CHAPTER 4

Designing Unified CVP for High Availability

Last revised on: August 18, 2009

This chapter describes guidelines and best practices for designing a high-availability Unified CVP system.

The chapter covers the following topics:

What's New in This Chapter

Table 4-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 4-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:
H.323 gatekeeper	Gatekeeper, page 4-11
SIP call routing using static routes	SIP Proxy, page 4-5
UDP transport on SIP trunks	UDP Transport as a SIP Best Practice, page 4-9

Overview

A high-availability design provides the highest level of failure protection. Your solution may vary depending upon business needs such as:

- Tolerance for call failure
- Budget
- Topological considerations

Unified CVP can be deployed in many configurations that use numerous hardware and software components. Each solution must be designed in such a way that a failure impacts the fewest resources in the call center. The type and number of resources impacted depends on how stringent the business requirements are and which design characteristics you choose for the various Unified CVP components, including the network infrastructure. A good Unified CVP design is tolerant of most failures (defined later in this chapter), but sometimes not all failures can be made transparent to the caller.

Unified CVP is a sophisticated solution designed for mission-critical call centers. The success of any Unified CVP deployment requires a team with experience in data and voice internetworking, system administration, and Unified CVP application configuration.

Before implementing Unified CVP, use careful preparation and design planning to avoid costly upgrades or maintenance later in the deployment cycle. Always design for the worst possible failure scenario, with future scalability in mind for all Unified CVP sites.

In summary, plan ahead and follow all the design guidelines and recommendations presented in this guide and in the latest version of the *Cisco Unified Communications Solution Reference Network Design (SRND) Based on Cisco Unified Communications Manager*, available at:

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_implementation_design_guides_list.html

For assistance in planning and designing your Unified CVP solution, consult your Cisco or certified Partner Systems Engineer (SE).

A Note About the Unified CVP Call Server Component

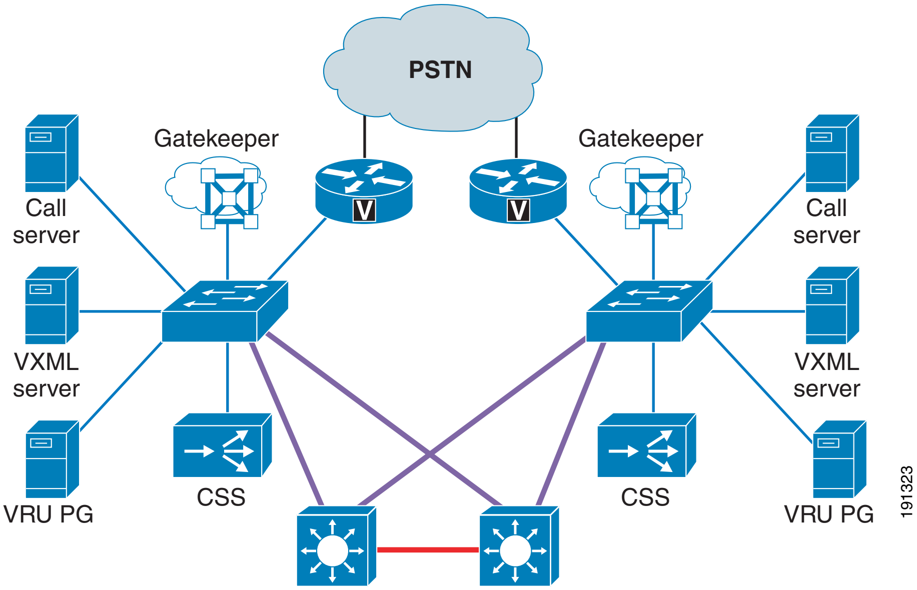
The other chapters of this document treat the Unified CVP Call Server as a single component because those chapters have no need to examine it in any more depth than that. When discussing Unified CVP high availability however, it is important to understand that there are actually several parts to this component:

- H.323 Service — Responsible for H.323 processing of incoming and outgoing calls as well as registering with the gatekeeper. The H.323 Service was known as the Unified CVP Voice Browser in previous versions of Unified CVP.
- SIP Service — Responsible for processing incoming and outgoing calls via SIP.
- ICM Service — Responsible for the interface to ICM. The ICM Service communicates with the VRU PG using GED-125 to provide ICM with IVR control. The ICM Service was part of the Application Server in previous releases of Unified CVP, but now it is a separate component.
- IVR Service — Responsible for the conversion of Unified CVP Microapplications to VoiceXML pages, and vice versa. The IVR Service was known as the Application Server in previous Unified CVP versions.

Layer 2 Switch

Figure 4-1 shows a high-level layout for a fault-tolerant Unified CVP system. Each component in the Unified CVP site is duplicated for redundancy. The quantity of each of these components varies based on the expected busy hour call attempts (BHCA) for a particular deployment. The following sections describe the failover strategy for each of these components.

Figure 4-1 Redundant Unified CVP System



In Figure 4-1, two switches provide the first level of network redundancy for the Unified CVP Servers:

- If one switch fails, only a subset of the components becomes inaccessible. The components connected to the remaining switch can still be accessed for call processing.
- If a Content Services Switch (CSS) is used, its redundant partner must reside on the same VLAN in order to send keep-alive messages to each other via Virtual Router Redundancy Protocol (VRRP), a protocol similar to Hot Standby Router Protocol (HSRP). If one of the switches fails, the other CSS is still functional.

For more information on data center network design, refer to the Data Center documentation available at

http://www.cisco.com/go/designzone

Note NIC teaming is not currently supported in the Unified CVP solution.

Note Cisco recommends that the NIC card and ethernet switch be set to 100 MB full duplex for 10/100 links, or set to auto-negotiate for gigabit links.

Originating Gateway

The function of the originating gateway in a Unified CVP solution is to accept calls from the PSTN and direct them to Unified CVP for call routing and IVR treatment.

This section covers the following topics:

- Configuration, page 4-4
- Call Disposition, page 4-14

Configuration

For the most current information on how to provide redundancy and reliability for originating gateways and T1/E1 lines, refer to the latest version of the *Cisco Unified Contact Center Enterprise Solution Reference Network Design (SRND)*, available at

http://www.cisco.com/en/US/products/sw/custcosw/ps1844/products_implementation_design_guides_list.html

In addition, consider the following issues when designing gateways for high availability in a Unified CVP solution:

- When used in ICM-integrated models, the originating gateway communicates with Unified CVP using H.323 or SIP. Unlike MGCP, SIP and H.323 do not have redundancy features built into the protocol. Instead, SIP and H.323 rely on the gateways and call processing components for redundancy.
- When configuring the gateway, it is best to bind the H.323 or SIP signaling to the virtual loopback interface, as illustrated in the following configuration examples:

  H.323:

```
interface Loopback0
 ip address 10.0.0.10 255.255.255.255
 h323-gateway voip interface
```

```
     h323-gateway voip id sj-gk ipaddr 10.0.1.100 1719 <<- GK IP
     h323-gateway voip h323-id sj-gw1
     h323-gateway voip bind srcaddr 10.0.0.10
```

SIP:

```
    voice service voip
     sip
      bind control source-interface Loopback0
      bind media source-interface Loopback0
```

This configuration allows call signaling to operate independent of the physical interfaces. In this way, if one interface fails, the other interface can handle the traffic. Each gateway interface should be connected to a different physical switch to provide redundancy in the event that one switch or interface fails. Each interface on the gateway is configured with an IP address on a different subnet. The IP Router(s) for the network is then configured with redundant routes to the Loopback address through the use of static routes or a routing protocol. If a routing protocol is used, pay careful attention to the number of routes being exchanged with the gateway, and consider using filters to limit the routing updates so that the gateway is only advertising the loopback address and not receiving routes.

Call Disposition

If the originating gateway fails, the following conditions apply to call disposition:

- Calls in progress are dropped. There is nothing that can be done to preserve these calls because the PSTN switch has lost the D-channel to all T1/E1 trunks on this gateway.
- New calls are directed by the PSTN carrier to a T1/E1 at an alternate gateway, provided the PSTN switch has its trunks and dial plan configured to do so.

SIP Proxy

A SIP Proxy server plays a similar role to the gatekeeper in a Unified CVP solution. The SIP Proxy server provides dial plan resolution on behalf of SIP endpoints, allowing dial plan information to be configured in a central place instead of statically on each SIP device. A SIP Proxy server is not required in a Unified CVP solution, but it is used in most solutions because of the benefits of centralized configuration and maintenance. Multiple SIP Proxy servers can be deployed in the network to provide load balancing, redundancy, and regional SIP call routing services. In a Unified CVP solution, the choices for SIP call routing are:

- SIP Proxy Server
  - Advantages:

    Weighted load balancing and redundancy.

