Migrating to Multiservice Networks— A Planning Primer

Multiservice Networking

Multiservice networking is the integration of several networks—data, voice, and video—into a single packet or cell-based infrastructure. Multiservice networking delivers the infrastructure needed to deploy new Web-based applications that represent a fundamental shift in how a business stays competitive and improves how a company interacts with customers, suppliers, employees, and industry partners. These applications include unified messaging, collaborative data sharing, video streaming and video conferencing, push-to-talk Web pages, network-enabled call centers and more.

Infrastructure convergence is not just about replacing existing telephone systems. Today's telephone systems work well but the transport of voice traffic can be much more efficient, and newer technologies exist for integrated multimedia and Web-based voice and video applications. Expanding the capabilities of a network to be ready for the new Web-based multimedia applications is part of the reason why companies are making the transition to multiservice networking.

Migrating from yesterday's multiple networks into multiservice networks usually happens in phases. Deployment is a decision based on the demands of a particular industry, market, and business requirements. For example, many companies only need to make minor changes to existing data network installations to enable voice integration. This may include adding multiservice-capable switches and routers to a world-class data network built out over the last few years, or may simply mean turning on the video and voice capabilities of an existing Cisco multiservice-capable router. Planning ahead can ease the transition on budgets and staff, as well as speed deployment time. When planning new workgroups or installations, such as new buildings or remote offices, a company can cost-justify the immediate deployment of a fully integrated multiservice network.

Multiservice network integration happens at two levels, and are often phased sequentially during planning and deployment: first the infrastructure, then the applications. Readying the network infrastructure for real-time traffic such as voice is a necessary step for deploying multimedia applications successfully. The infrastructure stage is often referred to as Voice IP Transport, or toll bypass, meaning the existing Key Systems and Private Branch Exchanges (PBXs) providing voice service to end users remain in place and the converged network is leveraged to transport voice and data traffic between the PBXs or the Public Switched Telephone Network (PSTN). The second stage, also known as IP Telephony, or Desktop Telephony, involves IP Phones, voice-capable computer applications and Web-based multimedia applications that integrate voice and data to the desktop. Even if both stages are deployed simultaneously, the infrastructure piece is a key part of the planning process to ensure successful applications deployment.

Multiservice Migration

This planning guide reviews the decisions to consider when evolving data-only networks toward a more robust multiservice infrastructure that can include data, voice, and video traffic. There are many ways to approach network migration, and this white paper outlines the recommended steps for planning a multiservice infrastructure:

1. Network audit
2. Network objectives
3. Technology and services review
4. Technical design and capacity planning
5. Financial analysis
6. Network rollout logistics

This approach begins with an evaluation of the current network, then sets objectives and goals, evaluates available technologies, provides technical design considerations for engineering a network for the unique characteristics of real-time communications, gives examples of how to do a financial analysis, and ends with guidelines to plan the deployment of the network.

Step 1—Network Audit

The first step in designing a multiservice network is to take stock of what currently exists in the data-only and voice-only networks. Review the existing equipment and evaluate its capabilities and operating costs. Determine existing facility costs and whether the current networks meet planned voice and data needs. Identify upcoming multimedia applications and new services that are needed to keep the company competitive, and determine their impact on the network as much as possible.

Determine the service quality of both voice and data to the user community, and which areas need improvement. Perhaps a new Web-based financial application does not have the user response time required to make employees productive. A traffic study may be necessary to look at current voice and data traffic patterns, as well as to determine which applications and protocols are using the data network today and what are their bandwidth needs. Perhaps bandwidth on some links in the network can be lowered, while others need to be increased.

Business policies should be reviewed or established of what traffic types (applications) on the network should get priority, which applications are mission-critical to the business, what bandwidth will be allocated to each, and what traffic on the network should be disallowed. The results of this analysis is at the basis of bandwidth provisioning as well as the Quality of Service (QoS) techniques that will be deployed in the network to ensure certain traffic types, including voice, meet the needs of the end user community and support business goals.

Step 2—Network Objectives

Once the current traffic and network use baseline is established, the next step is to set objectives for the integrated network. First, determine the dominant traffic types the integrated network is expected to support, and the service level expectations and agreements (even if informal) with user communities within the organization. Also, consider how closely voice and data functionality will be tied together. These appraisals will help in the selection of the appropriate technologies. Setting voice quality objectives will establish your organization's acceptable delay and compression limits.

Determine the voice traffic load the network can absorb and still meet the baseline data networking requirements. Determine the policies (priority and bandwidth) regarding mission-critical traffic types on the network.

There is usually not a single reason for migrating to a converged infrastructure, but some reasons are more important than others to a particular business. Being clear about these reasons will ensure meeting business goals when technologies are selected and network design decisions are made. Determine the top and lower priority reasons for migration in your organization. These may include:

- *Deploying new applications*—to enhance business efficiency and/or employee productivity
- *Cost savings*—on communications facilities, equipment purchase and maintenance, bandwidth efficiency, and long-distance or international voice call expenses

- *Increase competitiveness*—including e-business interworking with suppliers, partners and customers, offering new services and products, and improving customer care
- *Consolidate network management*—including simplifying equipment vendor relationships, leveraging cross-team skills, consolidating network management platforms and applications

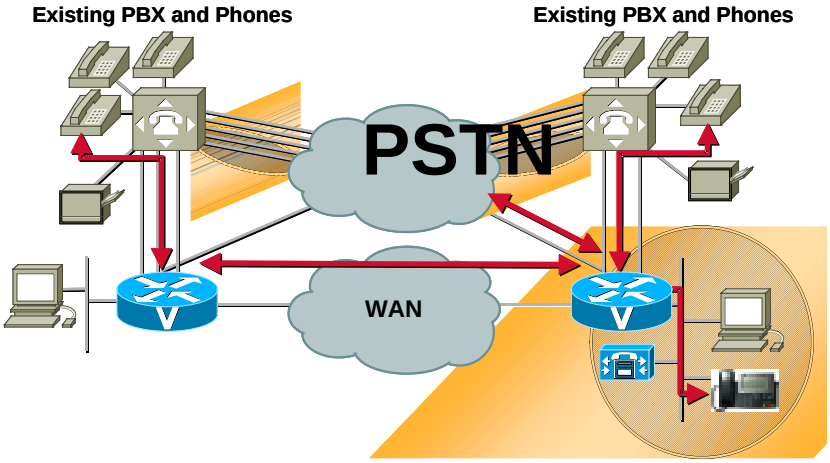
Determine the technical requirements of the converged network. Is any-to-any calling important, or only remote office to headquarters? Will fax traffic be carried? Should the dialing plan currently in place be retained or can it change? Will the converged network only carry On-net calls, or also be leveraged for On-net to Off-net (and vice versa) calling? Must the PBX configuration and software level remain unchanged, or will these be upgraded at the same time? Crystallizing these objectives will guide technology and network design choices.

Step 3—Technology and Services Review

The third step evaluates available technologies and services and selects the technology and connectivity choices that best meet the network objectives set in Step 2.

Multiservice IP transport systems follow a general model, as shown in Figure 1.

Figure 1 Multiservice Model



The transport network, which may be based on native IP, or IP over Frame Relay or ATM, forms the traditional “cloud”. At the edges of this network are devices or components called “voice gateways”. It is the responsibility of these devices to change the voice information from its traditional telephony form to one suitable for packet transmission. The network then forwards the packet to a voice gateway serving the destination or called party. Native IP devices such as IP Phones and PC applications also communicate with the gateways.

Voice over Packet Alternatives

One of the key technology choices is what type of voice over packet to deploy: Voice over IP (VoIP), Voice over Frame Relay (VoFR) or Voice over ATM (VoATM). VoIP is a layer 3 technology and can leverage Frame Relay and ATM layer 2 networks at the transport level. Appendix A discusses these technologies in greater detail. For VoIP, there are also multiple voice communications protocols that must be considered: H.323, MGCP and SIP.

The main decision point between VoIP and the layer 2 voice transport alternatives VoFR and VoATM, is interworking with other voice or multimedia applications. VoFR and VoATM are good WAN transport technologies, generally speaking more bandwidth efficient than VoIP, and some customers feel more comfortable with these technologies, but these can not be deployed over LANs or to the desktop. VoIP is the predominant form of Voice over Packet deployed today, and for voice application deployment this is the only choice even if the first step in network deployment is pure transport between existing PBXs. Leveraging the emerging employee productivity, customer care, e-commerce and business efficiency multimedia applications will require VoIP.

VoIP also leverages the entire Internet and Intranet IP infrastructure in terms of routing, making any-to-any calling in a VoIP network easy to design. Other benefits include that it is the only available Internet voice technology, and offers multi-vendor interworking which is more difficult to achieve with VoFR and VoATM solutions where standards have only recently emerged.

