Enterasys Design Center Networking – Connectivity and Topology Design Guide

Demand for application availability has changed how applications are hosted in today's datacenter. Evolutionary changes have occurred throughout the various elements of the data center, starting with server and storage virtualization and also network virtualization.

Motivations for server virtualization were initially associated with massive cost reduction and redundancy but have now evolved to focus on greater scalability and agility within the data center. Data center focused LAN technologies have taken a similar path; with a goal of redundancy and then to create a more scalable fabric within and between data centers.

Business requires next generation networks to change focus from redundancy to resiliency.

While it may seem that redundancy and resiliency are one and the same, they are not. Redundancy simply requires duplication of systems. Resiliency is the ability of the solution to "adapt" to the consequences of failure. Today's data center must meet a number of business requirements and overcome several design obstacles in order to truly achieve resiliency.

Business requirements

- Improve application performance
- Regulatory compliance
- Business (IT) agility

Design obstacles

- Density increases with a rapid pace
    - On an ongoing basis new applications are deployed on new server systems
    - Increases in server performance results in a large number of virtual machines per server
    - Increases in the number of virtual machines per server increases the traffic per server
- Dynamic application provisioning and resource allocation

Resiliency is not achieved by simply implementing new technologies. It also requires investment in architectures and tools along with a ready workforce that can operate these networks without requiring extensive vendor-specific training.

This paper will provide the reader with key concepts for designing a standards-based data center fabric to meet the requirements of today and tomorrow.

Table of Contents

Data Center Network Design Goals

Derived from the business objectives and the requirements of the applications hosted today in the data center the common design goals include:

- Performance
- Scalability and agility
- Flexibility to support various services
- Security
- Redundancy/High availability
- Manageability
- Lower OPEX and CAPEX
- Long term viability

There is no single solution that can be applied to all. What we will propose is a set of guidelines from which a solution can be designed which will meet the unique needs and goals of the organization. Additionally, the design architecture will emphasize criteria which are standard-based without compromising critical functionality.

Data center LANs are constantly evolving. Business pressures are forcing IT organizations to adopt new application delivery models. Edge computing models are transitioning from applications at the edge to virtualized desktops in the data center. The evolution of the data center from centralized servers to a private cloud is well underway and will be augmented by hybrid and public cloud computing services.

With data center traffic becoming less client-server and more server-server centric, new data center topologies are emerging.

With data center traffic becoming less client-server and more server-server centric, new data center topologies are emerging. Yesterday's heavily segmented data center is becoming less physically segmented and more virtually segmented. Virtual segmentation allows for the reduction of physical equipment, leading to both capital and operational expense (CAPEX/OPEX) savings.

New Enterasys connectivity solutions provide the ability to compress the traditional 3-tier network into a physical 2-tier network by virtualizing the routing and switching functions into a single tier. Virtualized routing provides for greater resiliency and fewer switches dedicated to just connecting switches. Reducing the number of uplinks (switch hops) in the data center improves application performance as it reduces latency throughout the fabric.

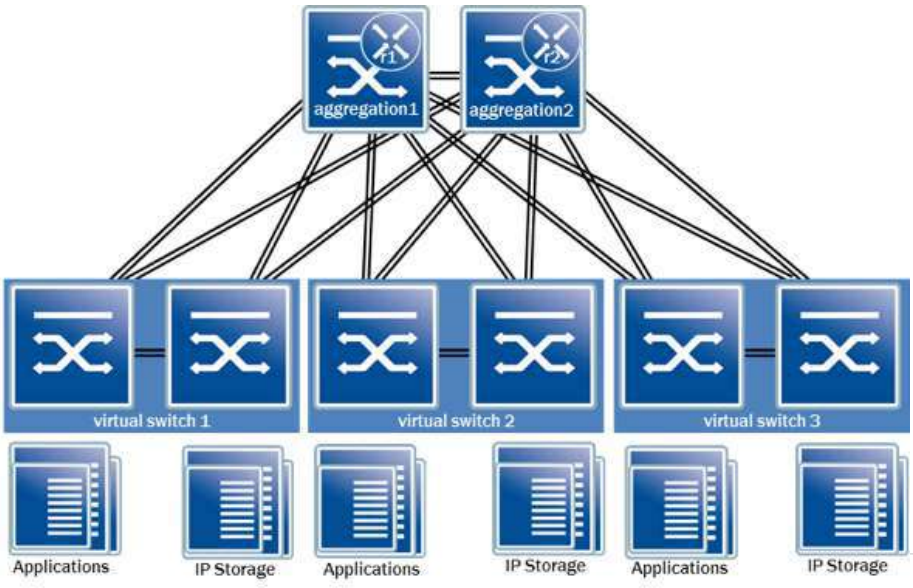


Figure 1: Two-tier Data Center Design

The design above shows virtual switches used at the data center LAN access layer providing connectivity for both applications and IP storage – iSCSI or NFS attached. The virtual switches leverage a Layer 2 meshed network for interconnectivity. The aggregation and core are merged into a single layer by virtualizing the router function in the data center LAN switch.

In addition to the transport layer, Enterasys provides an industry-leading solution for centralized command and control of the infrastructure. The Enterasys Network Management Suite (NMS) products, including Data Center Manager, simplify data center LAN management by enabling the deployment of a consistent configuration throughout the data center (both physical and virtual) and enterprise LAN. Management integration with industry-leading virtualization vendors provides multi-vendor hypervisor support that orchestrates virtual server/desktop operations with the virtual and physical networks, ultimately providing flexibility for the customer.

Data Center Connectivity Trends

10G, 40G, 100G

As the 40G/100G Ethernet standard (IEEE 802.3ba) was ratified in June 2010, the biggest market for 40G Ethernet is projected to be within the data center and for data center interconnects. Early adoption of 100G Ethernet will be used in a few bandwidth hotspots in carrier core networks and for network aggregation of 10G and 40G Ethernet links. Even with the ratification of the new Ethernet standard, there are a number of reasons 10G Ethernet is still growing and will continue to have significant growth for at least another 5 years:

- The cost of the technology is still high (as of 2013). It could take at least two more years before 100GE prices will be closer to that of 10x10GE.
- 40G/100G Ethernet is still new and it will take time until the technology is widely available. This is especially true when deploying the technology in a data center and core network, which is always planned to grow with a certain amount of multi-vendor equipment.

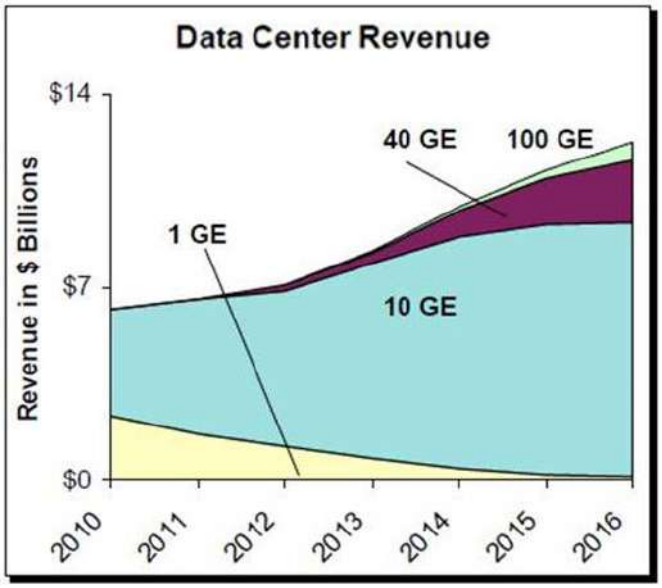


Figure 2 "source: Delloro"

The decision to implement a particular technology depends upon organizational needs, budget, and projected company business growth. However, the selected network device should at least incorporate the latest design architecture, a solid plan/road map, and enough capacity to support 40G/100GE in the future.

Storage I/O Consolidation

Within the industry, there is a stated long-term goal to establish Ethernet as the transport for a "converged" data and storage solution, thereby reducing TCO within a converged Ethernet data center. Storage connectivity today is a mix of Fibre Channel (FC), iSCSI and NFS (both iSCSI and NFS are already Ethernet and IP based). If FC is deployed, it requires two different sets of hardware, cables, tools and skill sets. Storage connectivity in the future will be based on a single converged network with an intermediate step of a single converged interface from the server to the access switch, all with new protocols and hardware. This will results in fewer adapters, cables, and nodes, resulting in more efficient network operations.

"Gartner research shows that a converged Data Center network requires more switches and ports, is more complex to manage and consumes more power and cooling than two well-designed separate networks."

The Enterasys solution is able to co-exist with FC environments, enabling the organization to continue to leverage existing investments. Enterasys provides support for Data Center Bridging (DCB) in multiple phases with different hardware and software requirements as the underlying technology to transport Fibre Channel over Ethernet (FCoE). Industry analyst firm Gartner has published a report regarding some of the myths of FCoE technology. In announcing the report ("Myth: A Single FCoE Data Center Network = Fewer Ports, Less Complexity and Lower Costs" ID Number: G00174456), Gartner notes the traditional architecture of separate storage and network systems still has merit:

As an alternative, Enterasys offers a simple, yet highly effective approach to enable, optimize and secure iSCSI SAN or NFS NAS deployments. The Enterasys S-Series modular switch is a key component of our overall solution, delivering an easy and effective way to optimize communications through automatic discovery, classification, and prioritization of SANs. In addition, the Enterasys solution will identify and automatically respond to security threats against virtual storage nodes, enforce role-based network access control policies, and comply with regulations for monitoring and auditing.

The IEEE Data Center Bridging task group, a working group of IEEE 802.1 working group, is focused on defining a new set of standards which will enable Ethernet to effectively deliver data center transport for both server and storage traffic. Terms commonly associated with DCB are "Data Center Ethernet", also known as DCE, and Convergence Enhanced Ethernet (CEE). It should be understood that DCB is the task group and term commonly being used to describe tomorrow's Data Center LANs.

Data Center Bridging is focused primarily on three (3) IEEE specifications:

- IEEE 802.1Qaz – ETS & DCBX – bandwidth allocation to major traffic classes (Priority Groups); plus DCB management protocol
- IEEE 802.1Qbb – Priority PAUSE. Selectively PAUSE traffic on link by Priority Group
- IEEE 802.1Qau – Dynamic Congestion Notification

In addition to these protocols people often include layer 2 meshing technologies when they refer to DCE or CEE.

Right now, FCoE only addresses the first five feet of connectivity in the data center, the five feet from the server to the network access switch. The transformation to a converged data and storage environment is no small challenge and will continue well into 2015 and beyond.

Main Components of the Data Center

This main focus of this paper is on data center network infrastructure design; however, we will briefly cover some of the other data center components. A data center is a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant power supplies, data communications connections, environmental controls (e.g., air conditioning, fire suppression, etc.) and security devices. For our purposes we will focus on the servers, storage and connectivity elements of the data center.

Servers

Servers deployed in the data center today are either full featured and equipped rack-mount servers or blade servers. A blade server is a stripped down server with a modular design optimized to minimize the use of physical space and energy. Whereas a standard server can function with (at least) a power cord and network cable, blade servers have many components removed to save space, minimize power consumption and other considerations, while still having all the functional components to be considered a computer. A blade enclosure, which can hold multiple blade servers, provides services such as power, cooling, networking, various interconnects and management. Together, blades and the blade enclosure form the blade system.

