preview uncompleted article - work in progress preview

XENYA

Tehnical report

DataCenter and Enterprise redundant network design with high performance modern compact routing switches

peter.reinhardt@xenya.si

Sumary

Several advanced network architectures will be presented that can provide redundant and very high throughput communication suitable for modern data centers and enterprise networks. Redundancy is provided without using slow converging protocols like spanning-tree and can still implement larger L2 network sections that can simplify internal management within datacenter and are required in today's datacenters for simple implementation of live virtual machines migration. Each architecture will be presented together with detailed configuration examples to simplify practical implementation.

Two versions of Spine/Leaf architectures will be presented - a standard L3 based Spine/Leaf architecture that uses ECMP routing to provide data load balancing and redundancy with ability to expand to very large networks and mlag based Spine/Leaf architecture that is able to implement both L3 and L2 networks for small to midsized networks. Proposed datacenter network architectures support fully redundant implementation that includes redundant server connections. An additional, more cost effective, architecture for connecting large number of standard, non redundant clients, suitable for enterprise network implementation will be also presented.

Both main architectures are a "fabric" type design that provides equal delays and throughput between any port pairs in the network.

Spine Leaf architecture

Spine/leaf architecture is an optimization of classical two or three tier architecture that was built around two big core switches. In case of Spine / Leaf architecture functions of central backbone switches can be dispersed to many spine switches and this enable us to use low cost and power efficient compact switches built around standard switch SoC based hardware. Such dispersed network functions are more evenly distributed across all network, so HW or connection failures has smaller impact than a failure of a large central switch. With this design we get network with lower power consumption, better scalability, and lower cost.

Basic Spine / Leaf architecture is shown on Figure 1

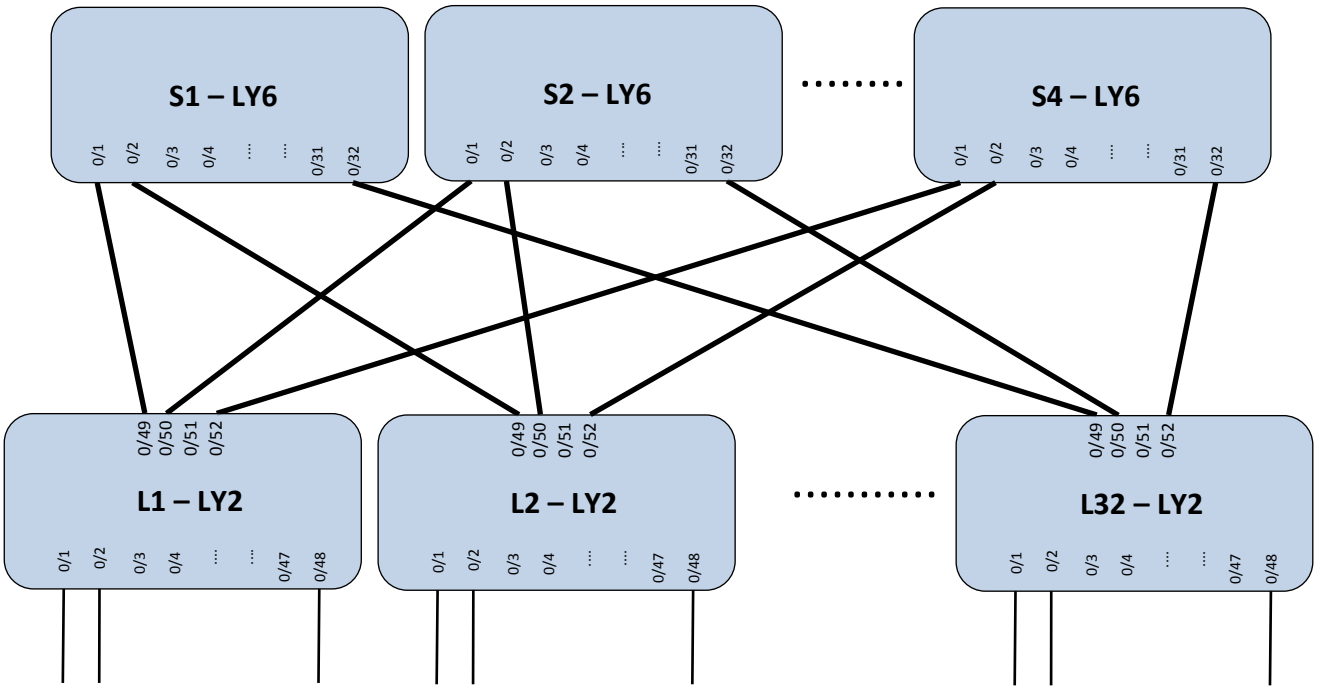


Figure 1 Example of basic Spine/Leaf architecture using 40Gb uplinks

Depending on desired target system scale, we have multiple options on choosing the interface configurations and protocols used. Most commonly Spine/Leaf architectures are based on standard ECMP (equal cost multiple path) routing as protocol of choice to provide the required redundancy with fast recovery and to provide data load balancing between all interconnections between switches. Standard routing protocols enable scaling to practically any size, so the end scale of Spine / Leaf architecture is determined by the number of ports in Spine switches and the number client ports in Leaf switches, while throughput depends on number of connected uplink ports on Leaf switches and optionally number of Spine switches.

In practice usually the number of ECMP routes is also limited and this also limits the number of connections to spine switches, since over each of these connection one of the equal cost paths are established, but this is typically significantly larger number than number of available uplink ports on leaf switches.

While routing protocols assure for redundancy of connections between switches, they are usually not supported on client (server) side connection. Server connections usually support only simple L2 protocols like link aggregation with NIC teaming or in best case LACP. This is the reason why we need additional protocols that provide redundant connections to south side. In old architectures L2 redundancy was provided by spanning tree protocols. These have many undesirable properties, like slow conversion, blocking of complete L2 domain due to topology change and transferring data on typically only half of the connections in highly redundant topologies. In modern networks we use instead, technologies that provide link aggregation and load sharing of links from multiple devices. These way clients can be connected with standard link aggregation protocols and all links can be used for data transfer.

Implementation description

We designed two specific implementations of Spine/Leaf architecture with Quanta/Xenya switches LY2R (or LY8) for Leaf switches and LY6 for spine switches.

Architecture A1

Architecture A1 is shown on **Figure 2**. This implementation uses 40 Gb connections between each Spine and Leaf switch. Shown implementation is an example of high bandwidth and moderate scalability. With different Spine/Leaf connections arrangement we can create larger systems with same throughput using more of same type switches.

Main Features

Provides L3 networking only. Each Leaf switch must have one or more local L3 subnets for clients that are defined in separate VLANs. Communication between clients on different client switches (actually different MLAG domains) can communicate only over L3 protocols.

Implements core of the network as standard Spine/Leaf architecture using L3 ECMP routing.

This design is highly scalable. Depending on way interconnections between Spine and Leaf switches are made It can scale up to 4 Spine switches and 32 Leaf switches resulting in 1536 10Gb client ports, when uplink ports are used as 40Gb ports. It can scale to 16 Spine switches and up to 128 Leaf switches with up to 6144 (or 3072 redundant) 10G ports - when interconnected with (same) uplink ports that are used as 16x10Gb ports. – all with same type of hardware and same throughput between client ports and same oversubscription ratio.

There is 1:3 oversubscription ratio of client ports aggregated bandwidth compared to uplink aggregated bandwidth - 160Gb aggregated uplink speed and 480Gb aggregated client port speed.

