Bandwidth, Latency, and QoS Considerations

Bandwidth, Latency, and QoS for Core Components

Estimating Bandwidth Consumption

Bandwidth plays a large role in deployments involving:

- The centralized call processing model (Unified CCX at the central site)
- Any call deployment model that uses call admission control or a gatekeeper
- Any deployments involving email processing model (via SocialMiner).

Remote Agent Traffic Profile

Unified CCX signaling represents only a very small portion of control traffic (Agent and Supervisor Desktop to and from the Unified CCX Server) in the network. For information on TCP ports and Differentiated Services Code Point (DSCP) marking for Unified CCX and CTI traffic.

Bandwidth estimation becomes an issue when voice is included in the calculation. Because WAN links are usually the lowest-speed circuits in an IP Telephony network, particular attention must be given to reducing packet loss, delay, and jitter where voice traffic is sent across these links. G.729 is the preferred codec for use over the WAN because the G.729 method for sampling audio introduces the least latency (only 30 milliseconds) in addition to any other delays caused by the network.

Where voice is included in bandwidth, system architects should consider the following factors:

- Total delay budget for latency (taking into account WAN latency, serialization delays for any local area network traversed, and any forwarding latency present in the network devices). The generally agreed-upon limit for total (one-way) latency for applications in a network is 150 milliseconds.
- Impact of delays inherent in the applications themselves. The average Unified CCX agent login time with no WAN delay is 8 seconds. This includes the exchange of approximately 1,000 messages between the agent application and various servers. The overall time to log in agents increases by approximately 30 seconds for each 30 milliseconds of WAN delay.

- Impact of routing protocols. For example, Enhanced Interior Gateway Routing Protocol (EIGRP) uses quick convergence times and conservative use of bandwidth. EIGRP convergence also has a negligible impact on call processing and Unified CCX agent logins.
- Method used for silently monitoring and recording agent calls. The method used dictates the bandwidth load on a given network link.

Use the following table to estimate the number of Unified CCX agents that can be maintained across the WAN (with IP Telephony QoS enabled). These numbers are derived from testing where an entire call session to Unified CCX agents, including G.729 RTP streams, is sent across the WAN. Approximately 30 percent of bandwidth is provisioned for voice. Voice drops are more of an issue when you are running RTP in conjunction with Cisco Finesse and other background traffic across the WAN. These voice drops might occur with a specific number of agents at a certain link speed, and those possible scenarios are denoted by the entry N/A (not applicable) in the following table.

Table 1: Remote Agents Supported by Unified CCX Across a WAN Link

Frame Relay	128 KB	256 KB	512 KB	768 KB	T1
G.729	3	7	15	25	38
G. 711	N/A	N/A	N/A	N/A	14

In remote agent deployments, QoS mechanisms should be used to optimize WAN bandwidth utilization. Advanced queuing and scheduling techniques should be used in distribution and core areas as well. For provisioning guidelines for centralized call processing deployments, refer to the *Cisco IP Telephony Solution Reference Network Design* documentation, available online at: http://www.cisco.com/go/ucsrnd.

IP Call Bandwidth Usage

An IP phone call consists of two streams of data. One stream is sent from phone A to phone B. The other stream is sent from phone B to phone A. The voice data is encapsulated into packets that are sent over the network. The amount of data required to store a voice stream is dependent upon the Codec used to encode the data.

The voice data itself is transmitted over the network using the Real-Time Transport Protocol (RTP). The RTP protocol supports *silence suppression*. When silence suppression is used, no voice packets are sent over the network if there is no sound.

Otherwise, even packets that contain silence are sent. This situation lowers the average required bandwidth for a call. Although it supports silence suppression, the lower bandwidth requirements for silence suppression should not be used when provisioning the network because the worst case scenario would be where there is not silence in the call, requiring the full call bandwidth as if silence suppression was not enabled.

When calculating bandwidth for an IP call, you must use the size of the RTP packet plus the additional overhead of the networking protocols used to transport the RTP data through the network.

For example, G.711 packets carrying 20 ms of speech data require 64 kbps (kilobytes per second) of network bandwidth per stream. These packets are encapsulated by four layers of networking protocols (RTP, UDP, IP, and Ethernet). Each of these protocols adds its own header information to the G.711 data. As a result, the G.711 data, once packed into an Ethernet frame, requires 87.2 kbps of bandwidth per data stream as it travels over the network. Because an IP phone call consists of two voice streams, in this example, a call would require 174.4 kbps.

