Towards an Open Data Center with an Interoperable Network: Enterprise Networking using Open Industry Standards

Casimer DeCusatis, IBM Corporation, 2455 South Road, Poughkeepsie, NY 12603

Abstract

Recently there has been an increased focus on transforming data center networks to meet the requirements of next generation, highly virtualized data centers. This paper describes a network architecture based on open industry standards which addresses many of the concerns facing traditional Ethernet, storage, and wide area networks. Various aspects of this architecture will be discussed, including layer 3 equal cost multipathing, software-defined networking and OpenFlow, lossless Ethernet, and extended distance connectivity with ultra-low latency networks. Examples will be given based on multiple vendor products which have demonstrated interoperability based on this approach.

1. Introduction

Modern data center networks face an unprecedented array of challenges, including cost effective scaling, support for new applications such as highly virtualized data centers, more reliable deliver of Ethernet data frames, and much more. In order to contain both capital and operating expense, this network transformation should be based on open industry standards as opposed to proprietary, single-vendor solutions [1]. A recently proposed network reference architecture describes best practices for designing an open data center with an interoperable network (ODIN). The ODIN reference architecture describes best practices for creating a flat, converged, virtualized data center network (or fabric) based on open industry standards, and provides a description of networking best practices [2]. In this paper, we will examine the issues which ODIN addresses in next generation data center networks.

In recent years there have been many fundamental and profound changes in the architecture of modern data centers, which host the computational power, storage, networking, and applications that form the basis of any modern business [3,4]. The traditional data center architecture and compute model, especially in the case of rack or blade servers using an x86 based processor, is shown in figure 1. Historically, Ethernet was first used to interconnect “stations” (dumb terminals) through repeaters and hubs; eventually this evolved into switched topologies for a campus network, which came to form the basis for traditional data center networks. Conventional Ethernet data center networks are characterized by access, aggregation, services, and core layers, which could have 3, 4, or more tiers. Data traffic flows from the bottom tier up through successive tiers as required, and then back down to the bottom tier, providing connectivity between servers. To reduce cost and promote scaling, oversubscription is typically used for all tiers of the network. Layer 2 and 3 functions are separated within the access layer of the network. Serviced dedicated to each application (firewalls, load balancers, etc.) are placed in vertical silos dedicated to a group of application servers. Finally the network

management is centered in the switch operating system; over time, this has come to include a wide range of complex and often vendor proprietary features and functions.

There are many problems with applying conventional networks to modern data center designs. Conventional data centers have consisted of lightly utilized servers running a bare metal operating system or a hypervisor with a small number of virtual machines (VMs). The servers may be running a mix of different operating systems, including Windows, Linux, and UNIX. The network consists of many tiers, where each layer duplicates many of the IP/Ethernet packet analysis and forwarding functions. This adds cumulative end-to-end latency (each network tier can contribute anywhere from 2 – 25 microseconds) and requires significant amounts of processing and memory. Oversubscription, in an effort to reduce latency and promote cost-effective scaling, can lead to lost data and is not suitable for storage traffic. Multiple networks are provided for both Ethernet and Fibre Channel (and to a lesser degree, not shown here, for other specialized applications such as server clustering or other protocols such as InfinIBand). Each of these networks may require its own dedicated management tools, in addition to server, storage, and appliance management. Servers typically attach to the data center network using lower bandwidth links, such as 1 Gbit/s Ethernet and either 2, 4, 8, or 16 Gbit/s Fibre Channel storage area networks (SANs).

