Sizing and Operating Conditions for Reference Designs

Contact Center Basic Traffic Terminology

It is important to be familiar with, and to be consistent in the use of, common contact center terminology. Improper use of these terms in the tools used to size contact center resources can lead to inaccurate sizing results.

The terms listed in this section are the most common terms used in the industry for sizing contact center resources. There are also other resources available on the internet for defining contact center terms.

Busy Hour or Busy Interval

A busy interval can be one hour or less (such as 30 minutes or 15 minutes, if sizing is desired for such smaller intervals). The busy interval occurs when the most traffic is offered during this period of the day. The busy hour or interval varies over days, weeks, and months. There are weekly busy hours and seasonal busy hours. There is one busiest hour in the year. Common practice is to design for the average busy hour (the average of the 10 busiest hours in one year). This average is not always applied, however, when staffing is required to accommodate a marketing campaign or a seasonal busy hour such as an annual holiday peak. In a contact center, staffing for the maximum number of agents is determined using peak periods, but staffing requirements for the rest of the day are calculated separately for each period (usually every hour) for proper scheduling of agents to answer calls versus scheduling agents for offline activities such as training or coaching. For trunks in most cases it is not practical to add or remove trunks or ports daily, so these resources are sized for the peak periods. In some retail environments, additional trunks can be added during the peak season and disconnected afterwards.

Busy Hour Call Attempts (BHCA)

The BHCA is the total number of calls during the peak traffic hour (or interval) that are attempted or received in the contact center. For the sake of simplicity, we assume that all calls offered to the Voice Gateway are received and serviced by the contact center resources. Calls normally originate from the PSTN, although calls to a contact center can also be generated internally, such as by a help-desk application.

Calls Per Second as reported by Call Router (CPS)

These are the number of call routing requests received by the Unified CCX Call Router per second. Every call will generate one call routing request in a simple call flow where the call comes in from an ingress gateway and is then sent to an Agent; however, there are conditions under which a single call will need more than one routing request to be made to the Unified CCX Call Router to finally get to the right agent.

Servers

Servers are resources that handle traffic loads or calls. There are many types of servers in a contact center, such as PSTN trunks and gateway ports, agents, and voicemail ports.

Talk Time

Talk time is the amount of time an agent spends talking to a caller, including the time an agent places a caller on hold and the time spent during consultative conferences.

Wrap-Up Time (After-Call Work Time)

After the call is terminated (the caller finishes talking to an agent and hangs up), the wrap-up time is the time it takes an agent to wrap up the call by performing such tasks as updating a database, recording notes from the call, or any other activity performed until an agent becomes available to answer another call.

Average Handle Time (AHT)

AHT is the mean (or average) call duration during a specified time period. It is a commonly used term that refers to the sum of several types of handling time, such as call treatment time, talk time, and queuing time. In its most common definition, AHT is the sum of agent talk time and agent wrap-up time.

Erlang

Erlang is a measurement of traffic load during the busy hour. The Erlang is based on having 3600 seconds (60 minutes, or 1 hour) of calls on the same circuit, trunk, or port. (One circuit is busy for one hour regardless of the number of calls or how long the average call lasts.) If a contact center receives 30 calls in the busy hour and each call lasts for six minutes, this equates to 180 minutes of traffic in the busy hour, or 3 Erlangs (180 min/60 min). If the contact center receives 100 calls averaging 36 seconds each in the busy hour, then total traffic received is 3600 seconds, or 1 Erlang (3600 sec/3600 sec).

Use the following formula to calculate the Erlang value:

Traffic in Erlangs = (Number of calls in the busy hour * AHT in sec) / 3600 sec

The term is named after the Danish telephone engineer A. K. Erlang, the originator of queuing theory used in traffic engineering.

Busy Hour Traffic (BHT) in Erlangs

BHT is the traffic load during the busy hour and is calculated as the product of the BHCA and the AHT normalized to one hour:

BHT = (BHCA * AHT seconds) / 3600, or
BHT = (BHCA * AHT minutes) / 60

For example, if the contact center receives 600 calls in the busy hour, averaging 2 minutes each, then the busy hour traffic load is (600 * 2/60) = 20 Erlangs.

