NUTANIX
YOUR ENTERPRISE CLOUD

Mellanox Networking with Nutanix

Nutanix Solution Note

Version 1.3 • September 2019 • SN-2063

Copyright

Nutanix, Inc.
1740 Technology Drive, Suite 150
San Jose, CA 95110

Contents

1. Executive Summary

This solution note addresses the fundamental differences between traditional three-tier and modern leaf-spine networking architectures and details the configuration elements required when coupling the Nutanix Enterprise Cloud OS with a Mellanox networking solution.

This document presents solutions illustrating the various ways you can lay out a leaf-spine network to achieve scale and density, using configurations ranging from 2 switches and 3 servers up to 50 switches and 720 servers. Nutanix nodes are equipped with redundant 10/25/40 GbE NICs to service virtual (VM, CVM, management, and migration) and physical network connectivity.

With the flexibility that virtualization offers, administrators can dynamically configure, balance, and share logical components across various traffic types. However, when architecting a network solution, we must also take the physical topology into consideration. Designing and implementing a resilient and scalable network architecture ensures consistent performance and availability and complies with security policies and regulatory requirements when scaling Nutanix hyperconverged appliances.

Unless otherwise stated, the solution described in this document is valid on all supported AOS releases.

Table 1: Document Version History

Version Number	Published	Notes
1.0	January 2017	Original publication.
1.1	August 2017	Updated platform overview and added references specific to ESXi 5.5 and AHV.
1.2	September 2018	Updated Nutanix overview.
1.3	September 2019	Added Mellanox SN2010 10/25 Gbps ToR switch and its use case.

2. Nutanix Enterprise Cloud Overview

Nutanix delivers a web-scale, hyperconverged infrastructure solution purpose-built for virtualization and cloud environments. This solution brings the scale, resilience, and economic benefits of web-scale architecture to the enterprise through the Nutanix Enterprise Cloud Platform, which combines three product families—Nutanix Acropolis, Nutanix Prism, and Nutanix Calm.

Attributes of this Enterprise Cloud OS include:

- Optimized for storage and compute resources.
- Machine learning to plan for and adapt to changing conditions automatically.
- Self-healing to tolerate and adjust to component failures.
- API-based automation and rich analytics.
- Simplified one-click upgrade.
- Native file services for user and application data.
- Native backup and disaster recovery solutions.
- Powerful and feature-rich virtualization.
- Flexible software-defined networking for visualization, automation, and security.
- Cloud automation and life cycle management.

Nutanix Acropolis provides data services and can be broken down into three foundational components: the Distributed Storage Fabric (DSF), the App Mobility Fabric (AMF), and AHV. Prism furnishes one-click infrastructure management for virtual environments running on Acropolis. Acropolis is hypervisor agnostic, supporting two third-party hypervisors—ESXi and Hyper-V—in addition to the native Nutanix hypervisor, AHV.

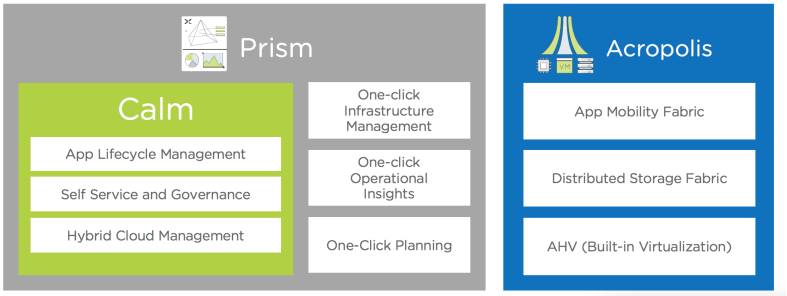

Figure 1: Nutanix Enterprise Cloud

2.1. Nutanix Acropolis Architecture

Acropolis does not rely on traditional SAN or NAS storage or expensive storage network interconnects. It combines highly dense storage and server compute (CPU and RAM) into a single platform building block. Each building block delivers a unified, scale-out, shared-nothing architecture with no single points of failure.

The Nutanix solution requires no SAN constructs, such as LUNs, RAID groups, or expensive storage switches. All storage management is VM-centric, and I/O is optimized at the VM virtual disk level. The software solution runs on nodes from a variety of manufacturers that are either all-flash for optimal performance, or a hybrid combination of SSD and HDD that provides a combination of performance and additional capacity. The DSF automatically tiers data across the cluster to different classes of storage devices using intelligent data placement algorithms. For best performance, algorithms make sure the most frequently used data is available in memory or in flash on the node local to the VM.

To learn more about the Nutanix Enterprise Cloud, please visit the Nutanix Bible and Nutanix.com.

3. Mellanox Networking for Enterprise Cloud

Traditional datacenter networks have a three-layer topology:

- Core: Where everything comes together (L3).
- Aggregation or distribution: Where the access switches connect (L2).
- Access layer: Where servers are connected (L1, often located at the top of a rack).

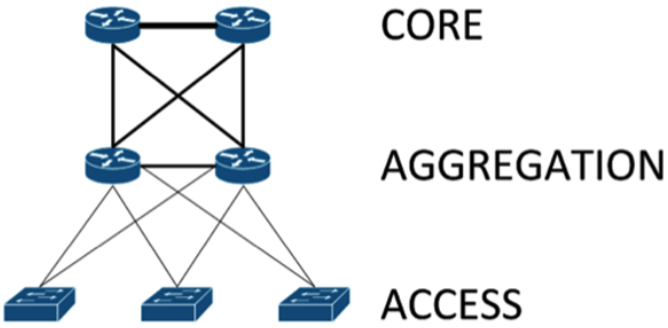

Figure 2: Traditional Network Tiers

When adding nodes or appliances at the access layer in this framework, you identify the network port density requirements at each layer and estimate how capex and opex costs are likely to increase. The three-layer architectural approach enables you to physically scale datacenter ports by simply adding switches at L2 and trunking them upstream to the existing aggregation layer.

Although this is a straightforward method for scaling the network, you must consider the rate of oversubscription to the upstream aggregation and core layers so you can provide sufficient bandwidth for L3 and above.

Traffic flows in the three-tier topology are predictable, as they are predominantly north-south flows—traffic that moves directly between the access and core layers (see the following figure). L2 traffic segments are contained in the same access layer and then switch upstream at the core for L3 routing or network services.

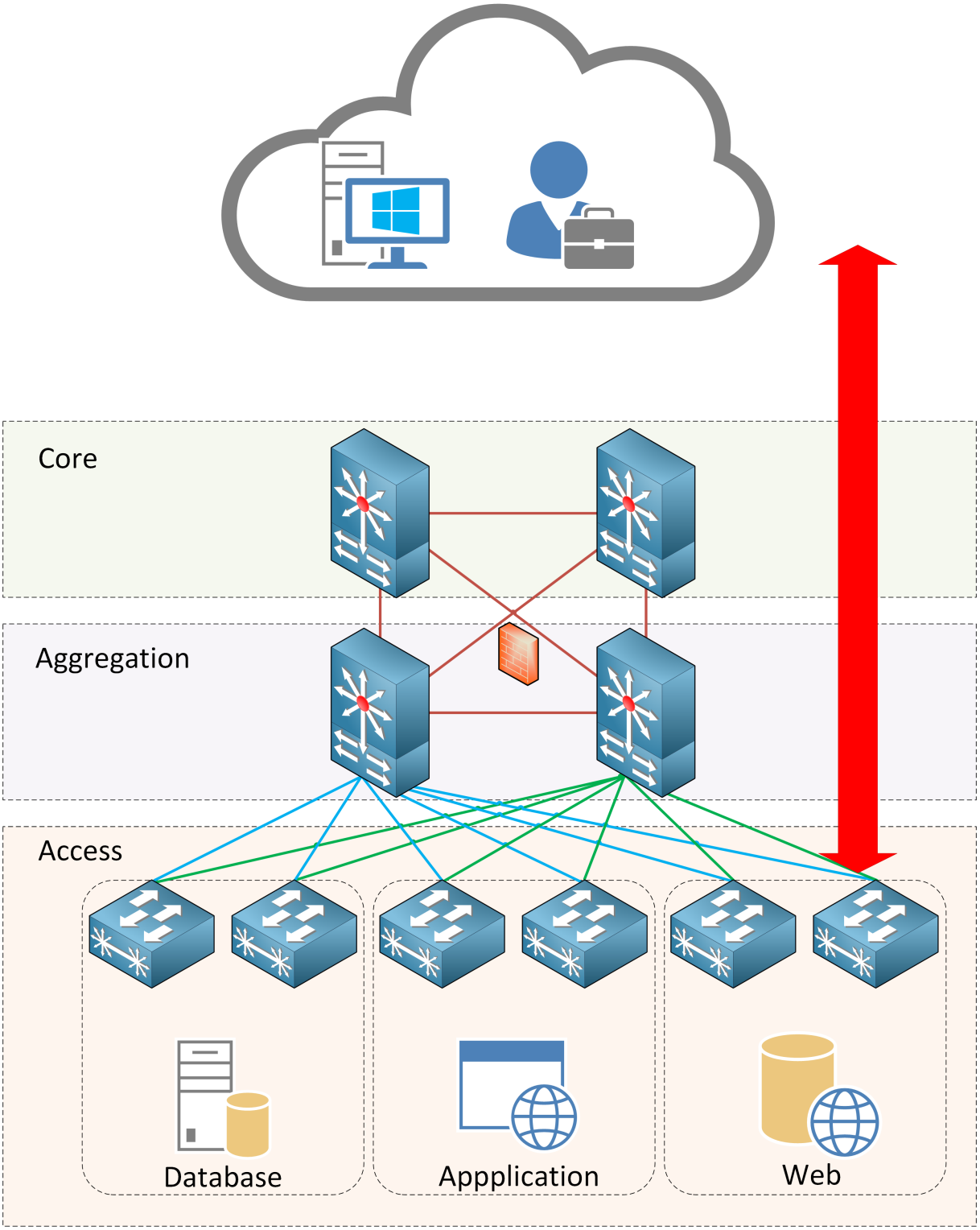

Figure 3: Traditional Three-Tier Network with North-South Traffic

This topology suits a datacenter composed of physical servers. However, two industry trends are driving new designs for datacenter network topology: virtualization and hyperconvergence.

