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.

The data in this table is based on the Cisco Collaboration Infrastructure Requirements and is provided only for planning purposes. The results may vary based on the hardware, load, and other factors that are applicable to your deployment.

Table 1: 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	2985	1698	1690	5156	1414	1385	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	5387	527	1518	3380	669	643	
AW-HDS	2635	862	996	NA			
HDS-DDS	7719	454	985				
Cisco Identity Server(Ids)	260	56	224				
Live Data	24723	18273	20950				

Rogger	3786	1006	1625	683	337	328	4,000 agents; 30 CPS; ECC; 5 scalars @ 40 bytes each; 200 Reporting users at max query load.
AW-HDS-DDS	2550	859	1206	NA			
Cisco Identity Server(Ids)	667	50	386				
Live Data	9393	4938	6614				
Rogger	3129	934	1394	1938	574	749	2,000 agents; 15 CPS; ECC; 5 scalars @ 40 bytes each; 100 Reporting users at max query load and 7 CTI Clients
AW-HDS-DDS	2922	761	1229	NA			
CUIC-LD-Ids	32123	3501	5581				
CUIC	2315	1471	1734	NA			200 Reporting users running 2 Real Time reports with 100 rows retrieved for each report that is refreshed every 15 seconds, and 2 Historical Reports with 2000 rows retrieved for each report that is refreshed every 30 minutes, during call load for all deployments.
Finesse	2984	2869	2931				1000 agents; 15 CPS; applicable for all deployments.
CVP Call/VXML	1790	1120	1186				7.5 CPS; applicable for all deployments.
CVP Reporting	735	709	718				Applicable for data write during load for all deployments.
PG	2755	2541	2566	1378	1357	1370	2000 agents; 15 CPS; applicable for all deployments.

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 2: 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

VXML Document Type	Approximate Size in bytes
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
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.

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.

Latency Considerations

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 3: 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 4: 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 5: 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 6: 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 20 and Bandwidth, Latency, and QoS for Unified Intelligence Center, on page 22.

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 and Latency Considerations for Cloud Connect

When the publisher and subscriber nodes of the Cloud Connect servers are geographically distributed, the maximum latency between the nodes is 200 milliseconds one way.

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.

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 and Latency Considerations for Webex Connect Integration

Ensure that the maximum average latency between Cloud Connect and Webex Connect is 200 milliseconds or lesser. However, we recommend that the latency is less than 100 milliseconds to ensure there is no issues with the customer experience and/or conversations appearing on the Agent Desktop without delay. If the latency exceeds 200 milliseconds, you may notice that the webhook notifications get piled up in the Digital Routing service that can lead to performance issues.

The maximum latency between Agent Desktop and Webex Engage nodes is 200 milliseconds.

The digital channels interaction using Webex Connect requires higher client bandwidth. See Finesse Bandwidth Calculator.

Bandwidth and Latency Considerations for Virtual Voice Agent

In Virtual Agent Voice feature, VVB Speech Server relays the voice media towards Google Dialogflow or CCAI Orchestrator via gRPC protocol based on your deployment. This requires a bandwidth of 106 Kbps per VAV call. For information on the recommended round-trip latency for maximum reliability, see the Google documentation at https://cloud.google.com/contact-center/ccai-platform/docs/Getting_Started. If this latency value goes higher, significant delays may occur during VAV calls, impacting functionality.

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.