
# Sizing

**Revised: April 2, 2015**

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 for each Cisco Collaboration product. Most products are deployed with virtualization only, but some products such as the Cisco TelePresence Server can also be deployed as an appliance or blade, depending on the requirements.
- 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 of the *Cisco Collaboration SRND* and product documentation available at http://www.cisco.com/go/ucsrnd.

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 Core Applications. 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 *Unified Communications in a Virtualized Environment*, available at http://www.cisco.com/go/uc-virtualized.

## What's New in This Chapter

Table 6-1 lists the topics that are new in this chapter or that have changed significantly from previous releases of this document.

*Table 6-1 New or Changed Information Since the Previous Release of This Document*

| New or Revised Topic | Described in: | Revision Date |
|---|---|---|
| Multiparty Media 400v capacity limits | Table 6-7 | April 2, 2015 |
| Recommended TelePresence Server platforms and capacities | TelePresence Server Platform Sizing, page 6-7 | January 22, 2015 |

# 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 6-2 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 6-2, 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 of the *Cisco Collaboration SRND* available at http://www.cisco.com/go/ucsrnd and in the product documentation available at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/tsd-products-support-series-home.html.

*Table 6-2 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:**<br>1 Publisher, 2 TFTP, 1 call processing pair (2 call processing subscribers) |
| Between 5,000 and 10,000 users or devices | **7 nodes:**<br>1 Publisher, 2 TFTP, 2 call processing pairs (4 call processing subscribers) |

Table 6-2 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 www.cisco.com/go/uc-virtualized.

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 6-2. 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 6-2:

- 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.
- 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.
- 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 TelePresence Servers 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 6-2).

- 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
- The Contact Source for Jabber is not based on Unified CM User Data Service (UDS) in this design, but rather Basic Directory Integration (BDI) or Enhanced Directory Integration (EDI). If Unified CM UDS is configured for the Contact Source, the maximum number of users per Unified CM call processing pair is reduced to 3,750.
- Up to 2,500 concurrent mobile and remote access endpoints per Unified CM call processing pair.

Other capacity limits that are applicable to the Cisco Collaboration solution and that are documented in 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 – Up to 2,000 per cluster.
- Extension Mobility (EM) – Up to 250 EM users per Unified CM call processing node, or 375 per cluster across two active call processing nodes.

## IM and Presence Sizing

For IM and Presence, simplified sizing guidance covers deployments with up to 15,000 users. The IM and Presence Service supports more users by adding IM and Presence node pairs, but this is outside of the simplified sizing guidance provided in this chapter. Table 6-3 describes the simplified sizing deployments. Again, if the number of users in your deployment is outside of the values in Table 6-3, do not use these simplified sizing deployments, but rather perform the normal sizing procedure documented in the sizing chapter of the *Cisco Collaboration SRND* and product documentation.

***Table 6-3 IM and Presence Simplified Sizing Deployments***

| Deployment Size | IM and Presence Nodes to be Deployed |
|---|---|
| Less than 2,000 users | One IM and Presence pair using the 2k-user OVA template |
| Between 2,000 and 5,000 users | One IM and Presence pair using the 5k-user OVA template |
| Between 5,000 and 15,000 users | One IM and Presence pair using the 15k-user OVA template |

These OVA virtual machine configuration templates require a full UC 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 www.cisco.com/go/uc-virtualized.

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

## 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 6-4 provides capacity examples for only three platforms. For information on other SRST platforms, including information on the required amount of DRAM and flash memory, refer to the SRST documentation available at

http://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cusrst/requirements/guide/srs10spc.html

***Table 6-4 SRST Sizing Examples***

| Platform | Maximum Number of Phones | Maximum Number of DNs |
|---|---|---|
| Cisco 2901 Integrated Service Routers | 35 | 200 |
| Cisco 3925 Integrated Service Routers | 730 | 1,000 |
| 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 TelePresence Servers. 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, TelePresence Servers, and other servers are installed, and one central location for EMEA.
- Preference for TelePresence Server platforms — Virtualized or non-virtualized
- 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.

## 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 6-5 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 6-5 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 6-5 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 6-5 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.

