SRUTHI THIYAGARAJAN

Student ID: 00001632730

COEN 233 – COMPUTER NETWORKS

Winter 2022

DATA CENTER

Audience

This document provides an overview of Data Center by covering the Data Center Architecture ranging from the legacy to the latest trends, Infrastructure, and Virtualization. The Virtualization section discusses in detail the VXLAN technology and the deployment in Spine-Leaf Network Architecture of the Data Center.

The reader is expected to have a good understanding of the Network fundamentals and the routing and switching mechanisms.

This document can be used by network architects, designers, developers, academicians, researchers, technical authors and by any person seeking to know about the Data Center Networks.

Table of Contents

Table of Figures

Table of Tables

1

1. Introduction

Data Centers encloses a wide variety of technologies and are constantly evolving. Data Centers are complex systems that have critical computing resources in controlled environments and under centralized management. This enables enterprises to operate according to the business needs and all the time. They are widely deployed in the digital commerce and electronic communication sector and form the backbone infrastructure of the Internet.

The computing resources or the main components of the Data center include Servers, Storage, Network, Application, and operating system software. Different types of application are financial, human resources, e-commerce, and business-to-business. In addition, several servers support network operations, and network-based applications.

Before the emergence of cloud computing which uses data center, network resources, storage and computing were separated on individual physical machines. Systems and applications that interacted with such resources were kept physically directly accessible by the Information Technology (IT) administrators for security reasons. This started to change with the introduction of inexpensive resources in data center environments. At the time of data center emergence, they were used to host basic database servers, email servers and other types of application servers. Resource management at lower costs were provided by the DCs. A new communications paradigm shift started when the concept of hosting a complete virtual Operating System (OS) on an another in- compatible OS (i.e., running Linux on Windows OS) was materialized in software virtualization solutions. System virtualization became omnipresent in data center environments, where hundreds of virtual machines can be controlled by a single logical console. Hence, physical servers were turned into individual isolated virtual machines which can be cloned, paused, and destroyed on-demand.

2

2. Overview of Data Center

2.1 Data Center Architecture

Data center architecture and requirements can differ significantly based on the need of constructing it. For example, data center built for a cloud service provider such as Microsoft concentrates on the **facility, infrastructure, and security requirements** whereas a **private data center** such as for a government entity may be a one with a dedicated purpose of classifying data.

Regardless of the classification, it is optimal to achieve datacenter operation through a balanced investment in facility and the equipment. Guarding the system from intruders and cyberattacks is an essential one. The primary elements of the data center are as follows:

- **Facility**
    - Space available for IT equipment. Data centers are the world's most energy-consuming facilities as they provide round-the-clock access to information. Optimization of space and environmental control to keep equipment within specific temperature/humidity ranges are emphasized.

- **Core components**
    - Equipment and software for IT operations and storage of data and applications which includes storage systems; servers; network infrastructure, such as switches and routers; and various information security elements, such as firewalls.

- **Support Infrastructure**
    - Equipment contributing for secure sustenance for highest availability. Some of the equipment are:
        - Uninterruptible Power Sources (UPS) – battery banks, generators, and redundant power sources.
        - Environmental control – computer room air conditioners (CRAC); heating, ventilation, and air conditioning (HVAC) systems; and exhaust systems.
        - Physical security systems – biometrics and video surveillance systems.

- **Workforce/Operation Staff**

- Personnel to monitor operations and maintain IT and infrastructure equipment around the clock.

Figure 1 depicts the topology of the Redundant Enterprise Data Center Network.

The core connectivity functions supported by Data Centers are Internet Edge, campus, and server-farm connectivity.

Internet Edge Connectivity:

Internet edge provides connectivity between enterprise and Internet with redundancy and security functions which includes the following:

- Routing – (external and internal) using the IBGP and EBGP [Exterior Border Gateway Protocol and Interior Border Gateway Protocol]
- Connecting different service providers with redundant connection
- Edge security – to and from the Internet
- Access control to the Internet from the enterprise clients

Campus Core Switches:

The campus core switches provide connectivity between Internet Edge and the server farms(intranet) as in Figure-1. The core switches are physically connected to the devices that provide access to other major networks such as WAN edge routers, server farm aggregation switches and campus distribution switches.

Server Farm:

A centralized location to host applications are provided by the Data Center. The application architecture has changed significantly from master-slave method to client-server model. The client-server model in use today have undergone evolution resulting in n-tier application model where the functions are decoupled to several layers. The evolution includes the migration from **thick client** in client-server model, where it is the responsibility of the client to handle a major part of the processing and rendering the data which are mostly proprietary to each application, to the **thin client in n-tier model,** where major processing takes place in the server and use of HTML and HTTP makes the rendering process generic and simple. The server functions are performed by several computers or processes specialized operations.

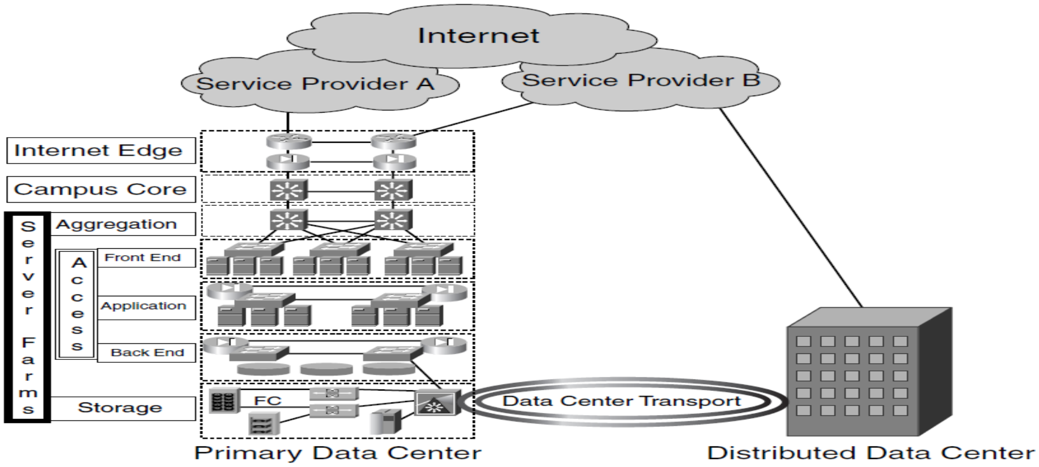


Figure-1 Topology of the Enterprise Data Center Architecture

Figure-2 represents the client-server application Interaction and Figure-3 represents the n-tier application interaction

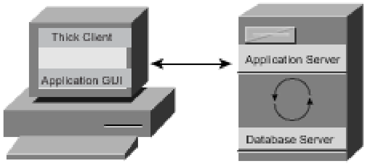


Figure-2 Client-Server Application Interaction

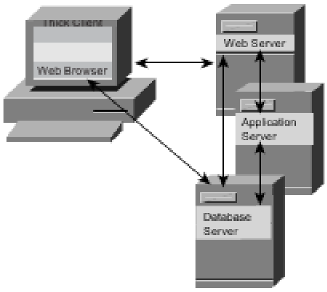


Figure-3 n-tier Application interaction

Networking is one of the important areas in Data center. It is the one which provides connectivity in and out of the data center. Below are the network layers of the server farm:

- Aggregation layer
- Access layer
  - Front-end segment
  - Application segment
  - Back-end segment
- Storage layer

- Data Center transport layer

Each of the layers is discussed in detail in Chapter 3.

2.2 Data Center Goals

The goals are:

- Traditional business-oriented goals like
  - Resiliency - support for business operations round the clock
  - Total Cost of Ownership - lowering the whole cost of operation and therefore the maintenance needed to sustain the business functions
  - Flexibility - rapid deployment of applications and consolidation of computing resources
- IT Initiatives:
  - Business maintenance and continuance
  - Increased security within the Data Center
  - Consolidation of Application, server, and Data Center
  - Integration of applications whether client/server and multitier (n-tier), or web services-related applications
  - Storage consolidation

The answer to the question “What is these IT initiatives and in what way are they different from the traditional business-oriented goals” is that they are a mix to address short-term and long-term problems. The long-term issue is addressed by establishing a strategic direction. All these require an architectural approach to design the Data Center Network flexible enough to accommodate future changes and avoid instability. The design criteria should include the following:

- Availability
- Scalability
- Security
- Performance
- Manageability

These criterions are logically separated and mapped to distinct functional areas of Data Center as explained below:

- **Infrastructure services:**
  - Routing, switching, and server-farm architecture
- **Application services:**
  - Load balancing, Secure Socket Layer (SSL) offloading, and caching
- **Security services:**
  - Packet filtering and inspection, intrusion detection, and intrusion prevention
- **Storage services:**

    - SAN architecture, Fiber Channel switching, backup, and archival
- **Business continuance:**
    - SAN extension, site selection, and Data Center interconnectivity

2.3 Role of Data Center in the Enterprise

Data Center are integral part of the enterprise supporting business applications and providing services such as:

- Data storage, management, backup, and recovery
- Productivity applications, such as email
- High-volume e-commerce transactions
- Powering online gaming communities
- Big data, machine learning and artificial intelligence

Every organization and government build and maintains its own data center or has access to other data center. The entity can have access to the data center either by renting servers at a colocation facility, use data center services managed by third-party, or use public cloud-based services from Amazon, Microsoft, Sony, and Google.