In response to these trends, many organizations have introduced a “virtualize-first” policy, in which applications run inside a virtual machine (VM) on a physical host and multiple isolated VMs share the underlying hardware. Another key virtualization feature is the ability to migrate VMs across hosts using policy-based automation.

Hyperconvergence natively converges storage and compute into a standard x86 appliance. Each device contains local direct-attached storage through software-defined storage logic. Clustering all servers using IP addresses creates a single unified storage pool (the aggregate of all storage devices across appliances) and allows hosts in the cluster to access it.

In an HCI environment, flows now move in an east-west pattern, because we have hosts spanning multiple access layers and racks, using the existing network for both storage and application services. Using a three-tier networking model in this environment can limit a network’s performance and scalability, affecting the overall solution. As virtualization and hyperconvergence now expand across datacenter boundary and into the scope of multiclouds, these limitations of three-tier networking are exacerbated.

3.1. Performance and Scalability Limitations of Three-Tier Architecture

A three-tier architecture relies on interswitch links to provide network connectivity across access layer segments. Link oversubscription can arise when the spanning tree protocol (STP) blocks redundant links to prevent network loops on the L2 segments.

Adding links to a bond is one way to reduce link oversubscription; however, additional links are only a temporary solution for a network where communication stretches across all access layers from multiple hosts at a given time.

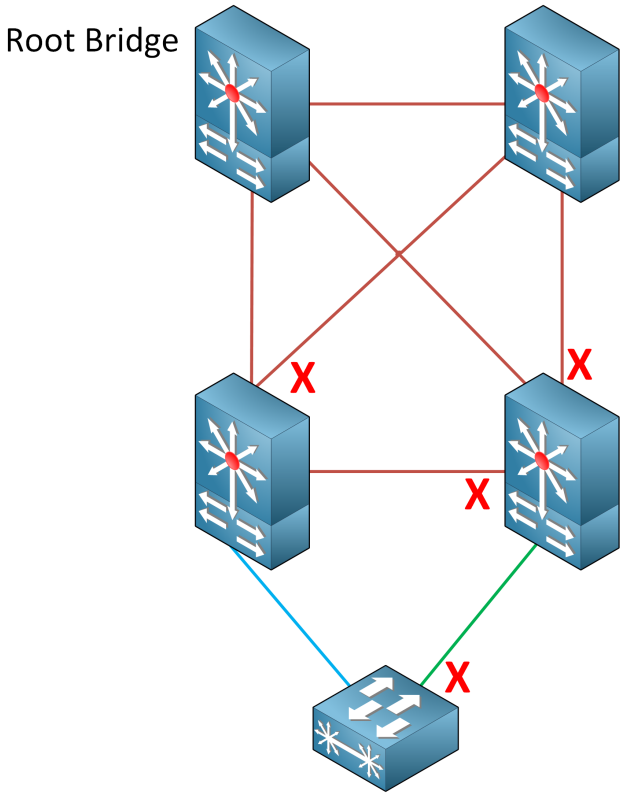

Figure 4: Spanning Tree Port Blocking

Furthermore, such stretched networks often experience suboptimal routing, which occurs when packets must leave the access layer, travel upstream to be switched at the core, and then go back to the access layer for processing, increasing latency. Inefficient routing like this can lead to an inability to scale and difficulty achieving consistent latency between the different points in the network.

Although these limitations can impact any data flow, they are particularly harmful to storage traffic, such as the internode data transfer in a hyperconverged infrastructure or software-defined storage.

3.2. Leaf-Spine Networks

An alternative to a three-tier framework, a leaf-spine architecture uses multiple access layer leaf switches that connect directly to every spine, providing horizontal scalability without the limitations of a spanning tree.

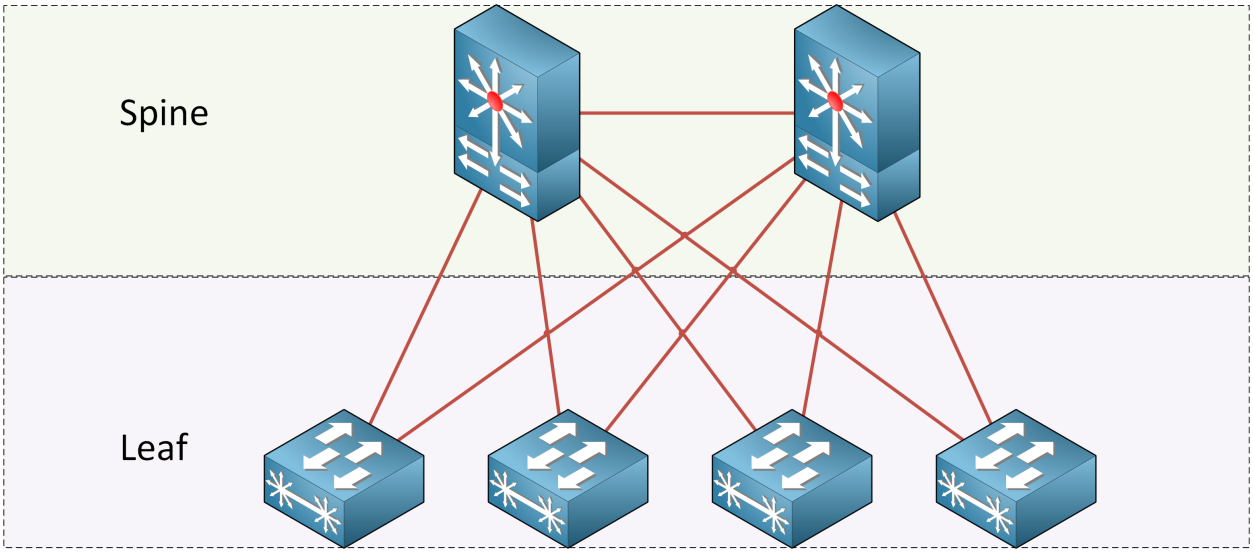

Figure 5: Leaf-Spine Network

You can use leaf switches for either layer 2 or layer 3 service. In an layer 2 design, utilizing modern protocols like TRILL or SPB, a loop-free leaf-spine topology minimizes latency between endpoints connected to the leaf, because it has a single hop for communication—from any leaf to the spine to any other leaf. This architecture allows consistent and predictable latency. Adding spine switches or adding more links between each leaf and its spine provides additional bandwidth between leaf switches.

When using a leaf switch for layer 3 service, each link becomes routed. In an layer 3 design, using protocols like OSPF and BGP with ECMP, all link bandwidth is utilized dynamically with load balancing and redundancy. With routed links, overlay technologies like VXLANs (Virtual Extensible Local Area Networks) or VLANs (Virtual Local Area Networks) can increase efficiency, because traffic does not need to traverse the spine for layer 3 services.

Overlay technologies provide L2 extension across datacenter boundaries and allow you to isolate traffic across individual leaf switches or spread traffic across multiple leaf switches, independent of the physical infrastructure address space.

3.3. Leaf-Spine Architecture with Mellanox Networking

In the following design, we demonstrate how to achieve a leaf-spine topology using Mellanox SN2xxx Series switches. This reference architecture consists of the hardware and software components listed in the following table.

Table 2: Hardware and Software Components

Switch #	Make	Model	Specification	Role
1	Mellanox	SN2700	32x 100 GbE	Spine or leaf
2	Mellanox	SN2010	18x 25 GbE + 4x 100 GbE	Leaf
3	Mellanox	SN2100	16x 100 GbE	Leaf
4	Mellanox	SN2100B	16x 40 GbE	Leaf
5	Mellanox	AS4610	48x 1 GbE	Out-of-band management

Mellanox switches allow you to create a network fabric that offers predictable, low-latency switching while achieving maximum throughput and linear scalability. Based on Mellanox Spectrum ASIC (application-specific integrated circuit), Mellanox switches deliver 10/25/40/50/100 Gbps speeds with the industry's lowest port-to-port latency (approximately 300 ns) and zero avoidable packet loss. In particular, the half-width, 1RU (rack unit) ToR (top-of-rack) switches are purpose built for Nutanix HCI racks with the right port count, the most compact design for high availability, and the rich feature set supporting EVPN-based VXLAN for datacenter interconnect (DCI) and RDMA over Converged Ethernet for performance acceleration. Mellanox switches are open to run various network operating systems (NOS), including Mellanox's own ONYX NOS and Cumulus Linux NOS.

