Lenovo

Lenovo Networking Best Practices for CNOS: Layer 3 Technology

Describes the configurations of topologies with Layer 3 routing

Describes the use of routing protocols such as BGP and OSPF

Include designs for fault tolerance with Layer 3

Explains functions that support multi-tenancy with Layer 3 topologies

Scott Lorditch

LENOVO PRESS

Abstract

This paper includes preferred practices for implementing common Layer 3 technologies with some of the options in the Lenovo Networking products. Its intended audience includes network architects and designers as well as technical sales professionals whose focus includes networking. The paper includes examples where multiple functions and protocols that enable Layer 3 designs are used together.

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

It is very common for network designs and topologies to include the use of Layer 3 forwarding, most commonly known as routing. Any network which is larger than the number of devices supported by a single IP subnet will use routing.

Routes can be configured manually (referred to as static routes), or Layer 3-capable switches can exchange routing information with each other (referred to as dynamic routes). Recommended topologies and designs which include the use of these protocols are described in this paper.

A historical aside: Years ago, there was an axiom in network design that stated “switch where you can, and route where you must.” This was because at the time, Layer 3 (router) ports were relatively scarce, relatively slow, and relatively quite expensive when compared to switch ports. With current technologies, this axiom no longer applies. The switching ASICs used in today’s Lenovo switches can perform both Layer 2 switching and Layer 3 routing with equal speed, and all ports on current switches can be used for routing.

OSPF with VRRP and vLAG

Open Shortest Path First (OSPF) is one of the most common enterprise Layer 3 technologies that are used within a data center for dynamic distribution of Layer 3 routes in a medium to large environment. OSPF has many advantages that use several operational characteristics to makes it efficient.

For more information about OSPF and its capabilities within the Lenovo® Networking Products, see the Application and Command Reference guides that are available in the Lenovo Information Center:

https://systemx.lenovofiles.com/help/index.jsp?topic=%2Fcom.lenovo.systemx.common.nav.doc%2Foverview_rack_switches.html&cp=0_5

Note: It will be necessary to select one of the networking products to navigate to the documents.

When OSPF and Virtual Router Redundancy Protocol (VRRP) are introduced on a pair of Lenovo Layer 3-capable switches, a highly redundant environment is created that allows for a floating Layer 3 IP Gateway Address. VRRP allows for various options that include the ability to define how and when a Layer 3 IP virtual address changes ownership between the pair of redundant switches. VRRP was originally developed to act as a Master/Slave relation where the Master switch is the only switch that responds to client ARP requests.

When vLAG is introduced with VRRP, this configuration enables active/active teaming of a third device and VRRP to be active/active.

Figure 1 shows an environment with all three options that are configured in which both switches allow for OSPF and VRRP and can act as Layer 3 active devices with the enabled vLAG ports. By allowing for both switches to act as a Layer 3 device, traffic can be split across both active pairs to reduce the total amount of bandwidth on any one Layer 3 switch.

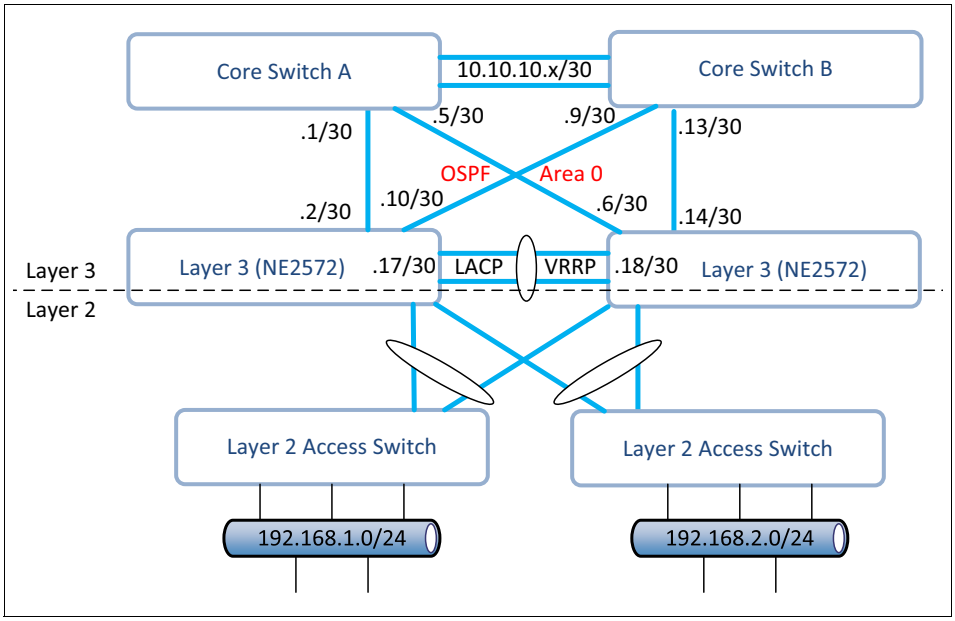

Figure 1 OSPF with VRRP and vLAG

Notes on using VRRP with vLAG:

- Servers attached to the L2 access switches at the bottom of Figure 1 should have their default gateways configured to use the VRRP virtual addresses which are shared between the two L3 switches.
- The L3 switches will not advertise the virtual addresses using OSPF to the core switches at the top of the diagram. Instead, there will be two equal cost routes, each of which will have one of the L3 switches as the next hop, for each subnet that is configured with VRRP in this environment. This is shown in Example 1.