    Centralized dial-plan configuration.

    SIP Proxy may already exist or be used for other applications for dial-plan resolution or intercluster call routing.
  - Disadvantages:

    Additional server and/or hardware required for SIP Proxy if not already deployed.
- Static routes using DNS SRV records on a DNS Server
  - Advantages:

    Weighted load balancing and redundancy.

  - Disadvantages:

    Might not be able to use an existing DNS server, depending on the location of the DNS server.

    The ability to share or delegate DNS server administration rights might not be possible in some organizations.

    Dial-plan configuration needs to be configured on each device individually (Cisco Unified Communications Manager, Unified CVP, and gateways).

    DNS SRV lookup is performed for each and every call by Unified CVP. If the DNS server is slow to respond, is unavailable, is across the WAN, or so forth, this will affect performance.

- Static routes using local DNS SRV records
  - Advantages:

    Weighted load balancing and redundancy.

    Does not depend on an external DNS Server, thus eliminating a point of failure, latency, and DNS Server performance concerns.

  - Disadvantages:

    Dial plan must be configured on each device individually (Cisco Unified Communications Manager, Unified CVP, and gateways).

Note The options for static routes using DNS SRV records on a DNS Server and static routes using local SRV records can introduce some unexpected long delays during failover and load balancing with UDP transport on the Unified CVP Call Server when the primary destination is shut down or off the network. With UDP, the per-hop delay is 3.5 seconds (if retry count is 2) or 7.5 seconds (if retry count is 3). This delay is on every call or every other call (on average) during failure, depending on load balancing.

- Static routes using IP addresses
  - Advantages:

    Does not depend on any other device (DNS or Proxy) to deliver calls to the destination.

  - Disadvantages:

    No redundancy possible for SIP calls from Unified CVP.

    Dial plan must be configured on each device individually.

  This option makes sense only for environments that do not have redundancy (single server) or for lab deployments.

Each device in the Unified CVP solution can use the above methods for determining where to send a call. The Unified CVP Call Server interface to the SIP network is through the Unified CVP SIP Service, which is discussed in the section on Unified CVP SIP Service, page 4-9.

Due to long delays when DNS is used with the Cisco Unified Presence proxy server, Cisco recommends that you disable the DNS server on the Cisco Unified Presence proxy server.

Configuration

The following sections discuss configuration of the SIP Proxy Server and Cisco IOS Gateways using SIP. It is not meant to be an exhaustive list of configuration options but only highlights certain configuration concepts.

SIP Proxy Server Configuration

The SIP Proxy Server should be configured with static routes that point at the appropriate devices (Call Servers, VoiceXML gateways, Cisco Unified Communications Manager cluster, and so forth). The SIP Proxy Server configuration allows you to specify the priority of the routes. In the case where there are multiple routes to the same destination, you can configure the SIP Proxy to load-balance across the destinations with equal priority or to send the calls in a prioritized manner using different priorities.

The Cisco Unified Presence Server SIP Proxy cannot use DNS SRV for outbound calls; it must be configured with multiple static routes in order to do load balancing and failover. (The Cisco Unified Presence Server does support the DNS SRV feature, but it has not been tested in Unified CVP deployments.) The static routes can point to an IP address or a regular DNS A host record.

To reduce the impact of a Proxy Server failure, Cisco recommends that you disable the RecordRoute header from being populated by the SIP Proxy Server. In this way, the inbound calls route through a SIP Proxy; but once they reach the Unified CVP Call Server, the signaling is exchanged directly between the originating device and the Call Server, and a SIP Proxy failure will not affect the calls in progress.

Cisco also highly recommends using UDP instead of TCP for SIP signaling. TCP stack timeout delays can cause significant delays to the caller during failures.

Cisco IOS Gateway Configuration

With Cisco IOS gateways, dial-peers are used to match phone numbers, and the destination can be a SIP Proxy Server, DNS SRV, or IP address. The following example shows a Cisco IOS gateway configuration to send calls to a SIP Proxy Server using the SIP Proxy's IP address.

```
sip-ua
 sip-server ipv4:10.4.1.100:5060

dial-peer voice 1000 voip
 session target sip-server
 ...
```

The **sip-server** command on the dial-peer tells the Cisco IOS gateway to use the globally defined sip-server that is configured under the **sip-ua** settings. In order to configure multiple SIP Proxies for redundancy, you can change the IP address to a DNS SRV record, as shown in the following example. The DNS SRV record allows a single DNS name to be mapped to multiple servers.

```
sip-ua
 sip-server dns:cvp.cisco.com

dial-peer voice 1000 voip
 session target sip-server
 ...
```

Alternatively, you can configure multiple dial-peers to point directly at multiple SIP Proxy servers, as shown in the following example. This configuration allows you to specify IP addresses instead of relying on DNS.

```
dial-peer voice 1000 voip
 session target ipv4:10.4.1.100
 preference 1
 ...
dial-peer voice 1000 voip
 session target ipv4:10.4.1.101
 preference 1
 ...
```

In the preceding examples, the calls are sent to the SIP Proxy server for dial plan resolution and call routing. If there are multiple Unified CVP Call Servers, the SIP Proxy server would be configured with multiple routes for load balancing and redundancy. It is possible for Cisco IOS gateways to provide load balancing and redundancy without a SIP Proxy Server. The following example shows a Cisco IOS gateway configuration with multiple dial-peers so that the calls are load-balanced across three Unified CVP Call Servers.

```
dial-peer voice 1001 voip
 session target ipv4:10.4.33.131
 preference 1
 ...
dial-peer voice 1002 voip
 session target ipv4:10.4.33.132
 preference 1
 ...
dial-peer voice 1003 voip
 session target ipv4:10.4.33.133
 preference 1
 ...
```

DNS SRV records allow an administrator to configure redundancy and load balancing with finer granularity than with DNS round-robin redundancy and load balancing. A DNS SRV record allows you to define which hosts should be used for a particular service (the service in this case is SIP), and it allows you to define the load-balancing characteristics among those hosts. In the following example, the redundancy provided by the three dial-peers configured above is replaced with a single dial-peer using a DNS SRV record. Note that a DNS server is required in order to do the DNS lookups.

```
ip name-server 10.4.33.200

dial-peer voice 1000 voip
 session target dns:cvp.cisco.com
```

With Cisco IOS gateways, it is possible to define DNS SRV records statically, similar to static host records. This capability allows you to simplify dial-peer configuration while also providing DNS SRV load balancing and redundancy. The downside of this method is that, if the SRV record needs to be changed, it must be changed on each gateway instead of on a centralized DNS server. The following example shows the configuration of static SRV records for SIP services handled by cvp.cisco.com, and the SIP SRV records for cvp.cisco.com are configured to load-balance across three servers.

```
ip host cvp4cc2.cisco.com 10.4.33.132
ip host cvp4cc3.cisco.com 10.4.33.133
ip host cvp4cc1.cisco.com 10.4.33.131
```

(SRV records for SIP/TCP)

```
ip host _sip._tcp.cvp.cisco.com srv 1 50 5060 cvp4cc3.cisco.com
ip host _sip._tcp.cvp.cisco.com srv 1 50 5060 cvp4cc2.cisco.com
ip host _sip._tcp.cvp.cisco.com srv 1 50 5060 cvp4cc1.cisco.com
```

(SRV records for SIP/UDP)

```
ip host _sip._udp.cvp.cisco.com srv 1 50 5060 cvp4cc3.cisco.com
ip host _sip._udp.cvp.cisco.com srv 1 50 5060 cvp4cc2.cisco.com
ip host _sip._udp.cvp.cisco.com srv 1 50 5060 cvp4cc1.cisco.com
```

Cisco highly recommends the use of UDP instead of TCP for SIP signaling. TCP stack timeout delays can cause significant delays to the caller during failures.

UDP Transport as a SIP Best Practice

Note that TCP transport is the default selection on the Unified Communications SIP trunk. Setting the transport to UDP for the trunk that interconnects with Unified CVP can resolve timing issues seen with TCP.

Cisco highly recommends the use of UDP instead of TCP for SIP signaling. TCP stack timeout delays can cause significant delays to the caller during failures.

Call Disposition

Calls are handled as indicated for the following failure scenarios:

- Primary SIP Proxy Server fails

  Active calls are preserved. Subsequent transfers of calls are successful, provided the backup SIP Proxy is available and the RecordRoute header is not being populated by the SIP Proxy. If the RecordRoute header is populated, signaling to the gateway will not be possible and subsequent transfer attempts will fail.