2.4 Geographical Distribution of Data Center

Data centers are distributed geographically for the subsequent reasons:

- To increase the reliability of hosted services
- To offer better cloud access performance to customers which in turn decreases the network distance between users and computing facilities.

Modular DC architectures are designed and evaluated to support this evolution. The conventional design goal is to permit building DCs incrementally starting by regular small basic building blocks, grouping some switches to interconnect several virtualization servers, using regular wiring schemes. As against legacy hierarchical architectures, modular DCs better accommodate horizontal traffic between virtualization servers in support of various IaaS operations like

VM migrations and storage synchronization.

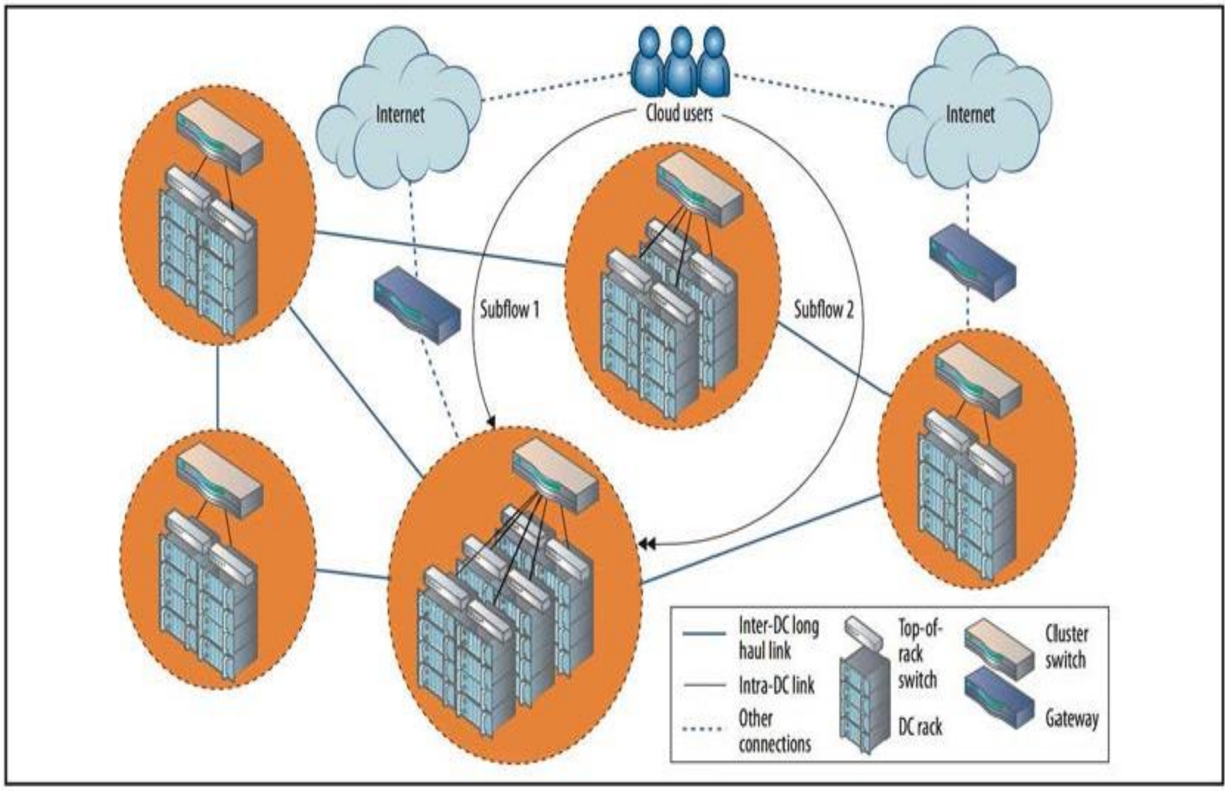


Figure-4 Geographical Distribution of Data Center

3

3. Infrastructure Design

3.1 Infrastructure Design

Infrastructure design is the core work as it encompasses every feature for the Data Center to function and serve as the foundation. The infrastructure features are listed as follows:

- Layer 2 - Switching
- Layer 3 - Routing

These features are the network layers of the server which are termed as follows:

- Aggregation layer
- Access layer
    - Front-end segment
    - Application segment
    - Back-end segment
- Storage layer
- Data Center transport layer

These layers form the basis to build highly available and scalable enterprise Data Centers. Many modifications and adaptions have been made to the need. Many new protocols, the way of assembling the aggregation and access layer have been introduced to reduce the wait time, to increase redundancy, and to support flexible business continuance.

Aggregation Layer:

Service to all server farms connected is provided by this layer and is the center point.
These devices which perform support services across all servers can be firewalls, load balancers, multilayer switches, and other devices. The multilayer switches perform the aggregation function and hence also known as aggregation switches. The service devices are shared amongst the servers. Some server farms span multiple access switches for redundancy, which in turn makes the aggregation switches, the logical point of connection for the services devices instead of the access layer switches.

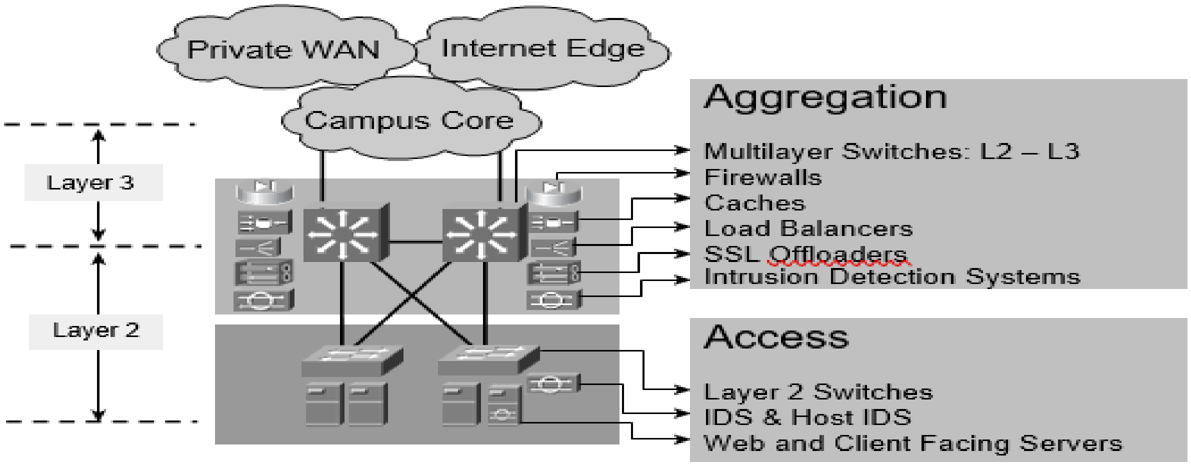


Figure-5 Aggregation and Access Layer Switches

Access Layer:

Layer 2 connectivity and services of the OSI model to the servers is provided by this layer. Additionally, the access layer is segregated into segments and in a multi-tier server farm, each server function can be located on different access switches on different segments.

Front-End Segment:

Layer 2 switches, security features in layer 2 and the front-end server farms comprise the front-end segment. The front-end server farms are with the following protocols:

- FTP
- Telnet
- SMTP
- Web Servers
- DNS
- Business Application Servers
- Network based application servers like IPTV and IP telephony – These are not placed at the aggregation/top layer because of port density.

The specific network features required at the front-end segment are determined by the server functionality. Servers using clustering mechanism for high availability or to communicate on the same subnet, layer 2 connectivity is required. Segregated server farms through load balancers or firewalls need connectivity through VLANs. This implies that layer adjacency can be provided by multiple access switches supporting front-end servers.

Security features has mechanisms such as

- Address Resolution Protocol (ARP) inspection
- broadcast suppression

- private VLANs and more to protect from L2 attacks.

Security devices are

- Network-based intrusion detection systems (IDSs) and
- Host based IDSs which monitors and detects intruders. Also prevents vulnerabilities from being exploited.

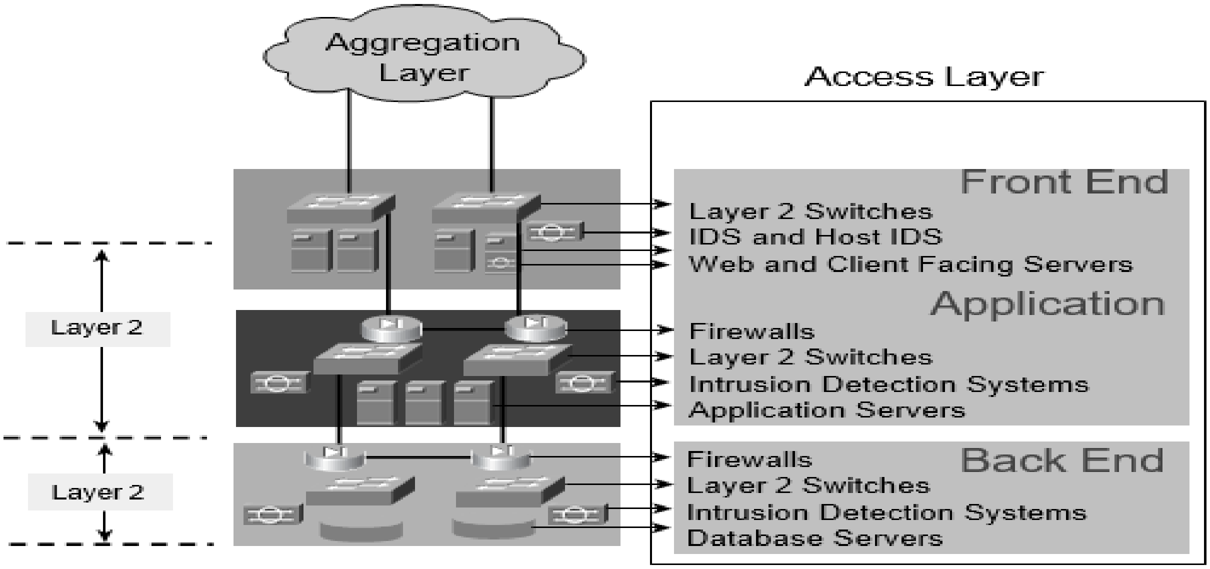


Figure-6 Access Layer Segment View

Application Segment:

This segment has application servers in addition to the network infrastructure elements as in front end segment. The feature required by this segment is the additional layer of security as the application servers directly communicate with the database systems. The security features can be of the form:

- Firewalls between web and application servers
- IDSs and Host IDSs

Application servers are referred to as the middleware or business logic as they have the responsibility of translating the user request from front end servers to commands, understandable by the database systems at the back end. Also, they run a part of the software used by business application. Scaling mechanisms include load balancers.

