

Sizing

Revised: November 30, 2020

Sizing the components of the Preferred Architecture for Enterprise Collaboration solution is an important part of the overall solution design.

For a given deployment, the goal of the sizing process is to determine:

- The type of platform to be used.
- The specifications and number of instances to be deployed for each Cisco Collaboration product.

  For the products that are deployed with virtualization, this corresponds to the selection of the virtual machine hardware specification defined in the Open Virtual Archive (OVA) template and the number of virtual machines. For the products that are not deployed with virtualization, this corresponds to the type and number of appliances or blades.

Sizing can be a complex exercise because of numerous parameters to take into considerations. In order to simplify the sizing exercise, this chapter provides some sizing examples with corresponding assumptions. We will refer to these sizing examples as *simplified sizing deployments*. If the requirements of your particular deployment are within those assumptions, then you can use the simplified sizing deployments in this document as a reference. If not, then the normal sizing calculations have to be performed as described in the *Sizing* chapter in the latest version of the *Cisco Collaboration SRND* and product documentation available at https://www.cisco.com/go/srnd.

Once the sizing is done for the products that are deployed with virtualization, determine how to place the virtual machines on Cisco Unified Computing System (UCS) servers, and consider the co-residency rules. Ultimately, this virtual machine placement process determines how many UCS servers are required for the solution.

This chapter explains sizing for all modules that are covered in this document, namely: Call Control, Conferencing, Collaboration Edge, and Voice Messaging. This chapter also covers Virtual Machine Placement and Platforms.

For products that are deployed as virtual machines, this document does not provide details on the virtual machine OVA template specification. For that information, refer to the documentation on *Cisco Collaboration Virtualization*, available at https://www.cisco.com/go/virtualized-collaboration.

What's New in This Chapter

Table 9-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

Table 9-1 New or Changed Information Since the Previous Release of This Document

New or Revised Topic	Described in:	Revision Date
IM and Presence Sizing Update	IM and Presence Sizing, page 9-4	November 30, 2020
Cisco Meeting Management Suite	Cisco Meeting Management Suite, page 9-9	January 23, 2019
Other minor changes and corrections	Various sections of this chapter	January 23, 2019
Cisco Prime License Manager has been replaced by Cisco Smart Software Manager and is no longer covered in this chapter.	For information on Cisco Smart Software Manager, see the chapter on Collaboration Management Services.	August 30, 2017
Sizing for Cisco Unified Border Element	Cisco Unified Border Element Sizing, page 9-11	August 30, 2017

Call Control

As discussed in the Call Control chapter, the Cisco Unified Communications Manager (Unified CM) and IM and Presence Service are provided through a Unified CM cluster and an IM and Presence cluster.

A Cisco Unified CM cluster consists of one publisher node, two dedicated TFTP servers, and one or multiple call processing node pairs. The number of call processing pairs depends on the size of the deployment and is discussed later in this section. The call processing nodes are deployed in pairs for 1:1 redundancy.

IM and Presence nodes are also deployed in pairs. The number of IM and Presence pairs also depends on the size of the deployment, and this will be discussed later in this section. The IM and Presence nodes are deployed in pairs for 1:1 redundancy.

Unified CM Sizing

For Unified CM, the simplified sizing guidance covers deployments with up to 10,000 users and 10,000 devices. Unified CM supports more users and more devices under different assumptions or by adding more call processing pairs, but this is outside the scope of the simplified sizing guidance provided in this chapter. Table 9-2 describes the simplified sizing deployments. The assumptions made for those deployments are documented below this table. If the number of users or endpoints in your deployment is outside of the values in Table 9-2, or if the requirements of your specific deployment fall outside of the assumptions, do not use these simplified sizing deployments, but rather perform the normal sizing procedure documented in the *Sizing* chapter in the latest version of the *Cisco Collaboration SRND* available at https://www.cisco.com/go/srnd and in the Unified CM product documentation available at

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

Table 9-2 Unified CM Simplified Sizing Deployments

Deployment Size	Unified CM Nodes to be Deployed (7.5k-User OVA Template Used for Each Unified CM Node)
Up to 5,000 users or devices	**5 nodes:** 1 Publisher, 2 TFTP, 1 call processing pair (2 call processing subscribers)
Between 5,000 and 10,000 users or devices	**7 nodes:** 1 Publisher, 2 TFTP, 2 call processing pairs (4 call processing subscribers)

Table 9-2 sizes deployments based on the maximum number of users and devices, whichever number is greater. For example, in a deployment with 5,000 users and an average of two devices per user (for example, each user has a desk phone and a Jabber client in softphone mode), the 7-node deployment is required because there are 10,000 devices in total.