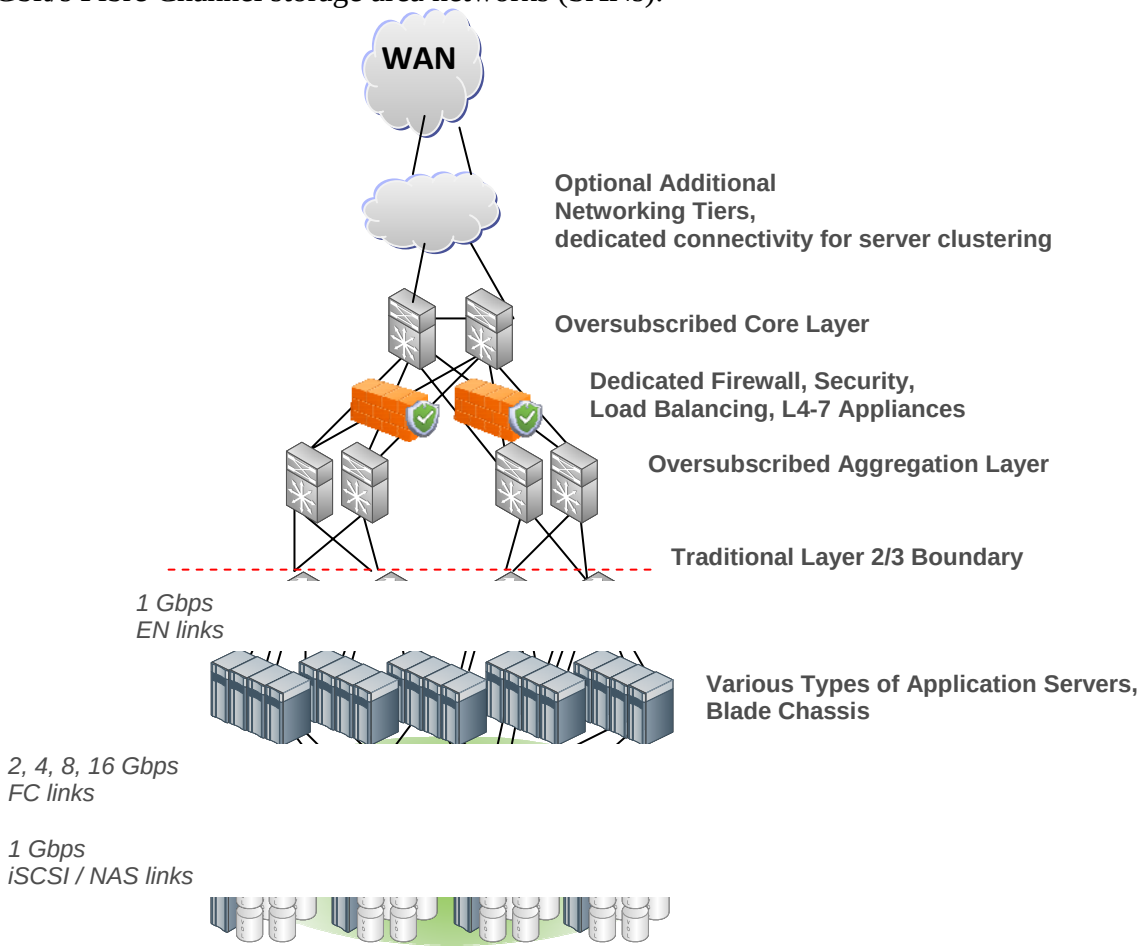

Figure 1 – Design of a conventional multi-tier data center network

Conventional networks do not scale in a cost effective or performance effective manner. Scaling requires adding more tiers to the network, more physical switches, and more physical service appliances. The physical network must be rewired to handle changes in the application workloads, and the need to manually configure features such as security access makes these processes prone to operator error. Further, there are unique problems associated with network virtualization (significantly more servers can be dynamically created, modified, or destroyed, which is difficult to manage with existing tools). Conventional networks do not easily provide for VM migration (which would promote high availability and better server utilization), nor do they provide for cloud computing applications such as multi-tenancy within the data center. Installation and maintenance of this physical compute model requires both high capital and operating expense. As a result, the management tasks have been focused on maintaining the infrastructure and not on enhancing the services that are provided by the infrastructure to add business value.

Modern data centers are undergoing a major transition toward a more dynamic infrastructure. This allows for the construction of flexible IT capability that enables the optimal use of information to support business initiatives. For example, a dynamic infrastructure would consist of highly utilized servers running many VMs per server, using high bandwidth links to communicate with virtual storage and virtual networks both within and across multiple data centers. As part of the dynamic infrastructure, the role of the data center network is also changing in many important ways, causing many clients to re-evaluate their current networking infrastructure. Many new industry standards have emerged, and are being implemented industry wide. The accelerating pace of innovation in this area has also led to many new proposals for next generation networks to be implemented within the next few years.

There are many factors to consider when modernizing the design of a data center network. For example, an ever increasing amount of data center traffic is between servers (so-called east-west traffic), as opposed to between clients and servers or between adjacent server pods (so-called north-south traffic). It has been estimated that as much as 75% of the traffic in cloud computing environments and highly virtualized data centers moves between servers. The modern data center network facilitates server-to-server communication by using new technologies to reduce the network from the three or four tier design that has become the industry norm to flatter, two tier designs. Modern data center networks will also take increasing advantage of virtualization at various points in the network. Finally, the industry has begun incrementally moving toward the convergence of fabrics which used to be treated separately. Each of these approaches is a non-trivial extension of the existing data center network; collectively, they present a daunting array of complex network infrastructure changes, with far reaching implications for data center design. The ODIN reference architecture should enable users to treat data center computing, storage, services, and network resources as fully fungible pools that can be dynamically and rapidly partitioned. Another key capability of this new compute model involves providing a family of integrated offerings, from platforms which offer server, storage, and networking resources combined into a simple, turnkey solution to network and virtualization building blocks that scale to unprecedented levels to enable future cloud computing systems.

The following sections will describe key elements of the current ODIN reference architecture, including Layer 3 equal cost multi-path networks, software-defined networking and OpenFlow, lossless Ethernet, and WAN connections with ultra-low latency.

