CHAPTER 29

Unified Communications Design and Deployment Sizing Considerations

Revised: July 31, 2012; OL-21733-18

An accurate estimation of the type and quantity of hardware platforms is a prerequisite to a successful deployment of Unified Communications products. Adequate computing and network resources must be provided so that the expected service goals are met.

Each Unified Communications product publishes its capacity limits on each hardware server platform where it runs. These published limits are obviously an important part of determining the needed amount of hardware resources. Individual products, however, may publish only their best-case performance numbers, or may publish numbers for a typical deployment. Both of these numbers are very useful but insufficient for a real-world sizing exercise. For example, Cisco Unified Communications Manager (Unified CM) publishes the maximum number of endpoints that a cluster consisting of Cisco MCS-7845-I3 servers can support. This number may assume average call rates and the absence of any other major activity in the cluster. In an actual usage scenario the call rate might be higher than that assumed, or there might be a requirement to support other services, and even though the nominal number of phones is not exceeded, a single cluster might be inadequate.

Another complexity arises from the fact that each product is rarely used just by itself. Most products are used as part of a larger deployment containing other Unified Communications products. For example, Cisco Unity Connection is likely used with Unified CM and gateways. Larger, more complex deployments may consist of several Unified Communications products, including those from the Cisco Contact Center portfolio (Cisco Unified Contact Center Enterprise, Unified Customer Voice Portal, Unified Intelligence Center, and others), which must work with Unified CM, gateways, Cisco Unified MeetingPlace, Cisco Unity Connection voice messaging, and network management applications. Interaction of each of these components on the others must be taken into account. For example, Unified CM might have to manage not only its regular phones but also the ones assigned to agents who can experience much higher call volumes. Also, gateways might have to handle VXML calls in addition to the regular voice calls. All of these interactions must be taken into account for an accurate sizing estimation.

This chapter discusses the sizing of individual Unified Communications components as well as systems consisting of several components communicating with each other. This chapter also discusses the performance impact of the different functions that the various Unified Communications products support, and it explains why "designing by datasheets" is not the preferred way to deploy a complex Unified Communications network. In addition, this chapter provides insights on how to work with the various sizing tools available, mostly notably the Cisco Unified Communications Sizing Tool.

What's New in This Chapter

This chapter is a new addition for this release of the *Cisco Unified Communications System 8.x SRND*. It borrows some information from the capacity planning sections of other chapters in this document, and it goes into greater detail about sizing in the context of the whole system deployment.

Table 29-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 29-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Collaborative conferencing capacity information	Collaborative Conferencing, page 29-39	July 31, 2012
CTI resources	Applications and CTI, page 29-23	July 31, 2012
Presence capacity information	Table 29-22	July 31, 2012
Maximum number of servers per cluster was corrected from 12 to 20	Server and Cluster Maximums, page 29-18	April 30, 2012

Factors That Affect Sizing

Unified Communications products are designed to be scalable. Capacity of a particular service can generally be increased either by stepping up to a higher-capacity server or by increasing the number of servers. Each product lists the servers it supports and its scalability model. The products also list their tested limits on the servers they support. In theory, one can simply follow these limits and models and come up with the required number of servers for a particular deployment.

In practice, however, sizing is not so simple. For one thing, there are several limits that apply to any deployment. For example, a Unified CM server may be qualified to register 2,500 users and define up to 500 regions. The Unified CM cluster composed of such servers will need more servers if either of these limits is exceeded while the other values are still within limits. Moreover, some of these limits are not absolute but change dynamically based on what else has been configured in the system.

The other major challenge in a sizing exercise is the interaction among components. Unified CM plays a central role in almost all Unified Communications deployments, and it is affected by how customers choose to use other systems. For example, the addition of Cisco Unified MeetingPlace to enable conferencing would tend to concentrate a large number of call setups into a short period (at the beginning of conferencing sessions) and thereby increase the stress on Unified CM during that short period, and this must be accounted for in Unified CM sizing.

Server variations also need to be considered. For example, Unified CM running on a Cisco MCS-7815 or MCS-7816 server is only a standalone entity and may not be clustered. Similarly, 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.

From a customer perspective, the sizing exercise consists of itemizing all of the functions that are expected in the proposed deployment. Some of these performance factors are obvious, but others are not. For example, one may correctly surmise that the busy hour call attempts (BHCA) that the system is expected to handle is a key factor of performance expectations. But there are nuances even in BHCA that need consideration, such as the types of calls. 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. The expected number

of phones and users is another example of an obvious factor that would affect sizing. Here again, the type of phones, the number of lines that they are configured with, and whether they are in secure mode, among other things, have an impact on Unified CM sizing.

Because of all these factors and possible variations, a proper sizing exercise is complex and must be well understood, especially for large deployments. This chapter provides guidance on the significant factors that consume resources, and their impact on the system, which must be estimated accurately in order to do a complete and accurate sizing.

Cisco Unified Communications Sizing Tools

You should not expect to be able to perform sizing for complex systems after reading this chapter. On the contrary, manual calculations of all the sizing factors is not practical. However, this chapter will enable you to gain an appreciation of the factors that significantly affect the performance of the system as a whole and that must be accounted for in any sizing effort.

To assist in accurate sizing, Cisco provides several tools that do the calculations based on these significant performance factors. These tools take into account data from testing, individual server performance, advanced and new features in product releases, design recommendations from this SRND, and other factors. The tools allow you to enter specific deployment information, and they apply their sizing algorithms on your supplied data to recommend a set of hardware resources. Obviously, for a desired deployment, this recommendation is only as good as the accuracy of the input data. User guides for the tools contain an explanation of the inputs and how they can best be collected from an existing system or estimated for a system still in the design stage.

The sizing tools are available at http://tools.cisco.com/cucst and they include the following:

- Cisco Unified Communications Sizing Tool — Guides users through a complete system deployment consisting of Cisco Unified CM, voice messaging, conferencing, gateways, Cisco intercompany Media Engine (IME), Cisco TelePresence Management Suite (TMS), and Cisco Unified Contact Center components.
- Cisco Unified Communications Manager Session Management Edition (SME) Sizing Tool — A specialized tool that focuses on the specific functions of a Unified CM Session Management Edition deployment.
- Cisco VXI Sizing and Configuration Tool — A specialized tool for sizing the Cisco Virtual Experience Infrastructure (VXI).

Access to the sizing tools is limited to users who have a qualified Cisco login account. 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

Unified Communications Sizing Compared with PBX Sizing

PBX sizing in the past has mostly been about PSTN access trunks. Consequently, the processes for determining how many trunk circuits are required for a given user base and the desired level of service are well documented. Well known models such as the Erlang B, Extended Erlang B, Erlang C, and other models are used for that purpose. However, sizing a Unified Communications system is inherently more complex for the following reasons:

- Unified Communications is not a monolithic system. Rather, it is composed of several servers doing different things but communicating with each other.
- Unified CM performs many more functions and provides many more services than a PBX.

Definition of Terms

The following terms are used throughout this chapter:

Simultaneous Calls

The number of calls that are all active in the system at the same time.

Maximum Simultaneous Calls

The maximum number of simultaneous calls in active (talk) state that the system can handle at one time.

Calls per Second

The call arrival rate, described as the number of calls that arrive (that is, new call setup attempts) in one second. Call arrival rates are also often quoted in calls per hour, but this metric is looser in the sense that 100 calls arriving in the last five seconds of an hour provides an average call arrival rate of 100 calls per hour (which is an extremely low rate for a communications system), while it also provides an arrival rate of 20 calls per second (which is a high rate). Sustaining 20 calls per second for an entire hour would result in 72,000 calls per hour. Therefore calls-per-hour is not a very useful metric for ascertaining a system's ability to handle bursty call arrival traffic patterns.

Busy Hour

The busiest hour of the day when people are most likely to use their phones. This hour varies from organization to organization and from industry to industry. But for most it is likely to be either in the morning (for example, 9AM to 10AM) or in the afternoon (for example, 2PM to 3PM).

Busy Hour Call Attempts (BHCA)

The number of calls attempted during the busiest hour of the day (the peak hour). This is the same as the calls-per-second (cps) rating for the busiest hour of the day, but it is expressed over a period of an hour rather than a second. For example, 10 cps would be equal to 36,000 calls per hour. There is also a metric for Busy Hour Call Completions (BHCC), which can be lower than the BHCA (call attempts) under the assumption that not all calls are successful (as when a blocking factor exists). This chapter assumes 100% call completions, so that BHCA = BHCC.

