jamk.fi

Overlay Technologies and Microsegmentation in Data Centers

Masi Takamäki

Master's Thesis
April 2018
School of Technology, Communication and Transport
Master's Degree Programme in Information Technology
Cyber Security

jamk.fi

Description

Author(s)	Type of publication	Date
Takamäki, Masi	Master's Thesis	April 2018
		Language of publication: English
	Number of pages 69	Permission for web publication: x

Title of publication

Overlay Technologies and Microsegmentation in Data Centers

Degree Programme

Master´s Degree Programme in Information Technology, Cyber Security

Supervisor(s)

Häkkinen Antti, Huotari Jouni

Assigned by

Rivinoja Jarkko, Cygate Oy

Abstract

The massive growth of data, availability and security requirements sets new challenges for organizations running a data center business nowadays. Traditional data center designs are not able meet these challenges anymore. Layer 2 is generally seen as a failure domain; nevertheless, there are still many demands for Layer 2 connectivity between the data centers. In addition, advanced cyber security attacks against organizations take place every day with a frightening success rate.

A lab environment in Cygate's premises was created for testing modern overlay technologies and microsegmentation. A software based assessment tool with the capability to simulate real application traffic and ability to easily deliver key performance metrics was used to verify the functionalities used in the lab environment.

The research was based on qualitative research method. The research strategy was an experimental research, which enabled exploring the influence and interaction of a certain phenomenon in a controlled environment.

The thesis resulted in an understanding of how modern overlay technologies function, how microsegmentation should be implemented and what the benefits for these two are.

Organizations running a data center business should seriously consider abandoning legacy Ethernet Fabrics and consider migrating to Leaf and Spine based IP Fabrics. If Layer 2 connectivity is still required, an overlay technology such as VXLAN should be implemented. Additionally, traditional perimeter firewalling is not efficient and does not provide the necessary protection against today's cyber security attacks. Microsegmentation can significantly improve the overall security level.

Keywords/tags (Data Center, Overlay, VXLAN, Microsegmentation)

Miscellaneous

jamk.fi

Kuvaus

Tekijä(t)	Julkaisun laji	Päivämäärä
Takamäki, Masi	Opinnäytetyö, ylempi AMK	Huhtikuu 2018
		Julkaisun kieli Englanti
	Sivumäärä 69	Verkkojulkaisulupa myönnetty: x

Työn nimi

Päällysteknologiat ja Mikrosegmentointi Konesaleissa

Tutkinto-ohjelma

Master´s Degree Programme in Information Technology, Cyber Security

Työn ohjaaja(t)

Häkkinen Antti, Huotari Jouni

Toimeksiantaja(t)

Rivinoja Jarkko, Cygate Oy

Tiivistelmä

Datamäärien valtava kasvu, vaatimukset saatavuuden ja tietoturvan osalta asettavat aivan uudenlaisia haasteita konesaliliiketoiminnalle. Perinteiset konesaliarkkitehtuurit eivät pysty enää vastaamaan näihin haasteisiin. Siirtokerrosta pidetään yleisesti vikaantumisalueena, tästäkin huolimatta sille on vielä monia tarpeita konesalien välillä. Tämän lisäksi kehittyneet kyberhyökkäykset yrityksiä vastaan ovat arkipäivää ja ne onnistuvat pelottavan suurella prosentilla.

Cygaten tiloihin rakennettiin testiympäristö modernien päällysteknologioiden ja mikrosegmentoinnin testaamista varten. Sovelluspohjaista arviointityökalua jonka ominaisuuksiin kuului aidon sovellusliikenteen simulointi sekä kyky helposti toimittaa keskeisiä suorituskykytietoja käytettiin testiympäristön toiminnallisuuksien vahvistamiseen.

Tutkimus perustui laadulliseen tutkimukseen. Tutkimusstrategiana oli kokeellinen tutkimus, joka mahdollisti tietyn ilmiön vaikutuksien tutkimuksen kontrolloidussa ympäristössä.

Tutkimuksesta saatiin ymmärrys siitä miten modernit päällysteknologiat toimivat, miten mikrosegmentointia tulisi toteuttaa ja mitä hyötyjä nämä kaksi tarjoavat.

Konesaliliiketoimintaa harjoittavien yrityksien kannatta vakavissaan harkita Ethernet pohjaisten ratkaisujen hylkäämistä ja siirtyä täysin IP pohjaisiin ratkaisuihin. Mikäli siirtokerrosta vaaditaan, päällysteknologioita kuten VXLAN:ia kannattaa toteuttaa. Lisäksi reunalla toteutettava palomuuraus on tehotonta, eikä tarjoa vaadittavaa suojaa tämän päivän kyberhyökkäksiä vastaan. Mikrosegmentointi voi parantaa tilannetta merkittävästi.

Avainsanat (Data Center, Overlay, VXLAN, Microsegmentation)

Muut tiedot

Acronyms

API	Application Programming Interface
BGP	Border Gateway Protocol
BUM	Broadcast, Unknown Unicast and Multicast
CE	Customer Edge Router
DCI	Data Center Interconnect
DWDM	Dense Wavelength Division Multiplexing
ECMP	Equal-Cost Multi-Path
ESI	Ethernet Segment Identifier
EVI	EVPN Instance
EVPN	Ethernet VPN
FEC	Forwarding Equivalence Class
IDS	Intrusion Detection System
IPS	Intrusion Prevention System
IT	Information Technology
LACP	Link Aggregation Control Protocol
LAN	Local Area Network
LER	Label Edge Router
LDP	Label Distribution Protocol
LSP	Label Switched Path
LSR	Label Switching Router
MC-LAG	Multi-Chassis Link Aggregation Group
MPLS	Multiprotocol Label Switching
NGFW	Next Generation Fire Wall
NLRI	Network Layer Reachability Information
NVP	Network Virtualization Platform
OS	Operating System
OVSDB	Open Virtual Switch Database
P	Provider Router
PE	Provider Edge Router
RSVP	Resource Reservation Protocol
SDDC	Software Defined Data Center
SDN	Software Defined Network
SLA	Service Level Agreement
STP	Spanning Tree Protocol
TE	Traffic Engineering
URL	Uniform Resource Locator
UTM	Unified Threat Management
VM	Virtual Machine
VNI	VXLAN Network Identifier
VPLS	Virtual Private LAN Service
VPN	Virtual Private Network
VTEP	VXLAN Tunnel End Point
VXLAN	Virtual Extensible LAN

Contents

Figures

Tables

1 Introduction

1.1 Clouds within Data Centers

For a long time, Information Technology (IT) has been considered as a cost in as situation, when all the time increasing requirements and shrinking budgets fail to meet the new business demands. This has now changed as organizations have realized the business opportunities that agile, scalable and flexible IT services offer. To support this, organizations are increasing the use of public, private and hybrid cloud services instead of smaller on-premise facilities managed by internal IT staff. What type of cloud or clouds to use depends on business and compliancy requirements. For example, for highly regulated documents such as medical research or health records the public cloud might not be the best location. For a certain type of lesser important data such as photos or videos, the actual location itself might not be that significant. Whatever cloud type is suitable for a certain organization, the magic behind the cloud remains the same: it is just someone else's computer or a collection of computers, i.e. more commonly known as a data center.

Data centers are the center of modern technologies, serving a critical role for organizations that seek new ways of enabling their business. They are not just huge box-like buildings in the middle of a desert, but industry's sharpest spear tip what comes to the latest technology, power consumption and cooling among other things. The possibilities that data centers offer are not only available for large organizations, but also for private consumers. These days buying or renting a virtual server, domain name and a public IP address from a public cloud is not a complicated task. Anyone can start up a new service in a cloud with minimal effort and cost and expand it easily if it bears fruit. However, cloud is not just for business, many people use cloud or cloud-based services every day, perhaps without even knowing it. Instant messaging and video calls in WhatsApp, entertainment services such as Netflix and Spotify, social media services such as Facebook, Twitter and Instagram just to name few. In addition, new phenomena such as Big Data, Internet of Things (IoT), Artificial Intelligence (AI), Machine Learning, 5G and Mobile Gaming will change or have already changed the world significantly. Where will all this massive amount of data

be most likely located? The answer is: in the cloud, and the clouds live inside data centers.

All these factors have caused a massive growth of total data stored and it continues to grow exponentially. According to IDC's Digital Universe Study (Gantz & Reinsel 2018), there is a 50-fold growth in all data from the beginning of 2010 to the end of 2020 (see Figure 1). To clarify the scale, one Exabyte is equal to 1 000 000 terabytes.

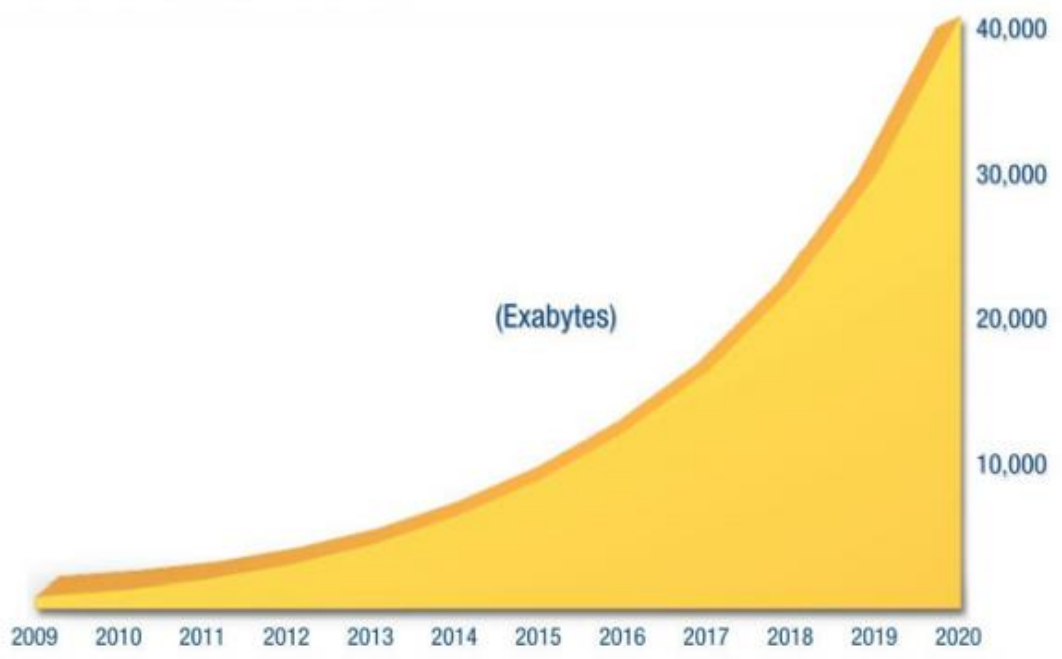

Figure 1. Growth of Data (Gantz & Reinsel 2018)

Massive growth of data, availability requirements for business-critical services and security requirements present evolving challenges for organizations running a data center business. Traditional data centers do not anymore provide the necessary scalability, flexibility and rapid deployment for modern and demanding applications and services.

1.2 Research Objective

The objective of this research is to create a basic directive architecture for a modern data center, which provides robustness, scalability, mobility and high throughput with minimal latency and jitter. This is to be achieved by outgrowing from Layer 2 solutions and using only Layer 3 technologies in the underlying network. Modern overlay technologies are then used to provide Layer 2 connectivity if it is required.

Another objective is more security oriented and examines the use of microsegmentation. No matter how many layers of different security controls there are, they only have to fail once which may lead to a compromised system inside a data center. The only way to efficiently stop lateral movement of an unauthorized user that has already penetrated all the defenses is to use microsegmentation. What kind of benefits does microsegmentation offer and how it may change the way applications and services are developed?

1.3 Research Method

The research is based on qualitative research method. The research strategy is experimental research, which enables exploring the influence and interaction of a certain phenomenon in a controlled environment. More specific, a lab environment was built with modern alternative technologies to find out if they provide the same or even better functionalities than traditional architectures. Noteworthy is also if the built lab environment provides enhanced security and if it reduces the failure domain size. An efficient software-based assessment tool with the capability to simulate real application traffic and ability to easily deliver key performance metrics was used to qualify the lab environment.

1.4 Research Questions

The research questions chosen for this thesis are the following:

- What kind of architectural changes do overlay technologies enable for the underlying network?
    - What should be considered when planning to change the underlying network architecture?
    - What are the key benefits of modern overlay technologies over traditional designs?
- What kind of architectural changes does microsegmentation enables within network and security perspective?
    - What are the key benefits and use scenarios for microsegmentation?

The first research questions should clarify the need for overlay technologies in modern networks as well as the key benefits of changing the underlying network architecture from Ethernet Fabric to IP Fabric. The last research questions should

give answer to how microsegmentation can change the way applications and services can be designed in a more flexible way, without forgetting the most important factor, security. Overall, these research questions together should provide a comprehensive view to the reader regarding what kind of different things should be taken into consideration when designing new data centers these days.

1.5 Structure of the Thesis

This thesis is divided into four different parts. After the introduction, the theory and background part starts at Chapter 2, which goes through the traditional data center design that is still widely used, as well as the more modern way of designing data center architectures. Chapters 3 & 4 focuses on more specific technologies by certain vendors, both that are widely used in data centers generally, as well as in this thesis's research and are therefore in a crucial role. Part two, Chapter 5 contains the actual research. Part three, Chapter 6 presents the results and the final part Chapter 7 presents the conclusions of and answers to the research questions of this thesis.

2 Data Center Architectures

2.1 Background

Organizations running data center business live in a state where technology evolves all the time, changes to the current environment are constant and yet they should be able to provide incredibly stable and flexible services for customers. In a provider perspective, this means that new features and improvements should be introduced all the time and this has to be done in line with the change management process. However, none of these operations should have any impact on continuous services provided to the customers. A typical SLA between the provider and the customer might be something like “five nines” (99.999 %). To clarify this, five nines equals 5.26 minutes of downtime per year or 25.9 seconds per month. That does not leave much room for mistakes, regardless of the root cause for a certain incident (human error a flaw in a software). In a customer perspective, provisioning new services or modifying the current should be as easy as it is in public clouds nowadays (just a few clicks via user-friendly interface). This requires a great amount of automation, which is not equally easy in all different areas. Many organizations already provide automated Virtual Machine provisioning, however, providing a single user interface for customers, which makes it possible to provision anything from new networks to firewall policies, load balancer rules or any added value services for virtual machines is currently available only from the major public cloud organizations.

2.2 Traditional Data Center Designs

A look at the history of the data communications network design, 100 % of the traffic was between desktop and mainframe computers. When the first workgroup servers were introduced, it changed this pattern to reflect an 80/20 rule: only 20 % of the traffic was intent to the data centers and 80 % of the traffic remaining in the workgroup. As enterprises realized the value of the data stored on these servers, this changed the communications ratio to 20/80 split. Nowadays the traffic patterns are not explainable anymore by a simple ratio of traffic to and from the workgroup.

Instead, they contain server-to-server traffic, Internet traffic as well as the legacy client-server traffic (Southwick, Marsche & Reynolds 2011).

2.2.1 Inside a Data Center

During the migration from 80/20 to the 20/80 traffic pattern the three-layer data center design was developed. This data center design was hierarchical, consisting of access, aggregation and core layers (see Figure 2.). The design was based on the current traffic flow patterns, limitations of that period's equipment design and the need for security (Southwick, Marsche & Reynolds 2011).

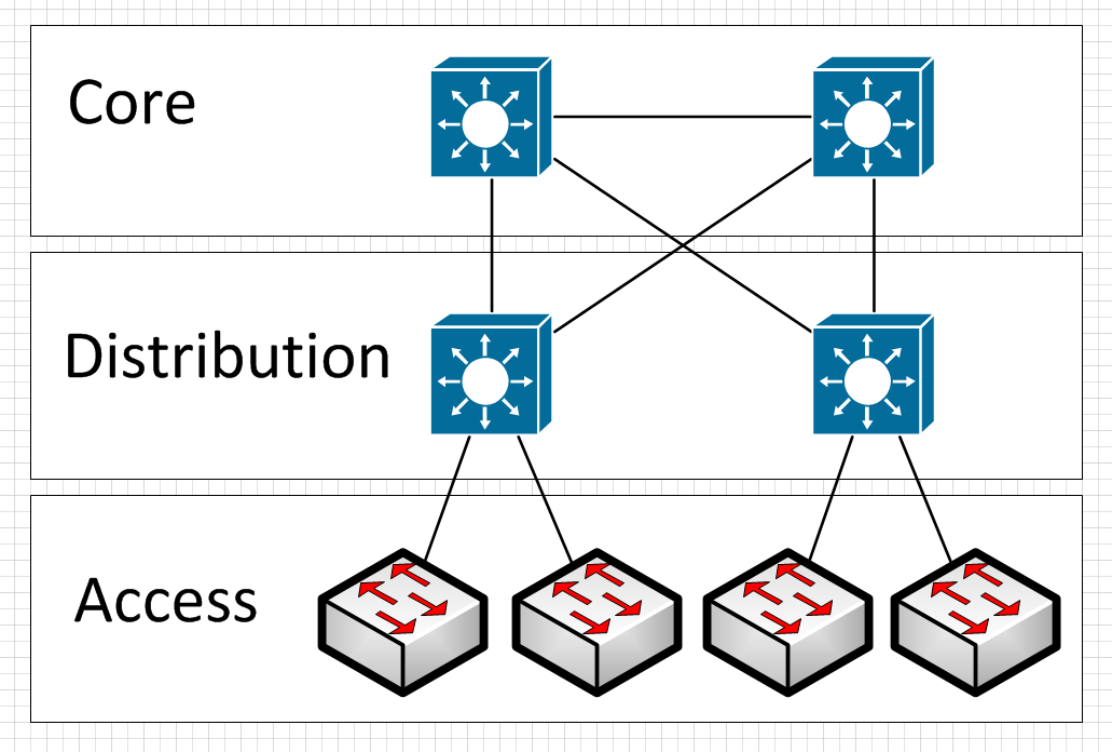

Figure 2. Three-Layer Network Design

In this design, the core layer is a high-speed switched backbone (since routers at that time could influence performance, switching was preferred). The distribution layer (sometimes called the aggregation layer) provides connectivity between the core and access layers. This layer consists of routers and firewalls that interconnect the access layers to high-speed core switches. Distribution layer is responsible for security between layers, as well as summarizing and aggregating of routes between subnets found in the access layer. The access layer is like the core, a switched layer that offers reliable connectivity to the distribution layer. It also provides connectivity to

the workstations and servers. The WAN layer is responsible of inter-site communications and Internet access. It is usually a routed layer, connected to the core layer using redundant links. The benefits of a hierarchical network design include:

- Modularity (facilitates change)
- Function to layer mapping (isolates faults)

To survive from both link and node failures in a three-layer design, backup routed links and multiple switched links are used. Redundant switched links create the possibility of broadcast storms with their associated outages. To avoid this situation, the Spanning Tree Protocol (STP) is used. STP ensures a network without the possibility of a network loop at a cost of efficiency of the links. This means that some of the redundant links are blocked from carrying traffic. STP recovers from links failures by a means of a series of timers that monitor the link and node health status. Convergence time for STP can cause disruption to communications for 10 seconds or more. Rapid Spanning Tree Protocol (RSTP) and other STP implementations can shorten the disruption time, but the mechanism is still the same and these long outages in a modern enterprise network is just not acceptable anymore. In addition to STP and RSTP limitations, the scalability of the three-layer design is quite limited, due to full mesh connectivity requirement of the core layer, number of interfaces and uplinks between the distribution and access layer (Southwick, Marsche & Reynolds 2011).

