Overview

This chapter contains the following sections:

Information About High Availability

The purpose of High Availability (HA) is to limit the impact of failures—both hardware and software— within a system. The Cisco NX-OS operating system is designed for high availability at the network, system, and service levels.

The following Cisco NX-OS features minimize or prevent traffic disruption in the event of a failure:

- Redundancy— redundancy at every aspect of the software architecture.
- Isolation of processes— isolation between software components to prevent a failure within one process disrupting other processes.
- Restartability—Most system functions and services are isolated so that they can be restarted independently after a failure while other services continue to run. In addition, most system services can perform stateful restarts, which allow the service to resume operations transparently to other services.

- Supervisor stateful switchover— Active/standby dual supervisor configuration. The state and configuration remain constantly synchronized between two Virtual Supervisor Modules (VSMs) to provide a seamless and stateful switchover in the event of a VSM failure.

Starting with Release 4.2(1)SV2(1.1), the high availability functionality is enhanced to support the split active and standby Cisco Nexus 1000V Virtual Supervisor Modules (VSMs) across two data centers to implement the cross-DC clusters and the VM mobility while ensuring high availability.

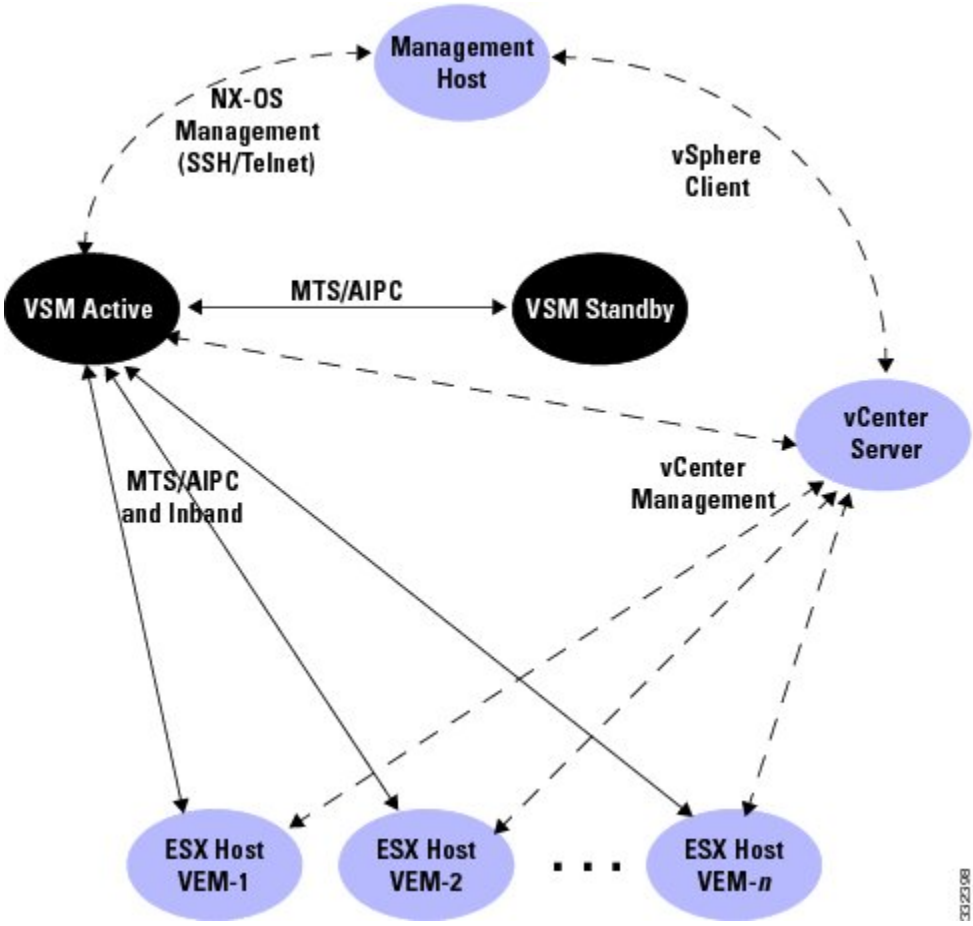
System Components

The Cisco Nexus 1000V system is made up of the following:

- One or two VSMs that run within Virtual Machines (VMs)
- Virtual Ethernet Modules (VEMs) that run within virtualization servers. VEMs are represented as modules within the VSM.
- A remote management component. For example, the VMware vCenter Server is a remote management component.