Example 1 IP route table display limited to routes learned via OSPF

```
G8332-1#sh ip route ospf
IP Route Table for VRF "default"
O       10.4.0.5/32 [110/2] via 10.10.10.2, Ethernet1/5/1, 20:56:00
O       10.4.0.6/32 [110/2] via 10.10.10.6, Ethernet1/5/2, 20:55:54
O       10.10.10.8/30 [110/2] via 10.10.10.2, Ethernet1/5/1, 20:51:25
O       10.10.10.12/30 [110/2] via 10.10.10.6, Ethernet1/5/2, 20:52:11
O       10.10.10.16/30 [110/11] via 10.10.10.6, Ethernet1/5/2, 20:55:54
                       [110/11] via 10.10.10.2, Ethernet1/5/1, 20:55:54
O       192.168.20.0/24 [110/11] via 10.10.10.6, Ethernet1/5/2, 20:55:54
                        [110/11] via 10.10.10.2, Ethernet1/5/1, 20:55:54
O       192.168.30.0/24 [110/11] via 10.10.10.6, Ethernet1/5/2, 20:55:54
                        [110/11] via 10.10.10.2, Ethernet1/5/1, 20:55:54
```

- Addresses beginning 192.168 are those which are configured onto the access switches at the bottom of the diagram in Figure 1.
- The two L3 switches will both accept traffic whose destination is the VRRP shared MAC address when vLAG is active, and will accept traffic for the VRRP shared virtual IP

address, unless vLAG is configured to disable this capability. Normal VRRP behavior is for only the “master” switch to accept traffic for those MAC and IP addresses.

Example 2 below shows the relevant portions of a configuration switch that is shown in Figure 1 on page 4.

Example 2 Partial output of a **`show run`**

```
!
version "10.9.1.0"
!
hostname P4-G8272-1
!
vlag tier-id 101
vlag isl port-channel 5
vlag hlthchk peer-ip 172.16.194.6 vrf management
vlag enable
vlag instance 1 port-channel 12
vlag instance 1 enable
!
vlan 1
!
vlan 10
!
vlan 20
!
vlan 30
!
vlan 40
!
interface Ethernet1/1
 switchport mode trunk
 switchport trunk allowed vlan 1,10,20,30
 channel-group 12 mode active
!
interface Ethernet1/2
 switchport mode trunk
 switchport trunk allowed vlan 1,10,20,30
 channel-group 12 mode active
....
!
interface Ethernet1/47
 no switchport
 ip address 10.10.10.2/30
 ip router ospf 0 area 0.0.0.0
!
interface Ethernet1/48
 no switchport
 ip address 10.10.10.10/30
ip router ospf 0 area 0.0.0.0
!
...
!
interface Ethernet1/53
 switchport mode trunk
 switchport trunk allowed vlan 10,20,30
 channel-group 5 mode active
!
interface Ethernet1/54
 switchport mode trunk
 switchport trunk allowed vlan 10,20,30
```

```
 channel-group 5 mode active
!
interface loopback0
 no switchport
 ip address 10.4.0.5/32
 ip router ospf 0 area 0.0.0.0
!
...
!
interface Vlan10
 no switchport
 ip address 10.10.10.17/30
 ip router ospf 0 area 0.0.0.0
!
interface Vlan20
 no switchport
 ip address 192.168.20.2/24
 no ip redirects
 vrrp 20
  address 192.168.20.1
  priority 101
  no shutdown
  ip router ospf 0 area 0.0.0.0
!
interface Vlan30
 no switchport
 ip address 192.168.30.2/24
 no ip redirects
 vrrp 30
  address 192.168.30.1
  priority 101
  no shutdown
 ip router ospf 0 area 0.0.0.0
!
interface port-channel5
 switchport mode trunk
 switchport trunk allowed vlan 10,20,30
!
interface port-channel12
 switchport mode trunk
 switchport trunk allowed vlan 1,10,20,30
!
router ospf 0
router-id 10.4.0.5
!
```

BGP with VRRP and vLAG

Although Border Gateway Protocol (BGP) is most common in the wide area network, it also has its merits in an enterprise Layer 3 technology that is used within some of the larger data centers that are often used to divide between boundaries.

For more information about BGP and its capabilities within Lenovo Networking products, see the Application and Command Reference guides that are available in the Lenovo Information Center: (note you will need to navigate to a a specific switch model to access the documents):

https://systemx.lenovofiles.com/help/index.jsp?topic=%2Fcom.lenovo.systemx.common.nav.doc%2Foverview_rack_switches.html&cp=0_5

When BGP and VRRP are introduced on a pair of Layer 3-capable switches, a highly redundant environment is created for a floating Layer 3 IP Gateway Address. VRRP enables various options to define how and when a Layer 3 IP elects to change ownership between a pair of redundant switches. VRRP was originally developed to act as a Master/Slave relation where the Master switch is the only device that responds to client ARP requests.

When introducing vLAG with VRRP, this configuration enables active/active teaming of a third device and VRRP to act as an active/active Layer 3 pair of switches.

Figure 2 shows an environment with all three options that are configured in which both switches allow for BGP and VRRP and the switches can act as Layer 3 active devices with the enabled vLAG ports. By allowing for both switches to act as a Layer 3 device, traffic can be split across both active pairs to reduce the total amount of bandwidth on any one Layer 3 switch.

