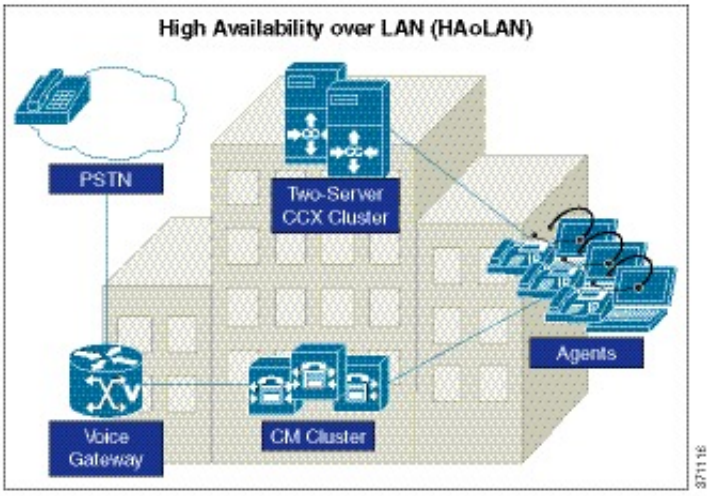
High Availability and Network Design

Unified CCX High Availability over LAN

Unified CCX supports high availability over LAN to provide redundancy over LAN. The following figure depicts the deployment for Unified CCX high availability over LAN.

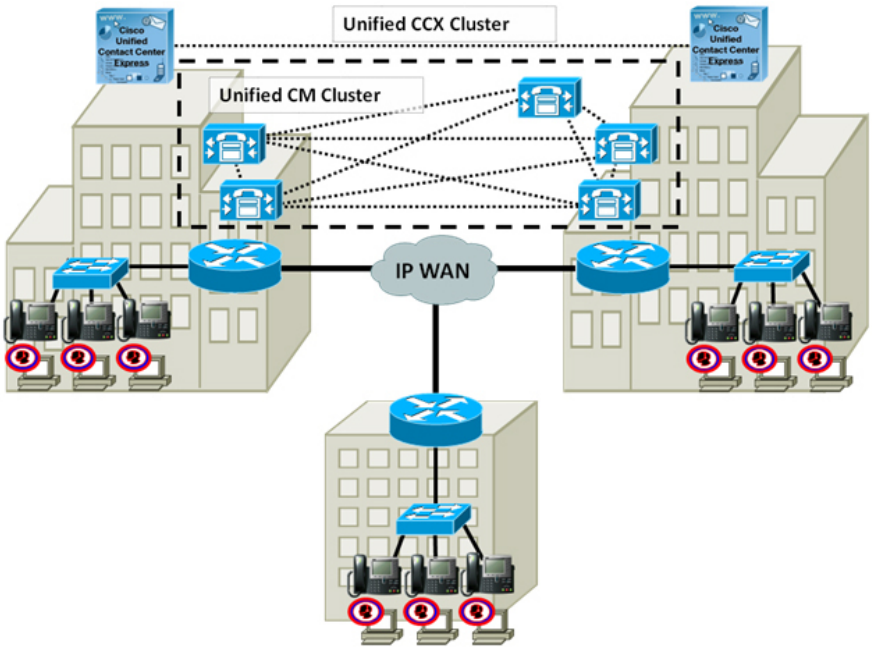
Figure 1: Unified CCX High Availability over LAN Deployment



Unified CCX High Availability over WAN

Unified CCX supports high availability over WAN to provide site redundancy. In this deployment, the Unified CCX servers are located in two different sites across the WAN. Each site should have at least one Unified CM server that is running CTI Manager with which Unified CCX communicates. The following figure depicts the deployment for Unified CCX high availability over WAN.

Figure 2: Unified CCX High Availability over WAN Deployment



Network Requirements

Observe the network requirements described in this section when deploying Unified CCX HA over WAN.

Delay

The maximum allowed round-trip time (RTT) between Unified CCX servers is 80 ms.

The maximum allowed round-trip time (RTT) between the Unified CCX server and the Unified CM server is 60 ms.

Note Do not use the ping utility on the Unified CCX server to verify RTT as it will not provide an accurate result. The ping is sent as a best-effort tagged packet and is not transported using the same QoS-enabled path as the WAN traffic. Therefore, verify the delay by using the closest network device to the Unified CCX servers, ideally the access switch to which the server is attached. Cisco IOS provides an extended ping capable of setting the Layer 3 type of service (ToS) bits to make sure that the ping packet is sent on the same QoS-enabled path that the WAN traffic will traverse. The time recorded by the extended ping is the round-trip time (RTT), or the time it takes to traverse the communications path and return. Refer to the Cisco IOS document available at

http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080093f22.shtml#extend_ping for more detail.