## Screen Licenses and Port Capacity

Video resolution determines the quality of users' video experience and the number of video connections that a Cisco TelePresence Server can support. For optimal experiences, we recommend enabling high definition (HD) video calls at a minimum resolution of 720p and 30 frames per second (fps). Depending on the budget and capability of an organization's endpoints and network, HD video calls might not always be possible. Table 6-6 shows TelePresence Server port capacity based on video quality, assuming the video streaming rate is 30 fps. The number of audio ports per screen license is not shown and is equal to 52, with a maximum of 200 audio ports supported per TelePresence Server.

*Table 6-6 TelePresence Server Port Capacity Based on Video Quality*

| Screen Licenses[1] | 1080p Ports[2] | 720p Ports[3] | 480p Ports[3] | 360p Ports[3] |
|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 |
| 5 | 5 | 10 | 15 | 20 |
| 10 | 10 | 20 | 30 | 40 |
| 20 | 20 | 40 | 60 | 80 |
| 48 | 48 | 96 | 144 | 192 |

1. The number of screen licenses that can be deployed on a TelePresence Server depends on the platform.
2. Assumes a separate content channel sharing at a maximum of 720p resolution and 15 fps.
3. Assumes a separate content channel sharing at a maximum of 720p resolution and 5 fps.

**Note** With Cisco TelePresence Conductor and TelePresence Server, a single conference resource can host multiple simultaneous conferences with different resolution limits. There is no need to dedicate a TelePresence Server to a single resolution.

As can be seen from Table 6-6, the desired video quality has a direct effect on the amount of resources consumed on a TelePresence Server and, as a result, a direct impact on the number of TelePresence Servers required for the deployment.

## TelePresence Server Platform Sizing

Cisco TelePresence Server is available in several different models and platforms with differing conference support and scalability. Table 6-7 lists the recommended TelePresence Server platforms for enterprise deployments, along with some of their associated port capacities. For more details, for information on other TelePresence Server platforms, or for information on other video and data channel resolutions, refer to the *Cisco TelePresence Data Sheet*, available at

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

***Table 6-7 TelePresence Server Platforms and Capacities***

| TelePresence Server Platform[1] | Cluster Support | HD 1080p Port Capacity[2] | HD 720p Port Capacity[3] | SD 480p Port Capacity[3] | SD 360p Port Capacity[3] |
|---|---|---|---|---|---|
| Multiparty Media 400v | No | 18 | 36 | 54 | 72 |
| TelePresence Server MSE 8710 | Yes, up to four 8710s can be clustered. | 12 per blade.<br>Up to 48 per cluster. | 24 per blade.<br>Up to 97 per cluster. | 36 per blade.<br>Up to 146 per cluster. | 48 per blade.<br>Up to 195 per cluster. |

1. TelePresence Servers support a maximum of 200 audio connections for any standalone deployment or cluster and with any audio codec.
2. Assumes content sharing at 720p resolution and 15 frames per second (fps).
3. Assumes content sharing at 720p resolution and 5 frames per second (fps).

There are other considerations to keep in mind too. For example:

- A TelePresence Server supports a maximum of 200 calls on any standalone server or cluster, with up to 104 calls in each conference.
- Screen licenses can be purchased in single units and applied to a device in any amount up to the maximum supported by that device.

## TelePresence Conductor Sizing

The total number of TelePresence Servers for non-scheduled conferences is limited by the capacity of TelePresence Conductor. Table 6-8 lists TelePresence Conductor capacities.

***Table 6-8 TelePresence Conductor Capacities***

| OVA Template | Total Number of TelePresence Servers | Total Number of Concurrent Participants Across All TelePresence Servers |
|---|---|---|
| Small OVA template | 30 | 50 |
| Large OVA template or appliance | 30 | 2,400 |

Clustering provides only high availability; it does not increase the maximum number of conference bridges or concurrent calls that can be supported.

If a deployment grows beyond the capacity of a single TelePresence Conductor cluster, it is possible to create additional independent TelePresence Conductor clusters and continue to add TelePresence Servers there.