The Mellanox networking solution includes What Just Happened (WJH, a streaming telemetry engine), and NEO (a REST API–based network management and orchestration software). Combined with WJH, Mellanox NEO provides automated VM-level network provisioning and monitoring of the end-to-end Mellanox network, especially through API integration with Nutanix AHV. Refer to this Mellanox community post for information on how to install and use the NEO plugin for Nutanix.

3.4. Mellanox Setup and Deployment

Mellanox provides a streamlined deployment model with a full document set to facilitate networking configurations ranging from basic to advanced. For a starter, refer to the following quick reference guides for Mellanox SN2010 switch:

- Quick Starter Guide with CLI
- Quick Starter Guide with NEO

For DCI deployment across a multicloud environment, refer to this reference deployment guide.

Mellanox networking is also used as the underlay network solution for Nutanix AI. Refer to the Nuntanix AI reference architecture for more details.

For this solution note, we demonstrate how to configure MLAG with CLI.

Combined with the features and intelligence of the NOS, multilink aggregation groups (MLAGs) create a highly available layer 2 fabric across Mellanox networking appliances to ensure that you can meet even the most stringent SLAs.

MLAGs aggregate ports across multiple physical switches. Configuring link aggregation between physical switch ports and Nutanix appliances enables the Nutanix Controller Virtual Machine (CVM) to utilize all pNICs actively load balancing user VM traffic. This capability is a key advantage, particularly in all-flash clusters.

The Mellanox OS provides a streamlined deployment model with a full documentation set to facilitate networking configurations ranging from basic to advanced.

Managing and updating each switch independently with MLAGs mitigates the single point of failure that typically results from using stacking techniques in the switches. For most deployments, we recommend using MLAGs instead of switch stacking.

MLAGs do not disable links to prevent network loops, as with STP. Although STP is still enabled to prevent loops during switch startup, once the switch is initialized and in a forwarding state, the MLAG disables STP and ensures that all the links are available to pass traffic and benefit from the aggregated bandwidth.

Although MLAGs add some slight management overhead, we can reduce that overhead significantly by using automation frameworks such as Puppet, Ansible, Chef, or CFEngine. MLAGs thus clearly have the advantage, because you can update the switch firmware independently without causing any disruption to the rest of the network.

Tip: To set up high avaliability on ESXi 5.5 with Mellanox adapters and switches, consult the Mellanox KB article on the topic. To set up high avalability on AHV with Mellanox adapters and switches, consult the Nutanix Connect Blog post Network Load Balancing with Acropolis Hypervisor.

3.5. Mellanox MLAG Configuration via CLI

You can find more details on MLAG configuration in the Quick Starter Guides listed or on the Mellanox community site.

Note: Before you start, make sure that both switches have the same software version. Run **show version** to verify. In addition, we recommend upgrading both switches to the latest MLNX-OS software release.

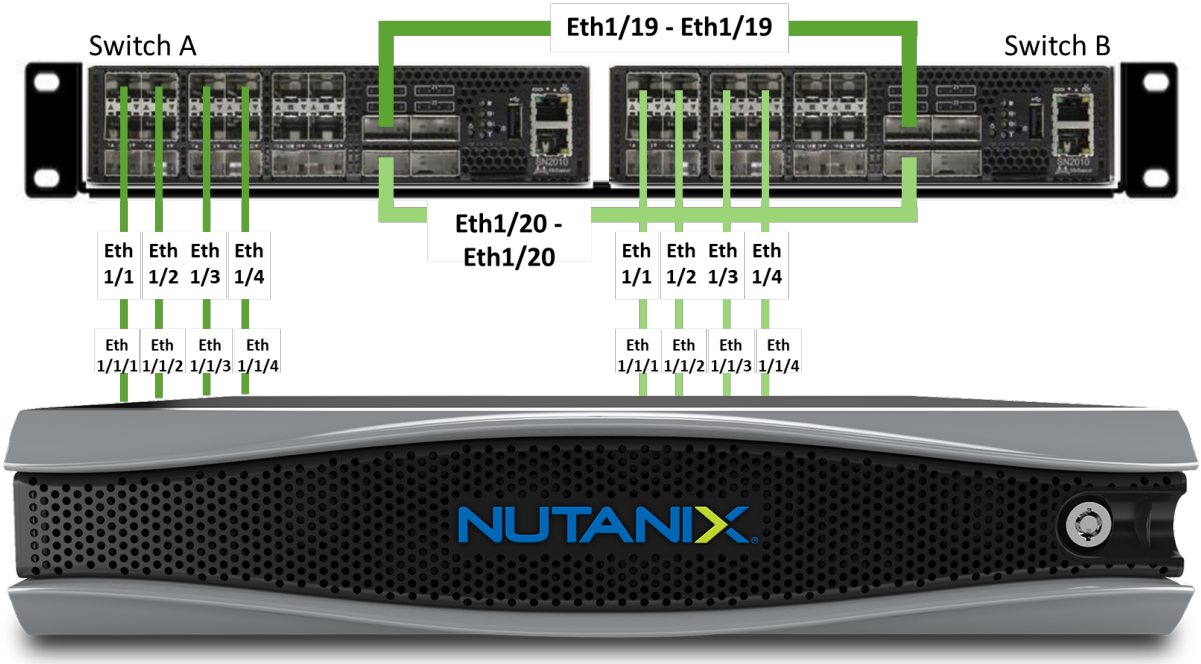

Figure 6: Configuring MLAGs on Mellanox SN2010

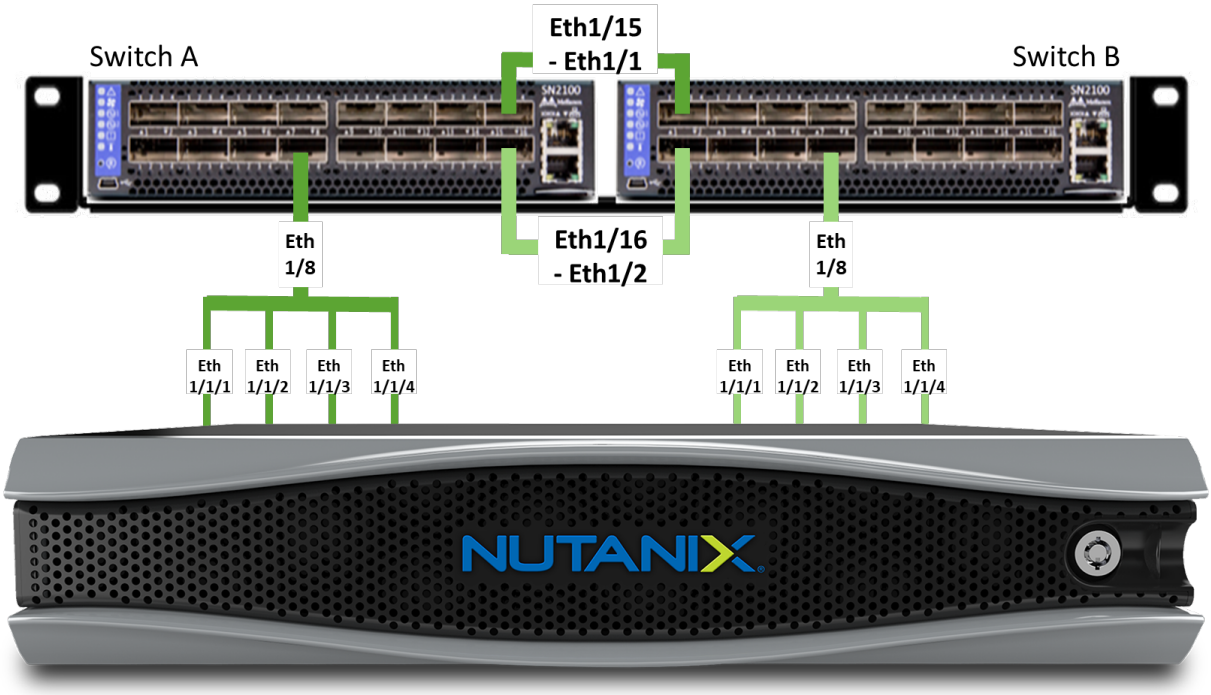

Figure 7: Configuring MLAGs on Mellanox SN2100

General

With a Nutanix cluster that can have up to 18 servers in a single rack, we recommend two Mellanox SN2010 switches. The following configurations all use these switches for 18-server configurations. These configurations can be scaled up for bigger deployments with SN2100 and SN2700 switches.

Run the following commands on both switches:

```
Switch (config) # lacp
Switch (config) # no spanning-tree
Switch (config) # ip routing
Switch (config) # protocol mlag
Switch (config) # dcb priority-flow-control enable force
```

Tip: You can find more details on how to split ports on the Mellanox community site.

