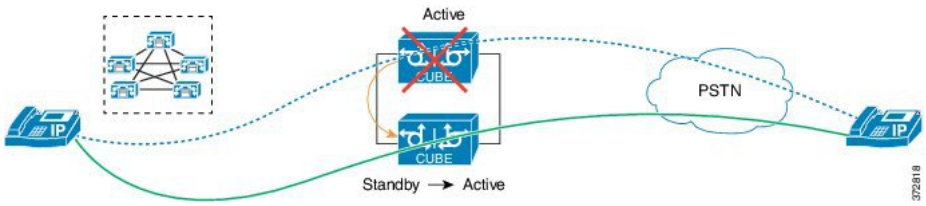
High Availability on Cisco CSR 1000V or C8000V Cloud Services Routers

Note Cisco Cloud Services Router 1000V Series (CSR 1000V) is no longer supported from Cisco IOS XE Bengaluru 17.4.1a onwards. If you are using CSR 1000V, you have to upgrade to Cisco Catalyst 8000V Edge Software (Catalyst 8000V). For End-of-Life information on CSR 1000V, see End-of-Sale and End-of-Life Announcement for the Select Cisco CSR 1000v Licenses.

The High Availability (HA) feature allows you to benefit from the failover capability of Cisco Unified Border Element (CUBE) on two routers, one active and one standby. When the active router goes down for any reason, the standby router takes over seamlessly, preserving and processing your calls.

Figure 1: Cisco CUBE High Availability



About vCUBE High Availability on CSR 1000V or C8000V Cloud Services Routers

Cisco Unified Border Element running on Cisco CSR 1000v Series Cloud Services Router and C8000V is called Virtual CUBE (vCUBE). vCUBE leverages Redundancy Group (RG) Infrastructure to provide high availability. HA is configured between two vCUBE Cisco CSR 1000v or C8000V instances running on either the same host or across different hosts that are connected through the same switch.

You can configure vCUBE on Cisco CSR 1000v or C8000V running on virtualized hosts listed in the Cisco Unified Border Element Data Sheet.

Box-to-Box Redundancy

Box-to-box redundancy enables configuring a pair of routers to act as back up for each other. In the router pair, the active router is determined based on the failover conditions. The router pair continuously exchange status messages. CUBE session information is checkpointed across the active and standby router. This enables the standby router to immediately take over all CUBE call processing responsibilities when the active router becomes unavailable.

