Chapter 1

Cisco Collaboration Solution Multisite Deployment Considerations

When deploying Cisco Unified Communications (UC) in a multisite environment, some unique aspects and design considerations need to be addressed and considered. Any information technology (IT) professional would likely admit that *complexity* is the last word anyone wants to hear in his or her data center and IT environment. Unified Communications can be broken down into components or building blocks. A multisite deployment implementation can be achieved if you properly plan and follow best practices.

Upon completing this chapter, you will be able to explain issues pertaining to a UC multisite deployment. Once you understand the issues, possible solutions are provided. Where applicable, best practices are mentioned according to Cisco Solutions Reference Network Designs (SRNDs) as well as Cisco Validated Designs (CVDs). Each recommended architecture is explained in greater detail throughout the remainder of this book.

Deploying a multisite UC environment requires a deep understanding of how to craft a proper multisite dial plan that allows for scalability, proper planning for bandwidth allocation for not only IP phones but also video endpoints, quality of service (QoS) design and implementation, and a highly available wide-area network (WAN) and local-area network (LAN) architecture, including survivable remote site telephony (SRST). This chapter helps identify issues that arise in multisite UC deployments.

Upon completing this chapter, you will be able to meet these objectives:

- Describe aspects that pertain to multisite deployment
- Describe QoS aspects in a multisite deployment
- Describe bandwidth aspects in a multisite deployment
- Describe availability in a multisite deployment
- Describe dial plan aspects in a multisite deployment
- Describe fixed-length versus variable-length numbering plans

- Describe how to optimize call routing and implement PSTN backup solutions
- Describe overlapping and nonconsecutive numbering plans
- Describe various PSTN requirements
- Describe how to create a scalable dial plan architecture
- Describe NAT and possible security issues in modern unified communications

Multisite Deployment Issues Overview

The goal of any successful business is to grow; usually this entails expansion and possibly adding sites or locations. In today's modern IT environments, the pace of expansion and the pressure of delivering new technologies to business units can be overwhelming at times. This is only compounded by a different type of end user coming into the workforce (the bring your own device [BYOD] end users who want to connect their personal devices to the corporate network and work in a manner that is efficient and effective for them). Figure 1-1 illustrates several issues with multisite deployments, including availability, quality, and bandwidth concerns, dial plan issues, and security concerns.

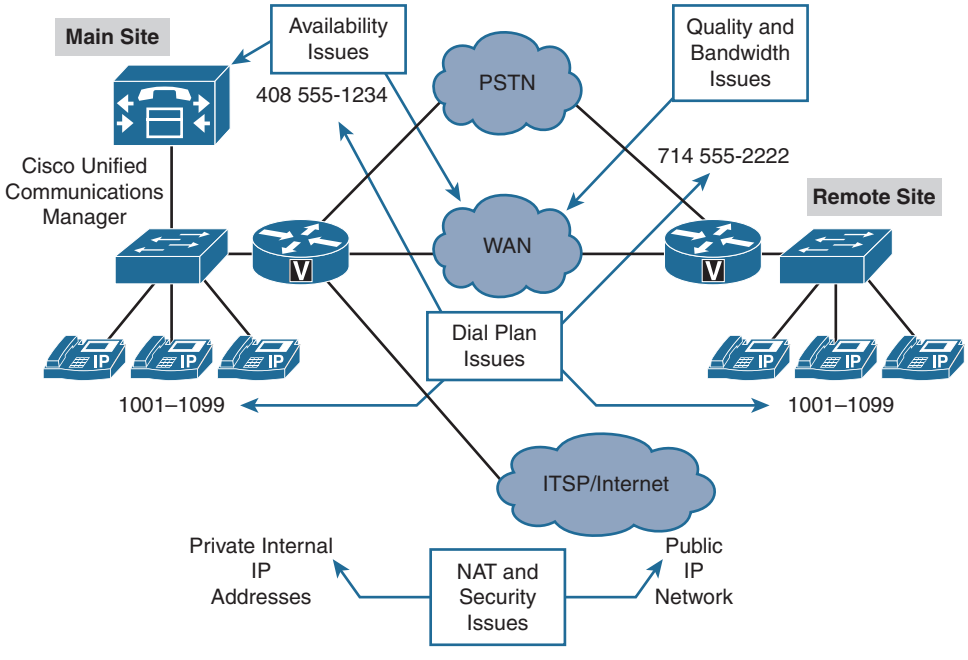

Figure 1-1 *Common Issues in Multiple-Site Deployments*

To provide workers with unified communications, video-capable devices, instant messaging (IM), voicemail, call centers, and enterprise-grade communication features, IT shops are challenged to support multiple systems and oftentimes multiple sites. In a multisite deployment, several issues can occur if not properly planned for:

- Quality issues
- Bandwidth
- Availability
- Dial plan
- NAT and security

Voice and video communications are considered real-time communications; they utilize the User Datagram Protocol (UDP), and more specifically the Real-time Transport Protocol (RTP). Think of RTP as a "fire and forget" protocol; if a packet is transmitted, it needs to be prioritized over a packet-switching network. It will not tolerate delay and will not be retransmitted. All traffic is treated equally by default in routers and switches. Due to voice and video being delay-sensitive packets, they must be given priority over other network traffic. QoS is a network and Unified Communications engineer's best friend when it comes to mitigating call or video quality issues. QoS allows you to prioritize voice and video over other types of network traffic. Cisco has a best practice architecture for QoS deemed Medianet. The Enterprise Medianet Quality of Service design principles are beyond the scope of this text; however, QoS is an important topic in Unified Communications and a properly planned collaboration solution.

Cisco Unified Communications (UC) can include voice and video RTP streams, signaling traffic, management traffic, and application traffic (such as rich media conferencing). The additional bandwidth that is required when deploying a Cisco UC solution has to be calculated and provisioned for to ensure that data applications and Cisco UC applications do not overload the available bandwidth. Bandwidth reservations can be made at a network level through proper QoS deployment and technologies such as Resource Reservation Protocol (RSVP). Bandwidth reservations can also be made at the Unified Communications application level by implementing Call Admission Control (CAC) and selecting the proper codec for voice and video calls. As of Unified Communications Manager 9x./10.x/11.x, newer technologies such as Enhanced Locations CAC, Intracluster Enhanced Locations CAC, and Intercluster Enhanced Locations CAC are available. These newer technologies are discussed in Chapter 7, "Call Admission Control (CAC) Implementation."

When deploying Cisco Unified Communications Manager (CUCM) with centralized call processing (servers in a main or headquarters site or data center with multiple branch or remote sites without local call processing servers), IP phones register with CUCM over the IP LAN and potentially over the WAN. If voice gateways such as Integrated Service Routers (ISRs) or Aggregation Services Routers (ASRs) in remote sites are using Media

Gateway Control Protocol (MGCP) as a signaling protocol, they also depend on the availability of CUCM acting as an MGCP call agent. Certain analog voice cards such as voice interface cards (VICs), which provide plain old telephone system (POTS) capability as well as high density analog devices (such as VG 350s), can register to CUCM using Skinny Client Control Protocol (SCCP), which is dependent on the communication path to CUCM. It is important to implement fallback solutions for IP phones and gateways in scenarios in which the connection to the CUCM servers is broken because of IP WAN failure. One common technique is to implement a highly available WAN as well as provide a feature on the ISR/ASR routers called survivable remote site telephony (SRST). SRST allows a gateway at a remote site to become the call-processing engine in the event of a WAN failure. The ISR/ASR router provides registration and call-processing capabilities to Cisco IP phones as well as certain virtual interface cards (VICs) and HD analog gateways. Fallback solutions also apply to H.323 or Session Initiation Protocol (SIP) gateways but require the correct dial peers to support this functionality. Each failover technology is examined in later chapters.

The goal of a properly designed Unified Communications dial plan is to limit "dial plan overlap," meaning users typically have unique extensions or directory numbers (DNs). There are techniques in CUCM in which the same extension can exist inside the same partition. In the event this occurs, the DN is considered a shared line. Unified Communications engineers typically design a single site in which each user has a unique DN inside a common partition for that site. When you design a multisite deployment or global deployment, DNs can overlap across multiple sites, the difference being these DNs are often in separate partitions and are separated out logically in the CUCM dial plan and database. A partition is a logical container (think of it as a padlock over a container) in which DNs, route patterns, meet-me numbers, voicemail ports, and so on can be placed. A calling search space (CSS) is the key by which IP phones and video endpoints are granted permission that allows them to dial certain numbers or unlock those partitions. A design challenge arises in multisite deployments regarding overlapping DNs, variable-length dial plans, various public switched telephone network (PSTN) access codes, and nonconsecutive numbers. Each of these challenges can be solved by designing a robust multisite dial plan. Some techniques used to mitigate these issues include site access codes, a properly planned extensions length, translation patterns, and proper route patterns. Each technique is examined in later sections. In general, avoid overlapping numbers across sites whenever possible for an efficient design.

