Collaboration Solutions Design and Deployment Sizing Considerations

Revised: February 28, 2014; OL-29367-05

This chapter describes system sizing for Cisco Collaboration products and systems. Sizing involves providing an accurate estimate of the required hardware platforms for the system, based on the number of users, traffic mix, and features that the system will provide.

Accurate sizing is critical to ensure that the deployed system will meet the expected service quality for call volumes and throughput. For standalone products, manual calculation of the system size may be feasible (as covered in the section on Sizing for Standalone Products, page 27-48). However, there are many sizing factors to consider in a complex system deployment. For example, multiple products may be distributed across different locations and may include video endpoints, call centers, and voice/video conferencing. Cisco Systems provides a set of sizing tools to handle the resulting complexity.

This chapter provides a general introduction to system sizing methodology and the factors that affect sizing, and also provides information about how to use the sizing tools.

Note This chapter should be read in conjunction with the product descriptions and design and deployment considerations covered in other chapters of this document. A good understanding of both of these aspects is required for a successful deployment.

This chapter includes the following major sections:

What's New in This Chapter

Table 27-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 27-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Unified CM server and cluster capacity limits	Server and Cluster Maximums, page 27-14	February 28, 2014
Capacity planning for mobility	Mobile Unified Communications, page 27-20	February 28, 2014
Cisco Business Edition 6000 capacity limits	Cisco Business Edition, page 27-48	August 23, 2013
Re-correction to conference bridge capacity limit	Table 27-10	August 23, 2013
Correction to conference bridge capacity limit	Table 27-10	June 28, 2013
Cisco Jabber Clients	Cisco Collaboration Clients and Applications, page 27-16	May 24, 2013
Cisco UC Integration™ for IBM Sametime	Cisco UC Integration™ for IBM Sametime, page 27-19	May 24, 2013
Cisco Unified IP Phone Services	IP Phone Services, page 27-25	May 24, 2013
WebEx Meetings Server	Cisco WebEx Meetings Server, page 27-40	April 2, 2013
General reorganization of the chapter, as well as other updates and corrections	All sections of this chapter	April 2, 2013
Collaborative conferencing	Collaborative Conferencing, page 27-38	October 31, 2012
CTI resources	Applications and CTI, page 27-22	August 31, 2012
Sizing information for Cisco Unified Mobility	Cisco Unified Mobility for Cisco Business Edition, page 27-50	June 28, 2012
Sizing information for Unified CM with Cisco Collaboration Clients and Applications	Cisco Collaboration Clients and Applications, page 27-16	June 28, 2012
Sizing information for LDAP directory integration	LDAP Directory Integration, page 27-29	June 28, 2012
Other minor updates for Cisco Unified Communications System Release 9.0	Various sections throughout this chapter	June 28, 2012

Methodology for System Sizing

To ensure accurate system sizing, Cisco follows a methodology that is driven by actual performance test results and that incorporates industry-standard traffic engineering models to estimate the maximum expected traffic that the system needs to handle.

The following sections describe the sizing methodology:

- Performance Testing, page 27-3
- System Modeling, page 27-4
- Traffic Engineering, page 27-5

Performance Testing

Each product performs a set of functions, and each function utilizes a number of resources (such as CPU and memory). Cisco designs and executes performance tests that allow us to measure resource usage accurately for each function at different usage levels.

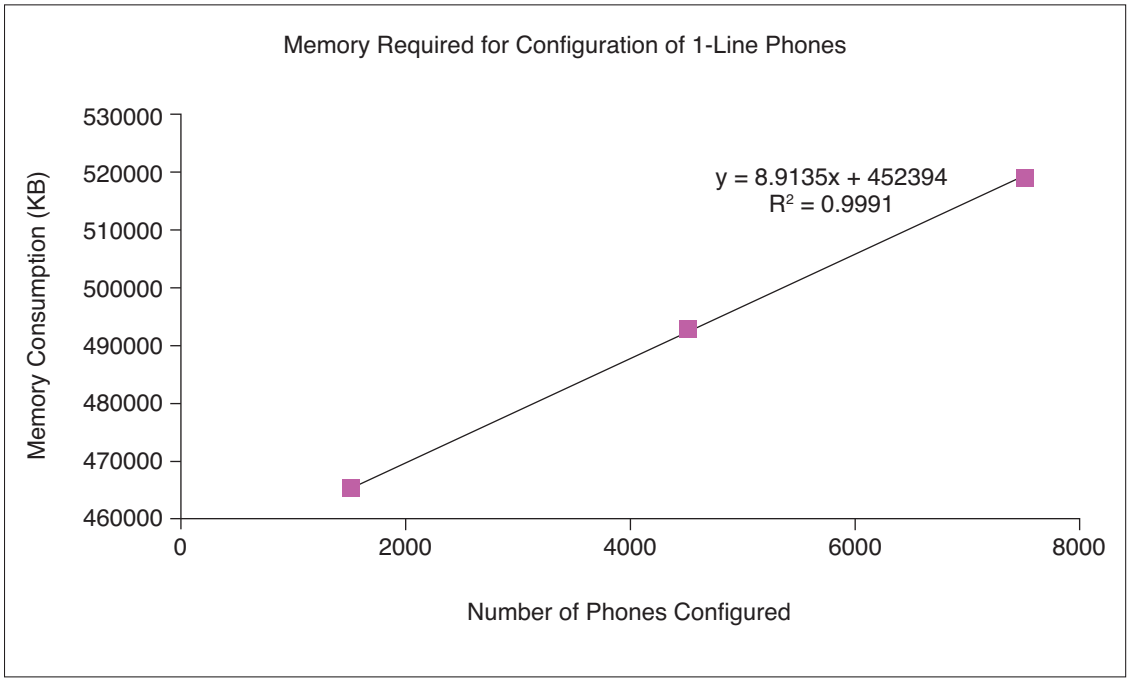
Most systems exhibit linearity within a certain range, beyond which the system performance can become unpredictable. Cisco sets the usage levels for each performance test to identify the linear range of the resource usage for each function. The results for each test can be graphed. If required, Cisco performs additional tests (at intermediate load levels) to confirm the linear section of the graph.

The linear section of the graph can be expressed as a formula to model the resource usage for each incremental addition of work. The R^2 value is a measure of variance between the formula and the measured data. If the R^2 value is close to 1, the formula is a close match for the data.

For example, Figure 27-1 shows the results of a test conducted to determine the memory requirements for configuring single-line IP phones. It shows the memory consumed by configuring 1500, 4500, and 7500 IP phones in Unified CM. The graph shows that the equation of the trend line is linear and can be used to predict the dependent variable (in this case, memory) based on the control variable (the number of phones).

In this particular test, the R^2 value is extremely close to 1. From the equation, we can compute that the memory consumed with no phones is 462,000 Kbytes (the Y-intercept) and that each additional one-line phone configured in the system consumes an additional 8.91 Kbytes.

Figure 27-1 Memory Required for Configuration of One-Line Phones



System Modeling

Cisco uses the performance test results to create a system model. A system model is a mathematical model that calculates the maximum resource usage for a specified set of features, endpoints, and traffic mix, which are provided as inputs to the model.

To develop a system model for a given product, Cisco performs the following steps:

1. Itemize all of the functions that the product performs. Identify variations of the function that need to be tested. For example, each type of call will potentially use a different amount of the measured resources.
2. Determine the resources of interest. Generally this includes memory and CPU. Specific products may have additional resources that impact system sizing.
3. Run the performance tests (as described in the previous section) to determine the resource usage for each function.
4. For each function, use the linear range to define the formula for resource usage.

We may need to repeat these steps a number of times because other factors (such as software release, call mix, and types of endpoints) can impact resource usage.

The system model for the product consists of aggregating the formulas for each function supported by the product. The model can be fairly simple for some products, but it can be very complex for a product that supports multiple functions, multiple endpoint types, and multiple call types.

Specific considerations for memory and CPU resource types are described in the following sections.

Memory Usage Analysis

The system model differentiates between static and dynamic memory, which have different usage characteristics. There is also system memory, which is reserved for the operating system and other processes. These three memory types are described in the following sections:

Static memory

Static memory is consumed even when there is zero call traffic. Static memory usage includes the data for system configuration and the data for registered endpoints. Static memory also includes configuration for the dial plan (which covers items such as partitions, translation patterns, route lists and groups). Static memory also includes the memory used for CTI and other applications. In a large system, static memory is mainly a function of the number of configured endpoints and the size of the dial plan.

Note that each type of endpoint may consume a different amount of memory. Memory usage may also depend on the device protocol (SIP or SCCP), the number of line appearances, security capabilities, and other factors. Each of these variants must be measured and incorporated into the model.

Dynamic memory

Dynamic memory is used for transient activities, such as saving the context of each active call. In a large system, dynamic memory is primarily a function of the number of concurrent calls.

The number of concurrent calls is determined by the average call holding time (ACHT). A longer ACHT results in more dynamic memory use because there will be a larger number of concurrent active calls.

Memory usage may vary considerably for different types of calls and different protocols (such as SCCP and SIP).

System memory

System memory is required by the operating system (OS) and by other processes and services. In addition, some memory may be reserved for transient spikes in usage. System memory reduces the amount of memory available for applications running on the platform.

CPU Usage Analysis

An inactive system exhibits some CPU activity, but most of the CPU utilization occurs during transaction processing, such as setting up and tearing down calls. Therefore, one of the key determinants of CPU usage is the offered call rate.

CPU usage can vary considerably depending on the type of calls. Calls can originate and terminate within the same server, or they can originate and terminate on two different servers or clusters. Calls can also originate from the Unified CM cluster and terminate to a PSTN gateway or trunk.

CPU usage analysis must account for the different cost of a call originating versus terminating on Unified CM, the protocols in use, and whether security features are enabled. CPU usage also depends on factors such as the configuration database complexity and whether CDRs or CMRs are being generated.

CPU usage will vary substantially depending on the actual hardware platform. Therefore, the same performance tests must be repeated on all platforms that are supported for each product.

CPU usage is also affected by CPU-intensive call operations such as call transfers, conferences, and media resource functions such as MTP or music on hold. Shared lines consume additional CPU resources, because each call to a shared line is offered to all of the phones that share the line.

Traffic Engineering

Cisco uses industry-standard traffic engineering models to estimate the dynamic load on the system.

Traffic engineering provides mathematical models that calculate the maximum traffic level expected for a set of users. The models also determine the amount of a shared resource (such as PSTN trunks) that is required to support a given traffic load.

The following sections describe traffic engineering considerations for different types of traffic:

Definitions

Traffic engineering defines the following terms:

Maximum Simultaneous Calls

The maximum number of simultaneous active calls that the system can handle at one time.

Calls per Second

The number of new call attempts that arrive at the system in one second. This unit can be used to define the average calls per second that the system expects to receive across a busy hour (this number would be equivalent to calls-per-hour divided by 60).

This unit can also be used to define the maximum burst of traffic that the system needs to handle.

Busy Hour

The hour in a given 24-hour period during which the maximum total traffic occurs. This hour varies depending on the organization and the type of traffic. For business voice traffic, the busy hour is often in the morning (for example, 10 AM to 11 AM).

Busy Hour Call Attempts (BHCA)

The user BHCA represents the average number of calls that a user initiates or receives during the busy hour. Typically, BHCA will be calculated as the average of the busy hour call attempts from the busiest 30 days of the year). System BHCA is the User BHCA multiplied by the number of users.

Blocking Factor

Indicates a grade of service, expressed as the probability that a call will be blocked during the busy hour due to lack of resources. For example, a blocking factor of 1% indicates that one out of every 100 calls may be blocked due to lack of resources required to process the call.

Average Call Hold Time

This is the average period of time that the resource is busy. For example, on a voice call the ACHT is the period of time between call setup and call tear-down when there is an open speech path between the two parties. A hold time of 3 minutes (180 seconds) is an industry average used for traffic engineering of voice systems.

Erlang

The Erlang is a measure of traffic load on a system. To calculate Erlangs, multiply calls per hour by the average holding time (in hours). Resource requirements can be derived from Erlangs by using the appropriate Erlang model.

The number of Erlangs handled by a resource (such as a trunk group) is equal to the number of simultaneous calls. The Erlang value is usually averaged over a one-hour period of time.

Erlang B Model

The Erlang B model can determine the number of trunks required to handle a traffic load (in Erlangs) with a specified blocking factor. The Extended Erlang B model includes the modeling of retries (for calls that are blocked). The retry percentage is an additional input to the Extended Erlang B model.

Erlang C Model

The Erlang C model incorporates queuing of incoming calls, and is therefore very useful for modeling call center traffic.

Bursty Traffic

Traffic models assume a fairly steady arrival rate for the call attempts, which is a valid assumption for a large number of subscribers acting independently. However, in a real system, a number of calls could arrive over a very short period of time. Such a traffic burst will consume the system resources very quickly, and can result in a high number of blocked calls. Products may specify the size and duration of traffic bursts that they can handle.

Voice Traffic

Standard voice traffic is characterized by specifying the busy hour call attempts (BHCA) and the average call holding time (ACHT). For example, if the system BHCA is 200 and the average call duration is 3 minutes, the system is being used for a total of 600 minutes, which is 10 Erlangs.

To calculate the usage of a shared resource (such as a PSTN trunk group), the blocking factor must also be specified. For example, given an Erlang value and the blocking factor, we can use an Erlang calculator or lookup tables to calculate the number of voice circuits that will be required on PSTN gateways.