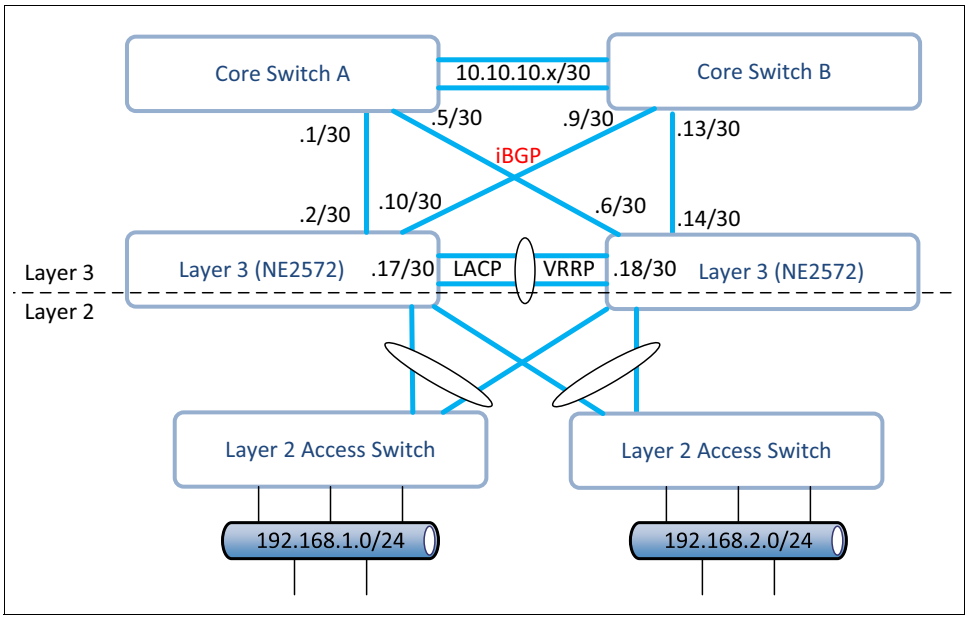

Figure 2 BGP and VRRP with vLAG

The configuration of a BGP implementation is similar to that of an OSPF implementation except for defining an eBGP or iBGP relationship with its peers. Note that iBGP means that the remote AS number is the same as the local one and eBGP means that different AS numbers are used.

Additional options for BGP are numerous and are described in the manuals listed above. One key feature is the ability to redistribute routes that originate elsewhere, including static routes and those from other protocols.

The command syntax for this is:

```
redistribute [direct|ospf|static] [route-map <name>]
```

The optional route-map can be used to filter and only redistribute some of the available routes.

Example 3 on page 8 shows the BGP portion of the configuration only. An alternative version of the configuration using dynamic BGP peers and an address range is also shown.

Also note that the **maximum-paths** configuration command is required to have both routes be active (ECMP). The **maximum-paths** option is also available if EBGP is used, where the switches would have different autonomous system (AS) numbers.

Note that the **address family ipv4 unicast** option is necessary for BGP to function, but if it is omitted, there will be no error diagnostic generated.

The rest of the configuration is described in “OSPF with VRRP and vLAG” on page 3.

Route maps and their configuration are discussed in “Route maps” on page 10.

Example 3 Output of the iBGP portion of a show run

```
router bgp 61001
 address-family ipv4 unicast
  maximum-paths ibgp 4
  redistribute direct
 neighbor 10.4.0.2 remote-as 61001
  update-source loopback0
  address-family ipv4 unicast
 neighbor 10.4.0.5 remote-as 61001
  update-source loopback0
  address-family ipv4 unicast
 neighbor 10.4.0.6 remote-as 61001
  update-source loopback0
  address-family ipv4 unicast

Using dynamic peers and address range
router bgp 61001
 address-family ipv4 unicast
  maximum-paths ibgp 4
  redistribute direct
 neighbor 10.4.0.0/24 remote-as 61001
  update-source loopback0
  address-family ipv4 unicast
```

The routing tables that are generated from the BGP configuration shown in Example 3 are shown in Example 4. Addresses are as follows:

- The 10.4.0.x addresses are used as the router-id addresses for BGP and are the loopback addresses for the respective switches.
- The 10.10.10.x addresses are those on the links which connect the switches.
- Addresses beginning 192.168 are those which are configured onto the access switches at the bottom of the diagram in Figure 2 on page 7.

Example 4 BGP route table entries

```
G8332-2#sh ip route bgp

IP Route Table for VRF "default"
B       10.10.10.0/30 [200/0] via 10.4.0.5 (recursive via 10.10.10.10 ), 00:00:58
B       10.10.10.4/30 [200/0] via 10.4.0.6 (recursive via 10.10.10.14 ), 00:00:58
B       10.10.10.16/30 [200/0] via 10.4.0.5 (recursive via 10.10.10.10 ), 20:56:04
                       [200/0] via 10.4.0.6 (recursive via 10.10.10.14 ), 20:56:04
B       192.168.20.0/24 [200/0] via 10.4.0.5 (recursive via 10.10.10.10 ), 20:56:04
                        [200/0] via 10.4.0.6 (recursive via 10.10.10.14 ), 20:56:04
B       192.168.30.0/24 [200/0] via 10.4.0.5 (recursive via 10.10.10.10 ), 20:56:04
                        [200/0] via 10.4.0.6 (recursive via 10.10.10.14 ), 20:56:04
```

ECMP with static and dynamic routes

Equal Cost Multipath (ECMP) provides load sharing across multiple static or dynamic routes to a single destination. Figure 3 shows how dynamic ECMP routes can be used to loadshare traffic.

