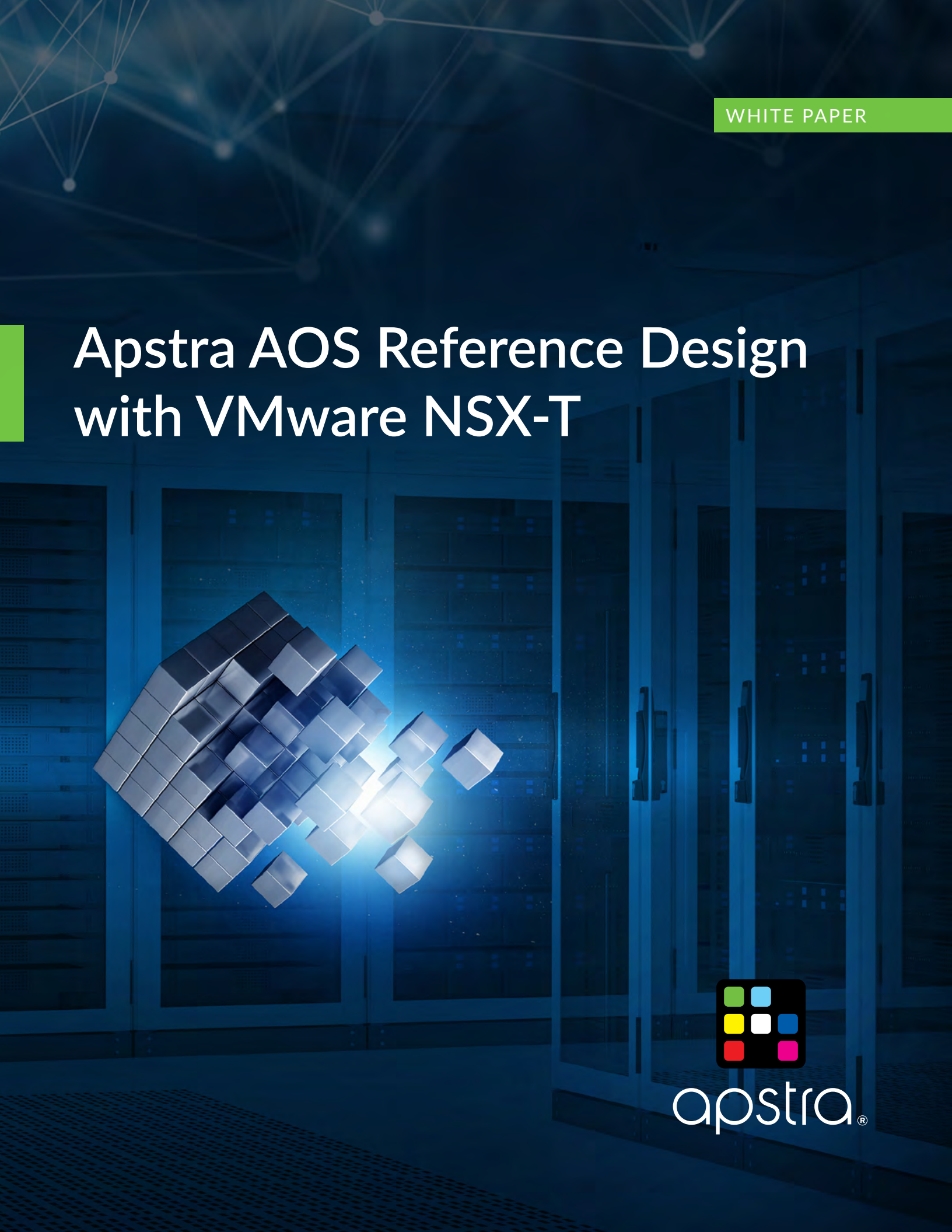
WHITE PAPER

Apstra AOS Reference Design with VMware NSX-T



apstra®

Table of Contents

Introduction

Apstra AOS automates the complex task of designing, building and operating a datacenter network through a revolutionary combination of technologies. AOS provides complete lifecycle management from first boot (ZTP) to simple device OS upgrades. Everything in AOS can be managed from a separate automation platform (ex. vRealize) by using the REST APIs. System health and remediation can be ensured in real time with the advanced capabilities of Intent Based Analytics (IBA).

