CHAPTER 14

Sizing

This chapter discusses how to determine how many physical machines to order and, in the case of gateways and gatekeepers, what kind to order.

This chapter covers the following topics:

Sizing Overview

When sizing a contact center, first determine the worst-case contact center profile in terms of the number of calls that are in each state. In other words, if you were to observe the contact center at its busiest instant in the busiest hour, how many calls would you find are in each of the following states:

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

In counting the number of calls that are in the talking state, count only calls that are using Unified CVP or gateway resources. To determine whether a talking call is using resources, you must consider how the call gets transferred to that VRU or agent. If the call was transferred via VoIP, it continues to use an ingress gateway port and it continues to use a Unified CVP resource because Unified CVP continues to monitor the call and provides the ability to retrieve it and re-deliver it at a later time. The same is true of calls that are tromboned to a TDM target, using both an incoming and an outgoing TDM port on the same gateway or on a different gateway (that is, toll bypass). Calls that are transferred to VRUs or agents in this manner should be counted as talking calls.

However, if the call was transferred via *8 TNT, hookflash, Two B Channel Transfer (TBCT), or an ICM NIC, neither the gateway nor Unified CVP play any role in the call. Both components have reclaimed their resources, therefore such calls should not be counted as talking calls.

Finally, include in the overall call counts those calls that have been transferred back into Unified CVP for queuing or self-service, via either blind or warm methods. Because these calls usually do not amount to more than 5% or 10% of the overall call volume, it is easy to overlook them.

The definitions of these call states differ somewhat from the definitions used for port licensing purposes (see Licensing, page 15-1). The use of automatic speech recognition (ASR) or text-to-speech (TTS) has nothing to do with delineating which calls are in which state, whereas it does for licensing purposes.

Similarly, the call state determination has nothing to do with whether the agents are Unified CCE agents or ACD agents, nor does it matter whether the customer intends to use Unified CVP's ability to retrieve and re-deliver the call to another agent or back into self-service.

In addition to the overall snapshot profile of calls in the contact center, you must also consider the busiest period call arrival rate in terms of calls per second. You will need this information for the contact center as a whole. Because it is hard to identify a true maximum arrival rate, you can use statistical means to arrive at this number. Except in fairly small implementations, this is seldom the critical factor in determining sizing.

With the above data, you can begin sizing each component in the network. This section next considers the Unified CVP products: Call Server and VoiceXML Server followed by the gateways, gatekeepers, and content switches. This section deals entirely with the number and type of physical components required to support the Unified CVP system, but it does not include any discussion of redundancy. For an understanding of how to extend these numbers to support higher reliability, see Designing Unified CVP for High Availability, page 4-1.

Note Unless otherwise noted, the information in this chapter applies to all deployment models, including Model #1: Standalone Self-Service.

Unified CVP Call Server

Note The Unified CVP Call Server is not used in Model #1: Standalone Self-Service. This section does not apply to such deployments.

Unified CVP Call Servers are sized according to the number of calls they can handle, in addition to their maximum call arrival rate.

Each Unified CVP Call Server can handle 850 SIP calls or 500 H.323 calls. Each Unified CVP Call Server is further limited to a sustained call arrival rate of 7 calls per second (cps). However, Model #4 is exempt from this limitation because the Unified CVP Call Server in that model does not perform any H.323 or SIP processing.

Specifically, the number of Unified CVP Call Servers required is the larger of:

((Self Service) + (Queue and Collect) + Talking) / 850 [or 500 for H.323], rounded up

or

(Average call arrival rate) / 7, rounded up [except in Model #4]

Unified CVP VoiceXML Server

Unified CVP VoiceXML Server sizing is simple: one VoiceXML Server can handle up to 750 calls. If you are using Unified CVP VoiceXML Servers, you should size those machines according to the following formula:

Calls / 750, rounded up

where *Calls* refers to the number of calls that are actually in Unified CVP VoiceXML Server self-service applications at that busy moment snapshot in time.