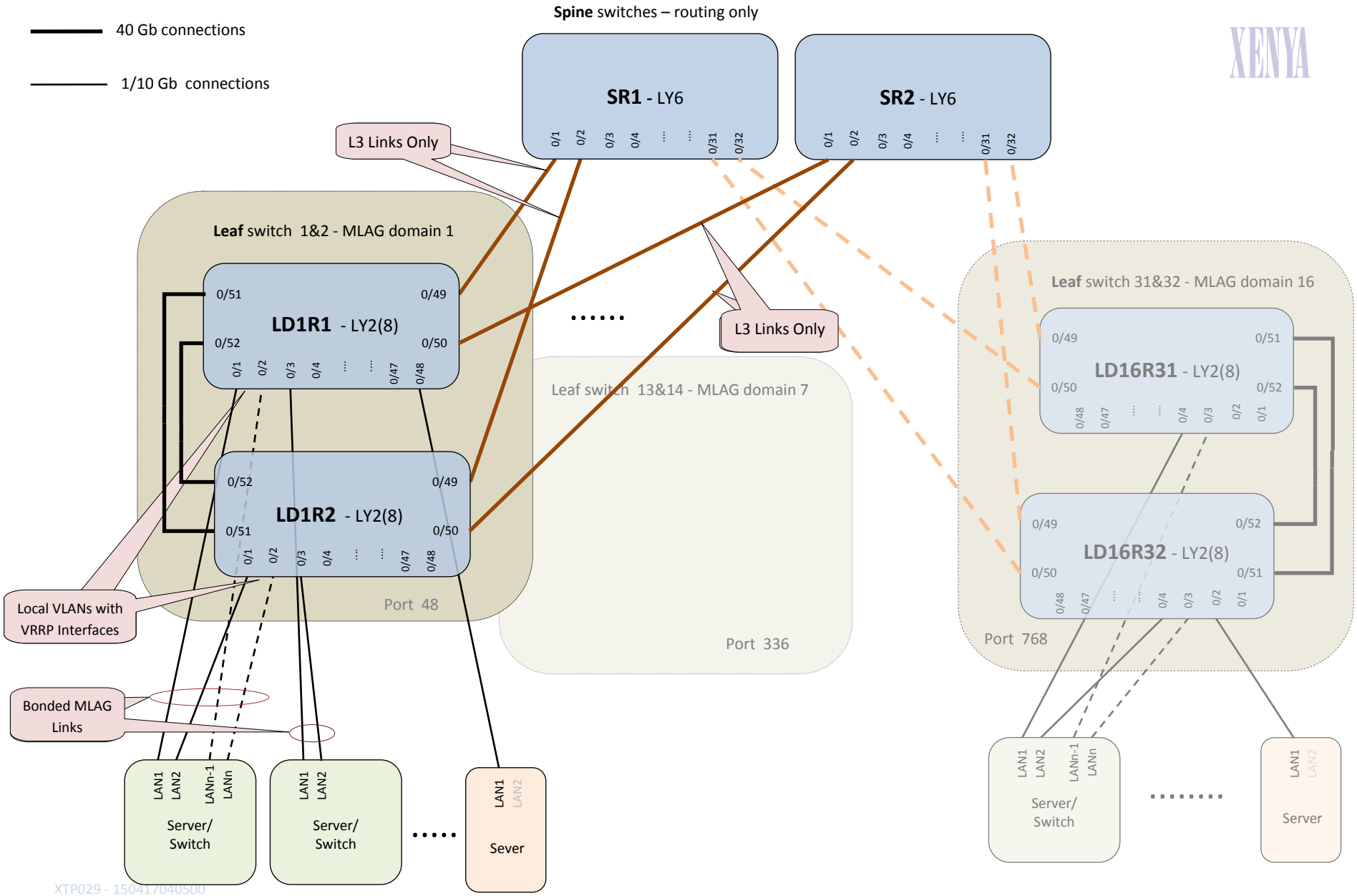
Pairs of two Leaf switches are bonded in MLAG domains and form up to 48 pairs of redundant L2 interfaces for clients that can be in one or more VLANs.

Standard client connection is a 10 Gb port that can work also as 1Gb interface. Clients can be connected with redundant connections 2x 10 Gb that are during normal operation both active. Higher bandwidth client connections are also possible with up to 16x 10 Gb ports in single aggregated connection.

On Leaf switches VLANs are defined, each with its own L3 subnet as client subnet. L3 Interfaces are defined with VRRP to provide dynamic resolution of virtual IP to local MAC addresses of both switches.

Client connections are implemented over two switches interconnected in a MLAG domain, where ports from two physical devices create single logical aggregated client connection.

preview uncompleted article - work in progress preview



XTP029 - 150417040500

Figure 2 Architecture A1 - Example of Spine/Leaf architecture with L3 ECMP and MLAG for redundant L2 south bound

preview uncompleted article - work in progress preview

A1 Configuration

In following passage, for clarity, only parts of switch configuration defining specific functions required for presented architecture are described. Complete detailed configuration for each switch is available in separate zip file.

Configuration of Spine Switches

Main task of spine switches in this architecture is ECMP routing on all connections to Leaf switches. On each of these connections there is dedicated /30 ip subnet defined for routing purposes only. OSPF is running on all interfaces to Leaf switches and aggregates routes from Leaf switches. These are routes to L3 interfaces in VLANs on Leaf switches. Using ECMP routing it provides traffic load balancing on multiple connections to each Leaf switch and also provides fast recovery in case of link or port failure.

In our example implementation first **switch SR1** provides all routing subnets in 192.168.255.0/25 subnet staring with first ip at 192.168.255.1 as gateway on interface 0/1 each subsequent interface implements next /30 subnet. Second Spine **switch SR2** implements all its routes in subnet 192.168.255.128/25 with gw 192.168.255.129 on port 0/1. Slightly truncated configurations are shown below:

Port configuration for **Spine Router 1 Q7-SR1**:

```
hostname "Q7-SR1-LY6"
!
!disable spanning tree
no spanning-tree bpdu-forwarding
no spanning-tree
!
! enable routing and vrrp
ip routing
ip vrrp
!
! configure L3 interfaces on each port directly
interface 0/1
   routing
   ip address 192.168.255.1 255.255.255.252
 exit
interface 0/2
   routing
   ip address 192.168.255.5 255.255.255.252
 exit
interface 0/3
   routing
!
Similar configuration  for  /30 subnet is applied for other 30 ports
!
!

! Configure ospf routing
 router ospf
  router-id 0.0.0.7
  network 192.168.255.0 0.0.0.127 area 0
  exit
```

```
end

Port configuration for Spine Router 2  Q8-SR2:
configure
hostname "Q8-SR2-LY6"
!
!disable spanning tree
no spanning-tree bpdu-forwarding
no spanning-tree
!
! enable routing and vrrp
ip routing
ip vrrp
!
! configure L3 interfaces on first 7 ports
interface 0/1
   routing
   ip address 192.168.255.129 255.255.255.252
 exit
interface 0/2
   routing
   ip address 192.168.255.133 255.255.255.252
 exit
interface 0/3
   routing
   ip address 192.168.255.137 255.255.255.252
 exit
! configure ospf routing
 router ospf
  router-id 0.0.0.8
  network 192.168.255.128 0.0.0.127 area 0
  exit
end
```

Configuration of Leaf Switches

Leaf Switches are implemented with a pair of switches connected in mlag domain. In upstream direction each of these switches acts as independent router routing over 2 independent 40 Gb connections to Spine switches. Routing is done from Local L3 interfaces defined in VLANs that spans both switches within MLAG domain. Since we have two switches/routers routing the same subnet we have defined a virtual gateway using vrrp protocol that implements a virtual ip that acts as gateway in Client subnet.

In (south) downstream direction a pair of switches creates a MLAG domain that acts as single L2 switch with aggregated connections to each server. This connection uses at least one port from each switch aggregated in single port-channel interface that supports LACP protocol and enables transmission with full bandwidth of both ports (2 or 20Gb) to each sewer/switch when all equipment is working, and with bandwidth of single port (1 or 10Gb) in case of malfunction of single switch, single connection or single server LAN card. Since we are interfacing mlag aggregated connection to routed network, we also enabled "`mlag peer-gateway`" function that enables both switches in mlag to route packets received with MAC address of either of these routers. This way each packet, regardless on which physical interface it is received, is routed immediately and is not passed to neighboring router for routing. This enables consistent routing between VLANs within mlag domain

and passes minimum amount of traffic over the peer -link connection that directly connects the two switches in mlag domain.

Following is the main part of the configuration of one of mlag switches explained step-by step.

First we have an initialization part that defines default modes of operation and default parameters for all ports. Most of these settings depend on actual application (i.e. jumbo packets with mtu 12288, may not be needed), some are optional (i.e like sending /receiving lldp packets, this just helps discover some interconnection faults during setup…), on uplinks we do not need spanning tree protocol, on downlinks it is optional, but is disabled in our example (since we use mlag for redundancy) , but we just detect possible customer created loops with stp.

```
configure
hostname "Q5-LR1-LY2"
!
!-----------------------------------------------
! configure default values for all ports
!-----------------------------------------------
! Define default parameters for 1/10Gb ports
! values below are just example
interface range 0/1 - 0/48
 no spanning-tree port mode
 mtu 12288
 lldp transmit
 lldp receive
 lldp transmit-tlv port-desc sys-name sys-desc
exit
!
! Define port mode parameters for 40Gb ports (required just once, then reload)
interface 0/49
 port-mode 1x40g
exit
interface 0/50
 port-mode 1x40g
exit
interface 0/51
 port-mode 1x40g
exit
interface 0/52
 port-mode 1x40g
exit
!
! Define default parameters for 40Gb ports
interface range 0/49 - 0/52
 no spanning-tree port mode
 ! udld – unidirectional detection is especially recommended for links that
 !   use MM QSFP optical connections with MTU connecting patch cables
 udld port
 mtu 12288
 lldp transmit
 lldp receive
 lldp transmit-tlv port-desc sys-name sys-desc
exit
end
```

Routing over uplinks is defined the same way as on Spine switches, only that second IP in a /30 subnet is used; while Spine switches use the first IP.

```
config
! enable routing and vrrp
ip routing
ip vrrp
!
! configure L3 interfaces on ports connected to Spine switches
interface 0/49
   description “Connected to SR1 port 0/1”
   routing
   no spanning-tree port
   ip address 192.168.255.2 255.255.255.252
 exit
interface 0/50
   description “Connected to SR2 port 0/1”
   routing
   no spanning-tree port
   ip address 192.168.255.130 255.255.255.252
 exit
end
```

OSPF configuration will be shown in following section describing L3 customer interfaces configuration, since specific parts of OSPF configuration is related to these L3 interfaces that are defined in VLANs spanning only mlag domain (the two switches that create mlag domain). For each customer subnet a VLAN must be created with ip addresses within client subnet that are defined on both switches. Within a mlag domain we have a common switch and basically two routers with two (ip and mac) addresses. To provide a single gateway for L3 clients we need to configure a virtual IP address using vrrp protocol that will provide redundancy in case of failure of one of switches in a mlag domain. In following example 3 VLANs are defined for 3 separate customer subnets. Only the L3 part of definition is shown. Assigning ports to these VLANs will be shown in section describing mlag definition. Any unique ip subnet can be used for customer subnet, but if there is a large number of switches/mlag domains with same VLAN id (these can be same since they are local to each mlag domain) then some standardized addressing scheme is recommended to enable easier debugging of potential miss configuration and to be able to define access rules that filter traffic based on location (i.e. mlag domain), VLAN and subnet (i.e. application). In demo configuration we created addresses for customer subnets in following way 10.<mlag_domain_id>.100+<VLAN_id>.n where n is 1 for virtual ip , 2 for first router ip and 3 for second router ip – in this case for example in mlag domain1 (switches LSR1 & LSR2) we have in VLAN 2 virtual address 10.1.102.1 and ip 10.1.102.2 on LSR1 and ip 10.1.102.3 on LSR2.