Unique to the AOS/NSX-T integration is the ability to use any network vendor equipment for the physical network. Several vendors provide reference architectures for their own technology, however only Apstra offers the flexibility to select from all of the major hardware providers, with complete abstraction from the complexities of the CLI and proprietary operational challenges. This hardware-agnostic approach translates to substantial decreases in both CAPEX and OPEX.

VMware NSX-T Use Case

Virtually all modern enterprises use VMware for their virtualization and application hosting needs. ESX is the dominant hypervisor for businesses that desire high availability, maximum compatibility with open hardware platforms, and compute feature richness. VMware NSX is the market leader in server based network virtualization. NSX provides virtual network transport for virtual workloads as well as bare metal network provisioning for specific hardware vendors through the automated management of a hardware virtual tunnel endpoint (VTEP).

Because of this, AOS naturally complements vSphere and NSX, as AOS can rapidly instantiate and manage the physical IP fabric needed by vSphere to connect hosts to resources. AOS simplifies the creation and operation of the physical fabric and augments all of the capabilities within vSphere and NSX-T by abstracting the physical network into a single logical entity that is managed from a single point.

AOS MANAGES VSPHERE PODS AS AOS "BLUEPRINTS".

ESX hosts are treated as L2 connected servers in an AOS blueprint. AOS can manage extremely large blueprints, supporting more than 10,000 connected servers. These ESX servers should adhere to VMware's best practices for sizing the Datacenter entity, HA/DRS clusters, vSAN groups, etc.

NSX-T Design Variants

NSX-T offers architects a broad selection of network services and feature richness to satisfy the vast majority of requirements. While every implementation is typically unique, the following groupings cover the major variations in designs:

- **NO NSX-T OVERLAYS**
  Pure VLAN backed logical switches but making use of NSX-T firewall and other services

  Requirements

  - Full L3 cross fabric connectivity
  - IP routing for underlay SVI
  - Optimization and policy based management of external router peering links

- **NSX-T OVERLAYS ON NON-EVPN UNDERLAY**
  Network designs without the need for double encap and reducing the overall complexity

  Requirements

  - All previously listed requirements
  - Virtual networking supporting bare metal port membership
  - Double encapsulation
  - Route isolation with VRFs
  - Complex multitenancy designs

- **NSX-T OVERLAYS ON EVPN UNDERLAY**
  Maximum flexibility and most feature richness

  Requirements

  - All previously listed requirements
  - Virtual networking supporting bare metal port membership
  - Double encapsulation
  - Route isolation with VRFs
  - Complex multitenancy designs

AOS can simplify the design, build, deploy and operation phases of all of these variations, and with an easy template design workflow, can instantiate a physical IP fabric to satisfy these needs in minutes. For example, enabling EVPN in the underlay network is a simple checkbox during the guided network design phase, with full customization of the control plane available via operational workflows.