Table 27-2 illustrates the relationship between number of trunks, blocking probability, and Erlangs of traffic.

Table 27-2 Erlang B Traffic Table (Number of Circuits Required)

Number of Erlangs	Blocking Probability					
	0.05%	1%	2%	3%	4%	5%
10	19	18	17	16	15	15
20	32	30	28	27	26	26
30	44	42	39	38	37	36

From Table 27-2 we can determine the following information:

- Given an Erlang requirement of 20 and a blocking factor of 1%, the system will need 30 circuits.
- Additional circuits are required to provide a lower blocking factor (such as 1%) than to provide a higher blocking factor (such as 5%).

Contact Center Traffic

Contact centers demonstrate a unique pattern of traffic, because these systems typically handle large volumes of calls that are handled by a small number of agents or interactive voice response (IVR) systems. Contact centers are engineered for high resource utilization, therefore their agents, trunks, and IVR systems are kept busy while they are in operation, which usually is 24 hours a day. Call queuing is typical (when incoming call traffic exceeds agent capacity, calls wait in queue for the next available agent), and the agents are usually dedicated during their work shifts to taking contact center calls.

Average call holding times for contact centers are often shorter than for normal business calls. Many calls interact only with the IVR system and never need to speak to a human agent. These calls are known as self-service calls. The average holding time for self-service calls is about 30 seconds, while a call serviced by an agent may have an average holding time of 3 minutes (the same as normal business traffic), making the overall average holding time in the contact center shorter than for normal business traffic.

The goal of contact center management is to optimize resource use (including IVR ports, PSTN trunks, and human agents), therefore resource utilization will be high.

A call center usually has a higher call arrival rate than a typical business environment. These call arrival rates can also peak at different times of day (not the usual busy hours) and for different reasons than normal business traffic. For example, when a television advertisement runs for a particular holiday package with a 1-800 number, the call arrival rate for the system will experience a peak of traffic for about 15 minutes after the ad airs. This call arrival rate can exceed the average call arrival rate of the contact center by an order of magnitude.

As noted earlier, contact center sizing uses the Erlang C model to account for calls waiting in queues. Contact centers require additional resources, such as interactive voice response (IVR) ports. The time that calls wait in queues needs to be factored in when sizing the PSTN gateways (see Gateway Sizing for Contact Center Traffic, page 27-33).

Note For additional information about Cisco Unified Contact Center deployments, refer to the *Cisco Unified Contact Center Enterprise SRND*, available at http://www.cisco.com/en/US/products/sw/custcosw/ps1844/products_implementation_design_guides_list.html.

Video Traffic

Point-to-point video traffic demonstrates similar characteristics to its voice equivalents for call arrival rates, peak usage times, and call durations. Also, signalling for call setup and take-down is similar to voice calls.

Video traffic requires significantly higher network bandwidth than voice because the payload in video packets is much larger than in voice packets. Also, video traffic can be much burstier than voice. Voice packet sizes are usually fairly consistent (specifics depend on the encoding algorithm in use), whereas video frames can vary considerably in size, depending on how much change has occurred since the previous frame. The resulting RTP packet stream can therefore exhibit bursts of traffic.

Implications for video conferencing are covered in the next section.

Conferencing and Collaboration Traffic

Conferencing traffic has considerably different characteristics than point-to-point voice/video calls. The traffic model for conferencing traffic needs to accommodate the following differences:

- Call arrivals

  A traditional traffic model assumes a Poisson distribution of busy-hour call arrivals throughout the busy hour. However, most participants join their conference call within 5 to 10 minutes of the meeting start time, and most conference calls are scheduled to start at the beginning of the hour. Therefore, the call arrival rate will exhibit a single burst at the top of the hour rather than a Poisson distribution throughout the hour.

- Peaks

  Business voice traffic typically has a distinct peak in the morning (between 10:00 and 11:00 AM) and another peak in the afternoon (between 1:00 and 2:00 PM). However, conference facilities are generally a limited resource, resulting in meetings that are distributed more evenly throughout the business day, with less of a pronounced peak at peak times.

- Call durations

  The average business voice call duration is 3 minutes. The average conference call duration may be closer to 50 minutes (depending on the mix of 30 minute, 60 minute, and longer meetings).

- Video conferencing

  Specialized equipment is required to provide the switching or combining of video streams. Therefore, expected usage of video endpoints is an important factor in the model.

System Sizing Considerations

For large and complex deployments, the system designer will need to consider a number of design and deployment factors that influence system sizing. These factors are described in the following sections:

- Network Design Factors, page 27-9
- Other Sizing Factors, page 27-10

Network Design Factors

Solution sizing is affected by the following network design factors:

- Cluster sizes

  A major design decision is whether to create a large centralized Cisco Unified CM cluster or to create a cluster at each major location. The central cluster may have a higher utilization, but you may be forced into a second cluster if a cluster limit is exceeded.

  Some system limits are not absolute and can change dynamically based on the sizing of other services configured in the system.

- Interaction between individual products

  Unified CM plays a central role in most Cisco Collaboration deployments, and it is affected by other components in the system. For example, the addition of Cisco Unified MeetingPlace will tend to concentrate a large number of call setups into a short period (at the beginning of conferencing sessions). Depending on the other functions covered by Unified CM, this may require the addition of Unified CM servers.

- Server capabilities

  Each type of server or router supports different capabilities. For example, some server platforms might not support clustering, and other platforms might not support the redundancy features that you require.

  As another example, different models of Cisco Integrated Service Routers (ISR) have restrictions on the number and types of network modules or Services Ready Engine (SRE) modules they can host.

- Optional capabilities and features

  The system sizing can be impacted if you enable options such as call detail recording (CDR) or call management record (CMR) generation.

Other Sizing Factors

The following additional factors also affect system sizing:

- Mix of call types:

  There are variations in resources consumed by each call type: calls between phones in the same server, calls between two servers in the same cluster, calls between two clusters, and calls that flow to and from the PSTN. Even calls from different types of phones and gateways are different, depending on the protocol and services such as video.

- Mix of endpoint types

  The expected number of phones and users is another example of an obvious factor that affects sizing. Here again, the type of phones, the number of lines configured on them, and whether they are in secure mode, among other things, have an impact on system sizing.

- System release

  System resource usage can vary between system releases. Sometimes, new capabilities in a release can cause an increase in resource usage. In other cases, software improvements can result in a decrease in resource usage.

- Use of external applications

  External applications can communicate with the call processing agent by using an interface such as CTI. This load needs to be factored into the system sizing.

- Anticipated system growth

  If system usage is expected to grow in the next year or two, it would be preferable to build that growth into the original system rather than face a potentially disruptive upgrade in the near future.

- Average and peak usage

  Ensure that the system sizing is based on a realistic view of peak usage. If the peak is underestimated, the system could experience service degradation or equipment outages when the actual peak traffic is encountered.

Because of all the factors and possible variations, the accurate sizing of a large system deployment is a complex undertaking. For this reason, Cisco strongly recommends using the system sizing tools described in the following sections.

Sizing Tools Overview

Cisco provides several sizing tools to assist with accurate solution sizing. The sizing tools are available at the following location (only Cisco employees and certified partners can access this site):

http://tools.cisco.com/cucst

Cisco recommends that you use the sizing tools to perform system sizing. These tools take into account data from performance testing, individual product limits and performance ratings, advanced and new features in product releases, design recommendations from this SRND, and other factors. Based on input provided by the system designer, the tools apply their sizing algorithms to the supplied data to recommend a set of hardware resources.

If you do not have access to the sizing tools, please contact your Cisco account representative or Cisco partner integrator to obtain system sizing information.

Tool-specific sections below contain explanations of the inputs required for the tool and how the inputs can best be collected from an existing system or estimated for a system still in the design stage. Obviously, the sizing recommendations generated by the tools are only as accurate as the input data you provide.

Cisco provides the following sizing tools:

- Cisco Unified Communications Sizing Tool

  This tool guides the user through a complete system deployment. The tool covers the following products and components:

  - Cisco Unified Communications Manager (Unified CM)
  - IM and Presence services
  - Voice messaging
  - Conferencing
  - Gateways
  - Cisco Intercompany Media Engine (IME)
  - Cisco Unified Communications Management Suite
  - Cisco Unified Contact Center components

- Cisco Unified Communications Manager Session Management Edition (SME) Sizing Tool

  This is a specialized tool that focuses on the specific functions of a Unified CM Session Management Edition deployment.

- Cisco VXI Sizing and Configuration Tool

  This is a specialized tool for sizing the Cisco Virtual Experience Infrastructure (VXI).

For more information on these tools and their access privileges, refer to the *Unified Communications Sizing Tool Frequently Asked Questions (FAQ)*, available at

http://tools.cisco.com/cucst/help/ucst_faq.pdf

Caution If any parameter of your system design exceeds the range of values that the above sizing tools allow you to enter, consult your Cisco account team or a Cisco Systems Engineer (SE) about your design before proceeding further.

Using the SME Sizing Tool

The Session Management Edition (SME) is a Unified CM operating in a specific deployment mode. In a pure SME deployment, call traffic runs only across trunk interfaces and the SME hosts no line interfaces.

An SME cluster follows the same topology as a regular Unified CM cluster. A publisher server provides the master configuration repository. A TFTP server may be co-resident with the publisher server if the number of phones or MGCP gateways in the cluster is relatively small. A redundancy ratio of 1:1 is recommended for call processing subscribers.

To size an SME cluster, you must consider the functionality that it is expected to perform. In a base configuration, the SME acts as a routing aggregation point for a number of leaf clusters. It also provides centralized PSTN access for all of the leaf clusters connected to it. In more advanced configurations, the

SME may also host centralized voice messaging, mobility, and conferencing services. The performance of the SME is influenced by the type of trunk protocols that the leaf clusters use to connect to it and by the BHCA across those trunks.

The SME sizing tool requires the following input parameters:

- The various types of trunk interfaces that the cluster services. The following trunk protocols are supported by the SME:
  - SIP
  - H.323
  - MGCP (Q.931)
  - SIP (Q.SIG)
  - H.323 Annex M1
  - MGCP (Q.SIG)
- The number of users that access SME cluster services through each type of trunk interface
- BHCA per user for each trunk interface to leaf clusters for intercluster calls
- BHCA per user for each trunk interface to leaf clusters for off-net (PSTN) calls
- The type of trunk interface used by the SME cluster to connect to the PSTN
- Average holding time for calls
- Number of route and translation patterns

If the SME acts as a service aggregation point, you must consider the following additional sizing parameters:

- For centralized voice messaging, the percentage of calls that are sent to voice mail
- For mobility, the number of users and the remote destinations per user
- For conferencing service, the conferencing dial-in interval

The performance of the SME is measured as calls-per-second across each pair of protocols. There are variations across the hardware platforms and software versions.

Using the VXI Sizing Tool

Cisco Virtualization Experience Infrastructure (VXI) is a systems approach that unifies virtual desktops, voice, and video, to provide a superior virtual workspace experience. The Cisco VXI Sizing Tool assists with the task of sizing components for a Virtualization Experience Infrastructure solution.

Using the Cisco Unified Communications Sizing Tool

Cisco Unified Communications Sizing Tool covers sizing for a number of products and components. For a complete list of components and versions supported by tool, see the release notes that are included in the sizing tool installation package.

The following sections describe the significant factors that influence sizing of the individual products and also how these individual products can influence the sizing considerations of other products in the system deployment:

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is the hub of any Unified Communications deployment. It performs key functions such as controlling endpoints, routing calls, enforcing policies, and hosting applications. Unified CM provides coordination for the other Unified Communications products such as PSTN gateways, Cisco Unity Connection, Cisco Unified MeetingPlace, and Cisco Unified Contact Center. The coordination function has an impact on Unified CM performance, and therefore must be accounted for in Unified CM sizing.

A number of factors affect Unified CM performance and must be considered when sizing a Unified CM deployment. These factors are described in the following sections:

Server and Cluster Maximums

The sizing tool applies the following server and cluster maximums. These values can vary depending on Unified CM software versions, hardware platforms, and OVA templates.

- Within a cluster, a maximum of 8 call processing subscriber nodes can be enabled with the Cisco CallManager Service. Other servers may be used for more dedicated functions such as publisher, TFTP subscribers, and media resource subscribers.
- Each cluster can support configuration and registration for a maximum of 40,000 secured or unsecured SCCP or SIP endpoints.
- A cluster consisting of server node instances running on VMware can support different capacities depending on the OVA template that is chosen. For most Cisco MCS server classes, there is a corresponding OVA template that provides a Unified CM instance with the same capacities (number of phones, gateways, locations, regions, CTI connections, and so forth) as the MCS server class. Because multiple virtual machine instances can run on the same blade or server, the total capacity on a blade or server can therefore be higher than on an MCS server.

Deployment Options

The following deployment options are overall settings that affect all operations in the system, and they are independent of how many endpoints are registered or how many calls are in progress.

Database Complexity

The CPU usage is considerably higher when the configuration database in Unified CM is considered to be complex. There is no one metric to determine whether the database is simple or complex. As a general rule, the database is complex if you have configured more than a few thousand endpoints and more than a few hundred dial plan elements such as translation and route patterns, hunt pilots, and shared lines.

Number of Regions and Locations

Configuration of regions and locations in the Unified CM cluster requires both database and static memory. The number of gateways that can be defined in the cluster is also tied to the number of locations that can be defined. Table 27-3 lists these limits for some of the Unified CM server platforms.