The amount of voice data in a single packet also influences the size of the packet and bandwidth. The example above used packets containing 20 milliseconds of speech for its calculations, but this value can be changed in the Unified CM configuration for each supported Codec. Configuring packets to contain more speech information reduces the number of packets sent over the network and reduces the bandwidth because there are fewer packets containing the additional networking headers, but the packet sizes increase.

The following table shows the bandwidth required for a phone call for the different combinations of Codec and amount of speech per packet.

Table 2: Per-call Packet Size Bandwidth Requirements

Codec	Milliseconds of speech per packet	Bandwidth required (Kbps) for a call
G.711	10	220.8
G.711	20	174.4
G.711	30	159.0
G.729	10	108.8
G.729	20	62.4
G.729	30	47.0
G.729	40	39.2
G.729	50	34.6
G.729	60	31.4

Note

- The calculations are based on G.711 using a sampling rate of 64 kbps speech encoding and the G.729 using 8 kbps. This means one second of speech encoded into the G.711 Codec requires 65,536 bits (or 8,192 bytes) to represent one second of sound.
- For full-duplex connections, the bandwidth speed applies to both incoming and outgoing traffic. (For instance, for a 100-Mbps connection, there is 100 Mbps of upload bandwidth and 100 Mbps of download bandwidth.) Therefore, an IP phone call consumes the bandwidth equivalent of a single stream of data. In this scenario, a G.711 IP phone call with no silence suppression and containing 20 milliseconds of speech per packet requires 87.2 kbps (174.4 / 2) of the available bandwidth.
- Unified CCX supports a-law and μ-law for G.711.
- If a prompt is recorded with G711 a-law phones and uploaded, you may encounter an error while playing the recorded prompt.

Bandwidth Available for Monitoring and Recording

The following tables display the percentage of total bandwidth available, based on the network connection, which is required for simultaneous monitoring sessions handled by a VoIP provider.

Table 3: Available Upload Bandwidth Percentage for Simultaneous Monitoring Sessions with G.729 Codec

Number of Simultaneous Monitoring Sessions	Percentage of Available Bandwidth Required (No Silence Suppression)							
	100 Mbps	10 Mbps	1.544 Mbps	640 kbps	256 kbps	128 kbps	64 kbps	56 kbps
Call only	0.1	0.9	5.6	13.6	34.1	68.1	Not supported (NS)[1]	
1	0.3	2.6	16.8	40.9	NS	NS	NS	NS
2	0.4	4.4	28.1	68.1	NS	NS	NS	NS
3	0.6	6.1	39.3	95.4	NS	NS	NS	NS
4	0.8	7.8	50.5	NS	NS	NS	NS	NS
5	1.0	9.6	61.7	NS	NS	NS	NS	NS
6	1.1	11.3	72.9	NS	NS	NS	NS	NS
7	1.3	13.1	84.2	NS	NS	NS	NS	NS
8	1.5	14.8	95.4	NS	NS	NS	NS	NS
9	1.7	16.6	NS	NS	NS	NS	NS	NS
10	1.8	18.3	NS	NS	NS	NS	NS	NS

[1] The bandwidth of the connection is not large enough to support the number of simultaneous monitoring sessions.

Table 4: Available Upload Bandwidth Percentage for Simultaneous Monitoring Sessions with G.711 Codec

Number of Simultaneous Monitoring Sessions	Percentage of Available Bandwidth Required (No Silence Suppression)						
	100 Mbps	10 Mbps	1.544 Mbps	640 kbps	256 kbps	128 kbps	64 kbps
Call only	0.0	0.3	2.0	4.9	12.2	24.4	48.8
1	0.1	0.9	6.0	14.6	36.6	73.1	Not supported (NS)[2]
2	0.2	1.6	10.0	24.4	60.9	NS	NS
3	0.2	2.2	14.1	34.1	85.3	NS	NS
4	0.3	2.8	18.1	43.9	NS	NS	NS
5	0.3	3.4	22.1	53.6	NS	NS	NS
6	0.4	4.1	26.1	63.4	NS	NS	NS
7	0.5	4.7	30.1	73.1	NS	NS	NS

Number of Simultaneous Monitoring Sessions	Percentage of Available Bandwidth Required (No Silence Suppression)						
	100 Mbps	10 Mbps	1.544 Mbps	640 kbps	256 kbps	128 kbps	64 kbps
8	0.5	5.3	34.1	82.9	NS	NS	NS
9	0.6	5.9	38.1	92.6	NS	NS	NS
10	0.7	6.6	42.2	NS	NS	NS	NS

[2] The bandwidth of the connection is not large enough to support the number of simultaneous monitoring sessions.