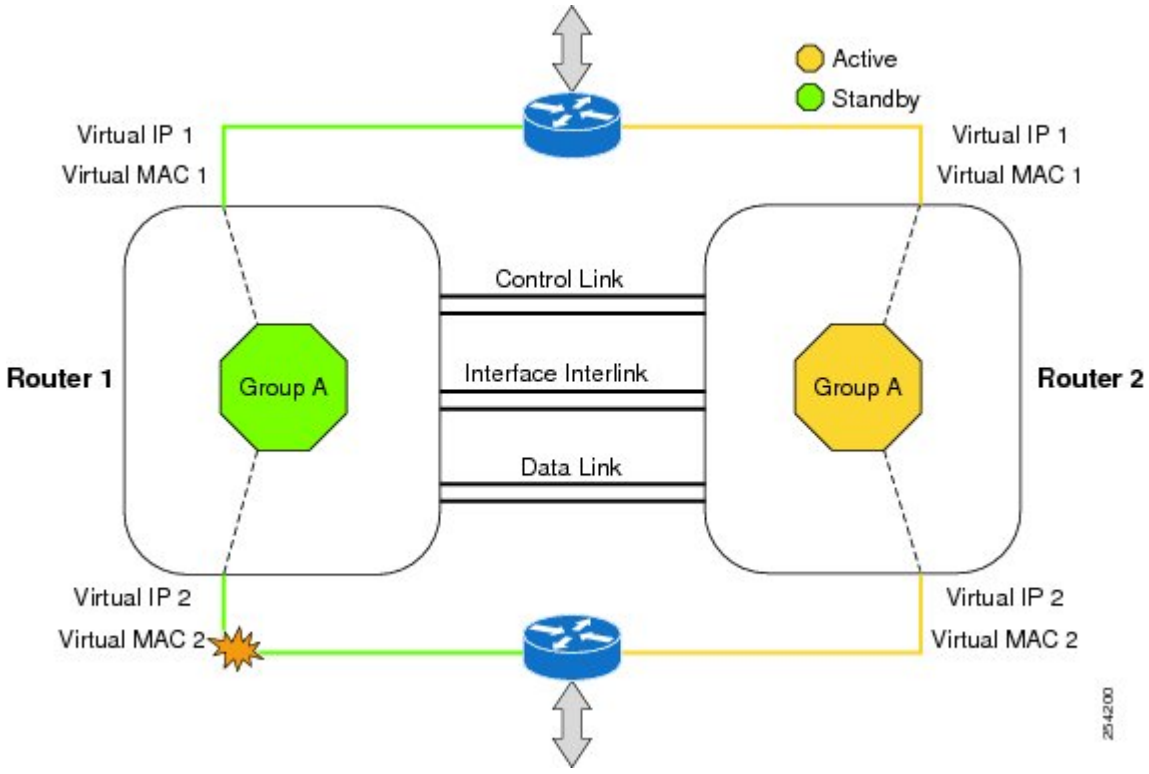
Redundancy Group (RG) Infrastructure

A group of redundant interfaces form a Redundancy Group. The active and standby routers are connected by a configurable control link and data synchronization link. The control link is used to communicate the redundancy state for each router. The data synchronization link is used to transfer stateful information to synchronize the stateful database for the calls and media flows. Each pair of redundant interfaces is configured with the same unique ID number, also known as the Redundancy Interface Identifier (RII).

A Virtual IP address (VIP) is configured on interfaces that connect to the external network. All signaling and media is sourced from and sent to the Virtual IP address. External devices such as Cisco Unified Communication Manager, uses VIP as the destination IP address for the calls traversing through Cisco UBE.

The following figure shows the redundancy group configured for a pair of routers with a single outgoing interface.

Figure 2: Redundancy Group Configuration



Network Topology

Figure 3: Virtual CUBE High Availability

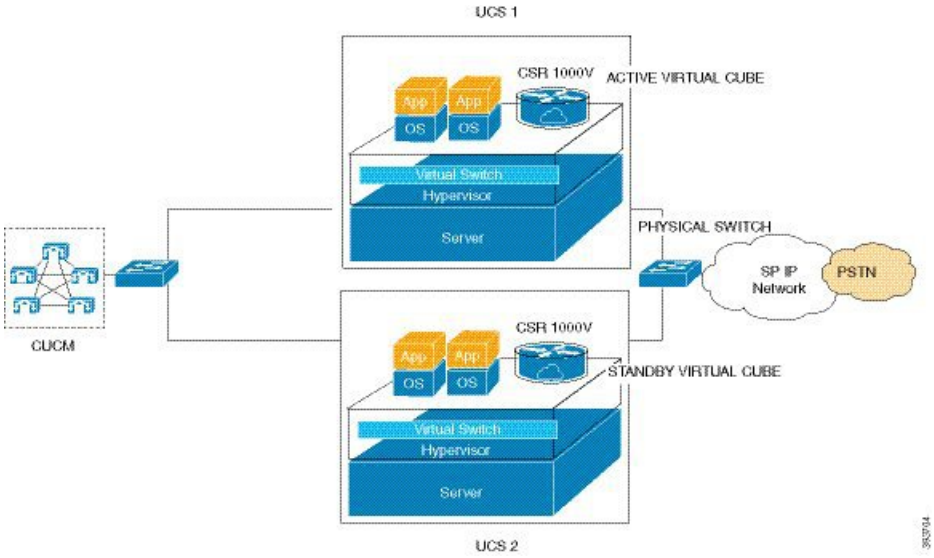
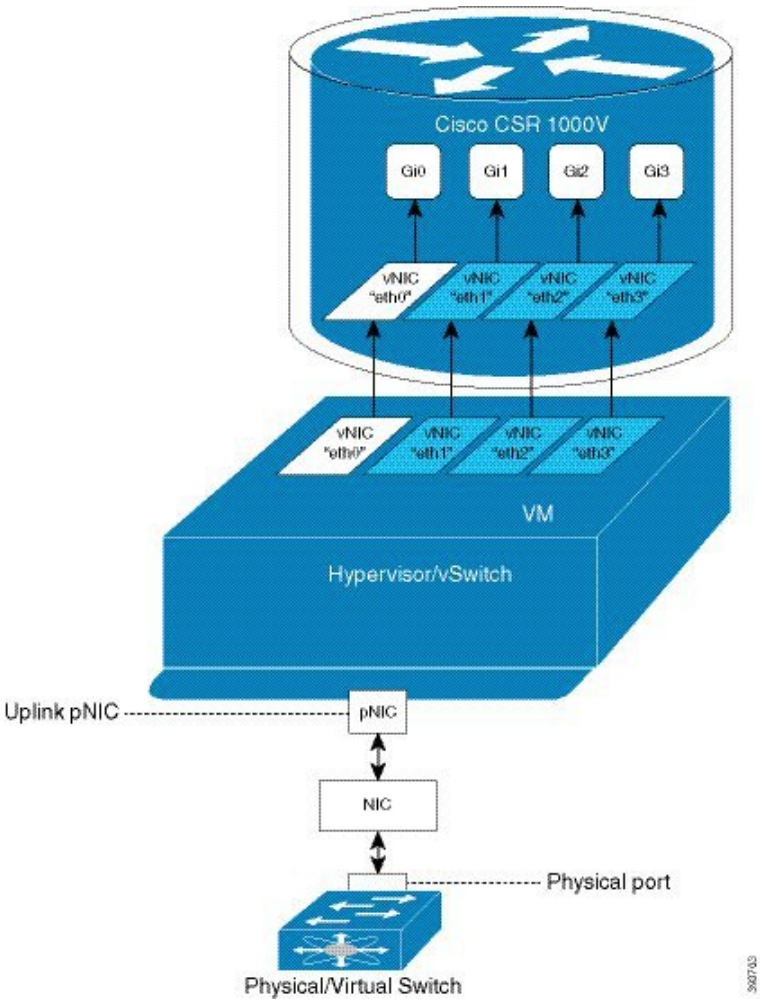


