150310

ADVANCED COMMUNICATION USING VOIP

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of
Dokuz Eylül University
In Partial Fulfillment of the Requirements for
the Degree of Master of Science in Electrical and Electronics Engineering

by
Tolga NARBAY

August, 2004
İZMİR

150 310

M.Sc THESIS EXAMINATION RESULT FORM

We certify that we have read this thesis and **"ADVANCED COMMUNICATION USING VoIP"** completed by **TOLGA NARBAY** under supervision of **ASSOC. PROF. DR. ZAFER DİCLE** and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc.Prof Dr. Zafer DİCLE

Supervisor

Doc Dr. Yalçın ÇEBİ

(Committee Member)

Yrd. Doç. Dr. Hacer Özturk

(Committee Member)

Approved by the

Graduate School of Natural and Applied Sciences

Prof.Dr. Cahit HELVACI

Director

ACKNOWLEDGEMENTS

I am very grateful to Assoc.Prof Dr. Zafer Dicle in answering questions and providing guidance as to the proper path both for this thesis and my knowledge of VoIP .

I am also grateful to Utku Ergül , for his willingness to help and great assisstance for the accomplishment of this thesis.

ABSTRACT

The consolidation of voice networks and data networks offers an opportunity for major savings. Accordingly, the challenge of integrating voice and data networks is becoming a rising priority for many network owners. Organizations are looking for solutions which will enable them to take advantage of capacity on broadband networks for voice and data transmission, as well as utilize the Internet and company Intranets as alternatives to more costlier dedicated mediums .

Voice over IP (VoIP) means the transmission of voice traffic in packets . Voice over IP meets the challenges of combining legacy voice networks and packet networks by allowing both voice and signaling information to be transported over the packet network.

Providing high quality telephony over IP networks is one of the key steps in the convergence of voice, fax, video, and data communications services. Adding voice to packet networks requires an understanding of how to deal with system level challenges such as interoperability, packet loss, delay, scalability, and reliability.

This thesis explains how telephony infrastructure is built and works today, major concepts concerning voice and data networking, transmission of voice over data, and IP signaling protocols used to interwork with current telephony systems. It also answers the questions about IP , the process of voice signalling in telephone networks today , various IP signalling protocols , the various caveats of converging voice and data networks and how does QoS (Quality of Service) ensure good voice quality in a network.

Keywords : VoIP , IP Signalling protocols, Packet switched transmission , Circuit switched transmission ,QoS in VoIP

ÖZET

Ses ve data ağlarının birleştirilmesi büyük tasarruflar için fırsat sunar. Bundan dolayı, pek çok ağ sahibi için ses ve data ağlarının birleştirilmesi yükselen bir öncelikte yer almaya başlamıştır. Organizasyonlar, Internet ve şirket intranetlerinden daha az masraflı alternatifler olarak yararlanmanın yanında ağlarındaki kapasiteden ses ve data iletimi içinde yararlanmalarını sağlayacak yeni çözüm arayışı içindedirler.

VOIP , IP üzerinden ses trafiğinin paketler içinde iletimi demektir. IP üzerinden ses, ses ve sinyalleşme bilgisinin paket ağı üzerinden taşınmasına olanak sağlayarak ses ağları ve paket ağlarının birleştirilmesi görevini yerine getirir.

IP ağları üzerinden yüksek kalitede telefon servisi sağlamak, ses, fax, video ve data iletişim servislerinin birleştirilmesinde anahtar adımlardan biridir. Paket ağlarına sesin eklenmesi birlikte çalışma, paket kaybı, gecikme, ölçülebilirlik, güvenilirlik gibi sistem seviyesindeki görevlerin de nasıl yapılabileceğinin de anlaşılmasını gerektirir.

Bu tez, telefon altyapısının ne şekilde oluşturulduğunu ve nasıl çalıştığını, ses ve data ağları için ana kavramları , sesin data ile beraber iletimini, ve mevcut telefon sistemleri ile beraber çalışması için kullanılan IP sinyalleşme protokollerini ele almaktadır. Aynı zamanda IP, mevcut telefon ağlarındaki ses sinyalleşme işlemleri, çeşitli IP sinyalleşme protokolleri, çeşitli ses ve data ağlarını birleştirme çalışmaları ve Servis Kalitesi'nin (QoS) ağda iyi ses kalitesini nasıl sağladığı hakkındaki soruları cevaplamaktadır.

Anahtar Kelimeler : Ses iletişimi , IP Sinyalleşme Protokolleri , Paket iletimli haberleşme, Devre iletimli haberleşme , VoIP Servis Kalitesi

CONTENTS

Chapter Two

ENTERPRISE TELEPHONY TODAY

Chapter Three

BASIC TELEPHONY SIGNALLING

Chapter Four

VOICE OVER IP BENEFITS AND APPLICATIONS

Chapter Five
IP TUTORIAL

Chapter Six
VOIP: AN IN-DEPTH ANALYSIS

Chapter Seven

QUALITY OF SERVICE

Chapter Eight

VoIP CONFIGURATION ISSUES

Chapter Nine

MARKET FORECASTS AND FUTURE

Chapter Ten
APPLICATION OF VOIP

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER ONE

INTRODUCTION

1 Introduction

The concept of Voice over IP has been in focus for the last couple of years , growing to be a major technology. Voice over IP is the concept of being able to use one single network, the TCP/ IP network, as the common carrier for not only data, but also voice, and eventually any other communication form.

With the same intension, we thought that the challenge of integrating voice and data networks for Dokuz Eylül University Campuses , which are apart , will enable us to take the advantage of capacity on backbone data networks to be also used for voice transmission. Dedicated medium for data transmission will provide us an alternative to a more costlier dedicated medium like Turk Telekom's PSTN network . The details of this work is explained later in the thesis.

This introductary part contrasts the similarities and differences between traditional TDM networks and networks running packetized voice.

1.1 The Beginning of the PSTN

The first voice transmission, sent by Alexander Graham Bell, was accomplished in 1876 through what is called a ring-down circuit. A ring-down circuit means that there was no dialing of numbers, Instead, a physical wire connected two devices. Basically, one person picked up the phone and another person was on the other end .

Over time, this simple design evolved from a one-way voice transmission, by which only one user could speak, to a bi-directional voice transmission, whereby both users could speak. Moving the voices across the wire required a carbon microphone, a battery, an electromagnet, and an iron diaphragm. It also required a physical cable between each location that the user wanted to call. The concept of dialing a number to reach a destination, however, did not exist at this time.

To further illustrate the beginnings of the PSTN, see the basic four-telephone network shown in Figure 1.1. As you can see, a physical cable exists between each location.

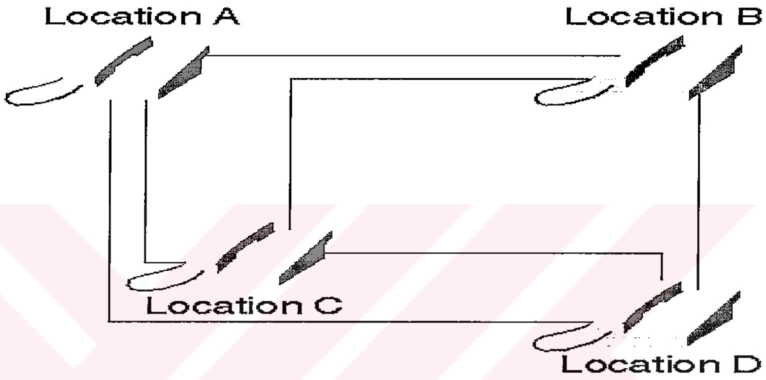

Figure 1. 1 Basic Four-Phone Network

Place a physical cable between every household requiring access to a telephone, however, and you'll see that such a setup is neither cost-effective nor feasible. To determine how many lines you need to your house, think about everyone you call as a value of N and use the following equation: N x (N–1)/2. As such, if you want to call 10 people, you need 45 pairs of lines running into your house.

Due to the cost concerns and the impossibility of running a physical cable between everyone on Earth who wanted access to a telephone, another mechanism was developed that could map any phone to another phone. With this device, called a switch , the telephone users needed only one cable to the centralized switch office.

At first, a telephone operator acted as the switch. This operator asked callers where they wanted to dial and then manually connected the two voice paths.After while the human switch is replaced by electronic switches. At this point, you can learn how the modern PSTN network is built.

1.2 Understanding PSTN Basics

Although it is difficult to explain every component of the PSTN, this section explains the most important pieces that make the PSTN work. The following sections discuss how your voice is transmitted across a digital network, basic circuit-switching concepts, and why your phone number is 10 digits long.

1.2.1 Analog and Digital Signaling

Everything you hear, including human speech, is in analog form. Until several decades ago, the telephony network was based on an analog infrastructure as well.

Although analog communication is ideal for human interaction, it is neither robust nor efficient at recovering from line noise. (Line noise is normally caused by the introduction of static into a voice network.) In the early telephony network, analog transmission was passed through amplifiers to boost the signal. But, this practice amplified not just the voice, but the line noise as well. This line noise resulted in an often unusable connection.

Analog communication is a mix of time and amplitude. Figure 1.2 , which takes a high-level view of an analog waveform, shows what your voice looks like through an oscilloscope.

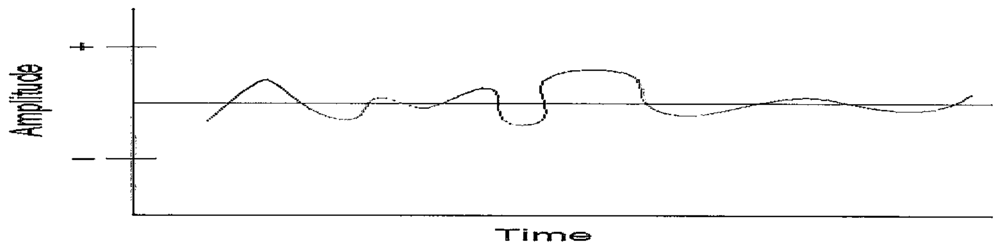

Figure 1. 2 Analog Waveform

If you were far away from the end office switch (which provides the physical cable to your home), an amplifier might be required to boost the analog transmission (your voice). Analog signals that receive line noise can distort the analog waveform and cause garbled reception. This is more obvious to the listener if many amplifiers are located between your home and the end office switch. Figure 1.3 shows that an amplifier does not clean the signal as it amplifies, but simply amplifies the distorted signal. This process of going through several amplifiers with one voice signal is called accumulated noise .

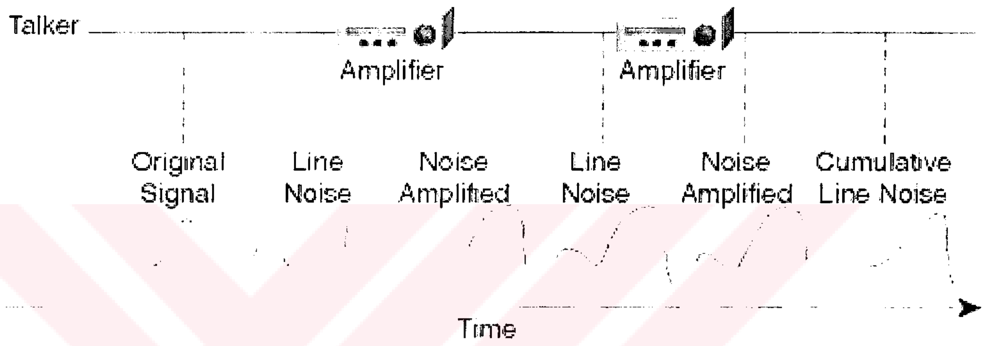

Figure 1. 3 Analog Line Distortion

In digital networks, line noise is less of an issue because repeaters not only amplify the signal, but clean it to its original condition. This is possible with digital communication because such communication is based on 1s and 0s. So, as shown in Figure 1.4, the repeater (a digital amplifier) only has to decide whether to regenerate a 1 or a 0.

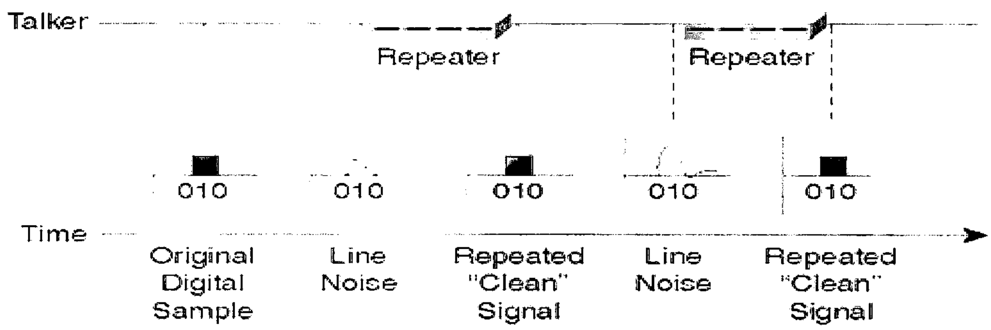

Figure 1. 4 Digital Line Distortion

Therefore, when signals are repeated, a clean sound is maintained. When the benefits of this digital representation became evident, the telephony network migrated to pulse code modulation (PCM).

1.2.2 Digital Voice Signals

PCM is the most common method of encoding an analog voice signal into a digital stream of 1s and 0s. All sampling techniques use the Nyquist theorem , which basically states that if you sample at twice the highest frequency on a voice line, you achieve good-quality voice transmission.

The PCM process is as follows: Analog waveforms are put through a voice frequency filter to filter out anything greater than 4000 Hz. These frequencies are filtered to 4000 Hz to limit the amount of crosstalk in the voice network. Using the Nyquist theorem, you need to sample at 8000 samples per second to achieve good-quality voice transmission. The filtered analog signal is then sampled at a rate of 8000 times per second.(Ericsson Training Book,2000)

After the waveform is sampled, it is converted into a discrete digital form. This sample is represented by a code that indicates the amplitude of the waveform at the instant the sample was taken. The telephony form of PCM uses eight bits for the code and a logarithm compression method that assigns more bits to lower-amplitude signals.

If you multiply the eight-bit words by 8000 times per second, you get 64,000 bits per second (bps). The basis for the telephone infrastructure is 64,000 bps (or 64 kbps).

Two basic variations of 64 kbps PCM are commonly used: μ-law, the standard used in North America; and a-law, the standard used in Europe. The methods are similar in that both use logarithmic compression to achieve from 12 to 13 bits of linear PCM quality in only eight-bit words, but they differ in relatively minor details.

The μ-law method has a slight advantage over the a-law method in terms of low-level signal-to-noise ratio performance, for instance.

1.2.3 Local Loops, Trunks, and Interswitch Communication

The telephone infrastructure starts with a simple pair of copper wires running to your home. This physical cabling is known as a local loop . The local loop physically connects your home telephone to the central office switch (also known as a Class 5 switch or end office switch).

The communication path between the central office switch and your home is known as the phone line, and it normally runs over the local loop.

The communication path between several central office switches is known as a trunk . Just as it is not cost-effective to place a physical wire between your house and every other house you want to call, it is also not cost-effective to place a physical wire between every central office switch.

