Collaboration Solution Sizing Guidance

Revised: March 1, 2018

This chapter describes system sizing for Cisco Collaboration products and systems. Sizing involves providing an accurate estimate of the required hardware platforms for the system, based on the number of users, traffic mix, 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 call volumes and throughput. For standalone products, manual calculation of the system size may be feasible (as covered in the section on Sizing for Standalone Products, page 25-49). However, there are many sizing factors to consider in a complex system deployment. For example, multiple products may be distributed across different locations and may include video endpoints, call centers, and voice/video conferencing. Cisco Systems provides a set of sizing rules to handle the resulting complexity.

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

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

This chapter includes the following major sections:

Note For simplified sizing guidance without the use of the Collaboration Sizing Tool, refer to the latest version of the *Cisco Preferred Architecture for Enterprise Collaboration CVD*, available at https://www.cisco.com/go/pa.

What's New in This Chapter

Table 25-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 25-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
Sizing for Cisco Jabber clients	Cisco Jabber Clients, page 25-18	March 1, 2018
Sizing for centralized IM and Presence clusters	Centralized IM and Presence, page 25-35	March 1, 2018

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, page 25-2
- System Modeling, page 25-3
- Traffic Engineering, page 25-5

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.

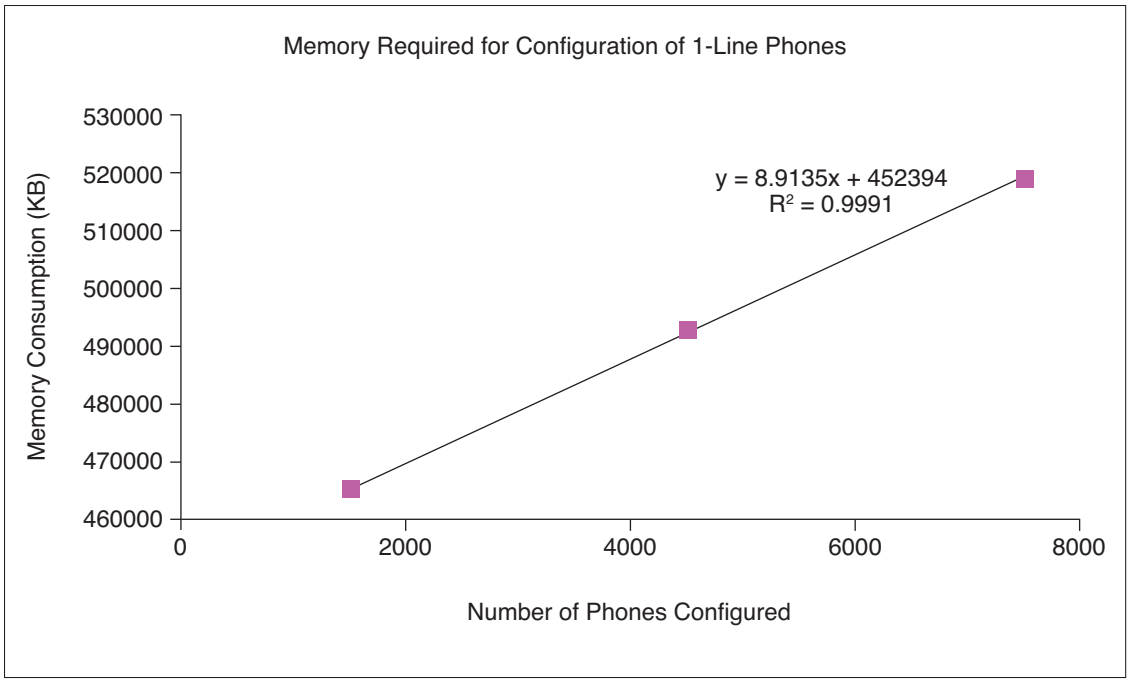
Most systems exhibit linearity within a certain range, beyond which the system performance can become unpredictable. Cisco sets the usage levels for each performance test to identify and confirm the linear range of the resource usage for each function. The results for each test can be graphed using a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

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 25-1 shows the results of a test conducted to determine the memory requirements for configuring single-line IP phones. It shows the memory consumed by configuring 1500, 4500, and 7500 single-line IP phones in Unified CM. The graph shows that the equation of the trend line is linear and can be used to predict the dependent variable (in this case, memory) based on the control variable (the number of phones).

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

Figure 25-1 Memory Required for Configuration of One-Line Phones



System Modeling

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

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

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

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

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

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

Memory Usage Analysis

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

Static memory

Static memory is consumed even when there is 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 saving the context of each active call. In a large system, dynamic memory is primarily a function of the number of concurrent calls.

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

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

System memory

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

CPU Usage Analysis

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

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

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

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

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

Traffic Engineering

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

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

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

Definitions

Traffic engineering defines the following terms:

Maximum Simultaneous Calls

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

Calls per Second

The number of new call attempts that arrive at the system in one second, 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 60.)

