Cisco Meeting Server

Load Balancing Calls Across Cisco Meeting Servers

White Paper

23 May 2019
Contents

Change History .................................................................................................................. 4

Introduction ....................................................................................................................... 5

Load balancing across Cisco Meeting Servers ................................................................. 6
   Call Bridge Groups ......................................................................................................... 6
   Configuring Call Bridges for load balancing incoming calls ......................................... 7
      Creating Call Bridge Groups ....................................................................................... 7
      Specifying the load limit on a cluster and enabling load balancing ....................... 8
      Fine-tuning the load balancing .................................................................................. 9
      How load balancing uses the settings ...................................................................... 9
   Load balancing outbound SIP calls ................................................................................ 10
      How to enable load balancing of outbound SIP calls ............................................. 11
      How to set up an outbound dial plan rule for load balancing outbound SIP calls ...... 11
      How to supply the Call Bridge Group or specific Call Bridge to use for outbound SIP calls to participants ............................................. 11
      Handling load balancing of active empty conferences ......................................... 11
   Load balancing Cisco Meeting App calls ..................................................................... 12
      Disabling load balancing Cisco Meeting App participants ....................................... 13

Example deployments of load balancing incoming calls using Cisco Expressway .......... 14
   Deployment using neighbor zones to reach remote Meeting Servers (example 1) .......... 14
      Dial plan configuration ............................................................................................. 15
      Load balancing calls over local Call Bridges ........................................................... 15
      Configuring failover to remote Call Bridges via Cisco Expressway clusters ........... 15
      Call flow for calls load balanced between servers in the same location ............... 16
      Call flow for calls redirected within a location ....................................................... 17
      Call flow for calls redirected between regions ...................................................... 18
   Deployment with zones direct to remote Call Bridges (example 2) ......................... 19
      Dial plan configuration ............................................................................................. 20
      Load balancing calls over local Call Bridges ........................................................... 20
      Configuring failover to remote Call Bridges via direct connections ..................... 21
      Call flow for calls load balanced between servers in the same location ............... 21
      Call flow for calls redirected within a location ....................................................... 22
      Call flow for calls redirected between regions ...................................................... 23

Example deployments of load balancing incoming calls using Cisco Unified Com
munications Manager .......................................................... 26
Deployment uses an SME for centralized call routing between Cisco Unified Com-
munications Managers (example 1) ........................................... 26
Dial plan configuration ............................................................ 27
Load balancing calls over local Call Bridges ................................. 27
Configuring failover to remote Call Bridges using an SME ................. 28
Call flow for calls load balanced between servers in the same location ........... 29
Call flow for calls redirected within a location .................................. 30
Call flow for calls redirected between regions ................................... 31
Using Local Route Groups ........................................................ 33
Deployment with centralized Meeting Servers and an SME to route calls (example 2) ... 37
Dial plan configuration ............................................................ 37
Call flow for a call being load balanced ........................................ 38
Call flow for call being redirected ............................................... 39
Deployment without an SME to route calls (example 3) ....................... 41
Dial plan configuration ............................................................ 42
Load balancing calls over local Call Bridges .................................. 42
Configuring failover to remote Call Bridges without using an SME .......... 43
Call flow for a call load balanced between servers in the same location ........... 44
Call flow for a call redirected within a location .................................. 44
Call flow for a call redirected between regions ................................... 45
Examples of load balancing outbound calls ..................................... 47
Load balancing outbound calls with local dial plan rule ....................... 47
Load balancing outbound calls with no local dial plan rule ..................... 47
Load balancing outbound calls with explicit choice of Call Bridge Group .......... 48
Load balancing outbound calls with explicit choice of Call Bridge ................ 49
Appendix A  Configuring a Cisco Expressway dial plan .......................... 50
Appendix B  Setting up Local Route Groups on Cisco Unified Communications Manager . 53
Appendix C  Growth in scaling deployments ...................................... 64
Cisco Legal Information ........................................................... 67
Cisco Trademark ........................................................................ 68
## Change History

<table>
<thead>
<tr>
<th>Date</th>
<th>Change Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 23, 2019</td>
<td>Minor addition</td>
</tr>
<tr>
<td>January 08, 2019</td>
<td>Minor corrections to appendix on scaling deployments</td>
</tr>
<tr>
<td>January 02, 2019</td>
<td>Added appendix on growth in scaling deployments. No changes for version 2.5.</td>
</tr>
<tr>
<td>September 25, 2018</td>
<td>Minor correction</td>
</tr>
<tr>
<td>September 21, 2018</td>
<td>Added load balancing incoming calls to clustered Meeting Servers using Expressway as the call control (from version 2.4)</td>
</tr>
<tr>
<td>January 22, 2018</td>
<td>Minor corrections.</td>
</tr>
<tr>
<td>December 18, 2017</td>
<td>Added load balancing Cisco Meeting App calls (from version 2.3)</td>
</tr>
<tr>
<td>May 8, 2017</td>
<td>Added load balancing outbound calls (from version 2.2)</td>
</tr>
<tr>
<td>December 20, 2016</td>
<td>First version covering load balancing incoming calls (from version 2.1)</td>
</tr>
</tbody>
</table>
Introduction

The Cisco Meeting Server software is a scalable software platform for voice, video and web content.

This white paper discusses how to increase the scalability and resilience of the Meeting Servers within a deployment, using Call Bridge grouping to load balance incoming and outgoing calls across clustered Call Bridges. Load balancing is achieved by trying to place calls for a single conference onto as few Call Bridges as possible. This reduces the number of distribution links required to connect the participants in the conference, and therefore reduces the overall load across the system.

Load balancing occurs via two mechanisms. Firstly calls to and from existing conferences are preferentially handled by Call Bridges currently hosting the conference. Secondly, the choice of the first Call Bridge, and subsequent Call Bridges, is done to preferentially use those with lower load.

The examples include using Cisco Unified Communications Manager and Cisco Expressway as the call control device. The primary role of call control is to move SIP calls between Call Bridge Groups as instructed by the Cisco Meeting Servers. Only these two call control systems are supported for load balancing SIP calls.

This white paper also covers load balancing media to and from the Cisco Meeting App. A call control device is not required for load balancing calls in deployments where only Cisco Meeting App is used to make calls (no SIP calls).

Note: Throughout this white paper, the Cisco Meeting Server is referred to as the Meeting Server, and the term Expressway refers to the Cisco VCS or Cisco Expressway running X8.11 or later software.
Load balancing across Cisco Meeting Servers

A typical large scale deployment consists of several Meeting Servers deployed at multiple offices/data centers. For scalability and resilience of the conferencing service, the Call Bridges will typically be configured as a cluster.

This white paper explains how to use Call Bridge grouping to load balance incoming and outbound calls on the Meeting Servers, and avoid overloading individual Meeting Servers in the cluster.

With Call Bridge Groups configured, a Meeting Server cluster can intelligently load balance calls across the Call Bridges within the same location or across nodes in different locations. The intelligent decision making behind where calls end up, is handled by the Meeting Servers. The call control system needs to be able to handle SIP messages from the Meeting Servers, in order to move calls to the correct location. This functionality has been tested using Cisco Unified Communications Manager and Cisco Expressway as call control systems. These are the only Cisco supported call control systems for this functionality. For load balancing with Cisco Expressway, use Cisco Expressway release X8.11 or later with Cisco Meeting Server release 2.4 or later. This white paper covers five scenarios in detail, the technology also works for other scenarios as well; the techniques below can be adapted to multiple Cisco Unified Communications Manager or Cisco Expressway topologies.

**Note:** There are different call capacities for Meeting Servers in a Call Bridge Group compared to a single or cluster of Meeting Servers. Appendix C provides an overview of the difference in call capacities.

