aruba

a Hewlett Packard Enterprise company

VSX overview

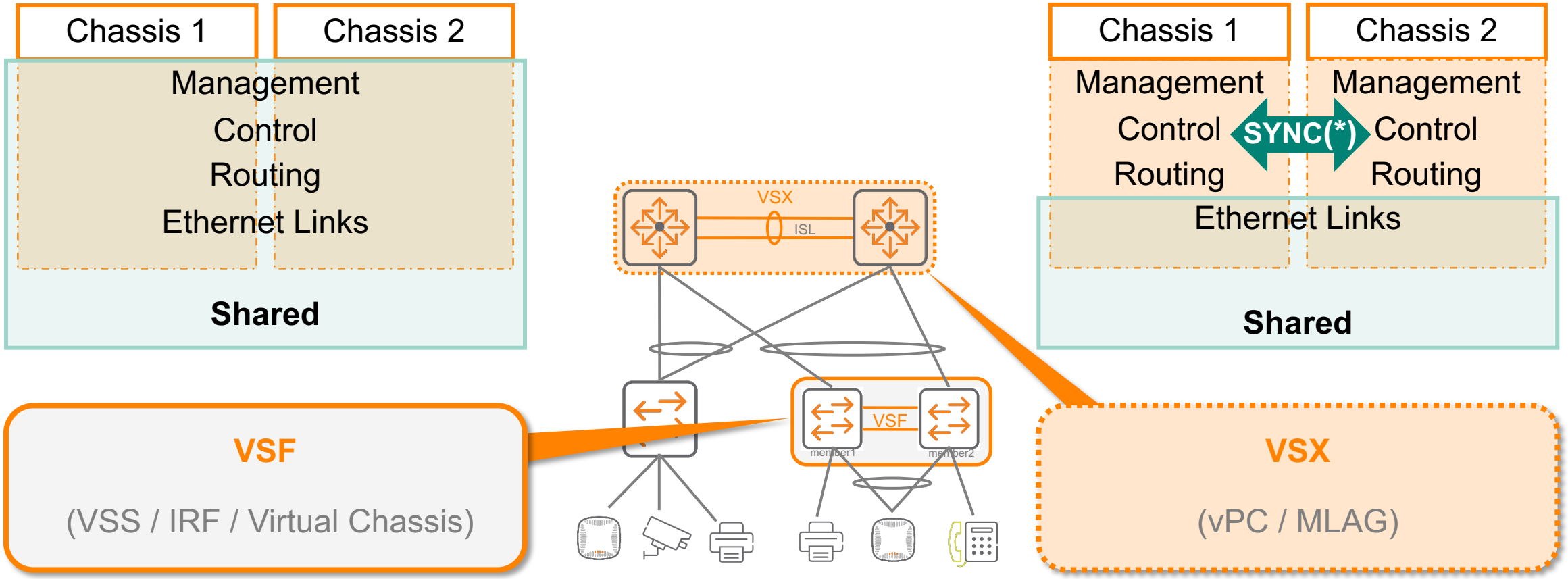
Core/Aggregation Virtualization Solutions

Market landscape

Aggregation / Core	8400/8320/8325 **VSX** 5400R **VSF**	Catalyst 68xx/45xx **VSS** Catalyst 9500/3850 **Stackwise Virtual*** Nexus 3/5/7/9xxx **vPC**	75xx Modular 7xxx Fixed **MLAG**	EX 8/9xxx **Virtual Chassis**	59xx/75xx/105xx 79xx/129xx **IRF**
Access	2930F/5400R **VSF**	Cat45xx **VSS** Cat2/3/9k **Stacking**		EX 2/3/4xxx **Virtual Chassis**	5500/5100/3600 **IRF**

Switch Virtualization Solutions

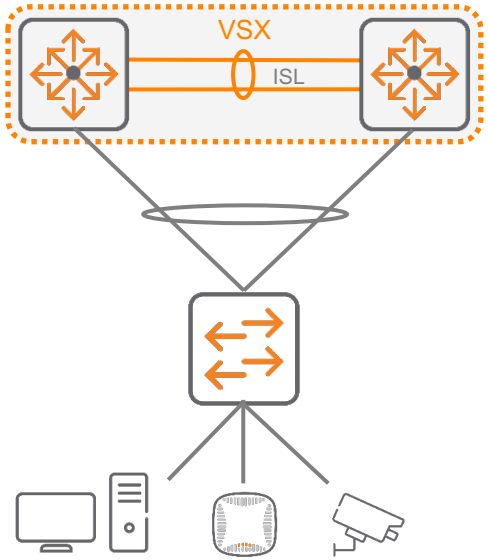
Comparison



(*) different levels of synchronization

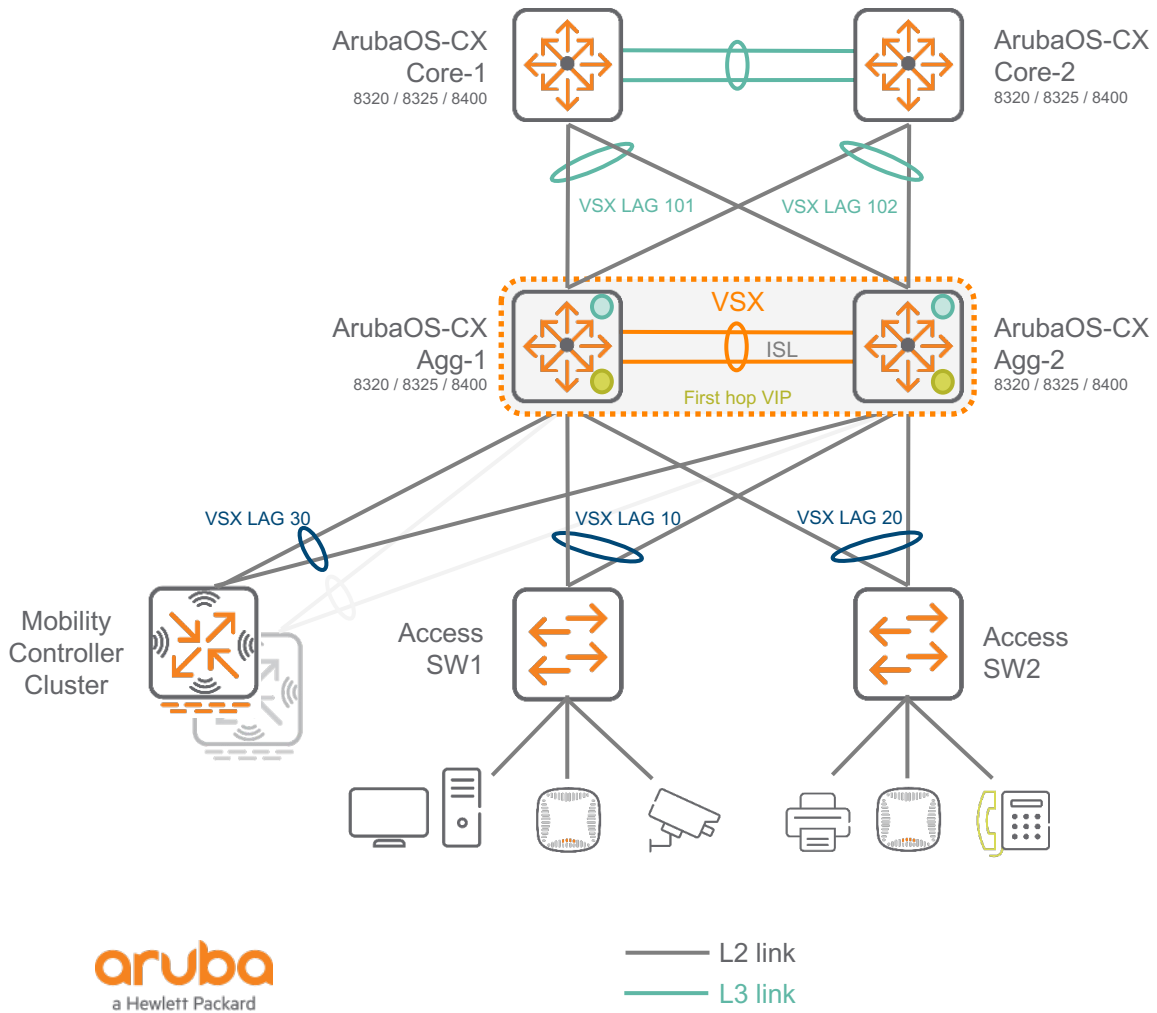
Aruba Virtual Switching Extension

VSX Key Principle: **High Availability by design during upgrades**



- Bind 2 same ArubaOS-CX switches to operate as one device for L2 but as independent nodes for L3.
- Support for active-active data-path:
  - Active-active L2
  - Active-active L3 unicast
  - Active-active L3 multicast
- Operational simplicity and usability:
  - for configuration
  - for troubleshooting
- Similar VSF benefits with better HA during upgrade

VSX Benefits



Control & Management Plane

- Dual control plane for best resiliency
- Unified management (synchronized configuration and easy troubleshooting)
- Independently software upgradable with near zero downtime
- In-chassis redundancy (8400) & device level redundancy

L2 Distributed LAGs (Agg to Acc)

- No spanning-tree
- Loop-free L2 multi-pathing (active-active)
- Rapid failover
- Simple configuration

L3 Distributed LAGs (Core to Agg)

- Distributed L3 over VSX pair (various options: ROP, SVIs or LAG'd SVIs)
- Unified data path (active-active first hop gateway)
- L3 ECMP + L2 VSX LAG (highly fault tolerant) with active-forwarding

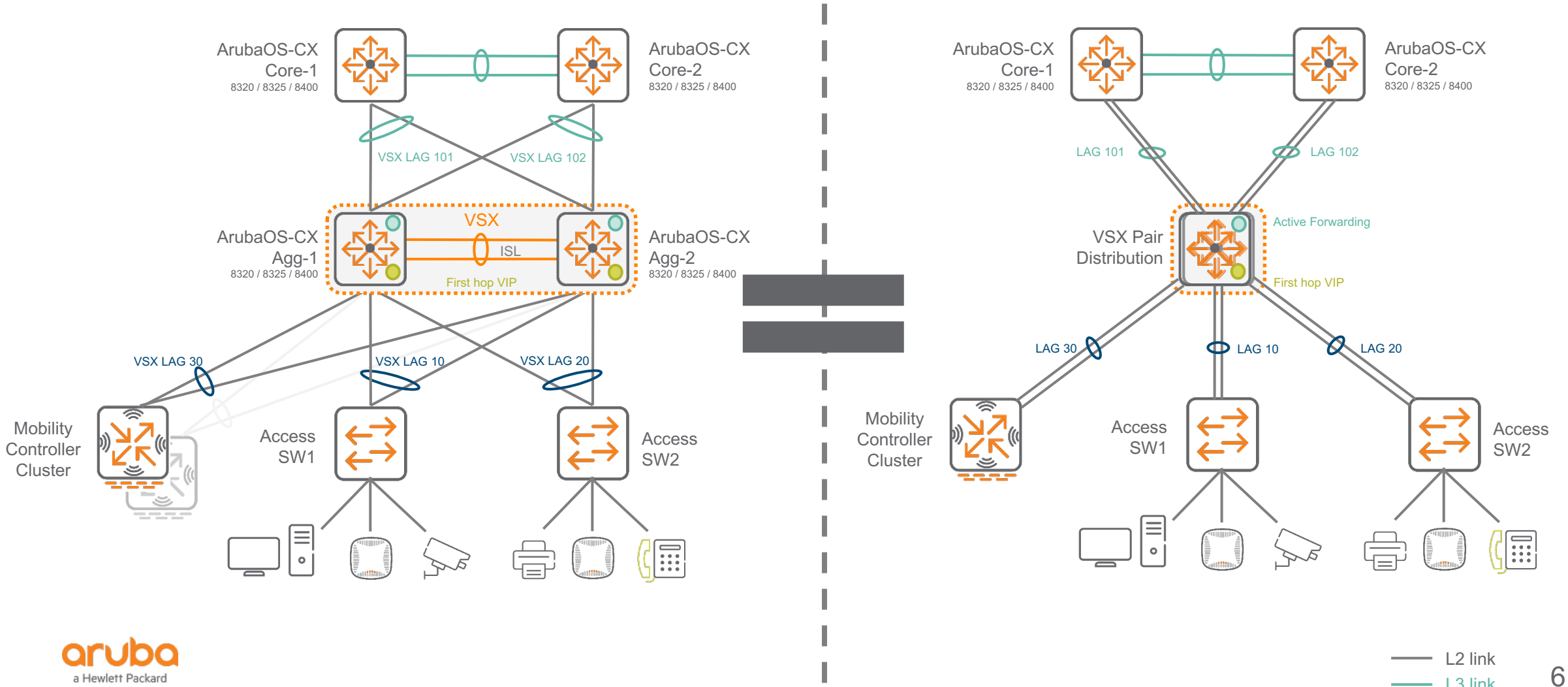
Active Gateway

- Active-Active first hop gateway (VIP)
- No VRRP/HSRP
- Simple configuration (1 command)
- No gateway protocol overhead
- DHCP relay redundancy

