Copyright

Copyright 2018 Nutanix, Inc.

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

All rights reserved. This product is protected by U.S. and international copyright and intellectual property laws.

Nutanix is a trademark of Nutanix, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.
1. Executive Summary

This solution note addresses the fundamental differences between traditional three-tier and modern leaf-spine networking architectures, while also detailing 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 5 switches and 3 servers up to 50 switches and 720 servers. Nutanix nodes are equipped with redundant 10 GbE and 1 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 when scaling Nutanix hyperconverged appliances.

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

<table>
<thead>
<tr>
<th>Version Number</th>
<th>Published</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>January 2017</td>
<td>Original publication.</td>
</tr>
<tr>
<td>1.1</td>
<td>August 2017</td>
<td>Updated platform overview and added references specific to ESXi 5.5 and AHV.</td>
</tr>
<tr>
<td>1.2</td>
<td>September 2018</td>
<td>Updated Nutanix overview.</td>
</tr>
</tbody>
</table>
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 three third-party hypervisors—ESXi, Hyper-V, and XenServer—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 with the Enterprise Cloud

Traditional datacenter networks have a three-layer topology:

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

![Traditional Network Tiers](image)

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 fairly 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 within 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 figure below). L2 traffic segments are contained within 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 and a VM that provides the software-defined storage logic. Clustering all servers using IP creates a single unified storage pool (the aggregate of all storage devices across appliances) and allows hosts within the cluster to access it.

Because of these two industry trends, traditional north-south flows now move in an east-west pattern, because we have hosts spanning multiple access layers and racks, utilizing the existing network for both storage and application services. Continuing to use a three-tier networking model can thus limit a network’s performance and scalability, affecting the overall solution.

### 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.
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 utilizes multiple access layer leaf switches that connect directly to every spine, providing horizontal scalability without the limitations of a spanning tree.
You can use leaf switches for either L2 or L3 service. In an L2 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 L3 service, each link becomes routed. With routed links, overlay technologies like VXLANs (Virtual Extensible Local Area Networks) or VLANs (Virtual Local Area Networks) can increase efficiency when routing packets, because traffic no longer needs to traverse the spine for L3 services.

Overlay technologies also 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 utilizing Mellanox SN2xxx Series switches. This reference architecture consists of the hardware and software components listed in the table below.
### 3.4. Mellanox Setup and Deployment

Mellanox switches allow you to create a network fabric that offers predictable, low-latency switching while achieving maximum throughput and linear scalability. Combined with the features and intelligence of the Mellanox operating system (OS), multilink aggregation groups (MLAGs) create a highly available L2 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 and actively load balances user VMs on TCP streams. 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. The Mellanox Spectrum ASIC (application-specific integrated circuit) delivers 100 GbE port speed with the industry’s lowest port-to-port latency (approximately 300 ns, or about 0.6 us leaf to spine).

In the examples that follow, we deploy the leaf-spine topology using MLAGs. Managing and updating each switch independently with MLAGs mitigates the single point of failure that typically results from employing stacking techniques within the switches.

When stacking switches, administrators group individual switches into a set for holistic management using a single IP address. Although each switch is physically separate, they are logically grouped as a single entity, displaying the characteristics of a single switch but having the port capacity of the sum of the combined switches.

Some situations might be suited for switch stacking—for example, if you require multiple stacks and the host links to the stacks themselves are redundant. However, for most networking deployments, we recommend using MLAGs.

