Cisco Collaboration Sizing Guide for Release 15

A Cisco Preferred Architecture (PA) design reference guide.

Published: December 2023

Cisco Systems, Inc.
www.cisco.com

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco website at www.cisco.com/go/offices.

Cisco Collaboration Sizing Guide

Collaboration Sizing Guide for Release 15

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

Accurate sizing is critical to ensure that the deployed system will meet the expected service quality for the required number of users and traffic volume. This ensures the applications deployed will be stable, performant, and fully functional at advertised capacities. However, many sizing factors must be considered in a complex system deployment. For example, multiple products may be distributed across different locations, including video endpoints, call centers, and voice/video conferencing. Cisco Systems provides a set of sizing rules to handle the resulting complexity.

This document briefly introduces system sizing methodology and the factors that affect sizing. It also provides information about how to use the sizing tools.

Note This document should be read in conjunction with the product information and design and deployment guidance covered in other documents, including product documentation (available at cisco.com) and the Preferred Architecture documents. A good understanding of these aspects is required for a successful deployment.

This document includes the following major sections:

Note For simplified sizing guidance without using the Collaboration Sizing Tool, refer to the latest version of the *Cisco Preferred Architecture for Enterprise Collaboration CVD*.

Methodology for System Sizing

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

The following sections describe the sizing methodology:

Performance Testing

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

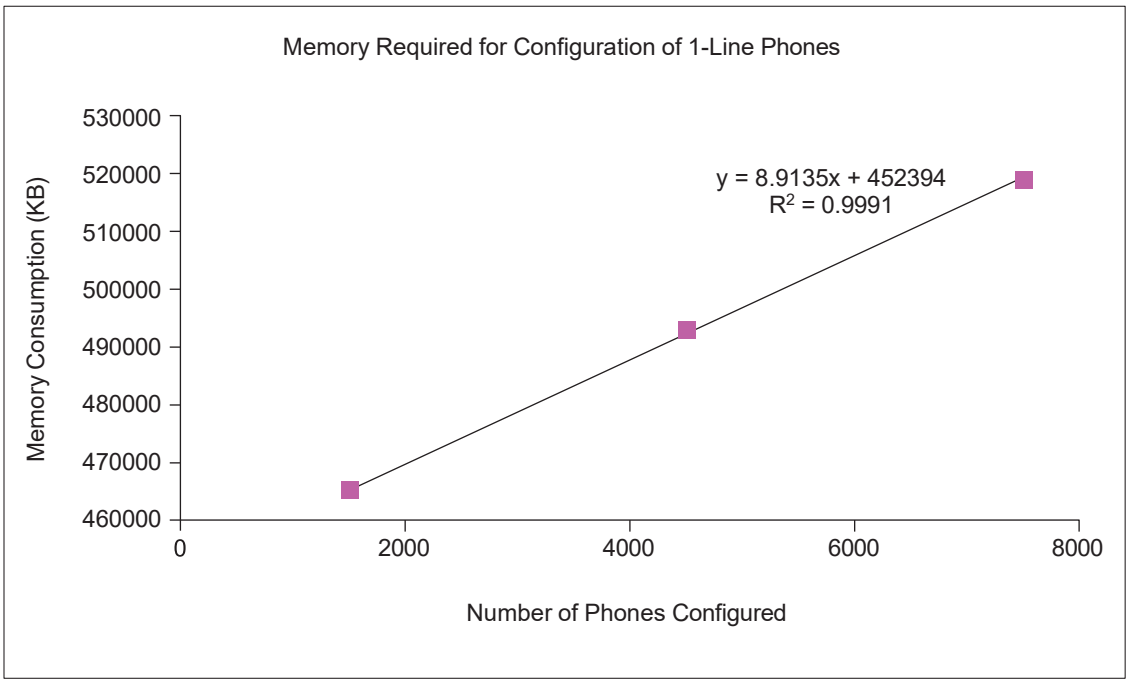
Most systems exhibit linearity within a certain range, beyond which the system performance can become unpredictable. Cisco sets the usage levels for each performance test to identify and confirm the linear range of the resource usage for each function. The results for each test can be graphed using a minimal number of data points. Additional data points (at intermediate load levels) are obtained to define the actual system behavior if required.

The slope of the linear section of the graph defines the resource usage and/or cost for each incremental addition of work. The R^2 value is used to estimate the closeness of the fit. If the R^2 value is close to 1, the formula is a close match for the data.

For example, Figure 1-1 shows the results of a test conducted to determine the memory requirements for configuring single-line IP phones. It shows the memory allocated when configuring 1,500, 4,500, and 7,500 single-line IP phones in Unified CM. The graph shows that the trend line equation is linear and can be used to predict the dependent variable (in this case, memory allocated) based on the control variable (the number of phones).

In this particular test, the R^2 value is extremely close to 1. From the equation, we can compute that the memory consumed with the configuration of 7,500 one-line phones is approximately 519,000 Kbytes. Each additional configured single-line endpoint in the system consumes an additional 8.91 Kbytes.

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



System Modeling

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

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

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

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

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

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

Memory Usage Analysis

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

Static memory

Static memory is consumed even when there is no traffic on the system. Static memory usage includes the data for system configuration and the data for registered endpoints. Static memory also includes configuration for the dial plan (which covers items such as partitions, translation patterns, route lists and groups). In addition, static memory includes the memory allocated for CTI and other applications. In a large system, static memory is mainly a function of the number of configured endpoints and the size of the dial plan.

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

Dynamic memory

Dynamic memory is used for transient activities, such as allocating memory for each active call (call in progress). In a large system, dynamic memory is primarily a function of the number of concurrent calls.

The number of concurrent calls is proportional to the average call holding time (ACHT). Longer ACHT results in more dynamic memory use because there will be a larger number of concurrent or active calls.

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

System memory

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

CPU Usage Analysis

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

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

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

Applications like Unified CM are often busy doing additional tasks (for example, trace, CDR generation and writing, etc.), so even when idle, there is a minimum steady state load that must be accounted for. Further, during runtime, there are scenarios where resource usage may burst above the typical steady state load at certain times, for example, top and bottom of busy hour load, system startup, system recovery after power outage, system upgrade and backups, CDR writes, etc.

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

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

Finally, CPU vendor & industry performance benchmarks for physical CPU rarely if ever correlate directly to collaboration applications' resource usage and capacity/performance. When conducting capacity planning, avoid making comparisons between CPU speed and older and newer versions of Cisco Collaboration software and do not expect linear upscaling or downscaling based on increased or reduced user and device capacity.

Traffic Engineering

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

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

It's worth noting that collaboration applications and particularly voice and video traffic load do not tend to match loads observed on other business applications like email, file/print services, and web browser sessions. Collaboration applications tend to have client-to-client traffic patterns versus the typical client-to-server or server-to-server patterns found with many other business applications and services.

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

- Definitions, page 6
- Voice Traffic, page 9
- Contact Center Traffic, page 9
- Video Traffic, page 9
- Conferencing and Collaboration Traffic, page 10

Definitions

Traffic engineering defines the following terms:

Simultaneous Calls

The number of simultaneous calls is the average number of calls active at a given time.

Calls per Second

The number of new call attempts that arrive at the system in one second, plus the number of existing calls torn down during that same one second interval. This unit can be used to define the average calls per second that the system expects to handle during a busy hour. (This number is equivalent to the busy hour calls divided by 3600.)

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

Busy Hour

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

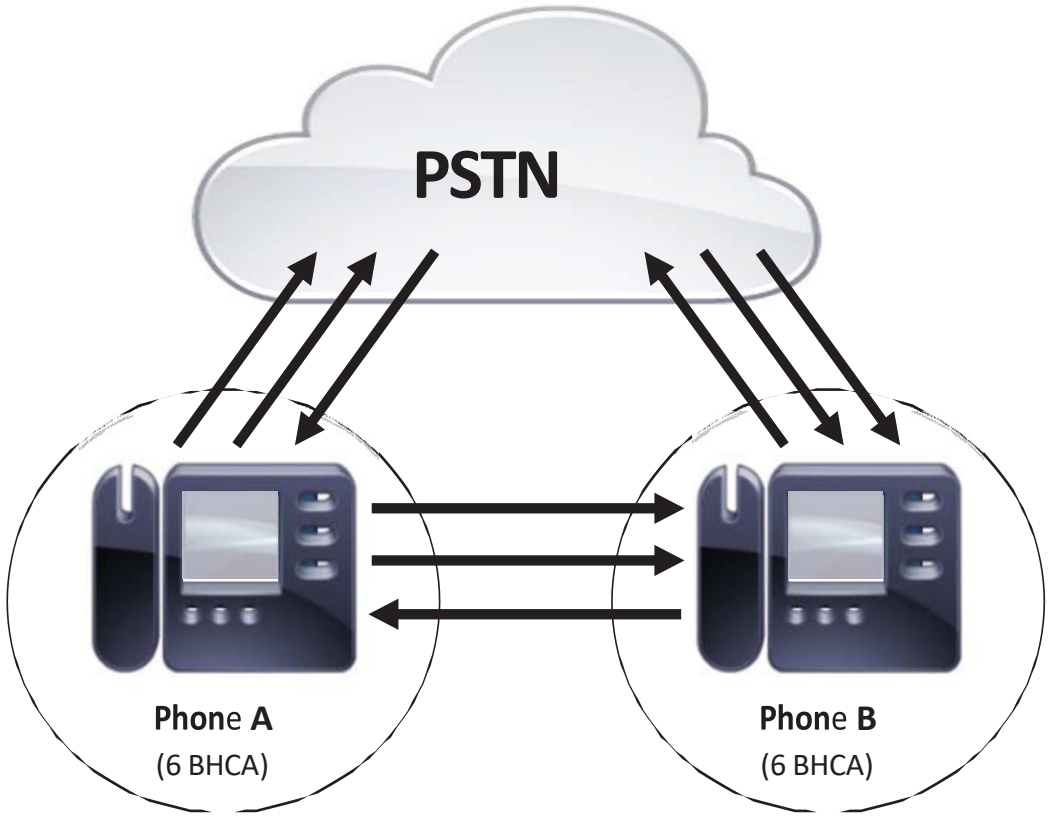
Busy Hour Call Attempts (BHCA)

The user BHCA represents the average number of calls that a user initiates or receives during the busy hour. BHCA is defined as the number of attempted calls at the sliding 60-minute period during which the maximum load occurs in a given 24-hour period (BHCA). Higher BHCA results in higher processor

utilization. System BHCA is always less than or equal to the product of the user BHCA and the number of users.

Figure 1-2 shows an example of a mix of PSTN and intracluster calls made by two users at a 6 BHCA rate. With phone A and phone B users making and receiving calls from the PSTN and to and from each other at 6 BHCA, there is a total of 12 user BHCA, which must be accounted for when sizing the system. As noted, this example translates to 9 system BHCA since the three IP-to-IP calls (between phone and phone B) are seen as 3 BHCA from a system perspective, even though the user BHCA for these IP-to-IP intracluster calls is 3 per user for a total of 6 BHCA.

Figure 1-2 Example: 6 BHCA per User with Mix of PSTN and Intracluster Calls



Average Call Hold Time

This is the average period of time that the user is busy. For example, on a voice call the ACHT is the period of time between call setup and call tear-down when there is an open speech path between the two parties. A hold time of 3 minutes (180 seconds) and the BHCA rate of 4 is an industry average used for traffic engineering of voice systems.