**Note:** Dual-homed conferences with a Meeting Server cluster are not currently supported with Expressway X8.11 as the edge for the Meeting Server, unless at least some of the Microsoft traffic flows directly between one of the Meeting Servers in the cluster and the Microsoft infrastructure (and not through Expressway). Dual-homing is supported with Expressway X8.11 as the edge for standalone Meeting Servers.

**Call Bridge Groups**

The load balancing of calls occurs between a group of Call Bridges that exist in the same location. To configure which Call Bridges are in each location, the concept of Call Bridge Groups is used. A Call Bridge Group defines a subset of cluster nodes that are more closely linked and should be treated as equivalent. This could refer to those in a single data center, or those in the same continent. The decision of how to group Call Bridges will depend on the specifics of network configuration and the desired behavior.
For the load balancing feature to work correctly, a Round Trip Time (RTT) of less than 100 ms is required for the servers in a Call Bridge Group. The maximum RTT between any two nodes in the same cluster remains as 300 ms.

When using Cisco Unified Communications Manager, call routing relies on the use of route patterns, route groups and route lists across the Cisco Unified Communications Manager deployment. It is assumed that these concepts are understood. For information on configuring these features, please consult the Cisco Unified Communications Manager documentation.

When using Cisco Expressway, call routing relies on the use of dial plans and zones, and it is assumed that these concepts are understood. For information on configuring dial plans and zones, please consult the Cisco Expressway documentation.

**Configuring Call Bridges for load balancing incoming calls**

There are three aspects to setting up the load balancing of calls across a Meeting Server cluster:

- creating the Call Bridge Groups,
- enabling load balancing,
- and optionally, fine-tuning the load balancing on each Call Bridge. In most deployments this will not be necessary.

In addition, load balancing incoming calls involves outbound calls from Call Bridges to Cisco Unified Communications Manager or Cisco Expressway. For these outbound calls to work, outbound dial plan rules must be configured, see Load balancing outbound SIP calls.

**Note:** If load balancing incoming calls involves outbound calls from Call Bridges to Cisco VCS, instead of Cisco Expressway, then a traversal license is required on the VCS. There is no requirement for a Rich Media Session license on Cisco Expressway for any load balanced Meeting Server deployments.

**Note:** If you are not using load balancing with Call Bridge Groups, then calls will not be rejected, but the quality of all calls will be reduced when the load limit is reached. If this happens often, we recommend that you buy additional hardware.

**Creating Call Bridge Groups**

1. For each Meeting Server cluster, decide how to group the Call Bridges, for instance by data center or by country or region.
2. Create a Call Bridge Group using a POST to the `/callBridgeGroups` with a name for the Call Bridge Group.
3. Assign a Call Bridge to a group by setting the `callBridgeGroup` parameter for that Call Bridge under `/callBridges`. Set the `callBridgeGroup` parameter of an individual Call Bridge to the GUID of the target `callBridgeGroup` you are assigning it to.

4. Repeat for all other Call Bridge Groups.

**Specifying the load limit on a cluster and enabling load balancing**

1. On each Call Bridge, issue a PUT to the `/system/configuration/cluster` node with the value for `loadLimit`. This sets a load limit for the maximum load on the server, see the table below for load limits:

<table>
<thead>
<tr>
<th>System</th>
<th>Load Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting Server 2000</td>
<td>500000 (see note below)</td>
</tr>
<tr>
<td>Meeting Server 1000</td>
<td>96000</td>
</tr>
<tr>
<td>X3</td>
<td>250000</td>
</tr>
<tr>
<td>X2</td>
<td>125000</td>
</tr>
<tr>
<td>X1</td>
<td>25000</td>
</tr>
<tr>
<td>VM</td>
<td>1250 per vCPU</td>
</tr>
</tbody>
</table>

Setting a load limit on any Call Bridge means it will reject calls based on the current load. By default, the rejection of calls from new participants occurs at 80% of the load limit to allow for the distribution of calls. This value can be fine tuned, see below.

**Note:** Although from version 2.4, a single Cisco Meeting Server 2000 supports an increased number of HD calls (700 from 500), a Cisco Meeting Server 2000 with Call Bridge Groups enabled only supports 500 HD calls. The `loadLimit` for a Cisco Meeting Server 2000 is set at 500,000 due to the total number of video calls supported remaining at 1,000. The Meeting Server does not offer a way to limit the minimum resolution of calls, so it will always be possible to have a mix of HD and lower resolution calls. There is no mechanism to make use of the increased scale in HD or full HD in a load balanced configuration.

2. Enable load balancing.
   a. For Cisco Unified Communications Manager deployments:
      
      PUT to the `/callBridgeGroups/<call bridge group>` with `loadBalancingEnabled=true`
   
   b. For Cisco Expressway deployments:
      
      PUT to the `/callBridgeGroups/<call bridge group>` with `loadBalancingEnabled=true` and
      
      PUT to the `/callBridgeGroups/<call bridge group>` with `loadBalanceIndirectCalls=true`
Tip: If you have only one Call Bridge, and you want to reject calls rather than reducing quality, then create a Call Bridge Group with a single Call Bridge and enable load balancing.

Fine-tuning the load balancing

It is possible to fine tune the load balancing parameters. This should be done with care, since it could impact the availability of the solution. Changing the default values could lead to overloading of servers and a degradation of video quality. This could occur due to either conferences becoming fragmented over multiple Call Bridges, or conferences using too many resources on a single Call Bridge.

Load balancing calls on a Call Bridge is controlled by 3 parameters:

- **loadLimit** – a numeric value for the maximum load on the Call Bridge, as set above.
- **newConferenceLoadLimitBasisPoints** – a numeric value for the basis points (1 in 10,000) of the load limit at which incoming calls to non-active conferences will be disfavored, ranges from 0 to 10000, defaults to 5000 (50% load). Value is scaled relative to **LoadLimit**.
- **existingConferenceLoadLimitBasisPoints** – a numeric value for the basis points of the load limit at which incoming calls to this Call Bridge will be rejected, ranges from 0 to 10000, defaults to 8000 (80% load). Value is scaled relative to **LoadLimit**.

To change the default threshold values for a Call Bridge:

1. Issue a PUT to the `/system/configuration/cluster` node with the value for **newConferenceLoadLimitBasisPoints** and the value for **existingConferenceLoadLimitBasisPoints**.

**Note:** distribution calls are always accepted, and will consume additional resources. If modifying the load balancing parameters, ensure that any necessary overhead for these calls has been included in the calculations.

How load balancing uses the settings

Within each Call Bridge Group there is a particular preference order in which Call Bridges will be chosen for each space. Any call for a space landing anywhere in the Call Bridge Group will be preferentially redirected to Call Bridges based on this order. The redirection is based on two thresholds: the existing conference threshold and the new conference threshold.

The thresholds are defined as:

existing conference threshold = existingConferenceLoadLimitBasisPoints/10000 × loadLimit
new conference threshold = newConferenceLoadLimitBasisPoints/10000 × loadLimit
When a call lands on a Call Bridge the local load limit is checked. If this is above the existing conference threshold, then the call is rejected. Note calls can also be rejected for other reasons. Rejected calls should be redirected by the call control device.

If the load is below this threshold, then the call will be answered and any IVRs traversed. Once the conference is known then the Call Bridge preference order within the group can be determined. This order is used to decide between Call Bridges in cases where there are multiple Call Bridges that could be chosen.

If any Call Bridges within the group are already running the conference, then the load of these Call Bridges are checked. If any of these are below the existing conference threshold, then one of these will be used.

If the conference isn’t running anywhere, or all Call Bridges currently running this conference are above the existing conference threshold, then the load of other Call Bridges are checked. If the load level of any of these Call Bridges is below the new conference threshold, then the one of these is chosen.

If no Call Bridge has yet been chosen, then one of the Call Bridges with a load less than the existing conference threshold is chosen.

