Configure IM and Presence Server High Availability

Contents

Introduction

This document describes how Instant Message and Presence (IM&P) High Availability works in an enterprise IM&P environment and how to troubleshoot it.

Prerequisites

Requirements

Cisco recommends that you have knowledge of these topics:

- Cisco Unified IM&P
- Cisco Jabber clients

Components used

- Cisco Unified IM&P 10.0 and later
- Cisco Jabber clients 9.6 and later

The information in this document was created from the components in a specific lab environment. All of the components used in this document started with a cleared (default) configuration. If your network is live, ensure that you understand the potential impact of any command.

IM and Presence High Availability (HA)

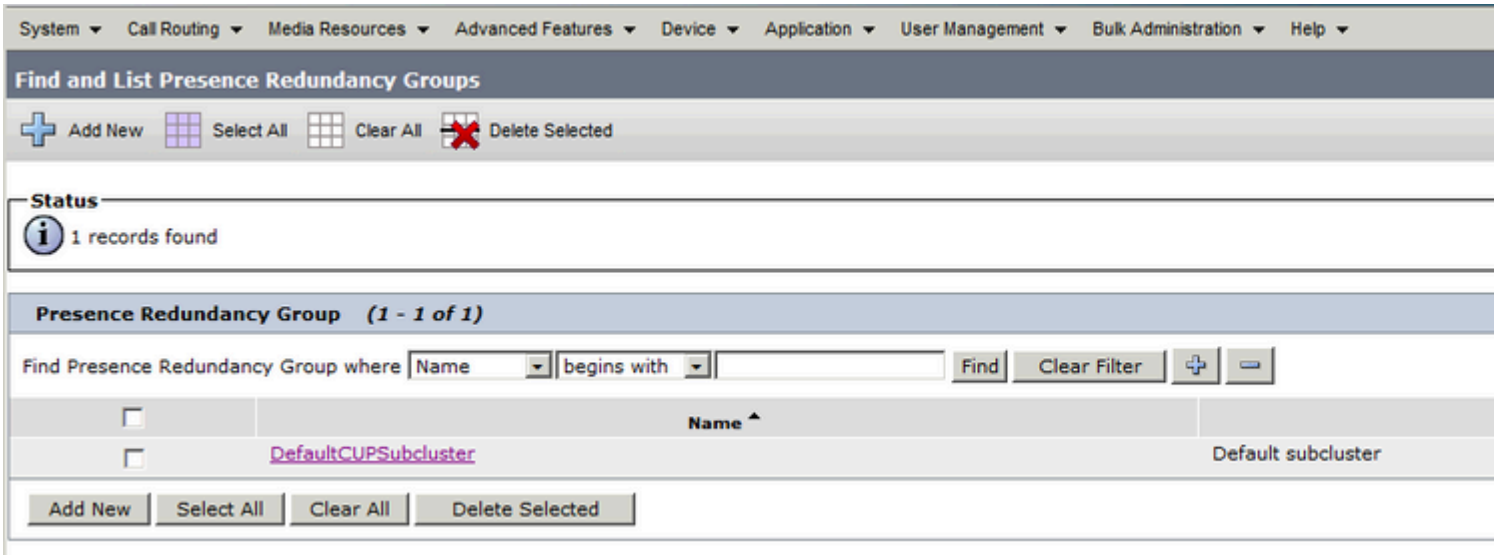
The IM and Presence Service Server offers high availability or redundancy in the form of logical server groups in the CUCM configuration. This configuration is passed to IM and Presence and then utilized to allow for redundancy in the event of an IM and Presence Service or server failure. When a HA event takes

place, the end user's sessions are moved from the failed server to the backup. When the server has been restored to a normal state, user sessions are then moved back either automatically or manually by the administrator.