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

Busy Hour

The hour in a given 24-hour period during which the maximum total traffic occurs. This hour varies depending on the organization and the type of traffic. For business voice traffic, the busy hour is 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. Typically, BHCA will be calculated as the average of the busy hour call attempts from the busiest 30 days of the year). System BHCA is the User BHCA multiplied by the number of users.

Blocking Factor

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

Average Call Hold Time

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

Erlang

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

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

Erlang B Model

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

Erlang C Model

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

Bursty Traffic

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

Voice Traffic

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

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

Table 25-2 illustrates the relationship between number of trunks, blocking probability, and Erlangs of traffic.

Table 25-2 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 25-2 we can determine the following information:

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

Contact Center Traffic

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

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

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

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

As noted earlier, contact center sizing uses the Erlang C model to account for calls waiting in queues. Contact centers require additional resources, such as interactive voice response (IVR) ports. The time that calls wait in queues needs to be factored in when sizing the PSTN gateways (see Gateway Sizing for Contact Center Traffic, page 25-39).

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*, available at https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-enterprise/products-implementation-design-guides-list.html.

Video Traffic

Point-to-point video traffic demonstrates similar characteristics to its voice equivalents for call arrival rates, peak usage times, and call durations. Also, 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 5 to 10 minutes of the meeting start time, and most conference calls are scheduled to start at the beginning of the hour. Therefore, the call arrival rate will exhibit a single burst at the top of the hour rather than a Poisson distribution throughout the hour.

- Peaks

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

- Call durations

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

- Video conferencing

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

Sizing a deployment for conferencing primarily involves deciding how many concurrent connections are required. For example, sizing for TelePresence Servers would include the following considerations:

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

Conference resources are generally dedicated to a region in order 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

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 WebEx Meetings 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.

- Server 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.

- 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 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/landing

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

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

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

Cisco provides the following sizing tools:

- Cisco Collaboration Sizing Tool

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

  – Cisco Unified Communications Manager (Unified CM)
  – IM and Presence services
  – Voice messaging
  – Conferencing
  – Gateways
  – Cisco Unified Communications Management Suite
  – Cisco Unified Contact Center components

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

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

- Cisco VXI Sizing and Configuration Tool

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

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

https://cucst.cloudapps.cisco.com/help/UC_Sizing_Tools_FAQ.pdf

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

In addition to these sizing tools, a Virtual Machine Placement Tool is available to Cisco partners and customers with a valid login account. The Virtual Machine Placement Tool is a graphical tool that allows you to select Tested Reference Configurations (TRC) or specifications-based hardware, and to drag and drop the various Cisco Collaboration application virtual machines on those servers. Some placeholders representing third-party application virtual machines are also available when deploying Cisco Collaboration applications co-resident with third-party applications. The sizing tools determine how large the servers need to be and how many virtual machines are necessary. This information can then be entered as an input to this Virtual Machine Placement Tool in order to determine how to place the various virtual machines and to determine how many servers would need to be deployed. Even though some of the co-residency rules are implemented in the tool, Cisco recommends verifying the rules by using the guidelines documented at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html

The Virtual Machine Placement Tool is available (with proper login authentication) at

https://www.cisco.com/go/vmpt

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

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

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

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

Using the VXI Sizing Tool

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

Using the Cisco 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

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:

- Applications and CTI, page 25-23
- Media Resources, page 25-28

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 40,000 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 40,000 CTI connections.
- The number of call processing subscribers in a cluster cannot exceed 4, plus 4 standby, for a total of 8 call processing 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.
- The name of a Unified CM virtual machine (VM) configuration corresponds to the maximum number of users, assuming that on average, each user has one phone. If this is not the case, the VM configurations would indicate the maximum number of endpoints registered to a Unified CM node. For example, a 10k-user VM configuration supports a maximum of 10,000 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 10k-user VM configuration would support a maximum of 5,000 users with 10,000 devices. This same principal applies for the smaller Unified CM VM configurations 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 complex if you have configured more than a few thousand endpoints and more than a few hundred dial plan elements such as translation and route patterns, hunt pilots, and shared lines.

Number of Regions and Locations

Configuration of regions and locations in the Unified CM cluster requires both database and static memory. The number of gateways that can be defined in the cluster is also tied to the number of locations that can be defined. Table 25-3 lists these limits for some of the Unified CM VM configurations.

Table 25-3 Maximum Number of Regions, Locations, Gateways, and Trunks

VM Configuration	Maximum Number of Regions	Maximum Number of Locations	Maximum Number of Trunks and Gateways
1,000 or 2,500 Users	1,000	1,000	1,100
7,500 or 10,000 Users	2,000	2,000	2,100

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

Call Detail and Call Management Records

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

High Availability

After you determine the minimum number of nodes required for the specified deployment, add the desired number of additional subscriber nodes to provide redundancy. Redundancy options are described in the chapter on Call Processing, page 9-1.

