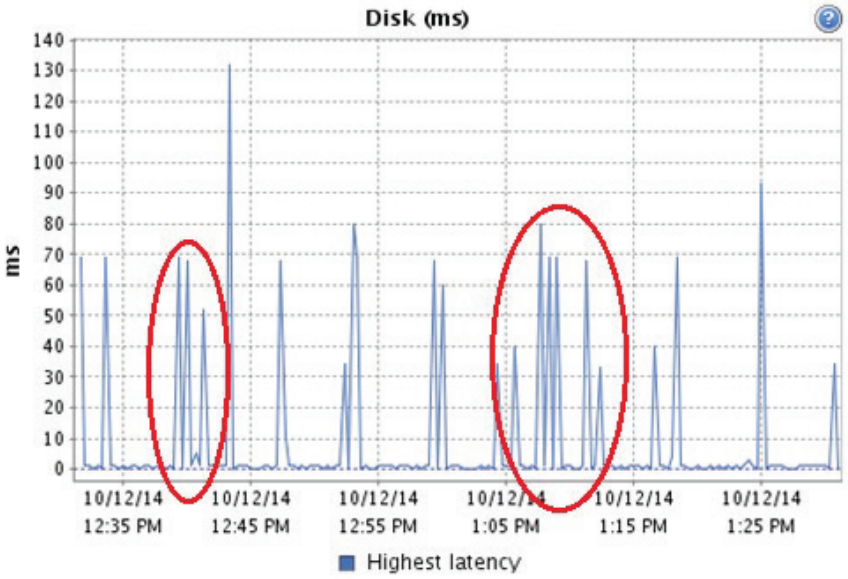

Figure 7.
DVD Store ESXi Disk I/O latency, Spirent traffic spikes.

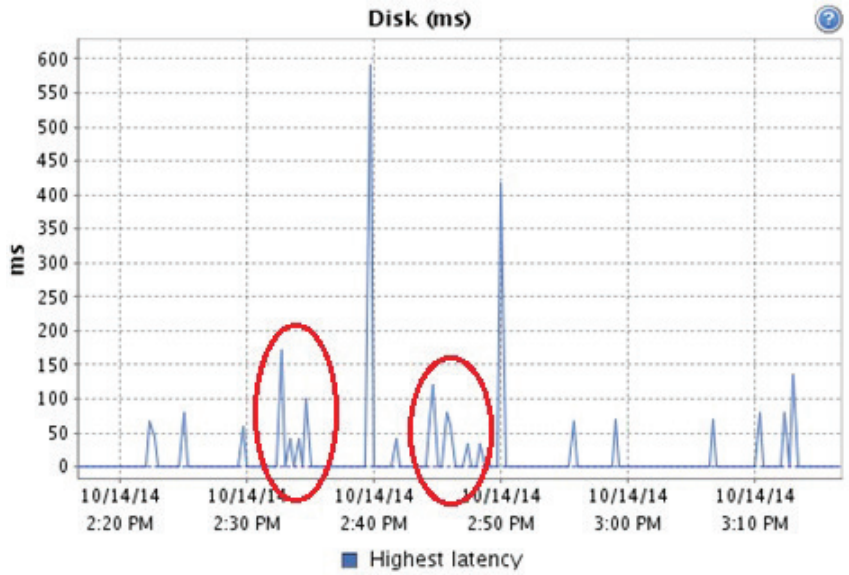

Figure 8.
Olio ESXi Disk I/O latency, Spirent traffic spikes.

The latency for the DVD Store application shows sustained spikes for several minutes, which are on the order of 50 to 70 milliseconds (ms), or 10 to 70 times the latency seen in the base case. Note that some of the spikes cluster for several minutes (as shown by the red ovals) before they go back down to the levels of the base case.

The Olio application shows much worse latency, rising as high as 580 ms for as long as a minute. This application also shows clusters of high latency that last several minutes, from 40 to 170 milliseconds.

For both applications, transaction response is slowed down by factors between 10 and almost 100. When these delays occur at especially busy times, such as during a special sale or the holidays, the economic impact can be damaging to business.

Figure 9 and Figure 10 shows the bandwidth measurements for the DVD Store and Olio applications when the Spirent is injecting period traffic spikes.