2. Layer 3 Spine-Leaf Designs with ECMP

In this section, we will describe the basic approach to a Layer 3 "Fat Tree" design (or CLOS network) using Equal Cost Multi-Pathing (ECMP). As shown in figure 2, a Layer 3 ECMP design creates multiple load balanced paths between nodes in a network. Bandwidth can be adjusted by adding or removing paths up to the maximum allowed number of links. Unlike a Layer 2 network, no links are blocked with this approach. Broadcast loops are avoided by using different VLANs, and the network can route around link failures. In a preferred design, all attached servers are dual homed (each server has two connections to the first network switch using active-active NIC teaming). This approach is known as a spine and leaf architecture, where the switches closest to the server are "leaf" switches which interconnect with a set of "spine" switches using a set of load balanced paths. This example illustrates a 4 way ECMP with 16 IP subnets per rack and 64 IP subnets per uplink, for a total of 80 IP subnets. Using a two tier design with a reasonably sized (48 port) leaf and spine switch such as the IBM G8264 and relatively low oversubscription (3:1), it is possible to scale this network up to around 1,000 – 2,000 ports of 10G traffic. Note that the design does not require a larger form factor core switch, although we could use core switches to replace the spine switches in this example.

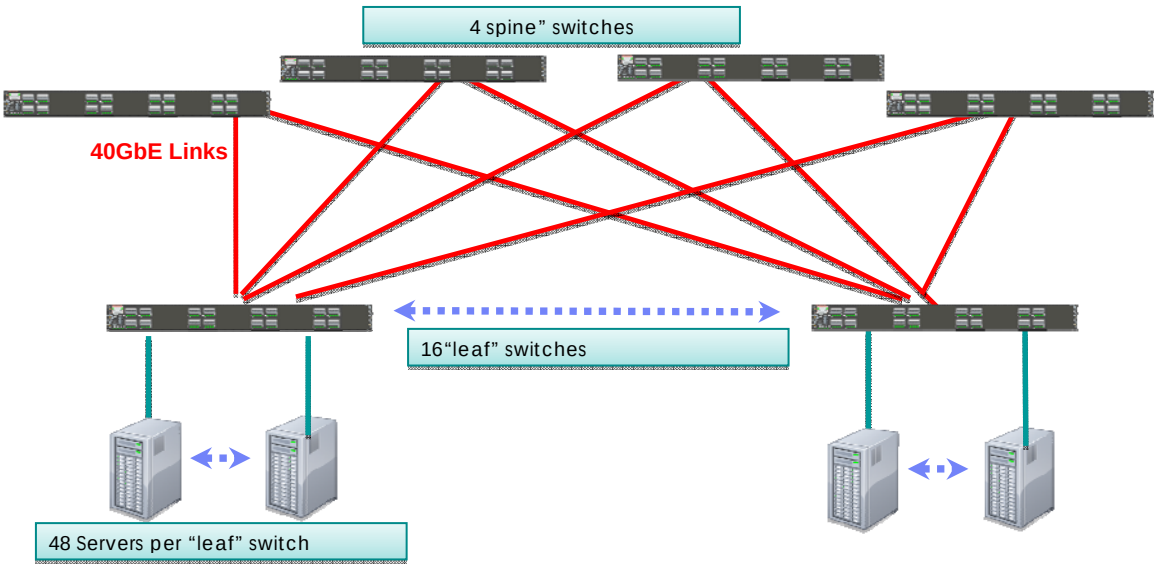

Figure 2 – Example Layer 3 ECMP leaf-spine design

A Layer 3 ECMP design can be enhanced by using Virtual Link Aggregation Groups (VLAGs), as shown in figure 3. If devices attached to the network support Link Aggregation Control Protocol (LACP) it becomes possible to logically aggregate multiple

connections to the same device under a common vLAG ID. It is also possible to use vLAG inter-switch links (ISLs) combined with VRRP protocols to interconnect switches at the same tier of the network. VRRP supports IP Forwarding between subnets, and protocols such as OSPF or BGP can be used to route around link failures. Server pods can be constructed as shown in this example, and VMs can be migrated to any server within the pod.