Switches are currently deployed in hierarchies. End office switches (or central office switches) interconnect through trunks to tandem switches (also referred to as Class 4 switches). Higher-layer tandem switches connect local tandem switches. Figure 1.5 shows a typical model of switching hierarchy.

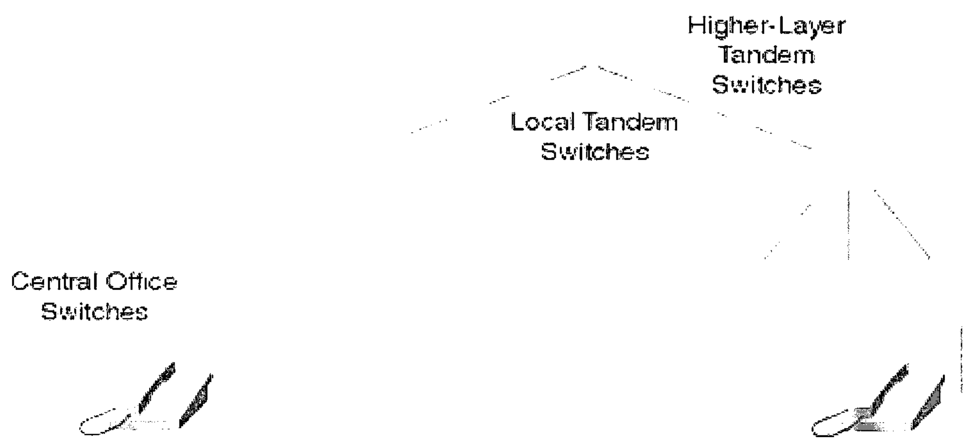

Figure 1. 5 Circuit-Switching Hierarchy

Central office switches often directly connect to each other. Where the direct connections occur between central office switches depends to a great extent on call patterns. If enough traffic occurs between two central office switches, a dedicated circuit is placed between the two switches to offload those calls from the local tandem switches. Some portions of the PSTN use as many as five levels of switching hierarchy. .(Ericsson Training Book,2000)

Now that you know how and why the PSTN is broken into a hierarchy of switches, you need to understand how they are physically connected, and how the network communicates.

1.2.4 PSTN Signaling

Generally, two types of signaling methods run over various transmission media. The signaling methods are broken into the following groups:

- **User-to-network signaling**— This is how an end user communicates with the PSTN.
- **Network-to-network signaling**— This is generally how the switches in the PSTN intercommunicate.

1.2.4.1 User-to-Network Signaling

Generally, when using twisted copper pair as the transport, a user connects to the PSTN through analog, Integrated Services Digital Network (ISDN), or through a E1 carrier.

The most common signaling method for user-to-network analog communication is Dual Tone Multi-Frequency (DTMF) . DTMF is known as in-band signaling because the tones are carried through the voice path. Figure 1.6 shows how DTMF tones are derived.

Dual Tone Multi-Frequency	1209	1336	1477	1633
697	1	2	3	A
770	4	5	6	B
852	7	8	9	C
941	*	0	#	D

Figure 1. 6 Dual Tone Multi-Frequency

When you pick up your telephone handset and press the digits (as shown in Figure 1.6), the tone that passes from your phone to the central office switch to which you are connected tells the switch what number you want to call.

ISDN uses another method of signaling known as out-of-band . With this method, the signaling is transported on a channel separate from the voice. The channel on which the voice is carried is called a bearer (or B channel) and is 64 kbps. The channel on which the signal is carried is called a data channel (D channel) and is 16 kbps. Figure 1.7 shows a Basic Rate Interface (BRI) that consists of two B channels and one D channel.

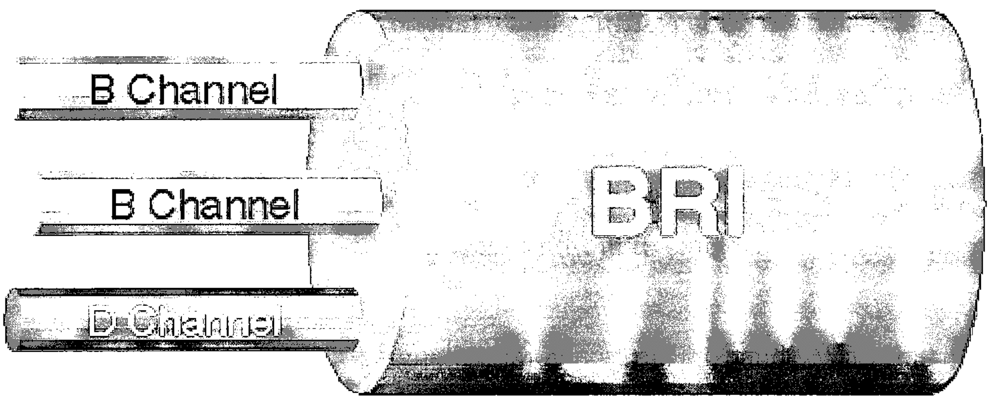

Figure 1. 7 Basic Rate Interface

Out-of-band signaling offers many benefits, including the following:

- Signaling is multiplexed (consolidated) into a common channel.
- Glare is reduced (glare occurs when two people on the same circuit seize opposite ends of that circuit at the same time).
- A lower post dialing delay.
- Additional features, such as higher bandwidth, are realized.
- Because setup messages are not subject to the same line noise as DTMF tones, call completion is greatly increased.

In-band signaling suffers from a few problems, the largest of which is the possibility for lost tones . This occurs when signaling is carried across the voice path and it is a common reason why you can sometimes experience problems remotely accessing your voice mail.

1.2.4.2 Network-to-Network Signaling

Network-to-network communication is normally carried across the following transmission media:

- T1/E1 carrier over twisted pair
  T1 is a 1.544-Mbps digital transmission link normally used in North America and Japan.
  E1 is a 2.048-Mbps digital transmission link normally used in Europe.
- T3/E3, T4 carrier over coaxial cable
  T3 carries 28 T1s or 672 64-kbps connections and is 44.736 Mbps.
  E3 carries 16 E1s or 512 64-kbps connections and is 34.368 Mbps.
  T4 handles 168 T1 circuits or 4032 4-kbps connections and is 274.176 Mbps.
- T3, T4 carrier over a microwave link
- Synchronous Optical Network (SONET) across fiber media
  SONET is normally deployed in OC-3, OC-12, and OC-48 rates, which are 155.52 Mbps, 622.08 Mbps, and 2.488 Gbps, respectively.

Network-to-network signaling types include in-band signaling methods such as Multi-Frequency (MF) and Robbed Bit Signaling (RBS). These signaling types can also be used to network signaling methods.

Digital carrier systems (T1, T3) use A and B bits to indicate on/off hook supervision. The A/B bits are set to emulate Single Frequency (SF) tones (SF typically uses the presence or absence of a signal to signal A/B bit transitions).

These bits might be robbed from the information channel or multiplexed in a common channel (the latter occurs mainly in Europe). More information on these signaling types is found in Chapter 3, "Basic Telephony Signaling."

MF is similar to DTMF, but it utilizes a different set of frequencies. As with DTMF, MF tones are sent in-band. But, instead of signaling from a home to an end office switch, MF signals from switch to switch.

Network-to-network signaling also uses an out-of-band signaling method known as Signaling System 7 (SS7) (or C7 in European countries). SS7 is beneficial because it is an out-of-band signaling method and it interconnects to the Intelligent Network (IN). Connection to the IN enables the PSTN to offer Custom Local Area Signaling Services (CLASS) services.

SS7 is a method of sending messages between switches for basic call control and for CLASS. These CLASS services still rely on the end-office switches and the SS7 network. SS7 is also used to connect switches and databases for network-based services (for example, 800-number services and Local Number Portability [LNP]).

Some of the benefits of moving to an SS7 network are as follows:

- Reduced post-dialing delay: There is no need to transmit DTMF tones on each hop of the PSTN. The SS7 network transmits all the digits in an initial setup message that includes the entire calling and called number. When using in-band signaling, each MF tone normally takes 50 ms to transmit. This means you have at least a .5-second post-dialing delay per PSTN hop. This number is based on 11-digit dialing (11 MF tones x 50 ms = 550 ms).
- Increased call completion: SS7 is a packet-based, out-of-band signaling protocol, compared to the DTMF or MF in-band signaling types. Single packets containing all the necessary information (phone numbers, services, and so on) are transmitted faster than tones generated one at a time across an in-band network.
- Connection to the IN: This connection provides new applications and services transparently across multiple vendors' switching equipment as well as the capability to create new services and applications more quickly.

To further explain the PSTN, visualize a call from house A to house B. This call traverses an end office switch, the SS7 network (signaling only), and a second end office switch. Figure 1.8 displays the call flow from this house to the house away .

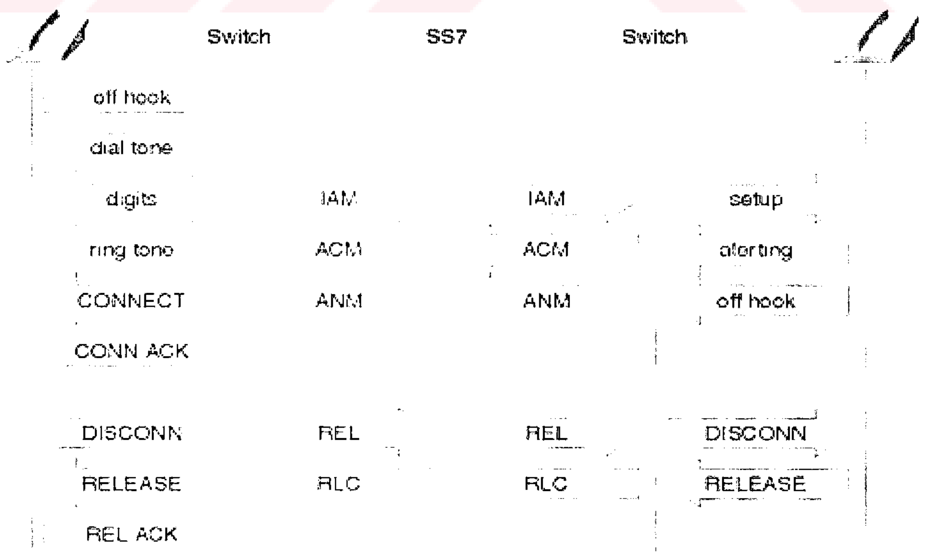

Figure 1. 8 PSTN Call Flow from House A to House B

To better explain the diagram in Figure 1.8 , it is explained step by step below;

1. The phone A is picked up and an off-hook indication to the end office switch is sent.
2. The switch sends back a dial tone.
3. The digits to call the house B is dialled (they are sent in-band through DTMF).
4. The switch interprets the digits and sends an Initial Address Message (IAM, or setup message) to the SS7 network.
5. The SS7 network reads the incoming IAM and sends a new IAM to the other switch.
6. The switch of B sends a setup message to phone B (it rings the phone).
7. An alerting message (alerting is the same as the phone ringing) is sent from B switch (not from B phone) back to the SS7 network through an Address Complete Message (ACM).
8. The SS7 network reads the incoming ACM and generates an ACM to switch A.
9. A ringing sound can be heard from A indicating that Phone B is ringing (The ringing is not synchronized; your local switch normally generates the ringing when the ACM is received from the SS7 network.)
10. Phone B is picked up, sending an off-hook indication to switch B.
11. Switch B sends an ANswer Message (ANM) that is read by the SS7, and a new ANM is generated to switch A.
12. A connect message is sent to phone A (only if it's an ISDN phone) and a connect acknowledgment is sent back (again, only if it's an ISDN phone). (If it is not an ISDN phone, then on-hook or off-hook representations signal the end office switch.)
13. Phone A can now talk to Phone B until the phone is hung up (on-hook indication).

If phone B was busy, phone A could use an IN feature by which it could park on phone B's line and have the PSTN call phone A back after phone B got off the phone.

The next section discusses services and applications that are common in the PSTN.

1.3 PSTN Services and Applications

As with almost every industry, it is usually better and easier to acquire additional business from current customers than it is to go out and get new customers. The PSTN is no different. Local Exchange Carriers (LECs) have been increasing the features they offer to create a higher revenue stream per consumer.

Numerous services are now available, for example, which were not available just a few years ago. These services come in two common flavors: custom calling features and CLASS features.

Custom calling features rely upon the end office switch, not the entire PSTN, to carry information from circuit-switch to circuit-switch. CLASS features, however, require SS7 connectivity to carry these features from end to end in the PSTN.

The following list includes a few of the popular custom calling features commonly found in the PSTN today:

- Call waiting—Notifies customers who already placed a call that they are receiving an incoming call.
- Call forwarding—Enables a subscriber to forward incoming calls to a different destination.
- Three-way calling—Enables conference calling.

With the deployment of the SS7 network, advanced features can now be carried end to end. A few of the CLASS features are mentioned in the following list:

- Display—Displays the calling party's directory number, or Automatic Number Identification (ANI).
- Call blocking—Blocks specific incoming numbers so that callers are greeted with a message saying the call is not accepted.
- Calling line ID blocking—Blocks the outgoing directory number from being shown on someone else's display. (This does not work when calling 800-numbers or certain other numbers.)
- Automatic callback—Enables you to put a hold on the last number dialed if a busy signal is received, and then place the call after the line is free.
- Call return (*69)—Enables users to quickly reply to missed calls.

A majority of these features are possible due to the use of SS7 and the IN. Many inter-exchange carriers (IXCs) also offer business features, such as the following:

- Circuit-switched long distance—Basic long-distance services (normally at a steeply discounted rate).
- Calling cards—Pre-paid and post-paid calling cards. You dial a number, enter a password, and then call your destination.
- 800/888/877 numbers—The calling party is not charged for the call; Rather, the party called is charged (normally at a premium rate).
- Virtual Private Networks (VPNs)—The telephone company manages a private dialing plan. This can greatly reduce the number of internal Information Service (IS) telecommunications personnel.
- Private leased lines—Private leased lines from 56 kbps to OC-48s enable both data and voice to traverse different networks. The most popular speed by far in North America is T1.
- Virtual circuits (Frame Relay or Asynchronous Transfer Mode [ATM])—The tele-phone carrier (IXC or LEC) switches your packets. It does this packet by packet (or cell by cell in ATM), not based upon a dedicated circuit.

This list of IXC business features is merely a sampling of the more popular features and applications available in the PSTN. Although the PSTN is evolving and consumers are using more of its features, the basic user experience has remained somewhat consistent since the inception of digital networking for telephony communications.

1.3.1 PSTN Numbering Plans

One feature that slowly changed over time is the dial plan. The addition of second lines for Internet access, cell phones, and fax machines has created a relative shortage of phone numbers. The next section delves into how the PSTN dial plan is put together and what you can expect over the next few years.

In some places in the United States, it is necessary to dial 1+10 digits for even a local call. This will become more and more prevalent as more devices require telephone numbers. The need to dial 1+10 digits for a local number is normally due to an overlay. An overlay can result in next-door neighbors having different area codes.