The following notes apply to the bandwidth requirements shown in the previous tables:

- The bandwidth values are calculated based on the best speed of the indicated connections. A connection's true speed can differ from the maximum stated due to various factors.
- The bandwidth requirements are based on upload speed. Download speed affects only the incoming stream for the IP phone call.
- The values are based upon each voice packet containing 20 milliseconds of speech.
- The number of bytes in each packet include the entire Ethernet encapsulation.
- The data represents the Codecs without silence suppression. With silence suppression, the amount of bandwidth used may be lower.
- The data shown does not address the quality of the speech of the monitored call. If the bandwidth requirements approach the total bandwidth available and other applications must share access to the network, latency (packet delay) of the voice packets can affect the quality of the monitored speech. However, latency does not affect the quality of recorded speech.
- The data represents only the bandwidth required for monitoring and recording. It does not include the bandwidth requirements for Cisco Finesse.

Web Chat Feature

When deploying the Unified CCX along with Cisco SocialMiner, observe the following network requirements:

Delay—The maximum allowed round-trip time (RTT) between the Unified CCX server and SocialMiner is 150 ms.

Bandwidth—In addition to the requirements for the Unified CCX and Unified CM clusters, provision sufficient bandwidth for SocialMiner, the customer web server, and remote agent or supervisor desktops to deploy Web Chat successfully.

Consider the bandwidth required for the following components:

- **Unified CCX and SocialMiner**—If SocialMiner and the Unified CCX are not co-located, there is an additional bandwidth requirement for the communication and contact signaling.
- **SocialMiner and Cisco Finesse Agent Desktop**—When a chat session starts, depending on the chat transcript size and communication frequency, there is an additional bandwidth requirement between SocialMiner and the Cisco Finesse Agent Desktop.

- **SocialMiner and Customer Website**—The customer website transfers all new chat contact requests to SocialMiner. When a chat contact request reaches SocialMiner, an active chat session is maintained by SocialMiner to carry chat messages between SocialMiner and the browser. After the chat contact request is transferred to SocialMiner, the customer website server is no longer a part of the active chat session.

The following table shows the minimum bandwidth requirement for the Unified CCX and SocialMiner when they are not located in the same network.

Note These numbers depend on overall network efficiency.

	Between Unified CCX and SocialMiner (KBps)	Between Unified CCX and Agent Desktop (KBps)	Between SocialMiner and Agent Desktop (KBps)	Between Customer Web Server and SocialMiner (KBps)
Actual data bandwidth	3.35 [1]	4.02 [2]	12 [3]	12 [3]
Data bandwidth considering HTTP traffic and other factors	40	40	100	100

[1] Allocate network bandwidth for signal communication based on this formula:

Signaling network bandwidth (in KBps) = Number of new chat sessions per second × Number of messages per chat session × Average message size

Example

If you have a Busy Hour Call Completion (BHCC) of 2400 (maximum supported) with an average chat duration of 3 minutes, you have 0.67 chat sessions per second. On average, if each chat session has five messages for state signaling and contact injection and each message is 1 KB (500 characters), then bandwidth utilization is 0.67 × 5 ×1 KBps = 3.35 KBps.

[2] Allocate network bandwidth for signal communication based on this formula:

Signaling network bandwidth (in KBps) = Number of new chat sessions per second × Number of messages per chat session × Average message size

Example

If you have a BHCC of 2400 (maximum supported) with an average chat duration of 3 minutes, you will have 0.67 chat sessions per second. On average, if each chat session has 3 messages and each message is 2 KB (1000 characters), bandwidth utilization is 0.67 × 3 × 2 KBps = 4.02 KBps.

[3] Allocate network bandwidth for chat based on this formula:

Chat network bandwidth (in KBps) = Chat sessions sending message per second × Average message size

Example

If all of the120 sessions are active and 10 percent of the chat sessions are sending messages every second, 120 × 10 / 100 = 12 chat sessions are sending a message each second.