Cisco Unified Communications and Unified IP Phones/Video endpoints use IP and private IP addresses primarily to communicate within the enterprise. One issue arises in a multiple-site deployment when the various UC systems need to interact and communicate with devices or businesses on the public Internet. Some UC examples include instant messaging (IM) in the form of Cisco Jabber and video business-to-business (B2B) communication in the form of Cisco Expressways or Video Communications Servers (VCS). Last but not least is the Internet telephony service providers (ITSPs), which rely on SIP trunks versus primary rate interface (PRI) or POTS telephone lines to provide communication paths into modern IT environments. SIP trunks are likely terminated onto Cisco Unified Border Elements (CUBEs), which can be a demarcation point between

the private and public networks. Security and firewall concerns have become paramount recently with spikes in global hacking. To provide secure communications, the private IP addresses within the enterprise must be translated into public IP addresses. Public IP addresses make the IP phones and video endpoints visible from the Internet and therefore subject to attacks. Network Address Translation (NAT) is one of the preferred technologies of allowing public devices and connections through the firewall and security policies to communicate with internal IP phones and video endpoints. NAT challenges and design considerations are discussed in depth in later sections.

Note The challenge of NAT and security is not limited to multisite deployments. Voice over IP (VoIP) and communications protocols such as Media Gateway Control Protocol (MGCP), Skinny Client Control Protocol (SCCP), H.323, and Session Initiation Protocol (SIP) all require design considerations any time their traffic is subjected to NAT and their traffic traverses through a CUBE or firewall. In addition, some larger environments may invoke security in the data center in the form of virtual firewalls to segment traffic from the network to various sections of the data center. Special design considerations are required for voice and RTP any time a NAT translation occurs. Video devices are especially problematic to NAT traversal and translation, and separate techniques are addressed for video devices.

Voice and Video Call Quality Issues

IP networks were not originally designed to carry real-time traffic. Instead, they were designed for resiliency and fault tolerance. Transmission Control Protocol (TCP) is a great example of this; if a packet fails to be delivered, we simply retransmit. This technique does not work with User Datagram Protocol (UDP) and Real-time Transport Protocol (RTP) protocols, which carry voice and video over IP. It makes no sense to receive the same word over and over again in a conversation just because it was delayed or, worst case, dropped. Each packet is processed separately by a router or Layer 3 switch in an IP network, sometimes causing different packets in a communications stream or word to take different paths to the destination. Imagine a scenario where a branch office has redundant MPLS providers back to a main site, using various router load-balancing protocols and high availability. It is entirely possible the traffic would take different paths to the same destination. The different paths in the network may have a different amount of packet loss, delay, and delay variation (jitter) because of bandwidth, distance, and congestion differences. The destination must be able to receive packets out of order and sequence them. This challenge is solved by the use of RTP sequence numbers, ensuring proper reassembly and playout to the application. When possible, it is best to not rely solely on these RTP mechanisms. Proper network design, using Cisco router Cisco Express Forwarding (CEF) switch cache technology, performs per-destination load sharing by default. Per-destination load sharing is not a perfect load-balancing paradigm, but it ensures that each IP flow (voice call) takes the same path.

Another common design consideration is that bandwidth is shared by multiple users and applications; the amount of bandwidth required for an individual IP flow varies significantly during short lapses of time. Most data applications are bursty by nature, whereas Cisco real-time audio communications with RTP use the same continuous-bandwidth stream. The bandwidth available for any application, including CUCM and voice-bearer traffic, is unpredictable. During peak periods, packets need to be buffered in queues waiting to be processed because of network congestion. *Queuing* is a term that anyone who has ever experienced air flight is familiar with. When you arrive at the airport, you must get in a line (queue) because the number of ticket agents (bandwidth) available to check you in is less than the flow of traffic arriving at the ticket counters (incoming IP traffic). If congestion occurs for too long, the queue (packet buffers) gets filled up, and passengers are annoyed. (Packets are dropped.) Higher queuing delays and packet drops are more likely on highly loaded, slow-speed links such as WAN links used between sites in a multisite environment. Quality challenges are common on these types of links, and you need to handle them by implementing QoS. Without the use of QoS, voice packets experience delay, jitter, and packet loss, impacting voice quality. It is critical to properly configure Cisco QoS mechanisms end to end throughout the network for proper audio and video performance.

During peak periods, packets cannot be sent immediately because of interface congestion. Instead, the packets are temporarily stored in a queue, waiting to be processed. The amount of time the packet waits in the queue, called the queuing delay, can vary greatly based on network conditions and traffic arrival rates. If the queue is full, newly received packets cannot be buffered anymore and get dropped (tail drop). Figure 1-2 illustrates tail drop. Packets are processed on a first-in, first-out (FIFO) model in the hardware queue of all router interfaces. Voice conversations are predictable and constant (sampling is every 20 milliseconds by default), but data applications are bursty and greedy. Voice, therefore, without any special QoS or queuing mechanism, is subject to degradation of quality because of delay, jitter, and packet loss.

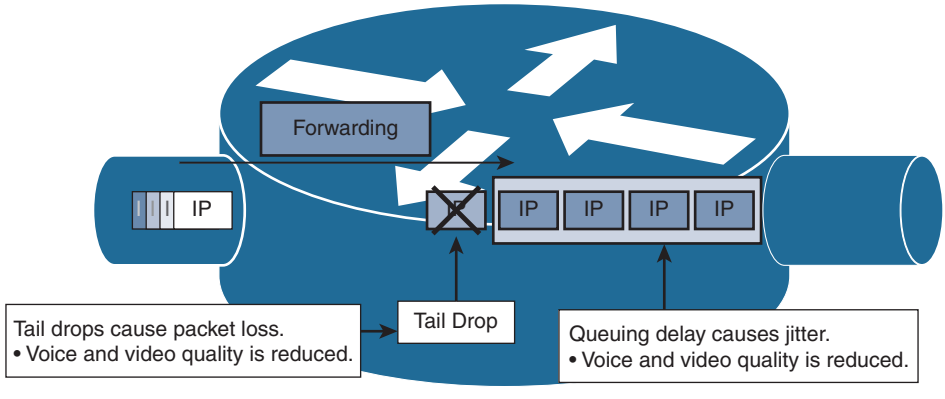

Figure 1-2 *Quality of Service Issues Example: Jitter and Packet Drop*

Bandwidth Challenges

Each site in a multisite deployment is usually interconnected by an IP WAN, or occasionally by a metropolitan-area network (MAN), such as Metro Ethernet. Within the past 10 to 15 years, various WAN technologies have emerged such as MPLS, SONET, Frame Relay, ATM, T1, and satellite, to name a few. Bandwidth on WAN links is limited and relatively expensive. The goal is to use the available bandwidth as efficiently as possible. Unnecessary traffic should be removed from the IP WAN links through content filtering, firewalls, and access control lists (ACLs). IP WAN acceleration methods for bandwidth optimization should be considered as well, such as Cisco Wide Area Application Services (WAAS), Cisco Intelligent WAN (IWAN) technologies, and perhaps caching technologies such as Akamai. Because available bandwidth on the WAN can become scarce, any period of congestion could result in service degradation unless QoS is deployed throughout the network. Figure 1-3 demonstrates the Cisco WAAS solution.

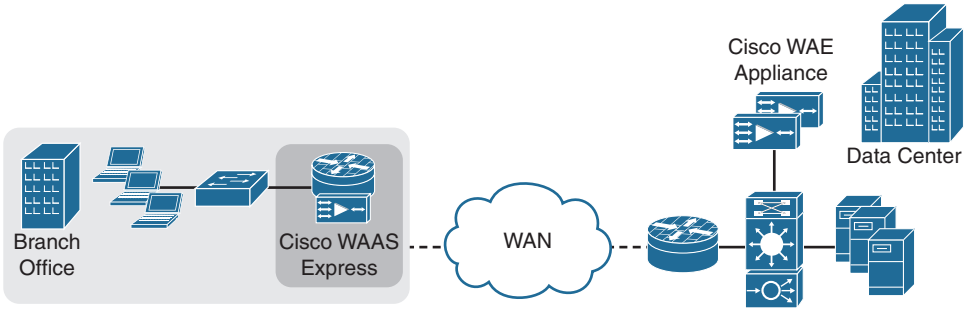

Figure 1-3 *Cisco WAAS Example*

Voice RTP streams produced by Cisco IP phones and video endpoints are a constant and predictable packet size. They are small in size but sent at a very high frequency rate (that is, a high number of small sized packets going across the wire or network link). In bandwidth-challenged locations or slow-speed WAN links, voice streams can be considered wasteful if the wrong voice codec is selected. G.711 uses a consistent 64 kbps for the payload size plus Layer 2 overhead. G.729, however, uses an 8-kbps payload size plus Layer 2 overhead. The Layer 2 overhead of packetization, the encapsulation of digitized voice into an RTP, UDP, IP, and Layer 2 header, is extremely high compared to the payload size. The more voice packets that are sent, the more headers are added to the RTP payload, and thus the more bandwidth required on the link.

The G.711 audio codec requires 64 kbps for the payload or RTP stream, whereas packetizing the G.711 voice sample in an IP/UDP/RTP header every 20 ms requires an additional 16 kbps for overhead. The overhead consists of a 12-byte RTP header, an 8-byte UDP header, and a 20-byte IP header. The header to payload ratio is 1:4; the bandwidth required is 80 kbps. This metric only considers the encapsulation to IP packets, the actual Layer 2 encapsulation (which varies based on WAN technologies)

is not considered. For example, the header size of a generic routing encapsulation (GRE) tunnel or IPsec virtual private network (VPN) across a Layer 2 transport is much higher.

