

Remaking the Data Center

Low-latency switches are the foundation for building a unified-fabric data center

By Robin Layland, Network World
September 17, 2009

A major transformation is sweeping over data center switching. Over the next few years the old switching equipment needs to be replaced with faster and more flexible switches.

Three factors are driving the transformation: server virtualization, direct connection of Fibre Channel storage to the IP switching and enterprise [cloud computing](#).

They all need speed and higher throughput to succeed but unlike the past it will take more than just a faster interface. This time speed needs to be coupled with lower latency, abandoning spanning tree and supporting new storage protocols.

Without these changes, the dream of a more flexible and lower cost data center will remain just a dream. Networking in the data center must evolve to a unified switching fabric.

Times are hard, money is tight; can a new unified-fabric really be justified? The answer is yes. The cost savings from supporting [server virtualization](#) along with merging the separate IP and storage networks is just too great. Supporting these changes is impossible without the next evolution in switching. The good news is that the switching transformation will take years, not months, so there is still time to plan for the change.

The Drivers

The story of how server virtualization can save money is well known. Running a single application on a server commonly results in utilization in the 10% to 30% range. Virtualization allows multiple applications to run on the server within their own image, allowing utilization to climb into the 70% to 90% range. This cuts the number of physical servers required; saves on power and cooling and increases operational flexibility.

The storage story is not as well known, but the savings are as compelling as the virtualization story. Storage has been moving to IP for years, with a significant amount of storage already attached via NAS or iSCSI devices. The cost saving and flexibility gain is well known.

The move now is to directly connect Fibre Channel storage to the IP switches, eliminating the separate Fibre Channel storage-area network. Moving Fibre Channel to the IP infrastructure is a cost saver. The primary way is by reducing the number of adapters on a server. Currently servers need an Ethernet adapter for IP traffic and a separate storage adapter for the Fibre Channel traffic. Guaranteeing high availability means that each adapters needs to be duplicated resulting in four adapters per server. A unified fabric reduces the number to two since the IP and Fibre Channel or iSCSI traffic share the same adapter. The savings grow since halving the number of

adapters reduces the number of switch ports and the amount of cabling. It also reduces operational costs since there is only one network to maintain.

The third reason is internal or enterprise cloud computing. In the past when a request reached an application, the work stayed within the server/application. Over the years, this way of design and implementing applications has changed. Increasingly when a request arrives at the server, the application may only do a small part of the work; it distributes the work to other applications in the data center, making the data center one big internal cloud.

Attaching storage directly to this IP cloud only increases the number of critical flows that pass over the switching cloud. It becomes critical that the cloud provide very low latency with no dropped packets. A simple example shows why low latency is a must. If the action took place within the server, then each storage get would only take a few microseconds to a nanosecond to perform. With most of the switches installed in enterprises the get can take 50 to 100 microseconds to cross the cloud, which depending on the number of calls adds significant delays to processing. If a switch discards the packet, the response can be even longer. The only way internal cloud computing works is with a very low latency and non-discarding cloud.

What is the problem for the network? Why change the switches?

Why can't the current switching infrastructure handle virtualization, storage and cloud computing? Compared with the rest of the network the current data center switches provide very low latency, discard very few packets and support 10 Gigabit Ethernet interconnects. The problem is that these new challenges need even lower latency, better reliability, higher throughput and support for Fibre Channel over Ethernet (FCoE) protocol.

The first challenge is latency. The problem with the current switches is that they are based on a store-and-forward architecture. Store-and-forward is generally associated with applications such as e-mail where the mail server receives the mail, stores it on a disk and then later forwards it to where it needs to go. Store-and-forward is considered very slow. How are layer 2 switches, which are very fast, store-and-forward devices?

Switches have large queues. When a switch receives a packet, it puts it in a queue, and when the message reaches the front of the queue, it is sent. Putting the packet in a queue is a form of store-and-forward. A large queue has been sold as an advantage since it means the switch can handle large bursts of data without discards.