Back-End Segment:

The backend segment supports the database server connectivity and more security in addition to the functionalities mentioned in the above two segments. Medium sized servers to high end servers with hardware supporting database systems are used either with directly attached storage or disk arrays attached to a SAN. On the separation of storage, the database server is connected to both the Ethernet switch and the SAN.

Storage Layer:

It consists of the storage infrastructure such as Fiber Channel switches and routers that support small computer system interface over IP or Fiber channel over IP. Connectivity to servers, storage devices such as disk subsystems, and tape subsystems are provided by storage network devices. The network used by the storage devices is known as SAN. High-speed communication between servers and storage are required in DC environment. In order to provide high-speed environments, block-level access to the information supported by SAN technology is required. File-level access is needed for applications that use Network Attached Storage (NAS) technology.

Data Center Transport Layer:

Role of this layer:

- Communication between distributed DCs for rerouting traffic of client-server
- Communication between distributed server farms located in distributed DCs for the purpose of remote mirroring, replication, or clustering.

Distributed data center exists to increase availability and redundancy. The main reason for this distribution is the natural calamities and business continuance. There have been many trends in the data center to accommodate short term goals and long-term goals. Some of them are:

Blade Servers
This came into use to reduce the operational costs and do better in computing capacity at a relatively lower cost. This is achieved by using a different topology with ethernet switches inside the blade chassis, requiring proper planning of slot and port density, redundancy, connectivity, rack space, power consumption, heat dissipation, weight, and cabling.

Grid computing
Group of internetworked computers which together look to the outside world as a massive single computer(virtualized) to perform large tasks.

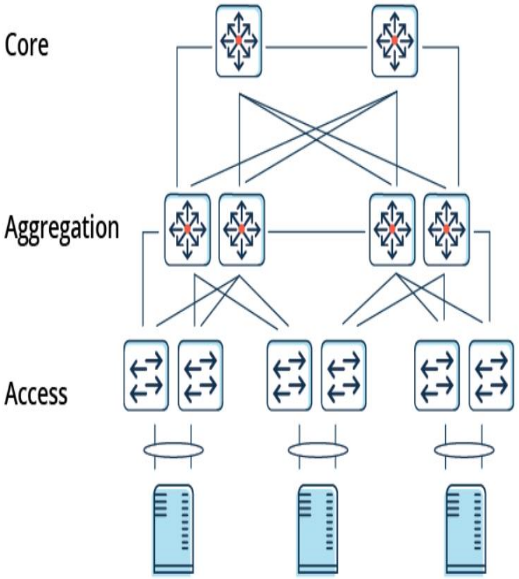
Web Services
Support for faster, easier, and secure access to the web-based applications. Firewalls is added to increase security. Web services refers to a secure environment for online processes from security and privacy perspective.

Service oriented Data Centers
Interoperability and manageability of the service devices are the priority in this service-oriented data centers. The interoperability comes from the use of standard interfaces and the support of trends like virtualization.

3.2 Spine-Leaf Network:

An application of the design feature seen in section 3.1 is the spine-leaf network topology which is very prevalent in the current data center deployments.

This network topology consists of two switching layers – a spine and leaf. Access switches that aggregate traffic from servers and connect directly into the spine or network core comprises the leaf layer.



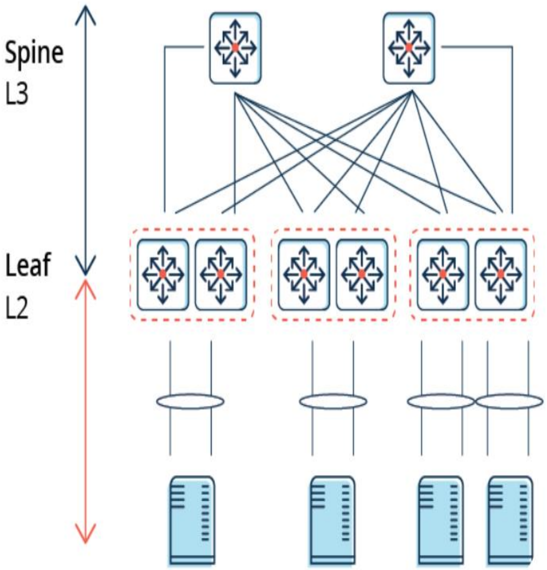


Figure-7 Traditional 3-tier architecture
Figure-8 Spine and Leaf Architecture

The traditional architecture consists of core routers, aggregation routers and access switches as depicted in Figure-7. Layer 2 Spanning tree protocol will be used between aggregation and access switches to build a loop free topology. The advantages of using STP are as follows:

Simple, easy to use, plug-and-play technology requiring little configuration.

Pod refers to the servers connected within each access switch. Pods use VLANs such that servers can move inside each pod without changing IP address and default gateway configurations. The limitation is it cannot support parallel forwarding paths (bandwidth restriction) and blockage of redundant paths in a VLAN.

Many proprietary protocols by different network organization were developed to overcome the limitation like Cisco developed virtual-port-channel(vpc) technology to overcome the limitation of STP. Vpc worked well in environment consisting of northbound and southbound traffic which is the traffic between client and servers. The limitation was it couldn't support east-west communication that is between the servers in different pods.

The interoperability and evolution lead to the extended L2 segments across all pods. This allowed more flexible resource pool that can be reallocated based on needs. Server virtualization which involves hosting of virtual machines can move freely across servers in different pods without the need to change the operating parameters. The applications are deployed in distributed fashion which led to an increase in east west traffic. There was limitation with 3 tier architecture, and proprietary protocols, bandwidth, and latency. Spine-leaf network came into existence to overcome these limitations.

In this spine-leaf network, every leaf switch is connected to the spine switch in a full mesh topology. The leaf switches are connected to the servers. The spine switches are the backbone of the network and is connected to every switch. Hence, the path can be chosen randomly, and the traffic load is distributed evenly among the spine. Fail over mechanism is supported as even if one of the spine fails, it would not affect the whole data center.

Additional spine can be added to the existing topology easily and can be connected to every leaf which results in an increase of interlayer bandwidth supporting more traffic which in terms is the Salability. Additional leaf switch can be added with adding same network configuration to address device port capacity.

Latency is predictable in this architecture, as the payload only has to hop between a spine and leaf to reach the destination.

Chapter 4 discusses virtualization and overlay networks.

3.3 Latest Infrastructure Models

Converged Infrastructure

Converged infrastructure (CI) is a form of datacenter management that combines legacy infrastructure components like storage arrays, servers, network switches, and virtualization onto a single unit that makes purchasing and deployment easier and more predictable.

Systems are designed and integrated by a vendor, packaged into a distinct set of pre-configured options with this CI. Converged infrastructure combines pre-integrated hardware components with software to orchestrate and provision these resources through a unified system instead of buying the components separately and working through compatibility and integration challenges manually.

CI aims to reduce the complexity in the datacenter management that is prevalent in the legacy, multi-tiered infrastructure. Reduction in hardware incompatibility issues and ease of deployment is the key of its design as it reduces time and resources utilized in integrating and deploying datacenter infrastructure.

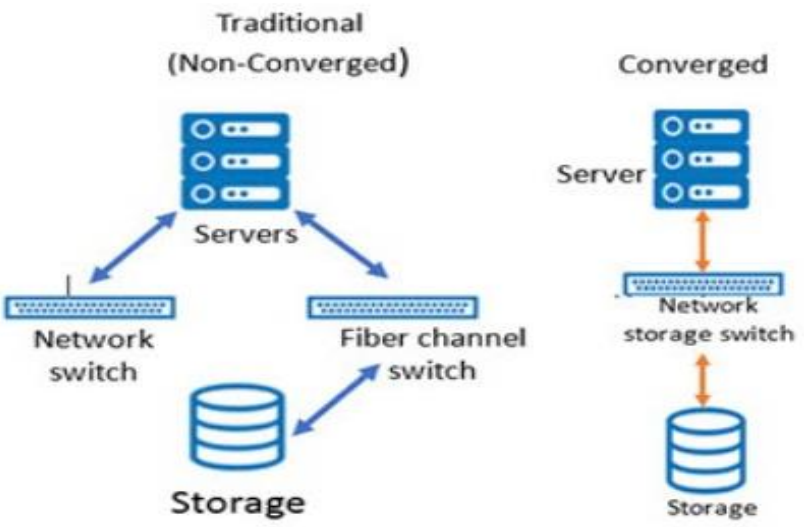


Figure-9 Traditional vs Converged infrastructure Interaction

Hyper-Converged Infrastructure

Hyperconverged infrastructure (HCI) incorporates intelligent distributed software to combine resources pool of storage and server into a complete software-defined solution. It builds one unified distributed system, making a highly scalable datacenter replacing legacy infrastructure components like separate servers, storage networks, and storage arrays.