Erlang

The Erlang is a measure of traffic load on a system. To calculate Erlangs, multiply calls per hour by the average holding time in minutes and divide by 60. Resource requirements can be derived from the system load requirements by using the appropriate Erlang model.

The number of Erlangs is also equal to the number of simultaneous calls. The number of simultaneous calls is calculated as a product of the average calls per second (cps) rate and the average call holding time in seconds.

Erlang B Model

The Erlang B model can determine the number of trunks required to handle a traffic load (in Erlangs) with a specified blocking factor - probability of not having an available circuit/trunk for a call. The Extended Erlang B model includes the modeling of retries (for calls that are blocked). The retry percentage is an additional input to the Extended Erlang B model More details about the Erlang model and traditional Traffic Analysis can be found at

Table 1-2 illustrates the relationship between number of trunks, blocking probability, and Erlangs of traffic.

Table 1-2 Erlang B Traffic Table (Number of Circuits Required)

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

From Table 1-2 we can determine the following information:

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

Erlang C Model

The Erlang C model incorporates queuing of incoming calls and is therefore used when modeling call center traffic.

Bursty Traffic

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

Voice Traffic

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

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

Contact Center Traffic

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

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

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

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

As noted earlier, contact center sizing uses the Erlang C model to account for calls waiting in queues. Contact centers require additional resources like interactive voice response (IVR) ports. The time that calls wait in queues needs to be factored in when sizing the PSTN gateways.

Note For additional information about Cisco Unified Contact Center deployments, refer to the latest version of the *Solution Design Guide for Cisco Unified Contact Center Enterprise*.

Video Traffic

Point-to-point video traffic (On-Net/Non-PSTN) is similar to its voice equivalents for call arrival rates, peak usage times, and call durations. Also, signaling for call setup and take-down is similar to voice calls.

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

Implications for video conferencing are covered in the next section.

Conferencing and Collaboration Traffic

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

- Call arrivals

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

- Peaks

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

- Call durations

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

- Video conferencing

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

Sizing a deployment for conferencing is primarily a function of the required dial-in interval and the estimated number of meeting participants. For example, sizing of Cisco Meeting Servers (CMS) would include the following considerations:

- Geographical location — Each region served by Unified CM could have dedicated conferencing resources and different bandwidth considerations.
- Preference for the type of Cisco Meeting Server platforms — Hardware or software
- Cisco Meeting Server platform capacities
- Type of conferencing — Audio and/or video; scheduled and/or non-scheduled.
- Conference video resolution — Higher video quality conferences use more resources.
- Large conference requirements — For example, all-hands meetings

Conference resources are generally dedicated to a region to keep as much of the conference media on the regional network; therefore, sizing can be considered on a region-by-region basis.

System Sizing Considerations

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

- Network Design Factors, page 11
- Other Sizing Factors, page 12

Network Design Factors

Solution sizing is affected by the following network design factors:

- Cluster sizes

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

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

- Interaction between individual products

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

- Physical hardware capabilities

  Each type of server or router supports different capabilities. For example, more powerful servers might have a higher number of network ports compared to Cisco Business Edition 6000 platforms or a Cisco Integrated Services Router (ISR).

  As another example, different models of Cisco Integrated Services Routers have restrictions on the number and types of network modules or Cisco Unified Computing System (UCS) E-Series blade servers they can host.

  Many applications are storage-IO-intensive but not storage-GB-intensive. In order to support higher user/traffic capacity, faster storage than normally used for other business applications may be required. Likewise, to support higher user/traffic capacity, faster CPU base frequencies may also be required. Static fixed speed CPUs are generally preferred over variable speed CPUs.

  In some cases, maximum capacity may be limited by the physical RAM available on the VM platform and additional VM hosts may be required. Keep in mind that hypervisor versions require their own RAM and this additional RAM must be accounted for.

- Optional capabilities and features

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

Other Sizing Factors

The following additional factors also affect system sizing:

- Mix of call types:

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

- Mix of endpoint types

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

- System release

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

- Use of external applications

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

- Anticipated system growth

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

- Average and peak usage

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

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

Sizing Tools Overview

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

https://cucst.cloudapps.cisco.com/

Cisco recommends that you use the sizing tools to perform system sizing. These tools consider data from performance testing, individual product limits and performance ratings, advanced and new product-released features, design recommendations from this document, and other factors. Based on input provided by the system designer, the tools apply their sizing algorithms to the supplied data to recommend a set of server requirements.

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

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

Cisco provides the following sizing tools:

- Cisco Collaboration Sizing Tool

  This tool guides the user through a complete system deployment. It is Enterprise grade — provides sizing of large and complex deals including Megacluster. This tool covers the following products and components:

  - Cisco Unified Communications Manager (Unified CM)
  - IM and Presence services
  - Voice messaging
  - Conferencing/Web Conferencing
  - Cisco Emergency Responder
  - Media Streaming and Recording
  - Collaboration Management
  - Gateways
  - Cisco Unified Communications Management Suite
  - Cisco Unified Contact Center components

  Sizing tool output provides application sizing and bill of materials for included components.

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

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

- Cisco Expressway Sizing Tool

  This is a specialized tool that focuses on the specific functions of Expressway mobile and remote access (MRA) and business to business (B2B) deployments.

- Quote Collab

  This tool focuses on sizing, configuration, and quoting of on-premises and hybrid deployments with between 500 and 10,000 users or endpoints. It assists users with application and hardware sizing as well as VM placement for enterprise solutions. Recommendations are BE7000-based. Quote Collab delivers a solution diagram, server diagram, bill of materials, high-level quote, and an editable Powerpoint summary. In addition, the tool enables export to Cisco Commerce Workspace for validation and order.

  Note The former Virtual Machine Placement Tool (VMPT) has been decommissioned and replaced by the Quote Collab tool (with “Servers Only” option).

  Quote Collab is available at

  https://cqc.cloudapps.cisco.com/

For more information on these tools and their access privileges, refer to the *Collaboration Sizing Tool Frequently Asked Questions*.

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

Using the SME Sizing Tool

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

An SME cluster follows the same topology as a regular Unified CM cluster. A publisher node provides the master configuration repository. The TFTP service can run on the publisher node if the number of phones or MGCP gateways in the cluster is relatively small. A redundancy ratio of 1:1 is recommended for call processing subscribers.

To size an SME cluster, you must consider the functionality that it is expected to perform. In a base configuration, the SME acts as a routing aggregation point for a number of leaf clusters. It also provides centralized PSTN access for all of the leaf clusters connected to it. In more advanced configurations, the SME may also host centralized voice messaging, mobility, and conferencing services. The performance of the SME is influenced by the type of trunk protocols that the leaf clusters use to connect to it and by the BHCA across those trunks.

The SME sizing tool requires the following input parameters:

- The SIP trunks that the cluster services.
- The number of users that access SME cluster services through the trunks
- BHCA per user for each trunk to leaf clusters for intercluster calls.
- BHCA per user for each trunk to leaf clusters for off-net (PSTN) calls.
- The trunks used by the SME cluster to connect to the PSTN.
- Average holding time for calls
- Number of route and translation patterns

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

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

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

Using the Expressway Sizing Tool

The Expressway Sizing Tool is used for sizing the Expressway platform with focus on mobile and remote access (MRA) and business to business (B2B) for Unified CM deployments.

To size an Expressway cluster, you must consider the functionality that it is providing and the utilization of that functionality. The performance of the Expressway is influenced by the functionality enabled (MRA, B2B, or both) as well as the connectivity of devices through the platform and the BHCA by the device users.

The Expressway sizing tool requires the following input parameters:

- Number of system users with and without MRA-enabled endpoints.
- Percentage of MRA users working on-premises.
- Overall average busy hour call attempts (BHCA) and average call hold time (AHT) per user.
- Percentage of user making B2B calls.
- Percentage of users using video in calls.
- Percentage of user in meetings.
- Meeting solution deployed (Webex v. on-premises).
- Percentage of users in meetings using Webex Edge Audio.

If Expressway is deployed with MRA functionality, you must consider the following additional endpoint sizing parameters:

- Total number of hardware endpoints deployed over MRA.
- Total number of Jabber clients deployed over MRA.
- Total number of Webex clients with Unified CM calling deployed over MRA.

The performance of the Expressway is measured in the number of MRA registered devices and clients and the number of active calls. There are variations across the hardware platforms and software versions.

Using the Cisco Collaboration Sizing Tool

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

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

- Cisco Unified Communications Manager, page 15
- Media Resources, page 29
- Cisco Unified CM Megacluster Deployment, page 33
- Cisco IM and Presence, page 33
- Cisco Expressway, page 37
- Emergency Services, page 36
- Gateways, page 38
- Voice Messaging, page 39
- Collaborative Conferencing, page 40
- Cisco Prime Collaboration Management Tools, page 43
- Cisco Unified Communications Manager Express, page 44
- Cisco Business Edition, page 45

Cisco Unified Communications Manager

Cisco Unified Communications Manager (Unified CM) is the hub of any Unified Communications deployment. It performs key functions such as controlling endpoints, routing calls, enforcing policies,

and hosting applications. Unified CM provides coordination for the other Unified Communications products such as PSTN gateways, Cisco Unity Connection, Cisco Unified Communications Manager IM and Presence Service, and Cisco Unified Contact Center. The coordination function has an impact on Unified CM performance, and therefore must be accounted for in Unified CM sizing.

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

Virtual Nodes and Cluster Maximums

The sizing tool applies the following server node and cluster maximums. These values can vary depending on Unified CM software version:

- Each cluster can support configuration and registration for a maximum of 50,000 secured or unsecured SCCP or SIP phones.
- Two TFTP server nodes are required, in addition to a dedicated publisher, 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 50,000 CTI connections.
- The number of call processing subscribers in a cluster cannot exceed 4, plus 4 standby, for a total of 8 call processing subscriber nodes. Also, the total number of server nodes in a cluster, including the publisher, TFTP, and media servers, may not exceed 21 servers as the maximum allowed in a cluster.
- There are three recommended Unified CM virtual machine (VM) configurations or Open Virtualization Archives (OVAs): Small, Medium, and Large. Based on the amount of virtual machine resources (vCPU, RAM, and so on) each OVA size supports a maximum number of users or devices assuming that on average, each user has one phone. If this is not the case, then the limit per OVA is the maximum number of endpoints registered to a Unified CM node. For example, the Large platform OVA supports a maximum of 12,500 users, assuming one device per user. However, if you plan to deploy multiple devices per user, then the maximum number of supported users is reduced. For example, if you have 2 devices per user, then the large OVA would support a maximum of 6,250 users with 12,500 devices total. This same principal applies for the Medium and Small Unified CM platform OVAs as well.

Deployment Options

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

Database Complexity

The CPU usage is considerably higher when the configuration database in Unified CM is considered to be complex. There is no one metric to determine whether the database is simple or complex. As a general rule, the database is considered complex if you have configured more than 10,000 endpoints and more than a few hundred dial plan elements such as translation and route patterns, hunt pilots, and shared lines.

Number of Regions and Locations

Configuration of regions and locations in the Unified CM cluster requires both database and static memory. The number of gateways that can be defined in the cluster is also tied to the number of locations that can be defined. Table 1-3 lists these limits for some Unified CM VM configurations.