Blocking Factor

The maximum percentage or fraction of call attempts that may be blocked during the busy hour. A blocking factor or 0.0 would mean that the number of circuits is equal to the number of callers, which is unrealistic for most deployments.

Average Hold Time

This is the period of "talk time" on a voice call; that is, the period of time between call setup and 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. The shorter the hold time on the average call, the greater the percentage of system CPU time spent on setting up and tearing down calls compared to the CPU time spent on maintaining the speech path.

Bursty Traffic

Steady arrival means the call attempts are spaced more or less equally over a period of time. For example, 60 calls per hour at a steady arrival rate would present one call attempt roughly every minute (or approximately 0.02 cps). With bursty arrival, the calls arriving over a given period of time (such as an hour) are not spaced equally but are clumped together in one or more spikes. In the worst case, an arrival rate of 60 calls per hour could offer all 60 calls in a single second of the hour, thus averaging 0 cps for most of the hour with a peak of 60 cps for that one second. This kind of traffic is extremely stressful to communications systems.

Erlang

An Erlang is a unit of measure for communications traffic. It is used to represent the utilization of a resource over a one-hour period. One Erlang means that one resource was used 100% of the hour. This could be due to a single call of one-hour duration or multiple sequential calls whose durations total to one hour. Therefore, if 10 Erlangs are required, it is necessary to have 10 resources to ensure that all traffic is serviced.

Designing for Performance

After analyzing the functional requirements and determining the appropriate products for a Unified Communications system, the next major question is how to design the network so that it is able to adequately deliver acceptable performance as measured by availability, reliability, response time, and quality of service. Can the system cope with the real-time performance requirements, support the desired number of users, and still scale up to meet the increasing needs of the foreseeable future?

To aid the Unified Communications network designers with answers to these questions, Cisco tests each of the products for its performance characteristics. The results are published and broad recommendations are made regarding the size and number of clusters, servers, and other components that should be deployed for supporting the given number of users. To a large extent these test results, combined with the design recommendations in this document, provide sufficient information for most Unified Communications deployments. For others, however, the system designers will need a deeper understanding of how each product works and how users will use it before a viable hardware set can be selected. The selection of such a set can also be complicated by the following concerns that should be addressed:

- System release
- Complexity of the configuration
- Utilization of options such as trace compression, call detail recording (CDR), call management record (CMR) generation, and so forth
- Interaction between individual products
- Anticipated growth
- Use of external applications
- Average and peak usage

Quantitative Analysis of Performance

Testing for performance analyzes the product under test for a set of basic functions it is designed to perform. For example, Unified CM performs many functions and each function requires a finite amount of CPU and memory. Unified CM handles endpoint registrations, user initiated calls, database queries, and many other functions. Performance testing involves testing of each of these basic functions in isolation, measuring the computing resources that are utilized as these functions are executed in an increasing volume.

A quantitative analysis of the performance characteristics of a software system given the hardware platform is done in a series of tests that aim to determine the linear range of the system operations. A linear range is where the amount of resources used and the throughput achieved vary in direct proportion with each other. This range is critical because, if the system does not exhibit linear behavior, its performance is unpredictable. Most systems exhibit linearity within a certain range, beyond which the system's performance becomes unpredictable. Therefore, the design must ensure that the system operates within the parameters of the linear range.

Conversely, putting together a system for deployment consists of decomposing the requirements into sets of basic functions, comparing them against the published test results, and determining the set of servers that meets the performance needs in their linear range of operation.

Performance Modeling of Computer Systems

The first step in determining how much a computer system can accomplish is to itemize the various tasks it is called upon to perform. For example, Unified CM may be required to do all of the following tasks:

- Initialize configured values such as those for endpoints, directory numbers, dial plans, and so forth.
- Perform endpoint registrations, which requires handling the initial registration messages, looking up databases to find their configuration information, and creating configuration files for the endpoints to download.
- Maintain endpoint registrations by handling periodic registration messages
- Handle new call requests, which can be a fairly complex process consisting of ensuring user entitlement, analyzing dialed digits, determining the destination (either another phone, gateway, or trunk), assembling the correct signaling based on rules stored in the database, and transmitting and receiving call signaling messages.
- Provide mid-call feature requests such as transfer and conference.
- Offer user management and requests for functions such as Do Not Disturb, Call Forward, and so forth.

Each of the functions that a computer system performs requires it to spend some of its resources consisting of CPU, memory, and disk I/O.

The linear operating range of the system under test is determined by subjecting the system to a battery of tests. Some tests that attempt to find this range are described in this section. From this linear range of operation, the cost incurred in terms of CPU, memory, and disk I/O can be determined for each incremental unit of the operation.

For example, memory utilization of each additional endpoint of a certain type can be determined from the slope of the line depicting the amount of memory used for a range of endpoints. Similarly, memory utilization for each registering endpoint and for each additional call can be quantified by using the same techniques.

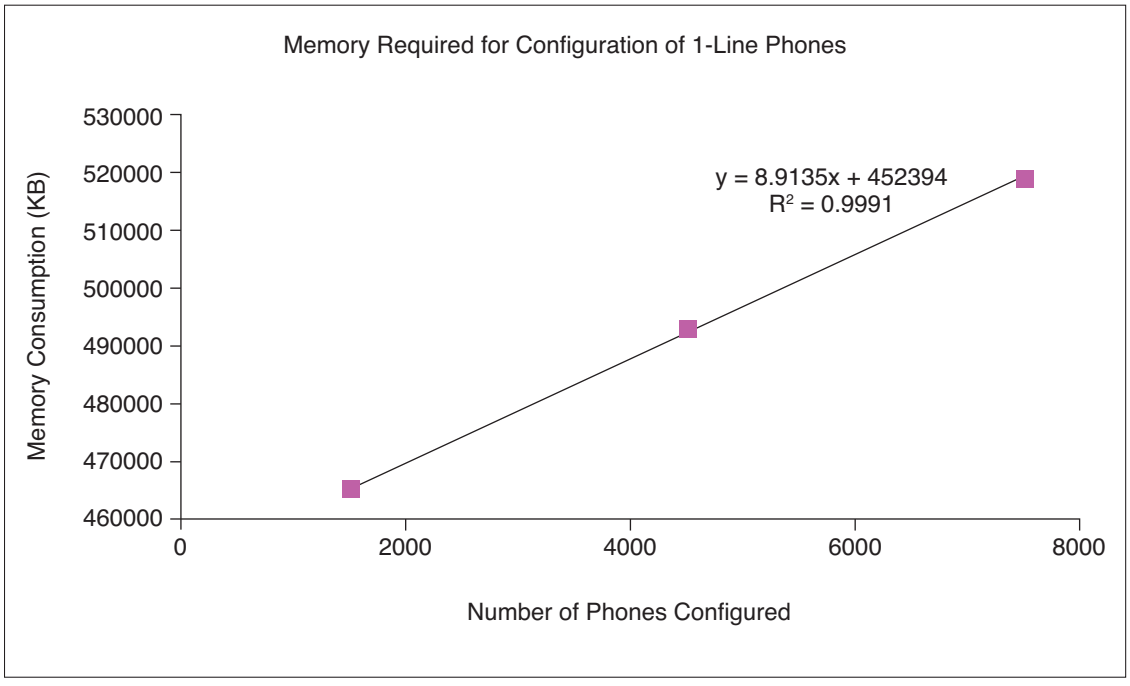
Memory Usage Analysis

Two types of memory usage are identified in the system: static and dynamic. Static memory is defined as the amount of memory that is in use even when there is zero call traffic. This usage of memory arises from configuration data, registration of endpoints, and other factors. Dynamic usage of memory results from call activity. Each active call requires its context to be saved, which results in a certain amount of memory being utilized for the duration of the active call. Thus, whereas static memory is a function of the number of endpoints, dynamic memory is a function of the number of concurrent calls, which itself is a function of the call rate (calls per second) and the average hold time (AHT) per call.

In practice, system memory is also required by the operating system (OS) and by other processes, so the net memory available for operations (static and dynamic memories) is somewhat less than the total memory available on the platform. In addition, some memory is needed for other processes and services running in the system and for any unforeseen spikes in usage.