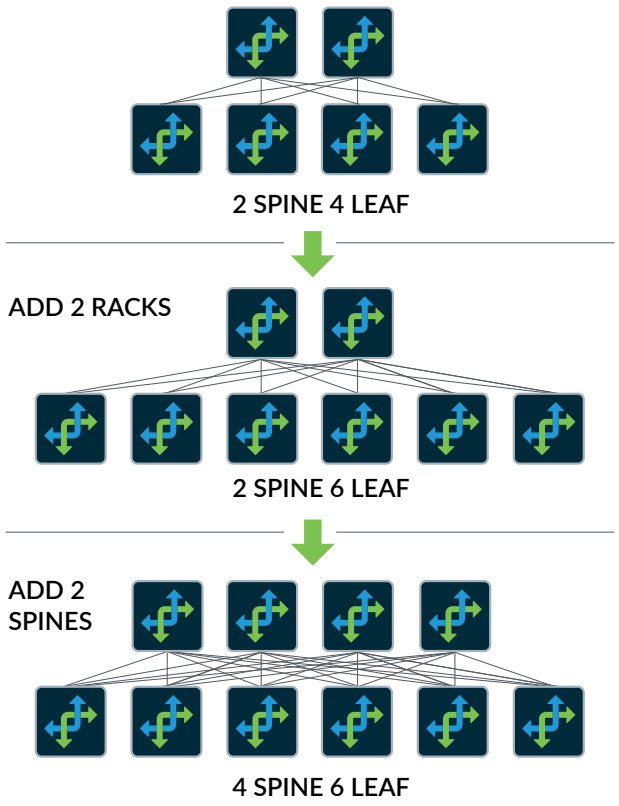
Physical Network Reference Design

Leaf-Spine Topology

AOS provides reference designs for a standard Spine-Leaf network topology that can be customized for size and bandwidth requirements. Spine-Leaf networks have proven to be the most popular and reliable topology and can be nested to create extremely large systems. Spine-Leaf networks in the data center began to replace a three-tier topology based on application evolution. In the past, most traffic was from external clients over the WAN or internet to a server in the data center (north-south). Now, as applications have evolved, most traffic is east-west. The oversubscription and extra hops of a three tier architecture were no longer suitable based on application traffic patterns. Leaf-Spine designs emerged, and proved uniquely suitable for modern application architectures.

In a Leaf-Spine system, each leaf is connected to a number of different spines with all links ideally running the same bandwidth. Ideally, the only thing that connects to spines is leaf switches. Leafs should not connect to other leafs (except in the case of MLAG peer links), and spines should not connect to other spines. This design, combined with eBGP as a routing protocol is well documented in RFC 7938 (https://trac.tools.ietf.org/html/rfc7938). This design provides multiple paths for redundancy levels based on the specific business needs. Equal bandwidth links enable a deterministic distribution of load across the spines. When new racks or servers are deployed onto the network, new leafs or top of rack switches (TORs) can be added without impacting the live network performance. When additional bandwidth is needed between racks or outbound from the pod, spines can be easily inserted without impacting the network. Operationally, the workflows to perform these changes are simple and can be performed during normal business hours as the tasks are non-intrusive.

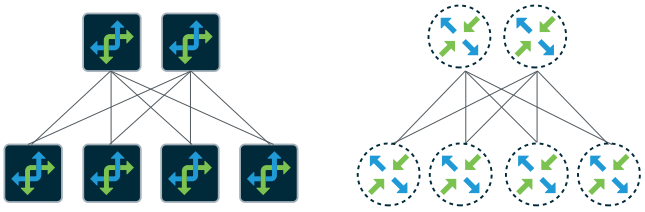
AOS can manage multiple Leaf-Spine fabrics from a single AOS server. Both 3 stage and 5 stage fabrics are supported, either one can be selected based on the sizing and scaling needs of the deployment.



ADDING CAPACITY TO A LEAF-SPINE SYSTEM IS SIMPLE AND DETERMINISTIC.

Layer-3 IP Routing

AOS uses BGP to create a routing domain for each deployed blueprint, leveraging eBGP specifically. These blueprints are connected to one or more core or upstream BGP routers, allowing multiple pods to communicate. AOS automates the configuration of the L3 routing protocol to ensure high levels of availability and redundancy within the spine.



