Lenovo Networking Best Practices for CNOS: Network Design and Topologies

Recommends physical and logical topologies for networks

Includes discussion of network robustness and redundancy

Includes examples of Lenovo "Easy Connect" configuration

Includes discussion of isolated management networks

Scott Lorditch

LENOVO PRESS

Abstract

This paper shows recommended network topologies with a focus on avoiding the use of Spanning Tree and providing maximal bandwidth and redundancy to serve a cluster of servers in a data center environment. The examples shown rely on the use of the CNOS firmware in Lenovo® top-of-rack switches. This paper is suitable for network designers and architects involved in designing or modifying a data center.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.

See a list of our most recent publications at the Lenovo Press web site:

http://lenovopress.com

Do you have the latest version? We update our papers from time to time, so check whether you have the latest version of this document by clicking the **Check for Updates** button on the front page of the PDF. Pressing this button will take you to a web page that will tell you if you are reading the latest version of the document and give you a link to the latest if needed. While you're there, you can also sign up to get notified via email whenever we make an update.

Contents

Introduction

This paper describes several network topologies that are frequently used by customers and can be used for future deployments. These topologies will involve those that include the use of embedded servers and switches but will not focus on those.

The presentation of each topology includes information about its merits and constraints, and about when it is most appropriate to use that topology. Most of the topologies that are shown can also be extended to include Fibre Channel over Ethernet (FCoE) if desired.

Most of the examples that are shown can be implemented by using a feature that is called *Easy Connect*. Several of the examples in this chapter show the addition of an Easy Connect element. For more information about what Easy Connect represents and its various iterations and interactions, see “Easy Connect” on page 10.

In general, the criteria for selecting the topology includes the following items:

- The capabilities of the devices in the customer’s network, which are immediately upstream from the embedded switches. This criterion includes the bandwidth of the network (1 Gb, 10 Gb, 25 Gb, 50 Gb or 100 Gb).
- The customer’s standards and practices for network interface card (NIC) teaming on their servers.
- The customer’s preferences for stacking and management of the embedded switches.

Full mesh topology with Virtual Link Aggregation

This topology is preferred (except for when conditions prevent its use, such as those conditions that described later in this section). It provides connectivity with the following qualities:

- High availability, which enables the environment to survive the failure of one of two access switches (server-connected switches), one of two upstream aggregation switches (or the links that connect to them), or both.
- All of the links between an access switch and an aggregation switch are active and can carry production traffic. None of the links are blocked by Spanning Tree to prevent network loops.
- The server-facing ports on the pair of access switches can be channeled together. Aggregation-based Active/active NIC teaming modes are available if the server’s ports are channeled in this way.
- This is the preferred topology for leaf switches and devices in a spine-leaf topology. Such topologies are discussed at length in these documents:
  - *Introduction to Spine-Leaf Networking Designs*

    https://lenovopress.com/lp0573
  - *Configuring Lenovo Networking Switches in a Spine-Leaf Topology*

    https://lenovopress.com/lp0930

Do not implement this topology if any of the following conditions are true:

- The upstream switches do not support a form of cross-switch link aggregation, such as vPC, VSS, Virtual Link Aggregation (vLAG), MCLAG, or a stacking feature. In this case,

© Copyright Lenovo 2019. All rights reserved.

other designs should be used — these are described starting in , “Inverted U topology with failover” on page 5.

- The customer does not have two switches that they plan to use to connect to the configuration. A single upstream switch design is not recommended because it contains a single point of failure and can isolate the entire cluster of servers and devices from the remainder of the network if that single switch fails.

This topology is shown in Figure 1. Equivalent designs can be deployed using a Flex System™ chassis if desired.

