v2.6 | March 2023 | BP-2050

BEST PRACTICES

Physical Networking

NUTANIX
YOUR ENTERPRISE CLOUD

Copyright

Contents

1. Executive Summary

Nutanix uses modern datacenter network technology to provide a highly available virtualization, compute, and storage platform. The network is a key component in ensuring high performance and availability, and successful Nutanix deployments combine the right physical switches with the right network designs. At the same time, certain physical switches and network designs can lead to poor performance or poor availability, and this guide can help you avoid those risks.

2. Introduction

Well-designed networks are central to Nutanix Cloud Platform resilience and performance. A Nutanix cluster can tolerate multiple simultaneous failures because it maintains a set replication factor and offers features like block and rack awareness. However, this level of resilience requires a highly available, redundant network between the cluster's nodes. Protecting the cluster's read and write storage capabilities also requires highly available connectivity between nodes. Even with intelligent data placement strategies, if network connectivity between more than the allowed number of nodes breaks down, VMs on the cluster could experience write failures and enter read-only mode.

To optimize I/O speed, Nutanix clusters send each write to another dynamically selected node in the cluster. As a result, a fully populated cluster sends storage replication traffic in a full mesh, using network bandwidth between all Nutanix nodes. Because storage write latency directly correlates to the network latency between Nutanix nodes, any network latency increase adds to storage write latency.

Audience

This best practice guide is part of the Nutanix Solutions Library. We intend it for network and virtualization administrators and architects responsible for designing networks for a Nutanix environment. Readers of this document should already be familiar with top-of-rack networking concepts for datacenter environments and basic hypervisor networking.

Purpose

In this document, we cover the following topics:

- Choosing physical switches

- Network design requirements and recommendations:
  - Switch fabric scale
  - VLANs
  - Stretched networks and WANs
  - Oversubscription
  - Rack awareness and block awareness
- Recommended network designs:
  - Leaf-spine
  - Core-aggregation-access (multitier)
  - Multisite designs

Document Version History

Version Number	Published	Notes
1.0	March 2019	Original publication.
2.0	November 2019	QoS and VLAN updates.
2.1	March 2020	Updated the Nutanix Enterprise Cloud Overview and Choosing a Physical Switch sections.
2.2	January 2021	Updated the Choosing a Physical Switch section.
2.3	May 2021	Added spanning tree warning.
2.4	April 2022	Added proxy ARP warning.
2.5	February 2023	Added VLAN scope and flood guidance. Other minor text updates.
2.6	March 2023	Clarified proxy ARP warning.

3. Choosing a Physical Switch

A Nutanix environment should use datacenter switches designed for transmitting large amounts of server and storage traffic at low latency. Don't use switches meant for deployment at the campus access layer. Campus access switches may have 10 Gbps ports like datacenter switches, but they aren't usually made to transport a large amount of bidirectional storage replication traffic.

The deployment size and purpose also influence physical switch choice. Datacenter switches with large buffers are critical in a large AOS cluster that grows beyond eight nodes or hosts storage-intensive applications. In smaller clusters or ROBO deployments that have fewer than eight nodes or don't host write-intensive applications, the switch may not experience buffer contention and you can relax these switch restrictions. There are also some switch types you should never use for the data path of any Nutanix deployment because of oversubscription or other architecture choices; we list examples of these as well.

Datacenter switches have the following characteristics:

- Line rate: Ensures that all ports can simultaneously achieve advertised throughput.
- Low latency: Minimizes port-to-port latency, measured in microseconds or nanoseconds.
- Large per-port buffers: Handle speed mismatch from uplinks without dropping frames.
- Nonblocking, with low or no oversubscription: Reduces chance of drops during peak traffic periods.
- 10 Gbps or faster links for Nutanix Controller VM (CVM) traffic: Only use 1 Gbps links either for additional user VM traffic or when 10 Gbps or faster connections are not available, such as in a ROBO deployment. Limit Nutanix clusters using 1 Gbps links to eight nodes.

