MRA Maintenance

Maintenance Mode on the Expressway

Maintenance mode on the Expressway has been enhanced so that you can bring an MRA system down in a managed way.

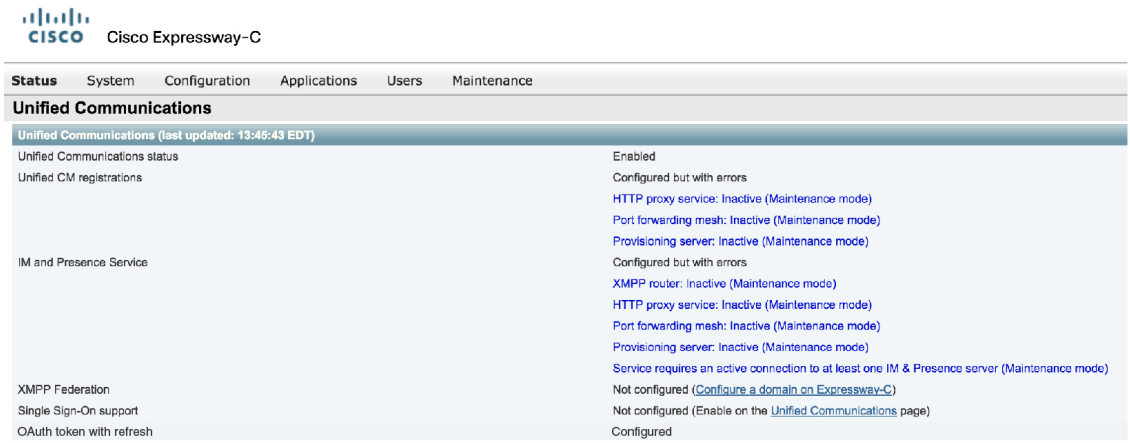
When you engage maintenance mode, the Expressway stops accepting new calls or proxy (MRA) traffic. Existing calls and chat sessions are not affected.

As users end their sessions normally, the system comes to a point when it is not processing any traffic of a certain type, and then it shuts that service down.

If users try to make new calls or start new chat sessions while the Expressway is in maintenance mode, the clients will receive a service unavailable response, and they might then choose to use another peer (if they are capable). This fail-over behavior depends on the client, but restarting the client should resolve any connection issues if there are active peers in the cluster.

The Unified Communications status pages also show (Maintenance Mode) in any places where MRA services are affected.

Figure 1: Maintenance Mode on Expressway-C



502281

Limitation for CE endpoints

Maintenance mode is not supported over MRA for endpoints running CE software. The Expressway drops MRA calls from these endpoints when you enable maintenance mode.

MRA Registration Counts

From X12.6.1 onward, the **Status** > **Overview** page on Cisco Expressway-E lets you monitor up-to-date usage information for SIP devices that are registered over MRA. The **Overview** page contains the following fields:

MRA Registration:

- **Current**—The total number of devices that are currently registered over MRA.
- **Peak**—The peak count for MRA registrations since the last Expressway restart.

Authorization Rate Control

The Expressway can limit the number of times that any user's credentials can be used, in a given configurable period, to authorize the user for collaboration services. This feature is designed to thwart inadvertent or real denial of service attacks, which can originate from multiple client devices authorizing the same user, or from clients that reauthorize more often than necessary.

Each time a client supplies credentials to authorize the user, the Expressway checks whether this attempt would exceed the **Maximum authorizations per period** within the previous number of seconds specified by the **Rate control period**.

If the attempt would exceed the chosen maximum, then the Expressway rejects the attempt and issues the HTTP error 429 “Too Many Requests”.

The authorization rate control settings are configurable in the **Advanced** section of the **Configuration** > **Unified Communications** > **Configuration** page.