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 Open Virtualization Archive (OVA) VM configurations that can be loaded onto a hypervisor. Different templates specify different capacities. For example, the 10,000 Users template defines a virtual machine that has a maximum capacity of 10,000 endpoints. There are also templates defined to support a maximum of 1,000, 2,500, and 7,500 endpoints.

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

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-sizing.html

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

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-unified-communications-manager.html

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:

http://www.cisco.com/go/virtualized-collaboration

Hardware and Virtualization Software Requirements

The following requirements are common to all applications. See each application's product documentation for additional requirements or restrictions.

- Details on supported and required virtualization hardware are available at:

  https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/collaboration-virtualization-hardware.html

- Details on supported and required virtualization software are available at:

  https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-software-requirements.html

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

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 gateways (H.323 or MGCP)

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

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

Table 25-4 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 25-4 Maximum Number of Endpoints Supported Per VM Configuration

VM Configuration	Maximum Endpoints per OVA Template[1]
10,000 Users	10,000
7,500 Users	7,500
2,500 Users	2,500
1,000 Users	1,000

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.

For Cisco Collaboration System Release (CSR) 12.*x*, the Unified CM deployments require all virtual nodes to increase their vRAM by 2 GB of memory for the following VM configuration templates:

- 1,000 users
  - 2 vCPU
  - 6 GB vRAM
  - 80 GB vDisk

- 2,500 users
  - 4 vCPU
  - 6 GB vRAM
  - 80 GB vDisk
- 7,500 users
  - 2 vCPU
  - 8 GB vRAM
  - 110 GB vDisk
- 10,000 users
  - 4 vCPU
  - 8 GB vRAM
  - 110 GB vDisk

For more details, refer to the documentation at:

http://www.cisco.com/go/virtualized-collaboration

Cisco Collaboration Clients and Applications

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

- Cisco Jabber Clients, page 25-18
- Cisco WebEx Connect, page 25-20
- Cisco UC Integration™ for Microsoft Lync, page 25-21
- Third-Party XMPP Clients and Applications, page 25-21

Cisco Jabber Desktop Client

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

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

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

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

- Device Configuration

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

  The Jabber Desktop Client uses the Cisco 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 Jabber Desktop Client in softphone mode uses the CCMCIP or UDS service to discover its device name for registration with Unified CM.

- Deskphone Mode

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

- Voicemail

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

- Authentication

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

- Contact Search

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

Cisco Jabber Clients

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

- Client scalability

  The Cisco IM and Presence Service VM configuration template determines the number of users a cluster can support. The Cisco Jabber Client deployment must balance all users equally across all nodes in the cluster. This can be done automatically by setting the User Assignment Mode Sync Agent service parameter to **balanced**.

- IMAP scalability

  The number of IMAP or IMAP-Idle connections is determined by the 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. For additional information, refer to the chapter on Cisco Rich Media Conferencing, page 11-1.

Cisco Jabber Clients are supported on iPhone, iPad, and Android as mobile clients and on Windows and Mac as desktop clients. When sizing your deployment with Jabber Clients, keep in mind that users may have any combination of desktop and mobile clients. If the Multiple Device Messaging (MDM) feature is enabled for users, then each client that is associates to a user counts as a device and thus counts toward the total number of users supported in both the Unified CM and IM and Presence VM templates.

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

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

- CTI scalability

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

- Call admission control

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

- Codec selection

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

- Cisco Unity Connection

  See the section on Managing Bandwidth, page 19-32, in the chapter on Cisco Voice Messaging, page 19-1.

- Cisco Jabber Clients are supported on iPhone, iPad, and Droid as mobile clients and on Windows PC and Mac as desktop clients.

  When sizing your deployment with Jabber Clients, keep in mind that users may have desktop and mobile clients, or multiple mobile clients or desktop clients. Users of the Multi Device Messaging (MDM) are more likely to request this feature. If it is enabled, then each client that is associates to a user counts as a device and hence counts against the total number of users supported by both the Unified CM and IM and Presence VM templates. If users have only Jabber desktop clients in desktop control mode, then they will count as only a single device due to the fact that the desk phone control utilizes CTI resources.

- Cisco WebEx Meetings Server

  Cisco WebEx Meetings Server provides WebEx conferencing services with voice, video, and collaboration sessions in a virtualized environment. For additional information about Cisco WebEx Meetings Server, refer to the *Cisco WebEx Meetings Server Planning Guide and System Requirements*, available at

  https://www.cisco.com/c/en/us/support/conferencing/webex-meetings-server/products-installation-and-configuration-guides-list.html

- 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 Series devices for contact source lookups. Using the UDS contact source to resolve contacts puts additional load on the system.

SAML SSO Cisco Jabber Client

Cisco Unified CM 10.*x* 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, Cisco Unity Connection, and 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, Cisco Unity Connection, and IM and Presence), all SAML requests must authenticate with the Identity Provider (IdP) for Cisco Jabber clients to login successfully.