There are some solutions to overcome these limitations. One is to use stacked switches, which means that two or more switches logically become a one switch (single control plane). In some scenarios, this may eliminate the need for xSTP protocols (due to use of stacked switches, there is no possibility for a loop within a network), simplifies the overall management (reduced amount of devices to manage) and flexibility (installing a new switch to a stack increases the amount of total interfaces available). Disadvantage of stacked switches are that usually only one or two switches in a stack can hold the control plane responsibilities. If the active switch fails, backup switch will take over. However, if the both two switches that are control plane capable fail for any reason, the whole switch stack may become unusable. In a

management perspective, firmware upgrades for stacked switches are always more likely have some issues compared to single switches. To decrease the failure domain size in a control plane perspective, the use of Multi-Chassis Link Aggregation Group (MC-LAG) in certain parts of the network is highly recommended. MC-LAG switches act as a single device in a data plane point of view (like stacked switches), but they are independent on a control plane perspective. A reasonable combination of stacked switches, MC-LAG and Link Aggregation Control Protocol (LACP) in a three-layer network design can significantly improve the fault tolerance and the overall performance of the network.

2.2.2 Data Center Interconnect

The Data Center Interconnect (DCI) describes the method how two or more data centers are connected together. This enables data centers to work together; share resources and pass workloads between one another to provide high availability. DCI can be a Layer 2 (extends VLANs) or a Layer 3 (uses IP routing), but a Layer 2 DCI is required to provide high availability for example firewall or load balancer clusters and for virtual machine mobilization between data centers. The idea is to physically expand the Local Area Network (LAN) and Storage Area Network (SAN) connections between two sites, logically creating a single unit.

Several types of networks that can provide DCI:

- Point-to-Point, private line, dark fiber
- IP, customer or service provider owned
- MPLS, customer or service provider owned

Point-to-Point transport is the most flexible, providing easy DCI if the physical distance between the two sites is reasonable. With any greater distances, dark fiber or a Dense Wavelength Division Multiplexing (DWDM) connection from service operator is required and these are not cheap. Point-to-Point connection appears as another switch-to-switch link and it can run any protocol (xSTP, routing protocols, etc.). IP transport means that an IP network separates the data centers. There are many transport options for DCI; however, MPLS backbone is preferred because of reliability, scalability and traffic engineering features (MPLS is covered in more detail in Chapter 2.2.3).

DCI between two data centers is relatively simple to achieve, and it can even provide relatively good availability. However, there is usually a need for a third or more data centers as well. In addition to high availability requirements, two or more data centers may be required for witness site (the decision maker in split-brain situation) purposes and for disaster recovery solutions (to fill certain compliance requirements). Hub and spoke topology is not usable in this case and a triangle topology between data centers would make the DCIs prone to all the Layer 2 issues like loops in the network as xSTP protocols are not considered as a suitable solution for DCI. Figure 3. illustrates a simple two data center high availability solution, where DWDM connection is responsible for DCI for Ethernet and Fiber Channel connections.

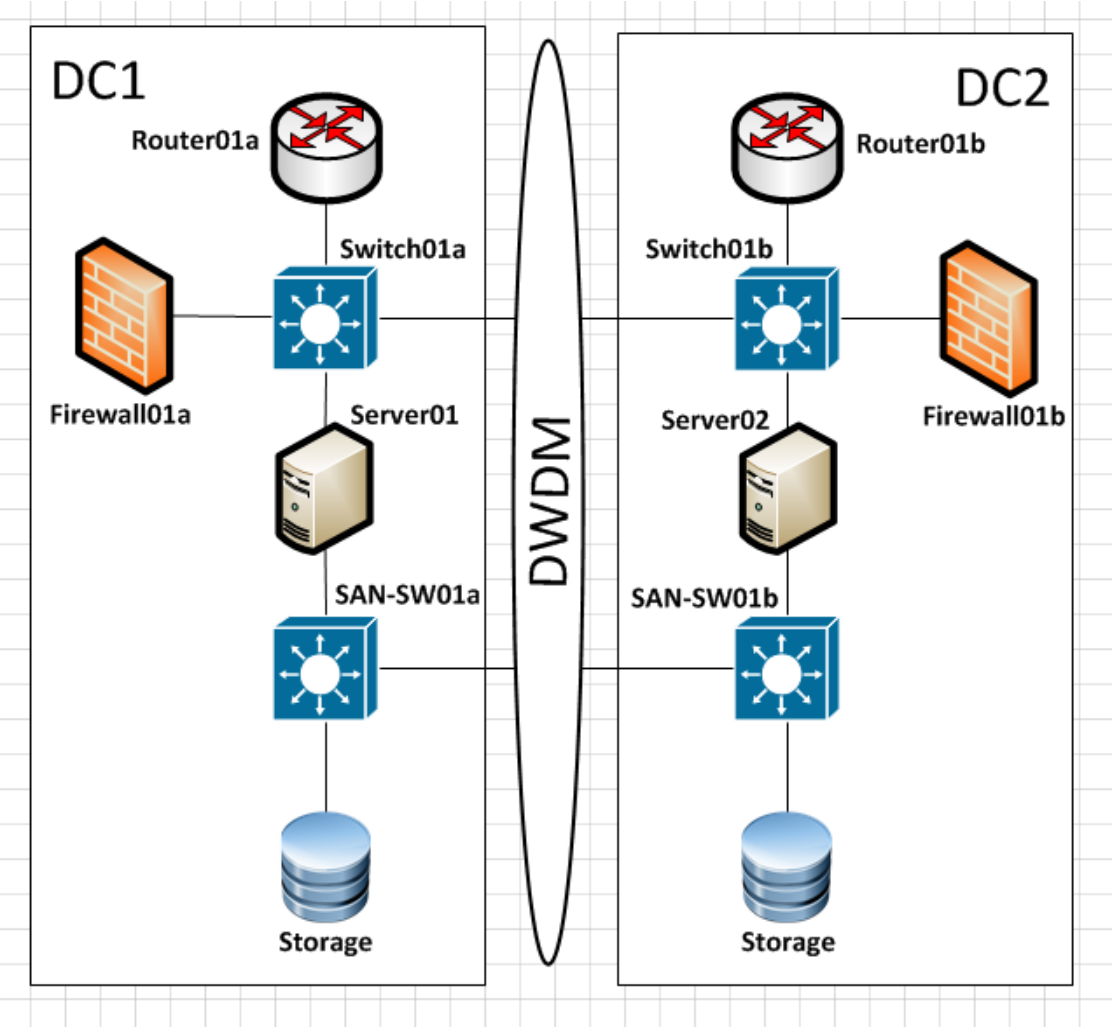

Figure 3. Layer 2 DCI using DWDM

2.2.3 MPLS

Multiprotocol Label Switching (MPLS) is a widely used packet forwarding technology in a service provider and enterprise networks, which uses labels in order to make forwarding decisions. In comparison with traditional IP networks, where routers running some routing protocol make independent forwarding decisions based on the IP packet's header. Each router analyses the header, looks on its own routing table and chooses the correct next hop. This lookup has to be done independently every time on every single IP packet, because the contents of the routing table can change occasionally. With MPLS, when packet enters from an IP network to a MPLS network first time, it is assigned to a certain Forwarding Equivalence Class (FEC) and a router appends a label to that packet. FEC is a subset of packets that are all treated in the same way by MPLS routers (treatment may be dependent on IP address, port information, DSCP etc.). Each router knows how to handle packets with a certain FEC, so once the packet arrives in the MPLS network there is no need to perform header analysis anymore. Routers use the label as an index into a table, which provides a FEC for the packet (see Table 1.). This makes MPLS networks considerably faster than traditional IP networks and it provides low latency services for a real-time traffic for example (RFC 2018a).

Table 1. Forwarding Equivalence Classes

IP Header Info	Label
xxx	10
yyy	11
zzz	12

After the FEC table lookup by the ingress Label Edge Router (LER), a certain label is added (or pushed in MPLS terminology) to the packet and then it is forwarded to MPLS domain. This label information is held in a special header sometimes called a shim header. This shim header is then inserted between the Layer 2 and the Layer 3 headers of the IP packet (see Figure 2.). Because of this, MPLS is sometimes referred as a Layer 2.5 service. The total length of the MPLS header is 4 bytes or 32 bits (see

Figure 4.). The first 20 bits forms the actual MPLS label; the next three bits were formerly known as experimental but are now renamed to Traffic Class (TC) and are used for Quality of Service (QoS) purposes. The next bit is the stack bit, which is called bottom-of-stack bit. This field is used to inform Label Switched Routers (LSR) if there is more than one label attached to the packet. Last eight bits are Time to Live (TTL), which is used the same way as in IP TTL byte in the IP header (mechanism to limit the lifetime of a packet).

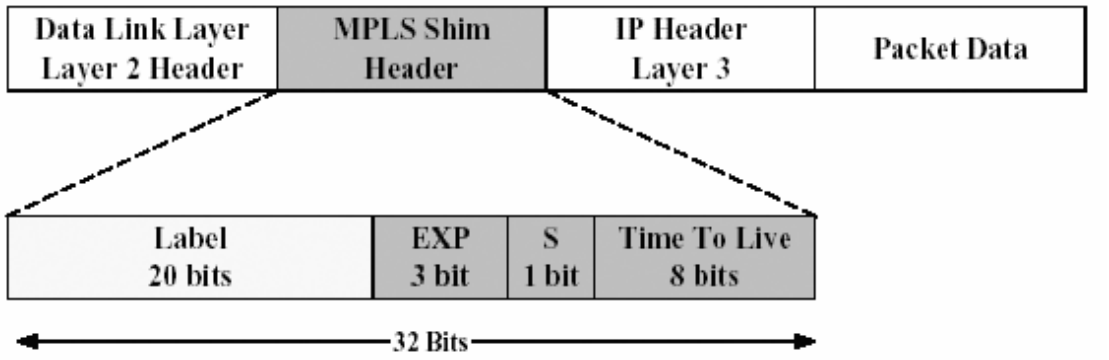

Figure 4. MPLS Shim Header (ResearchGate 2018)

Inside the MPLS domain packet arrives at one or more LSRs within its path. Each LSR performs a lookup for its Label Information Base (LIB), swaps the MPLS label to a new label and forwards it to MPLS domain. Finally, the packet arrives at the edge of MPLS domain and the egress LER removes (pops) the label and forwards the packet normally according to normal IP routing (RFC 2018a).

Another advantage of MPLS routing over normal IP packet routing is that the path across the MPLS network is established even before the packet starts its journey. This in advance known path is called a Label Switched Path (LSP), which is established by Label Distribution Protocol (LDP) or Resource Reservation Protocol Traffic Engineering (RSVP-TE). All LSPs are always unidirectional, so a return path is a separate LSP and may take a different route. Each router needs to know what label to use for a certain directly connected peer. This can be statically configured to each router, however, as this is not a scalable solution LDP or LDP and RSVP together are preferred. MPLS has very good traffic engineering capabilities, which makes the use of the network more efficient, as it can use certain characteristics like network topology and resources available within the LSP for optimization (RFC 2018a).

2.2.4 MPLS VPNs

MPLS is a technology that service providers and enterprises may market to organizations using any commercial terms, however, how can organizations benefit from MPLS? The answer is MPLS based Virtual Private Networks (VPNs). MPLS VPNs are biggest reason why organizations are using MPLS nowadays (the forwarding efficiency is not that relevant anymore due to fact that hardware performance has significantly increased). MPLS VPNs are private networks over a shared infrastructure. These VPNs usually consist of two different areas: the provider's network and the customer's network. Routers participating in an MPLS domain have different roles depending on which area they are located. Provider Router (P) is a core router part of provider's MPLS network forming LSPs. Provider Edge Router (PE) is a router at the edge of provider's MPLS network and they connect Customer Edge Router (CE). CE routers are part of a customer dedicated Virtual Routing and Forwarding (VRF) instance on the PE device.

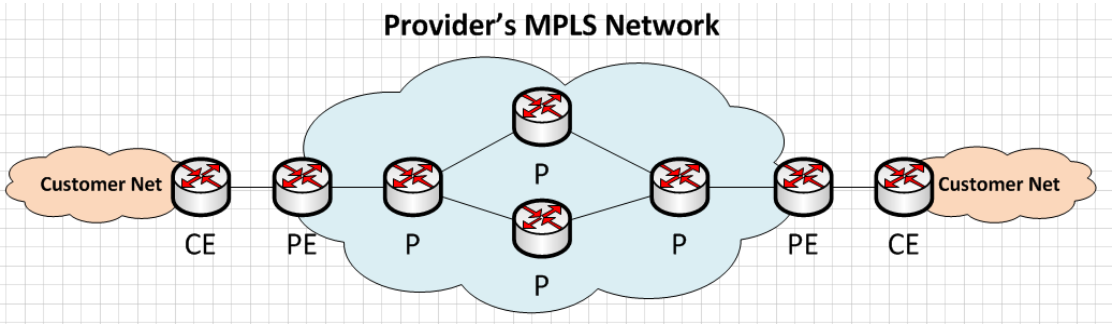

Figure 5. Three Different MPLS Router Roles

To ensure that VPNs remain private and isolated between different customers, MPLS network providers maintain policies that keep routing information separate from different VPNs. However, only LDP sessions can be authenticated between peers - MPLS VPNs as such do not provide authentication, integrity check and encryption such as IPSec VPNs do (Juniper 2018c).

VPN services consist of two basic components, data plane and control plane. Data plane describes the method how gateway encapsulates and decapsulates the original data. The control plane describes the process of learning (auto-discovery and/or signaling) performed by the gateways. Gateways can be statically configured which is not scalable, so usually a dynamic protocol (such as BGP) for control plane signaling

between gateways is used to exchange information. Multiprotocol BGP (MP-BGP) is an extension to BGP and it enables BGP to carry routing information for multiple address families (AFI) and network layers. MP-BGP is responsible for carrying Network Layer Reachability Information (NLRI) for different address families (like IPv4 and IPv6). MP-BGP session for signaling is established between the PE devices.

There are three different types of MPLS VPN services in use:

1. Point-to-point (Pseudo Wire)
2. Layer 2 MPLS VPN, or VPLS
3. Layer 3 MPLS VPN

The simplest one of these is an emulated Layer 2 point-to-point connection between two endpoints, better known as a Pseudo Wire. Pseudo Wire is intended to provide only the minimum necessary functionalities and is therefore not a suitable solution for any larger deployments. The most common Layer 2 VPNs in use are Kompella L2VPN (defined in RFC 6624) and Martini L2VPN (defined in RFC 4788). Both of them provide Virtual Private Wire Service (VPWS), but the signaling is different as Kompella uses BGP and Martini LDP. L2VPNs dynamically create point-to-point Pseudo Wires and they support many different Layer 2 technologies (Sanchez-Monge, Szarkovicz 2015).

Virtual Private LAN Service (VPLS) is an Ethernet based multipoint-to-multipoint Layer 2 VPN. VPLS overcomes other Layer 2 VPNs restrictions by offering a “switch in the cloud” style service that allows customers to connect to geographically spread sites together. This means that the behavior is similar as in the case that the remote sites were connected to the same LAN. VPLS MAC learning is performed in the forwarding plane (as comparison to VPWS, where MAC learning is not performed at all). If a PE device receives a frame with a known destination MAC address (MAC address is present in the MAC table), the frame is forwarded point-to-point to the destination host. All the other frames (Broadcast, Unknown unicast and Multicast) are flooded as in normal Layer 2 switches. To avoid loops in VPLS, split horizon is used. This means that a frame that is received on Pseudo Wire is never sent back on the same Pseudo Wire. In addition, a frame received on a Pseudo Wire is not forwarded on any other Pseudo Wire either (this is the default behavior). VPLS

signaling can implemented using BGP signaling, LDP signaling or both of them. Only BGP signaling supports auto-discovery (Juniper 2018c & Sanchez-Monge, Szarkovicz 2015).

The most commonly used MPLS VPN service is the Layer 3 VPN, or just MPLS VPN as in provider's language. In MPLS VPN, the entire service provider network acts like a distributed router from the customer perspective. In this scenario, the service provider creates VRFs on their PE routers and customers connecting from different CEs are then placed to the same VRF. After that, customer CE can exchange routing information with provider's PE and provider forwards these routes to other PEs on the same VRF and eventually to another CE of the customer. These Layer 3 MPLS VPNs based on RFC 4364 are also known as BGP/MPLS because BGP is used to exchange routing information across the provider's network, and MPLS is used to forward VPN traffic across the remote VPN sites (Juniper 2018c).

2.2.5 Traditional Data Center Limitations

Typical data centers imitate the three-layer network design, which makes sense because that is something the organizations have already done for many years and they have a lot of experience on that. Layer 2 is easy to implement, scale and manage when everything goes well. In addition, the mechanisms to reduce the Layer 2 failure domain size have evolved, yet they cannot remove the root cause: Layer 2 itself. All the customers claim that they have mission critical applications that require high availability, and high availability configurations usually require a network connectivity within the same broadcast domain. Is it reasonable to place these mission critical applications or services in a single Layer 2 domain that may spread across data centers, when even a single broadcast packet entering to a loop will probably bring both data centers down?

Despite these facts, Layer 2 might still be useful in small and constant environments. However, not for larger implementations as Layer 2 has the following fundamental problems:

- The basic behavior - flood when unknown
- No sufficient data plane loop detection mechanism
- No summarization at the boundary, when spanning across data centers

The earlier mentioned Layer 2 VPNs over MPLS (in Chapter 2.2.4) also suffer from these same limitations. For example, MAC learning in VPLS happens in the Forwarding Plane, which may have a crucial impact on reliability in case of any incident.

