Sizing and Operating Conditions for Reference Designs

Sizing for Reference Design Solutions

A contact center enterprise solution requires proper sizing of its resources. This chapter discusses the tools and methods for sizing those resources. This includes resources like:

- The required number of contact center agents (based on customer requirements such as call volume and service level desired)
- The number of VRU ports required for various call scenarios (such as call treatment, prompt and collect, queuing, and self-service applications)
- The number of Voice Gateway ports required to carry the incoming and outbound traffic volume

Proper sizing uses the traffic engineering principles encapsulated in the Erlang-B and Erlang-C models.

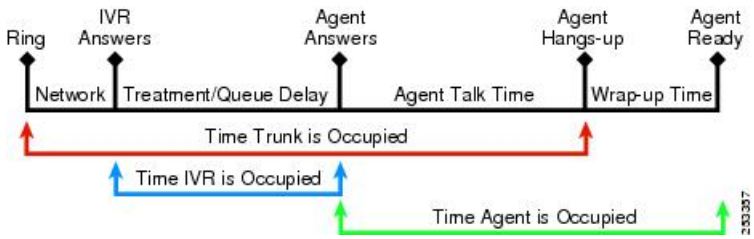
Resource Use During a Contact

When you size your contact center enterprise solution, the main resources that you look at are:

- Agents
- Gateway ports (PSTN trunks)
- VRU ports

To determine required resources, first look at this timeline of a typical inbound call and the resources that it requires at each step. This figure shows the main resources and the occupancy (hold and handle time) for those resources.

Figure 1: Inbound Call Timeline



If calls are not answered immediately, include ring delay time (network ring) in your call timeline. An average ring delay time is a few seconds. Add it to the trunk average handle time for your calculations.

Contact Center Traffic Terminology

These are the most common industry terms for sizing contact center resources.

Busy Hour or Busy Interval

A busy interval is 1 hour or less. The busy interval is when the most traffic occurs during a period of the day. The busy hour or interval varies due to circumstances like weekends and seasonal effects. Design for the average busy hour (the average of the 10 busiest hours in one year). This average is not always applied, however, when staffing is required to accommodate a marketing campaign or a seasonal busy hour such as an annual holiday peak. In a contact center, you staff for the maximum number of agents based on peak periods. But, you calculate the requirements for the rest of the day separately for each period (usually every hour). This gives proper scheduling of agents to answer calls versus scheduling agents for offline activities such as training or coaching. For trunks or VRU ports, it is not practical to add or remove trunks or ports daily, so these resources are sized for the peak periods. In some retail environments, extra trunks can be added during the peak season and disconnected afterwards.

Busy Hour Call Attempts (BHCA)

The BHCA is the total number of calls during the peak traffic hour (or interval) that are attempted or received in the contact center. For the sake of simplicity, we assume that the contact center resources (agents and VRU ports) receive and service all calls offered to the Voice Gateway. Calls usually originate from the PSTN, although calls to a contact center can also be generated internally, such as by a help-desk application.

Calls Per Second as reported by Call Router (CPS)

This is the rate at which the Unified CCE Router receives call routing requests. Every call generates one call routing request in a simple call flow from ingress gateway to VRU treatment to routing to an agent. However, some calls need more than one routing request to be made to the Router to finally get to the right agent.

An example of this is when the first agent who receives the call wants to transfer or conference to another agent by using a post route. This generates an extra routing request resulting in the same call generating two routing requests to the Router. A routing request is made to the Router whenever a resource is required for a call or task. These requests also include multimedia requests for Email, Chat, Callback and certain Outbound Calls. Call center administrators must account for these additional call routing requests when they size their contact center.

The maximum supported call rate is the call rate reported by the Router and not the BHCA at the ingress gateway. Factor these additional routing requests into the calculation of BHCA at the ingress gateway.

In general, the BHCA at the ingress gateway is lower than or equal to the corresponding CPS rate reported by the Router.

For example, consider the following situation. If the BHCA at the ingress gateway is 36,000, then the call rate at the ingress gateway is 10 CPS. If we assume that 10% of the calls are transferred through the Router, the CPS reported by Router is equal to 11 CPS. In this case, your solution needs a capacity of 11 CPS.

Servers

Servers are resources that handle traffic loads or calls. There are many types of servers in a contact center. Each type can require different resources.

Talk Time

Talk time is the amount of time an agent spends talking to a caller. This includes any time an agent places a caller on hold and any time spent during consultative conferences.

Wrap-Up Time (After-Call Work Time)

After the call terminates (the caller disconnects the call), and agent completes certain tasks to "wrap up" the call. The wrap-up time includes such tasks as updating a database, recording notes from the call, or any other activity performed until an agent becomes available to answer another call. Unified Contact Center Enterprise solutions sometimes call this period the *after-call work time*.

Average Handle Time (AHT)

AHT is the mean (or average) call duration during a specified time period. It refers to the sum of several types of handling time, such as call treatment time for self-service calls or talk time for calls to agents. In its most common definition, AHT is the sum of agent talk time and agent wrap-up time.

Erlang

Erlang is a measurement of traffic load during the busy hour. The Erlang is based on having 3600 seconds of calls on the same circuit, trunk, or port. (One circuit is busy for 1 hour regardless of the number of calls or how long the average call lasts.) The formula to calculate the Erlang value is:

```
Traffic in Erlangs = (Number of calls in the busy hour * AHT in sec) / 3600 sec
```

If a contact center receives 30 calls of 6-minute length in the busy hour, this equates to 180 minutes of traffic in the busy hour, or 3 Erlangs. If the contact center receives 100 calls averaging 36 seconds each in the busy hour, then total traffic received is 3600 seconds, or 1 Erlang (3600 sec/3600 sec).

Busy Hour Traffic (BHT) in Erlangs

BHT is the traffic load during the busy hour and is calculated as the product of the BHCA and the AHT normalized to 1 hour:

```
BHT = (BHCA * AHT seconds) / 3600
```

For example, if the contact center receives 600 calls in the busy hour, averaging 2 minutes each, then the busy hour traffic load is (600 * 2/60) = 20 Erlangs.

BHT is typically used in Erlang-B models to calculate resources such as PSTN trunks or self-service VRU ports.

Grade of Service (Percent Blockage)

This measurement is the probability that a resource or server is busy during the busy hour. In that case, the call is lost or blocked. This blockage typically applies to resources such as Voice Gateway ports, VRU ports, PBX lines, and trunks. In the case of a Voice Gateway, grade of service is the percentage of calls that are blocked or that receive busy tone (no trunks available) out of the total BHCA. For example,

a grade of service of 0.01 means that 1% of calls in the busy hour is blocked. A 1% blockage is a typical value to use for PSTN trunks, but different applications might require different grades of service.

Blocked Calls

A blocked call is a call that is not serviced immediately. Caller are blocked if they are rerouted to another route or trunk group, if they are delayed and put in a queue, or if they hear a tone (such as a busy tone) or announcement. The nature of the blocked call determines the model used for sizing the particular resources.

Service Level

The industry standard term for the percentage of the offered call volume (received from the Voice Gateway and other sources) that are answered within *X* seconds. A typical value for a sales contact center is 90% of all calls answered in less than 10 seconds (some calls are delayed in a queue). A support-oriented contact center might have a different service level goal, such as 80% of all calls answered within 30 seconds in the busy hour. Your service level goal determines the necessary agents, the percentage of queued calls, the average time calls spend in queue, and the necessary PSTN trunks and VRU ports.

Queuing

When agents are busy with other callers or are unavailable (after call wrap-up mode), subsequent callers must be placed in a queue until an agent becomes available. Your desired service level and agent staffing determines the percentage of calls queued and the average time spent in the queue. Contact center enterprise solutions use a VRU to place callers in queue and play announcements. The VRU initially handles all calls. It supplies call treatment and prompts for necessary information. The VRU handles self-service applications where the caller is serviced without needing to talk to an agent. Each of these scenarios requires a different number of VRU ports because each has a different average handle time and possibly a different call load. The number of trunks or gateway ports needed for each of these applications differs accordingly.

Erlang Calculators as Design Tools

Many traffic models are available for sizing telephony systems and resources. Choosing the right model depends on three main factors:

- Traffic source characteristics (finite or infinite)
- How lost calls are handled (cleared, held, delayed)
- Call arrival patterns (random, smooth, peaked)

For contact center enterprise solutions, you commonly use the Erlang-B and Erlang-C traffic models for sizing resources.