Note SSO is supported by Unified Communications services with SAML.

Cisco Jabber 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 Jabber 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 Cisco Jabber 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 soft phone). 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 10k-user VM configurations
- One pair of IM and Presence 5k-user VM configurations

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

Cisco WebEx Connect

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

For additional information on network and desktop requirements, refer to the latest version of the *Cisco WebEx Messenger Administration Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/webex-messenger/products-installation-guides-list.html

The Cisco Unified Communications integrations use Unified CM CTI Manager for click-to-call applications, as well as deskphone control mode with the Cisco Unified Client Services Framework. Therefore, observe the CTI limits as defined in the section on Applications and CTI, page 25-23. When Cisco UC Integration™ for Connect is operating in a softphone (audio on computer) mode, the Cisco Jabber Desktop Client is a SIP registered endpoint with Cisco Unified CM. When sizing a solution involving Cisco Unified Communications, you must include the CTI devices and the SIP endpoint devices utilizing resources on the Unified CM clusters.

Cisco UC Integration™ for Microsoft Lync

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

Third-Party XMPP Clients and Applications

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

Mobile Unified Communications

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

Cisco Unified Mobility

There are two parameters that are key to Unified CM's capacity to support single number reach (formerly Mobile Connect) and enterprise two-stage dialing (Mobile Voice Access and Enterprise Feature Access). For these functions to work appropriately, users must be enabled for mobility and remote destinations with shared lines must be defined for the users. Table 25-5 shows the limits for users and remote destinations and mobility identities in a cluster consisting of each class of Unified CM VM configurations.

Table 25-5 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
10,000 Users VM configuration	40,000	40,000 (or 10,000 per node)
7,500 Users VM configuration	30,000	30,000 (or 7,500 per node)
2,500 Users VM configuration	10,000	10,000 (or 2,500 per node)
1,000 Users VM configuration	4,000	4,000 (or 1,000 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 users primary device, and hence calls to that line use more CPU resources.
- If the remote destination or mobility identity is an external number (such as the user's cell phone or home), then gateway resources will be used to extend the call.

Call Traffic

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

It is important to differentiate between call types because call origination and termination are considered as distinct events in the half-call model. For endpoints registered on the same 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
- Shared directory numbers and the average number of endpoints that share the same DN
- Partitions, calling search spaces, translations, and transformation patterns
- Route patterns, route lists, and route groups
- Advertised and learned DN patterns
- Hunt pilots and hunt lists
- Circular, sequential, and broadcast line groups and their membership

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

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

Applications and CTI

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

Although the 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 25-6 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 25-6 CTI Resource Limits in Unified CM

VM Configuration	Maximum CTI Resources per Virtual Machine
1,000 Users	1,000
2,500 Users	2,500
7,500 Users	5,000 or 7,500[1]
10,000 Users	10,000

1. 7,500 CTI resources supported with Unified CM 10.5 and later releases; 5,000 CTI resources supported with Unified CM releases prior to 10.5.

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 25-7.

Table 25-7 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 25-8.

Table 25-8 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 25-7 and Table 25-8, 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 25-7 and Table 25-8, 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 25-7 and Table 25-8, 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 25-9 shows the maximum number of EM and EMCC logins per minute for each type of VM configuration.

Table 25-9 EM and EMCC Rates Per VM Configuration

VM Configuration	Maximum EM Login Rate (per Node)	Maximum EM Login Rate (Dual Nodes)	Maximum EMCC Login Rate (Per Node)	Maximum EMCC Login Rate (Dual Nodes)	Maximum Concurrent EMCC Devices
1,000 Users	200	300	60	70	333
2,500 Users	235	352	71	80	833
7,500 or 10,000 Users	250	375	75	90	2,500

Cisco Extension Mobility login and logout functionality can be distributed across a pair of subscriber nodes to increase login/logout cluster capacity. For example, when the EM load is distributed evenly between two virtual machines with the 7,500-user VM configuration, the maximum cluster-wide capacity is 375 sequential logins and/or logouts per minute.

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

Note Enabling EM Security does not diminish performance.

The EMCC login/logout process requires more processing resources than intracluster EM login/logout, therefore the maximum supported login/logout rates are lower for EMCC. In the absence of any intracluster EM logins/logouts, Unified CM supports a maximum rate of 75 EMCC logins/logouts per minute with a virtual machine using the 7,500-user or 10,000-user VM configuration. Most deployments will have a combination of intracluster and intercluster logins/logouts occurring. For this more common scenario, the mix of EMCC logins/logouts (whether acting as home cluster or visiting cluster) should be modeled for 40 per minute, while the intracluster EM logins should be modeled for 185 logins/logouts when using a single EM server node. The intracluster EM login rate can be increased to 280 logins/logouts per minute when using the 7,500-user or 10,000-user VM configuration in a dual EM node configuration. (See Table 25-9.)