The International Telecommunication Union's (ITU) H.323 standard for VoIP has been available for several years, is widely deployed today and many stable implementations exist. Newer, still emerging standards for VoIP, including MGCP and SIP, will soon become valid technology choices. Both H.323 and SIP follow a distributed model where enough basic call processing intelligence is built into the endpoints or gateways to achieve a VoIP call with no further assistance from other network elements. Various server-based value-added features are available—for H.323 in terms of Gatekeepers, and for SIP in terms of SIP proxies and other servers. MGCP follows a centralized model where there is a central Call Agent server that contains all the call processing functionality. At least two endpoints and one Call Agent are necessary to achieve a voice call.

Network Transport Alternatives

When considering a multiservice network, an evaluation of three WAN technology alternatives, ATM, Frame Relay, and IP, should be made.

- *IP*—is connectionless and uses higher layers protocols such as UDP and TCP to enable "sessions" across the network, including voice calls. Real-time Protocol (RTP) is widely used for real-time application deployments such as voice and video. IP has robust signaling, addressing, and routing functionality, integrates well with current data applications and is the most ubiquitous networking protocol. IP also has the distinct advantage of being a Layer 3 protocol, so it can leverage the benefits of a layer 2 Frame Relay or ATM networks. IP runs all the way to the desktop for the greatest flexibility in supporting new Web-based IP applications, open IP-based PBXs, and IP telephones. You can run Voice over IP (VoIP) over ATM and Frame Relay networks (or any other WAN protocol such as HDLC or PPP), as well as Ethernet.
- *Frame Relay (FR)*—networks are relatively inexpensive and a common low-speed WAN access mechanism in many parts of the world. Frame Relay services may provide permanent or switched virtual circuits (PVCs or SVCs), but lack sophisticated signaling, addressing and routing functionality making it essentially a point-to-point technology. PVCs can be strung together by FR switching at intermediate network nodes to build an end-to-end path for the traffic across the network. FR can be used around the edges of the WAN to connect remote offices into the Intranet, or can be used as the entire backbone technology for the WAN.
- *ATM*—was designed to handle time-sensitive traffic, such as voice, and is connection-oriented. Using fixed-length cells (instead of variable length packets or frames), ATM's switching function in particular is optimized for high-speed performance, allowing you to build connections based on meeting certain delay and delay variation criteria. ATM is typically deployed in the higher speed backbone of the WAN and seldom around the low-speed edges. A common WAN for an Enterprise network has low-speed FR links to the small remote offices, while using ATM in the backbone and a high-speed ATM drop (OC3 or higher) into larger or headquarters offices.

Your current network probably already consists of a hybrid of transport technologies. If it already has IP over it, then deploying VoIP is an easy next step.

Telephony Signalling

Another important technology choice is how the existing PBXs, Key Systems and PSTN entrypoints in your network will connect to the voice-enabled multiservice infrastructure. Cisco voice gateways support a wide range of telephony signalling protocols to allow you the most flexibility in your choices. The voice gateway also terminates the telephony signalling so that the signalling type between the voice gateway and the PBX on the left of Figure 1 can be different, and independent of, the signalling chosen between the voice gateway and PBX on the right of the figure. This allows you to select the signalling type that is most convenient, or most predominant in various parts of the world, for each office in your global network.

Telephony signalling interfaces come in various flavors. These include:

- *Analog interfaces*—such as phone set, analog trunks to the Central Office and analog E&M trunks to a PBX. Analog interfaces typically carry one voice call per port, and are fairly rudimentary in the signalling information, such as Caller ID, passed to the receiving equipment. Whenever analog voice interfaces are used in a network, echo must be considered in the network design.
- *Digital interfaces*—are T1 or E1-based interfaces that carry 24 or 30 voice calls respectively. Some protocols use the ABCD bit patterns of T1/E1 frames to convey signalling (these are called in-band or Channel Associated, CAS, protocols), while others use a dedicated channel for signalling information (these are out-of-band or Common Channel Signalling, CCS, protocols). Generally the CAS protocols, like the analog interfaces, carry very little information with the call with the exception of T1 Feature Group D. The CCS protocols are rich in supplementary features such as number and name displays, and disconnect or error reason information. CCS protocols include SS7, ISDN Q.931 PRI and BRI as well the inter-PBX ISDN protocol known as Q.SIG.

Interface types and protocols have regional considerations. T1 is used in North America and to some extent in Japan, while the rest of the world is E1-based. ISDN protocols including PRI, BRI and Q.SIG are dominant is Europe, while E1 R2 is frequently used in South America, Asia and the Middle East.

The voice interface between each voice gateway and its attached PBX or Key System should be chosen with the following in mind:

- The region of the world where this gateway is located. Especially for PSTN connections this is a major consideration.
- Matching the gateway to the PBX/Key System if it is desired that the existing telephony systems in the network should not change or be upgraded.
- The cost of the interface on both the gateway and the PBX, or the cost of getting such a connection to the PSTN.
- The call features desired (such as Caller ID displays) and the types of interfaces that can provide that.
- Error handling capabilities of the interface. The CCS protocols have far greater efficiency in reporting and recovering from error conditions than analog or CAS interfaces.
- The number of calls to be carried across the interface. Analog is often more cost effective at lower density and digital better at higher density.

Step 4—Technical Design and Capacity Planning

The previous step evaluated the available packet and telephony technologies and services. This section covers the salient technical design guidelines that should be considered to implement a network with the desired voice quality.

Voice Coding and Compression

The term voice coding refers to the process of converting an analog voice signal to its digital counterpart. Pulse code modulation (PCM) is the standard for representing digital voice as a 64-kbps bit stream and is used on most voice networks today. Voice compression is the method of reducing the amount of digital information below the traditional 64-kbps. Advances in technology have greatly improved the quality of compressed voice and have resulted in a spectrum of ITU standards for compression algorithms. When dealing with voice compression, a trade-off occurs between the level of voice quality delivered and the amount of bandwidth savings achieved. Through the use of voice compression and, therefore, the optimization of bandwidth, significant cost savings are possible.

The most common compression methods, and their associated quality scores, are listed in Table 1 below. The acronyms listed in the table refer to the algorithm name and more information on what these mean and how they work can be obtained in the ITU standard (the G series) listed for each method. Further discussion of these is beyond the scope of this paper.

Table 1 Compression Methods

Compression Method	ITU Standard	Payload Bandwidth	MOS Score	Delay
PCM	G.711	64-kbps	4.1	0.75 ms
MP-MLQ/ACELP	G.723.1	6.3-kbps/5.3-kbps	3.8/3.75	30 ms
ADPCM	G.726	32-kbps	3.85	1 ms
LD-CELP	G.728	16-kbps	3.61	3 to 5 ms
CS-ACELP	G.729	8-kbps	3.9	10 ms
CS-ACELP	G.729a	8-kbps	3.85	10 ms

Digital Signal Processing (DSP) is a computer discipline specializing in applications such as voice coding and compression. DSP chips perform these functions on streams of digital traffic, and are commonly referred to as codecs (a contraction of coding/decoding). The codec of choice in you network will be a trade-off between the bandwidth efficiency you desire and the voice quality acceptable to your user community. For example, G.729 obtains an 8-times bandwidth savings over G.711, but incurs only a marginal drop in perceived quality (4.1 to 3.9). G.729 is the predominant codec used for transporting voice over WAN networks, while G.711 or G.726/32 are frequently used in LAN environments where bandwidth efficiency is of less concern.

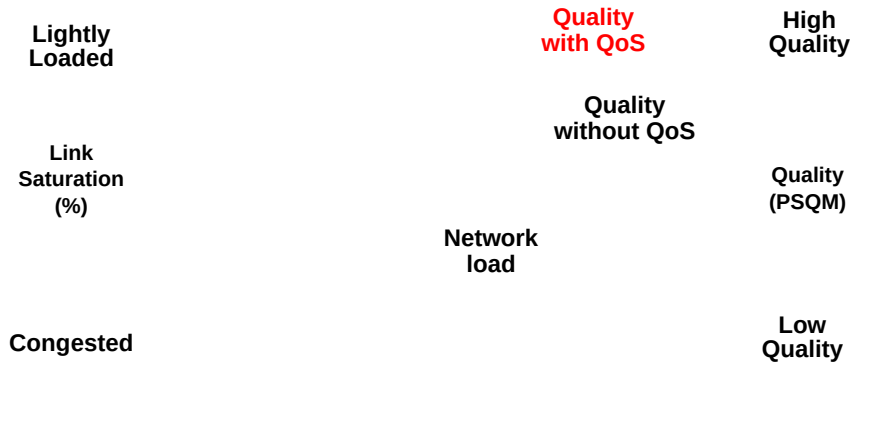
MOS, or Mean Opinion Score, is a widely quoted subjective measure of voice quality. Scores of 4 to 5 are deemed toll-quality, 3 to 4 communication quality, and less than 3, synthetic quality. Table 1 shows sample MOS scores for the commonly used codecs. MOS scores are not standard—they are based on the results of individual studies and although the ITU guides how studies should be conducted (P.800 series), they do not publish standardized MOS scores. The score for a particular codec can vary depending on the gender mix and language variety in the study. MOS scores typically rank codecs in the same order, even if the absolute scores vary. A newer method of objectively measuring voice quality, Perceptual Speech Quality Measurement (PSQM), has recently become available, and can be referenced as ITU standard P.861.