An overlay is when a region with an existing area code has another area code "overlayed." This offers the existing customers the benefits of not having to switch area codes, but forces everyone in that region to dial 10 digits to call anywhere.

Essentially, two numbering plans are used with the PSTN: the North American Numbering Plan (NANP) and the International Telecommunication Union Telecommunication Standardization Sector (ITU-T; formerly CCITT) International Numbering Plan. They are discussed in the following sections.

1.3.1.1 NANP

NANP is an 11-digit dialing plan that contains three parts: the Numbering Plan Area (NPA, also referring to as area code), Central Office Code (NXX), and Station Number. This plan is often referred to as NPA-NXX-XXXX.

NPA codes use the following format ; XX, where N is a value between 2–9 and X is a value between 0–9.

NANP is also referred to as 1+10. This means that when a 1 is the first number dialed, it will be proceeded by a 10-digit NPA-NXX-XXXX number. This enables the end office switch to determine whether it should expect a 7- or 10-digit telephone number.

End office switch keeps track of what long-distance provider is used in a static table. Each long-distance carrier has a code. This long-distance code is assigned by the North American Numbering Plan Association (NANPA) and is added to the number you call so that it is routed to the proper long-distance network carrier (or IXC).

1.3.1.2 ITU-T International Numbering Plan

ITU-T Recommendation E.164 specifies that a Country Code (CC), National Destination Code (NDC), and Subscriber Number (SN) be used to route a call to a specific subscriber.

The CC consists of one, two, or three digits. The first digit (1–9) defines world numbering zones. A list of all the defined CCs is found in ITU-T Recommendation E.164 Annex A.

NDC and SN vary in length based on the needs of the country. Neither one has more than 15 digits.Many other recommendations and specifications for international number plans are found in the E. recommendations from the ITU-T.

Although dial plans might not seem extremely important at the moment, they are crucial to the successful deployment and implementation of Voice over IP (VoIP) or traditional circuit-switched networks.

Turkey has chosen to use the ITU-T international numbering plan . For example , a number in İzmir has 90 as Country Code (CC-Turkey) , 232 as National Destination Code (NDC-İzmir) and a 7-digit number as Subscriber Number (SN-İzmir) to be used to route a call to a specific subscriber.

1.4 Drivers Behind the Convergence Between Voice and Data Networking

Understanding PSTN basics includes knowing why the existing PSTN does not fit all the needs of its builders or users. After you understand where today's PSTN is lacking, you will know where to look to find a solution. This section sets the stage for why the voice and data networks are merging into a signal network.

1.4.1 Drawbacks to the PSTN

Although the PSTN is effective and does a good job at what it was built to do (that is, switch voice calls), many business drivers are striving to change it to a new network, whereby voice is an application on top of a data network. This is happening for several reasons:

- Data has overtaken voice as the primary traffic on many networks built for voice.

  Data is now running on top of networks that were built to carry voice efficiently. Data has different characteristics, however, such as a variable use of bandwidth and a need for higher bandwidth. Soon, voice networks will run on top of networks built with a data-centric approach. Traffic will then be differentiated based upon application instead of physical circuits. New technologies (such as Fast Ethernet, Gigabit Ethernet, and Optical Networking) will be used to deploy the high-speed networks that needed to carry all this additional data.
- The PSTN cannot create and deploy features quickly enough.

  With increased competition due to deregulation in many telecommunications markets, LECs are looking for ways to keep their existing clientele. The primary method of keeping customers is by enticing them through new services and applications.The PSTN is built on an infrastructure whereby only the vendors of the equipment develop the applications for that equipment. This means you have one-stop shopping for all your needs. It is very difficult for one company to meet all the needs of a customer. A more open infrastructure, by which many vendors can provide applications, enables more creative solutions and applications to be developed.

It is also not possible with the current architecture to enable many vendors to write new applications for the PSTN. Imagine where the world would be today if vendors, such as Microsoft, did not want other vendors to write applications for its software.

- Data/Voice/Video (D/V/V) cannot converge on the PSTN as currently built.

  With only an analog line to most homes, you cannot have data access (Internet access), phone access, and video access across one 56-kbps modem. High-speed broadband access, such as digital subscriber line (DSL), cable, or wireless, is needed to enable this convergence. After the last bandwidth issues are resolved, the convergence can happen to the home. In the backbone of the PSTN, the convergence has already started.

- The architecture built for voice is not flexible enough to carry data.

  Because the bearer channels (B channels and T1 circuits), call-control (SS7 and Q.931), and service logic (applications) are tightly bound in one closed platform, it is not possible to make minor changes that might improve audio quality.

It is also important to note that circuit-switched calls require a permanent 64-kbps dedicated circuit between the two telephones. Whether the caller or the person called is talking, the 64-kbps connection cannot be used by any other party. This means that the telephone company cannot use this bandwidth for any other purpose and must bill the parties for consuming its resources.

Data networking, on the other hand, has the capability to use bandwidth only when it is required. This difference, although seemingly small, is a major benefit of packet-based voice networking.

1.4.2 Telecommunications Deregulation

So far, we have looked at the technical issues of how the PSTN operates, the basic hierarchy, and why we might need to converge voice and data networks. One important reason for this convergence is more political than technical.

Various countries throughout Europe, Asia, and the Americas are opening up their telecommunications markets to competition. In addition, in some cases, they are selling off the existing government-run telephone carriers to a private company (or many companies).

In many countries as in Turkey , the government ran the PSTN. This is changing as governments realize that communication is important to survival in the next century. These governments also realize that with communication comes knowledge, and with knowledge comes strength and prosperity.

Many new voice carriers are rushing to join these new deregulated markets. With the influx of fresh competition, pricing models are changing, and new, as well as old, carriers are considering deploying the latest technology to lower the cost of doing business.

The additional advantage of deploying new technology is the ability to offer value-added services and deploy these new services in a short amount of time. Services include bundled voice and Internet access, unified communications, Internet call waiting, and others.

1.5 Packet Telephony Network Drivers

The previous section discussed political drivers for competition in the PSTN. This section explains why a carrier might choose to develop a packet telephony network in lieu of a traditional circuit-switching network.

The integration of D/V/V is more than just a change in infrastructure. D/V/V integration also enables new features to be developed more quickly and opens up application development to thousands of Independent Software Vendors (ISVs). You can compare this integration of D/V/V to the change from mainframe computers, for which very few vendors developed applications, to client/servers, for which multiple vendors developed applications for distributed systems.

Figure 1.9 shows how the circuit-switching model is breaking into a new model by which open standards exist between all three layers. A packet infrastructure will carry the actual voice (media), the call-control layer will be separate from the media layer, and open APIs (Application Programming Interfaces) will enable new services to be created by ISVs.

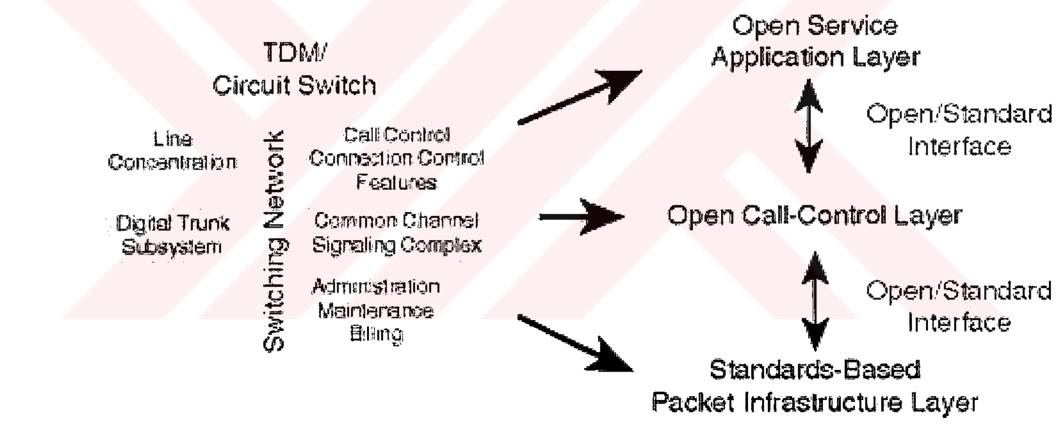

Figure 1. 9 Circuit Switching Versus Packet Switching

Figure 1.9 is an over-simplification of the changes that are actually happening. To further discuss these changes, you need to take a closer look at each of the three layers.

1.5.1 Standards-Based Packet Infrastructure Layer

The packet infrastructure replaces the circuit-switching infrastructure in this new model. This infrastructure most likely will be IP, although this model also works if

ATM is the underlying transport and IP rides across the top. IP is so attractive as the packet infrastructure because of its ubiquitous nature and the fact that it is the de facto application interface. This means that software applications running over IP do not have to be known. IP simply transports the data end to end, with no real interest in the payload.

Real-time Transport Protocol (RTP) is utilized in addition to a User Datagram Protocol (UDP)/IP header to provide timestamping . RTP runs atop UDP and IP and is commonly noted as RTP/UDP/IP. RTP is currently the cornerstone for carrying real-time traffic across IP networks. (Microsoft Netmeeting, for instance, utilizes RTP to carry audio and video communications.) To date, all VoIP signaling protocols utilize RTP/UDP/IP as their transport mechanism for voice traffic.

Often, RTP packet flows are known as RTP streams. This nomenclature is used to describe the audio path.

In IP networks, it is common and normal for packet loss to occur. In fact, Transmission Control Protocol/Internet Protocol (TCP/IP) was built to utilize packet loss as a means of controlling the flow of packets. In TCP/IP, if a packet is lost, it is retransmitted. In most real-time applications, retransmission of a packet is worse than receiving a packet due to the time-sensitive nature of the information.

The ITU-T recommends a one-way delay of no more than 150 ms. In a vendors VoIP network, the unidirectional delay might be 120 ms (currently, 65 ms to 85 ms of that 120-ms delay is derived from two VoIP gateways when using G.729). If the receiving station must request that a packet be re-transmitted, the delay will be too large, and large gaps and breaks in the conversation will occur.

RTP has a field that stamps the exact time the packet was sent (in relation to the entire RTP stream). This information is known as RTP timestamps and is used by the device terminating/receiving the audio flow. The receiving device uses the RTP timestamps to determine when a packet was expected, whether the packet was in order, and whether it was received when expected. All this information helps the receiving station determine how to tune its own settings to mask any potential network problems such as delay, jitter, and packet loss.

One of the main benefits of IP is the fact that properly built IP networks are self-healing . This means that because dynamic routing protocols are used and multiple possible destinations exist, a network can re-converge based upon the best route. It also means that it is possible for your voice (packetized in IP) to take multiple paths to the same destination. Currently you cannot nail up a single path between two destinations. Each individual packet takes the best path between sender and receiver. The fact that the packet layer is based upon open standards enables multiple vendors to provide solutions that are interoperable.

These open standards enable increased competition at this packet layer. The ITU-T, Internet Engineering Task Force (IETF), European Telecommunication Standards Institute (ETSI), and EIA-TIA are only a few of the standards bodies you might be familiar with.

One key component of having a standards-based packet infrastructure is the ability to have open standards to layers at the call-control layer. Referring to Figure 1.9, these open standards are provided by protocols such as H.323, SGCP, MGCP, SIP, and so on, which have open defined interfaces and are widely deployed into the packet infrastructure. One of the jobs of the call-control protocol is to tell the RTP streams where to terminate and where to begin. Call-control accomplishes this task by translating between IP addresses and phone numbering plans.

1.5.2 Open Call-Control Layer

Call-control, in a nutshell, is the process of making a routing decision about where a call needs to go and somehow making the call happen. In the PSTN today, these decisions are carried out by SS7 and are made by Service Control Points (SCPs). In the next chapters ,we will discuss how the different VoIP protocols work and how they solve different network design issues.

In this new model of separating the bearers (RTP streams) from the call-control layer and separating the call-control layer from the services, it is necessary to make sure that standards-based protocols are used. Data networking is unique in the fact that multiple protocols can co-exist in a network and you can tailor them to the particular needs of the network.

Many different IP routing protocols exist, for example, and each is specifically designed for a certain type of network. Some of these include the Router Information Protocol (RIP), Interior Gateway Routing Protocol (IGRP), Enhanced Interior Gateway Routing Protocol (EIGRP), Intermediary System to Intermediary System (IS-IS), Open Shortest Path First (OSPF), and Border Gateway Protocol (BGP). Each protocol solves a similar problem—routing updates.

Each routing problem is slightly different, however, and requires a different tool. In this case, the tool is a routing protocol, which solves each problem.

You can say the same of VoIP call-control protocols. They all solve a similar problem—phone numbering to IP address translation; however, they might all be used for slightly different purposes.

For instance, currently H.323 is the most widely deployed VoIP call-control protocol. H.323, however, is not widely seen as a protocol that is robust enough for PSTN networks. For these networks, other protocols such as Media Gateway Control Protocol (MGCP) and Session Initiation Protocol (SIP) are being developed.

Many VoIP call-control protocols are being developed, so it is possible that different protocols will be deployed throughout the coming years. Each protocol will be developed to fix a certain problem and serve a particular purpose. A leader will emerge from the mud, but only if there must be a winner. For the short term, at least, many protocols will be used, and there will be no need for a single call-control protocol.

1.5.2.1 VoIP Call-Control Protocols

As of this writing, the main VoIP call-control protocols are H.323, Simple Gateway Control Protocol (SGCP), Internet Protocol Device Control (IPDC), MGCP, and SIP. They are defined as follows:

- H.323 is the ITU-T recommendation with the largest installed base, simply because it has been around the longest and no other protocol choices existed before H.323
- SGCP was developed starting in 1998 to reduce the cost of endpoints (gateways) by having the intelligent call-control occur in a centralized platform (or gateway controller).
- IPDC is very similar to SGCP, but it has many other mechanisms for operations, administration, management, and provisioning (OAM&P) than SGCP. OAM&P is crucial to carrier networks because it covers how they are maintained and deployed.
- SIP is being developed as a media-based protocol that will enable end devices (endpoints or gateways) to be more intelligent, and enable enhanced services down at the call-control layer.

1.5.2.2 H.323

H.323 is an ITU-T recommendation that specifies how multimedia traffic is carried over packet networks. H.323 utilizes existing standards (Q.931, for example) to accomplish its goals. H.323 is a rather complex protocol that was not created for simple development of applications. Rather, it was created to enable multimedia applications to run over "unreliable" data networks. Voice traffic is only one of the applications for H.323. Most of the initial work in this area focused on multimedia applications, with video and data-sharing a major part of the protocol.

Applications require significant work if they are to be scalable with H.323; for example, to accomplish a call transfer requires a separate specification (H.450.2). SGCP and MGCP, on the other hand, can accomplish a call transfer with a simple command, known as a modify connection (MDCX), to the gateway or endpoint. This simple example represents the different approaches built into the protocol design itself—one tailored to large deployment for simple applications (MGCP), and the other tailored to more complicated applications but showing limitations in its scalability (H.323).

To further demonstrate the complexity of H.323, Figure 1.10 shows a call-flow between two H.323 endpoints.