The G.729 codec is used across the WAN in situations when bandwidth is a concern due to the smaller packet size. G.729 uses 8 kbps for the payload size and a sampling rate of every 20 ms yields 16 kbps plus Layer 2 overhead for the header; the bandwidth required is 24 kbps. In addition, G.729 has a 2:1 header to payload ratio as compared to G.711.

Note Sampling rate determines the bandwidth required per codec.

You may wonder how the 16-kbps value for the header bandwidth was calculated. The 40 bytes of header information must be converted to bits to figure out the packet rate of the overhead. Because a byte has 8 bits, 40 bytes * 8 bits in a byte = 320 bits. The 320 bits are sent 50 times per second based on the 20-ms rate (1 millisecond is 1/1000 of a second, and 20/1000 = .02). So:

.02 * 50 = 1 second

320 bits * 50 = 16,000 bits/sec, or 16 kbps

Note This calculation does not take Layer 2 encapsulation into consideration. For additional information, refer to *QoS Solution Reference Network Design (SRND)* (http://www.cisco.com/go/srnd) or *Cisco QoS Exam Certification Guide*, Second Edition (Cisco Press, 2004). For more information on QoS, go to http://www.cisco.com/go/qos.

Voice packets are benign compared to the bandwidth consumed by data applications. Data applications can fill the entire maximum transmission unit (MTU) of an Ethernet frame (1518 bytes or 9216 bytes if jumbo Ethernet frames have been enabled). In comparison to data application packets, voice packets are small (approximately 60 bytes for G.729 and 200 bytes for G.711 with the default 20-ms sampling rate).

Because of the inefficiency of voice packets, all unnecessary voice streams should be kept away from the IP WAN. A great example of this is media resources, in particular music on hold (MOH), conference bridges (CFB), and annunciators. Each of the types of resources requires additional bandwidth across the IP WAN. These media resources can be optimized in such a way that they do not have to traverse the IP WAN all the time, thereby saving bandwidth. You can achieve this optimization by placing local media resources at the remote sites where applicable.

In Figure 1-4, a conference bridge has been deployed at the main site. No conference bridge exists at the remote site. If three IP phones at a remote site join a conference, their RTP streams are sent across the WAN to the conference bridge. The conference bridge, whether using software or hardware resources, mixes the received audio streams and sends back three unique unicast audio streams to the IP phones over the IP WAN. The conference bridge removes the receiver's voice from his unique RTP stream so that the user does not experience echo because of the delay of traversing the WAN link and mixing RTP audio streams in the conference bridge.

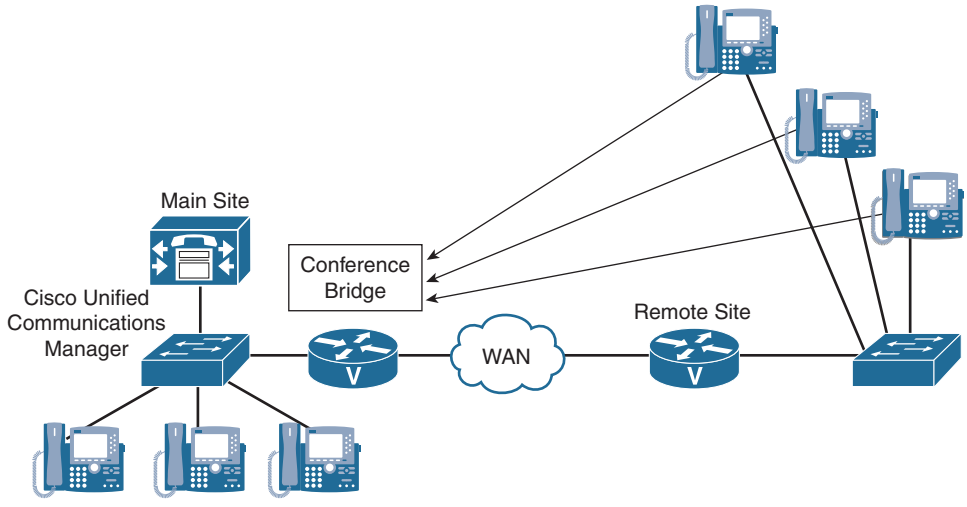

Figure 1-4 *Bandwidth Issue Example: Centralized Media Resources and Bandwidth*

Centralized conference resources cause bandwidth, delay, and capacity challenges in the voice network. Each G.711 RTP stream requires 80 kbps (plus the Layer 2 overhead), resulting in 240 kbps of IP WAN bandwidth consumption by this voice conference. If the conference bridge were not located on the other side of the IP WAN, this traffic would not need to traverse the WAN link, resulting in less delay and bandwidth consumption. If the remote site had a CUCM region configuration that resulted in calls with the G.729 codec back to the main site, the software conferencing resources of CUCM would not be able to mix the audio conversations. Software-based conferencing on CUCM can only handle the G.711 codec. Hardware conferencing or hardware transcoder media resources in a voice gateway are required to accommodate G.729 audio conferencing. Local hardware conference resources eliminate this need. All centrally located media resources (MOH, annunciator, conference bridges, videoconferencing, and media termination points) suffer similar bandwidth, delay, and resource-exhaustion challenges.

Cisco has a best practice architecture for media resources, which is available in the Cisco Validated Designs (CVDs) and Solutions Reference Network Designs (SRNDs). Chapter 6, "Cisco Collaboration Solution Bandwidth Management," is devoted to media resource and bandwidth management coverage. A general concept when planning media resources is *conference remotely* and *transcode centrally*. This is achieved by conferencing remotely using packet voice data modules (PVDMs), which are router/hardware-based conference resources at remote branches. Various Cisco applications such as Unity Connection and Unified Contact Center Express (UCCX) can only accept G.711 streams depending on how they are installed. These applications require a transcoding resource to convert a WAN G.729 codec into G.711. In a centralized call processing architecture, these applications are usually located in a headquarters or data center. You need PVDMs or digital signal processors (DSPs) located near these servers to

perform transcoding of calls coming from remote branches into these applications; the idea being to transcode centrally at main sites or data centers. In certain hybrid layouts, a mixture of local and remote transcoders and conference bridges are used to achieve the desired result.

Availability Challenges

When deploying CUCM in multisite environments, CUCM-based services are accessed over the IP WAN. Availability of the IP network, especially of the IP WAN that interconnects sites, is critical for several services and protocols. Protocols and services that are affected in the event of a WAN failure include the following:

- **Signaling in CUCM multisite deployments with centralized call processing:** Remote Cisco IP phones and video endpoints register with a centralized CUCM server. Remote MGCP gateways are controlled by a centralized CUCM server that acts as an MGCP call agent. VIC cards or high-density analog gateways provide POTS capabilities at remote branches and can register with a centralized CUCM server that acts as a SCCP call agent. SIP and H.323 protocols are peer-to-peer technologies and can survive in the event the WAN goes down provided proper dial peers and SRST configurations are in place.
- **Signaling in CUCM multisite deployments with distributed call processing:** In such environments, sites are connected via H.323 (non-gatekeeper-controlled, gatekeeper-controlled, or H.225), SIP trunks, or intercluster trunks (ICTs). In the event of a WAN failure, these connection types stop processing the signaling traffic between clusters.
- **Media exchange:** RTP streams sent between endpoints that are located at different sites rely on the IP WAN to be stable and available. In the event of an IP WAN failure, the audio paths for RTP stop functioning. This can be detrimental to any active calls across the WAN and to future calls placed between sites until functionality is restored.
- **Other services:** UC has a host of auxiliary services and protocols that all rely on the IP WAN. These include Cisco IP phone Extensible Markup Language (XML) services and access to applications such as attendant console, CUCM Assistant Cisco IP Manager Assistant (IPMA), VCS, or Expressway cluster signaling, media resources that register with CUCM using SCCP, centralized video conferencing using TelePresence Conductor and TelePresence Server, and centralized voicemail using Cisco Unity Connection. Scheduling of video resources such as meeting room reservations that rely on TelePresence Management Suite (TMS) to communicate with the endpoint and Microsoft Exchange are included in this category.

If the IP WAN connection is broken, these services are not accessible. The unavailability might be acceptable for some services, but strategic applications such as UC, voicemail,

video, and auxiliary services should be made available during WAN failure via backup mechanisms.

Figure 1-5 shows a UC network in which the main site is connected to a remote site via a centralized call-processing environment. The main site is also connected to a remote cluster through an ICT, representing a distributed call-processing environment. The combination of both centralized and distributed call processing represents a hybrid call-processing model in which small sites use the CUCM resources of the main site, but large remote offices have their own CUCM cluster. The bottom left of Figure 1-5 shows a SIP trunk terminated on a CUBE, which is typically implemented over a WAN connection such as MPLS to an ITSP. The benefit of the SIP trunk is that the ITSP provides the gateways to the public switched telephone network (PSTN) instead of you needing to provide gateways at the main site.

