# Network Infrastructure Capacity Planning



## Network Infrastructure Capacity Planning

Network infrastructure is key in the capacity planning process. The information contained in this section serves as reference material to help you plan for present and future capacity requirements for the Service Provider Cisco HCS Data Center.

The overall system capacity in terms of supported subscribers and customers can vary depending on the topology of the network deployed. It is important to regularly track network infrastructure metrics to manage potential growth as Cisco HCS expands.

Network infrastructure metrics are also very important during network integration through a fresh install or when growth occurs in the network that requires software or hardware upgrades.

## Network Resource Limits

Within the following documents, the verified scaling capabilities may list multiple features enabled at the same time. The numbers listed in many cases exceed those used by most customers in their topologies. The scale numbers listed may not be the maximum verified values if each feature is viewed in isolation.

The values provided in this guide should also not be interpreted as theoretical system limits for Cisco Nexus hardware or Cisco NX-OS software. These limits refer to values that have been validated by Cisco. They can

increase over time as more testing and validation is done. If the hardware is capable of a higher scale, future software releases may increase this verified maximum limit.

Cisco Nexus 5600 Series NX-OS Verified Scalability Guide

Cisco Nexus 7000 Series NX-OS Verified Scalability Guide

Cisco Nexus 9000 Series NX-OS Verified Scalability Guide

As a part of the capacity planning process, you must be aware of the number of tenants that the components support and consider the future growth plans. The overall number of supported tenants for each deployment can differ based on the resource-usage profile variations.

# Fixed Resource Reservations

You must have an accurate representation of how resources are potentially used. From a capacity planning perspective, routing resources can be used on a fixed or tenant basis. Consult the Architectural Capacity Guide for the number of fixed resources that should be reserved and subtracted from the total resources. The remaining resources can be used on a per-tenant basis and ultimately determine the supported number of tenants for the architecture used.

# Resource usage profiles (per-tenant basis)

Usage profiles on a per-tenant basis are divided into aggregation and core usage. Each interface has different requirements depending on the configuration used. The usage profile is different if you use Border Gateway Protocol (BGP) optimization.

BGP optimization refers to how the BGP connections are provisioned between the PE and core routing. When BGP optimization is used, the cross links between PE and core routing are removed. This provides a reduction in BGP peer usage on a per tenant basis.

BGP optimization reduces the amount of links (cross-links) required on the core layer. In cases where the limiting factor is the number of BGP peers, the optimization can be beneficial.

Consult the Cisco HCS design guides for more information:
http://www.cisco.com/c/en/us/support/unified-communications/hosted-collaboration-solution-hcs/products-implementation-design-guides-list.html

# Architectural Capacity Guide

The following tables provide a capacity guide for standard Cisco HCS deployments, as described in Architecture Capacity Planning. The tables provide resource reservation information as well as fixed software limits that are placed on individual network components that can be deployed within a given Cisco HCS architecture.

The overall number of supported tenants for each deployment can differ based on the resource-usage profile variations. These tables are for planning purposes only; your deployment numbers may vary. It is important to know the actual count for each individual deployment. These numbers reflect the system capacity before onboarding any customers. As a part of the capacity planning process, you must be aware of the number of tenants that the components support and consider the future growth plans.

*Table 1: Fixed resource reservation (overhead)*

| Reserved Resource | Count | HCS per Tenant<br>Resource Assumptions |
|---|---|---|
| **BGP Peers** | 20 | 3 |
| **VRFs** | 10 | 2 |
| **HSRP Instances** | 40 | 2 |
| **Static Routes** | 50 | 4 |
| **VLANs** | 100 | 2 |

*Table 2: Resource Usage Profiles*

| | BGP Peers | Static Routes | VLANs | HSRP | VRFs |
|---|---|---|---|---|---|
| **Standard resource usage profile** | | | | | |
| HCS Collapsed Agg | 3 | 4 | 2 | 2 | 2 |
| HCS Collapsed Agg + RA VPN | 4 | 4 | 2 | 2 | 2 |
| **BGP optimization resource usage profile** | | | | | |
| HCS Collapsed Agg | 2 | 4 | 2 | 2 | 2 |
| HCS Collapsed Agg + RA VPN | 3 | 4 | 2 | 2 | 2 |
| **BGP optimization and ASA as default route resource usage profile** | | | | | |
| HCS Collapsed Agg | 2 | 3 | 2 | 2 | 2 |
| HCS Collapsed Agg + RA VPN | 3 | 3 | 2 | 2 | 2 |

*Table 3: Nexus 5600 Tenant Capacity*

| Capacity (Tenant Limit) | Remote Access | BGP Peer Optimization | BGP Peer Optimization & ASA as Default Route |
|---|---|---|---|
| **112 (NX-OS 7.3)** | No | Yes | No |
| **78 (NX-OS 7.3)** | Yes | Yes | No |
| **118 (NX-OS 7.3)** | No | Yes | Yes |
| **78 (NX-OS 7.3)** | Yes | Yes | Yes |