Following are four steps in client VLAN/subnet configuration on Leaf switches:

1. Creating VLANs
2. Optionally naming VLANs
3. Defining VLAN ip configuration (interface ip and virtual ip)
4. Configuring OSPF routing

```
config
! First we need to create and optionally name the customers VLANs
!-------------------------------------------------------
! create VLANs for customer L3 client interfaces
interface vlan 2
exit
interface vlan 3
exit
interface vlan 4
exit
! optionally name client VLANs
vlan database
  vlan name 2 "vLD1R2_C1"
  vlan name 3 "vLD1R2_C2"
  vlan name 4 "vLD1R2_C3"
exit
! define ip interfaces for customers VLANs
interface vlan 2
  description 'Demo Routed customer VLAN 2'
  ip address 10.1.2.3 255.255.255.0
  ip vrrp 2
  ip vrrp 2 mode
  ip vrrp 2 ip 10.1.102.1
  ip vrrp 2 priority 150
  ; accept-mode - virtual ip should respond to ping
  ip vrrp 2 accept-mode
exit
!
interface vlan 3
  description 'Demo Routed customer VLAN 3'
  ip address 10.1.3.3 255.255.255.0
  ip vrrp 3
  ip vrrp 3 mode
  ip vrrp 3 ip 10.1.103.1
  ip vrrp 3 priority 250
  ip vrrp 3 accept-mode
exit
!

interface vlan 4
  description 'Demo Routed customer VLAN 4'
  ip address 10.1.4.1 255.255.255.0
  ip vrrp 4
  ip vrrp 4 mode
  ip vrrp 4 ip 10.1.104.1
  ip vrrp 4 priority 150
  ip vrrp 4 accept-mode
exit
! configuring OSPF is mlag domain specific
! start ospf - see end of config file
router ospf
  router-id 0.1.0.1
  ! this subnet is used for routing to Spine switches
  network 192.168.255.0 0.0.0.255 area 0.0.0.0
  ! enable ospf
  enable
  !
```

```
  ! here comes all the subnets of customer interfaces (specific to each mlag
domain)
  network 10.1.2.0 0.0.0.255 area 0.0.0.0
  network 10.1.3.0 0.0.0.255 area 0.0.0.0
  network 10.1.4.0 0.0.0.255 area 0.0.0.0
  ! do not run ospf on customer facing  VLAN interfaces
  ! still it has to be known to ospf in order to be reachable from other
routers
  passive-interface vlan 2
  passive-interface vlan 3
  passive-interface vlan 4
exit
end
```

In final section we will describe the configuration needed to implement mlag domain. Mlag communicates with external world through aggregated connections, so we need following steps to define mlag domain:

1. Create port-channels (aggregated connection): one for peer-link - a redundant link connecting the two switches in a mlag domain, and others for client aggregated connections. Creation it is done in the same way as in case of standard port-channels.
2. Assigning ports to port-channels – in most cases just one port is assigned on each switch to mlag port channel, but up to 8 ports can be assigned to single port-channel it is done the same as in case of standard port-channels.
3. Creating mlag, assigning mlag domain id. It must have same value in configurations of both switches within domain and must be different than in all other domains. Since we are using these aggregated line for L3 traffic we have to configure "`mlag peer-gateway`" that instructs both routers to accept packets with destination MAC address of any of the two routers. This eliminates unnecessary packet transfer over peer link when LAG hash load balancing algorithm assigns a link to other router as the current master in vrrp session, and it also enables correct routing between the VLANs within mlag domain.
4. Assigning port-channels to a mlag domain. Same mlag port-channel on both switches in a mlag domain must have same mlag id value.

```
! Create port-channels
!---------------------------------------------
! port channels for customer interfaces
interface port-channel 1
exit
interface port-channel 2
exit
interface port-channel 3
exit
interface port-channel 4
exit
interface port-channel 64
exit
!
!--------------------------------
! Assigning ports to port-channels
!--------------------------------
! assign ports for peer-gateway port channel
```

preview uncompleted article - work in progress preview

```
interface range 0/51 - 0/52
 channel-group 64 mode active
exit
!---------------------------------
! assign ports for customer port-channels
! example of 4 port aggregated port-channel - 2 on each switch
interface range 0/1 - 0/2
 channel-group 1 mode active
exit
!
interface 0/3
 channel-group 3 mode active
exit
!
interface 0/4
 channel-group 4 mode active
exit
!---------------------------------
! Create mlag domain, set domain parameters
mlag
 mlag domain 1
 mlag peer-gateway
 mlag keepalive-timeout 3
! delays mlag port activation until mlag is active
 mlag member-linkdown
 ! alternate redundant path for mlag control packets - i.e. through management
interface
 ! mlag peer-keepalive destination 10.1.11.116
 ! mlag peer-keepalive source 10.1.11.115
 ! set shortest keep-alive timer to provide fast conversion
 mlag peer-keepalive timeout 3
!
! assign default parameters to all port-channels
interface range port-channel 2 - 4
 !no spanning-tree port mode
 spanning-tree guard loop
 spanning-tree tcnguard
 mtu 12288
exit
!
! assign mlag IDs and other specific parameters to port channels
interface port-channel 64
 mlag peer-link
exit
!
interface port-channel 1
 mlag 1
 spanning-tree guard none
 switchport allowed vlan add untagged 2
 switchport native vlan 2
 switchport allowed vlan remove 1
exit
!
interface port-channel 3
 mlag 3
 switchport allowed vlan add untagged 3
 switchport native vlan 3
 switchport allowed vlan remove 1
```

```
exit
!
interface port-channel 4
 mlag 4
 switchport allowed vlan add untagged 4
 switchport native vlan 4
 switchport allowed vlan remove 1
exit
!
```

This completes the configuration of first switch in first mlag domain.

Configuration of second switch in a mlag domain is basically done with the same configuration file with following parameters modified:

1. Switch name and snmp name id
2. Ip addresses of uplink interfaces
3. Ip addresses of customer VLANs must be set to different ip within same subnet, while virtual ip addresses are the same
4. Ip address of switch management if it is not set by dhcp

Configuration for switches in other mlag domain must change all parameters above and also following parameters:

5. Mlag domain id have to be changed
6. Ip addresses of customer VLANs AND virtual ip addresses must be changed (set to IPs in different ip subnet)
7. Port-channels to port mappings may change
8. Names of VLANs ought to be changed

These changes will not be described here in more detail, since they are either self explanatory, or have been described above during description of configuration of the first Leaf switch.

Configuration files for these switches are available in a *.zip* file, so if needed one can study the differences directly in configuration files.

There are some optional configurations that might be present in configurations, but are not within the scope of this article: These are domain settings to resolve named addresses, snmp settings for common remote control & monitoring, sntp settings to sync time settings, syslog settings to monitor events, authorization settings, spanning tree settings ... or even more complex settings like datacenter bridging, that are used to create a unified data/storage network….

Some of these configuration settings are described in other technical notes.