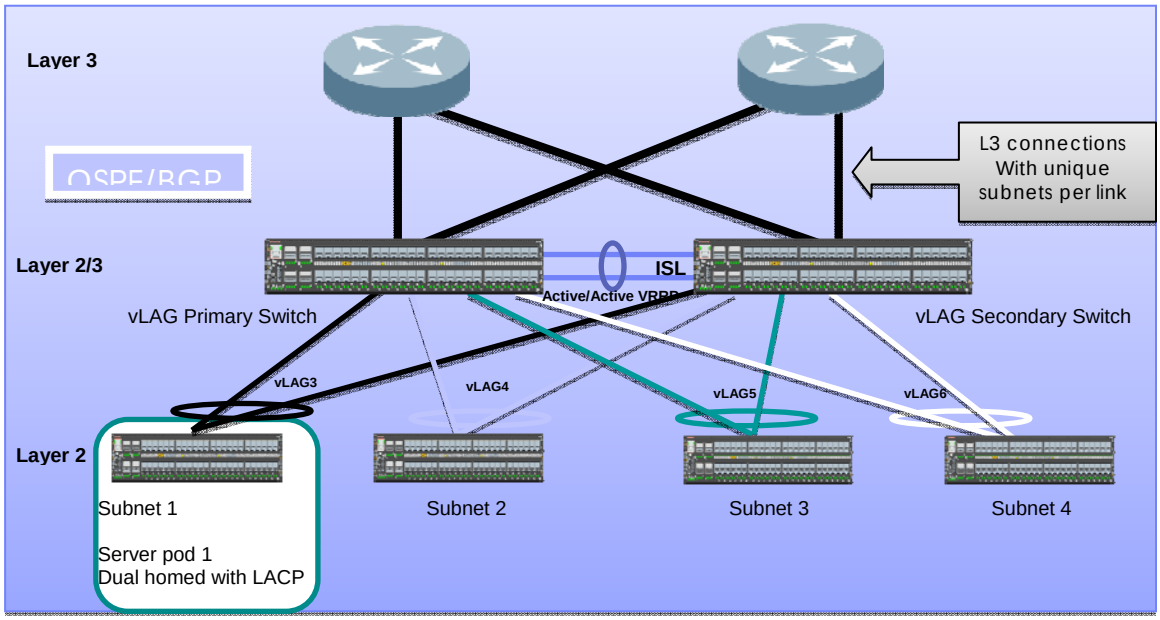

Figure 3 – Spine-Leaf ECMP design example

Layer 3 ECMP designs offer several advantages. They are based on proven, standardized technology which leverages smaller, less expensive rack or blade switches (virtual switches typically do not provide Layer 3 functions and would not participate in an ECMP network). The control plane is distributed, and smaller fault domains are possible using the pod design approach. These networks scale well (up to 1-2 thousand ports with a slightly oversubscribed 2 tier topology, higher with more tiers).

There are also some tradeoffs when using a Layer 3 ECMP design. The native Layer 2 domains are relatively small, which limits the ability to perform live VM migrations from any server to any other server. Such designs can be fairly complex, and may present complications with multicast domains. In the examples shown earlier, scaling is limited by the control plane, which can become unstable in some conditions (for example, if all the servers attached to a leaf switch boot up at once, the switch's ability to process ARP and DHCP relay requests will be a bottleneck in overall performance). In a Layer 3 design, the size of the ARP table supported by the switches can become a limiting factor in scaling the design, even if the MAC address tables are quite large. Finally, complications may result from the use of different hashing algorithms on the spine and leaf switches.

3. Software-Defined Networking and OpenFlow

The Open Networking Foundation (ONF) is driving the standardization of OpenFlow as part of its mission to create software-defined networking (SDN) standards. The ONF is led by a board of directors consisting of companies that own and operate some of the largest networks in the world (including Deutsche Telekom, Facebook, Google, Microsoft, Verizon, Yahoo, Goldman Sachs, and NTT). Many of the industry's leading networking companies are members of the ONF, including IBM and all of the ODIN participating companies.

SDN is a broad concept, including both OpenFlow and network overlays. SDN is used to simplify network control and management, automate network virtualization services, and provide a platform from which to build agile network services. In particular, OpenFlow provides a separation of the data and control planes in network switches. There are many benefits of a standard which opens the control plane of the switch network, and a flow paradigm that offers granular traffic control. OpenFlow also offers a global view of the network, including traffic statistics, and is fully compatible with existing Layer 2 and 3 protocols. In contrast to a traditional switch, OpenFlow allows direct access and manipulation of the forwarding or data plane of network switches and routers, both physical and virtual (hypervisor-based). In this environment, networking services (security, multi-pathing, and more) run like apps on a software-defined network stack. The use of OpenFlow to enable an ecosystem of network apps development, as opposed to the closed, vendor proprietary approach used today, represents an important change in the way networks services will be deployed in the future.