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

Cisco Unified CM Assistant

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

The following limits apply to Unified CM Assistant:

- A maximum of 10 Assistants can be configured per Manager.
- A maximum of 33 Managers can be configured for a single Assistant (if each Manager has one Unified CM Assistant-controlled line).
- A maximum of 3,500 Assistants and 3,500 Managers (7,000 total users) can be configured per cluster using the 7,500-user or 10,000-user 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. (For more information, see Unified CM Assistant, page 18-19.)

Cisco WebDialer

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

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

- Each WebDialer service can handle up to 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 25-10 lists the maximum values that may be configured for each of these services.

Table 25-10 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 25-10:

- 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 or to the chapter on Media Resources, page 7-1.

Music on Hold

Table 25-11 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 25-11 Music on Hold Maximum Per-Node Stream Capacity

Unified CM OVA Template	Unified CM 10.5(2) and Later		Unified CM 10.5(1) and Earlier	
	Co-resident MoH Streams (non-sRTP)[1]	Standalone MoH Streams	Co-resident MoH Streams	Standalone MoH Streams
1,000 User	500	750	500	500
2,500 User			1,000	1,000
7,500 User	750	1,000	1,000	1,000
10,000 User				

1. All capacities based on non-sRTP streams.

As shown in Table 25-12, beginning with Unified CM 10.5(2) 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 25-12 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 25-12 Music on Hold Maximum Per-Cluster Audio Source Capacity

Unified CM OVA Template	Unified CM 10.5(2) and Later		Unified CM 10.5(1) and Earlier	
	Co-resident MoH Sources	Standalone MoH Sources	Co-resident MoH Sources	Standalone MoH Sources
1,000 User	100	250	50	
2,500 User				
7,500 User	250	500		
10,000 User				

The capacity limits described in Table 25-11 and Table 25-12apply 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 (1,000-user or 2,500-user 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 1,000-user or 2,500-user 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 7,500-user 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 25-30, 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.

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 160,000. To optimize directory synchronization performance, Cisco recommends considering the following points:

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

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

Cisco Unified CM Megacluster Deployment

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

- Size of the configuration database
- Number of locations and regions
- Maximum number of LDAP synchronized or provisioned end users (160,000 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)
- Presence-related activities performed by typical users
  - Contact list size and composition (intracluster, intercluster, and federated). The Cisco IM and Presence system architecture is based on an average contact list size of 75 contacts per user on a fully populated system. While per-user contact list size will vary across the system, if significant numbers of users on the system exceed the average list size of 75 contacts, system performance will be impacted. 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 call related and user initiated)
- Deployment model
  - 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 25-13.

Table 25-13 Maximum Number of Users Supported per IM and Presence Cluster[1]

VM Configuration	Maximum Users Supported in Full Unified Communications Mode
500 Users	1,500
1,000 Users	1,000
2,000 Users	6,000
5,000 Users	15,000
15,000 Users	45,000
25,000 Users	75,000

1. Maximum supported sub-clusters is 3.

Roster Management

The number of contacts and watchers a user has, does impact the system performance. Due to the potential severity of the impact, the system administrator must monitor the usage to ensure that the cluster average per user does not exceed 75 contacts and/or watchers.

By default the service parameters are set to a maximum of 200 contacts and 200 watchers per user. The intent of this default parameter setting is to provide options for users who require a higher number of contacts. This does not imply that all 15,000 presence users on an IM and Presence node may each have 200 contacts and watchers.

We recommend that all IM and Presence deployments do not exceed a cluster average of 75 contacts and/or watchers per user, even though the service parameter is set to 200 for both.

For example, assume that we have the 15,000 Users VM configuration template in a fully loaded cluster with 3 sub-clusters and 45,000 presence-enabled users. If we want to maintain an average of 75 contacts for every user in the cluster, then the maximum number of contacts allowed for the entire cluster would be:

(45,000 users) * (75 contacts/user) = 3.375M contacts allowed for the IM and Presence cluster

Some users in this cluster may have up to 200 contacts while other users have fewer contacts, as long as the total number of contacts for all users in the cluster does not exceed 3.375M.

As another example, assume that we have a deployment of 5,000 IM and Presence users, and 50 of those users need 1,000 contacts each. The maximum number of contacts allowed for this deployment would be:

(5,000 users) * (75 contacts/user) = 375,000 contacts allowed for the entire deployment

The 50 heavy users would need: (50 users) * (1,000 contacts/user) = 50,000 contacts. That would leave (375,000 – 50,000) = 325,000 contacts available for the remaining 4,950 users, or:

325,000/4,950 = approximately 65 contacts on average for each of the other 4,950 users

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*, available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-device-support-tables-list.html

The formal definitions of the VM configurations for Cisco IM and Presence are available at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-ucm-im-presence.html

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 5k-user IM and Presence VM configuration used with a Unified CM 10k-user 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.