Bandwidth

Sufficient bandwidth must be provisioned for Unified CCX cluster, Unified CM cluster, and other optional components to deploy HA over WAN.

The following components must be accounted for, while calculating the bandwidth requirements:

- Unified CCX Cluster and Unified CM Cluster

  Unified CCX cluster consumes bandwidth between the Unified CCX servers in high availability. If the Unified CM running CTI Manager that Unified CCX communicates with is remote, there would be additional bandwidth utilized by Unified CCX.

  Unified CM could consume significantly higher bandwidth for Intra-Cluster Communication Signaling (ICCS) between sites when deploying with Unified CCX. This is due to the additional number of call redirects and CTI/JTAPI communications encompassed in the intra-cluster communications.

  Unified CCX can be deployed as ACD to route and queue contacts for available agent or as IP-IVR to perform self-service. The bandwidth requirements for Unified CCX and Unified CM clusters are different depending on the deployment type.

  The following table shows the minimum bandwidth requirement for Unified CCX and Unified CM clusters when deploying HA over WAN.

Table 1: Unified CCX HA over WAN Bandwidth Requirement

Deployment type	Unified CCX Cluster		Unified CM Cluster	
	Between Unified CCX Servers	Between Unified CCX and Remote Unified CM Servers	Database[1]	ICCS
ACD	1.2 Mbps	800 kbps	1.544 Mbps (T1)	70 kbps per 100 BHCA[2]
IP-IVR	1.2 Mbps	200 kbps	1.544 Mbps (T1)	25 kbps per 100 BHCA

[1] This column shows the database bandwidth required for Unified CM clustering over WAN and could be subject to change. For the final authorized value, refer to *Cisco Unified Communications Solution Reference Network Design (SRND)* available at: http://www.cisco.com/go/ucsrnd
[2] BHCA (Busy Hour Call Attempt) is the number of calls entering the system in the busy hour for Unified CCX or IP-IVR.

For Unified CCX Cluster in the preceding table:

- The traffic between Unified CCX servers includes database replication, heartbeat, and other communication between the Unified CCX HA servers.
- The traffic between Unified CCX server and remote Unified CM server running CTI Manager is the JTAPI call signaling.

For Unified CM Cluster in the preceding table:

- *Database* column includes traffic for database and other inter-server traffic for every Cisco Unified CM subscriber server remote to the Unified CM publisher.
- *ICCS* column shows all the ICCS traffic between CallManager/CallManager services and CallManager/CTI Manager services running in the Unified CM nodes across sites.

As an example, assume the Unified CCX HA over WAN deployment has two sites and is used as ACD. Site 1 has the Unified CCX, one Unified CM publisher and two Unified CM subscribers. Site 2 has the other Unified CCX and two Unified CM subscribers. Unified CCX in site 1 communicates with Unified CM subscriber in site 2 for JTAPI signaling. In the busy hour, there are 1500 calls coming into Unified CCX that get routed or queued for agents.

Note

- The maximum supported RTT between the Unified CCX server and the Microsoft Exchange server is 80 ms.
- A minimum bandwidth of 64 Mbps must be provisioned between two nodes of Unified CCX for intra-cluster communication and for fetching Unified Intelligence Center reports from non-master node. The maximum latency allowed is 80 ms. For bandwidth calculation, see the Cisco Unified CCX Bandwidth Calculator located at, https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-technical-reference-list.html.

For Unified CCX cluster, bandwidth required is:

1.2 Mbps + 800 kbps (0.8 Mbps) = 2 Mbps

For Cisco Unified CM cluster, there are two Unified CM subscribers remote from the Unified CM publisher and the BHCA is 1500. Bandwidth required is:

1.544 Mbps × 2 + 70 kbps × 15 (1.05 Mbps) = 4.138 Mbps

In total, 6.138 Mbps between sites is required for this deployment.

- Agents and Supervisors

  In HA over WAN deployment, agents and supervisors could reside in either Unified CCX sites or they could be remote depending on the location of active Unified CCX server at the time of operation.