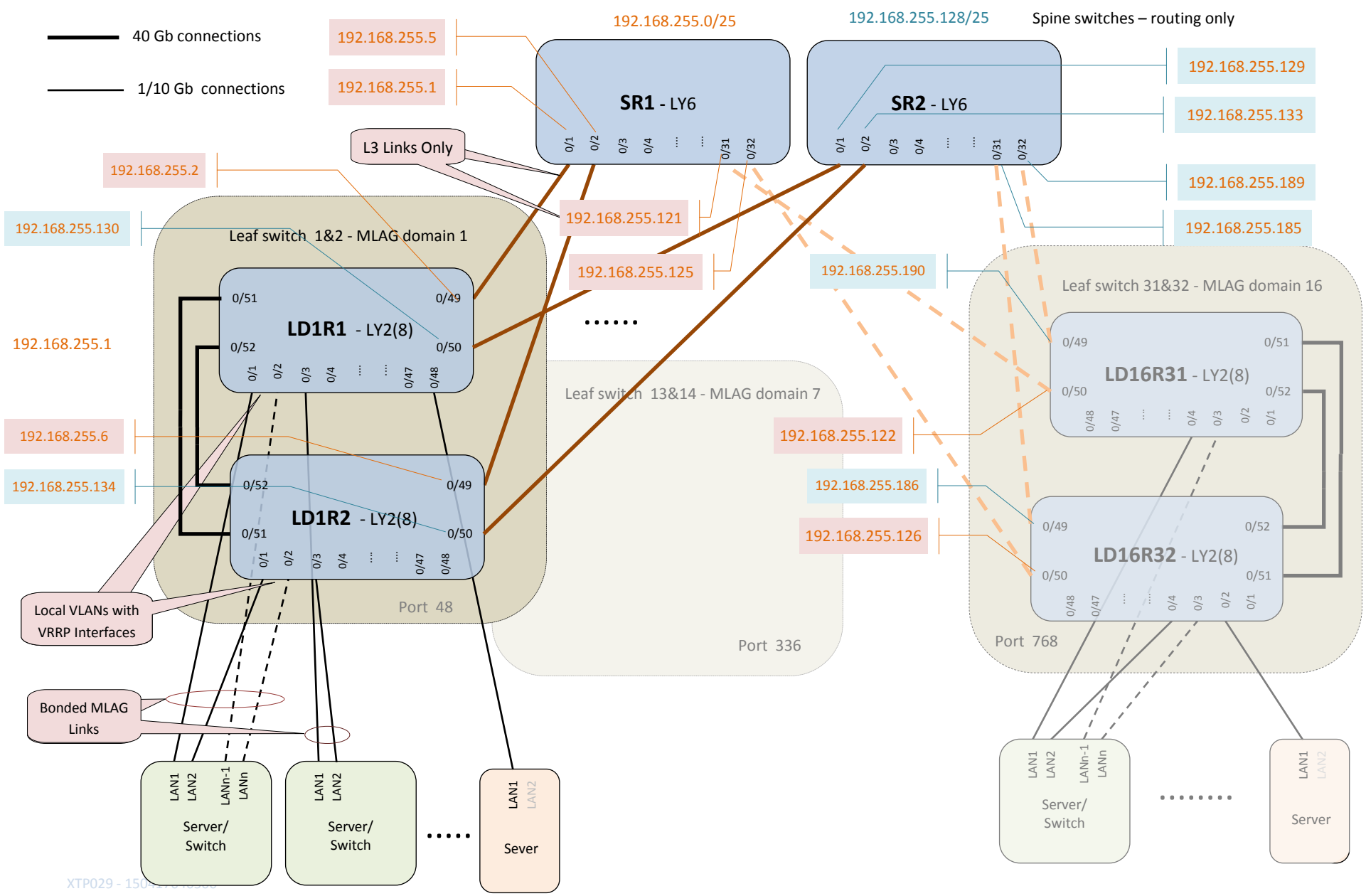


Figure 3 Architecture 1 - Example of Spine/Leaf architecture with L3 ECMP and MLAG for redundant L2 south bound connection - with uplink ip addresses defined

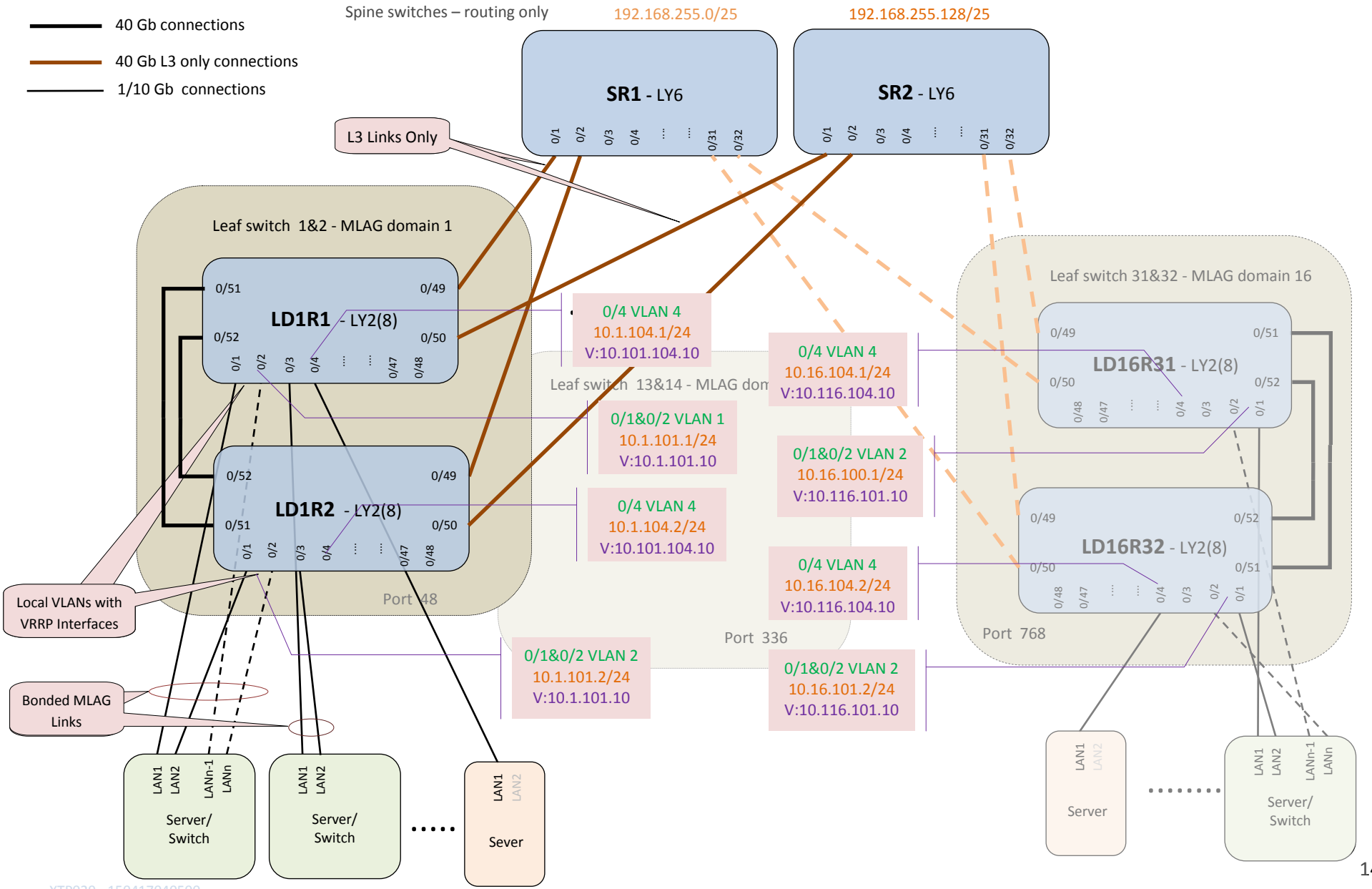


Figure 4 Architecture 1 - Example of Spine/Leaf architecture with L3 ECMP and MLAG for redundant L2 south bound connection with client ip, vlan and port info added for some interfaces

Architecture A2

Architecture A2 enables creation of relatively large redundant L2 networks without using protocols like spanning tree that can cause multiple problems, like slow conversion, blocking of traffic in complete L2 domain during conversion as a result of any topology change and inefficient use of redundant connections, where up to half of connections may be unused during normal network operation. Additionally, proposed design provide symmetric throughput and fairly balanced delays between different parts of the network, which simplifies optimization of applications that are spread over multiple servers or can move between servers.

Architecture A2 is shown on Figure **5**. Basic difference compared to Architecture A1 is that in Architecture A2 the Spine switches are also connected in a mlag configuration and are able to bridge (and route) the traffic between Leaf mlag domains.

With this architecture we can implement L2 and L3 connectivity for small to midsize networks with up to (15*48*2) 1440 10Gb ports for clients, when using LY2R switches (48x1/10Gb and 4x40Gb) or LY8 switches (48x1/10Gb and 6x40Gb) for Leaf switches and LY6 switches (32x40Gb) for Spine switches and with 1:3 oversubscription where all available 40Gb uplinks on LY2R are used.

Using mlag in Spine switches means that we cannot use more than two switches for Spine. This limits the scaling of this network if we want to keep the design in two tier Spine/Leaf design.

Leaf switches provide only L2 connection to Spine. As in previous design, Leaf switches are organized in pairs, forming mlag domains, but in this architecture mlag domains are connected with aggregated L2 connection (4x4Gb) to Spine mlag switches.

If L3 connectivity is needed it is implemented with routing within spine switches. All client L3 interfaces are defined as IP subnets within VLANs defined on Spine switches. As in previous examples, where we combined mlag and routing, we have within Spine mlag domain logically a switch and two routers, so we have to define two IP configurations on both routers within L3 client subnet and with virtual gateway using vrrp.

Routing is done only on Spine switches. All routing is local, but since we have multiple routers in network (at least two) dynamic routing is used using OSPF. Optimized use of internal link of and correct routing within mlag domain requires that "`mlag peer-gateway`" command is defined in Spine mlag domain.

Any access rules that are required are applied to customer VLAN interfaces primarily on Spine switches.