There are pros and cons for each server type. This discussion is not a focus of this paper.

Virtualization has introduced the ability to create dynamic data centers and with the added benefit of "green IT." Server virtualization can provide better reliability and higher availability in the event of hardware failure. Server virtualization also allows higher utilization of hardware resources while improving administration by having a single management interface for all virtual servers.

Storage

Storage requirements vary by server type. Application servers require much less storage than database servers. There are several storage options – Direct Attached Storage (DAS), Network

Attached Storage (NAS), or Storage Area Network (SAN). Applications that require large amounts of storage should be SAN attached using Fibre Channel or iSCSI. In the past, Fibre Channel offered better reliability and performance but needed highly-skilled SAN administrators. Dynamic data centers, leveraging server virtualization with Fibre Channel attached storage, will require the introduction of a new standard, Fibre Channel over Ethernet (FCoE). FCoE, requires LAN switch upgrades due to the nature of the underlying requirements, as well as Data Center Bridging Ethernet standards. FCoE is also non-routable, so it may cause issues when it comes to the implementation of disaster recovery/large geographical redundancy that L2 connectivity cannot yet achieve. On the other hand, iSCSI provides support for faster speeds and improved reliability, making it more attractive. iSCSI offers increased flexibility and a more cost effective solution by leveraging existing network components (NICs, switches, etc.). In addition, Fibre Channel switches typically cost 50% more than Ethernet switches. Overall, iSCSI is easier to manage than Fibre Channel, considering most IT personnel familiarity with the management of IP networks.

Connectivity

The networking component provides connectivity to the data center, for example, L2/L3 switches and WAN routers. As stated earlier, motivated by server virtualization, data center connectivity design is moving to network virtualization. Next, we'll take a look at some topology considerations when it comes to network connectivity in the data center.

Data Center Connectivity and Topology

Topology – Physical Designs

Two-Tier Design

A two-tier design is very popular in data center networks today. Access switches for server connectivity are collapsed in high density aggregation switches which provide the switching and routing functionality for access switching interconnections and the various server VLAN's. It has several benefits:

- Design simplicity (fewer switches and so fewer managed nodes)
- Reduced network latency (by reducing number of switch hops)
- Typically a reduced network design oversubscription ratio
- Lower aggregate power consumption

However, a disadvantage of a two-tier design includes limited scalability: when the ports on an aggregation switch pair are fully utilized, then the addition of another aggregation switch/router pair adds a high degree of complexity. The connection between aggregation switch pairs must be fully meshed with high bandwidth so no bottlenecks are introduced into the network design. Since an aggregation switch pair is also running routing protocols, more switch pairs means more routing protocol peering, more routing interfaces and complexity introduced by a full mesh design.

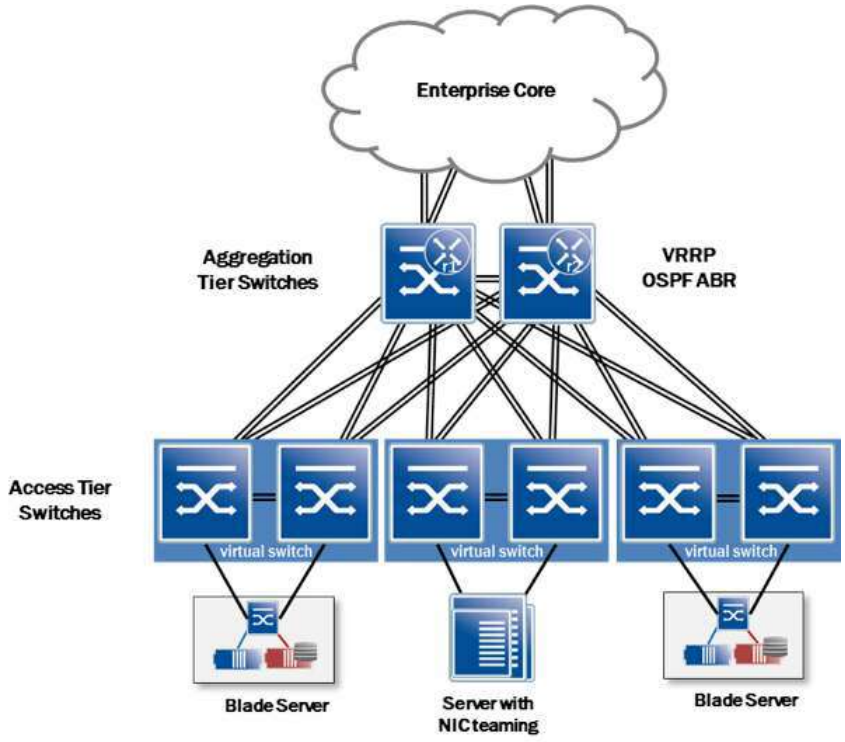


Figure 3: Two-tier data center design

Three-Tier Design

The three-tier data center design is comprised of access switches connected to servers, aggregation switches for access switch aggregation and data center core switches providing routing to and from the enterprise core network. The three-tier design is based on a hierarchical design so its main benefit is scalability. One could add new aggregation switch pairs with no need to modify the existing aggregation pairs. With routing being done by data center core switches, no full mesh is required. The disadvantages of three-tier design are higher latency due to the additional layer, additional congestion/oversubscription in the design (unless bandwidth between nodes is dramatically increased), more managed nodes (adding a certain amount of complexity for operation & maintenance), higher energy consumption and the need for additional rack space. Figure 3 shows a typical three tier data center architecture.

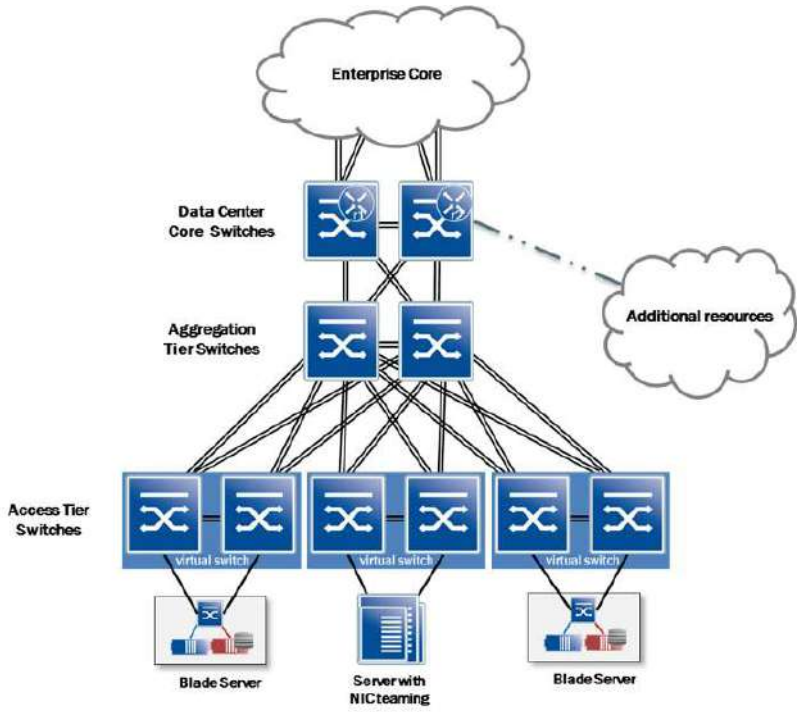


Figure 4: 3-tier data center design

Top of Rack (ToR)

Top of Rack (ToR) designs are often deployed in data centers today. Their modular design makes staging and deployment of racks easy to incorporate with equipment life-cycle management. Also cabling is often perceived to be easier when compared to an End of Row (EoR) design, especially when a large amount of Gigabit Ethernet attached servers are deployed.

Additionally cabling, cooling, rack space, power and services costs must also be carefully evaluated when choosing an architecture.

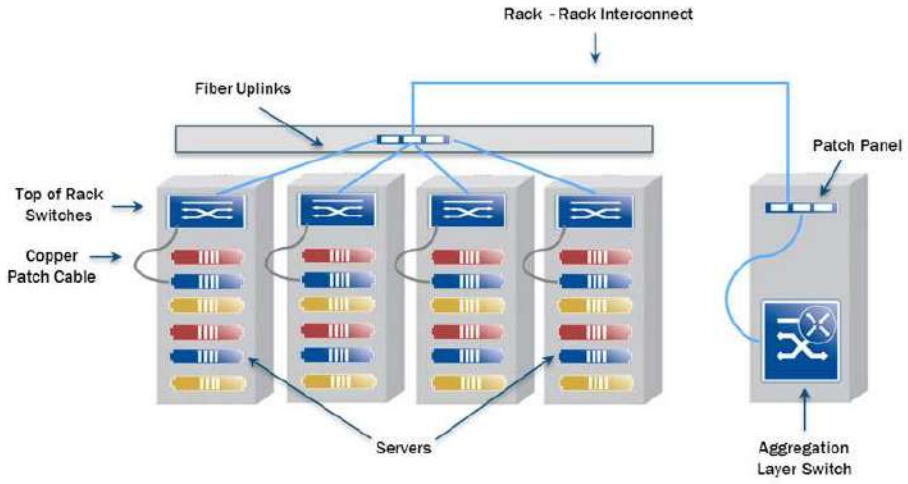


Figure 5: Top of Rack design

But ToR also has some disadvantages, such as:

- ToR can introduce additional scalability concerns, specifically congestion over uplinks and and shallow packet buffers which may prevent a predictable Class of Service (CoS) behavior.
    - In an EoR scenario this can be typically achieved by adding new line cards to a modular chassis
- Upgrades in technology (i.e. 1G to 10G, or 40G uplinks) often result in the complete replacement of a typical 1 Rack Unit (RU) ToR switch
- Number of servers in a rack varies over time, thus varying the number of switch ports that must be provided
    - Unused CAPEX sitting in the server racks is not efficient
- Number of unused ports (aggregated) will be higher than in an End of Row (EoR) scenario
    - This can also result in higher power consumption and greater cooling requirements compared to an EoR scenario

These caveats may result in an overall higher Total Cost of Ownership (TCO) for a ToR deployment compared to an EoR deployment. Additionally cabling, cooling, rack space, power and services costs must also be carefully evaluated when choosing an architecture. Lastly a ToR design results in a higher oversubscription ratio towards the core and potentially a higher degree of congestion. A fabric-wide quality of service (QoS) deployment (with the emerging adoption of DCB) cannot fully address this concern today.

End of Row (EoR)

Another data center topology option is an End of Row chassis-based switch for server connectivity. This design will place chassis-based switches at end of a row or the middle of a row to allow all the servers in a rack row to connect back to the switches.