The following figure shows the HA components and the communication links between them.

Figure 1: HA Components and Communication Links



Service-Level High Availability

Isolation of Processes

The Cisco NX-OS software has independent processes, known as services, that perform a function or set of functions for a subsystem or feature set. Each service and service instance runs as an independent, protected process. This way of operating provides a highly fault-tolerant software infrastructure and fault isolation between services. A failure in a service instance will not affect any other services running at that time. Additionally, each instance of a service can run as an independent process, which means that two instances of a routing protocol can run as separate processes.

Process Restartability

Cisco NX-OS processes run in a protected memory space independently of each other and the kernel. This process isolation provides fault containment and enables rapid restarts. Process restartability ensures that process-level failures do not cause system-level failures. In addition, most services can perform stateful restarts, which allows a service that experiences a failure to be restarted and to resume operations transparently to other services within the platform and to neighboring devices within the network.

System-Level High Availability

The Cisco Nexus 1000V supports redundant VSM virtual machines—a primary and a secondary—running as an HA pair. Dual VSMs operate in an active/standby capacity in which only one of the VSMs is active at any given time, while the other acts as a standby backup. The VSMs are configured as either primary or secondary as a part of the Cisco Nexus 1000V installation. The state and configuration remain constantly synchronized between the two VSMs to provide a stateful switchover if the active VSM fails.

Network-Level High Availability

The Cisco Nexus 1000V high availability at the network level includes port channels and the Link Aggregation Control Protocol (LACP). A port channel bundles physical links into a channel group to create a single logical link that provides the aggregate bandwidth of up to eight physical links. If a member port within a port channel fails, the traffic previously carried over the failed link switches to the remaining member ports within the port channel.

Additionally, LACP lets you configure up to 16 interfaces into a port channel. A maximum of eight interfaces can be active, and a maximum of eight interfaces can be placed in a standby state.

For additional information about port channels and LACP, see the *Cisco Nexus 1000V Layer 2 Switching Configuration Guide*.

VSM to VSM Heartbeat

The primary and secondary VSM use a VSM to VSM heartbeat to do the following within their domain:

- Broadcast their presence
- Detect the presence of another VSM
- Negotiate active and standby redundancy states

When a VSM first boots up, it broadcasts discovery frames to the domain to detect the presence of another VSM. If no other VSM is found, the booting VSM becomes active. If another VSM is found to be active, the booting VSM becomes standby. If another VSM is found to be initializing (for example, during a system reload), the primary VSM has priority over the secondary to become active.

After the initial contact and role negotiation, the active and standby VSMs unicast the following in heartbeat messages at one-second intervals:

- Redundancy state

- Control flags requesting action by the other VSM

The types heartbeat messages that are sent at specific time intervals are listed in the following table.

Interval	Description
1 second	Interval at which heartbeat requests are sent.
3 seconds	Interval after which missed heartbeats indicate degraded communication on the control interface so that heartbeats are also sent on the management interface.
6 seconds	Interval of communication loss after which the active VSM instructs the standby to reload.

Control and Management Interface Redundancy

If the active VSM does not receive a heartbeat response over the control interface for a period of 3 seconds, then communication is seen as degraded and the VSM also begins sending requests over the management interface. In this case, the management interface provides redundancy only in the sense that it acts to prevent both VSMs from becoming active, also called an active-active or split-brain situation.

Note The communication is not fully redundant, however, because the management interface only handles heartbeat requests and responses.

AIPC and the synchronization of data between VSMs is done through the control interface only.

Partial Communication

If the communication over the control interface is interrupted, the secondary VSM is not rebooted. The heartbeat is attempted over the management interface. Assuming that the heartbeats can pass over the management interface, the VSMs enter into a degraded mode, as displayed in the **show system internal redundancy trace** command output.

Note A transition from active to standby always requires a reload in both the Cisco Nexus 1000V and the Cisco Nexus 1010.

