Building Networks with the Lenovo/Juniper Virtual Data Center Pod Architecture

Introduces expandable topologies for a Virtual Data Center using Lenovo and Juniper networking

Describes network designs using VSAN and iSCSI storage

Provides commands used to implement key Lenovo and Juniper switch features

Includes a bill of materials showing the componets we used in our lab tests

Scott Lorditch

Rick Bartosik

LENOVO PRESS

Abstract

This paper presents two topologies for a cluster or pod of servers and storage which can be deployed in an existing or new data center. The topologies presented are VSAN and iSCSI and can be scaled to considerable size if desired. Juniper and Lenovo switching hardware is used in the designs described; the Juniper switches are deployed as core switches and the Lenovo switches as edge and aggregation switches.

This paper is for networking engineers, designers, and architects -- those working for Lenvo, Lenovo business partners and Lenovo customers.

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

Overview

The traditional model of acquiring data center hardware has been one-at-a-time or a-few-at-a-time dedicated server purchases where the capital expense was often borne by the department whose applications would run on it. In such models, networking gear and other infrastructure in the data center (including power, cooling) were typically bought separately from the servers. This acquisition model is changing, in part, because of the increased use of virtualized servers with hypervisors, and will likely continue to change with the advent of *hyperconverged* systems where storage is also virtualized.

A newer model of data center hardware acquisition is for the data center(s) or portions of them to be viewed as similar to a private cloud – where the use of resources is charged for as an expense but the capital costs are borne by the business as a whole, likely under a capital budget for the IT department. This model would call for hardware to be bought or leased as needed, to add capacity or to upgrade or modernize existing hardware.

In this paper, we present a model for a Virtual Data Center – a set of “pods” which can provide IT services and which can be expanded as needed. This Virtual Data Center can reside in a portion of an existing facility, a new facility or a remote, auxiliary facility. While it might be initially deployed for a specific set of applications, there is nothing in the architecture which requires this.

The key attributes of the Virtual Data Center are:

- It is scalable and expandable, as discussed above, up to 1536 total nodes with 10Gb connectivity.
- Storage is Ethernet-based and can be hyperconverged – residing on servers rather than on dedicated, specialized storage devices. Various storage protocols such as NAS as NFS as well as iSCSI and VSAN which are discussed below can be supported by these topologies.
- The architecture uses standards-based components, including Lenovo and Juniper networking hardware and Lenovo servers. These components have been tested together and have been shown to interoperate; the results of this testing are included in this document.
- The architecture is robust: by design, it will survive single failures in a network connection, or a network device – all without significant service interruption. Networking failures during our testing were recovered from in less than 1 second.
- The environment can be managed by deploying Lenovo XClarity™ Administrator.

We are offering a viable option for acquisition and deployment of a pod or cluster which can grow and whose architecture can scale to considerable size. The design provides for considerable redundancy and flexibility. This architecture, and/or minor variations of it, can grow to encompass an entire data center.

VSAN architecture

This design is for a Virtual Data Center which uses VMware’s VSAN hyperconverged storage capabilities. It thus does not have any devices specifically dedicated to storage; instead, each server’s on-board storage is shared across the cluster of servers. The servers each have considerable on-board storage capacity.

The VSAN design is shown in Figure 1 on page 4.