The result of all the queues is that it can take 80 microseconds or more for a large packet to cross a three-tier data center. The math works as follows. It can take 10 microseconds to go from the server to the switch. Each switch to switch hop adds 15 microseconds and can add as much as 40 microseconds. For example, assume two servers are at the "far" end of the data center. A packet leaving the requesting server travels to the top of rack switch, then the end-of-row switch and onward to the core switch. The hops are then repeated to the destination server. That is four switch-to-switch hops for a minimum of 60 microseconds. Add in the 10 microseconds to reach each server and the total is 80 microseconds. The delay can increase to well over 100

microseconds and becomes a disaster if a switch has to discard the packet, requiring the TCP stack on the sending server to time out and retransmit the packet.

Latency of 80 microseconds each way was acceptable in the past when response time was measured in seconds, but with the goal to provide sub-second response time, the microseconds add up. An application that requires a large chunk of data can take a long time to get it when each get can only retrieve 1,564 bytes at a time. A few hundred round trips add up. The impact is not only on response time. The application has to wait for the data resulting in an increase in the elapsed time it takes to process the transaction. That means that while a server is doing the same amount of work, there is an increase in the number of concurrent tasks, lowering the server overall throughput.

The new generation of switches overcomes the large latency of the past by eliminating or significantly reducing queues and speeding up their own processing. The words used to describe it are: lossless transport; non-blocking; low latency; guaranteed delivery; multipath and congestion management. Lossless transport and guaranteed delivery mean they don't discard packets. Non-blocking means they either don't queue the packet or have a queue length of one or two.

The first big change in the switches is the design of the way the switch forwards packets. Instead of a store-and-forward design, a cut-through design is generally used, which significantly reduces or eliminates queuing inside the switch. A cut-through design can reduce switch time from 15 to 50 microseconds to 2 to 4 microseconds. Cut-through is not new, but it has always been more complex and expensive to implement. It is only now with the very low latency requirement that switch manufacturers can justify spending the money to implement it.

The second big change is abandoning spanning tree within the data center switching fabric. The new generation of switches use multiple paths through the switching fabric to the destination. They are constantly monitoring potential congestion points, or queuing points, and pick the fastest and best path at the time the packet is being sent. Currently all layer 2 switches determine the "best" path from one end-point to another one using the spanning tree algorithm. Only one path is active, the other paths through the fabric to the destination are only used if the "best" path fails. Spanning tree has worked well since the beginning of layer 2 networking but the "only one path" is not good enough in a non-queuing and non-discarding world.

A current problem with the multi-path approach is that there is no standard on how they do it. Work is underway within standard groups to correct this problem but for the early versions each vendor has their own solution. A significant amount of the work falls under a standard referred to as Data Center Bridging (DCB). The reality is that for the immediate future mixing and matching different vendor's switches within the data center is not possible. Even when DCB and other standards are finished there will be many interoperability problems to work out, thus a single vendor solution may be the best strategy.

Speed is still part of the solution. The new switches are built for very dense deployment of 10 Gigabit and prepared for 40/100 Gigabit. The result of all these changes reduces the trip time

mentioned from 80 microseconds to less than 10 microseconds, providing the needed latency and throughput to make fiber channel and cloud computing practical.

Virtualization curve ball

Server virtualization creates additional problems for the current data center switching environment. The first problem is each physical server has multiple virtual images, each with their own media access control (MAC) address. This causes operational complications and is a real problem if two virtual servers communicate with each other. The easiest answer is to put a soft-switch in the VM, which all the VM vendors provide. This allows the server to present a single MAC address to the network switch and perform the functions of a switch for the VMs in the server.

There are several problems with this approach. The soft switch needs to enforce policy and access control list (ACL); make sure VLANs are followed and implement security. For example, if one image is compromised, it should not be able to freely communicate with the other images on the server, if policy says they should not be talking to each other.

If they were on different physical servers the network would make sure policy and security procedures were followed. The simple answer is that the group that maintains the server and the soft switch needs to make sure all the network controls are followed and in place. The practical problem with this approach is the coordination required between the two groups and the level of knowledge of the networking required by the server group. Having the network group maintain the soft switch in the server creates the same set of problems.