Table 27-3 Maximum Number of Regions, Locations, Gateways, and Trunks

Server Platform	Maximum Number of Regions	Maximum Number of Locations	Maximum Number of Trunks and Gateways
MCS 7816	100	100	110
MCS 7825	1,000	1,000	1,100
MCS 7835 or Open Virtualization Archive (OVA) equivalent	1,000	1,000	1,100
MCS 7845 or OVA equivalent	2,000	2,000	2,100

Whether or not you can actually define the maximum number of locations and regions in a cluster depends on how "sparse" your codec matrix is. If you have too many non-default values in the inter-region codec setting, you might not be able to scale the system to its full capacity for regions and locations. As a general rule, the change from default should not exceed 10% of the maximum number.

Call Detail and Call Management Records

Generation of call detail records (CDR) and call management records (CMR) places a heavier burden on the CPU.

High Availability

After you determine the minimum number of servers required for the specified deployment, add the desired number of additional subscriber servers to provide redundancy. Redundancy options are described in the chapter on Collaboration Deployment Models, page 10-1. Note that some of the server models are better suited for redundant configuration than others.

Number of Servers per Cluster

You can configure a regular cluster with up to four subscriber pairs. In a distributed topology, there may be multiple clusters even when none of the clusters has reached the maximum.

For a centralized topology, there is generally one cluster unless the server limit is reached. Note that other system limits might force a new cluster even if the per-server utilization is not at the limit.

Choice of Servers and UCS Platforms

Unified CM is supported on a variety of Cisco Media Convergence Server (MCS) and Unified Computing System (UCS) platforms. For defining the Unified CM Virtual Machine on a UCS platform, Cisco provides Open Virtualization Archive (OVA) templates that can be loaded onto a hypervisor. Different templates specify different capacities. For example, the 10000 template defines a virtual machine that has a maximum capacity of 10,000 endpoints. There are also templates defined to support a maximum of 1,000, 2,500, and 7,500 endpoints.

The formal definitions of the OVA templates for Unified CM and other Unified Communications products are available at the following location:

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

Note Choice of placement of virtual machines running Unified CM and other Unified Communications products can have an impact on performance and availability. For a discussion of these and other considerations for Unified Communications on UCS deployments, refer to the documentation at http://www.cisco.com/go/uc-virtualized.

Endpoints

The number of endpoints are an important part of the overall load that the system must support. There are different types of endpoints, and each type imposes a different load on Unified CM. Endpoints can be differentiated by:

- Digital (IP) or analog (using an adaptor)
- Software-based or hardware
- The protocol supported (SIP or SCCP)
- Whether the endpoint is configured with security
- Dialing modes (en-bloc or overlap)
- Audio only or audio and video
- Other devices such as gateways (H.323 or MGCP)

Each endpoint configured in the system uses system resources (such as static memory) just by being defined and registered. The endpoint consumes CPU and dynamic memory based on its call rate.

An endpoint can also place additional load on the Unified CM by running applications such as CTI that interact with services running in the Unified CM.

Table 27-4 shows the maximum number of endpoints supported by different server types. Note that these values are guidelines only. A given system may support less than these maximum amounts because of other applications included in the deployment.

Table 27-4 Maximum Number of Endpoints Per Server Platform or OVA Template

Server Platform Characteristics	Maximum Endpoints per Server or OVA Template
Cisco MCS 7845-I3 or OVA equivalent	10,000
Cisco MCS 7845 (All other supported models) or OVA equivalent	7,500
Cisco MCS 7835 (All supported models) or OVA equivalent	2,500
Cisco MCS 7825 (All supported models) or OVA equivalent	1,000
Cisco MCS 7816 (All supported models)	500

Cisco Collaboration Clients and Applications

Cisco Collaboration Clients include the following software applications that run on user desktops or other access devices:

- Cisco Jabber Clients, page 27-17
- Cisco WebEx Connect, page 27-18
- Cisco UC Integration™ for IBM Sametime, page 27-19
- Cisco UC Integration™ for Microsoft Lync, page 27-19
- Third-Party XMPP Clients and Applications, page 27-19

In addition, the following client provides an integrated telephony client with virtualized desktop access:

- Cisco Virtualization Experience Clients, page 27-19

Cisco Jabber Desktop Client

Cisco Jabber provides the underlying services layer for several clients, including Cisco Jabber Clients for Windows and Mac and Cisco UC Integration™ for Microsoft Lync.

The Jabber Desktop Client provides two modes of operation, each of which uses different resources in Unified CM. When it operates in softphone mode, the Jabber Client acts as a SIP registered endpoint and contributes to the total number of endpoints in the system. When it operates in desk phone mode, the Jabber Client acts as a CTI agent and therefore uses CTI resources on Unified CM.

Users may switch the Jabber-based clients to work in either mode. Therefore, it is necessary to properly account for the system resources needed for the anticipated usage.

The following additional items must be considered for a Jabber Desktop Client deployment:

- Device Configuration

  When configured in softphone mode, a Jabber Desktop Client configuration file is downloaded through TFTP or HTTP to the client for Unified CM call control configuration information. In addition, any application dial rules or directory lookup rules are also downloaded through TFTP or HTTP to Jabber Desktop Client devices.

  The Jabber Desktop Client uses the Cisco CallManager Cisco IP Phone (CCMCIP) or UDS service to gather information about the devices associated with a user, and it uses this information to provide a list of IP phones available for control by the client in deskphone control mode. The Jabber Desktop Client in softphone mode uses the CCMCIP or UDS service to discover its device name for registration with Unified CM.

- Deskphone Mode

  When configured in deskphone mode, the Jabber Desktop Client establishes a CTI connection to Unified CM upon login and registration to allow for control of the IP phone. Unified CM supports up to 40,000 CTI connections. If you have a large number of clients operating in deskphone mode, make sure that you evenly distribute those CTI connections across all Unified CM subscribers running the CTIManager service. This can be achieved by creating multiple CTI Gateway profiles, each with a different pair of CTIManager addresses, and distributing the CTI Gateway profile assignments across all clients using deskphone mode.

- Voicemail

  When configured for voicemail, the Jabber Desktop Client updates and retrieves voicemail through an IMAP or REST connection to the mailstore.

- Authentication

  Client login and authentication, contact profile information, and incoming caller identification are all handled through a query to the LDAP directory, unless stored in the local Jabber Desktop Client cache.

- Contact Search

  There are several contact sources that can be used with the Jabber Desktop Client. For example, the UDS service can be used by clients to search for contacts in the Unified CM User database. Alternatively, LDAP integration can be used. If the requested contact cannot be found in the local Jabber Desktop Client cache, UDS or LDAP contact searches take place.

Cisco Jabber Clients

When designing and sizing a solution for Cisco Jabber Clients, you must consider the following scalability impacts for all the components:

- Client scalability

  The Cisco IM and Presence Service hardware deployment determines the number of users a cluster can support. The Cisco Jabber Client deployment must balance all users equally across all servers in the cluster. This can be done automatically by setting the User Assignment Mode Sync Agent service parameter to **balanced**. The maximum number of contacts in the contact list is 200.

- IMAP scalability

  The number of IMAP or IMAP-Idle connections is determined by the messaging integration platform. For specific configuration sizing, refer to the Cisco Unity Connection product documentation available at http://www.cisco.com.

- Audio, video, and web conferencing

  Clients can access the conferencing services that are provided in your network. You need to account for these users when sizing the number of concurrent participants for these services. For additional information, refer to the chapter on Cisco Collaboration Services, page 22-1.

- Cisco Jabber Video for TelePresence

  This client impacts the sizing of the Cisco TelePresence Video Communication Server (Cisco VCS). For more information on Cisco Jabber Video for TelePresence, refer to the documentation at

  http://www.cisco.com/en/US/products/ps11328/tsd_products_support_series_home.html

The Cisco Jabber Clients interface with Unified CM. Therefore, the following guidelines for the current functionality of Unified CM apply when Cisco Jabber Client voice or video calls are initiated:

- CTI scalability

  In Desk Phone mode, calls from Cisco Jabber Clients use the CTI interface on Unified CM. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 9-1. You must include these CTI devices when sizing Unified CM clusters.

- Call admission control

  Cisco Jabber Client applies call admission control for voice and video calls by means of Unified CM locations or RSVP.

- Codec selection

  Cisco Jabber Client voice and video calls utilize codec selection through the Unified CM regions configurations.

- Cisco Unified MeetingPlace voice, video, and web collaboration sessions

  See Cisco Unified MeetingPlace, page 22-31.

- Cisco Unity Connection

  See the section on Managing Bandwidth, page 19-30, in the chapter on Cisco Voice Messaging, page 19-1.

Cisco WebEx Connect

A single end-user requires only a 56 kbps dial-up Internet connection to be able to log in to the Cisco WebEx Messenger service and get the basic capabilities such as presence, instant messaging, and VoIP calling. However, for a small office or branch office, a broadband connection with a minimum of 512 kbps is required in order to use the advanced features such as file transfer and screen capture.

For additional information on network and desktop requirements, refer to the Cisco WebEx administrator's guide available at

http://www.webex.com/webexconnect/orgadmin/help/index.htm

The Cisco Unified Communications integrations use Unified CM CTI Manager for click-to-call applications, as well as deskphone control mode with the Cisco Unified Client Services Framework. Therefore, observe the CTI limits as defined in the section on Applications and CTI, page 27-22. When Cisco UC Integration™ for Connect is operating in a softphone (audio on computer) mode, the Cisco Jabber Desktop Client is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Network Requirements

Cisco WebEx Messenger service deployment network requirements are available at:

http://www.webex.com/webexconnect/orgadmin/help/17161.htm

Cisco UC Integration™ for IBM Sametime

With Cisco UC Integration™ for IBM Sametime, instant messaging and presence services are provided by IBM rather than by Cisco Unified Communications services.

Cisco UC Integration™ for IBM uses Unified CM CTI Manager for click-to-dial applications, as well as deskphone control mode with the underlying Cisco Unified Communications services Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 9-1. When Cisco UC Integration™ for IBM is operating in a softphone (audio/video on computer) mode, the client is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Cisco UC Integration™ for Microsoft Lync

Cisco UC Integration™ for Microsoft Lync uses Unified CM CTI Manager for click-to-dial applications and deskphone control mode. Therefore, observe the CTI limits as defined in the chapter on Call Processing, page 9-1. When Cisco UC Integration™ for Microsoft Lync is operating in a softphone (audio on computer) mode, the client is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Third-Party XMPP Clients and Applications

Third-party Extensible Messaging and Presence Protocol (XMPP) clients may be used with both the WebEx Messenger service platform and the Cisco IM and Presence Service. Voice, video, and other collaboration mechanisms (except for instant messaging and chat) are typically not supported with these clients. Depending on their capabilities, these clients may be counted against the device capacities supported by the above products on their servers.

Cisco Virtualization Experience Clients

All Cisco Virtualization Experience Clients are deployed with a Virtual Desktop Infrastructure (VDI) component, while some of the deployments may also contain a Unified Communications component. Capacity planning and datacenter resource utilization for VDI when using the Cisco Virtualization Experience Clients is covered as part of the Virtualization Experience Infrastructure (VXI) sizing. For details, refer to the VXI documentation available at

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns1100/landing_vxi.html

Capacity planning for the Unified Communications components depends on which Virtualization Experience Client is deployed:

- Cisco VXC 2111 and 2112 integrated form factor zero clients are paired with a Cisco Unified IP Phone 8961, 9951, or 9971. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the Cisco Unified IP Phone; therefore, Computer Telephony Integration (CTI) planning guidelines must be followed for each client deployed.

- Cisco VXC 2211 and 2212 standalone form factor zero clients can be deployed as VDI-only or as a fully integrated voice, video, and virtual desktop with a number of different Cisco Unified IP Phones. When deployed in a Unified Communications environment, the Cisco client running in the user's virtual desktop uses the deskphone control mode of the Cisco Unified IP Phone; therefore, CTI planning guidelines must be followed for each client deployed.
- Cisco VXC 4000 software appliance is a software-only VXC deployment option. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the VXC 4000; therefore, CTI planning guidelines must be followed for each VXC 4000 deployed.
- Cisco VXC 6215 thin client running in VDI-only mode follows VDI capacity planning; however, when the VXC 6215 is deployed as a fully integrated voice, video, and virtual desktop, additional Unified Communications capacity must be accounted for. The Cisco client running in the user's virtual desktop uses the deskphone control mode of the Virtualization Experience Media Engine (VXME) running locally on the Linux thin client; therefore, CTI planning guidelines must be followed for each client deployed.

  VXME registers as a SIP line-side registered device on Cisco Unified CM; therefore, for each VXC 6215 thin client running as a fully integrated voice, video, and virtual desktop, a SIP line device and CTI connection is used.

Mobile Unified Communications

Mobility in Unified Communications is multi-faceted. Each of the different aspects of mobile communications consumes different Unified CM resources and must be accounted for both independently and as a part of the whole system. The following sizing considerations apply to mobility, but note that aspects of mobility that do not affect Unified CM are not discussed here.

Cisco Unified Mobility

There are two parameters that are key to Unified CM's capacity to support single number reach (Mobile Connect) and enterprise two-stage dialing (Mobile Voice Access and Enterprise Feature Access). For these functions to work appropriately, users must be enabled for mobility and remote destinations with shared lines must be defined for the users. Table 27-5 shows the limits for users or remote destinations and mobility identities in a cluster consisting of each class of Unified CM server or OVA template when the global service parameter **Matching Caller ID with Remote Destination** is set to **Complete Match**.