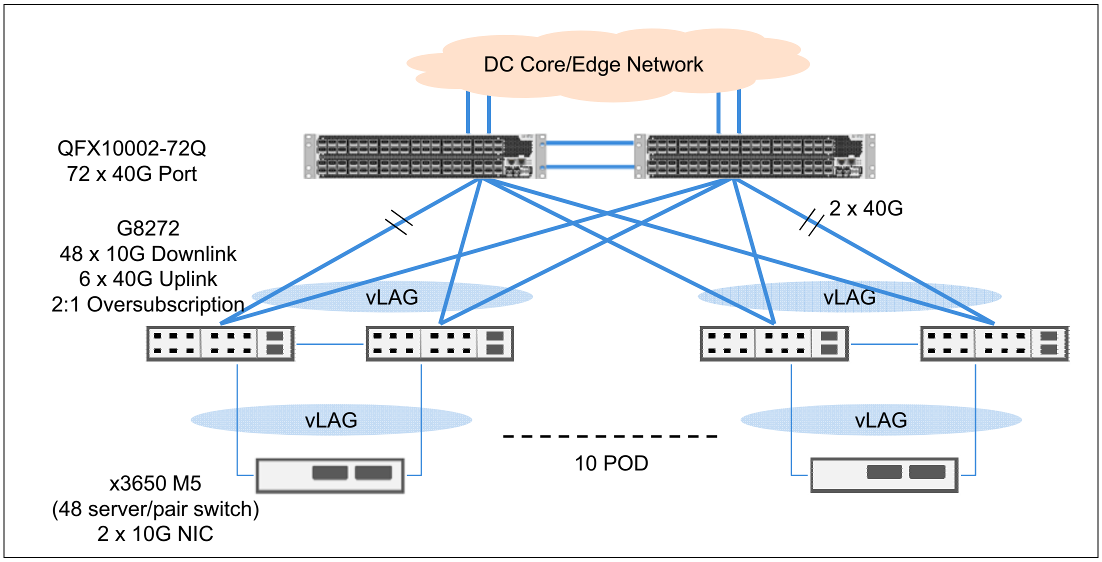

Figure 1 VSAN Architecture as tested

Design

The network design for this architecture provides sufficient bandwidth and redundancy to accommodate production traffic, including traffic from clients such as web browsers, as well as traffic used by the VSAN facility to ensure that stored data is replicated. The VSAN facility will also ensure that data is local to where it is being used, as much as possible.

Redundancy is provided in this design for both storage and networking. VMware HA (high availability) capabilities can also be deployed to protect against server failures. Network redundancy extends down to the servers and can be configured using active/active or active/standby teaming modes. Our testing was done with all links active.

The network in this design includes Lenovo top-of-rack switches, connected to a pair of core Juniper switches. The VDC cluster can be expanded by provisioning additional servers and additional top-of-rack switches. The core Juniper switches can connect to an existing data center network or can serve existing servers in the same data center real estate but which are not part of the VDC cluster.

In the test environment, the Lenovo switches were loaded with the new CNOS firmware, version 10.0. There are significant differences in the functionality and CLI which are introduced with this new release, but none of the functional differences are significant for this topology.

Technical details

The topology used in the VSAN architecture provides the ability to have all links active even though there will be loops in the topology. At the same time, the design provides rapid recovery from failures in switches, ports, or links. Our testing – performed with an IXIA traffic generator and running links at full line rate – showed sub-second recovery from these failures.

The topology relies on two proprietary features, one from Juniper (MC-LAG) and an equivalent from Lenovo (vLAG) – which allow full and partial mesh topologies with link aggregation. Both of these features allow a link aggregation group to connect to two paired switches at the other end while appearing to be a typical standards-compliant LAG to the downstream device. By using these features together on a pair of core Juniper switches and multiple pairs of Lenovo switches, the network meets the objectives mentioned above.

> **Cross connections**: Both vLAG and MC-LAG require a cross-connection between the paired switches that share the aggregated links. In our testing, we used 40Gb links for these connections.

The servers in this design can also use NIC teaming; each operating system and hypervisor in common use has support for active/active and active/standby teaming modes. A detailed discussion of NIC teaming is the subject of a Lenovo Press paper, *An Introduction to NIC Teaming with Lenovo Networking Switches*, available from:

https://lenovopress.com/redp5245

VMware's default teaming modes, on both their standard switch and Virtual Distributed Switch, do not require any special configuration on the adjacent switches (such as vLAG). An available teaming mode which does not assign a guest VM to a specific physical NIC but rather hashes packets based on fields in their headers does require vLAG, and can provide better sharing of available bandwidth especially when there are some guest VMs which use significantly more bandwidth than others.

iSCSI architecture

This VDC cluster design uses the standard iSCSI protocol to connect servers to storage. It is therefore not hyperconverged, but it can nonetheless be deployed using specialized servers to hold storage or using dedicated storage devices available from several vendors. Storage connectivity is provided by 10Gb or 40Gb Ethernet rather than a dedicated Fiber Channel storage network (SAN).

The design is shown in Figure 2 on page 5.

