

Data Centre Reference Design Solution Guide

Alcatel·Lucent Enterprise

Table of Contents

About this document 4
- Purpose 4
- Audience 4
- Scope 4

Acronyms 4

Related documents 6

Introduction 7

Solution overview 7
- Spine-and-leaf topology 7
- Wire-rate switching 8
- Oversubscription ratio 9
- Spine-and-leaf physical requirements 9
- Border spines and border leaves 10
- Hypervisor attachment 11

Single-site designs 12
- Layer 2 VC-based architecture 13
- Layer 3 VC-based architecture 14
- SPB-based architecture 15

Split-site design 16
- SPB-based architecture 17
- Layer 2 VC-based architecture 18
- Layer 3 VC-based architecture 19

Multi-site designs 19
- Border spine versus border leaves DCI 19

Routing 21
- Routing within the data centre 21
- Routing into and out of the data centre 24

Multi-site design 25
Global stretched subnets – L2 DCI 25
Local subnets – L3 DCI 28
Local stretched subnets – L3 DCI 29
Two-site data centre designs with DCI – VC-based 30
Two-site data centre designs with DCI – SPB-based 32
Overlay architecture considerations 33
Summary 34

About this document

Purpose

This document is a reference guide for Alcatel-Lucent Enterprise solutions to support provide Data Centre (DC) customers. This guide includes use cases, business drivers, technical requirements, and solution overviews along with the value proposition for each solution set.

Audience

This guide is intended for Alcatel-Lucent Enterprise Business Partner Sales and Pre-sales staff, as well as customers. It assumes the reader has a fundamental knowledge of IP switching and routing. It is intended to provide guidelines and best practices for customer deployments, to networking professionals involved in the design and deployment of enterprise networks.

Scope

This document will focus on traditional DC architecture solutions and certain considerations to be taken into account when implementing some overlay technologies, and it can be used as a reference document for designing DCs. This document will not provide in-depth product specifications as these are already provided in datasheets and specification guides.

The document is divided into individual modules for each solution set to enable the reader to focus on sections deemed most relevant to them.

Acronyms

Acronyms	
ACL	Access Control List
ARP	Address Resolution Protocol
AS	Autonomous System
AWS	Amazon Web Services
BGP	Border Gateway Protocol
BUM	Broadcast Unknown-unicast and Multicast
BVLAN	Backbone VLAN
CVLAN	Customer VLAN
DC	Data Centre
DCI	Data Centre Interconnect
DHCP	Dynamic Host Configuration Protocol
DLR	Distributed Logical Router
DNS	Domain-Name System
DR	Disaster Recovery
DR	Distributed Routers
DRS	Distributed Resource Scheduler
DWDM	Dense Wavelength-Division Multiplexing

Acronyms	
ECMP	Equal-Cost Multi-Path Routing
EMP	Ethernet Management Port
ESG	Edge Services Gateway
ESXi	Elastic Sky X integrated
EVPN	Ethernet VPN
GENEVE	Generic Network Virtualisation Encapsulation
IGMP	Internet Group Management Protocol
IP	Internet Protocol
ISID	Instance Service Identifier
IS-IS	Intermediate-System to Intermediate-System
IT	Information Technology
KVM	Kernel-based Virtual Machine
LACP	Link Aggregation Control Protocol
LAG	Link-Aggregation
LB	Load Balancer
LBD	Loopback Detection
MAC	Media Acces Control
MACSec	Media Access Control security
MED	Multi-Exit Discriminator
M-LAG or MC-LAG	Multi-Chassis Link Aggregation
MPLS	Multi-protocol label switching
MTU	Maximum Transmission Unit
MVRP	Multiple VLAN Registration Protocol
NAT	Network Address Translation
NIC	Network Interface Card
OOB	Out-Of-Band
OSPF	Open Shortest Path First
OVSDB	Open vSwitch Database Management Protocol
PBB	Provider Backbone Bridging
PIM	Protocol Independent Multicast
PW	Pseudowire
QSFP	Quad Small Form-factor Pluggable
RAM	Random Access Memory
RCD	Remote Chassis Detection
RU	Rack Unit
SAP	Service Access Point
SFP	Small Form-factor Plugable

Acronyms	
SMLT	Split Multi-Link Trunking
SP	Service Provider
SPB	Shortest Path Bridging
SR	Service Routers
STP	Spanning Tree Protocol
TCP	Transmission Control Protocol

Related documents

Alcatel-Lucent OmniSwitch® 6900 Datasheet

Shortest Path Bridging Architecture Guide

Network infrastructure solutions: Security best practices

Hybrid Cloud Architecture - AWS

Introduction

Data centres are not just next-generation network upgrades. They signify a transformation from an information technology (IT)-centric infrastructure to a service-centric infrastructure.

But to meet that objective, all components, including servers, storage, network and applications are virtualised to cost-effectively deliver computational elasticity, and data and application serviceability.

Based on business objectives and the type of cloud applications deployed, the key goals for any DC architecture are:

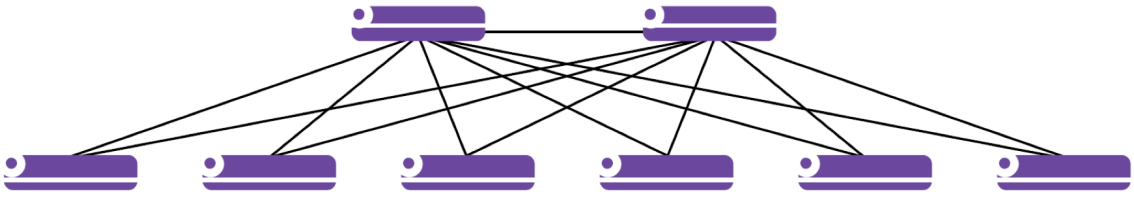
- Deterministic latency
- Redundancy/high availability
- Manageability
- Scalability

Alcatel-Lucent Enterprise offers a broad range of solutions, each addressing the fundamental requirements of data centre networks, for both today and tomorrow. This document provides guidelines to design solutions to meet the needs of any organisation.

Solution overview

Spine-and-leaf topology

Figure 1 – Spine-and-leaf topology



Spine-and-leaf, or Clos architecture is a two-tiered networking design. It is named after Charles Clos, the Bell Labs engineer who formalised this architecture for telephone, circuit-switched networks. This architecture is equally applicable to packet-switched data communication in a DC.

The main building blocks are leaves and spines.

What are the reasons that make the spine-and-leaf or Clos architecture so successful?

This architecture was developed to overcome the limitations of the three-tier architecture. With the increase of cloud and containerised infrastructure, and the increase of East-to-West traffic, this topology is the prevalent choice when building a modern DC.

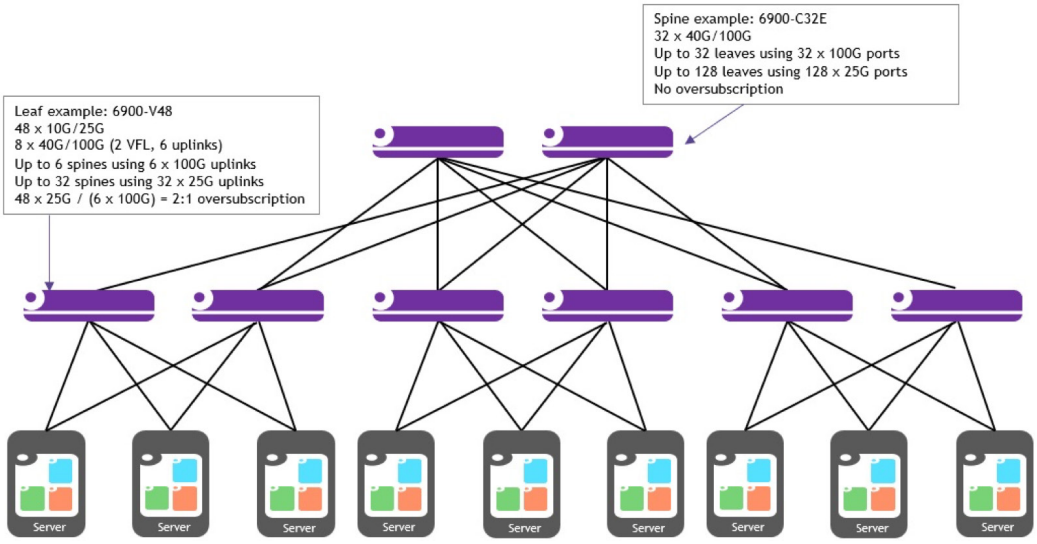
This architecture is the topology of choice for today's DCs for several reasons:

- **Simpler design** and less cabling
- **Cost-effectiveness**: It is based on cheaper compact, fixed, single Rack Unit (RU) switches, and not on expensive modular chassis'. Cost scales linearly with the number of ports
- **Bandwidth consistency**: Bandwidth is consistent between any two nodes
- **Latency**: Single hop between any pair of leaves which improves latency

- **Scalability**: Simple horizontal scalability; easy to scale within boundary
  - ¬ If better oversubscription ratio is required: Add more uplinks
  - ¬ If more ports are required: Add more leaves/spines as needed

Spines forward traffic along optimal paths between nodes at Layer 2 or Layer 3, while leaves control the flow of traffic between locally connected servers. Cross-sectional interconnect bandwidth can be improved though Link Aggregation (LAG) links, or by employing Layer 3 Equal-Cost Multi-Path (ECMP). There is single-hop latency for server to-server communication within the leaves. Additional latency is a factor when traffic needs to bridge the spine for communication, with a maximum of three hops for any-to any communication.

Figure 2 – Sample single-site spine-and-leaf topology



Let's refer to the example architecture displayed in **Figure-2** and analyse the architecture.

Wire-rate switching

To ensure high performance in the DC network, usually wire-rate switches are deployed. As a reference, we can consider switch models OS6900-V48 as leaves and OS6900-C32E as spines. These switches can be confirmed to be forwarding at wire-rate as per the following calculations:

Table 1 - OS6900 Performance Data

Item	OS6900-V48	OS6900-C32E
Number of Ports: 100G QSFP28	8	32
Number of Ports: 25G SFP28	48	N/A
Switching Capacity as per Datasheet	4 Tb/s	6.4 Tb/s

The formula below can be used to calculate the total bandwidth provided by all the interfaces:

Total bandwidth = Number of interfaces x Interface speed x 2 (full duplex)

To validate that the switch is forwarding at wire-rate, the total bandwidth should be less than or equal to the switch capacity:

OS6900-C32E - Total bandwidth = (32 x 100G) x 2 ~ 6.4 Tb/s

OS6900-V48 – Total bandwidth = (8 x 100G + 48 x 25G) x 2 ~ 4 Tb/s

Oversubscription ratio

Next, let's analyse this topology in terms of oversubscription ratio. We will analyse the limits set forth by the topology itself before other considerations (for example routing, virtual chassis (VC)) are factored in. We will look at those aspects in a later section.

When we talk about oversubscription ratio, we are referring to the ratio of uplink-port capacity to access-port capacity.

The oversubscription ratio is determined as follows:

(Number of downlink ports * downlink speed)/(Number of uplink ports * uplink speed)

Typical networks are designed with oversubscription levels ranging from 2:1 to 6:1, but in most cases an acceptable oversubscription ratio is 3:1. This depends on the traffic patterns on your network which will help decide on an acceptable oversubscription ratio.

Spine-and-leaf physical requirements

The design of your DC network should always be driven by the requirement of the applications, such as bandwidth requirements and number of access ports required. After this has been identified, you should determine an acceptable oversubscription ratio if you are designing a spine-and-leaf based architecture.

Leaf switches

When selecting a leaf switch, it should meet your oversubscription requirements, for example it should have the required number of access and uplink ports. The number of required leaf switches can be calculated as follows:

Number of leaf switches = (Number of access ports)/(Number of access ports per switch)

After that the number of uplinks should be determined. It can be calculated as follows:

Number of uplinks = (Number of leaf switches) * (Number of uplinks)

Keep in mind in case you are using Virtual Chassis (VC) based architecture, which will be discussed later on, you will require two uplink ports for VFL and the number of leaf switches selected should be even.

Spine switches

The spine switches minimum number of ports should be equal to the number of leaf switches since each spine will connect to all leaf switches. The number of spines required can be calculated as follows:

Number of spine switches = (Number of uplinks)/(Number of access ports per switch)

For redundancy purposes, it might be required to have a minimum of four spine switches, but it will be shown it as two spines for the rest of this document for simplicity.

Using **Figure 2** as a reference, let's analyse the oversubscription ratio by looking at the leaf nodes. A typical leaf node has 48 access ports and 8 uplink ports. The uplink-port speed is 4 times the access-port speed and these ports can be split into 4 logical ports with the same speed as an access port. Let's assume that we reserve 2 of those uplink ports as Virtual Fabric Link (VFL) ports and that leaves 6 uplink ports towards the spines. In this case, using OS6900-V48, the oversubscription ratio is 2:1.