> **Note:** You cannot mix stacking and MLAGs together.
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 of 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 availability on ESXi 5.5 with Mellanox adapters and switches, consult the Mellanox KB article on the topic. To set up high availability 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 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

**General**

Split the ports on the SN2100 switches:

- Connect the breakout cable on both SN2100 switches.
- Connect the breakout cable on the 4x 10 GbE side to the Nutanix servers.
- Assuming we are splitting from port Eth 1/1 to Eth 1/10, run the following commands on both SN2100 switches:

```
Switch (config) # interface ethernet 1/1-1/10 shutdown
Switch (config) # interface ethernet 1/1-1/10 module-type qsfp-split-4 force
```

- 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 within this LAG. In this example, we are configuring the IPL on ports 1/15 and 1/16 of two SN2100 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 1/15 channel-group 1 mode active
  Switch (config) # interface ethernet 1/16 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 within the same subnet as the management interface (mgmt0).
• Configure the following on both switches:
  Switch (config) # mlag-vip my-mlag-vip-domain ip 10.209.20.200 /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 40 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 1-40
  Switch (config) # mtu 9216 force
  Switch (config interface port-channel 1-2 ) # exit

• On switch A:
  Switch-A (config) # interface ethernet 1/1/1 mlag-channel-group 1 mode on
  Switch-A (config) # interface ethernet 1/1/2 mlag-channel-group 2 mode on
  Switch-A (config) # interface ethernet 1/1/3 mlag-channel-group 3 mode on
  Switch-A (config) # interface ethernet 1/1/4 mlag-channel-group 4 mode on
  Switch-A (config) # interface ethernet 1/2/1 mlag-channel-group 5 mode on
  Switch-A (config) # interface ethernet 1/2/2 mlag-channel-group 6 mode on
  Switch-A (config) # interface ethernet 1/2/3 mlag-channel-group 7 mode on
  Switch-A (config) # interface ethernet 1/2/4 mlag-channel-group 8 mode on
  Switch-A (config) # interface ethernet 1/10/1 mlag-channel-group 37 mode on
  Switch-A (config) # interface ethernet 1/10/2 mlag-channel-group 38 mode on
  Switch-A (config) # interface ethernet 1/10/3 mlag-channel-group 39 mode on
  Switch-A (config) # interface ethernet 1/10/4 mlag-channel-group 40 mode on
  Switch-A (config) # interface mlag-port-channel 1-24 no shutdown
On switch B:

Switch-B (config) # interface ethernet 1/1/1 mlag-channel-group 1 mode on
Switch-B (config) # interface ethernet 1/1/2 mlag-channel-group 2 mode on
Switch-B (config) # interface ethernet 1/1/3 mlag-channel-group 3 mode on
Switch-B (config) # interface ethernet 1/1/4 mlag-channel-group 4 mode on
Switch-B (config) # interface ethernet 1/2/1 mlag-channel-group 5 mode on
Switch-B (config) # interface ethernet 1/2/2 mlag-channel-group 6 mode on
Switch-B (config) # interface ethernet 1/2/3 mlag-channel-group 7 mode on
Switch-B (config) # interface ethernet 1/2/4 mlag-channel-group 8 mode on
Switch-B (config) # interface ethernet 1/10/1 mlag-channel-group 37 mode on
Switch-B (config) # interface ethernet 1/10/2 mlag-channel-group 38 mode on
Switch-B (config) # interface ethernet 1/10/3 mlag-channel-group 39 mode on
Switch-B (config) # interface ethernet 1/10/4 mlag-channel-group 40 mode on
Switch-B (config) # interface mlag-port-channel 1-24 no shutdown

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

**Uplinks**

In the example above, you can use port numbers 1/13 and 1/14 for uplinks to connect to the spine layer. You can use port numbers 1/11 and 1/12 for an out-of-band management switch MLAG.
4. Example Deployment Scenarios

Combining Mellanox SN2100 and SN2700 switches allows you to create a scalable network design that provides predictable and consistent low latency and high throughput from end to end in the network. Configuring the MLAG is a fairly 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 that connect to it.

4.1. Scalable Architecture 1. Small to Medium: 480 Nodes

The following figure is a high-level diagram for a small- or 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. We also use an AS4610 switch for 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, using QSFP-to-4xSFP+ splitter cables (10 GbE).

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.
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. As each SN2100 Series switch contains 16 ports, when we utilize 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 10, 25, 40, 50, or 56 GbE 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.
4.2. Scalable Architecture 2. 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. We also use 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 utilize the advantages 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 would include 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.