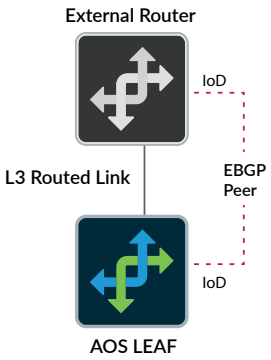
PHYSICAL L3 FABRIC

UNDERLAY EBGP DEFAULT VRF

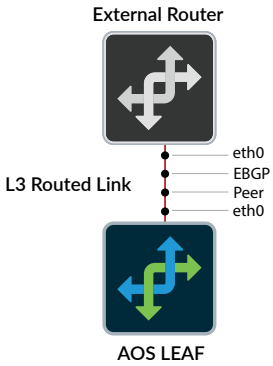
External Routing Connectivity Points

AOS helps network architects connect server environments to the rest of the network through defined Connectivity Points. These are ports on leaf/TOR switches that connect to a non-AOS managed device (or set of devices). AOS offers the following external connectivity options:

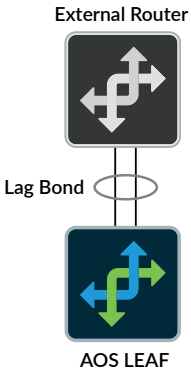
- **L3 ROUTED PORTS W/ LOOPBACK PEERING**
  By default, AOS expects to peer with the external routers by establishing an EBGP session with the loopback of each external router. This is accomplished by AOS automatically adding ebgp-multihop commands and an associated static route to the external router loopback with the next hop set to the address of the connected interface of the external router. This address is determined during the build process when you associate an IP pool with the links to the external routers. For example, if you assign 10.10.10.4/31, AOS will place 10.10.10.4/32 on the AOS managed interface and will expect 10.10.10.5/32 as the IP address on the external router. This address will be used in the static route for the loopback.



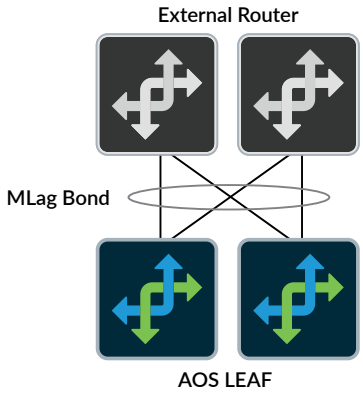
- **L3 ROUTED PORTS W/ FIRST HOP PEERING**
  AOS can also be optionally configured to peer with the local neighbor interface IP, this can be useful for custom configurations needed by firewall or load balancer VIPs/VRRP.



- **L2 SWITCHED PORTS W/ LAGS**
  Multiple L2 connections to a single external router can be bonded together for redundancy and bandwidth purposes. In fact, these links can connect to an MLAG pair as the AOS managed leaf does not know if it is connected to one switch or two.



- **L2 SWITCHED PORTS W/ MLAG**
  If two AOS managed leaf switches are already part of an MLAG pair, they can peer with one external router or even a pair of external routers in an MLAG pair. This is typically used when connecting two AOS leafs to a firewall pair. In this situation the AOS leafs will have a static route to the VIP provided on the L2 segment by the firewalls. MLAG bonds can only be created from racks that contain 2 leaf switches that support MLAG peer links.



External connectivity and routing can be customized with additional options:

- Per-VRF external connectivity options
- Per-peer route filtering, summarization, and route injections

PHYSICAL NETWORK ASSURANCE

AOS is packaged with several prebuilt IBA probes that perform complex system checks related to fabric health and network services. Examples of elements that can be validated include:

- MLAG imbalances
- ECMP imbalances
- Sustained packet loss
- East-West bandwidth utilization
- MTU consistency

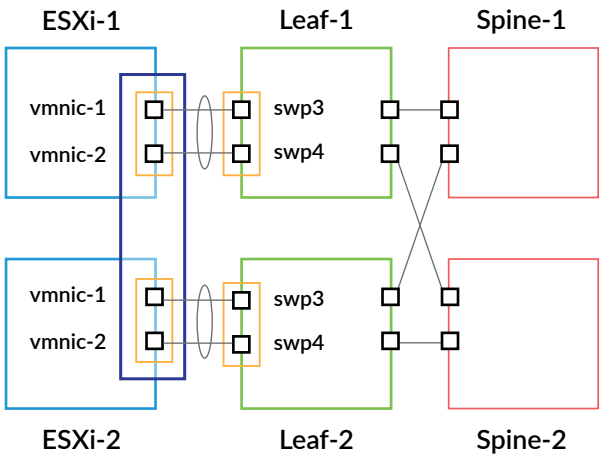
These probes can be instantiated in AOS through the “Deploy Prebuilt Probes” workflow.