Compared to a ToR design the servers can be placed anywhere in the racks so hot areas due to high server concentration can be avoided. Also the usage of the EoR equipment is optimized compared to a ToR deployment, with rack space, power consumption, cooling and CAPEX decreased as well. The number of switches that must be managed is reduced with the added advantages of a highly available and scalable design. Typically chassis switches also provide more features and scale in an EoR scenario compared to smaller platforms typical of ToR designs. On the other hand, cabling can be more complex as the density in the EoR rack increa

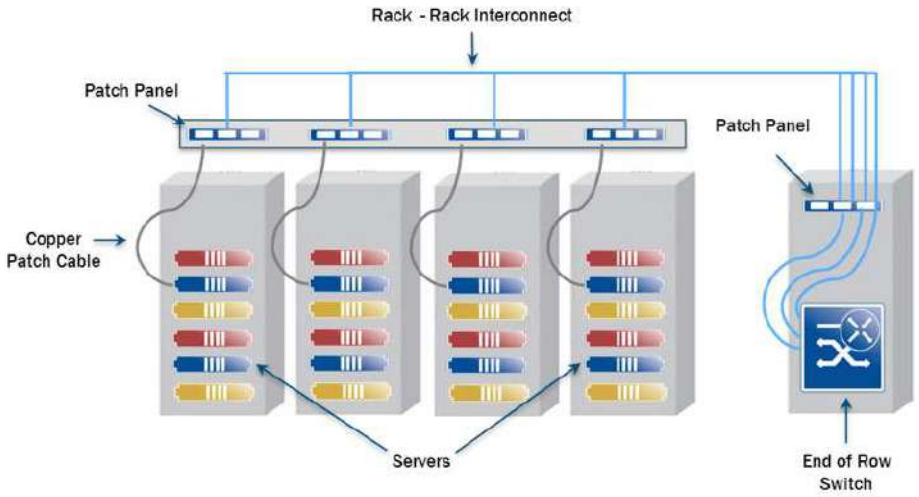


Figure 6: End of Row design

Data Center LAN Fabric Resiliency

Server virtualization has changed the requirements for how systems are connected to the network. Regardless of physical topology of the network (EoR or ToR) and the hypervisor vendor being used, there is a set of basic requirements which these systems demand from the network. As the consolidation of servers increases, so does the need for resiliency.

Server connectivity has several requirements:

- Must have redundant connections
- Should be load sharing (active-active)
- Must be highly automated

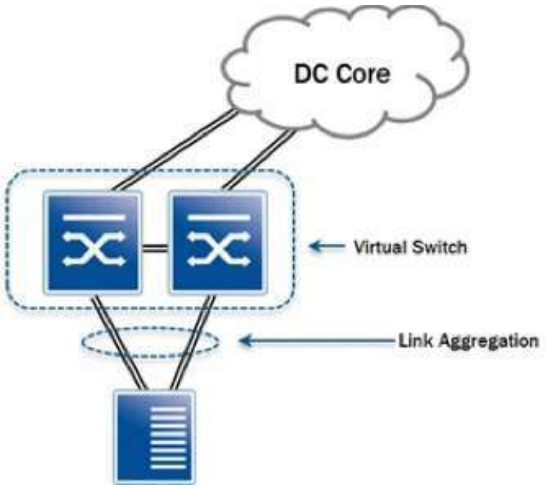


Figure 6: End of Row design

Link Aggregation (aka NIC teaming, or bonding depending on the vendor) is defined in the IEEE 802.1AX/802.3ad standard, which defines active load-sharing and redundancy between two nodes using an arbitrary number of links. Solutions have been developed by NIC card vendors in the past to prevent single points of failure by using special device drivers that allow two NIC cards to be connected to two different access switches or different line cards on the same access switch. If one NIC card fails, the secondary NIC card assumes the IP address of the server and takes over operation without connectivity disruption. The various types of NIC teaming solutions include active/standby and active/active. All solutions require the NIC cards to have Layer 2 connectivity to each other.

The resilient network is able to automatically adapt to failures in the system and provide assured application connectivity and performance.

Link Aggregation across Two Switches

Server and hypervisor manufacturers in general recommend two switches for server connectivity, addressing the first set of requirements for server connectivity and redundancy.

Redundancy does not necessarily meet the second requirement of load sharing. To do this, vendors would traditionally use NIC Teaming (TLB, SLB and the like) and manually configure the server to allocate virtual servers to specific ports or use stackable switches that form a single switch unit through the stack interconnect.

A resilient network meets all of the challenges above, incorporating redundant connections that dynamically distribute bandwidth across all available paths and automates the provisioning of systems connectivity. The resilient network is able to automatically adapt to failures in the system and provide assured application connectivity and performance. Enterasys virtual switching provides

a resilient infrastructure option in conjunction with Link Aggregation to the connected servers.

All of the server attachment technologies are NIC dependant. A standard mechanism to use is IEEE 802.3ad Link Aggregation but this does not work with two different switches unless these switches present themselves to the server as single entity. This can be accomplished as part of a stackable switch (such as the Enterasys B-Series or C-Series) or via virtual switching functionality currently provided by the Enterasys S-Series , 7100-Series or K-Series in the future.

Capacity and Performance Planning Considerations for the Data Center

High Availability Requirements

High Availability (HA) is crucial to data center networks. Data center failure costs include both revenue lost and business creditability. System availability is simply calculated by "system uptime" divided by "total time."

Availability =(MTBF)/(MTBF+MTTR) where MTBF is Mean Time Between Failure, MTTR is Mean Time To Repair

The table below shows availability percentage and down time per year.

Availability	Down time per year		
99.000%	3 days	15 hours	36 minutes
99.500%	1 days	19 hours	48 minutes
99.900%		8 hours	46 minutes
99.950%		4 hours	23 minutes
99.990%			53 minutes
99.999%			5 minutes
99.9999%			30 seconds

Figure 8: Calculated network down time per year

Typically, network architects expect to see 4 or 5 "nines" system availability. Each additional "9" can raise deployment costs significantly. To achieve a data center with near zero down time, we need to consider both system/application resiliency and network resiliency. For connectivity itself, there are two aspects to consider:

- System level resiliency: increasing MTBF by using reliable and robust hardware and software designed specifically for HA and minimizing the MTTR by using resilient hardware.
- Network level resiliency: this is achieved by not only designing the network with redundant/load sharing paths between network equipment but also through the support of fast convergence/fast rerouting features.

Furthermore, one must also consider data center site redundancy:

- Warm standby: In this scenario, the primary data center will be active and provide services while a secondary data center will be in standby. The advantage to warm standby is simplicity of design, configuration and maintenance. However, the disadvantage is no load sharing between two sites, which leads to under utilization of resources, inability to verify that the failover to secondary site is fully functional when it is not used consistently during normal operation, and an unacceptable delay in the event that a manual cutover is

required. It is also difficult to verify that the "warm" failover is functional when it is not used during normal operation.

- Hot standby: In this set-up, both the primary and secondary data centers provide services in a load sharing manner, optimizing resource utilization. The disadvantage to this scenario is that it is significantly more complex, requiring the active management of two active data centers and implementation of bi-directional data mirroring (resulting in additional overhead and more bandwidth between the two sites).

Oversubscription in the Data Center Network

The acceptable oversubscription in a data center network, is highly dependent on the applications in use and is radically different than in a typical access network. Today's design of presentation/web server, application server and database server "layers" combined with the new dynamics introduced through virtualization make it hard to predict traffic patterns and load between given systems in the data center network. The fact is that servers which use a hypervisor to virtualize applications yield higher performance and the resulting average demand on the interfaces belonging to these systems will be higher than on a typical server.

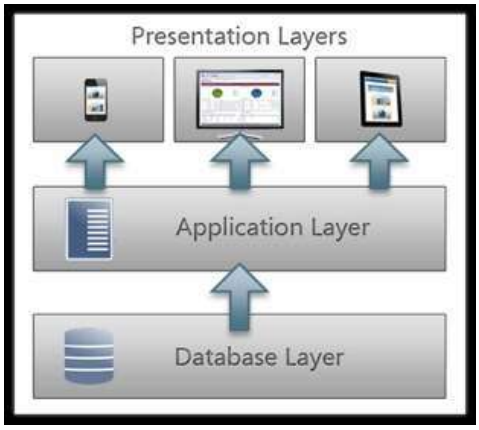


Figure 9: Typical application server design

Also, if virtual desktops are deployed, one has to carefully engineer the oversubscription and the quality of service architecture at the LAN access as well. Typically 0.5 to 1 Mbit/s per client must be reserved – without considering future streaming requirements.

Challenges with oversubscription include:

- Potential for congestion collapse
- Slow application performance
- Potential loss of control plane traffic

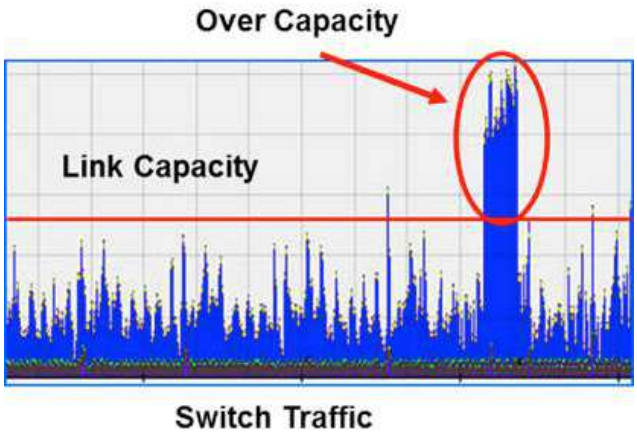


Figure 10: Oversubscription in the network

In general, oversubscription is simply calculated by using the ratio of network interfaces facing the downstream side versus the number of interfaces facing the upstream side (uplink) to the data center core. For example, in a server access switch that consists of 48 Gigabit Ethernet ports with two load sharing 10G Ethernet uplinks, the ratio of server interface bandwidth to uplink bandwidth is 48 Gbps/20Gbps, or 2.4:1, traditionally an acceptable ratio.