The contents of Table 1 makes clear the opportunity that exists to integrate voice and data networks while maintaining high voice quality. Note the trade-off between MOS scores, bandwidth use and delay. When designing networks, these factors must be balanced to ensure overall voice quality as well as meeting network objectives.

Voice Quality of Service (QoS)

Voice quality on a packet network is affected by more factors than codec choice. These include echo, delay and packet loss. The graph in Figure 2 shows how voice quality deteriorates with increased network load when no QoS features have been deployed in the network, whereas with QoS features the voice quality can be maintained regardless of data traffic load.

Figure 2 Voice QoS vs. Network Load



Echo—appears in a network whenever analog voice is used. Standards-based echo cancellation (G.165) is used to compensate for this.

Delay—has two components, fixed delay and variable delay, also known as jitter. The sections below will discuss the components of delay and how to plan for these. The objective of many QoS features in a packet network is to minimize delay and jitter for voice traffic.

Packet Loss—results when there is congestion in the network.To compensate for this, the network must be designed to selectively drop packets only from traffic types that can tolerate it. The codecs have auto-fill algorithms that will mask the loss of the occasional packet to the end user's ear.

This chart shows clearly that current voice over Internet products should not be confused with the practicality of carrying voice over IP. If QoS measures are designed into a network, toll-quality voice can be achieved with current technology.

Delay in Voice Networks

Two common network characteristics that affect quality are delay and jitter. Delay can cause two potential impairments to speech. First, long absolute delays cause both speakers to tend to begin to talk at the same time. Second, delay exacerbates echo, which is the reflection of the original signal back to the sender. Echo is indiscernible under low delay conditions. It is noticeable to the point of distraction when the delay becomes too great. Jitter causes gaps in the speech pattern that cause the quality of voice to be "jerky" or "clipped". Line quality also affects voice quality, but is outside the scope of this paper.

Table 2 shows a summary of the ITU recommendations for voice delay guidelines. One-way delays below 150 milliseconds (ms) are considered acceptable for most applications. Delays ranging from 150 to 400 ms are also acceptable subject to current voice quality. For example, a 300 ms delay from Chicago to New York is unacceptable, given experiences with public networks. On the other hand, a 300 ms delay from Chicago to Singapore may be acceptable given current conditions. Furthermore, higher delays may be acceptable if cost savings are taken into account. Generally, voice networks should be designed to achieve less than 200ms one-way delay.

Table 2 ITU Delay Recommendations

One-Way Delay	Description
0 to 150 ms	Acceptable for most user applications
150 to 400 ms	Acceptable provided that administrators are aware of the transmission time impact on the transmission quality of user applications
400+ ms	Unacceptable for general network planning purposes; however, it is recognized that in some exceptional cases this limit will be exceeded (for example, with satellite connections)

Fixed Delay

The components of fixed delay will now be explored, summarized in Figure 3.

Figure 3 Fixed Delay Components

•*Propagation Delay*—is based on the distance between source and destination. As a planning number, 6 microseconds per kilometer can be used. Propagation delay can be a significant factor in inter-continental connections.

•*Serialization Delay*—is the process of placing bits on the circuit. The higher the circuit speed, the less time it takes to place the bits on the circuit. So, the higher the speed the less serialization delay. For example, it takes 125 ms to place one byte on a 64-kb circuit. The same byte placed on an OC3 circuit will take 0.05 ms. Serialization delay is a significant factor on link speeds slower than 768-kbps.

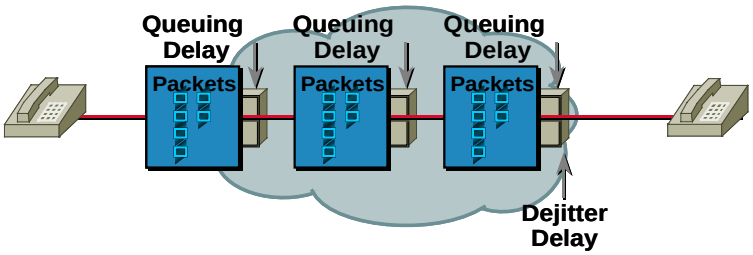
- *Processing Delay*—includes both algorithmic and packetization delay. Algorithmic delay is the time it takes the codec to produce a sample of voice. This varies between 1 and 30 ms for typical codecs. This function is often performed in DSP hardware and firmware, although it can be done in software as well. Packetization delay is the process of holding the digital voice samples for placement into the IP packet until enough samples are collected to fill the packet or cell payload.

It is important to keep in mind that some fixed delay components can be controlled by the network designer, while others can not. Propagation delay is purely a factor of distance and cannot be mitigated without adjusting the laws of physics. Serialization delay can be tuned by adjusting packet sizes, or by deploying packet fragmentation QoS techniques. Algorithmic delay is fixed per codec, but codec choice can influence this delay in the network design. Packetization delay is a configurable parameter on Cisco voice gateways, and can be adjusted based on the desired delay characteristics of the network.

Jitter

The delay components depicted in Figure 4 are variable in nature and result in delay variation, which is the difference in the interval between the arrival of subsequent packets belonging to the same voice conversation at the destination gateway.

Figure 4 Variable Delay Components



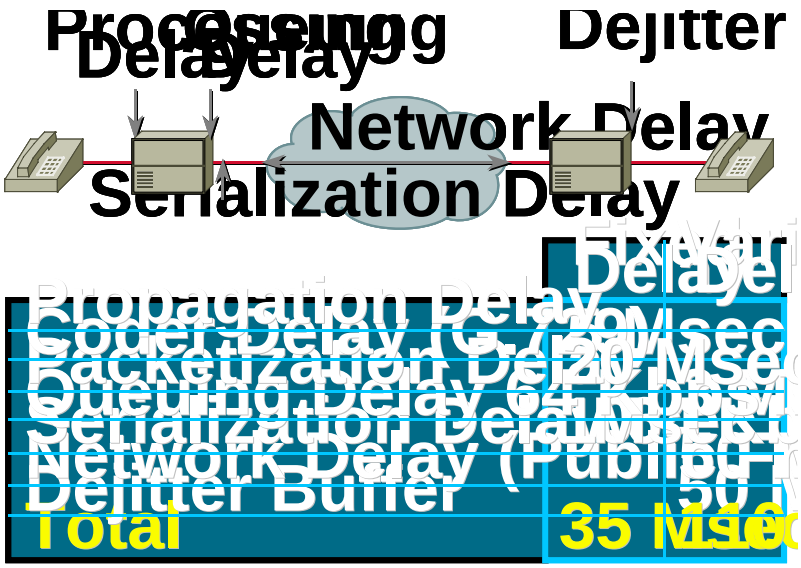
•*Queuing Delay*—is the delay caused by waiting on other packets to be serviced first on an egress interface. At any point in time a voice packet may be waiting in queue for a variable amount of time while awaiting access to the interface. Queuing delay is only incurred when there is contention on the interface, i.e. when there is more traffic waiting to go out than the link speed can accommodate. Queuing delays are therefore more likely and more severe on slower speed interfaces. Waiting time is also based on the size of the packets ahead in the queue. Proper configuration of routers is important to minimize this type of delay.

- *Dejitter Buffers*—are used at the receiving end to smooth out the delay variability of the arriving voice packets. They also help on the first talk spurt to provide smooth playback of the voice. If these buffers are too shallow, underflows and speech clips results. If they are too deep they cause excessive delay. In effect, a dejitter buffer reduces or eliminates delay variation by converting it to constant delay. Dejitter buffers adapt dynamically to the actual jitter being experienced by the arriving packet stream and should be monitored. If they grow too deep, the network design should be examined for the elements that are inducing excessive jitter and the source of the problem must be adjusted rather than the dejitter buffer.

Calculating a Delay Budget

Given a general understanding of the fixed and variable delay components, the delay budget should be calculated before voice is placed on a packet network. The delay budget is the amount of delay permissible in the planned network while still meeting voice quality objectives, shown in Figure 5.

