Network Infrastructure

This chapter describes the requirements of the network infrastructure needed to build an IP telephony system in an enterprise environment. Figure 3-1 illustrates the roles of the various devices that form the network infrastructure of a large-scale enterprise network, and Table 3-1 summarizes the features required to support each of these roles.

IP telephony places strict requirements on IP packet loss, packet delay, and delay variation (or jitter). Therefore, you need to enable most of the Quality of Service (QoS) mechanisms available on Cisco switches and routers throughout the network. For the same reasons, redundant devices and network links that provide quick convergence after network failures or topology changes are also important to ensure a highly available infrastructure.

Note In general, this document focuses on standalone and multisite Cisco Unified CallManager Express (Cisco Unified CME) implementations. However, this chapter addresses many issues related to larger enterprise-sized networks. As such, it discusses issues related to Cisco Unified CME deployments featuring centralized call processing. This information is included for context, as Cisco Unified CME is also applicable in larger networks as part of a distributed environment. For more information, see the collection of design guides presentedat: http://www.cisco.com/en/US/products/sw/voicesw/ps4625/products_implementation_design_guides_list.html. The design guides presented there describe deployment and implementation considerations for Cisco Unified CallManager, network infrastructure, and other related topics.

Figure 3-1 Typical Campus Network Infrastructure

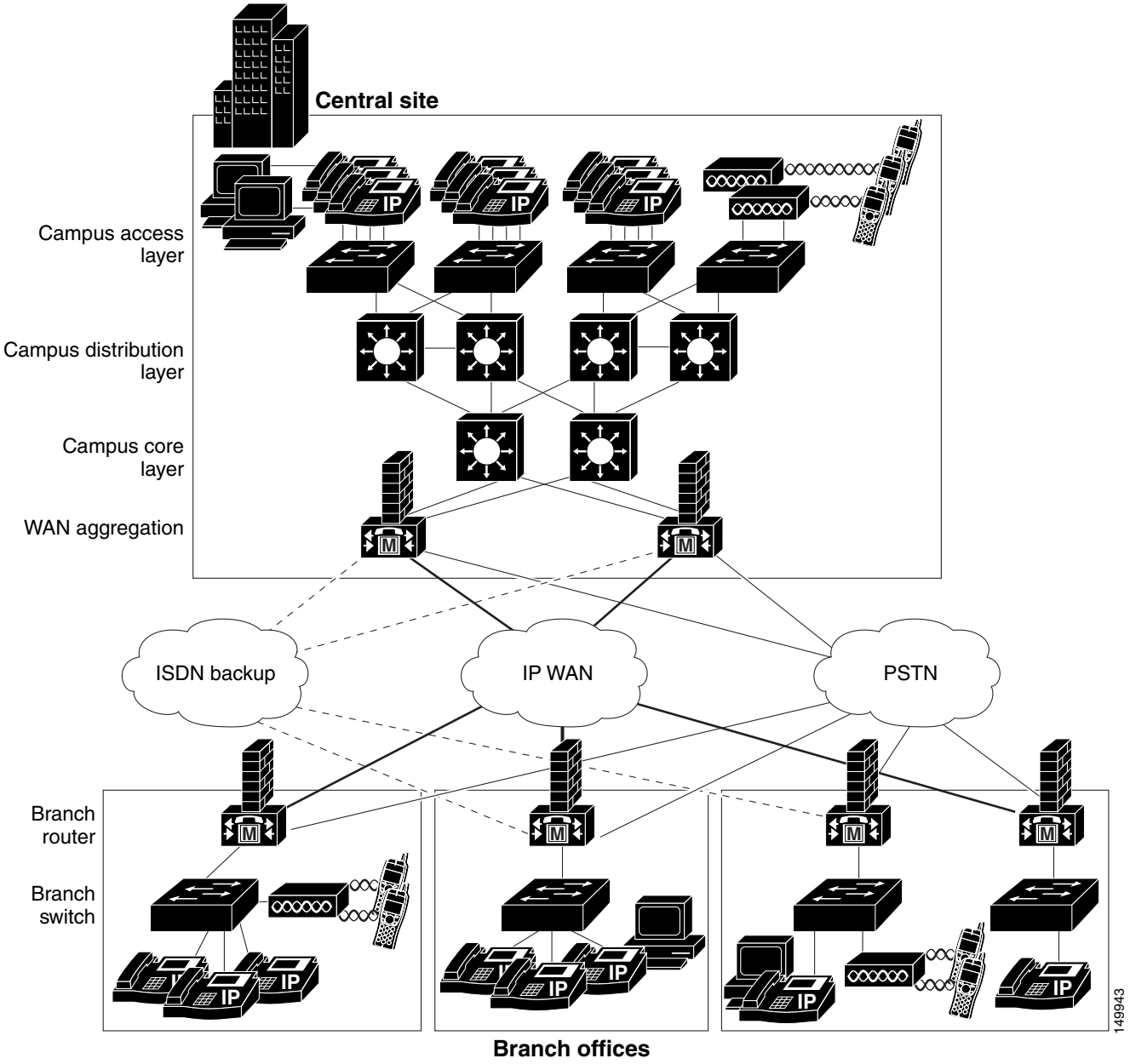


Table 3-1 Required Features for Each Role in the Network Infrastructure

Infrastructure Role	Required Features
Campus Access Switch	• In-Line Power • Multiple Queue Support • 802.1p and 802.1Q • Fast Link Convergence
Campus Distribution or Core Switch	• Multiple Queue Support • 802.1p and 802.1Q • Traffic Classification • Traffic Reclassification
WAN Aggregation Router (Site that is at the hub of the network)	• Multiple Queue Support • Traffic Shaping • Link Fragmentation and Interleaving (LFI) • Link Efficiency • Traffic Classification • Traffic Reclassification • 802.1p and 802.1Q
Branch Router (Spoke site)	• Multiple Queue Support • LFI • Link Efficiency • Traffic Classification • Traffic Reclassification • 802.1p and 802.1Q
Branch or Smaller Site Switch	• In-Line Power • Multiple Queue Support • 802.1p and 802.1Q

The following sections describe the network infrastructure features as they relate to:

- Cisco Unified CME Network Infrastructure Overview, page 3-4
- LAN Infrastructure, page 3-9
- WAN Infrastructure, page 3-20
- Wireless LAN Infrastructure, page 3-33

Note For additional information, see the "Related Documents and References" section on page xii.

Cisco Unified CME Network Infrastructure Overview

This publication focuses on two Cisco Unified CME implementations: standalone and multisite deployments. The general infrastructure considerations for networks supporting Cisco Unified CME are summarized in the following two sections:

- Standalone Network Infrastructure Overview, page 3-4
- Multisite Network Infrastructure Overview, page 3-6

Standalone Network Infrastructure Overview

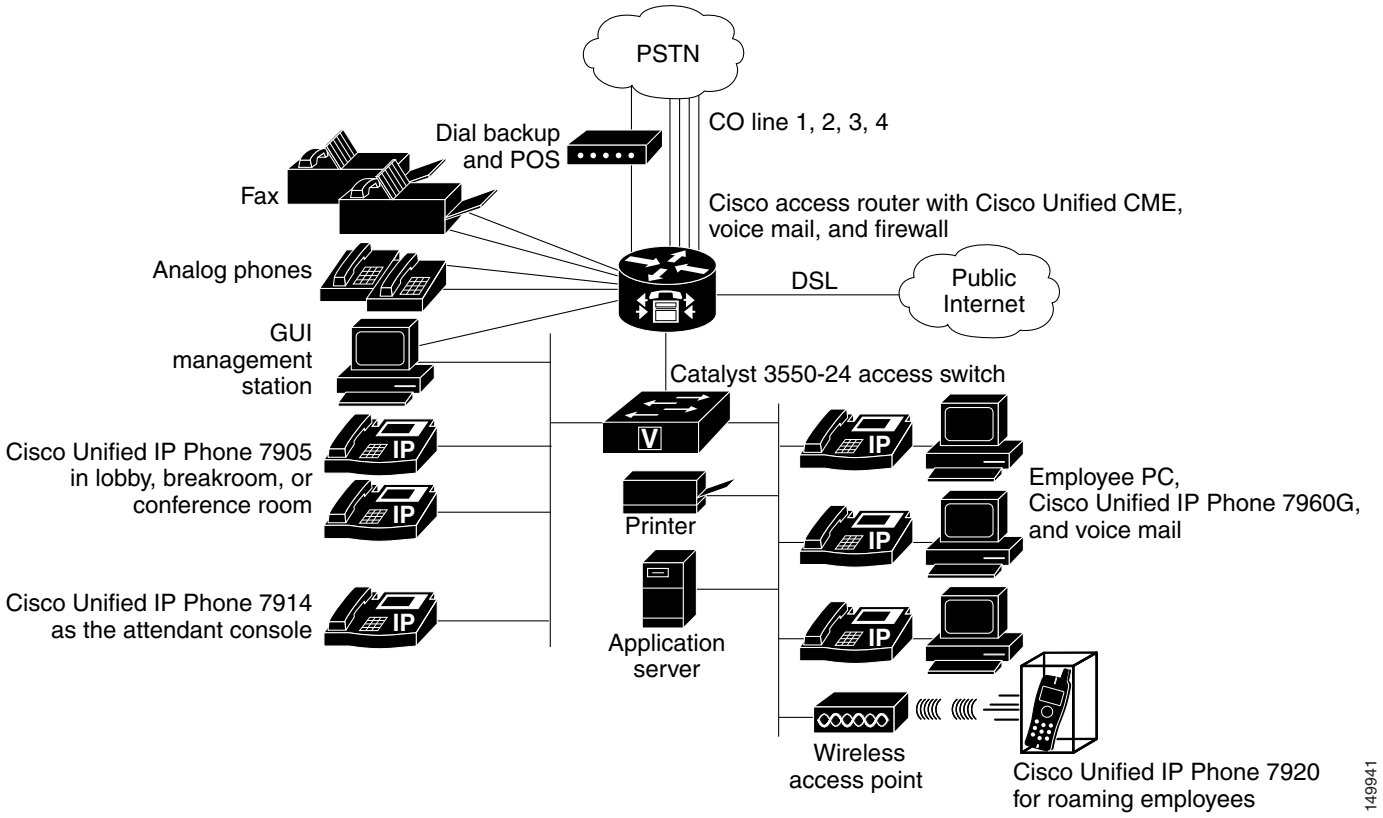
Cisco Unified CME is an excellent choice for a single-site, standalone office. In a world before IP telephony, such an office would have had an onsite router for data services and a separate key system or centrex for voice services. Now the router can be extended to provide converged data and voice services to the office. It also can be managed in the same way as before (either by an ISP or by a VAR or SI). Furthermore, both the business and the SP can realize cost, space, and management savings.

Savings just in wiring of a new office could be enough to make Cisco Unified CME cost-effective. Because the phones and computer equipment are all Ethernet-based, only Ethernet wiring is required in the office. Furthermore, only a single Ethernet wire or jack is required to each employee location or desktop. Computer equipment can be plugged into the back of the phone, and virtual LAN (VLAN) technology can be used to provide virtual separation (and therefore security) of voice from data traffic.

Leading-edge productivity features and improved customer service IP-based applications, such as XML services, can also be deployed easily over this converged infrastructure.

Figure 3-2 shows what such a single-site office's network might look like.

Figure 3-2 Standalone Office Network Topology



The network inFigure 3-2 has the following components:

- *Employee desktop*—Cisco 7960 IP Phones are provided for employees who work at a desk with a computer. The PC is connected via the phone's Ethernet switch. It also is connected via a single Ethernet cable to a LAN switch that provides inline power to the phones. In Figure 3-2, the LAN switch is a separate component, but a LAN switch that optionally provides inline power can also be integrated into the router chassis for offices requiring 50 or fewer LAN connections. The ability to connect computer equipment via the phone substantially reduces the overall number of switch ports required in the office. However, this might require that an existing LAN switch be upgraded to provide inline power for the IP phones. However, inline power is not a requirement for IPT deployments.
- *Internet connectivity*—This is provided via a DSL or a similar type of uplink to the local ISP, which also might host the company's e-mail services. For larger offices, DSL may not have sufficient bandwidth. Internet connectivity may then be deployed via fractional T1/E1 leased-line services, or even a grouping of multiple DSL or Basic Rate Interface (BRI) lines.
- *PSTN trunks*—These PSTN lines are analog Foreign Exchange Office (FXO) connections to the central office (CO). Each line carries a single incoming or outgoing phone call. Caller ID is typically delivered on such connections, but direct inward dial (DID) operation is not. A variation of this offering from the PSTN offers DID operation; this is technically known as analog DID service. It can have a different cost than the plain FXO service. The trunks can also be on a fractional T1/E1 or a full T1/E1 type of service that runs CAS or PRI services. Small businesses often prefer familiar key system operation. In this system, individual PSTN lines are mapped to buttons on the phones

labeled as Line1, Line2, Line3, and so on up to the number of lines coming in from the PSTN central office. (This arrangement is called key-system or square-keyswitch type of deployment.). These can also be used in the PBX-mode in which a user typically dials an access-code (like 9, commonly used in the US) for gain access to an outside PSTN line.

- *Attendant console*—Many small businesses with more than a handful of employees or considerable front-office customer interaction (such as a doctor's office) prefer that an attendant or receptionist answer incoming calls. Although these businesses might use an automated attendant (AA) for after-hours coverage, the typical preferred customer interaction during normal business hours is person-to-person. Attendant consoles can be a Cisco Unified IP Phone 7960 with one or two Cisco Unified IP Phone 7914s providing a total of 34 extensions that can be monitored. Attendant consoles can also be software based consoles from Cisco-certified third-party vendors.
- *Management station*—This is a web-based GUI management application for daily moves, adds, and changes to the system configuration. This can also be any one of the regular PCs used in the office. The only requirement is that it runs Internet Explorer Version 6 or later.
- *Other voice services*—One or more fax machines are used by almost every type of business. A small number of analog phones may also be used around the office, such as for emergency backup PSTN connectivity if power to the building fails.

Low-end IP phones, such as the Cisco Unified IP Phone 7902 or Cisco Unified IP Phone 7905, are scattered throughout the office in break rooms, health clinic exam rooms, lobbies, and perhaps conference rooms. These are often single-line phones that typically are not used to receive calls from the PSTN (they also do not have PC Ethernet ports). Instead, they are used for calls internal to the office or outgoing calls. Being IP phones, though, they participate in the intercom, paging, and display-based features often useful in a small office environment. Access to features, telephony interfaces, and calling plans can be controlled so that these phones are preventing from having access to outside lines.