In planning for oversubscription, a few things should be taken into consideration:

- Traffic flow/direction (client-to-presentation server and server-to-server traffic flows)
- Link failure scenarios

In using the spanning tree protocol, if the root port of a downstream switch fails and the backup link becomes forwarding, there is no change to the oversubscription rate. The diagram below explains how oversubscription is calculated with an RSTP design. Let's assume that each access switch has 24 Gigabit Ethernet ports with two 10G Ethernet uplinks. One port is forwarding and another is the alternative port. The oversubscription ratio at an edge switch is 24:10 (2.4:1)

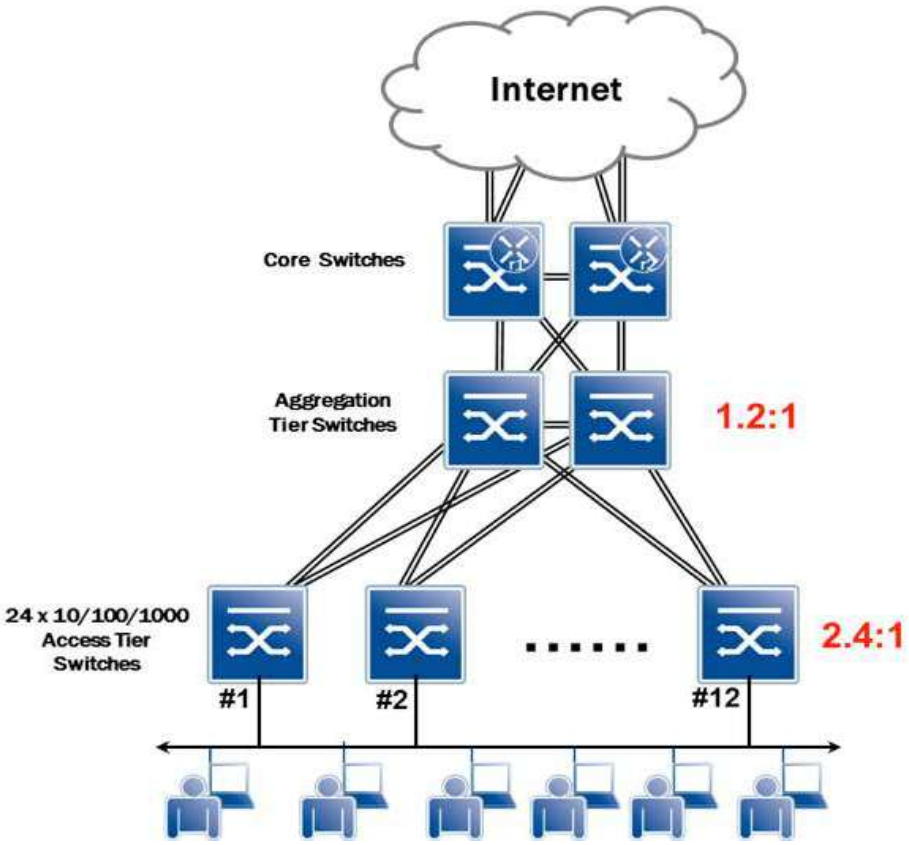


Figure 11: Oversubscription calculation with a spanning tree design

In the case of a virtual chassis/virtual switch, all the links between switches are active and allow traffic to flow through. In the diagram below, the oversubscription ratio at the edge switch is 24:20 (1.2:1). In the case of a single link failure between an edge switch and distribution switch, the oversubscription ratio at the edge switch will change to 2.4:1. If we assume that traffic utilization between an edge switch and distribution switch is at 70% or higher, the 2X oversubscription could cause serious congestion and packet drop even in a single link failure scenario. So if it is necessary to maintain the desired oversubscription rate in the event of single link failure, additional interfaces may be required in the design.

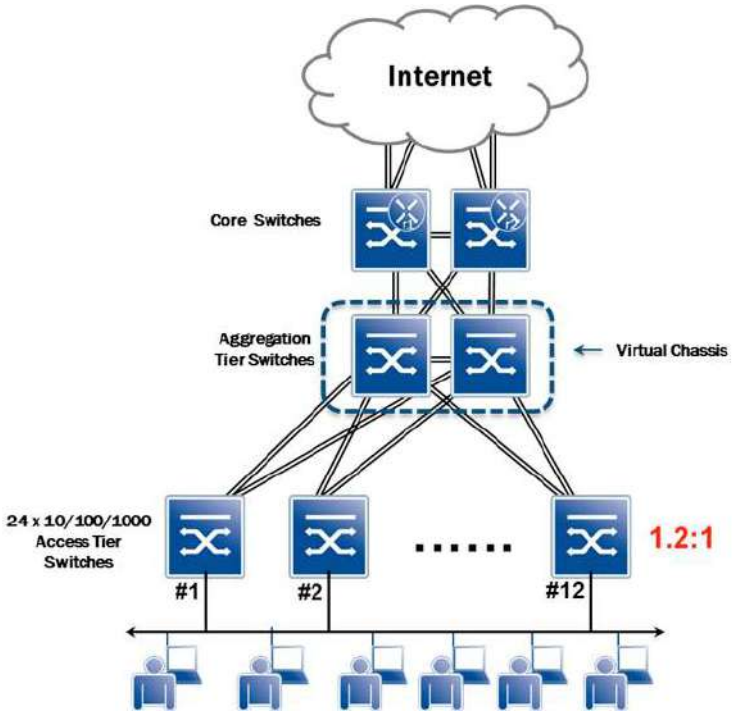


Figure 12: Oversubscription calculation with virtual chassis design

Incast, Micro Bursts, Buffering

Regardless of how the network oversubscription is designed, one has to be aware of the fact that storage technologies will create a completely different traffic pattern on the network than a typical user or VDI (Virtual Desktop Infrastructure) session. Storage traffic typically bursts to very high bandwidth in the presence of parallelization (especially within storage clusters which serve a distributed database). New standards like parallel Network File System (pNFS) increase that level of parallelization towards the database servers. This parallelization will often lead to the condition in which packets must be transmitted at the exact same time (which is obviously not possible on a single interface); this is the definition of an "incast" problem. The switch needs to be able to buffer these micro bursts so that none of the packets in the transaction get lost, otherwise the whole database transaction will fail. As interface speeds increase, large network packet buffers are required. The Enterasys S-Series is perfectly positioned with a packet buffer that exceeds 2 Gigabytes per I/O slot modules to solve this problem.

Topology – Logical Designs

Layer 2 Edge Designs

RSTP, MSTP

The original Spanning Tree (STP - IEEE 802.1D) algorithm was designed with maximum stability and safety in mind. In the event of a failure, all bridges adapt themselves to the new information sent by the root bridge, slowly unblocking their ports to ensure loop-free topology.

Rapid Spanning Tree Protocol (RSTP - IEEE 802.1w) has been designed to greatly improve convergence times. RSTP actively confirms that a port can safely transition to the forwarding state without having to rely on state machine timeouts as defined by IEEE 802.1D.

Multiple Spanning Tree Protocol (MSTP - IEEE 802.1s) was built on top of IEEE 802.1w so it inherits the fast re-convergence of the network with the added capability of improved link bandwidth utilization by separating spanning tree instances for a group of VLANs. To do this, any given bridge port could be in the "forwarding" state for some specific VLANs while in the "blocked" state for other VLANs.

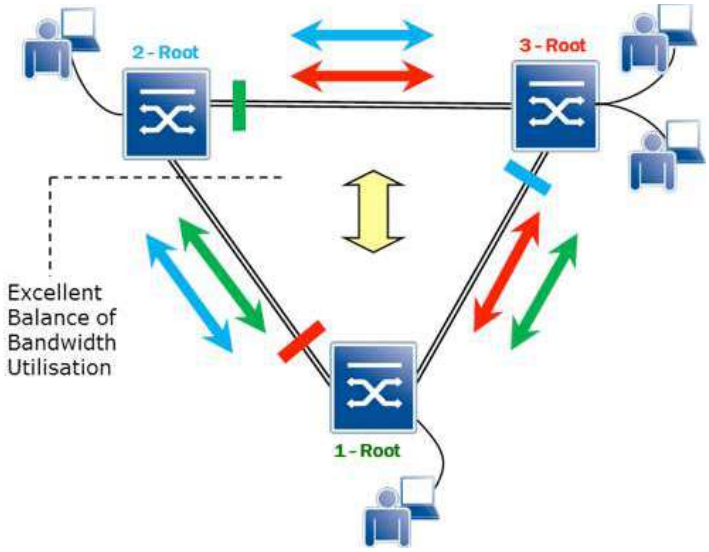


Figure 13: MSTP topology

With MSTP, multiple region designs are possible. With these designs, each MST region spans independently. This means link failures within a region would not cause re-span in other regions, which leads to increased stability of the network, especially for large networks.

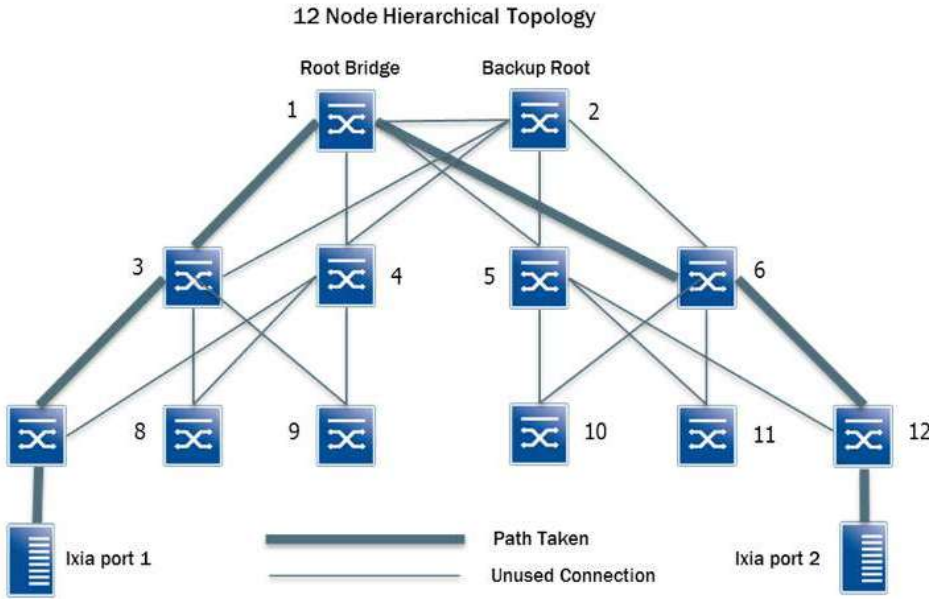


Figure 14: RSTP hierarchical topology

In this configuration 12 switches are connected in a hierarchical topology which is typical for a 3-tier data center design. A test was conducted by removing the link between bridges 1 and 3 in

which bridge 1 was the root and bridge 2 was the backup root (an Enterasys feature that enhances failover times in the case of root failures), and vice versa. In this case, the failover times averaged between 0.26 and 0.41 seconds, with an average of 0.40 seconds. The result shows that the use of RSTP/MSTP in today's data center networks is a viable and standards-based option depending on the failure recovery time requirements .

Shortest Path Bridging (SPB)

Shortest Path Bridging (SPB) IEEE 802.1aq was developed as an evolution of the various Spanning Tree protocols. SPB leverages the IS-IS link state protocol for building a global view of the switch topology and to control the layer 2 data plane. SPB and IS-IS build shortest path trees for each node to every other node within the SPB domain. These unique shortest path trees ensure efficient usage of available links within the SPB mesh by always using the shortest path between any two nodes in the domain. Where multiple equal cost paths exist, SPB provides Equal Cost Multipath (ECMP) algorithms to further distribute the load and efficiently utilize equal path links through the network. SPB's IEEE 802.1 heritage ensures full interoperability with the existing RSTP/MSTP topologies, in fact SPB leverages the spanning tree state machine for controlling forwarding on a per shortest path tree basis.