Figure 29-1 shows the results of a test conducted to determine the memory requirements for configuring one-line phones. It shows the memory consumed by simply configuring 1500, 4500, and 7500 IP phones in Unified CM. Linear regression techniques are used to draw a trend line through the data points. The equation of this trend line is then determined, as is the correlation coefficient R^2. A correlation coefficient of 1 or very close to it (at least 0.99) indicates that the trend is linear and that the equation of the trend line is valid and may be used to predict the dependent variable (in this case memory) based on the control variable (the number of phones).

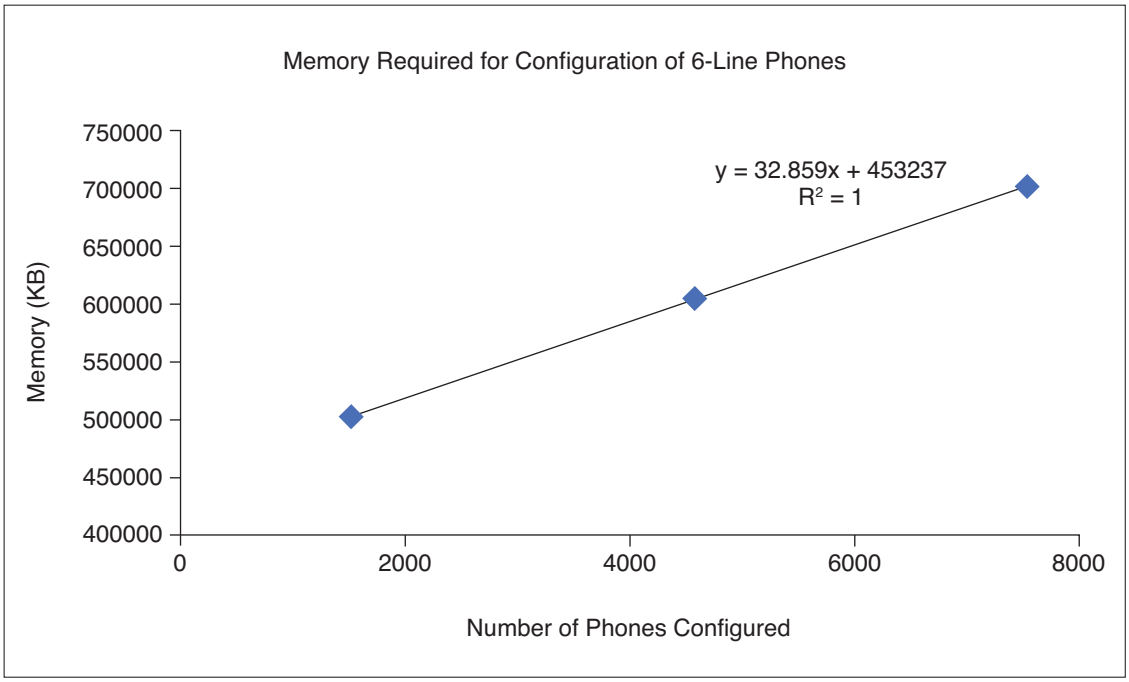
Figure 29-1 Memory Required for Configuration of One-Line Phones



In this particular experiment the R^2 value is extremely close to 1 (discounting small errors in measurement) and the equation for the trend line is valid. From the equation we can derive that the memory consumed with no phones is 452,394 Kbytes (the Y-intercept) and that each additional one-line phone configured in the system consumes 8.91 Kbytes.

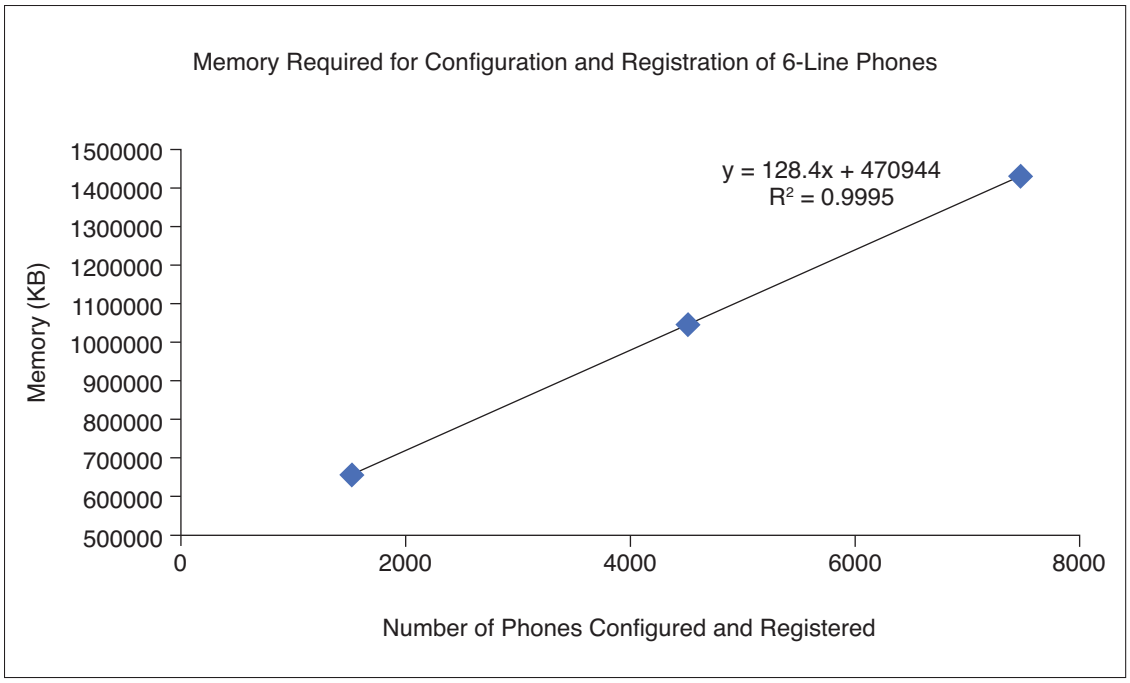
Figure 29-2 depicts the memory requirements for configuring six-line phones. In this chart R^2 is actually equal to 1, indicating that the trend line is a valid model. From the equation we can determine that configuring each six-line IP phone consumes approximately 33 Kbytes of memory.

Figure 29-2 Memory Required for Configuring Six-Line Phones



The other component that makes up the static memory – the memory required for registration of phones – can also be estimated in the same manner. Figure 29-3 shows the tested, measured, and plotted memory requirements for configuring and registering 1500, 4500, and 7500 phones, each with six lines. Note that R^2 is close enough to 1 to make the trend line a valid model. From the equation we can determine that registration of each six-line IP phone consumes approximately 128 Kbytes of memory.

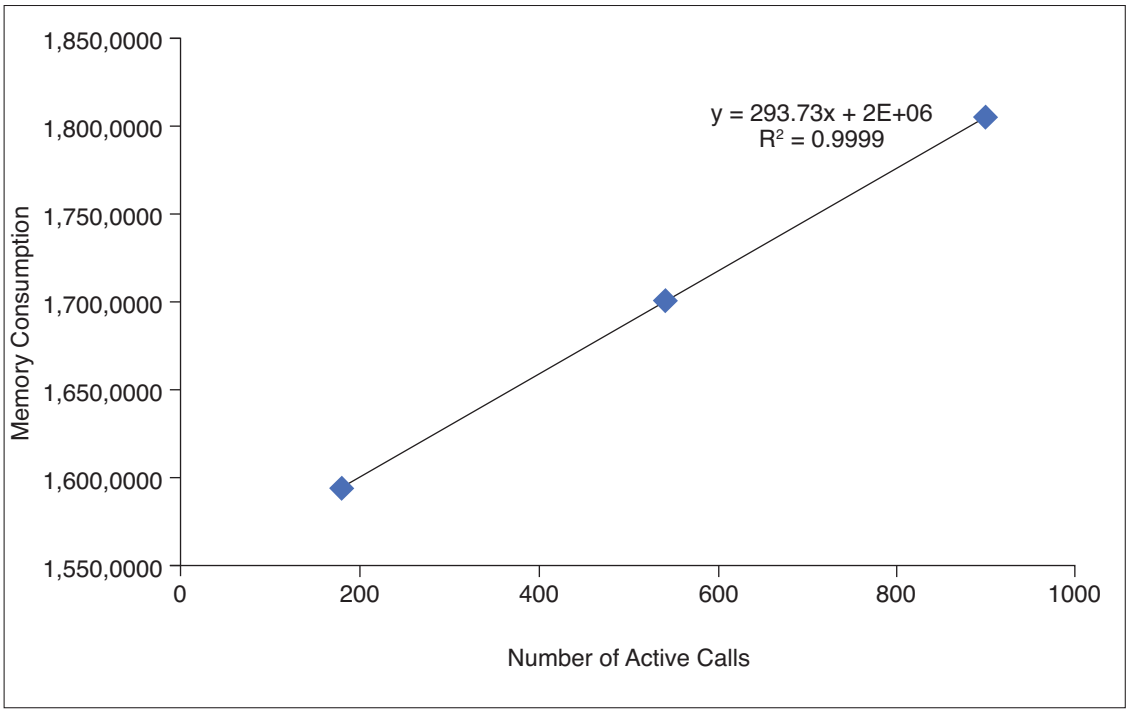
Figure 29-3 Memory for Configuration and Registration of Six-Line Phones