Figure 5 Delay Budget Example



In the Example, a delay budget of 200 ms will be used. Note that this example is of a public Frame Relay network, so the service provider's network delay figures should be used in the delay budget. If these are not available, they should be measured.

Voice packets will contain two 10-byte packets. The coder delay for G.729 voice compression is an initial 5 ms for a look-ahead, plus 20 ms for the two 10-byte samples.

Queuing delay is variable, and should be configured to be minimal. We will assume 3ms here.

Serialization delay is the time it takes to play out a packet onto the 64-kbps egress interface. The playout of the voice packet itself (typically a very short packet) is not as important as the playout of the data packet that precedes it. Fragmentation techniques should be deployed such that no data packet will delay a voice packet behind it for more than 10ms on a slow-speed egress interface.

Finally, a typical figure to use for a dejitter buffer is 50-60 ms. The total delay in this example is 145 ms. This is well within the delay guidelines set by the ITU and the target planning number of 200 ms.

Quality of Service Techniques

Many of the QoS tools in a packet network are geared towards minimizing delay and packet loss. This requires the network to recognize a voice packet and to apply differentiated treatment to voice packets vs. data packets. Planning for QoS in a network is really a broader exercise than just voice and data. All the traffic types using the networks must be understood, how important each is to the business, and what bandwidth and treatment requirements each imposes on the network to provide the service level expected by the end user.

The IETF's Architecture for Differentiated Services (RFC 2475) describes three broad types of traffic:

- *Low Latency, Guaranteed Delivery*—includes voice as well as any other traffic sensitive to delay and packet loss.
- *Guaranteed Delivery*—includes mission-critical traffic. This traffic is not latency sensitive (within reason) and can tolerate controlled amounts of packet loss.
- *Best Effort*—traffic that is either irrelevant to business needs (e.g. employee personal web surfing), or can be delayed significantly with minimal business impact (e.g. nightly backup traffic), or can recover from loss with reasonable end user results (restart a failedfile transfer session).

For the network components to provide the appropriate treatment to these categories of traffic, it is necessary that packets must be classified into the category where they belong. This classification is often based on business policy as much as the characteristics of the application under consideration.

QoS tools ensure the proper treatment of traffic in the network to deliver the desired quality and/or response time. There are various types of QoS tools, providing different functionality in the network:

- *Classification*—is the inspection of packets to determine what type of traffic it is. This is usually done close to or at the source of the traffic, e.g. on the ingress interface of a LAN segment to an Oracle server that hosts a mission-critical financial application. Or for voice traffic on a voice gateway. Many classification tools exist, and can classify traffic depending on various attributes such as protocol, port number, packet size, packet header content, interface etc.
- *Policing*—is the decision of what traffic to carry on the network. It could be that a particular type of traffic is disallowed (e.g. Napster traffic) and all packets dropped, or it is rate-limited meaning only a certain agreed amount of that traffic type will be allowed onto the network and packets exceeding that rate will be dropped.
- *Marking*—is inserting a mark in the header of a packet that has been classified and has passed the policing test. Marking is done so the packet doesn't have to be classified again—subsequent network nodes can simply look at the marking in the packet header and treat the packet appropriately. Packet inspection and classification can be processor intensive and should be done around the edges of the network, not in the backbone. Marking techniques include the Ethernet 802.1p/q Type of Service field, IP precedence or a DiffServ Code Point (DSCP) marking in an IP packet, and the DE (Discard Eligible) bit in a FR frame.
- *Queuing*—is an egress interface technique. The algorithm has two components, a decision on which queue to place an arriving packet in (queuing), and a decision of which queue's front packet to service next when the interface becomes free (scheduling). Queuing techniques provide the mechanism to prioritize one type of traffic over another when congestion exists on the egress interface. It is a key technique to minimize jitter in voice, by ensuring voice packets are serviced when present, while data packets wait in queue.
- *Traffic Shaping (TS)*—is another rate-limiting technique. Unlike policing, traffic shaping does not drop packets that exceed the given rate, but buffers or queues them. TS ensures a constant flow of egress traffic not exceeding the rate specified and smooths out the fluctuating arrival traffic to conform to the rate. This is often done upon entry into a Service Provider's network where there is a service level agreement of traffic rates that will be carried. Traffic shaping is also used as a technique to smooth out traffic over mismatching speeds on interfaces on each side of the network, another key task in minimizing voice jitter.

Dialing Plan Design

As discussed earlier, the telephony signalling is terminated by the voice gateway. This provides two important features, first the ability to switch a call anywhere in the network, and secondly to decouple the signalling dependencies on either side of the network. This implies that the voice gateway must understand the dialed digits in order to set up a voice call to the appropriate destination. In an IP network this amounts to translating the dialed digits to an IP address. To aid the voice gateway in this mapping, a dialing plan is implemented on the voice gateways in the network, similar to those on PBXs and in the PSTN.

Several factors must be considered regarding the dialing plan to ensure a smooth cutover and a manageable network. First, must the dialing plan from the user's perspective remain unchanged? Or will a new PBX "access code" be introduced to allow users to choose the IP network or the PSTN for their calls? Second, will the IP network only carry On-net calls, or also provide gateways to the PSTN? For On-net calls, only the company's private dialing plan needs to be implemented. If Off-net calls are also accommodated, the gateway may have to translate an internally dialed number to a publicly accessible number the PSTN will understand, including adding/deleting country and regional area or city codes.

Digit manipulation is a must in a network of any size and the Cisco voice gateways offer a plethora of tools to manipulate numbers such that the end user interface can be what the business desires, and digits can be changed to adapt to country specifics or public vs. private network access. Translation tools exist to manipulate both the calling and called digits of a call, which is often important for billing or tracking purposes.

Capacity Planning

Trunk provisioning is establishing the number of trunks (voice channels) from the PBX to the multiservice transport network. After establishing the number of trunks, the next step is to translate that number to the required network bandwidth.

The correct number of PBX or key system trunks will be determined by the traffic volume and flow, the selected grade of service or blocking factor, and other network-specific objectives. Some organizations will simply move trunks from the current network to the integrated multiservice network. Others will take this opportunity to update the traffic engineering information and conduct a traffic study. Either approach can work and is very familiar territory for voice engineering professionals.

In most cases you are only designing for On-net traffic. For example, the branch office of a bank has 30 employees. A large percentage of the voice traffic consist of local calls to customers. Typically, only 20 to 30 percent of the traffic is eligible for On-net. If this branch-office has a total of 10 outside lines, then only two to three voice channels are candidates for routing over the multiservice network.

Based on the proposed network design and the required number of trunks between locations, the required bandwidth can be calculated. Bandwidth calculations should take into account compression, overhead, and utilization. Each of these will vary, depending on which transport technology is chosen. Bandwidth efficiency techniques like RTP Header Compression (cRTP, applicable only to VoIP) and VAD (Voice Activity Detection) can be deployed. VAD, or Silence Suppression, keeps packets from being sent when there is no live voice present. For example, a conversation has a speechpath in both directions at all times, but typically only one person is speaking. VAD will suppress the listening party's "empty" packets from being sent onto the network. In addition, on a multiservice network, voice bandwidth is only needed when calls are actually being made. The benefit is that when calls are not being made, perhaps at night when the office is closed, the bandwidth is available for data applications like nightly backups.

It is important to design the network such that voice calls that will oversubscribe the allocated network bandwidth are kept off the network and rerouted via an alternate path like the PSTN. This is another voice QoS category called Call Admission Control. For example, if the WAN access link from a branch office is provisioned to carry no more than five simultaneous calls, the sixth call must not be allowed onto the network as it will impair voice quality.

Technical Guidelines Summary

The prior section provides a method for calculating your delay budget, and technical guidelines to be used when designing an integrated multiservice network. As in all designs, a balance must be found between quality and costs. Given the large mean opinion score (MOS) quality improvements in today's compression techniques, finding this balance is easier than ever.

The guidelines can be summarized as follows:

- Balance voice quality, delay, and bandwidth
- Determine acceptable delay and delay variation thresholds
- Calculate delay for the network and adjust design parameters to keep delay and jitter to within the budget
- Deploy QoS techniques to minimize several of the delay components
- Give due consideration to the dialing plan of the network
- Plan the capacity and the bandwidth required for the traffic expected on the network, and take steps to deal with excess traffic

Step 5—Financial Analysis

The objectives have been set, the packet voice technologies chosen, the network design planning completed, and the capacity determined to support the additional traffic within the delay budget. Now, the question remains: is the network cost justified?

A case study is presented in Appendix B to illustrate the steps required to perform a return on investment (ROI) analysis of a multiservice network. The case presents a small international firm based in San Jose, California, using a private line network with routers. This case was built using standard voice engineering principles and actual public switched telephone network (PSTN) rates and private line costs, and is presented to help you see how a ROI analysis is conducted.

