Solution Design Guide for Cisco Packaged Contact Center Enterprise, 12.5(1) & 12.5(2)

First Published: 2020-02-05

Last Modified: 2021-05-20

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

CONTENTS

Preface

Change History

This table lists the major changes made to this guide. The most recent changes appear at the top.

Changes	Section	Date
Added Support for Unified CM 15.0	Unified CCE Reference Designs	April, 2024
Added VAV and Agent Answers bandwidth information.	Bandwidth Provisioning and QoS Considerations for a WAN	Aug, 2023
Added a note for OVA when VVB fresh install is performed on a medium or small Open Virtual Appliance (OVA).	Cisco Virtualized Voice Browser Sizing	April, 2023
Increased configuration limit for active mobile agents with call-by-call connections per agent PG	Agent Limits PG Agent Capacity with Mobile Agents	March, 2021

Changes	Section	Date
Initial Release of Document for 12.5(1)		February, 2020
Added Reference Design layouts and VM specifications on Cisco UCS C240 M5SX and Cisco Hyperflex HX220c M5SX TRC servers for the following deployment types: • 2000 Agent Deployment • 4000 Agent Deployment • 12000 Agent Deployment	Contact Center Enterprise Reference Designs	
Added information on the new **Cisco Webex Experience Management Feature** feature.	Contact Center Enterprise Solutions Overview Design Considerations for Integrated Features	
Added information on the new **Customer Virtual Assistance Feature** feature.	Design Considerations for Integrated Features	
Added shared ACD Line support for both home and work phone on two shared ACD lines	Design Considerations for Integrated Features	
Increased configuration limits for the following: • Outbound dialer maximum calls per second per dialer • Outbound dialer maximum ports per SIP dialer • Number of preview campaigns per system • Number of predictive Campaigns per system (Agent or VRU based) • Predictive Campaign Skill Groups per Peripheral	Outbound Campaign Limits	
Removed information on Cisco MediaSense, Cisco Remote Expert, and Context Service which reached its end of maintenance support.		

Changes	Section	Date
Added a chapter for the Avaya and ICM-to-ICM Gateway Support	Avaya and ICM-to-ICM Gateway Support	
Updated the topics with support for Avaya and ICM-to-ICM Gateway	• Contact Center Enterprise Solution Comparison • Reference Designs and Deployment Types	

About This Guide

This guide provides design considerations and guidelines for deploying Cisco Unified Contact Center Enterprise (Unified CCE) solutions. The guide combines information for all the components that might be present in your solution. This guide assumes that you are familiar with basic contact center terms and concepts. Successful deployment of Unified CCE solutions also requires familiarity with the information presented in the *Cisco Collaboration System Solution Reference Network Designs*.

This guide focuses on the design process. Its goal is to present the necessary information to take your design from starting concept to final submission. Details of installation, configuration, and administration of your contact center enterprise solution are covered in other guides.

The first four chapters of the book give a broad perspective of the contact center enterprise solutions:

- Packaged Contact Center Enterprise
- Cisco Hosted Collaboration Solution for Contact Center
- Unified Contact Center Enterprise

For information about design considerations and guidelines specific to Packaged CCE, see the remaining chapters.

Audience

The first three chapters in this guide are for anyone who wants a broad overview of the contact center enterprise solutions.

The primary audience for the guide is people who design contact centers. The guide is also helpful for system administrators who want a deeper understanding of how the components in a contact center enterprise solution work together.

Related Documents

Consult these documents for details of these subjects that are not covered in this guide.

Subject	Link
Compatibility Matrix for information on which versions of which products are supported for a contact center enterprise solution.	https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-device-support-tables-list.html https://www.cisco.com/c/en/us/support/unified-communications/hosted-collaboration-solution-contact-center/products-device-support-tables-list.html
Cisco Unified Contact Center Enterprise Features Guide for detailed information on the configuration and administration of integrated features in your solution.	http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html
Cisco Collaboration Systems Solution Reference Network Designs for detailed information on the Unified Communications infrastructure on which your solution is built.	http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html
Cisco Packaged Contact Center Enterprise Features Guide	https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html
Cisco Packaged Contact Center Enterprise Administration and Configuration Guide for details on Avaya and ICM-to-ICM configurations.	

You can find the full documentation of each of the components in the contact center enterprise solutions at these sites:

Component	Link
Cisco Packaged Contact Center Enterprise	https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/tsd-products-support-series-home.html
Cisco Finesse	http://www.cisco.com/c/en/us/support/customer-collaboration/finesse/tsd-products-support-series-home.html
Cisco Customer Collaboration Platform	https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/tsd-products-support-series-home.html
Cisco Unified Customer Voice Portal	http://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/tsd-products-support-series-home.html
Cisco Unified Intelligence Center	http://www.cisco.com/c/en/us/support/customer-collaboration/unified-intelligence-center/tsd-products-support-series-home.html

Component	Link
Cisco Virtualized Voice Browser	http://www.cisco.com/c/en/us/support/customer-collaboration/virtualized-voice-browser/tsd-products-support-series-home.html

Communications, Services, and Additional Information

- To receive timely, relevant information from Cisco, sign up at Cisco Profile Manager.
- To get the business impact you're looking for with the technologies that matter, visit Cisco Services.
- To submit a service request, visit Cisco Support.
- To discover and browse secure, validated enterprise-class apps, products, solutions and services, visit Cisco Marketplace.
- To obtain general networking, training, and certification titles, visit Cisco Press.
- To find warranty information for a specific product or product family, access Cisco Warranty Finder.

Cisco Bug Search Tool

Cisco Bug Search Tool (BST) is a web-based tool that acts as a gateway to the Cisco bug tracking system that maintains a comprehensive list of defects and vulnerabilities in Cisco products and software. BST provides you with detailed defect information about your products and software.

Field Notice

Cisco publishes Field Notices to notify customers and partners about significant issues in Cisco products that typically require an upgrade, workaround, or other user action. For more information, see *Product Field Notice Summary* at https://www.cisco.com/c/en/us/support/web/tsd-products-field-notice-summary.html.

You can create custom subscriptions for Cisco products, series, or software to receive email alerts or consume RSS feeds when new announcements are released for the following notices:

- Cisco Security Advisories
- Field Notices
- End-of-Sale or Support Announcements
- Software Updates
- Updates to Known Bugs

For more information on creating custom subscriptions, see *My Notifications* at https://cway.cisco.com/mynotifications.

Documentation Feedback

To provide comments about this document, send an email message to the following address: contactcenterproducts_docfeedback@cisco.com

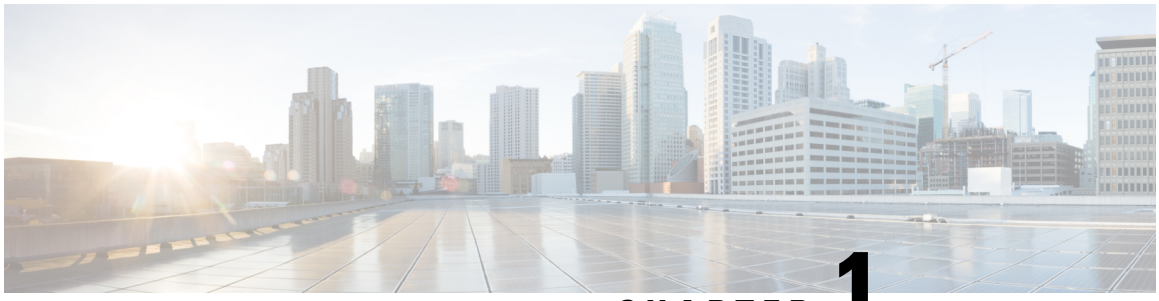
We appreciate your comments.

Conventions

This document uses the following conventions:

Table 2: Conventions

Convention	Description
boldface font	Boldface font is used to indicate commands, such as user entries, keys, buttons, folder names, and submenu names. For example: • Choose **Edit** > **Find**. • Click **Finish**.
italic font	Italic font is used to indicate the following: • To introduce a new term. Example: A *skill group* is a collection of agents who share similar skills. • A syntax value that the user must replace. Example: IF (*condition, true-value, false-value*) • A book title. Example: See the *Cisco Unified Contact Center Enterprise Installation and Upgrade Guide*.
`window font`	Window font, such as Courier, is used for the following: • Text as it appears in code or that the window displays. Example: `<html><title>Cisco Systems, Inc. </title></html>`
`< >`	Angle brackets are used to indicate the following: • For arguments where the context does not allow italic, such as ASCII output. • A character string that the user enters but that does not appear on the window such as a password.



CHAPTER 1

Cisco Unified Contact Center Solutions

Contact Center Enterprise Solutions

Note The first four chapters of this book are for anyone who wants to get familiar with the contact center enterprise solutions:

- Packaged Contact Center Enterprise
- Cisco Hosted Collaboration Solution for Contact Center
- Unified Contact Center Enterprise

For information about design considerations and guidelines specific to Packaged CCE, see the remaining chapters.

Cisco contact center enterprise solutions support several deployment models to meet various business needs. Choose the deployment model that fits your needs.

Cisco offers these contact center enterprise solutions:

- Cisco Packaged Contact Center Enterprise Solution—A predesigned solution for up to 12000 multichannel agents
- Cisco Hosted Collaboration Solution for Contact Center—For Service Providers who offer cloud contact center solutions for up to 24000 multichannel agents
- Cisco Unified Contact Center Enterprise Solution—For enterprise customers who need scale and flexibility for up to 24000 multichannel agents

All of the contact center enterprise solutions use a redundant architecture for high availability and solution serviceability. They provide a comprehensive feature set, including:

- Call processing and call control
- Web chat, email, and callback

- Social media monitoring
- Rich VRU and routing scripting
- Interactive voice and video response unit
- Voice and video recording and streaming
- Agent selection and queuing
- Web-based agent and supervisor desktops
- Comprehensive reporting

Packaged Contact Center Enterprise

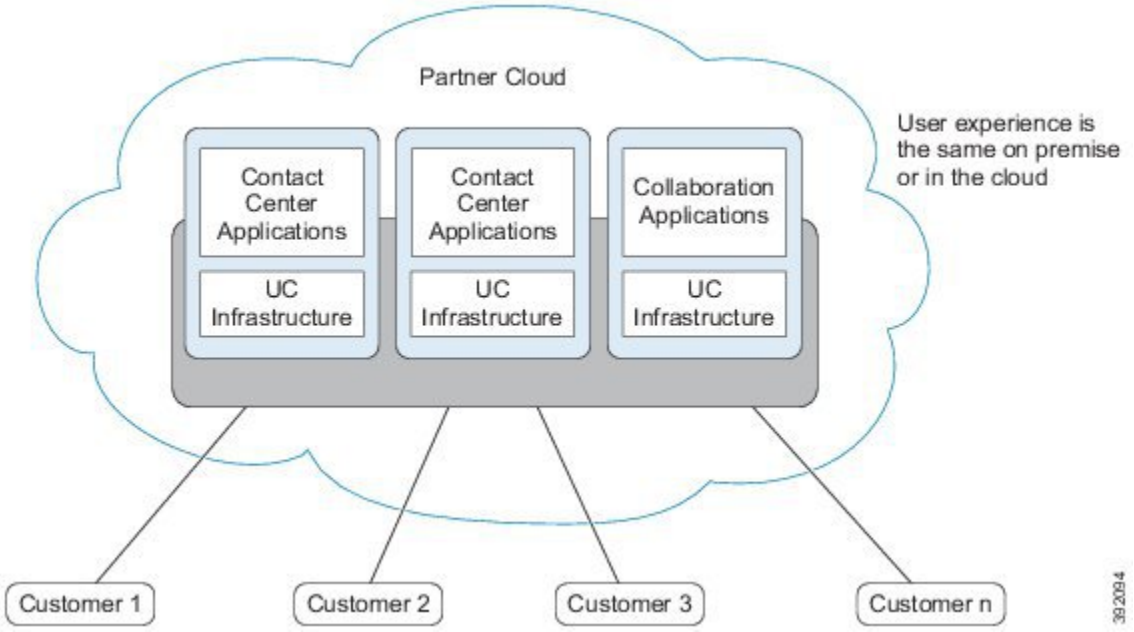
Cisco Packaged Contact Center Enterprise (Packaged CCE) is a predefined deployment.

If your requirements fit the boundaries of the solution, Packaged CCE simplifies some major features of your deployment. Enjoy the advantages of a simplified management interface and reduced time to install. The solution supports up to 12,000 concurrent active agents.

Hosted Collaboration Solution for Contact Center

Cisco Hosted Collaboration Solution for Contact Center (Cisco HCS for Contact Center) is designed for service providers who offer hosted contact center services to their end customers. Cisco HCS for Contact Center deployments can support small to large end customers. As a service provider, you operate the solution for your end customers. This contact center enterprise solution enables your end customers to tap into the applications and services of the Cisco Hosted Collaboration Solution. Cisco HCS for Contact Center enables you to deliver the capabilities of a contact center enterprise solution as a cloud service to the market with shared aggregation and administration layers across tenants, such as CUBE, ASA, Unified CCDM, UCDM, and Prime Collaboration Assurance..

Figure 1: Cisco HCS for Contact Center



Unified Contact Center Enterprise

Cisco Unified Contact Center Enterprise (Unified CCE) offers industry-leading routing capabilities and CTI capabilities. Unified CCE provides the options to support a single-site or a multisite environment.

Unified CCE supports multiple Unified CM clusters. You can also network multiple Unified CCE instances together for greater system capacity.

Contact Center Enterprise Solution Comparison

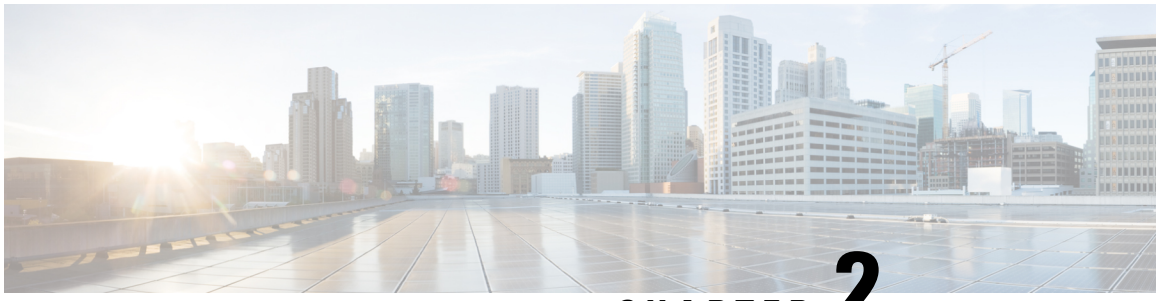
This table highlights the differences between the contact center enterprise solutions:

Table 3: Comparison of Contact Center Enterprise Solutions

Feature	Packaged CCE	Cisco HCS for Contact Center (for a single instance)	Unified CCE (for a single instance)
Deployments and Scalability	Preconfigured Reference Designs (2000, 4000, and 12000 agents) Avaya PG and ICM-to-ICM Gateway as a Non-Reference Design (4000 and 12000 agents)	Preconfigured Reference Designs (2000, 4000, 12000, and 24000 agents)[1]	Reference Designs (2000, 4000, 12000, 24000 agents, and Contact Director [24000 agents across 3 instances]) Non-Reference Design deployments

Feature	Packaged CCE	Cisco HCS for Contact Center (for a single instance)	Unified CCE (for a single instance)
Hardware	Tested Reference Configuration (B-Series and C-Series) Spec-based UCS support (B-Series and C-Series) Spec-based (Non-UCS), third-party hardware support	Tested Reference Configuration (B-Series and C-Series) Spec-based UCS support (B-Series and C-Series)	Tested Reference Configuration (B-Series and C-Series) Spec-based UCS support (B-Series and C-Series) Spec-based (Non-UCS), third-party hardware support
Architecture			
Configuration and Administration tools	Gadget-based CCE Administration and other tools	Partial gadget-based CCE Administration, Unified Contact Center Domain Manager, ICM Configuration Manager, and other tools	Partial gadget-based CCE Administration, Unified Contact Center Management Portal, ICM Configuration Manager, and other tools

[1] Includes the 500 Agent variation of 2000 Agents and the Small Contact Center variation of 4000 Agents. The only Non-Reference Design element that Cisco HCS for Contact Center supports is the Avaya PG.



CHAPTER 2

Unified CCE Reference Designs

Introduction to the Reference Designs

Note The first four chapters of this book are for anyone who wants to get familiar with the three contact center enterprise solutions:

- Packaged Contact Center Enterprise
- Cisco Hosted Collaboration Solution for Contact Center
- Unified Contact Center Enterprise

For information about design considerations and guidelines specific to Packaged CCE, see the remaining chapters.

The Contact Center Enterprise Reference Designs are a set of Cisco validated designs of our contact center enterprise solutions. The Reference Designs define the technologies and topologies that fit the needs for most deployments. The Reference Designs focus on simplifying the contact center enterprise solution design. They provide complete contact center functionality based on components that are strategic to Cisco.

We have defined the Reference Designs in the following table to cover most contact center needs:

Table 4: Reference Design Use by Contact Center Enterprise Solution

Reference Design	Packaged CCE	Cisco HCS for Contact Center	Unified CCE
2000 Agents	Yes	Yes	Yes
4000 Agents	Yes	Yes	Yes

Reference Design	Packaged CCE	Cisco HCS for Contact Center	Unified CCE
12000 Agents	Yes	Yes	Yes
24000 Agents	No	Yes	Yes
Contact Director	No	No	Yes

If your solution exceeds the configuration limits for a particular Reference Design, use a Reference Design with higher limits. For example, if your 2000-agent deployment requires 350 active reporting users, use the 4000 Agent Reference Design for your solution.

Reference Designs and Deployment Types

The Contact Center Enterprise Reference Designs are mapped to specific contact center solutions through deployment types. Deployment types are system codes that impose system limits and apply congestion control.

This table maps the Reference Designs and Non-Reference Designs with the deployment type that you use for each.

Table 5: Deployment Type Usage by Reference Design

Reference Design	Packaged CCE	Cisco HCS for Contact Center	Unified CCE
	Label	**Label**	**Label**
2000 Agent	Packaged CCE: 2000 Agents	HCS-CC: 2000 Agents	UCCE: 2000 Agents
4000 Agent	Packaged CCE: 4000 Agents	HCS-CC: 4000 Agents	UCCE: 4000 Agents
12000 Agent	Packaged CCE: 12000 Agents	HCS-CC: 12000 Agents	UCCE: 12000 Agents
24000 Agent	NA	HCS-CC: 24000 Agents	UCCE: 24000 Agents Router/Logger
Contact Director	NA	NA	Contact Director
Non-Reference Designs	Avaya PG and ICM-to-ICM Gateway Packaged CCE: 4000 Agents Packaged CCE: 12000 Agents	NA	ICM Rogger
			ICM Router/Logger
			UCCE: 8000 Agents Router/Logger
Lab Only Designs	Packaged CCE: Lab Mode	NA	UCCE: Progger (Lab Only)

Note After a Packaged CCE deployment is initialized, you cannot switch to another Packaged CCE deployment type. However, you can switch to a Unified CCE deployment type.

Benefits of a Reference Design Solution

Contact centers offer more possibilities with each new generation of software and hardware. New technology can make previously preferred methods obsolete for current contact centers. We created the Contact Center Enterprise Reference Designs to simplify your design choices and speed the development of your contact center. We expect that most new contact centers can use the Reference Designs to meet their needs.

By following the Reference Designs, you can:

- Guide your customers' expectations by presenting clear options.
- Streamline your design process with standard models.
- Avoid using components and features that are near the end of their lifecycle.
- Find powerful and efficient replacements for obsolete features.
- Align your designs with Cisco's vision of our future contact center developments.
- Enjoy quicker and easier approval processes.

Specifications for a Reference Design Solution

The Reference Designs define our vision of the functionality that most contact centers use. The Reference Designs consist of:

- **Core components**—Components that make up every contact center:
  - Ingress, Egress, and VXML Gateways
  - Unified Customer Voice Portal (Unified CVP)
  - Unified Contact Center Enterprise (Unified CCE)
  - Cisco Virtualized Voice Browser (VVB)
  - Unified Communications Manager (Unified CM)
  - Cisco Finesse
  - Cisco Unified Intelligence Center
- **Optional Cisco components**—Components that add functionality that not every contact center needs.
  - Customer Collaboration Platform
  - Cisco Unified SIP Proxy
  - Enterprise Chat and Email

- Cisco IdS
- Cloud Connect

- **Optional third-party components**—Third-party components that you can add to provide other features.
  - Load balancers
  - Recording
  - Speech servers - ASR/TTS
  - Wallboards
  - Workforce management

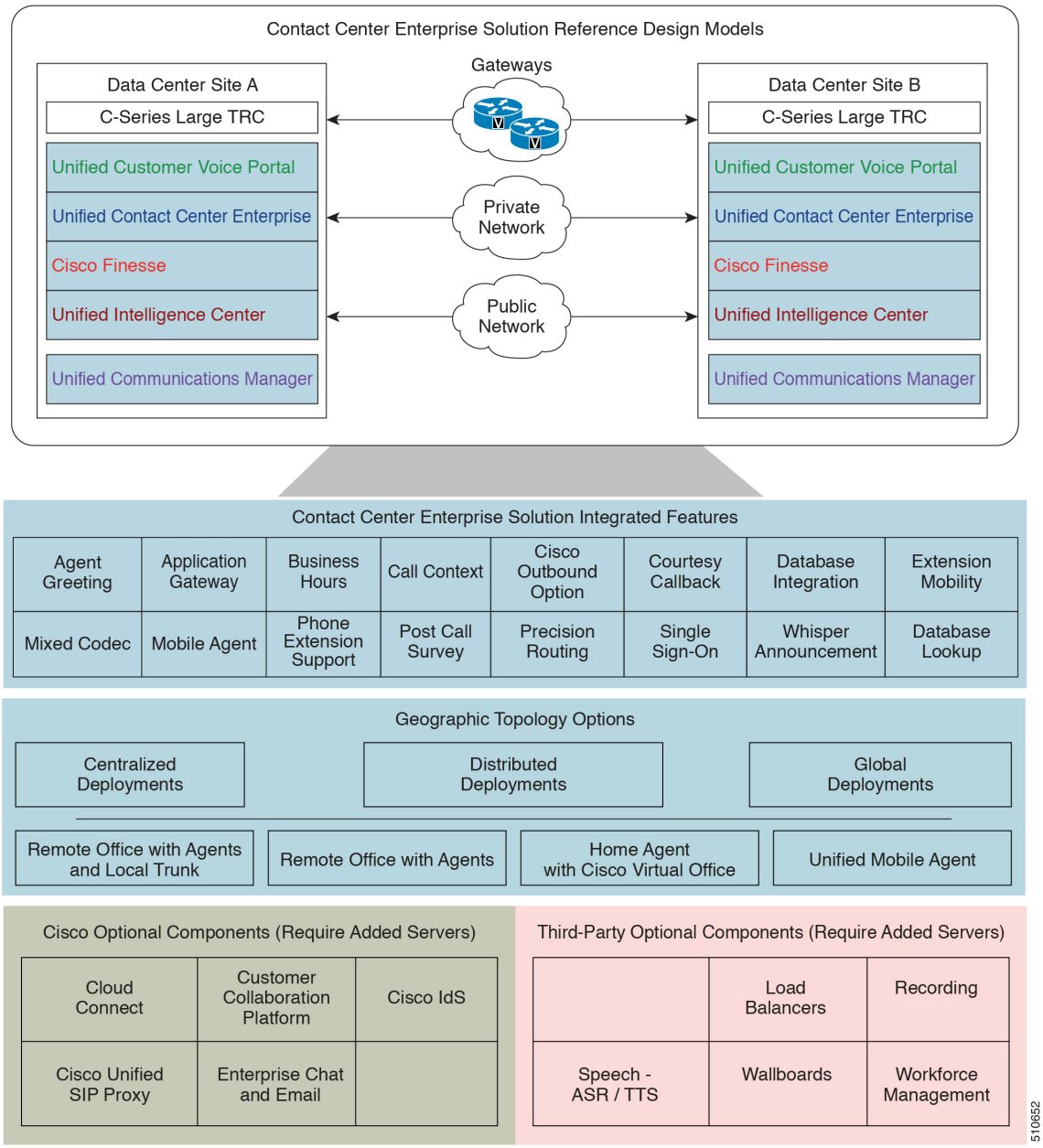
- **Integrated features**—These features do not require you to add an optional solution component to enable them. But, these features can require configuration in multiple solution components to activate them. They can affect your solution sizing and might have specific design considerations.
- **Call flows**—Standard contact handling and routing methods.
  - Inbound Calls:
    - New calls from a carrier
    - New internal calls
  - Supplementary services
    - Hold and resume
    - Transfers and conferences
    - Refer transfers
    - Network transfers
    - Requery and survivability

- **Topologies**—Standard layouts for your contact center components:
  - Centralized
  - Distributed
  - Global

Note In general, you cannot use the ICM-to-ICM Gateway in Reference Designs. Only the Contact Director Reference Design allows you to use that gateway.

This figure encapsulates the basic requirements of a Reference Design-compliant deployment:

Figure 2: Contact Center Enterprise Components and Features



Contact Center Enterprise Reference Designs

The following sections describe the Contact Center Enterprise Reference Designs.

The Reference Designs are supported for Cisco UCS C240 M5SX, Cisco UCS C240 M6SX, and Cisco HX220c-M5SX Tested Reference Configuration (TRC) servers as detailed in the Cisco Collaboration Infrastructure Requirements wiki: https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-infrastructure.html.

Note For more details on supported servers for the Reference Designs, see the *Cisco Collaboration Virtualization* page for your solution at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

The following notes apply to all the Reference Designs:

- Contact Center Enterprise solutions use vCPU oversubscription.
- The standard PG VM includes an Agent (Unified CM) PG, a VRU PG, and an MR PG. Unified CCE and Cisco HCS for Contact Center allow you to add more PGs and their peripherals onto this base layout.
- Cloud Connect, a component that allows you to use cloud services, is available only with Cisco Webex Experience Management.
- Cloud Connect can be on-box (as depicted in the following sections) for deployments on the Cisco HX220c-M5SX server, whereas, on the Cisco UCS C240 M5SX or Cisco UCS C240 M6SX servers, Cloud Connect must be off-box.
- CVP Reporting server, Cisco VVB, and Cloud Connect are optional components.
- The TRC layouts for Cisco UCS C240 M5SX and Cisco UCS C240 M6SX servers are identical. Note that only a single-socket 28-core CPU is used for the Cisco UCS C240 M6SX servers. If customers wish to use the additional socket on the Cisco UCS C240 M6SX servers with corresponding increase in cores, memory, and disks, the hardware will be supported under spec-based VM provisioning policies.
- Cisco HX220c-M6S servers are supported in accordance with spec-based policies only.
- CVP Reporting server and Cisco VVB are optional components.
- Based on your business and deployment requirements, you may distribute the VVB VMs on external servers, or as depicted in this section, deploy them on additional servers or nodes (in the case of M5-HX clusters).
- If the layout is on the Cisco HX220c-M5SX or Cisco HX220c-M6S server, you can deploy the additional VVB servers on HX nodes in the same cluster, or on external M5 or M6 servers, respectively.
- VVB with AppD enabled CPU MHz utilization spikes during services start up. VVB OVA profiles has upper threshold set as unlimited so there are no changes in OVA profile. This impact is only during services start up but not under general or load scenarios.
- An HX cluster can consist of a combination of compute and converged nodes, provided that all resource requirements and resource constraints are satisfied in accordance with the Virtual Machine Resource Provisioning Policy. This is supported only as a spec-based deployment model.
- For information on the data source allocation of the components in the Reference Design layouts, see the *Cisco Packaged Contact Center Enterprise Installation and Upgrade Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-installation-guides-list.html
- The Reference Design layouts in this section do not show off-box components like Customer Collaboration Platform.
- If you upgrade to Release 12.6(1) on Cisco UCS C240 M4SX server, we recommend that you install an additional 32 GB of memory on all servers to accommodate the increased memory requirements of the Release 12.6(1) VMs.

Virtual Machines Resource Provisioning Policy

Note The previously used Oversubscription policy is a part of the Virtual Machine (VM) Resource Provisioning Policy.

The Unified CCE Reference Designs support the virtual machine vCPU oversubscription of the physical CPU cores on a server. For the purposes of oversubscription, the hyper-thread cores do not count as physical cores. Whether or not you use oversubscription, use the VM Resource Provisioning policy. This policy limits the total available CPU MHz and the memory of a server that the host-resident VMs can consume.

Apply the VM Resource Provisioning policy when:

- You provision a Reference Design server for optional and third-party components that are not given a reference VM layout.
- You use UCS servers.
- You upgrade an existing solution and do not migrate to a Reference Design VM layout.

Note Apply the VM Resource Provisioning policy on a per-server basis. This policy does not apply to the Reference Design VM layouts. Your solution can contain servers that use the Reference Design VM layouts and other VM layouts that use the VM Resource Provisioning policy rules.

The application of the VM Resource Provisioning policy requires meeting the following conditions:

- You can use up to two vCPUs for every physical core on each server.
- You can use up to 65% of the total available CPU MHz on each server.
- You can use up to 80% of the total available memory on each server.

For more information on virtualization and specification-based server policies, see the *Cisco Collaboration Virtualization* at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

Note The Virtual Machine Placement Tool does not currently allow you to oversubscribe. This limitation is only an issue with the tool. You can oversubscribe within the limits that are provided here.

Note The custom remote server should have the same provisioning specification as Unified CVP server.

2000 Agent Reference Designs

All contact center enterprise solutions support the 2000 Agent Reference design on the Cisco UCS C240 M5SX or Cisco UCS C240 M6SX and the Cisco HX220c-M5SX Large TRC servers.

- In this Reference Design, Cisco Unified Intelligence Center, Live Data, and the Identity Service for Single Sign-On are coresident on a single VM. In the larger Reference Designs, they reside in separate VMs.
- You can optionally deploy the Unified Communications Manager Publisher and Subscribers on separate servers, instead of deploying them as shown in the 2000 Agent Reference Design layout. You should dedicate two of the subscribers to Unified CCE. All devices on these subscribers must be SIP.

  In 2000 Agent Reference Designs, a coresident Unified CM can support a maximum of 2000 phones. This includes your phones for all types of agents, whether contact center agents or back-office workers. If your solution requires more than 2000 phones, use a Unified CM on a separate server instead.
- In the global deployment topology, each remote site can have its own Unified CM cluster. A remote site cannot include a Cisco Unified Intelligence Center server.
- In Packaged CCE global deployments, you cannot create a remote site without PG VMs.
- You can deploy optional AW-HDS-DDS per site on external servers for longer data retention.
- In 2000 Agent Reference Designs, you can deploy ECE Data Server on-box for up to 400 agents. Deploy ECE off-box for up to 1500 agents.

  You can also deploy the ECE Data Server on a separate server.
- Deploy the ECE Web Server on an external server. You can place that server either in the same data center as the ECE Data Server or in a DMZ if customer chat interactions require that.

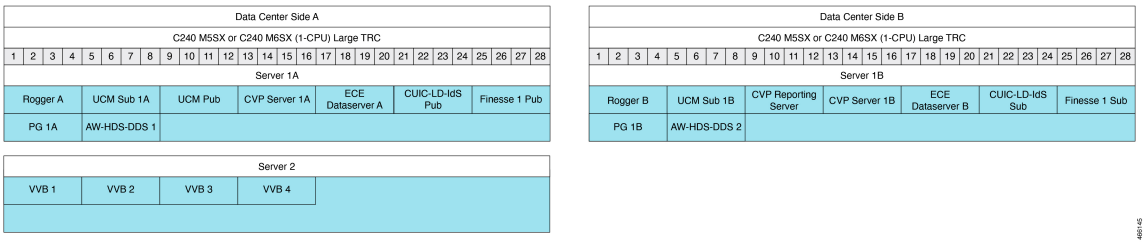
Note Adding more disks is not permitted in the Packaged CCE 2000 agent deployment. Any changes to the number of disks will result in a VM validation error.

Support on the Cisco UCS C240 M5SX and Cisco UCS C240 M6SX Large TRC Servers

Important If you plan to upgrade to 12.x on Cisco UCS C240 M4SX servers, deploy Unified CM and ECE HA VMs on external servers.

The following figure shows the base layout of the components in a 2000 Agent Reference Design on Cisco UCS C240 M5SX and Cisco UCS C240 M6SX Large TRC servers.

Figure 3: 2000 Agent Reference Design Model



This table lists the specifications for VMs.

Table 6: VM Specifications for 2000 Agent Reference Design

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2	vDisk 3
Rogger	4	5000	6	80	150	
Unified CM	4	7200	8[2]	110		
Unified CVP Server	4	3000	12	250		
Unified CVP Reporting Server	4	1800	6	80	438	
ECE Dataserver[3]	4	4000	20	80	50	300
CUIC-LD-IdS	4	5500	16	200		
AW-HDS-DDS	4	5000	16	80	750	
PG	2	4000	6	80		
Finesse	4	5000	10	146		
VVB	4	9000	10	146		

[2] The vRAM value for Unified CM 15.0 is 12 GB. The vRAM value in the Total Requirements table below will also vary and needs to be recalculated accordingly. For more details, see the Compatibility Matrix.

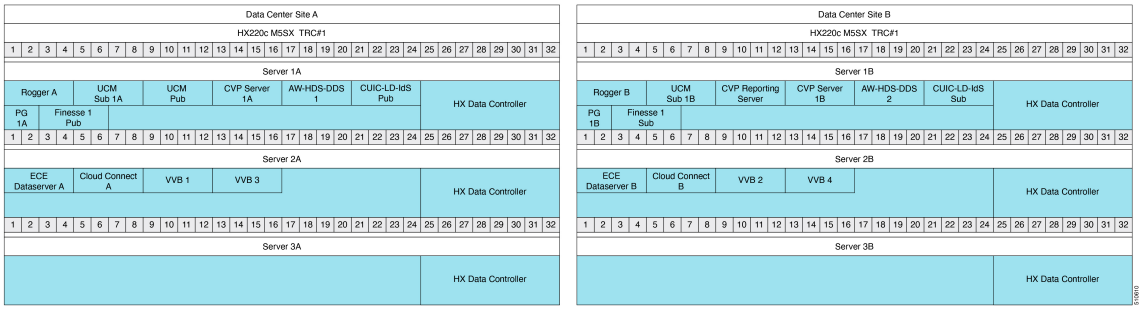
[3] For the latest VM specifications, see the row for 400 agents in the **Virtualization for Enterprise Chat and Email** page at https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-enterprise-chat-email.html.

Table 7: Total VM Requirements for 2000 Agent Reference Design

Server	vCPU	MHz	vRAM	vDisk
Data Center Site A	34	45900	102	2386
Data Center Site B	30	40500	100	2648
Server 2	16	36000	40	584

Support on the Cisco HX220c-M5SX TRC Server

This figure shows the base layout of the components in a 2000 Agent Reference Design on Cisco HX220c-M5SX TRC server.



This table lists the specifications for VMs.

Table 8: VM Specifications for 2000 Agent Reference Design

VM	vCPU	MHz	vRAM	vDisk1	vDisk2	vDisk 3
HX Data Controller	16	10800	48			
Rogger	4	5000	6	80	150	
Unified CM	4	7200	8[4]	110		
Unified CVP Server	4	3000	12	250		
Unified CVP Reporting Server	4	1800	6	80	438	
ECE Dataserver[5]	4	4000	20	80	50	300
CUIC-LD-IdS	4	5500	16	200		
AW-HDS-DDS	4	5000	16	80	750	
PG	2	4000	6	80		
Finesse	4	5000	10	146		
VVB	4	9000	10	146		
Cloud Connect	4	6000	10	146		

[4] The vRAM value for Unified CM 15.0 is 12 GB. The vRAM value in the Total Requirements table below will also vary and needs to be recalculated accordingly. For more details, see the Compatibility Matrix.

[5] For the latest VM specifications, see the row for 400 agents in the **Virtualization for Enterprise Chat and Email** page at https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-enterprise-chat-email.html.

Table 9: Total VM Requirements for 2000 Agent Reference Design

Server	vCPU	MHz	vRAM	vDisk
Data Center Site 1A	46	52700	130	1956

Server	vCPU	MHz	vRAM	vDisk
Data Center Site 1B	46	47300	128	2364
Data Center Site 2A	32	38800	98	868
Data Center Site 2B	32	38800	98	868

4000 Agent Reference Designs

All contact center enterprise solutions support the 4000 Agent Reference design on the following TRC servers:

- Cisco UCS C240 M5SX Large
- Cisco UCS C240 M6SX Large
- Cisco HX220c-M5SX

This model adds servers to scale up from the 2000 Agent Reference Design.

Note You can only deploy two AW-HDS-DDS per data center site in the 4000 Agent Reference Design. In larger solutions, you use a combination of HDS-DDS and AW-HDS.

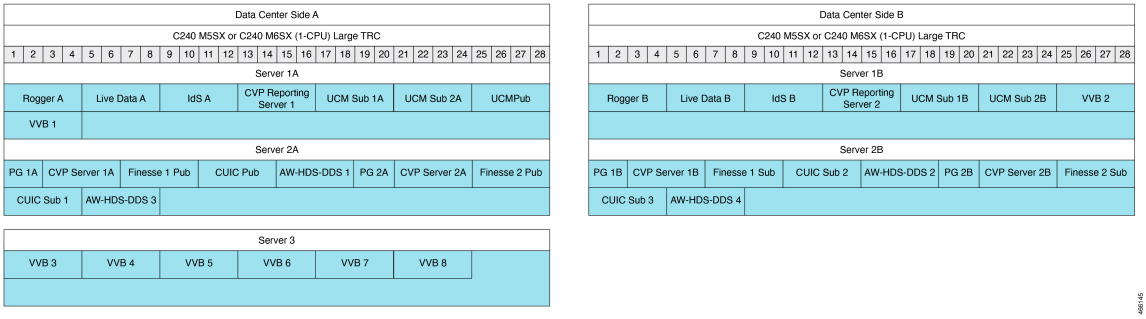
Support on the Cisco UCS C240 M5SX and Cisco UCS C240 M6SX TRC Servers

Important If you plan to upgrade to 12.x on Cisco UCS C240 M4SX servers, make the following changes to your servers and VM layouts:

- Deploy Unified CM and ECE HA VMs on external servers.
- Add 16 GB of physical RAM to each server that hosts Unified CVP call and VXML servers.
- Increase the memory reservations for the Unified CVP VMs to 12 GB.

This figure shows the base layout of the components in a 4000 Agent Reference Design on Cisco UCS C240 M5SX and Cisco UCS C240 M6SX TRC servers.

Figure 4: 4000 Agent Reference Design Model



This table lists the specifications for VMs.

Table 10: VM Specifications for 4000 Agent Reference Design

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2	
Rogger	4	5000	6	80	150	
Live Data	4	5500	30	146		
IdS	4	1500	10	146		
Unified CVP Reporting Server	4	1800	6	80	438	
Unified CM	4	7200	8[6]	110		
PG	2	4000	6	80		
Unified CVP Server	4	3000	12	250		
Finesse	4	5000	10	146		
Unified Intelligence Center	4	3600	16	200		
AW-HDS-DDS	4	5000	16	80	750	
VVB	4	9000	10	146		

[6] The vRAM value for Unified CM 15.0 is 12 GB. The vRAM value in the Total Requirements table below will also vary and needs to be recalculated accordingly. For more details, see the Compatibility Matrix.

Table 11: Total VM Requirements for 4000 Agent Reference Design

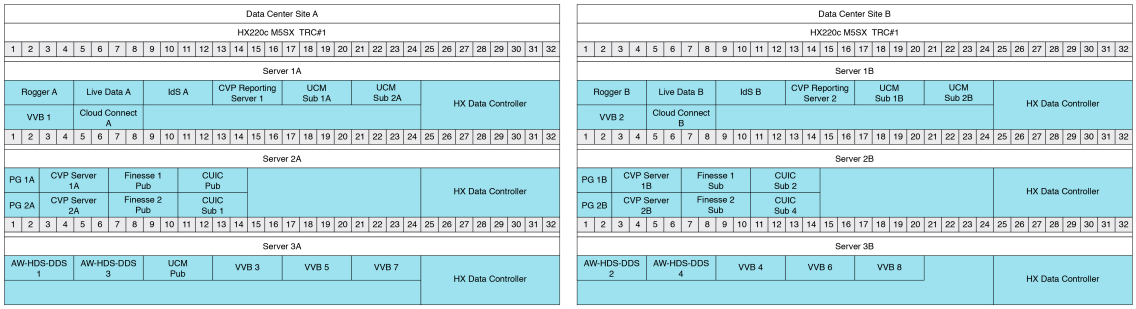
Server	vCPU	MHz	vRAM	vDisk
Data Center Site A - Server 1A	32	44400	86	1516
Data Center Site B - Server 1B	28	37200	78	1406

Server	vCPU	MHz	vRAM	vDisk
Data Center Site A - Server 2A	36	45000	120	2762
Data Center Site B - Server 2B	36	45000	120	2762
Server 3	24	54000	60	876

Support on the Cisco HX220c-M5SX TRC Server

This figure shows the base layout of the components in a 4000 Agent Reference Design on Cisco HX220c-M5SX TRC server.

Figure 5: 4000 Agent Reference Design Model



This table lists the specifications for VMs.

Table 12: VM Specifications for 4000 Agent Reference Design

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2	
HX Data Controller	16	10800	48			
Rogger	4	5000	6	80	150	
Live Data	4	5500	30	146		
IdS	4	1500	10	146		
Unified CVP Reporting Server	4	1800	6	80	438	
Unified CM	4	7200	8[7]	110		
PG	2	4000	6	80		
Unified CVP Server	4	3000	12	250		
Finesse	4	5000	10	146		
Unified Intelligence Center	4	3600	16	200		
AW-HDS-DDS	4	5000	16	80	750	

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2	
VVB	4	9000	10	146		
Cloud Connect	4	6000	10	146		

[7] The vRAM value for Unified CM 15.0 is 12 GB. The vRAM value in the Total Requirements table below will also vary and needs to be recalculated accordingly. For more details, see the Compatibility Matrix.

Table 13: Total VM Requirements for 4000 Agent Reference Design

Server	vCPU	MHz	vRAM	vDisk
Data Center Site A - Server 1A	48	54000	136	1552
Data Center Site B - Server 1B	48	54000	136	1552
Data Center Site A - Server 2A	48	50800	152	1932
Data Center Site B - Server 2B	48	50800	152	1932
Data Center Site A - Server 3A	24	44200	70	1958
Data Center Site B - Server 3B	20	37000	62	1848

12000 Agent Reference Designs

This Reference Design for a contact center enterprise solution supports 12000 agents on the following TRC servers:

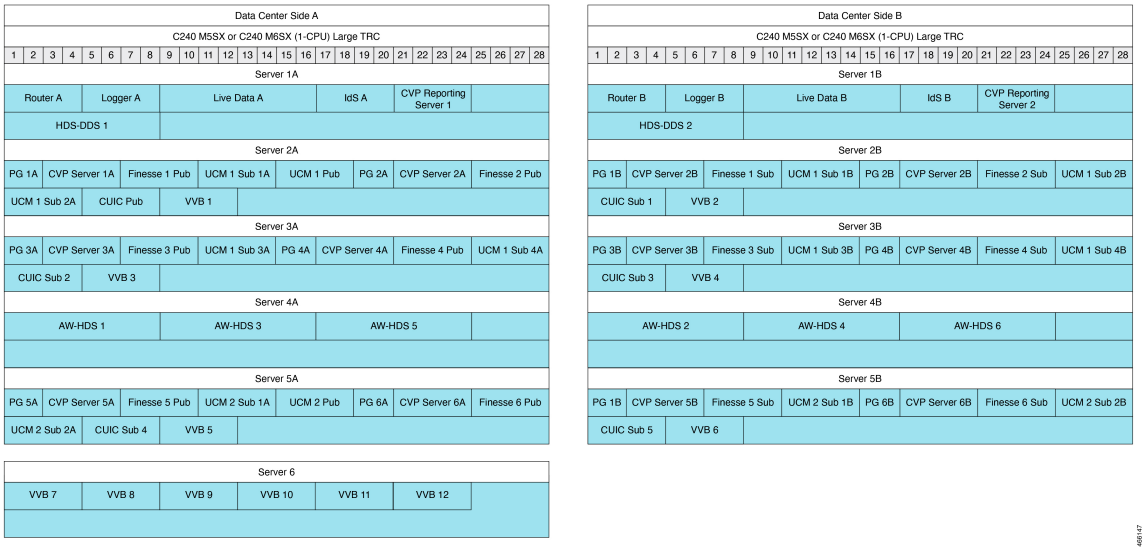
- Cisco UCS C240 M5SX Large
- Cisco UCS C240 M6SX Large
- Cisco HX220c-M5SX

This model adds servers to scale up from the 4000 Agent Reference Design.

Support on the Cisco UCS C240 M5SX and Cisco UCS C240 M6SX Large TRC servers

The following figure shows the base layout of the components in a 12000 Agent Reference Design on Cisco UCS C240 M5SX and Cisco UCS C240 M6SX Large TRC servers.

Figure 6: 12000 Agent Reference Design Model



This table lists the specifications for VMs.

Table 14: VM Specifications for 12000 Agent Reference Design

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2
Router	4	4000	8	80	
Logger	4	6000	8	80	500
Live Data	8	16500	30	146	
IdS	4	1500	10	146	
Unified CVP Reporting Server	4	1800	6	80	438
HDS-DDS	8	17500	16	80	500
AW-HDS	8	17500	16	80	500
PG	2	4000	6	80	
Unified CVP Server	4	3000	12	250	
Finesse	4	5000	10	146	
Unified CM	4	7200	8[8]	110	
Unified Intelligence Center	4	3600	16	200	
VVB	4	9000	10	146	

[8] The vRAM value for Unified CM 15.0 is 12 GB. The vRAM value in the Total Requirements table below will also vary and needs to be recalculated accordingly. For more details, see the Compatibility Matrix.

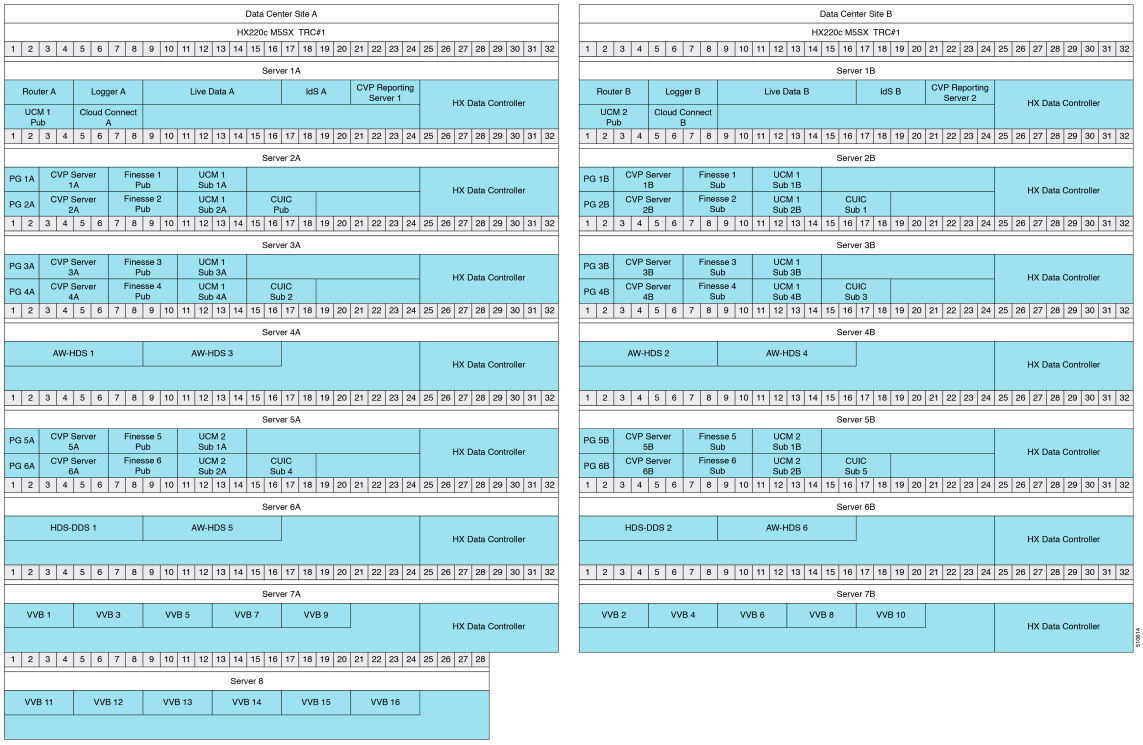
Table 15: Total VM Requirements for 12000 Agent Reference Design

Server	vCPU	MHz	vRAM	vDisk
Data Center Site A - Server 1A	32	47300	78	2050
Data Center Site B - Server 1B	32	47300	78	2050
Data Center Site A - Server 2A	40	60100	106	1628
Data Center Site B - Server 2B	36	52900	98	1518
Data Center Site A - Server 3A	36	52900	98	1518
Data Center Site B - Server 3B	36	52900	98	1518
Data Center Site A - Server 4A	24	52500	48	1740
Data Center Site B - Server 4B	24	52500	48	1740
Data Center Site A - Server 5A	40	60100	106	1628
Data Center Site B - Server 5B	36	52900	98	1518
Server 6	24	54000	60	876

Support on the Cisco HX220c-M5SX TRC Server

This figure shows the base layout of the components in a 12000 Agent Reference Design on Cisco HX220c-M5SX TRC server.

Figure 7: 12000 Agent Reference Design Model



This table lists the specifications for VMs.

Table 16: VM Specifications for 12000 Agent Reference Design

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2
HX Data Controller	16	10800	48		
Router	4	4000	8	80	
Logger	4	6000	8	80	500
Live Data	8	16500	30	146	
IdS	4	1500	10	146	
Unified CVP Reporting Server	4	1800	6	80	438
HDS-DDS	8	17500	16	80	420
AW-HDS	8	17500	16	80	500
PG	2	4000	6	80	
Unified CVP Server	4	3000	12	250	
Finesse	4	5000	10	146	

VM	vCPU	MHz	vRAM	vDisk 1	vDisk 2
Unified CM	4	7200	8[9]	110	
Unified Intelligence Center	4	3600	16	200	
VVB	4	9000	10	146	
Cloud Connect	4	6000	10	146	

[9] The vRAM value for Unified CM 15.0 is 12 GB. The vRAM value in the Total Requirements table below will also vary and needs to be recalculated accordingly. For more details, see the Compatibility Matrix.

Table 17: Total VM Requirements for 12000 Agent Reference Design

Server	vCPU	MHz	vRAM	vDisk
Data Center Site A - Server 1A	48	53800	128	1726
Data Center Site B - Server 1B	48	53800	128	1726
Data Center Site A - Server 2A	48	54700	136	1372
Data Center Site B - Server 2B	48	54700	136	1372
Data Center Site A - Server 3A	48	54700	136	1372
Data Center Site B - Server 3B	48	54700	136	1372
Data Center Site A - Server 4A	32	45800	80	1160
Data Center Site B - Server 4B	32	45800	80	1160
Data Center Site A - Server 5A	48	54700	136	1372
Data Center Site B - Server 5B	48	54700	136	1372
Data Center Site A - Server 6A	32	45800	80	1080
Data Center Site B - Server 6B	32	45800	80	1080
Data Center Site A - Server 7A	36	55800	98	730
Data Center Site A - Server 7B	36	55800	98	730
Server 8	24	54000	60	876

Reporting Users in the 12000 Agent Reference Design Model

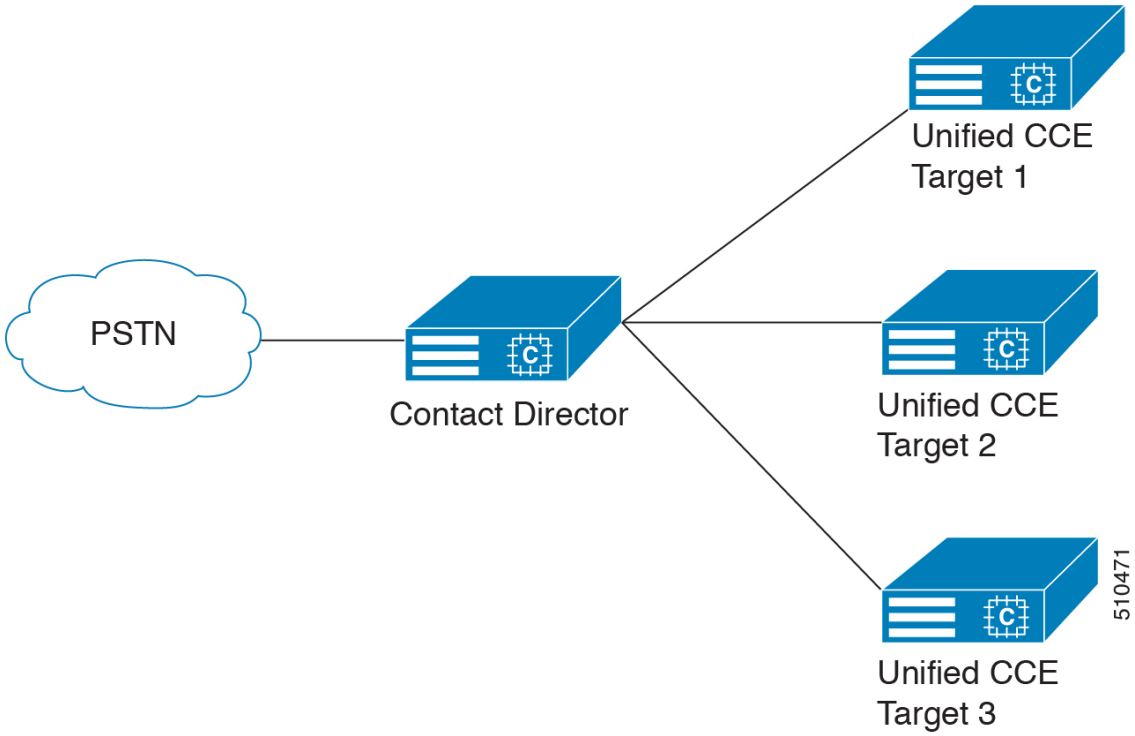
AW-HDS 3, AW-HDS 4, AW-HDS 5, and AW-HDS 6 in Servers 4A and 4B, are optional to support more than 400 reporting users. Servers 5A and 5B are optional to support more than 8000 agents. Servers 6A and 6B are optional to support more than 400 reporting users.

This Reference Design supports a maximum of six CUIC VMs and six AW-HDS VMs, three VMs on each site. This limit can accommodate a maximum of 1200 reporting users. If one site shuts down, the remaining site can only support 600 reporting users on its three nodes.

Contact Director

Only Unified CCE supports the Contact Director reference design. The Contact Director distributes incoming calls to other contact center instances. The targets can be Unified CCE instances or Unified ICM instances that connect to third-party contact centers. The Contact Sharing feature uses a Contact Director to distribute incoming contacts to a maximum of 3 Unified CCE instances.

Figure 8: Contact Director Solution with Two Unified CCE Target Instances

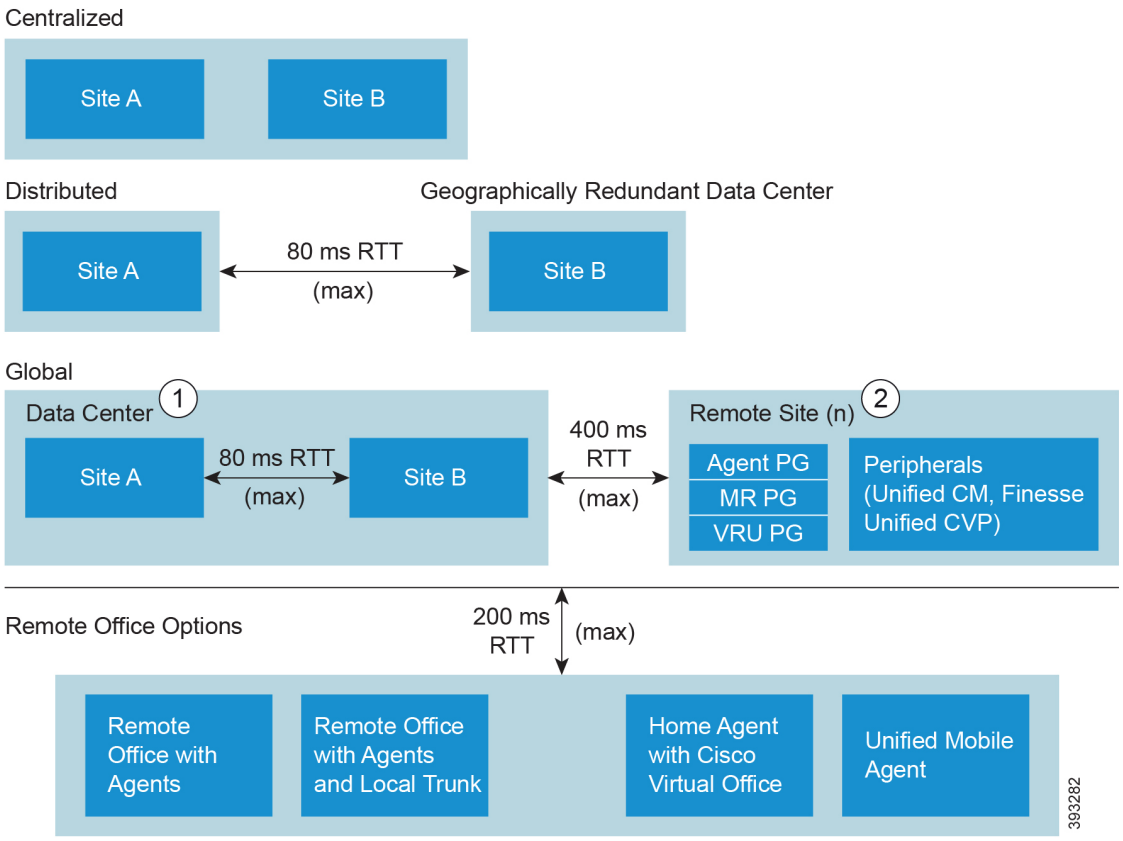


For information on the Contact Sharing feature, see the *Cisco Unified Contact Center Enterprise Features Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html.

Topologies for Reference Designs

The Contact Center Enterprise Reference Designs also define the allowed topologies for your deployment. The deployment topology consists of where you install the VMs for your data center and how your agents connect to the data center. This figure shows the basic topologies that you can use in a Reference Design.

Figure 9: Reference Design Topologies



1. The Main Site can use either a Centralized or a Distributed topology.
2. A Remote Site can be geographically colocated with the Main Site.

The Reference Designs allow the following topologies:

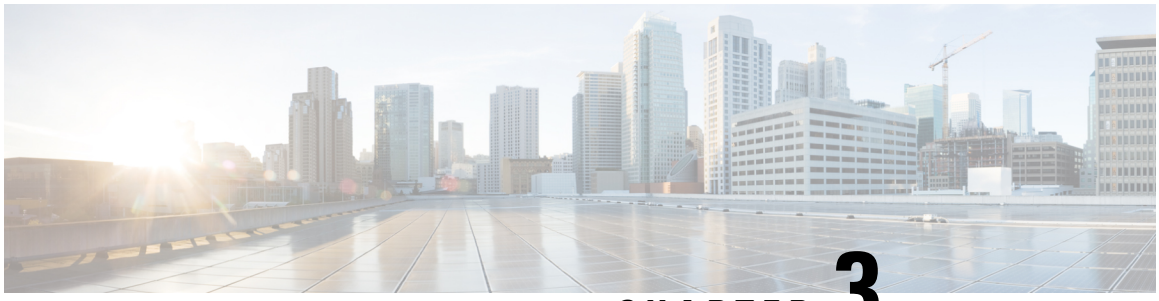
Topology	Description
Centralized	You host both sites of the redundant components in the same physical data center. Even when they are on the same LAN, the maximum round-trip time between the two sites is 80 ms. The data center includes the core contact center components and Unified CM.
Distributed	You host each site of the redundant components in a different geographical location. Distributed sites allow you to keep running on the other site if one site fails. You can also handle routing without sending a contact to a site in a different geographical region. The maximum round-trip time between the two sites is 80 ms.

Topology	Description
Global	You have a centralized or distributed main site. You also have a remote site that is generally in a different geographical location. The remote site gives you local access in that geographic region. The remote site allows you to handle your global work load without creating another contact center instance. The remote site requires a separate Unified CM cluster and a separate Cisco Finesse cluster if the RTT from the data center is greater than 80 ms. The maximum round-trip time between the main site and remote sites is 400 ms. **Note** A remote site cannot include a Cisco Unified Intelligence Center server. This topology fits the outsourcer model where the outsourcer has a separate peripheral gateway and a corresponding peripheral. **Note** Starting in Release 11.6, Packaged CCE supports this topology.

The Reference Designs allow the following methods for connecting your agents to a site:

Remote Office Topology	Description
Remote Office with Agents	A contact center office with agent workstations that connects to a site through a WAN router. The voice termination is at the site. All contacts go through the site first and then to the agents.
Remote Office with Agents and Local Trunk	A contact center office with a connection to the local PSTN. Contacts come in on the local trunk and the local gateway passes them to the data center for routing.
Home Agent with Broadband - Cisco Virtual Office (CVO)	An agent at a remote location with a VPN connection to a site. The agent has a Cisco IP Phone and a Cisco Finesse desktop. The agent can optionally use a Cisco Virtual Office (CVO) router for a permanent VPN connection.
Unified Mobile Agent	An agent who uses a PSTN phone.

Note The maximum allowed round-trip time between any remote office and the data center is 200 ms.



CHAPTER 3

Contact Center Enterprise Solutions Overview

Contact Center Solutions Architecture

Note The first four chapters of this book are for anyone who wants to get familiar with the contact center enterprise solutions:

- Packaged Contact Center Enterprise
- Cisco Hosted Collaboration Solution for Contact Center
- Unified Contact Center Enterprise

For information about design considerations and guidelines specific to Packaged CCE, see the remaining chapters.

Packaged CCE Solution Architecture

Packaged CCE is a predesigned, bounded deployment model of Unified CCE. The core components are deployed as on-box Virtual Machines (VMs) that are described by OVA files downloaded from https://www.cisco.com.

The Packaged CCE VMs provide the essential set of contact center functionality—call and non-voice task processing, prompts and rich VXML scripting, voice response collection, agent selection, queuing, and

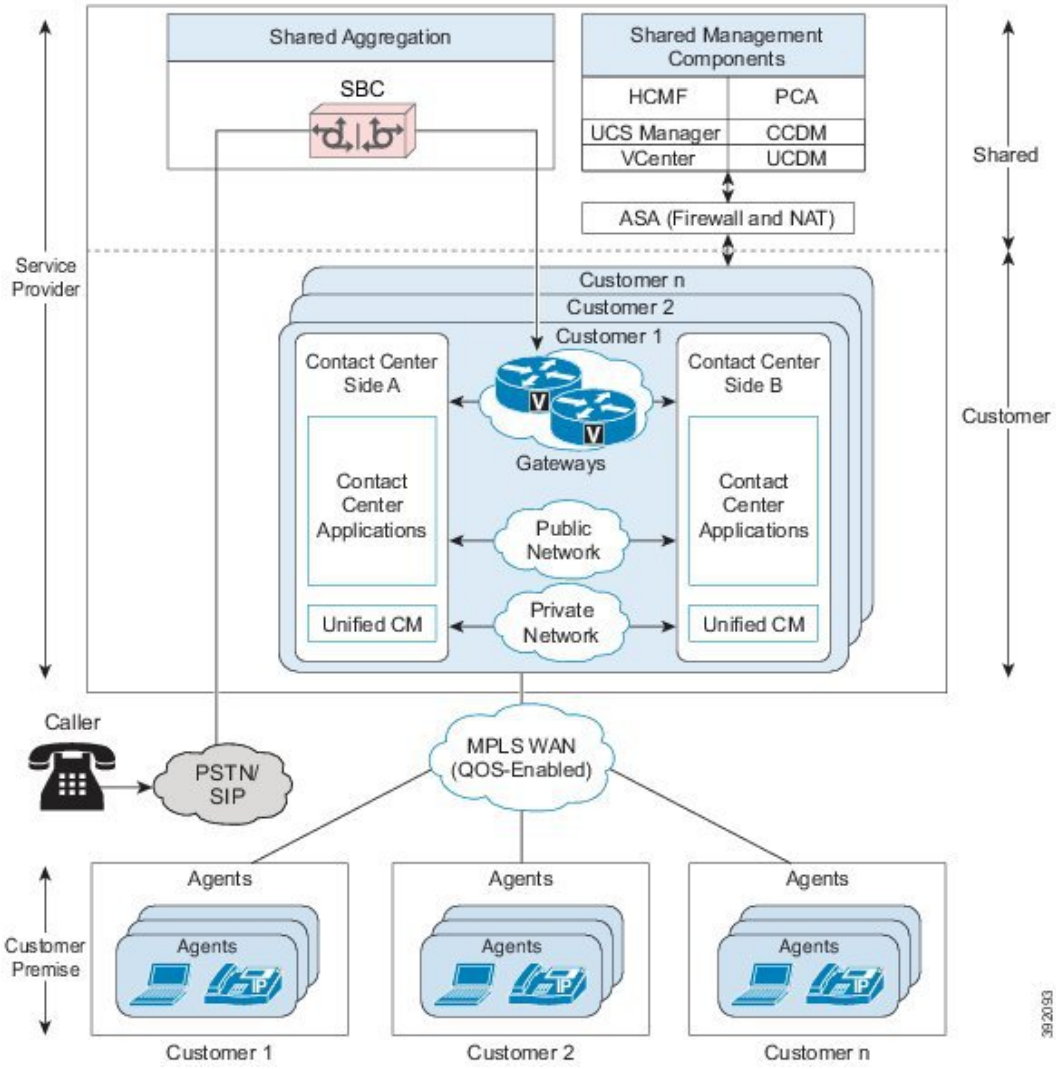
reporting. With its controlled environment and well-defined configuration and deployment boundaries, Packaged CCE is a robust solution with high availability and solution serviceability. Additional benefits are simplified ordering and deployment rollout, easier operation and maintenance, and Unified CCE Administration—a streamlined, browser-based administration interface for configuring the system and monitoring its health.

Cisco HCS for Contact Center Solution Architecture

Cisco HCS for Contact Center delivers in a hosted environment almost all of the components and features as a Unified CCE solution. Cisco HCS for Contact Center supports a subset of the Unified CCE models. You, as the service provider, manage the maintenance of the hosted environment. For your customers, this means lower hardware costs, easier and faster deployment, and no need to worry about upgrades, maintenance staff, and unpredictable costs.

Cisco HCS for Contact Center has an aggregation layer and a shared management layer. It combines Cisco Hosted Collaboration Solution components with the multiple network connections and route requests to the dedicated customer instances. The shared aggregation consists of a Hosted Collaboration Solution SBC for interfacing to a PSTN. The shared management consists of UCDM, Unified CCDM, HCM-F, Cisco Prime Collaboration Assurance (PCA), Cisco UCS Manager, VMware vCenter, and Cisco ASA (Firewall/NAT).

Figure 10: Cisco HCS for Contact Center



Unified CCE Solution Architecture

Cisco Unified Contact Center Enterprise (Unified CCE) is a solution that delivers intelligent call routing, network-to-desktop Computer Telephony Integration (CTI), and multichannel contact management to contact center agents over an IP network. Unified CCE combines software IP automatic call distribution (ACD) functionality with Cisco Unified Communications to enable companies to deploy an advanced, distributed contact center infrastructure rapidly.

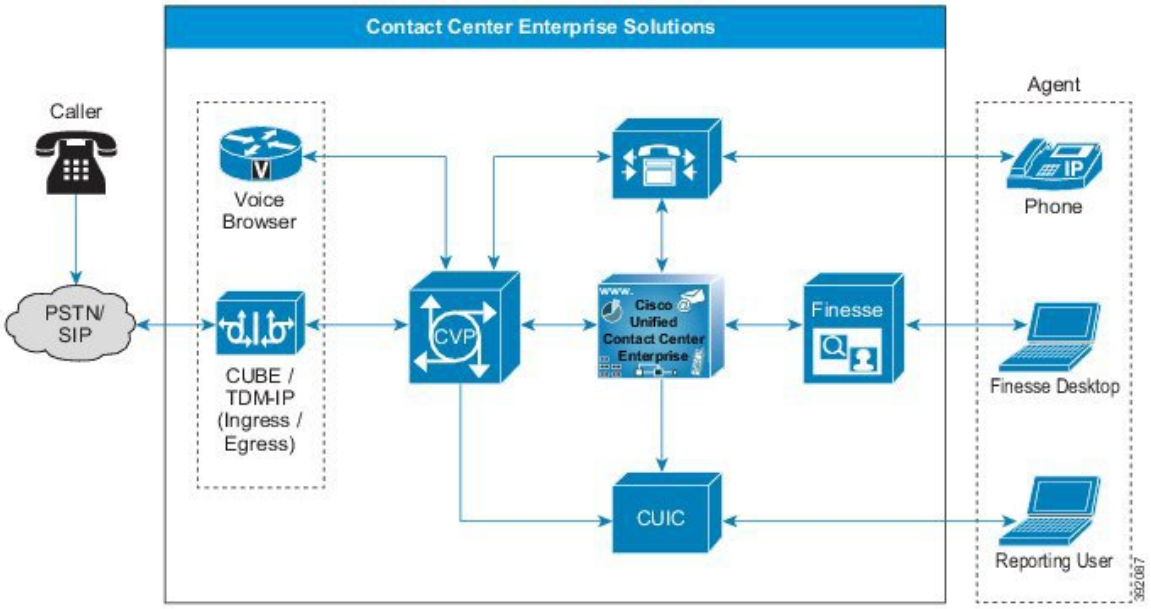
This design guide describes the deployment models and their implications including scalability, fault tolerance, and interaction between the solution components.

The Unified CCE product integrates with Cisco Unified Communications Manager, Cisco Unified Customer Voice Portal, Cisco VoIP Gateways, and Cisco Unified IP Phones. Together these products provide contact center solutions to achieve intelligent call routing, multichannel ACD functionality, voice response unit (VRU) functionality, network call queuing, and consolidated enterprise-wide reporting. Unified CCE can optionally

integrate with Cisco Unified Intelligent Contact Manager to network with legacy ACD systems while providing a smooth migration path to a converged communications platform.

The Unified CCE solution is designed for implementation in both single and multisite contact centers. Unified CCE uses your existing IP network to lower administrative expenses and to include branch offices, home agents, and knowledge workers in your contact center. The following figure illustrates a typical Unified CCE setup.

Figure 11: Typical Unified CCE Solution Deployment



The Unified CCE solution consists primarily of four Cisco software products:

- Unified Communications infrastructure—Cisco Unified Communications Manager
- Queuing and self-service—Cisco Unified Customer Voice Portal (Unified CVP)
- Contact center routing and agent management—Unified CCE. The major components are CallRouter, Logger, Peripheral Gateway, and the Administration & Data Server/Administration Client.
- Agent desktop software—Cisco Finesse

The solution is built on the Cisco IP Telephony infrastructure, which includes:

- Cisco Unified IP Phones
- Cisco Voice Gateways
- Cisco LAN/WAN infrastructure

Core Components

Requests coming into a contact center enterprise solution usually interact with the core components in the following order:

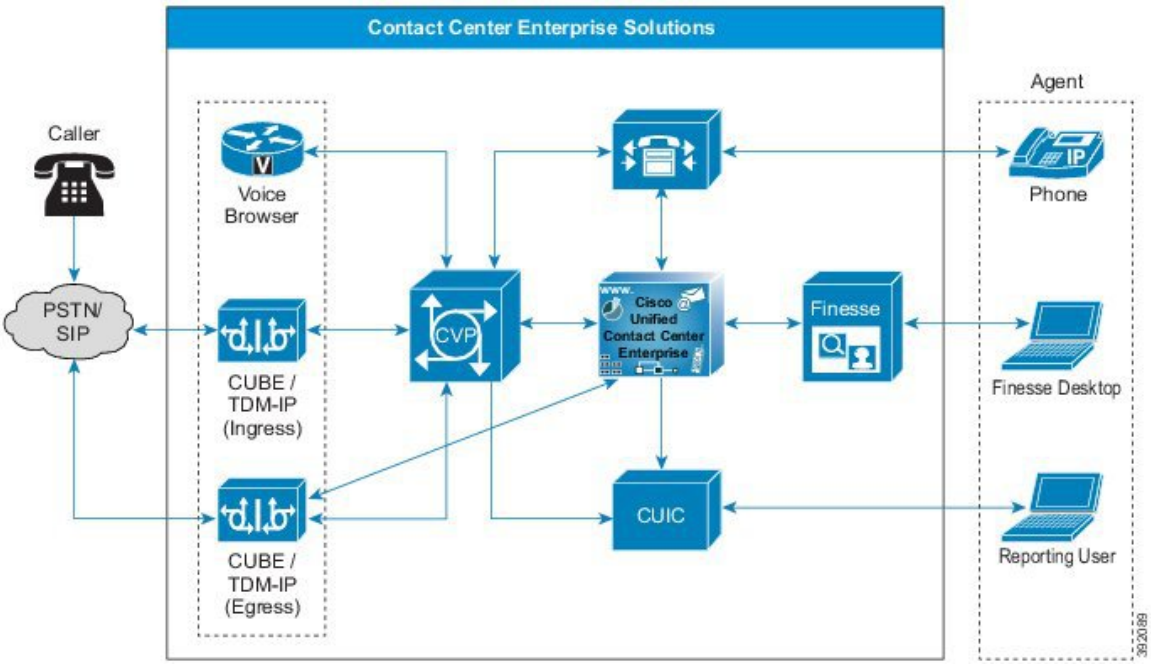
1. Cisco Ingress, Egress, and VXML Gateways
2. Cisco Unified Customer Voice Portal
3. Cisco Unified Contact Center Enterprise
4. Cisco Virtualized Voice Browser
5. Cisco Unified Communications Manager
6. Cisco Finesse
7. Cisco Unified Intelligence Center

Ingress, Egress, and VXML Gateways

You can use these gateways in your solution:

- Cisco Voice TDM gateway
- Cisco Unified Border Element
- Cisco VXML gateway

Figure 12: Ingress, Egress, and VXML Gateways



Note *Voice Browser* refers to either VXML Gateways or Cisco Virtualized Voice Browser (VVB).

TDM gateways and CUBE gateways can act as both ingress (for inbound calls) and egress gateway (for outbound calls) in a specific deployment.

These types of gateways can be colocated or exist on separate physical gateways.

Cisco IOS-XE does not support VXML gateway functionality.

Cisco TDM Voice Gateway

The Cisco Ingress Voice Gateway is the point at which an incoming call enters the contact center enterprise solution. It terminates time division multiplexing (TDM) calls on one side and implements VoIP on the other side. It serves as a pivot point for the extension of calls from the TDM environment to VoIP endpoints. This conserves WAN bandwidth because no hairpinning of the media stream occurs. The Cisco Ingress Voice Gateway also provides for call switching capabilities at the command of other contact center enterprise solution components.

You can use the Ingress Voice Gateway for the PSTN Voice Gateway. The Ingress Voice Gateway converts TDM speech to IP and converts DTMF digits to RFC2833 events.

Note Unified CVP does not support passing SIP-Notify DTMF events.

You can separate the VXML functionality from the Ingress Voice Gateway to provide a separate PSTN ingress layer. The separate PSTN layer and VXML enable the deployment to support many VXML sessions and PSTN interfaces. An ingress gateway that handles numerous ingress calls cannot also support that many VXML sessions. In such cases, you can off-load the VXML sessions to a separate farm of Voice Browsers, such as Cisco VVB.

Note You can use any TDM interface that your Cisco IOS gateway, IOS version, and the contact center enterprise components all support.

The Cisco Egress Voice Gateway is used only when calls are extended to TDM networks or equipment. For example, transferring a call to a PSTN or a TDM automatic call distributor (ACD). While the Real-time Transport Protocol (RTP) stream runs between the gateway ports, the signaling stream logically goes through the Unified CVP Server and Cisco Unified CCE. This allows subsequent call control (such as transfers).

Both TDM Ingress Gateways and Egress Gateways support Session Initiation Protocol (SIP).

Cisco Unified Border Element

The Cisco Unified Border Element (CUBE) is a Cisco router that runs as a Session Border Controller (SBC). SBCs interconnect independent Voice over IP (VoIP) and video over IP enterprise networks for data, voice, and video transport. SBCs are critical components for scaling networks from VoIP islands within a single customer network to an end-to-end IP community. SBCs are used both inside an enterprise and to communicate beyond an enterprise across service provider networks.

Note When this guide refers to CUBE, we always mean the Enterprise version, not the Service Provider version.

CUBE runs on Cisco Integrated Services Router (ISR) and Aggregation Service Router (ASR) routers. The Cisco Cloud Services Router (CSR) can run a virtual CUBE.

CUBE adds the following features to the Cisco IOS and IOS XE software image:

- A Network-to-Network Interface point for billing, security, call admission control, quality of service, and signaling interworking
- The feature set necessary to support the transition to SIP trunking
- The capability to act as a distinct demarcation point between two networks.
- The capability to intelligently allow or disallow real-time traffic between networks.

The use of third-party SIP trunks with contact center enterprise solutions is supported by using CUBE. CUBE performs the role of session border controller for SIP normalization and interoperability.

Virtual CUBE for Contact Center Solutions

In compatible Cisco IOS XE releases, Contact Center Enterprise (CCE) solutions support CUBE as a virtualized form factor. You can install virtual CUBE (vCUBE) on VMware ESXi hypervisors. CCE supports vCUBE in the following configurations:

Number of vCPUs	Memory Reservation	Concurrent WebSocket Forking Sessions
1	4 GB RAM	500
2	4 GB RAM	600
4	8 GB RAM	1000

For more details on CUBE sizing, see the Licensing Options section in the *Cisco Unified Border Element Version 14 Data Sheet* at https://www.cisco.com/c/en/us/products/collateral/unified-communications/unified-border-element/data-sheet-c78-729692.html

vCUBE supports most of the features available in CUBE. It supports Outbound Option without CPA. Features that manage the media plane do not work in the Cisco Cloud Services Router (CSR) router. vCUBE does not support the following Digital Signal Processor (DSP) features:

- Audio and Video Codec Transcoding or Transrating
- DTMF interworking
- Call Progress Analysis (CPA)
- Noise Reduction (NR), Acoustic Shock Protection (ASP), and Audio Gain
- IOS-based hardware MTP
- A mix of G.729 and G.711 during conferencing
- DSP high availability
- High availability protected mode (instances on the same host)

Note You can use multicodec, software conferencing, and MTP that are controlled by Unified CM instead of the DSP available in physical CUBEs. You can add a dedicated physical gateway if your solution requires CPA or mixed codecs for conferencing.

- Limited support for Voice Class Codec (VCC). The codec supported on peer leg is included in offer. Other codecs are filtered out.

For more details on support for vCUBE, see the vCUBE section in the *Cisco Unified Border Element Configuration Guide* at http://www.cisco.com/c/en/us/support/routers/cloud-services-router-1000v-series/products-installation-and-configuration-guides-list.html and the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html.

Cisco VXML Gateway

Centralized deployment models often include VXML Gateways. The VXML Gateway interprets VXML pages from the VXML Server.

Note The term *Voice Browser* can mean either a VXML Gateway or Cisco Virtualized Voice Browser (Cisco VVB).

You can cache audio prompts from a third-party media server in a VXML Gateway to reduce WAN bandwidth and prevent poor voice quality. The VXML document provides either a pointer to the location of the audio file or the address of a text-to-speech (TTS) Server to stream the audio. The VXML Gateway interacts with automatic speech recognition (ASR) and TTS Servers through Media Resource Control Protocol (MRCP).

You can deploy a Cisco IOS VXML Gateway on the same router as you deploy a Unified CVP Ingress Voice Gateway. This model is suitable for deployments with small branch offices. The Cisco IOS VXML Gateway can also run on a separate router platform. This model is suitable for deployments with large or multiple voice gateways, where only a small percentage of the traffic is for Unified CCE. This model allows shared public switched telephone network (PSTN) trunks between office users and contact center agents, and call routing based on the dialed number. VXML Gateway can store audio files on flash memory or on a third-party media server.

Unless a Cisco IOS VXML Gateway is combined with an Ingress Voice Gateway, the Cisco IOS VXML Gateway does not require TDM hardware. It interacts with VoIP on one side, and HTTP (carrying VXML or .wav files) and MRCP (carrying ASR and TTS traffic) on the other. As with Ingress Voice Gateways, Cisco IOS VXML Gateways are often deployed in farms for Centralized deployment models, or one in each office in Branch deployments.

As an alternative, you can deploy Cisco VVB on a separate virtual machine. This model is suitable for both standalone and comprehensive deployments. Cisco VVB communicates with ASR/TTS using MRCP.

Note Cisco IOS-XE does not have built-in voice browser capability. Therefore, deploying an IOS-XE ingress gateway with Unified CVP requires the use of a separate ISR G2 gateway or Cisco VVB to provide the voice browser.

Cisco Unified Customer Voice Portal

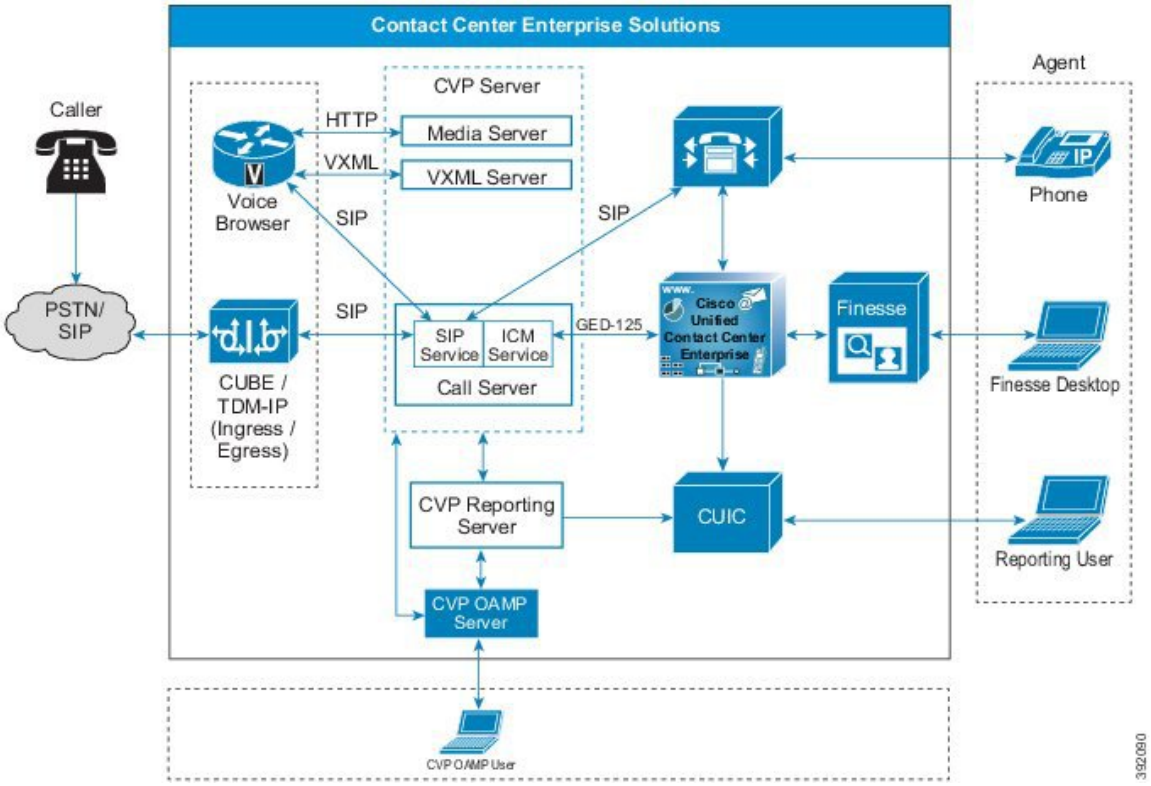
Cisco Unified Customer Voice Portal combines open-standards support for speech with intelligent application development and industry-best call control.

Unified Customer Voice Portal (Unified CVP) is a software application that runs on Cisco Unified Computing System (UCS) hardware or specification-based equivalents. Unified CVP provides prompting, collecting,

queuing, and call control services using standard web-based technologies. Its architecture is distributed, fault tolerant, and highly scalable. With CVP, voice terminates on Cisco Voice Browsers that interact with the Unified CVP application server using HTTP(S) (speech) and SIP (call control). Unified CVP includes the following subcomponents:

- CVP Call Server
- CVP VXML Server
- CVP Media Server
- CVP Reporting Server

Figure 13: Unified CVP in a Contact Center Enterprise Solution



The Unified CVP software tightly integrates with the Unified CCE software for application control. Unified CVP interacts with Unified CCE using the Voice Response Unit (VRU) Peripheral Gateway Interface. The Unified CCE scripting environment controls the initiation of building-block functions such as play media, play data, menu, and collect information. The Unified CCE script can invoke external VXML applications to be run by the CVP VXML Server.

The CVP Call Studio is an Eclipse-based IDE for developing VRU applications. The VXML Server is the application server which hosts those VRU applications. The VXML Server handles sophisticated, high-volume VRU applications. It can also interact with custom or third-party J2EE-based services. You can achieve load balancing with an optional CUSP server or the built-in SIP Server Group in CVP.

Unified CVP can support multiple grammars for prerecorded announcements in several languages. CVP can optionally provide automatic speech recognition and text-to-speech capability. CVP can also access customer databases and applications through the Unified CCE software.

Unified CVP also provides a queuing platform for the Unified CCE solution. Voice and video calls can remain queued on CVP until they are routed to a contact center agent (or external system). The system can play back music or videos while the caller is on hold. When Unified CCE routes the call to an agent, the agent can send videos to a caller from the agent desktop application.

CVP Call Server

The Call Server component provides the following independent services, which all run on the same Windows server:

- **SIP service**—This service communicates with the contact center enterprise solution components such as the SIP Proxy Server, Ingress Gateway, Unified CM SIP trunks, and SIP phones. The SIP service implements a Back-to-Back User Agent (B2BUA). This B2BUA accepts SIP invites from ingress voice gateways and typically directs those calls to an available Voice Browser port. After completing call setup, the Unified CVP B2BUA acts as an active intermediary for any subsequent call control. While the Unified CVP SIP signaling is routed through this service, this service does not touch the RTP traffic. Integrated into this B2BUA is the ability to interact with the Unified CCE through the ICM Service. This integration provides the ability for the SIP Service to query the Unified CCE for routing instruction and service control. This integration also allows Unified CCE to begin subsequent call control to do things such as transfers.
- **ICM service**—This service is responsible for all communication between Unified CVP components and Unified CCE. It sends and receives messages on behalf of the SIP Service and the IVR Service.

Note The IVR service is now part of the VXML Server.

CVP VXML Server

The VXML Server runs advanced VRU applications by exchanging VXML pages with the Voice Browser. Like almost all other Unified CVP product components, it runs within a Java 2 Enterprise Edition (J2EE) application server environment. Many customers add their own custom-built J2EE components to interact with back-end hosts and services. The VXML Server applications are written using Cisco Unified Call Studio and are deployed to the VXML Server for initiation of tasks. The applications are invoked on an as needed basis by a special Micro application which must be run from within the Unified CCE routing script.

The VXML Server can also be deployed in a standalone configuration that does not include any Unified CCE components. Applications are invoked as a direct result of calls arriving in the Voice Browser, and a single post application transfer is allowed.

Note The IVR service is now part of the VXML Server. So, it now uses a VXML server port license to run microapplication. In previous releases, the IVR service was part of the Call Server.

The IVR service creates VXML pages that implement the Unified CVP Micro-applications based on Run External Script instructions received from Unified CCE. The IVR Service functions as the VRU leg (in Unified CCE terminology). You transfer calls to it from the SIP Service to run Micro-applications. The VXML pages that this module creates are sent to the Voice Browser. The IVR service is also responsible for the conversion of Unified CVP Micro-applications to VXML pages, and the reverse.

CVP Media Server

The Media Server component is simply a web server which provides prerecorded audio files, external VXML documents, or external Automatic Speech Recognition (ASR) grammars to the gateway. Some of these files can be stored in local flash memory on the gateways. However, in practice, most installations use a centralized media server to simplify distribution of prerecorded customer prompt updates. Media Server functionality can also include a caching engine. The gateways themselves, however, can also do prompt caching when configured for caching.

Note The Media Server component in Unified CVP is installed by default, along with Unified CVP Call Server and Unified CVP VXML Server.

Media Servers can be deployed as a simplex operation, as a redundant pair, or with supported load balancers in a farm. The Voice Browser caches.wav files it retrieves from the Media Server. In most deployments, the Media Server encounters low traffic from Unified CVP.

CVP Reporting Server

The Unified CVP Reporting Server provides consolidated historical reporting for a distributed self-service deployment. The CVP Reporting server is optional, unless your solution requires it for Courtesy Callback, trunk group reporting, and VRU reporting.

The CVP Reporting Server runs on a Windows server that hosts an IBM Informix Dynamic Server (IDS) database management system. The database schema is preset, but you can develop custom reports through Unified Intelligence Center and other reporting solutions.

The Reporting Server should be local to the Call Servers and VXML Servers. Deploying the Reporting Server at a remote location across the WAN is supported if the latency is less than 80ms RTT between the CVP Reporting Server and the CVP Call Server that it serves for VXML reporting traffic. This assumes the WAN bandwidth is not a constraint. If you have Remote Site deployment with local CVP Call Server, then you need to have local CVP reporting server at the Remote Site. However, between Remote Sites, you can have the CVP reporting server across WAN serving the CVP Call Server at the other Remote Site if the latency between the Remote Sites is less than 80 ms RTT.

The Reporting Server receives reporting data from the SIP Service (if used), and the IVR Service of the VXML Server. The Reporting Server depends on the Call Server to receive call records.

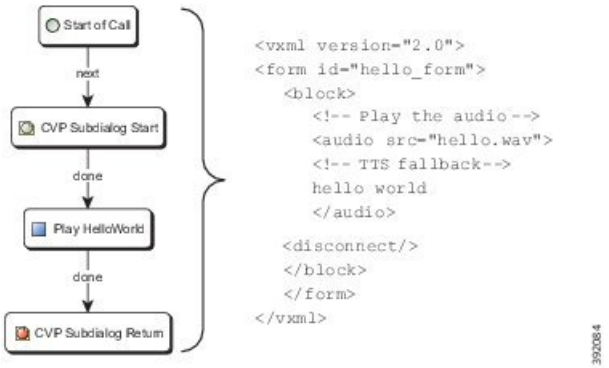
The Reporting Server does not perform database administrative and maintenance activities, such as backups or purging.

CVP Call Studio

Cisco Unified Call Studio is the service creation environment (script editor) for Unified CVP VXML Server applications. It is based on the open source Eclipse framework, which provides an advanced drag-and-drop graphical editing feature. Call Studio also provides options to insert vendor-supplied and custom-developed plug-ins that enable applications to interact with other services in the network. Call Studio basically is an offline tool. The only interaction with the Unified CVP VXML Server is to deliver compiled applications and plugged-in components to be run.

Call Studio provides an environment where you concentrate on your business logic. The tool handles the details of turning the logic into XML.

Figure 14: Call Studio Generates the Code for You



The Call Studio license is associated with the MAC address of the machine on which it is running. You typically designate one or more servers for that purpose. Cisco Unified Call Studio runs on a virtual machine or a Windows PC.

CVP Infrastructure

Unified CVP infrastructure includes the Web Services Manager, a services layer that supports a Diagnostic Portal API.

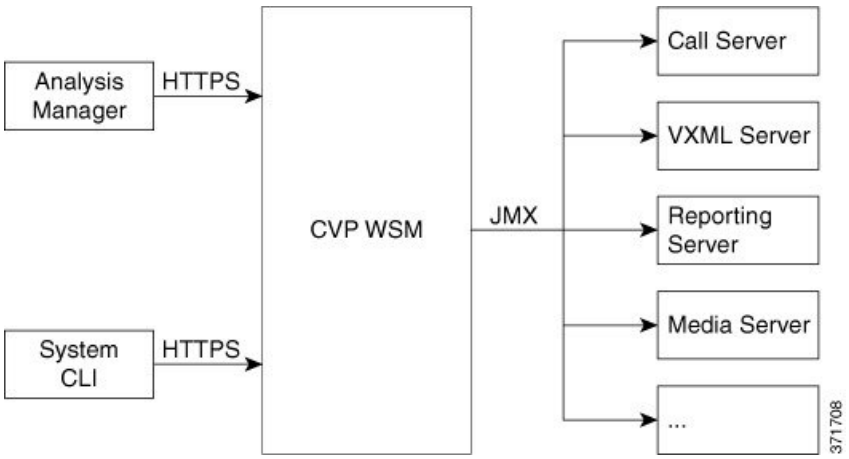
Unified CVP Infrastructure supports the following features:

- Diagnostic Portal API service support by the Web Services Manager.
- Unified System Command Line Interface (CLI) which is a client tool that supports the diagnostic portal API and other APIs for collecting diagnostic data.
- Licensing:
  - Common Licensing for all CVP components that support FlexLM.
  - Licenses are only valid if the license feature, `CVP_SOFTWARE`, is added. This feature is used to ensure if you are authorized to run the current version of CVP.
- Serviceability Across Products with enhanced Log and Trace messages.

The CVP WebServices Manager (WSM) is a component that is installed automatically on all Unified CVP Servers, including Remote Operations Manager (ROM)-only installations. WSM interacts with various subsystems and infrastructure handlers, consolidates the response, and publishes an XML response. WSM supports secure authentication and data encryption on each of the interfaces.

The following figure shows how the two interfaces interact with the Web Services Management (WSM) to provide information about Unified CVP components.

Figure 15: Typical Use of the Web Services Layer

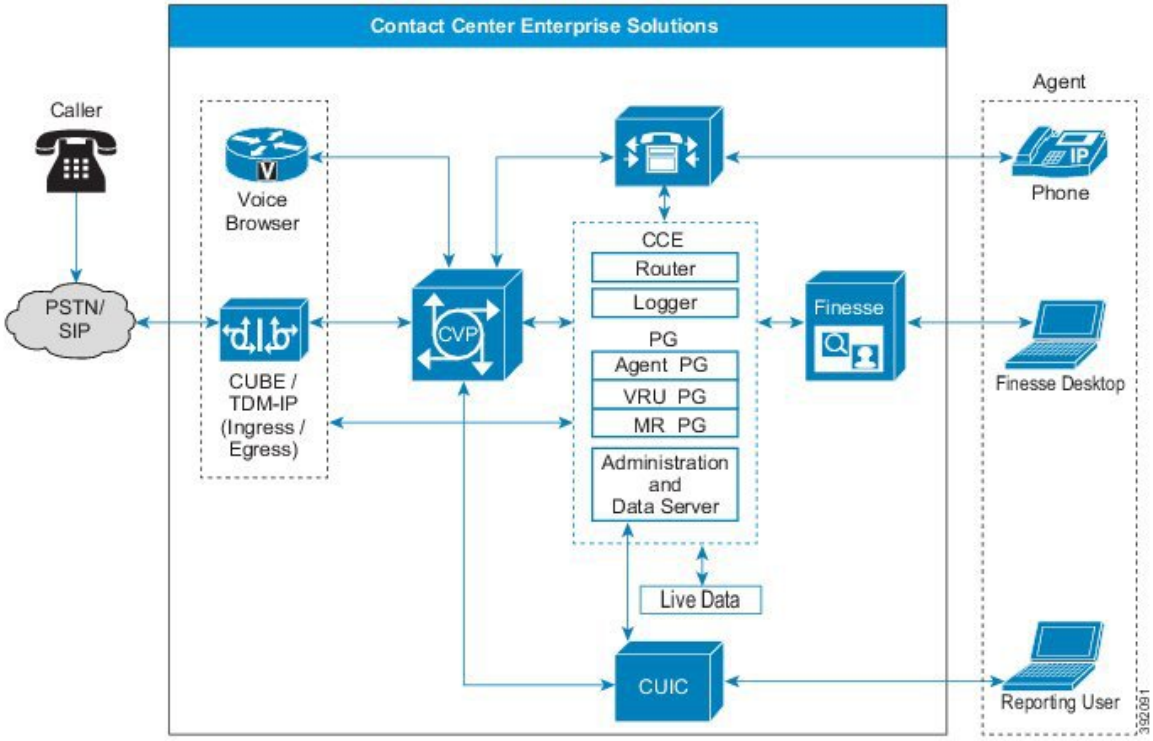


Contact Center Enterprise

Unified Contact Center Enterprise (Unified CCE) provides these contact center features:

- Agent state management
- Agent selection
- Call and task routing and queue control
- VRU interface
- CTI Desktop screen pops
- Contact center reporting data

Figure 16: Unified CCE in a Contact Center Enterprise Solution



Unified CCE runs in VMs on Cisco Unified Computing System servers or exact equivalents. This table lists the major components of Unified CCE:

Table 18: Unified CCE Core Components

Unified CCE Software Components	Description
CallRouter (Router)	Makes all routing decisions on how to route a call or customer contact. The Router is a part of the Central Controller.
Logger	The database server that stores contact center configuration data. The Logger also temporarily stores historical reporting data for distribution to the data servers. The Logger is a part of the Central Controller.

Unified CCE Software Components	Description
Peripheral Gateway (PG)	Interfaces to peripheral devices, like the Unified Communications Manager, VRU (Unified CVP), or multichannel products (Enterprise Chat and Email or third-party multichannel applications that use the Task Routing APIs). The standard layout for contact center enterprise solutions has the Agent PG, VRU PG, and MR PG coresident on a single VM. Each PG includes one or more Peripheral Interface Managers (PIMs) for the specific device interfaces. **Important** Your contact center enterprise solution can only use the new higher configuration limits with the standard three coresident PG layout.
Administration & Data Server	Provides the configuration interface and real-time and historical data storage. You can deploy this component in several configurations.
Live Data Server	Processes events from the Router and PGs for Unified CCE Live Data reports.

Terminology for Unified CCE Subcomponents

Combinations of these Unified CCE subcomponents are sometimes called by the following names:

Name	Description
CCE Central Controller	Router and Logger
CCE Rogger	Router and Logger running on same VM
CCE Call Server	Router and PG
CCE Data Server	Logger and AW

Unified CCE

Use Unified UCCE for advanced call control, such as IP switching and transfers to agents. Both provide call center agent-management capabilities and call scripting capabilities. Scripts running in either environment can access Unified CVP applications.

Unified CCE

Unified CCE is the standard version that most solutions use. In these solutions, Unified CCE selects the agent who handles the call. Unified CM acts as the ACD.

Router

The Router is the brain of Unified CCE. When a call or task arrives, it triggers a routing script that decides what happens to the contact. The Router directs contacts from one place to another based on the script's outcome and selects the agent to handle the contact. Routers work in redundant pairs, referred to as Side A

and Side B. Both sides are active. These separate, distributed instances use the Message Delivery Subsystem (MDS) to keep in lock-step with each other. Both sides share all data and control messaging so that both sides have the same data for routing decisions. The redundant deployment ensures that the system can operate even when one side fails. The opposite side continues routing contacts during an outage.

Logger

Unified CCE uses the Logger to store historical data and configuration data about the call center. The Logger collects the historical data and then distributes it later. Like the Router, you deploy the Logger as a redundant pair. Each side of the Logger only receives messages from the corresponding Router. For example, the Side A Router only sends messages to the Side A Logger. Because the routers run in lock-step, the Loggers on both sides receive the same messages during usual operation. After any outage, the Loggers resynchronize their data through the Routers. The Logger distributes historical data to the Historical Data Server (HDS). The Logger also distributes configuration and real time data to the Administration & Data Servers through Message Delivry Subsystem (MDS).

Depending on your solution, the Logger is on the same VM with the Router (a Rogger model) or on a separate VM (a Router/Logger model).

Peripheral Gateway

The peripheral gateway (PG) handles communication with telephony and multi-media devices through their CTI interfaces. PGs can communicate with ACDs, VRU devices, or IP PBXs. The PG normalizes the protocol of the assorted devices. The PG tracks the state of agents and calls that are on each device. The PG sends this status to the Router and forwards requests that require customer logic to the Router. A PG can include the following processes:

- Peripheral Interface Managers (PIMs)
- Computer Telephony Integration (CTI)
- Java Telephony API (JTAPI)

In the standard layout for the Contact Center Enterprise Reference Designs, the Agent PG, VRU PG, and MR PG are coresident on a single VM. The PIMs handle the protocol normalization. The PIMs communicate to the peripheral and translate the peripheral proprietary language into one that Unified CCE understands. The CTI Gateway (CG - CTI Server component) is also coresident with the PG.

Important Your contact center enterprise solution can only use the new higher configuration limits with the standard three coresident PG layout.

Unified CCE supports several types of PGs:

- Agent PG—Connects to Unified Communications Manager (Unified CM)
- Voice Response Unit (VRU) PG—Connects to CVP
- Media Resource (MR) PG—Connects to multimedia components, like Enterprise Chat and Email or Customer Collaboration Platform

As with the other Unified CCE core components, you deploy PGs in redundant pairs.

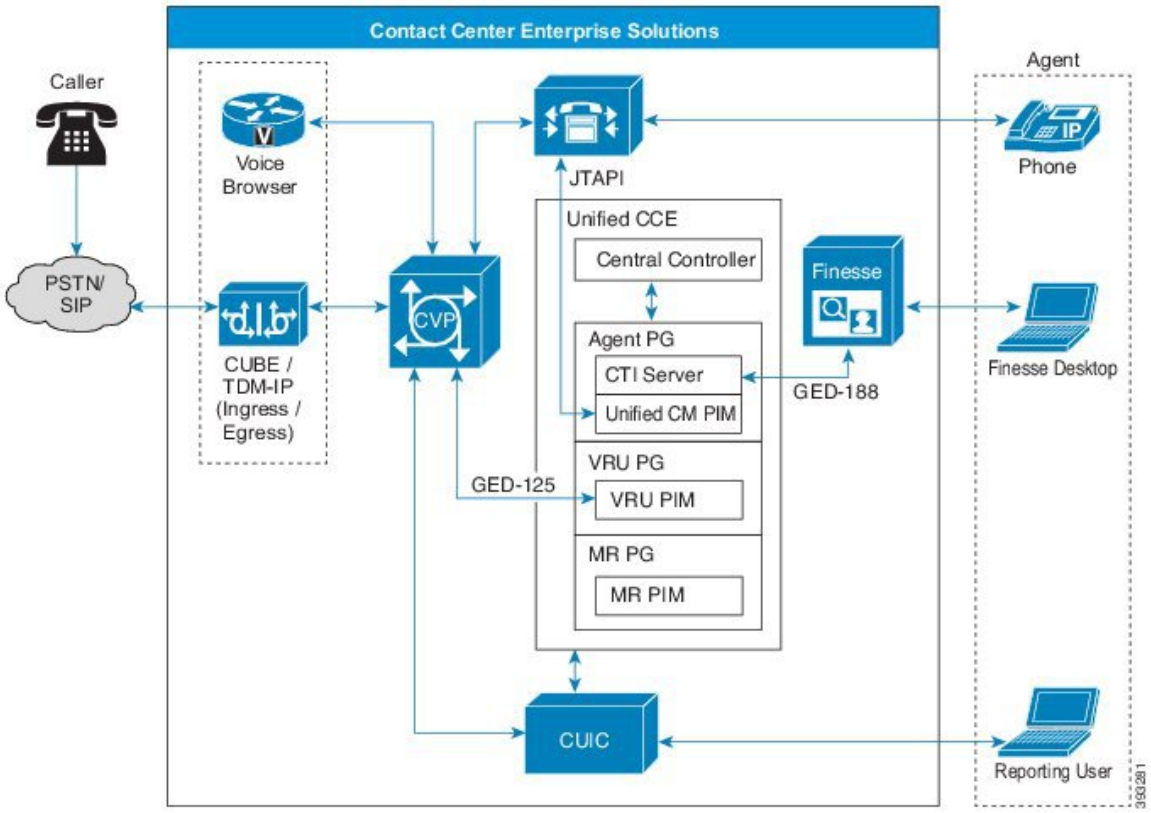
One class of PG talks to an ACD or a Unified CM that has agents on it. These PGs use a proprietary CTI protocol to the switch, and maintain the state of agents and calls in queue on the device. Another class of PG exposes client-neutral interfaces. The VRU PG exposes an interface that is tailored to voice calls. The MR PG exposes an interface for more generic task routing.

Unified CCE treats the VRU and Unified CM as separate peripherals. This separation provides flexibility. You can load balance between several VRUs.

Larger, multisite (multicluster) deployments include many Agent PGs. In these deployments, Unified CCE tracks all the agents and calls centrally. Unified CCE can route calls to the most appropriate agent, independent of the site or cluster that they use. This coordination makes a logical enterprise-wide contact center with one enterprise-wide queue.

The following figure shows the communications between the PG and the other solution components.

Figure 17: Communications Between the PG and the Other Components



Peripheral Interface Managers

For each Unified CM cluster, there is a Unified CM PIM on an Agent PG. Each redundant Agent PG pair can support a maximum of 2000 agents. For scalability, some deployments require multiple PIMs for the same cluster. Deploy each PIM on a different Agent PG. Deploy only one Agent PG on each VM.

CTI Server

Each Agent PG includes a CTI server. The CTI Server handles call control and agent requests from the agent desktops. On the Agent PG, CTI services connect to one side or the other, depending on which side is active. The CTI Server processes agent state requests and updates the Central Controller for consideration in routing

decisions. The PG forwards call control requests to the Unified CM, which monitors and controls the phone endpoints. The CTI Server keeps the agent desktop synchronized with the agent's IP phone state.

JTAPI Communications

The Unified CM PIM sign-in process establishes JTAPI communications between the Unified CM cluster and the application. The CTI Manager communicates through JTAPI to Unified CCE. Every subscriber within a cluster runs a CTI Manager instance. But, the Unified CM PIM on the PG communicates with only one CTI Manager (and thus one node) in the cluster. That connected CTI Manager passes CTI messages for the other nodes within the cluster. Each redundant pair of PGs shares a unique JTAPI user ID. The user ID is how the CTI Manager tracks the different applications.

For example, subscriber 1 connects to a Voice Gateway (VG) and subscriber 2 communicates with Unified CCE through the CTI Manager. When a call arrives at the VG, subscriber 1 sends an intra-cluster message to subscriber 2. Subscriber 2 sends a route request to Unified CCE to determine how to route the call.

The JTAPI communications between the cluster and Unified CCE include three distinct types of messaging:

- **Routing control**—Messages that enable the cluster to request routing instructions from Unified CCE.
- **Device and call monitoring**—Messages that enable the cluster to notify Unified CCE about state changes of a device (phone) or a call.
- **Device and call control**—Messages that enable the cluster to receive instructions from Unified CCE on how to control a device (phone) or a call.

Most calls use all three types of JTAPI communications within a few seconds. When a new call arrives, Unified CM requests routing instructions from Unified CCE. When a subscriber receives the routing response from Unified CCE, the subscriber sends the call to an agent phone. The subscriber notifies Unified CCE that the phone is ringing. That notification enables the answer button on the agent desktop. When the agent clicks the answer button, Unified CCE instructs the subscriber to make the phone go off-hook and answer the call.

In order for the routing control communication to occur, the subscriber needs a CTI Route Point. You associate a CTI Route Point with a specific JTAPI user ID. Through this association, the subscriber knows which application provides routing control for that CTI Route Point. Dialed Numbers (DNs) are then associated with the CTI Route Point. Then, the subscriber can generate a route request to Unified CCE when a new call to that DN arrives.

Note You cannot use the DN for a CTI Route Point on a different CTI Route Point in another partition. Ensure that DNs are unique across all CTI Route Points on all partitions.

Administration & Data Server

The Administration & Data Server is the main interface to the Unified CCE configuration. The Administration & Data Server includes a database with a copy of the configuration information from the Logger. The Administration & Data Server receives updates from the central controller to keep the database in sync. Clients can read the configuration from the database and send updates through the Central Controller. The main clients in the Administration & Data Server are the GUI configuration tools.

In production systems, install each Administration & Data Server on a separate VM from the Router and Logger to ensure no interruptions in the real-time call processing. In contact center enterprise lab systems, you can install the Administration & Data Server on the same VM as the Router and Logger.

For information about data storage in virtualized deployments, see the *Virtualization for Unified Contact Center Enterprise* at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-unified-contact-center-enterprise.html.

You can deploy the Administration & Data Server in a combination of roles to achieve the proper scalability for your deployment:

- Administration Server and Real-Time Data Server (AW)
- Administration Server and Historical Data Server (AW-HDS)
- Administration Server, Historical Data Server, and Detail Data Server (AW-HDS-DDS)
- Historical Data Server and Detail Data Server (HDS-DDS)

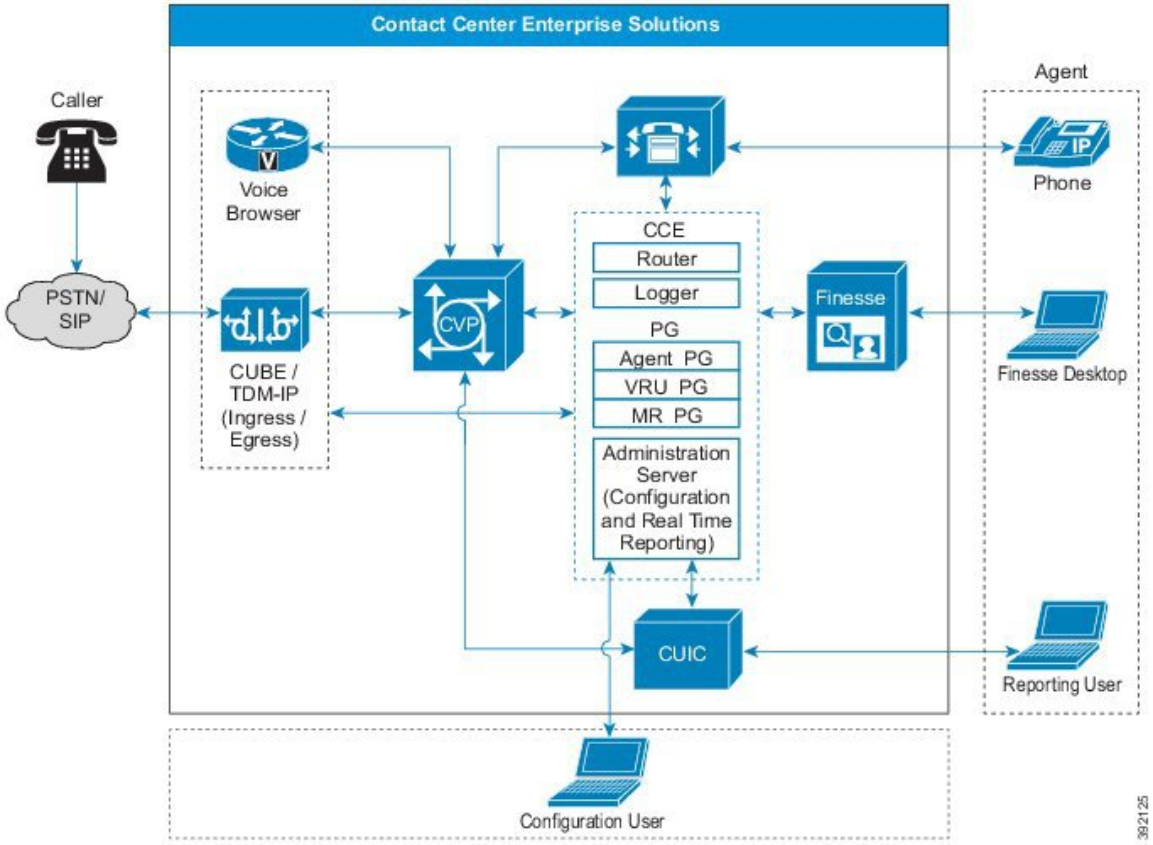
You do not deploy the Administration & Data Server in redundant pairs like the other core components. Instead, you deploy one Administration & Data Server for each Logger. If one Administration & Data Server fails, you can sign in your client AW to another server.

The AW acts as the authentication server for Cisco Finesse. In a Cisco Finesse deployment, the AW is mandatory and must run in high-availability mode (both a primary and backup AW).

Administration Server and Real-Time Data Server (AW)

This server handles configuration changes and real-time reporting with Cisco Unified Intelligent Center (Reporting client). The Real-Time Data Server portion of the AW uses the AW database to store real-time data and configuration data. Real-time reports combine these two types of data to present a near-current snapshot of the system. This role does not support historical reporting. System administrators generally use AWs to control access to what a configuration user can configure.

Figure 18: Configuration and Real-Time Reporting AW



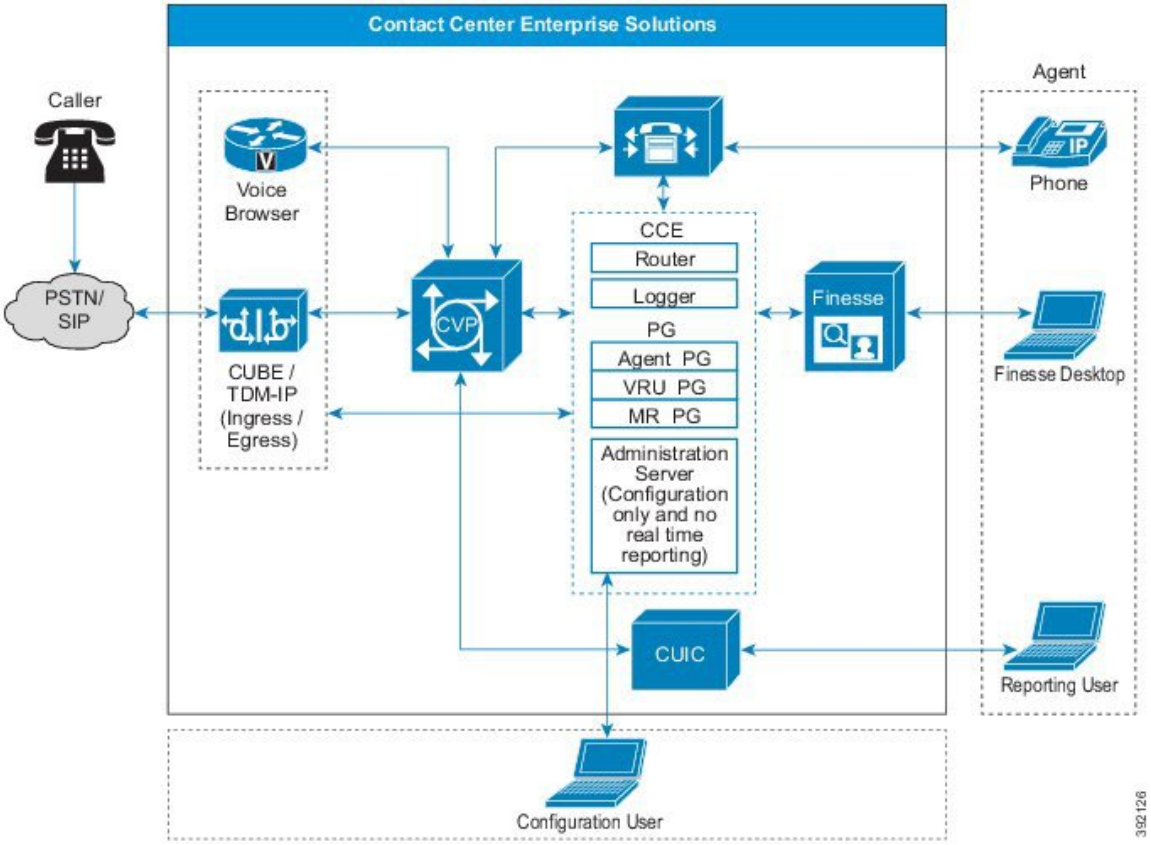
You can deploy an AW to handle only configuration tasks for scalability in these models:

- Configuration-Only Administration Server
- Administration Client (formerly called a *client AW*)

For these configuration-only models, real-time reporting is turned off.

This deployment role allows Unified CCMP to configure a specific Unified CCE Customer Instance. The load is low enough on such a lightweight Administration & Data Server that a single server is sufficient.

Figure 19: Configuration-Only AW



Configuration Only Administration Servers are the same as AWs, but without the real-time data. As such, Administration Clients cannot connect to them and they cannot display real-time data in Script Editor.

An Administration Client (formerly known as a *client AW*) serves the administration role but is deployed as a client to an Administration Server for scalability. The Administration Client can view and modify the configuration and receive real-time reporting data from the AW. But, it does not store the data itself and does not have a database.

The AW supports configuration tools for such tasks as creating agents, skill groups, precision queues, and routing scripts.

The primary AW communicates directly with the Central Controller for configuration data. You can set up secondary AWs to provide scaling for real-time reporting. During usual operation, the secondary AW connects to the primary AW for the data. If the primary AW fails, the secondary AW connects to the Central Controller.

You can deploy AWs coresident with the Central Controller or remotely. You can deploy the primary and secondary AWs together or separately.

If you use Administration Clients, you can deploy and connect multiple Administration Clients to either the primary or the secondary AWs. But, deploy them geographically local to their AW.

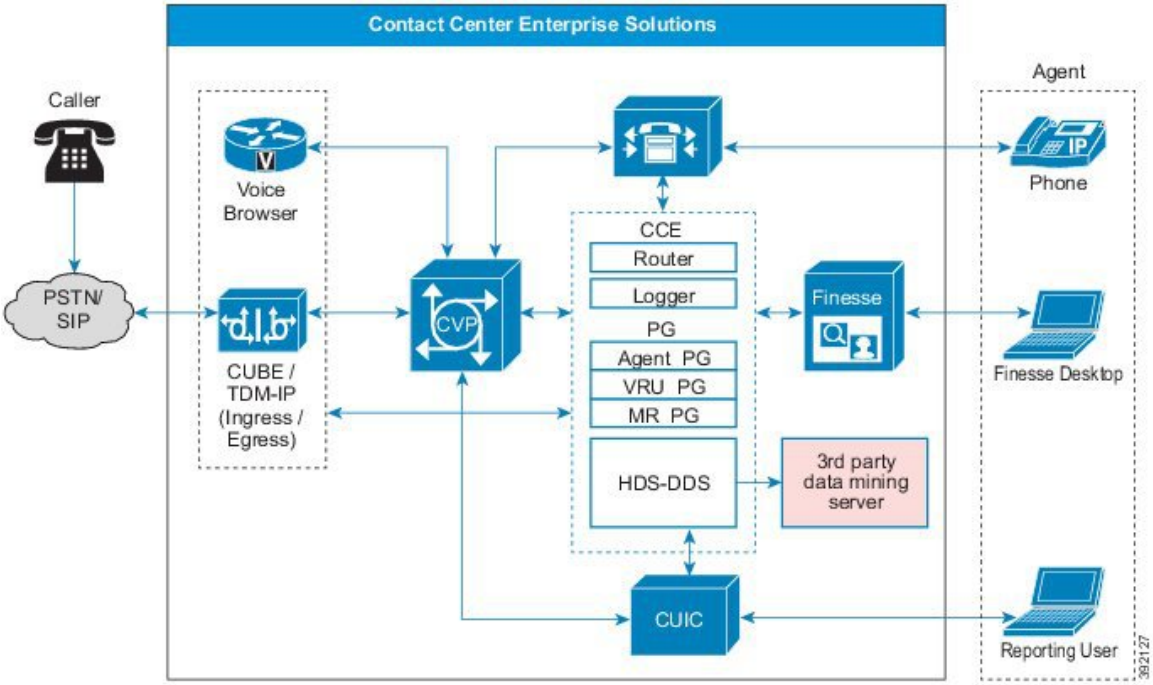
Note Administration Clients and Administration Workstations can support remote desktop access. But, only one agent can access a client or workstation at a time. Unified CCE does not support simultaneous access by several users on the same client or workstation.

Historical Data Server and Detail Data Server (HDS-DDS)

The role handles only data extraction and custom reports for call detail (TCD and RCD) records. You can only have one server of this type on each side of a redundant Logger pair. This role does not support these features:

- Real-time data reporting
- Configuration changes

Figure 20: Historical Data Server and Detail Data Server (HDS-DDS)



The Historical Data Server (HDS) and the Detail Data Server (DDS) provide longer-term historical data storage. The HDS stores historical data summarized in 15- or 30-minute intervals for reporting. The DDS stores detailed information about each call or call segment for call tracing. You can extract data from either source for warehousing and custom reporting.