Configure IPL

Configure the interpeer link (IPL), a link between switches that maintains state information, over an MLAG (port channel) with ID 1. For high availability, we recommend having more than one

physical link in this LAG. In this example, we are configuring the IPL on ports 1/19 and 1/20 of two SN2010 switches. All VLANs are open on these ports. The example uses VLAN 4000 for configuring the IP address.

- Run the following commands on both switches:

```
Switch (config) # interface port-channel 1
Switch (config interface port-channel 1 ) # exit
Switch (config) # interface ethernet 1/19 channel-group 1 mode active
Switch (config) # interface ethernet 1/20 channel-group 1 mode active
Switch (config) # vlan 4000
Switch (config vlan 4000) # exit
Switch (config) # interface vlan 4000
Switch (config interface vlan 4000 ) # exit
Switch (config) # interface port-channel 1 ipl 1
Switch (config) # interface port-channel 1 dcb priority-flow-control mode on force
```

- Configure the IP address for the IPL link on both switches on VLAN 4000. Enter the following commands on switch A:

```
Switch-A (config) # interface vlan 4000
Switch-A (config interface vlan 4000) # ip address 10.10.10.1 255.255.255.0
Switch-A (config interface vlan 4000) # ipl 1 peer-address 10.10.10.2
```

- Enter these commands on switch B:

```
Switch-B (config) # interface vlan 4000
Switch-B (config interface vlan 4000) # ip address 10.10.10.2 255.255.255.0
Switch-B (config interface vlan 4000) # ipl 1 peer-address 10.10.10.1
```

MLAG VIP and MAC

The MLAG VIP (virtual IP) is important for retrieving peer information.

Note: The MLAG VIP address should be in the same subnet as the management interface (mgmt0).

- Configure the following on both switches:

```
Switch (config) # mlag-vip my-mlag-vip-domain ip 10.209.28.200 /24 force
Switch (config) # mlag system-mac 00:00:5E:00:01:5D
Switch (config) # no mlag shutdown
```

MLAG Interface (Downlinks)

In this example, there are 18 MLAG ports—one for each host. Host 1 is connected to mlag-port-channel 1, and host 2 is connected to mlag-port-channel 2.

- Configure the following on both switches:

```
Switch (config) # interface mlag-port-channel 1-18
Switch (config) # mtu 9216 force
Switch (config interface port-channel 1-2 ) # exit
```

- On switch A:

```
Switch-A (config) # interface ethernet 1/1 mlag-channel-group 1 mode on
Switch-A (config) # interface ethernet 1/2 mlag-channel-group 2 mode on
Switch-A (config) # interface ethernet 1/3 mlag-channel-group 3 mode on
Switch-A (config) # interface ethernet 1/4 mlag-channel-group 4 mode on
:
For all 18 port channels
:
:
Switch-A (config) # interface ethernet 1/16 mlag-channel-group 16 mode on
Switch-A (config) # interface ethernet 1/17 mlag-channel-group 17 mode on
Switch-A (config) # interface ethernet 1/18 mlag-channel-group 18 mode on
Switch-A (config) # interface mlag-port-channel 1-18 no shutdown
```

- On switch B:

```
Switch-B (config) # interface ethernet 1/1 mlag-channel-group 1 mode on
Switch-B (config) # interface ethernet 1/2 mlag-channel-group 2 mode on
Switch-B (config) # interface ethernet 1/3 mlag-channel-group 3 mode on
Switch-B (config) # interface ethernet 1/4 mlag-channel-group 4 mode on
:
For all 18 port channels
:
:
Switch-B (config) # interface ethernet 1/16 mlag-channel-group 16 mode on
Switch-B (config) # interface ethernet 1/17 mlag-channel-group 17 mode on
Switch-B (config) # interface ethernet 1/18 mlag-channel-group 18 mode on
Switch-B (config) # interface mlag-port-channel 1-18 no shutdown
```

> **Note:** Set the MLAG interface in LACP mode, with run mode **active**.

Uplinks

In the previous example, you can use port numbers 1/21 and 1/22 for uplinks to connect to the spine layer.

4. Example Deployment Scenarios

Combining Mellanox SN2010 switches with SN2100 and SN2700 switches allows you to create a scalable network design with predictable and consistent low latency and high throughput from end to end in the network. Configuring the MLAG is a straightforward process that provides management flexibility by combining multiple physical interfaces into a single logical link between devices.

To achieve reliable, linear performance for hyperconverged web-scale environments, the network must be able to scale as seamlessly as the systems.

4.1. Scalable Architecture 1. Small: 192 Nodes

The following figure is a high-level diagram for a small-scale deployment. In this case, two Mellanox SN2010 switches serve as leaf switches for each rack, while a pair of Mellanox 2700 Series switches forms the spine. An AS4610 switch for out-of-band management connectivity.

The Mellanox SN2010 Series places two independent switches side by side in a 1RU platform to accommodate the highest rack performance. SN2100 offers 18 native 10/25 GbE ports and 4x 40/100 GbE ports.

This solution starts with a single rack (containing a minimum of three hyperconverged appliances or nodes) and can scale to 12 racks. Each rack has full 10/25 GbE redundant connectivity, with 1 GbE connections for out-of-band management.

Nutanix nodes connect to their respective leaf switches and to the AS4610, which provides 1 GbE out-of-band management connectivity.

The Mellanox SN2010 leaf switches connect to the SN2700 Series spine via QSFP cables. Depending on the leaf-switch model, these connections could provide 40 Gbps or 100 Gbps throughput per uplink back to the spine. As a result, oversubscription ratios may vary.

Using the Nutanix 3060 Series, with a total of four nodes or 16 hyperconverged blocks per rack, this deployment supports up to 192 servers across 12 racks.

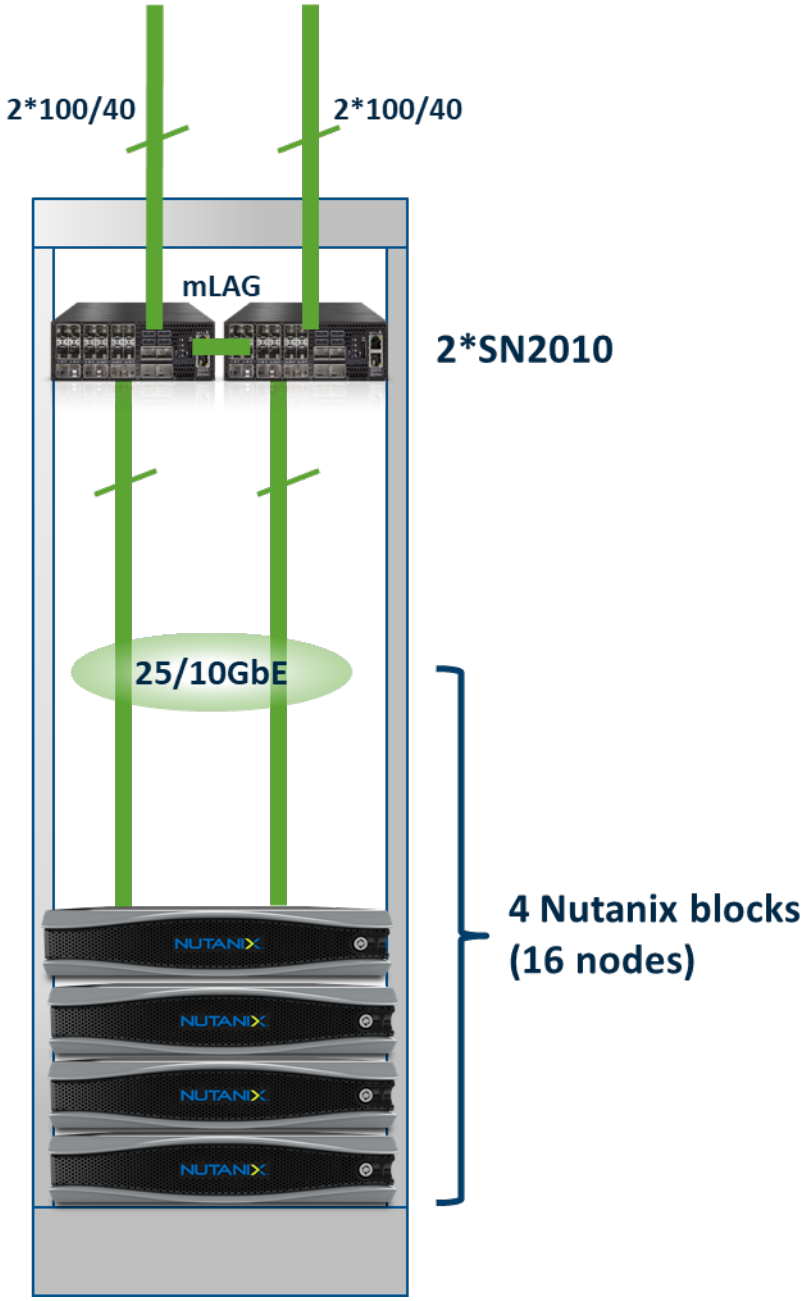

Figure 8: Small-Density Node Configuration

Leaf Switch Density Calculations: Small

- 4 Nutanix 3060 G6 blocks in each rack (4 nodes per block) = 16 nodes per rack.