The 7.5k-user virtual machine configuration (OVA template) is used in these simplified sizing deployments in order to optimize the overall resources consumed on the UCS server. This OVA template requires a full UC performance CPU platform such as the Cisco Business Edition 7000; and it is not supported on the Business Edition 6000, for example. For more information on those OVA virtual machine configuration templates and on the platform requirements, refer to the documentation at https://www.cisco.com/go/virtualized-collaboration.

A Unified CM call processing pair deployed with the 7.5k-user OVA template could support up to 7,500 users under some conditions. But in this design, we use some assumptions that put an additional load on Unified CM; for instance, we assume that each user can be configured with a Remote Destination Profile for Single Number Reach, each user can use Extension Mobility, each endpoint can be CTI controlled, some shared lines are configured, mobile and remote access is enabled, and so forth. Therefore the capacity per Unified CM call processing pair is reduced, as shown in Table 9-2. The following description provides more information on the assumptions used in this simplified sizing model.

Unified CM Assumptions

The following assumptions apply to the two simplified sizing deployments listed in Table 9-2:

- Average of up to 4 busy hour call attempts (BHCA, the number of call attempts during the busy hour) per user.
- Average of up to 2 DNs per device.
- Media and SIP signaling encryption can be enabled without changing this Unified CM simplified sizing.
- Up to 500 shared lines per call processing subscriber pair, each line being shared with an average of up to 3 devices.
- Jabber clients registering to Unified CM (softphone mode) must be counted against the device limit.
- Up to 3,000 partitions; 6,000 calling search spaces (CSSs); and 12,000 translation patterns per cluster.
- Per Unified CM cluster, up to 1,000 route patterns; 1,000 route lists; and 2,100 route groups. Per Unified CM call processing pair, up to 100 hunt pilots, 100 hunt lists, 50 circular/sequential line groups with an average of 5 members per line group, and 50 broadcast line groups with an average of 10 members per line group.
- Up to 500 CTI ports and 100 CTI route points per Unified CM call processing pair.
- GDPR/ILS is enabled when multiple Unified CM clusters are deployed.

- Extension Mobility (EM) — All users can use EM, but no Extension Mobility Cross Cluster (EMCC) users. Up to 250 EM logins/logouts per minute are supported. (This simplified sizing assumes the EM service is activated on one Unified CM node.)
- Unified CM media resources — Unified CM software conference bridges (software CFBs) and Unified CM media termination points (MTPs) should not be used in this design. Instead, use Cisco Meeting Server and Cisco IOS-based MTP, respectively.
- Average of up to one remote destination or mobility identity per mobility user. For example, in a deployment with 5,000 users, there can be up to 5,000 remote destinations or mobility identities.
- Up to 40,000 users synchronized with active directory (but only up to 5,000 or 10,000 active users would place or receive calls, depending on the simplified sizing deployment selected in Table 9-2).
- Up to 1,500 concurrent active calls (conferencing and non-conferencing sessions) per Unified CM call processing pair. For example, if all calls are conference calls and if the average number of participants in a conference is 10, then this design assumes up to 150 conference calls per Unified CM call processing pair.
- Up to 15 calls per second (cps) per Unified CM call processing pair.