The Cisco Unified IP Phone 7920 wireless phone can also be a great productivity enhancer for employees whose responsibilities demand both reachability and mobility, such as a retail floor supervisor, a warehouse supervisor, a bank branch manager, or a restaurant shift manager.

Note For information about wireless design for voice, see the *Cisco Wireless IP Phone 7920 Design and Deployment Guide* at the following location: http://www.cisco.com/en/US/docs/voice_ip_comm/cuipph/7920/5_0/english/design/guide/7920ddg.html

However, system installation, initial setup and configuration, software upgrades, and turning on new services are most likely done by the SP or the SI or VAR from whom the system was purchased or leased. If any trouble is experienced, these organizations are responsible for isolating the problem and working with the system's vendor to correct system operation.

Multisite Network Infrastructure Overview

Use Cisco Unified CME there are less than 200 users (so there is some space for growth) and when a centralized provisioning model is not needed. Also Cisco Unified CallManager must be used when implementing certain third-party (and some Cisco) applications that use JTAPI as the control interface. The exact point where a centralized Cisco Unified CallManager starts to make more sense depends on

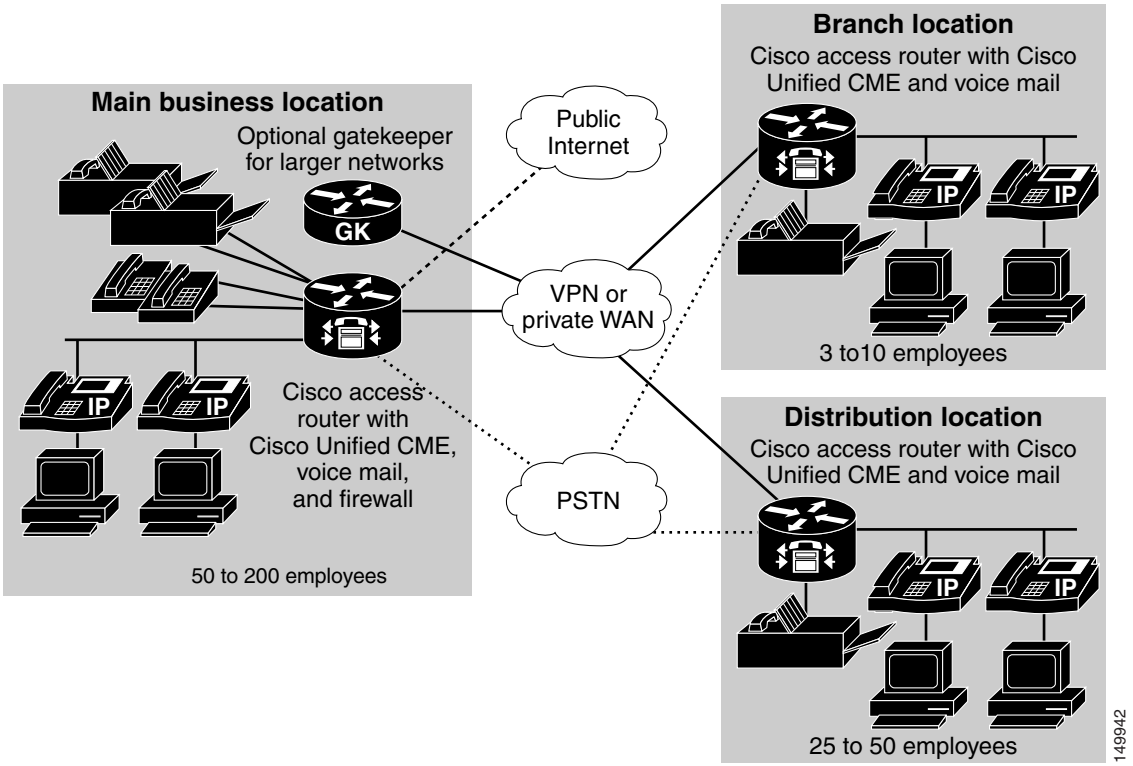
- The individual business
- Its management style
- The QoS readiness of the network between the sites

- The cost basis of the intersite connectivity
- How loosely or tightly coupled the sites are to one another in the normal course of a day's business

A PSTN-based network for voice access is generally recommended for a environment with a loosely coupled operational model and interconnected with only a minimal data network (bandwidth of less than 64 kbps and no QoS deployment). An example of such a business could be a restaurant chain. This network is essentially the same as the standalone model explored in the preceding section. Because the sites have only PSTN calling between them, no VoIP binds the sites together, and the network topology of each location would look like a standalone entity (from a voice traffic perspective). In contrast, a multisite enterprise model that is more tightly coupled (with access to inexpensive QoS) permits consideration of VoIP connectivity between the sites to gain toll savings and other management efficiency advantages. This basic premise of site coupling applies to both Cisco Unified CallManager and Cisco Unified CME solutions.

Figure 3-3 shows a sample network topology of what such an enterprise's branch office network might look like. This representation takes a general view of the branch office.

Figure 3-3 Multisite Distributed Cisco Unified CME Network Topology



There is significant similarity between the detailed layout of the small enterprise branch office and that of the standalone single-site office discussed earlier. The new or additional considerations are as follows:

- Employee desktop—Depending on the business the company conducts, the percentage of employee desktops varies. A retail organization has comparatively few desk-bound employees, whereas a bank or insurance company has a higher percentage. In each case, though, there is an employee who works on the floor or at a teller location, and these stations are often not equipped with individual phones

or computers. Instead, shared resources are deployed for use by these employees. Personal calls are probably made from a public payphone in the break room or from a small number of phones set aside in a shared employee space that employees can access during their breaks.

Desk-bound employees tend to have voice mail, whereas the employees on the retail floor are much less likely to find voice mail productive for their work environment and responsibilities. Sometimes voice mail is still deployed for these employees (again, accessed from a common phone or break room) for human resources or training purposes.

WAN connectivity—The network between the sites is likely to be a private WAN of some type. It could also be a virtual private network (VPN) using the public Internet as the transport, but as such it is not QoS-enabled and, therefore, is not a good fit for deploying VoIP traffic.

A VoIP-capable WAN is most likely either privately owned or provided as a single service to all the sites of the enterprise by a SP. A VPN may still be used on top of the basic network service. Each site's connectivity depends on the site's geographic location and its bandwidth needs. It could be DSL, BRI, fractional T1/E1 access, or even metro-Ethernet. Larger offices may require a full T1/E1 or may bind together multiple DSL or BRI physical access lines to provide larger bandwidth.

- The U.S. offering of integrated access, encompassing both voice and data channels sharing the same physical T1, is a very attractive offering for this type of office. The voice (PSTN) connection could be either T1 in-band signaling (T1 Channel Associated Signaling [T1 CAS]) or fractional PRI. The data connection is most likely Frame Relay.
- PSTN connectivity—PSTN connectivity also depends on the office's size and location. It could be low-density analog (FXO or analog DID) or BRI connections or higher-density fractional T1/E1, perhaps with (fractional) Primary Rate Interface (PRI) service.

  The business model and size of the office dictate whether the office might prefer key system operation (Line1, Line2, and so on appear on the buttons of each phone) or PBX-like operation with typically a single extension per phone and DID service from the CO. Smaller offices more often tend to use key system (shared-line) operation, because that is the traditional voice system they were likely to have had installed before migrating to IP telephony. In larger offices, it becomes impractical to have a button appearance for each incoming CO trunk. These sites tend to be better candidates for DID service. A human or AA provides receptionist services for general incoming business calls and directs clients to the correct department or employee extension.
- Other voice services—When a small number of sites (such as five or fewer) are interconnected, the on-net dial plan is often simple enough to be implemented directly at each site. However, this meshing of sites becomes increasingly complex to manage as the number of sites increases. For this purpose, a gatekeeper (GK) is shown at the main site in Figure 2-5. For enterprises of approximately ten or more locations, centralizing the dial plan management is well worth considering. An H.323 GK is the way to accomplish this when multiple Cisco Unified CME sites are interconnected. This way, the dial plan is administered in a single location and is not duplicated at each site, making changes to the dial plan easy to accomplish.

LAN Infrastructure

LAN infrastructure design is extremely important for proper IP telephony operation on a converged network. Proper LAN infrastructure design requires following basic configuration and design best practices for deploying a highly available network. Further, proper LAN infrastructure design requires deploying end-to-end QoS on the network. The following sections discuss these requirements:

LAN Design for High Availability

Properly designing a LAN requires building a robust and redundant network from the top down. By structuring the LAN as a layered model (see Figure 3-1) and developing the LAN infrastructure one step of the model at a time, you can build a highly available, fault tolerant, and redundant network. Once these layers have been designed properly, you can add network services such as DHCP and TFTP to provide additional network functionality. The following sections examine the infrastructure layers and network services:

Note For more information on campus design, see the *Gigabit Campus Network Design* white paper at http://www.cisco.com/warp/public/cc/so/neso/lnso/cpso/gcnd_wp.pdf

Campus Access Layer

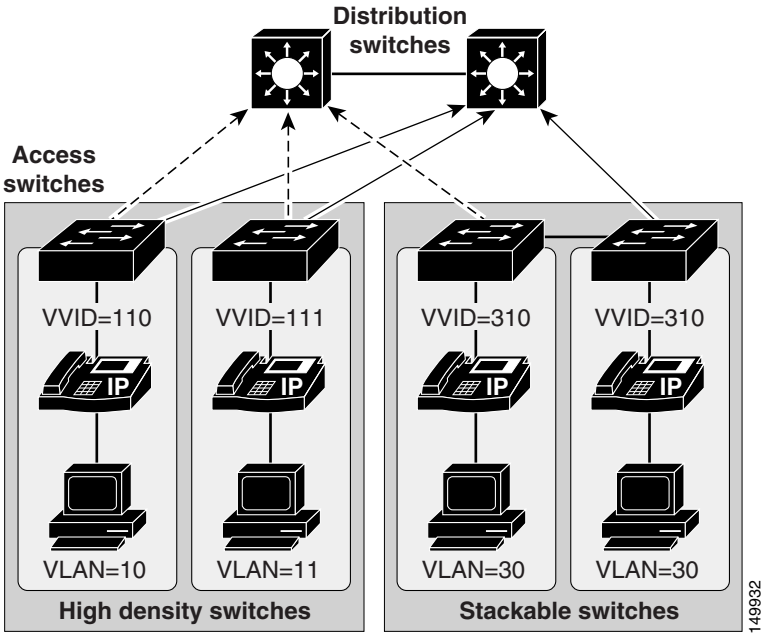
The key layer to consider when implementing a network to support Cisco Unified CME is a the *access layer*. The access layer of the LAN includes the portion of the network from the desktop port(s) to the wiring closet switch.

Note For more information about large scale VoIP deployments, see the following document: http://www.cisco.com/en/US/docs/voice_ip_comm/cucm/srnd/4x/42nstrct.html#wp1043366

Proper access layer design starts with assigning a single IP subnet per virtual LAN (VLAN). Typically, a VLAN should not span multiple wiring closet switches; that is, a VLAN should have presence in one and only one access layer switch (see Figure 3-4). This practice eliminates topological loops at Layer 2, thus avoiding temporary flow interruptions due to Spanning Tree convergence. However, with the introduction of standards-based IEEE 802.1w Rapid Spanning Tree Protocol (RSTP) and 802.1s Multiple Instance Spanning Tree Protocol (MISTP), Spanning Tree can converge at much higher rates. In situations where RSTP and/or MISTP can and have been configured on the access layer switch, there is no need for concern about topological loops. More importantly, confining a VLAN to a single access layer switch also serves to limit the size of the broadcast domain. There is the potential for large numbers of devices within a single VLAN or broadcast domain to generate large amounts of broadcast traffic periodically, which can be problematic. A good rule is to limit the number of devices per VLAN to about 512, which is equivalent to two Class C subnets (that is, a 23-bit subnet masked Class C address). Typical

access layer switches include the stackable Cisco Catalyst 2950, Cisco Catalyst 3500XL, Cisco Catalyst 3550, and Cisco Catalyst 3750, and also the larger, higher-density Catalyst 4000 and 6000 switches.

Figure 3-4 Access Layer Switches and VLANs for Voice and Data



When you deploy voice, we recommend that you enable two VLANs at the access layer: a native VLAN for data traffic (VLANs 10, 11, and 30 in Figure 3-4) and a voice VLAN under Cisco IOS software or Auxiliary VLAN under Catalyst Operating System for voice traffic (represented by VVIDs 110, 111, and 310 in Figure 3-4).

Note For implementations with 75 or fewer phones, the voice VLAN should be the same; the data VLAN should not unique for each switch. In the case of a smaller implementation, the VVID and VLAN should be the same.

We recommend separate voice and data VLANs the following reasons:

- Address space conservation and voice device protection from external networks

  Private addressing of phones on the voice or auxiliary VLAN ensures address conservation and ensures that phones are not accessible directly via public networks. PCs and servers are typically addressed with publicly routed subnet addresses; however, voice endpoints should be addressed using RFC 1918 private subnet addresses.

- QoS trust boundary extension to voice devices

  QoS trust boundaries can be extended to voice devices without extending these trust boundaries and, in turn, QoS features to PCs and other data devices.

- Protection from malicious network attacks

  VLAN access control, 802.1Q, and 802.1p tagging can provide protection for voice devices from malicious internal and external network attacks such as worms, denial of service (DoS) attacks, and attempts by data devices to gain access to priority queues via packet tagging.

- Ease of management and configuration

  Separate VLANs for voice and data devices at the access layer provide ease of management and simplified QoS configuration.

To provide high-quality voice and to take advantage of the full voice feature set, access layer switches should provide support for:

- 802.1Q trunking and 802.1p for proper treatment of Layer 2 CoS packet marking on ports with phones connected
- Multiple egress queues to provide priority queuing of RTP voice packet streams
- The ability to classify or reclassify traffic and establish a network trust boundary
- Inline power capability (Although inline power capability is not mandatory, we highly recommend for the access layer switches.)
- Layer 3 awareness and the ability to implement QoS access control lists (These features are required if you are using certain IP telephony endpoints, such as a PC running a softphone application, that cannot benefit from an extended trust boundary.)

Spanning Tree Protocol (STP)

To minimize convergence times and maximize fault tolerance at Layer 2, enable the following STP features:

- PortFast

  Enable PortFast on all access ports. The phones, PCs, or servers connected to these ports do not forward bridge protocol data units (BPDUs) that could affect STP operation. PortFast ensures that the phone or PC, when connected to the port, is able to begin receiving and transmitting traffic immediately without having to wait for STP to converge.