The switch manufacturer's datasheets, specifications, and white papers can help identify these characteristics. For example, a common datacenter switch datasheet may show a per-port buffer of 1 MB, while an access layer or fabric extension device has a per-port buffer of around 150 KB. During periods of high traffic or when using links with a speed mismatch (such as 40 Gbps uplinks to 10 Gbps edge ports), a smaller buffer can lead to frame drops, increasing storage latency. While there are some network designs that can achieve high throughput and low latency with very small switch buffers, these are generally very expensive or specialized environments, such as high-frequency stock trading, that aren't part of a common datacenter network.

The following list isn't exhaustive, but it gives some examples of model lines that meet the above requirements for high-performance or large clusters. Models similar to the ones listed are also great choices. You should use switches like these for large or high-performing clusters, but you can use them for smaller clusters and ROBO as well.

- Arista 7050X3, 7160, 7170, 7280: larger buffer models
- Aruba CX 8325 and CX 8360
- Cisco Nexus 9000, 7000, and 5000
- Dell S5200-ON
- HPE FM3810, FM3132Q
- Juniper QFX5100, QFX5200, QFX10K
- Lenovo NE2580O
- NVIDIA Mellanox SN2410, SN2100, and SN2010

The following are examples of switches that don't meet the high-performance datacenter switch requirements but are acceptable for ROBO clusters and clusters with low performance requirements or fewer than eight nodes. Avoid using these switches for the data path of large or high-performing clusters.

- Arista 7050 and 7150s: smaller buffer models
- Cisco Nexus 3000: smaller buffer model

- Cisco Catalyst 9300 and 9500: campus access switch
- Cisco Catalyst 3850: stackable multigigabit switch
- HPE FM2072

The following are examples of switches that are never acceptable for any Nutanix data path connection but are acceptable for out-of-band management.

- Cisco Nexus 2000 (Fabric Extender): highly oversubscribed with small per-port buffers
- 10 Gbps expansion cards in a 1 Gbps access switch: provide uplink bandwidth for the switch, not server connectivity

Each Nutanix node also has an out-of-band connection for IPMI, iLO, iDRAC, or similar management. Because out-of-band connections don't have the latency or throughput requirements of VM networking or storage connections, they can use any access layer switch.

> Note: Nutanix recommends an out-of-band management switch network separate from the primary network for high availability.

Nutanix maintains close partnerships with several networking vendors. For more information on the network integration and automation capabilities available with Nutanix, review our list of partners in the Networking and Security section of our Technology Alliances page.

Regardless of the switch vendor you choose, follow the general recommendations in this document. For vendor-specific configuration recommendations, refer to the References section in the Appendix.

4. Network Design Requirements and Recommendations

Maximum of Three Switch Hops

Nutanix nodes send storage replication traffic to each other in a distributed fashion over the top-of-rack network. One Nutanix node can therefore send replication traffic to any other Nutanix node in the cluster. The network should provide low and predictable latency for this traffic. Ensure that there are no more than three switches between any two Nutanix nodes in the same cluster.

A leaf-spine topology satisfies this requirement and is a popular choice.

Same Switch Fabric

A switch fabric is either a single leaf-spine topology or all switches connected to the same switch aggregation layer. The Nutanix VLAN shares a common broadcast domain within the fabric. Connect all Nutanix nodes that form a cluster to the same switch fabric. Don't stretch a single Nutanix cluster across multiple, disconnected switch fabrics.

Every Nutanix node in a cluster must be in the same layer 2 (L2) broadcast domain and share the same IP subnet.

WAN Links

A WAN (wide area network) or metro link connects different physical sites over a distance. As an extension of the switch fabric requirement, don't place Nutanix nodes in the same cluster if they're separated by a WAN. When Nutanix nodes are separated by a WAN, create multiple clusters and use disaster recovery replication between sites.

VLANs

To protect the Nutanix CVM and hypervisor traffic, place them together in their own dedicated VLAN, separate from other VM traffic. Don't place the CVM and hypervisor hosts in a VLAN shared with other VMs.