Other capacity limits that are applicable to the Cisco Collaboration solution and that are documented in the latest version of the *Cisco Collaboration SRND* and product documentation, also apply. For example:

- Computer Telephony Integration (CTI) — All devices can be enabled for CTI, with up to 5 lines per device and 5 J/TAPI applications monitoring the same CTI device.
- Annunciator – 48 per Unified CM call processing pair. Music on hold (MoH) – 250 concurrent MoH sessions per call processing pair. For a larger number of annunciators or concurrent MoH sessions, deploy standalone Unified CM subscribers as MoH servers.
- Gateway – Up to 2,100 per cluster.
- Locations and regions — When adding regions, select **Use System Default** for the Audio Codec Preference List and Audio and Session Bit Rate values. Changing these values for individual regions from the default has an impact on server initialization and publisher upgrade times. Hence, with a total of 2,000 regions you can modify up to 200 regions to use non-default values. With a total of 1,000 or fewer regions, you can modify up to 500 of them to use non-default values. A maximum of 2,000 locations is supported, and they do not have usage limitations like regions do.

IM and Presence Sizing

For IM and Presence, simplified sizing guidance covers deployments of a single Unified CM cluster and IM and Presence subcluster with up to 15,000 devices. Table 9-3 describes the simplified sizing deployments. If the number of users or logged-in Jabber endpoints in your deployment is outside of the values in Table 9-3, do not use these simplified sizing deployments, but rather perform the normal sizing procedure documented in the *Sizing* chapter in the latest version of the *Cisco Collaboration SRND* and product documentation.

Table 9-3 IM and Presence Simplified Sizing Deployments

Deployment Size	IM and Presence Nodes to be Deployed
Less than 5,000 users or logged-in Jabber endpoints	One IM and Presence pair using the 5k-user OVA template
Between 5,000 and 15,000 users or logged-in Jabber endpoints	One IM and Presence pair using the 15k-user OVA template

For example, if a deployment has 5,000 users and each user on average is logged on to two Jabber endpoints concurrently, then the capacity is limited by the 10,000 logged-in Jabber endpoints, and therefore this deployment requires one IM and Presence pair using the 15k-user OVA template. The two OVA virtual machine configuration templates in Table 9-3 require a full Unified Communications performance CPU platform such as the Cisco Business Edition 7000. For more information on those OVA virtual machine configuration templates and on the platform requirements, refer to the documentation available at https://www.cisco.com/go/virtualized-collaboration.

The two IM and Presence nodes are deployed as a pair in order to provide redundancy if one of the nodes fails.

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

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

The following IM and Presence deployment types and features will require 15k-user OVA or higher OVA template:

- Centralized IM and Presence deployments (25k-user OVA recommended) - deployments with one (or more) IM and Presence cluster and multiple
- Unified CM clusters
- Multi-cluster IM and Presence deployments - deployments with two (or more) Unified CM clusters each with IM and Presence sub-clusters or with two (or more) IM and Presence clusters.
- Persistent chat
- Message archiving
- 3rd party compliance
- Multiple device messaging (MDM)
- Managed file transfer (MFT)
- Outlook integration (Jabber client)

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

SRST Sizing

The number of phones and DNs supported on a Cisco Integrated Services Router (ISR) in Survivable Remote Site Telephony (SRST) mode depends on the platform. Table 9-4 provides a capacity example. For information on other SRST platforms, including information on the required amount of DRAM and flash memory, refer to the Cisco Unified SRST documentation available at

https://www.cisco.com/c/en/us/support/unified-communications/unified-survivable-remote-site-telephony/products-device-support-tables-list.html

Table 9-4 SRST Sizing Example

Platform	Maximum Number of Phones	Maximum Number of DNs
Cisco 4451-X Integrated Service Routers	1,500	2,500

Conferencing

Sizing a deployment for conferencing is primarily an exercise in deciding how many concurrent connections are required to Cisco Meeting Server. Considerations include:

- Geographical location — Each region served by Unified CM should have dedicated conferencing resources. For example, there could be one central location for the US where Unified CM, Cisco Meeting Server, and other servers are installed, and one central location for EMEA.
- Cisco Meeting Server 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.

Conference Port Usage Guidelines

Audio and video conference sizing depends heavily on specific details about the customer, their user base, and their conferencing habits. The guidelines in this section can be used as a basis for sizing a conferencing deployment, but user-to-port ratios will vary greatly depending on the deployment environment and the requirements of the organization.