Redundancy Group Configuration

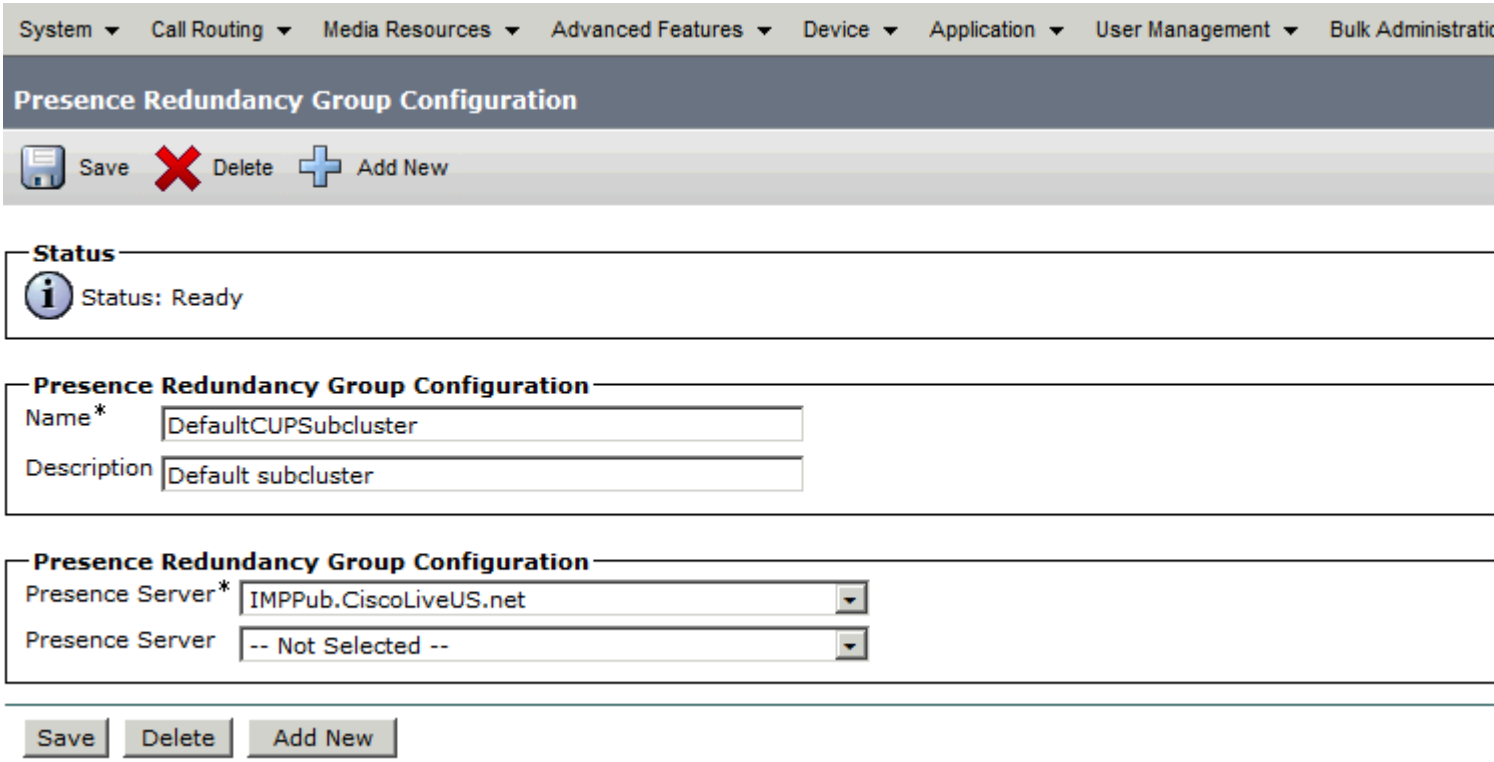
The redundancy group is the logical server pair that allows for the assignment of a server to the IM and Presence subcluster as well as the configuration for HA. In order to access this portion of the configuration, find it on the CUCM server web page.