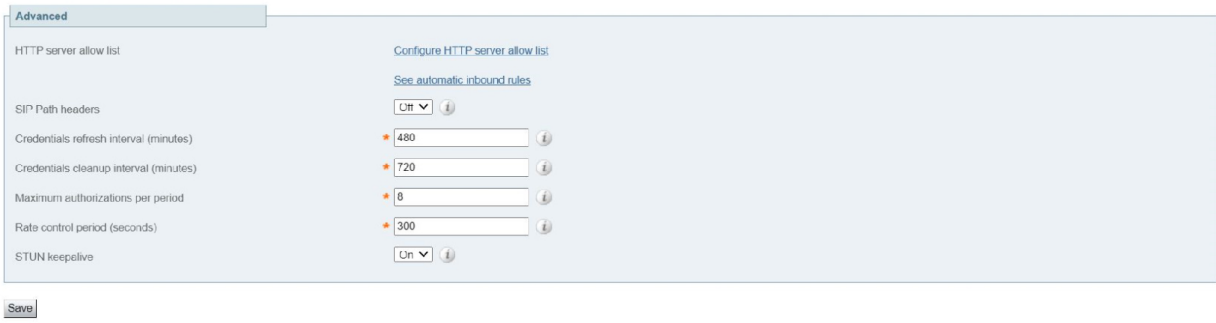
Credential Caching

Note These settings do not apply to clients that are using SSO (common identity) for authenticating via MRA.

The Expressway caches endpoint credentials which have been authenticated by Unified CM. This caching improves overall performance because the Expressway does not always have to submit endpoint credentials to Unified CM for authentication.

The caching settings are configurable in the **Advanced** section of the **Configuration** > **Unified Communications** > **Configuration** page.

Figure 2: Advanced Settings



Credentials refresh interval specifies the lifetime of the authentication token issued by the Expressway to a successfully authenticated client. A client that successfully authenticates should request a refresh before this token expires, or it will need to re-authenticate. The default is 480 minutes (8 hours).

Credentials cleanup interval specifies how long the Expressway waits between cache clearing operations. Only expired tokens are removed when the cache is cleared, so this setting is the longest possible time that an expired token can remain in the cache. The default is 720 minutes (12 hours).

SIP Registration Failover for Cisco Jabber

The SIP registration failover for Cisco Jabber applies if you deploy Expressway with Mobile and Remote Access (MRA).

Expressway X12.7 and later versions build on existing failover capabilities for clustered Expressways with a few MRA failover updates that improve substantially the failover time for Cisco Jabber clients that connect over MRA. Among the updates include adaptive routing, STUN keepalive support, and improved error reporting.

Note The registration failover feature uses STUN messages sent between the Unified CM and Expressway-C. This feature uses the same SIP connections along which SIP signaling messages traverse. In order to prevent filtering or removal of these STUN messages, disable SIP inspection on any firewall or Application Layer Gateway (ALG) device between Unified CM and Expressway C.

These new capabilities will allow Jabber clients to support MRA High Availability (failover) for voice and video.

Unified CM is able to resolve automatically added Expressway-C hostname

Unified CM does not respond to STUN requests when Expressway-C sends out STUN Keepalives on MRA SIP session.

Expressway-C nodes automatically add into Unified CM (under **Device** > **Expressway-C**) through the AXL API with the Expressway-C hostname (not FQDN) when the Unified CM is configured on Expressway-C for MRA solution.

Every 30 seconds, Expressway-C initiates MRA SIP session keepalive to Unified CM.

Before responding to the received keepalive, Unified CM tries to resolve the hostname of Expressway-C. If Unified CM fails to resolve through DNS it does not respond to STUN keepalive requests. This flaps the MRA SIP registration.

If Unified CM and Expressway-C are in different domains, make sure that the Unified CM can resolve the hostname of Expressway-C.

Adaptive routing

Adaptive routing updates in Expressway X12.7 and later versions allows Expressway to alter the routing path dynamically. If a node failure is detected, packets are rerouted to a peer node that is up and running. For example, assume that a remote Jabber client sends a SIP REGISTER that is intended to be routed through a specific Expressway-E (EXWY-E1), Expressway-C (EXWY-C1) and Unified CM (CUCM1) combination, but the designated Expressway-C node is either down or is in maintenance mode. In this case, the message is rerouted to a peer Expressway-C node (EXWY-C2) and then on to the intended Unified CM destination. After the registration, Cisco Jabber also updates its routing table so that future SIP messages use the registration path.

Note

- Failover does not include call preservation. The Jabber registration fails over to the new registration path, but active calls at the time of the failure are dropped.

STUN keepalive support

In addition to adaptive routing, Expressway X12.7 and later versions support the use of STUN keepalives by MRA connected Jabber clients. Remote Jabber clients send STUN keepalives into the enterprise network via Expressway-E to learn of connection issues ahead of time. As a result, if a node in the registration path fails, Jabber will learn of the failure after receiving the STUN response and can select a different route path for future SIP messages.

