VXLAN bridging with MLAG

eos.arista.com/vxlan-with-mlag-configuration-guide/

By Alex

Contents [hide]

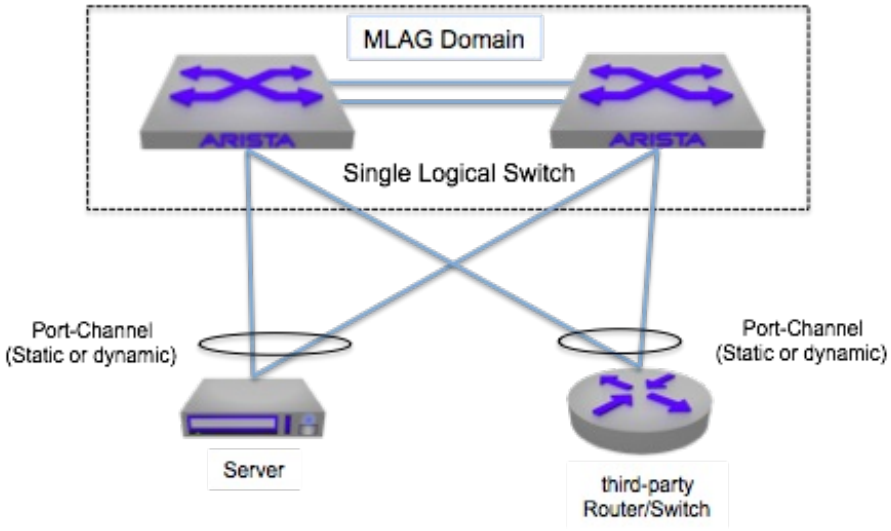
Introduction

This document describes the operation and configuration of VXLAN within an Multi-Chassis LAG (MLAG) deployment. The configuration and guidance within the document is based on the platforms and EOS release of table 1.0

Platform	Release	Functionality
Arista 7150S	4.14.5F	VXLAN bridging with MLAG support, using either HER or CVX control planes
Arista 7050X	4.14.5F	VXLAN bridging with MLAG support, using either HER or CVX control planes
Arista 7250X/7300	4.14.5F	VXLAN bridging with MLAG support, using either HER or CVX control planes
Arista 7280/7500	4.14.5F	VXLAN bridging with MLAG support, using either HER or CVX control planes

Arista MLAG technologyTable 1.0

Arista's Multi-Chassis LAG (MLAG) technology provides the ability to build a loop free active-active layer 2 topology. The technology operates by allowing two physical Arista switches to appear as a single logical switch (MLAG domain), third-party switches, servers or neighbouring Arista switches connect to the logical switch via a standard port-channel (static, passive or active) with the physical links of the port-channel split across the two physical switches of the MLAG domain. With this configuration all links and switches within the topology are active and forwarding traffic as no loop exists within the topology the configuration of spanning-tree becomes optional.



The focus of this document is the operation and configuration of an MLAG topology within a VXLAN overlay architecture.

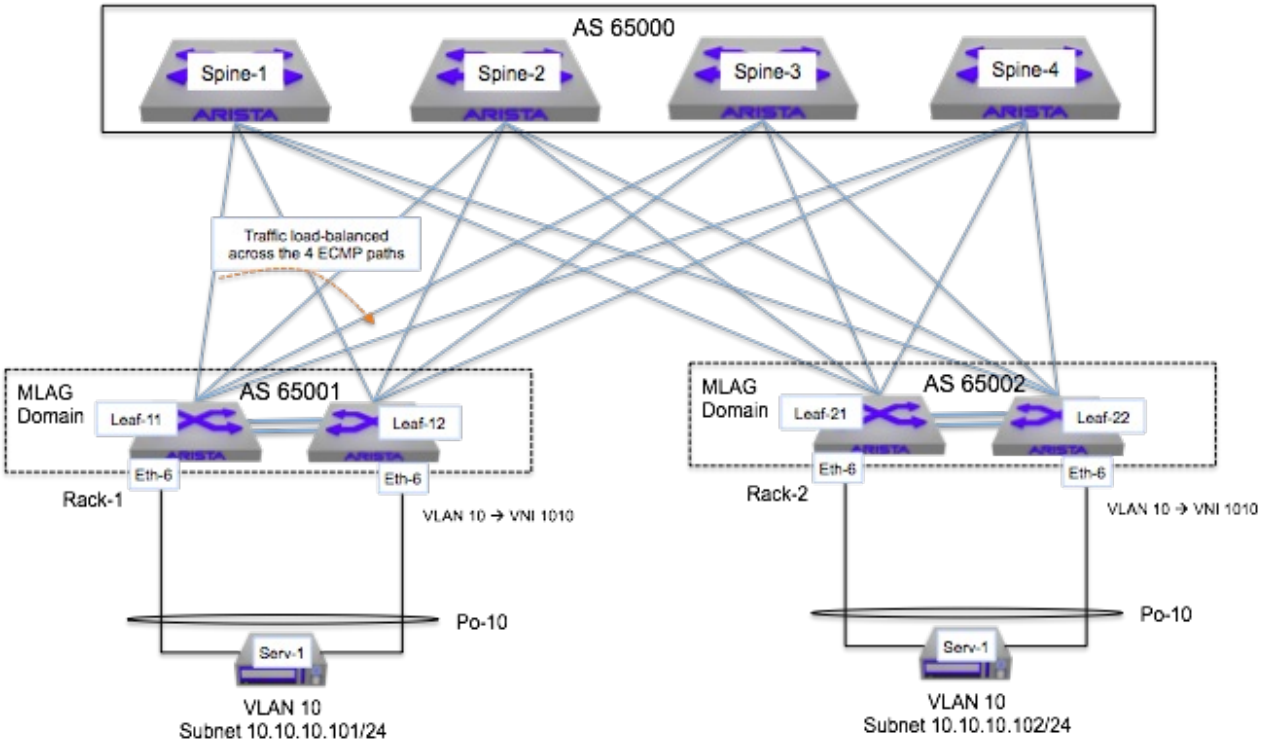
VXLAN with MLAG

The VXLAN protocol is an RFC (7348) standard, co-authored by Arista. The standard defines a MAC in IP encapsulation protocol allowing the construction of layer 2 overlay networks across a layer 3 IP infrastructure. The protocol is typically deployed as a data centre technology to create overlay network topologies for:

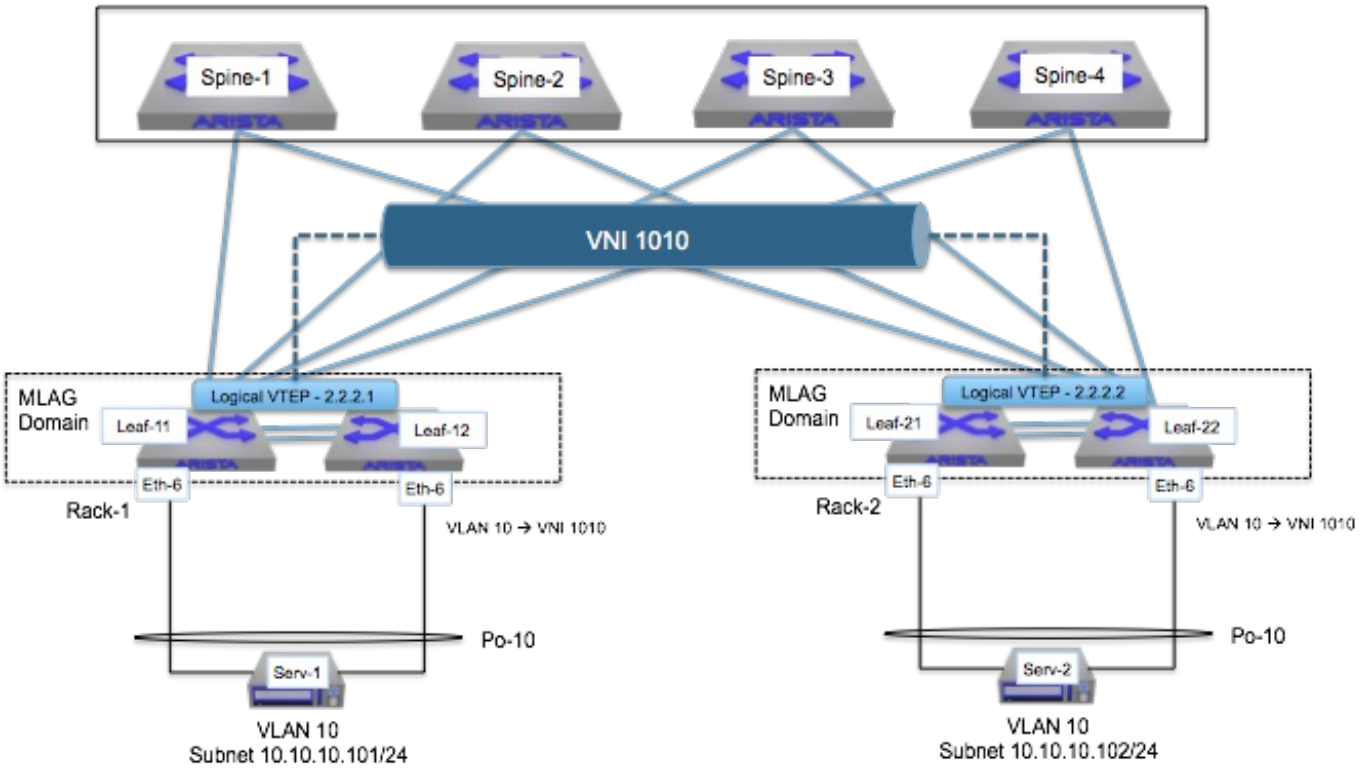
- Providing layer 2 connectivity between racks, or halls of the data centre without requiring an underlying layer 2 infrastructure,
- Linking geographically disperse data centres as a Data Centre Interconnect (DCI) technology.

Within these architectures, where the Arista switch is acting as the VXLAN virtual tunnel EndPoint (VTEP) into the overlay network, there is a requirement for the VTEP to provide a resilient L2 active-active topology similar to an MLAG deployment in a traditional VLAN based layer 2 network. This is achieved by combining Arista's MLAG technology with the hardware VTEP functionality of an Arista switch.

Figure 1, illustrates a typical layer 3 leaf/spine architecture deployed in a modern Data Centre to simplify scale while delivering consistent throughput and latency for east-to-west traffic. In this example eBGP is deployed between the leaf and spine switches for flexible control of the advertised routes but any standard dynamic routing protocol (OSPF, IS-IS etc) could be deployed. To provide traffic load-balancing across the four spine switches Equal Cost Multi-Pathing (ECMP) is configured. For illustration purposes the topology includes two racks of servers, with the servers within the rack dual-homed to a pair of leaf switches in an MLAG configuration.