**Load balancing outbound SIP calls**

From version 2.2, Call Bridge Groups supports the load balancing of outbound SIP calls, in addition to inbound SIP calls which was introduced in version 2.1.

To use this feature, do the following:

- enable load balancing of outbound SIP calls from spaces,
- set up outbound dial plan rules for load balancing outbound SIP calls,
- supply the Call Bridge Group or a specific Call Bridge for the outbound SIP calls, using the API object/calls/<call_id>/participants.

Once load balancing is enabled, outbound SIP calls follow the logic:

- Find the highest priority outbound dial plan rule that matches the domain,
  - if this applies to a local Call Bridge, then balance the call within the local Call Bridge Group.
  - if this only applies to remote Call Bridges, then load balance the call within the Call Bridge Group to which the Call Bridge is a member.

**Note:** Load balancing of calls from/to Lync clients, is not currently supported by Call Bridge Groups.
How to enable load balancing of outbound SIP calls

To configure the Call Bridges in a specific Call Bridge Group to attempt to load balance outgoing SIP calls from spaces, perform a PUT on the API object `/callBridgeGroups/<call bridge group id>` with the `loadBalanceOutgoingCalls` parameter set to true. Use POST if setting up a new Call Bridge Group.

For load balancing of outbound calls, each Call Bridge in the group must have the same dial plan rules.

How to set up an outbound dial plan rule for load balancing outbound SIP calls

This can be achieved in 3 ways:

1. Setting the `scope` parameter to `global` in all of the outbound dial plan rules. This ensures that all Call Bridges are able to use all of the outbound dial plan rules to reach a matching domain.

2. Creating identical outbound dial plan rules for each Call Bridge in the Call Bridge Group. Set the `scope` parameter set to `callBridge`. Use the `callBridge` parameter to set the ID of the Call Bridge.

3. Creating outbound dial plan rules for the specific Call Bridge Group. Set the `scope` parameter to `callBridgeGroup`, and set the `callBridgeGroup` parameter to the ID of the Call Bridge Group.

Before using load balancing of outbound calls, review the existing dial plan rules for each Call Bridge in the Call Bridge group. If the scope of existing rules needs to be altered, perform a PUT on the API object `/outboundDialPlanRules/<outbound dial plan rule id>` supplying the `scope` request parameter as defined above. Use POST on the API object `/outboundDialPlanRules` if setting up a new outbound dial plan rule.

How to supply the Call Bridge Group or specific Call Bridge to use for outbound SIP calls to participants

To make a call from a specific Call Bridge Group, perform a POST on the API object `/calls/<call id>/participants` with the parameter `callBridgeGroup` and the ID of the Call Group to use.

To make a call from a specific Call Bridge, perform a POST on the API object `/calls/<call id>/participants` with the parameter `callBridge` and the ID of the Call Bridge to use.

Handling load balancing of active empty conferences

The load balancing algorithm preferentially places new calls onto a Call Bridge where the conference is already active. An empty conference can be started on a Call Bridge by performing a POST on the API object `/calls`. By default these empty conferences are treated as active. This means that the first call to the empty conference is preferentially load balanced to
this Call Bridge. You can prevent the load balancing preferentially using the empty conferences, by setting the parameter `activeWhenEmpty` to `false` when performing the POST on the API object `/calls`.

### Load balancing Cisco Meeting App calls

Prior to version 2.3, only inbound and outbound SIP calls through Cisco Unified Communications Manager could be load balanced using Call Bridge Groups, calls using the Cisco Meeting App in the same deployment could not be load balanced; the media to and from the Cisco Meeting App always flowed through the Call Bridge that it first connected to.

From version 2.3, the existing load balancing algorithm has been extended to include Cisco Meeting App participants (including the WebRTC app users). This applies to:

- a Cisco Meeting App user joining as a member of the space,
- a Cisco Meeting App user joining as a non-member of the space, with and without a passcode
- a guest user joining the space,
- a participant added to a space from Cisco Meeting App.

By default, Cisco Meeting App participants are also load balanced if the `loadBalancingEnabled` parameter is set to true on the `/callBridgeGroups` API object (by default it is set to false). The decision on where to place the call is no longer restricted to the first Call Bridge which the Cisco Meeting App connects to.

The load balancing algorithm has been extended to include:

- Cisco Meeting App participants added via the API with a Call Bridge Group specified. The media will come from a Call Bridge in the specified Call Bridge Group, the Call Bridge chosen will be based on the existing algorithms
- Cisco Meeting App participants added via the API with a Call Bridge specified. The media will come from that Call Bridge.
- Cisco Meeting App participants simply joining a space without having been added to the space via the API. If this occurs, the Call Bridge that the Cisco Meeting App first connects to is determined, if that Call Bridge is part of a Call Bridge Group then the call is load balanced.

To load balance Cisco Meeting App calls, ensure that each Call Bridge in the Call Bridge Group has a connection to the XMPP cluster or single XMPP server, see the appropriate deployment guide for details on how to configure the connection.

**Note:** A call control device is not required for load balancing calls in deployments where only Cisco Meeting App is used to make calls (no SIP calls).
Disabling load balancing Cisco Meeting App participants

To disable load balancing Cisco Meeting App participants while continuing to load balance SIP calls, use the API to set the `loadBalanceUserCalls` request parameter on `/callBridgeGroups` to `false`. 
Example deployments of load balancing incoming calls using Cisco Expressway

This section of the white paper discusses two deployment examples to load balance incoming calls when using Cisco Expressway:

- Example 1 has the Meeting Servers trunked to their local Cisco Expressway cluster. The Cisco Expressways connect to each other via a neighbor zone.
- Example 2 has trunks from each Cisco Expressway cluster to every Meeting Server.

Figure 1: The two example deployments for load balancing incoming calls using Cisco Expressway

Deployment using neighbor zones to reach remote Meeting Servers (example 1)

This example deployment has four Meeting Servers, split over two offices (US, UK). The server names are simply CMS_OfficeNameNumber i.e. CMS_UK1. Cisco Expressway devices are named similarly.

Figure 2: Deployment model with neighbor zones to reach remote Meeting Servers.
This deployment uses neighbor zones to allow a Cisco Expressway to connect to a remote Meeting Server via another Cisco Expressway cluster. Meeting Servers are trunked to the local Cisco Expressway devices. Endpoints in the US all connect to the call control in the US (EXP_US), and similarly for the UK. Ideally calls originating in the US should use the US Call Bridges and similarly the UK endpoints should end up connected to the UK Call Bridges.

Dial plan configuration

To reduce inter-office bandwidth, the Cisco Expressways in these offices use dial plans to favor sending calls to the local Meeting Server resources. If the local resources are busy, or not available, then remote resources can be configured for use, if available. Only when nothing is available is a Busy response given.

Load balancing calls over local Call Bridges

Balancing calls over local Call Bridges is achieved by configuring a zone per local Call Bridge on the Cisco Expressway, the zone contains a single link to a single local conference resource in that location. The zone should be set up with a custom zone profile enabling Meeting Server load balancing. For more information see Appendix A

A search rule per local Call Bridge is then required. These are associated with zones created above. Each rule must have a different priority value, be set for “Source”= “Any” and “On successful match”= “Continue”.

The same rules can be configured on each cluster, see Figure 3.

Figure 3: Load balancing over local Call Bridges

Configuring failover to remote Call Bridges via Cisco Expressway clusters

Failover to remote Call Bridges is achieved by using additional search rules per Expressway cluster. These additional search rules target the neighbor zones to other Cisco Expressway
clusters. Note that to prevent call loops, the “source” value of these search rules may need to be set. This may mean multiple search rules need to be configured to cover all possible sources.