Static memory also includes other configuration items such as partitions, translation patterns, route lists and groups, as well as memory used for CTI and other applications.

Another type of memory called the dynamic memory is defined as the memory used for active calls. In contrast to static memory, which stays allocated all the time, dynamic memory is allocated for each call attempt and remains only until the end of the call. Figure 29-4 shows how the memory is utilized for 180, 540, and 900 active calls on one subscriber node of the Unified CM cluster. The graph shows that the trend line is a good fit and that approximately 294 Kbytes are used for each active call.

Figure 29-4 Memory Consumption Per Active Call



The preceding graphs and analysis are indicative of how memory is measured in the system. From a set of these observations, data may be interpolated that can start to build a memory model for various activities going on in the system. For example, we can estimate:

- Incremental memory required for configuring each additional line
- The maximum number of calls that can exist in the system before it runs out of memory and starts paging

A major determinant of dynamic memory usage is the average call holding time (ACHT), which is the average duration of each call. A longer ACHT means that more memory will be used in the system because there will be a larger number of active calls present at any time.

The description provided in this section has been simplified. Further complexities arise from the variety of phones that can be configured on Unified CM with different protocols, capabilities, security status, and other variables. Each of these variants is tested and analyzed. Furthermore, each of these variables depends on the software release, which could add improvements and new features. For active call measurements, the various types of calls that can be made between different destinations, such as between two SCCP phones or between a SIP phone and an MGCP gateway, is also considered.

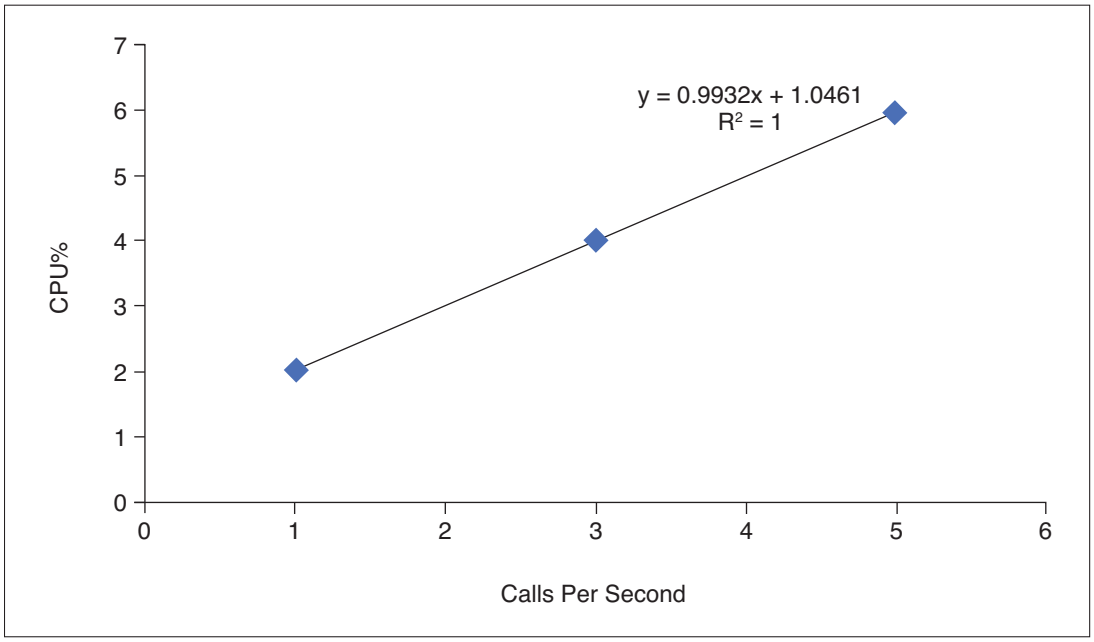
CPU Usage Analysis

Analysis of CPU usage follows the methodology used for memory analysis. While there is some CPU activity even when there are no calls being initiated or terminated, most of the CPU utilization occurs during the process of setting up or tearing down calls. Therefore, one of the key determinants of CPU usage is the call rate.

There are significant differences between the types of calls being made. Calls can originate and terminate within the same server, or they can be made between two servers. Calls can also originate from the Unified CM cluster and travel across a gateway or a trunk. All of these different call activity types impact the CPU differently, so it is important to consider them carefully.

Figure 29-5 shows CPU utilization as measured at 1, 3, and 5 calls per second. Because the trend line is linear, we can conclude that the CPU processing cost required to process one incoming call each second is about 1%.

Figure 29-5 CPU Consumption Per Call Setup



As with memory analysis, CPU usage involves many complexities that must also be considered. For example, CPU usage analysis must account for different costs of terminating and originating calls, different protocols, whether the calls are secure or not, and so forth. CPU usage also depends on whether or not the configuration database is complex or relatively simple, whether CDRs and/or CMRs are being generated, whether detailed tracing is being used, and so forth.

Whereas incremental memory usage is fairly independent of the actual server platform, CPU usage will vary substantially with the actual hardware being tested. Therefore, the same tests must be repeated on all servers that are supported.

Other CPU-intensive call operations such as call transfers, conferences, media resource functions such as MTP or music on hold, and so forth, should also be considered when sizing CPU resources.

Shared lines also consume CPU resources. Not only do shared lines count as extra lines on the phones that share DNs, but each call from or to any of the shared line phones is reflected on all of the other phones as well.

Fundamentals of Voice Traffic Engineering

Traffic engineering is the science of determining an optimum number of resources given the key usage data. In telephony this user data includes the busy hour call attempts (BHCA) and the average hold time (AHT). The BHCA measures all the calls that an average user initiates or receives during the busiest hour of the day. The AHT measures the time that the user spends on the phone for each initiated or received call. An individual's BHCA, when multiplied by the number of users, gives the volume of calls that the system must be able to handle. Once we have the total BHCA and the AHT, we can calculate the Erlang value that the system should be able to handle. One Erlang is a full hour of telephone conversation. For example, if the system BHCA is 10 and the calls last for 3 minutes each, then the system is being used for a total of 30 minutes and the equivalent Erlang value is 0.5.

Note This document assumes that traffic follows the Extended Erlang B model with random arrival pattern and that blocked callers make multiple attempts to complete their calls. For a more thorough discussion of the various Erlang models used in the industry, refer to the information at http://www.erlang.com/calculator/.

While this analysis reveals the total BHCA that the system must be able to handle at the given AHT, another key piece required for analysis is the blocking factor. It is well understood that deploying a telephony system that has enough capacity for all of the users to be on the phone all of the time would be prohibitively expensive, especially for larger systems. It follows, therefore, that if more than a certain number of callers try to access the system at the same time, some callers will necessarily be blocked. A key decision in system deployment is how many over-the-limit callers may be blocked during peak calling times. The amount of resources required for providing a smaller probability of being blocked, say 0.01 or 1%, would be more than the amount required to provide a blocking factor of 0.1 or 10%.

The Erlang value and the blocking factor are useful for calculating the amount of shared resources that must be provisioned in the system. For example, with these pieces of information one can figure out how many DS0s will be required on gateways for a system that has a given number of Erlangs of through traffic with the required blocking factor. This is generally done through an Erlang calculator or lookup tables. The number of required DS0s would increase with the number of Erlangs and decrease with an increase in the blocking probability.