ESX Host Connectivity Options

AOS supports a number of different server connectivity options for Layer 2 connected hosts. These systems use Ethernet, Spanning Tree Protocol (STP), autonegotiation, MLAG/vPC and 802.1q tags to participate in the network on a local VLAN or VXLAN at the switch port level. Layer 2 hosts are not directly managed by AOS so they appear gray in the topology views.

AOS supports multiple ethernet bonded links to support the ESX connectivity requirements. The specific host NIC configuration is highly dependent on the following:

- Application redundancy requirements
- Availability of switch maintenance windows
- Cost
- Cabling limitations
- VLAN and MTU consistency with the physical network



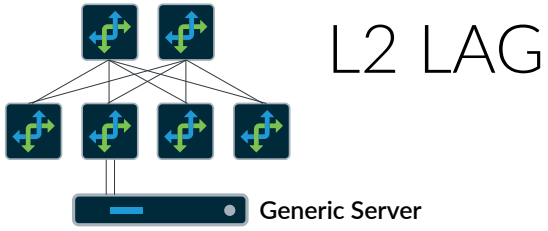
AOS supports traffic draining and taking leaf switches in and out of service. This enables the operator to gracefully remove a switch from service, perform maintenance like upgrading the NOS levels, and then selectively restore traffic once the system has been fully validated. To ensure the host and VMs still have network connectivity during these workflows, Apstra highly recommends the use of multiple TORs in an MLAG configuration for redundancy.

ESX Host Connectivity Examples

Each business needs to determine the exact NIC/switch configuration that best addresses their needs. Provided are two examples on either end of the spectrum for consideration:

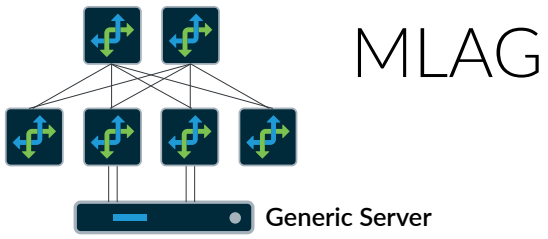
OPTION 1 - COST CONSCIOUS, LIMITED REDUNDANCY

- One leaf switch
- Single point of failure in the switch
- All major maintenance on switch will impact all connected hosts
- ESX hosts - 2 port 10G NIC - 1 LAG bond to TOR for NIC port redundancy
- LAG carries all traffic from ESX hosts over 802.1q tags



OPTION 2 - HIGH SERVICE AVAILABILITY, NEED FOR MAINTENANCE WINDOWS

- Two TOR leaf switches for ESX VM, vSAN, vMotion traffic
- One TOR leaf switch for management vmknic (ILO, OOB management)
- TORs participate in an MLAG group for redundancy
- ESX hosts - 2x 2 port 40G NICs - 1 LAG bond 2x40GB to each TOR for TOR and NIC redundancy
- Switch maintenance will have limited impact on applications including loss of 50% bandwidth and the loss of 1-2 packets during MLAG ARP flow cutover



DEFAULT GATEWAYS AND VMKERNEL TCP/IP STACKS

VMware official documentation describes VMKernel networking requirements. Apstra recommends following VMware's guidance on the use of multiple TCP/IP stacks.

QOS

For virtualized environments, the hypervisor sets the QoS values for the different traffic types. The physical switching infrastructure has to trust the values set by the hypervisor. No reclassification is necessary at the server-facing port of a top of rack switch. If there is a congestion point in the physical switching infrastructure, the QoS values determine how the physical network sequences, prioritizes, or potentially drops traffic.