Let's continue by looking at the spine nodes. A typical spine node has 32 ports whose speed matches the speed of the leaf uplink port and can also be split 4-ways. Therefore, up to 32 leaf nodes and 6 spine nodes are supported by this topology using full-bandwidth uplinks for a total of 1536 access ports at 2:1 oversubscription. When using split uplinks, up to 128 leaf nodes and 24 spine nodes are supported for a total of 6144 access ports at the same oversubscription ratio.

Bear in mind that these are the topology limits. But there are other considerations that influence the ability to reach those limits, such as routing, or the choice of Backbone VLANs (BVLANs), in the case of Shortest Path Bridging (SPB). We will examine those considerations later.

Furthermore, these figures do not factor in ports that may be needed to connect the DC to the outside world or to services such as firewalling and load-balancing, among others.

Border spines and border leaves

Border nodes connect the DC network to the outside world (for example the campus network, the WAN or the Internet) or to services such as firewalling and load-balancing.

In a spine-and-leaf architecture, there is a choice to make between border spines and border leaves. Let's examine these two options.

Figure 3 - Border spine architecture

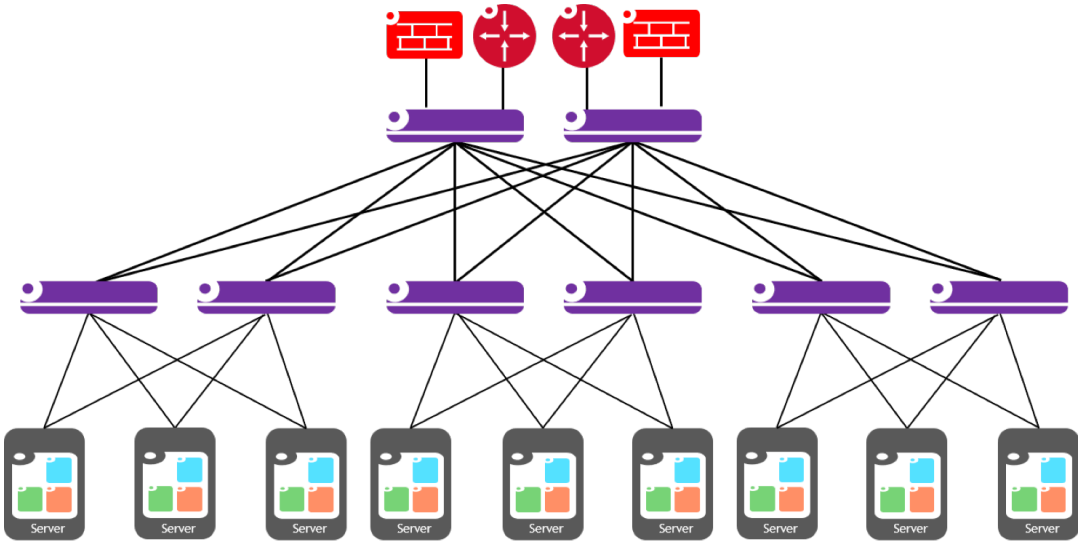
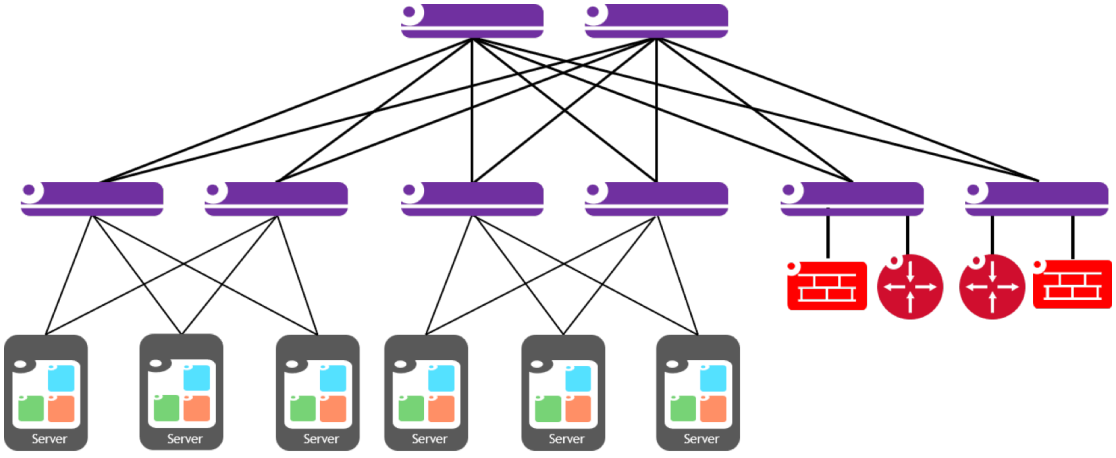


Figure 3 illustrates the case of border spines. As seen in the diagram, any external devices or links connect to the spines. Using border spines is a good option when there are two spines as this allows for active-active or active-standby redundant external connections and/or services.

However, a larger DC may require more than two spines. And using border spines may stop making sense if external connections or services only connect to two of those spines. This would result in unbalanced traffic because some spines would receive more traffic than others. In such case, using border leaves may make more sense.

Figure 4 - Border leaves architecture

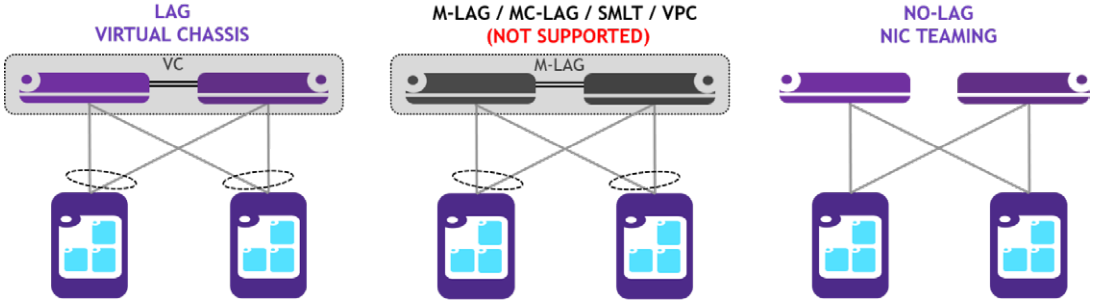


Border leaves are illustrated in **Figure 4**. Note that the diagram is still showing only two spines, this is just to keep it simple. In such architecture, inter-VLAN or inter-ISID (Instance Service Identifier) routing may still be performed at the spines while the border leaves perform external routing only. Alternatively, the network designer may choose to always route at the firewalls for security reasons, even between internal DC subnets, in which case the spines do not perform any routing at all.

Hypervisor attachment

In this section we will present different options for server or hypervisor attachment.

Figure 5 - Server attachment



These options are illustrated in **Figure 5.** Let's analyse them in detail:

- **VC + LAG**: This is the simplest, best and recommended option in most cases. Two leaf nodes are interconnected and form a VC. These two leaf nodes will typically be deployed as top-of-rack (TOR) switches on the same rack. However, they can be spread across contiguous racks also. Bare-metal servers or hypervisors connect to both leaf nodes through an Link Aggregation Control Protocol (LACP) aggregate. LACP provides active-active load balancing and fast failover in the event of link, NIC or leaf node failure.Traffic is load balanced according to a hashing function which can consider source and destination MAC, IP addresses or TCP/UDP ports. The exact hashing logic is configurable on both the switch and the hypervisor. VC virtualises both leaf nodes from a data, control and management plane point of view and at both Layer 2 (switching) and Layer 3 (routing). This virtualisation brings deployment and operational simplification. In addition, the Alcatel-Lucent OmniSwitch® side of the link aggregate is automatically configured as soon as it is configured on the hypervisor or server side, bringing additional operation simplification. While the VC feature is not licensed on OmniSwitch products, LACP may be licensed on the hypervisor or virtualisation platform.
- **MC-LAG + LAG**: This deployment option is no longer supported nor recommended. We are only describing it here for completeness. Note that other vendors may refer to this feature as M-LAG, VPC or SMLT. MC-LAG only provides the ability to terminate LACP aggregates on different leaf nodes and load balance across them. It does not virtualise both nodes from a Layer 3, control or management plane point of view like VC does. Both nodes are managed independently and are different entities at Layer 3. This introduces issues for DHCP relay, DHCP snooping, multicast, routing, and convergence issues. Today, many routers or firewalls are deployed as Virtual Machines (VMs). If so, imagine OSPF Hellos being sent (as multicast) on one of the members of the LAG ports. This means that the router will be adjacent to one of the leaf nodes only. In case of leaf node or link failure, the VM will need to rediscover the other leaf node which will affect convergence time.
- **No-LAG (NIC teaming)**: This option can be used when LACP is not supported on the hypervisor, or to avoid the additional licensing costs incurred in enabling this feature. Leaf nodes are not virtualised, they are two separate entities with independent control and management planes. The hypervisor's NICs are bonded in active-active mode without LACP. Traffic is balanced across both interfaces according to the VM's source MAC hash. This means that a given VM will always use the same interface until the interface or link fails. In such event, the affected VMs will failover to the remaining interface. For this reason, both interfaces need to be mapped to the same broadcast domains (for example, VLAN or SPB service). Therefore, these broadcast domains need to be extended to both leaf nodes. This is accomplished directly by interconnecting both nodes, or indirectly, by transiting the spine nodes. It should be noted that, except in the case of an SPB-based architecture, this option will require Spanning Tree Protocol (STP) since the same broadcast domains are extended across two nodes.

Enable Loopback Detection (LBD) on access ports to protect from loops that may be accidentally created inside the hypervisor (for example when deploying virtual network appliances).

Single-site designs

Let's begin by analysing the single-site data centre. The single-site data centre architecture can be VC-

based or SPB-based.

Layer 2 VC-based architecture

A Virtual Chassis (VC) is a group of switches managed through a single management IP address and that behave as a single bridge or router. It provides both node level and link level redundancy for devices connecting to the aggregation layer using dual-homed standard 802.3ad link aggregation mechanisms. The physical chassis is connected together through one VFL trunk, which may consist of up to 16 member ports, depending on the model.

In a Layer 2 VC-based architecture, the spines nodes are clustered together in a VC. Leaf nodes may or may not be clustered in a VC depending on whether the hypervisors will be attached through LACP aggregates or not. Please refer to **Figure 6** and **Figure 7**.

Figure 6 - VC spines and standalone leaves

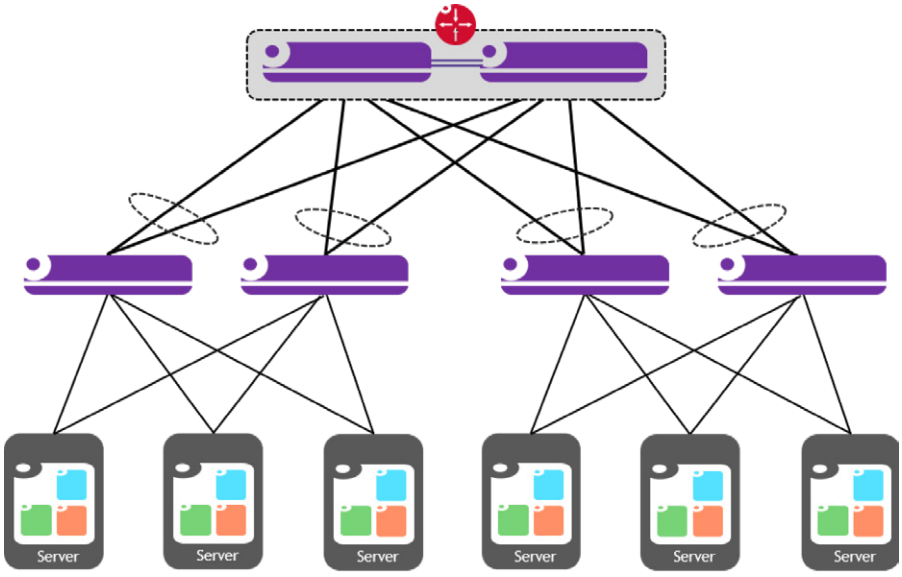
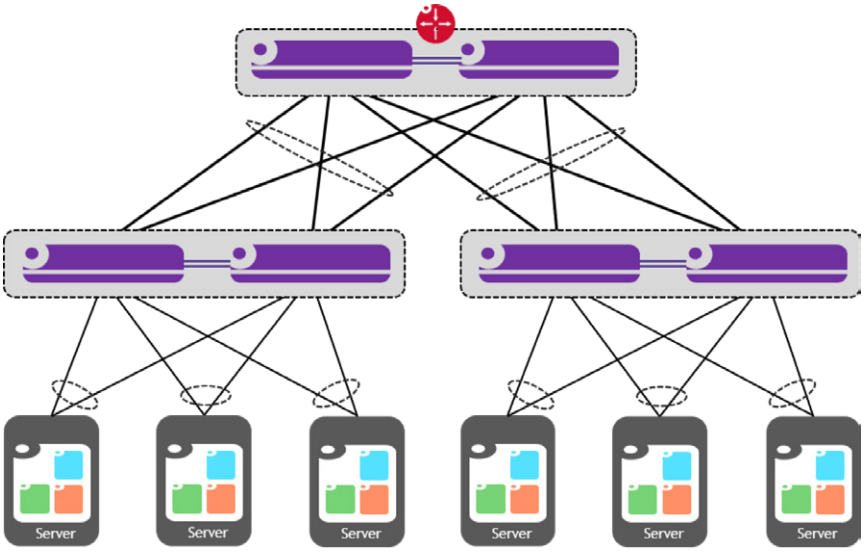


Figure 7 - VC spines and VC leaves



A VC-based architecture provides simple active-active and hash-based load balancing at both Layer 2 and Layer 3. Inter-subnet routing at the spine VC takes advantage of VC-based high-availability and load balancing without any additional redundancy protocols (for example, Virtual Router Redundancy Protocol (VRRP)). There is however a 6-spine limitation with fewer leaf nodes (links are required for VFL ports).