Settings

The STUN keepalive setting is configurable in the **Advanced** section of the **Configuration** > **Unified Communications** > **Configuration** page. See Figure 2: Advanced Settings.

Field	Description
STUN keepalive	Enable STUN keepalive for Unified CM High Availability. Default: *On*

Requirements

No specific configuration is required (subject of course to the necessary clustering/backup nodes existing). However, you must be running the following minimum releases:

Routing Feature	Minimum Releases Required
Adaptive routing	**1.** Expressway X12.7 **2.** Cisco Jabber 12.9 MR **3.** Cisco Webex App
STUN keepalives	**1.** Expressway X12.7 **2.** Cisco Unified Communications Manager 14 **3.** Cisco Jabber 12.9 MR **4.** Cisco Webex App

Note

- STUN keepalive is sent every 30 seconds from the client (Jabber) and if it didn't get the response within 3 seconds, then the client initiates failover.
- When Expressway is configured with a different domain from Unified CM, the Unified CM admin needs to update Expressway-C Hostname entry manually to FQDN, by appending the relevant system domain of Expressway-C.

Load Balance After Node Recovery

With MRA-HA, whenever there is a node(s) failure the load of the failed node(s) will be shifted to the other available nodes in the cluster. The following sections describe the load balancing procedure after the node(s) became active in the cluster.

Load Balance Expressway-C nodes

From X14.1 release, Expressway-C node load balancing is achieved by using Adaptive Routing on Expressway-E node.

After an Expressway-C node failure, the traffic/registrations will be handled by other nodes in the cluster. Once the failure node gets recovered and becomes active, even though new registrations go through that node, the existing load won't be handled by that node. To load balance the Expressway-C cluster in that scenario, we are introducing AR mechanism on Expressway-E.

There is a keep alive mechanism between Expressway-E nodes and Expressway-C nodes, in a mesh architecture. Within the keep alive message, Expressway-C sends resource usage/active registrations to Expressway-E. Then, Expressway-E evaluates the active registrations across all the nodes in Expressway-C and if it identifies an unbalanced load on the node, it triggers load balancing.

The load balancing is achieved by adaptively routing the Register messages (New/Refresh) to least loaded node. This will be done to the clients which supports Adaptive Routing. Once the load is balanced Expressway-E will stop the process. This ensures no node is idle and load is balanced.

Load Balance Expressway-E nodes

Expressway-E node maintains the count of total number of registrations of all nodes in the cluster. Whenever there is an imbalance in the cluster, the node with high number of registrations will respond to register messages with a warning header in the 200-response message, indicating load is imbalanced.

Note The load balance will not be shared equally or in a fixed ratio but will try to avoid the 0-100 share situation for a node.

Benefits with all software requirements

When all three components - clients, Expressway, Unified CM - are running updated software with MRA registration failover capabilities, the following benefits apply:

- No user action required for failover
- Faster failover times - down to 30-60 seconds from the previous standard of 120 seconds
- Route path updates dynamically to handle server failures
- More routes are available to reach the intended destination
- Remote Jabber clients can learn of server failures via STUN keepalives and adjust routing ahead of time

Adaptive routing benefit without Unified CM upgrade

Even without new Unified CM software (but with new Expressway and Jabber software), this feature has the benefit of allowing Jabber clients to detect path failures.

Note This action will take over 2 minutes, and Expressway may flag Unified CM servers as inactive in some scenarios where actually the server is just idle or has low use at the time.

Clustered Expressway Systems and Failover Considerations

You can configure a cluster of Expressway-Cs and a cluster of Expressway-Es to provide failover (redundancy) support as well as improved scalability.

Details about how to set up Expressway clusters are contained in Expressway Cluster Creation and Maintenance Deployment Guide and information about how to configure Jabber endpoints and DNS are contained in “Configure DNS for Cisco Jabber”.

Note that when discovering Unified CM and IM and Presence Service servers on Expressway-C, you must do this on the primary peer.