Erlang calculators help answer the following questions:

- How many trunks do I need?
- How many agents do I need?
- How many VRU ports do I need?

You need these figures for input to Erlang calculators:

- The busy hour call attempts (BHCA)
- Average handle time (AHT) for each of the resources

- Service level (percentage of calls that are answered within x seconds)
- Grade of service, or percent blockage, desired for trunks and VRU ports

The next sections present a brief description of the generic Erlang models in simple terms. They also describe the input and output of the Erlang models and which model to use for sizing particular resources. There are a variety of contact center sizing tools available. They all use the two basic traffic models, Erlang-B and Erlang-C.

Erlang-B Uses

Use the Erlang-B model to size PSTN trunks, gateway ports, or VRU ports. It assumes the following:

- Call arrival is random.
- If all trunks or ports are occupied, new calls are lost or blocked (receive busy tone) and not queued.

The input and output for the Erlang B model consists of the following three factors. If you have any two of these factors, the model calculates the third:

- Busy Hour Traffic (BHT). BHT is the product of the number of calls in the busy hour (BHCA) and the average handle time (AHT).
- Grade of Service
- Ports (lines)

Erlang-C Uses

Use the Erlang-C model to size agents in contact centers that queue calls before presenting them to agents. This model assumes:

- Call arrival is random.
- If all agents are busy, incoming calls are queued and not blocked.

The input parameters required for this model are:

- The number of calls that angents answer in the busy hour (BHCA)
- The average talk time and wrap-up time
- The delay or service level desired, expressed as the percentage of calls answered within a specified number of seconds

The output of this model gives the required number of agents, the percentage of calls delayed when agents are unavailable, and the average queue time.

Dynamic Configuration Limits for Unified CCE

Sometimes, you can exceed the standard limits for one resource by significantly reducing use of another resource. Test the specific trade-off that you plan to make before you incorporate it in your solution. The following sections provide guidance on how to balance certain resources.

Dynamic Limits for Skill Groups and Precision Queues Per Agent

The number of skill groups and precision queues per agent significantly affects the following subcomponents of Unified CCE:

- Cisco Finesse servers
- Agent PGs
- Router
- Logger

Note We use *queue* as a common term for skill groups and precision queues.

To maintain the performance of your solution, periodically remove unused queues.

The Reference Designs set a standard limit for the average queues per agent on each PG. On a particular PG, some agents can have more queues than other agents. As long as the average across all the agents on the PG is within the limit, you can still have the maximum active agents on that PG.

For example, assume that you have three groups of agents on a PG in a 4000 Agent Reference Design:

- Group A has 500 agents with five queues each.
- Group B has 1000 agents with 15 queues each.
- Group C has 500 agents with 25 queues each.

These three groups average to 15 queues per agent, so you can have them all on a single PG under the standard limits.

You can also exceed that standard limit if you reduce the number of agents on each PG and on the whole system.

Note See the configuration tables in the configuration limits chapter for the standard limits.

The Cisco Finesse server doesn't display statistics for unused queues. So, the active queues affect the performance of the Cisco Finesse server more than the total configured queues.

The Cisco Finesse desktop updates queue (skill group) statistics at 10-second intervals. The Cisco Finesse Desktop also supports a fixed number of queue statistics fields. You can't change these fields.

This table shows the approximate reduction in the number of agents your solution can support with more queues per agent:

Table 1: Dynamic Agents and Queues Limits

Average Queues per Agent	Maximum Agents per PG	Maximum Agents for 2000 Agent Reference Design	Maximum Agents for 4000 Agent Reference Design[1]	Maximum Agents for 12000 Agent Reference Design
15	2000	2000	4000	12000

Average Queues per Agent	Maximum Agents per PG	Maximum Agents for 2000 Agent Reference Design	Maximum Agents for 4000 Agent Reference Design[1]	Maximum Agents for 12000 Agent Reference Design
20	1500	1500	3000	9000
30	1000	1000	2000	6000
40	750	750	1500	4500
50	600	600	1200	3600

[1] You can't have more than 4000 Agents on a Rogger deployment.

Unified CCE supports a maximum of 50 unique skill groups across all agents on a supervisor's team, including the supervisor's own skill groups. If this number is exceeded, all skill groups that are monitored by the supervisor still appear on the supervisor desktop. However, exceeding this number can cause performance issues and isn't supported.

Note Each precision queue that you configure creates a skill group for each Agent PG and counts toward the supported number of skill groups per PG. The skill groups are created in the same Media Routing Domain as the precision queue.

Other Dynamic Sizing Factors

Many factors can affect your solution's server requirements and capacities. These sections call out the major sizing variables and how they affect your solution.

Busy Hour Call Attempts (BHCA)

As BHCA increases, the load on all your solution components increases, most notably on Unified CM, Unified CVP, and the Agent PG. The capacity numbers for agents assume up to 30 calls per hour per agent. If your solution requires a greater BHCA, decrease the maximum agents on the Agent PG.

Unified CM Silent Monitor

Each silently monitored call adds more processing for the PG and Unified CM. Each silently monitored call equals to two unmonitored agent calls. Leave room on your PGs to account for the percentage of the monitored calls.

Script Complexity

As the complexity and number of Unified CCE scripts increase, the processor and memory overhead on the Call Router and VRU PG increases significantly. As VRU script complexity increases with features such as database queries, the load placed on CVP and the Router also increases. The delay time between replaying RunExternalScript also has an impact.

The performance of complex scripts and database queries is hard to characterize. Test complex scripting in a lab to determine the response time of database queries under various BHCA. Adjust your sizing to account for their effects on the processor and memory of the Voice Browser, Unified CVP, the PGs, and the Router.

Third-Party Database and Cisco Resource Manager Connectivity

Carefully examine the connectivity of any Unified CCE solution component to external devices and software to determine the overall effect on the solution. Contact center enterprise solutions are flexible and customizable, but they are also complex. Contact centers are often mission-critical, revenue-generating, and customer-facing operations. Engage a Cisco Partner (or Cisco Advanced Services) with the appropriate experience and certifications to help you design your solution.

Expanded Call Context (ECC)

Your solution's use of ECC variables impacts the PGs, Router, Logger, and network bandwidth. You can configure and use ECC variables in many ways. The capacity impact varies based on the ECC configuration.

PG Agent Capacity with Mobile Agents

Mobile agent support capacity on the medium PG OVA are as follows:

- 2000 with nailed-up connections (1:1)
- 1500 with nailed-up connections if the average handle time is less than 3 minutes, or if agent greeting or whisper announcement features are used with the mobile agent (1.3:1)
- 1500 with call-by-call connections (1.3:1)

You can have a mix of mobile Agents and other agents on the same PG. Keep the respective weights of each type of agent in mind. For example, if you have 200 mobile agents with nailed-up connections, the PG can support 1800 other agents:

```
Additional Agents Allowed = (2000 – (200 * 1)) = 1800 Agents
```

If you plan to use 200 active mobile agents with call-by-call connections, the PG can support 1740 other agents:

```
Additional Agents Allowed = (2000 – (200 * 1.3)) = 1740 Agents
```

Configuration Limits for Reference Design Solutions

Sizing for Unified CVP

When you size your contact center, determine the worst-case profile for the number of calls that are in each state. At its busiest instant in the busiest hour, how many calls do you find in the following states:

- **Self-service**—Calls that are executing applications using the VXML Server.
- **Queue and collect**—Calls that are in queue for an agent or are executing prompt-and-collect self-service applications.
- **Talking**—Calls that are connected to agents or to third-party TDM VRU applications.

Note The definitions of these call states differ from the definitions used for port licensing purposes. You can ignore ASR and TTS processing when counting which calls are in which states for sizing purposes. However, ASR and TTS processing does come into call state counts for licensing.

Size the solution for the number of ports in use for calls in a talking state to agents. Even though you do not need licenses for those ports when using Unified CCE agents, TDM agents do require a Call Director license.

For calls in the talking state, count only calls that use Unified CVP or gateway resources. If the transfer uses VoIP, the call uses a Voice Browser port and Unified CVP resources. Unified CVP continues to monitor the call and enables you to retrieve it and redeliver it later. Unified CVP also continues to monitor calls to a TDM target. Those calls use both an incoming and an outgoing TDM port on the same gateway or on a different gateway (that is, toll bypass). Both of these types of calls count as talking calls.