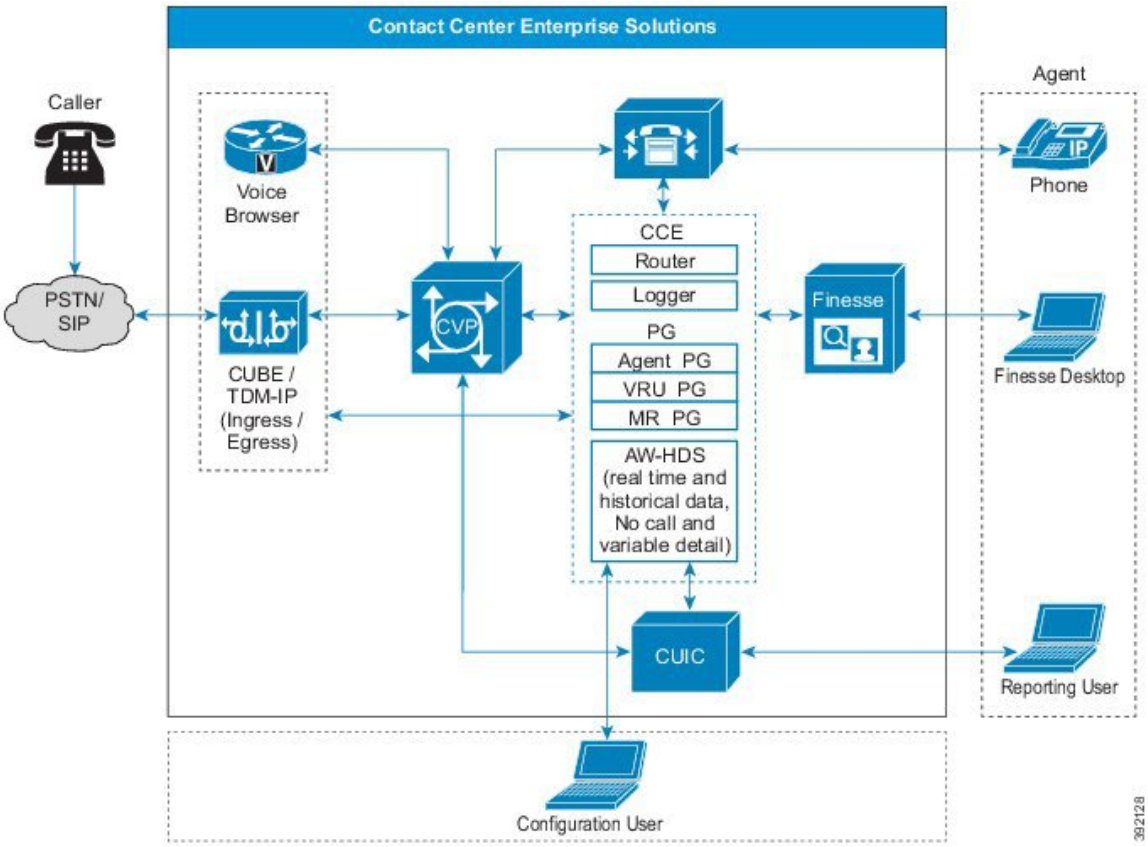
Typically, you deploy these Data Servers with a primary AW as a single server serving all three roles (AW-HDS-DDS). You use the HDS-DDS in large deployments where separating their function from the AW aids scalability.

Administration Server and Historical Data Server (AW-HDS)

This role handles configuration changes, real-time reporting, and historical reporting. This server uses the Cisco Unified Intelligent Center Reporting user for real-time and historical reporting. This role does not support these features:

- Call Detail, Call Variable, and Agent State Trace data
- Custom reporting data extraction

Figure 21: Administration Server and Historical Data Server (AW-HDS)



The Real-Time Data Server uses the AW database to store real-time data and configuration data. Real-time reports combine these two types of data to present a near-current snapshot of the system.

The Historical Data Server (HDS) provides longer-term historical data storage. The HDS stores historical data summarized in 15- or 30-minute intervals for reporting. You can extract data from the HDS for warehousing and custom reporting.

Figure 22: Communication Between Central Controller and Administration & Data Server

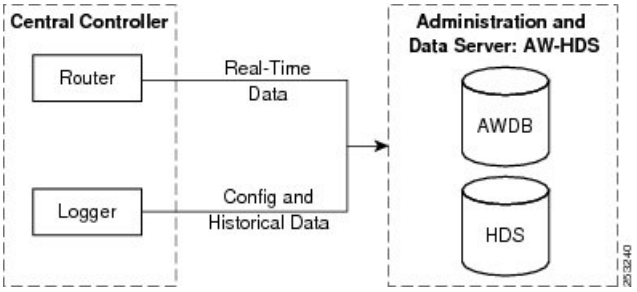
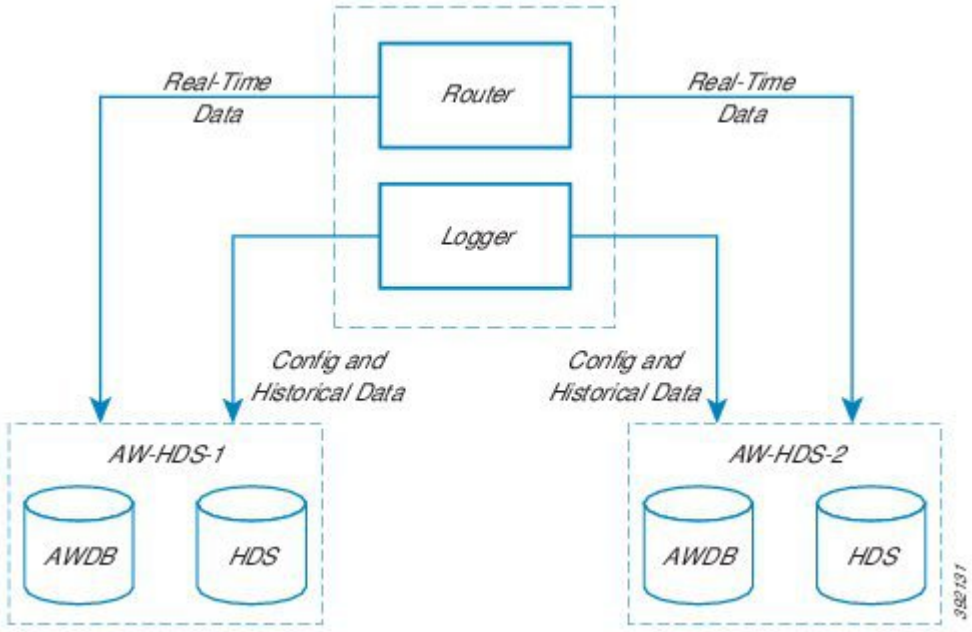


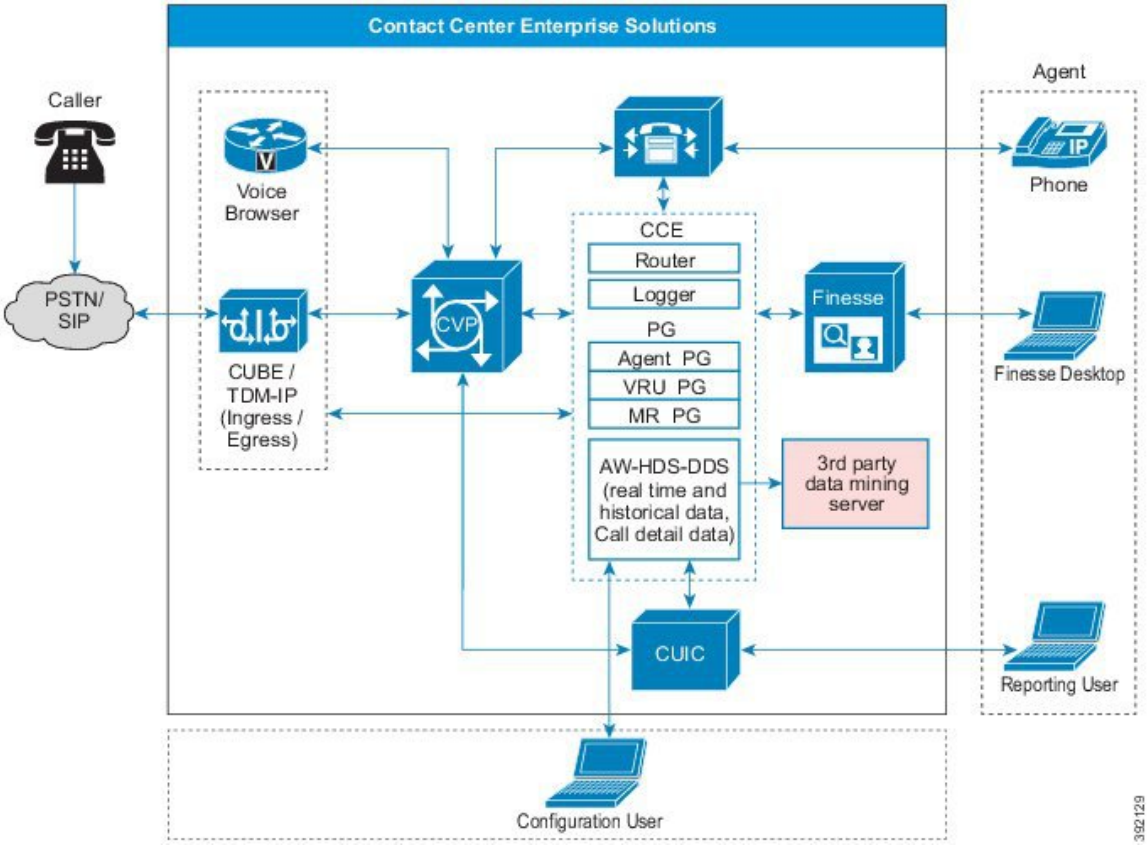
Figure 23: Communication Between Central Controller and Multiple Administration & Data Servers



Administration Server, Historical Data Server, and Detail Data Server (AW-HDS-DDS)

This role handles configuration changes, real-time reporting, and historical reporting, like the AW-HDS role. This server uses the Cisco Unified Intelligent Center (Unified Intelligence Center Reporting client) for real-time and historical reporting. This server also provides call detail and call variable data for custom reporting data extraction to feed historical data.

Figure 24: Administration Server, Historical Data Server, and Detail Data Server (AW-HDS-DDS)



The Real-Time Data Server uses the AW database to store real-time data and configuration data. Real-time reports combine these two types of data to present a near-current snapshot of the system.

The Historical Data Server (HDS) and the Detail Data Server (DDS) provide longer-term historical data storage. The HDS stores historical data summarized in 15- or 30-minute intervals for reporting. The DDS stores detailed information about each call or call segment for call tracing. You can extract data from either source for warehousing and custom reporting.

Data Purge

Data beyond the configured retention time is purged automatically at 12:30 AM and uses the time zone setting of the core server. The purge also triggers when the database reaches 80% and 90% of its maximum size.

Follow Cisco supported guidelines to run the purge at off-peak hours or during a maintenance window.

Note that you can control or change the automatic purge schedule through the command line interface. You can change it if the automated purge does not occur during your off-peak hours.

The purge has a performance impact on the Logger.

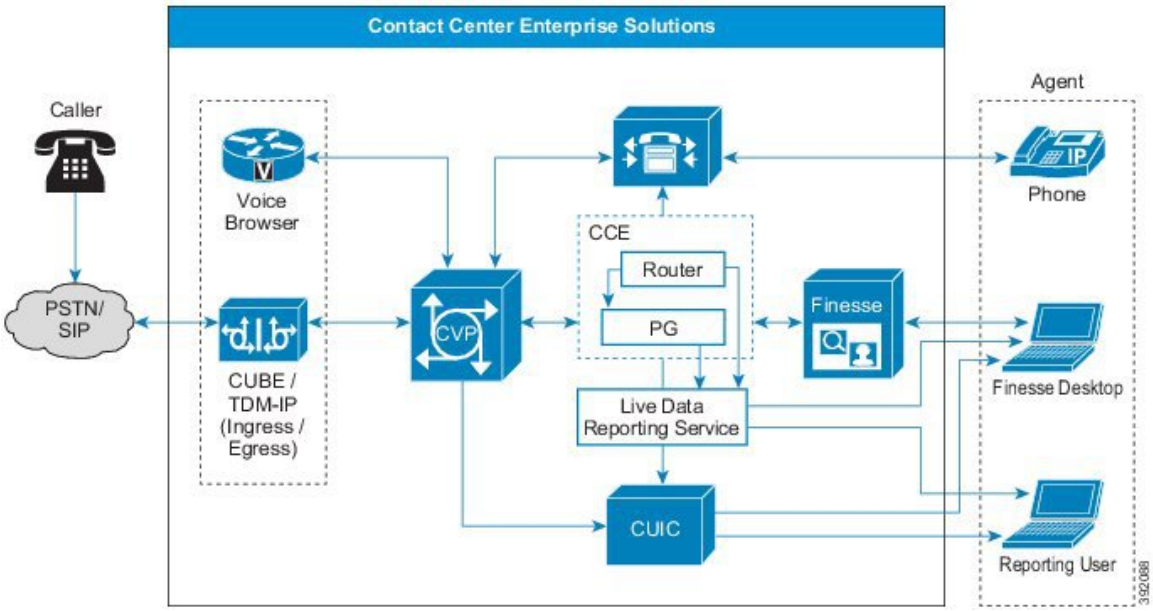
Live Data

Live Data is a data framework that processes real-time events with high availability for Live Data reports. Live Data continuously processes agent and call events from the peripheral gateway and the router. As events

occur, Live Data continuously pushes real-time updates to Unified Intelligence Center reporting clients. This table lists the placement of the Live Data services in the Reference Designs.

	2000 Agent	4000 Agent	12000 Agent	24000 Agent	Contact Director
Live Data placement	Colocated on a VM with Unified Intelligence Center and the Cisco Identity Service	Located on a standalone VM	Located on a standalone VM	Located on a standalone VM	The Contact Director does not have Live Data installed. Live Data is on the target Unified CCE instances.

Figure 25: Live Data Reporting



The PG and the Router push agent and call events to Live Data as the events occur. Live Data then continuously aggregates and processes the events in-stream and publishes the information. Unified Intelligence Center subscribes to the message stream to receive the events in real-time and continuously update Live Data reports. Individual state values, such as agent states, refresh as they happen. Other values, such as calls in queue, refresh approximately every 3 seconds.

Live Data resides in Unified CCE on a Cisco Voice Operating System (VOS) VM. You can embed Live Data reports in Finesse agent desktops.

Note Live Data requires that both Cisco Unified Intelligence Center and Cisco Finesse use the same transfer protocol. By default, both use HTTPS.

Cisco Virtualized Voice Browser

Cisco Virtualized Voice Browser (Cisco VVB) provides a platform for interpreting VXML documents. When an incoming call arrives at the contact center, Cisco VVB allocates a VXML port that represents the VoIP endpoint. Cisco VVB sends HTTP requests to the Unified CVP VXML server. The Unified CVP VXML server performs the request and sends back a dynamically generated VXML document.

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is the main call processing component of a Cisco Collaboration System. It manages and switches VoIP calls among IP phones. Unified CVP interacts primarily with Unified CM as a means for sending PSTN-originated calls to Unified CCE agents.

The following common scenarios require calls to Unified CVP to originate from Unified CM endpoints:

- An office worker (not an agent) on an IP phone dials an internal help desk number.
- An agent begins a consultative transfer that gets routed to a Unified CVP queue point.

Unified CM communicates with Unified CCE through the Java Telephony Application Programming Interface (JTAPI). In a fault-tolerant design, a Unified CM cluster supports thousands of agents. The number of agents and the number of busy hour call attempts (BHCA) supported within a cluster varies and must be sized according to Cisco guidelines.

Typically, when designing a Unified CCE solution, you first define the deployment scenario. You determine the arrival point (or points) for the voice traffic and the location (or locations) of the contact center agents. You then determine the sizing of the individual components within the Unified CCE design. This step includes determining how many Unified CM clusters and servers within each cluster are needed.

You can add a 2000 Agent Reference Design solution to an existing Unified CM deployment. In this case, the existing Unified CM cluster is an off-box replacement of the on-box cluster in the standard Reference Design layout. With this configuration, two of the subscribers must be dedicated to CCE. All devices on these subscribers must be SIP. In the global topology, each remote site can have its own Unified CM cluster.

Note

- Cisco Unified Communications Manager is supported on-box and off-box. Cisco Business Edition is supported off-box only.
- Move the CUCM VMs off-box before upgrading them to Release 12.5.

In a Unified CVP environment, Unified CM can be an Ingress or Egress Gateway. It is more common for Unified CM to be an Egress Gateway. Calls typically are from the PSTN, queued by Unified CVP, and then switched to Unified CM for handling by an agent. If the call is from an IP phone, not a PSTN, the Unified CM is an Ingress Voice Gateway from the perspective of Unified CVP.

Unified CM as an Egress Gateway

To deploy Unified CM with Unified CVP, use Unified CM call admission control for calls between the Ingress Voice Gateway and the agent IP phone. Unified CM recognizes the call coming from the centralized Unified CVP Call Server instead of from the Remote Ingress Voice Gateway.

Unified CM Ingress Gateway

When an IP phone initiates a call to Unified CVP, the Unified CM acts as the Ingress Voice Gateway to Unified CVP. A SIP trunk is used to send calls to Unified CVP.

Call Processing Nodes

Cisco Unified Communications Manager serves as the software-based call-processing component of the Cisco Unified Communications family of products.

The Unified CM system extends enterprise telephony features and functions to packet telephony network devices such as IP phones, media processing devices, voice-over-IP (VoIP) gateways, and multimedia applications. Unified CM provides signaling and call control services to Cisco-integrated telephony applications and third-party applications. Unified CM performs the following primary functions:

- Call processing
- Signaling and device control
- Dial plan administration
- Phone feature administration
- Directory services
- Operations, administration, maintenance, and provisioning (OAM&P)
- Programming interface to external voice-processing applications such as Cisco IP Communicator, Cisco Unified Customer Voice Portal (CVP)

The Unified CM system includes a suite of integrated voice applications that perform voice-conferencing and manual attendant console functions. This suite of voice applications means that no need exists for special-purpose voice-processing hardware. Supplementary and enhanced services such as hold, transfer, forward, conference, multiple line appearances, automatic route selection, speed dial, last-number redial, and other features extend to IP phones and gateways. Because Unified CM is a software application, enhancing its capabilities in production environments requires only upgrading software on the server platform, avoiding expensive hardware upgrade costs.

Distribution of Unified CM and all Cisco Unified IP Phones, gateways, and applications across an IP network provides a distributed, virtual telephony network. This architecture improves system availability and scalability. Call admission control ensures that voice quality of service (QoS) is maintained across constricted WAN link. It automatically diverts calls to alternate public switched telephone network (PSTN) routes when WAN bandwidth is not available.

A browser interface to the configuration database provides the capability for remote device and system configuration. This interface also provides access to HTML-based online help for users and administrators.

Unified CM, designed to work like an appliance, refers to the following functions:

- Unified CM servers can get preinstalled with software to ease customer and partner deployment. They automatically search for updates and notify administrators when key security fixes and software upgrades are available for the system. This process comprises Electronic Software Upgrade Notification.
- You can upgrade Unified CM servers while they continue to process calls, so upgrades take place with minimal downtime.
- Unified CM supports the Asian and Middle Eastern markets by supporting Unicode on higher resolution phone displays.

- Unified CM provides Fault, Configuration, Accounting, Performance, and Security (FCAPS).

TFTP and Music on Hold Nodes

A TFTP subscriber or server node performs two main functions as part of the Unified CM cluster:

- The serving of files for services to devices such as phones and gateways. This includes configuration files, binary files for upgrades, and various security files.
- Generation of configuration and security files. These are signed and sometimes encrypted before being made available for download.
- You can enable the Cisco TFTP service that provides this functionality on any server in the cluster. In a cluster with more than 1250 users, configuration changes that cause the TFTP service to regenerate configuration files can affect other services. In such clusters, dedicate a specific subscriber node to the TFTP service and MOH feature or any features that cause frequent configuration changes.
- Use the same hardware platform for the TFTP subscribers as used for the call processing subscribers.
- A Unified Communications Manager MoH server can generate a MoH stream from two types of sources, audio file and fixed source. Either source can be transmitted as unicast or multicast.

Cisco Finesse

Cisco Finesse is the next-generation agent and supervisor desktop for Cisco Unified Contact Center Enterprise, providing benefits across various communities that interact with your customer service organization. It is designed to improve collaboration by enhancing the customer and customer service representative experience.

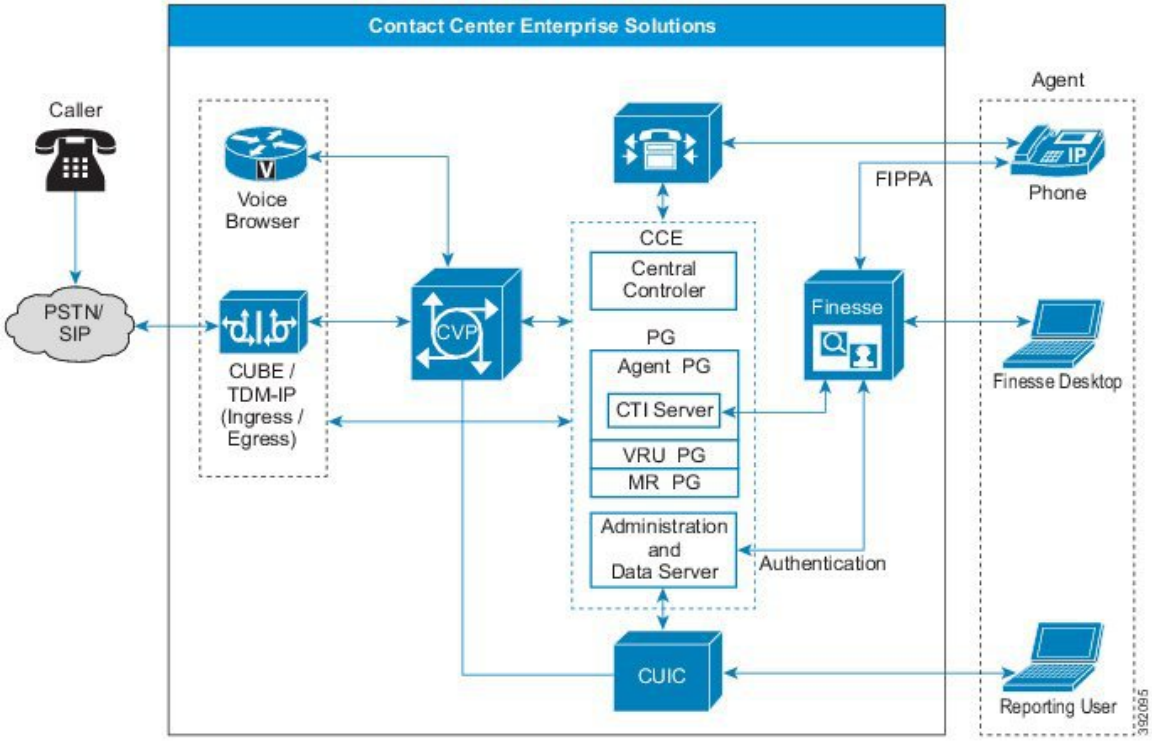
The Cisco Finesse agent and supervisor desktop for Cisco Unified Contact Center Enterprise integrates traditional contact center functions into a thin-client desktop. A critical characteristic is that every desktop is browser-based and implemented through a Web 2.0 interface. No client-side installations are required. This reduces the total cost of ownership (TCO).

Cisco Finesse also provides a Web 2.0 software development kit (SDK) and gadgets to enable developers to quickly implement the desktop.

You deploy the Cisco Finesse server on a dedicated VMware virtual machine (VM) that runs on the Cisco Voice Operating System (VOS) platform. The Cisco Finesse server is a required component for the Cisco Finesse desktop solution. The Cisco Finesse software is fault-tolerant and deploys on redundant VMs. Both Cisco Finesse servers are simultaneously active. One Cisco Finesse server acts as a publisher and replicates configuration data to the subscriber in the redundant pair.

The Cisco Finesse server connects to the CTI server on the Agent PG. Authentication with Unified CCE is provided over a connection to the Administration & Data Server. If you enable Single Sign-On (SSO), the Cisco Identity Service provides authentication.

Figure 26: Cisco Finesse in a Contact Center Enterprise Solution



Cisco Finesse requires that you deploy the Administration & Data Server with a backup Administration & Data Server. If the primary Administration & Data Server goes down, Cisco Finesse connects to the backup server for authentication so that agents can still sign in.

The Cisco Finesse server exposes supported client operations through a Representational State Transfer (REST) API. The REST API shields the developer from many of the details surrounding the CTI server wire protocol.

Cisco Finesse clients connect to the Cisco Finesse server over a web browser that points to the fully qualified domain name (FQDN) of the Cisco Finesse server.

You deploy the Cisco Finesse server in an active/active deployment, where both Cisco Finesse servers connect to the active CTI server on the Agent PG. The standard Cisco VOS replication mechanism provides redundancy for persistent configuration data on the Cisco Finesse servers.

Cisco Finesse Server Services

You can access the following Cisco Finesse services using the CLI:

- **Cisco Finesse Notification service**—This service is used for messaging and events. The Cisco Finesse desktop uses this service to view call events, agent state changes, and statistics.
- **Cisco Finesse Tomcat service**—This service contains all deployed Cisco Finesse applications. These applications include the following:
  - Cisco Finesse desktop application: This application provides the user interface for agents and supervisors.
  - Cisco Finesse IP Phone Agent application: This application allows agents and supervisors to perform Cisco Finesse operations on their Cisco IP Phone.

- Cisco Finesse REST API application: Cisco Finesse provides a REST API that enables client applications to access the supported server features. The REST API can use HTTPS to transport application data. The REST API also provides a programming interface that third-party applications can use to interact with Cisco Finesse. See the Cisco Finesse documentation at https://developer.cisco.com/site/finesse/ for more information on the REST API.
- Cisco Finesse administration application: This application provides the administrative operations for Cisco Finesse.
- Cisco Finesse Diagnostic Portal application: This application provides performance-related information for Cisco Finesse.

Agent Mobility

The Unified CCE deployment does not statically associate the agent desktop with any specific agent or IP phone extension. You configure agents and phone extensions within Unified CCE and associate them with a specific Unified Communications Manager cluster.

When agents sign in to their desktop, a dialog prompts for an agent ID or username, password, and the phone extension to use for that session. Then, the agent ID, phone extension, and agent desktop IP address are dynamically associated. The association is released when the agent signs out.

This mechanism allows an agent to work (or hot-desk) at any workstation. The mechanism also allows agents to take their laptops to any appropriately configured Cisco Unified IP Phone and sign in from that device.

Agents can also sign in to other phones using the Cisco Extension Mobility feature. For more information about this feature, see the Extension Mobility section of the *Feature Configuration Guide for Cisco Unified Communications Manager* at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html.

Cisco Unified Intelligence Center

Cisco Unified Intelligence Center (Unified Intelligence Center) is a web-based reporting application that provides easily consumable Live Data, real-time, and historical reporting for Unified CCE and Unified CVP. It allows supervisors and business users to report from a single interface on the details of multichannel contacts across the solution. You can extend the boundaries of traditional reporting to an information portal where you can integrate and share data throughout the organization.

You deploy the Unified Intelligence Center server on a dedicated VM that runs on the Cisco Voice Operating System (VOS) platform. In the 2000 Agent Reference Design, Unified Intelligence Center is coresident with Live Data and the Cisco Identity Service.

Unified Intelligence Center offers high scalability, performance, and advanced features such as data integration with other Cisco Unified Communications products or third-party data sources. Unified Intelligence Center incorporates a security model that defines different access and capabilities for specific users.

Cisco Unified Intelligence Center offers both a web-based reporting application and an administration interface. Unified Intelligence Center reporting capabilities include the following:

- Dashboard mashups
- Powerful grid presentations of reports with sorting and grouping
- Chart and gauge presentations of reports

- Association of multiple report displays with the same report definition
- Custom filters
- Custom thresholds to alert on the data
- Stock report templates for contact center enterprise data
- Ability to report data from MS SQL Server and Informix databases

Administrators can use Unified Intelligence Center to control access to features, reports, and data by granting privileges only to authorized individual users or groups of users. For example, you can assign each supervisor to a group of agents, skills, and call types that are the most relevant to them. This allows each report to provide focused, actionable insights into data that is appropriate to their role.

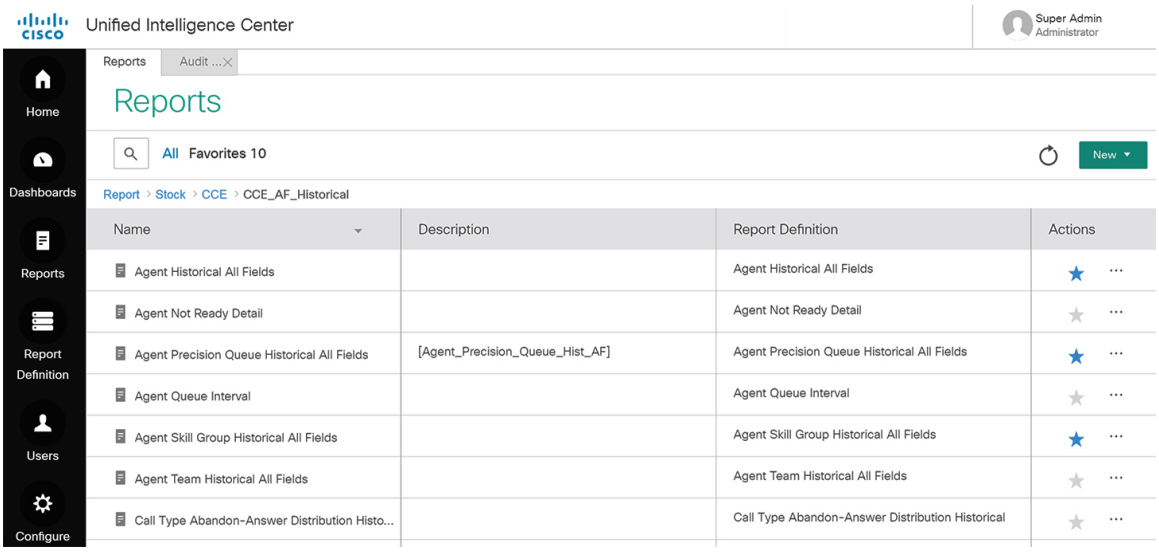
Several features in this product allow you to extend the Unified Intelligence Center platform beyond traditional reporting and into an enterprise-wide information portal. You can use data from nontraditional sources to improve business efficiency and effectiveness.

The Unified CCE Reporting solution provides an interface to access Live Data, real-time, and historical data for the contact center.

The reporting solution consists of the following components:

- Cisco Unified Intelligent Center—Reporting user interfaces
- Configuration and Reporting Data—Contained on one or more Administration & Data Servers

Figure 27: Unified Intelligence Center



Optional Cisco Components

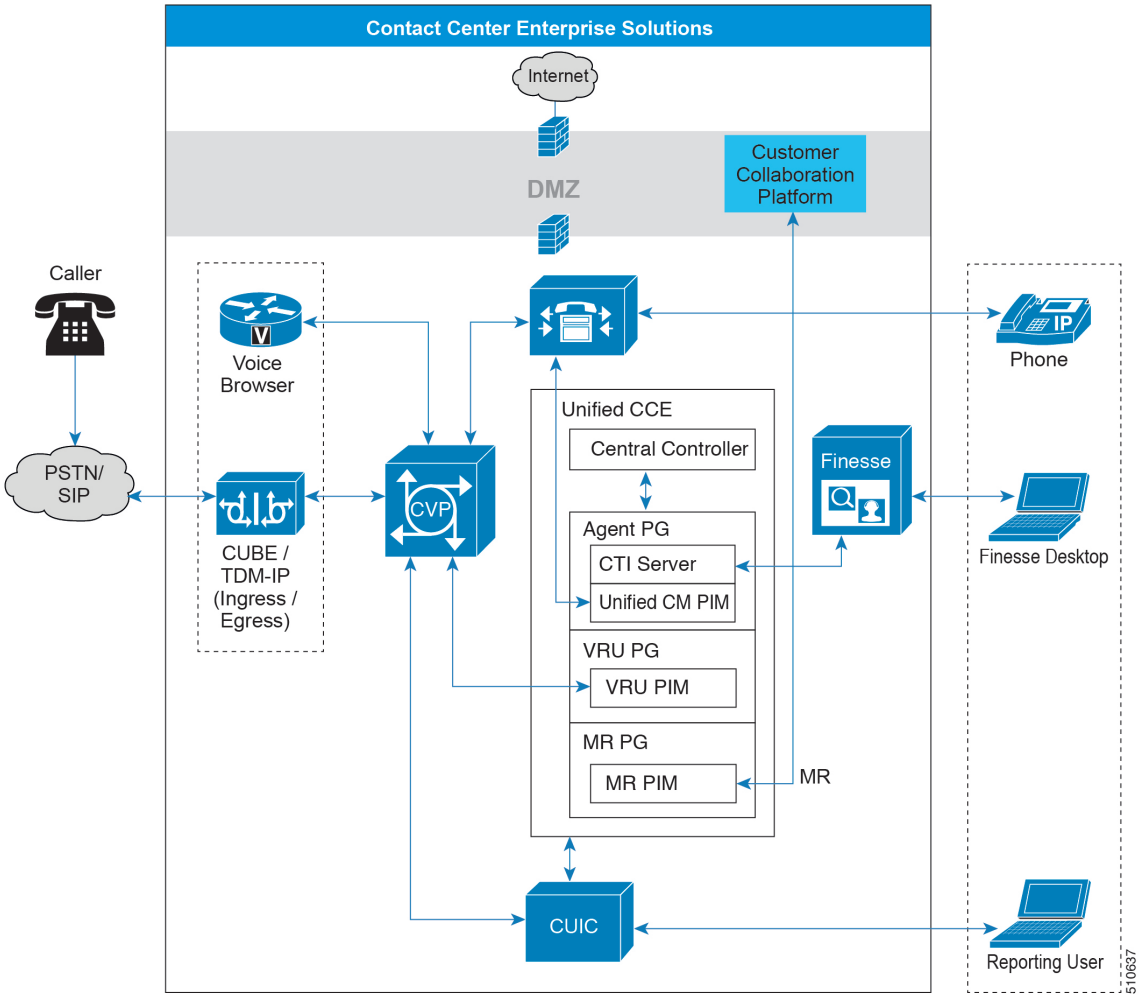
Some contact center enterprise solutions use these optional Cisco components. You add them to a solution when you want the functionality that they offer. Usually, these optional components require extra servers.

Cisco Customer Collaboration Platform

Cisco Customer Collaboration Platform provides the means to route digital media requests to agents in your contact center. Your solution can use Customer Collaboration Platform for the following:

- The Agent Request feature which allows a customer to initiate a request a call from an agent from a web site. For more information on this feature, see the *Cisco Unified Contact Center Enterprise Features Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html.
- The Task Routing APIs which you can use to integrate third-party multichannel applications.

Figure 28: Customer Collaboration Platform in Contact Center Enterprise Solutions



Task Routing

Task Routing describes the system's ability to route requests from different media channels to any agents in a contact center.

You can configure agents to handle a combination of voice calls, emails, chats, and so on. For example, you can configure an agent as a member of skill groups or precision queues in three different Media Routing Domains (MRD) if the agent handles voice, e-mail, and chat. You can design routing scripts to send requests to these agents based on business rules, regardless of the MRD from which the request came. Agents logged into multiple MRDs may switch media on a task-by-task basis.

The optional component Enterprise Chat and Email provides Task Routing out of the box. Third-party multichannel applications can use Task Routing by integrating with CCE through the Task Routing APIs.

Task Routing APIs provide a standard way to request, queue, route, and handle third-party multichannel tasks in CCE.

Contact Center customers or partners can develop applications using Customer Collaboration Platform and Finesse APIs in order to use Task Routing. The Customer Collaboration Platform Task API enables applications to submit nonvoice task requests to CCE. The Finesse APIs enable agents to sign into different types of media and handle the tasks. Agents sign into and manage their state in each media independently.

Cisco partners can use the sample code available on Cisco DevNet as a guide for building these applications (https://developer.cisco.com/site/task-routing/).

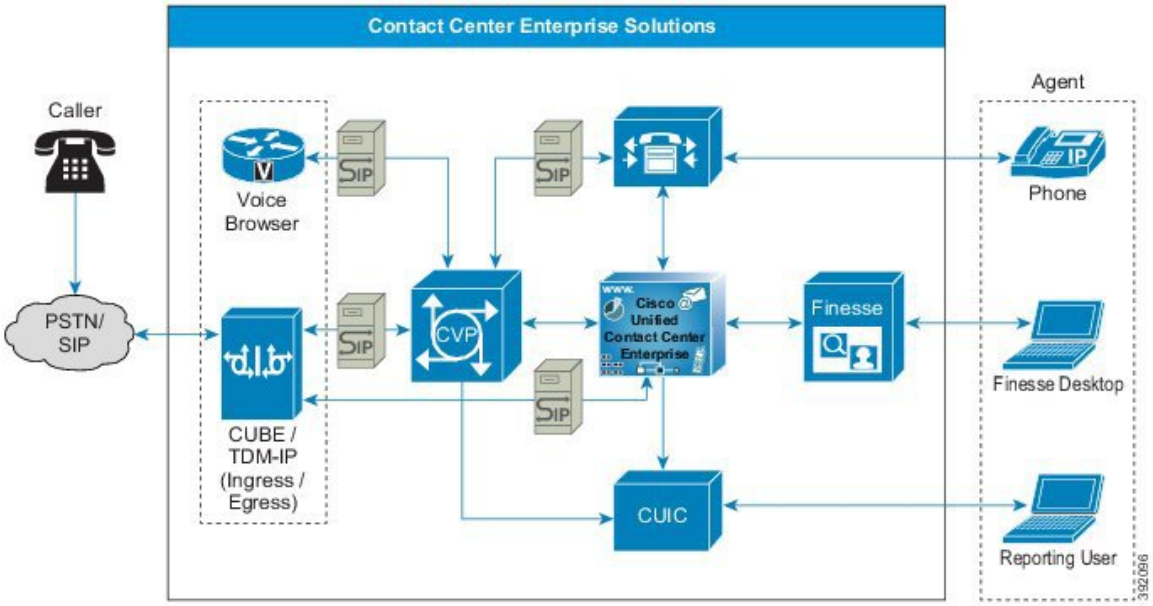
Cisco Unified SIP Proxy

The Cisco Unified SIP Proxy (CUSP) is a high-performance, highly available Session Initiation Protocol (SIP) server for centralized routing and SIP signaling normalization. By forwarding requests between call-control domains, CUSP enables you to route sessions within enterprise and service provider networks. The application aggregates SIP elements and applies highly developed routing rules. These rules enhance control, management, and flexibility of SIP networks.

Unified CVP supports only the CUSP Server.

In a Unified CVP deployment, a CUSP Server sees incoming calls from the TDM Gateway, from Unified CVP, and from the UCM SIP trunk. With a SIP back-to-back user agent in CVP, the initial call setup from the proxy involves an inbound call immediately followed by an outbound call (whether for VRU or to ACD). Later in the call, CVP may transfer the call to an agent, which involves an outbound leg, and reinvites to the inbound leg. A ringtone service setup is also available which also involves a separate outbound call and a reinvite to the caller. Reinvites on the caller leg occur at CVP transfer or during supplementary services.

Figure 29: CUSP in a Contact Center Enterprise Solution



The CUSP Server routes SIP messages among SIP endpoints. The CUSP Server enables solution wide SIP-endpoint high availability and load balancing. The CUSP Server is designed to support multiple SIP endpoints of various types and to implement load balancing and failover among these endpoints. Deployment of a SIP proxy in the solution enables a more centralized configuration of the dial plan routing configuration.

You can configure a SIP proxy with multiple static routes to do load balancing and failover with outbound calls. The static routes can point to an IP address or a DNS.

Domain Name System (DNS) Service Record (SRV) is not qualified for use on the CUSP Server. However, you can use it for the devices that must reach the CUSP Server, such as Unified CVP, Ingress Voice Gateway, and Unified CM.

You can deploy Unified CVP without a CUSP Server, depending on the design and complexity of the solution. In such cases, some of the functions that a CUSP Server provides are provided by the Unified CVP Server SIP service.

Following are the benefits of using a CUSP Server:

- You can use priority and weight routing with the routes for load balancing and failover.
- If a CUSP Server exists in your SIP network, then Unified CVP acts as an additional SIP endpoint. The Unified CVP fits incrementally into the existing SIP network.

If you do not use a CUSP Server, then the Ingress Voice Gateways and Unified CMs must point directly to Unified CVP. In such a deployment, perform the following tasks:

- Perform load balancing using DNS SRV lookups from gateway to DNS Server; balance SIP calls using this procedure.
- Perform load balancing of calls outbound from Unified CVP (outbound call leg) using DNS SRV lookups.

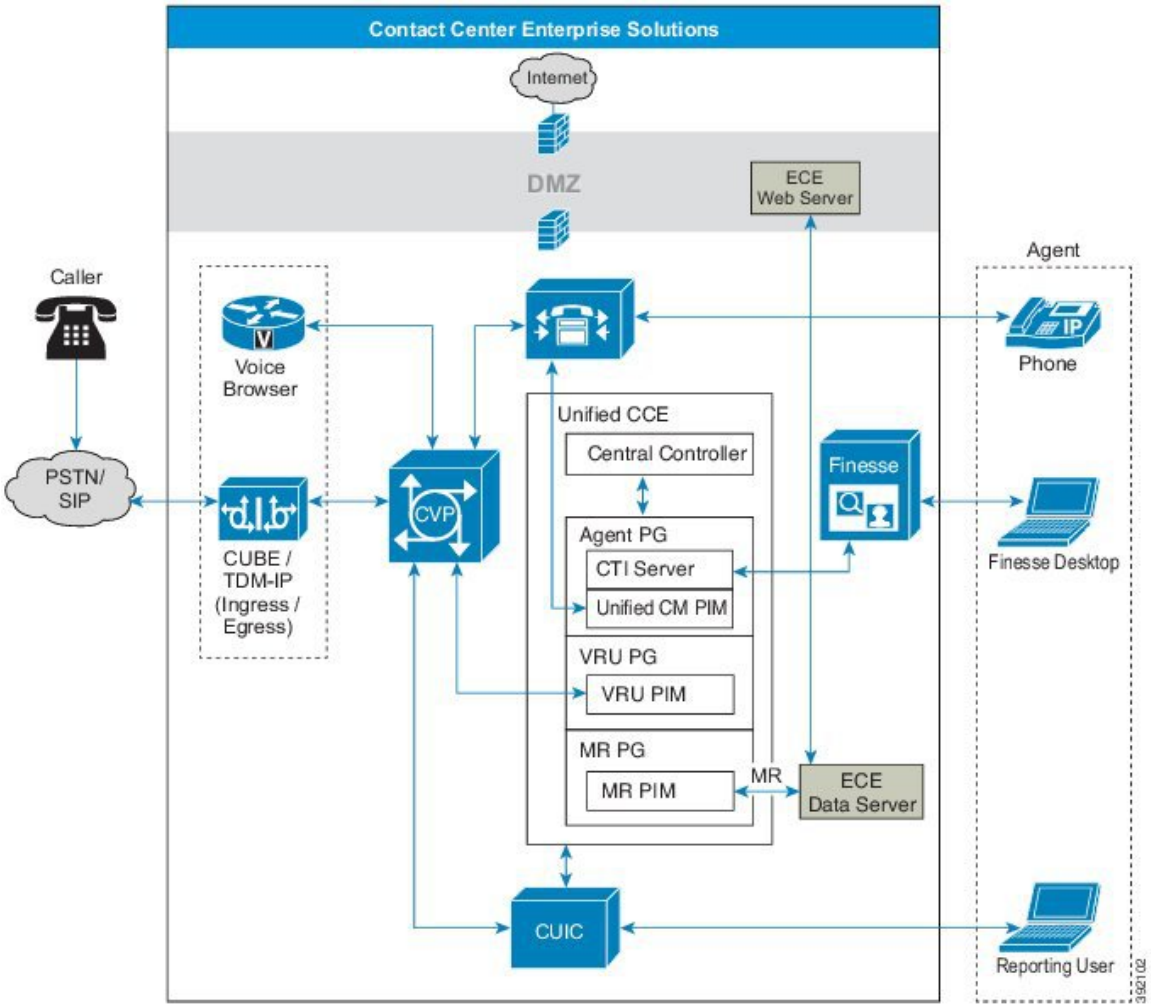
Enterprise Chat and Email

The contact center enterprise solutions use Enterprise Chat and Email (ECE) to provide a multichannel contact center.

For email, ECE enables organizations to intelligently route and process inbound emails, webform inquiries, faxes, and letters. For web-chat, ECE provides agents with a comprehensive set of tools for serving customers in real time. It enables call center agents to provide immediate personalized service to customers through text chat messaging and page-push abilities.

Deploy the ECE Web Server on an external server. You can place that server either in the same `site` as the ECE Data Server or in a DMZ if customer chat interactions require that.

Figure 30: ECE in Contact Center Enterprise Solutions



Enterprise Chat and Email Features

Following are the Enterprise Chat and Email (ECE) features.

Email

ECE supports email to create a communication channel between a customer and an agent. There are various steps involved in efficiently responding to emails from customers. Emails are first retrieved into the system and routed to appropriate users or queues. Once a response is created, it is processed through the system and sent to the customer.

Chat

It is an activity created for a chat session between a customer and an agent. A chat is a real time interaction between an agent and a customer during which they exchange text messages. As part of a chat, agents can also push web pages to customers. Based on how chat activities are routed to agents, they can be categorized as Standalone chats and Integrated chats. An integrated chat is routed to an integrated queue, and a message is sent to Unified CCE. Unified CCE processes the activity and assigns the chat to an available agent.

Web Callback and Delayed Callback

The Web Callback feature allows you to request a callback by submitting a form on a website. ECE processes the submitted information and connects the user with an agent. In the contact center enterprise integration, the ECE sends a message to Unified CCE requesting Unified CCE to route the callback request to an agent. Unified CCE sends a message to ECE. When an agent is available, the Call Router notifies the agent to begin the Web Callback.

The Delayed Callback feature is similar to the Web Callback feature. When the ECE receives the delayed callback request, it adds the request in the Delayed Callback table. ECE sends the HTML page to the caller that tells the timeframe for the callback. When the specified time arrives, ECE moves the request to the Unified CCE queue for routing to Unified CCE. The call is then processed the same way as for Web Callback.

Cloud Connect

Cloud Connect is a new component that allows customers to use cloud services such as Webex Experience Management. The administrator can configure the Cloud Connect server settings in Unified CCE Administration to contact the Cisco cloud services.

For information on how to configure Cloud Connect, see the *Cisco Packaged Contact Center Enterprise Administration and Configuration Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html

Third-Party Components

You can extend the functionality of your contact center enterprise solution with third-party components.

Load Balancers

In Contact Center Enterprise Reference Designs, load balancers are used in redirect mode only. You can use third-party load balancers for the following purposes in your contact center enterprise solution:

- For access to the Cisco Finesse sign-in page
- When you use the Finesse REST API directly
- With Unified CVP

- For access to the Unified CCE Administration tool sign-in page
- When you use the Unified CCE Administration REST API directly
- With Cisco Unified Intelligence Center
- With Cisco Unified Intelligence Center Administration Console

For more information on load balancer requirements, see the *Compatibility Matrix* for your contact center enterprise solution.

Recording

The Recording option provides network-based storage of media, including audio and video, with rich recording metadata. You can record, play back, and live stream the media. You can use this option for compliance, quality management, and agent coaching. The platform provides an efficient, cost-effective foundation for capturing, preserving, and mining conversations for business intelligence.

Note Unified CVP has a network-based recording (NBR) feature to support software-based forking for Real-time Transport Protocol (RTP) streams.

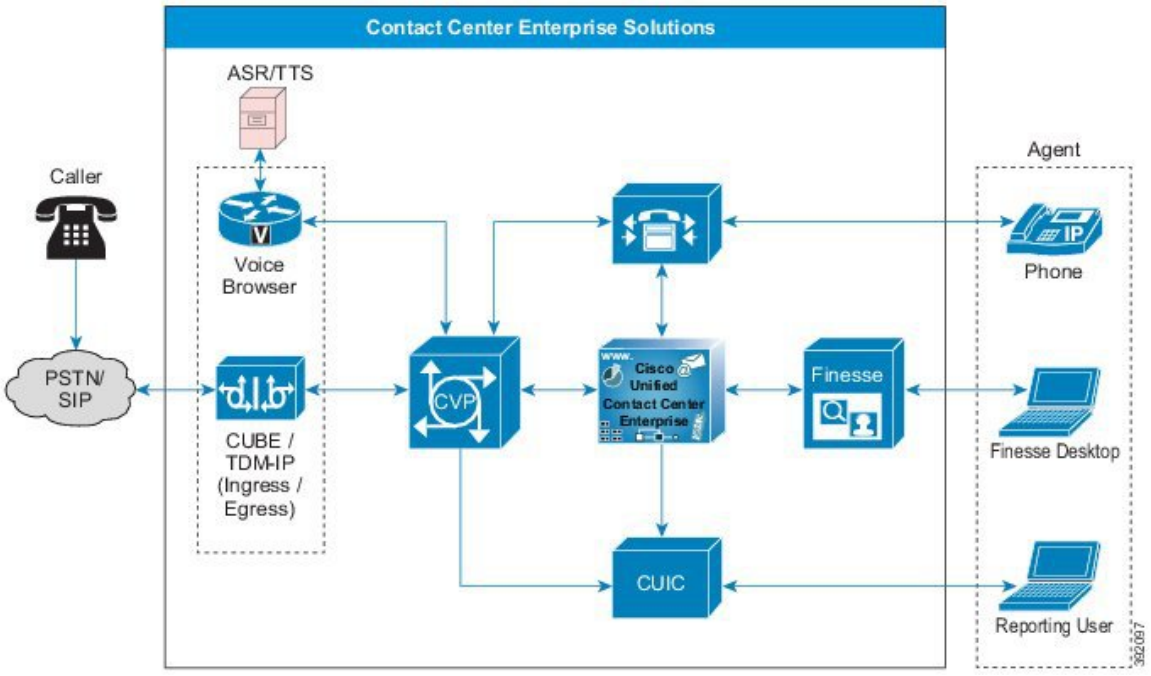
Note For ERSPAN support on UCS B Series for any third-party recording application, consult the vendor's application requirements.

Speech Servers - ASR/TTS

Automatic Speech Recognition (ASR) Server and Text-to-Speech (TTS) Server provides speech recognition services and text-to-speech services for a Voice Browser. Automatic Speech Recognition (ASR) enables callers to verbally choose menu options. For example, an Automated Attendant can ask who you are calling and then use your reply to connect the call. Text-to-Speech (TTS) converts plain text (UNICODE) into speech. For example, Voice Browsers can stream media from a text-to-speech (TTS) server.

ASR/TTS license use depends on what you use for a voice browser. The VXML Gateway does not release the ASR/TTS license until the end of a call. Cisco VVB releases the license when the script no longer requires it.

Figure 31: Speech Servers in Contact Center Enterprise Solutions



Communication between the ASR and TTS servers and the Voice Browser uses Media Resource Control Protocol (MRCP). See the *Compatability Matrix* for details on the support for MRCP versions.

The World Wide Web Consortium (W3C) provides a rich feature set to support the ASR grammars. You can implement and support inline grammars which pass the set of acceptable customer responses to the Voice Browser. You can also use external grammars, where Unified CCE passes a pointer to an external grammar source. The VXML Server adds this pointer to the VXML document that it sends to the Voice Browser. The Voice Browser then uses the grammar to check ASR input from the caller. In this case, the customer creates the grammar file. A third type of grammar is the built-in grammar. For a complete explanation of grammar formats, see the W3C website at http://www.w3.org/TR/speech-grammar/.

When the VXML Server directly passes the text for TTS to the gateway, we refer to the action as inline TTS. A separate server that communicates with the Voice Browser through MRCP performs the speech recognition and speech synthesis. The ASR and TTS engine also supports (with limitations) voice recognition and synthesis for multiple languages.

For information on third-party ASR or TTS software and servers, see your solution's *Compatibility Matrix*.

Wallboards

Wallboards enable you to monitor, in real time, the service that you are providing to your customers. Wallboards display information on customer service metrics such as number of calls waiting, waiting call length, and Service levels.

Workforce Management

Workforce Management (WFM) enables you to schedule multiple queues and sites. You can use a single WFM implementation worldwide. WFM also enables you to manage key performance indicators and real-time adherence to schedules.

Your users (agent, supervisor, scheduler, and administrator) can access WFM with a web browser. Because you avoid the installation of a thick client, WFM is ideally suited to a highly distributed workforce environment.

Integrated Features

The difference between optional components and integrated features is the ease of adding them to your solution. In general, an integrated feature does not require you to add a server or VM to your solution. You only configure it to activate it in your solution. But, remember that these features can have significant sizing or other design impacts.

You can find more information on various integrated features in your solution's *Feature Guide*.

Agent Greeting

With Agent Greeting, you can play a configurable, automated greeting to callers. Every caller receives a clear, well-paced, language-appropriate, and enthusiastic introduction from the answering agent. Agent Greeting relieves your agents from speaking opening scripts. Instead, your agents can spends the time reviewing the desktop screen pop-ups while the greeting plays.

Recording a greeting is much the same as recording a message for voice mail. Depending on how you set up the call center, agents record different greetings that play for different types of callers (for example, an English greeting for English speakers or an Italian greeting for Italian speakers).

Agent Greeting is available to agents and supervisors who use IP Phones with Built-in-Bridge (BiB) that are controlled by the Unified CCE and Unified CM.

Figure 32: Agent Greeting

Application Gateway

The Application Gateway provides an interface for the CCE routing engine to query an external service. It requires a custom application to be written that uses the Application Gateway protocol, GED-145, which is open to our development partners. For more information, see https://developer.cisco.com/site/devnet/home/index.gsp.

Application Gateway allows you to insert application gateway nodes in their scripts. These nodes help you to populate variables and send requests to the custom application, and retrieve relevant information. The information can be used in administrative scripts to open or close programs. It can also return relevant customer data in a routing script which can be sent to the agent.

Business Hours

The Business Hours feature lets you create schedules for regular working hours and extra working hours, and to close the contact center for holidays or emergencies. It provides the mechanism for routing these contacts to specific support teams based on the configured work hour schedules, holidays, emergency closures, or extra working hours. You can create Business Hour schedules for various scenarios for various contact center teams. This feature helps you create and apply several Business Hour schedules to the same team. On the other hand, you could apply the same Business Hour schedule to several support teams.

When a customer contacts the contact center, the response by the contact center is based on the status of the support team. This status is evaluated using the Business Hour configured for the team.

Cisco Outbound Option

In contact center enterprise solutions, agents can handle both inbound and outbound contacts. Contact center managers in need of outbound campaign solutions can take advantage of the enterprise view that Cisco Unified CCE maintains over agent resources. Cisco Outbound Option supports agent-based and VRU-based campaigns. For agent-based campaigns, it also supports transfer of calls to a VRU for answering machines or to meet regulatory requirements for abandoned calls. A VRU campaign does not use agents, instead the call is directed to a VRU which plays a recorded message to answered calls.

The Cisco Outbound Option Dialer provides outbound dialing functionality along with the existing inbound capabilities of the Cisco Unified Contact Center Enterprise. This application enables the contact center to dial customer contacts and direct contacted customers to agents. With Cisco Outbound Dialer, you can configure a contact center for automated outbound activities.

The Outbound Option Dialer is a software-only process that coresides on the Unified CM PG. The SIP Dialer process communicates with Voice Gateways or CUBE, Outbound Option Campaign Manager, CTI Server, and MR PIM. The Dialer communicates with the Campaign Manager to retrieve outbound customer contact records and to report outbound call disposition (including live answer, answering machine, RNA, and busy). The Dialer communicates with the Voice Gateway to place outbound calls. The Dialer communicates with the CTI Server to monitor skill group activity and to perform third-party call control for agent phones. The SIP Dialer communicates with the MR PIM to submit the route requests to select an available agent.

The Outbound Option Dialer can dial customers on behalf of all agents located on its peripheral. The Dialer is configured with routing scripts that can run in the following modes:

- Full blended mode-An agent can handle inbound and outbound calls
- Scheduled modes-For example, 8:00 a.m. to 12:00 p.m. (0800 to 1200) in inbound mode and 12:01 to 5:00 p.m. (1201 to 1700) in outbound mode

- Completely in outbound mode

If blended mode is enabled, the Dialer competes with inbound calls for agents. The Dialer does not reserve more agents than are configured in the administrative script Outbound Percent variable. If all agents are busy, then the Dialer does not attempt to reserve any additional agents.

You can deploy Outbound Option in several ways to achieve more or less high availability:

- **Single Campaign Manager, Outbound Option Import, and Database**—This is a non-fault tolerant configuration of the subcomponents that direct operation of the SIP Dialers. If the Campaign Manager or Outbound Option Import goes down, you lose outbound calling until they come back online. This configuration can direct multiple Dialers.
- **Redundant Campaign Managers, Outbound Option Import, and Databases**—This fault-tolerant configuration includes redundant subcomponents that operate in a warm-standby mode. If the active Campaign Manager or Outbound Option Import goes down, your solution fails over to the standby subcomponents. This configuration requires more bandwidth to keep the sides in synch. It also requires more disk space to maintain the duplicate records.
- **Redundant SIP Dialers**—Your solution can include one pair of redundant SIP Dialers for each Agent PG pair. You do not have to include a Dialer pair with every Agent PG pair. The Campaign Manager can load-balance across the available Dialers.
- **Multiple Voice Gateways and Unified SIP Proxy servers**—You can increase high availability by adding a Unified SIP Proxy pair for each Dialer. You can then add extra voice gateways for each Unified SIP Proxy pair. This enables you to increase the calls made by each Dialer to more than a single voice gateway can support. The solution balances the load across the available instances.

Cisco Outbound Option supports Call Progress Analysis (CPA) configuration on a campaign basis. When you enable this feature, the SIP Dialer instructs the Voice Gateway or CUBE to analyze the media stream. The gateway determines the nature of the call (such as voice, answering machine, modem, or fax detection).

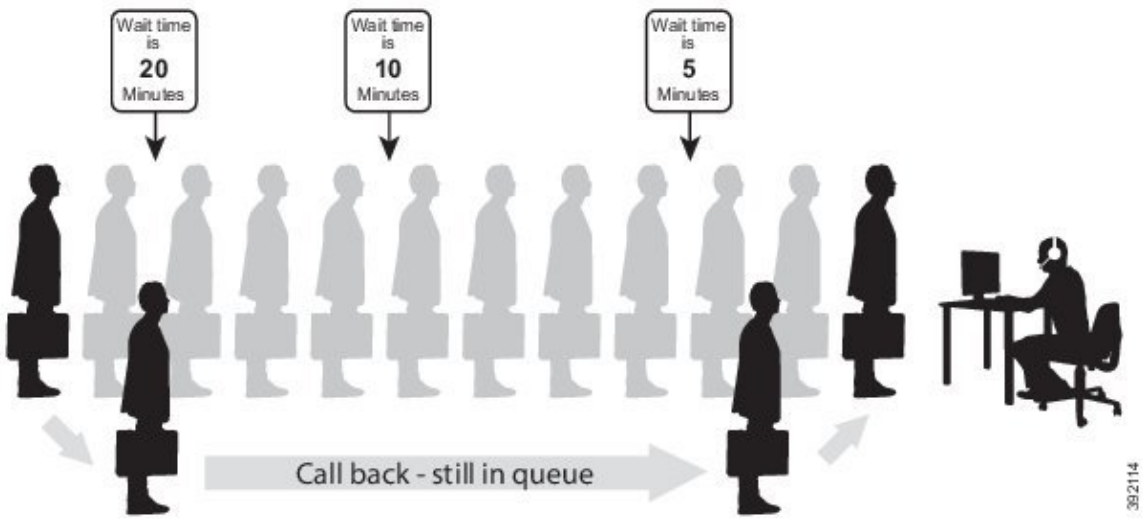
Note Virtual CUBE does not support CPA. Use a dedicated physical gateway if your solution needs CPA.

Courtesy Callback

Courtesy Callback gives a caller the option to have an agent return their call. This option limits the time a caller waits on the phone for an agent to answer.

Each call has a calculated Estimated Wait Time (EWT). When a caller's EWT approaches zero, the script places a call back to the caller. When the caller answers, the script inserts the caller back into the queue with their original order. The caller reaches an agent in the same time as if they had stayed on the phone.

Figure 33: Courtesy Callback



Call Context

Call Context refers to the attributes and data that are associated with a call.

Call Variables

You use call variables to pass business relevant data from Unified CVP to the agent desktop. Contact center enterprise solutions have a set of ten call variables. Each variable can contain 40 bytes of data.

Custom SIP Headers

With this feature, Unified CVP can pass selected SIP header information to and from Unified CCE for modification in the routing scripts. This feature gives you greater flexibility in providing SIP interoperability with third-party SIP trunks and gateways. You can pass information only in the header of the initial SIP INVITE, not for reinvites.

Be careful when modifying SIP headers. The tools do not check the syntax when you add or modify SIP headers.

Expanded Call Context Variables

Expanded Call Context (ECC) variables enable you to set business relevant data for transfer to the agent desktop. Unlike the call variables, you can configure the size, format, and the name of each ECC variable.

You can define as many ECC variables as necessary. But, you can only pass 2000 bytes of ECC variables on a specific interface at any one time. To aid you in organizing ECC variables for specific purposes, the solution has *ECC payloads*.

An ECC payload is a defined set of ECC variables with a maximum size of 2000 bytes. You can create ECC payloads to suit the necessary information for a given operation. You can include a specific ECC variable in multiple ECC payloads. The particular ECC variables in a given ECC payload are called its *members*.

You can use several ECC payloads in the same call flow, but only one ECC payload has scope at a given moment.

The solution includes an ECC payload named "Default" for backward compatibility. If your solution does not require more ECC variable space, you only need the Default payload. If your solution only has the Default payload, the solution automatically adds any new ECC variables to the Default payload until it reaches the 2000-byte limit.

User-to-User Information

User-to-user information (UUI) is the data that ISDN Supplementary Services provides as user-to-user services. UUI is an industry-standard field that enables info transfer between the contact center enterprise solutions and third-party solutions. The UUI feature transfers information between the calling and the called ISDN numbers during call setup and call disconnect.

In Unified CVP, you can use the UUI feature during transfers and disconnects to pass ISDN data from the PSTN to the Unified CCE router. You can also use UUI from Unified CCE to third-party ACDs.

The gateways can use application-specific UUI data in CTI applications and for better third-party ACD integration.

For example, you can pass data from an external system (such as caller-entered digits from a third-party VRU) to Unified CCE on an incoming call.

Note Unified CVP does not yet support the IETF UUI header. You can use the generic SIP header functionality to parse the standard UUI.

Database Integration

You can integrate your contact center with an external database. Database integration provides create, update, and retrieve operations on tables in the external database. Database integration uses the Database Element in the CVP Call Studio.

Database Lookup

Database Lookup is an optional feature that allows you to read data from an external database and use that information within a routing script or administrative script.

Database Lookup is only supported by Packaged CCE 4000 agent and 12000 agent deployment.

For example, create a script that uses an external SQL database to lookup a caller's ANI and determine if the caller is a silver or gold customer.

You must designate a single key column as the SQL primary key. Use an If node to reference database columns accessed by the DB Lookup node. In this example, use the If node to determine if the caller is a silver or gold customer.

When the DB Lookup node is run, it attempts to query a row of data from the external database. If the node is run as a part of an admin script, it will be called at regular intervals to check for changes as scheduled. If the node is run as a part a routing script, it will be a database query from the DB Worker thread.

For details on how to create a database and use it in the script, see https://www.cisco.com/c/en/us/support/docs/customer-collaboration/unified-contact-center-enterprise/116215-configure-dblookup-00.html

Note If the external remote database is on SQL Server 2017 version, you have to install the ODBC Driver 17 manually on the server hosting the external database. Download the ODBC Driver 17 from Microsoft.

Extension Mobility

To monitor and control the phones, the contact center solutions associate phones with a JTAPI user ID in Unified CM. When you use Extension Mobility or Extension Mobility Cross Cluster, you can associate an Extension Mobility device profile instead. In a Unified CCE environment, you associate the IP phones or the corresponding Extension Mobility device profiles with Unified CCE JTAPI user IDs. When an agent desktop signs in, the PIM requests a subscriber to allow the PIM to begin monitoring and controlling that phone. Until the agent signs in, the subscriber does not allow Unified CCE to monitor or control that phone. If the device or the corresponding Extension Mobility device profile is not associated with a Unified CCE JTAPI user ID, then the agent sign-in request fails.

Using Extension Mobility Cross Cluster (EMCC), when a Unified CCE PIM phone registers to the local cluster after Extension Mobility sign in, the phone looks like an agent situated across a WAN. The Unified CCE peripheral manages the agent devices based on the Extension Mobility profile rather than on a phone device in the Application User on the cluster. For more information, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

You can associate Extension Mobility devices using two methods; either by device or by user profile. Associate the Extension Mobility profile to the CCE Application User on Unified Communications Manager.

Configuring the EM Profile, instead of the device, provides more flexibility in which phones agents can use in the call center. Configuring the phone device limits which devices the agents can use. The option that you use in a contact center depends on the customer business case.

Mixed Codecs

By default, the contact center enterprise solutions accept incoming calls using the mu-law codecs. Your contact center can use the a-law codec instead. To use a-law, change the default values in CVP, Unified CM, and your VXML or Ingress Gateways. This table lists the audio codec support for various functions.

Table 19: Audio Codec Support

Function	Support
Inbound calls	Both G.711 (mu-law and a-law) and G.729 codecs
Outbound calls	G.711 (mu-law and a-law) only
VRU	G.711 (mu-law and a-law) only
Agents	Both G.711 (mu-law and a-law) and G.729 codecs

Note In order to avoid transcoders and universal transcoders, use both G.711 and G.729 codecs for inbound calls and agents. Use G.729 as the first codec in your preference list to save bandwidth on the WAN.

Cisco Outbound Option Dialer

SIP Dialers with CUBE can support a-law and u-law with specific design considerations.

Silent Monitor Support

The following silent monitoring solutions support both mu-law and a-law:

- Unified CM-based Silent Monitoring

No Support for Mixed Environments

You cannot mix codec use between instances of the following elements:

- All Mobile Agents on a peripheral are required to use the same codec.
- All CVP prompts are required to use the same codec.

Mobile Agent

Mobile Agent enables an agent to sign in from anywhere with any PSTN phone and a broadband VPN connection for agent desktop communications. The agent functions just as an agent sitting in your contact center with a Cisco IP Phone. Mobile agent uses a pair of CTI ports which serve as proxies to connect the agent and the caller.

Note Mobile Agent cannot use IPv6-enabled CTI ports.

Each PG can support fewer Mobile Agents than general agents. But, you can add extra PGs to support up to the maximum active agents that are allowed in the Reference Designs.

Phone Extension Support

Your contact center enterprise solution can support both general phone extensions and ACD (contact center) phone extensions. How you combine these types can affect your contact center.

You can assign phone lines to Unified CM clusters as follows:

- You can mix general and ACD extensions in the same cluster.
- You can separate the ACD extensions into specific clusters and the general extensions into other clusters.

You can also assign each agent's phone extensions to their device in several ways.

Note

- Unified CCE supports E.164 dial plans and provides partial support for the '+' prefix.
- In 2000 Agent Reference Designs, a coresident Unified CM can support a maximum of 2000 phones. This includes your phones for all types of agents, whether contact center agents or back-office workers. If your solution requires more than 2000 phones, use an off-box Unified CM instead.

Dual-Use Unified CM Clusters

You can use the same Unified CM cluster to support general IP telephony (office) extensions and ACD (contact center) extensions. However, consider the following points before choosing a dual-use cluster:

- Contact centers have strict maintenance windows. Maintenance might affect office extensions at inopportune times.
- Agents process far more calls than other office workers. Their devices place a higher load on the system than an average office worker. A cluster serving only office extensions can support many more extensions.
- All devices are required to meet the compatibility requirements for the contact center solution. See your solution's *Compatability Matrix*.

Because of these points, separate clusters for each type of extension offer better performance.

Phone Extensions for Different User Types

You can assign extensions differently to each agent's device to match their needs.

Unified CCE supports only one agent ACD extension on the IP phone. To enable Unified CCE to manage and control all calls on that extension, it cannot have voice-mail or call forwarding defined. Typically, the agent extension is not used as the agent's office extension. You can assign a separate extension to the agent's phone for that purpose. The office extension can have voice-mail and other calling features.

Typically, the connection defaults to the first extension on an IP phone when you pick up the handset. You want that first extension assigned to the extension that each person uses most often. Consider the following configurations based on the person's duties:

- **Contact Center Agent**—Assign the agent's ACD extension to the first position and their office extension to another position. This layout makes answering inbound ACD calls easiest. The contact center tracks any calls the agent places on the ACD extension as external calls. When the agent places a call on that extension, Unified CCE puts the agent in not-ready mode and does not route calls to that agent.
- **Knowledge Worker**—These agents don't directly handle many ACD calls. Assign their office extension to the first position and their ACD extension to another position. This layout avoids the contact center tracking their non-ACD calls. Because these agents place most calls on their office extensions, they must manually set their state to not-ready mode for most calls. That mode prevents Unified CCE from routing ACD calls to them during that time.
- **Single-line Worker**—These agents use the same extension for their ACD and office calls. This option enables you to see all agent activity and to avoid all interruptions for the agent. However, this option requires special care in your routing scripts to prevent agent-to-agent calls from interrupting customer calls. The routing employs CTI Route Points and a unique DN for each CTI Route Point.

- **Back-Office Agents**—These agents typically only use their office extension. Assign their office extension to the first position. If a back-office agent occasionally handles ACD calls, assign their ACD extension to the last position on their IP phone.

Post Call Survey

A Post Call Survey takes place after the call. Typically, you use the survey to determine whether a customer was satisfied with the call experience. You configure a call flow that sends the call to a DNIS for the Post Call Survey after the agent disconnects from the caller.

Your VRU asks callers whether they want to participate in a Post Call Survey. If they choose to do so, they are automatically transferred to the Post Call Survey after the call flow completes.

Cisco Webex Experience Management

Cisco Webex Experience Management is the platform for Customer Experience Management (CEM), integrated with powerful tools that allow you to see your business from your customers' perspective. Experience Management has all the sophisticated features and functionality including customer journey mapping.

With Experience Management integrated with Packaged CCE:

- Administrators can configure post call surveys to collect feedback directly from customers.
- Administrators can configure analytical gadgets, which can be viewed on Finesse desktop.
- Agents and supervisors can view pulse of the customers through industry standard metrics such as NPS, CSAT, and CES or other KPIs.

Note Currently, you can have surveys only for inbound ICD calls.

For information on how to configure Experience Management, see the *Webex Experience Management* chapter in the *Cisco Unified Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cust_contact/contact_center/icm_enterprise/icm_enterprise_12_0_1/Configuration/Guide/ucce_b_ucce-features-guide-12.html

Precision Routing

Precision Routing is a routing feature in Unified CCE. Precision Routing enhances and can replace traditional routing.

Traditional routing maps all an agent's skills into a hierarchy of business needs. However, traditional routing is restricted by its single dimensional nature. Precision Routing provides multidimensional routing with simple configuration, scripting, and reporting. The feature records varying proficiencies in a skill, rather than just possession of the skill. These multiple attributes with proficiencies more accurately expose the capabilities of each agent. The greater accuracy in routing brings more value to the business.

You can use a combination of attributes to create multidimensional precision queues. Unified CCE scripting can dynamically map the precision queues to match a caller's needs with the best available agent.

For more information on Precision Routing, see the *Cisco Unified Contact Center Enterprise Features Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html.

Single Sign-on (SSO)

The Single Sign-on (SSO) feature authenticates and authorizes agent and supervisor access to the contact center solution applications and services. The authentication process validates the identity of a user: "you are who you say you are." The authorization process confirms that an authenticated user is permitted to perform the requested action: "you can do what you are asking to do." When you enable SSO in the contact center solution, users only sign in once to gain access to all their Cisco browser-based applications and services. Access to Cisco administrator applications is not available through SSO.

SSO requires the following:

- A third-party Identity Provider (IdP)
- A Cisco Identity Service (Cisco IdS) cluster

When an SSO-enabled user signs in, the Cisco IdS interacts first with your IdP to authenticate the user. When the user is authenticated, the Cisco IdS confirms with the accessed Cisco services to confirm that the user is authorized for the requested role. When the user is both authenticated and authorized, the Cisco IdS issues an access token that allows the user to access the application. The access token enables the user to switch between the authorized contact center applications for that session without presenting credentials again.

SAML 2.0 Authentication

SSO uses Security Assertion Markup Language (SAML) to exchange authentication details between an Identity Provider (IdP) and a service provider. The identity provider authenticates user credentials and issues SAML assertions, which are pieces of security information transferred from the identity provider to the service provider for user authentication. Each assertion is an XML document that contains trusted statements about a subject including, for example, username and privileges. SAML assertions are usually digitally signed to ensure their authenticity.

A generic SAML authentication flow consists of:

- Client - A browser-based user client used to access a service.
- Service Provider - An application or service the user tries accessing.
- Identity Provider - An entity performing the user authentication.

The identity provider keeps actual credentials and authentication mechanism hidden. Based on the authentication process result, the identity provider issues SAML assertions.

Elements Used in SAML 2.0

The following is the list of elements that are used in SSO SAML 2.0 authentication:

- Client (the user's client)—A browser-based client or a client that can leverage a browser instance for authentication. For example, a system administrator's browser.
- Lightweight Directory Access Protocol (LDAP) users—Users are integrated with an LDAP directory. For example, Microsoft Active Directory or OpenLDAP.

- Security Assertion Markup Language (SAML) assertion—An assertion is an XML document that contains trusted statements about a subject. For example, a username. SAML assertions are digitally signed to ensure their authenticity. It consists of pieces of security information that are transferred from Identity Providers (IdPs) to the service provider for user authentication.
- Service Provider (SP)—An application or service that trusts the SAML assertion and relies on the IdP to authenticate the users. For example, Cisco Identity Service (IdS).
- An Identity Provider (IdP) server—This is the entity that authenticates user credentials and issues SAML assertions.
- SAML Request—An authentication request that is generated by a Cisco Identity Service (IdS). To authenticate the LDAP user, IdS delegates an authentication request to the IdP.
- Circle of Trust (Co-T)—It consists of the various service providers that share and authenticate against one IdP in common.
- Metadata—An XML file generated by the Cisco IdS (for example, Cisco Identity Service Management) and an IdP. The exchange of SAML metadata builds a trust relationship between the IdP and the service provider.
- Assertion Consumer Service (ACS) URL—A URL that instructs the IdPs where to post SAML assertions.

Cisco Identity Service (IdS)

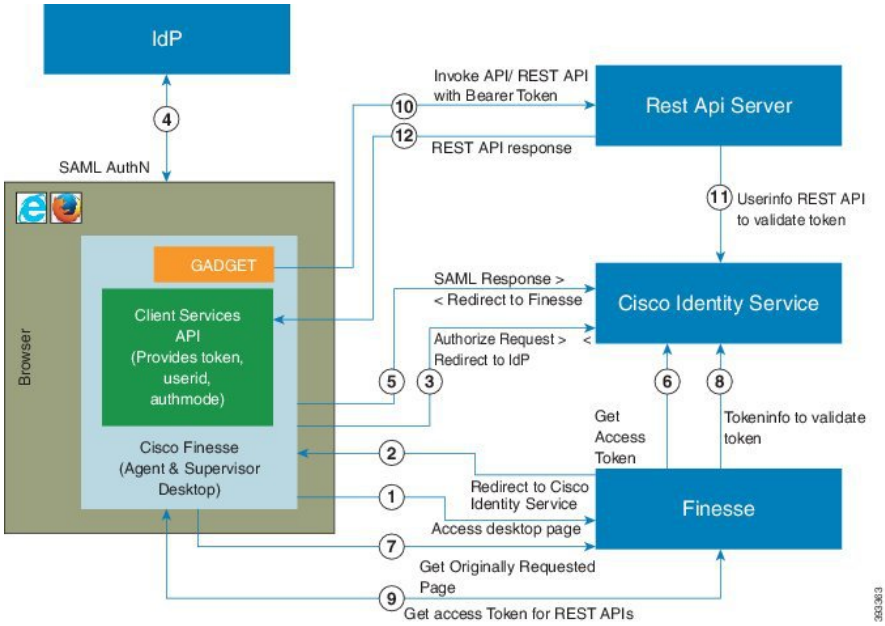
Authentication is managed for the contact center solution by the Cisco Identity Service (Cisco IdS). When an SSO-enabled user signs in, the Cisco IdS interacts first with the customer's Identity Provider (IdP) to authenticate the user. The IdP stores user profiles and provides authentication services to support SSO sign-ins. When the user is authenticated, the Cisco IdS exchanges information with the Cisco service the user is attempting to access to confirm that the user is authorized for the role they are requesting. When the user is both authenticated and authorized, the IdS issues an access token that allows the user to access the application. When the access is established during a particular session, the user can switch among contact center solution applications without presenting credentials again.

Authentication and Authorization Flow

The complete authentication and authorization flow has been simplified as:

- When you access an application with protected resources, the application will redirect you to the Cisco Identity Service for authentication. Cisco Identity Service leverages SAML and generates a SAMLRequest and redirects the browser to the Identity Provider.
- The browser authenticates directly against the Identity Provider. Applications are not involved in the authentication process and have no access to user credentials.
- The OAuth flow accesses the resource with a token which is then validated.
- Cisco Identity Service sends an authentication request through the browser to the identity provider.
- The user enters the login credentials to the identity provider for authentication. After the assertion is successful and the user attributes are read it will redirect to the original application that was accessed. Cisco Identity Service accompanied by an assertion that confirms successful authentication and includes user information and access rights for the web application.

Figure 34: Authentication and Authorization Flow



Whisper Announcement

Whisper Announcement plays a brief, prerecorded message to an agent just before the agent connects with each caller. The announcement plays only to the agent; the caller hears ringing while the announcement plays.

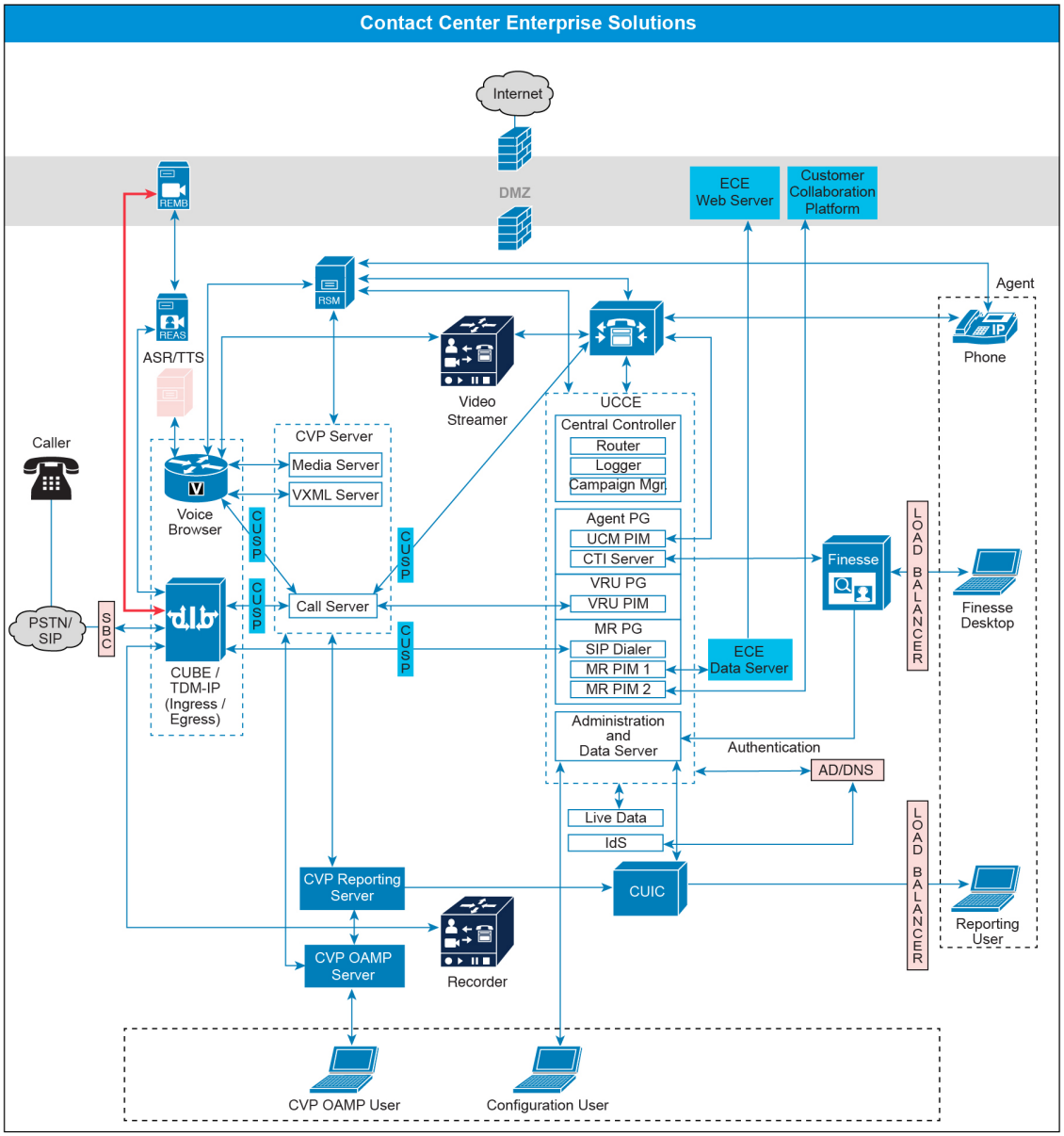
The announcement can contain information about the caller that helps prepare the agent to handle the call. The information can include caller language preference, choices the caller made from a menu (Sales, Service), customer status (Platinum, Gold, Regular), and so on.

After you enable Whisper Announcement, you specify which announcements to play in the call routing scripts. The script chooses which announcement to play based on various inputs. For example, different scripts might play for different dialed numbers, customer ID lookups in your customer database, or selections the caller made from a VRU menu.

Call Flows

Reference designs only supports Unified CVP comprehensive call flows. The comprehensive call flow includes VRU, queuing, and IP switching.

Figure 35: Logical Component Connectivity



Comprehensive

The Comprehensive call flow can route and transfer calls across your VoIP network. For example, you can use this model to offer VRU services, and to queue calls for routing to an agent. Callers reach a VRU initially. If they need help from an agent, their call receives queue treatment and transfers to an agent. You can also transfer calls between agents. Unified CVP and Unified CCE pass call data between these endpoints and provide reporting for all calls.

The Comprehensive call flow has the following features:

- Allows callers to access the contact center through local, long distance, or toll-free numbers terminating at the ingress voice gateways, and from VoIP endpoints.
- Provides VRU, including integrated self-service applications, queuing, and initial prompt and collect, and IP switching capabilities.
- Can route and queue calls to Unified CCE agents.
- Must use SIP.
- Provides the video VRU, video queuing, and video agent capabilities.
- Use an optional Unified CVP VXML Server.
- Prompt or collect data using optional ASR and TTS services.

Incoming Calls

Incoming calls can come from an outside carrier (either SIP or TDM) or an internal help desk. Congestion Control counts incoming calls against your CPS.

Note All new incoming calls always enter the Cisco IOS gateway (CUBE or TDM-IP gateway) and are associated with the Unified CVP survivability service.

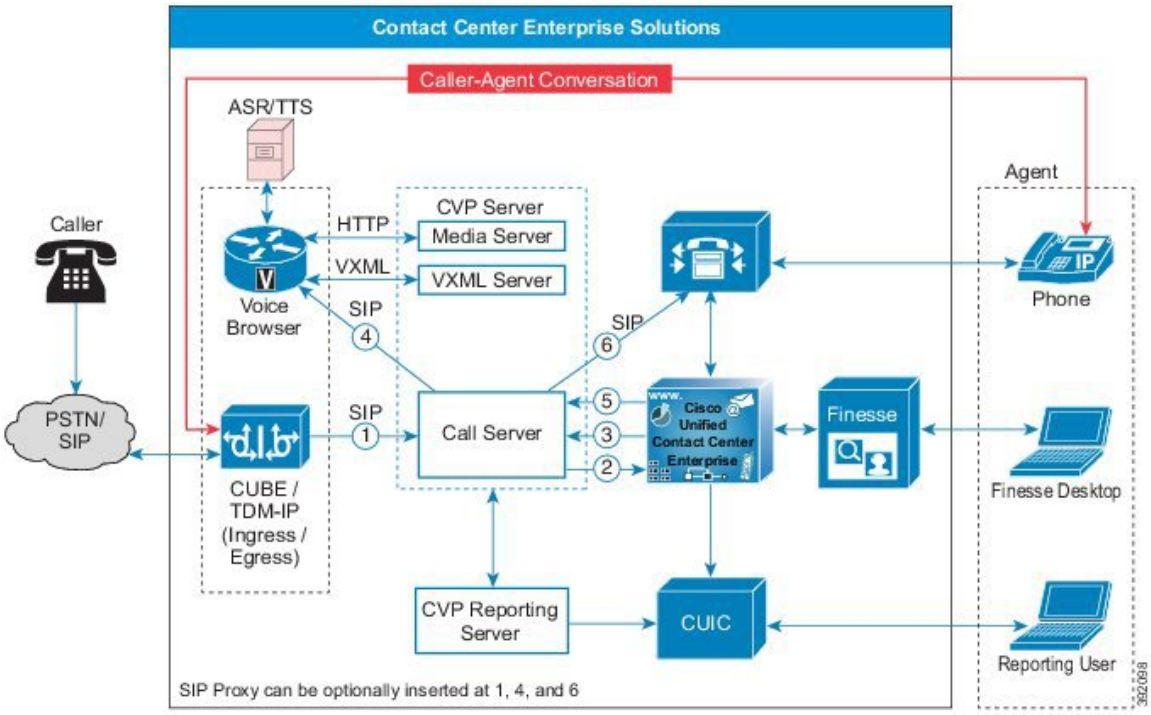
Incoming Calls from Carrier

The following table shows the basic SIP trunk or TDM-IP GW call flow.

Call Flow	Logical Call Routing
Incoming call from Carrier	**VRU:** Caller --> Carrier --> CUBE or TDM-IP GW--> Unified CVP --> Voice Browser **Agent:** Caller --> Carrier --> CUBE or TDM-IP GW--> Unified CVP -->Unified Communications Manager -> Agent 1 **Note** You can have calls front-ended by the carrier through a third-party SBC or Unified CM Session Management Edition (Unified CM SME). The incoming call flow in that solution is: Caller --> Unified CM SME (or an SBC) --> CUBE --> Unified CVP

The call flows in the following figure represent units of call flow functionality. You can combine these call flow units in any order during a call.

Figure 36: Basic Call Flow with VRU and Queue to an Agent



The call flow for an incoming call from the Carrier to a TDM Gateway or through the SBC to the CUBE gateway is as follows:

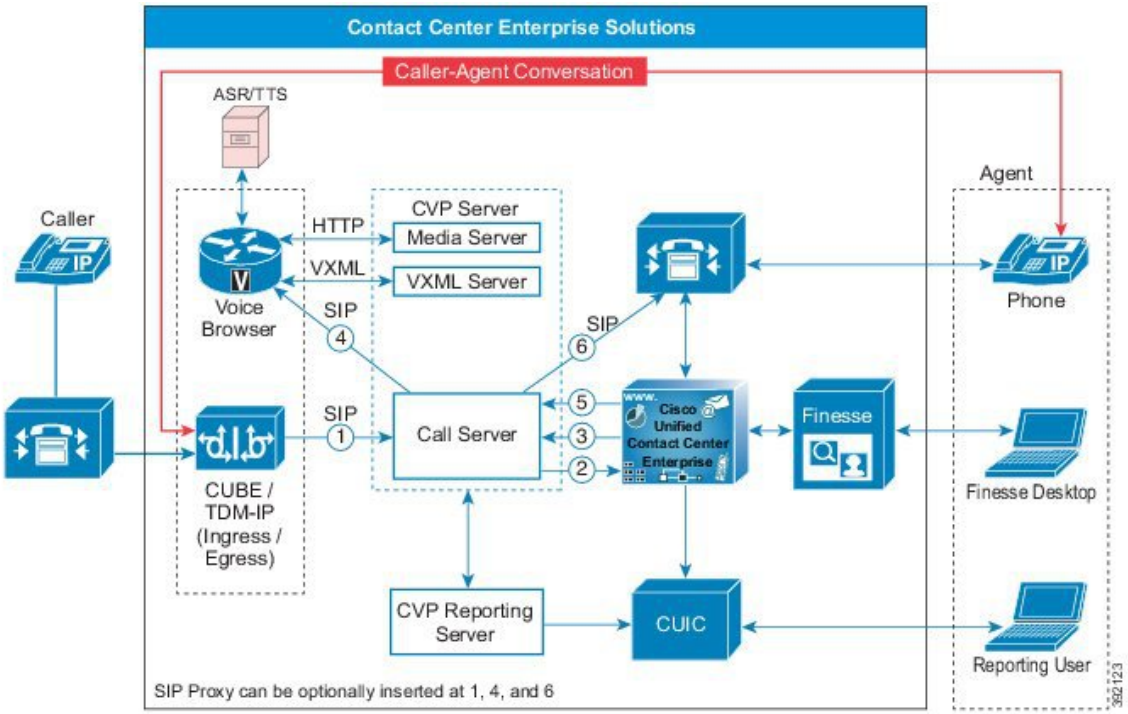
1. New incoming call from CUBE or TDM gateway to CVP.
2. New incoming call to Unified CCE from CVP.
3. Play "Hello World" Prompt.
4. CVP sends call to Voice Browser, and the caller hears the VRU.
5. When an agent is available, Unified CCE sends the agent number to CVP.
6. CVP sends the call to the agent phone through Unified CM.

Incoming Calls from Internal Help Desk

Enterprises that use IP phones can provide their employees with call-in self-service applications, for example, an application to sign up for health benefits. An employee might try to reach an agent, such as the IT help desk, and end up waiting in queue. Both of these scenarios result in calls originating from Unified CM to Unified CVP through CUBE.

Call Flow	Logical Call Routing
Incoming call from Unified Communications Manager (internal help desk)	**VRU:** Caller --> Unified CM --> CUBE(E) --> Unified CVP --> Voice Browser **Agent:** Caller --> Unified CM --> CUBE(E) --> Unified CVP --> Unified CM -> Agent1

Figure 37: Internal Help Desk Call Flow



The call flow for an incoming call from a phone that's registered with your Unified CM cluster:

1. New incoming call from an internal caller goes through CUBE or TDM gateway to CVP.
2. New incoming call to Unified CCE from CVP.
3. Play "Hello World" Prompt.
4. CVP sends call to Voice Browser, and the caller hears the VRU.
5. When an agent is available, Unified CCE sends the agent number to CVP.
6. CVP sends the call to the agent phone through Unified CM.

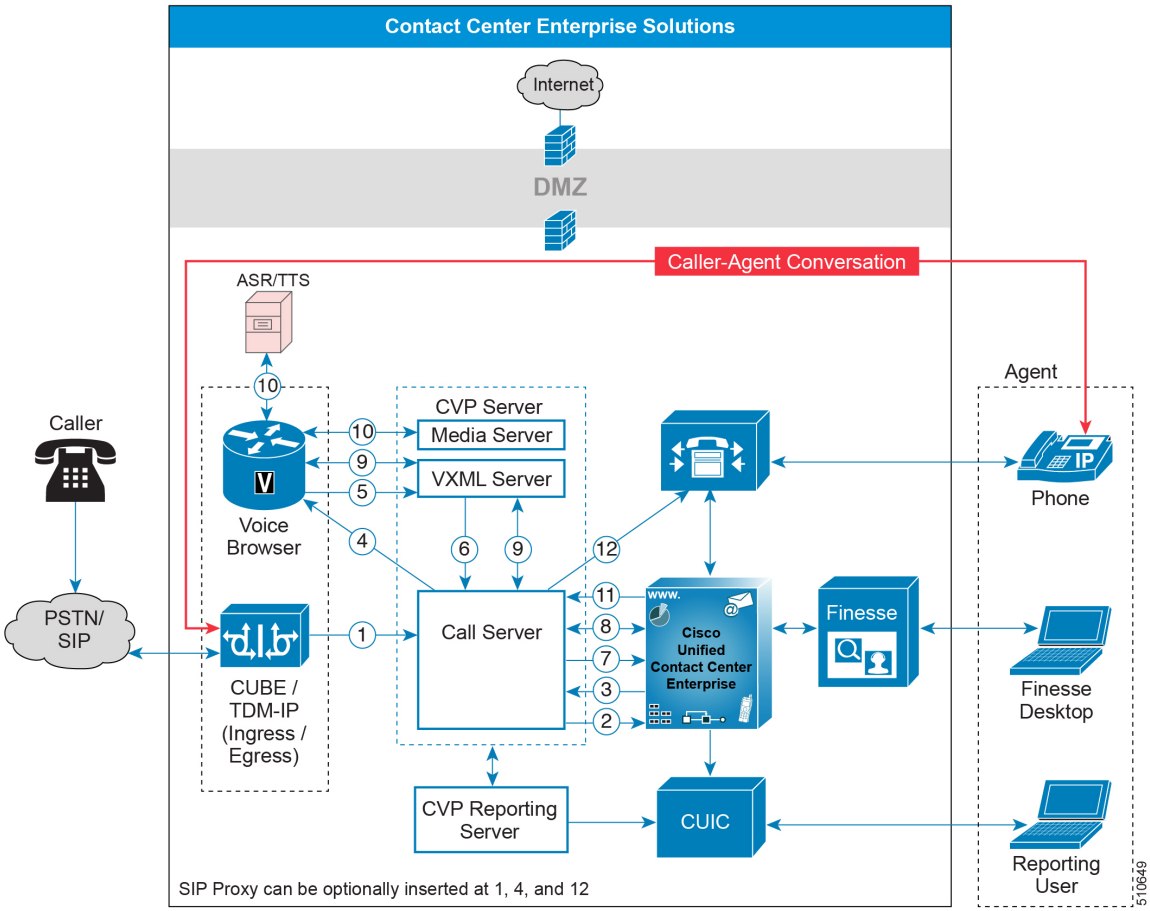
Note You can optionally insert the Cisco Unified SIP Proxy between the following core components:

- (CUBE or TDM-IP GW) to (CVP or Unified CM)
- CVP to (Voice Browser or Unified CM)
- Unified CM to (CUBE or TDM-IP GW or CVP)

Comprehensive with ICM Micro-Apps or CVP Call Studio Apps

When you use Micro-Applications or Call Studio applications, the call flow is as follows:

Figure 38: Detailed Call Flow for New Incoming Call



1. The new incoming call comes into a CUBE or a TDM-IP Gateway.
2. New incoming call to Unified CCE from CVP. The CVP Call Server sends a route request to Unified CCE through VRU PG. This route request for a DN invokes Unified CCE to run a routing script based on the DN and call type association.
3. The Unified CCE routing script uses either an implicit or explicit `Send to VRU` node to return a label to CVP Call Server. There is a pause in the script being run .

The label is a combination of the configured network VRU label for CVP and a random correlation id.

4. The CVP Call Server sends an SIP Invite message to the Voice Browser by translating the network VRU label to the browser's IP address. Optionally, this can pass through a SIP Proxy Server.
5. The Voice Browser sends an HTTP New Call message to the VXML Server with the network VRU label.
6. The VXML Server then sends the request to Call Server.
7. CVP Call Server then sends a request instruction message to Unified CCE, which then resumes the routing script.
8. The Unified CCE routing script uses `Run Script` nodes to instruct the CVP Call Server about the VRU treatment.

   Unified CCE can then send a `Run Script Request` message to run a VRU operation. The request can invoke the following:

   - **Micro-Application**—Use a Micro-Application for simple VRU operations. It supports basic operations like playing prompts and collecting digits. The Micro-Application is referenced in the Unified CCE Script and defined as part of a network VRU script.
   - **Call Studio Application**—Use a Call Studio Application for complex VRU call flows. You design it in the Call Studio Designer and deploy it in the VXML Server. You can then reference the application in a Unified CCE script.

9. The Call Server communicates with the VXML Server to invoke the specific application.

   Based on the Micro-Application or Studio Application, VXML Server generates the relevant VXML page. The Voice Browser renders the page to the caller. The VXML Server and Voice Browser communicate back and forth with each other until the end of the application.
10. The Voice Browser connects to one of the following services during the rendering of the VXML page:
    - For audio prompts, it connects over HTTP to the Media Server, which is coresident on the CVP Server.
    - For ASR/TTS, it establishes an MRCP connection with an external speech server to synthesize the text prompt or recognize a user speech for user input.
    - For video, it connects over SIP to an external server to play a video prompt.

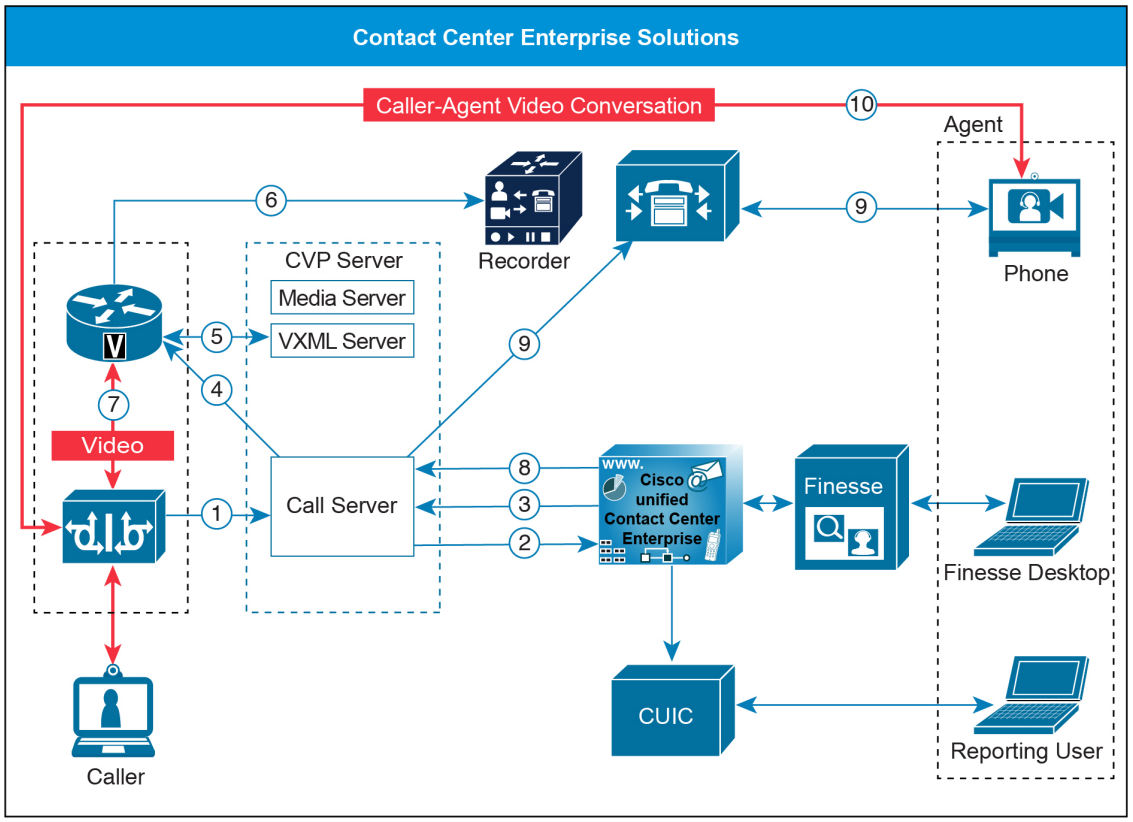
Note If there are any more applications to run, the call flow repeats Steps 8-10.

11. When an agent is available, Unified CCE sends the agent number to CVP. The VRU initiation stops once the Unified CCE script gets a Queuing node or Release node in the script. The SIP call leg with the Voice Browser terminates.
12. CVP sends the call to the agent phone through Unified CM.

Video Call Flow

This call flow runs through a video VRU before connecting to an agent.

Figure 39: Video Call Flow



1. Send a incoming call from Unified CM to CVP.
2. Send a incoming call from CVP to Unified CCE.
3. Play the CVP Studio video application.
4. CVP sends the call to CUBE and the VXML gateway.
5. CVP VXML Server instructs the VXML gateway to connect to DN XXXX.
6. CUBE sends the call to the Video Media Server with DN XXXX. Caller gets static video.
7. Agent is now available.
8. CVP sends the call to and agent.

Supplementary Services

Supplementary services include the following call flows:

Table 20: Supported System Call Flows

System Call Flows	Supported
Hold and Resume	Yes

System Call Flows	Supported
Consult Transfer and Conferences	Yes
Blind Transfer and Conferences	Yes
Router requery	Yes
Postroute using Unified CVP	Yes

Hold and Resume

Agents use Hold to suspend a call temporarily. If Music on Hold resources are available, the caller hears music while on hold. Otherwise, the caller hears a tone.

Multicast Music-on-Hold

As an alternative to the unicast Music-on-Hold (MOH), you can multicast MOH with supplementary services on Unified CM. You have these options when deploying MOH with this feature:

- With Unified CM multicasting the packets on the local LAN
- With the branch gateway multicasting on their local LAN

Use branch gateway multicasting when you have configured survivable remote site telephony (SRST) on the gateway. This method enables the deployment to use MOH locally and avoid MOH streaming over the WAN link.

Note For information about configuring MOH on the Call Manager Enterprise (CME), see https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucme/admin/configuration/manual/cmeadm/cmemoh.html#wpmkr1022205.

Transfers and Conferences

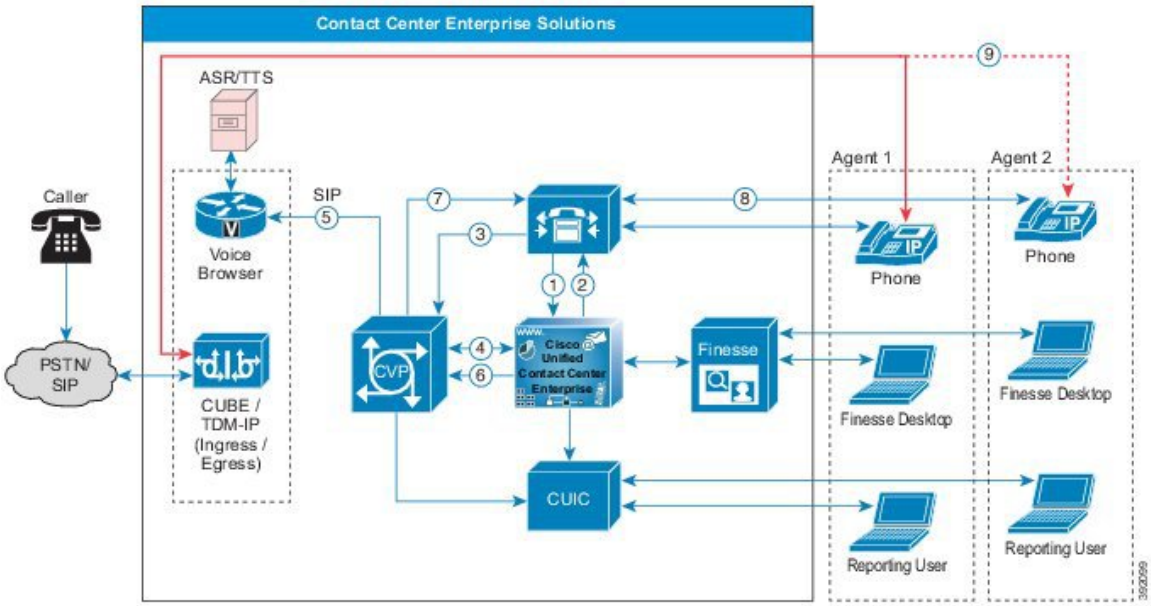
In most contact center solutions, agents can transfer calls to or start conferences with other agents. There are two ways to transfer or start a conference:

- Blind
- Consult (also known as a warm transfer)

Blind Transfers and Conferences

In a blind transfer, the first agent dials a number and ends the call. The caller then gets connected to the second agent or placed into a queue if necessary. This type of transfer does not involve a call originated by Unified CM.

Figure 40: Blind Transfer Call Flow with VRU and Queue to a Second Agent



1. Agent 1 begins a blind transfer request, an incoming call form Unified CM to Unified CCE.
2. Agent 2 is unavailable, which sends the call to the VRU.
3. Unified CM sends the call to Unified CVP.
4. Unified CCE instructs CVP to connect to the Voice Browser to play VRU or queue music.
5. Unified CVP sends the call to the Voice Browser. The caller hears the VRU or queue music.
6. When Agent 2 is available, Unified CCE sends the agent number to CVP.
7. Unified CVP sends a SIP call to Agent 2 through Unified CM. The VRU or queue music disconnects.
8. Unified CM sends the call to Agent 2 and the call data appears on the Cisco Finesse desktop.
9. The caller talks to Agent 2.

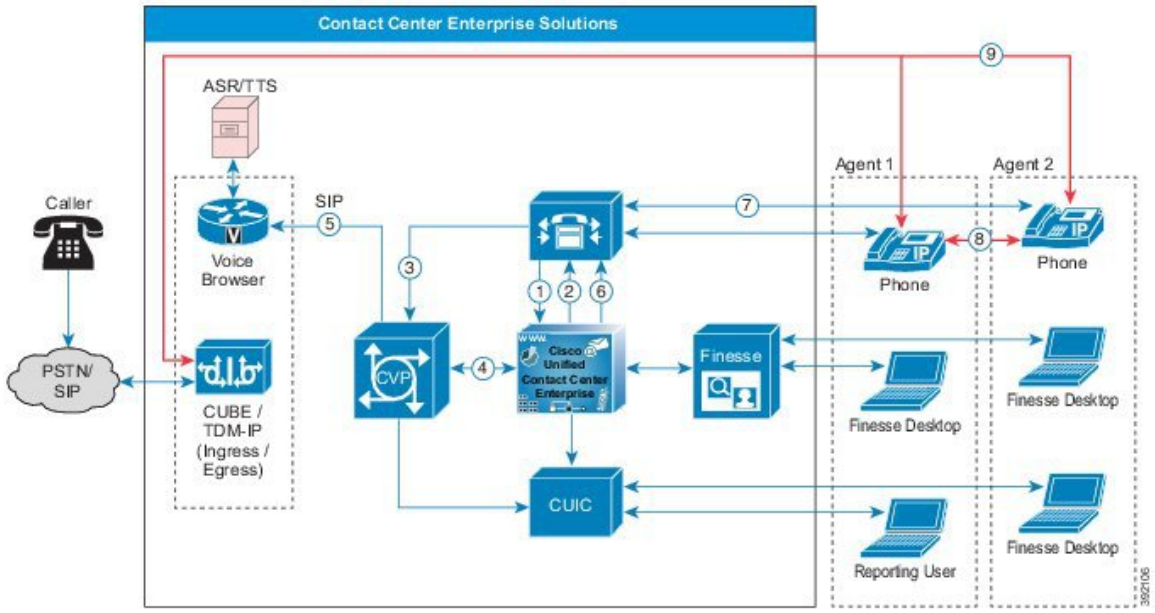
Consult Transfers and Conferences

In a warm transfer or conference, the agent dials a number and is connected to the second agent while the caller is placed on hold. The two agents can talk, then they can conference in the caller, and the first agent drops off. If the second agent is not available, the first agent (not the caller) is placed into a queue. All of this processing can take place without involving Unified CVP, unless the first agent gets queued. In that case, the first agent's call is transferred to Unified CVP, which creates a call originated by Unified CM.

Table 21: SIP Trunk Call Flow

Call Flow	Logical Call Routing
Post routed call from agent-to-agent	**VRU:** Agent 1 --> Unified CM --> Unified CVP --> Voice Browser **Agent:** Agent 1 --> Unified CM --> Unified CVP --> Unified CM--> Agent 2

Figure 41: Consult Call Flow with VRU and Queue to a Second Agent



1. Agent 1 begins a consult transfer request, an incoming call from Unified CM to Unified CCE.
2. Agent 2 is unavailable, which sends the call to the VRU.
3. Unified CM sends the call to Unified CVP.
4. Unified CCE instructs CVP to connect to the Voice Browser to play VRU or queue music.
5. While Agent 1 waits, they get treated with the VRU. The agent hears the VRU or queue music and the caller gets the Music on Hold (MOH).
6. When Agent 2 is available, Unified CCE sends the agent number to Unified CM.
7. Unified CM sends a SIP calls to Agent 2. The VRU disconnects.
8. Agent 1 consults with Agent 2.
9. Agent 1 completes the transfer. The caller speaks with Agent 2 and Agent 1 drops off.

Note Conference call flows are the same as consult call flows. Both conference call flows and consult call flows conference the call with the agents, rather than holding them during consult. Hold and Resume, Alternate and Reconnect, and Consult and Conference call flows invoke the session initiation protocol (SIP) ReINVITE procedure to move the media streams. A Conference to VRU call flow is similar to conference with no agent available call flow.

SIP Refer Transfer

In some scenarios, Unified CVP transfers a call to a SIP destination and does not have Unified ICM and Unified CVP retain any ability for further call control. Unified CVP can perform a SIP Refer transfer, which allows Unified CVP to remove itself from the call, and free licensed Unified CVP ports. The Ingress Voice Gateway port remains in use until the caller or the terminating equipment releases the call. SIP Refer transfers are used in both Comprehensive and Call Director deployments.

Invoke a SIP Refer transfer by any of the following methods:

- Unified ICM sends Unified CVP a routing label with a format of rfXXXX (For example, rf5551000).
- An application-controlled alternative is to set an ECC variable (user.sip.refertransfer) to the value **y** in the Unified ICM script, and then sends that variable to Unified CVP.

Note Direct Refer transfer using label works only if **Send To VRU** node is used before the Refer.

You can invoke the SIP Refer transfer after Unified CVP queue treatment has been provided to a caller. SIP Refer transfers can be made to Cisco Unified Communications Manager or other SIP endpoints, such as a SIP-enabled ACD.

Router requery on a failed SIP Refer transfer is supported using SIP with the Unified CVP, but only on calls where the survivability service is not handling the SIP Refer request.

Network Transfer

Unified CVP allows Network Transfer to transfer calls to another destination after an agent answers them.

There are two flags in Unified ICM to control the Network Transfer:

- NetworkTransferEnabled—This flag is part of the Unified ICM script. When enabled, it instructs the Unified ICM to save the information about the initial routing client (the routing client that sent the NewCall route request).
- NetworkTransferPreferred—This flag is enabled on the Unified CVP Peripheral Gateway configuration. When enabled, any route request from this routing client sends the route response to the initial routing client instead of the routing client that sent the route request.

The following points explain how you can do a network transfer:

- You can use Network Transfer to perform a blind transfer only from agent 1 to agent 2 through Unified CVP. In this case, Unified CCE instructs Unified CVP to route the contact back from Agent 1, and then route it either to a Voice Browser (for VRU treatment) or to another destination (for example, to Agent 2).

- You cannot use Network Transfer to perform a warm transfer or conference with Unified CVP. The call leg to Agent 1 must be active while Agent 1 performs a consultation or conference. Unified CVP cannot route the contact back from Agent 1 during the warm transfer or conference.

If a caller dials the same number regardless of a blind transfer, warm transfer, or conference, then perform the following tasks:

- Do not enable the NetworkTransferEnable flag in the Unified ICM script.
- Dial the CTI Route Point of the same Unified CCE Peripheral Gateway for any transfer or conference request to preserve the call context during the transfer. Dialing the Route Pattern or CTI Route Point of another Peripheral Gateway does not preserve the call context.
- Use SendToVru as the first node in the Unified ICM routing script.

Note Extra ports are used during the consultation, blind transfer, or conference calls. They are released after the originating consultation is terminated.

Requery and Survivability

Router requery allows the rerouting of calls due to any network failure connections. For example, Ring No Answer, Busy, and Network Unreachable trigger router requery. Only the QUEUE node and Label node in Unified CCE scripts support router requery. Define the rerouting logic in the script based on the error path from these nodes.

Call survivability on CVP runs on the ingress gateway. It triggers the survivability action when CVP detects any downstream failures. Based on the routing parameters for the survivability, you can have a failure trigger actions like a call restart or sending the calls to the local SRST phones.

Topologies

Cisco Unified Contact Center Enterprise (Unified CCE) is a solution that delivers intelligent call routing, network-to-desktop Computer Telephony Integration (CTI), and multichannel contact management over an IP network to contact center agents. Unified CCE adds software to create an IP automatic call distribution (ACD) onto a Cisco Unified Communications framework. This unified solution allows companies to rapidly deploy an advanced, distributed contact center infrastructure.

You can configure Unified CCE to sort customer contacts. Unified CCE monitors resource availability and delivers each contact to the most appropriate resource in the enterprise. The system profiles each customer contact using related data such as dialed number and calling line ID, caller-entered digits, data submitted on a web form, and information obtained from a customer database lookup. Simultaneously, the system monitors the resources available in the contact center to meet customer needs, including agent skills and availability, voice-response-unit (VRU) status, and queue lengths.

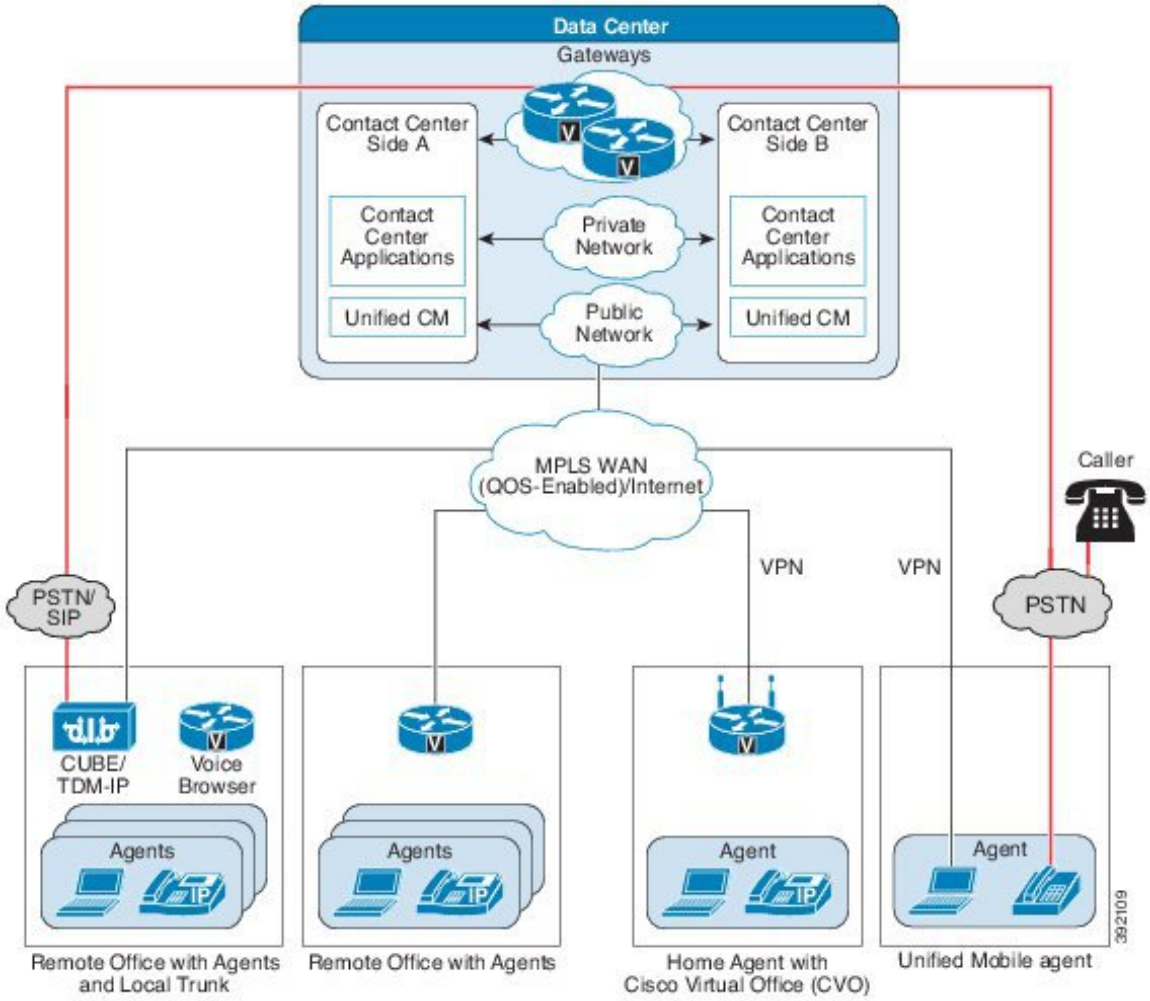
Unified CCE allows you to smoothly integrate inbound and outbound voice applications with internet applications such as real-time chat, web collaboration, and email. This integration enables a single agent to support multiple interactions simultaneously regardless of which communications channel the customer chooses.

The Unified CCE base model includes a common set of features that apply across supported Unified CCE models.

Contact Center Enterprise Architecture

The following figure shows the logical view of the contact center enterprise topology. Agents that are local to the `site` are not shown.

Figure 42: Contact Center Enterprise Solution Topology and Remote Office Options

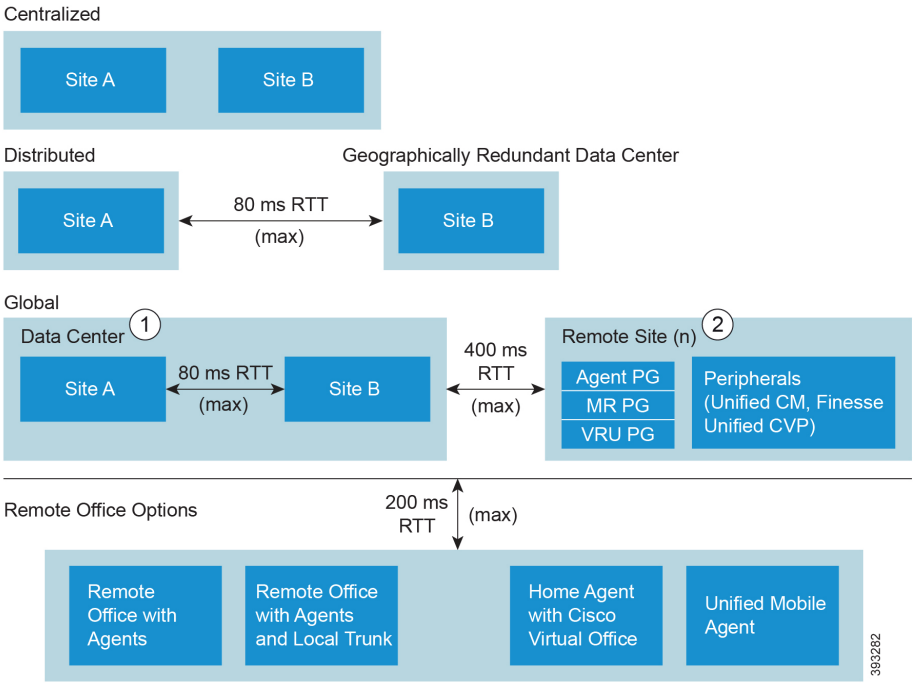


Topology Types

There are three topology models for contact center enterprise solutions:

- **Centralized Deployments**—Servers colocated in single main site
- **Distributed Deployments**—Servers distributed across different geographic sites
- **Global Deployments**—Remote Peripheral Gateway (PG) and peripheral

Figure 43: Topologies



Note

1. The Main Site can use either a Centralized or a Distributed topology.
2. A Remote Site can be geographically colocated with the Data Center. You can have up to 150 Remote Sites

Centralized Deployments

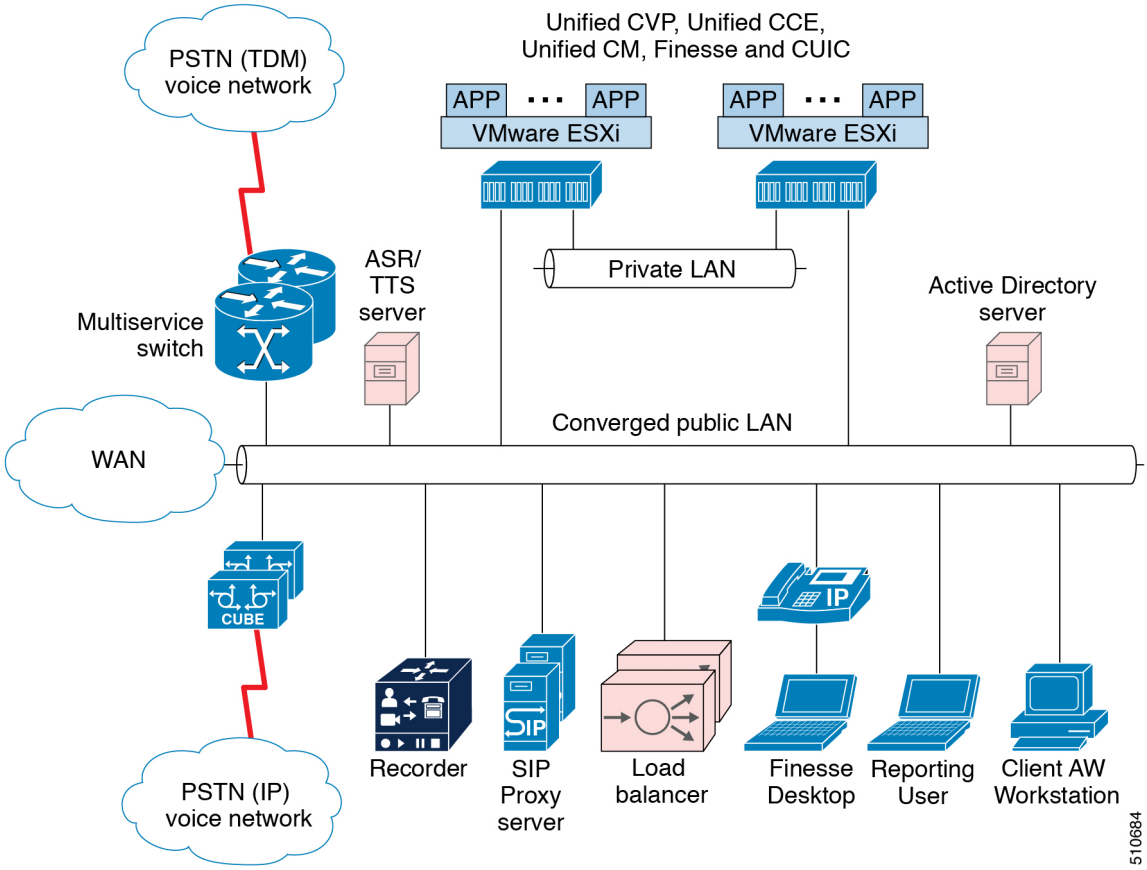
A centralized site can contain all the Unified CCE base model components. In a centralized data center, the agents, supervisors, and administrators are local to the data center. A centralized site can also include multiple agent locations.