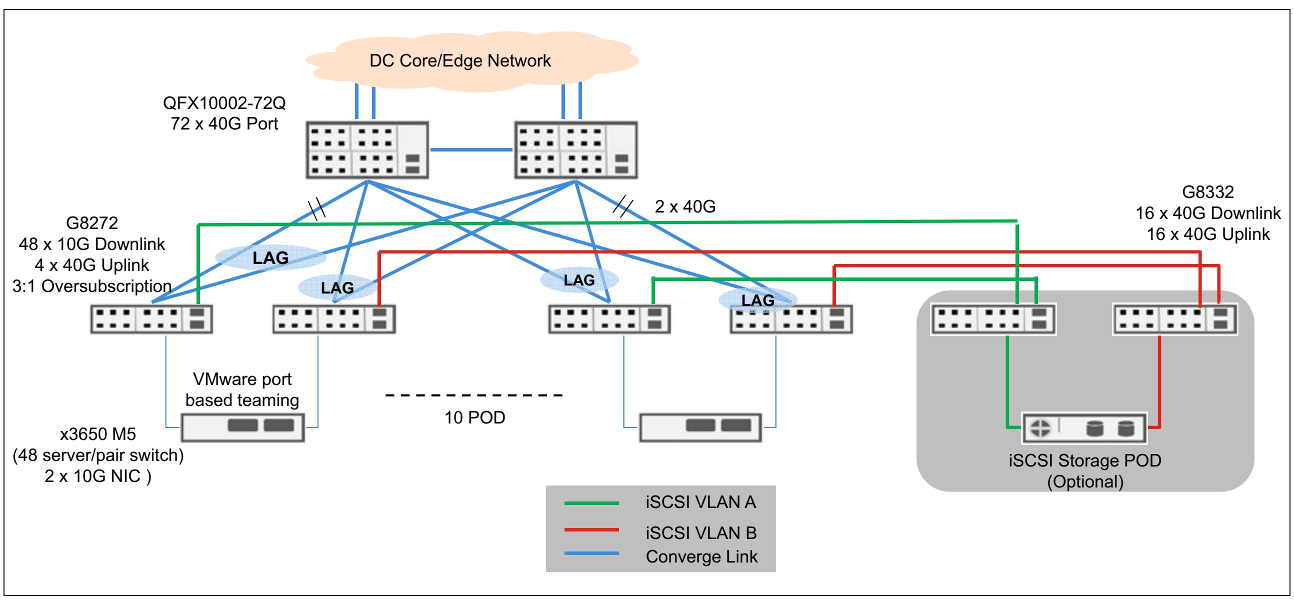

Figure 2 iSCSI architecture as tested

Design

Storage and production traffic run over Ethernet, which means that the same Lenovo and Juniper networking products can be used in this architecture as are used in the VSAN architecture described in "VSAN architecture" on page 3. The Virtual Data Center cluster can be expanded similarly to the VSAN architecture, except that both compute servers and storage servers would be part of an expansion.

The iSCSI architecture relies on redundancy techniques which are defined in the iSCSI standard (Multipath) as well as network redundancy. The network architecture is therefore slightly different from that in the VSAN design shown above. One key difference in the network design is that the Juniper switches are still configured for MC-LAG (shared aggregation) and the Lenovo switches are not since iSCSI is not using NIC teaming. (Note that regular production data can use teaming if a CNA is used.)

The Lenovo switches in this topology run the current 8.x release of firmware. There are no new functions in the 10.x release which would be significant for the testing that was performed.

Technical details

The iSCSI pod design is similar to the VSAN design except that the techniques used to enable failover at the server NIC level are different and partly as a result there are two distinct VLANs connecting the compute servers to the storage.

NIC teaming is not used in this design; iSCSI multipathing is used instead, where available. The iSCSI standard allows for redundancy by opening multiple concurrent TCP sessions on available ports on a server ("initiator") and connecting to any available IP address on the storage array ("target portal"). In our test bed and frequently in real implementations, two (or more) VLANs are used to carry iSCSI traffic between servers and storage; this provides a natural demarcation between the multiple paths between them, providing capacity and redundancy.

NIC teaming configurations would use a single IP address for multiple physical NIC ports and so would limit the diversity of the TCP connections used.

Some servers may be equipped with converged network adapters (CNAs) which provide an iSCSI hardware initiator. NIC teaming is typically not supported for iSCSI traffic on these adapters.

Juniper MC-LAG is used to allow the pair of Juniper core switches to connect to the Lenovo edge switches in such a way that the edge switches see a single link aggregation group. Because NIC teaming is not being used for the iSCSI traffic, vLAG is not used to make the pair of Lenovo switches appear as a single switch. Instead, the Lenovo L2 failover feature was used. This feature can be configured to deactivate server-facing ports if a switch has no available uplink path, so that all the traffic will use the surviving path if it is available. In testing this feature, recovery took less than one second.