However, if a transfer uses *8 TNT, hookflash, Two B Channel Transfer (TBCT), or an ICM NIC, the gateway and Unified CVP do not play a role. Both components reclaim their resources. Such calls do not count as talking calls.

Include in the overall call counts those calls that are transferred back to Unified CVP for queuing or self-service. For example, in a warm transfer, Unified CVP queues the agent during the post-route phase. The call uses two ports for the two separate call control sessions at Unified CVP. Transfers are usually a small part of the overall call volume, and you can easily overlook them.

In addition to the overall snapshot profile, also consider the CPS for the busiest period of call arrival. You need this information for the contact center enterprise solution because it is difficult to identify the exact maximum arrival rate. You can use statistical means to arrive at this number.

You size Unified CVP Servers for the number of handled calls and the maximum call arrival rate.

Table 2: CVP Server Call Rate for Call Flows

Call flow	Simultaneous Calls Supported	Calls Per Second
Comprehensive Call Flow with Secure / Non Secure SIP	3000	15
Comprehensive Call Flow with Secure / Non Secure SIP, Secure HTTP	2500	15
Standalone with / without Request ICM Label	3000	15
Standalone with / without Request ICM Label, Secure HTTP	2000	15

Note

- For the **Call Per Second**, this is the maximum call rate that is received at the CVP Call Server from all Ingress Gateways in the solution and assumes the worst case scenario where WAAG is enabled.
- It is always recommended to have enough VM resources for the garbage collection on CVP Servers to run adequately (VM memory usage should not go beyond 80%). If there are not enough system resources, garbage collection may take more time which can cause issues with the Call Server or VXML Server services.

CVP Call Server Sizing

The solution needs the greater number of Call Servers given by these equations:

```
(Self Service) + (Queue and Collect) + (Talking) / Simultaneous Calls Supported, rounded
up
OR
(Average call arrival rate) / Calls Per Second, rounded up.
```

Also, distribute the calls to the Unified CM cluster among the subscribers in the cluster. Do not exceed 2 CPS per subscriber.

See the table *CVP Server Call Rate for Call Flows* in section *Sizing for Unified CVP* for more details.

Log Directory Size Estimate

Use the following formula to calculate the estimated space per day (in gigabytes) for the Call Server Directory log file:

```
3.5 GB * R
```

Where *R* equals the number of calls per second.

For proper serviceability, reserve enough space to retain five to seven days of log messages.

CVP VXML Server Sizing

One VXML Server can handle calls as mentioned in the table "*CVP Server Call Rate for Call Flows* in the section "*Sizing for Unified CVP*". If you are using VXML Servers, size them according to the following formula:

```
Calls / Simultaneous Calls Supported, rounded up
```

Calls are the number of calls that are in VXML Server self-service applications at that snapshot in time.

Note For UCS performance numbers, see the *Virtualization for Cisco Unified Customer Voice Portal* page.

With an appropriate Cisco IOS release, you can configure Unified CVP to use HTTPS on the VXML Server and on the Unified CVP IVR Service.

Note Mainline Cisco IOS is not supported.

Performance of the CVP VXML Server varies with the complexity of your VXML application.

CUSP Performance Benchmarks

When your solution uses CUSP, remember the following points:

- CUSP baseline tests were done in isolation on the proxy. The capacity numbers from those tests are 450 TCP or 500 UDP transactions/second. Consider these figures to be the most stressed condition allowable.
- A CVP call from the proxy server requires, on average, four separate SIP calls: a caller inbound leg, a VXML outbound leg, a ringtone outbound leg, and an agent outbound leg.
- When a consultation with CVP queuing occurs, the session incurs four more SIP transactions, effectively doubling the number of calls.

Sizing Gateways for Contact Center Enterprise Solutions

Call capacities on Cisco gateways vary depending on whether they are doing ingress only, VXML only, or a combination of the two. Capacities on Voice Browsers also vary depending on factors like ASR/TTS services and type of VXML application. For instance, an intensive JavaScript application reduces call capacity.

In general, you can size gateways that perform ingress only to the maximum number of TDM cable attachment points.

Before sizing the voice gateways, use the Unified CCE Resource Calculator to determine the maximum number of trunks (DS0s) and VXML VRU ports to support your solution.

The following table provides sizing information for different versions of Cisco IOS. The sizing information is based on these factors:

- The overall CPU usage does not exceed 75 percent.
- Sizing represents the maximum number of concurrent VXML sessions and VoIP calls on the gateway.
- Sizing is based on Unified CVP VXML documents.
- Sizing includes active conferences and active transfers.
- For the VXML Only columns, sizing includes only basic routing and IP connectivity running on the gateway. If you intend to run extra applications such as fax or other noncontact center traffic, account for that traffic in your deployment's capacity. For the VXML + PTSN columns, the indicated number of VXML sessions and voice calls are supported simultaneously on the same gateway.
- Sizing is based on using either the Cisco Call Server or Cisco Unified CVP VXML Server.
- Each gateway is configured to share the load with its redundant pair during usual operations. Under usual operations, each gateway handles the load close to half of its capacity. During a failover scenario, each gateway operates with its maximum supported load.
- Each port provides TDM and VXML functionality including ASR/TTS.

Table 3: Maximum Number of VXML Sessions Supported by Cisco Voice Gateways (Cisco IOS Release 15.1.4.M7 and Later)

VXML Gateway CPU Capacity for Cisco IOS Release 15.1.4.M7 or Later					
Platform	**VXML Only**		**VXML + PSTN**		**Memory**
	DTMF	**ASR**	**DTMF**	**ASR**	**Recommended**
2901	12	8	9	6	2 GB
2911	60	40	47	31	2 GB
2921	90	60	71	48	2 GB
2951	120	80	95	64	2 GB
3925	240	160	190	127	2 GB
3945	340	228	270	180	2 GB
3925E	475	450	380	375	2 GB

VXML Gateway CPU Capacity for Cisco IOS Release 15.1.4.M7 or Later					
Platform	**VXML Only**		**VXML + PSTN**		**Memory**
	DTMF	**ASR**	**DTMF**	**ASR**	**Recommended**
3945E	580	550	460	450	2 GB
Based on ISO 15.1.4.M7, G.711, basic calls, Ethernet egress, CPU NTE 75%					

Note A single combination gateway cannot exceed the number of concurrent VXML sessions and VoIP calls.

Table 4: Maximum Number of VXML Sessions Supported by Cisco Voice Gateways Executing Intensive JavaScript Applications (Cisco IOS Release 15.1.4.M7 and Later)

Cisco Voice Gateway Platform	Dedicated VXML Gateway		Voice Gateway and VXML		
	VXML and DTMF	**VXML and ASR/TTS**	**VXML and DTMF**	**VXML and ASR/TTS**	**Memory Recommended**
AS5350XM	105	85	110	70	512 MB (default)
AS5400XM	105	85	110	70	512 MB (default)

Table 5: Maximum Number of VXML Sessions Supported by Cisco Voice Gateways Using HTTPS (Cisco IOS Release 15.1.4.M7 and Later)

Cisco Voice Gateway Platform	Dedicated VXML Gateway		Voice Gateway and VXML		
	VXML and DTMF	**VXML and ASR/TTS**	**VXML and DTMF**	**VXML and ASR/TTS**	**Memory Recommended**
3945E	510	342	408	270	2 GB
AS5350XM	155	120	138	95	512 MB (default)
AS5400XM	155	120	138	95	512 MB (default)

Note The performance numbers in the preceding table are only for selected models of Cisco Voice Gateways using HTTPS. Use the HTTPS performance numbers of the 3945E router, to estimate the performance numbers for router models that are not listed in Table 11.

See the section on sizing gateways for contact center traffic in *Cisco Collaboration System Solution Reference Network Designs* at https://www.cisco.com/c/en/us/support/unified-communications/

unified-communications-manager-callmanager/products-implementation-design-guides-list.html to ensure that the call arrival rates do not exceed the listed capacities.

CPU Usage

For all gateways, ensure that the overall CPU usage is less than 75 percent on average. The following factors affect CPU usage:

- Calls per second (CPS)
- Maximum concurrent calls
- Maximum concurrent VXML sessions
- Intensive JavaScript applications

Memory Considerations

Consider how much DRAM and flash memory to order. The capacity that comes with the machine by default is sufficient for most purposes. However, consider increasing the amount of DRAM in order to expand your flash memory if your application requires:

- Large numbers of distinct .wav files (as with complex self-service applications)
- Unusually large .wav files (as with extended voice messages or music files)

Note You can only extend HTTP cache to 100 MB in the current Cisco IOS releases.