Nutanix recommends that you configure the CVM and hypervisor host VLAN as the native, or untagged, VLAN on the connected switch ports. This native VLAN configuration allows easy node addition and cluster expansion while placing traffic in a VLAN when it enters the switch. By default, new Nutanix nodes send and receive untagged traffic. If you use a tagged VLAN for the CVM and hypervisor hosts instead, you must configure that VLAN while you provision the new node, before you add that node to the Nutanix cluster.

Within the Nutanix VLAN, nodes use IPv6 neighbor discovery protocol and IPv6 UDP broadcast messages. To simplify cluster expansion, you can disable multicast and broadcast flood optimizations that block unsolicited discovery messages in the Nutanix VLAN. Consult your network team before making these changes, as multicast and broadcast flood optimizations can limit potentially harmful traffic. If these protocols are blocked, you can add nodes manually from the command line.

Use tagged VLANs for all guest VM traffic and add the required guest VM VLANs to all connected switch ports for hosts in the Nutanix cluster. Limit VLANs for guest VM traffic to the smallest number of physical switches and switch ports possible to reduce broadcast network traffic load. If you no longer need a VLAN, remove it. Removing unnecessary VLANs helps to prevent accidental flooding or excessive broadcast and multicast traffic.

Nutanix AHV network automation streamlines VLAN provisioning for guest VMs. For a list of partners that support network automation, visit the Technology Alliances site and filter on networking and security.

Disable proxy ARP in the Nutanix VLAN before configuring network segmentation and other AHV virtual switch operations. Several Nutanix processes use ARP as a method to detect whether a connection or endpoint is active before proceeding. Proxy ARP replies to all ARP requests and invalidates

the results of these checks, causing unexpected failures for operations such as virtual switch updates and other tasks.

Stretched Layer 2 Networking

Datacenters have traditionally shared a VLAN with a layer 2 (L2) broadcast domain between different racks. Modern designs sometimes terminate this L2 boundary at the top of the rack so that each rack is a separate L2 domain and a different IP subnet. Because all nodes in a Nutanix cluster must share the same L2 broadcast domain, this approach presents a challenge.

This scenario—stretching an L2 network between two racks in the same datacenter—is the only case when using a stretched L2 network over a layer 3 (L3) network is acceptable for storage traffic within a Nutanix cluster, because the Nutanix cluster is still in the same switch fabric or aggregation layer.

> Note: Don't place Nutanix nodes in the same Nutanix cluster if the stretched L2 network spans multiple datacenters or availability zones or if there is a remote link between the two locations.

When you need compute and storage at multiple sites, use separate Nutanix clusters at each physical location. Use replication tools such as asynchronous disaster recovery, NearSync, and Metro Availability to share data between Nutanix clusters at different sites. These recommendations apply even if the two availability zones share close proximity or highly available network paths. Each site should be a cluster boundary to protect against network, power, or other failures. The boundary could be a building, a firewall, or even different racks in the datacenter, depending on your availability requirements.

Flow Virtual Networking and Subnet Extension

With Flow Virtual Networking, you can use a Virtual Private Cloud (VPC) to extend the L2 network for user VMs across an L3 boundary. VPCs allow a subnet to exist on multiple Nutanix clusters, even if the clusters are in different L3 subnets. You can't use VPCs to extend the subnet used for Nutanix CVMs.

Rack Awareness and Block Awareness

Block awareness and rack awareness provide smart placement of Nutanix cluster services, metadata, and VM data to help maintain data availability even if you lose an entire block or rack. The same network requirements for low latency and high throughput between servers in the same cluster still apply when you use block and rack awareness.

> Note: Don't use features like block or rack awareness to stretch a Nutanix cluster between different physical sites.

Oversubscription

Oversubscription occurs when an intermediate network device or link doesn't have enough capacity to allow line rate communication between the systems connected to it. For example, if a single 10 Gbps link connects two switches and four hosts connect to each switch at 10 Gbps, the connecting link is oversubscribed. Oversubscription is often expressed as a ratio—in this case 4:1, as the environment could potentially attempt to transmit 40 Gbps between the switches with only 10 Gbps available. Achieving a 1:1 ratio isn't always feasible, but you should keep the ratio as small as possible based on budget and available capacity.