When downstream nodes such as other switches or hypervisors are multi-homed to all VC units through an LACP aggregate, node-to-node traffic requires a single hop across a single VC unit and does not need to traverse the VFL. What this means is that traffic is forwarded at wire-rate and the VFL is not a bottleneck. This is because traffic forwarding to LAG member ports gives preference to local ports over remote (across the VFL) ports. This is true for unicast traffic only. Broadcast, Unknown-unicast, and Multicast (BUM) traffic may have to be forwarded across the VFL.

In case the hypervisors are not attached using LACP Link Aggregates, then Spanning Tree Protocol (STP) is required since the broadcast domain is required to be extended.

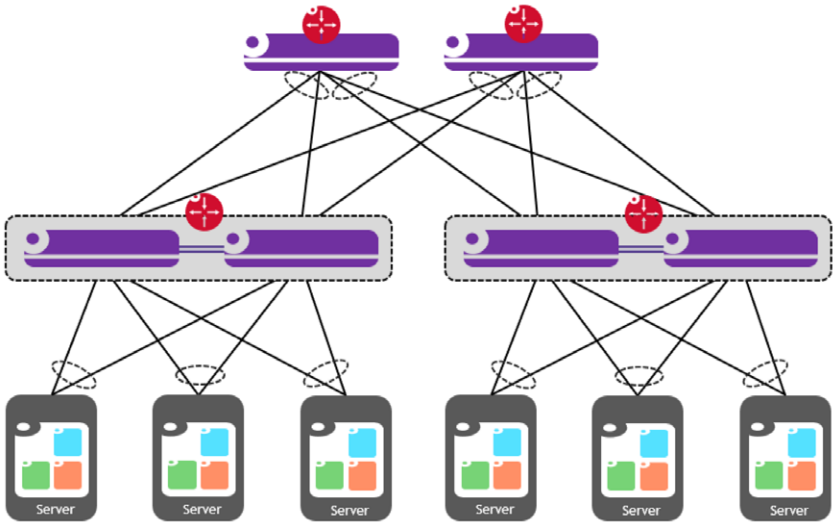
STP configuration

When required to configure STP in a Layer 2 spine-and-leaf architecture, you should ensured that only spine node(s) are configured as Root Bridges and that the root-guard feature is enabled on all downlinks. Since we are using the spine switches as a single VC, there will only be a single instance of STP running and no leaf-to-spine link blocking.

Layer 3 VC-based architecture

The design shown in **Figure 8** is a fully-routed architecture based on a VC configured on the leaf nodes. Point-to-point routed interfaces are configured between nodes, and you can configure a routing protocol such as OSPF, IS-IS, or BGP. Routing Designs within the DC are covered in a later section. Hypervisors can be configured with a LAG connection to a TOR VC leaf switch as discussed earlier in the Hypervisor attachment section. Spine nodes do not need to be in a VC. In the event you are using an Overlay architecture which uses VXLAN, each rack uses a different Virtual Tunnel Endpoint (VTEP) subnet which benefits BUM replication due to hierarchical replication. We will discuss VTEPs and BUM replication in the upcoming Overlay Architectures Considerations section. Access VLANs can be the same across all racks since it is locally significant and not stretched. Templates can be used for provisioning TOR leaf switches for easier management. VTEP IP addresses are assigned through DHCP, therefore the DHCP relay should be configured on the gateways. Use of LAG between the leaf nodes and the spines optimises the routing table and eliminates the need to forward traffic across the VFL.

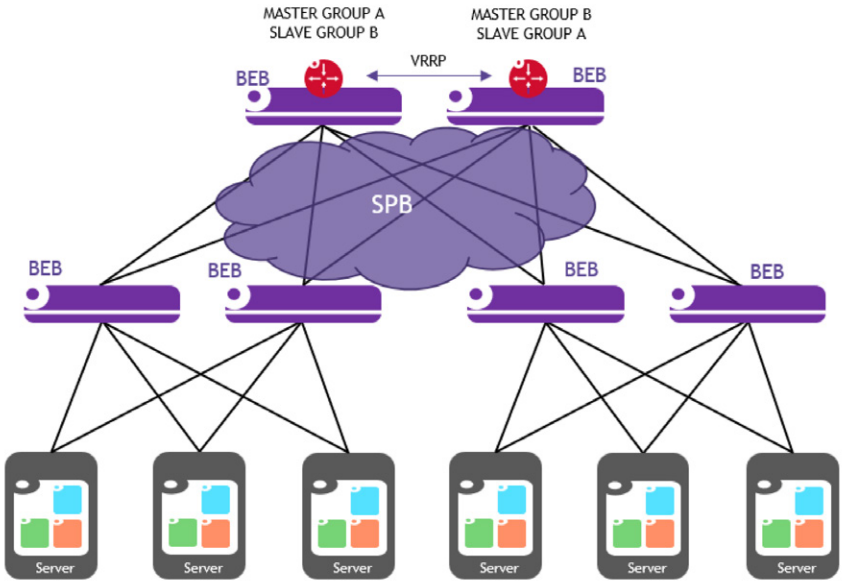
Figure 8 - Layer 3 fully routed VC leaves architecture



SPB-based architecture

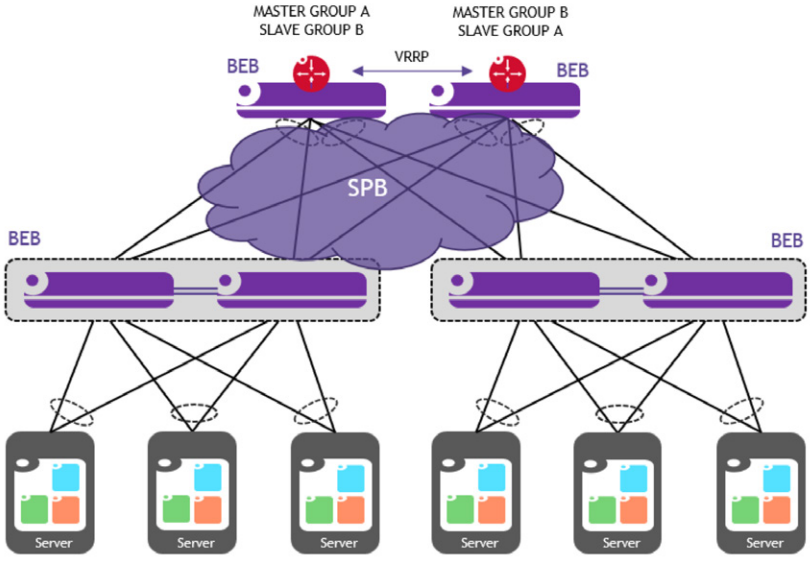
In the SPB-based design with standalone leaves, as seen in Figure 9, hypervisors are connected using NIC-teaming which performs per-VM load balancing. There will be as many BVLANs as there are spines. Dynamic User Network Profile (UNP) and Service Access Points (SAPs) will be configured on leaf nodes per the different Customer VLAN (CVLAN) of the VMs on the hypervisors, which will be load balanced across the BVLANs by default. To achieve load-balancing on the spines, VRRP is required on the spine nodes and subnets will be split between spines. There will be as many VRRP groups as there are spines, and each group priority will be set such that each spine is a master of one group. This will evenly balance traffic across BVLANs and VRRP groups. In case you are not using LACP Link Aggregate in the hypervisors, you should configure STP and enable LBD on the BEB leaf nodes.

Figure 9 - SPB-based architecture



Hypervisors can be configured with LAG connection to the leaves per the design in **Figure 10**. Here, the leaves will be configured as VCs of two leaf nodes, causing them to be considered as one logical node. SPB, by default, will only use one of the links from a VC connected to separate spines, which is not optimal. To load-balance traffic evenly, links should be bundled in a LAG from the spines to the leaves. To avoid unicast traffic forwarding on the VFL, each leaf node should have a link to each spine, and all hypervisors should be dual homed to every leaf. For routing, SPB and VRRP will be configured the same as the previous design.

Figure 10 - SPB-based architecture with VC leaf nodes



Split-site design

Some data centre sites are in separate physical locations, but are in close proximity whereby they can be designed and connected similar to a single-site design. Each site will have independent power and cooling, among other requirements.

SPB-based architecture

SPB-based split-site design, as shown in **Figure 11** and **Figure 12**, are similar to the SPB-based single-site design and the same comments apply.

Figure 11 - Split-site SPB-based architecture

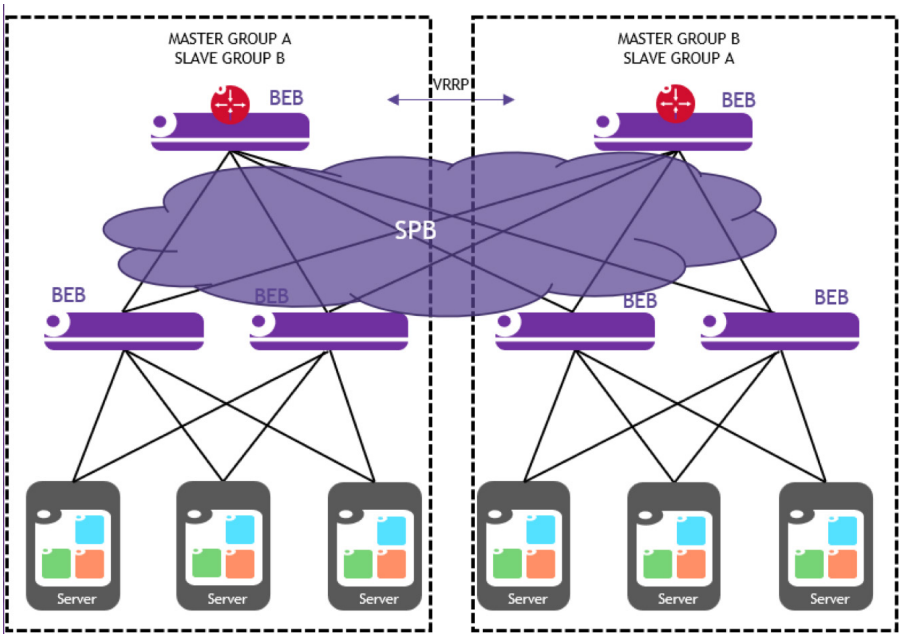
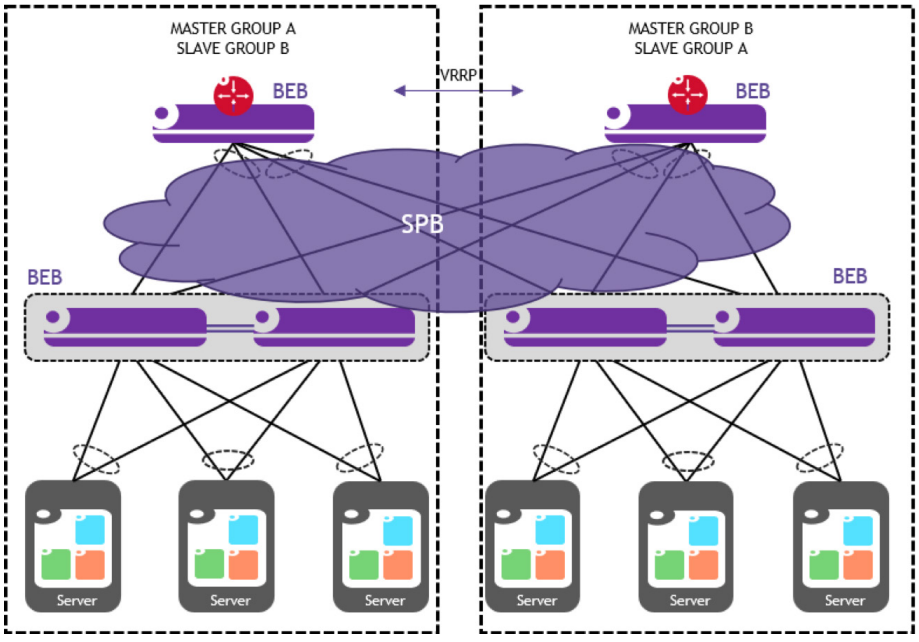


Figure 12 - Split-site SPB-based and VC leaf nodes architecture



Layer 2 VC-based architecture

For VC-based architectures, as shown in **Figures 13** and **14**, even though the sites may have independent power and cooling, you run the risk of a split-brain scenario since they have a single control brain. This split scenario is disruptive to the network as the conflicting MAC and IP addresses can lead to Layer 2 loops and L3 traffic disruption. Solutions such as Virtual-Chassis Split Protection (VCSP) and Out-Of-Band (OOB) Ethernet Management Port (EMP) Remote Chassis Detection (RCD) can be considered to detect and mitigate the split-brain scenario. In reality, however, usually fibre connections run in the same fibre conduit and a fibre cut will cause all fibre connections between both sites to be cut, resulting in a split-brain scenario where both nodes will act as Master and host the same subnets, causing a disruption to the network.

