Designing the Flat Data Center Network

Gary Lee, Fulcrum Microsystems

There has been a lot of discussion recently about flat data center networks, which is a major shift from the deployment of traditional multi-layer enterprise network architectures that are no longer viewed as an efficient solution for large data centers due to their high latencies and complex software. The proposed replacement data center network architectures will be relatively flat and interconnect basic units such as virtual machines and virtual storage volumes across large, switched Ethernet fabrics.

The evolution of the needs of the data center and of Ethernet technology have brought the industry to a place where Ethernet has become the most widely deployed networking technology for LANs, interconnecting everything from servers to storage. Most of this technology was borrowed from the enterprise network, where latency and packet loss are less of a concern. The typical enterprise network supports islands of clients typically formed around departmental or functional boundaries. Because of this, enterprise networks consist of multiple layers, including access, aggregation and core. For example, all the clients within a department may be connected to a given access switch, with multiple departments connected through an aggregation switch and multiple floors in the building connected through a core switch. Although this works fine for enterprise networks, it is not an ideal architecture for the data center.

Large data centers do not have the hierarchy seen in the enterprise. Data center administrators need to dynamically re-assign resources such as virtual machines or storage volumes from any point within the data center. To them, the data center network should look like one large homogeneous interconnect fabric. But as data centers have expanded to support tens of thousands of endpoints, the only choice has been to repurpose enterprise networking gear. This complex three-tier network architecture has high latency, and is costly to maintain. In addition, separate storage networks are typically used, adding further to complexity and cost.

New Data Center Networking Requirements

Several new technology trends are emerging that will allow data center network planners to implement flat, converged data center networks. By mid 2012, analysts predict that most servers will ship with 10GbE LAN on motherboard (LOM). This allows the convergence of protocols such as iSCSI and FCoE along with data traffic on a single server port. But this also puts pressure on the network to support more cross-sectional bandwidth. The Spanning Tree protocol, which has been used successfully in enterprise networks, leaves a lot of bandwidth unused, and does not provide fast fail-over mechanisms. Because of this, specialized data center protocols such as Transparent Interconnection of Lots of Links (TRILL) and shortest path bridging (SPB) are being proposed, which replace STP with multi-pathing and redundancy to increase overall network bandwidth and availability.

This additional bandwidth also allows network convergence; but for storage traffic to co-exist with data traffic on the same network, lossless operation and bounded latency must be provided. To support this, the IEEE has developed as set of Data Center Bridging (DCB) standards, which include Priority Flow Control (PFC) for lossless operation and Enhanced Transmission Selection (ETS) for bounded latency (minimum bandwidth guarantees). Other DCB standards include Quantized Congestion Notification (QCN) and DCB exchange protocol (DCBx). PFC, ETS and DCBx are being widely supported by the industry and are finding their way into the latest data center networking products.

Why Latency Matters in the Data Center

Large three-tier enterprise networks have latencies in the hundreds of microseconds, which is becoming unacceptable for many data center applications. An example of this is web content delivery. To maintain the best user experience, web content must be served up within certain time constraints. As pointed out in a recent paper from Microsoft and Stanford, “With such tight deadlines, network delays within the data center play a significant role in application design.”

This web service data center must complete an ever-increasing number of tasks within a limited time period. For example, as you search for products in an online catalog, the items you search for may be used in a profile algorithm that brings up advertisements for related products. These types of algorithms are only going to increase in number and complexity over time (think of the advanced search algorithms and specialized hardware that Google uses). The way the data center handles this is to spawn multiple workflows on various virtual machines that are connected through Ethernet switches. Given the user response time limits, low latency switches will allow more sophisticated workflows, increasing revenues for the web server client.

Another promising revenue source for the cloud service provider is compute clustering. Organizations are starting to look at using compute clusters from cloud service providers. As an example, Amazon started offering compute clustering services last year. The performance of these cluster services depends directly on the data center network latency; and to clients using these services, time is money.