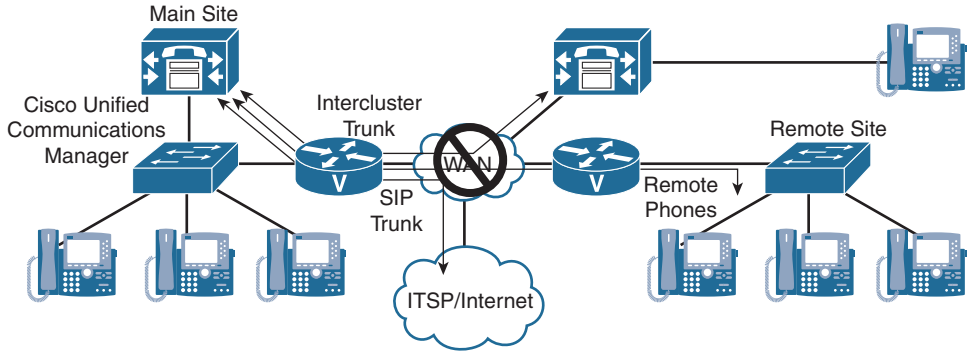

Figure 1-5 *Availability Issues Example: IP WAN Failure*

An IP WAN outage in Figure 1-5 will cause an outage of call-processing services for the remote site connected in a centralized fashion. The remote cluster will not suffer a call-processing outage, but the remote cluster will not be able to dial the main site over the IP WAN during the outage via the ICT. Mission-critical voice applications (voicemail, interactive voice response [IVR], and so on) located at the main site will be unavailable to any of the other sites during the WAN outage.

If the ITSP is using the same links that allow IP WAN connectivity, all calls to and from the PSTN are also unavailable.

Note A deployment like the one shown in Figure 1-5 is considered a bad design because of the lack of IP WAN fault tolerance and PSTN backup. A high availability (HA) design would include multiple redundant WAN links, HA routing protocols, multiple ICT trunks, and redundant CUBEs. The Cisco CVDs have detailed sections on providing fault tolerance and HA solutions in a UC environment.

Dial Plan Challenges

In a UC multisite deployment, with a single or multiple CUCM cluster, dial plan design requires the consideration of several issues that do not exist in single-site deployments, including the following:

- Overlapping numbers
- Nonconsecutive numbers
- Variable-length numbering
- Direct inward dialing (DID) ranges and E.164 addressing
- Optimized call routing
- Various PSTN requirements
- Scalability

Overlapping Numbers

Users located at different sites can have the same DNs assigned. An example of this is a user in Virginia with a DID of 804-424-1601; the UC administrator may configure a DN of 1601 on that user's phone. Another user in Colorado may have a DID of 303-860-1601; the UC administrator may configure a DN of 1601 on that user's phone as well. This can occur provided the two extensions are in different partitions inside the CUCM. Because DNs usually are unique only within a site, a multisite deployment requires a solution for overlapping numbers. In this example, how could the Virginia-based DN of 1601 dial a Denver-based DN of 1601? They are the same number in separate partitions. The solution is creative site codes or creative use of CSS design.

Note The solutions to the problems listed in this chapter are discussed in more detail in Chapter 2, "Understanding Multisite Deployment Solutions."

In Figure 1-6, Cisco IP phones at the main site use DNs 1001 to 1099, 2000 to 2157, and 2365 to 2999. At the remote site, 1001 to 1099 and 2158 to 2364 are used. These DNs have two issues. First, 1001 to 1099 overlap; these DNs exist at both sites, so they are not unique throughout the complete deployment. This causes a problem: If a user in the remote site were to dial only the four digits 1001, which phone would ring? This issue of overlapping dial plans needs to be addressed by digit manipulation. In addition, the nonconsecutive use of the range 2000 to 2999 (with some duplicate numbers at the two sites) requires a significant number of additional entries in call-routing tables because the ranges can hardly be summarized by one entry (or a few entries).

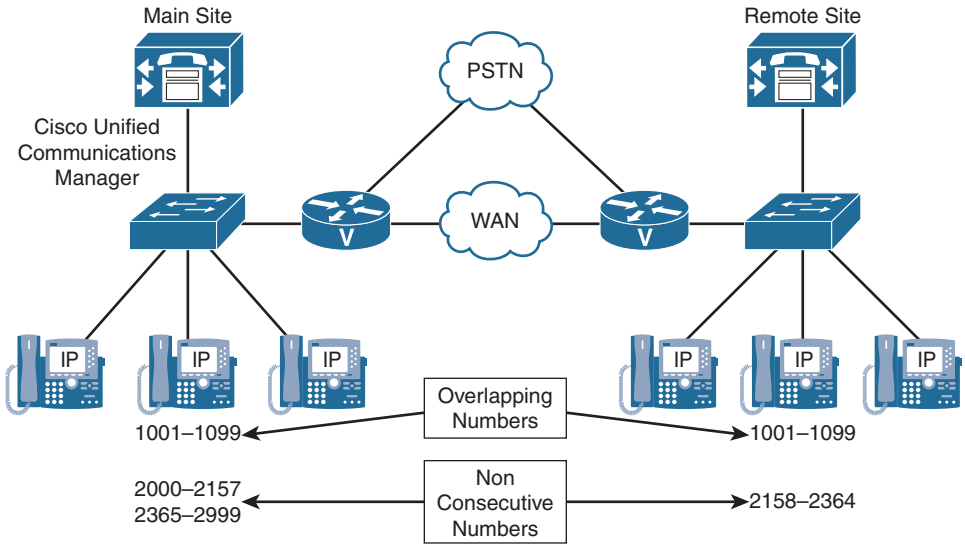

Figure 1-6 *Dial Plan Challenges: Overlapping and Nonconsecutive Numbers*

Nonconsecutive Numbers

Contiguous ranges of numbers are important to summarize call-routing information, analogous to contiguous IP address ranges for route summarization. For example, a remote branch may have a PSTN DID range of 757-466-1XXX, thus providing that branch with 1000 DNs from extension 1000 through 1999 (assuming a four-digit dial plan). In CUCM, you can summarize these patterns and do not need to enter all 1000 entries into the routing table/dial plan as simply 1XXX. Such blocks of extensions can be represented by one entry in the call-routing table, such as route patterns and translation patterns (in CUCM), dial peer destination patterns (in IOS), and voice translation rules (in IOS), which keep the routing table short and simple. If each endpoint requires its own entry in the call-routing table, the table gets too big, lots of memory is required, and lookups take more time. Therefore, nonconsecutive numbers at any site are not optimal for efficient call routing. A nonoptimal design is to skip ranges of numbers for this remote site. Imagine what the routing table would look like if it only had DN range 1000 to 1050 then 1190 to 1300 followed by 1550 to 1600. These "gaps" would require multiple routing entries in the CUCM database.

Variable-Length Numbering

Some countries, such as the United States and Canada, have fixed-length numbering plans for PSTN numbers. North America uses the North American Numbering Plan (NANP). This dictates that PSTN phone numbers are ten digits in the format XXX-XXX-XXXX, a three-digit number plan area code (NAA), followed by a three-digit exchange (NXX), followed by a four-digit subscriber code.

Others, such as Mexico and England, have variable-length numbering plans; their PSTN numbers vary in length. A problem with variable-length numbers is that the complete length of the number dialed can be determined in CUCM only by waiting for the interdigit timeout. Interdigit timeout refers to the time CUCM waits to determine you are done dialing a number. Waiting for the interdigit timeout, known as the T.302 timer, adds to the post-dial delay, which may annoy users. Further, the T.302 timer is a service parameter in CUCM that needs to be set on every server in the cluster; the default is 15 seconds, which for many people is far too long.

Throughout my consulting tenor, I have found that lowering the T.302 timer to around 7 to 8 seconds is the best solution for many organizations; lower and you may disconnect users mid-dial, any longer would annoy users who have completed dialing and are waiting on CUCM to connect the call. Ways to mitigate this issue are discussed later in this chapter. You can allow users to specify a terminating digit to represent they are done dialing a variable-length pattern.

Direct Inward Dialing (DID) Ranges and E.164 Addressing

When considering integration with the PSTN, internally used DNs have to be related to external PSTN numbers (public DIDs and E.164 addressing). In layman terms, it is how you coordinate the mapping of external DIDs to an internal DN scheme. A misconception among many junior UC engineers is that CUCM contains a screen or mechanism to track DID to DN mappings; this simply does not exist, and proper planning and a few Excel spreadsheets are often used. You can create translation patterns for every DID and translate them into DNs if you wanted to go to an extreme and where warranted, but that is adding complexity. Depending on the numbering plan (fixed or variable) and services provided by the PSTN, the following solutions are common:

- Each internal DN relates to a fixed-length PSTN number.
- Another solution is to not reuse any digits of the PSTN number, but to simply map each internally used DN to any PSTN number assigned to the company.

Each internal DN has its own dedicated PSTN number. The DN can, but does not have to, match the least-significant digits of the PSTN number. In countries with a fixed numbering plan, such as the NANP, this usually means that the four-digit office or subscriber codes are used as internal DNs. If these are not unique, office codes or administratively assigned site codes might be added to the number, resulting in five or more digits being used for internal DNs.