Third-Party VXML Application Considerations

If you are using a non-Cisco VXML application, your deployment must adhere to the CPU usage requirements. Ensure that adequate memory is available on Cisco gateways at full load when running external VXML applications.

Contact the provider of that application for performance and availability information. Cisco makes no claims or warranties regarding the performance, stability, or feature capabilities of a third-party VXML application when interoperating in a Cisco environment.

CUBE and Virtual CUBE Considerations

For information on sizing physical or virtual Cisco UBE, see *Cisco Unified Border Element Configuration Guide* at https://www.cisco.com/c/en/us/support/unified-communications/unified-border-element/products-installation-and-configuration-guides-list.html.

For session capacity information on CUBE, see the *Cisco Unified Border Element Data Sheet* at https://www.cisco.com/c/en/us/products/unified-communications/unified-border-element/datasheet-listing.html.

Note The CVP comprehensive call flow uses more than the standard 7 messages per call leg in a VoIP call flow with Unified CM. Because of this, size CUBE sessions for contact center enterprise solutions as follows:

- ISR G2 scalability is 40% less than the standard CUBE session capacity. Scalability for the ASR1K and ISR4K series is about 75% less.
- The amount of DSPs in the platform limits CPA numbers. Treat CPA as high complexity.
- CUBE establishes a SIP session for media forking to a call recording server. If you use CUBE-controlled recording or Unified CM-controlled recording, add an extra session for each recording to your overall sizing.
- You can have a mix of CVP sessions and recording sessions on the same CUBE. Add the sessions together to properly size the CUBE. For example, if you have 1000 CVP sessions and 1000 forking sessions for call recording, then the CUBE expected load on ISR G2 is roughly:

  ```
  (1000 CVP * 1.66 for performance impact) + 1000 Recording = 2660 total sessions
  ```

  On ASR 1K/ISR 4K, the expected load is roughly:

  ```
  (1000 CVP * 4 for performance impact) + 1000 Recording = 5000 total sessions
  ```

  Use the total sessions to size against the standard SIP sessions that your CUBE model supports.

Important Correctly sizing CUBE when you activate multiple services, such as transcoder and MTP resources, on CUBE is more complex. Consult your Cisco Account team to connect with someone from the CUBE team. They can help you properly size complex CUBE deployments.

CVP Basic Video Service Sizing

You can have video-capable agents in your contact center enterprise solution.

You can use the same Unified CVP Call Server for both video calls and traditional audio calls in a comprehensive call flow.

The basic video service uses the comprehensive call flow. It requires Call Server, VXML Server, and IOS VXML Gateways. You size these components in the same way that you do for audio calls.

Cisco Unified Video conferencing hardware, Radvision IVP, and Radvision iContact are not required for the basic video service.

Note Video call is not supported in Cisco VVB.

CVP Reporting Server Sizing

Sizing the CVP Reporting Server involves many variables. Different VXML applications have different characteristics that influence the amount of reporting data. Some of these factors are:

- The types of elements in the application
- The granularity of the required data

- The call flow through the application
- The length of calls
- The number of calls

To size the CVP Reporting Server, first estimate the amount of reporting data that your VXML application generates.

Once you determine the number of reporting messages from your application, complete the following steps for each VXML application:

1. Estimate the CPS that the application receives.
2. Estimate the number of reporting messages for your application.

This equation determines the number of reporting messages that a VXML application generates each second:

*A# = %VXML * CPS * MSG*

Where:

- *A#* is the number of estimated reporting messages per second for an application.
- *CPS* is the number of calls per second.
- *%VXML* is the percentage of calls that use this VXML application.
- *MSG* is the number of reporting messages that this application generates.

Total the values for each VXML application in your solution to get the estimated reporting messages per second for your solution.

Each CVP Reporting Server can handle 420 messages per second. If your solution requires more than one CVP Reporting Server, partition the VXML applications to use specific Reporting Servers.

Solutions with Multiple CVP Reporting Servers

In solutions that require more than one CVP Reporting Servers, you partition the deployment vertically.

When vertically partitioning to load balance reporting data, consider these requirements and guidelines:

- Associate each Call Server and VXML Server with only one CVP Reporting Server.
- Reports cannot span multiple Informix databases.
- Subdivide applications that generate more combined call processing and application messages than one CVP Reporting Server can support.
- You can filter VXML. Filtering out noninteresting data creates more usable data repositories that support higher message volume.
- Configure the dial plan and other available means to direct the incoming calls to the appropriate Call Server and VXML Server.

For more information on these requirements, see the *Reporting Guide for Cisco Unified Customer Voice Portal* available at http://www.cisco.com/c/en/us/support/customer-collaboration/unified-customer-voice-portal/products-installation-and-configuration-guides-list.html.

To combine data from multiple databases, you can use these possible options:

- Export the reporting data to another format, like a spreadsheet, and combine the data outside of the database.
- Export the reporting data to CSV files and import it into a customer-supplied database.
- Extract the data to a customer-supplied data warehouse and run reports against that data.

Message Details on CVP Reporting Servers

This table lists the number of reporting messages that various elements or activities generate.

Table 6: Number of Reporting Messages Per Element or Activity

Element or Activity	Number of Reporting Messages (Unfiltered)
Start	2
End	2
Subflow Call	2
Subflow Start	2
Subflow Return	2
Throw	2
Alert	2
Subdialog_start	2
Subdialog_return	2
Hotlink	2
HotEvent	2
Transfer w/o Audio	2
Currency w/o Audio	2
Flag	2
Action	2
Decision	2
Application Transfer	2
VXML Error	2
CallICMInfo (per call)	2
Session Variable (per change)	2
Custom Log (per item)	2

Element or Activity	Number of Reporting Messages (Unfiltered)
Play (Audio file or TTS)	2
LeaveQueue	2
Callback_Disconnect_Caller	3
Callback_Add	4
Callback_Get_Status	4
Callback_Set_Queue_Defaults	4
Callback_Update_Status	4
Callback_Enter_Queue	5
Callback_Reconnect	5
Get Input (DTMF)	5
Callback_Validate	6
Get Input (ASR)	9
Form	10
Digit_with_confirm	20
Currency_with_confirm	20
ReqICMLabel	30

Note Every application requires these elements. You cannot filter them.

Sizing for Unified CM Clusters

Unified CM clusters provide a mechanism for distributing call processing across a converged IP network infrastructure. Clusters also facilitate redundancy and provide feature transparency and scalability.

For a more detailed view of Unified CM clusters, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.

Cluster Sizing Concepts

Before attempting to size a Unified CM cluster for your solution, perform the following design tasks:

- Determine the different types of call flows.
- Determine the required deployment model (single site, centralized, distributed, clustering over the WAN, or remote branches within centralized or distributed deployments).
- Determine the protocols to be used.

- Determine redundancy requirements.
- Determine all other customer requirements for Cisco Unified Communications that share a cluster with a Unified CCE deployment. For example, consider Cisco Unified IP Phones, applications that are not part of Unified CCE, route patterns, and so forth.

After you complete these tasks, you can begin to accurately size the necessary clusters. Many factors affect the sizing of a cluster, and the following list mentions some of those factors:

- Number of office phones and the busy hour call attempt (BHCA) rate per phone
- Number of inbound agent phones and the BHCA rate per phone
- Number of CTI ports and the BHCA rate on those VoIP endpoints. (If you use Unified CVP for call treatment, self-service, and queuing, these factors might not apply.)
- Number of Voice Gateway ports and the BHCA rate on those VoIP endpoints
- Number of outbound agent phones, outbound dialing mode, and BHCA rate per phone
- Number of outbound dialer ports, number of VRU ports for outbound campaigns, and the BHCA rate per port for both
- Number of mobile agents and the BHCA rate per mobile agent
- Number of voicemail ports and the BHCA rate to those VoIP endpoints
- Signaling protocols used by the VoIP endpoints
- Percent of agent call transfers and conferences
- Dial plan size and complexity, including the number of dialed numbers, lines, partitions, calling search spaces, locations, regions, route patterns, translations, route groups, hunt groups, pickup groups, and route lists
- Amount of media resources needed for functions such as transcoding, conferences, encryption, and so forth
- Coresident applications and services such as CTI Manager, E-911, and Music on Hold
- Unified CM release (sizing varies per release)
- Type of Unified CM OVA

Other factors can affect cluster sizing, but theses are the most significant factors in terms of resource consumption.