If the average message size is 1 KB (500 characters), the chat network bandwidth is 12 KBps.

Agent Email Feature

When you deploy Unified CCX along with Cisco SocialMiner, observe the following network requirements:

Delay—The maximum allowed round-trip time (RTT) between the Unified CCX server and SocialMiner is 150 ms.

Bandwidth—In addition to the requirements for the Unified CCX and Unified Communications Manager clusters, you must provision sufficient bandwidth for SocialMiner, the mail server, and remote agent/supervisor desktops to deploy Agent Email successfully.

The following table shows the minimum bandwidth requirement for Unified CCX and SocialMiner when they are not located in the same network.

Note These numbers depend on overall network efficiency.

	Between Unified CCX and SocialMiner (KBps)
Actual data bandwidth	0.67 [1]
Data bandwidth considering HTTP traffic and other factors	40

[1] Allocate network bandwidth for signal communication based on this formula:

Signaling network bandwidth (in KBps) = Number of new email sessions per second × Number of messages per email session × Average message size

Example

If you have 400 emails (maximum supported) per hour, you will have 0.11 email sessions per second. On average, if each email session has six messages for state signaling and contact injection and each message is 1 KB (500 characters), then bandwidth utilization is 0.11 x 6 x 1 KB = 0.67 KBps.

	Between Unified CCX and Agent Desktop (KBps)
Actual data bandwidth	2.22 [2]
Data bandwidth considering HTTP traffic and other factors	40

[2] Allocate network bandwidth for signal communication based on this formula:

Signaling network bandwidth (in KBps) = Number of new email sessions per second × Number of messages per email session × Average message size

Example

If you have 400 emails (maximum supported) per hour and an agent can handle five concurrent emails, you will have 0.11 emails per second. The agent can requeue or respond to that email directly. Assuming on average 10% of email messages are requeued and there are 100 Email CSQs in the system, three messages, each 1 KB, and the requeue list message is 10 KB, the bandwidth requirement is calculated as follows:

network bandwidth (in KBps) = number of concurrent emails x number of new email sessions per second x [(number of messages per email session x average message size) + (percentage of emails requeued x size of requeue list message)]

5 x 0.11 x ((3 x 1 KB) + (0.1 x 10 KB)) = 2.22 KBps

Agent Email Flow

There are four types of Agent Email flows that exist between the Agent Desktop, SocialMiner, and the Exchange Server.

- Basic Email flow—No attachments and no requeue.
- Email with attachments flow—The customer's email contains attachments and the agent's reply has attachments.
- Email requeue flow—The customer's email is sent to another queue.
- Email requeue with attachments flow—The customer's email contains attachments. The email is requeued and the agent's reply contains attachments.

The flows listed above represent the extreme cases which makes the calculations conservative.

Requeues and attachments are expected to occur 10% of the time. The maximum number of emails per hour is 400. The occurrence of each type of flow is determined by first calculating the number of basic and requeue flows followed by the number of basic and requeue flows that contain attachments:

- Total basic email flow = Maximum email per hour – [maximum email per hour x (requeue percent / 100)]
  - Email with attachments flow = total basic email flow x (attachment percent / 100)
  - Basic email flow = total basic email flow – email with attachments flow
- Total email requeue flow = Maximum email per hour x (requeue percent / 100)
  - Email requeue with attachments flow = total email requeue flow x (attachment percent / 100)
  - Email requeue flow = total email requeue flow – email requeue with attachments flow

After considering the maximum values, the calculation is:

- Total basic email flow = 360
  - Email with attachments flow = 36
  - Basic email flow = 324
- Total email requeue flow = 40
  - Email requeue with attachments flow = 4
  - Email requeue flow = 36

Each of the flows has a set of messages with different bandwidth requirements. The requirements are derived based on the following constants:

- Customer email size = 6 KB
- Attachment size = 5120 KB
- Agent reply size = 6 KB
- SLA = 60 minutes
- Save draft interval = 3 minutes

Agent Email Routing Configuration

Emails are fetched on every polling interval configured by the admin. A snapshot age (in minutes) is defined by the admin. Based on this configuration, SocialMiner determines the oldest email to be fetched and emails are fetched, in the oldest first order from the mail server and then injected to the Unified CCX engine. Later, the Unified CCX engine presents the emails to the agent.