Figure 4: Load balancing over remote Call Bridges using neighbor zones

Call flow for calls load balanced between servers in the same location

Within each Call Bridge Group, the aim is to have calls for the same conference placed on the same server whenever possible. The call control system doesn’t need to know about the existing placement or loading, and will only need to redirect calls as required by the Meeting Server. This redirect will occur via a second Meeting Server node sending a new INVITE message back to the Cisco Expressway. This message contains the necessary information for the Cisco Expressway to replace any existing connection to the first Meeting Server with this new connection. The user’s device does not need to participate in this transfer at all, see Figure 5.
Figure 5: Call flow for calls load balanced between servers in the same location

Call flow for calls redirected within a location

If the first Call Bridge contacted is too busy to receive the initial call, then it will reject that call and then the call control system will reroute the call to another Call Bridge based on the dial plan, see Figure 6.
Figure 6: Call flow for calls redirected within a location 1.

In this example the call flow is:
1. An incoming call from a US endpoint to 8123456@example.com lands on EXP_US.
2. EXP_US matches this against a search rule, resolves that to “CMS_US1”.
3. EXP_US makes a call to “CMS_US1”.
4. This server returns a SIP 488 error code.
5. EXP_US continues to another search rule and makes a call to “CMS_US2”

**Note:** If CMS_US2 returns a SIP 488 error code, then the above process is repeated with the next search rule. If all Call Bridges in the group return a SIP 488 error code, then a BUSY will be sent back to the endpoint.

Call flow for calls redirected between regions

It is possible that all local Call Bridges are in use. In this case, it is possible to redirect the call to another Call Bridge or a Call Bridge Group. This is a deployment choice and the Meeting Servers can be deployed to either reject calls when all local resources are in use, or to try other Call Bridges, see Figure 7. The **Dial plan configuration** section explains how to use dial plans to redirect calls between regions.
Figure 7: Call flow for calls redirected between regions

In this example the call flow is:

1. An incoming call from a US endpoint to 8123456@example.com lands on EXP_US
2. EXP_US uses search rules to resolve this to “CMS_US1”.
3. EXP_US makes a call to “CMS_US1”.
4. This server returns a SIP 488 error code.
5. EXP_US continues to another search rule and makes a call to “CMS_US2”
6. This server returns a SIP 488 error code.
7. EXP_US continues to another search rule and makes a call to “EXP_UK”
8. EXP_UK starts searching its search rules and resolves this to “CMS_UK1”

Deployment with zones direct to remote Call Bridges (example 2)

This example deployment has four Meeting Servers, split over two offices (US, UK). The server names are simply CMS_OfficeNameNumber i.e. CMS_UK1. Cisco Expressway devices are named similarly, see Figure 8.
This deployment uses zones to allow a Cisco Expressway to directly connect to a remote Meeting Server. Meeting Servers are trunked to the local Cisco Expressway devices. Endpoints in the US all connect to the call control in the US (EXP_US), and similarly for the UK. Ideally calls originating in the US should use the US Call Bridges and similarly the UK endpoints should end up connected to the UK Call Bridges.

**Dial plan configuration**

To reduce inter-office bandwidth, the Cisco Expressways in these offices use dial plans to favor sending calls to the local Meeting Server resources. If the local resources are busy, or not available, then remote resources can be configured for use, if available. Only when nothing is available is a Busy response given.

**Load balancing calls over local Call Bridges**

Balancing calls over local Call Bridges is achieved by configuring a zone per local Call Bridge on the Cisco Expressway, the zone contains a single link to a single local conference resource in that location. The zone should be set up with a custom zone profile enabling Meeting Server load balancing. For more information see Appendix A.

A search rule per local Call Bridge is then required. These are associated with zones created above. Each rule must have a different priority value, be set for “Source” = “Any” and “On successful match” = “Continue”.

The same rules can be configured on each cluster, see Figure 9.
Configuring failover to remote Call Bridges via direct connections

Failover to remote Call Bridges is achieved by using additional search rules per Expressway cluster. These additional search rules target the neighbor zones directly to the remote Meeting Server nodes.

Call flow for calls load balanced between servers in the same location

Within each Call Bridge Group, the aim is to have calls for the same conference placed on the same server whenever possible. The call control system doesn’t need to know about the existing placement or loading, and will only need to redirect calls as required by the Meeting Server. This redirect will occur via a second Meeting Server node sending a new INVITE message back to the Cisco Expressway. This message contains the necessary information for the Cisco Expressway to replace any existing connection to the first Meeting Server with this new connection. The user’s device does not need to participate in this transfer at all, see Figure 11.
Figure 11: Call flow for calls load balanced between servers in the same location

Call flow for calls redirected within a location

If the first Call Bridge contacted is too busy to receive the initial call, then it will reject that call and then the call control system will reroute the call to another Call Bridge based on the dial plan, see Figure 12.
Figure 12: Call flow for calls redirected within a location

In this example the call flow is:
1. An incoming call from a US endpoint to 8123456@example.com lands on EXP_US
2. EXP_US matches this against a search rule, that resolves to “CMS_US1”.
3. EXP_US makes a call to “CMS_US1”.
4. This server returns a SIP 488 error code.
5. EXP_US continues to another search rule and makes a call to “CMS_US2”

**Note:** If CMS_US2 returns a SIP 488 error code, then the above process is repeated with the next search rule. If all Call Bridges in the group return a SIP 488 error code, then a BUSY will be sent back to the endpoint.

**Call flow for calls redirected between regions**

It is possible that all local Call Bridges are in use. In this case, it is possible to redirect the call to another Call Bridge or a Call Bridge Group. This is a deployment choice and the Meeting Servers can be deployed to either reject calls when all local resources are in use, or to try other Call Bridges, see Figure 13. The [Dial plan configuration](#) section explains how to use dial plans to redirect calls between regions.
In this example the call flow is:

1. An incoming call from a US endpoint to 8123456@example.com lands on EXP_US
2. EXP_US uses search rules to resolve this to “CMS_US1”.
3. EXP_US makes a call to “CMS_US1”.
4. This server returns a SIP 488 error code.
5. EXP_US continues to another search rule and makes a call to “CMS_US2”
6. This server returns a SIP 488 error code.
7. EXP_US continues to another search rule and makes a call to “CMS_UK1”
Example deployments of load balancing incoming calls using Cisco Unified Communications Manager

This section of the white paper discusses three deployment examples to load balance incoming calls using Cisco Unified Communications Manager:

- Example 1 has the Meeting Servers trunked to their local Cisco Unified Communications Manager. The Cisco Unified Communications Managers connect to a Cisco Unified Communications Manager Session Management Edition (SME) as leaf nodes. The SME routes calls between the nodes.

- Example 2 has centralized Meeting Servers trunked to an SME and a global Cisco Unified Communications Manager deployment.

- Example 3 has the Meeting Servers trunked to their local Cisco Unified Communications Manager. The Cisco Unified Communications Managers are simply trunked together, there is no SME to centrally route the calls.

Figure 14: The three example deployments for load balancing incoming calls

For any deployment, there are three options for how calls from different devices are mapped to specific resources:

- Multiple partitions with Calling Search Spaces used to select the correct partition.

- Single partition with Local Route Groups. Selection of routes is done via multiple device pools.

- Dial string manipulation within a single partition per cluster.

Each of these options can be used with any of the deployments.

The last option is easy to do for numeric dial plans, but for URI dialing, LUA scripts will be required. The other two options work equally well for numeric and URI dialing.
Deployment uses an SME for centralized call routing between Cisco Unified Communications Managers (example 1)

This example deployment has 8 Meeting Servers, split over three offices (US, UK and HK). The server names are simply CMS_OfficeNameNumber i.e. CMS_UK1. Cisco Unified Communications Manager devices are named similarly.