In addition to these, Layer 2 VLAN address space is a limited 12 bits field, providing only 4096 unique VLAN identifiers which is significant limitation for service providers of a certain size. Q-in-Q (IEEE standard 802.1q) and MAC-in-MAC (IEEE standard 802.1ah) provide more flexibility for large operators when it comes to the VLAN address space, merely using multiple tags to encapsulate multiple VLANs to a single VLAN.

2.3 Modern Data Center Designs

2.3.1 EVPN

To overcome all the limitations of traditional network designs, Software Defined Networking (SDN) has been growing all the time and different solutions have hit the market. However, there has been no standard way of signaling the creation of virtual networks and exchanging MAC addresses. In top of that, there are multiple different data plane encapsulations available for overlay networking:

- Virtual Extensible VLAN (VXLAN)
- Network Virtualization using Generic Routing Encapsulation (NVGRE)
- Stateless Transport Tunneling (STT)
- Multiprotocol Label Switching (MPLS)-over-MPLS
- MPLS-over-User Datagram Protocol (UDP)

This thesis focuses only on mostly used EVPN data plane encapsulations, VXLAN and MPLS. EVPN is implemented on MP-BGP, so it enables BGP to carry routing information for multiple address families and network layers. To overcome all the limitations of older technologies, EVPN was designed to be agnostic to the underlying data plane encapsulation, as it has nothing to do with the actual function and design of the data plane itself. However, binding EVPN to a particular data plane encapsulation can limit the EVPN's deployment and use case. The architecture of EVPN is very similar to MPLS L3VPN. A huge benefit of EVPN architecture is that

enterprises can now implement the same control plane protocol (MP-BGP) across the entire network: from data center to WAN. MP-BGP EVPN uses Address Family Identifier (AFI) of 25, which is the Layer 2 VPN address family. Subsequent Address Family Identifier (SAFI) is 70, which is the EVPN address family. BGP is a proven protocol in enterprise and service provider networks and it has ability to scale to millions of route advertisements. It is also very policy oriented (compared to other routing protocols), which gives administrator complete control over route advertisements (Hanks 2016 & Juniper 2018d).

As already mentioned, EVPN signaling uses MP-BGP, and it currently has five route types defined in RFC 7432 and four route types at a draft stage (see Table 2.)

Table 2. EVPN Route Types

Route Type	Description	Usage	RFC
0	Reserved		RFC 7432
1	Ethernet Auto-Discovery	Multipath and Mass Withdrawn	RFC 7432
2	MAC/IP Advertisment	MAC/IP Advertisement	RFC 7432
3	Multicast Route	BUM Flooding	RFC 7432
4	Ethernet Segment Route	ES Discovery and DF Election	RFC 7432
5	IP Prefix Route	IP Route Advertisement	draft-ietf-bess-evpn-prefix-advertisement-04
6	Selective Multicast Ethernet Tag Route		draft-ietf-bess-evpn-igmp-mld-proxy-00
7	IGMP Join Synch Route		draft-ietf-bess-evpn-igmp-mld-proxy-00
8	IGMP Leave Synch Route		draft-ietf-bess-evpn-igmp-mld-proxy-00

EVPN Route type 1 is required in All-Active Multihoming as Ethernet Auto-Discovery routes are advertised on per EVI or ESI. Route Type 2 is needed for advertising MAC addresses between PE routers for a certain ESI. Type 3 is responsible for BUM traffic delivery across the EVPN network, so the ingress router knows how to send BUM traffic to other PE devices in the same EVPN instance. Route type 4 is used in All-

Active Multihoming scenarios for the PE devices within the same ESI to discover each other. Type 5 is responsible for advertising IP prefixes for inter-subnet connectivity between data centers. These data packets are sent as Layer 2 Ethernet frames encapsulated using VXLAN header over IP network (Juniper 2018a, Juniper 2018e & RFC 2018d).

2.3.2 IP Fabric Inside a Data Center

To overcome Layer 2 limitations inside a data center, the only sustainable solution is to use IP Fabric solution as an underlying network architecture. IP Fabric is one of the most flexible and scalable solutions available for data centers nowadays. It is a full IP infrastructure with no Layer 2 switching or xSTP protocols, which makes it inherently stable and loop free. Instead, it uses standards-based Layer 3 protocols allowing enterprises to use any vendor's products (interoperability, topology can be a mix of devices). All devices in IP Fabric are basically just Layer 3 routers (or switches, performing Layer 3 operations) that rely on routing information to make forwarding decisions. Routers within IP Fabric are called Spine- and Leaf nodes. In a Spine/Leaf architecture, each Leaf node has a routed link connection to each Spine nodes. No direct physical connectivity exists between Spine nodes or between Leaf nodes. Spine or Leaf function is only a matter of the device's physical location in the IP Fabric; the device itself does not know if it is a Spine or Leaf node.

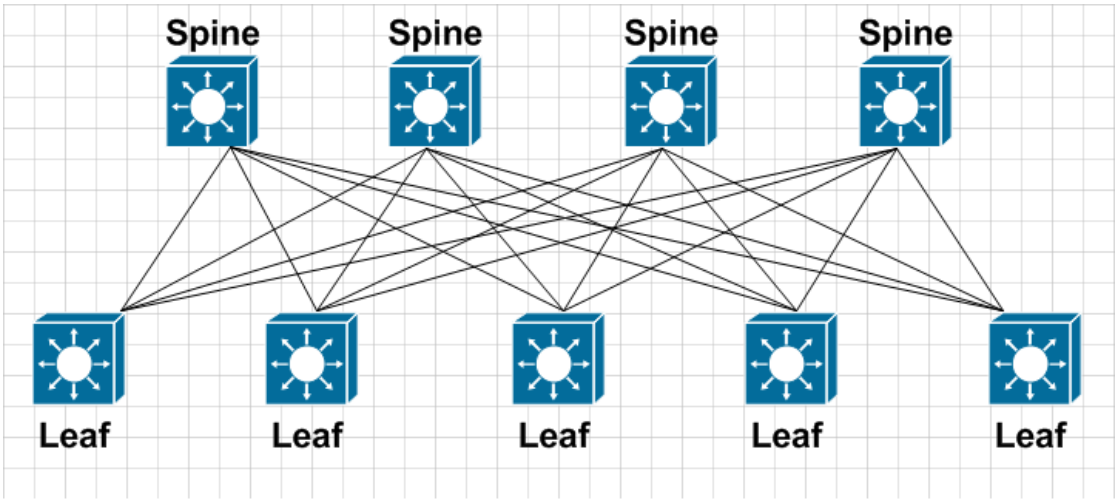

Figure 6. Spine and Leaf Architecture

Compute nodes (servers), firewalls, load balancers and edge routers inside a data center are only connected to the Leaf nodes (basic principle, nothing connects to a Spine expect a Leaf). This kind of a setup creates a resilient network, where all traffic has multiple paths to all other devices in the fabric (each server-facing interface on Leaf node is always two hops away from any other Leaf node-facing interface). This also creates predictable behavior in case of any failure and consistent latency across the whole fabric. Supported routing protocols for IP Fabric are BGP, ISIS and OSPF. The most commonly used and most suitable for very large implementations is BGP or EBGP more precisely, where all devices within IP Fabric are in a different AS. Every Leaf node has a peering session to every Spine node. To benefit from the fact that all routes in the fabric are in active state at the same time, Equal-Cost Multi-Path (ECMP) routing should be implemented to utilize all the connections. To optimize the failover time in a failure situation, the use of Bidirectional Forward Detection (BFD) over any dynamic routing protocol is highly recommended. ECMP and BFD together serve a crucial role in IP Fabric, as it provides increased throughput and minimal downtime for the underlying network.

Leaf and Spine topology-based architecture is ideal to “East-West” flow of data, instead of traditional “North-South” data flow. This means that traffic is mostly server-to-server instead of client-to-server, so the actual traffic may not travel outward of the actual data center itself. Within IP Fabric, each Leaf node has its own VLAN address space, so if Layer 2 mobility is required between Leaf nodes VXLAN is one solution (covered in Chapter 2.2.4). IP fabric can span between data centers without any issues (Juniper 2018d).

2.3.3 EVPN-MPLS for DCI

So far, the only somehow scalable overlay technology to provide Layer 2 DCI over WAN between multiple data centers has been VPLS, which has all the drawbacks of the Layer 2 switched network:

- No active-active multihoming
- No choice in data plane encapsulation
- No control plane
- MAC discovery, advertisement and flooding is inefficient

EVPN uses MAC and IP addresses as endpoint identifier, as MPLS/VPN only uses IP prefixes. EVPN was once called MAC VPN, because it implements MAC route advertising. Important aspect of MAC addresses (compared to IP) is that MAC addresses do not support any subnetting or aggregation. This means that advertisement in EVPN is always one MAC route per host. Due to rich features available in MP-BGP, EVPN brings a host of control plane policies into design of the network. At high level, three different types of Ethernet Services exist:

- VLAN-Based
- VLAN-Bundle
- VLAN-Aware

In EVPN VLAN-Based Service each VLAN is mapped directly to its own EVPN Instance (EVI), so there is a direct one-to-one mapping between VLAN IDs and EVIs. Because of this, there is a route target (RT) per VLAN ID. EVPN also assigns a label to each VLAN ID (in VXLAN VNID is used as a label), which makes it possible to change the VLAN ID between sites (LERs do not care of the VLAN ID, all forwarding and flooding happens with the EVPN label). The downside of VLAN-Based EVPN Service is that it requires many labels to scale (for example, 16 000 VLANs would mean 16 000 labels) and the benefit that it allows VLAN normalization. In VLAN-Bundle type of EVPN Service, all the VLAN IDs share the same EVI (so there is an n:1 ratio of VLAN IDs to an EVI). This saves on label space, but the drawback is that it does not support VLAN normalization or overlapping MAC addresses and flooding is very inefficient due to shared label. VLAN-Aware EVPN Service type is a hybrid design of the first two. Each VLAN share the same EVI in control plane perspective, but in the data plane point of view, each VLAN ID has its own label, which in effect creates its own broadcast domain per VLAN ID. This hybrid design makes the flooding much more efficient and supports VLAN normalization (Hanks 2016 & Sanchez-Monge, Szarkovicz 2015).

The main advantage of EVPN over VPLS is EVPN All-Active Multihoming support (Single-Homed and Single-Active are still supported). EVPN introduces a new router role, which is called the designated forwarder (DF). DF is responsible for forwarding all the BUM traffic for given Ethernet Segment Identifier (ESI). ESIs are created when a set of PE routers create a LAG to a CE device for redundancy using all-active links.

The other PE router assumes the role of backup designated forwarder or BDF (Sanchez-Monge, Szarkovicz 2015).

2.3.4 EVPN-VXLAN inside a Data Center

The unfortunate fact that some applications and services still require a Layer 2 connectivity enforces the use of some overlay technology over IP Fabric. Best scalable solution to achieve this is to use Virtual eXtensible Local Area Network (VXLAN) as a Layer 2 VPN. VXLAN was designed to address all the traditional Layer 2 network issues and limitations (STP, VLAN address space and how MACs are handled). Many vendors have chosen to support VXLAN as the Layer 2 VPN, as it does not rely on MPLS transport. VXLAN transport tunnels are IP-based running on top of UDP destination port of 4789 (by default). VXLAN is defined in RFC 7348 as follows: “It takes Ethernet Frames and encapsulates them into IP Packets”. VXLAN data plane component consists of the following:

- Encapsulation, adding an outer Ethernet header, outer IP header, outer UDP header and VXLAN header to the original L2 frame (original VLAN tag is usually removed)
- Decapsulation, removing all the above headers and forwarding the original L2 frame to its destination (if necessary, adding the appropriate VLAN tag)

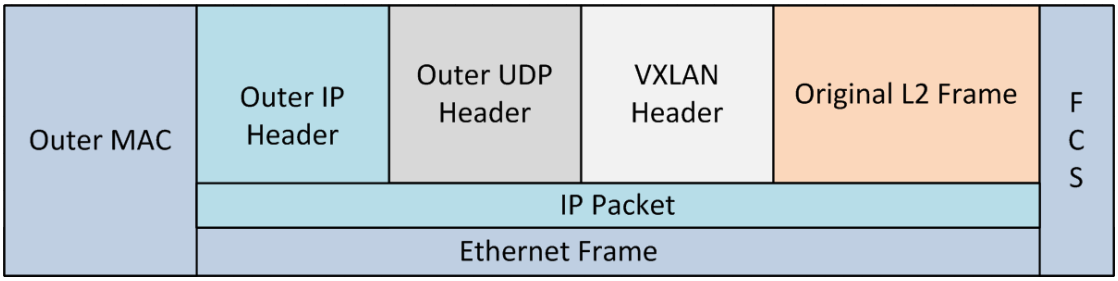

Figure 7. VXLAN Packet Format

1. Original L2 Frame, the frame that is being tunneled over the underlay network (without original VLAN tag)
2. VXLAN Header, 64 bits
3. Outer UDP Header, usually the well-known destination port of 4789 (might be something else, but this port is assigned for VXLAN by IANA)
4. Outer IP Header, source IP is the sending VXLAN Tunnel End Point (VTEP) and destination IP is the receiving VTEP IP
5. Outer MAC, as packet is sent normally over Layer 3 network, source and destination MAC changes at each hop
6. Frame Check Sequence (FCS), new FCS for outer Ethernet Frame

VXLAN header is a 64 bits field that has:

- Flags, 8 bits where the first bit must be set to "1" for valid VXLAN Network ID (VNI) and the other 7 bits are reserved fields, set to "0" on transmission and ignored by the receiver
- VXLAN Segment ID/VXLAN Network Identifier, 24 bits value which is designated to individual VXLAN overlay network on which communicating members are located
- Reserved fields, 24 and 8 bits, must be set to "0" on transmission and ignored by the receiver

VXLAN has an ID field with 24 bits meaning that there can be over 16 million of unique identifiers compared to 12 bits VLAN, which has the limitation of 4096 unique identifiers. Control plane component (the learning of remote VXLAN gateways):

- Static configuration or multicast using PIM
- MP-BGP EVPN
- Open Virtual Switch Database (OVSDB)

VTEP is the endpoint of VXLAN tunnel, it encapsulates Layer 2 frames using VXLAN encapsulation, sends them into the wire and decapsulates them when they reach their destination. VXLAN control plane describes the methods available for VTEPs to learn about other VTEPs and synchronize the information that each VTEP has. For example, VMware NSX supports only OVSDB as some other vendor's solution might support OVSDB and/or EVPN as a VXLAN control plane. The following figures (Figure 8. and 9.) explain how the data flow changes in a traditional switched network versus VXLAN encapsulated traffic over an IP Fabric (RFC 2018b, 2018c & Juniper 2018d).

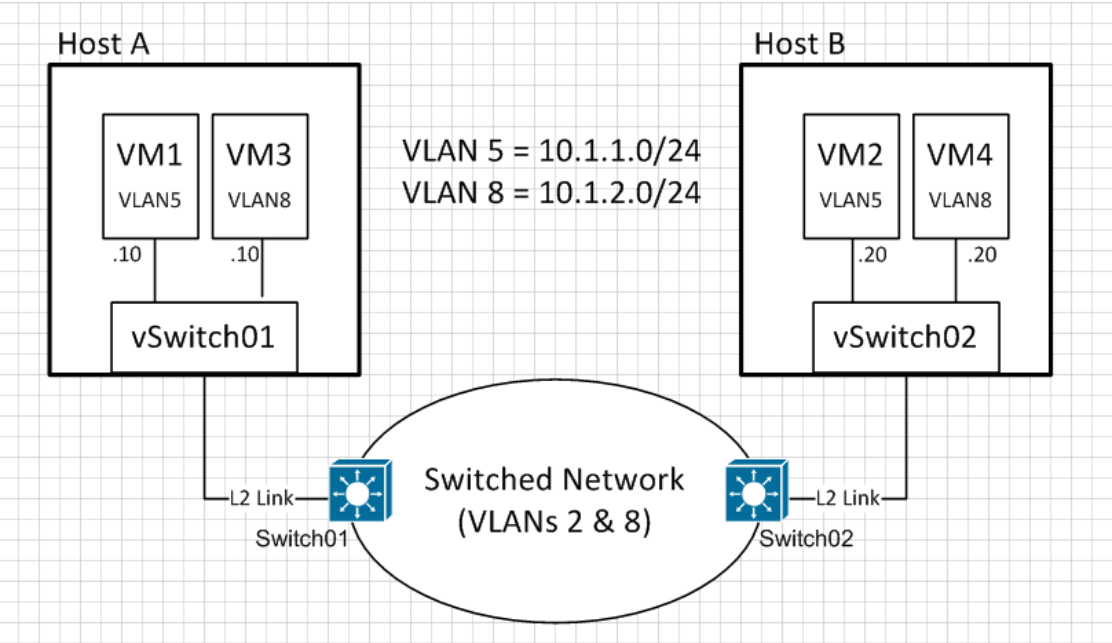

Figure 8. Traditional Switched Network

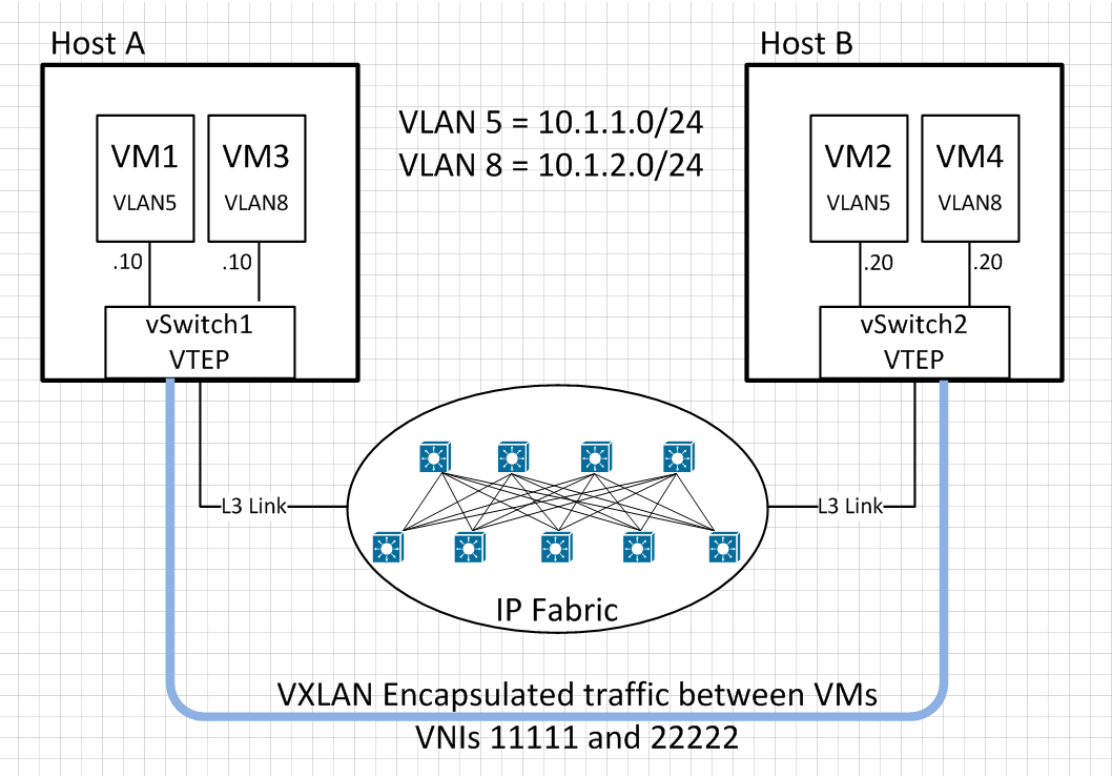

Figure 9. VXLAN Encapsulation over IP Fabric

In Figure 9. Hosts A and B have a virtual switch that supports VTEPs. VXLAN control plane could be a static mapping per VTEP (or VXLAN Gateway), where VLAN 5 maps to VNI 11111 and VLAN 8 maps to VNI 22222. The VLAN IDs on Host A and Host B does not need to be the same, as the VLAN ID is removed by the source VTEP and re-

added by the destination VTEP during VXLAN encapsulation process. If a VM needs to communicate with a bare metal server (i.e, an unvirtualized or a physical server) that does not support VXLAN, there the server needs to connect to a router that supports VXLAN Layer 2 Gateway function. Another form of gateway that a router might support is a VXLAN Layer 3 Gateway. A Layer 3 gateway acts as a default gateway for hosts in the same VXLAN segment. VTEPs that are part of a virtual switch are usually referred to as a software VTEP whereas routers acting as a VXLAN Gateway (Layer 2 or 3) are usually referred as a hardware VTEP. In larger implementations than just a few VTEPs, EVPN is the recommended control plane for VXLAN. EVPN signaling is based on BGP, so it is highly scalable, allows multipath forwarding to active/active multi-homed server (server can be attached to two or more different Leafs) and control plane MAC learning which considerably reduces BUM traffic. BUM traffic is a generic term describing Layer 2 Broadcast, Unknown Unicast and Multicast traffic (Juniper 2018d).