aruba
a Hewlett Packard Enterprise company

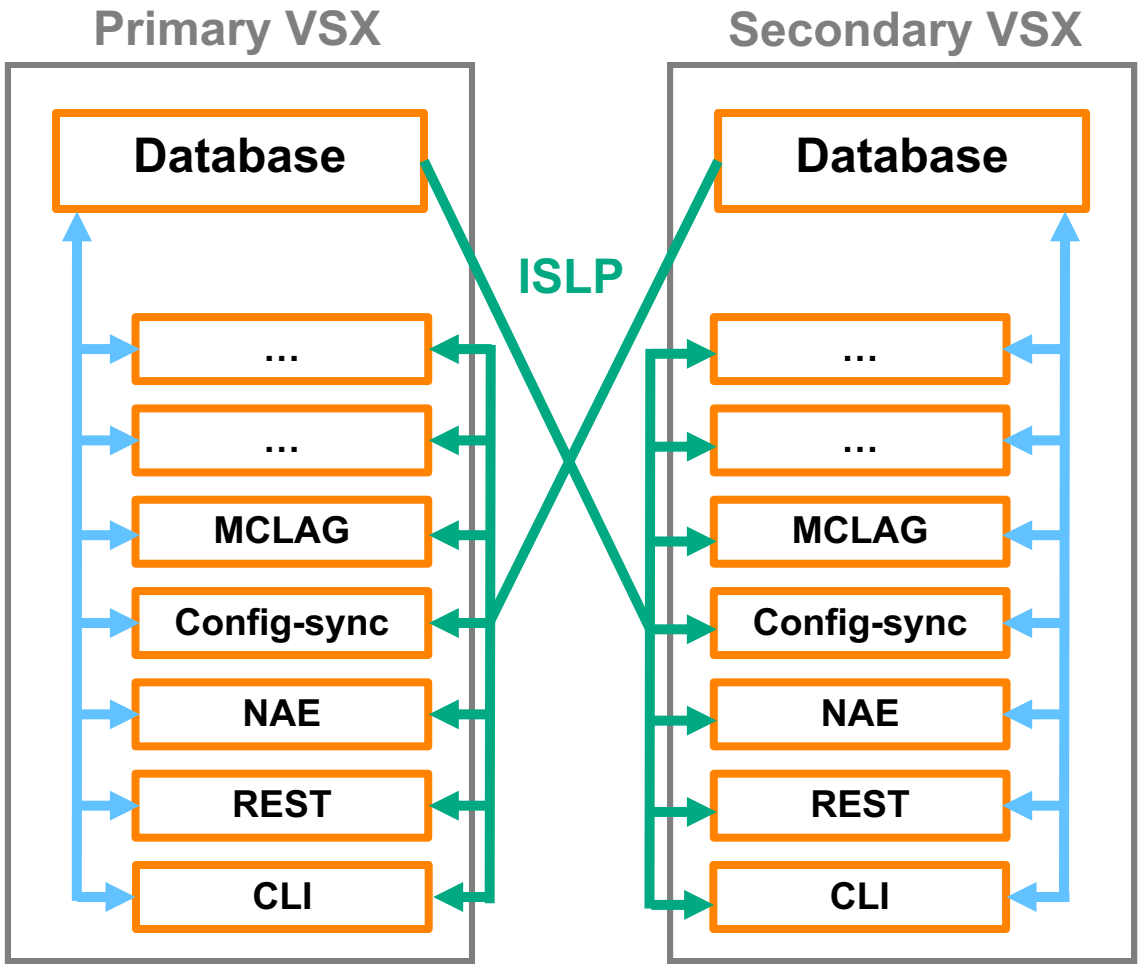
VSX - Data Plane Virtualization

Multi-Chassis Link Aggregation (MCLAG)



VSX - Management Plane Synchronization

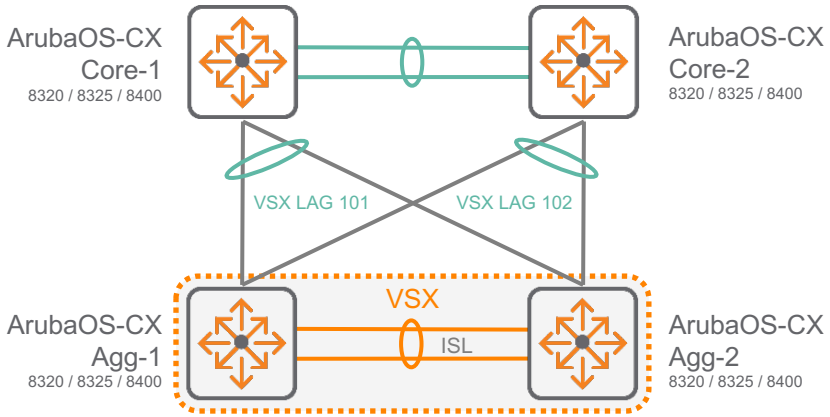
vsx-sync and vsx-peer

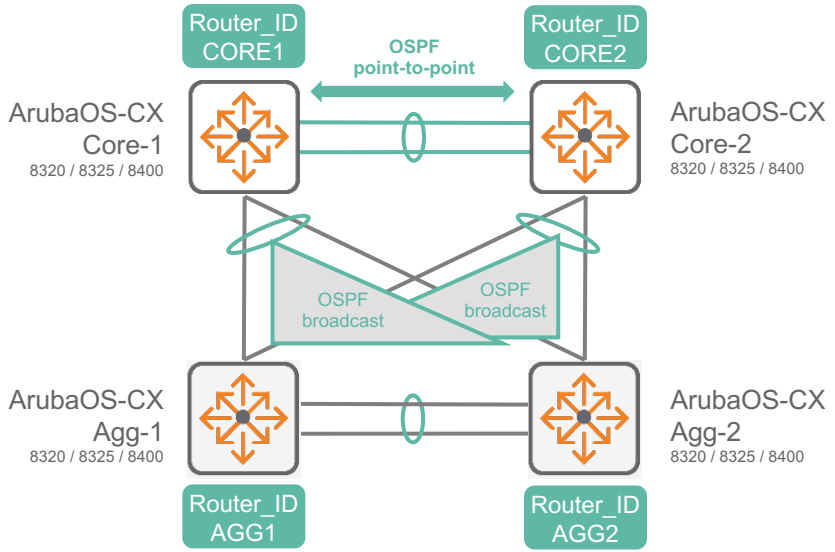


- **Database driven architecture (SystemStateDB)**
  - Allows active-active components to know the state of the peer
  - Enables CLI/REST/Web-UI to easily expose both control planes in a single place
  - Allows analytics across the VSX pair
- **Configuration and troubleshooting simplicity**
  - Continuous synchronization of the common configuration
  - Show commands that aggregate/contrast information from both switches for ease of troubleshooting
  - Show commands support for “vsx-peer” to show information from the peer device
  - Provide joint view of the VSX system
- **Hitless upgrade orchestration**
  - Allow traffic switchover from the switch that undergoes upgrade to the remaining forwarding switch
- **Active-Active Analytics**
  - NAE agents will cross monitor each others TS database
  - Detect discrepancies that remain for too long
  - Validate that the overall solution is healthy

Control Plane Separation

Each VSX node has its own router_ID

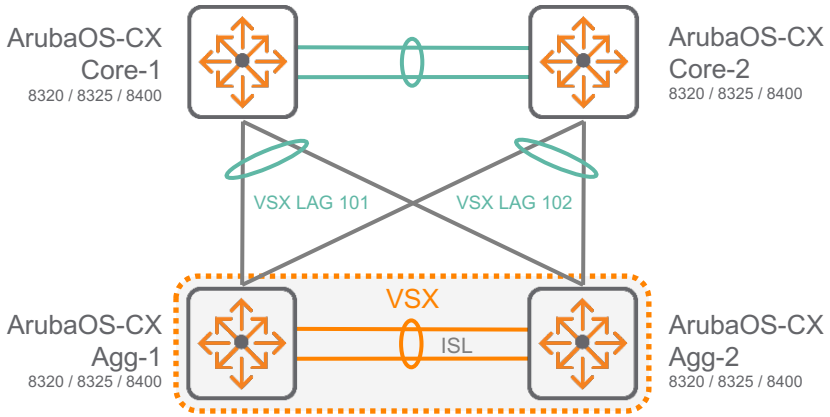


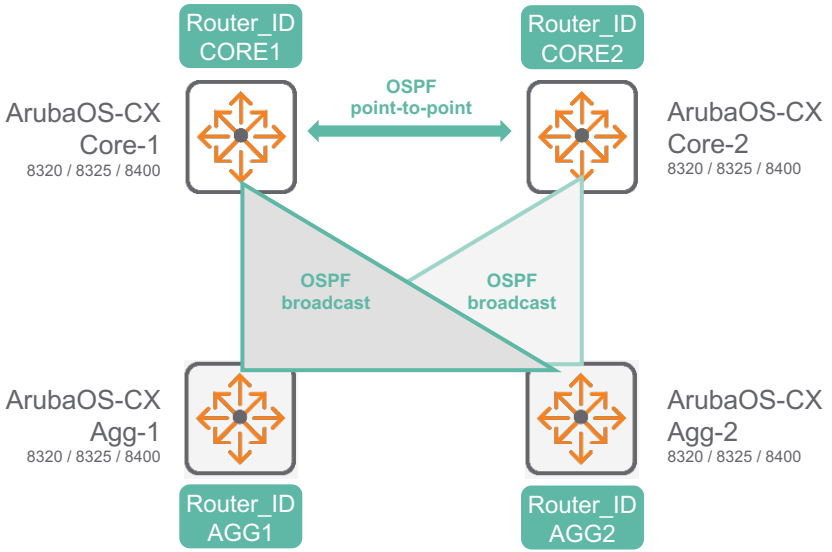


L2 link
L3 link

Control Plane Separation

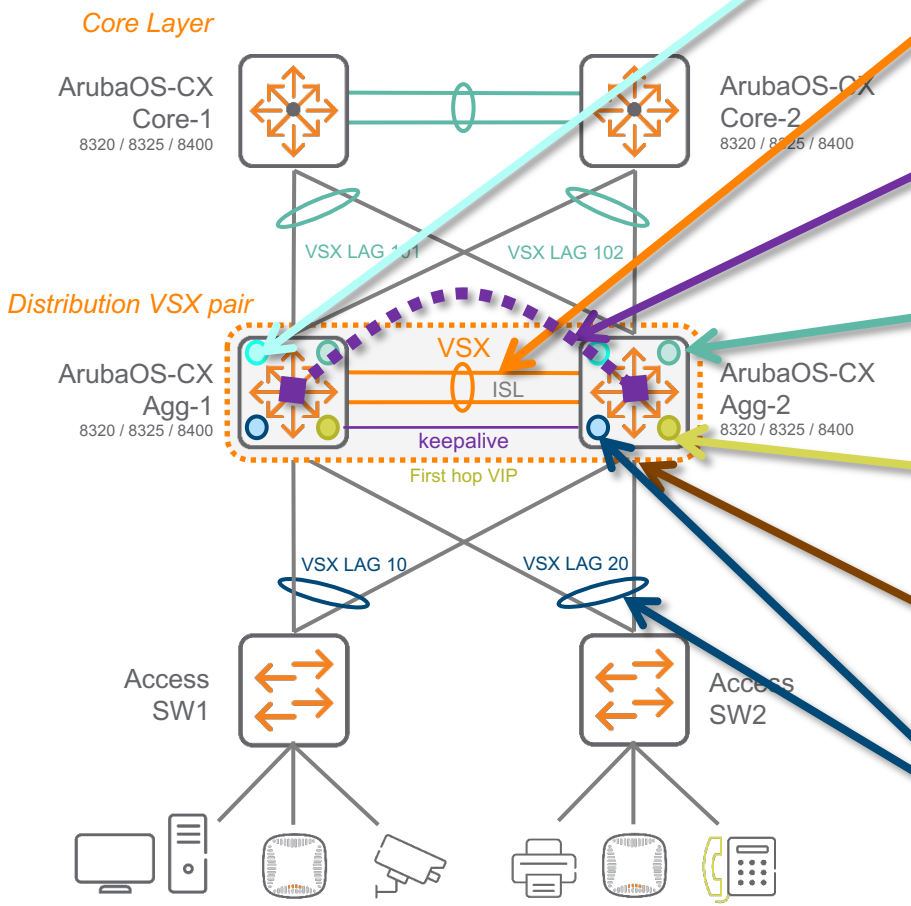
Logical view





L2 link
L3 link

VSX Components



PIM Dual-DR

Inter Switch Link (ISL)