In general, you estimate the resource consumption (CPU, memory, and I/O) for each of these factors to size the Unified CM cluster. You then choose VMs that satisfy the resource requirements. Gather information about these factors before you can size a cluster with any accuracy.

Cluster Guidelines

The following guidelines apply to all Unified CM clusters in your solution:

- All primary and backup subscribers must use the same OVF template. All subscribers in the cluster must run the same Unified CM software release and service pack.

- Within a cluster, you can enable a maximum of eight subscribers (four primary and four backup subscribers) with the Cisco Call Manager Service. You can use more VMs for dedicated functions such as TFTP, publisher, and music on hold.
- In a 4000 Agent Reference Design, a Unified CM cluster can support about 4000 agents. In a 12,000 Agent Reference Design, a Unified CM cluster with four primary and four backup subscribers can support about 8000 agents. These limits assume that the BHCA call load and all configured devices are spread equally among the eight call processing subscribers with 1:1 redundancy. These capacities can vary, depending on your specific deployment. Size your solution with the *Cisco Unified Communications Manager Capacity Tool*.

  A subscriber can support a maximum of 1000 agents. In a fail-over scenario, the primary subscriber supports a maximum of 2000 agents.

Note In a 4000 Agent Reference Design, a cluster with four subscribers (two primary and two backup) can support the maximum load. If you create clusters with more subscribers, do not exceed the maximum of 4000 agents for the cluster.

When sizing the cluster to support contact center solutions for the appropriate number of CTI resources, remember to account for the following:

- Configured phones from agents who are not signed in
- Applications which remotely control the device like Call Recording, Attendant Console, and PC-clients
- Other 3rd-party applications which consume CTI resources

Unified CM can support multiple concurrent CTI resources, for example, when multiple lines, the contact center, and recording are used concurrently. Those CTI resources follow the same CTI rules as described in the *Cisco Collaboration System Solution Reference Network Designs*:

- Devices (including phones, music on hold, route points, gateway ports, CTI ports, JTAPI Users, and CTI Manager) must never reside or be registered on the publisher. If there are any devices registered with the publisher, any administrative work on Unified CM impacts call processing and CTI Manager activities.
- Do not use a publisher as a fail-over or backup call processing subscriber in production deployments. Any deviations require review by Cisco Bid Assurance on a case-by-case basis.
- Any deployment with more than 150 agent phones requires a minimum of two subscribers and a combined TFTP and publisher. The load-balancing option is not available when the publisher is a backup call processing subscriber.
- If you require more than one primary subscriber to support your configuration, then distribute all agents equally among the subscriber nodes. This configuration assumes that the BHCA rate is uniform across all agents.
- Similarly, distribute all gateway ports and CTI ports equally among the cluster nodes.
- Some deployments require more than one Unified CCE JTAPI user (CTI Manager) and more than one primary subscriber. In these deployments, if possible, group and configure all devices that are monitored by the same Unified CCE JTAPI User (third-party application provider), such as Unified CCE route points and agent devices, on the same VM.

- Enable CTI Manager only on call processing subscribers, thus allowing for a maximum of eight CTI Managers in a cluster. To provide maximum resilience, performance, and redundancy, load-balance CTI applications across the various CTI Managers in the cluster. For more CTI Manager considerations, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.
- If you have a mixed cluster with Unified CCE and general office IP phones, if possible, group and configure each type on a separate VM (unless you need only one subscriber). For example, all Unified CCE agents and their associated devices and resources are on one or more Unified CM servers. Then, all general office IP phones and their associated devices (such as gateway ports) are on other Unified CM servers, as long as cluster capacity allows. If you use the *Cisco Unified Communications Manager Capacity Tool*, run the tool separately with the specific device configuration for each primary Unified CM server. You need to run it multiple times because the tool assumes that all devices are equally balanced in a cluster. Remember that with Unified CCE, you must use the 1:1 redundancy scheme.
- Use hardware-based conference resources whenever possible. Hardware conference resources provide a more cost-effective solution and allow better scalability within a cluster.
- Register all CTI route points for the Unified CCE Peripheral Gateway (PG) JTAPI user with the subscriber node running the CTI Manager instance that communicates with that Unified CCE PG.
- The *Cisco Unified Communications Manager Capacity Tool* does not currently measure CTI Manager impact on each VM separately. However, the CTI Manager does place an extra burden on the subscriber running that process. The tools report the resource consumption based on these subscribers. The actual resource consumption on the other Unified CM subscribers can be slightly lower.
- Even if a contact center agent does not use them, count all devices for a Unified CCE PG JTAPI user as an agent device. The PG is still notified of all device state changes for that phone, even though an agent does not use the phone. To increase cluster scalability, if your agents do not regularly uses a device, do not associate the device with the Unified CCE PG JTAPI user.
- CPU resource consumption by Unified CM varies, depending on the trace level enabled. Changing the trace level from Default to Full on Unified CM can increase CPU consumption significantly under high loads. The Cisco Technical Assistance Center does not support changing the tracing level from Default to No tracing.
- Under usual circumstances, place all subscribers from the cluster within the same LAN or MAN. Do not place all members of a cluster on the same VLAN or switch.
- If the cluster spans an IP WAN, follow the specific guidelines in the sections on clustering over the WAN in this guide and in the *Cisco Collaboration System Solution Reference Network Designs*.

For the most current information about supported releases, see the latest version of the your solution's *Compatibility Matrix*.

For more Unified CM clustering guidelines, see the *Cisco Collaboration System Solution Reference Network Designs* at http://www.cisco.com/go/ucsrnd.

Component and Feature Impacts on Scalability

Some optional components and features affect the scalability and capacity of your solution. This table lists some of these effects.

Component or feature	Impact
IPsec	When you enable IPsec: • The PG capacities decrease by 25% for agents, VRU ports, SIP Dialer ports, and call rate. • The maximum call rate (calls per second) decreases by 25%.
Mobile agents	Unified CCE does not directly control the phones of mobile agents. The two delivery modes, Call-by-Call and Nailed Connection, use resources differently.
Cisco Outbound Option	Your outbound resources can vary based on hit rate, abandon limit, and talk time for the campaigns. A quick, but inexact, estimate is that you require two ports for each outbound agent. While you can technically have 2000 agents per PG assigned to outbound calls, the Dialers probably cannot keep all those agents fully occupied. Use the sizing tool to determine outbound resources required for your campaigns.
Agent Greeting	The Agent Greeting feature affects the Router, Logger, and Unified CM. On the Router and Logger, this feature increases the route requests made. That effectively decreases the maximum call rate by about one third.
Extended Call Context (ECC)	Increased Extended Call Context (ECC) usage affects performance and scalability on critical components of Unified CCE. The capacity impact varies based on the ECC configuration, and requires professional guidance on a case-by-case basis.

Resource Requirements for Reporting

Do not run more than ten concurrent reports on any client machine. This is a combined limit for reports that run on the Unified Intelligence Center User Interface, Permalinks, and Dashboards on the client machine. See the *Maximum rows per report* row in the System Load Limits table for the maximum number of rows supported in reports.

Our capacity testing shows that 200 concurrent reporting users can each have the following running reports:

- Two Live Data reports with 100 rows of 10 fields.
- Two real-time reports with 100 rows of 10 fields, refreshing every 15 seconds.
- Two historical reports with 2000 rows of 10 fields, refreshing every 30 minutes.

That means 400 Realtime and 400 Historical reports can be run concurrently. These numbers include the reports run from permalinks, reports on dashboards, schedules and desktop gadgets.

For example, if you have 200 historical permalinks open and 100 supervisors are accessing one historical report each from the desktop gadget, you can run 100 more historical reports.

If you have fewer reporting users on a node, they can run proportionally more reports. But, no client machine can exceed the ten report limit.

Cisco Virtualized Voice Browser Sizing

The call capacity of Cisco VVB is based on the call support for ASR or TTS activities and on the type of VXML application. For instance, an intensive JavaScript application reduces call capacity and VVB with HTTPS has a lower call capacity than with HTTP.

Ensure that the average overall CPU usage is less than 65 percent. The following factors affect CPU usage:

- Calls per second (CPS)
- Maximum concurrent VXML sessions
- Complexity of VXML applications

Before sizing Cisco VVB, use the Unified CCE Resource Calculator to determine the maximum number of trunks (DS0s) and VXML VRU ports to support the entire solution.