System > Presence Redundancy Groups

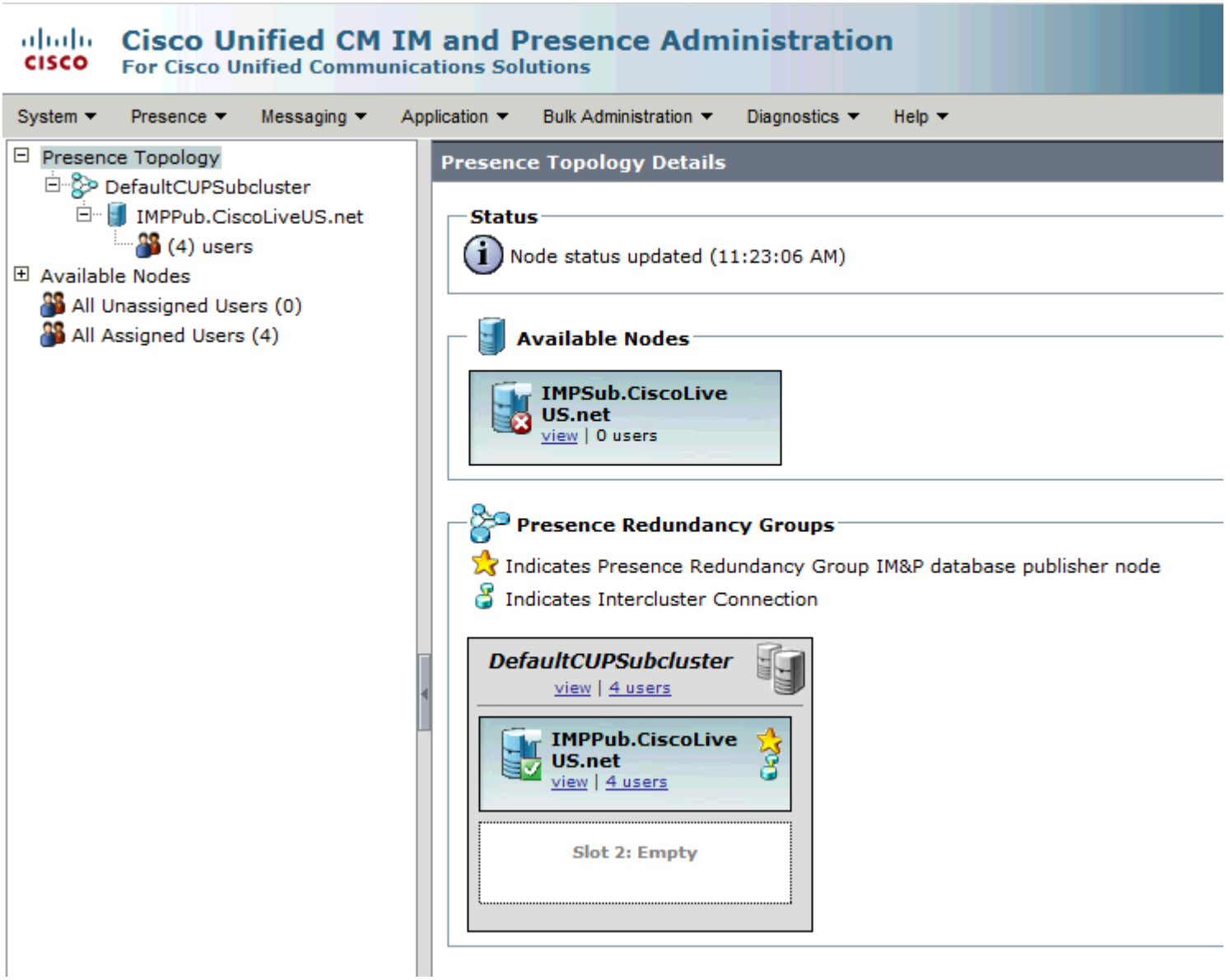


When the administrator adds the IM&P Publisher to the **System > Server** configuration on CUCM and the IM&P server is saved, the DefaultCUPSubCluster redundancy group gets created with the Publisher assigned to it.