Table 9-5 shows suggested ratios to start planning conference resource requirements. These numbers vary depending on the capabilities of deployed endpoints, availability of alternative audio conferencing such as Cisco WebEx, and users' comfort level in creating and joining conferences. As a starting point, the following formulas can be used to calculate port requirements:

- Audio ports = 50 + (<number of users> / 9)
- Video ports = 8 + (<number of users> / 15)

Table 9-5 Recommended Number of Conference Ports

Number of Users	Number of Audio Ports	Number of Video Ports
1,000	161	75
1,750	244	125
3,000	383	208
5,000	605	342
10,000	1,161	675

The numbers in Table 9-5 can be used for either scheduled or non-scheduled conferencing. It is expected that, for scheduled meetings, customers can use existing usage data to draw more definite conclusions about concurrent meeting usage.

Understanding what type of meetings a customer expects to take place will help further refine the number of ports required. The total number of ports can be calculated with the formula:

Total ports = <Average number of participants in a meeting> * <Concurrent meetings>

For example, with 3,000 users, Table 9-5 suggests 208 ports. This can, for instance, correspond to an average of 3 participants per meeting and 69 concurrent meetings, or an average of 6 participants per meeting and 34 concurrent meetings. By assessing the suggested port numbers in this manner, it is easier to determine whether the total number of ports is likely to be sufficient for the deployment.

Another important point to consider is what the maximum meeting size is likely to be. In most cases the largest meeting is an all-hands meeting type. For instance, if a customer has 1,000 users but has a requirement to join 96 systems in an all-hands TelePresence conference, this would override the 75 port suggestion.

Cisco Meeting Server Platform Sizing

Cisco Meeting Server is available in several different models and platforms with differing conference support and scalability. Table 9-6 lists the recommended Cisco Meeting Server platforms for enterprise deployments, along with their associated per-node port capacities. These numbers are valid with non-encrypted and encrypted media and signaling. For more information on Cisco Meeting Server clustering, for information on other Cisco Meeting Server platforms, or for information on other video and data channel resolutions, refer to the *Cisco Meeting Server and Cisco Meeting App Data Sheet*, available at

https://www.cisco.com/c/en/us/products/conferencing/meeting-server/datasheet-listing.html

Table 9-6 Cisco Meeting Server Platforms and Capacities

Cisco Meeting Server Platform[1]	Full HD 1080p30 Port Capacity[2]	HD 720p30 Port Capacity[2]	SD 480p30 Port Capacity[2]
Cisco Meeting Server 1000	48	96	192
Cisco Meeting Server 2000	350	700	1,000

1. Cisco Meeting Servers support a maximum of 3,000 audio connections for any standalone deployment or cluster and with any audio codec.
2. Assumes content sharing at 720p resolution and 5 frames per second (fps).

There are other considerations to keep in mind too. For example, a Cisco Meeting Server supports a maximum of 450 participants in each conference per node, and this capacity can be increased by adding Cisco Meeting Server nodes.

Cisco TelePresence Management Suite (TMS)

We recommend two simplified sizing deployments for Cisco TMS, illustrated in Table 9-7. There are other possible TMS deployments, but they are not covered in this guide. For instance, the single server deployment that has all TMS, TMSXE, and Microsoft SQL components residing in the same virtual machine is not described here because it does not provide redundancy.

The two deployments in Table 9-7 provide high availability. The redundant node is deployed for resiliency, not for scalability. A load balancer providing a single virtual IP address for the primary and backup nodes is also required.