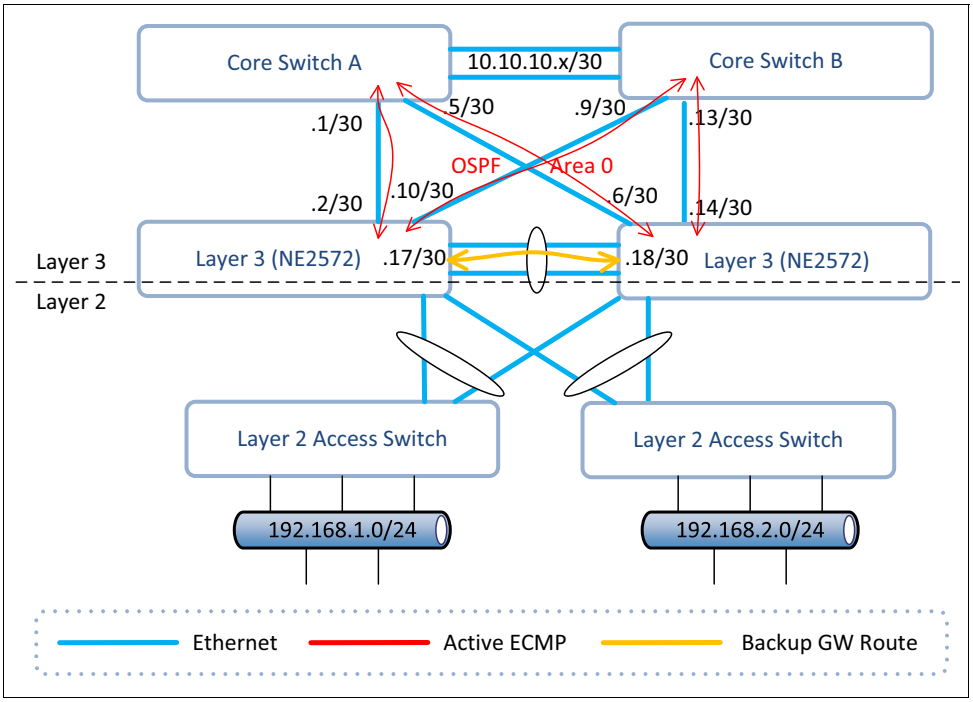

Figure 3 ECMP that uses dynamic routes

Static routes specify to which IP source network, mask, and destination address to send traffic. A second route with the same source network and mask but with a different destination address allows a Layer 3 switch to split the traffic across the multiple routes based on source IP, destination IP, and source and destination Layer 4 port (by default). For more information about static routes, see “Static Routing with vLAG” on page 12.

Example 5 shows two static routes that use the same source network and mask to two separate gateway addresses, which enables load share traffic.

Example 5 ECMP that uses static route configuration (CNOS)

```
!
ip route 10.10.10.0/24 192.168.50.1
ip route 10.10.10.0/24 192.168.60.1
[ip load-sharing <hash options>]
!
```

Dynamic routes can use ECMP by specifying multiple routes to the same physical destination and loadshare based on IP Source and/or, Destination addresses, and source and/or destination Layer 4 port numbers.

Note that OSPF will add multiple equal-cost routes to the active routing table (**`show ip route`**) by default, BGP requires that the **`maximum-paths [ibgp|ebgp] <n>`** command be configured, as shown in Example 6 on page 10.

Example 6 OSPF routes with multiple routes to some destinations

```
sho ip route ospf
IP Route Table for VRF "default"
O       10.4.0.2/32 [100/3] via 10.10.10.6, Ethernet1/5/2, 00:00:52
                    [100/3] via 10.10.10.2, Ethernet1/5/1, 00:00:52
O       10.4.0.5/32 [100/2] via 10.10.10.2, Ethernet1/5/1, 00:00:52
O       10.4.0.6/32 [100/2] via 10.10.10.6, Ethernet1/5/2, 00:00:52
O       10.10.10.8/30 [100/2] via 10.10.10.2, Ethernet1/5/1, 00:00:52
O       10.10.10.12/30 [100/2] via 10.10.10.6, Ethernet1/5/2, 00:00:52
O       10.10.10.16/30 [100/11] via 10.10.10.6, Ethernet1/5/2, 00:00:52
                       [100/11] via 10.10.10.2, Ethernet1/5/1, 00:00:52
O       192.168.20.0/24 [100/11] via 10.10.10.6, Ethernet1/5/2, 00:00:52
                        [100/11] via 10.10.10.2, Ethernet1/5/1, 00:00:52
O       192.168.30.0/24 [100/11] via 10.10.10.6, Ethernet1/5/2, 00:00:52
                        [100/11] via 10.10.10.2, Ethernet1/5/1, 00:00:52
```

ECMP static routes can also be weighted, with weight ranging as high as four. This means that the more heavily weighted routes would be placed in a round-robin rotation as many as four times, thus using them more frequently. This can be used when routes do not hash evenly, or simply when the routes do not have the same available bandwidth.

The syntax for weighed routes is as follows:

ip ecmp weight enable – to enable the use of weighting

ip ecmp weight <ip address> <weight 1-4> or
ip ecmp weight interface <interface> <weight 1-4>

For more information about dynamic routes, see “Dynamic Routing with vLAG” on page 12.

Example 7 shows the configuration of two dynamic routes from BGP that use the same source network and mask to two separate gateway addresses, which enables loadshare traffic.

Example 7 ECMP that uses dynamic route configuration (CNOS)

```
G8332-1#sho ip route bgp
IP Route Table for VRF "default"
B       10.10.10.8/30 [200/0] via 10.4.0.5 (recursive via 10.10.10.2 ), 00:00:45
B       10.10.10.12/30 [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 00:00:45
B       10.10.10.16/30 [200/0] via 10.4.0.5 (recursive via 10.10.10.2 ), 00:00:45
                       [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 00:00:45
B       192.168.20.0/24 [200/0] via 10.4.0.5 (recursive via 10.10.10.2 ), 00:00:45
                        [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 00:00:45
B       192.168.30.0/24 [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 00:00:45
B       192.168.100.0/24 [200/0] via 10.4.0.2 (recursive via 10.10.10.6
                                                      via 10.10.10.2 ), 00:01:19
```

Route maps

Route maps are used to filter routes to neighboring switches. They can be used to filter which routes are advertises to a neighboring switch and/or which received advertisements are added to the local routing table. Similarly, route-maps can be used to limit which routes are redistributed into a routing process for advertisements to neighboring switches. Route maps can be used for OSPF and BGP processes.

Example 8 below shows an iBGP with filtering so that the path to one subnet (192.168.30.2) is not advertised.

The resulting routing table in the neighbor router is shown in Example 9.

