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

As a basis for this exercise, first determine the customer's contact center profile in terms of the number of calls which are in each state, worst case.

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 which are either executing applications using the CVP VoiceXML Server;
- "Queue and Collect" – calls which are in queue for an agent, or which are executing prompt and collect type self service applications; and
- "Talking" – calls which are connected to agents or to third party TDM VRU applications.

In counting the number of calls which are in the Talking state, count only calls which are using 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 CVP resource, because 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 which are tromboned to a TDM target, using both an incoming and an outgoing TDM port on the same gateway or on a different gateway (ie., toll bypass). Calls which are transferred to VRUs or agents in this manner should be counted as "Talking" calls.

However, if the call was transferred via *8 TnT, hook flash, TBCT, or an ICM NIC, neither the gateway nor CVP play any role in the call. Both components have reclaimed their resources. Such calls should not be counted as "Talking" calls.

Finally, include in the overall call counts those calls that have been transferred back into 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.

You will also notice that the definitions of these call states differ somewhat from the similar definitions used for port licensing purposes (see Chapter 15, "Licensing"). The use of ASR or 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 IPCC agents or ACD agents, nor does it matter whether the customer intends to use 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, one more piece of information is required:

- Busiest period call arrival rate, in terms of calls per second

You will need this information on a contact-center-wide basis. You can use statistical means to arrive at this number, because you will never be able to identify a true maximum arrival rate. Except in fairly small implementations, this is seldom the critical factor in determining sizing.

Armed with the above data, you can now begin sizing each component in the network. This section first considers the CVP products – Call Server and VoiceXML Server, followed by the gateways, gatekeepers and content switches. Note that this section deals entirely with the number and type of physical components required to support the customer's system. It does not include any redundancy. For an understanding of how to extend these numbers to support higher reliability, please see Chapter 5, "Designing CVP for High Availability".

Note Unless otherwise noted, this entire Component Sizing chapter applies to all deployment models, including Model #1, Standalone Self-Service.

CVP Call Server

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

CVP Call Servers are sized according to the number of call legs they can handle, in addition to their maximum call arrival rate. A call might have up to two legs: one for call control activity ("switch leg"), and one for VRU activity ("VRU leg").

Each CVP Call Server can handle 400 call legs. Therefore, the number of calls each CVP Call Server can handle depends on the percentage of calls which are using two call legs. In the case of Deployment Model #3, CVP with ICM-Controlled Switching, all cases use only a switch leg. Therefore, each CVP Call Server can handle the full 400 calls. In the case of Deployment Model #4, IVR/Queuing via ICM with NIC-controlled routing, all calls use only a VRU leg. Each CVP Call Server can therefore again handle the full 400 calls. In Models #3a and #3b, the capacity depends on the percentage of calls which are in talking state.

Each CVP Call Server is further limited to a 3 CPS call arrival rate. However, Model #4 is exempt from this limitation because the CVP Call Server in this model does not perform any H.323 processing.

Specifically, the number of CVP Call Servers required is:

```
(2*Self Service + 2*Queue and Collect + Talking) / 400, rounded up
```

or

```
call arrival rate / 3, rounded up (except in Model#4)
```

whichever is larger.

CVP VoiceXML Server

CVP VoiceXML Server sizing is simple. One VoiceXML Server can handle up to 500 calls. If you are using CVP VoiceXML Server, you should size those machines according to this formula:

```
calls / 500, rounded up
```

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

Gateway

This section covers the following topics:

Choice of Gateway

CVP uses gateways for two purposes: TDM ingress and VoiceXML rendering. Any Cisco gateway which is supported by CVP can usually be used for either purpose or both. However, depending on your Deployment Model, you might need only one or the other:

- Model #1, Standalone Self-Service: All calls use both ingress and VoiceXML
- Model #2, CVP with ICM-Controlled Switching: All calls use ingress only
- Model #3a, DTMF Menu, Prompt and Collect: All calls use ingress, some calls use VoiceXML
- Model #3b, ASR/TTS and/or complex IVR application: All calls use ingress, some calls use VoiceXML
- Model #4, IVR/Queuing via ICM with NIC-controlled routing: All calls use VoiceXML only

In cases where both ingress and VoiceXML are required, you can choose to run both functions on the same gateways, or you can choose to designate some gateways for ingress and others for VoiceXML.