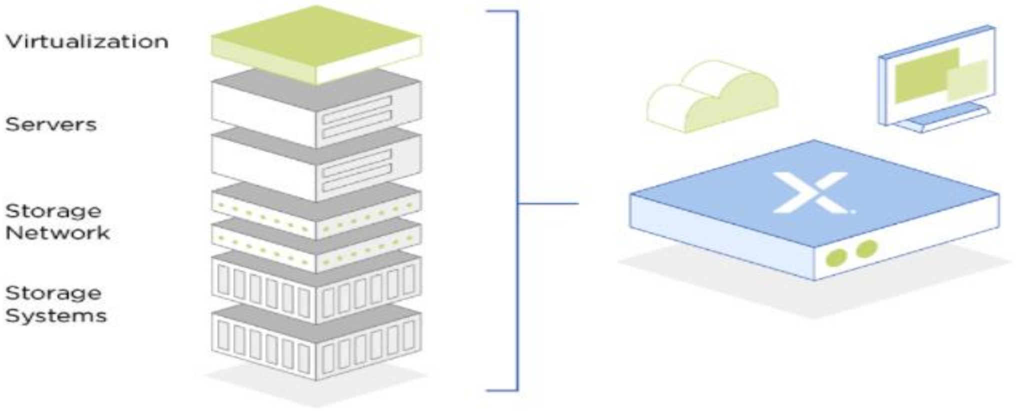


Figure-10 Hyperconverged Infrastructure

4

4. Virtualization and Overlay Network

4.1 Virtualization

Separating process of the software layer of a computer or server or device from the hardware layer of a computer or sever or device. A new layer is placed between the two as a mediator called the hypervisor. Virtualization can also be defined as a process for users to create multiple simulated environments within one piece of hardware. Virtualization uses software which acts as a hardware.

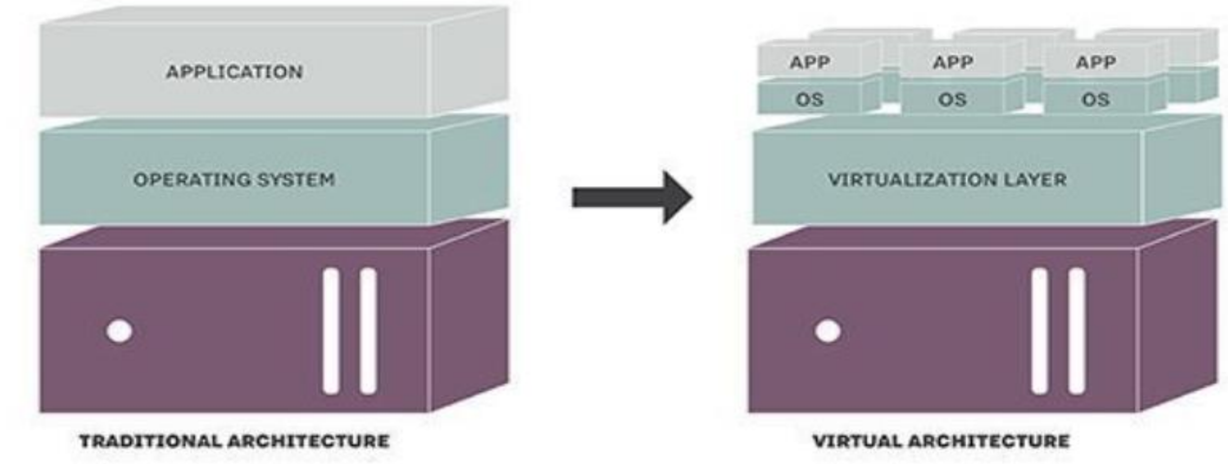


Figure-11 Virtualization conceptual view

The hypervisor is the software that isolates the hardware and software components of a system. The device in which the hardware resources are virtualized is called the **host**. Host hosts multiple **virtual machines**, which is a compute resource that uses software, in this case the hypervisor instead of a physical computer to run programs or deploy apps.

Multiple VMs which are guests share the system's physical compute resources such as bandwidth, memory space, processor cycles etc.

Type 1 Hypervisor / Bare-metal Hypervisor:

This software runs directly on top of the host system hardware. Pros are better performance, scalability, and stability as they have direct access to system hardware which implies high availability and resource management.

Examples of Type 1 Hypervisor: Citrix Xen Server, VMWare ESXi, Microsoft Hyper-v.

Type 2 Hypervisor / Hosted Hypervisor:

This software is placed over the host operating system. Known host OS provides the ease of system configuration and management tasks but at the same time with the limitation of performance and OS security flaws.

Examples of Type 2 Hypervisor: Oracle VM Virtual Box, VMWare Workstation

4.2 Data Center Virtualization:

Design, development, and implementation of data centers using virtual hardware is data center virtualization. Many areas of virtualization are possible in Data Center:

Server Virtualization:

Multiple operating systems run autonomously of each other on a single machine in the data center. Virtualization software called Hypervisors encapsulates a guest version of the operating system and emulates hardware resources and thereby reduces the need for physical hardware systems. Increase in resource utilization lowers server costs. Organization in which the use of only one operating system is required, a new technology called containerization, part of docker is used. Foregoing a hypervisor and sharing a single instance of the operating system and running on the "bare metal" of the server makes containerization more efficient than server virtualization.

IT and O&M costs are reduced to a great extent and improvement in service deployment flexibility is achieved.

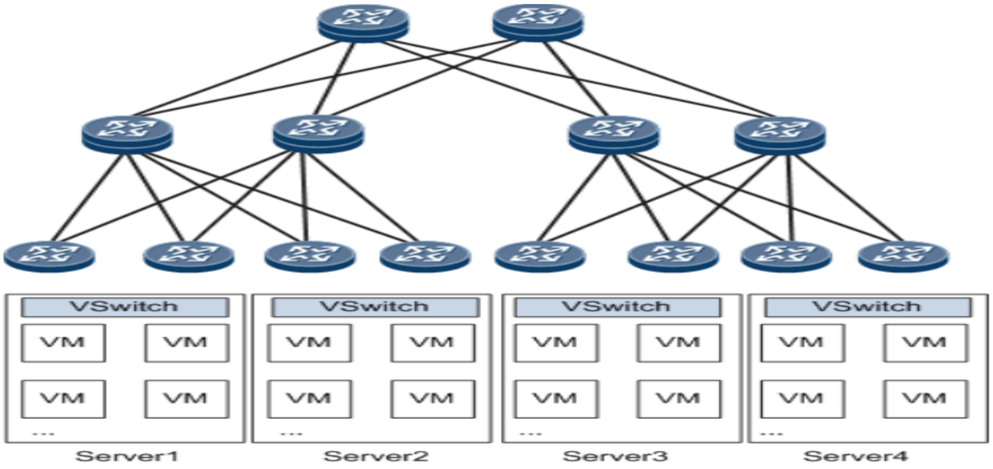


Figure-12 Server Virtualization

A massive increase in the number of hosts leads to the following problems on virtualizing a server:

VM scale is bounded by the network specification.

- On a Layer 2 network, data packets are forwarded at Layer 2 based on MAC entries. But MAC table capacity is limited. This limits the number of VMs.

Network isolation capabilities are limited.

- Network isolation is done using VLANs currently. On large-scale virtualized networks, deployment of VLANs has the below issues:
    - The VLAN tag field defined in IEEE 802.1Q has only 12 bits and can support only a maximum of 4094 VLANs, and hence cannot support user identification requirements of large Layer 2 networks. Dynamic network adjustment adaptability cannot be done.

VM migration scope is bounded by the network architecture.

- After a VM is started, when it faces resource shortage, it needs to be moved to a new server, for example, when the CPU usage is too high, or memory resources are inadequate. The IP address of the VM must remain unchanged for a smooth VM migration which requires the service network to be a Layer 2 network.
- Multipath redundancy, backup and reliability are also required.

VXLAN addresses the above problems on large layer 2 networks as it is a network virtualization technique.

Application Virtualization:

Application Virtualization Software provides access to remote applications. Apps are hosted in remote servers and delivered to an end-user giving performance as if they are on the user’s physical device. Client’s computers have many apps running in contained environments. Virtualization makes applications run through different operating systems and browsers without any dependencies.

Network Virtualization:

Network Virtualization Software (NV) provides a solution for moving VMs across different logical domains.

Rather than providing a connection between them in a network, NV connects the two domains by creating a tunnel through the existing network. Through this network administrators can move/migrate VMs independently of the existing infrastructure without having to reconfigure the network. Agility and optimization are obtained through network virtualization. Faster provisioning and the automation of manual processes can be done with network virtualization.

There are two methods to virtualize a whole data center. One is virtualizing switches, and another is to use VXLAN technology.

