Scalability and Performance

HCS PoD Tenant Scaling

Along with the recommendations described in Architecture Capacity Planning, the number of customers that can be accommodated in a Cisco HCS PoD configuration is based on the following key factors that are required for each customer:

- Two VRFs are defined for each HCS customer on the Nexus aggregation switch, one northbound and one toward the Adaptive Security Appliance (ASA)
- Two HSRP groups are defined, one toward the applications and ASA (inside), and one for the SBC (demarcation device) and ASA (outside)
- Two VLANs are provided for each HCS customer, one toward the application and one for ASA and Session Border Controller (*SBC*).
- Two static routes are provided inside and outside to connect to the ASA firewall
- One static route is defined to the outside VLAN to connect to the SBC
- One static route is defined to route traffic to the management domain. This route is for any traffic communication between the management domain and on-premise devices, such as voice gateway, and so on.

The ASA and SBC connect on the same outside VLAN. This means that the outside VLAN of the ASA is same as the inside VLAN of SBC.

The outside of the SBC does not pass through the ASA so that media does not go through the ASA for the inter-customer or off-net calls. Therefore, you can place the SBC and ASA on the same HSRP group and VLAN for SBC inside and ASAs outside VLAN.

You can extend the HSRP group used by the aggregation switch facing toward the application to the ASA (security appliance) on the inside; this saves one HSRP group on the Nexus aggregation switch.

Cisco recommends that you use the static route to connect to the ASA and SBC because dynamic routing (BGP) is not supported on the ASA. Use one static route to route calls from UC applications to the firewall and use one static route to route the incoming specific customer base traffic to the firewall. In HCS deployments, all the communication between an end device and the Cisco Unified Communications Manager goes through the firewall.

Note Signaling goes through the firewall, but no media goes through the firewall other than the MOH or voicemail.

Define a static route on the Nexus aggregation switch to route the outbound traffic to the SBC and define one static route to route the customer-specific management traffic from the customer premise to the management domain.

Based on the preceding numbers, static routes are the lowest common denominator. If you require four static routes for each Cisco HCS customer in your deployment and only 4,000 static routes are supported on the Nexus 7000, the HCS customer scale numbers can be determined using the following formulas:

- Number of customers = (Static Routes - 50 Spare)/Static Route per customer
- Number of customers = (BGP peers - 20 Spare) / BGP peers per customer
- Number of customers = HSRP Groups - 40 Spare) / HSRP per customer
- Number of customers = (VRF - 10 Spare) / VRF per customer
- Number of customers = (VLANs - 100 Spare) /VLAN per customer

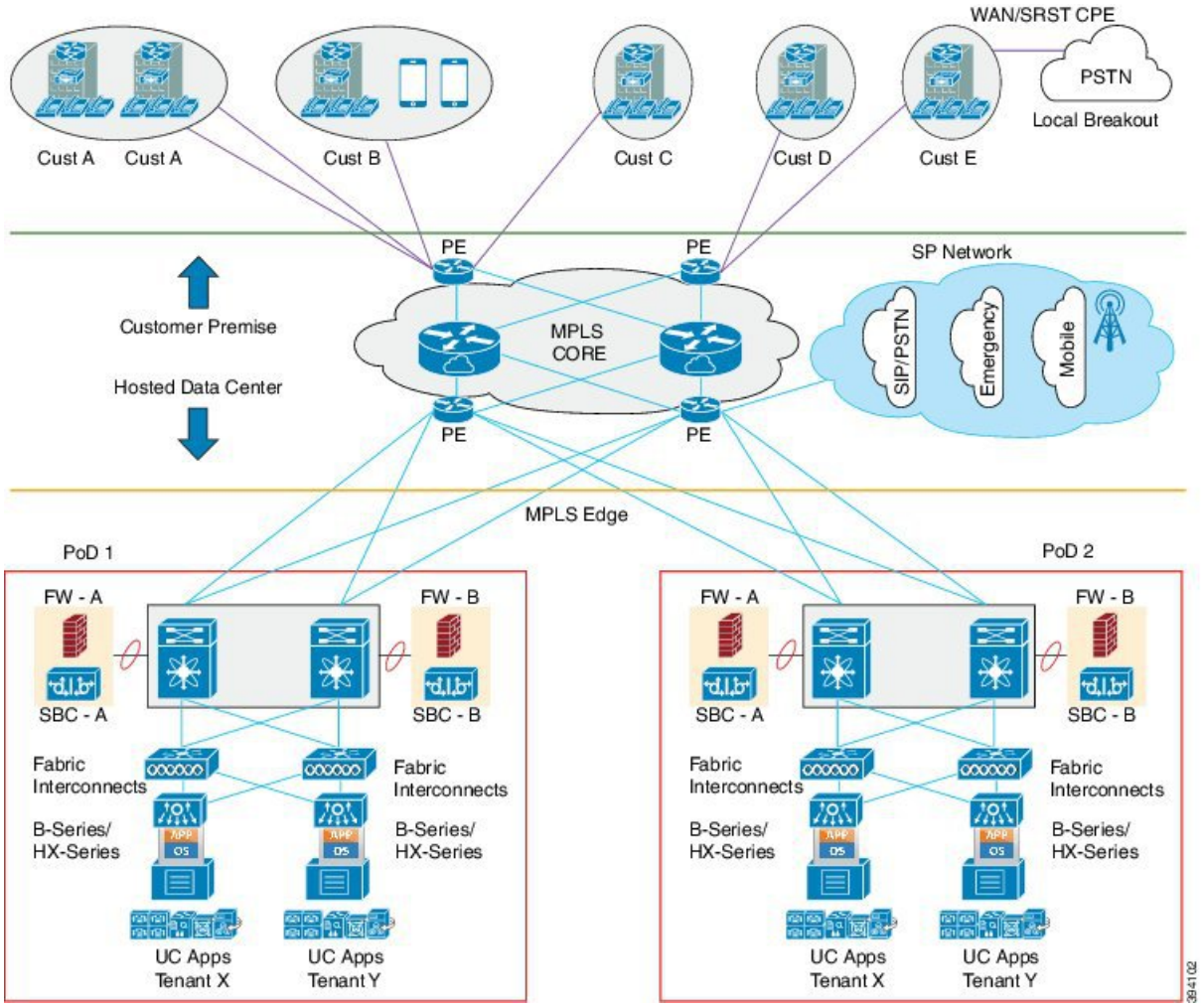
When deploying an over-the-Internet model for the same enterprises that have the MPLS-enabled HCS, there is no change to the maximum number of customers. If a service provider onboards only over-the-Internet customers, they still require four static routes per customer. Therefore, the maximum number of customers for the following deployments is the same:

- Cisco HCS deployment
- Cisco HCS deployment with over-the-top traffic (OTT)
- Cisco HCS deployment with TP
- Cisco HCS deployment with TP and OTT

PoD-Based Data Center Scale

To support the scalability of the Service Provider Cisco HCS data center based on the PoD deployment model, the solution must scale horizontally with a dedicated instance of applications and with end-to-end segregation and security. As a result, you must deploy multiple PoDs horizontally and connect them directly to the MPLS PE north bound.

Figure 1: Cisco HCS PoD Horizontal Scale Approach



1. To minimize the throughput impact to the storage traffic, you may connect the shared storage directly to the UCS Fabric Interconnects (FIs). This enables deployment of a small storage and better performance for storage as explained earlier. Smaller storage was mentioned in the context of serving the VMs deployed within a pair of FIs. Based on the current scale for Small PoD a single FI pair is sufficient and the Service Provider can start with a small storage. However, this means multiple PoDs will have to be provisioned with their own storage.
2. If you need to deploy another PoD at the same location, consider the migration of the storage to include a MDS/SAN switch. The details are not covered in this document. Connecting the storage to the Nexus 9300 will not help in storage scaling because a dedicated pair of Nexus 9300s are deployed per PoD.

PoD Bandwidth Distribution Recommendations

The recommendations for Cisco HCS Small PoD data center bandwidth distribution are as follows:

- Use a pair of Nexus 9300 Series Switches. Cisco Nexus 9300-EX and 9300-FX series switches offer a variety of interface options to transparently migrate existing data centers from 100-Mbps, 1-Gbps, and

10-Gbps speeds to 25- Gbps at the server, and from 10- and 40-Gbps speeds to 50- and 100- Gbps at the aggregation layer.

  - Cisco Nexus 9000 Fixed Switches Data Sheets: https://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/datasheet-listing.html

- Use a pair of Fabric Interconnects, 6332-16UP or 6454.
  - Cisco UCS 6300 Series Fabric Interconnects Data Sheets: https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-6300-series-fabric-interconnects/datasheet-listing.html
  - Cisco UCS Fabric Interconnect 6454 Data Sheet: https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-6400-series-fabric-interconnects/datasheet-listing.html
- Deploy one to ten UCS chassis with each UCS chassis containing up to eight B-200 blade servers.
  - Cisco UCS Manager Configuration Guides: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-installation-and-configuration-guides-list.html
  - Each physical server should contain either 192 or 256 GB RAM.
  - Two variations of virtual CPU (vCPU) to core ratio applicable for general compute are available: 1:1 for non-oversubscribed virtual machines (VMs), and 1:n for oversubscribed VMs.
- SAN Storage may be connected via Fabric Nexus 93000 Switch, Fabric Interconnects or MDS:

Cisco Data Center Platform	NPIV core	FC NPV	FCoE NPV
Cisco MDS 9700 Director	Yes	-	-
Cisco MDS 9500 Director	Yes	-	-
Cisco MDS 9396T	Yes	Yes	-
Cisco MDS 9396S	Yes	Yes	-
Cisco MDS 9250i	Yes	-	-
Cisco MDS 9222i	Yes	-	-
Cisco MDS 9148T	Yes	Yes	-
Cisco MDS 9148S	Yes	Yes	-
Cisco MDS 9148	Yes	Yes	-
Cisco MDS 9132T	Yes	Yes	-
Cisco MDS Blade Switches	Yes	Yes	-
Cisco Nexus 9000 Director	-	-	Yes
Cisco Nexus 9300 Switches	-	Yes	Yes
Cisco Nexus 7700 Directors	Yes	-	-
Cisco Nexus 7000 Directors	Yes	-	-
Cisco Nexus 6004	Yes	Yes	Yes
Cisco Nexus 5600	Yes	Yes	Yes
Cisco Nexus 5500	Yes	Yes	Yes

Cisco Data Center Platform	NPIV core	FC NPV	FCoE NPV
Cisco UCS FI 6454	Yes	Yes	Yes
Cisco UCS FI 6332/6334	Yes	Yes	Yes
Cisco UCS FI 6248UP/6296UP	Yes	Yes	Yes

- Use a pair of Firepower 2100 or 4100 Series NGFWs in ASA mode.
  - Cisco Firepower 2100 Series: https://www.cisco.com/c/en/us/support/security/firepower-2100-series/tsd-products-support-series-home.html
  - Cisco Firepower 4100 Series: https://www.cisco.com/c/en/us/support/security/firepower-4100-series/tsd-products-support-series-home.html
- Each SBC has 10 GB towards aggregation.
- Total aggregate bandwidth toward SBC is 20 GB (10 GB from each Nexus 9300).
- If the SBC is used, each SBC has a 1 GB pipe toward Nexus 9300 (Aggregation).
- The Site-to-Site VPN concentrator (ASR 1000) has 10 GB toward each aggregation switch.
- Total Aggregate bandwidth for each Site-to-Site VPN concentrator is 20 GB.
- Each Nexus 9300 has two 10 GB pipes toward the MPLS PEs.
- Total aggregated bandwidth at the MPLS PE with one pair of Nexus 9300 switches is 40 GB.

A geographically-redundant deployment of Small PoDs is supported. If clustering over WAN is needed across the two data centers, the latency between the data centers must be under 40 milliseconds.

Some of the bandwidth numbers required by the UC and management applications are provided below. This list is not complete and should not be taken as the definitive bandwidth requirements for a Small PoD deployment.

Cisco Unified Communications Manager

- A minimum of 1.544 Mbps (T1) bandwidth is required for Intra-Cluster Communication Signaling (ICCS) traffic between sites.
- In addition to the bandwidth required for Intra-Cluster Communication Signaling (ICCS) traffic, a minimum of 1.544 Mbps (T1) bandwidth is required for database and other inter-server traffic between the publisher node and every remote subscriber node.

See the *Cisco Collaboration System Solution Reference Network Designs (SRND)* at: https://www.cisco.com/c/en/us/support/unified-communications/unified-communications-manager-callmanager/products-implementation-design-guides-list.html

Cisco Unified Communications Manager IM and Presence Service

For deployments over the WAN, a minimum of 10 megabits per second of dedicated bandwidth is required for each IM and Presence Service cluster, with no more than an 80-millisecond round-trip latency. Any bandwidth less than this recommendation can adversely impact performance.

See *Configuration and Administration of the IM and Presence Service* at: https://www.cisco.com/c/en/us/support/unified-communications/unified-presence/products-installation-and-configuration-guides-list.html

Cisco Unity Connection

Cisco Unity Connection has various minimum bandwidth requirements depending on deployment type; when both servers are installed in the same building or site, or when both servers are in separate buildings or sites.

The bandwidth numbers are intended as guidelines to ensure proper operation of an active-active cluster with respect to synchronization traffic between the two servers. Additional conditions such as network congestion, CPU utilization, and message size may contribute to lower throughput than expected. Call-control and call-quality requirements are in addition to the guidelines and should be calculated using the bandwidth recommendations in the applicable Cisco Unified Communications SRND at http://www.cisco.com/c/en/us/solutions/enterprise/unified-communication-system/index.html.

See *System Requirements for Cisco Unity Connection Release 14.x* at: https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/products-installation-guides-list.html for additional details.