In a typical deployment where Nutanix nodes connect to redundant top-of-rack switches, storage replication traffic between CVMs traverses multiple devices. To avoid packet loss caused by link oversubscription, ensure that the switch uplinks consist of multiple interfaces operating at a faster speed than the Nutanix host interfaces. For example, for nodes connected at 10 Gbps, the interswitch connection should consist of multiple 10 Gbps or faster links.

Host Connections

Connect Nutanix hosts to redundant top-of-rack switches. Use the active-active configuration available in the hypervisor when possible. With vSphere, follow the VMware vSphere Networking best practices, using Route based on Originating Virtual Port or Route based on Physical NIC Load. With AHV, follow

the Nutanix AHV Networking best practices, using either the default active-backup configuration or LACP with balance-tcp.

On switch ports that face Nutanix hosts, designate the interfaces as spanning tree edge ports to minimize port downtime and prevent triggering spanning tree topology changes.

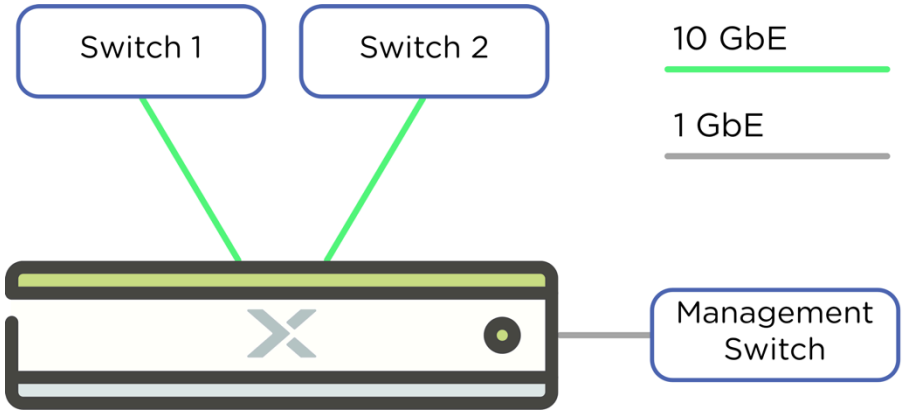

Figure 1: Host Network Connections

Quality of Service

When there is network traffic contention, quality of service (QoS) prioritizes traffic and reserves network resources for certain classes of traffic. QoS can mark important traffic with priority values or enable network switches to derive the traffic's class based on fields inside the traffic. Once you have defined traffic classes, each network device can reserve resources for certain classes or decide which classes to prioritize and which to drop.

With a fabric that eliminates contention for storage traffic, no QoS configuration is required. Nutanix doesn't recommend configuring any top-of-rack QoS for storage traffic. Instead, configure a nonblocking, low-latency switch fabric with no oversubscription between Nutanix nodes.

As a precaution, you can consider traffic policing on the physical switch ports to prevent excessive unicast, broadcast, or multicast flooding. The details of this configuration, sometimes called optimized multicast flood, differ between switch vendors, but it uses a pattern that matches all multicast and broadcast traffic. This configuration prevents excessive traffic from outside the Nutanix cluster from overwhelming the switch ports connecting the Nutanix cluster.

Nutanix doesn't recommend that you customize any QoS markings on traffic sourced from the CVM or the hypervisor host. If guest VMs pass frames with L2 CoS (class of service) or L3 DSCP (differentiated services code point) markers on them, these values pass from the hypervisor to the physical switch.

Nutanix AHV doesn't support configuration of enforcement for QoS prioritization or reservations. To guarantee available bandwidth for guest VMs in AHV, create an additional virtual switch using a dedicated set of physical adapters. For more information follow Nutanix AHV Networking best practices.

With Nutanix on ESXi, refer to the Nutanix on VMware vSphere Networking guide for Network I/O Control (NIOC) recommendations. If you need QoS at the hypervisor level, Nutanix recommends using shares without hard limits, so QoS with NIOC only applies when there is network contention.