Data Center Virtualization

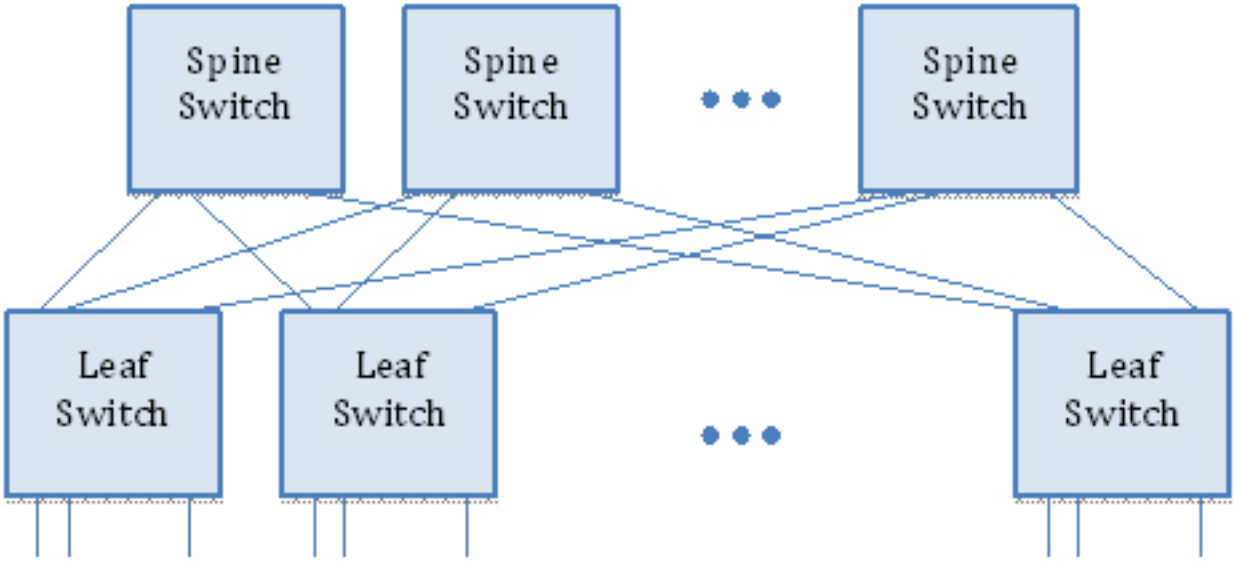
Data centers are being virtualized to improve efficiencies. This means that the network no longer connects only hardware blocks, but now must interconnect virtual machines and virtual storage volumes. With technologies such as VEPA and VNTag, Ethernet bridges outside the servers will orchestrate the movement of all traffic between virtual machines. The ultimate goal is to have a sea of resources such as VMs that can be easily repurposed or moved anywhere in the data center without impacting client services.

With traditional layered data center networks, moving a VM may not only effect the access switch configuration, it may require reconfiguration of the aggregation and core switches as well. From a latency point of view, a three-tier data center network is divided into silos, where maintaining low latency between VMs requires that they be moved only within the domain of an access switch, greatly limiting data center

flexibility. In addition, if low latency clustering services are requested by a client, the network administrator must find an access switch that has enough available servers to meet the client's requirements. In the flat data center, low latency can be maintained across all of the servers, providing very simple deployment scenarios.