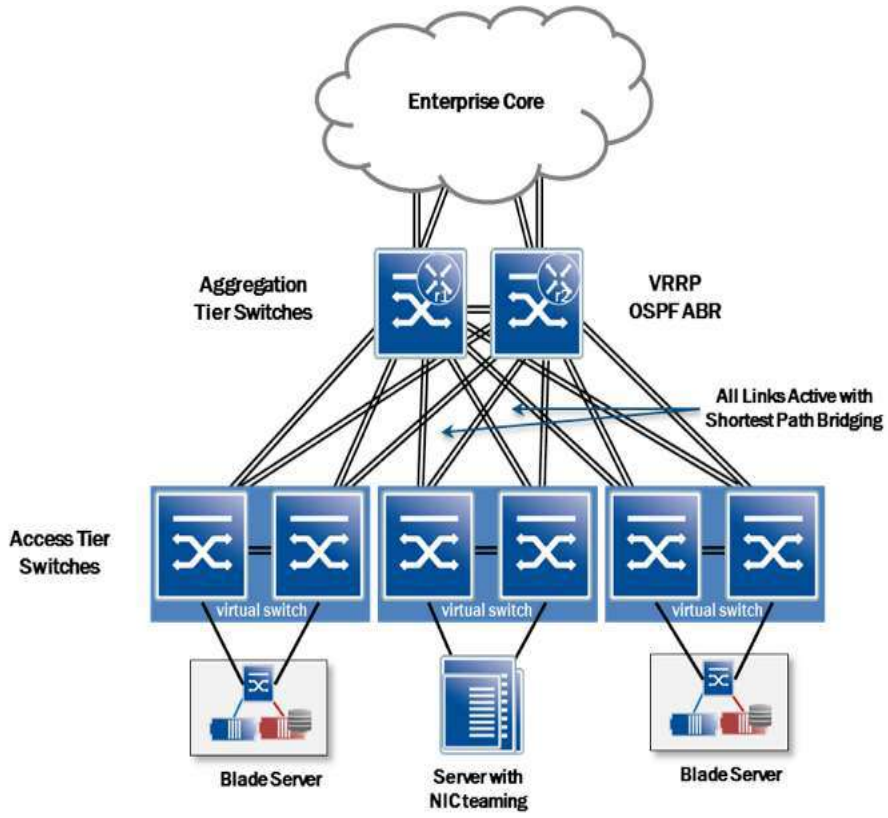


Figure 15: SPB topology

Fully meshed data center designs leveraging Shortest Path Bridging provide load-sharing through the efficient use of multiple paths through network.

Virtual Switching

Enterasys Virtual Switch Bonding merges physical switches into a single logical switch. This functionality provides redundancy at the server access layer and could also be used on the aggregation layer. Logically the virtual switch is a single managed system that dynamically provisions trunked server connectivity using IEEE 802.1AX/802.3ad link aggregation protocols. Dynamic trunk provisioning can lower OPEX overhead in comparison to static server NIC teaming. In virtualized configurations, assigning virtual hosts to an aggregated link provides better

application performance and reduces the need for hypervisor network configuration. Enterasys virtual switching provides:

- Automated link aggregation across physical switches
- Meshed L2 network uplink to data center aggregation/core switches
- Non-stop forwarding of application traffic
- Automated "host-specific" network/security profiles per virtual host, per port
- Support for thousands of virtual hosts per system

Enterasys Virtual Switch Bonding (VSB) is supported with the Enterasys S-Series platform and 7100-Series products. S-Series VSB allows two chassis to be fully virtualized to form a single entity via dedicated hardware ports. The S-Series depending on the model can use either multiple ordinary 10G ports or multiple dedicated VSB ports to form the high speed link between chassis. 7100-Series virtual switch bonding will allow up to eight switches to form a single entity.

One has to be aware that VSB (like other implementations) may reduce overall availability, especially when configuration errors by the network administrators occur. Outages due to misconfiguration of components are still happening today even as processes within the organizations look to eliminate them. Since a virtual switch acts like a single switch such a configuration error or other problems during software upgrades can result in decreased overall availability of the solution. Therefore, it is recommend to use independent switches in the core of a data center network that interact with each other via standard protocols like IEEE RSTP/MSTP, IEEE Shortest Path Bridging (SPB), IETF OSPF.

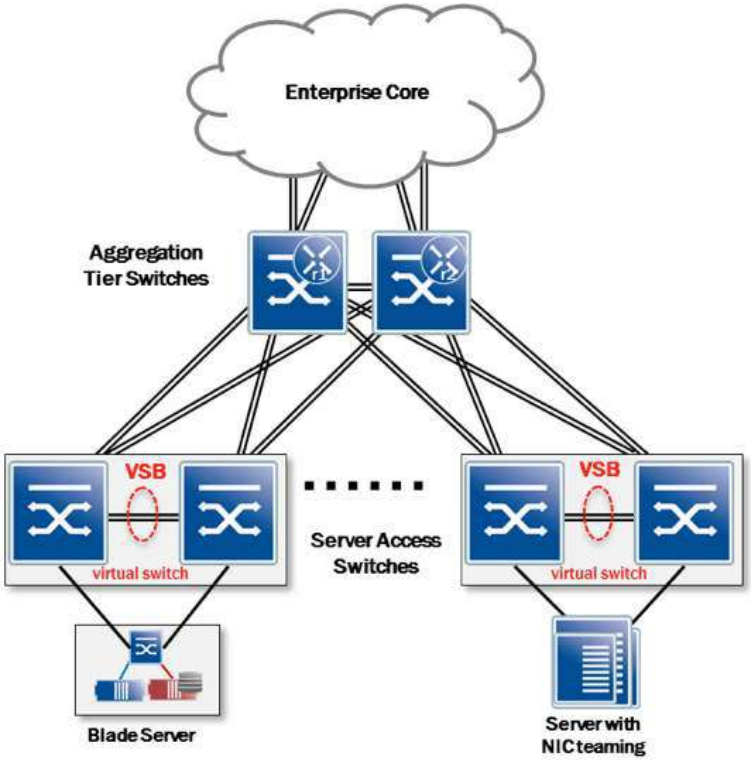


Figure 16: Enterasys virtual switching

Layer 2 Core Designs

If we are to meet the needs of the applications driving the business, resiliency is required across the entire network, not just at a particular layer. Today's Layer 2 meshed design derives its resiliency by leveraging industry standard protocols including IEEE 802.1s (MSTP) and IEEE 802.1w (RSTP). These protocols not only provide interoperability with existing data network equipment, their maturity provides the administrator with a rich set of tools and a trained workforce who can implement and maintain the network topology. Interoperability and maturity provide for greater uptime and lower OPEX when compared to proprietary solutions.

Implementing data center connectivity in a meshed design, MSTP helps attain the goals of redundancy, load-sharing and vendor interoperability. Best practices designs leveraging the benefits of MSTP can provide traffic-shaping, redundancy and automated resiliency within the Layer 2 meshed network. RSTP accelerates topology change events should they occur, ensuring application availability to the consumer.

Fully meshed data center designs leveraging Shortest Path Bridging provide load-sharing through the efficient use of multiple paths through network. Shortest Path Bridging builds upon the existing Data Center LANs and improves the resiliency of the networks because they:

- Have the ability to use all available physical connectivity
- Enable fast restoration of connectivity after failure
- Restrict failures so only directly affected traffic is impacted during restoration; all surrounding traffic continues unaffected
- Enable rapid restoration of broadcast and multicast connectivity simultaneously

Shortest Path Bridging comes in 2 versions – SPBV, using 802.1Q VLAN translation data plane forwarding and SPBM using 802.1ah MAC-in-MAC encapsulation for data plane forwarding. The SPB standard use of the IS-IS link state protocol as the topology discovery protocol for building a layer 2 mesh is a similar approach that Enterasys/Cabletron used back in 1996. At the time, Cabletron's VLSP (VLAN Link State Protocol) leveraged OSPF's link state functionality for MAC address forwarding to create a layer 2 full mesh. SPB interoperability is in the early stages as various vendors are implementing the standard and is gaining momentum.

Enterasys is committed to open standards, and these protocols show promise for delivering a more reliable and interoperable data center. This is especially true of SPB due to its full interoperability with RSTP/MSTP and standardization by the IEEE.

Customers considering a new data center network that are not ready for SPB, should consider a design built on a layer 2 core with standard RSTP/MSTP protocols as this design will enable an easy, non-disruptive migration toward Shortest Path Bridging when the time comes.

Layer 3 Core Designs

Layer 3 meshed core networks focus on two key principles, route availability and gateway availability. Two industry standard protocols provide Layer 3 networks with this capability, IETF OSPF-ECMP and IETF VRRP.

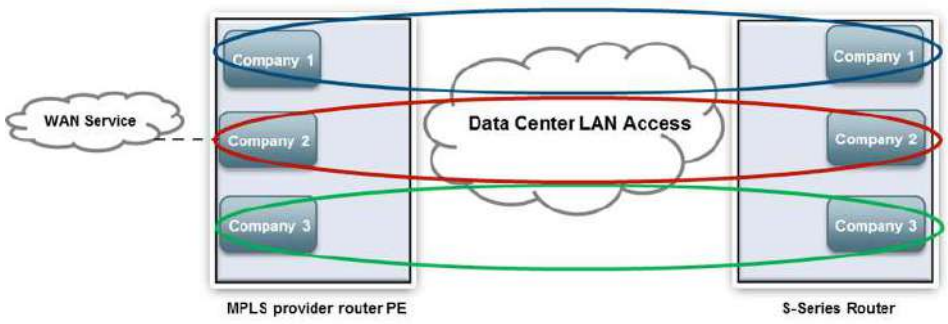
Load Sharing – OSPF, VRRP, Fabric Routing

OSPF-ECMP enables Layer 3 meshed networks to negotiate a Layer 3 (routed) meshed network and load balance connectivity across the available paths in the network. This allows network designs to leverage all paths for data traffic ensuring capital investments are leveraged and not just used for insurance. Additionally, OSPF-ECMP provides additional traffic engineering capabilities to the network operator to ensure that critical applications have the necessary bandwidth and circuit availability. Combining VRRP's automated gateway redundancy with OSPF-ECMP provides interoperable Layer 3 resiliency today with similar maturity of tools as the Layer 2 options previously described.

Central to all data center designs is the need for optimized traffic routing within the data center as well as between datacenters. Enterasys leverages standards based VRRP to provide a single virtual router gateway shared across multiple physical devices to provide redundancy and layer 3 resiliency. Enterasys Fabric Routing is an enhancement to VRRP that optimizes the flow of east/west traffic within the datacenter by allowing the closest router to forward the data regardless of VRRP mastership. Fabric Routing is an excellent solution for intra-data center traffic but does not solve the issue of optimizing external traffic flows that need to enter the data center. The inefficient and potential asymmetric traffic flow is solved by the implementation of host routing enhancements to the Fabric Routing functionality allowing IP host mobility. With this enhancement, Fabric Routing is extended such that a fabric router that forwards user traffic will distribute a specific host route into the respective routing protocols. This host route advertisement ensures efficient symmetric return path traffic flows into the data center.

Separation – VRF, MPLS, L3VPN

Virtual Routing and Forwarding (VRF) allows multiple independent routing instances to exist on a single router. It completely separates customers or departments based on routing domains, enabling secure, dedicated routing resources for critical applications. VRF provides a simple solution for campus LAN and data center applications. It is the natural extension of and was the precursor to provider MPLS VPN services into a data center, while not exposing the operator to the complexity of MPLS. On a larger scale the use of BGP/MPLS L3VPN allows the transport of customer VRF/VPN data without extending customer specific routing information across a common IP backbone by leveraging multi-protocol BGP and MPLS or IP tunneling as the encapsulation. The use of IP tunneling removes the complexity associated with implementing MPLS in the backbone.