*Table 4: Nexus 9300 Tenant Capacity*

| Capacity (Tenant Limit) | Remote Access | BGP Peer Optimization | BGP Peer Optimization & ASA as Default Route |
|---|---|---|---|
| **225 (NX-OS 7.0)**<br>**112 (NX-OS 9.2)** | No | Yes | No |
| **164 (NX-OS 7.0)**<br>**112 (NX 9.2)** | Yes | Yes | No |
| **225 (NX 7.0)**<br>**150 (NX 9.2)** | No | Yes | Yes |
| **164 (NX 7.0)**<br>**150 (NX 9.2)** | Yes | Yes | Yes |

*Table 5: Nexus 9500 Tenant Capacity*

| Capacity (Tenant Limit) | Remote Access | BGP Peer Optimization | BGP Peer Optimization & ASA as Default Route |
|---|---|---|---|
| **225 (NX-OS 7.0)**<br>**225 (NX-OS 9.2)** | No | Yes | No |
| **164 (NX-OS 7.0)**<br>**225 (NX-OS 9.2)** | Yes | Yes | No |
| **225 (NX-OS 7.0)**<br>**225 (NX-OS 9.2)** | No | Yes | Yes |
| **164 (NX-OS 7.0)**<br>**225 (NX-OS 9.2)** | Yes | Yes | Yes |

*Table 6: Nexus 7000 Tenant Capacity*

| Capacity (Tenant Limit) | Remote Access | BGP Peer Optimization | BGP Peer Optimization & ASA as Default Route |
|---|---|---|---|
| **987 (NX-OS 7.2)**<br>**987 (NX-OS 8.3)** | No | Yes | No |
| **660 (NX-OS 7.2)**<br>**826 (NX-OS 8.3)** | Yes | Yes | No |

| Capacity (Tenant Limit) | Remote Access | BGP Peer Optimization | BGP Peer Optimization & ASA as Default Route |
|---|---|---|---|
| **990 (NX-OS 7.2)**<br>**1240 (NX-OS 8.3)** | No | Yes | Yes |
| **660 (NX-OS 7.2)**<br>**826 (NX-OS 8.3)** | Yes | Yes | Yes |

# Unified Communications Application Capacity Planning

Accurate sizing is critical to ensure that deployed systems will meet the expected service quality for call volumes and throughput. For standalone products, manual calculation of the system size may be feasible. For large and complex deployments, the system designer will need to consider several design and deployment factors that influence system sizing. For example, multiple products may be distributed across different locations and may include video endpoints, call centers, and voice/video conferencing. There are also network factors to be considered including:

- Cluster sizes
- Interaction between individual products
- Server capabilities
- Optional capabilities and features

As well as other sizing factors including:

- Mix of call types
- Mix of endpoint types
- System release
- Use of external applications
- Anticipated system growth
- Average and peak usage

For more information on UC application sizing see, Unified CM Sizing of the ***Preferred Architecture for Cisco Collaboration 12.x Enterprise On-Premises Deployments, CVD*** *guide*.

# Methodology for UC Application System Sizing

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

### Performance Testing

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

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

### System Modeling

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

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

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

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

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

### Traffic Engineering

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

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

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

- Voice Traffic
- Contact Center Traffic
- Video Traffic
- Conferencing and Collaboration Traffic

For planning information, refer to the *Cisco Hosted Collaboration Solution End-to-End Planning Guide* and the *Cisco Hosted Collaboration Solution Reference Network Design Guide*.

# Capacity Planning for Client Services Framework

Cisco Unified Client Services Framework operates as either a SIP endpoint registered to Unified CM or as a deskphone controller of a Cisco Unified IP Phone using a CTI connection to Unified CM. When planning a deployment using the Client Services Framework, Cisco partners and employees can use the Cisco Unified Communications Sizing Tool (available at http://tools.cisco.com/cucst ) to assist in the appropriate sizing of SIP registered endpoints and CTI controlled devices.