Keep Alive Mechanism

Active-Forwarding

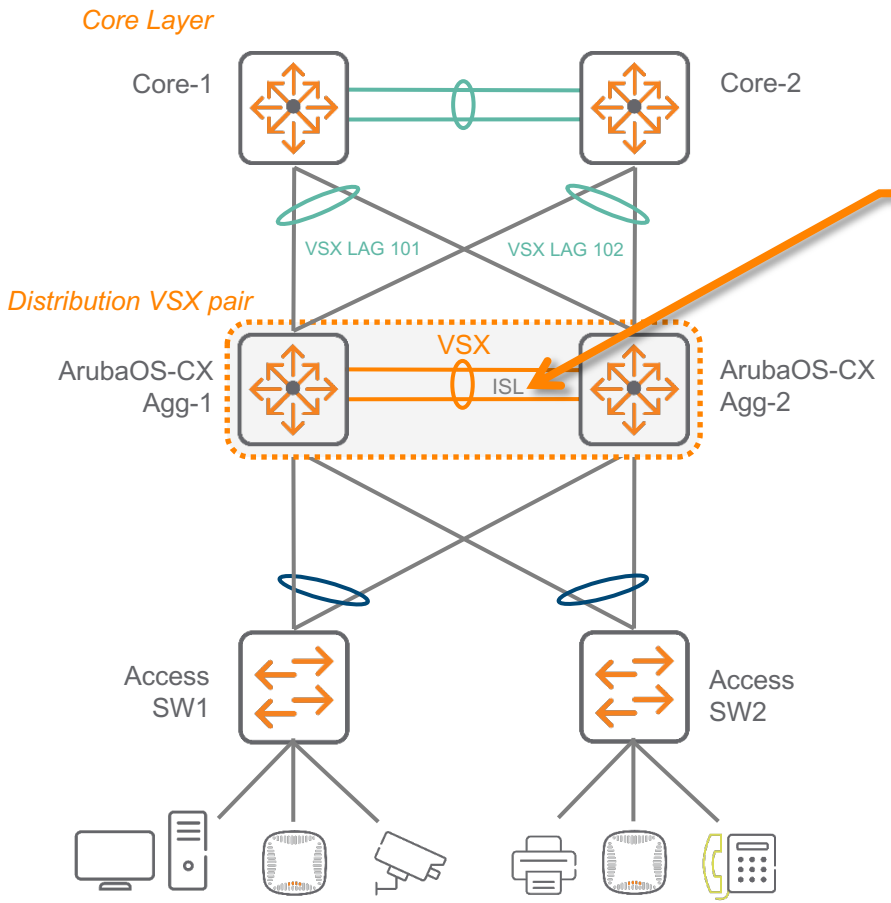
Active-Gateway + DHCP

Linkup Delay

VSX LAG
Multi-Chassis Link Aggregation
+ VSX system-mac

aruba
a Hewlett Packard Enterprise company

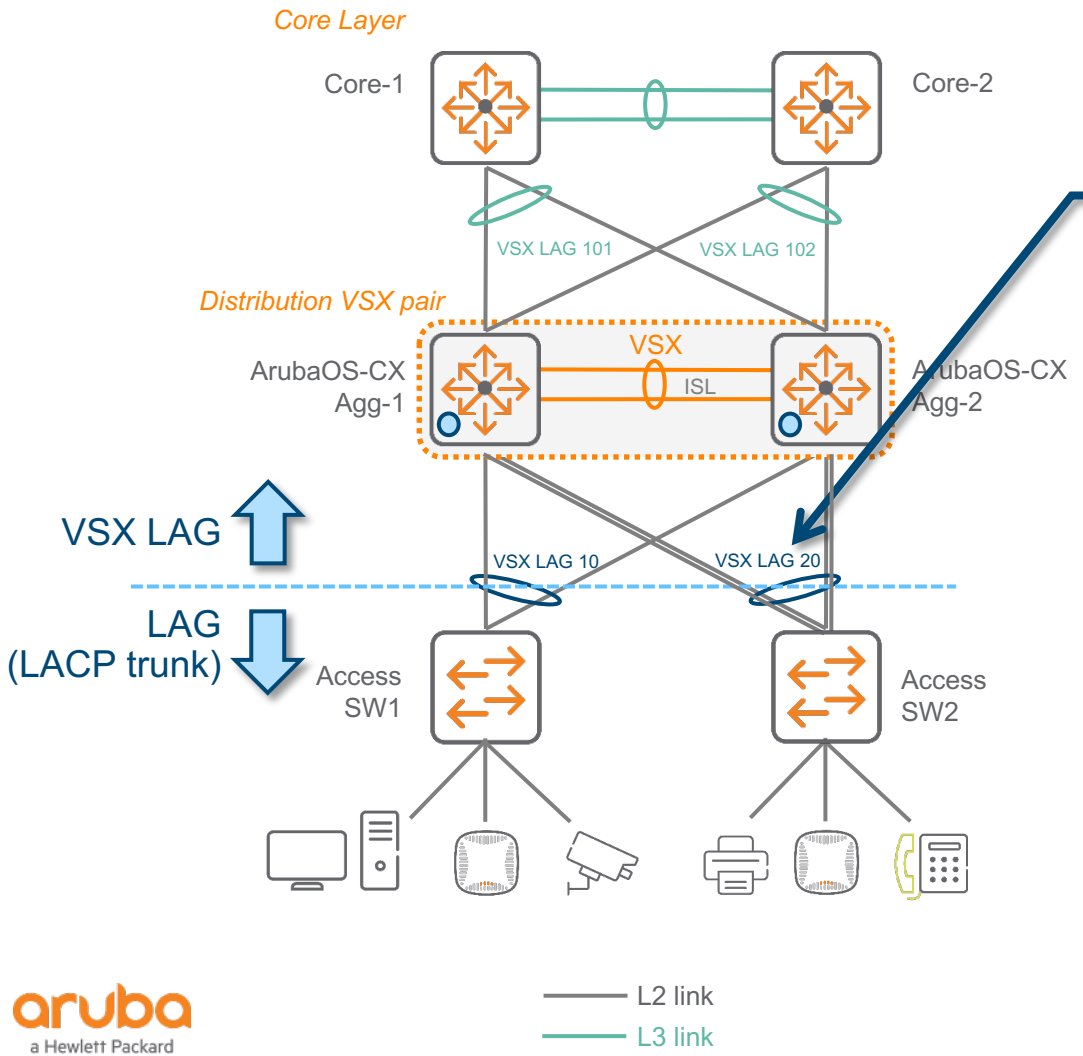
VSX Components
ISL



Inter Switch Link (ISL)

1. Each VSX switch needs to be configured with an ISL link that is **directly connected** to its peer VSX switch.
2. ISL can be a single circuit but aggregated circuits - **LAG is strongly recommended** (up to 8 physical links). Ports must have same speed.
3. Speed could be 10G, 40G or 100G. Prefer 40G or 100G. Example: 2x40G
4. ISL can span **long distances** (transceiver dependent).
5. ISL link is used for data path traffic forwarding, control plane VSX protocol exchange and management plane for peer management.
6. Traffic going over the ISL has **no additional encapsulation**.
7. ISLP is the protocol that runs over ISL and that is used to **synchronize LACP states, MSTP states, MAC and ARP tables** and configuration.
8. A **hello packet** is periodically exchanged just to make sure the peer's control plane is alive (range [1..5]s, def 1s). ISL also has a **dead-interval** range of 2..20 and default is 20. If a device does not receive a hello packet form its peer within the dead interval, it treats the peer device as dead and goes for a split detection.
   ISL port is treated as down when it stays down for the configured **Hold-time** (default=0s) interval.
9. All QoS/ACL policies that can be applied to network ports can be applied to ISL as well.

VSX Components
VSX Link Aggregation

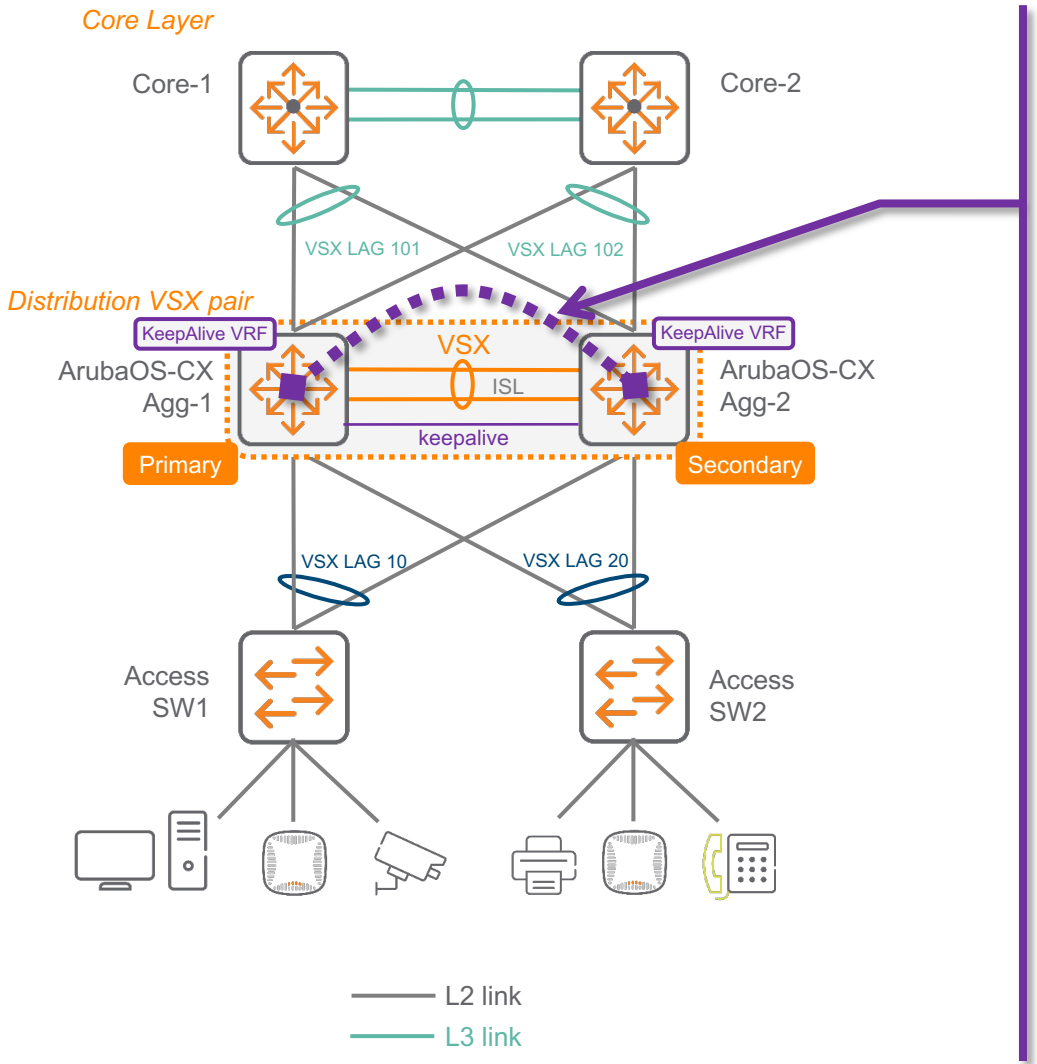


VSX LAG = MCLAG

1. VSX LAG is a Link Aggregation technique, where two or more links across **two** switches are aggregated together to form a LAG which will act as a **single logical interface**.
2. The two switches appear as a **single peer ID** to partner devices upstream and downstream that form a LAG with the VSX pair. The Sytem ID can be configured with **VSX system-mac**.
3. Ports must have **same speed**. You can not add one spare port (unused) that has lower speed than currently selected port.
4. Up to **4 physical links per chassis**, for **8 links max per VSX LAG**.
5. As of today, VSX LAG are **Layer2 only**.
6. VSX LAG is preferably **LACP based**. Non-LACP or static VSX LAG is also supported.
7. Both switches **synchronize their LAG states** using ISLP (ISL protocol).
8. User can **configure lag-hash** scheme that will then be used to spread traffic among the LAG links.
9. VSX lags are '**local optimized**'. i.e. when there are local links in a lag, they are alone used for forwarding. The configured LAG hash-scheme is constrained to just look at the local-links. ISL links are used only when local links of a VSX lag are down.
10. Number of VSX LAGs: 255 on 8400, 53 on 8320, 55 on 8325.

VSX Components

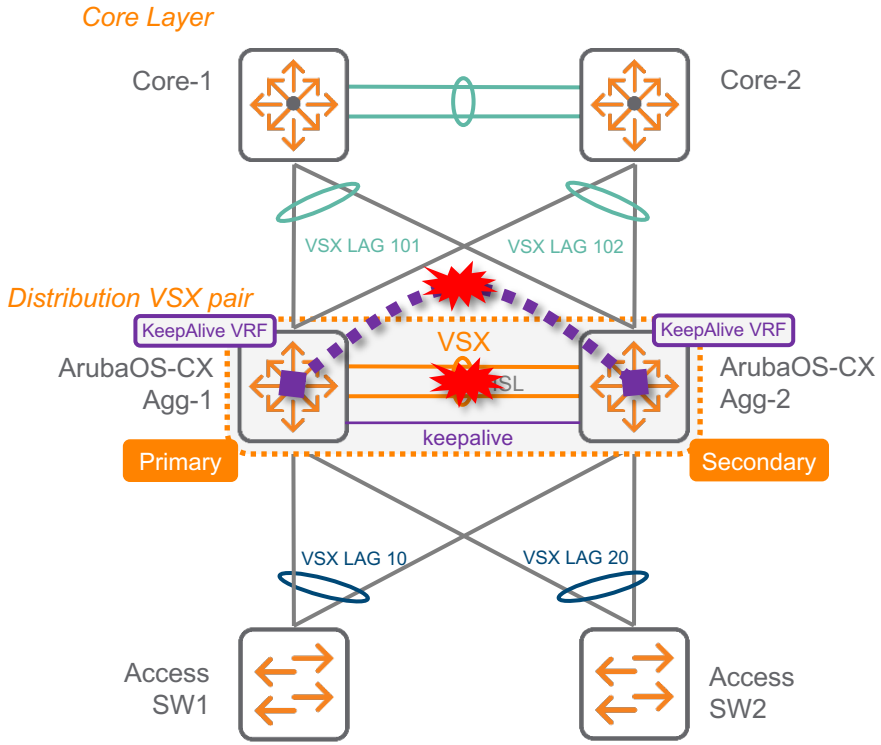
Split Brain Protection - Keepalive



Keepalive

1. Each VSX switch should be configured with a keep-alive connection to the other switch. When ISL goes down while VSX switches are still up, each VSX node uses keep-alive communication to identify split brain situation.
2. Under ISL failure condition, the user configured **Primary** VSX switch will keep its multi-chassis (VSX) LAG links **UP** and the ***Secondary*** VSX switch will force its multi-chassis (VSX) LAG links to go ***DOWN***.
3. The keep-alive communication is established over a routed network (IPv4).
4. Keep-alive packets are **UDP based**. UDP port is configurable. (Default port 7678)
5. Keep-alive mechanism **does not have to use a direct link** unlike ISL. Upstream L3 network can transport keep-alive messages.
6. Keep-alive packet can be sourced from loopback IP address.
7. It is **not recommended** to have keep-alive communication over ISL.
8. Hardening options: use a dedicated L3 link (can be 1G speed), a dedicated “Keep-Alive” VRF.
9. Keep-alive **hello** packets: sent every 1s (default). Configurable **hello-interval** range is 1 to 5s.
10. Keep-alive **dead-interval**: 3s (default). Configurable range of 2..20. If a device does not receive a keep-alive packet from its peer within the dead interval, it treats the peer device as out-of-service.