Expressway Automated Intrusion Protection

From X8.9 onwards, automated intrusion protection is enabled, by default, for the following categories:

- http-ce-auth
- http-ce-intrusion
- sshpfwd-auth
- sshpfwd-intrusion
- xmpp-intrusion

This change affects new systems. Upgraded systems keep their existing protection configuration.

On Expressway-C

The Expressway-C receives a lot of inbound traffic from Unified CM and from the Expressway-E when it is used for Mobile and Remote Access.

If you want to use automated protection on the Expressway-C, you should add exemptions for all hosts that use the automatically created neighbor zones and the Unified Communications secure traversal zone. The Expressway does not automatically create exemptions for discovered Unified CM or related nodes.

On Expressway-E

You should enable the Automated protection service (**System** > **System administration**) if it is not yet running.

To protect against malicious attempts to access the HTTP proxy, you can configure automated intrusion protection on the Expressway-E (**System** > **Protection** > **Automated detection** > **Configuration**).

We recommend that you enable the following categories on the Expressway-E:

- HTTP proxy authorization failure and HTTP proxy protocol violation. Do not enable the HTTP proxy resource access failure category.
- XMPP protocol violation

Note The Automated protection service uses Fail2ban software. It protects against brute force attacks that originate from a single source IP address.

Configure Exemptions

If you have Automated Intrusion Protection configured, use this procedure to configure exemptions for IP address ranges from one or more protection categories.

One example where you may need an exemption is if you have multiple MRA users connected behind a NAT using the same public IP address. This may trigger protection due to the incoming traffic from the single IP address.

Note This procedure assumes you have the Automated Intrusion Protection enabled on Expressway-E and disabled on Expressway-C, which is the recommended deployment.

Step 1 On Expressway-E, go to **System** > **Protection** > **Automated detection** > **Exemptions**.

Step 2 Click on the **Address** that you want to configure or click **New** to configure a new address.

Step 3 Enter the **Address** and **Prefix Length** to define the IP address range that you want to exempt.

Step 4 Select from the categories to which you want to apply the exemption. For the example situation where you have multiple users behind a NAT, the following categories would apply:

- HTTP Proxy Authentication Failure
- HTTP Proxy Resource Access Failure
- SIP Authentication Failure

Step 5 Click **Add Address**.

Check the Unified Communications Services Status

You can check the status of the Unified Communications services on both Expressway-C and Expressway-E.

Step 1 Go to **Status** > **Unified Communications**.

Step 2 Review the list and status of domains, zones and (Expressway-C only) Unified CM and IM and Presence Service servers.

The page displays any configuration errors along with links to the relevant configuration page that you access to address the issue.

Why You Need to Refresh the Discovered Nodes?

When the Expressway-C discovers a Unified Communications node, it establishes a connection to read the information required to create zones and search rules to proxy requests originating from outside of the network in towards that node. **This configuration information is static.** Expressway only reads it when you manually initiate discovery of a new node, or when you refresh the configuration of previously discovered nodes. If any related configuration has changed on a node after you discover it, the mismatch between the new configuration and what the Expressway-C knows of that node is likely to cause some kind of failure.

The information that the Expressway-C reads from the Unified Communications node is different for each node type/role. These are examples of UC configuration that you can expect to require a refresh from the

Expressway. The list is not exhaustive. If you suspect that a configuration change on a node is affecting MRA services, you should refresh those nodes to eliminate one known source of potential problems.

- Changing cluster (such as adding or removing a node)
- Changing security parameters (such as enabling Mixed Mode)
- Changing connection sockets (such as SIP port configuration)
- Changing TFTP server configuration
- Upgrading node software

Devices cannot connect during the refresh

It takes some time to restore services after a server refresh and while the refresh is in progress, Jabber clients and other endpoints are unable to connect over MRA. It is not possible to provide accurate timings as they vary depending on the deployment. For straightforward deployments the refresh typically takes 5 to 10 seconds, but very complex configurations may take upwards of 45 seconds.

Refresh Servers on the Expressway-C

You must refresh the Cisco Unified Communications Manager and Cisco Unity Connection nodes defined on the Expressway-C. This fetches keys that the Expressway needs to decrypt the tokens.

Step 1 For Unified CM, go to **Configuration** > **Unified Communications** > **Unified CM servers** and click **Refresh servers**.

Step 2 For Unity Connection, go to **Configuration** > **Unified Communications** > **Unity Connection servers** and click **Refresh servers**.