Table 1-3 Maximum Number of Regions, Locations, Gateways, and Trunks

VM Configuration	Maximum Number of Regions	Maximum Number of Locations	Maximum Number of Trunks and Gateways
Small	1,000	1,000	1,100
Medium or Large	4,000	4,000	4,100

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

Call Detail and Call Management Records

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

High Availability

After you determine the minimum number of nodes required for the specified deployment, add the desired number of additional subscriber nodes to provide redundancy.

Number of Virtual Server Nodes per Cluster

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

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

Choice of VM Configurations and Hardware Platforms

Cisco provides OVA VM configurations that can be loaded onto a hypervisor. Different templates specify different capacities. For example, the Large platform OVA defines a virtual machine that has a maximum capacity of 12,500 endpoints. There are also templates defined to support a maximum of 3,750 (Small OVA), and 10,000 endpoints (Medium OVA).

The formal definitions of the VM configurations for Unified CM and other Unified Communications products are available at the following location.

Specific information for Unified CM is available at the following location.

With Unified CM, some of the VM configurations are not supported on the low-end hardware platforms. To verify which VM configuration is supported on a hardware platform, refer to the documentation at the following location or use the QuoteCollab tool.

Endpoints

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

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

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

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

Table 1-4 shows the maximum number of endpoints supported by different VM configuration types. Note that these values are guidelines only. A given system may support less than these maximum amounts because of other applications included in the deployment.

Table 1-4 Maximum Number of Endpoints Supported Per VM Configuration

VM Configuration	Maximum Endpoints per OVA Template[1]
Large	12,500
Medium	10,000
Small[2]	3,750

1. These limits represent the maximum number of endpoints that can be configured in the database and registered per virtual subscriber node. All other registered devices such as media termination points (hardware of software) or SIP trunks do not count against these limits.
2. Capacity for the small OVA is dependent on hardware. Deploying the small OVA on the BE6000 platform reduces the maximum node capacity to 1,200 (BE6000M).

For more details including hardware requirements, please refer to the documentation at the following location.

Cisco Collaboration Software Clients and Mobility

Cisco collaboration clients are software applications that run on user desktops or other access devices. Mobility is a set of features which address users that are mobile and not always within the organization's network boundaries.

- Cisco Collaboration Software Clients, page 19
- Mobility, page 22

Cisco Collaboration Software Clients

When designing and sizing a solution for Cisco collaboration software clients, you must consider the following scalability impacts for all the components:

- Client scalability

  The deployed platform OVAs determines the number of devices a cluster can support. The Cisco software client deployment must balance client registration and other communications equally across all nodes in the cluster.

- IMAP scalability

  The number of IMAP or IMAP-Idle connections is determined by the voice messaging integration platform.

- Audio, video, and web conferencing

  Clients can access the conferencing services that are provided in your network. You need to account for these users when sizing the number of concurrent participants for these services.

Cisco Jabber and Cisco Webex client applications are supported on mobile devices (iPhone, iPad, and Android) as dual mode or tablet devices and on desktops (Windows and Mac) as client services framework (CSF) devices. When sizing your deployment with software clients, keep in mind that users may have any combination of desktop and mobile clients.

Note For the purposes of this discussion, references to Cisco software clients include Cisco Jabber as well as Cisco Webex when deployed in Calling in Webex (Unified CM) mode. Unless otherwise noted, functionality described applies to both software client (Jabber and Webex).

Desktop clients provide two modes of operation, each of which uses different resources in Unified CM. When it operates in softphone mode, the Cisco software client acts as a SIP registered endpoint and contributes to the total number of endpoints in the system. When it operates in deskphone control mode, the software client acts as a CTI agent and therefore uses CTI resources on Unified CM.

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

Note If a user has only a desktop client in deskphone control mode, then that will count as only a single device due to the fact that the deskphone control utilizes CTI resources and lines.

The software client's interface with Unified CM. Therefore, the following guidelines for the current functionality of Unified CM apply when software client voice or video calls are initiated:

- CTI scalability

  In deskphone control mode, calls from the software client use the CTI interface on Unified CM. Therefore, observe the CTI limits as defined in the section on Applications and CTI, page 24. You must include these CTI devices when sizing Unified CM clusters.

- Call admission control

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

- Codec selection

  Cisco software client voice and video calls utilize codec selection through the Unified CM regions configurations.

- Cisco Unified CM User Data Service (UDS)

  UDS is an umbrella of service APIs provided by Unified CM. UDS provides a contact source API that can be used by Jabber over Cisco edge services for contact source lookups. Using the UDS contact source to resolve contacts puts additional load on the system. In the case of Webex, the contact source is always the Webex cloud.

The following additional items must be considered for desktop client deployments:

- Device Configuration

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

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

- Deskphone Control Mode

  When configured in deskphone control mode, the client establishes a CTI connection to Unified CM upon login and registration to allow for control of the IP phone. Unified CM supports up to 50,000 CTI connections. If you have a large number of clients operating in deskphone control mode, make sure that you evenly distribute those CTI connections across all Unified CM subscribers running the CTIManager service. This can be achieved by creating multiple CTI Gateway profiles, each with a different pair of CTIManager addresses, and distributing the CTI Gateway profile assignments across all clients using deskphone mode.

- Voicemail

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

- Authentication

  Client login and authentication is handled based on the configured method within the deployment. Options include LDAP-based authentication in the case of Jabber, HTTPS web-based authentication in the case of Webex, and Single Sign-On (SSO) which is supported with both clients. Login credentials may be stored in the local client cache or in the case of OAuth-based authorization, tokens are stored to enable renewed secure connectivity.

- Contact Search

  There are several contact sources that can be used with the desktop clients. For example, the UDS service can be used by Jabber clients to search for contacts in the Unified CM User database. Alternatively, LDAP integration can be used by Jabber. In the case of Webex, contact search is done against the identity store for the Webex organization. In all cases, if the requested contact cannot be found in the local desktop client cache, contact searches take place against the appropriate directory source (LDAP, UDS, or Webex).

SAML SSO Cisco Jabber Client

Cisco Unified CM provides the Security Assertion Markup Language Single Sign-On (SAML SSO) feature, which enhances the end user experience by allowing users to log in only once to access all applications within the Cisco Collaboration solution.

SAML SSO provides secure mechanism to use credentials and relevant information of the end user to be leveraged across multiple Unified Communications applications (such as Unified CM, Unity Connection, and Unified CM IM and Presence). For the SAML Single Sign-On feature to work as expected, the network architecture must scale to support the number of users for each cluster.

For a Unified Communications deployment across multiple applications (such as Unified CM, Unity Connection, and Unified CM IM and Presence), all SAML requests must authenticate with the Identity Provider (IdP) for Cisco software clients to login successfully.

Note SSO is supported by Unified Communications services with SAML.

Software clients with SAML SSO logins should also be factored into system sizing because the numbers of users logging into the system in a typical day at the same time could have an impact on the time it takes for user(s) to log in. This is expected due to the limiting factor of how many requests the system can process at one time. The current maximum login rate for software client users is 2.7 logins per second (about 166 logins per minute) or 10,000 logins within one hour. This is assuming that all users and devices are evenly distributed across all nodes and that the software client is in softphone mode.

There are many interdependent variables that can affect Unified CM cluster scalability (such as regions, locations, gateways, media resources, and so forth). Therefore, it is vital to determine the number of users, endpoints, and calls per user per hour, to deploy efficiently so that resources are available to handle the required load.

As an example, consider a deployment with redundant subscriber pairs supporting 5,000 users, each associated with two devices (desk phone and softphone). This deployment would require the following number of virtual machines and VM configurations (assuming high availability and redundancy):

- One pair of Unified CM subscribers with Large OVA or VM configurations
- One pair of Unified CM IM and Presence 5k-user VM configurations

The Unified CM IM and Presence 5k-user VM configuration pair would support the 5,000 users, and a pair of Unified CM Large VM configurations would support the 10,000 devices.

Mobility

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

Cisco Unified Mobility

There are two parameters that are key to Unified CM's capacity to support single number reach and enterprise two-stage dialing (Mobile Voice Access and Enterprise Feature Access). For these functions to work appropriately, users must be enabled for mobility and remote destinations with shared lines must be defined for the users. Table 1-5 shows the limits for users and remote destinations and mobility identities in a cluster consisting of each class of Unified CM VM configurations.

Table 1-5 Maximum Number of Mobility Users and Remote Destinations and Mobility Identities per Cluster

Cluster Nodes	Maximum Number of Users Enabled for Mobility per Cluster	Maximum Number of Remote Destinations and Mobility Identities per Cluster
Large VM configuration	50,000	50,000 (or 12,500 per node)
Medium VM configuration	40,000	40,000 (or 10,000 per node)
Small VM configuration	15,000	15,000 (or 3,750 per node)

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

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

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

Call Traffic

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

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

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

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

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

Call traffic can arise from other sources as well. Each time a call is redirected in a transfer or to voicemail, it requires processing by the CPU. If a directory number is configured on multiple phones, an incoming call to that number needs to be presented to all of those phones, thus increasing CPU usage at call setup time. If advanced features are being used, calls made using this technology, and the percentage of these calls that need to be redirected to the PSTN because of call quality, must also be accounted for.

Dial Plan

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

- Directory numbers (DNs)
- Shared directory numbers and the average number of endpoints that share the same DN.
- Partitions, calling search spaces, translations, and transformation patterns.
- Route patterns, route lists, and route groups
- Global Dial Plan Replication (GDPR)
- Hunt pilots and hunt lists
- Circular, sequential, and broadcast line groups and their membership

Note Directory numbers (DNs) are part of the Unified CM in-memory database.

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

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

Applications and CTI

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

Although the large Unified CM VM configurations are able to support CTI for all of their registered devices, the smaller VM configurations do not scale that high. Table 1-6 lists the maximum number of CTI resources supported for each Unified CM VM configuration. 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 maximum number of CTI resources for a VM configuration corresponds to the endpoint capacity of that VM configuration.

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 1-6 CTI Resource Limits in Unified CM

VM Configuration	Maximum CTI Resources per Virtual Machine
Small	3,750
Medium	10,000
Large	12,500

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

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

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

Determining CTI Resources Required for a Unified CM Cluster

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

Step 1 Determine the total CTI device count.

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

Step 2 Determine the CTI line factor.

Determine the CTI line factor of all devices in the cluster, according to Table 1-7.

Table 1-7 CTI Line Factor

Number of Lines per CTI Device	CTI Line Factor
1 to 5 lines	1.0
6 lines	1.2
7 lines	1.4
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 1-8.

Table 1-8 CTI Application Factor

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

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

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

The following examples illustrate the process.

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

(500 CTI Devices) * (Greater of {1.8 Line Factor or 1.0 Application Factor}) (500 CTI Devices) * (1.8 Line Factor) = 900 total CTI resources required

Example 2: 2,000 CTI devices deployed with an average of 5 lines per device and an average of 9 applications per device. According to the factor lists in Table 1-7 and Table 1-8, the 5 lines per device renders a line factor of 1.0, while 9 applications per device renders an application factor of 1.8. Applying these values in the formula from Step 4 yields:

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

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

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

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

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

IP Phone Services

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

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

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

Cisco Extension Mobility and Extension Mobility Cross Cluster

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

- Creation of EM profiles requires both disk database space and static memory.
- The rate at which users may log into their EM accounts affects both CPU and memory usage. Unified CM nodes 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 Unified CM node 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 node multiplied by the number of nodes 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 node.

Table 1-9 shows the maximum number of EM and EMCC logins per minute for each type of VM configuration.