Table 9-7 Cisco TMS Simplified Deployments and Capacities

Deployment Model	Deployment	Cisco TMS	Cisco TMSXE
Regular Deployment	2 nodes total: both TMS and TMSXE running on each node Additional servers for Microsoft SQL	< 200 controlled systems (endpoints added to TMS for scheduling) < 100 concurrent participants < 50 concurrent ongoing scheduled conferences	< 50 endpoints bookable in Microsoft Exchange
Large Deployment	4 nodes total: 2 nodes with TMS and 2 nodes with TMSXE Additional servers for Microsoft SQL	< 5,000 controlled systems (endpoints added to TMS for scheduling) < 1,800 concurrent participants < 250 concurrent ongoing scheduled conferences	< 1,800 endpoints bookable in Microsoft Exchange or < 1,000 endpoints bookable in Office 365 or a combination of on-premises Exchange and Office 365

Other factors that influence Cisco TMS performance and scaling include:

- The number of users accessing the Cisco TMS web interface.
- Concurrency of scheduled or monitored conferences.
- Simultaneous usage of the Cisco TMS Booking API (TMSBA) by multiple extensions or custom clients. Booking throughput is shared by all scheduling interfaces, including the Cisco TMS New Conference page.

For more information on sizing Cisco TMS, refer to the *Cisco TelePresence Management Suite Installation and Upgrade Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-tms/products-installation-guides-list.html

Cisco Meeting Management Suite

There are two VM configurations for Cisco Meeting Management Suite, depending on the number of Cisco Meeting Server call bridges, the number of call legs started at peak time across all call bridges, and the number of users signed into Meeting Management at the same time. For more information, refer to the latest version of the *Cisco Meeting Management Installation and Configuration Guide*, available at

https://www.cisco.com/c/en/us/support/conferencing/meeting-management/products-installation-guides-list.html

Collaboration Edge

This section covers sizing of Cisco Expressway and Cisco Unified Border Element, two key components of the Collaboration Edge.

Cisco Expressway Sizing

Table 9-8 shows the maximum capacity that a single Expressway node can handle at any point of time when using the medium OVA template.

The Expressway nodes are clustered together to provide redundancy and larger scalability. The cluster configurations that are recommended and that are covered in this document consist of clusters of 2, 3, or 6 nodes. Table 9-9 shows the cluster capacity for those recommended deployments. It is important to note that all of the deployment models account for redundancy. With a cluster of 2 or 3 nodes, one node can fail without impacting the cluster capacity (N+1 redundancy). With a full cluster of 6 nodes, two nodes can fail without impacting the cluster capacity (N+2 redundancy).

In order to better understand the relationship between the cluster capacity and the level of redundancy, the following example analyses the video capacity during normal operations and after a failover, using the medium OVA template:

The maximum video call capacity per node is 100 sessions. In a 3-node cluster in a non-resilient deployment, the video call cluster capacity is 300, but it would be reduced by one-third if one node fails. In order to provide resiliency and maintain the cluster capacity if one of the three nodes fails, the recommended high-available 3-node cluster capacity is limited to 200 video sessions. During normal operations, video calls are load-balanced across the cluster, with each node handling approximately 66 video calls. If one node fails, the remaining nodes can then handle all 200 video sessions because each node can handle 100 video sessions, and therefore the cluster capacity is maintained.

Table 9-8 Expressway Node Capacity

OVA Template	Mobile and Remote Access Proxy Registrations per Node[1]	Video Calls Capacity per Node	Audio-Only Calls Capacity per Node
Virtual machine with medium OVA template	2,500	100	200

1. Proxy registration considerations apply only to mobile and remote access, not to business-to-business communications.

Table 9-9 Cisco Expressway Simplified Sizing Deployments and Associated Cluster Capacity

Deployment Model	Expressway Cluster Deployment	Redundancy Model	Mobile and Remote Access Proxy Registrations per Cluster[1]	Video Calls Capacity per Cluster	Audio-Only Calls Capacity per Cluster
Virtual machine with medium OVA template					
Deployment 1	2 nodes	N+1	2,500	100	200
Deployment 2	3 nodes	N+1	5,000	200	400
Deployment 3	6 nodes	N+2	10,000	400	800

1. Proxy registration considerations apply only to mobile and remote access, not to business-to-business communications.

Note There are two other OVA templates available, the small and the large OVA templates. The small OVA template is designed to run on the Cisco Business Edition 6000M or 6000H. The large OVA template is not supported with the Cisco Business Edition 7000, and it is supported only with limited hardware. There is also an option to use a hardware appliance, the Cisco Expressway CE1200. Refer to the documentation at https://www.cisco.com/go/virtualized-collaboration for more information.