An OpenFlow switch consists of three parts, as illustrated in figure 4. First, a Flow Table tells the switch how to process each data flow by associating an action with each flow table entry. Second, a Secure Channel connects the switch to a remote control processor (called the Controller) so commands and packets can be sent between the controller and the switch. Finally, the OpenFlow Protocol provides an open, standardized interface for the controller to communicate with the switch and to remove, add, or change flow control entries

The OpenFlow Protocol allows entries in the Flow Table to be defined by a server external to the switch. For example, a flow could be a TCP connection, all the packets from a particular MAC or IP address, or all packets with the same VLAN tag. Each flow table entry has a specific action associated with a particular flow, such as forwarding the flow to a given switch port (at line rate), encapsulating and forwarding the flow to a controller for processing, or dropping a flow's packets (for example, to help prevent denial of service attacks). There are many applications for OpenFlow in modern networks. For example, a network administrator could create on-demand 'express lanes' for voice and data traffic that are time-sensitive. Software could also be used to combine several fiber optic links into a larger virtual pipe to handle a particularly heavy flow of traffic temporarily. When the data rush is over, the links would automatically separate under the supervision of the controller. In cloud computing environments, OpenFlow

improves scalability and enables resources to be shared efficiently among different services in response to the number of users.

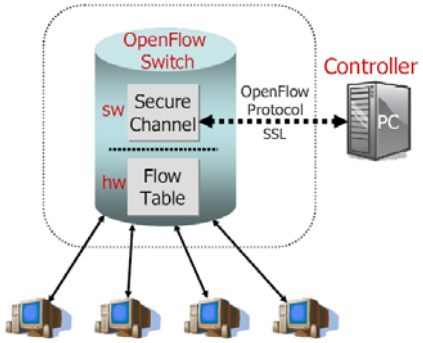

Figure 4 – Basic OpenFlow architecture

The IBM G8264 has been demonstrated in an OpenFlow configuration using the NEC pFlow controller and switch; the high level system architecture is shown in figure 5. This configuration was demonstrated at InteOp 2011 and at the Open Network Summit 2012 [5], where it was used to demonstrate functions such as control of the bisection bandwidth and oversubscription ratio for the attached OpenFlow switches; the use of intelligent flow-based multipathing for active load balancing; and high availability with fast failover in case or a port or switch failure. The NEC controller demonstrated the ability to create multiple virtual networks on a single multi-vendor switch infrastructure and create intelligent flow filtering, routing; and monitoring. This configuration was used to demonstrate several practical use cases for OpenFlow, including how to extend a single management domain across a physical and virtual switch infrastructure composed of devices from a variety of vendors. This topology was also used to define network-based quality of service (QoS) for voice over IP (VoIP) applications.

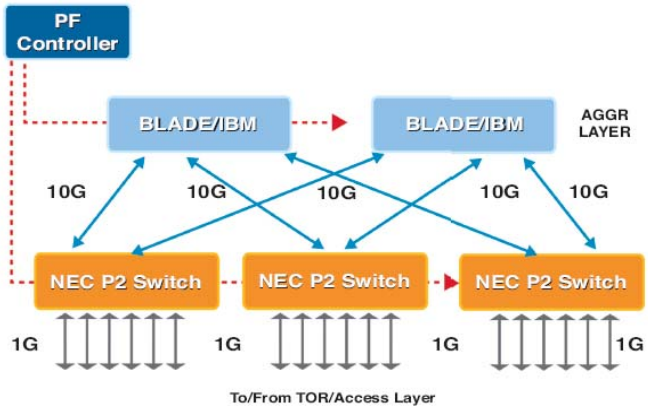

Figure 5 – High level reference architecture for OpenFlow networking

4. Lossless Ethernet Protocol

Traditional Ethernet is a lossy protocol; that is, data frames can be dropped or delivered out of order during normal operation. In an effort to improve the performance of Ethernet, the IEEE has developed a new standard under the 802.1 work group to create

"lossless" Ethernet. This new form of Ethernet can be deployed along with conventional Ethernet networks, or as the foundation for advanced features including RDMA over Converged Ethernet (RoCE) and Fibre Channel over Ethernet (FCoE). Historically, this work grew out of a series of proposals made by a consortium of networking companies called the Converged Enhanced Ethernet (CEE) Author's Group; for this reason, the resulting standard is also known as CEE. However, the formal name for this standard as defined by the IEEE is Data Center Bridging (DCB). There are three key components which are required to implement lossless Ethernet, namely Priority-based Flow Control (PFC), Enhanced Transmission Selection (ETS) , and Data Center Bridging Exchange protocol (DCBx). A fourth, optional component of the standard is congestion notification (QCN).

PFC is defined by IEEE 802.1Qbb; it creates eight different traffic priority levels, based on a field added to the frame tags called the priority code point field. This enables the control of data flows on a shared link, with granularity on a per-frame basis. Traffic which can be processed on a lossy link may be handled appropriately while it shares a link with lossless traffic (such as FCoE).

ETS is defined by IEEE 802.1Qaz; it allows data to be organized into groups, or traffic classes, each of which is assigned a group identification number. This makes it possible to allocate different fractions of the available link bandwidth to different traffic classes, (bandwidth allocation). Traffic from different groups can be provisioned for its desired data rate (for example, FCoE traffic may operate at 8 Gbit/s while Ethernet may operate at 1 Gbit/s). This feature enables quality of service to be implemented based on the application requirements. It also prevents any one traffic flow from setting all of its frames to the highest traffic priority level and consuming the full bandwidth of the link.