- Root guard or BPDU guard

  Enable root guard or BPDU guard on all access ports to prevent the introduction of a rogue switch that might attempt to become the Spanning Tree root, thereby causing STP re-convergence events and potentially interrupting network traffic flows. Ports that are set to **errdisable** state by BPDU guard must either be re-enabled manually or the switch must be configured to re-enable ports automatically from the errdisable state after a configured period of time.

- UplinkFast and BackboneFast

  Enable these features where appropriate to ensure that, when changes occur on the Layer 2 network, STP converges as rapidly as possible to provide high availability. When using stackable switches such as the Cisco Catalyst 2950, Cisco Catalyst 3550, or Cisco Catalyst 3750, enable Cross-Stack UplinkFast (CSUF) to provide fast failover and convergence if a switch in the stack fails.

- UniDirectional Link Detection (UDLD)

  Enable this feature to reduce convergence and downtime on the network when link failures or misbehaviors occur, thus ensuring minimal interruption of network service. UDLD detects, and takes out of service, links where traffic is flowing in only one direction. This feature prevents defective links from being mistakenly considered as part of the network topology by the Spanning Tree and routing protocols.

Note With the introduction of RSTP 802.1w, features such as PortFast and UplinkFast are not required because these mechanisms are built in to this standard. If RSTP has been enabled on the Catalyst switch, these commands are not necessary.

Network Services

Once a highly available, fault-tolerant, multi-layer campus network has been built, network services such as DNS, DHCP, TFTP, and NTP can be deployed. These topics are addressed in the following individual sections:

Domain Name System (DNS)

DNS enables the mapping of host names to IP addresses within a network or networks. DNS server(s) deployed within a network provide a database that maps hostnames to IP addresses. Devices on the network can query the DNS server and receive IP addresses for other devices in the network, thereby facilitating communication between network devices.

Relying on DNS, however, can be problematic. If the DNS server becomes unavailable and a network device is relying on that server to provide a hostname-to-IP-address mapping, communication can and will fail. For this reason, do not rely on DNS for communication between Cisco Unified CME and the IP telephony endpoints.

Configure Cisco Unified CME systems, gateways, and endpoint devices to use IP addresses rather than hostnames. We do *not* recommend configuration of DNS parameters such as DNS server addresses, hostnames, and domain names. If you eliminate DNS configuration within the IP telephony network, telephony devices and applications do not have to rely on the DNS server.

Dynamic Host Configuration Protocol (DHCP)

DHCP is used by hosts on the network to get initial configuration information, including IP address, subnet mask, default gateway, and TFTP server. DHCP eases the administrative burden of manually configuring each host with an IP address and other configuration information. DHCP also provides automatic reconfiguration of network configuration when devices are moved between subnets. The configuration information is provided by a DHCP server located in the network, which responds to DHCP requests from DHCP-capable clients.

You should configure IP telephony endpoints to use DHCP to simplify deployment of these devices. Any RFC 2131 compliant DHCP server can be used to provide configuration information to IP telephony network devices. When deploying IP telephony devices in an existing data-only network, all you have to do is add DHCP voice scopes to an existing DHCP server for these new voice devices. Because IP telephony devices are configured to use and rely on a DHCP server for IP configuration information, you must deploy DHCP servers in a redundant fashion. At least two DHCP servers should be deployed within the telephony network such that, if one of the servers fails, the other can continue to answer DHCP client requests. You should also ensure that DHCP server(s) are configured with enough IP subnet addresses to handle all DHCP-reliant clients within the network.

Note The preceding information applies to large-scale networks. For most Cisco Unified CME deployments, the Cisco IOS router running Cisco Unified CME can also be the DHCP server.

DHCP Option 150

IP telephony endpoints can be configured to rely on DHCP Option 150 to identify the source of telephony configuration information, available from a server running the Trivial File Transfer Protocol (TFTP).

In the simplest configuration, where a single TFTP server is offering service to all deployed endpoints, Option 150 is delivered as a single IP address pointing to the system's designated TFTP server.

We highly recommend using a direct IP address (that is, not relying on a DNS service) for Option 150 because doing so eliminates dependencies on DNS service availability during the phone boot-up and registration process.

DHCP Lease Times

Configure DHCP lease times as appropriate for the network environment. Given a fairly static network in which PCs and telephony devices remain in the same place for long periods of time, we recommend longer DHCP lease times (for example, one week). Shorter lease times require more frequent renewal of the DHCP configuration and increase the amount of DHCP traffic on the network. Conversely, networks that incorporate large numbers of mobile devices, such as laptops and wireless telephony devices, should be configured with shorter DHCP lease times (for example, one day) to prevent depletion of DHCP-managed subnet addresses. Mobile devices typically use IP addresses for short increments of time and then might not request a DHCP renewal or new address for a long period of time. Longer lease times will tie up these IP addresses and prevent them from being reassigned even when they are no longer being used.

Cisco Unified IP Phones adhere to the conditions of the DHCP lease duration as specified in the DHCP server's scope configuration. Once half the lease time has expired since the last successful DHCP server Acknowledgment, the IP phone will request a lease renewal. This DHCP client Request, acknowledged by the DHCP server, will allow the IP phone to retain use of the IP scope (that is, the IP address, default gateway, subnet mask, DNS server (optional), and TFTP server (optional)) for another lease period. If the DHCP server becomes unavailable, an IP phone will not be able to renew its DHCP lease, and as soon as the lease expires, it will relinquish its IP configuration and will thus become unregistered from Cisco CallManager until a DHCP server can grant it another valid scope.

DHCP Network Deployments

There are two options for deploying DHCP functionality within an IP telephony network

- Centralized DHCP Server

  Typically, for a single-site campus IP telephony deployment, the DHCP server should be installed at a central location within the campus. As mentioned previously, redundant DHCP servers should be deployed. If the IP telephony deployment also incorporates remote branch telephony sites, as in a centralized multi-site Cisco CallManager deployment, a centralized server can be used to provide DHCP service to devices in the remote sites. This type of deployment requires that you configure the **ip helper-address** on the branch router interface. If redundant DHCP servers are deployed at the central site, both servers' IP addresses must be configured as **ip helper-address**. Also note that, if branch-side telephony devices rely on a centralized DHCP server and the WAN link between the two sites fails, devices at the branch site will be unable to send DHCP requests or receive DHCP responses.

Note By default, **service dhcp** is enabled on the Cisco IOS device and does not appear in the configuration. Do not disable this service on the branch router because doing so will disable the DHCP relay agent on the device, and the **ip helper-address** configuration command will not work.

- Centralized DHCP Server and Remote Site Cisco IOS DHCP Server

  When configuring DHCP for use in a centralized multi-site Cisco Unified CME deployment, you can use a centralized DHCP server to provide DHCP service to centrally located devices. Remote devices could receive DHCP service from a locally installed server or from the Cisco IOS router at the remote site. This type of deployment ensures that DHCP services are available to remote telephony devices even during WAN failures. The following example lists the basic Cisco IOS DHCP server configuration commands.

```
service dhcp                                ! Activate DHCP Service on the IOS Device

ip dhcp excluded-address <ip-address>|<ip-address-low> <ip-address-high>
                                            !  Specify an IP Address or IP Address Range to
                                            !  be excluded from the DHCP pool

ip dhcp pool <dhcp-pool name>               ! Specify DHCP pool name
   network <ip-subnet> <mask>               ! Specify DHCP pool IP subnet and mask
   default-router <default-gateway-ip>      ! Specify the Default-gateway
   option 150 ip <tftp-server-ip-1> ...     ! Specify TFTP servers (up to four) -
                                            ! IP phones use only the first two addresses in
                                            ! the array.
```

Trivial File Transfer Protocol (TFTP)

Within a Cisco Unified CME system, endpoints (such as IP phones running the SCCP protocol) rely on a TFTP-based process to acquire configuration information. The endpoints request a configuration file whose name is based on the requester's MAC address (for example, for an IP phone with MAC address ABCDEF123456, the filename would be SEPABCDEF123456.cnf.xml). The configuration file includes the version of software that the phone must run and a list of Cisco Unified CME servers that the phone should register with.

If the configuration file instructs the phone to run a software file other than the one it currently uses, the phone will request the new version of software from the TFTP server. The phone goes through this process once per reboot of the phone or router, before registering.

Centralized call processing deployments require remote phones to download configuration files and phone software through the branch's WAN link. When scheduled maintenance involves the downloading of new software, download times are a function of the number of phones requiring upgrades, the file size, and the WAN link's bandwidth and traffic utilization.

Network Time Protocol (NTP)

NTP allows network devices to synchronize their clocks to a network time server or network-capable clock. NTP is critical for ensuring that all devices in a network have the same time. When troubleshooting or managing a telephony network, it is crucial to synchronize the time stamps within all error and security logs, traces, and system reports on devices throughout the network. This synchronization enables administrators to recreate network activities and behaviors based on a common timeline. Billing records and call detail records (CDRs) also require accurate synchronized time.

Cisco Unified CME NTP Time Synchronization

Time synchronization is especially critical on Cisco Unified CME devices. Configure automatic NTP time synchronization on all Cisco Unified CME servers within the network.

Cisco IOS and Catalyst Operating System NTP Time Synchronization

Time synchronization is also important for other devices within the network. Cisco IOS routers and Catalyst switches should be configured to synchronize their time with the rest of the network devices via NTP. This is critical for ensuring that debug, syslog, and console log messages are time-stamped appropriately. Troubleshooting telephony network issues is simplified when a clear timeline can be drawn for events that occur on devices throughout the network.

The following example illustrates the configuration of NTP time synchronization on Cisco IOS and Catalyst Operating System devices.

Cisco IOS software configuration:

```
ntp server 64.100.21.254
```

Catalyst Operating System configuration:

```
set ntp server 64.100.21.254
set ntp client enable
```

To ensure proper NTP time synchronization on routers and switches, it may be necessary to configure time zones using the **clock timezone** command (in Cisco IOS software) and/or **set timezone** command (in Catalyst Operating System).

When an external connection to an NTP server is not available, Cisco IOS software can be used as a reference server for other devices so that all devices including phones use the same time reference. This can be done by using the global Cisco IOS software **ntp master** command in configuration mode. Also set the clock and time zone on the router.

Power over Ethernet (PoE)

PoE (or inline power) is 48 Volt DC power provided over standard Ethernet unshielded twisted-pair (UTP) cable. Instead of using wall power, IP phones and other inline powered devices (PDs) such as the Aironet Wireless Access Points can receive power provided by inline power-capable Catalyst Ethernet switches or other inline power source equipment (PSE). Inline power is enabled by default on all inline power-capable Catalyst switches.

Deploying inline power-capable switches with uninterruptable power supplies (UPS) ensures that IP phones continue to receive power during power failure situations. Provided the rest of the telephony network is available during these periods of power failure, then IP phones should be able to continue making and receiving calls. You should deploy inline power-capable switches at the campus access layer within wiring closets to provide inline-powered Ethernet ports for IP phones, thus eliminating the need for wall power.

Cisco PoE is delivered on the same wire pairs used for data connectivity (pins 1, 2, 3, and 6). If existing access switch ports are not capable of inline power, you can use a power patch panel to inject power into the cabling. (In this case pins 4, 5, 7, and 8 are used.) Additionally, power injectors may be used for specific deployment needs.

Caution The use of power injectors or power patch panels can damage some devices because power is always applied to the Ethernet pairs. PoE switch ports automatically detect the presence of a device that requires PoE before enabling it on a port-by-port basis.

In addition to Cisco PoE inline power, we now support the IEEE 802.3af PoE standard. Not all access switches and phones comply with 802.3af. The Cisco Catalyst 6500, Cisco Catalyst 4500, and Cisco Catalyst 3750 are capable of supporting 802.3af. For information about which Cisco Unified IP Phones support the 802.3af PoE standard, see the information regarding PoE switches provided at the following:

Category 3 Cabling

The use of Category 3 cabling is supported for IP Communications under the following conditions:

- Phones with a PC port and a PC attached to it (Cisco Unified IP Phone 7970, Cisco Unified IP Phone 7960, Cisco Unified IP Phone 7940, and Cisco Unified IP Phone 7910+SW) should be set to 10 Mbps, full-duplex.

  This setting requires hard-coding the upstream switch port, the phone switch and PC ports, and the PC NIC port to 10 Mbps, full-duplex. No ports should be set to AUTO negotiate. If desired, you can hard-code the phone's PC port to 10 Mbps half-duplex, thereby forcing the PC's NIC to negotiate to 10 Mbps half-duplex (assuming the PC's NIC is configured to AUTO negotiate). This configuration is acceptable as long as the uplink between the phone and the upstream switch port is set to 10 Mbps full-duplex.

- Phones with no PC ports and with 10 Mbps switch ports (Cisco Unified IP Phone 7902, Cisco Unified IP Phone 7905, and Cisco Unified IP Phone 7910 IP Phones) should be allowed to auto-negotiate to 10 Mbps, half-duplex.

  Because these phones support only 10 Mbps Ethernet and their ports cannot be manually configured, the upstream switch port should be set to either AUTO negotiate or 10 Mbps, half-duplex. In both cases, these phones will negotiate to 10 Mbps, half-duplex.

- Phones with a PC port but no PC attached to it (Cisco Unified IP Phone 7970, Cisco Unified IP Phone 7960, Cisco Unified IP Phone 7940, Cisco Unified IP Phone 7910+SW, and Cisco Unified IP Phone 7912) can be allowed to negotiate to 10 Mbps, half-duplex.

  If you leave these phones with the default switch port configuration of AUTO negotiate and configure the upstream switch port to 10 Mbps, half-duplex, these phones will revert to 10 Mbps, half-duplex.

Note The Cisco Unified IP Phone 7912 IP Phone should not be used with Category 3 cable when a PC is attached because the switch and PC ports on this phone cannot be forced to 10 Mbps, full duplex.

IBM Type 1A and 2A Cabling

The use of IBM Cabling System (ICS) or Token Ring shielded twisted-pair type 1A or 2A cabling is supported for IP Communications under the following conditions:

- Cable lengths should be 100 meters or less.
- Adapters without impedance matching should be used for converting from universal data connector (UDC) to RJ-45 Ethernet standard.

Note There are only two twisted pairs in the Token Ring cables. Therefore, inline power for IP phones can be supported, but mid-span power insertion cannot (with Cisco Inline Power and 802.3af) because it requires more than two pairs.