- All SIP Proxy Servers fail or are unreachable

  New calls arriving at the gateway are default-routed if survivability is configured on the gateway.

Unified CVP SIP Service

The Unified CVP SIP Service is the service on the Unified CVP Call Server that handles all incoming and outgoing SIP messaging and SIP routing. The Unified CVP SIP Service can be configured to use a SIP Proxy server for outbound dial plan resolution, or it can be configured to use static routes based on IP address or DNS SRV. Unified CVP Call Servers do not share configuration information about static routes; therefore, if a change needs to be made to a static route, then the change must be made on each Call Server's SIP Service. Cisco recommends that you use a SIP Proxy Server to minimize configuration overhead.

Configuration

If only a single SIP Proxy server is needed for outbound call routing from the Unified CVP Call Server, choose the SIP Proxy configuration when configuring the SIP Service. In the Operations Console, configure the following:

- Add a SIP Proxy Server and specify the IP address of the server.

Under the Call Server SIP Service settings, configure the following:

- Enable Outbound Proxy = True
- Use DNS SRV type query = False
- Outbound Proxy Host = SIP Proxy Server configured above

When using multiple SIP Proxy servers for outbound redundancy from the Call Server, configure the SIP Proxy with a DNS name and configure DNS SRV records in order to reach the SIP Proxy Servers. The DNS SRV records can exist on an external DNS Server, or they can be configured in a local DNS SRV record on each CVP server. In the OAMP Console, configure the following:

- Add a SIP Proxy Server and specify DNS name of the server.

Under the SIP Service configuration, configure the following:

- Enable Outbound Proxy = True
- Use DNS SRV type query = True
- The DNS SRV record should then be configured with the list of SIP Proxy Servers.

Cisco also highly recommends using UDP instead of TCP for SIP signaling. TCP stack timeout delays can cause significant delays to the caller during failures.

To configure the Local DNS SRV record on each server, perform the following steps:

1. Create a file named **srv.xml** in the **c:\cisco\cvp\conf** directory, with the following format:

```
<locater>
      <host name="vxml.cisco.com">
            <record weight="1" priority="1" destination="10.4.33.1" port="5060"/>
            <record weight="1" priority="1" destination="10.4.33.2" port="5060"/>
      </host>
      <host name="cm.cisco.com">
            <record weight="1" priority="1" destination="10.4.33.101" port="5060"/>
            <record weight="1" priority="2" destination="10.4.33.100" port="5060"/>
      </host>
</locater>
```

2. Under the SIP Service configuration, check **Resolve SRV records locally**.

Configuring High Availability for Calls in Progress

In the event that a Unified CVP Call Server fails with calls in progress, it is possible to salvage all calls if certain gateway configuration steps have been taken. A Call Server can fail in one of several ways:

- The server can crash.
- The process can crash.
- The process can hang.
- There can be a network outage.

The configuration discussed in this section protects against all of these situations. However, the following two situations cannot be protected against:

- Someone stops the process with calls in progress. This situation occurs when a system administrator forgets to put the Call Server out-of-service first to allow calls in progress to finish before stopping the process.
- The Call Server exceeds the recommended call rate. Although there is a throttle for the absolute number of calls allowed in the Call Server, there is no throttle for call rate. In general, exceeding 5 calls per second (cps) for an extended period of time can cause erratic and unpredictable call behavior on certain components of the CVP solution if one of the components is not sized correctly or if the call load is not balanced according to the weight and sizing of each call processing component.

For call survivability, configure the originating gateways as described in the latest version of the *Configuration and Administration Guide for Cisco Unified Customer Voice Portal (CVP)*, available at

http://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html

The survivability.tcl script itself also contains some directions and useful information.

In the event of most downstream failures (including a Unified CVP Call Server failure), the call is default-routed by the originating gateway. Note that survivability is not applicable in the Unified CVP Standalone and NIC-routing models because there is no Unified CVP H.323 or SIP Service involved anywhere in those models.

There is also a mechanism for detection of calls that have been cleared without Unified CVP's knowledge:

- Unified CVP checks every 2 minutes for inbound calls that have a duration older than a configured time (the default is 10 minutes).
- For those calls, Unified CVP sends an UPDATE message. If the message receives a rejection or is undeliverable, then the call is cleared and the license released.

The CVP SIP Service can also add the Session Expires header on calls so that endpoints such as the originating gateway may perform session refreshing on their own. RFC 4028 (*Session Timers in the Session Initiation Protocol*) has more details on the usage of Session Expires with SIP calls.

Call Disposition

Calls are handled as indicated for the following scenarios:

- Calls in progress

  If the Unified CVP SIP Service fails after the caller has been transferred (transfers include transfer to an IP phone, VoiceXML gateway, or other egress gateway), then the call continues normally until a subsequent transfer activity (if applicable) is required from the Unified CVP SIP Service. If the caller has not hung up and is awaiting further activity, there is a period of 9 to 18 seconds of silence before the caller is default-routed by survivability to an alternate location.

  If the call has not yet been transferred, the caller hears 9 to 18 seconds of silence before being default-routed by survivability to an alternate location. (Survivability does not apply in the Unified CVP Standalone and NIC-routing models.)

- New calls

  New calls are directed by the SIP Proxy to an alternate Unified CVP Call Server. If no Call Servers are available, the call is default-routed to an alternate location by survivability. (Survivability does not apply in the Unified CVP Standalone and NIC-routing models.)

Gatekeeper

An H.323 gatekeeper may be used as an optional component when using H.323 in any of the ICM-integrated deployment models except Model #4 (VRU Only with NIC Controlled Routing), which does not use Unified CVP for call control at all. Additionally, if SIP is used as the call control protocol, the gatekeeper is not required. An originating gateway can perform all of its H.323 call routing by using VoIP dial-peers that contain static IP addresses, whereas the Unified CVP H.323 Service must always perform a gatekeeper Remote Access Service (RAS) lookup to route calls.

Note In one particular situation, when using the VBAdmin SetTransferLabel option, the H.323 Service ignores the IP address returned from the gatekeeper and instead routes the IVR call leg back to the originating gateway from which the call arrived. This feature ensures that no WAN bandwidth is used during IVR treatment or queuing. A gatekeeper is still required in this situation because the H.323 Service has to perform the gatekeeper lookup function to obtain possible alternate endpoints in the event that the attempt to transfer the call to the originating gateway fails.

Unified CVP can use one of the following types of gatekeeper high-availability mechanisms:

- Gatekeeper Redundancy Using HSRP, page 4-12
- Gatekeeper Redundancy Using Alternate Gatekeeper, page 4-12

Only HSRP and alternate gatekeeper are supported by Unified CVP. Alternate gatekeeper support was introduced in Unified CVP 3.1 SR1.

Gatekeeper Redundancy Using HSRP

HSRP is a Cisco proprietary router redundancy protocol that allows two or more gatekeepers to share the same IP address in an active/standby fashion. Using HSRP, two gatekeepers work together to present the appearance of a single virtual IP address on the LAN.

The gatekeepers share the same IP and MAC addresses. Therefore, if one of the gatekeepers fails, the hosts on the LAN are able to continue forwarding packets to a consistent IP and MAC address. The process of transferring the routing responsibilities from one device to another is transparent to the user. The H.323 endpoints (such as the Unified CVP H.323 Service, Cisco Unified Communications Manager, and gateways) register to a virtual IP address that represents the HSRP gatekeeper pair.

If one gatekeeper fails, its partner assumes primary control. The major disadvantage of HSRP is that both gatekeepers in the HSRP failover pair must reside on the same IP subnet or VLAN, therefore they generally cannot be separated geographically. Gatekeepers using HSRP for redundancy also do not share any state information. Therefore, when a failover occurs, all of the devices must re-register with the gatekeeper from scratch.

As of Unified CVP 3.1 SR1, HSRP is no longer recommended. Instead gatekeeper clustering and alternate gatekeeper configuration on Unified CVP is the preferred method of gatekeeper redundancy.

Gatekeeper Redundancy Using Alternate Gatekeeper

The Unified CVP H.323 Service can be configured with a list of alternate gatekeepers (as many as needed; there is no limit). When the H.323 Service starts, it attempts to register to the first gatekeeper in the list. If the registration is not successful, it tries the remaining gatekeepers sequentially in the list until a successful registration occurs.