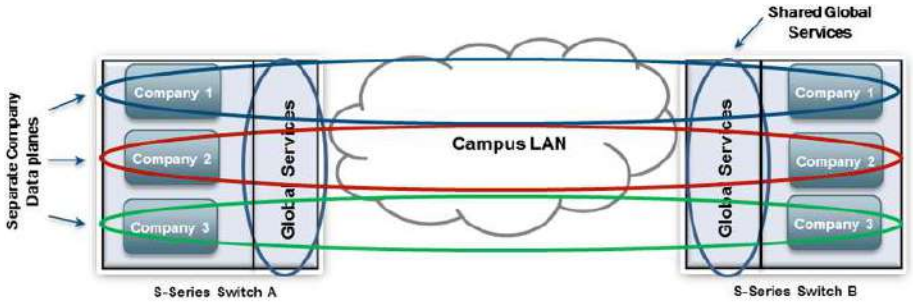


Figure 17: VRF/L3VPN design

Data Center Interconnect

The evolving traffic patterns of clusters, servers and storage virtualization solutions are demanding new redundancy schemes. These schemes provide the transport technology used for inter-data center connectivity and the geographical distances between data centers and are critical as the network design evolves to provide ever higher levels of stability, resiliency and performance.

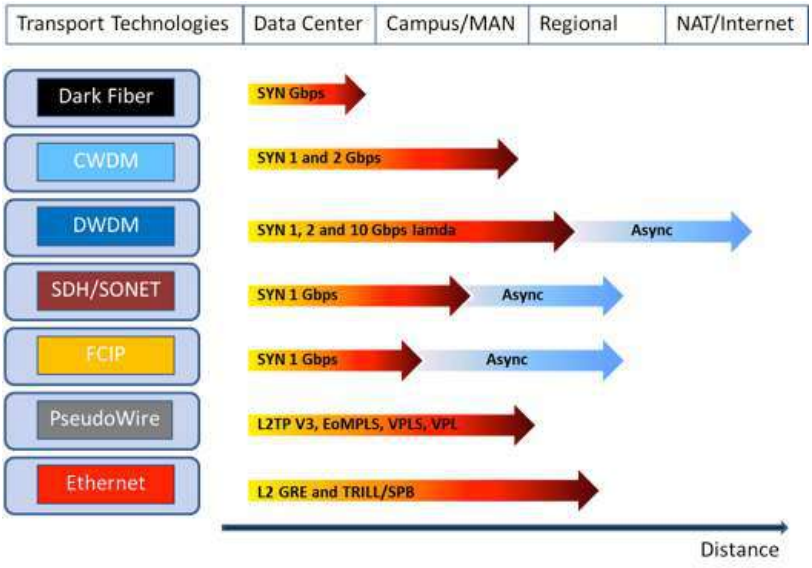


Figure 18: Physical logical DCI technologies

The transport technology of choice between data centers is dependent upon several requirements:

- Synchronous or asynchronous data replication
- Jitter and delay acceptance for virtualized applications and their storage
- Jitter and delay acceptance for cluster solutions
- Available bandwidth per traffic class
- Layer 2 or Layer 3 interconnect

An important issue when operating a load-balanced service across data centers and within a data center is how to handle information that must be kept across the multiple requests in a user's session. If this information is stored locally on one back end server, then subsequent requests going to different back end servers would not be able to find it. This might be cached information

that can be recomputed, in which case load-balancing a request to a different back end server just introduces a performance issue.

One solution to the session data issue is to send all requests in a user session consistently to the same back end server. This is known as "persistence" or "stickiness". A downside to this technique is its lack of automatic failover: if a backend server goes down, its persession information becomes inaccessible, and sessions depending upon it are lost. So a seamless failover cannot be guaranteed. In most cases dedicated hardware load balancers are required.

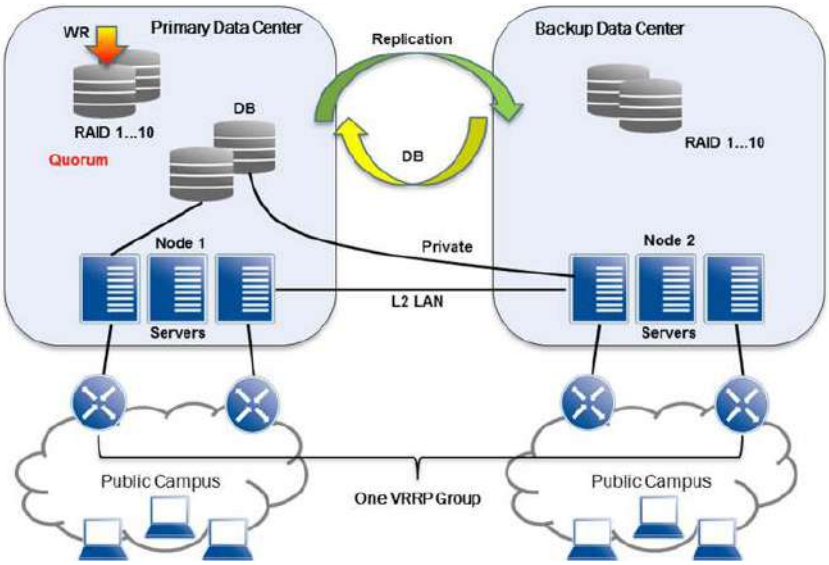


Figure 19: Cluster node separation across two data centers

The discussion about load balancing and persistence has a great impact on separation. Figure 19 shows a typical situation for cluster node separation across two redundant data centers. In this example the node separation of different clusters types with shared nothing and shared data bases are shown.

In many cases, the same subnet is used across both of the data centers, which is then route summarized. The "cluster" subnet will be advertised as an external route using "redistribute connected" and by filtering all subnets except the cluster subnet. While redistributing, the primary data center will be preferred to the remote data center by lower path cost until such time as the primary data center disappears completely.

The clients placed within the public campus network access the data center services across redundant routers which are grouped together in one VRRP group. In this configuration it is important to have greater VRRP priority for the primary data center. However this might cause problems in event of failover, when the traffic must be re-routed from the primary data center to the backup data center. This is especially true when traffic traverses stateful firewalls, when one has to make sure that traffic on both directions passes the same firewall system. Techniques for VRRP interface or next hop tracking can make sure that this is covered appropriately. To provide database access across both data centers at any time, connectivity between access switches and storage systems must be duplicated. Replication of databases must be achieved through Layer 2 techniques, such as VPLS, GRE, SPB, or with 802.1Q and RSTP/MSTP along with 802.3ad Link Aggregation or possibly through switch clustering/bonding techniques. In all cases one will face huge demand for bandwidth and performance that can be quite expensive for WAN links and must be properly sized.

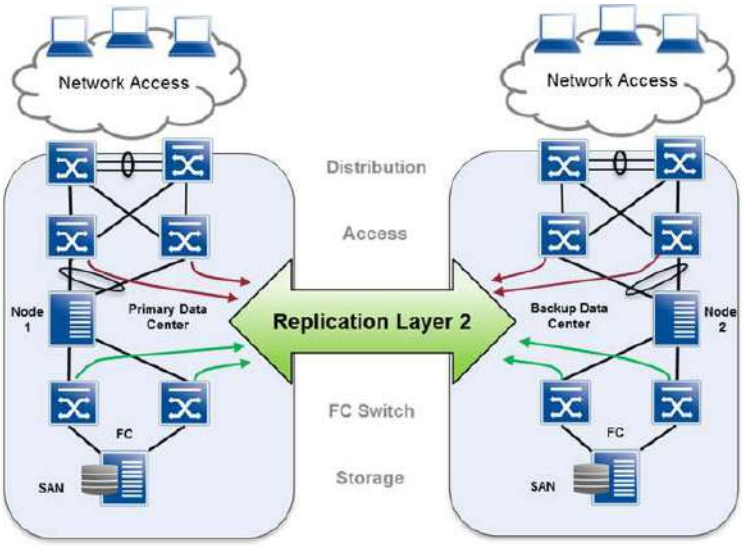


Figure 20: Data replication across data centers

Service and Security Layer Insertion

A modularized service and security layer should also reside within the data center and not in the core network itself. The aggregation/distribution layer is the best suited enforcement point for additional services like VPN, IPS, firewall security and others. All servers can access these services with short but predictable latency and bandwidth in an equal fashion. High performance and intelligent Layer 4-7 application switches, such as the Enterasys S-Series, can be connected to aggregation/distribution for always-on, highly scalable and secure business critical applications or be part of that layer itself.

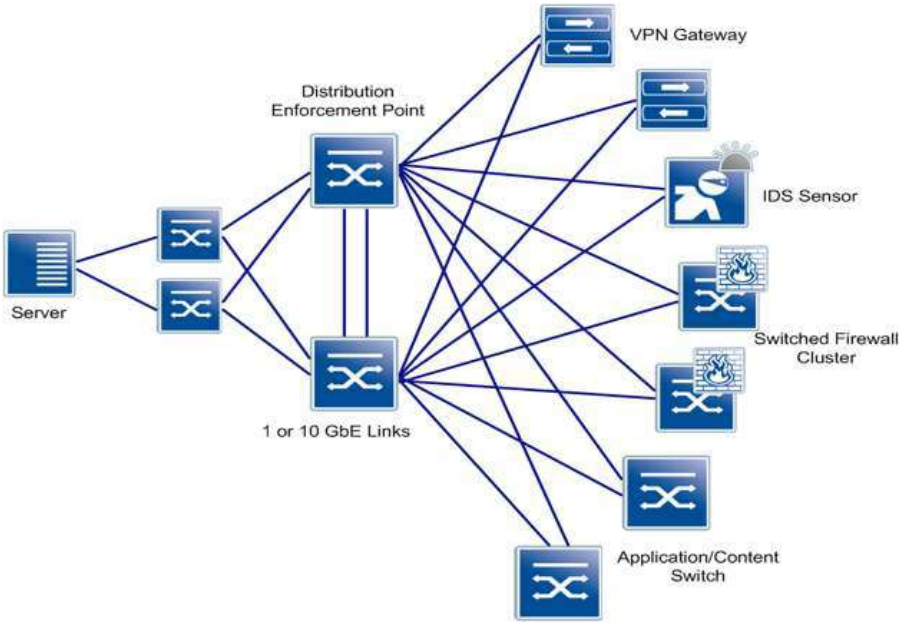


Figure 21: Security within the data center

Resources within the data center are segmented (virtually or physically) according to the services they supply and the security zones they serve. This segmentation provides an opportunity for further optimization of security and monitoring solutions. Figure 21 shows an example of application server pooling for different services such as:

- Web services – portals, web-based warehouses
- Applications services – enterprise resource planning
- Core service – DNS, DHCP, NTP, FTP, RADIUS
- Data base services – MS SQL, Oracle, Sybase

This segmentation allows the design to benefit from a Service Oriented Architecture (SOA), which includes the following advantages:

- Zones can be hosted by different managed service providers
- Borders between application categories, or zones, can be protected by effective security measures like firewalls, session border controllers and/or Intrusion Prevention Systems (IPS)
- Application performance is more predictable
- Distribution of malware or hacker attacks is limited to one zone
- Outages, failures and administration errors are restricted to one zone only

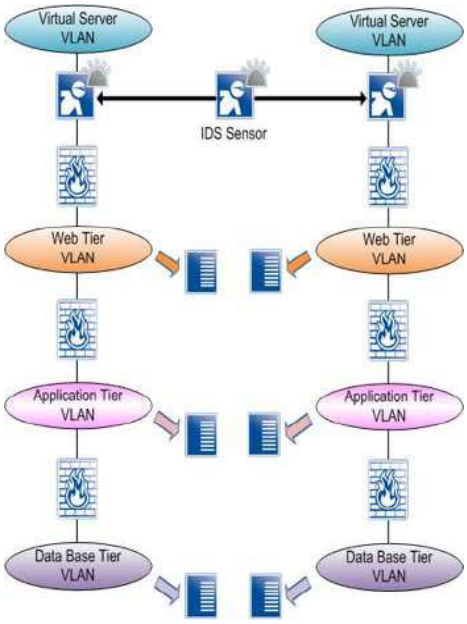


Figure 22: Segmented application security

Potential Concerns with Service and Security Layer Products

Network architects typically configure service modules and appliances to be in transparent (pass-through) mode, since these modules need to be able to be removed without requiring a reconfiguration of the entire system. When these modules are put in-line (all traffic passes through them), module throughput must be calculated so that the service modules will not introduce significant congestion into the system. One must avoid adding additional points of oversubscription whenever possible. For example, while traffic from clients to servers must pass through an IPS, traffic between servers may not need to. In addition to raw bandwidth, the number of concurrent sessions and the rate of connections per second that a security device supports can introduce additional performance issues. The number of concurrent sessions or connections per second can be calculated from the total number of servers and end users. While there's no general rule for this calculation, vendors will typically supply a recommendation based upon the use model and configuration.

The Enterasys S-Series provides LSNAT support on a per VRF basis allowing multiple tenants to each utilize the virtualization and load balancing capabilities separately on the same device.

Load Balancing with the Enterasys S-Series

Using the unique capabilities of Enterasys S-Series switches, a load balancing solution can be implemented without requiring any additional hardware. LSNAT (as defined in RFC 2391) allows an IP address and port number to be transformed into a Virtual IP address and port number (VIP) mapped into many physical devices. The Enterasys S-Series provides LSNAT support on a per VRF basis allowing multiple tenants to each utilize the virtualization and load balancing capabilities separately on the same device. When traffic destined to the VIP is seen by the LSNAT device, the

device translates it into a real IP address and port combination using a selected algorithm such as Round Robin, Weighted Round Robin, Least Load or Fastest Response. This allows the device to choose from a group of real server addresses and replace the VIP with the selected IP address and port number.

The LSNAT device then makes the appropriate changes to packet and header checksums before passing the packet along. On the return path, the device sees the source and destination pair with the real IP address and port number and knows that it needs to replace this source address and source port number with the VIP and appropriate checksum recalculations before sending the packet along. Persistence is a critical aspect of LSNAT to ensure that all service requests from a particular client will be directed to the same real server. Sticky persistence functionality provides less security but increased flexibility, allowing users to load balance all services through a virtual IP address. In addition, this functionality provides better resource utilization and thus increased performance.

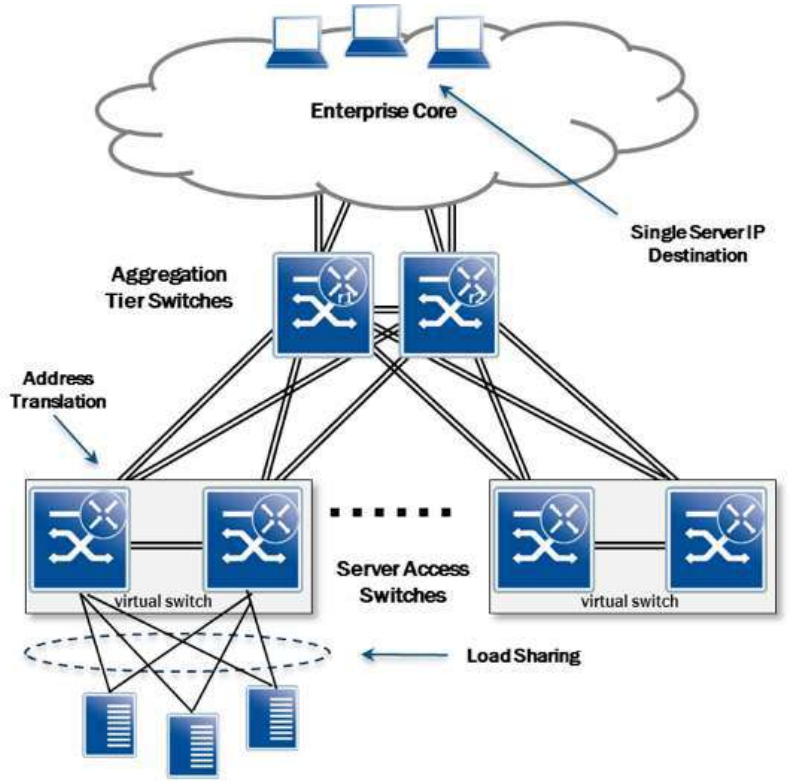


Figure 23: Example of server load balancing

A feature rich platform with a future-proofed backplane of more than 9.5 Tbps switching capacity, Enterasys switches are the best fit for large scale data center designs.

An essential benefit of using LSNAT is that it can be combined with routing policies. Configuring different costs for OSPF links, a second redundant server farm can be made reachable by other metrics. In this way, load balancing is achieved in a much more cost effective manner.

Data Center Connectivity – Best Practices with Enterasys Products

Enterasys recommends a 2-Tier design with a collapsed data center core for small to medium size data centers due to the many benefits gained with this solution as described in previous sections. However, the 3-tier design is preferable with very large data center designs in order to obtain better

Data Center Design Best Practices

Layer 2 Best Practices

- SPB, MSTP or virtual switch technology should be implemented in data center networks to gain full utilization of inter-switch links.
- Spanning tree is always needed. Even with the implementation of a virtual switch, STP is required as a backup mechanism in case of configuration errors that could disable the virtual switch or the aggregation links. SPB is built upon and fully interoperable with STP. SPB uses STP as a backup mechanism.
- With a spanning tree L2 design, STP root bridge and backup root bridge must be explicitly defined on data center core switches (or on a collapsed core) to achieve an optimal path for traffic flow, to and from servers.
- When implementing link aggregation between switches, oversubscription should be planned carefully. When one or more physical links in a link aggregation group go down, spanning tree or a routing protocol will not re-calculate the L3 forwarding path. Hence the same path is used with fewer physical ports, and the new oversubscription ratio could lead to congestion.
- When deploying a blade server chassis with integrated switches, it is not recommended to connect those integrated switches to a Layer 2 access switch. Doing so may increase latency and oversubscription. It is recommended to connect the blade servers directly to the aggregation switch.

scalability. A feature rich platform with a future-proofed backplane of more than 9.5 Tbps switching capacity, Enterasys switches are the best fit for large scale data center designs.

Enterasys' premier product for the data center, the S-Series provides the ability to collapse the traditional 3-tier network into a physical 2-tier network by virtualizing the routing and switching functions within a single tier. Virtualized routing provides for greater resiliency and fewer switches dedicated to pure switch interconnects. Reducing the number of uplinks (switch hops) in the data center improves application performance, reduces CAPEX and reduces meantime-to-repair (MTTR). This reduction in CAPEX includes not only the lower administrative costs but also the reduction of overall power consumption and cooling requirements. The fact that the S-Series can be deployed in the data center, core and distribution layer of the network reduces the overall cost to manage and maintain a network infrastructure dramatically with reduced spare parts, training, cooling costs, etc.

The Enterasys S-Series has all the advantages of Top of Rack virtual switching solution without requiring an independent chassis. The S-Series chassis implements a distributed switching architecture without a dedicated supervisor engine. In essence, the S-Series chassis is a virtual switch cluster with fully redundant switching and power systems. The S-Series provides a highly resilient distributed switching and routing architecture with management and control functions embedded in each module, delivering unsurpassed reliability, scalability, and fault tolerance for data center deployments. Organizations can cost-effectively add connectivity as needed while scaling performance capacity with each new module. The highly available architecture makes forwarding decisions, and enforces security policies and roles while classifying and prioritizing traffic at wire speed. All I/O modules provide the highest Quality of Service (QoS) features for critical applications such as voice and HD video even during periods of high network traffic load, while also proactively preventing Denial of Service (DoS) attacks and malware propagation.

The S-Series implements an industry-leading, flow-based switching architecture to intelligently manage individual user and application conversations, far beyond the capabilities of switches that are limited to using VLANs, ACLs, and ports to implement role-based access controls. Its classification capability from Layer 2 to Layer 4 will soon be extended beyond Layer 4 by using the S-Series unique flowbased ASIC technology, CoreFlow2. Users are identified and roles are applied to ensure each individual user can access their business-critical applications no matter where they connect to the network. S-Series role-based access rules can intelligently sense and automatically respond to security threats while improving reliability and quality of the end-user experience.

Visibility is key in the new data center network, visibility that integrates network topology, VM and services (application, presentation, database) location and performance requirements. The raw data is typically provided by non-sampled NetFlow records by the S-Series product suite in an Enterasys designed data center. An intelligent network management platform is needed to provide a comprehensive view of the entire data center infrastructure in a single comprehensive view. Enterasys NMS with Enterasys Data Center Manager is a powerful unified management system that provides IT administrators a transparent, crossfunctional service provisioning process that bridges the divide among the server, networking and storage teams. By presenting an integrated view of virtual server and network environments, this solution provides significant operational efficiencies among teams in the IT organization. With a unique vendor-agnostic approach, DCM supports a variety of virtualization, storage and server platforms, enabling the unified management of the physical and virtual network and ensuring networks will have the high availability necessary for critical applications and business data.

Furthermore, the S-Series is the only enterprise switch to support multi-user, multi-method authentication on every port, absolutely essential when you have virtual machines as well as devices such as IP phones, computers, printers, copiers, security cameras and badge readers connected to the data center network.