DCBx is defined by IEEE 802.1Qaz; it is a protocol which discovers resources connected to the network, initializes the connection between these resources and the rest of the network, and thus establishes the scope of the network that supports lossless Ethernet. The local configuration for lossless Ethernet switches (including PFC, ETS, and relevant application parameters which tell the end station which priority to use for a given application type) is distributed to other compatible switches in the network using DCBx. The DCBx protocol also detects configuration errors between peer switches. To accomplish this, DCBx makes use of the capabilities defined in IEEE 802.1AB (link layer discovery protocol).

An optional (but desirable) component of lossless Ethernet is Queue Congestion Notification (QCN), defined by IEEE 802.1Qau. QCN is an end-to-end congestion management protocol which detects congestion in the fabric and throttles network traffic at the edge of the fabric.

IBM has implemented lossless Ethernet in all of its System Networking products, including rack switches (G8264), blade chassis switches, and virtual switches (5000v). Most other companies in the industry, including other ODIN participants, have either announced support for this interface or are expected to do so in the future. Using this

protocol, IBM supports FCoE transport and FCoE multi-hop between a blade chassis and rack switch. Further, IBM System Networking products and PureSystems solutions can attach to other vendor's lossless Ethernet networks. For example, Juniper's Qfabric products are PureSystems Ready [6]; Qfabric can also be used as an FCoE gateway or transit switch when attached to IBM System Networking solutions. Of course, this does not mean that Fibre Channel storage area networks (SANs) will be replaced anytime soon.

As an encapsulation protocol, FCoE will perform its functions with some performance overhead above that of the native FC protocol that it encapsulates. Data centers with genuine need for high-performing 8 G FC will question the benefits of sharing a 10 GbE link with LAN traffic. For this reason, it is expected that FCoE will most likely be deployed first in environments currently using 4 Gbps FC and 1 GbE links, and that deployment on 10G links will come afterwards.

5. Multi-Site Connectivity

Wide area networks (WANs) are used to interconnect multiple data centers for business continuity and backup/recovery applications, and are thus an important part of the overall data center strategy. It is common for the volume of WAN traffic to increase at an annual rate of 30% of more, and this traffic volume is expected to increase even further with the advent of larger cloud data centers and multi-site enterprise disaster recovery solutions. In the past, data centers didn't extend broadcast domains over long distance. Filtering was required for traffic intended to go outside a given broadcast domain. In a more modern environment, there may be tens to hundreds or thousands of virtual servers on a single domain; if this is extended over distance, it would require a huge amount of WAN bandwidth (otherwise, it might take a very long time to move a VM and its associated data). Higher data rates on the WAN and service provider network would also drive disproportionately higher data rates on switches within the data center and at the WAN edge, which does not lend itself to cost effective scaling.

Multi-site connectivity can be implemented in a number of ways. Public Internet connections with IPSec secure tunneling are readily available and low cost, but do not provide the quality of service and performance guarantees required for many larger enterprises. Managed data connectivity services provide additional layers of security and performance running over a public or private Internet connection. Leased line data services are available from service providers which include options for private management of point-to-point networks (known as private circuits or Layer 2 VPN) or full mesh connectivity (Layer 3 VPN). In areas where leased optical fiber (or "dark fiber") is available, it is often cost effective for larger enterprises to use dedicated optical wavelength division multiplexing (WDM) solutions. The cost of WDM is falling rapidly, and it is becoming available as an integrated option on some large Ethernet switches.

A backbone using industry standard multi-protocol label switching (MPLS) for site to site connectivity is compatible with a dual homed Ethernet architecture in the data center. For storage applications, dark fiber WDM solutions are preferred, and may be supplemented with Fibre Channel over IP (FC-IP) solutions. For example, IBM has

tested and qualified multi-site data center solutions using WDM equipment from ODIN participants Adva and Huawei, as well as other companies. IBM has also recently announced extended distance versions of their storage volume controller (SVC) technology using WDM or FC-IP technology from Brocade, as shown in figure 6 [7]. The use of WDM as a channel extension option has also been demonstrated [8].