If a customer chooses to implement a configuration like this and deploys iSCSI hardware initiators using a CNA, it is possible to modify this topology so that NIC teaming can be used for data traffic (which would flow through the Ethernet NIC function of the CNA) while using iSCSI multipathing to provide redundancy for the iSCSI traffic (which would flow through the iSCSI HBA function of the CNA). This variation of the architecture was not part of our testing.

Juniper feature configuration

In this section, we list extracts from the Juniper configuration used in our testing which configures the Multichannel Aggregated Ethernet (MC-AE) feature (equivalent to the RackSwitch vLAG feature).

The extracts shown are from the VSAN topology; there are small differences when the iSCSI topology was used which primarily resulted from the use of two distinct iSCSI VLANS in that design.

The following are key points regarding the configuration extract shown in Figure 3:

- There are also minor variations between the configurations of the two Juniper switches (QFX-1 and QFX-2).
- QFX-1 and QFX-2 are configured with the same LACP system-id and admin-key. This allows QFX-1 and QFX-2 to appear as a single switch to the edge switch when negotiating LACP. The Lenovo platform views the MC-AE interface as a standard LACP connection.
- QFX-1 and QFX-2 are configured with the same mc-ae-id. This represents the unique MC-AE interface connecting to the edge switch.
- QFX-1 and QFX-2 are configured with the same redundancy-group. This represents a group of one or more MC-AE interfaces that share common VLANs.
- The chassis identifier for QFX-1 and QFX-2 must be unique; 0 and 1.
- The MC-LAG mode on QFX-1 and QFX-2 is set to active-active, meaning that all links to the edge switch are active (from an LACP perspective) and can forward traffic.
- The status-control parameter is set such that QFX-1 is the “Active” or “Master” peer and QFX-2 is the “Standby” peer. This parameter determines the state of QFX-1 and QFX-2 in the event of and ICCP failure
- QFX-1 and QFX-2 are configured with a single access VLAN: 100. In the case of VSAN, the TCP based traffic move East to West and is not influenced by the VLAN configuration on the QFX. VLAN association is required for any North-South traffic, inbound/outbound from the Lenovo platform through the QFX.

Figure 3 shows the Juniper configuration for Multichannel Aggregated Ethernet (MC-AE) feature.

```
set interfaces ae1 aggregated-ether-options lacp active
set interfaces ae1 aggregated-ether-options lacp system-id 00:00:00:00:00:a1
set interfaces ae1 aggregated-ether-options lacp admin-key 3
set interfaces ae1 aggregated-ether-options mc-ae mc-ae-id 1
set interfaces ae1 aggregated-ether-options mc-ae redundancy-group 1
set interfaces ae1 aggregated-ether-options mc-ae chassis-id 0 (partner is 1)
set interfaces ae1 aggregated-ether-options mc-ae mode active-active
set interfaces ae1 aggregated-ether-options mc-ae status-control active (partner is “standby”)
set interfaces ae1 aggregated-ether-options mc-ae init-delay-time 360
set interfaces ae1 unit 0 family ethernet-switching interface-mode access
set interfaces ae1 unit 0 family ethernet-switching vlan members 100
set interfaces et-0/0/1 ether-options 802.3ad ae1
set interfaces et-0/0/2 ether-options 802.3ad ae1
set vlans v100 vlan-id 100
set vlans v100 l3-interface irb.100
```

Figure 3 Portion of the Juniper configuration for Multichassis-Aggregated-Ethernet (MC-AE)

The configuration extract in Figure 4 sets up the cross-connection between the pair of QFX switches and the protocols which run across it. This is used to keep the two switches in synch and to enable failover to happen when needed.