The following additional items must be considered for a Client Services Framework deployment:

- TFTP: When configured in softphone mode, a Client Services Framework device configuration file is downloaded through TFTP to the client for Unified CM call control configuration information. In addition, any application dial rules or directory lookup rules are also downloaded through TFTP to Client Services Framework devices.
- CTI: When configured in deskphone mode, the Client Services Framework establishes a CTI connection to Unified CM upon login and registration to allow for control of the IP phone. Unified CM supports up to 40,000 CTI connections. If you have a large number of clients operating in deskphone mode, make sure that you evenly distribute those CTI connections across all Unified CM subscribers running the CTIManager service. This can be achieved by creating multiple CTI Gateway profiles, each with a different pair of CTIManager addresses, and distributing the CTI Gateway profile assignments across all clients using deskphone mode.
- CCMCIP: The Client Services Framework uses the Cisco CallManager Cisco IP Phone (CCMCIP) service to gather information about the devices associated with a user, and it uses this information to provide a list of IP phones available for control by the client in deskphone control mode. The Client Services Framework in softphone mode uses the CCMCIP service to discover its device name for registration with Unified CM.
- IMAP: When configured for voicemail, the Client Services Framework updates and retrieves voicemail through an IMAP connection to the mailstore.
- LDAP: Client login and authentication, contact profile information, and incoming caller identification are all handled through a query to the LDAP directory, unless stored in the local Client Services Framework cache.
- UDS: The UDS service can be used by clients to search for contacts in the Unified CM User database. Like LDAP directory searches, UDS contact searches take place if the requested contact cannot be found in the local Client Services Framework cache.

# High Availability for Client Services Framework

Cisco Unified Client Services Framework provides primary and secondary servers for each of the following configuration components: TFTP server, CTIManager, CCMCIP server, voicemail server, UDS server, and LDAP server. When operating in softphone mode, the Client Services Framework is registered with Cisco Unified CM as a SIP endpoint, and it supports all of the registration and redundancy capabilities of a registered endpoint of Unified CM. When operating in deskphone mode, the Client Services Framework is controlling a Cisco Unified IP Phone using CTI, and it supports configuration of a primary and secondary CTIManager in the CTIManager Profile.

# Prime Collaboration Assurance Capacity Planning

Cisco Prime Collaboration Assurance (PCA) is a comprehensive video and voice service assurance and management system with a set of monitoring and reporting capabilities that help you receive a consistent, high-quality video and voice collaboration experience. PCA is available in two modes: Cisco Prime Collaboration Assurance Advanced—Enterprise and MSP mode.

- The Enterprise mode provides a single enterprise view or multiple domains view within your enterprise. This option is usually used in a standard single enterprise environment.

- The MSP mode provides multiple customer views. This option is used in managed service provider environments. This view allows you to view the devices of multiple customers that are being managed.

For more information on the these modes, see the *Cisco Prime Collaboration Assurance - Advanced and Analytics Guide* available at:
http://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-user-guide-list.html

To install Cisco Prime Collaboration Assurance - Advanced (without Cisco Prime Collaboration Analytics), you need only one virtual machine. If you want to enable Cisco Prime Collaboration Analytics during the Cisco Prime Collaboration Assurance Advanced installation, the number of virtual machines that are required to install Cisco Prime Collaboration Assurance depends on the number of endpoints that you want to manage in Cisco Prime Collaboration Analytics:

- If you have fewer than or equal to 80,000 endpoints (small, medium, and large deployment models), you need one virtual machine where you can install both the database and application.
- If you have more than 80,000 endpoints (very large deployment model), you need two virtual machines to install the database and application separately on each machine.

For information on installing Advanced Assurance, see the *Cisco Prime Collaboration Assurance and Analytics Install and Upgrade Guide* available at:
https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-collaboration/products-installation-guides-list.html

# Expressway Capacity Planning

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

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

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

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

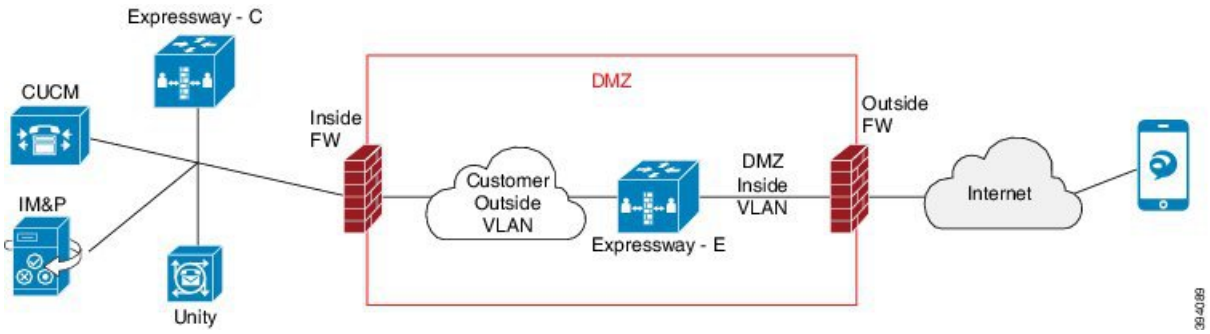
The standard deployment of the Cisco Collaboration Edge architecture involves deploying at least one Expressway-C and Expressway-E pair for secure mobile device and remote VPN-less access back to enterprise collaboration services. An Expressway-C with a trunk and line-side connection to Unified CM, and an Expressway-E deployed in the DMZ and configured with a traversal zone to an Expressway-C.