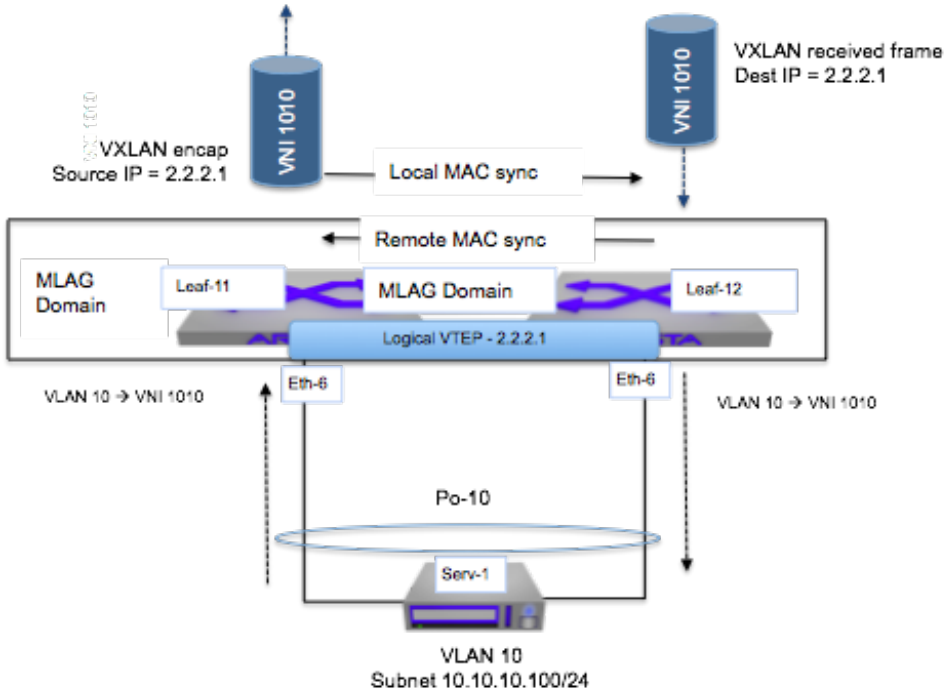


To provide the Layer 2 connectivity between the racks VXLAN is introduce as the overlay technology, this achieved by configuring a VXLAN VTEP on the leaf switches. Given the servers are dual-homed to a pair of leaf switches in an MLAG configuration, a single logical VTEP is created for each MLAG domain. This is achieved by configuring the VTEP on both MLAG peers with the same Virtual Tunnel Interface (VTI) ip address, this ensures both MLAG peers VXLAN encapsulate any locally received traffic with the same source IP address.



The logical VTEP in combination with MLAG provides an active-active VXLAN topology, native traffic received by either peer switch can be VXLAN encapsulated and any VXLAN encapsulated traffic (received via a spine switch)

can be locally de-encapsulated and forwarded to the end device. To enable this behaviour the local MAC table of a peer switch is synchronised across the peer link, along with any remote MAC's learnt via VXLAN , this information would include the remote hosts MAC address along with the associated remote VTEP IP address.



VXLAN with MLAG configuration

The following steps illustrate the configuration for creating the MLAG domain between the leaf switch pair leaf-11 and leaf-12 and it's logical VTEP (2.2.2.1), allowing active-active layer 2 connectivity between the servers in the two racks, across the IP fabric using VXLAN encapsulation. The configuration steps for creating the logical vtep for rack-2 (VTEP-2, 2.2.2.2) are identical and therefore for brevity are not detailed

Step 1: Create port-channel (port-1000) between the two leaf switches which will be used as the MLAG peer link.

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#interface eth 1-2 Leaf-11(config-if-Et1-2)#channel-group 1000 mode active Leaf-11(config-if-Et1-2)#interface port-channel 1000 Leaf-11(config-if-Po1000)#switchport mode trunk	Leaf-12(config)#interface eth 1-2 Leaf-12(config-if-Et1-2)#channel-group 1000 mode active Leaf-12(config-if-Et1-2)#interface port-channel 1000 Leaf-12(config-if-Po1000)#switchport mode trunk

Step 2: Create the peer VLAN (4094) and peer link IP addresses on both switches, the peer IP is used for heartbeat and MAC address synchronisation between the peers.

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#vlan 4094 Leaf-11(config)#int vlan 4094 Leaf-11(config-if-Vl4094)#ip address 172.168.10.1/30 Leaf-11(config-if-Vl4094)#int port-channel 1000 Leaf-11(config)#no spanning-tree vlan 4094	Leaf-12(config)#vlan 4094 Leaf-12(config)#int vlan 4094 Leaf-12(config-if-Vl4094)#ip address 172.168.10.2/30 Leaf-12(config-if-Vl4094)#int port-channel 1000 Leaf-12(config)#no spanning-tree vlan 4094

Step 3: Configure the MLAG domain on both peers, using the configured port-channel as the peer link and interface

VLAN 4094 as the peer address.

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#mlag Leaf-11(config-mlag)#Domain-id Rack-1 Leaf-11(config-mlag)#peer-link port-Channel 1000 Leaf-11(config-mlag)#local-interface vlan 4094 Leaf-11(config-mlag)#peer –address 172.168.10.2	Leaf-12(config-mlag)#mlag Leaf-12(config-mlag)#Domain-id Rack-1 Leaf-12(config-mlag)#peer-link port-Channel 1000 Leaf-12(config-mlag)#local-interface vlan 4094 Leaf-12(config-mlag)#peer –address 172.168.10.1

Step 4: Configure the MLAG port-channel (port-channel 10) on interface ethernet 6 of both peers, for connecting server-1 to the MLAG domain in a redundant manner of the rack and configure the port-channel as an access port in VLAN 10.

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#vlan 10 Leaf-11(config)#interface eth 6 Leaf-11(config-if-Et6)#channel-group 10 mode active Leaf-11(config-if-Po10)#int port-Channel 10 Leaf-11(config-if-Po10)#switchport mode trunk Leaf-11(config-if-Po10)#mlag 10 Leaf-11(config-if-Po10)#switchport mode access Leaf-11(config-if-Po10)#switchport access vlan 10	Leaf-12(config)#vlan 10 Leaf-12(config)#interface eth 6 Leaf-12(config-if-Et6)#channel-group 10 mode active Leaf-12(config-if-Po10)#int port-Channel 10 Leaf-12(config-if-Po10)#switchport mode trunk Leaf-12(config-if-Po10)#mlag 10 Leaf-12(config-if-Po10)#switchport mode access Leaf-12(config-if-Po10)#switchport access vlan 10

Step 5: With the MLAG domain created, the virtual VTEP can be configured across the two peer switches. Create the loopback interface to be used as the source-ip address for the VTEP, for the logical VTEP the same IP address will be defined on both MLAG peers. For the leaf-11 and leaf-12 switches the logical VTEP ip address will be 2.2.2.1/32:

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#int loopback 1 Leaf-11(config-if-Lo1)#ip address 2.2.2.1/32	Leaf-12(config)#int loopback 1 Leaf-12(config-if-Lo1)#ip address 2.2.2.1/32

Step 6: Assign the Loopback 1 interface to the Virtual Tunnel Interface (VTI) of the VTEP on both leaf switches

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#int vxlan 1 Leaf-11(config-if-Vx1)#vxlan source-interface loopback 1	Leaf-12(config)#int vxlan 1 Leaf-12(config-if-Vx1)#vxlan source-interface loopback 1

Step 7: Map VLAN 10 to the VNI 1010 on both peer switches and add the logical VTEP of leaf-21 and leaf-22 to the flood-list of VNI (i.e. 2.2.2.2/32). This will mean the MLAG domain created on the peers leaf-21 and leaf-22, will receive any broadcast, multicast or unknown unicast frames for the VNI, allowing the learning of MAC address and forwarding of broadcast between the two VTEPs. In the configuration of the logical VTEP on leaf-21 and leaf-22, the logical VTEP 2.2.2.1 would be added to the flood-list on both switches.

