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.

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. Microsoft SCVMM Server.

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 does not affect any other services that are 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. These stateful restarts allow 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 allows you to 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.

VSM NIC Ordering

The VSM creates interfaces in an ascending MAC order of the virtual NIC offered by Microsoft Hyper-V. Currently, Microsoft Hyper V provides no guarantees that this order is the same as displayed at the VSM VM Settings panel, but this is usually the case. VSM always uses its first interface as control0 and its second interface as mgmt0. The network profiles for these two interfaces may need to use different VLAN. Therefore, the users should verify that the interfaces are selected by the VSM in the same order as displayed in the Settings panel, to select profiles appropriately.

If the order is not the same, the users can use the following commands to specify their preferred MAC to control0 / mgmt0 interface mappings.

- **system internal control-mac XXXX.XXXX.XXXX**
- **system internal mgmt-mac XXXX.XXXX.XXXX**

These commands require **copy running-config startup-config** to be run afterwards to make the change persistent and effective after the next VSM reload.

show system internal interface mac-address

```
Sample output:
Interface Preferred MAC
--------- --------------
mgmt0 cccc.bbbb.aaaa
control0 aaaa.bbbb.cccc
```

Note Actual MAC values for currently selected control0/mgmt0 interfaces can be displayed with existing command “show interface mac-address”. After a VSM reload, both mappings should be the same, unless the system malfunctions.

If any of the preferred MAC for control0/mgmt0 selected by users is not available at VSM boot up, the driver ignores it and it picks another interface instead (following MAC ascending order). In that case, the system logs an error with a syslog as follows:

```
%KERN-3-SYSTEM_MSG: Preferred MAC (aaaa.bbbb.cccc) for control0 not found - kernel
```

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 the standby VSM. If another VSM is found to be initializing (for example, during a system reload), the primary VSM has priority over the secondary to become the active VSM.

Note From SV3(1.1) onwards, VSM validates the source MAC address of High Availability (HA) packets it receives on control and management interfaces. During initial contact VSM learns the peer VSM MAC addresses and stores in a permanent location. So only HA packets from a learnt VSM will be accepted.

After the initial contact and role negotiation, the active and standby VSMs unicast the following in heartbeat messages:

- Redundancy state
- Control flags requesting action by the other VSM

The following intervals apply when sending heartbeat messages.

Interval	Description
1 second	Interval at which heartbeat requests are sent.
8 seconds	Interval after which missed heartbeats indicate degraded communication on the control interface so that heartbeats are also sent on the management interface.
varies	The standby VSM resets automatically when the active VSM is no longer able to synchronize with it. This means the interval varies depending on how long the active VSM can buffer the data to be synchronized when the communication is interrupted.

Control and Management Interface Redundancy

The VSM communicates with the peer VSM over layer 2 only on the control and management interfaces. If the active VSM does not receive a heartbeat response over the control interface for a period of half of the inter-VSM maximum heartbeat loss interval (eight heartbeats by default), communication is seen as degraded and the VSM begins sending requests over the management interface in addition to the control interface. In this case, the management interface provides redundancy by preventing both VSMs from becoming active. This process is 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

The secondary VSM is not immediately rebooted when communication over the control interface is interrupted because the HA mechanism tolerates brief interruptions in communication. When communication is first interrupted on the control interface, the heartbeat messages are sent over the management interface. If communication over the management interface is successful, the VSMs enter into a degraded mode, as displayed in the **show system internal redundancy trace** command output. If communication is interrupted on both interfaces for too long, the two VSMs get out of synchronization and the standby VSM is forced to reboot.

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

Loss of Communication

When there is no communication between redundant VSMs , they cannot detect the presence of the other. The standby VSM will be removed from the list of inserted modules at the active VSM. The standby interprets the lack of heartbeats as a sign that the active has failed and it also becomes active. This process is what is referred to as active-active or split-brain, as both are trying to control the system by connecting to SCVMM and communicating with the VEMs.

Because redundant VSMs use the same IP address for their management interface, remote Secure Shell (SSH)/Telnet connections might fail, as a result of the path to this IP address changing in the network. For this reason, we recommend that you use the consoles during a split-brain conflict.

The following parameters are used to select the VSM to be rebooted during the split-brain resolution: the module count, the last configuration time, and the last active time.

VSM-VEM Communication Loss

Depending on the specific network failure that caused it, each VSM might reach a different, possibly overlapping, subset of Virtual Ethernet Modules (VEMs). When the VSM that was in the standby state becomes a new active VSM, 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, 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 VSM. However, if the VEM also receives a request to switch from the new active VSM, it detaches from the original active VSM and attaches to the new VSM.

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, which occurs approximately 15 seconds after it stops receiving heartbeats from the original, active VSM. At that point, the VEM attaches to the new active VSM if it has connectivity to it.

Note If a VEM loses the connection to its VSM, Live Migrationsthat particular VEM is blocked. The VEM shows SCVMM a degraded (yellow) status.

One-Way Communication

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

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

In this scenario, the redundancy state is inconsistent (**show system redundancy state**) and the two VSMs lose synchronization. When two-way communication is resumed, the standby VSM replies to the active VSM and asks to be reset.

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

Split-Brain Resolution

When the connectivity between two Virtual Supervisor Modules (VSMs) is broken, this loss of communication can cause both VSMs to take the active role. This condition is called active-active or split-brain condition. When the communication is restored between the VSMs, both VSMs exchange information to decide which one would have a lesser impact on the system, if rebooted.

Both primary and secondary VSMs process the same data to select the VSM (primary or secondary) that needs to be rebooted. When the selected VSM is rebooted and attaches itself back to the system, high availability is back to normal. The VSM uses the following parameters in order of precedence to select the VSM to be rebooted during the split-brain resolution:

1. Module count—The number of modules that are attached to the VSM.
2. Last configuration time—The time when the last configuration is done on the VSM.
3. Last standby-active switch—The time when the VSM last switched from the standby state to the active state. (The VSM with a longer active time gets higher priority.)

Checking the Accounting Logs and the Redundancy Traces

During the split-brain resolution, when a VSM reboots, the accounting logs that are stored on the VSM are lost. You can 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.

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, if a secondary VSM is configured or installed with the same role as the primary VSM and with the same domain ID, the secondary VSM and the primary VSM exchange heartbeats to discover each other. Both VSMs detect and report a role collision when they exchange heartbeats. When a collision is detected, the VSMs report the MAC address of the VSM with which the local VSM is colliding.

Due to this issue, the HA-paired VSM cannot communicate with the correct VSM. This problem can occur on a primary VSM 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 maximum number of colliding VSMs reported is eight.

Note After the eighth role collision, the problem is still logged and the MAC address entry is overwritten. The **show system redundancy status** command displays the overwrite details.

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

Enter the **show system redundancy status** command on both the primary and secondary VSM consoles to display the MAC addresses of the detected VSMs with the same role and domain ID, if any. When the VSM stops communicating in the domain, the collision time is not updated anymore. After an hour elapses since the last collision, the collision MAC entries are removed.

Displaying the Role Collision

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

Procedure

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

This example shows how to display the detected traffic collision:

```
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.
```