Route-maps are configured with IP prefix lists which can be set to `permit` (allowed to advertise as normal to the neighbor(s)) or `deny` (suppress the advertisement). They need to be specified on the configuration of BGP neighbor(s) or redistribution commands.

When configured on a BGP neighbor, route maps can filter to block advertisements which would otherwise be sent (`out`) or cause some received advertisements to be discarded (`in`). Route maps can also be used to limit which routes are to be redistributed from another protocol.

Example 8 Route map to filter advertisement of 192.168.30.x and BGP neighbor configuration

```
route-map thirty permit 10
 match ip address prefix-list block-thirty
ip prefix-list block-thirty seq 10 deny 192.168.30.0/24
ip prefix-list block-thirty seq 20 permit any
!
router bgp 61001
 address-family ipv4 unicast
  redistribute direct
 neighbor 10.4.0.1 remote-as 61001
  update-source loopback0
  address-family ipv4 unicast
    route-map thirty out
```

The resulting routing table shown in Example 9 below does not include the “30” subnet advertisement from the switch at 10.10.10.2 (whose partial configuration is shown in Example 8) but it does include the 192.168.20.0 subnet from both of the neighboring routers 10.10.10.2 and .6 and the “30” subnet from switch at address 10.10.10.6 where the route map is not configured.

Example 9 Routing table showing absence of suppressed route

```
! #sh ip route bgp
IP Route Table for VRF "default"
B       10.10.10.8/30 [200/0] via 10.4.0.5 (recursive via 10.10.10.2 ), 00:00:44
B       10.10.10.12/30 [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 00:00:44
B       10.10.10.16/30 [200/0] via 10.4.0.5 (recursive via 10.10.10.2 ), 21:31:18
                       [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 21:31:18
B       192.168.20.0/24 [200/0] via 10.4.0.5 (recursive via 10.10.10.2 ), 21:31:18
                        [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 21:31:18
B       192.168.30.0/24 [200/0] via 10.4.0.6 (recursive via 10.10.10.6 ), 00:00:44
```

Layer 3 with vLAG and limitations

Static and Dynamic Layer 3 with VRRP and vLAG are all supported on the same pair of switches at the same time. Layer 3 VRRP with vLAG can provide for an active/active environment on the Layer 2 and Layer 3 portions of the Switch pair. Although it is still preferred practice to enable and use Spanning Tree to prevent Layer 2 loops, having vLAG enabled can reduce or eliminate the number of blocked paths, which effectively reduces the requirements for Spanning Tree. However, it is still advisable to have Spanning Tree enabled to help prevent any future potential broadcast storms.

Dynamic Routing with vLAG

Dynamic Routing, such as OSPF and BGP, are supported by vLAG on the same pair of switches. However, vLAG cannot be a member of a port that also is peering with a neighbor. For example, if OSPF (or BGP) is used on a port to form an adjacency with a neighbor, it is not supported to also enable vLAG on that specific port.

Figure 4 shows the use of dynamic routes and vLAG on the same pair of switches. The OSPF point-to-point connections are single port adjacencies while the Layer 2 portion in this example is using PortChannels and vLAG to provide an even distribution to both vLAG enabled Layer 3 NE2572 switches.

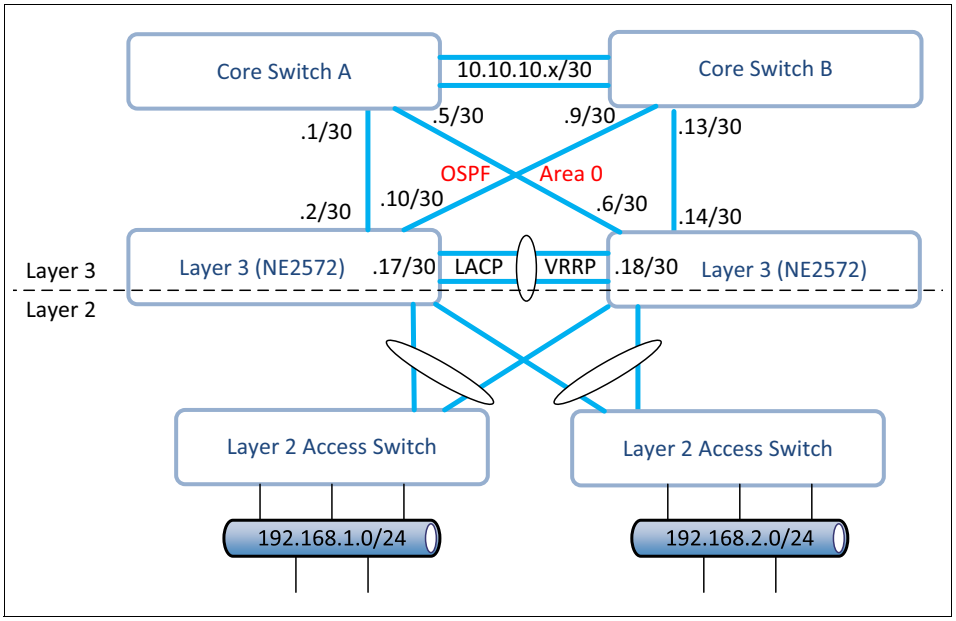

Figure 4 Pair of NE2572 Layer 3 switches with vLAG and Dynamic Layer 3 routing enabled

Static Routing with vLAG

Unlike Dynamic routing, Static routing can support vLAG on the same switch ports. This support means that a pair of vLAG enabled switches with VRRP and default gateways that point upstream towards the core can also be connected to the core via a pair of vLAG ports to provide for Layer 2 and Layer 3 active/active across the same set of ports.

Figure 5 shows the use of static routes and vLAG over the same ports.