Following are some guidelines for determining whether they should be combined or split:

- In classical branch office deployments, where the call needs to be queued at the branch where it arrived, ingress and VoiceXML functions must always be combined.
- In cases where a large number of non-CVP PSTN connections will share the gateways, it is recommended to dedicated Ingress for that purpose, and use separate VXML gateways.
- AS5850eRSC and Cisco Catalyst 6500 Communication Media Module (CMM) can be used for ingress only; they cannot be used for VoiceXML.
- VXML-only gateways are less costly, since they do not require DSP farms or TDM cards. Use a spreadsheet to determine which way you obtain the best price.

- With relatively low call volume, it is usually better to combine them for redundancy purposes. Two combined gateways are better than one of each, since the loss of one gateway still allows calls to be processed, though at a lower capacity.

The next decision is whether to use ISR gateways (2xxx or 3xxx) or the AS5xxx gateways. Guidelines state that ISR gateways should only be used in branch office sites, whereas AS5xxxx gateways should be used in centralized data center sites.

Sometimes it is difficult to determine what constitutes a *branch office*, and therefore which gateway should be used. The following guidelines might help:

- The classical definition of branch offices, for which you should use ISR gateways, includes:
  - Multiple sites where TDM calls will be arriving from the PSTN
  - Those sites are separated from the data centers where most of the CVP equipment lie
  - You will be placing only one gateway at each site
- If you have sites where you will be stacking multiple gateways for any reason, then those sites are data center sites and should use 5xxx gateways.
- Anything which does not fit these two descriptions: use your best judgement.

Finally, note that IP-IP Gateway functionality is not a supported configuration with CVP. None of the deployment models described in Chapter 3, "Deployment Models" require this functionality, but some customers have very complex call flows which do not clearly fit those models, and some customers want to integrate with non-Cisco VoIP in a way where IP-IP gateway might be helpful.

Note Refer to technical specifications for the AS5xxx series gateways at: http://www.cisco.com/en/US/products/hw/iad/index.html and for the Integrated Service Routers (ISRs) at: http://www.cisco.com/en/US/products/hw/routers/index.html.

Gateway Sizing

Individual Cisco gateways can handle different call capacities depending on whether they are doing ingress only, VoiceXML only, or a combination of the two. Gateways doing VoiceXML activities also have different call capacities depending on whether or not they are supporting ASR or TTS activities.

In general, gateways performing ingress only can be sized according to the number of TDM cables that can be physically connected. For gateways which are combined or VXML-only, it is important to ensure that the overall CPU usage will be less than 70% on average. The factors affecting CPU usage are:

- Calls per second (cps)
- Maximum concurrent calls
- Maximum concurrent VoiceXML sessions

Before sizing the voice gateways, use the IPCC Resource Calculator to determine the maximum number of trunks (DS0s) and VoiceXML IVR ports needed to support the entire solution.

For almost all CVP deployment models, sizing is based on the Maximum concurrent VoiceXML sessions & VoIP calls, described in Table 4-1.

Table 4-1 Cisco Voice Gateway VXML Session Support

Platform	Dedicated VXML Sever		Voice Gateway and VXML		Memory Recommended
	VXML and DTMF	VXML and ASR/TTS	VXML and DTMF	VXML and ASR/TTS	
Cisco 2801	7	6	6	4	256MB
Cisco 2811	30	24	24	20	256MB
Cisco 2821	48	36	36	30	256MB
Cisco 2851	60	56	56	48	512MB
Cisco 3725	68	50	50	38	512MB
Cisco 3745	100	80	77	60	512MB
Cisco 3825	120	96	96	72	512MB
Cisco3845	150	144	144	96	512MB
AS5400HPX	96	90	90	72	512MB
AS5350XM	240	192	192	192	Default
AS5400XM	240	192	192	192	Default

These numbers assume that the only activities running on these gateways are VoiceXML with basic routing and connectivity. If you intend to run additional applications, such as fax, security, normal business calls, etc., the capacity numbers presented here should be prorated accordingly. These figures apply to IOS version 12.4 (mainline).