The following assumptions are used for the Expressway simplified sizing deployments in Table 9-9:

- All video calls are encrypted. The average call rate across all the video calls is 768 kbps. For example, half of the video calls could be at 384 kbps and the other half at 1152 kbps.
- All audio calls are encrypted, and the average bandwidth across all audio calls is 64 kbps.
- For virtual machines using the medium OVA template, the call rate is up to 5 calls per second (cps) per node.

The following guidelines apply when clustering Cisco Expressway:

- Expressway clusters support up to 6 nodes (cluster capacity up to 4 times the node capacity).
- Expressway-E and Expressway-C nodes cluster separately; an Expressway-E cluster consists of Expressway-E nodes only, and an Expressway-C cluster consists of Expressway-C nodes only.
- Expressway peers should be deployed in equal numbers across Expressway-E and Expressway-C clusters. For example, a three-node Expressway-E cluster should be deployed with a three-node Expressway-C cluster.
- The capacity of all nodes across and within each Expressway-E and Expressway-C cluster pair must be the same. For example, an Expressway-E node using the small OVA template must not be deployed if the other nodes in the Expressway-E cluster or in the corresponding Expressway-C cluster are using the medium OVA template.
- 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.
- Multiple Expressway-E and Expressway-C clusters may be deployed to increase capacity.

For more information on Expressway, refer to the *Cisco Expressway Administrator Guide*, available at

https://www.cisco.com/c/en/us/support/unified-communications/expressway-series/products-maintenance-guides-list.html

Cisco Expressway Sizing Example

A company has 6,000 users, and on average 1,000 users are traveling at any given time. 80% of the mobile users require mobile and remote access at any given time. In this case, Expressway has to be sized to allow for 800 concurrent registrations (80% of 1,000).

Moreover, 10% of the mobile users are in a call at the same time. 5% of these users are calling through Expressway, while the remaining 5% are calling through the cellular network, so that the number of concurrent calls to the Expressway is 40 (5% of 800).

In the corporate network, 1% of the users are on a business-to-business calls at the same time. This accounts for an additional 50 calls (1% of (6,000 – 1,000)).

In this case we need to size the cluster to support 800 concurrent registrations and 90 concurrent calls (40+50).

Table 9-8 shows that a medium OVA template supports up to 100 concurrent calls and 2,500 concurrent registrations. We can therefore deploy an Expressway-C cluster consisting of two nodes using the medium OVA template, and an Expressway-E cluster also consisting of two nodes using the medium OVA template. Each Expressway server node can manage the whole amount of 800 registrations and 90 calls at the same time, as shown by Deployment 1 in Table 9-9. Clustering is needed because, if one of the two Expressway nodes goes down, the other node can handle the whole amount of traffic. Under normal conditions, calls and registrations are load-balanced between the two nodes of the Expressway-C and Expressway-E clusters.

After some time, the business-to-business calls in this example increase from 1% to 3%. We now need to account for 190 concurrent calls (40+150) instead of 90. The maximum that a medium OVA template can handle is 100 calls, so we need to deploy a larger cluster in this case. Table 9-9 shows that Deployment 2 can account for 200 concurrent calls even in case of a server failure. Therefore, the administrator in this example decides to add another medium OVA node to the Expressway-C and Expressway-E clusters, for a total of 3 nodes per cluster.

Cisco Unified Border Element Sizing

Cisco Unified Border Element is supported on a wide range of Cisco routing platforms, including platforms such as the Cisco 4400 Series Integrated Services Routers (ISR) and the Cisco 1000 Series Aggregation Service Routers (ASR). Cisco Unified Border Element also provides redundancy on the following platforms:

- The Cisco ISR platforms, which can provide box-to-box redundancy with both signaling and media preservation for active calls.
- The Cisco ASR platforms, which can provide box-to-box or in-box redundancy with media and signaling preservation (stateful failover) for active calls.