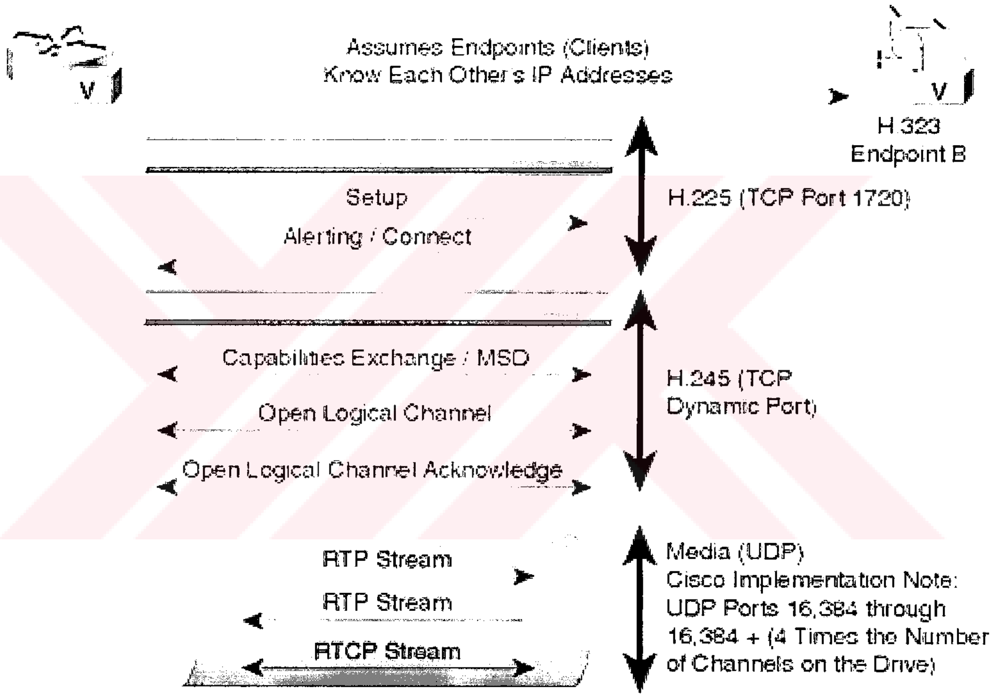

Figure 1. 10 H.323 Call-Flow

Figure 1.10 illustrates the most basic H.323 call-flow. In most cases, more steps are needed because gatekeepers are involved.

To better explain Figure 1.10, let's step through the call-flow:

1. Endpoint A sends a setup message to Endpoint B on TCP Port 1720.

2. Endpoint B replies to the setup message with an alerting message and a port number to start H.245 negotiation.
3. H.245 negotiation includes codec types (G.729 and G.723.1), port numbers for the RTP streams, and notification of other capabilities the endpoints have.
4. Logical channels for the UDP stream are then negotiated, opened, and acknowledged.
5. Voice is then carried over RTP streams.
6. Real Time Transport Control Protocol is used to transmit information about the RTP stream to both endpoints.

This call-flow is based on H.323 v1. H.323 v2, however, enables H.245 negotiation to be tunneled in the H.225 setup message. This is known as *fast-start* , and it cuts down on the number of roundtrips required to set up an H.323 call. It does not, however, make the protocol any less complex.

1.5.3 Open Service Application Layer

By far the most interesting layer of any networking protocol is the application layer. Without good applications, the network infrastructure is built for naught. When moving to a new infrastructure, it is not necessary to carry over all the features that are on the old infrastructure. Only the features or applications that customers need are required.

When building a network that has open interfaces from the packet layer to the call-control layer and from the call-control layer to the application layer, vendors no longer have to develop applications. Now, they can simply write to these standard APIs and have access to a whole new infrastructure. When a new packet infrastructure is built, opportunities for new applications become widely available.

Legacy applications such as call-centers for enterprise networks, and standard PSTN applications such as call waiting and call forwarding, must be ported onto a new infrastructure without the end user realizing that the change occurred. After these legacy applications are ported, literally thousands of new enhanced applications can be specifically developed for packet infrastructures. These include (but are not limited to) Internet call waiting, push to talk, find me-follow me, and unified messaging. These applications are discussed in next chapters.

1.6 New PSTN Network Infrastructure Model

As discussed in the previous sections, the new infrastructure will focus on the ability to separate the old stagnant infrastructure into a model by which multiple vendors can develop applications and features quickly for the consumer. Figure 1.11 shows how vendors of VOIP equipment wants to carry this model forward.

Figure 1.11 clearly shows the relationship between all three layers as well as the relationship between these layers and the components that would be used in a live network. Carriers will enjoy this method, as it means they won't be locked into a single solution for any of their layers. They will be able to mix and match all three layers to offer the services, functionality, and time-to-market that they need.

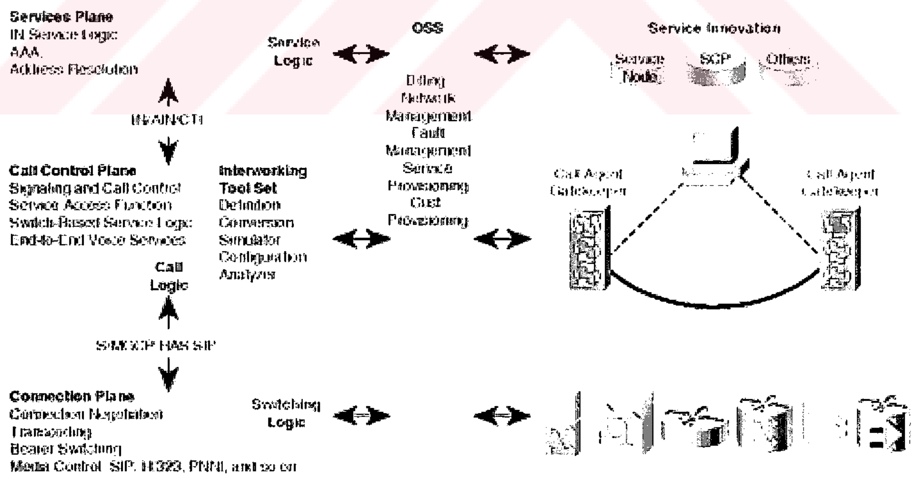

Figure 1. 11 Elements of Packet Telephony

Some carriers might be hesitant to utilize more than one equipment vendor to cut down on their integration timeframe, but many service providers will partner with a minimum of two vendors to ensure competition.

The reality of Figure 1.11 is that the bearers, connection plane, or media transport will be either IP gateways or ATM gateways, or a combination of both. Multiple vendors will be in this space initially, but most likely, they will consolidate to three to five major players.

The call-control plane is an extremely important piece of the new PSTN network infrastructure model, as it must gracefully coexist with both the connection plane and the service (application) layers. Many vendors are building MGC technology.

In fact, approximately 15 vendors are working to ensure compatibility from the connection plane into the call-control and service/application plane.

Many vendors will continue to be in the call-control plane, as service providers will more than likely use several vendors for this key technology, depending upon what service they decide to deploy. The onus on the Call Agent vendors will be to ensure compatibility from one Call Agent to another. Call Agent interoperability is one of the components that could keep service providers from using large-scale, packet-based voice networks.

The service or application plane is where the innovation in the network will happen. One major issue affecting the service plane is its reliance upon soft-switch vendors to open APIs that are useful enough to develop services. For this reason, you will see many application vendors attempting to develop Call Agent technology until APIs into the top Call Agent vendors are fully open and service-friendly.

The service plane is where thousands of ISVs will converge to develop new and revenue-enhancing applications. This is comparable to the client/server revolution in which Microsoft removed the barriers of having to code video drivers, and so on, and enabled ISVs to concentrate on applications. This same revolution is happening in the PSTN today and will change the way services and telephony/multimedia networks are designed, built, and deployed.

CHAPTER TWO
ENTERPRISE TELEPHONY TODAY

2 Enterprise Telephony Today

Enterprise Telephony (ET) is a business telephone system because it provides basic business features, such as hold, three-way calling, call transfer, and call forwarding. ET shares many similarities with it's big brother, the Public Switched Telephone Network (PSTN), but it also has many differences. This chapter details both the similarities and differences between these two networks, the ways in which they interoperate, and typical ET network designs.

2.1 Similarities Between PSTN and ET

PSTN and ET are similar in the following ways:

- Circuit Switching—Both networks are based on the switching of 64-kbps circuits.
- Common Infrastructure Model—Bearers, call-control, and service planes are contained in one platform. These features are described in previous chapter.
- Local Loop—Phones can plug directly into the switch and receive a dial tone, place and receive phone calls, and so on.
- Services Offered—Both networks can provide basic services such as call hold, three-way calling, call transfer, and call forwarding.

Both networks also switch 64-kbps circuits, however, the scale at which each does so is much different. The PSTN uses a Class 5 switch that can support hundreds of thousands of local loops. The ET equivalent to a Class 5 switch, the Private Branch eXchange (PBX), supports from five to several thousand local loops.

The primary task of a Class 5 switch is to provide residential telephony, but it also offers a few basic business features, such as call waiting and call return. A PBX, however, usually offers more features, including call hold, three-way calling, call transfer, voice mail, and many others.

2.2 Differences Between PSTN and ET

PSTN and ET are different from each other in the way they treat signaling and in the types of features they offer.

2.2.1 Signaling Treatment

Although the PSTN uses signaling interfaces developed by industry bodies, PBX manufacturers often create proprietary protocols to enable their PBXs to intercommunicate and carry additional features transparently throughout their voice network.

Chapter 1 discusses how the PSTN uses Signaling System 7 (SS7), ISDN, and in-band signaling as its primary signaling links. These are well-documented standards that have been evolving for many years. Although these signaling protocols cannot solve all signaling problems today, anyone can develop software to interface into the PSTN network.

Many PBXs in ET use CAS and PRI for signaling. In many cases, computer telephony integration (CTI) links also are used to enable a third-party computer application to control some of the PBX's operations. Most PBX vendors, however, implement a proprietary signaling mechanism. This forces enterprise networks to consolidate on one brand of PBX. Although this can be good for the manufacturer, the enterprise business customer is now locked into one vendor.

Many vendors are starting to implement standards-based signaling protocols that enable interoperability between different vendors' PBXs. A list of these protocols is as follows:

- Q Signaling (QSIG)—This is an open standard designed to enable multiple vendors to agree on supplementary services, dial plans, and much more. (The "Q" comes from the International Telecommunication Union Telecommunication Standardization Sector [ITU-T] Q.xxx set of standards.)
- Digital Private Network Signaling System (DPNSS)—This is a British standard designed to enable cross-vendor, inter-PBX communication. This standard was rolled into QSIG.

2.2.2 Advanced Features

Providing advanced features is also an important differentiation between ET and PSTN. Business requirements for telephone networks are much greater than the average home user. Enterprise customers have the need for high-use, feature-rich systems that enable applications such as the following:

- Inbound and outbound call centers—ET networks with this feature usually contain a CTI link that enables new applications—for instance, a screen pops up on the representative's computer screen that gives the representative the caller-ID information, as well as other information about that caller (buying habits, shipping address, and so on).
- Financial Enterprise Telephony—ET networks with this feature often include a network known as *hoot-n-holler*, in which one person speaks and many people listen. This is common in stock brokerage.

ET customers can use the PSTN to service basic PBX needs, but the PSTN does not have advanced applications such as call centers. Also, using PSTN is usually more costly than using ET, and the PSTN might not have all the necessary functionality that the enterprise customer needs.

2.3 Common ET Designs

ET designs generally consist of an inter-working between PSTN and the enterprise network. This inter-working can be as simple as an analog line from the PSTN or a leased line between two PBXs. Or, it can be as complex as an Asynchronous Transfer Mode (ATM) connection using an inter-exchange carrier's (IXC's) public ATM network. This section covers the various methods and network designs commonly used in most ET networks.

There are five methods that businesses can choose, each of which uses slightly different components. These methods include the following:

- Simple business line—This method involves using a line directly from the PSTN as a business line. This line is similar to a residential line, however the business customer is normally charged a higher monthly rate. This simple business line is usually used for very small businesses that do not need many telephony features. This service is provided and managed by the Local Exchange Carrier (LEC) or Competitive LEC (CLEC).
- PBX—A PBX provides many features (such as hold, transfer, park, and so on)that business customers require. This switch often connects to the PSTN through a T1 or E1 circuit. These systems often integrate voice mail, local lines, and PSTN trunks.
- Key-system—This is a smaller version of a PBX and is generally used in offices of fewer than 50 people.
- Centrex line—Provided and managed by the LEC or CLEC, this line offers additional services similar to a PBX, but an additional monthly charge is involved. These services include transfer, three-way calling, and a closed user-dialing plan.
- Virtual Private Networks (VPNs)—With a VPN, the PSTN contains a private dial plan for the enterprise customer. LECs, CLECs, and IXCs can provide VPNs. A local PBX can provide additional features, however.

These methods are broken into two groups: those that the PSTN provides and manages, and those that are privately owned and merely need to interconnect with the PSTN. Each category is discussed in the following sections.

2.3.1 ET Networks Provided by PSTN

If a business has little capital resources for an internal department to manage the telephone network, it often looks to PSTN to provide telephony services. A business might also use PSTN because it is too large for an internal Information Services (IS) department to efficiently manage the entire network, so the telephony network is outsourced through a VPN to the PSTN carrier. The three PSTN-provided telephony networks include the following:

- A simple business line
- A Centrex line
- A VPN

2.3.1.1 Simple Business Line

The most basic of these methods is a simple business line. This service is usually used by small businesses of one or two people who do not need additional phone services.

A landscaping company with one owner and one employee, for example, does not need more than one telephone line with an answering machine attached. Such a company does not need features such as call hold or call transfer. A simple business line is similar to a residential line, but it usually has a higher monthly fee than a residential line. The local carrier charges this additional cost because it assumes the business line is used more often than a residential line.

2.3.1.2 Centrex Line

As a business begins to grow, it starts to require additional services, such as call transfer, call hold, and call waiting. The business can purchase a key-system or PBX, which starts at around U.S. $2000, or it can simply pay a few extra dollars every month (U.S. $20–$30) to the PSTN for additional services.

These services enable the PSTN to offer features in a Closed User Group (CUG). A CUG describes a situation where all the phones within the business become a virtual switch and can dial one another with only four or five digits, transfer calls, and put callers on hold. This service offers more functionality than a simple business line, but it usually becomes cost-prohibitive to implement as the company grows.

2.3.1.3 VPN

Another option available to business users is VPN. VPNs offer enterprise customers the benefits of a private network (CUG) without the administration or equipment hassles of a large tie-line network (a tie-line is simply a permanent circuit between two points).

A VPN enables an enterprise customer to dial a specific number, which then directs the PSTN to treat the customer as a CUG.

2.3.2 Private ET Networks

By far, the most popular option for ET is for businesses to purchase their own key-system or PBX to provide local telephone access to their employees. This method provides many benefits, including the following:

- No recurring charges—Owning a PBX costs less per month than purchasing Centrex services from the PSTN.