Switch Leaf-11	Switch Leaf-12
Leaf-11(config-if-Vx1)#vxlan vlan 10 vni 1010 Leaf-11(config-if-Vx1)#vxlan vlan 10 flood vtep 2.2.2.2	Leaf-12(config-if-Vx1)#vxlan vlan 10 vni 1010 Leaf-12(config-if-Vx1)#vxlan vlan 10 flood vtep 2.2.2.2

Step 8: To achieve IP connectivity between the logical VTEPs in the two racks, the logical VTEP's loopback address needs to be advertised to the spine switches, in this topology via eBGP. Configuration below is for the physical switches within rack-1, the logical VTEP loopback address for rack-2 also needs to be advertised into BGP on leaf-21 and leaf-22

Switch Leaf-11	Switch Leaf-12
Leaf-11(config)#router bgp 65001 Leaf-11(config-router-bgp)#network 2.2.2.1/32	Leaf-12(config)#router bgp 65001 Leaf-12(config-router-bgp)#network 2.2.2.1/32

With the logical VTEP's loopback address advertised into BGP, due to ECMP being enabled, the two leaf switches in rack-2 will learn a path via each of the four spines to the logical VTEP (2.2.2.1) in rack-1 and vice versa for the leafs switches in rack-1, which will each have 4 paths to the logical VTEP (2.2.2.2) in rack-2. Thus as traffic is VXLAN encapsulated and routed between the two logical VTEPs, it will be load-balanced on a per flow basis across the four spine switches.

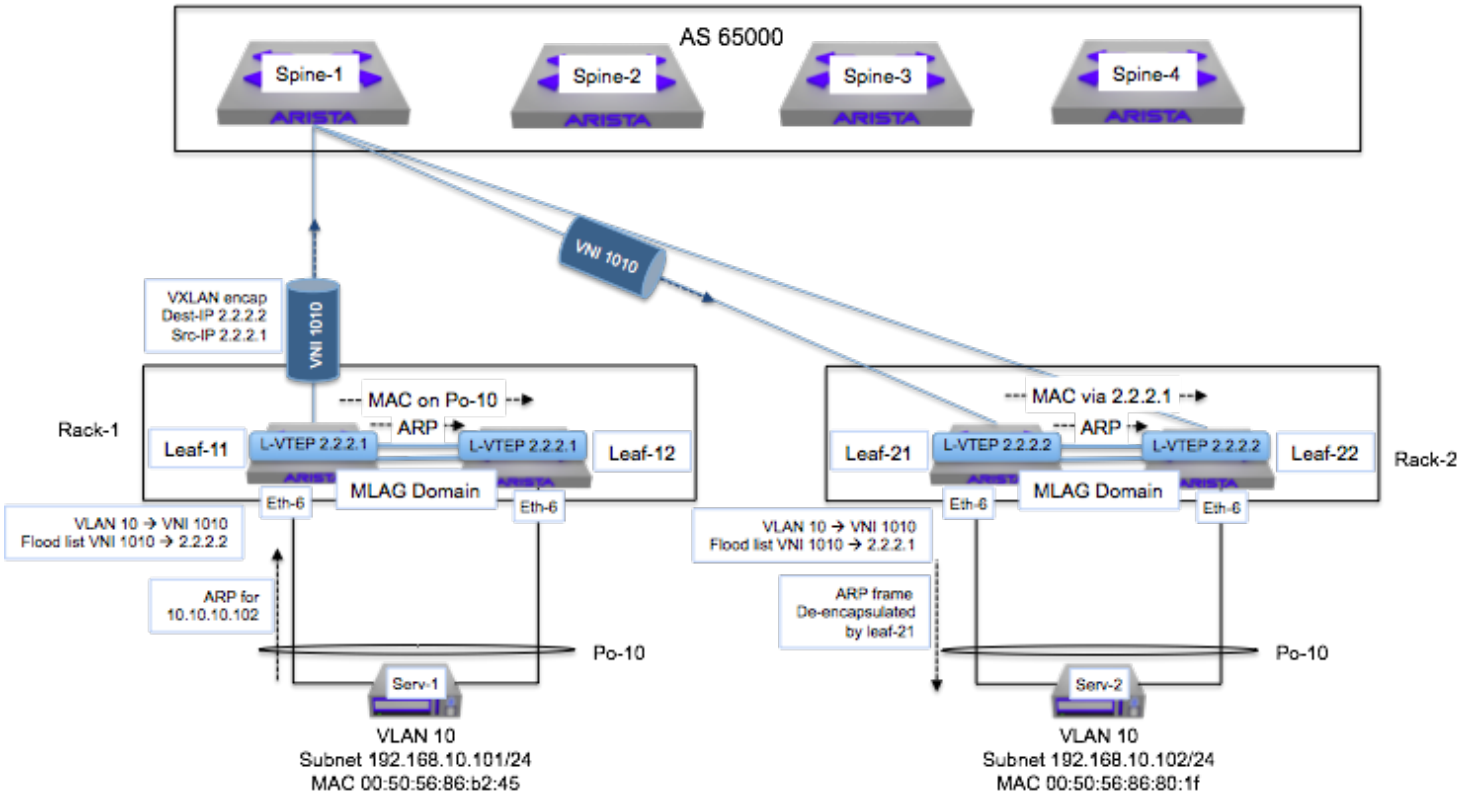
Following is the routing tables for each of the leaf switches in the topology:

Switch Leaf-11 routing table	Switch Leaf-12 routing table
Leaf-12(config)#show ip route Codes: C - connected, S - static, K - kernel, O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1, E2 - OSPF external type 2, N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP, R - RIP, I - ISIS, A B - BGP Aggregate, A O - OSPF Summary, NG - Nexthop Group Static Route Gateway of last resort is not set C 2.2.2.1/32 is directly connected, Loopback1 **B E 2.2.2.2/32 [20/0] via 172.168.2.1, Ethernet2 → Spine-1** **via 172.168.2.5, Ethernet3 → Spine-2** **via 172.168.2.9, Ethernet4 → Spine-3** **via 172.168.2.13, Ethernet5 → Spine-4** C 10.10.10.0/30 is directly connected, Vlan4094 C 10.10.10.4/30 is directly connected, Vlan4093 C 172.168.2.0/30 is directly connected, Ethernet2 C 172.168.2.4/30 is directly connected, Ethernet3 C 172.168.2.8/30 is directly connected, Ethernet4 C 172.168.2.12/30 is directly connected, Ethernet5 C 192.168.0.0/22 is directly connected, Management1	Leaf-12(config)#show ip route Codes: C - connected, S - static, K - kernel, O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1, E2 - OSPF external type 2, N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP, R - RIP, I - ISIS, A B - BGP Aggregate, A O - OSPF Summary, NG - Nexthop Group Static Route Gateway of last resort is not set C 2.2.2.1/32 is directly connected, Loopback1 **B E 2.2.2.2/32 [20/0] via 172.168.2.1, Ethernet2 → Spine-1** **via 172.168.2.5, Ethernet3 → Spine-2** **via 172.168.2.9, Ethernet4 → Spine-3** **via 172.168.2.13, Ethernet5 → Spine-4** C 10.10.10.0/30 is directly connected, Vlan4094 C 10.10.10.4/30 is directly connected, Vlan4093 C 172.168.2.0/30 is directly connected, Ethernet2 C 172.168.2.4/30 is directly connected, Ethernet3 C 172.168.2.8/30 is directly connected, Ethernet4 C 172.168.2.12/30 is directly connected, Ethernet5 C 192.168.0.0/22 is directly connected, Management1

Switch Leaf-21 routing table	Switch Leaf-22 routing table
Leaf-21(config-if-Vx1)#show ip route Codes: C - connected, S - static, K - kernel, O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1, E2 - OSPF external type 2, N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP, R - RIP, I - ISIS, A B - BGP Aggregate, A O - OSPF Summary, NG - Nexthop Group Static Route Gateway of last resort is not set C 2.2.2.2/32 is directly connected, Loopback1 **B E 2.2.2.1/32 [20/0] via 172.168.3.1, Ethernet2 → Spine-1** **via 172.168.3.5, Ethernet3 → Spine-2** **via 172.168.3.9, Ethernet4 → Spine-3** **via 172.168.3.13,Ethernet5 → Spine-4** C 10.10.10.0/30 is directly connected, Vlan4094 C 10.10.10.4/30 is directly connected, Vlan4093 C 172.168.3.0/30 is directly connected, Ethernet2 C 172.168.3.4/30 is directly connected, Ethernet3 C 172.168.3.8/30 is directly connected, Ethernet4 C 172.168.3.12/30 is directly connected, Ethernet5 C 192.168.0.0/22 is directly connected, Management1	Leaf-22(config-router-bgp)#show ip route Codes: C - connected, S - static, K - kernel, O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1, E2 - OSPF external type 2, N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type2, B I - iBGP, B E - eBGP, R - RIP, I - ISIS, A B - BGP Aggregate, A O - OSPF Summary, NG - Nexthop Group Static Route Gateway of last resort is not set C 2.2.2.2/32 is directly connected, Loopback1 **B E 2.2.2.1/32 [20/0] via 172.168.4.1, Ethernet2 → Spine-1** **via 172.168.4.5, Ethernet3 → Spine-2** **via 172.168.4.9, Ethernet4 → Spine-3** **via 172.168.4.13,Ethernet5 → Spine-4** C 10.10.10.0/30 is directly connected, Vlan4094 C 10.10.10.4/30 is directly connected, Vlan4093 C 172.168.4.0/30 is directly connected, Ethernet2 C 172.168.4.4/30 is directly connected, Ethernet3 C 172.168.4.8/30 is directly connected, Ethernet4 C 172.168.4.12/30 is directly connected, Ethernet5 C 192.168.0.0/22 is directly connected, Management1

Traffic Forwarding Behaviour

The active-active forwarding behaviour of the logical VTEP is illustrated in THE figure below. The example shows the traffic flow for Serv-1 in rack-1 communicating with Serv-2 in rack-2 via VXLAN. In the configuration, Serv-1 and Serv-2 are both members of VLAN 10 in each of the racks, VLAN 10 is mapped to VNI 1010 on the logical VTEPs providing layer 2 connectivity between the servers across the leaf spine fabric. The Servers are dual-homed to both leaf switches in their respective racks, using a port-channel (port-channel 10) which is split across the two physical switches of the rack, configured in an MLAG domain.