For almost all Unified CVP deployment models, sizing is based on these factors:

- The maximum concurrent VXML sessions and VoIP calls
- The CPS that Cisco VVB handles

Note

- The performance numbers listed in ASR and TTS columns are applicable only for MRCPv1 and v2.
- When Open Virtual Appliance (OVA) and VVB are already installed and the customer wants to change a profile from small to medium or vice versa, the existing OVA must be deleted and a new install with a fresh OVA specification must be done.

System Specification	CPS	DTMF (Non-Secure)	TTS / ASR (Non-Secure)	DTMF / TTS / ASR (Secure)
Medium OVA (4 CPU, 10-GB RAM)	16	600	480	480
Small OVA (4 CPU, 8-GB RAM)	16	480	380	380
Medium OVA with AppD (8 CPU, 10-GB RAM)	16	600	480	480
Small OVA with AppD (8 CPU, 8-GB RAM)	16	480	380	380
KVM [4451 (2 CPU - Gladen), 8-GB RAM]	6	120	96	96
KVM [4431 (6 CPU - Gladen), 8-GB RAM]	3	80	70	70
KVM [4351 (6 CPU - Rangley), 8-GB RAM]	3	60	50	50
KVM [4331 (6 CPU - Rangley), 8-GB RAM]	2	40	30	30

Note

- TLS/SRTP reduces CPS up to 25% for small or medium profile.
- TLS/SRTP with ASR/TTS is not currently supported.
- Secure: Secured Transport over HTTPS/TLS/SRTP
- These values represent the performance with VXML pages from Unified CVP Call Studio applications running on the Unified CVP VXML Server. Other VXML applications can perform differently. These figures are for a system running VXML v2.0 and MRCPv1 or v2 with CPU utilization of less than 65 percent.

  These values reflect testing of moderately complex VXML applications on the Cisco Unified CVP VXML Server. Performance varies with different applications. Performance from external VXML applications (such as Nuance OSDMs) is not representative of the performance when interoperating with non-Cisco applications. Ensure that adequate memory is available on Cisco VVB at full load when running external VXML applications. Contact the application provider for performance and availability information.
- We make no claims or warranties regarding the performance, stability, or feature capabilities of an external VXML application added to your contact center enterprise solution.
- You can extend the HTTP cache to 512 MB in Cisco VVB.
- When calculating CPS at Cisco VVB, consider every call (VRU, Ringback, and WAAG) received. When you calculate the CPS into Cisco VVB for your solution, first determine the services which each incoming call at CVP uses.

  For example, if you disable WAAG for all agents, the total CPS at Cisco VVB is (2 x incoming rate) at CVP. That is one call for VRU and one call for Ringback. If you enable WAAG for all agents, the total CPS at Cisco VVB is (4 x incoming rate) at CVP, because WAAG adds two more calls.

Sizing for Cisco Finesse

Cisco Finesse supports up to 1800 agents and 200 supervisors (for a total of 2000 users) over HTTPS for each Cisco Finesse server pair.

Sizing for Congestion Control

Congestion Control protects the Router from overload conditions caused by high call rates. When faced with extreme overload, congestion control keeps the system running close to its rated capacity.

Congestion Control provides satisfactory service during an overloaded condition to a smaller percentage of calls, rather than a highly degraded service to all calls. The feature keeps the system within its capacity by rejecting calls at the call entry point. Throttling the capacities ensures that the routed calls receive acceptable service without timeouts.

In the discussion of Congestion Control, "calls" include nonvoice tasks from third-party multichannel applications that use the Universal Queuing APIs and voice calls. Congestion Control treats these calls and tasks the same. It monitors and throttles incoming calls and tasks, and does not drop calls or tasks once they are in the system. This means that transfers and RONAs are counted toward Congestion Control, but are not throttled or rejected.

Another exception is picking tasks like email in course of multi-tasking on a voice call. These pick task requests also do not get rejected owing to Congestion Control.

Note Pull task requests will get rejected if the system is congested except in the cases where the tasks may be waiting in the Unified CCE queue. Requests to pull tasks out of the Unified CCE queue is allowed even during congestion because this helps decongest the Unified CCE system.

Note For Enterprise Chat and Email, forwarded email tasks are considered new tasks, and are subject to throttling.

The measured CPS at the Router is the trigger for identifying congestion. The deployment type sets the supported CPS capacity for your solution. The Router measures the new incoming call requests from all the routing clients and computes a moving weighted average. If the average CPS exceeds the thresholds, the congestion levels change and the reduction percentage increases. The congestion control algorithm has three congestion levels. It rejects or treats the incoming calls at the value for that level. The system notifies the routing clients of changes in the congestion level.

In a Contact Director Reference Design, the congestion control is based on the call rate measured at each instance. The Contact Director receives information on the congestion level of each target. It applies any necessary reduction in its routing decisions. The INCRP routing client also applies congestion control to calls before sending them to the target instance.

Deployment Type Descriptions

After upgrading or installing the system, configure the system to a valid deployment type. The following table lists the supported deployment types with guidelines for selecting a valid deployment type.

For more information on the requirements referred to in this table, see your solution page on the *Cisco Collaboration Virtualization* site at http://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/cisco-collaboration-virtualization.html.

Table 7: Deployment Types

Deployment Type Code	Deployment Name	Guidelines for Selection
0	Not Specified	This is a system default deployment type. You cannot select this option; is the default setting after fresh install or upgrade.
1	NAM (Deprecated)	Select this deployment type for NAM instance in a Contact Director deployment. The system should be distributed deployment with Router and Logger installed on different VMs, which meets the specified requirements. No agents are allowed in this deployment type. If agents are configured and signed in, the capacity is adjusted to maximum capacity of a Unified CCE 12000 Agents solution.

Deployment Type Code	Deployment Name	Guidelines for Selection
2	Contact Director	Select this deployment type for ICM instance which is dedicated to self-service call flows using Unified CVP or third-party VRU systems. The system should be distributed deployment with Router and Logger installed on different VMs which meets the specified requirements. No agents are allowed in this deployment type. If agents are configured and signed in, the capacity is adjusted to maximum capacity of an Enterprise Contact Center (Unified CCE 12000 Agents Router/ Logger).
3	NAM Rogger (Deprecated)	Select this deployment type for NAM instance in a Contact Director deployment. The Router and Logger colocated on a single VM meet the specified requirements. No agents are allowed in this deployment type. If agents are configured and signed in, the capacity is adjusted to maximum capacity of an Enterprise Contact Center (Unified CCE 12000 Agents Router/ Logger).
4	ICM Router/Logger	Select this deployment for type ICM Enterprise system where both Legacy TDM ACD PGs and CCE PGs are deployed. The system should be distributed deployment with Router and Logger installed on different VMs which meet the specified requirements.
5	UCCE: 8000 Agents Router/Logger	Select this deployment for type CCE Enterprise system where only CCE PGs are deployed. The system should be distributed deployment with Router and Logger installed on different VMs which meet the specified requirements for 8000 CCE agents.
6	UCCE: 12000 Agents Router/Logger	Select this deployment type for CCE Enterprise system where only CCE PGs are deployed. The system should be distributed deployment with Router and Logger installed on different VMs which meet the specified requirements for 12000 CCE agents.
7	Packaged CCE: 2000 Agents	Select this deployment type for a Packeged CCE production deployment.
8	ICM Rogger	Select this deployment type for ICM Enterprise system where both Legacy TDM ACD PGs and CCE PGs are deployed. The Router and Logger are colocated on a single VM which meets the specified requirements.
9	UCCE: 4000 Agents Rogger	Select this deployment type for CCE Enterprise system where only CCE PGs are deployed. The Router and Logger are colocated on a single VM which meets the specified requirements.
10	Packaged CCE: Lab Mode	Select this deployment type for a Packaged CCE lab deployment.

Deployment Type Code	Deployment Name	Guidelines for Selection
13	UCCE: Progger (Lab Only)	For all lab deployments, select this type although the Router, Logger, and PG are not on the same VM. **Note** This deployment type is not supported for production systems.
16	UCCE: 2000 Agents	Select this deployment type for the 2000 Agent Reference Design in a Unified CCE solution.
17	Packaged CCE: 4000 Agents	Select this deployment type for the 4000 Agent Reference Design in a Packaged CCE solution.
18	Packaged CCE: 12000 Agents	Select this deployment type for the 12000 Agent Reference Design in a Packaged CCE solution.
19	UCCE: 24000 Agents Router/Logger	Select this deployment type for the 24000 Agent Reference Design in a Unified CCE solution.