Standard VLANs (802.1q and QinQ) span across complete network – all mlag domains. All VLAN IDs must be unique for each VLAN used in the system. In this design one cannot reuse same VLAN IDs in different mlag domains not even for L3 configured VLANs.

Although there may be many switches in such a network (up to 32 when LY2 & LY6 switches are used with 1:3 oversubscription), adding a VLAN manually is not as difficult as it may seem at first glance. One must add a new VLAN on one switch in each mlag domain where the customer interfaces are present and on one of the Spine switches (so typically only on 3 switches and max on 8 switches

config must be changed) and on all switches the added VLAN configuration is identical, only on spine switches optional L3 interface must be added. Of course there are other mechanisms for dynamic VLAN manipulations, like GVRP and VTP which can be used to distribute VLAN information, and optionally L3 overlay protocols that tunnel L2 connections like VXLAN/NVGRE (LY6 and LY8 switches) can also be used …

Combining architecture A1 & A2

Both architectures shown so far can be easily combined in single network as shown on Figure 6. In such combined network we can have the benefits of both designs – relatively big redundant L2 sections and standard L3 interface to rest of the network.

In both network architectures mlag domain was used as a tool to assure L2 redundant connections to client equipment and in some cases within network. While this is excellent solution with many advantages, the protocols to implement mlag have so far not been standardized, so all mlag (or mlag like) solutions are proprietary.

In Architecture A1 – a standard spine/Leaf design - internal connections between Leaf and Spine switches are standard routed L3 links. Since these connections use only standard protocols, this is the place where networks from different vendors can be connected. This enables to reuse existing equipment and/or to combine equipment that is most suitable to implement different parts of the network. If we need larger L2 domains with such a network we can combine the two architectures.

In combined network design we have same scaling as in the A2 design if we want to preserve same delays, throughput and oversubscription ratio in both parts of the network. With lower bandwidth connection between the two sections and/or with longer delays (more hops) between the sections, we can modify the combined architecture so that it can scale the L3 only section to significantly larger number of nodes.

preview uncompleted article - work in progress preview

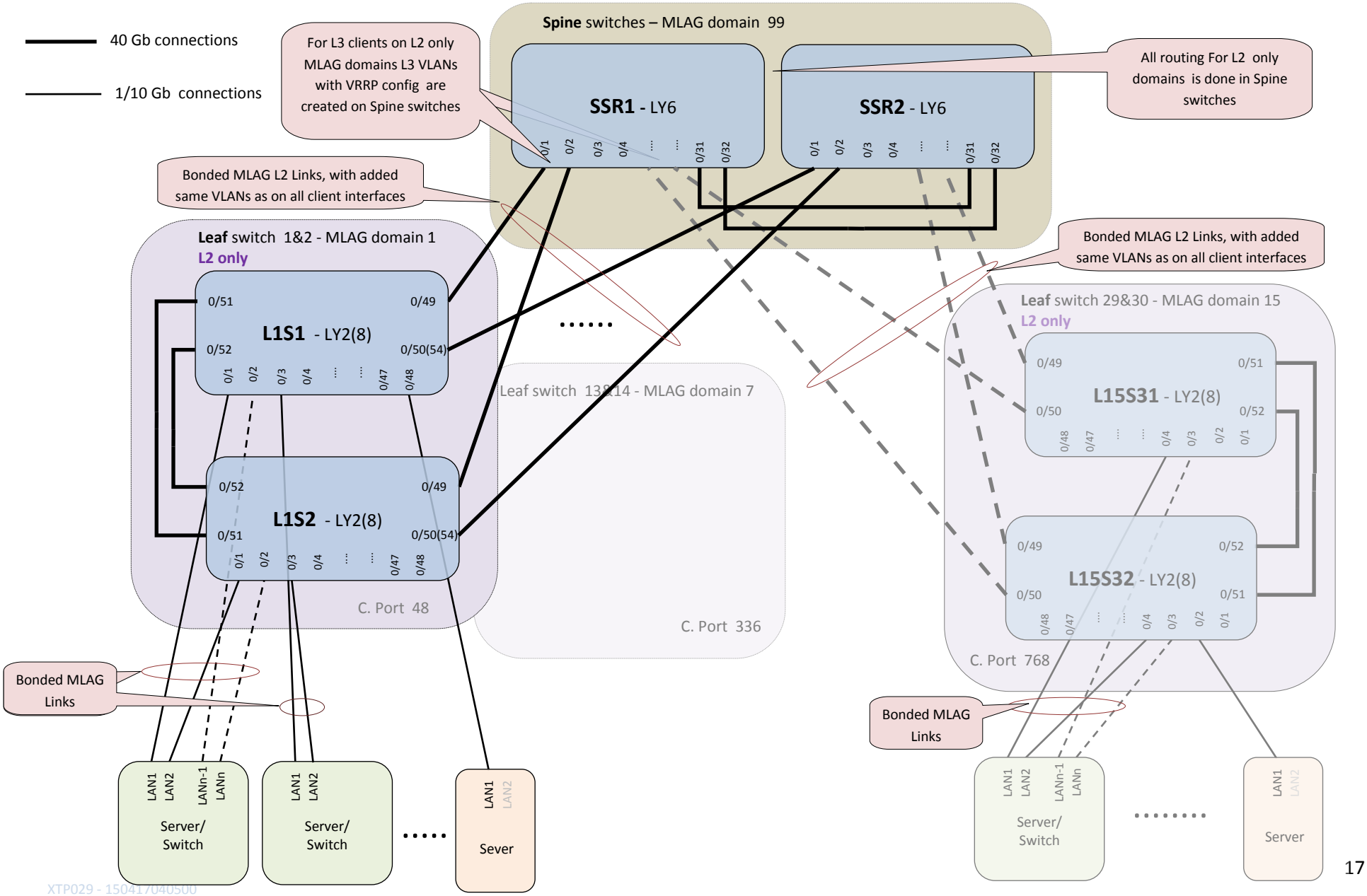


Figure 5 Architecture A2 - Example of Spine/Leaf architecture that can provide L2 and L3 connectivity between client interfaces

Configuration of architecture A2 and combined architecture A1&A2

Configuration in this architecture differs since different switches perform different tasks as in A1 design. Main differences are that bridging is done on Spine and Leaf tier and that in A2 design on Leaf tier there is no routing. Since in this design we bridge on both tires we also use mlag on both tires.

Configuration of Spine Switches

Functions that Spine switch performs in A2 design –routing and redundant bridging - are similar to functions that Leaf switch performs in A1 design, so the configuration is similar to configuration of leaf switch in previous design. We need to do the following:

1. General settings
2. Create mlag domain and aggregated interfaces (port-channels)
3. Configure client VLANs for all client interfaces and IP addresses with vrrp configuration for L3 client interfaces.
4. Configure dynamic routing.
5. For combined A1/A2 architecture we also have to configure L3 point to point routing interfaces on some physical ports

Here are the details:

General Settings

This section contains basic settings like, management, snmp, ntp, syslog and initial port configurations, but here only basic port configuration is shown.

```
!-----------------------------------------------
! configure default values for all ports
!-----------------------------------------------
! Define default parameters for 1/10Gb ports
! some settings below (mtu lldp) are just example
interface range 0/1 - 0/32
 ! Define port mode parameters for 40Gb ports (required just once, then reload)
 port-mode 1x40g
 ! rest of port params aplied to all ports
 no spanning-tree port mode
 mtu 12288
 ; bidi detection is recommended if QSFP+ with MTO connectors are used
 udld port
 lldp transmit
 lldp receive
 lldp transmit-tlv port-desc sys-name sys-desc
exit
!
!interface 0/52
! port-mode 1x40g
!exit
!
```

Configuring mlag domain

1. Create port-channels (aggregated connection): one for peer-link - a redundant link connecting the two switches in a mlag domain, one for uplink connection, and others for client aggregated connections - one for each client.
2. Assigning ports to port-channels – for peer link port-channel at least two ports must be added (in our example 2x 40Gb), in uplink port-channel all remaining uplink ports are added (2x 40Gb) to single common aggregated uplink port-channel. For client connections typically each client requires one port-channel with just one port connected to each switch in mlag port-channel.
3. Creating mlag, assigning mlag domain id. It must have same value in configurations of both Spine switches (in this case 99).
4. Assigning port-channels to a mlag domain. Same mlag port-channel on both switches in a mlag domain must have same mlag id value.

```
!-------------------------------------------------------
! Configure mlag domain and interfaces
!-----------------------------------------------
! Create  portchannels
!------------------------------
! port channels for customer interfaces
interface port-channel 1
 description link to mlag domain 1
exit
!
interface port-channel 2
 description link to mlag domain 2
exit
!
interface port-channel 3
 description link to mlag domain 3
exit
interface port-channel 4
 description link to mlag domain 4
exit
!
! .....
!
! port channel for peer-gateway
interface port-channel 64
 description link to peer switch
exit
!
!--------------------------------
! Assigning ports to port-channels
!--------------------------------
! assign ports for peer-gateway port channel
interface range 0/31 - 0/32
 channel-group 64 mode active
exit
!--------------------------------
! assign ports for leaf mlag 1 port-channels
interface range 0/1 - 0/2
 channel-group 1 mode active
exit
```

```
!
! assign ports for leaf mlag 2 port-channels
interface range 0/3 - 0/4
 channel-group 2 mode active
exit
!
! assign ports for leaf mlag 3 port-channels
interface range 0/5 - 0/6
 channel-group 3 mode active
exit
!
! assign ports for leaf mlag 4 port-channels
interface range 0/7 - 0/8
 channel-group 4 mode active
exit
!
!---------------------------------
! Create mlag domain, set domain parameters
mlag
 mlag domain 99
 ! must be added to enable correct L3 packet forwarding within spine mlag
domain
 ! accept and route packets with MAC address of neighboring router in mlag
domain
 mlag peer-gateway
 ! set shortest keep-alive timer to provide fast conversion
 mlag peer-keepalive timeout 3
 ! following command must be removed before FW upgrade to minimize interference
 !   following command delays mlag port activation until mlag is active
 mlag member-linkdown
 ! you can define alternate redundant path for mlag control packets –
 !  i.e. through management interface (not set)
 ! mlag peer-keepalive destination 10.1.11.116
 ! mlag peer-keepalive source 10.1.11.115
 !
 ! assign default parameters to all port-channels
interface range port-channel 2 - 4
 no spanning-tree port mode
 mtu 12288
exit
!
! assign mlag IDs and other specific parameters to port channels
interface port-channel 64
 mlag peer-link
exit
!
! assign mlag id numbers to each client port-channels
interface port-channel 1
 mlag 1
exit
interface port-channel 2
 mlag 2
exit
interface port-channel 3
 mlag 3
exit
interface port-channel 4
 mlag 4
```