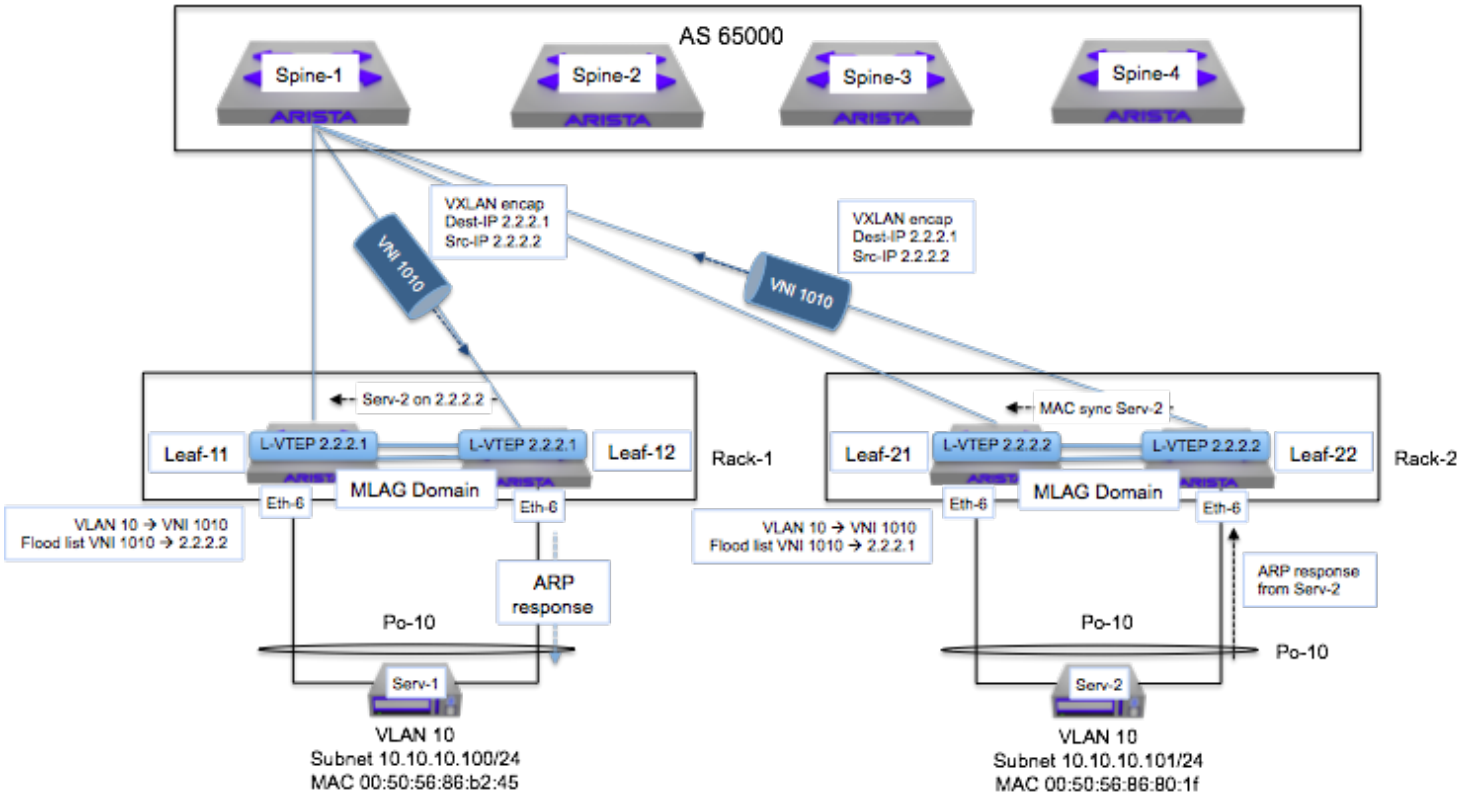


The following highlights the steps on how Serv-1 discovers Serv-2, where both servers reside on same L2 subnet 10.10.10.0/24, but physically located in different racks.

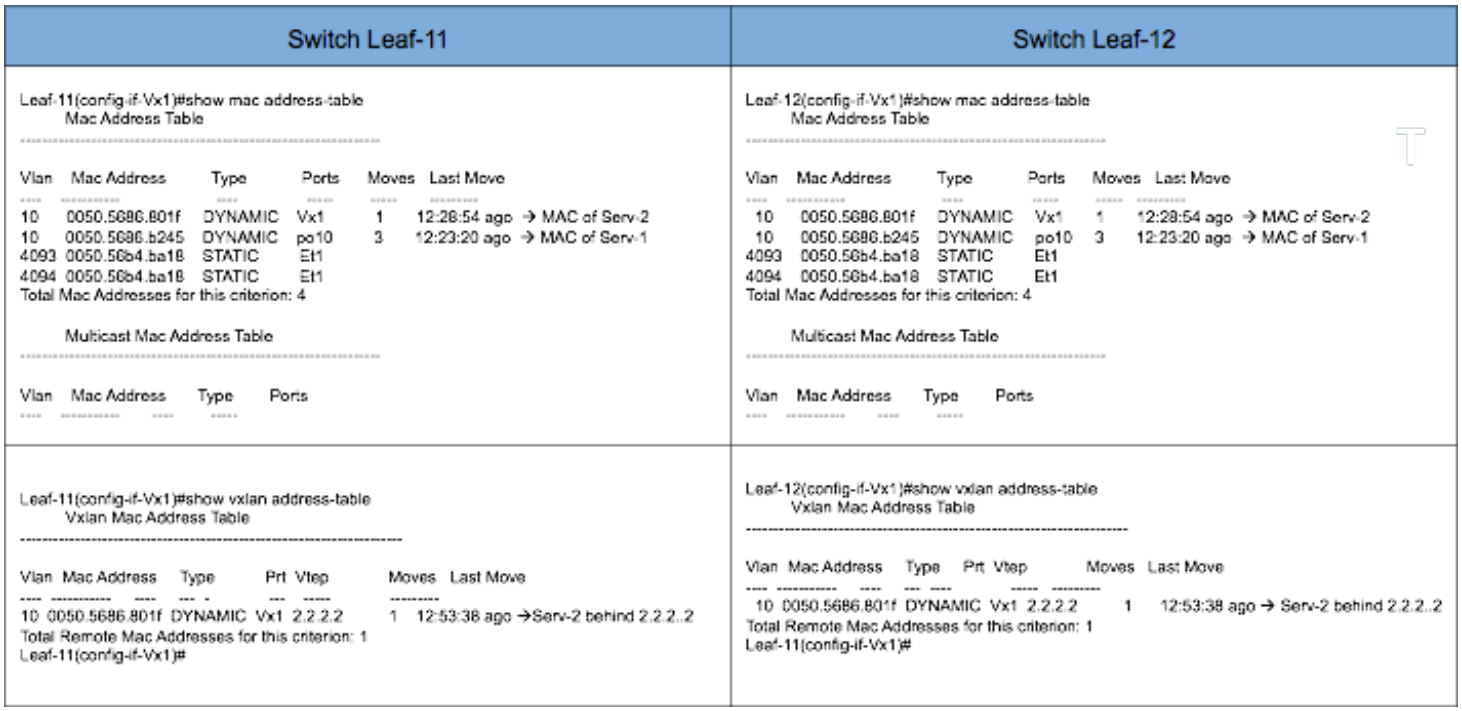
1. Due to load-balancing algorithm of Serv-1's port-channel, the initial ARP packet for Serv-2 is received by leaf-11, as a broadcast frame the ARP packet would be flooded to any local ports on leaf-11 which are a member of Vlan-10.
2. As standard MLAG behaviour the frame is also flooded across the MLAG peer link for reception by any single-homed devices on leaf-12. The MAC address of Serv-1 is synchronised with leaf-12 across the peer link, allowing leaf-12 to learn the MAC of Serv-1 on it's local port member of MLAG port-channel-10.
3. As VLAN-10 is mapped to VNI 1010, the ARP frame is also flooded to all VTEPs which are a member of the VNI (defined by the configured flood list of the VTEP), which in this example would be 2.2.2.2 the logical VTEP of rack-2. The ARP frame is VXLAN encapsulated by leaf-11 with a source IP address equal to logical VTEP (2.2.2.1) and a destination IP address of the logical VTEP on rack-2 (2.2.2.2).
4. Leaf-11 switch has four potential paths via each of the spine switches to the logical VTEP 2.2.2.2. For a specific flow ECMP hashing algorithm will pick one of the four paths. In this case lets presume the frame arrives at spine-1.
5. The spine-1 switch, subsequently routes the VXLAN frame to leaf-21. Note Spine-1 switch has two paths to 2.2.2.2 (leaf-21 and leaf-22), the path chosen is based on the result of the ECMP hashing algorithm for the flow on spine-1.
6. Receiving the frame, leaf-21 will remove the VXLAN header and in the process learn the MAC address of