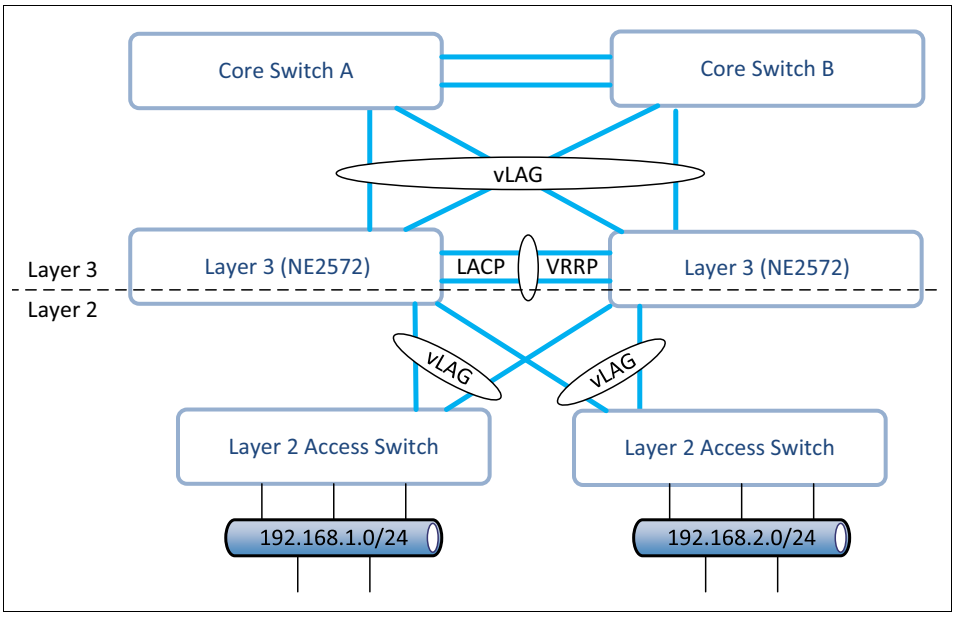

Figure 5 Pair of NE2572 switches with vLAG and Static Layer 3 routing enabled

Spanning Tree with vLAG

Spanning Tree with vLAG is supported on the same pair of switches if it is a form of multi-Spanning Tree. For example, although MSTP and PVRST+ support vLAG, RSTP and vLAG cannot be enabled on the same switch at the same time.

The following parameters are described in the current CNOS Applications Guides for the various models of Lenovo switches:

- When VLAG is configured or changes are made to your VLAG configuration, consider the following VLAG behavior:
  - When a static Mrouter is added on VLAG links, ensure that you also add the Mrouter on the ISL link to avoid VLAG link failure. If the VLAG link fails, traffic cannot be recovered through the ISL.
  - When you enable VLAG on the switch, the ISL shuts down if an MSTP region mismatch is detected with the VLAG peer. In such a scenario, correct the region on the VLAG peer and manually enable the ISL.
  - If you enabled VLAG on the switch and you must change the STP mode, ensure that you first disable VLAG and then change the STP mode.
  - When VLAG is enabled, you might see two root ports on the secondary VLAG switch. One of these ports is the actual root port for the secondary VLAG switch and the other port is a root port that is synced with the primary VLAG switch.
  - The LACP key used must be unique for each VLAG in the entire topology.
  - The STG to VLAN mapping on both VLAG peers must be identical.

- The following parameters must be identically configured on the VLAG ports of both of the VLAG peers:
  - VLANs
  - Native VLAN tagging
  - STP mode
  - BPDU Guard setting
  - STP port setting
  - MAC aging timers
  - Static MAC entries
  - ACL configuration parameters
  - QoS configuration parameter

Policy Based Routing (CNOS 10.9 and above)

Policy Based Routing (PBR) is a feature that increases the flexibility of routing decisions by allowing the consideration of several additional parameters. It is configured through the use of route maps which match on one or more of the available parameters and set the next hop for the packet to a specific address. It is possible to provide a list of next-hop addresses and to verify their availability before using them, but ECMP-style load balancing is not supported.

Available match criteria in the route maps are: source and destination IP, protocol, source and destination TCP/UDP ports, precedence, and DSCP value.

It may be useful to think of this feature as an enhancement over what ACLs can do: instead of just being able to allow or deny a packet, it is possible to choose where the packet is to be forwarded. Note that the selected destination still must be reachable for this to work, and the use of the health-checks is recommended. The feature is useful to provide increased granularity and allow traffic to be forwarded based on criteria – as mentioned above – which could relate to applications or security zones.

The configuration of Policy Based Routing is similar to configuring ACLs, with the following additions:

1. Configure an ACL to select which traffic is to be routed according to the configured policy. Note that “permit” traffic is subject to the configured policy while “denied” traffic is forwarded according to the normal routing tables even if that results in the traffic being discarded.
2. Configure a PBR route-map which consists of “match” clauses referring to the ACL from step 1 and one or more “set” clauses that specify the desired next hop. These can (and are recommended to) include the **`verify availability`** option.
3. Attach the route map to a L3 interface (VLAN interface, or **`no switchport`** routed interface)

An example of this is shown in Example 10.

Example 10 Policy Based Routing configuration

```
! note – the below line is required
feature pbr
!
ip access-list pbr-acl-allow-2-10
10 permit udp any host 192.168.2.10
! match on udp traffic to 2.10
20 deny tcp any host 192.168.2.10
!not tcp traffic
30 permit any any 192.168.2.0 0.0.0.255
```

```
! match on other traffic to subnet
ip access-list pbr-acl-deny-3-20
10 deny any any host 192.168.3.20
! no match on 3.20
20 permit any any any
!
route-map pbr-map permit 10
match ip address pbr-acl-allow-2-10
set ip next hop 10.10.10.2 10.10.10.3
set ip next-hop verify 40.40.40.2 seq 1 interval 10 retry 3
set ip next-hop verify 50.50.50.2 seq 5
exit
!
route-map pbr-map permit 20
match ip address pbr-acl-deny-3-20
set ip next-hop 10.10.10.2
exit
!
interface eth 1/39 (just as an example)
ip address 39.39.39.2/24
ip policy route-map pbr-map
```

The configuration above does not use ACLs as filters to allow some traffic and block other traffic. Rather, the traffic that matches the ACLs is forwarded to the next-hop address specified, and traffic that does not match is routed normally, using existing static or dynamic routes in the switch's routing table.