Figure 4: vNICs Mapped to Cisco CSR 1000V Router Interfaces



We recommend that you keep the following in mind when enabling this topology:

- Connect the Cisco CSR 1000v or C8000V running on the server to the virtual switch within the virtualized host. Then connect the virtual switches to external switches using the physical host interfaces. The virtual switch routes the traffic internally between the virtual machines and also connects the external networks.
- Configure high-availability connectivity using redundancy on virtual switch to avoid checkpointing failures.

  In a scenario where the physical switch is down and there is no redundancy configured on virtual switch, the active router continues to process calls as it tracks only the status of virtual switch (which is up). At the same time, the standby router assumes the role of active router as it does not receive keepalive messages from the active router through the physical switch. Hence checkpointing fails. To avoid such scenarios, we recommend you to configure high availability connectivity using redundancy on virtual switch.
- Do not track the switches that are used to connect non-networking end devices or LAN, to determine uplink failures.

- Connect the redundancy group control and data interfaces in the CUBE HA pair to the same physical switch to avoid any latency in the network.
- The RG control and data interfaces of the CUBE HA pair can be connected through a back-to-back cable or using a switch. However, it is recommended to use Portchannel for the RG control and data interfaces for redundancy. A single connection using back-to-back cable or switch presents a single point of failure due to a faulty cable, port, or switch, resulting in error state where both routers are Active.
- If the RG ID is the same for the two different CUBE HA pairs, keepalive interface for check-pointing the RG control and data, and traffic must be in a different subnet or VLAN.

Note This recommendation is applicable only if you connect using a switch, not by back-to-back cables.

- You can configure a maximum of two redundancy groups. Hence, there can be only two Active and Standby pairs within the same network.

Note This recommendation is applicable only if you connect using a switch, not by back-to-back cables.

- Source all signaling and media from and to the virtual IP address.
- Always save the running configuration to avoid losing it due to router reload during a failover.
- Virtual Routing and Forwarding
  - Define Virtual Router Forwarding (VRF) in the same order on both Active and Standby routers for an accurate synchronization of data.
  - You can configure VRFs only on the traffic interface (SIP and RTP). Do not configure VRF on redundancy group control and data interface.
  - VRF configurations on both the Active and Standby router must be identical. VRF IDs checkpoints for the calls before and after switchover (includes VRF-based RTP port range).
- Manually copy the configurations from one router to the other.
- Replicating the configuration on the Standby router does not commit to the startup configuration; it is the running configuration. You must run the **write memory** command to commit the changes that are synchronized from the Active router on the Standby router.