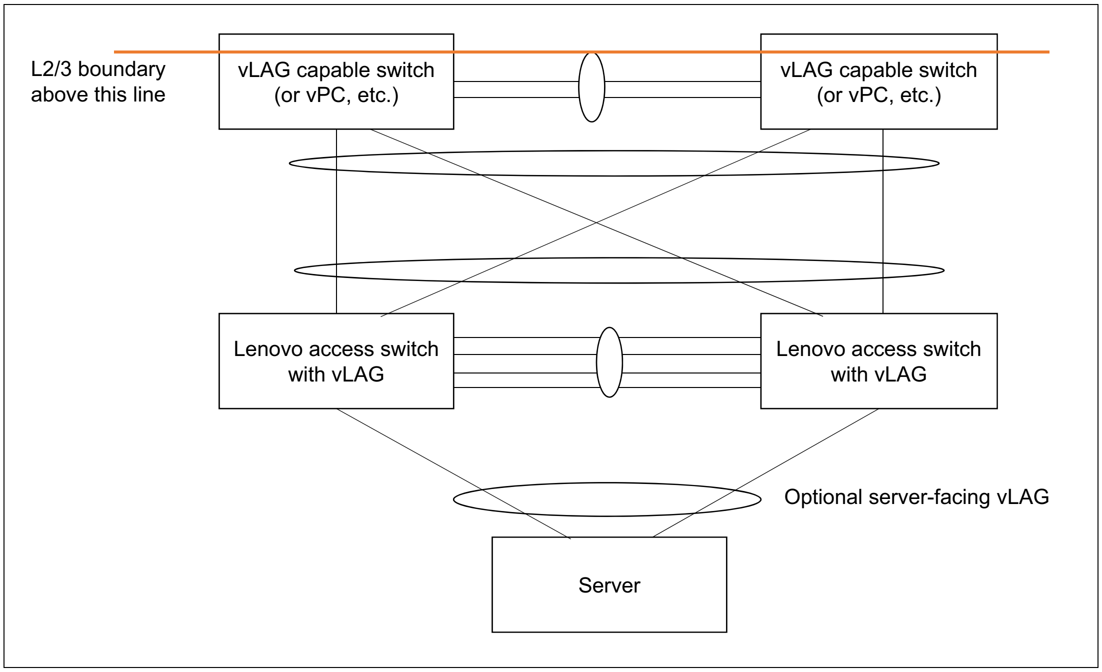

Figure 1 Full mesh network design

Key portions of the Lenovo switch configuration for this topology are shown in Figure 2. This configuration fragment shows the portions of the overall configuration that relates to vLAG on the access switches.

```
vlag enable
vlag tier-id 50
vlag isl port-channel 99
vlag instance 1 port-channel 10
vlag instance 1 enable
vlag instance 2 port-channel 20
vlag instance 2 enable
vlag hlthchk peer-ip 10.1.1.2 [vrf management]
```

Figure 2 Configuration commands for vLAG

Specific information for vLAG

When you are using vLAG in this topology or elsewhere, it is important to consider the requirements for a successful vLAG deployment. Such factors as proper sizing of the Interswitch Link (ISL), configuration of a health check network, and so on, are critical for a healthy vLAG environment.

In general, the ISL should carry a multiple of the bandwidth used for downlink connections to edge devices. Most traffic through a vLAG topology will not use the ISL except if there is a port or other outage, but the design should allow for this.

The use of a health check is not required but it is strongly recommended to avoid network problems if the ISL should fail and both switches remain active.

Inverted U topology with failover

This configuration is an equivalent for Ethernet to the common practice in SAN designs of having two parallel networks from the adapter, which is host bus adapter (HBA) for Fibre Channel, to the storage device. This configuration is often referred to as a SAN-A/SAN-B design. This design has a “left side” network that is tied to one NIC port on the server, flowing through one (“left side”) access switch, and connecting up to the left aggregation/core switch in the customer environment. The second NIC port similarly connects through the “right side” network.

Because there are two distinct switching devices at every tier, this topology cannot support active/active NIC teaming modes that are switch dependent (aggregation). It supports and should be used with, switch independent mode active/active teaming (for example, Linux bonding mode 5 or VMware ESXi default teaming mode, route based on originating virtual port ID), and all active/standby teaming modes such as Linux bonding mode 1.

This topology provides connectivity to the upstream network with the following qualities:

- High availability is provided in that the environment survives the failure of a single access switch or a single upstream aggregation switch.
- For servers in which NIC teaming or bonding is not used, this configuration does not provide L2 high availability. Serves that are not using NIC teaming or bonding can use other tools, such as local routing, to achieve high availability; however, in general, this architecture is most commonly used without teaming or bonding on the servers.
- Assuming the server is using teaming or bonding, the *failover* feature is required when this design is used and must be explicitly configured. The failover feature administratively disables server-facing (internal) ports when the external ports that connect the switch to its upstream neighbor fail. Without the failover feature, if the uplinks out of the switch failed, the server is unaware of this failure and continues to send traffic to the switch, which drops traffic when there is no upstream path available.