Loss of Communication

When there is no communication between redundant VSMs or Cisco Nexus 1010s, neither can detect the presence of the other. The active drops the standby. The standby interprets the lack of response as a sign that the active has failed and it also becomes active. This is what is referred to as active-active or split-brain, as both are trying to control the system by connecting to Virtual Center and communicating with the VEMs.

Since redundant VSMs or Cisco Nexus 1010s use the same IP address for their management interface, remote SSH/Telnet connections may fail, as a result of the path to this IP address changing in the network. For this reason, it is recommended that you use the consoles during a split-brain conflict.

Starting with Release 4.2(1)SV2(1.1), the high availability mechanism on Cisco Nexus 1000V is enhanced to address this issue. The following parameters are used in selecting the VSM to be rebooted during the split-brain resolution: the module count, the vCenter connectivity status, the last configuration time, and the last active time.

Refer to the Split-Brain Resolution, on page 7 section for more information.

VSM-VEM Communication Loss

Depending on the specific network failure that caused it, each VSM may reach a different, possibly overlapping, subset of VEMs. When the VSM that was in standby state becomes a new active, it broadcasts a request to all VEMs to switch to it as the current active device. Whether a VEM switches to the new active VSM or not, depends on the following:

- The connectivity between each VEM and the two VSMs.
- Whether the VEM receives the request to switch.

A VEM remains attached to the original active VSM even if it receives heartbeats from the new active. However, if it also receives a request to switch from the new active, it detaches from the original active and attaches to the new one.

If a VEM loses connectivity to the original active device and only receives heartbeats from the new one, it ignores those heartbeats until it goes into headless mode. This occurs approximately 15 seconds after it stops receiving heartbeats from the original active. At that point, the VEM attaches to the new active if it has connectivity to it.

Note If a VEM loses the connection to its VSM, the Vmotions to that particular VEM are blocked. The VEM shows up in the VCenter Server as having a degraded (yellow) status.

One-way Communication

If there is a network communication failure where the standby VSM receives heartbeat requests but the active does not receive a response, the following occurs:

- The active VSM declares that the standby VSM is not present.
- The standby VSM remains in standby state and continues receiving heartbeats from the active VSM

The redundancy state is inconsistent (**show system redundancy state**) and the two VSMs lose synchronization.

Note If a one-way communication failure occurs in the active to standby direction, it is equivalent to a total loss of communication. This is because a standby VSM only sends heartbeats in response to active VSM requests.

Split-Brain Resolution

When the connectivity between the Virtual Supervisor Modules (VSMs) is broken, the loss of communication can cause both VSMs to take the active role. This condition is also called as the active-active or the split-brain resolution. When the communication is restored between the VSMs, both VSMs attempt to resolve the split-brain resolution by rebooting the primary VSM.

The method of resolving the split-brain resolution by rebooting the primary VSM helps in the scenarios, when the secondary VSM has a proper configuration and all the VEMs are properly connected. If the secondary VSM does not have a proper configuration, it may overwrite the valid configuration of the primary VSM. As a result, both VSMs may have an invalid configuration after the split-brain resolution.

Starting with Release 4.2(1)SV2(1.1), the high availability functionality on Cisco Nexus 1000V is enhanced to address this issue. Both primary and secondary VSMs process the same data to select the VSM (primary/secondary) that needs to be rebooted. When the selected VSM is rebooted and attaches back itself, the high availability functionality comes back normal. The following parameters are used in order of their precedence to select the VSM to be rebooted during the split-brain resolution:

- Last configuration time: The time when the last configuration is done on the VSM.
- Module count: The number of modules attached to the VSM.
- vCenter (VC) status: Status of the connection between the VSM and vCenter
- Last active time: The time when the VSM becomes active.

Checking the Accounting Logs and the Redundancy Traces

During the split-brain resolution, when a VSM is rebooted, the accounting logs that are stored on the VSM are lost. Starting with Release 4.2(1)SV2(1.1), new CLI commands are supported to display the accounting logs that were backed up during the split-brain resolution. You can also check the redundancy traces that are stored on the local and remote VSMs.

Use the following commands to check the accounting logs that were backed up and the redundancy traces that were stored during the split-brain resolution:

Procedure

Step 1 n1000v# **show system internal active-active accounting logs**
Displays the accounting logs that are stored on a local VSM during the last split-brain resolution.