Bandwidth should be provisioned for remote agents between sites using the maximum number of agents from the two sites. Estimate the required bandwidth using the Unified CCX Bandwidth Calculator available at:

https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-technical-reference-list.html

- Optional Components

  Customers might have the following optional components deployed across the WAN from Unified CCX or Unified IP IVR. Ensure to account for the additional bandwidth required in their HA over WAN deployment.

  - **Wallboard Server**— Determine the amount of data that is retrieved from Unified CCX database to the remote wallboard server.
  - **Enterprise Database**— Estimate the total amount of data that is retrieved through the database steps from the remote enterprise database.
  - **SMTP Server**— If the SMTP server is remote from the Unified IP IVR, determine the average size of each outgoing email and calculate the total.

- To calculate bandwidth for Finesse, see the *Unified CCX Bandwidth Calculator* , available at:

  https://www.cisco.com/c/en/us/support/customer-collaboration/unified-contact-center-express/products-technical-reference-list.html

Quality of Service

Quality of Service (QoS) must be enabled and engineered correctly on the network to provide consistent and predictable end-to-end levels of service. Unified CCX software does not mark any network packet, so ensure that you mark the traffic at the network edge routers.

The following table shows the QoS markings for Unified CCX HA over WAN deployment.

Table 2: QoS Considerations for Unified CCX HA Over WAN

Traffic	QoS Marking
JTAPI Call Signaling	IP Precedence 3 (DSCP 24 or PHB CS3)
Database Replication between Unified CCX nodes[3]	IP Precedence 0 (DSCP 0 or PHB BE)

[3] The database traffic may be reprioritized to a higher priority data service (for example, IP Precedence 2 [DSCP 18 or PHB AF21] if required by the particular business needs). An example of this is the usage of outbound dialer in Unified CCX, which relies on writing data to the Config Datastore.

For more information on QoS requirements of VoIP, refer to the Enterprise QoS Solution Reference Network Design Guide available here:

http://www.cisco.com/en/US/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND/QoSIntro.html#wp46447

Deployment Considerations

Consider the following when deploying high availability over WAN with Unified CCX:

- Deploy the ASR or TTS server locally in each Unified CCX site
- Set up Unified CCX to use the local Unified CM servers for both primary and secondary in the following configurations. If this is not possible, at least the primary Unified CM server should be local.
  - AXL Service Provider
  - JTAPI Provider for Unified CM Telephony Subsystem
  - JTAPI Provider for Resource Manager/Contact Manager Subsystem

Note Significant delays in agent login will occur during Unified CCX failover if AXL and JTAPI communications are over WAN, especially under load conditions.

- Assign the two sets of CTI Ports (one for the master and the other for the standby engine) to different device pools, regions and locations, in the CTI Port Group.
- Data in Historical Datastore and Repository Datastore start merging after the network partition is restored. This situation could potentially generate heavy data traffic over WAN. Restore the WAN link during after hours to minimize the performance impact.
- Do not support VPN tunneling across the WAN.

Unified CCX-Finesse deployment

Cisco Finesse is supported in both single-node deployment and high-availability deployment over LAN and WAN.

Unified CCX-Cisco Customer Collaboration Platform deployment

Customer Collaboration Platform (CCP) doesn't have an HA but works with a Unified CCX in HA.

Engine Redundancy

Any incoming call arriving at Cisco Unified Communications Manager that are destined for Unified CCX route points can be accepted by the Unified CCX engine if all Unified CCX call treatment and ACD routing services are operational.

If the active Unified CCX server fails, the ACD subsystem will not be able to route calls to the agents until the automatic logging in process completes. The agents are then logged in back to the same state (Ready or Not Ready) that they were in before the failover. However, if the agent was in an active call, they are logged back into the Not Ready State and the call continues uninterrupted.

When the Master Engine is Down

Once the master engine goes down, the engine on the other node will be selected as the new master. Calls which were queued by the previous primary engine are dropped after a failover. New calls coming in while agents are re-logging will stay in the queue until agents log in. Historical data will be written to the new master engine’s local database.

Automatic Call Distribution (ACD)

The HA failure of the active server is detected and the ACD subsystem can automatically fail over from the active to the standby server. All ACD functions are restored on the standby server within 5 seconds.

Interactive Voice Response

When an active server fails in a HA system, IVR subsystem will automatically failover.

All calls in queue and calls receiving IVR call treatment will be lost. Calls already transferred to the agent will be preserved.