You might want to choose another design for the following reasons:

- High availability support for servers that are not configured with some form of NIC teaming (bonding) is not available. Other designs provide more robustness for servers that are configured in such a manner.
- You want and are using products that can provide a stacking or fabric design that spans multiple chassis.
- The customer has only one aggregation or core switch (avoid a single upstream switch design because it represents a single point of failure).

This topology is shown in Figure 3. Equivalent network designs can be deployed with a Flex System chassis.

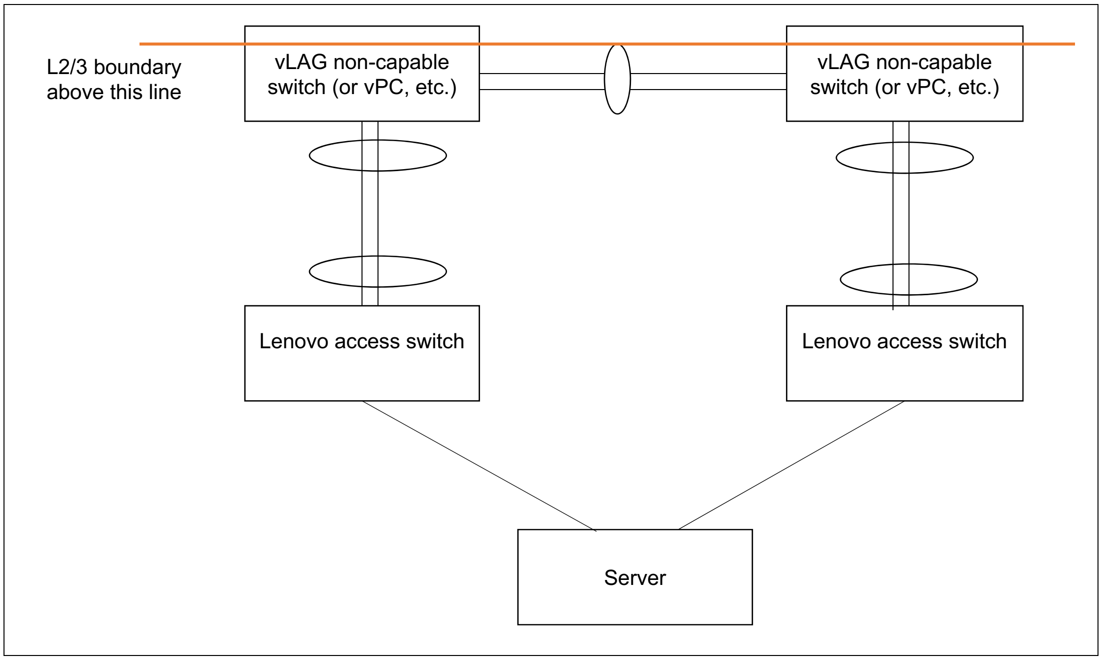
L2/3 boundary above this line
vLAG non-capable switch (or vPC, etc.)
vLAG non-capable switch (or vPC, etc.)
Lenovo access switch
Lenovo access switch
Server

Figure 3   Inverted U network design

The failover feature, which is needed when this topology is used, is configured to specify which upstream aggregation switch facing ports are used as uplinks and which server-facing ports are administratively disabled when the uplinks are down. An example configuration of the failover feature is shown in Figure 4.

```
teaming enable
teaming profile 1 mmon monitor interface port-channel 200
teaming profile 1 mmon control interface port-channel 1
teaming profile 1 enable
```

Figure 4   Configuration commands for failover with auto-monitor tracking an LACP uplink key

Traditional STP design with blocking

This topology was commonly used when functions such as those provided by vLAG were not available. It uses a partial mesh between the server access switches and two upstream switches that are cross-connected to each other. The loops, which are built into this design, are blocked by STP, which puts some ports into a blocking status to prevent a broadcast storm. Operationally, this design resembles an inverted-U topology; however, the blocked links can take over if there are failures with the switch or link.

The major drawback of this design is that it does require the use of STP and results in wasted bandwidth owing to blocked links. For more information about STP, including the multiple versions of the STP protocol, see the Lenovo Press paper, *Lenovo Networking Best Practices for CNOS: Layer 2 Design and Configuration*, `https://lenovopress.com/LP1005`.

Because ports that are blocked by STP do not carry production traffic, sufficient bandwidth must be built into the topology to carry the expected loads with these ports idle. This topology uses the available links inefficiently. In some cases, all of the available links can be used by setting STP parameters (link cost and switch priority) in such a way that some links are blocked for about half of the VLANs in use and other links are blocked for the remaining VLANs.