In the local agent deployment scenario, the agents, supervisors, and administrators are local to the site.

Local Agent Architecture

The following figure shows the physical view of a local agent.

Figure 44: Local Agent—Physical View



Local Agent Components

The local agent deployment scenario includes the following components in addition to the core solution components:

- Unified Intelligence Center browser clients for local access to reporting
- Administration tools, such as, Unified CCE configuration tools, Internet Script Editor, or the local Administrative Workstation
- Optional third-party recording server for VoIP capture of agent or customer calls
- Agent phones with Built-In Bridge (BIB) to support features like Silent Monitoring.

Local Agent Benefits

The local agent deployment scenario provides the following benefits:

- Does not require location-based call admission control
- Simple codec setup

Local Agent Design Requirements

The following table describes the design requirements for a local agent.

Table 22: Local Agent Design Requirements

	Requirement	Notes
Infrastructure	Location-based call admission control is not required	Local agents use LAN bandwidth, which is typically sufficient for all Unified CCE traffic.
Desktop	Cisco Finesse Customer Relationship Management	
Codec	Transcoding is not required.	If all agents are local to the data center (no required WAN connectivity), you do not need to use G.729 or any other compressed RTP stream.
Recording	Unified CM-based BIB Unified CM Network Based Recording with Cisco Unified Border Element and the recorder. The Unified CM NBR feature allows for setting preference and fallback of CM controller media-forking at the originating Cisco Unified Border Element or the IP Phone's BIB.	By default, you can only record all agents constantly. Selective recording requires extra integration work.
Silent Monitoring	Unified CM-based BIB	

The following table describes the media resources for a local agent.

Table 23: Local Agent Media Resources

Resource	Method	Notes
Music on Hold	Unicast Unified Communications Manager	
Conference bridges	IP phone with BIB Hardware-based, located at voice gateways	
Media Termination Points	Not supported	
Transcoders	Hardware-based, located at voice gateways	Required for SIP trunks with a-law.

Distributed Deployments

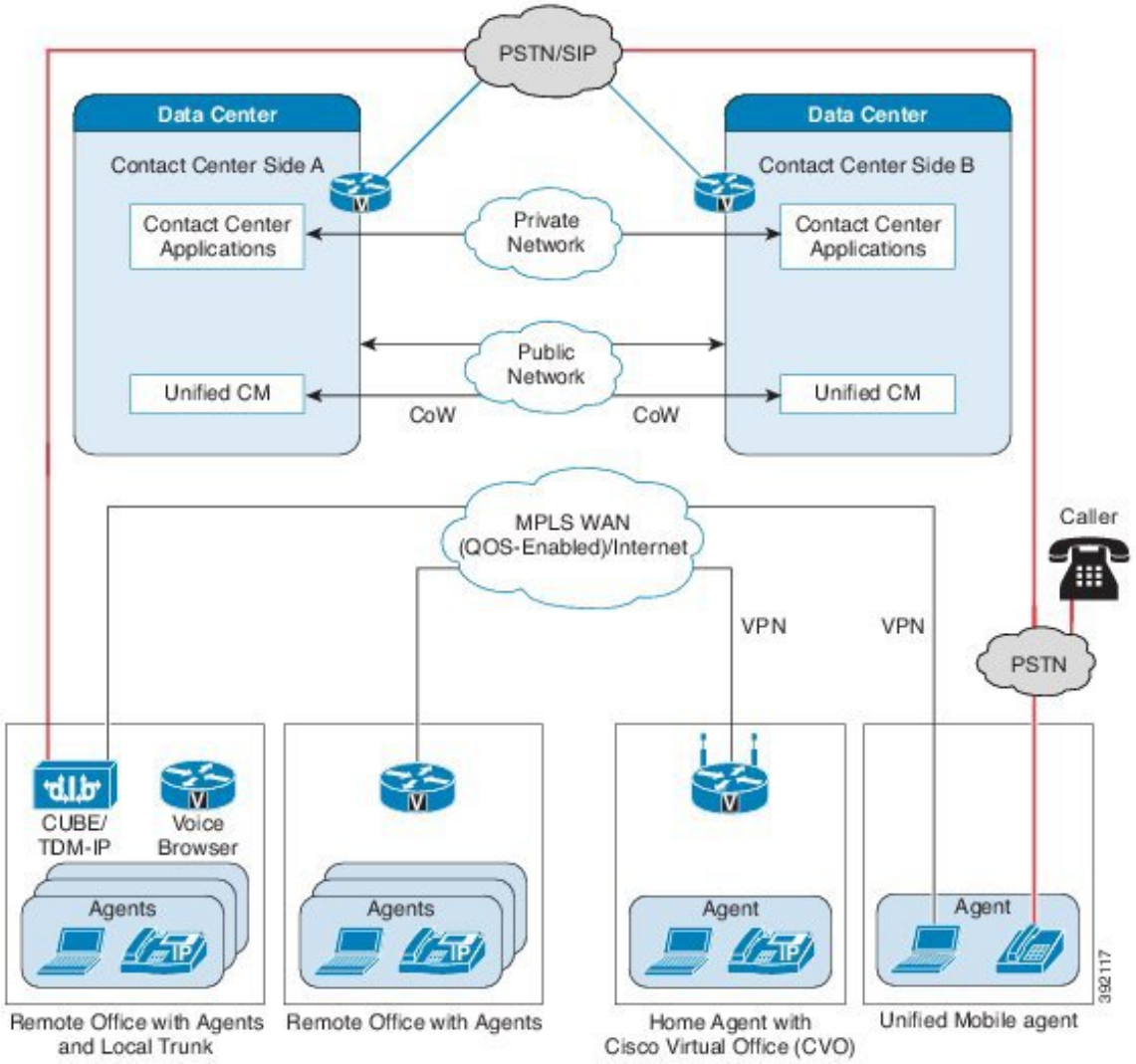
Globalization, security, and disaster recovery considerations are driving business to diversify locations across multiple regions. In addition, organizations want to distribute workloads between servers, share network resources effectively, and increase the availability of critical applications. Geographically redundant sites split critical applications across two data centers. Enterprises deploy geographically redundant sites to minimize planned or unplanned downtime and share data across regions.

Geographically redundant sites have a load balancer in each data center.

Clustering Over the WAN

The following figure shows geographically redundant sites with clustering over the WAN.

Figure 45: Geographically Redundant Sites with Clustering over WAN



Geographically redundant sites provide clustering over the WAN, distributed Unified Communications Manager clusters, and 1:1 redundancy for Unified CVP, SIP proxy, voice gateways, and Cisco Unified Intelligence Center.

Latency requirements across the high-availability (HA) WAN must meet the current Cisco Unified Communications requirements for clustering over the WAN. Unified CM allows a maximum latency of 40 ms one way (80-ms round trip).

Keep the public and private traffic on separate routes within the network and respect standard latency and bandwidth. Use independent physical circuits for the public and the private traffic.

Global Deployments

Global Deployments enable the Service Provider to deploy a single contact center available worldwide with a centralized main site and global access. This reduces deployment costs by eliminating multiple customer instances.

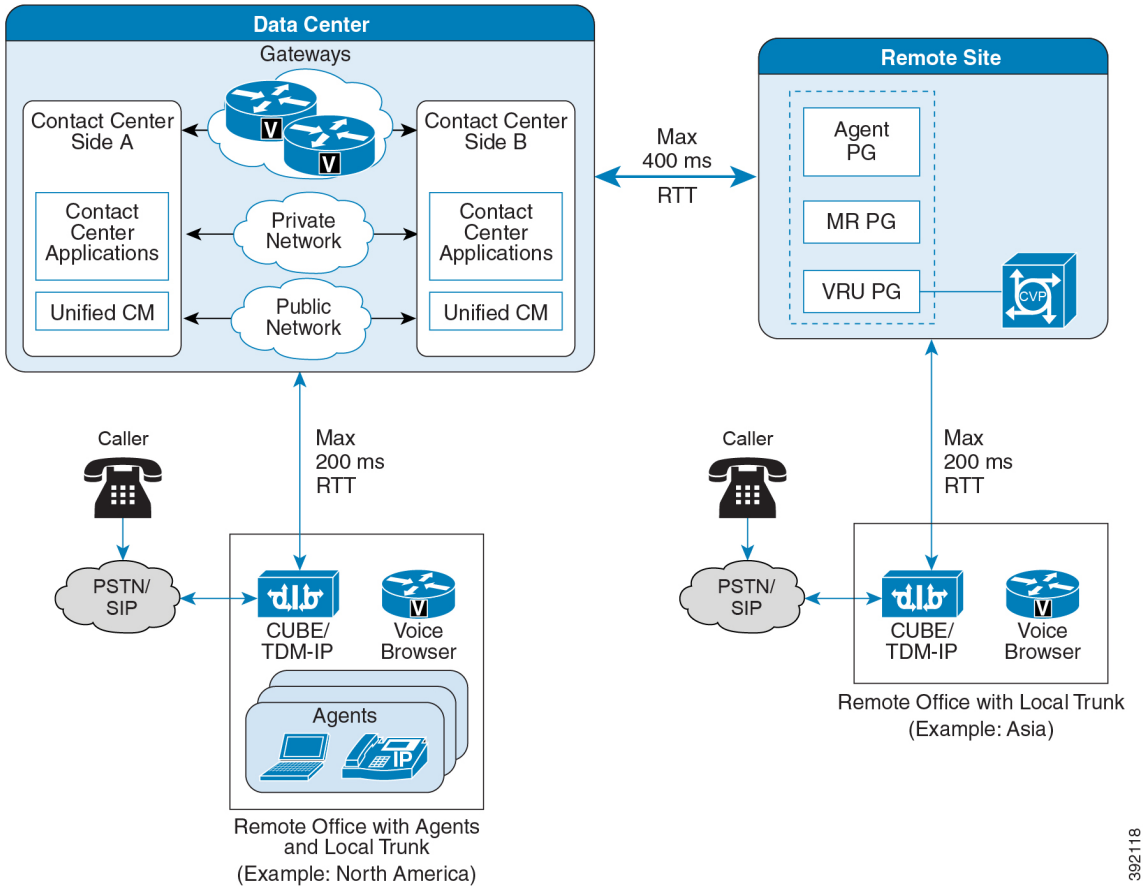
You can locate the Unified CM in a centralized or remote site or a customer premise. The following global deployment topologies are supported:

- Remote CVP deployment
- Remote Unified CM deployment
- Remote CVP and Unified CM deployment
- Remote MR PG deployment with multichannel options

Remote CVP Deployment

The topology shown in the illustration shows a simple example of Remote CVP deployment. In certain cases, contact center enterprise solutions use this topology for widely distributed sites. This topology provides global access to a centralized main site. This deployment requires extra Unified CVP servers with Unified CCE VRU PG Servers at remote sites. The maximum RTT with the central controller over the WAN is 400 ms.

Figure 46: Remote CVP Deployment Topology

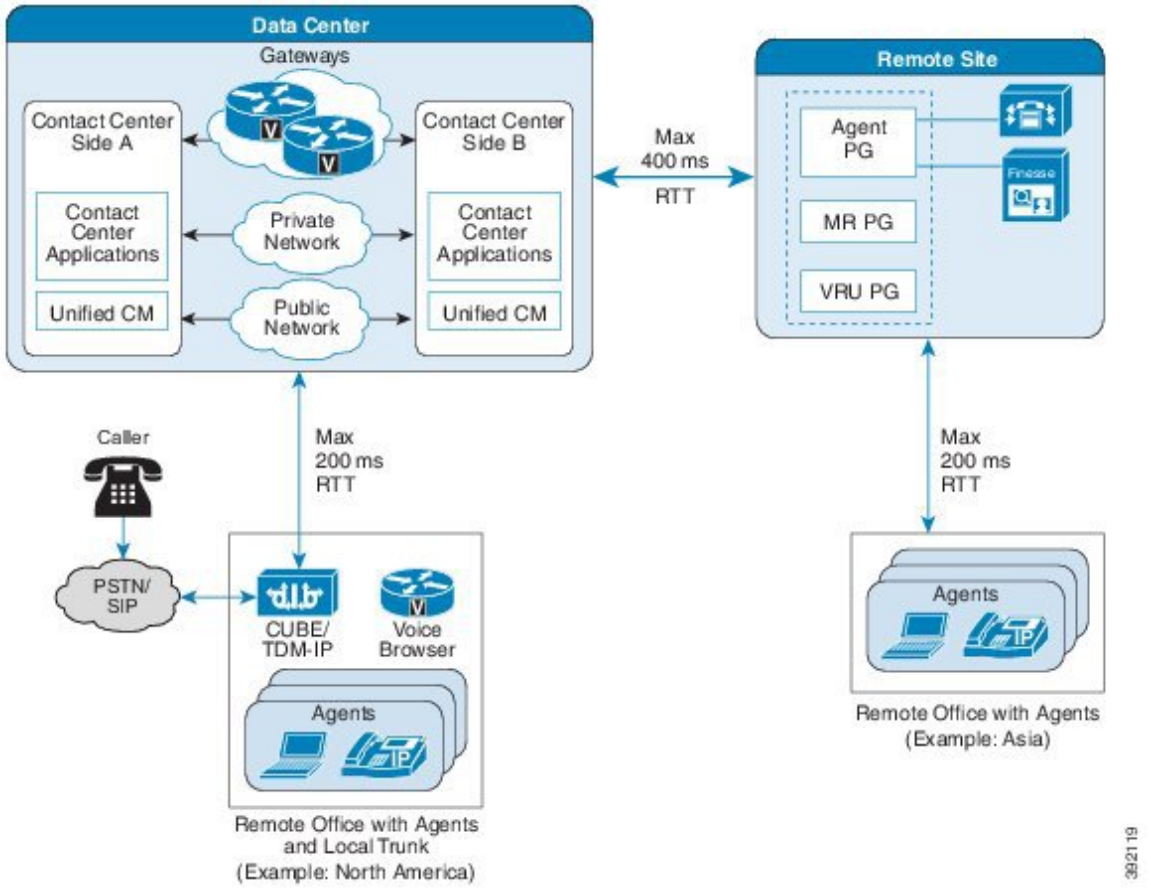


Remote Unified CM Deployment

If you have a remote office with agents, gateways, and Unified Communications Manager clusters, the Unified Communications clusters at the sites are typically independent. In this distributed call processing model, each site has its own Unified Communications cluster, with its own agents and PG pairs.

The following figure shows three Unified Communications Manager clusters. The remote office has a WAN connection back to the main site. Each Unified Communications Manager cluster is independent, with its own agents and PG pairs. Each site uses subscribers that are local to the site because JTAPI is not supported over the WAN. For example, site A cannot use the subscribers in site B. The Unified CCE central controller, Unified Intelligence Center, load balancer, SIP proxy server, and Unified CVP are located in the main site. TDM and VXML voice gateways are located at the remote office with local PSTN trunks.

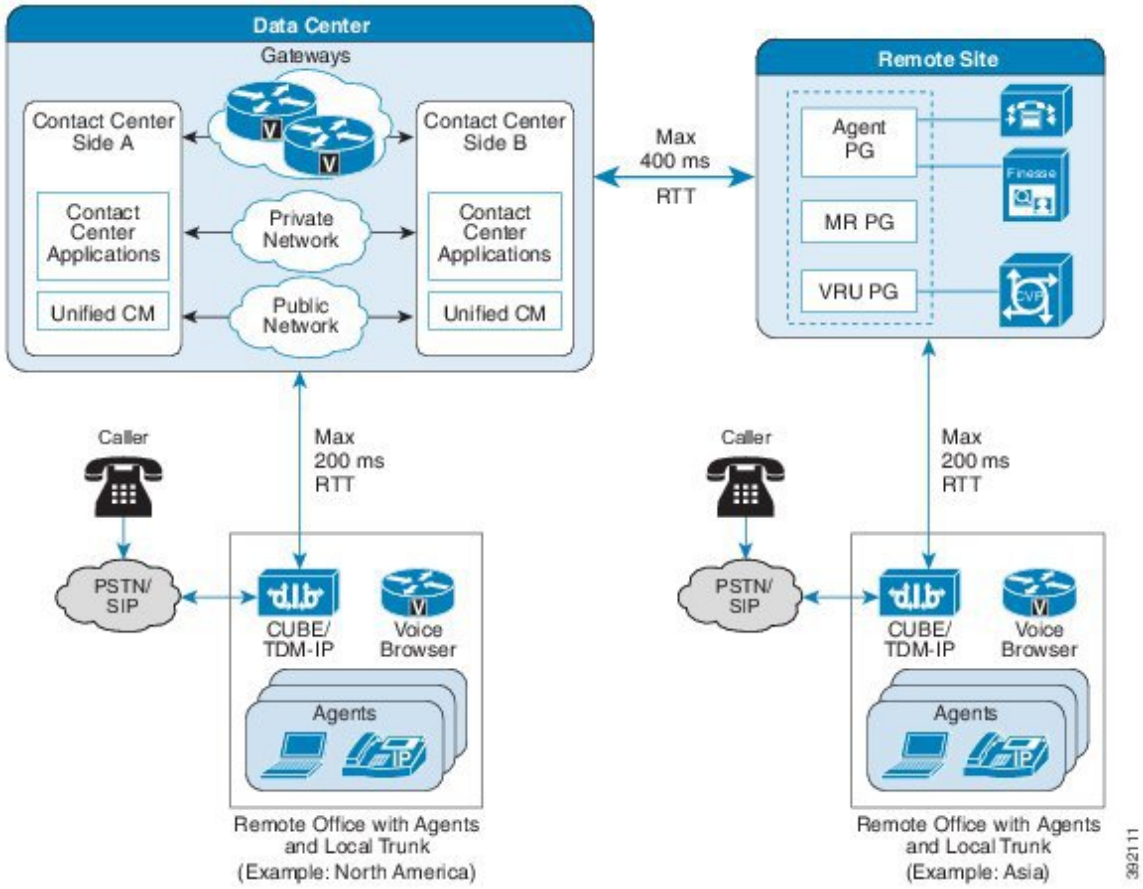
Figure 47: Remote Unified Communications Manager Clusters Topology



Remote CVP and Unified CM Deployment

The topology shown in the illustration shows a simple example of Remote CVP deployment. This deployment requires extra Unified CVP and Unified CM servers with Unified CCE Generic PG Servers at remote sites. The maximum RTT with central controller over the WAN is restricted up to 400ms.

Figure 48: Global Deployment Topology



Remote Office Options

Remote agent support provides Computer Telephony Integration (CTI), contact distribution, and reporting capabilities to remote agents in branch offices or at home, through either a broadband network connection or their home phone line. Unified CCE provides identical user interfaces and feature functions to agents regardless of agent location.

The Unified Mobile Agent feature gives the contact center the flexibility to adapt to a fast-moving mobile workforce. Agents can choose their destination phone number during sign-in time and change the number as often as they want. Agents can be on any phone device on any third-party switch infrastructure.

Unified CCE remote office features help companies to use existing and on-demand resources and fully extend CTI functions across the extended enterprise.

Remote office options include:

- Office with Unified CCE agents
- Office with agent and a local trunk
- Cisco Virtual Office
- Mobile Agent

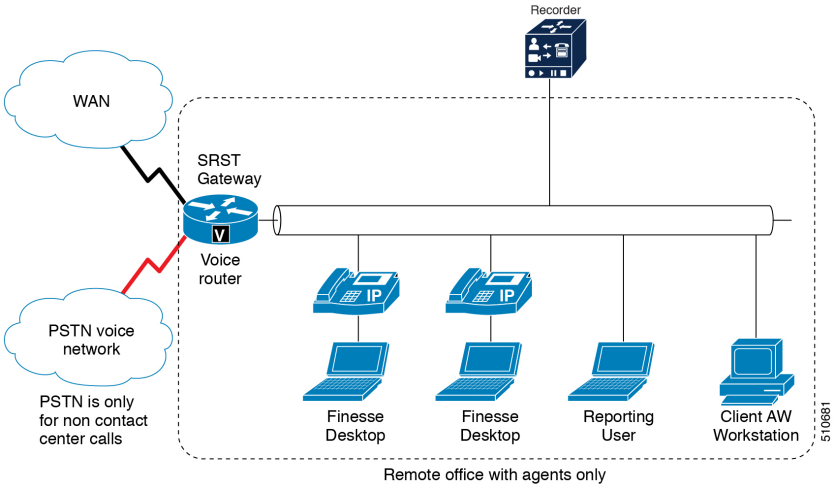
Remote Office with Agents

A remote office with agents is located either at the central office or at a branch office.

Remote Office with Agents

The following figure shows the physical view of a remote office with agents.

Figure 49: Remote Office with Agents—Physical View



Remote Office with Agents Components

A remote office with agents includes the following components:

- Unified Intelligence Center browser clients for local access to reporting
- Administration tools: Unified CCE configuration tools, Internet Script Editor, or the local Administrative Workstation
- Agent phones with BIB for Unified CM-based Silent Monitoring support

Remote Office with Agents Benefits

A remote office with agents provides the following benefits:

- Requires only a small data switch and router, IP phones, and agent desktops at remote sites for a few agents.
- Requires only limited system and network management skills at remote sites.
- Small remote sites and offices do not require PSTN trunks.
- PSTN trunks for incoming traffic connect to main site for efficiency.
- Unified CCE queue points (Unified CVP) are aggregated for efficiency.
- Does not use VoIP WAN bandwidth while calls queue. Calls extend over the WAN only when an agent is available for the caller.

Remote Office with Agents Design Requirements

The following table describes the design requirements for a remote office with agents.

Table 24: Remote Office with Agents Design Requirements

	Requirement	Notes
Infrastructure	Location-based call admission control	A failure of Unified CM location-based call admission control results in a disconnected routed call. Allow for adequate bandwidth to the remote sites and design a Quality of Service WAN.
	Bandwidth	Plan bandwidth capacity for the following traffic: • RTP (caller to agent) • Unified CM signaling to IP phones • Client desktop to PG (CTI data) • ISE client to ISE server • Administration Client • Unified Intelligence Center client to Unified Intelligence Center server • Silent Monitoring RTP • Recording RTP (if there is no recording server in the remote office) • Music on Hold traffic for calls that are on hold when you use Unified CM Unicast Music on Hold • Live Data **Note** Adequate bandwidth and QoS provisioning are critical for client desktop to PG links.
	Customer contact numbers	Customers might need to dial a long-distance number rather than a local PSTN number to reach the central office. You can offer customers a toll-free number, but the contact center incurs toll-free charges.
Desktop	Cisco Finesse Customer Relationship Management	
Codec	G.711 or G.729a	G.711 requires more bandwidth than G.729a.
Recording	BIB Network-based Recording	Audio forking requires Unified Border Element.

	Requirement	Notes
Silent Monitoring	Unified CM-based BIB	

The following table describes the media resources for a remote office with agents.

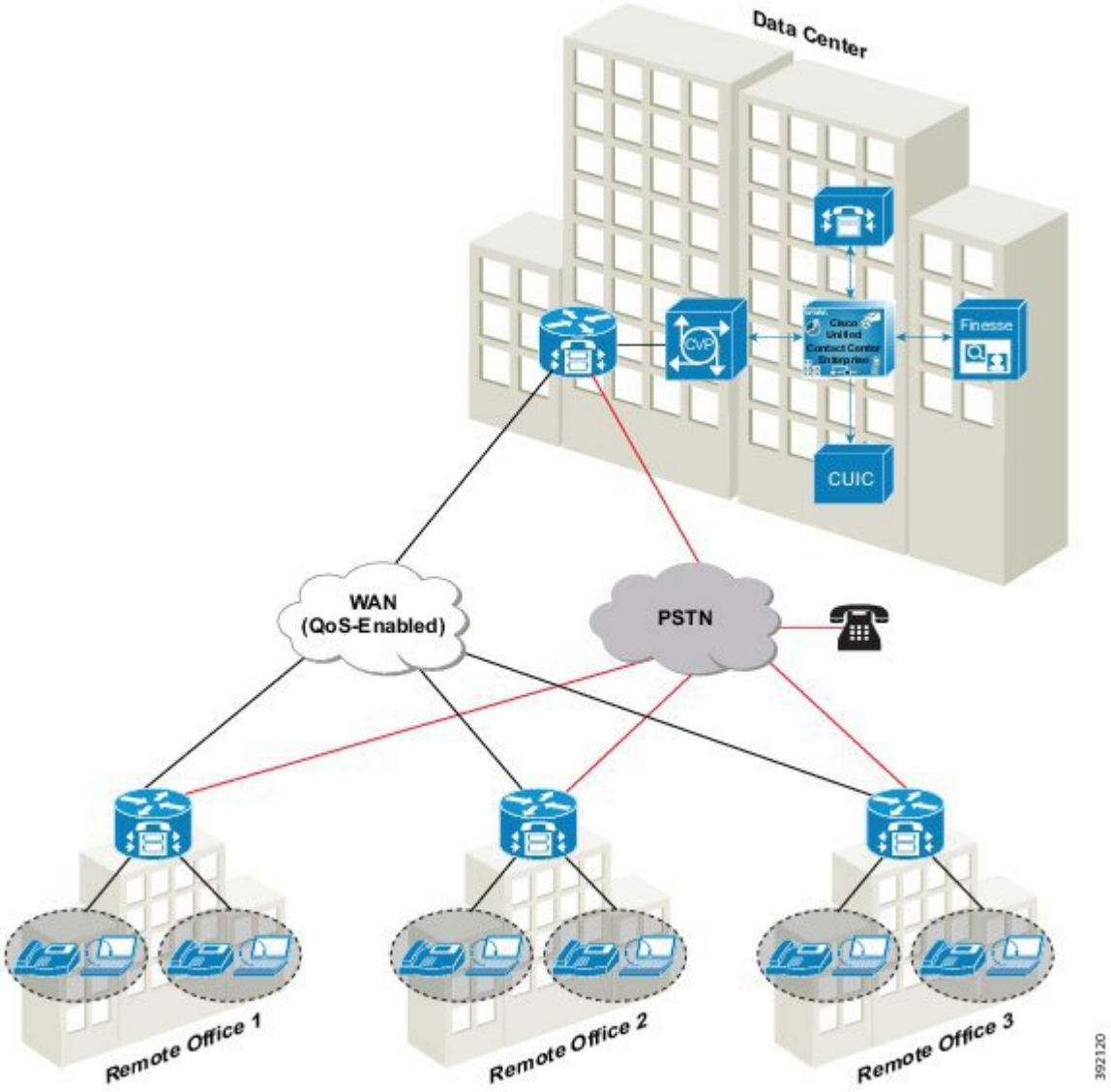
Table 25: Remote Office with Agents Media Resources

Resource	Method	Notes
Music on Hold	Unicast using Unified CM	
Conference bridges	Hardware-based, located at voice gateways	Conference bridges use local Unified Survivable Remote Site Telephony (SRST).
Media Termination Points	Hardware-based, located at voice gateways	For Unified Mobile Agents, MTPs are required only at the main site.
Transcoders	Hardware-based, located at voice gateways	Transcoders use local Unified SRST.

Remote Office with Agents and a Local Trunk

Use the remote office with agents and voice gateway deployment for contact centers with sites that each require local PSTN trunks for incoming calls. This deployment provides local PSTN connectivity for local calling and access to local emergency services.

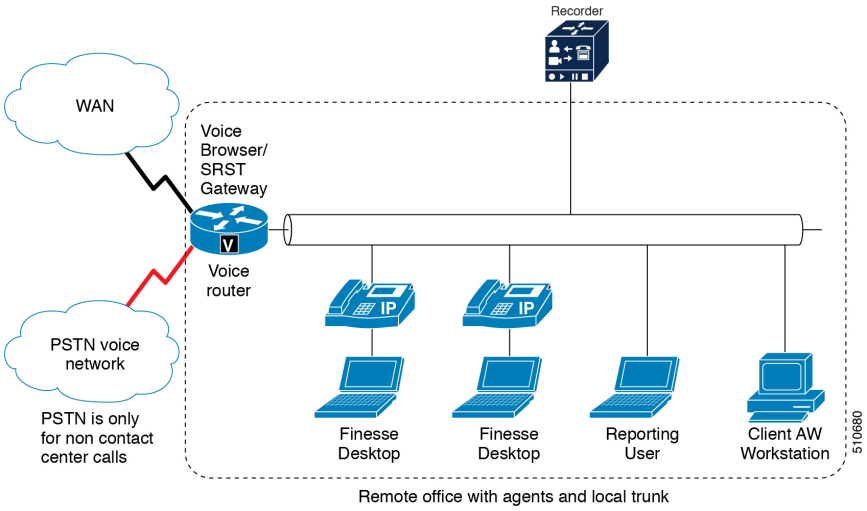
Figure 50: Remote Offices with Agents and Local Trunks



Remote Office with Agents and Voice Gateway Architecture

The following figure shows the physical view of a remote office with agents and voice gateway.

Figure 51: Remote Office with Agents and Voice Gateway—Physical View



Remote Office with Agents and Voice Gateway Components

A remote office with agents and voice gateway includes the following components:

- Integrated Services Router (ISR) voice gateway for ingress voice customer calls under Unified CVP control with local PSTN. Unified SRST backup requires trunks.
- Unified Intelligence Center browser clients for local access to reporting.
- Administration tools: Unified CCMP browser clients, Internet Script Editor, or the local Administrative Workstation.
- Optional third-party recording server for VoIP capture of agent or customer calls.
- Agent phones with BIB for Unified CM-based Silent Monitoring support.

Remote Office with Agents and Voice Gateway Benefits

A remote office with agents and voice gateway provides the following benefits:

- Requires only limited systems management skills for remote sites because most servers, equipment, and system configurations are managed from a centralized location.
- Does not require WAN RTP traffic for calls that arrive at the remote site and agents handle there.
- Unified CVP uses the VXML browser in Cisco IOS on the voice gateway to provide call treatment and queueing at the remote site. This call treatment and queueing eliminate the need to move the call over the VoIP WAN to a central queue and treatment point. VVB can provide the same capability locally.

Remote Office with Agents and Voice Gateway Design Requirements

The following table describes the design requirements for a remote office with agents and voice gateway.

Table 26: Remote Office with Agents and Voice Gateway Design Requirements

	Requirement	Notes
Infrastructure	Location-based call admission control	A failure in Unified CM location-based call admission control results in a disconnected routed call. Allow for adequate bandwidth to the remote sites and design a QoS WAN.
	Bandwidth	Plan bandwidth capacity for the following traffic: • RTP for calls transferred to other remote offices, or if calls are not restricted to the remote office where the calls arrive. • Unified CM signaling to IP phones • Client desktop to PG (CTI data) • Unified Intelligence Center client to Unified Intelligence Center server • Silent Monitoring RTP • Recording RTP (if a recording server is not located in the remote office) • Voice Browser (VXML documents and VXML file retrieval) • Music on Hold for calls that are on hold when you use Unified CM Unicast Music on Hold • ISE client to server • Administration client to the Administration Server and Real-Time Data Server • Live Data
Desktop	Cisco Finesse Customer Relationship Management	
Codec	G.711 or G.729a	G.711 requires more bandwidth than G.729a.

	Requirement	Notes
Recording	BIB	Audio forking requires a Unified Border Element.
Silent Monitoring	Unified CM-based BIB	

The following table describes the media resources for a remote office with agents and voice gateway.

Table 27: Remote Office with Agents and Voice Gateway Media Resources

Resources	Method	Notes
Music on Hold	Unicast using Unified CM	
Conference bridges	Hardware-based, located at voice gateways	Conference bridges use local Unified SRST.
Media Termination Points	Hardware-based, located at voice gateways	For Unified Mobile Agents, MTPs are required only at the main site.
Transcoders	Hardware-based, located at voice gateways	Transcoders use local Unified SRST.

Call Admission Control Considerations

Call admission control can be considered as a solution and not just a Unified CVP component. These considerations are most evident in the distributed branch office model where there are other voice services, such as Unified CM, sharing the same gateways with Unified CVP and the amount of bandwidth between the sites is limited. Be sure that, call admission control methods are in place on the network so that the same call admission control method is used for all the calls traversing the WAN from that site. If two call admission control methods can admit four calls each and the WAN link can handle only four calls, then it is possible for both call admission control entities to admit four calls onto the WAN simultaneously. This control method impairs the voice quality. If a single call admission method cannot be implemented, then each call admission control method must have bandwidth allocated to it. This situation is not desirable because it leads to inefficient bandwidth overprovisioning.

Two call admission control methods can be used in a Unified CVP environment: Unified CM Locations and Unified CM RSVP Agent. In a single-site deployment, call admission control is not necessary.

Unified CM performs call admission by assigning devices to certain locations and track of the number of calls that are active between these locations. Unified CM tracks the bandwidth that is used and, depending on the codec, can determine the number of calls.

Unified CM Call Administration Control

If Unified CM sends or receives calls from Unified CVP and there are Unified CVP gateways and IP phone agents collocated at remote sites, it is important to understand the call flows in order to design and configure call admission control correctly.

Resource Reservation Protocol

Resource Reservation Protocol (RSVP) is used for Call Admission Control, and it is used by the routers in the network to reserve bandwidth for calls. RSVP is not qualified for call control signaling through the Unified CVP Call Server in SIP. The solution for CAC is to use the Locations configuration on Unified CVP and in Unified CM.

Call Admission Control Deployment

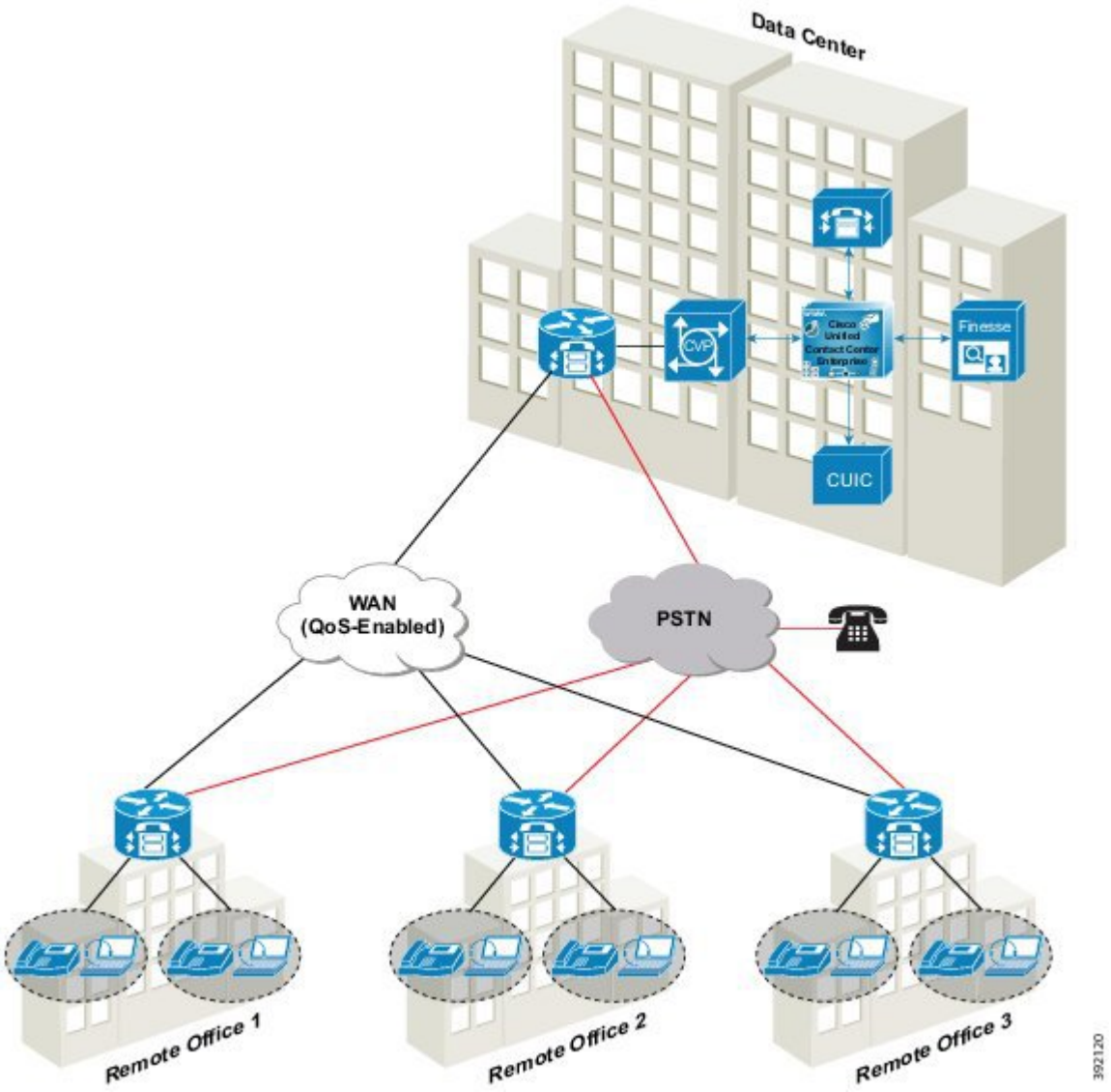
Call admission control is the function for determining if there is enough bandwidth available on the network to carry an RTP stream. Unified CM can use its own locations function or RSVP to track bandwidth between the Ingress Gateway and destination IP phone locations.

In networks, Resource Reservation Protocol (RSVP) is a protocol used for call admission control, and it is used by the routers in the network to reserve bandwidth for calls. RSVP is not qualified for call control signaling through the Unified CVP Call Server in SIP. As an alternative, the solution for Call Admission Control is to employ locations configuration on Unified CVP and in Unified CM.

Queue-at-the-Edge Branch Office Deployment

The following figure illustrates a typical branch office deployment.

Figure 52: Typical Branch Office Deployment.



You can deploy Unified CVP in a single cluster Unified CM deployment to provide queue-at-the-edge functionality. In this deployment, use branch-located Ingress Gateways to give callers access by local phone numbers rather than centralized or nongeographic numbers. This consideration is especially important in international deployments spanning multiple countries. The goal of this deployment is to first route the calls locally to an agent available in the branch office, if possible. This keeps the media streams local.

You locate the Egress Gateways at the branches to provide either localized PSTN breakout or integration of decentralized TDM platforms (ACDs) into the solution. Apart from the gateways, all other CVP subcomponents are at the main site. WAN links provide data connectivity from each branch location to the main site. (Although the media server is centrally located, commonly used VRU media is cached at the local branch.)

In this deployment, the branch office only has an Ingress Gateway (optionally acting as a Voice Browser also), IP phones for agents, IPT phones, and agent desktops.

You can configure Unified CCE Skill Groups, dial plans, and routing priorities so that incoming calls at each branch preferentially connect to agents at the same branch. Then, the RTP traffic flows directly from the Ingress Gateway to the IP phone. The RTP traffic does not need to traverse the WAN (although signaling and data might traverse the WAN).

If a local agent is not available, only the call gets routed to a remote agent over the WAN link. The originating call and the initial VRU treatment are still done locally.

In a WAN link failure, the CVP survivability application running on the POTS dial-peer for TDM originated calls can still route incoming calls locally.

Enhanced Location Call Admission Control Feature

ELCAC Concepts

The following definitions are important to the ELCAC feature:

- **Phantom Location**—A default location with unlimited bandwidth used when calculating calls that are hairpinned over a SIP trunk. You also use a phantom location when the SIP call is queued at the local branch to enable correct bandwidth calculations. Assign the phantom location to the gateway or trunk for CVP.
- **Location Routing Code**—The Location Routing Code is a string of numbers that Unified CVP appends to the label it receives from Unified ICM. Depending on the Location Routing Code, configure the dial plan to route the call to a destination, like the branch Voice Browser or Egress Gateway, or a Unified CM node. You can append the Location Routing Code at the front of the label, between label and the correlation ID., or not at all. This configuration is separate from the Unified CM location configuration, and is specific to Unified CVP. The Location Routing Code indicates the real location of the call and enables you to deduct the bandwidth from the correct location. A Location Routing Code is unique across multiple Unified CM clusters. Multiple Location Routing Codes can still route to the same branch office (if needed) by mapping the unique Location Routing Codes to same branch gateways in proxy routes.
- **Shadow Location**—This new location is used for intercluster trunks between two Cisco Unified Communications Manager clusters. This location is not used as intercluster ELCAC is not supported in Unified CVP.

Locations are created in Unified CM. Unified CVP gets these locations when you synchronize the location information from the Unified CM on Packaged CCE. You can associate a Location Routing Code for these locations on Packaged CCE and then associate your sites and gateways to these locations. Packaged CCE also enables you to create new locations. Based on this configuration, CVP creates two hash objects. One hash would map location to a Location Routing Code and the second hash would store mapping of GW IP address

to location name and Location Routing Code. These hash objects enable routing the call to appropriate GW to provide edge queuing (using Location Routing Code). They also pass around the location information on the call legs for Unified CM to do proper CAC calculations.

For branch office deployments, the following considerations apply:

- Control the number of calls that goes over the WAN link to branch offices based on the available bandwidth of the WAN link.
- For the queue-at-the-edge functionality, route the call originating from a specific branch office to a local Voice Browser on priority.

For Unified CVP intracluster Enhanced Location CAC, control the number of calls that go over the WAN link to branch offices. The decision to admit calls is based on the CAC computations, which represent the bandwidth used by the call. These computations are valid whether the calls are IP calls between two phones within Cisco Unified Communications Manager, calls over SIP trunks, or calls originated from TDM-IP Gateway.

For queue-at-the-edge functionality, the call originating from a specific branch office must be routed to a local Voice Browser based on priority. That is, always choose a local branch agent if possible.

Unified CVP supports topology modeling with Enhanced Location Call Admission Control (ELCAC) for intracluster. It does not support intercluster Enhanced Location CAC. Location Bandwidth Manager is enabled for intracluster CAC, but disabled for intercluster CAC. For more information on ELCAC topology modeling, see the Cisco Unified Communications SRND based on Cisco Unified Communications Manager, available at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

Comparison of Enhanced Location Call Admission Control Feature

The Enhanced Location Call Admission Control (ELCAC) feature addresses two important issues with the prior CAC feature:

1. Bandwidth miscalculations in CAC with IP originated callers and with any post transfers from agents.
2. Inability to deterministically select a local Voice Browser for VRU treatment at the branch office. This occurs during warm transfers from an agent when there is no correlation between the two calls at consult.

With ELCAC, because the location information is in the call leg, calls are routed to a local gateway on agent transfer. On the VRU leg, CVP appends the site id to the call based on this location. The SIP Proxy can then use the site id to get to the local gateway.

Router Requery with ELCAC

When Unified CM rejects a call for insufficient bandwidth, a SIP message 488, Not Acceptable Here, is returned to Unified CVP. The message triggers a router requery over the GED-125 interface to the VRU peripheral. The Unified CCE Router can return another agent label if requery is configured properly.

Design Considerations

The following considerations apply when using ELCAC:

- Associate the SIP trunk between Unified CVP and Unified CM with a Phantom location.

Note Unified CM also has a *shadow location* for intercluster ELCAC. CVP does not support this.

- In multicluster Unified CM deployments, consider oversubscribing bandwidth on WAN links based on the anticipated peak call volume. You can also choose a centralized branch office deployment, as intercluster ELCAC is not supported on Unified CVP.
- In single-cluster Unified CM deployments, ELCAC is supported only for Hub and Spoke topology with Unified CVP.
- The ELCAC feature does not work with a trunk configured that requires MTP. When MTP is inserted, the media terminates between the device and MTP resource, not between the two devices.
- If the Unified CM media layer inserts a MTP/Transcoder/TRP media resource, the incoming location information is not used.
- If the intercluster call is not looped back to the same cluster, the former behavior of Location CAC logic applies.
- Each site has unique Location Routing Code. Align all gateways at the same site to the same Location Routing Code. If two clusters use the same location name, then two Location Routing Codes can map to the same physical branch.
- A second Unified CM cluster may have the same location as the first cluster, but be required to use a unique Location Routing Codeon Unified CVP. You can define a route in the proxy server to send all calls at the same location to a common Voice Browser that both clusters use.
- Each cluster would manage the bandwidth for devices in its cluster. If two clusters happen to use the same physical location, then they each separately manage the bandwidth for the phones that they manage.

Distributed Network Options

You can distribute the gateways in the following options:

- **Combined Branch Gateways**—Enables call treatment at the edge and integration of locally dialed numbers into the enterprise virtual contact center. You have both the Ingress Gateway and the Voice Browser at the branch. If you use a Cisco IOS Voice Gateway, you can combine the Ingress Gateway and Voice Browser functions on it.
- **Branch Ingress Voice Gateways with Centralized Voice Browsers**—Enables integration of locally dialed numbers and resource grouping of Voice Browser. This option supports organizations with many branches, with a few contact center calls in each branch. The VRU announcements in the centralized Voice Browsers traverse the WAN to the Ingress Gateway.
- **Branch Egress Gateways**—Enables agents to transfer calls across the WAN to remote TDM terminations.

You can also use a combination of these distributed options.

Home Agent with Cisco Virtual Office

Cisco Virtual Office solutions boost flexibility and productivity by delivering secure, comprehensive, and manageable network services to teleworkers. They supply full IP phone, wireless, data, and video services

over an encrypted VPN. Cisco Virtual Office delivers a transparent, office-caliber experience. Video playback is smooth, voice doesn't stutter, and wireless connectivity is effortless.

In a Cisco Virtual Office, the VPN router requires QoS capability for the desktop. Include in your calculations the bandwidth for Unified Intelligence Center, the agent desktop, and extra call flows such as recording.

Remember that broadband has no guarantee on bandwidth. Because of this, your broadband link needs greater capacity than the minimum requirement for the contact center traffic. The greater bandwidth enables the agent to stay active during peak times.

Unified Mobile Agent

Unified Mobile Agent supports call center agents using phones that Unified CCE does not directly control. A mobile agent can be physically located either outside or inside the contact center.

- **Outside the contact center**—The agent uses an analog phone in the home or a mobile phone.
- **Within the contact center**—The agent uses an IP phone connection that Unified CCE or Unified Communications Manager does not control.

In addition, a mobile agent can be available through different phone numbers at different times; the agent enters the phone number at sign-in time. The agent can access Unified Mobile Agent using any phone number, as long as the agent can dial the number using the Unified CM Dial Plan.

System administrators configure the Unified Mobile Agent to use a nailed (permanent) or call-by-call connection. Mobile agents can participate in outbound campaigns, but they can only use the nailed connection mode for all outbound dialing modes.

Unified Mobile Agent Components

The Unified Mobile Agent deployment scenario includes the following components:

- Cisco Virtual Office cable/DSL router for secure VPN data connectivity to the sites (no voice)
- Agent uses local phone with traditional local phone service to accept inbound calls
- Cisco Finesse desktops connect to Cisco Virtual Office cable/DSL router
- Administration tools: Unified configuration tools, Internet Script Editor, or the local Administrative Workstation

Unified Mobile Agent Benefits

The Unified Mobile Agent deployment scenario provides the following benefits:

- Unified Mobile Agent can send calls to any PSTN or mobile phone. This extends the reach of a centralized IP contact center.
- Contact centers can hire skilled employees where they live and integrate remote workers into geographically dispersed teams with access to equivalent corporate applications.
- Contact centers can reduce startup costs by bringing temporary agents online during seasonal high call volume. Agents can choose their destination phone number during sign-up time. They can change the number as often as they want, giving the contact center the flexibility to adapt to a fast-moving mobile workforce.

- The mobile agents have equal access to applications and services as agents at the central site. These geographically dispersed agents create a built-in backup plan to keep business processes functioning in unforeseen circumstances.

Unified Mobile Agent Design Requirements

The following table describes the design requirements for Unified Mobile Agent.

Table 28: Unified Mobile Agent Design Requirements

	Requirement	Notes
Configuration	Dial plan	For mobile agents on a dedicated gateway, all calls from the CTI ports go through a specific gateway at the site regardless of which phone number is called.
		Define the local CTI port directory number (DN), which is the routing label when the agent is selected.
		To keep the mobile agent signed in, set the values for both the Maximum Call Duration timer and Maximum Call Hold timer to 0. To configure these timers, use the Unified CM Administration web page for service parameters using Unified Communications Service.
		The Cisco Unified Mobile Agent connect tone provides an audible indication when a call is delivered to the nailed connection mobile agent. The connection tone is two beeps, which the nailed connection mobile agent hears when answering a call. This feature is turned off by default. Use the PG registry key PlayMAConnectTone to enable the Cisco Unified Mobile Agent connect tone.
	SIP trunk (CUBE)	CUBE dynamically changes the media port during the call. If you use the Mobile Agent feature, the SIP trunk that connects to the agent endpoint requires MTP resources.
Codec	G.711 or G.729	Ingress and egress voice gateways can be G.711 or G.729 but not a mix of both. All CTI ports for a PG must advertise the same codec type. All mobile agents should use the same codec, but local agents on the supervisor's team can use a mix of codecs. Configure the gateway MTPs to do a codec pass-through because the Mobile Agent uses G.729 and the rest of the components support all the codecs.

	Requirement	Notes
Infrastructure	DNS	You must have a DNS entry for the mobile agent desktop. If you do not have a DNS entry for the mobile agent desktop, the agent cannot connect to a CTI server.
	Firewall	If an agent with a nailed connection is idle longer than the firewall idle timeout value, the firewall can block the media stream. To prevent this, increase the firewall idle timeout value.
	Bandwidth	Minimum supported bandwidth speed: • 256-kbps upload • 1.0-Mbps download Use bandwidth calculators to ensure that you provide sufficient bandwidth. QoS is enabled only at the remote agent router edge. Currently, service providers do not provide QoS.
	Latency	The mobile agent round-trip delay to the Unified CCE site must not exceed 200 ms. The mobile agent jitter delay must not exceed 60 ms.
	Voice gateways	Use egress gateways for mobile agents.
	Call control	Use RONA when a mobile agent is signed in and ready, but is unavailable to answer a call.
		A mobile agent on one PG can only make blind transfers and conferences to a mobile agent on another PG in the same Unified CM cluster.
	Phones	Disable agent phone call features such as call waiting, call forwarding, and voicemail.
	Agent workstation	Set up the mobile agent workstation to use DHCP.
	Security	Enable security features on the remote agent router.
Desktop	Cisco Finesse	Cisco Finesse does not support Switched Port Analyzer (SPAN) port silent monitoring.
Recording	SPAN port Network-based Recording	Recording server in the site.
Silent Monitoring	Not available	

The following table describes Unified Mobile Agent media resources.

Table 29: Unified Mobile Agent Media Resources

Resource	Method	Notes
Music on Hold	Unified CM unicast	If the MoH server does not stream using a G.729 codec, then set up a transcoder to enable outside callers to receive Music on Hold.
Conference bridges	Voice gateways in the site	Agent greeting requires a conference bridge.
Media Termination Points	Voice gateways in the site	Assign two MTPs for each Unified Mobile Agent: • MTP for remote CTI port • MTP for local CTI port CTI ports do not support in-band Dual-Tone Multifrequency (DTMF) RFC 2833. The MTPs perform the conversion. Do not place MTPs at the egress gateway. If you use SIP trunks, configure Media Termination Points (MTPs). Enabling the use of an MTP on a trunk affects all calls that traverse that trunk, even noncontact center calls. Ensure that the number of available MTPs can support the number of calls traversing the trunk.
Transcoders	Voice gateways in the site	All mobile agents must have the same codec: G.711 or G.729.

Solution Administration

The contact center enterprise solutions offers several sets of primary administration tools.

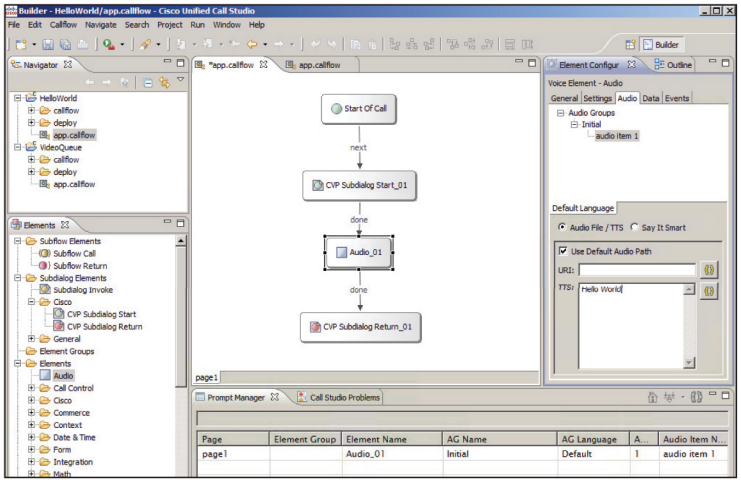
Service Creation Environments

The contact center enterprise solutions include two service creation environments:

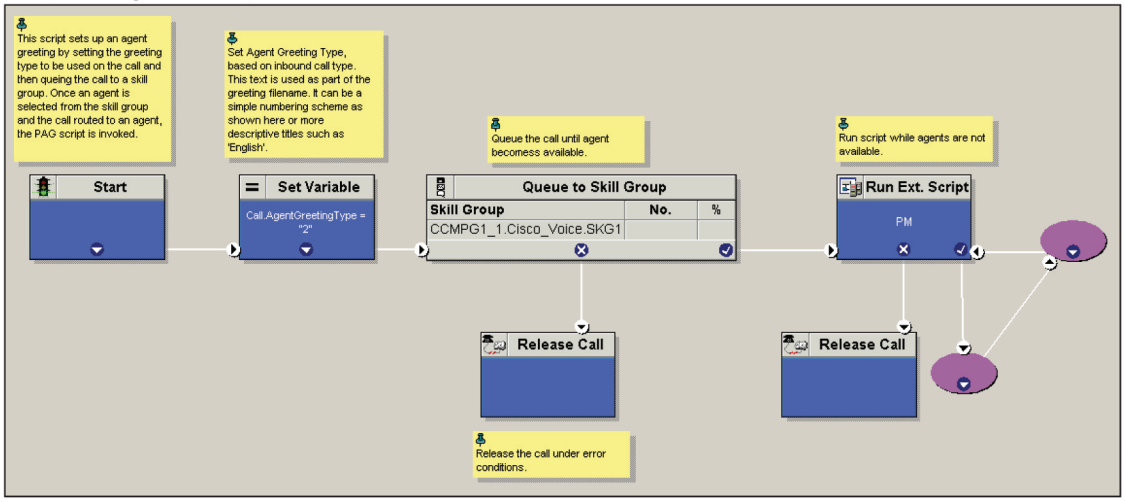
- **Unified CVP Call Studio**—The Call Studio is a platform for rapidly creating, managing, and deploying sophisticated dynamic VXML self-service applications. The Call Studio application runs in the Eclipse framework. You do not need knowledge of Eclipse to work with the Call Studio. The Call Studio includes plug-ins for voice application development, Java programming, and many other features provided by Eclipse.
- **Unified CCE Script Editor**—The Script Editor is a tool for creating, updating, scheduling, and monitoring your routing scripts and administrative scripts.

Figure 53: Service Creation Environment

Script using Unified Call Studio



Script using ICM Script Editor



Solution Serviceability and Monitoring

The contact center enterprise solutions support several solution serviceability tools. These tools leverage similar interfaces (SNMP, Syslog, Diagnostic REST/SOAP API, telnet/SSH CLI interface) from each component of the solution but provide unique features and functionality.

- Analysis Manager
- Prime Collaboration Assurance
- Unified System CLI

Also, you could use third-party SNMP and network management tools as well to monitor and perform solution serviceability.

Prime Collaboration Manager

For managing a Unified Communications deployment, customers are encouraged to use the Cisco Prime Collaboration Assurance product. Cisco Prime Collaboration Assurance is a member of the Cisco Unified Communications family of products and provides a comprehensive and efficient solution for network management, provisioning, and monitoring of Cisco Unified Communications deployments.

Cisco Prime Collaboration Assurance monitors and evaluates the current status of both the IP communications infrastructure and the underlying transport infrastructure in your network. Cisco Prime Collaboration Assurance uses open interfaces such as SNMP and HTTP to remotely poll data from different devices in the IP communications deployment.

Cisco Prime Collaboration Assurance is a comprehensive video and voice assurance and management system with a set of monitoring, troubleshooting, and reporting capabilities that help ensure end users receive a consistent, high-quality video and voice collaboration experience. You deploy Prime Collaboration in Managed Service Provider (MSP) mode. The following are the key features of Cisco Prime Collaboration.

- Voice and Video Unified Dashboard
- Device Inventory Management
- Voice and Video Endpoint Monitoring
- Diagnostics
- Fault Management
- Reports
- Live Contact Center topology with link status, device status, device performance, device 360
- Contact Center device discovery
- Contact Center devices real time performance monitoring
- Events and Alarms along with root cause analysis
- Contact Center device Dashboards - Prebuilt and custom
- Threshold, Syslog, Correlation and System Rules - Prebuilt and custom
- Multi-tenancy and logged-in agent licensing information

Analysis Manager

The Analysis Manager functionality integrated with the Unified Communications Manager Real-Time Monitoring Tool (RTMT) is provided as the client-side tool to collect diagnostic information from this diagnostic framework.

Using the Analysis Manager, the administrator connects to one or more Unified Communications devices to set trace levels, collect trace and log files, and gather platform and application configuration data as well as version and license information. The Analysis Manager is the one tool that allows administrators to collect diagnostic information from all Cisco Unified Communications applications and devices.

The Analysis Manager offers local user and domain security for authentication and secure HTTP to protect data exchanged by it and the diagnostic framework.

The Web Service Manager supports all diagnostic (health and status) requests from the Analysis Manager. The Analysis Manager is part of UCM RTMT tool. It provides users an interface for collecting health and status information for all devices in its network topology. If Unified CVP is configured as a part of the solution, you can leverage the WSM through the Analysis Manager to collect diagnostic details, such as server map, version information, licenses, configuration, components, logs, traces, performance factors, platform information for each CVP Device on a component and subcomponent level. You can set or reset debug levels using the Analysis Manager on a component and subcomponent level.

A new user with the wsmadmin username is created during installation with the same password as the Operations Console Server administrator user. Use wsmadmin to control access to the diagnostic portal services.

Unified System CLI

In addition to the Analysis Manager, a command line interface-Unified System CLI tool-is available that allows a client to access the diagnostic framework on any Unified Communications server. The Unified System CLI can be accessed without a remote desktop.

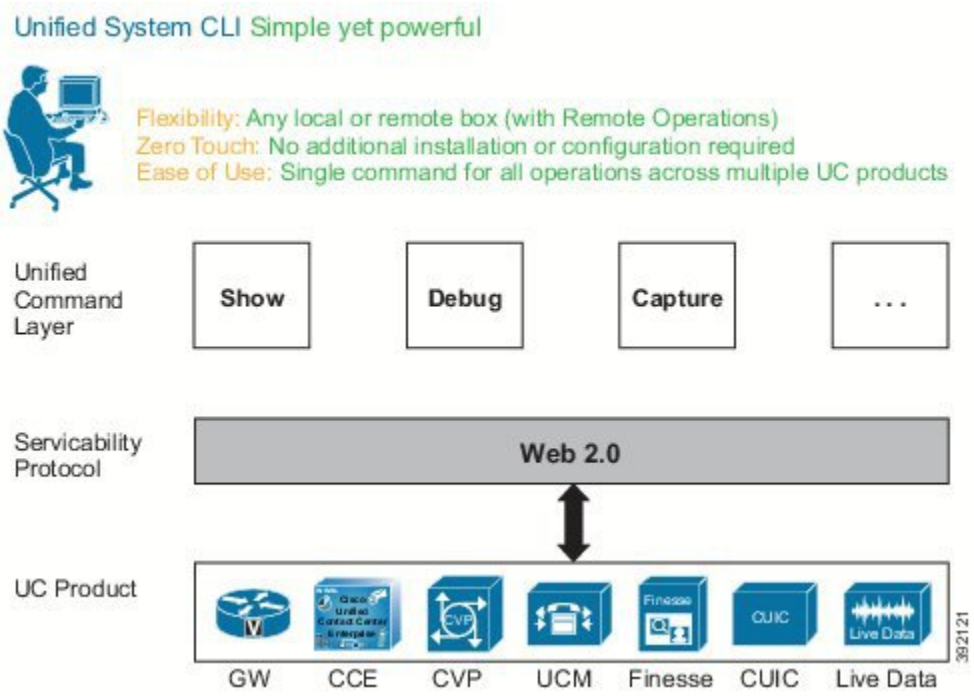
When an issue arises in your solution, use the System CLI tool to collect data for review by Cisco engineers. For example, you can use the System CLI if you suspect a call is handled incorrectly. In this case, you can use the **show tech-support** command to collect data and send the data to Cisco support.

Unified System CLI has the following features:

- Is automatically installed on all Unified CVP Servers as part of the infrastructure. No additional installation is required on any Unified CVP server.
- Uses a consistent command across the solution.
- Can be run as a Windows scheduled job.

The following figure shows the high-level commands for the Unified System CLI and shows the interaction of devices and Unified Cisco products.

Figure 54: High-Level Commands for Unified System CLI



Unified System CLI runs at a low priority; it uses idle CPU time on the system. It should not affect call processing even if run on a system running under load.

The response time from the given CLI command varies depending on the load of the system and the server response time. The response time when there is no running load should be below 5 seconds for each server for operations, such as show version, show license, show debug, and show perf. The response time when there is no running load for show platform operation should be below 10 seconds for each server.

However, the response time cannot be determined for commands, such as show trace, show log, show sessions, show all, and show tech-support. The response for these commands can vary depending on the data being transferred by the server.

Unified System CLI Modes of Operation

The Unified System CLI operates as an interactive user interface and can also be used as a batch command. This feature allows the Unified System CLI to be used in scheduled jobs.

The Unified System CLI can operate interactively as follows:

- **Local mode**—In this mode, the Unified System CLI only interacts with a single device. For example, the **show version** command shows only the version for a single device.

Analysis Manager vs Unified System CLI

Analysis Manager and Unified System CLI access the Diagnostic Portal API. Both the Analysis Manager and the Unified System CLI have similar features, except for the differences shown in the table.

Table 30: Differences Between Analysis Manager and Unified System CLI

Analysis Manager	Unified System CLI
Is a GUI-based client that is part of the Unified CM Real-Time Monitoring Tool (RTMT). The Analysis Manager has a user-friendly interface due to its GUI-based design.	Is a command line based tool. The Unified System CLI is more flexible because it can be used in a batch file to perform more complex tasks.
Is neither bundled with CVP nor installed by Unified CVP installer.	Is bundled with Unified CVP installer, and is also bundled with the Unified CCE installer.

Third-Party Network Management Tools

Unified CCE is managed using the Simple Network Management Protocol (SNMP). Unified CCE devices have a built-in SNMP agent infrastructure that supports SNMP v1, v2c, and v3 and it exposes instrumentation defined by the CISCO-CONTACT-CENTER-APPS-MIB. This MIB provides configuration, discovery, and health instrumentation that you can monitor with standard SNMP management stations. Unified CCE provides a rich set of SNMP notifications that alerts administrators of any faults in the system. Unified CCE also provides a standard syslog event feed (conforming to RFC 3164) if you need a more verbose set of events.

Unified CVP and Unified Intelligence Center support SNMP v2 and v3.

Cisco Finesse and Customer Collaboration Platform only support SNMP from the VOS platform. You cannot use SNMP directly from the Cisco Finesse and Customer Collaboration Platform applications.

You can use Simple Network Management Protocol (SNMP) station to monitor the solution deployment status.

Unified CCE has a built-in web-based (REST-like) interface for diagnostics called the Diagnostic Framework, which is resident on every Unified CCE server.

System Performance Monitoring Guidelines

Supporting and maintaining an enterprise solution requires many steps and procedures. Depending on the customer environment, the support procedures vary. System performance monitoring is one procedure that helps maintain the system. This section provides a guide for monitoring Unified CCE to ensure that the system is performing within system tolerances. System monitoring is especially critical for customers as they expand or upgrade their system. Monitor the system during times of heavy activity.

The following system components are critical to monitor:

- CPU
- Memory
- Disk
- Network

The following table highlights some of the important counters for the critical system components, along with their threshold values:

Table 31: Monitoring Threshold Values

Monitored Resource	Thresholds
CPU	%Processor Time; the threshold of this counter is 60%. ProcessorQueueLength; this value must not exceed (2 * [the total number of CPUs on the system]).
Memory	% Committed Bytes; this value must remain less than (0.8 * [the total amount of physical memory]). Memory\Available MByte; this value must not be less than 16 MB. Page File %usage; the threshold for this counter is 80%.
Disk	AverageDiskQueueLength; this value must remain less than (1.5 * [the total number of disks in the array]). %Disktime; this value must remain less than 60%.
Network	NIC\bytes total/sec; this value must remain less than (0.3 * [the bandwidth of the NIC]). NIC\Output Queue Length; the threshold for this counter is 1.
Unified CCE	Cisco Call Router(_Total)\Agents Logged On Cisco Call Router(_Total)\Calls in Progress Cisco Call Router(_Total)\calls /sec

In general, the 95th percentile for your busy hour traffic should not exceed these thresholds.

Note These performance counters for CPU, memory, disk, and network are applicable to all Windows-based applications within the deployment. The sample rate is 15 seconds.

For more information on monitoring your VMs, see *Cisco Collaboration Virtualization* at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

End-to-End Individual Call Tracking

When a call arrives at the Ingress Gateway, Cisco IOS assigns that call a 36-digit hexadecimal Global Unique Identifier (GUID), which identifies the call. The contact center carries that GUID through all of the components that the call encounters, as follows:

- Ingress gateway—Shown in Cisco IOS log files.
- Voice Browser—Shown in Cisco IOS and Cisco VVB log files.
- Unified CVP components—Shown in Unified CVP log files.
- Unified CCE—Shown in the Extended Call Context (ECC) variable `user.media.id` and stored with all Termination Call Detail (TCD) and Route Call Detail (RCD) records.

- Automatic speech recognition (ASR) and text-to-speech (TTS) servers—Shown in logs as the logging tag.
- Cisco Unified Communications Manager (Unified CM)—Appears in the detailed logs.

With proper levels of logging enabled, you can trace a call through all these components.

Localization

The contact center enterprise solutions concentrate on providing localization support for the agent and supervisor desktops. Most of the administration tools use exclusively English. The tools accept characters from the appropriate Windows code page for your SQL collation in these values:

- Agent names
- Peripheral variables
- ECC variables
- Description fields of ICM tables
- Wrap-up data
- Reason codes

However, you always enter characters from left to right in the tools. Each Unified CCE instance can only support one Windows code page in the database. The *Compatibility Matrix* for your contact center enterprise solution lists the supported localized versions of Microsoft Windows Server and SQL Server that you can use with your solution.

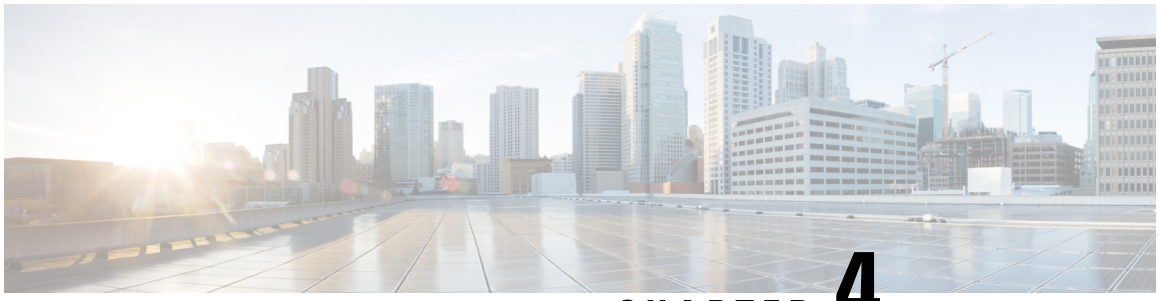
For example, with Latin1_General for the SQL collation, agent names can contain any language written in the Western European character set (Windows code page 1252). These include Afrikaans, Basque Catalan, Georgian, Indonesian, Irish, and Malay. With Cyrillic_general for the SQL collation, agent names can contain any languages written in Cyrillic (Windows code page 1251). These include Bulgarian, Kyrgyz, Mongolian, Uzbek, Serbian, and Ukrainian.

For more information on localization of the Finesse desktop, see the *Cisco Finesse Administration Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-user-guide-list.html.

Multi-Language Support

The Voice Browser and the Media Resource Control Protocol (MRCP) specification do not restrict support for multiple languages. However, your Automatic Speech Recognition (ASR) or TTS Server might have restrictions for this. Check with your preferred ASR or TTS vendor about their support for your languages before preparing a multilingual application.

You can dynamically change the ASR server value with the **cisco property com.cisco.asr-server** command in the VXML script. This property overrides any previous value set by the VXML script. Similarly, you can change the TTS server with **cisco property com.cisco.tts-server** command in the VXML script.



CHAPTER 4

Configuration Limits and Feature Availability for Reference Designs

Reference Design Configuration Limits

Note The first four chapters of this book are for anyone who wants to get familiar with the contact center enterprise solutions:

- Packaged Contact Center Enterprise
- Cisco Hosted Collaboration Solution for Contact Center
- Unified Contact Center Enterprise

For information about design considerations and guidelines specific to Packaged CCE, see the remaining chapters.

The following tables list key configuration limits for Contact Center Enterprise Reference Designs solutions.

Some of these limits are interdependent and dynamically change depending on the elements in your solution. For example, the number of skills per agent affects the maximum number of agents.

Limits that are listed as "per PG" always refer to a redundant pair of PGs.

Important Your contact center enterprise solution can only use the new higher configuration limits with the standard three coresident PG layout.

Agent Limits

Note The figures in the Contact Director column refer to what are configured on the Contact Director. The figures do not include what is configured on the target systems to which the Contact Director connects.

Table 32: Agent Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Active Agents[10]	2000	4000	12,000	24,000	24,000 (cumulative on 3 target systems)
Active Agents on each Unified CM cluster	2000	4000	8000	8000	NA
Configured Agents	12,000	24,000	72,000	72,000	NA
Configured Agents per PG	12000	12000	12000	12000	NA
Agents with TraceON enabled	100	100	400	400	NA
Agent Desk Settings	2000	4000	12,000	12,000	NA
Active Mobile Agents per Agent PG[11] [12]	2000 with nailed-up connections Or 1500 with call-by-call connections	2000 with nailed-up connections Or 1500 with call-by-call connections	2000 with nailed-up connections Or 1500 with call-by-call connections	2000 with nailed-up connections Or 1500 with call-by-call connections	NA
Active ECE Multimedia Agents	1500[13]	4000[14]	12,000[15]	24,000[16]	NA
Agents per team	50	50	50	50	NA
Teams to which an agent can belong	1	1	1	1	NA
Skills per agent	15 Refer to the section on dynamic sizing for details.	15 Refer to the section on dynamic sizing for details.	15 Refer to the section on dynamic sizing for details.	10 Refer to the section on dynamic sizing for details.	NA
Number of agents in a skill group	12,000	24,000	72,000	72,000	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Attributes per agent	50	50	50	50	NA

[10] This includes Outbound and Multichannel agents. However, the number of agents that you can keep occupied is based on the Outbound Option dialer and Customer Collaboration Platformlimits.
[11] 1500 with nailed-up connections if average handle time is less than 3 minutes, or if Agent greeting or Whisper Announcement features are used in conjunction with Mobile Agent.
[12] The Large PG OVA supports 2000 agents with call-by-call connections. The Packaged CCE 2000 agent deployment does not support large PGs.
[13] When ECE is colocated, the limit is 400 agents. The limit of 1500 applies when ECE is on a separate server.
[14] This limit requires multiple ECE clusters. Each Agent PG can support either a 400 agent colocated cluster or a 1500 agent cluster on a separate server.
[15] This limit requires multiple ECE clusters. Each Agent PG can support either a 400 agent colocated cluster or a 1500 agent cluster on a separate server.
[16] This limit requires multiple ECE clusters. Each Agent PG can support either a 400 agent colocated cluster or a 1500 agent cluster on a separate server.

Supervisor and Reporting User Limits

Table 33: Supervisor and Reporting User Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	Unified CCE 24000 Agent Reference Design Model	Contact Director Reference Design Model
Active Supervisors[17]	200	400	1200	2400[18]	NA
Configured Supervisors	1200	2400	7200	7200	NA
Active teams	200	400	1200	2400	NA
Configured teams	1200	2400	7200	7200	NA
Supervisors per Team	20	20	20	20	NA
Teams per supervisor	20	20	20	20	NA
Agents per supervisor	1000	1000	1000	1000	NA
Active Cisco Unified Intelligence Center Reporting users	200	400	1200[19]	1200[20]	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	Unified CCE 24000 Agent Reference Design Model	Contact Director Reference Design Model
Configured Cisco Unified Intelligence Center Reporting users	1200	2400	7200	7200	NA
Reporting users per CUIC node	100	200	200	200	NA

[17] Supervisors count against the agent limits. Ten percent of your active agents can be supervisors.
[18] Because there can only be 1200 Active Reporting users, all Active Supervisors cannot concurrently use Cisco Unified Intelligence Center reports.
[19] During a Central Controller failover, this limit drops to 600 until both sides are active again.
[20] During a Central Controller failover, this limit drops to 600 until both sides are active again.

Access Control Limits

Table 34: Access Control Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Active Administrators per distributor[21]	50	50	50	50	50
Configured Web Administrators	100	100	100	100	NA
Roles—Packaged CCE only	30	30	30	NA	NA
Departments—Packaged CCE only	200	200[22]	200	NA	NA
Department per Administrator—Packaged CCE only	10	10	10	NA	NA
Machines in inventory	1000	1000	1000	1000	NA

[21] Because Packaged CCE, CCMP, and CCDM use web administration, this limit does not apply with them.
[22] Departments limit also applies to Cisco HCS for Contact Center Small Contact Center solutions.

Outbound Campaign Limits

Table 35: Outbound Campaign Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Outbound dialer per system	1 per Agent PG	1 per Agent PG	1 per Agent PG	1 per Agent PG	NA
Outbound dialer maximum calls per second	60	120	240	240	NA
Outbound dialer maximum calls per second per dialer[23]	60	60	60	60	NA
Outbound dialer maximum ports on each SIP dialer	3000	3000	3000	3000	NA
Outbound dialer maximum ports on each system (total)	3000	6000	12000	12000	NA
Number of Preview Campaigns per System	1500 campaigns Preview and Direct Preview modes support up to 750 campaign skill groups on a Medium PG VM and 1500 campaign skill groups on a Large PG VM.	1500 campaigns Preview and Direct Preview modes support up to 750 campaign skill groups on a Medium PG VM and 1500 campaign skill groups on a Large PG VM.	1500 campaigns Preview and Direct Preview modes support up to 750 campaign skill groups on a Medium PG VM and 1500 campaign skill groups on a Large PG VM.	1500 campaigns Preview and Direct Preview modes support up to 750 campaign skill groups on a Medium PG VM and 1500 campaign skill groups on a Large PG VM.	NA
Number of Predictive Campaigns per system (Agent or VRU based)	375	750	1500	1500	
Campaign skill groups per Campaign	20	20	20	20	NA
Predictive Campaign Skill Groups per Peripheral	375	375	375	375	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Maximum Outbound Skills per Agent	5	5	5	5	NA
Do Not Call Records per Import	1,000,000	20,000,000	60,000,000	60,000,000	NA

[23] This figure assumes a 30% transfer rate to a VRU or an agent.

Precision Queue and Skill Groups Limits

Note Each Precision Queue has an associated Skill Group. Each Precision Queue effectively has a weight of two Skill Groups.

Table 36: Precision Queue and Skill Group Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Skill Groups per System	16,000 [24]	16,000	27,000	48,000	54,000
Enterprise Skill Groups	4000	4000	4000	4000	4000
Maximum combined configured Skill Groups and Precision Queues per peripheral	4000	4000	4000	4000	NA
Configured Precision Queues per system	4000 [25]	4000 [26]	The smaller of: 4000 Or 27,000 divided by the number of agent peripherals	The smaller of: 4000 Or 48,000 divided by the number of agent peripherals	8000 of the maximum 54,000 queues
Precision Queue steps	10,000	10,000	10,000	10,000	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Precision Queue term per Precision Queue	10	10	10	10	NA
Precision steps per Precision Queue	10	10	10	10	NA
Unique attributes per Precision Queue	10	10	10	10	NA
Max Unique Skills per Team	50	50	50	50	NA
Configured labels	100,000	100,000	160,000	160,000	160,000
Precision Routing Attributes on each system	10,000	10,000	10,000	10,000	NA
Precision Routing Attributes for each Agent	50	50	50	50	NA
Skill Group statistics refresh rate	10 seconds (default)	10 seconds (default)	10 seconds (default)	10 seconds (default)	NA
Skill Groups per PG	4000	4000	4000	4000	NA
Queues[27] per Contact Sharing Group	NA	NA	NA	NA	100
Contact Sharing Rules	NA	NA	NA	NA	100
Contact Sharing Groups	NA	NA	NA	NA	1000

[24] In most Packaged CCE 2000 Agent topologies, you can only have 4000 Skill Groups because there is only 1 Agent PG. In the Global topology, using remote sites, Packaged CCE supports 16,000 skill groups, system wide. Each remote site with an Agent PG adds 4000 skill groups. The 16,000 maximum requires 3 remote sites.

[25] In a Non-Reference Design deployment (when you use more agent PGs than what is supported by your CCE reference design), use this formula to calculate the maximum number of Precision Queues per system: lesser of 4000 or 27000 / total number of Agent PGs.

[26] In a Non-Reference Design deployment (when you use more agent PGs than what is supported by your CCE reference design), use this formula to calculate the maximum number of Precision Queues per system: lesser of 4000 or 27000 / total number of Agent PGs.

[27] This term includes both Skill Groups and Precision Queues.

Task Routing Limits

Table 37: Task Routing Limits

Resource	2000 Agent Reference Design	4000 Agent Reference Design	12000 Agent Reference Design	24000 Agent Reference Design	Contact Director Reference Design
Maximum active agents assigned to tasks per system	2000	2000	2000	2000	NA
Maximum reserved and active tasks per agent[28]	15	15	15	15	NA
Maximum incoming tasks/sec across all MRDs[29]	5	5	5	5	NA
Task Routing API request/hr through Customer Collaboration Platform	15,000	15,000	15,000	15,000	NA

[28] This figure includes paused and interrupted tasks. Tasks that are still in queue or are transferred out by an agent do not count towards this limit.

[29] Customer Collaboration Platform throttles the task submission rate to Unified CCE to 5 tasks per second. Customer Collaboration Platform holds a maximum of 10,000 tasks in the queue for submission. If the queue exceeds 10,000 tasks, then Customer Collaboration Platform discards the additional tasks with the disposition code NOTIFICATION_RATE_LIMITED. Once the queue is ready again, additional tasks are added to the queue.

Dialed Number Limits

Note In the Global topology, each remote site can support the full limit of Dialed Numbers as mentioned in the table.

Table 38: Dialed Number Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Dialed Numbers on each CVP peripheral (External Voice and Post Call Survey) [30]	2000[31]	4000	12,000	12,000	12,000

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Dialed Number on each Unified CM peripheral (Internal Voice)	2000	2000	2000	2000	NA
Dialed Number on each MR peripheral (Multichannel)	1000	1000	1000	1000	NA
Dialed Number on each Unified CM peripheral (Outbound Voice)	1000	1000	1000	1000	NA

[30] You cannot exceed the system maximum total of 240,000 DN records across all routing client types.
[31] 2000 agent reference model supports a total of 2000 external dialed numbers, of which 500 is for Post Call Survery and the remainder is for other CVP dialed numbers.

System Load Limits

Table 39: System Load Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
VRU Ports in a Reference Layout[3233]	3000	6000	18,000	36,000	36,000
Maximum VRU Ports with Added PGs[34]	6000	12,000	36,000	48,000	72,000
Maximum Inbound Calls per Second (CPS)	15	30	90	90	300, of which Contact Sharing can handle 120 and the remainder is for self-service and line-of-business direct routing.
Congestion Control CPS[35]	18	35	105	105	300
Maximum Inbound CPS per VRU PG[36]	15	15	15	15	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Maximum VRU PIM per VRU PG	2	2	2	2	NA
Dynamic Reskilling (operations/hr.)	7200	7200	7200	7200	NA
Maximum Queued Calls and Tasks	15,000	15,000	15,000	15,000 [37]	15,000
Media Routing Domains per system	20	20	20	20	NA
Agent Callback requests through Customer Collaboration Platform(requests/hr.)	1000	1000	1000	1000	NA
ECE Email or Chat requests per hour for 400 agent deployment	6 per agent	6 per agent	6 per agent	6 per agent	NA
ECE Email or Chat requests per hour for 1500 agent deployment[38]	6 per agent	6 per agent	6 per agent	6 per agent	NA
Incoming Messages per Second for CVP Reporting Server	420	420	420	420	420
Reports per user For more details, see Resource Requirements for Reporting, on page 395	2 Live Data reports 2 AW-RealTime reports 2 historical reports	2 Live Data reports 2 AW-RealTime reports 2 historical reports	2 Live Data reports 2 AW-RealTime reports 2 historical reports	2 Live Data reports 2 AW-RealTime reports 2 historical reports	NA
Maximum rows per report[39]	3000 for real-time 8000 for historical	3000 for real-time 8000 for historical	3000 for real-time 8000 for historical	3000 for real-time 8000 for historical	NA
Configured Business Hour Objects[40]	1000	1000	1000	1000	1000
Configured Schedule Objects per Business Hours Object[41]	50	50	50	50	50

[32] These figures assume that your solution has an equal number of redundant ports. The actual number of ports is twice these figures.
[33] The total calls at agents or the VXML server in the basic layout for each Reference Design model. The added components in a global deployment increase these numbers.
[34] These figures assume that your solution has an equal number of redundant ports. The actual number of ports is twice these figures.
[35] Inbound calls per second figures assume 10% of agents are supervisors who are not directly answering calls. The figures also assume a distribution of calls with 10% transfers and 5% conferences.
[36] If one of the CVP Call Servers is down, the maximum inbound CPS per VRU PIM is also 15.
[37] You can increase this to 27,000 by changing the `ICM\<inst>\Router[A/B]\Router\CurrentVersion\Configuration\Queuing\MaxCalls` registry setting.
[38] For more details on email or chat sizing considerations, see the Enterprise Chat and Email Design Guide at https://www.cisco.com/c/en/us/support/customer-collaboration/cisco-enterprise-chat-email/products-implementation-design-guides-list.html.
[39] **Large Schedules** that are configured in Cisco Unified Intelligence Center have an upper limit of 25000 rows. For more information, see Cisco Unified Intelligence Center User Guide.
[40] Cisco HCS for Contact Center does not support the Business Hours feature.
[41] Daily schedules account for 7 of these schedule objects. You can use the remainder for holidays and exceptions.

Call Variable Limits

Table 40: Call Variable Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Persistent Enabled Expanded Call Variables (Default)[42]	5	5	5	5	5
Persistent Enabled Expanded Call Variable Arrays	0	0	0	0	0
Maximum Contents per ECC (Expanded Call Context) Variable (bytes)	210	210	210	210	210
Maximum Total ECC Contents Size per ECC Payload (bytes)	2000	2000	2000	2000	2000
Maximum ECC Variable Name (bytes without null character)	32	32	32	32	32

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Maximum Total Contents and Name Size for ECC Variables per ECC Payload (bytes)	2500	2500	2500	2500	2500
Maximum ECC Variables Contents per Call (bytes)	6000	6000	6000	6000	6000
Maximum System-wide ECC Variable Contents (bytes)[43]	90,000,000	90,000,000	90,000,000	90,000,000	NA
Number of Peripheral Variables	10	10	10	10	10
Call Context for Peripheral Variables 1-10 (bytes)	40	40	40	40	40

[42] See the "Call Context" section for details.
[43] This limit is the maximum per call limit multiplied by the maximum queued calls and tasks for the system.

Other Limits

Table 41: Other Limits

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
Maximum Agent PGs with Live Data, Precision Queueing, or Single Sign-On enabled[44]	4[45]	4 [46] 12 (when using the 12000 agents Live Data OVA)	12 24 (when using the 24000 agents Live Data OVA)	24	50
Maximum PGs [47]	30	100	150	150	NA
Maximum Agent PGs on each VM	1	1	1	1	NA
Maximum Cisco Finesse server pairs per PG pair	1	1	1	1	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
MR PIMs on each MR PG	4	4	4	4	NA
Custom Application Gateway	20	20	20	20	20 per enterprise system
Bucket Intervals	2000	4000	12,000	12,000	NA
Configured Call Types	4000	8000	15,000	15,000	15,000
Call Type Skill Group per Interval[48]	70,000	70,000	70,000	70,000	NA
Active Routing Scripts	1000	2000	6000	6000	6000
Configured Routing Scripts	2000	4000	12,000	12,000	12,000
Network VRU Scripts	2000	4000	12,000	12,000	12,000
System-wide Maximum Configured Reason Codes and Labels	2800, plus 21 system-defined	3800, plus 21 system-defined	7800, plus 21 system-defined	7800, plus 21 system-defined	NA
Not-ready Reason Codes	100 global codes 100 associated reason codes for each team 500 configured team reason codes	100 global codes 100 associated reason codes for each team 1000 configured team reason codes	100 global codes 100 associated reason codes for each team 3000 configured team reason codes	100 global codes 100 associated reason codes for each team	NA
Sign-out Reason Codes	100 global codes 100 associated reason codes for each team 500 configured team reason codes	100 global codes 100 associated reason codes for each team 1000 configured team reason codes	100 global codes 100 associated reason codes for each team 3000 configured team reason codes	100 global codes 100 associated reason codes for each team	NA
Wrap-up Reason labels[49]	100 global labels 1500 team labels	100 global labels 1500 team labels	100 global labels 1500 team labels	100 global labels 1500 team labels	NA
Administration Bulk Jobs[50]	200	200	200	200	NA

Resource	2000 Agent Reference Design Model	4000 Agent Reference Design Model	12000 Agent Reference Design Model	24000 Agent Reference Design Model	Contact Director Reference Design Model
CTI AllEventClients[51]	7/Medium PG 20/Large PG[52]	7/Medium PG 20/Large PG [53]	7/Medium PG 20/Large PG [54]	7/Medium PG 20/Large PG	NA
Real-Time Only Distributors (for configuration only)	4 (2 on each side)	4 (2 on each side)	10 (5 on each side)	10 (5 on each side)	10 (5 on each side)
Agent Targeting Rule (ATR)	1000	1000	1000	1000	NA

44 Deploy only one Agent PG, one VRU PG, and one MR PG on each VM. Use the Medium PG OVA or Large PG OVA, depending on your need for CTI All-Event Clients.
45 For Packaged CCE 2000 Agent, you have only 1 Agent PG, 1 VRU PG, and 1 MR PG. You can extend to the 4 maximum, if you use the Global topology with 3 remote sites.
46 A Cisco HCS for Contact Center Small Contact Center solution can support 50 PGs with SSO enabled. But, it can only support 12 PGs with Precision Queueing and Live Data enabling.
47 The maximum PG count includes the maximum Agent PG count (specified in the previous row).
48 Exceeding this limit causes gaps in your reporting.
49 A team cannot use more than 100 of the total team wrap-up reason labels.
50 This covers the SSO Migration Tool and the Packaged CCE Bulk Tool. It does apply to legacy bulk configuration tools.
51 The CTI AllEventClients limit includes Cisco Finesse, Enterprise Chat and Email, and Outbound Dialer connections. These limits do not apply for CTI OS desktops.
52 Does not apply for Packaged CCE, which does not use the Large PG OVA.
53 Does not apply for Packaged CCE, which does not use the Large PG OVA.
54 Does not apply for Packaged CCE, which does not use the Large PG OVA.

The following table lists the configuration limits for adding external machines in the Packaged CCE deployments.

External Machines	2000 Agent Reference Design Model		4000 Agent Reference Design Model		12000 Agent Reference Design Model	
	Main Site	Remote Site	Main Site	Remote Site	Main Site	Remote Site
Gateways	0 or more	0 or more	0 or more	0 or more	0 or more	0 or more
Customer Collaboration Platform	0 or 1	None	0 or 1	None	0 or 1	None
Cisco Unified CVP Reporting	0 or 1	0 or 1	0 or 1	0 or more	0 or 1	0 or more
Cisco Enterprise Chat and Email (ECE)	0 or 1	4[55]	0 or 1	0 or more	0 or 1	0 or more

External Machines	2000 Agent Reference Design Model		4000 Agent Reference Design Model		12000 Agent Reference Design Model	
	Main Site	Remote Site	Main Site	Remote Site	Main Site	Remote Site
Third-party Multichannel	0 or 1	4[56]	0 or 1	0 or more	0 or 1	0 or more
Media Server	0 or more	0 or more	0 or more	0 or more	0 or more	0 or more

[55] Subject to the availability of MR PIMS (for which maximum limit is four). For example, if you have configured two ECEs, you will be able to configure only two more ECEs or Third-party Multichannels.

[56] Subject to the availability of MR PIMS (for which maximum limit is four). For example, if you have configured two Third-party multichannels, you will be able to configure only two more Third-party multichannels or ECEs.

Feature Availability for Reference Designs

These sections summarize the features available in contact center solutions that follow the Contact Center Enterprise Reference Designs.

Agent and Supervisor

Capability	Supported	Notes

Call Flows	Post-route by CVP
	Comprehensive call flow:
	• Inbound and outbound calls
	• Supplementary services
	  • Hold and resume
	  • Blind, consult, and refer transfers and conferences
	  • Router requery	

Outbound campaigns	Cisco Outbound Option supports these dialing modes: • Predictive • Preview • Direct Preview • Progressive	The SIP Dialer uses the UDP transfer protocol for SIP.
Mobile Agent	Nailed-up and Call-by-call modes	

Capability	Supported	Notes
Silent Monitoring	Unified CM-based (BiB)	You cannot monitor mobile agents with Unified CM-based silent monitoring.
Recording	Unified CM-based Network-based Recording CUBE(E)-based TDM gateway-based	
CRM Integration	CRM integration is available through the Cisco Finesse Web API, Finesse gadgets, and existing CRM connectors.	You can integrate with a CRM using the following methods: • CRM iFrame in the Finesse container. This method is simple and easy but does not provide deep CRM integration. • Third-party gadget in the Finesse container. This method achieves full CRM integration but requires custom development using third-party and Finesse APIs. • Finesse gadgets in a CRM browser-based desktop. This method provides lightweight integration into the CRM application. • Finesse Web API s or the CTI Server protocol to integrate into a CRM application. This method provides deep CRM integration but requires custom development.
Desktop	Cisco Finesse Finesse IP Phone Agent	FIPPA only supports a subset of Finesse's features.
Desktop Customization	Cisco Finesse API	

Voice and Infrastructure

Capability	Supported	Notes
Music on Hold	Unicast with Unified CM subscriber or voice gateway Multicast using voice gateway	

Capability	Supported	Notes
Proxy / Cisco Unified SIP Proxy (CUSP)	SIP Proxy is an optional component.	Instead of using CUSP, some deployments can achieve High Availability (HA) and load balancing using these solution components: • Time Division Multiplexing (TDM) gateway and Unified CM, which use the SIP Options heartbeat mechanism to perform HA. • Unified CVP servers, which use the SIP server group and SIP Options heartbeat mechanism to perform HA and load balancing. • Outbound Option. The Outbound dialer can connect to only one physical gateway, if SIP proxy is not used.
Ingress Gateways	See the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html for the supported hardware.	
Protocol	Session Initiation Protocol (SIP) over Transmission Control Protocol (TCP) Session Initiation Protocol (SIP) over User Datagram Protocol (UDP) for Outbound Option SIP Dialer to egress voice gateway. All subsequent transfers to endpoints must use SIP TCP. Secure SIP to SIP signaling	Contact center enterprise solutions do not support H.323. You can use SIP over UDP only for the Outbound Dialer. From the Outbound Option SIP Dialer to the egress gateway has to use UDP.
Codec	For VRU: G.711 mu-law and G.711 a-law For voice agents: G.711 mu-law, G.711 a-law, G.729, and G.729a For video: • Video track: H.264	Contact center enterprise solutions do not support iSAC or iLBC. Mixed codecs for Mobile Agent. Remote and Local ports must use the same codec. Mixed codecs for CVP prompts. CVP prompts must all use the same codec.

Capability	Supported	Notes
Media Resources	Gateway or Unified CM based: • Conference bridges • Transcoders and Universal Transcoders • Hardware and IOS Software Media Termination Points	For Unified CM-based resources, appropriately size Unified CM for this load.

IP Phone Support

For a list of supported phones, see the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html. Supported phones need the Built-In-Bridge (BIB), CTI-controlled features under SIP line side.

SCCP-based line side protocol is not supported in newer phones.

Administration Interfaces

Capability	Supported	Notes
Core Component Provisioning	• Gateways - CLI • Unified CVP - Web-based Packaged CCE Administration • Unified CCE - Web-based administration and thick client configuration tools • Unified CCMP for Unified CCE solutions, Unified CCDM for Cisco Hosted Collaboration Solution for Contact Center solutions • Cisco VVB - Web-based Packaged CCE Administration and operations console • Unified CM - Web-based administration • Cisco Finesse - Web-based Packaged CCE Administration • Unified Intelligence Center - Web-based administration	For provisioning, Packaged CCE does not support CCMP or CCDM.

Capability	Supported	Notes
Service Creation Environment	Unified CCE Internet Script Editor Unified CCE Script Editor CVP Call Studio	
Serviceability	Cisco Prime Collaboration - Assurance Unified System Command Line Interface (CLI) RTMT Analysis Manager Diagnosis SNMP syslog	Contact center enterprise solutions do not support RTMT Analysis Manager Analyze Call Path. Finesse supports RTMT only for log collection.

VRU and Queueing

This table lists the VRU and call queuing features that optimize inbound call management.

Capability	Supported	Notes
Voice Response Unit (VRU)	Unified CVP Comprehensive Model Type 10	
Caller Input	DTMF - RFC2833 Automatic Speech Recognition and Text-to-speech (ASR/TTS)	
Video	CVP and Video Basic CVP Video in Queue	
CVP Media Server	The CVP Media Server uses the third-party Microsoft Internet Information Services (IIS). The CVP installer adds the CVP Media Server coresident on the Unified CVP Server.	

Reporting

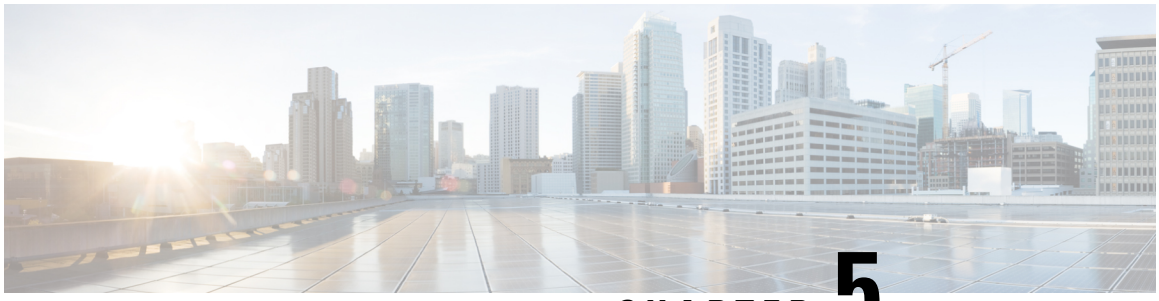
Capability	Supported	Notes
Reporting tools	Cisco Unified Intelligence Center Third-party reporting applications Custom reporting	For Packaged CCE, Exony VIM is supported for reporting only. Packaged CCE does not support Exony VIM provisioning features.

Capability	Supported	Notes
Database sources	Unified CCE AW-HDS-DDS Unified CCE Live Data Unified CVP Reporting	For a typical 1000 agent deployment with an average rate of 8 calls per second, the retention period is approximately 24 months. For a longer retention period, install an external HDS. To size the needs for your deployment, use the DB Estimator tool in the ICMDBA tool.
Database Integration	CVP Database Element	Unified CVP VXML Server supports connections to third-party Microsoft SQL Server databases.
Retention	All contact center enterprise solutions have a fixed retention size for the AW-HDS-DDS. For more retention, you need an external HDS-DDS node. Use the DB Estimator Tool in the ICMDBA tool to calculate the vDisk size based on your solution sizing and customer retention requirements. The DB vDisk of the AW-HDS-DDS can be custom-sized when you deploy the OVA. A 2000 Agent Reference Design can have up to 4 external HDS. For more information about the HDS sizing, see the *Cisco Collaboration Virtualization* page for your solution at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.	

Capability	Supported	Notes
Report capacities	Two hundred Unified Intelligence Center users can concurrently run: • Two real-time reports with 100 rows per report, with 10 columns each. • Two historical reports with 2000 rows, with 10 columns each. • Two live data reports with 100 rows, with 10 columns each. (Adjust this based on the deployment type whether LD runs or not). This is applicable for both Unified CCE and Packaged CCE solutions. **Note** • Do not run more than ten concurrent reports on any client machine. This is a combined limit for reports that run on the Unified Intelligence Center User Interface, Permalinks, and Dashboards on the client machine. • However, you cannot run ten concurrent reports for the 200 maximum reporting users on each node. • You have fewer reporting users on a node, they can run proportionally more reports. But, no client machine can exceed the ten report limit.	In addition, 30 users each running one real-time XML permalink and one historical XML permalink is supported. (This results in approximately 7200 real-time XML permalink executions per hour and 60 Historical XML permalink executions per hour.) The real-time reports have the capacity of 100 rows per report, with 10 columns each and the historical reports have the capacity of 2000 rows, with 10 columns each.

Third-Party Integrations

Option	Notes
Recording	Recording Methods: • CUCM-based (BiB) • Network-based Recording • CUBE Forking Optionally, you can use a third-party recording server integration.
Wallboards	Wallboard provide real-time monitoring of your service to customers. They display information on customer service metrics, such as number of calls waiting, waiting call length, and service levels.
Workforce Management	WFM allows the scheduling of multiple Contact Service Queue (CSQs) and sites. You can use a single WFM implementation worldwide.
Cisco Solution Plus	Refer to the Cisco Solution Plus program for supported options.
Automated Call Distributor (ACD)	You cannot use a third-party ACD in a Reference Design.



CHAPTER 5

Packaged Contact Center Enterprise Solution Design Considerations

Core Components Design Considerations

General Solution Requirements

Data Backup for Your Solution

Run data backup tools only during a scheduled maintenance window. If you use local SQL backups, make sure that the local machine has sufficient capacity. If not, back up to remote storage on the network.

NTP and Time Synchronization

Finesse Time Synchronization:

While time drift occurs naturally, it is critical to configure NTP to keep solution components synchronized. Cisco Finesse server and the Desktop client machines should be time synchronized with the same NTP server (Linux-based NTP v4) for the Duration fields within the Live Data reports to be updated correctly.

Live Data Time Synchronization:

Contact center enterprise solutions require that all parts of the solution have the same time. To prevent time drifts on Live Data reports, the NTP settings on the followinge VMs must be synchronized:

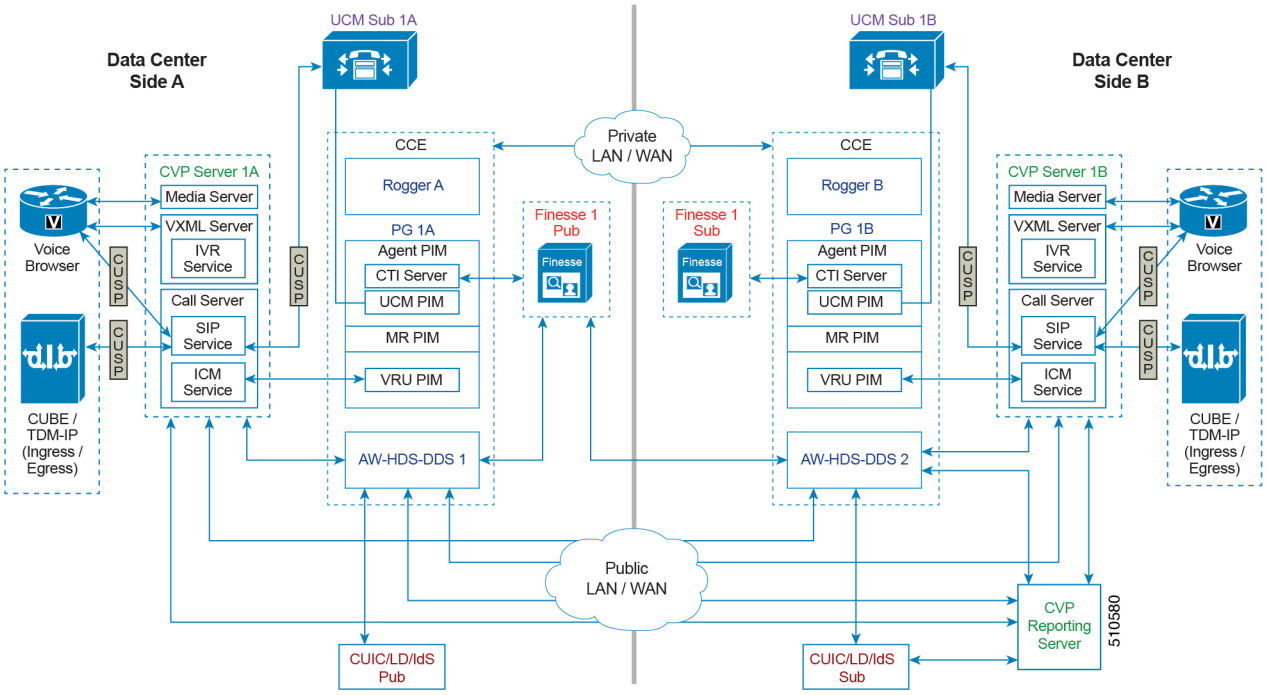
- Router
- Logger
- Administration & Data Server
- Unified Intelligence Center Publisher and Subscriber

Detailed Contact Center Enterprise Reference Design Topologies

Detailed 2000 Agent Reference Design

This figure shows the logical connections under general operating conditions between the sides in a redundant data center.

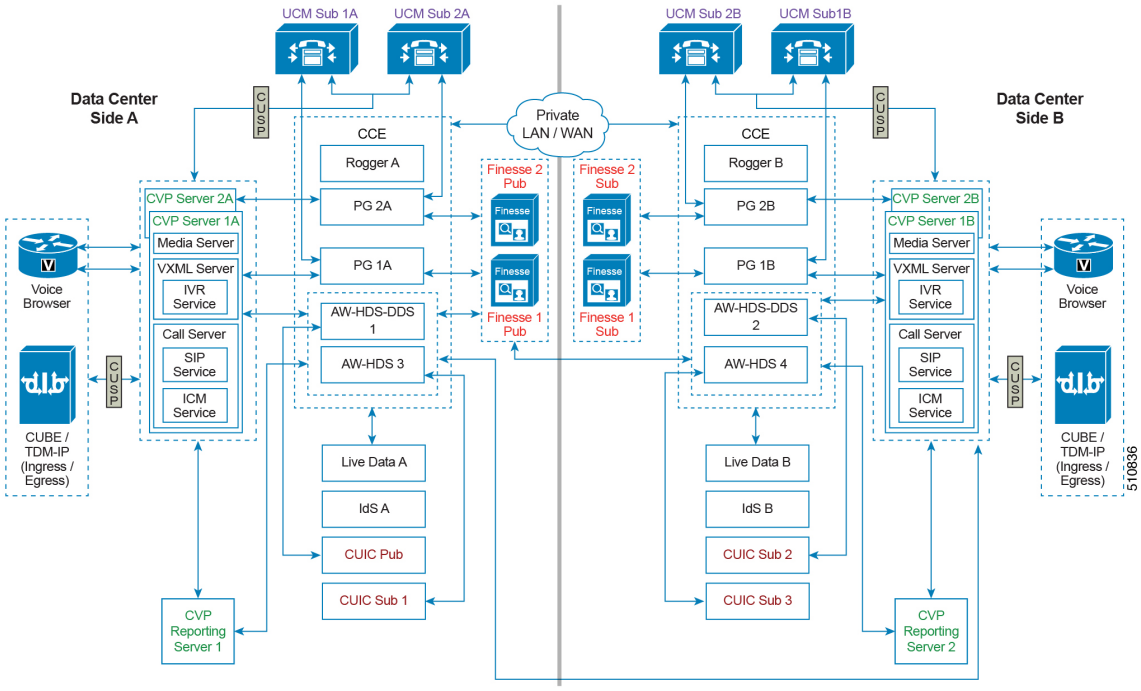
Figure 55: Detailed 2000 Agent Reference Design



Detailed 4000 Agent Reference Design

This figure shows the logical connections under general operating conditions between the sides in a redundant data center.

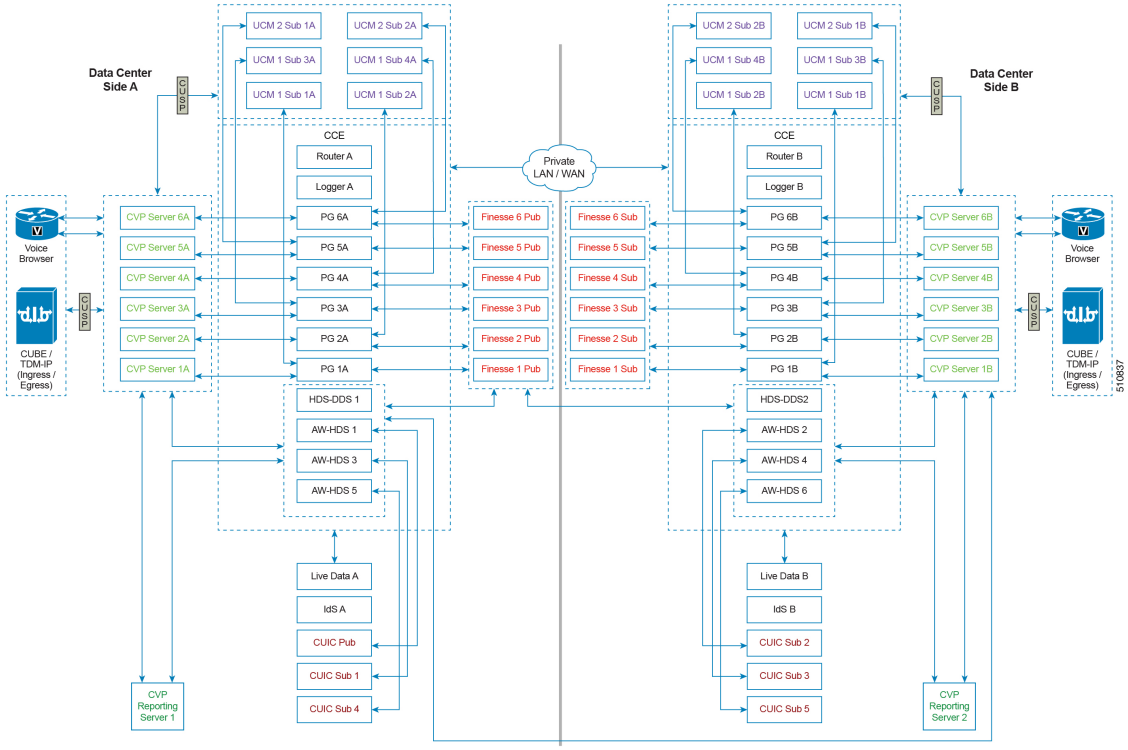
Figure 56: Detailed 4000 Agent Reference Design



Detailed 12000 Agent Reference Design

This figure shows the logical connections under general operating conditions between the sides in a redundant data center.

Figure 57: Detailed 12000 Agent Reference Design



Ingress, Egress, and VXML Gateways Design Considerations

IOS Gateway Roles

The contact center enterprise solutions use IOS Gateways for TDM ingress and VXML rendering. You can usually use any Cisco gateway that the solutions support can for either purpose or both. This table lists which call flows use each function:

Table 42: IOS Gateway Function Use by Call Flows

Call Flows	TDM Ingress	VXML Rendering
Reference Design		
Comprehensive Using Unified ICM Micro-Apps	Yes	Some
Comprehensive Using Unified CVP VXML Server	Yes	Some

Note You can use Cisco Virtualized Voice Browser as an alternative for VXML gateways.

When both Ingress and VXML are required, you can run both functions on the same gateways or designate some gateways for ingress and others for VXML. Use the following guidelines to determine whether to combine or split the functions:

- In branch office deployments, where the call is queued at the branch where it arrived, always combine the ingress and VXML functions.
- Where many non-CVP PSTN connections share the gateways, use separate gateways for each function.
- VXML-only gateways are less costly because they do not require DSP farms or TDM cards.
- For low call volumes, you generally combine the functions for redundancy purposes. If one combined gateway fails, the other gateway can still process calls at a reduced capacity.

The next decision is whether to use Cisco Integrated Service Router (ISR) or ISR-G2 Gateways.

The classic branch office in which to use ISR Gateways, includes:

- One of several sites where TDM calls arrive from the PSTN
- A site that is separated from the main site where most of your solution's equipment resides
- Each site uses one gateway

TDM-IP Gateway Design Considerations

For the most current information about the various digital (T1/E1) and analog interfaces supported by the various voice gateways, see the latest product documentation available at the following sites:

- **Routers**—http://www.cisco.com/cisco/web/psa/default.html?mode=prod&level0=268437899
- **Unified Communications Gateways**—http://www.cisco.com/cisco/web/psa/default.html?mode=prod&level0=278875240

Cisco Unified Border Element Design Considerations

The Cisco Unified Border Element (CUBE) is a session border controller (SBC) that provides connectivity between IP voice networks with SIP. Your solution can use physical CUBEs or virtual CUBEs. Your solution can only use CUBE in flow-through mode, where all calls are routed through CUBE.

Note Unlike flow-through mode, flow-around mode loses the ability to do DTMF interworking, transcoding, and other key functions, such as telephone and media capabilities.

Your solution needs a CUBE when replacing a TDM voice circuit with an IP voice trunk from a telephone company. CUBE serves as a feature demarcation point for connecting enterprises to service providers over IP voice trunks.

Note For outbound calls, physical CUBE supports Call Progress Analysis (CPA). Virtual CUBE does not support CPA.

Our testing shows that you can use CUBE in the following scenarios:

- SIP-to-SIP connectivity between a third-party SIP device and Unified CVP over Cisco-certified SIP trunks.
- SIP-to-SIP connectivity between Unified CM and Unified CVP.
- Coresidency of VoiceXML Gateway and CUBE for any of the above scenarios but with the limitation that the call flow does not work when the configurations listed here occur at the same time on the CUBE:
  - Survivability TCL script and incoming translation rules are configured under the same incoming dial-peer.
  - Header-passing between the call legs is enabled globally.

For more information about using the CUBE with contact center enterprise solutions, see *Cisco Unified Border Element for Contact Center Solutions* at http://cisco.com/en/US/docs/voice_ip_comm/unified_communications/cubecc.html.

Note For a listing of the maximum sessions that each CUBE supports, see the *Cisco Unified Border Element Configuration Guide* at http://www.cisco.com/c/en/us/support/routers/cloud-services-router-1000v-series/products-installation-and-configuration-guides-list.html.

Note Due to a limitation in Cisco IOS, the CUBE does not support midcall escalation or descalation from audio to video or the reverse.

Note Currently, CVP does not check the Allowed-Methods in the SIP message. As a result, it passes the `UPDATE` message from Ingress to Outbound leg although Outbound does not support `UPDATE` method.

Workaround: Disable the `UPDATE` message in SBC in the Ingress leg.

CUBE Deployment Restrictions

Note the following restrictions when deploying CUBE with SIP Trunks:

- Configure CUBE in media pass-through mode, the default mode on the dial-peer, in the Unified CVP deployment. Media flow-around mode is not supported or validated.
- CUBE does not support passing the Refer-To header URI designation from CVP when a REFER call flow is initiated. CUBE rewrites the destination address based on the dial peer configuration. Therefore, configure the dial plan on CVP and CUBE.
- You cannot use REFER passthrough with Survivability. The script does not let REFER messages be relayed to a SIP service provider regardless of other CUBE configuration.
- You cannot use REFER consume with Survivability and Router Requery. Survivability always accepts the REFER, even if the transfer does not complete. Unified CCE deems the transfer successful and does not attempt to requery.

- You cannot use survivability with service provider Alternate Destination Routing (ADR). Manipulation in the script does not let error messages (ring-no-answer or busy) reach the service provider. Use manipulation in the Remote-Party-ID header instead.
- If GTD is present on the incoming call or if Unified CCE sets a value for the UUI variable, Unified CVP sends a BYE immediately after outpulsing digits in a DTMF transfer. If you need a delay between the digits, use a comma at the end of the label.
- If GTD is not present on the incoming call, Unified CCE does not set a value for the UUI variable. Then, the service provider does not disconnect a call after receiving digits in a DTMF transfer. Unified CVP sends a BYE request after the SIP.ExternalTransferWait timer expires.
- Solutions with Courtesy Callback require survivability.

Note Call Survivability is supported on CUBE HA mode with the following restrictions:

- If there is a courtesy callback (CCB) registered with CVP, then post switchover CCB is not supported.
- Only call survivability TCL script is supported with CUBE high availability. Other TCL based services are not supported.
- Only the active calls will be check pointed. (Calls which are connected - 200OK / ACK transaction completed). Calls in transition state will not be check pointed.

Cisco ASR 1000 Series as a Unified Border Element

Unified CVP supports Cisco IOS XE software with the following limitations:

- The ASR 1000 Series gateways do not support VXML. So, route the VRU leg of the call to a separate VXML Gateway. Do not use the `Send To Originator` setting on the CVP Call Server to route the VRU leg of the callback to the originating ASR CUBE Gateway. Route standalone CVP calls to a separate VXML Gateway.
- Unified CVP does not support the global `Pass Thru SDP` setting on the ASR 1000 Series gateways.
- The Courtesy Callback call flow does not work if you configure the ASR as CUBE for the media flow-around, instead of the media flow-through.
- Typically, you position a proxy server behind the session border controller. If the proxy is in front of the ASR session border controller, use the proxy servers to perform UDP to TCP Up-Conversion when receiving large packet SIP messages. In this case, turn off the proxy servers to ensure that UDP transport is used for the connection on the inbound call.
- Do not use the following Survivability.tcl options on the ASR. These options are traditionally for POTS dial peers:
  - ani-dnis-split.
  - takeback-method.
  - -- *8.
  - -- hf.
  - icm-tbct.

- digital-fxo.

- The following Survivability.tcl options are not supported:
  - aa-name—This option is not supported because ASR does not support the CME auto-attendant service.
  - standalone—This option is not supported because ASR does not support VXML.
  - standalone-isntime—This option is not supported because ASR does not support VXML.
- Due to ASR limitations, the following features are not supported:
  - Refer with Re-query
  - Legacy Transfer Connect using DTMF *8 label
- ASR 1000 does not terminate the TDM trunks. Therefore, the following TDM Gateway features do not apply to ASR 1000:
  - PSTN Gateway trunk and DS0 information for SIP calls to Unified CCE
  - Resource Availability Indication (RAI) of DS0 trunk resources through the SIP OPTIONS message to Unified CCE

Note If your solution uses ASR 1000 Series gateways, it requries an Assessment to Quality (A2Q) review. This review is required for new contact center enterprise solutions and existing solutions that are upgrading to the ASR 1000.

Cisco ISR as a Unified Border Element

Unified CVP supports ISR with the following limitations:

- The Courtesy Callback call flow does not work with ISR as CUBE configured for the media flow-around. Configure it for the media flow-through instead.

VXML Gateway Design Considerations

Note You can use Cisco Virtualized Voice Browser as an alternative for VXML gateways.

VXML Gateway with DTMF or ASR/TTS

The VXML Gateway allows customers to interact with the VXML browser through DTMF tones or ASR/TTS. Because the gateway does not have PSTN interfaces, voice traffic is sent using Real-Time Transport Protocol (RTP) to the VXML Gateway. The RFC 2833 uses in-band signaling in RTP packets to transmit DTMF tones. A VXML with DTMF or ASR and TTS allows you to increase the scale of the deployment and support hundreds of VXML sessions.

In a branch office topology, you can deploy a separate PSTN Gateway and a VXML Gateway to provide an extra layer of redundancy In addition, provide support for Survivable Remote Site Telephony (SRST) at the branch office.

VXML Over HTTP

The VXML Server and Voice browser communicate with request-response cycles using VXML over HTTP. Uniform Resource Identifiers (URI) link the VXML documents together. Users input information by web forms similar to HTML. The forms contain input fields that the user edited and sent back to a server.

Resources for the Voice browser are located on the VXML Server. These resources are VXML files, digital audio, instructions for speech recognition (Grammars), and scripts. The VXML browser begins every communication process with the Voice application as a request to the VXML Server. The VXML files contain grammars that specify expected words and phrases. A link contains the URL for the Voice application. The browser connects to that URL when it recovers a match between spoken input and one of the grammars.

Note The CVP installer installs the CVP Call Server, the CVP VXML Server, and the Media Server together.

The following points are key to determining the VXML Server performance:

- QoS and network bandwidth between the Web application server and the voice gateway
- Performance on the VXML Server
- Use of prerecorded audio versus Text-to-Speech (TTS)

  Voice user-interface applications tend to use prerecorded audio files wherever possible. Recorded audio sounds better than TTS. Choose the quality of the prerecorded audio fileso that it does not impact download time and browser interpretation. Make the recordings in the 8-bit mu-law 8 kHz format.

- Audio file caching

  Ensure that the voice gateway is set to cache audio content. Caching prevents delays in downloading files from the media source.

- Use of Grammars

  You can discover problems in a voice application only through formal usability testing or observation of the application in use. Poor speech recognition accuracy is a common problem with voice applications. It is most often caused by poor grammar implementation. When users mispronounce words or say things that the grammar designer did not expect, the recognizer cannot match their input against the grammar. Another common problem for grammars is many difficult-to-distinguish entries. These entries result in many incorrectly recognized inputs and decrease the performance of the VXML Server. Improve the recognition accuracy by analysizing its performance and tuning the grammar appropriately.

Distributed Gateways

These sections discuss the types of voice gateways and their effects in a distributed deployment.

Ingress or Egress Voice Gateways at the Branch

Your solution can use Ingress Voice Gateways at a branch office to provide callers with access by local phone numbers, instead of by centralized or nongeographic numbers. This capability is important for solutions that span multiple countries.

Your solution can use Egress Gateways at branches either for a localized PSTN breakout or to integrated decentralized TDM platforms into your solution.

The other components of your solution are centrally located. The WAN links provide data connectivity from each branch location to the main site.

Ingress or VXML Gateway at the Branch

Other voice services that run at the branch can affect the Ingress or VXML Gateways. For example, if the branch is a remote Unified CM site, Unified CM can support both ACD agent lines and nonagent lines. This deployment uses the PSTN gateway for new contacts and traffic from the nonagent lines. When a branch has the VXML and Voice Gateway functions on separate devices, ensure that the dial plan sends the VRU leg to the local VXML resource. This is because the Unified CVP Call Server `settransferlabel` label applies only to coresident VXML and Voice Gateway configurations.

Colocated VXML Servers and VXML Gateways

Your solution can either have all gateways and servers centralized or have a set of colocated Unified CVP VXML Servers and VXML Gateways at each site.

Colocation has the following advantages:

- A WAN outage does not affect self-service applications.
- VXML uses no WAN bandwidth.

Colocation has the following disadvantages:

- Replicated branch offices require extra Unified CVP VXML Servers.
- Deploying applications to multiple Unified CVP VXML Servers creates more overhead.

Gateways at Branch with Centralized VXML Server

Advantages of centralized VXML:

- Administration and reporting are centralized.
- Branch offices can share Unified CVP VXML Server capacity.

Disadvantages of centralized VXML:

- Branch survivability is limited.
- Requires more WAN bandwidth for VXML over HTTP traffic.

Local Trunks in Contact Center Enterprise Solutions

Contact center enterprise solutions have two options for local trunks at the customer premise:

- Cisco Unified Border Element—Enterprise at the customer premise
- TDM gateway at the customer premise

Note Transcoding resources are not deterministically picked from the local customer premise gateway.

CVP Design Considerations

CVP Call Server Design Considerations

Unified CVP Algorithm for Routing

When you set up a dial plan and call routing, you can combine Unified CVP features to achieve the required effect. For example, you can use Location Based CAC, SigDigits, SendToOriginator, LocalSRV, and Use Outbound Proxy.

CVP uses this process to formulate the destination SIP URI for the outbound calls from Unified CVP. This description covers CONNECT messages that include labels from the Unified CCE (for example, VXML Gateway, and Unified CM). It also applies for calls to the ringtone service, recording servers, and error message playback service.

Note This process only describes calls using the SIP subsystem, which includes audio only and basic video SIP calls.

CVP supports the `SendToOriginator` algorithm only for a colocated IOS VXML Gateway and Ingress Voice Gateway. Cisco Virtualized Voice Browser (VVB) does not support this algorithm because the gateways cannot be colocated when you use Cisco VVB.

The process for creating the destination SIP URI host portion for outbound calls, which includes the Unified CCE label, is as follows:

1. The process starts with the Unified CCE label. The Unified CCE subsystem might already have inserted the Location siteID. If you're using SigDigits, they are prepended. For network VRU labels, the Unified CCE subsystem passes in the entire prefix and correlation ID as the label.
2. If `SendtoOriginator` is matched for the Unified CCE label, the Unified CVP algorithm uses the IP or hostname of the caller (Ingress Voice Gateway). The gateway returns the SIP URI.