Figure 13 - VC-based architecture with standalone leaves

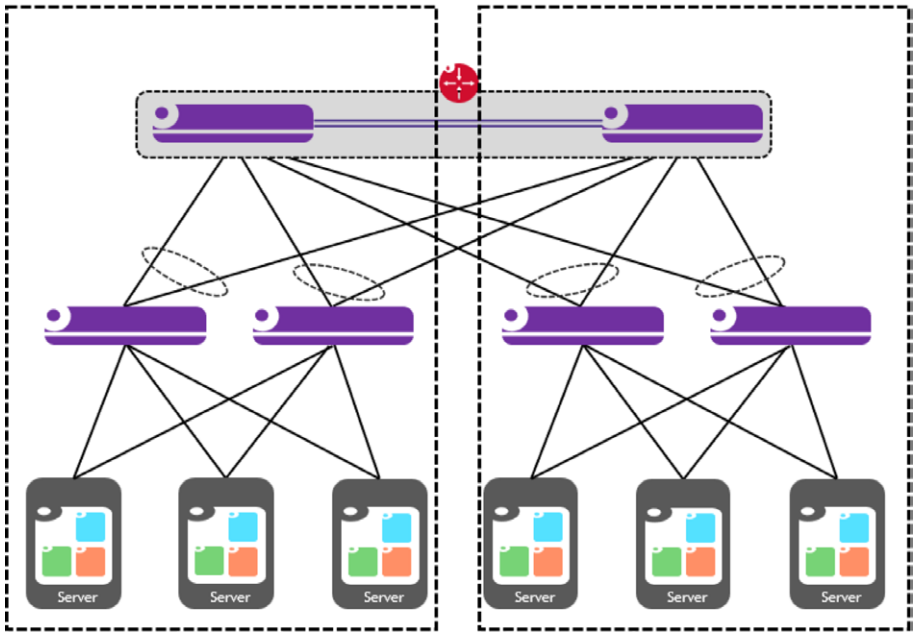
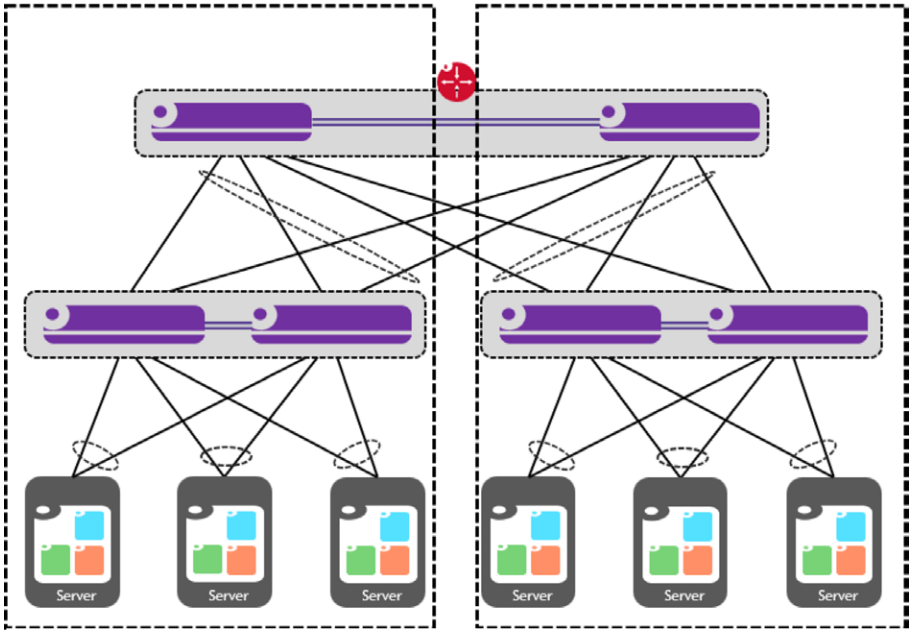


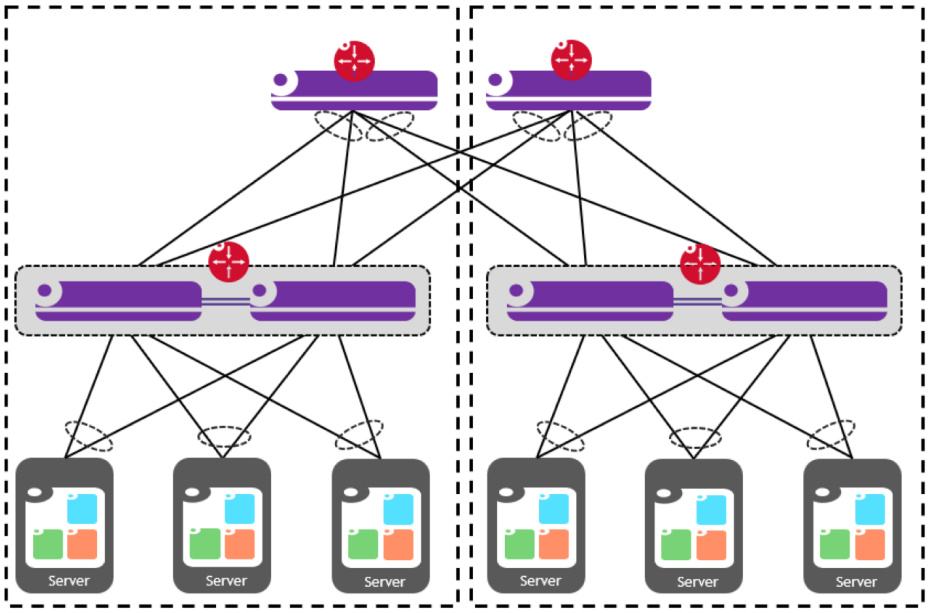
Figure 14 - VC-based architecture with VC leaves



Layer 3 VC-based architecture

To avoid a split-brain scenario you may use a VC on the leaf nodes or separate spine nodes with a Layer 3 fully-routed architecture, as shown in **Figure 15**. The same comments apply as a single-site VC-based L3 fully routed architecture.

Figure 15 - Split-Site fully routed VC-based leaf nodes architecture



Multi-site designs

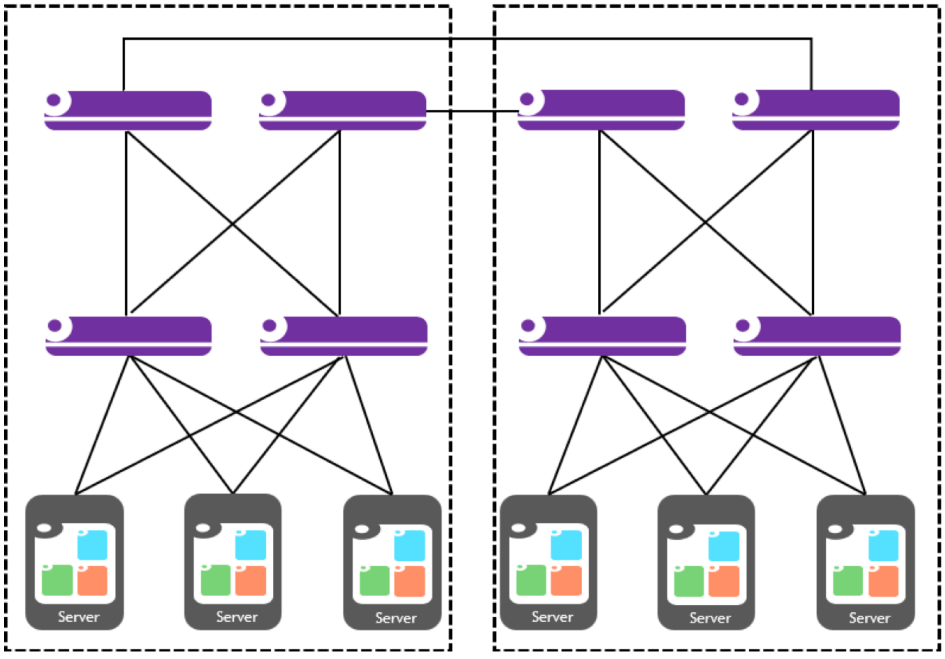
Border spine versus border leaves DCI

This type of design requires a Data Centre Interconnect (DCI), whether dark fibre connectivity or optical fibre connectivity such as Dense Wavelength-Division Multiplexing (DWDM). It could also be configured through a Layer 2 service such as Pseudowire (PW), Virtual Private LAN Service (VPLS), or Multi-Protocol Label Switching (MPLS).

Considerations need to be taken when designing a two-site architecture with DCI:

- When using dark fibre, DWDM or PW, LACP can be configured between both sides. This depends on the Service Provider (SP), otherwise it can be configured statically.
- MTU needs to be verified since DC Traffic may have overhead such as 802.1Q tags, or a service such as VXLAN or SPB
- Traffic can be encrypted using MACSec between the two sites

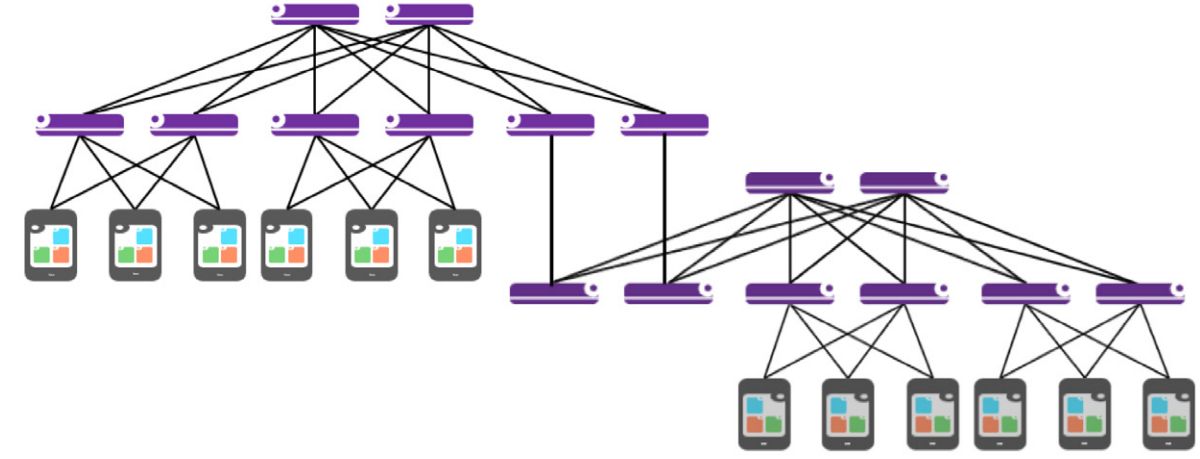
Figure 16 - Two-site border spines DCI topology



The design shown in **Figure 16** is optimal when there are two spines since there are only two hops between any pair of leaves, and two DCI links interconnecting the spines. If four spines are required, it will require four DCI links and will be costly and impractical. Otherwise if two DCI links are used only, this will cause the traffic to be unbalanced and there will be a variable number of hops between leaves at different sites.

Another consideration is to connect the two sites using DCI in border leaves as shown in **Figure 17**. This will have a balanced traffic flow. However, with two DCI links we will have four hops between the different sites.

Figure 17 - Two-site border leaves DCI topology



Routing

The standard dynamic protocols used commonly for routing are OSPF, IS-IS and BGP. The Open Shortest Path First (OSPF) protocol is a link-state routing protocol widely used in large enterprise networks. Intermediate System to Intermediate System (IS-IS) is also a link-state routing protocol, but is more common in service provider networks. Both protocols compute "shortest-path trees" for routes using Dijkstra's algorithm. They have fast-convergence times and are generally easier to deloy than BGP.

Border Gateway Protocol (BGP) is an Exterior Gateway Protocol (EGP) path-vector routing protocol used to exchange reachbility information between Autonomous Systems (ASs). BGP allows for granular control of network policies and is used commonly in service provider networks.

Routing within the data centre

Layer 2-based architecture

In case a Layer 2 based design is selected where inter-VLAN routing is performed at the spines, then you may use static or any dynamic routing protocol at the spines to route towards the outside world.

Layer 3-based architecture

In the event a Layer 3 architecture is selected, then there are multiple routing protocols options that can be used; including BGP, OSPF or IS-IS.

The design should be simplified by using as few routing protocols as possible and avoiding complex routing protocol features in the spine-and-leaf fabric. The use of multi-area OSPF designs or multi-level IS-IS and summarisation should not be used unless required.