Unified CVP Co-Residency

The following components can be installed on the same physical server (co-resident):

- Unified CVP Call Server
- Unified CVP VoiceXML Server
- Media Server

A SIP-based co-resident server can handle 750 SIP calls as well as 750 VoiceXML Server sessions simultaneously, and it can handle a sustained call arrival rate of 6 calls per second. An H.323 co-resident server can handle 500 H.323 calls as well as 500 VoiceXML Server sessions simultaneously, and it can handle a sustained call arrival rate of 6 calls per second.

The number of Unified CVP Call Servers required is the larger of:

((Self Service) + (Queue and Collect) + Talking) / 750 [or 500 for H.323], rounded up

or

(Average call arrival rate) / 6, rounded up [except in Model #4]

The co-resident media server can be used for up to 750 calls [or 500 for H.323], assuming that prompt caching is enabled in the VoiceXML gateways. If multiple co-resident servers are to be used, you must load-balance across the co-resident media servers in order to spread the load of the calls across all of the servers. To reduce the administrative overhead of managing content on multiple media servers, separate dedicated media servers can be used.

Co-Resident Reporting Server and Call Server

The Unified CVP Reporting Server can also be co-resident with the Call Server, but only for Standalone VoiceXML deployments. The Call Server is normally not needed in a Standalone VoiceXML deployment; but if reporting is desired, a Call Server is required in order to send the reporting data from the VoiceXML Server to the Reporting Server. Thus, when the Unified CVP Reporting server is co-resident with a Call Server, the Call Server is not processing any SIP or H.323 calls but is simply relaying reporting data from the VoiceXML Server.

The co-resident Call Server does not have a significant impact on performance in this model, therefore the sizing information in the section on the Unified CVP Reporting Server, page 14-4, does not change.

Unified Presence Server

The Cisco Unified Presence server is the SIP Proxy Server provided by Cisco for use with Unified CVP. Table 14-1 outlines the performance of the various server types.

Table 14-1 Call Handling Capacities for Cisco Unified Presence Servers

Cisco Server Model	Recording Function	UDP	TCP
MCS-7825	Record-Route On	200 cps	100 cps
	Record-Route Off	300 cps	300 cps
MCS-7835	Record-Route On	200 cps	100 cps
	Record-Route Off	300 cps	300 cps
MCS-7845	Record-Route On	600 cps	200 cps
	Record-Route Off	1100 cps	500 cps

The capacities in Table 14-1 are measured in calls per second (cps). However, one call coming in from the PSTN is not equivalent to one call through Cisco Unified Presence. Multiple calls are actually generated per inbound customer call for queuing, ringback, and subsequent agent transfers. A typical incoming call will be transferred by Unified CVP four times, so the inbound PSTN call rate should be multiplied by 4.

Example:

> If Unified CVP receives 20 PSTN calls per second, Cisco Unified Presence will see about 80 calls per second.

Unified CVP Reporting Server

There are many variables to take into account when sizing the Unified CVP Reporting Server. Different VoiceXML applications have different characteristics, and those characteristics play a large part in the amount of reporting data generated. Some of these factors are:

- The types of elements used in the application
- The granularity of data required
- The call flow users take through the application
- The length of calls
- The number of calls

To size the Reporting Server, you must first estimate how much reporting data will be generated by your VoiceXML application. The example applications and the tables in subsequent sections of this chapter will help you to determine the number of reporting messages generated for your application.

Once you have determined the number of reporting messages generated by your application, complete the following steps *for each VoiceXML application*:

1. Estimate the number of minutes customers will spend receiving VoiceXML call treatment by that application.
2. Estimate the calls per second that the application will receive.
3. Estimate the number of reporting messages for your application.

Use the following equation to determine the number of reporting messages generated per second for each VoiceXML application:

$$A\# = \%CPS * CPS * MSG / Min / 60$$

Where:

A# = the number of estimated reporting messages per second for an application. Complete one calculation per application (A1, A2, …, A*n*).

CPS = the number of calls per second.

%CPS = the percentage of calls that use this VoiceXML application.