Apstra recommends that ESX operators adhere to the recommendations provided by VMware (https://www.vmware.com/pdf/vmware-validated-design-20-reference-architecture-guide.pdf) when configuring Quality of Service settings on both the pNICs and the physical network. QOS settings for the switches can be configured with Configlets.

ESX Host Network Configuration Assurance

The AOS Server can communicate directly with the NSX-T server to gather information about host configuration. AOS has several prebuilt IBA probes that provide real time assurance through complex system checks.

These checks include:

- **LAG Configuration Validation**
  Validates LAG settings and the presence of a working bond to the ESX hosts
- **MTU Settings**
  Checks that the underlay MTU supports jumbo frames
- **LLDP Settings**
  Checks that hosts have LLDP enabled

These probes can be instantiated in AOS through the "Deploy Prebuilt Probes" workflow.

NSX-T Overlay Egress to Underlay

Placement of Edge Clusters (North-South Data Flows)

NSX-T reference architecture recommends the connection of Edge Clusters on dedicated physical switches that have the appropriate bandwidth and redundancy to carry all of the North-South traffic flows out of the NSX overlay. In AOS, this can be managed with logical racks and the rack based template designer. Two racks within a template can be designed with multiple 100G uplinks and MLAG for high throughput and load sharing, while the other racks in the template can use smaller 40G connected switches with 10G ports facing the servers.

The NSX-T Service Router (SR) should be instantiated on the edge cluster and provided with the recommended physical resources.

There are a number of different design considerations for an NSX-T Tier-0 SR/DR deployment. The most important decisions are related to:

- L2 redundancy design for Edge Cluster Nodes (Active/Active vs. Active/Passive)
- Number and speed of NICs
- L2 vs L3 Egress
- Network and route isolation (VLAN/VXLAN/VRFs)

Edge HA

Edge nodes are sometimes deployed in active-standby on two different racks and this presents some networking requirements in terms of heartbeat traffic and GARP that happens on failover. AOS can provide VXLAN encapsulation for these workloads provided the traffic can be tagged and switched onto an 802.1q trunk. These services can also include bare-metal network devices including physical firewalls and load balancers.

Bridging of NSX-T segment to Underlay VLAN/VXLANs

For virtual networks that require direct L2 connectivity to the physical network, NSX-T can be configured to bridge traffic directly to the Leaf TOR using AOS managed VLANs/VXLANs.

Consider mapping this to a specific underlay VRF to constrain the scope of visibility of the stretched subnets.

Underlay Multitenancy

Use of VLAN vs. VXLAN in the underlay

VMware ESX supports management, storage, vmotion and overlay traffic traversing a Layer-3 network. These virtual networks can be satisfied with 4 VLANs on individual racks without use of any VXLAN/EVPN in the underlay.

VXLAN may be required for

- Inter-rack L2 between bare-metal workloads
- Bridging NSX-T logical segment to bare-metal servers
- Extending multi-tenancy/isolation into underlay (ex: advertising overlay routes into a specific EVPN VRF)
- Supporting EVPN hand-off in future versions of NSX-T (to support EVPN)

Use of VRFs in Underlay

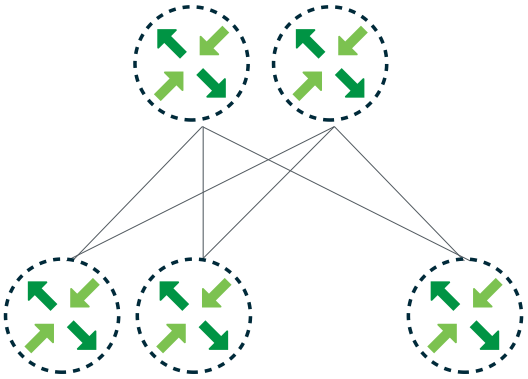
Also, for virtual network isolation for strict multi-tenancy, it is possible to map VRF-based tenancy to NSX logical separation.

It is highly recommended that NSX-T Edge Gateways be placed in a non-default VRF.

