Overview of Call Control and Routing

Revised: June 15, 2015

Once the network infrastructure has been put in place for your Cisco Unified Communications and Collaboration System, call control and routing applications, components, and services can be layered on top of that infrastructure. There are numerous applications and features that can, and in some cases must, be deployed on the network infrastructure:

- Bandwidth management — Provides mechanisms for ensuring voice and video quality and for preventing oversubscription of network bandwidth by limiting the number of calls that are allowed on the network at a given time. With a combination of packet marking and re-marking, and advance queuing mechanisms such as low latency or priority queuing, voice and video quality is guaranteed. Similarly, call admission control enforces the overall call capacity of the call processing components and available network bandwidth.

- Dial plan — Provides endpoint numbering, dialed digits analysis, and classes of restriction to limit the types of calls that a user can make.

- Emergency services — Provide essential information about the caller's location and emergency situation to the appropriate Public Safety Answering Point (PSAP) so that the caller receives a swift response and the necessary help (for example, police, fire, or ambulance teams).

- Directory and identity management services — Lightweight Directory Access Protocol (LDAP) provides applications with a standard method for accessing and potentially modifying the information stored in a directory. Likewise, identity management and single sign-on ensure that user identification and access are secure. These capabilities enable companies to centralize all user information in a single repository available to multiple applications, resulting in better access to the information and a reduction in maintenance costs through the ease of making adds, moves, and changes.

The chapters in this part of the SRND cover the features, components, and services mentioned above. Each chapter provides an introduction to the component or service, followed by discussions surrounding architecture, high availability, and design considerations. The chapters focus on design-related aspects of the applications and services rather than product-specific support and configuration information, which is covered in the related product documentation.

This part of the SRND includes the following chapters:

- Bandwidth Management, page 13-1

  This chapter examines bandwidth management techniques and the potential for oversubscribing IP links, which causes the voice and video quality of calls to become unacceptable. It also examines the use of call admission control to allow only a certain number of simultaneous calls on the network at a given time to prevent oversubscription. This chapter covers quality of service (QoS) and call admission control types, including location-based call admission control, as well as design and deployment guidelines for successfully deploying QoS and admission control services.

- Dial Plan, page 14-1

  This chapter explores dial plan features and functions that enable the call processing application to route calls to appropriate destinations. The chapter considers various aspects of dial plan services, including dial plan constructs, dial plan numbering options and design considerations, classes of restriction, inbound and outbound calling features, and dial plan and call routing redundancy mechanisms.

- Emergency Services, page 15-1

  This chapter discusses accessing emergency services through Public Safety Answering Points (PSAPs) on the PSTN from within the enterprise IP communications environments, an important aspect of most deployments due to possible critical needs for medical, fire, and other emergency response services. The chapter provides an overview of the various emergency service components both inside and outside the enterprise. It also discusses planning, 911 network service providers, gateway interfaces, and number-to-location mapping.

- Directory Integration and Identity Management, page 16-1

  This chapter covers aspects of Unified Communications and Collaboration integration with the LDAP directories, including the Cisco Unified Communications Manager directory architecture itself, as well as design considerations for LDAP synchronization and authentication. Directory access from Unified Communications and Collaboration endpoints as well as security considerations such as single sign-on are also explored.

Architecture

Call routing components and services such as call processing agents and IP and PSTN gateways rely on the underlying network infrastructure for network connectivity and access. By connecting to the underlying network infrastructure, call routing components and features are able to leverage end-to-end network connectivity and quality of service to access both the enterprise and public telephone networks. In turn, call routing applications and services provide basic Unified Communications and Collaboration functions such as call control, dial plan, call admission control, and gateway services to other applications and services in the deployment. For example, a Unified CM cluster connects to the IP network through a switch in order to communicate with other devices and applications within the network as well as to access other devices and services in other locations. At the same time, the Unified CM cluster provides services such as phone registration and media resource provisioning and allocation to call control components and services such as IP phones.

Further, just as call routing components rely on the network infrastructure for network connectivity, call routing components and services are also often dependent upon each other for full functionality. For example, while Unified CM provides registration and call routing services to various IP endpoints within the network, it is completely dependent upon gateways and gateway services to route calls beyond the enterprise.