Spanning Tree

In traditional networks, using the spanning tree protocol prevents network loops and enforces a loop-free network switch topology. Not all networks use the spanning tree protocol, but if yours does, treat the switch ports facing Nutanix hosts as server or host ports, sometimes referred to as edge ports. A Nutanix server acts like an end host, not a switch. Nutanix recommends either disabling spanning tree on the ports facing Nutanix servers or configuring the Nutanix-facing ports with features such as type edge or portfast, so these ports skip the normal spanning tree phases and immediately forward traffic.

> Note: Configure switch ports facing Nutanix servers as portfast or type edge. Failure to do so may cause service interruption while switch ports transition through the listening and learning phases, dropping all production traffic for up to 30 seconds during spanning tree topology changes.

5. Nutanix-Recommended Network Designs

Leaf-Spine

The leaf-spine network design is popular in new datacenter deployments because it's easy to deploy and easy to scale after deployment. A leaf-spine topology requires at least two spine switches and two leaf switches. Every leaf connects to every spine using uplink ports. There are no connections between the spine switches or between the leaf switches in the conventional leaf-spine design.

Use uplinks that are a higher speed than the edge ports to reduce uplink oversubscription. To increase uplink capacity, add spine switches or uplink ports as needed.

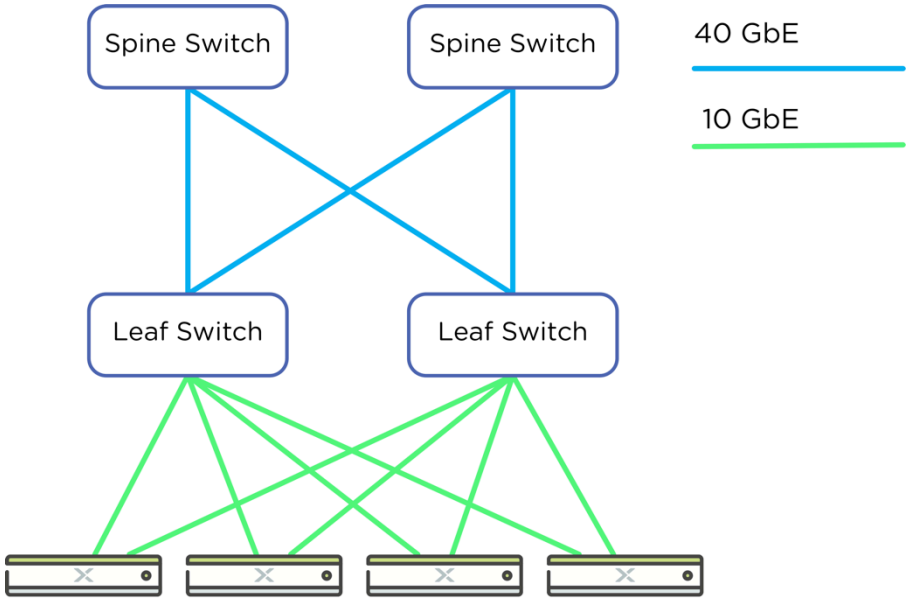

Figure 2: Leaf-Spine Network

To scale the leaf-spine network, add leaf and spine switches. Because there are no more than three switch hops between any two Nutanix nodes in this design, a Nutanix cluster can easily span multiple racks and still connect to the same switch fabric.