BHT is typically used in Erlang-B models to calculate resources such as PSTN trunks. Some calculators perform this calculation transparently using the BHCA and AHT for ease of use and convenience.

Grade of Service (Percent Blockage)

This measurement is the probability that a resource or server is busy during the busy hour. All resources might be occupied when a user places a call. In that case, the call is lost or blocked. This blockage typically applies to resources such as Voice Gateway ports, PBX lines, and trunks. In the case of a Voice Gateway, grade of service is the percentage of calls that are blocked or that receive busy tone (no trunks available) out of the total BHCA. For example, a grade of service of 0.01 means that 1% of calls in the busy hour is blocked. A 1% blockage is a typical value to use for PSTN trunks, but different applications might require different grades of service.

Blocked Calls

A blocked call is a call that is not serviced immediately. Callers are considered blocked if they are rerouted to another route, if they are delayed and put in a queue, or if they hear a tone (such as a busy tone) or announcement. The nature of the blocked call determines the model used for sizing the particular resources.

Service Level

This term is a standard in the contact center industry, and it refers to the percentage of the offered call volume (received from the Voice Gateway and other sources) that are answered within x seconds, where x is a variable. A typical value for a sales contact center is 90% of all calls answered in less than 10 seconds (some calls are delayed in a queue). A support-oriented contact center might have a different service level goal, such as 80% of all calls answered within 30 seconds in the busy hour. Your contact center's service level goal determines the number of agents needed, the percentage of calls that are queued, the average time calls spend in queue, and the number of PSTN trunks needed.

Queuing

When agents are busy with other callers or are unavailable (after call wrap-up mode), subsequent callers must be placed in a queue until an agent becomes available. The percentage of calls queued and the average time spent in the queue are determined by the service level desired and by agent staffing. Cisco's Unified CCX solution uses a IVR to place callers in queue and play announcements. It can also be used to handle all calls initially (call treatment, prompt and collect such as DTMF input or account numbers or any other information gathering) and for self-service applications where the caller is serviced without needing to talk to an agent (such as obtaining a bank account balance, airline arrival/departure times, and so forth). Each of these scenarios requires a different number of IVR ports to handle the different applications because each has a different average handle time and possibly a different call load.

Server Capacities and Limits

OVA Profile

The following table displays Open Virtualization Alliance (OVA) configuration settings to be used for Unified CCX:

Table 1: OVA Settings

Agent Capacity	vCPU	vRAM	vDisk
100 agents	2	10 GB	1 x 146GB
300 agents	2	10 GB	2 x 146GB
400 agents	4	16 GB	2 x 146GB

The following table provides a selected list of capacity limits when deploying Unified CCX.

Table 2: Capacity Limits

Deployment	Capacity
Maximum number of teams	8 **Note** This maximum of eight teams is mentioned considering that each team has five supervisors. However, more teams can be created if the number of supervisors are less for each team. **For example**: If one team is assigned with one supervisor, you can have a maximum of 40 teams.
Maximum number of supervisors in a team	5
Maximum number of inbound agents	400
Maximum number of preview outbound agents	150
Maximum number of remote agents	100
Maximum number of concurrent supervisors	42
Maximum number of teams that a supervisor can be assigned	5
Maximum number of agents in a team	50
Maximum number of IVR ports	400
Maximum number of outbound IVR ports	150
Maximum number of progressive and predictive outbound agents	150

This table shows absolute limits. Reaching the limits for multiple criteria in a specific configuration might not be possible. Use the Cisco Unified Communications Sizing Tool to validate your configuration. This tool is available at:

http://tools.cisco.com/cucst

The Cisco Unified Communications Sizing Tool is available to Cisco partners only. For more details and to validate your configuration, contact your Cisco sales engineer or Cisco partner to access this tool.

For information on capacity and sizing of Cisco Workforce Optimization, refer to the *Cisco Workforce Optimization System Configuration Guide*.

The summary overview of system maximums for inbound and outbound voice listed in the table is for reference only.