Note It is important to set the proper deployment type for your solution during the configuration. If you select the wrong deployment type, your solution is either unprotected from overload or it rejects and treats calls based on incorrect capacity settings.

Congestion Treatment Mode

The system has five options to handle the calls that are rejected or treated due to congestion. You can choose any of the following options to handle the calls:

- **Treat Call with Dialed Number Default Label**—The rejected calls are treated with the default label of the dialed number on which the incoming call arrived.
- **Treat call with Routing Client Default Label**—The rejected calls are treated with the default label of the routing client on which the incoming call arrived.
- **Treat call with System Default Label**—The rejected calls are treated with the system default label set in Congestion Control settings.
- **Terminate call with a Dialog Fail or RouteEnd**—Terminates the incoming call dialog with a dialog failure.
- **Treat call with a Release Message to the Routing Client**—Terminates the incoming call dialog with a release message.

You set the treatment options in the congestion settings either at the routing client or at the global level. If you select a treatment mode at the routing client, it takes precedence over the system congestion settings.

Note If you choose to return a label back to treat the call with an announcement, use an announcement system external to the Unified CCE instance. Never return a treated call to the Unified CCE instance for further processing.

Call Treatment for Outbound Option

Outbound Option is a special case for call treatment with Congestion Control. When you integrate the Media Routing Peripheral Gateway (MR PG) for Outbound Option, configure the PG's routing client to always send the dialog failure. The dialer retries the rejected reservation calls after a specified period.

Congestion Control Levels and Thresholds

The Congestion Control algorithm works in three levels. Each level has onset and abatement values. When the average CPS exceeds one level's onset value, the system moves to a higher congestion level. For example, if the system is at level 0 and the CPS exceeds the Level 2 onset capacity, the system moves directly to Level 2. The congestion level reduces when the average CPS falls below the current level's abatement value. Congestion levels can rise several levels at once. However, the congestion level reduces only one level at a time.

Table 8: Congestion Levels

Congestion Levels	Threshold (Percent of Capacity)	Description
Level1Onset	110%	If the average CPS exceeds this value, the congestion level moves to Level 1.
Level1Abate	90%	If the average CPS goes below this value, the congestion level moves back to Level 0 (Normal operating Level).
L1Reduction	10%	The percentage of incoming calls that are rejected at Level 1 congestion.
Level2Onset	130%	If the average CPS exceeds this value, the congestion level moves to Level 2.
Level2Abate	100%	If the average CPS goes below this value, then the congestion level moves back to Level 1.
Level2Redution	30%	The percentage of incoming calls that are rejected in Level 2 congestion.
Level3Onset	150%	If the average CPS exceeds this value, the congestion level moves to Level 3.
Level3Abatement	100%	If the average CPS goes below this value, the congestion level moves back to Level 2.

Congestion Levels	Threshold (Percent of Capacity)	Description
Level3Reduction	Variable reduction from 100% to 30%	The percentage of incoming calls that are rejected in Level 3 congestion. Depending on the incoming call rate, the reduction percentage varies from 30% to 100% when the congestion level enters Level 3.

Note You cannot configure the onset, abatement, and reduction settings. These values are defined as a percentage of the standard CPS capacity for the system.

Congestion Control CPS Limits

This table lists the maximum supported calls per second (CPS) for the supported deployment types.

Table 9: Deployment Types

Deployment type	Maximum calls per second	Notes
NAM (Deprecated)	300	Deprecated as of 11.5.
Contact Director	300	Reference Design
NAM Rogger (Deprecated)	150	Deprecated as of 11.5.
UCCE: 12000 Agents	105	Reference Design
Packaged CCE: 2000 Agents	18	Reference Design
UCCE: 4000 Agents	35	Reference Design
Packaged CCE: Lab Mode	1	Not supported for production environments.
UCCE: Progger (Lab Only)	4	Not supported for production environments.
UCCE: 2000 Agents	18	Reference Design
Packaged CCE: 4000 Agents	35	Reference Design
Packaged CCE: 12000 Agents	105	Reference Design
UCCE: 24000 Agents Router/Logger	105	Reference Design

Operating Considerations for Reference Design Compliant Solutions

Solution-Wide Support for Transport Layer Security

The contact center enterprise solutions use TLS 1.2 by default. For most components, you can enable earlier versions of TLS if necessary.

Time Synchronization for Your Solution

To ensure accurate operation and reporting, all the components in your contact center solution must use the same value for the time. You can synchronize the time across your solution using a Simple Network Time Protocol (SNTP) server. The following table outlines the needs of various component types in your solution.

Important Use the same NTP sources throughout your solution.

Type of component	Notes
Domain controllers	Domain controllers must all point to the same NTP servers.
ESXi hosts	All ESXi hosts must point to the same NTP servers as primary domain controllers.
Windows components in the contact center domain	Windows machines in the domain point to, and are automatically in synch with, the primary domain controller for NTP. They require no configuration for NTP.
Windows components not in the contact center domain	Follow the Microsoft documentation to synchronize directly with the NTP server.
Non-Windows components	Components such as Unified Intelligence Center, Cisco Finesse, Customer Collaboration Platform, and Unified Communications must point to the same NTP servers as the domain controllers.
Cisco Integrated Service Routers	To provide accurate time for logging and debugging, use the same NTP source as the solution for the Cisco IOS Voice Gateways.

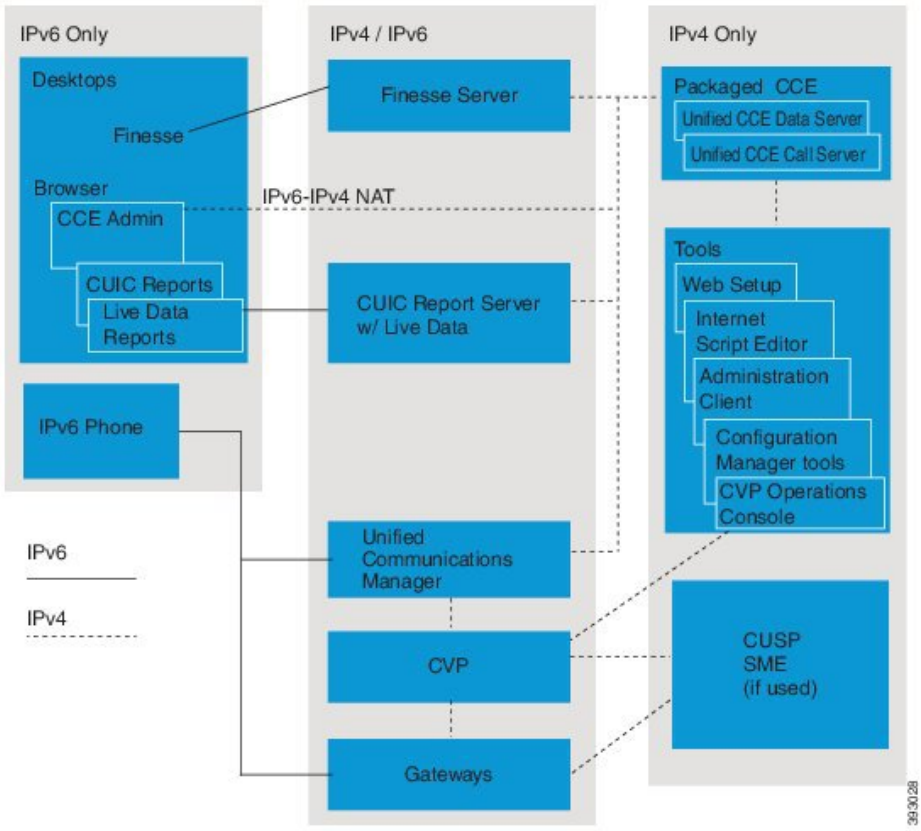
Contact Center Enterprise Solution Support for IPv6

Unified Contact Center solutions can support IPv6 connections for agent and supervisor Finesse desktops and phones. This support means that most of the endpoints in your deployment can use IPv6 addresses.

Your IPv6-enabled deployment can use either IPv6-only or a mix of IPv4 and IPv6 endpoints. Servers that communicate with those endpoints can now accept IPv6 connections, in addition to IPv4 connections. Communication between servers continues to use IPv4 connections.