Table 27-5 Maximum Number of Mobility Users or Remote Destinations and Mobility Identities per Cluster with Complete Caller ID Matching

Cluster Nodes	Maximum Number of Users Enabled for Mobility per Cluster with Complete Match	Maximum Number of Remote Destinations and Mobility Identities per Cluster with Complete Match
MCS 7845-I3 or 10,000 User OVA	40,000	40,000 (or 10,000 per node)
MCS 7845 (All other models) or 7,500 User OVA	30,000	30,000 (or 7,500 per node)
MCS 7835 or 2,500 User OVA	10,000	10,000 (or 2,500 per node)
MCS 7825 or 1,000 User OVA	4,000	4,000 (or 1,000 per node)

Note The above capacities also apply when the **Matching Caller ID with Remote Destination** service parameter is set to **Partial Match** and the Unified CM version is 9.1(2)SU1 or later. However, if **Partial Match** is set and the version of Unified CM is below 9.1(2)SU1, then the maximum number of mobility-enabled users or remote destinations and mobility identities per cluster is only 15,000 (or 3,750 per node) for all MCS 7845 server models and the 10,000 and 7,500 User OVA templates. Capacities for other MCS server models and OVA templates remain the same whether **Complete Match** or **Partial Match** is set.

Note A mobility-enabled user is defined as a user that has a remote destination profile and at least one remote destination or dual-mode device and a mobility identity configured.

Each remote destination and mobility identity defined in the system affects Unified CM in several ways:

- The remote destination or mobility identity occupies static memory and configuration space in the database.
- Each occurrence uses a shared line with the users primary device, and hence calls to that line use more CPU resources.
- If the remote destination or mobility identity is an external number (such as the user's cell phone or home), then gateway resources will be used to extend the call.

Call Traffic

The quantity and quality of call traffic is a very significant factor in sizing Unified CM.

It is important to differentiate between call types because call origination and termination are considered as distinct events in the half-call model. For endpoints registered on the same server, that server handles both call halves for calls between these endpoints. For calls made between two servers in the same cluster, each of the participating servers will handle either the call origination or call termination. For calls made between endpoints registered on different clusters, each cluster will handle only half of each call. For calls made between an endpoint in a cluster and the PSTN, a PSTN gateway will handle half of the call, and these call types form the basis for sizing the gateways.

For accurate sizing of call traffic, you must consider the following factors:

- Overall Busy Hour Call Attempts (BHCA) per user
- Average Call Holding Time (ACHT) per call
- BHCA from and to the PSTN using MGCP, H.323, and SIP protocols
- BHCA from and to other clusters using H.323 intercluster trunks or SIP protocols
- BHCA from and to other enterprises using Cisco Intercompany Media Engine (IME)
- BHCA within the cluster

Each different type of call takes a different amount of CPU resources to set up. The number of busy hour call attempts determines the CPU usage. CPU requirements vary directly with the call placement rate. The ACHT determines the dynamic memory requirements to sustain calls for their duration. A longer ACHT means that more dynamic memory must remain allocated, thus increasing the memory requirement.

Call traffic can arise from other sources as well. Each time a call is redirected in a transfer or to voicemail, it requires processing by the CPU. If a directory number is configured on multiple phones, an incoming call to that number needs to be presented to all of those phones, thus increasing CPU usage at

call setup time. If advanced features such as the Intercompany Media Engine (IME) are being used, calls made using this technology, and the percentage of these calls that need to be redirected to the PSTN because of call quality, must also be accounted for.

Dial Plan

The dial plan in Unified CM consists of configuration elements that determine call routing and associated policies. In general, dial plan elements occupy static memory space in Unified CM servers. The following dial plan elements impact the amount of memory required:

- Directory numbers
- Shared directory numbers and the average number of endpoints that share the same DN
- Partitions, calling search spaces, translations, and transformation patterns
- Route patterns, route lists, and route groups
- Advertised and learned DN patterns
- Hunt pilots and hunt lists
- Circular, sequential, and broadcast line groups and their membership

There are no hard limits enforced by Unified CM for any of the dial plan elements, but there is a fixed amount of shared system memory available.

Most of the dial plan elements do not have a direct effect on CPU usage. The exception is shared lines, such as hunt lists and line groups. Each shared line multiplies the CPU cost of a call setup because the call is presented to all of the endpoints that share a particular directory number.

Another factor (for a large dial plan) is the space required to store the elements of the dial plan in the Informix Database System. There is a fixed amount of disk space available to hold the entire configuration of Unified CM, and extra-large dial plans can exceed the maximum space. In this case, you might have to break up the dial plan and store it in multiple clusters.

Applications and CTI

In the context of Unified CM, applications are the "extra" functions beyond simple call processing provided by Unified CM. In general these applications make use of Computer Telephone Integration (CTI), which allows users to initiate, terminate, reroute, or otherwise monitor and treat calls. Features such as Cisco Unified CM Assistant, Attendant Console, Contact Center, and others, depend on CTI to function.

Although the high-end server platforms are able to support CTI for all of their registered devices, the lower-end platforms do not scale that high. Table 27-6 lists the maximum number of CTI resources supported by each type of server platform. These maximum values apply to the following types of CTI resources:

- The maximum number of CTI controlled and/or monitored endpoints that can be registered to a Unified CM subscriber node.
- The maximum number of endpoints that a Unified CM subscriber node running the CTI Manager service can monitor or control.
- The maximum number of TAPI/JTAPI application instances that can connect to a Unified CM subscriber node running the CTI Manager service. The TAPI/JTAPI application instances that can connect to a Unified CM subscriber node running the CTI Manager service are sometimes referred as CTI connections.

Note that the numbers for the high end of each class of server equal the number of devices that the class can support.

In addition to native applications provided by Unified CM, third-party applications may also be deployed that use Unified CM CTI resources. When counting CTI ports and route points, be sure to account for the third-party applications as well.

Table 27-6 CTI Resource Limits in Unified CM

Server Platform	Maximum CTI Resources per Server
MCS 7815	150
MCS 7816-I2/I3/I4	400
MCS 7816-I5	500
MCS 7825-I1/I2	800
MCS 7825-I3/I4	900
MCS 7825-I5 and OVA equivalent	1,000
MCS 7835-I1/I2	2,000
MCS 7835-I3 and OVA equivalent	2,500
MCS 7845-I1	2,500
MCS 7845-I2 and OVA equivalent	5,000
MCS 7845-I3 and OVA equivalent	10,000

In addition to the maximum number of connections and devices, CTI limits are also influenced by:

- The number of lines on each of the controlled devices (up to 5 lines per controlled device)
- The number of shared occurrences of a line controlled by CTI (up to 5 per line)
- The number of active CTI applications (up to 5 for any device)
- A maximum of 6 BHCA per controlled device

The CTI resources available on Unified CM are reduced if any of these values is exceeded.

Determining CTI Resources Required for a Unified CM Cluster

Use the following steps to determine the required number of CTI resources for a Unified CM cluster.

Step 1 Determine the total CTI device count.

Count the number of CTI devices that will be in use on the cluster.

Step 2 Determine the CTI line factor.

Determine the CTI line factor of all devices in the cluster, according to Table 27-7.

Table 27-7 CTI Line Factor

Number of Lines per CTI Device	CTI Line Factor
1 to 5 lines	1.0
6 lines	1.2
7 lines	1.4

Table 27-7 CTI Line Factor (continued)

Number of Lines per CTI Device	CTI Line Factor
8 lines	1.6
9 lines	1.8
10 lines	2.0

Note If there are multiple line factors for the devices within a cluster; determine the average line factor across all CTI devices in the system.

Step 3 Determine the application factor.

Determine the application factor of all devices in the cluster, according to Table 27-8.

Table 27-8 CTI Application Factor

Number of Applications per CTI Device	CTI Application Factor
1 to 5 applications	1.0
6 applications	1.2
7 applications	1.4
8 applications	1.6
9 applications	1.8
10 applications	2.0

Step 4 Calculate the required number of CTI resources according to the following formula:

Required Number of CTI Resources = (Total CTI Device Count) * (The greater of {the CTI Line Factor or the CTI Application Factor})

The following examples illustrate the process.

Example 1: 500 CTI devices deployed with an average of 9 lines per device and an average of 4 applications per device. According to the factor lists in Table 27-7 and Table 27-8, the 9 lines per device renders a line factor of 1.8, while 4 applications per device renders an application factor of 1.0. Applying these values in the formula from Step 4 yields:

(500 CTI Devices) * (Greater of {1.8 Line Factor or 1.0 Application Factor})

(500 CTI Devices) * (1.8 Line Factor) = 900 total CTI resources required

Example 2: 2,000 CTI devices deployed with an average of 5 lines per device and an average of 9 applications per device. According to the factor lists in Table 27-7 and Table 27-8, the 5 lines per device renders a line factor of 1.0, while 9 applications per device renders an application factor of 1.8. Applying these values in the formula from Step 4 yields:

(2000 CTI Devices) * (Greater of {1.0 Line Factor or 1.8 Application Factor})

(2000 CTI Devices) * (1.8 Application Factor) = 3,600 total CTI resources required

Example 3: 5,000 CTI devices deployed with an average of 2 lines per device and an average of 3 applications per device. According to the factor lists in Table 27-7 and Table 27-8, the 2 lines per device renders a line factor of 1, while 3 applications per device renders an application factor of 1. Applying these values in the formula from Step 4 yields:

(5,000 CTI Devices) * (Greater of {1 Line Factor or 1 Application Factor})

(5,000 CTI Devices) * (1 Line or Application Factor) = 5,000 total CTI resources required

IP Phone Services

Cisco Unified IP Phone Services are applications that utilize the web client and/or server and XML capabilities of the Cisco Unified IP Phone. The Cisco Unified IP Phone firmware contains a micro-browser that enables limited web browsing capability. These phone service applications provide the potential for value-added services and productivity enhancement by running directly on the user's desktop phone.

Cisco Unified IP Phone Services act, for the most part, as HTTP clients. In most cases they use Unified CM only as a redirect server to the location of the subscribed service. Because Unified CM acts only as a redirect server, there typically is minimal performance impact on Unified CM unless there is a large number of requests (hundreds of requests per minute or more).

With the exception of IP Phone Services for the integrated Extension Mobility and Unified CM Assistant applications, IP Phone Services must reside on a separate web server. Running phone services other than Extension Mobility and Unified CM Assistant on the Unified CM server is not supported.

Cisco Extension Mobility and Extension Mobility Cross Cluster

Using Extension Mobility (EM) impacts the system performance in the following ways:

- Creation of EM profiles requires both disk database space and static memory.
- The rate at which users may log into their EM accounts affects both CPU and memory usage. Servers have bounds on the maximum number of logins per minute that they can support.
- Extension Mobility Cross Cluster (EMCC) has a higher impact on resources. There is a limit on the number of EMCC users that a server can support. The maximum EMCC login rates supported are lower than those supported for EM. In addition, there is a trade-off between EM and EMCC login rates. If both are occurring at the same time, then the maximum capacity for each will be reduced.
- EM and EMCC login rates per cluster are not simply the login rate of each server multiplied by the number of servers in the cluster, because profiles in a shared database have to be accessed. The maximum login rate in a cluster consisting of more than one call processing subscriber should be limited to 1.5 times that of a single server.

Table 27-9 shows the maximum number of EM and EMCC logins per minute for each type of server.

Table 27-9 EM and EMCC Rates Per Server Type

Server Types	Maximum EM Login Rate (per Server)	Maximum EM Login Rate (Dual Servers)	Maximum EMCC Login Rate (Per Server)	Maximum EMCC Login Rate (Dual Servers)	Maximum Concurrent EMCC Devices
MCS 7815, MCS 7816	15	22	5	7	100 (MCS 7815) or 167 (MCS 7816)
MCS 7825 and OVA equivalent	200	300	60	70	333

Table 27-9 EM and EMCC Rates Per Server Type (continued)

Server Types	Maximum EM Login Rate (per Server)	Maximum EM Login Rate (Dual Servers)	Maximum EMCC Login Rate (Per Server)	Maximum EMCC Login Rate (Dual Servers)	Maximum Concurrent EMCC Devices
MCS 7835 (I2/H2, I3/H3) and OVA equivalent	235	352	71	80	833
MCS 7845 and OVA equivalent	250	375	75	90	2,500

Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. For example, when the EM load is distributed evenly between two MCS 7845-H2/I2 servers, the maximum cluster-wide capacity is 375 sequential logins and/or logouts per minute.

Note The Cisco Extension Mobility service can be activated on more than two nodes for redundancy purposes, but Cisco supports a maximum of two subscriber nodes actively handling logins/logouts at any given time.

Note Enabling EM Security does not diminish performance.

The EMCC login/logout process requires more processing resources than intracluster EM login/logout, therefore the maximum supported login/logout rates are lower for EMCC. In the absence of any intracluster EM logins/logouts, Unified CM supports a maximum rate of 75 EMCC logins/logouts per minute with Cisco MCS 7845-H2/I2 and MCS 7845-I3 servers. Most deployments will have a combination of intracluster and intercluster logins/logouts occurring. For this more common scenario, the mix of EMCC logins/logouts (whether acting as home cluster or visiting cluster) should be modeled for 40 per minute, while the intracluster EM logins should be modeled for 185 logins/logouts when using a single EM login server. The intracluster EM login rate can be increased to 280 logins/logouts per minute when using MCS 7845-H2/I2 or MCS 7845-I3 servers in dual EM server configuration. (See Table 27-9.)