A VLAN is when a LAN is virtualized and broken into multiple distinct subnets for different types of traffic that have distinct requirements. A VXLAN takes the VLAN methodology, but the traffic goes through the layer 3 routing architecture and uses the layer 2 data packet addresses to deliver the content. This is type of network is called Overlay network. Next heading discusses on the overlay network and VXLAN protocol in detail.

4.3 Overlay Network

Network overlays are networks of interconnected nodes that are virtual, and which shares an underlying physical network. Deployment of applications requiring specific network topologies is done without modifying the underlying network. Overlay networks consist of a series of virtual or physical computers spread as a layer on top of an existing network. The purpose/aim of the overlay network is to add missing functionality without a complete network redesign. These networks communicate to the existing network through virtual or physical nodes.

New encapsulation frame formats have been specifically built for data center supporting overlay networks. These include Virtual Extensible LAN(VXLAN), Network Virtualization using Generic Routing Encapsulation (NVGRE), Transparent Interconnection of Lots of Links (TRILL), and Location Identifier Separation Protocol (LISP). Of all, VXLAN is widely used because of the incredible advantages.

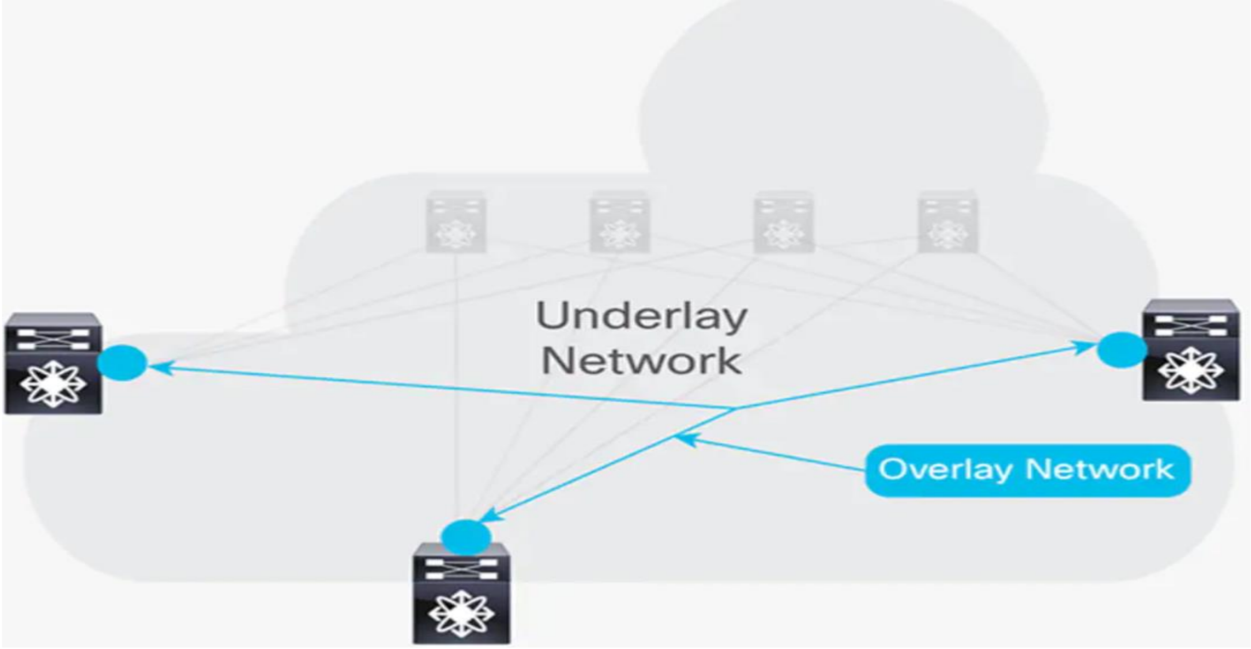


Figure-13 Underlay Overlay Network

4.4 VXLAN Protocol and deployment in Spine-leaf network

4.4.1 VXLAN Overview

VXLAN is a network virtualization methodology that uses MAC-in-UDP encapsulation that adds a UDP header and a VXLAN header before a raw Ethernet packet.

It is an industry standard Overlay protocol (RFC 7348). It extends Layer 2 segments over a Layer 3 infrastructure to build Layer 2 overlay logical networks. Encapsulation of Ethernet frames into IP User Data Protocol (UDP) headers and transportation of the encapsulated packets through the underlay network to the remote VXLAN tunnel endpoints (VTEPs) using the conventional IP routing and forwarding mechanism takes place. These **VTEPs** are often **virtual bridges** in a hypervisor, VXLAN-aware applications within a virtual machine (VM), or within the hardware of a switch. VTEPs, called VXLAN gateways are key to virtualizing networks across the existing data center infrastructure which takes care of encapsulation and decapsulation of VXLAN headers within the packets.

VXLAN protocols requires the below protocols for its functioning:

- Multicast support: IGMP and PIM
- Layer 3 routing protocol: OSPF, BGP

Although some networking organization has come up with network OS which will support VXLAN without multicast. Example: Cisco

The UDP-IP Packet is transported across an IP network after the encapsulation of original layer 2 frame with VXLAN header. Each VXLAN network segment has a unique **24bit VXLAN Network Identifier**, or VNI. The 24-bit address provides scaling of virtual networks 16.7 million compared to 4096 supported by VLAN. Multicast and network hardware restrictions reduces the useable number of virtual networks in most deployments. VMs in an exceedingly logical L2 domain use the identical subnet and are mapped to a common VNI. It's the L2 to VNI mapping that makes VMs talk with one another. Layer 3 addressing schemes is not changed by the virtual networks. IP addressing rules apply to the virtual networks too.

VXLANs make VM identification unique by combining the VM's MAC address and its VNI. Duplicate MAC addresses are allowed in a datacenter domain because of this with a limitation that duplicate MACs cannot exist on the same VNI. Virtual machines on a VNI subnet do not require any special configuration to support VXLAN because the encapsulation/decapsulation and VNI mapping are managed by the VTEP built into the hypervisor.

VXLAN Tunnel Endpoint:

VXLAN uses VXLAN tunnel endpoint (VTEP) devices to map tenants' end devices to VXLAN segments. Each VTEP function has two interfaces:

- Switch interface on the local LAN segment to support endpoint communication through bridging.
- IP interface to transport to the IP network.

The IP interface has a unique IP address that identifies the VTEP device on the IP network. The VTEP device uses this IP address to encapsulate Ethernet frames and transmits the encapsulated packets to the transport network through the IP interface. A VTEP device also discovers the remote VTEPs for its VXLAN segments and learns remote MAC Address-to-VTEP mappings through its IP interface.

Configuration Requirements:

- The VTEP must be configured with the layer-2 or IP subnet to VNI network mappings.
- VNI to IP multicast mapping. Each VTEP device is independently configured with this multicast group and participates in PIM routing. Incases, where the underlay network is required to be multicast free, VXLAN VTEP can use a list of IP addresses of other VTEPs in the network to send broadcast and unknown unicast traffic.

Outcome/Reason:

- VTEP to VNI mapping builds forwarding tables for VNI/MAC traffic flows.
- VNI to IP multicast mapping allows VTEPs to emulate broadcast/multicast functions across the overlay network.
- Synchronization of VTEP configurations can be automated with common configuration management tools like RANCID, or can be managed through VMware's vCenter Orchestrator, Open vSwitch or other systems.

The VXLAN segments are independent of the underlying network topology; conversely, the underlying IP network between VTEPs is independent of the VXLAN overlay. It routes the encapsulated packets based on the outer IP address header, which has the initiating VTEP as the source IP address and the terminating VTEP as the destination IP address.

4.4.2 VXLAN Header

Figure-14 depicts the VXLAN packet format details. Table-1 describes the packet in detail.

FIELD	DESCRIPTION
VXLAN header	○ VXLAN Flags (8 bits): The value is 00001000. ○ VNI (24 bits): VXLAN network identifier used to identify a VXLAN segment. ○ Reserved fields (24 bits and 8 bits): must be set to 0.
Outer UDP header	○ DestPort: destination port number, which is 4789 for UDP.

	o Source Port: source port number, which is calculated by performing the hash operation on inner Ethernet frame headers.
Outer IP header	o Protocol: This is set to 0 × 11 to indicate it's a UDP packet. o IP SA: source IP address, which is the IP address of the local VTEP of a VXLAN tunnel. o IP DA: destination IP address, which is the IP address of the remote VTEP of a VXLAN tunnel. If unknown/unlearned or is a broad/multi-cast address, then VXLAN simulates a network broadcast using its multicast group.
Outer Ethernet header	o MAC DA: destination MAC address, which is the MAC address mapped to the next hop IP address based on the destination VTEP address in the routing table of the VTEP on which the VM that sends packets resides. o MAC SA: source MAC address, which is the MAC address of the VTEP on which the VM that sends packet resides. o 802.1Q Tag: VLAN tag carried in packets. This field is optional. o Ethernet Type: Ethernet frame type.