Split Brain
Scenario

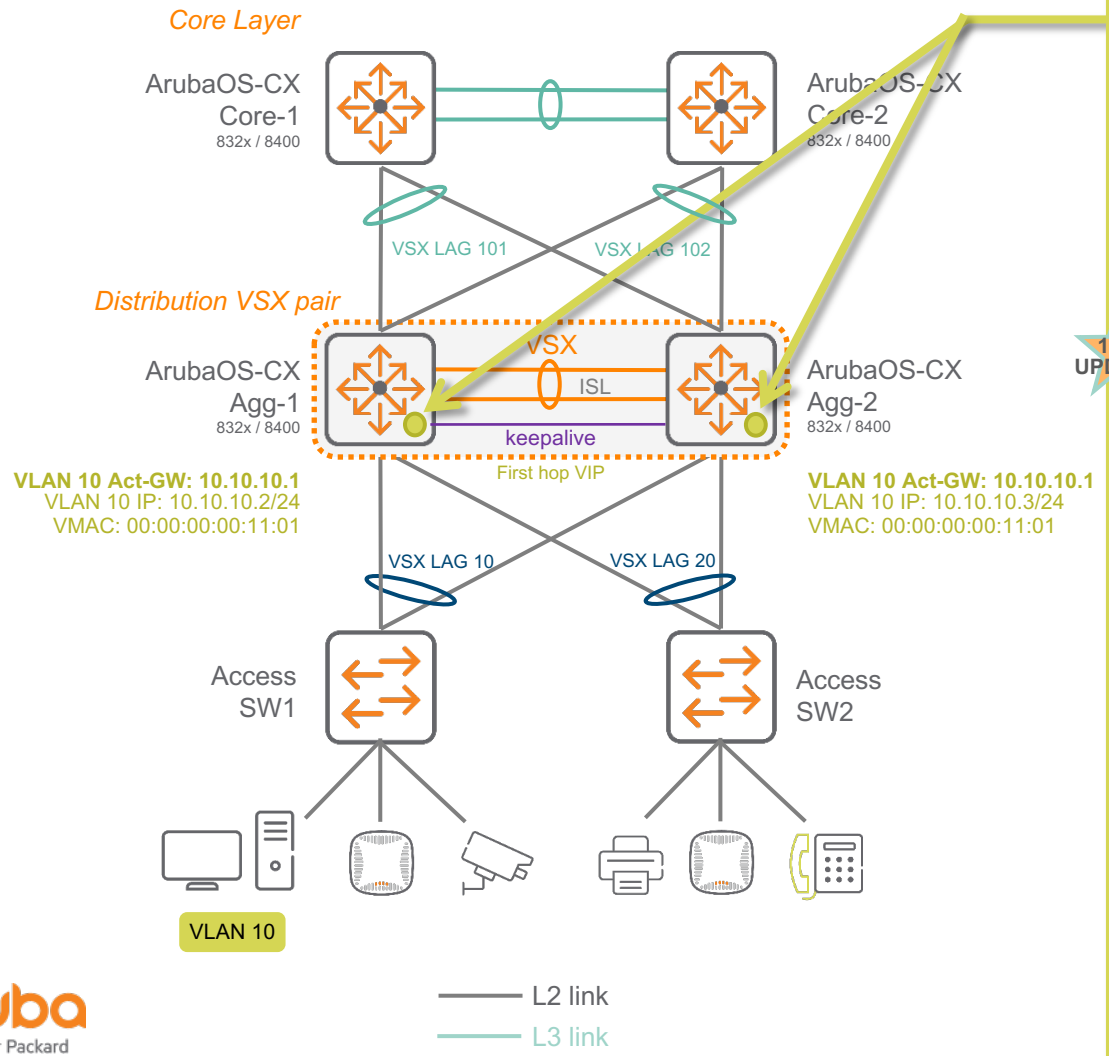


	Keepalive-Established	Keepalive-Init
ISL "In-Sync"	**Normal Operation**	No traffic impact. No protection.
ISL "Out-of-Sync"	On Secondary only: VSX LAG Down and associated member ports	**Split Brain:** **Both out-of-synced nodes forwarding.**

When ISL is restored, there is no reboot of secondary node.

Secondary VSX LAGs are brought up after linkup-delay timer.

VSX Components
Active-Gateway



Active-active Layer 3 default gateway

1. VSX switches can be configured with a shared virtual IP address (VIP)and a shared virtual MAC address (VMAC) on the Layer3 VLAN interface.
2. VIP/VMAC serves as the default gateway for the access VLANs.
3. The first VSX device that receives traffic from the access layer (based on LAG hash on access switch) will route it across to the L3 domain.
4. No need for VRRP and it is mutually exclusive.
5. Like VRRP, routed traffic from the VSX node is sourced from the switch interface MAC (and not the VMAC).
6. Each active-gateway sends a periodic broadcast hello packet to avoid VMAC aging on the access switches.
7. Up to 4K SVIs with 1x IPv4 + 1x IPv6 active-gateways can be configured. **10.2 UPDATE**
8. VIP and VMAC must be the same on both VSX switches.
9. Up to 16 different VMACs per VSX pair, not shareable between IPv4 and IPv6. Example: max 8 VMACs for IPv4 simultaneously with max 8 VMACs for IPv6.
10. Key Differences between VRRP and Active GW:

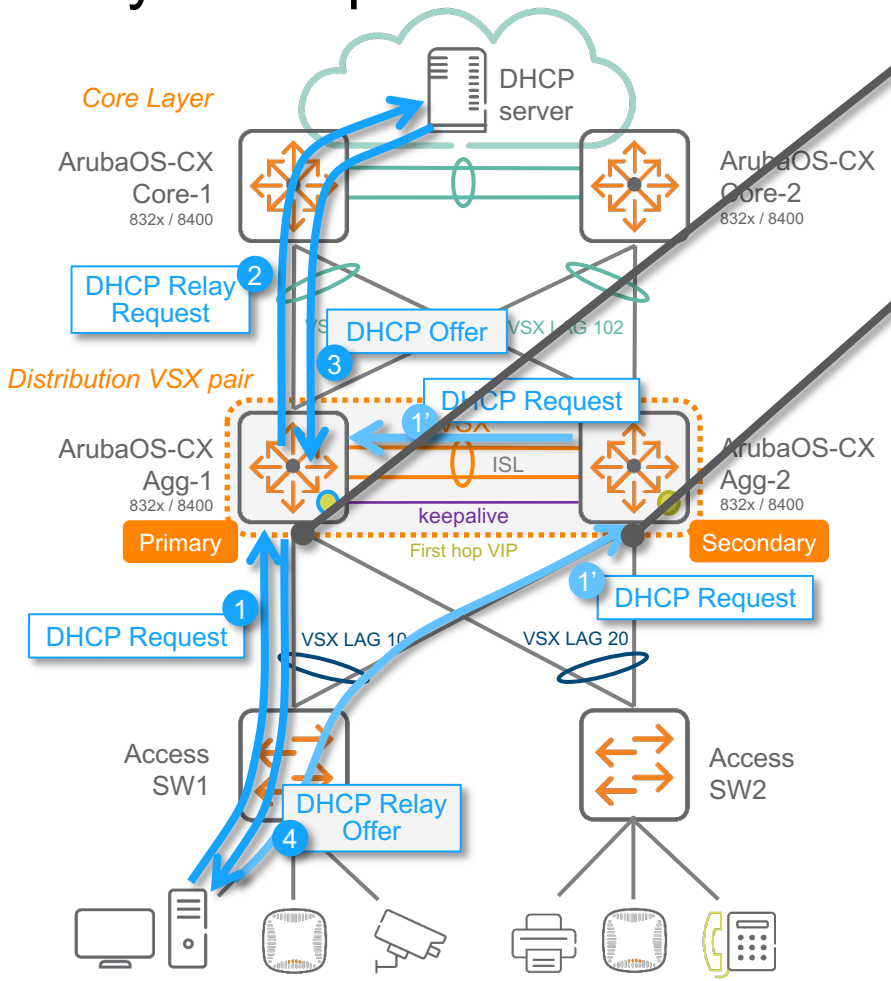
    a) *Configuration*:
    - Active-gateway is a single line configuration of Virtual IP/MAC configuration on an interface VLAN.
    - VRRP requires VRIDs configuration per VLAN, roles per VRID, VRRP advertisement timer configuration, etc…
    - Active-gateway does not support secondary IP address.

    b) *Data plane*:
    - VRRP is an active-standby data plane and only the VRRP primary device is forwarding traffic.
    - With Active GW, both devices are forwarding traffic.

aruba
a Hewlett Packard Enterprise company

Active-Gateway

DHCP relay backup



Outbound traffic

```
551 8272.589186 15.136.40.60    10.0.111.50                DHCP  358 DHCP Offer    - Transaction ID 0x23c0203b
552 8272.791318 HewlettP_68:64:d2 Vmware_8e:61:91          ARP    60 Who has 10.0.111.50? Tell 10.0.111.2
553 8273.794010 HewlettP_68:64:d2 Vmware_8e:61:91          ARP    60 Who has 10.0.111.50? Tell 10.0.111.2
554 8274.796564 HewlettP_68:64:d2 Vmware_8e:61:91          ARP    60 Who has 10.0.111.50? Tell 10.0.111.2
555 8275.602482 15.136.40.60    10.0.111.50                DHCP  358 DHCP ACK      - Transaction ID 0x23c0203b
556 8275.729760 00:00:00_00:11:01 Vmware_8e:61:91          ARP    60 10.0.111.1 is at 00:00:00:00:11:01
557 8275.729761 00:00:00_00:11:01 Vmware_8e:61:91          ARP    60 10.0.111.1 is at 00:00:00:00:11:01
558 8276.135769 HewlettP_68:1e:c2 Vmware_8e:61:91          ARP    60 Who has 10.0.111.50? Tell 10.0.111.4
559 8280.735560 HewlettP_68:64:d2 Broadcast                ARP    60 Who has 10.0.111.50? Tell 10.0.111.2
```