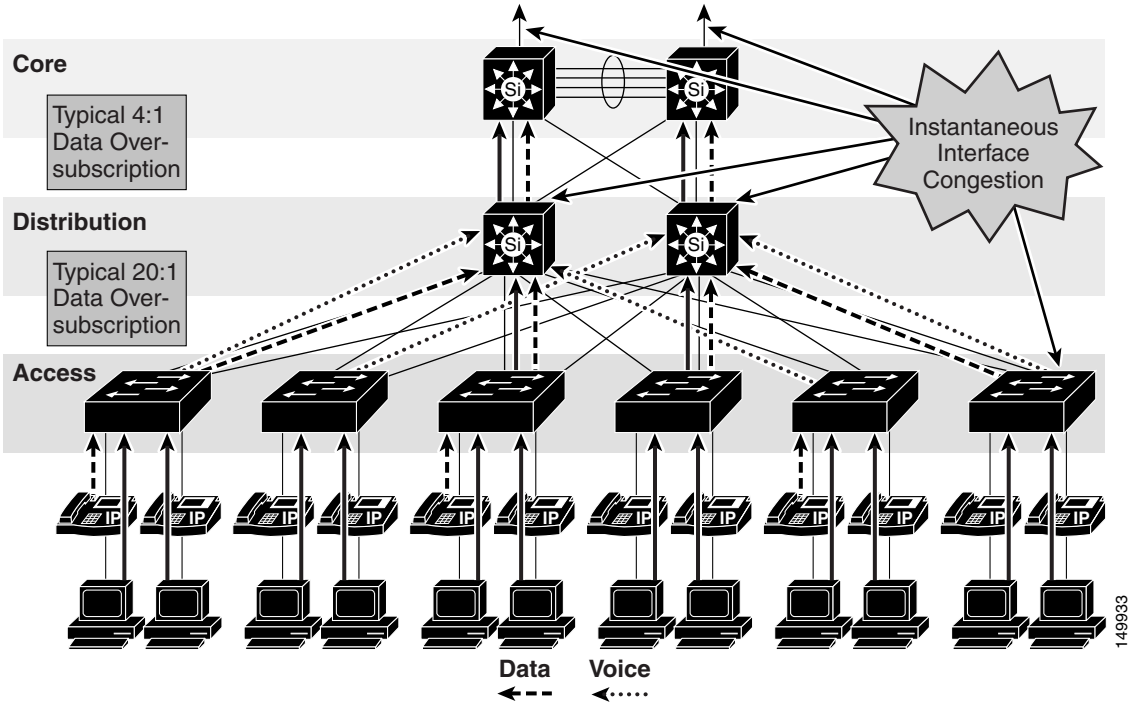
Running data over the network is not always a sufficient test of the quality of the cable plant because some non-compliance issues might not be apparent. Therefore, customers might want to perform a cable plant survey to verify that their type 1A and 2A cabling installation is compliant with Ethernet standards.

LAN Quality of Service (QoS)

Until recently, quality of service was not an issue in the enterprise campus because of the asynchronous nature of data traffic and the ability of network devices to tolerate buffer overflow and packet loss. However, with new applications such as voice and video, which are sensitive to packet loss and delay, buffers and not bandwidth are the key QoS issue in the enterprise campus.

Figure 3-5 illustrates the typical oversubscription that occurs in LAN infrastructures.

Figure 3-5 Data Traffic Oversubscription in the LAN



This oversubscription, coupled with individual traffic volumes and the cumulative effects of multiple independent traffic sources, can result in the egress interface buffers becoming full instantaneously, thus causing additional packets to drop when they attempt to enter the egress buffer. The fact that campus switches use hardware-based buffers, which compared to the interface speed are much smaller than those found on WAN interfaces in routers, merely increases the potential for even short-lived traffic bursts to cause buffer overflow and dropped packets.

Applications such as file sharing (both peer-to-peer and server-based), remote networked storage, network-based backup software, and emails with large attachments, can create conditions where network congestion occurs more frequently and/or for longer durations. Some of the negative effects of recent worm attacks have been an overwhelming volume of network traffic (both unicast and broadcast-storm based), increasing network congestion. If no buffer management policy is in place, loss, delay, and jitter performance of the LAN may be affected for all traffic.

Another situation to consider is the effect of failures of redundant network elements, which cause topology changes. For example, if a distribution switch fails, all traffic flows will be reestablished through the remaining distribution switch. Prior to the failure, the load balancing design shared the load between two switches, but after the failure all flows are concentrated in a single switch, potentially causing egress buffer conditions that normally would not be present.

For applications such as voice, this packet loss and delay results in severe voice quality degradation. Therefore, QoS tools are required to manage these buffers and to minimize packet loss, delay, and delay variation (jitter).

The following types of QoS tools are needed from end to end on the network to manage traffic and ensure voice quality:

- Traffic classification

  Classification involves the marking of packets with a specific priority denoting a requirement for class of service (CoS) from the network. The point at which these packet markings are trusted or not trusted is considered the trust boundary. Trust is typically extended to voice devices (phones) and not to data devices (PCs).

- Queuing or scheduling

  Interface queuing or scheduling involves assigning packets to one of several queues based on classification for expedited treatment throughout the network.

- Bandwidth provisioning

  Provisioning involves accurately calculating the required bandwidth for all applications plus element overhead.

The following sections discuss the use of these QoS mechanisms in a campus environment:

- Traffic Classification, page 3-18
- Interface Queuing, page 3-19
- Bandwidth Provisioning, page 3-19
- Impairments to IP Communications if QoS is Not Employed, page 3-20

Traffic Classification

It has always been an integral part of the Cisco network design architecture to classify or mark traffic as close to the edge of the network as possible. Traffic classification is an entrance criterion for access into the various queuing schemes used within the campus switches and WAN interfaces. The IP phone marks its voice control signaling and voice RTP streams at the source, and it adheres to the values presented in Table 3-2. As such, the IP phone can and should classify traffic flows.

Table 3-2 lists the traffic classification requirements for the LAN infrastructure.

Table 3-2 Traffic Classification Guidelines for Various Types of Network Traffic

Traffic Type	Layer 2 Class of Service (CoS)	Layer 3 IP Precedence	Layer 3 Differentiated Services Code Point (DSCP)	Layer 3 Per-Hop Behavior (PHB)
Voice Real-Time Transport Protocol (RTP)	5	5	46	EF
Voice control signaling[1]	3	3	24	CS3
Video conferencing	4	4	34	AF41
Data	0, 1, 2	0, 1, 2	10 to 22	BE to AF23

1. The recommended DSCP/PHB marking for voice control signaling traffic has been changed from 26/AF31 to 24/CS3. A marking migration is planned within Cisco to reflect this change, however many products still mark signaling traffic as 26/AF31. Therefore, in the interim, we recommend that both AF31 and CS3 be reserved for call signaling.

Interface Queuing

After packets have been marked with the appropriate tag at Layer 2 (CoS) and Layer 3 (DSCP or PHB), it is important to configure the network to schedule or queue traffic based on this classification, so as to provide each class of traffic with the service it needs from the network. By enabling QoS on campus switches, you can configure all voice traffic to use separate queues, thus virtually eliminating the possibility of dropped voice packets when an interface buffer fills instantaneously.

Although network management tools may show that the campus network is not congested, QoS tools are still required to guarantee voice quality. Network management tools show only the average congestion over a sample time span. While useful, this average does not show the congestion peaks on a campus interface.

Transmit interface buffers within a campus tend to congest in small, finite intervals as a result of the bursty nature of network traffic. When this congestion occurs, any packets destined for that transmit interface are dropped. The only way to prevent dropped voice traffic is to configure multiple queues on campus switches. For this reason, we recommend always using a switch that has at least two output queues on each port and the ability to send packets to these queues based on QoS Layer 2 and/or Layer 3 classification. Cisco Catalyst 6000, Cisco Catalyst 4000, Cisco Catalyst 3750, Cisco Catalyst 35XX, and Cisco Catalyst 2950 switches all support two or more output queues per port.

Bandwidth Provisioning

In the campus LAN, bandwidth provisioning recommendations can be summarized by the motto, *Over provision and under subscribe*. This motto implies careful planning of the LAN infrastructure so that the available bandwidth is always considerably higher than the load and there is no steady-state congestion over the LAN links.

The addition of voice traffic onto a converged network does not represent a significant increase in overall network traffic load; the bandwidth provisioning is still driven by the demands of the data traffic requirements. The design goal is to avoid extensive data traffic congestion on any link that will be traversed by telephony signaling or media flows. Contrasting the bandwidth requirements of a single G.711 voice call (approximately 86 kbps) to the raw bandwidth of a FastEthernet link (100 Mbps) indicates that voice is not a source of traffic that causes network congestion in the LAN, but rather it is a traffic flow to be protected from LAN network congestion.

Impairments to IP Communications if QoS is Not Employed

If QoS is not deployed, packet drops and excessive delay and jitter can occur, leading to impairments of the telephony services. When media packets are subjected to drops, delay, and jitter, the user-perceivable effects include clicking sound, harsh-sounding voice, extended periods of silence, and echo.

When signaling packets are subjected to the same conditions, user-perceivable impairments include unresponsiveness to user input (such as delay to dial tone), continued ringing upon answer, and double dialing of digits due to the user's belief that the first attempt was not effective (thus requiring hang-up and redial). More extreme cases can include endpoint re-initialization, call termination, and the spurious activation of Cisco SRST functionality at branch offices (leading to interruption of gateway calls).

These effects apply to all deployment models. However, single-site (campus) deployments tend to be less likely to experience the conditions caused by sustained link interruptions because the larger quantity of bandwidth typically deployed in LAN environments (minimum links of 100 Mbps) allows for some residual bandwidth to be available for the IP Communications system.

In any WAN-based deployment model, traffic congestion is more likely to produce sustained and/or more frequent link interruptions because the available bandwidth is much less than in a LAN (typically less than 2 Mbps), so the link is more easily saturated. The effects of link interruptions impact the users, whether or not the voice media traverses the packet network.

WAN Infrastructure

Proper WAN infrastructure design is important for proper IP telephony operation on a converged network with two or more Cisco Unified CME systems or Cisco Unified CME systems along with Cisco Unified CallManager systems. If VoIP calls are exchanged between sites, WAN considerations are important.

Proper infrastructure design requires following basic configuration and design best practices for deploying a WAN that is as highly available as possible and that provides guaranteed throughput. Furthermore, proper WAN infrastructure design requires deploying end-to-end QoS on all WAN links. The following sections discuss these requirements:

- WAN Design and Configuration Best Practices, page 3-20
- WAN Quality of Service (QoS), page 3-22

WAN Design and Configuration Best Practices

Properly designing a WAN requires building fault-tolerant network links and planning for the possibility that these links might become unavailable. By carefully choosing WAN topologies, provisioning the required bandwidth, and approaching the WAN infrastructure as another layer in the network topology, you can build a fault-tolerant and redundant network. The following sections examine the required infrastructure layers and network services:

- Deployment Considerations, page 3-21
- Guaranteed Bandwidth, page 3-21
- Best-Effort Bandwidth, page 3-22

Deployment Considerations

WAN deployments for voice networks must follow a hub-and-spoke topology, with a central hub site and multiple remote spoke sites connected into the central hub site. In this scenario, each remote or spoke site is one WAN link hop away from the central or hub site and two WAN link hops away from all other spoke sites.

WAN links should, when possible, be made redundant to provide higher levels of fault-tolerance. Redundant WAN links provided by different service providers or located in different physical ingress/egress points within the network can ensure backup bandwidth and connectivity in the event that a single link fails. In non-failure scenarios, these redundant links may be used to provide additional bandwidth and offer load balancing of traffic on a per-flow basis over multiple paths and equipment within the WAN.

Voice and data should remain converged at the WAN, just as they are converged at the LAN. QoS provisioning and queuing mechanisms are typically available in a WAN environment to ensure that voice and data can interoperate on the same WAN links. Attempts to separate and forward voice and data over different links can be problematic in many instances because the failure of one link typically forces all traffic over a single link, thus diminishing throughput for each type of traffic and in most cases reducing the quality of voice. Furthermore, maintaining separate network links or devices makes troubleshooting and management difficult at best.

When deploying voice in a WAN environment, we recommend that you use the lower-bandwidth G.729 codec for any voice calls that will traverse WAN links because this practice will provide bandwidth savings on these lower-speed links. Furthermore, media resources such as MoH should be configured to use multicast transport mechanism when possible because this practice will provide additional bandwidth savings.

Finally, recommendation G.114 of the International Telecommunication Union (ITU) states that the one-way delay in a voice network should be less than or equal to 150 milliseconds. It is important to keep this in mind when implementing low-speed WAN links within a network. Topologies, technologies, and physical distance should be considered for WAN links so that one-way delay is kept at or below this 150-millisecond recommendation.

Guaranteed Bandwidth

Because voice is typically deemed a critical network application, it is imperative that bearer and signaling voice traffic always reaches its destination. For this reason, it is important to choose a WAN topology and link type that can provide guaranteed dedicated bandwidth. The following WAN link technologies can provide guaranteed dedicated bandwidth:

- Leased Lines
- Frame Relay
- Asynchronous Transfer Mode (ATM)
- ATM-to-Frame Relay Service Interworking
- Multiprotocol Label Switching (MPLS)
- Cisco Voice and Video Enabled IP Security VPN (IP Sec V3PN)

These link technologies, when deployed in a dedicated fashion or when deployed in a private network, can provide guaranteed traffic throughput. All of these WAN link technologies can be provisioned at specific speeds or bandwidth sizes. In addition, these link technologies have built-in mechanisms that help guarantee throughput of network traffic even at low link speeds. Features such as traffic shaping,

fragmentation and packet interleaving, and committed information rates (CIR) can help ensure that packets are not dropped in the WAN, that all packets are given access at regular intervals to the WAN link, and that enough bandwidth is available for all network traffic attempting to traverse these links.

Best-Effort Bandwidth

There are some WAN topologies that are unable to provide guaranteed dedicated bandwidth to ensure that network traffic will reach its destination, even when that traffic is critical. These topologies are extremely problematic for voice traffic, not only because they provide no mechanisms to provision guaranteed network throughput, but also because they provide no traffic shaping, packet fragmentation and interleaving, queuing mechanisms, or end-to-end QoS to ensure that critical traffic such as voice will be given preferential treatment.