Table-1 VXLAN Packet Description

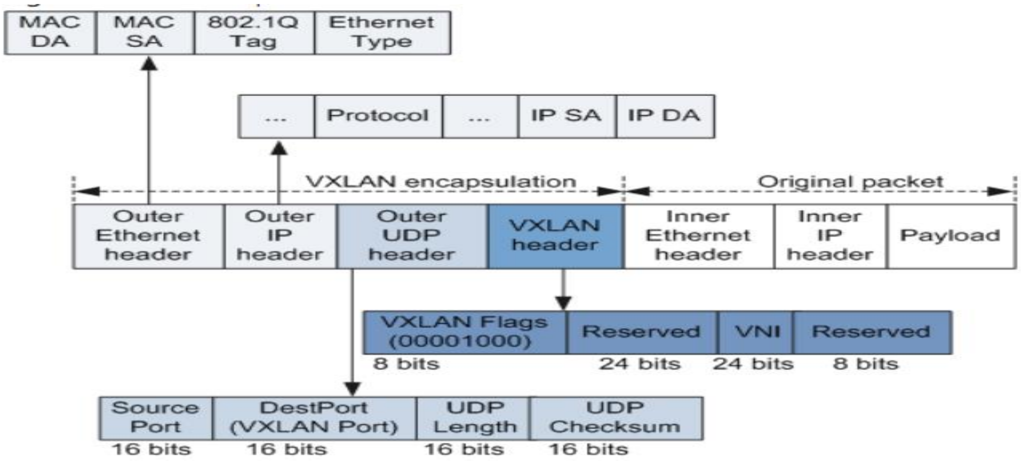


Figure-14 VXLAN Packet format details

Let's see into the details of Outer IP header.

- Destination IP is overridden by the IP multicast group corresponding to the VNI of the source virtual machine.
- Frame is multicast and all VTEPs on the VNI multicast group receive the frame. They decapsulate the frame, learn the source ID and VNI mapping for future use and based on the frame type and local forwarding table information forward or drop the packet.
- The VTEP hosting the target virtual machine will encapsulate and forward the virtual machines reply to the sourcing VTEP.

- The source VTEP receives the response and caches the ID and VNI mapping for future use.

4.4.3 VXLAN deployment in Spine-leaf Network

Figure-15 and Figure-16 shows the underlay and overlay network in a separate view.

Underlay network:

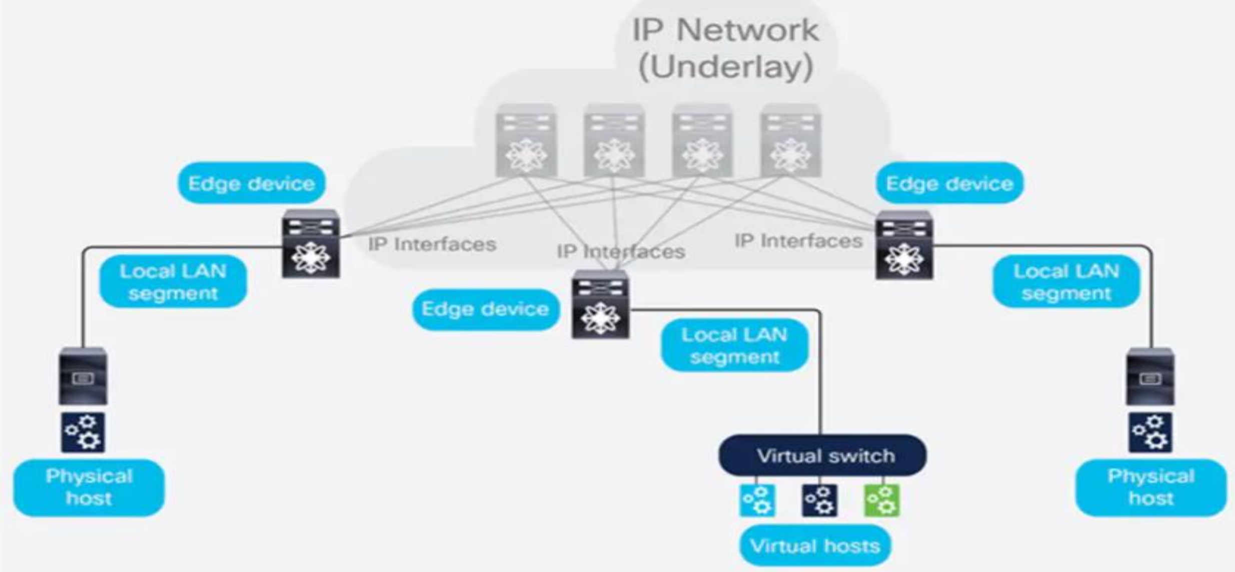


Figure-15 Underlay IP network

This is a normal IP network in a spine and leaf topology.

Overlay Network:

The most common and efficient deployment of VXLAN is with a control plane as they help in distribution of end-host reachability information among the VTEPs as end-host information learning and VTEP discovery both are data-plane based. Most used control plane protocol is the BGP-EVPN. This technology provides control-plane and data-plane separation and a unified control plane for both Layer 2 and Layer 3 forwarding in a VXLAN overlay network.

Deployment without control plane mechanism, like flood-learn suffers from the flooding challenges as the number of hosts increases in the broadcast domain.

The overlay network uses the BGP EVPN for the control plane for the VXLAN overlay network.

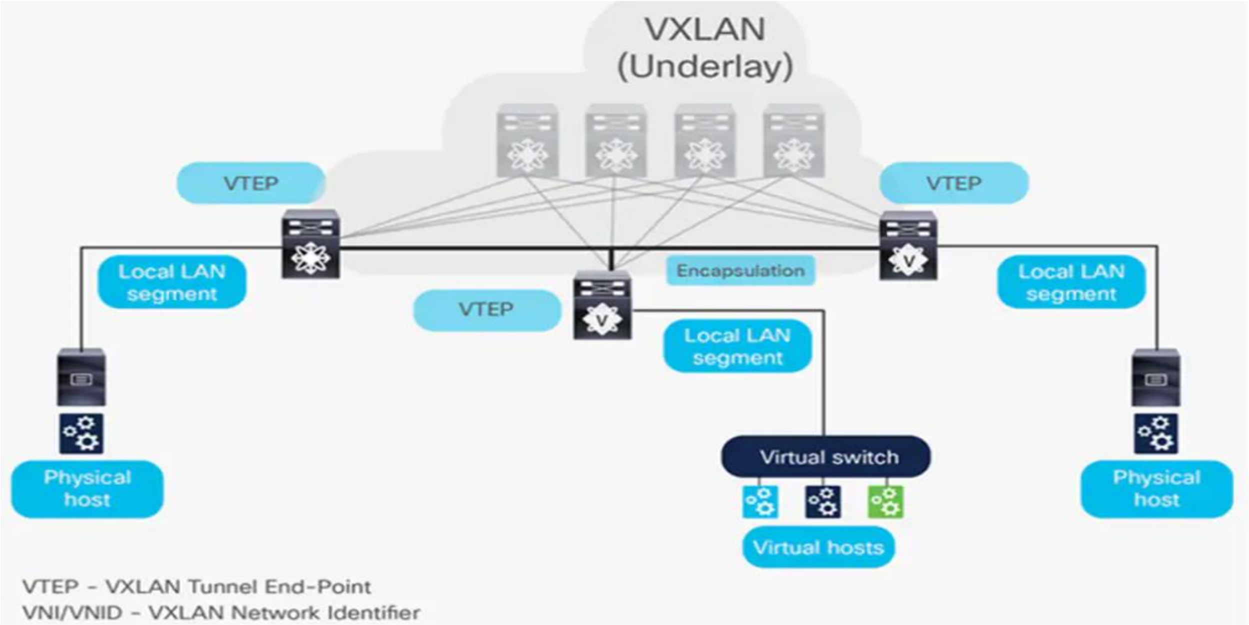


Figure-16 VXLAN overlay network

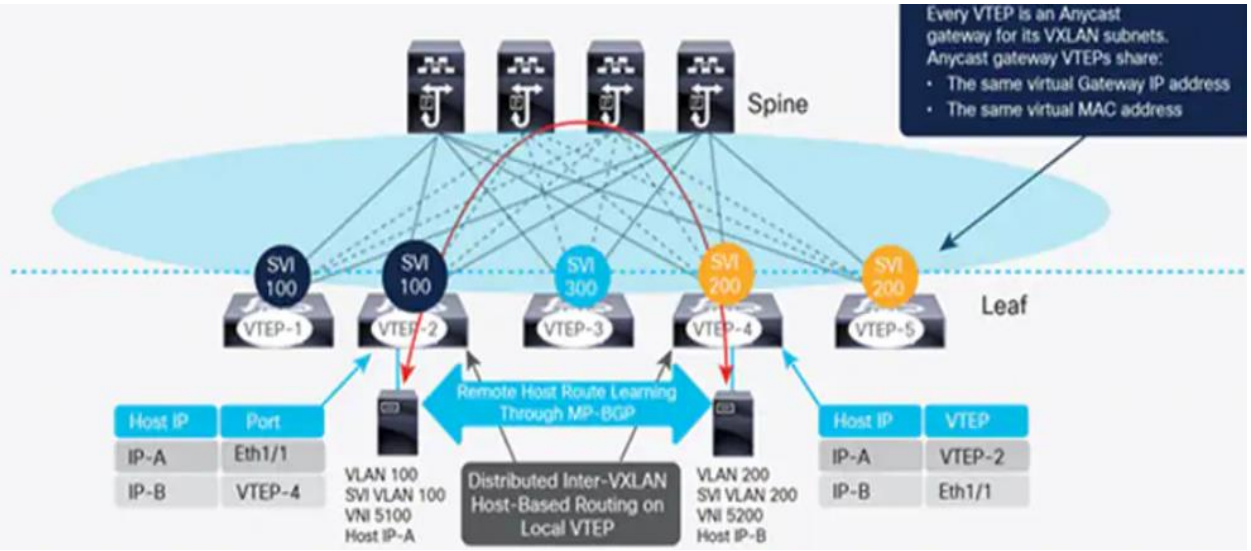
Layer 3 Routing Function:

Routing between VXLAN segments or from a VXLAN segment to a VLAN segment will be required in many situations. In the VXLAN spine-and-leaf network design, the leaf Top-of-Rack (ToR) switches are enabled as VTEP devices to extend the Layer 2 segments between racks. These VTEPs are Layer 2 VXLAN gateways for VXLAN-to-VLAN or VLAN-to-VXLAN bridging.