Table 3: Reference Capacities for Inbound Deployment

Inbound-Only Deployment- Maximum Capacities						
	Standalone Server			**Two-Server Cluster**		
OVA profile	3	2	1	3	2	1
Agents	400	300	100	400	300	100
Supervisors	42	32	10	42	32	10
Web Chat volume per hour	2400[1]	2400 [2]	1200 [3]	2400 [4]	2400 [5]	1200 [6]
Silent Monitoring	42	32	10	42	32	10
Recording and Playback using Finesse	The recording limit is based on the number of recording licenses deployed on Unified CCX.					
Customer service queues	250	250	35	250	250	35
Skills	250	250	250	250	250	250
Historical reporting sessions	8	8	3	16	16	10
IVR ports[7]	400	300	100	400	300	100
ASR ports	100	100	50	100	100	50
TTS ports	160	160	40	160	160	40
VoiceXML ports	80	80	40	80	80	40
Busy Hour Call Completions (BHCC)	6000	5000	2000	6000	5000	2000
Number of skills with which an agent can associate	50	50	50	50	50	50
Number of CSQs with which an agent can associate (includes total combined email CSQs and voice CSQs)	25	25	25	25	25	25
Number of skills with which a CSQ can associate	50	50	50	50	50	50

Inbound-Only Deployment- Maximum Capacities						
Number of CSQs for which a call can queue	25	25	25	25	25	25
Number of agents per team	50	50	50	50	50	50

1 Large profile of SocialMiner is supported.
2 Large profile of SocialMiner is supported.
3 Small profile of SocialMiner is supported.
4 Large profile of SocialMiner is supported.
5 Large profile of SocialMiner is supported.
6 Small profile of SocialMiner is supported.
7 The number of IVR ports is also limited by the maximum number supported for a given server platform. In case of virtualized deployment, the maximum number of IVR ports is limited by the maximum number supported for a given virtual machine template.

Table 4: Reference Capacities for Email Deployment

	Standalone Server			Two-Server Cluster		
OVA Profile	3	2	1	3	2	1
Total Agents	400	300	100	400	300	100
Agents assigned to handle Emails	120	120	60	120	120	60
Email volume per hour (MS Exchange) with Small Attachments[8]	400	300	100	400	300	100
Email volume per hour (Office 365 or Gmail) with Small Attachments[9]	300	200	75	300	200	75
Email volume per hour (MS Exchange, Office 365 or Gmail) with Larger Attachments[10]	100	75	25	100	75	25
Maximum CSQs for Agent Email	100	100	100	100	100	100

8 (a). Maximum size of each attachment is less than 2 MB. (b). Maximum size of combined attachments in an email sent is 5 MB and 10 MB in a received email. (c). Maximum number of attachments in an email is 10.
9 (a). Maximum size of each attachment is less than 2 MB. (b). Maximum size of combined attachments in an email sent is 5 MB and 10 MB in a received email. (c). Maximum number of attachments in an email is 10.
10 (a). Maximum size of each attachment can range between 2-10 MB. (b). Maximum size of combined attachments in an email can range between 10-20 MB. (c). Maximum number of attachments in an email is 10. The limits have been tested and validated for 15% of total Emails with maximum attachment size.

Note The maximum Web Chat Concurrent sessions for any type of OVA profile used must not exceed 120.

Table 5: Reference Capacities for Blended Deployments

Blended Deployment- Maximum Capacities						
	Standalone Server			**Two-Server Cluster**		
Agents	400	300	100	400	300	100
Supervisors	42	32	10	42	32	10
Silent Monitoring	42	32	10	42	32	10
Customer service queues	250	250	35	250	250	35
Skills	250	250	250	250	250	250
IVR ports	400	300	100	400	300	100
ASR ports	100	100	50	100	100	50
TTS ports	160	160	40	160	160	40
VoiceXML ports	80	80	40	80	80	40
Web Chat volume per hour	2400 [11]	2400 [12]	1200 [13]	2400 [14]	2400 [15]	1200 [16]
Blended or Preview Agents	150	150	75	150	150	75
Blended or Progressive/Predictive Agents	150	150	75	150	150	75
Preview Outbound BHCC	6000	5000	2000	6000	5000	2000
Progressive and Predictive Outbound BHCC	6000	5000	2000	6000	5000	2000
Outbound IVR BHCC	6000	5000	2000	6000	5000	2000
Total BHCC[17]	6000	5000	2000	6000	5000	2000
Number of skills with which an agent can associate	50	50	50	50	50	50
Number of CSQs with which an agent can associate	25	25	25	25	25	25
Number of skills with which a CSQ can associate	50	50	50	50	50	50
Number of CSQs for which a call can queue	25	25	25	25	25	25