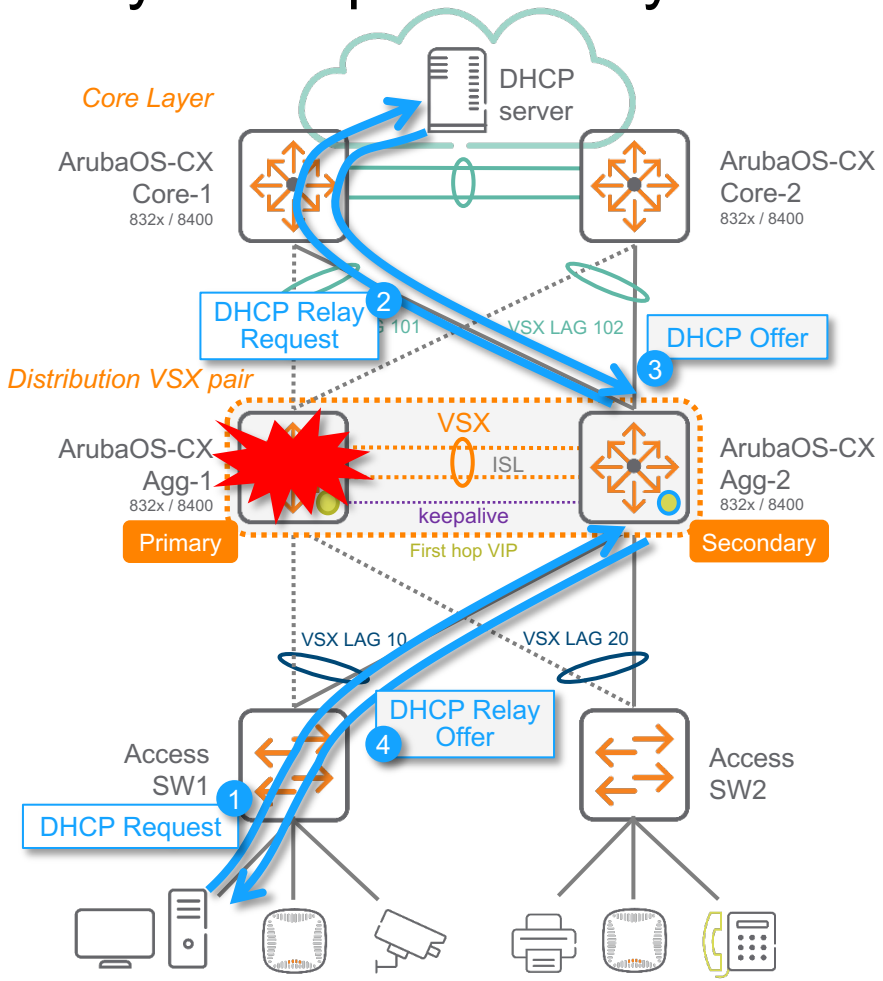
```
483 8242.164143 HewlettP_68:44:38 LLDP_Multicast           LLDP  113 TTL = 120 System Name = 8320-2 System Description = Aruba JL479A  TL.10.02.0001AK
484 8246.845066 00:00:00_00:11:01 Vmware_8e:61:91          ARP    60 10.0.111.1 is at 00:00:00:00:11:01
485 8246.845068 00:00:00_00:11:01 Vmware_8e:61:91          ARP    60 10.0.111.1 is at 00:00:00:00:11:01
486 8246.845261 10.0.111.3      10.0.111.50                ICMP  109 Destination unreachable (Network unreachable)
487 8246.845262 10.0.111.3      10.0.111.50                ICMP  109 Destination unreachable (Network unreachable)
488 8247.167516 HewlettP_68:0e:06 Vmware_8e:61:91          ARP    60 Who has 10.0.111.50? Tell 10.0.111.5
```

L2 link

L3 link

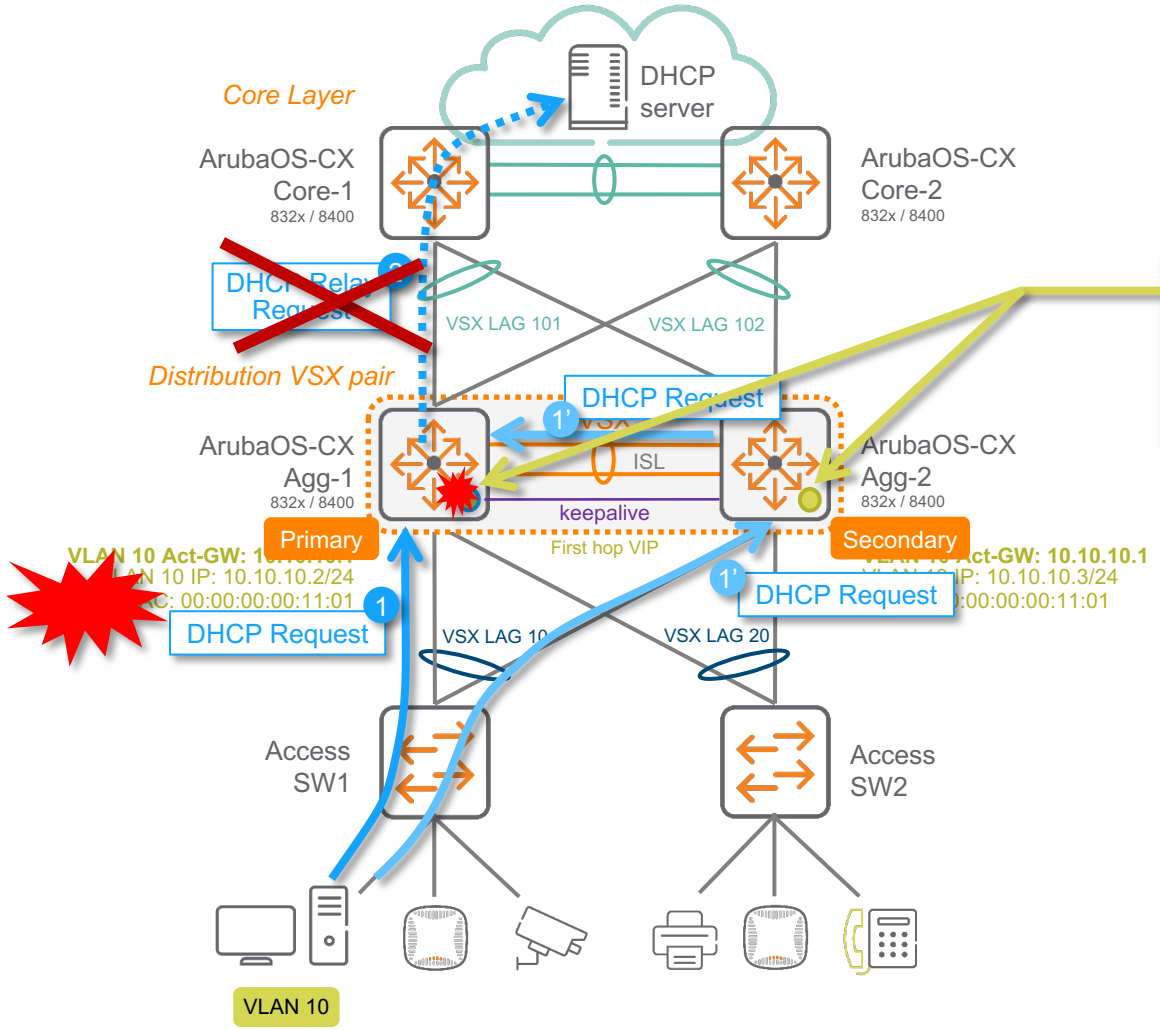
Active-Gateway

DHCP relay backup – Primary failure



Active-Gateway : DHCP relay backup

Caveat – SVI state



DHCP relay failure if SVI down on primary

1. Only primary VSX node relays DHCP request to upstream DHCP server
2. Shutting down associated SVI on Primary VSX node will prevent any DHCP requests to be relayed.

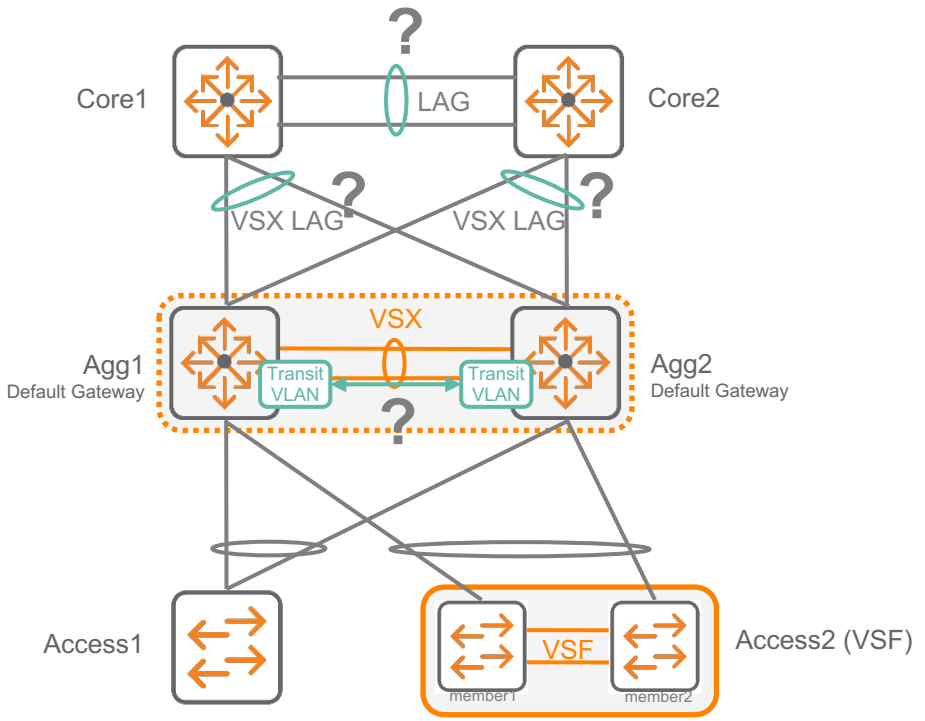
L2 link

L3 link

VSX LAG and upstream unicast routing

Constraints Diversity

- L2 or L3 links with upstream core nodes ?
- Static, OSPF, BGP ?
- Single VRF / Multiple VRF
- Sizing / limitations
- Best practice for HA



- In all scenarios, **both VSX switches run independent control planes** (separate OSPF/BGP processes) and present themselves as different routers with their own Router_IDs in the network.

- In the **data path**, they function as a single router and support **active-active forwarding**.

Definition
SVI / ROP

- **SVI**:
  - A **Switched Virtual Interface** (SVI) is a logical Layer 3 interface configured per VLAN (one-to-one mapping) that perform all Layer 3 processing for packets to or from all switch ports associated with that VLAN.

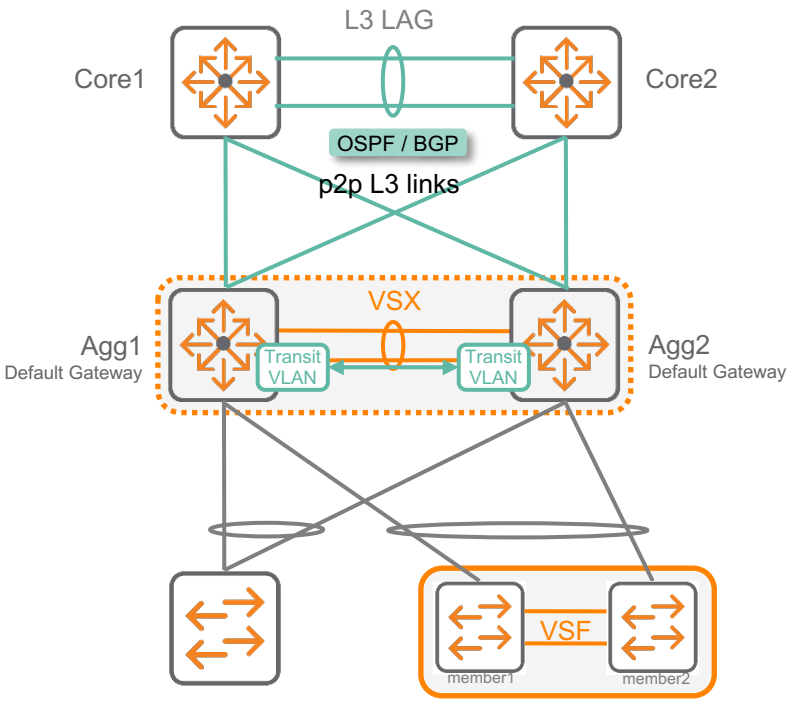
- **ROP**:
  - A **Routed Only Port** is a physical port on a switch that process all Layer 3 functions for packets to or form the said port without any binding to VLAN processing.

L2 link
L3 link

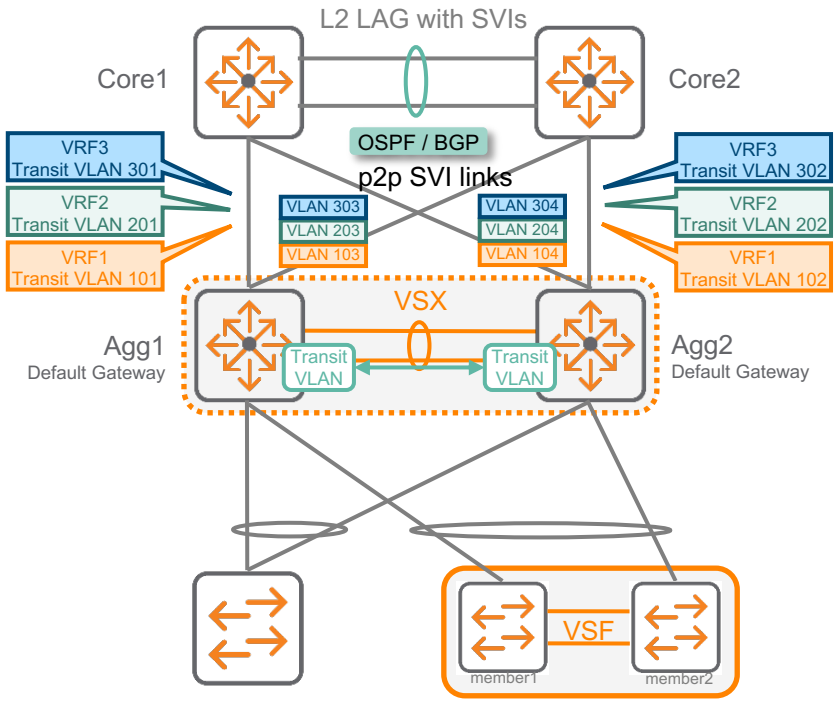
Upstream Connectivity Options

ROP, SVIs, VSX LAG SVIs

ROP (single VRF)