When created, the Redundancy Group looks like this:



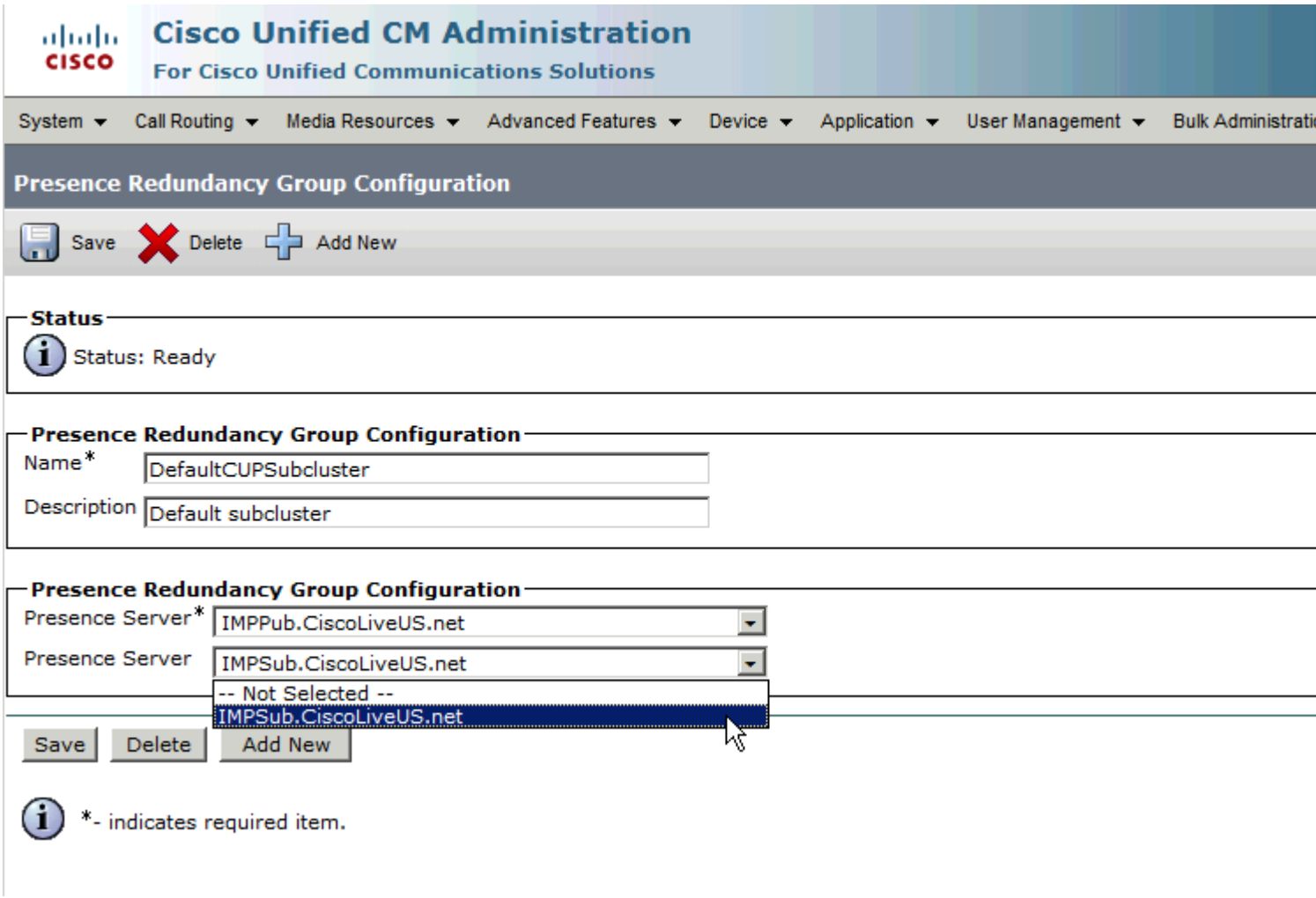
This Redundancy Group translates to the IM and Presence subcluster. In the current state of the Redundancy Group configuration in CUCM, this would be what it would look like in the IM and Presence Cluster Topology web page:



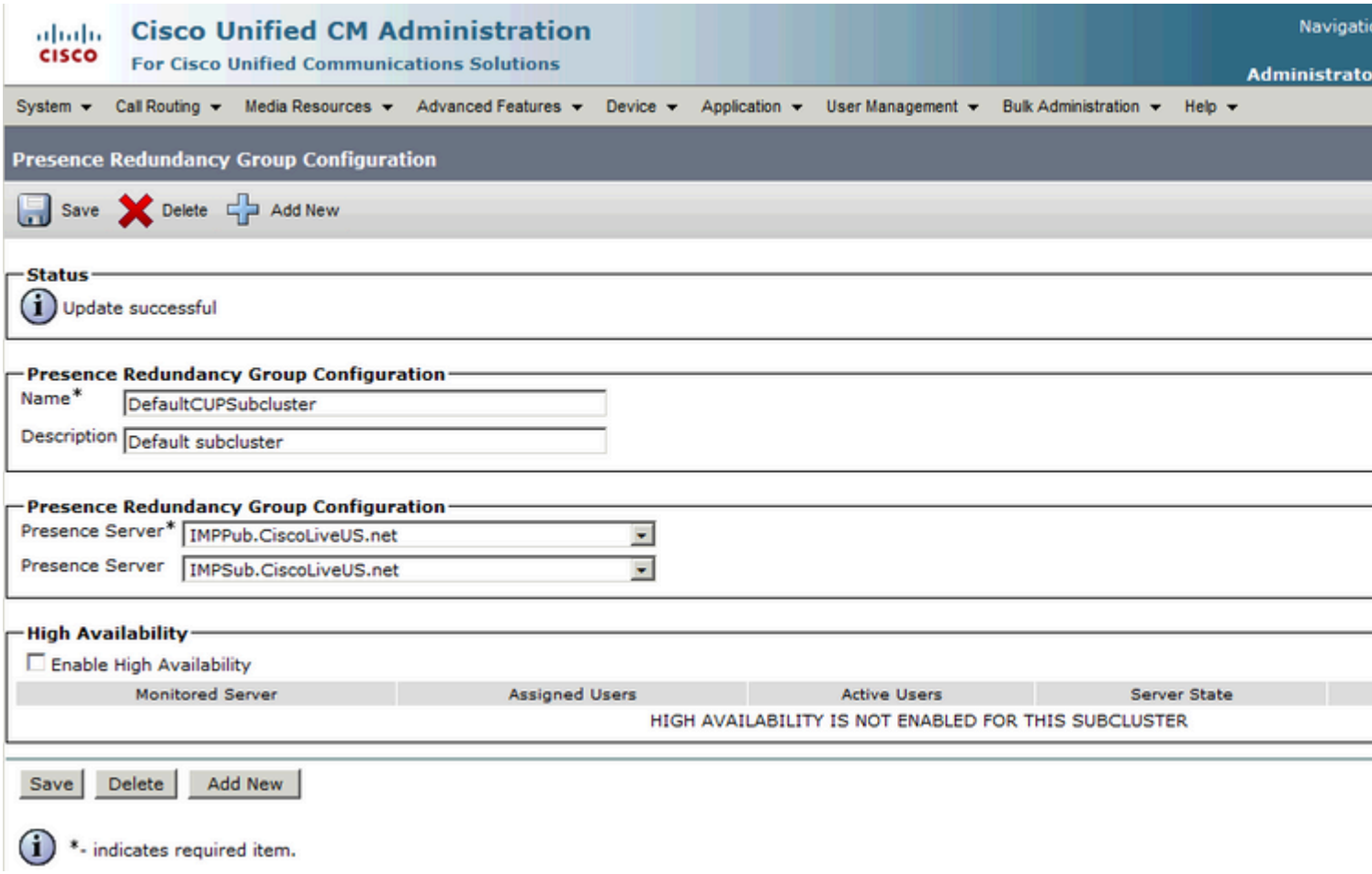
You see that the IM&P Publisher is assigned to the DefaultCUPSubcluster and the Subscriber server is not. This is because the IM&P Subscriber server is not assigned to the Redundancy Group in the CUCM configuration.

Assign the Subscriber to the Redundancy Group.

In order to assign the Subscriber server to the Redundancy Group, simply choose the subscriber server from the dropdown menu, then **Save** the configuration change.