- Control over adds, moves, and changes—There is no need to contact the PSTN carrier to add new lines, move a phone, or change subscriber information.

2.3.2.1 PBX Networks

Figure 2.1 shows the relationship between having individual lines from the PSTN, or using a PBX to lower the number of lines (trunks) from the PSTN. Because most users of the telephone system are not calling externally at the same time (depending upon the business type), cost savings on PSTN trunks are realized.

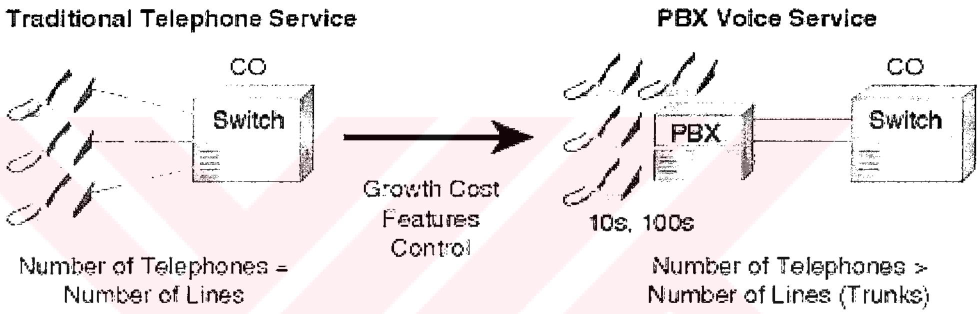

Figure 2. 1 PSTN Compared to a PBX or Key-System

Another advantage to enterprise customers who have their own circuit-switch (PBX) is the control such a setup offers. If you need to add a new user, change a feature, or move a user to a different location, there is no need to contact the PSTN carrier.

The PBX adds another level of complexity, however. The enterprise customer must now deal with the additional burden of configuring and maintaining call routing on the PBX.Figure 2.2 shows a sample block diagram of a user now dialing outside the PBX to the PSTN.

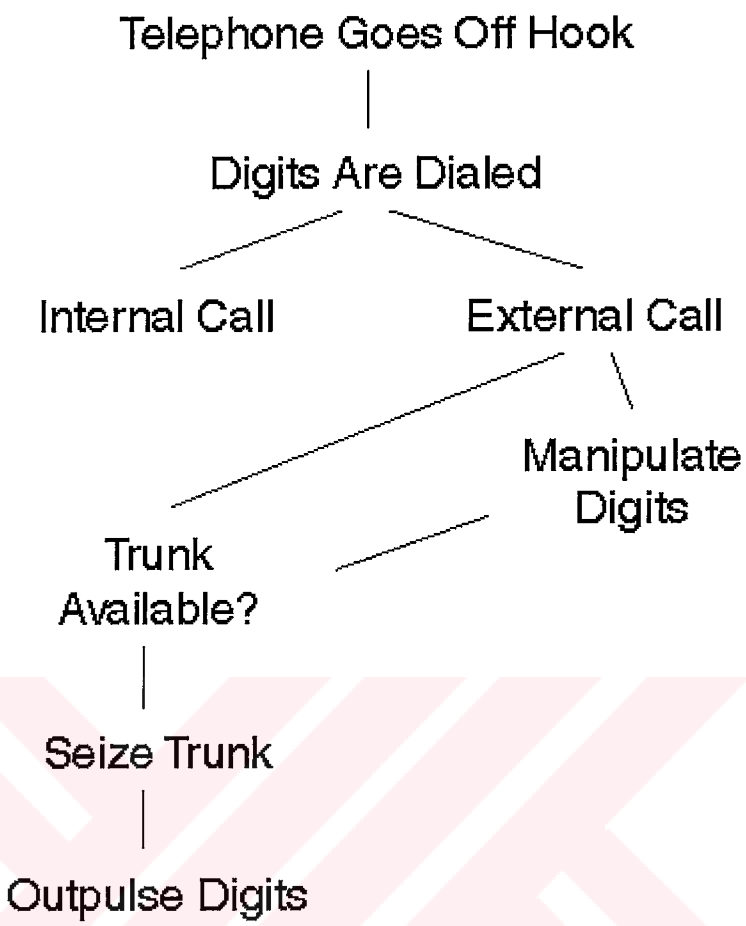

Figure 2. 2 PSTN Call Through a PBX

Figure 2.2 details how a PBX makes a basic call-routing decision regarding when to route the call to the PSTN or to an internal phone extension. This process can be hidden from the user (all calls starting with a "1" use an outbound trunk, for example), or the user can be "trained" (forcing the user to dial "9" for an outbound trunk, for instance) to assist the PBX to choose the proper path.

In many cases, the user decides to route the call to the PSTN based on an "escape" digit (this is usually "9" in the U.S. and "0" in Europe). Other times, the user is unaware that the call is routed over the PSTN.

As an example, consider a five-digit dialing plan for a company that has locations over a large geographical area. Each PBX can be programmed to translate that five-digit number to a 1+10 (ITU-T Recommendation E.164) number and route the call over the PSTN, as shown in Figure 2.3. This 1+10 number also can be referenced as an E.164 number, as it follows that ITU-T recommendation.

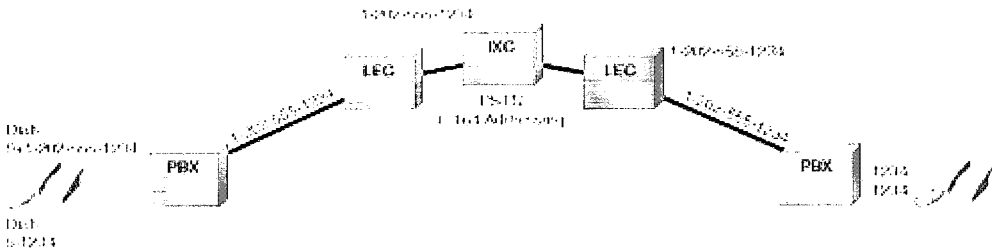

Figure 2. 3 Number Translation Through a PBX

In Figure 2.3, the following occurs:

- A user dials 5-1234, which the local PBX translates to 1-202-555-1234 and sends to the LEC switch.
- The LEC passes the 1+10 number to the IXC, which passes it to another LEC.
- The LEC in area code 202 passes the entire 10-digit number to the remote PBX.
- The remote PBX modifies the incoming number from 202-555-1234 to a four-digit number and rings the appropriate line (1234).

This process of digit manipulation enables the PBX user to dial the least amount of digits possible. This not only saves users time, but it also makes it easier for users to remember frequently used extensions.

Tie-Lines for PBX Interconnection

If a business has two sites and they have a large call volume between them, the business usually purchases a tie-line. Recall that a tie-line is simply a permanent circuit between two points (T1, E1, fractional T1/E1, or some other transport). For this scenario to be cost-effective, it must cost less to run a call between site A and site B over the PSTN than it does to send a call over a permanent circuit.

Figure 2.4 shows two sites (one in San Jose, California, and one in Dallas, Texas), with a T1 circuit between them.

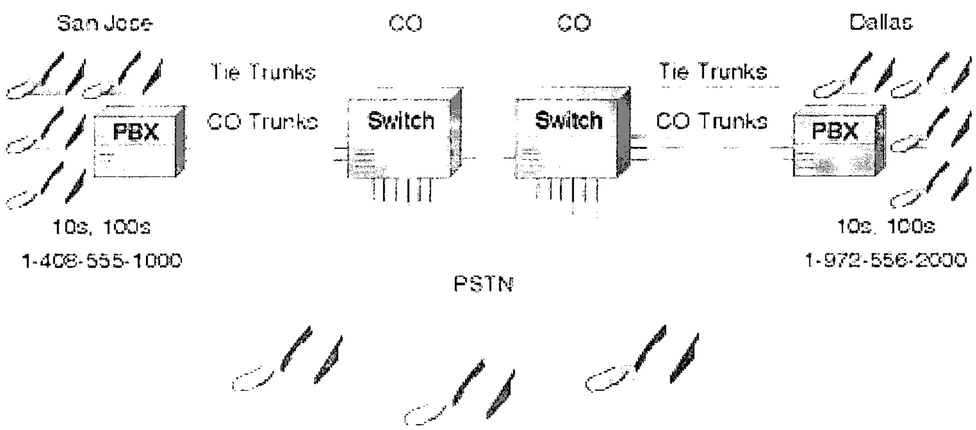

Figure 2. 4 Tie-Line Between San Jose and Dallas

This tie-line still uses the PSTN, but the business pays a flat rate for the dedicated use of the circuit between San Jose and Dallas.

The PBX uses a preprogrammed Automatic Route Selection (ARS) table to determine which trunk should be used. Referring back to Figure 2.4 , the PBX is configured to use the tie-line between San Jose and Dallas. If that tie-line becomes full, the PBX uses the Central Office (CO) trunks as overflow to the PSTN.

To determine whether having a tie-line is cost-effective, a careful analysis of the call volume and cost between San Jose and Dallas as compared to the cost of the T1 circuit must be performed. (Subramanian,2000)

CHAPTER THREE
BASIC TELEPHONY SIGNALLING

3 Basic Telephony Signalling

Many corporations find it advantageous to operate their own voice networks, and they do so by connecting dedicated links between Private Branch eXchanges (PBXs) for inter-office communication, or by using Virtual Private Networks (VPNs) for voice.

Originally, PBXs were connected to the Public Switched Telephone Network (PSTN) for voice services, or they were interconnected using analog tie-lines to transfer voice. When the need for more voice trunks and the technology matured, analog tie-lines were replaced with higher-speed digital facilities capable of accessing sophisticated and feature-rich networks. This chapter analyzes the signaling techniques that traverse analog and digital facilities in corporate and interexchange networks.

This chapter also discusses channel-associated signaling (CAS) systems, such as Consultative Committee for International Telegraph and Telephone (CCITT) No. 5, R1, and R2, and it reviews how these CAS systems operate.

It also describes access protocols, such as Integrated Services Digital Network (ISDN), Q Signaling (QSIG).These protocols deliver PBX signaling through a network to distant PBXs.

Private ISDN networks use the PSTN for connectivity and services. QSIG is an inter-PBX signaling system similar to ISDN that enables corporate PBXs to connect, thus creating a private voice network.

3.1 Signaling Overview

Before covering signaling methods and standards, it's important to discuss some basic concepts. These basic concepts are applied in the individual signaling methods further along in the chapter.

3.1.1 Analog and Digital Signaling

Originally, PBXs were connected by simple analog lines that enabled the transmission of voice-band information. Analog systems are not as common today as they used to be, however, and in many cases, they have been replaced by higher-speed digital facilities that cost less than their analog counterparts.

Digital signaling is the most common type of telephony signaling used in today's corporate and service provider networks. In digital networks, many forms of signaling techniques are used.

One form is robbed-bit signaling. With this method, a bit is "robbed" from designated frames to use for signaling purposes. Robbed-bit signaling inserts the signaling information into the digital voice stream without affecting voice quality.

This signaling technique is discussed in more detail in the "CAS" section later in this chapter. In addition to CAS, other digital protocols include R1, R2, ISDN, QSIG.

3.1.2 Direct Current Signaling

This form of signaling relies on direct current (DC) to signal the end switch or office. DC signaling indicates transition state changes by toggling on or off the flow of DC. These end office switches use current detectors to identify changes in state. DC signaling is used in the following two signaling arrangements:

- Subscriber Loop—This is a simple form of DC signaling between the subscriber and the local end office. When a subscriber goes off-hook, DC (-48V) flows across the line or loop between the telephone and the local end office switch. Line cards in the local office are equipped with current detectors to determine when a connection is being requested. When a subscriber goes on-hook, the capacitor in the telephone blocks the flow of current.

Similarly to off-hook, the change in DC signals to the end office switch that the call was terminated. In this case, the same pair of wires is used to provide the voice and signaling path.

- recEive and transMit (E&M)—This trunking arrangement uses a form of DC signaling to indicate state changes on trunks or tie-lines. With E&M, two leads—one called "E" and the other called "M"—are dedicated to signaling. You can detect the toggling of E&M leads by applying either ground (earth) or a voltage potential (magneto). This form of signaling is covered in the "E&M Signaling" section later in this chapter.

DC signaling has some limitations. Signaling is limited to the number of states you can represent by DC, for instance. Also, when you use the same pair of wires for voice and signaling, the lines or trunks are kept busy even when the two subscribers are not connected.

3.1.3 In-Band and Out-of-Band Signaling

In-band signaling uses tones in place of DC. These tones are transmitted over the same facility as voice and, therefore, are within the 0–4kHz voice band. The tones include Single Frequency, Multi-Frequency (MF), and Dual-Tone Multi-Frequency (DTMF), described here:

- Single Frequency—This tone is used for interoffice trunks and has two possible states: on-hook or idle, and off-hook or busy. The Single Frequency tone is based on a single frequency of 2600 Hz and is used to identify a change in state. Therefore, no tone is present when a connection or circuit is up. When either party hangs up, however, a 2600 Hz tone is sent over the circuit, notifying all interoffice exchanges of the disconnect.

At one time, the Single Frequency tone was used to gain fraudulent long-distance services from service providers. The perpetrator attached a "blue box" to the subscriber line and used it to fool interoffice exchanges into interpreting the 2600 Hz tone as a clear-forward signal. The interoffice switch then accepted the called party number and believed that the local switch would charge for the call. Access to the interoffice switch was accomplished by dialing 0 and fooling the interoffice switch before the operator answered. Service providers eventually curbed this activity by implementing certain protective measures.

- MF—This tone is used by interoffice trunks to indicate events, such as seizure, release, answer, and acknowledge, and to transmit information, such as the calling party number. MF signaling uses a combination of pulses specified by frequencies to signal across a network. These frequencies are system-specified and are covered in more detail in the "CAS," "R1," and "R2" sections later in this chapter. MF signaling uses the same facilities as the voice path and, therefore, is less efficient than common channel signaling (CCS) systems, such as Signaling System 7 (SS7).
- DTMF—This form of addressing is used to transmit telephone number digits from the subscriber to the local office. With the development of DTMF came the replacement of transistor oscillators in telephones with keypads and dual-tone oscillators. DTMF tones identify the numbers 0 through to 9 and the "*" and "#" symbols. When a subscriber presses one of these keys, the oscillator sends two simultaneous tones. Digits are represented by a particular combination of frequencies: one from the low group (697, 770, 852, and 941 Hz) and one from the high group (1290, 1336, 1447, and 1633 Hz). Sixteen possible combinations exist; however, only 12 are implemented on the keypad.

3.1.4 Loop-Start and Ground-Start Signaling

The two most common methods for end-loop signaling are loop-start and ground-start signaling.