In addition, the following considerations should be taken into account:

1. Use one dedicated VRF for all NSX-T traffic to contain the scope of overlay route advertisements, instead of leaking them into the fabric
2. Use different VRFs for explicit functionality requirements:
    a. vMgmt: One for NSX-T management
    b. vInfra: One for storage, vmotion
    c. vData: One for overlay and edge transit
3. Map VRFs to one or more tenants (aka T0 routers), so that multi-tenancy is preserved and can be extended in the future for multi-site communications.

OVERLAY EVPN VRF TENENT: GREEN

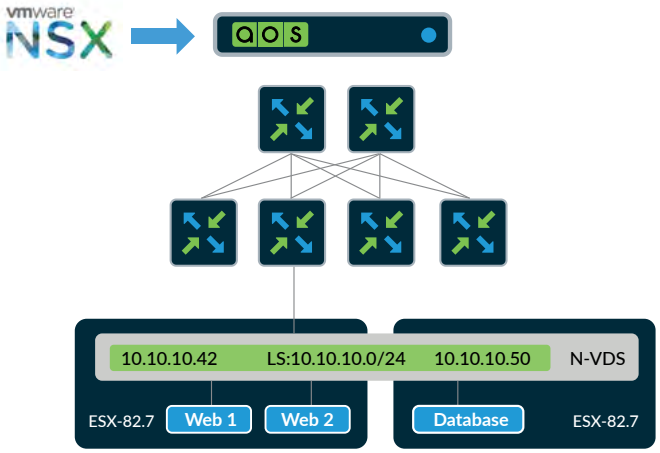


Tenant VXLAN ports active on Leafs 1,2,4

NSX-T & AOS Integration

AOS is able to connect to the NSX-T API with Read-Only access to gather information about the hosts, clusters, VMs, portgroups, vDS/N-vDS, and NICs within the NSX-T environment. This collection is done as an AOS extensible telemetry collector. The collector runs in an offbox agent which connects directly to NSX-T. On first connect, it downloads all of the necessary info and thereafter polls the controller every 60 seconds for new updates. The collector pushes info into the AOS Graph Datastore. This allows AOS to do VM query and raise alerts on physical/virtual network mismatch.

AOS needs to identify ESX/ESXi hosts on the fabric as well as the VMs connected to AOS managed leaf switches. To accomplish this, LLDP information transmitted by the ESX/ESXi hosts is used to associate host interfaces with leaf interfaces. For this feature to work, LLDP transmit has to be turned on within the dVS. A prebuilt AOS IBA probe will check to make sure LLDP is enabled on all hosts managed by the NSX-T controller.