Table 1-9 EM and EMCC Rates Per VM Configuration

VM Configuration	Maximum EM Login Rate (per Node)	Maximum EM Login Rate (per Cluster)	Maximum EMCC Login Rate (Per Node)	Maximum EMCC Login Rate (per Cluster)	Maximum Concurrent EMCC Devices
Small	235	352	71	80	833
Medium or Large	500	750	75	225	7,000

Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. For example, when the EM load is distributed evenly between two virtual machines with the Medium or Large VM configuration, the maximum cluster-wide capacity is 750 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.

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 Medium or Large virtual machines.
- A maximum of three pairs of primary and backup Unified CM Assistant nodes 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 server nodes 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.

Cisco WebDialer

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

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

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

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

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

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

Example: Calculating WebDialer Calls per Second

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

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

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

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

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

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

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

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

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

Attendant Console

The integration of Cisco Unified CM with the Attendant Console utilizes CTI resources. The server-based attendant console monitors 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

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

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

- Co-resident deployment

  In a co-resident deployment, the streaming application runs on any server node (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 node or virtual machine.

- Standalone deployment

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

The Cisco IP Voice Media Streaming Application can provide MTP, annunciation, and conferencing capabilities, but a more scalable design is to place these functions on external Cisco Integrated Service Routers (ISRs). The music-on-hold functionality of this application is, however, not so easily placed on external sources. Table 1-10 lists the maximum values that may be configured for each of these services.

Table 1-10 Cisco IP Voice Media Streaming Application Capacity Limits

Media Device Type	Default Quantity	Maximum Number of Streams or Devices	Supported Codecs
Annunciator	48	750	G.711, G.729, L16WB
Software Conference Bridge	48	256	G.711, L16WB
Music on Hold	250	1,000	G.711, G.729, L16WB
Software Media Termination Point (MTP)	48	512	G.711, L16WB, passthrough

The following notes apply to Table 1-10:

- All values represent the number of callers supported per media device. For instance, 48 software conference bridges can support 16 three-party conferences.
- These devices can be co-resident with the call processing nodes when using default settings or near to default settings.
- When increasing capacities to the maximum values, Cisco recommends deploying the media devices on standalone nodes (not with call processing).
- If MoH audio sources are used with initial (greeting) announcements, Cisco recommends keeping the initial announcements less than 15 seconds in duration, otherwise you might need to reduce the maximum number of MoH streams per MoH server node to between 500 and 700 due to extra file I/O.
- Each media device may be disabled/enabled via the IPVMS Service Parameter (MoH is on the MoH device configuration page). It is possible to configure an MoH-only Unified CM node, and so forth.

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.

Music on Hold

Table 1-11 lists the VM configurations and the maximum number of simultaneous music-on-hold (MoH) streams each node can support. You should ensure that the actual usage does not exceed these limits, because once MoH maximum stream capacity has been reached, additional load could result in poor MoH quality, erratic MoH operation, or even loss of MoH functionality. Add additional MoH nodes (co-resident or dedicated) to increase Unified CM cluster MoH stream capacity.

Table 1-11 Music on Hold Maximum Per-Node Stream Capacity

Unified CM OVA Template	Co-resident MoH Streams (non-sRTP)[1]	Standalone MoH Streams
Small	500	750
Medium Large	750	1,000

1. All capacities based on non-sRTP streams.

As shown in Table 1-12, you can define a maximum of 500 unique sources of audio for Music on Hold in a Unified CM cluster. The maximum audio source capacities shown in Table 1-12 are per-cluster based on the VM configuration size and MoH server type (co-resident or standalone) used in the cluster. Adding MoH nodes to a Unified CM cluster increases only MoH stream capacity but does not increase audio source capacity. Audio source capacity can be increased only by moving from co-resident to standalone MoH nodes, increasing the cluster-wide node VM configuration size, or adding additional Unified CM clusters.

Table 1-12 Music on Hold Maximum Per-Cluster Audio Source Capacity

Unified CM OVA Template	Co-resident MoH Sources	Standalone MoH Sources
Small	100	250

Table 1-12 Music on Hold Maximum Per-Cluster Audio Source Capacity (continued)

Unified CM OVA Template	Co-resident MoH Sources	Standalone MoH Sources
Medium Large	250	500

The capacity limits described in Table 1-11 and Table 1-12apply to any combination of unicast, multicast, or simultaneous unicast and multicast streams.

Performance Considerations

To maximize the number of MoH audio sources and streams, you must reduce the number of some other media devices, such as disabling software MTPs and/or software conference bridges. The Cisco IP Voice Media Streaming Application service does not support maximum settings for all the media devices simultaneously. Oversubscribing the system resources (for example, CPU usage and disk I/O) with media devices would impact the overall system performance. An IPVMS alarm is issued if a media device is unable to meet provisioned capacity.

For low-end configurations (Small VM configuration) and MoH co-resident with moderate call processing, MoH is limited to a maximum of 500 streams, 100 MoH audio sources, and 48 to 64 annunciator streams with MTPs and conference bridges set at default values or disabled.

A dedicated Small VM configuration MoH node is required to support 750 MoH streams with 250 MoH audio sources and 250 annunciator streams.

To support a maximum of 1,000 MoH streams, 500 MoH audio sources, and 750 annunciators, the minimum requirement is a Medium OVA dedicated standalone MoH server.

Use of sRTP for MoH and/or annunciator will reduce the maximum number of MoH callers by 25%, and a dedicated IPVMS server for MoH and annunciator is highly recommended in this case.

The Unified CM MoH server supports four codecs: G.711 ulaw, G.711 mulaw, G729a, and Wideband audio. With unicast MoH, because the codec is negotiated during call setup, the number of MoH streams depends not on the number of MoH codecs enabled but on the number of endpoints that are on hold with unicast MoH. In the case of multicast MoH, each multicast-enabled audio source generates one MoH steam for each MoH codec enabled. For example, if 2 codecs are enabled and all 500 MoH sources are multicast-enabled, then 1,000 multicast MoH streams would be active even if no endpoints are on hold. In this scenario, if any endpoints are placed on unicast MoH, then additional MoH streams capacity would be required.

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.

Call Queuing (Hunt Pilot Queuing)

The maximum number of media streams that can be sent for call queuing is the same as with Music on Hold streams. See Music on Hold, page 30, for details.

The maximum number of hunt pilots with call queuing enabled is 100 per Unified CM subscriber node. The maximum number of simultaneous callers in queue for each hunt pilot is 100. The maximum number of members across all hunt lists does not change when call queuing is enabled.

LDAP Directory Integration

The Unified CM Database Synchronization feature provides a mechanism for importing a subset of the user configuration data (attributes) from the LDAP store into the Unified CM publisher database. Once synchronization of a user account has occurred, the copy of each user's LDAP account information may then be associated to additional data required to enable specific Unified Communications features for that user. When authentication is also enabled, the user's credentials are used to bind to the LDAP store for password verification. The end user's password is never stored in the Unified CM database when enabled for synchronization and/or authentication.

Note To ensure timely and successful client user authentication against LDAP, the required LDAP bind transaction between Unified CM and LDAP for authentication must be completed within 150 ms. This includes the time for network communication between Unified CM and LDAP. As such, LDAP bind transaction times that exceed 150 ms (whether due to long network round trip times or LDAP server performance) can result in queued authentication requests and delayed or failed authentication for clients. In cases of poor LDAP performance and/or slow response times, ensure the network round trip time between Unified CM nodes and LDAP nodes is well below the required 150 ms total transaction time to compensate for delayed LDAP responses.

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

The maximum number of users for a Unified CM cluster is limited by the maximum size of the internal configuration database that gets replicated between the cluster members. Currently the maximum number of users that can be configured or synchronized is 220,000. To optimize directory synchronization performance, Cisco recommends considering the following points:

- Directory lookup from phones and web pages may use the Unified CM database or the IP Phone Service SDK. When directory lookup functionality uses the Unified CM database, only users who were configured or synchronized from the LDAP store are shown in the directory. If a subset of users is synchronized, then only that subset of users is seen on directory lookup.
- When the IP Phone Services SDK is used for directory lookup, but authentication of Unified CM users to LDAP is needed, the synchronization can be limited to the subset of users who would log in to the Unified CM cluster.
- If only one cluster exists, if the LDAP store contains fewer than the maximum number of users supported by the Unified CM cluster, and if directory lookup is implemented to the Unified CM database, then it is possible to import the entire LDAP directory.
- If multiple clusters exist and if the number of users in LDAP is less than the maximum number of users supported by the Unified CM cluster, it is possible to import all users into every cluster to ensure directory lookup has all the entries.

- If the number of user accounts in LDAP exceeds the maximum number of users supported by the Unified CM cluster and if the entire user set should be visible to all users, it will be necessary to use the Unified IP Phone Services SDK to off-load the directory lookup from Unified CM.
- If both synchronization and authentication are enabled, user accounts that have either been configured or synchronized into the Unified CM database will be able to log in to that cluster. The decision about which users to synchronize will impact the decision on directory lookup support.

Note Synchronizing more user accounts can lead to starvation of disk space, slower database performance, and longer upgrade times.

Cisco Unified CM Megacluster Deployment

A Unified CM cluster is considered to be a megacluster when the number of call processing subscribers exceeds the normal cluster maximum of 4 pairs (whether due to increased user/device scale or just to provide more geographic coverage regardless of scale). A megacluster may have up to 8 pairs of call processing subscribers and no more than 21 server nodes in a single megacluster.

For example, you may have the publisher, TFTP, TFTP backup, MoH, MoH backup, 8 primary, and 8 backup servers counted toward the 21-server limit.

Note IM and Presence does not count toward the 21-server limit for a megacluster deployment.

Cisco IM and Presence has introduced a VM configuration template to align with megacluster deployments using a 25,000-user VM configuration.

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 twice the number of a normal cluster (8 call processing subscriber pairs).
- Maximum number of CTI devices and connections also doubles.

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

- Size of the configuration database
- Number of locations and regions
- Maximum number of LDAP synchronized or provisioned end users (220,000 users per cluster)

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

Cisco IM and Presence

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

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

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

- Number and type of users
  - Clients employed by users to obtain presence services.
  - Operating mode for users (instant messaging only or full Unified Communications facilities)
  - Average number of devices per user
- Presence-related activities performed by typical users.
  - Manual and call related presence changes including calendaring. Presence change impact is directly proportional to the average watcher list size (equivalent to contact list size) and composition (intracluster, intercluster, and federated). By default, the maximum contact list size is 200. If some users will exceed 200 contacts, this maximum contact list size can be changed by modifying the Presence Settings of the IM and Presence cluster.
  - 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 calls related, and user initiated)
- Deployment model
  - Whether centralized IM&P (one IM&P cluster to many Unified CM clusters) or distributed IM&P (one IM&P cluster **per** Unified CM cluster).
  - Whether intercluster presence is supported
  - Whether federation is supported
  - Whether high availability is desired
- Server preferences
  - The desired VM configuration size
- System options
  - Whether compliance recording is required

Once the system requirements are quantified, the number of required virtual machines can be determined from the data in Table 1-13.

Table 1-13 Maximum Number of Users Supported per Unified CM IM and Presence Cluster[1]

Unified CM IM & P OVA Template	Maximum Users	Maximum Client Devices Supported in Full Unified Communications Mode
Extra Small	1,000 Users	3,000
Small	5,000 Users	15,000
Medium	15,000 Users	45,000
Large	25,000 Users	75,000

1. Maximum supported sub-clusters are 3.

In some cases, IM and Presence nodes may require additional resources and thus larger OVA templates to operate effectively. IM and presence features have significant impact on system performance above and beyond the number of users assigned to IM and Presence and the number of devices per user.