An example to provide clarity is a remote branch in California with a DID range of 415-586-7200 through 7299 may choose to assign internal DNs or DNs to phones using a four-digit extension from 7200 to 7299. Assume there is an additional remote branch in Chicago with DID range 312-733-7200 to 7299. You could create DNs on the phones in Chicago with extensions 7200 to 7299 and place the DNs in separate partitions (logically separating them in CUCM). How does one dial between sites now? One common technique is to use site codes; in the dial plan for San Francisco, users would append a site code to the four-digit extension if they were trying to reach Chicago

phones. For example, a user in San Francisco may dial 557200 to reach an extension of 7200 in Chicago. The site code would be 55, representing all phones in Chicago. CUCM can uniquely route the call to a site based on the site code and using a translation pattern or digit-stripping mechanisms in CUCM.

Another solution is to not reuse any digits of the PSTN number, but to simply map each internally used DN to any PSTN number assigned to the company. In this case, the internal and external numbers do not have anything in common. If the internally used DN matches the least-significant digits of its corresponding PSTN number, significant digits can be set at the gateway or trunk. Also, general external phone number masks, transformation masks, or prefixes can be configured. This is true because all internal DNs are changed to fully qualified PSTN numbers in the same way.

An example of this technique is a UC dial plan in which sites have contiguous blocks of DNs, a site in New York may receive extensions 1000–1999, a site in San Diego may receive blocks 2000–2999, and so on. The internal numbering scheme has nothing to do with the DID ranges from the PSTN. New York DIDs may be 212-618-6750 through 212-618-6799. To map the public DID to an internal DN, you need to invoke digit manipulation in the form of translation patterns, transforms, significant digits, or other techniques in CUCM to mask and change the DID to fit the internal scheme. This approach can be laborious because a one-for-one translation is required.

What if a remote site has no DIDs in fixed-length numbering plans? To avoid the requirement of having one DID number per internal DN when using a fixed-length numbering plan, it is common in some organizations to disallow DIDs to internal extensions. Instead, the PSTN trunk has a single number, and all PSTN calls routed to that number are sent to an attendant, an auto-attendant, a receptionist, or a secretary. From there, the calls are *transferred* to the appropriate internal extension.

Internal DNs are part of a variable-length number. In countries with variable-length numbering plans, a typically shorter "subscriber" number is assigned to the PSTN trunk, but the PSTN routes all calls *starting* with this number to the trunk. The caller can add digits to identify the extension. There is no fixed number of additional digits or total digits. However, there is a maximum, usually 32 digits, which provides the freedom to select the length of DNs. This maximum length can be less.

For example, in E.164 the maximum number is 15 digits, not including the country code. A caller simply adds the appropriate extension to the company's (short) PSTN number when placing a call to a specific user. If only the short PSTN number without an extension is dialed, the call is routed to an attendant within the company. Residential PSTN numbers are usually longer and do not allow additional digits to be added; the feature just described is available only on trunks.

Optimized Call Routing

Having an IP WAN between sites with local PSTN access at all sites allows for PSTN toll bypass by sending calls between sites over the IP WAN instead of using the PSTN. In such scenarios, the PSTN should be used as a backup path only in case of WAN failure.

Another solution, which extends the idea of toll bypass and can potentially reduce toll charges, is to also use the IP WAN for PSTN calls. With tail-end hop-off (TEHO), the IP WAN is used as much as possible, and the gateway that is closest to the dialed PSTN destination is used for the PSTN breakout. An example is a New York-based IP phone dialing a San Diego PSTN number. Provided the enterprise has an IP WAN between the sites and a local gateway in San Diego with a PSTN circuit (POTS, PRI, or SIP trunk), you could in fact route that call across the WAN to go out the San Diego gateway as a local call, thus bypassing costly long-distance charges if the call were sent out the New York gateways as a long-distance call. In certain areas of the world, TEHO or toll bypass is illegal, because the telephone companies are often government regulated. Check the telephone laws in your specific country or locality to determine any legal issues that could arise from this technique.

Various PSTN Requirements

Various countries and sometimes even several PSTN providers within the same country can have numerous requirements regarding the PSTN dial rules. This situation can cause issues when calls can be routed via multiple gateways. If the requirements of a primary gateway are different from the requirements of a backup gateway, numbers must be transformed accordingly. In the United States, smaller cities and localities often will allow the use of a seven-digit local dialing plan, meaning you can dial seven digits for local PSTN calls within that area. In larger metropolitan areas, this is often expanded and mandated that ten-digit dialing is used to represent a local call. Imagine if you have multiple paths for a call to go out with redundant gateways, one being in a small locality and the backup gateway being in a major city. Digit manipulation would be required to expand a local call from seven to ten digits if it routed out the backup gateway in a site where ten-digit dialing is mandated.

Additional PSTN considerations surrounding automatic number identification (ANI) also need to be addressed. The ANI of calls that are being received from the PSTN can be represented in various ways: as a seven-digit subscriber number, as a ten-digit number including the area code, or in international format with the country code in front of the area code. To standardize the calling number for all calls (which can be displayed in the call logs or display screens on IP phones and video endpoints), the format that is used must be known, and the number must be transformed accordingly. In countries where PSTN numbers do not have fixed lengths, it is impossible to detect the type of number (local, national, or international) by looking at only the length of the number. In such cases, the type of number must be specified in signaling messages (for example, by the ISDN type of number [TON]). *Type of number* is an ISDN term in which the calling number or ANI of the call contains additional information elements (IE) describing the number as a local, national, or international call. Gateways can read this information and append or strip the appropriate number of digits and pass a uniform ANI length to IP phones and video endpoints.

Note In the United States, TON is mainly used on PRIs; SIP trunks may not support TON depending on the carrier.

Scalability

In large or very large deployments, dial plan scalability issues arise. When interconnecting multiple CUCM clusters or CUCM Express routers via trunks, it is difficult to implement a dial plan on an any-to-any basis where each device or cluster needs to know the numbers or prefixes that are found at every other system. In addition to the need to enter almost the same dial plan at each system, a static configuration does not reflect true reachability. If there are any changes, the dial plans at each system must be updated. Although there are solutions that allow centralized dial plan configuration (for example, Cisco Session Management Edition [SME] or H.323 gatekeepers), in very large deployments a dynamic discovery of DN ranges and prefixes simplifies the implementation and provides a more scalable solution.

Fixed Versus Variable-Length Numbering Plans

A fixed numbering plan features fixed-length area codes and local numbers. The United States utilizes fixed-length dial plans. An open numbering plan or variable-length numbering plan features variance in length of area code or local number, or both, within the country. Figure 1-7 illustrates an international deployment with various numbering schemes in place including fixed and variable length. Take a moment to familiarize yourself with the dialing habits in both schemes.

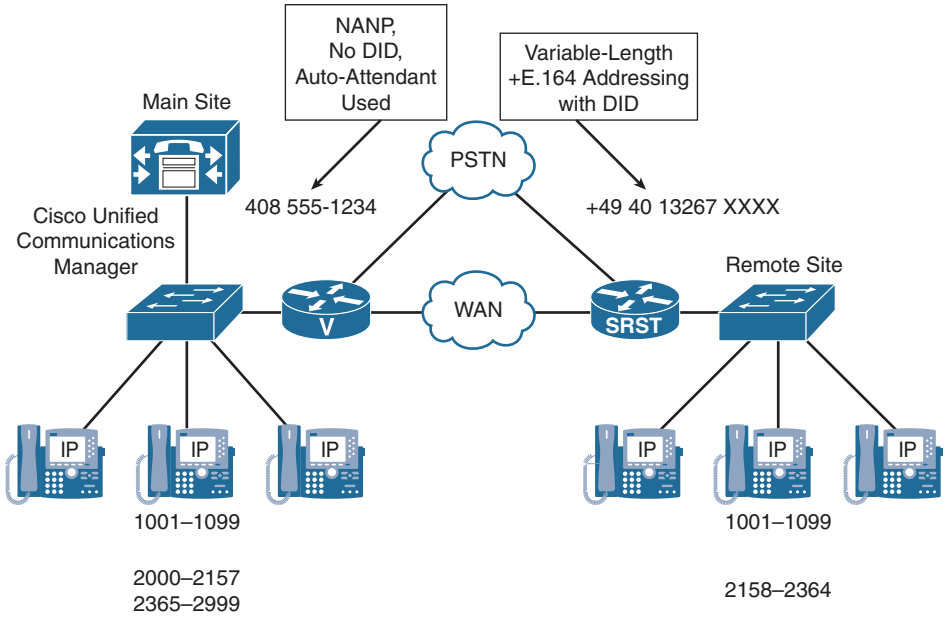

Figure 1-7 *Variable-Length Numbering, +E.164 Addressing, and DID Example*

Figure 1-7 features a main site in the United States. The NANP PSTN number is 408 555-1234. DIDs are not used. All calls placed to the main site are managed by an attendant. There is a remote site in Germany with a PSTN number of +49 404 13267.

The German location uses four-digit extensions, and DID is allowed, since digits can be added to the PSTN number. When calling the German office attendant (not knowing a specific extension), users in the United States dial +9 011 49 404 13267. Note that the + is replaced by the international prefix 011 and the access code 9. If they know that they want to contact extension 1001 directly, they dial +9 011 49 404 13267 1001.

Table 1-1 contrasts the NANP and a variable-length numbering plan (Germany's numbering plan).