EMCC logged-in devices (visiting phones) consume twice as many resources as any other endpoint in a cluster. The maximum supported number of EMCC logged-in devices is 2,500 per cluster, but this also decreases the theoretical maximum number of other devices per cluster from 30,000 to 25,000. Even if the number of other registered devices in the cluster is reduced, the maximum supported number of EMCC logged-in devices is still 2,500.

Cisco Unified CM Assistant

The Cisco Unified CM Assistant application uses CTI resources in Unified CM for line monitoring and phone control. Each line (including intercom lines) on a Unified CM Assistant or Manager phone requires a CTI line from the CTIManager. In addition, each Unified CM Assistant route point requires a CTI line instance from the CTIManager. When you configure Unified CM Assistant, the number of required CTI lines or connections must be considered with regard to the overall cluster limit for CTI lines or connections.

The following limits apply to Unified CM Assistant:

- A maximum of 10 Assistants can be configured per Manager.
- A maximum of 33 Managers can be configured for a single Assistant (if each Manager has one Unified CM Assistant-controlled line).
- A maximum of 3,500 Assistants and 3,500 Managers (7,000 total users) can be configured per cluster using the Cisco MCS 7845 server.
- A maximum of three pairs of primary and backup Unified CM Assistant servers can be deployed per cluster if the **Enable Multiple Active Mode** advanced service parameter is set to **True** and a second and third pool of Unified CM Assistant servers are configured.

In order to achieve the maximum Unified CM Assistant user capacity of 3,500 Managers and 3,500 Assistants (7,000 users total), multiple Unified CM Assistant server pools must be defined. (For more information, see Unified CM Assistant, page 18-20.)

Cisco WebDialer

Cisco WebDialer provides a convenient way for users to initiate a call. Its impact on Unified CM is fairly limited because extra resources are required only at call initiation and are not tied up for the duration of the call. Once the call has been established, its impact on Unified CM is just like any other call.

The WebDialer and Redirector services can run on one or more subscriber nodes within a Unified CM cluster, and they support the following capacities:

- Each WebDialer service can handle up to 4 call requests per second per node.
- Each Redirector service can handle up to 8 call requests per second.

The following general formula can be used to determine the number of WebDialer calls per second (cps):

(Number of WebDialer users) * ((Average BHCA) / (3600 seconds/hour))

When performing this calculation, it is important to estimate properly the number of BHCA per user that will be initiated specifically from using the WebDialer service. The following example illustrates the use of these WebDialer design calculations for a sample organization.

Example: Calculating WebDialer Calls per Second

Company XYZ wishes to enable click-to-call applications using the WebDialer service, and their preliminary traffic analysis resulted in the following information:

- 10,000 users will be enabled for click-to-call functionality.
- Each user averages 6 BHCA.
- 50% of all calls are dialed outbound, and 50% are received inbound.
- Projections estimate 30% of all outbound calls will be initiated using the WebDialer service.

Note These values are just examples used to illustrate a WebDialer deployment sizing exercise. User dialing characteristics vary widely from organization to organization.

10,000 users each with 6 BHCA equates to a total of 60,000 BHCA. However, WebDialer deployment sizing calculations must account for placed calls only. Given the initial information for this sizing example, we know that 50% of the total BHCA is for placed or outbound calls. This results in a total of 30,000 placed BHCA for all the users enabled for click-to-call using WebDialer.

Of these placed calls, the percentage that will be initiated using the WebDialer service will vary from organization to organization. For the organization in this example, several click-to-call applications are made available to the users, and it is projected that 30% of all placed calls will be initiated using WebDialer.

(30,000 placed BHCA) * 0.30 = 9,000 placed BHCA using WebDialer

To determine the number of WebDialer servers required to support a load of 9,000 BHCA, we convert this value to the average call attempts per second required to sustain this busy hour:

(9,000 call attempts / hour) * (hour/3,600 seconds) = 2.5 cps

Each WebDialer service can support up to 4 cps, therefore one node can be configured to run the WebDialer service in this example. This would allow for future growth of WebDialer usage. In order to maintain WebDialer capacity during a server failure, additional backup WebDialer servers should be deployed to provide redundancy.

Attendant Console

The integration of Cisco Unified CM with the Cisco Unified Department, Unified Business, and Unified Enterprise Attendant Consoles centers on their CTI resource usage. These applications monitor the last 2,000 users to whom the attendant sent calls, thus increasing CTI resource usage. In addition, each call uses a number of CTI route points and ports for greetings, queuing, and so forth.

Media Resources

The Unified CM server offers the Cisco IP Voice Media Streaming Application, which provides certain media functions that are performed in software only and do not require hardware resources. Unified CM can act as a media termination point (MTP), as a conference bridge, as an annunciator (for playing announcements), or as a source of music-on-hold streams. Although the capabilities of Unified CM are limited compared to similar functions provided by Cisco Integrated Service Routers (ISRs), they are generally the key source of music-on-hold streams (both unicast and multicast).

The Cisco IP Voice Media Streaming Application may be deployed in one of two ways:

- Co-resident deployment

  In a co-resident deployment, the streaming application runs on any server (either publisher or subscriber) in the cluster that is also running the Unified CM software.

  Note The term *co-resident* refers to two or more services or applications running on the same server.

- Standalone deployment

  In a standalone deployment, the streaming application runs on a dedicated server within the Unified CM cluster. The Cisco IP Voice Media Streaming Application service is the only service enabled on the server, and the only function of the server is to provide media resources to devices within the network.

The Cisco IP Voice Media Streaming Application can provide MTP, annunciation, and conferencing capabilities, but a more scalable design is to place these functions on external Cisco Integrated Service Routers (ISRs). The music-on-hold functionality of this application is, however, not so easily placed on external sources. Table 27-10 lists the maximum values that may be configured for each of these services.

Table 27-10 Cisco IP Voice Media Streaming Application Capacity Limits

Service	Maximum Number of Streams
Annunciator	400
Conference Bridge	256
Media Termination Point	512

Note To calculate the capacities of each of the media functions on the DSPs supported by each individual ISR, refer to the Cisco ISR product data sheets or to the chapter on Media Resources, page 7-1.

Music on Hold

Table 27-11 lists the server platforms and the maximum number of simultaneous music-on-hold (MoH) sessions each can support. You should ensure that the actual usage does not exceed these limits because, once MoH sessions have reached these limits, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality.

Table 27-11 Music on Hold Capacity Limits

Server Platform	Codecs Supported	Maximum Number of MoH Sessions
MCS 7816 MCS 7825 MCS 7878 and OVA equivalent	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident or standalone server: 500 MoH sessions
MCS 7835 MCS 7845 and OVA equivalent	G.711 (A-law and mu-law) G.729a Wideband audio	Co-resident or standalone server: 1,000 MoH sessions

You can define a maximum of 51 unique sources of Music on Hold on a Unified CM cluster. Considering that each MoH source may be streamed in up to four encodings, there can be a maximum of 204 multicast streams in the cluster. The limits described in Table 27-11 apply to any combination of unicast, multicast, or simultaneous unicast and multicast sessions.

Impact on Unified CM

Whether deployed in co-resident or standalone mode, the Cisco IP Voice Media Streaming Application consumes CPU and memory resources. This impact must be considered in the overall sizing of Unified CM.

In general, usage of media resources can be considered to add to the BHCA that needs to be processed by Unified CM.

LDAP Directory Integration

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for

that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

User account information is cluster-specific. Each Unified CM publisher server maintains a unique list of those users receiving Unified Communications services from that cluster. Synchronization agreements are cluster-specific, and each publisher has its own unique copy of user account information.

The maximum number of users for a Unified CM cluster is limited by the maximum size of the internal configuration database that gets replicated between the cluster members. Currently the maximum number of users that can be configured or synchronized is 80,000. To optimize directory synchronization performance, Cisco recommends considering the following points:

- Directory lookup from phones and web pages may use the Unified CM database or the IP Phone Service SDK. When directory lookup functionality uses the Unified CM database, only users who were configured or synchronized from the LDAP store are shown in the directory. If a subset of users is synchronized, then only that subset of users is seen on directory lookup.
- When the IP Phone Services SDK is used for directory lookup, but authentication of Unified CM users to LDAP is needed, the synchronization can be limited to the subset of users who would log in to the Unified CM cluster.
- If only one cluster exists, if the LDAP store contains fewer than the maximum number of users supported by the Unified CM cluster, and if directory lookup is implemented to the Unified CM database, then it is possible to import the entire LDAP directory.
- If multiple clusters exist and if the number of users in LDAP is less than the maximum number of users supported by the Unified CM cluster, it is possible to import all users into every cluster to ensure directory lookup has all the entries.
- If the number of user accounts in LDAP exceeds the maximum number of users supported by the Unified CM cluster and if the entire user set should be visible to all users, it will be necessary to use the Unified IP Phone Services SDK to off-load the directory lookup from Unified CM.
- If both synchronization and authentication are enabled, user accounts that have either been configured or synchronized into the Unified CM database will be able to log in to that cluster. The decision about which users to synchronize will impact the decision on directory lookup support.

Note Cisco supports the synchronization of user accounts up to the limit mentioned above, but it does not enforce this limit. Synchronizing more user accounts can lead to starvation of disk space, slower database performance, and longer upgrade times.

Cisco Unified CM Megacluster Deployment

A Unified CM cluster is considered to be a megacluster when the number of call processing subscribers exceeds the normal cluster maximum of 4 pairs. A megacluster may have up to 8 pairs of call processing subscribers and no more than 21 servers in all.

A Unified Communications deployment can be simplified in certain cases with a Unified CM megacluster. The following limits increase with such a deployment:

- Maximum number of endpoints supported is now twice the number in a normal cluster (up to 80,000 using MCS 7845-I3 or OVA equivalent).
- Maximum number of CTI devices and connections also doubles.

However, some cluster-wide constants do not increase. Chief among these are:

- Size of the configuration database
- Number of locations and regions

Note Due to the many potential complexities surrounding megacluster deployments, customers who wish to pursue such a deployment must engage either their Cisco Account Team or their certified Cisco Unified Communications Partner.

Cisco Intercompany Media Engine

The sizing of servers used for running the Cisco Intercompany Media Engine (IME) depends only on the quantity of directory numbers enrolled for the IME service. Table 27-12 lists the capacity of each supported server.

Table 27-12 IME Server Supported Capacities

Server Platform	Maximum Number of Enrolled DIDs
MCS 7825-I2/H2 and 7825-I4/H4	20,000
MCS 7845-I2/H2 and 7845-I3	40,000

Because all IME call media (audio and video) flow through the IME-enabled Cisco Adaptive Security Appliance (ASA), capacity depends on the type and number of calls flowing through it. The IME-enabled ASA monitors only the audio stream incoming from the internet for voice quality. The video media is not monitored for voice quality, but it does flow through the IME-enabled ASA for RTP-to-SRTP conversion, and the bandwidth of the video directly affects the number of sessions each ASA can handle. Table 27-13 provides capacity limits for the ASA-5550 and ASA-5580. Performance limits of other ASA models have not been validated yet.

Table 27-13 Maximum Number of IME Calls per Type of Call and ASA Model

ASA Model	Voice G.711	Video 300 kbps	Video 800 kbps	Video 1 Mbps
ASA-5500 4 GB	480 Calls	240 Calls	120 Calls	80 Calls
ASA-5580-20 4 GB	900 Calls	600 Calls	300 Calls	200 Calls

Impact of IME on Unified CM

Unified CM does not have a limit on the number of IME calls it can handle, but IME calls should be factored into the overall call capacity provided by the cluster. In addition, some calls through IME might need to be re-routed mid-call through gateways if the call quality is not considered acceptable. The expected number of calls re-routed this way should be considered both for Unified CM processing and for number of calls through the gateways.

Emergency Services

The Cisco Emergency Responder tracks the locations of phones and the access switch ports to which they are connected. The phones may be discovered automatically or entered manually into the Emergency Responder. Table 27-14 shows the server platforms that support the Emergency Responder and their maximum capacities.

Table 27-14 Cisco Emergency Responder Server Platforms and Capacities

Server Platform	Maximum Number of Automatically Tracked Phones	Maximum Number of Manually Configured Phones	Maximum Number of Roaming Phones	Maximum Number of Switches	Maximum Number of Switch Ports	Maximum Number of Emergency Response Locations
MCS 7816	6,000	1,000	600	200	12,000	1,000
MCS 7825 and OVA equivalent	12,000	2,500	1,200	500	30,000	3,000
MCS 7835 and OVA equivalent	20,000	5,000	2,000	1,000	60,000	7,500
MCS 7845 and OVA equivalent	30,000	10,000	3,000	2,000	120,000	10,000

The formal definitions of the OVA templates for Cisco Emergency Responder and other Unified Communication products are available at the following location:

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

There can be only one Emergency Responder active per Unified CM cluster. Therefore, choose a server platform that has sufficient resources to provide emergency coverage for all of the phones in the cluster.

For more details on network hardware and software requirements for Emergency Responder, refer to the *Cisco Emergency Responder Administration Guide*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps842/prod_maintenance_guides_list.html

Gateways

PSTN gateways handle traffic between the Unified Communications system and the PSTN. The amount of traffic determines the resource usage (CPU and memory) and the number of PSTN DS0 circuits required for the gateways.