Step 6—Network Roll-out Logistics

If step 5 results in an acceptable ROI, the time has come to implement the network. In a network of any realistic size, this does not happen overnight, and current applications can not be disrupted while the new network is cut over. Many customers prefer to start with a proof of concept laboratory test where a PBX can be connected to voice-enabled routers over a simulated backbone. Experience is gained with the new technology and configurations independent of the live network, and the design can be proven or adjusted. Next a small pilot test phase may be rolled out to a select number of users who have been prepared or trained to report on the voice quality they receive and perhaps have an alternate way of dialing a call if the network does not behave the way they need to conduct business. Many configuration and feature adjustments may happen during this pilot phase to tune the network, services and design.

When a successful pilot phase has been completed, a roll-out plan for the remaining sites can be put in place and each can be cut over at a time that is convenient for the support group. Existing equipment may have to be upgraded, either software or hardware or both, to be voice-ready before the cutover of that site can take place. The logistics of the cutover phase depends on the user community, the geographic challenges of getting to a site, the potential impact of disrupting service to a particular application or site and many other factors.

Conclusion

Multiservice networking is not just about convenience. It is about preparing companies to be more competitive and efficient in today's global economy. It is about being ready to deliver the Web-based multimedia services needed to assure continued survival and success. The cost savings and increased bandwidth from multiservice network integration can deliver an ROI that not only justifies the initial consolidation of voice and data networks, but also makes the funds available to develop the new applications and build a more robust network needed for competitive advantage.

By planning and executing the migration for existing networks in stages, network managers can easily roll out new network services and Web-based multimedia applications as needed. New locations can benefit from having multiservice-capable routers and switches installed from the start and having new services turned on when ready.

Cisco is the market leader in voice over packet solutions and has the products that you need today to begin deploying your multiservice network. From a small branch-office to a large enterprise campus or a service provider network, Cisco has a range of flexible and scalable multiservice products that will meet your business requirements.

In addition, Cisco is committed to delivering world-class service and support as well. Cisco has made significant investments to ensure that you have the support you need when you deploy a multiservice network, and has service and support offerings at every point of contact, whether direct or through Cisco authorized resellers and service providers.

Cisco has established teams specializing in voice and high-availability systems within each support delivery organization to ensure access to experts with the knowledge and experience for successful implementations of multiservice networking. Throughout the life cycle of a multiservice network, from planning and design through implementation and operation, companies have access to expertise and experience through any Cisco support products.

Appendix A—Overview of Voice over IP, Frame Relay, and ATM

The following transport technologies can be used to create a multiservice network: IP, Frame Relay, and ATM. Every network must, in some fashion, perform addressing, routing, and signaling duties. Addressing is done on several levels, but at the very least is required to identify the calling and the called party. These high level addresses will be translated to lower level addresses understood by various network elements. Routing is the procedure used to find a path through the network from source to destination and moves the information through the network. ATM and FR are layer 2 protocols and build a path through the network by concatenating a number of point-to-point virtual circuits. Signaling carries information about the session being setup from source to destination, including alerts, status and the actions necessary to establish a connection.

A certain degree of communications knowledge has been assumed, and the discussion is limited to those areas most pertinent to data and voice integration.

I. Voice over IP (VoIP)

IP is a layer 3 network protocol, and uses various layer 2 point-to-point or link layer protocols such as HDLC, PPP, FR or ATM for its transport. IP routing forwards a packet between one link and the next towards its destination.

Voice over IP Standards and Protocols

Table 3 depicts the relationship between the International Organization for Standards (ISO) reference model and the protocols used in IP voice network elements.

Table 3 ISO Reference Model and VoIP Standards and Protocols

ISO Protocol Layer	Standards and Protocols
Presentation	Codecs/Netmeeting/Applications
Session	H.323/SIP/MGCP
Transport	RTP/TCP/UDP
Network	IP
Link	FR, ATM, Ethernet, MLPPP, PPP, HDLC and more

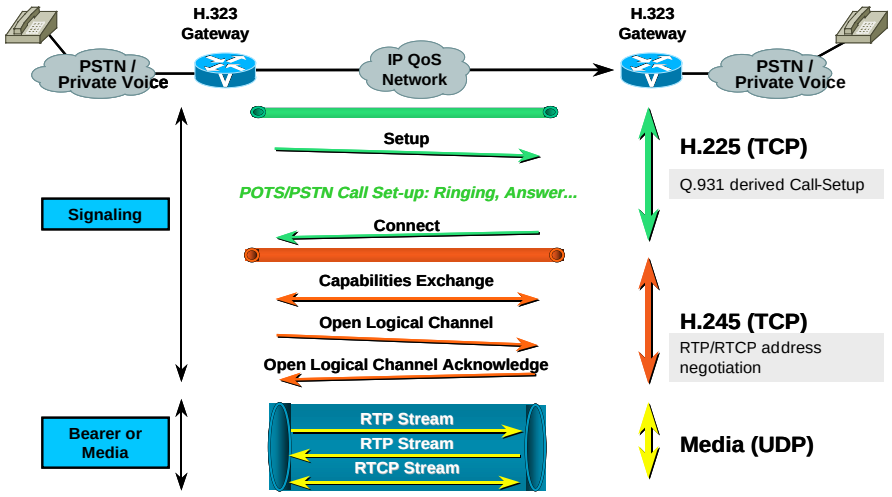
- *H.323*—ITU Standard for conducting voice over IP. It includes several related standards such as H.225 (call control), H.245 (media path and parameter negotiation), H.235 (security), H.225 Registration, Admission and Status (H.323 Gatekeeper protocol), and H.450 (supplementary services).
- *SIP*—Session Initiation Protocol is defined via IETF RFC 2543.
- *MGCP*—Media Gateway Control Protocol, an IETF draft standard for controlling voice gateways through IP networks.

VoIP protocols typically use Real-time Protocol (RTP) for the media stream (the speechpath). RTP uses User Datagram Protocol (UDP) as it's transport protocol. Voice signalling traffic often uses Transmission Control Protocol (TCP) as its transport medium. The IP layer provides routing and network level addressing, while the link layer protocol controls and directs the transmission of the information over the physical medium.

H.323 Call Set-up

H.323 is the dominant VoIP protocol deployed today and the most mature of the VoIP technologies. Figure 6 shows a basic call set-up between two voice gateways using H.323.

Figure 6 H.323 Call Set-up



The phone on the left of the figure initiates a call. The dialed digits are delivered to the voice gateway via a telephony signalling protocol such as ISDN PRI. The voice gateway translates the digits to a destination IP address and initiates an H.225 call set-up to that network node.

H.225 is a protocol very similar to Q.931. The two voice gateways will negotiate parameters such as UDP ports to be used for the media stream via H.245 before the call set-up completes. When the call is answered, the speechpath is established via RTP.

H.225 is again used, as in ISDN Q.931, to tear down the call when the parties hang up on the call.

H.323 version 2 contains a FastConnect call set-up that cuts down significantly on the number of individual messages exchanged by the gateways.

Voice over IP QoS

Unclassified IP traffic is treated as “best effort” by the network, meaning that incoming IP traffic is allowed to be transmitted on a first-come, first-served basis. To maintain voice quality, voice packets must be prioritized over other IP traffic. Voice packets should be classified and marked with IP Precedence 5 or DSCP Expedited Forwarding (EF). The packet marking controls the queuing and forwarding features on the network nodes to ensure voice is treated with the priority necessary to minimize delay and jitter.

Cisco voice gateways and network nodes implement a sophisticated queuing technique called Low Latency Queuing (LLQ) that accomplishes this by giving strict priority to voice packets and at the same time ensuring the data traffic cannot be starved by voice traffic.

If VoIP is implemented across low-speed (less than T1 speeds) WAN access links, link fragmentation and interleaving (LFI) techniques must also be deployed to minimize serialization delay. MLPPP can be used over leased lines or ATM, and FRF.12 can be used over FR links.

Voice over IP Routing

One of the strengths of IP is the maturity and sophistication of its routing protocols. A routing protocol, such as the Enhanced Interior Gateway Routing Protocol (EIGRP), is able to take delay into consideration in calculating the best path. These are also fast-converging routing protocols, which allow voice traffic to take advantage of the self-healing capabilities of IP networks.

Advanced features, such as policy routing and access lists, make it possible to create highly sophisticated and secure routing schemes for voice traffic. RSVP can be invoked by Cisco VoIP gateways to ensure that voice traffic is able to get a guaranteed path through the network.