Table 29-2 illustrates the relationship between number of circuits, blocking probability, and busy hour traffic.

Table 29-2 Erlang C Traffic Table (Maximum Offered Load)

Number of Circuits	Blocking Probability							
	0.01%	0.05%	0.1%	0.5%	1.0%	2.0%	5.0%	10.0%
1	0.0001	0.0005	0.0010	0.0050	0.0100	0.0200	0.0500	0.1000
2	0.0142	0.0319	0.0452	0.1025	0.1465	0.2103	0.3422	0.5000
3	0.0860	0.1490	0.1894	0.3339	0.4291	0.5545	0.7876	1.0400
4	0.2310	0.3533	0.4257	0.6641	0.8100	0.9939	1.3190	1.6530
5	0.4428	0.6289	0.7342	1.0650	1.2590	1.4970	1.9050	2.3130

From Table 29-2 we can determine the following information:

- The number of Erlangs that the system can handle increases with the number of circuits and with the blocking factor. Whereas the first relationship is obvious, the second can be understood by realizing that a greater number of calls are being blocked.
- Given an Erlang requirement of 0.50 and a blocking factor of 0.1%, the system would need 5 circuits.
- Assuming we have 5 circuits and a blocking factor of 1%, there would be 1.259 Erlangs available. It then follows that if we have 10 users, each user can talk for (1.259∗3600/10) = 453.24 seconds during the busy hour.

Note Specifically for Cisco Unified Contact Center deployments, there might be other resources that have to be sized according to the same principles. For example, requirements for the number of interactive voice response (IVR) ports and agents are modeled using similar quantitative analysis. Some of the considerations here besides average hold time and BHCA include time waiting in queues and other factors, which means that a higher number of DS0 circuits will be required. For a full description, 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.

Sizing by Product

This section discusses significant factors that influence sizing of the following individual products and describes how these individual products influence the sizing considerations of other products in the system deployment:

- Cisco Unified Communications Manager Express, page 29-14
- Cisco Business Edition, page 29-14
- Cisco Unified Communications Manager, page 29-18
- Cisco Unified CM Megacluster Deployment, page 29-30
- Cisco Unified CM Session Management Edition, page 29-30
- Cisco Intercompany Media Engine, page 29-31
- Emergency Services, page 29-32
- Media Resources, page 29-28
- Gateways, page 29-32
- Voice Messaging, page 29-38
- Collaborative Conferencing, page 29-39
- Cisco Unified Presence, page 29-45
- Cisco Unified Communications Management Suite, page 29-46

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 1861 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 are performing, 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

The three models of Cisco Business Edition – Business Edition 3000, 5000, and 6000 – offer different capacities measured in terms of number of users, number of endpoints, and maximum call volumes. Table 29-3 describes the pertinent performance characteristics of the three models.

Table 29-3 Capacities of Cisco Business Edition Models

Model	Maximum Number of Users	Maximum Number of Endpoints	Maximum BHCA
Business Edition 3000 (MCS 7816)	300	400	2,200
Business Edition 5000 (MCS 7828)	500	575	3,600
Business Edition 6000 (UCS C200)	1,000	1,200	5,000

Busy Hour Call Attempts (BHCA) for Cisco Business Edition

As shown in Table 29-3, Business Edition 3000 supports a maximum of 2,200 BHCA, Business Edition 5000 supports a maximum of 3,600 BHCA, and Business Edition 6000 supports a maximum of 5,000 BHCA. When calculating your system usage, stay at or below the BHCA maximum shown in Table 29-3 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 supports a maximum number of endpoints as indicated in Table 29-3, actual endpoint capacity depends on the total system BHCA.

Note Business Edition 3000 supports a limited set of endpoints. For a list of the supported endpoints, refer to the *Administration Guide for Cisco Business Edition 3000*, available at http://www.cisco.com/en/US/products/ps11370/prod_maintenance_guides_list.html.

A trunk device carries multiple calls to more than one endpoint. It can be any trunk or gateway device such as a SIP trunk, a gatekeeper-controlled H.323 trunk, or in the case of Business Edition 3000 an MGCP backhauled PRI trunk.

Business Edition 5000 and 6000 both support intercluster trunking as well as H.323, SIP, and MGCP trunks or gateways and analog gateways. However, Business Edition 3000 does not support intercluster trunking. Business Edition 3000 trunk and gateway support is limited to the Cisco 2901 Integrated Services Router (ISR) for MGCP PSTN connectivity over a maximum of two E1/T1 PRIs. Business Edition 3000 also supports the Cisco VG224 Analog Voice Gateway for analog phones.

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 BHCA maximum specified in Table 29-3.

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 BHCA as shown in Table 29-3.

If you are using Cisco Unified Mobility for Mobile Connect (also known as single number reach, or SNR) on Business Edition 5000 or 6000, or if you are using the Reach Me Anywhere feature (also SNR), 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 25-63, 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).

Note Cisco Business Edition 3000 does not support media authentication or encryption.

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.

Business Edition 5000 with Cisco Unified Contact Center

For this example, assume that Cisco Unified Contact Center Express (Unified CCX) is integrated with Business Edition 5000 and that the system has the following characteristics:

- The required specification is for 15 contact center agents with a maximum of 30 calls per hour during the busiest hour.
- There are 96 non-agent users with average usage of 4 BHCA, and each user has the ability to configure one remote destination for single number reach with Cisco Unified Mobility.
- There are 36 non-agent users with heavy usage of 10 BHCA, and each also has the ability to configure one remote destination for single number reach.
- There are 20 extra shared lines, 10 of which are shared across 10 users from the average usage pool as well as 10 in the heavy usage pool.
- There are 7 T1 trunks (allowing for up to 161 simultaneous calls) with a total of 1200 BHCA across all trunks.

Note Cisco Business Edition 5000 is not supported with Cisco Unified Contact Center Enterprise.

Note This example groups the BHCA for all gateway trunks into a single total trunk BHCA value. This method would be typical for a single-site deployment. However, in a multisite deployment, the various sites' trunks could have different BHCA requirements and thus require different BHCA groupings.

The BHCA calculations for this system are as follows:

15 contact center agents at 30 BHCA = 450 BHCA

96 average-usage users at 4 BHCA = 384 BHCA

36 heavy-usage users at 10 BHCA = 360 BHCA

10 shared lines in the 4 BHCA group = 20 BHCA

10 shared lines in the 10 BHCA group = 50 BHCA

Total of 1200 BHCA for all T1 trunks = 1200 BHCA

One remote destination for single number reach across each of the 96 average-usage users at 4 BHCA = 192 BHCA. (See Capacity Planning for Cisco Unified Mobility, page 25-63, for details on this calculation.)

One remote destination for single number reach across each of the 36 heavy-usage users at 10 BHCA = 180 BHCA. (See Capacity Planning for Cisco Unified Mobility, page 25-63, for details on this calculation.)

Total BHCA for all endpoint devices in this case is:

(450 + 384 + 360 + 20 + 50 + 192 + 180 + 1200) = 2,836 BHCA

This level of usage is acceptable because it is below the system maximum of 3,600 BHCA, and it allows for future growth of approximately 800 BHCA.

This sizing example applies exclusively to Business Edition 5000. Business Edition 3000 is not capable of trunking to Unified Contact Center deployments. Business Edition 6000 runs Unified Contact Center Express co-resident; and although sizing considerations are similar, this example is specifically related to Business Edition 5000.

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

- Cisco Business Edition 3000

  http://www.cisco.com/en/US/products/ps11370/tsd_products_support_series_home.html
- Cisco Business Edition 5000

  http://www.cisco.com/en/US/products/ps7273/tsd_products_support_series_home.html
- 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

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is the hub of any Unified Communications deployment. It performs the most basic functions and controls endpoints, routes calls, enforces policies, hosts applications, and in general anchors other Unified Communications products such as gateways, Cisco Unity Connection, Cisco Unified MeetingPlace, Cisco Unified Contact Center suite of products, and others. These applications depend on Unified CM to function and in turn affect Unified CM's performance, which must be accounted for in Unified CM sizing.