PSTN traffic is generated by the endpoints registered to Unified CM, but there may be other sources such as interactive voice response (IVR) applications and other parts of a contact center deployment.

Gateways can also perform other functions that require resources (such as CPU, memory, and DSP). These functions include media processing such as media termination point (MTP), transcoding, conference bridge, and RSVP Agents.

Gateways, especially those based on the Cisco Integrated Service Routers (ISRs), can provide other functions such as serving as VXML processing engines, acting as border elements, doubling as Cisco Unified Communications Manager Express or Survivable Remote Site Telephony (SRST), or performing WAN edge functions. All of these activities need to be taken into account when calculating the gateway load.

Gateway Groups

When considering the number of gateways, you also need to consider the geographical placement of physical gateway servers. In a deployment model where PSTN access is distributed, you need to size those gateways as a group by themselves and assign the appropriate amount of load to each such group.

A grouping might also be appropriate if certain gateways are expected to be dedicated for certain functions and share common characteristics.

Therefore, to accurately estimate the number of gateways required, the following information is required:

- Groups of gateways that share a common group profile. The common profiles will depend on the complexity of the deployment.
- For each group, the traffic patterns, platform, blocking probability, and so forth, that make up the profile.
- The individual gateway platform that makes up the group. In deciding on a particular gateway model, ensure that the model can support the capabilities and the capacity that is expected of it. Note that more than one gateway might be required in a gateway group, depending on the ability of the selected platform to meet the performance requirements.

PSTN Traffic

PSTN circuits are shared by all users of the system, and there are usually many more users than PSTN circuits. The number of circuits required is estimated by using the traffic management principles described in the section on call traffic (Call Traffic, page 27-21).

The amount of external traffic received and generated by your business determines the number of PSTN circuits required. When converting from a TDM-based system, many customers will continue to use the same number of circuits for their IP-based communications system as they had used for the previous system. However, you may want to perform a new traffic analysis, which will detect if the system is over-provisioned for the current levels of traffic (and, therefore, the customer is paying for circuits that are not needed). If the system is under-provisioned, users will experience an unacceptable number of blocked and/or lost calls, in which case increasing the number of circuits will remedy the situation.

The number of PSTN circuits determines the DSP requirements for the gateways. DSP resources are required to perform conversion between IP and TDM voice (PSTN circuits use TDM encoding).

One key input is the blocking factor, which determines the percentage of call attempts that may not be serviced at peak traffic levels. A lower blocking factor means that more call attempts will succeed, but the system will require more circuits than for a higher blocking factor.

Gateway Sizing for Contact Center Traffic

Short call durations as well as bursty call arrival rates impact the PSTN gateway's ability to process the traffic. Under these circumstances the gateway needs more resources to process all calls in a timely manner, compared to calls of longer duration that are presented more uniformly over time. Because gateways have varying capabilities to deal with these traffic patterns, careful consideration should be

given to selecting the appropriate gateway for the environment in which it will operate. Some gateways support more T1/E1 ports than others, and some are more able than others to deal with multiple calls arriving at the same time.

For a traffic pattern with multiple calls arriving in close proximity to each other (that is, high or bursty call arrival rates), a gateway with a suitable rating of calls per second (cps) is the best fit. Under these conditions, using calls with 15-second hold times, the Cisco AS5400XM Universal Gateway can maintain 16 cps with 250 calls active at once, the Cisco 3845 Integrated Services Router can maintain 13 cps with 200 calls active at once, and the Cisco 3945 Integrated Services Router can maintain 28 cps with 420 calls active at once. The performance of the Cisco AS5350XM Universal Gateway is identical to that of the AS5400XM in terms of calls per second.

For traffic patterns with a steady arrival rate, the maximum number of active calls that a gateway can handle is generally the more important consideration. Under these conditions, using calls with 180-second hold times, the Cisco AS5400XM Universal Gateway can maintain 600 simultaneously active calls with a call arrival rate of up to 3.3 cps, the Cisco 3845 Integrated Services Router can maintain 450 simultaneously active calls with a call arrival rate of up to 2.5 cps, and the Cisco 3945 Integrated Services Router can maintain 720 simultaneously active calls with a call arrival rate of up to 4 cps.

These numbers assume that all of the following conditions apply:

- CPU utilization does not exceed 75%.
- PSTN gateway calls are made with ISDN PRI trunks using H.323.
- The Real Time Control Protocol (RTCP) timer is set to the default value of 5 seconds.
- Voice Activity Detection (VAD) is off.
- G.711 uses 20 ms packetization.
- Cisco IOS Release 15.0.1M is used.
- Dedicated voice gateway configurations are used, with Ethernet (or Gigabit Ethernet) egress and no QoS features. (Using QoS-enabled egress interfaces or non-Ethernet egress interfaces, or both, will consume additional CPU resources.)
- No supplementary call features or services are enabled – such as general security (for example, access control lists or firewalls), voice-specific security (TLS, IPSec and/or SRTP), AAA lookups, gatekeeper-assisted call setups, VoiceXML or TCL-enabled call flows, call admission control (RSVP), and SNMP polling/logging. Such extra call features use additional CPU resources.

Voice Activity Detection (VAD)

Voice Activity Detection (VAD) is a digital signal processing feature that suppresses the creation of most of the IP packets during times when the speech path in a particular direction of the call is perceived to be silent. Typically only one party on a call speaks at a time, so that packets need to flow in only one direction, and packets in the reverse (or silent) direction need not be sent except as an occasional keepalive measure. VAD can therefore provide significant savings in the number of IP packets sent for a VoIP call, and thereby save considerable CPU cycles on the gateway platform. While the actual packet savings that VAD can provide varies with the call flow, the application, and the nature of speaker interactions, it tends to use 10% to 30% fewer packets than would be sent for a call made with VAD turned off.

VAD is most often turned off in endpoints and voice gateways deployed in Unified CM networks; VAD is most often turned on in voice gateways in other types of network deployments.

Codec

Both G.711 and G.729A use as their default configuration a 20 ms sampling time, which results in a 50 packets-per-second (pps) VoIP call in each direction. While a G.711 IP packet (200 bytes) is larger than a G.729A packet (60 bytes), this difference has not proven to have any significant effect on voice gateway CPU performance. Both G.711 and G.729 packets qualify as "small" IP packets to the router, therefore the packet rate is the salient codec parameter affecting CPU performance.

Performance Overload

Cisco IOS is designed to have some amount of CPU left over during peak processing, to handle interrupt-level events. The performance figures in this section are measured with the processor running at an average load of approximately 75%. If the load on a given Cisco IOS gateway continually exceeds this threshold, the following results will occur:

- The deployment will not be supported by Cisco Technical Assistance Center (TAC).
- The Cisco IOS Gateway will display anomalous behavior, including Q.921 time-outs, longer post-dial delay, and potentially interface flaps.

Cisco IOS Gateways are designed to handle a short burst of calls, but continual overloading of the recommended call rate (calls per second) is not supported.

Note With any gateway, you might be tempted to assign unused hardware ports to other tasks, such as on a Cisco Communication Media Module (CMM) gateway where traffic calculations have dictated that only a portion of the ports can be used for PSTN traffic. However, the remaining ports must remain unused, otherwise the CPU will be driven beyond supported levels.

Performance Tuning

The CPU utilization of a Cisco IOS Voice Gateway is affected by every process that is enabled in a chassis. Some of the lowest level processes such as IP routing and memory defragmentation will occur even when there is no live traffic on the chassis.

Lowering the CPU utilization can help to increase the performance of a Cisco IOS Voice Gateway by ensuring that there are enough available CPU resources to process the real-time voice packets and the call setup instructions. Table 27-15 describes some of the techniques for decreasing CPU utilization.

Table 27-15 Techniques for Reducing Gateway CPU Utilization

Technique	CPU Savings	Description
Enable Voice Activity Detection (VAD)	Up to 20%	Enabling VAD can result in up to 45% fewer voice packets in typical conversations. The difficultly is that, in scenarios where voice recognition is used or there are long delays, a reduction in voice quality can occur. Voice appears to "pop" in at the beginning and "pop" out at the end of talk spurts.
Disable Real Time Control Protocol (RTCP)	Up to 5%	Disabling RTCP results in less out-of-band information being sent between the originating and terminating gateways. This results in lower quality of statistics displayed on the paired gateway. This can also result in the terminating gateway having a call "hang" for a longer period of time if RTCP packets are being used to determine if a call is no longer active.

Table 27-15 Techniques for Reducing Gateway CPU Utilization (continued)

Technique	CPU Savings	Description
Disable other non-essential functions such as: Authentication, Authorization, and Accounting (AAA); Simple Network Management Protocol (SNMP); and logging	Up to 2%	Any of these processes, when not required, can be disabled and will result in lower CPU utilization by freeing up the CPU to provide faster processing of real-time traffic.
Change the call pattern to increase the length of the call (and reduce the number of calls per second)	Varies	This can be done by a variety of techniques such as including a long(er) introduction prompt played at the beginning of a call or adjusting the call script at the call center.

Additional Information

A full discussion of every gateway, its capabilities, and call processing capacities is not possible in this chapter. For more information on Cisco Voice Gateways, refer to the following documentation:

- Cisco Voice Gateway Solutions:

  http://www.cisco.com/en/US/products/sw/voicesw/index.html#~all-prod

- Gateway protocols supported with Cisco Unified Communications Manager (Unified CM):

  http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/admin/8_0_1/ccmsys/a08gw.html

- Interfaces and signaling types supported by the following Cisco Voice Gateways:
  - Cisco 3900 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps10536/products_relevant_interfaces_and_modules.html
  - Cisco 2900 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps10537/products_relevant_interfaces_and_modules.html
  - Cisco 3800 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps5855/products_relevant_interfaces_and_modules.html
  - Cisco 2800 Series Integrated Services Routers

    http://www.cisco.com/en/US/products/ps5854/products_relevant_interfaces_and_modules.html

- Gateway features supported with MGCP, SIP, and H.323:

  http://www.cisco.com/en/US/prod/collateral/routers/ps259/product_data_sheet0900aecd8057f2e0.pdf

- SIP gateway RFC compliance:

  http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/gatecont/ps6831/product_data_sheet0900aecd804110a2.html

- Skinny Client Control Protocol (SCCP) feature support with FXS gateways:

  http://www.cisco.com/en/US/prod/collateral/voicesw/ps6790/gatecont/ps2250/ps5516/product_data_sheet09186a00801d87f6.html

- Gateway capacities and minimum releases of Cisco IOS and Unified CM required for conferencing, transcoding, media termination point (MTP), MGCP, SIP, and H.323 gateway features:

  http://www.cisco.com/en/US/prod/collateral/routers/ps259/product_data_sheet0900aecd8057f2e0.pdf

- Various voice traffic calculators, including Erlang calculators:

  http://www.erlang.com/calculator/

Voice Messaging

Voice messaging is an application that needs to be sized not only by itself but also for its effect on other Unified Communications components, mainly Unified CM.

Total number of users is the key factor for sizing the voice messaging system. Other factors that affect sizing for voice messaging are:

- Number of calls during the busy hour that the application has to handle
- Average length of messages left on the servers
- Number of users who check their messages during the busy hour
- Average length of user sessions
- Any advanced operations such as voice recognition or text-to-speech sessions
- Any media transcoding
- Ports on the voice messaging system are analogous to the DS0s on a gateway and are shared resources that need to be optimized. The same considerations of probabilistic arrival and the need for blocking apply to both types of resources.

Table 27-16 shows the applicability of the various voice messaging solutions to the scalability requirements of the deployment.

Table 27-16 Scaling Voice Messaging Solutions

	Maximum Number of Users Supported on a Single Server (or Failover or Clustered Deployment)				**Maximum Number of Users Supported in a Digital Networking Solution**	
Solutions	**500**	**1,000**	**15,000**	**20,000**	**100,000**	**250,000**
Cisco Unity Express	Y	N	N	N	Y	Y
Cisco Business Edition 6000	Y	Y	N	N	N	N
Cisco Unity Connection (Unified/Integrated Messaging)	Y	Y	Y	Y	Y	N

Table 27-17 shows the maximum limits of various functions of different servers running Cisco Unity Connection.

Table 27-17 Servers and Capacities for Cisco Unity Connection

Server Platform	Maximum Number of Ports	Maximum Voice Recognition Sessions	Maximum Text to Speech Sessions	Maximum Number of Voicemail Users
MCS 7825	48	48	48	2,000
MCS 7835	150	150	150	4,000
MCS 7845	250	250	250	20,000
OVA Template for 5,000 users	100	100	100	5,000
OVA Template for 10,000 users	150	150	150	10,000
OVA Template for 20,000 users	250	250	250	20,000

The formal definitions of the OVA templates for Cisco Unity Connection and other Unified Communication products are available at

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

Impact on Unified CM

The impact of a voice messaging system on Unified CM can be gauged by considering the extra processing that Unified CM needs to do. These extra call flows add to the sizing load of Unified CM as follows:

- Calls that need to be forwarded to the voice messaging system when the user is not present or if the user deliberately forwards the calls using Do Not Disturb (DND) or other features.
- Calls from users who dial the voice messaging pilot number to access their voice messages go through Unified CM, and these calls must be added to the calls being handled by Unified CM, including both the number and the duration of these calls.