   The setting for `SendtoOriginator` only applies to callers on Cisco Ingress Voice Gateways (the SIP UserAgent header is selected). Non-Cisco IOS Gateways do not have the CVP bootstrap service that the Cisco IOS VXML Gateway uses.
3. If `use outbound proxy` is set, then use the host of the proxy and return SIP URI.
4. If `local static route` is found for the label, return the SIP URI.

Note If `local static route` is not found, the algorithm throws a `RouteNotFoundException` exception.

Consider these points for calls using the SIP subsystem:

- To avoid complex Dialed Number strings, do not use the SigDigits feature with Locations CAC siteIDs.
- You can specify an Outbound Proxy FQDN as a Server Group FQDN (local SRV FQDN). You can also configure a local static route destination as a Server Group FQDN.
- Ringtone DN (91919191), Recording Server (93939393), and Error message services (92929292) follow the same process.

- `SendtoOriginator` can work with a REFER label.
- A REFER label can work with the SigDigits setting.

CVP VXML Server Design Considerations

The complexity of your VXML applications affect the performance of the VXML Server. Load test your application for memory leaks and application deadlocks to maintain acceptable VXML Server performance.

CVP Media Server Design Considerations

Voice Prompt Deployment and Management

You can deploy voice prompts with the following methods:

- Local File System

  Store the voice prompt files on a local system. Audio prompt retrieval uses no bandwidth. With this method, Voice Browsers do not have to retrieve audio files for playing prompts, so WAN bandwidth is not affected. However, to change a prompt, you change it on every Voice Browser.

  - **IOS VXML Gateway**—Prompts are deployed on flash memory.

    An IOS VXML Gateway is either a VXML Gateway or a PSTN Gateway, which has the Ingress Voice Gateway and VXML Gateway colocated. Store only critical prompts here, such as error messages or messages for when the WAN is down.

    When recorded in G.711 mu-law format, typical prompts are about 10 to 15 KB in size. For these gateways, size the flash memory by factoring in the number of prompts and their sizes. Also leave space for storing the Cisco IOS image.

  - **Cisco VVB**—Prompts are uploaded on the local file system.

    Cisco VVB includes built-in CVP prompts. You can change the `Error` tone default prompt through the **Cisco VVB Administrator console**.

- Media Server

  Each local Voice Browser, if configured properly, can cache many prompts, depending on the size of the prompts. Cisco VVB can cache up to 512 MB and Cisco IOS can cache up to 100 MB. To test whether your Media Server is appropriately serving the media files, specify the URL of a prompt on the Media Server in a browser. Your web browser downloads and plays the `.wav` file without any authentication.

The design of a Media Server deployment depends on the following factors:

- The number of media files that each gateway plays
- The network connectivity between the gateway and the Media Server
- How often you change the media files

Design Considerations for Large Numbers of Media Files

If your gateway plays many different media files to your customers, the gateway might not have space to cache all the media files.

For example, consider an enterprise with many agents. Each agent has their own agent greeting file. You cannot cache all those files in the gateway flash memory.

Colocated Media Server with Voice Browser

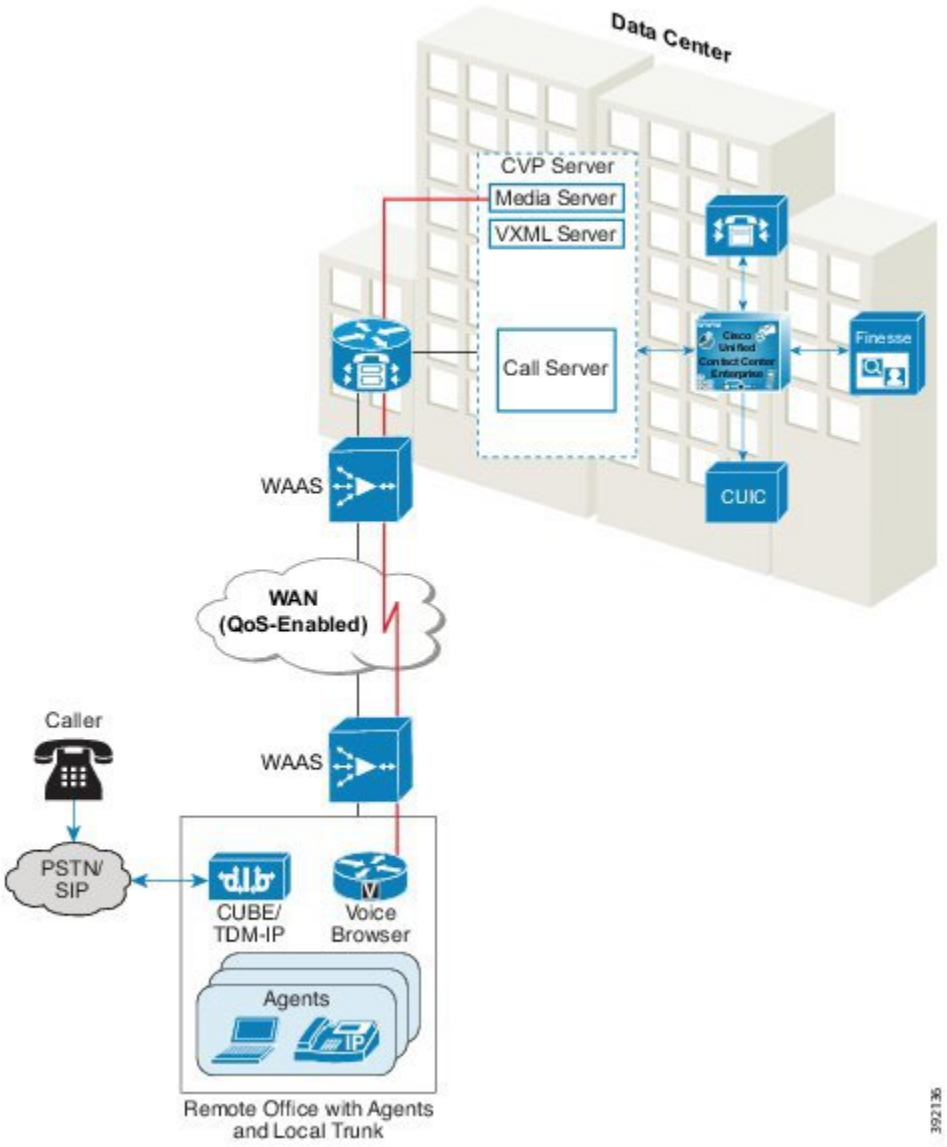
One approach to this problem is to colocate the Media Server with the Voice Browser. When a Media Server and Voice Browser coexist on a LAN with plenty of bandwidth, the download of prompts does not add noticeable delay.

Media Server and Voice Browser Distributed Over a WAN

Your solution can have a Media Server separated from a Voice Browser across a WAN.

This figure shows a distributed deployment over WAN.

Figure 58: Distributed Deployment Over WAN



The download of the media files across a high latency WAN to the Voice Browser can add noticeable delays. The delay can greatly affect the user experience. The delay is proportional to the size and number of media files that are transported across the WAN. You can optimize the delay with Cisco Wide Area Application Services (WAAS).

Design Considerations for Media Streaming

Consider the following factors for both the LAN deployment and the WAN accelerator deployment:

- Maximum network round-trip time (RTT) delays of 200 ms.

  An example is the transfer of files from the CVP Operations Console to the Ingress or the VXML Gateway using Bulk Administration File Transfer (BAFT).

- Maximum number of streaming sessions supported on each gateway with no additional overhead of video with media forking.

The following table lists the preferred media streaming method for various deployments:

Scenario	Frequency of Change	Over LAN	Over WAN
Small number of files	Rare	Cached	Cached
Small number of files	Often	Streamed or Cached	Streaming with WAAS
Large number of files	Rare	Streamed	Streaming with WAAS
Large number of files	Often	Streamed	Streaming with WAAS

Design Considerations in using VVB Media Streaming

Streaming service provides continuous audio streaming to callers.

Earlier, continuous audio streaming was supported in IOS Voice XML gateway. In Cisco VVB, the support for audio streaming started in release 12.5 via HTTP(S).

Cisco VVB creates a single connection with a streaming server and broadcasts the audio to all callers while they wait for the call to be answered by an agent. The audio streamed can be a live music or promotional audio. Caching is disabled for the audio streaming.

Cisco VVB requires an internet connection to establish a connection with the streaming server. By default, the audio is streamed for a maximum of 30 minutes per caller. Since there's only one connection with the streaming server, whenever callers connect, they hear the same audio as the rest of the callers (not from the beginning of the audio streaming).

The maximum number of connected callers for a single connection streaming is 50. The application developer can use the streaming service to play a streaming URL from a local or cloud streaming server.

Note

- Cisco VVB currently supports HTTP(S) streaming for G711 A-law / U-law 8k wav format.
- You can configure the maximum streaming time the connection is active using the *com.cisco.voicebrowser.streaming.timeout* property.
- You can identify the user using the *http.streaming.useragent* property. This is an optional property when enabling the VVB audio streaming feature.
- If there are any network issues, the streaming stops. However, the application flow continues.

For more information on the streaming property, see the *Custom VoiceXML Properties* section in the Element Specifications for Cisco Unified CVP VXML Server and Call Studio.

Design Considerations for Media File Deployment

No Support for TCP Socket Persistence

Unified CVP does not support TCP socket persistence.

WAN Acceleration Support

The Cisco Wide Area Application Services (WAAS) system is a set of devices called Wide Area Application Engines (WAEs). The WAEs work together to optimize TCP traffic over your network. Cisco WAAS uses TCP optimization techniques and application acceleration features to overcome the most common challenges in transporting traffic over a WAN. When deployed at the periphery of the network on the VXML Gateway side, Cisco WAAS performs these functions:

- Changes the TCP header to optimize the traffic.
- Acts as a large HTTP cache located locally.
- Uses compression algorithms to further reduce the traffic.
- Reduces traffic with Data Redundancy Elimination (DRE) techniques.

Cisco WAAS is deployed in inline mode where whole data is forced to pass through the Cisco WAAS.

Media File Deployment on IOS Gateway

Nonstreaming and Streaming Modes

In nonstreaming mode, the VXML gateway downloads the entire audio file from the HTTP server before the Media Player can start playing the prompt. This can cause a delay for the caller. For small files, the delay is only a few milliseconds. You can avoid the delay for larger files by using either caching or streaming mode.

In streaming mode, the Media Player streams the audio in media chunks from the HTTP server to the caller. The Media Player can start playing a prompt when it recieves the first chunk. In streaming mode, the size of the audio prompt does not add any delay for the caller. However, the back-and-forth interactions to fetch the media file in chunks can degrade performance.

Caching the audio files in memory reduces the advantage of streaming large files directly from the HTTP server.

Media File Cache Types

There are two types of cache for storing media files:

HTTP Client cache
In nonstreaming mode, the HTTP Client cache stores the entire media file. In streaming mode, the HTTP Client cache stores the first chunk of the media file. The HTTP Client cache stores 100 MB of prompts, in either mode. Any file that is larger than the configured HTTP Client memory file size is not cached.

VRU Media Player cache
Nonstreaming mode never uses the VRU Media Player cache. In streaming role, the VRU Media Player cache stores all chunks of the file. In nonstreaming mode, the VRU Media Player cache stores 16 MB. In streaming mode, it can store 32 MB.

Query URL Caching

A query is a URL that has a question mark (?) followed by one or more **name=value** attribute pairs in it. The Unified CVP VXML Server uses query URLs when generating its dynamic VXML pages. Because each call is unique, data retrieved from a query URL can waste the cache memory. The data is also a possible security risk, because the query URL can contain information such as account numbers or PINs.

Cisco IOS disables Query URL caching by default. To ensure that it is disabled, enter a **show run** command in Cisco IOS and ensure that the following Cisco IOS command does not appear:

```
Gateway configuration: http client cache query
```

Media File Deployment on Cisco VVB

Cisco VVB includes an HTTP client. The client fetches VXML documents, audio files, and other file resources and stores them in flash memory.

A caching property is associated with VXML resources, audio prompts, grammar files, and script files.

By default, Query URLs are not cached. A query is a URL that has a question mark (?) followed by one or more **name=value** attribute pairs in it.

Cache Aging

The HTTP Client manages its cache by the freshness of each cached entry. Whether a cached entry is fresh or stale depends on its `Age` and `FreshTime`. `Age` is the elapsed time since the file was last downloaded from the server. `FreshTime` is the expected time for the file to stay in the HTTP Client cache since the file was last downloaded.

Several variables affect the `FreshTime` of a file, such as HTTP message headers from the server and the cache refresh value.

The `FreshTime` of a file is determined in the following sequence:

1. When downloaded, if the file has an HTTP message header with the `Cache Control: max-age` header, the `FreshTime` is the `max-age`.
2. If Step 1 does not apply, the `FreshTime` is the `Expires` header minus the `Date` header.

Note The HTTP/1.1 specification, *RFC 2616 (HyperText Transport Protocol)*, recommends the use of either the `Cache Control: max-age` header or the `Expires` header.

3. If the previous headers are not present, the `FreshTime` is 10% of the `Date` header minus the `Last-Modified` header.

For the Cisco IOS VXML Gateway, you can assign a `FreshTime` value to the files with the `http client cache refresh` command. But, that value only applies if the previous sequence fails to set a value.

Stale files are refreshed only when needed. A stale cached entry stays in the cache until it is removed to make room for a new file, based on these conditions:

- The cached entry becomes stale.
- Its refresh count is zero (0); that is, the cached entry is not being used.
- The cache needs the memory space to make room for other entries.

When the Age exceeds the FreshTime and the file needs to be played, the HTTP Client checks with the media server to determine whether or not the file has been updated. When the HTTP Client sends a GET request to the server, it uses a conditional GET to minimize its impact on network traffic. The GET request includes an If-Modified-Since in the headers sent to the server. With this header, the server returns a 304 response code (Not Modified) or returns the entire file if the file was updated recently. When the Age exceeds the FreshTime and the file needs to be played, the HTTP Client checks with the media server to determine whether or not the file has been updated. When the HTTP Client sends a GET request to the server, it uses a conditional GET to minimize its impact on network traffic. The GET request includes an If-Modified-Since in the headers sent to the server. With this header, the server returns a 304 response code (Not Modified) or returns the entire file if the file was updated recently.

This conditional GET applies only to nonstreaming mode. In streaming mode, the HTTP Client always issues an unconditional GET. There is no If-Modified-Since header included in the GET request that results in an unconditional reload for each GET in streaming mode.

CVP Reporting Server Design Considerations

The CVP Reporting Server houses the Reporting Service and hosts an IBM Informix Dynamic Server (IDS) database management system. The database's schema is available to enable you to write custom reports for the database. The Reporting Service does not itself perform the database administrative and maintenance activities, such as backups or purges. However, Unified CVP provides access to such maintenance tasks through the Operations Console.

The Reporting Service:

- Provides historical reporting of self-service activity in your contact center. The service summarizes call activity for the contact center managers.
- Can also provide operational analysis of various VRU applications.
- Receives reporting data from the IVR Service of VXML server and the SIP Service. The Reporting Service transforms and writes this data into the Informix database.

Your solution can use either a single of multiple CVP Reporting Servers. A single Reporting Server does not necessarily represent a single point of failure. The database management system provides data safety and security. Your solution can tolerate temporary outages due to persistent buffering of information on the source components.

If your solution uses multiple Reporting Servers, you can associate each CVP Call Server with only one Reporting Server. Also, your reports cannot span multiple Informix databases.

Note Unified CVP subcomponents cannot synchronize the machine time themselves. Provide a cross-component time synchronization feature, such as NTP, to assure accurate time stamps for logging and reporting.

CVP Reporting Server Features

Consider the following points when designing your solution with the CVP Reporting Server:

- You can size the Informix database up to 100 GB. You cannot use a 2 GB or smaller database in a production environment.
- The Reporting Server supports the Analysis Manager tool. The Analysis Manager can query the Reporting Server with an authenticated user's credentials.
- The Reporting Server aggregates Unified CVP data in 15-minute increments. Cisco Unified Intelligence Center provides templates to display call data and dominant path information at 15-minute, daily, and weekly intervals.
- All metadata for administrative processes is in the `Ciscoadmin` database. This location removes the tables from the general view of reporting users.
- All database backup files are compressed and stored on the Reporting Server. The backup file is called **cvp_backup_data.gz** and is stored on the %INFORMIXBACKUP% drive in the **cvp_db_backup** folder.
- Using the system CLI, you can make the request to list log files on the Reporting Server (**show log**). This request includes the Informix Database Server Engine logs. The **show tech-support** command also includes these files.
- With the debug level 3 (or 0) command from within the System CLI, you can turn on and turn off the debug. When turned on, this command generates trace files for all administrative procedures, Purge, Statistics, and Aggregator.

Note After the command is turned on, trace files place an elevated burden on the database.

- Log data for administrative procedures are written on a nightly basis to the **%CVP_HOME%\logs** folder.
- All the **StartDateTime**, **EndDateTime**, and **EventDateTime** values are stored as UTC in Reporting Server tables.
- Transfer Type data and Transfer Labels for SIP call events are stored in the call event table.
- Summary purge results are logged in the log table.
- Three new scheduled tasks have been added to the Reporting Server scheduler:
  - **CVPSummary**, which builds summary tables.
  - **CVPCallArchive**, which archives Callback data to maintain callback database performance.
  - **CVPLogDump**, which extracts the administrative logs on a nightly basis.

CVP Backup and Restore

Using the Operations Console, you can schedule daily database backups or run database backups on-demand. In a major failure, you can restore the database manually to the last backup time. This limits the loss of data to 24 hours at most.

CVP Call Studio Design Considerations

When you design applications in CVP Call Studio, keep the applications small and closely mapped to your business flows. Large applications are harder to maintain and work with. Maintain a balance between subflows and independent applications.

Unified CVP Coresidency

To calculate the number of servers required with SIP call control, use this formula:

```
(Self Service + Queue and Collect + Talking) / 3000, rounded up
```

Where:

Self Service is the number of calls that require SIP call control and run an application on the VXML Server.

Queue and Collect is the number of calls that require SIP call control and run an application using Microapps only on the Call Server.

Talking is the number of calls at agents.

The following example applies for VXML and HTTP sessions only.

```
((3000) + (500) + 3700) / 3000 = 3 servers
```

If you use CUBE as a Session Border Controller (SBC) for flow-through calls to handle VXML requirements, use the sizing information provided in the example.

If you use CUBE as a Session Border Controller (SBC) to handle flow-through calls only (no VXML), then consider Voice Activity Detection (VAD) and see the sizing information in the *Cisco Unified Border Element Ordering Guide*, available at http://www.cisco.com/c/en/us/products/collateral/unified-communications/unified-border-element/order_guide_c07_462222.html.

Contact Center Enterprise Design Considerations

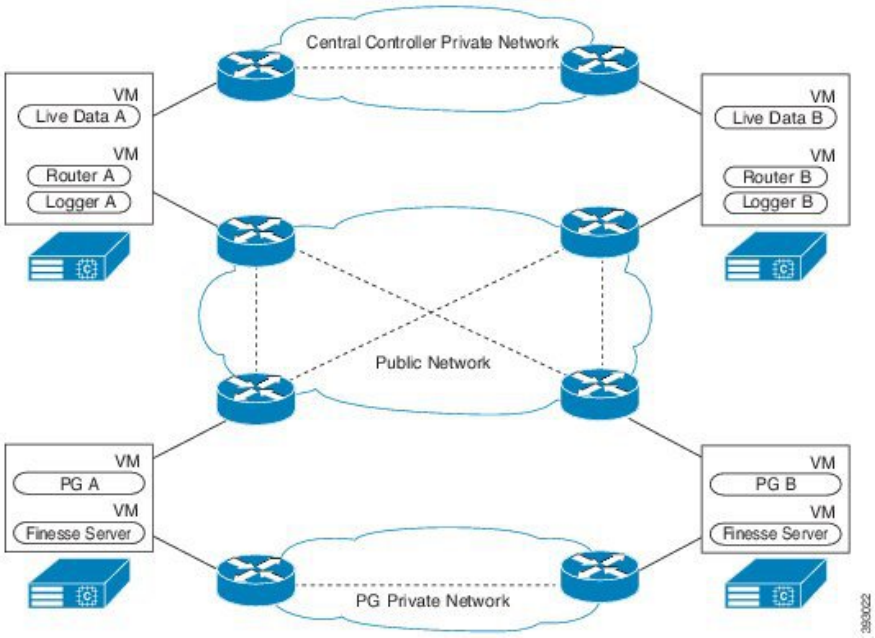
The contact center enterprise software provides enterprise-wide distribution of multichannel contacts. It can support inbound and outbound phone calls, web collaboration requests, email messages, and chat requests. It can also support geographically separated contact centers. The contact center enterprise software is an open standards-based solution that includes routing, queuing, monitoring, and fault tolerance capabilities.

Note Unified CCE is the name for both one of the contact center enterprise solutions and one of the core components for all of the solutions.

Router Design Considerations

You can geographically distribute redundant Unified CCE servers or locate them at the same physical site. In a production deployment, the Router and Logger must connect over a private network that is isolated from the public network.

Figure 59: High Availability Design for Central Controller



Note You can use the same private network path for the Central Controller and PGs.

Logger Design Considerations

The design of the Logger database holds two weeks of data usually. This period allows enough time for the data to replicate to the AW-HDS-DDS.

The Logger uses the same private network path as its Router.

Peripheral Gateway Design Considerations

Agent Peripheral Gateway Design Considerations

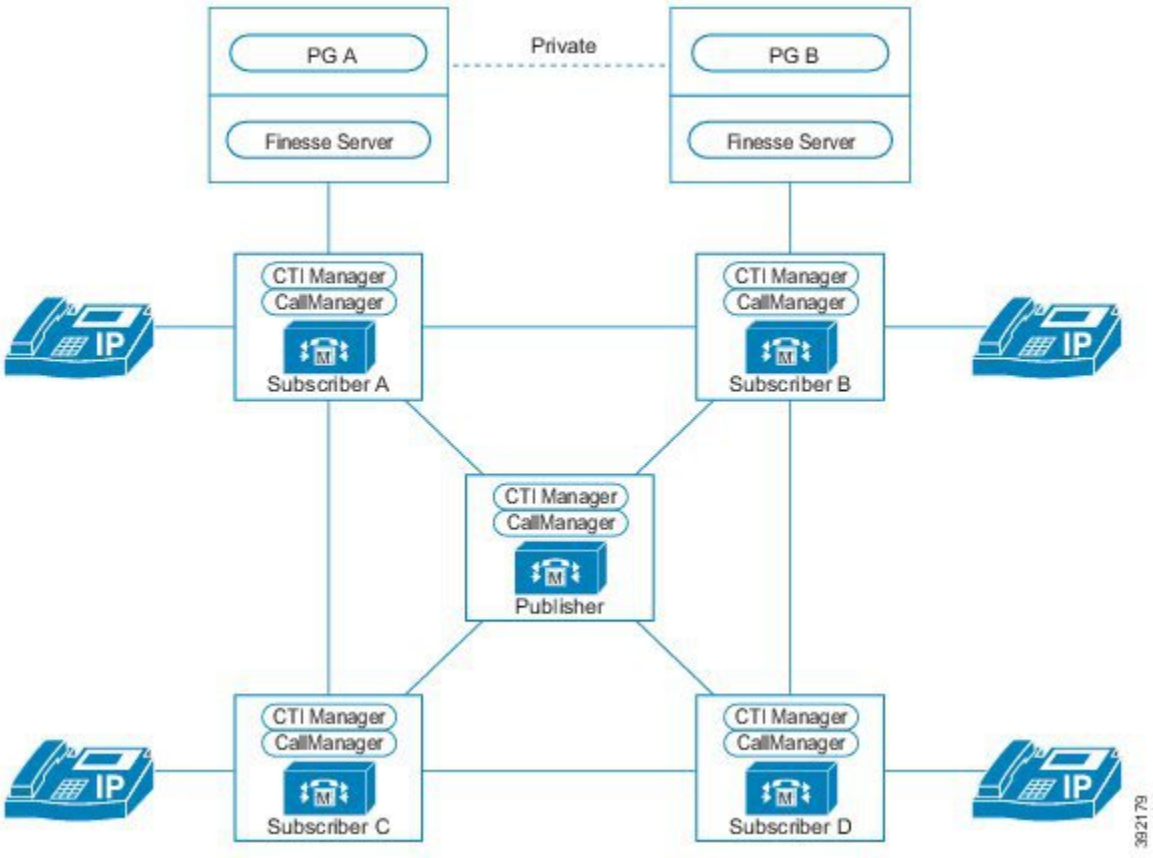
The Agent PG communicates with the Unified CM cluster through the CTI Manager. An Agent PG can control agent phones and CTI route points anywhere in the cluster. The Agent PG registers with the CTI Manager on a Unified CM subscriber in the cluster. The CTI Manager accepts all JTAPI requests from the PG for the cluster. When the PG requests a phone or route point on another subscriber, the CTI Manager forwards the request to the other subscriber.

Note The *Agent PG* is the PG that includes the Unified CM PIM. It is sometimes called a Unified CM PG.

A fault-tolerant design deploys Agent PGs in a redundant configuration, because a PG only connects to the cluster through a single CTI Manager. If that CTI Manager fails, the PG cannot communicate with the cluster. A redundant PG provides a secondary pathway through a different CTI Manager on a different subscriber in the cluster.

The minimum design for a high-availability cluster is one publisher and two subscribers. If the primary subscriber fails, the devices rehome to the secondary subscriber and not to the publisher for the cluster.

Figure 60: High Availability Design for Unified CM Cluster



The redundant PGs keep in synchronization through a private network that is isolated from the public network. If the two PG servers are geographically distributed, use a separate WAN connection for the private network. To avoid a single point of failure in the network, do not use the same circuits or network gear as for the public network.

Within the Agent PG, the JTAPI Gateway and Unified CM PIM manage the connectivity to the cluster. The JTAPI Gateway handles the JTAPI socket connection protocol and messaging between the PIM and the CTI Manager. The PIM manages the interface between Unified CCE, the JTAPI Gateway, and the cluster. It requests specific objects to monitor and handle route requests from the cluster. The PG starts the JTAPI Gateway and PIM automatically as node-managed processes. The PG monitors the processes and automatically restarts them if they fail.

The JTAPI services from both redundant Agent PGs sign in to the CTI Manager after initialization. Agent PG-A signs in to the primary CTI Manager; Agent PG-B signs in to the secondary CTI Manager. Only one PG in each pair actively registers and monitors phones and CTI route points. The redundant PG runs in hot-standby mode. The redundant PG signs into the secondary CTI Manager only to initialize the interface and make it available for a failover. This arrangement significantly decreases the time for the failover.

When the system starts, the PG that first connects to the Router server and requests configuration information is the active PG. The Router ensures that the PG with the best connection becomes active. The nominal designations of “Side A” and “Side B” do not affect which PG becomes active. During a PG failover caused

by a private link failure, a weighting mechanism chooses which PG is active to minimize the impact on the contact center.

If calls arrive at the CTI Route Points before the PIM is operational, the calls fail unless you set up a recovery number. Place the recovery number in the route points' `Call Forward on Unregistered` or `Call Forward on Failure` setting. For example, you can set the recovery numbers to the Cisco Unity voicemail system for the Auto Attendant.

Note You cannot use the DN for a CTI Route Point on a different CTI Route Point in another partition. Ensure that DNs are unique across all CTI Route Points on all partitions.

Active PG Shutdowns

Avoid shutting down an active peripheral gateway service in your production environment. This causes a service interruption of a minute or more while the other side connects and activates. The length of the interruption depends on the size of the configuration and the type of peripheral. For example, the VRU peripheral usually takes less time. The other side for the VRU might take 30 seconds or less to reactivate.

Voice Response Unit Peripheral Gateway Design Considerations

In the standard three PG model, the VRU PG includes two PIMs with a 1:1 pairing to a CVP servers on Side A and Side B.

Media Resource Peripheral Gateway Design Considerations

In the standard three PG model, the MR PG includes PIMs to support these functions:

- Outbound Option
- Enterprise Chat and Email
- Customer Collaboration Platform—This PIM handles Task Routing and Agent Request.
- Third-party integrations

Administration & Data Server Design Considerations

Administration & Data Server Limits by Reference Design

You can deploy only so many Administration & Data Servers for each Logger.

Table 43: Administration & Data Server Deployment Limits Per Logger

Component on each Logger side	2000 Agent	4000 Agent	12,000 Agent
AW-HDS-DDS	1 per side, installed on the same server with the core components.	2 per side	NA
HDS-DDS	NA	NA	1 per side

Component on each Logger side	2000 Agent	4000 Agent	12,000 Agent
AW-HDS	Optionally, either 1 AW-HDS or AW-HDS-DDS per side, installed on a separate server from the core components	NA	3 per side
Real-Time Distributors only[57]	2 per side	2 per side	5 per side

[57] These AWs are for configuration only. You install them off the servers shown in the Reference Design layouts.

Note Each Real-Time Distributor can support 64 users.

Live Data Server Design Considerations

Reporting Clients by Live Data Configuration

Use the Live Data coresident configuration for 2000 Agent Reference Designs. For solutions with a standalone Live Data server, you typically use the small Live Data deployment configuration with a Unified CCE Rogger deployment. You typically use the large Live Data deployment configuration with a separate Router and Logger.

Note The standard Cisco Finesse agent desktop includes the Live Data gadget.

Cisco Virtualized Voice Browser Design Considerations

The number of Virtualized Voice Browsers that your solution requires depends on the VRU ports that your solution needs on the VXML Gateway. Install Cisco VVB depending on the number of SIP sessions required for your solution. This table lists feature support by Cisco VVB:

Platform or Feature	Cisco Virtualized Voice Browser Considerations
Voice Codec	G711
Call Flows	Standalone and Comprehensive with call survivability are supported.
ASR/TTS	Supported
Courtesy Callback	Supported
HTTP	Supported
HTTPS	Supported
Local Prompts	Supported

Platform or Feature	Cisco Virtualized Voice Browser Considerations
Local Hostname Resolution	Supported
MRCP v1 and v2	Supported
VXML 2.0 and 2.1	Supported
RTSP Streaming	Not Supported
Video	Not Supported

Unified Communications Manager Design Considerations

Note The Reference Design layouts use a 7500 user Unified CM OVA which supports 2000 contact center enterprise agents on each redundant pair of subscribers. If you use a different Unified CM OVA, move the cluster off the servers in the Reference Design layout or comply with the specification-based hardware policy. See the *Cisco Collaboration Virtualization* page for your solution at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html for details on the specification-based policy.

Cisco Unified Communications Manager (Unified CM) connects calls passed from Unified CVP to the agent that Unified CCE chose. Unified CVP transfers callers to Unified CCE agent phones or desktops using SIP. The Unified CVP Call Server receives an agent label from Unified CCE and routes the call using SIP proxy. The call is then sent to the appropriate Unified CM subscriber in the cluster, which connects the caller to the agent. The 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.

All contact center enterprise solutions use redundant Unified CM, Unified CCE, and Unified CVP components. Because of the redundancy, your solution can lose half of its core systems and be still operational. In that state, a solution handles calls by rerouting them through Unified CVP to a VRU session or an agent on the still-operational components. Where possible, deploy Unified CCE so that no devices, call processing, or CTI Manager services run on the Unified CM publisher.

To enable automatic failover and recovery, pairs of redundant components interconnect over private network paths. The components use TCP heartbeat messages for failure detection. Unified CM uses a cluster design for failover and recovery. Each cluster contains a Unified CM publisher and multiple UM subscribers. Agent phones and computers register with a primary target but automatically reregister with a backup target if the primary fails.

To set up a Unified CM cluster for your contact center enterprise solution:

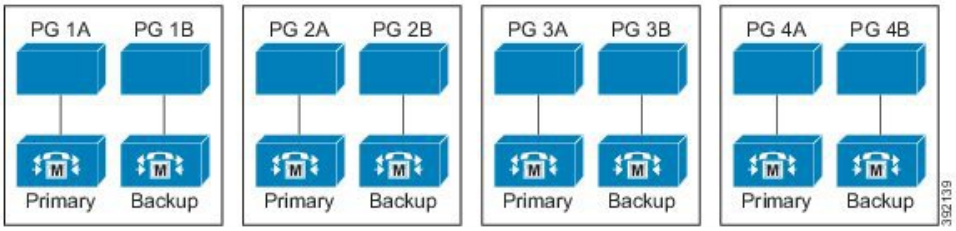
- Configure a SIP trunk in Unified CM.
- When configuring agent labels, consider which device is the routing client. When the label is returned directly to Unified CM, Unified CM is the routing client. When the label is sent to Unified CVP, associate the labels with each of the Unified CVP Switch leg Call Servers.

Unified CM Connection to the Agent PG

You can deploy Agent PGs to connect to a Unified Communications Manager cluster in the following ways:

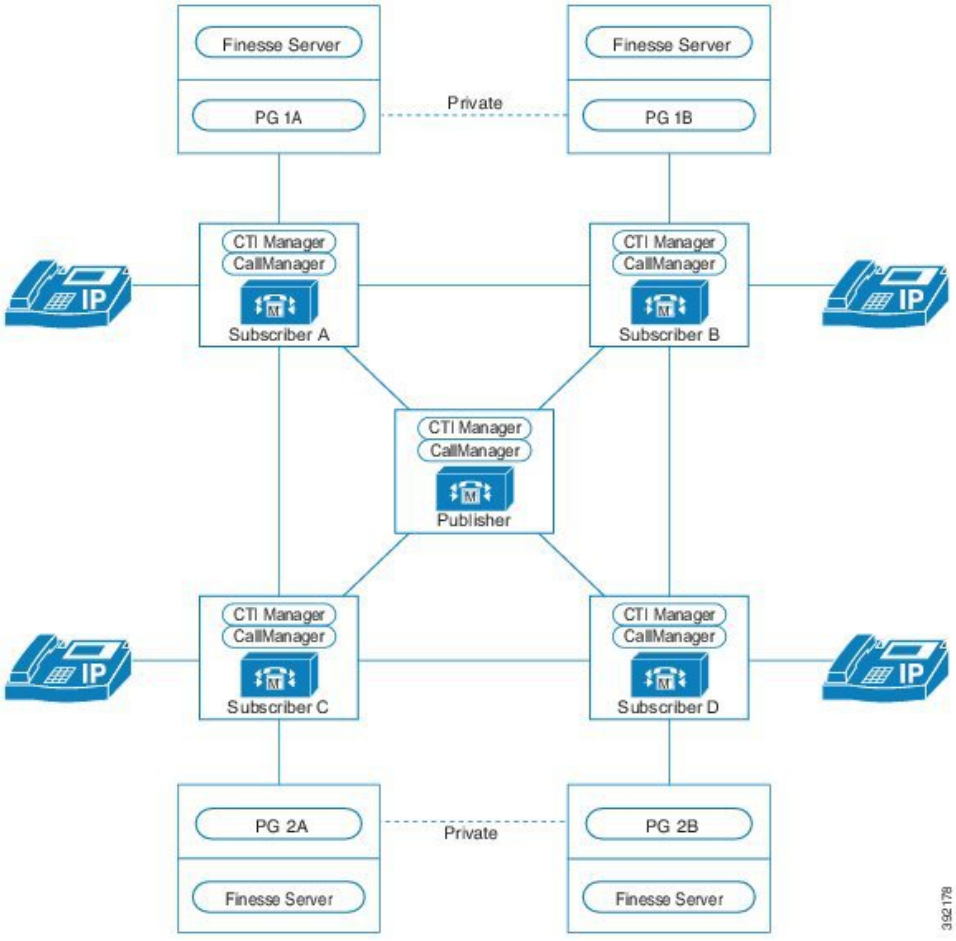
- Deploy an Agent PG for each pair of subscribers. Each subscriber runs the CTI Manager service. Each Agent PG connects to a CTI Manager running on its corresponding subscriber pair. This figure shows an example where four primary subscribers are required and four backup subscribers are deployed to provide 1:1 redundancy.

Figure 61: Deploy Agent PG for Each Pair of Subscribers in a Cluster



This figure shows the connections between the components in a solution with two Agent PG pairs and a cluster with four subscribers.

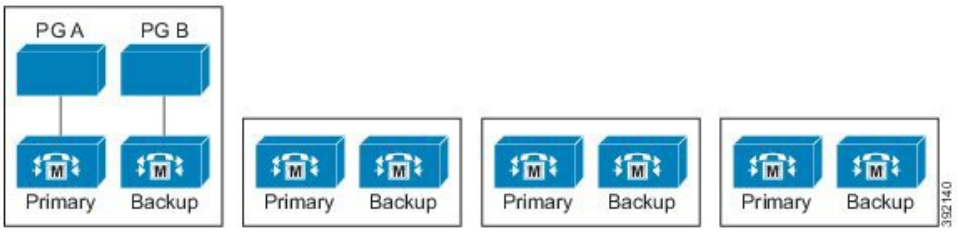
Figure 62: 2 Agent PG Pairs for Unified CM Cluster



- Deploy a single Agent PG for the entire cluster. This type of deployment requires a single pair of subscribers running CTI Manager. Spread agent phone registration among all the subscribers, including

the subscribers running the CTI Manager service. The following diagram shows an example where four primary subscribers are required and four backup subscribers are deployed to provide 1:1 redundancy.

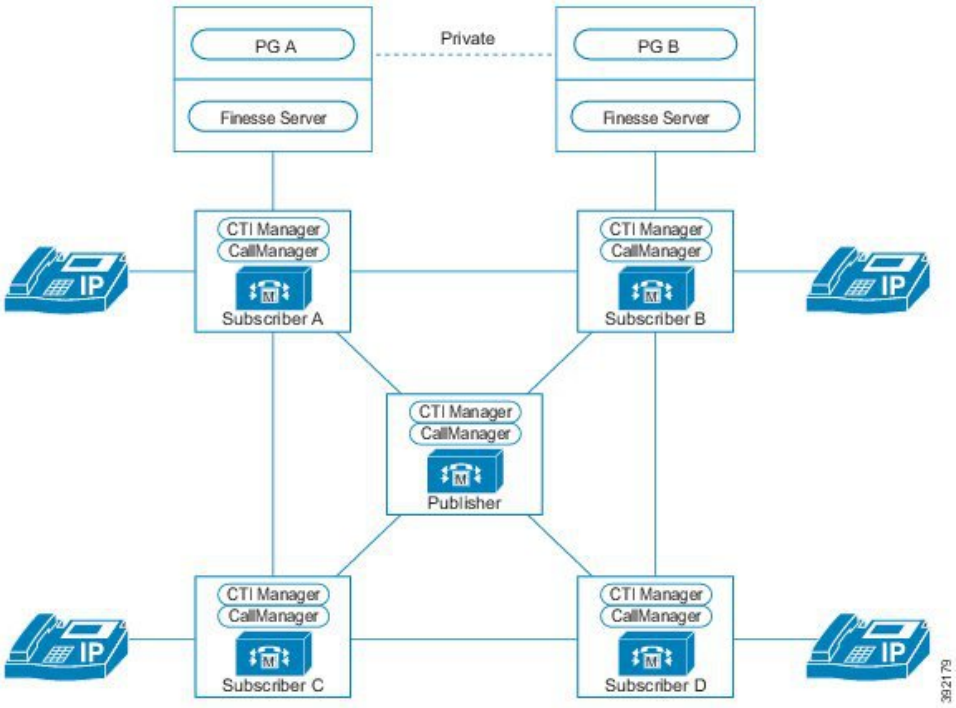
Figure 63: Deploy Single Agent PG for Entire Cluster



Note Use this option when your cluster supports both back-office phones and agent phones.

This figure shows the connections between the components in a solution with two Agent PG pairs and a cluster with four subscribers.

Figure 64: Single Agent PG Pair for Entire Unified CM Cluster



This model reduces the server count for the PG. Another benefit is that there is a single PIM for the entire cluster. So, you can create teams that span across many subscribers. This allows supervisors, for example, to monitor agent phones registered to any subscriber in the cluster. However, this deployment can have slightly higher resource usage on the cluster. Use the *Cisco Unified Communications Manager Capacity Tool* to size the Unified CM servers for your solution.

Single-line and Multi-line Feature Support

Single-line and Multi-line Support

Unified CCE supports Unified CM-based monitoring of a single line and multiple agent lines. Multi-line supports the following capabilities:

- Monitoring and reporting of calls on all lines.
- Other than call initiation, all other call controls on the non-ACD extensions are supported from multi-line capable desktops. Calls initiated from the desk phone can be controlled after initial call setup.
- Requires a maximum of two call appearances.
- Supports a maximum of four lines per phone, one ACD line and up to three non-ACD lines.
- Supports shared lines on ACD and non-ACD lines. Please see table below for differences.
- You can configure multiple devices with a shared non-ACD line, but can only sign in an Agent into one device. Unified CCE supports Shared lines to enable agents with voice facilities at both home and work to share a voicemail line.

Note A shared non-ACD Line does not support **Non ACD Line Impact** configuration in PG Explorer.

- Unified CCE may not be backward compatible with third-party CTI applications when Multi-Line Agent Mode is enabled. Validate Multi-line support with the third-party vendor.

Single-line Versus Multi-line Behavior

Action	Single-line behavior	Multi-line behavior
Accept a routed call while call is on second line?	Yes	Yes, when Non ACD Line Impact is set to ensure there is no impact for the deployment.
Supervisor Monitor using Unified CM-based silent monitor	Yes	Yes. **Note** Non ACD lines do not support Unified CM-based silent monitoring.
Call park	Supported on unmonitored second line	Not supported because all lines are monitored.
Call Waiting / Busy trigger > 1	Supported	No longer supported. Hard-coded to **1** on 69xx series phones (must be configured before enabling multi-line).

Action	Single-line behavior	Multi-line behavior
Reporting on second line calls	Use CDRs in Unified CM	Termination Call Detail Records for call to or from an agent's Non ACD line with an unmonitored device or another agent's Non ACD line is reported with a Non ACD Peripheral Call Type. Reporting for all calls on the Non-ACD line is captured in the Agent Interval table for that agent.
Number of configured lines on phone	No limit described (only monitoring one line)	Do not configure more than four lines. If you do, the agent cannot sign in on any of the lines. This generates a configuration alert.
Shared ACD Line	Shared lines are supported on the ACD line for up to two devices. Device control is selected via the device itself, by going off-hook or making or receiving a call.	Shared lines are supported on the ACD line for up to two devices. Device control is selected via the device itself, by going off-hook or making or receiving a call.
Non-ACD shared line You can configure multiple devices with a shared non-ACD line, but can only sign in an Agent into one device. Unified CCE supports Shared lines to enable agents with voice facilities at both home and work to share a voicemail line. **Note** A shared non-ACD Line does not support **Non ACD Line Impact** configuration in PG Explorer.	Supported on unmonitored line; no configuration limitations	Support for non-ACD lines is limited to having one agent login to a device that has a common second line. The agent cannot sign in on both phones at the same time. Another agent cannot sign into another device which has the same common line.

MTP Usage on the Unified CM Trunk

If your solution uses the Unified CM SIP Trunk, certain call flows, such as Cisco Unity Voice Mail or Mobile Agent, might require an MTP resource.

This is necessary when the negotiated media capabilities of the endpoints do not match, such as with the DTMF in-band versus out-of-band capability. In this case, Unified CM dynamically allocates an MTP due to the DTMF media capabilities mismatch.

Your solution might also require MTPs when interoperating with third–party devices.

Mobile and Remote Access

The Cisco Collaboration Edge architecture includes Unified Communications Mobile and Remote Access (MRA) to enable access by devices that are not in the enterprise network. MRA uses Expressway to provide secure firewall traversal and support for Unified CM registrations. Unified CM can then provide supported devices with call control, provisioning, messaging, and presence services.

For details on Collaboration Edge, see the documentation at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-system/tsd-products-support-series-home.html. For details on Expressway deployment and configuration, see the documentation at http://www.cisco.com/c/en/us/support/unified-communications/expressway-series/tsd-products-support-series-home.html. See the *Compatability Matrix* for your solution for details of device support for MRA.

If your solution uses MRA, consider these points:

- The connection between the Cisco Finesse client and server is over a VPN, not over the MRA connection.
- Certain phones do not support Extension Mobility over MRA.
- Contact center enterprise video deployments do not support MRA.
- If you have VPN split-tunneling configured, you can use Jabber with MRA and the Finesse desktop on the same client machine. See https://www.cisco.com/c/en/us/support/security/anyconnect-secure-mobility-client/products-installation-and-configuration-guides-list.html for Cisco AnyConnect Mobility Client Split-Tunneling configuration.
- If VPN split-tunneling is not available, you can run after splitting them onto two clients.
    - A remote agent who runs Jabber with MRA on one client machine and the Finesse desktop with a VPN connection on a second client machine.
    - A remote agent who runs a Jabber softphone on a laptop that is connected over MRA and runs the Finesse desktop as a Xenapp thin client.

Cisco Finesse Design Considerations

Cisco Finesse is a supervisor and agent desktop for use with contact center enterprise solutions. You install the Cisco Finesse server on a VM. Clients then use a web browser to point to the Cisco Finesse server. No Cisco Finesse software is installed on the client, which speeds and simplifies installation and upgrade.

See the *Contact Center Enterprise Compatibility Matrix* for supported browsers and operating systems for Cisco Finesse clients (administration console, agent desktop, and supervisor desktop).

The Cisco Finesse desktop application consists of the client and server components. The client uses standard web programming elements (HTML, JavaScript) that are distributed as gadgets using the OpenSocial 1.0 specification. You can configure the agent desktop to use Cisco and third-party gadgets through a layout management mechanism.

Cisco Finesse is part of a class of applications called Enterprise Mashups. An Enterprise Mashup is a web-centric method of combining applications on the client side. The gadget-based architecture of Cisco Finesse enables client-side mashup and easier integration. There are fewer version compatibility dependencies because gadget upgrades are handled independently.

You can customize the agent and supervisor desktops through the Cisco Finesse administration console. Administrators can define the tab names that appear on the desktops and configure which gadgets appear on each tab.

This table summarizes the capabilities of Cisco Finesse:

Table 44: Desktop Features

Desktop Functionality	Cisco Finesse
Desktop Chat	Yes
Team Message	Yes
Browser-based desktops	Yes
Custom development	Yes (using standard web components such as HTML, JavaScript)
Desktop security	Yes **Note** Cisco Finesse supports HTTPS for up to 2000 agents on each PG pair.
Workflow automation	Yes
Mobile (remote) agents	Yes
Silent monitoring	Yes
Monitor mode applications	NA
Outbound calls	Yes
Microsoft Terminal Services support	NA
Citrix presentation server support	NA
Agent mobility	Yes
Agent Greeting	Yes

Cisco Finesse REST API

Cisco Finesse provides a REST API for client applications to access the supported server features. The REST API transports XML payloads over HTTPS.

Cisco Finesse also provides a JavaScript library and sample gadget code to aid third-party integration. You can find the developer documentation for the REST API, the JavaScript library, and sample gadgets on the Cisco Developer Network at https://developer.cisco.com/site/finesse/.

Cisco Finesse Agent Desktop

Out of the box, the agent desktop provides the following features:

- Basic call control (answer, hold, retrieve, end, and make a call)
- Advanced call control (consultation, transfer after consult, conference after consult)

- Single-step transfer (agents can transfer a call without first initiating a consultation call)
- Queue statistics gadget (to view information about the queues to which the agent is assigned)
- View of the agent's Call History and State History
- Not Ready and Sign Out reason codes
- Contact lists
- Workflows
- Mobile agent support
- Progressive, Predictive, Preview Outbound, and Direct Preview Outbound
- Desktop Chat
- Team Message

Cisco Finesse Supervisor Desktop

The Cisco Finesse supervisor features extend the agent desktop with more supervisor-only gadgets. These features include the following:

- Team performance gadget to view the agent status
- Queue statistics gadget to view queue (skill group) statistics for the supervisor's queues
- View of the supervisor's Call History and State History
- View the Call History and State History of any agent in the supervisor's team
- Agent Call Information from Team Performance Gadget (TPG)
- Unified CM Silent Monitoring
- Barge-in
- Intercept
- Change agent state (A supervisor can sign out an agent, force an agent into Not Ready state, or force an agent into Ready state.)

Cisco Finesse IP Phone Agent

With Cisco Finesse IP Phone Agent (IPPA), agents can access Cisco Finesse capabilities on their Cisco IP Phone instead of through the browser. Cisco Finesse IPPA only provides a subset of Cisco Finesse features that are available on the browser. It enables agents and supervisors to receive and manage Cisco Finesse calls without access to a PC.

Note Supervisors can only perform agent tasks on their IP Phones. Cisco Finesse IPPA does not support supervisor tasks, such as monitor, barge, and intercept.

Cisco Finesse IPPA supports the following functionality:

- Sign in and out
- Call variables display
- Pending state
- Wrap-up reasons
- Optional wrap-up
- Not Ready reasons
- State change using reason codes
- One Button Sign In

Cisco Finesse Administration Console

Cisco Finesse includes an administrative application that allows administrators to configure the following:

- Connections to the CTI server and the Administration & Data Server database
- Cluster settings for VOS replication
- Not ready and sign out reason codes
- Wrap-up reasons
- Contact lists
- Workflows and workflow actions
- Call variable and ECC variable layouts
- Desktop layout
- Desktop Chat server settings
- Team resources
- Cisco Finesse IP Phone Agent (IPPA)

Reason codes, wrap-up reasons, contact lists, workflows, and desktop layouts can be global (apply to all agents) or assigned to specific teams.

Cisco Finesse Deployment Considerations

Cisco Finesse and the Multiline Feature

Cisco Finesse supports the configuration of multiple lines on agent phones if Unified CCE is configured for multiline. You can configure several secondary lines on an agent phone. However, the Cisco Finesse server blocks any events that the CTI server sends in response to operations on secondary lines. The Cisco Finesse server does not publish these events to the Cisco Finesse clients. Information about calls on secondary lines does not appear on the Cisco Finesse desktop.

If your agents use 8900 Series or 9900 Series phones, enable Multi-Line on the Unified CM peripheral. This configuration option is a peripheral-wide option. If you enable Multi-Line for even one agent with an 8900 Series or 9900 Series phone, enable it for all agents.

To support multiline, configure all phones with the following settings:

- Set Maximum number of calls to 2.
- Set Busy trigger to 1.

Cisco Finesse with Citrix

Contact center enterprise solutions support running the Cisco Finesse desktop within a Citrix environment. Cisco Finesse supports Citrix XenApp and XenDesktop.

Note AWs, Configuration-Only Administration Servers, and Administration Clients can operate only as a single remote instance on a given VM.

For more information about supported versions, see the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html.

Cisco Finesse with NAT and Firewalls

Some solutions have two or more disjointed networks that interconnect with Network Address Translation (NAT).

Cisco Finesse provides limited support for NAT. Cisco Finesse supports basic NAT (one-to-one IP address mapping) between the Cisco Finesse servers and clients.

The following caveats apply to Cisco Finesse and NAT:

- You cannot use PAT/NPAT (one-to-many address mapping that uses ports) between the Cisco Finesse servers and clients.
- You cannot use NAT between the Cisco Finesse servers and any of the servers to which they connect (such as Unified CCE or Unified CM servers).
- Cisco Finesse IP Phone Agent (IPPA) does not support NAT.

Note For more information about NAT and firewalls, see the chapter on solution security.

IP Phone and IP Communicator Support

Cisco Finesse supports the use of Cisco IP hardware phones and the Cisco IP Communicator software phone.

For more information about the supported phone models and IP Communicator versions, see the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html.

IP Phones and Silent Monitoring

Silent monitoring supports both IP hardware phones and Cisco IP Communicator.

IP Phones and Mobile Agent

The Mobile Agent feature does not require any specific type of phone. You can even use analog phones with this feature.

IP Phones and Citrix or MTS

Cisco Finesse supports both IP hardware phones and Cisco IP Communicator when using Citrix or MTS.

In these environments, install Cisco IP Communicator on the agent desktop PC. You cannot deploy Cisco IP Communicator on the Citrix or MTS server.

Cisco Finesse and Cisco Jabber

Cisco Finesse supports Cisco Jabber for Windows as a contact center enterprise voice endpoint. Cisco Finesse supports the following Jabber functionality:

- Voice and Video
- Built-In Bridge (BIB) for silent monitoring
- IM and Presence

Note Agents cannot use Jabber to transfer or conference calls. Agents must use the Cisco Finesse desktop for transfer and conference.

To use Jabber with Cisco Finesse, change the default Jabber configuration as follows:

- Change Maximum number of calls from 6 to 2.
- Change Busy trigger from 2 to 1.

For more information on support for Jabber, see the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html.

Desktop Chat

Desktop Chat is a XMPP browser based chat, which is powered by Cisco Instant Messaging and Presence (IM&P) service. Desktop Chat allows agents, supervisors, and Subject Matter Experts (SMEs) within the organization to chat with each other.

For more details see, https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

Instant Messaging and Presence (IM&P) provides presence and chat capabilities within the Unified CM platform. The Desktop Chat interface is hosted by the Finesse Agent desktop and requires a separate log in to the IM&P service.

Note In HCS deployments, the Desktop Chat user search functionality shows all the users available from LDAP and is not restricted to the users from the corresponding customer whose agent or supervisor initiated the search.

Desktop Chat does not support Cisco Mobile Remote Agent /VPN based access to the IM&P server. Desktop Chat requires direct access to the IM&P server to connect to the chat service.

For more information, see *Desktop Chat Server Settings* section in *Cisco Finesse Administration Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-maintenance-guides-list.html.

Cisco Instant Messaging and Presence (IM&P)

IM&P incorporates the Jabber platform and supports XMPP protocol and can track the user's presence via multiple devices. IM&P pulls its user list from users who have been enabled for chat capabilities, from Unified CM (or LDAP if LDAP integration is enabled). Only Unified CM users enabled for chat capability can login to IM&P.

Cisco IM&P supports multiple forms of clustered deployment to provide high availability. Finesse is configured with a specific Agent PG. This Agent PG is connected to a Unified CM that associated with a specific IM&P cluster. Configure Finesse to connect to this IM&P cluster.

Identity, Presence, Jabber

A User is identified in the IM&P service with a unique identity which is in the form of username@FQDN.com.

A user is described in terms of the identity of the user, presence status, (available, unavailable, or busy) and the presence capabilities of the user.

The presence status of the user is not related to the Agent Status and has to be managed independently by the user post login.

Cisco IM&P service combines the presence status of user across multiple devices and publishes them for subscribers who have added the contact in their contact list.

IM&P supports a composed presence for the users, which is derived from the state matrix of all the devices that the agent is logged into. Cisco IM&P takes sources of presence from the XMPP client for the user, on-hook and off-hook status from CUCM, and in a meeting status from Microsoft Exchange to generate the users overall composed presence. Desktop Chat displays the composed presence of the user. For details about how to arrive at the composed presence, refer to the *Cisco IM&P User Guide* at: https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-user-guide-list.html .

Irrespective of the deployment type, the Desktop Chat requires an explicit login using the IM&P identity of the user after logging into the Finesse Desktop.

SSO is not supported with Desktop Chat and thus an explicit login is required in SSO mode.

Desktop Chat presence indicates the availability of users to communicate across the configured devices.

Desktop Chat availability will also be reflected in the combined IM&P presence of the user.

Logging into Desktop Chat, by default sets the users state as available.

An agent logging into Desktop Chat can thus be seen as available in Jabber or other XMPP platforms connected with IM&P and can communicate with these users.

Note File transfer is supported only for users communicating using Desktop Chat. For more information on the supported file types and the maximum size of file attachments see, *Desktop Properties CLIs* section in the Cisco Unified Contact Center Express Administration and Operations Guide.

Example for Desktop Chat availability:

A Desktop Chat user can be logged into the Desktop Chat and Jabber at the same time. Incoming chats will be relayed to all the logged in clients including Desktop Chat. However, Desktop Chat does not support Multi-Device-Messaging. So messages being sent from other XMPP clients like Jabber will not be displayed within the Desktop Chat. Once alternate clients are used to respond to incoming chats, further messages are not shown in Desktop Chat until the user starts responding using the Desktop Chat.

For more information on network designs, refer to the *Solution Reference Network Design* guide https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

Cisco IM&P Design Considerations

Finesse browser makes a separate connection to Cisco IM&P over HTTPS, after it retrieves the chat server URI from the Finesse server. This requires separate certificates to be accepted if self-signed certificates are employed, in an HTTPS deployment.

The chat interaction happens over XMPP protocol, on the HTTP connection with long polling or BOSH established with Cisco IM&P.

There are no other interactions between Finesse server and browser for chat related capabilities, except for retrieving the Cisco IM&P server configurations.

Chat log persistence is available with the browser during the desktop session.

User search capabilities require Unified CM LDAP integration. In its absence, remote contacts have to be manually added by the user.

If the user is an existing Jabber user, the same contacts are shared between the Desktop Chat and Jabber which are also persisted across sessions.

There are no limits on the number of ongoing chats or the contacts in Desktop Chat apart from the restrictions or guidelines advised by Cisco IM&P. For the limit on the number of ongoing chats or the contacts and how to configure the Cisco IM&P server for chat, see the IM&P Solution Reference Networking Guide.

Note Desktop Chat requires the Cisco IM and Presence certificates to be trusted. For more information on accepting certificates, see the *Accept Security Certificates* section, in the *Common Tasks* chapter of *Cisco Finesse Agent and Supervisor Desktop User Guide for Cisco Unified Contact Center Express* at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-user-guide-list.html.

Cisco IM&P Deployment Considerations

Finesse is configured to the primary and secondary IM&P chat servers through the Cisco Finesse Administration interface.

Desktop Chat automatically discovers the appropriate IM&P node, configured for the user, by connecting to the configured servers and connects to the appropriate nodes in IM&P. This resolution is only performed for the first time chat is loaded and subsequently uses the same nodes, until the browser cache is cleared by the user.

Note Desktop Chat does not use DNS_SRV* records unlike Jabber and cannot automatically configure itself based on the network configurations. The explicit chat URI configuration from Administrative pages is required for chat server discovery.

For details on Cisco IM&P deployment, see Unified CM Solution Reference Network Design guide at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab12/collab12/presence.html.

See Configuration and Administration of the IM and Presence Service on Cisco Unified Communications Manager guide for details about the following:

- How to install and configure IM&P services.
- How to configure IM&P to enable chat services for end users.
- How to configure clusters and high availability deployment.
- How to configure IM&P Federation.

Desktop Chat Server Settings

Desktop Chat is an XMPP browser based chat, which is powered by Cisco Instant Messaging and Presence (IM&P) service. It provides presence and chat capabilities within the Unified CM platform. For more details, see *Configuration and Administration of the IM and Presence Service* at https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-and-configuration-guides-list.html.

Desktop Chat connects to Cisco IM&P servers over port 5280 from the browser hosting the agent desktop. IM&P server visibility and port accessibility needs to be ensured if clients intend to use this feature. The Desktop Chat gadget configures the IM&P host BOSH URL's used by the desktop to communicate with the IM&P server over BOSH HTTP.

IM&P has a clustered design, where users are distributed across multiple nodes in the cluster. The Desktop Chat initially discovers the IM&P nodes that a user has configured, caches this information and communicates with the actual server for subsequent login, until the browser cache is cleared. To spread the initial discovery load, it is advisable to configure the nodes in a round robin fashion if the deployment has more than one Finesse cluster. For example, if there are 5 IM&P nodes configure Finesse cluster A with node 1 & 2, Finesse cluster B with nodes 3 & 4, and so on.

Node availability should be considered while configuring the IM&P URL. The secondary node will be available for discovery in scenarios where the first node is not reachable. The secondary node will be connected for discovery only if the primary node is unreachable.

For the URL to be configured, refer Cisco Unified Presence Administration service, in *System, Service Parameters*. Choose the required IM&P server, select Cisco XCP Web Connection Manager. The URL binding path is listed against the field *HTTP Binding Path*. The full URL to be configured in Finesse is `https://<hostname>:5280/URL-binding-path`.

Use the Desktop Chat Server Settings to configure chat settings for the Finesse desktop. The following table describes the fields on the Desktop Chat Server Settings gadget.

Field	Explanation
Primary Chat Server	Enter the IM&P primary server URL of Desktop Chat.
Secondary Chat Server	Enter the IM&P secondary server URL of Desktop Chat.

Actions on the Desktop Chat Server gadget:

- **Save:** Saves your configuration changes
- **Revert:** Retrieves the most recently saved server settings

Important For Desktop Chat to work without any issues, ensure the following services are running on IM&P:

- Cisco Presence Engine
- Cisco XCP Text Conference Manager
- Cisco XCP Web Connection Manager
- Cisco XCP Connection Manager
- Cisco XCP Directory Service
- Cisco XCP Authentication Service
- Cisco XCP File Transfer Manager

Note Desktop Chat requires the Cisco IM and Presence certificates to be trusted. To start the Desktop Chat without experiencing an exception, you must add the certificate to the browser trust store, or configure IM and Presence with CA-signed certificate, or push self-signed certificate through group policies in supported browsers. For more information on accepting certificates, see the *Accept Security Certificates* section, in the *Common Tasks* chapter of *Cisco Finesse Agent and Supervisor Desktop User Guide for Cisco Unified Contact Center Express* at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-user-guide-list.html.

For more information on adding certificates to the browser trust store, see Certificate Management.

Note Desktop Chat is not supported with the unrestricted versions of IM&P.

Cisco Unified Intelligence Center Design Considerations

Unified Intelligence Center Deployments

A Unified Intelligence Center deployment consists of the following:

- One or more Unified Intelligence Center reporting (member) nodes in a cluster
- Real-time and historical data sources
- Live Data sources
- Other optional data sources

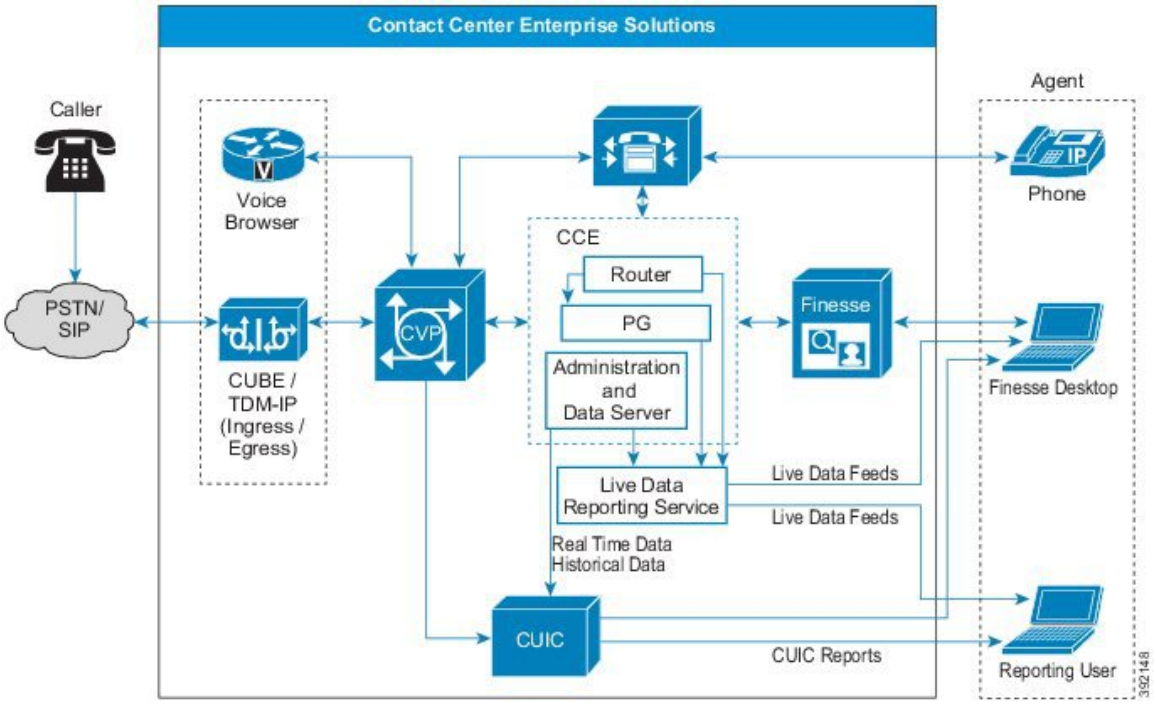
Note Ensure that the Unified Intelligence Center and the data source servers are in synch with the same NTP server.

Unified Intelligence Center nodes are deployed in standalone VMs in contact center enterprise solutions. Unified Intelligence Center supports Historical, Real-time, and Live Data reports.

The data flow for a historical or real-time report runs like this:

1. The web client makes an HTTPS request for a Unified Intelligence Center historical or real-time report.
2. The web server on the Unified Intelligence Center reporting node receives the request.
3. The reporting node pulls the data for the report from the data source server.
4. The reporting node sends the report to the web client through the web server.

Figure 65: Unified Intelligence Center Deployment



The client updates Live Data reports from a Live Data event stream from the Live Data service. For more information on Live Data control and data flows, see the *Serviceability Guide for Cisco Unified ICM/Contact Center Enterprise* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-installation-and-configuration-guides-list.html.

Unified Intelligence Center Reporting Node

The reporting node is the core of Unified Intelligence Center and contains all the features for reporting. You can deploy reporting nodes in the following configurations:

- A standalone reporting node on a controller node
- A controller node with up to seven reporting member nodes in a cluster

A reporting node includes the following applications:

- A firewall
- A web server that runs the Unified Intelligence Center application
- JAVA services and JSP pages that translate the web requests into HTML
- A Unified Intelligence Center Database (Informix) with replication support within the cluster
- The Administration (OAMP) application (on the publisher node)

Unified Intelligence Center Database (Informix)

Each reporting node includes the Unified Intelligence Center application database. The Unified Intelligence Center database is the main data store for the Unified Intelligence Center reporting web application. It holds configuration information relating to users, reports, and user access rights for each node in the cluster.

In a Unified Intelligence Center cluster, each database server connects to all other database servers. Data immediately replicates from any server to all other servers.

An automated daily purge runs at midnight and handless database maintenance activities. You can change the purge schedule as needed. Purge and backup are the only local database maintenance tasks for local Unified Intelligence Center databases.

Unified Intelligence Center Data Sources

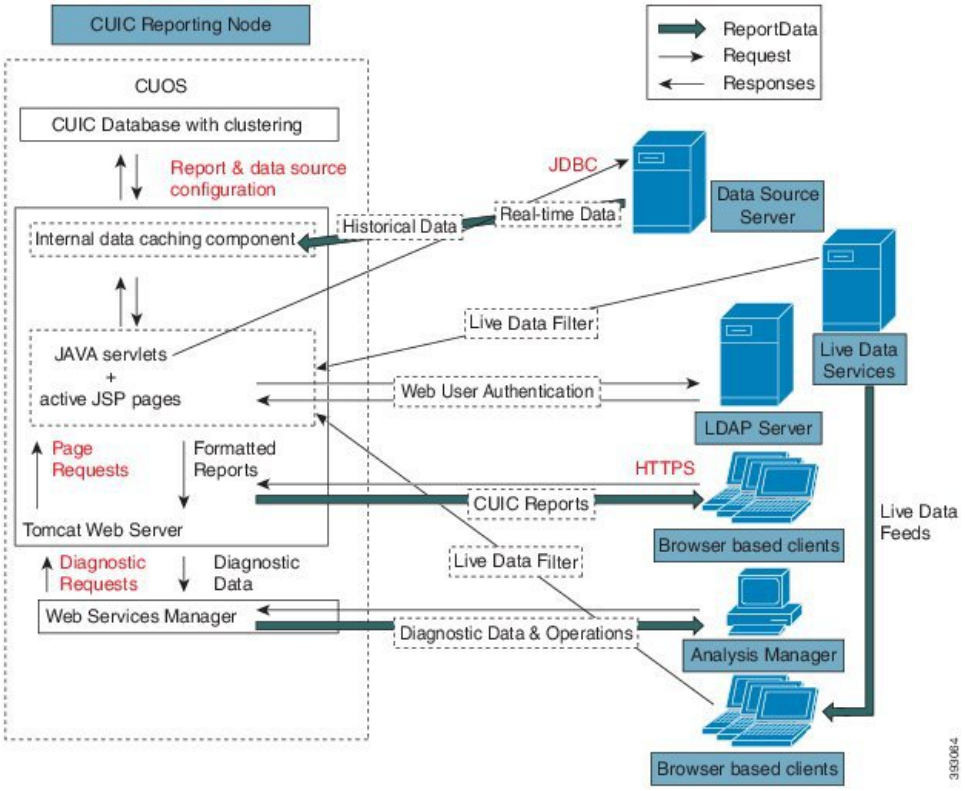
Unified Intelligence Center supports these data sources:

- Microsoft SQL-based or Informix data sources for Real-time and Historical reports. A SQL-based data source is a valid JDBC-compliant database and the schema that store the reporting data.
- Streaming data sources for Live Data reports

These data source servers are supported for these reports:

- Contact center enterprise reports, including those displayed in Cisco Finesse gadgets
- Importable Unified CVP reports
- Customer Collaboration Platform reports
- Enterprise Chat and Email reports

Figure 66: Unified Intelligence Center Architecture



Live Data with Unified Intelligence Center

Unified CCE publishes real-time updates in agent, skill group, and calltype states through WebSockets. Unified Intelligence Center reports consume these messages directly from Live Data Services and display the updates in real time.

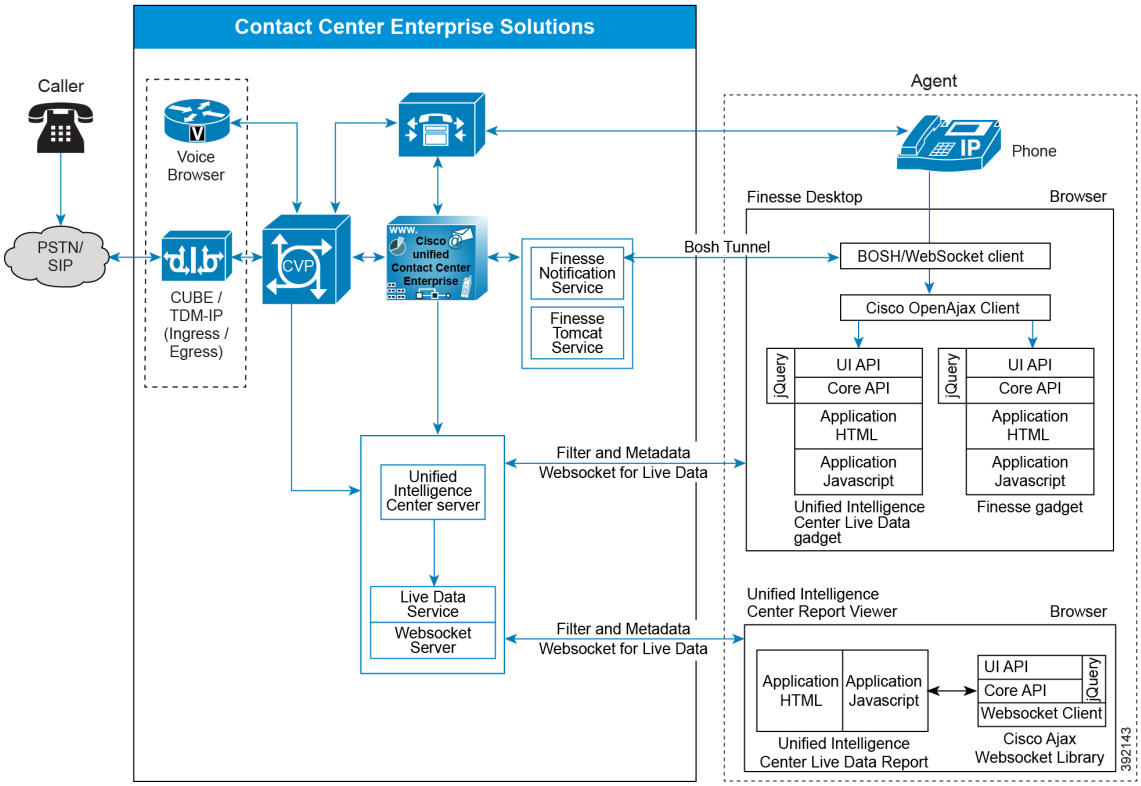
In Packaged CCE, each Unified Intelligence Center node hosts a Live Data Service instance to serve Live Data clients signed in to that node.

Live Data reports can run in the Unified Intelligence Center Report Viewer. Cisco Finesse desktops can show Live Data reports in gadgets. They use a WebSocket tunnel from the Cisco Finesse desktop parent container to one of the Live Data Services. The gadget creates the tunnel during loading.

If a WebSocket connection fails, the Live Data reports automatically fail over to the currently connected Live Data node.

This figure shows the architecture for Live Data reporting in a contact center enterprise deployment:

Figure 67: Live Data Reports in Contact Center Enterprise



Administration & Data Server as the Data Source

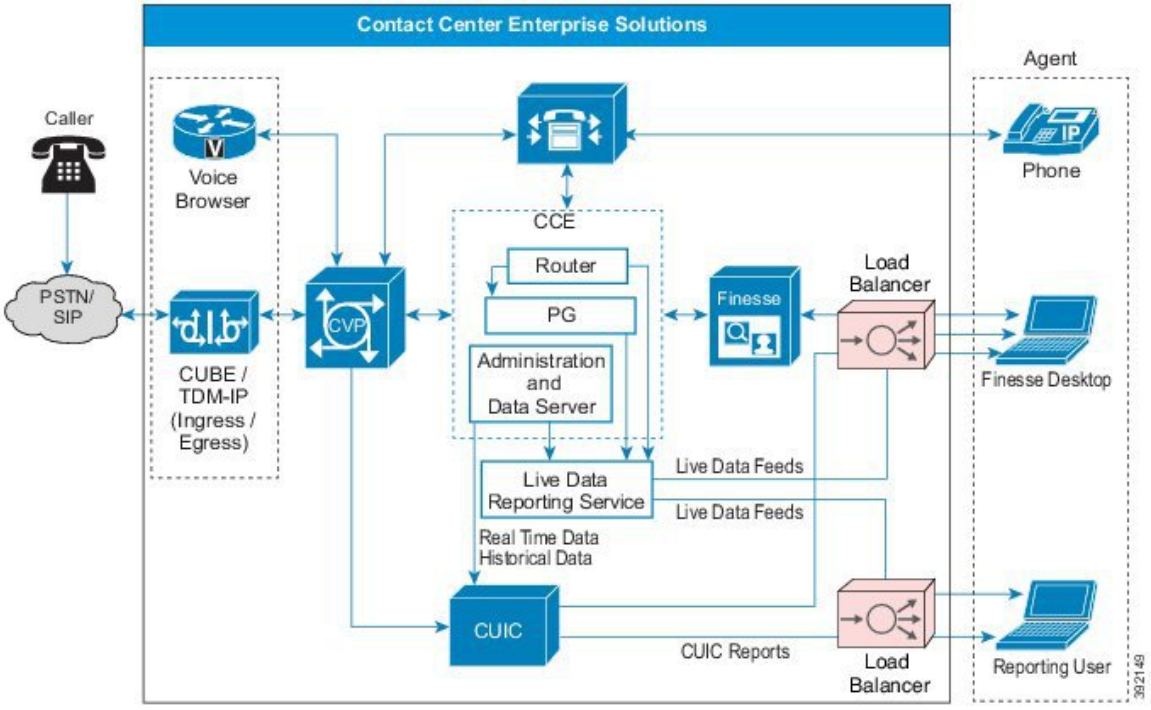
The Unified Intelligence Center stock reports use the Administration & Data Server as their data source. A contact center enterprise deployment can have multiple Administration & Data Servers. Unified Intelligence Center uses the database and its views as the tables for the data source queries.

The Unified Intelligence Center installation adds two data sources to the reporting (member) nodes:

- The historical data source, which supplies data for the historical reports and user integration
- The realtime data source, which supplies data for the realtime reports

Your deployment can use the same AW-HDS for both data sources, or you can configure a different server for each data source. Unified Intelligence Center requires an AW-HDS as a data source for standard historical reports. But, it can also use an AW as a data source for standard real-time reports. Unified Intelligence Center requires an AW-HDS-DDS or HDS-DDS as the data source for custom reports on TCD records.

Figure 68: Unified Intelligence Center Deployment with Unified CCE for Historical and Real-Time Reporting



Unified CVP as the Data Source

For deployments that import reports from Unified CVP, Unified Intelligence Center uses the CVP Reporting Server as the data source.

The CVP Reporting Server houses the Reporting Service and hosts an IBM Informix Dynamic Server (IDS) database management system. The database's schema is available to enable you to write custom reports for the database. Your solution can include several CVP Reporting Servers.

Unified Intelligence Center only connects to the CVP Reporting Server. The CVP Reporting Server mediates between the other Unified CVP subcomponents and Unified Intelligence Center.

Unified Intelligence Center in WAN Deployments

You can distribute the Unified Intelligence Center cluster over the WAN. Each node in a cluster requires a connection to every other node. In a cluster, either over a LAN or WAN, the configuration objects that are created on one node automatically replicate to the other nodes. The replication uses bandwidth across the WAN. But, since you create configuration objects infrequently, this affects the WAN bandwidth less often than running reports.

An object is instantly available users on the local node. It can take a few seconds before the object replicates to other nodes. The WAN bandwidth for replication depends on the configuration of the cluster.

Site Organization

A Unified Intelligence Center cluster can have a maximum of eight nodes. To have fully redundant clustering, each site is at most four nodes. As qualified, each Unified Intelligence Center supports up to 200 reporting

users under the standard reporting profile. Although you can have up to eight nodes in a cluster, you cannot exceed the Reference Design limit for the maximum reporting users.

The primary (controller) node is at the primary site. The primary node hosts the following services:

- Administration application
- Scheduler

WAN Failures

Each Unified Intelligence Center node buffers replication data to send to other nodes in the cluster. During a connection loss, the node queues the data until contact with the other nodes is restored. Each node continues to work independently during a connectivity failure. The queue holds up to 1600 MB. If connectivity is restored before the buffer exhaustion, the node synchronizes at a rate proportional to the amount of queued data and the connection bandwidth.

When the node nears buffer exhaustion, it sends an alarm. If connectivity is not restored before the buffer exhaustion, then replication is reset. By resetting replication, the node can continue to run reports and work independently. A secondary node that resets replication requires full synchronization with the primary node (primary database backup and restore on secondary node) after reconnecting with the primary node. If replication is reset, then all created or modified objects on the secondary node are rolled-back to the state of the primary database.

If the primary node fails, reinstall and revert to a saved backup. Back up the primary node periodically to avoid data loss.

For more information on data replication, see the *Administration Console User Guide for Cisco Unified Intelligence Center*.

Unified Intelligence Center Administration

The Unified Intelligence Center Administration server provides operations, administration, maintenance, and provisioning (OAMP) functions. The Administration server is the primary interface for configuring and provisioning devices in a Unified Intelligence Center cluster. You deploy and access the administration functions on the primary (controller) node in the cluster.

Cisco Unified Intelligence Center uses Hazelcast for application clustering. Hazelcast provides a second-level cache for the Unified Intelligence Center application layer. When any entity (for example: report, report definition, and so on) cached by Hazelcast is updated in one of the Unified Intelligence Center nodes, it must be invalidated and reloaded in all the other Unified Intelligence Center nodes in the cluster. The Hazelcast cluster automatically takes care of it by publishing clusterwide notifications containing the identifiers of such entities which must be invalidated.

In Unified Intelligence Center, the default mechanism for Hazelcast cluster discovery or formation is UDP multicast. Unified Intelligence Center uses the Multicast group IP address 224.2.2.3 and port 54327. You cannot change these settings in Unified Intelligence Center.

The UDP multicast based discovery mechanism will not work for the customer in the following scenarios:

- When the network has multicasting disabled.
- If the nodes in the Unified Intelligence Center cluster are in different subnets.

In such scenarios, you can change the discovery mechanism to TCP/IP. You can form the CUIC application cluster using TCP/IP instead of the default UDP Multicast based discovery mechanism.

For more information on the administration functions or on cluster configuration using Hazelcast, see the *Administration Console User Guide for Cisco Unified Intelligence Center* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-intelligence-center/products-maintenance-guides-list.html.

Note You do not use the Administration application for system-level functions such as network options, certificates, upgrades, and SNMP and Alert settings.

Throttling for Historical and Real-Time Reports

The Unified Intelligence Center throttling mechanism prevents servers from freezing or encountering an Out-of-Memory situation under extreme load.

Note Do not use the throttling mechanism for any sizing purposes. The throttling mechanism only prevents an Out-of-Memory situation. Throttling does not ensure a good quality of service. If your deployment is overused, the level of service can degrade substantially before the throttling mechanism activates.

Always use the sizing calculator to determine the proper reporting resources for your solution.

Processing report data consumes the most memory in Unified Intelligence Center. The throttling mechanism controls memory consumption due to reporting activity.

Unified Intelligence Center measures reporting activity through the *report row*. This measure of reporting activity gives you flexibility. You can run a few large reports or many small ones and the throttling mechanism is equally effective without requiring any tuning.

Unified Intelligence Center measures reporting activity through the *report row*. This measure of reporting activity gives you flexibility. You can run a few large reports or many small ones and the throttling mechanism is equally effective without requiring any tuning.

Tests with the stock reports show that 2 KB is a conservative estimate for the size of a report row. Based on that estimate, a Unified Intelligence Center server can load a maximum of 250,000 report rows into memory before the server runs out of memory.

To enforce this limit, each Unified Intelligence Center keeps count of the report rows currently loaded into memory. All reporting operations check that count to determine if they can load more report rows into memory. When you reach the limit, reporting operations fail and display an error as follows:

- **Violations while fetching data from the data source**—The report cancels and marks the report as failed. Unified Intelligence Center does not take partial results. The system either reads all the data for a request or marks the report as failed and stores none of the data.
- **Violations while preparing an HTTPS response for a browser**—Unified Intelligence Center rejects the display request. An error message says that the server is low on resources and cannot render the report.

Reference Design and Topology Design Considerations

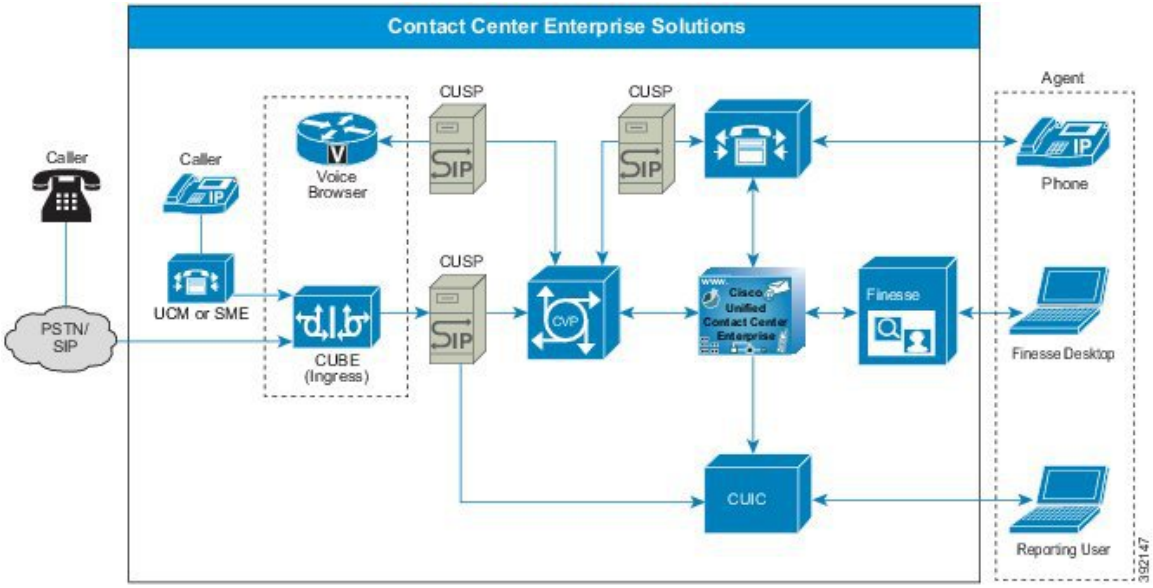
Unified CM SME Deployment

Cisco Unified Communications Manager Session Management Edition (Unified CM SME) integrates with Unified CVP as a dial peer configurator or aggregator to connect to multiple Unified CM clusters in a contact center enterprise solution.

Configure Unified CM SME as a back-to-back SIP user agent. As an aggregator of multiple Unified CM clusters, it routes the call to the appropriate cluster based on the dial plan.

This figure shows the Unified CM SME deployment.

Figure 69: Unified CM SME Deployment



Unified CM SME does not support high availability and is a single point of failure. Consider these design points to minimize the effect of network connectivity or component failures in Unified CM SME.

- Deploy Unified CM SME in redundant clustered mode (at least 1+1 publisher subscriber) at the egress side of Unified CVP.
- Configure **Session Refresh** and **Session Timer** in the Gateway and CUBE. This configuration clears call sessions from the gateway and releases Unified CVP Call Server ports if there is a Unified CM SME failure.
- In a Unified CM SME failure, all call server ports are cleared after the customer drops the call.

Note Call supplementary services do not work for the already established calls once the Unified CM SME fails.

A momentary network connectivity failure to Unified CM SME results in the following limitations:

- Unified CM SME does not clear the call if the agent ends the call during a momentary connectivity failure. This results in a stale cached entry and ports in the Unified CVP application. In such cases, the caller should drop the call to clear the stale cached entry.
- The call does not get cleared from the agent desktop and the agent cannot receive any incoming calls. The agent remains in the talking state and cannot clear the call from the desktop. In such cases, manually clear the call from the phone.
- Because of a delay in call clearance, the call reporting data can reflect inaccurate details for call duration and reason code.

For more information about Unified CM SME configuration, see *Configuration Guide for Cisco Unified Customer Voice Portal* available at: http://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-installation-and-configuration-guides-list.html.

Global Deployments Considerations

- The maximum Round Trip Time (RTT) between the main site and a remote site is restricted to 400 milliseconds.
- The maximum RTT between the Side A components and Side B site is restricted to 80 milliseconds.

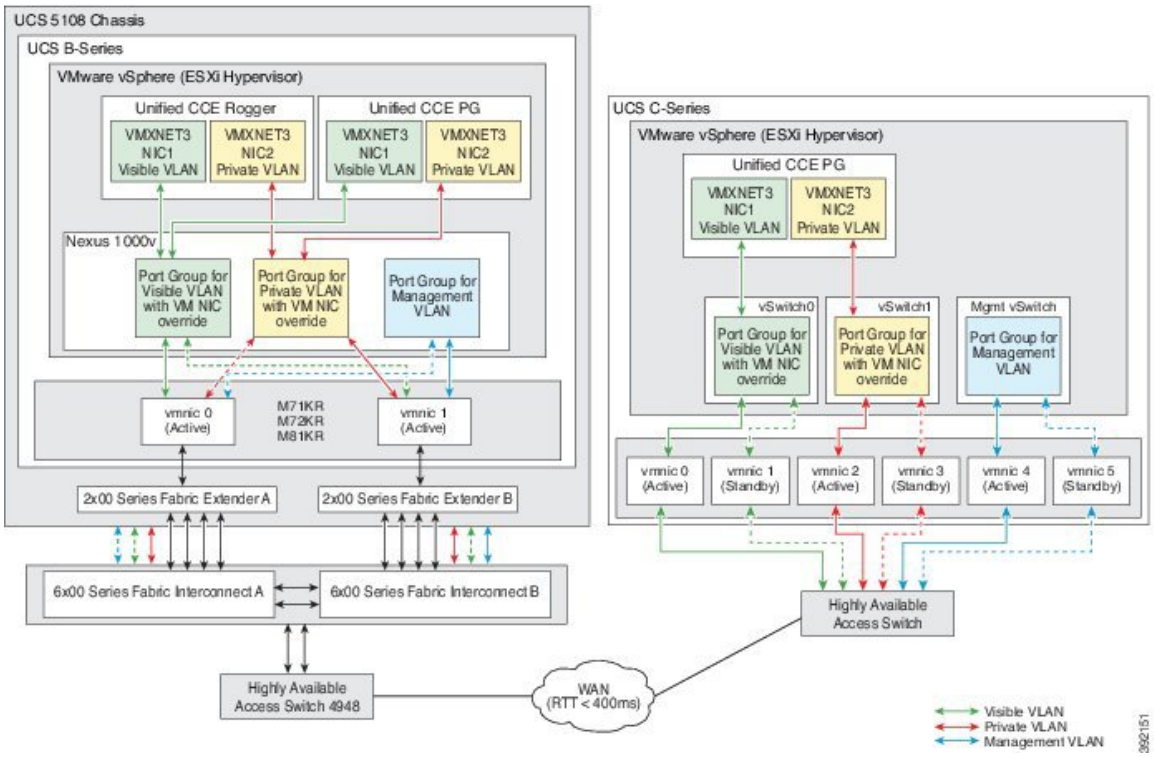
Note If you use Precision Queues, you can have a maximum of 12 Unified CM PIMs combined between the central and remote site.

- Use the hostname for CVP Media Servers and configure the IOS gateways to point to the local CVP servers.

UCS Network Design for Global Deployments

This figure shows the default design for a global deployment to meet Public and Private Network communications requirements. The Main Site uses UCS B Series blades and the **Remote Site** uses a UCS C-series server.

Figure 70: UCS Network Reference Design for Global Deployment



Call Survivability in Distributed Deployments

Distributed deployments require design guidelines for other voice services at the branch. For example, take a branch that is a remote Unified CM site supporting both ACD agent and nonagent phones. In this deployment, the PSTN Gateway handles not only ingress calls for Unified CVP. It also handles ingress or egress calls for the regular non-ACD phone.

Branch reliability in WANs may be an issue in a centralized Unified CVP model because they are typically less reliable than LAN links. The call survivability function must be considered for both the Unified CVP and non-CVP calls. For Unified CM endpoint phones, survivability is accomplished by using a Cisco IOS feature known as Survivable Remote Site Telephony (SRST).

For Unified CVP calls, a combination of services from a TCL script (survivability.tcl) and SRST functions handle call survivability. The survivability TCL script monitors the SIP connection for all calls that ingress through the remote gateway. If a signaling failure occurs, the TCL script takes control of the call and redirects it to a configurable destination. The destination choices for the TCL script are configured as parameters in the Cisco IOS Gateway configuration.

Note When the called number is in "E164" format, the survivability script removes the "+" sign from the called number before forwarding it to Unified CVP. Unified CVP or ICM does not support the "+" sign in the beginning of DNIS.

Alternate destinations for this transfer include another IP destination (including the SRST call agent at the remote site), call restart, call restart with a new destination, *8 TNT, or hookflash. With transfers to the SRST call agent at the remote site, the most common target is an SRST alias or a basic ACD hunt group.

Voice mail and recording servers do not send Real-Time Control Protocol (RTCP) packets in reverse direction toward the caller (TDM Voice Gateway). This can falsely trigger the media inactivity timer of the survivability script. It is important to apply the survivability.tcl script carefully to the dial peers. A call might drop if it goes to the voice mail or to a recording element. One method is to use a separate dial peer for voice mail or recording calls, and not associate the Unified CVP survivability script for those dial peers. Another method is to disable the media inactivity on the survivability script associated with the voice mail or recording dial peers.

For further information on configuration and application of these transfer methods, see the *Configuration Guide for Cisco Unified Customer Voice Portal* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-installation-and-configuration-guides-list.html.

Note To take advantage of alternate routing on signaling failures, use the survivability service on all gateways pointing to Unified CVP. Always use this service, unless you have a specific implementation that prevents using it.

Router requery is not supported when using SIP REFER with Unified CVP Comprehensive Call Flow when the survivability service is handling the REFER message from Unified CVP. Other call flows can support router requery with REFER when Cisco IOS handles the REFER without the survivability service or if Unified CM handles the REFER. For third-party SIP trunks, the support of router requery with REFER depends on their implementation and support for SIP REFER.

Optional Cisco Components Design Considerations

Customer Collaboration Platform Design Considerations

You deploy Customer Collaboration Platform as a large single server. Customer Collaboration Platform does not support redundant topologies for high availability.

You can deploy the server inside or outside the corporate firewall in "Intranet" and "Internet" topologies:

- The Intranet topology provides the additional security of your network firewall to reduce the risk of an external party accessing the system. Use this topology when Customer Collaboration Platform accesses internal sites, such as an internal forum site.

  The Intranet topology complicates proxy configuration, but it simplifies directory integration.

- The Internet topology puts Customer Collaboration Platform outside of your network firewall. It relies on the built-in security capabilities of the Customer Collaboration Platform appliance.

  Whether this topology's security is acceptable or not depends on how you use the system and your corporate policies.

  The Internet topology can complicate directory integration.

You can deploy Customer Collaboration Platform so that some users access the server through a firewall or proxy.

Customer Collaboration Platform Administration UI Access

By default, access to Customer Collaboration Platform administration user interface is restricted. Administrator can provide access by allowing the client's IP addresses and revoke access by removing the client's IP from the allowed list.

To allow client's IP addresses, see the commands provided in the *Task Routing* section of the Cisco Packaged Contact Center Enterprise Features Guide available at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html

Customer Collaboration Platform MR PG Encryption

When you configure Customer Collaboration Platform as an External Machine in the Unified CCE Administration System Inventory, configure the MR PG server, open the CCE Configuration for Multichannel Routing drawer on the Customer Collaboration Platform UI.

To enable the secured transport mode for media routing interface use the **Secured (TLS):** check box:

- If checked, the MR communication will be secured over TLSv1.2 channel.
- If unchecked, the MR communication will be over plain TCP channel.

Task Routing Considerations

Task Routing

Task Routing describes the system's ability to route requests from different media channels to any agents in a contact center.

You can configure agents to handle a combination of voice calls, emails, chats, and so on. For example, you can configure an agent as a member of skill groups or precision queues in three different Media Routing Domains (MRD) if the agent handles voice, e-mail, and chat. You can design routing scripts to send requests to these agents based on business rules, regardless of the media. Agents signed into multiple MRDs may switch media on a task-by-task basis.

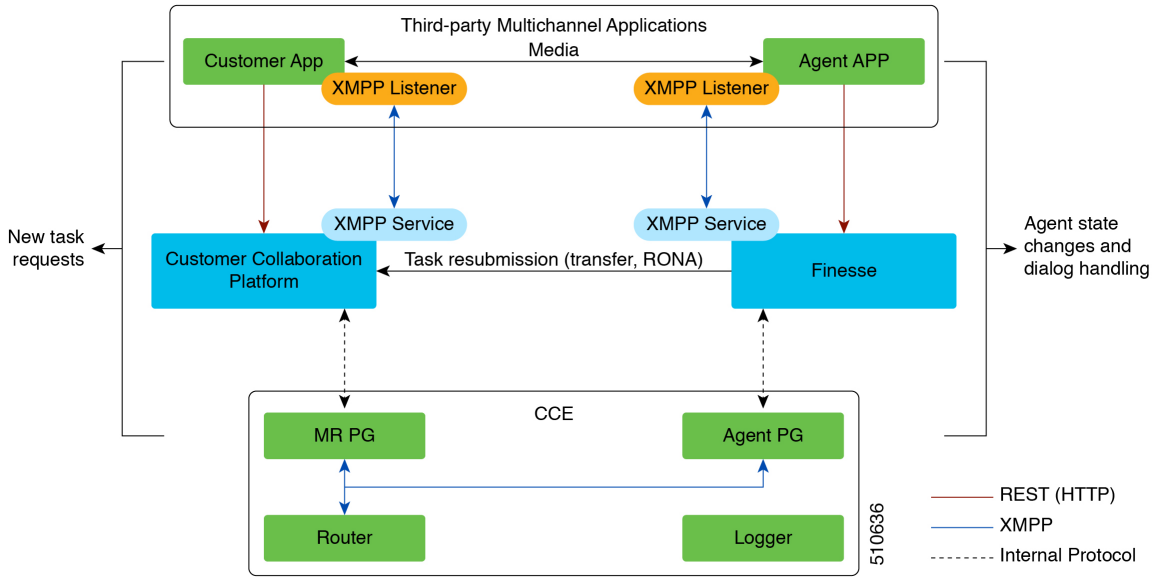
Enterprise Chat and Email provides universal queue out of the box. Third-party multichannel applications can use the universal queue by integrating with CCE through the Task Routing APIs.

Task Routing APIs provide a standard way to request, queue, route, and handle third-party multichannel tasks in CCE.

Contact Center customers or partners can develop applications using Customer Collaboration Platform and Finesse APIs in order to use Task Routing. The Customer Collaboration Platform Task API enables applications to submit nonvoice task requests to CCE. The Finesse APIs enable agents to sign into different types of media and handle the tasks. Agents sign into and manage their state in each media independently.

Cisco partners can use the sample code available on Cisco DevNet as a guide for building these applications (https://developer.cisco.com/site/task-routing/).

Figure 71: Task Routing for Third-party Multichannel Applications Solution Components



Customer Collaboration Platform and Task Routing

Third-party multichannel applications use Customer Collaboration Platform'sTask API to submit nonvoice tasks to CCE.

The API works in conjunction with Customer Collaboration Platform task feeds, campaigns, and notifications to pass task requests to the contact center for routing.

The Task API supports the use of Call variables and ECC variables for task requests. Use these variables to send customer-specific information with the request, including attributes of the media such as the chat room URL or the email handle.

Note CCE solutions support only the Latin 1 character set for Expanded Call Context variables and Call variables when used with Finesse and Customer Collaboration Platform. Arrays are not supported.

CCE and Task Routing

CCE provides the following functionality as part of Task Routing:

- Processes the task request.
- Provides estimated wait time for the task request.
- Notifies Customer Collaboration Platform when an agent has been selected.
- Routes the task request to an agent, using either skill group or precision queue based routing.
- Reports on contact center activity across media.

Finesse and Task Routing

Finesse provides Task Routing functionality via the Media API and Dialog API.

With the Media API, agents using third-party multichannel applications can:

- Sign into different MRDs.
- Change state in different MRDs.

With the Dialog API, agents using third-party multichannel applications can handle tasks from different MRDs.

Task Routing Use Cases

Use the Task Routing APIs to request, queue, route, and handle third-party multichannel tasks in CCE.

We support these use cases:

- Agents handling multiple concurrent tasks in a single Media Routing Domain (MRD), and across MRDs.
- Interruptible MRDs. Agents handling tasks in those MRDs can be interrupted by tasks in other MRDs. For example, you can set an email MRD to be interruptible, meaning that an agent handling an email task can be interrupted by a task from another MRD, such as a voice call or a chat.
- Blind transfer to a specified script selector. Direct transfer is not supported.
- Agent sign out with assigned tasks. Those tasks are either closed or transferred, depending on the agent's dialogLogoutAction setting in the Finesse Media API.
- RONA. If an agent does not accept an offered task within the Start Timeout threshold for the MRD, Finesse resubmits the task for routing and makes that agent not routable.

Task Routing Task Flow

This task flow describes a typical multichannel scenario, in which an agent is configured to handle email and chat tasks.

The email Media Routing Domain (MRD) is interruptible, and the agent is set to accept interrupts in that MRD. Because the email MRD is interruptible, an agent handling an email task can be interrupted with tasks from other MRDs. Because the agent is set to accept interrupts, the state of the agent, email task, and Finesse dialog in the email MRD change to INTERRUPTED if the agent is assigned a task in another MRD. The agent also cannot perform work on the email task while interrupted.

The chat MRD is non-interruptible.

See the *Cisco Packaged Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html for more information about MRD interruptibility and accepting or ignoring interrupts.

The partner has developed email and chat applications that are integrated with the Finesse and Customer Collaboration Platform APIs.

1. A customer sends an email to the company. The email application submits a new email task request to CCE. The task request includes a script selector and Call and ECC variables with customer-specific information, including the handle to the email.
2. CCE maps the script selector to a call type, which determines which routing script to run. The routing script queues the email task to the appropriate skill group or precision queue in the email MRD.

3. The email application polls for status and Estimated Wait Time (EWT).
4. An agent signs in to the email MRD and changes to Ready state.
5. CCE assigns the email to the agent. The Call and ECC variables used to create the task are included in the dialog's media properties. The application uses these variables to allow the agent to reply to the email. The agent starts work on the email dialog in Finesse.
6. Another customer sends a chat request to the company. The chat application submits a new chat task request to CCE. The Call and ECC variables include the chat room URL.
7. CCE maps the script selector submitted with the chat request to a call type, which determines which routing script to run. The routing script queues the chat task to the appropriate skill group or precision queue in the chat MRD.
8. The chat application polls for status and Estimated Wait Time (EWT).
9. While working on the email task, the same agent signs in to the chat MRD and changes to Ready state.
10. CCE assigns the chat task to the agent. The application uses the Call and ECC variables to add the agent to the chat room with the customer. The agent starts work on the chat dialog in Finesse.
11. The agent's state in the email MRD changes to Interrupted, and the email dialog state changes to Interrupted. The application disables actions for the email dialog.
12. The agent transfers the chat task to a different script selector. Finesse closes the chat dialog and resubmits the task to Customer Collaboration Platform. The application closes the chat room.
13. The agent is not handling other non-interruptible dialogs, and the email dialog becomes active.
14. The agent continues working on the email dialog. The agent pauses the dialog to take a short break, and then resumes the dialog.
15. When the email reply is complete, the agent performs wrap up work for the dialog. The agent closes the dialog. Finesse sends a handle event to Customer Collaboration Platform for the email task. The application sends the email reply to the customer.

Task Routing Design Impacts

Enterprise Reference Design Task Routing Support

All solutions support task routing, with the exception of the Small Contact Center solution.

Task Routing Deployment Requirements

Task Routing for third-party multichannel applications deployment requirements:

- Finesse and Customer Collaboration Platform are required. Install and configure Finesse and Customer Collaboration Platform before configuring the system for Task Routing.

  See the Finesse documentation and Customer Collaboration Platform documentation.

  By default, access to the Customer Collaboration Platform administration user interface is restricted. Administrator can provide access by unblocking the IP addresses of the clients. For more details, see the *Control Customer Collaboration Platform Application Access* topic in the *Cisco Customer Collaboration Platform Installation and Upgrade Guide* guide.

- You can install only one Customer Collaboration Platform machine in the deployment.

- Customer Collaboration Platform must be geographically colocated with the Unified CCE PG on one side.
- Install Customer Collaboration Platform in a location from which CCE, Finesse, and the third-party multichannel Customer Collaboration Platform Task Routing application can access it over the network.

  If you install Customer Collaboration Platform in the DMZ, open a port for CCE and Finesse to connect to it. The default port for CCE to connect to Customer Collaboration Platform is port 38001. Finesse connects to Customer Collaboration Platform over HTTPS, port 443.

  Install the third-party multichannel application locally with Customer Collaboration Platform, or open a port on the Customer Collaboration Platform server for the application to connect to it.

Task Routing Sizing and Capacity Limits

The sizing and capacity limits for Task Routing for third-party multichannel applications are the following:

- Maximum multimedia agents per system: 2,000 agents
- Maximum tasks per hour: 15,000 tasks
- Maximum concurrent tasks: 20,000 tasks
- Maximum tasks per agent across all an agent's MRDs: 5 tasks
- Maximum tasks per agent in a single MRD: 5 tasks
- Task submission rate: 5 tasks per second

  Customer Collaboration Platform throttles the task submission rate to CCE to 5 tasks per second. Customer Collaboration Platform holds a maximum of 10,000 tasks in the queue for submission. If the queue exceeds 10,000 tasks, then Customer Collaboration Platform discards the additional tasks with the disposition code NOTIFICATION_RATE_LIMITED. Once the queue is ready again, additional tasks are added to the queue.

Task Routing Bandwidth, Latency, and QoS Considerations

For Customer Collaboration Platform, the network must have sufficient bandwidth to reliably support HTTP. Task Routing Rest API requests carry only metadata; they do not carry media. If the customer application connects to Customer Collaboration Platform via XMPP to receive task status notifications, the network must reliably support a persistent TCP connection with the Customer Collaboration Platform XMPP server. Connecting toCustomer Collaboration Platform via XMPP does not significantly impact network bandwidth.

To calculate the required bandwidth for Task Routing tasks for the Finesse desktop, use the *Finesse Bandwidth Calculator*, available at: https://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-technical-reference-list.html.

For CCE, bandwidth, latency, and QoS considerations are the same for Task Routing tasks as they are for voice calls. A Task Routing task uses the same bandwidth as a voice call.

Unified SIP Proxy Design Considerations

Consider these points when adding Cisco Unified SIP Proxy (CUSP) into your solution:

- CUSP is a VM that can reside on UCS B, C, and E series servers.

- A standard CUSP topology consists of 2 redundant, geographically separated gateways. The gateways have one proxy module each. They use SRV priority for redundancy of proxies. They do not use HSRP.
- CUSP can coreside with VXML or TDM gateways.
- Configure TDM gateways with SRV or with Dial Peer Preferences to use the primary and secondary CUSP proxies.
- Set up CUSP with Server Groups to find the primary and back up Unified CVP, Unified CM, and VXML gateways.
- Set up Unified CVP with a Server Group to use the primary and secondary CUSP proxies.
- Set up Unified CM with a Route Group with multiple SIP Trunks, to use the primary and secondary CUSP proxies

Performance Matrix for CUSP Deployment

CUSP baseline tests in isolation on the proxy give a maximum capacity of 450 TCP transactions per second. Use this as the highest benchmark and most stressed condition allowable. From the proxy server perspective, a CVP call entails an average of four separate SIP calls:

- Caller inbound leg
- VXML outbound leg
- Ringtone outbound leg
- Agent outbound leg

When a consult with CVP queuing occurs, the session incurs an extra four SIP transactions, effectively doubling the number of calls.

`Record Route` is turned off by default on CUSP.

Note Always turn the `Record Route` setting off on the proxy server. This avoids a single point of failure, allows fault tolerant routing, and increases the performance of the Proxy server. Using that setting on the proxy server doubles the impact to performance. It also breaks the high availability model. The proxy becomes a single point of failure for the call, if the proxy goes down.

Call Disposition with CUSP

The following sections discuss configuration of Cisco IOS Gateways using SIP. These examples highlight certain configuration concepts.