II. Voice over Frame Relay (VoFR)

Frame Relay is quite common and comparatively affordable in most parts of the world. Frame Relay is an interface specification, whereas ATM and TCP/IP are architectural specifications. Consequently, Frame Relay will likely be used solely as a transport mechanism. Cisco has enhanced the local management interface (LMI) and the congestion notification methods between Cisco Frame Relay switches and routers to improve VoFR functionality.

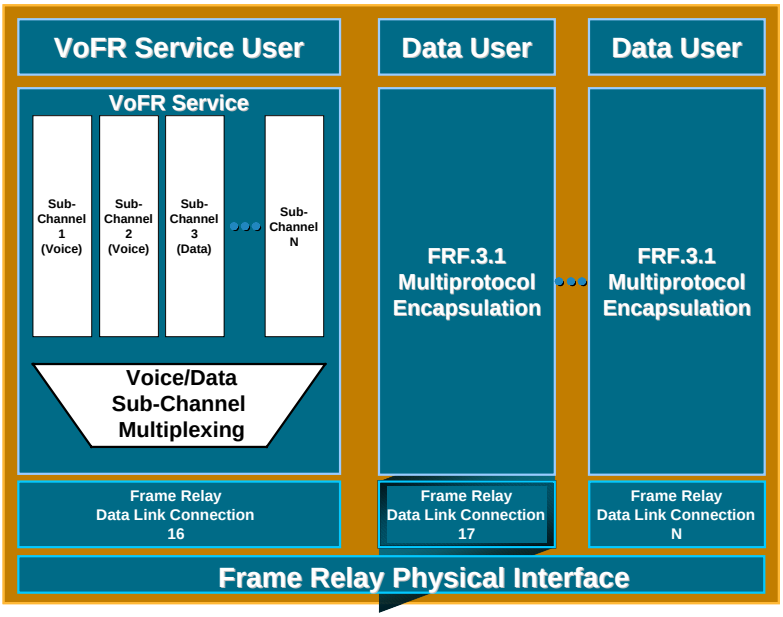
Voice over Frame Relay Standards and Protocols

Frame Relay call setup tends to be proprietary by vendor because the Frame Relay Forum voice standard FRF.11 is a point-to-point protocol and does not specify a call set-up procedure. Cisco voice gateways implement pure FRF.11 for interworking with other vendors’ equipment, but also implement extensions to FRF.11 to enable switched voice networks to be built over FR networks.

Address mapping is handled through statically mapping the dialed digits to a FR PVC. The traffic will be switched trough the network based on the FR switching implemented between PVCs on network nodes, and a voice call will likely traverse several PVC hops before reaching its destination. A full mesh of PVCs between all network nodes is not necessary to carry voice, but delay must be taken into account when designing the network to ensure a voice call does not incur too much delay by traversing too many hops.

Figure 7 shows the Frame Relay Forum Model for multiplexing voice and data on PVCs on a FR interface.

Figure 7 VoFR PVC Multiplexing Model



Voice and data can share the same PVC and this is a major advantage if the service provider charge to the customer is per PVC.

To accomplish this FRF.11 specifies sub-channels on a PVC where each sub-channel can be of a different media: voice, fax, signalling or data. PVC 16 in the figure is defined as a VoFR PVC and therefore uses FRF.11. PVCs 17 and up are pure data PVCs and are governed by FRF3.1.

Voice over FR QoS

Unlike VoIP, VoFR traffic does not need to be explicitly classified as the FR software can recognize a voice packet from a data packet by the protocol and type fields in the FR header. However, Cisco voice gateways use the same queuing technique, Low Latency Queuing (LLQ), to prioritize VoFR packets over data packet on the egress interface.

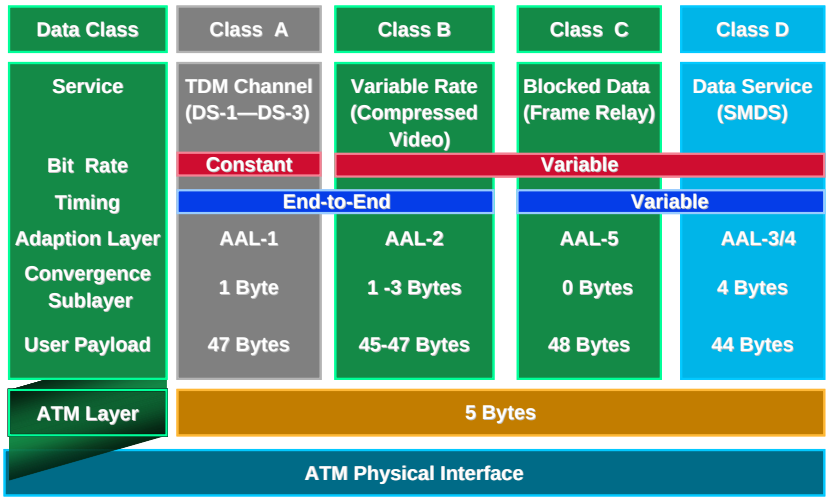
FRF.11 Annex C also specifies a link fragmentation and interleaving (LFI) technique that must be deployed to minimize serialization delay on FR links of less than T1 speeds. The FR traffic shaping parameters (CIR, Bc and Be) values must be carefully tuned to minimize delay and jitter and voice traffic streams. Data (FRF3.1) PVCs sharing the same interface with a VoFR PVC must also be fragmented—this is done via FRF.12—as serialization delay is incurred at the interface level, not at the PVC level.

III. Voice over ATM (VoATM)

The ATM Forum has specified different classes of services to represent different possible traffic types. Designed primarily for voice communications, constant bit rate (CBR) and variable bit rate (VBR) classes have provisions for passing real-time traffic and are suitable for guaranteeing a certain level of service. CBR is typically used for circuit emulation services (CES), while VBR is used for voice over packet VoATM. Unspecified bit rate (UBR) and available bit rate (ABR) are more suitable for data applications. UBR in particular, makes no guarantees about the delivery of the data traffic.

The method of transporting information through an ATM network depends upon the nature of the traffic. Different ATM adaptation types have been developed for different traffic types, each with its benefits and detriments. ATM Adaptation Layer 1 (AAL1) is the most common adaptation layer used with CBR services, while ATM Adaptation Layers 2 and 5 (AAL2 and AAL5) are widely used for voice over packet transports. Figure 8 shows the ATM Forum's Protocol Model.

Figure 8 VoFR PVC Multiplexing Model



ATM uses a fixed-length 53-byte cell to transport traffic. Five bytes are used for the ATM cell header, leaving 48 bytes for the AAL encoding. This, as shown in the figure, in all cases except AAL5, uses a few more bytes for header information. The fixed length cell of ATM provides for high-speed hardware-based switching capabilities making ATM a robust backbone technology. Variable length IP packets are segmented into 53-byte portions and reassembled on the egress side of the ATM network. This is called the ATM Segmentation and Reassembly (SAR) function. Multiple IP packets will not be contained in the same ATM cell, so for short IP packets, ATM cell fill efficiency is an important network design consideration. A typical G.729 VoIP packet is 60 bytes, which requires two ATM cells (assuming AAL5 encoding) with 96 bytes payload (48 bytes payload in each AAL5 cell) of which 36 are not used. VoIP packet sizes should be tuned to optimize ATM cell fill if ATM is used at the transport technology.

Voice over ATM QoS

The definition of ATM service types (CBR, VBR etc.) also include various QoS parameters, including the following:

- Peak Cell Rate (PCR)—the maximum data rate a connection can carry without losing data
- Sustainable Cell Rate (SCR)—the average cell throughput
- Maximum Burst Size (MBS)—the size of the maximum burst of cell allowed
- Minimum Cell Rate (MCR)—the rate of an application's ability to handle latency
- Maximum Cell Transfer Delay (MCTD)—how long the network can take to transmit a cell
- Cell Delay Variation Tolerance (CDVT)—jitter

ATM VBR service is predominantly used for packetized VoATM and VoIP-over-ATM services (excluding circuit emulation services). VBR is often offered in two flavors, VBR-rt for real-time traffic and VBR-nrt for non-real-time traffic.

Because ATM is typically used over high-speed interfaces (speeds of T1 and higher), serialization delay is not usually an issue in ATM networks. It may become a design consideration if the network mixes FR circuits at the edges and ATM at the high-speed core or drops into large sites. For these networks, LFI must be deployed.