The H.323 Service stays registered to that gatekeeper until either of the following events occurs:

- That gatekeeper has some type of failure. The H.323 Service recognizes a gatekeeper failure in the following ways:
    - The periodic RAS Registration Request (RRQ) to the gatekeeper times out or is rejected.
    - An Admission Request (ARQ) on a transfer times out.
    - The gatekeeper pro-actively tells the H.323 Service to unregister, such as when the administrator does a shutdown on the gatekeeper configuration.

- The user does another setGK from VBAdmin. This causes the H.323 Service to register with the first gatekeeper in the list, if that gatekeeper is available; otherwise, it once again does a sequential attempt.

Gatekeeper clustering is not required in order to use Unified CVP alternate gatekeeper. It is possible to have two gatekeepers identically configured and also configure Unified CVP with alternate gatekeepers to provide redundancy.

The Unified CVP H.323 Service does not support gatekeeper clustering messages, but there is no reason that the gatekeepers cannot be part of a GUP cluster. In this way, other H.323 endpoints that natively support clustering (such as Cisco Unified Communications Manager and Cisco IOS gateways) can take advantage of the benefits of gatekeeper clustering. Unified CVP simply ignores clustering messages, such as when one of the gatekeepers in the cluster becomes overloaded or when Unified CVP registers with the gatekeeper.

Because Unified CVP does not automatically learn the other members of the gatekeeper cluster when it registers to the gatekeeper, it is necessary to define the gatekeeper cluster members statically in Unified CVP. Unified CVP uses one or more of the gatekeepers in the cluster as the alternate gatekeepers in its list and detects failure according to the rules mentioned earlier in this section.

Configuration

This section covers the following topics:

- HSRP Configuration, page 4-13
- Alternate Gatekeeper, page 4-14

HSRP Configuration

On the primary gatekeeper, enter these commands:

```
interface ethernet 0
 ip address 10.0.1.98 255.255.255.0
 ! Unique IP address for this GK
 standby 1 ip 10.0.1.100
 ! Member of standby group 1, sharing virtual address 10.0.1.100
 standby 1 preempt
 ! Claim active role when it has higher priority.
 standby 1 priority 110
 ! Priority is 110.
```

On the backup gatekeeper, enter these commands:

```
interface ethernet 0
 ip address 10.0.1.99 255.255.255.0
 standby 1 ip 10.0.1.100
 standby 1 preempt
 standby 1 priority 100
```

On both gatekeepers, enter identical gatekeeper configurations. For example:

```
gatekeeper
 ! Enter gatekeeper configuration mode.
 zone local gk-sj cisco.com 10.0.1.100
 ! Define local zone using HSRP virtual address as gatekeeper RAS address.
```

For more information, refer to the latest version of the *Configuration and Administration Guide for Cisco Unified Customer Voice Portal (CVP)*, available at

http://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html

Alternate Gatekeeper

Configure alternate gatekeepers using Unified CVP VBAdmin, as shown in the following examples:

```
set GK "10.0.1.100, 10.0.2.100, 10.0.3.100"
```

This example sets up three gatekeepers to which the H.323 Service could possible register. In each case, the H.323 Service registers to the first local zone that is configured in that gatekeeper. It also uses the default RAS port 1719.

```
setGK "10.0.1.100:zone1:1718, 10.0.2.100"
```

This example causes the H.323 Service to attempt to register first to gatekeeper 10.0.1.100 on port 1718 with local zone zone1. If that gatekeeper fails, the H.323 Service subsequently attempts to register to 10.0.2.100 on port 1719, with the first local zone defined on that gatekeeper.

Call Disposition

The call dispositions presented in this section apply to both HSRP and alternate gatekeeper.

A gatekeeper can fail in any of the following ways:

- The primary gatekeeper fails
  - Some calls in progress may not be transferred during the period that the endpoints are re-registering to the backup gatekeeper. After the failed transfer, an error is returned to the ICM. If the ICM script is coded to return an error (an END node does this) *and* survivability is configured on the gateway, the call is default-routed.
  - New calls arriving at the incoming gateway and Unified CVP are serviced correctly, although it is possible that some of the calls might invoke survivability during the period that the endpoints are re-registering to the backup gatekeeper.
- All gatekeepers fail
  - The Unified CVP H.323 Service goes out of service.
  - Calls in progress are not transferred. After the failed transfer, an error is returned to the ICM. If the ICM script is coded to return an error (an END node does this) *and* survivability is configured on the gateway, the call is default-routed.
  - New calls arriving at the gateway are default-routed if survivability is configured on the gateway.
- The primary gatekeeper degrades but does not fail
  - There are two conditions that usually cause this behavior: low memory due to memory leaks or excessive debug levels causing CPU overload.
  - In this situation, call processing behavior is unpredictable due to the fact that there might be no clean failover to the backup gatekeeper. If survivability is configured on the gateway, calls are default-routed.

Unified CVP H.323 Service

When multiple Unified CVP Call Servers are used for redundancy and scalability purposes in Unified CVP, Cisco recommends using a gatekeeper for load balancing and failover services. The H.323 Service is the component of the Unified CVP Call Server that processes H.323 messages and registers with the gatekeeper.

While it is possible for the ingress PSTN gateways to send H.323 calls to the H.323 Service using dial-peers with the specific IP address of the Call Server, doing so results in delays to callers during a failure scenario. In this scenario, a dial-peer is statically configured on the ingress gateways to load-balance across Unified CVP Servers, or in a prioritized fashion so that the primary server is always used under normal conditions. When the H.323 Service is no longer reachable for whatever reason, the dial-peer will attempt to send the call to the failed server and wait for a timeout to occur before proceeding to the next dial-peer configured, and this process occurs for each new call.

When a gatekeeper is used instead, the gateway dial-peer simply points to the gatekeeper, and the gatekeeper is responsible for determining which Call Servers are active and it load balances across them. The gatekeeper's registration process enables it to know which servers are available and does not suffer from the same timeouts as dial-peers. Therefore, Cisco recommends using a gatekeeper instead of static Cisco IOS dial-peers for redundancy and load balancing.

Configuration

Unified CVP H.323 configuration for high availability is performed primarily on the ingress gateways, but it is also necessary to configure the H.323 Service to register to the gatekeeper.

Configuring High Availability for New Calls

The gatekeeper knows which Unified CVP Call Servers are in service or out of service. It is therefore important to let a gatekeeper route incoming calls to a Unified CVP Call Server. By default, Unified CVP H.323 Services register to the gatekeeper with a technology prefix (tech-prefix) of **2#**. The Unified CVP H.323 Service must register with a tech-prefix, and it is not possible to configure the H.323 Service without a tech-prefix.

A technology prefix is a way for the gatekeeper to categorize registering endpoints by functionality. In general, no additional configuration is needed on the gatekeeper for incoming calls. The H.323 Service registers to the gatekeeper with 2#, and the originating gateway prepends a 2# to the incoming Dialed Number Identification Service (DNIS) digits. The gatekeeper automatically knows how to match the gateway request to an available Call Server. On the gatekeeper, the command **show gatekeeper gw-type-prefix** displays the route plan that the gatekeeper uses to route calls.

On the originating gateways, define the dial-peer for the Unified CVP Call Servers as follows:

```
dial-peer voice 11111 voip
 session target ras
 tech-prefix 2#
```

The command **session target ras** instructs the gateway to send the call to its gatekeeper. The command **tech-prefix 2#** instructs the gateway to prepend 2# to the DNIS number when sending the call to the gatekeeper.

Configuring High Availability for Calls in Progress

In the event that a Unified CVP Call Server fails with calls in progress, it is possible to salvage all calls if certain gateway configuration steps have been taken. A Call Server can fail in one of the following ways:

- The server can crash.
- The process can crash.
- The process can hang.
- There can be a network outage.

The configuration discussed in this section protects against all of these situations. However, the following two situations cannot be protected against:

- Someone stops the process with calls in progress. This situation occurs when a system administrator forgets to put the Call Server out-of-service first to allow calls in progress to finish before stopping the process.
- The Call Server exceeds the recommended call rate. Although there is a throttle for the absolute number of calls allowed in the Call Server, there is no throttle for call rate. In general, exceeding 5 calls per second (cps) for an extended period of time causes the Call Server to have erratic and unpredictable behavior. This situation can be prevented by proper sizing of the Unified CVP system.

For call survivability, configure the originating gateways as described in the latest version of the *Configuration and Administration Guide for Cisco Unified Customer Voice Portal (CVP)*, available at

> http://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html

The survivability.tcl script itself also contains some directions and useful information.

In the event of most downstream failures (including a Unified CVP Call Server failure), the call is default-routed by the originating gateway. Note that survivability is not applicable in the Unified CVP Standalone and NIC-routing models because there is no Unified CVP Call Server involved anywhere in those models.