High Availability

As with the network infrastructure, critical Unified Communications and Collaboration call routing services should be made highly available to ensure that required features and functionality remain available if failures occur in the network or with individual call routing components. It is important to understand the various types of failures that can occur and the design considerations around those failures. In some cases, the failure of a single server or component (for example, a subscriber node in a Unified CM cluster) might have little or no impact due to the redundant nature of the Unified CM clustering mechanism. However, in other cases a single failure can impact multiple components or services. For example, the failure of a PSTN or IP gateway could result in loss of access to the public telephone network, and even though a call processing agent such as Unified CM is still available and able to provide most features and services, it cannot route calls to the PSTN because there is no path available if the gateway fails. To avoid these types of situations, you should deploy multiple PSTN gateways to provide redundant gateway services, and you should configure the call processing agent to handle call routing to both gateways as needed.

For features and services such as dial plan and bandwidth management, high availability considerations include temporary loss of functionality due to network connectivity or call processing agent application server failures, which result in the inability of the call agent to route calls and therefore the inability for callers to make calls. Oversubscription of the network could also occur if QoS and other call admission control services are not available to the endpoints initiating calls. For example, if the call admission control agent fails or loses connectivity to the network, the call may still go through but without the call admission control service being aware of the call, thus potentially resulting in poor quality. To avoid these types of scenarios, provide call admission control resiliency by deploying multiple call admission control agents.

High availability considerations are also a concern for components and services such as video endpoints and remote site survivability. For deployments with network-attached remote sites where devices are leveraging call processing services from an agent in a central site, remote site survivability using SRST, for example, can ensure that local phones within the remote site will still receive call processing services in the event of a connectivity failure to the central site. Likewise, to ensure that video endpoints are highly available, you can deploy more than one multipoint control unit (MCU) in case one fails.

Capacity Planning

The network infrastructures must be designed and deployed with consideration for the capacity and scalability of the individual components and the overall system. Similarly, deployments of call routing components and services must also be designed with attention to capacity and scalability considerations. When deploying various call routing applications and services, not only is it important to consider the scalability of the applications and services themselves, but you must also consider the scalability of the underlying network infrastructure. Certainly the network infrastructure must have available bandwidth and be capable of handling the additional traffic load that the call routing components will create. Similarly, the call routing infrastructure and its components must be capable of handling all the required device configurations and registrations as well as the call load or busy hour call attempts (BHCA),

For example, with call processing agents such as Unified CM, it is critical to assess the size of the deployment in terms of number of users, endpoints, and calls per user per hour, and to deploy sufficient resources to handle the required load. If a call processing agent is undersized and does not have sufficient resources, features and services will begin to fail as the load increases. Two of the chief considerations when attempting to size a call processing deployment are the call processing type and the call processing hardware. Both of these are critical for sizing the system appropriately given the number of users, locations, devices, and so on. As an example, Cisco Unified Communications Manager has a much

higher capacity than Cisco Unified Communications Manager Express and should therefore be used for larger deployments. In addition, the server platform selected to run the call processing agent will, in many cases, determine the maximum load.

Capacity planning for remote site survivability is much the same in that it relies on backup call processing hardware. Selecting the appropriate Cisco IOS platform to provide backup or survivable call processing services typically begins with determining the number of devices and users that must be supported at that site in the event that connectivity to the central site is disrupted. Equally critical in this sizing exercise are the local PSTN gateway services. In the event of a central site connection failure, will the local PSTN gateway have sufficient circuits to be able to route all calls without blocking during the busiest hour? If the answer is no, adding more gateways or trunks will be necessary to size the remote site appropriately for backup call processing.

PSTN and IP gateways must also be sized appropriately for a deployment, so that sufficient capacity is available to handle all calls in the busiest hour. In some cases, you might have to deploy multiple PSTN or IP gateways to provide enough resources.

When configuring and sizing QoS and call admission control services, ensure that sufficient bandwidth is available over network connections and in priority queues to support the required number of calls. If sufficient bandwidth is not available, additional network capacity, gateways, and IP or telephony trunks might be required.

Sizing dial plan services is also important. However, in most cases dial plan capacity in terms of the number of endpoints or phone numbers, route patterns, or other dial plan constructs, is completely dependent upon the type of call processing agent and platform used.

For components and services such as video telephony, appropriate sizing is just as critical. Capacity planning considerations for video telephony focus mainly on network bandwidth, available video ports, and MCU sessions. In most cases additional capacity can be added by increasing the number of application servers and MCUs or by upgrading server or MCU hardware with higher-scale models, assuming the underlying network infrastructure is capable of handling the additional load.

For a complete discussion of system sizing, capacity planning, and deployment considerations related to sizing, refer to the chapter on Collaboration Solution Sizing Guidance, page 25-1.