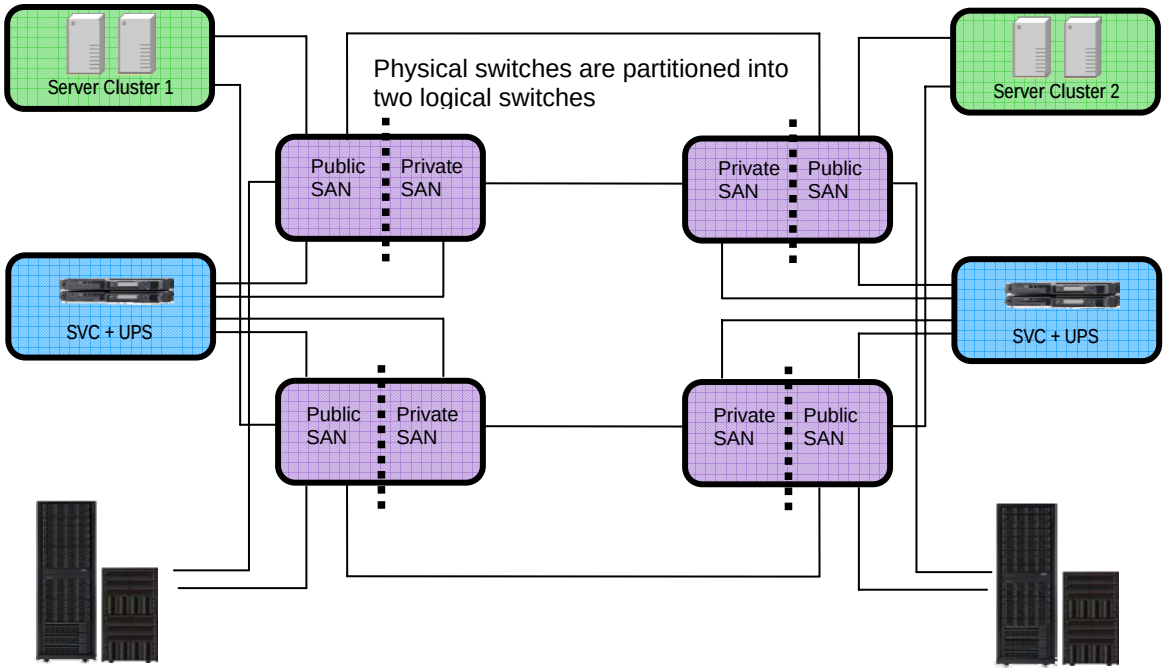

Figure 6 – SVC stretch clusters using either FC-IP or WDM

Routers and firewalls should be deployed in an active/active configuration and use separate WAN links, cross-connected to provide high availability. Load balancing across redundant connections is optional depending on traffic volumes and availability requirements of the application. Other emerging protocols including IPv6 and OpenFlow are beginning to make inroads into the WAN, as well. One application which has recently received considerable interest is the design of data centers to accommodate extremely low latency applications [9].

6. Conclusions

By implementing the recommendations of the ODIN reference architecture, traditional data center networks can evolve to the next generation design shown in figure 7. These networks are characterized by a flattened, 2 tier design (with embedded Blade switches and virtual switches within the servers), to provide lower latency and better performance. This design can scale without massive oversubscription and lower total cost of ownership to thousands of physical ports (using Layer 3 ECMP) and potentially tens of thousands of VMs. As a recent example, the IBM PureSystems solutions announced in April 2012 provide up to 896 processor cores, 43 TB memory, and 480 TB storage, with each compute or storage node served by 80 GB of network bandwidth. This facilitates over 26

million IO operations per second, per rack (the system scales up to 4 racks using current technology). Further, this traffic is optimized for east-west transport between servers; over 75% of data traffic flows east-west in this design. By limiting the traffic flow to TOR switches, overall latency is reduced to half the value of previous implementations. By integrating networking with servers and storage, PureSystems reduces the total number of networking devices, promoting high availability and low energy costs as well as simplifying cabling within and between racks. The use of virtual overlay networks mean that the network cables can be wired once and dynamically reconfigured through software-defined networking, which also enables pools of service appliances shared across multi-tenant environments. The network used in PureSystems employs the G8264 rack switch, which is OpenFlow compliant as discussed previously. Large Layer 2 domains enable VM mobility across different physical servers. The network state resides in the virtual switches (IBM 5000v), which are enabled for automated configuration and migration of port profiles (VMs can be moved either through the hypervisor vSwitch or through an external network switch). VM migration is also facilitated by the support of IEEE 802.1Qbg standards across all networking within PureSystems. Management functions are centralized, requiring fewer management instances with less manual intervention and more automation (with less opportunity for operator error). While PureSystems initial deployment relies on proved Fibre Channel storage, the design is enabled to support FCoE as an option (this approach can also be leveraged on network architectures other than PureSystems). Desirable features such as disjoint fabric paths and multi-hop communications enable the fabric as an on-ramp for cloud computing.