Additional Cisco IOS Gateway Configuration

The command in the following example disables the TCP timeout for H.225 signaling on the gateway:

```
voice service voip
 h323
  no h225 timeout keepalive
```

This action allows the gateway to lose connectivity with the Call Server or Cisco Unified Communications Manager but still retain active calls. If you do no t use this command, calls that are still active that are otherwise unaffected by the failure (that is, the RTP stream is still streaming between the endpoints) will be disconnected when the TCP session times out.

The following commands specify the RTP media timeout:

```
ip rtcp report interval 2000

gateway
 timer receive-rtcp 4
```

When the gateway detects that RTCP messages have not been received in the specified interval, the call is disconnected.

Call Disposition

If the Unified CVP H.323 Service fails, the following conditions apply:

- Calls in progress

  If the Unified CVP H.323 Service fails after the caller has been transferred (transfers include transfer to an IP phone, VoiceXML gateway, or other egress gateway), then the call continues normally until a subsequent transfer activity (if applicable) is required from the Unified CVP H.323 Service. If the caller has not hung up and is awaiting further activity, there is a period of 9 to 18 seconds of silence before the caller is default-routed by survivability to an alternate location.

  If the call has not yet been transferred, the caller hears 9 to 18 seconds of silence before being default-routed by survivability to an alternate location. (Survivability does not apply in the Unified CVP Standalone and NIC-routing models.)

- New calls

  New calls are directed by the gatekeeper to an alternate Unified CVP Call Server. If no Call Servers are available, the call is default-routed to an alternate location by survivability. (Survivability does not apply in Unified CVP Standalone and NIC-routing models.)

Unified CVP IVR Service

With Unified CVP 3.1 and earlier, the IVR Service (previously called the Application Server) was treated independently of the H.323 Service (previously called the Voice Browser) and VoiceXML gateways. High availability was achieved by configuring the Unified CVP Voice Browser and VoiceXML gateways with a list of application server IP addresses and/or using the Content Services Switch (CSS). With Unified CVP 4.0 and later releases, the IVR Service is tightly coupled with the SIP Service or H.323 Service. If the IVR Service goes out of service, the H.323 or SIP Service will be taken out of service as well so that no further calls are accepted by the Call Server.

Configuration

No additional configuration is needed in order to tell the H.323 or SIP Service which IVR Service to use. By default, the H.323 and SIP Service use the IVR Service that resides on the same server. It is also no longer necessary to configure the VoiceXML gateway with the IP address of the Call Server's IVR Service. When SIP is used, the SIP Service inserts the URL of the Call Server's IVR Service into a header in the SIP INVITE message when the call is sent to the VoiceXML gateway. The VoiceXML gateway extracts this information from the SIP INVITE and uses it when determining which Call Server to use. When H.323 is used, the VoiceXML gateway examines the source IP address of the incoming call from the Call Server. This IP address is then used as the address for the Call Server's IVR Service.

The following example illustrates the VoiceXML bootstrap service that is invoked when a call is received:

```
service bootstrap flash:bootstrap.tcl
  paramspace english index 0
  paramspace english language en
  paramspace english location flash
  paramspace english prefix en
```

Unlike Unified CVP 3.1 and earlier releases, with Unified CVP 4.0 and later releases you do not have to configure the IP address of the Call Server. The bootstrap.tcl learns the IP address of the source Call Server and uses it as its call server. There is no need for a CSS or backup Call Server configuration because receiving a call from the Call Server means that the server is up and operational.

The following files in flash memory on the gateway are also involved with high availability: handoff.tcl, survivability.tcl, recovery.vxml, and several .wav files. Use Trivial File Transfer Protocol (TFTP) to load the proper files into flash. Configuration information for each file can be found within the file itself. For more information, refer to the latest version of the *Configuration and Administration Guide for Cisco Unified Customer Voice Portal (CVP)*, available at

> http://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html

Call Disposition

If the Unified CVP IVR Service fails, the following conditions apply to the call disposition:

- Calls in progress are default-routed to an alternate location by survivability on the originating gateway. (Survivability does not apply in the Unified CVP Standalone and NIC-routing models.)
- New calls are directed to an in-service Unified CVP IVR Service.

VoiceXML Gateway

The VoiceXML gateway parses and renders VoiceXML documents obtained from one or several sources: the Unified CVP Call Server (IVR Service), the Unified CVP VXML Servers, or some other external VoiceXML source. Rendering a VoiceXML document consists of retrieving and playing prerecorded audio files, collecting and processing user input, and/or connecting to an ASR/TTS server for voice recognition and dynamic text-to-speech conversion.

The VoiceXML gateway does not support G.729 encoded prompts and can support only G.711 encoded prompts. An IP originated call from a G.729 region would require a transcoder during the IVR treatment at the VoiceXML gateway. For details about the transcoder and sizing, refer to the *Cisco Unified Communications SRND Based on Cisco Unified Communications Manager*, available at

> http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_implementation_design_guides_list.html

Configuration

High availability configuration for VoiceXML gateways is controlled by the gatekeeper for H.323, the SIP proxy for SIP, and/or the Unified CVP Call Server. Whether the VoiceXML gateways are distributed or centralized also influences how high availability is achieved.

In the event that a Unified CVP Call Server is unable to connect to a VoiceXML gateway, an error is returned to the ICM script. In the ICM script, separate the Send to VRU node from the first Run External script node in order to catch the VoiceXML gateway connection error. If an END script node is used off the X-path of the Send to VRU node, the call is default-routed by survivability on the originating gateway. (Survivability does not apply in the Unified CVP Standalone and VRU-only models.) A Queue to Skill group node could also be used, but that method is effective only if there is an agent available.

Otherwise, ICM tries to queue the caller, and that attempt fails because the Unified CVP Call Server is once again unable to connect to a VoiceXML gateway. An END node could then also be used off the X-path of the Queue to Skill Group node to default-route the call.

Centralized VoiceXML Gateways

In this configuration, the VoiceXML gateways reside in the same data center as the Unified CVP Call Server.

H.323 VoiceXML Gateways

On the gatekeeper, configure a zone prefix list that contains the H.323 IDs of all VoiceXML gateways at the data center. For example, assume that there are three VoiceXML gateways in the data center with H.323 IDs of VoiceXMLgw1, VoiceXMLgw2, and VoiceXMLgw3, and that the ICM label for the Network VRU is 5551000. In this example, the gatekeeper distributes calls in essentially a round-robin scheme among all three VoiceXML gateways, provided they are all in service:

```
zone prefix gkzone-name 5551000* gw-priority 10 VoiceXMLgw1 VoiceXMLgw2 VoiceXMLgw3
```

SIP VoiceXML Gateways

If you are using Cisco Unified Presence Server: On the SIP proxy server, configure a static route for the Network VRU label for each gateway. If the VRU label is 5551000, the static route pattern would be 5551000* in order to allow for the correlation-id to be appended and routed to the VoiceXML gateway.

If you are using SIP static routes on the Unified CVP Call Server: Under the SIP Service configuration for the Call Server, configure a static route for each Network VRU label and gateway. If the VRU label is 5551000, the static route pattern would be 5551000>. The > is a wildcard representing one or more digits, and it is needed so that the correlation-id appended to the DNIS number can be passed to the VoiceXML gateway correctly.

In the case of both SIP proxy or Unified CVP static routes, the next-hop address of the route can be either the IP address of the gateway or a DNS SRV record. If you are using an IP address, you must create multiple static routes, one for each VoiceXML gateway. In the case of DNS SRV, only one route per Network VRU label is needed, and the SRV record will provide for load-balancing and redundancy.

Distributed VoiceXML Gateways (Co-Resident Ingress Gateway and VoiceXML)

In this configuration, the gateway that processes the incoming call from the PSTN is separated from the Unified CVP servers by a low-bandwidth connection such as a WAN, and the VoiceXML gateway that is used is the same as the ingress gateway. The purpose of this configuration is to keep the media stream at the edge to avoid consuming bandwidth on the WAN.

H.323 VoiceXML Gateways

Use the SetTransferLabel in VBAdmin (not gatekeeper zone prefixes) to control the selection of the VoiceXML gateway. The SetTransferLabel command is specified per Network VRU label. When the Unified CVP Call Server receives a label from ICM that matches what is configured in the SetTransferLabel, the Unified CVP Call Server performs a gatekeeper lookup but ignores the destination gateway returned by the gatekeeper and sends the call back to the gateway that originated the call. The H.323 Service determines the originating gateway by looking at the source IP address of the H.323 signaling.