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 utilize 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 10, 25, 40, 50, or 56 GbE using the appropriate QSFP+ Optic breakout cable.

- To satisfy the connectivity requirements for each rack, we need 4x 40 GbE or 100 GbE ports per rack (for the two leaf switches) and 4x 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.
Can go up to – 24 / 2 = 12 Racks
(As on spine, 32 ports. 2 ports will go on IPL, 6 will go for out of band management network, remaining 24 will go with 2 ports every rack)

Figure 10: Medium- to Large-Density Spine Configuration
4.3. 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, thus 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. Designs based on this fundamental framework provide predictability and low-latency switching, while achieving maximum throughput and linear scalability.

Throughput and scalability are especially crucial in light of how SSD technology is transforming performance within the datacenter. With the recent introduction of NVMe-based SSDs and Intel’s Optane™ Technology, the network must be able to supply sufficient bandwidth for these devices to perform optimally without impacting existing applications or services.

With these innovations in mind, investing in and deploying a Mellanox solution future-proofs your network, ensuring that it can support advances in network interface cards beyond the scope of 10 GbE NICs (to 25, 40, 50, or 100 GbE and beyond).

Coupled with a software-defined networking solution, Mellanox network switches offer such benefits as manageability, scalability, performance, and security, while delivering 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>
<thead>
<tr>
<th>Table 3: Product Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mellanox Part Number</strong></td>
</tr>
<tr>
<td>MC2609130-003</td>
</tr>
<tr>
<td>MCP1600-C001</td>
</tr>
<tr>
<td>MFA1A00-C030</td>
</tr>
<tr>
<td>MC2210130-001</td>
</tr>
<tr>
<td>MC2210411-SR4L</td>
</tr>
<tr>
<td>MSN2100-BB2F</td>
</tr>
<tr>
<td>MSN2700-BS2F</td>
</tr>
</tbody>
</table>
### Bills of Materials

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

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

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

<table>
<thead>
<tr>
<th>Mellanox OPN</th>
<th>Quantity (Per Rack)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSN2100-BB2F</td>
<td>2</td>
<td>SN2100B</td>
</tr>
<tr>
<td>MTEF-KIT-D</td>
<td>1</td>
<td>Rail installation kit</td>
</tr>
<tr>
<td>Mellanox OPN</td>
<td>Quantity (Per Rack)</td>
<td>Details</td>
</tr>
<tr>
<td>-------------------</td>
<td>--------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>SUP-SN2100-1S</td>
<td>1</td>
<td>One-year Mellanox technical support and warranty for SN2100</td>
</tr>
<tr>
<td>MC2609130-003</td>
<td>20</td>
<td>QSFP-to-4xSFP+ splitter cables: 3 m passive copper (10 GbE cables to Nutanix nodes)</td>
</tr>
<tr>
<td>MC2210130-001</td>
<td>2</td>
<td>QSFP cables: ETH 40 GbE, 40 Gbps, 1 m passive copper (for MLAG peering)</td>
</tr>
<tr>
<td>MC2210411-SR4L</td>
<td>4</td>
<td>QSFP cables: ETH 40 GbE, 40 Gbps, up to 30 m (for uplink)</td>
</tr>
<tr>
<td>Edge core AS4610</td>
<td>1</td>
<td>1 GbE out-of-band management switch (with required RJ45 cables and 10 GbE uplink cables), RJ45 cables as needed</td>
</tr>
</tbody>
</table>

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

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

Appendix   |   31

### Mellanox OPN

<table>
<thead>
<tr>
<th>Mellanox OPN</th>
<th>Quantity (Per Rack)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge core AS4610</td>
<td>1</td>
<td>1 GbE out-of-band management switch (with required RJ45 cables and 10 GbE uplink cables), RJ45 cables as needed</td>
</tr>
</tbody>
</table>

### 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 table below presents a list of the general parameters needed for MLAG service.