When multiple next-hop addresses are configured, their priority is based on the order that they appear in the configuration.

Addresses configured as "next-hop verify-availability" are periodically checked to confirm that they are up, and are not used if they are down. Their priority is based on the sequence number in their configuration; next hop addresses with lower configuration numbers are used before higher ones.

Notes on PBR:

- PBR can be used within a VRF (discussed in the next section) but can not be used to forward traffic from one VRF to another.
- If PBR is used on an interface and the configuration of that interface changes it to a different VRF, then the PBR configuration will be applied in the new VRF.
- When PBR is used with vLAG and/or VRRP, it functions only on the active switch of the pair. If vLAG and VRRP are used together, it will function on both peer switches if configured on both of them.

VRF – Virtual Routing and Forwarding

VRF (virtual routing and forwarding) allows a switch to be partitioned into multiple L3 routing domains. The different domains each have their own routing tables, and it is possible for subnets to be duplicated and/or overlap in different domains. The current version of CNOS has a limit of 64 production data VRFs plus the management VRF.

Individual interfaces are configured to specify which VRF they are a member of; an interface can only participate in one VRF at a time. Only Layer 3 interfaces can be configured for membership in a VRF. This includes interfaces on which an IP address can be configured:

routed ports, and VLAN interfaces. Port-channel interfaces can not currently be configured with an IP address; to work around this, an otherwise unused VLAN would need to be configured.

Many CLI commands have operands such as **use-vrf** to specify which tables and routes to use. This is shown in the Command Reference manual, and may change in future releases.

In the current release, VRF routes from BGP and static routes are supported except in the default VRF. OSPF routes are only supported in the 'default' VRF. OSPF support is planned to be added in a future release.

VRFs are created and configured as shown in the examples in the following section.

Creation of VRFs

A VRF is created using the **vrf context** command, as shown in Example 11. Each VRF requires a route-distinguisher, a value used to allow multiple VRFs to add routes to the same destination to the routing table without conflict. This is configured with the **rd** command, and is arbitrary but must be unique to the switch on which it is used. Multiple formats for the rd are supported, including ASN2, ASN4, and IPv4 address.

Example 11 VRF creation

```
vrf context pod-1
rd 64001:1
!
vrf context pod-2
rd 64002:2
```

Assigning interfaces to a VRF

A layer 3 interface (one which can be assigned an IP address) can be a member of only one VRF. Unless specified otherwise, interfaces will be members of the `default` VRF except that the mgmt0 interface is always assigned to the `management` VRF.

An example of assigning other interfaces to the two VRFs created above in Example 11 is shown below in Example 12. Note that the same IP addresses are used for the interfaces in VRF pod-1 and those in VRF pod-2.

Example 12 Interface VRF assignment

```
NE10032-1-1#sho run int eth 1/13/1-4
!
interface Ethernet1/13/1
 no switchport
 vrf member pod-1
 ip address 10.1.1.1/30
!
!
interface Ethernet1/13/2
 no switchport
 vrf member pod-1
 ip address 10.1.2.1/30
!
!
interface Ethernet1/13/3
 no switchport
```

```
 vrf member pod-1
 ip address 10.1.3.1/30
!
!
interface Ethernet1/13/4
 no switchport
 vrf member pod-1
 ip address 10.1.4.1/30
 !
NE10032-1-1#sho run int eth 1/15/1-4
!
interface Ethernet1/15/1
 no switchport
 vrf member pod-2
 ip address 10.1.1.1/30
!
!
interface Ethernet1/15/2
 no switchport
 vrf member pod-2
 ip address 10.1.2.1/30
!
!
interface Ethernet1/15/3
 no switchport
 vrf member pod-2
 ip address 10.1.3.1/30
!
!
interface Ethernet1/15/4
 no switchport
 vrf member pod-2
 ip address 10.1.4.1/30
!
```

BGP Configuration with multiple VRFs

BGP can be used as a dynamic routing protocol simultaneously on more than one VRF, and can support the same route entries as part of different VRFs. This can enable support for multi-tenancy, although additional configuration parameters would also be required.

Just as interfaces are part of a VRF, BGP neighbors also are configured to show their VRF membership. However, only a single local autonomous system (60001) is supported, which can be seen in Example 13 below. Having neighbors in different autonomous systems – and thus deploying external BGP rather than internal BGP – is supported.

Note also that the same neighbor IP address is used for one neighbor in pod-1 and also in pod-2, but these neighbors are in fact physically distinct neighboring switches.

Example 13 BGP configuration with two VRFs

```
router bgp 60001
 vrf pod-1
  address-family ipv4 unicast
   redistribute direct
  neighbor 10.1.1.2 remote-as 60001
   address-family ipv4 unicast
  neighbor 10.1.2.2 remote-as 60001
   address-family ipv4 unicast
```

```
!
 vrf pod-2
  address-family ipv4 unicast
   redistribute direct
  neighbor 10.1.1.2 remote-as 60001
   address-family ipv4 unicast
  neighbor 10.1.4.2 remote-as 60001
   address-family ipv4 unicast
```

The routing tables that result from the above configuration and attaching to the four neighbors specified are shown in Example 14 below. Some of the routes are learned from being configured on the neighboring switches.