Note OVA size refers to the total number of devices and does not reflect the impact the above features have on IM and Presence.

The following IM and Presence deployment types and features will require Medium OVA template or higher:

- Centralized IM and Presence deployments (Large OVA recommended) — deployments with one (or more) IM and Presence cluster and multiple.
- Unified CM clusters
- Multi-cluster IM and Presence deployments — deployments with two (or more) Unified CM clusters each with IM and Presence sub-clusters or with two (or more) IM and Presence clusters.
- Persistent chat
- Message archiving
- 3rd party compliance
- Multiple device messaging (MDM)
- Managed file transfer (MFT)
- Outlook integration (Jabber client)

Failure to provide additional resources by using a larger OVA template for IM and Presence deployment types and features above will result in higher system CPU, IM and Presence service core dumps, persistent chat and other performance related issues.

For additional information on Cisco IM and Presence, refer to the latest version of the *Compatibility Matrix for Cisco Unified Communications Manager and the IM and Presence Service*.

The formal definitions of the VM configurations for Cisco IM and Presence.

Impact on Unified CM

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

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

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

IM and Presence VM configurations differ from Unified CM VM configurations. IM and Presence

templates are user based while Unified CM templates are device based. For example, a Small IM and Presence VM configuration used with a Unified CM Large VM configuration would support 5,000 users with 2 devices each. All IM and Presence nodes within the same cluster must use the same type of VM configuration.

Centralized IM and Presence

Cisco IM and Presence supports a centralized deployment option. A centralized IM and Presence cluster can provide presence service for users on multiple remote Unified CM clusters; however, the total number of users across all the remote Unified CM clusters must not exceed 75,000, assuming that each user has a single client. Multiple clients per user would reduce this limit.

Note The centralized IM and Presence cluster requires a Unified CM publisher node, for a total of 7 servers in the cluster: 3 IM and Presence sub-cluster pairs (6 servers) + the Unified CM publisher node.

For deploying a centralized IM and Presence cluster, we recommend using the Large IM and Presence VM template for all the IM and Presence nodes in the cluster and using the Large Unified CM VM template for the Unified CM publisher node of that centralized cluster.

The centralized IM and Presence deployment can be clustered over the WAN, subject to the following restrictions:

- All remote Unified CM clusters must be within 80 ms round-trip-time (RTT) of the centralized IM and Presence cluster.
- A centralized IM and Presence cluster may be connected to another centralized IM and Presence cluster by means of an intercluster trunk with a maximum latency of 300 ms RTT.
-

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 1-14 shows the VM configurations that support the Emergency Responder and their maximum capacities.

Note These limits apply to standalone Emergency Responder deployments, and they assume that Native Emergency Services are not being used.

Table 1-14 Cisco Emergency Responder VM Configurations and Capacities

VM Configuration	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
20,000 Users	20,000	5,000	2,000	1,000	60,000	7,500
30,000 Users	30,000	10,000	3,000	2,000	120,000	10,000
40,000 Users	40,000	12,500	4,000	2,500	150,000	12,500

The formal definitions of the VM configurations for Cisco Emergency Responder and other Unified Communication products.

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

cluster.

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

Cisco Expressway

Note Cisco Expressway is no longer sized within the Cisco Collaboration Sizing Tool. Instead use the Cisco Expressway Sizing Tool available from the *Collaboration Sizing, Configuration & Quote* site at https://cucst.cloudapps.cisco.com/.

Cisco Expressway deployments rely on Cisco Unified CM as the component for call control, including remote endpoint registration. When sizing such a system, consider the function it performs as well as its impact to Unified CM.

When sizing Cisco Expressway, you typically must consider the following parameters to determine the required number of Cisco Expressway-C and Expressway-E node pairs:

- Number of endpoint registrations through each pair of Expressway-C and Expressway-E nodes during peak usage time
- Expected number of simultaneous voice-only and video calls traversing each pair of Expressway-C and Expressway-E nodes

Expressway-C and Expressway-E clusters support a maximum of 6 nodes.

Mobile and remote access (MRA) does not require any specific licenses, but business-to-business communication requires rich media licenses. Licenses in the form of rich media sessions are shared across an Expressway cluster. Each Expressway node in the cluster contributes its assigned rich media sessions to the cluster database, which is then shared across all of the nodes in the cluster. This model results in any one Expressway node being able to carry many more licenses than its physical capacity.

Cisco Expressway Capacity Planning

Table 1-15 lists the Cisco Expressway proxy registrations and call capacities for Cisco Expressway-C and Expressway-E server node pairs and clusters.

Table 1-15 Cisco Expressway-C and Expressway-E Node and Cluster Capacities

Platform	Proxy Registrations[1]	Video Calls	Audio-only Calls
CE1300	7,500 per node 30,000 per cluster	500 per node 2,000 per cluster	1,000 per node 4,000 per cluster
Large OVA	4,000 per node 16,000 per cluster	500 per node 2,000 per cluster	1,000 per node 4,000 per cluster
Medium OVA	3,000 per node 12,000 per cluster	150 per node 600 per cluster	300 per node 1,200 per cluster
Small OVA (Business Edition 6000)[2]	200 per node 200 per cluster	20 per node 20 per cluster	40 per node 40 per cluster

1. Proxy registration applies only to mobile and remote access (MRA) connections, not business-to-business communications.
2. Cisco Expressway-C and Expressway-E can be clustered across multiple Business Edition 6000 nodes for redundancy purposes; however, there is no increased capacity when clustering with Business Edition 6000.

Note The capacity numbers in Table 1-15 assume Fast Path Registration for MRA is enabled on Expressway-E

Note The large OVA template is supported only with limited hardware. Refer to the documentation on Cisco Collaboration Virtualization for more information.

The following guidelines apply when clustering Cisco Expressway:

- Expressway clusters support up to 6 nodes (cluster capacity up to 4 times the node capacity).
- The capacity of all nodes across and within each Expressway-E and Expressway-C cluster pair must be the same. For example, an Expressway-E node using the large VM configuration must not be deployed if other nodes in the Expressway-E cluster or in the corresponding Expressway-C cluster are using the medium size VM configuration.
- Expressway peers should be deployed in equal numbers across Expressway-E and Expressway-C clusters. For example, a three-node Expressway-E cluster should be deployed with a three-node Expressway-C cluster.
- An Expressway-E and Expressway-C cluster pair can be formed by a combination of nodes running on an appliance or running as a virtual machine, as long as the node capacity is the same across all nodes.
- The Expressway node VM configurations or Expressway Appliances must match across and within Expressway Series cluster pairs.
- Multiple pairs of Expressway Series clusters may be deployed to increase capacity.

Note There is a dependency between Cisco Expressway clusters and Cisco Unified CM clusters. Expressway capacity planning must also consider the capacity of the associated or dependent Unified CM cluster(s).

For more information about Cisco Expressway capacity planning considerations, including sizing limits, capacity planning, and deployment considerations, refer to the Cisco Expressway product documentation.

Gateways

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

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

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

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

Gateway Groups

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

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

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

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

Voice Messaging

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

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

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

Table 1-16 shows the applicability of the various voice messaging solutions to the scalability requirements of the deployment.

Table 1-16 Scaling Voice Messaging Solutions

	Maximum Number of Users Supported on a Single Node (or Failover or Clustered Deployment)				Maximum Number of Users Supported in a Digital Networking Solution	Maximum Number of Users Supported in an HTTPS Networking Solution
Solutions	**500**	**1,000**	**15,000**	**20,000**	**100,000**	**100,000**
Cisco Unity Express	Yes	No	No	No	Yes	No
Cisco Business Edition	Yes	Yes	No	No	No	No
Cisco Unity Connection (Unified/Integrated Messaging and Cisco Business Edition 7000)	Yes	Yes	Yes	Yes	Yes	Yes

Table 1-17 shows the maximum limits of various functions of different VM configurations running Cisco Unity Connection.

Table 1-17 VM Configurations and Capacities for Cisco Unity Connection

VM Configuration	Maximum Number of Ports	Maximum Voice Recognition Sessions	Maximum Text to Speech Sessions	Maximum Number of Voicemail Users
1,000 Users	24	24	24	1,000
5,000 Users	100	100	100	5,000
10,000 Users	150	150	150	10,000
20,000 Users	250	250	250	20,000

Refer to the formal definitions of VM configurations for Cisco Unity Connection and the Cisco Unity Connection Supported Platforms List for more information.

Impact on Unified CM

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

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

Collaborative Conferencing

Collaborative conferencing system can be cloud based, on-premises based, or hybrid. In case of on premise based or hybrid system, Cisco Unified CM servers (VMs) are needed in addition to conferencing servers (VMs).

When sizing conferencing servers, the following parameters should be considered in order to determine the type and number of servers/nodes:

- Number of registered conferencing system
- Expected number of conference participants (audio, video, and web) during the peak usage interval
- Required dial-in duration for all the participants to join different conferences on top at the peak usage interval (usually top of the busiest hour)
- Video and audio quality (resolution, codec type)

General Sizing Guidelines for Conferencing

There are several methods for calculating conferencing resource requirements:

- Calculation based on average monthly usage.

  If the average conferencing usage statistics (minutes per month spent on conference calls) is available, Table 1-18 can be used to roughly calculate the conferencing capacity requirements in terms of the number of ports needed.

Table 1-18 Conferencing Capacity Based on Average Monthly Usage

Average Monthly Usage (minutes)	Baseline Usage (minutes per port per month)	Estimated Number of Ports
20,000 to 50,000	500	40 to 100
50,001 to 500,000	1,000	101 to 550
500,0001 to 1,000,000	2,000	551 to 800
1,000,001 to 2,000,000	3,000	801 to 1,133
2,000,001 to 8,000,000	4,000	1,134 to 2,633

- Calculation based on number of conferencing system users.

  Assuming light usage of the conferencing system, one should plan on having one port for every 20 to 50. This would assume 2-5% of users on a conference call during the peak usage interval. Assuming average conferencing usage, we would assume 5-10% of users on conference calls during peak internal, thus one port for every 10-20 users. Heavy conferencing usage can assume one port for every 5-10 users needed – 10-20% of users on conference calls during peak hours. For example, on a system with 6,000 users, you should provision 120-1,200 ports depending on system usage.

- Calculation based on actual peak interval usage.

  Actual conferencing usage during peak hours usually can be obtained from existing conferencing system logs or service provider bills. Conferencing statistics usually provide the number of conferences and the average number of participants per conference. Product of these two numbers would define the average number of active conference participants, thus the number of ports needed. The actual number of ports needed can be actually much higher depending on the deployments size – for the smaller deployment, the higher max to average ratio would be required.

Sizing Guidelines for Voice, Video, and Web Conferencing

When sizing modern collaboration conferencing system, the product capacity data sheet provides supported limits under different usage conditions: audio/video, recording, streaming, and content sharing.

The key factor for proper sizing (determination of the number of servers/systems needed) is the accurate estimation of users simultaneously on collaboration meetings together with the duration of interval during which users join meetings at the beginning of the busy interval.

Note: percentage of video vs. audio calls is not an important factor since audio and audio/video capacity is the same. Video quality might be a factor and lead to reduced capacity should higher video quality is required.