XTP029 - 150417040500

```
exit
```

Create client VLANs

In this architecture all VLANs span complete L2 network, so unique VLAN IDs must be used for each client VLAN. In L2/L3 portion of network all ip configuration is done on Spine switches. This means that VLAN configuration may be copied to all switches in in all L2/L3 domains. If mlag is already established, creating a VLAN on one switch creates it on whole domain (on both switches), so only VLAN descriptions need to be copied and ip configuration set on other switch. In combined architecture A2&A1 L3 only interfaces intended for routing are defined on Spine switches (as shown in A1 design). For L3 only sections client L3 interfaces are as in A1 design defined on Leaf switches and VLAN IDs for these interfaces may also be reused.

Following there are few examples of VLAN and IP interface configuration on Spine switches.

```
!-----------------------------------------------
! enable routing and vrrp
!-----------------------------------------------
ip routing
ip vrrp
!-------------------------------------------------------
! Configure client VLANs
!-------------------------------------------------------
! Define VLANs for L3 client interfaces
! Note: Client VLANs span complete network, VLAN IDs may be passed to
!  client interfaces, so careful planning of VLAN ID and ip address space must
be done
!
interface vlan 2
 description "First client routed VLAN"
exit
interface vlan 3
 description "First client routed VLAN"
exit
interface vlan 45
 description "First client bridged VLAN"
exit
interface vlan 220
 description "additional client routed VLAN"
exit
! optionally name client VLANs
vlan database
vlan name 2 "vSA2_C1"
vlan name 3 "vSA2_C2"
vlan name 45 "vSA2_C3"
vlan name 220 "vSA2_C220"
exit
!
! Configure VLANs
interface vlan 1
 shutdown
exit
!
! define ip interfaces for customeres VLANs
! each of these VLANs create a L2 interface for customer
```

```
interface vlan 2
  routing
  description 'Demo Routed customer VLAN 2'
  ip address 10.11.2.2 255.255.255.0
  ip vrrp 2
  ip vrrp 2 mode
  ip vrrp 2 ip 10.11.2.1
  ip vrrp 2 priority 150
  ip vrrp 2 accept-mode
exit
!
interface vlan 3
  routing
  description 'Demo Routed customer VLAN 3'
  ip address 10.1.32.2 255.255.255.0
  ip vrrp 3
  ip vrrp 3 mode
  ip vrrp 3 ip 10.1.23.1
  ip vrrp 3 priority 250
  ip vrrp 3 accept-mode
exit
!
interface vlan 45
  description 'Demo bridged customer VLAN 45'
  ! L2 only
  shutdown
exit
!
interface vlan 220
  routing
  description 'Demo Routed customer VLAN 220'
  ip address 192.158.4.2 255.255.255.0
  ip vrrp 4
  ip vrrp 4 mode
  ip vrrp 4 ip 192.158.220.1
  ip vrrp 4 priority 250
  ip vrrp 4 accept-mode
exit
!
```

All Client network must be added in OSPF definitions. This can be done with each network separately or as aggregated network prefix for multiple or all client networks.

```
!
! start ospf - see end of config file
router ospf
  router-id 0.1.0.1
  network 192.168.255.0 0.0.0.255 area 0.0.0.0
  enable
  !
  ! here comes all the networks of customer interfaces, may be aggregated
  network 10.11.2.0 0.0.0.255 area 0.0.0.0
  network 10.1.23.0 0.0.0.255 area 0.0.0.0
  network 192.158.4.0 0.0.0.255 area 0.0.0.0
  ! do not run ospf on customer interfaces
  passive-interface vlan 2
  passive-interface vlan 3
  passive-interface vlan 4
exit
```

```
end
```

Configuration on second switch in mlag domain is the same as shown above only IP addresses on L3 interfaces and router id terminate with .2 instead .1 as shown above. All virtual IP addresses are the same.

Add client VLANs on client port-channel interfaces.

All Spine port channels transfer tagged client VLANs. Each new VLAN must be added to all downstream port-channels each is connected to different Leaf mlag domain. Following configuration adds all VLANs above to all downlink port-channels:

```
! add all client VLANs to all downstream port-channels
interface range port-channel 1 - 4
 switchport allowed vlan add tagged 2-3,45,220
 switchport native vlan 3
 switchport allowed vlan remove 1
exit
!
```

Same can be defined also in the way as is displayed in configuration:

```
! add all client VLANs to all downstream port-channels
interface range port-channel 1 - 4
 switchport allowed vlan add 2-3,45,220
 switchport tagging 2-3,45,220
 switchport native vlan 3
 switchport allowed vlan remove 1
exit
!
```

Configuration of Leaf Switches

In A2 design leaf switches do only bridging, so only L2 configuration is needed. This includes creating mlag domain, creating VLANs, adding physical ports that connect client devices to client VLANs and adding all client VLANs to uplink aggregated port.

There are two dedicated port-channels: first one (po 64) is used as connection to other switch in mlag domain (peer-link) and the second one in the uplink port-channel (po 63) that connects a Leaf switch to Spine switches. All other port channels are used to connect to client devices – this part is application specific. In our example we assume that each client is connected with one port to each switch in a mlag.

Configuration steps are following:

1. Configure ports
2. Create mlag domain
3. Create client VLANs
4. Add ports to client VLANs and add all VLANs to uplink port-channel

preview uncompleted article - work in progress preview

Configuration:

```
!-----------------------------------------------
! configure default values for all ports
!-----------------------------------------------
! Define default parameters for 1/10Gb ports
! values below are just example
interface range 0/1 - 0/48
 mtu 12288
 udld port
 lldp transmit
 lldp receive
 lldp transmit-tlv port-desc sys-name sys-desc
exit
!
!-------------------------------------------------------
! Configure mlag and VLANs to clients
!-------------------------------------------------------
! Define VLANs for L3 client interfaces
! Note: Client VLANs and names span complete network
!
interface vlan 2
 description "First client routed VLAN"
exit
interface vlan 3
 description "First client routed VLAN"
exit
interface vlan 45
 description "First client bridged VLAN"
exit
interface vlan 220
 description "additional client routed VLAN"
exit
! optionally name client VLANs
vlan database
vlan name 2 "vSA2_C1"
vlan name 3 "vSA2_C2"
vlan name 45 "vSA2_C3"
vlan name 220 "vSA2_C220"
exit
!
! Create port-channels
!-----------------------------------------------
! port channels for customer interfaces
interface port-channel 1
 description link to mlag domain 1
exit
!
interface port-channel 2
 description link to mlag domain 2
exit
!
interface port-channel 3
 description link to mlag domain 3
exit
interface port-channel 4
 description link to mlag domain 4
exit
```

preview uncompleted article - work in progress preview

```
!
! .....
!
! port channel for uplink
interface port-channel 63
 description uplink port to Spine switches
exit
! port channel for peer-gateway
interface port-channel 64
 description link to peer switch
exit
!
!---------------------------------
! Assigning ports to port-channels
!---------------------------------
! assign ports for peer-gateway port channel
interface range 0/51 - 0/52
 channel-group 64 mode active
exit
!---------------------------------
! assign ports for uplink port channel
interface range 0/49 - 0/50
 channel-group 63 mode active
exit
!---------------------------------
! assign ports for client port-channels
! this is mlag domain specific part of configuration
interface range 0/1
 channel-group 1 mode active
exit
!
! assign ports for leaf mlag 2 port-channels
interface range 0/2
 channel-group 2 mode active
exit
!
! assign ports for leaf mlag 3 port-channels
interface range 0/3
 channel-group 3 mode active
exit
!
! assign ports for leaf mlag 4 port-channels
interface range 0/4
 channel-group 4 mode active
exit
!
!---------------------------------
!
! Create mlag domain, set domain parameters
mlag
 mlag domain 01
 ! set shortest keep-alive timer to provide fast conversion
 mlag peer-keepalive timeout 3
 ! following command must be removed before FW upgrade to minimize interference
 !   it delays mlag port activation until mlag is active
 mlag member-linkdown
 ! define alternate redundant path for mlag control packets - i.e. through
management interface
```