Flat Network Scaling

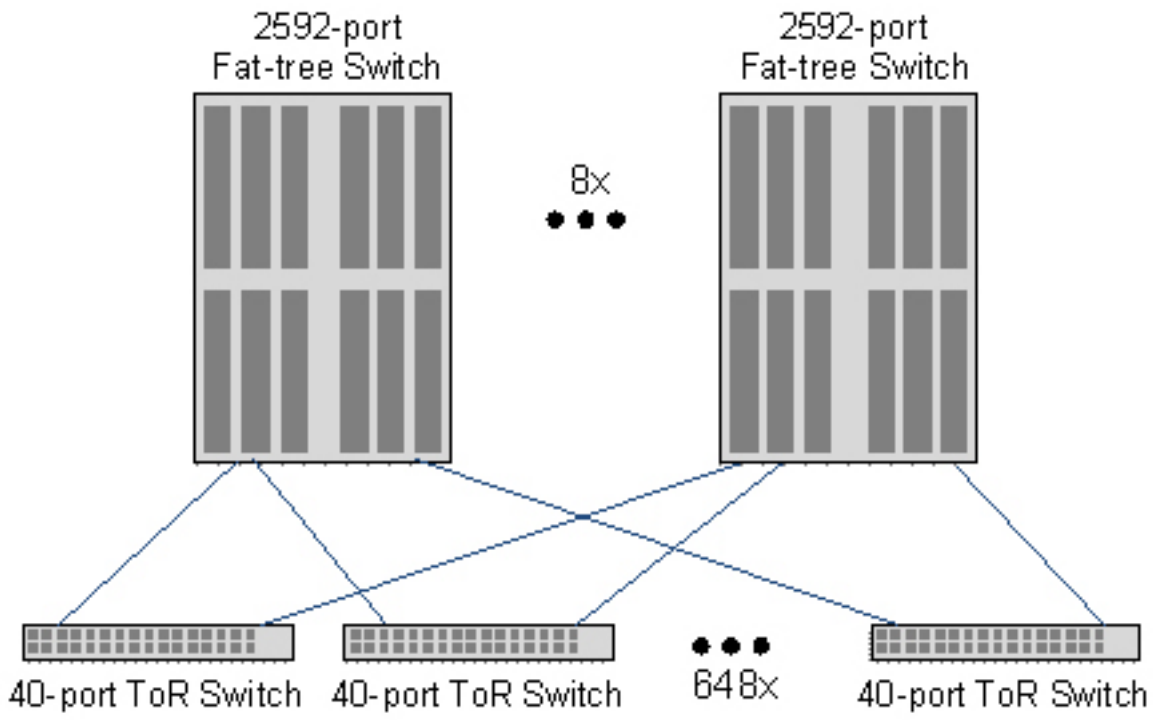
In order to solve the problems described above, the data center network equipment must easily scale to support tens of thousands of servers and up to hundreds of thousands of VMs while maintaining low latency. One way to do this is to merge the aggregation and core layers into a single large switch using a fat tree architecture as shown in Figure 1.



To maintain cross-sectional bandwidth, the leaf switches connect half of their ports to the spine switches and the other half as external switch ports. So for a switch chip containing N ports, the total two-stage non-blocking fabric will have up to N2/2 ports. In order for this to work, the leaf switches must uniformly distribute their traffic load across the spine switches. This can be done using advanced load distribution techniques.

Flat Network Design Example

The FocalPoint series of switches from Fulcrum Microsystems provides several features that allow system designers to develop large flat data center fabrics. For example, the new FM6000 Series contains up to seventy-two 10 GbE ports or 18 40 GbE ports. At 72 ports, a single chassis using the fat tree architecture shown in Figure 1 could support up to 2,592 10 G ports with less than 1 uS of port-port latency. This can replace the aggregation and core layers in the enterprise architecture, and top-of-rack (ToR) switches can serve as the access layer as shown in the figure below. Assuming a slightly over-subscribed data center network using 40-port 10GbE top-of-rack switches that also provide 320 G of uplink bandwidth, eight of these large chassis could connect up to 648 ToR switches providing more than 25,000 10G network connections with, at most, five switch hops. The latency in this flat architecture will be almost two orders of magnitude better than an enterprise-oriented network.