In the upcoming sections we will discuss the multiple routing protocols and options that can be used when selecting a Layer 3-based design. We will assume that a VC-based design is used at the leaf nodes to allow for redundant server connectivity.

Layer 2 versus Layer 3 architecture

If a Layer 3-based design and VM Mobility is required, the bridging domain must be extended using an overlay technology such as VXLAN. VXLAN requires a Layer 3 underlay transport network as it runs over IP/UDP. This will be further discussed in the Overlay Architecture Considerations section. In case a redundant server attachment is considered for a Layer 3-based network, then a VC should be configured in the leaf nodes.

A Layer 2 architecture with VC on the leaf nodes and an SPB fabric provides you with a simplified architecture with inherent VM mobility and redundant hypervisor attachment.

Using a Layer 2-based design without SPB allows for VM mobility, but STP is required on the entire fabric and the DC will be in a single flooding and failure domain. Inter-subnet traffic would be required to cross the fabric even if the destination traffic lives in the same hypervisor. This can be better improved when using VC on the spines nodes since it provides active-active and hash-based load balancing at both Layer 2 and Layer 3.

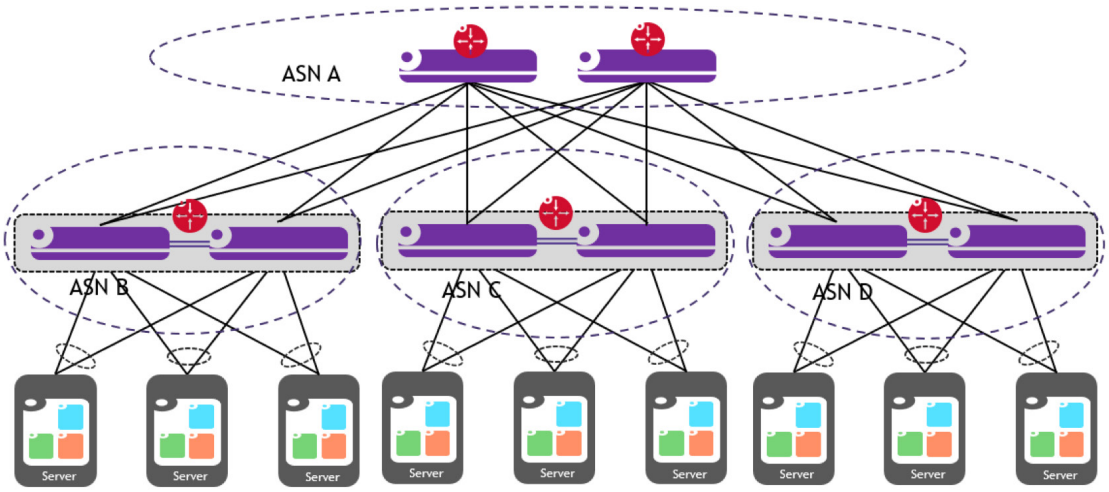
Routing convergence design

In case a Layer 3-based architecture is selected, you can improve the convergence time of detecting a link failure by using Bidirectional Forwarding Detection (BFD). Depending on the signalling of the physical layer is slow to inform the upper Layer protocols. BFD can detect link failures within milliseconds and help improve convergence time.

eBGP

A recommended design option in a Layer 3 routed network is to use eBGP between nodes where all the spine switches are in a single Autonomous System (AS) and the leaf switches are in separate ASs. Using private AS numbers 64512-65534 is also recommended to avoid leaking out of internal BGP prefixes to an external network. This is shown in **Figure 18**. As mentioned earlier, we will assume Layer 3 VC-based design is selected to allow for redundant server connectivity.

Figure 18 - eBGP routing design - Single AS spines



This design is best selected to avoid the "Path Hunting" issue, which is similar to the count-to-infinity issue in distance-vector protocols. This is caused when there are a large number of protocol updates when a link fails. Since a switch does not know the physical state of every other switch in the network, it cannot determine if the prefix is unreachable or not using another path. This is when the switch tries to find reachability to the prefix through the other available paths. This overhead can increase as the network scales. You can avoid this issue by using a single AS in the spines. Each spine will have a single path to each prefix (or two in case you are using a hypervisor double-attachment with NIC teaming). In case the path is received from other leaf switches, they are rejected since it will consider it as a loop since the AS PATH contains the same AS.

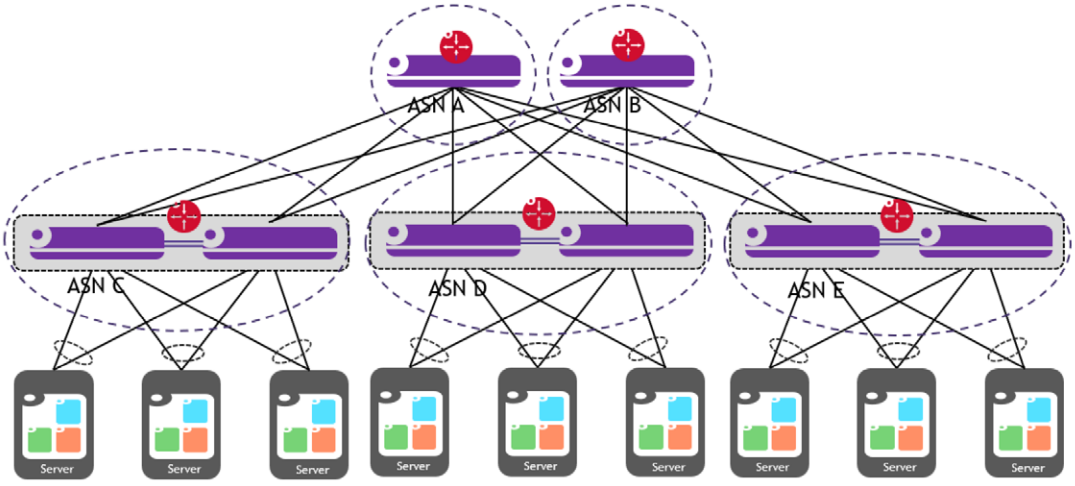
Fine-tuning the convergence time is also required in case BGP is selected as the routing protocol. The default timers used in BGP are not optimised for DC routing. Multipath should be enabled to allow for load-balancing. When multipath support is enabled and the path selection process determines that multiple paths are equal when the router-id is disregarded, then all equal paths are installed in the hardware forwarding table. When multipath support is disabled, only the best route entry is installed in the hardware forwarding table. Adjusting the default BGP advertisement-interval should also be considered to allow for fast convergence. This interval is used by the peer to send BGP UPDATE messages to external peers. Other BGP timers that should be adjusted are KEEPALIVE and Hold time. The keep alive interval can never be more than one-third the value of the hold-time interval. When the hold interval is reached without receiving keep alive or other update messages, the peer is considered dead. The last feature that should be enabled is fast external failover (FEFO). When enabled, this command allows BGP to take immediate action when a directly connected interface, on which an external BGP session is established, goes down. Normally BGP relies on TCP to manage peer connections. FEFO improves upon TCP by resetting connections as soon as they go down.

Table 2 - Recommended BGP features

Parameter	Default value	Recommended value
Multipath	Disabled	Enabled
KEEPALIVE	30 sec	3 sec
Hold time	90 sec	9 sec
Advertisment interval	30 sec	0 sec
FEFO	Disabled	Enabled

Another option is to use different AS numbers for the spines, as shown in Figure 19:

Figure 19 - eBGP routing design - Separate AS spines

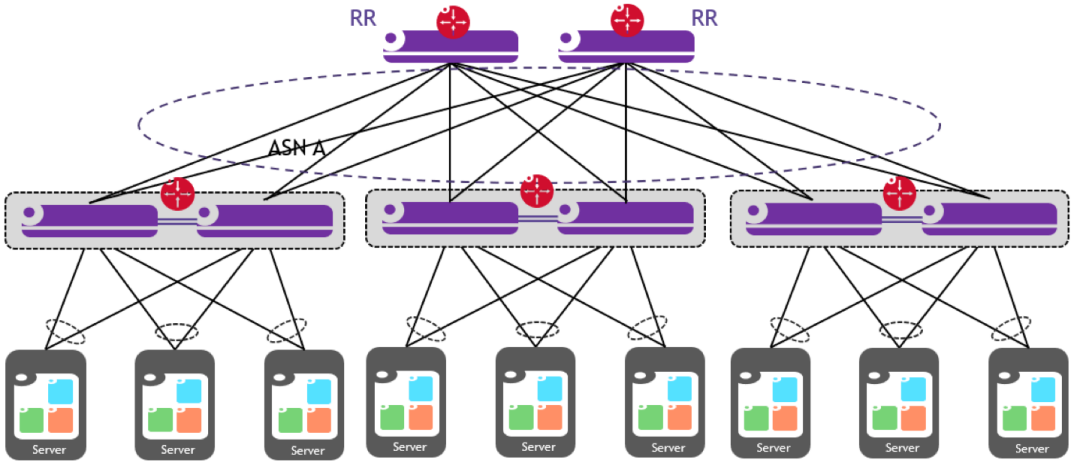


To avoid the "Path Hunting" issue, you can use route-maps to announce only locally originated routes on leaf switches and spine routes. The same comments for convergence timers applies to this design as well.

iBGP

Another option is to use iBGP with Route Reflectors (RR) on the spine nodes, as shown in **Figure 20**. Since iBGP requires either a full-mesh of peers or RR, using RR is more scalable in a DC design.

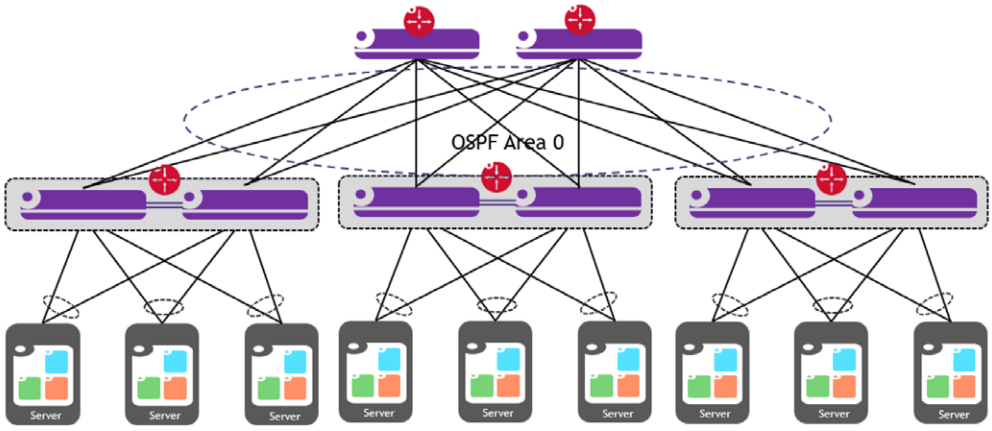
Figure 20 - iBGP design with Route Reflectors



OSPF

In case OSPF is selected, using a single-area OSPF configuration should satisfy most requirements. It is a best practice to configure point-to-point interfaces in the spine-leaf links and use ECMP for load-balancing.

Figure 21 - OSPF design single area



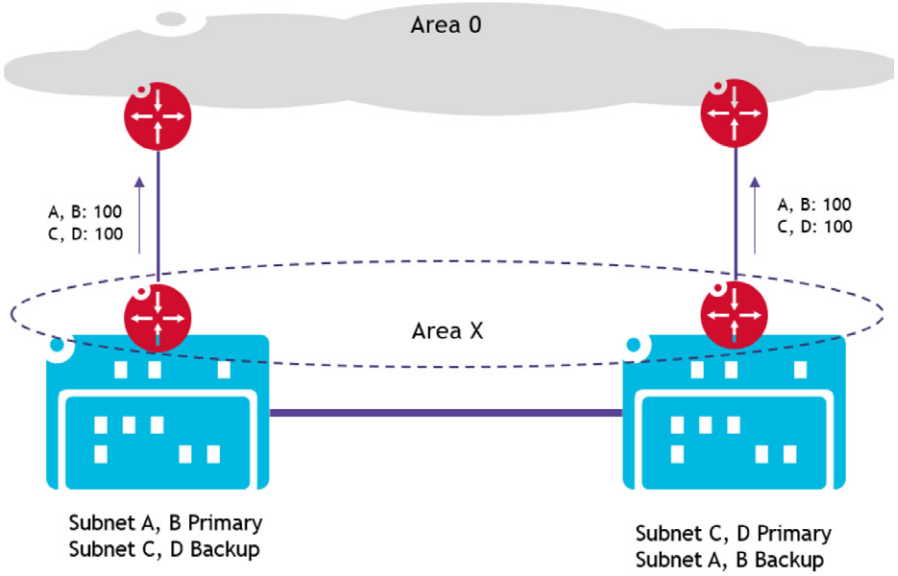
Routing into and out of the data centre

When routing traffic in and out of the DC, protocols such as OSPF/IS-IS should not be considered for several reasons:

- OSPF does not allow for prefix-based policy control
- The only metric is cost and is set on the interface and is not prefix-based
- With firewalls, it can only work as active-standby
- OSPF requires separate protocol instances for IPv4 and IPv6

OSPF can be acceptable when the DCI is very fast and not over long distances. Even with inefficient traffic flows, we will have plenty of bandwidth and low latency. OSPF can also be acceptable within the same building or campus.