- Because each node contains 2x 10 GbE ports, you need a total of 8x 10 GbE ports per four-node block, with a total of 80 ports per rack. Because each SN2010 Series switch contains 18 native 10/25 GbE ports, when we use 2 of them as leaf switches, we use 32 direct 10 GbE cables for connections to 16 nodes. This configuration leaves 2x 10/25 GbE ports on each leaf switch.
- Two 100 GbE ports from our leaf switches form an MLAG peering between the pair, while two more 100 GbE ports uplink to their spine switch.
- We need 16x 1 GbE ports to satisfy our out-of-band connectivity requirements. One AS4610 switch provides 48x 1 GbE connectivity per switch and 4x 10 GbE ports per switch. We use two of these 10 GbE ports for establishing uplinks to the respective spine switches.

Spine Density Calculations: Small

- The spine, consisting of two SN2700 Series switches, contains 32x 100 GbE ports and has the ability to convert a 100 GbE port into a 10, 25, 40, 50, or 56 GbE port using the appropriate QSFP+ Optic breakout cable.
- To satisfy the connectivity requirements for each rack, we need four 40 GbE or 100 GbE ports per rack (for the two leaf switches) and two 10 GbE ports (for the one out-of-band AS4610 switch), all of which are trunked to our Mellanox SN2700 spine switches.
- Subtracting the two ports required for the MLAG between our SN2700 Series switches (2x 100 GbE or 2x 40 GbE), each switch has 30x 100 GbE or 40 GbE ports available for leaf connectivity.
- Therefore, scaling the solution to 12 racks, with 2 ports per rack, requires 24 switch ports per spine for leaf connectivity, leaving 6 ports available per spine. Of these 6 ports, we need 3 ports to provide out-of-band switch connectivity utilizing the QSFP-to-4xSFP+ breakout cables (3 x 4 = 12 ports), so we have 3 ports remaining per spine.

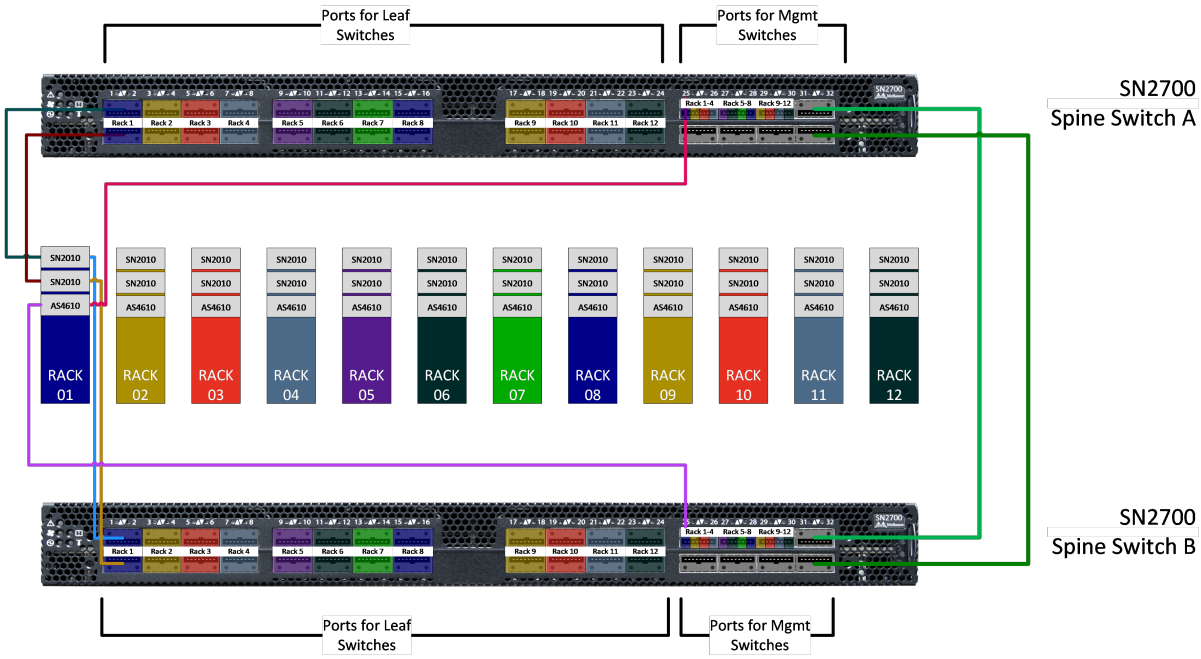

Figure 9: Small-Density Spine Configuration

4.2. Scalable Architecture 2. Medium: 480 Nodes

The following figure is a high-level diagram for a medium-scale deployment. In this case, two Mellanox SN2100 or SN2100B switches serve as leaf switches for each rack, while a pair of Mellanox 2700 Series switches forms the spine. An AS4610 switch provides out-of-band management connectivity.

The Mellanox SN2100 Series places two independent switches side by side in a 1RU platform to accommodate the highest rack performance. This series is available in two primary configurations: SN2100 offers 16x 100 GbE nonblocking ports, while SN2100B offers 16x 40 GbE nonblocking ports.

This solution starts with a single rack (containing a minimum of three hyperconverged appliances or nodes) and can scale to 12 racks. Each rack has full 10 GbE redundant connectivity, with 1 GbE connections for out-of-band management.

Nutanix nodes connect to their respective leaf switches and to the AS4610, which provides 1 GbE out-of-band management connectivity.

The Mellanox SN2100 or SN2100B leaf switches connect to the SN2700 Series spine via QSFP cables. Depending on the leaf-switch model, these connections could provide 40 Gbps or 100 Gbps throughput per uplink back to the spine. As a result, oversubscription ratios may vary.

Using the Nutanix 3060 Series, with a total of 40 nodes or 10 hyperconverged blocks per rack, this deployment supports up to 480 servers across 12 racks.