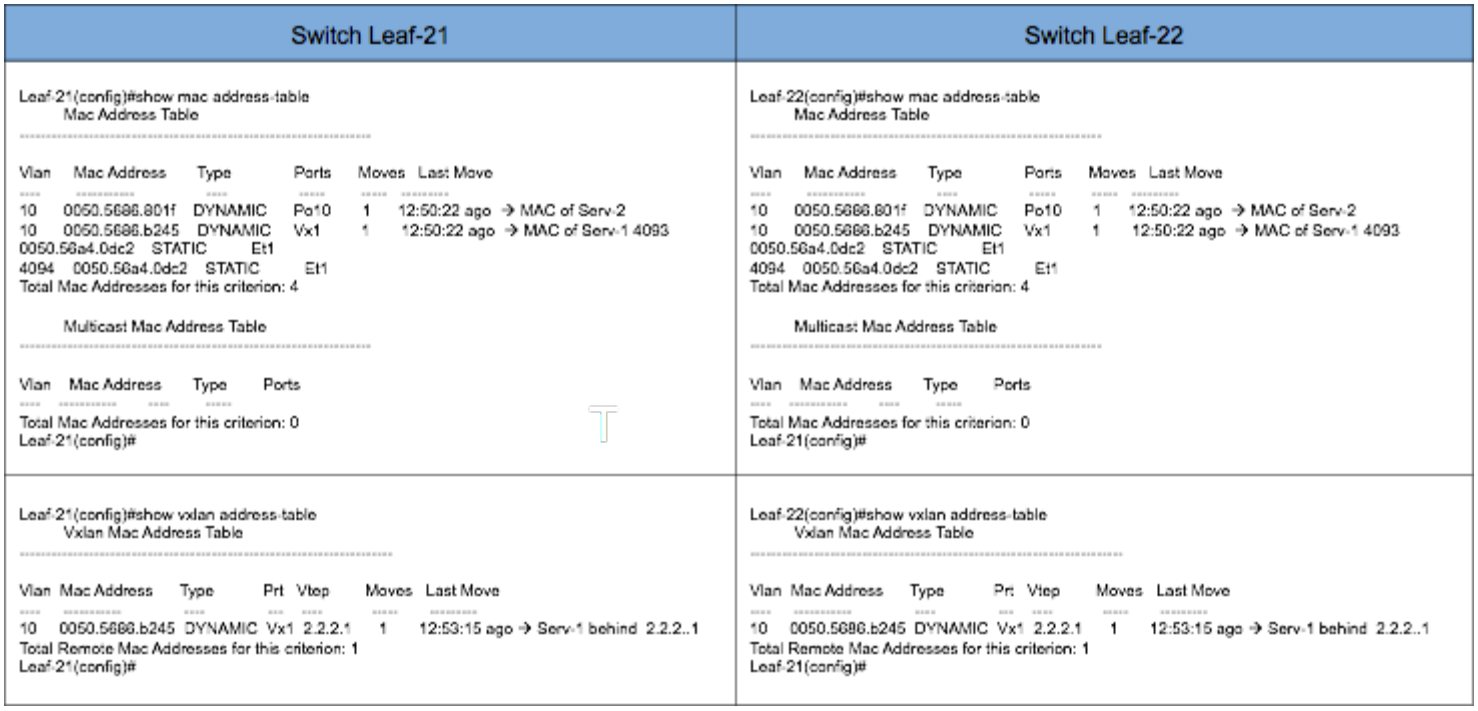
Serv-1 as residing behind the logical VTEP of rack-1 (i.e. 2.2.2.1).

7. This remote MAC entry will then be added to MAC table of leaf-21 and synchronised with leaf-22 over the MLAG peer link. The frame is then flooded to all member ports of VLAN 10 on leaf-21 and as standard MLAG behaviour flooded across the peer link for reception by any single-homed devices on leaf-22, this is on the assumption that the Serv-2 MAC has not yet been learnt



1. Receiving the ARP packet, Serv-2 responds to ARP request with unicast ARP response frame direct to the MAC of Serv-1.
2. Due to the initial ARP request, both leaf-21 and leaf-22 have learnt the MAC of Serv-1 as residing behind the logical VTEP of rack-1 (2.2.2.1), leaf-21 via the traffic flow and leaf-22 from the MLAG MAC sync process. Thus regardless of which link of the port-channel Serv-2 hashes the ARP response, either MLAG switch will have a MAC entry for Serv-1
3. The leaf-21 switch receiving the ARP response. will VXLAN encapsulate the frame with destination IP address of the logical VTEP of rack-1 (2.2.2.1). Due the ECMP hashing in this example, the frame is routed to the Spine-1 switch.
4. The spine-1 switch, subsequently routes the VXLAN frame to leaf-12. Note Spine-1 switch has two paths to 2.2.2.1 (leaf-11 and leaf-12), the path chosen is based on the result of the ECMP hashing algorithm for the flow on spine-1.
5. Receiving the VXLAN encapsulated frame, leaf-12 learns the MAC of Serv-2 behind the logical VTEP of rack-2 (2.2.2.2), this information is sync'd across the MLAG peer link to leaf-11.
6. The VXLAN frame is then de-encapsulated and forwarded down local link of port-channel-10 where the MAC of Serv-1 has been learnt.