Figure 22 - Two-site data centre OSPF routing design

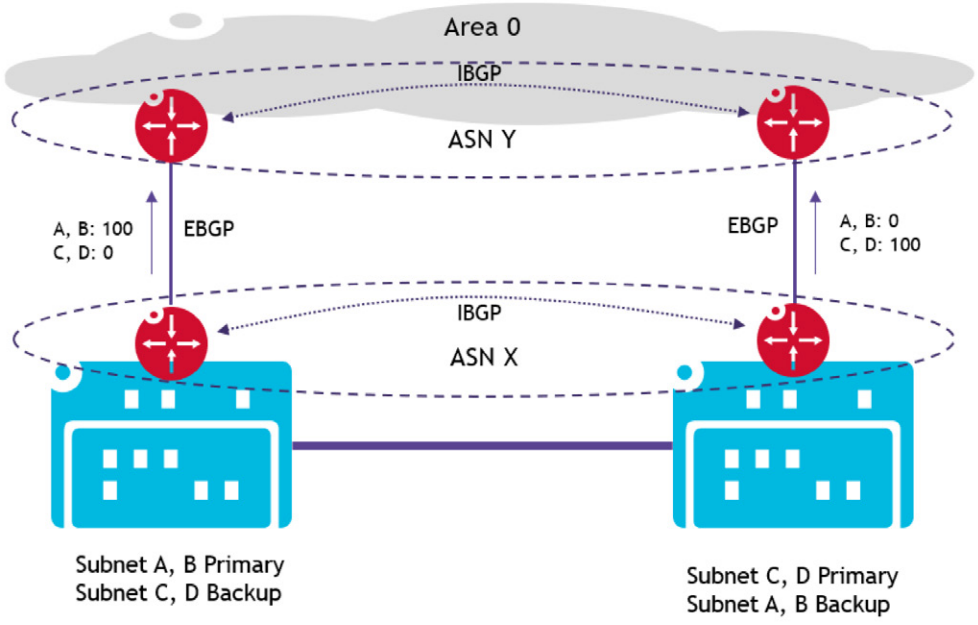


A better option is to use BGP for the following reasons:

- BGP allows for granular policy-control of prefixes
- It allows better load-balancing with firewalls
- It allows better load-balancing on per-prefix basis
- The same protocol (MP-BGP) can be used for IPv4 and IPv6 address families
- If SPB is used, whether on campus or DC side, it is not required to add another routing protocol within the site

In this design, as shown in **Figure 23**, iBGP will be used between the DC sites and eBGP between the DC and the campus. Traffic can be engineered for load-balancing using the Multi-Exit Discriminator (MED) and the local preference BGP attributes such that some subnets will be advertised with better preference from one DC and with a worse preference on the other to ensure both links are utilised.

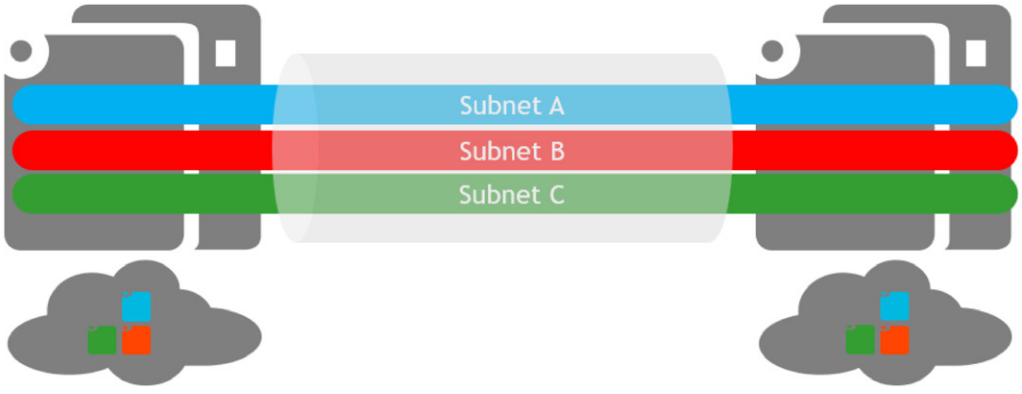
Figure 23 - Two-site data centre BGP routing design



Multi-site design

Global stretched subnets – L2 DCI

Figure 24 - Two-site data centre global stretched subnets



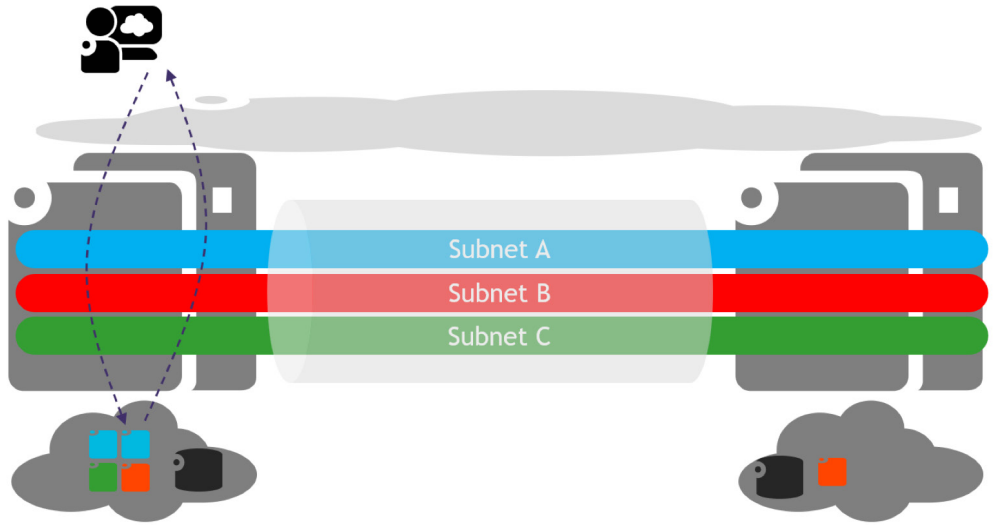
Stretching subnets or VLANs is sometimes considered between two-site DCs for several reasons:

- Subnet mobility for Disaster Recovery (DR) purposes
- VM Mobility used by automatic resource schedulers (such as Dynamic Resource Scheduler (DRS) for VMWare) between clusters to optimise compute resources and load balancing
- VM Mobility for maintenance purposes

Stretching the Layer 2 domain in general not only increases the failure domain, it also introduces other drawbacks, such as sub-optimal traffic flows, traffic tromboning and traffic bottlenecks in the DCI link.

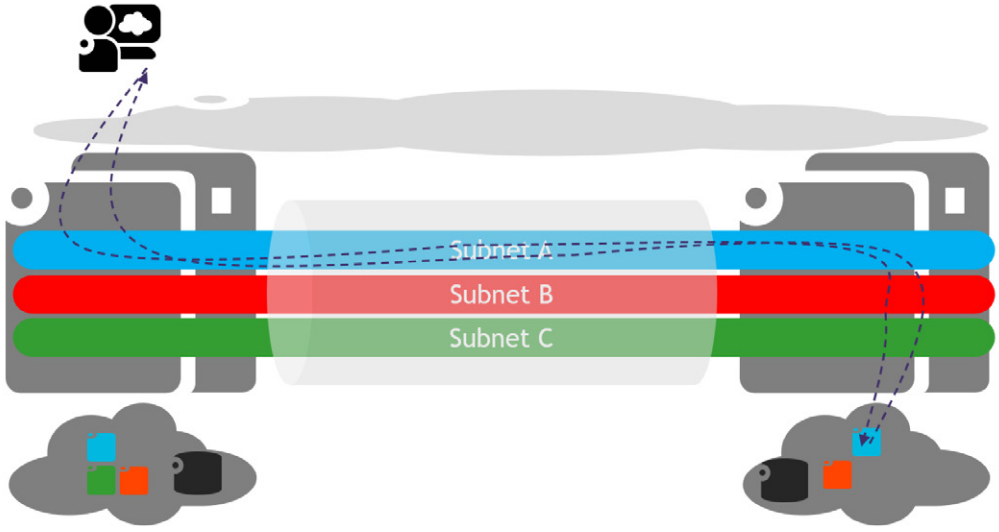
For example, if an external user is accessing an application hosted in DC-A, traffic will flow normally in both directions, as shown in **Figure 25**.

Figure 25 - Two-Site data centre global stretched subnets - External user traffic



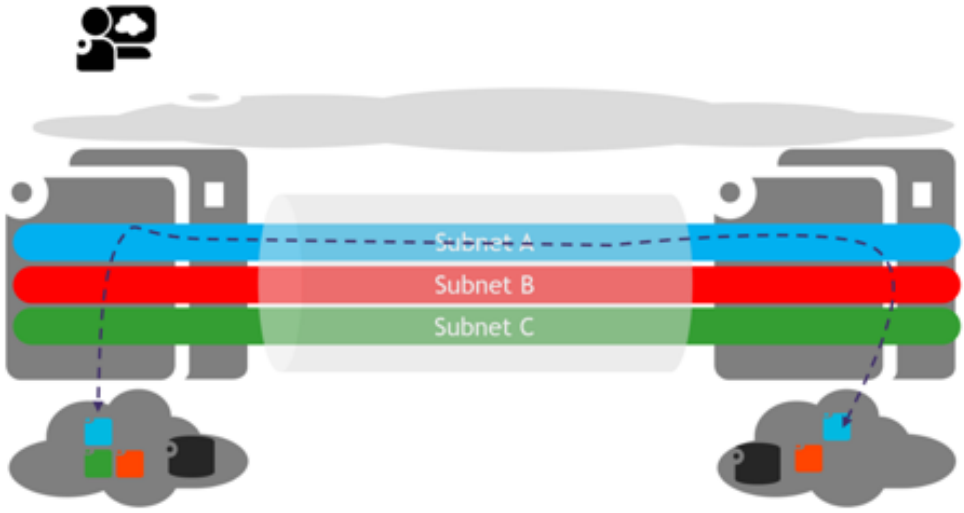
Let us consider that the VM was moved from DC-A to DC-B for maintenance purposes or DRS determined it is optimal for resource utilisation. Since firewalls and load-balancers require session persistence, traffic will still arrive in DC-A and traverse the DCI connection to reach the application in DC-B. This is shown in **Figure 26**.

Figure 26 - Two-site data centre global stretched subnets - Suboptimal external user traffic



Inter-site traffic between different VMs in the same subnet will also traverse the DCI connection, as shown in **Figure 27**.

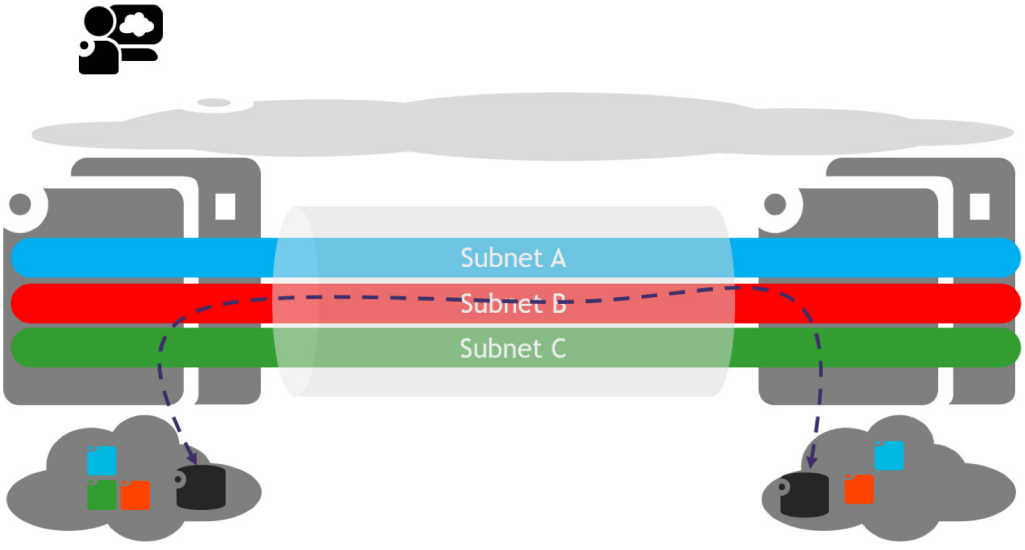
Figure 27 - Two-site data centre global stretched subnets - Inter-VM traffic



If the application, which was moved to DC-B, requires reads or writes to the database, which was originally in DC-A, this will also traverse the DCI connection. Even if there is a replica of the database in DC-B, this will still require primary writing to the DC-A database.