- Loop-Start Signaling—This is the simplest and least intelligent of the two signaling protocols. It also is the most common form of subscriber loop signaling. This protocol basically works in the same way as the telephone and the local end office, whereby the creation of a loop initiates a call and the closure of a loop terminates a call. Loop-start signaling is not common for PBX signaling and has one significant drawback, in that glare can occur. *Glare* occurs when two endpoints try to seize the line at the same time, and it often results in two people being connected unknowingly. The person picking up the phone thinks he has a dial tone, but unbeknownst to him he is connected to someone who called him.
- Ground-Start Signaling—This signaling protocol differs from loop-start signaling, in that it provides positive recognition of connects and disconnects. Current-detection mechanisms are used at each end of the trunk, enabling end office switches to agree on which end is seizing the trunk before it is seized. This form of signaling minimizes the effect of glare and costs the same as loop-start signaling. As such, it is the preferred signaling method for PBXs.

3.1.5 CAS and CCS

CAS exists in many networks today. CAS systems carry signaling information from the trunk in the trunk itself. CAS systems were originally developed by different equipment vendors and, therefore, exist in many versions or variants.

Today's telecommunication networks require more efficient means for signaling, however, so they are moving to common channel-type systems, such as CCS.

CCS uses a common link to carry signaling information for a number of trunks. This form of signaling is cheaper, has faster connect times, and is more flexible than CAS. The first generation of CCS is known as SS6; the second generation, SS7.

3.1.6 E&M Signaling

E&M is a common trunk-signaling technique used on telephony switches and PBXs. The signaling and voice trunks in E&M are separated. In E&M, voice is transmitted over either two or four-wire circuits, with six methods for signaling. E&M signaling methods are referred to as Types I, II, III, IV, and V; they also are known by the British Telecom (BT) standard, SSDC5.

3.2 CAS

CAS exists in many varieties that operate over various analog and digital facilities. The analog facilities are either two- or four-wire and the digital facilities are either North American T1 or European E1. This section discusses, CCITT No. 5, R1, and R2 CAS systems.

The main areas of discussion for each CAS system are supervision signaling and address signaling over analog and digital facilities. CCITT No. 5 was designed for analog trunks and uses different MF signals for supervision and address signaling.

It is important to cover a few points before proceeding with a discussion of CAS systems. When a call is placed from Exchange A toward Exchange B, Exchange A is considered the outgoing exchange and Exchange B the incoming exchange.

One-way trunks are trunks on which only Exchange A or Exchange B can initiate a call. Exchanges A and B can initiate a call over two-way trunks. Double seizures can occur over two-way trunks when both exchanges try to seize the trunk at the same time, however. When this occurs, mechanisms such as timers are used to detect and resolve such events.

Three groups of signals are present in channel-associated interexchange signaling systems:

- Supervision Signals—These signals represent events that occur on a trunk and can be specific to the CAS variant. Signals include seizure, wink, and answer; they also are referred to as line signals.
- Address Signals—These signals typically represent the digits dialed or called party number and, in some instances, other information. In this chapter, address signals are based on MF signaling and can be system- or variant-specific.
- Tones and Announcements—These include tones such as ringing and busy tones and announcements such as, "The number you have dialed is no longer in service."

One more concept to cover before moving forward is that of service circuits. Service circuits are used in most exchanges to send and receive address signals and tones, as well as to play announcements. These circuits are typically system-specific; the processor connects a path from the trunk to the appropriate service circuit inside the switch. The pools of service circuits are temporarily used to send and receive tones or to play announcements.(Keshav,1997)

3.2.1 CCITT No. 5 Signaling

The CCITT adopted the CCITT No. 5 signaling system in the 1960s for use in international networks. This signaling system is still used today, usually on long international trunks and, in some cases, over transoceanic and satellite links. This signaling system was designed to operate over analog trunks equipped with Time Assignment Speech Interpolation (TASI). TASI is similar to voice activity detection (VAD), in that it enables unused bandwidth (silences or pauses in speech) to be used for other phone conversations. Link-by-link and in-band signaling are used for both supervision and address signaling. (Keshav,1997)

3.2.1.1 Supervision Signaling

Supervision signaling is accomplished by two frequencies, sent either individually or in combination. CCITT No. 5 uses compelled supervision signaling, whereby the signaling tone is left on until an acknowledgment is received.
The two in-band frequencies are f1, which equals 2400 Hz, and f2, which equals 2600 Hz. The combination of f1 and f2 produces a composite signal.

3.2.1.2 Address Signaling

In CCITT No. 5, address signaling is based on the combination of two frequencies. The address signaling sequence starts with KP1 for national numbers and KP2 for international numbers.

3.2.2 R1

The CAS system known as R1 is available in the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) Q.310 to Q.332 specifications. This signaling system is almost identical to Bell System MF signaling and, therefore, is not further discussed. (Keshav,1997)

3.2.3 R2

R2 signaling is a CAS system developed in the 1960s that is still in use today in Europe, Latin America, Australia, and Asia. Originally known as multi-frequency code (MFC) signaling, R2 signaling exists in several country versions or variants and in an international version called CCITT-R2.

R2 signaling operates over two- or four-wire analog and digital trunks and does not operate over TASI-equipped trunks or satellite links. R2 signaling is more suitable for relatively short international trunks. One of the differentiating aspects of this system compared to R1 is its register or inter-register signaling.

This section focuses on supervision and inter-register signaling for CCITT-R2 and National R2 signaling systems. (Keshav,1997)

3.3 ISDN

ISDN has been available to the public since the 1980s. International Telecommunication Union (ITU; formerly CCITT) I series recommendations define the international standards for ISDN. This subscriber or user-based interface protocol provides single access to multiple services.

ISDN signaling is compatible with SS7 and inter-works with the ISDN User Part (ISUP) protocol. This inter-working enables ISDN subscribers to access the same services and intelligence as they can on the SS7 network. ISDN also enables PBXs to connect over the PSTN and create VPNs. This is accomplished by delivering PBX signaling over the network to distant PBXs.

The ISDN suite defines the specifications for access to the network. The following list outlines some ISDN functions and capabilities:

- ISDN provides circuit-based (voice and data) communications and packet-based communications to its users.
- Many new services can be extended to users.
- ISDN includes two access methods: Basic Rate Interface (BRI) and Primary Rate Interface (PRI).
- ISDN includes single access for PSTN, Direct-Inward-Dial (DID), Direct-Outward-Dial (DOD), 800, Foreign Exchange (FX), tie-lines, packet-switched data, circuit-switched data, and dedicated data.
- ISDN is capable of adding additional channels for high-speed data communications.
- ISDN is capable of transmitting voice and data on the same facility.
- ISDN uses separate channels for signaling.
- ISDN signaling is compatible with SS7.

- ISDN enables the creation of VPNs.

3.3.1 ISDN Services

The following communication services are available in circuit-switched ISDN networks:

- Bearer Services—Three types of bearer services are available for a call. They include speech, 3.1 kHz audio (for modem data), and 64 kbps digital data.

Bearer services are specified by the calling user in the call setup message and are transferred over the network to the called user. The exchanges within the network also use this information when selecting the appropriate outgoing trunk. In the case of speech, exchanges can use analog or digital trunks for interconnection, whereas 64 kbps digital data requires digital trunks.

- Teleservice—This service enables the calling user to specify the type of data service for 3.1 kHz audio and 64 kbps digital data. The teleservice information (fax, telex, and so on) is transmitted transparently across the network to the called user. The called user processes the information to select the appropriate terminal equipment (TE) function to terminate the incoming call.
- Supplementary Services—The ISDN service offering also provides many supplementary services. These services also are typically found on PBXs and virtual private voice networks. The following are examples of supplementary services: calling line identification (caller id), closed users groups, call waiting, user-to-user signaling, advice of charge, call forward, and call hold.

When a user requests these services, supplementary service messages are sent to the network to invoke the requested processes. In the case of user-to-user signaling, the two ISDN users send signaling information transparently during the call setup and teardown parts of the call.(Shenker et al,1998)

3.3.2 ISDN Access Interfaces

Before discussing ISDN access methods, it is important to cover the concept of B and D channels:

- B Channel—The B channel is a 64 kbps channel that carries user information streams. No signaling information is carried in the B channel. B-channel user streams include speech encoded at 64 kbps according to ITU G.711, data at or less than 64 kbps, and voice encoded at lower bit rates.
- D Channel—The D channel is used primarily to carry signaling for circuit switching by ISDN networks. D-channel bit rates are different depending on the access method. The D channel also is capable of transmitting user packet data up to 9.6 kbps.

Two types of access methods exist for ISDN:

- BRI
- PRI

3.3.2.1 BRI

BRI delivers two bi-directional 64 kbps B channels and one bi-directional 16 kbps D channel over standard two-wire telephone lines. Basic rate ISDN service typically is used for residential and small office, home office (SOHO) applications. Each B channel can transmit speech or data; the D channel transmits the signaling or call control messages.

The configuration and reference points for BRI are specified in Figure 3.1

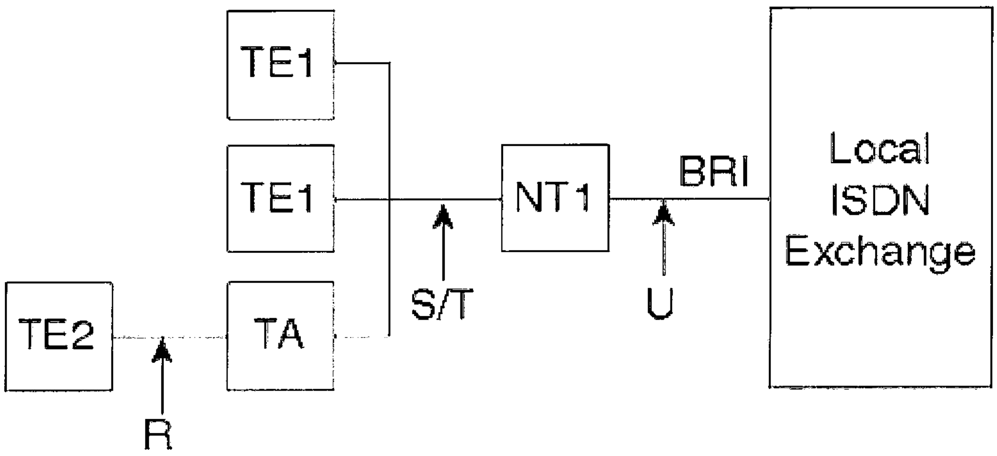

Figure 3. 1 ISDN BRI Reference Points

The reference configuration for ISDN is defined in the ITU specification I.411. The reference points specify the transmission medium, interface, and connectors (if used).

- U Reference Point—The U reference point specifies the transmission characteristics of the local loop. For BRI, this two-wire interface operates at 160 kbps (2B + D + 16 kbps for overhead) over standard copper-twisted wires.
- S/T Reference Point—For basic rate access, this interface provides a four-wire connection to ISDN-compatible terminals or terminal adapters. The interface operates at 144 kbps (2B + D) between the ISDN device and the network termination device. You can connect up to eight ISDN devices to the S/T interface.
- R Reference Point—The R reference point provides connection for non-ISDN devices. Such devices connect to the terminal adapter using interfaces such as RS-232 and V.35.

This reference configuration also specifies the set of functions required to access ISDN networks:

Network Termination 1 (NT1)—Outside the United States, NT1 is on the network side of the defined user-network interface and is considered part of the service provider network. NT1s terminate the two-wire local loop and provide four-wire S/T bus for ISDN terminal equipment (TE).

- TE1—TE1s are ISDN-compatible devices that connect directly to the S/T connector on the NT1.
- TE2—TE2s are non-ISDN compatible devices that require terminal adapter (TA) interconnection.
- TA—TAs provide an ISDN-compliant interface to NT1s and standard interfaces for TE2s. These standard interfaces include RS-232, V.35, RS-449, and X.21.

3.3.2.2 PRI

PRI corresponds to two primary rates: 1.544 Mbps (T1) and 2.048 Mbps (E1). PRIs typically are used in medium to large business applications. PRI is comprised of B channels and one 64 kbps D channel. The interface structure for T1 is 23B + D (North America and Japan). The interface structure for E1 is 30B + D (Europe).

The configuration and references for PRI are specified in Figure 3.2

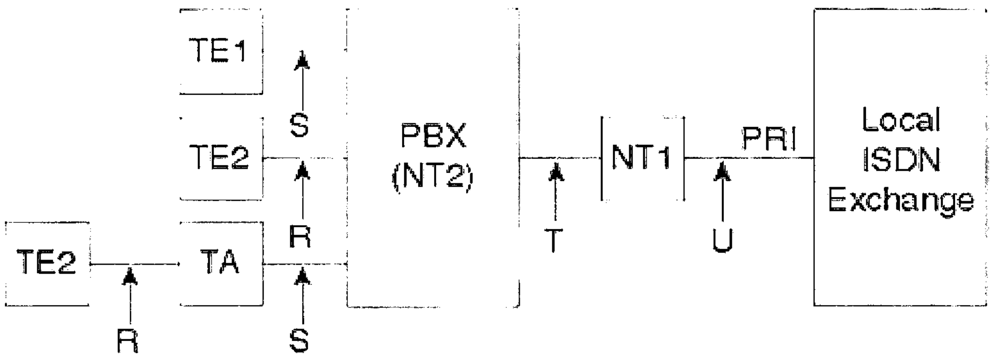

Figure 3. 2 ISDN PRI Reference Points

The configuration and reference points for PRI are similar to those for BRI. The differences between the two reference models are discussed here.

- U Reference Point—For PRI, the U interface is four-wire and operates at either T1 (1.544 Mbps) or E1 (2.048 Mbps) PRI rates.
- T Reference Point—For PRI, the T interface provides access to the Network Termination 2 (NT2) device.
- NT2—PBX equipment can provide such NT2 functions as Layer 2 (L2) and Layer 3 (L3) protocol handling as well as multiplexing, switching, interface termination, and maintenance. NT2s also can provide connections to ISDN-compatible TE1s and non-ISDN compatible TE2s.

3.3.3 ISDN L2 and L3 Protocols

ISDN user-network interface L2 and L3 specifications also are referred to as Digital Subscriber Signaling System No. 1 (DSS1). L2 provides error-free and secure connections for two endpoints across the ISDN reference configuration. L3 provides the mechanism for call establishment, control, and access to services. The L2 protocol for ISDN is Q.920/921, and the L3 protocol is Q.930/931. Q.932 enables general procedures for accessing and controlling supplementary services.

The specifications for L2 are referred to as Link Access Procedures on the D channel (LAPD). This protocol provides the reliable transfer of frames between the local exchange and the TE. The specifications for Q.920 and Q.921 are extensive and are available from the Q series of ITU recommendations.

The specifications for L3 define the messages that pass between the local exchange and the TE. These messages are used for call setup, call supervision, call teardown, and supplementary services. The next section discusses the specifics of ISDN messaging.

3.3.4 Q.931 Call Control Messages

The message structure and signaling elements of Q.931 are used in ISDN networks to provide call control capabilities. Q.931 messages are sent from the network to the user and from the user to the network. They are referred to as user-network and network-user messages.

The message type field in the general format of the Q.931 message is used to determine the type of message being sent.

3.3.5 Basic ISDN Call

This section outlines a typical ISDN call between two users served by the same local exchange. The signaling sequence between User A (TE-A), the local exchange, and User B (TE-B) is illustrated in Figure 3.3.