The following factors affect Unified CM performance and must be considered when sizing a Unified CM deployment:

- Server and cluster maximum capacities
- System-level settings such as database complexity, trace level, and so forth
- Number and types of endpoints that are registered on Unified CM
- Number of users
- Traffic mix
- Dial plan
- Applications within Unified CM (Extension Mobility, WebDialer, and other CTI-enabled applications)
- Media resources hosted by the subscribers using the Cisco IP Voice Media Streaming Application

Server and Cluster Maximums

Although it is not practical to list every minute detail needed to accurately determine the number of Unified CM servers required for a particular sizing calculation, there are certain server and cluster maximums that must be observed, and some of these values change with Unified CM software version:

- Each cluster can support configuration and registration for a maximum of 40,000 secured or unsecured SCCP or SIP phones with Unified CM 8.6(1) and later releases.
- Each cluster can support configuration and registration for a maximum of 30,000 secured or unsecured SCCP or SIP phones with Unified CM 8.5 and earlier releases.
- Two TFTP servers are required if the number of endpoints in the cluster exceeds 1,250.
- Support for CTI connections has improved over the last several releases, and each cluster can support a maximum of 40,000 CTI connections.
- The number of call processing subscribers in a cluster cannot exceed 4, plus 4 standby, for a total of 8 call processing servers. Also, the total number of servers in a cluster, including the publisher, TFTP, and media servers, may not exceed 20.

The following sections describe how each of these components of Unified CM affects its sizing and therefore must be considered in an analysis of a given system description.

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.

Tracing Level

The system supports two tracing levels: default and detailed. When the level is set to detailed, about 20% more CPU resources are required as compared to the default option.

Database Complexity

There is really no one measurement to determine if the database of configuration information in Unified CM should be considered as simple or complex. As a general rule, if you have more than a few thousand endpoints and more than a few hundred dial plan elements such as translation and route patterns, hunt pilots, shared lines, and so forth, then the resulting database should be considered complex. The CPU usage is considerably higher when the underlying database is complex.

Call Detail and Call Management Records

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

Trace Compression

Beginning with Cisco Unified CM 8.0, traces are always compressed and the compression may not be turned on or off. For earlier releases, turning on compression saves disk space but adds to CPU utilization.

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 29-4 lists these limits for some of the Unified CM server platforms.

Table 29-4 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-7815 and 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.

High Availability

Deploying redundant servers increases the number of total servers required in the solution. After figuring out the minimum number of servers required for the specified deployment, add the desired number of subscriber servers. Redundancy options are described in the chapter on Unified Communications Deployment Models, page 5-1. Note that some servers do not lend themselves well to redundancy.

Number of Servers per Cluster

A cluster may be configured to consist of from one to four subscriber pairs. Reducing the number of subscriber pairs per cluster may increase the number of clusters, and hence the number of total servers, required for a given sizing analysis. An increase in the number of clusters can sometimes be desirable if the deployment consists of geographically distributed equally large locations or if any cluster-wide limit is forcing a new cluster even if the per-server utilization is low.

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 the hypervisor. Different templates specify different capacities. For example, the 10000 template defines a virtual machine with 4 virtual CPUs, 6 GB of RAM, and 160 GB of hard disk space that has a maximum capacity of serving up to 10,000 endpoints. There are similar templates defined for 1000, 2500, and 7500 endpoints as well.

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

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 and Users

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

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

Each type of endpoint defined in the system uses system resources – for example, static memory just by being defined and registered, and CPU and dynamic memory based on its call rate. Each endpoint could also place additional load on Unified CM by running applications interacting with services running inside of Unified CM.

There are defined maximum supported quantities of endpoints for a given server, as shown in Table 29-5. Note that these values are guidelines only, and it is possible that the system may not be able to support these maximum amounts because of other applications running in the deployment.

Table 29-5 Maximum Number of Endpoints Per Server Platform or OVA Template

Server Platform Characteristics	Maximum Endpoints per Server or OVA Template	High-Availability Server
Cisco MCS 7845-I3 or OVA equivalent	10,000	Yes
Cisco MCS 7845 (All other supported models) or OVA equivalent	7,500	Yes
Cisco MCS 7835 (All supported models) or OVA equivalent	2,500	Yes
Cisco MCS 7825 (All supported models) or OVA equivalent	1,000	No
Cisco MCS 7816 (All supported models)	500	No
Cisco MCS 7815 (All supported models)	300	No

The designation of High Availability in Table 29-5 indicates whether those servers may be paired for high availability within a cluster consisting of those servers.

Some endpoints may operate in one of two modes. Endpoints such as Cisco Unified Personal Communicator and those based on Common Services Framework (CSF), such as Cisco WebEx Connect, Cisco UC Integration™ for Microsoft Lync, and others, can work either as soft-phones registered directly with Unified CM as phones, or in desk-phone control mode where they act as applications that use CTI to communicate with Unified CM to control a desk phone. Either way, they use Unified CM resources (endpoints or CTI applications) but count against different operating limits.

Along with the endpoints, the number of busy hour users must also be taken into account. The number of users and their collective usage of the endpoints determine the call processing load on the system.

Call Traffic

Next to the number of endpoints, the quantity and quality of call traffic places the second biggest requirement on 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. A single server needs to handle both halves for calls made between two endpoints registered on it. For calls made between two servers in the same cluster, each of the participating servers needs to handle only half of the call. For calls made between endpoints registered on different clusters, a server and the cluster as a whole need to handle only half of each call. For calls made between an endpoint in a cluster and the PSTN, a PSTN gateway needs to handle half of the call, and these calls form the basis for sizing the gateways themselves.

When considering call traffic, other complexities arise from calls between endpoints that work on different protocols, such as between SIP and SCCP-based phones, if calls are transferred, and if conferencing is invoked.

In general, the following factors require consideration:

- 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 rate of call placement, or the BHCA, 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 higher 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. As another example, 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 static configuration elements that determine call routing and associated policies. In general, dial plan elements occupy static memory space in Unified CM servers, and 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, and translation 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 only a limited amount of shared system memory available.

Of the above dial plan elements, the number of lines shared across multiple endpoints is of particular interest. Each shared line multiplies the CPU cost of a call setup because the call has to be presented to all the endpoints that share that particular directory number.

Another aspect of a large dial plan that comes into play is the space required to hold the elements of the plan in the Informix Database System. There is only a finite amount of disk space available to hold the entire configuration of Unified CM, and extra-large dial plans can overwhelm it. In this case, the only option may be to break up the dial plan and use its parts 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, and Contact Center depend on CTI to function.

Historically CTI interactions have been relatively expensive operations in Unified CM that severely limited system scalability, but recent optimizations have reduced their impact on scalability. Although the high-end server platforms for Unified CM 8.6(1) and later releases are able to support CTI for all of their registered devices, the lower-end platforms do not scale that high. Table 29-6 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 29-6 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 with Unified CM 8.6 and later releases; otherwise 400
MCS 7825-I1/I2	800
MCS 7825-I3/I4	900
MCS 7825-I5 and OVA equivalent	1,000 with Unified CM 8.6 and later releases; otherwise 900
MCS 7835-I1/I2	2,000
MCS 7835-I3 and OVA equivalent	2,500 with Unified CM 8.6 and later releases; otherwise 2,000
MCS 7845-I1	2,500
MCS 7845-I2 and OVA equivalent	5,000
MCS 7845-I3 and OVA equivalent	10,000 with Unified CM 8.6 and later releases; otherwise 5,000

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 with Unified CM 8.6 and later releases; otherwise up to 2 lines per controller device)
- The number of shared occurrences of a line controlled by CTI (up to 5 per line with Unified CM 8.6 and later releases; otherwise up to 2 per line)
- The number of active CTI applications (up to 5 for any device with Unified CM 8.6 and later releases; otherwise up to 2 for any device)
- A maximum BHCA of 6 per controlled device

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

Table 29-7 lists the number of supported CTI devices for Cisco Business Edition.

Table 29-7 Users and CTI Devices in Cisco Business Edition

Model	Maximum Number of Users	Maximum CTI Devices
Business Edition 3000	300	400
Business Edition 5000	500	575
Business Edition 6000	1,000	1,200

Determining CTI Resources Required 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 29-8.

Table 29-8 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
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 29-9.