Considerations and Restrictions

The following is a list of further considerations and restrictions you should know before configuring this topology:

Considerations

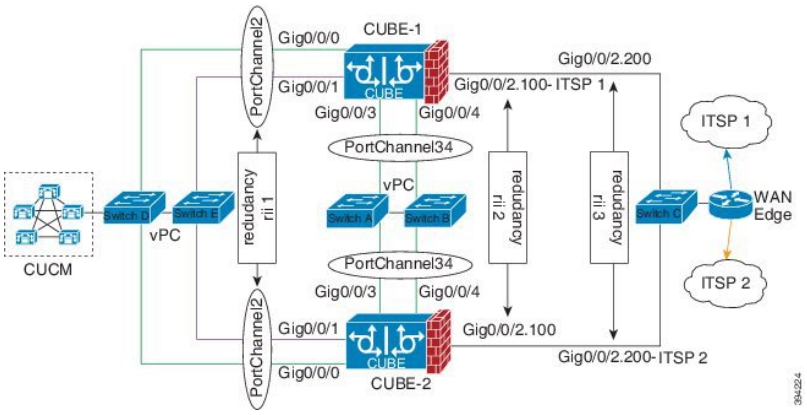
Before you configure High Availability feature on Cisco UBE, understand its behavior, and also know the Cisco IOS XE Software version that is required for supporting the call processing features to work seamlessly after switchover.

- Only active calls are checkpointed (Calls that are connected with 200 OK or ACK transaction completed).
- When you apply and save the configuration for the first time, the platform must be reloaded.
- For H.323, and TCP-based calls, media preservation is supported after the failover, but session signaling is not preserved.
- If you have Cisco Unified Customer Voice Portal (CVP) in your network, we recommend that you configure TCP session transport for the SIP trunk between CVP and CUBE.
- Upon failover, the previously active CUBE reloads by design.
- CUBE uses the virtual IP address to communicate Smart Licensing information.
- For SIP-SIP TLS calls, configure both the active and standby CUBE as trust points to a common external CA Server.
- TCP sessions are not preserved during the failover. Remote user agents are expected to reestablish TCP sessions (using port 5060) before sending subsequent messages.
- Call Admission Control (CAC) state is maintained through switchover. After Stateful Switchover, no calls are allowed if the CAC limit is reached before the switchover.
- Up to six multimedia lines in the SDP are checkpointed for CUBE high availability. From Cisco IOS XE Release 3.17 onwards, SDP Passthru (up to two m-lines) calls are also checkpointed.
- Survivability.tcl preservation is supported from Cisco IOS XE Release 3.17 onwards for Unified Customer Voice Portal (CVP) deployments.
- SRTP-RTP, SRTP-SRTP, and SRTP Passthru are supported.

Note Redundancy control traffic that is exchanged between CUBE-1 and CUBE-2 is not secured natively and displays SRTP encryption keys in cleartext. If SRTP is used, you must secure this traffic by configuring a transport IPsec tunnel between the two interfaces that are used as the redundancy control link.

- Port channel is supported for both RG control data and traffic interfaces only from Cisco IOS XE 16.3.1 onwards.

Figure 5: Additional Supported Options for CUBE HA



- While deploying High Availability pair with Application Centric Infrastructure (ACI), perform one of the following:
  - Disable IP data plane learning on the VRF.

    Refer to IP Data-plane Learning for details.
  - Use an intermediate Layer 3 switch between the High Availability pair and the ACI deployment. This Layer 3 switch prevents the ACI from directly learning the CUBE IP address and its associated MAC addresses.