XTP029 - 150417040500

```
 ! mlag peer-keepalive destination 10.1.11.116
 ! mlag peer-keepalive source 10.1.11.115
!
!
! assign default parameters to all port-channels
interface range port-channel 2 - 4
 no spanning-tree port mode
 mtu 12288
exit
!
interface range port-channel 63 - 64
 no spanning-tree port mode
 mtu 12288
exit
!
! assign mlag IDs and other specific parameters to port-channels
interface port-channel 64
 mlag peer-link
exit
!
interface port-channel 63
 mlag 63
exit
!
! assign mlag id numbers to each port-channel
interface port-channel 1
 mlag 1
exit
interface port-channel 2
 mlag 2
exit
interface port-channel 3
 mlag 3
exit
interface port-channel 4
 mlag 4
exit
!
! add all client VLANs to upstream port-channel
interface port-channel 63
 switchport allowed vlan add tagged 2-3,45, 220
 switchport native vlan 1
exit
!
!
interface vlan 1
exit
!
! define specific config customers VLANs if any
! no L3 config in this part
interface vlan 2
exit
!
interface vlan 3
exit
!
interface vlan 45
exit
```

preview uncompleted article - work in progress preview

```
!
interface vlan 220
exit
!
end
!
!-------------------------------------------------------------------
!
! enable all ports - after when adjecent switch is disabled
interface range 0/1 - 0/52
no shutdown
exit
```

Adding VLANs to client client port-channel interfaces.

In many cases a single untagged VLAN is added to client interface – a mlag port-channel. For multiple services on single client interface, multiple tagged client VLANs can be added to client interface.

```
!Single untagged VLAN configuration
interface port-channel 4
 mlag 4
 switchport allowed vlan add 33
 switchport native vlan 33
 switchport allowed vlan remove 1
exit
!
! or Multiple tagged VLANs configuration
interface port-channel 4
 mlag 4
 switchport allowed vlan add tagged 33,15,20-25
 switchport native vlan 33
 switchport allowed vlan remove 1
exit
!
or Multiple tagged VLANs configuration with single untagged VLAN
interface port-channel 4
 mlag 4
 switchport allowed vlan add tagged 15,20-25
 switchport allowed vlan add untagged 33
 switchport native vlan 33
 switchport allowed vlan remove 1
exit
```

Configuration of other Leaf switches

If we use physical ports with same port numbers on both switches to create mlag port-channels, than the configurations on both switches in mlag domain are identical and can be copied from one switch to the other unchanged, with exception of switch hostname, and management interface IP address and gateway if they are manually configured.

Leaf switches in additional domains have also identical configurations to the one presented, with exception of mlag domain id and optionally different assignments of client ports to client port-channels and management ip addresses.

Leaf switches in L3 only parts of network have configuration as was described in A1 design above.

preview uncompleted article - work in progress preview

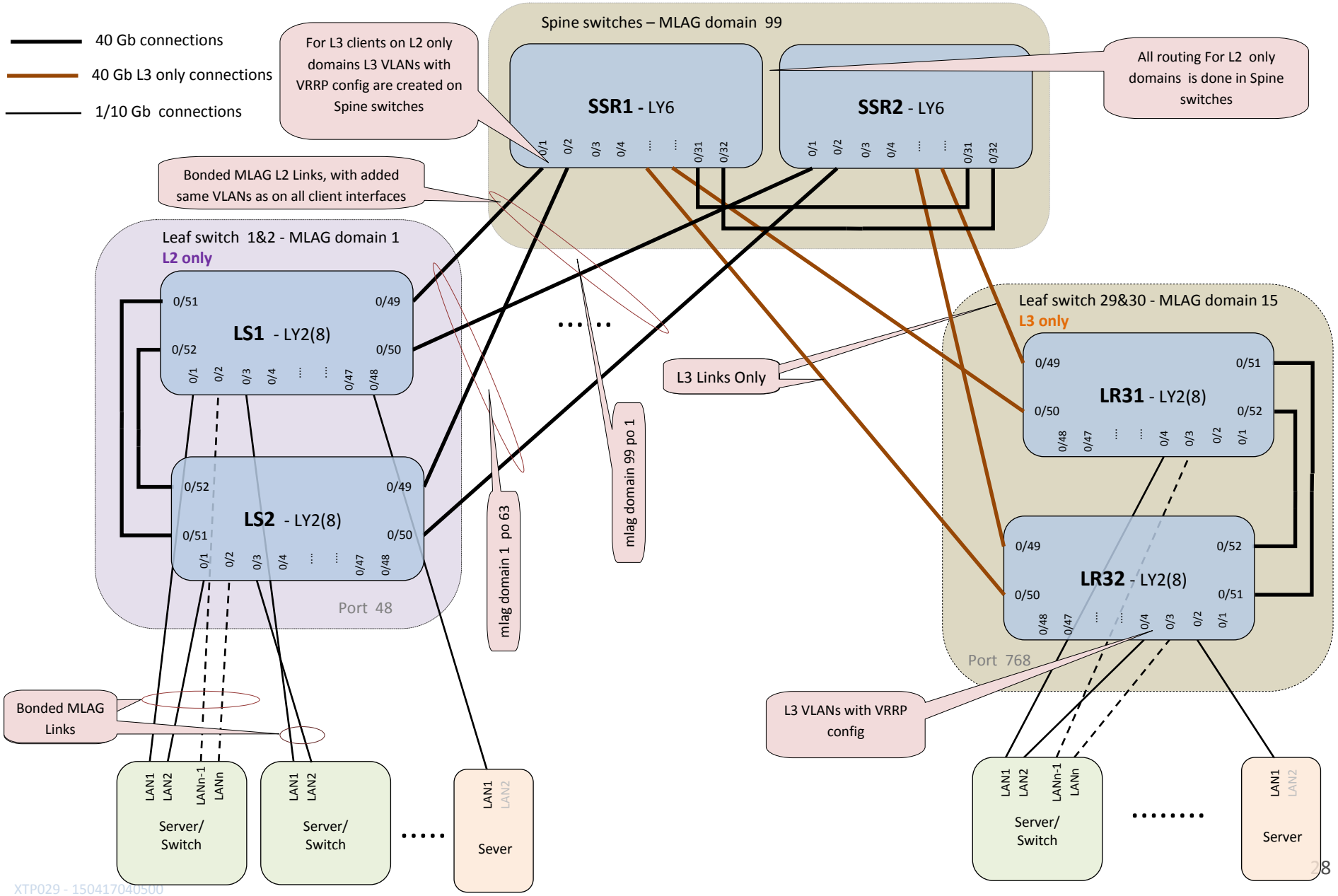


Figure 6 combined A1/A2 Spine/Leaf architecture combining L2/L3 sections and L3 only sections

XTP029 - 150417040500

Delays, Throughput and Scaling

All so far described architectures A1 & A2 and combined A1&A2 were designed as “fabric topologies” where bandwidth and delays are equal for communication between any two clients. While this is always true for L3 only traffic in A1 architecture there may be slightly different delays for asymmetrically connected clients in L2 network, since there will occasionally be one or two additional hops within mlag domain needed, depending on lag load balancing hash algorithm. (in L2 there may be different number of hops within the L2 network for different users due to L2 LAG load balancing algorithms (3 to 5), in L3 there are always 3 hops)

All shown architectures using LY2R switches have aggregated uplink speed of 4 x 40 GB for each mlag domain with 2x48 10Gb client ports. These designs provide oversubscription ratio of 1:6. If higher throughput is required then we can use LY8 switches as Leaf switches, that have 48 x 10Gb and 6 x 40Gb ports and can in same designs provide up to 8x40Gb uplinks for each mlag domain, thus providing oversubscription ratio of 1:3. As a consequence with LY8 switches with double uplink bandwidths In same designs can scale just to half the number of mlag domains as with LY2R switches – up to 8 mlag domains in A2 and A1&A2 design and still with up to 64 mlag domains (128 Leaf switches) with A1 architecture (using uplinks as 4x10G connections), but if we want to scale to this number we also have to double the number of Spine switches.

Non Fabric Designs

Combined design can be implemented also on many other ways that do not preserve fabric topology, but have other benefits.

In combined network A1&A2 design we have same scaling as in the A2 design if we want to preserve same delays, throughput and oversubscription ratio in both parts of the network. With lower bandwidth connection between the two sections and/or with longer delays (more hops) between the sections, we can modify the combined architecture so that it can scale the L3 design to significantly larger number of nodes.

Equal delay lower bandwidth solutions, can simply add additional spine switches to implement L3 only routing for additional leaf switches (as in A1 design). Leaf switches have just a portion of uplinks connected to switches providing L2/L3 connectivity, and the rest of them to other spine switches that provide L3 only connectivity. In such design bandwidth between L2/L3 and L3 only domain is limited to bandwidth of uplinks connecting from L3 only switches to L2/L3 spine switches.