Collaborative Conferencing

Cisco Collaborative Conferencing systems include Cisco Unified CM as a component for call control. When sizing such a system, the function it performs as well as its impact to Unified CM should be considered.

When sizing such conferencing systems, you typically have to consider the following parameters to determine the type and number of servers:

- Number of users who could use the system at any one time
- Number of audio, video, and web users on the system at the peak usage time
- Required dial-in duration
- Video resolution and audio codec requirements

Sizing Guidelines for Audio Conferencing

Cisco recommends the following methods for calculating audio conferencing capacity:

- Calculation based on average monthly usage

  If you know the average voice conferencing usage (average minutes per month), use Table 27-18 to calculate the audio conferencing capacity.

Table 27-18 Audio Conferencing Capacity Based on Average Monthly Usage

Average Monthly Usage (minutes)	Baseline Usage (minutes per port per month)	Estimated Number of Ports
20,000 to 50,000	1,500	15 to 35
50,000 to 500,000	2,000	25 to 250
500,000 to 1,000,000	3,000	165 to 335
1,000,000 to 2,000,000	3,500	285 to 570
2,000,000 to 8,000,000	4,000	500 to 2,000

- Calculation based on number of users

  You should plan on having one port for every 20 users with average usage. If the users are heavy conference users, then provision one port for every 15 users. For example, in a system with 6000 users, you should provision 300 audio ports; however, if those users heavily use conferencing, then plan for 400 audio ports.

- Calculation based on actual peak usage

  Actual voice conferencing usage during peak hours usually can be obtained from existing voice conferencing system logs or service provider bills. Cisco recommends provisioning 30% extra capacity based on the actual peak usage in order to protect against extra conferencing volume.

Factors Affecting System Sizing

In addition to the estimates provided by the methods described above for the system baseline port requirement, the following factors also affect system sizing:

- When migrating from an "operator-scheduled" model to a user-scheduled model, you might need to add another 20% to the baseline.
- The default average meeting size is 4.5 callers per meeting. Use the value that is applicable to your case if it is different than the default.
- Increase the baseline estimate accordingly if the following condition applies:

  (Estimated meetings per day) * (Estimated users) > 80% of baseline

- If the largest single meeting exceeds 20% of the estimated capacity, increase the estimate accordingly.
- If there are continuous meetings with dedicated ports, then you must add those additional ports ((Meetings) * (Dedicated callers)) to the baseline.

The total number of ports will include all the above factors in addition to the baseline. Plan for conferencing system capacity expansion if the total estimated port capacity exceeds 80% of the maximum supported ports.

Sizing Guidelines for Video Conferencing

Cisco recommends the following three methods for calculating video conferencing capacity:

- Calculation based on number of knowledgeable workers

  Cisco recommends provisioning a video user license for every 40 knowledgeable workers.

- Calculation based on number of voice conferencing user licenses

  Cisco recommends provisioning video conferencing capacity in the range of 17% to 25% of existing audio user licenses. The percentage depends on business requirements regarding video conferencing and on the size of the conferencing system.

- Calculation based on existing video Multipoint Control Unit (MCU)

  Cisco recommends deploying a direct replacement for an existing video conferencing system.

Impact on Unified CM

The impact to Unified CM can be analyzed based on the extra call traffic that the conferencing system generates. The most impact occurs when conference users dial into their meetings that are typically scheduled at the top of the hour or half-hour. A large amount of call traffic within a few minutes of conference start times increases the load on Unified CM for just those few minutes and must be designed in appropriately. In addition, if conference users include callers from the PSTN or from other clusters, those parameters must also be considered to gauge their impact on the gateways.

Cisco WebEx Meetings Server

The Cisco WebEx Meetings Server provides WebEx conferencing services using enterprise-provided servers (a Cisco UCS server clusters in the enterprise data center).

Cisco WebEx Meetings Server is offered in different configurations, which the sizing tool chooses based primarily on the number of knowledge workers that have access to the conferencing service.

For each configuration, Cisco recommends a standard Cisco UCS server type with specific configurations of hardware and VMware products. However, Cisco WebEx Meetings Server is designed to work on any equivalent or better Cisco UCS Server that meets or exceeds these specifications.

This product is packaged as a VMware vSphere compatible OVA virtual appliance and not as a collection of software packages on a DVD. Therefore, Cisco WebEx Meetings Server requires the vCenter product to deploy the OVA and install the Cisco WebEx Meetings Server product.

Currently, Cisco WebEx Meetings Server does not operate in co-resident mode on the Cisco UCS server. Cisco WebEx Meetings Server requires a dedicated UCS server.

For additional information about Cisco WebEx Meetings Server, refer to the *Cisco WebEx Meetings Server System Requirements*, available at

http://www.cisco.com/en/US/products/ps12732/products_installation_and_configuration_guides_list.html

Sizing Factors

The sizing tool uses the following inputs to calculate system capacity:

- Number of knowledge users

  The number of knowledge users is defined as the set of employees that can access the conferencing system (to initiate a conference or join a conference).

Many knowledge users share the available conferencing ports. The assumption is that only a small percentage of users are active in a conference call at any time. Based on this percentage, we can estimate of the number of conferencing ports required to support these users.

The sizing tool defines light usage (3.3% of users active at any one time), average usage (5% active) and heavy usage (10% active). Therefore, a system operating with average usage will support twice as many users as a system with heavy usage.

- User minutes per month

  The user minutes per month is the total number of minutes of active conferences for the month, across all ports. This value is expressed in thousands of minutes. This factor is significant for calculating the size of the recording server.

- Actual peak usage

  Actual peak usage is defined as the maximum number of concurrent users of the system. This number is significant in determining the required number of conferencing ports. Cisco recommends provisioning enough capacity to handle 30% more users than the actual peak usage, to ensure that adequate conferencing ports are available during peak usage times.

- Video

  The percent of conferences with video and high-quality video will impact the network bandwidth required by the system. Up to 50% of the users can be using high-quality video.

- Traffic mix

  Different call types require different Unified CM resources. For accurate assessment of the Unified CM impact, the tool requires estimates of the following call types:

  – Percent of conference calls incoming via enterprise IP phones. This call leg is handled by Unified CM and therefore has an impact on Unified CM capacity.

  – Percent of external call legs, which impacts sizing for PSTN gateways.

- Access by external users

  If external users need to access the system, additional virtual machines are configured to provide reverse proxy functionality. If the system is intended for internal users only, these additional virtual machines are not required.

- Disaster recovery

  For disaster recover, you can configure a cold-standby system in a second data center. If the primary system is configured for high availability, you can optionally choose to configure high availability for the disaster recovery system.

- High availability

  The system can be configured in non-redundant mode or in high-availability (HA) mode. In HA mode, the cluster is provisioned with one or more backup servers (the specific configuration depends on the system size).

System Capacities

Cisco WebEx Meetings Server is offered in four system sizes, as listed in Table 27-19. System size is expressed as the maximum number of concurrent users of the system. Maximum concurrent users defines the maximum number of users who can participate in conference calls at any given time.

Table 27-19 Servers and Capacities for Cisco WebEx Meeting Server

Maximum	50 Concurrent Users	250 Concurrent Users	800 Concurrent Users	2,000 Concurrent Users
Audio and web users (combined)	50	250	800	2,000
High-quality video and video sharing (combined)	25	125	400	1,000
Participants in a single meeting	50	100	100	100
Playback recordings of meetings that have ended	13	63	200	500
Recordings of meetings in progress	5	25	80	200
Number of conferences (average of 2 participants per meeting)	25	125	400	1,000
Calls per second	1	3	8	20

Note that the following optional capabilities can be used without any impact on system capacity:

- Encrypted audio (sRTP)
- Secured Meeting Center Web (SSL)
- Different audio codecs
- Low-resolution video

Recordings

Meetings for up to 5% of the ports (or 10% of meetings) can be recorded. You need to provision an NFS-mounted hard drive of sufficient size to store the recorded meetings. One meeting will generate a file with a size of 50 to 100 MB.

Network Bandwidth

To estimate the bandwidth required on the LAN and WAN, the sizing tool makes the following assumptions:

- Each port will use 1 Mbps of network bandwidth.
- The user mix will be 80% internal to the enterprise and 20% external.

Therefore, the required bandwidth (in Mbps) on the LAN is 0.8 * (Number of ports), and on the WAN is 0.2 * (Number of ports)

Collaborative Conferencing with MeetingPlace

The capacity of a given Cisco Unified MeetingPlace solution depends on the platform on which the Unified MeetingPlace Meeting Directors and Application server with Express Media Server (EMS) are installed. For example, with the Unified MeetingPlace Application server installed on a Cisco

MCS 7845-I3 (or equivalent) server, voice conferencing can scale to 1,200 ports (G.711) in a single system or conferencing node. However, with the Unified MeetingPlace Application server installed on a Cisco MCS 7835-I3, voice conference can scale only to 400 ports (G.711) with the same configuration.

Video usage characteristics such as bandwidth and resolution are an important aspect for the sizing of Express Media Server.

Additional Factors Affecting System Sizing

Consider the following recommendations to maintain the maximum capacity with Cisco Unified MeetingPlace:

- If an audio codec other than G.711 is desired, use transcoders based on Cisco Integrated Services Routers (ISR) to achieve maximum capacity.
- Use Line Echo Cancellation (LEC) provided by an external device such as an ISR, rather than the build-in LEC from Unified MeetingPlace.

Express Media Server

The Cisco Unified MeetingPlace Express Media Server (EMS) capacity is directly related to codec and video bandwidth because it is installed co-resident with the Unified MeetingPlace Application server. When the Unified MeetingPlace Application server is installed on a Cisco MCS 7835-H2/I2 server, the overall system capacity decreases for both EMS and HMS deployments. Standards-based video as well as G.729 and G.722 audio codecs all affect the capacity of the EMS system. For the detailed capacity numbers, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

The EMS introduces the concept of System Resource Units (SRUs), where the system capacity (or the Total SRUs value) is based on the type of hardware platform on which the Unified MeetingPlace Application Server resides and the speed and number of processors on that system. The system immediately consumes some of these SRUs from the total for normal operation, and it puts the remaining resources in an SRU pool and makes them available for enhanced audio and video features. Table 27-20 shows the number of total SRUs available for enhanced audio and video per supported platform.

Table 27-20 Total System Resource Units per Supported EMS Platform

Server Platform	Total System Resource Units (SRUs) Available for Enhanced Audio and Video
MCS 7835-I3	400
MCS 7845-I2/H2	500
MCS 7845-I3	1,200
UCS B200 or C210 Series	1,200 (with or without Meeting Director co-resident)
UCS C200 Series	500 (2 nodes with redundancy)

Table 27-21 Number of System Resource Units Consumed for Various Audio Codecs and Video Bandwidths

Session Type	Number of SRUs Used
One G.711 audio port	1
One G.729 or one G.722 audio port	6

Table 27-21 Number of System Resource Units Consumed for Various Audio Codecs and Video Bandwidths (continued)

Session Type	Number of SRUs Used
One video port at 320 kbps[1]	1
One video port at 384 kbps	1
One video port at 768 kbps	2
One video port at 2,000 kbps	6

1. The lowest rate that is guaranteed for a video license is 320 kbps.

As shown by the data in Table 27-20 and Table 27-21, on an MCS 7845-I3 server handling only G.711 audio calls, the EMS supports 1,200 audio sessions. Alternatively, it supports 600 video sessions at up to 384 kbps with G.711 audio (a video session also consumes SRUs for the audio session).

In Unified CM, the regions setting of the SIP trunk used for call delivery to Unified MeetingPlace can be configured to control the audio codec and video bandwidth of calls sent to the EMS. Understanding the nature and capabilities of the endpoints dialing into Unified MeetingPlace is critical to proper design. For more information on EMS capacity planning, refer to the latest version of the *Planning Guide for Cisco Unified MeetingPlace*, available at

http://www.cisco.com/en/US/products/sw/ps5664/ps5669/products_implementation_design_guides_list.html

Unified MeetingPlace Web Server

The Unified MeetingPlace Web Server is required only for Unified MeetingPlace scheduling deployments to schedule and attend meetings from the Web user interface, for Lotus Notes integrations, or for accessing the recording storage. There is no capacity planning consideration for these servers. Cisco MCS 7835 servers are sufficient for the largest Unified MeetingPlace deployment, but MCS 7845 servers may be used as well.

Cisco IM and Presence

As with all other applications, sizing for Cisco IM and Presence is accomplished in the following way:

- Decompose the system into its most elemental services.
- Measure the unit cost of each of these services.
- Analyze the given system description as an aggregation of the identified services and arrive at a net system cost.
- Determine the number of required servers based on system cost and deployment options.

For IM and Presence, the following system variables in the system under analysis are relevant and must be considered for accurate sizing:

- Number and type of users
  - Clients employed by users to obtain presence services
  - Operating mode for users (instant messaging only or full Unified Communications facilities)
- Presence-related activities performed by typical users
  - Contact list size and composition (intracluster, intercluster, and federated)

  - Number of instant messages (directly between two users) per user during the busy hour
  - Chat support with number of chat rooms, users per chat room, and instant messages per user per chat room
  - State changes per user (both call related and user initiated)
- Deployment model
  - Whether intercluster presence is supported
  - Whether federation is supported
  - Whether high availability is desired
- Server preferences
  - The class of server or voice messaging platform desired
- System options
  - Whether compliance recording is required

Once the system requirements are quantified, the number of required servers can be determined from the data in Table 27-22.