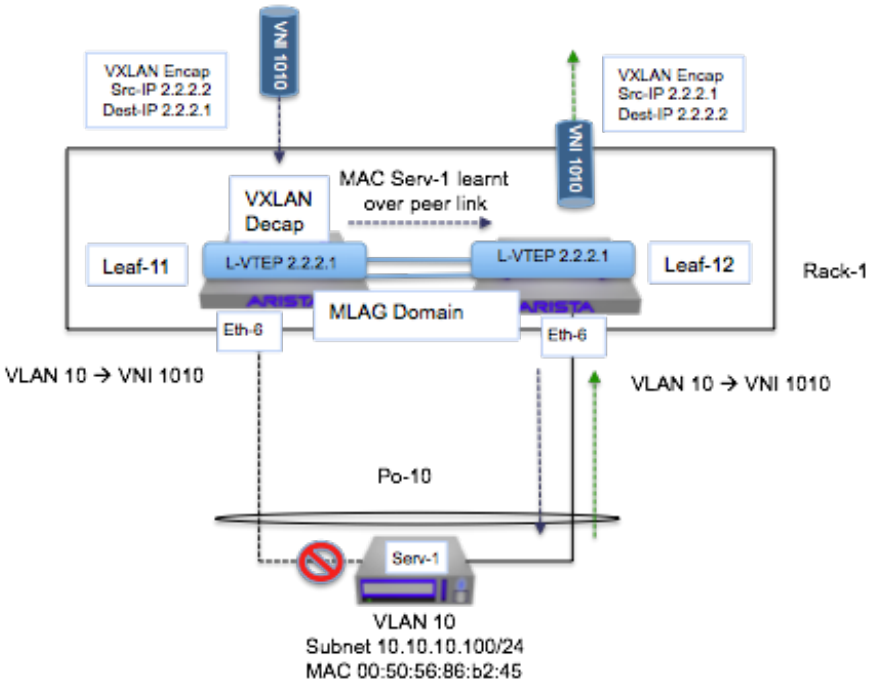




The MAC and VXLAN address table of all four leaf switches after the traffic flow.

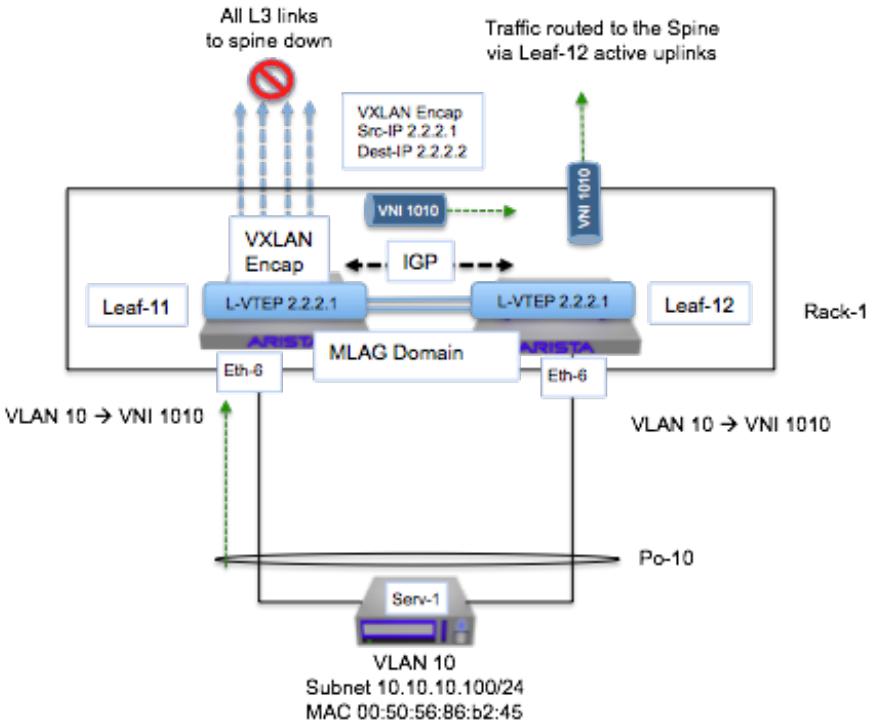
Traffic Failover Behaviour

Traffic forwarding during a failure scenario follows standard MLAG behaviour, if a link of the server's port-channel fails traffic will be forwarded across one of the remaining active links of the port-channel (ethernet 6 of leaf-12 in the example below). When leaf-12 receives the frame from Serv-1, with the destination being Serv-2, leaf-12 will VXLAN encapsulate the frame and route it over the IP fabric via it's own local routing. If the returning traffic is received on leaf-11 due to the ECMP hash of the spine, leaf-11 well de-encapsulate the frame and based on it's local MAC table switch the frame across the peer link for forwarding to Serv-1 via leaf-12's local ethernet 6 interface.



Traffic that is to be either VXLAN encapsulated or de-encapsulated, will always be done by the first VTEP switch receiving the frame, any subsequent processing will be based on the switches local MAC table (decapsulated and switch) or routing table (encapsulate and route).

In the unlikely event that a leaf switch loses all four of it's uplinks to the spine, it's recommended that a routing protocol is enabled across the peer link to ensure traffic can still be routed to spine layer. This concept is illustrated below, were switch leaf-11 has all four links to the spine inactive, traffic received on ethernet 6 of leaf-11 destined for the logical VTEP of rack-2 will be VXLAN encapsulated with a destination IP address of 2.2.2.2. As the local links to the spine are in-active, it will learn a route to 2.2.2.2 via the IGP running across the peer link, forcing the traffic to be routed across the MLAG peer link and forwarded across the spine based on the routing table of leaf-12. As best practice it is recommend the IGP protocol is configured on a dedicated VLAN of the peer link rather than use the peer link VLAN of the MLAG configuration.


All L3 links
to spine down
Traffic routed to the Spine
via Leaf-12 active uplinks
VXLAN Encap
Src-IP 2.2.2.1
Dest-IP 2.2.2.2
VNI 1010
VNI 1010
VXLAN
Encap
IGP
Leaf-11
L-VTEP 2.2.2.1
L-VTEP 2.2.2.1
Leaf-12
Rack-1
MLAG Domain
Eth-6
Eth-6
VLAN 10 → VNI 1010
VLAN 10 → VNI 1010
Po-10
Serv-1
VLAN 10
Subnet 10.10.10.100/24
MAC 00:50:56:86:b2:45