The layer 3 VXLAN routing function needs to be enabled on the VTEPs for routing of traffic between VXLAN segments or between VXLAN-VLAN segment. The VXLAN BGP EVPN spine-and-leaf network provides Layer 3 internal VXLAN routing as well maintain connectivity with the external networks like the campus network, WAN, and Internet. VXLAN BGP EVPN uses distributed anycast gateways for internal routed traffic. The common designs used are internal and external routing on the spine layer, and internal and external routing on the leaf layer.

Internal Routing:

Uses the distributed anycast gateway mechanism. Any VTEP in a VNI can be the distributed anycast gateway for end hosts in its IP subnet by supporting the same virtual gateway IP address and the virtual gateway MAC address (shown in Figure 17). With the anycast gateway function in EVPN, end hosts in a VNI can always use their local VTEPs for this VNI as their default gateway to send traffic out of their IP subnet. This enables optimal forwarding for northbound traffic from end hosts in the VXLAN overlay network. A distributed anycast gateway also offers the benefit of transparent host mobility in the VXLAN overlay network. Because the gateway IP address and virtual MAC address are identically provisioned on all VTEPs in a VNI, when an end host moves from one VTEP to another VTEP, it doesn’t need to send another ARP request to relearn the gateway MAC address.



Every VTEP is an Anycast gateway for its VXLAN subnets.
Anycast gateway VTEPs share:

- The same virtual Gateway IP address
- The same virtual MAC address

Host IP	Port
IP-A	Eth1/1
IP-B	VTEP-4

Host IP	VTEP
IP-A	VTEP-2
IP-B	Eth1/1

Figure-17 Internal Routing – Distributed anycast gateway

External Routing:

At the Border Leaf:

Figure 18 shows a pair of border leaf switches connected to outside routing devices. The border leaf switch runs BGP EVPN on the inside with the other VTEPs in the VXLAN fabric and exchanges EVPN routes with them. At the same time, it runs the normal IPv4 or IPv6 unicast routing in the tenant VRF instances or the router with the external routing device on the outside. The routing protocol can be regular eBGP or any routing protocol. The border leaf switch learns external routes and advertises them to the EVPN domain as EVPN routes so that other VTEP leaf nodes can also learn about the external routes for sending outbound traffic.

The border leaf switch can also be configured to send EVPN routes learned in the Layer 2 EVPN address family to the IPv4 or IPv6 unicast address family and advertise them to the external routing device. By this design, user/external traffic needs to take two underlay hops (VTEP to spine to border leaf) to reach the external network.

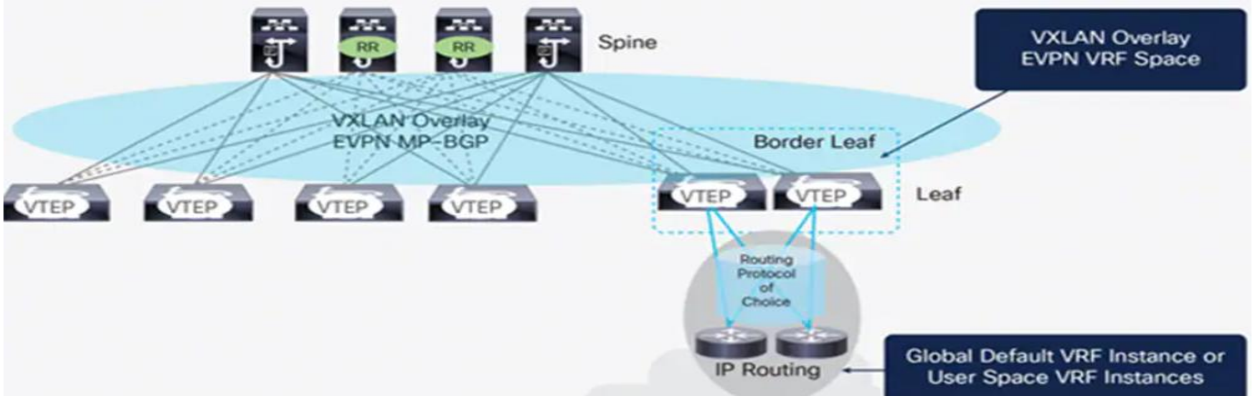


Figure-18 External routing at the border leaf

Hence, the spine switch only needs to run the BGP-EVPN control plane and IP routing; it does not need to support the VXLAN VTEP function.

<u>At the border Spine:</u>

Figure 19 shows a pair of spine switches connected to the outside routing devices. In this, the spine switch needs to support VXLAN routing. The spine switch runs BGP EVPN on the inside with the other VTEPs in the VXLAN fabric and exchanges EVPN routes with them. At the same time, it runs the normal IPv4 or IPv6 unicast routing in the tenant VRF instances/router with the external routing device on the outside. The routing protocol can be regular eBGP or any external routing protocol. The spine switch learns external routes and advertises them to the EVPN domain as EVPN routes so that other VTEP leaf nodes can also learn about the external routes for sending outbound traffic.

The spine switch can also be configured to send EVPN routes learned in the Layer 2 EVPN address family to the IPv4 or IPv6 unicast address family and advertise them to the external routing device. By this design, tenant/user traffic needs to take only one underlay hop (VTEP to spine) to reach the external network. But the spine switch needs to run the BGP-EVPN control plane and IP routing and the VXLAN VTEP function.

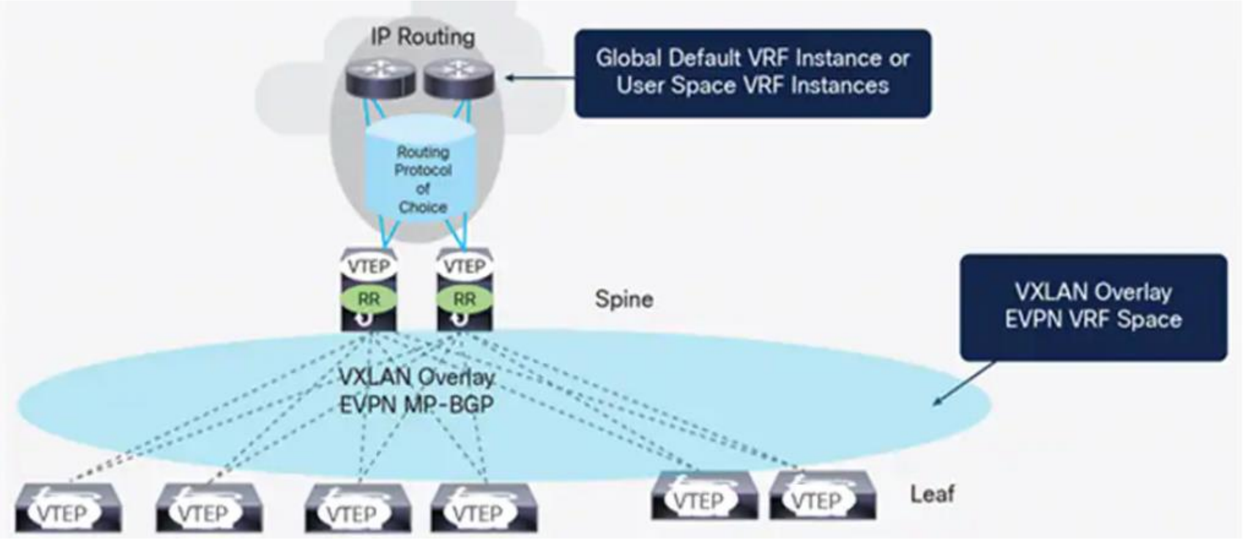

IP Routing
Global Default VRF Instance or User Space VRF Instances
Routing Protocol of Choice
VTEP
RR
VTEP
RR
Spine
VXLAN Overlay EVPN VRF Space
VXLAN Overlay EVPN MP-BGP
VTEP
VTEP
VTEP
VTEP
VTEP
VTEP
Leaf

Figure-19 External Routing at the Spine

Multitenancy:

In BGP EVPN, multiple tenants can co-exist and share a common IP transport network while having their own separate VPNs in the VXLAN overlay network (Figure 20). In the VXLAN BGP EVPN spine-and-leaf network, VNIs define the Layer 2 domains and enforce Layer 2 segmentation by not allowing Layer 2 traffic to traverse VNI boundaries. Likewise, Layer 3 segmentation is obtained by Layer 3 VRF technology and by using a separate Layer 3 VNI mapped to each VRF instance. Each tenant has its own VRF routing instance. IP subnets of the VNIs for a tenant are in the same Layer 3 VRF instance.