Unified CCX Outbound Dialer

Behavior Under High Availability

The Config Data Store (CDS) is required for general operation of outbound for call status and call result updates of contact records. When deploying in a two-node high availability system, the CDS must be running on both nodes to enable the database write operation. The Outbound subsystem will be operational as long as the Publisher CDS is up and running. In a high availability environment, only the dialer in the master node is active.

If a contact is imported for a campaign and failover occurs before the contact is dialed out, then the contact is retried the next day. The number of contacts that can be retried for each campaign is as mentioned below:

- For direct preview campaigns, the count is the maximum value that is configured for Contact Records Cache Size field.
- For IVR-based progressive and predictive campaigns, the count is the Number of Dedicated Ports multiplied with the Lines Per Port (LPP) values configured.
- For agent-based progressive and predictive campaigns, the count is 45 for medium or large VM profiles and is 15 for small VM profiles.

Failover Scenarios for Preview Outbound:

- If a preview outbound call not in reserved state is waiting for the agent to accept the call and when the master engine goes down, the agent is automatically logged out and the preview call disappears from the agent desktop. If the master engine restarts during failover, the call status for that contact record is set to unknown. If the master engine does not restart during failover, the contact is called when the campaign starts and there are available agents.
- If a preview outbound call not in reserved state is accepted by the agent and the call is ringing on the customer phone, there is no change on the call. However, the agent is logged out and will be able to use call control capabilities only through the phone.

Failover Scenarios for Progressive and Predictive IVR-Based Outbound:

- The CTI ports on the master engine will go out of service on a failover and the calls that are in progress between customers and CTI ports will be disconnected. The standby server will continue dialing out the remaining contacts in the campaign after the failover.

Failover Scenarios for Progressive and Predictive Agent-Based Outbound:

- If an agent is currently on an outbound call and Cisco Finesse service restarts or agent closes the browser and reopens, then the agent is automatically logged in after 60 seconds and the state of the agent is set to Not Ready. If the customer is still on the call, then the agent continues to handle the call but outbound specific options will not be available on the agent desktop.

WAN Link Failure Between Sites—Island Mode

Connectivity failure creates a scenario called 'Island mode' where each node (on either side of the network) assumes mastership and handles calls. Each node behaves as if the other side has failed and declares itself master (Engine and Data Stores components). The node that was already the master, continues as is. Phones and Finesse need to register with Unified CM and server on the same side of the network. This operation happens automatically. The following lists the failover behaviors:

- Historical data is written to local Data Stores
- Real Time Reporting (RTR) shows the status of each node independently
- No configuration changes are allowed
- Enterprise Database access across the network is not possible
- Outbound will be impaired as these do not support high availability

If the Island mode occurs for more than four days, DB replication between the nodes will be broken and will need to be reestablished from Unified CCX Administration web interface when the WAN link is restored.

Note Backup scripts are executed on the publisher, and it backs up the database that has mastership. In Island mode, only one node gets backed up and the data getting collected on the other node does not get backed up. The backup is inconsistent, and if restored, there will be loss of data.

When Connectivity is Restored

Once the network connectivity is restored, convergence of engine mastership occurs. Two masters cannot exist and one of the nodes will drop mastership. All active calls being handled by that node will be dropped.

Similarly, convergence happens for the data stores with no disruption in call activity. All data will be replicated as soon as convergence is done only if the link was up within a predetermined replication retention period, otherwise, the customer needs to initialize the replication from datastore control center pages.

Tip You can use the Unified CCX Administration Datastore control center pages or the CLI to check the replication status.

WAN Link and Single Engine Failure

When the WAN goes down, CTI functionality, which was provided by Unified CM Sub 1 across the WAN is no longer available. The master engine on node 2 fails over to Unified Communications Manager Sub 2. All calls still in the queue are dropped.

Some agents will remain in Not Ready state since the corresponding agent's phones are registered with the Unified Communications Manager Sub 1. There is no automatic function to force phones to re-register.

This situation is corrected when the WAN link is restored.

Chat and Email

With high availability, failure of the active server can be detected and the Web Chat subsystem automatically fails over from the active server to the standby server. All unanswered chats are moved to the new active server.

An active chat session is available until the browser gets redirected to the standby server. The chat session is terminated after the redirect is complete and the message is displayed as, 'The Chat has Ended.' During an engine failover, the agent gets a message that, 'Chat and Email are temporarily down due to Outages.' All queued contacts are discarded in Chat whereas it is reinjected in Email.