It is recommended to deploy Expressway-C and Expressway-E in clusters. Each cluster can have up to six Expressway nodes and a maximum of N+2 physical redundancy. All nodes are active in the cluster. In a multitenant deployment, the Expressway-E's capacity is shared across all the customers, whereas the Expressway-C cluster's capacity is dedicated to the customer.

Since Expressway-C is deployed in the internal network and Expressway-E in the DMZ, Expressway-C must be connected to Expressway-E through a Unified Communications traversal zone for mobile and remote

access. Business-to-business calls require a separate traversal zone, which retains the name of traversal client zone for Expressway-C and traversal server zone for Expressway-E. The traversal server, traversal client, and Unified Communications traversal zones include all the nodes of Expressway-C and Expressway-E, so that if one of the nodes is not reachable, another node of the cluster will be reached instead.

***Figure 1: Expressway Firewall Traversal***




The following guidelines apply when clustering Cisco Expressway for mobile and remote access:

- Expressway clusters support up to 6 nodes (4 active and 2 backup).
- All nodes of the Expressway-E and Expressway-C cluster pairs must use identical OVA templates. For example, an Expressway-E node using the large OVA template must not be deployed if other nodes in the Expressway-E cluster or in the corresponding Expressway-C cluster are using the medium size OVA template.
- Expressway peers should be deployed in equal numbers across Expressway-E and Expressway-C clusters. For example, a three-node Expressway-E cluster should be deployed with a three-node Expressway-C cluster.
- An Expressway-E and Expressway-C cluster pair can be formed by a combination of nodes running on an appliance or running as a virtual machine, if the node capacity is the same across all nodes.
- The Expressway node OVA templates or Expressway Appliances must match across and within Expressway Series cluster pairs.
- Multiple pairs of Expressway Series clusters may be deployed to increase capacity.

For details about cluster configuration, refer to the latest version of the *Cisco Expressway Cluster Creation and Maintenance Deployment Guide* available at:
https://www.cisco.com/c/en/us/support/unified-communications/expressway/model.html?dtid=osscdc000283

For more information about Cisco Expressway capacity planning considerations, including sizing limits, capacity planning, and deployment considerations, refer to the Cisco Expressway product documentation available at:
http://www.cisco.com/c/en/us/support/unified-communications/expressway-series/tsd-products-support-series-home.html

For more information about Cisco Expressway sizing considerations, refer to Cisco Expressway Sizing of the *Preferred Architecture for Cisco Collaboration 12.x Enterprise On-Premises Deployments, CVD* guide.

# Capacity Planning for Dial Plan Profiles

The dial plan is one of the key elements of a Unified Communications and Collaboration system, and an integral part of all call processing agents. Generally, the dial plan is responsible for instructing the call processing agent on how to route calls. Highly complex dial plans can include multiple line appearances as well as large numbers of partitions, calling search spaces, route patterns, translations, route groups, hunt groups, pickup groups, route lists, call forwarding, co-resident services, and other co-resident applications. All these functions can consume additional resources within the Unified CM system and can have an impact on the call processing capabilities of the system and in some cases can reduce the overall capacity.

Refer to *Cisco Collaboration SystemSolution Reference Network Designs (SRND)* for information regarding:

- Dial Plan Fundamentals - General concepts commonly used in enterprise voice and video dial plans.
- Dial Plan Elements - Various dial plan elements available in the architectural elements of an enterprise collaboration solution, including Cisco Unified Communications Manager (Unified CM) and Cisco TelePresence Video Communication Server (VCS).
- Recommended Design - Design guidelines related to multisite collaboration networks, endpoint addressing, and building classes of service. Also, dial plan integration between Unified CM and VCS is covered.
- Special Considerations
    - Automated Alternate Routing
    - Device Mobility
    - Extension Mobility
    - Time-of-Day Routing
    - Logical Partitioning

# ACI Capacity Planning

This section contains the maximum verified scalability limits for Cisco Application Centric Infrastructure (Cisco ACI) parameters in the following releases:

Cisco Application Policy Infrastructure Controller (Cisco APIC), Release 5.0(1)

Verified Scalability Guide for Cisco APIC, Release 5.0(1), Multi-Site, Release 3.0(1), and Cisco Nexus 9000 Series ACI-Mode Switches, Release 15.0(1)

Multi-instance capability allows you to run your container instances that use a subset of resources of the security module/engine. Multi-instance capability is supported for the Firepower Threat Defense, it is not supported for the ASA. To understand more about Firepower Threat Defense 4100/9300, see Cisco FTD Multi Instance.