The following WAN network topologies and link types are examples of best-effort bandwidth technology:

- The Internet
- DSL
- Cable
- Satellite
- Wireless

In most cases, these link types can provide the guaranteed network connectivity and bandwidth required for critical voice and voice applications. However, these technologies might be suitable for personal or telecommuter-type network deployments. At times, these topologies can provide highly available network connectivity and adequate network throughput; but at other times, these topologies can become unavailable for extended periods of time, can be throttled to speeds that render network throughput unacceptable for real-time applications such as voice, or can cause extensive packet losses and require repeated retransmissions. In other words, these links and topologies are unable to provide guaranteed bandwidth, and when traffic is sent on these links, it is sent best-effort with no guarantee that it will reach its destination. For this reason, we recommend that you do *not* use best-effort WAN topologies for voice-enabled networks that require enterprise-class voice services and quality.

Note There are some new QoS mechanisms for DSL and cable technologies that can provide guaranteed bandwidth; however, these mechanisms are not typically deployed by service providers, and these services are still significantly oversubscribed.

WAN Quality of Service (QoS)

Before placing voice and video traffic on a network, it is important to ensure that there is adequate bandwidth for all required applications. After this bandwidth has been provisioned, voice priority queuing must be performed on all interfaces. This queuing is required to reduce jitter and possible packet loss if a burst of traffic oversubscribes a buffer. This queuing requirement is similar to the one for the LAN infrastructure.

Next, the WAN typically requires additional mechanisms such as traffic shaping to ensure that WAN links are not sent more traffic than they can handle, which could cause dropped packets.

Finally, link efficiency techniques can be applied to WAN paths. For example, link fragmentation and interleaving (LFI) can be used to prevent small voice packets from being queued behind large data packets, which could lead to unacceptable delays on low-speed links.

The goal of these QoS mechanisms is to ensure reliable, high-quality voice by reducing delay, packet loss, and jitter for the voice traffic. Table 3-3 lists the QoS features and tools required for the WAN infrastructure to achieve this goal.

Table 3-3 QoS Features and Tools Required to Support IP Telephony for each WAN Technology and Link Speed

WAN Technology	Link Speed: 56-to-768 kbps	Link Speed: Greater than 768 kbps
Leased Lines	• Multilink Point-to-Point Protocol (MLP) • MLP Link Fragmentation and Interleaving (LFI) • Low Latency Queuing (LLQ) • Optional: Compressed Real-Time Transport Protocol (cRTP)	• LLQ
Frame Relay	• Traffic Shaping • LFI (FRF.12) • LLQ • Optional: cRTP • Optional: Voice-Adaptive Traffic Shaping (VATS) • Optional: Voice-Adaptive Fragmentation (VAF)	• Traffic Shaping • LLQ • Optional: VATS
Asynchronous Transfer Mode (ATM)	• TX-ring buffer changes • MLP over ATM • MLP LFI • LLQ • Optional: cRTP (requires MLP)	• TX-ring buffer changes • LLQ
Frame Relay and ATM Service Interworking (SIW)	• TX-ring buffer changes • MLP over ATM and FR • MLP LFI • LLQ • Optional: cRTP (requires MLP)	• TX-ring buffer changes • MLP over ATM and FR • LLQ
Multiprotocol Label Switching (MPLS)	• Same as above, according to the interface technology • Class-based marking is generally required to remark flows according to service provider specifications	• Same as above, according to the interface technology • Class-based marking is generally required to remark flows according to service provider specifications

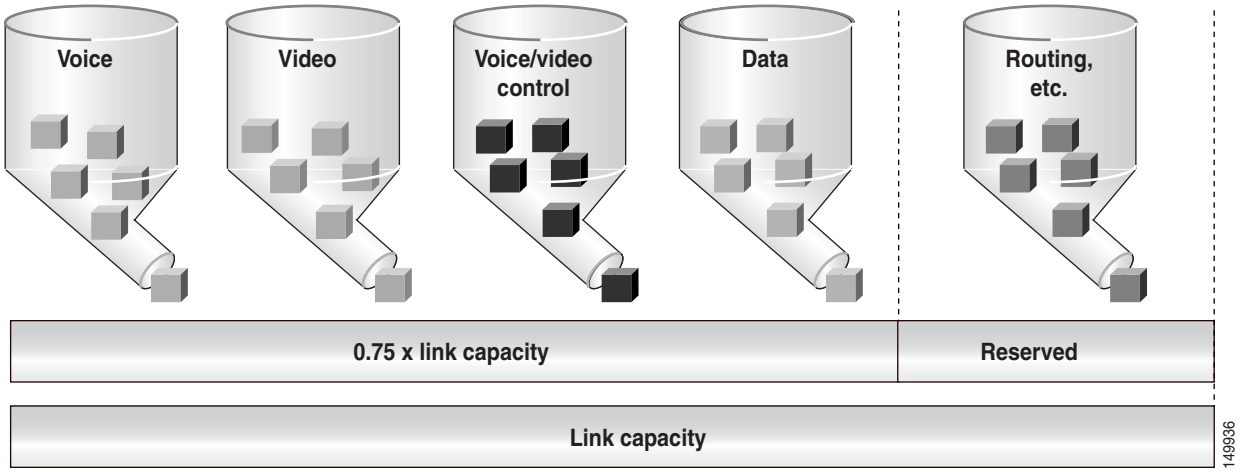
The following sections highlight some of the most important features and techniques to consider when designing a WAN to support both voice and data traffic:

- Bandwidth Provisioning, page 3-24
- Traffic Prioritization, page 3-27
- Link Efficiency Techniques, page 3-28
- Traffic Shaping, page 3-31

Bandwidth Provisioning

Properly provisioning the network bandwidth is a major component of designing a successful IP network. You can calculate the required bandwidth by adding the bandwidth requirements for each major application (for example, voice, video, and data). This sum then represents the minimum bandwidth requirement for any given link, and it should not exceed approximately 75 percent of the total available bandwidth for the link. This 75 percent rule assumes that some bandwidth is required for overhead traffic, such as routing and Layer 2 keepalives. Figure 3-6 illustrates this bandwidth provisioning process.

Figure 3-6 Link Bandwidth Provisioning



In addition to using no more than 75 percent of the total available bandwidth for data, voice, and video, the total bandwidth configured for all LLQ priority queues should typically not exceed 33 percent of the total link bandwidth. Provisioning more than 33 percent of the available bandwidth for the priority queue can be problematic for a number of reasons. First, provisioning more than 33 percent of the bandwidth for voice can result in increased CPU usage. Because each voice call will send 50 packets per second (with 20 ms samples), provisioning for large numbers of calls in the priority queue can lead to high CPU levels due to high packet rates. In addition, if more than one type of traffic is provisioned in the priority queue (for example, voice and video), this configuration defeats the purpose of enabling QoS because the priority queue essentially becomes a first-in, first-out (FIFO) queue. A larger percentage of reserved priority bandwidth effectively dampens the QoS effects by making more of the link bandwidth FIFO. Finally, allocating more than 33 percent of the available bandwidth can effectively starve any data queues that are provisioned. Obviously, for very slow links (less than 192 kbps), the recommendation to provision no more than 33 percent of the link bandwidth for the priority queue(s) might be unrealistic because a single call could require more than 33 percent of the link bandwidth. In these situations, and in situations where specific business needs cannot be met while holding to this recommendation, it may be necessary to exceed the 33 percent rule.

From a traffic standpoint, an IP telephony call consists of two parts:

- The voice carrier stream, which consists of Real-Time Transport Protocol (RTP) packets that contain the actual voice samples.
- The call control signaling, which consists of packets belonging to one of several protocols, according to the endpoints involved in the call (for example, H.323, MGCP, SCCP, SIP, or TAPI). Call control functions are, for instance, those used to set up, maintain, tear down, or redirect a call.

Bandwidth provisioning should include not only the voice stream traffic but also the call control traffic. In fact, in multisite WAN deployments, the call control traffic (and also the voice stream) must traverse the WAN, and failure to allocate sufficient bandwidth for it can adversely affect the user experience.

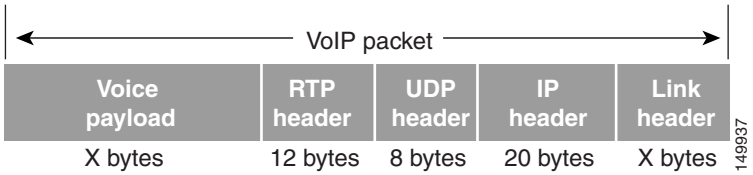
The next three sections describe the bandwidth provisioning recommendations for the following types of traffic:

- Voice bearer traffic in all multisite WAN deployments (see the "Provisioning for Voice Bearer Traffic" section on page 3-25)
- Call control traffic in multi-site WAN deployments with distributed call processing (see the "Provisioning for Call Control Traffic with Distributed Call Processing" section on page 3-26)

Provisioning for Voice Bearer Traffic

As illustrated in Figure 3-7, a voice-over-IP (VoIP) packet consists of the payload, IP header, User Datagram Protocol (UDP) header, Real-Time Transport Protocol (RTP) header, and Layer 2 Link header. At the default packetization rate of 20 ms, VoIP packets have a 160-byte payload for G.711 or a 20-byte payload for G.729. When Secure Real-Time Transport Protocol (SRTP) encryption is used, the payload for each packet is increased by 4 bytes. At the default packetization rate of 20 ms, SRTP VoIP packets have a 164-byte payload for G.711 or a 24-byte payload for G.729. The IP header is 20 bytes, the UDP header is 8 bytes, and the RTP header is 12 bytes. The link header varies in size according to the Layer 2 media used.

Figure 3-7 Typical VoIP Packet



The bandwidth consumed by VoIP streams is calculated by adding the packet payload and all headers (in bits), then multiplying by the packet rate per second (default of 50 packets per second). Table 3-4 details the bandwidth per VoIP flow at a default packet rate of 50 packets per second (pps). Table 3-4 does not include Layer 2 header overhead and does not take into account any possible compression schemes, such as compressed Real-Time Transport Protocol (cRTP).

Table 3-4 lists the bandwidth consumed by the voice payload and IP header only, at a default packet rate of 50 packets per second (pps) and at a rate of 33.3 pps for both unencrypted and encrypted payloads.

Table 3-4 Bandwidth Consumption for Voice Payload and IP Header Only

Codec	Sampling Rate	Voice Payload in Bytes	Packets per Second	Bandwidth per Conversation
G.711	20 ms	160	50.0	80.0 kbps
G.711 (SRTP)	20 ms	164	50.0	81.6 kbps
G.711	30 ms	240	33.3	74.7 kbps
G.711 (SRTP)	30 ms	244	33.3	75.8 kbps
G.729A	20 ms	20	50.0	24.0 kbps
G.729A (SRTP)	20 ms	24	50.0	25.6 kbps

Table 3-4 Bandwidth Consumption for Voice Payload and IP Header Only (continued)

Codec	Sampling Rate	Voice Payload in Bytes	Packets per Second	Bandwidth per Conversation
G.729A	30 ms	30	33.3	18.7 kbps
G.729A (SRTP)	30 ms	34	33.3	19.8 kbps

A more accurate method for provisioning is to include the Layer 2 headers in the bandwidth calculations. Table 3-5 lists the amount of bandwidth consumed by voice traffic when the Layer 2 headers are included in the calculations.

Table 3-5 Bandwidth Consumption with Layer 2 Headers Included

	Header Type and Size						
Codec	Ethernet 14 Bytes	PPP 6 Bytes	ATM 53-Byte Cells with a 48-Byte Payload	Frame Relay 4 Bytes	MLPPP 10 Bytes	MPLS 4 Bytes	WLAN 24 Bytes
G.711 at 50.0 pps	85.6 kbps	82.4 kbps	106.0 kbps	81.6 kbps	84.0 kbps	81.6 kbps	89.6 kbps
G.711 (SRTP) at 50.0 pps	87.2 kbps	84.0 kbps	106.0 kbps	83.2 kbps	85.6 kbps	83.2 kbps	N/A
G.729A at 50.0 pps	29.6 kbps	26.4 kbps	42.4 kbps	25.6 kbps	28.0 kbps	25.6 kbps	33.6 kbps
G.729A (SRTP) at 50.0 pps	31.2 kbps	28.0 kbps	42.4 kbps	27.2 kbps	29.6 kbps	27.2 kbps	N/A

Provisioning for Call Control Traffic with Distributed Call Processing

In distributed call processing deployments, several sites are connected through an IP WAN. Each site contains a Cisco Unified CME system and can follow either the single-site model or the centralized call processing model. A gatekeeper can be used for keeping dial-plans consistent and easily manageable between sites.

The following considerations apply to this deployment model:

- The signaling protocol used to place a call across the WAN is H.323 or SIP.
- Control traffic is exchanged between the Cisco IOS gatekeeper and the Cisco Unified CME systems at each site, and also between the Cisco Unified CME systems themselves.

Therefore, bandwidth for control traffic must be provisioned on the WAN links between Cisco Unified CME systems and between each Cisco Unified CME and the gatekeeper. Because the topology is limited to hub-and-spoke, with the gatekeeper typically located at the hub, the WAN link that connects each site to the other sites usually coincides with the link that connects the site to the gatekeeper.

The control traffic that traverses the WAN belongs to one of the following categories:

- Quiescent traffic, which consists of registration messages periodically exchanged between each Cisco Unified CME and the gatekeeper

- Call-related traffic, consisting of H.225 or H.245 signaling traffic, exchanged between two Cisco Unified CME systems when a call needs to be set up, torn down, forwarded, and so on

Because the total amount of control traffic depends on the number of calls that are set up and torn down at any given time, it is necessary to make some assumptions about the call patterns and the link utilization. The WAN links that connect each of the spoke sites to the hub site are normally provisioned to accommodate different types of traffic (for example, data, voice, and video). Using a traditional telephony analogy, we can view the portion of the WAN link that has been provisioned for voice as a number of *virtual tie lines*.

Assuming an average call duration of 2 minutes and 100 percent utilization of each virtual tie line, we can derive that each tie line carries a volume of 30 calls per hour. This assumption allows us to obtain the following formula that expresses the recommended bandwidth for call control traffic as a function of the number of virtual tie lines.

Recommended Bandwidth Based on Number of Virtual Tie Lines.