Note Prior to IM and Presence release 11.5, concurrent user logins were limited to a maximum of 80% of the IM and Presence VM template capacity. With IM and Presence 11.5 and later releases, 100% of the presence users can log in through Jabber at the same time. For example, in a deployment of 45,000 presence-enabled users, IM and Presence releases prior to 11.5 support only 36,000 (80% of 45,000) concurrent logins, while IM and Presence 11.5 and later releases support all 45,000 users logged in at the same time (assuming only one Jabber client per user login). This enhancement also increases the allowed number of concurrent Jabber users by 20%.

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 25k-user IM and Presence VM template for all the IM and Presence nodes in the cluster and using the 10k-user 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 25-14 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 25-14 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
12,000 Users	12,000	2,500	1,200	500	30,000	3,000
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 are available at the following location:

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-emergency-responder.html

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*, available at

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

Cisco Expressway

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 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 25-15 lists the Cisco Expressway proxy registrations and call capacities for Cisco Expressway-C and Expressway-E server node pairs and clusters.

Table 25-15 Cisco Expressway-C and Expressway-E Node and Cluster Capacities

Platform	Proxy Registrations[1]	Video Calls	Audio-only Calls
Large OVA (or Expressway Appliance)	2,500 per node 10,000 per cluster	500 per node 2,000 per cluster	1,000 per node 4,000 per cluster
Medium OVA (or Expressway Appliance)	2,500 per node 10,000 per cluster	100 per node 400 per cluster	200 per node 800 per cluster
Small OVA (Business Edition 6000)	2,500 per node 2,500 per cluster[2]	100 per node 100 per cluster[2]	200 per node 200 per cluster[2]

1. Proxy registration applies only to mobile and remote access 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 large OVA template is supported only with limited hardware. Refer to the documentation at https://www.cisco.com/go/virtualized-collaboration 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 available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/tsd-products-support-series-home.html

Gateways

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

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

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

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

Gateway Groups

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

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

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

- Groups of gateways that share a common group profile. The common profiles will depend on the complexity of the deployment.

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

PSTN Traffic

PSTN circuits are shared by all users of the system, and there are usually many more users than PSTN circuits. The number of circuits required is estimated by using the traffic management principles described in the section on call traffic (Call Traffic, page 25-22).

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

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

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

Gateway Sizing for Contact Center Traffic

Short call durations as well as bursty call arrival rates impact the PSTN gateway's ability to process the traffic. Under these circumstances the gateway needs more resources to process all calls in a timely manner, compared to calls of longer duration that are presented more uniformly over time. Because gateways have varying capabilities to deal with these traffic patterns, careful consideration should be given to selecting the appropriate gateway for the environment in which it will operate. Some gateways support more T1/E1 ports than others, and some are more able than others to deal with multiple calls arriving at the same time.

For a traffic pattern with multiple calls arriving in close proximity to each other (that is, high or bursty call arrival rates), a gateway with a suitable rating of calls per second (cps) is the best fit. Under these conditions, for example, the Cisco 3945 Integrated Services Router can maintain 28 cps with 420 calls active at once.

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

These numbers assume that all of the following conditions apply:

- CPU utilization does not exceed 75%.
- PSTN gateway calls are made with ISDN PRI trunks using H.323.
- The Real Time Control Protocol (RTCP) timer is set to the default value of 5 seconds.
- Voice Activity Detection (VAD) is off.

- G.711 uses 20 ms packetization.
- Cisco IOS Release 15.0.1M is used.
- Dedicated voice gateway configurations are used, with Ethernet (or Gigabit Ethernet) egress and no QoS features. (Using QoS-enabled egress interfaces or non-Ethernet egress interfaces, or both, will consume additional CPU resources.)
- No supplementary call features or services are enabled – such as general security (for example, access control lists or firewalls), voice-specific security (TLS, IPSec and/or SRTP), AAA lookups, gatekeeper-assisted call setups, VoiceXML or TCL-enabled call flows, call admission control (RSVP), and SNMP polling/logging. Such extra call features use additional CPU resources.

Voice Activity Detection (VAD)

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

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

Codec

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

Performance Overload

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

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

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

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

Performance Tuning

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

Lowering the CPU utilization can help to increase the performance of a Cisco IOS Voice Gateway by ensuring that there are enough available CPU resources to process the real-time voice packets and the call setup instructions. Table 25-16 describes some of the techniques for decreasing CPU utilization.

Table 25-16 Techniques for Reducing Gateway CPU Utilization

Technique	CPU Savings	Description
Enable Voice Activity Detection (VAD)	Up to 20%	Enabling VAD can result in up to 45% fewer voice packets in typical conversations. The difficultly is that, in scenarios where voice recognition is used or there are long delays, a reduction in voice quality can occur. Voice appears to "pop" in at the beginning and "pop" out at the end of talk spurts.
Disable Real Time Control Protocol (RTCP)	Up to 5%	Disabling RTCP results in less out-of-band information being sent between the originating and terminating gateways. This results in lower quality of statistics displayed on the paired gateway. This can also result in the terminating gateway having a call "hang" for a longer period of time if RTCP packets are being used to determine if a call is no longer active.
Disable other non-essential functions such as: Authentication, Authorization, and Accounting (AAA); Simple Network Management Protocol (SNMP); and logging	Up to 2%	Any of these processes, when not required, can be disabled and will result in lower CPU utilization by freeing up the CPU to provide faster processing of real-time traffic.
Change the call pattern to increase the length of the call (and reduce the number of calls per second)	Varies	This can be done by a variety of techniques such as including a long(er) introduction prompt played at the beginning of a call or adjusting the call script at the call center.