#### Table 7: Parameters Needed for MLAG Service

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Use any name (4–20 characters). Should be unique if you have more than one MLAG in your network.</td>
<td>MLAG</td>
</tr>
<tr>
<td>Description</td>
<td>Use any description (in text).</td>
<td>MLAG-Service</td>
</tr>
<tr>
<td>Port channel</td>
<td>IPL port channel (for example, on ports 1/35–1/36 on both switches).</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Use any number in the range 1–65335.</td>
<td></td>
</tr>
<tr>
<td>Parameter</td>
<td>Description</td>
<td>Example</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>VLAN ID</td>
<td>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.</td>
<td>2</td>
</tr>
<tr>
<td>Virtual system MAC</td>
<td>Virtual MAC to be used as the MLAG’s virtual IP; used for LACP if enabled. Use any Unicast MAC address.</td>
<td>AA:AA:AA:AA:AA:AA</td>
</tr>
<tr>
<td>IPL port range</td>
<td>The range of ports used for the IPL (the switches are on one). Format: 1/&lt;port&gt;–1/&lt;port&gt;. The number of ports used depends on the level of high availability needed. We recommend using two or more links.</td>
<td>1/35–1/36</td>
</tr>
<tr>
<td>IPL peer port range</td>
<td>The range of ports used for the IPL (on the peer switch). Format: 1/&lt;port&gt;–1/&lt;port&gt;. The number of ports used depends on the level of high availability needed. We recommend using two or more links.</td>
<td>1/35–1/36</td>
</tr>
<tr>
<td>Device IP</td>
<td>The management IP for one of the switches.</td>
<td>10.20.2.43</td>
</tr>
<tr>
<td>Peer device IP</td>
<td>The peer switch management IP.</td>
<td>10.20.4.131</td>
</tr>
<tr>
<td>Parameter</td>
<td>Description</td>
<td>Example</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>MLAG virtual IP</td>
<td>The virtual IP of the MLAG switch pair. Assign an IP address from the management network subnet.</td>
<td>10.20.2.150</td>
</tr>
<tr>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>MLAG virtual IP mask</td>
<td>The mask of the management subnet. In this example, it is /16.</td>
<td>16</td>
</tr>
<tr>
<td>IPL IP address</td>
<td>This IP address (assigned to one of the switches) is internal, for passing MLAG control packets between the switches.</td>
<td>1.1.1.1</td>
</tr>
<tr>
<td></td>
<td>Use any IP.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>This IP is not distributed externally or outside of the switches, but it should not be part of any other switch addressing.</td>
<td></td>
</tr>
<tr>
<td>IPL IP address mask</td>
<td>As there are only two addresses, a subnet mask of /30 could work here (four addresses).</td>
<td>30</td>
</tr>
</tbody>
</table>
### 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. HowTo 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

Figure 1: Nutanix Enterprise Cloud................................................................. 5
Figure 2: Traditional Network Tiers............................................................... 7
Figure 3: Traditional Three-Tier Network with North-South Traffic............. 8
Figure 4: Spanning Tree Port Blocking......................................................... 10
Figure 5: Leaf-Spine Network...................................................................... 11
Figure 6: Configuring MLAGs on Mellanox................................................. 14
Figure 7: Small- to Medium-Density Node Configuration.............................. 19
Figure 8: Small- to Medium-Density Spine Configuration............................. 21
Figure 9: Medium- to Large-Density Node Configuration............................. 22
Figure 10: Medium- to Large-Density Spine Configuration......................... 24
Figure 11: Medium- to Large-Density Spine Calculation............................... 25
List of Tables

Table 1: Document Version History .................................................................................. 4
Table 2: Hardware and Software Components ............................................................... 12
Table 3: Product Details .................................................................................................. 28
Table 4: Small to Medium Density with SN2100 and 100 GbE Uplinks ......................... 29
Table 5: Small to Medium Density with SN2100B and 40 GbE Uplinks ......................... 29
Table 6: Medium to Large Density with SN2700 and 100 GbE Uplinks ......................... 30
Table 7: Parameters Needed for MLAG Service ............................................................. 31