2.3.5 Benefits of the Modern Overlay Technologies

As the requirements have dramatically changed and the predominant amount of traffic is now East-West instead of North-South, the three-layer network design is well outdated. Inside a data center, Spine and Leaf architecture scales extremely well. An enterprise can start up with just a few devices (for example, two Spines and two Leafs) and flexible add more if the number of interfaces or total bandwidth is not enough. IP Fabric based solutions provide reliable, low latency and high throughput Layer 3 network without any Layer 2 disadvantages. Use of Layer 3 routing protocols with ECMP improves the total available bandwidth by utilizing all available links. If a Layer 2 connectivity is required, an overlay technology like VXLAN in top of the underlying network is an excellent choice. Enterprises running their own MPLS Network with VPLS for DCI should really consider changing to EVPN, as it provides much more effective MAC handling capabilities and intelligent control plane learning. Modern overlay technologies also make it much more easier to troubleshoot if there are any issues, as there is no more need to for MAC tracing (switch by switch) or capturing traffic from multiple points of the network and then manually comparing the results.

2.4 Traditional Data Center Firewall Designs

Organizations have been using routers or Layer 3 switches as the default gateway for a long time, but as the security has gained more weight and implementing access list to separate networks was no more satisfactory, firewalls (or load balancers) are often implemented as the default gateway for all the networks. This change in traffic flows has placed new demands for the firewall throughput, as all the traffic between networks is being inspected (even though it may not be necessary for all the traffic). In a price perspective, a firewall with 1G or 10G connectivity versus a router or Layer 3 switch with same specifications might be ten times more expensive. In addition, from the service operator perspective, customers need to be logically separated by a virtual router or a virtual system. For example, small or medium sized firewalls usually have a quite limited amount of virtual systems available, meaning that the service provider may need to over-size the firewall in capacity perspective to obtain the required number of virtual systems.

2.4.1 Traditional Three Tier Architecture and Traffic Flow

Security inside a data center has been traditionally performed at the perimeter, because most of the traffic flows were North-South only. Most of the applications running inside a data center have been designed to respect these traffic flows. This Three Tier Architecture consist of the following tiers:

- Presentation Tier (Web)
- Application Tier (App)
- Database Tier (DB)

To be able to provide necessary security controls between different tiers, these tiers have been placed in separate network segments, so the traffic crossing between tiers can be enforced to go through a firewall and/or a load balancer. This architecture is in line with the three-layer network design (covered in Chapter 2.2.1), which consists of the access, distribution and core layers. Usually data centers have dedicated perimeter firewalls inside the data center as well as in the data center border that connects to the outside world.

2.4.2 Disadvantages

Any application or service running inside a data center designed using three-tier architecture has a few disadvantages. For example, in a situation where a certain application is publicly available from the Internet to any user, the incoming connection will generate a relatively much traffic. The following devices would probably have to participate in a single query from the end user:

- Internet Firewall or Load balancer, at the Data Center border
- Firewall or Load Balancer, before Web Tier
- Firewall or Load Balancer, before App Tier
- Firewall or Load Balancer, before DB Tier

Assuming that this single query creates 100 KB of initial traffic and a single session, this would be multiplied by at least three or four times due to three-tier architecture design (creating a snowball effect, as the traffic bounces multiple times in the physical wire). This may not seem much in a bandwidth perspective, however, if there are 10 000 users this would generate 1 GB of traffic and many sessions. In addition, if any server in a certain tier is compromised, the attacker is able to freely move inside that tier without any possibility of administrative control or visibility.

2.5 Data Center Firewall Design using Microsegmentation

Microsegmentation inside a data center or in a cloud could enhance security significantly. It is mostly implemented inside hypervisors to enforce security controls between two Virtual Machines, even though those two VMs are located in a same network segment (in comparison, in a non-virtualized environment this traffic would be switched locally with no possibility to enforce the traffic to a perimeter firewall). This kind of a design transfers workloads from the perimeter firewall to the hypervisor firewall, which is responsible for microsegmentation. This contributes the perimeter firewall to focus more strictly on the border of the data center.

2.5.1 Microsegmentation and Change to Traffic Flow

If the same application or service that was used as an example in Chapter 2.4.2 would be implemented using microsegmentation, all the three Virtual Machines

(Web, App and DB) could be located in the same network segment. If these three VMs were located in a single hypervisor, the packet would need to enter the hypervisor only once as all the firewall and load balancer requirements could be implemented using distributed services. The traffic between the tiers would never have to leave to the physical wire. In addition, as the traffic flows nowadays are mostly inside a data center, microsegmentation together with Leaf and Spine architecture is an optimal design.

2.5.2 Advantages of Microsegmentation

The main advantage of microsegmentation is to be able to control and have a visibility to the lateral movement (East-West traffic) inside the data center, if a malware or a malicious user has been able to gain unauthorized access to any system. With microsegmentation, it is possible to protect all the traffic flows allowing only the flows that are required for a certain service or application to function. This approach is known as the Zero Trust Model. Distributed firewalling at a hypervisor level also offers a very high throughput and performance compared to perimeter firewalls, assuming that the server running these services has reasonable amount of capacity. Microsegmentation can change the way how enforcing security policies is done (which is currently mostly based on per IP addresses), as it can support more dynamic factors, like Virtual Machine tags, virtual switch membership or a folder where a certain VM is located in the hierarchy.

Enforcing security controls as close to the source as possible is very efficient and it provides easier means for automation and orchestration, as all the configurations are done to virtual firewalls that are most likely already controlled by some management tool. Utilizing microsegmentation should be carefully planned before implementation, because poorly designed solution can have an overwhelming administrative burden and it might not increase the overall security at all.

3 Next Generation Firewalls

The purpose of a firewall is to protect an organization's networks and assets. Originally, firewalls were just an IP and port-based gatekeepers allowing legal traffic to pass and denying everything else. Nowadays firewall industry is a highly competitive field offering various vendors' products with multiple different capabilities, where a traditional firewall's mission is only a one small piece of a puzzle.

3.1 The First Generation

The first generation of firewalls were stateless, i.e. each packet or frame needed to be processed individually against the set of user defined rules. These firewalls or routers (also known as packet filters) examined each packet based on the following criteria:

- Source IP address
- Destination IP address
- Source TCP/UDP port
- Destination TCP/UDP port

As these devices were unaware of the connection state, an administrator needed to create a rule allowing the incoming packet (TCP SYN for example) and then create a second rule allowing reply packet (TCP SYN ACK). This of course causes additional administrative overhead and in hardware perspective, processing each packet separately is quite a CPU intensive process. Packet filters are susceptible to IP spoofing, where a malicious user is trying to gain unauthorized access by sending messages with a spoofed IP address. In addition, they rarely provide sufficient logging or reporting capabilities (Tech Republic 2018).

The first stateful firewall was introduced more than 20 years ago by Check Point. Compared to stateless, these stateful firewalls were able to keep track of the connection state (TCP session or UDP communication). An initial request for a connection (TCP SYN) is evaluated against the set of rules. If allowed, reply packets (TCP SYN ACK and ACK) are also allowed without a need for a second rule. This successful three-way handshake will finalize the connection state to be established.

The established connection state is held in the memory (state table) and communication between hosts is now freely available. Keeping record of the connection state makes it more efficient in terms of packet inspection but also enables new kind of attack vectors for a malicious user (Check Point, 2018).

3.2 The Second Generation

The second generation attempted to increase the level of security by adding a software layer to intercept connections. These devices are more commonly known as application proxy or gateway firewalls. The proxy is transparent for a client and server, evaluating all the data that is sent through an established connection between these two endpoints. The software layer was able to enforce additional user defined policies instead of just IP addresses and ports, such as URL filtering (Uniform Resource Locator). In addition, they usually offered better logging and reporting capabilities for administrators, however, for a user experience; they might cause additional latency and reduce the overall throughput (Tech Republic 2018).

3.3 The Third Generation

The third generation of firewalls are able to perform real-time traffic inspection on the wire without affecting the throughput, or at least this is what most vendors claim. Third generation firewalls are usually single device boxes with UTM (Unified Threat Management) capabilities, which might include any of the following:

- Firewall Capabilities
- Intrusion Detection/Prevention System (IDS/IPS)
- Antimalware
- Spam
- Content Filtering
- IPSec VPNs
- Identity Based Control
- SSL/SSH Inspection
- Application Awareness

The concept of UTM is to add multiple of these critical security technologies, integrated into a single appliance provided by a single vendor. Consolidating many features to a single device does save expenses and simplifies management, however,

as a drawback, these devices can become a single point of failure, and usually enabling more and more UTM features affects the overall performance of the UTM appliance (Information Week 2018).

3.4 Palo Alto Next-Generation Security Platform

Palo Alto Networks was founded in 2005 by Nir Zuk, who already had previous history with vendors such as Check Point and NetScreen (later acquired by Juniper). By 2017, this next-generation security company had more than 45 000 customers in over 150 countries. Palo Alto Next Generation Firewall is a zone-based firewall, which means that security policies are applied between zones. A zone is a group of interfaces (physical or virtual) that represent a segment in the network that share the same security requirements. In a hardware architecture perspective, Palo Alto firewall contains a separate Control Plane and Data Plane. By separating these two, Palo Alto ensures that each plane runs independently and they have their own dedicated processors, memory and hard drives (Palo Alto 2018a).

3.4.1 Palo Alto's Single-Pass Architecture

According to Palo Alto, all UTM devices are capable of performing firewall functions on a low latency and high throughput; however, adding more security features will eventually lead to decreased performance and increased latency. This is because a sequence of different functions (UTM functions) will occur individually within a UTM appliance. This kind of approach is less flexible than the one in which all functions share the same information and enforcement mechanisms. To address this, Palo Alto has developed a Single-Pass Architecture that tackles these performance and flexibility issues with a unique single-pass approach to packet processing. This means that all operations are performed only once per packet by the single pass software. The Key processing task are the following:

- **Networking and Management Functionality**, the foundation of all traffic processing
- **User-ID**, maps IP addresses to Active Directory users and groups, to enable visibility and policy enforcement by user and group
- **App-ID**, Application identification, combination of application signatures, protocol detection and decryption, protocol decoding and heuristics
- **Content-ID,** Scans traffic for data, threats and URL categorization

- **Policy Engine,** Based on the networking, management, User-ID, App-ID, Content-ID information, the policy engine enforces a single security policy to traffic

With these features, Palo Alto has made it possible to add multiple key security functions to their next generation firewall instead of adding another physical security device (Palo Alto 2018b).

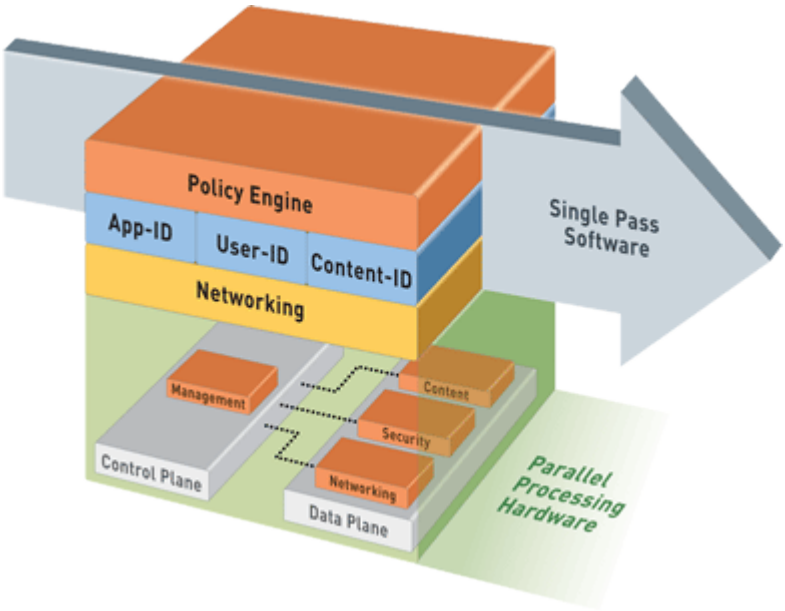

Figure 10. Palo Alto Single Pass Parallel Processing Architecture (Palo Alto 2018c)

3.4.2 Platforms

Palo Alto firewall portfolio offers a various range of products from a small remote site usage into the heart of a data center. In addition to physical firewalls, all the same functionalities are available if an organization chooses to use virtual firewalls instead. Virtual firewalls designed for hypervisors have a bit restricted features (i.e routing), as the main purpose of these devices is slightly different. The only supported interface type for VM series firewall running on any hypervisor is a virtual wire. Virtual wire logically binds two interfaces together (so no switching or routing takes place) and it requires no changes to adjacent network devices. For example, interfaces ethernet1/1 and ethernet1/2 are bound together through a virtual wire and they use the NetX dataplane API to communicate with the hypervisor. Palo Alto's Panorama is the centralized management and logging tool for all Palo Alto devices.

4 VMware Software-Defined Data Center

According to VMware, in order to be able to meet today's new challenges regarding IT efficiency and performance, organizations must virtualize their data centers. By doing that, all infrastructure services become as easy to provision, manage and scale as virtual machines. VMware Software-Defined Data Center (SDDC) architecture enables a fully automated and zero-downtime infrastructure for any application or service. SDDC components can be deployed all together or in single units. These components consist of the following:

- Compute Virtualization
  - vSphere ESX
  - vSphere vCenter
- Network Virtualization
  - NSX
- Software-Defined Storage Technologies
  - vSAN
  - Site Recovery Manager
  - Virtual Volumes
- Automated Management Framework (VMware Cloud Management Solutions)
  - vRealize Automation
  - vRealize Orchestrator
  - vRealize Operations Management
  - vRealize Log Insight
- VMware, Partner or Community Solutions

All of these SDDC architecture components can be implemented to private, public or hybrid cloud. SDDC also makes it possible to easily create a replica from organization's current production environment for testing and development purposes. This Thesis focuses only on compute and network virtualization part of the VMware portfolio (VMware 2018a).

4.1 Server Virtualization

Server virtualization has been around now for more than a decade; however, still many organizations consider it as a new technology. However, those who have already adopted it could not live without it. Virtualizing servers has been a game changer providing efficiency, availability and capabilities that would not be possible with physical world's limitations.

VMware vSphere is a virtualization platform that enables organizations to invent new products, services and new business models faster than ever. Most of the physical servers are operating at less than a 15 percent of capacity, especially when running a single service or application per server. Even if this is a waste of resources in server perspective, it is still understandable in segmentation and security perspective. Running a single service or application per server affects directly the number of physical servers required and leads to a management complexity. Server virtualization addresses this inefficiency and provides additional availability, scalability and security. Each virtualized service or application and its operating system (OS) are encapsulated in a separate, isolated software container called virtual machine (VM). Multiple VMs can run simultaneously on a single physical server and all the VMs have access to the underlying server's computing resources, putting the majority of hardware resources into effective use. Of course, some servers are not reasonable to virtualize, for example due to some license or support agreements, which do not permit virtualization (VMware 2018b).

4.2 Network Virtualization

VMware's network virtualization started in 2002 with introduction of the first virtual switch (vSwitch) to the ESX hypervisor. VMware continued to develop virtualization technologies and the primary goal was the realization of a complete virtualization platform for the data center. The goal of network virtualization is the same as in server virtualization; reproduction of the physical network/server in a software but with increased availability and security (decoupling hardware and software). Applications should run exactly the same way on a virtual network as on a physical network. Network virtualization presents logical network and security services to connected workloads:

- Logical ports
- Switches
- Routers
- Firewalls
- Load Balancers
- SSL-VPN

All these components run independent of the physical hardware (underlying network). Administrators can create and provision these components within minutes, rather than days or weeks which is the usual case with physical equipment (VMware 2018b & 2018d).

4.3 NSX Overview

To strengthen its software defined data center (SDDC) strategy, VMware acquired a network virtualization company called Nicira in August 2012. Nicira's network virtualization platform (NVP) makes it possible to dynamically create virtual network infrastructure and services that are separated and independent from the underlying network. VMware's own SDDC product was called "vCloud Networking and Security", however, after acquiring Nicira these two products were combined, which was the starting point of VMware NSX. NSX brings a number of unique advantages, that are extremely challenging for competitors to replicate, the primary advantage being microsegmentation, which significantly reduces the attack surface and vectors if utilized reasonable along with other security methods. Another advantage is that it allows organizations to extend their workloads or services to public cloud with the same level of security and control that they currently have in their own data center or private cloud. Another advantage is VXLAN capability, which is completely independent from physical network. NSX has currently two different versions, "NSX-V" (NSX for vSphere) and "NSX-T" (NSX for Transformers). This thesis focuses only on NSX-V, which is later referred to as NSX only (VMware 2018c).