Table 1-1 *Fixed- Versus Variable-Length Numbering Plans*

Component	Description	Fixed-Length Numbering Plan (NANP)	Variable-Length Numbering Plan (Germany)
Country code	A code of one to 3 digits is used to reach the particular telephone system for each nation or special service. Obtain the E.164 standard from http://itu.org to see all international country codes.	1	49
Area code	Used within many nations to route calls to a particular city, region, or special service. Depending on the nation or region, it may also be called a numbering plan area, subscriber trunk dialing code, national destination code, or routing code.	3 digits	3 to 5 digits
Subscriber number	Represents the specific telephone number to be dialed, but it does not include the country code, area code (if applicable), international prefix, or trunk prefix.	3-digit exchange code plus a 4-digit station code	3 or more digits
Trunk prefix	The initial digits to be dialed in a domestic call, before the area code and the subscriber number.	1	0
Access code	A number that is traditionally dialed first "to get out to the PSTN," used in private branch exchanges (PBXs) and VoIP systems.	9	0

Table 1-1 *continued*

Component	Description	Fixed-Length Numbering Plan (NANP)	Variable-Length Numbering Plan (Germany)
International prefix	The code dialed before an international number (country code, area code if any, and then subscriber number).	011	00 or + (+ is used by cell phones)

An area code is used within many countries to route calls to a particular city, region, or special service. Depending on the country or region, it may also be referred to as one of the following:

- Numbering plan area (NPA)
- Subscriber destination code
- National destination code
- International Country code

The subscriber number represents the specific telephone number to be dialed, but does not include the country code, area code (if applicable), international prefix, or trunk prefix.

A trunk prefix refers to the initial digits that are dialed in a call within the United States, preceding the area code and the subscriber number.

An international prefix is the code that is dialed before an international number (the country code, the area code if any, and then the subscriber number).

The table contrasts the NANP and a variable-length numbering plan (the German numbering plan, in this example).

Some examples include the following:

- **Within the U.S.:** 9-1-408-555-1234 or 408-555-1234 (within the same area code)
- **U.S. to Germany:** 9-011-49-404-132670
- **Within Germany:** 0-0-404-132670 or 0-132670 (within the same city code)
- **Germany to the U.S.:** 0-00-1-408-555-1234 (Note: The 1 in 00-1-408 is the U.S. country code, not the trunk prefix.)

Note In the examples shown following Table 1-1, dialing out from the United States illustrates the common practice of dialing 9 first as an access code to dial out. This use is common, but optional, in a dial plan. However, if the access code is used, the 9 must be stripped before reaching the PSTN, whereas the other dialed prefixes must be sent to the PSTN for proper call routing.

It is worth noting that the logic of routing calls by CUCM over the WAN or through the PSTN is appropriately transparent to the phone user. The phone user has no idea if the call is being routed via a PSTN circuit or the IP WAN. Appropriate digit manipulation needs to occur in either scenario.

Detection of End of Dialing in Variable-Length Numbering Plans

There are three ways to detect end of dialing in variable-length numbering plans:

- Interdigit timeout
  - Simple to configure
  - Least convenient
- Use of # key
  - Different implementation in Cisco IOS Software (simple) versus CUCM (complex)
  - Convenient
  - Requires users to be aware of this option
- Use of overlap sending and receiving
  - Convenient
  - Must be supported by PSTN
  - Complex implementation

As the preceding list shows, one issue that can arise is how to detect that a user is done dialing a number when using a variable-length dial plan. You must ensure that you give ample time for IP phone users to complete their call without disconnecting them prematurely. Also confirm that there is a mechanism in place to allow the users to signify they are done dialing and that CUCM should start routing the call immediately.

From an implementation perspective, the simplest way to detect end of dialing is to wait for an interdigit timeout to expire. This approach, however, provides the least comfort to the end user because it adds post-dial delay. In an environment with only a few numbers of variable length (for example, NANP, where only international calls are of variable length), waiting for the interdigit timeout might be acceptable. However, even in such an environment, it might make sense to at least reduce the value of the timer, because the default value in CUCM is high (15 seconds).

Note In CUCM, the interdigit timer is set by the cluster-wide Cisco Call Manager service parameter T302 timer that is found in CUCM Administration by navigating to **System > Service Parameters** under the Cisco Call Manager Service.

In Cisco IOS Software, the default for the interdigit timeout is 10 seconds. You can modify this value using the voice-port **timeouts interdigit** command.

Another solution for detecting end of dialing on variable-length numbers is the use of the # key. An end user can press the # key to indicate that dialing has finished. The implementation of the # key is different in CUCM versus Cisco IOS Software. In Cisco IOS gateways, the # is seen as an instruction to stop digit collection. It is not seen as part of the dialed string. Therefore, the # is not part of the configured destination pattern. In CUCM, the # is considered to be part of the dialed number and, therefore, its usage has to be explicitly permitted by the administrator by creating patterns that include the #. If a pattern includes the #, the # has to be used; if a pattern does not include the #, the pattern is not matched if the user presses the # key. Therefore, it is common in CUCM to create a variable-length pattern twice: once with the # at the end, and once without the #. An example is the following two route patterns inside CUCM: 9.011! and 9.011!#. Note that the "9." would be discarded with the strip PreDot discard digit command at the route pattern level. You can additionally specify the discard PreDot and trailing pound instruction for the 9.011!# route pattern; thereby allowing both patterns inside CUCM—one for variable length and the other including a terminating digit of pound.

An alternative way to configure such patterns is to end the pattern with ![0–9#]. In this case, a single pattern supports both ways of dialing (with and without the #). However, be aware that the use of such patterns can introduce other issues. For example, this can be a concern when using discard digits instructions that include trailing-# (for example, PreDot-trailing-#). This discard digit instruction will have an effect only when there is a trailing # in the dialed number. If the # is not used, the discard digit instruction is ignored. Therefore, the PreDot component of the discard digit instruction is also not performed. PreDot is a form of digit manipulation in CUCM that strips off all digits before the dot.

Allowing the use of the # to indicate end of dialing provides more comfort to end users than having them wait for the interdigit timeout. However, this possibility has to be communicated to the end users, and it should be consistently implemented. As previously mentioned, it is automatically permitted in Cisco IOS Software, but not in CUCM.

The third way to indicate end of dialing is the use of overlap send and overlap receive. If overlap is supported end to end, the digits that are dialed by the end user are sent one by one over the signaling path. Then, the receiving end system can inform the calling device after it receives enough digits to route the call (number complete). Overlap send and receive is common in some European countries, such as Germany and Austria. From a dial plan implementation perspective, overlap send and receive is difficult to implement when different PSTN calling privileges are desired. In this case, you have to collect enough digits locally (for example, in CUCM or Cisco IOS Software) to be able to decide to permit or deny the call. Only then can you start passing digits on to the PSTN one by one using overlap. For the end user, however, overlap send and receive is comfortable because each call is processed as soon as enough digits have been dialed. The number of digits that are sufficient varies per dialed PSTN number. For example, one local PSTN destination might be reachable by a seven-digit number, whereas another local number might be uniquely identified only after receiving nine digits.

Optimized Call Routing and PSTN Backup

Using an IP WAN enables savings on the cost of long-distance or international PSTN calls in a multisite environment. There are two ways to save costs on long-distance or international PSTN calls in a multisite deployment:

- **Toll bypass:** Calls between sites that use the IP WAN instead of the PSTN are toll-bypass calls. The PSTN is used only when calls over the IP WAN are not possible (either because of a WAN failure or because the call is not admitted by CAC). An example is dialing between IP phones at two sites; the call traverses the IP WAN for RTP and signaling versus going across the PSTN.
- **TEHO:** TEHO extends the concept of toll bypass by also using the IP WAN for calls to the remote destinations in the PSTN. With TEHO, the IP WAN is used as much as possible and PSTN breakout occurs at the gateway that is located closest to the dialed PSTN destination. Local PSTN breakout is used as a backup in case of an IP WAN or CAC failure. An example is dialing a San Diego number from New York. Provided you have an IP WAN connecting both sites, the call can flow across the IP WAN and exit a voice gateway or CUBE in San Diego as a local call, thereby saving costly PSTN charges.

Note Some countries do not allow the use of TEHO. When implementing TEHO, ensure that the deployment complies with legal requirements.

When using the IP WAN to reach remote PSTN destinations or internal DNs at a different site, it is important to consider backup paths. When the IP WAN is down or when not enough bandwidth is available for an additional voice call, calls should be routed via the local PSTN gateways as a backup path.

In the example shown in Figure 1-8, a call from Chicago to San Jose would be routed as shown in the following steps:

Step 1. A Chicago user dials 9 1 408 555-6666, the number for a PSTN phone that is located in San Jose.

Step 2. The call is routed from the CUCM Express in Chicago to the CUCM cluster in San Jose over the IP WAN.

Step 3. The CUCM in San Jose routes the call to the San Jose gateway, which breaks out to the PSTN with a (now) local call to the San Jose PSTN.

Step 4. The San Jose PSTN phone rings.

If the WAN were unavailable for any reason before the call, the Chicago gateway would have to be properly configured to route the call with the appropriate digit manipulation through the PSTN at a potentially higher toll cost to the San Jose PSTN phone.