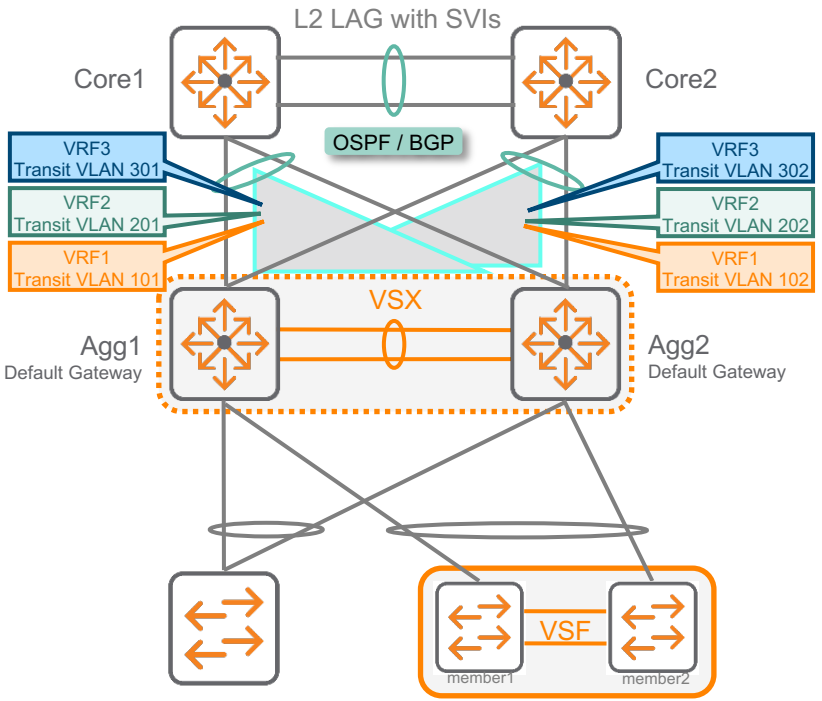


SVIs (multiple VRFs)



VSX LAG SVIs (multiple VRFs)

VSX LAG + L3 ECMP



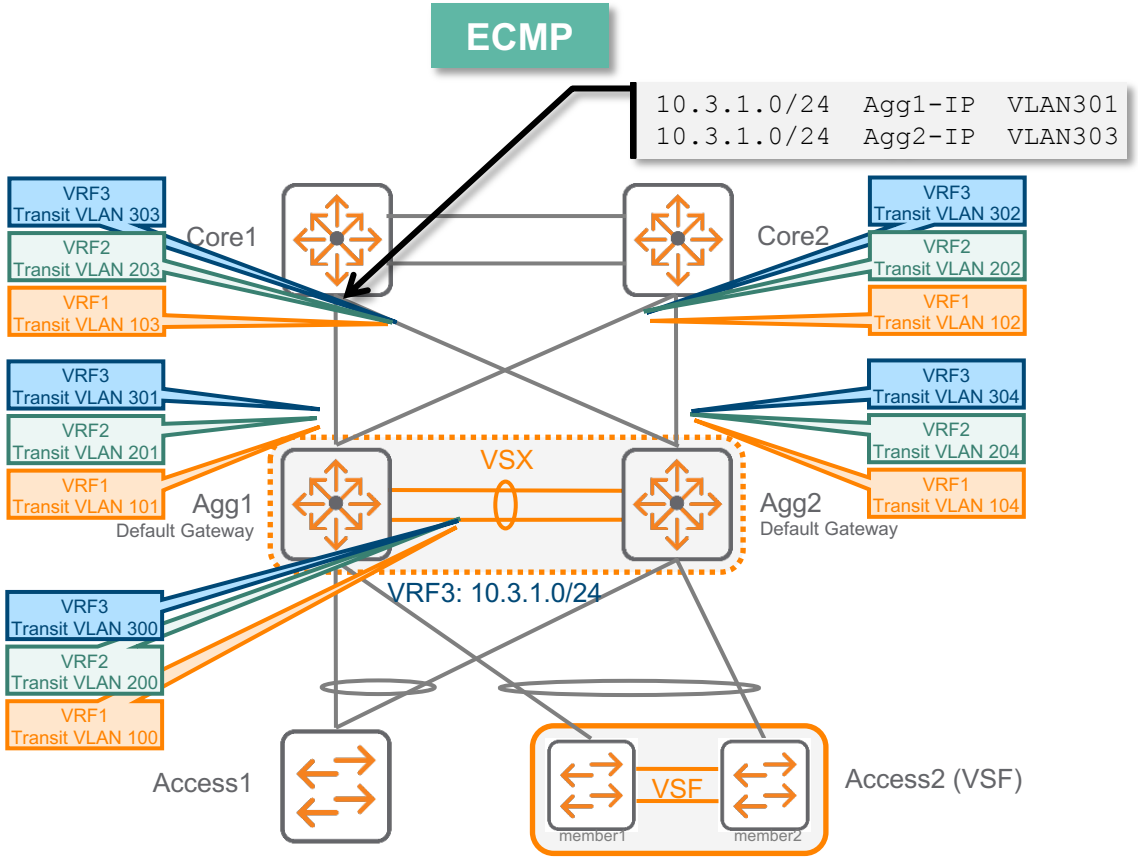
Optimize IPAM with usage of /31 for L3 point-to-point

/29 for Transit VLAN

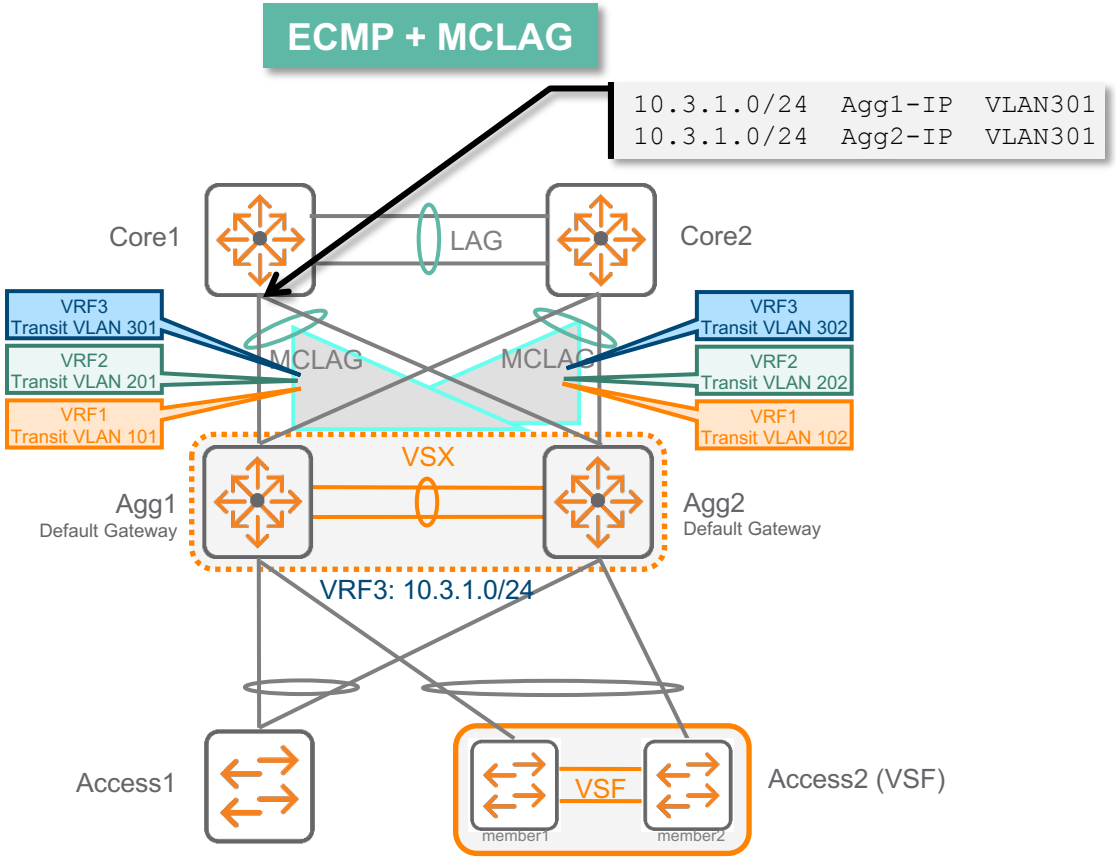
aruba
a Hewlett Packard Enterprise company

