Cisco Unity Connection Clusters (Active/Active High Availability)

Cisco Unity Connection clusters (active/active high availability) and disaster recovery are two key customer requirements for preserving voice messaging services in the event of a system outage or disaster. For information on disaster recovery, see the Disaster Recovery System and COBRAS chapter.

✎

Note The Unity Connection cluster feature is supported only with Cisco Business Edition 6000/7000.

Unity Connection Cluster Overview

Unity Connection supports a cluster configuration of two Unity Connection servers to provide high availability and redundancy. The Unity Connection servers handle calls, HTTP, and IMAP requests. If only one server in the Unity Connection cluster is functioning, the remaining server preserves the system functionality by handling all calls, HTTP requests, and IMAP requests for the Unity Connection cluster. Note that each server in the Unity Connection cluster must have enough voice messaging ports to handle all calls for the Unity Connection cluster.

The first server installed is the publisher server for the Unity Connection cluster; the second server installed is the subscriber server. These terms are used to define the database relationship during installation. The separation of roles is consistent with the Cisco Unified Communications Manager cluster schema in which there is always one publisher server and multiple subscriber servers. (Note that Unity Connection runs on the Cisco Unified CM platform). Unlike a Cisco Unified CM cluster, however, Unity Connection supports only two Unity Connection servers in the Unity Connection cluster.

Note It is recommended to perform provisioning only on the Publisher server in Active-Active mode and on Subscriber (Acting Primary) in case of cluster failover. The password change and password setting modification for User PIN/Web application should be provisioned on Publisher server in Active-Active mode.

For a network diagram of a Unity Connection cluster integrated with Cisco Unified CM, see Figure.

A Unity Connection cluster server pair supports up to 20,000 users. In this configuration, both servers can support up to 250 voice messaging ports each for a cumulative total of 500 voice messaging ports when both servers are active. If only one server is active, the port capacity is lowered to a maximum of 250 ports.

For more information on capacity planning for a Unity Connection cluster, see the *Cisco Unity Connection 15 Supported Platforms List* at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/connection/15/supported_platforms/b_15cucspl.html.

Note A Unity Connection cluster server pair supports up to 20,000 IMAP Idle clients (250 sessions).If the IMAP clients that connect to the Unity Connection server do not support IMAP Idle, each of these clients must be counted as 4 IMAP Idle clients. For example, deploying 4 non-IMAP Idle clients is the same as deploying 16 IMAP Idle clients.

Publisher Server

The publisher server is required in a Unity Connection cluster, and there can be only one publisher server in a Unity Connection cluster server pair. The publisher server is the first server to be installed, and it provides the database and message store services to the subscriber server in the Unity Connection cluster server pair.

For information on installing a Unity Connection cluster server pair, see the "Installing Cisco Unity Connection" chapter of the Install, Upgrade, and Maintenance Guide for Cisco Unity Connection, Release 15, available at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/connection/15/install_upgrade/guide/b_15cuciumg.html.

As a best practice, you should direct the majority of client traffic (for example, IMAP and the Cisco Personal Communications Assistant) and administration traffic (for example, Cisco Unity Connection Administration, the Bulk Administration Tool, and backup operations) to the publisher server in a Unity Connection cluster server pair. However, the majority of call traffic (for example, SCCP, SIP, or PIMG/TIMG) should be directed to the subscriber server in a Unity Connection cluster server pair rather than to the publisher server. Additional call traffic can be directed to the publisher server, if needed, but the call traffic should be directed to the subscriber server first.

Subscriber Server

When installing the subscriber server in a Unity Connection cluster server pair, you provide the IP address or hostname of the publisher server. After the software is installed, the subscriber server subscribes to the publisher server to obtain a copy of the database and message store. There can be only one subscriber server in a Unity Connection cluster server pair.

As a best practice, you should direct the majority of call traffic (for example, SCCP, SIP, or PIMG/TIMG) to the subscriber server in a Unity Connection cluster server pair. Additional call traffic can be directed to the

publisher server, if needed, but the call traffic should be directed to the subscriber server first. Most of the client traffic (for example, IMAP and the Cisco Personal Communications Assistant) and administration traffic (for example, Cisco Unity Connection Administration, the Bulk Administration Tool, and backup operations) should be directed to the publisher server in a Unity Connection cluster server pair. Additional client and administration traffic can be directed to the subscriber server, if needed, but the client and administration traffic should be directed to the publisher server first.