The fault tolerance for Web Chat is provided in the Unified CCX. In an HA deployment, Customer Collaboration Platform is configured to communicate with both the Unified CCX nodes. When a new contact arrives at Customer Collaboration Platform, both the Unified CCX nodes are notified.

In the case of a failover, all emails that were previously queued and were assigned to an agent get requeued and get assigned to the agents.

Attention Cisco Customer Collaboration Platform does not support HA deployment options. Chat and email will not be available if Customer Collaboration Platform is down.

Note Web Chat and email do not support the Island mode scenario.

Cisco Finesse High Availability Considerations

This section describes Cisco Finesse Desktop behavior during the failover of one of the following:

- Unified CCX Engine failure
- Cisco Finesse Server failure
- Notification Service failure
- CM failure
- Netwrok failure between two centers
- Network failure between agent and supervisor desktop

For HA deployment the Cisco Finesse is in service on both the nodes, based on the following requirements:

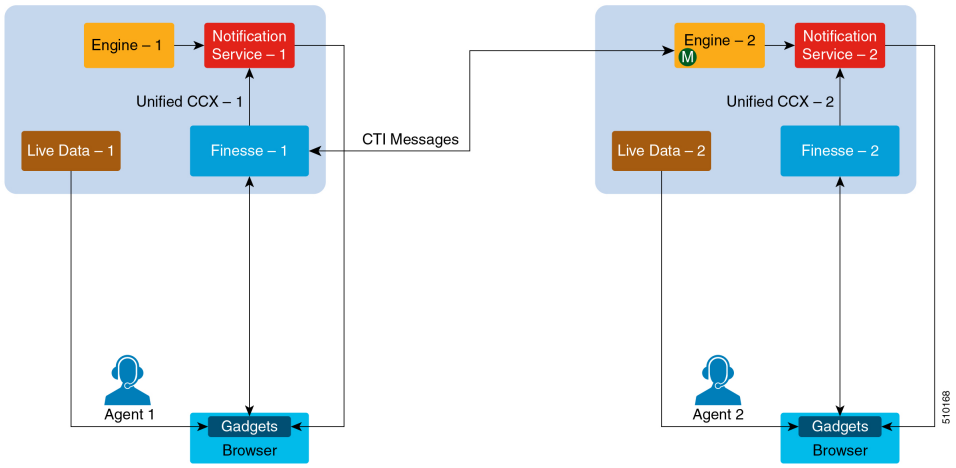
- On Unified CCX Master node:
  - Finesse is in In Service if its connected to the Unified CCX engine, and Notification service on the same node.

- On Unified CCX Non-Master Node:
  - Finesse is in In Service if it is able to connect to the Master CCX engine (remote).
  - Unified CCX engine is running in slave mode on the Non-Master node.
  - Notification Service is running on Non-Master node.

This enables the desktop failover to happen independent of the engine mastership status.

The following image describes the high level view of the Cisco Finesse Desktop Failover scenario.

Figure 3: High Level View Architecture Diagram of Cisco Finesse Desktop Failover Scenario



Note Agent logging into both the Finesse nodes simultaneously is not supported.

Unified CCX does not support load balancing of agent login. All agents must log in to the Master Node only. The enhancement in behavior is for failover support only.

For voice calls, Finesse connects to CTI server on master CCX Engine.

For agents to log in to Finesse in the secondary node, Notification Service, Engine and LiveData services (Socket.IO Service) have to be running on the secondary node.

Failure Scenarios in HA Deployment

This table describes failure scenarios that you might encounter in high availability deployment when the Unified CCX Node 1: Master, Unified CCX Node 2: Non-master before failover.

Failure Scenario	Unified CCX HA Behavior	Cisco Finesse Service on Node 1	Cisco Finesse Service on Node 2	Cisco Finesse Client Behavior