Increased delays (hop count) designs can be implemented with an additional tier of switches either above spine switches or below leaf switches in current A2 design. In this case there are at least some paths that have longer delays (more hops) between different sections of the network. Additional delays are in range from 600ns to few us (or less with cut through switching). Weather this is a problem, depends on applications running on such networks.

Architecture of enterprise access network

Most of enterprise client devices connect to network with non redundant connection, still we want to preserve as much redundancy in the network as possible also in implementation of client access network.

In our solution we used similar design as one mlag Leaf structure of Spine/Leaf architecture A1 with an additional layer (tier) of Edge switches that connect redundantly to both routers in this Leaf mlag structure that acts as access routers. Edge switches directly connect to clients, typically with UTP and are connected to access routers with optical 10G uplinks. They are typically installed distributed in wiring closets close to clients.

Access routers that form mlag domain route traffic between all customer VLANs and between customer VLANs and rest of the network. All access rules are implemented on these routers. They limit access between different sections of the network. Since there is just one set of access rules (in two copies) that control this, it is easy to maintain them.

If more complex statefull rules are required, external firewall may be added and redundantly connected to access routers. This way also WAN connection can be implemented. Connection to multiple separate firewalls (to optimize cost) and/or application delivery servers and load balancers for different applications (to provide adequate throughput) are also supported.

Design of this section (access network) is very simple and is shown on Figure **9**.

This design implements following goals:

- Edge switches are the only non redundant part in the system. Failure of this switch affect only users directly connected to the failed switch.
- Concentrates the main part of access routing to single redundantly implemented device, where it is easy to configure and maintain access rules that limit access of specific users to certain network resources.
- Eliminates use of spanning tree within the network. It is configured on edge switches just to guard against accidental loops, caused by users. This protection uses also other mechanism to prevent undesired consequences of loops caused by users like – loop detection, that effectively detects loops within a switch and limited mac moving threshold with port blocking, storm control thresholds (broadcast, multicast, unknown uncast)...
- Easily integrates in common datacenter network with standard L3 connections to the rest of the network. This enables independent implementations of each part of network (i.e. with equipment from different vendors and/or reuse of existing equipment)
- On links to Edge switches only standard protocols (LACP) are implemented and standard functions, like VLAN support are needed. Most managed switches can support this. For some users, that do not require high bandwidths existing switches may be reused.

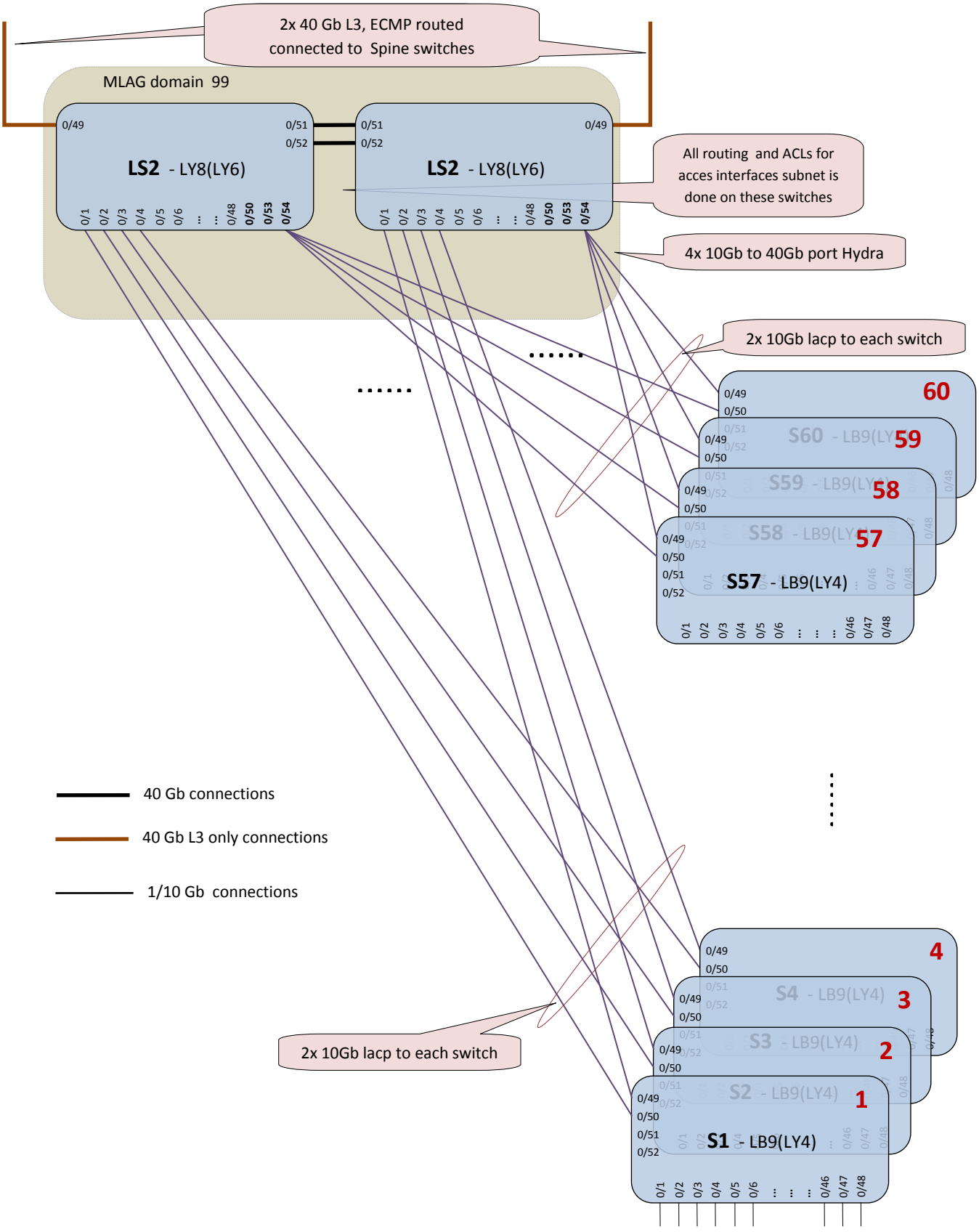


Figure 7 Access network implementation

preview uncompleted article - work in progress preview

Configuration of acces network

Configurations of switches LS1 & LS2 are very similar to configurations of leaf switches in A1 design so we will not display again all detail here. Edge switches are just L2 switches, that implement aggregated uplink connection with LACP protocol and VLANs needed to connect the clients. Actually to simplify configuration, all VLANs may be defined on all Edge switches, so that except of actual port assignments to VLAN(s) and switch names and management IPs the rest of configuration is identical on all switches.

All steps to configure the access routers were already described so here we just list them.
create port-channels

Edge switches:

1. Configure ports
2. Create uplink port-channel
3. Create client VLANs and add them to uplink port-channel
4. Adding uplink ports to all client VLANs
5. Adding client ports to client VLANs

Access routers (mlag domain):

6. Configure ports (including change of port mode for ports that use 40Gb ports as 4x 10Gb)
7. Create mlag domain
8. Create client VLANs
9. Optionally naming VLANs
10. Defining VLAN IP interfaces with vrrp configuration (interface ip and virtual ip)
11. Adding (all) client ports to client VLANs and adding all VLANs to uplink port-channel
12. Configuring dynamic routing (OSPF routing)

preview uncompleted article - work in progress preview

Conclusion

Described were just some of possible architectures that cover the needs of wide range of applications – From medium sized, simple to manage L2 networks, to highly scalable L3 only networks suitable for largest data centers. All provided redundancy without using slow converging protocols. Some of the benefits of these designs are:

- Providing network solutions, where we can balance between the scaling capability and bandwidth capability
- Distributed network design provides very resilient network, where even multiple failures have impact on just a small portion of the network and in most cases just degrade performance.
- Ability to gradually expand the network in small increments as requirements demand, without affecting work in part of the network that is already running.
- Use of fast low cost highly integrated switching equipment minimizes initial investment, overall cost and allows simple maintenance, since we can have all network components on stock. These can be used both for maintenance or expansion of the network whenever required.
- These networks have very low power consumption and needed power scales together with the size of the network. Networks with power consumption of less than 10W/10Gb connections can easily be built this way.

Improvements

Using latest switches and some newer overlay protocols like VXLAN/NGRE we can create similar architectures that can scale to even larger systems with redundant connections and provide dynamically connected L2 subnets, where actual redundancy is assured with fast converging L3 routing protocols and ECMP routing and L2 connectivity with overlaying L2 tunneling protocol. These solutions include all the advantages in a single architecture that are especially important in large installations – scaling to large networks, and providing large AND dynamically connected L2 networks all for tens of thousands of nodes.

Configuration and use of overlaying protocols in networks described will be presented in one of the next technical papers.

Converged networks

Modern networks transfer different types of traffic, most commonly - storage and standard IP data. Where this is required we have to employ additional configurations that process different categories of traffic differently, assure lossless traffic for storage protocols and that control traffic flows on the edges of the network and preferably per flow type. These are functions of Datacenter Bridging set of protocols. Description of use and configuration of Datacenter protocols will be presented in one of the next technical papers.