Figure 15: Deployment model with Meeting Servers on leaf Cisco Unified Communications Manager nodes when using an SME for centralized routing of calls

This deployment uses a Cisco Unified Communications Manager Session Management Edition (SME) for centralized routing of conferencing traffic, all other Cisco Unified Communications Managers connect to the SME as leaf nodes. Meeting Servers are trunked to the local Cisco Unified Communications Manager devices. Endpoints in the US all connect to the call control in the US (CUCM_US), and similarly for the UK. Ideally calls originating in the US should use the US Call Bridges and similarly the UK endpoints should end up connected to the UK Call Bridges.
Dial plan configuration

To reduce inter-office bandwidth, the Cisco Unified Communications Manager in these offices use dial plans to favor sending calls to the local Meeting Server resources. If the local resources are busy, or not available, then remote resources can be configured for use, if available. Only when nothing is available is a Busy response given.

Load balancing calls over local Call Bridges

Balancing calls over local Call Bridges is achieved by configuring a route group per location on the Cisco Unified Communications Manager, the route group contains links to the local conference resources in that location. The route group should be set up with circular distribution to load balance calls across the Meeting Servers.

Each trunk to a local Call Bridge must be configured to use a SIP Trunk Security Profile that has the “Accept Replaces Header” checkbox selected. For more information see the Security Guide for Cisco Unified Communications Manager.

A route list is then required for calls that originate in that location. This is associated with a route pattern as dialed by the users. This route list contains the route group.

Route patterns should be pointed to each of these route lists. If no failover to remote Call Bridges is required then only one partition is required per cluster. The same route pattern can be configured on each cluster, see Figure 16.

Figure 16: Load balancing over local Call Bridges
Configuring failover to remote Call Bridges using an SME

Failover to remote Call Bridges is achieved by configuring additional route groups into the route lists per location. Then additional route groups and lists are added for calls to and from the SME.

Two partitions should be used per location for the leaf nodes, and one partition per location on the SME. These allow the same route patterns to be used everywhere, with a Calling Search Space used on the local calls and also on the trunks between each location.

Each trunk to a Call Bridge needs to have the rerouting Calling Search Space set to include a partition containing the trunks to the Call Bridges.

On each leaf cluster there is a new route group consisting of the trunk to the SME. This is used for local calls that could not be handled on a local resource.

On the leaf nodes there are route lists for calls coming from the SME. These only point to the route group containing local resources.

The SME has a route group per leaf cluster. This route group contains trunks to all the other leaf clusters. Incoming calls from one leaf cluster are mapped to a route group that contains the trunks to all other leaf clusters.
Figure 17: Load balancing over remote Call Bridges using the SME

Call flow for calls load balanced between servers in the same location

Within each Call Bridge Group, the aim is to have calls for the same conference placed on the same server whenever possible. The call control system doesn’t need to know about the existing placement or loading, and will only need to redirect calls as required by the Meeting Server.

This redirect will occur via a second Meeting Server node sending a new INVITE message back to the Cisco Unified Communications Manager. This message contains the necessary information for the Cisco Unified Communications Manager to replace any existing connection.
to the first Meeting Server with this new connection. The user’s device does not need to participate in this transfer at all, see Figure 18.

Figure 18: Call flow for calls load balanced between servers in the same location

Call flow for calls redirected within a location

If the first Call Bridge contacted is too busy to receive the initial call, then it will reject that call and then the call control system will reroute the call to another Call Bridge based on the dial plan, see Figure 19.
Figure 19: Call flow for calls redirected within a location

1. An incoming call from a US endpoint to 8123456@example.com lands on CUCM_US
2. CUCM_US matches this against a route pattern, resolves that to “RL_US” and then to the “RG_US”.
3. CUCM_US makes a call to one of the servers in that route group “CMS_US1”.
4. This server returns a SIP 488 error code.
5. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US2”

**Note:** If CMS_US2 returns a SIP 488 error code, then the above process is repeated with the next Call Bridge in the group. If all Call Bridges in the group return a SIP 488 error code, then a BUSY will be sent back to the endpoint.

**Call flow for calls redirected between regions**

It is possible that all local Call Bridges are in use. In this case, it is possible to redirect the call to another Call Bridge or a Call Bridge Group. This is a deployment choice and the Meeting Servers can be deployed to either reject calls when all local resources are in use, or to try other Call Bridges, see Figure 20. **Dial plan configuration** on page 27 explains how to use dial plans to redirect calls between regions.
In this example,

1. An incoming call from a US endpoint to 8123456@example.com lands on CUCM_US.
2. CUCM_US uses Calling Space Search to match this against a route pattern, resolves that to “RL_US” and then to the “RG_US”.
3. CUCM_US makes a call to one of the servers in that route group “CMS_US2”.
4. This server returns a SIP 488 error code.
5. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US3”.
6. This server returns a SIP 488 error code.
7. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US1”.
8. This server returns a SIP 488 error code.
9. CUCM_US understands that this call may be routable to another element in the route list. All of “RG_US” has been tried, so it moves to the second item in the route list, “RG_SME”.
10. “RG_SME” only contains one entry, so CUCM_US sends the call to SME.
11. SME uses Calling Search Space rules to determine that the call should use route list “RL_SME_xUS”. It then tries making a call to the first route group in this list (“RG_SME_xUS”) and
calls CUCM UK.

12. CUCM UK uses Calling Search Space to determine that this call should use route list “RL_SME_xUK”. This only contains one route group “RG_SME_xUK” which has just the local bridges.

13. CUCM UK makes a call to one of the servers in that route group “CMS_UK3”.

Using Local Route Groups

The above configuration introduces the use of the same Route Pattern in multiple partitions, each one addressed through a specific Calling Search Space for each trunk. The number of Partitions and Calling Search Spaces on the SME cluster equals the number of trunks. There are also two partitions and Calling Search Spaces on each leaf cluster. In order to simplify the configuration and reduce to a single Partition and Calling Search Space per cluster, it’s possible to use the Local Route Group approach.

A Local Route Group is a variable that can assume any configured Route Group value based on the Device Pool setting for the calling device.

A Route Pattern or SIP Route Pattern is connected to a Route List, the Route List includes a series of Local Route Groups (instead of Route Groups), and the Local Route Groups are populated accordingly to the settings on the Device Pool of the calling device. In this case the actual route groups are unchanged from the above configuration.

To set this up, the administrator needs to create two Local Route Groups. In this example they are called LRG_Conferencing_Local and LRG_Conferencing_Remote. See Figure 40 in Appendix B. Once these have been created by the administrator, they appear as a list in any Device Pool. These Local Groups should be set up on the SME and all leaf clusters, and for simplicity the same Local Route Group names are used on all systems.

On each CUCM cluster the administrator should create a route pattern pointing to a Route List containing these two Local Route Groups. See Figure 41 and Figure 42. For this example ensure the configuration is identical on every cluster.

**Figure 21: Local Remote Groups**

```
<table>
<thead>
<tr>
<th>Route Pattern</th>
<th>RL_Conferencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>812xxxx</td>
<td>LRG_Conference_Local</td>
</tr>
<tr>
<td></td>
<td>LRG_Conference_Remote</td>
</tr>
</tbody>
</table>
```

Route groups should then be defined on each cluster. These Route Groups are the same as above.
Table 1: Route Groups used in examples

<table>
<thead>
<tr>
<th>Route Group</th>
<th>Trunks to</th>
<th>Cluster where configuration is active</th>
</tr>
</thead>
<tbody>
<tr>
<td>RG_US</td>
<td>CMS_US1, CMS_US2, CMS_US3</td>
<td>US</td>
</tr>
<tr>
<td>RG_UK</td>
<td>CMS_UK1, CMS_UK2, CMS_UK3</td>
<td>UK</td>
</tr>
<tr>
<td>RG_HK</td>
<td>CMS_HK1, CMS_HK2</td>
<td>HK</td>
</tr>
<tr>
<td>RG_SME</td>
<td>SME</td>
<td>US, UK, HK</td>
</tr>
<tr>
<td>RG_SME_xUS</td>
<td>CUCM_US, CUCM_HK</td>
<td>SME</td>
</tr>
<tr>
<td>RG_SME_xUK</td>
<td>CUCM_US, CUCM_HK</td>
<td>SME</td>
</tr>
<tr>
<td>RG_SME_xHK</td>
<td>CUCM_US, CUCM_UK</td>
<td>SME</td>
</tr>
</tbody>
</table>

For each device pool, the values for LRG_Conferencing_Local and LRG_Conferencing_Remote should be set. Where appropriate, the first of these should point to a Route Group containing local trunks to resources that will be used first. The second should point to a Route Group containing trunks to other CUCM nodes through which conferencing resources could be used. Not all device pools will have both set. Incoming trunks from SME to a leaf node should not loop back to the SME, and there are no local resources on the SME in this example. See Figure 43 to Figure 51 in Appendix B.