Deployment Models for Cisco HCS

Deployment Models for Cisco Hosted Collaboration Solution drive the actual capacity and performance calculations. There are two high-level deployment models supported for the HCS product:

1. Full Cisco HCS offering - Large PoD
2. Small or medium business offering - Small PoD

The number of users per customer is assumed to be much lower. This offering focuses on reducing the number of virtual machines (VMs) required. This is achieved primarily by relaxing requirements for application redundancy and HA. A dedicated music on hold server is not required; instead, each Cisco Unified Communications Manager VM contains the MOH function (termed co-resident MOH). Additionally, some customers (particularly in the Asia-Pacific market) may not require voicemail, which reduces the number of required VMs.

Knowledge of the following input parameters is necessary to compute the basic required resources for the various applications:

1. Number of end users (subscribers).
2. Number of endpoints (physical or virtual devices) per subscriber.

You can use the preceding input to determine the appropriate Open Virtualization Archive (OVA) file, which specifies resource allocation on the virtual machine for each application. Each OVA supported by Cisco ensures that adequate resources are provisioned for the required usage.

The domain manager provisions all the subscribers in the system at the beginning and also provides end-user self-care and other functionality for users who are already active. Some of the other management applications monitor on a per-VM basis, so it is imperative that you compute this number to determine the sizing in these instances.

For more information on supported OVAs for the following applications, refer to the Cisco Hosted Collaboration Solution Compatibility Matrix, available at www.cisco.com/go/hcsmatrix:

- Unified Communications applications
- Cisco Unity Connection
- Cisco Unified Communications Manager IM and Presence Service
- Management applications

Note As a rule, without using the UC Sizing Tool, the maximum number of users is 10K for a cluster with full redundancy using the 2500 User OVA, 30K for the 7500 User OVA, and 40K for the 10K User OVA deployment models of Cisco Unified Communications Manager. These numbers may vary depending on locations, routes, dial plan, and so on.

Standard Cisco Unified Communications Manager call processing deployments employ a cluster with up to eight subscriber nodes; assuming full redundancy at maximum capacity, that amounts to four active call processing nodes per cluster. For more information, see:

https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab12/collab12.html

https://www.cisco.com/c/dam/en/us/td/docs/voice_ip_comm/uc_system/virtualization/virtualization-cisco-unified-communications-manager.html#2500userVM

Contact Center Deployment Model for Cisco HCS

For design, configuration, installation, and upgrade information on the Contact Center Agent deployment models, see *Solution Design Guide for Cisco Hosted Collaboration Solution for Contact Center* at Cisco Hosted Collaboration Solution for Contact Center.

UC Application Distribution on Cisco UCS

The UCS Chassis houses the VMs that contain the Unified Communications Manager, Cisco Unity Connection, and Cisco Unified Communications Manager IM and Presence Service.

Cisco HCS system capacity is determined by system performance requirements rather than by strict limits, with some of the following exceptions. The overall system performance is impacted by many key parameters such as customer size, network topology, subscriber feature profile, redundancy, and call profile.

The Cisco HCS system can grow and scale to meet capacity demands through available hardware and software:

- Add UCS chassis (increase VMs)
- Upgrade Fiber Interconnect (6200>6300>6400) (Connectivity)
- IOPS capacity (migrate to SSD-based storage)
- UCS chassis supports many blade densities
- UC per cluster capacity:
  - Unified Communications Manager: dedicated cluster can support up to 40K phones
  - Cisco Unified Communications Manager IM and Presence Service: can support up to 75K users in Full UC mode
  - Cisco Unity Connection: 100K total users (20K users per server)

When designing any deployment, you should utilize the appropriate sizing tools on a per-case basis. This is necessary because system limits are multidimensional and depend on multiple factors, all of which you should discuss and analyze in advance of the actual deployment. Refer to http://tools.cisco.com/ucs and the Collaboration Sizing Tool, available at http://cucst.cloudapps.cisco.com.

Average Number of Devices Per User

The size of a Unified Communications Manager cluster is a function of a number of Unified Communications Manager configuration parameters where one of the most important is the number of devices. To derive the number of devices from a customer size defined in end users, you must define an average number of devices per user.

We assume that a foundation user consumes two devices on Unified Communications Manager and three in Standard 3 average. Given the proportion of Standard/Foundation, the average number of devices per user is 3.5.

Cisco Unity Connection Scale and Performance Limits

Design Guide for Cisco Unity Connection

For limitations details on Tenant Partitioning, see the topic Limitations of Tenant Partitioning in the Design Guide for Cisco Unity Connection available at https://www.cisco.com/c/en/us/support/unified-communications/unity-connection/tsd-products-support-series-home.html.