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 Reporting 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 the dial peer configuration while also providing DNS SRV load balancing and redundancy. The disadvantage 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
```

SIP Proxy Dial-Plan Configuration

If you have a SIP Proxy, use different VRU labels for the Unified CM routing client and the CVP routing clients. The Unified CM routing client uses its VRU label to send a call to the CVP Call Server to hand off call control first. The CVP routing client uses its VRU label to send a call to the VXML Gateway for treatment. When a call comes to CVP, CVP transfers to the CVP routing client's VRU label. It then delivers the call to the VXML Gateway for queuing treatment.

Structure the dial plan in your SIP Proxy as follows:

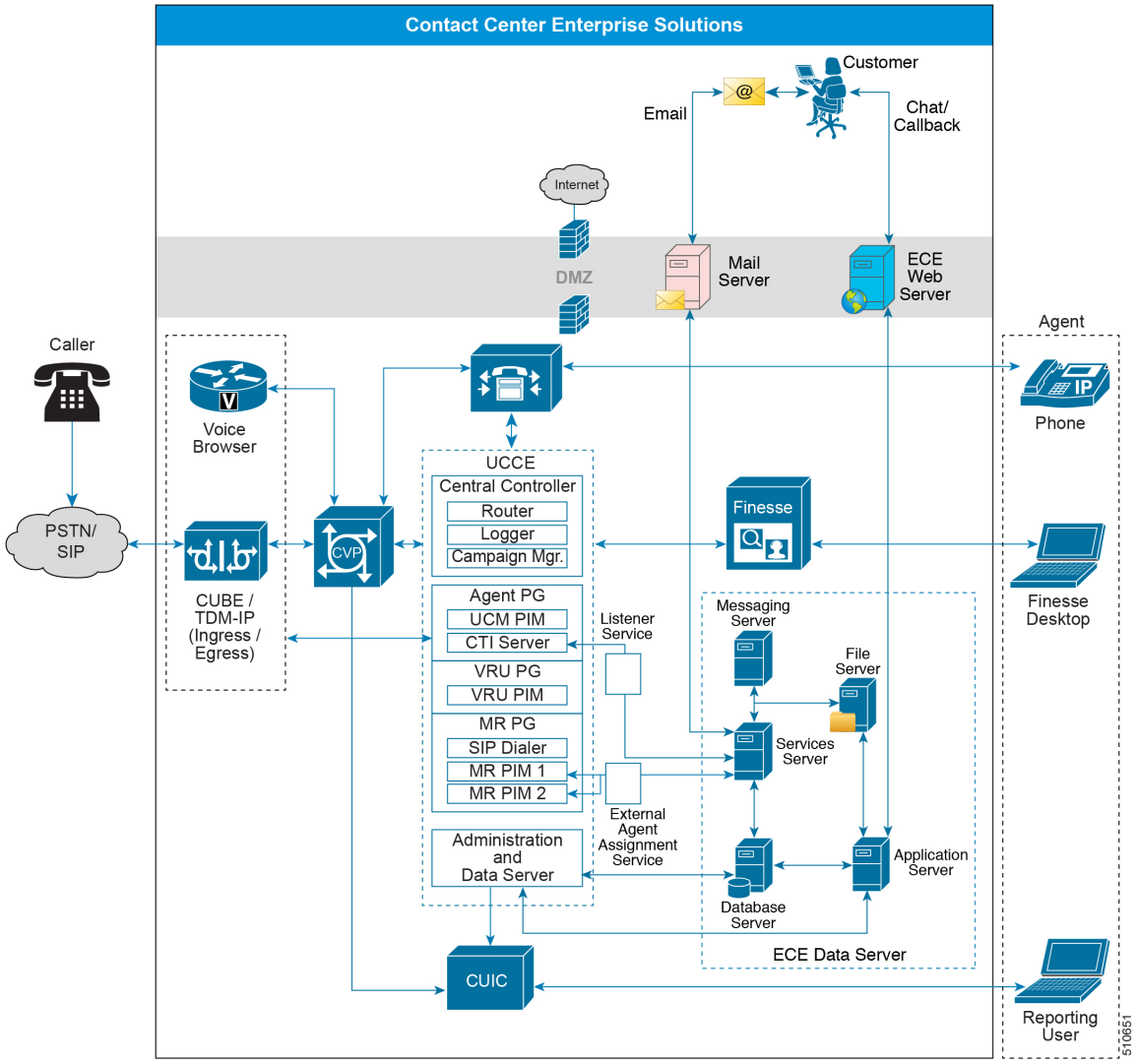
```
[Unified CM routing client VRU label + correlation-id]: pointing to CVP servers
[CVP routing client VRU label + correlation-id]: pointing to VXML Gateways
```

Enterprise Chat and Email Design Considerations

Enterprise Chat and Email provides web and email interaction management through a common set of web servers and pages for agents and administrators. It integrates with the contact center enterprise solution to provide universal queuing of contacts to agents from different media channels.

For more architectural and design information, see the *Enterprise Chat and Email Design Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/cisco-enterprise-chat-email/products-implementation-design-guides-list.html.

Figure 72: Enterprise Chat and Email Design Considerations



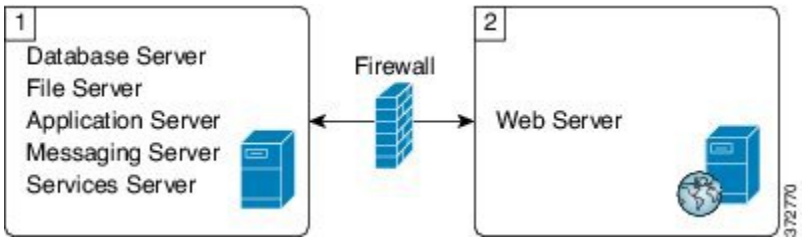
Enterprise Chat and Email Deployment Options

Due to the modular, component-based architecture, Enterprise Chat and Email (ECE) caters to the growing demands for concurrent user loads. To provide the flexibility to suit deployments of varied sizes, ECE supports various components that you can distribute across various servers in your solution.

Collocated Deployment

In a collocated deployment, you install the web server on one server and install all other components on a separate server. You can install the web server outside the firewall, if necessary.

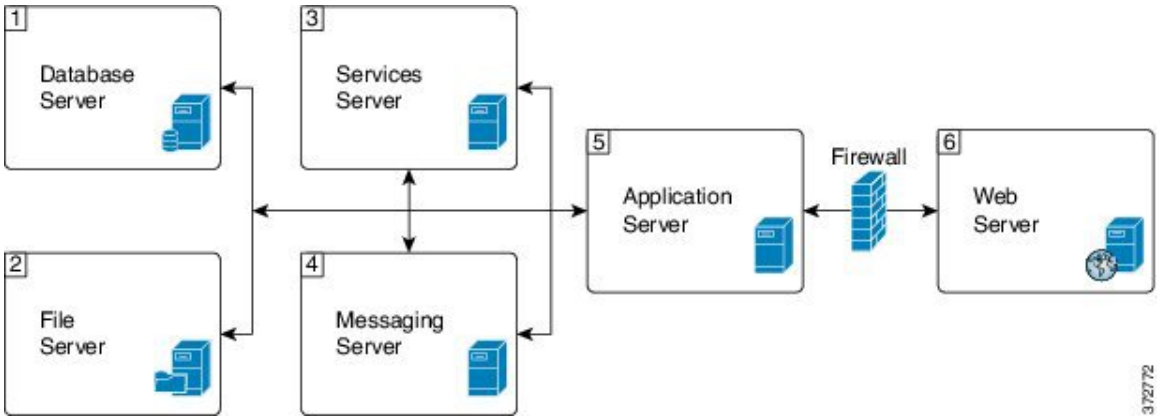
Figure 73: Collocated Deployment



Distributed-Server Deployment

In a distributed-server deployment, you install each component on a separate server, with the web server outside the firewall. You can restart the application, messaging, services, and web servers independently in this configuration without restarting any other servers.

Figure 74: Distributed-Server Deployment



Silent Monitoring Design Considerations

For supervisors managing teams, Unified CM-based Silent Monitoring provides a non-invasive mechanism to monitor agent voice calls.

Unified CM-based Silent Monitoring Design Considerations

Unified CM-Based Silent Monitoring Call Flow

Cisco Finesse provides silent monitoring functionality through Unified CM Silent Monitoring. Cisco Finesse works with Unified CM Silent Monitoring as follows:

1. The supervisor application sends a REST request to the Cisco Finesse server to begin silent monitoring.
2. The Cisco Finesse server sends the AgentSuperviseCall() message to Unified CCE to start the silent monitoring session.
3. Unified CCE sends the CallStartMonitor() message to Unified CM.
4. Unified CM instructs the supervisor phone to call the Built-In Bridge (BIB) on the agent phone.

5. The supervisor phone calls the BIB on the agent phone.
6. The agent phone forwards a mix of the agent voice stream and customer voice stream.
7. Unified CM sends call events for the silently monitored call to Unified CCE.
8. Unified CCE sends update events to the Cisco Finesse server.
9. The Cisco Finesse server sends XMPP updates to the Cisco Finesse supervisor application.

Cisco Finesse does not support silent monitoring of mobile agents. Supervisors cannot silent monitor mobile agents and mobile supervisors cannot perform silent monitoring.

Supervisors cannot perform silent monitoring from a Cisco Finesse IP Phone. Supervisors can only perform silent monitoring from the Cisco Finesse desktop.

Unified CM-Based Silent Monitoring Impacts

Unified CM accomplishes silent monitoring with a call between the supervisor (monitoring) device and the agent (monitored) device. The agent phone mixes and sends the agent's conversation to the supervisor phone, where it plays to the supervisor.

Unified CCE supports the Silent Monitoring functionality available in Unified CM. Unified CM Silent Monitoring supports only one silent monitoring session and one recording session for the same agent phone.

Note Unified CM Silent Monitoring does not support mobile agents.

Unified CM Silent Monitoring can monitor any Unified CCE agent desktop, if the following conditions exist:

- The monitored agent uses a compatible Cisco Unified IP phone or Cisco IP Communicator. For more details, see the *Compatibility Matrix* for your solution.
- The contact center uses a compatible version of Cisco Unified CM. For more information, see the *Compatibility Matrix* for your solution.

Supervisors can use any Cisco IP Phone, including Cisco IP Communicator, to silently monitor agents.

Unified CM Silent Monitoring works the same as other call control functionality provided by Unified CM (such as conference and transfer). When the silent monitoring session begins, the desktop sends a message through Unified CCE, through Unified CM, and out to the phones where silent monitoring runs.

Messaging through Unified CCE and Unified CM impacts Unified CCE performance.

Call Transcript Design Considerations

The Call Transcript feature enables the transcript of the call between an agent and a caller to be made available to the agent or the Admin on the user desktop. The agent can view the transcript and summary of the call and configured keywords on the desktop. The transcript is available for future reference and enables the agent to take note of action items.

Call Transcript Architecture

The Call Transcript feature leverages the Network-Based Recording (NBR) feature of Cisco Unified Call Manager (CUCM) to fork media streams of the caller and the agent to the Voicea SIP endpoints. However, the basic deployment architecture of NBR does not work as the recording servers are based on static recording endpoints that can be configured upfront on network elements. The Voicea SIP endpoints are constructed dynamically for each instance of the media stream, and require additional interactions between the solution components as depicted in the call flow diagram.

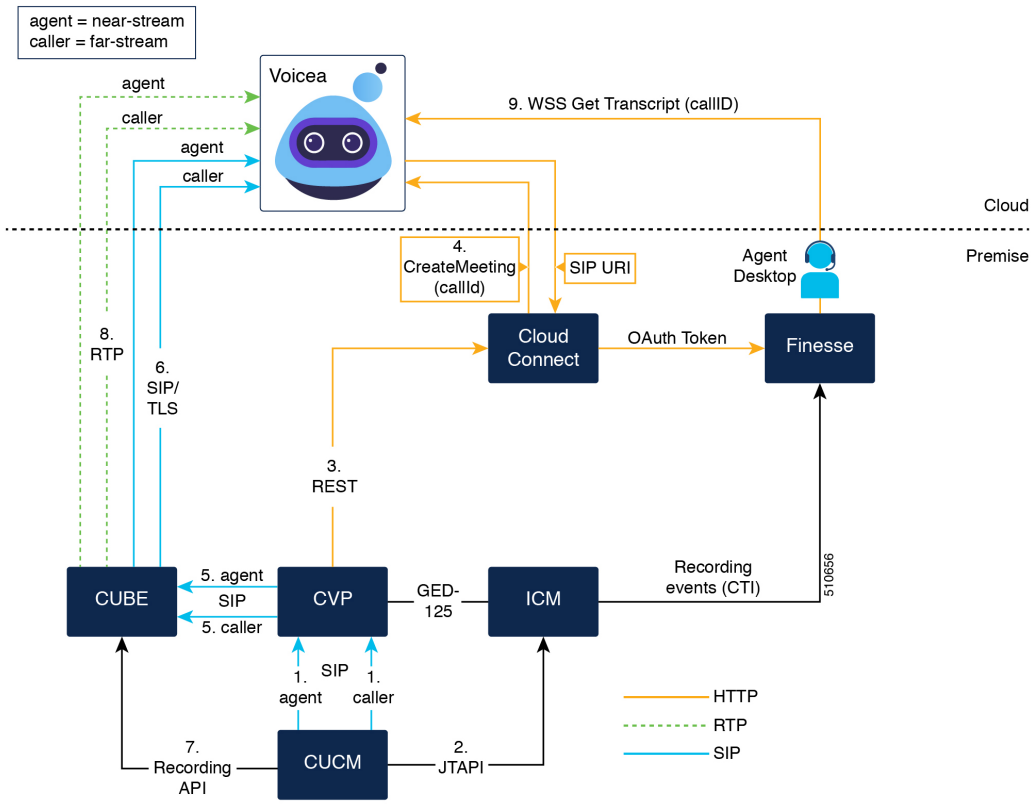
Every instance of a media stream is represented by a Meeting object in the Voicea cloud. Therefore, a basic call between a caller and an agent comprising two media streams (the far-end media stream corresponding to the caller and the near-end media stream corresponding to the agent) are represented by two different Meeting object instances in the Voicea cloud. These Meeting objects must be created before the actual media stream is transmitted to the Voicea cloud. When the Meeting object is created, a unique dynamic SIP URL is generated. The media stream corresponding to the Meeting instance must be forked to this URL. In addition, the two Meeting instances need to be associated together to generate a single merged transcript for the call.

To facilitate this association, a global callGUID generated by CUCM to link the two legs of the call for recording purposes is also used to associate the two Meeting instances for the call in the Voicea cloud.

Call Transcript works by configuring NBR on CUCM-registered endpoints with media forking set to **Gateway-preferred**. This implies that the transcript functionality can be controlled only by endpoint configuration, and not at the agent, team, or skill group level. CVP is configured as the static recording server from the CUCM perspective, and essentially abstracts the dynamic nature of the SIP URL from CUCM.

Once a call is established between caller and agent and is in the connected state, the sequence of the data and control flows to obtain the transcript for the call is detailed in the following diagram.

Figure 75: Call Transcript Call Flow



Call Transcript Sequential Call Flow

1. A call comes into the Ingress Gateway (CUBE).

Note If automatic recording is enabled on an endpoint and a call to that endpoint goes to the connected state, CUCM sends two invites on its configured SIP trunk to CVP - one for the agent, and the other for the caller.

2. The invite from CUCM contains callGUID, which CVP uses to initiate media forking for the two endpoints. It invokes the Cloud Connect API to proxy through the endpoint meta-data and callGUID to Call Transcript. Call Transcript returns two dynamic SIP URLs corresponding to the two endpoints of the call.
3. Once the URLs are obtained, CVP initiates the SIP invites to initiate media forking to Call Transcript directly, or uses CUBE to forward the invites to Call Transcript.
4. Call Transcript returns the RTP address and port information for media transmission to CUBE/CVP, and from there to CUCM.
5. Media forking results in the generation of two JTAPI events which are delivered as NETWORK_RECORDING_STARTED_EVENT and NETWORK_RECORDING_TARGET_INFO_EVENT CTI events to Finesse. The

NETWORK_RECORDING_TARGET_INFO_EVENT event has callGUID, which is used by CVP to create the meetings in Call Transcript.

6. In response to the NETWORK_RECORDING_TARGET_INFO_EVENT, Finesse obtains a Voicea OAuth token from Cloud Connect and sends the token and callGUID to the User Desktop in a Dialog event.
7. The Call Transcript gadget on the User Desktop (Finesse) uses callGUID and the token to retrieve the WSS URLs corresponding to the two meetings associated with this call.
8. The gadget then creates two Websocket channels to Call Transcript cloud to receive the live transcript - one for the agent and the other for the caller.
9. The CTS gadget receives the live transcripts and merges them on the gadget. The merged live transcript is displayed to the agent with relative timestamps and speaker tags.

The availability of the call transcript for a particular call depends on whether the corresponding agent endpoint is configured for NBR with gateway-preferred forking. The latency that Cloud Connect experiences on the createMeeting API invocations has a direct bearing on its call-handling capacity. These latencies are logged in CVP as well as in Cloud Connect.

The channel between CUBE and Voicea is not secured. If a secure channel is desired, an IPSEC or secure MPLS tunnel needs to be created between CUBE and Voicea SIP endpoint, since Voicea currently does not support secure RTP and SIP/TLS.

Since CUCM can be configured only with one recorder, to have both the call transcript as well as a conventional recording solution in place, a CUBE media proxy needs to be deployed, which is capable of forking the media stream to 5 concurrent destinations. For more information, refer to https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/voice/cube/configuration/cube-book/voi-cube-media-proxy.html.

Call Transcript functionality reduces the overall call-handling capacity of the CUBE as every incoming call now requires 2 additional media streams for forking – the same as any network-based recording solution. This needs to be taken into account for scaling CUBE to the call rate requirements of the deployment.

Third-Party Component Design Considerations

Use the following design considerations when any third-party component connects to Unified CCE through CTI Server.

All-Event Client Limits

The CTI server uses All-Event clients.

Maximum for Two vCPU Small Agent PG OVAs

On each Agent PG, you can have a maximum of seven All-Event clients on the CTI server. Cisco Finesse uses two of these clients.

Maximum for Four vCPU Large Agent PG OVAs

VMs built from the large OVA with four vCPU can support more All-Event Clients and monitor-mode connections. On each Agent PG, you can have a maximum of 20 All-Event clients on the CTI server. Cisco Finesse uses two of these clients.

All-Event Clients

The Cisco Finesse desktop solution uses two of the available All-Event clients. Some of the other possible consumers of the clients are:

- Outbound Dialer
- Real-Time Adherence (2)
- Some third-party recording vendors (2)
- Enterprise Chat and Email (2)
- B+S CRM Connectors

DNS Server Deployment Considerations

Consider the following when configuring a DNS servers for your solution:

- Configure the DNS servers for reverse lookup.
- Do not configure the DNS servers beyond a NAT network boundary.
- Deploy redundant DNS servers with low latency on the connection with the servers performing lookups.

Load Balancer Design Considerations

Load Balancers for Cisco Finesse Sign-In

After agents sign in to the Cisco Finesse desktop, the Cisco Finesse desktop client caches the IP address of both Cisco Finesse servers. If a Cisco Finesse server goes out of service, the Cisco Finesse client automatically redirects and signs the agent in to the other Cisco Finesse server. Given this client-side logic, the use of a load balancer for failover purposes is not supported.

However, the following are two scenarios in which you can use a load balancer with Cisco Finesse.

Note These scenarios only apply to the Cisco Finesse desktop. These scenarios do not apply to Cisco Finesse IP Phone Agents.

Navigation to the Cisco Finesse Sign-In Page

When an agent navigates to a Cisco Finesse server that is down or not reachable, the agent cannot access the sign-in page. The agent receives an error and must manually sign in to the other Cisco Finesse server. To avoid this manual step, you can use a load balancer with URL redirect mode to direct the agent to an active Cisco Finesse server. The Cisco Finesse SystemInfo REST API provides the status of the Cisco Finesse server. For details about this API, see the *Cisco Finesse Web Services Developer Guide*.

When you configure a load balancer to determine the status of the Cisco Finesse servers, use this call flow:

1. The agent points their browser to the load balancer.
2. The load balancer redirects the agent browser to an appropriate Cisco Finesse server.

3. The agent signs in to the Cisco Finesse server directly. At this stage, the load balancer is no longer part of the call flow.

Direct Use of the Cisco Finesse API

If you use the Cisco Finesse REST API directly, the call flow cannot use the Cisco Finesse client-side failover logic. You can opt to use a load balancer to manage high availability. The load balancer is part of a custom application which, like all custom applications, Cisco does not support. You or a Cisco partner provide the required support for the load balancer.

Remember that there are two connections between Cisco Finesse clients and the Cisco Finesse server:

- A REST channel for requests and responses
- An XMPP channel to send notifications from the server to the client

Both channels for a given client must connect to the same Cisco Finesse server. You cannot connect the load balancer to the REST connection for one Cisco Finesse server and to the XMPP channel connection for the other Cisco Finesse server. This setup provides unpredictable results and is not supported.

Load Balancers for Cisco Unified Intelligence Center (CUIC)

Unified Intelligence Center can be accessed using load balancers. The following conditions apply:

- Live Data reports cannot be accessed through the load balancer.
- Load balancer is not supported when CUIC nodes and browser clients are split across a WAN.

Load Balancers for CVP

You can use load balancers with the Unified CVP solution components to provide load distribution and high availability for HTTP, HTTPS, and MRCP traffic. Load balancers can spread the rendering of the VXML pages between VXML Gateways and VXML Server. Load balancers can also spread the fetching of the media files for VRU scripts run from media servers.

Note If your solution is MRCPv2, use CUSP for load balancing.

CVP now supports any third-party load balancer that meets these requirements:

- Supports both SSL offloading and SSL pass-through
- Supports load balancer high availability
- Does not have mandatory session stickiness
- Uses cookie-insert for persistence
- Distribution algorithm is round-robin

Load Balancers for the Unified CCE Administration Tool

You can use a load balancer with the Unified CCE Administration tool in these scenarios.

Navigation to the Unified CCE Administration Sign-In Page

When administrators or supervisors navigate to the Unified CCE Administration tool on a server that is down or not reachable, they cannot access the sign-in page. They receive an error and must manually sign in to Unified CCE Administration on the other server. To avoid this manual step, you can use a load balancer with URL redirect mode to direct their sessions to an active server.

Usage scenario:

1. Users point their browsers to the load balancer.
2. The load balancer redirects the browser sessions to an appropriate Unified CCE Administration server.
3. The users sign in to the Unified CCE Administration server directly.

Direct Use of the Unified CCE Administration API

If you use the Unified CCE Administration REST API directly, you can opt to use a load balancer to manage high availability. The load balancer is part of a custom application which, like all custom applications, Cisco does not support. You or a Cisco partner provide the required support for the load balancer.

Load Balancing for Unified CCE Administration for Packaged CCE Initialization

You cannot use a load balancer as part of the fresh install process for Packaged CCE. The user must sign in to Unified CCE Administration on the Side A Unified CCE AW-HDS-DDS to initialize the Packaged CCE deployment type.

Load Balancers with Enterprise Chat and Email

Do not use the load balancer's redirect mode with Enterprise Chat and Email

Recording Design Considerations

Network-Based Recording Design Considerations

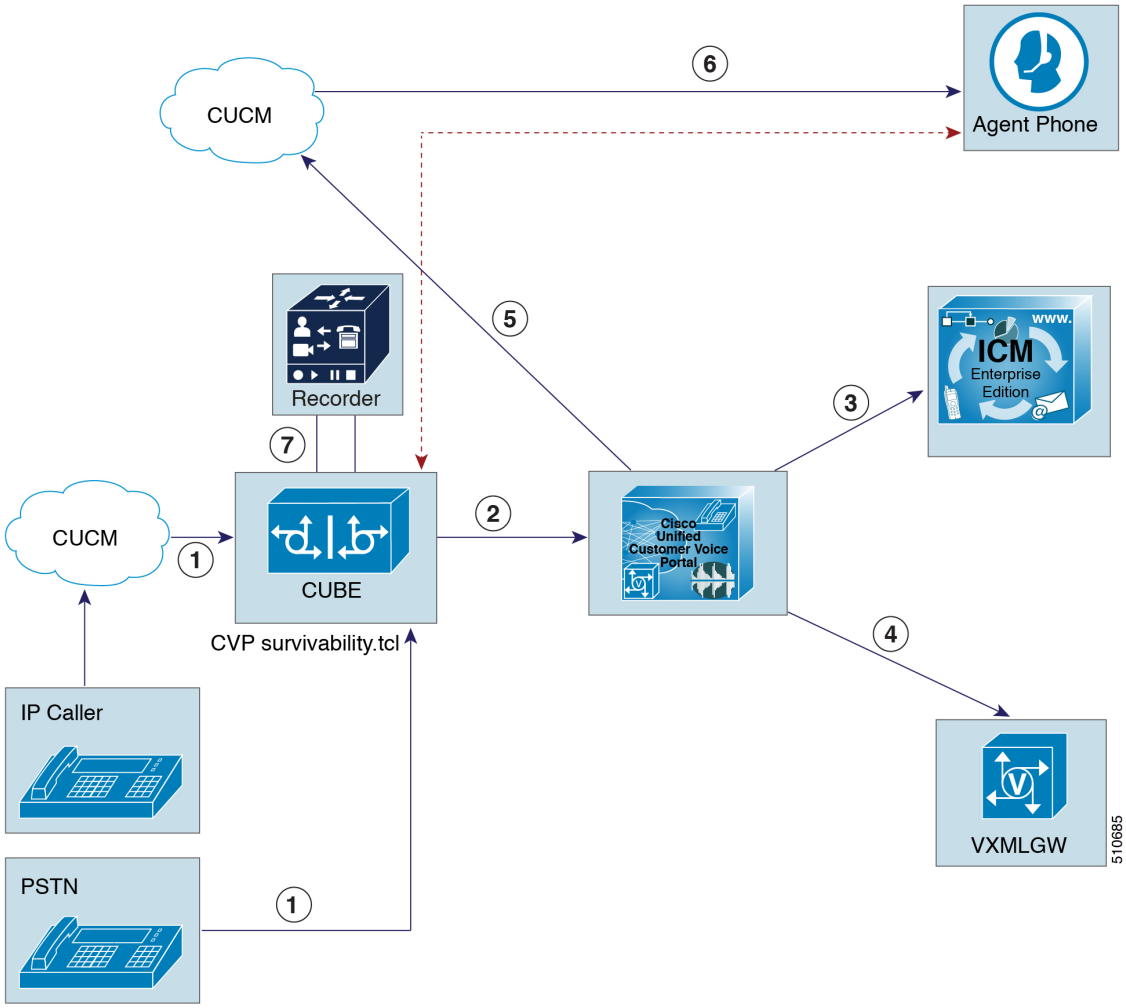
The network-based recording (NBR) feature supports software-based forking for Real-time Transport Protocol (RTP) streams. With media forking, you can create midcall multiple streams (or branches) of audio and video for a single call. You can then send the streams of data to different destinations. To enable network-based recording using CUBE, refer to its configuration guide. You can configure specific commands or use a call agent. CUBE acts as a recording client.

Note Network-based recording works with the call survivability feature.

This figure shows the call flow for network-based recording.

Figure 76: Network-Based Recording Call Flow

Figure 77: Network-Based Recording Call Flow



A typical call flow for network-based recording is as follows:

1. The incoming call arrives at CUBE.
2. The Ingress Gateway sends the call to Unified CVP.
3. Unified CVP sends the incoming call request to Unified CCE and gets a VRU label.
4. Unified CVP sends the call to the VXML Gateway. The caller hears the VRU. However, the call is not recorded.
5. After the agent is available, Unified CVP connects the caller to the agent.
6. Network-based recording starts for this conversation.

Network-Based Recording Limitations

- For agent to agent call transfer, network-based recording does not work but phone-based recording does. If you want to use network-based recording, you can use an ISR gateway between Unified CVP and Unified CM.
- The NBR feature is supported only on selected IOS image trains. For more information about the supported IOS image trains, see the *Compatibility Matrix* for your solution.

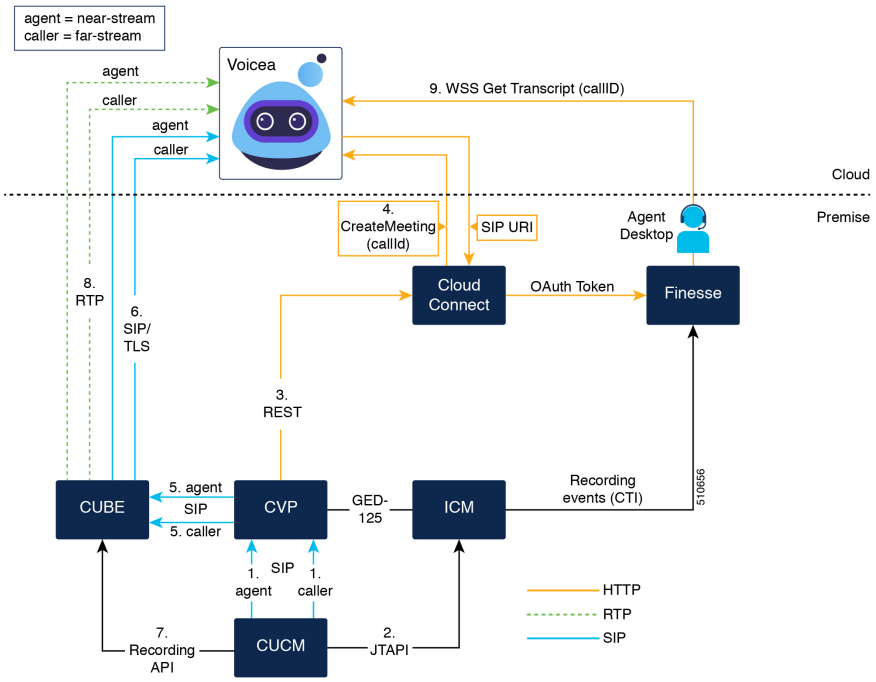
Call Transcript Design Considerations

The Call Transcript feature enables the transcript of the call between an agent and a caller to be made available to the agent or the Admin on the user desktop. The agent can view the transcript and summary of the call and configured keywords on the desktop. The transcript is available for future reference and enables the agent to take note of action items.

Call Transcript Design Considerations

The Call Transcript feature leverages the Network-Based Recording (NBR) feature of CUCM to fork media streams originating from endpoints (caller and agent) to the Voicea SIP endpoints. However, the basic deployment architecture of NBR as outlined in the previous section does not work as they are based on static recording endpoints that can be configured up front on network elements. The Voicea SIP endpoints are constructed dynamically for each and every instance of the media stream, and require additional interactions between the solution components as depicted in the diagram below.

Figure 78: Call Transcript Call Flow Diagram



Every instance of a media stream is represented by a Meeting object in the Voicea cloud. Therefore, a basic call between caller and agent comprising two media streams (the far-end media stream corresponding to the caller and the near-end media stream corresponding to the agent) is represented by two different Meeting object instances in the Voicea cloud. These Meeting objects have to be created prior to the transmission of

the actual media stream to the Voicea cloud, since the creation of the Meeting object results in the generation of a unique dynamic SIP URL to which the media stream corresponding to the Meeting instance needs to be forked. In addition, the two Meeting instances need to be associated together in order to generate a single merged transcript for the call.

In order to facilitate this association, a global callGUID generated by CUCM in order to link the two legs of the call for recording purposes is also used to associate the two Meeting instances for the call in the Voicea cloud.

Call Transcript works by configuring NBR on CUCM registered endpoints with media forking set to 'Gateway-preferred". This implies that transcript functionality can be controlled only by endpoint configuration, and not at the agent, team, or skill group level. CVP is configured as the static recording server from the CUCM perspective, and essentially abstracts the dynamic nature of the SIP URL from CUCM. Once a call is established between caller and agent and is in the connected state, the sequence of data and control flows in order to obtain the transcript for the call is as follows:

1. CUCM detects that the near-stream endpoint is configured for NBR with gateway-preferred forking, and sends 2 SIP INVITES to CVP corresponding to the near-stream (agent) and far-stream (caller) with a common callGUID linking the two streams.
2. CUCM also generates corresponding JTAPI network recording events that are handled by the JTAPI gateway in the Agent PG and converted to CTI recording events that are consumed by Finesse. The recording events alert Finesse to the fact that the call is being recorded and Finesse retrieves the common callGUID for the recording session from the CTI recording events.
3. CVP uses the callGUID in the SIP INVITE messages from CUCM to invoke 2 createMeeting API calls on Cloud Connect VM.
4. Cloud Connect VM uses the stored tenant Id credentials to invoke the corresponding createMeeting API calls on the Voicea service endpoints. A dynamic SIP URL is returned for each invocation of the createMeeting API call along with the required authentication information. Cloud Connect VM returns this information to CVP in response to its API invocation on Cloud Connect.
5. CVP now formulates two new SIP INVITEs corresponding to each of the media streams with the dynamic SIP URL obtained from Voicea (via Cloud Connect) and sends it to CUBE.
6. CUBE in turn sends the SIP INVITE messages to the Voicea SIP endpoint and obtains the RTP endpoints for media transfer. This requires the CUBE endpoint to be unblocked at the Voicea cloud. CUBE obtains the RTP endpoints for media transfer in response to the SIP INVITE messages from Voicea, which passes them to CVP, which in turn passes the information to CUCM, in response to the SIP INVITE messages that it received from CUCM in Step 1.
7. CUBE is configured as a XMF application in CUCM. After CUCM receives the RTP endpoint information from CVP, it instructs CUBE via the Kayuga API to commence media forking for the near and far-end srteams.
8. CUBE forks the media to the Voicea cloud.

The availability of the call transcript for a particular call depends on whether the corresponding agent endpoint is configured for NBR with gateway-preferred forking. The latency that CVP experiences on the createMeeting API invocations has a direct bearing on its call handling capacity. These latencies are logged in CVP and the responses to these calls must be obtained in <XXX> milliseconds or less.

The channel between CUBE and Voicea is not secured. If a secure channel is desired, an IPSEC or secure MPLS tunnel needs to be created between CUCE and Voicea SIP endpoint, since Voicea currently does not support secure RTP and SIP/TLS.

Since CUCM can be configured only with one recorder, in order to have both the call transcript as well as a conventional recording solution in place, a CUBE media proxy needs to be deployed, which is capable of forking the media stream to 5 concurrent destinations. (https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/voice/cube/configuration/cube-book/voi-cube-media-proxy.html)

Call transcript functionality reduces the overall call handling capacity of the CUBE as every incoming call now requires 2 additional media streams for forking–the same as any network-based recording solution. This needs to be taken into account for scaling CUBE to the call rate requirements of the deployment.

Call Transcript Sequential Call Flow

1. A call comes into the Ingress Gateway (CUBE).

Note If automatic recording is enabled on an endpoint and a call to that endpoint goes to the connected state, CUCM sends two invites on its configured SIP trunk to CVP - one for the agent, and the other for the caller.

2. The invite from CUCM contains callGUID, which CVP uses to initiate media forking for the two endpoints. It invokes the Cloud Connect API to proxy through the endpoint meta-data and callGUID to Call Transcript. Call Transcript returns two dynamic SIP URLs corresponding to the two endpoints of the call.
3. Once the URLs are obtained, CVP initiates the SIP invites to initiate media forking to Call Transcript directly, or uses CUBE to forward the invites to Call Transcript.
4. Call Transcript returns the RTP address and port information for media transmission to CUBE/CVP, and from there to CUCM.
5. Media forking results in the generation of two JTAPI events which are delivered as NETWORK_RECORDING_STARTED_EVENT and NETWORK_RECORDING_TARGET_INFO_EVENT CTI events to Finesse. The NETWORK_RECORDING_TARGET_INFO_EVENT event has callGUID, which is used by CVP to create the meetings in Call Transcript.
6. In response to the NETWORK_RECORDING_TARGET_INFO_EVENT, Finesse obtains a Voicea OAuth token from Cloud Connect and sends the token and callGUID to the User Desktop in a Dialog event.
7. The Call Transcript gadget on the User Desktop (Finesse) uses callGUID and the token to retrieve the WSS URLs corresponding to the two meetings associated with this call.
8. The gadget then creates two Websocket channels to Call Transcript cloud to receive the live transcript - one for the agent and the other for the caller.
9. The CTS gadget receives the live transcripts and merges them on the gadget. The merged live transcript is displayed to the agent with relative timestamps and speaker tags.

The availability of the call transcript for a particular call depends on whether the corresponding agent endpoint is configured for NBR with gateway-preferred forking. The latency that Cloud Connect experiences on the createMeeting API invocations has a direct bearing on its call-handling capacity. These latencies are logged in CVP as well as in Cloud Connect.

The channel between CUBE and Voicea is not secured. If a secure channel is desired, an IPSEC or secure MPLS tunnel needs to be created between CUBE and Voicea SIP endpoint, since Voicea currently does not support secure RTP and SIP/TLS.

Since CUCM can be configured only with one recorder, to have both the call transcript as well as a conventional recording solution in place, a CUBE media proxy needs to be deployed, which is capable of forking the media stream to 5 concurrent destinations. For more information, refer to https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/voice/cube/configuration/cube-book/voi-cube-media-proxy.html.

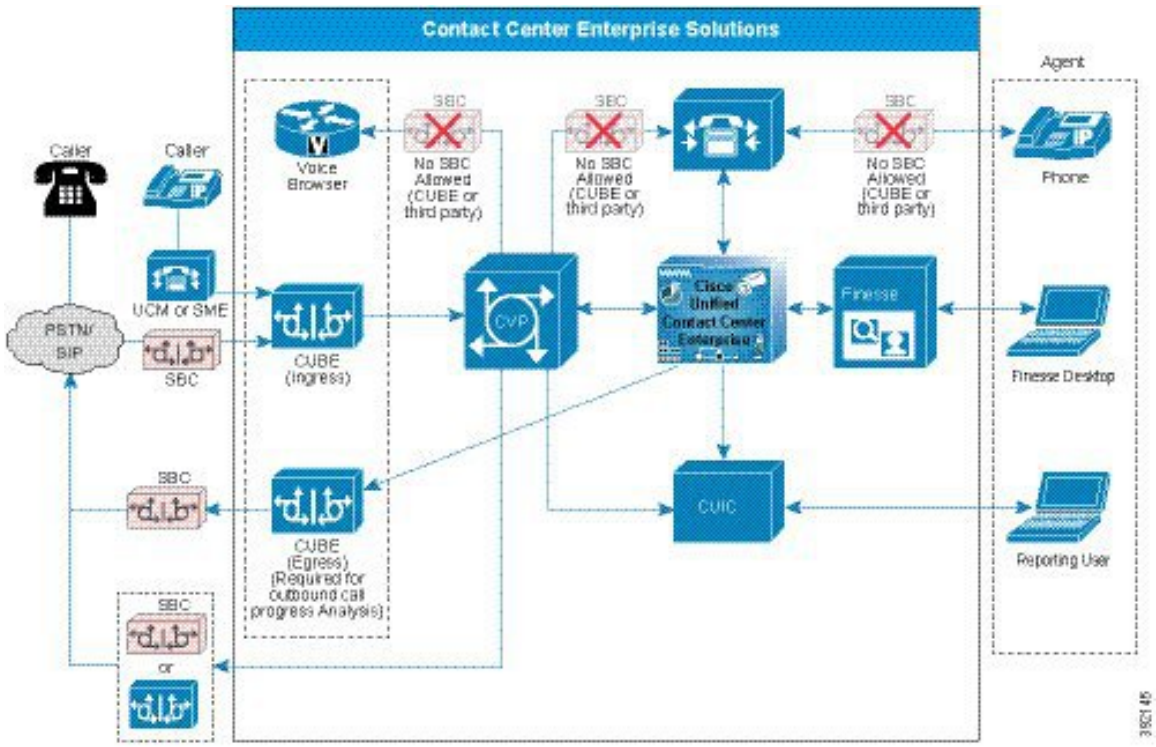
Call Transcript functionality reduces the overall call-handling capacity of the CUBE as every incoming call now requires 2 additional media streams for forking – the same as any network-based recording solution. This needs to be taken into account for scaling CUBE to the call rate requirements of the deployment.

Session Border Controllers

You can use a third-party Session Border Controller (SBC) as an ingress or egress border element between the PSTN and your contact center. Most solutions require CUBE between a third-party SBC and Unified CVP to ensure support for all contact center enterprise features. You cannot use an SBC in most other communication legs in your deployment.

The following diagram shows where you can and cannot use an SBC:

Figure 79: SBC Implementation



Note CUBE includes both SBC functionality and features for contact center enterprise solutions. You can replace CUBE with a third-party SBC where it supplies only SBC functions. Your solution still requires CUBE for the contact center enterprise features.

We do not actively test third-party SBCs.

Consider the following general design guidelines and limitations when you consider adding an SBC to your deployment.

Support Policy

Cisco may ask you to remove an SBC from the call flow temporarily for any interoperability issues. If the issue is not found without the SBC, then Cisco TAC hand over the case to the manufacturer of that SBC.

Third-Party SBC Use Without CUBE

When you connect to a third-party SIP device, including a SIP PSTN service provider, use a CUBE. If you do not place a CUBE between Unified CVP and the SIP device, ensure that both sides are compatible with thorough integration testing.

When connecting to a PSTN SIP Trunking service without a CUBE, carefully consider how to secure the connection between the contact center and the service provider. Also consider how to accomplish NAT and address hiding. Otherwise, the service-provider network can have full access to the contact center network. As the service-provider interconnect interface provided by Cisco, CUBE addresses both of these concerns.

For solutions that do not use certain CVP features, you might connect a third-party SBC directly to CVP. Such solutions require careful testing to ensure interoperability.

Important Designs that directly connect a third-party SBC to Unified CVP require a special exception from Cisco to deploy.

Supported Features

Without specific testing, support for any feature is not guaranteed. Past testing generally showed support for these features:

- G.711ulaw, G711alaw, and G.729 (no Annex B) codecs
- DNIS and ANI presentation
- SIP/TCP on the SBC's internal interface and SIP/UDP on the external interface
- CVP-based Queuing
- CVP applications with DTMF
- CVP-based intrasite transfers using re-INVITE
- Unified CM-based intrasite transfers and conferences
- SBC-based midcall codec negotiation. Basic call flows generally work, but complicated call flows are less likely to work.
- Cisco Unified Communications Manager (Unified CM) midcall codec negotiation (with transcoder insertion where needed). Basic call flows generally work, but complicated call flows are less likely to work.
- SBC converting SIP INFO messages from CVP to RFC2833 tones (for DTMF-based transfers)

Note Timing issues between CVP and the SBC can result in CVP disconnecting the call before the SBC completes the transfer.

- REFER transfers with SBC in REFER

Note CVP expects a BYE or a final NOTIFY that terminates the subscription from the SBC when the transfer is complete. If the SBC does not send either of these, the call ends for the duration of the subscription and the CVP license is not released.

- SIP 302 Redirect responses with SBC in consume mode
- CVP-based Redirect on No Answer
- Call hold

Unsupported Features

Without specific testing, support for any feature is not guaranteed. Past testing generally showed that these features are unsupported:

- Outbound Option with SIP Call Progress Analysis
- Courtesy Callback
- Call survivability and associated features (survivability.tcl script, local branch SRST and TOD routing, Hookflash, TBCT)
- Handling of VRU PG failure and any downstream failure handling through survivability
- Queue at the edge (using CVP `SendToOriginator` feature)
- Video call flows
- SIP over TLS and SRTP
- Locations-based CAC
- REFER with Replaces
- SBC configured as a SIP proxy (instead of CUSP) for messages between Cisco components
- Network Trunk Group Utilization and Reporting
- Trunk group utilization
- SIP Resource Availability Indicator (RAI) dynamic call routing
- SIP dial-peer based recording

Generally, features are unsupported because various Unified CCE and CVP features rely on specific CUBE functions that are not present in third-party SBCs.

Other Caveats

The following are other caveats to consider if you add a third-party SBC to your deployment:

- Most third-party SBCs do not generate a Cisco-Guid header which Unified CCE uses for end-to-end call tracking.
- If you use SIP over TCP between the SBC and CVP, some calls can drop if an SBC switches over through its high-availability feature. For each CVP server used, exactly one call can drop after the switch over occurs. All other calls in progress at the given CVP server stay active with stateful signaling and media. The dropped call is the first that sends a SIP message to the SBC after the switch over. SIP over UDP does not display this behavior.
- In solutions that use third-party SBCs, the network might have a firewall between the SBC and the subnet with the contact center solution components. The firewall configuration allows any SIP-related communication between the SBC and CVP.

  For the SIP traffic over TCP, CVP creates an outgoing connection with the SIP port for the SBC. On CVP, an idle TCP connection remains in the ESTABLISHED state for 4 hours, even if there are no calls between the SBC and CVP.

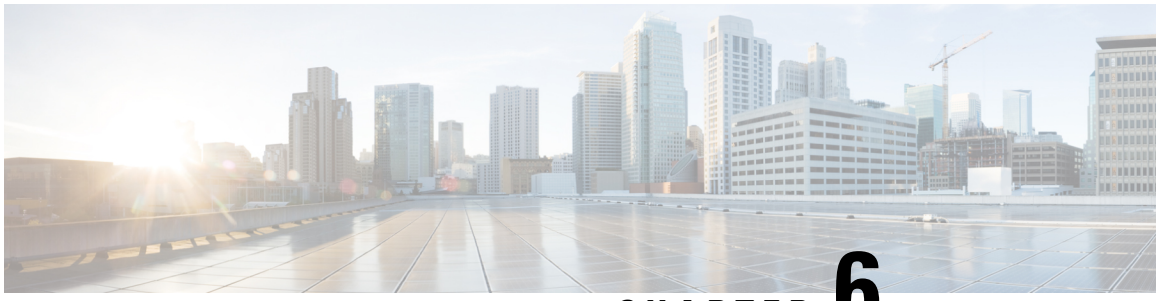
  The firewall configuration might free such idle connections without CVP detecting it. When CVP next receives incoming calls, a few calls fail because CVP cannot send the SIP requests on the outgoing TCP connection to the SBC. Calls can fail until CVP establishes a new connection to the SBC.
- In some scenarios, CVP should send busy and ring-no-answer notifications to the SBC. In such cases, use the Remote-Party-ID header and manipulate the header to include "--CVP" at the end of the display name.
- Not all SIP service providers support advanced features such as REFER, 302 Redirect Messages, DTMF-based take-back-and-transfer, or data transport (UUI, GTD, NSS, and so on).
- Unified CM can use an UPDATE method for session refresh on its signaling path to CVP. This use case is untested, and therefore unsupported, when CVP connects directly to a third-party SBC.
- CVP supports both IPv6 SDP and IPv4 SDP. Use Alternate Network Address Format (ANAT) when the SBC uses both IPv4 and IPv6 in the session description.

Cisco Virtualized Voice Browser

We have not tested Cisco VVB with any third-party SBC.

Speech Recognition and Text to Speech

Automatic Speech Recognition (ASR) or Text-to-Speech (TTS) Server cannot use silence suppression and must use the G.711 codec.



CHAPTER 6

High Availability and Network Design

High Availability Designs

Cisco contact center enterprise solutions have high availability features by design. Your solution design must include redundancy for the core components. The redundant components fail over automatically and recover without manual intervention. Your design can include more than that basic high availability capability. A successful deployment requires a team with experience in data and voice internetworking, system administration, and contact center enterprise solution design and configuration.

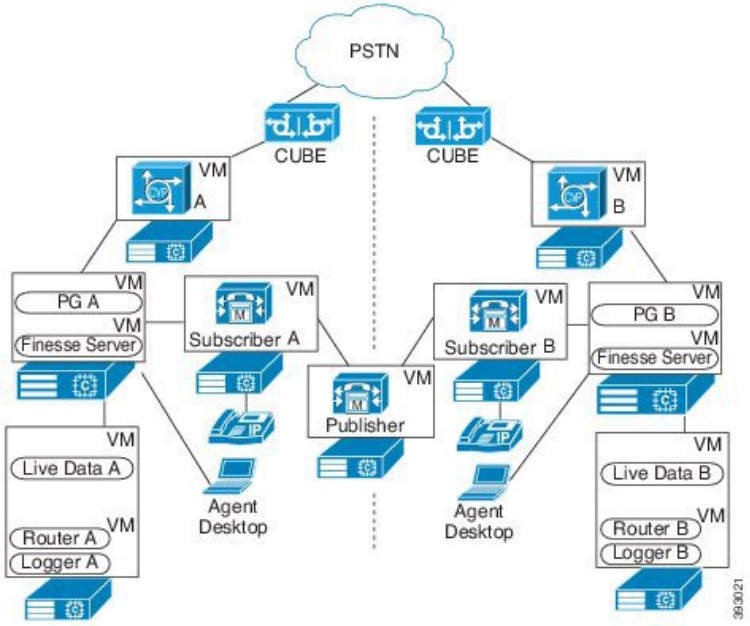
Each change to promote high availability comes at a cost. That cost can include more hardware, more software components, and more network bandwidth. Balance that cost against what you gain from the change. How critical is preventing disconnects during a failover scenario? Is it acceptable for customers to spend a few extra minutes on hold while part of the system recovers? Would the customer accept losing context for some calls during a failure? Can you invest in greater fault tolerance during the initial design to position the contact center for future scalability?

Plan carefully to avoid redesign or maintenance issues later in the deployment cycle. Always design for the worst failure scenario, with future scalability in mind for all deployment sites.

Note This guide focuses on design of the contact center enterprise solution itself. Your solution operates in a framework of other systems. This guide cannot provide complete information about every system that supports your contact center. The guide concentrates on the Cisco contact center enterprise products. When this guide discusses another system, it does not offer a comprehensive view. For more information about the complete Cisco Unified Communications product suite, see the Cisco solutions design documents at http://www.cisco.com/en/US/docs/voice_ip_comm/uc_system/design/guides/UCgoList.html.

The following figure shows a fault-tolerant Unified CCE single-site deployment.

Figure 80: Unified CCE Component Redundancy



Note The contact center enterprise solutions do not support nonredundant (simplex) deployments in production environments. You can only use non-redundant deployments in a testing environment.

This design shows how each component is duplicated for redundancy. All contact center enterprise deployments use redundant Unified CM, Unified CCE, and Unified CVP components. Because of the redundancy, your deployment can lose half of its core systems and be operational. In that state, your deployment can reroute calls through Unified CVP to either a VRU session or an agent who is still connected. Where possible, deploy your contact center so that no devices, call processing, or CTI Manager services are running on the Unified CM publisher.

To enable automatic failover and recovery, redundant components interconnect over private network paths. The components use heartbeat messages for failure detection. The Unified CM uses a cluster design for failover and recovery. Each cluster contains a publisher and multiple subscribers. Agent phones and desktops register with a primary target, but automatically reregister with a backup target if the primary fails.

High Availability and Virtualization

In a virtualized deployment, place the components carefully to maintain high availability. The mechanisms that support high availability are the same. But, distribute the components to minimize multiple failovers from a single failure. When you deploy on Direct Attached Storage (DAS) only systems, consider the following points:

- Failure of a VM brings down all the components that are installed on the VM.
- Failure of a physical server brings down all the VMs that are installed on that VMware vSphere Host.

Deployments on systems with shared storage can use some of the VMware High Availability features for greater resiliency. For specific information about supported VMware features, see the *Cisco Collaboration Virtualization* at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

To minimize the impact of hardware failures, follow these guidelines:

- Avoid placing a primary VM and a backup VM on the same physical server, chassis, or site.
- Avoid placing all the active components in a failover group on the same physical server, chassis, or site.
- Avoid placing all VMs with the same role on the same physical server, chassis, or site.

Server Failovers

When a server or blade fails over, active calls become inactive and incoming calls are disrupted. Processing resumes when the backup components become active. When the primary server recovers, processing of active and incoming calls returns to the primary server.

Virtualization Do's and Don'ts

Keep the following points in mind when planning your virtualization:

- Consider which components can be coresident and which components must be coresident on the same VMs. For more information about placement of components in virtual environments, see the virtualization web page for your solution.
- The contact center enterprise solutions do not support NIC teaming for the Guest OS (Windows or VOS).
- Set your NIC card and Ethernet switch to autonegotiate.

VMware High Availability Considerations

High availability (HA) provides failover protection against hardware and operating system failures within your virtualized contact center enterprise environment. You can use VMware's HA settings for contact center application VMs only if your solution uses SAN storage.

Consider the following when deploying your solution with VMware HA enabled:

- Cisco does not support VMware Distributed Resource Scheduler (DRS).
- In vCenter, select **Admission Control Policy** > **Specify a failover host**. When an ESXi host fails, all of the VMs on this host fail over to the reserved HA backup host. The failover host Admission Control

Policy avoids resource fragmentation. The Contact Center Enterprise Reference Design models assume a specific VM colocation within your solution. This VM colocation requirement guarantees system performance, based on contact center enterprise capacity requirements.

- HA Backup hosts must be in the same data center with the primary server, but not in the same physical chassis as the contact center blades. Use 10-GB networking connectivity for vSphere management.
- In vCenter, select the **VM monitoring status** > **VM Monitoring Only**.
- In vCenter, for the **Host Isolation response**, select the appropriate option to shutdown all the Virtual Machines.
- Configure your VMs with the **VM restart priority** as listed here:

Table 45: VM Settings

VM	VM Restart Priority
Cisco Unified Intelligence Center	Low
Contact Center Management Portal or Contact Center Domain Manager	Low
Unified CVP Reporting Server	Low
Unified CCE PGs	Medium
Cisco Finesse	Medium
Unified CVP Servers	High
Unified CCE Routers and Loggers	High
Cisco IdS	Medium
Cisco Unified Call Manager	High
Cisco Live Data	Low
Cisco Customer Collaboration Platform	Low

LAN and WAN Communications in Packaged CCE

Cisco requires VMware NIC Teaming to connect to the redundant physical switches for the Packaged CCE public and private networks. This requirement is due to the use of VMware virtual switch (vSwitch) and the nature of the fault and recovery mechanism. See the *Cisco Packaged Contact Center Enterprise Installation and Upgrade Guide* for more on this design.

Follow the network uplink designs in the *Virtualization for Cisco Packaged CCE* at http://docwiki.cisco.com/wiki/Virtualization_for_Cisco_Packaged_CCE and in the *Cisco Packaged Contact Center Enterprise Installation and Upgrade Guide*. Other network uplink design can cause unexpected fault scenarios that affect the recovery and operation of your contact center.

This fault-tolerant design uses two VMware vSwitches, one each for the public and private network VLANs. Each vSwitch has one active and one standby VMNIC. Split each pair of VMNICs across the physical NIC

adapters and alternate active paths to two data center switches. For example, put the active onboard and the standby on the add-in card. Configure the two data center switches for both public and private network communications. The uplinks ensure that no single failure (physical NIC card, cable, physical switch) causes the simultaneous loss of both the public and private network communications. They also prevent the simultaneous loss of both the active and standby VMNIC for the same vSwitch.

Note Configure VMware and the server management vSwitch separate from either the public or private network communications paths, following VMware best practices.

Latency Requirements

Packaged CCE deployments that you split over a campus LAN, Metro Area Network (MAN), or WAN require highly available public network communications between those data centers. These data centers also require a private network communications path with no "single point of failure" within that data center and WAN infrastructure results. Your network latency cannot result in the loss of both public and private network communications for more than 500 ms.

Network	Maximum Round-Trip Time Supported
Public	400 ms
Private	100 ms

Quality of Service and Bandwidth Provisioning

Your WAN must support QoS.

For video calls, enable QoS for video RTP communication and the data channel. Your network must have sufficient bandwidth to accommodate the following:

- Video communication between the endpoints
- Communication to the VXML/CUBE gateway and the recorder for the video playback
- Video recording by the recorder

For more information about video call bandwidth considerations, refer to the "Cisco Collaboration Solutions Design and Deployment Sizing Considerations" chapter of the *Cisco Collaboration System Solution Reference Network Designs*, at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/tsd-products-support-series-home.html.

Network Design for Reference Design Compliant Solutions

Tested Reference Configurations

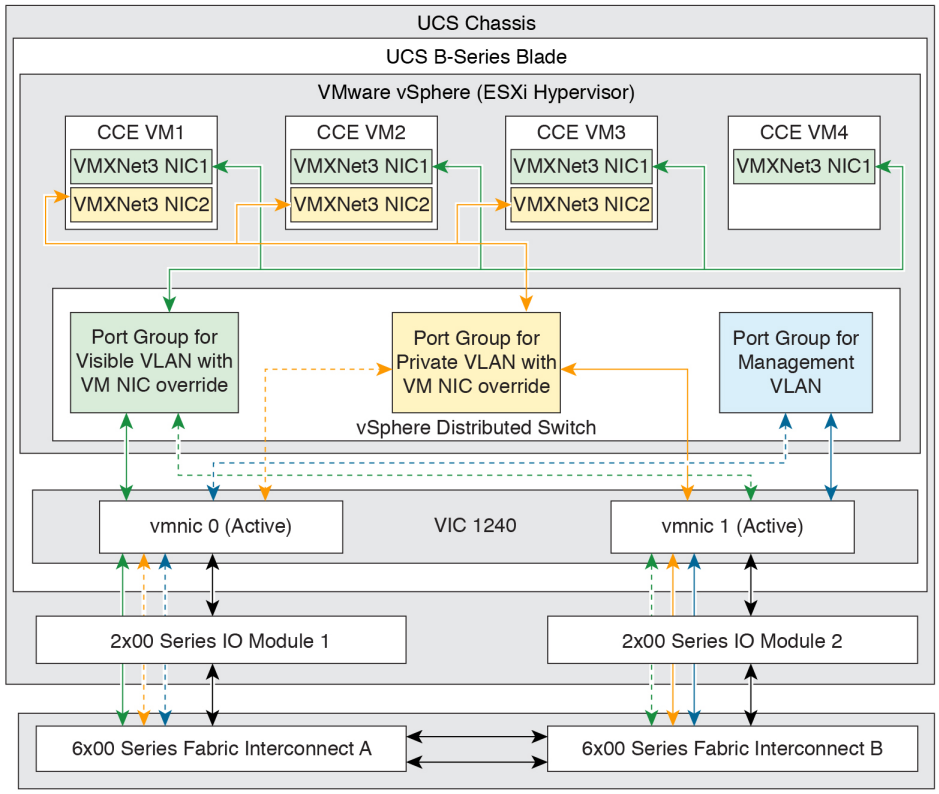
This section provides guidance on the network configuration of UCS deployments. It includes information on fault tolerance and redundancy.

Network Requirements for Cisco UCS B-Series Servers

The following figure shows the virtual to physical communications path from application local OS NICs to the data center network switching infrastructure.

This design uses a single virtual switch with two VMNICs in Active/Active mode. The design has public and private network path diversity aligned through the Fabric Interconnects using the Port Group VMNIC override mechanism of the VMware vSwitch. The design requires path diversity of the public and private networks to avoid failure of both networks from a single path loss through the Fabric Interconnects.

Figure 81: Network Requirements for Cisco UCS B-Series Servers



Reference Design	CCE VM1	CCE VM2	CCE VM3	CCE VM4
2000 Agent	Rogger	PG	NA	AW-HDS-DDS
4000 Agent	Rogger	PG	NA	AW-HDS-DDS AW-HDS
12000 Agent	Router	Logger	PG	HDS-DDS AW-HDS

510800

Contact Center with UCS B Fabric Interconnect requires the following:

- Fabric must be in end-host Mode.
- Ethernet interfaces must be 1/10 GB and connected to Gigabit Ethernet switches.
- No Fabric Failover must be enabled for vNICs in UCS Manager.

Note The Nexus 1000v and other Cisco distributed virtual switches based on the Nexus 1000v are not compatible with ESXi 6.5 after Update 1. See the VMware article on the discontinuation of third-party vSwitches for more details.

Data Center Switch Configurations

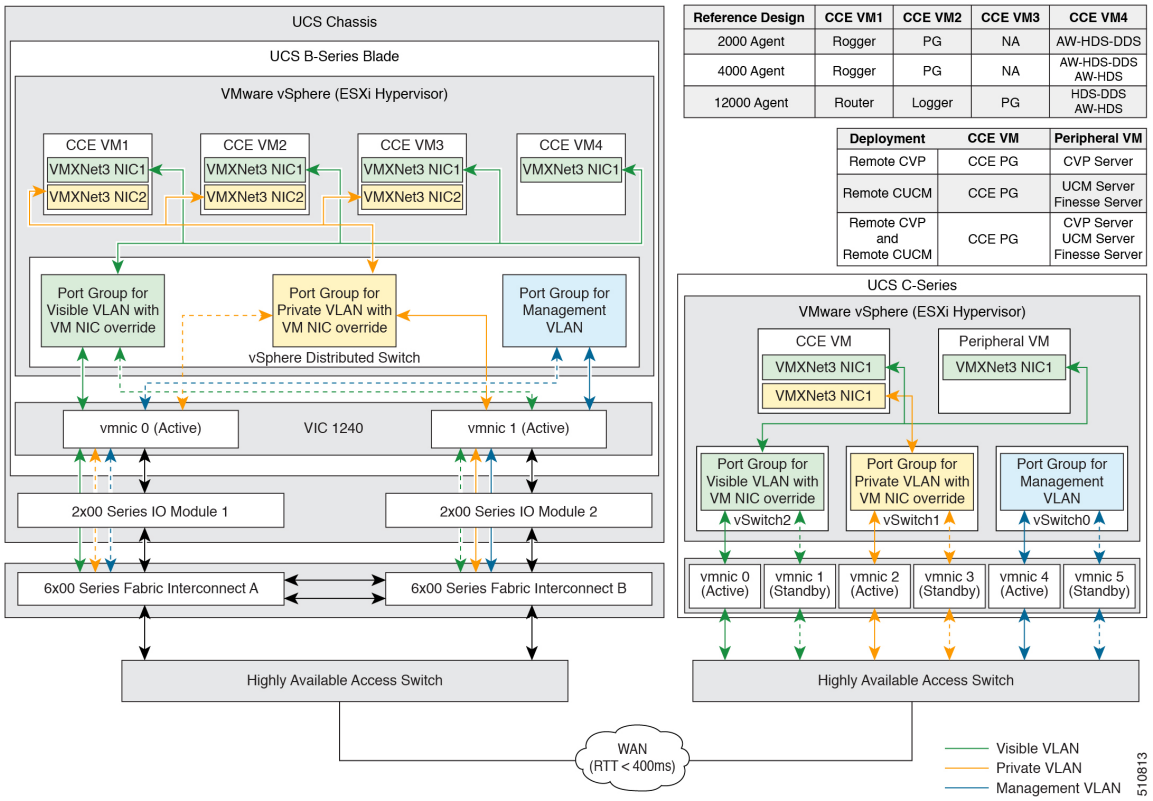
The contact center enterprise supports several designs for configuring Ethernet uplinks from UCS B-Series Fabric Interconnects to the data center switches. Your design requires Virtual Switch VLAN Tagging. Depending on data center switch capabilities, you can use either EtherChannel / Link Aggregation Control Protocol (LACP) or Virtual PortChannel (vPC).

The required design for public and private network uplinks from UCS Fabric Interconnects uses a Common-L2 design, where both VLANs are trunked to a pair of data center switches. Service Provider also may choose to trunk other management (including VMware) and enterprise networks on these same links, or use a Disjoint-L2 model to separate these networks. Both designs are supported, though only the Common-L2 model is used here.

C Series

This figure shows the reference design for all solutions on UCS C-series servers and the network implementation of the vSphere vSwitch design.

Figure 82: Network Requirements for Cisco UCS C-Series Servers



This design uses the VMware NIC Teaming (without load balancing) of virtual machine network interface controller (VMNIC) interfaces in an active/standby configuration. It uses alternate and redundant hardware paths to the network.

Your network side implementation can vary from this design. But, it requires redundancy and cannot have single points of failure that affecting both public and private network communications.

Ethernet interfaces must be 1/10 GB and connected to Gigabit Ethernet switches.

For more details on UCS C-series networking, see the *Cisco Collaboration Virtualization* page for your solution at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

PSTN Network Design Considerations

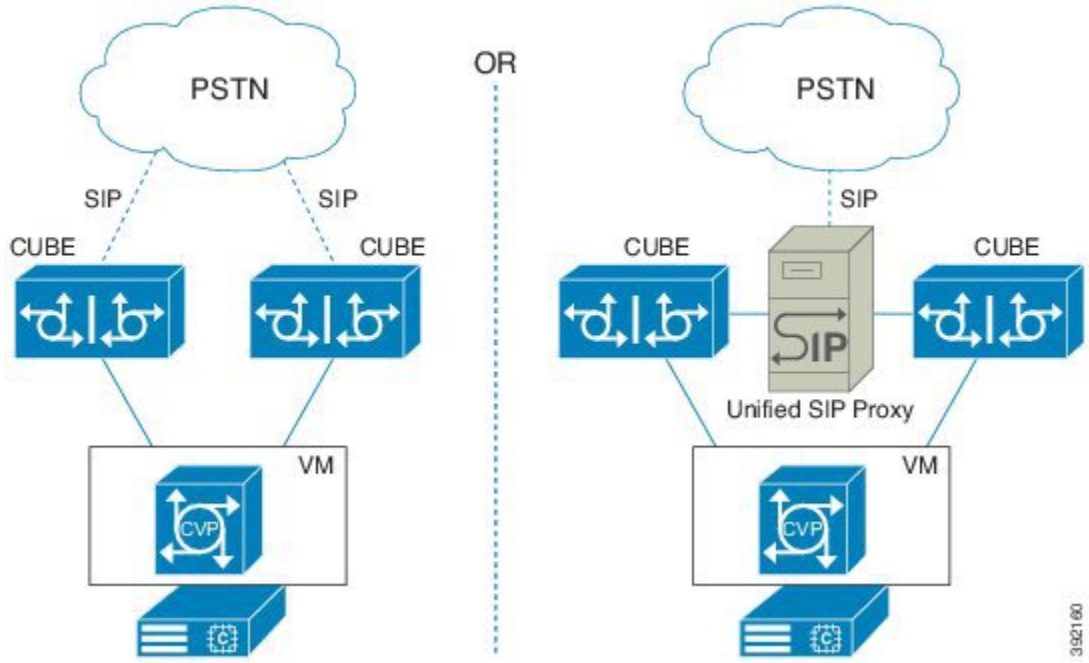
Highly available contact center designs start with the network infrastructure for data, multimedia, and voice traffic. A "single point of failure" in your network infrastructure devalues any other high availability features that you design into the contact center. Begin from the PSTN and ensure that incoming calls have multiple paths for reaching Unified CVP for initial treatment and queuing.

Ideally, design with at least two SIP trunks each connecting to a separate Cisco Unified Border Element (CUBE). If any CUBE or SIP trunk fails, the PSTN can route all traffic through the remaining SIP trunks. The PSTN route either by configuring all the SIP trunks as a large trunk group or by configuring rerouting or overflow routing to the other SIP trunks. You can also connect a redundant CUBE to each SIP trunk to preserve capacity when a Cisco UBE fails and the SIP trunk is still functional.

In some areas, the PSTN does not provide multiple SIP trunks to a single site. In that case, you can connect the SIP trunk to a Cisco Unified SIP Proxy (CUSP). Then, you could connect multiple CUBEs to the CUSP to provide some redundancy.

The CUBE passes calls to Unified CVP for initial treatment and queuing. Register each CUBE with a separate Unified CVP for load balancing. For further fault tolerance, you can register each CUBE with a different Unified CVP as a backup or configure a SIP Server group in CUBE. If a CUBE cannot connect with a Unified CVP, you can also use TCL scripts to provide some call processing. A TCL script can reroute the calls to another site or dialed number. The script can also play a locally stored .wav file to the caller and end the call.

Figure 83: High Availability Ingress Points



For more information about CUBE, Unified CVP, and voice networks in general, see the *Cisco Collaboration System Solution Reference Network Designs* at https://www.cisco.com/en/US/docs/voice_ip_comm/uc_system/design/guides/UCgoList.html.

Cisco Unified Survivable Remote Site Telephony (SRST)

Voice gateways using the Cisco Unified Survivable Remote Site Telephony (SRST) option for Unified CM follow a similar failover process. If the gateway is cut off from its controlling subscriber, the gateway fails over into SRST mode. The failover drops all voice calls and resets the gateway into SRST mode. Phones rehome to the local SRST gateway for local call control.

While running in SRST mode, Unified CCE operates as if the agents have no CTI connection from their desktops. The routing application detects the agents as not ready and sends no calls to these agents. When the gateway and subscriber reestablish their connection, the subscriber takes control of the gateway and phones again, allowing the agents to reconnect.

Active Directory and High Availability

Consider the following points that affect high availability when the network link fails between your contact center enterprise setup and Active Directory:

- Call traffic is not impacted during the link failure.
- The VMs in the domain restrict sign in using the domain controller credentials. You can sign in using cached credentials.
- If you stop Unified CCE services before the link fails, you must restore the link before starting the Unified CCE subcomponents.
- You cannot access the local PG Setup or sign in to the Unified CCE Web Setup.

- If the link fails while the Unified CCE services are active, access to Unified CCE Web Setup, configuration tools, and Script Editor fails.
- The administrator and super-users can access or configure any attribute, except the Reporting Configuration, in the Cisco Unified Intelligence Center OAMP portal.
- Supervisors cannot sign in to the Cisco Unified Intelligence Center Reporting portal. However, supervisors who are already signed in can access the reports.

Contact Center Enterprise Network Architecture

Cisco contact center enterprise solutions are distributed, resilient, and fault-tolerant network applications that rely on their network infrastructure meeting real-time data transfer requirements. A properly designed contact center enterprise network requires proper bandwidth, low latency, and a prioritization scheme that favors specific UDP and TCP traffic. The design requirements ensure the fault-tolerant message synchronization between redundant subcomponents. These requirements also ensure the delivery across the system of time-sensitive status data (routing messages, agent states, call statistics, trunk information, and so forth).

In your solution, WAN and LAN traffic comes in the following categories:

Voice and video traffic

Voice calls (voice carrier stream) consist of Real-Time Transport Protocol (RTP) packets that carry the actual voice samples between various endpoints such as PSTN gateway ports, Unified CVP ports, and IP phones. This traffic includes the voice streams for silently-monitored or recorded agent calls.

Call control traffic

Call control traffic includes data packets in several protocols (MGCP or TAPI/JTAPI), depending on the endpoints of the call. Call control includes functions to set up, maintain, tear down, or redirect calls. Call control traffic includes routing and service control messages that route voice calls to peripheral targets (such as agents or services) and other media termination resources (such as Unified CVP ports). Control traffic also includes the real-time updates of peripheral resource status.

Data traffic

Data traffic can include email, web activity, SIP signalling, and CTI database application traffic for the agent desktops. Priority data includes data for non-real-time system states, such as reporting and configuration update events.

This section discusses the data flows between the following:

- A remote Peripheral Gateway (PG) and the Unified CCE Central Controller (CC)
- The sides of a PG or a CC redundant pair
- The desktop application and the Finesse server

For more information on media (voice and video) provisioning, see the *Administration Guide for Cisco Unified Contact Center Enterprise* at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-maintenance-guides-list.html.

Network Link High Availability Considerations

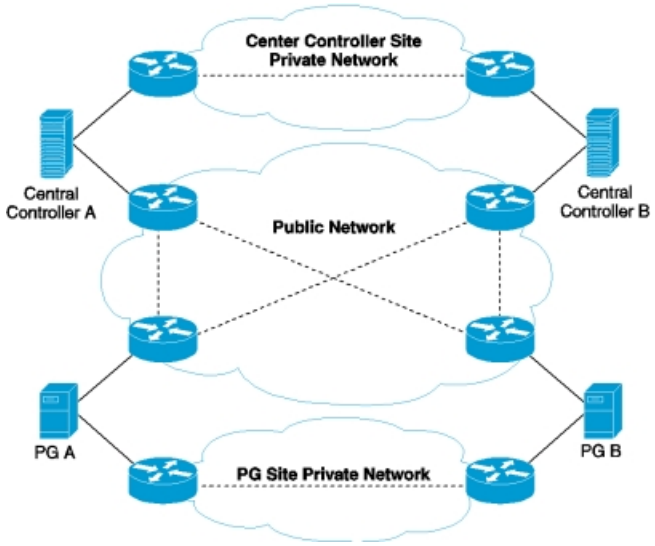
The fault-tolerant architecture employed by Unified CCE requires two independent communication networks. These networks are separate physical networks. The private network carries traffic necessary to maintain and

restore synchronization between the components. It also handles client communication through the Message Delivery Subsystem (MDS). The public network (using a separate path) carries traffic between the Central Controllers and PGs. The public network also serves as an alternate network for the fault-tolerance software to distinguish between component failures and network failures. For high availability, include redundant connections in your public network. Ideally, each connection uses a different carrier.

Note The public network is also called the visible network occasionally.

The figure below illustrates the network segments for a contact center enterprise solution. The redundant pairs of PGs and Central Controllers are geographically separated.

Figure 84: Example of Public and Private Network Segments for a Unified CCE System



In this case, the public network carries traffic between the Central Controller, PGs, and Administration & Data Servers. The public network never carries synchronization control traffic. Public network WAN links must have adequate bandwidth to support the PGs and Administration & Data Servers. You must use either IP-based priority queuing or QoS to ensure that the contact center traffic is processed within acceptable tolerances for both latency and jitter.

The private network carries traffic between the redundant sides of a Central Controller or a PG. This traffic consists primarily of synchronized data and control messages. The traffic also conveys the state transfer necessary to re-synchronize redundant sides when recovering from an isolated state. When deployed over a WAN, the private network is critical to the overall responsiveness of your contact center enterprise solution. The network has aggressive latency requirements. So, you must use either IP-based priority queuing or QoS on the private network links.

To achieve the required fault tolerance, the private WAN link is fully independent from the public WAN links (separate IP routers, network segments or paths, and so forth). Independent WAN links ensure that a single point of failure is truly isolated between the public and the private networks. Deploy public network WAN segments that traverse a routed network so that you maintain the route diversity between the PG and the Central Controller throughout the network. Avoid routes that result in common path selection and a common point of failure for the multiple sessions.

PGs and Administration & Data Servers local to one side of the Central Controller connect to the local Central Controller side through the public Ethernet and to the remote Central Controller side over public WAN links. Optionally, you can deploy bridges to isolate PGs and Administration & Data Servers from the Central Controller LAN segment to enhance protection against LAN outages.

Public Network Traffic Flow

The active PG continuously updates the Central Controller call routers with state information for agents, calls, queues, and so forth. This traffic is real-time traffic. The PGs also send up historical data at intervals based on their configuration. The historical data is low priority, but it must reach the central site before the start of the next interval.

When a PG starts, the central site supplies its configuration data so that it knows which resources to monitor. This configuration download can cause a significant spike in network bandwidth usage.

The public traffic can be summarized as follows:

- **High-priority traffic**—Includes routing and Device Management Protocol (DMP) control traffic. It is sent in TCP with the public high-priority IP address.
- **Heartbeat traffic**—UDP messages with the public high-priority IP address and in the port range of 39500 to 39999. Heartbeats are transmitted at 400-ms intervals in both directions between the PG and the Central Controller.
- **Medium-priority traffic**—Includes real-time traffic and configuration requests from the PG to the Central Controller. The medium-priority traffic is sent in TCP with the public high-priority IP address.
- **Low-priority traffic**—Includes historical data traffic, configuration traffic from the Central Controller, and call close notifications. The low-priority traffic is sent in TCP with the public non-high-priority IP address.

Private Network Traffic Flow

The private network carries critical Message Delivery Service (MDS) traffic.

The private traffic can be summarized as follows:

- **High-priority traffic**—Includes routing, MDS control traffic, and other traffic from MDS client processes such as the PIM CTI Server, Logger, and so forth. It is sent in TCP with the private high-priority IP address.
- **Heartbeat traffic**—UDP messages with the private high-priority IP address and in the port range of 39500 to 39999. Heartbeats are transmitted at 100-ms intervals bi-directionally between the duplexed sides.
- **Medium-priority and low-priority traffic**—For the Central Controller, this traffic includes shared data sourced from routing clients as well as (non-route control) call router messages, including call router state transfer (independent session). For the OPC (PG), this traffic includes shared non-route control peripheral and reporting traffic. This class of traffic is sent in TCP sessions designated as medium priority and low priority, respectively, with the private non-high priority IP address.
- **State transfer traffic**—State synchronization messages for the Router, OPC, and other synchronized processes. It is sent in TCP with a private non-high-priority IP address.

Merged Network Connections

Unified CCE components use a public network and a private network to communicate. These networks must be separate physical networks. For high availability, include redundant connections in your public network. Ideally, each connection uses a different carrier.

If QoS and bandwidth are configured correctly, your design can merge a public or private WAN link with other corporate traffic. If you use a link that merges non-contact-center traffic, keep the public and private traffic on different networks. However, never split private network traffic onto low-priority and high-priority data paths. The same link must carry all private network traffic for a given component. Sending low-priority and high-priority traffic on different links disables the component failover behavior. Similarly, all low- and high-priority traffic from each peripheral gateway to the low- and high-priority addresses of the call router must take the same path.

During a public network failure, you can temporarily fail over the public Unified CM traffic to the private network. Size the private network to accommodate the extra traffic. When the public traffic fails over to the private network, restore the public network as quickly as possible to return to usual operation. If the private network also fails, your contact center can experience instability and data loss, including the corruption of one Logger database.

IP-Based Prioritization and Quality of Service

Contact center enterprise solutions require QoS on all private networks. On public links, you can use QoS in the 2000 Agent and 4000 Agent Reference Designs. For public links in a 12,000 Agent Reference Design, QoS can delay the detection of server failures.

If large amounts of low-priority traffic get in front of high-priority traffic, the delay can trigger the fault tolerance behavior. To avoid these delays, you need a prioritization scheme for each of the WAN links in the public and private networks. Contact center enterprise solutions support IP-based prioritization and QoS.

In a slow network flow, the time a single large (for example, 1500-byte) packet consumes on the network can exceed 100 ms. This delay would cause the apparent loss of one or more heartbeats. To avoid this situation, the contact center uses a smaller Maximum Transmission Unit (MTU) size for low-priority traffic. This allows a high-priority packet to get on the wire sooner. (MTU size for a circuit is calculated based on the circuit bandwidth, as configured at PG setup.)

An incorrectly prioritized network generally leads to call time-outs and loss of heartbeats. The problems increase as the application load increases or when shared traffic is placed on the network. You can also see application buffer pool exhaustion on the sending side, due to extreme latency conditions.

Contact center enterprise solutions use three priorities: high, medium, and low. Without QoS, the network recognizes only two priorities identified by source and destination IP address (high-priority traffic sent to a separate IP destination address) and, for UDP heartbeats, by a specific UDP port range. IP-based prioritization configures IP routers with priority queuing to give preference to TCP packets with a high-priority IP address and to UDP heartbeats over the other traffic. When using this prioritization scheme, 90% of the total available bandwidth is granted to the high-priority queue.

A QoS-enabled network applies prioritized processing (queuing, scheduling, and policing) to packets based on QoS markings as opposed to IP addresses. The contact center provides a marking capability of Layer-3 DSCP for private and public network traffic. Traffic marking implies that configuring dual IP addresses on each Network Interface Controller (NIC) is no longer necessary because the network is QoS-aware. However, if you mark the traffic at the network edge instead, you still require dual-IP configuration to differentiate packets by using access control lists based on IP addresses.

Note Layer-2 802.1p marking is also possible if Microsoft Windows Packet Scheduler is enabled (for PG/Central Controller traffic only). However, this is not supported. Microsoft Windows Packet Scheduler is not suited to Unified CCE. 802.1p markings are not widely used, nor are they required when DSCP markings are available.

For more information about proper network design for data traffic, see the network infrastructure and Quality of Service (QoS) documentation at http://www.cisco.com/c/en/us/solutions/enterprise/design-zone-borderless-networks/index.html.

UDP Heartbeat and TCP Keep-Alive

The UDP heartbeat design detects if a public network link has failed. Detection can be made from either end of the connection, based on the direction of heartbeat loss. Both ends of a connection send heartbeats at periodic intervals (every 400 milliseconds) to the opposite end. Each end looks for analogous heartbeats from the other. If either end does not receive a heartbeat after five times the heartbeat period, that end assumes that something is wrong and the application closes the socket connection. At that point, a TCP Reset message is typically generated from the closing side. Various factors can cause loss of heartbeats, such as:

- The network failed.
- The process sending the heartbeats failed.
- The VM with the sending process is shut down.
- The UDP packets are not properly prioritized.

There are several parameters associated with heartbeats. In general, leave these parameters set to their system default values. Some of these values are specified when a connection is established. Other parameters can be set in the Windows registry. The two values of most interest are:

- The amount of time between heartbeats
- The number of missed heartbeats (currently hard-coded to five) that indicate a failure

The default value for the private heartbeat interval between redundant components is 100 milliseconds. One side can detect the failure of the circuit or the other side after 500 ms. The default heartbeat interval between a central site and a peripheral gateway is 400 ms. In this case, it takes 2 seconds to reach the circuit failure threshold.

The contact center enterprise QoS implementation uses a TCP keep-alive message to replace the UDP heartbeat. The public network interface enforces a consistent heartbeat or keep-alive mechanism. But, the private network interface enforces the keep-alive. When QoS is enabled on the public network interface, a TCP keep-alive message is sent; otherwise UDP heartbeats are retained.

The TCP keep-alive feature, provided in the TCP stack, detects inactivity and then causes the server or client side to terminate. The TCP keep-alive feature sends probe packets across a connection after the connection is idle for a certain period. The connection is considered down if a keep-alive response from the other side is not heard. On a Windows server, you can specify keep-alive parameters on a per-connection basis. For contact center enterprise public connections, the keep-alive timeout is set to (5 * 400) ms, matching the failure detection time of 2 seconds with the UDP heartbeat.

Our reasons for moving to TCP keep-alive with QoS enabled are:

- In a converged network, router algorithms to handle network congestion conditions can have different effects on TCP and UDP. As a result, delays and congestion experienced by UDP heartbeat traffic can result in connection failures from timeouts.
- The use of UDP heartbeats creates deployment complexities in a firewall environment. With the dynamic port allocation for heartbeat communications, you open a large range of port numbers which weakens the security of your firewall.

Note You cannot use WAN accelerators on a WAN that carries contact center traffic. WAN accelerators can send signals that effectively disable the failure detection function.

HSRP-Enabled Networks

If your solution network uses the Hot Standby Router Protocol (HSRP) on the default gateways, follow these requirements:

- Set the HSRP hold time and its associated processing delay lower than five times the heartbeat interval (100 ms on the private network and 400 ms on the public network). This level avoids private network communication outage during the switch-over of the HSRP active router.

  With convergence delays that exceed private or public network outage notification, HSRP failover times can exceed the detection threshold and result in a failover. If the HSRP configuration has primary and secondary designations and the primary path router fails over, HSRP reinstates the primary path when possible. That reinstatement can lead to a second outage detection.

  Do not use primary and secondary designations with HSRP convergence delays near 500 ms for the private network and 2 seconds for the public network. However, convergence delays below the detected threshold (which result in HSRP failovers that are transparent) do not mandate a preferred path configuration. This approach is preferable. Keep enabled routers symmetrical if path values and costs are identical. However, if available bandwidth and cost favor one path (and the path transition is transparent), then designation of a primary path and router is advised.
- Our fault-tolerant design requires that the private network is physically separate from the public network. Therefore, do not configure HSRP to fail over one type of network traffic to the other network link.
- The bandwidth requirement for the contact center must always be guaranteed with HSRP, otherwise the system behavior is unpredictable. For example, if you configure HSRP for load sharing, ensure that sufficient bandwidth remains on the surviving links in the worst-case failure situations.

Unified CCE Failovers During Network Failures

Network failures simultaneously affect any components that send traffic across the affected network. Unified CCE subcomponents use both private and public network links to communicate.

The traffic on the private network performs these functions:

- State transfer during component startup
- Synchronization of redundant pairs of Routers
- Synchronization of redundant Logger databases

- Synchronization of redundant pairs of PGs

The public network carries the rest of the traffic between the subcomponents: voice data, call context data, and reporting data. The public network includes all the public network links between the Unified CCE subcomponents.

Note In virtualized contact centers, network failures can arise from failures in the virtual environment, like a virtual NIC, or from failures of physical resources.

Response to Private Network Failures

When a private network fails, the contact center quickly enters a state where one or both sides transition into isolated-enabled operation. The isolated operation continues until the Routers detect the restoration of the private network. The redundant pairs of Routers and PGs then resynchronize and resume usual operation.

Assume that Router A is the pair-enabled side and Router B is the pair-disabled side. When the private network fails, the Router A behaves as follows:

- If Router A has device majority, it transitions to the isolated-enabled state and continues handling traffic.
- If Router A does not have device majority, it transitions to the isolated-disabled state and stops processing traffic.

When the private network fails, Router B behaves as follows:

- If Router B does not have device majority, it transitions to the isolated-disabled state and does not process traffic.
- If Router B does have device majority, it enters a test state. Router B instructs its enabled PGs to contact Router A over the public network. Then, Router B responds as follows:
  - If no PG can contact Router A to determine its state, Router B transitions to the isolated-enabled state and begins handling traffic. This case can result in both Router A and Router B running in isolated-enabled state.
  - If any PG contacts Router A and finds it in the isolated-disabled state, Router B transitions to the isolated-enabled state and begins handling traffic.
  - If any PG contacts Router A and finds it in the isolated-enabled state, Router B transitions to the isolated-disabled state and does not process traffic.

During the Router failover processing, any Route Requests for the Router are queued until the surviving Router is in isolated-enabled state. A Router failure does not affect any in-progress calls that have already reached a VRU or an agent.

The corresponding Logger shuts down when its Router goes idle. Each Logger communicates only with its own Router. If the private network connection is restored, the isolated-enabled Router's Logger uses its data to resynchronize the other Logger. The system automatically resynchronizes the Logger configuration database if the private network connection is brought back up before the 14 day retention period of the Config_Message_Log table. And, if the private network connection remains down for more than the 14 day retention period, you must resynchronize the configuration data on Loggers using the Unified ICMDBA application as described in the *Administration Guide*, at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-maintenance-guides-list.html.

In each redundant pair of PGs, there is also an enabled PG and a disabled PG. At system start, the first PG to connect becomes the enabled PG. However, after a private network failure, the PG with the greatest weight in the redundant pair becomes the enabled PG. The other PG becomes the disabled PG.

Response to Public Network Failures

Highly available networks generally include redundant channels for the public network. When one channel fails, the other channel takes over seamlessly. The contact center detects a public network failure only when all channels fail between two subcomponents.

Note In contact centers without redundant public networks, the contact center detects a failure when the single channel fails.

How the contact center responds to a public network failure depends on number and function of the sites and how the sites are linked. The following sections look at some of the more common or significant scenarios.

Failures between Unified Communication Managers

The scenario that can cause the most problems involves the Unified CM subscribers losing their public link. Because the functioning private network keeps the Routers and Agent PGs in synch, the Routers can still detect all agent devices. In this situation, a Router can assign a call to an agent device that is registered on the subscriber on the other side of the public network failure. However, the local CVP cannot pass the connection information to the agent device on the other side of the public network failure. The call fails, but the routing client marks the call as routed to the agent device on the remote subscriber.

Failures in Clustering over the WAN

Failures between Sites

In the clustering over the WAN topology, you need a highly available, highly resilient WAN with low latency and sufficient bandwidth. The public network is a critical part of the contact center's fault tolerance. A highly available WAN is fully redundant with no single points of failure, usually across separate carriers. During a partial failure of the WAN, the redundant link needs the capability to handle the full load for the sites within the QoS parameters. As an alternative to redundant WANs, you can employ Metro Area Networks (MAN), Dense Wavelength Division Multiplexing (DWDM), or Ethernet WANs. For more information about designing highly available, highly resilient WANs, see *Design Zone for Branch WAN* at https://www.cisco.com/c/en/us/solutions/enterprise/design-zone-branch-wan/index.html in the Cisco Design Zone.

Note You cannot use a Wireless WAN, like WiMAX, Munincipal Wi-Fi, or VSAT, for you contact center enterprise solution.

If the public network fails between the sites, the system responds in this manner:

1. The Unified CM subscribers detect the failure. The subscribers continue to function locally with no impact to local call processing and call control. However, any calls that were set up over the public network fail.
2. The Routers and PGs detect the failure. The PGs automatically realign their data communication stream to their local Router. The local Router then passes data to the Router on the other side over the private

network to continue call processing. The altered data path does not cause a failover of the PG or the Router.

The impact of the public network failure on agents depends on where their phones and desktops registered:

- The most common case is that the agent desktop and agent phone are both registered to the PG and a subscriber on the same side (Side A for example). When the public link between the sites fails, the agent can continue handling calls usually.
- In some cases, the agent desktop (Side A for this example) and the agent phone (Side B for this example) can end up registered on different sides. In those cases, the CTI Manager directs phone events over the public network to the PG on the opposite side. When the public network between the sites fails, the phone does not rehome to Side A of the cluster. The phone remains operational on Side B. The PG on Side A cannot detect this phone. Because the Unified CCE subcomponents can no longer direct calls to the agent phone, Unified CCE automatically signs out the agent.
- Usually, the redundant desktop server pair load balances agent desktop connections. So, half of the desktops register on a desktop server that connects to a PG with an active CTI Server across the public network. When the public network fails, the desktop server loses connection with the remote CTI Server. The desktop server disconnects the active agent desktops to force them to rehome to the redundant desktop server at the remote site. The agent desktop automatically uses the redundant desktop server. The agent desktop remains disabled until it connects to the redundant desktop server.

Failures to Agent Sites

The contact center enterprise topology for clustering over the WAN assumes that the agents are remotely located at multiple sites. Each agent site requires access to both sites through the public network for redundancy. In a complete network failure, these connections also provide basic SRST functionality, so that the agent site can still make emergency (911) calls.

If the agent site loses the public network connection to one of the sites, the system responds in this manner:

1. Any IP phones that are homed to the Unified CM subscribers at the disconnected site automatically rehome to subscribers at the othersite. To use the rehoming behavior, configure a redundancy group.
2. Agent desktops that are connected to the desktop server at that disconnected site automatically realign to the redundant server at the other site. (Agent desktops are disabled during the realignment process.)

If the agent site loses the public network connection to both of the sites, the system responds in this manner:

1. The local Voice Gateway (VG) detects the failure of the communications path to the cluster. The VG then goes into SRST mode to provide local dial-tone functionality.
2. With Unified CVP, the VGs detect the loss of connection to the Unified CVP Server. Then, the VGs run their local survivability TCL script to reroute the inbound calls.
3. If an active call came in to the disconnected agent site on a local PSTN connection, the call remains active. But, the PG loses access to the call and creates a TCD record.
4. The Finesse server detects the loss of connectivity to the agent desktop and automatically signs the agent out of the system. While the IP phones are in SRST mode, they cannot function as contact center enterprise agents.

Response to Failures of Both Networks

Individually, parts of the public and private networks can fail with limited impact to the agents and calls. However, if both of these networks fail at the same time, the system retains only limited functionality. This failure is considered catastrophic. You can avoid many such failures by careful WAN design with built-in backup and resiliency.

A simultaneous failure of both networks within a site shuts down the site.

If both the public and private networks simultaneously fail between two sites, the system responds in this manner:

1. Both Routers check for device majority. Each router enters isolated-enabled mode if the router has device majority or isolated-disabled mode if the router does not have device majority.
2. The PGs automatically realign their data communications, if necessary, to their local Router. A PG that cannot connect to an active Router becomes inactive.
3. The Unified CM subscribers detect the failure and continue to function locally with no impact to local call processing and call control.
4. Any in-progress calls that are sending active voice path media over the public WAN link fail with the link. When the call fails, the PG creates a TCD record for the call.
5. In a clustering over the WAN topology, the Unified CM subscribers on each side operate with access only to local components.
6. The call routing scripts automatically route around the off-line devices using peripheral-on-line status checks.
7. Agents with both their phones and desktops registered with local Unified CM subscribers are not affected. All other agents lose some or all functionality while their phones and desktops rehome. Those agents might also find themselves signed out, depending on the exact system configuration.
8. Unified CCE does not route new calls that come into the disabled side. But, you can redirect or handle those calls with the standard Unified CM redirect on failure for their CTI route points or with the Unified CVP survivability TCL script in the ingress Voice Gateways.

Ingress, Egress, and VXML Gateway High Availability Considerations

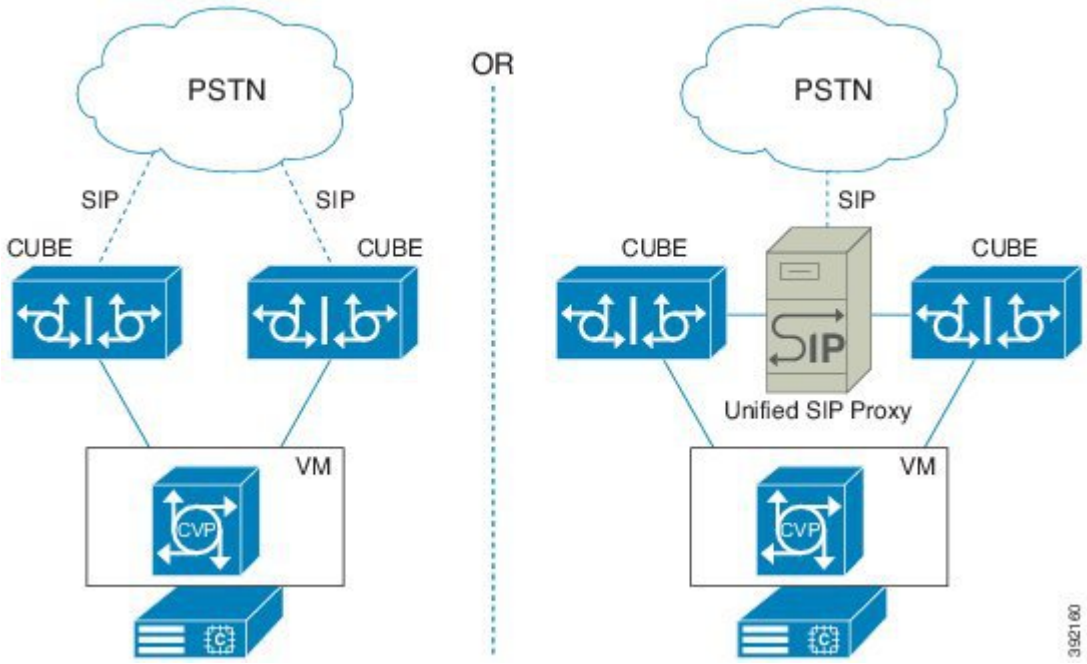
Highly available contact center designs start with the network infrastructure for data, multimedia, and voice traffic. A "single point of failure" in your network infrastructure devalues any other high availability features that you design into the contact center. Begin from the PSTN and ensure that incoming calls have multiple paths for reaching Unified CVP for initial treatment and queuing.

Ideally, design with at least two SIP trunks each connecting to a separate Cisco Unified Border Element (CUBE). If any CUBE or SIP trunk fails, the PSTN can route all traffic through the remaining SIP trunks. The PSTN route either by configuring all the SIP trunks as a large trunk group or by configuring rerouting or overflow routing to the other SIP trunks. You can also connect a redundant CUBE to each SIP trunk to preserve capacity when a Cisco UBE fails and the SIP trunk is still functional.

In some areas, the PSTN does not provide multiple SIP trunks to a single site. In that case, you can connect the SIP trunk to a Cisco Unified SIP Proxy (CUSP). Then, you could connect multiple CUBEs to the CUSP to provide some redundancy.

The CUBE passes calls to Unified CVP for initial treatment and queuing. Register each CUBE with a separate Unified CVP for load balancing. For further fault tolerance, you can register each CUBE with a different Unified CVP as a backup. If a CUBE cannot connect with a Unified CVP, you can also use TCL scripts to provide some call processing. A TCL script can reroute the calls to another site or dialed number. The script can also play a locally stored .wav file to the caller and end the call.

Figure 85: High Availability Ingress Points



For more information about CUBE, Unified CVP, and voice networks in general, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/en/US/docs/voice_ip_comm/uc_system/design/guides/UCgoList.html.

Cisco Unified Survivable Remote Site Telephony (SRST)

Voice gateways using the Cisco Unified Survivable Remote Site Telephony (SRST) option for Unified CM follow a similar failover process. If the gateway is cut off from its controlling subscriber, the gateway fails over into SRST mode. The failover drops all voice calls and resets the gateway into SRST mode. Phones rehome to the local SRST gateway for local call control.

While running in SRST mode, Unified CCE operates as if the agents have no CTI connection from their desktops. The routing application detects the agents as not ready and sends no calls to these agents. When the gateway and subscriber reestablish their connection, the subscriber takes control of the gateway and phones again, allowing the agents to reconnect.

High Availability for Ingress and Egress Gateways

The ingress gateway accepts calls from the PSTN and directs them to Unified CVP for VRU treatment and call routing. The same gateway can act as an egress gateway in certain call flows.

Note The ingress gateway is sometimes called the originating gateway.

In the contact center enterprise Reference Designs, the ingress gateway uses SIP to communicate with Unified CVP. The SIP protocol does not have built-in redundancy features. SIP relies on the gateways and call processing components for redundancy. You can use the following techniques to make call signalling independent of the physical interfaces. If one interface fails, the other interface handles the traffic.

Dial-Peer Binding

With dial-peer level bind, you set up a different bind for each dial peer. You do not need to have a single interface that is reachable from all subnets. The dial peer helps to segregate the traffic from different networks (for example, the SIP trunk from service provider and the SIP trunk to Unified CM or CVP). This example shows a dial peer level binding:

```
Using voice-class sip bind
 dial-peer voice 1 voip
 voice-class sip bind control source-interface GigabitEthernet0/0
```

Global Binding

For other gateways, you can use global binding. Connect each gateway interface to a different physical switch to provide redundancy. Each gateway interface has an IP address on a different subnet. The IP routers use redundant routes to the loopback address, either by static routes or by a routing protocol.

You can use a routing protocol to review the number of routes that are exchanged with the gateway. In that case, use filters to limit the routing updates. Have the gateway only advertising the loopback address and not advertise the receiving routes. Bind the SIP signaling to the virtual loopback interface, as shown in this example:

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

Call Survivability During Failovers

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

- **Calls in progress**— The PSTN switch loses the D-channel to all T1/E1 trunks on this gateway. The active calls cannot be preserved.
- **Incoming calls**— The PSTN carrier directs the calls to a T1/E1 at an alternate gateway. The PSTN switch has to have its trunks and dial plan properly configured.

High Availability for VXML Gateways

The VXML Gateway parses and renders VXML documents from the Unified CVP VXML Servers or an external VXML source. Rendering a VXML document can include the following:

- Retrieving and playing prerecorded audio files
- Collecting and processing user input
- Connecting to an ASR/TTS Server for voice recognition and dynamic text-to-speech conversion.

You cannot have a load balancer on the path between the VXML Gateway and the Unified CVP Call Server.

In topologies that separate the ingress gateway from the Unified CVP Call Server, colocate the VXML Gateway at the Call Server site. This arrangement keeps the media stream from using bandwidth across the WAN.

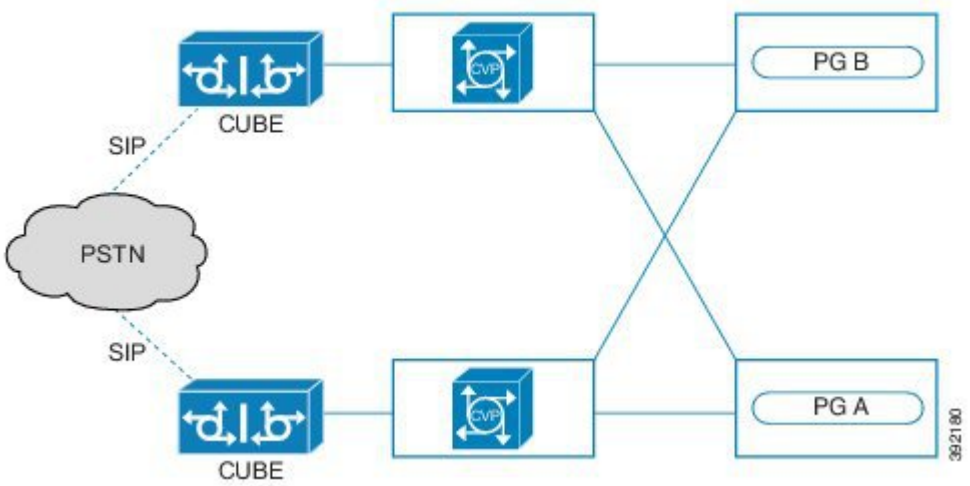
If the VXML Gateway fails, calls are affected as follows:

- **Calls in progress**—The ingress gateway's survivability features route calls in progress to an alternate location by default.
- **Incoming calls**—Incoming calls find an alternate VXML gateway.

CVP High Availability Considerations

The Contact Center Enterprise Reference Designs use Unified CVP for the call treatment and queuing. CVP uses SIP for the call control, rather than relying on Unified CM for JTAPI call control.

Figure 86: Unified CVP High Availability Deployment



Unified CVP can use the following system components:

- Cisco Unified Border Element (CUBE) supports the transition to SIP trunking. CUBE provides interworking, demarcation, and security services between the PSTN and your contact center.
- Cisco Voice Gateway (VG) terminates TDM PSTN trunks to transform them into IP-based calls on an IP network. Unified CVP uses specific Cisco IOS Voice Gateways that support SIP to enable more flexible call control. VGs controlled by Unified CVP can also use the Cisco IOS built-in Voice Extensible Markup Language (VXML) Browser to provide the caller treatment and call queuing. CVP can also leverage the Media Resource Control Protocol (MRCP) interface of the Cisco IOS VG to add automatic speech recognition (ASR) and text-to-speech (TTS) functions.

- The CVP Server provides call control signaling when calls are switched between the ingress gateway and another endpoint gateway or a Unified CCE agent. The CVP Server also provides the interface to the Unified CCE VRU Peripheral Gateway (PG). The CVP Server translates specific Unified CCE VRU commands into VXML code for rendering on the VG. The CVP Server can communicate with the gateways using SIP as part of the solution. For high availability discussions, you can view the CVP Server as these subcomponents:
  - **SIP Service**—Responsible for all incoming and outgoing SIP messaging and SIP routing.

Note You can configure the Call Server to use a SIP Proxy Server for outbound dial plan resolution. SIP Proxy Server minimizes the configuration overhead.

You can also configure it to use static routes based on an IP address or DNS SRV. Call Servers do not share configuration information about static routes. When you change a static route, you must change it on each Call Server's SIP Service.

  - **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 CVP Media Server acts as a web server that provides predefined audio files to the voice browsers as part of their VXML processing. You can cluster media servers using the Cisco Content Services Switch (CSS) products. With clustering, you can pool multiple media servers behind a single URL for access by all the voice browsers.
- The CVP Server hosts the Unified CVP VXML runtime environment. The VXML service creation environment uses an Eclipse toolkit browser in the CVP Call Studio application. The runtime environment runs the dynamic VXML applications and processes Java and Web Services calls for external systems and database access.
- Cisco Unified SIP Proxy (CUSP) servers that are used with CVP can select voice browsers and associate them with specific dialed numbers. When a call comes into the network, the VG queries the Unified SIP Proxy to determine where to send the call based on the dialed number.

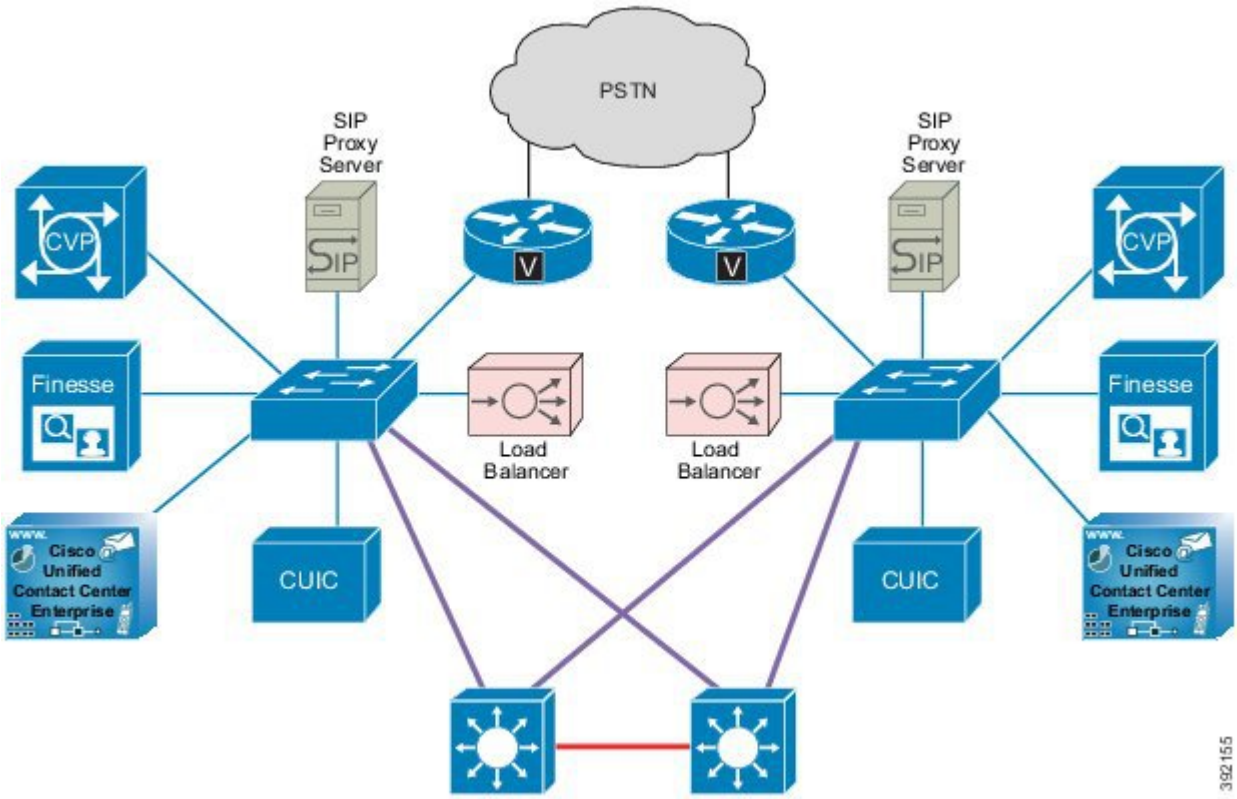
Important Contact center enterprise solutions do not support Unified CM's intercluster Enhanced Location Call Admission Control (ELCAC) feature.

These methods can increase the high availability of CVP:

- To provide the automatic call balancing across the CVP Servers, add redundant CVP Servers under control of the Unified CCE PGs.
- To handle conditions where the gateway cannot contact the CVP Server, add survivability TCL scripts to the gateway. For example, you can redirect calls to another CVP Server on another CVP-controlled gateway.
- To load balance the audio file requests across multiple CVP Media Servers and VXML URL access across multiple servers, add a Cisco Content Server.

This figure 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.

Figure 87: Redundant Unified CVP System



The two switches shown in the figure provide the network redundancy for the Unified CVP Servers. If one switch fails, only a subset of the components becomes inaccessible. The components that are connected to the remaining switch are still accessible for the call processing.

High Availability Factors to Balance

You can make your contact center enterprise solution more highly available by adding the following components and subcomponents for CVP:

- **Multiple gateways, Unified CVP Servers, Unified CVP VXML Servers, and VRU PGs**—Enables inbound and outbound call processing and VRU services to continue during individual component failures.
- **Unified CVP Media Servers**—The VVoice Browser sends requests to the backup media server if the primary media server is unreachable. Ensure that Whisper Announcement and Agent Greeting audio files are duplicated on all media servers for proper failover behavior.
- **Multiple call processing locations**—Enables call processing to continue if a call processing location goes dark.
- **Redundant WAN links**—Enables Unified CVP call processing to occur if individual WAN links fail.

Call Survivability During Failovers

The following sections describe how the failure of contact center enterprise components and CVP subcomponents affect the call survivability.

Voice Browser

The Voice Browser parses and renders VXML documents obtained from one or several sources. If the VXML gateway fails, the following happens:

- **Calls in progress**—The ingress gateway's survivability features route calls in progress to an alternate location by default.
- **Incoming calls**—Incoming calls find an alternate VXML gateway.

Unified CVP IVR Service—The CVP IVR Service creates the VXML pages that implement the Unified CVP Micro applications. The micro applications are based on RunExternalScript instructions that are received from Unified CCE. If the IVR Service fails, the following happens:

- **Calls in progress**—Calls in progress are routed by default to an alternate location by survivability on the originating gateway.
- **Incoming calls**—Incoming calls are directed to an in-service IVR Service.

Unified CM

The CVP Call Server recognizes when the Unified CM fails and the following happens:

- **Calls in progress**—The server assumes that it should preserve the active calls, and maintains the signaling channel to the originating gateway. The originating gateway is not aware that Unified CM has failed. More activities in the active calls (such as hold, transfer, or conference) are not possible. After the call ends, the phone routes to another Unified CM server.
- **Incoming calls**—Incoming calls are directed to an alternate Unified CM server in the cluster.

CVP Call Server

The CVP Call Server contains the following services which handle call survivability during failovers.

- **Unified CVP SIP Service**—The CVP SIP Service handles all incoming and outgoing SIP messaging and SIP routing. If the SIP Service fails, the following happens:
  - **Calls in progress**—If the CVP SIP Service fails after the caller is transferred (including transfers to an IP phone or Voice Browser), then the call continues usually. But, the CVP SIP Service cannot transfer that call again. If the failure happens before the caller is transferred, then the default survivability routing transfers the call to an alternate location.
  - **Incoming calls**—Unified SIP Proxy directs incoming calls to an alternate Unified CVP Call Server. If no Call Servers are available, the call is default-routed to an alternate location by survivability.

CVP Media Server

Store the audio files locally in flash memory on the VXML gateway or on an HTTP or TFTP file server. Audio files stored locally are highly available. However, HTTP or TFTP file servers provide the advantage of centralized administration of audio files.

If the media server fails, the following happens:

- **Calls in progress**—Calls in progress recover automatically. The high-availability configuration techniques make the failure transparent to the caller. If the media request fails, use scripting techniques to work around the error.
- **Incoming calls**—Incoming calls are directed transparently to the backup media server, and service is not affected.

Note You can locate the Media Server across a WAN from the VXML Gateway. If the WAN connection fails, the gateway continues to use prompts from the gateway cache until the requested prompt expires. The gateway then attempts to reacquire the media, and the call fails if survivability is not enabled. If survivability is enabled, the calls are default-routed.

CVP VXML Server

The Unified CVP VXML Server runs advanced VRU applications by exchanging VXML pages with the Voice Browser. If the CVP VXML Server fails, the following happens:

- **Calls in progress**—You can recover calls in progress in a Unified CCE-integrated deployment with scripting techniques. For example, configure the script to first connect to Unified CVP VXML Server A. If the application fails out the X-path of the Unified CVP VXML Server ICM script node, try Unified CVP VXML Server B.
- **Incoming calls**—Incoming calls are directed transparently to an alternate CVP VXML Server.

CVP Reporting Server

Failure of a CVP Reporting Server has no impact on call survivability.

The Reporting Server does not perform any database administrative and maintenance activities such as backups or purges. However, the Unified CVP provides access to such maintenance tasks through the Operations Console. The single CVP Reporting Server does not necessarily represent a single point of failure. Data safety and security are provided by the database management system. Temporary outages are tolerated due to persistent buffering of information on the source components.

More Call Survivability Points

Consider the following points when you plan for call survivability in your solution:

- There are scenarios in which call recovery is not possible during a failure:
  - Someone stops the process with calls in progress. For example, a system administrator forgets to do a Call Server graceful shutdown. In this case, the CVP Call Server terminates all active calls to release the licenses.

- The Call Server exceeds the recommended call rate. There is a limit for the number of calls allowed in the Call Server. But, there is no enforced limit for the call rate. In general, exceeding the recommended calls per second (CPS) for a long period can cause erratic and unpredictable call behavior. Size your solution correctly and balance the call load appropriately across each call processing component.

- Configure the originating gateways for call survivability as described in the *Configuration Guide for Cisco Unified Customer Voice Portal* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-installation-and-configuration-guides-list.html. The `survivability.tcl` script also contains some directions and useful information.
- You can detect calls that are 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 120 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 can perform session refreshing on their own. RFC 4028 (Session Timers in the Session Initiation Protocol) contains more details on the usage of Session expires with SIP calls.
- During failovers, calls under Unified CVP control get treatment from the survivability TCL script in their ingress Voice Gateways. In these cases, the routing dialog in the Unified CCE Central Controller stops. If the survivability scripts redirect the calls to another active Unified CCE component, the call appears as a "new call" to the system with no relationship to the original call for reporting or tracking purposes.

SIP Proxy Servers with CVP

The SIP Proxy Server provides the dial plan resolution for the SIP endpoints. You can configure the dial plan information in a central place, instead of statically on each SIP device. You do not need a SIP Proxy Server in your solution. Consider one for the centralized configuration and maintenance benefits. By deploying multiple SIP Proxy Servers, you can achieve load balancing, redundancy, and regional SIP call routing services. Your solution has the following choices for SIP call routing.

SIP Proxy Server

SIP Proxy Servers provide these advantages:

- Weighted load balancing and redundancy.
- Centralized dial-plan configuration.
- If you already have a SIP proxy or one is used by other applications for dial-plan resolution or intercluster call routing, you might leverage existing assets.

However, you might require another server for the SIP Proxy Server.

Static Routes Using Server Groups (DNS SRV Records) on a DNS Server

You can achieve weighted load balancing and redundancy with this kind of static routing.

However, you might find these disadvantages with this method:

- Ability to use an existing server depends on the location of the DNS server.
- Some organizations limit the ability to share or delegate DNS server administration rights.
- You must configure dial plans on each device individually (Unified CM, Unified CVP, and gateways).
- Unified CVP performs a DNS SRV lookup for every call. Performance is an issue if the DNS server is slow to respond, is unavailable, or is across the WAN.

Static Routes Using Local DNS SRV Records

You can achieve these advantages with this type of static routing:

- Weighted load balancing and redundancy.
- Eliminates concerns over latency, DNS Server performance, and a point of failure by not depending on an external DNS Server.

However, you must configure dial plans on each device individually (Unified CM, Unified CVP, and gateways).

Note Static routes using SRV with a DNS Server, or using Server Groups, can cause unexpected, long delays during failover and load balancing. This happens with TCP or UDP transport on the Unified CVP Call Server when the primary destination is shut down or is off the network. With UDP, when a hostname has elements with different priorities in the Server Group (srv.xml), Unified CVP tries twice for each element, with a 500-msec delay. The delay is on every call during failure, depending on load balancing, and is in accordance with section 17.1.1.1 of RFC 3261 regarding the T1 timer. If server group heartbeats are turned on, then the delay may only be incurred once, or not at all, depending on the status of the element.

Cisco Unified SIP Proxy Support

Cisco Unified SIP Proxy (CUSP) is our implementation of a SIP Proxy Server. CUSP is a dedicated SIP Proxy Server that runs on the gateway or on a virtual machine.

CUSP Deployment Options

These sections describe your options for deploying CUSP in your contact center enterprise solution.

Redundant SIP Proxy Servers

In this option, you have two gateways and each has one proxy VM. The gateways are geographically separated for redundancy. They use SRV priority for redundancy of proxies and do not use HSRP.

Note these points when you select this option:

- CUSP can coreside with VXML or TDM Gateways.
- You can configure TDM Gateways with SRV or with Dial Peer Preferences to use the primary and secondary CUSP proxies.
- CUSP is set with Server Groups to find the primary and back up Unified CVP, Unified CM, and Voice Browsers.

- Unified CVP is set up with a Server Group to use the primary and secondary CUSP proxies.
- Unified CM is set up with a Route Group with multiple SIP Trunks to use the primary and secondary CUSP proxies.

In this example, ISR1 is on the east coast and ISR2 is on the west coast. The TDM Gateways use the closest ISR, and only cross the WAN when failing over to the secondary priority blades.

The SRV records look like this:

```
east-coast.proxy.atmycompany.com
blade 10.10.10.10 priority 1 weight 10 (this blade is in ISR1 on east coast)
blade 10.10.10.20 priority 2 weight 10 (this blade is in ISR2 on west coast)

west-coast.proxy.atmycompany.com
blade 10.10.10.20 priority 1 weight 10 (this blade is in ISR2 on west coast)
blade 10.10.10.10 priority 2 weight 10 (this blade is in ISR1 on east coast)
```

Double Capacity Redundant SIP Proxy Servers

In this option, you have two gateways and each has two proxy VMs. All four proxy servers are in active mode with calls being balanced between them. The gateways are geographically separated for redundancy. They use SRV priority to load balance across proxies with priority.

Note these points when you select this option:

- Due to platform validation restrictions on CUSP, the ISR is dedicated to the proxy blade function. The ISR is not collocated as a Voice Browser nor as a TDM Gateway.
- You can configure TDM Gateways with SRV or with Dial Peer Preferences to use the primary and secondary CUSP proxies.
- CUSP is set with Server Groups to find the primary and back up Unified CVP, Unified CM, and Voice Browsers.
- Unified CVP is set up with a Server Group to use the primary and secondary CUSP proxies.
- Unified CM is set up with a Route Group with multiple SIP Trunks to use the primary and secondary CUSP proxies.

In this example, ISR1 is on the east coast and ISR2 is on the west coast. The TDM Gateways use the closest ISR, and only cross the WAN when failing over to the secondary priority blades.

The SRV records look like this:

```
east-coast.proxy.atmycompany.com
blade 10.10.10.10 priority 1 weight 10 (this blade is in ISR1 on east coast)
blade 10.10.10.20 priority 1 weight 10 (this blade is in ISR1 on east coast)
blade 10.10.10.30 priority 2 weight 10 (this blade is in ISR2 on west coast)
blade 10.10.10.40 priority 2 weight 10 (this blade is in ISR2 on west coast)