Unified CCX Engine fails over from master to Non-Master node	Engine Mastership changes from master to non-master.	Finesse goes **OUT_OF_SERVICE** and will return to **IN_SERVICE** as soon as it connects to the new master engine.	Finesse goes **OUT_OF_SERVICE** and will return to **IN_SERVICE** as soon as it connects to the new master engine.	Agent would see the red disconnection bar, and will automatically relogin into the Finesse side that comes to **IN_SERVICE** first. It can be either Node 1 or Node 2.
Unified CCX Engine Failure on Non-Master node	Unified CCX Engine on Node 1 continues to be Master, thus no change.	Finesse will continue to be in **IN_SERVICE**.	Finesse goes **OUT_OF_SERVICE**.	Agents connected to Node 1 will continue to be logged in. Agents connected to Node 2 will be temporarily disconnected and will connect to the Finesse Service on the node that is **IN_SERVICE**.
Finesse Service Out of Service on Node 1	Engine mastership is not affected.	**OUT_OF_SERVICE**	Finesse on Node2 will continue to be in **IN_SERVICE**.	Any agents connected to Node 1 will be temporarily disconnected and will connect to Finesse on Node 2. Agents connected to Node 2 will not be impacted.
Finesse Service Out of Service on Node 2	Engine mastership is not affected.	Finesse on Node 1 will continue to be in **IN_SERVICE**.	**OUT_OF_SERVICE**	Any agents connected to Node 2 will be temporarily disconnected and will connect to Finesse on Node 1. Agents connected to Node 1 will not be impacted.
Unified CCX Notification Service Failure on Node1	Engine mastership is not affected	**OUT_OF_SERVICE**	Finesse on Node 2 will continue to be **IN_SERVICE**.	Any agents connected to Node1 will be temporarily disconnected and will connect to Finesse on Node 2. Agents connected to Node 2 will not be impacted.

Unified CCX Notification Service Failure on Node 2	Engine mastership is not affected	Finesse on Node 1 will continue to be **IN_SERVICE**	Finesse service on the node where notification service failed, remain **OUT_OF_SERVICE** until notification service comes up.	Any agents connected to Node 2 will be temporarily disconnected and will connect to Finesse on Node 1. Agents connected to Node1 will not be impacted.
Island Mode	Both HA nodes become Master	Finesse on Node 1 will continue to be **IN_SERVICE** and will be connected to Engine on Node 1.	Finesse goes Out Of Service and will return to **IN_SERVICE** as soon as it connects to the engine on Node 2 which is also the master.	Agents connected to Node 1 will continue to be logged in. Agents connected to Node 2 will be temporarily disconnected and will connect to the Finesse Service on the Node 2.

Finesse IP Phone Agent Failure Behavior

The Finesse IP Phone Agent does not automatically failover to the alternate Finesse server. To ensure continued operations in a failure situation, you must configure at least two Finesse IP Phone services in Unified CM, each pointing to different Finesse servers.

When the Finesse server fails, Finesse IPPA attempts to reconnect to it every 5 seconds. After three attempts, if the Finesse server is not in service, Finesse IPPA displays a server unavailable message to the agent. The total time to go out of service is approximately 15 seconds.

In a failure scenario, the Finesse IPPA agents must exit from the current Finesse service and manually sign in to another configured Finesse service that points to an alternate Finesse server. After they successfully sign in to an alternate Finesse service, the agents can resume usual operations.

Cisco Unified Intelligence Center High Availability Considerations

Server is Down

In a two-node high availability (HA) setup, you can connect to any node to access reports. If the node you are connected to goes down, then manually log in to the other node to access reports as this doesn't happen automatically.

Island Mode

If WAN is down, the nodes function in Island mode and both of the nodes independently assume mastership (engine and data stores components). You can access reports from either of the nodes.

Note There will be a data discrepancy in the reports as there is no data replication between the nodes till the connectivity is restored.

Standalone CUIC has no high availability.

Customer Collaboration Platform High Availability Considerations

Cisco Customer Collaboration Platform (CCP) does not support high availability.

CCP uses either a small or large, single-server, all-in-one, deployment. You cannot use a load-balancing, split site deployment.

ASR TTS High Availability Considerations

solution supports redundant ASR/TTS servers. In a basic configuration, the VXML Gateway first passes all incoming requests to the primary ASR/TTS server. If the primary server is unreachable, the gateway then passes that request to the backup server. Any request that reaches the backup server stays on that server for the duration of the request.

You can add a load balancer to spread the incoming requests across your ASR/TTS servers.

Cisco IM&P High Availability Considerations

Failover is supported for Desktop Chat and any Cisco IM&P node failure results in automatic connection to the node pair peer, as configured for the user.

Desktop Chat Failover

The following table lists the desktop chat failover scenarios:

Failover Type	Desktop Chat Behavior
Cisco IM&P server failover	The desktop chat status is retained, and all active chat sessions are lost.
Finesse server failover	The desktop chat status is retained, and all active chat sessions are lost.
server failover	The desktop chat status and all chat sessions are retained.

See the Cisco Finesse Administration Guide for failover details with Desktop Chat.