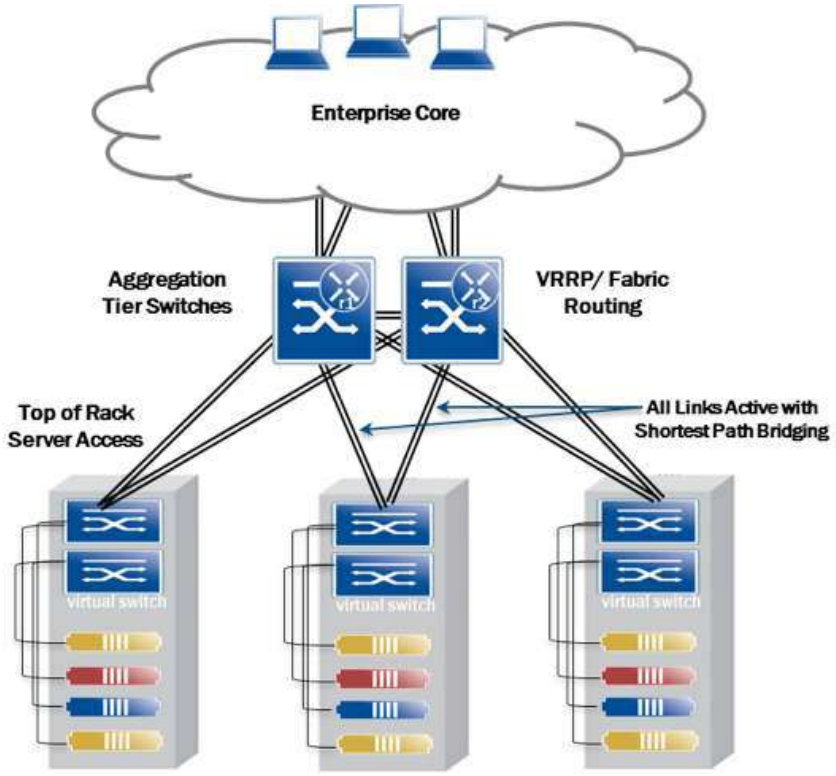
Layer 3 Best Practices

- Route summarization should always be implemented at data center aggregation or core switches at the interconnection to the enterprise core.
- Fabric and host routing should be used for optimized traffic routing within and between data centers.
- Routing protocol authentication should be implemented to protect the routing domain from unauthorized or misconfigured routers, which leads to service interruptions due to routing re-convergence.
- In a multi-vendors network, OSPF path cost should be tuned to be the same among L3 switches/routers inside the routing domain.
- Passive interfaces should be used on networks which have no routing peers.
- Point-to-point L3 fully meshed links should be implemented to immediately detect peer failures.

The following design examples demonstrate Enterasys best practices for Top of Rack, End of Row and Data Center Interconnect deployments.

Enterasys 2-tier Design – Top of Rack

With an Enterasys Top of rack design, a user could deploy a pair of 7100-Series switches in a virtual switch bond as virtual top of rack (ToR) switch. The ToR switches can be connected to an Enterasys S8 collapsed data center core via 10G or 40G uplinks. Figure23 depicts a Top of Rack solution hosting multiple dual homed servers. Optionally, 10G attached servers and blade centers can be directly connected to the aggregation switches – resulting in a hybrid ToR and EoR deployment. This design can leverage link aggregation from the virtual bonded switches and is relevant to MSTP or SPB environments.



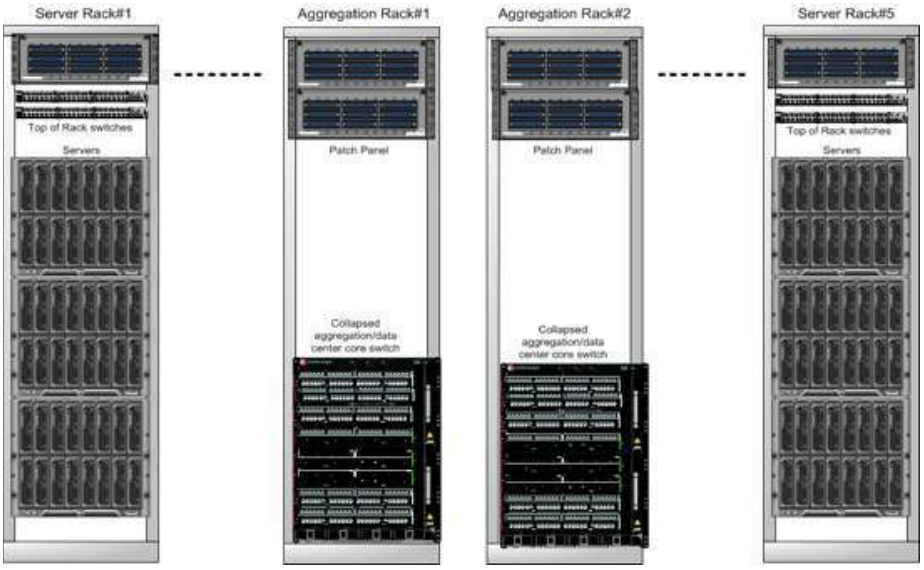


Figure 24: Enterasys Top of Rack (ToR) design

Enterasys 2-tier Design – End of Row

An End of Row solution can be implemented based on S-Series technology. Instead of using 1 RU Top of Rack switches, a user would implement a pair of modular chassis switches per server row. Figure 24 demonstrates a design example of multiple server access groups, using pairs of chassis based switches in a virtual switch bond supporting dual homed servers. Each access switch pair is connected to two aggregation/data center core switches. This design can leverages link aggregation from the virtual bonded switches and is relevant to MSTP or SPB environments.

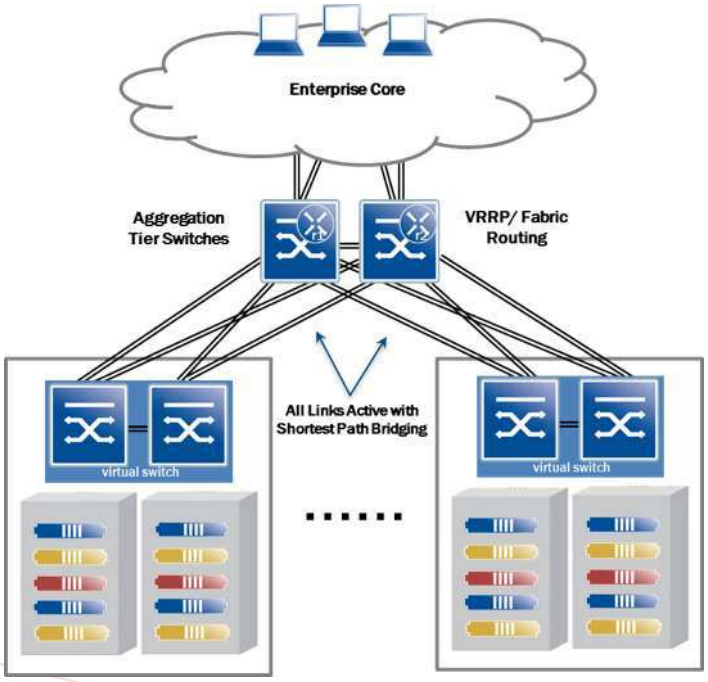


Figure 25: Enterasys End of Row design

Enterasys Data Center Interconnect

To meet the needs of the ever expanding server virtualization trends and redundancy demands, today's data center designs have specific requirements for inter-data center connectivity. Regardless of the geographical distance between data centers, layer 2 and layer 3 interconnect schemes are viable solutions supported today by Enterasys products. In many redundant data center designs the same subnet is used across both of the data centers, where the primary data center is specified with a lower path cost. The same VRRP IP and MAC are used in both locations to allow common gateway redundancy and allow seamless mobility. In this scenario, the primary data center will be preferred to the remote data center by lower path cost until such time as the primary data center disappears completely.

Layer 2 DCI

Redundant data center designs where the same subnet is used across multiple data centers may need direct layer 2 connectivity to perform functions such as data replication or hot VM movement. The simplest method to connect multiple data centers together is to extend the common VLAN(s) across backbone extending the layer 2 domain. This solution is a viable option for many smaller deployments but it should be considered on how this extension will impact the size of the spanning tree domain.

Another method for Layer 2 data center interconnect is to create a layer 2 tunnel between the data center sites. This solution allows layer 2 traffic to be transported across the layer 3 infrastructure transparently, with the added benefit of not extending the size of the size of the spanning tree domain. Enterasys S and K-Series leverages standard IP/GRE tunneling to interconnect the data centers. In this scenario both data centers see each other as part of a common layer 2 domain. Leveraging Enterasys Fabric routing and host routing functionality, devices or virtual machines can easily be moved between datacenters in a hot or cold manner. After moving to the new location, the VM will be reachable via its new location as a result of the VM host route advertisement by the local fabric router in the new location. Fabric and host routing optimize the flow of traffic into and between data centers by providing direct access to and from each data center symmetrically. The traffic optimization limits the amount of traffic that traverses the interconnect links to traffic that needs to go between data centers providing the added benefit of conserving bandwidth on potentially expensive data center interconnect links.

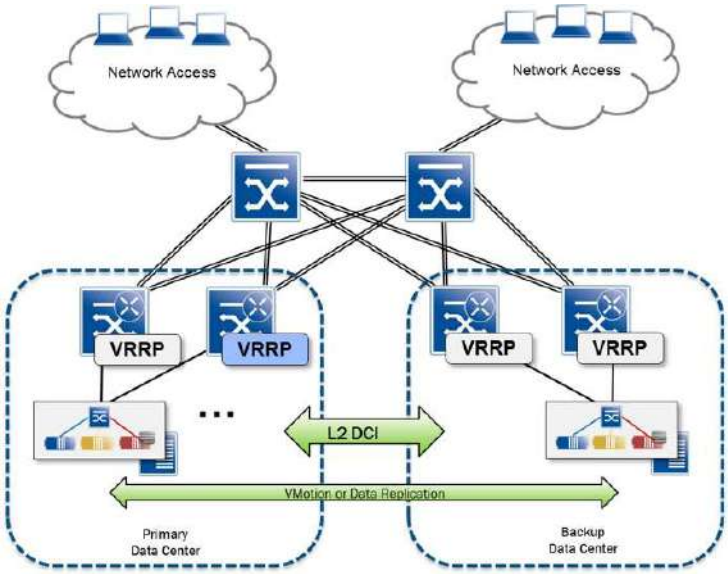


Figure 26: Enterasys Layer 2 DC Interconnect

Layer 3 DCI

Considering redundant data center designs where the same subnet is used across the primary and backup data center, a standard routed layer 3 interconnect may be suitable. This is environment is suitable in the scenario where traffic does not need to have direct layer 2 connectivity between the respective data centers, such as when physical movement of server connectivity to a new data center is desired.

Leveraging Enterasys Fabric routing and host routing functionality virtual machines can easily be physically moved to a redundant datacenter with a higher path cost. After moving to the new location, the VM will be reachable via the same subnet, now in the redundant datacenter. This is made possible by the advertisement of the VM host route by the local fabric router in the new redundant data center location.

Conclusion

The emergence of technologies such as cloud computing and virtualization have forced organizations to take another look at how they design their data centers. In order to support the demanding availability requirements of today's applications, data centers need to go beyond the redundancy requirements of yesterday to a more future-proofed resilient infrastructure that will serve them well down the road. This requires organizations to support new technologies and standards, and also choose a solution that will provide an open and flexible enough architecture to support the evolving needs of the business. Enterasys delivers a simplified data center LAN that improves application performance and increases business agility, providing customers with a future-proofed approach to data center design best practices. To learn more, visit http://www.enterasys.com/solutions/DataCenter.aspx.

5608 - 0713