A network that uses the Traditional STP design is shown in Figure 5. Equivalent designs can be deployed with a Flex System chassis.

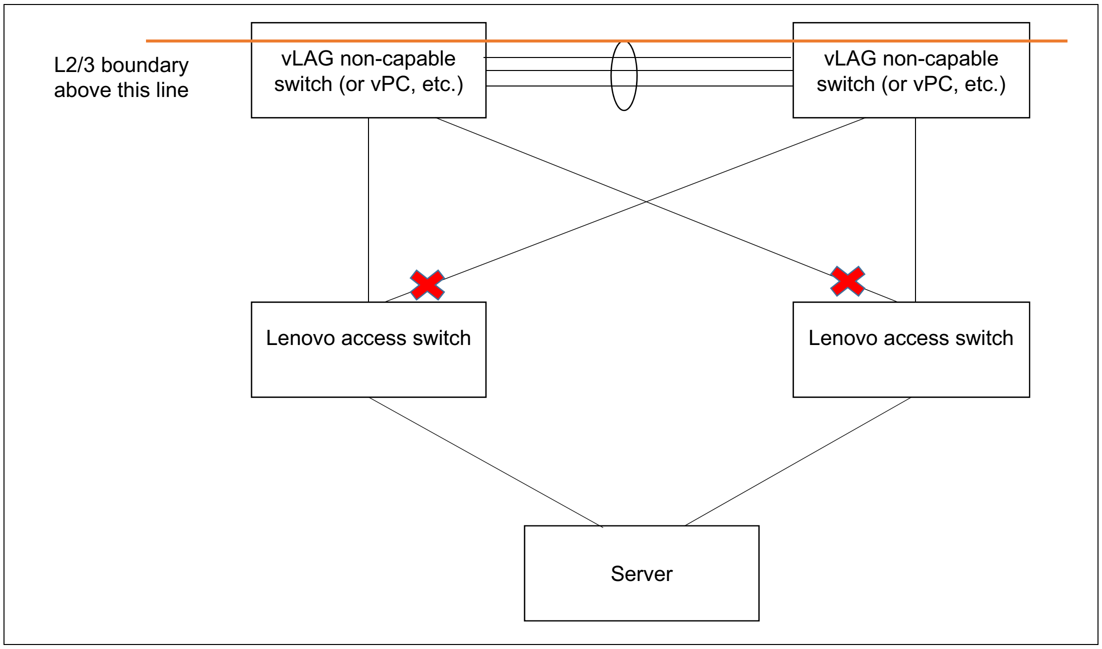

Figure 5 Network topology with STP and blocked uplinks

Figure 6 shows an excerpt of a configuration which implements this design. Note that the default behavior under CNOS (not under ENOS) is to automatically create STP instances when VLANs are created and the configuration would reflect this. The default spanning-tree protocol is compatible with Cisco's PVRST+, with an instance for each VLAN. The MST standard can also be chosen and allows one instance of spanning-tree to support multiple VLANs, as long as their topologies are or are nearly identical.

```
interface eth 1/1, eth 1/10-11
switchport mode trunk
switchport trunk allowed vlan 10,20
[spanning tree stp 10 vlan 10]
[spanning tree stp 20 vlan 20]
```

Figure 6 Spanning tree configuration commands

In this configuration, vLAG can still be used on server-facing ports, and then aggregation-based active/active NIC teaming or bonding on the server can be used. Traditionally, this configuration includes configuring the server to use switch independent mode active/active NIC teaming (for example, Linux mode 5), or active/standby (such as Linux `bonding mode=1`) teaming.

Isolated management network

The use of a separate management network is always a preferred practice for the isolation of data and management environments. An isolated management network can be used in a lights-out environment to provide out-of-band connectivity to locate and troubleshoot issues that might span across the data network and multiple servers. In today's data centers, this network often consists of a dedicated management switch that uses 1 Gb connectivity.

Figure 7 shows Lenovo Switching products that are connecting to a separate 1 Gb Management for out-of-band management. Server management ports (XCC and IMM2), as well as switch management ports, should be connected to this out-of-band network so that they can be reachable in the event of an outage or another issue on the main data network.