4.4 NSX Architecture and Components

VMware NSX relies on a set of components that allow network and security virtualization systems to run successfully. These components along with reliable physical underlay network must be designed and deployed with care for the system to work properly. NSX has the following required components:

- vCenter Server, management component and interface for all NSX related products
- NSX Manager, responsible for the Management Plane
- NSX Controller Cluster, responsible for the Control Plane
- ESX Cluster (NSX prepared), responsible for the Data Plane
- Distributed Logical Router, responsible for VXLAN routing

- NSX Edge, provides routing between VLAN and VXLAN as well as FW/LB/SSL-VPN services

NSX installation requires a vCenter Server, because the NSX Manager must be registered to a vCenter Server. In addition, vCenter is the only solution that can provide some other required functionalities or features like Distributed Virtual Switch (DVS) and ESX clusters. NSX Manager is a virtual appliance (deployed from OVA) that can be accessed through a web interface to manage the appliance configuration. It is also the entry point of the REST API. All the switching, routing and security configurations of the logical network can only be performed using vCenter Web Client (legacy client is not supported). NSX Controllers are Virtual Machines responsible for NSX Control Plane: they handle all the logical networking functions. Three Controllers must be deployed for high availability purposes (best practice is to have dedicated hosts in different racks to minimize the failure domain size). NSX Controllers must be able to connect to the NSX Manager; however, they do not require Layer 2 connectivity. NSX Manager deploys VMware Infrastructure Bundles (VIBs) to the ESX hosts in a cluster. These VIBs run in the vmkernel and are responsible for the following functions:

- Logical Switching
- Distributed Routing
- Distributed Security

VMware recommendation is to run a separate management cluster for vCenter Server (and its required components), NSX Manager and Controllers, a compute cluster for end user Virtual Machines and NSX Edge cluster for Edge Service Gateway and Distributed Logical Router Control VMs. From a network perspective, VMware NSX works with the existing three-layer network design. However, Leaf and Spine underlay architecture is preferred. All the NSX capabilities are independent of the underlying network and they work with any reliable IP network that is configured to support MTU size of 1600. NSX does not currently support fragmentation, so creating a VXLAN tunnel between a private and a public cloud over Internet is not usually possible due to MTU requirements. A VXLAN port group is automatically created on an ESX host when NSX is deployed to a cluster. This port group is then used by the

VTEP. VMware NSX VXLAN data replication between ESX hosts supports the following methods:

- Multicast
- Unicast
- Hybrid

If multicast mode is chosen, the underlying network is responsible for passing multicast traffic between the VTEPs. The unicast replication mode uses unicast traffic to send information to all the VTEPs. The only configuration required in unicast mode is to deploy the NSX Controllers. However, unicast mode does not scale very well and should only be used in a lab or proof of concept (PoC) environments. The Hybrid mode replication is the default and recommended. It uses multicast replication within same Layer 2 segment and unicast to go across Layer 3 devices (VMware 2018e).

4.5 NSX With Palo Alto

NSX can be integrated with multiple 3rd party products, however, in this thesis Palo Alto is the chosen vendor and solution. VMware and Palo Alto have collaborated on an integrated offering to enable companies to realize the full potential of the SDDC. The NSX built-in firewall has only capabilities to inspect traffic up to Layer 4, whereas Palo Alto's firewall provides a full Layer 7 inspection features. However, the NSX distributed firewall works in the kernel and can provide up to line-rate performance compared to 3rd party products that have to work through NSX API, which has some limitations regarding a maximum throughput per session.

In addition to vCenter Server and NSX Manager, Palo Alto's Panorama server is a required component. Panorama serves as the central point of management for the Palo Alto VM-series firewalls (VM-FW) running on NSX. Whenever a new VM-FW is deployed in NSX, it automatically communicates with Panorama to obtain a license and receives all the necessary configurations and policies. The REST-based XML API integration is the key component that enables Panorama to synchronize with the NSX Manager and the firewalls to allow the use of dynamic address groups. Dynamic address groups are then used in the security policies. They allow the creation of

policies that automatically adapt to changes – adds, moves or deletes servers (or virtual machines generally). It also enables the flexibility to apply multiple different rules to a single server based on tags that define its role on the network, operating system or what kind of traffic it processes. For example, a newly created VM with a tag “Web” might automatically get a policy with the following statements:

- New VM is allowed to communicate to a certain Microsoft SQL database server using application “mssql-db” (instead of just TCP port 1433)
- New VM is reachable from Internet using application “web-browsing” (instead of just TCP port 80)
- New VM is applied as a source address in a NAT rule, which does a bi-directional static source IP translation and is now reachable from Internet with a public IP address (assuming, that the VM received a private IP address via DHCP)

The Palo Alto VM series firewall for NSX strictly focuses on securing East-West traffic inside a data center. Combining Palo Alto VM series and perimeter firewalls (all managed by Panorama) can considerably increase the overall security and visibility inside a data center.

5 Research

The thesis research was conducted in Cygate's lab premises using currently available hardware owned by Cygate. The environment was built to support all the requirements that modern data center designs may have. Despite these many requirements, the idea was to keep the topology and the number of devices as simple as possible. The main focus of this research was on scenarios 1 and 2 (overlay technologies and microsegmentation) and scenarios 3 and 4 were added just for light evaluation and demonstration of NSX capabilities and performance in a throughput perspective.

5.1 The Lab Environment

The lab environment for this thesis was built using the following assets (more detailed device configurations and show command outputs can be found in Appendices 1 and 2):

- Two Juniper MX80 Routers, with Junos version 15.1R6.7
  - labiaasmx01 and labiaasmx02
- Two Juniper EX3300 Switches, with Junos version 14.1X53-D46.7
  - cdciaassw01 (SW01) and cdciaassw02 (SW02)
- Two Palo Alto 3050 Firewalls, with PANOS 8.0.8
  - cdciaasfw01a and cdciaasfw01b
- Two Palo Alto VM300 Firewalls, with PANOS 8.0.8
  - nsxlab-fw01 (1) and nsxlab-fw01 (2)
- Four Dell Servers running VMware ESX 6.0.0 build 6921384
  - iaasesx01, iaasesx02, iaasesx03 (ESX03) and iaasesx05 (ESX05)
- One Ixia ixChariot Server, version 9.5.14.13
  - ixchariot.lab.cygate.fi
- Four Ixia ixChariot Endpoints (running on CentOS 6.8), version 9.5
  - nsx-test-vm01, nsx-test-vm02, nsx-test-vm03 and nsx-test-vm04
- VMware vCenter Server Appliance 6.0.0.30400
  - nsxvc01.lab.cygate.fi (this VM was running on a separate lab environment)
- VMware NSX Manager 6.2.8
  - nsxmgr01.lab.cygate.fi
- Three NSX Controllers

All the physical connections in the lab use 1Gbit copper cabling. However, some of the connections are bundled together using LACP. ESX hosts are connected to DC local switch with multiple interfaces (dedicated interfaces for both scenarios, all

these individual interfaces are not visible in the Figure 11). All the devices were managed via out of band management network, which is not shown in the diagram.

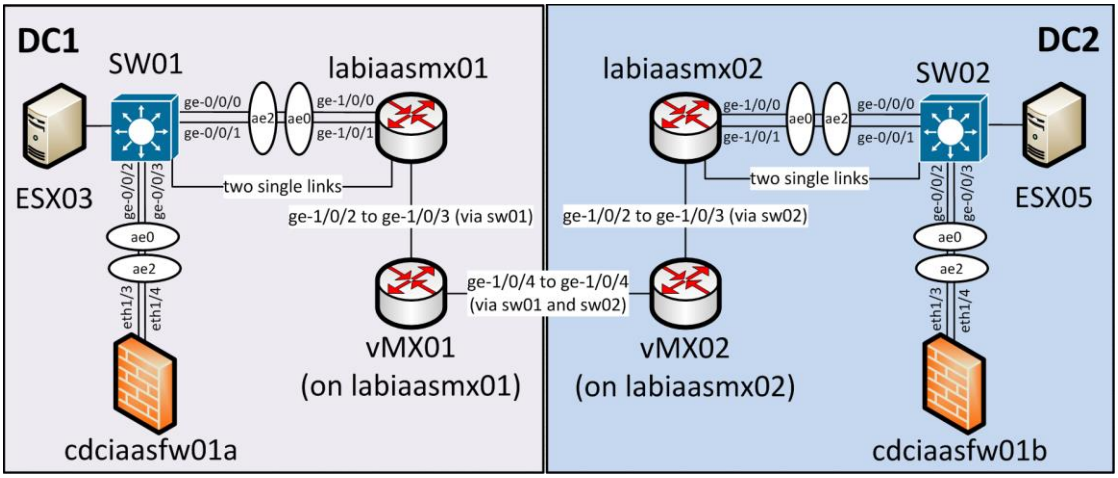

Figure 11. Lab Physical Topology

VMware ESX hosts ESX03 and ESX05 were compute nodes, one in each data center. All the management components (VMware vCenter, NSX components etc.) were running on a separate management cluster. Both ESX hosts for computing had two virtual machines, one for each scenario.

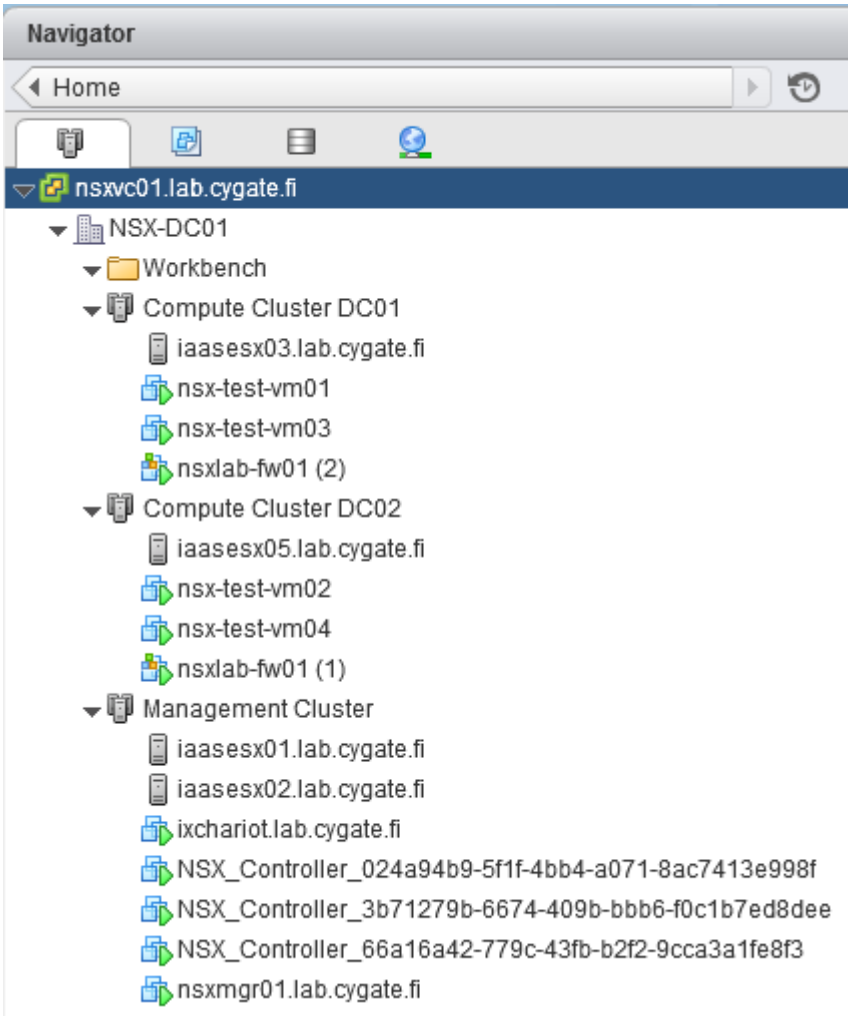

Figure 12. Lab ESX hosts and VMs in vCenter

The MPLS backbone that was used in both scenarios, was built using two Juniper MX routers (MX01 and MX02 in Figure 13.) and each of them was running an additional logical instance (vMX01 and vMX02 in Figure 13.) to create an MPLS core total of four routers. All the connections between routers were implemented via switches using dedicated physical interfaces, so it would have been possible to add any device in the wire (for packet capture or to create latency for example). Router roles in MPLS network are listed as follows:

- MX01, Provider Edge Router at DC1
- vMX01, Provider Router
- vMX02, Provider Router
- MX02, Provider Edge Router at DC2

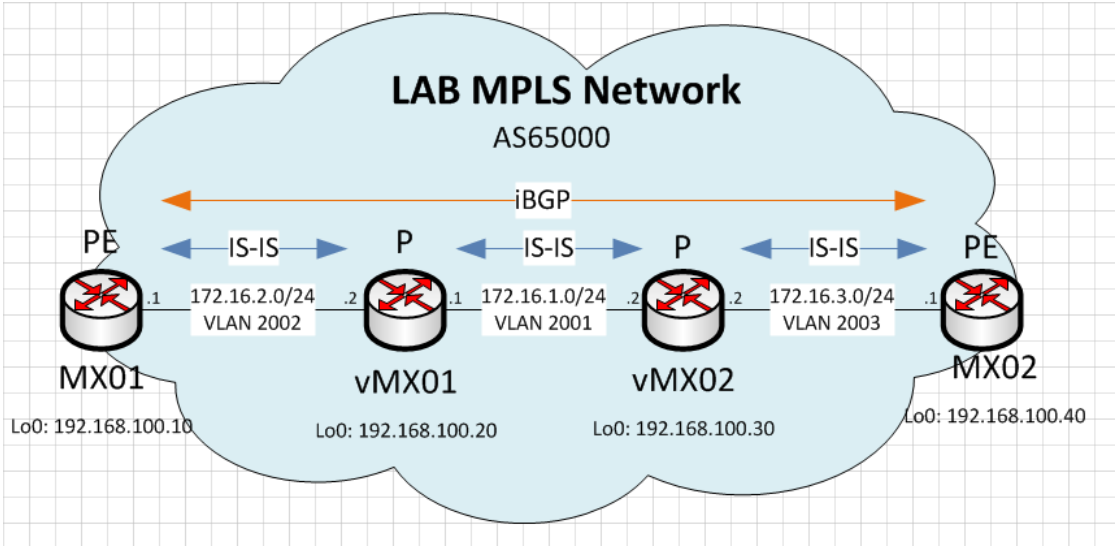

Figure 13. Lab MPLS Network

IS-IS routing was used for IGP, advertising loopback IPs and transit networks between routers. LDP was responsible for MPLS label exchange between routers. An iBGP session with MP-BGP capabilities was established between PE routers to exchange information.

Firewall policies on physical Palo Alto appliance cdciaasfw01 and on virtual Palo Alto appliances nsxlab-fw01 (1) and nsxlab-fw01 (2) were identical. The only policy in place allowed all traffic, with any source/destination/service/application. Logging was enabled at session end and the security profile in use had the following settings:

- Antivirus, action alert
- Anti-Spyware, alert for all severities (with packet capture set to single-packet)
- Vulnerability Protection, alert for all severities (with packet capture set to single-packet)
- URL Filtering, alert for all categories

The NSX distributed firewall had only rule, which allows all traffic and logs the session at the end.

ESX hosts shared the same storage cluster, which was implemented using iSCSI and dedicated interfaces on ESX hosts. This simulated active/active storage would have made it possible to create a stretched ESX cluster between data centers.

All the tests were performed between two endpoints that were running Ixia's IxChariot Endpoint software. The endpoints communicate and report to Ixia IxChariot

Server using a dedicated out of band management network that is not shown in the lab network diagrams. Total of six different Ixia IxChariot test sets were defined and implemented in this lab. Each test duration was two minutes.

Test 1: TCP Low Performance

- TCP, 16KB chunks, Data Rate: Unlimited, 1 User
- Skype-VoIP-G711, 1 User

The purpose of Test 1 is to find out the overall throughput using 16KB TCP chunks and Skype-VoIP-G711 call for latency and jitter measurement.

Test 2: TCP Baseline Performance

- TCP, 32KB chunks, Data Rate: Unlimited, 1 User
- Skype-VoIP-G711, 1 User

The purpose of Test 2 is to find out the overall throughput using 32KB TCP chunks and Skype-VoIP-G711 call for latency and jitter measurement.

Test 3: TCP High Performance

- TCP 64KB chunks, Data Rate: Unlimited, 1 User
- Skype-VoIP-G711, 1 User

The purpose of Test 3 is to find out the overall throughput using 64KB TCP chunks and Skype-VoIP-G711 call for latency and jitter measurement.

Test 4: Mixed UDP and TCP Performance

- TCP, 64KB chunks, Data Rate: 20 Mbps, 15 Users
- TCP, 64 Byte chunks, Data Rate: 1 Mbps, 15 Users
- UDP, 32KB chunks, Data Rate: 20 Mbps, 15 Users
- Skype-VoIP-G711, 1 User

The purpose of Test 4 is to demonstrate “real life traffic” (Layer 4 traffic only) with different sized UDP and TCP chunks with limited bandwidth and limited number of users. Skype-VoIP-G711 call is used again for latency and jitter measurement.

Test 5: Application Mix

- Facebook, 5 Users
- Gmail, 5 Users
- Youtube, 10 Users
- Twitter, 5 Users
- SMTP, 5 Users
- Outlook, 5 Users
- HTTP, 14 Users
- Skype-VoIP-G711, 1 User

The purpose of Test 5 is to demonstrate real life traffic (Layer 7) usage with multiple users and Skype-VoIP-G711 call for latency and jitter measurement. The first four test were just dummy UDP or TCP traffic, however these applications behave as a real user used them. Palo Alto firewalls recognize these applications correctly.

Test 6: UDP Small Packets Performance

- UDP, 64 Byte chunks, Data Rate: Unlimited, 1 User
- Skype-VoIP-G711, 1 User

The purpose of Test 6 is to find out the overall throughput using 64 Byte UDP chunks and Skype-VoIP-G711 call for latency and jitter measurement. This kind of traffic is closer to a DDoS than a real-life traffic and is very resource intensive for devices to handle.

5.2 Scenario 1: DCI with EVPN-MPLS

The first scenario simulates an enterprise network that has two data centers (DC1 and DC2), and Data Center Interconnect is implemented with EVPN using enterprise owned MPLS core. It would have been possible to achieve the same functionality with VPLS; however, EVPN was chosen as the modern overlay technology. Internet connection is not implemented in this lab (as it provides no added value), yet the most logical place for that would be for example behind the firewall using a dedicated VRF on MPLS core.