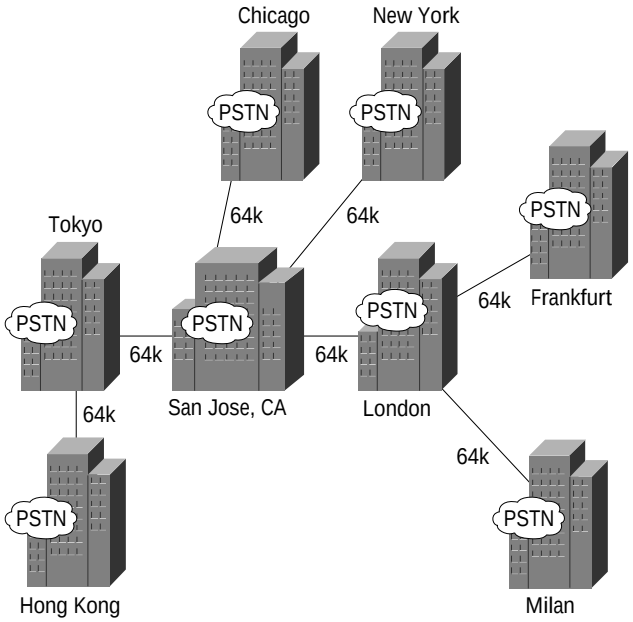
If voice and data are mixed over the same ATM VC, then sophisticated queuing techniques like LLQ may also be necessary.

Appendix B—Case Study

This case study will give you an example of a financial analysis of the benefits of a multiservice network migration.

With its headquarters in San Jose, California, this company has branch-offices in Hong Kong, Tokyo, Chicago, New York, London, Milan, and Frankfurt. There are about 15 people in the branch-offices except for London and New York, where there are approximately 45. The network topology and leased-line circuit speeds are shown in Figure 9.

Figure 9 Case 1 Initial Network Topology



Most branch calls are between branch employees and customers in the local area. Calls between headquarters and branch employees account for only 20 percent of the total call volume. Although the branch-to-headquarters traffic volume is in the minority, it is also the most expensive on a per-minute basis because it is billed at international rates. Consequently, for international long distance services, this firm pays approximately US$32,000 per month, or about US$390,000 annually. Because the company wants to enable branch-to-branch dialing, instead of just branch-to-headquarters, a translate model will be used.

Voice Network

Key systems and small PBXs connected over the PSTN provide voice services. In the analysis that follows it is assumed that the firm generates enough call volume to obtain a VPN contract from a carrier at about a 15 percent discount from standard PSTN rates. This is to ensure the financial analysis is conservative.

In the branch-offices, each individual spends approximately two and a half hours each day communicating via telephone or fax. About 20 percent of this traffic is between the branch and headquarters. Table 4 shows the potential on-Net voice and fax traffic volume and expense.

Table 4 PSTN Volume and Expense

Location	Purpose	Number of People	Average Minutes per Person per Day	On-Net Percent to Headquarters	Number of Workdays per Month	Total Minutes per Person per Month	Total Minutes per Office per Month	Cost per Minute [1]	Monthly Cost per Office
San Jose, CA	Headquarters								
Frankfurt	Branch	15	150	20%	21.67	650	9,752	$0.54	$5,266
Milan	Branch	15	150	20%	21.67	650	9,752	$0.48	$4,681
London	Branch	45	150	20%	21.67	650	29,255	$0.29	$8,484
New York	Branch	45	150	20%	21.67	650	29,255	$0.07	$2,048
Chicago	Branch	15	150	20%	21.67	650	9,752	$0.07	$683
Tokyo	Branch	15	150	20%	21.67	650	9,752	$0.52	$5,071
Hong Kong	Branch	15	150	20%	21.67	650	9,752	$0.63	$6,095

Location	Purpose	Number of People	Average Minutes per Person per Day	On-Net Percent to Headquarters	Number of Workdays per Month	Total Minutes per Person per Month	Total Minutes per Office per Month	Cost per Minute [1]	Monthly Cost per Office
Total							107,267		$32,326

1. Assumptions
–This is the average of the cost per minute in each direction and assumes 50 percent of calls are to the Headquarters and 50 percent from the Headquarters.
–The cost of a voice call is based on a carrier quote with customer discount.
–All costs in US dollars.

Data Network

An eight-node data network, leased by the firm, utilizes routers, and is hubbed out of its San Jose headquarters. A number of branches connect through other branches to reach the headquarters: this arrangement was set up to keep leased-line costs low.

Network Redesign

It is essential that redesign of the data network to support the added voice traffic be accomplished without adversely affecting performance and quality. The plan is to have the PSTN cost reductions pay for the multiservice enhancements. Though any of the packet voice technologies could be used to build this multiservice network, given the firm's infrastructure and expertise in IP, a VoIP network is chosen. Considering the assumptions, the following redesign is conservative.

First, the additional bandwidth required on the data network to support the voice and fax traffic is determined. The best way is to collect traffic information from both the key system or PBX and the router, and then graphically add the voice and data traffic together. This enables one to see how often the combined voice and data traffic would exceed the available bandwidth. But this kind of traffic information is often unavailable. If this information is not available, the easiest method to establish the proper upgrade would be to estimate whether (and how much) extra bandwidth would be required, and provision that amount. Then the voice traffic can be added to the data stream while tracking two measures of performance: user-reported voice quality and data latency. If either performance measure appears to be insufficient, then more bandwidth should be added or possibly a different voice codec should be used.

Data and voice traffic frequently peak at different times in the day. The data network will frequently benefit from the added bandwidth.

In redesigning the network the following assumptions were used:

- There are approximately 15 people per small branch, 45 per large branch.
- The bidirectional voice and fax call volume totals about two-and-a-half hours per person, per day, per branch.
- About 20 percent of the total call volume is between headquarters and each branch location, or branch to branch.
- A busy-hour loading factor of 17 percent is appropriate.
- The Cisco voice compression module uses 8-kbps, plus 3-kbps overhead per call. It was assumed that a 64-kbps trunk circuit supports five calls, rather than seven. This is a conservative estimate if voice activity detection is used.
- In the small branches, one key system trunk module would be required, whereas two cards would be necessary for the large branches.

Using the above assumptions and the following calculations, the amount of voice and fax traffic at each branch office that can optimally be diverted from the PSTN to the multiservice network. For information on Erlangs and to use on-line Erlang calculators, please refer to http://www.erlang.com.

Voice and fax traffic calculations:

- 2.5 hours call volume per user per day X 15 users = 37.5 hours daily call volume per office
- 37.5 hours X 60 minutes per hour = 2,250 minutes per day
- 2,250 minutes X 17% (busy hour load) = 382.5 minutes per busy hour
- 382.5 minutes per busy hour X 1 Erlang/60 minutes per busy hour = 6.375 Erlangs
- 6.375 Erlangs X 20% of traffic to headquarters = 1.275 Erlangs volume proposed

To determine the appropriate number of trunks required to carry the traffic, traffic engineering tables are consulted next, given the desired P grade of service. This firm chose a P.05 grade of service. Table 5 shows the applicable sections of the Erlang C tables.

Table 5 Case 1 Erlang C Table

Blocking Probability (Grade of Service)	Small Branch Traffic to Headquarters (1.275 Erlangs)	Large Branch Traffic to Headquarters (3.825 Erlangs)
P = 0.01	5 trunks	10 trunks
P = 0.05	4 trunks	8 trunks
P = 0.10	3 trunks	7 trunks
P = 0.20	3 trunks	6 trunks

Using the calculated Erlangs and Table 5, it turns out that four trunks are required in the smaller offices and eight trunks in the larger ones. Table 6 summarizes the calculations and trunking requirements, along with the figures for the larger branch-office.

The London and Frankfurt private lines were increased from 64-kbps to 128-kbps to support the added voice traffic based on the above conclusions. Since the circuit from New York to Paris must carry, at a maximum, the four compressed voice streams from Chicago and the eight compressed voice streams from New York, it is increased to 192-kbps. At 12-kbps apiece, this added traffic at its maximum would be 12 voice streams x 11-kbps/voice streams = 144-kbps. This still leaves 58-kbps for the data traffic when all voice circuits are busy. Since most of the time some of the voice circuits would be unused, there would often be more bandwidth available for data traffic, and the added bandwidth would provide enhanced data performance.