Table 29-9 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 29-8 and Table 29-9, 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 29-8 and Table 29-9, 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 29-8 and Table 29-9, 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

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 29-10 shows the maximum number of EM and EMCC logins per minute for each type of server.

Table 29-10 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
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. 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 8.*x* 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 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 29-10.)

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 19-19.)

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 2 call requests per second (7,200 calls per hour) 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 2 cps, therefore 2 nodes should 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, by the virtue of the Cisco IP Voice Media Streaming Application, may be used for certain media functions that can be performed in software only and do not require hardware resources. Unified CM can act as a media termination point (MTP), as a conference bridge, or as a source of music-on-hold streams. Although the capabilities of Unified CM are limited in comparison 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

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

While the Cisco IP Voice Media Streaming Application provides MTP, annunciation, and conferencing capabilities, you might find it more scalable to place this functionality 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 29-11 lists the maximum values that may be configured for each of these services.

Table 29-11 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 17-1.

Music on Hold

Table 29-12 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 29-12 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: 250 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: 500 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 29-12 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.

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 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

Therefore, care should be taken when deciding whether or not to deploy a megacluster.

Cisco Unified CM Session Management Edition

The Session Management Edition (SME) is Unified CM in a specific deployment mode. Thus, all the sizing discussion for Unified CM applies to SME as well. The big distinction is that the call traffic in a pure SME deployment is solely through trunk interfaces rather than through line interfaces. Therefore, sizing an SME cluster is in general a simpler exercise than for Unified CM as a whole.

An SME cluster follows the same guidance as that for 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, assess 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 advances configurations, the SME may also host centralized voice messaging, mobility, and conferencing. 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 these trunks.

The following considerations apply when sizing an SME cluster:

- 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, the following relevant sizing parameters come into play as well:

- If using centralized voice messaging, the percentage of calls that are sent to voice mail
- If using mobility, the number of users and the remote destinations per user
- If using conferencing, 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.

For SME sizing calculations, use the Cisco Unified CM SME Sizing Tool, available at http://tools.cisco.com/cucst.

Cisco Intercompany Media Engine

The sizing of servers used for running the Cisco Intercompany Media Engine (IME) depends solely on the quantity of directory numbers enrolled for the IME service. Table 29-13 lists the capacity of each supported server.

Table 29-13 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 29-14 provides capacity limits for the ASA-5550 and ASA-5580. Performance limits of other ASA models have not been validated yet.

Table 29-14 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 29-15 shows the server platforms that support the Emergency Responder and their maximum capacities.

Table 29-15 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

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

The capacity of Emergency Responder also affects the Unified CM cluster size. There can be only one Emergency Responder active per cluster. Therefore, choose the server that has sufficient resources to provide emergency coverage to 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

The expected traffic in and out of gateways is the key to calculating the number of DS0s required. To calculate this traffic, consider the sources where the traffic can originate and terminate. The endpoints registered to Unified CM are, of course, the major traffic consumers, but there may be others such as interactive voice response (IVR) applications and other parts of a contact center deployment.

In addition to voice call termination, gateways can also perform a variety of other functions that require resources (either CPU and memory or DSP). These functions include media processing such as media termination point (MTP), transcoding, conference bridge, RSVP Agents, and others.

Gateways, especially those based on the Cisco Integrated Service Routers (ISRs), can provide services beyond just terminating PSTN traffic, 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 other activities that the gateway is performing 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

The discussion on voice traffic analysis earlier in this chapter (Call Traffic, page 29-21) is particularly pertinent to gateways in deciding the number of PSTN circuit DS0s that are required for time-division multiplexing (TDM) voice termination. Because there are likely to be fewer PSTN circuits than the number of system users, a decision has to be made about the optimum number of such circuits, and consequently the DSP requirements, for the gateways. The blocking factor determines the percentage of calls that may not be serviced at peak traffic levels. A smaller blocking factor requires more circuits.

Traffic is measured in Erlangs, and an Erlang is defined as one call lasting for one hour. This section does not go into any further detail on Erlangs other than to say that there are mathematical tables (Erlang-B and Erlang-C) that are used to calculate how many circuits are required for a given amount of offered traffic.

The amount of traffic received and generated by your business determines the size of the external circuits required. However, many customers typically continue to use the same number of circuits for their IP-based communications system as they previously used for a TDM-based system. While this sizing method might work if no issues are encountered, the process of ongoing system traffic analysis should be part of any routine maintenance practices. Traffic analysis can show that the system is over-provisioned for the current levels of traffic (and, therefore, the customer is paying for circuits that are not needed) or, conversely, that the system is under-provisioned and might be suffering from occasional blocked and/or lost calls, in which case increasing the number of circuits will remedy the situation.

Normal Business Traffic Profile

Most customers have a normal traffic profile, which means that they typically have two busy hours per day, one occurring during the morning from 10:00 to 11:00 and the other in the afternoon from 14:00 to 15:00. These busy-hour patterns can often be attributed to such things as employees starting the work day or returning from a lunch break. The calls tend to have longer hold times, and they tend to arrive and leave in a steady manner. A generally accepted industry average holding time to use for traffic calculations is 3 minutes.

Assuming that the communications system is engineered with the busy-hour traffic in mind, no issues should arise. Engineering a system below these levels will result in blocked and/or lost calls, which can have a detrimental effect on business.

Contact Center Traffic Profile

Contact centers present somewhat different patterns of traffic in that these systems typically handle large volumes of calls for the given number of agents or interactive voice response (IVR) systems available to service them. Contact centers want to get the most out of their resources, therefore their agents, trunks, and IVR systems are kept busy all the 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.

Call holding times in contact centers are often of a shorter average duration than normal business calls. Contributing to the shorter average call holding time is the fact that many calls interact only with the IVR system and never need to speak to a human agent (also termed self-service calls). A representative holding time for self-service calls is about 30 seconds, while a call that talks to an agent has 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 centers to optimize resource use (including IVR ports, PSTN trunks, and human agents), combined with the fact that contact centers are systems dedicated to taking telephone calls, also presents the system with higher call arrival rates than in a typical business environment. These call arrival rates can also peak at different times of day and for different reasons (not the usual busy hour) 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 where those calls are received 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.

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, as compared to gateways that receive 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 29-16 describes some of the techniques for decreasing CPU utilization.

Table 29-16 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 29-16 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.

In sizing hardware for the voice messaging system itself (either Cisco Unity or Cisco Unity Connection), the total number of users in the system should be considered. Other items that impact messaging hardware are as follows:

- 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 29-17 shows the applicability of the various voice messaging solutions to the scalability requirements of the deployment.

Table 29-17 Scaling Voice Messaging Solutions

	Maximum Number of Users Supported on a Single Server (Failover or Clustered Deployment)			Maximum Number of Users in a Digital Networking Solution	
Solutions	**500**	**15,000**	**20,000**	**100,000**	**250,000**
Cisco Unity Express	Y	N	N	Y	Y
Cisco Business Edition	Y	N	N	N	N
Cisco Unity Connection (unified/integrated messaging)	Y	Y	Y	Y	N
Cisco Unity (unified and voice messaging)	Y	Y	N	Y	Y

Table 29-18 shows the maximum limits of various functions of different servers running Cisco Unity Connection.

Table 29-18 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 and is 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

Like voice messaging, a conferencing system in a Unified Communications environment should be considered not only for the functions it performs but also for its impact to Unified CM.

There are many deployment options possible using a combination of Cisco Unified MeetingPlace, WebEx as Software-as-a-Service (SaaS), WebEx Node for Aggregation Services Router (ASR), and WebEx Node for Media Convergence Server (MCS), to perform audio, video, and content sharing functions. Sizing consideration for a conferencing system can vary significantly based on the selected deployment option. Typically an enterprise would need to size only its on-premises equipment.

A conferencing system may consist of some or all of the following:

- Meeting directors
- Application servers with Express Media Server (EMS)
- Audio blades (optional)
- Video blades (optional)
- MeetingPlace Web servers

When sizing a conferencing system, you typically will have to consider the following parameters to determine the type and number of application 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
- Whether Cisco WebEx or Unified MeetingPlace will be used to schedule meetings

The following parameters should be considered for determining whether a Hardware Media Server (HMS) is required or if an Express Media Server (EMS) would be sufficient:

- Usage of iLBC or other high-complexity audio codecs. Usage of these codecs would require an HMS.
- Media options such as video continuous presence and echo cancellation. Both of these options need an HMS.