The following are key points regarding the configuration extract shown in Figure 4:

- QFX-1 and QFX-2 are configured with multi-chassis-protection which specifies the peer's ICCP IP address and the ICL link (ae0) used for protection in the event the MC-AE interface goes down.
- QFX-1 and QFX-2 are configured with the same service-id number. The switch service ID is used to synchronize applications, IGMP, ARP, and MAC learning across MC-LAG members.
- Define the local and peer IP loopback addresses used to form the TCP-based connection. In this case, a static route was used for reachability.
- The ICCP peer session-establishment-hold-time is modified from the default of 300 seconds to 50 seconds. During this time interval, the ICCP TCP client attempts to connect to the peer TCP server, and the timer is reset once the ICCP link is up. Configuring session establishment hold time helps in faster ICCP connection establishment. The recommended value is 50 sec.
- Specify one or more Redundancy Group IDs that contain MC-AE interfaces managed by this session. In this configuration there is only one group.
- Configure BFD liveness-detection to provide MC-LAG connectivity in the event of an ICCP failure. The recommended configuration is to leverage the management port of each QFX switch as the liveness-detection-backup-peer.
- This connection is represented by VLAN=500 and IP Subnet of 10.2.1.0/24. To ensure failover in the case of an LACP or interface failure, this connection must also support the Lenovo VLAN=100 requiring the interface to support VLAN Trunking.
- A single port or Ethernet bundle is supported as the ICCP/ICL. Juniper recommends the use of an Ethernet bundle, per the following use of 2x40G links.

Figure 4 shows the remainder of the MC-AE configuration - including the inter-chassis link and VLAN configuration.

```
set multi-chassis multi-chassis-protection 10.2.1.200 interface ae0 (partner is .100)
set protocols iccp local-ip-addr 10.2.1.100 (partner is .200)
set protocols iccp peer 10.2.1.200 redundancy-group-id-list 1
set protocols iccp peer 10.2.1.200 backup-liveness-detection backup-peer-ip 10.1.1.200
set protocols iccp peer 10.2.1.200 liveness-detection minimum-receive-interval 50
set protocols iccp peer 10.2.1.200 liveness-detection transmit-interval minimum-interval 1000
set iccp peer 3.3.3.1 session-establishment-hold-time 50
set switch-options service-id 10
set interfaces irb unit 500 family inet address 10.2.1.100/24
set interfaces ae0 unit 0 family ethernet-switching interface-mode trunk
set interfaces ae0 unit 0 family ethernet-switching vlan members 100
set interfaces ae0 unit 0 family ethernet-switching vlan members 500
set interfaces et-0/0/23 ether-options 802.3ad ae0
set interfaces et-0/0/24 ether-options 802.3ad ae0
set vlans v500 vlan-id 500
set vlans v500 l3-interface irb.500
```

Figure 4 Portion of the Juniper configuration for Multichassis-Aggregated-Ethernet (MC-AE)

Lenovo feature configuration

In this section, we list the commands used to configure vLAG.

The new Lenovo switch firmware, CNOS (10.0), uses slightly different command syntax than the previous (8.x) firmware but the underlying concepts for vLAG are the same.

vLAG is used in this configuration in concert with Juniper MC-LAG, which is configured between the pair of QFX10000 switches so that each pair of switches presents a single link aggregation group to the other, using the LACP standard.

Both vLAG and MC-LAG require a link (or aggregation) between the two partner switches so that learned MAC addresses can be shared with the partner and traffic can be forwarded when necessary. Both vLAG and MC-LAG also have an additional link between the switches, called the health check (hlthchk) in vLAG and multi-chassis-protection in MC-LAG. This additional link serves to allow the partner switches to know if the other switch is active in the event the main link between them should fail.

Figure 5 shows the commands to configure vLAG and LACP aggregation on the Lenovo switches running CNOS.

```
/* Configure Link aggregation group */
interface Ethernet1/51
 aggregation-group 100 mode active
 lacp timeout short
exit
interface Ethernet1/52
 aggregation-group 100 mode active
 lacp timeout short
exit
interface Ethernet1/53
 aggregation-group 1000 mode active
exit
interface Ethernet1/54
aggregation-group 1000 mode active
exit
interface port-aggregation1000
 lacp suspend-individual
exit

/* Configure vLAG */
vlag tier-id 10
vlag priority 120 (partner uses 100)
vlag isl port-aggregation 1000
vlag hlthchk peer-ip 1.1.1.2 vrf management (partner switch uses 1.1.1.1)
vlag startup-delay 0
vlag enable
vlag instance 1 port-aggregation 100
vlag instance 1 enable
```

Figure 5 Configuring vLAG and LACP aggregation on the Lenovo switches running CNOS

iSCSI configurations

Failover, also called Layer-2 Failover, is a feature of Lenovo switches which disables downlink ports, typically server-facing (edge) ports, when the uplink connections from the switch have failed. This serves to trigger NIC failover on the servers, which is done differently depending on the OS installed on the server.