Today, the answer is to learn to deal with confusion and develop procedures to make the best of the situation and hope for the best. A variation on this is to use a soft switch from the same vendor as the switches in the network. The idea is that coordination will be easier since the switch vendor built it and has hopefully made the coordination easier. Cisco is offering this approach with VMware.

The third solution is to have all the communications from the virtual server sent to the network switch. This would simplify the switch in the VM since it would not have to enforce policy, tag packets or worry about security. The network switch would perform all these functions as if the virtual servers were directly connected to the servers and this was the first hop into the network.

This approach has appeal since it keeps all the well developed processes in place and restores clear accountability on who does what. The problem is spanning tree does not allow a port to receive a packet and send it back on the same port. The answer is to eliminate the spanning tree restriction of not allowing a message to be sent back over the port it came from.

Spanning Tree and virtualization

The second curve ball from virtualization is ensuring that there is enough throughput to and from the server and that the packet takes the best path through the data center. As the number of processors on the physical server keep increasing, the number of images increase, with the result

that increasingly large amounts of data need to be moved in and out of the server. The first answer is to use 10 Gigabit and eventually 40 or 100 Gigabit. This is a good answer but may not be enough since the data center needs to create a very low latency, non-blocking fabric with multiple paths. Using both adapters attached to different switches allows multiple paths along the entire route, helping to ensure low latency.

Once again spanning tree is the problem. The solution is to eliminate spanning tree, allowing both adapters to be used. The reality is the new generation layer 2 switches in the data center will act more like routers, implementing their own version of OSPF at layer 2.

Storage

The last reason new switches are needed is Fibre Channel storage. Switches need to support the ability to run storage traffic over Ethernet/IP such as NAS, ISCSI or FCoE. Besides adding support for the FCoE protocol they will also be required to abandon spanning tree and enable greater cross sectional bandwidth. For example Fibre Channel requires that both adapters to the server are active and carrying traffic, something the switch's spanning tree algorithm doesn't support. Currently the FCoE protocol is not finished and vendors are implementing a draft version. The good news is that it is getting close to finalization.

Current state of the market

How should the coming changes in the data center affect your plan? The first step is to determine how much of your traffic needs very low latency right now. If cloud computing, migrating critical storage or a new low latency application such as algorithmic stock trading is on the drawing board, then it is best to start the move now to the new architecture. Most enterprises don't fall in that group yet but they will in 2010 or 2011 and thus have time to plan an orderly transformation.

The transformation can also be taken in steps. For example, one first step would be to migrate Fibre Channel storage onto the IP fabric and immediately reduce the number of adapters on each server. This can be accomplished by replacing just the top of the rack switch. The storage traffic flows over the server's IP adapters and to the top of the rack switch which send the Fibre Channel traffic directly to the SAN. The core and end of rack switch do not have to be replaced. The top of the rack switch supports having both IP adapters active for storage traffic only with spanning tree's requirement of only one active adapter applying to just the data traffic. Brocade and Cisco currently offer this option.

If low latency is needed, then all the data center switches need to be replaced. Most vendors have not yet implemented the full range features needed to support the switching environment described here. To understand where a vendor is; it is best to break it down into two parts. The first part is whether the switch can provide very low latency. Many vendors such as Arista Networks, Brocade, Cisco, Extreme, Force 10 and Voltaire have switches that can.

The second part is whether the vendor can overcome the spanning tree problem along with support for dual adapters and multiple pathing with congestion monitoring. As is normally the

case vendors are split on whether to wait until standards are finished before providing a solution or provide an implementation based on their best guess of what the standards will look like. Cisco and Arista Networks have jumped in early and provide the most complete solutions. Other vendors are waiting for the standards to be completed in the next year before releasing products.

What if low latency is a future requirement, what is the best plan? Whenever the data center switches are scheduled for replacement they should be replaced with switches that can support the move to the new architecture and provide very low latency. This means it is very important to understand the vendor's plans and migration schemes that will move you to the next generation unified fabric.

Layland is head of Layland Consulting. He can be reached at robin@layland.com.