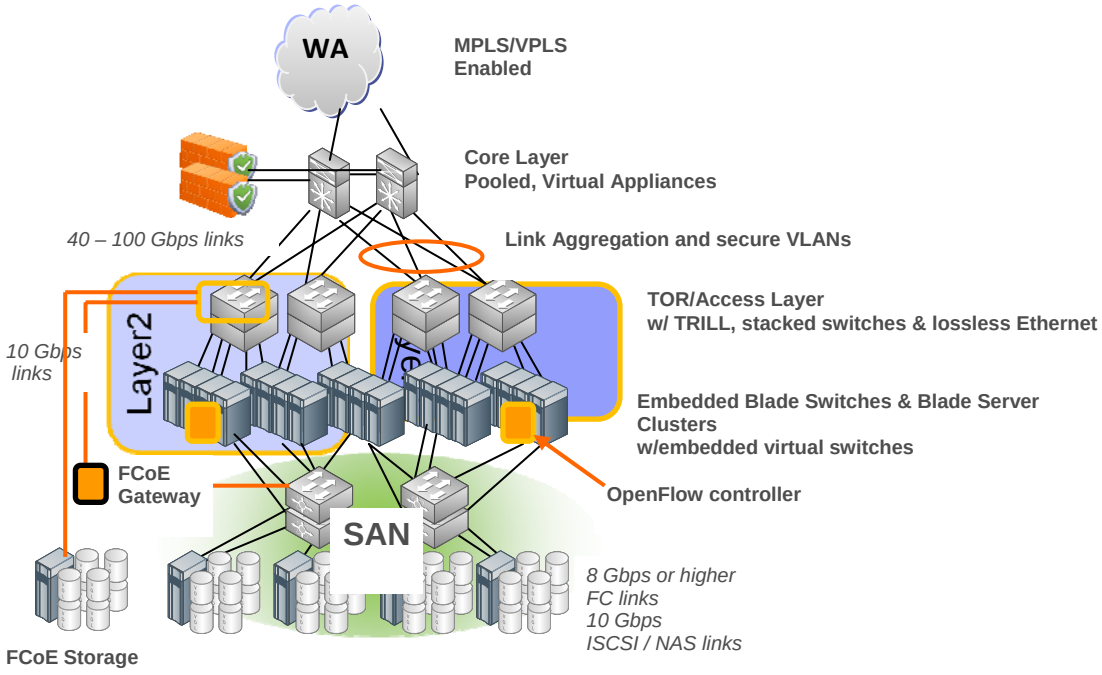

Figure 7 An Open Data Center with an Interoperable Network

The most cost effective approach to redesigning a next generation data center network involves the use of open industry standards and interoperable, multi-vendor networks. In addition to enhanced versions of the standards described here, many new standards for

data center networking are under development by the IEEE, IETF, ONF, and other bodies. The ODIN reference architecture will continue to evolve in response to these new requirements, including recommendations on the best ways to combine different standards together into a common network fabric.

References

1) Gartner Group report, "Debunking the myth of the single-vendor network", 17 Nov. 2010 http://www.dell.com/downloads/global/products/pwcnt/en/Gartner-Debunking-the-Myth-of-the-Single-Vendor-Network-20101117-published.pdf

2) C. DeCusatis, "Towards an open data center with an interoperable network (ODIN): volumes 1-5; published at InterOp 2012, May 8, Las Vegas, NV http://www-03.ibm.com/systems/networking/solutions/odin.html ; see also IBM Data Networking blog https://www-304.ibm.com/connections/blogs/roller-ui/allblogs?userid=2700058MPY&lang=en_us and Twitter feed @IBMCasimer

3) T. Benson et.al., "Network traffic characteristics of data centers in the wild", Proc. IMC conference, published by the ACM (2010) http://pages.cs.wisc.edu/~tbenson/papers/imc192.pdf

4) C. DeCusatis, "Optical networking in smarter data centers: 2015 and beyond", invited paper, 2012 OFC/NFOEC Annual Meeting, paper OTu1G.7, Los Angeles, CA (March 4-8, 2012) http://www.ofcnfoec.org/Mobile/Home/Conference-Program/Invited-Speakers.aspx

5) IBM White Paper QCW03010-USEN-00, "OpenFlow: the next generation in network interoperability", InterOp conference (May 2011) http://www-03.ibm.com/systems/networking/solutions/odin.html

6) L. King, "IBM PureSystems: accelerating cloud leveraging Juniper networks Qfabric and vGW solutions", http://forums.juniper.net/t5/Architecting-the-Network/IBM-PureSystems-Accelerating-Cloud-leveraging-Juniper-Networks/ba-p/137625

7) B. Larson and C. DeCusatis, "SVC Stretch Clusters", Edge 2012, Orlando, FL (June 2012)

8) T. Bundy, M. Haley, F. Street, C. DeCusatis, "The impact of data center convergence, virtualization, and cloud on DWDM optical networks both today and into the future", Proc. Pacific Telecommunications Council 2012 Annual Meeting, Honolulu, Hawaii, (January 2012)

9) C., DeCusatis, "Enterprise networks for low latency, high frequency financial trading", Proc. Enterprise Computing Community conference, June 12-14, 2011, Marist College, Poughkeepsie, NY (2011) http://ecc.marist.edu/conf2011/