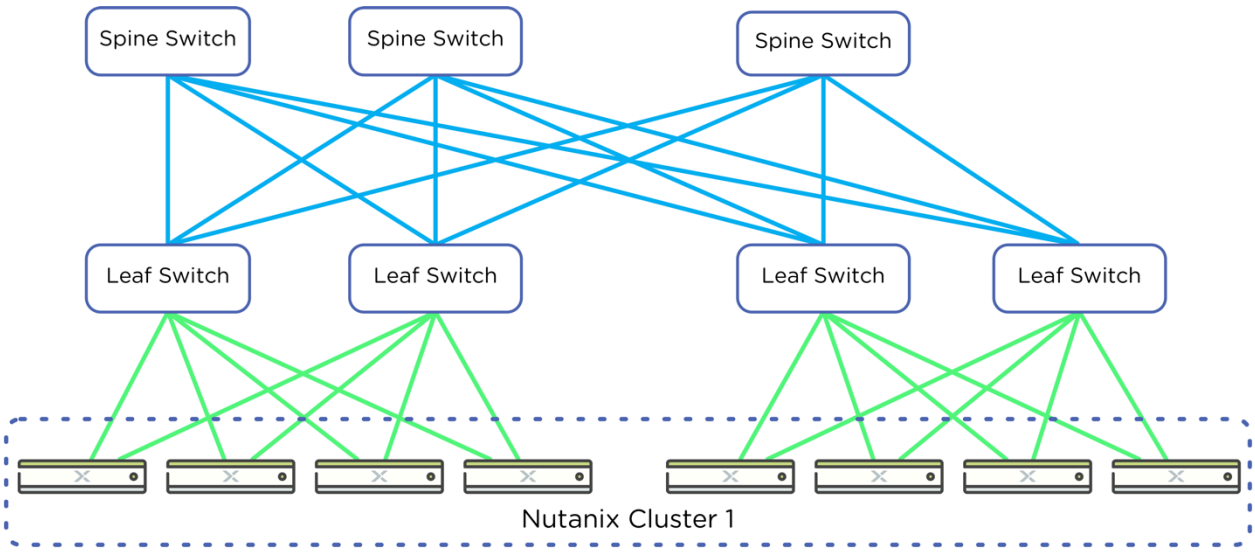

Figure 3: Scaling the Leaf-Spine Network

Core-Aggregation-Access

The core-aggregation-access (or three-tier) design is a modular layout that allows you to upgrade and scale layers independently. Ensure that all nodes in a Nutanix cluster share the same aggregation layer to comply with the three-switch-hop rule.

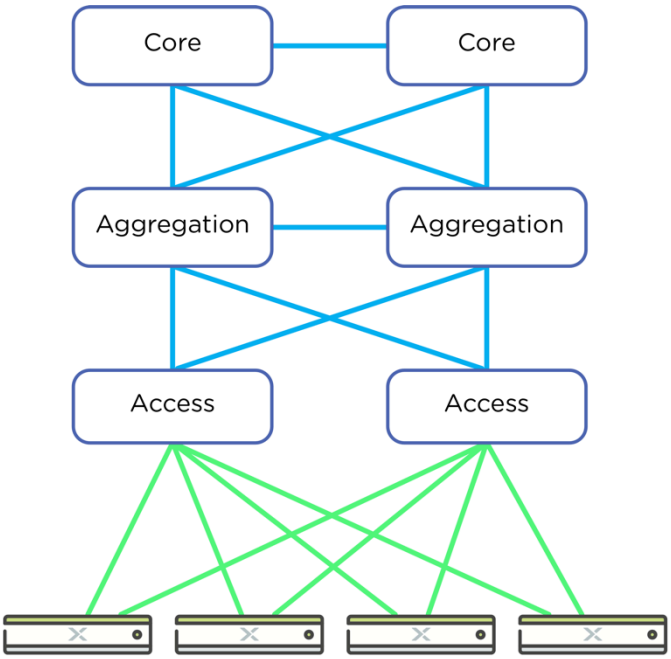

Figure 4: Core-Aggregation-Access Network

Scaling the three-tier network design may require adding another aggregation and access layer to the core. In this case, there would be more than three switch hops between the two access layers. Ensure that you add Nutanix nodes in separate aggregation and access layers to separate clusters to keep the number of switch hops between nodes in the same cluster to three or fewer. In the following example, Cluster 1 connects to one aggregation layer and Cluster 2 connects to another.