Also, note that if you run VXML on one of the 28/37/3800 gateways, additional licenses, FL-VXML-1 or FL-VXML-12 are required.

Table 4-2 should also be consulted in order to ensure that the concurrent call load and call arrival rates do not exceed these rated capacities.

Table 4-2 Maximum Calls on a Platform

Platform	Maximum Calls per Second	Maximum VoIP Calls
2801	1	32
2811	1	80
2821	1.2	128
2851	2	192
3725	2	192
3745	4	384
3825	4	384
3845	8	540
AS5400HPX	7	384
AS5350XM	20	192
AS5400XM	20	648

In addition to these capacities, also consider how much DRAM and flash memory to order. The capacity which comes with the machine by default is usually sufficient for most purposes. However, if your application requires large numbers of distinct .wav files (as with complex self service applications), or if your application has unusually large .wav files (as with extended voice messages or music files), you might want to increase the amount of DRAM in order to accommodate more cache space. .Wav files are recorded at 8kb per second. Additionally, if you plan to use the flash memory itself rather than a media server to house media files, you might want to expand your flash memory order. The use of DRAM for prompt caching is discussed in detail in Chapter 14, "Media File Options."

Using MGCP Gateways

Cisco Customer Voice Portal requires the deployment of H.323 voice gateways. However, customers might require the deployment of MGCP 0.1 Voice Gateways with Cisco CallManager-based IPC deployments. These requirements could be survivability, overlap sending, or simplified deployment.

There are three design considerations to deploy Cisco Customer Voice Portal in this environment:

- Design and Plan a phased migration of each MGCP voice gateway to H.323.
- Implement both MGCP 0.1 and H.323.

  There are some considerations with this option. First, only one signaling protocol might be assigned to a PTSN interface module. This means you need dedicated PSTN interfaces/modules to each CVP and to Cisco CallManager. If calls are being sent over the Wide Area Network, you must examine you Call Admission design. Optimally, have a single call admission control mechanism for all calls as well as the dialplan.
- Deploy a second H.323 voice gateway at each CVP location.

  This means you need additional PSTN lines. If calls are being sent over the Wide Area Network (WAN), you must examine your Call Admission design. Optimally, have a single call admission control mechanism for all calls as well as the dialplan.

Gatekeeper

Gatekeepers are required components in all deployments except Model #1, Standalone Self-Service, and Model #4, NIC-based Call Control with CVP Queue, Collect and Self-Service. A gatekeeper is an H.323 entity on a LAN that provides address translation and control access to the LAN for H.323 terminals and gateways. The gatekeeper can provide other services to the H.323 terminals and gateways, such as bandwidth management and locating gateways.

Generally, when you need to add a gatekeeper for the CVP design, you should use the Cisco 2811 for smaller deployments, the Cisco 3825 for medium deployments, and the Cisco 3845 or 7301 for larger deployments. Again, CPU utilization is the determining factor when sizing gatekeepers. Memory usage is seldom a critical factor, unless you are also using the same gatekeepers to register very large numbers of endpoints outside of the CVP implementation.

The factors that affect CPU utilization are:

- The overall number calls the gatekeeper has to support
- The length of time the gatekeeper has to monitor the call
- The number of H.323 endpoints registering with the gatekeeper
- The call arrival rate

However in almost all cases the call arrival rate will be the determining factor in CVP implementations. Please see Table 4-3.

Table 4-3 Maximum Call Per Second on Router

Router	Max Calls per Second
7301	100
72XX	80
3845	70
3825	60
3745	50
3725	40
2851	30
2821	20
2811	10

Note The 5xxx routers cannot be used as gatekeepers. Also, note that gatekeeper functionality should not be combined on the same router as gateway functionality, except in non-redundant lab environments (and a more expensive IOS feature set is required in this case as well).

For the most current information about the Cisco integrated services routers, refer to the latest product documentation available at the following sites:

- Cisco 2800 Series

  http://www.cisco.com/univercd/cc/td/doc/product/lan/cat2800/
- Cisco 3800 Series

  http://www.cisco.com/univercd/cc/td/doc/product/access/acs_mod/3800/
- Cisco 7301

  http://www.cisco.com/univercd/cc/td/doc/product/core/7301/