All trunks and local devices should be added to the appropriate Device Pool as shown in the table below.

Table 2: Device Pools used in examples

<table>
<thead>
<tr>
<th>Device Pool</th>
<th>LRG_Conferencing_Local</th>
<th>LRG_Conferencing_Remote</th>
<th>Cluster where configuration is active</th>
</tr>
</thead>
<tbody>
<tr>
<td>US endpoints</td>
<td>RG_US</td>
<td>RG_SME</td>
<td>US</td>
</tr>
<tr>
<td>UK endpoints</td>
<td>RG_UK</td>
<td>RG_SME</td>
<td>UK</td>
</tr>
<tr>
<td>HK endpoints</td>
<td>RG_HK</td>
<td>RG_SME</td>
<td>HK</td>
</tr>
<tr>
<td>Trunk to SME on US</td>
<td>RG_US</td>
<td>&lt;None&gt;</td>
<td>US</td>
</tr>
<tr>
<td>Trunk to and from SME</td>
<td>RG_UK</td>
<td>&lt;None&gt;</td>
<td>UK</td>
</tr>
<tr>
<td>Trunk to and from SME</td>
<td>RG_HK</td>
<td>&lt;None&gt;</td>
<td>HK</td>
</tr>
<tr>
<td>Calls to SME from US</td>
<td>&lt;None&gt;</td>
<td>RG_SME_xUS</td>
<td>SME</td>
</tr>
<tr>
<td>Calls to SME from UK</td>
<td>&lt;None&gt;</td>
<td>RG_SME_xUK</td>
<td>SME</td>
</tr>
<tr>
<td>Calls to SME from HK</td>
<td>&lt;None&gt;</td>
<td>RG_SME_xHK</td>
<td>SME</td>
</tr>
</tbody>
</table>
The Route List specifies which Local Route Group will be chosen based on dialled digits or URI. When a Route Pattern that points to RL_Conferencing is matched, the LRG_Conferencing_Local and LRG_Conferencing_Remote Route Groups will be populated with the corresponding content in the Device Pool.

For example,

1. an incoming call from a US endpoint to 8123456@example.com lands on CUCM_US
2. CUCM_US matches this against a route pattern, resolves that to “RL_Conferencing” and then to the “LRG_Conferencing_Local”.
3. CUCM_US uses the device pool of the endpoint to populate “LRG_Conferencing_Local” with “RG_US”
4. CUCM_US makes a call to one of the servers in that route group “CMS_US2”.
5. This server returns a SIP 488 error code.
6. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US3”
7. This server returns a SIP 488 error code.
8. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US1”
9. This server returns a SIP 488 error code.
10. CUCM_US understands that this call may be routable to another element in the route list. All of “RG_US” has been tried, so it moves to the second item in the route list “LRG_Conferencing_Remote”.
11. CUCM_US uses the device pool of the endpoint to populate “LRG_Conferencing_Remote” with “RG_SME”
12. This only contains one entry, so CUCM_US sends the call to SME
13. SME uses the route pattern to match this to the route list “RL_Conferencing” and then to the Local Route Groups “LRG_Conferencing_Local” and “LRG_Conferencing_Remote”.
14. SME uses the device pool of the trunk to populate these. “LRG_Conferencing_Local” is empty and “LRG_Conferencing_Remote” is “RG_SME_xUS”
15. SME selects one of the trunks in “RG_SME_xUS” and calls CUCM_UK.
16. CUCM_UK uses uses the route pattern to match this to the route list “RL_Conferencing” and then to the Local Route Groups “LRG_Conferencing_Local” and “LRG_Conferencing_Remote”.
17. CUCM_UK uses the device pool of the trunk to populate these. “LRG_Conferencing_Local” is “RG_UK” and “LRG_Conferencing_Remote” is empty.
18. CUCM_UK makes a call to one of the servers in that route group “CMS_UK3”.

Load Balancing Calls Across Cisco Meeting Servers : White Paper
If you want to know more on Local Route Groups, see the Preferred Architecture for Collaboration, Call Control section:

Deployment with centralized Meeting Servers and an SME to route calls (example 2)

This example deployment has 4 Meeting Servers all centrally deployed. The server names are simply CMS_Number i.e. CMS_1. Cisco Unified Communications Manager devices are named based on their location.

Figure 22: Deployment model with centralized Meeting Servers and an SME to route calls

This deployment uses a Cisco Unified Communications Manager Session Management Edition (SME) to route calls between the Cisco Unified Communications Manager and the centralized Meeting Servers. All trunks to Meeting Servers come from the SME rather than local Cisco Unified Communications Manager devices. All Meeting Servers exist in a single Call Bridge Group in a common location. Endpoints in the US all connect to the call control in the US (CUCM_US), and similarly for the UK.

Dial plan configuration

When all conferencing resources are centrally located in a single Call Bridge Group the balancing over the Call Bridges is achieved by configuring a single route group on the SME containing all of the Meeting Servers. The route group should be set up with circular distribution to load balance calls across the Meeting Servers.
Each trunk to a local Call Bridge must be configured to use a SIP Trunk Security Profile that has the “Accept Replaces Header” checkbox selected. For more information see the Security Guide for Cisco Unified Communications Manager.

A single route list is then required for all calls. This is associated with a route pattern as dialed by the users. This route list contains the route group.

**Figure 23: Load balancing over centrally located Call Bridges**

<table>
<thead>
<tr>
<th>Route Pattern 1</th>
<th>Route List Route Group</th>
<th>Route Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>812xxxx</td>
<td></td>
<td>CMS_1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CMS_2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CMS_3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CMS_4</td>
</tr>
</tbody>
</table>

**Call flow for a call being load balanced**

The aim is to have calls for the same conference placed on the same Meeting Server whenever possible. The call control system doesn’t need to know about existing placement or loading, and will only need to redirect calls as required.

This redirect will occur via a second Meeting Server node sending a new INVITE message back to the SME. This message contains the necessary information for the SME to replace any existing connection to the first Meeting Server with this new connection. The user’s device does not need to participate in this transfer at all.
Figure 24: Call flow for load balancing a call

Call flow for call being redirected

If the first Call Bridge contacted is too busy to receive the initial call, then it will reject that call and then the call control system will reroute the call to another Call Bridge based on the dial plan, see Figure 25.
Figure 25: Call flow for call redirected within a location

1. An incoming call from a US endpoint to 8123456@example.com lands on CUCM_US.
2. CUCM_US matches this against a route pattern, resolves that to trunk to SME.
3. SME matches this against a route pattern, resolves that to “Route List” and then to the “Route Group”.
4. SME makes a call to one of the servers in that route group “CMS_1”.
5. This server returns a SIP 488 error code.
6. SME understands that this call may be routable to another element in the route group and makes a call to “CMS_2”.