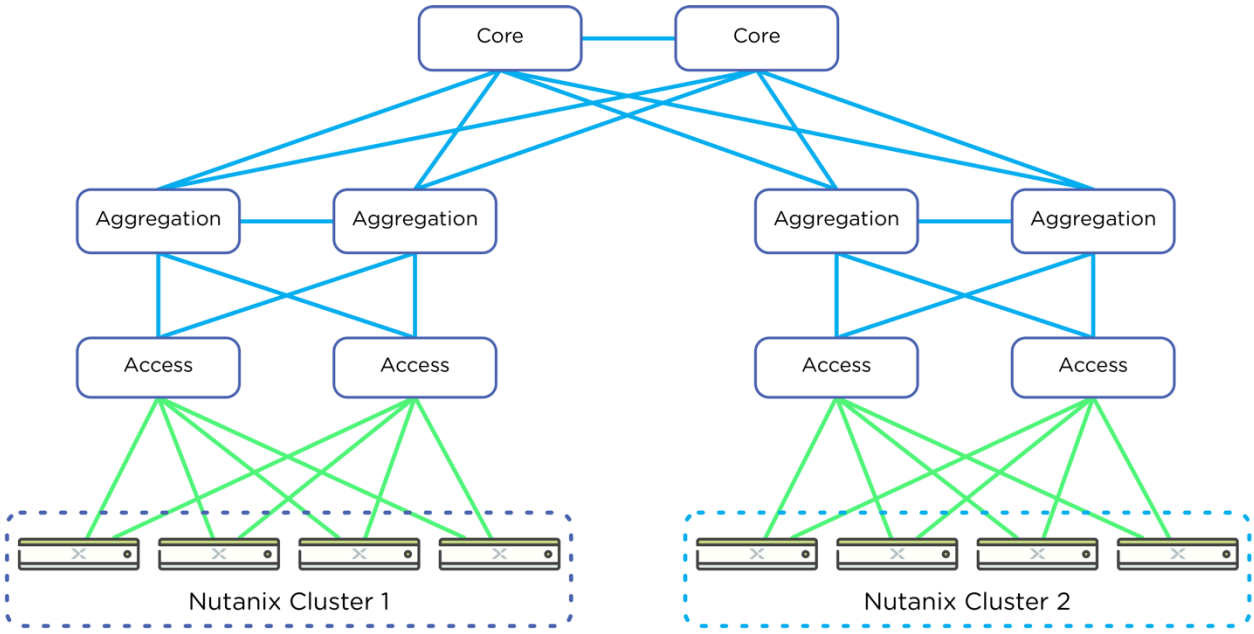

Figure 5: Scaling the Core-Aggregation-Access Network

There are many ways to connect switches in the three-tier design, so your deployment may look slightly different than the one shown above.

Multisite Designs

When two or more physical sites or physical availability zones exist, a single Nutanix cluster shouldn't span them. Instead, create multiple Nutanix clusters (one per availability zone) and connect them with tools such as asynchronous disaster recovery, NearSync, and Metro Availability. This design provides high availability for data and applications and eliminates the possibility of a split-brain scenario, in which a Nutanix cluster is partitioned when the two sites lose connectivity.