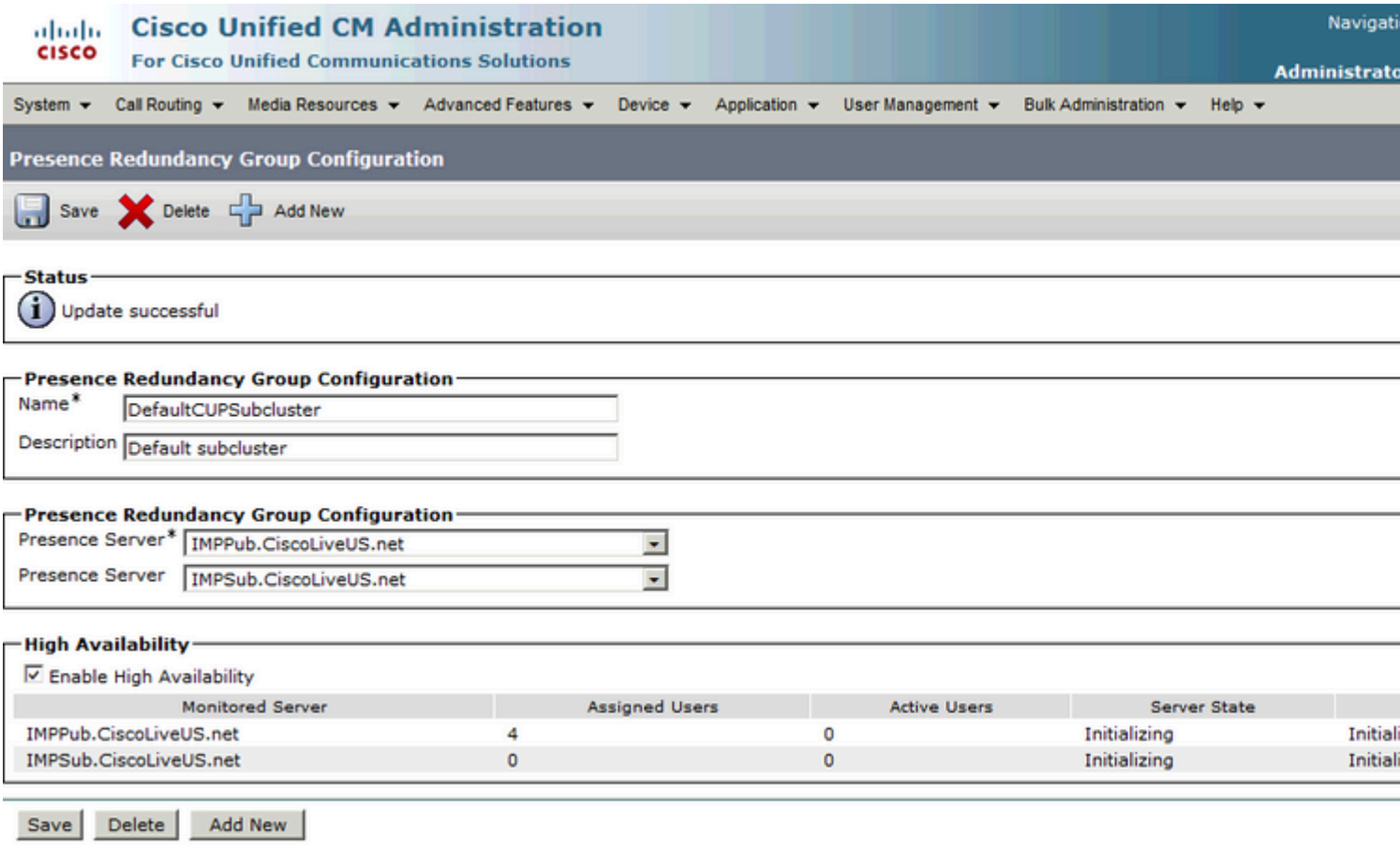


After the IM&P Subscriber is added to the Redundancy Group:



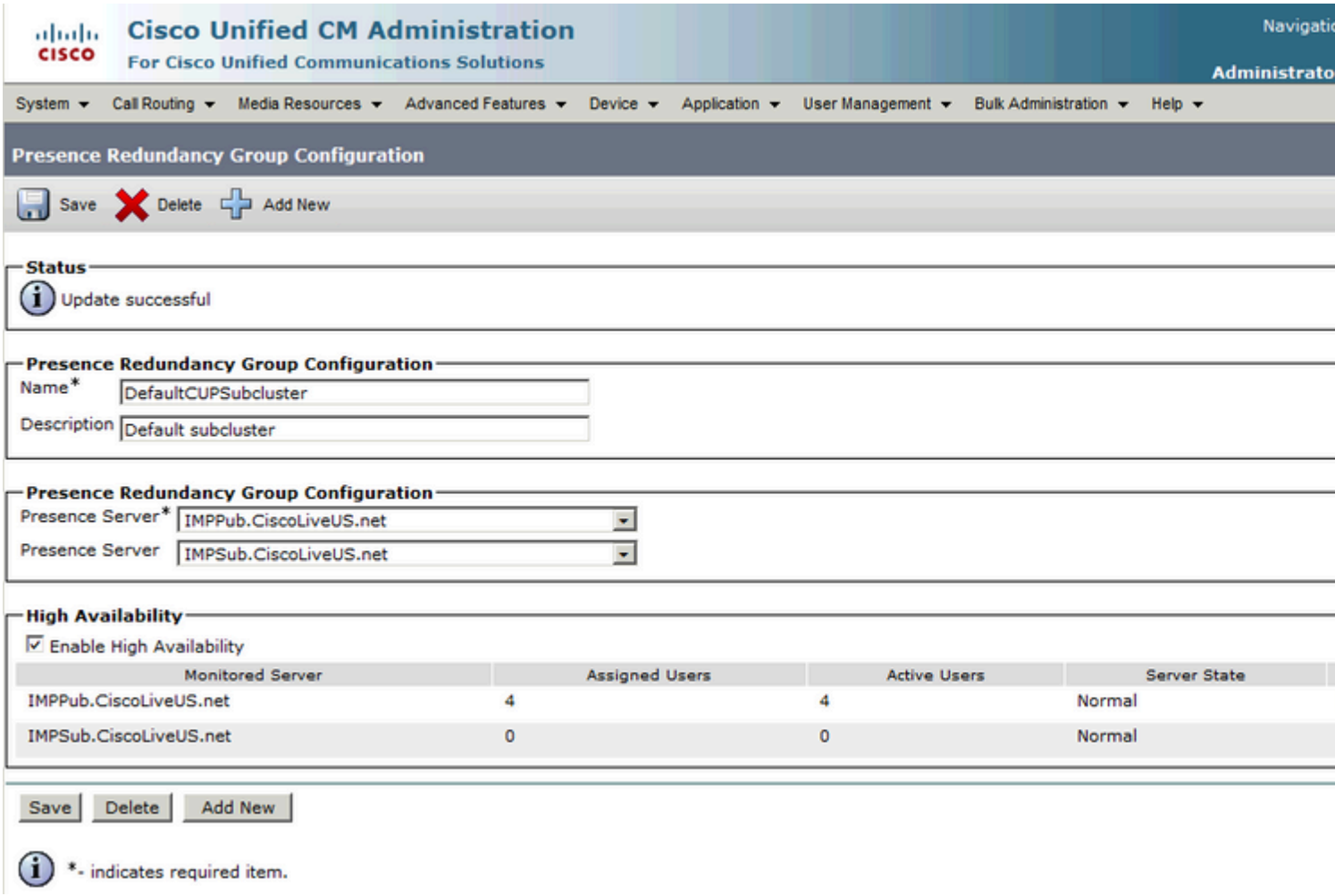
You see after the addition of the secondary node (the subscriber) that the High Availability option can be selected. In order to enable High Availability, you simply need to choose the **Enable High Availability** checkbox and **Save** the configuration change.