In the event of any disruption (SocialMiner or CCX Engine or network connectivity between CCX and SocialMiner goes down and comes back), SocialMiner re-injects the latest 200 email contacts to Unified CCX.

Agent Desktop and SocialMiner

If SocialMiner and Unified CCX are not co-located, additional bandwidth is required for communication and contact signaling.

	Between Agent Desktop and SocialMiner
Actual data bandwidth	3560160 KB per hour
Data bandwidth considering HTTP traffic and other factors	1024 KBps

Example

Using an email size of 6 KB and an agent reply of 6 KB, the bandwidth requirements for the set of messages between the Agent Desktop and SocialMiner for each flow is as follows:

- Basic email flow = 6288 KB
  - Load UI files into Finesse = 6144 KB
  - Signaling = 6 KB
  - Get email body = 6 KB
  - Save draft = agent reply size x (SLA / save draft interval) = 120 KB
  - send reply = customer email size + agent reply size = 12 KB
- Email with attachments flow = 31888 KB
  - Load UI files into Finesse = 6144 KB
  - Signaling = 6 KB
  - Get email body with attachments = customer email size + attachment size = 5126 KB
  - Save draft = agent reply size x (SLA / save draft interval) = 120 KB

    - Customer attachments download = attachment size = 5120 KB
    - Agent attachments upload = attachment size = 5120 KB
    - Agent attachments download = attachment size = 5120 KB
    - Send reply with attachments = customer email size + agent reply size + attachment size = 5132 KB

- Email requeue flow = 6300 KB
    - Load UI files into Finesse = 6144 KB
    - Signaling (get contact + reserve contact) = 4 KB
    - Get email body = 6 KB
    - Requeue = 2 KB
    - Signaling (requeue + get contact + reserve contact) = 6 KB
    - Get email body = 6 KB
    - Save draft = agent reply size x (SLA / save draft interval) = 120 KB
    - Send reply = customer email size + agent reply size = 12 KB
- Email requeue with attachments flow = 37020 KB
    - Load UI files into Finesse = 6144 KB
    - Signaling (get contact + reserve contact) = 6 KB
    - Get email body with attachments = customer email size + attachment size = 5126 KB
    - Signaling (requeue + get contact + reserve contact) = 6 KB
    - Get email body with attachments = customer email size + attachment size = 5126 KB
    - Save draft = agent reply size x (SLA / save draft interval) = 120 KB
    - Customer attachments download = attachment size = 5120 KB
    - Agent attachments upload = attachment size = 5120 KB
    - Agent attachments download = attachment size = 5120 KB
    - Send reply with attachments = (customer email size + agent reply size + attachment size) = 5132 KB

At 400 emails per hour, the bandwidth for each flow is as follows:

- Basic email flow = 6288 KB x 324 = 2037312 KB
- Email with attachments flow = 31888 KB x 36 = 1147968 KB
- Email requeue flow = 6300 KB x 36 = 226800 KB
- Email requeue with attachments flow = 37020 KB x 4 = 148080 KB

The total bandwidth is 3560160 KB per hour. The bandwidth in KBps between the Agent Desktop and SocialMiner is 1024 KBps.

SocialMiner and Mail Server

SocialMiner must retrieve emails, save drafts, and send emails from the Agent Desktop to the mail server. If the mail server is not on the same network as SocialMiner, you must ensure that the bandwidth requirement is based on mean per-session email traffic.

	Between SocialMiner and Mail Server (KBps)
Actual data bandwidth	1516720 KB per hour
Data bandwidth considering HTTP traffic and other factors	512 KBps

Example

Using an email size of 6 KB and an agent reply of 6 KB, the bandwidth requirements for the set of messages between the SocialMiner and mail server for each flow is as follows:

- Basic email flow = 156 KB
  - Initial fetch of customer email = 6 KB
  - Get email body = 6 KB
  - Save draft = agent reply size x (SLA / save draft interval) = 120 KB
  - send = 2 x (customer email size + agent reply size) = 24 KB
- Email with attachments flow = 35996 KB
  - Initial fetch of customer email with attachments = customer email size + attachment size = 5126 KB
  - Get email body with attachments = customer email size + attachment size = 5126 KB
  - Save draft = agent reply size x (SLA / save draft interval) = 120 KB
  - Customer attachments download = attachment size = 5120 KB
  - Agent attachments upload = attachment size = 5120 KB
  - Agent attachments download = attachment size = 5120 KB
  - Send reply with attachments = 2 x (customer email size + agent reply size + attachment size) = 10264 KB
- Email requeue flow = 162 KB
  - Initial fetch of customer email = 6 KB
  - Get email body = 6 KB
  - Get email body after requeue = 6 KB
  - Save draft = agent reply size x (SLA / save draft interval) = 120 KB

    - Send reply = 2 x (customer email size + agent reply size) = 24 KB

- Email requeue with attachments flow = 41122 KB
    - Initial fetch of customer email with attachments = customer email size + attachment size = 5126 KB
    - Get email body with attachments = customer email size + attachment size = 5126 KB
    - Get email body with attachments afer requeue = customer email size + attachment size = 5126 KB
    - Save draft = agent reply size x (SLA / save draft interval) = 120 KB
    - Customer attachments download = attachment size = 5120 KB
    - Agent attachments upload = attachment size = 5120 KB
    - Agent attachments download = attachment size = 5120 KB
    - Send reply with attachments = 2 x (customer email size + agent reply size + attachment size) = 10264 KB

At 400 emails per hour, the bandwidth for each flow is as follows:

- Basic email flow = 156 KB x 324 = 50544
- Email with attachments flow = 35996 KB x 36 = 1295856
- Email requeue flow = 162 KB x 36 = 5832 KB
- Email requeue with attachments flow = 41122 KB x 4 = 164488 KB

The total bandwidth is 1516720 KB per hour. The bandwidth in KBps between SocialMiner and the mail server is 512 KBps.

- Ensure maximum RTT between SocialMiner and Office 365 SMTP and IMAP hosts are within 100 ms (including network traversals via SOCKS5 proxy if applicable).
- Follow best practices for provisioning Office 365 accounts geographically based on recommendations from Microsoft support.

Context Service Performance Considerations

In general, typical Context Service requests within the same geographical area take from 100ms to 300ms and across the globe can take as long as 1.2 to 1.5 seconds.

QoS and Call Admission Control

Quality of service (QoS) becomes an issue when more voice and application-related traffic is added to an already growing amount of data traffic on your network. Accordingly, Unified CCX and time-sensitive traffic such as voice need higher QoS guarantees than less time-sensitive traffic such as file transfers or emails (particularly if you are using a converged network).

QoS should be used to assign different qualities to data streams to preserve Unified CCX mission-critical and voice traffic. The following are some examples of available QoS mechanisms:

- Packet classification and usage policies applied at the edge of the network, such as Policy Based Routing (PBR) and Committed Access Rate (CAR).
- End-to-end queuing mechanisms, such as Low Latency Queuing (LLQ). Because voice is susceptible to increased latency and jitter on low-speed links, Link Fragmentation and Interleaving (LFI) can also be used to reduce delay and jitter by subdividing large datagrams and interleaving low-delay traffic with the resulting smaller packets.
- Scheduling mechanisms such as Traffic Shaping to optimize bandwidth utilization on output links.

Unified CCX and Application-Related Traffic

The table lists TCP ports and DSCP markings for use in prioritizing Unified CCX and Unified CM mission-critical CTI traffic. The DSCP markings for call signaling traffic between Unified CCX and Cisco Unified Communication Manager and for voice traffic played from the Unified CCX server are set by default according to the traffic classification guidelines documented in *Cisco Unified Communications System Design Guidance,* available here:

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

Unified CCX does not mark any network traffic other than those mentioned here. As a result, traffic should be marked and prioritized at the edge according to the values in the table.