Recommended Bandwidth (bps) = 116 * (Number of virtual tie lines)

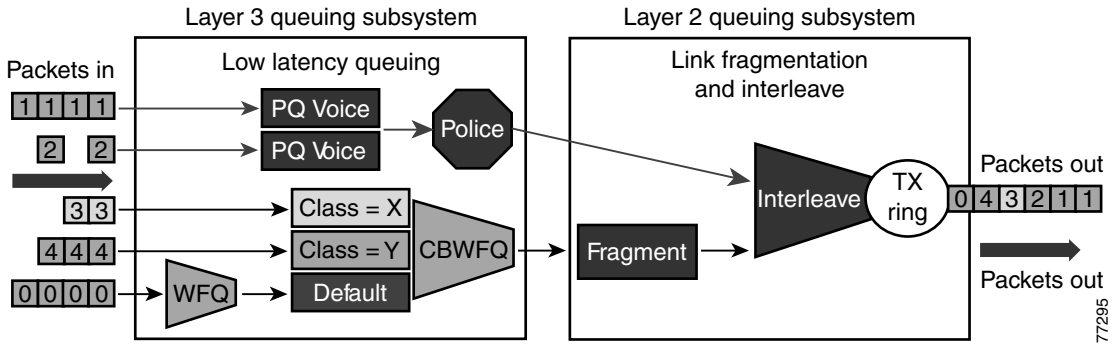
If we take into account the fact that 8 kbps is the smallest bandwidth that can be assigned to a queue on a Cisco IOS router, we can deduce that a minimum queue size of 8 kbps can accommodate the call control traffic generated by *up to 70 virtual tie lines*. This amount should be sufficient for most large enterprise deployments.

Traffic Prioritization

In choosing from among the many available prioritization schemes, the major factors to consider include the type of traffic involved and the type of media on the WAN. For multiservice traffic over an IP WAN, We recommend low-latency queuing (LLQ) for all links. This method supports up to 64 traffic classes, with the ability to specify, for example, priority queuing behavior for voice and interactive video, minimum bandwidth class-based weighted fair queuing for voice control traffic, additional minimum bandwidth weighted fair queues for mission critical data, and a default best-effort queue for all other traffic types.

Figure 3-8 shows an example prioritization scheme.

Figure 3-8 Optimized Queuing for VoIP over the WAN



We recommend the following prioritization criteria for LLQ:

- The criterion for *voice* to be placed into a priority queue is the differentiated services code point (DSCP) value of 46, or a per-hop behavior (PHB) value of EF.

- The criterion for *video conferencing* traffic to be placed into a priority queue is a DSCP value of 34, or a PHB value of AF41. However, due to the larger packet sizes of video traffic, these packets should be placed in the priority queue only on WAN links that are faster than 768 kbps. Link speeds below this value require packet fragmentation, but packets placed in the priority queue are not fragmented, thus smaller voice packets could be queued behind larger video packets. For links speeds of 768 kbps or lower, video conferencing traffic should be placed in a separate class-based weighted fair queue (CBWFQ).

Note One-way video traffic, such as the traffic generated by streaming video applications for services such as video-on-demand or live video feeds, should always use a CBWFQ scheme because that type of traffic has a much higher delay tolerance than two-way video conferencing traffic.

- As the WAN links become congested, it is possible to starve the *voice control* signaling protocols, thereby eliminating the ability of the IP phones to complete calls across the IP WAN. Therefore, voice control protocols, such as H.323, MGCP, and Skinny Client Control Protocol (SCCP), require their own class-based weighted fair queue. The entrance criterion for this queue is a DSCP value of 24 or a PHB value of CS3.

Note We have begun to change the marking of voice control protocols from DSCP 26 (PHB AF31) to DSCP 24 (PHB CS3). However many products still mark signaling traffic as DSCP 26 (PHB AF31); therefore, in the interim, we recommend that you reserve both AF31 and CS3 for call signaling.

- In some cases, certain data traffic might require better than best-effort treatment. This traffic is referred to as *mission-critical data*, and it is placed into one or more queues that have the required amount of bandwidth. The queuing scheme within this class is first-in-first-out (FIFO) with a minimum allocated bandwidth. Traffic in this class that exceeds the configured bandwidth limit is placed in the default queue. The entrance criterion for this queue could be a Transmission Control Protocol (TCP) port number, a Layer 3 address, or a DSCP/PHB value.
- All remaining traffic can be placed in a default queue for best-effort treatment. If you specify the keyword **fair**, the queuing algorithm will be weighted fair queuing (WFQ).

Link Efficiency Techniques

The following link efficiency techniques improve the quality and efficiency of low-speed WAN links.

Compressed Real-Time Transport Protocol (cRTP)

You can increase link efficiency by using Compressed Real-Time Transport Protocol (cRTP). This protocol compresses a 40-byte IP, User Datagram Protocol (UDP), and RTP header into approximately two to four bytes. cRTP operates on a per-hop basis. Use cRTP on a particular link only if that link meets *all* of the following conditions:

- Voice traffic represents more than 33 percent of the load on the specific link.
- The link uses a low bit-rate codec (such as G.729).
- No other real-time application (such as video conferencing) is using the same link.

If the link fails to meet any one of the preceding conditions, then cRTP is not effective and you should not use it on that link. Another important parameter to consider before using cRTP is router CPU utilization, which is adversely affected by compression and decompression operations.

cRTP on ATM and Frame Relay Service Inter-Working (SIW) links requires the use of Multilink Point-to-Point Protocol (MLP).

Note that cRTP compression occurs as the final step before a packet leaves the egress interface; that is, after LLQ class-based queueing has occurred. Beginning in Cisco IOS Release 12.(2)2T and later, cRTP provides a feedback mechanism to the LLQ class-based queueing mechanism that allows the bandwidth in the *voice* class to be configured based on the compressed packet value. With Cisco IOS software releases earlier than 12.(2)2T, this mechanism is not in place, so the LLQ is unaware of the compressed bandwidth and, therefore, the *voice* class bandwidth has to be provisioned as if no compression is taking place. Table 3-6 shows an example of the difference in *voice* class bandwidth configuration given a 512-kbps link with G.729 codec and a requirement for 10 calls.

Note that Table 3-6 assumes 24 kbps for non-cRTP G.729 calls and 10 kbps for cRTP G.729 calls. These bandwidth numbers are based on voice payload and IP/UDP/RTP headers only. They do not take into consideration Layer 2 header bandwidth. However, actual bandwidth provisioning should also include Layer 2 header bandwidth based on the type WAN link used.

Table 3-6 LLQ Voice Class Bandwidth Requirements for 10 Calls with 512 kbps Link Bandwidth and G.729 Codec

Cisco IOS Release	With cRTP Not Configured	With cRTP Configured
Before 12.2(2)T	240 kbps	240 kbps[1]
12.2(2)T or later releases	240 kbps	100 kbps

1. 140 kbps of unnecessary bandwidth must be configured in the LLQ *voice* class.

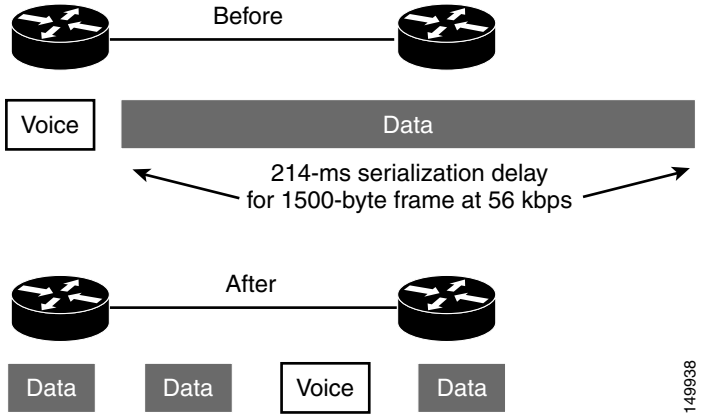
It should also be noted that, beginning in Cisco IOS Release 12.2(13)T, cRTP can be configured as part of the voice class with the Class-Based cRTP feature. This option allows cRTP to be specified within a class, attached to an interface via a service policy. This new feature provides compression statistics and bandwidth status via the **show policy interface** command, which can be very helpful in determining the offered rate on an interface service policy class given the fact that cRTP is compressing the IP/RTP headers.

Note For additional recommendations about using cRTP with a Voice and Video Enabled IP Sec VPN (V3PN), see the V3PN documentation available at http://www.cisco.com/application/pdf/en/us/guest/netsol/ns171/c649/ccmigration_09186a008074f2cb.pdf

Link Fragmentation and Interleaving (LFI)

For low-speed links (less than 768 kbps), use of link fragmentation and interleaving (LFI) mechanisms is required for acceptable voice quality. This technique limits jitter by preventing voice traffic from being delayed behind large data frames, as illustrated in Figure 3-9. The two techniques that exist for this purpose are Multilink Point-to-Point Protocol (MLP) LFI (for Leased Lines, ATM, and SIW) and FRF.12 for Frame Relay.

Figure 3-9 Link Fragmentation and Interleaving (LFI)



Voice-Adaptive Fragmentation (VAF)

In addition to the LFI mechanisms mentioned above, voice-adaptive fragmentation (VAF) is another LFI mechanism for Frame Relay links. VAF uses FRF.12 Frame Relay LFI; however, once configured, fragmentation occurs only when traffic is present in the LLQ priority queue or when H.323 signaling packets are detected on the interface. This method ensures that, when voice traffic is being sent on the WAN interface, large packets are fragmented and interleaved. However, when voice traffic is not present on the WAN link, traffic is forwarded across the link unfragmented, thus reducing the overhead required for fragmentation.

VAF is typically used in combination with voice-adaptive traffic shaping (se0 the "Voice-Adaptive Traffic Shaping" section on page 3-32). VAF is an optional LFI tool, and you should exercise care when enabling it because there is a slight delay between the time when voice activity is detected and the time when the LFI mechanism engages. In addition, a configurable deactivation timer (default of 30 seconds) must expire after the last voice packet is detected and before VAF is deactivated, so during that time LFI will occur unnecessarily. VAF is available in Cisco IOS Release 12.2(15)T and later releases.

Traffic Shaping

Traffic shaping is required for multiple-access, non-broadcast media such as ATM and Frame Relay, where the physical access speed varies between two endpoints and several branch sites are typically aggregated to a single router interface at the central site.

Figure 3-10 illustrates the main reasons why traffic shaping is needed when transporting voice and data on the same IP WAN.

Figure 3-10 Traffic Shaping with Frame Relay and ATM

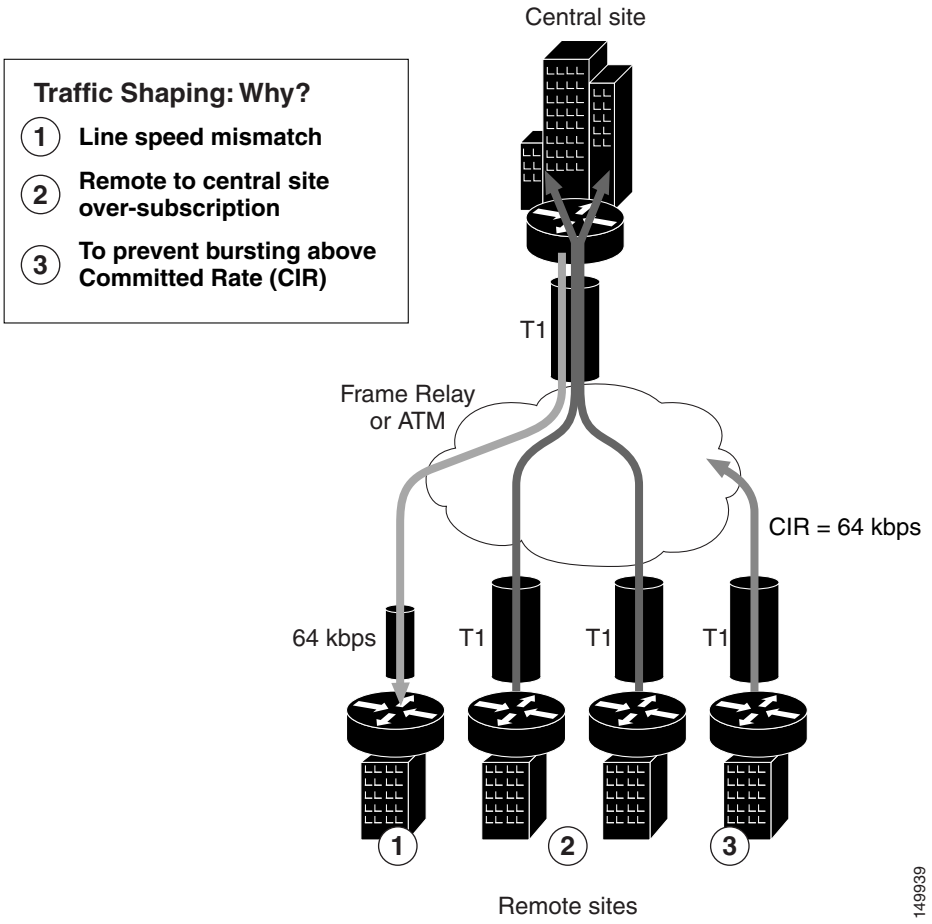


Figure 3-10 shows three different scenarios:

1. Line speed mismatch

   While the central-site interface is typically a high-speed one (such as T1 or higher), smaller remote branch interfaces may have significantly lower line speeds, such as 64 kbps. If data is sent at full rate from the central site to a slow-speed remote site, the interface at the remote site might become congested and degrade voice performance.

2. Oversubscription of the link between the central site and the remote sites

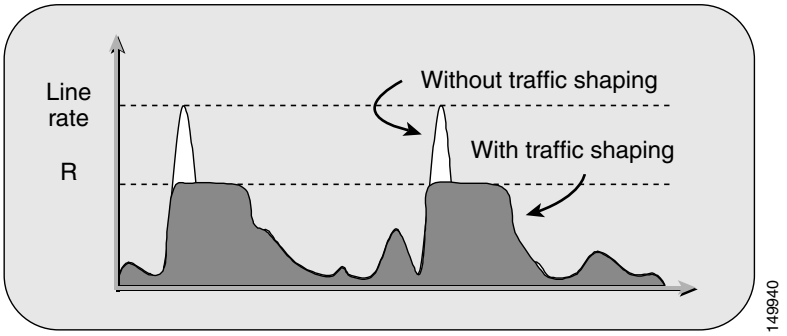
   It is common practice in Frame Relay or ATM networks to oversubscribe bandwidth when aggregating many remote sites to a single central site. For example, there may be multiple remote sites that connect to the WAN with a T1 interface, yet the central site has only a single T1 interface. While this configuration allows the deployment to benefit from statistical multiplexing, the router interface at the central site can become congested during traffic bursts, thus degrading voice quality.