An independent TelePresence Conductor cluster should be used per regional Unified CM cluster. Using the topology example in this document (see the Call Control chapter), there would be one TelePresence Conductor cluster for the US Unified CM cluster and another one for the EMEA Unified CM Cluster.

# Collaboration Edge

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

## Cisco Expressway Sizing

Cisco Expressway simplified sizing and licensing guidance covers only a few configurations: clusters of 2, 3, or 6 nodes. There are other possible configurations that are not covered in this document; refer to the Cisco Expressway product documentation for details.

Table 6-9 shows the maximum capacity that can be handled at any point of time by a single node.

Table 6-10 shows the recommended cluster capacity for the simplified sizing and licensing 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 or the licensing capacity (N+1 redundancy). With a cluster of 6 nodes, two nodes can fail without impacting the cluster capacity or the licensing capacity (N+2 redundancy).

Mobile and remote access does not require any specific licenses, but business-to-business communications 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 that 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 stated in Table 6-9. To support N+1 and N+2 redundancy models, the total number of rich media sessions in the cluster should not exceed the physical capacity of the remaining N nodes in the cluster.

In order to better understand the relationship between the cluster capacity, license capacity, and 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 150 sessions. In a two-node cluster in a non-resilient deployment, the video call cluster capacity is 300, but it would be reduced by half if one node fails. In order to provide resiliency and maintain the cluster capacity if one of the two nodes fails, the recommended high-available two-node cluster capacity is limited to 150 video sessions. During normal operations, video calls are load-balanced across the cluster; and with business-to-business communications, rich media session licenses are shared across the cluster. If one node fails, the remaining node is licensed to handle all 150 cluster video sessions because of license sharing. Because the node capacity is also 150 video sessions, the remaining node can then handle all 150 video sessions, and therefore the cluster capacity is maintained.

*Table 6-9 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 or appliance CE500 | 2,500 | 150 | 300 |
| Virtual machine with large OVA template or appliance CE1000 | 2,500 | 500 | 1,000 |

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

*Table 6-10 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 or appliance CE500** | | | | | |
| Deployment 1 | 2 nodes | N+1 | 2,500 | 150 | 300 |
| Deployment 2 | 3 nodes | N+1 | 5,000 | 300 | 600 |
| Deployment 3 | 6 nodes | N+2 | 10,000 | 600 | 1,200 |
| **Virtual machine with large OVA template or appliance CE1000** | | | | | |
| Deployment 4 | 2 nodes | N+1 | 2,500 | 500 | 1,000 |
| Deployment 5 | 3 nodes | N+1 | 5,000 | 1,000 | 2,000 |
| Deployment 6 | 6 nodes | N+2 | 10,000 | 2,000 | 4,000 |

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

**Note** The large OVA template is supported only with limited hardware. Refer to the documentation at http://www.cisco.com/go/uc-virtualized for more information.

The following assumptions are used for the Expressway simplified sizing deployments in Table 6-10:

- 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 or for CE500 appliances, the call rate is up to 5 calls per second (cps) per node.
- For virtual machines using the large OVA template or appliance CE1000, the call rate is up to 10 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 large OVA template must not be deployed if the nodes in the Expressway-E cluster or in the corresponding Expressway-C cluster are using the medium OVA template.
- 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

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

#### Cisco Expressway Sizing Example

A company has 8,000 users, and on average 2,000 users are traveling at any given time. 80% of the mobile users require mobile and remote access. In this case, Expressway has to be sized to allow for 1,600 concurrent registrations (80% of 2,000).

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 80 (5% of 1,600).

In the corporate network, 1% of the users are on a business-to-business calls at the same time. This accounts for an additional 60 calls (1% of (8,000 – 2,000)).

In this case we need to size the cluster to support 1,600 concurrent registrations and 140 concurrent calls (80+60).