Data center LANs are pure Layer 2, so no IP fabric is in use. VMware ESX and firewall cluster nodes as well as routers have been placed in both data centers to provide high availability in case of a single device or even a whole data center failure. All

three networks (RTR-FW-Transit, DC-VM-Network01 and 02 in figure 14.) have been spanned between data centers using EVPN/MPLS.

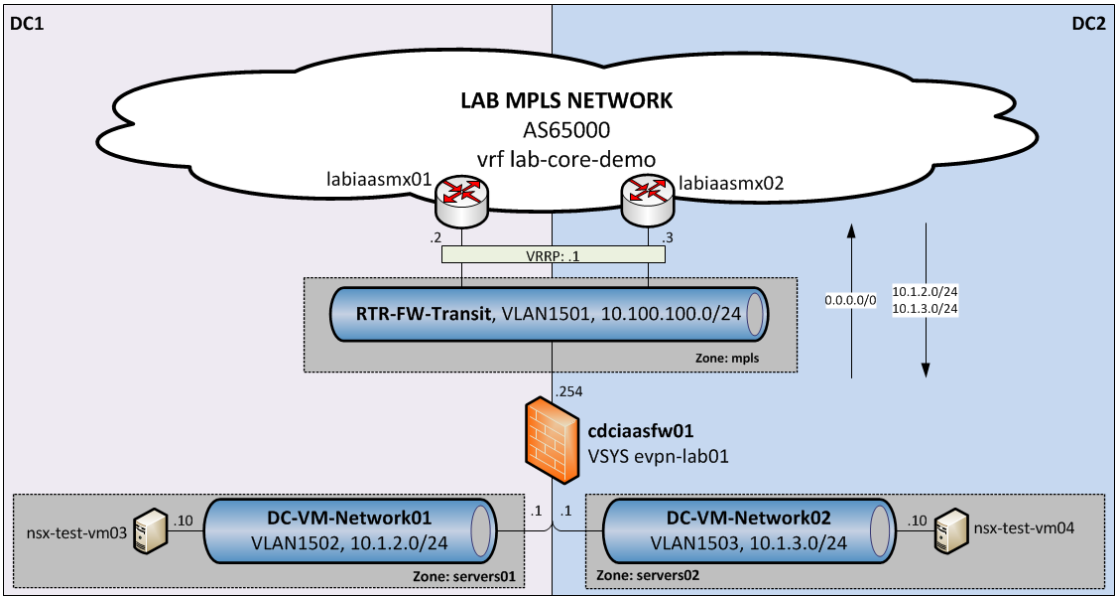

Figure 14. Scenario 1 Logical Topology

Virtual machines have been placed in a different Layer 3 networks so firewall policies can be applied for all the traffic. The firewall cluster is an Active/Standby solution and the active node is located in DC1. This means that all the traffic to and from nsx-test-vm04 (located in DC2) must travel via DC1 where the default gateway is located. This is achieved using EVPN (EVPN instance “evpn-cdc-dci” in Lab MPLS Core). In an overlay technology perspective (EVPN), the control plane is MP-BGP and the data plane is MPLS labels in this scenario. The following list contains the addressing of the main components used in this scenario:

nsx-test-vm03 in DC1: IP address 10.1.2.10/24, MAC address 00:50:56:A7:C2:D2

nsx-test-vm04 in DC2, IP address 10.1.3.10/24, MAC address 00:50:56:A7:A7:3E

cdciaasfw01, VLAN 1502, IP address 10.1.2.1/24, MAC address 00:1b:17:00:01:30

cdciaasfw01, VLAN 1503, IP address 10.1.3.1/24, MAC address 00:1b:17:00:01:30

5.3 Scenario 2: DCI with NSX VXLAN and Palo Alto Microsegmentation

The second scenario has again two data centers (DC1 and DC2), however, this time each data center has completely independent Layer 3 networks. This simulates an

enterprise network that has been implemented using IP Fabric (one fabric that is located in both data centers or one fabric per data center). To be able to stretch networks inside VMware ESX hosts into both data centers, overlay technology such as VXLAN could be used. In this lab, the VXLAN is implemented using VMware NSX and firewalling is done using distributed hypervisor firewalls (so physical firewalls are not implemented). The distributed firewall in use is a Palo Alto VM-300 per ESX host. VXLAN tunnel (VNI 100500) was established from VTEP on host ESX03 to VTEP on host ESX05 via MPLS Core (VRF "lab-core-nsx" in Figure 15). VTEP replication between ESX hosts was done using unicast mode. This lab environment did not include Internet connection, so neither distributed logical router (DLR), NSX Edge router or physical firewalls for the perimeter were deployed. In an overlay technology perspective (VXLAN), the control plane is NSX Controllers and the data plane is IP via MPLS L3VPN in this scenario.

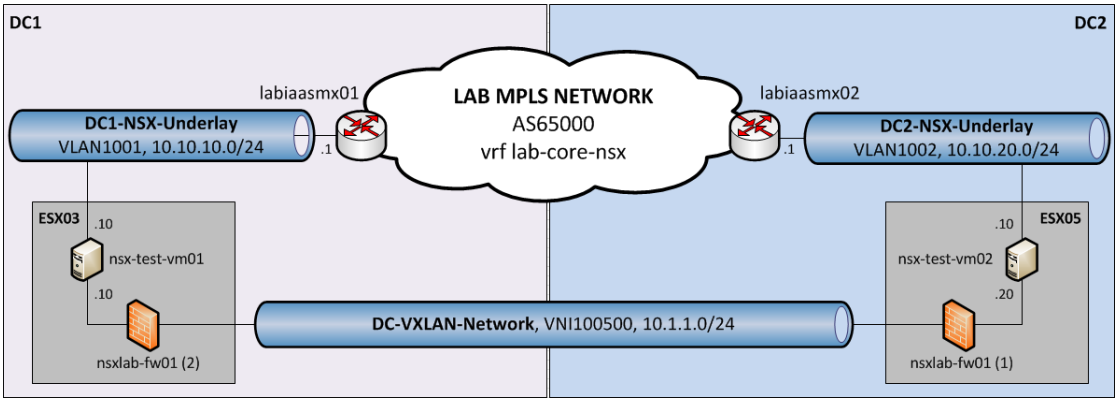

Figure 15. Scenario 2 Logical Topology

Virtual machines are running on a separate ESX hosts in a different data centers. Despite this, VMs are in a same network segment and the same firewall policy is applied to all traffic using microsegmentation. The following list contains the addressing of the main components used in this scenario:

nsx-test-vm01: IP address 10.1.1.10/24, MAC address 00:50:56:A7:C2:AF

nsx-test-vm02, IP address 10.1.1.20/24, MAC address 00:50:56:A7:9B:57

iaasesx03, VTEP IP address 10.10.10.10/24, MAC address 00:50:56:6C:72:14

iaasesx05, VTEP IP address 10.10.20.10/24, MAC address 00:50:56:66:E8:77

5.4 Scenario 3: DCI with NSX VXLAN and Microsegmentation

The third scenario is exactly the same as in the second, with the exception that distributed firewalling is achieved using NSX firewall instead of Palo Alto's. This was implemented so the impact on performance using 3rd party firewall product for microsegmentation could be compared to the VMware NSX native firewall capability.

5.5 Scenario 4: NSX VXLAN and Microsegmentation in a single DC

The fourth scenario is identical to the third, however, this time both Virtual Machines (nsx-test-vm01 and nsx-test-vm02) reside on same ESX host within a single data center. This was implemented to demonstrate the performance that can be achieved between two virtual machines on a single ESX host. Compared to the traditional ESX setup with only two VMs, firewall policy is enforced for all the communication between the VMs. These virtual machines could be located in a different network segments (in another VXLAN for example), however this provides no benefit as any firewall policy can be enforced for Virtual Machines within the same network segment using microsegmentation.

6 Evaluation of Results

6.1 Background

The testing was performed using Ixia's ixChariot server and endpoint clients. All six different test sets were executed for each scenario (scenarios from A to D were explained in more detail in Chapter 5). The main purpose of the Ixia IxChariot tool is to quickly evaluate something that was just built to find out the possible limitations and bottlenecks, before entering to a production state. For a continuous testing or monitoring of the network, many other Ixia products are more suitable.

These scenarios do not directly compare modern overlay technologies versus traditional designs or architectures; nonetheless they show and prove the functionality itself. It is good to keep in mind that the physical restrictions in the lab environment was one Gbps, due to physical cablings used. The Juniper EX3300 series switches that were implemented, have only a 3MB buffer size per Packet Forwarding Engine (PFE), so they are not designed for data center usage. This most likely caused some packet loss on some test cases. Other physical devices (Juniper MX80 routers and Palo Alto 3050 firewalls) that were used should not have any difficulties under these test cases (in throughput perspective).

6.2 The Results

This Chapter presents all the test results in a table view, one test at a time.

Table 3. TCP Low Performance Results

	TP Min	TP Max	TP Avg	Bytes TX	Bytes RX	Bytes Lost	Jitter Avg	1-Way Delay Avg
A	703 Mbps	928 Mbps	920 Mbps	13.78 GB	13.78 GB	0 B	0 ms	15.8 ms
B	400 Mbps	419 Mbps	410 Mbps	6.149 GB	6.149 GB	480 B	0.08 ms	12.5 ms
C	407 Mbps	417 Mbps	409 Mbps	6.146 GB	6.146 GB	0 B	0 ms	0.1 ms
D	263 Mbps	16.9 Gbps	12.4 Gbps	186.3 GB	186.3 GB	0 B	0 ms	0.1 ms

The first test shows that scenario A throughput average is very close to the theoretical maximum, which is one Gbps. As the average throughput in scenarios B and C is about the same (which is still only 50% compared to case A), the bottleneck

is not the Palo Alto VM300 firewall used in scenario B. However, the use of Palo Alto VM300 increases the latency remarkable. The lower performance in NSX scenarios might be due to insufficient ESX server performance (CPU). Scenario D performance compared to others is overwhelming, as the packets do not need to leave the physical host at all.

Table 4. TCP Baseline Performance Results

	TP Min	TP Max	TP Avg	Bytes TX	Bytes RX	Bytes Lost	Jitter Avg	1-Way Delay Avg
A	716 Mbps	931 Mbps	920 Mbps	13.8 GB	13.8 GB	160 B	0 ms	19.6 ms
B	370 Mbps	416 Mbps	408 Mbps	6.126 GB	6.126 GB	480 B	0.03 ms	11.7 ms
C	406 Mbps	418 Mbps	410 Mbps	6.152 GB	6.152 GB	3 KB	0 ms	11.7 ms
D	8.5 Gbps	17.1 Gbps	13.9 Gbps	208.7 GB	208.7 GB	0 B	0 ms	0.1 ms

The second test results are very similar to the first test. The only difference is that the latency is now on the same level on scenarios B and C.

Table 5. TCP High Performance Results

	TP Min	TP Max	TP Avg	Bytes TX	Bytes RX	Bytes Lost	Jitter Avg	1-Way Delay Avg
A	649 Mbps	930 Mbps	919 Mbps	13.786 GB	13.786 GB	0 B	0.02 ms	22.8 ms
B	373 Mbps	414 Mbps	409 Mbps	6.130 GB	6.130 GB	0 B	0.02 ms	1.0 ms
C	406 Mbps	419 Mbps	410 Mbps	6.152 GB	6.152 GB	5.2 KB	0 ms	11.9 ms
D	605 Mbps	16.8 Gbps	12.0 Gbps	180.9 GB	180.9 GB	0 B	0 ms	0.1 ms

The third test shows the same pattern as the first two tests, nothing anomalous.

Table 6. Mixed UDP and TCP Performance Results

	TP Min	TP Max	TP Avg	Bytes TX	Bytes RX	Bytes Lost	Jitter Avg	1-Way Delay Avg
A	566 Mbps	616 Mbps	612 Mbps	9.179 GB	9.176 GB	2.6 MB	1.5 ms	4.8 ms
B	442 Mbps	640 Mbps	604 Mbps	9.095 GB	9.052 GB	42.2 MB	1.7 ms	11.6 ms
C	522 Mbps	548 Mbps	539 Mbps	8.104 GB	8.097 GB	7.82 MB	0.2 ms	1.9 ms
D	433 Mbps	630 Mbps	565 Mbps	8.536 GB	8.480 GB	55.5 MB	1.0 ms	14.6 ms

Mixed UDP and TCP Performance test generated a little more than 600 Mbps of total traffic. On scenarios A and B the average throughput is very close to this amount of

traffic. However, surprisingly scenarios C and D throughput is lower than in scenarios A and B. Quite heavy packet loss was most probably caused due to insufficient buffer sizes in the Juniper EX3300 switches.

Table 7. Application Mix Results

	TP Min	TP Max	TP Avg	Bytes TX	Bytes RX	Bytes Lost	Jitter Avg	1-Way Delay Avg
A	322 Mbps	393 Mbps	358 Mbps	5.367 GB	5.367 GB	0 B	0 ms	0.1 ms
B	182 Mbps	264 Mbps	237 Mbps	3.551 GB	3.551 GB	0 B	2.2 ms	13.3 ms
C	-	-	-	-	-	-	-	-
D	-	-	-	-	-	-	-	-

The license that was available in Ixia only allowed 50 simultaneous users, so the application test in throughput perspective is deficient. However, this would have been one of the best test scenarios, as all the application traffic is real traffic enforcing the firewalls to analyze the contents of each packet. This test was not implemented in scenarios C and D, as the NSX distributed firewall supports only Layer 4 inspection (whereas Palo Alto supports up to Layer 7).

Table 8. UDP Small Packets Performance

	TP Min	TP Max	TP Avg	Bytes TX	Bytes RX	Bytes Lost	Jitter Avg	1-Way Delay Avg
A	0.06 Mbps	111 Mbps	103 Mbps	1624 MB	1550 MB	73 MB	0.2 ms	1.9 ms
B	26.0 Mbps	58.1 Mbps	35.1 Mbps	1152 MB	527 MB	625 MB	2.4 ms	12.3 ms
C	61.7 Mbps	78.1 Mbps	74.8 Mbps	1137 MB	1122 MB	14 MB	0 ms	2.0 ms
D	14.2 Mbps	31.5 Mbps	23.0 Mbps	753 MB	345 MB	407 MB	0.5 ms	2.1 ms

The last test was about sending a very small UDP packets as many as possible, simulating a DDoS inside a data center. This kind of traffic is very resource intensive to handle, which is obvious from the results: low overall throughput and heavy packet loss in all scenarios. If this kind of an attack is originating from the Internet, it is vital to stop or suppress it at the perimeter before entering the data center itself.

7 Conclusions and discussion

7.1 Answering the Research Questions

Modern overlay technologies inside a data center enables the underlying network to outgrow from legacy Ethernet Fabrics to IP Fabrics. The optimal architecture for IP Fabric is a Leaf and Spine topology. Changing the legacy three-layer data center design (access, distribution and core) to Leaf and Spine changes the traffic flows in a way that makes the use of traditional perimeter firewalls and load balancers ineffective. There is still a vital role for perimeter firewall and load balancer, which is at the border of a data center. For the data center interconnect (DCI), modern overlay technologies enable intelligent MAC address learning and reduces the possibilities to stumble for Layer 2 risks and drawbacks.

When it comes to planning to change the underlying network, it all comes to the requirements. If the data center is relatively small and there are no strict SLA requirements, Ethernet Fabric might be reasonable choice. Larger data center deployments that may even span across the globe, IP Fabrics with overlay technology such as VXLAN responsible for Layer 2 connectivity is a mandatory choice.

The separation of underlay and overlay networking provides better opportunities for automation and orchestration. The underlying network is the foundation that should be implemented to be resilient, scalable and to provide high availability with minimum downtime in case of any failures. Overlay networks are then implemented in top of that foundation via use of automation (so called programmable network). Overlay networks are also easier to extend to a public cloud from a private data center (or private cloud). This kind of approach only requires standardized changes to be done for the underlying network, minimizing the risks as all more complex changes are targeted to the overlay network.

Changing the data center topology to Leaf and Spine solution is an ideal choice for microsegmentation as well, because it provides high throughput and low latency, which is not the case with traditional perimeter firewalls. With microsegmentation, all the server components participating the same service chain can be located within a single network segment, if that is intended.

7.2 Summary

Data centers designed using purely Ethernet Fabrics will eventually fail or suffer from many restrictions in today's strict competition. IP Fabric based solutions will take the dominion, as they significantly reduce the failure domain size and enable SDN like services via use of overlay technologies. It does not matter what type of a cloud organizations are using (private, public or hybrid), overlay technologies provide seamless integration between all of them. The reason why organizations are moving their services to the cloud is that it reduces the provision time of services and reduces IT costs. Private clouds (data centers) must provide the same agility and resiliency.

Storage and compute power have evolved exponentially in the last ten years, whereas network has not. Spine and Leaf topology with IP Fabric solution using ECMP can considerably narrow down this gap. The complete data center solution is of course not just about the network; however, it constructs the solid foundation and it is relatively easy for example to build the storage traffic to use IP instead of fiber channel. One identified challenge might be providing connectivity between bare metal servers and virtual servers, if the data center solution is full IP based solution with VXLAN as an overlay technology.

So who should consider migrating from Ethernet Fabric to IP Fabric inside a data center? The answer is: everyone running a data center business if the total amount of switches in use is two or more (creating a possibility for a loop). The change from legacy data center designs to modern architectures also influences the way in which data centers are operated. The physical cablings in IP Fabrics are somewhat more laborious than in the three-tier design. Network administrators are used to troubleshoot traditional Layer 2 issues such as MAC address tracing and capturing traffic from multiple points of switched network to be able to find the root cause; now they will have to change their habits and attitudes dramatically. Due to separation of underlay and overlay networking, an administrator needs to isolate the incident to either one and use different troubleshooting methods. In addition, some administrators might consider technologies used in IP Fabric more as a "service provider thing" and they may need to strengthen their skills. Overall, the whole IT

infrastructure is divided into different silos (storage, compute, and network) and they all seem to speak using different languages. Maybe the SDN could be the common language for all of them?

If an organization currently has its own MPLS core with L2VPN implementations, migration to EVPN-MPLS based solution is a relatively easy task to accomplish. However, the actual benefits from that migration are significant; one common control plane (MP-BGP) with intelligent MAC address handling eliminates most of the Layer 2 fundamental issues. If an organization is not using MPLS now and plans to extend a data center to another location, they should really weigh the possibility to use pure IP based services like VXLAN as it makes troubleshooting and configurations simpler.