After High Availability is enabled:

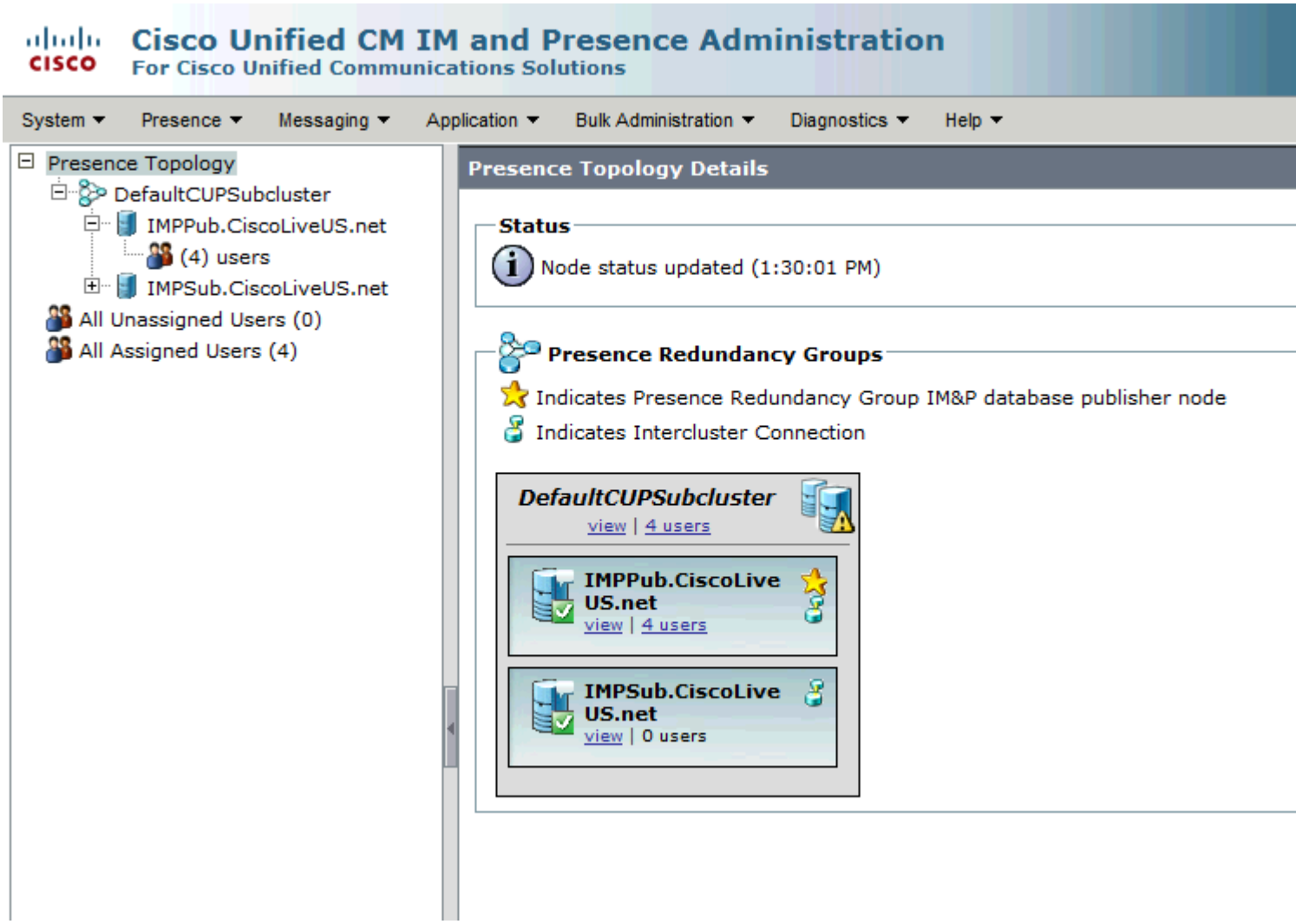


The page then auto-refreshes the server state and reason. When the server is in an initialization state, this means that the two servers are able to communicate. The servers would then verify service status before the state transitions to a Normal state. If the two servers can connect to each other and all monitored services are up on both, you would then get a Normal-Normal state. This means that all monitored services are active on the IM&P Servers.

Normal-Normal Redundancy Group State:



Normal-Normal High Availability State in the IM&P Topology Page:



Monitored IM and Presence Services

Since you could have various deployment models: IM Only, IM with SIP/XMPP Federation, IM with Compliance, IM with persistent chat, Remote Call Control Only, and so on, the actual list of which of these processes to monitor is dynamic. By default these items are always monitored when HA is enabled:

- IDS Database
- Presence Engine (if activated)
- XCP Router

The Server Recovery Manager checks to determine if compliance (Message Archiver), persistent chat (Text Conference Manager), SIP federation (SIP Federation Connection Manager), XMPP federation (XMPP Federation Connection Manager) are configured and activated.

If they are both configured and activated, the Server Recovery Manager(SRM) monitors those services as well.

> **Caution**: Before you proceed with a restart of one or more of the monitored services, you are required to disable the High Availability from the Presence Redundancy Groups on the CUCM server. The same applies when a reboot of one or more of the IM&P nodes is performed.

User Failover Process

When a failover takes place (automatic or manual), the major point to remember is that the user account is not moved from one server to the other, but only the user session in Presence Engine is moved. In pre-10 versions of IM and Presence, the user assignment was moved from one server to the other. This user move was very expensive to server resources, and added to the load that was on the server. In 10.X and later, the user stays homed on the server that they are assigned to, and the backend user session in the Presence Engine is moved from the failed node to the functional node. The user does not have to exit Jabber and re-log in when the change happens with Server Recovery Manager (SRM).

Jabber Client Re-Login Timer

In order for the user session to become fully active on the secondary IM&P node after a failover event, the user must attempt to log in to that server via SOAP (Client Profile Agent). This happens automatically with the one-time password that is passed from the IMDB database. Since log ins are extremely expensive to resources on the IM and Presence server, there must be a way to throttle log ins when a failover event occurs. This throttle or buffer allows all users to log in to the secondary node without service disruption for users on the secondary node. The mechanisms that are used to throttle user log ins are the Client Re-Login Lower Limit and Client Re-Login Upper Limit Server Recovery Manager (SRM) service parameters.

Client Re-Login Lower Limit - the parameter that defines the minimum amount of time (in seconds) that the Jabber client waits before the client attempts to log in to the secondary server in the event of an HA event.

Client Re-Login Upper Limit - the parameter that defines the maximum amount of time (in seconds) that the Jabber client waits before the client attempts to log in to the secondary server in the event of an HA event.