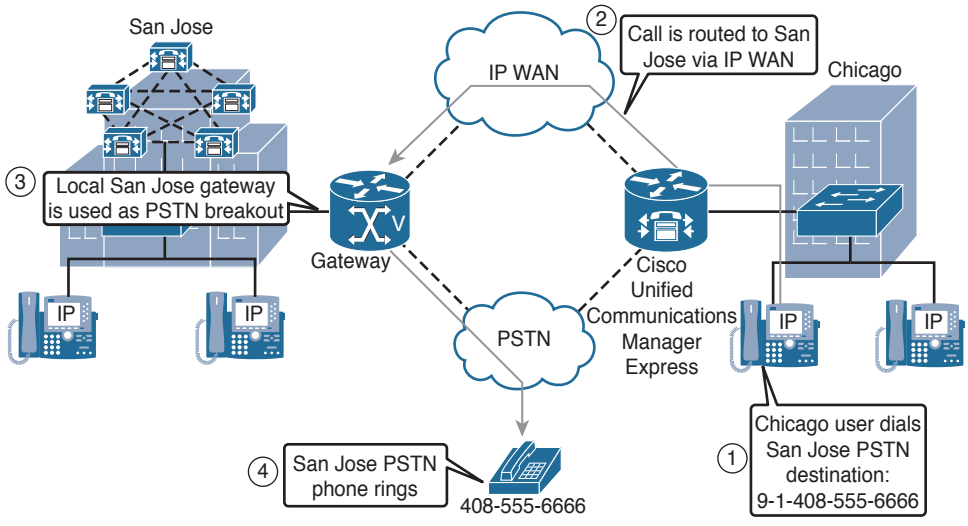

Figure 1-8 *TEHO Example*

Note The primary purpose of implementing TEHO is because of a reduction of operating costs from calling through the PSTN. In the TEHO example shown in Figure 1-8, there would be potential cost savings. However, costs savings are typically considerably higher when the remote location and destination call are international.

PSTN Requirements

Various countries can have different PSTN dialing requirements, which makes it difficult to implement dial plans in international multisite deployments.

There are several design challenges regarding PSTN access, including the following:

- Dial rules for the called-party number on outbound PSTN calls
  - Dial PSTN access code
  - Dial national access code
  - Dial international access code
- Dial presentation of called- and calling-party numbers on inbound and outbound calls
  - Dial length of number and its components
  - Dial ISDN number types

- Dial overlap send and overlap receive
- Dial + prefix on E.164 numbers

- Dial emergency dialing

One of the issues in international deployments is various PSTN dial rules. For example, in the United States, the PSTN access code is 9, whereas in most countries in Europe 0 is used as the PSTN access code. The national access code in the United States is 1, whereas 0 is commonly used in Europe. The international access code is 011 in the United States, and 00 is used in many European countries. Some PSTN provider networks require the use of the ISDN TON, but others do not support it. Some networks allow national or international access codes to be combined with ISDN TON. Others require you to send the actual number only (that is, without any access codes) when setting the ISDN TON.

The same principle applies to the calling-party number. As mentioned earlier, in variable-length numbering plans, the TON cannot be detected by its length. Therefore, the only way to determine whether the received call is a local, national, or international call is by relying on the availability of the TON information in the received signaling message.

Some countries that have variable-length numbering plans use overlap send and overlap receive. With overlap send, a number that is dialed by an end user is passed on to the PSTN digit by digit. Then, the PSTN indicates when it has received enough numbers to route the call. Overlap receive describes the same concept in the opposite direction: When a call is received from the PSTN in overlap mode, the dialed number is delivered digit by digit, and not en bloc. Some providers that use overlap send toward their customers do not send the prefix that is configured for the customer trunk, but only the additional digits that are dialed by the user who initiates the call.

When dialing PSTN numbers in E.164 format (that is, numbers that start with the country code), the + sign is commonly prefixed to indicate that the number is in E.164 format. The advantage of using the + sign as a prefix for international numbers is that it is commonly accepted as a symbol for internationally formatted telephone numbers around the world. In contrast, PSTN access codes such as 011 (used in the NANP) or 00 (often used in Europe) are known only in the respective countries.

Finally, emergency dialing can be an issue in international deployments. Because various countries have several emergency numbers and numerous ways to place emergency calls, users are not sure how to dial the emergency number when roaming to other countries. An international deployment should allow roaming users to employ their home dialing rules when placing emergency calls. The system should then modify the called number as required at the respective site.

Issues Caused by Different Methods of PSTN Dialing

Different local PSTN dial rules can cause several issues, especially in international deployments. Imagine an executive flies from New York to Paris. Typically, the executive has not been properly trained on how to dial a local call according to the Paris

PSTN dialing requirements. Vice versa, a European executive travels to the United States. He or she may not know the proper format of calls in the NANP dialing plan, which is fixed length.

The following list outlines how to store PSTN contacts so that they can be used from any site:

- Different ways to store or configure PSTN destinations:
  - Dial speed dials
  - Dial fast dials
  - Dial address book entries
  - Dial call lists
  - Dial AAR targets
  - Dial call-forward destinations
- Stored number can be used at multiple sites (countries) because of roaming users using local PSTN gateways:
  - Dial Cisco Extension Mobility
  - Dial Cisco Device Mobility
  - Dial PSTN backup
  - Dial TEHO and LCR

The main problem that needs to be solved in international environments is how to store contacts' telephone numbers. Address book entries, speed and fast dials, call list entries, redial capability, and other numbers should be in a format that allows them to be used at any site, regardless of the local dial rules that apply to the site where the user is currently located. Call-forwarding destinations should also be in a universal format that allows the configured number to be used at any site.

The main reason for a universal format is that a multisite deployment has several features that make it difficult to predict which gateway will be used for the call. For example, a roaming user might use Cisco Extension Mobility or Device Mobility. Both features allow an end user to use local PSTN gateways while roaming. If no universal format is used to store speed dials or address book entries, it will be difficult for the end user to place a PSTN call to a number that was stored according to the NANP dial rules while in countries that require different dial rules. Even when not roaming, the end user can use TEHO or least cost routing (LCR), so that calls break out to the PSTN at a remote gateway, not at the local gateway. If the IP WAN link to the remote gateway is down, the local gateway is usually used as a backup. How should the number that is used for call routing look in such an environment? It is clearly entered according to local dial rules by the end user, but ideally it is changed to a universal format before call routing is performed. After the call is routed and the egress gateway is selected, the number could then be changed as required by the egress gateway.

Dial Plan Scalability Issues

In large CUCM deployments, it can be difficult to implement dial plans, especially when using features such as TEHO with local PSTN backup. Dial plans are difficult to implement in large Cisco UC deployments, and the following list outlines several scalability issues to take into consideration:

- Dial static configuration for multiple sites or domains is very complex because of any-to-any call-routing requirements.
- Dial centralized H.323 gatekeepers or SIP network services offer dial plan simplification.
  - Dial less configuration because of any-to-one call-routing topology
  - Dial static configuration nevertheless (no dynamic recognition of routes, no automatic PSTN rerouting)
  - Dial no built-in redundancy
- Dial while, an optimal solution, is desirable for large deployments. Services such as Global Dial Plan Replication (GDPR) and Service Advertisement Framework (SAD) and Call-Control Discovery (CCD) allow dynamic learning of dial plans in large networks. These concepts and their implementation are discussed in later chapters.

In large CUCM deployments, it can be difficult to implement scalable and easy-to-use dial plans, especially when using features such as TEHO with local PSTN backup or globalized +E.164 dial plans.

The main scalability issue of large deployments is that each call-routing domain (for example, a CUCM cluster or a CUCM Express router) must be aware of how to get to all other domains. Suppose that you have three CUCM clusters spread across the globe handling global communications; you have users traveling between sites who would like to retain their local dialing habits, features, and functionality.

Such a dial plan can become very large and complex, especially when multiple paths (for example, a backup path for TEHO) must be made available. Because each call routing domain must be aware of the complete dial plan, a static configuration does not scale. For example, any changes in the dial plan must be applied individually at each call-routing domain.

Centralized H.323 gatekeepers or SIP network services can be used to simplify the implementation of such dial plans, because there is no need to implement the complete dial plan at each call-routing domain. Instead of an any-to-any dial plan configuration, only the centralized component must be aware of where to find each number. This approach, however, means that you rely on a centralized service. If the individual call-routing entities have no connectivity to the centralized call-routing intelligence, all calls will fail. Further, the configuration is still static. Any changes at one call-routing domain (for example, new PSTN prefixes because of changing the PSTN provider) must also be implemented at the central call-routing component.

In addition, these centralized call-routing services do not have built-in redundancy. Redundancy can be provided, but requires additional hardware, additional configuration, and so on. Redundancy is not an integrated part of the solution.

The ideal solution for a large deployment is to allow an automatic recognition of routes. Internal as well as external (for PSTN backup) numbers should be advertised and learned by call-routing entities. A dynamic routing protocol for call-routing targets addresses scalability issues in large deployments. A new technique and technology has emerged with the advent of CUCM Version 10.x called Global Dial Plan Replication (GDPR). GDPR is a feature that is based on the concepts in previous CUCM releases. In Version 9.x, Cisco introduced the Intercluster Lookup Service (ILS) and Call Control Discovery (CCD). Think of ILS and CCD as a mechanism in which one CUCM cluster can advertise its DNs to another CUCM cluster via the IP network. It does this by broadcasting its dial plan using Service Advertisement Framework (SAF). GDPR, ILS, CCD, and Cisco SAF are explained in more detail in Chapter 16, "Cisco Service Advertisement Framework (SAF) and Call Control Discovery (CCD)."