Table 9-10 provides capacity examples for a few platforms. This table shows the maximum number of SIP trunk sessions, which corresponds to the maximum number of end-to-end PSTN SIP-SIP calls. It provides limits without media and signaling encryption and limits with RTP/SRTP interworking, where traffic is encrypted inside the corporate network and not encrypted for the connection to the SIP service

provider. For information on other platforms and for more detailed, information including required amount of DRAM and flash memory, refer to the *Cisco Unified Border Element Data Sheet* and the *Cisco Unified Border Element and Gatekeeper Ordering Guide*, both available at

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

Table 9-10 Cisco Unified Border Element Capacity Examples

Platform	Maximum SIP Trunk Sessions with Non-Encrypted Media and Signaling	Maximum SIP Trunk Sessions with Encrypted Media and Signaling
Cisco 4451-X Integrated Service Router	6,000	1,400
Cisco 1004 and 1006 Aggregation Services Routers	16,000	5,000

Cisco Unified Border Element Sizing Example

A company has 10,000 users and has media and signaling encryption enabled in the corporate network. During the busiest hour, 10% of them are in a call at the same time. 8% of these users are calling external destinations, while the remaining users are engaged in internal calls. The Telecom carrier and the enterprise have agreed that G.711 can be used on all calls, therefore no transcoding is needed. For this deployment, 800 SIP sessions (8% of 10,000) are needed. Table 9-10 shows that a Cisco 4451-X ISR can support up to 1,400 sessions with encryption. Thus, for this example two Cisco 4451-X ISRs can be deployed, one active and one standby to provide redundancy.

Voice Messaging

This section covers sizing for Cisco Unity Connection.

As discussed in the section on the Cisco Unity Connection Deployment Process, the recommended Unity Connection deployment in this design consists of one publisher and one subscriber in active/active mode.

This guide covers three simplified sizing deployments for Unity Connection, depending on the number of users and the number of Jabber endpoints. These deployments are shown in Table 9-11. For example, if a deployment has 10,000 users and 1,000 Jabber endpoints total, then at a minimum the 10k-user OVA template has to be deployed. Or for example, if a deployment has 6,000 users and 2,000 Jabber endpoints, then at a minimum the 10k-user OVA template has be deployed. There are other possible deployments with Unity Connection, but they are not covered in this guide. Refer to the latest version of the *Cisco Collaboration SRND* and product documentation for information on the other possible deployments.

Table 9-11 Cisco Unity Connection Simplified Sizing Deployments

Deployment Size	Unity Connection Nodes to be Deployed for Active/Active
Up to 5,000 users or up to 1,000 Jabber endpoints	One Unity Connection pair using 5k-user OVA template
5,000 to 10,000 users or up to 2,000 Jabber endpoints	One Unity Connection pair using 10k-user OVA template
10,000 to 20,000 users or up to 5,000 Jabber endpoints	One Unity Connection pair using 20k-user OVA template

Cisco Unity Connection Assumptions

- High availability is implemented for all Cisco endpoints, including Jabber endpoints.
- Media and SIP signaling encryption can be enabled without changing this Unity Connection simplified sizing.
- There is a single inbox for all users (Unified Messaging).
- Notifications of voice messages (new message, message update, and message deleted) use HTTP (not HTTPS).

The OVA template limits should not be exceeded. For example, with the 5k-user OVA template, there is a limit of 200 ports with G.711 or 50 ports with G.722. For more information on the OVA template limits, refer to:

- Cisco Unity Connection virtualization information at https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-unity-connection.html
- *Cisco Unity Connection Supported Platforms List* at https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-installation-guides-list.html

It is also important to consider the amount of storage required to store voice mail. The message storage depends on the size of the virtual disk. For example, the approximate message storage using the G.711 codec is 137k minutes with the 5k-user OVA template, which is defined with one vDisk of 200 GB. Note that with the 10k-user OVA template, different vDisk sizes are available to address different message storage requirements. For more information, refer to the latest version of the *Cisco Unity Connection Supported Platforms List*.