Requirements for Unity Connection Cluster

For current Unity Connection cluster requirements, see the System Requirements for Cisco Unity Connection *Release 15* at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/connection/15/requirements/b_15cucsysreqs.html.

Following are the requirements when both the servers in a cluster are in separate buildings or sites:

- Both servers must meet specifications according to the *Cisco Unity Connection 15 Supported Platforms List* at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/connection/15/supported_platforms/b_15cucspl.html.

- For a cluster with two virtual machines, both must have the same virtual platform overlay.

- Depending on the number of voice messaging ports on each Unity Connection server, the path of connectivity must have the following guaranteed bandwidth with no steady-state congestion:

  - For 50 voice messaging ports on each server—7 Mbps

  - For 100 voice messaging ports on each server—14 Mbps

  - For 150 voice messaging ports on each server—21 Mbps

  - For 200 voice messaging ports on each server—28 Mbps

  - For 250 voice messaging ports on each server—35 Mbps

Note The bandwidth numbers above 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 above and should be calculated using the bandwidth recommendations in the applicable *Cisco Unified Communications SRND* at http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns818/landing_uc_mgr.html.

- When both the subscriber and publisher are taking calls, the maximum round-trip latency must be not more than 100 ms. When only the publisher is taking calls, subscriber is idle but replicating with publisher, the maximum round-trip latency must be not more than 150 ms.

- The network must use the following load-balancing techniques for connections to the Unity Connection servers:

  - The Unity Connection servers are assigned a common DNS name with the Unity Connection publisher server first.

- All user client and administrator sessions connect to the Unity Connection publisher server. If the Unity Connection publisher server stops functioning, the user client and administrator sessions must connect to the Unity Connection subscriber server.

- Phone systems must attempt to route incoming calls to the Unity Connection subscriber server. If no voice messaging ports are available to answer calls on the Unity Connection subscriber server, the phone systems must route calls to the Unity Connection publisher server.

- The TCP and UDP ports of the firewall must be open as listed in the "IP Communications Required by Cisco Unity Connection" chapter of the Security Guide for Cisco Unity Connection, Release 15 available at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/connection/15/security/guide/b_15cucsecx.html.

- Both Unity Connection servers must have the same software and engineering-special versions installed.

- Both Unity Connection servers must have the same enabled features and configurations.

- Both Unity Connection servers must be configured to be in the same time zone.

- Both Unity Connection servers must connect to the same phone system(s).

- If the Unity Connection servers contain dual NICs, the two NICs on each Unity Connection server must be configured for fault tolerance using a single IP address, or one of the NICs must be disabled. Configuring the two NICs with distinct IP addresses for network load balancing is not supported.

- For selected servers supported for earlier versions of Unity Connection, a memory upgrade. To determine whether your server requires a memory upgrade, see the applicable server-specific table in the *Cisco Unity Connection 15 Supported Platforms List*.

- For selected servers supported for earlier versions of Unity Connection, replacement hard disks. To determine whether your server requires hard-disk replacement, see the applicable server-specific table in the *Cisco Unity Connection 15 Supported Platforms List*.

- The Unity Connection cluster feature is not supported for use with Cisco Business Edition.

Balancing the Load of Calls Unity Connection Servers Handle

Although it is possible to balance the load of calls that the Unity Connection servers handle in a Unity Connection cluster, most of the call traffic should be directed to the subscriber server. This configuration follows the Cisco Unified Communications Manager cluster model of allowing call traffic only on subscriber servers.

Cisco Unified Communications Manager by Skinny Client Control Protocol (SCCP)

When integrating Unity Connection with Cisco Unified CM by Skinny Client Control Protocol (SCCP), it is possible to balance the voice traffic that the Cisco Unity Connection server pair handles using one of the following methods:

- In Cisco Unified Communications Manager Administration (on the Call Routing > Route/Hunt > Line Group page), use Top Down as the distribution algorithm for the line group that contains directory numbers of ports that answer calls on both servers in the Unity Connection cluster.