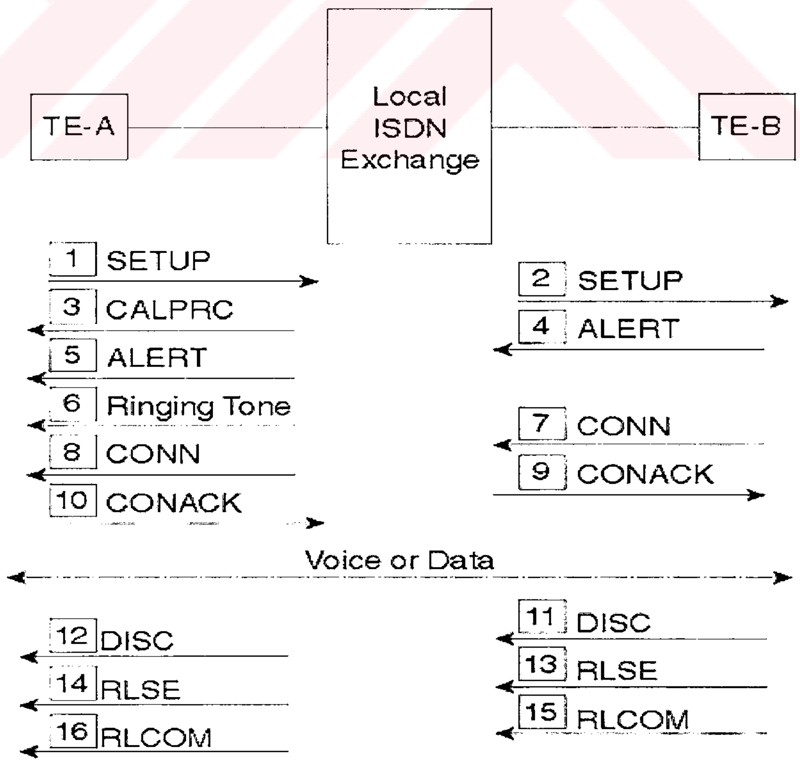

Figure 3. 3 Basic ISDN Call

3.3.5.1 Call Setup

TE-A initiates the call by sending a SETUP message to TE-B. The SETUP message contains the complete called party number (also known as the en-bloc signal). The local exchange then sends a SETUP message to TE-B and includes in the message the B-channel assignment.

At this point, the local exchange sends a CALPRC message to TE-A indicating that the call setup started. If TE-B accepts the incoming call, an ALERT message is returned. The local exchange then sends an ALERT to TE-A, and if this is a speech call, a ringing tone is applied to the B channel.

When TE-B answers the call, a CONN message is sent to the exchange where the B channels are connected; a CONN message also is sent to TE-A. The local exchange acknowledges TE-B's CONN with a CONACK; TE-A also can acknowledge the CONN with a CONACK.

3.3.5.2 Call Disconnect

Consider the example in which TE-B is the first to initiate a disconnect. TE-B sends a DISC, and the local exchange then sends a DISC to TE-A.

At this point, the local exchange clears the B channel to TE-B and sends an RLSE message to TE-B. Next, TE-B releases the endpoint B channel and sends an RLCOM message. The same release procedure also occurs between TE-A and the local exchange.

3.4 QSIG

QSIG is a peer-to-peer signaling system used in corporate voice networking. Internationally, QSIG is known as Private Signaling System No. 1 (PSS1). This open standard is based on the ITU-T Q.9XX series of recommendations for basic service and supplementary services. Therefore, as well as providing inter-PBX communications, QSIG is compatible with public and private ISDN. (Keshav,1997)

QSIG also has one important mechanism known as Generic Functional Procedures (QSIG GF). This mechanism provides a standard method for transporting features transparently across a network.

The following are attributes of the QSIG global signaling system:

- It is a standards-based protocol enabling the interconnection of multivendor equipment.
- It enables inter-PBX basic, feature transparency, and supplementary services.
- It is interoperable with public and private ISDNs.
- It operates in any network configuration (star, mesh, and so on) and is compatible with many PBX-type interfaces.
- It does not impose restrictions on private numbering plans.

QSIG is an important signaling system. The remainder of this section covers the following key aspects of QSIG:

- Services
- Architecture and Reference Points
- Protocol Stack
- Basic Call Setup and Teardown

3.4.1 QSIG Services

QSIG supports a suite of services and features for corporate PBX networks. The three main service groups include basic services, generic functional procedures, and supplementary services.

- Basic service (QSIG BC)—This service provides the capabilities to set up, manage, and tear down a call. Similar to an ISDN bearer service, basic services include speech, 3.1 kHz audio, and 64 kbps unrestricted.
- QSIG GF—This is a standardized method for transporting nonstandard features, thus providing feature transparency. This mechanism enables the exchange of signaling information for the control of supplementary and additional network features over a corporate network.
- Supplementary services—This category includes services and additional network features (ANFs). Supplementary services and ANFs include call completion, call forward, call diversion, call transfer, call waiting, caller id, and advice of charge.

3.4.2 QSIG Architecture and Reference Points

It is necessary to extend the ISDN reference model to include PBX-to-PBX signaling for corporate networks. To accommodate these two new reference points, "Q" and "C" were identified by the standard, as illustrated in Figure 3.3.

The Q reference point defines the logical signaling between PBXes, and the C reference point identifies the physical interconnection. A corporate network can have dedicated analog or digital channels, or it can have VPN switched connections.

Typically, it is assumed that a T1 or E1 digital interface is used to connect to the network. QSIG end-to-end signaling is maintained from PBX to PBX, and ISDN and ISUP inter-working is critical for end-to-end signaling in the ISDN network. As mentioned previously, QSIG is compatible with ISDN; these reference points also are noted in Figure 3.4.

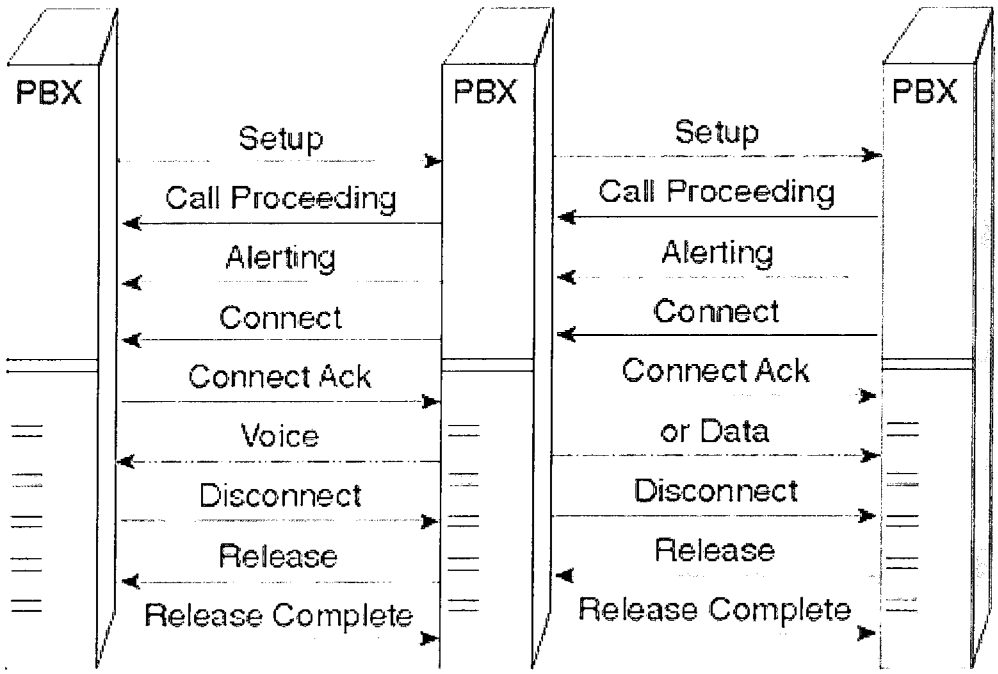

Figure 3. 4 Reference Model for Corporate Networks

The T reference point defines access to the NT2 device for ISDN PRI. The C reference point is the physical interconnection point to the PBX. It is compatible with many interfaces, including two- and four-wire analog, BRI, PRI, and radio and satellite links. The Q reference point specifies the logical signaling point between two PBXs. This reference point is used to specify signaling-system and related protocols.

3.4.3 QSIG Protocol Stack

The QSIG protocol stack specifies a signaling system at reference point Q. QSIG has an identical structure to that of ISDN, and at L1 and L2, these protocols can be the same. They differ at L3, however, where QSIG is split into the following three sublayers:

- QSIG BC—With this symmetrical protocol, the interfaces and messages for the user and network sides are identical. The messages and sequences of this protocol are more easily understood and demonstrated in the example at the end of this section.

- QSIG GF—This protocol specifies the control entities for supplementary services and ANFs. This protocol does not have the capability to control these services, but it does provide the generic layer capabilities to enable them. The protocol provides a connection-oriented and connectionless mechanism between the application entities of different PBXs.

- QSIG Supplementary Service and ANF Protocols—These define the procedures for individual or specific services and features. These services and ANFs are defined and detailed in separate specifications. Such organizations as the European Computer Manufacturers Association (ECMA) and the European Telecommunication Standards Institute (ETSI) are developing these protocol standards.

3.4.4 QSIG Basic Call Setup and Teardown Example

The QSIG BC protocol provides the basic capabilities for call setup and teardown. This protocol extends the ISDN access protocol for use in a corporate network or private ISDN. QSIG BC is a symmetrical protocol whereby both the network and user sides of the interface are identical. The message sequence for a basic call is demonstrated in Figure 3.5. The QSIG BC messages are functionally similar to the messages discussed in the "ISDN" section of this chapter.

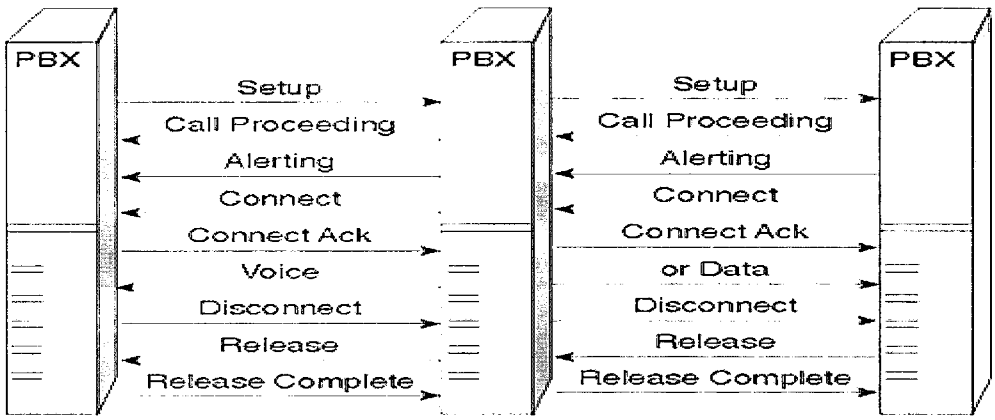

Figure 3. 5 QSIG BC Message Sequence

CHAPTER FOUR

VOICE OVER IP BENEFITS AND APPLICATIONS

4 Voice over IP Benefits and Applications

This chapter discusses how the applications currently available in the PSTN, along with other new applications, work in a packet-based voice network.

4.1 Key Benefits of VoIP

One of the key drivers of combining voice and data networks is monetary savings. If you look strictly at minute-to-minute costs, the savings realized by going with VoIP might not be large enough to justify the expense of rolling out this service.

Price savings can vary based on your geographic location. In countries other than North America, for instance, a minute-to-minute cost comparison between VoIP and the traditional PSTN (a local call in some countries can be around $1 a minute) more than justifies the expense of the new network.

In some countries , however, many large corporations pay 3 cents or less per minute for long-distance calls they make within the same country. For such corporations, it is hard to justify to accounting that rolling out a new infrastructure will provide a Return on Investment (ROI) that will pay off quickly—that is, unless they factor in items other than per-minute charges.

For enterprise networks, for instance, consolidating voice and data networks might mean the ET customer can order fewer circuits from the PSTN. Also, an IP infrastructure requires fewer adds, moves, and changes than a traditional voice or data network. This is because, with one infrastructure, you can use such data features as Dynamic Host Configuration Protocol (DHCP).

DHCP enables a device (a PC or an IP phone) to dynamically receive an IP address (that is, the IP address does not need to be statically configured into the device). So, for instance, if you have an IP phone configured with DHCP, you can move the phone wherever you need and still keep the same phone number. This is similar to moving your laptop from office to office and still being able to log in to the same network server.

Many large enterprises have determined that it costs several hundred dollars just to move a telephone today (this is due to such factors as labor costs and the cost of reconfiguring the switch). Such costs are not incurred in an IP infrastructure, however, because your IP phone profile is set up, and the IP network doesn't care where you are located.

An additional benefit of VoIP is the ability to have one Information Services (IS) department that supports both voice and data networks (as the networks are now one entity). This can initially cause tension between these two infrastructures, but as with any technological revolution, one must enhance one's skills to survive. This has been the case with the introduction of most new technologies—from the cotton gin to robots.

One benefit of VoIP that enterprises and service providers often overlook is the fact that common infrastructure tools are now no longer needed. These include such tools as physical ports for services such as voice mail. In a circuit-switched voice network, voice mail is sold based on the number of mailboxes and the number of physical ports needed to support simultaneous users. With VoIP, physical circuit-switched ports are not necessary. The voice mail server need only have an IP connection (Ethernet, Asynchronous Transfer Mode [ATM], and so on).

Also, VoIP enables voice mail systems to be put on standards-based platforms (such as PCs and UNIX machines). After a feature is on a standards-based platform, price gouging is much less likely to occur.

What if your voice mail server was the same as your e-mail server and you could decide whether to download your voice mail over a telephone or use your e-mail client to peruse your voice mail? Those who travel will truly appreciate benefits such as the capability to download voice mail and respond electronically, and to forward voice mail to a group. Such technology exists today and will soon be available and widely used through enterprise and service provider networks.(Black,1999)

4.2 Packet Telephony Call Centers

In most call centers today, the largest costs are for the brick and mortar holding the building together. You can drastically cut the actual costs of renting a building, putting a phone at each desk, and purchasing the required infrastructure (call-routing technology, PCs, and so on) by using a Packet Telephony Call Center (PTCC).

Each call center is different, but for many call centers, the ability to grow the business as needed (perhaps as discretely as one station at a time) is a great benefit. Currently, call centers must grow in chunks. The size of these chunks depends on how many ports the call centers can purchase for their Private Branch eXchange (PBX) at a time. This is a great disadvantage because call centers usually need to be flexible and be able to grow and shrink as the number of required stations changes.

Many call centers are unable to grow in smaller chunks because the hardware necessary to provide desktop phone services is sold only in larger units (such as growing one to several T1s or E1s at a time, instead of a phone at a time). This prevents the call centers from being able to grow quickly based on seasonal or natural growth.

Circuit-Switching Call Centers (CSCCs) enable users to work from home and still take calls, but this equipment is expensive. With PTCCs, users can log in to a phone no matter where they are and have access to the exact same features as if they were at their desk, and the costs are much lower.