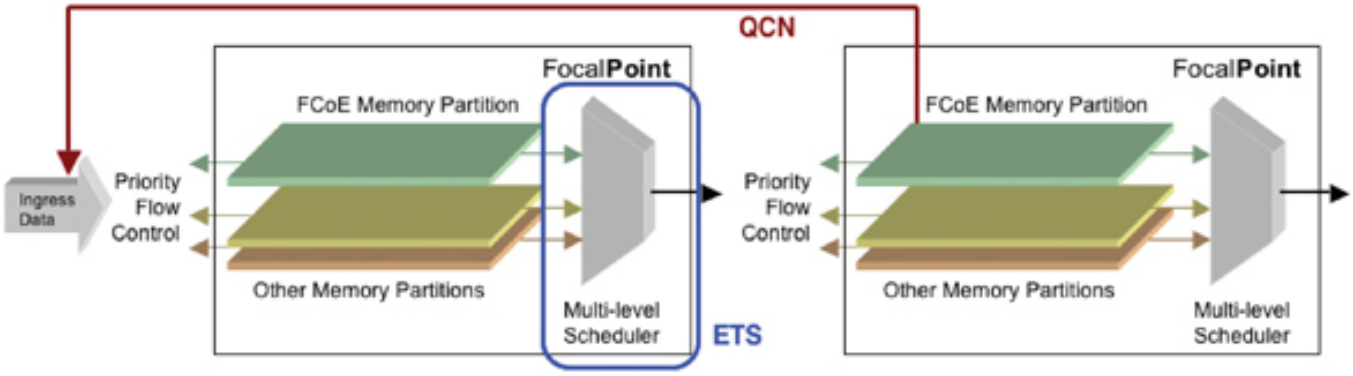
The FM6000 also provides several advanced load distribution mechanisms using 16 programmable hash key profiles. Up to three 16-bit hash keys can be generated from a wide variety of L2/L3 header fields. This can provide very uniform traffic load distribution over the second stage of a fat tree, or between the ToR switch and core.

Support for Server Virtualization
The switch uses 16-bit identifiers called global resource tags (Glorts), which can be used to identify up to 64K virtual machines (ports) in a large flat data center fabric. These virtual port identifiers can be used for:
• Forwarding/learning/aging/mirroring by virtual port
• Ingress/egress ACLs by virtual port
• Security by virtual port
• Policing and statistics by virtual port
In addition, the FocalPoint API supports non-disruptive updates, allowing mobility of ACL rules and security policies as virtual machines are moved throughout the data center. Using these virtual port identifiers, long with a configurable frame processing pipeline called FlexPipeTM, the FM6000 series can support emerging virtualization standards such as VEPA and VNTag and adapt to any future changes in these standards.

Support for Convergence
As the size and density of datacenters increase, the cost, area, power and support of multiple switch fabric installations cannot be tolerated. Because of this, the IEEE is developing several new standards that will enable Ethernet as the single unified fabric for data, storage and HPC traffic. The industry has coined these new initiatives Datacenter Bridging (DCB).

Figure 3 shows an overview of several features that are used to support DCB in the

FM6000 series. Priority Flow Control (PFC) is used as a link level flow control for lossless operation using multiple logical memory partitions within an output queued shared memory architecture. Memory partitions can be independently flow controlled to protect traffic such as storage from frame loss. Traffic can also be classified and assigned to one of 12 CoS queues at each egress port. Enhanced Transmission Selection (ETS) provides minimum bandwidth guarantees for special traffic classes such as FCoE using a DRR scheduler. Quantized Congestion Notification (QCN) is a congestion feedback mechanism used to reduce fabric congestion between stages in a large flat network architecture. These features allow lossless operation with bounded latency for iSCSI, FCoE or HPC traffic.



Traditional data center fabrics have evolved from the enterprise, which typically use three layers of fabric hierarchy and compartmentalize fabric performance. The new data centers are large virtualized environments that desire large flat converged switch fabrics for ease of configuration, lower cost, low latency and uniform performance. One network switch family allows data center network designers to achieve these goals by proving advanced features such as efficient scalability, server virtualization support and mechanisms for fabric convergence.

Source URL (retrieved on *09/21/2014 - 12:35am*):
http://www.ecnmag.com/articles/2011/03/designing-flat-data-center-network?qt-recent_content=0