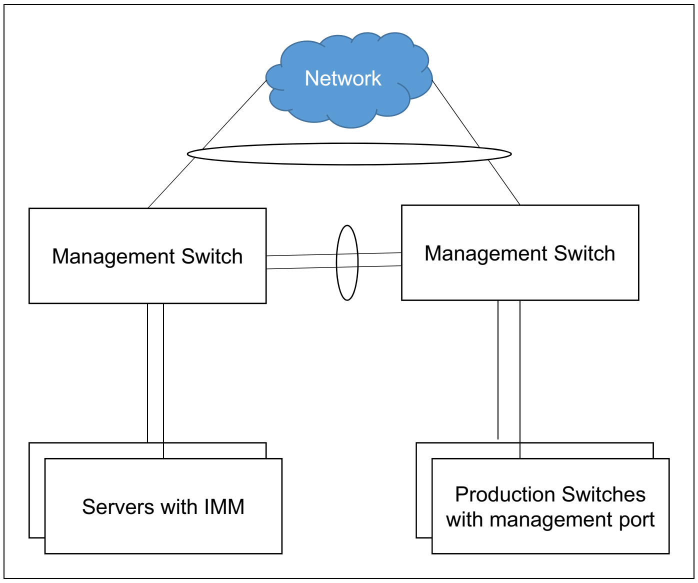

Figure 7 Out of Band 1G Management connectivity

When you use a separate management environment for out-of-band connectivity, it is important to remember that the management network still can consist of high availability features, such as vLAG. Consider the following points about the management network:

- If an HA environment is used in the management network between a pair of Rack Switches, vLAG and Spanning Tree can be enabled.
- Spanning Tree `edge` and `bpdu-guard` should be enabled on ports which connect to all access devices, such as server or storage NICs and management components.

Figure 8 shows the configuration of a G8052 1G management switch.

Figure 8 Example configuration of a G8052 1G management switch

```
hostname G8052-MGMT-Sw1
!
interface port 51,52
   description vLAG-ISL
   switchport mode trunk
   switchport trunk allowed vlan 1021,4090
   switchport trunk native vlan 4090
   spanning-tree guard loop
   lacp key 5152
   lacp mode active
   exit
!
vlan 4090
   name vLAG-ISL
   exit
!
vlan 999
   name Native_VLAN
   exit
!
vlan 1021
   name Management_VLAN
   exit
!
vlan 1022
   name vLAG-HC_VLAN
   exit
!
no spanning-tree stp 6 enable
spanning-tree stp 6 vlan 1022
!
spanning-tree stp 26 vlan 4090
no spanning-tree stp 26 enable
!
interface port 49,50
   description Uplink-To-Network
   switchport mode trunk
   switchport trunk allowed vlan 1021,999
   switchport trunk native vlan 999
   lacp key 4950
   lacp mode active
   exit
!
interface port 1-47
   switchport access vlan 1021
   spanning-tree portfast
   bpdu-guard
   exit
!
interface port 48
   description Health-check-link
   switchport access vlan 1022
   spanning-tree portfast
   exit
!
interface ip 1
   ip address 1.1.1.1 255.255.255.0
   vlan 1022
   enable
   exit
!
vlag ena
vlag hlthchk peer-ip 1.1.1.2
vlag tier-id 100
vlag isl adminkey 5152
vlag adminkey 4950 enable
```

Easy Connect

Easy Connect is a simple configuration mode that is implemented on Lenovo Networking Ethernet switches to enable easy integration of Lenovo switches with Cisco, Juniper, and other vendor data center networks. Easy Connect makes connecting to core networks simple while enabling advanced in-system connectivity at the network edge. It also allows administrators to allocate bandwidth and optimize performance.

Easy Connect is available under CNOS and has long been supported using the older ENOS firmware as well. In both cases, it is implemented using double tagging or q-in-q tagging. However, it is not a full implementation of the use of multiple tags, which some vendors allow to use more than two 802.1q tags.

Easy Connect types and uses cases

With Easy Connect enabled, the switch becomes a simple I/O device that connects servers and storage with the core network. It aggregates compute node ports and behaves similarly to Cisco Fabric Extension (FEX) by appearing as a “dumb” device to the upstream network. With Easy Connect enabled, the upstream network and the attaching hosts are responsible for managing all VLAN assignments and tagging. This loop-free connectivity requires no extra configuration and helps provide economical bandwidth use with prioritized pipes and network virtualization.