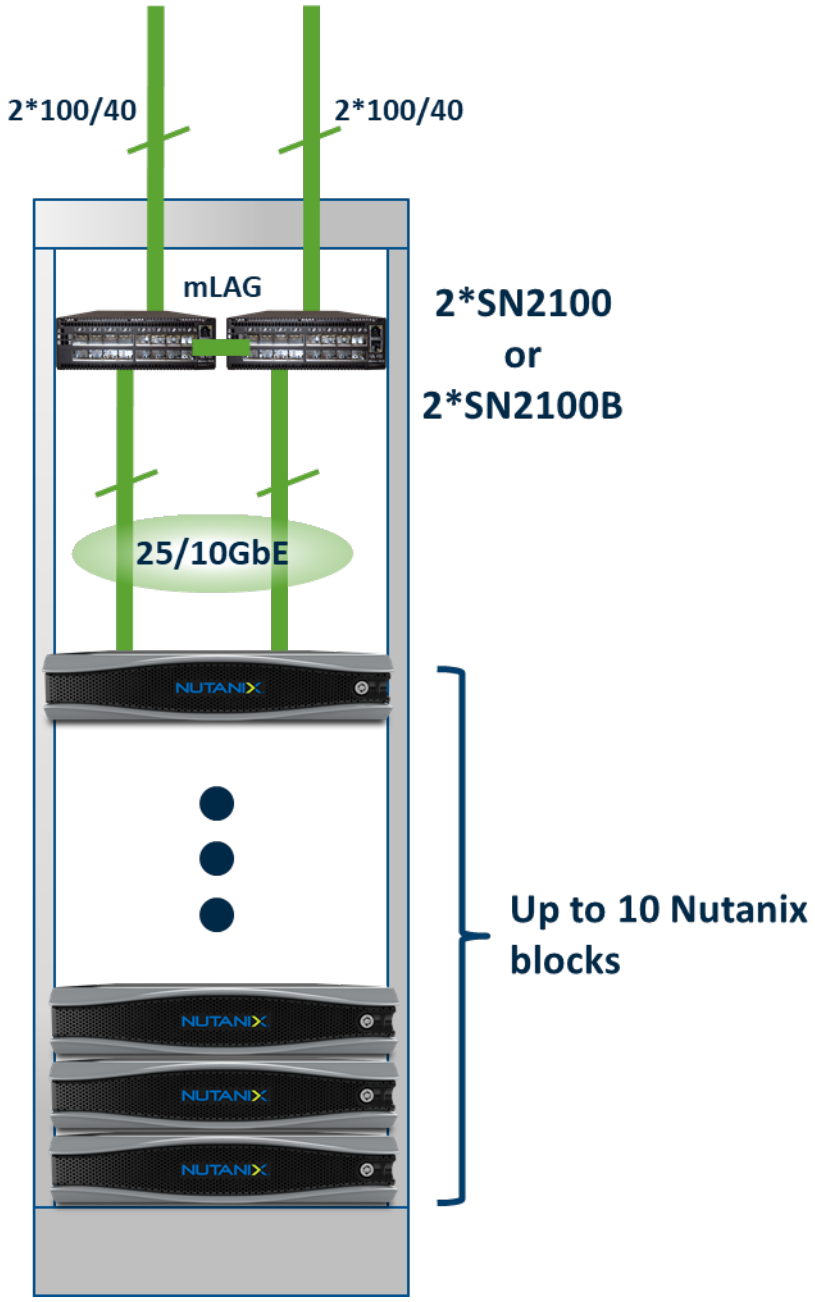

Figure 10: Medium-Density Node Configuration

Leaf Switch Density Calculations: Small to Medium

- 10 Nutanix 3060 G6 blocks in each rack (4 nodes per block) = 40 nodes per rack.
- Because each node contains 2x 10 GbE ports, you need a total of 8x 10 GbE ports per four-node block, with a total of 80 ports per rack. Because each SN2100 Series switch contains 16 ports, when we use 2 of them as leaf switches (SN2100 or SN2100B), we can meet our connectivity requirements by using a Mellanox QSFP-to-4xSFP+ cable to convert a single 100 GbE or 40 GbE port to 4x 10 GbE ports; 20 of these cables thus meet our host connectivity requirements of 80 ports. 10 ports with QSFP-to-4xSFP+ cables from each leaf switch provide 40x 10 GbE uplinks per switch, or 80 uplinks total between both switches. This configuration leaves 6 ports on each leaf switch.
- Two additional ports from our leaf switches form an MLAG peering between the pair, while two more ports uplink to their spine switch.
  - # Now, each leaf switch has two spare ports available.
- We need 40x 1 GbE ports to satisfy our out-of-band connectivity requirements. One AS4610 switch provides 48x 1 GbE connectivity per switch as well as 4x 10 GbE ports per switch. We use two of these 4x 10 GbE ports for establishing uplinks to the respective spine switches.

Spine Density Calculations: Small to Medium

- The spine, consisting of two SN2700 Series switches, contains 32x 100 GbE ports and the ability to convert a 100 GbE port into a 10, 25, 40, 50, or 56 GbE port using the appropriate QSFP+ Optic breakout cable.
- To satisfy the connectivity requirements for each rack, we need four 40 GbE or 100 GbE ports per rack (for the two leaf switches) and two 10 GbE ports (for the one out-of-band AS4610 switch), all of which are trunked to our Mellanox SN2700 spine switches.
- Subtracting the two ports required for the MLAG between our SN2700 Series switches (2x 100 GbE or 2x 40 GbE), each switch has 30x 100 GbE or 40 GbE ports available for leaf connectivity.
- Therefore, scaling the solution to 12 racks, with 2 ports per rack, requires 24 switch ports per spine for leaf connectivity, leaving 6 ports available per spine. Of these 6, we need 3 ports to provide out-of-band switch connectivity utilizing the QSFP-to-4xSFP+ breakout cables (3 x 4 = 12 ports), so we have 3 ports remaining per spine.

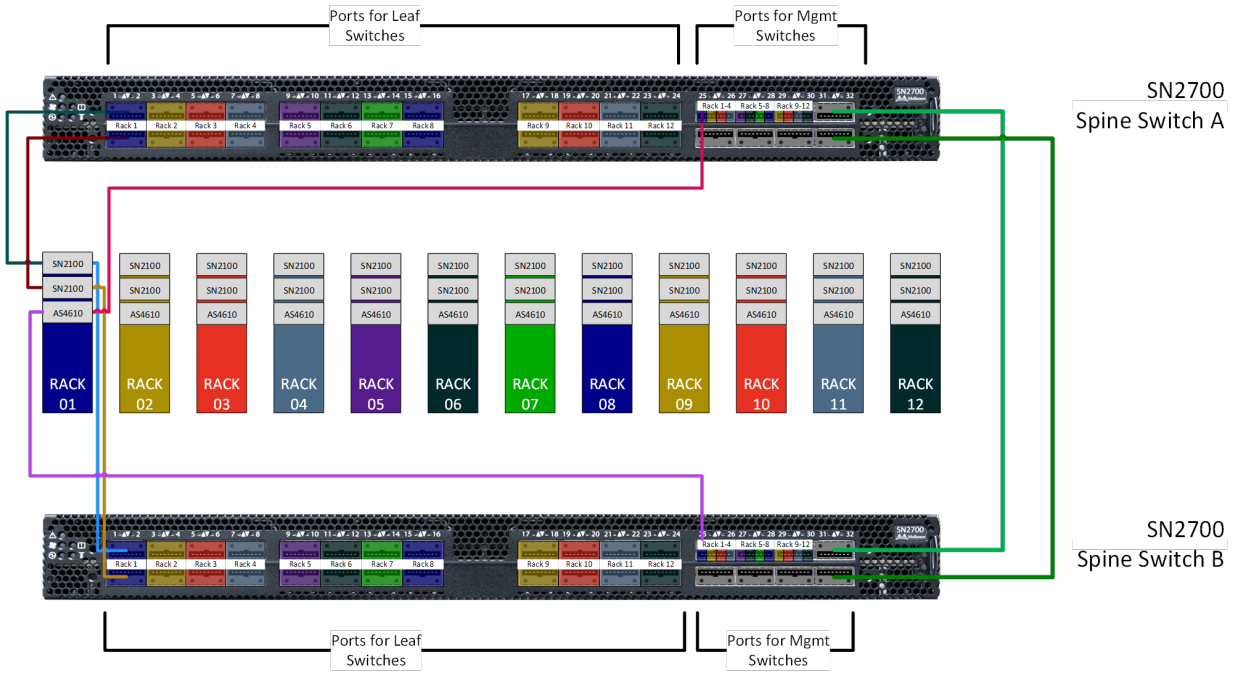

Figure 11: Medium-Density Spine Configuration

4.3. Scalable Architecture 3. Medium to Large: 720 Nodes

The following figure is a high-level diagram for a medium- to large-scale deployment. In this case, we use two Mellanox 2700 Series switches for each leaf and another two for the spine and two AS4610 switches for out-of-band management connectivity.

Mellanox SN2700 Series switches accommodate the highest rack performance in a condensed 1RU footprint. The SN2700 Series is an ONIE (Open Network Install Environment) platform, so you can mount a variety of operating systems on it and take advantage of open networking and the capabilities of the Mellanox Spectrum ASIC.

This series is available in two primary configurations: SN2700 offers 32x 40, 56, or 100 GbE nonblocking ports, while SN2700B offers 32x 10 or 40 GbE nonblocking ports.

This solution starts with a single rack (containing a minimum of three hyperconverged appliances or nodes) and can scale to 12 racks. Each rack includes full 10 GbE redundant connectivity, with 1 GbE connections for out-of-band management.

Nutanix appliances connect to their respective leaf switches via QSFP-to-4xSFP+ splitter cables (10 GbE) as well as two AS4610 switches, which provide 1 GbE out-of-band management connectivity.

The Mellanox SN2700 Series leaf switches connect to the spine via QSFP cables. Depending on the leaf-switch model, these connections could provide 40 Gbps or 100 Gbps throughput per uplink back to the spine, so oversubscription ratios may vary.

Using the Nutanix 3060 Series, with a total of 60 servers or 15 hyperconverged blocks per rack, this deployment supports up to 720 servers across 12 racks.