Blended Deployment- Maximum Capacities						
Number of email CSQs	100	100	100	100	100	100
Outbound IVR ports	150	150	75	150	150	75
Maximum number of configured agents	2000	2000	2000	2000	2000	2000

[11] Large profile of SocialMiner is supported.
[12] Large profile of SocialMiner is supported.
[13] Small profile of SocialMiner is supported.
[14] Large profile of SocialMiner is supported.
[15] Large profile of SocialMiner is supported.
[16] Small profile of SocialMiner is supported.
[17] For high-availability (HA) deployments, the BHCC listed in the table is for LAN deployments. For WAN deployments, BHCC is 5000, 750, and 750 for OVA profile 3 and 2, and 1, respectively. In addition, the BHCC contributed by the preview outbound dialer should not exceed 1000, 750 and 750 for OVA profile 3, 2, and 1, respectively. The BHCC contributed by Outbound IVR should not exceed 1000 and 750 for OVA profile 3 and 2 respectively. These reduced BHCCs apply only to HA over WAN deployments.

Note All the capacities stated in this section are system maximums.

Operating Considerations for Reference Design Compliant Solutions

Time Synchronization

To ensure accurate operation and reporting, all the components in your contact center solution must use the same value for the time. You can synchronize the time across your solution using a Simple Network Time Protocol (SNTP) server. The following table outlines the needs of various component types in your solution.

Important Use the same NTP sources throughout your solution. When you configure the Unified CCX node ensure to point to a Stratum-1, Stratum-2, or Stratum-3 NTP server to ensure that the cluster time is synchronized correctly with an external time source. The NTP information for second node is pulled from the first node.

Type of component	Notes
ESXi hosts	All ESXi hosts must point to the same NTP servers.
Unified CCX components	Components such as Standalone Unified Intelligence Center, SocialMiner, and Unified Communications must point to the same NTP servers.

Type of component	Notes
External components used in Unified CXX solution	MS Exchange and any Identity Provider (IdP) that is configured with Unified CCX. To point to Time Synchronized common NTP source as that of CCX components. Follow the Microsoft documentation to synchronize directly with the NTP server.
Cisco Integrated Service Routers	To provide accurate time for logging and debugging, use the same NTP source as the solution for the Cisco IOS Voice Gateways.

IPv6 Support

Unified CCX can be deployed as part of a dual stack IPv4 and IPv6 solution. Unified CCX servers and other optional servers (for example, ASR/TTS, WFM, QM etc) should be running in IPv4 segment. However, Unified CM, MediaSense servers, IP Phones and Gateways can be configured as either IPv4 or IPv6. If the calling device is in IPv6 and the receiving device is in IPv4, Unified CM dynamically inserts a media termination point (MTP) to convert the media between the two devices from IPv4 to IPv6 or vice versa. This would have an impact on Unified CM performance.

For more information on IPv6 deployment with Unified CM, refer to the document *Deploying IPv6 in Unified Communications Networks with Cisco Unified Communications Manager* available here:

http://www.cisco.com/go/ucsrnd

SIP Support

Unified CCX CTI ports are notified of caller-entered digits (DTMF input) via JTAPI messages from Unified CM. Unified CCX does not support any mechanism to detect in-band DTMF digits where DTMF digits are sent with voice packets. In deployments with voice gateways or SIP phones that only support in-band DTMF or are configured to use in-band DTMF, an MTP resource must be invoked by Unified CM to convert the in-band DTMF signaling so that Unified CM can notify Unified CCX of the caller-entered digits. Ensure to enable out-of-band DTMF signaling when configuring voice gateways in order to avoid using the previous MTP resources. For detailed design consideration related to DTMF handling, media resources and voice gateway deployments, see the Cisco Unified Communications Solution Reference Network Design at http://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html.