Relevant IBA Probes for NSX-T

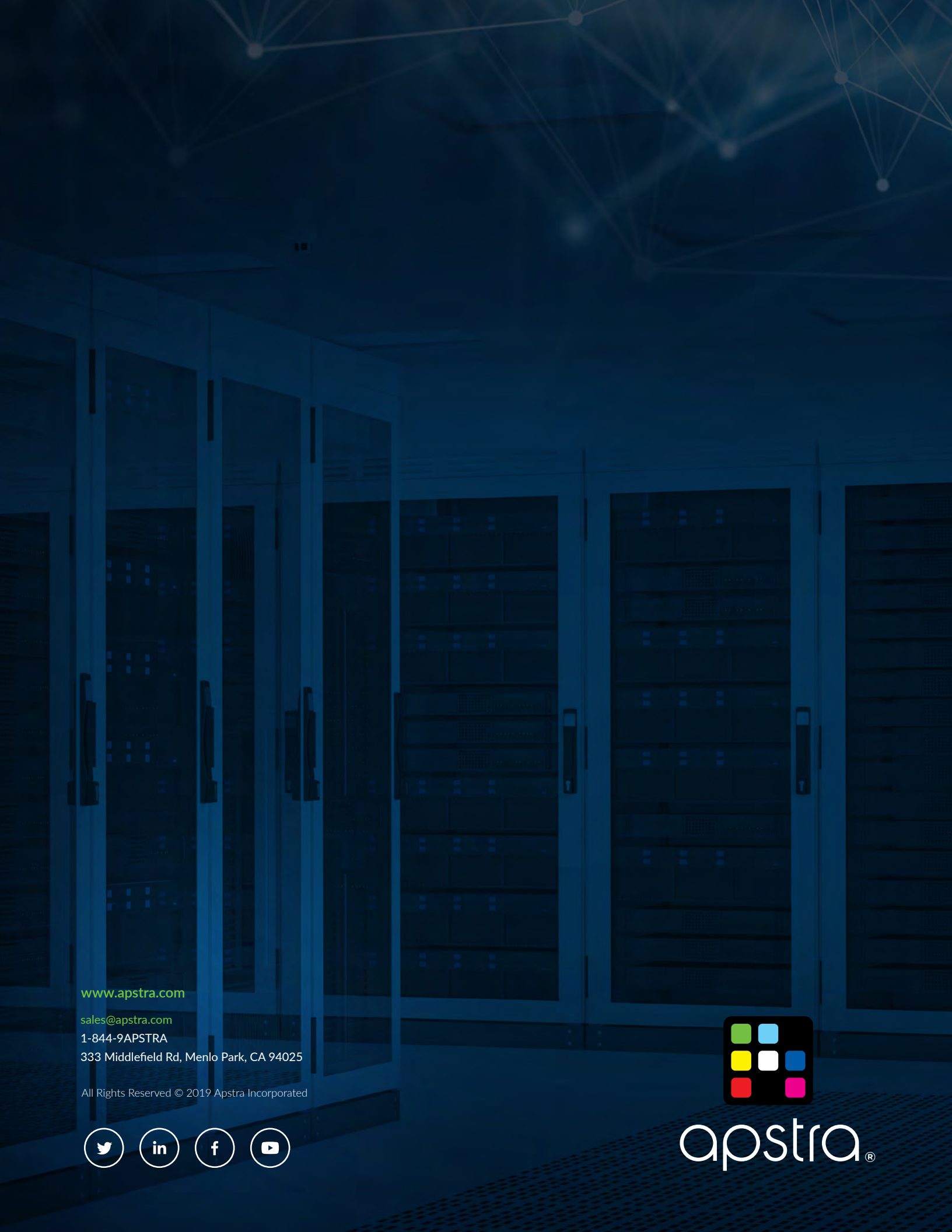
The following probes provide network assurance of consistency between the physical and logical networks:

- **Hypervisor and Fabric LAG config mismatch**
  This probe ensures that the LAG configuration for the physical network matches the LAG bond for the host pNICs.
- **Hypervisor missing LLDP config**
  This probe ensures that the hosts managed by the NSX-T controller have LLDP enabled. This is required to identify hypervisors and locate VMs within the fabric.
- **VMs without Fabric configured VLANs**
  This probe determines if there are VLANs on the portgroups that are not present on the AOS managed fabric. If the VLANs are not present, an anomaly is generated and a link to a Remediation Workflow is present in the probe view. The remediation creates a matching switch local VLAN on the fabric matching the host's requirements.

These probes can be instantiated in AOS through the "Deploy Prebuilt Probes" workflow

Summary

Apstra's AOS enables administrators to quickly deploy and operate complex datacenter networks that are optimized to support NSX-T workloads. Apstra has embedded many best practices for a VMware environment to ensure that operators can deploy and manage these networks without large investments of time and money. Every datacenter environment that is built with AOS automatically follows these practices and standards, freeing the architect from finding the latest white papers and reference guides. These reference designs can be rapidly instantiated in minutes, enabling faster time to market and lower operational costs without sacrificing reliability or performance.



www.apstra.com

sales@apstra.com
1-844-9APSTRA
333 Middlefield Rd, Menlo Park, CA 94025

All Rights Reserved © 2019 Apstra Incorporated

apstra®