Additional Information

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

- Cisco Voice Gateway Solutions:

  https://www.cisco.com/c/en/us/products/unified-communications/communications-gateways/index.html

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

    https://www.cisco.com/c/en/us/products/routers/3900-series-integrated-services-routers-isr/relevant-interfaces-and-modules.html

  - Cisco 2900 Series Integrated Services Routers

    https://www.cisco.com/c/en/us/products/routers/2900-series-integrated-services-routers-isr/relevant-interfaces-and-modules.html

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

  https://www.cisco.com/c/dam/en/us/products/collateral/routers/2800-series-integrated-services-routers-isr/product_data_sheet0900aecd8057f2e0.pdf

- SIP gateway RFC compliance:

  https://www.cisco.com/c/en/us/products/collateral/unified-communications/ios-gateways-session-initiation-protocol-sip/product_data_sheet0900aecd804110a2.html

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

  https://www.cisco.com/c/en/us/products/collateral/unified-communications/vg-series-gateways/product_data_sheet09186a00801d87f6.html

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

  https://www.cisco.com/c/dam/en/us/products/collateral/routers/2800-series-integrated-services-routers-isr/product_data_sheet0900aecd8057f2e0.pdf

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 25-17 shows the applicability of the various voice messaging solutions to the scalability requirements of the deployment.

Table 25-17 Scaling Voice Messaging Solutions

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
	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 25-18 shows the maximum limits of various functions of different VM configurations running Cisco Unity Connection.

Table 25-18 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
100 Users	8	8	8	100
500 Users	16	16	16	500
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

The formal definitions of the VM configurations for Cisco Unity Connection are available at

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-unity-connection.html

Impact on Unified CM

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

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

Collaborative Conferencing

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

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

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

Sizing Guidelines for Audio Conferencing

Cisco recommends the following methods for calculating audio conferencing capacity:

- Calculation based on average monthly usage

  If you know the average voice conferencing usage (average minutes per month), use Table 25-19 to calculate the audio conferencing capacity.

Table 25-19 Audio Conferencing Capacity Based on Average Monthly Usage

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

- Calculation based on number of users

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

- Calculation based on actual peak usage

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

Factors Affecting System Sizing

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

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

  (Estimated meetings per day) * (Estimated users) > 80% of baseline
- If the largest single meeting exceeds 20% of the estimated capacity, increase the estimate accordingly.
- If there are continuous meetings with dedicated ports, then you must add those additional ports ((Meetings) * (Dedicated callers)) to the baseline.

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

Sizing Guidelines for Video Conferencing

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

- Calculation based on number of knowledgeable workers

  Cisco recommends provisioning a video user license for every 40 knowledgeable workers.
- Calculation based on number of voice conferencing user licenses

  Cisco recommends provisioning video conferencing capacity in the range of 17% to 25% of existing audio user licenses. The percentage depends on business requirements regarding video conferencing and on the size of the conferencing system.
- Calculation based on existing video Multipoint Control Unit (MCU)

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

Impact on Unified CM

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

Cisco WebEx Meetings Server

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

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

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

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

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

For additional information about Cisco WebEx Meetings Server, refer to latest version of the *Cisco WebEx Meetings Server Planning Guide and System Requirements*, available at

https://www.cisco.com/c/en/us/support/conferencing/webex-meetings-server/products-installation-and-configuration-guides-list.html

Sizing Factors

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

- Number of knowledge users

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

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

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

- User minutes per month

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

- Actual peak usage

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

- Video

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

- Traffic mix

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

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

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

- Access by external users

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

- Disaster recovery

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

- High availability

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

System Capacities

Cisco WebEx Meetings Server is offered in four system sizes, as listed in Table 25-20. System size is expressed as the maximum number of concurrent users of the system. Maximum concurrent users defines the maximum number of users who can participate in conference calls at any given time.

Table 25-20 VM Configurations and Capacities for Cisco WebEx Meeting Server

Maximum	50 Concurrent Users	250 Concurrent Users	800 Concurrent Users	2,000 Concurrent Users
Audio and web users (combined)	50	250	800	2,000
Video and video sharing (combined)	25	125	400	1,000
Participants in a single meeting	50	100	100	100
Playback recordings of meetings that have ended	12	63	200	500
Recordings of meetings in progress	3	13	40	100
Number of conferences (average of 2 participants per meeting)	25	125	400	1,000
Calls per second	1	3	8	20

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

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

Recordings

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

Network Bandwidth

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

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

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

Cisco Prime Collaboration Management Tools

Cisco Prime Collaboration offers a set of integrated tools to test, deploy, and monitor Cisco Unified Communications and TelePresence systems. Cisco Prime Collaboration includes the following products: Prime Collaboration Provisioning, Prime Collaboration Assurance, and Prime Collaboration Analytics.