west-coast.proxy.atmycompany.com
blade 10.10.10.30 priority 1 weight 10 (this blade is in ISR2 on west coast)
blade 10.10.10.40 priority 1 weight 10 (this blade is in ISR2 on west coast)
blade 10.10.10.10 priority 2 weight 10 (this blade is in ISR1 on east coast)
blade 10.10.10.20 priority 2 weight 10 (this blade is in ISR1 on east coast)
```

CUSP Design for High Availability

The following points affect high availability for CUSP:

- **Do not use Proxy Server Record Route**—This option impacts the performance of the proxy server and creates a "single point of failure." Do not turn on this option.

  When the RecordRoute header is not populated, the signaling bypasses CUSP once the inbound call reaches the Unified CVP Call Server. From that point in the routing, the signaling runs directly from the originating device to the CVP Call Server.

- **Upstream Element Routing with SIP Heartbeats**—CUSP treats any response to an INVITE or OPTIONS as a good response. So, CUSP does not mark an element as down when it receives a response. If the response is configured in the failover response code list for the server group, then CUSP fails over to the next element in the group. Otherwise, CUSP sends the response downstream as the final response.

Server Groups and CVP High Availability

A Server Group is a dynamic routing feature. Through a Server Group, the originating endpoint can check the status of the destination address before sending the SIP INVITE. A heartbeat method tells the originating SIP user about the status of the destination. This feature allows faster failover on call control by eliminating delays due to failed endpoints.

A Server Group consists of one or more destination addresses (endpoints). The Server Group has a domain name, which is also known as the SRV cluster domain name or FQDN. Server Groups work like a local SRV implementation (`srv.xml`), but the Server Group adds the extra heartbeat method to the SRV as an option. This feature only covers outbound calls from Unified CVP. To cover the inbound calls to Unified CVP, the SIP Proxy Server can send similar heartbeats to Unified CVP, which can respond with status responses.

Note

- Server Groups in Unified CVP and SIP Proxy Servers functions in the same way.
- A Server Group can only send heartbeats to endpoints defined in it.
- With record routes set to OFF, any mid-dialog SIP message bypasses the elements defined in Server Group. These messages include REFERs or REINVITES. These messages are delivered directly to the other endpoint in the dialog.
- Dialed number pattern updates that use a SIP Server Group are not recommended. These updates have to be done when no calls are running or in a maintenance window.

Unified CCE High Availability Considerations

The subcomponents of Unified CCE can recover from most failure scenarios without manual intervention. The redundant architecture ensures that your solution continues handling calls in single subcomponent failure scenarios. Only a rare simultaneous failure of several subcomponents interrupts your business operations.

Redundancy and Fault Tolerance

You deploy the Router and Logger in a paired redundant fashion. The two sides of the redundant deployment are referred to as Side A and Side B. For example, Router A and Router B are redundant instances of the Router running on two different VMs. In usual operation, both sides are running. When one side is down, the

configuration is running in stand-alone mode. These modes are occasionally referred to as duplex and simplex modes.

Note Stand-alone (simplex) deployments of the Router and Logger are not supported in production environments. You *must deploy* these components in redundant pairs.

The two sides are for redundancy, not load-balancing. Either side can run the full load of the solution. Sides A and B both perform the same set of messages and produce the same result. Logically, there is only one Router. The synchronized performance means that both sides process every call. During a failure, the surviving Router takes over the call midstream and continues without user intervention.

The Peripheral Gateway (PG) components run in hot-standby mode. Only one PG is active and controlling the Unified CM or the appropriate peripheral. When the active side fails, the surviving side automatically takes over processing. During a failure, the surviving side runs in stand-alone mode until the redundant side is restored. Then, the PGs automatically return to redundant operation.

The Administration & Data Servers, which handle configuration and real-time data, are deployed in pairs for fault tolerance. You can deploy multiple pairs for scalability. The Administration & Data Servers for historical data follow an N+1 architecture for redundancy and scalability. Each Administration & Data Server has a Logger (Side A or B) as its preferred and primary data source.

Router High Availability Considerations

Device Majority and Failovers

Device majority determines whether a Router enters a disabled state. The Router checks for device majority when it loses its connection with its redundant Router. Each Router determines device majority for itself. None, one, or both Routers can have device majority simultaneously.

To have device majority, a Router must meet one of these conditions:

- The Router is the Side A router and it can communicate with *at least half* of its total enabled PGs.
- The Router is the Side B router and it can communicate with *more than half* of its total enabled PGs.

Router Failover Scenarios

CTI Manager with Agent PG Link Fails

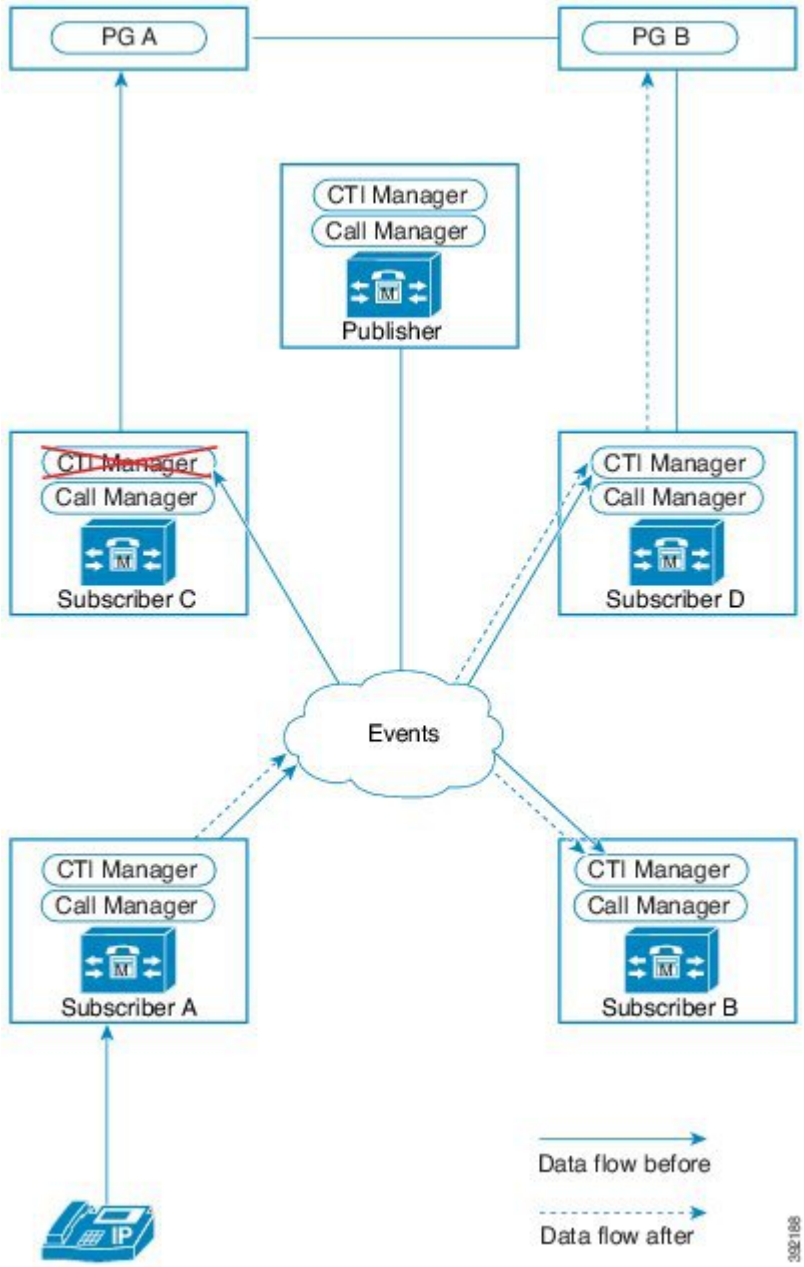
Each Agent PG can support only one CTI Manager connection. While each subscriber has a CTI Manager, only two subscribers connect to the Agent PGs. You would have to add another pair of Agent PGs to enable all subscribers in a four-subscriber cluster to connect directly to an Agent PG.

The following figure shows the failure of a CTI Manager with a connection to the Agent PG. Only subscribers C and D are configured to connect to the Agent PGs.

The following conditions apply to this scenario:

- For redundancy, all phones and gateways that are registered with subscriber A use subscriber B as their backup server.
- The CTI Managers on subscribers C and D provide JTAPI services for the Agent PGs.

Figure 88: CTI Manager with Agent PG Connection Fails



Failure recovery occurs as follows:

1. When the CTI Manager on subscriber C fails, the Agent PG Side A detects the failure and induces a failover to PG Side B.
2. Agent PG Side B registers all dialed numbers and phones with the CTI Manager on subscriber D and call processing continues.
3. In-progress calls stay active, but the agents cannot use phone services, like transfers, until the agents sign back in.

4. When the CTI Manager on subscriber C recovers, Agent PG Side B continues to be active and uses the CTI Manager on subscriber D. The Agent PG does not fail back in this model.

Subscriber Without CTI Manager Link to Agent PG Fails

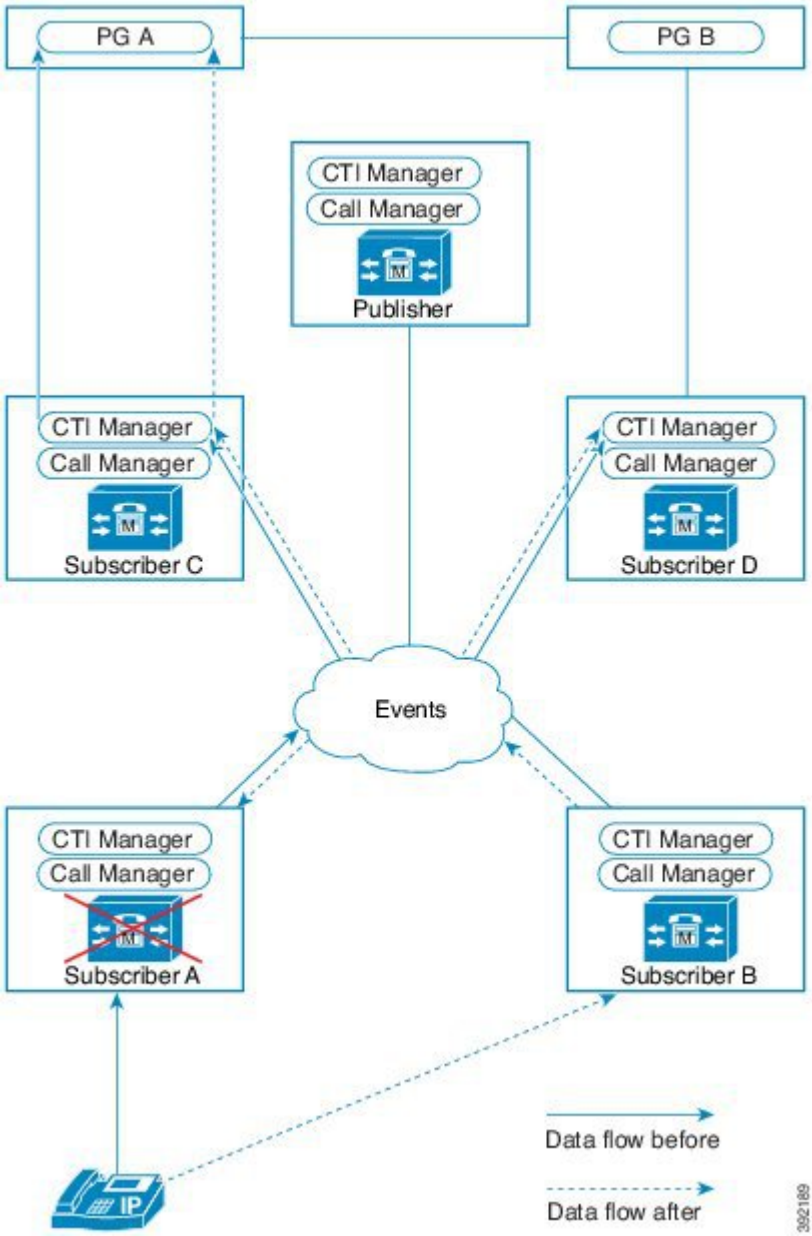
Each Agent PG can support only one CTI Manager connection. While each subscriber has a CTI Manager, only two subscribers connect to the Agent PGs. You would have to add another pair of Agent PGs to enable all subscribers in a four-subscriber cluster to connect directly to an Agent PG.

The following figure shows a failure on subscriber A, which does not have a direct connection to an Agent PG.

The following conditions apply to this scenario:

- For redundancy, all phones and gateways that are registered with subscriber A use subscriber B as their backup server.
- Subscribers C and D connect to the Agent PGs and their local instance of the CTI Manager provides JTAPI services for the PGs.

Figure 89: Unified Communications Manager Without Link to Agent PG Fails



Failure recovery occurs as follows:

1. If subscriber A fails, its registered phones and gateways rehome to the backup subscriber B.
2. Agent PG Side A remains active and connected to the CTI Manager on subscriber C. The PG does not fail over, because the JTAPI-to-CTI Manager connection has not failed. But, the PG detects the phone and device registrations automatically switching from subscriber A to subscriber B.
3. Call processing continues for any devices that are not registered to subscriber A.

4. While the agent phones are not registered, the Agent PG disables the agent desktops. This response prevents the agents from using the system without a subscriber connection. The Agent PG signs the agents out during this transition to avoid routing calls to them.
5. Call processing resumes for the phones after they reregister with their backup subscriber.
6. In-progress calls continue on phones that were registered to subscriber A, but the agents cannot use phone services, like transfers, until the agents sign back in.
7. When the in-progress call ends, that phone reregisters with the backup subscriber. The Agent PG signs the agents out during this transition to avoid routing calls to them.
8. When subscriber A recovers, phones and gateways rehome to it. You can set up the rehoming on subscribers to return groups of phones and devices gracefully over time. Otherwise, you can require manual intervention during a maintenance window to redistribute the phones to minimize the impact to the call center. During this rehoming process, the CTI Manager notifies the Agent PG of the registrations switching from subscriber B back to the original subscriber A.
9. Call processing continues after the phones and devices return to their original subscriber.

Multiple Failure Scenarios

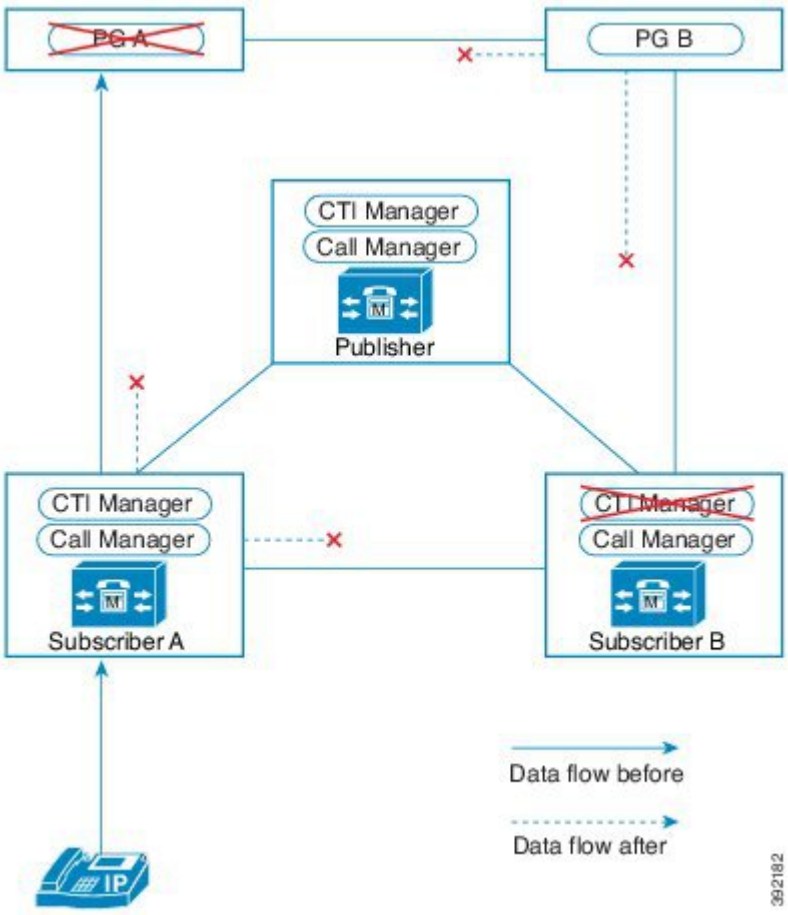
When more than one component fails, Unified CCE might not fail over as seamlessly as during a single-component failure. The following sections discuss how Unified CCE responds to multicomponent failures.

CTI Manager and Agent PG Fail

A CTI Manager connects only with its local subscriber and a single Agent PG. There is no direct communication with the other CTI Manager in the cluster. The CTI Managers are kept in synch by data from the other components.

If the Agent PG on one side and the CTI Manager on the other side both fail, Unified CCE cannot communicate with the cluster. This scenario prevents the system from connecting to the agents on this cluster. The cluster remains disconnected until the Agent PG or the backup CTI Manager come back online.

Figure 90: Agent PG Cannot Connect to Backup CTI Manager



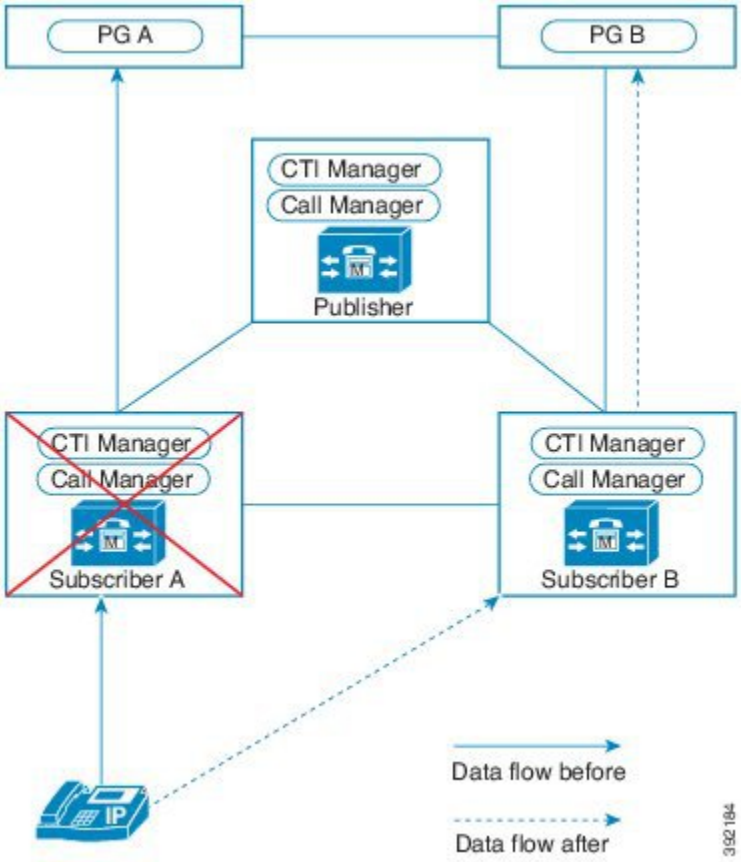
Unified CM Subscriber and CTI Manager Both Fail

The scenario shows recovery from a complete failure of the Unified CM subscriber A server.

The following conditions apply to this scenario:

- Subscriber A has the primary CTI Manager.
- For redundancy, all phones and gateways that are registered with subscriber A use subscriber B as their backup server.

Figure 91: Unified Communications Manager and CTI Manager Fail



Failure recovery occurs as follows:

1. When subscriber A fails, all inactive registered phones and gateways reregister to subscriber B.
2. The in-progress calls remain active, but the agents cannot use phone services, like transfers.
3. Agent PG Side A detects a failure and induces a failover to Agent PG Side B.
4. Agent PG Side B becomes active and registers all dialed numbers and phones. Call processing continues.
5. As each in-progress call ends, that agent phone and desktop reregister with the backup subscriber. The exact state of the agent desktop varies depending on the configuration and desktop.
6. When subscriber A recovers, all idle phones and gateways reregister to it. Active devices wait until they are idle before reregistering to the primary subscriber.
7. Agent PG Side B remains active using the CTI Manager on subscriber B.
8. After recovery from the failure, the Agent PG does not fail back to Side A of the redundant pair. All CTI messaging is handled using the CTI Manager on subscriber B which communicates with subscriber A to obtain phone state and call information.

Logger High Availability Considerations

Logger Fails

The Unified CCE Logger and Database Server maintain the system database for the configuration (agent IDs, skill groups, call types) and scripting (call flow scripts). The server also maintains the recent historical data from call processing. The Loggers receive data from their local Router. Because the Routers are synchronized, the Logger data is also kept synchronized.

The Logger failure has no immediate impact on the call processing. The redundant Logger receives a complete set of call data from its local Router. If the system restores the failed Logger, the Logger automatically requests all the transactions for when it was offline from the backup Logger. The Loggers maintain a recovery key that tracks the order of the recorded entries in the database. The redundant Logger uses these keys to identify the missing data.

If the Logger remains offline for more than the 14 day retention period of the Config_Message_Log table, the system does not resynchronize the Logger configuration database automatically. The system administrator can manually resynchronize the Loggers using the Unified ICMDBA application as described in the *Administration Guide*, at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-maintenance-guides-list.html. The manual process allows you to choose a convenient time to transfer the configuration data across the private network.

The Logger replication process sends data from the Logger database to the HDS database on the Administration and Data Servers. The replication process also automatically replicates each new row that the Logger database records after the Logger synchronization takes place.

In deployments that use Cisco Outbound Option with only a single Campaign Manager, the Campaign Manager is loaded only on the primary Logger. If that platform is out of service, any outbound calling stops while the Logger is down.

Reporting Considerations

The Unified CCE reporting feature uses real-time, 5 minute, and reporting-interval (15 or 30 minute) data to build its reporting database. At the end of each 5 minute and reporting interval, each PG gathers its local data and sends it to the Routers. The Routers process the data and send the data to their local Logger for historical data storage. The Logger replicates the historical data to the HDS/DDS database.

The PGs provide buffering (in memory and on disk) of the 5-minute data and reporting-interval data. The PGs use this buffered data to handle slow network response and automatic retransmission of data after network services are restored. If both PGs in a redundant pair fail, you can lose the 5-minute data and reporting-interval data that was not sent to the Central Controller.

When agents sign out, all their reporting statistics stop. When the agents next signin, the real-time statistics for the agents start from zero. Depending on the agent desktop and what an agent is doing during a failure, some failovers can cause the contact center to sign out agents. For more information, see the *Reporting Concepts for Cisco Unified ICM/Contact Center Enterprise* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-user-guide-list.html.

Peripheral Gateway High Availability Considerations

PG Weight

During a failover for a private link failure, a weighted value determines which PG becomes the enabled PG. The number and type of active components on each side determines the weighted value of the PG. The weight assigned to each component reflects the recovery time of that component and the disruption to the contact center when the component is down. Agent PIMs have higher weights than VRU PIMs and the CTI Server. The component weights are not configurable.

Record Keeping During Failovers

The call data that gets recorded during a failover depends on which component fails. Depending on the failure condition, some call data is lost. The Router can lose access to active calls because of the failure. The active calls are still active, but the Router responds as if the calls have dropped. Usually, the Agent PG creates a Termination Call Detail (TCD) record in the Unified CCE database.

Calls that are already connected to an agent can continue during a failover. The Agent PG creates another TCD record for such calls when they end.

Agent PG Fails

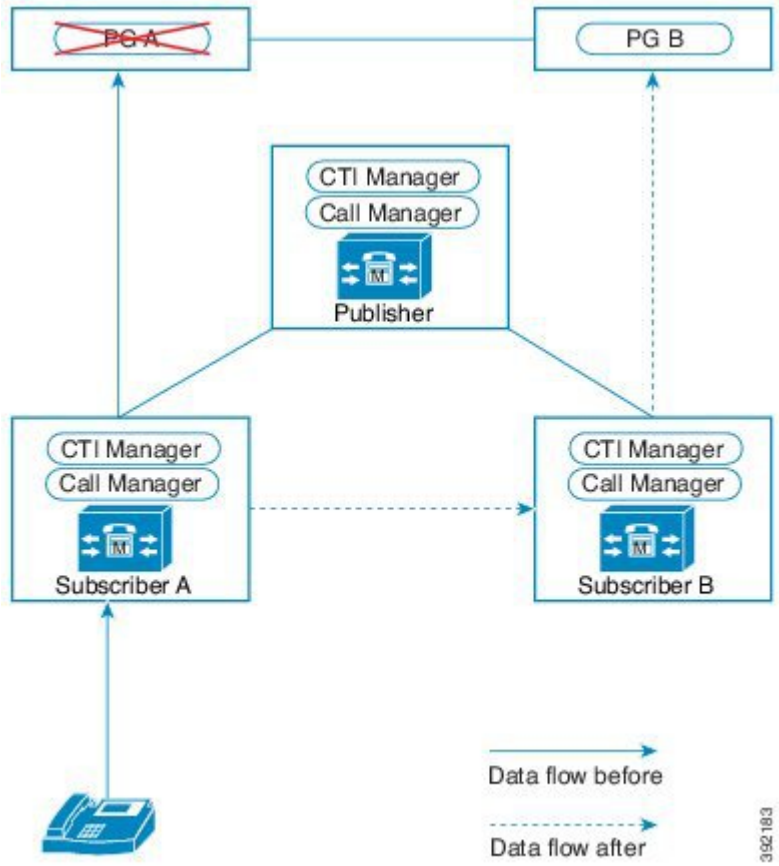
This scenario shows recovery from a PG Side A failure.

The following conditions apply to this scenario:

- Unified CM subscriber A has the primary CTI Manager.
- For redundancy, all phones and gateways that are registered with subscriber A use subscriber B as their backup server.

The following figure shows a failure on PG Side A and a failover to PG Side B. All CTI Manager and Unified Communications Manager services continue running as usual.

Figure 92: Agent PG Side A Fails



Failure recovery occurs as follows:

1. PG Side B detects the failure of PG Side A.
2. PG Side B registers all dialed numbers and phones. Call processing continues through PG Side B.
3. Phones and gateways stay registered and operational with subscriber A; they do not fail over.
4. The in-progress calls remain active on agent phones, but the agents cannot use phone services, like transfers, until the agents sign back in.
5. During the failover to PG Side B, the states of unoccupied agents and their desktops can change depending on their configuration. Options for three-party calls can be affected. In some cases, agents have to sign back in or manually change their state after the failover completes.
6. After recovery from the failure, PG Side B remains active and uses the CTI Manager on subscriber B. The PG does not fail back to Side A, and call processing continues on the PG Side B.

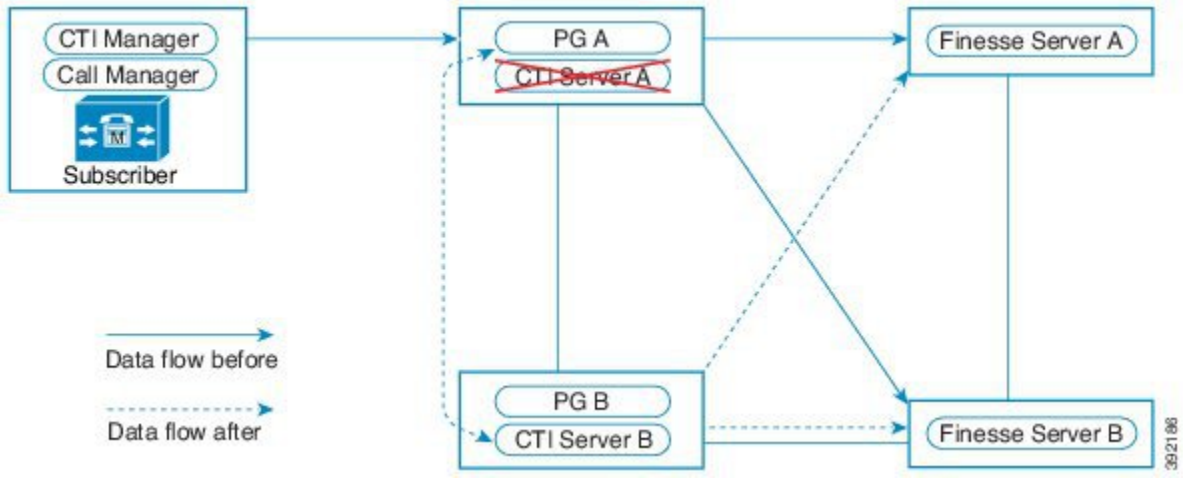
CTI Server Fails

The CTI Server monitors the Agent PG traffic for specific CTI messages (such as call ringing or off-hook events). The CTI Server makes those messages available to CTI clients such as the Cisco Finesse server. The

CTI Server also processes third-party call control messages (such as make call or answer call) from the CTI clients. The CTI Server sends those messages through the Agent PG to Unified CM for processing.

You deploy the CTI Server in redundant pairs. Each half of the redundant pair is coresident on a VM with one half of a redundant Agent PG pair. On failure of the active CTI Server, the redundant CTI Server becomes active and begins processing call events.

Figure 93: CTI Server Fails



The Finesse server is a client of the CTI Server. The desktop server, rather than the CTI Server, maintains agent state during a failover. Finesse partially disables agent desktops when the CTI Server fails. In some cases, an agent must sign in again after the failover completes.

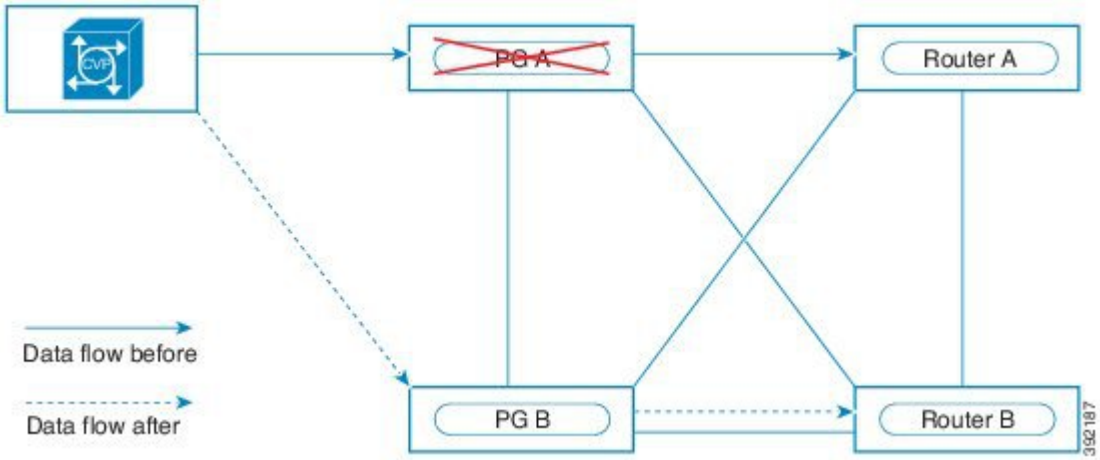
Note If no clients are connected to the active CTI Server, a mechanism forces a failover after a preset period. This failover isolates any spurious reasons that prevent the CTI clients from connecting to the active CTI Server.

VRU PG Fails

When a Voice Response Unit (VRU) PG fails, calls in progress or queued in Unified CVP do not drop. The Survivability TCL script in the Voice Gateway redirects the calls to a secondary Unified CVP or a number in the SIP dial plan, if available.

After failover, the redundant VRU PG connects to the Unified CVP and begins processing new calls. On recovery of the failed VRU PG side, the currently running VRU PG continues to operate as the active VRU PG. Redundant VRU PGs enable Unified CVP to function as an active queue point or to provide call treatment.

Figure 94: VRU PG Fails



Administration & Data Server High Availability Considerations

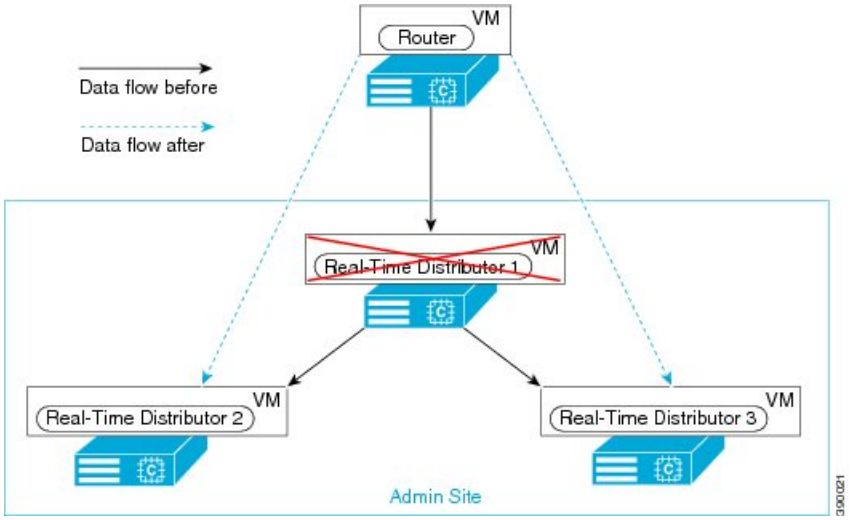
Administration and Data Server Fails

The Administration and Data Server provides the user interface to the system for making configuration and scripting changes. The server can also host the web-based reporting tool and the Internet Script Editor. Unlike other Unified CCE components, the Administration and Data Server does not operate in redundant pairs. If you want to provide redundancy for the functions on this server, you can include more Administration and Data Servers in your design. But, there is no automatic failover behavior.

The Administration and Data Server receives a real-time feed of data from across Unified CCE from the Router through a Real-Time Distributor. If you have several Administration and Data Servers at the same site, you can configure the Real-Time Distributors into a single Administrator Site. The Administrator Site has a primary distributor and one or more secondary distributors. The primary distributor registers with the Router and receives the real-time feed across the network from the router. The secondary distributors use the primary distributor as their source for the real-time feed. This arrangement reduces the number of real-time feeds that the router supports and saves bandwidth.

If the primary real-time distributor fails, the secondary real-time distributors register with the router for the real-time feed as shown in the following figure. Administration clients that cannot register with the primary or secondary Administration and Data Server cannot perform any tasks until the distributors are restored.

Figure 95: Primary Real-Time Distributor Fails



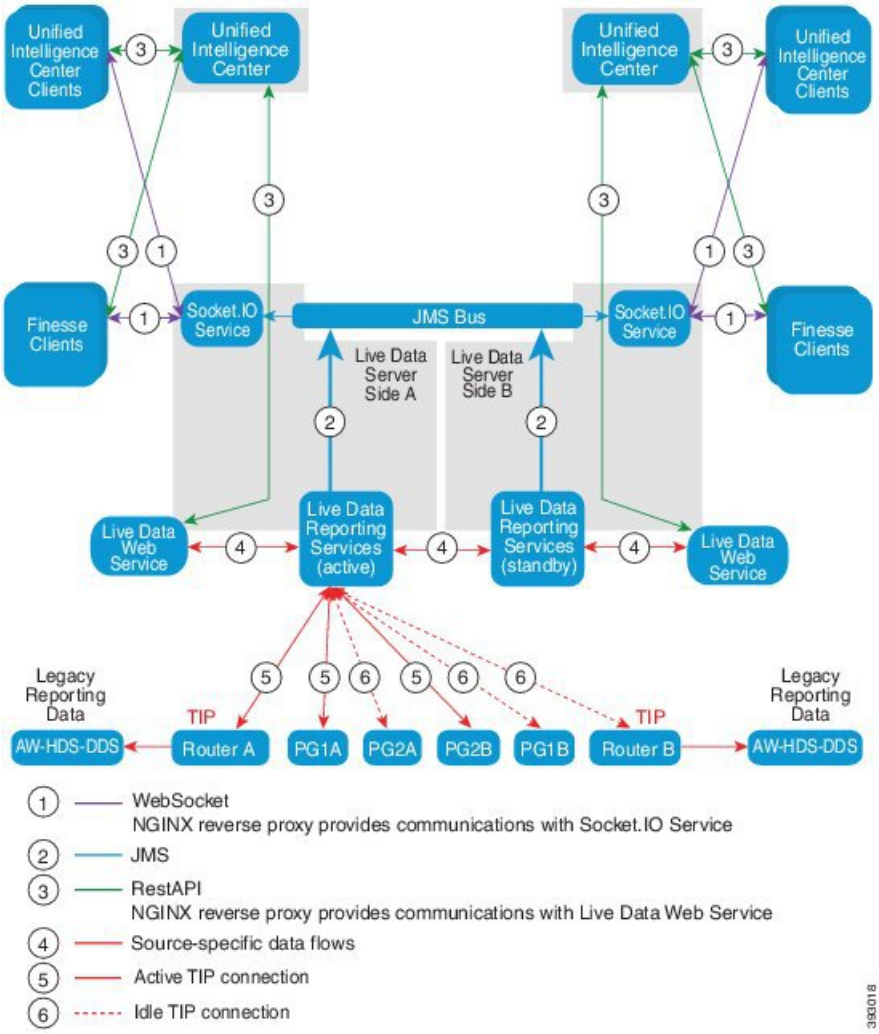
Live Data High Availability Considerations

Live Data is a highly available system deployed as two Live Data systems, one on Side A and one on Side B. By design, a Live Data deployment can tolerate any single point of failure. There are three layers of failover:

- Server
- TIP
- Socket.IO stream

Note Live Data server failover is also called Live Data cluster failover. This document uses the term Live Data server failover.

Figure 96: Live Data Reporting Topology



Live Data Server Failover

The Live Data servers work in cold-active or standby mode. Only one Live Data server is active at any time. The other Live Data server is standby. The standby Live Data server constantly monitors the status of the active server. When the active server fails, the standby server takes over and becomes active. The failing server becomes the standby server when it is ready to serve.

A weighted algorithm determines which Live Data server is active in the following two scenarios:

Scenario 1: When both Live Data servers start simultaneously, the servers use the same device majority calculation as the Routers.

Scenario 2: The active Live Data server can lose connectivity to some of the PGs. The standby server then detects that loss. If it has 130% PGs more than the active server for two minutes, it requests to assume the active status. The standby server becomes the active server, and the server that was previously active becomes the standby server.

TIP Failover

Live Data uses the TIP transport protocol to communicate with the Router and PG servers. The active Live Data server establishes TIP connections to both sides of the Router and PGs. The standby Live Data server does not establish any TIP connections. Only one TIP connection is active at a time, either to Side A or to Side B. When the active TIP connection fails, the active Live Data server recovers to the idle TIP connection.

Socket.IO Failover

A Socket.IO client connects to either side of the Live Data server to receive the Live Data report event stream (Socket.IO stream). Unified Intelligence Center clients are an example of a Socket.IO client. The standby Live Data server also produces the Socket.IO stream by proxy from the active server. Socket.IO client heartbeat losses results in a Socket.IO connection failure. The Socket.IO client then fails over to the other Live Data server.

Virtualized Voice Browser High Availability Considerations

Cisco Virtualized Voice Browser (VVB) is a single node with no built-in high availability for active redundancy. To improve the level of availability and to eliminate a single point of failure, deploy more VVBs. You can build passive redundancy by including the extra VVBs in the CVP SIP Server group. By deploying more VVBs, you can manage unscheduled and scheduled downtime of one of the VVBs.

During a VVB failure, all active calls on the failed VVB disconnect and all the call data is lost. After CVP detects a failure of a VVB in the SIP Server group, CVP routes incoming calls to the remaining active VVBs. When the CVP heartbeat mechanism detects the recovery of the failed VVB, CVP starts routing calls to the recovered VVB.

Unified CM High Availability Considerations

After you design the data network, design the Cisco Unified Communications infrastructure. Before you can deploy any telephony applications, you need the Unified CM cluster and CTI Manager in place to dial and receive calls.

Several services that are important to your solution run on each Unified CM server:

- Unified CM
- CTI Manager
- CallManager service
- TFTP

For details on the architecture of all these services, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/en/US/docs/voice_ip_comm/uc_system/design/guides/UCgoList.html.

High availability design for a cluster requires that you understand how the Unified CM, CTI Manager, and CallManager services interact. Unified CM uses the CTI Manager service to handle its CTI resources. CTI Manager acts as an application broker that abstracts the physical binding of applications to a particular Unified CM server. The CallManager service registers and monitors all the Cisco Unified Communications devices.

The CTI Manager accepts messages from the Agent PG, a CTI application, and sends them to the appropriate resource in the cluster. The CTI Manager acts like a JTAPI messaging router using the Cisco JTAPI link to

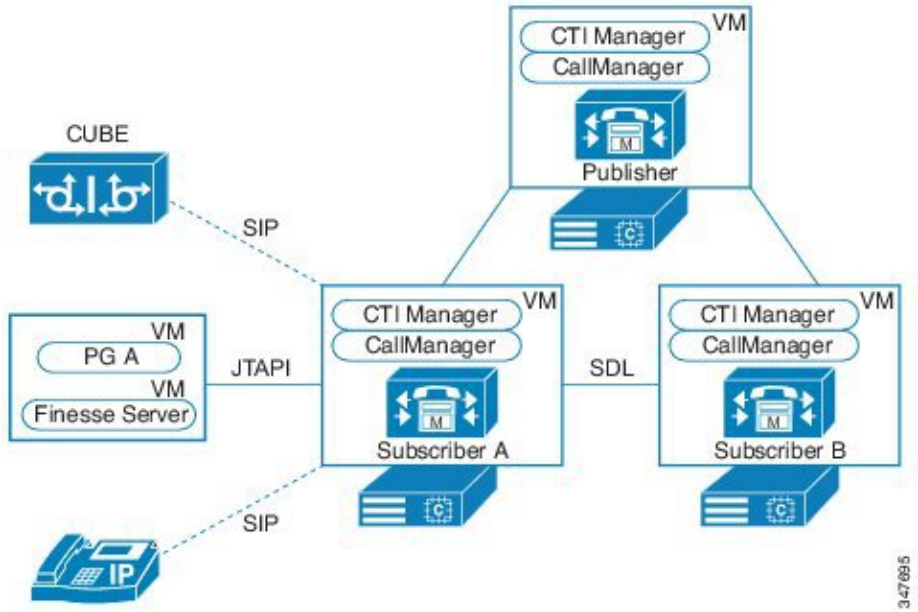
communicate with Agent PGs. The JTAPI client library in Unified CM connects to the CTI Manager instead of connecting directly to the CallManager service.

The CallManager service acts as a switch for all the Cisco Unified Communications resources and devices in the system. The CallManagers on each Unified CM server link themselves across the public network with the Signal Distribution Layer (SDL). This link keeps the cluster in sync. Each CTI Manager connects with the Unified CM and CallManager services on its server. CTI Managers do not connect directly with other CTI Managers in the cluster.

Agent PGs use a CTI-enabled user account in Unified CM, typically called the "JTAPI user" or "PG user". The Agent PGs sign in to the CTI Manager to connect to the devices for that user. If the appropriate device is resident on the local CallManager, the CTI Manager handles the request for that device. If the device is not resident on its local subscriber, then the CallManager service forwards the request to the appropriate subscriber through the private link to the other CallManager services.

The following figure shows the connections in a cluster.

Figure 97: Connections in Unified Communications Manager Cluster



For high availability, distribute device registrations across all the subscribers in the cluster. If you concentrate the registrations on a single subscriber, the traffic puts a high load on that subscriber. The memory objects that the Agent PGs use to monitor registered devices also add to the device weights on the subscribers.

If the PG that is connected to a subscriber fails, the redundant PG that takes over and sends all the requests to another subscriber. Then, the local CallManager service must route the CTI Manager messaging for those requests across the cluster to the original subscriber. The additional messaging in this failover condition creates greater load on the cluster.

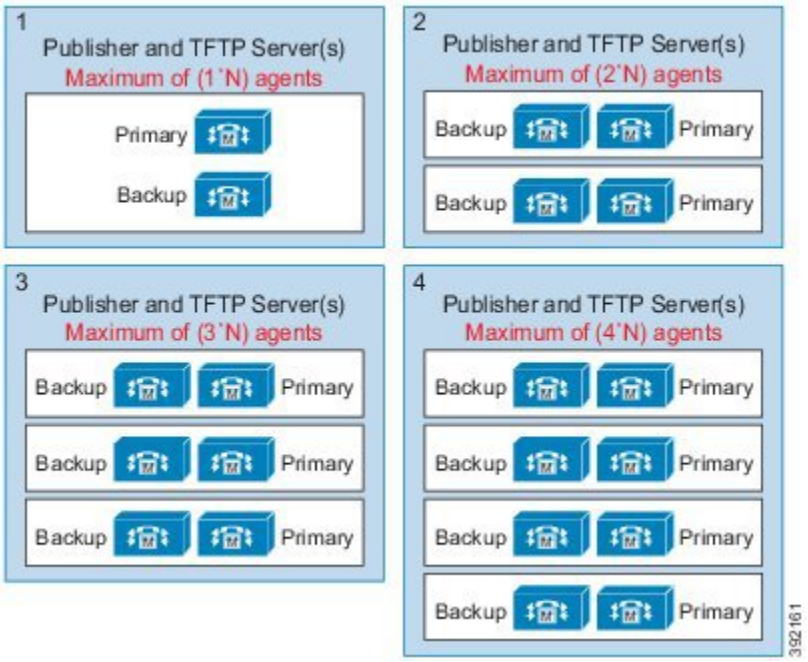
Unified CM Redundancy

Some Unified CM deployments use a 2:1 redundancy scheme. Each pair of primary subscribers shares a single backup subscriber. But, because of the higher phone usage in contact centers and to simplify upgrade processes,

contact center enterprise solutions uses a 1:1 redundancy scheme for subscribers. Each primary subscriber requires its own backup subscriber.

This figure shows different size clusters. For a contact center enterprise solution that uses Unified CVP, *N* is equal to 2000/pair of subscribers in this figure.

Figure 98: Redundancy Configuration Options



Unified CM Load Balancing

The 1:1 redundancy scheme for Unified CM subscribers lets you balance the devices over the primary and backup subscriber pairs. Generally, a backup subscriber has no devices registered unless its primary subscriber is unavailable.

You can enable load balancing through Unified CM redundancy groups and device pool settings. You can move up to half of the device load from the primary to the secondary subscriber. In this way, you can reduce by half the impact of any server becoming unavailable. To minimize the effect of any outage, distribute all devices and call volumes equally across all active subscribers.

Cisco Finesse High Availability Considerations

You deploy the Cisco Finesse server in redundant pairs in contact center enterprise solutions. Both Cisco Finesse servers are always active. When a Cisco Finesse server goes out of service, the agents on that server are put into a NOT READY or pending NOT READY state. They are redirected to the sign-in page of the other server. This can happen when the following situations occur:

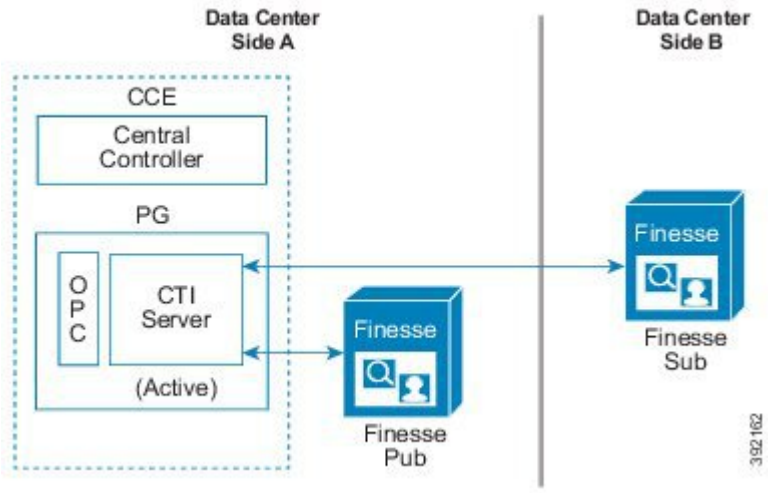
- The Cisco Finesse Tomcat Service goes down.
- The Cisco Notification Service goes down.

- Cisco Finesse loses connection to both CTI servers.

If a client disconnects, it tries to reconnect to one of the two available Cisco Finesse servers. If the reconnect takes longer than 2 minutes, Cisco Finesse signs out the agent. The agent then has to sign in when the client reconnects.

A single Agent PG supports one instance of a Cisco Finesse cluster, consisting of two servers, a publisher and subscriber. Multiple Finesse clusters cannot communicate with the same Agent PG/CTI Server. Each Cisco Finesse server can support the maximum of 2,000 users that the CTI server supports. This capacity enables one Cisco Finesse server to handle the full load if the other server fails. The total number of users between the two Cisco Finesse servers cannot exceed 2,000. Each Cisco Finesse server requires a single CTI connection, as shown in the following figure:

Figure 99: Multiple Cisco Finesse Servers



When deploying Cisco Finesse, follow the coresidency policies outlined in the *Cisco Collaboration Virtualization* at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

Cisco Finesse IP Phone Agent Failure Behavior

Unlike the desktop, the Cisco Finesse IP Phone Agent (Cisco Finesse IPPA) does not automatically failover to the alternate Cisco Finesse server. For proper failover behavior, configure at least two Cisco Finesse IP Phone services in Unified CM. Each service should use different Cisco Finesse servers.

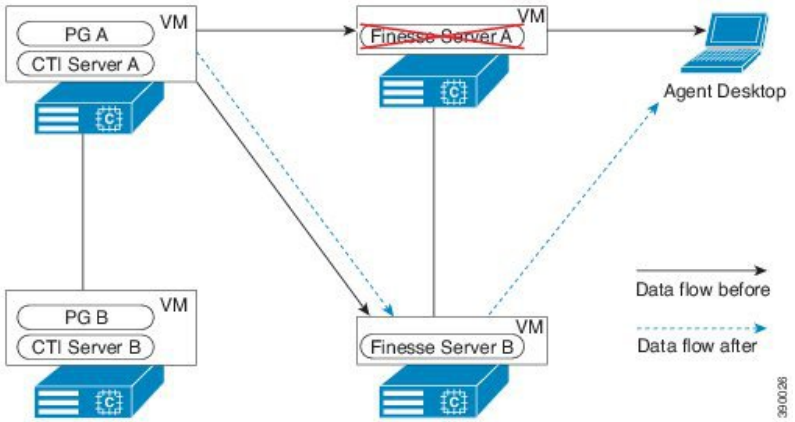
When the Cisco Finesse server fails, Cisco Finesse IPPA attempts to reconnect to it every 5 seconds. After three failed attempts, Cisco Finesse IPPA displays a server unavailable message to the agent. The total time to go out of service is approximately 15 seconds.

In a failure scenario, the agents must sign out and then sign in to an alternate Cisco Finesse server. The agents can then resume usual operations.

Cisco Finesse Server Fails

You deploy the Cisco Finesse server in redundant pairs in dedicated virtual machines. Both Cisco Finesse servers run in active mode all the time.

Figure 100: Cisco Finesse Server Fails



When a Cisco Finesse server fails, failure recovery occurs as follows:

1. Agent desktops that are signed in to the server detect a loss of connection and fail over to the redundant server.
2. Agents are automatically signed in on the new server after the failover.

Note When Cisco Finesse server fails over, the Desktop Chat Status is retained and all active chat sessions are lost.

3. Third-party applications that use the Cisco Finesse REST API must perform the failover within their application logic to move to the redundant server.
4. The Cisco Finesse server does not automatically restart. After you restart the failed server, new agent desktop sessions can sign in on that server. Agent desktops that are signed in on the redundant server remain on that server.

Note If Cisco Finesse Tomcat for one side fails, that Cisco Finesse server fails over.

Cisco Finesse Behavior When Other Components Fail

The following sections describe Cisco Finesse behavior when other Unified CCE components fail.

Agent PG Fails or CTI Server Fails

Cisco Finesse servers connect to the active Agent PG which is coresident with and connects to the CTI server. If the active Agent PG fails or the CTI server fails, Cisco Finesse tries to connect to the redundant CTI server. If the redundant server is unavailable, then Finesse keeps trying to connect to either of the servers until it is successful. Then Finesse clears all its agent, skill group, and Call data. Cisco Finesse is out of service until all current configuration is received from the redundant CTI server including the agent and call states. Agents see a red banner on the desktop when it loses connection, followed by a green banner when it reconnects.

Benchmark Parameters:

Cisco Finesse, Release 12.5(1) optimizes CTI failover and Desktop failover performances.

- CTI Failover—The maximum time taken for the CTI server or Agent PG failover varies from 35 seconds to 75 seconds when deployed with Agent PG 12.5(1).
- Desktop Failover—The maximum time taken for the desktop failover with the default desktop layout varies from 50 seconds to 110 seconds when deployed with Agent PG 12.5(1).

The failover time varies depending on the WAN bandwidth, the number of signed-in users, latency, CPU count, and the number of the gadgets configured on the Finesse desktop.

For more information on deployment practices and guidelines to ensure optimal failover performance, see *Guidelines for Optimal Desktop Failover* and *Failover Planning* sections in *Cisco Finesse Administration Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-maintenance-guides-list.html.

For more information on ensuring how the custom gadgets improve failover performance, see *Best Practices for Gadget Development* section in *Cisco Finesse Web Services Developer Guide* at https://developer.cisco.com/docs/finesse/#!rest-api-dev-guide.

Administration & Data Server Fails

Cisco Finesse uses the Administration & Data Server to authenticate agents. The Cisco Finesse administrator configures the settings for the primary Administration & Data Server (and optionally, the backup Administration & Data Server) in the Cisco Finesse administration user interface. If the primary Administration & Data Server fails and a backup Administration & Data Server is not configured, Cisco Finesse agents cannot sign in to the desktop. Agents who are signed in when the failover occurs can no longer perform operations on the desktop.

If the backup Administration & Data Server is configured, Cisco Finesse tries to connect to the backup server. After Cisco Finesse connects to the backup Administration & Data Server, agents can sign in and perform operations on the desktop.

Cisco IM&P Server Fails

Cisco Finesse uses the Instant Messaging and Presence (IM&P) servers for the Desktop Chat feature. IM&P pulls its user list from users who have been enabled for chat capabilities, from Unified CM (or LDAP if LDAP integration is enabled).

Failover is supported for Desktop Chat and any IM&P node failure results in automatic connection to the node pair peer, as configured for the user.

Note When Cisco Finesse server fails over, the Desktop Chat Status is retained and all active chat sessions are lost.

Unified Intelligence Center High Availability Considerations

Cisco Unified Intelligence Center uses a cluster model with a publisher and up to 7 subscribers for high availability. Configuration replicates within the cluster. Processing automatically spreads between the active nodes, bypassing any failed nodes.

Data source failover

Cisco Unified Intelligence Center supports failover for data sources while a report is run. If the report which is run on the data source fails with the following error codes, the report is attempted to run on the failover data source.

Error codes

```
08000, CONNECTION_EXCEPTION
08003, CONNECTION_DOES_NOT_EXIST
08006, CONNECTION_FAILURE
08002, CONNECTION_NAME_IN_USE
08001, SQL_CLIENT_UNABLE_TO_GET_CONNECTION
08004, SQL_SERVER_REJECTED_CONNECTION
08007, TRANSACTION_RESOLUTION_UNKNOWN
08S01, CONNECTION_EXCEPTION_COMM_FAILURE
```

Note When *ValueList* data source failover is supported, administrator has to manually switch data source using the switch button in the data source card. Data source failover takes approximately 1 minute, during this period, the report failure may be observed.

Unified CM-based Silent Monitoring High Availability Considerations

For existing calls, there is no high availability. For incoming calls, call processing and silent monitoring move to the backup Unified CM subscriber.

Customer Collaboration Platform High Availability Considerations

Cisco Customer Collaboration Platform (CCP) does not support high availability.

CCP uses either a small or large, single-server, all-in-one, deployment. You cannot use a load-balancing, split site deployment.

Unified SIP Proxy High Availability Considerations

With `RecordRoute` disabled, Unified SIP Proxy can handle the failover of active calls. In an active call during failover, the backup SIP Proxy server handles new transactions.

Enterprise Chat and Email High Availability Considerations

Enterprise Chat and Email (ECE) provides high availability. The colocated ECE deployments and 1500 Agent clusters, both support Geo-redundancy using the following techniques:

- Use a load balancer to distribute incoming requests across multiple web servers. If a server goes down, the load balancer detects the failure and redirects requests to another application server. This capability supports 300 agents on each web server for up to 5 web servers. It does not provide any redundancy at maximum capacity.
- Keep all subcomponents of ECE within the network round trip times. For more information, see the *Enterprise Chat and Email Design Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/cisco-enterprise-chat-email/products-implementation-design-guides-list.html.
- To support ECE failover when Unified CCE components fail, ensure that the Agent and MR PGs are implemented as per the high availability requirement. This technique ensures that a single subcomponent failure does not block processing of all sessions.
- Install the database using Microsoft SQL Server's Availability Group clustering configuration. For more information, see the *Enterprise Chat and Email Installation and Configuration Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/cisco-enterprise-chat-email/products-installation-guides-list.html.

Load-Balancing Considerations for Enterprise Chat and Email

You can load balance the web service component of an ECE deployment to serve many agents. You can set up the web (or web and application) servers behind the load balancer with a virtual IP address. When an agent accesses ECE with the virtual IP address, the load balancer sends a request to one of the servers behind the address. The load balancer then sends a response back to the agent. In this way, from a security perspective, the load balancer also serves as a reverse proxy server.

The load balancer must support sticky sessions with cookie-based persistence. After maintenance tasks, verify that all Web and application servers are available to share the load. If you allow agents access without all servers being available, the sticky connection feature can cause an overload on the first Web and application server.

Using other parameters, you can define a load-balancing algorithm to meet the following objectives:

- Equal load balancing
- Isolation of the primary Web and application server
- Send fewer requests to a low-powered Web and application server.

The load balancer monitors the health of all Web and application servers in the cluster. During a failure, the load balancer removes that server from the available pool of servers.

ECE Behavior When Other Components Fail

Failures of other solution components generally have no effect on active sessions. All active sessions continue uninterrupted.

These sections describe the ECE behavior for Incoming sessions when other solution components fail.

Agent PG Failover

If the Agent PG to which an ECE server connects fails over, the effect on incoming sessions is as follows:

- **Web Callback Sessions**—During the failover period, customers cannot schedule a Web Callback session to the agents on the failed PG. If there are other Agent PGs, ECE can assign the Web Callback session to an agent on those PGs. If there are no available Agents, the Web Callback session can queue for resources to become available. After the failover to the redundant PG completes, all incoming sessions can use the available agents on that PG.
- **Delayed Callback Sessions**—Processing of the callback switches to the redundant PG. When the specified delay elapses, the callback goes through.
- **Chat Sessions**—The incoming chat session reaches an agent after failover to the redundant PG completes.
- **Email**—Processing of incoming email resumes after failover to the redundant PG completes.

MR PG Failover

If the MR PG to which an ECE server connects fails over, the effect on incoming sessions is as follows:

- **Queued Sessions**—Sessions that are already in queue remain in queue. After the failover to the redundant PG completes, ECE reissues the previously queued sessions to the PG.
- **Web Callback Sessions**—The new session is established between the customer and the agent after failover to the redundant PG completes.
- **Delayed Callback Sessions**—Processing of the callback switches to the redundant PG. When the specified delay elapses, the callback goes through.
- **Chat Sessions**—The incoming chat session reaches an agent after failover to the redundant PG completes.
- **Email**—Processing of incoming email resumes after failover to the redundant PG completes.

CTI Manager Failover

If the CTI Manager through which an ECE server connects fails over, the effect on incoming sessions is as follows:

- **Web Callback Sessions**—The new session cannot be placed and the customer receives the message, "System cannot assign an Agent to the request."
- **Delayed Callback Sessions**—Processing of the callback switches to the redundant CTI Manager. When the specified delay elapses, the callback goes through.
- **Chat Sessions**—The incoming chat session reaches an agent after failover to the redundant CTI Manager completes.
- **Email**—Processing of incoming email resumes after failover to the redundant CTI Manager completes.

Router Failover

If the active Router fails over, the redundant Router seemlessly handles all incoming sessions.

ASR TTS High Availability Considerations

solution supports redundant ASR/TTS servers. In a basic configuration, the VXML Gateway first passes all incoming requests to the primary ASR/TTS server. If the primary server is unreachable, the gateway then

passes that request to the backup server. Any request that reaches the backup server stays on that server for the duration of the request.

You can add a load balancer to spread the incoming requests across your ASR/TTS servers.

Outbound Option High Availability Considerations

The Cisco Outbound Option includes these subcomponents:

Subcomponent	Location	Redundancy	Description
Campaign Manager	Logger A and B	Redundant	Manages the dialing lists and rules that are associated with the calls.
Outbound Option Import	Logger A and B	Redundant	Imports campaign records.
Outbound Option database	Logger A and B	Redundant	Holds the records for the calling campaigns.
SIP Dialer	Agent PG A or B MR PG A or B	Redundant	Performs the dialing tasks that the Campaign Manager assigns according to the campaign. The SIP Dialer transfers calls that connect to the assigned agents.

To improve high availability in the Cisco Outbound Option, you can also use a redundant CUSP pair to connect to multiple voice gateways. The redundant gateways ensure that the Dialers have enough trunks available to place calls if a gateway fails. If outbound calling is the primary application, you can dedicate these gateways to outbound calling only.

Outbound Option High Availability supports two-way replication between the Outbound Option database on Logger Side A and the Outbound Option database on Logger Side B. The two-way replication is performed over the public network between the Loggers.

Your solution supports multiple Dialers and a redundant pair of Campaign Managers that control the Dialers, in warm-standby mode. The redundant pairs of SIP Dialers operate in a warm-standby mode similar to the PG fault-tolerance model.

SIP Dialer Design Considerations

The SIP Dialers run in warm-standby mode. The Campaign Manager activates one SIP Dialer in the Ready state from its registered SIP Dialer pool. If the activated SIP Dialer changes state from Ready to Not Ready or loses its connection, the Campaign Manager activates the standby SIP Dialer. The Campaign Manager returns all outstanding records to Pending status after a timeout period.

The active SIP Dialer fails over if it loses connection to the CTI Server, Agent PG, or SIP server. The SIP server can be a voice gateway or CUSP. Connect each dialer in a redundant pair to a different SIP server.

For regulatory compliance, the SIP Dialer does not automatically re-attempt calls that were in progress during a failover. Instead, the Dialer sends all active and pending customer records to the Campaign Manager. If the Campaign Manager is not available, the dialer closes them internally.

The CUSP server provides weighted load balancing and redundancy in a multiple-gateway deployment by configuring each gateway as part of the Server group configuration. If a gateway is overloaded or loses its WAN link to the PSTN network, CUSP can resend an outbound call to the next available gateway.

The Campaign Manager and SIP Dialer already include warm-standby functionality. Because of this, do not use the Hot Swappable Router Protocol (HSRP) feature for CUSP servers that are dedicated for Outbound Option.

Outbound Option Record Handling During Fail Over

The Dialer updates the Campaign Manager with the intermediate status of the customer records. This ensures that the Campaign Manager tracks the next set of actions when the Dialer fails over.

When the Dialer calls a customer by sending out a SIP Invite, it sends a state update message for the customer record to the Campaign Manager. The Campaign Manager then updates the CallStatus of the record to the Dialed state in the DialingList (DL) table.

The Campaign Manager again updates the state of the customer records in the following events:

- **When the call is successful**: The Campaign Manager updates the customer records to the Closed state.
- **When the connection fails between the Dialer and Campaign Manager**: All the Dialed state records remain in the Dialed state. The Active state records move to the Unknown state.
- **When the connection fails between the Dialer and the CTI server**: The Campaign Manager updates the customer records to the Closed state. Next, the Campaign Manager sends the dialer-disconnected status and all the Active state records move to the Unknown state. The Dialed state records remain in the Dialed state.
- **When the connection fails between the Dialer and the SIP gateway (GW)**: The Campaign Manager receives a Close customer record message once the call is released from the Agent Desktop. In this condition, all the Dialed state records moves to Closed state when the call is released from Agent Desktop. The Active state records move to the Unknown state.
- **When the connection fails between the dialer and the MR PIM**: The Campaign Manager receives only the Dialer status message with a connected status. After it receives the Close customer record message, it updates the records to the Closed state.
- **When the Campaign Manager Fails**: All the Dialed state records move to the Closed state. The Active state records move to the Unknown state.

Campaign Manager High Availability Considerations

Campaign Manager runs in warm-standby mode as a redundant pair, Side A and Side B. By default, the Campaign Manager on Side A (Campaign Manager A) is set as the Active Campaign Manager. The Campaign Manager B is set as the Standby Campaign Manager. Each of the redundant pair of Loggers has its own deployment of Campaign Manager and Outbound Option Import.

When you enable Outbound Option high availability, the processes initiate bidirectional database replication for the contact table, dialing lists, Do Not Call table, and Personal Callbacks (PCB).

At system startup, the Outbound Option Import and Dialers initiate connections to the Campaign Managers. The standby Campaign Manager accepts the Outbound Option Import connections from the standby side and sets the Outbound Option Import to standby state. However, the standby Campaign Manager refuses the Dialer

connections including the Dialer connections from its resident side. The active Campaign Manager accepts the Outbound Option Import and the Dialer connections, including the Dialers from the standby side.

The Outbound Option (Blended Agent) Import process on a Logger side communicates only to the Campaign Manager on the same Logger side. Therefore, the status of the Campaign Manager and Blended Agent Import process, on the respective sides, are in synchronization with each other.

If either of the two, the Dialer or the Blended Agent Import process fails to connect to Campaign Manager within EMTClientTimeoutToFailover interval, the Campaign Manager switches over.

The Campaign Manager fails over when any of the following failures occur:

- The connection to the Outbound Option Import fails.
- The connections to all the dialers fail.

Note The active Campaign Manager does not fail over if it is connected to even one dialer.

- All the Dialers report Not Ready state to the active Campaign Manager.
- The connection to the router fails.

When the failed Campaign Manager comes back online, it is set to standby state. The active Campaign Manager continues in the active state.

When one Campaign Manager in the redundant pair fails, the other side stores replication transactions in a series of files in a replication folder. Take this into account when you size the disk space on the Loggers.

Note If Outbound Option high availability is not enabled, the Router recognizes the deployment type accordingly and discards the failover messages.

Dialer Behavior during Campaign Manager Failover

Dialers connect to the active Campaign Manager. A Dialer alternates connection attempts between the Side A and Side B Campaign Managers until it connects to the active Campaign Manager. When the active Campaign Manager goes down, the standby Campaign Manager becomes active after a configurable interval (default is 60 seconds).

The Dialer failover behavior is as follows:

- **At system startup**: The Side A Campaign Manager becomes active. The Dialers send connection requests first to the Side A Campaign Manager. If that Campaign Manager does not connect, the Dialers send requests to the Side B Campaign Manager after the configurable interval. The active Campaign Manager connections are accepted and established.
- **When the Dialer detects disconnection from the active Campaign Manager**: The Dialer sends a connection request to the standby Campaign Manager. If the Campaign Manager failover is complete, then the standby Campaign Manager becomes the active Campaign Manager and connects to the Dialer.

If the Campaign Manager failover has not occurred, the standby Campaign Manager rejects the connection request. The Dialer then alternates connection requests between the Campaign Managers until one becomes active and accepts the connection request.

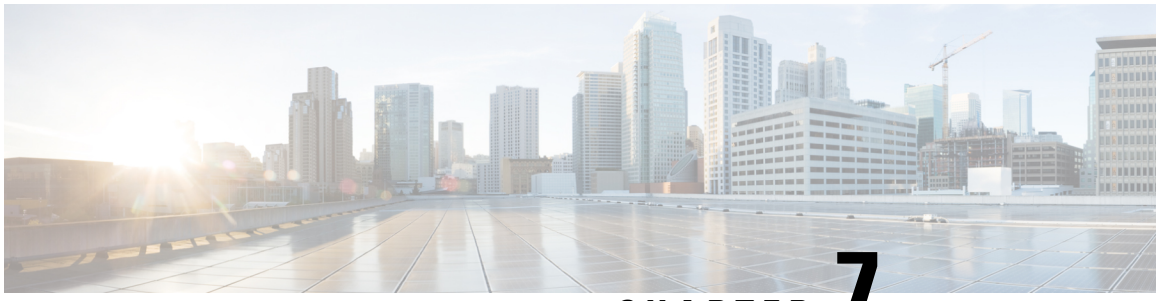
The Microsoft Windows Event Viewer, SYSLOG, and SNMP capture the disconnection and connection attempts of the Dialers.

Single Sign-On High Availability Considerations

You deploy the Cisco Identity Service (Cisco IdS) as a cluster. The cluster contains a publisher and a subscriber. The cluster nodes automatically replicate configuration data and authorization codes across the cluster. When a node reconnects, the cluster determines the most recent configuration and authorization code data and replicates that across the cluster.

A contact center application can authenticate and authorize an agent or supervisor if it can reach any node. The contact center applications query their local Cisco IdS node by default. If that node is unavailable, the applications query any configured remote node. When the local node reconnects to the cluster, the applications return to querying the local node.

If the packet loss on your network exceeds 5 percent, a node might not obtain an access token using an authorization code that the other node issued. In this case, the user has to sign in again. If the packet loss becomes too great or the connection is lost, the Cisco IdS functions as a solo node. The cluster automatically reforms when network connectivity improves.



CHAPTER 7

Design Considerations for Integrated Features

Agent Greeting Considerations

Consider these points when you add Agent Greeting to your solution:

- Agent Greeting does not support outbound calls made by an agent. The announcement plays for inbound calls only.
- Only one Agent Greeting file plays per call.
- Supervisors cannot listen to agent recorded greetings.
- Agent Greetings do not play when the router selects the agent through a label node.

- Agent Greeting supports Unified CM-based Silent Monitoring with this exception: Supervisors cannot hear the greetings themselves. If a supervisor starts a silent monitoring session while a greeting plays, a message appears that a greeting is playing and to try again shortly.
- Use either G.711 a-law or mu-law for the VRU leg on the Voice Browser dial-peer. Do not use the voice-class codec.
- In general, Agent Greeting feature requires shorter latency across the system. For example, the public network has a maximum round-trip latency of 100 ms to support Agent Greeting feature as designed.

Agent Greeting requires the following:

- The phones have the BIB feature.
- The phones must run the latest firmware version delivered with Unified Communications Manager.
- The phones must be have BIB enabled in Unified Communications Manager.

Agent Greeting with Whisper Announcement

You can use Agent Greeting with the Whisper Announcement feature. Consider these points when using them together:

- The Whisper Announcement always plays first.
- To shorten your call-handling time, use shorter Whisper Announcements and Agent Greetings than if you were using either feature by itself. A long Whisper Announcement followed by a long Agent Greeting equals a long wait before an agent actively handles a call.
- If you use a Whisper Announcement, your agents probably handle different types of calls: for example, “English-Gold Member-Activate Card,” “English-Gold Member-Report Lost Card,” “English-Platinum Member-Account Inquiry.” Ensure that greetings your agents record are generic enough to cover the range of call types.

Agent Greeting Phone Requirements for Local Agents

Agent Greeting is available to agents and supervisors who use IP Phones with Built-In Bridge (BIB). These agents are typically located within a contact center. Phones used with Agent Greeting must meet these requirements:

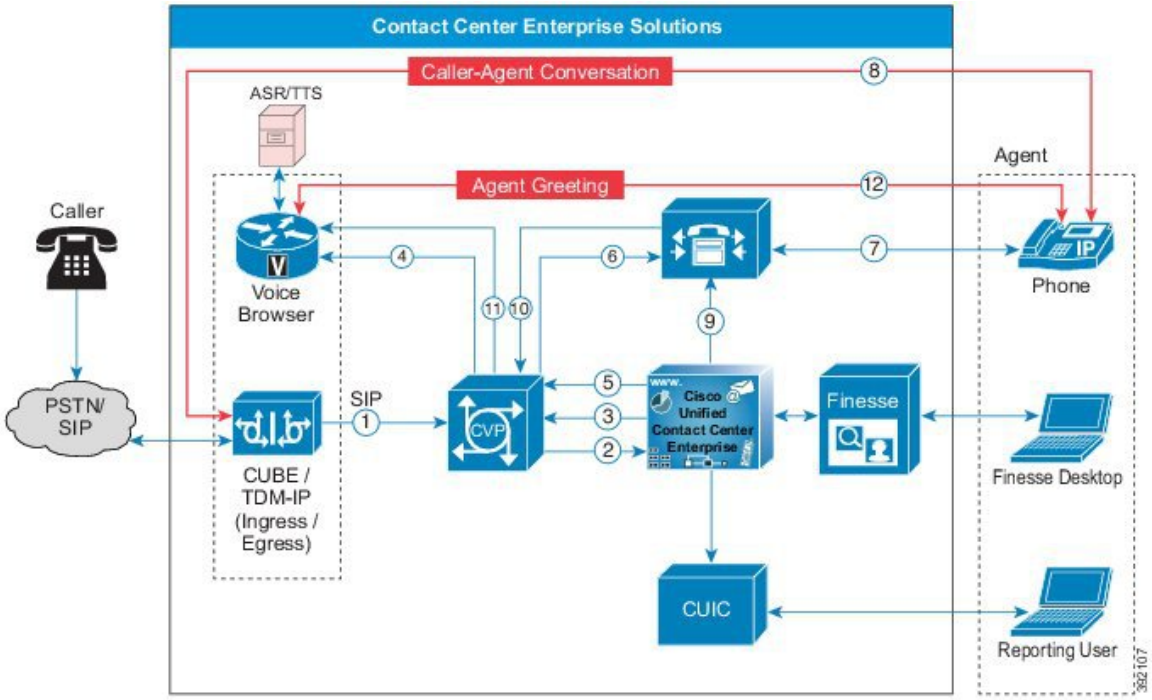
- The phones must have the BIB feature.

Note If you disable BIB, the system attempts to use a conference bridge for agent greeting call flow and raises a warning event.

- Ensure that the phone's firmware is up to date. (Usually, phone firmware upgrades automatically when you upgrade your Unified CM installation.)
- For a list of supported phones for contact center enterprise solutions, see the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html.

Agent Greeting Call Flows

Figure 101: Agent Greeting Call Flow



1. The incoming call arrives from CUBE or a TDM gateway at CVP.
2. CVP sends the incoming call to Unified CCE.
3. Unified CCE instructs CVP to queue the call.
4. CVP sends the call to the Voice Browser for VRU treatment.
5. When an agent is available, Unified CCE sends the agent number to CVP.
6. CVP sends the call to Unified CM.
7. Unified CM establishes the connection to the agent phone.
8. The caller connects to the agent phone and stops hearing the ringback.
9. Unified CCE determines which CVP to invoke, and instructs Unified CM to tell the phone BIB to open a stream to CVP.
10. Unified CCE and CVP shake hands to set the trigger for CVP to let it know which greeting to play.
11. CVP instructs the Voice Browser to have the Media Server play the greeting.
12. The phone's BIB mixes the greeting. After the greeting plays, CVP disconnects and the agent speaks with the caller.

Agent Greeting Design Impacts

Sizing Considerations with Agent Greeting

Agent Greeting invokes conference resources to bring the greeting into the call. For most phones, it uses the Built-In Bridge feature on the phone. For Mobile Agent, it uses conference resources. This adds a short but extra call leg to every call, which has impacts on several components.

Voice Browser and CVP

Agent Greeting uses CVP and Voice Browser resources. Agent Greeting has a profile of short calls but at a high call rate. Account for these calls when sizing your solution.

Router and Logger

Agent Greeting has an impact of up to 1.5 regular calls on the Router and Logger. That lowers the maximum call rate for your solution by a third. Each Agent Greeting involves an additional route request. The Router PerfMon counter reflects this extra request as a higher call rate.

Peripheral Gateway

The impact of Agent Greeting on the PG resource usage does not reduce the supported agent capacity per PG.

Unified CM

When Agent Greeting can affect the number of agents that a Unified CM subscriber supports.

Mobile Agent

If you enable Agent Greeting with Mobile Agent, it uses extra Conference Bridge and MTP resources. To properly size the Conference Bridge and Unified CM resources, add a conference for each inbound call in place of the Agent Greeting.

Sizing the Agent Greeting Prompt Cache

If you enable Agent Greeting, properly size the prompt cache.

Consider the following example for a 1-minute long file in the G.711 mu-law codec:

The following calculation shows that the prompt uses approximately 1/2 MB:

```
Prompt size = 8 kb/sec (g711uLaw bit rate) * 60 seconds - 480 kb
```

On a Cisco IOS router, the maximum prompt cache is 100 MB. The maximum size of a single file should be 600 KB.

This table gives some example sizing for prompt caches on an IOS router:

Table 46: Agent Greeting Prompt Cache Sizing

Greeting Duration	Greeting Size	Total Greetings
5 second	40 KB	2000 agent greetings with 80-percent space reserved for Agent Greeting

Greeting Duration	Greeting Size	Total Greetings
60 second	480 KB	100 agent greetings with 50-percent space used for Agent Greeting

Note For Cisco VVB, the maximum cache size is 512 MB which allows you to cache more greetings.

Agent Greeting Impact on the Call Server

The maximum CPS for contact center enterprise solutions assumes that you use Agent Greeting. The impact of this feature is already accounted for in the CPS limit.

Enabling Agent Greeting also affects the port usage. The required ports are calculated based on the CPS and duration of agent greeting.

Agent Greeting Impact on the Voice Browser

Agent Greeting increases the Voice Browser sessions required for your solution. You calculate the Voice Browser sessions based on CPS and the duration of the agent greeting. The agent greeting counts as one extra call to the Voice Browser.

Use the following formula to determine the total sessions including the extra sessions required for the Agent Greeting feature:

```
Total sessions = Inbound sessions + ((Greeting Duration / Total call duration) * Inbound
sessions)
```

For example, 120 calls with a 60-second duration is a rate of 2 CPS and requires 120 inbound sessions. If the agent greeting duration is 5 seconds, then the overall rate is 4 CPS, but the number of sessions required is 130.

```
Total sessions = 120 inbound sessions + [(5-second agent greeting duration / 60-second total
 call duration) *
120 inbound sessions] = 130 total sessions.
```

Application Gateway Considerations

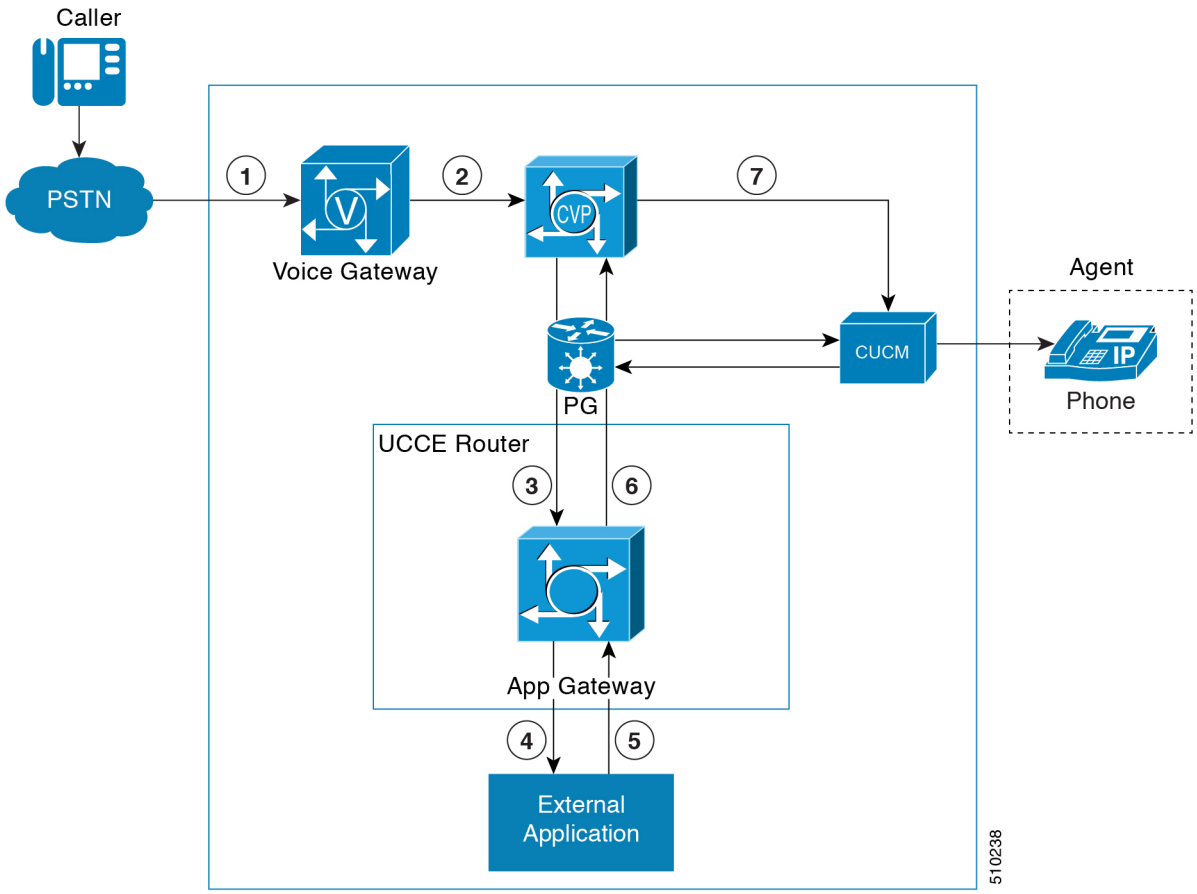
The custom application requires to be written to conform to the specifications described in the application gateway protocol spec, GED-145.

The application gateway has several options for fault tolerance models which are required to be considered while designing and deploying the application.

Application Gateway Call Flows

The basic Contact Sharing call flow runs as shown in this diagram:

Figure 102: Application Gateway Call Flows



Application Gateway Design Impacts

The target server for the application gateway is required to run on a separate virtual machine.

Application Gateway Sizing Considerations

The maximum call rate in Application Gateway parallels to the maximum call rate for the system.

Business Hours Considerations

Administrators can easily configure Business Hours from CCE Administration web interface rather than the regular CCE Configuration tools such as Agent Explorer Tools, Announcement tool and so on. A partner can use the API to incorporate this functionality in their tools.

Business Hours Use Cases

Use the Business Hours feature to manage the incoming customer calls or digital channel communications, by routing these contacts based on the Business Hours you configure.

Use the Business Hour status in an IF node in scripts to control the call and digital channel contacts, such as email and chat, and notify the customers accordingly.

You can have Business Hours scripts for the following treatments:

- When the business is open, route calls and digital contacts to the applicable skill groups and precision queues.
- When the business is closed, play the message for the closed status with the appropriate Status Reason and terminate the call. Route the digital contacts to the appropriate queues.
- When the business is not open 24x7, route the calls to skill groups and precision queues for after-hours support or play the after-hours message.
- When the business is open 24x7, at a predefined time before the end of each shift, route the calls and digital contacts to the appropriate queues for the next shift.
- When the business is closed for an emergency on a working day, notify the customers contacting your contact center appropriately about the emergency closing.

Based on reason code and status, the customers will hear appropriate prompts on the call.

Business Hours Design Impacts

Packaged CCE supports Business Hours for these reference designs:

- 2000 Agents
- 4000 Agents
- 12000 Agents

Customer Virtual Assistant Considerations

CVA feature enables the IVR Platform to integrate with cloud-based speech services. This feature supports human-like interactions that enable customers to resolve issues quickly and more efficiently within the IVR thereby, reducing the calls directed towards actual agents.

Concepts for CVA

Following are the concepts that are to be handled in a CVA application. For more information on these concepts, see Google's Dialogflow documentation at https://cloud.google.com/dialogflow/docs/.

Concepts	Description
Speech Operations	• **TTS**: Synthesise the prompt in text form to audio form. For more information see https://cloud.google.com/text-to-speech/. • **ASR**: Recognise the user voice and transcribe voice into text. For more information, see https://cloud.google.com/speech-to-text/.
Context Management	Decide which Dialogflow Intent parameter is to be filled and determine the correlation between multiple dialogs to fill multiple parameters for an intent.
Session Management	Manage relationship between one intent and another within an IVR Session.
Business Logic	Define the sequential flow of events, such as which event follows which event.
Fulfilment	Perform the action for the user intent in terms of invoking the REST API, database operation, or customer action.
Natural Language Understanding (NLU)	Process the caller's speech in the form of text and use the words uttered by the caller to understand the intent and identify the parameters.
NLU Design / Training	Design the NLU and its training. This includes identifying intents, entities, slots, and training the NLU engine for different scenarios with training data.

Note The CVA feature is supported only on VVB. It is not supported with the VXML Gateway.

CVA Call Flows and Architecture

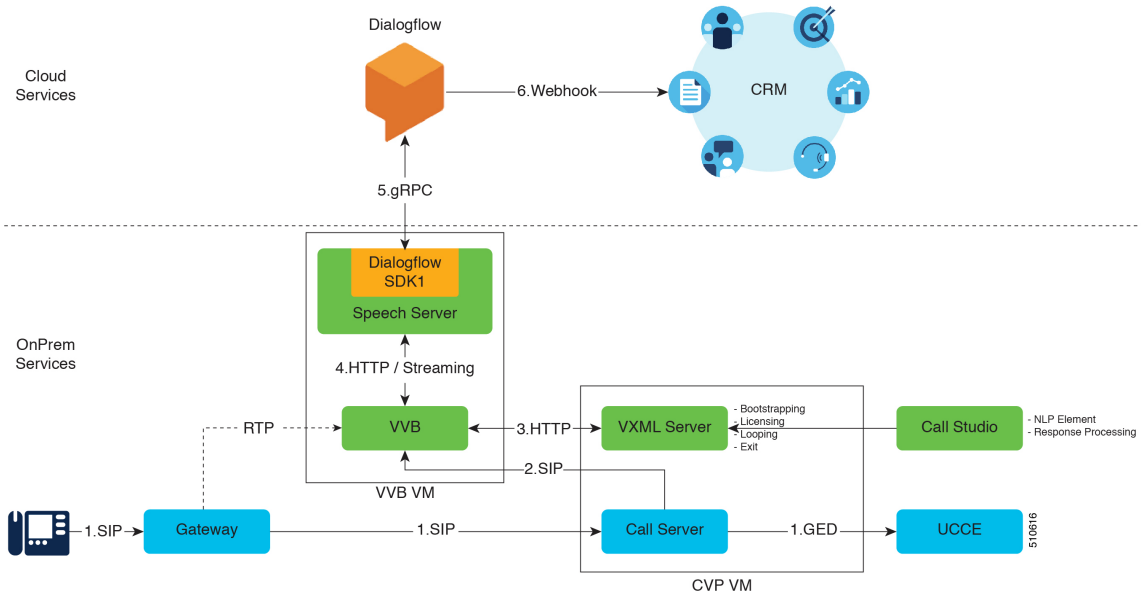
Dialogflow Element Call Flow

This call flow integrates Call Studio application in the easiest and quickest way.

In this call flow:

- CVP acts as a platform and controls the Bootstrapping, Licensing, Looping, and Exit logic.
- Dialogflow Agent performs Speech operations, Context Management, Business Logic, Fulfilment, and NLU operations.

Figure 103: Architecture



1. Call Server receives a SIP call and gets the IVR treatment instruction from UCCE.
2. Call Server invokes VVB for IVR treatment.
3. VVB fetches the VXML page from VXML Server (based on Call Studio application) and creates a VXML page for CVA / non-CVA elements.
4. VVB, when required to interact with TTS / ASR services, streams media to Speech Server.
5. Speech Server engages the Dialogflow plugin and relays the media to Dialogflow.
6. Dialogflow does following processing:
   - Transcribes the voice and converts it to text internally.
   - Processes the text and identifies the intents.
   - Engages a CRM by a webhook based API.
   - Returns the subsequent prompt/fulfilment result (based on `AudioOutput` property configuration in Call Studio Application) to VVB in the form of audio or text.

     If the output is audio, Dialogflow returns the audio payload in an API response. VVB plays this audio directly and no action is required in Call Studio. If the output is in the form of text, Dialogflow returns the text prompt in a response that needs to be synthesized by engaging TTS service in a separate `Audio` element in the call flow.

DialogflowIntent/DialogflowParam Element Call Flow

This call flow provides finer control on the flow.

In this call flow:

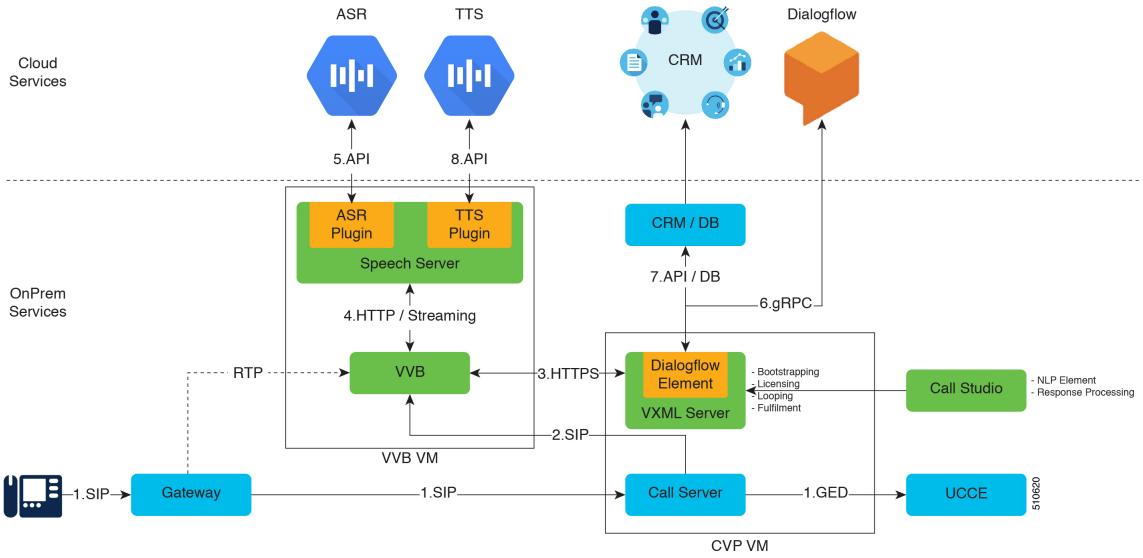
- Call Studio application performs Context Management, Session Management, Business Logic, and Fulfillment.
- Separate speech services perform the different speech operations.
- Dialogflow performs all the NLU operations.

Note When the intent is created in Dialogflow, the **Required** check box must be deselected so that the intent can be controlled in the Call Studio application.

This call flow can be used when:

- Existing Call Studio application implements different custom integrations with Database / API to be extended to support CVA capabilities.
- Prompt is to be generated dynamically or hosted on media server and controlled from Call Studio application.
- Parameter sequencing and handling are to be controlled from Call Studio application.

Figure 104: Architecture



1. Call Server receives a SIP call and gets the IVR treatment instruction from UCCE.
2. Call Server invokes VVB for IVR treatment.
3. VVB fetches the VXML page from VXML Server (based on Call Studio application) and creates a VXML page for CVA / non-CVA elements.
4. Based on the Call Studio script, VVB determines whether to collect DTMF or to interact with TTS / ASR services and then streams the media to Speech Server.
5. Speech Server engages the ASR service and relays the media to Dialogflow.
   a. ASR returns the recognized text back to VVB.

b. VVB returns the recognized text to VXML Server.

6. VXML Server calls the Dialogflow API and sends the user utterance text to it for processing.

   VXML Server interacts with Dialogflow in the following two ways:

   - **DialogflowIntent**: The user utterance is used to identify the intent.
   - **DialogflowParam**: The user utterance is used to fill the param (slot) for an intent.

7. VXML Server repeats step 3 to step 6 until all the parameters for an intent are filled, after which VXML Server can fulfil the intent through API / Custom Code / DB Operation.
8. After every operation to prompt the customer, VXML Server generates the text that is to be synthesized to customer. This text goes to the TTS service through VVB and Speech Server, and the prompt is synthesized to the customer. VVB then caches the static prompts for further usage.

Note **Bandwidth Considerations**: Bandwidth calculations for CVA-based call flows is quite different from traditional DTMF-based call flows. So, to calculate actual bandwidth usage for CVA-based calls, consider per call bandwidth consumption of 20 kbps.

VVB Scale with CVA: Call capacity for Virtualized Voice Browser with CVA feature remains inline to already published scale number for traditional IVR with ASR / TTS services.

Cisco Outbound Option Considerations

Cisco Outbound Option for Unified CCE places outbound calls through a Voice Gateway. The Outbound Option Dialer does not require telephony cards to generate tones or to detect tones or voices.

The Cisco Outbound Option involves the following processes:

- Campaign Manager and Import processes manage campaigns.
- Depending on your fault tolerance strategy, you can have one Campaign Manager or a redundant pair.
- The Dialer process dials customers and connects them with properly skilled agents or available VRUs. The Dialer reports the results of all contact attempts back to the Campaign Manager. The active Campaign Manager manages all Dialer processes. The Dialer is installed on the same platform as the Agent PG.
- A Media Routing Peripheral is required for the Dialer to reserve agents for outbound use. It can coreside on other servers in a Unified CCE deployment.
- Mobiles agents are supported only with a nailed connection for outbound campaigns.

Note Precision Routing does not support Cisco Outbound Option. Outbound campaigns use skill groups. However, an agent involved in an outbound campaign (through an outbound skill group) can sign in to a Precision Queue and handle inbound Precision Routing calls.

Cisco Outbound Option provides the following benefits:

- Enterprise-wide dialing, with IP Dialers placed at multiple call center sites. The Campaign Manager server is located at the central site.
- Centralized management and configuration through the Unified CCE Administration & Data Server.
- Call-by-call blending of inbound and outbound calls.
- Flexible outbound mode control. Use the Unified CCE script editor to control the type of outbound mode and percentage of agents within a skill to use for outbound activity.
- Integrated reporting with outbound specific reporting templates.

The time required to complete a call transfer of a customer call to an agent depends on the telephony environment. The following factors can add to transfer times:

- **Improperly configured Cisco Unified Communications infrastructure**—Port speed mismatches between servers or inadequate bandwidth.
- **WAN**—WAN unreliable or not configured properly.
- **IP Communicator**—Media termination running on a desktop does not have the same system priority as with a desk phone. Use desk phones instead of IP Communicator for Outbound Option.
- **Call Progress Analysis**—Call Progress Analysis (CPA) takes a half second to differentiate between voice and an answering machine if the voice quality is good. When calling mobile phones, the voice quality is often less than optimal, so it takes the dialer or Voice Gateway longer to differentiate.

  You cannot use Virtual CUBEs with CPA.

Outbound Option Dialing Modes

Outbound Option has several dialing modes.

Note All dialing modes reserve an agent at the start of every outbound call cycle by sending a reservation call to the agent.

Predictive Dialing

In predictive dialing, the dialer determines the number of customers to dial per agent based on the abandon rate. The agent must take the call if that agent is signed in to a campaign skill group.

A Predictive Dialer is designed to increase the resource utilization in a call center. It is designed to dial several customers per agent. After reaching a live contact, the Predictive Dialer transfers the customer to a live agent along with a screen pop to the agent's desktop. The Predictive Dialer determines the number of lines to dial per available agent based on the target abandoned percentage.

Outbound Option predictive dialing works by keeping outbound dialing at a level where the abandon rate is below the maximum allowed abandon rate. Each campaign is configured with a maximum allowed abandon rate. In Predictive mode, the dialer continuously increments the number of lines it dials per agent until the abandon rate approaches the preconfigured maximum abandon rate. The dialer begins lowering the lines per agent until the abandon rate goes below the preconfigured maximum. In this way, the dialer stays just below the preconfigured maximum abandon rate. Under ideal circumstances, the dialer internally targets an abandon

rate of 85% of the preconfigured maximum abandon rate. Due to the random nature of outbound dialing, the actual attainable abandon rate at any point in time may vary for your dialer.

Preview Dialing

Preview dialing reserves an agent prior to initiating an outbound call and presents the agent with a popup window. The agent may then Accept, Skip, or Reject the call with the following results:

- **Accept** - The customer is dialed and transferred to the agent.
- **Skip** - The agent is presented with another customer call.
- **Skips-Close** - The customer is not called again, and the agent is presented with another customer call.
- **Reject** - The agent is released. The system delivers another call to the agent, either another preview outbound call, or a new inbound call.
- **Rejects-Close** - The agent is released and the record is closed so it is not called again. The system delivers another call to the agent, either another Preview outbound call or a new inbound call.

Direct Preview Dialing

The Direct Preview mode is similar to the Preview mode, except that the dialer automatically calls from the agent's phone after the agent accepts. Because the call is initiated from the agent's phone, the agent hears the ringing, and there is no delay when the customer answers. However, the agent must deal with answering machines and other results that the Dialer Call Progress Analysis (CPA) handles in other modes.

Note

- The CPA and the transfer to IVR features are not available while using Direct Preview Dialing mode
- A zip tone is a tone that announces incoming calls. There is no zip tone in Direct Preview mode

Progressive Dialing

Progressive Dialing is similar to predictive dialing. But, in Progressive Dialing mode, Outbound Option does not calculate the number of lines to dial per agent. It allows you to configure a fixed number of lines that are always dialed per available agent.

Personal Callback Mode

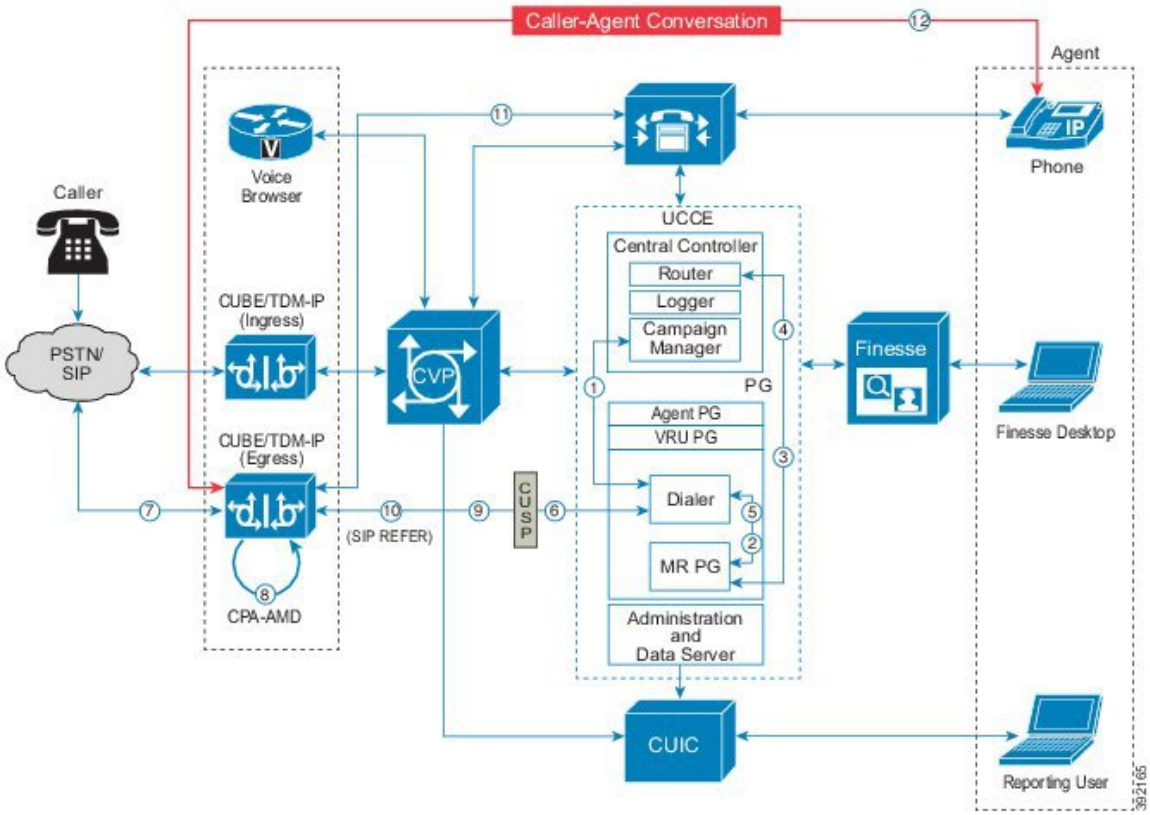
When the person who is called requests to be called back later, the agent can specify that the callback is directed to the same agent. The system then calls the customer back at a prearranged time established between the requested agent and the customer.

Cisco Outbound Option Call Flows

Call Flow for Agent Campaign

The following figure illustrates a transfer to agent call flow in an Outbound Option deployment with a SIP dialer.

Figure 105: SIP Dialer Agent Campaign Call Flow



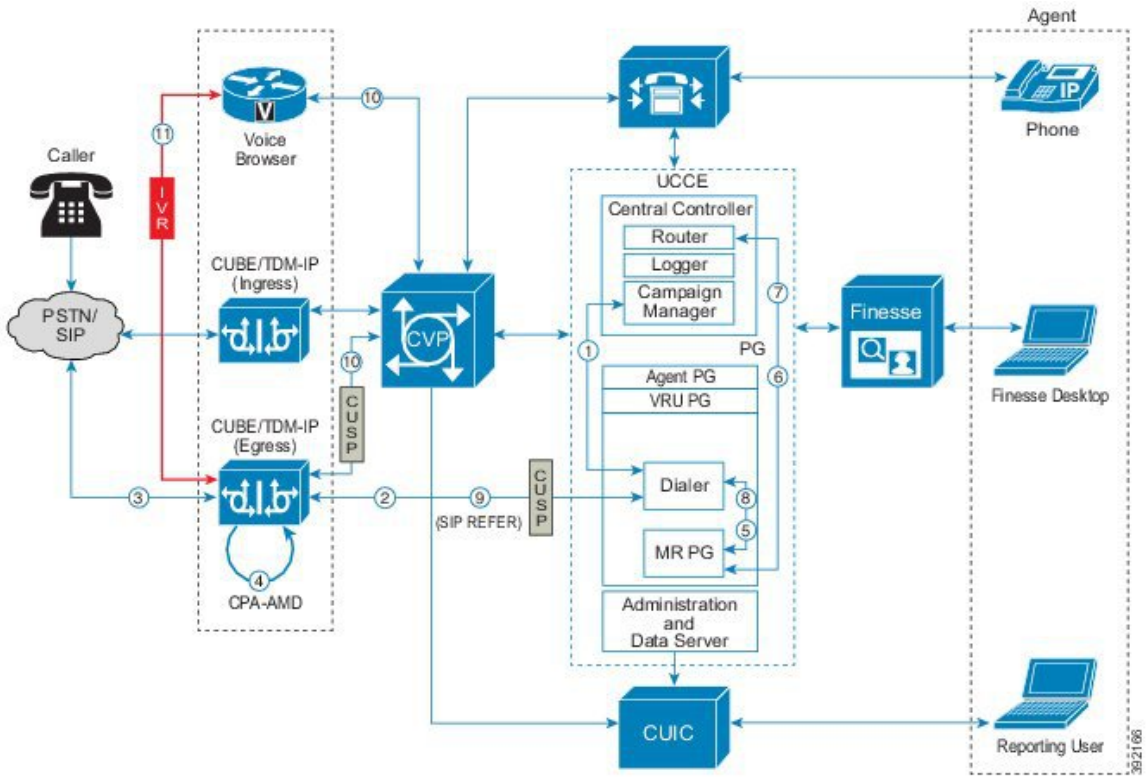
The following steps describe this call flow in detail:

1. The import is scheduled and the campaign starts. The records are delivered to the dialer.
2. The dialer looks for an available agent through the media routing interface.
3. The media routing peripheral gateway (MR PG) forwards the request to the router.
4. The routing script identifies an agent and responds to the MR PG.
5. The media routing PIM notifies the dialer that the agent is available.
6. The dialer signals the gateway to call the customer.
7. The gateway calls the customer, and the dialer is notified of the attempted call.
8. Call Progress Analysis (CPA) is done at the gateway.
9. When voice is detected, the dialer is notified.
10. The dialer asks the voice gateway using SIP REFER to transfer the call to the reserved agent by its agent extension.
11. The gateway directs the call to the agent through Unified CM (using dial peer configuration to locate the Unified CM).
12. Media are set up between the gateway and the agent's phone.

Call Flow Diagram for VRU Campaign

The following figure illustrates a transfer-to-VRU call flow in an Outbound Option deployment with a SIP dialer.

Figure 106: SIP Dialer Unattended VRU Campaign Call Flow



The following steps describe this call flow in detail:

1. An unattended VRU campaign starts, scheduling an import. Customer records are delivered to the dialer.
2. The dialer sends a SIP INVITE to the voice gateway to start a call to a customer.
3. The gateway places the customer call.
4. The voice gateway does Call Progress Analysis (CPA) and detects an answering machine (AMD). The dialer is notified.
5. The dialer sends a VRU route request to the MR PG.
6. The MR PG forwards the route request to the router and the routing script is invoked.
7. The router sends the route response with the network VRU label to the MR PG.
8. The MR PG forwards the route response to the dialer.
9. The dialer sends a SIP REFER request for the label to the voice gateway.
10. The voice gateway transfers the call to Unified CVP. CVP takes control of the call, handshakes with Unified CCE to get call context, and invokes the Voice Browser.
11. Media is set up between CUBE or the TDM-IP gateway and the Voice Browser.

Cisco Outbound Option Design Impacts

Follow these requirements when implementing Cisco Outbound Option:

- Configure abandon to VRU in agent-based campaigns. Telemarketing laws often require this behavior.
- Schedule large imports of the contact list and Do-Not-Call list during off-hours because the Campaign Manager runs on the same system as the Logger.
- Do not use Cisco IP Communicator softphone for agents configured for Cisco Outbound Option. IP Communicator can introduce an extra delay in transferring customer calls to the agent.
- An IPv6 client cannot import to Outbound Option.
- Finesse IP Phone Agent (IPPA) does not support Cisco Outbound Option.
- If you use the redundant Campaign Manager, Outbound Option Importer, and Database, your databases are larger:
    - Do Not Call records require more space. For comparison, 60 million DNC records require about 1 GB of extra disk space.

SIP Dialer Design Considerations

Cisco Outbound Option enables an agent to participate in outbound campaigns and take inbound calls through a SIP software dialer.

Follow these requirements when implementing the SIP Dialer:

- T1 PRI, E1 PRI and CUBE interfaces to the PSTN are supported for Outbound Option SIP dialers. BRI, FXO, E1R2 will not work with Dialer.
- Cisco Finesse supports Progressive, Predictive, Preview, and Direct Preview modes.
- For redundant SIP Dialers, use a Media Routing PIM on each redundant MR PG. One SIP Dialer is active while another SIP Dialer is in warm standby mode. One MR PIM is for each SIP Dialer. In a redundant MR PG environment, each PG side has only one PIM that connects to the local dialer when the Dialer becomes active.
- Use the G.711 codec in the dialer peer configuration of the gateway when the campaign configuration enables recording for the SIP Dialer.
- Enable SIP Dialer call throttling to prevent overloading the Voice Gateways.
- The Voice Gateway dial peers and CUSP routing policies are used for SIP Dialers to place outbound calls. This enables calls to be placed using gateways that are deployed to leverage toll-bypass and lower local calling rates.
- Configure CVP to send calls back to the gateway that they came from to reduce network DSP resource usage and to improve media transfer. This is important when the SIP Dialer and CVP share a Voice Browser that places outbound calls for VRU treatments.
- The Outbound Option Dialer uses IPv4 to place calls. Use IPv6 NAT at the voice gateway to translate the calls to IPv6.
- Although the SIP Dialer does not advertise the a-law codec, SIP Dialers with CUBE support a-law with specific design considerations. This deployment uses DSP resources on CUBE during the initial negotiation

(no media) between the SIP Dialer and the SIP service provider. During a REFER from the Dialer to the agent, CUBE renegotiates the codec with the agent's endpoint to use a-law. CUBE then releases the Transcoder.

Note The CUBE allocates a DSP for each outbound dialer call, whether or not the CPA is enabled.

Outbound Option Deployments

The SIP Dialer offers high scalability by offloading call process resources and call progress analysis to the gateway. Furthermore, the SIP Dialer has no Unified CM or gateway proximity requirements.

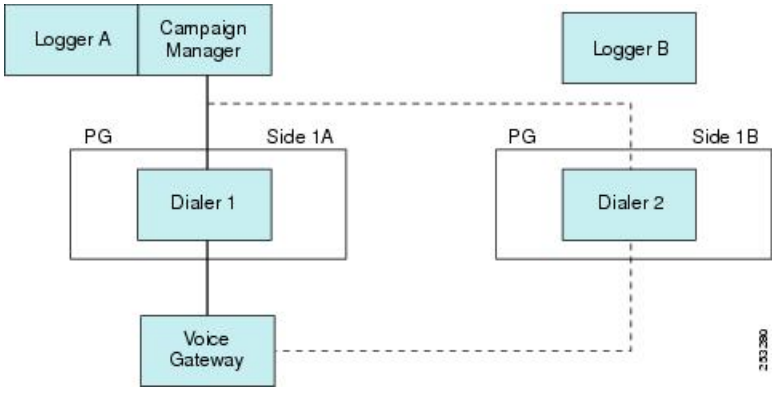
You can deploy the SIP dialer on the VM with the MR PG. Redundant MR PGs and Agent PGs are required. Run Outbound Option on a VM that meets the minimum requirements specified in the *Virtualization Wiki* for your solution.

The redundant Agent PG supports only redundant SIP Dialers; one dialer is active and another dialer is in warm-standby mode. For redundant SIP Dialer installations, each SIP Dialer connects to the MR PIM on the same MR PG side (Side A or Side B).

SIP Dialer with Single Gateway Deployment

This figure shows the installation of redundant SIP Dialers with a single Gateway. The Dialers are shown to be installed on Side A and Side B of the redundant PGs. The port capacity depends on the type of Cisco Voice Gateway deployed. This deployment model is used when scaling and high availability are not factors.

Figure 107: Single Gateway Deployment for SIP Dialer



Note This figure does not show the optional redundant Campaign Manager.

The SIP Dialer architecture supports only one active SIP Dialer per peripheral. Configure only one SIP Dialer. You install two Dialers on separate PG platforms, but you use the same Dialer Name.

For Unified CCE deployments, the SIP Dialer and Media Routing PG processes can run on a separate VM or on the same VM as the Agent PG. For a deployment with redundant SIP Dialers and MR PGs on the Agent PGs, each MR PG has one MR PIM that connects to the coresident SIP Dialer.

The SIP Dialer uses the local static route file to place and transfer outbound calls when **Sip Server Type** is set to **Voice Gateway** in the Dialer setup dialog. These outbound calls are transferred to CVP or outbound agents. Make sure that the SIP Dialer uses the local static route file for single gateway deployments.

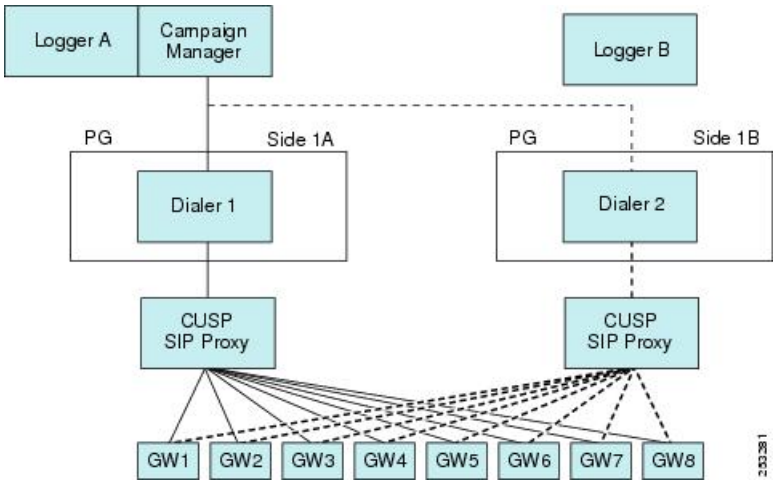
The SIP Dialer uses the Unified SIP Proxy server to place and transfer outbound calls when **Sip Server Type** is set to **CUSP Server** in the Dialer setup dialog. These calls are placed or transferred to CVP or outbound agents.

Note Codec configuration (G.729 versus G.711) impacts port capacity and CPU utilization of gateways. Configuring G.729 requires more DSP and CPU resources for gateways.

SIP Dialer with Multiple Gateways Deployment

The following figure shows the deployment model for Unified SIP Proxy and eight Voice Gateways. The active Dialer points to the Unified SIP Proxy server. The proxy handles load balancing and failover. The SIP Dialer supports Unified SIP Proxy on the Cisco 3845 Integrated Services Router.

Figure 108: Multiple Gateway Deployment for SIP Dialer



Note This figure does not show the optional redundant Campaign Manager.

In a multiple gateway deployment, the SIP Dialer requires Server Group and Route Table configurations on Unified SIP Proxy servers to identify the gateways. It also requires numbers so that the gateways can transfer customer calls to CVP or agents for the Dialer. Setting the **Sip Server Type** radio button to **SIP Proxy** in the Dialer setup dialog is required for multiple gateway deployment.

Outbound Option and Clustering Over the WAN

The deployment model for clustering Unified CCE over the WAN allows for improved high availability by deploying redundant components on the other end of the WAN. The "Single Campaign Manager, Importer, and Database" high-availability model differs from the model for clustering over the WAN. When deploying Outbound Option with clustering over the WAN, keep in mind that you only gain benefits with redundant Outbound Option components.

Distributed Deployments of Outbound Option

A distributed deployment model involves a central Unified CCE system and Unified Communications Manager cluster located at one site, with the Campaign Manager installed on the logger at this site, and a second site reachable over a WAN, which consists of the dialer, a PG, and a second cluster with Cisco Outbound Option.

For SIP Dialer deployment, a Unified SIP Proxy server is installed for one SIP Dialer on each PG side, and the Side A/Side B Dialer is targeting the same set of Voice Gateways through its own Unified SIP Proxy server. Multiple Voice Gateways can be installed locally to customer phones, or each Voice Gateway can be installed locally to an area so that tolls are not encountered if leased circuits or IP MPLS WAN circuits are available.

The Campaign Manager sends dialer records over the WAN, and the dialer places calls to local customers. The second site would support inbound agents as well.

The following bandwidth options are available between India and the US in customer environments:

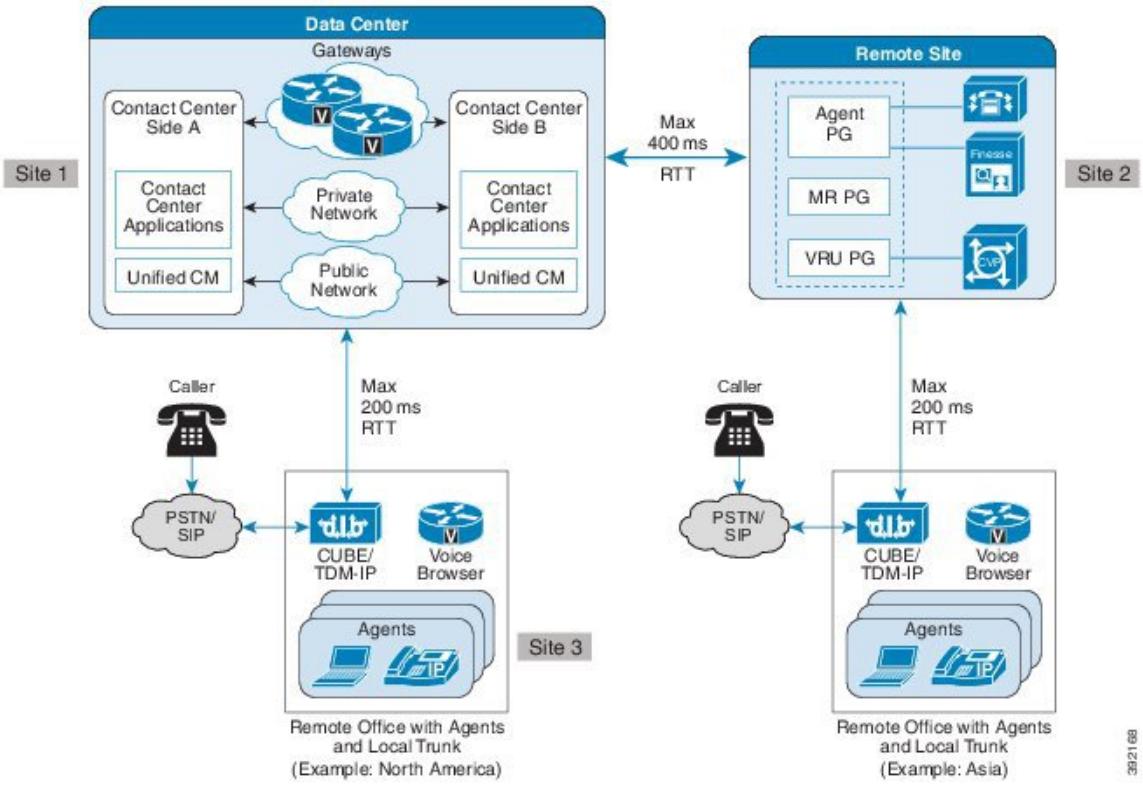
1. Terrestrial P2P leased 2 Mbps circuits
2. Terrestrial P2P DS3 (44 Mbps) leased circuits
3. IP MPLS WAN circuits. Varying speeds are available from the service provider depending on customer needs. Typical usage is 44 Mbps.
4. The service provider hands off PRI (E1) trunks to India. The WAN cloud is usually built on SIP by the service provider. The service provider converts TDM to IP at the ingress/egress point in the United States and converts IP to TDM in India.

Options 1 and 2 above are the most common. Option 3 is becoming more popular with outsourcers because the MPLS cloud can connect to several of their customers. For example, the diagrams in the following sections show that the Outbound Contact Center System is deployed across multiple sites in the United States and India for various agent-based campaigns or transfer to a VRU campaign. The customers are in one country; for example, in the United States.

Distributed Deployment for Agent-Based Campaigns

This figure shows an example of a distributed deployment for an agent-based campaign.

Figure 109: Distributed Deployment Example for Agent-Based Campaign



In this distributed deployment example for an agent-based campaign:

- The Voice Gateway and Router and Logger A servers are distributed between two sites (Site 1 and Site 3) in the United States.
- The Unified Communications Manager cluster is located at Site 2 in India along with the Agent PG.
- The redundant MRPG/Dialer and redundant Agent PGs are installed on the same VM at Site 2 in India.
- The SIP Dialer at Site 2 uses the Voice Gateways that are located at Site 3 in the United States.
- The Voice Gateways are included in the diagram with CT3 interface at Site 3 in the United States. These routers provide 1:1 redundancy for Dialer calls.
- The Unified SIP Proxy servers are locally redundant at Site 2 to avoid the WAN SIP signaling traffic for transfering live outbound calls.
- Each SIP Dialer connects to its own Unified SIP Proxy server at Site 2.
- Each Unified SIP Proxy server controls the set of Voice Gateways at Site 3 in the United States.
- Each Unified SIP Proxy server controls the set of Voice Gateways at Site 3 in the United States.

If recording is enabled at the SIP Dialer, the bandwidth requirements are as follows:

- Answered outbound calls require the following bandwidth for each agent call:
  - G.711 Codec calls require a WAN bandwidth of 80 kbps

- G.729 Codec calls require a WAN bandwidth of 26 kbps

- Alerting outbound calls require the following bandwidth for each agent call:
  - G.711 Codec calls require a WAN bandwidth of 80 kbps
  - G.729 Codec calls require a WAN bandwidth of 26 kbps

Sizing for Outbound Option

Cisco Outbound Option can run fully blended campaigns in which agents can handle inbound and outbound calls alternately.

Note See the chapter on configuration limits and feature availability for other limits that impact sizing.

When sizing your deployment, do not use the maximum outbound agents on a PG without also looking at expected hit rate, lines dialed per agent, and average handle times.

SIP Dialer targets the support of 1000 outbound agents for one PIM per PG. The number of supported agents is smaller when deploying mobile agents. To support this number of agents, the deployment must have at least five high-end gateways dedicated to outbound dialing.

SIP Dialer can support 3000 ports and 60 calls per second (CPS).

Each port can dial two calls per minute, assuming an average 30 seconds per call attempt, so 30 ports can handle one call per second for the Dialer. If the time to get all ports busy exceeds the average port busy time, then some ports are always idle.

Dialer Port Calculations

The following formula can be used to calculate the number of dialer ports that are required to achieve targeted call rate:

```
Number of Ports = [target call rate * average call duration * (1 + hit rate %)]
```

This table shows the required ports to achieve targeted outbound call rates. These figures assume an average of 30 seconds per outbound call and a 20% hit rate.

Table 47: Ports Required to Achieve Targeted Outbound Call Rates

Targeted outbound calls per second	Number of ports required
10	360
20	720
30	1080

Voice Gateway Considerations

The most powerful Voice Gateway supports about 12 calls per second, even under the most favorable conditions. Five gateways can support an aggregate spike of up to 60 calls per second when evenly distributed. However, even distribution does not account for occasions when ports are tied up with agent or VRU calls after the

transfer. So assuming a 50% transfer rate and using a conservative estimate, eight Voice Gateways are required to support a spike of up to 60 calls per second.

For the most current information about Voice Gateway models and releases that the SIP Dialer supports, see the *Compatibility Matrix* for your solution.

For gateway sizing considerations, see the published Cisco gateway performance data.

Agent PG Considerations

The Unified Communications Manager PIM can support up to 15 calls per second.

If the voice hit rate for the campaign is 15%, then the PG can sustain dialing at a rate of 100 calls per second.

Unified CM Considerations

The Unified CM subscriber supports a certain rate of outbound calls per second. To support a larger CPS at the Agent PG, distribute the Dialer across multiple subscribers using a Unified SIP Proxy server.

CUSP Considerations

A typical outbound call requires two transactions, if the call is transferred to an agent or VRU. A typical outbound call requires one transaction, if the call is not transferred to an agent or VRU.

CVP Considerations

Calls can be distributed to Unified CVP using translation routes. Any load balancing across Unified CVPs happens in the routing script.

Since four SIP Proxy transactions are required for some outbound call scenarios with Unified CVP, give Unified CVP its own Unified SIP Proxy server in large-scale deployments.

Mobile Agent Considerations

The SIP Dialer supports 500 unified mobile agents per Agent PG. With the SIP Dialer solution, the outbound calls have the same impact on Unified Communications Manager as inbound calls. Maintain a 2:1 ratio for number of inbound agents versus outbound agents. Since the SIP Dialer solution supports 1000 outbound regular agents per Agent PG, the SIP Dialer supports 500 outbound mobile agents per Agent PG.

SIP Dialer Throttling

In a single or multiple gateway deployment, the SIP Dialer raises an alarm if any gateway is overloaded. If you enable the autothrottle mechanism, the dialer also automatically throttles the dialing rate of overloaded gateways down to 10 percent of the configured port throttle value per 5000 customer attempts until 50 percent of the correction is met. 50 percent of the correction means that the SIP Dialer stops autothrottling when it reaches 50 percent of the configured port throttle value.

Note The autothrottle mechanism is disabled by default. To automatically throttle overloaded gateways, enable the autothrottle mechanism by setting the value of registry key **EnableThrottleDown** to 1.

The SIP Dialer always raises an alarm when a gateway is overloaded, even when the autothrottle mechanism is disabled.

You can control SIP Dialer throttling with the field **Port Throttle** in the dialer configuration. Port Throttle indicates the number of ports to throttle per second. Setting the value to Port Throttle = 5 allows SIP Dialer to dial outbound calls at a rate of five calls per second per Dialer.

When the SIP Dialer connects to the Voice Gateway directly in the deployment, limit the dialer port throttle by the maximum dialer call setup rate listed on the gateway sizing table.

When the SIP Dialer connects through the CUSP in the deployment, the port throttle setting on the dialer must not exceed the total gateway capacity under assumption. Calls are load-balanced through CUSP and each gateway reaches its maximum available capacity. Limit the port throttle by the CUSP maximum transaction. Currently, the dialer maximum throttle setting is 60 calls per second. Under general transfer rate, calls through CUSP do not exceed maximum CUSP transaction rate given that CUSP is exclusively used by outbound deployments.

Set the port throttle value to 5 for Cisco 2800 Series Integrated Services Routers, set the port throttle value to 15 for Cisco 3800 Series Integrated Services Routers, and set this value to 20 for Cisco Access Servers and Universal Gateways.

Single Gateway Deployment

Use the following formula to calculate the Port Throttle if the gateway is dedicated 100% for outbound campaigns:

```
Port Throttle = (Value for Gateway)
```

Use the following formula to calculate the Port Throttle if the gateway is shared by multiple SIP Dialers for outbound campaigns:

```
Port Throttle = (Value for Gateway) / (Number of SIP Dialers)
```

Use the following formula to calculate the Port Throttle if the gateway is shared by multiple components (Unified CM, Unified CVP, and SIP Dialer) for inbound and outbound calls:

```
Port Throttle = (Value for Gateway) * (Percentage of outbound calls) * (1 – Hit Rate)
```

Multiple Gateway Deployment

Use the following formula to calculate the Port Throttle if the gateways are dedicated 100% for outbound campaigns:

```
Port Throttle = Total Values for Gateways
```

Use the following formula to calculate the Port Throttle if the gateways are shared by multiple SIP Dialers for outbound campaigns:

```
Port Throttle = (Total Values for Gateways) / (Number of SIP Dialers)
```

Use the following formula to calculate the Port Throttle, if the gateways are shared by multiple components (Unified CM, Unified CVP, and SIP Dialer) for inbound and outbound calls:

```
Port Throttle = (Total Values for Gateways) * (Percentage of outbound calls) * (1 – Hit
Rate)
```

The throttling mechanism in the SIP Dialer process is not aware of which gateway the Unified SIP Proxy server selects to place outbound calls. Calculate the appropriate weight for each gateway in the Server Group configuration of the Unified SIP Proxy server for the load balance.

```
Weight = (Value for Gateway) / (Port Throttle) * 100
```

For example, assume a Cisco 3800 Series Gateway (192.168.10.3) and a Cisco 2800 Series Gateway (192.168.10.4) are used in a multiple gateway deployment. The following configuration allows that 3800

Series gateway in the cucm.example.com server group to receive 75 percent of the traffic and the 2800 Series gateway to receive 25 percent.