3. Bursting above Committed Information Rate (CIR)

   Another common configuration is to allow traffic bursts above the CIR, which represents the rate that the service provider has guaranteed to transport across its network with no loss and low delay. For example, a remote site with a T1 interface might have a CIR of only 64 kbps. When more than 64 kbps worth of traffic is sent across the WAN, the provider marks the additional traffic as "discard eligible." If congestion occurs in the provider network, this traffic will be dropped with no regard to traffic classification, possibly having a negative affect on voice quality.

Traffic shaping provides a solution to these issues by limiting the traffic sent out an interface to a rate lower than the line rate, thus ensuring that no congestion occurs on either end of the WAN. Figure 3-11 illustrates this mechanism with a generic example, where R is the rate with traffic shaping applied.

Figure 3-11 Traffic Shaping Mechanism



Voice-Adaptive Traffic Shaping

Voice-adaptive traffic shaping (VATS) is an optional dynamic mechanism that shapes traffic on Frame Relay permanent virtual circuits (PVCs) at different rates based on whether voice is being sent across the WAN. The presence of traffic in the LLQ voice priority queue or the detection of H.323 signaling on the link causes VATS to engage. Typically, Frame Relay shapes traffic to the guaranteed bandwidth or CIR of the PVC at all times. However, because these PVCs are typically allowed to burst above the CIR (up to line speed), traffic shaping keeps traffic from using the additional bandwidth that might be present in the WAN. With VATS enabled on Frame Relay PVCs, WAN interfaces are able to send at CIR when voice traffic is present on the link. However, when voice is not present, non-voice traffic is able to burst up to line speed and take advantage of the additional bandwidth that might be present in the WAN.

When VATS is used in combination with voice-adaptive fragmentation (VAF) (see the "Link Fragmentation and Interleaving (LFI)" section on page 3-29), all non-voice traffic is fragmented and all traffic is shaped to the CIR of the WAN link when voice activity is detected on the interface.

As with VAF, exercise care when enabling VATS because activation can have an adverse effect on non-voice traffic. When voice is present on the link, data applications will experience decreased throughput because they are throttled back to below CIR. This behavior will likely result in packet drops and delays for non-voice traffic. Furthermore, after voice traffic is no longer detected, the deactivation timer (default of 30 seconds) must expire before traffic can burst back to line speed. It is important, when

using VATS, to set end-user expectations and make them aware that data applications will experience slowdowns on a regular basis due to the presence of voice calls across the WAN. VATS is available in Cisco IOS Release 12.2(15)T and later.

For more information on the voice-adaptive traffic shaping and fragmentation features and how to configure them, see documentation at:
http://www.cisco.com/en/US/docs/ios/12_2t/12_2t15/feature/guide/ft_vats.html

Wireless LAN Infrastructure

Wireless LAN infrastructure design becomes important when IP telephony is added to the wireless LAN (WLAN) portions of a converged network. With the addition of wireless IP telephony endpoints such as the Cisco Wireless IP Phone 7920, voice traffic has moved onto the WLAN and is now converged with the existing data traffic there. Just as with wired LAN and wired WAN infrastructure, the addition of voice in the WLAN requires following basic configuration and design best-practices for deploying a highly available network. In addition, proper WLAN infrastructure design requires understanding and deploying QoS on the wireless network to ensure end-to-end voice quality on the entire network. The following sections discuss these requirements:

- WLAN Design and Configuration, page 3-33
- WLAN Quality of Service (QoS), page 3-40

Note For more information about the Cisco Unified Wireless IP Phone 7920, see the following URL:
http://www.cisco.com/en/US/products/hw/phones/ps379/ps5056/index.html

WLAN Design and Configuration

Properly designing a WLAN requires, first and foremost, ensuring that the existing wired network is deployed in a highly available, fault-tolerant and redundant manner. Next, an understanding of wireless technology is required. Finally, by configuring and deploying wireless APs and wireless telephony endpoints in an effective way, you can build a flexible, secure, redundant, and highly scalable network.

The following sections examine the WLAN infrastructure layers and network services:

- Wireless Infrastructure Considerations, page 3-33
- Wireless AP Configuration and Design, page 3-36
- Wireless Security, page 3-38

Wireless Infrastructure Considerations

The following sections provide guidelines and best practices for designing the WLAN infrastructure:

- VLANs, page 3-34
- Roaming, page 3-34
- Wireless Channels, page 3-34
- Wireless Interference, page 3-36
- Multicast on the WLAN, page 3-36

VLANs

Just as with a wired LAN infrastructure, when deploying voice in a wireless LAN, you should enable at least two virtual LANs (VLANs) at the Access Layer. The Access Layer in a wireless LAN environment includes the access point (AP) and the first-hop access switch. On the AP and access switch, you should configure both a native VLAN for data traffic and a voice VLAN (under Cisco IOS software) or Auxiliary VLAN (under Catalyst Operating System) for voice traffic. This voice/auxiliary VLAN must be separate from all the other wired voice VLANs in the network. In addition, as with voice endpoints on wired LANs, wireless voice endpoints should be addressed using RFC 1918 private subnet addresses. When deploying a wireless infrastructure, we recommend configuring a separate management VLAN for the management of WLAN APs. This management VLAN should not have a WLAN appearance; that is, it should not have an associated service set identifier (SSID) and it should not be directly accessible from the WLAN.

Roaming

Another very important consideration for wireless infrastructure is wireless endpoint roaming. When wireless devices roam at Layer 2, they keep their IP address and network configuration. For this reason, roaming can occur fairly quickly (in 100 to 400 ms). All that is required is re-authentication, if Cisco LEAP or Extensible Authentication Protocol (EAP) is used, and the passing of Inter-Access Point Protocol (IAPP) messages between the last AP and the new AP to indicate that the endpoint has roamed. Layer 2 roaming is typically unnoticeable to the end user.

When devices roam at Layer 3, they move from one AP to another AP and cross a subnet boundary. With the release of the new Cisco Catalyst 6500 Series Wireless LAN Services Module (WLSM), the Cisco Unified Wireless IP Phone 7920 now supports call-survivable Layer 3 roaming while using Static WEP. Cisco Centralized Key Management (Cisco CKM) enables the Cisco 7920 phone to achieve full Layer 3 mobility while using LEAP. For details about the Cisco WLSM, see the product documentation available at:
http://www.cisco.com/en/US/products/hw/modules/ps2706/products_implementation_design_guide09186a00807d592c.html

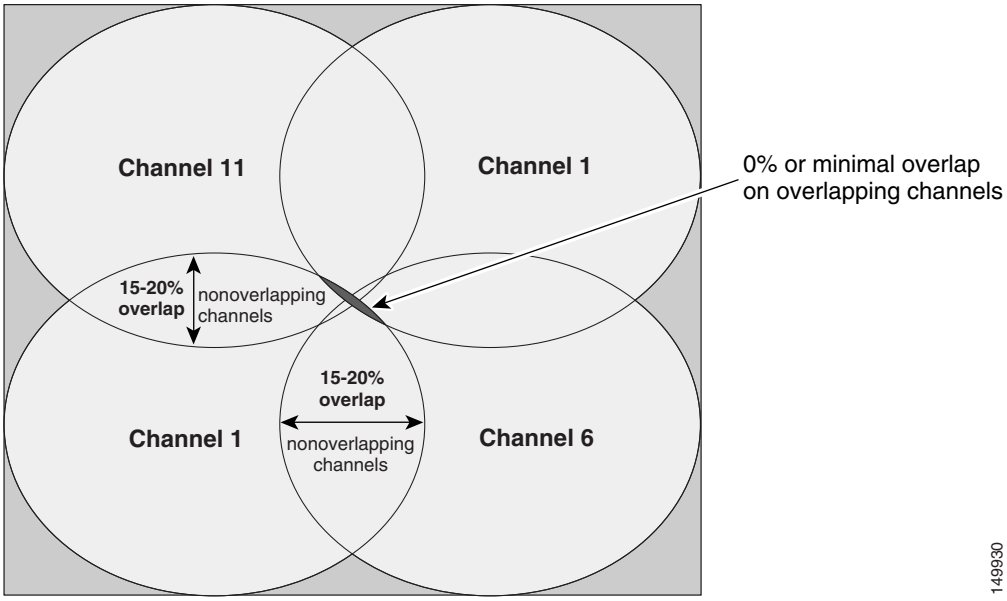
Note If Cisco Catalyst 4000 Series switches are used as Layer 3 devices at the distribution layer, a minimum of a Supervisor Engine 2+ (SUP2+) or Supervisor Engine 3 (SUP3) module is required. The Supervisor Engine 1 or 2 (SUP1 or SUP2) modules can cause roaming delays. The Cisco Catalyst 2948G, 2948G-GE-TX, 2980G, 2980G-A, and 4912 switches are also known to introduce roaming delays. we do not recommend using these switches in a wireless voice network.

Wireless Channels

Wireless endpoints and APs communicate via radios on particular channels. When communicating on one channel, wireless endpoints typically are unaware of traffic and communication occurring on other nonoverlapping channels. Optimal channel configuration for 2.4 GHz 802.11b requires a five-channel spread to prevent interference or overlap between channels. In North America, channels 1, 6, and 11 are the three usable nonoverlapping channels for APs and wireless endpoint devices. In the European Union, the nonoverlapping usable channels for 802.11b are 1, 6, and 11 or 12 or 13. In Japan these channels are 1, 6, and 11 or 12, 13, or 14.

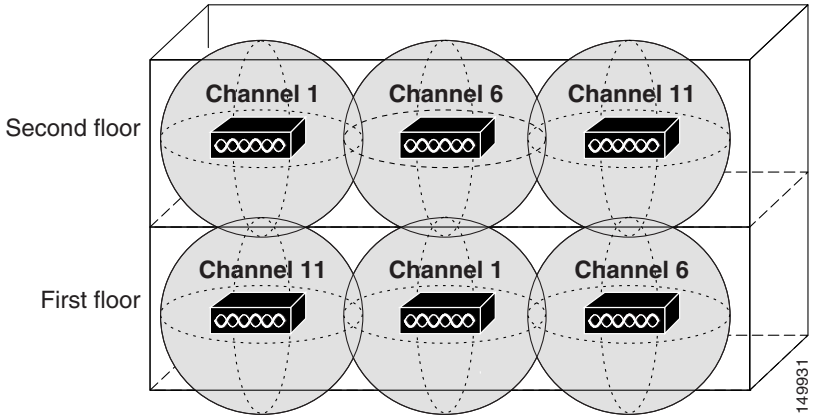
AP coverage should be deployed so that minimal or no overlap occurs between APs configured with the same channel (for example, see Channel 1 in Figure 3-12). However, proper AP deployment and coverage on *nonoverlapping* channels (1, 6, and 11 in North America) require an overlap of 15 to 20 percent. This amount of overlap ensures smooth roaming for wireless endpoints as they move between AP coverage cells. Overlap of less than 15 to 20 percent can result in slower roaming times and poor voice quality; while overlap of more than 15 to 20 percent can result in too frequent or constant roaming. Figure 3-12 illustrates appropriate AP overlap for both overlapping and nonoverlapping channels.

Figure 3-12 Wireless 802.11b Channel Overlap



Deploying wireless devices in a multistory building such as an office high-rise or hospital introduces a third dimension to wireless AP and channel coverage planning. The 2.4 GHz wave form of 802.11b can pass through floors and ceilings and walls. For this reason, not only is it important to consider overlapping cells or channels on the same floor, but it is also necessary to consider channel overlap between adjacent floors. With only three channels, proper overlap can be achieved only through careful three-dimensional planning. Figure 3-13 shows the potential for channel overlap when considering the three-dimensional aspects of 802.11b wireless coverage.

Figure 3-13 Wireless 802.11b Channel Overlap Considerations (in Three Dimensions)



Note Careful deployment of APs and channel configuration within the wireless infrastructure are imperative for proper wireless network operation. For this reason, we require that a complete and thorough site survey be conducted before deploying wireless networks in a production environment. The survey should include verifying nonoverlapping channel configurations, AP coverage, and required data and traffic rates; eliminating rogue APs; and identifying and mitigating the impact of potential interference sources.

Wireless Interference

Interference sources within a wireless environment can severely limit endpoint connectivity and channel coverage. In addition, objects and obstructions can cause signal reflection and multipath distortion. Multipath distortion occurs when traffic or signaling travels in more than one direction from the source to the destination. Typically, some of the traffic arrives at the destination before the rest of the traffic, which can result in delay and bit errors in some cases. You can reduce the affects of multipath distortion by eliminating or reducing interference sources and obstructions, and by using diversity antennas so that only a single antenna is receiving traffic at any one time. Interference sources should be identified during the site survey and, if possible, eliminated. At the very least, interference impact should be alleviated by proper AP placement and the use of location-appropriate directional or omni-directional diversity radio antennas.

Possible interference sources include:

- Other APs on overlapping channels
- Other 2.4 GHz appliances, such as 2.4 GHz cordless phones, personal wireless network devices, sulphur plasma lighting systems, microwave ovens, rogue APs, and other WLAN equipment that takes advantage of the license-free operation of the 2.4 GHz band
- Metal equipment, structures, and other metal or reflective surfaces such as metal I-beams, filing cabinets, equipment racks, wire mesh or metallic walls, fire doors and fire walls, concrete, and heating and air conditioning ducts
- High-power electrical devices such as transformers, heavy-duty electric motors, refrigerators, elevators, and elevator equipment

Multicast on the WLAN

We do not recommend forwarding multicast traffic on a WLAN with voice devices because of the following reasons:

- Multicast packets are buffered on the AP if devices associated with the AP are in power-save mode.

  When devices such as the Cisco Wireless IP Phone 7920 go into power-save mode, an AP buffers all multicast packets until the next active interval for these devices. This buffering causes packet delay and affects all devices associated with the AP even if they are not in power-save mode. This condition can be extremely problematic for real-time multicast applications such as music on hold and streaming video.

- Multicast packets on the WLAN are unacknowledged and are not retransmitted if lost or corrupted.

  AP and wireless endpoint devices use acknowledgements on the link layer to ensure reliable delivery. When packets are not received or acknowledged, they are retransmitted. This retransmission does not occur for multicast traffic on the WLAN. Given that wireless networks have a higher instance of bit error than wired networks, this lack of retransmission results in a larger number of lost packets when compared to a wired LAN.

Before enabling multicast applications on the wireless network, we recommend testing these applications to ensure that performance and behavior are acceptable.

Wireless AP Configuration and Design

Proper AP selection, deployment, and configuration are essential to ensure that the wireless network handles voice traffic in a way that provides high-quality voice to the end users.