Table 1-19 lists important sizing considerations for audio, video, and web conferencing system capacity planning. To accurately size a conferencing deployment, it is important to quantify and understand things like the maximum number of concurrent users, the maximum number of concurrent conferences, and the maximum number of participants in a single conference supported by the system.

Table 1-19 Audio, Video, and Web Conferencing System Sizing Considerations

Sizing Consideration	Details
Maximum Concurrent Meeting Connections (Audio, video, and web users)	The number of people participating in concurrent meetings on the system at any given time.
Maximum Simultaneous Audio Connections (Teleconference phone calls and voice connection using computer from meeting clients)	The system capacity should remain the same, regardless of what combination of audio codec (G.711, G.722, G.729), IP version (IPv4 or IPv6), or traffic encryption (TCP or TLS, RTP or SRTP).
Maximum Concurrent Video and Video Sharing Users	This is the maximum number of concurrent meeting connections (or participants) allowed to use video sharing on the system at the same time. Note that in most cases, if one participant in a meeting uses video, then all other users in the same meeting are counted as video users (even if they are not using video). Likewise, desktop sharing is generally not considered video for the purposes of concurrent video and sharing.
Maximum Participants in One Meeting	The maximum number of participants that can attend a meeting on the system.
Maximum Meetings that can be Recorded	This is the total number of meetings that can be recorded simultaneously by the system at any given time.
Maximum Concurrent Recording Playback Sessions	This is the total number of recording playback sessions that can be handled by the system simultaneously. This refers to recording that are stored on the system and not recordings that are downloaded by the user to their desktop.
Maximum Concurrent Meetings	The number of separate meetings that can be active concurrently on the system.
Maximum Call Rate (calls/second)	This is the average number of users that can join a meeting during a one second time period. After the system reaches this number, the next user(s) that attempt to join may experience an additional few seconds wait before connecting to the meeting.

Table 1-19 Audio, Video, and Web Conferencing System Sizing Considerations (continued)

Sizing Consideration	Details
Maximum Concurrent Sign-in	This is the average number of users who can simultaneously sign into the system during a one second time period. After the system reaches this number, the next few users to sign in the system might experience an additional few seconds wait before they are fully signed into the system.
Maximum Aggregate Bandwidth Utilization	This is the maximum bandwidth the system can handle in aggregate at any point in time.

Note The percentage of video v audio call is not typically an important factor for conference system sizing since audio/video capacity tends to be the same. However, video quality is an important factor to consider as higher quality video will reduce the overall system conferencing capacity.

Conferencing Impact on Unified CM

Extra call volume increases capacity/resource requirements for Unified CM. Call processing resources are impacted during the few minutes period at the beginning and at the end of busy conferencing period. Assuming 10% of busy hour users on collaboration calls, supported call per second rate need to be at least 40% higher compared to no users on collaboration calls and 4 BHCA per user.

Memory requirement is also higher. 10% of users on conference calls will also increase the number of simultaneous calls by at least 40%.

Sizing for Standalone Products

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

- Cisco Unified Communications Manager Express, page 44
- Cisco Business Edition, page 45

Cisco Unified Communications Manager Express

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

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

Cisco Business Edition

Cisco Business Edition is a packaged collaboration solution that is preloaded with premium services for voice, video, mobility, messaging, conferencing, instant messaging and presence, and contact center applications.

Cisco Business Edition 6000 and 7000 both have platform model options to choose from.

Cisco Business Edition 6000 is currently available in one hardware platform option:

- BE6000M — Maximum capacity of 1,000 users; 1,200 devices; and 100 contact center agents. Typically supports five collaboration application options in a single virtualized server platform. Maximum of 5,000 BHCA.

To learn more about Cisco Business Edition 6000 solutions.

Cisco Business Edition 7000 is available in two hardware platform options:

- BE7000H — This high-density model typically supports five to ten collaboration applications in deployments sized for 1,000 to 5,000 users with 3,000 to 15,000 devices and multiple sites.
- BE7000M — This medium-density model typically supports four to six collaboration applications in deployments sized for 1,000 to 5,000 users with 3,000 to 15,000 devices and multiple sites.

To learn more about Cisco Business Edition 7000 solutions.

Busy Hour Call Attempts (BHCA) for Cisco Business Edition

As mentioned above, Business Edition 6000M supports a maximum of 5,000 BHCA. When calculating your system usage, stay at or below this BHCA maximum to avoid oversubscribing Cisco Business Edition 6000. 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 6000M

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 8800 Series or other Collaboration voice and video endpoints, a software client such as Cisco Jabber, an analog phone port, or an H.323 client. While Cisco Business Edition 6000 supports a maximum of 1,200 endpoints on a medium-density server, as indicated above, actual endpoint capacity depends on the total system BHCA.

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

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

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

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

100 phones at 4 BHCA is 100∗4 = 400

80 phones at 12 BHCA is 80∗12 = 960

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

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

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

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

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

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

100 phones at 8 BHCA = 800 BHCA

150 phones at 4 BHCA = 600 BHCA

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

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

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

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

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

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

If you are using Cisco Unified Mobility for single number reach (SNR) on Business Edition 6000, keep in mind that calls extended to remote destinations and mobility identities, 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.

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

Another aspect of capacity planning to consider for Cisco Business Edition 6000 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 6000. 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 6000, Cisco recommends configuring no more than three members of a hunt or line group when a broadcast distribution algorithm is required. Depending on the load of the system, doing so could greatly affect the BHCA of the system and possibly oversubscribe the platform's resources. The number of hunt or line groups that have a distribution algorithm of broadcast should also be limited to no more than three. These are best practice recommendations meant to prevent over-subscription of the system BHCA. Exceeding these recommendations within a deployment is supported as long as the overall BHCA capacity of the system is not exceeded.

Mixing different types of hardware platforms within a Unified CM cluster is also allowed. However, because not all VM configurations are supported on all server platforms, mixing VM configurations will impact the overall cluster capacity.

Cisco Unified Mobility for Cisco Business Edition 6000

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

Each configured remote destination or mobility identity has potential BHCA implications. For every remote destination or mobility identity configured for a user, one additional call leg is used. Because each call consists of two call legs, one remote destination ring is equal to half (0.5) of a call. Therefore, you can use the following formula to calculate the total remote destination BHCA:

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

For example:

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

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

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

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

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

Simplified Sizing Examples

This section introduces a set of simplified sizing examples meant to provide guidance for four different sized deployments. It begins by summarizing the simplified sizing examples and then providing a general set of assumptions in terms of validated design and deployment best practices which apply to all of the sizing examples. These assumptions and the sizing examples apply to the calling, IM & presence, voice messaging, edge, and meeting workloads supported by the deployment. Next, for each sizing example a set of deployment size specific assumptions is provided followed by a virtual machine placement diagram showing the type and quantity of the various required virtual machines distributed over some number of Business Edition 7000 (BE7000) platform servers. These simplified sizing examples are based on the guidance and best practices documented in the Enterprise On-Premises Preferred Architecture (PA) for release 15.

Note The hardware used for these simplified sizing examples is specific to the hardware models referenced (BE7000M M6 / BE7000H M6 and CMS 1000 M6 / CMS 2000 M6) and available at time of publication. If newer higher density, higher performance hardware is used, VM density and therefore layout may change given overall platform capacity increase.

The following simplified sizing examples are available:

- Small, page 52
- Medium, page 54
  - Medium #1, page 54
  - Medium #2, page 55
- Large, page 57

Table 1-20 below provides a brief summary of the four simplified sizing examples covered here. This table covers the size and quantity of the virtual machine (VM) host platforms, the collaboration application VMs and appliances, and the total vCPUs, vRAM, and vDisk requirements for each sizing example.

Table 1-20 Simplified Sizing Example Summary

	Small	Medium #1	Medium #2	Large
Deployment Size	Up to 1,000 users or devices	Up to 5,000 users or devices	Up to 10,000 users or devices	Up to 20,000 users or devices
VM Host Platform[1]	BE7000M	BE7000M	BE7000M	BE7000H
Total VM Host Platforms (minimum)	4	4	6	6
Total VMs	15	18	26	36
Total vCPUs	32 (of 64)	44 (of 64)	76 (of 96)	136 (of 168)
Total vRAM[2]	98 GB	166 GB	250 GB	380 GB
Total vDisk	1.72 TB	2.23 TB	3.85 TB	5.69 TB
Cisco Meeting Server (CMS) Platform[3]	CMS 1000	CMS 1000	CMS 1000	CMS 2000
Total CMS Platforms	2	2	6	2

Table 1-20 Simplified Sizing Example Summary

	Small	Medium #1	Medium #2	Large
CUBE Platform	ASR 1000 / Catalyst 8000 / ISR 1100 / ISR 4000			
Total CUBE Platforms	Variable based on locations/sites and redundancy - up to 1,000 locations	Variable based on locations/sites and redundancy - up to 2,000 locations		

1. BE7000M M6 with 16 cores or BE7000H M6 platform with 28 total cores assumed for all examples as indicated. Newer BE7000 models with larger numbers of cores will provide improved density.
2. ESXi overhead not included. An additional 8 GB of memory per host is required for ESXi 7. And additional 12 GB of memory per host is required for ESXi 8.
3. CMS 1000 M6 or CMS 2000 M6 assumed for all examples as indicated. Newer CMS 1000 and CMS 2000 platforms may provide increased capacity.

Note High availability (HA) is enabled and assumed for all configured components and endpoints (including Cisco Jabber).

Note The numbers in Table 1-20 are based on current VM OVA templates and platforms and published capacities. Because new VM OVA templates and platforms are introduced and existing platform capacities change, it is always a good idea to refer to the latest product data sheets and the Cisco Collaboration Virtualization information page.

General Assumptions

Table 1-21 below lists the base set of design and deployment assumptions applying to the simplified sizing examples (of all sizes) cover here.

Table 1-21 Simplified Sizing General Assumptions for Collaboration Deployments

Workload - Application / Platform	Assumptions (for all example sizes)
Calling - Unified CM	• All servers are placed at the Headquarter site within single centralized cluster (Centralized Call Processing deployment model). • Average of up to 4 busy hour call attempts (BHCA, the number of call attempts during the busy hour) per user. Call holding time does not exceed 3 minutes. • Traffic mix: 50% of BHCA traffic to/from PSTN via SIP trunks, , 50% of BHCA traffic is intracluster. • Average of up to 2 DNs per device. • Media and SIP signaling encryption may be enabled. • Up to 3,000 partitions; 6,000 calling search spaces (CSSs); and 12,000 translation patterns. • Up to 1,000 route patterns; 1,000 route lists; and 2,100 route groups. • Unified CM media resources: – Unified CM software conference bridges (software CFBs) and Unified CM media termination points (MTPs) are not included in these examples. Instead, Cisco Meeting Server and Cisco IOS-based MTP are included. – 48 annunciators per call processing pair, 250 concurrent music on hold (MoH) sessions per call processing pair. For a larger number of annunciators or concurrent MoH sessions, deploy standalone Unified CM subscribers as MoH servers. – 5% of users simultaneously receiving unicast MoH streams. • Average of up to one remote destination / mobility identity per mobility user. • Computer Telephony Integration (CTI) - All devices can be enabled for CTI, with up to 5 lines per device and 5 J/TAPI applications monitoring the same CTI device. • Gateway - Up to 2,100 per cluster. • Locations and regions - When adding regions, select Use System Default for the Audio Codec Preference List and Audio and Session Bit Rate values. Changing these values for individual regions from the default has an impact on server initialization and publisher upgrade times. Hence, with a total of 2,000 regions you can modify up to 200 regions to use non-default values. With a total of 1,000 or fewer regions, you can modify up to 500 of them to use non-default values. • Extension Mobility (EM) - All users can use EM. No Extension Mobility Cross Cluster (EMCC) users. • All users can use Web Dialer. • Up to 50,000 users synchronized from LDAP, but active BHCA user up to specified example deployment size.
IM & Presence - Unified CM IM&P	• All Jabber users are active IM and presence users.