```
netmod(cusp-config)> server-group sip group cucm.example.com enterprise
netmod(cusp-config-sg)> element ip-address 192.168.10.3 5060 tls q-value 1.0 weight 75
netmod(cusp-config-sg)> element ip-address 192.168.10.4 5060 tls q-value 1.0 weight 25
netmod(cusp-config-sg)> lbtype weight
netmod(cusp-config-sg)> end server-group
```

SIP Dialer Recording

The SIP Dialer can record ("Recording") or enable the recording of Call Progress Analysis by third-party applications ("Media Termination") to be used for CPA troubleshooting. It does not record the full conversation.

There is usually no media stream between the SIP Dialer and the Voice Gateway. But when the recording or media termination is enabled in the Campaign configuration, the SIP Dialer requests the Voice Gateways to send the media stream to the SIP Dialer. The media stream is in G.711 or G.729 codec, depending on the dial peer configuration on the Voice Gateway. The SIP Dialer can record the media stream only with G.711 codec, but it can receive media streams for both G.711 and G.729 codecs to allow a third recording server to perform SPAN-based recording for outbound calls.

When “Recording” is enabled in the Campaign configuration, the SIP Dialer receives media streams, decodes RTP packets in G.711 codec, and writes them into a recording file. The SIP Dialer will send an alarm if the media stream is G.729 codec. The SIP Dialer has been tested to be able to support a maximum of 100 recording sessions per Dialer server due to CPU resource and disk I/O limitations.

When “Media Termination” is enabled in the Campaign configuration, the SIP Dialer will only receive the media stream to allow a third-party recording server to perform SPAN-based recording.

There is a limit for Media Termination Sessions because of a thread resource limitation per process. The SIP Dialer has to create a thread to listen on the media stream. The current limit for Media Termination Sessions is 200.

The SIP Dialer uses the following Registry keys to allow users to manage recording sessions and disk space:

Table 48: SIP Dialer Registry Keys

Name	Data Type	Description	Default Value
MaxRecordingSessions	DWORD	The maximum recording sessions per SIP Dialer, if the recording is enabled in the Campaign configuration.	100
MaxMediaTerminationSessions	DWORD	The maximum media termination sessions per SIP Dialer, if the recording is enabled in the Campaign configuration.	200
MaxAllRecordFiles	DWORD	The maximum recording file size (bytes) per SIP Dialer.	500,000,000
MaxPurgeRecordFiles	DWORD	The maximum recording file size (bytes) that SIP Dialer will delete when the total recording file size, MaxAllRecordFiles, is reached.	100,000,000

Outbound Option Bandwidth, Latency, and QoS Considerations

In many Outbound Option deployments, all components are centralized; therefore, there is no WAN network traffic to consider.

For some deployments, if the outbound call center is in one country (for example, India) and the customers are in another country (for example, US), then consider the WAN network structure under the following conditions:

- In a distributed Outbound Option deployment, when the Voice Gateways are separated from the Outbound Option Dialer servers by a WAN.
- When using Unified CVP deployments for transfer to a VRU campaign, and the Unified CVP servers are separated from the Outbound Option Dialer servers by a WAN. Provide Unified CVP with its own Cisco Unified SIP Proxy Server in the local cluster to reduce the WAN traffic.
- When deploying a SIP Dialer solution for transfer to a VRU campaign, and the Cisco Unified SIP Proxy Servers for the SIP Dialers are separated from the Outbound Option Dialer servers by a WAN.
- When the third-party Recording Server is separated from the Outbound Option Dialer servers by a WAN, configure the recording server local to the Voice Gateways.

Adequate bandwidth provisioning is an important component in the success of the Outbound Option deployments.

Impact of Redundant Campaign Manager, Outbound Option Importer, and Database

When using these redundant components, consider the following points:

- Certain deployments can have increased WAN network traffic.
- Messaging and record replication increases the bandwidth used between Side A and Side B.

Distributed SIP Dialer Deployment

SIP is a text-based protocol; therefore, the packets used are larger than some protocols. The typical SIP outbound call flow uses an average of 12,500 bytes per call that is transferred to an outbound agent. The average hit call signaling bandwidth usage is:

Hit Call Signaling Bandwidth = (12,500 bytes/call) (8 bits/byte) = 100,000 bits per call = 100 Kb per call

The typical SIP outbound call flow uses about 6,200 bytes per call that is disconnected by the outbound dialer. Those outbound calls can be the result of a busy ring no-answer, an invalid number, and so forth. The average non-hit call signaling bandwidth usage is:

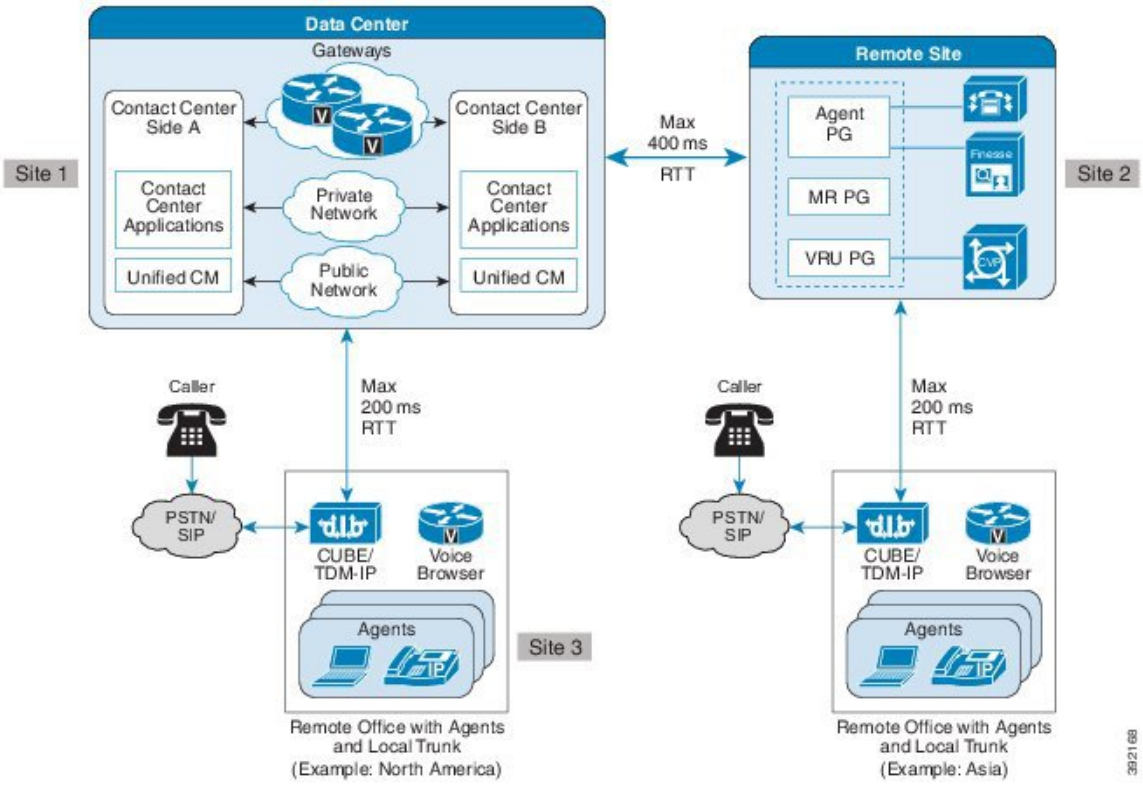
Non-Hit Signaling Call Bandwidth = (6,200 bytes/call) (8 bits/byte) = 49,600 bits per call = 49.6 Kb per call

Codec Bandwidth = 80 Kbps per call for G.711 Codec,
or 26 Kbps per call for G.729 Codec

Agent-Based Campaign - No SIP Dialer Recording

This figure shows an example of the distributed Outbound SIP Dialer deployment for an agent-based campaign.

Figure 110: Distributed Outbound SIP Dialer Deployment for an Agent-Based Campaign



The average WAN bandwidth usage in this case is:

```
WAN Bandwidth = Calls Per Second * (Hit Rate * (Codec Bandwidth * Average Call Duration
+ Hit Call Signaling Bandwidth) + (1 – Hit Rate) * Non-Hit Call Signaling Bandwidth) = Kbps
```

Example 1

With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the agent-based campaign, and a WAN link with G.711 codec and average call duration of 40 seconds, the bandwidth usage is:

```
60 * (20% * (80 * 40 + 100) + (1 – 20%)*49.6) = 41980.8 kbps = 41.98 Mbps
```

Example 2

With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the agent-based campaign, and a WAN link with G.729 codec and average call duration of 40 seconds, the bandwidth usage is:

```
60 * (20% * (26 * 40 + 100) + (1 – 20%)*49.6) = 16060.8 kbps = 16.06 Mbps
```

Agent-Based Campaign - SIP Dialer Recording

The average WAN bandwidth usage in this case is:

```
WAN Bandwidth = Calls Per Second * (Codec Bandwidth * Average Call Duration + Hit Rate
* Hit Call Signaling Bandwidth + (1 - Hit Rate) * Non-Hit Call Signaling Bandwidth) = Kbps
```

Example 3

With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the agent campaign, and a WAN link with average G.711 codec and average call duration of 40 seconds, the bandwidth usage is:

```
60 * (80 * 40 + 20% *100 + (1 – 20%)*49.6) = 199180.8 kbps = 199.18 Mbps
```

Example 4

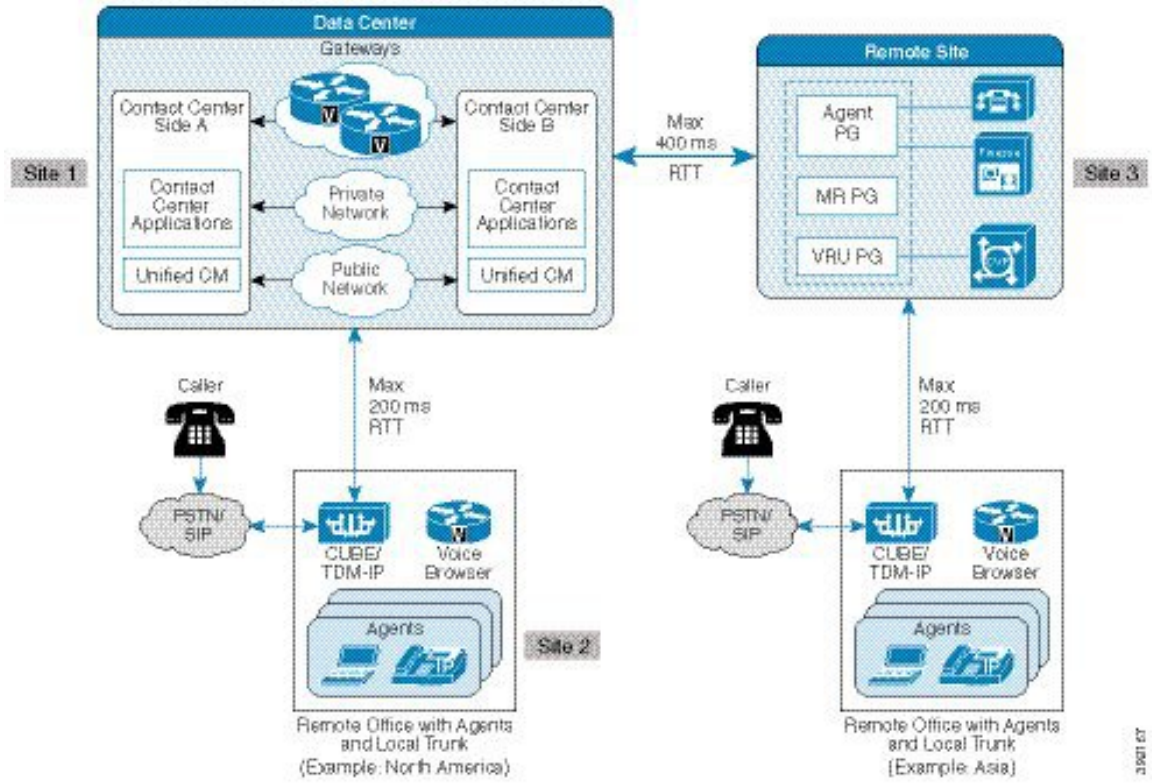
With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the agent campaign, and a WAN link with average G.729 codec and average call duration of 40 seconds, the bandwidth usage is:

```
60 * (26 * 40 + 20% *100 + (1 – 20%)*49.6) = 67660.8 kbps = 67.66 Mbps
```

Transfer-To-VRU Campaign - No SIP Dialer Recording

The following figures show examples of the distributed Outbound SIP Dialer deployment for transfer to a VRU campaign.

Figure 111: Distributed Outbound SIP Dialer Deployment for Transfer to a VRU Campaign with CVP



The average WAN bandwidth usage in this case is:

```
WAN Bandwidth = Calls Per Second * Hit Rate * Hit Call Signaling Bandwidth + Calls Per
Second
* (1 - Hit Rate) * Non-Hit Call Signaling Bandwidth = Kbps
```

Example 5

With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the transfer-to-IVR campaign, and a WAN link with G.711 codec, the bandwidth usage is:

```
60 * 20% * 100 + 60 *(1 – 20%)*49.6= 3600 kbps = 3.6 Mbps
```

Transfer-To-VRU Campaign - SIP Dialer Recording

The average WAN bandwidth usage in this case is:

```
WAN Bandwidth = Calls Per Second * (Codec Bandwidth * Average Call Duration + Hit Rate
* Hit Call Signaling Bandwidth + (1 - Hit Rate) * Non-Hit Call Signaling Bandwidth) = Kbps
```

Example 6

With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the agent campaign, and a WAN link with G.711 codec and average call duration of 40 seconds, the bandwidth usage is:

```
60 * (80 * 40 + 20% *100 + (1 – 20%)*49.6) = 199180.8 kbps = 199.18 Mbps
```

Example 7

With call throttling of 60 cps on the SIP Dialer, a 20% hit rate for the transfer-to-VRU campaign, and a WAN link with G.729 codec and average call duration of 40 seconds, the bandwidth usage is:

```
60 * (26 * 40 + 20% *100 + (1 – 20%)*49.6) = 67660.8 kbps = 67.66 Mbps
```

Courtesy Callback Considerations

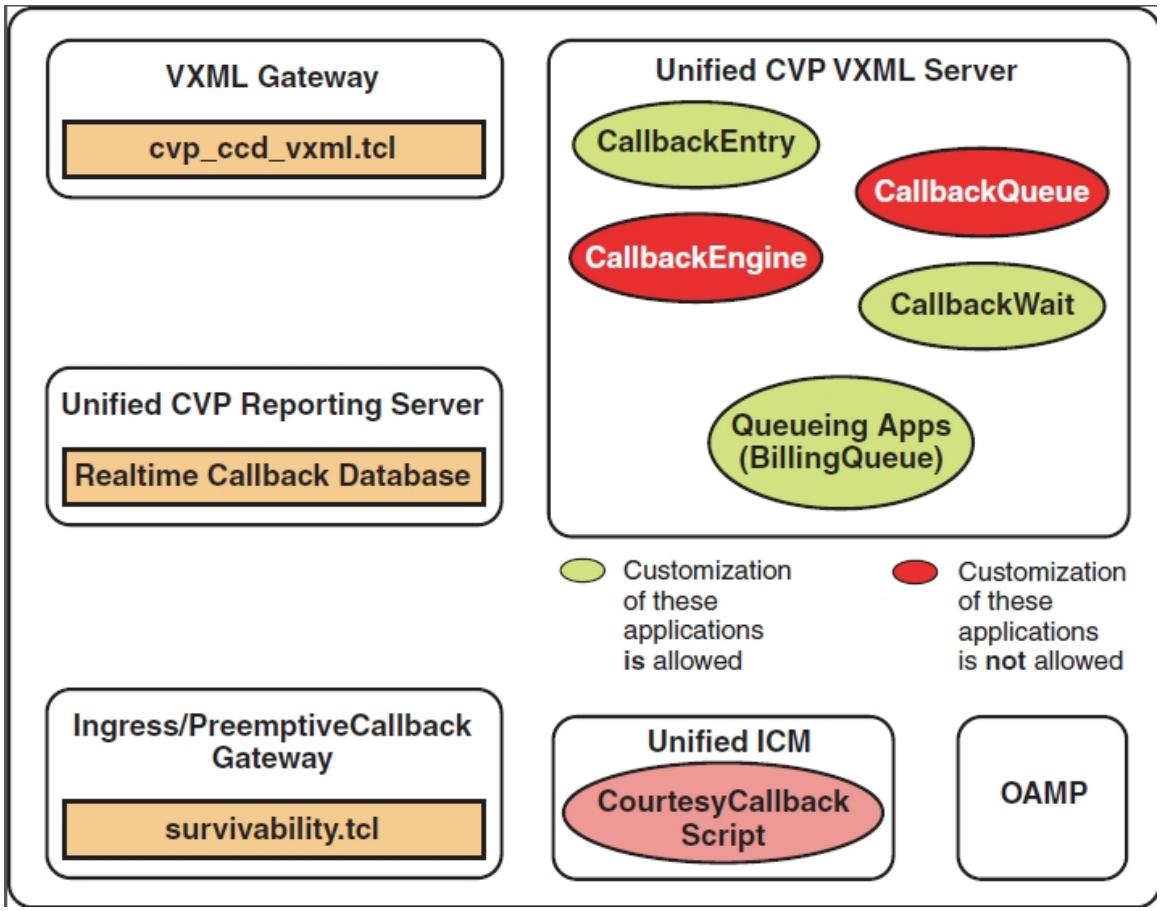
Courtesy Callback reduces the time callers have to wait on hold or in a queue. Your solution can call back callers who meet your criteria, instead of having them wait on the phone for an agent. The caller who has been queued by Unified CVP can end the call. The solution then calls them back when an agent is close to becoming available (preemptive callback).

Preemptive callback does not change the time that a caller waits for an agent. It enables the caller to end the call and not remain in queue listening to music. Callers who remain in queue and those who opt for the callback treatment appear the same to agents answering the call.

Note Scheduling a callback to occur at a specified time is not part of this feature.

Figure 112: Courtesy Callback Components

This figure shows the components used for the Courtesy Callback feature.



Note Do not allow the caller to invoke the Courtesy Callback applications more than once for the same call on the VXML Server.

Courtesy Callback uses the TCL service on IOS Voice Gateway and a built-in feature on Cisco VVB.

Courtesy Callback Use Case

In your callback script, you can establish criteria for offering a caller courtesy callback. These are examples of callback criteria that you can establish:

- Expected wait for a customer in queue exceeds some maximum number of minutes, based on your average call handling time per customer.

Note The included sample scripts use this method for determining callback eligibility.

- Assigned status of a customer. You can offer gold customers the opportunity to be called back, instead of remaining on the line.

- The particular service that a customer requests. You can establish sales calls or system upgrades as callback criteria.

Courtesy Callback Call Flows

If the caller opts for a callback, they leave their name and phone number. Their request remains in the system. The system places a callback to the caller when the Estimated Wait Time (EWT) reaches the correct value. The caller answers the call and confirms that they are the original caller, and the system connects the caller to the agent after a short wait.

Note Courtesy Callback is also supported for IP-originated calls.

A typical call flow for this feature follows this pattern:

1. The call arrives at Unified CVP and the call is treated in the VRU environment.
2. The Call Studio and Unified CCE Courtesy Callback scripts determine if the caller is eligible for a callback based on your rules.
3. If the caller is eligible, the system announces the EWT and offers the caller a callback when an agent is available.
4. The caller chooses what to do:
   a. If the caller chooses not to use the callback feature, queuing continues.
   b. If the caller chooses to receive a callback, the system prompts the caller to record their name and to key in their phone number.
5. The system writes a database record to log the callback information.

Note If the database is not accessible, the system does not offer a callback to the caller.

6. The caller disconnects from the TDM side of the call. However, the IP side of the call in Unified CVP and Unified CCE is still active. This keeps the call in the same queue position. No queue music plays, so Voice Browser resources used during this time are less than for a caller actually in the queue.
7. When an appropriate agent is close to being available (as determined by your callback scripts), then the system calls the person back. The system announces the recorded name when the callback is made to ensure that correct person accepts the call.
8. A VRU session asks the caller to confirm that they are the correct person and that they are ready for the callback.

   If the system cannot reach the callback number (for example, busy lines, RNA, or network problems), then the call is not sent to an agent. The call also does not go to the agent if the caller does not confirm that they are the correct person. The agent is guaranteed that someone is waiting when they take the call. The system assumes that the caller is already on the line by the time the agent gets the call.

This feature is called preemptive callback because the system assumes that the caller waits a minimal time for the agent and the caller is on the line when the agent answers.

9. The system presents the call context on the agent screen-pop.

If the system cannot reach the caller after a configurable number and frequency of retries, the callback cancels and the database status updates appropriately. You can run reports to determine if any manual callbacks are necessary based on your business rules.

See the *Configuration Guide for Cisco Unified Customer Voice Portal* at https://www.cisco.com/en/US/products/sw/custcosw/ps1006/products_installation_and_configuration_guides_list.html for a call flow description of the scripts providing the Courtesy Callback feature.

Courtesy Callback Design Impacts

Consider the following design impacts for Courtesy Callback feature:

- The callback uses the same Ingress Gateway through which the call arrived. The outbound calls cannot be made on any other Egress Gateway.
- Queue calls that allow callback on a Unified CVP VXML Server.
- Courtesy Callback requires the Unified CVP Reporting Server.
- Answering machine detection is not available for this feature. During the callback, the caller is prompted with a brief VRU session message and acknowledge with DTMF that they are ready to take the call.
- Calls that are transferred to agents using DTMF *8, TBCT, or hookflash cannot use Courtesy Callback.
- Courtesy Callback doesn't support Agent call transfers to the CCB Queue, over a computer telephony integration (CTI) route point.
- Callbacks are a best-effort function. After a limited number of attempts to reach a caller during a callback, the callback is terminated and marked as failed.
- Configure the allowed or blocked numbers that Courtesy Callback uses to place calls through the Unified CVP Operations Console .
- The media inactivity detection feature on the Voice Browser can affect waiting callback calls. For more information, see the *Configuration Guide for Cisco Unified Customer Voice Portal*.
- Courtesy Callback call flow routed via CUCM is not supported, even though they originate at CUBE.
- Courtesy Callback requires an accurate EWT calculation for its optimal behavior.

  Do the following to optimize the EWT, when using Precision Queues for Courtesy Callback:

  - Queue the calls to a single Precision Queue
  - Do not include a `Consider If` expression when you configure a step.
  - Do not include a wait time between steps or use only one step in the Precision Queue.

Note Use simple Precision Queue definitions (for example, with one step and one-to-one agent mapping). The complexity of Precision Queues makes calculating accurate EWT difficult.

- Courtesy Callback supports 900 calls with 10 CPS. One reporting server can be configured to support 900 CCB calls simultaneously with standard reporting enabled.

Note CCB does not support the use of SRTP.

Callback Time Calculations

The following sections provide an overview of how callback time is determined.

These are some definitions of key terms used:

- **Wait Time**—The interval of time between when the call enters the queue and when the call leaves the queue.
- **Reconnect Time**—The interval between when the callback starts and when the caller accepts the callback and is waiting for an agent.
- **Callback in Queue Time**—The interval between when the caller reconnects and when the call leaves the queue.
- **Service Level Agreement (SLA)**—Average of Callback in Queue Time. Average means that roughly 50 percent of calls are within the service level and 50 percent are outside the service level.
- **Average Dequeue Time**—The average number of seconds that it takes for a call to leave the queue.
- **Remaining Time**—The number of seconds left to count down to call back the caller.

Callback in Queue Time

The average Callback in Queue Time after a callback is based on an agreed service level. Courtesy Callback also avoids calling back too early or too late, as both scenarios are undesirable. If callers are called back too early, they are more likely to have to wait in the queue for a longer time. If the callback is made too late, there is a greater chance that your agents could be idle and waiting for calls.

The remaining time changes when the dynamics of a call center change. Such changes include when more or fewer agents are available or when the average handle time changes. Courtesy Callback calculates the Average Dequeue Time based on various factors, such as calls in queue, average handle time, and agents in ready and talking states.

The Average Dequeue Time updates when a call enters the queue and when it leaves the queue. Calculations use this information to reduce the Callback in Queue Time and minimize times when your agents wait for calls.

Process Details and Calculation Methods

Courtesy Callback uses the following formula to determine the Average Dequeue Time and to update the remaining time for all Courtesy Callback calls in the queue.

Note Courtesy Callback can support a default wait time of 30 minutes with a maximum exception of 90 minutes.

Average Dequeue Time Calculation

The Average Dequeue Time (D) is calculated using the formula:

```
D = (EWT + F)/N
```

Where:

- *EWT* is the estimated wait time for a new Courtesy Callback call.
- *F* is the number of seconds that the first call is already in position in the queue.
- *N* is the number of calls in queue.

Note The Dequeue Time plays a significant role in the optimal behavior of the Courtesy Callback feature. The average Dequeue Time is calculated based on factors such as call volume, agent availability, and the average handle time for a particular skill group.

The Estimated Wait Time (EWT) is an approximation. The uniformity of average handling time and agent availability for a particular skill group drive its accuracy. If these factors are not uniform, it leads to a difference between the announced wait time and the actual callback time. The use of microapps can insert calls into the queue that were not included in the EWT calculation. For scripting of calls that include Courtesy Callback, queue all calls on the VRU using VxmlScripting, instead of microapps.

Remaining Time Calculation

The remaining time for a callback in the queue is calculated using this formula:

```
R(p) = p*D - F - C
```

Where:

- *p* is the current queue position of the call from 1 to N.
- *R(p)* is the remaining time for the Pth queue position Courtesy Callback call.
- *C*, the post-callback time, is the sum of the time to get the Courtesy Callback caller back on the phone and the SLA time.

Note

- With time in first place (*RPT.ewtWithFirstInQueueTime* in *reporting.properties*) = true,

  The remaining time is calculated as:

  ```
  R(p) = p*(EWT + F)/ N - F - C
  ```

- With time in first place (*RPT.ewtWithFirstInQueueTime* in *reporting.properties*) = false (default),

  The remaining time is calculated as:

  ```
  R(p) = p*(EWT)/ N - F - C
  ```

Example Scripts and Audio Files

This feature uses Unified CCE scripts. Modifiable example scripts are provided on the Unified CVP install media in `\CVP\Downloads and Samples\`. These scripts determine whether to offer a callback to the caller. The files provided are:

- `CourtesyCallback.ICMS`, the Unified CCE script
- `CourtesyCallbackStudioScripts.zip`, a collection of Call Studio scripts

Sample audio files for these scripts are installed to `<CVP_HOME>\OPSConsoleServer\CCBDownloads\CCBAudioFiles.zip` and also as part of the Media Files installation option.

If you use `CCBAudioFiles.zip`, unzip the contents onto the media server. `CCBAudioFiles.zip` has Courtesy Callback-specific application media files under `en-us\app` and media files for **Say It Smart** under `en-us\sys`. If you already have media files for **Say It Smart** on your media server, then you only require the media files under `en-us\app`.

Note The default prompts work for most of the default Call Studio scripts. Review and provision the **Say It Smart** plugin prompts for specific cases that the default prompts do not cover.

The sample scripts use the default location of `http://<server>:<port>/en-us/app`. Change the default location of the sample audio files in the sample scripts for your environment. (That is, substitute the media server IP address and port in <server> and <port>).

The following example scripts are provided:

- **BillingQueue**—This script plays queue music to callers that either choose not to have a callback or who reenter the queue after receiving a callback. You may customize this script to suit your business needs.
- **CallbackEngine**—This script keeps the VoIP leg of a callback alive between when a caller opts for a callback and when a caller receives the callback.

Important Do not customize this script.

- **Callback Entry**—This script handles the initial VRU when a caller enters the system and provides the caller the opportunity to receive a callback. You may customize this script to suit your business needs.
- **CallbackQueue**—This script handles the keepalive function of a call while a caller is in queue and listening to the music.

Important Do not customize this script.

- **CallbackWait**—This script handles the VRU portion of a call when a customer is called back. You may customize this script to suit your business needs.

Note The Courtesy Callback sample files and scripts are also available on DevNet (**Customer Voice Portal (CVP)** > **Downloads** > **Courtesy Callback Sample Scripts**) at https://developer.cisco.com/site/customer-voice-portal/ .

Call Context Considerations

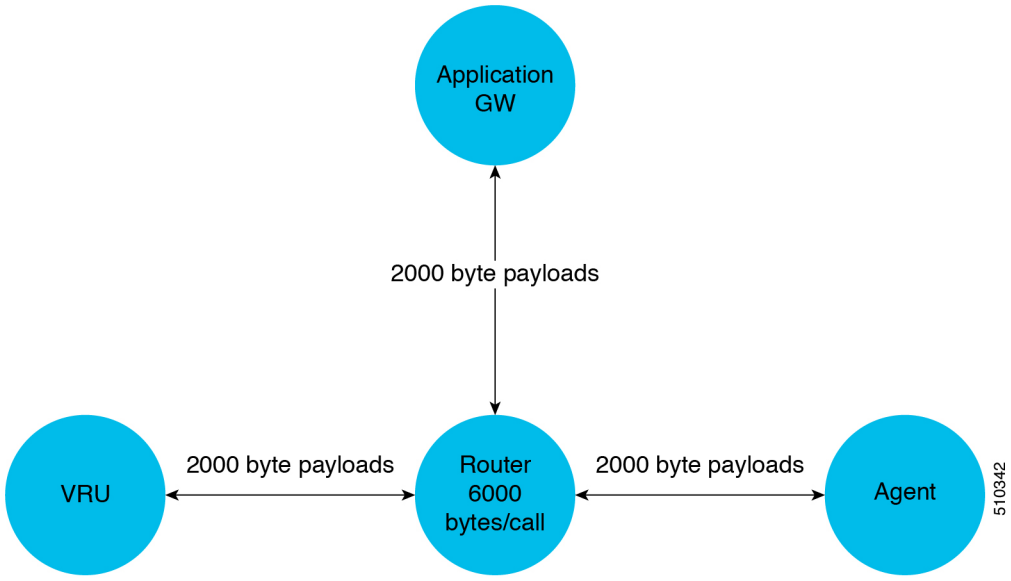
Expanded Call Context Variable Considerations

Expanded Call Context (ECC) variables enable you to set business relevant data for transfer to the agent desktop. Unlike the call variables, you can configure the size, format, and the name of each ECC variable. You use an *ECC payload* to pass defined sets of ECC variables, the *members* of that ECC payload. An ECC payload holds 2,000 bytes. The Router can store approximately 6,000 bytes of ECC variables for each call. If you exceed the system-wide limit on ECC variable contents, the solution generates events to warn you.

The solution includes an ECC payload named "Default" for backward compatibility. If your solution does not require more ECC variable space, you only need the Default payload. If your solution only has the Default payload, the solution automatically adds any new ECC variables to the Default payload until it reaches the 2000-byte limit.

Note You cannot delete the Default payload. But, you can change its members. CVP, Outbound Option, Multi-Channel, and the ECC variables consume some of the space in the Default payload.

Create your ECC payloads to carry the ECC variables for a particular interface:



Note For ECC payloads to a CTI client, the size limit is 2,000 bytes plus an extra 500 bytes for the ECC variable names. Unlike other interfaces, the CTI message includes ECC variable names.

Only one ECC payload can have scope at any time during a call. You set which ECC payload has scope in the scripting environment. The solution uses the Default payload unless you override it.

For conferences and transfers, in general, use the same ECC payload through the call flow. The variable merging behavior is more complex when merging two calls with different ECC payloads. For more information, see the *Scripting and Media Routing Guide for Cisco Unified ICM/Contact Center Enterprise*.

ECC Payload Use by Interface

This table summarizes the use of ECC payloads in various operations:

Condition	ECC Payload That Is Used
Routing to VRU	Default payload If an ECC payload is specified in the configuration of that VRU, it overrides the Default payload.
Routing to Application Gateway	ECC payload that currently has scope in the script
Routing to Agent PG (including the Unified CM PG and Avaya PG)	ECC payload that currently has scope in the script
Routing to Media Routing PG	Default payload If an ECC payload is specified in the configuration of the VRU for that MR PG, it overrides the Default payload.
Routing to pre-12.0 PG	Always Default payload
Routing to System PG (Agent or VRU)	Always Default payload
Routing to Avaya Aura Symposium PG	Always Default payload
Routing to Aspect PG	Always Default payload
Contact Director to target Unified CCE	ECC payload that currently has scope in the script
Routing to INCRP NIC	ECC payload that currently has scope in the script
Pre-route to Gateway PG on Parent in Parent/Child	Always Default payload

Note If you do not create another ECC payload, the solution uses the Default payload for everything.

Use of UUI in Contact Center Enterprise Solutions

You can set UUI by Unified CCE scripts and extract it by Unified CVP for resending in SIP messages.

UUI processing scenarios:

- You can have GTD (generic type descriptor) data in the inbound call leg of the SIP INVITE message in the mime body format for GTD. In this case, Unified CVP saves the GTD data as inbound GTD and passes the UUI portion (if present) to Unified CCE.

  Cisco IOS gateways support this GTD format on outbound VoIP dial peers with SIP transport.

  If Unified CCE modifies the data, it sends the modified UUI back to Unified CVP. Unified CVP converts the UUI data from Unified CCE into hex, modifies the UUS (if present), and overwrites the inbound GTD value. Unified CVP only modifies the UUS portion, using the format:

  ```
  UUS,3,<converted Hex value of data from Unified CCE>
  ```

  Unified CVP preserves the rest of the GTD parameter values, saving the values as they arrived from the caller GTD.

- If the inbound call leg has no GTD, Unified CVP prints a message on the trace stating "No GTD Body present in Caller Body." The call then continues as a regular call.

Note

- Unified CCE passes the modified UUI in the *user.microapp.uui* ECC variable or the *Call.UsertoUserInfo* variable.
- If you use both variables, the *Call.UsertoUserInfo* variable takes precedence.

Modified GTD is set in the outbound INVITE mime body from CVP SIP B2BUA, which includes IP originated callers and TDM callers. If a DTMF label for outpulse transfer is received on a connected call, then the BYE message is sent with the GTD only if Unified CCE passes UUI. The BYE message comes immediately after the SIP INFO with DTMF.

Note You cannot use the UUI data transfer feature with Hookflash or Two B Channel Transfer (TBCT).

UUI in Unified CCE Scripts

To extract the UUI in your Unified CCE Script, look at the *user.microapp.uui* Call ECC variable and the *Call.UserToUserInfo* variable. By using the SET node on either one of these variables, you can set the variable on the outbound direction of the call.

Setting *Call.UserToUserInfo* variable takes precedence over using the ECC variable.

Note Unified CVP sends a BYE message on the DTMF label only if Unified CCE passes UUI.

If a BYE message is received, then the GTD from the received BYE is used to send it on the other leg.

Configure the Ingress Gateway with signaling forward unconditional, so that GTD with UUI and UUS are forwarded on the VoIP side. For example:

```
voice service voip
       signaling forward unconditional
```

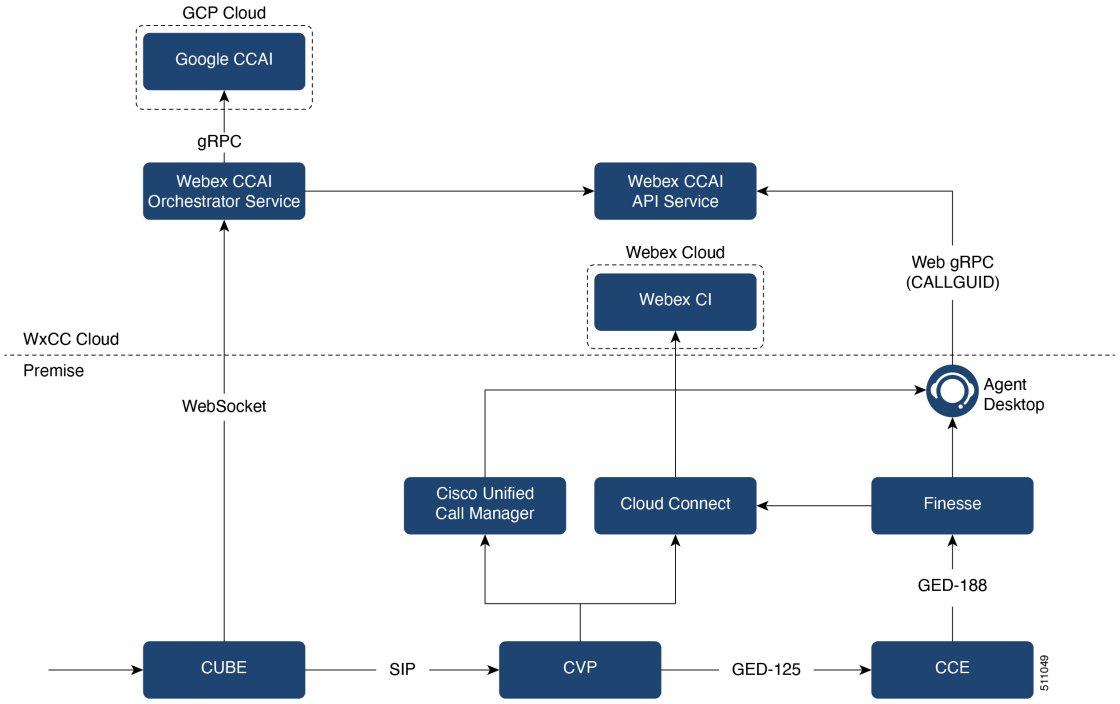
UUI in REFER and 302 Redirect Responses

If you use a REFER call flow, you can configure UUI in the Unified CCE script. The UUI is in a mime body and hex-encoded according to an ATT IP Toll Free NSS format. This placement of UUI also applies to 302 redirect responses.

```
VER,1.00
PRN,t1113,*,att**,1993
FAC,
UUS,0,(hex encoded UUI string here)
```

Contact Center AI Services Considerations

Figure 113: Contact Center AI Services Architecture



Unified CCEPackaged CCE solution leverages Artificial Intelligence (AI) and Natural Language Understanding (NLU) to provide Contact Center AI Services that assist agents via the Agent Answers and Call Transcript gadgets on the Cisco Finesse desktop. Unified CCE solution integrates with the following to provide the Contact Center AI Services:

- Google Contact Center AI (Google CCAI) is an extension of services that help create contact center solutions. These extensions, which are built upon Google's machine learning technology, assist human agents in conversations with end users by providing real-time documents and response suggestions. For more details, see https://cloud.google.com/solutions/contact-center.

- Webex Contact Center AI (Webex CCAI) services offer AI capabilities and mediation services between different AI service providers:
  - Webex CCAI Orchestrator service routes the voice streams of a conversation to the chosen AI Service provider (such as Google CCAI) in real time.
  - Webex CCAI API service receives the chosen AI service provider's response from the Webex CCAI Orchestrator service. Webex CCAI API service provides these responses to the Agent Answers and Call Transcript gadgets.
- Webex Contact Center (Webex CC) is a multi-tenant cloud solution which manages the tenant and its configuration.
- Webex Common Identity (Webex CI) authenticates the tenant before allowing access to Webex CC/Webex CCAI services.
- CUBE acts as a gateway in the call flow and forks the media streams of the agent and the caller towards the Webex CCAI Orchestrator service via the WebSocket protocol.

In CCE, you associate the Contact Center AI Services with incoming calls by setting two new ECC variables in a CCE routing script. This enables the Contact Center AI Services for all the calls that are routed through that script. If these configurations are present in the CCE script and gadget is enabled in Cisco Finesse, the gadgets are displayed to all the agents. The CCE routing script defines whether the Agent Answers has to be invoked for an agent. During an incoming call, when the agent speaks with the caller:

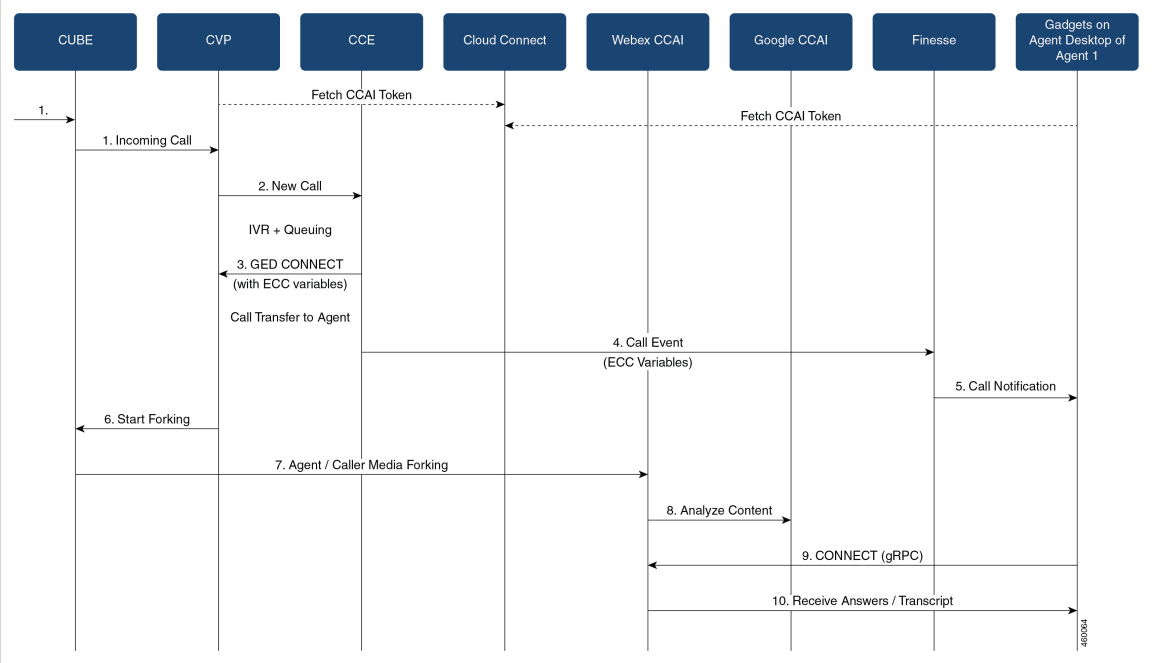
- The Agent Answers gadget displays relevant suggestions and recommendations in real time for the agent to consider. The suggestions and recommendations are based on the ongoing conversation between the caller and the agent.
- The Call Transcript gadget dynamically converts the ongoing conversation to text and presents the text to an agent for real-time viewing and reference.

Contact Center AI Services Call Flow

Figure 114: Contact Center AI ServicesCall Flow Diagram

Note Unified CCE solution components fetch and cache the CCAI authentication token from Cloud Connect to establish communication with the Webex CCAI Orchestrator service. The dotted lines in the preceding image indicate that the authentication tokens are fetched only once from Cloud Connect and not for every call.

Figure 115: Contact Center AI ServicesCall Flow Diagram



1. An incoming call arrives from CUBE at CVP.
2. CVP requests for call treatment instruction from CCE. CCE instructs CVP to subject the call to the standard CVA/IVR treatment, following which the call is queued for an agent to become available.
3. When an agent is available, CCE sends the GED Connect message to CVP. The GED Connect message includes two new ECC variables, one associated with the Contact Center AI Config ID and the other with the CALLGUID.

   Contact Center AI Config ID is the unique identifier assigned to your Google CCAI account during the account setup. A CALLGUID is the unique identifier that CCE assigns to a call when the call arrives at CVP.

   For more details on the Contact Center AI Config ID, see https://help.webex.com/en-us/npbt02j/Configure-Contact-Center-AI
4. CVP transfers the call to the agent through Unified CM. CCE sends the call information (which includes details on Contact Center AI Config ID and CALLGUID) to the Cisco Finesse server.
5. The call arrives at the Agent Desktop from Cisco Finesse. If the Agent Answers and Call Transcript gadgets are configured on the Agent Desktop, the gadgets are displayed to the agent in the Agent Desktop.
6. Using the Contact Center AI Config ID in the ECC variable (received at step 3), CVP requests CUBE to start forking the media streams of the agent and the caller. The forking request informs CUBE about which AI services are enabled for that call and where the forked media streams should be sent to.
7. When CUBE receives the forking request, it attempts to establish a WebSocket connection with the Webex CCAI WebSocket server. When the WebSocket connection is successful, CUBE starts forking the media streams of the agent and the caller towards Webex CCAI.

8. The Webex CCAI Orchestrator Service passes the media streams of the agent and the caller to Google CCAI (or other AI services). Google CCAI processes the media streams and generates the transcripts and suggestions as response.

   The Webex CCAI Orchestrator Service publishes the responses received from Google CCAI to the Webex CCAI API Service.

9. The Agent Answers and Call Transcript gadgets request for Contact Center AI Services by invoking the Webex CCAI API Service.

10. The Agent Answers and Call Transcript gadgets start displaying the response received from the Webex CCAI API Service to the agent in real-time.

For details on how to configure Contact Center AI Services, see the Agent Answers chapter and the Call Transcription chapter in the *Cisco Packaged Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html.

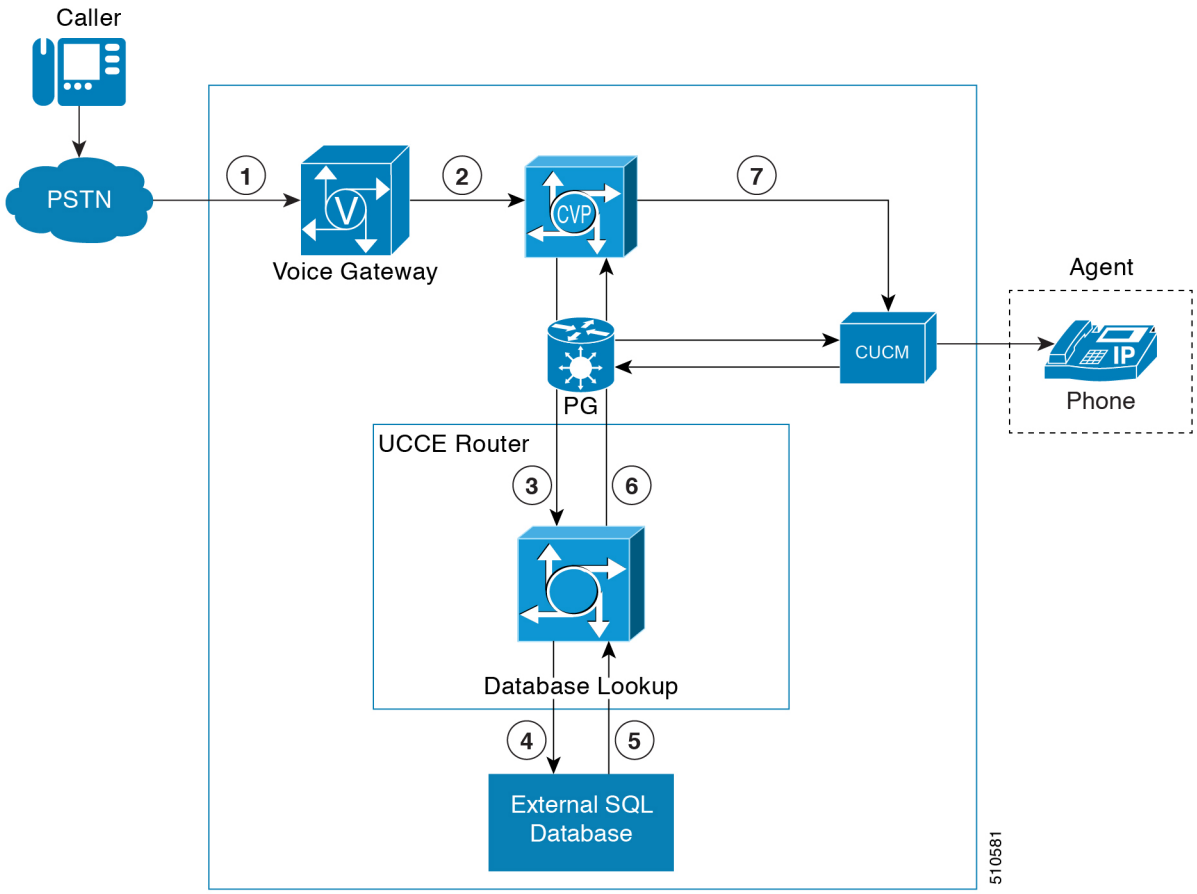
Database Lookup Design Considerations

- Application Gateway has more overhead in setting up an integration with the GED 145 interface but offers more flexibility in how data is accessed.
- CVP APIs offer REST API or database or custom integration options which also offload processing on the primary routing engine in the solution. The downside is that CVP processing across multiple nodes adds to complexity in coordination and maintenance. It also does not have all of the context available while running in routing script.
- Database Lookup has a lower cost of setup than application gateway and makes the information available within the script were reporting objects are accessible, but there are some limitations on how data is indexed, and what data will be available. Database Lookup also provides a way to access external data brought into the script without utilizing ECC variables as are commonly used in Application Gateway or CVP integrations to push data to the router.

Database Lookup Call Flows

The basic Database Lookup call flow runs as shown in this diagram.

Figure 116: Database Lookup Call Flows



Database Lookup Sizing Considerations

The supported Database Lookup rate aligns with the maximum call rate for the system.

Database Lookup Design Impacts

- The external database is required to be hosted on a virtual machine that is separate from the virtual machines the Packaged CCE solution is hosted on.
- The external database must be running a compatible version of the Microsoft SQL Server.
- The timeouts configured for Database Lookup should be consistent and align with other request timeouts. If utilized in a Contact Director route to a target instance, the Database Lookup should be 25% of the route request timeout.
- The Database Lookup node is based on a single primary key. Complex queries are not supported.
- The total size of the data from all the columns must not exceed 3500 bytes.

Mixed Codec Considerations

Contact center enterprise solutions support G.711 codec only for VRU. The SIP carrier or TDM-IP gateway sends the capability as G.711 and G.729, with a higher priority for G.729. The prompts at the Voice Browser should be G.711. The agents support both G.711 and G.729, with a higher priority for G.729. This configuration avoids the use of transcoders for VRU and connecting calls to agents. You can avoid the use of universal transcoders for Whisper Announcement by defining dual codecs for the ingress gateway and Unified CM.

VRU is negotiated as G.711. The solution automatically renegotiates the caller-agent conversation as G.729 to save bandwidth over WAN links.

You can use either G.711 mu-law or a-law prompts, but configure the entire solution for the same format.

G.711 a-law supports the following features:

- Agent Greeting
- Whisper Announcement
- Unified CM-Based Silent Monitoring
- Outbound SIP Dialer
- Courtesy Callback
- Post Call Survey
- Mobile Agents

Note SIP Dialers with CUBE can support a-law with specific design considerations. The SIP Dialer does not advertise a-law. The solution needs DSP resources (transcoder) on CUBE during the initial negotiation (no media) between the SIP Dialer and the SIP service provider. During a REFER from the Dialer to the agent, CUBE renegotiates the code with the agent to use a-law. CUBE then releases the DSP resource (transcoder).

Mixed Codec Use Case

Use mixed codecs to avoid transcoders and DSP resources. Define a dual codec at the ingress and egress gateways and Unified CM. The VRU automatically negotiates the call as G.711. The system then renegotiates the call as G.729 or G.711 for the caller-agent conversation.

Mixed Codec Call Flows

Logical Flow During VRU

1. A call arrives at the ingress voice gateway (G.729, G.711). The gateway sends a SIP invite message to the SIP Proxy Server, which forwards the request to the Unified CVP SIP Service.
2. CVP sends the call to the Voice Browser.
3. The call is established with G.711 codec without the use of transcoders.

Logical Flow During Caller and Agent Conversation

1. A call arrives at the ingress voice gateway (G.729, G.711). The gateway sends a SIP invite message to the SIP Proxy Server, which forwards the request to the Unified CVP SIP Service.
2. CVP sends the call to Unified CM to route to a Unified CCE agent.
3. The call is renegotiated and established as G.729 without the use of transcoders.

Mixed Codec Design Impacts

Design impacts for mixed codec are:

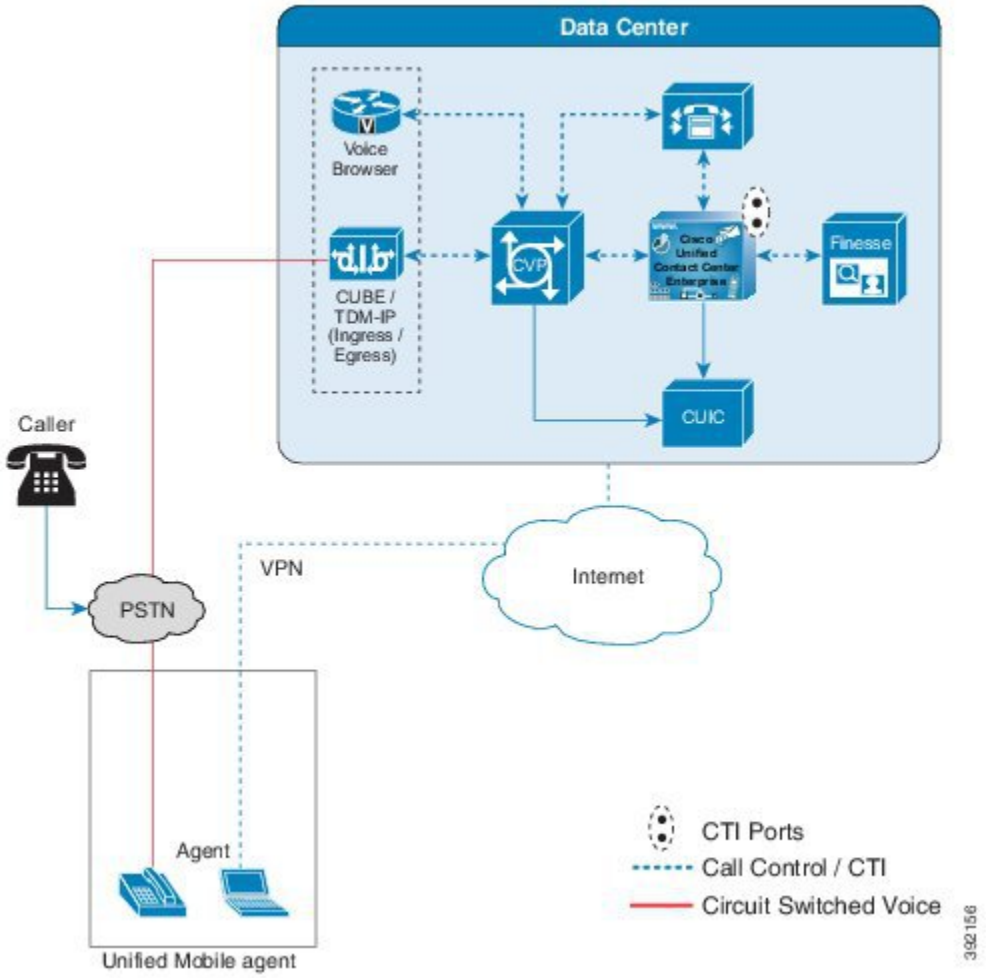
- A SIP trunk that only supports G.729 forces the use of transcoders.
- Certain features, such as Whisper Announcement, require universal transcoders for G.729 to G.729 interworking.
- If transcoders are required due to single G.729 codec, use Unified CM controlled transcoding resources. They trigger automatically for any codec mismatch.
- Sizing of appropriate transcoding and universal transcoding resources are required based on call flows, supplementary services, and certain integrated features.

Mobile Agent Considerations

Unified Mobile Agent enables an agent with any PSTN phone and a broadband VPN connection to function like a local agent in a formal contact center.

Unified Mobile Agent uses a pair of CTI ports that function as proxies for the mobile agent phone (or endpoint) and the caller phone (or endpoint). Every signed-in mobile agent requires two CTI ports, that's local and remote. The CTI ports take the place of the Cisco IP Phone that Unified CM JTAPI controls. The agent signs in to the local CTI port DN and Unified CM routes calls for the mobile agent to that DN. In a nailed connection, the remote CTI port calls the agent during sign-in. However, in a call-by-call connection, the Unified CM routes a call to the agent. By using media redirection, the CTI ports signal for the two VoIP endpoints to stream RTP packets directly. There's no further involvement from the CTI ports until further call control (transfer, conference, hold, retrieve, or release) is required. The agent performs any subsequent call control from the agent desktop. The PG transmits the necessary subsequent call control by JTAPI to Unified CM for the CTI ports to control the media of the call.

Figure 117: Cisco Unified Mobile Agent Architecture



The two CTI ports (local and remote) are logically and statically linked within the PG software. The PG registers the CTI ports at PG initialization. Call observers are added for these two CTI Ports when a mobile agent signs in. The PG provides call control for the CTI Ports and the call. The voice path is between the two voice gateways.

At the contact center, a mobile agent can sign in as a local agent from a JTAPI-controlled phone, using the same agent ID. Historical call reporting doesn't distinguish between calls handled as a mobile agent or a local agent.

Note

- Unified Mobile Agent can't use IPv6-enabled CTI ports.
- To record transferred calls in queue, from Unified Mobile Agent to Unified CVP, we recommend that you record the calls using the remote CTI port instead of the local CTI port.

Connection Modes

With Unified Mobile Agent, you can configure agents to use either call-by-call dialing, a nailed connection, or allow agents to make the choice during sign-in.

With call-by-call connections, consider these points:

- If the agent phone is configured for voicemail, then disable voicemail to let the RONA call processing to occur.
- An agent must answer the phone by going off the hook and end the call by disconnecting their phone. The Answer button on the agent desktop is disabled.
- An agent can't end one leg of a transfer without ending it at the other leg. The transfer must be completed or both legs must be dropped.
- Autoanswer isn’t possible. There’s no call control mechanism to make the mobile agent phone go off the hook.

With nailed connections, consider these points:

- A mobile agent can log off by using the desktop or by just disconnecting the phone.
- You can use autoanswer in a nailed connection.
- The following Unified CM timers can end a mobile agent call:
  - Maximum Call Duration timer (the default value is 720 minutes)
  - Maximum Call Hold timer (the default value is 360 minutes)

  To keep the mobile agent signed in, set the values for both of these timers to 0 so these timers never expire.
- Your firewall can block the media stream if:
  - An agent is idle for longer than the firewall idle timeout value.
  - The firewall idle timeout expires.

  To prevent this issue, increase the firewall idle timeout value.

Mobile Agent Call Flows

Call-By-Call Connection Call Flow

In call-by-call dialing, the agent’s remote phone is dialed for each incoming call. When the call ends, the agent’s phone is disconnected before the agent is made ready for the next call.

A basic call flow for this type of dialing is as follows:

1. At sign-in, a mobile agent specifies their agent ID, password, a local CTI port DN as the extension, and a phone number at which to call them. The administrator preselects the CTI port DN based on the agent's location.
2. The queueing process works the same for a mobile agent as for a local agent.

3. When a mobile agent is selected for the call, the new processing for a mobile agent begins. The Router uses the directory number for the agent's local CTI port as the routing label.
4. The incoming call rings at the agent's local CTI port. The Agent PG is notified that the local CTI port is ringing but does not answer the call immediately. The caller now hears ringing.
5. Simultaneously, a call to the agent is initiated from the remote CTI port for the selected agent. This process can take a while to complete, depending on the connection time. If the agent does not answer within the configured time, RONA processing starts.
6. When the agent answers their phone by going off-hook, this second call is temporarily placed on hold. Then, the original customer call is answered and directed to the agent call media address. The agent call is then taken off hold and directed to the customer call media address. The result is an RTP stream directly between the two VoIP endpoints.
7. When the call ends, both connections disconnect and the agent is set to ready, not ready, or wrap-up, as appropriate.

Nailed Connection Call Flow

In nailed connection mode, the agent is called once at sign-in, and the line stays connected through multiple customer calls.

A basic call flow for this type of connection is as follows:

1. At sign-in, a mobile agent specifies their agent ID, password, a local CTI port DN as the extension, and a phone number at which to call them. The administrator preselects the CTI port DN based on the agent's location. A remote CTI port is statically associated with the local CTI port.
2. The remote CTI port starts a call to the phone number that the mobile agent supplied. When the agent answers, the call is immediately placed on hold. The agent is not signed in and ready until this process completes.
3. The queueing process works the same for a mobile agent as for a local agent.
4. When a mobile agent is selected for the call, the new processing for a mobile agent begins.
5. The incoming call rings at the local CTI port for the mobile agent. The JTAPI gateway detects that the CTI port is ringing, but does not immediately answer the call. The caller now hears ringing.
6. The agent's desktop indicates that a call is ringing. The agent phone does not ring because it is already off hook. If the agent does not answer within the configured time, RONA processing begins.
7. When the agent presses the Answer button to accept the call, the customer call is answered and directed to the agent call media address. The agent call is then taken off hold and directed to the customer call media address.
8. When the call ends, the customer connection disconnects and the agent connection is placed back on hold. The agent is set to ready, not ready, or wrap-up, depending on agent configuration and agent desktop input.

Outbound Call Flow for Mobile Agent

Mobile agents can participate in outbound campaigns only on a nailed connection.

The call flow for predictive, progressive, or preview dialing is as follows:

1. The mobile agent signs in using the local CTI port DN as the agent phone number.
2. The standard Outbound Option call flow occurs.
3. When the Router selects the mobile agent, the MR PG returns the label (local CTI port DN) for an available agent to the dialer.
4. The dialer places a reservation phone call to the local CTI port DN and automatically places it on hold.
5. In progressive or predictive mode, when the dialer selects the mobile agent to handle a live call, the dialer transfers the call to the local CTI port.

   In preview mode, when the dialer reaches a live call on behalf of the mobile agent, the dialer transfers the call to the local CTI port.
6. The dialer auto-answers the transferred call for the agent through the CTI server. This quickly establishes the voice path between the customer and the agent. The dialer then disconnects the reservation call to the mobile agent.

Mobile Agent Design Impacts

Unified Mobile Agents can sign in to Unified CCE with any PSTN phone that gets routed to a Cisco Voice Gateway. Mobile agents also require an agent desktop.

You can use any Voice Gateway that Unified CCE supports for mobile agents. You can register the Voice Gateway with the same Unified CM cluster as the Agent PG or with another cluster. The caller (ingress) and mobile agent (egress) Voice Gateways can use either MGCP or SIP.

Note If you enable Silent Monitoring, use different Voice Gateways for ingress and egress.

Unified Mobile Agents can use a Cisco IP Phone that is configured for SIP. Calls to mobile agents can also originate from SIP IP Phones.

For improved Unified CM performance, use Extension Mobility, instead of Unified Mobile Agent, for mobile agents with IP Phones on the same cluster as the Agent PG. Because the IP Phone device is associated with the JTAPI user, there is a small performance hit on Unified CM for making that association.

Consider the following factors when designing a Unified Mobile Agent solution:

- If you use SIP trunks, configure Media Termination Points (MTPs). This requirement also applies if you use TDM trunks to interface with service providers. For detailed information, see *Cisco Unified Contact Center Enterprise Features Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html.
- Enabling the use of an MTP on a trunk affects all calls that traverse that trunk, even non-contact center calls. Ensure that the number of available MTPs can support the number of calls traversing the trunk.

Agent Location and Call Admission Control Design

Because a CTI port is a virtual type of endpoint, it can be located anywhere. But, always configure the CTI ports for a mobile agent with the same location as the agent's VoIP endpoint. The CTI port pair for a mobile agent must also be colocated with the Voice Gateway (or VoIP endpoint) that calls the agent. If you do not have both these conditions, call admission is not accounted for correctly.

Call admission control sees the mobile agent call as two separate calls. The first call leg is from the caller to the agent's local CTI port. The second call leg is from the remote CTI port to the agent. Because the CTI ports are colocated with the agent endpoint, call admission control counts only the call from the caller location to the agent location. This is why it is important for an agent to use CTI ports for their current location.

From the perspective of call admission control locations for the mobile agent CTI ports, there are three deployment scenarios:

- Use CTI ports colocated with the egress Voice Gateway that calls the mobile agent.
- Use CTI ports colocated with the ingress Voice Gateway.
- Use CTI ports colocated with the intercluster trunk between Unified CM clusters.

All pools of CTI ports are colocated with the VoIP endpoint type for the agent (Voice Gateway or IP phone).

Callers and agents can also use VoIP endpoints on another Unified CM cluster. This configuration enables agents in remote locations to be called from local Voice Gateways for a different Unified CM cluster.

Note If you use Silent Monitoring in this case, your solution requires a monitoring server at the remote site with the agent (egress) Voice Gateway.

For additional information about call admission control design, see the call admission control information in the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.

Dial Plans for Mobile Agent

Configure the Unified CM Dial Plan so that it routes calls from the remote CTI port to a Voice Gateway colocated with the mobile agent CTI ports. Otherwise, call admission control accounting does not work correctly.

You can also configure the Unified CM Dial Plan so that all calls from the CTI ports go through a specific gateway regardless of the called number. In that configuration, you have a dedicated gateway that all mobile agents use. It is more easily managed, but it might not be an efficient configuration from the perspective of PSTN trunk usage.

For additional information about dial plan design, see the *Cisco Collaboration System Solution Reference Network Designs*.

Codec Design for Mobile Agent

Media streams between the ingress and egress Voice Gateways can be G.711 or G.729. You cannot mix codecs, because all CTI ports for a PG must advertise the same codec type. This requirement can result in G.711 (instead of G.729) calls being sent across the WAN. If you route most calls to agents colocated with the ingress Voice Gateway, sending a few G.711 calls over the WAN might not be an issue. The alternative is to use G.729 for all mobile agent calls. If most Unified CCE calls cross a WAN segment, it probably makes sense to have all CTI ports configured for G.729. However, you cannot have G.711 for some mobile agent calls and G.729 for others. Your solution requires a dedicated region for the CTI ports to ensure that all calls to and from this region use the same encoding format.

For additional information about codec design considerations, see the *Cisco Collaboration System Solution Reference Network Designs*.

Music on Hold with Mobile Agent

You can use Music on Hold (MoH) for mobile agents just as you do for traditional agents. To let callers hear music, assign MoH resources to the Ingress Voice Gateway. Specify the user or network audio source on the local CTI port configuration. To let the agent hear music when on hold, assign MoH resources to the Egress Voice Gateway. Specify the user or network audio source on the remote CTI port configuration.

Note Always assign the MoH resources to the gateways. Do not assign MoH resources to local and remote CTI ports. It is unnecessary and can have a performance impact on the system.

A Mobile Agent remote call over a nailed connection is put on hold when there is no active call to the agent. In general, enable MoH to the mobile agent phone for nailed connection calls. If MoH resources are an issue, consider multicast MoH services.

For a nailed connection, disabling MoH for the remote phone might lead to the hold tone playing instead. This depends on the call processing agent that controls the remote phone. For Unified CM, the hold tone is enabled by default and is similar to the Mobile Agent connect tone. With the Unified CM hold tone enabled, it is difficult for the agent to identify if a call has arrived by listening for the Mobile Agent connect tone. Therefore, disable the hold tone for Unified CM by changing the setting of the Tone on Hold Timer service parameter on Unified CM.

For additional information about MoH design, see the *Cisco Collaboration System Solution Reference Network Designs*.

Cisco Finesse with Mobile Agent

Cisco Finesse supports mobile agents. The Cisco Finesse server needs no configuration to enable the Mobile Agent feature. To use the Mobile Agent feature, follow all configurations as outlined in *Cisco Unified Contact Center Enterprise Features Guide*.

Note Cisco Finesse IP Phone Agent does not support mobile agents.

On the Cisco Finesse sign-in page, if you select the mobile agent check box, the mobile agent options are presented to the agent. The mobile agent provides the local CTI port extension, a mode (Call by Call or Nailed Connection), and a dial number for the agent’s phone.

The agent's phone number must route to a VoIP endpoint (Voice Gateway, IP phone, or intercluster trunk) colocated with the CTI port pair for call admission control to work properly.

A Cisco Finesse mobile supervisor can perform all of the functions that a nonmobile supervisor can perform, except for Silent Monitoring. Cisco Finesse does not support Silent Monitoring of mobile agents.

DTMF Considerations with Mobile Agent

If mobile agents consult a VRU or other network component that uses DTMF to navigate, your solution might require MTP resources. The Mobile Agent feature relies on CTI ports, which do not support in-band DTMF (RFC 2833). If the agent's endpoints support only in-band DTMF (or if they are configured for in-band DTMF per RFC 2833), then Unified CM automatically inserts MTP resources to handle the mismatch. Ensure that sufficient MTP resources are available in this case.

Session Border Controllers with Mobile Agent

Some SIP devices, such as the Cisco Unified Border Element or other Session Border Controllers, can dynamically change the media port during a call. If this happens with a Mobile Agent call, your solution requires MTP resources on the SIP trunk that connects to the agent endpoint.

Fault Tolerance for Mobile Agent

The RTP stream for a mobile agent call is between the Ingress and Egress Voice Gateways. Because of this, a failure of Unified CM or Unified CCE does not affect call survivability. However, after a failover, subsequent call control (transfer, conference, or hold) might not be possible. The mobile agent's desktop notifies them of the failover and the agent must sign in again.

Sizing Considerations for Mobile Agent

Unified Mobile Agent uses conference bridge resources for Agent Greeting. With Agent Greeting, size each call as if it had a conference, rather than the greeting.

Unified Mobile Agent requires the use of two CTI ports per contact center call. One CTI port controls the caller endpoint, and the other CTI port controls the selected agent endpoint. The actual RTP stream is between the two endpoints and its bridged through these two CTI ports. However, there is extra call processing on Unified CM to set up calls to mobile agents through these two CTI ports.

Mobile agents can essentially sign in from any location (with the agent desktop) where they have a high-quality broadband connection and a PSTN phone. However, they are still associated logically with a particular Agent PG and Unified CM cluster, even if the Voice Gateway is registered with a different cluster. The agent is associated with a particular peripheral and cannot migrate freely to other peripherals without some custom modifications.

For specific subscriber and cluster sizing, use the Cisco Unified Communications Manager Capacity Tool. When sizing the cluster, input the maximum number of simultaneously signed-in mobile agents. To handle the configured mobile agents above your maximum simultaneous signed-in mobile agents, enter Type 1 CTI ports with a BHCA and BHT of 0 in the tool. This is similar to the method for accounting for local agent phones that are not signed in by using the CTI third-party controlled lines in the tool. As an alternative, you can input all mobile agents (signed-in and not signed-in) into the tool and adjust the BHCA and BHT per mobile agent accordingly. The total BHCA and BHT must remain the same as when considering simultaneous signed-in mobile agents with their actual BHCA and BHT.

Phone Extension Support Considerations

Consider these aspects for using phone extensions in your solution.

Your contact center can handle the following types of calls, each with its own considerations:

- Routed ACD Calls – Calls routed to an agent through a central queue that is based on skill or attribute.
- Routed Agent Calls – Calls routed to a particular agent where the customer has a specific business relationship with this agent.
- Non-routed calls – Direct dialed calls, possibly through a Direct Inward Dial extension or from unmonitored phones within the business.
- Agent to Agent calls – Calls placed between agents, either routed or not routed.

A phone extension can be either:

- ACD Extension – The extension the agent logs into, to which calls are routed.
- Secondary extension – Sometimes called a non-ACD extension. This is generally an extension where calls are not routed. The agent can use it for business or personal activity. The solution might monitor activity on this extension, depending on configuration, which impacts available features.

Monitored Secondary Extensions

Multi-Line Agent Control (MLAC) is a Cisco CCE agent peripheral setting which enables reporting and call control for calls on the secondary extensions. Use of MLAC imposes some restrictions, such as:

- No call waiting
- A limit to four extensions for each agent phone
- No shared lines between agents
- Applies to all phones on a given peripheral when enabled

Unmonitored Secondary Extensions

The solution does not track call activity on an unmonitored secondary extension. A secondary extension can have PBX functions that are not generally associated with contact center actions, like the use of shared lines.

Call Type Considerations for Phone Extensions

Non-Routed Direct Agent Dialing

If an ACD line receives calls that are not routed, or that are routed regardless of state, it can impact agent experience and reporting. A direct call from another agent can arrive while an agent is on a call or in a wrap-up state.

If your business model allows, the call flows are cleaner if all calls are routed to agents, even agent-to-agent calls. Write the routing script to take the agent state into account. Otherwise, there are possible race conditions that can impact reporting and agent experience.

Direct Agent Dialing

When agents have relationships with their customers and provide a direct number, there are a few options on how to deploy.

If you use the agent's primary extension for these direct dialed calls, then consider these options:

- Non-Routed — The calls go directly to the agent's ACD extension bypassing routing. Always enable call waiting on the phone. A VRU of routing script cannot add call context. But, if the calling number is not blocked, you can customize Finesse to perform an external database dip if the agent is signed in.
- Routed — If you route the calls to the agent's ACD extension, you get richer reporting, can include more call context, and can make routing decisions based on agent state:
  - You can queue the call until the agent is available using a Queue to Agent node.

- You can send the call to a signed-in agent regardless of state using an Agent to Agent node.
- You can send the call to an agent's extension or to voicemail using a Label node.
- You can use the requery option on those nodes if the agent doesn't answer or invokes a busy condition.

Shared ACD Line Support

Contact center enterprise solutions include shared ACD lines support for up to two devices. This support enables an agent with devices at different locations to utilize the same extension. When logging into the shared extension, the agent selects which device to make active via the device itself, by going off-hook or making or receiving a call. The active device is the device utilized when the agent answers a call or makes a call from the Finesse desktop. Please note, however, that all devices will ring when a call is sent to that extension.

Note UCM auto-answer is not supported when shared ACD lines are in use. Agent Desk Settings auto-answer is also not supported.

E.164 Dial Plan Design

Unified CCE supports E.164 dial plans and provides partial support for the '+' prefix as follows:

- Agent extensions cannot include the '+' character.
- Agent secondary lines cannot include the '+' character if the agent peripheral has "All Lines" Agent Control enabled.
- The VRU cannot include the '+' prefix unless you route DNs through a CTI Route Point.
- Dialer-imported contact numbers and campaign prefixes cannot include the '+' prefix.
- Agents can dial the '+' prefix with an E.164 number through their Cisco Finesse desktop.
- Agents can dial the '+' prefix with an E.164 number through their phones.

For contact centers that advertise the agent extension outside of the contact center, these considerations apply:

- Use transformation patterns to add the '+' prefix to the calling number on outgoing calls. You can use Calling Party Transformation CSS for phone configuration.
- To route incoming calls addressed to an E.164 number with the '+' prefix, use called party transformations on the translation patterns to strip the '+' prefix from the called number.
- The Attendant Console does not have visibility into the phone status.

Post Call Survey Considerations

A contact center typically uses a post call survey to gauge customer satisfaction with their experience. For example, did the customer receive the necessary information using the self-service or did they have a pleasant experience with the agent.

The Post Call Survey (PCS) feature enables a call flow that transfers the caller to a DNIS that prompts the caller with a post call survey.

The VRU treatment asks the caller if they want to participate in a post call survey. There are two responses a caller can have to a post call survey request:

1. If the caller chooses to do so, the call flow automatically transfers the caller to the survey call after the agent ends the conversation.
2. If the caller declines, your Unified CCE script uses an ECC variable to turn off the Post Call Survey on a per-call basis. By setting the ECC variable to *n*, the call does not transfer to the PCS DNIS.

For reporting purposes, the post call survey call has the same Call-ID and call context as the original inbound call.

Post Call Survey Use Case

The caller is typically asked if they want to participate in a survey after the call. Your solution can determine based on dialed numbers to invoke the post call survey at the end of a call. When the customer completes the conversation with an agent, the customer is automatically redirected to a survey. When the agent ends the call, it initiates the Post Call Survey.

A customer can use the keypad on a touch tone phone and voice with ASR/TTS to respond to questions asked during the survey. For the solution, the post call survey call is just like another regular call. The Post Call Survey retrieves the call context information from the original customer call.

Post Call Survey Design Impacts

Observe the following conditions when designing a Post Call Survey:

- A Post Call Survey initiates when the last agent ends the call. The call routing script launches a survey script.
- The mapping of a dialed number pattern to a Post Call Survey number enables the Post Call Survey feature for the call.
- The value of the expanded call variable **user.microapp.isPostCallSurvey** controls whether the call transfers to the Post Call Survey number.
    - If **user.microapp.isPostCallSurvey** is set to **y** (the implied default), the call transfers to the mapped post call survey number.
    - If **user.microapp.isPostCallSurvey** is set to **n**, the call ends.
    - To route all calls in the dialed number pattern to the survey, your script does not have to set the **user.microapp.isPostCallSurvey** variable. The variable is set to **y** by default.
- You cannot have a REFER call flow with Post Call Survey. REFER call flows remove Unified CVP from the call. But, Post Call Survey needs Unified CVP because the agent has already disconnected.
- For Unified CCE reporting purposes, the Post Call Survey call inherits the call context for the initial call. When a survey starts, the call context of the customer call that was transferred to the agent replicates into the call context of the Post Call Survey call.

- The expanded call variable **isPostCallSurvey** will be cached only when the UCCE router generates a label for CVP.

Webex Experience Management Considerations

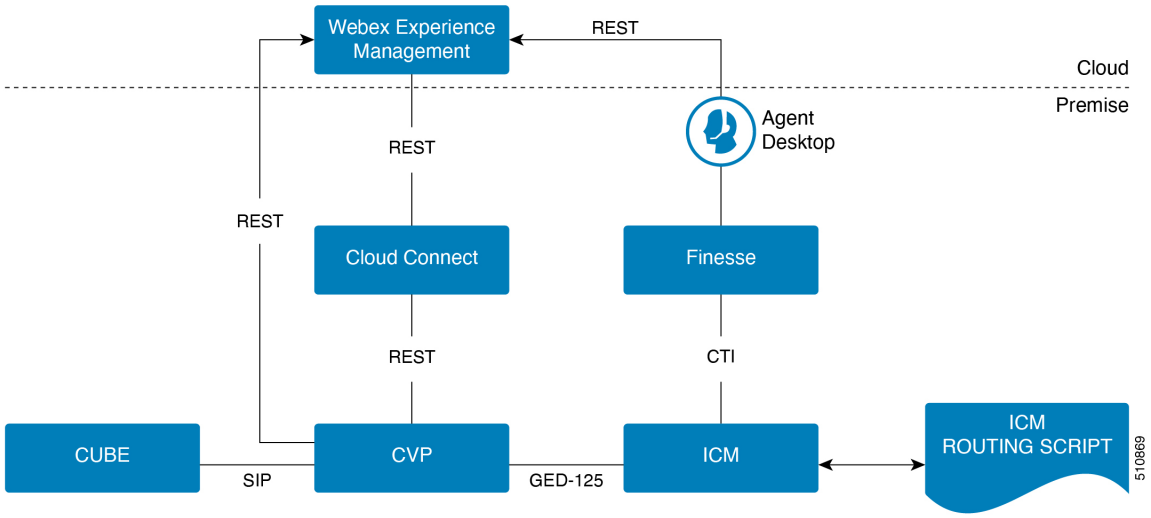
Webex Experience Management surveys use the same scripting and call flows as Post Call Surveys, with the exception that the questionnaire is provided by the cloud-based Experience Management service. The Call Studio survey application to be invoked is configured in the router script that runs during the survey leg of the call, and is passed to the CVP using an ECC variable. The Call Studio survey application fetches the questions from the Experience Management service, collects the answers from the caller, and submits them to the Experience Management service over REST APIs.

Experience Management can be configured to invoke surveys through the following channels:

- Voice
- Email
- SMS

Note The voice surveys can be triggered using the existing Post Call Survey and through Experience Management.

The following diagram shows the protocol interfaces between the solution components to deliver the Experience Management survey feature:



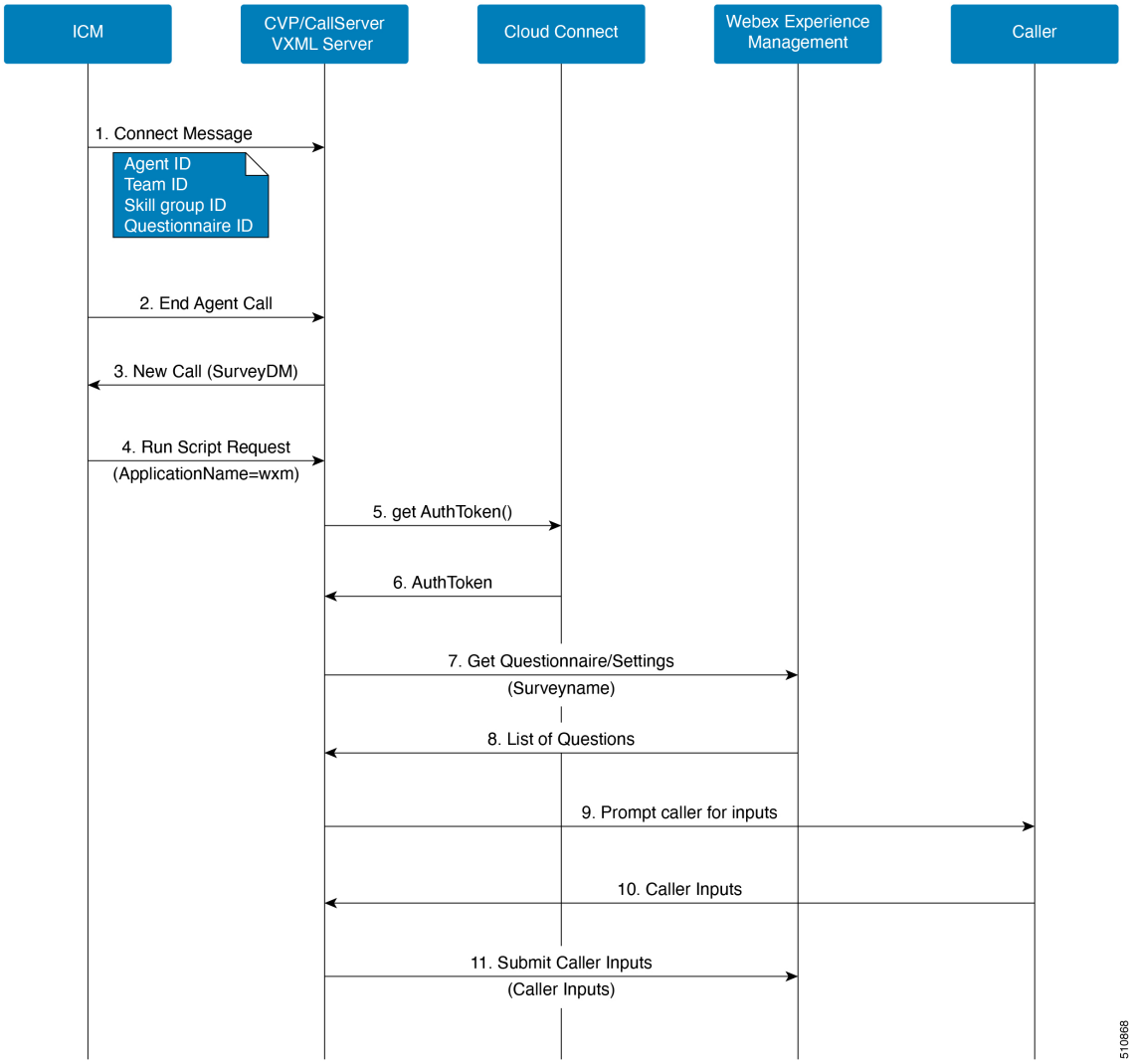
In addition to collecting the survey data, Experience Management provides rich context-sensitive analytics of the survey data. The analytics dashboard can be viewed on the agent and supervisor desktops. In order to associate each instance of the survey with the appropriate customer and call context, CVP sends the call context and any available customer context like ANI, in the survey leg of the call to Experience Management via APIs that are proxied via Cloud Connect. CVP receives the call and customer context information from CCE in the GED-125 connect message during the original leg of the call. Cloud Connect provides secure

storage for the tenant credentials required to invoke Experience Management APIs, and provides capabilities to monitor the status and latency of the Experience Management service.

Experience Management Call Flows

Experience Management Voice Survey

Figure 118: Voice Survey Call Flow



1. During an inbound call, when the ICM routing script allocates an agent, ICM sends the associated call context (Agent ID, Skill Group ID, Team ID, and QuestionnaireId) back to CVP in the connect message. The connect message is used as call context in the associated survey call (at Step 4).
2. The end of the agent call triggers a survey call to the configured Survey DN in the CVP.
3. The Survey DN is associated with a Call Type in the ICM that runs an associated routing script.

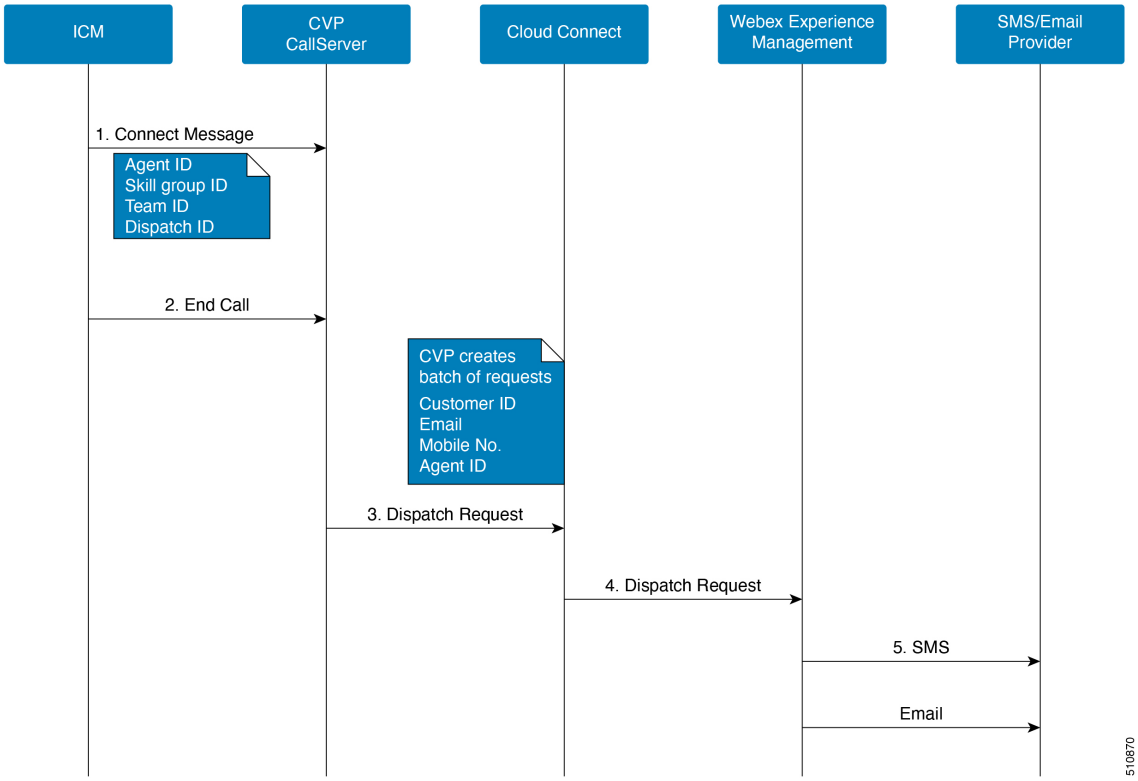
4. The associated routing script returns a run script request containing the VXML application (wxm) to be run along with the call context (Agent ID, Skill Group ID, Team ID, and DispatchId).
5. The VXML server invokes a getAuthToken() API call to Cloud Connect.
6. Cloud Connect uses the organization credentials of Experience Management (administrator credentials and API key) to invoke the getAuthToken() API and sends back the AuthToken to the CVP VXML server.
7. The VXML server then invokes the getQuestionnaire() and getSettings() API call with the AuthToken from Cloud Connect (received in step 6) and Survey Name from ICM (received in step 4) to Experience Management.
8. Experience Management returns the list of questions to the CVP VXML server.
9. The VXML server continues to prompt the caller with the questions.
10. The caller submits the input for the questions to the VXML server.
11. The VXML server submits the user answers and agent details received in step 4 to Experience Management.

Note For Experience Management voice survey to function:

- VXML server must be connected to the internet. Enable direct access to the internet or configure HTTP proxy settings in the VXML server. For more information, refer to the *Configure HTTP proxy settings in VXML server* section in the *Configuration Guide for Cisco Unified Customer Voice Portal* at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-installation-and-configuration-guides-list.html.
- Either configure TTS in the VVB, or upload appropriate audio prompts to the Experience Management PCS survey configuration.

Experience Management Email/SMS Survey

Figure 119: Email/SMS Survey Call Flow



1. During an inbound call, when the ICM routing script allocates an agent, ICM sends the associated call context (Agent ID, Skill Group ID, Team ID, and DispatchId) back to the CVP in the connect message. The connect message is used as call context in the associated survey call (at Step 3).

Note DispatchId tells the CVP call server that Email/SMS needs to be sent back to the caller after the call ends.

2. The end of agent call triggers the deferred Post Call Survey (sending Email/SMS) logic in the call server.
3. The CVP call server creates a batch of requests and sends it to Cloud Connect which contains DispatchId, CustomerId, Email, and Mobile (received in Step 1 from ICM) and calls the DispatchRequest() API on Cloud Connect.
4. Cloud Connect calls the DispacthRequest() API call on Experience Management.

For more information on how to configure Experience Management, see the *Webex Experience Management* chapter in the *Cisco Unified Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html

Network Considerations

Experience Management Voice Survey: Cloud Connect, CVP VXML, and Finesse Client must have access to the Experience Management cloud services.

Experience Management Email/SMS Survey: Cloud Connect and Finesse Client must have access to the Experience Management cloud services.

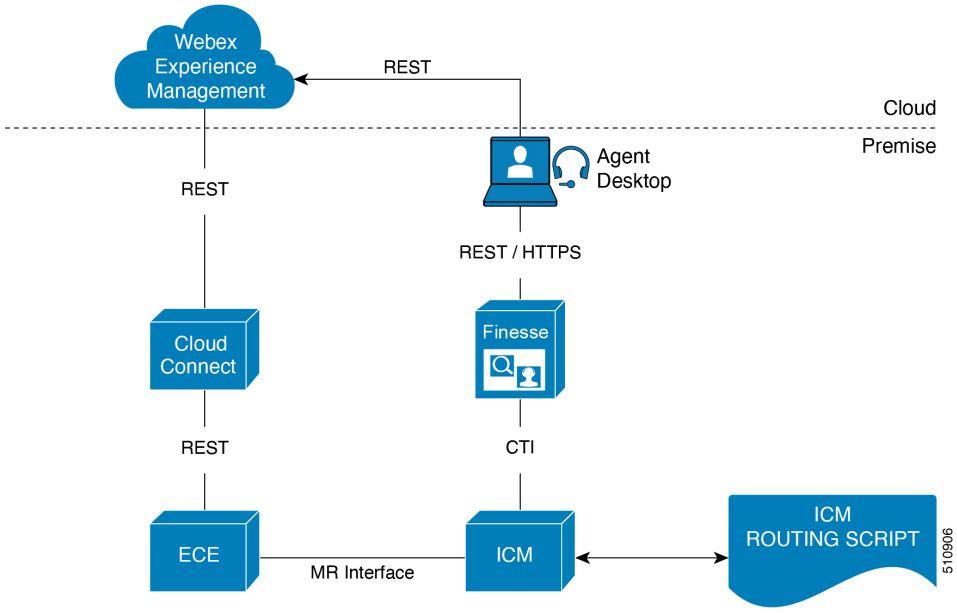
All Finesse clusters must be included in the allowed list of URLs on the Experience Management portal in the form of the following URL: *https://<finesse FQDN>:8445*.

Webex Experience Management Digital Channel Survey Considerations

Enterprise Chat and Email (ECE) receives the digital survey configuration and customer prefill data via ECC variable from ICM. It retrieves the digital channel survey questionnaire from the Experience Management using REST API calls proxied via Cloud Connect. It embeds the survey URL in the inline email response or in the chat window at the end of the chat.

The following diagram shows the protocol interfaces between the solution components to deliver the Digital Channel Survey feature:

Figure 120: Digital Channel Survey Architecture



In order to associate each instance of the survey implementation with the appropriate customer and agent (in case of transfer,the survey is associated with the last agent involved in the customer interaction), ECE sends available customer context like customer ID, email, and phone number to Experience Management when the survey is initiated.

ECE receives the customer context information from CCE in ECC variables and establishes a connection with Cloud Connect using the inventory information available with UCCE. Cloud Connect provides secure storage

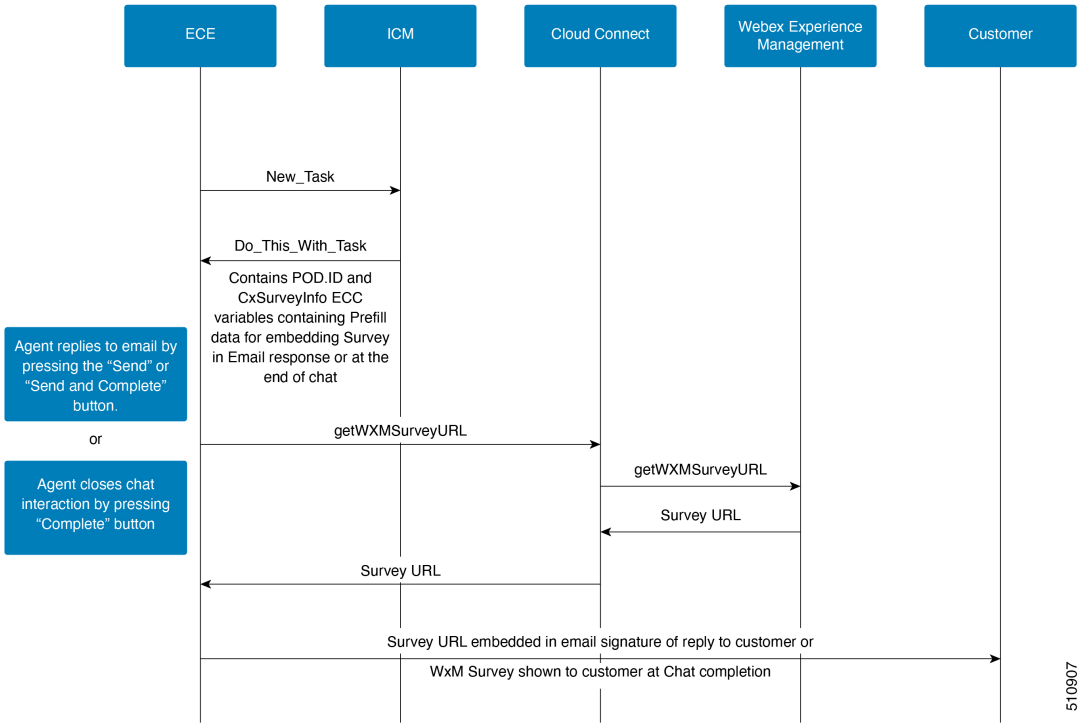
for the tenant credentials required to invoke Experience Management APIs, and provides capabilities to monitor the status and latency of the Experience Management service.

Digital Channel Survey Call Flows

This section provides the Digital Channel Survey call flow information.

Digital Channel Survey (Email/Chat)

Figure 121: Email/Chat Survey Call Flow



1. ECE initiates a new task request to ICM.
2. ICM routing script allocates the task to an agent. ICM sends the associated email/chat context (such as, Agent ID, Skill Group ID, Cisco Webex Teams ID, Questionnaire name, and Customer ID) and assigns the task to the agent.
3. When the ECE initiates the survey, it calls the `getWXMSurveyURL()` API to retrieve the survey URL from the Cloud Connect along with the questionnaire name.

   ECE sends the customer and agent context received from ICM in the **POD.ID** and **user.CxSurveyInfo**variables while calling the `getWXMSurveyURL()`. If ICM doesn't send the mobile and email address in the **POD.ID**, then the ECE refers its database for these and add them, if available, while calling the `getWXMSurveyURL()`.
4. Cloud Connect calls the Experience Management `getWXMSurveyURL ()` API to get the survey questionnaire.
5. ECE receives the Survey URL for the questionnaire.

6. ECE initiates the survey with the survey URL:
    - For Email - The survey URL is embedded in the email signature.
    - For Chat - The survey is shown to the customer at the end of chat.

For more information on how to configure Enterprise Chat and Email, see https://www.cisco.com/c/en/us/support/contact-center/enterprise-chat-email-12-5-1/model.html.

For more information on how to configure Experience Management, see the *Webex Experience Management Digital Channel Survey* chapter in the *Cisco Unified Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-feature-guides-list.html

For more information on how to add Experience Management gadgets into Finesse desktop layout for agents and supervisors, see Cisco Webex Experience Management Gadgets.

Network Consideration

- Cloud Connect and Finesse desktop browser must have access to the Experience Management cloud services.
- All Finesse VMs must be allowed on the Experience Management portal in the form of the following URL: *https://<finesse FQDN>:8445*.

Customer Journey Analyzer

Customer Journey Analyzer is a cloud service that processes historical contact center data from on-premise deployment to generate specific Business Metrics. It displays trends to help you identify patterns and gain insights for continuous improvement. You can view the Abandoned Contacts dashboard on the Customer Journey Analyzer, which enables supervisors and business analysts to identify where contacts are being abandoned and take appropriate action. You can use Customer Journey Analyzer to create visualizations using Customer Activity Records and Customer Session Records.

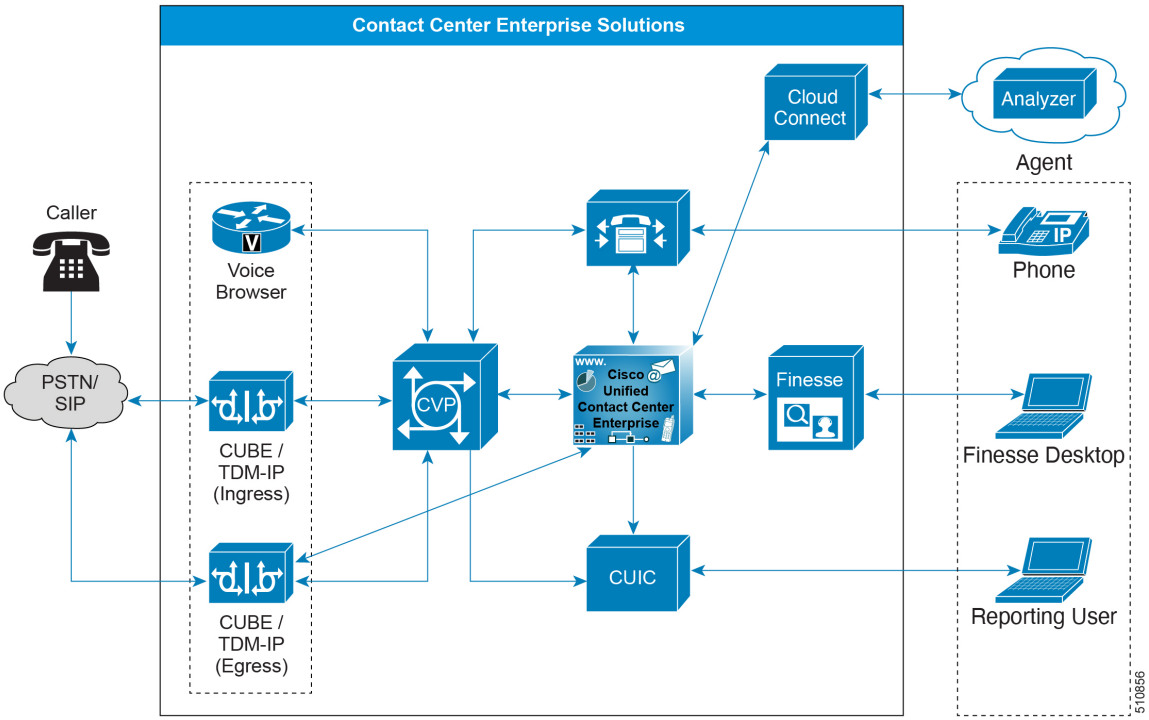
The following reports are part of the Abandoned Contacts dashboard:

For more information on the reports, refer to the **Online help** in the Dashboard.

- Total Abandoned Contacts
- Leading Abandonment Reason
- Call Back / Renewed Chat Rate
- Customer Journey
- Contacts Trend
- Abandoned Contacts by Stage
- Abandoned Contact Details

Customer Journey Analyzer Data Call Flow

Figure 122: Customer Journey Analyzer Call Flow



Cloud Connect allows Cisco Unified CCE customers to use cloud capabilities like **Business Metrics in Customer Journey Analyzer**. The **Customer Journey Analyzer** mines historical data from multiple data sources and systems to generate specific business views of data. It displays trends to help you identify patterns and gain insight for continuous improvement. To generate these business metrics, agent activities and contact (telephone, chat and email) activities records are published to **Customer Journey Analyzer**. Cloud Connect component is responsible for publishing the data to **Customer Journey Analyzer**.

Data Security for Customer Journey Analyzer

The Customer Journey Analyzer for this feature the following data is secured.

1. **System Configuration Data**:
   a. All configuration information is transferred from **Unified CCE Administration** through **REST APIs** with secured *HTTPS* connections.
   b. The user credentials used by the Cloud Connect are encrypted before caching them into the machine.
2. **Customer Contacts Data**: The Call and Agent activity records are published to Analyzer over secured *HTTPS* connection.

Precision Routing Considerations

Precision routing offers a multidimensional alternative to skill group routing. Precision queues are the key components of precision routing. Using Unified CCE scripting, you can dynamically map the precision queues to direct a call to the agent who best matches the caller's precise needs. Precision queues consist of one or more steps with configured expressions allowing a customer to find the precise needs of the caller.

There is no need to add an agent to a precision queue; agents become members of precision queues automatically based on their attributes. Consider a precision queue that requires an agent in Boston, who speaks fluent Spanish, and who is proficient in troubleshooting a specific piece of equipment. An agent with the attributes `Boston = True`, `Spanish = True`, and `Repair = 10` is automatically part of that precision queue. A Spanish-speaking caller in Boston who needs help with that equipment is routed to that agent.

Precision routing enhances and can replace traditional routing. Traditional routing looks at all of the skill groups to which an agent belongs and defines the hierarchy of skills to map business needs. However, traditional routing is restricted by its single-dimensional nature.

Precision routing provides multidimensional routing with simple configuration, scripting, and reporting. Agents have multiple attributes with proficiencies so that the capabilities of each agent are accurately exposed. This brings more value to the business.

If your routing needs are not too complex, consider using one or two skill groups. However, if you want to conduct a search involving as many as ten different proficiency levels in one easily managed queue, use precision queues.

Precision Routing Use Case

Unlike skill groups, a precision queue breaks down attribute definitions to form a collection of agents at an attribute level. The agents who match the attribute level of the precision queue become associated with that precision queue.

With precision queues, an English Sales queue involves defining the attributes English and Sales, and associating agents with those traits to those attributes. The precision queue English Sales dynamically maps all agents who have those traits to the precision queue. You can also define more complex proficiency attributes to associate with those agents. This enables you to build, in a single precision queue, multiple proficiency searches like `English Language Proficiency = 10` and `Sales Proficiency = 5`.

To match the English Sales queue with skill groups, you set up two separate skill groups, one for each of the attributes. With precision queues, you can refine agents by attributes. With skill groups, you define a skill group and then assign agents to it.

Precision Routing Call Flows

At a high level, consider a 5-step precision queue which first checks if the caller is Premium Member:

1. Attribute: Skill > 8 - Consider If: Caller is Premium Member
2. Attribute: Skill > 6
3. Attribute: Skill > 4
4. Attribute: Skill > 3

5. Attribute: Skill >= 1

John, who is not a premium customer, calls 1-800-repairs. The system sends John's call to this precision queue. The precision queue works like this:

1. Since John is not a premium customer, he is immediately routed out of Step 1 (because of the Consider If on Step 1).
2. The call moves into Step 2 where he waits for an agent with a Skill greater than 6 to answer his call.
3. After the Step 2 wait time expires, John's call moves to Step 3 to wait for an agent with a Skill greater than 4.
4. After the Step 3 wait time has expired, John's call moves to Step 4 to wait for an agent with a Skill greater than 3.
5. When it arrives at Step 5, John's call waits indefinitely for an available agent. This step applies to any call because there is no routing logic past this step.

The call goes through each successive step to expand the pool of available agents. Eventually, when you reach the last step, the call waits for the largest pool of potential agents. With each extra step, the chances increase that there is an available agent to handle the call. This also puts the most valuable and skilled agents in the earlier precision queue steps. Calls come to them first before moving on to the less appropriate agents in later steps.

Precision Routing Design Impacts

Precision Routing Attributes

Attributes identify a call routing requirement, such as language, location, or agent expertise. Each precision queue can have up to ten unique attributes, and you can use these attributes in multiple terms. You can create two types of attributes: Boolean or proficiency. Use Boolean attributes to identify an agent attribute value as true or false. Use proficiency attributes to establish a level of expertise in a range from 1 to 10, from lowest to highest.

When you create a precision queue, you identify which attributes are parts of that queue and then implement the queue in scripts. When you assign new attributes to an agent, the attribute values automatically associate the agent with any precision queue with matching criteria.

Precision Routing Limitations

Precision Routing is available only for Agent PGs on CCE.

Cisco Outbound Option does not support Precision Routing. However, agents who participate in an outbound campaign or nonvoice activities (by using Skill Groups) can also handle inbound calls from a precision queue.

Throttling During Precision Queue Changes

A configuration update on a precision queue (from the API or the Unified CCE Administration tool) can result in many agents with changed precision queue associations. These updates could overload the system if done all at once. Therefore, the system moves the agents into and out of the precision queues gradually, based on available system resources.

You can submit another precision queue configuration update before an earlier update completes. If you submit the updates too quickly, the new update can cause the pending configuration updates to queue in the system. To avoid a backlog, the system rejects new precision queue configuration updates after reaching five concurrent pending updates. Once the pending precision queue updates fall below the threshold, the system accepts new configuration updates.

To mitigate possible overload conditions on the agent peripheral during these operations, the system limits the number of calls to the peripheral during an overload condition. When an overload occurs, the system stops sending Precision Routing calls to that peripheral for a short time.

Single Sign-On (SSO) Considerations

The Single Sign-on (SSO) feature authenticates and authorizes agent and supervisor access to the contact center solution applications and services. The authentication process validates the identity of a user: "you are who you say you are." The authorization process confirms that an authenticated user is permitted to perform the requested action: "you can do what you are asking to do." When you enable SSO in the contact center solution, users only sign in once to gain access to all their Cisco browser-based applications and services. Access to Cisco administrator applications is not available through SSO.

SSO requires the following:

- A third-party Identity Provider (IdP)
- A Cisco Identity Service (Cisco IdS) cluster

Note Synchronize the time in Cisco IdS and IdP for SSO to work effectively. It is recommended that the Cisco IdS and IdP are time-synchronized using NTP Server.

The SSO feature requires an IdP that complies with the Security Assertion Markup Language 2.0 (SAML v2) Oasis standard. The IdP stores user profiles and provides authentication services to support SSO sign-ins. For a current list of supported Identity Provider products and versions, see the *Compatibility Matrix* for your solution at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-device-support-tables-list.html.

The Cisco IdS cluster manages authentication for the contact center solution. The individual SSO-enabled applications and services manage authorization. The Cisco IdS cluster is a redundant pair with a publisher and subscriber. You can only perform most administration tasks on the publisher, but either node can issue or refresh access tokens. The cluster replicates configuration and authorization codes between all nodes.

When an SSO-enabled user signs in, the Cisco IdS interacts first with your IdP to authenticate the user. When the user is authenticated, the Cisco IdS confirms with the accessed Cisco services to confirm that the user is authorized for the requested role. When the user is both authenticated and authorized, the Cisco IdS issues an access token that allows the user to access the application. The access token enables the user to switch between the authorized contact center applications for that session without presenting credentials again.

Note The user credentials are only presented to the IdP. The contact center solution applications and services only exchange tokens; they do not see the users' information.

SSO Component Support

The following contact center solution components support SSO:

- Unified CCE—Agent and Supervisor interfaces
- Cisco Finesse—Agent and Supervisor interfaces
- Cisco Unified Intelligence Center—Agent and Supervisor interfaces
- Customer Collaboration Platform—Task Routing interface
- Enterprise Chat and Email—Agent and Supervisor interfaces through the ECE gadget for Cisco Finesse

SSO Message Flow

SSO uses Security Assertion Markup Language (SAML) to exchange authentication and authorization details between the enterprise identity provider (IdP) and the Cisco IdS.

When a user browses to a web page for an SSO-enabled service, the authentication request is redirected to the Cisco IdS. Cisco IdS generates a SAML authentication request and directs it to the Identity Provider. The IdP presents a sign-in page to the user at the browser to collect the user's credentials. After the IdP authenticates the user, the IdP issues a SAML assertion to the Cisco IdS. The assertion contains trusted statements about the user.

When the SAML assertion is received, the Cisco IdS uses the Open Authorization (OAuth) protocol to complete authorization with the requested service. The service may present an approval page to the user to enable specific resources.

Together SAML and OAuth make it possible for a user to authenticate while only exposing user credentials to the authentication provider. The username and password are only presented to the IdP. The contact center solution applications and services do not see the user information. Only the SAML assertion and the OAuth token are exchanged.

SSO Design Impacts

Single Sign-On Support and Limitations

Note the following points that are related to SSO support:

- To support SSO, enable the HTTPS protocol across the enterprise solution.
- SSO supports agents and supervisors only. SSO support is not available for administrators in this release.
- SSO supports multiple domains with federated trusts.
- SSO supports only contact center enterprise peripherals.
- SSO support is available for Agents and Supervisors that are registered to remote or main site PG in global deployments.

Note the following limitations that are related to SSO support:

- SSO support is not available for third-party Automatic Call Distributors (ACDs).

- The SSO feature does not support Cisco Finesse IP Phone Agent (FIPPA).
- The SSO feature does not support Cisco Finesse Desktop Chat.
- In Hybrid mode,
  - When an agent in SSO mode tries to log in to CUIC, and if the agent does not exist in CUIC, the agent cannot log in to CUIC.
  - When a Supervisor in SSO mode tries to log in to CUIC, and if the Supervisor user does not exist in CUIC, the Supervisor cannot log in to CUIC. For the Supervisor to log in to CUIC, perform Unified CCE User Integration. For more information on Unified CCE User Integration, see *Administration Console User Guide for Cisco Unified Intelligence Center* at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-intelligence-center/products-maintenance-guides-list.html.

Contact Center Enterprise Reference Design Support for Single Sign-On

Packaged CCE supports single sign-on for these reference designs:

- 2000 Agents
- 4000 Agents
- 12000 Agents

Coresidency of Cisco Identity Service by Reference Design

Reference Design	Packaged CCE Solution
2000 Agent	Cisco IdS is coresident with Unified Intelligence Center and Live Data on a single VM.
4000 Agent	Standalone Cisco IdS VM
12000 Agent	Standalone Cisco IdS VM

Reference Design Topology Support for SSO

The deployment topology specifies where you install the VMs for your contact center and how your agents connect to the sites. SSO is supported for components in the following Reference Design topologies:

- **Centralized**—You host both sides of the redundant components in the same site.

  Even when they are on the same LAN, the maximum round-trip time between the two sides is 80 ms.

- **Distributed**—You host each side of the redundant components in a different geographical sites.

  The maximum round-trip time between the two sides is 80 ms.

You can use single sign-on with remote agents using any of the supported Remote Office Options:

- Office with Agents
- Office with Agents and a Local Trunk
- Home Agent with Cisco Virtual Office

- Unified Mobile Agent

The maximum allowed round-trip time between any remote office and the main site is 200 ms.

User Management for SSO

In Unified CCE Administration, you can select three different SSO modes for your system:

- **SSO**—Enables SSO for all Agents and Supervisors.

  All users sign in using the IdS for authentication and authorization.

- **Non-SSO**—Disables SSO for all Agents and Supervisors.

  All users sign in using the existing Unified CCE local authentication and Active Directory.

- **Hybrid**—Supports a mixture of SSO-enabled and non-SSO users. In hybrid mode, you set the SSO mode for individual users using Unified CCE Configuration Manager tools. Each user signs in using their configured method.

  If you are enabling SSO in an existing deployment, use the Hybrid mode to gradually migrate agents to SSO while other agents continue to use local authentication.

The contact center enterprise user sign-in name must match the configured SAML claim rule for the Cisco IdS in your IdP.

- If your deployment is in a single domain, the sign-in name can be a simple user ID or a sign-in name in email format: user@cisco.com.
- If your deployment is across multiple domains, the sign-in name must be in email format. If your user sign-in names are simple User IDs, configure the agent LoginName in the Unified CCE database to email format.

The Unified CCE Administration Bulk Configuration tools include an SSO Migration tool. You can migrate groups of agents and supervisors to SSO accounts and, if necessary, change their usernames with that tool. The tool downloads a content file that includes records for agents and supervisors who have not been migrated to SSO accounts. In the content file, you specify SSO usernames for existing agents and supervisors and submit the file. When you update their usernames, the sign-in names in the database are also updated and the users are automatically enabled for SSO.

Qualified Identity Providers

If you use any Identity Provider (IdP) outside of the listed IdPs in the table below, Cisco IdS supports the IdP as long as the IdP is SAML 2.0 compliant and meets the following requirements described in the subsequent SAML Request and Response sections:

- SAML Request Attributes
- Expectations from SAML Response

IdP Metadata Schema

When you configure IdS and exchange Metadata between Cisco Identity Service (IdS) and the Identity Provider (IdP), ensure that the IdP Metadata file should confirm to the SAML metadata schema at:

https://docs.oasis-open.org/security/saml/v2.0/saml-schema-metadata-2.0.xsd

SAML Request Attributes

SAML request supports the following SAML 2.0 bindings:

- **HTTP-POST** binding
- NameIDFormat in SAML request must be **urn:oasis:names:tc:SAML:2.0:nameid-format:transient**

```
<samlp:AuthnRequest xmlns:samlp="urn:oasis:names:tc:SAML:2.0:protocol"
       ID="s25f4fb66688cf429e430034f4cceac00b6124570d" Version="2.0"
       IssueInstant="2018-10-29T10:01:39Z"
Destination="https://win-adfs30-151.uccxteam.com/adfs/ls/"
       ForceAuthn="false" IsPassive="false"
ProtocolBinding="urn:oasis:names:tc:SAML:2.0:bindings:HTTP-POST"
       AssertionConsumerServiceURL="https://ccxssodemo1.cisco.com:8553/ids/saml/response">
    <saml:Issuer
xmlns:saml="urn:oasis:names:tc:SAML:2.0:assertion">ccxssodemo1.cisco.com</saml:Issuer>
    <samlp:NameIDPolicy xmlns:samlp="urn:oasis:names:tc:SAML:2.0:protocol"
           Format="urn:oasis:names:tc:SAML:2.0:nameid-format:transient"
          SPNameQualifier="ccxssodemo1.cisco.com" AllowCreate="true"></samlp:NameIDPolicy>
</samlp:AuthnRequest>
```

Expectations from SAML Response

The following are the expectations from SAML Response:

- The entire SAML response (message and assertion) is signed or only the message is signed but not the SAML assertion alone is signed.
- SAML Assertion must not be encrypted.
- SAML response must be signed using **SHA-128**.
- NameIDFormat in SAML response must be **urn:oasis:names:tc:SAML:2.0:named-format:transient**.
- **uid** and **user_principal** attributes should be present in SAML assertion in the AttributeStatement section.

  The "uid" attribute value must be the user Id using which users log in to Cisco contact centre applications that are SSO enabled and the "user_principal" attribute value must be in uid@domain format.

```
<samlp:Response xmlns:samlp="urn:oasis:names:tc:SAML:2.0:protocol"
    Consent="urn:oasis:names:tc:SAML:2.0:consent:unspecified"
Destination="https://ids-ssp-node.cisco.com:8553/ids/saml/response"
    ID="_6a309495-d3c2-4a28-b8e3-289f8f5355bd"
InResponseTo="s21c84ba20862f573f5daec121c305ba6aac877843"
    IssueInstant="2017-08-10T13:20:26.556Z" Version="2.0">
    <Issuer
xmlns="urn:oasis:names:tc:SAML:2.0:assertion">http://ADFSServer.cisco.com/adfs/services/trust

    </Issuer>
    <ds:Signature xmlns:ds="http://www.w3.org/2000/09/xmldsig#">
        <ds:SignedInfo>
            <ds:CanonicalizationMethod Algorithm="http://www.w3.org/2001/10/xml-exc-c14n#"
 />
            <ds:SignatureMethod Algorithm="http://www.w3.org/2000/09/xmldsig#rsa-sha1" />
            <ds:Reference URI="#_6a309495-d3c2-4a28-b8e3-289f8f5355bd">
            ..........
            </ds:Reference>
        </ds:SignedInfo>
        ..........
        ..........
```

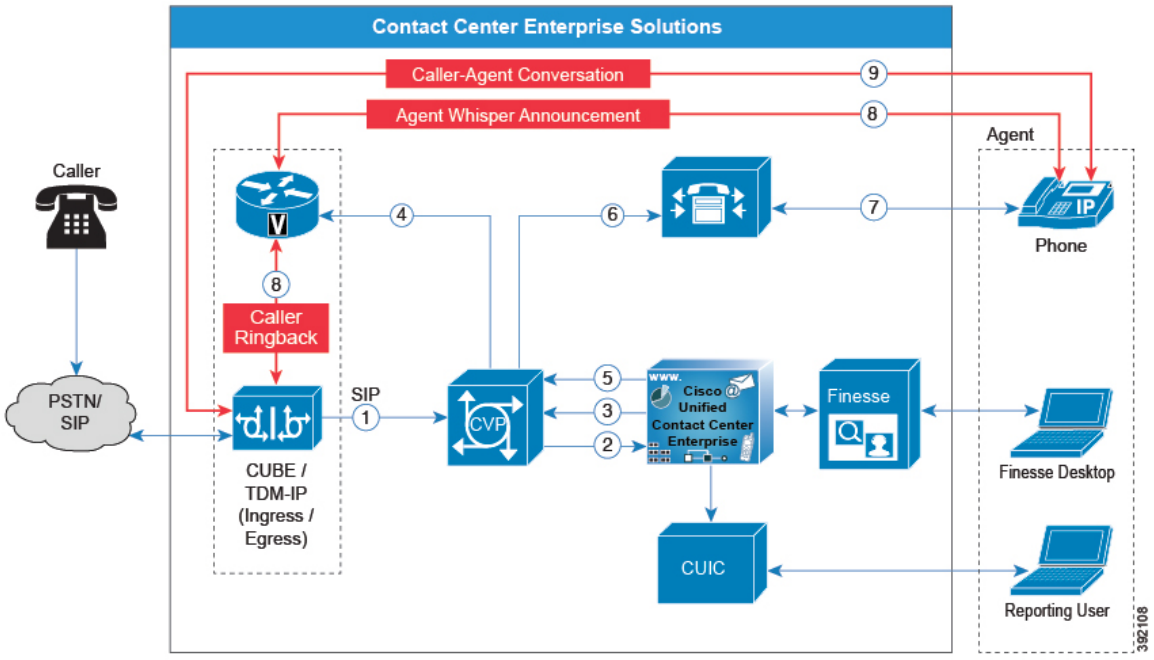
```
     </ds:Signature>
    <samlp:Status>
        <samlp:StatusCode Value="urn:oasis:names:tc:SAML:2.0:status:Success" />
    </samlp:Status>
    <Assertion xmlns="urn:oasis:names:tc:SAML:2.0:assertion"
ID="_df3bdbcf-a225-4e97-b00a-a199bdda3d2c"
        IssueInstant="2017-08-10T13:20:26.556Z" Version="2.0">
        <Issuer>http://ADFSServer.cisco.com/adfs/services/trust</Issuer>
        ............