AP Selection

Cisco recommend the following APs for deploying wireless voice:

- Cisco Aironet 350 Series AP
- Cisco Aironet 1100 Series AP
- Cisco Aironet 1200 Series AP

For these APs, Cisco IOS Release 12.2(13)JA3 or later releases should be used. We do *not* recommend the VxWorks operating system for APs when deploying wireless voice because new features are not being added to VxWorks, but some of those new features are required for voice deployments.

The Cisco Integrated Services Routers (ISRs)—including the Cisco 800, Cisco 1800, Cisco 2800, and Cisco 3800 series routers—also support access point functionality. For the fixed Cisco 800s and fixed Cisco 1800s (with a built-in access point), use Cisco IOS Release 12.3(8)YI or later. For the modular Cisco 1841, Cisco 28xx, and Cisco 38xx routers, use a HWIC-AP to provide access point functionality along with Cisco IOS Release 12.4(2)T or later releases. Features supported on the wireless ISR APs are similar to the Aironet APs, with a few exceptions. The following links documents features supported on the wireless ISR APs:

- Cisco 870

  http://www.cisco.com/en/US/prod/collateral/routers/ps380/ps6200/product_data_sheet0900aecd8028a976.html

- Cisco 1800 (fixed hardware configuration)

  http://www.cisco.com/en/US/prod/collateral/routers/ps5853/ps6184/product_data_sheet0900aecd8028a95f_ps5853_Products_Data_Sheet.html

- HWIC-AP

  http://www.cisco.com/en/US/prod/collateral/modules/ps5949/ps6246/product_data_sheet0900aecd8028cc7b.html

- Configuration Guide

  http://www.cisco.com/en/US/docs/ios/wlan/configuration/guide/12_4t/wl_12_4t_book.html

AP Deployment

When deploying Cisco APs, do not associate more than 15 to 25 devices to a single AP at any time. This number will vary depending on usage profiles. The number of devices on an AP affects the amount of time each device has access to the medium. As the number of devices increases, the traffic contention increases. Associating more than 15 to 25 devices to an AP can result in poor AP performance and slower response times for associated devices.

AP Configuration

When deploying wireless voice, observe the following specific AP configuration requirements:

- Enable Address Resolution Protocol (ARP) caching.

  ARP caching is required on the AP because it enables the AP to answer ARP requests for the wireless endpoint devices without requiring the endpoint to leave power-save or idle mode. This feature results in extended battery life for the wireless endpoint devices.

- Match the transmit power on the AP to that on the wireless voice endpoints.

  When possible, the transmit power on the AP and the voice endpoints should match. Matching transmit power on the AP and voice endpoints helps eliminate the possibility of one-way audio traffic. If the transmit power configuration on the APs must vary, the transmit power on all voice endpoints should be configured to match the AP with the highest transmit power.

Note Beginning with version 1.0(8) of the Cisco Unified Wireless IP Phone 7920 firmware, the phone will take advantage of the Dynamic Transmit Power Control (DTPC) feature by automatically adjusting its transmit power based on the Limit Client Power (mW) setting of the current AP.

- Set the data rate to 11 Mbps.

  Configuring the maximum 11 Mbps data rate ensures the best level of throughput for voice devices and the largest number of active calls per AP.

- Manually configure the RF channel selection. (Do *not* use the option **Search for Least Congested Channel**).

  To control wireless network channels and eliminate channel overlap, it is important to configure a channel number manually on each AP based on its location.

- Assign a Service Set Identifier (SSID) to each VLAN configured on the AP.

  SSIDs enable endpoints to select the wireless VLAN they will use for sending and receiving traffic. These wireless VLANs and SSIDs map to wired VLANs. For voice endpoints, this mapping ensures priority queuing treatment and access to the voice VLAN on the wired network.

- Enable **QoS Element for Wireless Phones** on the AP.

  This feature ensures that the AP will provide QoS Basic Service Set (QBSS) information elements in beacons. The QBSS element provides an estimate of the channel utilization on the AP, and Cisco wireless voice devices use it to help make roaming decisions and to reject call attempts when loads are too high.

- Configure two QoS policies on the AP, and apply them to the VLANs and interfaces.

  Configure a voice policy and a data policy with default classifications for the respective VLANs to ensure that voice traffic is given priority queuing treatment. (See the "Interface Queuing" section on page 3-40 for more information).

Wireless Security

Another important consideration for a wireless infrastructure is security. Wireless endpoints, including wireless phones, can connect to the wireless network using one of the following security mechanisms:

- Cisco LEAP

  Cisco LEAP requires the wireless endpoint to provide a user name and password to authenticate with the network. Once this authentication occurs, a dynamic key is generated, and traffic to and from the wireless device is encrypted. This method requires an EAP-compliant Remote Authentication Dial-In User Service (RADIUS) authentication server such as the Cisco Secure Access Control Server (ACS), which provides access to a user database for authenticating the wireless devices. Cisco LEAP is the recommend security mechanism for use with wireless voice because it requires the highest level of security for access to voice VLANs.

- Static Wire Equivalent Privacy (WEP)

  Static WEP requires the exchange of a statically configured 40-bit or 128-bit character key between the wireless endpoint and the AP. If the keys match, the wireless device is given access to the network. Be aware that there are known weaknesses in the WEP encryption algorithm. These weaknesses, coupled with the complexity of configuring and maintaining static keys, can make this security mechanism undesirable in many cases.

- Open authentication

  This method requires no authentication and provides no security for traffic traveling between the wireless endpoint device and the wireless network. This method requires only that the wireless device be configured with the proper SSID for the wireless VLAN to which the device wishes to connect. We do not typically recommend this method for wireless voice because it requires no authentication for access to voice VLANs and provides no encryption for voice traffic.

Cisco LEAP Authentication and ACS Deployment Models

As indicated previously, Cisco LEAP is the preferred method of wireless device authentication (especially voice devices) because it provides the most secure and robust mechanism for access to the network and voice VLAN(s). Because an EAP-compliant RADIUS server is required, we recommend the use of Cisco Secure ACS for Windows Server Version 3.1 or later releases.

When deploying Cisco LEAP for wireless authentication and encryption, carefully consider the placement of the ACS within the network, and select one of the following ACS deployment models:

- Centralized ACS

  ACS server or servers are located in a centralized place within the network and are used to authenticate all wireless devices and users within the network.

- Remote ACS

  In networks where remote locations are separated from the central site by low-speed or congested WAN links, an ACS server can be located at the remote site and remote wireless devices or users can be authenticated by this server locally, thus eliminating the potential for delayed authentication via a centralized ACS across the WAN link.

- Local and fallback RADIUS server on the Cisco AP

  In networks where remote locations are separated from a central site by low-speed WAN links, local wireless devices can authenticate against local Cisco IOS APs. APs running Cisco IOS Release 12.2(11)JA or later releases can authenticate Cisco LEAP users and devices locally without relying on an external ACS. A single AP can support up to 50 users with this functionality.

  The Cisco integrated services routers (ISR) also support local authentication via LEAP. The number of users supported depends on the platform.

  This feature can be used in lieu of a centralized or local ACS, or in the case of a WAN or ACS failure in which the remote site users are unable to contact a local ACS or the central site ACS.

When choosing a deployment model for the ACS, it is imperative to make authentication services redundant so that the ACS does not become a single point of failure when wireless devices attempt to access the network. For this reason, each ACS server should replicate its database to a secondary server. Furthermore, it is always a good idea to provide a local ACS or an on-AP RADIUS server at remote sites to ensure that remote wireless devices can still authenticate in the event of a WAN failure.

In addition to ACS server placement, it is also important to consider the implications of user database location in relation to the ACS server. Because the ACS server must access the user database to authenticate wireless devices, the location of the user database affects the amount of time the authentication will take. If the user database is a Microsoft Active Directory (AD) server located on the network, the ACS must send an authentication request to the AD server and wait for a response. To ensure the fastest response times for wireless voice endpoints attempting to authenticate to the network, we recommend defining users locally on the ACS server. Remote databases have unknown response times and can adversely affect authentication times.

WLAN Quality of Service (QoS)

Just as QoS is necessary for LAN and WAN wired network infrastructure in order to ensure high voice quality, QoS is also require for wireless LAN infrastructure. Because of the bursty nature of data traffic and the fact that real-time traffic such as voice is sensitive to packet loss and delay, QoS tools are required to manage wireless LAN buffers, limit radio contention, and minimize packet loss, delay, and delay variation.

However, unlike most wired networks, wireless networks are a shared medium, and wireless endpoints do not have dedicated bandwidth for sending and receiving traffic. While wireless endpoints can mark traffic with 802.1p CoS, DSCP, and PHB, the shared nature of the wireless network means limited admission control and access to the network for these endpoints.

Wireless QoS involves the following main areas of configuration:

- Traffic Classification, page 3-40
- Interface Queuing, page 3-40
- Bandwidth Provisioning, page 3-41

Traffic Classification

As with wired network infrastructure, it is important to classify or mark pertinent wireless traffic as close to the edge of the network as possible. Because traffic marking is an entrance criterion for queuing schemes throughout the wired and wireless network, marking should be done at the wireless endpoint device whenever possible. Marking or classification by wireless network devices should be identical to that for wired network devices, as indicated in Table 3-2.

In accordance with traffic classification guidelines for wired networks, the Cisco Unified Wireless IP Phone 7920 marks voice media traffic or RTP traffic with DSCP 46 (or PHB EF) and voice signaling traffic (SCCP) with DSCP 26 (or PHB AF31). Once this traffic is marked, it can be given priority or better than best-effort treatment and queuing throughout the network. All wireless voice devices should be capable of marking traffic in this manner. All other traffic on the wireless network should be marked as best-effort or with some intermediary classification as outlined in wired network marking guidelines.

Interface Queuing

Once marking has occurred, it is necessary to enable the wired network APs and devices to provide QoS queuing so that voice traffic types are given separate queues to reduce the chances of this traffic being dropped or delayed as it traversed the wireless LAN. Queuing on the wireless network occurs in two directions, upstream and downstream. Upstream queuing concerns traffic traveling from the wireless endpoint up to the AP and from the AP up to the wired network. Downstream queuing concerns traffic traveling from the wired network to the AP and down to the wireless endpoint.

Unfortunately, there is little upstream queuing available in a wireless network. While wireless devices such as the Cisco Unified Wireless IP Phone 7920 can provide queuing upstream as the packets leave the device, there is no mechanism in place to provide queuing among all clients on the wireless LAN because wireless networks are a shared medium. Therefore, although voice media packets might receive priority treatment leaving the wireless endpoint, these packets must contend with all the other packets that other wireless devices may be attempting to send. For this reason, it is extremely important to follow the guideline of no more than 15 to 25 wireless clients per AP. Going beyond the upper limit of this guideline can result in additional voice packet delay and jitter.

As for downstream QoS, Cisco APs currently provide up to eight queues for downstream traffic being sent to wireless clients. The entrance criterion for these queues can be based on a number of factors including DSCP, access control lists (ACLs), and VLAN. Although eight queues are available, we recommend using only two queues when deploying wireless voice. All voice media and signaling traffic should be placed in the highest-priority queue, and all other traffic should be placed in the best-effort queue. This ensure the best possible queuing treatment for voice traffic.

To set up this two-queue configuration, create two QoS policies on the AP. Name one policy **voice** and configure it with the class of service **Voice <10 ms Latency (6)** as the **Default Classification for all packets on the Vlan**. Name the other policy **data** and configure it with the class of service **Best Effort (0)** as the **Default Classification for all packets on the Vlan**. Then assign the **data** policy to the incoming and outgoing radio interface for the data VLAN(s), and assign the **voice** policy to the incoming and outgoing radio interfaces for the voice VLAN(s). With the QoS policies applied at the VLAN level, the AP is not forced to examine every packet coming in or going out to determine the type of queuing it should receive. This configuration ensures that all voice media and signaling are given priority queuing treatment in a downstream direction.

Bandwidth Provisioning

Another QoS requirement for wireless networking is the appropriate provisioning of bandwidth. Bandwidth provisioning involves the bandwidth between the wired and wireless networks as well as the number of simultaneous voice calls that an AP can handle. Wireless APs typically connect to the wired network via a 100 Mbps link to an Access Layer switch port. While the ingress Ethernet port on the AP can receive traffic at 100 Mbps, the maximum throughput on an 802.11b wireless network is 11 Mbps. After taking into account the half-duplex nature of the wireless medium and the overhead of wireless headers, the practical throughput on the 802.11b wireless network is about 7 Mbps. This mismatch in throughput between the wired and wireless network can result in packet drops when traffic bursts occur in the network.

Rather than allowing traffic bursts to send excessive traffic toward the AP only to have it dropped by the AP, it is a good idea to rate-limit or police this traffic to a rate that the wireless network can handle. Forcing the AP to drop excessive traffic causes increased CPU utilization and congestion at the AP. Instead, limiting the traffic rate to 7 Mbps on the link between the wired Access Layer switch and the wireless AP ensures that traffic is dropped at the Access Layer switch, thus removing the burden from the AP. Note that, depending on the wireless network deployment, the practical throughput might be less than 7 Mbps, especially if more than the recommended number of devices are associated to a single AP.

Based on wireless voice network testing, we determined that a single wireless AP can support up to seven (7) G.711 voice calls or eight (8) G.729 voice calls. If these limits are exceeded, voice quality will suffer and voice calls might be dropped. While there is no true call admission control mechanism or method for provisioning wireless bandwidth for voice traffic, Cisco 7920 Wireless IP Phones can provide a simplified version of call admission control or bandwidth provisioning based on channel utilization information received from APs on the network. This information can be sent by the AP to the phone via a beacon that includes the QoS Basic Service Set (QBSS). The QBSS provides an estimation of the percentage of time the RF channel is in use by that AP. The higher the QBSS element value, the higher the channel utilization and the less likely the channel and AP can provide sufficient bandwidth for additional wireless voice devices. If the QBSS element value is 45 or higher, then any calls attempted by the wireless IP phone will be rejected with a “Network Busy” message and/or fast busy tone. In addition, the wireless IP phone considers the QBSS element in its roaming algorithm and will not roam to an AP that is sending beacons with a QBSS element of 45 or higher.

Note The QBSS value is simply an estimation of the channel utilization for a particular AP. The real channel utilization can be much higher than indicated. For this reason, the wireless voice device might still be able to place a voice call on an AP that has already reached the limit of 7 or 8 calls, thus still resulting in dropped calls or poor voice quality.

The QBSS information element is sent by the AP only if **QoS Element for Wireless Phones** has been enable on the AP. (See the "Wireless AP Configuration and Design" section on page 3-36.)