Table 1-21 Simplified Sizing General Assumptions for Collaboration Deployments

Workload - Application / Platform	Assumptions (for all example sizes)
Voice Messaging - Unity Connection	• Media and SIP signaling encryption can be enabled without changing this Unity Connection simplified sizing. • All Jabber users leverage visual voicemail with redundancy. • There is a single inbox for all users (Unified Messaging). • Notifications of voice messages (new message, message update, and message deleted) use HTTP (not HTTPS). • G.711 Codec is used. • Voicemail is being recorded by up to 20% of users at the time. • Voicemail recording length up to 1 minute.
Edge - Expressway	• Cisco Expressway enables mobile and remote access (MRA) remote connectivity for users working outside of the office and business-to-business (B2B) calling for communication with other organizations. • Expressway-E VMs are expected to be placed on DMZ located host servers, however, to better summarize the overall VM requirements for the virtual machine placement examples, the Expressway-E VMs have been included on the same set of BE7000 servers as all the other VMs. In a production deployment the Expressway-E VMs would reside on separate hardware in the DMZ (BE7000 or other hardware). • The Expressway Large OVA template is not supported on the BE7000 platform, as such the Large OVA template is out of scope for these simplified sizing examples. • All video calls are encrypted. The average call rate across all the video calls is 768 kbps. • For example, half of the video calls could be at 384 kbps and the other half at 1152 kbps. • All audio calls are encrypted, and the average bandwidth across all audio calls is 64 kbps. • Expressway clusters support up to 6 nodes (cluster capacity up to 4 times the node capacity). • Expressway-E and Expressway-C nodes cluster separately; an Expressway-E cluster consists of Expressway-E nodes only, and an Expressway-C cluster consists of Expressway-C nodes only. Expressway peers should be deployed in equal numbers across Expressway-E and Expressway-C clusters. For example, a three-node Expressway-E cluster should be deployed with a three-node Expressway-C cluster. • The capacity of all nodes across and within each Expressway-E and Expressway-C cluster pair must be the same. For example, an Expressway-E node using the large OVA template must not be deployed if the nodes in the Expressway-E cluster or in the corresponding Expressway-C cluster are using the medium OVA template.
Edge - CUBE	• The CUBE serves as PSTN gateway (IP or TDM) as well as SRST router in cases of lost remote site connectivity. • CUBE is one of the following platform series: ASR 1000, Catalyst 8000, ISR 4000, or ISR 1100.

Table 1-21 Simplified Sizing General Assumptions for Collaboration Deployments

Workload - Application / Platform	Assumptions (for all example sizes)
Meetings - CMS / CMM / TMS	• One conference session per user (regardless of per user endpoint count). • Full HD conference quality assumed. • High availability N+1 schema. High availability deployment requires a minimum of 3 CMS Database VMs. • Maximum of 8 CMS Bridges (CMS Conferencing servers) per cluster without BU approval. • For the purposes of these sizing examples the CMM VMs are deployed on the BE7000 VM hosts. Optionally, these VMs could instead be hosted on the CMS nodes themselves.
Management - Prime Collaboration Deployment (PCD)	• PCD is deployed in all examples as it assists with deployment and installation as well as ongoing application node maintenance (upgrades, hardware moves, etc.) of Unified CM, Unified CM IM&P, and Unity Connection application nodes. • A single PCD VM is deployed without redundancy for all sizing examples.

Virtual Machine Placement

The virtual machine (VM) placement examples for the simplified sizing examples illustrate one way to layout the required VMs, however, there are numerous ways to distribute the VMs across the BE7000 platform. The important part is to separate redundant VM nodes for each application onto different BE7000s such that a single BE7000 failure does not fully eliminate all instances of a required application in the deployment.

While there is spare capacity on each BE7000 in these examples to accommodate deployment growth requiring more application VMs, this spare capacity also provides room for temporarily moving application VMs between the BE7000 servers in order to free up any one server for software and/or hardware maintenance operations. Additional BE7000 VM host servers may be deployed to accommodate additional collaboration application VMs (for example, Cisco Emergency Responder) as well as future growth of existing applications and ongoing VM and server maintenance.

Note Cisco Meeting Server (CMS) and CUBE are not shown in the virtual machine placement layout figures because they are deployed on CMS 1000 or CMS 2000 and CUBE IOS/IOS-XE router platforms, respectively.

Small

The small, simplified sizing example is for deployments of up to 1,000 users or devices and includes calling, IM and presence, voice messaging, edge services, and meeting workloads. All sizing assumes collaboration application virtual machines are hosted on the Business Edition 7000M (BE7000M) platform.

This sizing example is based on the following workload specific assumptions:

- Calling
  - Deployment is based on the Unified CM Small OVA nodes deployed with '1:1 Redundancy'.
  - Up to 1,000 locations, regions, and device pools.
  - Total number of DNs 2,000 with up to 100 DNs shared across average of 3 additional endpoints.

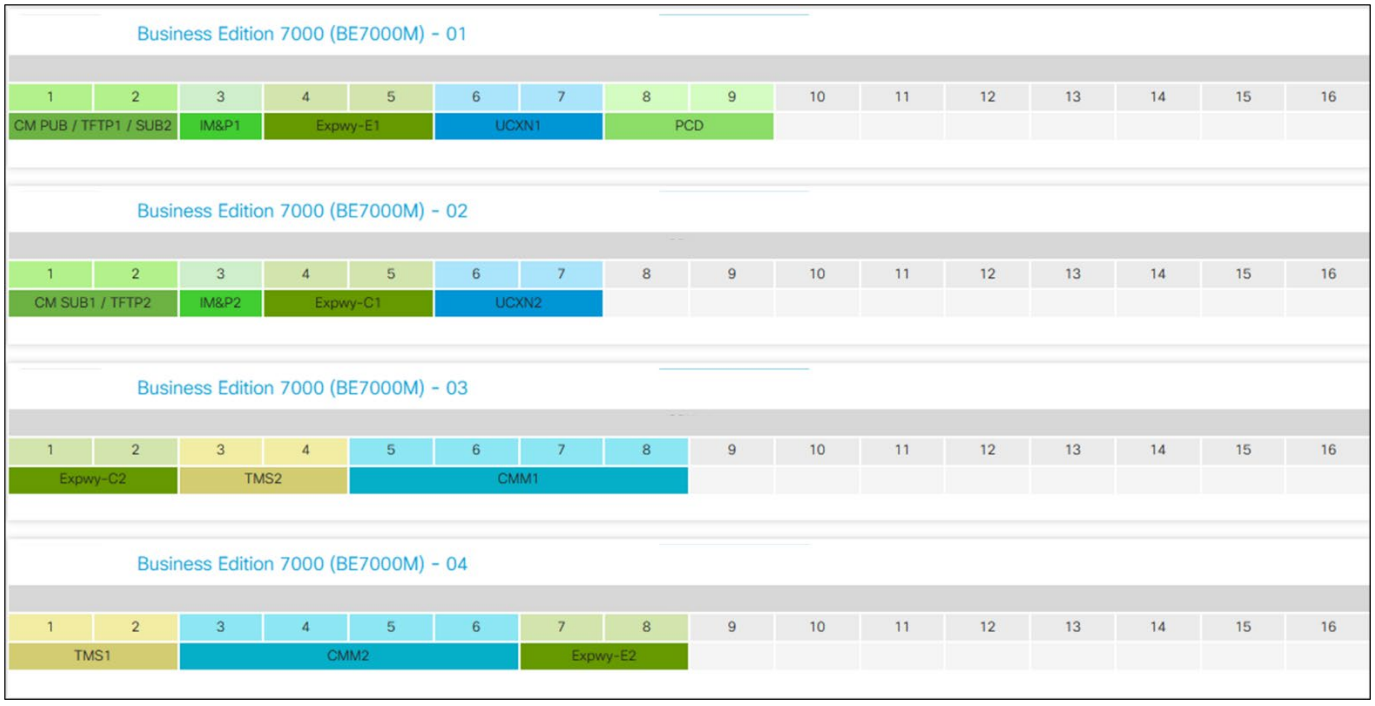
  - Up to 25 hunt pilots, 25 hunt lists, 15 circular/sequential line groups with an average of 5 members per line group, and 15 broadcast line groups with an average of 10 members per line group.
  - Up to 100 CTI ports and 25 CTI route points.
  - Up to 125 EM logins/logouts per minute supported.
- IM & Presence
  - Unified CM IM&P nodes are deployed on the Extra-Small OVA/VM.

Note With this example, advanced IM&P features (including persistent chat and managed file transfer (MFT)) are not enabled. Enabling these features would at a minimum require the use of Small OVA/VMs.

- Voice Messaging
  - Unity Connection nodes are deployed on the 1,000 user OVA/VM.
- Edge
  - Expressway-C and Expressway-E nodes are deployed on the Small OVA template.
- Meetings
  - Cisco Meeting Server (CMS) is deployed on the CMS 1000 platform.
  - Cisco Meeting Management (CMM) nodes are deployed on the Small OVA/VM.
  - TelePresence Management Suite (TMS) nodes are deployed on the Regular OVA.

Figure 1-3 shows a sample layout of the collaboration application virtual machines required for this small deployment example.

Figure 1-3 Virtual Machine Placement for Small Simplified Sizing Example



Medium

The medium sizing examples are for deployments of up to 10,000 users or devices and includes calling, IM and presence, voice messaging, edge services, and meeting workloads. All sizing assumes collaboration application virtual machines are hosted on the Business Edition 7000M (BE7000M) platform.

Note The primary difference between the medium #1 example VM layout in Figure 1-4 and the medium #2 example VM layout in Figure 1-5 is that the medium #2 example adds a second Unified CM subscriber pair (CM SUB3 and CM SUB4) and another Expressway-C/Expressway-E pair (Expwy-C3 and Expwy-E3) to handle the increased user/device load. In addition, the medium #2 example requires larger Unity Connection OVAs (UCXN1 and UCXN2) and additional and larger TMS OVAs (TMS1, TMS2, TMS-SQL1, TMS-SQL2, TMSXE1, and TMSEXE2).