        ............
            <NameID Format="urn:oasis:names:tc:SAML:2.0:nameid-format:transient"
                NameQualifier="http://ADFSSserver.cisco.com/adfs/services/trust"
                SPNameQualifier="ids-ssp-node.cisco.com">CISCO\Admin121</NameID>
            <SubjectConfirmation Method="urn:oasis:names:tc:SAML:2.0:cm:bearer">
                <SubjectConfirmationData
                    InResponseTo="s21c84ba20862f573f5daec121c305ba6aac877843"
                    NotOnOrAfter="2017-08-10T13:25:26.556Z"
                    Recipient="https://ids-ssp-node.cisco.com:8553/ids/saml/response" />
            </SubjectConfirmation>
        </Subject>
        <Conditions NotBefore="2017-08-10T13:20:26.556Z"
            NotOnOrAfter="2017-08-10T14:20:26.556Z">
            <AudienceRestriction>
                <Audience>ids-ssp-node.cisco.com</Audience>
            </AudienceRestriction>
        </Conditions>
        <AttributeStatement>
            <Attribute Name="user_principal">
                <AttributeValue>Admin121@cisco.com</AttributeValue>
            </Attribute>
            <Attribute Name="uid">
                <AttributeValue>Admin121</AttributeValue>
            </Attribute>
        </AttributeStatement>
        <AuthnStatement AuthnInstant="2017-08-10T13:18:12.086Z"
            SessionIndex="_df3bdbcf-a225-4e97-b00a-a199bdda3d2c">
            <AuthnContext>

<AuthnContextClassRef>urn:oasis:names:tc:SAML:2.0:ac:classes:PasswordProtectedTransport</AuthnContextClassRef>

            </AuthnContext>
        </AuthnStatement>
    </Assertion>
</samlp:Response>
```

Whisper Announcement Considerations

Whisper Announcement plays a brief, prerecorded message to an agent just before the agent connects with each caller. Use the announcement to quickly orient the agent to the type of call. The announcement plays only to the agent; the caller hears ringing while the announcement plays.

Whisper Announcement Call Flows

Figure 123: Whisper Announcement Call Flow



The standard call flow with Whisper Announcement is as follows:

1. Incoming call arrives at CVP from the carrier.
2. CVP sends the call to Unified CCE.
3. Unified CCE instructs CVP to queue the call.
4. CVP sends the call to the Voice Browser.
5. Unified CCE sends the agent label with the whisper announcement prompt.
6. CVP sends the call to Unified CM.
7. Unified CM sends the call to the agent phone.
8. The caller continues to hear ringback. The agent hears the whisper announcement.
9. When the whisper announcement ends, the caller connects to the agent.

Whisper Announcement Design Impacts

Whisper Announcement has these limitations:

- Announcements do not play for outbound calls made by an agent. The announcement plays for inbound calls only.
- For Whisper Announcement to work with agent-to-agent calls, use the SendToVRU node before you transfer the call to the agent. Transfer the call to Unified CVP before you transfer the call to another

agent. Then, Unified CVP can control the call and play the announcement, regardless of which node transfers the call to Unified CVP.

- CVP Refer Transfers do not support Whisper Announcement.
- Whisper Announcement supports Silent Monitoring. However, for Unified Communications Manager-based Silent Monitoring, supervisors cannot hear the announcements themselves. The supervisor desktop dims the Silent Monitor button while an announcement plays.
- Only one announcement can play for each call. While an announcement plays, you cannot put the call on hold, transfer, or conference; release the call; or request supervisor assistance. These features become available again after the announcement completes.
- The codec settings for Whisper Announcement recording and the agent's phone must match. For example, if Whisper Announcement is recorded in G.711 ALAW, the phone must also be at G.711 ALAW. If Whisper Announcement is recorded in G.729, the phone must support or connect using G.729.
- In an IPv6-enabled environment, Whisper Announcement might require extra Media Termination Points (MTPs).

Whisper Announcement Media Files

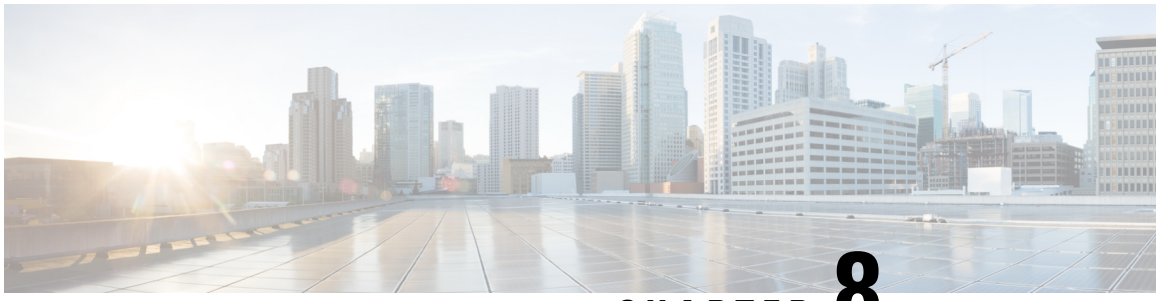
You store and serve your Whisper Announcement audio files from the Unified Contact Center Enterprise (Unified CCE) media server. This feature supports only the wave (`.wav`) file type. The maximum play time for a Whisper Announcement is subject to a timeout. Playback terminates at the timeout regardless of the actual length of the audio file. The timeout is 15 seconds. In practice, you may want your messages to be much shorter than that, 5 seconds or less, to shorten your call-handling time.

Whisper Announcement with Transfers and Conferences

When an agent transfers or starts a conference call to another agent, the second agent hears an announcement if the second agent's number supports Whisper Announcement. For consultative transfers or conferences, while the announcement plays, the caller hears whatever generally plays during hold. The first agent hears ringing. In the case of blind transfers, the caller hears ringing while the announcement plays.

Whisper Announcement Sizing Considerations

The impact of Whisper Announcement on solution component sizing is not as significant as the impact caused by Agent Greeting.



CHAPTER 8

Bandwidth, Latency, and QoS Considerations

Bandwidth, Latency, and QoS for Core Components

Sample Bandwidth Usage by Core Components

This table presents some sample bandwidth usage by the core components from our test environment.

Table 49: Sample Bandwidth Usage

Contact Center Enterprise Components	Public Network bandwidth (KBps)			Private Network bandwidth (KBps)			Operating Conditions
	Peak	Average	95th Percentile	Peak	Average	95th Percentile	
Router	2307	1189	1173	1908	1048	1024	12,000 agents; 105 CPS(Includes 10% Transfer and 5% Conference), ECC: 5 scalars @ 40 bytes each; 200 Reporting users at max query load.
Logger	14624	2696	8718	12351	2184	7795	
AW-HDS	4113	1522	3215	NA	NA	NA	
HDS-DDS	3323	512	1627	NA	NA	NA	
Cisco Identity Server(IdS)	35	26	33	NA	NA	NA	
Large Live Data(4K to 12K)	47073	6018	8079	NA	NA	NA	

Contact Center Enterprise Components	Public Network bandwidth (KBps)			Private Network bandwidth (KBps)			Operating Conditions
	Peak	Average	95th Percentile	Peak	Average	95th Percentile	
Rogger	6410	2314	3498	5875	1987	3185	4,000 agents; 30 CPS; ECC; 5 scalars @ 40 bytes each; 200 Reporting users at max query load.
AW-HDS-DDS	4891	2476	3651	NA	NA	NA	
Cisco Identity Server(IdS)	112	88	105	NA	NA	NA	
Small Live Data(upto 4K)	25086	3998	5487	NA	NA	NA	
Rogger	2561	1040	1338	2141	789	1443	2,000 agents; 15 CPS; ECC; 5 scalars @ @ 40 bytes each; 200 Reporting users at max query load.
AW-HDS-DDS	4014	2881	3174	NA	NA	NA	
CUIC-LD-Ids	105544	8496	10236	NA	NA	NA	
Medium PG	12012	7702	10486	8795	6478	7846	
CUIC	10620	4351	7947	NA	NA	NA	
Finesse	18449	16125	17381	NA	NA	NA	
CVP Call/VXML	2198	2141	2196	NA	NA	NA	
CVP Reporting	2008	1980	2004	NA	NA	NA	

Bandwidth, Latency, and QoS for Ingress, Egress, and VXML Gateway

Your network latency between the Voice Browser and CVP VXML Server cannot exceed 200-ms RTT. You can use the Global Deployment to help maintain the necessary latency.

Bandwidth, Latency, and QoS for Unified CVP

Bandwidth Considerations for Unified CVP and VVB

The ingress gateway and Voice Browser are separated from the servers that provide them with media files, VXML documents, and call control signaling. These factors make the bandwidth requirement for the Unified CVP.

For example, assume that all calls have 1 minute of VRU treatment and a single transfer to an agent for 1 minute. Each branch has 20 agents. If each agent handles 30 calls per hour, you have 600 calls per hour per branch. The call average rate is 0.166 calls per second (CPS) per branch.

Even a small change in these variables can have a large impact on sizing. Remember that 0.166 calls per second is an average for the entire hour. Typically, calls do not come in uniformly across an entire hour, and there are usually peaks and valleys within the busy hour. External factors can also affect the call volume. For example, bad weather increases call volumes for business like airlines and promotions can increase call volumes for retailers. Find the busiest traffic period for your business, and calculate the call arrival rate based on the worst-case scenario.

VXML Documents

A VXML document is generated for every prompt that is played to the caller. This document is generated based on voice application scripts that you write using either Unified ICM scripts or Cisco Unified Call Studio, or both. A VXML document varies in size, depending on the type of prompt being used. For example, menu prompts with multiple selections are larger in size than prompts that play announcements only.

Note The approximate size of a VXML document for a Call Server or a VXML Server and the gateway is 7 kilobytes.

You can calculate bandwidth in the following ways:

Bandwidth Estimated by Prompts

You can estimate the bandwidth for a branch office as follows:

```
CPS * Bits per Prompt * Prompts per call = Bandwidth in bps
```

For the previous example, consider a VXML document of 7 kilobytes:

```
7,000 bytes * 8 bits/byte = 56,000 bits per prompt
(0.166 calls/second) * (56,000 bits/prompt) * (Number of prompts / call) = bps per
branch
```

Bandwidth Estimated by VXML Documents

Use the VXML document sizes listed in the following table to calculate the required bandwidth. The document sizes in the following table are measured from the VXML Server to the Voice Browser.

Table 50: Approximate Size of VXML Document Types

VXML Document Type	Approximate Size in bytes
Root document (one required at beginning of a call)	19,000
Subdialog_start (at least one per call at beginning of a call)	700
Query gateway for Call-ID and GUID (one required per call)	1300
Menu (increases in size with the number of menu choices)	1000 + 2000 per menu choice
Play announcement (a simple .wav file)	1100

VXML Document Type	Approximate Size in bytes
Cleanup (one required at the end of a call)	4000

Note For more complex solutions, this second method yields a better estimate of the required bandwidth than estimating by prompts.

Media File Retrieval

You can store media files, or *prompts*, locally in flash memory for IOS Voice Gateway and in the file system for Cisco VVB. Storing them locally eliminates bandwidth considerations. However, it is difficult to maintain these prompts because a prompt that requires changes must be replaced on every router or VVB. Local storage of these prompts on an HTTP media server (or an HTTP cache engine) enables the gateway to locally cache voice prompts after retrieval. An HTTP media server can cache multiple prompts, depending on the number and size of the prompts. The refresh period for the prompts is defined on the HTTP media server. The bandwidth usage is limited to the initial load of the prompts at each gateway, including the periodic updates after the expiration of the refresh interval.

Note You cannot disable the HTTP Cache in VVB.

Not caching prompts at the VXML Gateway has significant impacts:

- It degrades Cisco IOS performance by 35-45%.
- It requires extra bandwidth. For example, if you have 50 prompts with an average size of 50 KB and a refresh interval of 15 minutes, the average bandwidth usage is:

```
(50 prompts) * (50,000 bytes/prompt) * (8 bits/byte) = 20,000,000 bits
(20,000,000 bits) / (900 second) = 22.2 kbps per branch
```

Note Bandwidth considerations for VVB include bandwidth for VXML documents, Media File retrieval and RTP streams for G.711 and G.729 voice traffic.

Network Link Considerations for Unified CVP

For Unified CVP, you can group WAN and LAN traffic into the voice traffic, the call control traffic, and the data traffic.

Voice Traffic

Voice calls consist of Real-Time Transport Protocol (RTP) packets. These packets contain voice samples that are transmitted into the following:

- Between the PSTN Ingress Gateway or originating IP phone over a WAN or LAN connection and one of the following:

- Another IP phone whether or not collocated (located on the same LAN) with the Ingress Gateway or calling IP phone.
- A front-end Egress Gateway for a TDM ACD (for legacy ACDs or VRUs). The Egress Gateway might or might not be collocated with the Ingress Gateway.
- A Voice Browser that performs prompt-and-collect treatment. The Voice Browser can be the same or a different Ingress Gateway. In either case, both the Ingress Gateway and Voice Browser are collocated.

- Between the Voice Browser and the ASR or TTS Server. The RTP stream between the Voice Browser and ASR/TTS server must be G.711.

Call Control Traffic with SIP

Unified CVP works in Call Control mode or Signaling mode with three types of VoIP endpoints: Cisco IOS Voice Gateways and Unified Communications Manager. Call Control traffic flows over a WAN or LAN between the following endpoints:

- **Call Server and Inbound Calls**—The inbound call can come from Unified CM, a Cisco IOS Voice Gateway, or another SIP device.
- **Call Server and Outbound Calls**—The outbound call can come from Unified CM or a Cisco IOS Voice Gateway. The Egress Gateway can be a VXML Gateway that provides prompt-and-collect treatment to the caller. It can also be the target of a transfer to an agent (CCE or TDM) or a legacy TDM VRU.

Call Control Traffic with VRU PG

The Call Server and the CCE VRU PG communicate using the GED-125 protocol. The GED-125 protocol includes the following features:

- Notification messages that control the caller experience when a call arrives.
- Instructions to transfer or disconnect the caller.
- Instructions that control the VRU treatment the caller experiences.

The VRU PG connects to Unified CVP over a LAN connection. However, in deployments that use clustering over the WAN, Unified CVP can connect to the redundant VRU PG across the WAN.

The bandwidth between the Central Controller and VRU PG is similar to the bandwidth between the VRU PG and Unified CVP.

If the redundant VRU PG pair is split across the WAN, the total bandwidth is double. You need the reported bandwidth for the Central Controller-to-VRU-PG connection. You need the same amount of bandwidth for the VRU-PG-to-Unified-CVP connection.

Media Resource Control Protocol Traffic

The VXML Gateway and Cisco Virtualized Voice Browser communicate with ASR/TTS Servers using both Media Resource Control Protocol (MRCP) v1.0 and v2. This protocol establishes connections to the ASR/TTS Server, such as Nuance. The connection can be over LAN or WAN.

Note Cisco does not test or qualify speech applications in WAN environment. For guidelines on design, support over WAN and associated caveats, see the vendor-specific documentation. TAC provides limited support (as in the case of any third-party interoperability certified products) on issues related to speech applications.

Central Controller to VRU PG Traffic

There is no sizing tool for communications between the Central Controller and the VRU PG. However, the tool for estimating bandwidth between the Central Controller and the IP IVR PG produces accurate measurements for Unified CVP, if you substitute one value.

For the **Average number of RUN VRU SCRIPT nodes** field, substitute the number of CCE script nodes that interact with Unified CVP. Nodes that can interact with Unified CVP are:

- Run External Script
- Label
- Divert Label
- Queue to Skill Group
- Queue to Agent
- Agent
- Release
- Send to VRU
- Translation Route to VRU

The connection can be over a WAN or a LAN.

Data Traffic

Data traffic includes VXML documents and prerecorded media files that are returned as a result of HTTP requests. Voice Browser runs the following requests:

- **Media files in an HTTP request to a Media File Server**—The Media File Server response returns the media file in the body of the HTTP message. The Voice Browser then converts the media files to Real-Time Transport Protocol (RTP) packets and plays them to a caller. The connection can be over a WAN or a LAN.
- **VXML documents from the CVP Server**—In this case, the connection can be over a WAN or a LAN.

Bandwidth Sizing

Generally, a distributed topology is the most bandwidth intensive for Unified CVP. The Ingress Gateway and Voice Browser are separated from the servers that provide the media files, VXML documents, and call control signaling.

Note Recall the earlier example of all calls have 1 minute of VRU treatment and a single transfer to an agent for 1 minute. Each branch has 20 agents, and each agent handles 30 calls per hour for a total of 600 calls per hour per branch. The call average rate is 0.166 calls per second (CPS) per branch.

SIP Signaling

SIP is a text-based and signaling communications protocol for controlling multimedia communication sessions, such as VoIP networks. You also use SIP to create, modify, and terminate sessions consisting of media streams. These sessions include internet phone calls, multimedia distribution, and multimedia conferences. You can use SIP for two-party (unicast) or multiparty (multicast) sessions.

A typical SIP call flow uses about 17,000 bytes per call. Using the previous bandwidth formulas based on calls per second, the average bandwidth usage is:

```
(17,000 bytes/call) * (8 bits/byte) = 136,000 bits per call
(0.166 calls/second) * (136 kilobits/call) = 22.5 average kbps per branch
```

G.711 and G.729 Voice Traffic

Unified CVP supports both G.711 and G.729 codecs. However, both call legs and all VRUs on a given call must use the same voice codec. For speech recognition, the ASR/TTS server only supports G.711. For information on the voice RTP streams, see *Cisco Collaboration Systems Solution Reference Network Designs (SRND)* at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

Network Latency

After the proper application bandwidth and QoS policies are in place, consider the network latency in a distributed CVP deployment. With sufficient network bandwidth, the primary contributor to latency is the distance between the Voice Browser and the Call Server or VXML Server. In distributed CVP deployments, minimize the latency and understand its effect on solution performance.

Network latency affects a distributed CVP deployment in the following ways:

- It affects the end-user calling experience when the network latency is between CVP components. Call signaling latency with SIP between the Call Servers and voice gateways affects the call setup time. Latency can add a period of silence during this setup. It includes the initial call setup and subsequent transfers or conferences that are part of the final call flow.
- It significantly affects the download time for VXML application documents, and has a pronounced effect on the ultimate caller experience.

The following system configuration changes can reduce WAN delays from geographic separation of the Voice Browser from the VXML Server:

1. Provide audio to the caller during periods of silence.

   The following settings provide ringback and audio during times of dead air so that the caller does not disconnect:

   - To add a ringback tone during longer than usual call setup times with VRU, on the survivability service, keep the `wan-delay-ringback` setting at 1.

- Add the VRU subsystem settings for `IVR.FetchAudioDelay` and `IVR.FetchAudioMinimum`. These WAN Delay settings are required when the root document fetch is delayed over the WAN link.
- Specify the value for `IVR.FetchAudio` as follows: `IVR.Fetchaudio= flash:holdmusic.wav`. Leave the default empty so that nothing is played in a usual scenario.
- Retain the default setting of 2 to avoid a blip sound in a usual network scenario.
- Set WAN Delay to zero to play a `holdmusic.wav` immediately for a minimum of 5 seconds.
- Use ECC variables, such as `user.microapp.fetchdelay`, `user.microapp.fetchminimum`, and `user.microapp.fetchaudio`, to override ECC variable values in between invocations of `getSpeechExternal` microapps.

Note You cannot use ECC variables while a call is at the Virtualized Voice Browser.

2. Enable Path MTU Discovery on the IOS Voice Gateways.

   On the IOS Voice Gateways, add the `ip tcp path-mtu-discovery` command.

   The Path MTU Discovery method maximizes the use of available bandwidth in the network between the endpoints of a TCP connection.

3. Minimize round trips between the VXML Server and the ICM script.

   When control is passed from a running VXML Server application back to the ICM script, you incur a significant WAN delay.

   After the VXML Server application starts to run, minimize the number of trips back to the Unified CCE script. Each round trip between the VXML Server and the Unified CCE script incurs a delay. It establishes two new TCP connections and HTTP retrieval of several VXML documents, including the VXML Server root document.

4. Decrease the size of the VXML Server root document.

   On the VXML Server, in your gateway adapter plugin.xml file change:

   ```
   <setting name="vxml_error_handling">default</setting>
   ```

   To:

   ```
   <setting name="vxml_error_handling">minimal</setting>
   ```

   For example, the location of the `plugin.xml` file for the CISCO DTMF 1 GW adapter is `Cisco\CVP\VXMLServer\gateways\cisco_dtmf_01\6.0.1\plugin.xml`.

Note HTTP transfers VXML documents and other media files that are played to the caller. For the best end-user calling experience, treat the HTTP traffic with a priority higher than that of usual HTTP traffic in an enterprise network. If possible, treat this HTTP traffic the same as CVP call signaling traffic. As a workaround, you can move the VXML Server to the same local area as the Voice Browser, or use Wide Area Application Service (WAAS).

Port Usage and QoS Settings for Unified CVP

The Call Server marks only the QoS DSCP for SIP messages, if done via Windows policy. If you need QoS for CVP traffic across a WAN, configure network routers for QoS using the IP address and ports to classify and mark the traffic. The following table outlines the necessary configuration.

The CVP-Data queue and the Signaling queue are not a priority queue in Cisco IOS router terminology. Use the priority queue for the voice or other real-time traffic. Reserve some bandwidth based on the call volume for call signaling and CVP traffic.

Component	Port	Queue	PHB	DSCP	Maximum Latency (Round Trip)
Media Server	TCP 80	CVP-Data	AF11	10	1 s
Unified CVP Call Server, SIP	TCP or UDP 5060	Call Signaling	CS3	24	200 ms
Unified CVP IVR Service	TCP 8000	CVP-Data	AF11	10	1 s
Unified CVP VXML Server	TCP 7000	CVP-Data	AF11	10	1 s
Ingress Voice Gateway, SIP	TCP or UDP 5060	Call Signaling	CS3	24	200 ms
Voice Browser, SIP	TCP or UDP 5060	Call Signaling	CS3	24	200 ms
SIP Proxy Server	TCP or UDP 5060	Call Signaling	CS3	24	200 ms
MRCP	TCP 554	Call Signaling	CS3	24	200 ms

Bandwidth Provisioning and QoS Considerations for a WAN

Some CVP deployments have all the components centralized. Those deployments use a LAN structure, so WAN network traffic is not an issue. A WAN might impact your bandwidth and QoS for CVP in the following scenarios:

- A distributed CVP deployment with a WAN between the Ingress Gateways and the Unified CVP servers
- A CVP deployment with a WAN between the Ingress Gateway and the agent.

CVP considers QoS in the following way:

- CVP has no private WAN network structure. When required, WAN activity is conducted on a converged WAN network structure.
- CVP does not use separate IP addresses for high- and low-priority traffic.

Note Resource Reservation Protocol (RSVP) is used for call admission control. Routers also use it to reserve bandwidth for calls. RSVP is not qualified for call control signaling through the Unified CVP Call Server in SIP. For call admission control, the solution is to employ Locations configuration on CVP and Unified CM.

VAV and Agent Answers

The following are the bandwidth details:

- Bandwidth per VAV call is 106 Kbps.
- Bandwidth per Agent Answers call is 183 Kbps.

Bandwidth, Latency, and QoS for Packaged CCE

Packaged CCE Bandwidth and Latency Requirements

The amount of traffic sent between the Central Controllers (routers) and PGs is largely based on the call load at that site. Transient boundary conditions, like configuration loads, and specific configuration sizes also affect the amount of traffic. Bandwidth calculators and sizing formulas can project bandwidth requirements more accurately.

Bandwidth calculations for a site with an ACD and a VRU must account for both peripherals. Use 1000 bytes per call as a rule, but monitor the actual behavior once the system is operational to ensure that enough bandwidth exists. Based on that rule, a site that has four peripherals, each taking 10 calls per second, requires 320 kbps of bandwidth. (Packaged CCE meters data transmission statistics at both the Central Controller and PG sides of each path.)

As with bandwidth, Packaged CCE requires specific latency on the network links to function as designed. The private network between redundant Central Controller and PG nodes has a maximum round-trip latency of 80 ms. The PG-to-CC public network has a maximum round-trip latency of 400 ms to perform as designed. Meeting or exceeding these latency requirements is important for Packaged CCE post-routing and translation routes.

Note In general, Agent Greeting feature requires shorter latency across the system. For example, the public network has a maximum round-trip latency of 100 ms to support Agent Greeting feature as designed.

Packaged CCE bandwidth and latency design depends on an underlying IP prioritization scheme. Without proper prioritization in place, WAN connections fail.

Depending on the final network design, your IP queuing strategy in a shared network must achieve Packaged CCE traffic prioritization concurrent with other non-DNP traffic flows. This queuing strategy depends on traffic profiles and bandwidth availability. Success in a shared network cannot be guaranteed unless the stringent bandwidth, latency, and prioritization requirements of the solution are met.

Agent Desktop to Call Servers and Agent PGs

There are many factors to consider for the traffic and bandwidth requirements between desktops and CCE Call Servers and Agent PGs. The VoIP packet stream bandwidth is the predominant contributing factor to

bandwidth usage. But, there are other factors such as the call control, agent state signaling, silent monitoring, recording, and statistics.

To calculate the required bandwidth for the Cisco Finesse desktop, see the *Finesse Bandwidth Calculator* at http://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-technical-reference-list.html.

The latency between the server and agent desktop is 400-ms round-trip time for Cisco Finesse.

Central Controller Components

CCE Central Controllers (Routers and Loggers) require a separate private network link between the redundant pairs. Latency across the private network must not exceed an 80-ms round trip.

Private Network Bandwidth Requirements for Packaged CCE

Use this worksheet to help compute the link and queue sizes for the private network.

Note Minimum link size in all cases is 1.5 Mbps (T1).

Component	Effective BHCA (bps)	Multiplication Factor	Recommended Link (bps)	Multiplication Factor	Recommended Queue (bps)	
Central Controller		* 30		* 0.8		Total Central Controller High-Priority Queue Bandwidth
Unified CM PG		* 100		* 0.9		Add these numbers together to get the total PG High-Priority Queue Bandwidth
Unified VRU PG		* 120		* 0.9		
Unified CVP Variables		* ((Number of Variables * Average Variable Length)/40)		* 0.9		
		Total Link Size				Total PG High-Priority Queue Bandwidth

For a single private network link between the sites, add all link sizes together and use the Total Link Size at the bottom of the table. Otherwise, use the first row for the Central Controller private network and the total of the other rows for the PG private network.

Effective BHCA (effective load) on all similar components that are split across the WAN is defined as follows:

- **Central Controller**—This value is the total BHCA on the call center, including conferences and transfers. For example, 10,000 BHCA ingress with 10% conferences or transfers is an effective 11,000 BHCA.

- **Unified CM PG**—This value includes all calls that come through CCE Route Points that the Unified CM controls and that are transferred to agents. This assumes that each call comes into a route point and is eventually sent to an agent. For example, 10,000 BHCA ingress calls to a route point and transferred to agents, with 10% conferences or transfers, is an effective 11,000 BHCA.
- **Unified VRU PG**—This value is the total BHCA for the call treatment and queuing coming through CVP. The calculation assumes 100% treatment. For example, 10,000 BHCA ingress calls, with all of them receiving treatment and 40% being queued, is an effective 14,000 BHCA.
- **Unified CVP Variables**—The number of Call and ECC variables and the variable lengths for all calls routed through CVP.

Example of a Private Bandwidth Calculation

The following table shows an example calculation for a combined dedicated private link with the following characteristics:

- BHCA coming into the contact center is 10,000.
- CVP treats all calls and 40% are queued.
- All calls are sent to agents unless abandoned. 10% of calls to agents are transfers or conferences.
- There are four Unified CVPs used to treat and queue the calls, with one PG pair supporting them.
- There is one Unified CM PG pair for a total of 900 agents.
- Calls have ten 40-byte Call Variables and ten 40-byte ECC variables

Component	Effective BHCA (bps)	Multiplication Factor	Recommended Link (bps)	Multiplication Factor	Recommended Queue (bps)	
Central Controller	11,000	* 30	330,000	* 0.8	264,000	Total Central Controller High-Priority Queue Bandwidth
Unified CM PG	11,000	* 100	1,100,000	* 0.9	990,000	Add these numbers together to get the total PG High-Priority Queue Bandwidth
Unified VRU PG	0	* 120	0	* 0.9	0	
Unified CVP Variables	14,000	* ((Number of Variables * Average Variable Length)/40)	280,000	* 0.9	252,000	
		Total Link Size	1,710,000		1,242,000	Total PG High-Priority Queue Bandwidth

For the combined dedicated link in this example, the results are as follows:

- Total Link Size = 1,710,000 bps
- Central Controller high-priority bandwidth queue of 264,000 bps
- PG high-priority queue bandwidth of 1,242,000 bps

If this example is for a solution with two separate links, Central Controller private and PG private, the link sizes and queues are as follows:

- Central Controller link of 330,000 bps (actual minimum link is 1.5 Mb, as defined earlier), with a high-priority bandwidth queue of 264,000 bps
- PG link of 1,380,000 bps, with a high-priority bandwidth queue of 1,242,000 bps

When using Multilink Point-to-Point Protocol (MLPPP) for private networks, set the following attributes for the MLPPP link:

- Use per-destination load balancing instead of per-packet load balancing.
- Enable Point-to-Point Protocol (PPP) fragmentation to reduce serialization delay.

Note You must have two separate multilinks with one link each for per-destination load balancing.

Bandwidth Requirement for Clustering over WAN

Bandwidth must be guaranteed across the highly available (HA) WAN for all CCE private, public, CTI, and Unified Communications Manager intracluster communication signaling (ICCS). Moreover, bandwidth must be guaranteed for any calls going across the highly available WAN. Minimum total bandwidth required across the highly available WAN for all CCE signaling is 2 Mbps.

VRU PG to Unified CVP

Currently, no tool exists that specifically addresses communication between the VRU PG and Unified CVP. However, the tool mentioned in the previous section produces a fairly accurate measurement of the needed bandwidth. The bandwidth consumed between the CCE Central Controller and VRU PG is similar to the bandwidth consumed between the VRU PG and CVP.

If the VRU PGs are split across the WAN, the total bandwidth required is double what the tool reports. You need the reported bandwidth for the Central-Controller-to-PG link and again for the PG-to-Unified-CVP link.

CTI Server to Cisco Finesse

To determine the bandwidth required where Cisco Finesse connects to the CTI server over a WAN link, use the *Finesse Bandwidth Calculator* at http://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-technical-reference-list.html.

Unified CM Intracluster Communication Signaling (ICCS)

Contact center enterprise solutions require more bandwidth for Intracluster Communication Signaling (ICCS) between subscribers than a Unified Communications Manager-only deployment. CCE requires more call redirects and extra CTI/JTAPI communications for the intracluster communications. Use the following formulae to calculate the required bandwidth for the ICCS and database traffic between subscribers in CCE:

- Intracluster Communications Signaling (ICCS)

  ```
  Total Bandwidth (Mbps) = (Total BHCA) / 10,000) * [1 + (0.006 * Delay)]
  ```

  Where *Delay* = Round-trip-time delay in msec

  This value is the bandwidth required between each Unified CM subscriber that is connected to Voice Gateways, agent phones, and Agent PGs. The minimum value for this link is 1.544 Mbps.

Note This formula assumes a BHCA of 10,000 or more. For a BHCA of less than 10,000, use the minimum of 1.544 Mbps.

- Database and other communications

  1.544 Mbps for each subscriber remote from the publisher

  The BHCA value to use for this ICCS formula is the total BHCA for all calls coming into the contact center.

- CTI ICCS

  Bandwidth (Mbps) = (Total BHCA/10,000) * 0.53

These bandwidth requirements assume proper design and deployment. Inefficient design (for example, if ingress calls to Site 1 are treated in Site 2) causes more intracluster communications, possibly exceeding the defined bandwidth requirements.

QoS Considerations for Packaged CCE

This section presents QoS marking, queuing, and shaping guidelines for both the Packaged CCE public and private network traffic. Provisioning guidelines are presented for the network traffic flows over the WAN, including how to apply proper Quality of Service (QoS) to WAN traffic flows. Adequate bandwidth provisioning and implementation of QoS are critical components in the success of contact center enterprise solutions.

Generally, your contact center enterprise WAN network structure uses separate links for both its Private and Public networks. For optimal network performance characteristics (and route diversity for the fault-tolerant fail-overs), Packaged CCE requires dedicated private facilities, redundant IP routers, and appropriate priority queuing.

Enterprises deploying networks that share multiple traffic classes prefer to maintain their existing infrastructure rather than revert to an incremental, dedicated network. Convergent networks offer both cost and operational efficiency, and such support is a key aspect of Cisco Powered Networks.

You can deploy Packaged CCE with a convergent QoS-aware public network and a convergent QoS-aware private network environment. But, your solution must meet the stringent latency and bandwidth requirements.

Packaged CCE uses the Differentiated Services (DiffServ) model for QoS. DiffServ categorizes traffic into different classes and applies specific forwarding treatments to the traffic class at each network node.

Where to Mark Traffic

In planning QoS, a question often arises about whether to mark traffic in CCE or at the network edge. Each option has its pros and cons. Marking traffic in CCE saves the access lists for classifying traffic in IP routers and switches.

There are several disadvantages to marking traffic in CCE. First, you change each PG separately to change the marking values for the public network traffic. Second, you enable QoS trust on the access-layer routers and switches, which can open the network to malicious packets with inflated marking levels.

Note In Windows, you can use the Group Policy Editor to apply a QoS policy to apply DSCP Level 3 markings to packets. You can also administer these policies through the Active Directory Domain Controller. This may simplify the administration issue. For more information, see appropriate Microsoft documentation.

In contrast, marking traffic at the network edge allows for centralized and secured marking policy management. There is no need to enable trust on access-layer devices. You have a little overhead to define access lists to recognize CCE packets. Although they are provided in the tables for reference purposes, do not use port numbers in the access lists for recognizing CCE traffic. The port numbers make the access lists complex. You must modify the access lists every time that you add a new customer instance to the system.

How to Mark Traffic

The default CCE QoS markings can be overwritten if necessary. These tables show the default markings, latency requirement, IP address, and port for each priority flow. In these tables, *i#* is the customer instance number. In the public network, the medium-priority traffic is sent with the high-priority public IP address and marked the same as the high-priority traffic. But, in the private network, the medium-priority traffic is sent with the non-high-priority private IP address and marked the same as the low-priority traffic.

For details about Cisco Unified Communications packet classifications, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/c/en/us/td/docs/voice_ip_comm/uc_system/design/guides/UCgoList.html.

Note Cisco has begun to change the marking of voice control protocols from DSCP 26 (PHB AF31) to DSCP 24 (PHB CS3). However, many products still mark signaling traffic as DSCP 26 (PHB AF31). Therefore, in the interim, reserve both AF31 and CS3 for call signaling.

Table 51: Public Network Traffic Markings (Default) and Latency Requirements

Priority	Server-Side IP Address and Port	One-Way Latency Requirement	DSCP / 802.1p Marking
High	IP address: Router's high-priority public IP address TCP port: • 40003 + (i# * 40) for DMP high-priority connection on A • 41003 + (i# * 40) for DMP high-priority connection on B UDP port: 39500 to 39999 for UDP heartbeats.	200 ms	AF31 / 3

Priority	Server-Side IP Address and Port	One-Way Latency Requirement	DSCP / 802.1p Marking
Medium	IP address: Router's high-priority public IP address TCP port: • 40017 + (i# * 40) for DMP high-priority connection on A • 41017 + (i# * 40) for DMP high-priority connection on B UDP port: 39500 to 39999 for UDP heartbeats.	1000 ms	AF31 / 3
Low	IP address: Router's non-high-priority public IP address TCP port: • 40002 + (i# * 40) for DMP high-priority connection on A • 41002 + (i# * 40) for DMP high-priority connection on B UDP port: 39500 to 39999 for UDP heartbeats.	5 seconds	AF11 / 1

Table 52: Router Private Network Traffic Markings (Default) and Latency Requirements

Priority	Server-Side IP Address and Port	One-Way Latency Requirement	DSCP / 802.1p Marking
High	IP address: Router's high-priority private IP address TCP port: 41005 + (i# * 40) for MDS high-priority connection UDP port: 39500 to 39999 for UDP heartbeats	40 ms	AF31 / 3

Priority	Server-Side IP Address and Port	One-Way Latency Requirement	DSCP / 802.1p Marking
Medium	IP address: Router's non-high-priority private IP address TCP port: 41016 + (i# * 40) for MDS medium-priority connection	1000 ms	AF11/1
Low	IP address: Router's non-high-priority private IP address TCP port: • 41004 + (i# * 40) for MDS low-priority connection • 41022 + (i# * 40) for CIC StateXfer connection • 41021 + (i# * 40) for CLGR StateXfer connection • 41023 + (i# * 40) for HLGR StateXfer connection • 41020 + (i# * 40) for RTR StateXfer connection	1000 ms	AF11/1

Table 53: PG Private Network Traffic Markings (Default) and Latency Requirements

Priority	Server-Side IP Address and Port	One-Way Latency Requirement	DSCP / 802.1p Marking
High	IP address: PG high-priority private IP address TCP port: • 43005 + (i# * 40) for MDS high-priority connection of PG no.1 • 45005 + (i# * 40) for MDS high-priority connection of PG no.2 UDP port: 39500 to 39999 for UDP heartbeats	40 ms	AF31/3
Medium	IP address: PG's non-high-priority private IP address TCP port: • 43016 + (i# * 40) for MDS medium-priority connection of PG no.1 • 45016 + (i# * 40) for MDS medium-priority connection of PG no.2	1000 ms	AF11/1

Priority	Server-Side IP Address and Port	One-Way Latency Requirement	DSCP / 802.1p Marking
Low	IP address: PG's non-high-priority private IP address TCP port: • 43004 + (i# * 40) for MDS low-priority connection of PG no.1 • 45004 + (i# * 40) for MDS low-priority connection of PG no.2 • 3023 + (i# * 40) for OPC StateXfer of PG no.1 • 45023 + (i# * 40) for OPC StateXfer of PG no.2	1000 ms	AF11/1

QoS Enablement in Packaged CCE

QoS is enabled by default on Private network traffic.

Disable QoS for the Public network traffic. For most deployments, disabling QoS for the Public network traffic ensures timely failover handling.

You can add QoS markings outside the contact center applications with a Windows Group Policy or by enabling marking on the IP edge routers.

For information about enabling QoS on the router during install, see the install documentation for your solution.

QoS Performance Monitoring

Once the QoS-enabled processes are up and running, the Microsoft Windows Performance Monitor (PerfMon) can be used to track the performance counters associated with the underlying links. For details on using PerfMon, see the Microsoft documentation. For more information on performance counters for QoS, see *Serviceability Guide for Cisco Unified ICM/Contact Center Enterprise* at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-installation-and-configuration-guides-list.html.

QoS for Virtualized Voice Browser

The following table outlines the default QoS for RTP and SIP for Cisco VVB. If needed, you can change the defaults as outlined."

Component	DSCP	Port
Cisco VVB RTP	CS0 (Default) • Use Platform CLI for setting Expedited Forwarding (EF) • Set **set dscp marking ipvms EF** and **set dscp enable ipvms** to enable DSCP for RTP	RTP 24576-32767
Cisco VVB SIP	CS0 (Default) • Use Platform CLI for setting to *CS3* • Set **set dscp marking UnifiedSIPSSTCP CS3** and **set dscp enable UnifiedSIPSSTCP** to enable DSCP for SIP over TCP • Set **set dscp marking UnifiedSIPSSUDP CS3** and **set dscp enable UnifiedSIPSSUDP** to enable DSCP for SIP over UDP	TCP/UDP 5060
Cisco VVB TCP/UDP to servers such as VXML server, Call server, Media server, ASR, and TTS	CS0 (Default) • Use the following Platform CLI commands to set the DSCP value to CS3 for outgoing TCP connections on ephemeral ports: **set dscp marking tcp_ephemeral CS3** **set dscp enable tcp_ephemeral** • Use the following Platform CLI commands to set the DSCP value to CS3 for outgoing UDP connections on ephemeral ports: **set dscp marking udp_ephemeral CS3** **set dscp enable udp_ephemeral**	TCP/UDP 32768-61000

Bandwidth, Latency, and QoS for Unified CM

Bandwidth for Agent Phones to Unified CM Cluster

The required bandwidth for phone-to-Unified CM signaling is 150 bps for each phone.

For example, in a 1000 agent solution, each contact center site requires approximately 150 kbps.

Bandwidth, Latency, and QoS for Cisco Finesse

For Cisco Finesse, the largest bandwidth usage is during the agent or supervisor sign-in. This operation includes the web page load, the CTI sign-in, and the display of the initial agent state. After the desktop web page loads, the required bandwidth is less.

Supervisor desktops use more bandwidth at sign-in because of its additional gadgets. We do not mandate a minimum bandwidth for the sign-in operations. Determine the required bandwidth for your solution based on how long you want the sign-in to take. Cisco Finesse has a bandwidth calculator (http://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-technical-reference-list.html) to estimate the required bandwidth for a specified client sign-in time.

During failover, agents are redirected to the alternate Cisco Finesse server and are signed in automatically, and desktop is reloaded. Expected bandwidth utilization reaches up to approximately 250 Mbps for 90 seconds (peak), to ensure all 2000 agents failover successfully from one side to another. The bandwidth requirements increase depending on the type and number of gadgets configured for teams.

For more information, see *Finesse Bandwidth Calculator for Unified Contact Center Enterprise* at https://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-technical-reference-list.html.

Note The Cisco Finesse bandwidth calculator does not include the bandwidth required for any third-party gadgets in the Cisco Finesse container. It also does not consider any other applications running on the agent desktop client that might compete for bandwidth.

Because Cisco Finesse is a web application, caching can significantly affect the required bandwidth. After the initial agent sign-in, caching significantly reduces the bandwidth required for any subsequent sign-ins. To minimize the required bandwidth for sign-in, enable caching in the browser.

After sign-in is complete, the most intensive operation for both an agent and a supervisor is making an outbound call to a route point. For the supervisor, updates to the **Team Performance** and **Queue Statistics** gadgets may be occurring concurrently. You can use the Cisco Finesse bandwidth calculator to calculate the total bandwidth required for connections between all Cisco Finesse clients and the Cisco Finesse server.

Ensure that your solution has the required bandwidth available after accounting for other applications' needs, including any voice traffic that shares this bandwidth. The performance of the Cisco Finesse interface, and potentially the audio quality of calls, can degrade if sufficient bandwidth is not continuously available.

Cisco Finesse Desktop Latency

You can locate Agent and Supervisor desktops remotely from the Agent PG. In a poorly designed deployment, high time-out values can cause an extreme delay between the desktop server and desktop clients. Large latency affects the user experience and lead to confusing or unacceptable results for the agent. For example, the phone can start ringing before the desktop updates. Limit the latency between the server and agent desktop to 400-ms round-trip time for Cisco Finesse.

Cisco Finesse also requires that you limit latency between the Cisco Finesse server and the PG to 200-ms round-trip time. Limit latency between Cisco Finesse servers to 80-ms round-trip time.

QoS for Cisco Finesse

Cisco Finesse does not support configuration of QoS settings in network traffic. Generally, have the QoS classification and marking of traffic done at the switch or router level. You can prioritize signaling traffic there, especially for agents who are across a WAN.

Bandwidth and Latency Considerations for Cisco IM&P

Cisco IM&P service is closely integrated with Unified CM and it depends on Unified CM for user management and service enabling and authentication.

Cisco IM&P can be deployed as a cluster to guarantee availability and the users must be pre-configured to specific node pairs within the cluster. Details of Cisco IM&P installation and cluster deployment can be found here https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-installation-guides-list.html.

For more details on the latency requirements for IM&P server refer, Unified CM SRND at https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

The Desktop Chat feature using Cisco IM&P requires higher client bandwidth. See the Finesse Bandwidth calculator at: https://www.cisco.com/c/en/us/support/customer-collaboration/finesse/products-technical-reference-list.html

The maximum latency supported between Finesse and IM&P nodes is 200 ms.

Bandwidth, Latency, and QoS for Unified Intelligence Center

Parameters for Reporting Bandwidth

The following parameters have a combined effect on the responsiveness and performance of the Cisco Unified Intelligence Center on the desktop:

- Real-time reports—Simultaneous real-time reports run by a single user.
- Refresh rate for realtime reports—If you have a Premium license, you can change the refresh rate by editing the Report Definition. The default refresh rate for Unified Intelligence Center is 15 seconds. The deafult refresh rate for Live Data is 3 seconds.
- Cells per report—The number of columns that are retrieved and displayed in a report.
- Historical report—Number of historical reports run by a single user per hour.
- Refresh rate for historical reports—The frequency with which report data is refreshed.
- Rows per report—Total number of rows on a single report.
- Charts per dashboard—Number of charts (pie, bar, line) in use concurrently on a single dashboard.
- Gauges per dashboard—Number of gauges (speedometer) in use concurrently on a single dashboard.

Network Bandwidth Requirements

The required bandwidth varies based on the refresh frequency, the number of rows and columns in each report, and other factors. Across a WAN, Unified Intelligence Center requires a latency of 200 ms or less.

You can use the *Cisco Unified Intelligence Center Bandwidth Calculator* (http://www.cisco.com/c/en/us/support/customer-collaboration/unified-intelligence-center/products-technical-reference-list.html) to calculate the bandwidth requirements for your Unified Intelligence Center implementation.

Unified Intelligence Center Sample Bandwidth Requirement

This sample data came from a test on a LAN with a local AWDB database and a client machine to run the reports.

The load for this test used a single Unified Intelligence Center user running the following:

Two hundred Unified Intelligence Center users, each concurrently running:

- Two realtime reports with 100 rows per report, with 10 columns each.
- Two historical reports with 2000 rows, with 10 columns each.
- Two live data reports with 100 rows, with 10 columns each. (Adjust this based on the deployment type whether LD runs or not).

This table gives the observed bandwidth usage for the test:

Table 54: Observed Bandwidth Usage During Test

Connection	Bandwidth
Unified Intelligence Center <– –> AWDB	3.4 mbps
Unified Intelligence Center <– –> Browser-based Reporting Client	5.5 mbps

The required bandwidth differs based on such parameters as the number of rows in each report and the number of concurrent report implementations.

Disk Bandwidth Requirements in Virtual Environments

When Unified Intelligence Center runs in a VM on C-series servers, in addition to the CPU and memory reservations, provision the I/O subsystem for 25 KB/s. On average, Unified Intelligence Center at full load consumes 10 KB/s of this bandwidth. The peak I/O throughput requirement reaches 25 KB/s.

Bandwidth, Latency, and QoS for Cisco Live Data

Bandwidth Considerations for Live Data

The amount of traffic, and therefore, the bandwidth usage between Central Controllers, PGs and Live Data are largely based on the call load at a site.

The bandwidth usage between Live Data and the desktop clients depends on the call rate and the number of active subscriptions to the reports. The number of active subscriptions is based on the following:

- The number of Live Data reports that are being viewed on CUIC.
- The number of agents that are signed in.
- The number of skill groups and PQs that each agent is a member of.

Bandwidth Considerations for Cisco IdS

The amount of traffic, and therefore, the bandwidth usage between Cisco IdS and any of the following components depends only on the number of signed-in agents:

- Cisco Finesse
- Cisco Unified Intelligence Center

The Finesse Bandwidth calculators and the Unified Intelligence Center Bandwidth calculators factor in the marginal increase in API calls when the agent shifts begin.

For more details on the bandwidth, latency, and QoS considerations of Finesse and Unified Intelligence Center, see Bandwidth, Latency, and QoS for Cisco Finesse , on page 366 and Bandwidth, Latency, and QoS for Unified Intelligence Center, on page 368.

Bandwidth, Latency, and QoS for Optional Cisco Components

Bandwidth, Latency, and QoS for Enterprise Chat and Email

The minimum required bandwidth for an agent sign-in to the Enterprise Chat and Email servers is 384 kbs. After sign-in, the required bandwidth is 40 kbs or more.

A 5050-KB attachment is supported within this required bandwidth. While downloading larger attachments, you can experience a temporary slow down in the agent user interface.

For more information on the bandwidth, latency, and QoS requirements for Enterprise Chat and Email, see *Enterprise Chat and Email Design Guide* at http://www.cisco.com/c/en/us/support/customer-collaboration/cisco-enterprise-chat-email/products-implementation-design-guides-list.html.

Bandwidth, Latency, and QoS for Silent Monitoring

Bandwidth, Latency, and QoS for Unified CM-Based Silent Monitoring

With Silent Monitoring, supervisors can listen to the agent calls in CCE call centers. Voice packets sent to and received by the monitored agent's IP hardware phone are captured from the network and sent to the supervisor desktop. At the supervisor desktop, these voice packets are decoded and played on the supervisor's system sound card. Silent Monitoring of an agent consumes approximately the same network bandwidth as an extra voice call. If a single agent requires bandwidth for one voice call, then the same agent being silently monitored requires bandwidth for two concurrent voice calls. To calculate the total bandwidth required for your call load, multiply the number of calls by the per-call bandwidth for your codec and network protocol.

Bandwidth, Latency Consideration for Customer Journey Analyzer

Cloud Connect component connects to **Customer Journey Analyzer** and **Unified CCE Historical** database. For quick reading of the data from Unified CCE-HDS database, the Cloud Connect should be installed in same network as HDS.

The network bandwidth requirement between the Cloud Connect and the Customer Journey Analyzer depends on the call volume of your deployment. For each call, 4000 bytes of data is transmitted to the Analyzer.

For 2000 agent deployment model, the bandwidth is as shown :

Calls per Second	Bytes for Call	Bandwidth in KBps
15	4000	60000

To account for transfer and conference call scenarios, consider those calls additional call legs. For example a transfer call or a conferenc call should be considered as two calls for bandwidth calculation.

The latency requirement between Cloud Connect and Analyzer should not exceed 400 ms round trip time.

Bandwidth and Latency Considerations for Cisco Answers

Agent Answers requires CUBE to fork the media streams to the CCAI WebSocket Connector service using the WebSocket protocol. Currently, only the vCUBE supports the WebSocket protocol. For more details on vCUBE, see Virtual CUBE for Contact Center Solutions, on page 33.

CUBE encodes the media streams using the g711 u-law codec and forks the media of both the customer and the agent call legs towards the CCAI WebSocket Connector service. The g711 u-law encoded media streams and the WebSocket protocol overheads require 183 Kbps of bandwidth per call.

Note If you are using G.729 codec for agents, switch to G.711 u-law codec to use Agent Answers and allocate additional bandwidth over the WAN links.

The actual number of calls for which media is forked for Agent Answers and the bandwidth requirements for the media forking traffic depends on the Agent Answers configuration of the corresponding CallTypes and agents. For details on how to configure Agent Answers, see the Agent Answers chapter in the *Cisco Packaged Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html.

vCUBE can combine call media traffic on a single WebSocket connection. This setting is called the Call Threshold parameter and is set to three by default, implying that the media streams corresponding to three concurrent calls, six streams in total, can be multiplexed over a single connection. The reliability of such multiplexing depends on the network latency between vCUBE and the CCAI WebSocket Connector service. For maximum reliability, round-trip latency under 50 msec is recommended. If this latency value is higher, for example, around 100 msec, it's recommended that you set the Call Threshold parameter in vCUBE to 1, that is, only the media streams for a single call are sent over each connection. If the latency exceeds 150 msec, significant disruptions may occur for the Agent Answers and Transcript features on the Agent Desktop.

Note The default call threshold on CUBE is 3. If your data centers are distributed across various geographic regions, set the Call Threshold in vCUBE to 1 for better results. The call threshold value can be modified by running the "connection call-threshold" CLI command under the stream service media profile. Setting the Call threshold to 1 on vCUBE helps avoid delays in the answers feature, but there may be a few forking failures if the round trip time is more.

Bandwidth, Latency, and QoS for Optional Third-Party Components

Bandwidth, Latency, and QoS for ASR/TTS

Automatic Speech Recognition (ASR) or Text-to-Speech (TTS) Server cannot use silence suppression and must use the G.711 codec.

ASR and TTS in WAN Configurations

ASR or TTS is bandwidth-intensive. ASR or TTS RTP and MRCP traffic is not tagged with QoS DSCP markings. Use access control lists (ACLs) to classify and re-mark the traffic at the remote site and central site.

Note Cisco does not test or qualify speech applications in a WAN environment. For guidelines on design, support over WAN, and associated caveats, see the vendor-specific documentation.

The Cisco Technical Assistance Center provides limited support (as in the case of any third-party interoperability-certified products) on issues related to speech applications.

Classifying RTP Media Traffic Between Voice Browsers and ASR or TTS Servers

The Voice Browser uses the Cisco IOS RTP UDP port range of 16384 to 32767. However, the RTP UDP port range for ASR or TTS servers can vary between operating systems and vendors. You can construct an ACL to match the traffic from the ASR or TTS server based on the Voice Browser UDP port range. However, if possible, use ASR ports or TTS Server as well. Mark the RTP traffic with DSCP EF so that it is placed in the priority queue with other voice traffic.

Configure the QoS priority queue to support the maximum number of anticipated ASR or TTS sessions. Keep the QoS priority queue bandwidth separate from any bandwidth for a call admission control method, such as Unified CM locations or Resource Reservation Protocol (RSVP). To support two ASR or TTS G.711 sessions (80 kbps each) and four IP phone calls using G.729 (24 kbps each), the priority queue bandwidth is 256 kbps. Limit the locations call admission control or RSVP bandwidth to the IP telephony bandwidth (96 kbps in this example) only. If you configure that bandwidth across the entire 256 kbps, IP calls can use all of the bandwidth and conflict with the ASR or TTS sessions.

Classifying MRCP Traffic Between Voice Browsers and ASR or TTS Servers

The MRCP traffic is easy to classify. ASR or TTS Servers listen on a TCP port that can be configured based on the vendor for MRCP requests. So, use this port in ACLs to classify the traffic. The bandwidth for MRCP can vary depending on the frequency of the application using the ASR or TTS resource. MRCP uses about 2000 bytes per interaction. If there is an ASR or TTS interaction every 3 seconds per call, you can calculate the average bandwidth as follows:

```
(2000 bytes/interaction) * (20 interactions/minute) * (8 bits/byte) = 320,000 bits per
minute per call
(320,000 bits per minute) / (60 seconds/minute) = 5.3 average kbps per branch
```

If you configure a maximum of 6 ASR or TTS sessions at any given time, then you use 32 average kbps per branch.

Limiting the Maximum Number of ASR or TTS-Enabled Calls

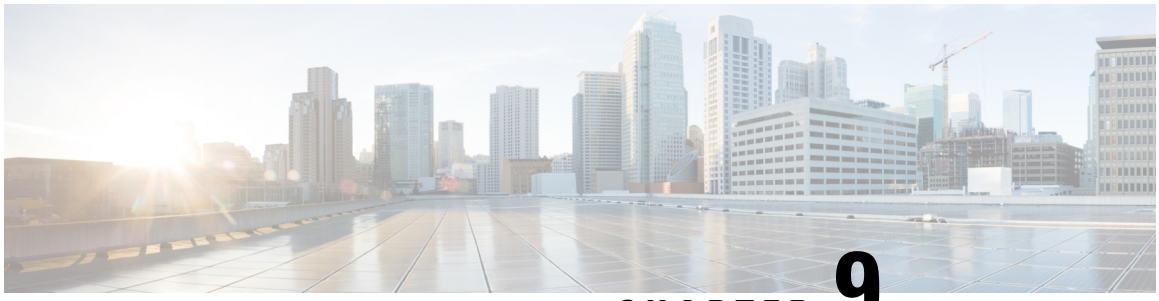
Limit the number of calls enabled for ASR or TTS. When the limit is reached, use regular DTMF prompt-and-collect instead of rejecting the call altogether. In the following example, assume 5559000 is the ASR or TTS DNIS and 5559001 is the DTMF DNIS. You can configure the Ingress Gateway to do the ASR load limiting for you. Change the DNIS when you exceed maximum connections allowed on the ASR or TTS VoIP dial peer.

```
voice translation-rule 3 rule 3 /5559000/ /5559001/
!
voice translation-profile change
```

```
  translate called 3
!
!Primary dial-peer is ASR or TTS enabled DNIS in ICM script
dial-peer voice 9000 voip
  max-conn 6
  preference 1
  destination-pattern 55590..
  ...
!
!As soon as 'max-conn' is exceeded, next preferred dial-peer will change
the DNIS to a DTMF prompt & collect ICM script
dial-peer voice 9001 voip
  translation-profile outgoing change
  preference 2
  destination-pattern 55590..
  ...
!
```

Note 80 kbps is the rate for G.711 full-duplex with no Voice activity detection, including IP/RTP headers and no compression. The rate for G.729 full-duplex with no VAD is 24 kbps, including IP/RTP headers and no compression. For information on VoIP bandwidth usage, see Voice Codec Bandwidth Calculator

Note Because Cisco VVB does not have a dial-peer to the ASR, you cannot use this technique with Cisco VVB.



CHAPTER 9

Sizing and Operating Conditions for Reference Designs

Sizing for Reference Design Solutions

A contact center enterprise solution requires proper sizing of its resources. This chapter discusses the tools and methods for sizing those resources. This includes resources like:

- The required number of contact center agents (based on customer requirements such as call volume and service level desired)
- The number of VRU ports required for various call scenarios (such as call treatment, prompt and collect, queuing, and self-service applications)
- The number of Voice Gateway ports required to carry the incoming and outbound traffic volume

Proper sizing uses the traffic engineering principles encapsulated in the Erlang-B and Erlang-C models.

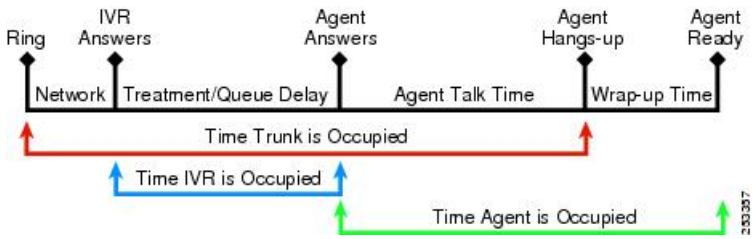
Resource Use During a Contact

When you size your contact center enterprise solution, the main resources that you look at are:

- Agents
- Gateway ports (PSTN trunks)
- VRU ports

To determine required resources, first look at this timeline of a typical inbound call and the resources that it requires at each step. This figure shows the main resources and the occupancy (hold and handle time) for those resources.

Figure 124: Inbound Call Timeline



If calls are not answered immediately, include ring delay time (network ring) in your call timeline. An average ring delay time is a few seconds. Add it to the trunk average handle time for your calculations.

Contact Center Traffic Terminology

These are the most common industry terms for sizing contact center resources.

Busy Hour or Busy Interval

A busy interval is 1 hour or less. The busy interval is when the most traffic occurs during a period of the day. The busy hour or interval varies due to circumstances like weekends and seasonal effects. Design for the average busy hour (the average of the 10 busiest hours in one year). This average is not always applied, however, when staffing is required to accommodate a marketing campaign or a seasonal busy hour such as an annual holiday peak. In a contact center, you staff for the maximum number of agents based on peak periods. But, you calculate the requirements for the rest of the day separately for each period (usually every hour). This gives proper scheduling of agents to answer calls versus scheduling agents for offline activities such as training or coaching. For trunks or VRU ports, it is not practical to add or remove trunks or ports daily, so these resources are sized for the peak periods. In some retail environments, extra trunks can be added during the peak season and disconnected afterwards.

Busy Hour Call Attempts (BHCA)

The BHCA is the total number of calls during the peak traffic hour (or interval) that are attempted or received in the contact center. For the sake of simplicity, we assume that the contact center resources (agents and VRU ports) receive and service all calls offered to the Voice Gateway. Calls usually originate from the PSTN, although calls to a contact center can also be generated internally, such as by a help-desk application.

Calls Per Second as reported by Call Router (CPS)

This is the rate at which the Unified CCE Router receives call routing requests. Every call generates one call routing request in a simple call flow from ingress gateway to VRU treatment to routing to an agent. However, some calls need more than one routing request to be made to the Router to finally get to the right agent.

An example of this is when the first agent who receives the call wants to transfer or conference to another agent by using a post route. This generates an extra routing request resulting in the same call generating two routing requests to the Router. A routing request is made to the Router whenever a resource is required for a call or task. These requests also include multimedia requests for Email, Chat, Callback and certain Outbound Calls. Call center administrators must account for these additional call routing requests when they size their contact center.

The maximum supported call rate is the call rate reported by the Router and not the BHCA at the ingress gateway. Factor these additional routing requests into the calculation of BHCA at the ingress gateway.

In general, the BHCA at the ingress gateway is lower than or equal to the corresponding CPS rate reported by the Router.

For example, consider the following situation. If the BHCA at the ingress gateway is 36,000, then the call rate at the ingress gateway is 10 CPS. If we assume that 10% of the calls are transferred through the Router, the CPS reported by Router is equal to 11 CPS. In this case, your solution needs a capacity of 11 CPS.

Servers

Servers are resources that handle traffic loads or calls. There are many types of servers in a contact center. Each type can require different resources.

Talk Time

Talk time is the amount of time an agent spends talking to a caller. This includes any time an agent places a caller on hold and any time spent during consultative conferences.

Wrap-Up Time (After-Call Work Time)

After the call terminates (the caller disconnects the call), and agent completes certain tasks to "wrap up" the call. The wrap-up time includes such tasks as updating a database, recording notes from the call, or any other activity performed until an agent becomes available to answer another call. Unified Contact Center Enterprise solutions sometimes call this period the *after-call work time*.

Average Handle Time (AHT)

AHT is the mean (or average) call duration during a specified time period. It refers to the sum of several types of handling time, such as call treatment time for self-service calls or talk time for calls to agents. In its most common definition, AHT is the sum of agent talk time and agent wrap-up time.

Erlang

Erlang is a measurement of traffic load during the busy hour. The Erlang is based on having 3600 seconds of calls on the same circuit, trunk, or port. (One circuit is busy for 1 hour regardless of the number of calls or how long the average call lasts.) The formula to calculate the Erlang value is:

```
Traffic in Erlangs = (Number of calls in the busy hour * AHT in sec) / 3600 sec
```

If a contact center receives 30 calls of 6-minute length in the busy hour, this equates to 180 minutes of traffic in the busy hour, or 3 Erlangs. If the contact center receives 100 calls averaging 36 seconds each in the busy hour, then total traffic received is 3600 seconds, or 1 Erlang (3600 sec/3600 sec).

Busy Hour Traffic (BHT) in Erlangs

BHT is the traffic load during the busy hour and is calculated as the product of the BHCA and the AHT normalized to 1 hour:

```
BHT = (BHCA * AHT seconds) / 3600
```

For example, if the contact center receives 600 calls in the busy hour, averaging 2 minutes each, then the busy hour traffic load is (600 * 2/60) = 20 Erlangs.

BHT is typically used in Erlang-B models to calculate resources such as PSTN trunks or self-service VRU ports.

Grade of Service (Percent Blockage)

This measurement is the probability that a resource or server is busy during the busy hour. In that case, the call is lost or blocked. This blockage typically applies to resources such as Voice Gateway ports, VRU ports, PBX lines, and trunks. In the case of a Voice Gateway, grade of service is the percentage of calls that are blocked or that receive busy tone (no trunks available) out of the total BHCA. For example,

a grade of service of 0.01 means that 1% of calls in the busy hour is blocked. A 1% blockage is a typical value to use for PSTN trunks, but different applications might require different grades of service.

Blocked Calls

A blocked call is a call that is not serviced immediately. Caller are blocked if they are rerouted to another route or trunk group, if they are delayed and put in a queue, or if they hear a tone (such as a busy tone) or announcement. The nature of the blocked call determines the model used for sizing the particular resources.

Service Level

The industry standard term for the percentage of the offered call volume (received from the Voice Gateway and other sources) that are answered within *X* seconds. A typical value for a sales contact center is 90% of all calls answered in less than 10 seconds (some calls are delayed in a queue). A support-oriented contact center might have a different service level goal, such as 80% of all calls answered within 30 seconds in the busy hour. Your service level goal determines the necessary agents, the percentage of queued calls, the average time calls spend in queue, and the necessary PSTN trunks and VRU ports.

Queuing

When agents are busy with other callers or are unavailable (after call wrap-up mode), subsequent callers must be placed in a queue until an agent becomes available. Your desired service level and agent staffing determines the percentage of calls queued and the average time spent in the queue. Contact center enterprise solutions use a VRU to place callers in queue and play announcements. The VRU initially handles all calls. It supplies call treatment and prompts for necessary information. The VRU handles self-service applications where the caller is serviced without needing to talk to an agent. Each of these scenarios requires a different number of VRU ports because each has a different average handle time and possibly a different call load. The number of trunks or gateway ports needed for each of these applications differs accordingly.

Erlang Calculators as Design Tools

Many traffic models are available for sizing telephony systems and resources. Choosing the right model depends on three main factors:

- Traffic source characteristics (finite or infinite)
- How lost calls are handled (cleared, held, delayed)
- Call arrival patterns (random, smooth, peaked)

For contact center enterprise solutions, you commonly use the Erlang-B and Erlang-C traffic models for sizing resources.

Erlang calculators help answer the following questions:

- How many trunks do I need?
- How many agents do I need?
- How many VRU ports do I need?

You need these figures for input to Erlang calculators:

- The busy hour call attempts (BHCA)
- Average handle time (AHT) for each of the resources

- Service level (percentage of calls that are answered within x seconds)
- Grade of service, or percent blockage, desired for trunks and VRU ports

The next sections present a brief description of the generic Erlang models in simple terms. They also describe the input and output of the Erlang models and which model to use for sizing particular resources. There are a variety of contact center sizing tools available. They all use the two basic traffic models, Erlang-B and Erlang-C.

Erlang-B Uses

Use the Erlang-B model to size PSTN trunks, gateway ports, or VRU ports. It assumes the following:

- Call arrival is random.
- If all trunks or ports are occupied, new calls are lost or blocked (receive busy tone) and not queued.

The input and output for the Erlang B model consists of the following three factors. If you have any two of these factors, the model calculates the third:

- Busy Hour Traffic (BHT). BHT is the product of the number of calls in the busy hour (BHCA) and the average handle time (AHT).
- Grade of Service
- Ports (lines)

Erlang-C Uses

Use the Erlang-C model to size agents in contact centers that queue calls before presenting them to agents. This model assumes:

- Call arrival is random.
- If all agents are busy, incoming calls are queued and not blocked.

The input parameters required for this model are:

- The number of calls that angents answer in the busy hour (BHCA)
- The average talk time and wrap-up time
- The delay or service level desired, expressed as the percentage of calls answered within a specified number of seconds

The output of this model gives the required number of agents, the percentage of calls delayed when agents are unavailable, and the average queue time.

Dynamic Configuration Limits for Unified CCE

Sometimes, you can exceed the standard limits for one resource by significantly reducing use of another resource. Test the specific trade-off that you plan to make before you incorporate it in your solution. The following sections provide guidance on how to balance certain resources.

Dynamic Limits for Skill Groups and Precision Queues Per Agent

The number of skill groups and precision queues per agent significantly affects the following subcomponents of Unified CCE:

- Cisco Finesse servers
- Agent PGs
- Router
- Logger

Note We use *queue* as a common term for skill groups and precision queues.

To maintain the performance of your solution, periodically remove unused queues.

The Reference Designs set a standard limit for the average queues per agent on each PG. On a particular PG, some agents can have more queues than other agents. As long as the average across all the agents on the PG is within the limit, you can still have the maximum active agents on that PG.

For example, assume that you have three groups of agents on a PG in a 4000 Agent Reference Design:

- Group A has 500 agents with five queues each.
- Group B has 1000 agents with 15 queues each.
- Group C has 500 agents with 25 queues each.

These three groups average to 15 queues per agent, so you can have them all on a single PG under the standard limits.

You can also exceed that standard limit if you reduce the number of agents on each PG and on the whole system.

Note See the configuration tables in the configuration limits chapter for the standard limits.

The Cisco Finesse server doesn't display statistics for unused queues. So, the active queues affect the performance of the Cisco Finesse server more than the total configured queues.

The Cisco Finesse desktop updates queue (skill group) statistics at 10-second intervals. The Cisco Finesse Desktop also supports a fixed number of queue statistics fields. You can't change these fields.

This table shows the approximate reduction in the number of agents your solution can support with more queues per agent:

Table 55: Dynamic Agents and Queues Limits

Average Queues per Agent	Maximum Agents per PG	Maximum Agents for 2000 Agent Reference Design	Maximum Agents for 4000 Agent Reference Design[58]	Maximum Agents for 12000 Agent Reference Design
15	2000	2000	4000	12000

Average Queues per Agent	Maximum Agents per PG	Maximum Agents for 2000 Agent Reference Design	Maximum Agents for 4000 Agent Reference Design[58]	Maximum Agents for 12000 Agent Reference Design
20	1500	1500	3000	9000
30	1000	1000	2000	6000
40	750	750	1500	4500
50	600	600	1200	3600

[58] You can't have more than 4000 Agents on a Rogger deployment.

Unified CCE supports a maximum of 50 unique skill groups across all agents on a supervisor's team, including the supervisor's own skill groups. If this number is exceeded, all skill groups that are monitored by the supervisor still appear on the supervisor desktop. However, exceeding this number can cause performance issues and isn't supported.

Note Each precision queue that you configure creates a skill group for each Agent PG and counts toward the supported number of skill groups per PG. The skill groups are created in the same Media Routing Domain as the precision queue.

Other Dynamic Sizing Factors

Many factors can affect your solution's server requirements and capacities. These sections call out the major sizing variables and how they affect your solution.

Busy Hour Call Attempts (BHCA)

As BHCA increases, the load on all your solution components increases, most notably on Unified CM, Unified CVP, and the Agent PG. The capacity numbers for agents assume up to 30 calls per hour per agent. If your solution requires a greater BHCA, decrease the maximum agents on the Agent PG.

Unified CM Silent Monitor

Each silently monitored call adds more processing for the PG and Unified CM. Each silently monitored call equals to two unmonitored agent calls. Leave room on your PGs to account for the percentage of the monitored calls.

Script Complexity

As the complexity and number of Unified CCE scripts increase, the processor and memory overhead on the Call Router and VRU PG increases significantly. As VRU script complexity increases with features such as database queries, the load placed on CVP and the Router also increases. The delay time between replaying RunExternalScript also has an impact.

The performance of complex scripts and database queries is hard to characterize. Test complex scripting in a lab to determine the response time of database queries under various BHCA. Adjust your sizing to account for their effects on the processor and memory of the Voice Browser, Unified CVP, the PGs, and the Router.

Third-Party Database and Cisco Resource Manager Connectivity

Carefully examine the connectivity of any Unified CCE solution component to external devices and software to determine the overall effect on the solution. Contact center enterprise solutions are flexible and customizable, but they are also complex. Contact centers are often mission-critical, revenue-generating, and customer-facing operations. Engage a Cisco Partner (or Cisco Advanced Services) with the appropriate experience and certifications to help you design your solution.

Expanded Call Context (ECC)

Your solution's use of ECC variables impacts the PGs, Router, Logger, and network bandwidth. You can configure and use ECC variables in many ways. The capacity impact varies based on the ECC configuration.

PG Agent Capacity with Mobile Agents

Mobile agent support capacity on the medium PG OVA are as follows:

- 2000 with nailed-up connections (1:1)
- 1500 with nailed-up connections if the average handle time is less than 3 minutes, or if agent greeting or whisper announcement features are used with the mobile agent (1.3:1)
- 1500 with call-by-call connections (1.3:1)

You can have a mix of mobile Agents and other agents on the same PG. Keep the respective weights of each type of agent in mind. For example, if you have 200 mobile agents with nailed-up connections, the PG can support 1800 other agents:

```
Additional Agents Allowed = (2000 – (200 * 1)) = 1800 Agents
```

If you plan to use 200 active mobile agents with call-by-call connections, the PG can support 1740 other agents:

```
Additional Agents Allowed = (2000 – (200 * 1.3)) = 1740 Agents
```

Configuration Limits for Reference Design Solutions

Sizing for Unified CVP

When you size your contact center, determine the worst-case profile for the number of calls that are in each state. At its busiest instant in the busiest hour, how many calls do you find in the following states:

- **Self-service**—Calls that are executing applications using the VXML Server.
- **Queue and collect**—Calls that are in queue for an agent or are executing prompt-and-collect self-service applications.
- **Talking**—Calls that are connected to agents or to third-party TDM VRU applications.

Note The definitions of these call states differ from the definitions used for port licensing purposes. You can ignore ASR and TTS processing when counting which calls are in which states for sizing purposes. However, ASR and TTS processing does come into call state counts for licensing.

Size the solution for the number of ports in use for calls in a talking state to agents. Even though you do not need licenses for those ports when using Unified CCE agents, TDM agents do require a Call Director license.

For calls in the talking state, count only calls that use Unified CVP or gateway resources. If the transfer uses VoIP, the call uses a Voice Browser port and Unified CVP resources. Unified CVP continues to monitor the call and enables you to retrieve it and redeliver it later. Unified CVP also continues to monitor calls to a TDM target. Those calls use both an incoming and an outgoing TDM port on the same gateway or on a different gateway (that is, toll bypass). Both of these types of calls count as talking calls.

However, if a transfer uses *8 TNT, hookflash, Two B Channel Transfer (TBCT), or an ICM NIC, the gateway and Unified CVP do not play a role. Both components reclaim their resources. Such calls do not count as talking calls.

Include in the overall call counts those calls that are transferred back to Unified CVP for queuing or self-service. For example, in a warm transfer, Unified CVP queues the agent during the post-route phase. The call uses two ports for the two separate call control sessions at Unified CVP. Transfers are usually a small part of the overall call volume, and you can easily overlook them.

In addition to the overall snapshot profile, also consider the CPS for the busiest period of call arrival. You need this information for the contact center enterprise solution because it is difficult to identify the exact maximum arrival rate. You can use statistical means to arrive at this number.

You size Unified CVP Servers for the number of handled calls and the maximum call arrival rate.

Table 56: CVP Server Call Rate for Call Flows

Call flow	Simultaneous Calls Supported	Calls Per Second
Comprehensive Call Flow with Secure / Non Secure SIP	3000	15
Comprehensive Call Flow with Secure / Non Secure SIP, Secure HTTP	2500	15
Standalone with / without Request ICM Label	3000	15
Standalone with / without Request ICM Label, Secure HTTP	2000	15

Note

- For the **Call Per Second**, this is the maximum call rate that is received at the CVP Call Server from all Ingress Gateways in the solution and assumes the worst case scenario where WAAG is enabled.
- It is always recommended to have enough VM resources for the garbage collection on CVP Servers to run adequately (VM memory usage should not go beyond 80%). If there are not enough system resources, garbage collection may take more time which can cause issues with the Call Server or VXML Server services.

CVP Call Server Sizing

The solution needs the greater number of Call Servers given by these equations:

```
(Self Service) + (Queue and Collect) + (Talking) / Simultaneous Calls Supported, rounded
up
OR
(Average call arrival rate) / Calls Per Second, rounded up.
```

Also, distribute the calls to the Unified CM cluster among the subscribers in the cluster. Do not exceed 2 CPS per subscriber.

See the table *CVP Server Call Rate for Call Flows* in section *Sizing for Unified CVP* for more details.

Log Directory Size Estimate

Use the following formula to calculate the estimated space per day (in gigabytes) for the Call Server Directory log file:

```
3.5 GB * R
```

Where *R* equals the number of calls per second.

For proper serviceability, reserve enough space to retain five to seven days of log messages.

CVP VXML Server Sizing

One VXML Server can handle calls as mentioned in the table "*CVP Server Call Rate for Call Flows* in the section "*Sizing for Unified CVP*". If you are using VXML Servers, size them according to the following formula:

```
Calls / Simultaneous Calls Supported, rounded up
```

Calls are the number of calls that are in VXML Server self-service applications at that snapshot in time.

Note For UCS performance numbers, see the *Virtualization for Cisco Unified Customer Voice Portal* page.

With an appropriate Cisco IOS release, you can configure Unified CVP to use HTTPS on the VXML Server and on the Unified CVP IVR Service.

Note Mainline Cisco IOS is not supported.

Performance of the CVP VXML Server varies with the complexity of your VXML application.

CUSP Performance Benchmarks

When your solution uses CUSP, remember the following points:

- CUSP baseline tests were done in isolation on the proxy. The capacity numbers from those tests are 450 TCP or 500 UDP transactions/second. Consider these figures to be the most stressed condition allowable.
- A CVP call from the proxy server requires, on average, four separate SIP calls: a caller inbound leg, a VXML outbound leg, a ringtone outbound leg, and an agent outbound leg.
- When a consultation with CVP queuing occurs, the session incurs four more SIP transactions, effectively doubling the number of calls.

Sizing Gateways for Contact Center Enterprise Solutions

Call capacities on Cisco gateways vary depending on whether they are doing ingress only, VXML only, or a combination of the two. Capacities on Voice Browsers also vary depending on factors like ASR/TTS services and type of VXML application. For instance, an intensive JavaScript application reduces call capacity.

In general, you can size gateways that perform ingress only to the maximum number of TDM cable attachment points.

Before sizing the voice gateways, use the Unified CCE Resource Calculator to determine the maximum number of trunks (DS0s) and VXML VRU ports to support your solution.

The following table provides sizing information for different versions of Cisco IOS. The sizing information is based on these factors:

- The overall CPU usage does not exceed 75 percent.
- Sizing represents the maximum number of concurrent VXML sessions and VoIP calls on the gateway.
- Sizing is based on Unified CVP VXML documents.
- Sizing includes active conferences and active transfers.
- For the VXML Only columns, sizing includes only basic routing and IP connectivity running on the gateway. If you intend to run extra applications such as fax or other noncontact center traffic, account for that traffic in your deployment's capacity. For the VXML + PTSN columns, the indicated number of VXML sessions and voice calls are supported simultaneously on the same gateway.
- Sizing is based on using either the Cisco Call Server or Cisco Unified CVP VXML Server.
- Each gateway is configured to share the load with its redundant pair during usual operations. Under usual operations, each gateway handles the load close to half of its capacity. During a failover scenario, each gateway operates with its maximum supported load.
- Each port provides TDM and VXML functionality including ASR/TTS.

Table 57: Maximum Number of VXML Sessions Supported by Cisco Voice Gateways (Cisco IOS Release 15.1.4.M7 and Later)

VXML Gateway CPU Capacity for Cisco IOS Release 15.1.4.M7 or Later					
Platform	**VXML Only**		**VXML + PSTN**		**Memory**
	DTMF	**ASR**	**DTMF**	**ASR**	**Recommended**
2901	12	8	9	6	2 GB
2911	60	40	47	31	2 GB
2921	90	60	71	48	2 GB
2951	120	80	95	64	2 GB
3925	240	160	190	127	2 GB
3945	340	228	270	180	2 GB
3925E	475	450	380	375	2 GB

VXML Gateway CPU Capacity for Cisco IOS Release 15.1.4.M7 or Later					
Platform	**VXML Only**		**VXML + PSTN**		**Memory**
	DTMF	**ASR**	**DTMF**	**ASR**	**Recommended**
3945E	580	550	460	450	2 GB
Based on ISO 15.1.4.M7, G.711, basic calls, Ethernet egress, CPU NTE 75%					

Note A single combination gateway cannot exceed the number of concurrent VXML sessions and VoIP calls.

Table 58: Maximum Number of VXML Sessions Supported by Cisco Voice Gateways Executing Intensive JavaScript Applications (Cisco IOS Release 15.1.4.M7 and Later)

Cisco Voice Gateway Platform	Dedicated VXML Gateway		Voice Gateway and VXML		
	VXML and DTMF	**VXML and ASR/TTS**	**VXML and DTMF**	**VXML and ASR/TTS**	**Memory Recommended**
AS5350XM	105	85	110	70	512 MB (default)
AS5400XM	105	85	110	70	512 MB (default)

Table 59: Maximum Number of VXML Sessions Supported by Cisco Voice Gateways Using HTTPS (Cisco IOS Release 15.1.4.M7 and Later)

Cisco Voice Gateway Platform	Dedicated VXML Gateway		Voice Gateway and VXML		
	VXML and DTMF	**VXML and ASR/TTS**	**VXML and DTMF**	**VXML and ASR/TTS**	**Memory Recommended**
3945E	510	342	408	270	2 GB
AS5350XM	155	120	138	95	512 MB (default)
AS5400XM	155	120	138	95	512 MB (default)

Note The performance numbers in the preceding table are only for selected models of Cisco Voice Gateways using HTTPS. Use the HTTPS performance numbers of the 3945E router, to estimate the performance numbers for router models that are not listed in Table 11.

See the section on sizing gateways for contact center traffic in *Cisco Collaboration System Solution Reference Network Designs* at https://www.cisco.com/c/en/us/support/unified-communications/

unified-communications-manager-callmanager/products-implementation-design-guides-list.html to ensure that the call arrival rates do not exceed the listed capacities.

CPU Usage

For all gateways, ensure that the overall CPU usage is less than 75 percent on average. The following factors affect CPU usage:

- Calls per second (CPS)
- Maximum concurrent calls
- Maximum concurrent VXML sessions
- Intensive JavaScript applications

Memory Considerations

Consider how much DRAM and flash memory to order. The capacity that comes with the machine by default is sufficient for most purposes. However, consider increasing the amount of DRAM in order to expand your flash memory if your application requires:

- Large numbers of distinct .wav files (as with complex self-service applications)
- Unusually large .wav files (as with extended voice messages or music files)

Note You can only extend HTTP cache to 100 MB in the current Cisco IOS releases.

Third-Party VXML Application Considerations

If you are using a non-Cisco VXML application, your deployment must adhere to the CPU usage requirements. Ensure that adequate memory is available on Cisco gateways at full load when running external VXML applications.

Contact the provider of that application for performance and availability information. Cisco makes no claims or warranties regarding the performance, stability, or feature capabilities of a third-party VXML application when interoperating in a Cisco environment.

CUBE and Virtual CUBE Considerations

For information on sizing physical or virtual Cisco UBE, see *Cisco Unified Border Element Configuration Guide* at https://www.cisco.com/c/en/us/support/unified-communications/unified-border-element/products-installation-and-configuration-guides-list.html.

For session capacity information on CUBE, see the *Cisco Unified Border Element Data Sheet* at https://www.cisco.com/c/en/us/products/unified-communications/unified-border-element/datasheet-listing.html.

Note The CVP comprehensive call flow uses more than the standard 7 messages per call leg in a VoIP call flow with Unified CM. Because of this, size CUBE sessions for contact center enterprise solutions as follows:

- ISR G2 scalability is 40% less than the standard CUBE session capacity. Scalability for the ASR1K and ISR4K series is about 75% less.
- The amount of DSPs in the platform limits CPA numbers. Treat CPA as high complexity.
- CUBE establishes a SIP session for media forking to a call recording server. If you use CUBE-controlled recording or Unified CM-controlled recording, add an extra session for each recording to your overall sizing.
- You can have a mix of CVP sessions and recording sessions on the same CUBE. Add the sessions together to properly size the CUBE. For example, if you have 1000 CVP sessions and 1000 forking sessions for call recording, then the CUBE expected load on ISR G2 is roughly:

  ```
  (1000 CVP * 1.66 for performance impact) + 1000 Recording = 2660 total sessions
  ```

  On ASR 1K/ISR 4K, the expected load is roughly:

  ```
  (1000 CVP * 4 for performance impact) + 1000 Recording = 5000 total sessions
  ```

  Use the total sessions to size against the standard SIP sessions that your CUBE model supports.

Important Correctly sizing CUBE when you activate multiple services, such as transcoder and MTP resources, on CUBE is more complex. Consult your Cisco Account team to connect with someone from the CUBE team. They can help you properly size complex CUBE deployments.

CVP Basic Video Service Sizing

You can have video-capable agents in your contact center enterprise solution.

You can use the same Unified CVP Call Server for both video calls and traditional audio calls in a comprehensive call flow.

The basic video service uses the comprehensive call flow. It requires Call Server, VXML Server, and IOS VXML Gateways. You size these components in the same way that you do for audio calls.

Cisco Unified Video conferencing hardware, Radvision IVP, and Radvision iContact are not required for the basic video service.

Note Video call is not supported in Cisco VVB.

CVP Reporting Server Sizing

Sizing the CVP Reporting Server involves many variables. Different VXML applications have different characteristics that influence the amount of reporting data. Some of these factors are:

- The types of elements in the application
- The granularity of the required data

- The call flow through the application
- The length of calls
- The number of calls

To size the CVP Reporting Server, first estimate the amount of reporting data that your VXML application generates.

Once you determine the number of reporting messages from your application, complete the following steps for each VXML application:

1. Estimate the CPS that the application receives.
2. Estimate the number of reporting messages for your application.

This equation determines the number of reporting messages that a VXML application generates each second:

*A# = %VXML * CPS * MSG*

Where:

- *A#* is the number of estimated reporting messages per second for an application.
- *CPS* is the number of calls per second.
- *%VXML* is the percentage of calls that use this VXML application.
- *MSG* is the number of reporting messages that this application generates.

Total the values for each VXML application in your solution to get the estimated reporting messages per second for your solution.

Each CVP Reporting Server can handle 420 messages per second. If your solution requires more than one CVP Reporting Server, partition the VXML applications to use specific Reporting Servers.

Solutions with Multiple CVP Reporting Servers

In solutions that require more than one CVP Reporting Servers, you partition the deployment vertically.

When vertically partitioning to load balance reporting data, consider these requirements and guidelines:

- Associate each Call Server and VXML Server with only one CVP Reporting Server.
- Reports cannot span multiple Informix databases.
- Subdivide applications that generate more combined call processing and application messages than one CVP Reporting Server can support.
- You can filter VXML. Filtering out noninteresting data creates more usable data repositories that support higher message volume.
- Configure the dial plan and other available means to direct the incoming calls to the appropriate Call Server and VXML Server.

For more information on these requirements, see the *Reporting Guide for Cisco Unified Customer Voice Portal* available at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-installation-and-configuration-guides-list.html.

To combine data from multiple databases, you can use these possible options:

- Export the reporting data to another format, like a spreadsheet, and combine the data outside of the database.
- Export the reporting data to CSV files and import it into a customer-supplied database.
- Extract the data to a customer-supplied data warehouse and run reports against that data.

Message Details on CVP Reporting Servers

This table lists the number of reporting messages that various elements or activities generate.

Table 60: Number of Reporting Messages Per Element or Activity

Element or Activity	Number of Reporting Messages (Unfiltered)
Start	2
End	2
Subflow Call	2
Subflow Start	2
Subflow Return	2
Throw	2
Alert	2
Subdialog_start	2
Subdialog_return	2
Hotlink	2
HotEvent	2
Transfer w/o Audio	2
Currency w/o Audio	2
Flag	2
Action	2
Decision	2
Application Transfer	2
VXML Error	2
CallICMInfo (per call)	2
Session Variable (per change)	2
Custom Log (per item)	2

Element or Activity	Number of Reporting Messages (Unfiltered)
Play (Audio file or TTS)	2
LeaveQueue	2
Callback_Disconnect_Caller	3
Callback_Add	4
Callback_Get_Status	4
Callback_Set_Queue_Defaults	4
Callback_Update_Status	4
Callback_Enter_Queue	5
Callback_Reconnect	5
Get Input (DTMF)	5
Callback_Validate	6
Get Input (ASR)	9
Form	10
Digit_with_confirm	20
Currency_with_confirm	20
ReqICMLabel	30

Note Every application requires these elements. You cannot filter them.

Sizing for Unified CM Clusters

Unified CM clusters provide a mechanism for distributing call processing across a converged IP network infrastructure. Clusters also facilitate redundancy and provide feature transparency and scalability.

For a more detailed view of Unified CM clusters, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.

Cluster Sizing Concepts

Before attempting to size a Unified CM cluster for your solution, perform the following design tasks:

- Determine the different types of call flows.
- Determine the required deployment model (single site, centralized, distributed, clustering over the WAN, or remote branches within centralized or distributed deployments).
- Determine the protocols to be used.

- Determine redundancy requirements.
- Determine all other customer requirements for Cisco Unified Communications that share a cluster with a Unified CCE deployment. For example, consider Cisco Unified IP Phones, applications that are not part of Unified CCE, route patterns, and so forth.

After you complete these tasks, you can begin to accurately size the necessary clusters. Many factors affect the sizing of a cluster, and the following list mentions some of those factors:

- Number of office phones and the busy hour call attempt (BHCA) rate per phone
- Number of inbound agent phones and the BHCA rate per phone
- Number of CTI ports and the BHCA rate on those VoIP endpoints. (If you use Unified CVP for call treatment, self-service, and queuing, these factors might not apply.)
- Number of Voice Gateway ports and the BHCA rate on those VoIP endpoints
- Number of outbound agent phones, outbound dialing mode, and BHCA rate per phone
- Number of outbound dialer ports, number of VRU ports for outbound campaigns, and the BHCA rate per port for both
- Number of mobile agents and the BHCA rate per mobile agent
- Number of voicemail ports and the BHCA rate to those VoIP endpoints
- Signaling protocols used by the VoIP endpoints
- Percent of agent call transfers and conferences
- Dial plan size and complexity, including the number of dialed numbers, lines, partitions, calling search spaces, locations, regions, route patterns, translations, route groups, hunt groups, pickup groups, and route lists
- Amount of media resources needed for functions such as transcoding, conferences, encryption, and so forth
- Coresident applications and services such as CTI Manager, E-911, and Music on Hold
- Unified CM release (sizing varies per release)
- Type of Unified CM OVA

Other factors can affect cluster sizing, but theses are the most significant factors in terms of resource consumption.

In general, you estimate the resource consumption (CPU, memory, and I/O) for each of these factors to size the Unified CM cluster. You then choose VMs that satisfy the resource requirements. Gather information about these factors before you can size a cluster with any accuracy.

Cluster Guidelines

The following guidelines apply to all Unified CM clusters in your solution:

- All primary and backup subscribers must use the same OVF template. All subscribers in the cluster must run the same Unified CM software release and service pack.

- Within a cluster, you can enable a maximum of eight subscribers (four primary and four backup subscribers) with the Cisco Call Manager Service. You can use more VMs for dedicated functions such as TFTP, publisher, and music on hold.
- In a 4000 Agent Reference Design, a Unified CM cluster can support about 4000 agents. In a 12,000 Agent Reference Design, a Unified CM cluster with four primary and four backup subscribers can support about 8000 agents. These limits assume that the BHCA call load and all configured devices are spread equally among the eight call processing subscribers with 1:1 redundancy. These capacities can vary, depending on your specific deployment. Size your solution with the *Cisco Unified Communications Manager Capacity Tool*.

  A subscriber can support a maximum of 1000 agents. In a fail-over scenario, the primary subscriber supports a maximum of 2000 agents.

Note In a 4000 Agent Reference Design, a cluster with four subscribers (two primary and two backup) can support the maximum load. If you create clusters with more subscribers, do not exceed the maximum of 4000 agents for the cluster.

When sizing the cluster to support contact center solutions for the appropriate number of CTI resources, remember to account for the following:

- Configured phones from agents who are not signed in
- Applications which remotely control the device like Call Recording, Attendant Console, and PC-clients
- Other 3rd-party applications which consume CTI resources

Unified CM can support multiple concurrent CTI resources, for example, when multiple lines, the contact center, and recording are used concurrently. Those CTI resources follow the same CTI rules as described in the *Cisco Collaboration System Solution Reference Network Designs*:

- Devices (including phones, music on hold, route points, gateway ports, CTI ports, JTAPI Users, and CTI Manager) must never reside or be registered on the publisher. If there are any devices registered with the publisher, any administrative work on Unified CM impacts call processing and CTI Manager activities.
- Do not use a publisher as a fail-over or backup call processing subscriber in production deployments. Any deviations require review by Cisco Bid Assurance on a case-by-case basis.
- Any deployment with more than 150 agent phones requires a minimum of two subscribers and a combined TFTP and publisher. The load-balancing option is not available when the publisher is a backup call processing subscriber.
- If you require more than one primary subscriber to support your configuration, then distribute all agents equally among the subscriber nodes. This configuration assumes that the BHCA rate is uniform across all agents.
- Similarly, distribute all gateway ports and CTI ports equally among the cluster nodes.
- Some deployments require more than one Unified CCE JTAPI user (CTI Manager) and more than one primary subscriber. In these deployments, if possible, group and configure all devices that are monitored by the same Unified CCE JTAPI User (third-party application provider), such as Unified CCE route points and agent devices, on the same VM.

- Enable CTI Manager only on call processing subscribers, thus allowing for a maximum of eight CTI Managers in a cluster. To provide maximum resilience, performance, and redundancy, load-balance CTI applications across the various CTI Managers in the cluster. For more CTI Manager considerations, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.
- If you have a mixed cluster with Unified CCE and general office IP phones, if possible, group and configure each type on a separate VM (unless you need only one subscriber). For example, all Unified CCE agents and their associated devices and resources are on one or more Unified CM servers. Then, all general office IP phones and their associated devices (such as gateway ports) are on other Unified CM servers, as long as cluster capacity allows. If you use the *Cisco Unified Communications Manager Capacity Tool*, run the tool separately with the specific device configuration for each primary Unified CM server. You need to run it multiple times because the tool assumes that all devices are equally balanced in a cluster. Remember that with Unified CCE, you must use the 1:1 redundancy scheme.
- Use hardware-based conference resources whenever possible. Hardware conference resources provide a more cost-effective solution and allow better scalability within a cluster.
- Register all CTI route points for the Unified CCE Peripheral Gateway (PG) JTAPI user with the subscriber node running the CTI Manager instance that communicates with that Unified CCE PG.
- The *Cisco Unified Communications Manager Capacity Tool* does not currently measure CTI Manager impact on each VM separately. However, the CTI Manager does place an extra burden on the subscriber running that process. The tools report the resource consumption based on these subscribers. The actual resource consumption on the other Unified CM subscribers can be slightly lower.
- Even if a contact center agent does not use them, count all devices for a Unified CCE PG JTAPI user as an agent device. The PG is still notified of all device state changes for that phone, even though an agent does not use the phone. To increase cluster scalability, if your agents do not regularly uses a device, do not associate the device with the Unified CCE PG JTAPI user.
- CPU resource consumption by Unified CM varies, depending on the trace level enabled. Changing the trace level from Default to Full on Unified CM can increase CPU consumption significantly under high loads. The Cisco Technical Assistance Center does not support changing the tracing level from Default to No tracing.
- Under usual circumstances, place all subscribers from the cluster within the same LAN or MAN. Do not place all members of a cluster on the same VLAN or switch.
- If the cluster spans an IP WAN, follow the specific guidelines in the sections on clustering over the WAN in this guide and in the *Cisco Collaboration System Solution Reference Network Designs*.

For the most current information about supported releases, see the latest version of the your solution's *Compatibility Matrix*.

For more Unified CM clustering guidelines, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.

Component and Feature Impacts on Scalability

Some optional components and features affect the scalability and capacity of your solution. This table lists some of these effects.

Component or feature	Impact
IPsec	When you enable IPsec: • The PG capacities decrease by 25% for agents, VRU ports, SIP Dialer ports, and call rate. • The maximum call rate (calls per second) decreases by 25%.
Mobile agents	Unified CCE does not directly control the phones of mobile agents. The two delivery modes, Call-by-Call and Nailed Connection, use resources differently.
Cisco Outbound Option	Your outbound resources can vary based on hit rate, abandon limit, and talk time for the campaigns. A quick, but inexact, estimate is that you require two ports for each outbound agent. While you can technically have 2000 agents per PG assigned to outbound calls, the Dialers probably cannot keep all those agents fully occupied. Use the sizing tool to determine outbound resources required for your campaigns.
Agent Greeting	The Agent Greeting feature affects the Router, Logger, and Unified CM. On the Router and Logger, this feature increases the route requests made. That effectively decreases the maximum call rate by about one third.
Extended Call Context (ECC)	Increased Extended Call Context (ECC) usage affects performance and scalability on critical components of Unified CCE. The capacity impact varies based on the ECC configuration, and requires professional guidance on a case-by-case basis.

Resource Requirements for Reporting

Do not run more than ten concurrent reports on any client machine. This is a combined limit for reports that run on the Unified Intelligence Center User Interface, Permalinks, and Dashboards on the client machine. See the *Maximum rows per report* row in the System Load Limits table for the maximum number of rows supported in reports.

Our capacity testing shows that 200 concurrent reporting users can each have the following running reports:

- Two Live Data reports with 100 rows of 10 fields.
- Two real-time reports with 100 rows of 10 fields, refreshing every 15 seconds.
- Two historical reports with 2000 rows of 10 fields, refreshing every 30 minutes.

That means 400 Realtime and 400 Historical reports can be run concurrently. These numbers include the reports run from permalinks, reports on dashboards, schedules and desktop gadgets.

For example, if you have 200 historical permalinks open and 100 supervisors are accessing one historical report each from the desktop gadget, you can run 100 more historical reports.

If you have fewer reporting users on a node, they can run proportionally more reports. But, no client machine can exceed the ten report limit.

Cisco Virtualized Voice Browser Sizing

The call capacity of Cisco VVB is based on the call support for ASR or TTS activities and on the type of VXML application. For instance, an intensive JavaScript application reduces call capacity and VVB with HTTPS has a lower call capacity than with HTTP.

Ensure that the average overall CPU usage is less than 65 percent. The following factors affect CPU usage:

- Calls per second (CPS)
- Maximum concurrent VXML sessions
- Complexity of VXML applications

Before sizing Cisco VVB, use the Unified CCE Resource Calculator to determine the maximum number of trunks (DS0s) and VXML VRU ports to support the entire solution.

For almost all Unified CVP deployment models, sizing is based on these factors:

- The maximum concurrent VXML sessions and VoIP calls
- The CPS that Cisco VVB handles

Note

- The performance numbers listed in ASR and TTS columns are applicable only for MRCPv1 and v2.
- When Open Virtual Appliance (OVA) and VVB are already installed and the customer wants to change a profile from small to medium or vice versa, the existing OVA must be deleted and a new install with a fresh OVA specification must be done.

System Specification	CPS	DTMF (Non-Secure)	TTS / ASR (Non-Secure)	DTMF / TTS / ASR (Secure)
Medium OVA (4 CPU, 10-GB RAM)	16	600	480	480
Small OVA (4 CPU, 8-GB RAM)	16	480	380	380
KVM [4451 (2 CPU - Gladen), 8-GB RAM]	6	120	96	96
KVM [4431 (6 CPU - Gladen), 8-GB RAM]	3	80	70	70
KVM [4351 (6 CPU - Rangley), 8-GB RAM]	3	60	50	50
KVM [4331 (6 CPU - Rangley), 8-GB RAM]	2	40	30	30

Note

- TLS/SRTP reduces CPS up to 25% for small or medium profile.
- TLS/SRTP with ASR/TTS is not currently supported.
- Secure: Secured Transport over HTTPS/TLS/SRTP
- These values represent the performance with VXML pages from Unified CVP Call Studio applications running on the Unified CVP VXML Server. Other VXML applications can perform differently. These figures are for a system running VXML v2.0 and MRCPv1 or v2 with CPU utilization of less than 65 percent.

  These values reflect testing of moderately complex VXML applications on the Cisco Unified CVP VXML Server. Performance varies with different applications. Performance from external VXML applications (such as Nuance OSDMs) is not representative of the performance when interoperating with non-Cisco applications. Ensure that adequate memory is available on Cisco VVB at full load when running external VXML applications. Contact the application provider for performance and availability information.
- We make no claims or warranties regarding the performance, stability, or feature capabilities of an external VXML application added to your contact center enterprise solution.
- You can extend the HTTP cache to 512 MB in Cisco VVB.
- When calculating CPS at Cisco VVB, consider every call (VRU, Ringback, and WAAG) received. When you calculate the CPS into Cisco VVB for your solution, first determine the services which each incoming call at CVP uses.

  For example, if you disable WAAG for all agents, the total CPS at Cisco VVB is (2 x incoming rate) at CVP. That is one call for VRU and one call for Ringback. If you enable WAAG for all agents, the total CPS at Cisco VVB is (4 x incoming rate) at CVP, because WAAG adds two more calls.

Sizing for Cisco Finesse

Cisco Finesse supports up to 1800 agents and 200 supervisors (for a total of 2000 users) over HTTPS for each Cisco Finesse server pair.

Sizing for Congestion Control

Congestion Control protects the Router from overload conditions caused by high call rates. When faced with extreme overload, congestion control keeps the system running close to its rated capacity.

Congestion Control provides satisfactory service during an overloaded condition to a smaller percentage of calls, rather than a highly degraded service to all calls. The feature keeps the system within its capacity by rejecting calls at the call entry point. Throttling the capacities ensures that the routed calls receive acceptable service without timeouts.

In the discussion of Congestion Control, "calls" include nonvoice tasks from third-party multichannel applications that use the Universal Queuing APIs and voice calls. Congestion Control treats these calls and tasks the same. It monitors and throttles incoming calls and tasks, and does not drop calls or tasks once they are in the system. This means that transfers and RONAs are counted toward Congestion Control, but are not throttled or rejected.

Another exception is picking tasks like email in course of multi-tasking on a voice call. These pick task requests also do not get rejected owing to Congestion Control.

Note Pull task requests will get rejected if the system is congested except in the cases where the tasks may be waiting in the Unified CCE queue. Requests to pull tasks out of the Unified CCE queue is allowed even during congestion because this helps decongest the Unified CCE system.

Note For Enterprise Chat and Email, forwarded email tasks are considered new tasks, and are subject to throttling.

The measured CPS at the Router is the trigger for identifying congestion. The deployment type sets the supported CPS capacity for your solution. The Router measures the new incoming call requests from all the routing clients and computes a moving weighted average. If the average CPS exceeds the thresholds, the congestion levels change and the reduction percentage increases. The congestion control algorithm has three congestion levels. It rejects or treats the incoming calls at the value for that level. The system notifies the routing clients of changes in the congestion level.

In a Contact Director Reference Design, the congestion control is based on the call rate measured at each instance. The Contact Director receives information on the congestion level of each target. It applies any necessary reduction in its routing decisions. The INCRP routing client also applies congestion control to calls before sending them to the target instance.

Deployment Type Descriptions

After upgrading or installing the system, configure the system to a valid deployment type. The following table lists the supported deployment types with guidelines for selecting a valid deployment type.

For more information on the requirements referred to in this table, see your solution page on the *Cisco Collaboration Virtualization* site at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

Table 61: Deployment Types

Deployment Type Code	Deployment Name	Guidelines for Selection
0	Not Specified	This is a system default deployment type. You cannot select this option; is the default setting after fresh install or upgrade.
1	NAM (Deprecated)	Select this deployment type for NAM instance in a Contact Director deployment. The system should be distributed deployment with Router and Logger installed on different VMs, which meets the specified requirements. No agents are allowed in this deployment type. If agents are configured and signed in, the capacity is adjusted to maximum capacity of a Unified CCE 12000 Agents solution.

Deployment Type Code	Deployment Name	Guidelines for Selection
2	Contact Director	Select this deployment type for ICM instance which is dedicated to self-service call flows using Unified CVP or third-party VRU systems. The system should be distributed deployment with Router and Logger installed on different VMs which meets the specified requirements. No agents are allowed in this deployment type. If agents are configured and signed in, the capacity is adjusted to maximum capacity of an Enterprise Contact Center (Unified CCE 12000 Agents Router/ Logger).
3	NAM Rogger (Deprecated)	Select this deployment type for NAM instance in a Contact Director deployment. The Router and Logger colocated on a single VM meet the specified requirements. No agents are allowed in this deployment type. If agents are configured and signed in, the capacity is adjusted to maximum capacity of an Enterprise Contact Center (Unified CCE 12000 Agents Router/ Logger).
4	ICM Router/Logger	Select this deployment for type ICM Enterprise system where both Legacy TDM ACD PGs and CCE PGs are deployed. The system should be distributed deployment with Router and Logger installed on different VMs which meet the specified requirements.
5	UCCE: 8000 Agents Router/Logger	Select this deployment for type CCE Enterprise system where only CCE PGs are deployed. The system should be distributed deployment with Router and Logger installed on different VMs which meet the specified requirements for 8000 CCE agents.
6	UCCE: 12000 Agents Router/Logger	Select this deployment type for CCE Enterprise system where only CCE PGs are deployed. The system should be distributed deployment with Router and Logger installed on different VMs which meet the specified requirements for 12000 CCE agents.
7	Packaged CCE: 2000 Agents	Select this deployment type for a Packeged CCE production deployment.
8	ICM Rogger	Select this deployment type for ICM Enterprise system where both Legacy TDM ACD PGs and CCE PGs are deployed. The Router and Logger are colocated on a single VM which meets the specified requirements.
9	UCCE: 4000 Agents Rogger	Select this deployment type for CCE Enterprise system where only CCE PGs are deployed. The Router and Logger are colocated on a single VM which meets the specified requirements.
10	Packaged CCE: Lab Mode	Select this deployment type for a Packaged CCE lab deployment.

Deployment Type Code	Deployment Name	Guidelines for Selection
11	HCS-CC: 2000 Agents	Select this deployment type for the 2000 Agent Reference Design in a Cisco HCS for Contact Center solution. Includes the Cisco HCS for Contact Center 500 Agent design, a variation of the 2000 Agent Reference Design.
13	UCCE: Progger (Lab Only)	For all lab deployments, select this type although the Router, Logger, and PG are not on the same VM. **Note** This deployment type is not supported for production systems.
14	HCS-CC: 4000 Agents	Select this deployment for Unified CCE system where only Unified CCE PGs are deployed. This deployment is for distributed systems with the Router and Logger on other servers that meet the requirements for 4000 Unified CCE agents.
15	HCS-CC: 12000 Agents	Select this deployment type for the 12000 Agent Reference Design in a Cisco HCS for Contact Center solution.
16	UCCE: 2000 Agents	Select this deployment type for the 2000 Agent Reference Design in a Unified CCE solution.
17	Packaged CCE: 4000 Agents	Select this deployment type for the 4000 Agent Reference Design in a Packaged CCE solution.
18	Packaged CCE: 12000 Agents	Select this deployment type for the 12000 Agent Reference Design in a Packaged CCE solution.
19	UCCE: 24000 Agents Router/Logger	Select this deployment type for the 24000 Agent Reference Design in a Unified CCE solution.
20	HCS-CC: 24000 Agents	Select this deployment type for the 24000 Agent Reference Design in a HCS-CC solution.

Note It is important to set the proper deployment type for your solution during the configuration. If you select the wrong deployment type, your solution is either unprotected from overload or it rejects and treats calls based on incorrect capacity settings.

Congestion Treatment Mode

The system has five options to handle the calls that are rejected or treated due to congestion. You can choose any of the following options to handle the calls:

- **Treat Call with Dialed Number Default Label**—The rejected calls are treated with the default label of the dialed number on which the incoming call arrived.

- **Treat call with Routing Client Default Label**—The rejected calls are treated with the default label of the routing client on which the incoming call arrived.
- **Treat call with System Default Label**—The rejected calls are treated with the system default label set in Congestion Control settings.
- **Terminate call with a Dialog Fail or RouteEnd**—Terminates the incoming call dialog with a dialog failure.
- **Treat call with a Release Message to the Routing Client**—Terminates the incoming call dialog with a release message.

You set the treatment options in the congestion settings either at the routing client or at the global level. If you select a treatment mode at the routing client, it takes precedence over the system congestion settings.

Note If you choose to return a label back to treat the call with an announcement, use an announcement system external to the Unified CCE instance. Never return a treated call to the Unified CCE instance for further processing.

Call Treatment for Outbound Option

Outbound Option is a special case for call treatment with Congestion Control. When you integrate the Media Routing Peripheral Gateway (MR PG) for Outbound Option, configure the PG's routing client to always send the dialog failure. The dialer retries the rejected reservation calls after a specified period.

Congestion Control Levels and Thresholds

The Congestion Control algorithm works in three levels. Each level has onset and abatement values. When the average CPS exceeds one level's onset value, the system moves to a higher congestion level. For example, if the system is at level 0 and the CPS exceeds the Level 2 onset capacity, the system moves directly to Level 2. The congestion level reduces when the average CPS falls below the current level's abatement value. Congestion levels can rise several levels at once. However, the congestion level reduces only one level at a time.

Table 62: Congestion Levels

Congestion Levels	Threshold (Percent of Capacity)	Description
Level1Onset	110%	If the average CPS exceeds this value, the congestion level moves to Level 1.
Level1Abate	90%	If the average CPS goes below this value, the congestion level moves back to Level 0 (Normal operating Level).
L1Reduction	10%	The percentage of incoming calls that are rejected at Level 1 congestion.
Level2Onset	130%	If the average CPS exceeds this value, the congestion level moves to Level 2.

Congestion Levels	Threshold (Percent of Capacity)	Description
Level2Abate	100%	If the average CPS goes below this value, then the congestion level moves back to Level 1.
Level2Redution	30%	The percentage of incoming calls that are rejected in Level 2 congestion.
Level3Onset	150%	If the average CPS exceeds this value, the congestion level moves to Level 3.
Level3Abatement	100%	If the average CPS goes below this value, the congestion level moves back to Level 2.
Level3Reduction	Variable reduction from 100% to 30%	The percentage of incoming calls that are rejected in Level 3 congestion. Depending on the incoming call rate, the reduction percentage varies from 30% to 100% when the congestion level enters Level 3.

Note You cannot configure the onset, abatement, and reduction settings. These values are defined as a percentage of the standard CPS capacity for the system.

Congestion Control CPS Limits

This table lists the maximum supported calls per second (CPS) for the supported deployment types.

Table 63: Deployment Types

Deployment type	Maximum calls per second	Notes
NAM (Deprecated)	300	Deprecated as of 11.5.
Contact Director	300	Reference Design
NAM Rogger (Deprecated)	150	Deprecated as of 11.5.
UCCE: 12000 Agents	105	Reference Design
Packaged CCE: 2000 Agents	18	Reference Design
UCCE: 4000 Agents	35	Reference Design
Packaged CCE: Lab Mode	1	Not supported for production environments.
HCS-CC: 2000 Agents	18	Reference Design
UCCE: Progger (Lab Only)	4	Not supported for production environments.
HCS-CC: 4000 Agents	35	Reference Design

Deployment type	Maximum calls per second	Notes
HCS-CC: 12000 Agents	105	Reference Design
UCCE: 2000 Agents	18	Reference Design
Packaged CCE: 4000 Agents	35	Reference Design
Packaged CCE: 12000 Agents	105	Reference Design
UCCE: 24000 Agents Router/Logger	105	Reference Design
HCS-CC: 24000 Agents	105	Reference Design

Operating Considerations for Reference Design Compliant Solutions

Solution-Wide Support for Transport Layer Security

The contact center enterprise solutions use TLS 1.2 by default. For most components, you can enable earlier versions of TLS if necessary.

Time Synchronization for Your Solution

To ensure accurate operation and reporting, all the components in your contact center solution must use the same value for the time. You can synchronize the time across your solution using a Simple Network Time Protocol (SNTP) server. The following table outlines the needs of various component types in your solution.

Important Use the same NTP sources throughout your solution.

Type of component	Notes
Domain controllers	Domain controllers must all point to the same NTP servers.
ESXi hosts	All ESXi hosts must point to the same NTP servers as primary domain controllers.
Windows components in the contact center domain	Windows machines in the domain point to, and are automatically in synch with, the primary domain controller for NTP. They require no configuration for NTP.

Type of component	Notes
Windows components not in the contact center domain	Follow the Microsoft documentation to synchronize directly with the NTP server.
Non-Windows components	Components such as Unified Intelligence Center, Cisco Finesse, Customer Collaboration Platform, and Unified Communications must point to the same NTP servers as the domain controllers.
Cisco Integrated Service Routers	To provide accurate time for logging and debugging, use the same NTP source as the solution for the Cisco IOS Voice Gateways.

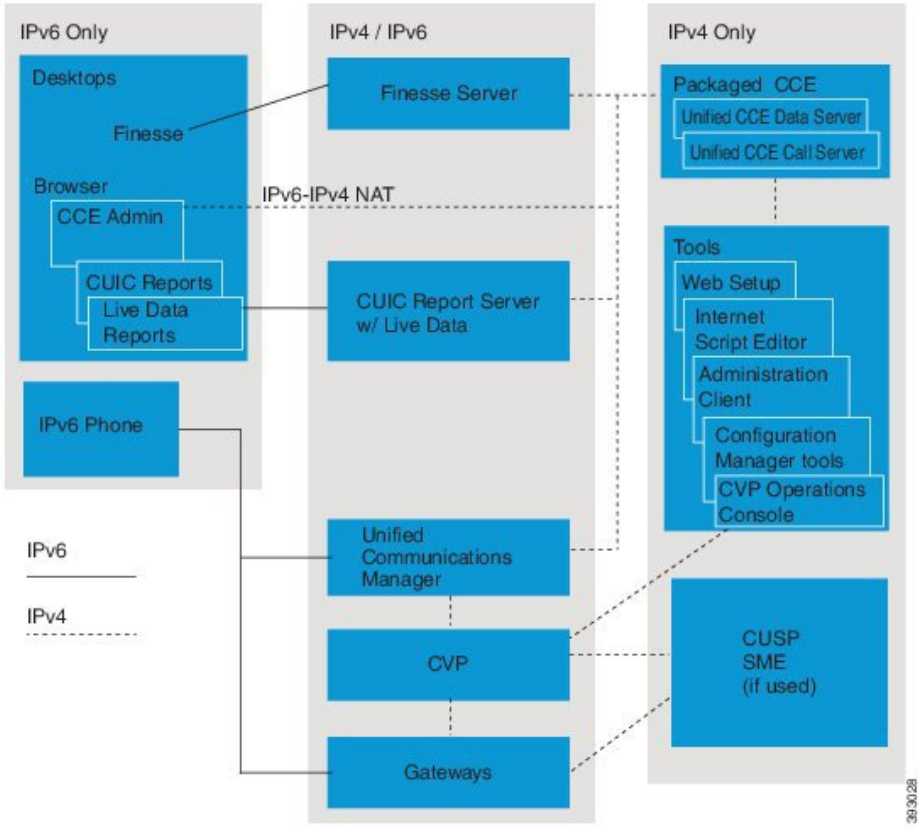
Contact Center Enterprise Solution Support for IPv6

Unified Contact Center solutions can support IPv6 connections for agent and supervisor Finesse desktops and phones. This support means that most of the endpoints in your deployment can use IPv6 addresses.

Your IPv6-enabled deployment can use either IPv6-only or a mix of IPv4 and IPv6 endpoints. Servers that communicate with those endpoints can now accept IPv6 connections, in addition to IPv4 connections. Communication between servers continues to use IPv4 connections.

This diagram shows a logical view of a deployment with only IPv6 desktops and phones:

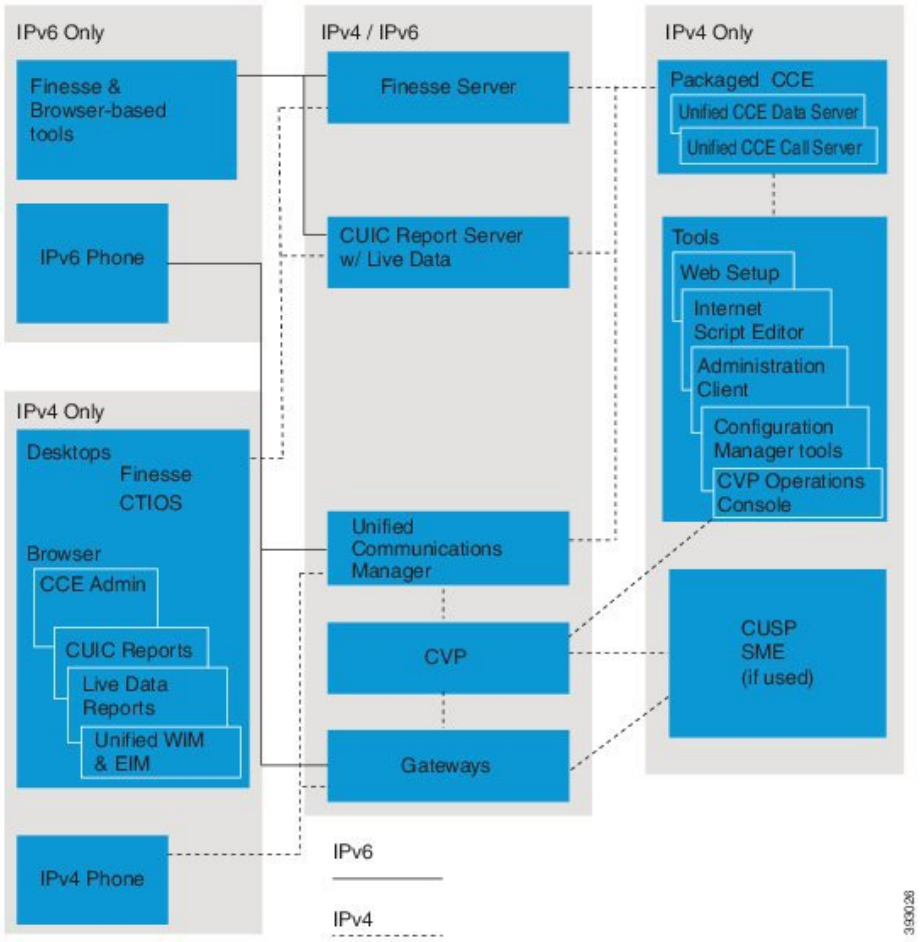
Figure 125: Packaged CCE Deployment with Only IPv6 Agents



In these IPv6-only deployments, agents and supervisors use Finesse and browser-based tools that connect to dual-stack interfaces on the servers. The ingress gateways and Unified CM also use dual-stack interfaces to handle the voice traffic. These deployments require IPv4-based Administration Workstations to run the configuration tools that you do not access through a browser.

This diagram shows a logical view of a mixed deployment with both IPv6 and IPv4 endpoints:

Figure 126: Packaged CCE Deployment with Both IPv4 and IPv6 Agents



The Finesse desktop can support either IPv4 or IPv6 connections. Agents and supervisors who use the CTI OS desktops must use IPv4 connections. Enterprise Chat and Email agents must use IPv4 connections.

For a list of endpoints that support IPv6, see your solution *Compatibility Matrix*.

For information on enabling IPv6 in the Cisco Unified Communications Manager, see *Deploying IPv6 in Unified Communications Networks with Cisco Unified Communications Manager* at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

General IPv6 Design Considerations

You cannot enable IPv6 on all the component servers in your contact center. For example, your deployment might use IPv6 phones, but only use IPv4 desktops. In that deployment, you enable IPv6 on the component servers that connect to the desktops.

When an IPv4 endpoint communicates with an IPv6 endpoint, Unified Communication Manager invokes Media Termination Points (MTPs) to negotiate the mismatch. As a result, IPv4-only endpoints like the VXML browser require extra MTP devices. Most uses of CVP features require MTPs for the IPv4-to-IPv6 negotiation.

You do not have to set up IPv6 at installation. You can enable IPv6 at any time. You can also revert to IPv4 from IPv6 if necessary.

You assign IPv6 addresses to hosts that have a dual IP stack. You use the Fully Qualified Domain Name (FQDN), rather than the IPv6 address, in the solution's user interface.

CVP Features with IPv6

When you enable IPv6 in your solution, CVP requires the following conditions to support its features:

- **Call Survivability**—Use only IPv4 for the incoming trunk to the gateway.
- **Courtesy Callback and Refer**—When a trunk carries both the incoming and outbound dialing traffic, the session target in the Ingress Gateway dialpeer uses the same protocol as the incoming trunk.

Desktop and Tool Support

This table lists which desktops support each connection type:

Desktop	IPv6 Connections	IPv4 Connections
Finesse	Yes	Yes
	No	Yes

A supervisor's team can include a mix of agents using Finesse desktops with either IPv4 or IPv6 connections and CTI OS desktops with IPv4 connections.

You cannot use IPv6 to connect to a Finesse desktop through Citrix XenApp.

Desktops with either IPv4 or IPv6 connections can access the following tools:

- Unified CCE Administration web tool (using NAT64).
- Finesse configuration tools
- Cisco Unified Intelligence Center (Cisco Unified IC) configuration tools and reports

You require an IPv4 connection to access the following tools:

- Enterprise Chat and Email
- Web Setup
- Script Editor
- Internet Script Editor
- Diagnostic Portico
- Configuration Manager and its associated tools

IPv6 Design Considerations for Video Endpoints

If you use video endpoints, consider the following points when enabling IPv6:

- Configure the incoming trunk to gateway in IPv4 mode only.
- Disable ANAT in the Ingress Gateway.

- Agent devices can use either IPv4 or dual IP mode.

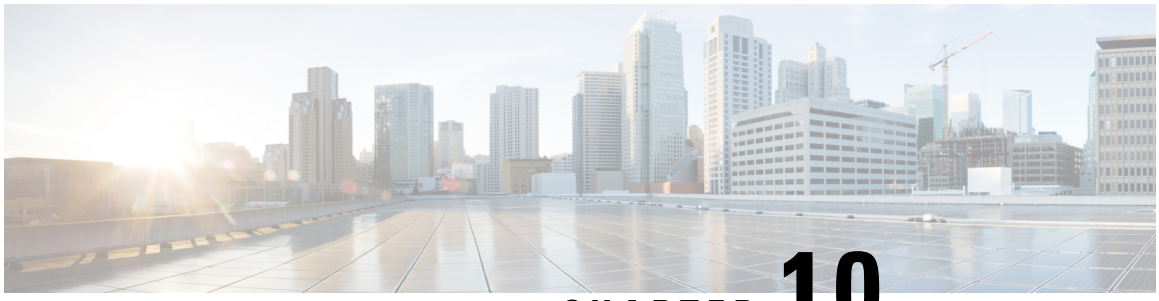
Other Component and Feature Support

This table lists the connection type that each component or feature supports in an IPv6-enabled environment:

Component or Feature	Supported Connections in IPv6-enabled Environment		Notes
	IPv6	IPv4	
Enterprise Chat and Email	No	Yes	
Mobile Agent	No	Yes	The CTI ports for Mobile Agent can only have an **IP Addressing Mode** of **IPv4 Only**.
Outbound Option	No	Yes	The Outbound Option Dialer uses IPv4 to place calls. A voice gateway that supports both IPv4 and IPv6 renegotiates call signaling and media to IPv6 during referral to an IPv6 agent. You cannot use an IPv6-only voice gateway with Outbound Option. An IPv6 client cannot import to Outbound Option.
Customer Collaboration Platform	No	Yes	
Unified CM Silent Monitoring	Yes	Yes	
Virtualized Voice Browser	No	Yes	You cannot use Cisco VVB in an IPv6-enabled environment.

For more information on enabling IPv6 in a contact center enterprise solution, see your solution's *Installation and Upgrade Guide*. These documents have more details for specific products:

Component	Documents
Unified CVP	*Configuration Guide for Cisco Unified Customer Voice Portal*
Cisco Finesse	*Cisco Finesse Installation and Upgrade Guide* *Cisco Finesse Administration Guide*
Cisco Unified Intelligence Center	*Administration Console User Guide for Cisco Unified Intelligence Center*

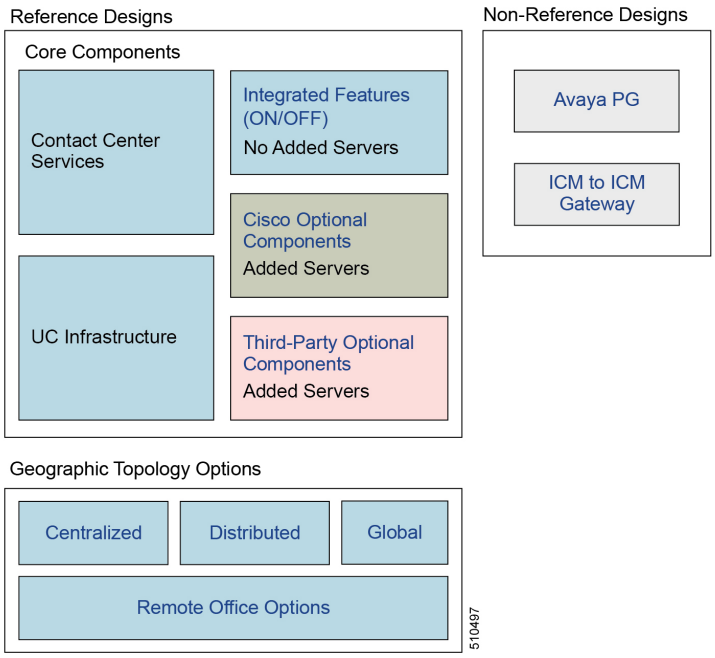


CHAPTER 10

Avaya and ICM-to-ICM Gateway Support

Introduction

Packaged CCE supports Avaya PG and ICM-to-ICM Gateway as a Non-Reference Design solution. You must deploy Avaya PG on a separate VM.



Packaged CCE offers several call flows to support different needs. Use the following in a Non-Reference Design:

- Pre-route call flows

- Translation route
- Post-route call flows

For more information , see the *Cisco Packaged Contact Center Enterprise Features Guide* at https://www.cisco.com/c/en/us/support/customer-collaboration/packaged-contact-center-enterprise/products-maintenance-guides-list.html.

Unified CVP Type 10 is the supported Network VRU.

Configuration Limits and Scalability Constraints

The following table specifies the configuration limits and scalability constraints for the Non-Reference parameters that are supported in the Packaged CCE 4000 and 12000 Agents Deployment.

When you design your contact center, ensure that your design is deployed within these limits. Consult Cisco if you have special configuration requirements that might exceed specific parameters.

Note System configuration limits are defined in the Reference Designs. For more information, see Reference Design Configuration Limits, on page 121.

Table 64: Packaged CCE Configuration Limits and Scalability Constraints

Parameter	Limit Value	Comments
	>450 - <=12,000 agents	
Avaya PIMs on each Avaya Peripheral Gateway (PG)	5	Multiple PIMs on a PG affect performance. Compared to a single PIM on each PG, multiple PIMs lower the total number of agents, VRU ports, and supported call volume. There is a maximum of one PIM on each Avaya PG with CTI OS coresident.
Maximum number of CTI servers per PG	1	
Skill groups on each Avaya PG	4000	

Additional Sizing Factors

Many variables in the Packaged CCE configuration and deployment options can affect the server requirements and capacities. This section describes the major sizing variables and how they affect the capacity of the various Packaged CCE components.

Average Skill Groups Per Agent

The number of skill groups per agent (which is independent of the total number of skills per system) significantly affects the CTI OS servers.

Limit the number of skill groups per agent to 5 or fewer, when possible. Periodically remove unused skill groups so that they do not affect the system's performance. You can also manage the effects on the CTI OS Server by increasing the value for the frequency of statistical updates.

Note CTI OS monitor mode applications are supported only at 20 or lower skill groups per agent.

CTI OS Monitor Mode Applications

A CTI OS Monitor Mode application can affect the performance of the CTI OS Server. CTI OS supports only two such applications per server pair. Depending on the filter specified, the impact on the CPU utilization might degrade the performance of the Agent PG.

CTI OS Skill Group Statistics Refresh Rate

The skill group statistics refresh rate can also affect the performance of CTI OS Server. Cisco requires that you do not lower the refresh rate below the default value of 10 seconds.

Translation Route Pool

Sizing the translation route pool depends on the expected call arrival rate. Use the following formula to size the translation route pool:

Translation route pool = 20 * (Calls per second)

This calculation is specific to Packaged CCE. For more general Packaged CCE deployments, consult your Cisco Account Team or Partner.

ACD Call Deployments and Sizing Implications

The information in this section applies to ACD integrations that use Unified CVP. The ACD device shares the following characteristics:

- Manage agents and transfer calls to the destination.
- Route requests and be switch leg devices. However, the device cannot handle Correlation ID and more than one transfer.

An ACD user issues a Route Request for one of the following reasons:

- Connect to another agent in a particular skill group
- Reach a self-service application
- Blind-transfer a previously received call to one of the above entities

Each of the above calls invokes a routing script. The script searches for an available destination agent or service and if an appropriate destination is found, it sends the corresponding label either back to the ACD or, if blind-transferring an existing call, to the original caller's Switch leg device. If it needs to queue the call or if the ultimate destination is intended to be a self-service application rather than an agent or service, the script sends a VRU translation route label either back to the ACD or, if transferring an existing call through blind-transfer, to the original caller's Switch leg device.

If the above sequence results in transferring the call to Unified CVP's VRU leg device, a second transfer is done to deliver it to a Voice Browser. To ensure that these events take place, the following configuration elements are required:

- For new calls from the ACD or warm transfers of existing calls:
  - Configure the Unified CVP peripheral to be associated with a Type 10 Network VRU.
  - Associate the dialed number that the ACD dialed with a Customer Instance that is associated with a Type 10 Network VRU.
  - When an ACD is not configured , the routing script that is invoked by the ACD dialed number must contain a TranslationRouteToVRU node to get the call to Unified CVP's Switch leg, followed by a SendToVRU node to get the call to the Voice Browser and Unified CVP's VRU leg.
  - Associate all the VRU scripts that are run by that routing script with the Type 10 Network VRU.
- For blind transfers of existing calls:
  - The Unified CVP peripheral can be associated with any Network VRU.
  - The dialed number that the ACD dialed must be associated with a Customer Instance that is associated with a Type 10 Network VRU.
  - The routing script that is invoked by the ACD dialed number must contain a SendToVRU node to send the call to the Voice Browser and Unified CVP's VRU leg.
  - All the VRU scripts that are run by that routing script must be associated with the Type 10 Network VRU.

When Packaged CCE chooses an agent or ACD destination label for a call, it tries to find one that lists a routing client that can accept that label. For calls originated by an ACD that are not blind transfers of existing calls, the only routing client is the ACD , after the call is transferred to Unified CVP, because of the handoff operation, the only routing client is the Unified CVP Switch leg. However, in the case of blind transfers of existing calls, two routing clients are possible:

- The Call Server switch leg that delivered the original call.
- The ACD . For calls that originate through Unified CVP, you can prioritize Unified CVP labels above ACD labels by checking the **Network Transfer Preferred** check box in the Packaged CCE screen for the Unified CVP peripheral.

Agent Desktops

A Packaged CCE deployment requires an agent desktop application. The agent uses this desktop for agent state control and call control. In addition to these required features, the desktop can provide other useful features.

Cisco offers the following agent desktop application for Avaya agents:

- CTI Toolkit Desktop: An agent desktop application built with the CTI Toolkit. The desktop supports full customization and integration with other applications, customer databases, and Customer Relationship Management (CRM) applications.

Cisco partners offer the following types of agent desktop applications:

- Partner agent desktops: Custom agent desktop applications are available through Cisco Technology Partners. These applications are based on the CTI Toolkit and are not discussed individually in this document. The Finesse REST API also enables partner desktop integration.
- Prepackaged CRM integrations: CRM integrations are available through Cisco Unified CRM Technology Partners. These integrations are based on the CTI Toolkit and are not discussed individually in this document.

Desktop applications typically run on agent desktops, Administration & Data servers, or administration clients. Services that support the desktop applications can run on the Avaya Peripheral Gateway (PG) server or on their own server. For each PG, there is one set of active desktop services.

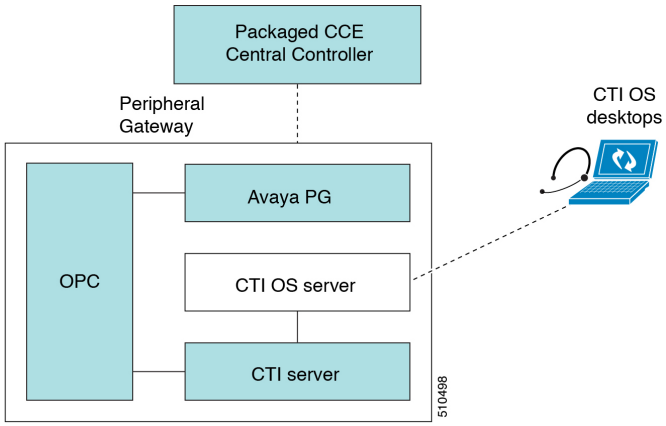
CTI Object Server

The CTI Object server (CTI OS) is a high-performance, scalable, fault-tolerant, server-based solution for deploying CTI applications. CTI OS is a required component for the CTI Toolkit Desktop. The CTI OS server runs as a redundant pair, one server on each VM that hosts an Avaya PG.

Desktop applications pass communications, such as agent state change requests and call control, to the CTI OS server. CTI OS is a single point of integration for CTI Toolkit Desktops and third-party applications, such as Customer Relationship Management (CRM) systems, data mining, and workflow solutions.

The CTI Object server connects to the CTI server over TCP/IP and forwards call control and agent requests to the CTI server.

Figure 127: CTI OS Desktop Architecture



The CTI OS server also manages CTI Toolkit desktop configuration and behavior information, simplifying customization, updates, and maintenance, and supporting remote management.

CTI Object Server Services

- Desktop security: Supports secure socket connections between the CTI Object server on the PG and the agent, supervisor, or administrator desktop PC. Any CTI application built using the CTI OS Desktop Toolkit (CTI Toolkit) C++/COM CIL SDK can use the desktop security feature.

Note Desktop Security is not currently available in the .NET and Java CILs.

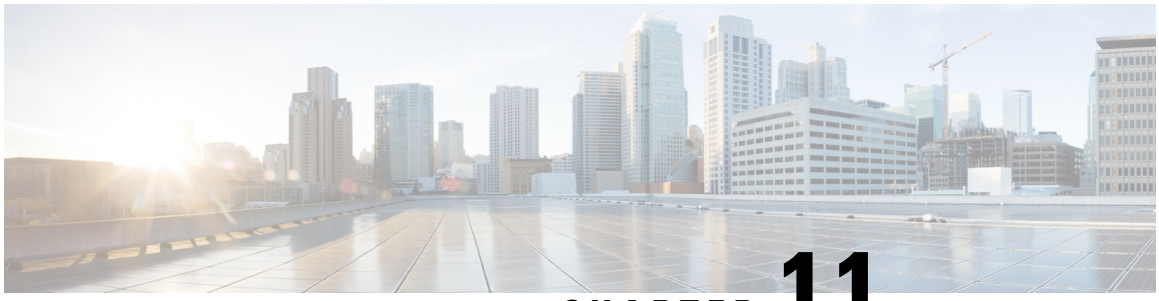
- Quality of Service (QoS): Supports packet prioritization with the network for desktop call control messages.

Note QoS is not currently available in the .NET and Java CILs.

- Failover recovery: Supports automatic agent sign-in upon failover.
- Chat: Supports message passing and the text chat feature between agents and supervisors.

You deploy the CTI Object server in redundant pairs, one on Avaya PG A and one on Avaya PG B. Both CTI OS servers are active simultaneously. The CTI Toolkit desktop applications randomly connect to one of the two servers. If the connection to the original server fails, the desktops automatically fail over to the alternate server.

Note The CTI OS server interfaces to any desktop application built using the CTI Toolkit SDK.



CHAPTER 11

Solution Security

Decouple CCE Authorizations from Active Directory

Prior to Release 12.0(1), Packaged CCE uses Microsoft Active Directory Security Groups to control user access rights to perform setup and configuration tasks. Microsoft AD also grants permissions for system components to interact; for example, it grants permissions to a Distributor to read the Logger database. Microsoft AD manages the user privileges that are associated with the Security Groups - Setup, Config, and Service. Thus, Microsoft AD handled both authentication and authorization. In such cases, Microsoft AD must assign user privileges to the Security Groups. To accomplish this, Packaged CCE solution administration requires write permissions to Microsoft AD for authorization.

By default, Packaged CCE now decouples authentication and authorization functions.

Decoupling authentication and authorization removes the need to use Microsoft AD to manage authorization in Packaged CCE components. The Packaged CCE solution requires that you add user IDs to the local user groups on each local machine for authorizations. User privileges are provided by memberships to local user groups in the local machines. Microsoft AD is only used for authentication.

To authorize a user ID that is already present in the Microsoft AD, you associate or add the user ID to the local user groups:

- Associate the user ID with the local `UcceService` security group to provide the SQL server authorizations to the user ID for read/write operations in the SQL database. Use the Service Account Manager tool to assign a domain user as a service account user.
- Add the user ID to the local Administrators group for Packaged CCE Setup operations. Add the user ID to the local `UcceConfig` security group for Packaged CCE configuration operations using the Configuration Manager tools.

ADSecurityGroupUpdate Registry Key

This Registry key allows or disallows updates to the Config and Setup security groups in the Domain under an instance Organizational Unit (OU).

The key has two values as follows:

- 0—Indicates that the Administrator gadget only updates the User_Role column in the User Group table in the database schema and *not* the Config and Setup security groups in the domain under instance OU.
- 1—Indicates that the Administrator gadget updates the User_Role column in the User Group table in the database schema *and* the Config and Setup security groups in the domain under instance OU.

The default value is 0.

User Health in Service Account Manager

After upgrade, the Service Account Manager checks the users in the `UcceService` local group. If the users are not in the `UcceService` local group, then the Service Account Manager displays the status as *Unhealthy*. In such a case, run **Fix Group Membership** to make the status healthy. Alternatively, provide the new domain user in the Service Account Manager (SAM) tool or in `Websetup`

For more information about the enhancements, see the following guides:

- The chapter on the Service Account Manager in the Staging Guide for Cisco Unified ICM/Contact Center Enterprise.
- The sections on adding components to Packaged CCE instances, configuring permissions in the local machine, and migrating databases in the *Cisco Packaged Contact Center Enterprise Administration and Configuration Guide*.

Organizational Units

Application-Created OUs

When you install the solution software, the AD Domain in which the VMs are members must be in Native Mode. The installation adds several OU objects, containers, users, and groups for the solution. You need delegated control over the Organizational Unit in AD to install those objects. You can locate the OU anywhere in the domain hierarchy. The AD Administrator determines how deeply nested the contact center enterprise solution OU hierarchy is created and populated.

Note All created groups are Domain Local Security Groups, and all user accounts are domain accounts. The Service Logon domain account is added to the Local Administrators' group of the application servers.

The contact center enterprise installation integrates with a Domain Manager tool. You can use the tool standalone for preinstalling the OU hierarchies and objects required by the software. You can also use it when the Setup program is invoked to create the same objects in AD. The AD/OU creation can be done on the domain in which the running VM is a member or on a trusted domain.

Active Directory Administrator-Created OUs

An administrator can create certain AD objects. A prime example is the OU container for Unified CCE Servers. This OU container is manually added to contain the VMs that are members of a given domain. You move these VMs to this OU once they are joined to the domain. This segregation controls who can or cannot

administer the servers (delegation of control). Most importantly, the segregation controls the AD Domain Security Policies that the application servers in the OU can or cannot inherit.