This process could be repeated until all Call Bridges return the SIP 488 error code, in which case a BUSY will be sent back to the endpoint.
Deployment without an SME to route calls (example 3)

This example deployment has 8 Meeting Servers, split over three offices (US, UK and HK). The server names are simply CMS_OfficeNameNumber i.e. CMS_UK1. Cisco Unified Communications Manager devices are named similarly. In this example, no SME is deployed.

Figure 26: Deployment model of Meeting Servers without an SME to route calls

Call control is configured to preferentially route calls from endpoints to the local Call Bridge Group. Endpoints in the US all connect to the call control in the US (CUCM_US), and similarly endpoints in the UK connect to CUCM_UK, and endpoints in HK connect to CUCM胡K.
Dial plan configuration

To reduce inter-office bandwidth, the Cisco Unified Communications Manager in these offices use dial plans to favor sending calls to the local Meeting Server resources. If the local resources are busy, or not available, then remote resources can be configured for use, if available. Only when nothing is available is a Busy response given.

Load balancing calls over local Call Bridges

Load balancing calls over local Call Bridges is achieved by configuring a route group per location containing the local resources. The route group should be set up with circular distribution to load balance calls across the servers.

Each trunk to a local Call Bridge must be configured to use a SIP Trunk Security Profile that has the “Accept Replaces Header” checkbox selected. For more information see the Security Guide for Cisco Unified Communications Manager.

A route list is then required for calls that originate in that location. This is associated with a route pattern as dialed by the users. This route list contains the route group.

On each cluster, a route pattern should point to the route list, see Figure 27.

Figure 27: Load balancing over local Call Bridges

Key:
RP = Route Pattern
RL = Route List
RG = Route Group
Configuring failover to remote Call Bridges without using an SME

Failover to remote Call Bridges is achieved by configuring a route group per location containing the local resources, and a second route group containing the trunks to the other Cisco Unified Communications Manager clusters. Both route groups should be set up with circular distribution to load balance calls across the servers and locations.

Two route lists are required. The first is used for calls that originate in that location. This is associated with a route pattern as dialed by the users. This route list contains both route groups. Since route lists have an order in which route groups are used, the group containing local resources must be first. The second route group lists the trunks to the other locations. Calls using this second route group must be transformed before sending. In this example the calls are prefixed with an ‘8’ before being forwarded to other clusters.

The second route list is for calls that originate in other locations. For these only the local resources should be queried, and no further distribution to other trunks should be possible. This is done by only adding the route group for local resources to this route list. This route list should be associated with the transformed dial pattern created when forwarding a call between regions. Any changes to add digits or other transforms must be reversed before the call reaches the Meeting Servers.

On each cluster, route patterns should be pointed to each of these route lists, see Figure 28.

Figure 28: Load balancing over remote Call Bridges without using an SME
Call flow for a call load balanced between servers in the same location

Within each Call Bridge Group, the aim is to have calls for the same conference placed on the same server whenever possible. The call control system doesn’t need to know about the existing placement or loading, and will only need to redirect calls as required by the Meeting Server.

This redirect will occur via a second Meeting Server node sending a new INVITE message back to the Cisco Unified Communications Manager. This message contains the necessary information for the Cisco Unified Communications Manager to replace any existing connection to the first Meeting Server with this new connection. The user’s device does not need to participate in this transfer at all, see Figure 29.

Figure 29: Call flow for load balancing between servers in the same location

Call flow for a call redirected within a location

If the first Call Bridge contacted is too busy to receive the initial call, then it will reject that call and then the call control system will reroute the call to another Call Bridge based on the dial plan, see Figure 30.
Figure 30: Call flow for call redirected within a location

1. An incoming call from a US endpoint to 8123456@example.com lands on CUCM_US.
2. CUCM_US matches this against a route pattern, resolves that to "RL_US" and then to the "RG_US".
3. CUCM_US makes a call to one of the servers in that route group "CMS_US1".
4. This server returns a SIP 488 error code.
5. CUCM_US understands that this call may be routable to another element in the route group and makes a call to "CMS_US2".

Note: If CMS_US2 returns a SIP 488 error code, then the above process is repeated with the next Call Bridge in the group. If all Call Bridges in the group return a SIP 488 error code, then a BUSY will be sent back to the endpoint.

Call flow for a call redirected between regions

It is possible that all local Call Bridges are in use. In this case, it is possible to redirect the call to another Call Bridge or a Call Bridge group. This is a deployment choice and the Meeting Servers can be deployed to either reject calls when all local resources are in use, or to try other Call Bridges, see Figure 31. Dial plan configuration on page 42 explains how to use dial plans to redirect calls between regions.
In this example,

1. An incoming call from a US endpoint to 8123456@example.com lands on CUCM_US
2. CUCM_US matches this against a route pattern, resolves that to “RL_US” and then to the “RG_US”.
3. It makes a call to one of the servers in that route group “CMS_US2”.
4. This server returns a SIP 488 error code.
5. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US3”
6. This server returns a SIP 488 error code.
7. CUCM_US understands that this call may be routable to another element in the route group and makes a call to “CMS_US1”
8. This server returns a SIP 488 error code.
9. CUCM_US understands that this call may be routable to another element in the route list. All of “RG_US” has been tried, so it moves to the second item in the route list, “Route Group US 2”, transforms the dial string to add a leading ‘8’ and makes a call to “CUCM_UK”
10. CUCM_UK matches the call to 88123456@example.com to a route pattern, and then to “RL_UK” and in turn to “RG_UK”. It also strips the leading 8.
11. It makes a call to one of the servers in that route group “CMS_UK3”.
Examples of load balancing outbound calls

Load balancing outbound calls with local dial plan rule

In this example, the API client is making a call to a user that is reachable from any server. The API client sends an API request to one of the Meeting Server nodes (CMS_US2). This determines that the node can make the call based on the domain. Since the node is part of a Call Bridge Group, it communicates with the other nodes in this group and then selects a node to make the outbound call from. This call behaves in the same way as any outbound call.

Figure 32: Call flow for load balancing outbound calls with local dial plan rule

Load balancing outbound calls with no local dial plan rule

In this example, the API client is making a call to a user that is reachable from only servers based in Hong Kong. The API client sends an API request to one of the Meeting Server nodes (CMS_US2). This determines that the node can’t make the call based on the domain. Using the dial plan rules the node determines that nodes in HK can make the call. It communicates with the other nodes in that Call Bridge Group and then selects a node to make the outbound call from. This call behaves in the same way as any outbound call.
Figure 33: Call flow for load balancing outbound calls with no local dial plan rule

Load balancing outbound calls with explicit choice of Call Bridge Group

In this example, the API client is making a call to a user, but the API client wants to only use servers based in Hong Kong. The API client sends an API request to one of the Meeting Server nodes (CMS_US2). Based on the parameters in this API request, the node communicates with the other nodes in the specified Call Bridge Group and then selects a node to make the outbound call from. This call behaves in the same way as any outbound call.

Figure 34: Call flow for load balancing outbound calls with explicit choice of Call Bridge Group
Load balancing outbound calls with explicit choice of Call Bridge

In this example, the API client is making a call to a user, but the API client wants to only use a specific server. The API client sends an API request to one of the Meeting Server nodes (CMS_US2). Based on the parameters in this API request, it communicates with the specified node and uses this node to make the outbound call from. This call behaves in the same way as any outbound call.

Figure 35: Call flow for load balancing outbound calls with explicit choice of Call Bridge
Appendix A  Configuring a Cisco Expressway dial plan

Call routing in Cisco Expressway deployments relies on the use of dial plans and zones, and it is assumed that these concepts are understood. For information on configuring dial plans and zones, please consult the Expressway documentation.

Dial plans are used by the call control system to favor sending calls to local Meeting Server resources, thereby reducing inter-office bandwidth.