Video usage characteristics such as bandwidth and resolution are an important aspect for the sizing of both kinds of media servers (HMS and EMS).

The capacity of a given Cisco Unified MeetingPlace solution depends on the platform on which the Unified MeetingPlace Meeting Directors, Application servers with EMS or HMS, or WebEx Node for MCS or ASR servers are installed, followed by the capacity of the Unified MeetingPlace Media Servers deployed. 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) with EMS or 2,000 ports (G.711) with HMS in a single system or conferencing node.

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 29-19 shows the number of total SRUs available for enhanced audio and video per supported platform.

Table 29-19 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

Table 29-19 Total System Resource Units per Supported EMS Platform (continued)

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

Table 29-20 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
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 29-19 and Table 29-20, 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

Hardware Media Server

The Cisco Unified MeetingPlace Hardware Media Server (HMS) uses some different settings than the EMS. The Global Audio Mode setting in Unified MeetingPlace Application Administration directly affects the voice capacity of Unified MeetingPlace HMS audio blades. The Global Audio Mode can be configured in either of the following ways:

- G.711 and G.729 without Line Echo Cancellation (LEC)

  With this configuration setting, a single audio blade in the HMS can support a maximum of 250 voice ports. It would require 8 audio blades to reach the maximum supported system limit of 2,000 concurrent audio sessions.

- G.711, G.722, iLBC, or G.729 with Line Echo Cancellation (LEC)

  With this configuration, a single audio blade can support a maximum of 166 voice ports. With 8 audio blades, the maximum supported number of concurrent audio sessions using these additional codecs is 1,328.

The Global Video Mode setting in Unified MeetingPlace Application Administration determines the video capacity of Unified MeetingPlace HMS video blades. The Global Video Mode can be configured in either of the following ways:

- Standard Rate (video call speed up to 384 kbps)

  In this mode, a video blade in the HMS can support a maximum of 40 video ports.

- High Rate (video call speed up to 2,048 kbps)

  In this mode, a video blade can support a maximum of 20 video ports.

For a complete list of the video formats supported by Unified MeetingPlace, refer to the latest version of the *Compatibility Matrix for Cisco Unified MeetingPlace*, available at

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

Unified MeetingPlace Hardware Media Servers can be either the Cisco Unified MeetingPlace 3515 or the Cisco Unified MeetingPlace 3545 chassis. The Unified MeetingPlace 3515 is a fixed platform that comes with one audio blade and one video blade pre-installed. The Unified MeetingPlace 3545 is a modular platform consisting of a chassis that supports four audio blades or video blades in various combinations.

Cascading of Audio and Video Blades

If multiple audio blades and video blades are installed in the Unified MeetingPlace 3545, the media server uses virtual cascading to overflow voice and video streams from one audio or video blade to another. The audio blade has built-in cascading ports that do not decrease the audio session capacity. With a single video blade deployed in the Unified MeetingPlace system, all video ports are available for video conferencing. With multiple video blades deployed, the media server will automatically reserve video ports for cascading purposes. For Standard Rate video, 8 video ports are reserved for cascading, leaving 40 video ports available. For High Rate video, 4 video ports are reserved for cascading, leaving 20 video ports available.

Example 29-1 Unified MeetingPlace Audio Conference

A Unified MeetingPlace 3545 Media Server is deployed with two audio blades and two video blades. A meeting is scheduled with 350 audio ports and the Global Audio Mode is configured for G.711 with LEC. In this case:

- The media server allocates 251 ports from the first audio blade, out of which 250 ports are used for audio participants and one port is used for voice cascading or connecting to the second audio blade.
- The media server allocates 101 ports from the second audio blade, out of which 100 ports are used for audio participants and one port is used for voice cascading.

Example 29-2 Unified MeetingPlace Video Conference

A Unified MeetingPlace 3545 Media Server is deployed with two audio blades and two video blades. For this example, assume a meeting is scheduled with 65 video ports and the Global Video Mode is configured for Standard Rate video. In this case:

- The media server allocates 41 ports from the first video blade, out of which 40 ports are used for video participants and one port is used for video cascading or connecting to the second video blade.
- The media server allocates 26 ports from the second video blade, out of which 25 ports are used for video participants and one port is used for video cascading.

Sizing Guidelines for Unified MeetingPlace Audio Conferencing

Cisco recommends the following methods for calculating Unified MeetingPlace audio conferencing capacity:

- Calculation based on average monthly usage

  If you know the average voice conferencing usage (average minutes per month), use Table 29-21 to calculate the Unified MeetingPlace audio conferencing capacity.

Table 29-21 Unified MeetingPlace Audio Conferencing Capacity Based on Average Monthly Usage

Average Monthly Usage (minutes)	Baseline Usage (minutes per user license 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 or reservationless model on Cisco Unified MeetingPlace, 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.

When planning for Unified MeetingPlace capacity expansion, also consider whether the following conditions apply to your system:

- The total estimated port capacity exceeds 80% of the maximum supported ports as listed in 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

- An audio codec other than G.711 is desired. However, transcoders based on Cisco Integrated Services Routers (ISR) can be used if needed to achieve maximum capacity for other codec types in the meetings.
- Line Echo Cancellation (LEC) is provided by an external device such an ISR, rather than Unified MeetingPlace providing echo cancellation functionality.

Sizing Guidelines for Unified MeetingPlace Video Conferencing

Cisco recommends the following three methods for calculating Unified MeetingPlace 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 Unified MeetingPlace system.

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

  Cisco recommends deploying a direct replacement for an existing video conferencing system. A video conferencing license on the existing system can be replaced by a Cisco Unified MeetingPlace user license.

Unified MeetingPlace Web Server

The Unified MeetingPlace Web Server is required only for Unified MeetingPlace scheduling deployments to schedule and attend meetings from a Web user interface, for Lotus Notes integrations, or for processing 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.

WebEx Node for MCS

Web conferencing optionally using Cisco WebEx Node for MCS can accommodate up to 500 web sessions, depending on the type of hardware on which the WebEx Node for MCS resides. A maximum of four WebEx Nodes for MCS can be deployed per solution, or an unlimited number of nodes can be deployed with WebEx Node for ASR, allowing for scalability up to 2,000 web sessions on-premises with redundancy. WebEx Nodes can be distributed anywhere in a customer network, but Cisco recommends deploying them closest to the larger groups of web users. Only internal users will have web sessions with

the WebEx Node for MCS or ASR; external users will always connect to the cloud. For detailed capacities of the Cisco WebEx Node for MCS or ASR, 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

Impact on Unified CM

The impact to Unified CM can be analyzed by 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 that 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 Unified Presence

As with all other applications, sizing for Cisco Unified Presence is accomplished in the following manner:

- Decomposing the system into its most elemental services
- Measuring the unit cost of each of these services
- Analyzing the given system description as an aggregation of the identified services and arriving at a net system cost
- Determining the number of required servers based on system cost and deployment options

For Unified 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 (intra-cluster, inter-cluster, 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 29-22.

Table 29-22 Maximum Number of Users Supported per Unified Presence Server

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

The formal definitions of the OVA templates for Cisco Unified 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 Unified Presence server 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 affect of CTI control of phones, however, must be counted against CTI limits.

Cisco Unified Communications Management Suite

The Cisco Unified Communications Management Suite consists of four applications. 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 Unified Communications Management Suite 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 server, for example, can host from one to all four of the Unified Communications Management Suite applications, and can be used to manage up to 10,000 endpoints. The specification for such a co-resident server is:

- Processor: 3 GHz, 8 Core
- Memory: 16 GB
- Disk space: 320 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 on Cisco 1040 Sensor and Network Analysis Modules (NAMs) to measure voice quality.

Table 29-23 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 NAM modules)
- 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

Conclusion

In summary, it can be challenging to determine the hardware composition of a large Unified Communications system consisting of a number of separate applications working together. However, an understanding of the functional requirements of the software and the performance capabilities of the hardware platforms certified for running the software, can be very helpful in making an accurate estimation of the servers required. Tools developed by Cisco for this purpose are available as described in this chapter. For further assistance, contact your Cisco Partner or Cisco Systems Engineer, and they can use the Cisco Unified Communications Sizing Tool (http://tools.cisco.com/cucst) to validate all designs.