Table 6-9 shows that a medium OVA template supports up to 150 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 1,600 registrations and 140 calls at the same time, as shown by Deployment 1 in Table 6-10. 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 2%. We now need to account for 200 concurrent calls (80+120) instead of 140. The maximum that a medium OVA template can handle is 150 calls, so we need to deploy a larger cluster in this case. Table 6-10 shows that Deployment 2 can account for 300 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 2900, 3900, and 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 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 6-11 provides capacity examples for a few platforms. 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

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

***Table 6-11 Cisco Unified Border Element Capacity Examples***

| Platform | Maximum SIP Trunk Sessions |
|---|---|
| Cisco 2901 Integrated Service Router | 100 |
| Cisco 3925 Integrated Service Router | 800 |
| Cisco 4451-X Integrated Service Router | 4,000 |
| Cisco 1004 and 1006 Aggregation Services Routers | 16,000 |

### Cisco Unified Border Element Sizing Example

A company has 8,000 users. 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, 640 SIP sessions (8% of 8,000) are needed. Table 6-11 shows that a Cisco 3925 ISR can support up to 800 sessions. Thus, for this example two Cisco 3925 ISRs with Cisco Unified Border Element software are selected, one active and one standby to provide redundancy.

# Core Applications

This section covers sizing for the applications discussed in the Core Applications chapter: namely, Cisco Unity Connection and Cisco TelePresence Management Suite (TMS).

## 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. These deployments are shown in Table 6-12. There are other possible deployments with Unity Connection, but they are not covered in this guide. Refer to the *Cisco Collaboration SRND* and product documentation for information on the other possible deployments.

***Table 6-12 Cisco Unity Connection Simplified Sizing Deployments***

| Deployment Size | Unity Connection Nodes to be Deployed for Active/Active |
|---|---|
| 1,000 users | One Unity Connection pair using 1k-user OVA template |
| 1,000 to 5,000 users | One Unity Connection pair using 5k-user OVA template |
| 5,000 to 10,000 users | One Unity Connection pair using 10k-user OVA template |

### Cisco Unity Connection Assumptions

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 http://docwiki.cisco.com/wiki/Virtualization_for_Cisco_Unity_Connection
- Cisco Unity Connection product documentation available at http://www.cisco.com/c/en/us/support/unified-communications/unity-connection-version-10-x/model.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 *Cisco Unity Connection Supported Platforms List*.

## Cisco TelePresence Management Suite (TMS)

We recommend two simplified sizing deployments for Cisco TMS, illustrated in Table 6-13. There are other possible TMS deployments, but they are not covered in this guide. For instance, the single server deployment that has all TMS, TMSPE, 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 6-13 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 6-13 Cisco TMS Simplified Deployments and Capacities*

| Deployment Model | Deployment | Cisco TMS | Cisco TMSXE | Cisco TMSPE |
|---|---|---|---|---|
| Regular Deployment (2 vCPU OVA template) | 2 nodes total: each with TMS, TMSPE, and TMSXE<br>Additional servers for Microsoft SQL | < 200 controlled systems (endpoints added to TMS for scheduling)<br>< 100 concurrent participants<br>< 50 concurrent ongoing scheduled conferences | < 50 endpoints bookable in Microsoft Exchange | < 1,000 Collaboration Meeting Rooms (CMRs) |
| Large Deployment (4 vCPU OVA template) | 4 nodes total: 2 each with both TMS and TMSPE; and 2 with TMSXE only<br>Additional servers for Microsoft SQL | < 5,000 controlled systems (endpoints added to TMS for scheduling)<br>< 1,800 concurrent participants<br>< 250 concurrent ongoing scheduled conferences | < 1,800 endpoints bookable in Microsoft Exchange | < 48,000 Collaboration Meeting Rooms (CMRs) |