Table 27-22 Maximum Number of Users Supported per IM and Presence Cluster

Server Platform	Maximum Users Supported in Full Unified Communications Mode	Maximum Users Supported in Instant Messaging Only Mode
MCS 7816	3,000	7,500
MCS 7825 and OVA equivalent	6,000	6,000
MCS 7835 and OVA equivalent	15,000	37,500
MCS 7845 and OVA equivalent	45,000	75,000

For additional information, refer to the latest version of *Hardware and Software Compatibility Information for IM and Presence Service on Cisco Unified Communications Manager*, available at

http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_device_support_tables_list.html

The formal definitions of the OVA templates for Cisco IM and Presence and other Unified Communication products are available at

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Downloads_(including_OVA/OVF_Templates)

Impact on Unified CM

The Cisco IM and Presence Service influences the performance of Unified CM in the following ways:

- User synchronization through an AXL/SOAP interface
- Presence information through a SIP trunk
- CTI traffic to enable phone control

In general, the impact of user synchronization (except for a one-time hit) and that of presence information through the SIP trunk are negligible. The effect of CTI control of phones, however, must be counted against CTI limits.

Cisco Unified Communications Management Tools

Cisco Prime Collaboration offers a set of integrated tools to test, deploy,and monitor Cisco Unified Communications and TelePresence systems. Cisco Prime Collaboration includes the following products: Cisco Prime Provisioning Manager, Cisco Prime Operations Manager, and Cisco Prime Service Monitor. An additional management product, Cisco Unified Service Statistics Manager, provides advanced statistics analysis and reporting capabilities.

Sizing for these applications is relatively simple and depends directly on the number of endpoints or network devices that they are expected to manage. These applications can work either in a standalone mode hosted on separate hardware servers or in a co-resident environment on a single server.

The server characteristics to host the management applications are generally stated in terms of hardware specifications: CPU characteristics (processor speed and number of cores), memory, and disk space for each level of desired capacity.

The co-resident servers, for example, can host from one to all four of the management applications, and they are specified in two configurations – one for managing up to 2,000 endpoints and a larger configuration for managing up to 10,000 endpoints. The specifications for such co-resident servers are as follows:

- Large co-resident configuration:
  - Processor: 3 GHz, 8 Core
  - Memory: 16 GB
  - Disk space: 320 GB
- Small co-resident configuration:
  - Processor: 3 GHz, 4 Core
  - Memory: 8 GB
  - Disk Space: 100 GB

These hardware characteristics can be mapped to the equivalent Cisco MCS or UCS servers.

Cisco Prime Unified Provisioning Manager

The Cisco Prime Unified Provisioning Manager (Unified PM) can support up to 60,000 phones and can be implemented either on a single machine or on two machines. A two-machine deployment is recommended when the number of phones exceeds 30,000.

Hardware resources required for various levels of performance are described in the *Cisco Unified Provisioning Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps7125/products_data_sheets_list.html

Cisco Prime Unified Operations Manager

The Cisco Prime Unified Operations Manager (Unified OM) can manage phones and other network devices such as routers and switches. The Unified Operations Manager operates in a single machine configuration. The Unified OM supports up to 45,000 phones and 2,000 other IP devices.

Hardware resources required for various levels of performance are described in the *Cisco Unified Operations Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps6535/products_data_sheets_list.html

Cisco Prime Unified Service Monitor

The Cisco Prime Unified Service Monitor (Unified SM) consists of not only the server to run the Unified SM software but also the Cisco 1040 Sensor and Network Analysis Modules (NAMs) to measure voice quality. Table 27-23 lists the maximum number of concurrent RTP streams supported on the 1040 Sensor and various NAMs.

Table 27-23 Performance for 1040 Sensor and Different NAM Types

	Cisco Network Analysis Module Type				
	1040 Sensor	**NME-NAM**	**NAM-2**	**NAM 2204 Appliance**	**NAM 2220 Appliance**
Maximum number of concurrent RTP streams supported	100	100	400	1,500	4,000

Hardware resources required for various levels of performance are described in the *Cisco Unified Service Monitor Data Sheet*, available at

http://www.cisco.com/en/US/products/ps6536/products_data_sheets_list.html

Unified SM supports the following voice quality monitoring capacities:

- Up to 50 Cisco 1040 Sensors
- Up to 45,000 IP phones
- Up to 5,000 sensor-based RTP streams per minute (with Cisco 1040 Sensors or NAMs)
- Up to 1,600 Cisco Voice Transmission Quality (CVTQ) calls per minute
- Up to 1,500 RTP streams and 666 CVTQ calls per minute

Cisco Unified Service Statistics Manager

The Cisco Unified Service Statistics Manager (Unified SSM) operates in a single-server mode and can scale to manage up to 45,000 phones.

Hardware resources required for various levels of performance are described in the *Cisco Unified Service Statistics Manager Data Sheet*, available at

http://www.cisco.com/en/US/products/ps7285/products_data_sheets_list.html

Sizing for Standalone Products

The following products are not included in the sizing tools, but the following sections describe how to size these products:

- Cisco Unified Communications Manager Express, page 27-48
- Cisco Business Edition, page 27-48

Cisco Unified Communications Manager Express

Cisco Unified Communications Manager Express (Unified CME) runs on one of the Cisco IOS Integrated Services Router (ISR) platforms, from the low-end Cisco 881 ISR to the high-end Cisco 3945E ISR 2. Each of these routers has an upper limit on the number of phones that it can support. The actual capacity of these platforms to do call processing may be limited by the other functions that they perform, such as IP routing, Domain Name System (DNS), Dynamic Host Control Protocol (DHCP), and so forth.

Unified CME can support a maximum of 450 endpoints on a single Cisco IOS platform; however, each router platform has a different endpoint capacity based on the size of the system. Because Unified CME is not supported within the Cisco Unified Communications Sizing Tool, it is imperative to follow the capacity information provided in the Unified CME product data sheets available at

http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_data_sheets_list.html

Cisco Business Edition

Cisco Business Edition 6000 (UCS C200 or C220) offers the following capacities:

- Maximum of 1,000 users
- Maximum of 1,200 endpoints on a medium-density server or 2,500 endpoints on a high-density server
- Maximum of 5,000 BHCA

Busy Hour Call Attempts (BHCA) for Cisco Business Edition

As mentioned above, Business Edition 6000 supports a maximum of 5,000 BHCA. When calculating your system usage, stay at or below this BHCA maximum to avoid oversubscribing Cisco Business Edition. The BHCA consideration becomes significant when the usage for any phone is above 4 BHCA. A true BHCA value can be determined only by taking a baseline measurement of usage for the phone during the busy hour. Extra care is needed when estimating this usage without a baseline.

Device Calculations for Cisco Business Edition

Devices can be grouped into two main categories for the purpose of this calculation: phone devices and trunk devices.

A phone device is a single callable endpoint. It can be any single client device such as a Cisco Unified IP Phone 7900 Series, a software client such as Cisco IP Communicator, an analog phone port, or an H.323 client. While Cisco Business Edition 6000 supports a maximum of 1,200 endpoints on a medium-density server or 2,500 endpoints on a high-density server, as indicated above, actual endpoint capacity depends on the total system BHCA.

A trunk device carries multiple calls to more than one endpoint. It can be any trunk or gateway device such as a SIP trunk or a gatekeeper-controlled H.323 trunk. Business Edition 6000 supports intercluster trunking as well as H.323, SIP, and MGCP trunks or gateways and analog gateways.

The method for calculating BHCA is much the same for both types of devices, but trunk devices typically have a much higher BHCA because a larger group of endpoints is using them to access an external group of users (PSTN or other PBX extensions).

You can define groups of devices (phone devices or trunk devices) with usage characteristics based on BHCA, and then you can add the BHCA for each device group to get the total BHCA for the system, always ensuring that you are within the supported maximum of 5,000 BHCA.

For example, you can calculate the total BHCA for 100 phones at 4 BHCA each and 80 phones at 12 BHCA each as follows:

100 phones at 4 BHCA is 100∗4 = 400

80 phones at 12 BHCA is 80∗12 = 960

Total BHCA = (100∗4) + (80∗12) = 1,360 BHCA for all phones

For trunk devices, you can calculate the BHCA on the trunks if you know the percentage of calls made by the devices that are originating or terminating on the PSTN. For this example, if 50% of all device calls originate or terminate at the PSTN, then the net effect that the device BHCA (1360 in this case) would have on the gateways would be 50% of 1360, or 680 BHCA. Therefore, the total system BHCA for phone devices and trunk devices in this example would be:

Total system BHCA = 1,360 + 680 = 2,040 BHCA

If you have shared lines across multiple phones, the BHCA should include one call leg (there are two call legs per each call) for each phone that shares that line. Shared lines across multiple groups of devices will affect the BHCA for that group. That is, one call to a shared line is calculated as one call leg per line instance, or half (0.5) of a call. If you have different groups of phones that generate different BHCAs, use the following method to calculate the BHCA value:

Shared line BHCA = 0.5 ∗ (Number of shared lines) ∗ (BHCA per line)

For example, assume there are two classes of users with the following characteristics:

100 phones at 8 BHCA = 800 BHCA

150 phones at 4 BHCA = 600 BHCA

Also assume 10 shared lines for each group, which would add the following BHCA values:

10 shared lines in the group at 8 BHCA = 0.5∗10∗8 = 40 BHCA

10 shared lines in the group at 4 BHCA = 0.5∗10∗4 = 20 BHCA

The total BHCA for all phone devices in this case is the sum of the BHCA for each phone group added to the sum of the BHCA for the shared lines:

800 + 600 + 40 + 20 = 1,460 total BHCA

Note that the total BHCA in each example above is acceptable because it is below the system maximum of 5,000 BHCA.

If you are using Cisco Unified Mobility for Mobile Connect (also known as single number reach, or SNR) on Business Edition 6000, keep in mind that calls extended to remote destinations or off-system phone numbers affect BHCA. In order to avoid oversubscribing the appliance, you have to account for this SNR remote destination or off-system phone BHCA. To calculate the BHCA for these SNR features, see Capacity Planning for Cisco Unified Mobility, page 23-62, and add that value to your total BHCA calculation.

Note Media authentication and encryption using Secure RTP (SRTP) impacts the system resources and affects system performance. If you plan to use media authentication or encryption, keep this fact in mind and make the appropriate adjustments. Typically, 100 IP phones without security enabled results in the same system resource impact as 90 IP phones with security enabled (10:9 ratio).

Another aspect of capacity planning to consider for Cisco Business Edition is call coverage. Special groups of devices can be created to handle incoming calls for a certain service according to different rules (top-down, circular hunt, longest idle, or broadcast). This is done through hunt or line group configuration within Cisco Business Edition. BHCA can also be affected by this factor, but only as it pertains to the line group distribution broadcast algorithm (ring all members). For Business Edition, Cisco recommends configuring no more than three members of a hunt or line group when a broadcast distribution algorithm is required. Depending on the load of the system, doing so could greatly affect the BHCA of the system and possibly oversubscribe the platform's resources. The number of hunt or line groups that have a distribution algorithm of broadcast should also be limited to no more than three. These are best practice recommendations meant to prevent oversubscription of the system BHCA. Exceeding these recommendations within a deployment is supported as long as the overall BHCA capacity of the system is not exceeded.

Cisco Unified Mobility for Cisco Business Edition

The capacity for Cisco Unified Mobility users on Cisco Business Edition 6000 systems depends exclusively on both the number of remote destinations per user and the BHCA of the users enabled for Unified Mobility, rather than on server hardware. Thus, the number of remote destinations supported on Cisco Business Edition 6000 depends directly on the BHCA of these users. The guidelines for sizing Cisco Unified Mobility for Cisco Business Edition 6000 are as follows:

- No more than 4 remote destinations can be configured per user. Given a maximum of 1,000 users per Cisco Business Edition 6000 system, the theoretical limit is 4,000 remote destinations. However, given the maximum of 5,000 BHCA per Business Edition 6000, it is possible that the system might not be able to support 4,000 remote destinations. Instead BHCA calculations should be used to properly size the number of remote destinations that can be handled by the system.
- Each configured remote destination has potential BHCA implications. For every remote destination configured for a user, one additional call leg is used. Because each call consists of two call legs, one remote destination ring is equal to half (0.5) of a call. Therefore, you can use the following formula to calculate the total remote destination BHCA:

Total remote destination BHCA = 0.5 * (Number of users) *
(Number of remote destinations per user) * (User BHCA)

For example:

Assuming a system of 300 users at 5 BHCA each, with each user having one remote destination (total of 300 remote destinations), the calculation for the total remote destination BHCA would be:

Total remote destination BHCA = 0.5 * (300 users) * (1 remote destination per user) * (5 BHCA per user) = 750 BHCA

Total user BHCA in this example is [(300 users) * (5 BHCA per user)], which is 1,500 total user BHCA. By adding the total remote destination BHCA of 750 to this value, we get a total system BHCA of 2,250 (1,500 total user BHCA + 750 total remote destination BHCA).

If other applications or additional BHCA variables are in use on the system in the example above, the capacity might be limited. (See the preceding sections for further details.)

For more information on Cisco Business Edition 6000 capacity planning as well as all other Business Edition product information, refer to the following product documentation for Cisco Business Edition 6000:

- http://docwiki.cisco.com/wiki/Cisco_Unified_Communications_Manager_Business_Edition_6000
- http://www.cisco.com/en/US/products/ps11369/tsd_products_support_series_home.html