Step 2 n1000v# **show system internal active-active redundancy traces**
Displays the redundancy traces that are stored on a local VSM during the last split-brain resolution.

Step 3 n1000v# **show system internal active-active remote accounting logs**
Displays the remote accounting logs that are stored on a remote VSM during the last split-brain resolution.

Step 4 n1000v# **show system internal active-active remote redundancy traces**
Displays the remote redundancy traces that are stored on a remote VSM during the last split-brain resolution.

Step 5 n1000v# **clear active-active accounting logs**
Clears the accounting logs that are stored on a local VSM during the split-brain resolution.

Step 6 n1000v# **clear active-active redundancy traces**
Clears the redundancy traces that are stored on a local VSM during the split-brain resolution.

Step 7 n1000v# **clear active-active remote accounting logs**
Clears the remote accounting logs that are stored on a remote VSM during the split-brain resolution.

Step 8 n1000v# **clear active-active remote redundancy traces**
Clears the remote redundancy traces that are stored on a remote VSM during the split-brain resolution.

VSM Role Collision Detection

In the Cisco Nexus 1000V switch, if a VSM is configured or installed with the same role as the existing VSM and with the same domain ID, the new VSM and the existing VSM exchange heartbeats to discover each other. Both the VSMs detect a role collision upon processing the exchanged heartbeats.

Due to this issue, the remote management component, VEM, and the HA-paired VSM cannot communicate with the correct VSM. This issue can occur on a primary or a secondary VSM depending on whether the newly configured or the installed VSM has the primary or the secondary role assigned to it.

The collisions are detected on the control and the management interfaces. The number of role collisions is restricted to 8.

Note The colliding VSMs may also report a collision detection from the original VSM. Since the colliding VSMs may use the same IP address for their management interfaces, the remote SSH/Telnet connections may fail. Therefore, we recommend that you use the consoles during a role collision detection.

After the successful high availability configuration, execute the CLI command **show system redundancy status** on both primary and secondary VSM consoles. When the colliding VSM stops communicating in the domain, the collision time is not updated anymore. After an hour has elapsed since the last collision, the collision MAC entries are removed.

An appropriate action, for example, changing the domain or shutting down the switch, should be taken on the colliding VSMs to ensure the proper operation of the Cisco Nexus 1000V switch.

Displaying Role Collision

Use the **show system redundancy status** CLI command to display the VSM role collision:

Procedure

n1000v# **show system redundancy status**
When a role collision is detected, a warning is highlighted in the CLI output. Along with the MAC addresses, the latest collision time is also displayed. If no collisions are detected, the highlighted output is not displayed.

In the following example, the detected traffic collision is displayed with the highlighted output:

```
n1000v# show system redundancy status

Redundancy role
---------------
      administrative:   secondary
         operational:   secondary

Redundancy mode
---------------
      administrative:   HA
         operational:   HA

This supervisor (sup-2)
-----------------------
    Redundancy state:   Active
    Supervisor state:   Active
      Internal state:   Active with HA standby

Other supervisor (sup-1)
------------------------
    Redundancy state:   Standby
    Supervisor state:   HA standby
      Internal state:   HA standby
```

WARNING! Conflicting sup-2(s) detected in same domain
--
MAC Latest Collision Time
00:50:56:97:02:3b 2012-Sep-11 18:59:17
00:50:56:97:02:3c 2012-Sep-11 18:59:17
00:50:56:97:02:2f 2012-Sep-11 18:57:42
00:50:56:97:02:35 2012-Sep-11 18:57:46
00:50:56:97:02:29 2012-Sep-11 18:57:36
00:50:56:97:02:30 2012-Sep-11 18:57:42
00:50:56:97:02:36 2012-Sep-11 18:57:46
00:50:56:97:02:2a 2012-Sep-11 18:57:36

NOTE: Please run the same command on sup-1 to check for conflicting(if any) sup-1(s) in the same domain.

Recommended Reading

- *Cisco Nexus 1000V Installation and Upgrade Guide*
- *Cisco Nexus 1000V Port Profile Configuration Guide*