The advantages of microsegmentation are unquestionable; fictitious situation where a service has been implemented using the three-tier architecture (Web, App and DB) and a web server is running Apache that has a 0-day exploit. Next thing one knows, some other party suddenly manages all one's servers in that segment because they were able to use SSH protocol to connect (another 0-day exploit in SSH or brute force attack) to servers next to one's Apache server. Unless, microsegmentation is implemented denying SSH traffic within that segment. Regardless of the microsegmentation, traditional perimeter firewalling still has its place inside a data center: its primary function is to be a gatekeeper between the public internet and your data center, stopping all the non-legitimate traffic and volumetric or session-based attacks before entering the data center.

References

Check Point. 2018, Check Point, Our History. Accessed 7.1.2018. Retrieved from https://www.checkpoint.com/about-us/our-history/

Gantz, J. & Reinsel D., The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the far East. Accessed 23.1.2018. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Hanks, D. 2016. Juniper QFX10000 Series. A Comprehensive Guide to Building Next-Generation Data Centers.

Information Week. 2018. The Evolution of Firewalls Past Present and Future. Accessed 12.1.2018. Retrieved from https://www.informationweek.com/partner-perspectives/the-evolution-of-firewalls-past-present-and-future/a/d-id/1318814?

Juniper. 2018a. Juniper, EVPN Multihoming Overview. Accessed 7.1.2018. Retrieved from https://www.juniper.net/documentation/en_US/junos/topics/concept/evpn-bgp-multihoming-overview.html

Juniper. 2018b. Juniper, MPLS Overview. Accessed 7.1.2018. Retrieved from https://www.juniper.net/documentation/en_US/junos/topics/concept/mpls-security-overview.html

Juniper. 2018c. Juniper VPN Types. Accessed 8.2.2018. Retrieved from https://www.juniper.net/documentation/en_US/junos/topics/concept/vpn-types.html

Juniper. 2018d. Advanced Data Center Switching Student Guide Volume 2. Course Material.

Juniper. 2018e. Understanding EVPN Pure Route Type-5 on QFX Switches. Accessed 28.4.2018. Retrieved from https://www.juniper.net/documentation/en_US/junos/topics/concept/evpn-route-type5-understanding.html

Metzler, J. Kubernan Brief, Vol.1, number 6. Next Generation Firewalls – The Policy and Security Control Point. Accessed on 12.1.2018. Retrieved from https://www.paloaltonetworks.com/content/dam/pan/en_US/assets/pdf/white-papers/kubernan-notes-vol1-no6.1.pdf

Sanchez-Monge A., Szarkovicz K. G., 2015. MPLS in the SDN Era.

Palo Alto. 2018a. Palo Alto, Our Company. Accessed 7.1.2018. Retrieved from https://www.paloaltonetworks.com/company

Palo Alto. 2018b. Palo Alto, Single-Pass Architecture. Accessed 7.1.2018. Retrieved from https://www.paloaltonetworks.com/technologies/single-pass-architecture

Palo Alto. 2018c. Palo Alto, The Four Key Elements of Security for the Software-defined Data Center. Accessed 28.4.2018. Retrieved from https://researchcenter.paloaltonetworks.com/2015/06/the-four-key-elements-of-security-for-the-software-defined-data-center/

ResearchGate. 2018. MPLS Shim Header. Accessed 8.2.2018. Retrieved from https://www.researchgate.net/figure/MPLS-shim-header_fig2_42803342

RFC3031. 2018a. Multiprotocol Label Switching Architecture. Accessed 8.2.2018. Retrieved from https://tools.ietf.org/html/rfc3031

RFC7348. 2018b. Virtual eXtensible Local Area Network. Accessed 2.3.2018. Retrieved from https://tools.ietf.org/html/rfc7348

RFC7047. 2018c. The Open vSwitch Database Management Protocol. Accessed 24.3.2018. Retrieved from https://tools.ietf.org/html/rfc7047

RFC7423. 2018d. BGP MPLS-based Ethernet VPN. Accessed 24.3.2018. Retrieved from https://tools.ietf.org/html/rfc7432

Southwick, P., Marschke, D. & Reynolds, H. 2011. Junos Enterprise Routing.

Tech Republic. 2018. Understand the Evolution of Firewalls. Accessed 12.1.2018. Retrieved from https://www.techrepublic.com/article/understand-the-evolution-of-firewalls/

VMware. 2018a. VMware, SDDC Getting Started. Accessed 12.1.2018. Retrieved from https://code.vmware.com/sddc-getting-started

VMware. 2018b. VMware, Virtualization. Accessed 12.1.2018. Retrieved from https://www.vmware.com/solutions/virtualization.html

VMware. 2018c. VMware, VMware and Nicira. Accessed 12.1.2018. Retrieved from https://www.vmware.com/company/acquisitions/nicira.html

VMware. 2018d. VMware, The History of NSX and the Future of Network Virtualization. Accessed 12.2.2018. Retrieved from https://www.vmware.com/radius/history-nsx-future-network-virtualization/

VMware. 2018e. VMware, NSX: Design And Deploy. Lecture Material NSX 6.2

Appendices

Appendix 1. Device show commands

Scenario 1

```
[root@nsx-test-vm03 ~]# arp -a -i eth1 -n
? (10.1.2.1) at 00:1b:17:00:01:30 [ether] on eth1

[root@nsx-test-vm04 ~]# arp -a -i eth1 -n
? (10.1.3.1) at 00:1b:17:00:01:30 [ether] on eth1

cyadmin@cdciaassw01> show interfaces descriptions
Interface      Admin Link Description
ge-0/0/14      up    up   cdciaasesx03, vmnic5
ge-0/0/22      up    up   cdciaassw02, ge-0/0/23
ae0          up    up   cdciaasfw01, ae1
ae2          up    up   cdciaasmx01, ae0

cyadmin@cdciaassw01> show ethernet-switching table vlan vlan1501
Ethernet-switching table: 2 unicast entries
  VLAN          MAC address       Type       Age Interfaces
  vlan1501        *               Flood        - All-members
  vlan1501        00:00:5e:00:01:0f Learn        0 ae2.0
  vlan1501        00:1b:17:00:01:30 Learn       33 ae0.0

cyadmin@cdciaassw01> show ethernet-switching table vlan vlan1502
Ethernet-switching table: 2 unicast entries
  VLAN          MAC address       Type       Age Interfaces
  vlan1502        *               Flood        - All-members
  vlan1502        00:1b:17:00:01:30 Learn        0 ae0.0
  vlan1502        00:50:56:a7:c2:d2 Learn        0 ge-0/0/14.0

cyadmin@cdciaassw01> show ethernet-switching table vlan vlan1503
Ethernet-switching table: 2 unicast entries
  VLAN          MAC address       Type       Age Interfaces
  vlan1503        *               Flood        - All-members
  vlan1503        00:1b:17:00:01:30 Learn        0 ae0.0
  vlan1503        00:50:56:a7:a7:3e Learn        0 ae2.0

cyadmin@cdciaassw02> show interfaces descriptions
Interface      Admin Link Description
ge-0/0/14      up    up   cdciaasesx05, vmnic5
ge-0/0/23      up    up   cdciaassw01, ge-0/0/22
ae0          up    up   cdciaasfw02, ae1
ae2          up    up   cdciaasmx02, ae0

cyadmin@cdciaassw02> show ethernet-switching table vlan vlan1501
Ethernet-switching table: 2 unicast entries
  VLAN          MAC address       Type       Age Interfaces
  vlan1501        *               Flood        - All-members
  vlan1501        00:00:5e:00:01:0f Learn        0 ge-0/0/8.0
  vlan1501        00:1b:17:00:01:30 Learn        0 ae2.0

cyadmin@cdciaassw02> show ethernet-switching table vlan vlan1502
Ethernet-switching table: 2 unicast entries
  VLAN          MAC address       Type       Age Interfaces
  vlan1502        *               Flood        - All-members
  vlan1502        00:1b:17:00:01:30 Learn        0 ae2.0
  vlan1502        00:50:56:a7:c2:d2 Learn     2:45 ae2.0

cyadmin@cdciaassw02> show ethernet-switching table vlan vlan1503
Ethernet-switching table: 2 unicast entries
  VLAN          MAC address       Type       Age Interfaces
  vlan1503        *               Flood        - All-members
  vlan1503        00:1b:17:00:01:30 Learn        0 ae2.0
  vlan1503        00:50:56:a7:a7:3e Learn        0 ge-0/0/14.0

cyadmin@cdciaasmx01> show interfaces descriptions
Interface      Admin Link Description
```

```
ae0          up    up   cdciaassw01, ae2

cyadmin@cdciaasmx01> show evpn instance evpn-cdc-dci extensive
Instance: evpn-cdc-dci
  Route Distinguisher: 192.168.100.10:5000
  Per-instance MAC route label: 300080
  MAC database status               Local  Remote
    Total MAC addresses:              4      2
    Default gateway MAC addresses:    0      0
  Number of local interfaces: 1 (1 up)
    Interface name  ESI                            Mode           Status
    ae0.0           00:00:00:00:00:00:00:00:00:00  single-homed   Up
  Number of IRB interfaces: 0 (0 up)
  Number of bridge domains: 3
    VLAN ID  Intfs / up    Mode           MAC sync  IM route label
    1501        1   1      Extended       Enabled   300208
    1502        1   1      Extended       Enabled   300224
    1503        1   1      Extended       Enabled   300240
  Number of neighbors: 1
    192.168.100.40
      Received routes
        MAC address advertisement:          2
        MAC+IP address advertisement:       0
        Inclusive multicast:                3
        Ethernet auto-discovery:            0
  Number of ethernet segments: 0

cyadmin@cdciaasmx01> show evpn database instance evpn-cdc-dci
Instance: evpn-cdc-dci
VLAN  MAC address       Active source              Timestamp       IP address
1501  00:00:5e:00:01:0f  192.168.100.40             Apr 28 20:04:16
1501  00:1b:17:00:01:30  ae0.0                      Apr 28 20:09:32
1502  00:1b:17:00:01:30  ae0.0                      Apr 28 20:09:32
1502  00:50:56:a7:c2:d2  ae0.0                      Apr 28 20:07:55
1503  00:1b:17:00:01:30  ae0.0                      Apr 28 20:09:32
1503  00:50:56:a7:a7:3e  192.168.100.40             Apr 28 20:04:16

cyadmin@cdciaasmx02> show interfaces descriptions
Interface      Admin Link Description
ae0          up    up   cdciaassw02, ae2

cyadmin@cdciaasmx02> show evpn instance evpn-cdc-dci extensive
Instance: evpn-cdc-dci
  Route Distinguisher: 192.168.100.40:5000
  Per-instance MAC route label: 299824
  MAC database status               Local  Remote
    Total MAC addresses:              2      4
    Default gateway MAC addresses:    0      0
  Number of local interfaces: 1 (1 up)
    Interface name  ESI                            Mode           Status
    ae0.0           00:00:00:00:00:00:00:00:00:00  single-homed   Up
  Number of IRB interfaces: 0 (0 up)
  Number of bridge domains: 3
    VLAN ID  Intfs / up    Mode           MAC sync  IM route label
    1501        1   1      Extended       Enabled   300016
    1502        1   1      Extended       Enabled   300032
    1503        1   1      Extended       Enabled   300064
  Number of neighbors: 1
    192.168.100.10
      Received routes
        MAC address advertisement:          4
        MAC+IP address advertisement:       0
        Inclusive multicast:                3
        Ethernet auto-discovery:            0
  Number of ethernet segments: 0

cyadmin@cdciaasmx02> show evpn database instance evpn-cdc-dci
Instance: evpn-cdc-dci
VLAN  MAC address       Active source              Timestamp       IP address
1501  00:00:5e:00:01:0f  ae0.0                      Feb 26 09:02:43
1501  00:1b:17:00:01:30  192.168.100.10             Apr 28 20:08:54
1502  00:1b:17:00:01:30  192.168.100.10             Apr 28 20:08:54
1502  00:50:56:a7:c2:d2  192.168.100.10             Apr 28 20:07:17
```

```
1503  00:1b:17:00:01:30  192.168.100.10          Apr 28 20:08:54
1503  00:50:56:a7:a7:3e  ae0.0                  Apr 28 19:44:43
```

Scenario 2

```
[root@nsx-test-vm01 ~]# arp -a -i eth0 -n
? (10.1.1.20) at 00:50:56:a7:9b:57 [ether] on eth0

[root@nsx-test-vm02 ~]# arp -a -i eth0 -n
? (10.1.1.10) at 00:50:56:a7:c2:af [ether] on eth0

cyadmin@cdciaasmx01> show arp no-resolve
MAC Address       Address         Interface       Flags
00:50:56:6c:72:14 10.10.10.10     ge-1/0/6.1001           none

cyadmin@cdciaasmx02> show arp no-resolve
MAC Address       Address         Interface       Flags
00:50:56:66:e8:77 10.10.20.10     ge-1/0/6.1002           none

cyadmin@cdciaassw01> show ethernet-switching table vlan vlan1001
Ethernet-switching table: 2 unicast entries
  VLAN              MAC address       Type        Age Interfaces
  vlan1001          *                 Flood       - All-members
  vlan1001          00:50:56:6c:72:14 Learn       2:06 ae1.0
  vlan1001          64:87:88:5a:b4:5e Learn       2:05 ge-0/0/9.0

cyadmin@cdciaassw02> show ethernet-switching table vlan vlan1002
Ethernet-switching table: 2 unicast entries
  VLAN              MAC address       Type        Age Interfaces
  vlan1002          *                 Flood       - All-members
  vlan1002          00:50:56:66:e8:77 Learn       47 ae1.0
  vlan1002          5c:5e:ab:02:06:66 Learn       47 ge-0/0/9.0

nsx-controller # show control-cluster logical-switches vtep-table 100500
VNI     IP              Segment         MAC               Connection-ID
100500  10.10.20.10     10.10.20.0      00:50:56:66:e8:77 18
100500  10.10.10.10     10.10.10.0      00:50:56:6c:72:14 16
nsx-controller # show control-cluster logical-switches mac-table 100500
VNI     MAC               VTEP-IP         Connection-ID
100500  00:50:56:a7:c2:af 10.10.10.10     16
100500  00:50:56:a7:9b:57 10.10.20.10     18

cyadmin@cdciaasmx01> show route table lab-core-nsx.inet.0

lab-core-nsx.inet.0: 3 destinations, 3 routes (3 active, 0 holddown, 0 hidden)
+ = Active Route, - = Last Active, * = Both

10.10.10.0/24      *[Direct/0] 8w6d 01:12:58
                    > via ge-1/0/6.1001
10.10.10.1/32      *[Local/0] 8w6d 01:12:58
                      Local via ge-1/0/6.1001
10.10.20.0/24      *[BGP/170] 8w6d 00:43:55, localpref 100, from 192.168.100.40
                      AS path: I, validation-state: unverified
                    > to 172.16.2.2 via ge-1/0/2.2002, Push 17, Push 299792(top)

cyadmin@cdciaasmx02> show route table lab-core-nsx.inet.0

lab-core-nsx.inet.0: 3 destinations, 3 routes (3 active, 0 holddown, 0 hidden)
+ = Active Route, - = Last Active, * = Both

10.10.10.0/24      *[BGP/170] 8w6d 00:44:22, localpref 100, from 192.168.100.10
                      AS path: I, validation-state: unverified
                    > to 172.16.3.2 via ge-1/0/2.2003, Push 19, Push 299808(top)
10.10.20.0/24      *[Direct/0] 8w6d 01:40:50
                    > via ge-1/0/6.1002
10.10.20.1/32      *[Local/0] 8w6d 01:40:50
                      Local via ge-1/0/6.1002
```

Appendix 2. Device configurations

cdciaassw01:

```
set version 14.1X53-D46.7
set system host-name cdciaassw01
set interfaces ge-0/0/0 description "cdciaasmx01, ge-1/0/0"
set interfaces ge-0/0/0 ether-options 802.3ad ae2
set interfaces ge-0/0/1 description "cdciaasmx01, ge-1/0/1"
set interfaces ge-0/0/1 ether-options 802.3ad ae2
set interfaces ge-0/0/2 description "cdciaasfw01, eth1/3"
set interfaces ge-0/0/2 ether-options 802.3ad ae0
set interfaces ge-0/0/3 description "cdciaasfw01, eth1/4"
set interfaces ge-0/0/3 ether-options 802.3ad ae0
set interfaces ge-0/0/4 description "cdciaasmx01, ge-1/0/2"
set interfaces ge-0/0/4 mtu 9192
set interfaces ge-0/0/4 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/4 unit 0 family ethernet-switching vlan members 2002
set interfaces ge-0/0/5 description "cdciaasmx01, ge-1/0/3"
set interfaces ge-0/0/5 mtu 9192
set interfaces ge-0/0/5 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/5 unit 0 family ethernet-switching vlan members 2002
set interfaces ge-0/0/6 description "cdciaasmx01, ge-1/0/4"
set interfaces ge-0/0/6 mtu 9192
set interfaces ge-0/0/6 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/6 unit 0 family ethernet-switching vlan members 2001
set interfaces ge-0/0/8 description "cdciaasmx01, ge-1/0/5"
set interfaces ge-0/0/8 mtu 9192
set interfaces ge-0/0/8 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/8 unit 0 family ethernet-switching vlan members 1501
set interfaces ge-0/0/9 description "cdciaasmx01, ge-1/0/6"
set interfaces ge-0/0/9 mtu 9192
set interfaces ge-0/0/9 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/9 unit 0 family ethernet-switching vlan members 1001
set interfaces ge-0/0/12 description "cdciaasesx03, vmnic7"
set interfaces ge-0/0/12 ether-options 802.3ad ae1
set interfaces ge-0/0/13 description "cdciaasesx03, vmnic6"
set interfaces ge-0/0/13 ether-options 802.3ad ae1
set interfaces ge-0/0/14 description "cdciaasesx03, vmnic5"
set interfaces ge-0/0/14 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/14 unit 0 family ethernet-switching vlan members 1502-1503
set interfaces ge-0/0/20 description "cdciaasmx01, fxp0"
set interfaces ge-0/0/20 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/20 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/21 description "cdciaasfw01, mgmt"
set interfaces ge-0/0/21 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/21 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/22 description "cdciaassw02, ge-0/0/23"
set interfaces ge-0/0/22 mtu 9192
set interfaces ge-0/0/22 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/22 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/22 unit 0 family ethernet-switching vlan members 2001
set interfaces ge-0/0/23 description "hki-per-labmgmtsw01, fa0/5"
set interfaces ge-0/0/23 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/23 unit 0 family ethernet-switching vlan members 3
set interfaces ae0 description "cdciaasfw01, ae1"
set interfaces ae0 mtu 9192
set interfaces ae0 aggregated-ether-options minimum-links 1
set interfaces ae0 aggregated-ether-options link-speed 1g
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family ethernet-switching port-mode trunk
set interfaces ae0 unit 0 family ethernet-switching vlan members 1501-1503
set interfaces ae1 description "cdciaasesx03, LAG1"
set interfaces ae1 mtu 9192
set interfaces ae1 aggregated-ether-options minimum-links 1
set interfaces ae1 aggregated-ether-options link-speed 1g
set interfaces ae1 aggregated-ether-options lacp active
```

```
set interfaces ae1 aggregated-ether-options lacp periodic fast
set interfaces ae1 unit 0 family ethernet-switching port-mode trunk
set interfaces ae1 unit 0 family ethernet-switching vlan members 1001
set interfaces ae2 description "cdciaasmx01, ae0"
set interfaces ae2 mtu 9192
set interfaces ae2 aggregated-ether-options minimum-links 1
set interfaces ae2 aggregated-ether-options link-speed 1g
set interfaces ae2 aggregated-ether-options lacp active
set interfaces ae2 aggregated-ether-options lacp periodic fast
set interfaces ae2 unit 0 family ethernet-switching port-mode trunk
set interfaces ae2 unit 0 family ethernet-switching vlan members 1501-1503
set vlans vlan1001 description DC1-NSX-Underlay
set vlans vlan1001 vlan-id 1001
set vlans vlan1501 description RTR-FW-Transit
set vlans vlan1501 vlan-id 1501
set vlans vlan1502 description DC-VM-Network01
set vlans vlan1502 vlan-id 1502
set vlans vlan1503 description DC-VM-Network02
set vlans vlan1503 vlan-id 1503
set vlans vlan2001 description P1-to-P2
set vlans vlan2001 vlan-id 2001
set vlans vlan2002 description PE1-to-SW01
set vlans vlan2002 vlan-id 2002
```