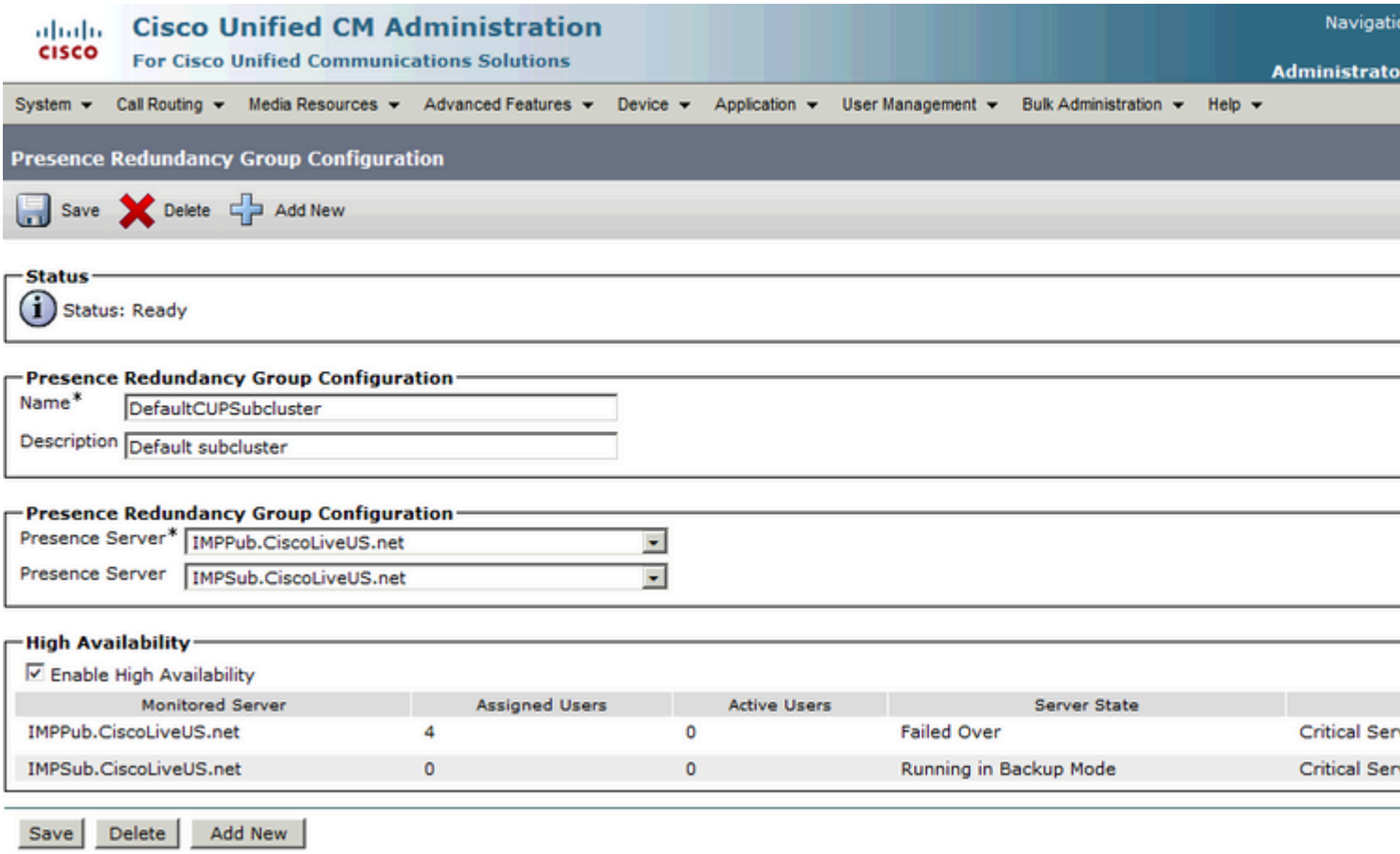
The Jabber client receives these parameters at log in to the server and caches the values for future use. When you receive a HA event from the IM&P server, the client chooses a random number of seconds between the upper and lower limits, and waits that amount of time before the Jabber client attempts to log in to the secondary. Once the timer expires, the client then attempts SOAP log in to the secondary node.

IM and Presence Fallback Types

If there is user failover, there must be user fallback when service is restored on the problematic server. There are two types of server fallback:

Manual Fallback

Manual fallback (default configuration for Server Recovery Manager) takes place when service has been restored and the redundancy group allows the Fallback button. When this button is selected, the user sessions that were moved to the secondary node are moved back to their homed node. The Jabber client then applies the re-log in upper and lower limits for the fallback.



Automatic Fallback

Automatic fallback takes place when the server monitors the services and the Server Recovery Manager (SRM) service automatically fallback users to their homed nodes. The key in this configuration is that the Server Recovery Manager (SRM) service waits 30 minutes for a failed service/server to remain active before an automatic fallback is initiated. Once this 30-minute uptime is established, user sessions are moved back to their homed nodes. The Jabber client then applies the re-log in upper and lower limits for the fallback.

> **Note**: Automatic fallback is not the default configuration, but it can be enabled. To enable automatic fallback, change the Enable Automatic Fallback parameter in the Server Recovery Manager Service Parameters to value True.

Troubleshoot

This section provides the information you can use in order to troubleshoot your configuration.

When troubleshooting High Availability on the IM&P Service Server, there are two important timers that you must keep in consideration.

- The Servers exchange 4 keepalives every 60 seconds. If there is no response after the 60 seconds, the Cisco Service Recovery Manager (SRM) considers that the unresponsive node went off-line and triggers a Fail Over command. As the next snippet shows, the last heartbeat occurred 62 seconds ago.

```
2021-05-13 02:48:48,244 INFO[HS]rsrm.RsrmHeartBeatHandler - RsrmHeartBeatHandler: peer down, time since
```

```
2021-05-13 02:48:48,244 INFO [HS] rsrm.RsrmAutomaticFallback - RsrmAutomaticFallback: peer states vecto
```

Tip: For this scenario, if you have found some latency in your network, it is recommended to increase the heartbeat timeout timer from 60 to 90 seconds.
Navigate to **CUCM Administration web page > System > Service parameters configuration > Select the IM&P Server> Select Cisco Recovery ManagerSettings.** On the Keep Alive (Heartbeat) timeout, increase the number to 90 seconds.

- The IM&P Subscriber server waits 90 seconds. If it detects that one or more of the monitored services is down, the Subscriber server takes over.

Logs to Collect for Troubleshooting

- Server Recover Manager (SRM) logs from before and after the failover event (debug level if possible).
- The output of the command via IM&P command-line interface **run sql select * from enterprisesubcluster**.
    - The enterprisesubcluster table in IM&P houses the Redundancy Group configuration.
- The output of the command via IM&P command-line interface **run sql select * from enterprisenode**.
    - The enterprisenode table displays the node information and subcluster assignment of the node.
- If the failed over is produced by a service being stopped, gather:
    - Event viewer system logs
    - Event viewer application logs
    - Logs from the service that are stopped.