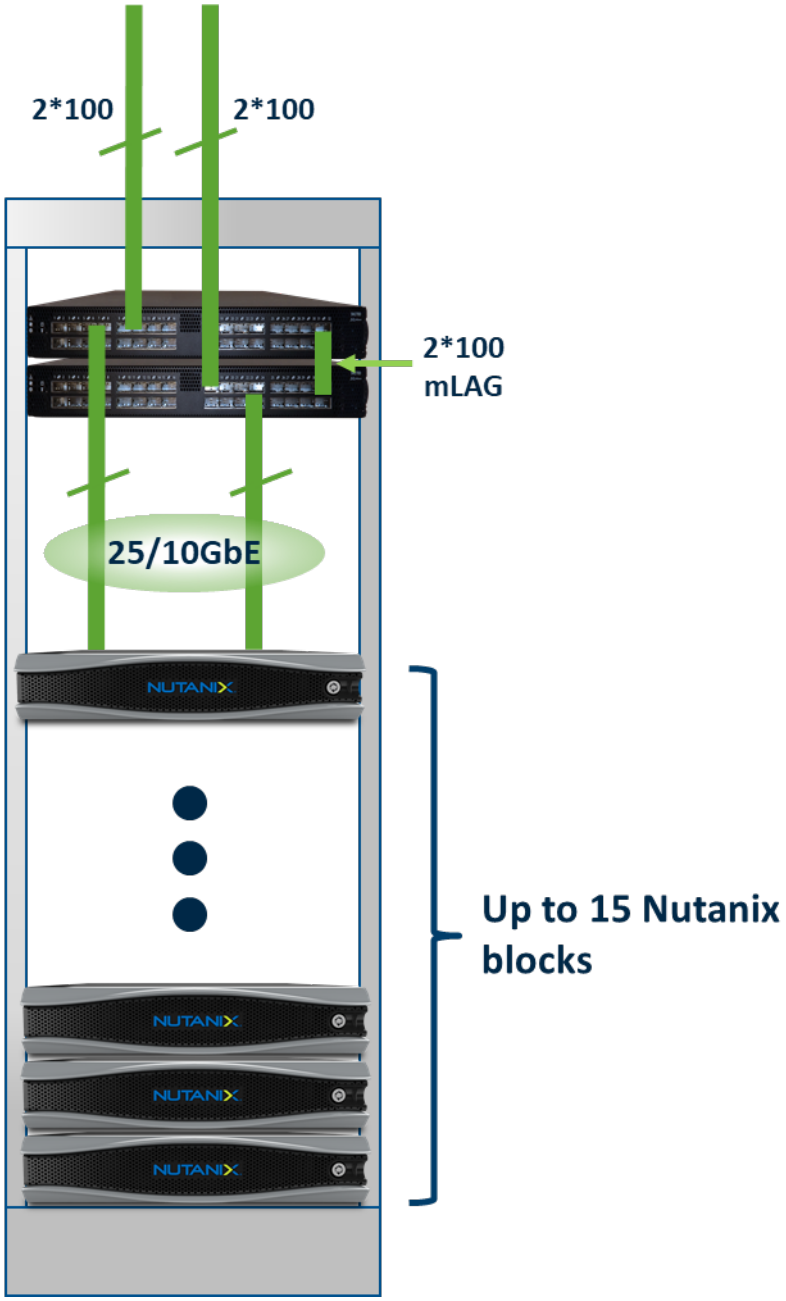

Figure 12: Medium- to Large-Density Node Configuration

Leaf Switch Density Calculations: Medium to Large

- 15 Nutanix 3060 G6 blocks in each rack (4 nodes per block) = 60 nodes per rack.

- Because each node contains 2x 10 GbE ports, you need 8x 10 GbE ports per block, for a total of 120 ports per rack (60 nodes, with two ports per node). Because each Mellanox SN2700 Series switch contains 32 ports, when we use 2 of them as leaf switches, we can meet our connectivity requirements by using a Mellanox QSFP-to-4xSFP+ cable to convert a single 100 GbE or 40 GbE port to 4x 10 GbE ports; 30 of these cables thus meet our host connectivity requirement for 120 ports. Using 15 ports per leaf leaves 17 ports remaining per leaf switch.
- Two additional ports from our leaf switches form an MLAG peering between the pair, while two more ports uplink to their spine switch.
  # As a result, each leaf switch has 13 spare ports available.
- We require 60x 1 GbE ports to satisfy our out-of-band connectivity requirements for 60 nodes. Two AS4610 switches provide 48x 1 GbE connectivity per switch (96x 1 GbE total) as well as 4x 10 GbE ports per switch. We use 2 of these 4x 10 GbE ports to establish our MLAG peer between the AS4610 switches and the other 2 for our uplinks to the respective spine switches.

Spine Density Calculations: Medium to Large

- The spine, consisting of SN2700 Series switches, contains 32x 100 GbE ports and the ability to convert a 100 GbE port into a 10, 25, 40, 50, or 56 GbE port using the appropriate QSFP+ Optic breakout cable.
- To satisfy the connectivity requirements for each rack, we need four 40 GbE or 100 GbE ports per rack (for the two leaf switches) and four 10 GbE ports (for the two out-of-band AS4610 switches), all of which are trunked to our Mellanox SN2700 Series spine switches.
- Subtracting the two ports required for the MLAG between our SN2700 Series switches (2x 100 GbE or 2x 40 GbE), each switch has 30x 100 GbE or 40 GbE ports available for leaf connectivity.
- Therefore, scaling the solution to 12 racks requires 24 switch ports per spine for leaf connectivity, plus 6 ports for out-of-band switch connectivity utilizing the QSFP-to-4xSFP+ breakout cables (6 x 4 = 24 ports).
- Scaling the solution to 12 racks uses all available ports on the spines.

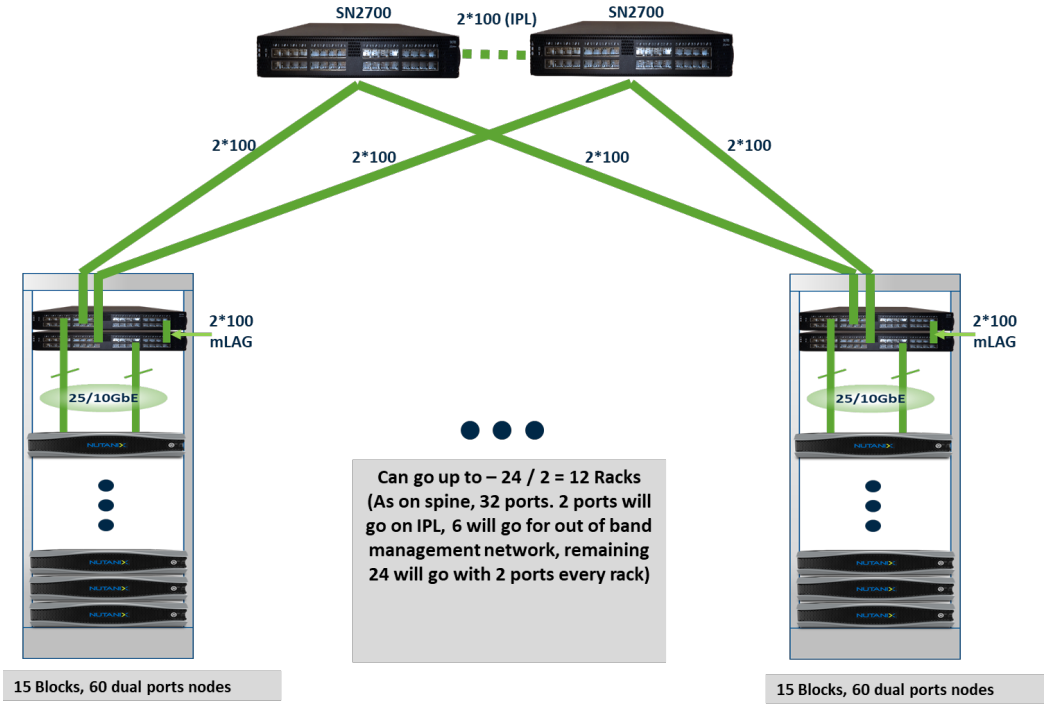

Figure 13: Medium- to Large-Density Spine Configuration

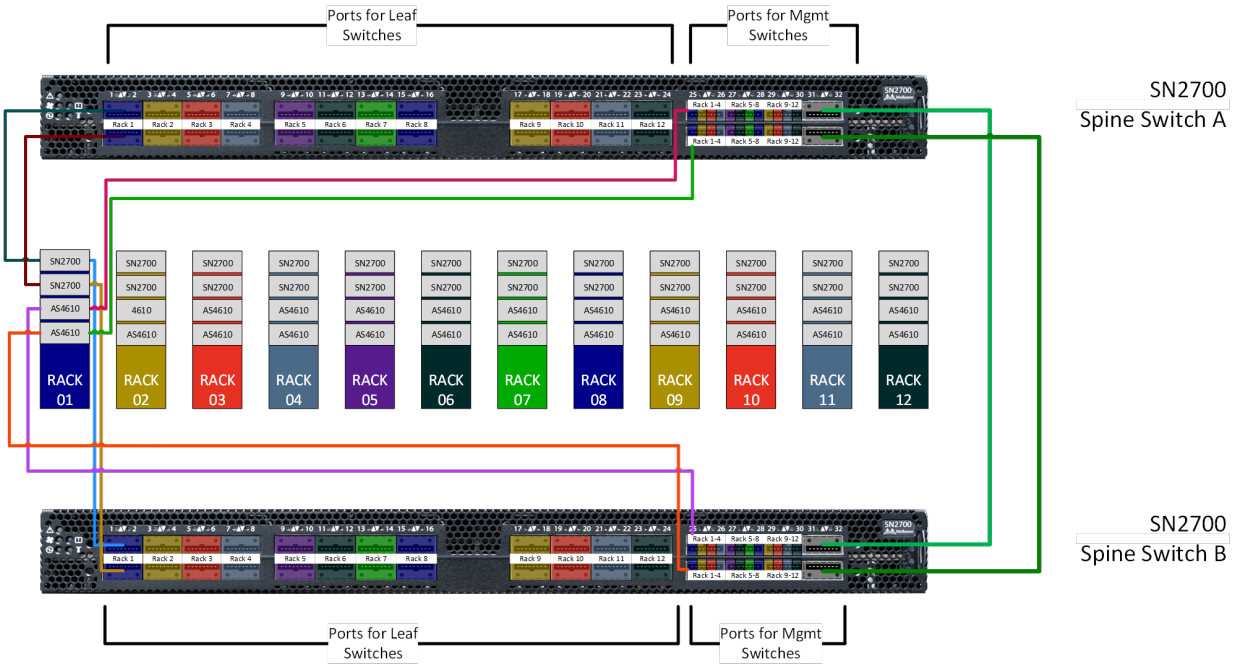

Figure 14: Medium- to Large-Density Spine Calculation

4.4. Spine Scalability

When considering either topology (small-medium or medium-large), it is important to understand conditions that require you to add more spine switches to the network.