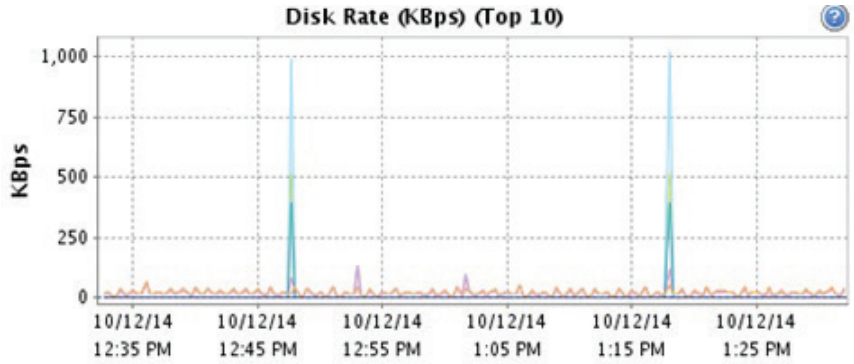

Figure 9.
Olio ESXi Disk I/O latency, Spirent traffic spikes.

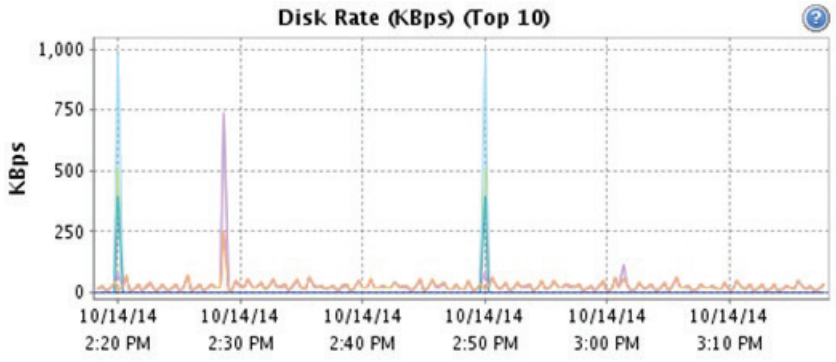

Figure 10.
Olio ESXi disk bandwidth, base case.

Notice that the bandwidth did not change compared to the base test. This demonstrates that relying only on bandwidth as a measure of how well the network meets an application's storage I/O requirements is not a complete assessment. The latency measurement for the I/O requests is also required, to see everything that is occurring. Although networks are designed to deliver bandwidth, the impact of latency on critical IP storage workloads is much more important, and commonly overlooked. This is an important difference between designing a network for printer traffic and web applications and designing a network for tier 1 applications storage traffic.

As the test data illustrates, even small momentary congestion events on ISL ports can result in large, detrimental reductions in application transaction performance even when the average bandwidth required is only 50 percent to 60 percent of available link bandwidth. Although the average bandwidth required does not exceed the average capacity, the short duration traffic spikes result in support calls and complaints about the "slow" response of business applications.

CONCLUSION

In conclusion, for larger environments, there are compelling advantages to having a separate physical network for IP storage traffic, regardless of whether you use block storage (iSCSI) or NAS storage (CIFS and NFS). These advantages include:

- Low, deterministic latency
- Guaranteed delivery
- Smaller administrative domain, which is easier to troubleshoot
- Fewer configuration compromises
- Better fault isolation
- Less complexity to upgrade and maintain

Of these advantages, low and deterministic latency is dramatically affected by small-duration traffic spikes resulting in dramatic decreases in the transaction rate for applications as they wait for storage I/O to complete. TCP is required to achieve guaranteed delivery, but the side effects of the slow start mechanism can cause orders of magnitude increases in I/O latency for minutes at a time, dramatically slowing application response. The process of diagnosing and eliminating this behavior can be very complicated and time consuming. Therefore, simplifying the data center network design by using a dedicated IP storage network for both block and NAS storage pays dividends. This is why leading application vendors recommend this as a design best practice.

A Brocade VCS fabric is well suited for IP storage traffic and is a particularly good choice for attaching NAS servers and the highest I/O NAS clients, because a VCS fabric can minimize the impact of transient congestion events as compared to classic Ethernet with STP. Interoperability with classic Ethernet 802.1 protocols and 802.3ad standard LAG simplifies upgrading critical portions of existing networks with a VCS fabric.

ABOUT BROCADE

Brocade networking solutions help organizations achieve their critical business initiatives as they transition to a world where applications and information reside anywhere. Today, Brocade is extending its proven data center expertise across the entire network with open, virtual, and efficient solutions built for consolidation, virtualization, and cloud computing. Learn more at www.brocade.com.

Corporate Headquarters
San Jose, CA USA
T: +1-408-333-8000
info@brocade.com

European Headquarters
Geneva, Switzerland
T: +41-22-799-56-40
emea-info@brocade.com

Asia Pacific Headquarters
Singapore
T: +65-6538-4700
apac-info@brocade.com

BROCADE