Collaboration Management Services

This section covers sizing for the following management services used in the Enterprise Collaboration Preferred Architecture:

- Cisco Prime Collaboration Provisioning
- Cisco Prime Collaboration Deployment

Cisco Prime Collaboration Provisioning

The recommended deployment for Cisco Prime Collaboration Provisioning in this design consists of one node. There is no redundant node in this deployment. Back up your Cisco Prime Collaboration Provisioning virtual machine instead.

This guide covers two sizing deployments for Cisco Prime Collaboration Provisioning, depending on the number of endpoints. These deployments are listed in Table 9-12. There are other possible deployments for Cisco Prime Collaboration Provisioning, but they are not covered in this guide. Refer to the latest version of the *Cisco Collaboration System 11.x SRND* and the Cisco Prime Collaboration Provisioning product documentation for information on the other possible deployments.

Table 9-12 Cisco Prime Collaboration Provisioning Deployment Sizing

Deployment Size	Cisco Prime Collaboration Provisioning Node to be Deployed
Up to 3,000 devices	One node using the small (3k devices) OVA template
Up to 20,000 devices	One node using the medium (20k devices) OVA template

Cisco Prime Collaboration Deployment

Cisco Prime Collaboration Deployment is deployed as one node. There is no redundant node is this deployment. Back up your Cisco Prime Collaboration Deployment virtual machine instead. The single Cisco Prime Collaboration Deployment node can support a deployment of any size.

Virtual Machine Placement and Platforms

With Cisco Collaboration products that are deployed with virtualization, after sizing the deployment, the next step is to determine how to place the virtual machines together on the Cisco Unified Computing System (UCS) servers, which will ultimately determine how many UCS servers are required for the solution. This process is performed with the Collaboration Virtual Machine Placement Tool (VMPT), which requires a cisco.com login and which is available at https://www.cisco.com/go/vmpt.

Figure 9-1 shows an example of using VMPT for a deployment with 5,000 users and 5,000 total endpoints (including 1,000 Jabber endpoints). This example assumes that Cisco Business Edition 7000M is deployed. It does not include the Cisco Meeting Servers; we assume they are deployed on the Cisco Meeting Server 1000 platform.

Figure 9-1 Virtual Machine Placement Example Using VMPT



In general, in addition to using VMPT, it is a good practice to validate the virtual machine placement by ensuring that the deployment meets all the co-residency requirements documented at

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

The main placement and co-residency rules are:

- No over-subscription — All virtual machines require a one-to-one mapping between virtual hardware and physical hardware. For example, with the CPU there must be a one-to-one mapping between virtual hardware and physical hardware, even when hyper-threading is enabled.
- VMware latency sensitivity, available with vSphere 5.5 and later versions, should be set to **High** for the Unity Connection virtual machines. If not, one spare physical core must be reserved for the ESXi scheduler on each ESXi host where Unity Connection is installed.
- Most of the applications discussed in this guide support co-residency with third-party applications, which means they can be installed on the same UCS server. However, it is important to understand that, with co-residency of third-party applications, the third-party applications must follow the same rules as Cisco collaboration applications. For example, once a third-party application is installed on the same host as a Cisco collaboration application, CPU over-subscription is not supported with that third-party application, a physical core needs to be reserved for the ESXi scheduler when deploying Unity Connection, and so forth. With Cisco Business Edition platforms, the ESXi license also dictates some of the co-residency options. For example, with the Cisco UC Virtualization Hypervisor/Foundation, there is a limit on the number of third-party applications that can be co-resident.

Redundancy Consideration

Even though the hardware platforms can be highly redundant, it is good practice to plan for hardware redundancy. For example, do not deploy the primary and backup application virtual machines on the same UCS server, as shown in the example in Figure 9-1. Instead, deploy primary and backup virtual machines on different servers to provide redundancy in case a host fails.

Platforms

For the products that are deployed with virtualization, Cisco Business Edition 7000 can be an excellent solution. It is easy to order and easy to deploy. VMware vSphere Hypervisor (ESXi) is pre-installed. Business Edition 7000 is also pre-loaded with the Cisco Collaboration software set and some of the Cisco Collaboration applications are also pre-installed.