- Port density

  Because each leaf switch requires connectivity back to each spine, spine switch port density is a limiting factor. Therefore, when the spine no longer has any spare ports available, you should introduce a new spine switch. When extending the spine, keep in mind that each existing leaf switch must have spare ports available to connect to the newly introduced spine switch.

- Oversubscription

  When the oversubscription rates between leaf switches or network traffic conditions exceed the bandwidth capacity between the leaf and spine switches (for example, 40 Gbps), you need to deploy additional spine switches. A key advantage of the Nutanix Enterprise Cloud is data locality. Data locality minimizes east-west traffic between nodes by keeping a VM's data close to its compute, so all read operations occur locally.

Architecting a Nutanix cluster within the boundaries of two leaf switches can further reduce the likelihood of oversubscription becoming an issue. This design contains storage traffic within the failure domain of a single rack, significantly reducing the traffic load on the spine switches.

5. Conclusion

Leaf-spine architectures offer lower management overhead than traditional three-tier network infrastructures, thanks to complementary technology and integration with automation platforms such as Salt, Ansible, Chef, or Puppet. Not only do management advances help deliver consistent configuration, they also reduce the manual configuration of multiple switches. Leaf-spine networks provide predictability and low-latency switching, while achieving maximum throughput and linear scalability.

Low latency and high throughput networking is especially crucial with NVMe-based SSDs and Intel's Optane Technology raising the bar. These devices need high performance networks to perform optimally without impacting other applications and services.

Deploying a Mellanox solution future-proofs your network, ensuring that it can support advances in network interface cards (NICs) beyond the scope of 10 GbE NICs (to 25, 40, 50, or 100 GbE and beyond), and expansion to multiclouds with EVPN-based layer 3 networking.

Coupled with a software-defined networking solution, Mellanox network switches offer such benefits as manageability, scalability, performance, and security to deliver a unified network architecture with lower opex.

Appendix

Terminology

For a detailed explanation of terminology involved in Mellanox configuration, including interpeer link (IPL), MLAG cluster, MLAG interface, and virtual system ID (VSID), visit the Mellanox community site.

Product Details

Table 3: Product Details

Mellanox Part Number	Description	Technical Description	Details
MC2609130-003	Splitter cable	40 Gbps to 4x 10 Gbps QSFP+ to 4xSFP+	Product spec
MCP1600-C001	100 GbE cable	100 Gbps QSFP28	Product spec
MFA1A00-C030	100 GbE optics cable	100 Gbps QSFP28	Product spec
MC2210130-001	40 GbE cable	40 Gbps QSFP+	Product Spec
MC2210411-SR4L	Optical module	40 Gbps QSFP+	Product spec
MSN2100-BB2F	SN2100 switch	Spectrum-based half switch	Product spec
MSN2700-BS2F	SN2700 switch	Spectrum-based full switch	Product spec

Mellanox Part Number	Description	Technical Description	Details
AS4610	Edge core AS4610	Out-of-band management switch	Product spec

Bills of Materials

Table 4: Small to Medium Density with SN2100 and 100 GbE Uplinks

Mellanox OPN	Quantity (Per Rack)	Details
MSN2100-CB2F	2	SN2100
MTEF-KIT-D	1	Rail installation kit
SUP-SN2100-1S	1	One-year Mellanox technical support and warranty for SN2100
MC2609130-003	20	QSFP-to-4xSFP+ splitter cables: 3 m passive copper (10 GbE cables to Nutanix nodes)
MCP1600-C001	2	Passive copper cable: ETH 100 GbE, 100 Gbps, QSFP, 1 m (for MLAG peering)
MMA1B00-C100D	4	Optical module for 100 GbE uplinks: QSFP28, up to 100 m (for uplink)
Edge core AS4610	1	1 GbE out-of-band management switch (with required RJ45 cables and 10 GbE uplink cables), RJ45 cables as needed

Table 5: Small to Medium Density with SN2100B and 40 GbE Uplinks

Mellanox OPN	Quantity (Per Rack)	Details
MSN2100-BB2F	2	SN2100B
MTEF-KIT-D	1	Rail installation kit

Mellanox OPN	Quantity (Per Rack)	Details
SUP-SN2100-1S	1	One-year Mellanox technical support and warranty for SN2100
MC2609130-003	20	QSFP-to-4xSFP+ splitter cables: 3 m passive copper (10 GbE cables to Nutanix nodes)
MC2210130-001	2	QSFP cables: ETH 40 GbE, 40 Gbps, 1 m passive copper (for MLAG peering)
MC2210411-SR4L	4	QSFP cables: ETH 40 GbE, 40 Gbps, up to 30 m (for uplink)
Edge core AS4610	1	1 GbE out-of-band management switch (with required RJ45 cables and 10 GbE uplink cables), RJ45 cables as needed

Table 6: Medium to Large Density with SN2700 and 100 GbE Uplinks

Mellanox OPN	Quantity (Per Rack)	Details
MSN2700-CS2F	2	SN2700 with rail kits
SUP-SN2700-1S	1	One-year Mellanox technical support and warranty for SN2700
MC2609130-003	120	QSFP-to-4xSFP+ splitter cables: 3 m passive copper (10 GbE cables to Nutanix nodes)
MCP1600-C001	2	100 GbE cables between SN2700s: QSFP28 cables, 1 m passive copper (for MLAG peering).
MMA1B00-C100D	4	Optical module for 100 GbE uplinks: QSFP28, up to 100 m (for uplink)

Mellanox OPN	Quantity (Per Rack)	Details
Edge core AS4610	1	1 GbE out-of-band management switch (with required RJ45 cables and 10 GbE uplink cables), RJ45 cables as needed

Configurations Using Mellanox NEO

For more information on how to create an MLAG switch pair using Mellanox NEO, please see the related article, which can be found on the Mellanox community site.

MLAG Configuration Planning

This section can help you plan out a MLAG configuration for your own deployment. Before you start the configuration itself, design your network. The following table presents a list of the general parameters needed for MLAG service.

Table 7: Parameters Needed for MLAG Service

Parameter	Description	Example
Name	Use any name (4–20 characters). Should be unique if you have more than one MLAG in your network.	MLAG
Description	Use any description (in text).	MLAG-Service
Port channel	IPL port channel (for example, on ports 1/35–1/36 on both switches). Use any number in the range 1–65,335.	1

Parameter	Description	Example
VLAN ID	The IPL VLAN ID. Use any VLAN ID other than the default VLAN (normally VLAN ID 1). Note: It is possible to use VLAN ID 1, if you change the default VLAN on the switch to a different number.	2
Virtual system MAC	Virtual MAC to be used as the MLAG's virtual IP; used for LACP if enabled. Use any Unicast MAC address.	AA:AA:AA:AA:AA:AA
IPL port range	The range of ports used for the IPL (the switches are on one). Format: 1/<port>–1/<port>. The number of ports used depends on the level of high availability needed. We recommend using two or more links.	1/35–1/36
IPL peer port range	The range of ports used for the IPL (on the peer switch). Format: 1/<port>–1/<port>. The number of ports used depends on the level of high availability needed. We recommend using two or more links.	1/35–1/36
Device IP	The management IP for one of the switches.	10.20.2.43
Peer device IP	The peer switch management IP.	10.20.4.131

Parameter	Description	Example
MLAG virtual IP	The virtual IP of the MLAG switch pair. Assign an IP address from the management network subnet. In this example, the switches have an IP address in the range 10.20.X.X/16. Do not assign this address to any other network element.	10.20.2.150
MLAG virtual IP mask	The mask of the management subnet. In this example, it is /16.	16
IPL IP address	This IP address (assigned to one of the switches) is internal, for passing MLAG control packets between the switches. Use any IP. This IP is not distributed externally or outside of the switches, but it should not be part of any other switch addressing.	1.1.1.1
IPL IP address mask	As there are only two addresses, a subnet mask of /30 could work here (four addresses).	30

Parameter	Description	Example
Peer IPL IP address	This IP address (assigned to the peer switch) is internal, for passing MLAG control packets between the switches. Use any IP. This IP is not distributed externally or outside of the switches, but it should not be part of any other switch addressing.	1.1.1.2

References

1. Mellanox Scale-Out Open Ethernet Products
2. HowTo Setup High Availability on ESXi 5.5 with Mellanox Adapters and Switches
3. Mellanox LinkX Ethernet DAC Splitters Cables
4. Network Load Balancing with Acropolis Hypervisor
5. How To Configure MLAG on Mellanox Switches
6. HowTo Enable MLAG Switch Pair Using Mellanox NEO
7. Mellanox Breakout Cables 40G > 4x10G and 100G > 4x25G
8. Mellanox SN2100 Open Ethernet Switch Datasheet
9. Mellanox SN2700 Open Ethernet Switch Datasheet

About Nutanix

Nutanix makes infrastructure invisible, elevating IT to focus on the applications and services that power their business. The Nutanix Enterprise Cloud OS leverages web-scale engineering and consumer-grade design to natively converge compute, virtualization, and storage into a resilient, software-defined solution with rich machine intelligence. The result is predictable performance, cloud-like infrastructure consumption, robust security, and seamless application mobility for a broad range of enterprise applications. Learn more at www.nutanix.com or follow us on Twitter @nutanix.

List of Figures

List of Tables