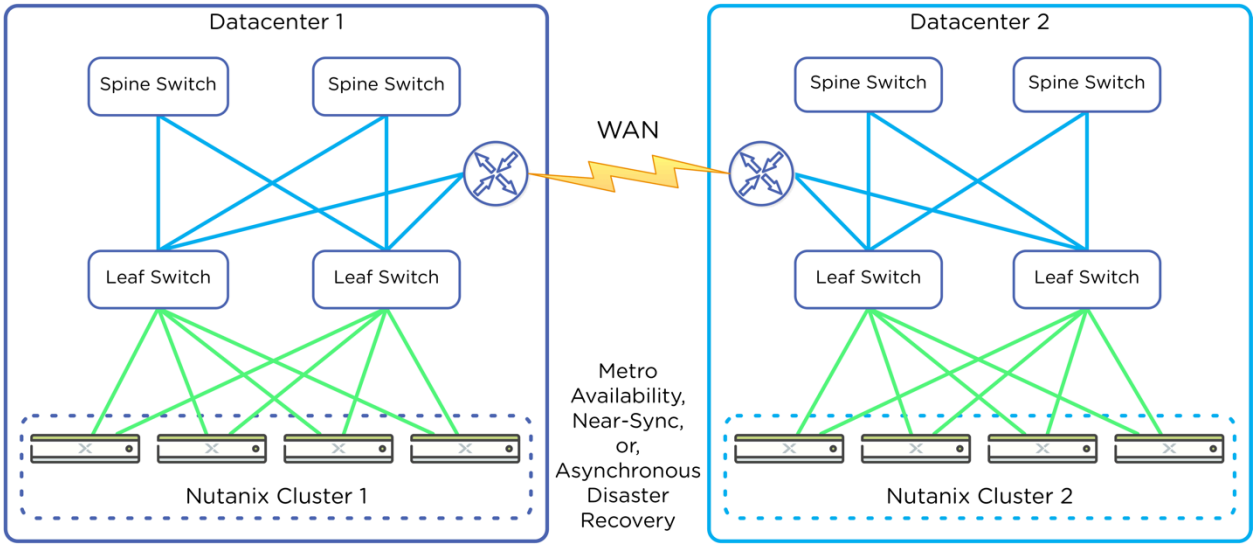

Figure 6: Multisite Network Design Using Replication

6. Conclusion

We often talk to customers concerned about the network for their datacenters as well as for their individual Nutanix clusters. Although networking may seem daunting, it requires attention to only three key points:

1. Procure the right switches.
2. Select the right network design.
3. Use the right replication technologies.

If you connect your Nutanix cluster to a well-designed datacenter network, your storage platform can be highly available while still providing you with excellent performance.

For feedback or questions, please contact us using the Nutanix NEXT Community forums.

7. Appendix

Physical Networking Best Practices Checklist

A single Nutanix cluster must meet the following requirements:

- Have a maximum of three switch hops between any two Nutanix nodes in the same cluster.
- Connect all Nutanix nodes in the same cluster to the same switch fabric (leaf-spine network) or aggregation layer.
- Avoid WAN or remote links between Nutanix nodes in the same Nutanix cluster.
- Separate Nutanix CVM and hypervisor hosts into a dedicated VLAN that doesn't include any VM traffic.
- After consulting your network team, disable flood optimization in this Nutanix VLAN to allow discovery.
- Trunk only the required VLANs to switch ports facing Nutanix servers and remove all other VLANs.
- Consider policing or flood optimization in non-Nutanix VLANs to prevent excessive broadcast or multicast traffic.
- Disable proxy ARP in the Nutanix VLAN before configuring network segmentation and other AHV virtual switch operations to prevent unexpected failures.
- Don't place Nutanix nodes in the same Nutanix cluster if the stretched L2 network spans multiple datacenters or availability zones or if there is a remote link between the two locations.

- Only use a stretched L2 network over L3 when the Nutanix cluster remains in the same switch fabric or aggregation layer, such as a L2 network stretched between two racks in the same datacenter.
  - Nutanix VPCs can stretch the L2 network for user VMs but not for CVMs.
- Don't use features like block or rack awareness to stretch a Nutanix cluster between different physical sites.
- Configure adequate uplinks between switches or interswitch links for east-west storage traffic to minimize port-to-port oversubscription. For example, use multiple 40 Gbps or faster uplinks (or interswitch links).
- Connect hosts using redundant links.
- Configure switch ports facing Nutanix servers as spanning tree portfast or edge to skip the listening and learning phases and prevent cluster outages due to spanning tree topology changes.

These requirements and recommendations keep latency between nodes minimal and predictable. Networks that have too many switch hops or introduce WAN links between only some of the nodes introduce variable latency, which has a negative impact on storage latency, and variable throughput. Networks that are larger than required also add components (datacenters, zones, switches, and switch fabrics) that can all impact availability.

References

1. Data Protection and Disaster Recovery
2. Data Protection for AHV-Based VMs
3. Metro Availability
4. Nutanix AHV Networking
5. VMware vSphere Networking
6. VMware NSX for vSphere
7. Cisco ACI™ with Nutanix
8. NVIDIA Networking with Nutanix
9. Networking: Arista EOS Recommended Practices
10. Cisco Nexus Recommended Practices

About Nutanix

Nutanix is a global leader in cloud software and a pioneer in hyperconverged infrastructure solutions, making clouds invisible and freeing customers to focus on their business outcomes. Organizations around the world use Nutanix software to leverage a single platform to manage any app at any location for their hybrid multicloud environments. Learn more at www.nutanix.com or follow us on Twitter @nutanix.

List of Figures