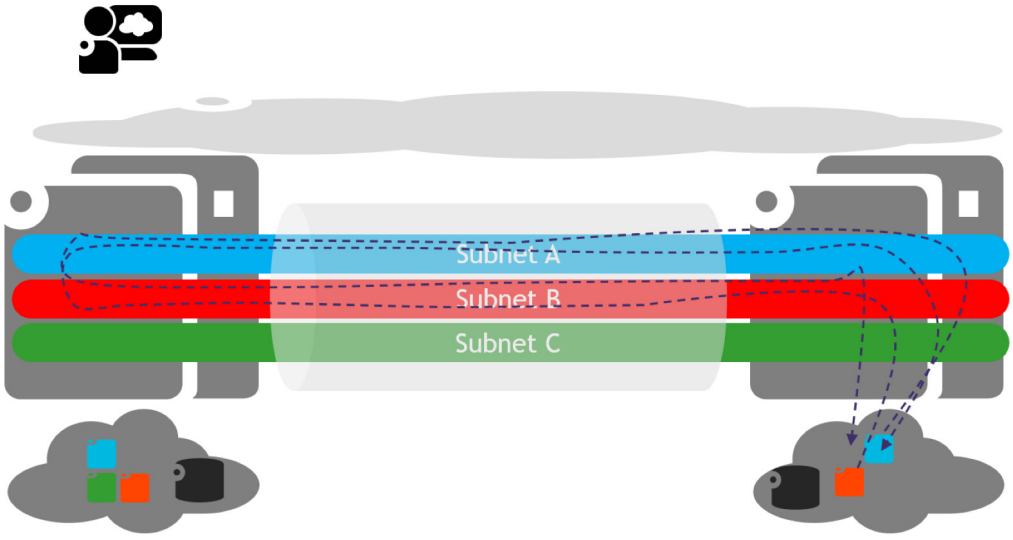
For applications, continuous access to storage is required, and moving VMs between different sites requires synchronous replication between sites. This will also traverse the DCI connection as shown in **Figure 28**. Also, depending on the storage technology, it is possible that DC-B storage will be read-only and writes need to be done on DC-A storage.

Figure 28 - Two-site data centre global stretched subnets - Inter-site storage replication



In case the application which was moved to DC-B needs to communicate with a VM in DC-B, traffic flows may suffer from the "trombone" effect, as shown in **Figure 29**. Since the VM is required to communicate with the default gateway, which may still be in DC-A, this will cause traffic to traverse the DCI connection to DC-A and route back to DC-B. However, this can be solved with active-active routing with VC or active-active VRRP.

Figure 29 - Two-site data centre global stretched subnets - Trombone effect

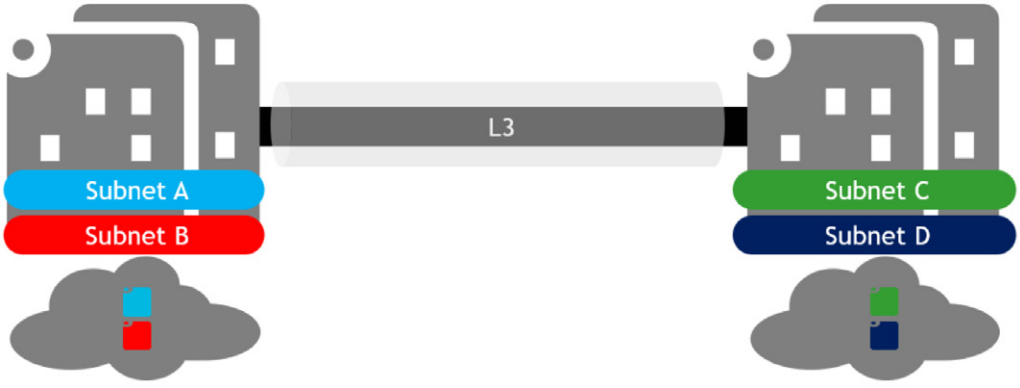


The DCI link usually does not have enough bandwidth or capacity to handle stretched subnets. Moving a few VMs with a few Terabytes of RAM may take hours or days depending on the speed and amount of data moved and BUM traffic will be flooded across both sites causing DCI congestion.

In conclusion, subnet stretching may look appealing until the downsides are analysed. This design may no longer be required since clustering no longer requires Layer 2 connectivity.

Local subnets – L3 DCI

Figure 30 - Two-site data centre local subnets



Using local subnets for each site is a better design solution for multiple reasons:

- Stretching subnets is no longer required since clustering no longer requires Layer 2 connectivity and L3 clustering across sites is possible. This can also be achieved using a load balancer (LB).
- There is better fault isolation. For example, in the event of a broadcast storm, it will not extend to the other site.
- Traffic flows are optimised and DCI connection will be less congested
- Inter-site non-live VM migration can be achieved by changing DNS entries

This design maybe optimal as highlighted above, but will be less appealing to the Server Infrastructure team.

Local stretched subnets – L3 DCI

A compromise solution is to use local subnets which can be stretched when required. This solution requires the use of VRRP between both sites. DC-A will be the master of its own subnets, and DC-B will be the Master of its own subnets, but the slave interfaces are administratively disabled on the other DC. The slave interfaces will be enabled only when and if required, such as a maintenance or a DR event. This is shown in **Figure 31** before stretching the subnets and **Figure 32** after stretching the subnets.

Figure 31 - Two-site data centre local stretched subnets - Before stretching

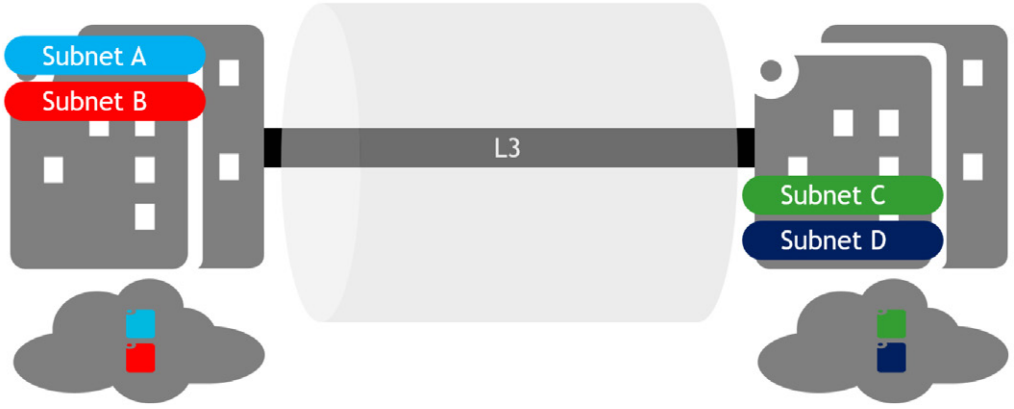
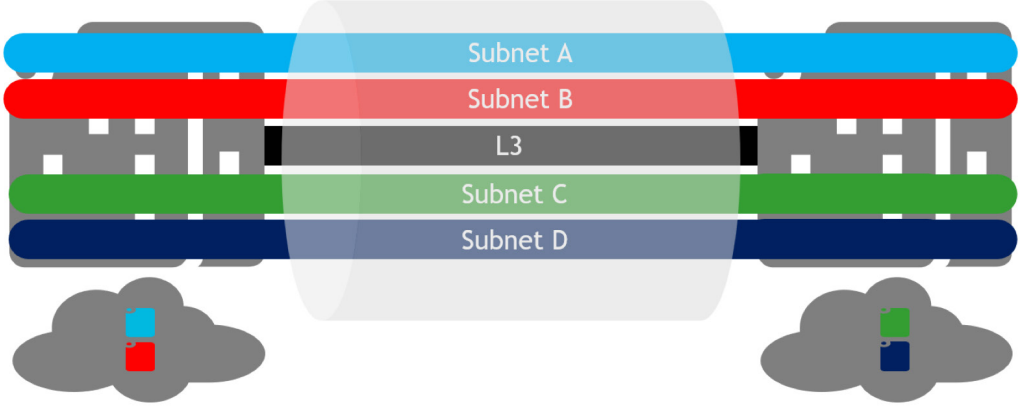


Figure 32 - Two-site data centre local stretched subnets - After stretching



In the event we are required to stretch the subnets and we are using SPB, we will move the VMs from DC-A and administratively enable the DC-B interfaces and the services will self-configure. If we use VC and LAG it will work in the same way, provided we use UNP and Multiple VLAN Registration Protocol (MVRP).

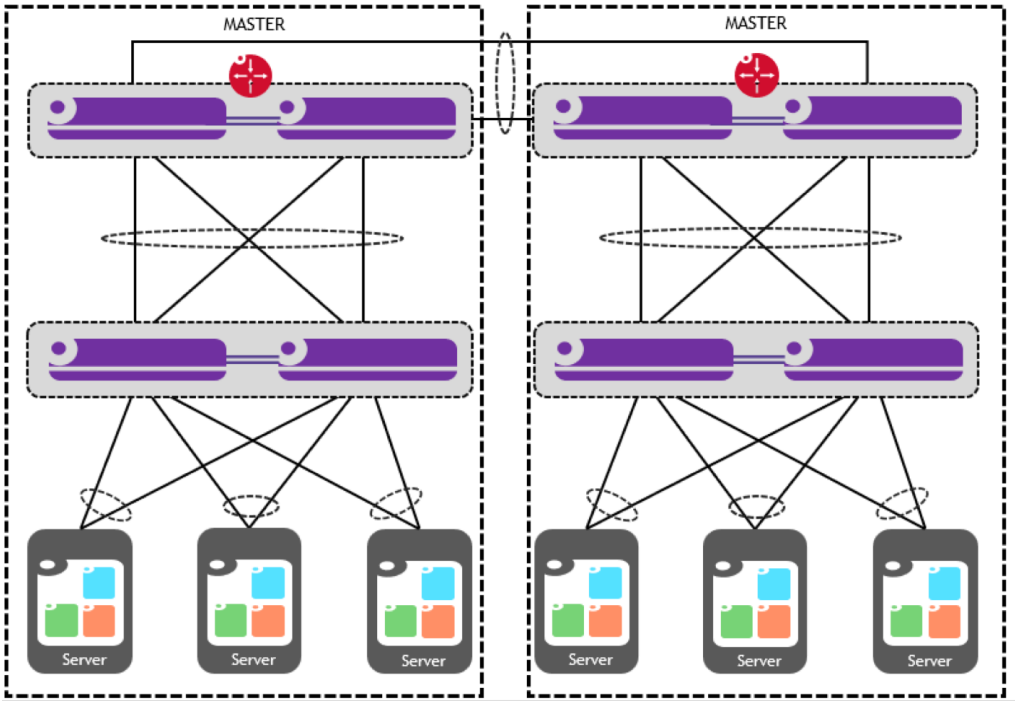
The same caveats apply however if, for example, we are required to move many VMs across the sites, as in the event of a DR event, where the DCI connection will be congested due to the large amount of data.

Two-site data centre designs with DCI – VC-based

If stretched subnets are still chosen, the design shown in **Figure 33** and **Figure 34** is an option. When using dark fibre, DWDM, or pseudo wire for the DCI connection. LACP Link Aggregate, if supported, can be configured between both sides. Otherwise, it can be configured statically. To optimise traffic flow, we can use active-active VRRP. This is achieved by configuring both sites as VRRP Master. This can be done by blocking VRRP Hellos and ARP requests over the DCI with Access Control Lists (ACLs). ARP requests sent locally will be replied to by the local gateway VC.

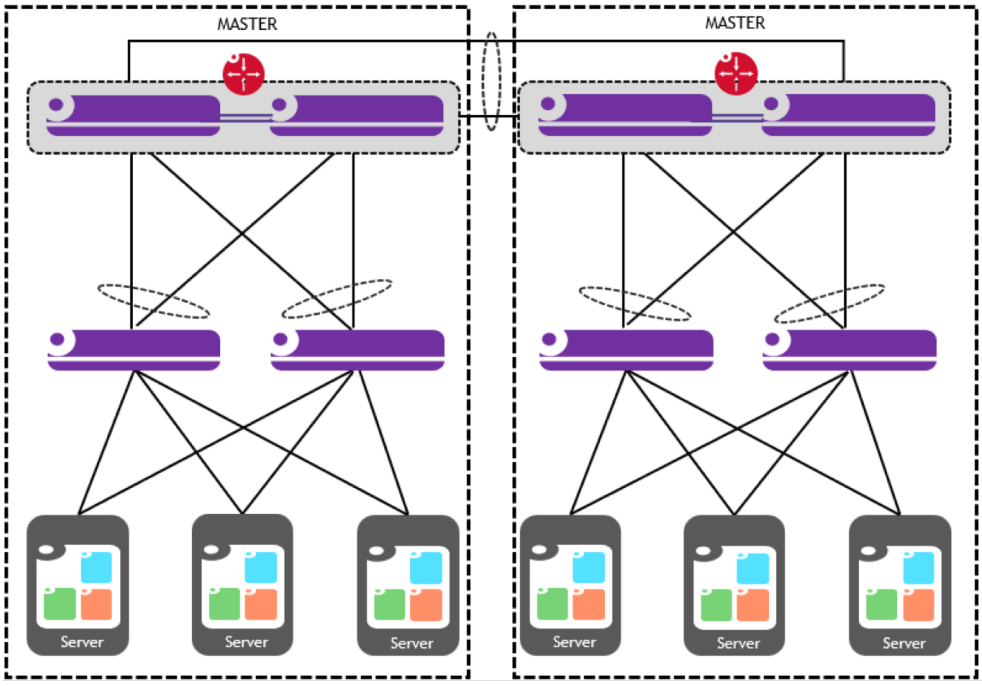
To avoid traffic traversing the VFL, we need to attach the hypervisor to both leaves and attach each leaf to all spines as shown in the design in **Figure 33**.