SIP VoiceXML Gateways

With SIP, the equivalent of the SetTransferLabel command is the Send to Originator configuration under the SIP Service. If the Network VRU label is 5551000, the Send to Originator pattern would be 5551000>. The > is a wildcard pattern representing one or more digits. The SIP Service determines the originating gateway by looking at the Remote-Party-ID header in the SIP INVITE message.

Distributed VoiceXML Gateways (Separate Ingress Gateway and VoiceXML)

In this configuration, the gateway that processes the incoming call from the PSTN is separated from the Unified CVP servers by a low-bandwidth connection such as a WAN, and the VoiceXML gateway that is used is different than the ingress gateway but located at the same site as the ingress gateway. The purpose of this configuration is to keep the media stream at the same site and not consume bandwidth on the WAN and to optimize VoiceXML gateway sizing when it is appropriate to separate ingress and VoiceXML gateways. In this case, setTransferLabel and Send to Originator cannot be used because you would not want the IVR leg of the call to go back to the ingress gateway. Additionally, it is also impractical to use a gatekeeper or SIP Proxy to control the call routing because you would have to configure separate Network VRUs, Network VRU labels, and customers in ICM for each remote site. Instead, use SetSigDigits functionality.

With this method, the Unified CVP Call Server strips the leading significant digit(s) from the incoming DNIS number. The value that is stripped is saved and prepended when subsequent transfers for the call occur.

H.323 VoiceXML Gateways

When H.323 is used, the significant digit is prepended with a # sign so that the gatekeeper treats it as a technology prefix. The VoiceXML gateway at the remote site should register to the gatekeeper with the same technology prefix as the leading significant digit(s) that were stripped from the DNIS number. The gatekeeper then routes the IVR leg of the call to the correct VoiceXML gateway. If you are using Cisco Unified Communications Manager (Unified CM), remember that Unified CVP indiscriminately prepends the sigdigits value to all transfers, including those to Unified CM. Therefore, when using Unified CM in this scenario, it is necessary to define a gatekeeper-controlled trunk for each of the VoiceXML gateway tech-prefixes and to add zone prefix configuration to the gatekeeper for the Unified CM agents, as illustrated in the following example.

Configuration of ingress gateway:

```
dial-peer voice 1000 voip
 tech-prefix 2#  (gets the call to CVP)
 translate-outgoing called 99
```

Apply a translation-rule to the incoming DNIS number to prepend the value 3:

```
translation-rule 99
 Rule 1 8002324444 38002324444
```

Assuming the DNIS number is 8002324444, the final DNIS string routed to Unified CVP is 2#38002324444.

Configuration in VB Admin:

```
setTechPrefix 2#
setSigDigits 1
```

Strip one digit from the DNIS number after stripping the 2# technology prefix.

Configuration of VoiceXML gateway:

Register to the gatekeeper with tech-prefix 3#:

```
h323-gateway voip tech-prefix 3#
```

Cisco Unified CM configuration (if used):

Create a separate gatekeeper-controlled trunk corresponding to each of the tech-prefixes used by the VXML gateways.

Gatekeeper configuration:

Define zone prefixes to route calls appropriately to Unified CM agents (only if using Cisco Unified CM).

Summary of call routing:

1. A call arrives at Unified CVP with a DNIS string of 2#38002324444.
2. Unified CVP first strips the tech-prefix (2#), leaving 38002324444.
3. Unified CVP then strips one digit (3) from the beginning of the DNIS string, leaving 8002324444.
4. 8002324444 is passed to ICM for call routing.
5. When it is time to transfer, assume ICM returns the label 5551000102. Unified CVP prepends 3#, giving 3#5551000102. This value is then passed to the gatekeeper for address resolution.
6. The gatekeeper resolves this label to the VoiceXML gateway that registered with tech-prefix 3#.
7. The VoiceXML gateway strips the 3#, leaving 5551000102 for the destination address.

SIP VoiceXML Gateways

When SIP is used, the significant digits are prepended to the DNIS number, and a SIP Proxy can be configured to route based on those prepended digits. The static routes in the SIP Proxy for the VoiceXML gateway should have the digits prepended. Because these prepended digits were originally populated by the ingress gateway, the SIP Proxy can use them to determine which VoiceXML gateway to use based on the incoming gateway. In this way, calls arriving at a particular site can always be sent back to that site for VoiceXML treatment, with the result that no WAN bandwidth is used to carry the voice RTP stream. Keep in mind that Unified CVP indiscriminately prepends the sigdigits value to all transfers, including those to Unified CM. Therefore, when using Unified CM in this scenario, it is necessary to strip the prepended digits when the call arrives so that the real DNIS number of the phone can be used by Unified CM to route the call, as illustrated in the following example.

Configuration of ingress gateway:

Apply a translation-rule to the incoming DNIS to prepend the value 3333:

```
translation-rule 99
 rule 1 8002324444 33338002324444

dial-peer voice 1000 voip
 translate-outgoing called 99
```

Assuming the DNIS number is 8002324444, the final DNIS string routed to Unified CVP is 33338002324444.

Configuration of Unified CVP SIP Service:

The Unified CVP SIP Service does not currently have a configuration field for setting the significant digits that should be stripped. Instead, you must edit the sip.properties file. The sip.properties file is located in the C:\Cisco\CVP\conf directory by default. Add the following line to the end of the sip.properties file (to strip four digits from the DNIS number):

```
SIP.SigDigits = 4
```

Configuration of VoiceXML gateway:

Configure the VXML gateway to match the DNIS string, including the prepended digits:

```
dial-peer voice 3000 voip
 incoming-called number 33335551000T
 service bootstrap
 ...
```

Configure the Unified CVP bootstrap.tcl application with the sigdigits parameter, telling it how many digits to strip off of the incoming DNIS string:

```
application
 service bootstrap flash:bootstrap.tcl
 param sigdigits 4
 ...
```

Cisco Unified CM configuration (if used):

Configure Unified CM to strip the prepended digits, either by using the Significant Digits configuration on the SIP Trunk configuration page or by using translation patterns.

SIP Proxy configuration:

Define static routes on the SIP Proxy, with the prepended digit present, to be sent to the appropriate VoiceXML gateway. Because transfers to agents on a Unified CM cluster will also have the digits prepended, the static routes for agent phones must also contain the prepended digits.

Summary of call routing:

1. A call arrives at Unified CVP with a DNIS number of 33338002324444.
2. Unified CVP then strips four digits (3333) from the beginning of the DNIS string, leaving 8002324444.
3. 8002324444 is passed to ICM for call routing.
4. When it is time to transfer, assume ICM returns the label 5551000102. Unified CVP prepends 3333, giving 33335551000102.
5. The SIP Service then resolves the address using the SIP Proxy or local static routes, and it sends the call to the VoiceXML gateway.
6. The VoiceXML gateway bootstrap.tcl will strip the 3333, leaving 5551000102 for the destination address.

H.323 Alternate Endpoints

In all cases for either centralized or distributed deployments, configure alternate endpoints for each of the VoiceXML gateways in case the VoiceXML gateway rejects the incoming request (perhaps due to error or overload):

```
endpoint alt-ep h323id VoiceXMLgw1 ip-address-VoiceXMLgw2
endpoint alt-ep h323id VoiceXMLgw2 ip-address-VoiceXMLgw3
endpoint alt-ep h323id VoiceXMLgw3 ip-address-VoiceXMLgw1
```

Call Disposition

If the VoiceXML gateway fails, the following conditions apply to the call disposition:

- Calls in progress are default-routed to an alternate location by survivability on the ingress gateway. (Survivability does not apply in Unified CVP Standalone and NIC-routing models.)
- New calls find an alternate VoiceXML gateway.

Hardware Configuration for High Availability on the Voice Gateways

The individual hardware components have the following high-availability options:

- Redundant power supplies and on-hand spares
- Separate components for higher availability
- Dedicated components, which have fewer interaction issues

Example 1: Separate PSTN Gateway and VoiceXML Gateway

A PSTN gateway and a separate VoiceXML gateway provide greater availability than a combine PSTN and VoiceXML gateway.

Example2: Duplicate Components for Higher Availability

- Two 8-T1 PSTN gateways provide greater availability than one 16-T1 PSTN gateway.
- Two 96-port Unified CVP VXML Servers provide greater availability than one 192-port Unified CVP VXML Server.
- Larger designs can use N+1 spares for higher availability.

Example3: Geographic Redundancy for Higher Availability

Geographical redundancy and high availability can be achieved by purchasing duplicate hardware for Side A and Side B.