On the call control device:

1. Configure a zone for each Call Bridge in the Call Bridge Group.
   a. Navigate to Configuration > Zones and create a New zone of type Neighbor, that this cluster will be directly communicating with. You need one zone per Call Bridge node.

   **Figure 36: Creating new zones on Expressway**

2. In the Advanced section, select the Zone profile as Custom and set the Meeting Server load balancing parameter to On.

   **Figure 37: Setting the Meeting Server load balancing parameter**
3. For each zone, create a search rule pointing to the zone. Navigate to Configuration > Dial plan > Search rules and create a New search rule. Each search rule must have a different priority, and On successful match must be set to Continue.

When choosing priorities it is important to consider the order in which you want this Cisco Expressway to use resources. Local resources would typically have higher priority than remote resources.

Figure 38: Creating a search rule for connections to local resources

If remote resources are going to be used via other Cisco Expressway clusters, configure search rules that resolve to these. These rules would typically have a lower priority than rules to local Meeting Servers, and typically have a source set to avoid call loops.
Figure 39: Creating search rules for connections to other Cisco Expressway clusters
Appendix B  Setting up Local Route Groups on Cisco Unified Communications Manager

This appendix provides example screen shots related to setting up Local Route Groups on Cisco Unified Communications Manager.

Figure 40: Adding local route groups
Figure 41: Adding a route list that uses local route groups
Figure 42: Creating a route pattern using the route list
Figure 43: Creating a route group
Figure 44: Setting the local route groups for the US Endpoint device pool
Figure 45: Setting the local route groups for the trunk to SME device pool
Figure 46: Setting the device pool on the trunk to/from SME

Figure 47: Partitions on a leaf cluster
Figure 48: Creating a route list for calls from the SME
Figure 49: Setting up a route pattern for a particular partition
Figure 50: Setting the partitions for a CSS
Figure 51: Setting the Calling Search Space for calls from a trunk
Appendix C  Growth in scaling deployments

Table 3 below demonstrates the expansion in maximum call capacities on Meeting Servers by upgrading to later software versions. Bold indicates a new feature in that software version. Note there are different capacities for a single or cluster of Meeting Servers compared to load balancing calls within a Call Bridge Group.

Table 3: Evolution in Meeting Server call capacity

<table>
<thead>
<tr>
<th>Software version</th>
<th>Cisco Meeting Server platform</th>
<th>2.0</th>
<th>2.1</th>
<th>2.2</th>
<th>2.3</th>
<th>2.4 and 2.5</th>
<th>2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1080p30 720p30 SD Audio</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Individual</td>
<td>48 96 192 3000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>48</td>
<td>250</td>
<td>48</td>
</tr>
<tr>
<td>Meeting Servers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td>500</td>
<td>96</td>
</tr>
<tr>
<td>or</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td>1000</td>
<td>96</td>
</tr>
<tr>
<td>Meeting Servers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>192</td>
<td>3000</td>
<td>192</td>
</tr>
<tr>
<td>in a cluster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td>3000</td>
<td>96</td>
</tr>
<tr>
<td>(notes 1,2,3 and 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>450</td>
<td>100</td>
</tr>
<tr>
<td>HD participants</td>
<td>48 96 192 3000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>48</td>
<td>250</td>
<td>48</td>
</tr>
<tr>
<td>per conference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td>500</td>
<td>96</td>
</tr>
<tr>
<td>per server</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96</td>
<td>1000</td>
<td>96</td>
</tr>
<tr>
<td>WebRTC connections per Web Bridge</td>
<td>100</td>
<td>NA</td>
<td>100</td>
<td>NA</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
## Appendix C  Growth in scaling deployments

<table>
<thead>
<tr>
<th>Software version</th>
<th>2.0</th>
<th>2.1</th>
<th>2.2</th>
<th>2.3</th>
<th>2.4 and 2.5</th>
<th>2.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meeting Servers in a Call Bridge Group</td>
<td>Call type supported</td>
<td>NA</td>
<td>NA</td>
<td>Inbound SIP</td>
<td>NA</td>
<td>Inbound SIP Outbound SIP</td>
</tr>
<tr>
<td>1080p30 720p30 SD Audio Load limit</td>
<td>NA</td>
<td>NA</td>
<td>48 96 192 3000 96,000</td>
<td>NA</td>
<td>48 96 192 3000 96,000</td>
<td>48 96 192 3000 96,000</td>
</tr>
<tr>
<td>Number of HD participants per conference per server</td>
<td>NA</td>
<td>NA</td>
<td>96</td>
<td>NA</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>WebRTC connections per Web Bridge</td>
<td>NA</td>
<td>NA</td>
<td>100</td>
<td>NA</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Note 1: Maximum of 24 Call Bridge nodes per cluster; cluster designs of 8 or more nodes need to be approved by Cisco, contact Cisco Support for more information.

Note 2: Clustered Cisco Meeting Server 2000’s without Call Bridge Groups configured, support integer multiples of maximum calls, for example integer multiples of 700 HD calls.

Note 3: Up to 16,800 HD concurrent calls per cluster (24 nodes x 700 HD calls).
Note 4: A maximum of 2600 participants per conference per cluster depending on the Meeting Servers platforms within the cluster.

Note 5: A Cisco Meeting Server 2000 with Call Bridge Groups enabled only supports 500 HD calls. The `loadLimit` for a Cisco Meeting Server 2000 is set at 500,000 due to the total number of video calls supported remaining at 1,000. The Meeting Server does not offer a way to limit the minimum resolution of calls, so it will always be possible to have a mix of HD and lower resolution calls. There is no mechanism to make use of the increased scale in HD or full HD in a load balanced configuration.

Note 5: A From version 2.6, the call capacity for Cisco Meeting Server 2000 with Call Bridge Groups enabled, has increased to 700 HD calls, and the `loadLimit` has increased from 500000 to 700000. The load calculation for the different call resolutions has been updated to match the new 700000 limit. Load limits for other Meeting Server platforms stay as they were previously; these changes only apply to the Cisco Meeting Server 2000.
Cisco Legal Information

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS IN THIS MANUAL ARE SUBJECT TO CHANGE WITHOUT NOTICE. ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS MANUAL ARE BELIEVED TO BE ACCURATE BUT ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. USERS MUST TAKE FULL RESPONSIBILITY FOR THEIR APPLICATION OF ANY PRODUCTS.

THE SOFTWARE LICENSE AND LIMITED WARRANTY FOR THE ACCOMPANYING PRODUCT ARE SET FORTH IN THE INFORMATION PACKET THAT SHIPPED WITH THE PRODUCT AND ARE INCORPORATED HEREIN BY THIS REFERENCE. IF YOU ARE UNABLE TO LOCATE THE SOFTWARE LICENSE OR LIMITED WARRANTY, CONTACT YOUR CISCO REPRESENTATIVE FOR A COPY.

The Cisco implementation of TCP header compression is an adaptation of a program developed by the University of California, Berkeley (UCB) as part of UCB’s public domain version of the UNIX operating system. All rights reserved. Copyright © 1981, Regents of the University of California.

NOTWITHSTANDING ANY OTHER WARRANTY HEREIN, ALL DOCUMENT FILES AND SOFTWARE OF THESE SUPPLIERS ARE PROVIDED “AS IS” WITH ALL FAULTS. CISCO AND THE ABOVE-NAMED SUPPLIERS DISCLAIM ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, THOSE OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OR ARISING FROM A COURSE OF DEALING, USAGE, OR TRADE PRACTICE.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

All printed copies and duplicate soft copies of this document are considered uncontrolled. See the current online version for the latest version.

Cisco has more than 200 offices worldwide. Addresses and phone numbers are listed on the Cisco website at www.cisco.com/go/offices.

© 2019 Cisco Systems, Inc. All rights reserved.
Cisco Trademark

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1721R)