In the tested iSCSI topology, the servers are connected to two G8272 switches, which are each in turn connected to the two Juniper QFX10002 core switches. If the uplinks from a G8272 fail, the failover feature will disable the server-facing ports, causing the server OS to use the other NIC (assuming it still has a functioning path) to send both data and iSCSI traffic.

The switch configuration for this feature is shown in Figure 6. The configuration shown will disable port 1 if both ports 51 and 52 fail, and will reactivate port 1 if either of 51 or 52 is restored.

```
/* Configure VLAN port mode and membership for interface */
interface port 1
        switchport mode trunk
        switchport trunk allowed vlan 1,200
        exit

interface port 49
        switchport access vlan 200
        exit

/* Configure LACP for  interface */
interface port 51
        lacp mode active
        lacp key 51

interface port 52
        lacp mode active
        lacp key 51

/* Configure L2 Failover triggers */
failover enable
failover trigger 1 mmon monitor admin-key 51
failover trigger 1 mmon control member 1
failover trigger 1 enable
```

Figure 6 Configuring Layer-2 Failover

Test results

Test cases were developed to replicate common real-world failures. For example:

- Failures in a Lenovo or Juniper switch due to a switch reboot or a switch power failure
- Failures in a link between switches (or the ports or transceivers on that link)

Testing was not done where two switches failed simultaneously; there are several obvious cases where this would cause an outage which would not be recoverable short of restoring at least one of the failed switches.

The result of the testing was recovery from failures was sub-second. There were no significant differences in failover behavior which arose from using Lenovo and Juniper products from what we've experienced in configurations with only one vendor's products used.

The topologies shown also did not experience interoperability issues resulting from the use of Lenovo and Juniper gear. VLANs, link aggregation, and vLAG/MC-LAG (also called MC-AE by Juniper) all worked together without issues.

A bill-of-materials showing the products and software releases used is in the Appendix.

Appendix

Table 1 lists the components we used in the testing of the two designs presented in this paper.

Table 1 Bill of Materials for the test environment

Component	iSCSI Topology	VSAN Topology
Server Type/Model	Lenovo System x3550 M5	Lenovo System x3650 M5
NIC Type	Dual-port 10Gb	Dual-port 10Gb
Juniper Switch Model	QFX10002-72Q	QFX10002-72Q
Lenovo Switch Model	G8272 – firmware 8.2.5.0 G8332 – firmware 8.3.4.0	G8272 – firmware 10.1.1.0
Storage Device	V7000	Server internal HDD/SSD
ESXi	ESXi 6.0, vCenter	ESXi 6.0, vCenter
Guest VM OS	Windows	Windows
Test tools	IXIA, iPerf, IOMeter, dt	IXIA, iPerf, IOMeter, dt
Transceivers and DAC cables used	Juniper: ► Fiber: 740-032986 QE067148 QSFP+-40G-SR4 ► DAC: 740-038624 12286660 QSFP+-40G-CU3M Lenovo: ► Fiber: BN-CKM-QP-SR4 ► DAC : 00D5803-N13692A	Juniper: ► Fiber: 740-032986 QE067148 QSFP+-40G-SR4 ► DAC: 740-038624 12286660 QSFP+-40G-CU3M Lenovo: ► Fiber: BN-CKM-QP-SR4 ► DAC : 00D5803-N13692A

Authors

This paper was produced by the following team of specialists:

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.

Rick Bartosik is a Consulting Engineer with Juniper Networks. He has over 20 years of systems engineering experience in various networking disciplines such as LAN/WAN design and optimization, VoIP/UC engineering, video conferencing, IPTV and protocol optimization. His experience includes working for a technical startup company as well as one of the largest manufacturers of Networking and UC applications in the world. Rick’s current responsibilities include focusing on next generation data center, automation and SDN technologies. He is a certified Juniper Networks Certified Enterprise Routing and Switching Professional (JNCIP-ENT) and a certified Juniper Networks Certified Design Specialist, Data Center (JNCDS-DC).

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

- Muhammad Ateeq Khan, Lenovo Networking Development
- David Watts, Lenovo Press

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.

This document was created or updated on October 7, 2016.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp0539

Trademarks