Restrictions

- Geographic stateful switchover is not supported.
- Calls in the transient state at the time of switchover are not preserved.
- IPv6 is not supported.
- All SCCP-based media resources (Conference bridge, Transcoding, Hardware MTP, and Software MTP) are not supported.
- Cisco Unified Survivable Remote Site Telephony (Cisco Unified SRST) or TDM Gateway colocation on CUBE HA is not supported.
- Routers connected through Metropolitan Area Network (MAN) Ethernet regardless of latency are not supported.
- Out-of-band DTMF (Notify or KPML) is not supported post switchover. Only rtp-nte to rtp-nte and voice-inband to voice-inband DTMF works after the switchover.
- Media-flow around and UC Services API (Cisco Unified Communications Manager Network-Based Recording) are not supported.
- You cannot terminate Wide Area Network (WAN) on CUBE directly or Data HA on either side. Both active and standby routers must be in the same Data Center and connected to the same physical switch.
- The Courtesy Callback (CCB) feature is not supported if a callback was registered with Cisco Unified Customer Voice Portal (CVP) and then a switchover was done on CUBE.

- You cannot configure a secondary IP address for the interfaces.
- If the redundancy group ID is same for the two different CUBE HA pairs, then the keepalive interface that is used for checkpointing RG control and data traffic must be in a different subnet or VLAN.

How to Configure vCUBE High Availability on Cisco CSR 1000v or C8000V

Before You Begin

- Use Cisco IOS-XE Release 3.11 and or later on both active and standby routers.
- Ensure that you have the required licenses for configuring high availability. For detailed information, see Cisco Unified Border Element Data Sheet.

Configure High Availability

Procedure

	Command or Action	Purpose
Step 1	Configure the Redundancy Group.	
Step 2	Configure the interfaces.	

Step 3	Configure SIP Binding.
	Example:
	
	`Router(config)#dial-peer voice 1 voip`
	`Router(config-dial-peer)#session protocol sipv2`
	`Router(config-dial-peer)#incoming called-number 2000`
	`Router(config-dial-peer)#voice-class sip bind control source-interface GigabitEthernet0/0/0`
	`Router(config-dial-peer)#voice-class sip bind media source-interface GigabitEthernet0/0/0`
	`Router(config-dial-peer)#codec g711ulaw`
	`Router(config-dial-peer)#!`
	
	`Router(config)#dial-peer voice 2 voip`
	`Router(config-dial-peer)#destination-pattern 2000`
	`Router(config-dial-peer)#session protocol sipv2`
	`Router(config-dial-peer)#session target ipv4:203.0.113.13`
	`Router(config-dial-peer)#voice-class sip bind control source-interface GigabitEthernet0/0/1`
	`Router(config-dial-peer)#voice-class sip bind media source-interface GigabitEthernet0/0/1`
	`Router(config-dial-peer)#codec g711ulaw`
		Configure CUBE to bind SIP messages to the interface that is configured with a virtual IP address (VIP) for the Redundancy Group employed.

	Command or Action	Purpose
Step 4	Configure the Punt Policing feature. **Example:** `Router(config)#platform punt-policer 60 40000` In the preceding example, the punt-rate of the virtual IP address (punt-cause 60) is increased from the default value of 2000–40000.	SIP packets toward the virtual IP address and physical IP address match different punt-cause codes. The punt-rate of the virtual IP address with a punt-cause of 60, is lower than the punt-rate of the physical IP address. To ensure that the behaviour of the SIP packets toward virtual and physical IP address remains the same, you must increase the punt-rate of the virtual IP address by using the **platform punt-policer** command in global configuration mode. **Note** For Cisco IOS XE Releases 16.6.7, 16.9.4, 16.11.1, 16.12.1, 17.1.1 and later releases, you do not need to increase the punt-rate. The following table provides details of the fields of the CLI. **Note** The default punt rate value of the virtual IP address and the physical IP address varies with the router platform. **Note** The default and maximum setting are platform-specific. Default value is optimal for most deployments. Change the rate only when suggested by Cisco Support.

Step 5	Configure the Redundancy Group under **voice service voip**. This Reduncy Group creation enables Box-to-Box CUBE high availability.
	Example:
	`Router#voice service voip`
	`Router(conf-voi-serv)#redundancy-group 1`
		For enabling protected mode:
		`Router#voice service voip`
		`Router(conf-voi-serv)#no redundancy-reload`

Step 6	Configure the Media Inactivity timer. **Example:**	The Media Inactivity Timer enables the Active and Standby router pair to monitor and disconnect calls if no Real-Time Protocol (RTP) packets are received within a configurable time period.