This diagram shows a logical view of a deployment with only IPv6 desktops and phones:

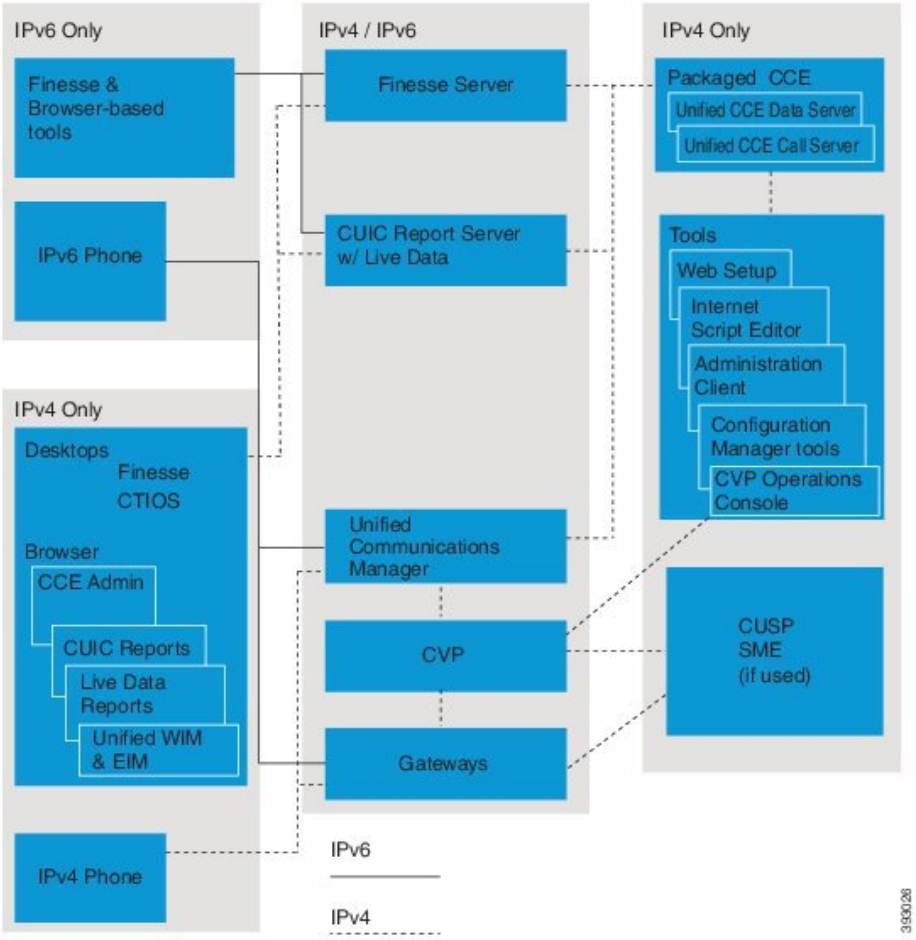
Figure 2: Packaged CCE Deployment with Only IPv6 Agents



In these IPv6-only deployments, agents and supervisors use Finesse and browser-based tools that connect to dual-stack interfaces on the servers. The ingress gateways and Unified CM also use dual-stack interfaces to handle the voice traffic. These deployments require IPv4-based Administration Workstations to run the configuration tools that you do not access through a browser.

This diagram shows a logical view of a mixed deployment with both IPv6 and IPv4 endpoints:

Figure 3: Packaged CCE Deployment with Both IPv4 and IPv6 Agents



The Finesse desktop can support either IPv4 or IPv6 connections. Agents and supervisors who use the CTI OS desktops must use IPv4 connections. Enterprise Chat and Email agents must use IPv4 connections.

For a list of endpoints that support IPv6, see your solution *Compatibility Matrix*.

For information on enabling IPv6 in the Cisco Unified Communications Manager, see *Deploying IPv6 in Unified Communications Networks with Cisco Unified Communications Manager* at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.

General IPv6 Design Considerations

You cannot enable IPv6 on all the component servers in your contact center. For example, your deployment might use IPv6 phones, but only use IPv4 desktops. In that deployment, you enable IPv6 on the component servers that connect to the desktops.

When an IPv4 endpoint communicates with an IPv6 endpoint, Unified Communication Manager invokes Media Termination Points (MTPs) to negotiate the mismatch. As a result, IPv4-only endpoints like the VXML browser require extra MTP devices. Most uses of CVP features require MTPs for the IPv4-to-IPv6 negotiation.

You do not have to set up IPv6 at installation. You can enable IPv6 at any time. You can also revert to IPv4 from IPv6 if necessary.

You assign IPv6 addresses to hosts that have a dual IP stack. You use the Fully Qualified Domain Name (FQDN), rather than the IPv6 address, in the solution's user interface.

CVP Features with IPv6

When you enable IPv6 in your solution, CVP requires the following conditions to support its features:

- **Call Survivability**—Use only IPv4 for the incoming trunk to the gateway.
- **Courtesy Callback and Refer**—When a trunk carries both the incoming and outbound dialing traffic, the session target in the Ingress Gateway dialpeer uses the same protocol as the incoming trunk.

Desktop and Tool Support

This table lists which desktops support each connection type:

Desktop	IPv6 Connections	IPv4 Connections
Finesse	Yes	Yes
	No	Yes

A supervisor's team can include a mix of agents using Finesse desktops with either IPv4 or IPv6 connections and CTI OS desktops with IPv4 connections.

You cannot use IPv6 to connect to a Finesse desktop through Citrix XenApp.

Desktops with either IPv4 or IPv6 connections can access the following tools:

- Unified CCE Administration web tool (using NAT64).
- Finesse configuration tools
- Cisco Unified Intelligence Center (Cisco Unified IC) configuration tools and reports

You require an IPv4 connection to access the following tools:

- Enterprise Chat and Email
- Web Setup
- Script Editor
- Internet Script Editor
- Diagnostic Portico
- Configuration Manager and its associated tools

IPv6 Design Considerations for Video Endpoints

If you use video endpoints, consider the following points when enabling IPv6:

- Configure the incoming trunk to gateway in IPv4 mode only.
- Disable ANAT in the Ingress Gateway.

- Agent devices can use either IPv4 or dual IP mode.

Other Component and Feature Support

This table lists the connection type that each component or feature supports in an IPv6-enabled environment:

Component or Feature	Supported Connections in IPv6-enabled Environment		Notes
	IPv6	IPv4	
Enterprise Chat and Email	No	Yes	
Mobile Agent	No	Yes	The CTI ports for Mobile Agent can only have an **IP Addressing Mode** of **IPv4 Only**.
Outbound Option	No	Yes	The Outbound Option Dialer uses IPv4 to place calls. A voice gateway that supports both IPv4 and IPv6 renegotiates call signaling and media to IPv6 during referral to an IPv6 agent. You cannot use an IPv6-only voice gateway with Outbound Option. An IPv6 client cannot import to Outbound Option.
Customer Collaboration Platform	No	Yes	
Unified CM Silent Monitoring	Yes	Yes	
Virtualized Voice Browser	No	Yes	You cannot use Cisco VVB in an IPv6-enabled environment.

For more information on enabling IPv6 in a contact center enterprise solution, see your solution's *Installation and Upgrade Guide*. These documents have more details for specific products:

Component	Documents
Unified CVP	*Configuration Guide for Cisco Unified Customer Voice Portal*
Cisco Finesse	*Cisco Finesse Installation and Upgrade Guide* *Cisco Finesse Administration Guide*
Cisco Unified Intelligence Center	*Administration Console User Guide for Cisco Unified Intelligence Center*

Solution-Wide Support for vMotion

Cisco Contact Center Enterprise solution components, including Cisco Unified Communications Manager (CUCM), support vMotion of live VMs. It is applicable on all contact center deployment models on Cisco HyperFlex servers. An HX node with adequate spare capacity to serve as the target for the vMotion operation

must be available on the HyperFlex HX cluster. Use vCenter to initiate and monitor the vMotion operation. This setup can be used to perform any hardware maintenance operation or move all VMs out of a HyperFlex HX node without disrupting the contact center operations in order to upgrade firmware or perform other maintenance operations on the HyperFlex node. We recommend that HyperFlex clusters be deployed with 10-GB Fabric Interconnects to avoid any latency issues during vMotion operations.

Note vMotion for Contact Center Enterprise solution components is not supported across HyperFlex clusters, or in HyperFlex Stretched Cluster deployments.