Content Services Switch (CSS)

The VoiceXML gateway is the only box in the Unified CVP system that makes requests to the CSS. When the VoiceXML gateway needs to make a request for media, ASR/TTS, or VoiceXML, it looks in its configuration to determine to where it should make the request. When a CSS is used, the IP address that is configured on the VoiceXML gateway is a virtual IP address that points to a service configured on the CSS. There are three types of services that the VoiceXML gateway client can request from the CSS:

- Media Server
- ASR/TTS
- Unified CVP VXML Server

If the primary CSS that is servicing these requests should fail, the client VoiceXML gateway must still be able to obtain media and VoiceXML from the servers.

Configuration

You can configure high availability for the CSS by using the Virtual IP (VIP) Redundancy method, as described in the latest version of the *Configuration and Administration Guide for Cisco Unified Customer Voice Portal (CVP)*, available at

> http://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html

Also refer to the latest version of the *CSS Redundancy Configuration Guide*, available at

> http://www.cisco.com/en/US/products/hw/contnetw/ps792/products_installation_and_configuration_guides_list.html

Essentially, a master/backup pair of CSSs functions very much like an HSRP gatekeeper pair. They must reside on the same VLAN and exchange heartbeats using Virtual Router Redundancy Protocol (VRRP), a protocol very similar to HSRP. If the primary CSS fails, the backup CSS recognizes the failure within three seconds and begins processing client requests to its configured virtual IP addresses. The configuration of the master and backup CSSs must always be kept in synchronization.

Call Disposition

If the master CSS fails, then the following conditions apply to the call disposition:

- Calls in progress encounter various behaviors, depending on the type of service the VoiceXML gateway client requested:
  - Media server requests are unaffected.

    The VoiceXML gateway has a very short-lived interaction with the CSS for audio files. Upon receiving a media server request from the gateway, the CSS simply provides an HTTP redirect IP address for the VoiceXML gateway. At that point, the gateway fetches the audio file directly from the media server, bypassing any further interaction with the CSS. Additionally, media file requests to the CSS are very infrequent because the VoiceXML gateway caches previously retrieved media files. Cisco recommends using the CSS to find the available VXML Server via heartbeat and to perform load balancing. Subsequent requests and responses between the VoiceXML gateway and the VXML Server should bypass the CSS.
  - Unified CVP IVR Service requests are unaffected.

    Only the initial VoiceXML document request to a Unified CVP IVR Service uses the CSS. The CSS first picks a Unified CVP IVR Service to service the request. The first document passes through the CSS on its return to the VoiceXML gateway. However, subsequent VoiceXML requests are made directly from the VoiceXML gateway client to the Unified CVP IVR Service. If the CSS fails during the very brief period that the first VoiceXML document is being returned, the VoiceXML gateway simply retries the request. If the backup (now primary) CSS selects the same Unified CVP IVR Service as the previous one, there is an error due to a duplicate call instance. In that case, the caller is default-routed by survivability on the originating gateway. (Survivability does not apply in the Unified CVP Standalone model.)

- ASR/TTS requests typically fail but might be recoverable.

  When the VoiceXML gateway first makes an ASR/TTS request to the CSS, a TCP connection is opened from the VoiceXML gateway to the Media Resource Control Protocol (MRCP) server. That TCP connection goes through the CSS and persists until the caller disconnects or is transferred to an agent. If the primary CSS fails, that TCP connection is terminated. The VoiceXML gateway returns an error code, which you could write a script to work around. The worst-case scenario is that the caller is default-routed to an alternate location by survivability on the originating gateway. (Survivability does not apply in the Unified CVP Standalone model.)

- Unified CVP VXML Server requests may fail.

  The VoiceXML gateway is "sticky" to a particular Unified CVP VXML Server for the duration of the VoiceXML session. It uses CSS cookies to provide that stickiness. If the CSS fails, the backup CSS has no knowledge of the cookie. Subsequent requests in the session might go to the correct Unified CVP VXML Server, but there is no guarantee. The VoiceXML gateway returns an error code, which you could write a script to work around. The worst-case scenario is that the caller is default-routed to an alternate location by survivability on the originating gateway. (Survivability does not apply in the Unified CVP Standalone model.) The Adaptive Session Redundancy (ASR) feature of CSS ensures that port licenses are not temporarily and needlessly unavailable on the VXML Server. The VXML Server is stateful, and the ASR feature minimizes VXML Server license port usage during a CSS failover.

- New calls are directed transparently to the VIPs on the backup CSS, and service is unaffected.

Media Server

Audio files can be stored locally in flash memory on the VoiceXML gateway or on an HTTP/TFTP file server. By definition, audio files stored locally are highly available. However, HTTP/TFTP file servers provide the advantage of centralized administration of audio files.

Configuration When Using Unified CVP Microapplications

The VoiceXML gateway sends HTTP requests to an HTTP media server to obtain audio files. It uses the following VoiceXML gateway configuration parameters to locate a server when not using a CSS:

```
ip host mediaserver <ip-address-of-primary-media-server>
ip host mediaserver-backup <ip-address-of-secondary-media-server>
```

The backup server is invoked only if the primary server is not accessible, and this is not a load-balancing mechanism. Once failover occurs, all calls continue to use the backup server until that server becomes inaccessible. Note that *mediaserver* is not a fixed name, and it needs to match whatever name was assigned to the media_server ECC variable in the ICM script.

The VoiceXML gateway also uses the following VoiceXML gateway configuration parameters to locate a server when using a CSS:

```
ip host mediaserver <ip-address-of-CSS-VIP-for-media-server>
iip host mediaserver-backup <ip-address-of-CSS-VIP-for-media-server>
```

Because the CSS almost always locates a media server on the first request, a backup server is rarely invoked. However, it is helpful to configure the backup server when using a CSS for deployments where there are multiple DataCenters with a CSS in each.

Call Disposition When Using Unified CVP Microapplications

If the media server fails, the following conditions apply to the call disposition:

- Calls in progress should recover automatically. The high-availability configuration techniques described above should make the failure transparent to the caller. If the media request does fail, use scripting techniques to work around the error (for example, retry the request, transfer to an agent or label, or use TTS).
- New calls are directed transparently to the backup media server, and service is not affected.
- If the media server is located across the WAN from the VoiceXML gateway and the WAN connection fails, the gateway continues to use prompts from gateway cache until the requested prompt becomes stale, at which time the gateway attempts to re-fetch the media and the call fails if survivability is not enabled. If survivability is enabled, the call are default-routed.

Configuration When Using Cisco Unified Call Studio Scripting

When scripting in Cisco Unified Call Studio, unlike with ICM scripting, there is no concept of "*-backup*" for media files. The best the script writer can do is to point **Properties->AudioSettings->Default Audio Path URI** in the application to a single media server or the CSS VIP address for a farm of media servers.

Unified CVP Video Media Server

To provide redundancy and load-balancing for the Video Media Server, use a CSS. For configuration details, refer to the CSS configuration and administration guide available at

http://www.cisco.com/en/US/products/hw/contnetw/ps792/products_installation_and_configuration_guides_list.html

Call Disposition

If the media server fails, the following conditions apply to the call disposition:

- Calls in progress that are streaming from the Video Media Server will stop receiving the media stream. Additionally, the Agent will lose the connection to the Video Media Server and will lose the ability to play media.
- New calls are directed to the backup media server, and service is not affected.

Unified CVP VXML Server

The VoiceXML gateway makes HTTP requests to the Unified CVP VXML Server to obtain VoiceXML documents.

Configuration

The Unified CVP VXML Server high-availability configuration and behavior differ between Standalone deployments and deployments that are integrated with ICM, as described in the following sections.

Standalone Self-Service Deployments

For instructions on configuring primary and backup Unified CVP VXML Servers, see the latest version of the *Cisco Unified CVP Configuration and Administration Guide*, available at http://www.cisco.com.

Specifically, it is the CVPPrimaryVXMLServer and CVPBackupVXMLServer gateway parameters that control the high availability characteristics of the Unified CVP VXML Server. If Unified CVP VXML Server load balancing and more robust failover capabilities are desired, a CSS may be used. (For configuration details, see the latest version of the *Cisco Unified CVP Configuration and Administration Guide*.) Load balancing can also be achieved without a CSS by varying the primary and backup Unified CVP VXML Server configurations across multiple gateways.

Deployments Using ICM

When a Unified CVP VXML Server is used in conjunction with ICM, the ICM script will pass a URL to the VoiceXML gateway in order to invoke the VoiceXML applications. You can configure the ICM script to first attempt to connect to Unified CVP VXML Server A, and if the application fails out the X-path of the Unified CVP VXML Server ICM script node, Unified CVP VXML Server B should be tried. The IP address in the URL can also represent Unified CVP VXML Server VIPs on the CSS.