Keyword	Description
platform punt-policer	Configures the Punt Policing feature.
60	*punt-cause*—Punt cause. Range is 1–107. Punt cause of the virtual interface is 60.
40000	*punt-rate*—Rate limit in packets per second. Range is 10–146484.

	Command or Action	Purpose

	`Router(config)#ip rtcp report interval 9000`
	`Router(config)#gateway`
	`Router(config-gateway)#media-inactivity-criteria all`
	`Router(config-gateway)#timer receive-rtp 1200`
	`Router(config-gateway)#timer receive-rtcp 5`
		For the SIP calls, the switched over calls are cleared with signaling (as signaling information is preserved for switched calls).
		The Media Inactivity Timer releases TCP-based calls. This is used to guard against any hung sessions resulting from the failover when a normal call disconnect does not clear the call.
		You must configure the same duration for the Media Inactivity Timer on both routers. The default value is 30 seconds for SIP calls. The sample configuration is as follows:
		SIP call legs are cleared once the RTCP timer expires.

Step 7	Reload the router.
	Example:
	`Router>enable`
	`Router#relaod`
		Once all the preceding configurations are completed, you must save the configurations, and reload the router.

Step 8	Configure the peer router.	Follow the preceding steps to configure the standby router. Make sure that you use the correct IP addresses.
Step 9	Point the attached devices to the CUBE Virtual IP (VIP) address.	The IP-PBX, Unified SIP Proxy, or service provider must route the calls to CUBE's virtual IP address. High availability configuration does not handle SIP messages to the CUBE's physical IP addresses.

Configuration Example

Active Router:

```
voice service voip
 no ip address trusted authenticate
 allow-connections sip to sip
 redundancy-group 1
sip
  bind control source-interface GigabitEthernet1
  bind media source-interface GigabitEthernet1
!
redundancy
 application redundancy
  group 1
   name cube_b2b_ha_1
   priority 125 failover threshold 75
   timers delay 30 reload 60
   control GigabitEthernet2 protocol 1
   data GigabitEthernet2
   track 1 shutdown
  protocol 1
   name cube_b2b_ha_1
   authentication text sol_ha1
!
```

```
track 1 interface GigabitEthernet1 line-protocol
!
interface GigabitEthernet1
 ip address 192.0.2.1 255.255.255.0
 negotiation auto
 no mop enabled
 no mop sysid
 redundancy rii 102
 redundancy group 1 ip 192.0.2.3 exclusive
!
interface GigabitEthernet2
 ip address 198.51.100.1 255.255.255.0
 negotiation auto
 no mop enabled
 no mop sysid
```

Standby Router:

```
 voice service voip
 no ip address trusted authenticate
 allow-connections sip to sip
 redundancy-group 2
sip
  bind control source-interface GigabitEthernet1
  bind media source-interface GigabitEthernet1
!
redundancy
 application redundancy
  group 2
   name cube_b2b_ha_1
   priority 100 failover threshold 75
   timers delay 30 reload 60
   control GigabitEthernet2 protocol 1
   data GigabitEthernet2
   track 1 shutdown
  protocol 1
   name cube_b2b_ha_1
   authentication text sol_ha1
!
track 1 interface GigabitEthernet1 line-protocol
!
interface GigabitEthernet1
 ip address 192.0.2.2 255.255.255.0
 negotiation auto
 no mop enabled
 no mop sysid
 redundancy rii 102
 redundancy group 2 ip 192.0.2.3 exclusive
!
interface GigabitEthernet2
 ip address 198.51.100.2 255.255.255.0
 negotiation auto
 no mop enabled
 no mop sysid
```

Troubleshoot High Availability Issues

Use the following show and debug commands to troubleshoot High Availability issues:

- **show redundancy application group all**
- **show redundancy application transport clients**

- **show redundancy client domain all | inc VOIP RG**
- **show voice high-availability summary**
- **show voip fpi stats**
- **debug voip rtp session**
- **debug voice high-availability all**
- **debug voip fpi all**
- **debug redundancey application group {config | faults | media | protocol | rii transport | vp}**

Note Do not turn on a large number of debugs on a system carrying high volume of active call traffic.