Medium #1

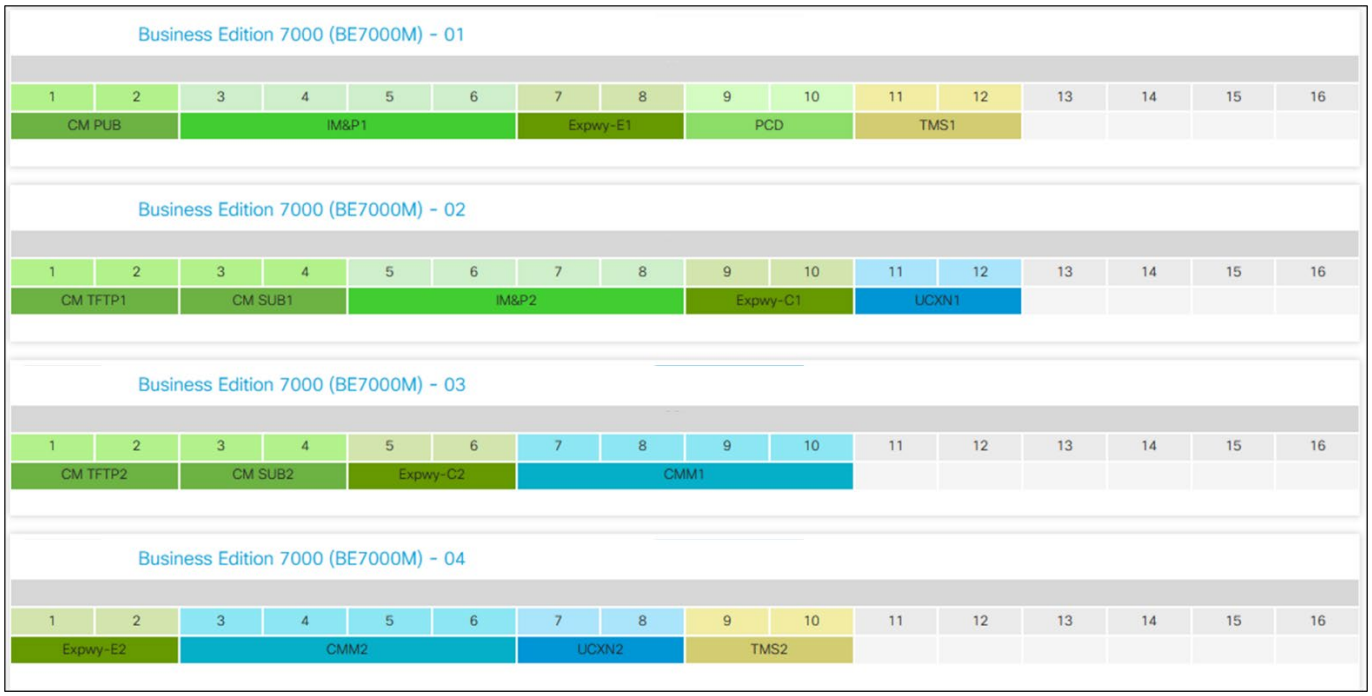
The first medium simplified sizing example is for deployments of up to 5,000 users This example is based on the following workload specific assumptions:

- Calling
  - Deployment is based on the Unified CM Medium OVA nodes deployed with '1:1 Redundancy'.
  - Up to 2,000 locations, regions, and device pools.
  - Total number of DNs 10,000 with up to 500 DNs shared across average of 3 additional endpoints.
  - Up to 100 hunt pilots, 100 hunt lists, 50 circular/sequential line groups with an average of 5 members per line group, and 50 broadcast line groups with an average of 10 members per line group.
  - Up to 500 CTI ports and 100 CTI route points
  - Up to 500 EM logins/logouts per minute supported.
- IM & Presence
  - Unified CM IM&P nodes are deployed on the Medium OVA/VM.
  - Managed file transfer (MFT) and persistent chat are both enabled for all IM&P users.
- Voice Messaging
  - Unity Connection nodes are deployed on the 5,000 user OVA/VM.
- Edge
  - Expressway-C and Expressway-E nodes are deployed on the Medium OVA template.
- Meetings
  - Cisco Meeting Server (CMS) is deployed on the CMS 1000 platform.
  - Cisco Meeting Management (CMM) nodes are deployed on the Small OVA/VM.
  - TelePresence Management Suite (TMS) nodes are deployed on the Regular OVA.

Note The general assumptions included in Table 1-20, the medium #1 example assumptions above, and the VM layout example shown in Figure 1-4 below aligns with the simplified sizing example to be included in the Sizing chapter of the forthcoming Enterprise On-Premises PA for Release 15 Cisco Validated Design (CVD) guide.

Figure 1-4 shows a sample layout of the collaboration application virtual machines required for the medium #1 deployment example.

Figure 1-4 Virtual Machine Placement for Medium #1 Simplified Sizing Example



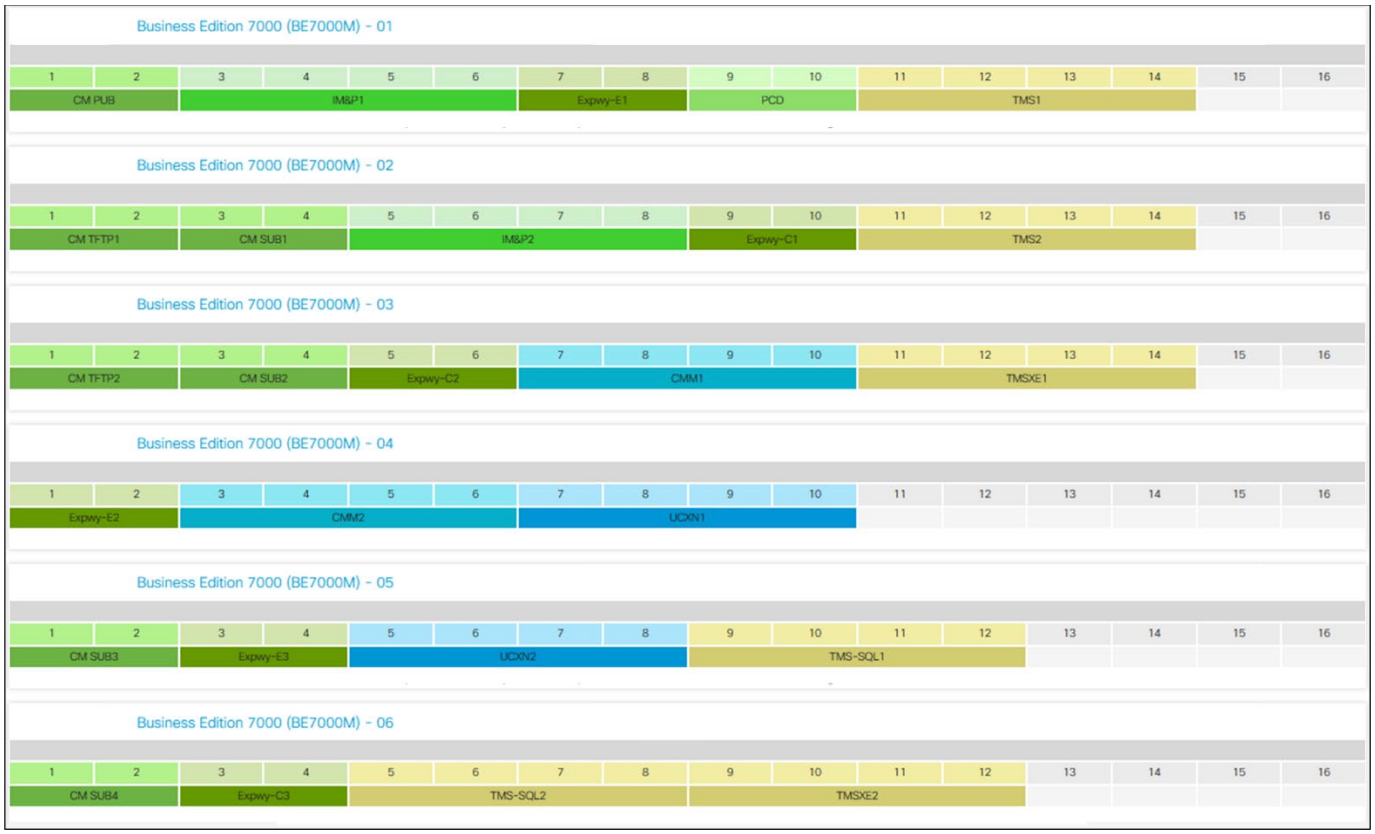
Medium #2

The second medium simplified sizing example is for deployments of up to 10,000 users or devices. This sizing example is based on the following workload specific assumptions:

- Calling
  - Deployment is based on the Unified CM Medium OVA nodes deployed with '1:1 Redundancy'.
  - Up to 2,000 locations, regions, and device pools.
  - Total number of DNs 20,000 with up to 1,000 DNs shared across average of 3 additional endpoints.
  - Up to 200 hunt pilots, 200 hunt lists, 100 circular/sequential line groups with an average of 5 members per line group, and 100 broadcast line groups with an average of 10 members per line group.
  - Up to 1,000 CTI ports and 200 CTI route points.
  - Up to 750 EM logins/logouts per minute supported.
  - Up to 150 Unified CM Assistants supporting up to 150 Unified CM Managers.
  - Up to two Attendant Console Servers with up to 5 attendants monitoring up to 1,000 DNs.
- IM & Presence
  - Unified CM IM&P nodes are deployed on the Medium OVA/VM.
  - Managed file transfer (MFT) and persistent chat are both enabled for all IM&P users.
- Voice Messaging
  - Unity Connection nodes are deployed on the 10,000 user OVA/VM.
- Edge
  - Expressway-C and Expressway-E nodes are deployed on the Medium OVA template.
- Meetings
  - Cisco Meeting Server (CMS) is deployed on the CMS 1000 platform.
  - Cisco Meeting Management (CMM) nodes are deployed on the Small OVA/VM.
  - TelePresence Management Suite (TMS) nodes are deployed on the Large OVA/VM.

Figure 1-5 shows a sample layout of the collaboration application virtual machines required for the medium #2 deployment example.

Figure 1-5 Virtual Machine Placement for Medium #2 Simplified Sizing Example



Large

The large, simplified sizing example is for deployments of up to 20,000 users or devices and includes calling, IM and presence, voice messaging, edge services, and meeting workloads. All sizing assumes collaboration application virtual machines are hosted on the Business Edition 7000H (BE7000H) platform.

This sizing example is based on the following workload specific assumptions:

- Calling
  - Deployment is based on the Unified CM Large OVA nodes deployed with '1:1 Redundancy'.
  - Up to 2,000 locations, regions, and device pools.
  - Total number of DNs 40,000 with up to 2,000 DNs shared across average of 3 additional endpoints.
  - Up to 500 hunt pilots, 500 hunt lists, 250 circular/sequential line groups with an average of 5 members per line group, and 200 broadcast line groups with an average of 10 members per line group.
  - Up to 2,000 CTI ports and 500 CTI route points
  - Up to 750 EM logins/logouts per minute supported.
  - Up to 300 Unified CM Assistants supporting up to 300 Unified CM Managers.
  - Up to four Attendant Console Servers with up to 10 attendants monitoring up to 2,000 DNs.
- IM & Presence
  - Unified CM IM&P nodes are deployed on the Large OVA/VM.
  - Managed file transfer (MFT) and persistent chat are both enabled for all IM&P users.
- Voice Messaging
  - Unity Connection nodes are deployed on the 20,000 user OVA/VM.
- Edge
  - Expressway-C and Expressway-E nodes are deployed on the Medium OVA template.
- Meetings
  - Cisco Meeting Server (CMS) is deployed on the CMS 2000 platform.
  - Cisco Meeting Management (CMM) nodes are deployed on the Large OVA/VM.
  - TelePresence Management Suite (TMS) nodes are deployed on the Large OVA/VM.

Figure 1-6 show a sample layout of the collaboration application virtual machines required for the large deployment example.

Figure 1-6 Virtual Machine Placement for Large Simplified Sizing Example