The performance criteria used in classifying this traffic includes:

- No packet drops on the outbound or inbound interface of the WAN edge router
- Voice (G.729) loss under 1percent
- One-way voice delay under 150 ms

A detailed description of QoS is not within the scope of this design guide. For QoS design considerations, refer to the quality of service design guide available here:

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

Table 5: QoS Classifications for Unified CCX Interfaces

Unified CCX Component	Interface / Protocol	Port	DSCP Marking
Unified CCX Engine — CTI-QBE messaging destined to Unified CM from Unified CCX	CTI-QBE	TCP 2748	CS3
Unified CCX Administration and BIPPA Service — HTTP traffic destined for web administration and BIPPA interface on Unified CCX	HTTP / HTTPS	TCP 8443	AF21
Unified CCX Engine and Unified CCX Administration — SOAP AXL HTTPS messaging destined to Unified CM from Unified CCX	HTTPS / SOAP	TCP 8443	AF21
Unified CCX Engine — CTI messaging destined to Unified CCX from CAD clients	CTI	TCP 12028	CS3
Unified CCX Engine — RTP voice bearer traffic (bi-directional)	RTP	UDP 16384 - 32767	EF

CAC and RSVP

Unified CM supports Resource-Reservation Protocol (RSVP) between endpoints within a cluster. RSVP is a protocol used for Call Admission Control (CAC) and is used by the routers in the network to reserve bandwidth for calls. The bandwidth being controlled is only for the voice streams, call signalling traffic is not part of CAC.

Before RSVP, each Unified CM cluster maintained its own calculation of how many active calls were traversing between locations in order to calculate bandwidth usage. If more than one Unified CM cluster shared the same link, bandwidth would have to be carved out and dedicated for each cluster, and this led to inefficient use of available bandwidth. RSVP also enables customers to deploy complex network topology while location-based CAC is limited to a hub-and-spoke type of topology.

RSVP solves this problem by tracing the path between two RSVP Agents that reside on the same LAN as the IP Phones. A software MTP or transcoder resource that runs on Cisco IOS routers can be RSVP Agents. The RSVP Agents are controlled by Unified CM and are inserted into the media stream between the two IP phones when a call is made. The RSVP Agent of the originating IP Phone will traverse the network to the destination IP Phone's RSVP Agent, and reserve bandwidth. Since the network routers (and not Unified CM) are keeping track of bandwidth usage, multiple phone calls can traverse the same RSVP controlled link even if the calls are controlled by multiple Unified CMs.

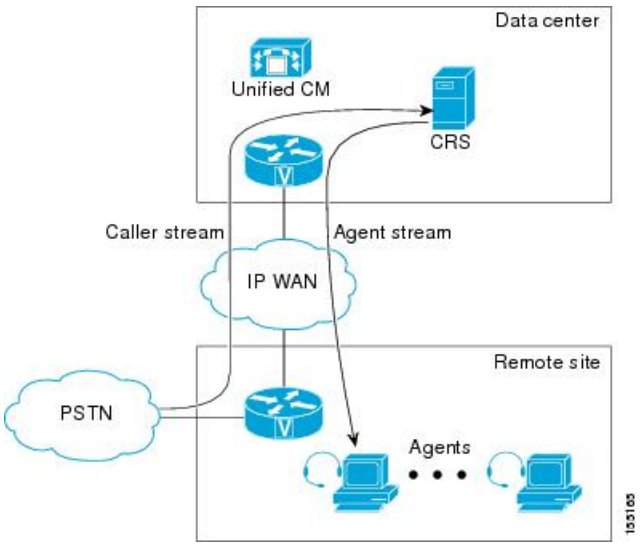
For more information, see the RSVP chapter in *Cisco Unified Communications Solution Reference Network Design (SRND)*.

Unified CCX selects a call center agent independent of the mechanism, using either RSVP or location-based CAC. Unified CCX routes a call to an available agent even though the agent phone might not be able to receive the call due to lack of bandwidth. Proper sizing of bandwidth between sites is very important.

For any call transfer, there are moments when two calls are active. If any of the active calls traverses between sites, then CAC is used. Even when the original call is placed on hold during a transfer, that call still takes up the same amount of bandwidth just like an active call.