Example 14 BGP routing tables for two VRFs

```
NE10032-1-1#sho ip route vrf pod-1
Codes: C - connected, S - static, R - RIP, B - BGP
       O - OSPF, IA - OSPF inter area, D - DHCP
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       * - candidate default


IP Route Table for VRF "pod-1"
B       1.1.10.0/24 [200/0] via 10.1.1.2, Ethernet1/13/1, 00:30:39
B       1.1.20.0/24 [200/0] via 10.1.1.2, Ethernet1/13/1, 00:30:39
B       1.1.30.0/24 [200/0] via 10.1.1.2, Ethernet1/13/1, 00:30:39
C       10.1.1.0/30 is directly connected, Ethernet1/13/1
C       10.1.2.0/30 is directly connected, Ethernet1/13/2
C       10.1.3.0/30 is directly connected, Ethernet1/13/3
C       10.1.4.0/30 is directly connected, Ethernet1/13/4

Gateway of last resort is not set

NE10032-1-1#sho ip route vrf pod-1
Codes: C - connected, S - static, R - RIP, B - BGP
       O - OSPF, IA - OSPF inter area, D - DHCP
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       * - candidate default


IP Route Table for VRF "pod-2"
B       1.1.10.0/24 [200/0] via 10.1.4.2, Ethernet1/15/4, 00:22:44
B       1.1.20.0/24 [200/0] via 10.1.4.2, Ethernet1/15/4, 00:22:44
B       1.1.30.0/24 [200/0] via 10.1.4.2, Ethernet1/15/4, 00:22:44
C       10.1.1.0/30 is directly connected, Ethernet1/15/1
C       10.1.2.0/30 is directly connected, Ethernet1/15/2
C       10.1.3.0/30 is directly connected, Ethernet1/15/3
C       10.1.4.0/30 is directly connected, Ethernet1/15/4

Gateway of last resort is not set
```

In the route tables shown in Example 14, note that the two VRFs have routes to identical destinations, but each VRF only has routes that flow via interfaces that are members of that VRF. For example, the routes in pod-1 all flow over interfaces Ethernet 1/13/1 through 4, which are configured in that VRF as shown in Example 12 on page 16.

Additional discussion of VRFs

IP ARP entries also belong to the VRFs corresponding to the interfaces where they were learned; this is shown in Example 15. Just as there are routes to the same destination IP address in the route table, there are ARP entries for the same address. These nonetheless are on different switches, and have different MAC addresses.

Example 15 IP ARP tables for two VRFs

```
NE10032-1-1#sho ip arp vrf pod-1

Flags: D - Static Adjacencies attached to down interface

Current ARP configuration
ARP refresh: enabled
Global ARP timeout: 1500

IP ARP Table for context pod-1
Total number of entries: 2
Address          Age      MAC Address     Interface       State
10.1.1.2         00:01:04 a48c.dbe6.b201  Ethernet1/13/1  REACHABLE
10.1.2.2         00:16:05 a48c.dbe6.c001  Ethernet1/13/2  REACHABLE
NE10032-1-1#sho ip arp vrf pod-2

Flags: D - Static Adjacencies attached to down interface

Current ARP configuration
ARP refresh: enabled
Global ARP timeout: 1500

IP ARP Table for context pod-2
Total number of entries: 2
Address          Age      MAC Address     Interface       State
10.1.1.2         00:00:16 a48c.dbe6.b001  Ethernet1/15/1  REACHABLE
10.1.4.2         00:11:05 a48c.dbe6.d801  Ethernet1/15/4  REACHABLE
```

There are numerous other commands which can take a VRF name as an operand. For example, ping will use the default VRF unless a different VRF is specified. An example of this is shown in Example 16 below; one can ping the same address with different VRFs in the ping command, but it may or may not be reachable.

Example 16 Ping with VRF operand

```
NE10032-1-1#ping 1.1.10.32 vrf pod-2
PING 1.1.10.32 (1.1.10.32) 56(84) bytes of data.
64 bytes from 1.1.10.32: icmp_seq=1 ttl=64 time=0.204 ms
64 bytes from 1.1.10.32: icmp_seq=2 ttl=64 time=0.266 ms
^C
--- 1.1.10.32 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1037ms
rtt min/avg/max/mdev = 0.204/0.235/0.266/0.031 ms


NE10032-1-1#ping 1.1.10.32 vrf pod-1
PING 1.1.10.32 (1.1.10.32) 56(84) bytes of data.
^C
--- 1.1.10.32 ping statistics ---
2 packets transmitted, 0 received, 100% packet loss, time 1062ms
```

```
NE10032-1-1#ping 1.1.10.41 vrf pod-1
PING 1.1.10.41 (1.1.10.41) 56(84) bytes of data.
64 bytes from 1.1.10.41: icmp_seq=1 ttl=64 time=0.200 ms
64 bytes from 1.1.10.41: icmp_seq=2 ttl=64 time=0.199 ms
64 bytes from 1.1.10.41: icmp_seq=3 ttl=64 time=0.202 ms
^C
--- 1.1.10.41 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2074ms
rtt min/avg/max/mdev = 0.199/0.200/0.202/0.011 ms


NE10032-1-1#ping 1.1.10.41 vrf pod-2
PING 1.1.10.41 (1.1.10.41) 56(84) bytes of data.
^C
--- 1.1.10.41 ping statistics ---
2 packets transmitted, 0 received, 100% packet loss, time 1019ms
```

In summary: VRFs allow the creation of multiple routing domains within a switch. However, this is not the same as the Switch Partition (SPAR) feature which was available on some switches running our older ENOS firmware. The SPAR feature partitioned the switch based on physical interfaces, at Layer 2; the VRF feature partitions routing and related functions using IP addresses at Layer 3.

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:

- 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.

This document was created or updated on April 22, 2019.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/lp1087

Trademarks