Call Disposition

If the Unified CVP VXML Server fails, the following conditions apply to the call disposition:

- Calls in progress in a Standalone deployment are disconnected. Calls in progress in an ICM-integrated deployment can be recovered using scripting techniques to work around the error as shown in the script above (for example, retry the request, transfer to an agent or label, or force an error with an END script node to invoke survivability on the originating gateway).
- New calls are directed transparently to an alternate Unified CVP VXML Server. Note that, without a CSS, callers might experience a delay at the beginning of the call and have to wait for the system to timeout while trying to connect to the primary Unified CVP VXML Server.

Radvision IVP Server

The Radvision IVP Server communicates with many components of the solution and is central to the redundancy and load-balancing of many of the components of the Full Service Video solution.

The Radvision IVP B2BUA SIP Server receives SIP calls from calling endpoints such as 3G gateways or Cisco Unified Communications Manager servers. A SIP Proxy server should be used to provide redundancy and load-balancing for calls being delivered to multiple Radvision IVP servers.

The Radvision IVP communicates with the Unified CVP Call Servers using an XML interface. The Unified CVP Call Servers register with the Radvision IVP Servers upon initialization, and keepalive messages are exchanged. The Unified CVP Call Servers should be configured to register with all of the Radvision IVP Servers, and the Radvision IVP Servers should be configured to load-balance in a round-robin pattern across all available Unified CVP Call Servers.

The Radvision IVP also communicates with the MCU and EMP modules. The MCU and EMP modules register to the Radvision IVP upon initialization and exchange keepalive messages. Multiple MCU and EMP modules should be pooled together in order to provide redundancy. The Radvision IVP chooses from available MCU and EMP resources as those resources are needed.

Call Disposition

In the event of a Radvision IVP server failure, new calls are delivered by the SIP Proxy to the backup IVP server. Calls in progress are dropped.

In the event of a Unified CVP Call Server failure, registration to the Radvision IVP server is lost and new calls are no longer delivered to the failed Unified CVP Call Server. Calls in progress are dropped. However, if there is a backup Unified CVP Call Server present in the network, then the Radvision IVP server can be configured to use the backup Unified CVP Call Server. This configuration is done via the IVR tab on the backup Unified CVP Call Server.

In the event of an MCU or EMP module failure, the MCU or EMP module will first try to recover itself if it goes offline or if there is an alarm condition. During the recovery period or in a situation where the MCU or EMP is not able to recover itself, it un-registers from the IVP server and is no longer used by IVP for new calls. Calls in progress are dropped.

In the event that the Video Media Server (VMS) is down, no new calls can receive video from agent or video announcements.

Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) Server

The VoiceXML gateway sends MRCP requests to the ASR/TTS servers in order to perform voice recognition and text-to-speech instructions that are defined in a VoiceXML document.

Configuration

The ASR/TTS high-availability configuration and behavior differ between Standalone and ICM-integrated deployments, as described in the following sections.

Standalone Self-Service Deployments

A CSS is required in Standalone deployments to provide failover capabilities for ASR/TTS. For instructions on configuring the CSS for ASR/TTS and on configuring the ASR/TTS Server in a Standalone deployment, see the latest version of the *Cisco Unified CVP Configuration and Administration Guide*, available at http://www.cisco.com.

Deployments Using ICM

The VoiceXML gateway uses gateway configuration parameters to locate an ASR/TTS server both when using a CSS and when not using a CSS. Note that the backup server is invoked only if the primary server is not accessible and if this is not a load-balancing mechanism. Once failover occurs, all calls continue to use the backup server until that server becomes inaccessible.

The hostnames (such as **asr-en-us**) are fixed and cannot be modified. The only portion that may be modified is the locale. In the following example, there is a set of primary and backup English ASR/TTS servers and a set of Spanish servers. Configure the CSS, if used, according to the instructions in the latest version of the *Configuration and Administration Guide for Cisco Unified Customer Voice Portal (CVP)*, available at

http://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html

When a CSS is used, the IP addresses mentioned the following example would be the virtual IP address for the ASR/TTS service on the CSS.

```
ip host asr-en-us <ip-address-of-primary-English-ASR-server>
ip host asr-en-us-backup <ip-address-of-secondary-English-ASR-server>
ip host tts-en-us <ip-address-of-primary-English-TTS-server>
ip host tts-en-us-backup <ip-address-of-secondary-English-TTS-server>
ip host asr-es-es <ip-address-of-primary-Spanish-ASR-server>
ip host asr-es-es-backup <ip-address-of-secondary-Spanish-ASR-server>
ip host tts-es-es <ip-address-of-primary-Spanish-TTS-server>
ip host tts-es-es-backup <ip-address-of-secondary-Spanish-TTS-server>
```

Call Disposition

If the ASR/TTS MRCP server fails, the following conditions apply to the call disposition:

- Calls in progress in Standalone deployments are disconnected. Calls in progress in ICM-integrated deployments can be recovered using scripting techniques to work around the error (for example, retry the request, transfer to an agent or label, switch to prerecorded prompts and DTMF-only input for the remainder of the call, or label or force an error with an END script node to invoke survivability on the originating gateway).
- New calls in Standalone or ICM-integrated deployments are directed transparently to an alternate ASR/TTS server if a CSS is being used. New calls in ICM-integrated deployments are directed transparently to an alternate ASR/TTS server if "-backup" gateway hostnames have been used.

Cisco Unified Communications Manager

Unified CVP transfers callers to Cisco Unified Contact Center Enterprise (Unified CCE) agent phones or desktops using H.323 or SIP. The Unified CVP Call Server receives an agent label from the ICM and routes the call using a gatekeeper or SIP proxy. The call is then sent to the appropriate Cisco Unified Communications Manager (Unified CM) in the cluster, which connects the caller to the agent. The Unified CVP Call Server proxies the call signaling, so it remains in the call signaling path after the transfer is completed. However, the RTP stream flows directly from the originating gateway to the phone. This fact becomes very significant in discussions of high availability.

Configuration

For the most current information on providing Unified CM for high availability, refer to the latest version of the *Cisco Unified Contact Center Enterprise Solution Reference Network Design (SRND)*, available at

http://www.cisco.com/en/US/products/sw/custcosw/ps1844/products_implementation_design_guides_list.html

Call Disposition

If the Unified CM process fails on the server that is either hosting the call or hosting the phone, the following conditions apply to the call disposition:

- Calls in progress are preserved. Skinny Client Control Protocol (SCCP) phones have the ability to preserve calls even when they detect the loss of their Unified CM. The caller-and-agent conversation continues until either the caller or agent goes on-hook. The Unified CVP Call Server recognizes that Unified CM has failed, assumes the call should be preserved, and maintains the signaling channel to the originating gateway. In this way, the originating gateway has no knowledge that Unified CM has failed. Note that additional activities in the call (such as hold, transfer, or conference) are not possible. Once the parties go on-hook, the phone then re-homes to another Unified CM server. When the agent goes on-hook, Real-Time Control Protocol (RTCP) packets cease transmitting to the originating gateway, which causes the gateway to disconnect the caller 9 to 18 seconds after the agent goes on-hook. If survivability has been configured on the gateway and the caller is waiting for some additional activity (the agent might think the caller is being blind-transferred to another destination), the caller is default-routed to an alternate location.
- New calls are directed to an alternate Unified CM server in the cluster.

Intelligent Contact Management (ICM)

Cisco Intelligent Contact Management (ICM) software provides enterprise-wide distribution of multichannel contacts (inbound/outbound telephone calls, Web collaboration requests, email messages, and chat requests) across geographically separated contact centers. ICM software is an open standards-based solution whose capabilities include routing, queuing, monitoring, and fault tolerance.

Configuration

For the most current information on configuring ICM for high availability, refer to the latest version of the *Cisco Unified Contact Center Enterprise Solution Reference Network Design (SRND)*, available at

http://www.cisco.com/en/US/products/sw/custcosw/ps1844/products_implementation_design_guides_list.html

Call Disposition

There are many components in Cisco ICM, and call disposition varies depending on the component that fails. Although there are a few exceptions, the following conditions apply to the call disposition:

- If the Voice Response Unit (VRU) Peripheral Gateway (PG) or any component on the VRU PG fails, calls in progress are default-routed by survivability on the originating gateway.
- If the Logger fails, calls in progress are unaffected.
- If the primary router fails, calls in progress are unaffected. If both the Side A and Side B routers fail, calls in progress are default-routed by survivability on the originating gateway.
- New calls are directed to the backup ICM component.