Figure 33 - Two-site data centre design - VC-based with stretched subnets, active-active VRRP and local gateway and VC leaves



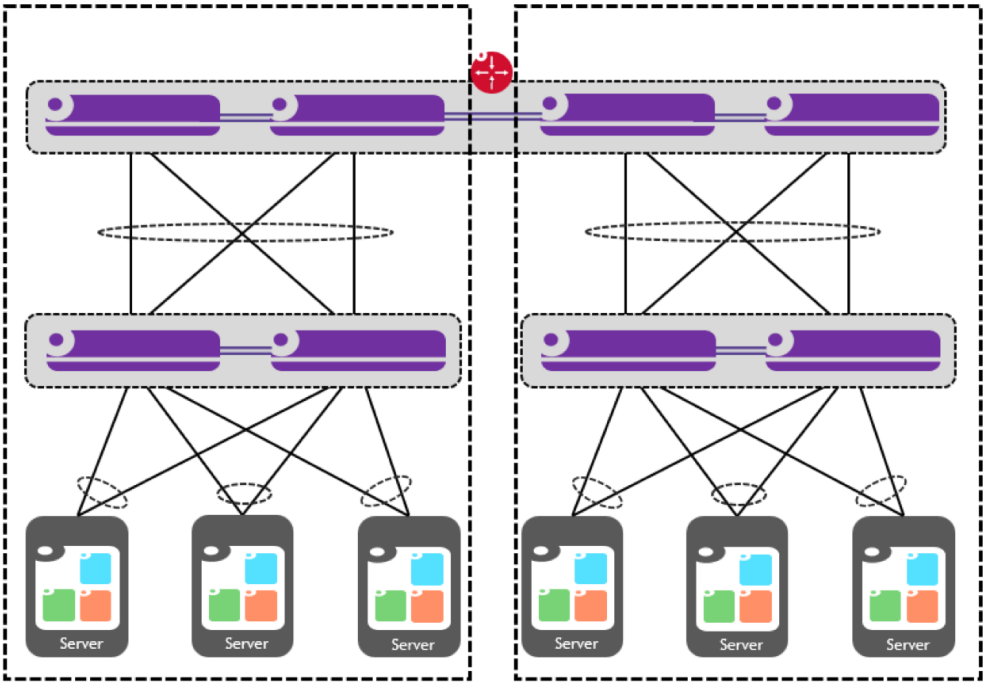
We can also consider standalone leaves without dual-attached hypervisors, as shown in the design in **Figure 34**.

Figure 34 - Two-site data centre design - VC-based with stretched subnets, active-active VRRP and local gateway with standalone leaves



Both designs will still have the previously mentioned traffic optimisation issues regarding stretched subnets, such as storage and database traffic flows and traffic into and out of the DC.

Figure 35 - Two-site data centre design - VC-based with stretched subnets and full VC



The design with Full VC shown in **Figure 35** can also be an option but will have the same caveats mentioned earlier in the Split-site VC-based design. The only difference is that the split-brain scenario can only be detected using OOB EMP RCD, and not using VCSP, since leaf nodes are only local and not connected to the other site. Even if a split-brain scenario is detected, this will cause the entire site to go down.

Two-site data centre designs with DCI – SPB-based

For an SPB-based design, you cannot use ACLs for active-active VRRP by configuring both sites as the VRRP Master. This is because traffic across the DCI is MAC-in-MAC encapsulated, and ACLs are required to look inside the Provider Backbone Bridging (PBB) encapsulation, which is not supported. Traffic could be optimised by load-balancing VRRP Groups such that each spine is the Master for one group of subnets and by using local subnets without stretching as shown in **Figure 36**. You could also use a VC in the leaf nodes with LAG as shown in **Figure 37**. In case it is supported in the future, active-active VRRP with SPB or Anycast Gateway design would be a better choice.

Figure 36 - Two-site data centre design - SPB-based with local subnets, VRRP and standalone leaves

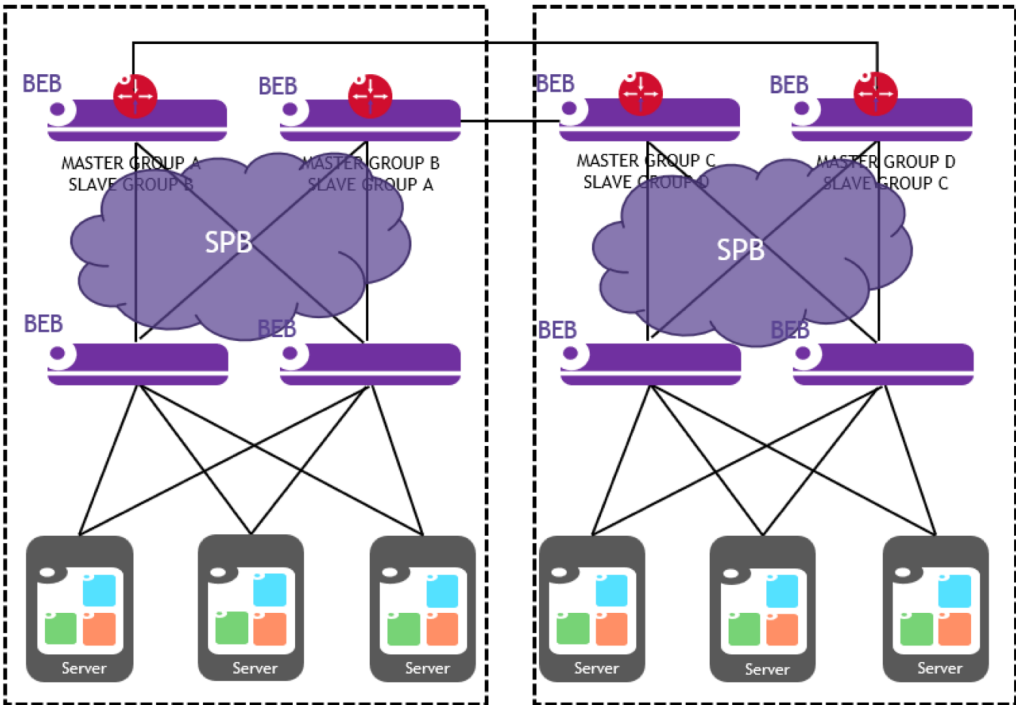
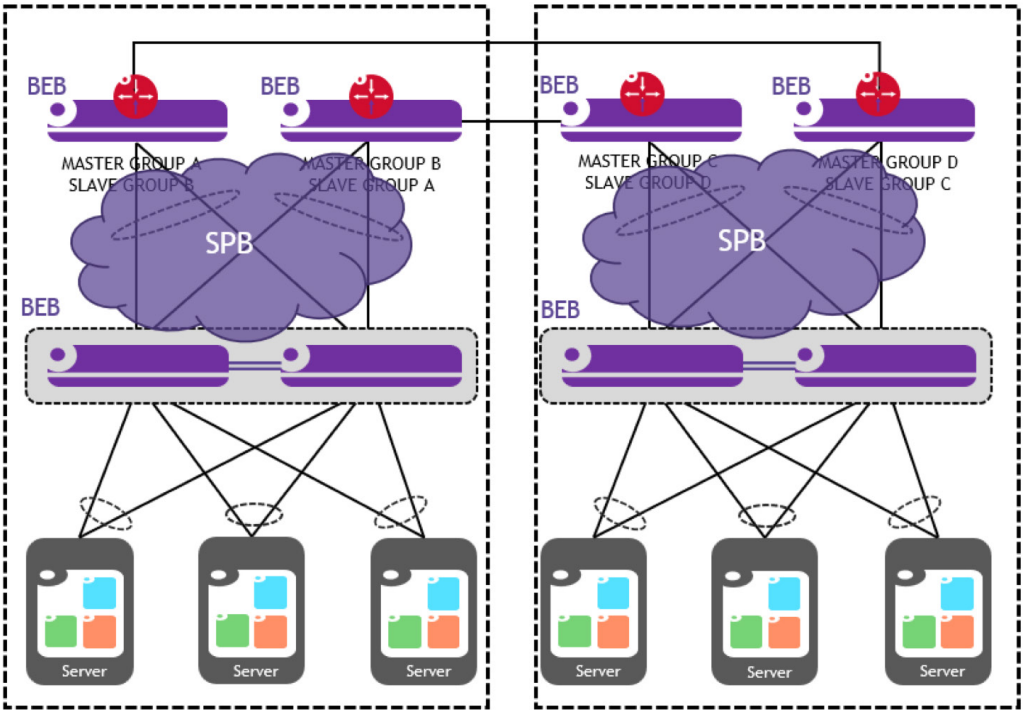


Figure 37 - Two-site data centre design - SPB-based with local subnets, VRRP and VC leaves



Overlay architecture considerations

In an overlay network, data is transmitted over virtual tunnels between nodes. This encapsulation and decapsulation may result in extra overhead, which should be taken into considerations. Overlay networking allows for high scalability and multi-tenancy. The underlay connectivity could be a Layer 2 or Layer 3 network. The underlay network usually does not have the visibility or intelligence to keep track of end-point connections.

There are many overlay network protocols which are implemented that require certain considerations in the underlay for handling the overlay encapsulation. Some are Virtual Extensible Local Area Network (VXLAN) with or without Ethernet Virtual Private Network (EVPN), Network Virtualisation using Generic Routing Encapsulation (NVGRE), Generic Routing Encapsulation (GRE), and Network Virtualisation over Layer 3 (NVO3). These technologies may require certain configurations in the underlay network for control plane communication and data plane forwarding of traffic.

Some of these overlay technologies are implemented in hardware, where the encapsulation is performed at the TOR leaf switches, while others are implemented in software at the hypervisor level.

BUM flooding is usually performed with IP multicast with the underlay network, head-end replication, or using service nodes (such as the case for VMWare NSX-T). If IP multicast is used, then multicast routing should be enabled in the underlay.

MTU

Considering that overlay technologies add extra overhead to the Ethernet frames, the Maximum Transmission Unit (MTU) will need to be adjusted accordingly in the underlay.

Management plane

In this plane, VLANs to Virtual Network Instances (VNI) are mapped. This can be configured statically on the hardware Virtual Tunnel End Point (VTEP) or using software-defined controllers. If hardware VTEPs are used, such as OmniSwitch models, Open vSwitch Database Management Protocol (OVSDB) is supported to allow management and configuration of the hardware VTEP using a management software such as NSX Manager and vCentre.

Control plane

Each technology uses a control plane mechanism to learn MAC addresses, to know where each end-point is connected. This is done through a control plane protocol, a controller such as VMWare vCentre (OVSDB), or through flood-and-learn mechanisms. Flooding can be done with IP multicast replication or through unicast (head-end) replication.

In the event multicast replication is required in the overlay network, VNIs are associated to multicast group addresses. When BUM traffic is received, it is encapsulated in the overlay technology frame where the destination IP is the multicast group of the VNI. All VTEPs with VMs in a given VNI join the multicast group and receive BUM traffic. Internet Group Management Protocol (IGMP) Snooping and Querier are configured on the hypervisor-facing interfaces. This is required to optimise the delivery of Layer 2 multicast traffic. Layer 3 Protocol Independent Multicast (PIM) protocol should also be enabled to ensure multicast traffic is also delivered to VTEPs in a different subnet from the source VTEP. PIM-Bidir is recommended since all VTEP devices communicate with each other. Each VNI should be assigned a multicast group and redundant Rendezvous Points (RPs) in the spine nodes should be used for load-balancing. In this mode the VTEP, MAC, or ARP tables are not maintained since the responsibility is given to the underlay. This replication mechanism is the same as Tandem replication. Another control plane protocol to map the MAC Address to the VTEP end-points is EVPN, which is a BGP-based control plane protocol.

Summary

We have presented different design topologies and their implications, whether VC-based or SPB-based topologies, as well as whether subnets should or should not stretched between data centre sites. Additionally, we have explored intra-DC and inter-DC routing, and the considerations that should be taken in the underlay architecture when deploying overlay technologies.

www.al-enterprise.com The Alcatel-Lucent name and logo are trademarks of Nokia used under license by ALE. To view other trademarks used by affiliated companies of ALE Holding, visit: www.al-enterprise.com/en/legal/trademarks-copyright. All other trademarks are the property of their respective owners. The information presented is subject to change without notice. Neither ALE Holding nor any of its affiliates assumes any responsibility for inaccuracies contained herein. © Copyright 2023 ALE International, ALE USA Inc.
All rights reserved in all countries. DID23052901EN (June 2023)

Alcatel·Lucent Enterprise