MSG = the number of reporting messages this application generates. To determine the number of reporting messages generated by your application, use the information provided in the sections on Reporting Message Details, page 14-6, and Example Applications, page 14-7.

Min = Amount of time spent in the application (in minutes).

60 = the number of seconds in one minute.

Next, estimate the total number of reporting messages that your deployment will generate per second by summing the values obtained from the previous calculation for each application:

A(total) = A1+ A2+…..+A*n*

This is the total number of reporting messages generated per second by your VoiceXML applications. The Cisco MCS-7845 reporting servers can handle 420 messages per second. If the total number of reporting messages per second for your deployment is less than 420, you can use a single reporting server. If it is greater, you need to use multiple reporting servers and partition the VoiceXML applications to use specific reporting servers.

How to Use Multiple Reporting Servers

If the number of messages per second (as determined in steps 1 and 2 above) exceeds the reporting server capacity, then the deployment must be partitioned vertically.

When vertically partitioning to load-balance reporting data, a Unified CVP system designer must consider the following requirements that apply to deployments of multiple reporting servers:

- Each Call server and each VoiceXML Server can be associated with only one Reporting Server.
- Reports cannot span multiple Informix databases.

For more information on these requirements, refer to the *Reporting Guide for Cisco Unified Customer Voice Portal Release 4.0(1)*, available at

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

When designing Unified CVP deployments with multiple reporting servers, observe the following guidelines:

- Subdivide applications that generate more combined call processing and application messages than are supported by one reporting server.
- VoiceXML can be filtered, and filtering out non-interesting data creates more usable data repositories that support higher message volume.
- Configure the dial plan and/or other available means to direct the incoming calls to the appropriate Call Server and VoiceXML Server.

If you need to combine data from multiple databases, possible options may include:

- Exporting reporting data to Excel, comma separated values (CSV) files, or another format that allows data to be combined out side of the database
- Exporting reporting data to CSV files and importing it into a customer-supplied database
- Extracting data to a customer-supplied data warehouse and running reports against that data

Reporting Message Details

Table 14-2 outlines the various elements or activities and the number of reporting messages generated by each.

Table 14-2 Number of Reporting Messages per Element or Activity

Element or Activity	Number of Reporting Messages (Unfiltered)
Start[1]	2
End[1]	2
Subdialog_start[1]	2
Subdialog_return[1]	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
Play (Audio file or TTS)	2
Get Input (DTMF)	5
Get Input (ASR)	9
Form	10
Digit_with_confirm	20
Currency_with_confirm	20
ReqICMLabel	30

[1] These elements are required in every application and cannot be filtered.

Example Applications

This section presents some examples of applications that can be used to estimate the number of reporting messages that will be generated by your particular application.

Low Complexity

Total: 16 reporting messages per minute per call.

Element Type	Approximate Number of Reporting Messages
Start	2
Subdialog_strart	2
Play element	2
Play element	2
Play element	2
Play element	2
Subdialog_end	2
End	2

Medium Complexity DTMF Only

Total: 39 reporting messages per minute per call.

Element Type	Approximate Number of Reporting Messages
Start	2
Subdialog_strart	2
Play element	2
Get input	5
Play element	2
Get input	5
Form	10
Input	5
Transfer with audio	2
Subdialog_end	2
End	2

Medium Complexity Using Automatic Speech Recognition (ASR)

Total: 51 reporting messages per minute per call.

Element Type	Approximate Number of Reporting Messages
Start	2
Subdialog_strart	2
Play element	2
Get input	9
Play element	2
Get input	9
Form	10
Input	9
Transfer with audio	2
Subdialog_end	2
End	2

High Complexity Using Automatic Speech Recognition (ASR)

Total: 107 reporting messages per minute per call.

Element Type	Approximate Number of Reporting Messages
Start	2
Subdialog_strart	2
Icmrequrestlabel	30
Form	10
ASR capture	9
Digit with confirm	20
Form	10
Digit with confirm	20
Subdialog_end	2
End	2