Upstream routing over VSX LAG SVI links

L3 ECMP + VSX LAG



OSPF point-to-point



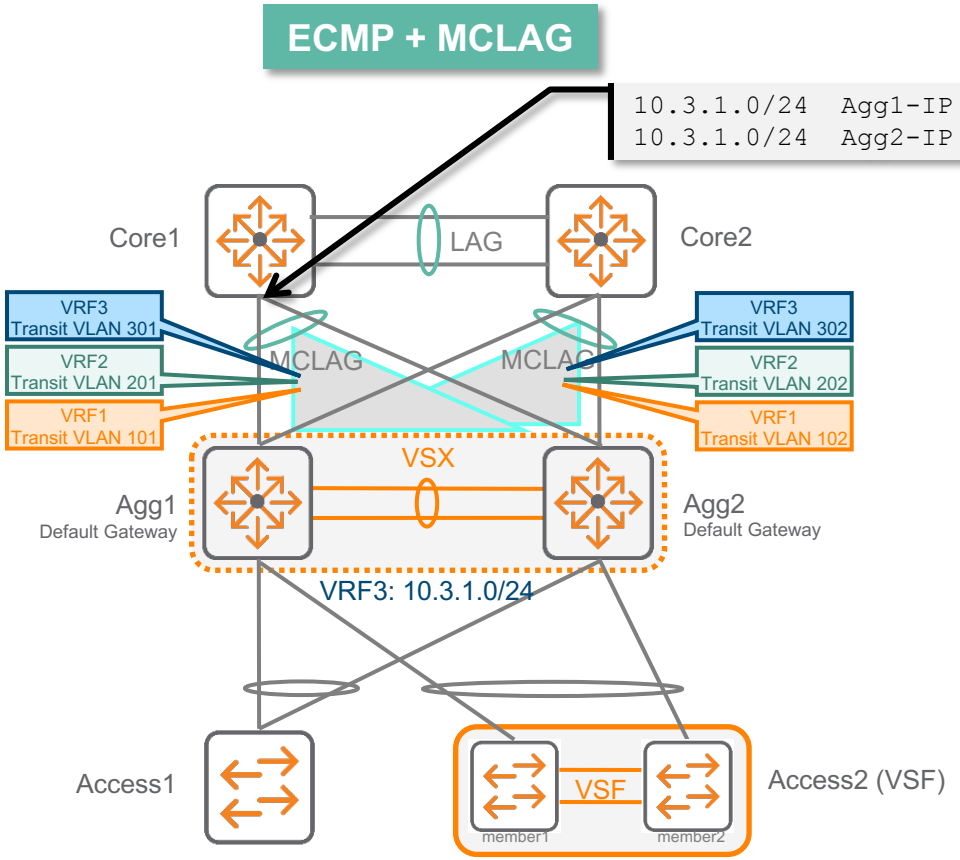
OSPF broadcast

L2 link

L3 link

Upstream routing over VSX LAG SVI links

L3 ECMP + VSX LAG

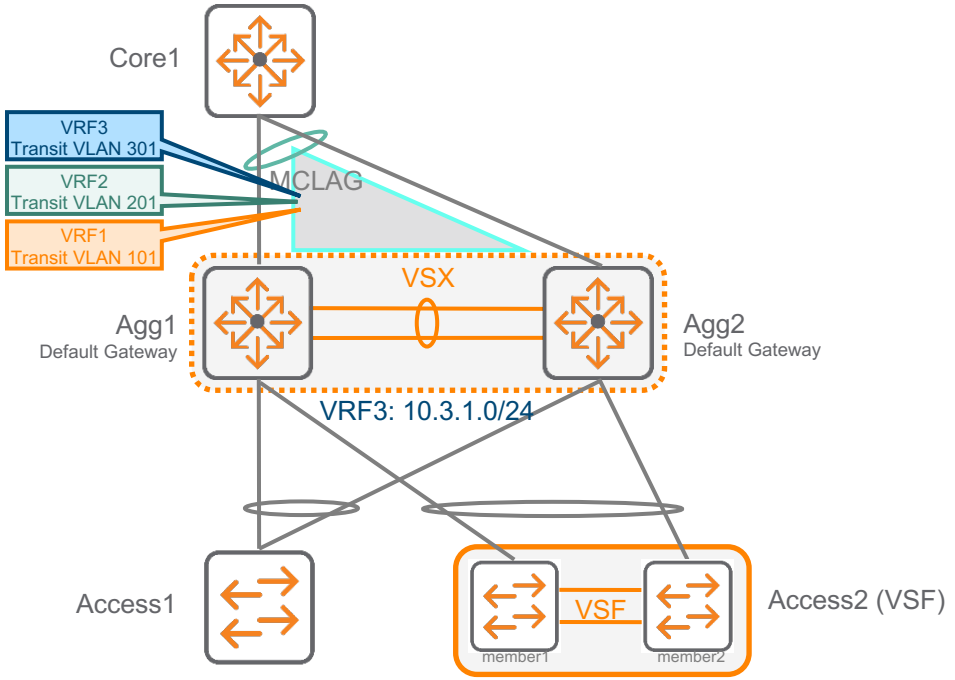


OSPF broadcast

aruba
a Hewlett Packard Enterprise company

Upstream routing over VSX LAG SVI links

L3 ECMP + VSX LAG

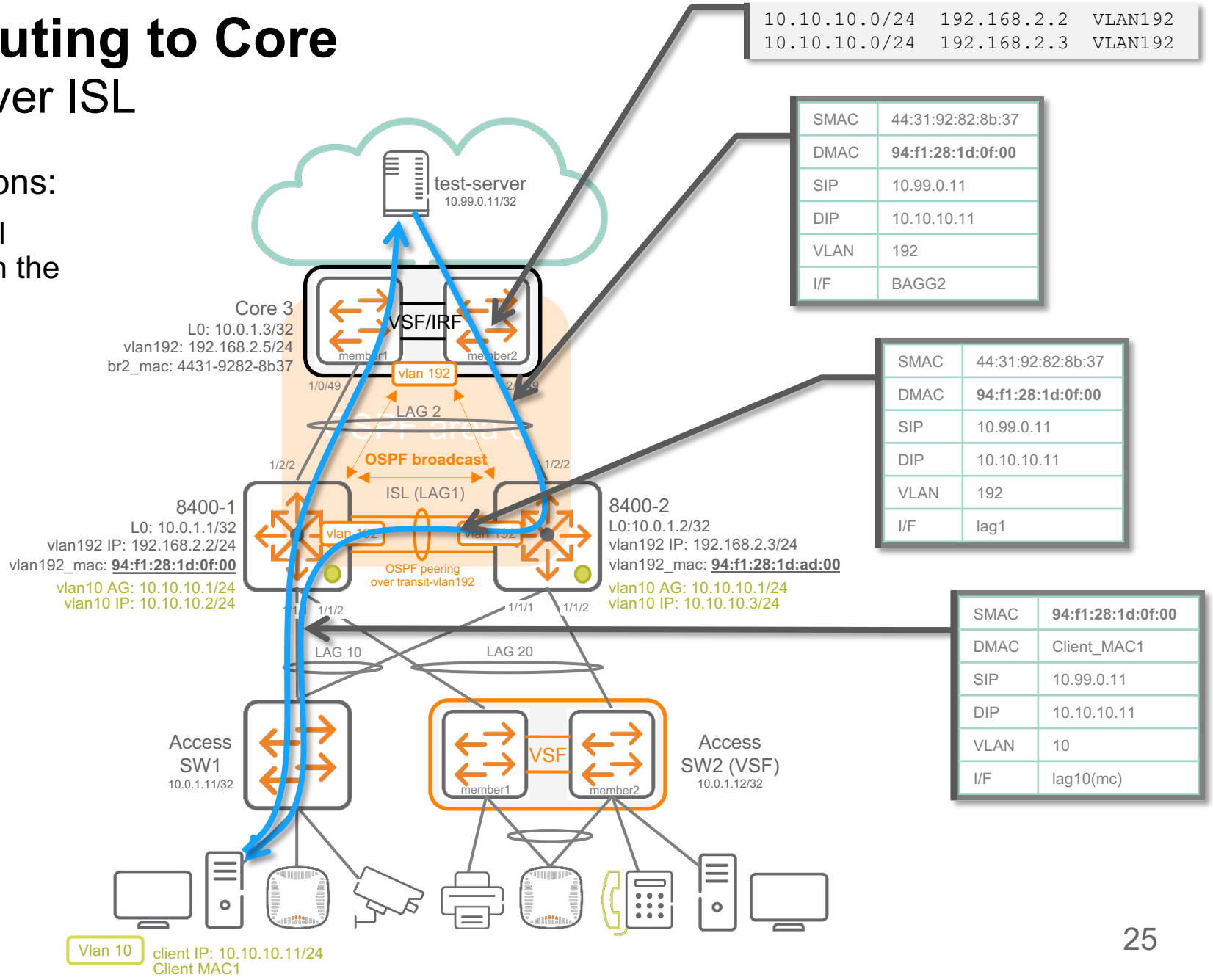


OSPF broadcast

Upstream unicast routing to Core

In nominal case: Traffic over ISL

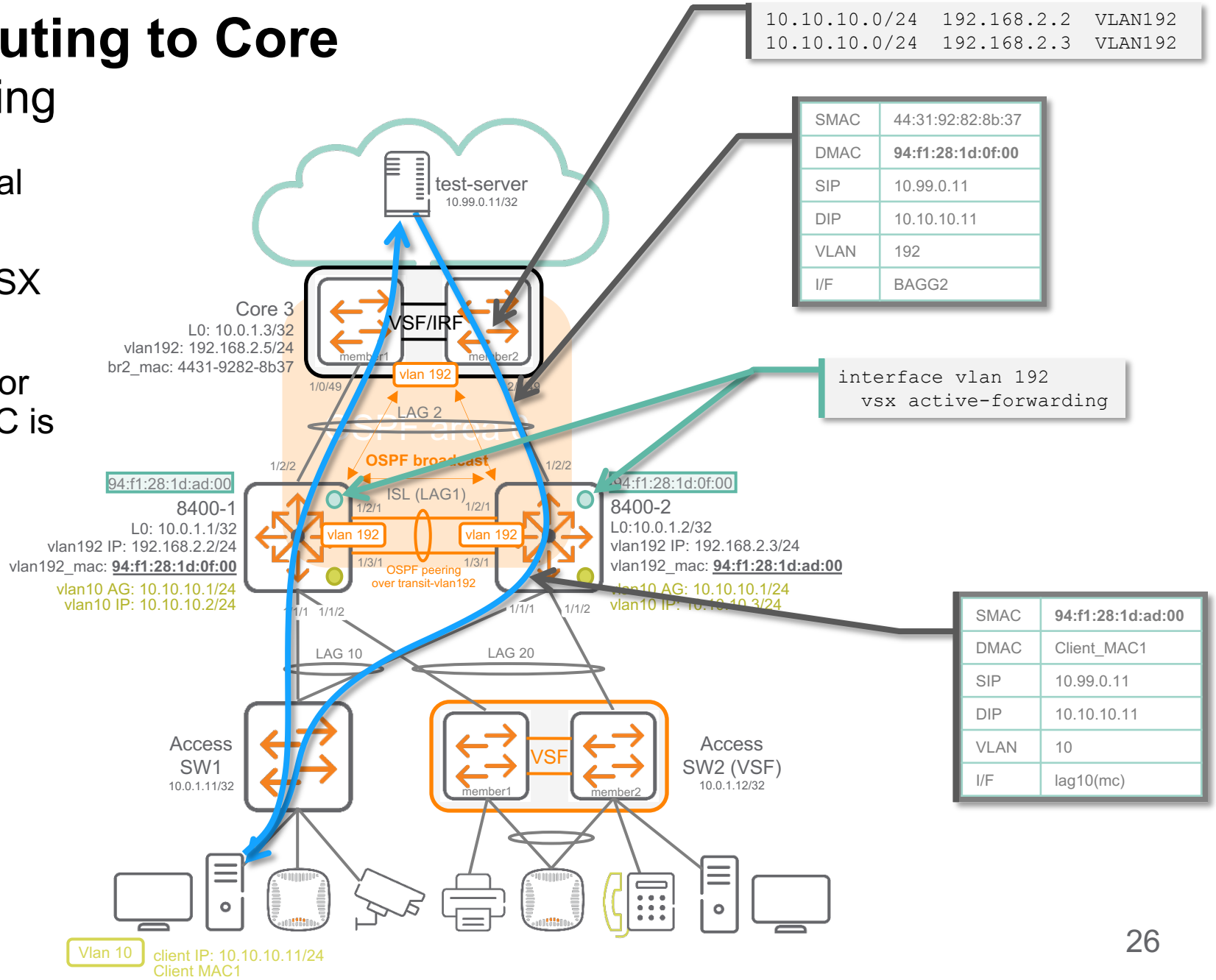
- Additional ISL sizing considerations:
  - Statistically, 50% of data traffic will cross ISL and add one L2 node on the data-path
  - ISL BW **á** uplinks BW



Upstream unicast routing to Core

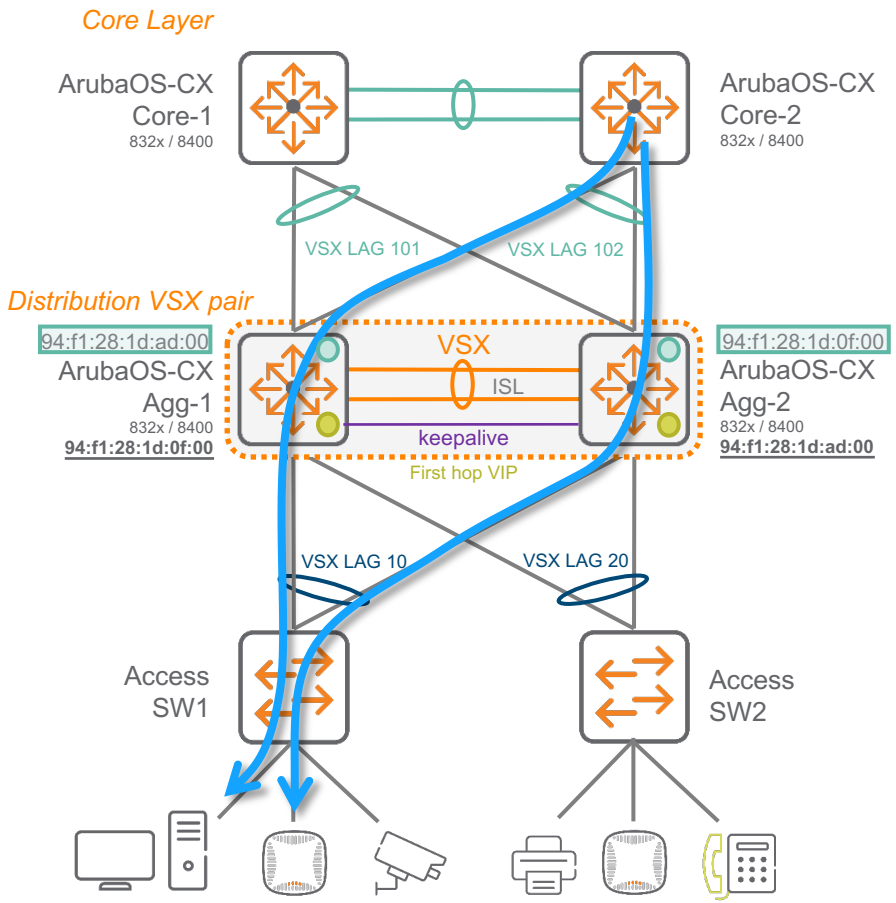
With VSX Active-Forwarding

- No data traffic over ISL in nominal case.
- Each VSX node configures its VSX peer MAC as its own MAC.
- 8400-2 will process L3 function for the received packet as the DMAC is equal the its VSX peer MAC.



VSX Active-Forwarding

North-South unicast traffic

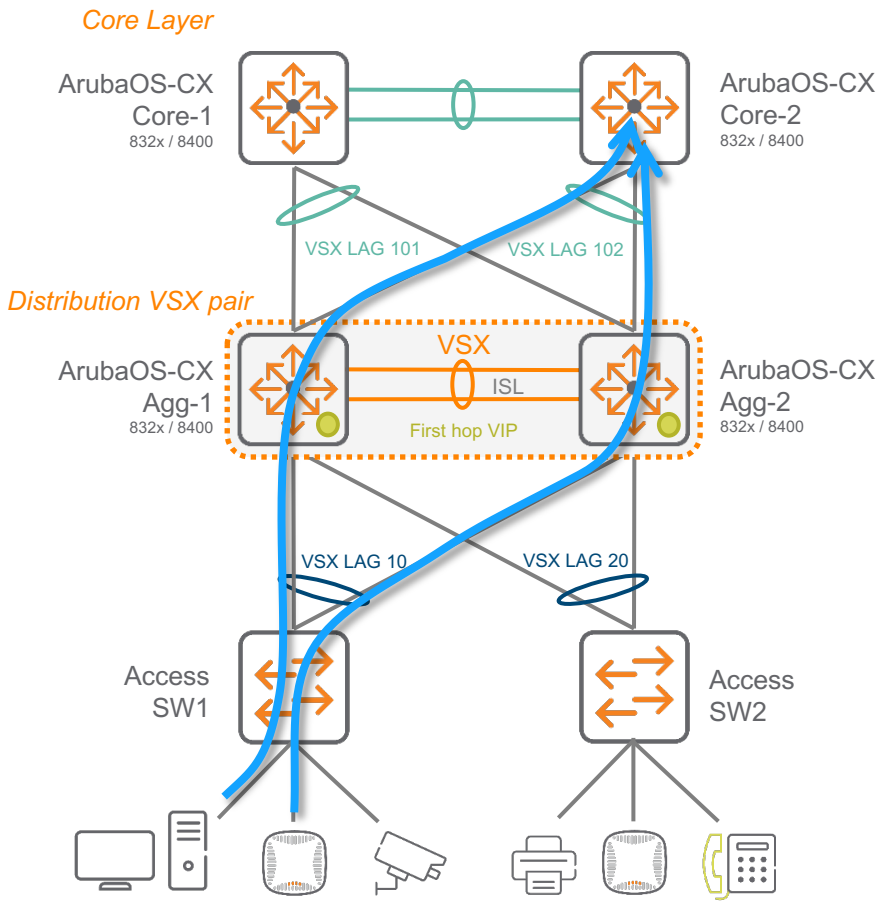


In nominal situation: **no traffic over ISL** thanks to:

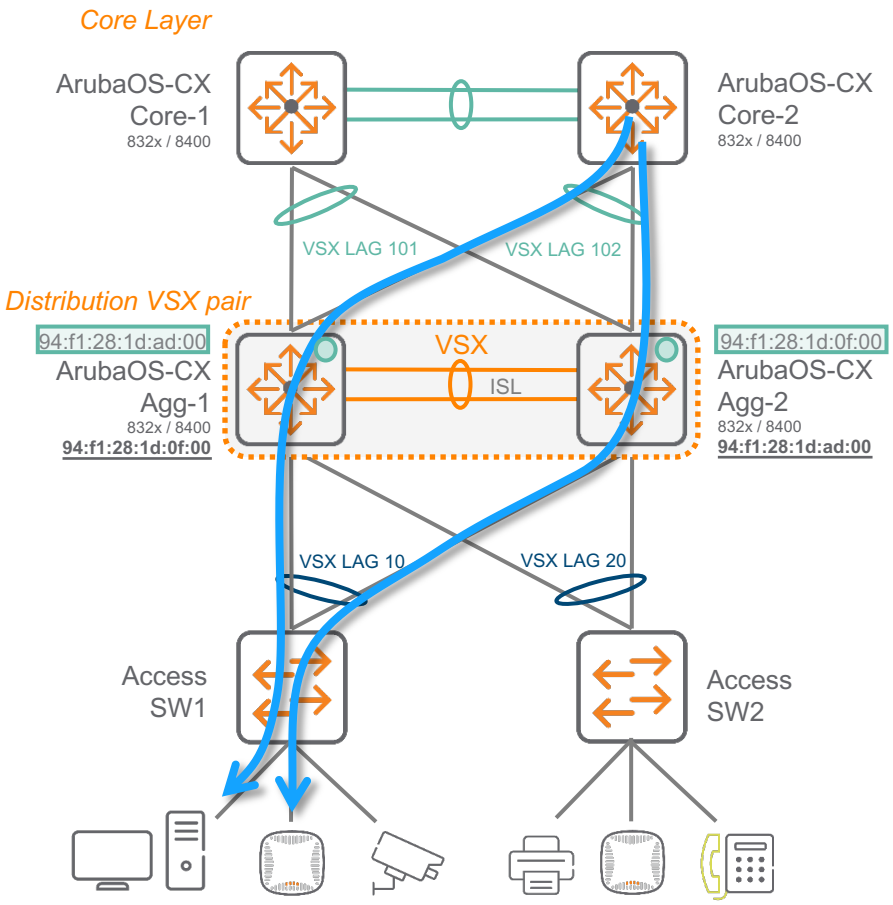
- VSX LAG local link optimized usage
- Active-forwarding

North/South unicast traffic: ISL off-load

Summary: Active-gateway and Active-Forwarding



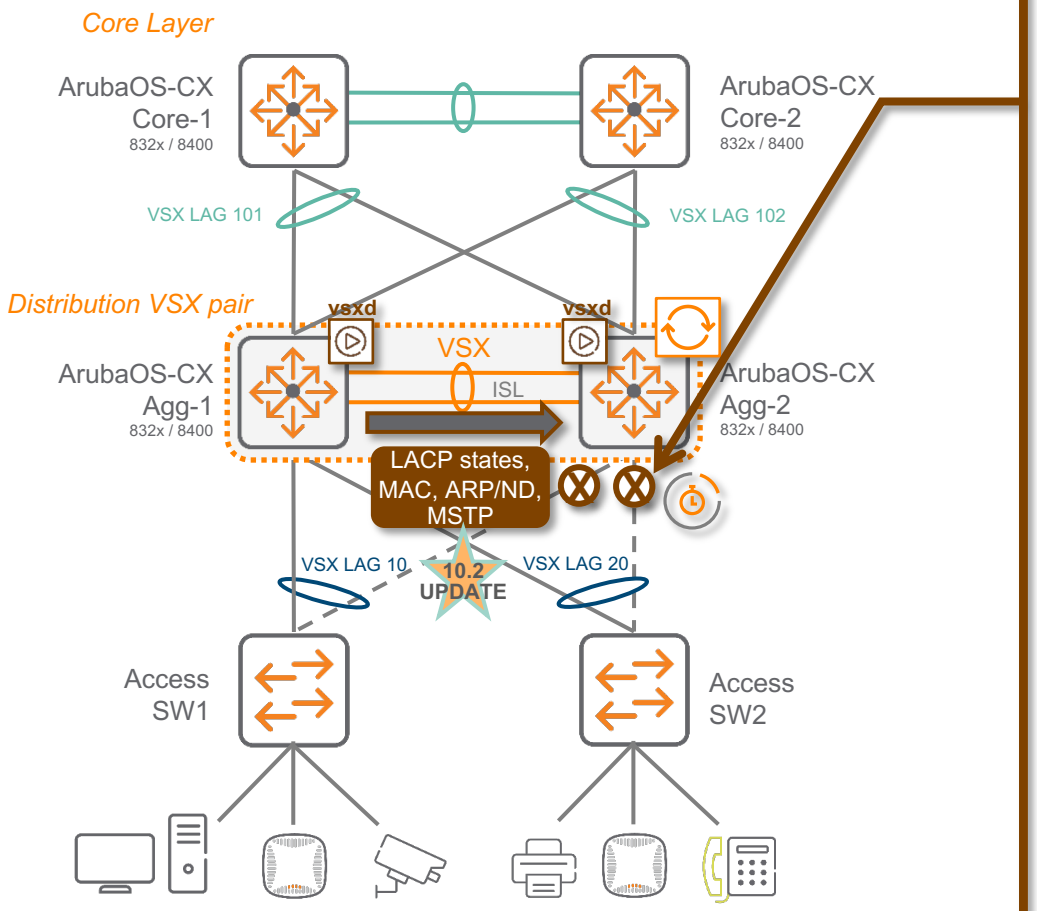
Active-gateway



Active-forwarding

aruba
a Hewlett Packard Enterprise company

VSX Components
Initial Sync and Linkup Delay



Linkup Delay

10.2 UPDATE

1. When a VSX device is rebooted, it has no entries for MAC, ARP, routes. If downstream VSX LAG ports are activated before all these information are re-learnt, traffic is dropped.

2. To avoid traffic drop, VSX LAGs on the rebooted device stay down until restore of LACP, MAC, ARP, MSTP databases.

3. The learning process has 2 phases:

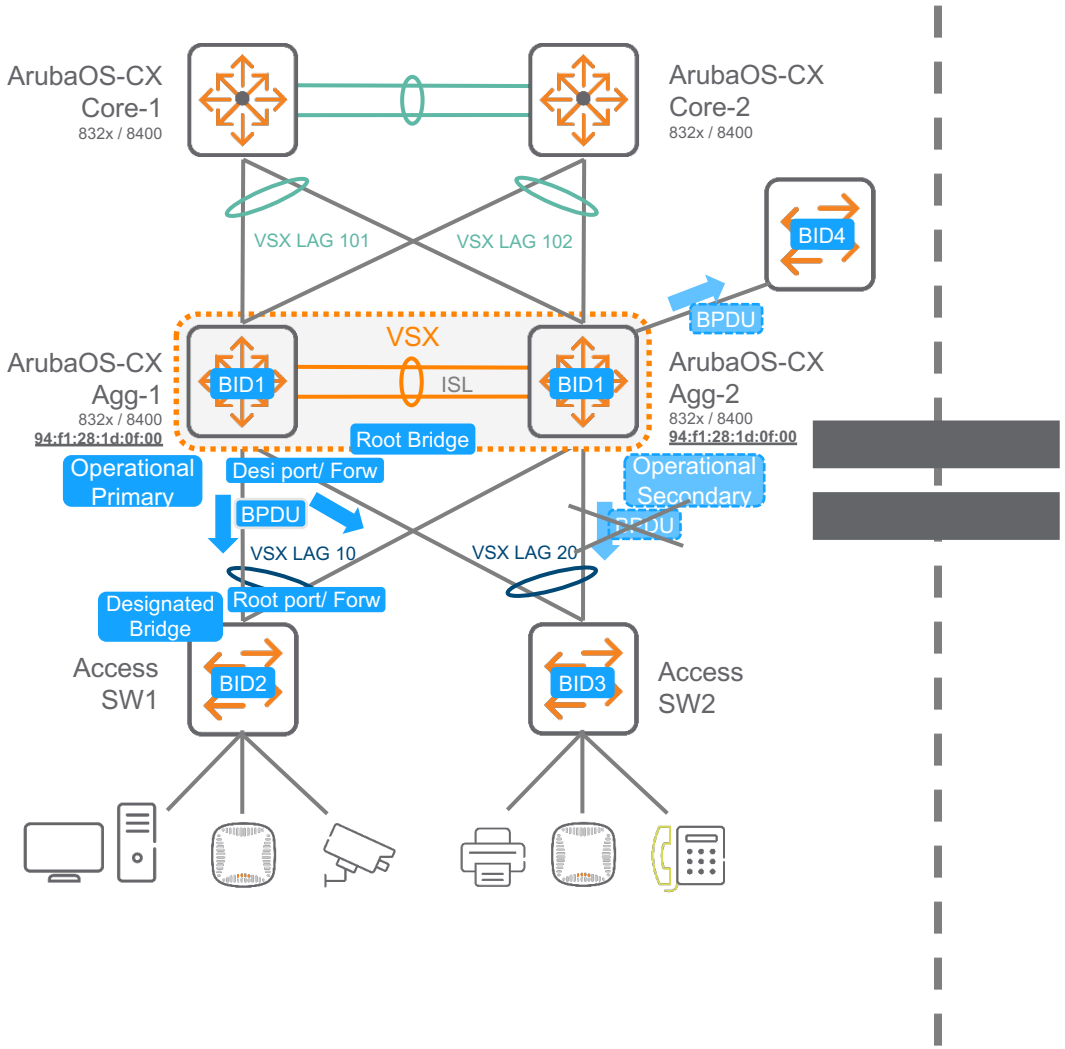
   a) Initial Sync Phase:
   - 10.2 UPDATE This is the download phase where the rebooted node learns all the **LACP+MAC+ARP+MSTP DB entries** from its VSX peer through ISLP.
   - Initial Sync timer is not configurable. It is the required time to download DB information from the peer.

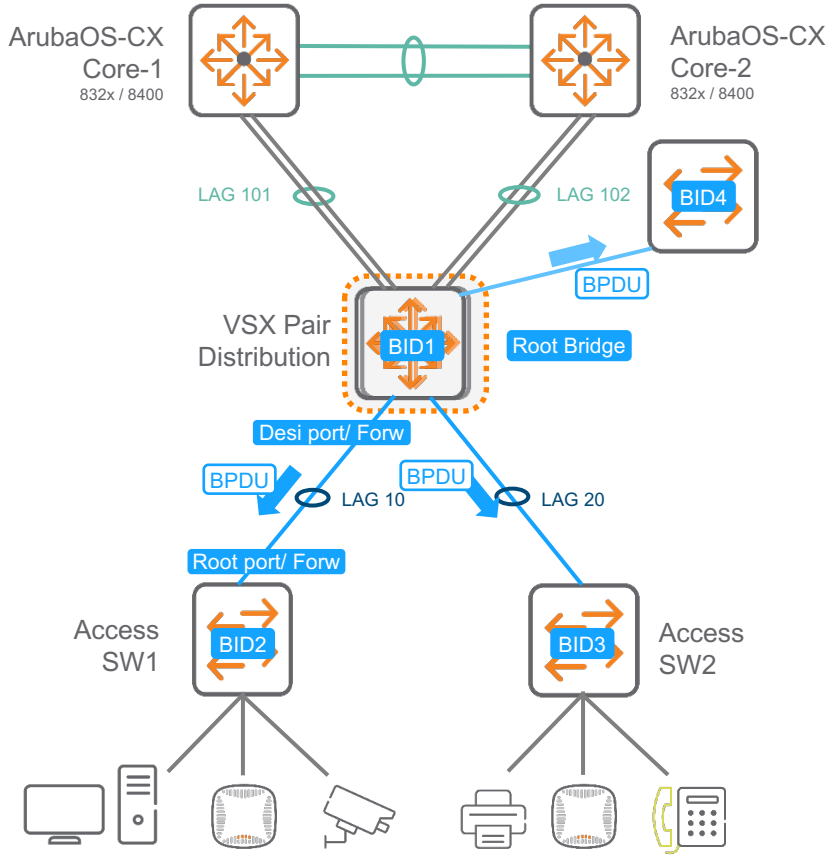
   b) Linkup Delay Phase:
   - This is the duration for:
     - installing the downloaded entries to the ASIC.
     - establishing router adjacencies with core nodes and learning upstream routes.
   - Linkup Delay timer default value is 180s. Depending on the network size, ARP / routing tables size, the timer might need to be set to higher value (max 600s).

4. When both VSX devices reboot, linkup-delay-timer is not used.

5. In order to get upstream router adjacencies established during Linkup Delay, the upstream LAG (ex: LAG 101) has to be excluded from the scope of the Linkup Delay. Until linkup delay timer, all SVIs that VSX LAGs are a member of are kept in a pseudo-shut state.

VSX + MSTP
MST0





L2 link
L3 link

Thank you!