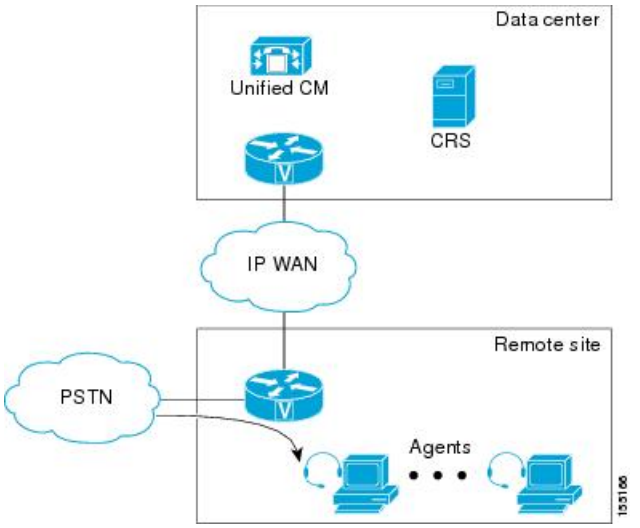
In the two examples that follow, the voice gateway and agents are at a remote site, while the Unified CCX server is at another site. A call from PSTN reaches the voice gateway at the remote site and connects to Unified CCX at the site. This takes one call bandwidth over the WAN link, which is represented by the caller stream. Once an agent is available and selected at the remote site, Unified CCX transfers the call to the agent.

Figure 1: Call From PSTN to Unified CCX Server to Agent



During the transfer, before the agent picks up the call, there is another call setup between Unified CCX and the agent phone. It takes up another call bandwidth over the WAN, and is represented by the agent stream in the previous example. Once the agent picks up the call, the voice traffic is between the voice gateway and the agent phone, which are both at the remote site. At that time, no bandwidth is reserved over the WAN, as illustrated in the following example. This example shows how call bandwidth is reserved in a contact center call that is eventually routed to an agent. Depending on where the voice gateway, the agents, and the Unified CCX server are located, proper WAN bandwidth should be provisioned.

Figure 2: After Agent Picks Up Call



Bandwidth, Latency, and QoS for Cisco Finesse

Bandwidth requirement for Cisco Finesse client to server

The agent and supervisor login operation involves loading web pages, and includes the CTI login and the display of the initial agent state. After the desktop web page loads, the required bandwidth is significantly less.

As Cisco Finesse is a web application, caching can significantly impact the required bandwidth. To minimize the amount of bandwidth required for login, make sure that caching is enabled in the browser.

To help you with the bandwidth calculation, Cisco Finesse provides a bandwidth calculator (Cisco Unified CCX Bandwidth Calculator) to estimate the bandwidth required to accommodate the client login time.

The bandwidth calculator does not include the bandwidth required for any third-party gadgets in the Cisco Finesse container or any other applications running on the agent desktop client.

The bandwidth listed in the bandwidth calculator must be available for Cisco Finesse after you account for the bandwidth used by other applications, including voice traffic that may share this bandwidth. The performance of the Cisco Finesse interface, and potentially the quality of voice sharing this bandwidth, may degrade if sufficient bandwidth is not continuously available.

Cisco Finesse Desktop Latency

Cisco Finesse Agent and Supervisor Desktops can be located remotely from Unified CCX. The round-trip time between the Unified CCX server and the agent desktop must not exceed 400 ms.

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, Latency, and QoS for Unified Intelligence Center

The two bandwidth considerations in a Unified Intelligence Center installation include the following:

- Bandwidth between the Unified Intelligence Center and data source
- Bandwidth between the user and Unified Intelligence Center

The Unified CCX database is local to the server. In a normal operating mode, the bandwidth between Unified Intelligence Center and the data source can be ignored.

Note Each report requires about 2.6 Mbps of bandwidth between the user and Unified Intelligence Center.

The configuration parameters that affect bandwidth include the following:

- Size of each row: 500 bytes
- HTML size overhead for each row: 500 bytes
- Time to transfer the rendered report from Unified Intelligence Center to the browser: 3 seconds

Reporting Scaling Considerations

Following are the reporting considerations:

- A maximum of eight reporting users logged in concurrently on Cisco Unified Intelligence Center can view:
  - Four Live Data reports with 50 rows of 10 fields refreshing every 3 seconds.
  - Two historical reports with 2000 rows with 10 fields each refreshing every 30 minutes.
- A maximum of 42 Finesse supervisors can view:
  - Three Live Data reports with 50 rows of 10 fields refreshing every 3 seconds.
- A maximum of 358 Finesse agents can view:
  - Three real-time reports with 20 rows of 10 fields refreshing every 3 seconds.

Bandwidth, Latency, and QoS for Optional Cisco Components

Bandwidth, Latency, and QoS for Cisco MediaSense

MediaSense requires gigabit LAN connectivity with 2 ms or less latency between servers in a cluster.

Bandwidth, Latency, and QoS for Optional Third-Party Components

None