Under CNOS, this is configured with the `switchport mode dot1q-tunnel` command on the interfaces where it is desired. Using the dot1q tunnel applies an “outer” 802.1q tag to all frames and processes them through the switching ASIC using the assigned tag value. Ingress frames will have this tag applied whether or not they are already tagged; this tag is typically removed on egress.

Figure 9 on page 11 shows how to successfully deploy `dot1q-tunnel` with vLAG. To successfully deploy this configuration, the upstream network components should also be running some type of Virtualization, such as Stacking, vLAG, vPC, or MC-LAG, depending on the platform of choice.

The following preferred practices are guidelines for implementing easy-connect with or without vLAG:

- All of the normal rules for the VLAG ISL apply.
- A non-production VLAN should be used as the tunnel ports’ Native VLAN or PVID. This VLAN should have Spanning Tree enabled to use some of the Spanning Tree protection, such as loop guard across the vLAG ISL and BPDU Guard on downlink ports.
- As shown in Figure 9 on page 11, it can be seen that regardless of whether you are running dot1q-tunnel or Layer 2 standard switching, both options support similar topologies.

```
hostname Sw1
spanning-tree mode disable
interface eth 1/9-10 *** ISL ports ***
   switchport mode trunk
   switchport trunk allowed vlan 3090,3091
   switchport trunk native vlan 3090

   channel-group 52 mode active
vlan 3090
   name ISL-Peer-Link
vlan 3091
   name EasyConnect
interface eth 1/11-12, eth 1/20-33 ** server facing ports and uplinks **
   switchport access vlan 4091
   switchport mode dot1q-tunnel
interface eth 1/1-2 ** uplinks **

   channel-group 43 mode active
interface mgmt0
   ip address 1.1.1.1/24
   enable
vlag ena
vlag tier-id 10
vlag hlthchk peer-ip 1.1.1.2 vrf management
vlag isl port-channel 52
vlag instance 1 port-channel 43 enable
```

Figure 9 Example configuration with tagpvid-ingress and vLAG

Easy Connect versus full L2 switching

Although Easy Connect is a simple and recommended approach to keeping simplicity and ease-of-use, it might not always meet all requirements. Full Layer 2 (L2) Switching that includes managing of VLANs, Spanning Tree, and possibly Layer 3 has more capability and allows for easier troubleshooting over that of Q-n-Q.

Consider the following advantages of each option:

- Layer 2 Switch Mode:
  - Allows for ongoing VLAN management and manipulation of Native VLAN/PVID ID on a per-port basis (inner VLAN aware).
  - Allows for Spanning Tree Management (option to disable)
- Easy Connect Mode
  - One time configure or Plug and Play options.
  - Allows for VLAN Independent for simple management (is not aware of inner tag).
  - Allows for Spanning Tree to be disabled.
  - Allows for vLAG.
- Combination of L2 and Easy Connect Mode:
  - Allows for L2 and Easy Connect to be configured on the same switch. Easy Connect and standard L2 networking require independent physical ports because Easy Connect converts a physical port into a Q-n-Q port. After the command `switchport mode dot1q-tunnel` was run on that interface, it removes the ability to allow older VLANs to be used on the same port.

Author

Scott Lorditch is a Consulting System Engineer for Lenovo. He performs network architecture assessments and develops designs and proposals for solutions that involve Lenovo Networking products. He also developed several training and lab sessions for technical and sales personnel. Scott joined IBM as part of the acquisition of Blade Network Technologies® and joined Lenovo as part of the System x® acquisition from IBM. Scott spent almost 20 years working on networking in various industries, as a senior network architect, a product manager for managed hosting services, and manager of electronic securities transfer projects. Scott holds a BS degree in Operations Research with a specialization in computer science from Cornell University.

Thanks to the following people for their contributions to this project:

- Philippe Cauche
- David Watts

This paper is based on a chapter in the Lenovo Press book, *Lenovo Networking Best Practices for Configuration and Installation*. Thanks to the authors:

- Scott Irwin
- Scott Lorditch
- Ted McDaniel
- William Nelson
- Matt Slavin
- Megan Gilge

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

© Copyright Lenovo 2019. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by Global Services Administration (GSA) ADP Schedule Contract

This document was created or updated on January 31, 2019.

Send us your comments via the **Rate & Provide Feedback** form found at http://lenovopress.com/lp1068

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at http://www.lenovo.com/legal/copytrade.html.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

- Blade Network Technologies®
- Flex System™
- Lenovo®
- Lenovo(logo)®
- System x®

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.