In Unity Connection Administration, all the ports that share the same device name prefix are in a one port group. (If there are ports that share a different device name prefix, they must be in a separate port group.) Beginning with the answering port that has the lowest number in its display name, assign half the answering ports to the subscriber server so that the subscriber server answers most incoming calls. Assign the remaining answering ports to the publisher server. Then beginning with the dial-out port that has the lowest number in its display name, assign half the dial-out ports to the primary server so that the primary server handles MWIs and notification calls. Assign the remaining dial-out ports to the subscriber server.

- In Cisco Unified Communications Manager Administration (on the Call Routing > Route/Hunt > Line Group page), use Longest Idle Time as the distribution algorithm for the line group that contains directory numbers of ports that answer calls on both servers in the Unity Connection cluster.

In Unity Connection Administration, all the ports are in a single port group. The first half of the answering ports and dial-out ports are assigned to the publisher server and the remaining ports are assigned to the subscriber server in the Unity Connection cluster.

✎

Note You can use the following CLI command to configure the "Wait for Blind Transfer Ringing Timer" counter.

run cuc dbquery unitydirdb execute procedure
csp_ConfigurationModify(pFullName='System.Telephony.WaitForBlindTransferLongTimeoutMs',pvaluelong='1000')

For more information on the CLI commands, see the applicable Command Line Interface Reference Guide for Cisco Unified Communications Solutions at
http://www.cisco.com/en/US/products/ps6509/prod_maintenance_guides_list.html.

Cisco Unified Communications Manager through a SIP Trunk

When integrating with Cisco Unified CM through a SIP trunk, it is possible to balance voice traffic that the Unity Connection cluster server pair handles using one of the following methods:

- Use a Route List in Cisco Unified CM.

- Use DNS-SRV – RFC 2782.

- Use a SIP gateway DNS-SRV.

TDM-Based (Circuit-Switched) Phone System through PIMG/TIMG Units

When integrating with a TDM-based (circuit-switched) phone system through PIMG/TIMG units, it is possible to balance the load of voice traffic that the Unity Connection cluster server pair handles using one of the following methods:

- Turn on load balancing on the PIMG/TIMG units.

- Use load balancing on the TDM based PBX.

Note   You should turn on fault tolerance on the PIMG/TIMG units. This allows the PIMG/TIMG units to redirect calls to either server in the Unity Connection cluster if one server is unavailable to take calls.

Load Balancing Clients in a Unity Connection Cluster

Although it is possible to balance client and administration requests that the Unity Connection cluster server pair handles (for example, from the Cisco Personal Communications Assistant (PCA), IMAP, and Cisco Unity Connection Administration), most client and administration traffic should be directed to the publisher server.

In order to balance client requests, it is necessary to use DNS A-records. DNS A-records allow client DNS lookups to resolve to either server in a round-robin fashion.

Note   If one server in a Unity Connection cluster server pair stops functioning and failover occurs, clients such as the Cisco PCA and IMAP clients may need to authenticate again by signing in.

You should not be using DNS to load balance with multiple A-records because this method does not account for server unavailability (for example, if one of the servers in a Unity Connection cluster server pair stops functioning). The DNS server cannot determine the availability of a server IP address that is listed in an A-record. It may be necessary for the clients to attempt DNS resolution multiple times before they connect to a functioning Unity Connection server in a Unity Connection cluster server pair.

Configuration for Dial-Out Voice Messaging Ports

Each Unity Connection server in a cluster must have voice messaging ports designated for the following dial-out functions in case either server has an outage:

- Sending message waiting indicators (MWIs).

- Performing message notifications.

- Allowing telephone record and playback (TRAP) connections.

As a best practice, you should dedicate an adequate number of voice messaging ports for these dial-out functions. These dedicated dial-out ports should not receive incoming calls and should not be enabled for answering calls.

For More Information

- For configuring Unity Connection ports and port groups to support a cluster and the various phone system integrations, see the applicable Cisco Unity Connection integration guide at http://www.cisco.com/en/US/products/ps6509/products_installation_and_configuration_guides_list.html.

- For configuring Unity Connection Clients to support a cluster, see the "Configuring Cisco Unity Connection Cluster" chapter of the Install, Upgrade, and Maintenance Guide for Cisco Unity Connection, Release 15, available at https://www.cisco.com/c/en/us/td/docs/voice_ip_comm/connection/15/install_upgrade/guide/b_15cuciumg.html.