These applications run on virtual machines. Cisco Prime Collaboration Provisioning runs on its own virtual machine, while Cisco Prime Collaboration Assurance and Cisco Prime Analytics run on the same virtual machine. Virtual machine sizing for these applications is relatively simple and depends directly on the number of endpoints or network devices that they are expected to manage.

Cisco Prime Collaboration Provisioning

Cisco Prime Collaboration Provisioning can support up to 150,000 endpoints and is implemented either on a single machine (for up to 10,000 endpoints) or on two machines (over 10,000 endpoints).

Virtual machine resources required for various levels of performance are described in the latest version of the Cisco Prime Collaboration install and upgrade guides, available at

https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-installation-guides-list.html

Cisco Prime Collaboration Assurance

Cisco Prime Collaboration Assurance can manage phones and other network devices such as routers and switches. It operates in a single machine configuration and supports up to 150,000 phones.

Virtual machine resources required for various levels of performance are described in the latest version of the *Cisco Prime Collaboration Quick Start Guide*, available at

https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-installation-guides-list.html

Cisco Prime Collaboration Analytics

Cisco Prime Collaboration Analytics runs on the same virtual machine as Cisco Prime Collaboration Assurance and works with Cisco Network Analysis Modules (NAMs) to measure voice quality.

Hardware resources required for various levels of performance are described in the latest version of the *Cisco Prime Collaboration Data Sheet*, available at

https://www.cisco.com/c/en/us/products/cloud-systems-management/prime-collaboration/datasheet-listing.html

Sizing for Standalone Products

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

- Cisco Unified Communications Manager Express, page 25-49
- Cisco Business Edition, page 25-49

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 available at

https://www.cisco.com/c/en/us/products/unified-communications/unified-communications-manager-express/datasheet-listing.html

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.

The Cisco Business Edition 4000 (BE4000) is the newest addition to the Business Edition Family. The BE4000 is powered by Cisco Unified Communication Manager Express and provides call processing services for small to medium single-site deployments and deployments in which local call processing at a remote site is needed.

The BE4000 is a dedicated cloud-managed platform that provides audio telephony and voicemail service for up to 200 audio telephony devices, with each device licensed for telephony and a voicemail port.

The BE4000 supports a maximum of 200 users with the following:

- Cisco IP Phone 7800 Series and 8800 Series SIP endpoints
- Cisco Unity Express Virtual Voicemail
- Maximum of 5 busy hour call attempts (BHCA)

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

Cisco Business Edition 6000 is available in three hardware platform options:

- BE6000H — Maximum capacity of 1,000 users; 2,500 devices; and 100 contact center agents. Supports nine collaboration application options in a single virtualized server platform. Maximum of 5,000 BHCA.
- BE6000M — Maximum capacity of 1,000 users; 1,200 devices; and 100 contact center agents. Supports five collaboration application options in a single virtualized server platform. Maximum of 5,000 BHCA
- BE6000S — Maximum capacity of 150 users and 300 devices. Supports five fixed collaboration applications in a single integrated router/gateway/virtualized blade server platform. Maximum of 750 BHCA.

To learn more about Cisco Business Edition 6000 solutions, visit https://www.cisco.com/go/be6000.

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, visit https://www.cisco.com/go/be7000.

Busy Hour Call Attempts (BHCA) for Cisco Business Edition

This section use Cisco Business Edition 6000H as an example to calculate capacity, but the information in this section also applies to BE6000M as well as the smaller 750 BHCA capacity BE6000S.

As mentioned above, Business Edition 6000H 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 6000H

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 300 endpoints on a BE6000S, 1,200 endpoints on a medium-density server, or 2,500 endpoints on a high-density server, as indicated above, actual endpoint capacity depends on the total system BHCA.

A trunk device carries multiple calls to more than one endpoint. It can be any trunk or gateway device such as a SIP trunk or a gatekeeper-controlled H.323 trunk. Business Edition 6000 supports intercluster trunking as well as H.323, SIP, and MGCP trunks or gateways and analog gateways. 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. To calculate the BHCA for these SNR features, see Capacity Planning for Cisco Unified Mobility, page 21-74, and add that value to your total BHCA calculation.

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

Another aspect of capacity planning to consider for Cisco Business Edition 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, as described in the section on Mixing Hardware Platforms and Business Edition Platforms, page 9-8.

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:

- https://www.cisco.com/go/be6000
- https://www.cisco.com/c/en/us/support/unified-communications/business-edition-6000/tsd-products-support-series-home.html