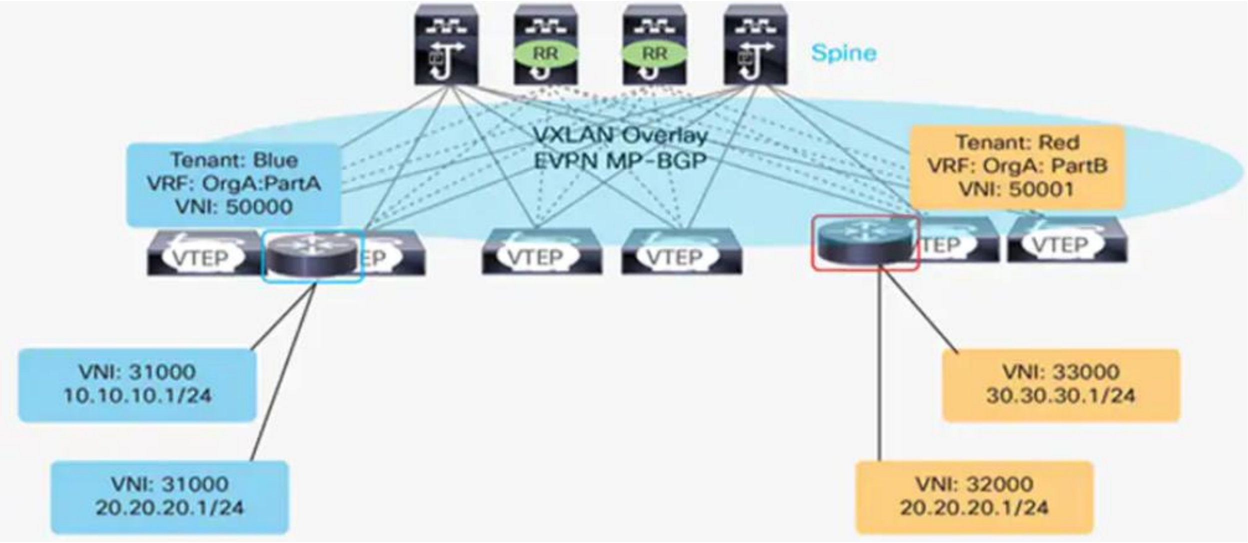


Figure-20 Multitenancy

4.4.4 VXLAN Advantage and Use cases

VXLAN's layer 2 tunneling feature overcomes IP subnetting restrictions which permits administrators to move VMS to any server in the data center, regardless of the data center's subnetting scheme. Supporting VM mobility across all the servers in the data center with reliable L3 is achievable. Application Examples:

- Hosting provider provisioning a cloud for its customer.
- VM Farm that has outgrown its IP address space but wants to preserve the data center network architecture.
- Cloud service provider whose multi-tenant offering needs to scale beyond 802.1q VLANS.
- VXLANs in enterprise data centers include preventing data collision, better control traffic rules, increasing LAN segments as more workloads are added.
- Security – Network isolation restricts the hacker's area or movement in that data center. Virtual machines on different networks work the same way if they are present on the same layer-2 subnet.

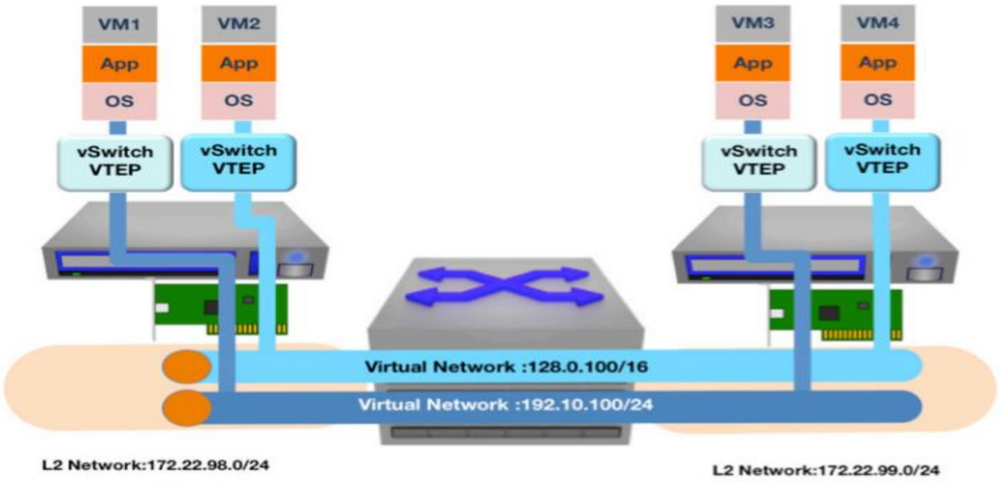

VM1
App
OS
VM2
App
OS
VM3
App
OS
VM4
App
OS
vSwitch VTEP
vSwitch VTEP
vSwitch VTEP
vSwitch VTEP
Virtual Network :128.0.100/16
Virtual Network :192.10.100/24
L2 Network:172.22.98.0/24
L2 Network:172.22.99.0/24

Figure-21 VMs on different L2 domain connected by VXLAN

5

5 Future Aspects

The future of the Data Center is smart: Converged Infrastructure and Hyper converged infrastructure is the future of the modern Data Centers reducing the O&M cost.

Hyper-converged Data Centers are software-defined Data Centers (SDDC) which are also known as smart Data Centers. Virtualization and convergence of all operational layers like computing, networking, and storage into a single box is done. With HCI, same server has every data center component giving improved efficiencies, reduced costs, and increased control.

Colocation: A collocation is a place where customers have better control over their private infrastructure. Colocation services (or Colo) are delivered by Data Center solution providers, which is the consolidation of network, storage, and server to enhance user experience. Modern colocation services are moving to Data Center-as-a-Service (DCaaS) as it provides a greater flexible deployment than Software as a Service, Platform as a Service, and Infrastructure as a Service models. A hybrid DCaaS colocation architecture consists of public IaaS platform, on hosted or on-premises private cloud and a Wide Area Network (WAN) to connect the two. A major advantage of DCaaS is greater reduction in cost.

Acronyms

OS	Operating System
IT	Information Technology
UPS	Uninterrupted Power Supply
CRAC	Computer Room Air Conditioners
HVAC	Heating Ventilation and Air Conditioning
IBGP	Interior Border Gateway Protocol
EBGP	Exterior Border Gateway Protocol
WAN	Wide Area Network
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
SSL	Secure Socket Layer
DC	Data Center
IaaS	Infrastructure as a Service
VLAN	Virtual Local Area Network
FTP	File Transfer Protocol
SMTP	Simple Mail Transfer Protocol
DNS	Domain Name System
IP	Internet Protocol
ARP	Address Resolution Protocol
L2	Layer 2
L3	Layer 3
IDS	Intrusion Detection Systems
SAN	Storage Area Network
NAS	Network Attached Storage
STP	Spanning Tree Protocol
vpc	Virtual Port Channel

CI	Converged Infrastructure
HCI	Hyper Converged Infrastructure
VM	Virtual Machine
MAC	Media Access Control
CPU	Control Processing Unit
VXLAN	Virtual Extensible Local Area Network
NV	Network Virtualization
NVGRE	Network Virtualization using Generic Routing Encapsulation
TRILL	Transparent Interconnection of Lots of Links
LISP	Location Identifier separation Protocol
UDP	User Datagram Protocol
VTEP	Virtual Tunnel End Point
IGMP	Internet Group Management Protocol
PIM	Protocol Independent Multicast
OSPF	Open Shortest Path First
BGP	Border Gateway Protocol
VNI	VXLAN Network Identifier
EVPN	Ethernet Virtual Private Network
VRF	Virtual Routing and Forwarding
SDDC	Software Defined Data Center
Colo	Colocation
DCaaS	Data Center as a Service

References

i. VXLAN RFC - https://datatracker.ietf.org/doc/html/rfc7348

ii. M. Chowdhury and R. Boutaba, “A Survey of Network Virtualization,” Computer Networks.

iii. Junji Kinoshita, Kazuhiro Maeda, Hitoshi Yabusaki, Ken Akune, Norihisa Komoda “Realization of VXLAN Gateway-Based Data Center Network Virtualization”, IEEE Conference

iv. DC Fundamentals - http://estigia.fi-b.unam.mx/maestria/Cisco%20Press%20-%20Data%20Center%20Fundamentals.pdf

v. https://www.arista.com/en/solutions/cloud-networking

vi. Waleed S. Alnumay, Uttam Ghosh “A network virtualization framework for information centric data center networks”, IEEE Conference

vii. Md. Faizul Bari, Raouf Boutaba;Rafael Esteves, Lisandro Zambenedetti Granville, Maxim Podlesny, Md Golam Rabbani, Qi Zhang, Mohamed Faten Zhani “Data Center Network Virtualization: A Survey” , IEEE Journal

viii. https://support.huawei.com/enterprise/en/doc/EDOC1100142674/65029582/

ix. https://www.sdxcentral.com/data-center/definitions/enterprise-data-center-networking/

x. https://www.nutanix.com/info

xi. Edison F. Naranjo, Gustavo D. Salazar Ch “Underlay and overlay networks: The approach to solve addressing and segmentation problems in the new networking era: VXLAN encapsulation with Cisco and open-source networks”, IEEE Conference

xii. Talvinder Singh, Varun Jain;G Satish Babu “VXLAN and EVPN for data center network transformation”, IEEE Conference

xiii. Naoki Oguchi, Motoyoshi Sekiya “Virtual data planes for easy creation and operation of end-to-end virtual networks”, IEEE Conference