cdciaassw02:

```
set version 14.1X53-D46.7
set system host-name cdciaassw02
set interfaces ge-0/0/0 description "cdciaasmx02, ge-1/0/0"
set interfaces ge-0/0/0 ether-options 802.3ad ae2
set interfaces ge-0/0/1 description "cdciaasmx02, ge-1/0/1"
set interfaces ge-0/0/1 ether-options 802.3ad ae2
set interfaces ge-0/0/2 description "cdciaasfw02, eth1/3"
set interfaces ge-0/0/2 ether-options 802.3ad ae0
set interfaces ge-0/0/3 description "cdciaasfw02, eth1/4"
set interfaces ge-0/0/3 ether-options 802.3ad ae0
set interfaces ge-0/0/4 description "cdciaasmx02, ge-1/0/2"
set interfaces ge-0/0/4 mtu 9192
set interfaces ge-0/0/4 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/4 unit 0 family ethernet-switching vlan members 2003
set interfaces ge-0/0/5 description "cdciaasmx02, ge-1/0/3"
set interfaces ge-0/0/5 mtu 9192
set interfaces ge-0/0/5 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/5 unit 0 family ethernet-switching vlan members 2003
set interfaces ge-0/0/6 description "cdciaasmx02, ge-1/0/4"
set interfaces ge-0/0/6 mtu 9192
set interfaces ge-0/0/6 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/6 unit 0 family ethernet-switching vlan members 2001
set interfaces ge-0/0/8 description "cdciaasmx02, ge-1/0/5"
set interfaces ge-0/0/8 mtu 9192
set interfaces ge-0/0/8 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/8 unit 0 family ethernet-switching vlan members 1501
set interfaces ge-0/0/9 description "cdciaasmx02, ge-1/0/6"
set interfaces ge-0/0/9 mtu 9192
set interfaces ge-0/0/9 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/9 unit 0 family ethernet-switching vlan members 1002
set interfaces ge-0/0/10 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/10 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/11 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/11 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/12 description "cdciaasesx05, vmnic7"
set interfaces ge-0/0/12 ether-options 802.3ad ae1
set interfaces ge-0/0/13 description "cdciaasesx05, vmnic6"
set interfaces ge-0/0/13 ether-options 802.3ad ae1
set interfaces ge-0/0/14 description "cdciaasesx05, vmnic5"
set interfaces ge-0/0/14 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/14 unit 0 family ethernet-switching vlan members 1502-1503
set interfaces ge-0/0/20 description "cdciaascentos01.lab.cygate.fi, vmnic0"
set interfaces ge-0/0/20 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/20 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/21 description "cdciaasmx02, fxp0"
```

set interfaces ge-0/0/21 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/21 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/22 description "cdciaasfw02, mgmt"
set interfaces ge-0/0/22 unit 0 family ethernet-switching port-mode access
set interfaces ge-0/0/22 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/23 description "cdciaassw01, ge-0/0/22"
set interfaces ge-0/0/23 mtu 9192
set interfaces ge-0/0/23 unit 0 family ethernet-switching port-mode trunk
set interfaces ge-0/0/23 unit 0 family ethernet-switching vlan members 3
set interfaces ge-0/0/23 unit 0 family ethernet-switching vlan members 2001
set interfaces ae0 description "cdciaasfw02, ae1"
set interfaces ae0 mtu 9192
set interfaces ae0 aggregated-ether-options minimum-links 1
set interfaces ae0 aggregated-ether-options link-speed 1g
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family ethernet-switching port-mode trunk
set interfaces ae0 unit 0 family ethernet-switching vlan members 1501-1503
set interfaces ae1 description "cdciaasesx05, LAG1"
set interfaces ae1 mtu 9192
set interfaces ae1 aggregated-ether-options minimum-links 1
set interfaces ae1 aggregated-ether-options link-speed 1g
set interfaces ae1 aggregated-ether-options lacp active
set interfaces ae1 aggregated-ether-options lacp periodic fast
set interfaces ae1 unit 0 family ethernet-switching port-mode trunk
set interfaces ae1 unit 0 family ethernet-switching vlan members 1002
set interfaces ae2 description "cdciaasmx02, ae0"
set interfaces ae2 mtu 9192
set interfaces ae2 aggregated-ether-options minimum-links 1
set interfaces ae2 aggregated-ether-options link-speed 1g
set interfaces ae2 aggregated-ether-options lacp active
set interfaces ae2 aggregated-ether-options lacp periodic fast
set interfaces ae2 unit 0 family ethernet-switching port-mode trunk
set interfaces ae2 unit 0 family ethernet-switching vlan members 1501-1503
set vlans vlan1002 description DC2-NSX-Underlay
set vlans vlan1002 vlan-id 1002
set vlans vlan1501 description RTR-FW-Transit
set vlans vlan1501 vlan-id 1501
set vlans vlan1502 description DC-VM-Network01
set vlans vlan1502 vlan-id 1502
set vlans vlan1503 description DC-VM-Network02
set vlans vlan1503 vlan-id 1503
set vlans vlan2001 description P1-to-P2
set vlans vlan2001 vlan-id 2001
set vlans vlan2003 description PE2-to-SW02
set vlans vlan2003 vlan-id 2003

cdciaasmx01:

set version 15.1R6.7
set system host-name cdciaasmx01
set interfaces ge-1/0/0 description "cdciaassw01, ge-0/0/0"
set interfaces ge-1/0/0 gigether-options 802.3ad ae0
set interfaces ge-1/0/1 description "cdciaassw01, ge-0/0/1"
set interfaces ge-1/0/1 gigether-options 802.3ad ae0
set interfaces ge-1/0/2 description "cdciaassw01, ge-0/0/4"
set interfaces ge-1/0/2 flexible-vlan-tagging
set interfaces ge-1/0/2 mtu 9192
set interfaces ge-1/0/2 unit 2002 description "MPLS/IP Core -> P1 (via sw01 ge-0/0/5)"
set interfaces ge-1/0/2 unit 2002 vlan-id 2002
set interfaces ge-1/0/2 unit 2002 family inet address 172.16.2.1/24
set interfaces ge-1/0/2 unit 2002 family iso mtu 9150
set interfaces ge-1/0/2 unit 2002 family mpls mtu 9150
set interfaces ge-1/0/3 description "cdciaassw01, ge-0/0/5"
set interfaces ge-1/0/3 flexible-vlan-tagging
set interfaces ge-1/0/3 mtu 9192
set interfaces ge-1/0/3 encapsulation flexible-ethernet-services
set interfaces ge-1/0/4 description "cdciaassw01, ge-0/0/6"
set interfaces ge-1/0/4 flexible-vlan-tagging
set interfaces ge-1/0/4 mtu 9192
set interfaces ge-1/0/4 encapsulation flexible-ethernet-services

```
set interfaces ge-1/0/5 description "cdciaassw01, ge-0/0/8"
set interfaces ge-1/0/5 flexible-vlan-tagging
set interfaces ge-1/0/5 mtu 9192
set interfaces ge-1/0/5 unit 1501 vlan-id 1501
set interfaces ge-1/0/5 unit 1501 family inet mtu 1500
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.2/24 vrrp-group 15 virtual-address 10.100.100.1
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.2/24 vrrp-group 15 priority 90
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.2/24 vrrp-group 15 preempt hold-time 600
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.2/24 vrrp-group 15 accept-data
set interfaces ge-1/0/6 description "cdciaassw01, ge-0/0/9"
set interfaces ge-1/0/6 flexible-vlan-tagging
set interfaces ge-1/0/6 mtu 9192
set interfaces ge-1/0/6 unit 1001 vlan-id 1001
set interfaces ge-1/0/6 unit 1001 family inet mtu 1600
set interfaces ge-1/0/6 unit 1001 family inet address 10.10.10.1/24
set interfaces ae0 description "cdciaassw01, ae2"
set interfaces ae0 flexible-vlan-tagging
set interfaces ae0 mtu 9192
set interfaces ae0 encapsulation flexible-ethernet-services
set interfaces ae0 aggregated-ether-options minimum-links 1
set interfaces ae0 aggregated-ether-options link-speed 1g
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge vlan-id-list 1501-1502
set interfaces ae0 unit 0 family bridge vlan-id-list 1503
set interfaces lo0 unit 0 family inet address 192.168.100.10/32 primary
set interfaces lo0 unit 0 family inet address 127.0.0.1/32
set interfaces lo0 unit 0 family iso address 49.0001.1921.6815.0010.00
set routing-options router-id 192.168.100.10
set routing-options autonomous-system 65000
set protocols mpls no-propagate-ttl
set protocols mpls icmp-tunneling
set protocols mpls interface ge-1/0/2.2002
set protocols mpls interface lo0.0
set protocols bgp local-address 192.168.100.10
set protocols bgp hold-time 20
set protocols bgp mtu-discovery
set protocols bgp out-delay 0
set protocols bgp log-updown
set protocols bgp family inet-vpn unicast
set protocols bgp group masi-lab-v4 type internal
set protocols bgp group masi-lab-v4 local-address 192.168.100.10
set protocols bgp group masi-lab-v4 family inet-vpn unicast
set protocols bgp group masi-lab-v4 family evpn signaling
set protocols bgp group masi-lab-v4 family route-target
set protocols bgp group masi-lab-v4 neighbor 192.168.100.40 description PE2
set protocols isis level 1 disable
set protocols isis interface ge-1/0/2.2002 point-to-point
set protocols isis interface ge-1/0/2.2002 level 2 metric 10
set protocols isis interface lo0.0 passive
set protocols ldp track-igp-metric
set protocols ldp transport-address router-id
set protocols ldp interface ge-1/0/2.2002
set protocols ldp interface lo0.0
set protocols lldp interface all
set routing-instances evpn-cdc-dci instance-type virtual-switch
set routing-instances evpn-cdc-dci interface ae0.0
set routing-instances evpn-cdc-dci route-distinguisher 192.168.100.10:5000
set routing-instances evpn-cdc-dci vrf-target target:65000:5000
set routing-instances evpn-cdc-dci protocols evpn extended-vlan-list 1501-1503
set routing-instances evpn-cdc-dci bridge-domains dci vlan-id-list 1501-1503
set routing-instances lab-core-demo instance-type vrf
set routing-instances lab-core-demo interface ge-1/0/5.1501
set routing-instances lab-core-demo route-distinguisher 192.168.100.10:3000
set routing-instances lab-core-demo vrf-target target:65000:3000
set routing-instances lab-core-demo vrf-table-label
set routing-instances lab-core-demo routing-options static route 10.1.2.0/24 next-hop 10.100.100.254
set routing-instances lab-core-demo routing-options static route 10.1.3.0/24 next-hop 10.100.100.254
set routing-instances lab-core-nsx instance-type vrf
set routing-instances lab-core-nsx interface ge-1/0/6.1001
set routing-instances lab-core-nsx route-distinguisher 192.168.100.10:4000
set routing-instances lab-core-nsx vrf-target target:65000:4000
```

set routing-instances lab-core-nsx vrf-table-label

cdciaasmx02:

set version 15.1R6.7
set system host-name cdciaasmx02
set interfaces ge-1/0/0 description "cdciaassw02, ge-0/0/0"
set interfaces ge-1/0/0 gigether-options 802.3ad ae0
set interfaces ge-1/0/1 description "cdciaassw02, ge-0/0/1"
set interfaces ge-1/0/1 gigether-options 802.3ad ae0
set interfaces ge-1/0/2 description "cdciaassw02, ge-0/0/4"
set interfaces ge-1/0/2 flexible-vlan-tagging
set interfaces ge-1/0/2 mtu 9192
set interfaces ge-1/0/2 unit 2003 description "MPLS/IP Core -> P1 (via sw01 ge-0/0/5)"
set interfaces ge-1/0/2 unit 2003 vlan-id 2003
set interfaces ge-1/0/2 unit 2003 family inet address 172.16.3.1/24
set interfaces ge-1/0/2 unit 2003 family iso mtu 9150
set interfaces ge-1/0/2 unit 2003 family mpls mtu 9150
set interfaces ge-1/0/3 description "cdciaassw02, ge-0/0/5"
set interfaces ge-1/0/3 flexible-vlan-tagging
set interfaces ge-1/0/3 mtu 9192
set interfaces ge-1/0/3 encapsulation flexible-ethernet-services
set interfaces ge-1/0/4 description "cdciaassw02, ge-0/0/6"
set interfaces ge-1/0/4 flexible-vlan-tagging
set interfaces ge-1/0/4 mtu 9192
set interfaces ge-1/0/4 encapsulation flexible-ethernet-services
set interfaces ge-1/0/5 description "cdciaassw02, ge-0/0/8"
set interfaces ge-1/0/5 flexible-vlan-tagging
set interfaces ge-1/0/5 mtu 9192
set interfaces ge-1/0/5 unit 1501 vlan-id 1501
set interfaces ge-1/0/5 unit 1501 family inet mtu 1500
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.3/24 vrrp-group 15 virtual-address 10.100.100.1
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.3/24 vrrp-group 15 priority 120
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.3/24 vrrp-group 15 preempt hold-time 600
set interfaces ge-1/0/5 unit 1501 family inet address 10.100.100.3/24 vrrp-group 15 accept-data
set interfaces ge-1/0/6 description "cdciaassw02, ge-0/0/9"
set interfaces ge-1/0/6 flexible-vlan-tagging
set interfaces ge-1/0/6 mtu 9192
set interfaces ge-1/0/6 unit 1002 vlan-id 1002
set interfaces ge-1/0/6 unit 1002 family inet mtu 1600
set interfaces ge-1/0/6 unit 1002 family inet address 10.10.20.1/24
set interfaces ae0 description "cdciaassw02, ae2"
set interfaces ae0 flexible-vlan-tagging
set interfaces ae0 mtu 9192
set interfaces ae0 encapsulation flexible-ethernet-services
set interfaces ae0 aggregated-ether-options minimum-links 1
set interfaces ae0 aggregated-ether-options link-speed 1g
set interfaces ae0 aggregated-ether-options lacp active
set interfaces ae0 aggregated-ether-options lacp periodic fast
set interfaces ae0 unit 0 family bridge interface-mode trunk
set interfaces ae0 unit 0 family bridge vlan-id-list 1501-1502
set interfaces ae0 unit 0 family bridge vlan-id-list 1503
set interfaces lo0 unit 0 family inet address 192.168.100.40/32 primary
set interfaces lo0 unit 0 family inet address 127.0.0.1/32
set interfaces lo0 unit 0 family iso address 49.0001.1921.6815.0040.00
set routing-options router-id 192.168.100.40
set routing-options autonomous-system 65000
set protocols mpls no-propagate-ttl
set protocols mpls icmp-tunneling
set protocols mpls interface ge-1/0/2.2003
set protocols mpls interface lo0.0
set protocols bgp local-address 192.168.100.40
set protocols bgp hold-time 20
set protocols bgp mtu-discovery
set protocols bgp out-delay 0
set protocols bgp log-updown
set protocols bgp family inet-vpn unicast
set protocols bgp group masi-lab-v4 type internal
set protocols bgp group masi-lab-v4 local-address 192.168.100.40
set protocols bgp group masi-lab-v4 family inet-vpn unicast
set protocols bgp group masi-lab-v4 family evpn signaling

set protocols bgp group masi-lab-v4 family route-target
set protocols bgp group masi-lab-v4 neighbor 192.168.100.10 description PE1
set protocols isis level 1 disable
set protocols isis interface ge-1/0/2.2003 point-to-point
set protocols isis interface ge-1/0/2.2003 level 2 metric 10
set protocols isis interface lo0.0 passive
set protocols ldp track-igp-metric
set protocols ldp transport-address router-id
set protocols ldp interface ge-1/0/2.2003
set protocols ldp interface lo0.0
set protocols lldp interface all
set routing-instances evpn-cdc-dci instance-type virtual-switch
set routing-instances evpn-cdc-dci interface ae0.0
set routing-instances evpn-cdc-dci route-distinguisher 192.168.100.40:5000
set routing-instances evpn-cdc-dci vrf-target target:65000:5000
set routing-instances evpn-cdc-dci protocols evpn extended-vlan-list 1501-1503
set routing-instances evpn-cdc-dci bridge-domains dci vlan-id-list 1501-1503
set routing-instances lab-core-demo instance-type vrf
set routing-instances lab-core-demo interface ge-1/0/5.1501
set routing-instances lab-core-demo route-distinguisher 192.168.100.40:3000
set routing-instances lab-core-demo vrf-target target:65000:3000
set routing-instances lab-core-demo vrf-table-label
set routing-instances lab-core-demo routing-options static route 10.1.2.0/24 next-hop 10.100.100.254
set routing-instances lab-core-demo routing-options static route 10.1.3.0/24 next-hop 10.100.100.254
set routing-instances lab-core-nsx instance-type vrf
set routing-instances lab-core-nsx interface ge-1/0/6.1002
set routing-instances lab-core-nsx route-distinguisher 192.168.100.40:4000
set routing-instances lab-core-nsx vrf-target target:65000:4000
set routing-instances lab-core-nsx vrf-table-label