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

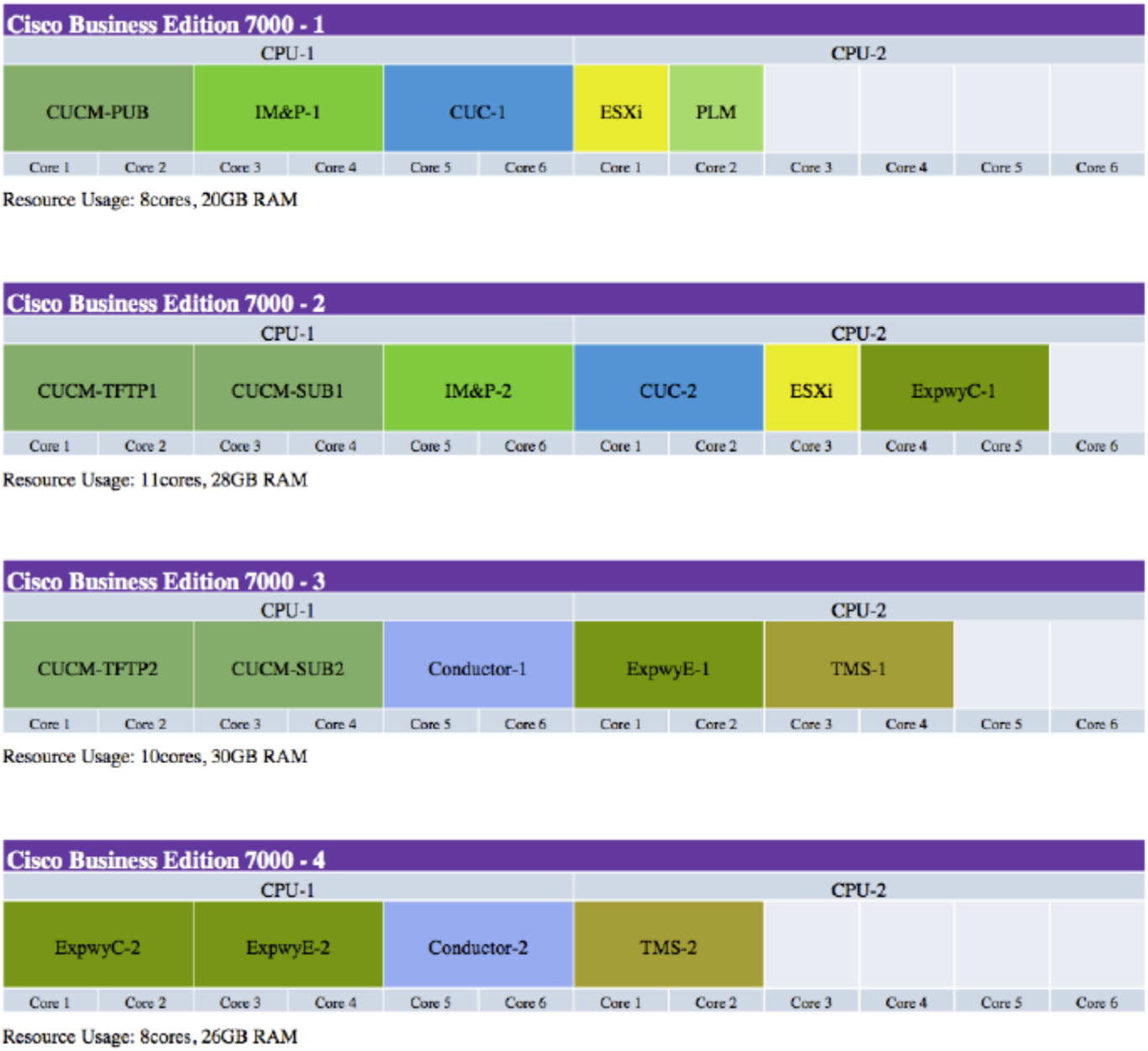
http://www.cisco.com/c/en/us/support/conferencing/telepresence-management-suite-tms/products-installation-guides-list.html

# 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 http://www.cisco.com/go/vmpt.

Figure 6-1 shows an example of using VMPT for a deployment with 5,000 users. This example assumes that Cisco Business Edition 7000M is deployed. It does not include the TelePresence servers, which could be deployed, for example, with the Multiparty Media 400v or TelePresence Server MSE 8710 platforms.

*Figure 6-1 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

http://docwiki.cisco.com/wiki/Unified_Communications_Virtualization_Sizing_Guidelines#Application_Co-residency_Support_Policy

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.
- Cisco Unity Connection requires a spare physical core to 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 6-1. 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. It includes the Cisco UCS server hardware and a hypervisor license. VMware vSphere Hypervisor (ESXi) is pre-installed. Business Edition 7000 is also pre-loaded with the Cisco Collaboration software set.