NAT and Security Issues

In single-site deployments, CUCM servers and IP phones usually use private IP addresses because there is no need to communicate with the outside IP world. NAT is not configured for the phone subnets, and attacks from the outside are impossible as they are behind the corporate firewall. In modern multisite environments, this requires a paradigm shift as users have multiple devices in multiple sites all requiring communication paths that may transmit across the public Internet. A great example of this is instant messaging utilizing Cisco Jabber and the IM and Presence Server. How do you allow IMs on devices that roam outside of the LAN? Another example is video devices such as a Cisco DX80 endpoint on an executive's desk in his home office. How would you get that voice and video traffic back into the LAN and through a firewall? These newer technologies raise interesting security and protocol issues that are outlined in the list that follows:

- In single-site deployments, CUCM and IP phones do not require access to public IP networks:
  - NAT is not required.
  - Not reachable from the outside.
  - Not subject to attacks from outside (except from ITSP environment).
- In multisite deployments, private links or VPN tunnels can be used:
  - Requires gateway configuration at each site
  - Allows only intersite communication
  - Blocks access to and from outside (unless traffic is tunneled)

- Access to public IP networks is required in some situations
  - Connections to ITSPs or destinations on the Internet.
  - NAT required; CUCM and IP phones are exposed to the outside.
  - CUCM and IP phones are subject to attacks.

As the preceding list shows, if you focus on multisite deployments, IPsec VPN tunnels can be used between sites. The VPN tunnels allow only intersite communication; access to the protected internal networks is not possible from the outside (only from the other site through the tunnel). Therefore, attacks from the outside are blocked at the gateway. To configure IPsec VPNs, the VPN tunnel must be configured to terminate on the two gateways in the different sites. Sometimes this is not possible. For instance, the two sites may be under different administration, or perhaps security policies do not allow the configuration of IPsec VPNs.

In these cases, or when connecting to a public service such as an ITSP, you must configure NAT for CUCM servers and IP phones. When CUCM servers and IP phones are reachable with public IP addresses, they are subject to attacks from the outside world, which introduces potential security issues.

In such a case, or when connecting to a public service such as an ITSP, NAT has to be configured for CUCM servers and IP phones. Cisco UC accomplishes this NAT traversal by utilizing a third-party session border controller or a Cisco CUBE solution.

In Figure 1-9, Company A and Company B both use IP network 10.0.0.0/8 internally. For the companies to communicate over the Internet, the private addresses are translated to public IP addresses. Company A uses public IP network A, and Company B uses public IP network B. All CUCM servers and IP phones are reachable from the Internet and communicate with each other.

As soon as CUCM servers and IP phones can be reached with public IP addresses, they are subject to attacks from the outside world, introducing potential security issues.

Recently released UC technologies from Cisco have attempted to address security concerns with NAT and UC. Cisco has CUBEs, which are specialized Integrated Services Routers (ISRs) or Aggregation Services Routers (ASRs) that terminate ITSP circuits across Multiprotocol Label Switching (MPLS) or the open Internet. These devices provide a demarcation point and firewall features such as NAT and ACLs to limit which public entities are allowed to communicate through these devices. In the United States since the mid 2000s, the use of session border controllers and SIP technologies is outpacing that of traditional PSTN technologies such as PRIs or POTS telephone lines. Recent trends have shown that by 2020 SIP trunks will surpass the number of PRIs in the United States. Many telcos are moving their backbones to an all IP-based network; this transition only makes sense for the customer-facing offerings. SIP trunks are an all IP solution with many benefits, which are discussed throughout this book.

Company A Private IP	Company A Public IP	Company B Public IP	Company B Private IP
10.0.0.0/8	Public IP A	Public IP B	10.0.0.0/8

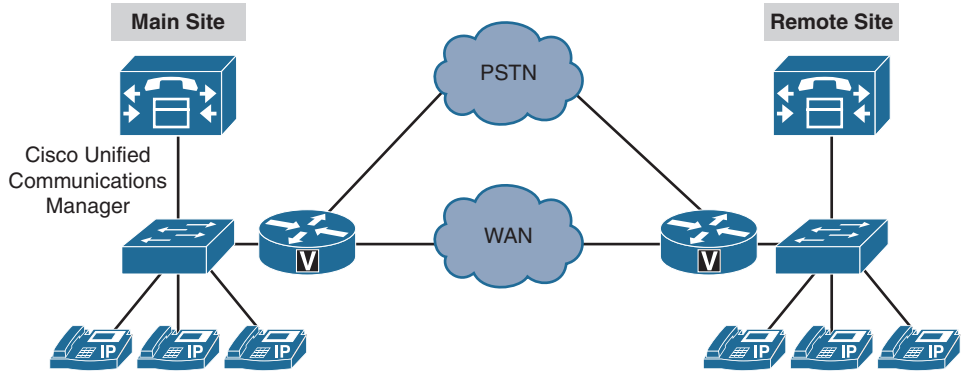

Figure 1-9 *Network Address Translation Security Issues for CUCM and IP Phones*

In addition, Cisco has launched an Expressway and VCS platform line that is capable of NAT traversal for video and IM devices. The technology uses a pair of servers (one on the LAN and another in the demilitarized zone [DMZ], which is a special network in the firewall for communications with external devices). These devices utilize a traversal zone or technology that establishes a communication path from the outside work into the DMZ, and a trust relationship is set up from the DMZ to the internal network. This allows for Cisco Jabber and video endpoints to register and communicate with internal devices without the use of a VPN client or VPN technology. Additional information about Cisco VCS and Expressway is covered in later chapters of this book.

Summary

Implementing a large communications network has never been a trivial task. Yet, with the right level of planning, leveraging expertise, and best practices, it is possible. Sticking to a flexible yet robust infrastructure and dial plan is the key to achieve enterprise-grade communications. The following key points were discussed in this chapter:

- Multisite deployment introduces issues and complexity, including call and video quality, bandwidth concerns, high availability, dial plan design, and NAT security.
- During congestion, voice and video packets have to be buffered or queued; otherwise, they may get dropped.
- Bandwidth in the IP WAN is limited and should be used as efficiently as possible.

- A multisite deployment has several protocols and services that depend on the availability of the IP WAN.
- A multisite dial plan has to address overlapping and nonconsecutive numbers, variable-length numbering plans, DID ranges, and ISDN TON and should minimize PSTN costs.
- When CUCM servers and IP phones need to be exposed to the outside, they can be subject to attacks from the Internet.

References

For additional information, refer to the following:

Cisco Systems, Inc. Cisco Collaboration Systems 10.x Solution Reference Network Designs (SRND), May 2014. http://www.cisco.com/c/en/us/td/docs/voice_ip_comm/cucm/srnd/collab10/collab10.html.

Review Questions

Use these questions to review what you have learned in this chapter. The answers appear in Appendix A, "Answers Appendix."

1. **Which of the following best describes DID?**
   a. E.164 international dialing
   b. External dialing from an IP phone to the PSTN
   c. VoIP security for phone dialing
   d. The ability of an outside user to directly dial into an internal phone

2. **Which of the following statements is the least accurate about IP networks?**
   a. IP packets can be delivered in the incorrect order.
   b. Buffering results in variable delays.
   c. Tail drops result in constant delays.
   d. Bandwidth is shared by multiple streams.

3. **Which statement most accurately describes overhead for packetized voice?**
   a. VoIP packets are large compared to data packets and are sent at a high rate.
   b. The Layer 3 overhead of a voice packet is insignificant and can be ignored in payload calculations.

c. Voice packets have a small payload size relative to the packet headers and are sent at high packet rates.

d. Packetized voice has the same overhead as circuit-switching voice technologies.

4. **What does the + symbol refer to in E.164?**

   a. Country code

   b. Area code

   c. International access code

   d. User's phone number

5. **Which two of the following are dial plan issues requiring a CUCM solution in multisite distributed deployments?**

   a. Overlapping directory numbers

   b. Overlapping E.164 numbers

   c. Variable-length addressing

   d. Centralized call processing

   e. Centralized phone configuration

6. **What is a requirement for performing NAT for Cisco IP phones between different sites through the Internet?**

   a. Use DHCP instead of fixed IP addresses

   b. Exchange RTP media streams with the outside world

   c. Use DNS instead of hostnames in CUCM

   d. Exchange signaling information with the outside world

7. **Which is the most accurate description of E.164?**

   a. An international standard for phone numbers including country codes and area codes

   b. An international standard for local phone numbers

   c. An international standard for dialing only local numbers to the PSTN

   d. An international standard for phone numbers for DID

8. **Which of the following is the most accurate description of TEHO?**

   a. Using the PSTN for cost reduction

   b. Using the IP WAN link for cost reduction

c. Using the IP WAN link for cost reduction with remote routing over the WAN, and then transferring into a local PSTN call at the remote gateway

d. Using the PSTN for cost reduction with minimal IP WAN usage

9. **What is the greatest benefit of toll bypass?**

a. It increases the security of VoIP.

b. It creates an effective implementation of Unified Communications.

c. It reduces operating costs by routing internal calls over WAN links as opposed to the PSTN.

d. It implements NAT to allow the routing of calls across the public Internet.