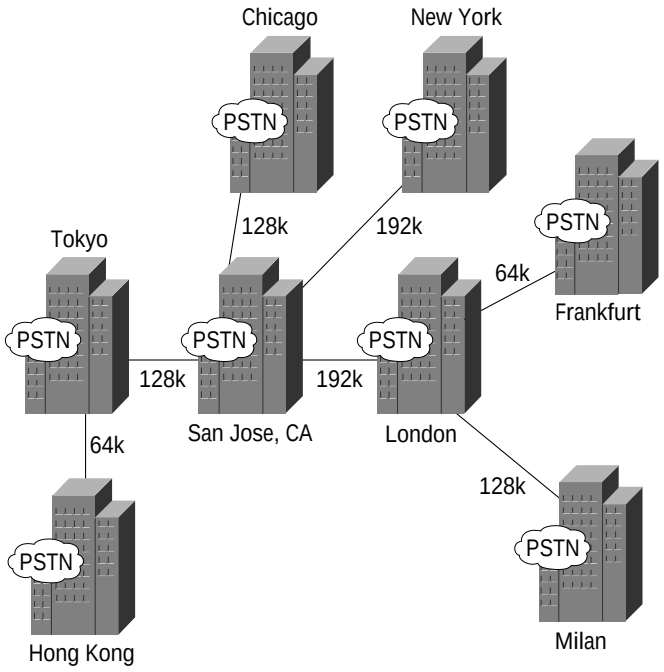
Table 6 Trunking Analysis

Site	Number of Users	Call Volume per Day	Minutes per Day	Busy Hour (17%)	Minutes per Busy Hour	Percent of Traffic to Headquarters	Total Erlangs to Headquarters	Required Trunks for 0.05 Blocking Probability
Large	45	2.5	6,750	0.17	1,147.5	20%	3.825	8
Small	15	2.5	2,250	0.17	382.5	20%	1.275	4
						2 large	7.65	
						5 small	6.375	
						Total Erlangs for Headquarters	14.025	23

The Chicago-to-New York circuit was not increased beyond its original 64-kbps. Even at its maximum, the added Chicago voice traffic can add no more than 36-kbps of traffic onto the data pipe, throttling the data traffic down to a minimum of 28-kbps. Like Chicago, many locations will find that their data traffic can tolerate the added delay created by this occasional reduction in bandwidth. The peak hours for voice calls frequently differ from the peak data-transfer times, and the two sets of traffic rarely interfere with each other, as mentioned previously.

The Hong Kong-to-Tokyo link did not require an upgrade from its original 64-kbps as is the case for the Chicago-to-New York link. The combined voice traffic from Tokyo to Paris (four channels from Hong Kong, four more from Tokyo) required a maximum of 72-kbps (and usually used less); so, only the Tokyo circuit was increased to 128-kbps. This leaves a minimum of 56-kbps for data traffic, and during less-than-peak voice calling moments, it offers more than the original 64-kbps for improved data performance. The redesigned network appears in Figure 10.

Figure 10 Redesigned Topology



The additional bandwidth is not free. Table 7 shows the incremental expense details. The complete financial picture is analyzed in the following section.

Table 7 Upgrade Expenses

San Jose to:	Original 64-kbps (Monthly Cost)	Redesigned 128-kbps (Monthly Cost)	Redesigned 192-kbps (Monthly Cost)	Incremental Cost (Monthly Cost)
Tokyo	$8,700	$13,400	–	$4,700
Tokyo to Hong Kong	$5,500	–	–	–
Chicago	$1,250	$2,050	–	$800
New York	$1,400	–	$3,100	$1,700
London	$6,400	–	$13,400	$7,000
London to Milan	$5,100	$8,150	–	$3,050
London to Frankfurt	$4,250	–	–	–
Total	–	–	–	**$17,250**

Cisco Multiservice Equipment

A Cisco 2610 modular access router was installed in the smaller branches and a Cisco 3640 modular access router was put in place in the two larger branches. At each of the smaller branch-offices, four key system FXO trunks were connected to the Cisco 2610 router (eight trunks to the Cisco 3640 in the larger branch locations). The reprogrammed key system would then preferentially direct approximately 95 percent of traffic destined for headquarters to one of the trunks connected to the Cisco router. The remaining overflow on-Net voice traffic, an estimated 5 percent, is directed to the PSTN.

The leased lines terminate at the San Jose headquarters, where the voice channels are decompressed and routed to the headquarter's PBX. Because 23 channels have been removed from the PSTN and are now sent over the router network, the San Jose headquarters can remove one of the PSTN T1 access lines.

The Cisco 2600s and Cisco 3600s transport the voice calls over the IP network compressed at 8-kbps, excluding overhead, using the G.729 CS-ACELP algorithm. Each voice channel utilizes a dedicated DSP to perform encoding and compression. The design of the Cisco 3600s and the dedicated DSPs enable the high performance and low latency that ensure very high voice quality.

To reliably deliver the real-time voice traffic, the Cisco IOS® software employs a number of techniques. Resource reservation protocol (RSVP) reserves bandwidth when the remote phone number is dialed. Compressed Real-Time Protocol (CRTP) compresses the overall header, thereby keeping both overhead low and payload throughput high.

It turns out that approximately 50 percent of normal voice conversation is silence. By not transmitting this silence, bandwidth is available for the data traffic. Cisco IOS software uses sophisticated, silence-suppression techniques to realize this savings, so that the receiver remains assured that the call remains connected, comfort noise is generated locally.

Commuting to work can be very difficult in Hong Kong and New York. Therefore, the Cisco 3600 routers will eventually be used as access servers for employees who will telecommute from their homes.

Note that not all key systems provide automatic route selection for preferential voice traffic switching. If plans call for this type of configuration, contact the key system or PBX vendor to confirm that this capability exists in the equipment.

In the next section the costs and savings resulting from the multiservice network migration are summarized.

Financial Analysis

Table 8 shows the capital costs for the branches and the headquarters. The required additional bandwidth indicated in the previous section cost US$17,250 per month. Comparing the savings to these additional expenses, the net monthly savings are illustrated in Table 9.

Table 8 Capital Costs

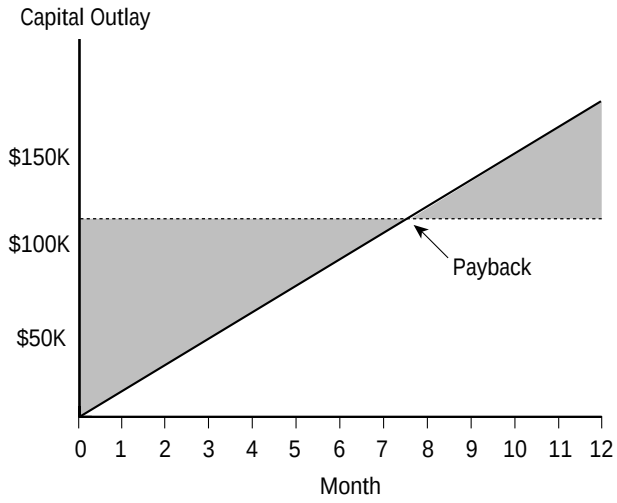
Equipment	Estimated Cost (in US dollars)
Cisco 2610 & 3640 Modular Access Routers	$77,000
Nine Branch Key System Modules ($700 each)	$6,300
One Headquarters PBX Trunk Module	$5,418
Total Capital Cost	$88,718

Table 9 Savings and Payback

Monthly PSTN Voice Savings (95% of total)	$30,710
E1 Removed from HQ to PSTN	$950
Required Added WAN Links	$(17,250)
Net Total Monthly Savings	$14,410
Net Total Annual Savings	$172,919
Capital Costs	$88,718
Installation (Estimate)	$15,000
Total Capital Costs	$103,718
Payback Period (Months)	7.19

This firm saves nearly US$175,000 per year by moving their internal voice traffic onto their router backbone. The payback period is less than eight months. The important numbers are summarized in Figure 11.

Figure 11 Financial Summary



Cost per Minute

An organization's voice cost per minute is a common expense measurement. The details required to calculate the original cost per minute are given in Table 4. The total on-Net minutes per month for all offices equals 107,267 minutes and the total monthly cost is US$32,326. Dividing US$32,326 by 107,267 yields an average cost per minute of US$0.30 for all calls between headquarters and the branches. To carry 95 percent of this traffic over the data network, that network must be upgraded at a cost of US$17,250 per month. Ninety-five percent of 107,267 equals 101,904 minutes. Dividing the monthly upgrade costs by the carried minutes yields an average cost per minute of US$0.17. In this case, the firm was able to reduce their average cost per minute from US$0.30 to US$0.17; a 43 percent reduction.

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems Europe s.a.r.l.
Parc Evolic, Batiment L1/L2
16 Avenue du Quebec
Villebon, BP 706
91961 Courtaboeuf Cedex
France
http://www-europe.cisco.com
Tel: 33 1 69 18 61 00
Fax: 33 1 69 28 83 26

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Headquarters
Nihon Cisco Systems K.K.
Fuji Building, 9th Floor
3-2-3 Marunouchi
Chiyoda-ku, Tokyo 100
Japan
http://www.cisco.com
Tel: 81 3 5219 6250
Fax: 81 3 5219 6001

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the Cisco Connection Online Web site at http://www.cisco.com/offices.

Argentina • Australia • Austria • Belgium • Brazil • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Czech Republic • Denmark • Dubai, UAE Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel • Italy • Japan • Korea • Luxembourg • Malaysia Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Singapore Slovakia • Slovenia • South Africa • Spain • Sweden • Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela

7/99 BW5228