A CSCC currently uses a device known as a PBX Extender, a remote piece of equipment that extends the features of the PBX to the user's premises. A PBX Extender can run upward of $1000 per user, and that's just for the equipment itself. You also have to purchase software that must be added to the central site; the circuit to the worker's residence; and Customer Premise Equipment (CPE) gear, such as the router, for the remote site.(Black,1999)

When you use a VoIP network, however, you don't need additional equipment for the remote site. You can take the same phone you use at work and have exactly the same functionality. Of course, the company still has to purchase the circuit to the worker's residence, as well as the CPE equipment.

Nevertheless, VoIP lowers the costs of locating stations anywhere geographically. In doing so, VoIP gives call-center operators a great advantage in terms of hiring skilled or unskilled workers, as well as growing and shrinking the number of stations needed at any given point in time.

In a packet telephony infrastructure, you can have a group of distributed virtual agents that you can locate anywhere, and you can still offer them the same tools that a traditional call center offers. Figure 4.1 shows ways you can use a common IP infrastructure to unite various methods, and it showcases one possibility of telecommuters as virtual agents.

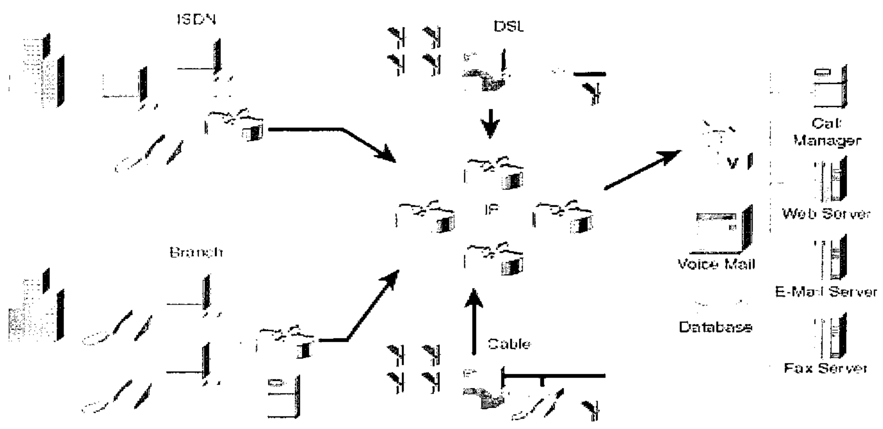

Figure 4. 1 Virtual Agents

Two of the challenges facing CSCCs today are cost and employee retention. Descriptions of these challenges are as follows:

- Many toll-free numbers—CSCCs must manage the number of circuits the enterprise uses. Using more circuits increases the cost of operating the CSCC and, therefore, can potentially decrease profits.
- Misrouted/rerouted calls—Each time a call must be routed to a different agent (because, for example, an agent might not have the correct skills to answer a customer's question, or he does not speak a customer's language), revenue is lost.
- Multiple centers—The capability to "follow the sun" increases the "brick-and-mortar" costs in a CSCC. Following the sun implies that different physical call centers must exist to keep workers working a normal shift. This also is known as time-of-day routing. (When the United States call-center operators are sleeping, for example, Australian operators can take the calls.)
- Percentage distribution/overflow routing—The capability to handle overflow between different physical locations increases the profitability in a peak call-flow time. But, if this overflow mechanism is not properly managed, it can cost more to overflow the call than to not service the incoming call.
- Employee turnover—Call-center work can be stressful, and, because of the repetitive nature of such work, keeping workers can be difficult.
- Seasonal staffing needs—Oftentimes, call centers experience more volume during certain periods. As such, they must hire people to accommodate the high-

volume periods, and then lay people off when volume drops. (This is a common plight of technical support staff during the holiday season, for instance.)

- Inconvenient busy hours—If the call center does not have a "follow the sun" model, it must hire staff to work inconvenient hours, such as the night shift, for instance.
- Regional call-center talent—Having skilled workers come into a brick-and-mortar facility can lower the number of possible workers in the pool of talent. Telecommuting so that regional workers can work within any geography in a specific time zone increases the number of workers in the available pool.

The CSCCs are adapting to meet these challenges, as well as meet new demands. One of the solutions to the previously mentioned challenges is increased efficiency. To become more efficient, practice the following principles:

- Computer Telephony Integration (CTI)—One application is for caller information (such as the caller's name, buying patterns, and address) to be "popped" onto the agent's screen so that the agent can handle the call more quickly.
- Skills/application-based routing—Routing calls to the proper agent based on technical skills, language, and any other skill can increase the speed by which the call is handled.
- Information duplication—Call agents can avoid asking the same question twice if transferred to a new agent. This is possible due to the information on the first agent's screen "popping" onto the new agent's screen when the call is transferred to the new agent.
- Interactive Voice Response (IVR)—This enables callers to input basic information (such as account information) so that calls can be handled more quickly.

CSCCs will upgrade to an integrated voice and data network initially based on cost. But, the true value (which might be hidden to some) is in the value-added services and applications that can be offered after this enhanced network is in place.

Some of these services and applications include having both your voice mail and e-mail integrated into one application; using Web-based customer support; having

CTI capability; being able to fax in and out from the desktop (and fax to your e-mail account); and being able to conduct desktop video conferencing with your customer.

Traditionally, an entire call center revolves around the PBX (as shown in Figure 4.2). As such, call centers are held ransom by the number of ports they can afford at any given point in time. Reliance on the PBX also forces the CSCC to deploy applications only when they are compatible with the PBX or when the CTI link enables the field to be passed.

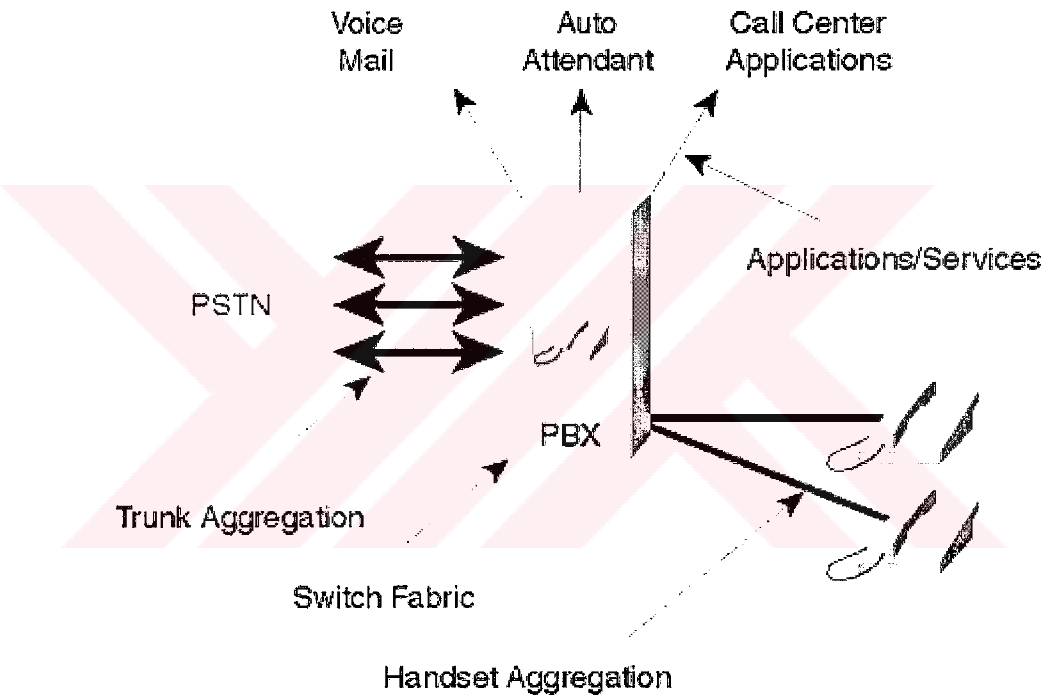

Figure 4. 2 Circuit-Switching Call Center

In a PTCC, the network is integrated and standards-based and does not rely on only one component or vendor to provide the entire solution.This enables the call center to have remote users for a fraction of the cost of PBX extenders

This also enables the business to grow in the increments that customers need, and enables the business to add new applications (such as data/voice collaboration) as needed.

Another important issue with CSCCs is the ability to retain and develop current employees. Studies show that giving employees options on schedules and "flex-hours" greatly increases the retention rates of many companies.

Although Figure 4.2 shows how a CSCC is efficient for a large centralized call center, the CSCC design lacks the flexibility to enable telecommuters, and it lacks a true integration into Internet telephony or unified communications (such as fax–to–e-mail).

PTCC enables you to retain a connection into the legacy PBX call center, but it also enables integration into the new network of Web support, Internet telephony, and unified communications. Figure 4.3 shows the components and network design of a PTCC.

This connection to the legacy PBX is accomplished by having an external call-processing engine that connects to the PBX and to the Call Manager through CTI links. The external call-processing engine enables telecommuters and PBX call agents to answer calls as though they are attached to the call center.

Also, with a connection from the legacy CSCC into the IP network, you can use enhanced features such as IP-based IVR systems (also known as Voice Response Units [VRUs]) and unified messaging services such as fax–to–e-mail, text-to-speech, speech-to-text, and so on.

As you can see in Figure 4.3 , the Call Center Corporation is no longer tied to physical ports for the VRU, and the entire messaging infrastructure (e-mail, voice mail, applications, and so on) is tied into one common infrastructure.

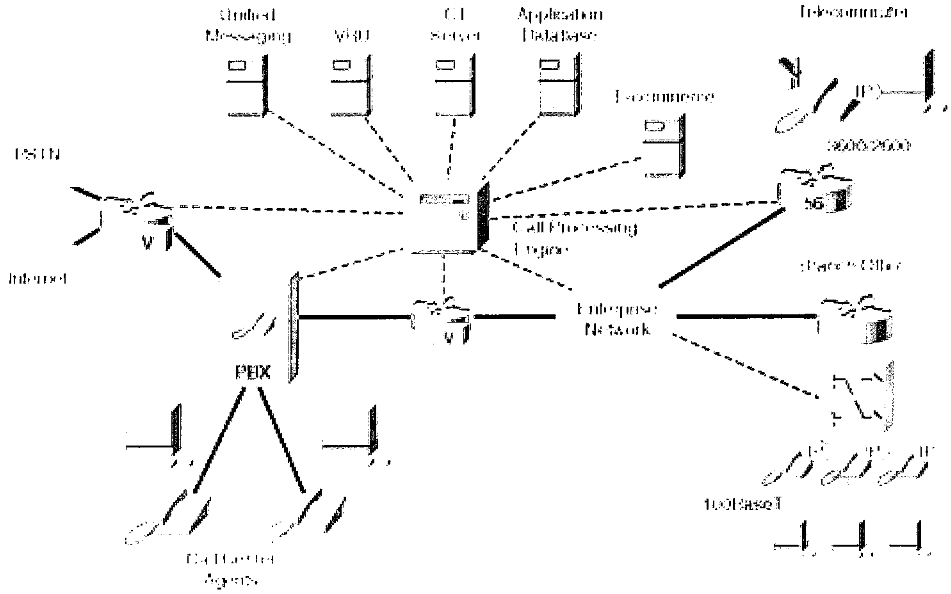

Figure 4. 3 Packet Telephony Call Center

The call-routing or call-processing engine is now just part of the data network and is removed from the PBX. This enables telecommuters, call-center agents, and branch office agents to have the same access to the same information. Access to a common infrastructure gives everyone equal footing, and it gives the customer a common look and feel, as shown in Figure 4.4

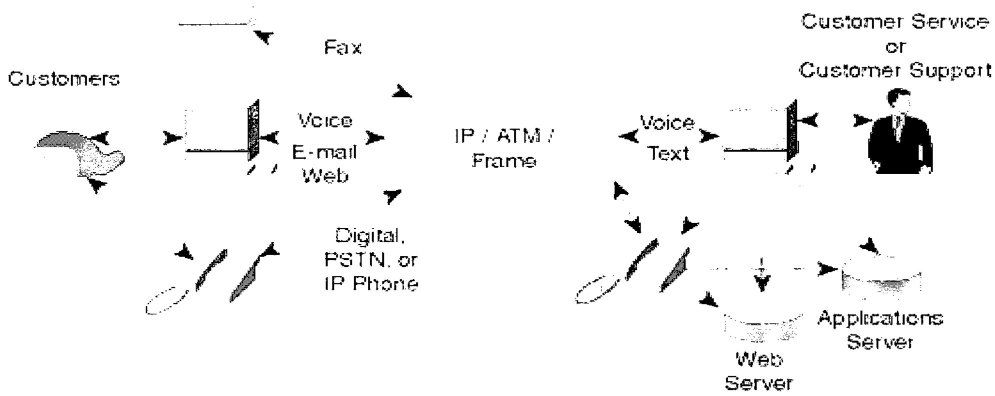

Figure 4. 4 Common Infrastructure for All Call Agents

This new architecture also enables you to stop making unnecessary expenditures on legacy CSCC gear and begin expanding into the packet telephony call-center space. This, in effect, frees you from having to grow your business in larger chunks.

This network also utilizes your existing wide-area network (WAN) data infrastructure and can provide a more efficient use of existing bandwidth. Calls to remote agents are now essentially "toll-free" because they are traveling over an IP infrastructure.

Another key benefit is Web integration. This means a call-center customer can request a call back from the Web site (also known as "click to call back"), which uses the PSTN or even places an Internet telephony call. This saves the call center money because it doesn't have to pay the incoming 800 toll charges.

4.3 Value-Added Services

After Internet Telephony service providers (ITSPs) have a VoIP network (possibly for a pre- or post-paid application) in place, they can begin to offer value-added services . Two of these value-added applications are Internet Call Waiting (ICW) and Virtual Second Line (V2L).

4.3.1 ICW

ICW is a service that enables subscribers to receive notification of an incoming voice call on their PCs while connected to their ISP. Subscribers are notified of the incoming call through a screen-pop on their PCs, at which point they can do the following:

- Send the call to voice mail.
- Receive the call on the PC using H.323 software (VoIP).
- Drop the Internet session and receive the call on the telephone (PSTN).
- Ignore the call (provide a busy signal or let it ring).

These enhanced services afford benefits to both the service provider and the customer. The service provider can leverage its existing infrastructure to provide more services, and it can have an existing potential customer base with its dial-up customers. The service provider can provide this new service without having to become an official telecommunications provider (such as a Competitive LEC, or CLEC).

The customer benefits in that he does not miss incoming calls while online, does not have to pay for a second line from the telephone company just for Internet access, and can handle incoming calls in many ways. He can still have access to caller-ID, for instance, and he can set up variables such as forwarding to voice mail, ignoring the call, or transferring the call to a cell-phone.

4.3.1.1 V2L

V2L is a simple service in that it enables Internet users to place and receive phone calls through their ISP only when they are connected through their Internet connection (modem, cable, digital subscriber line [DSL], and so on). In many V2L cases, the PC is actually assigned a valid E.164 number, although this is not a requirement.

All the benefits of ICW also exist for V2L. One key additional benefit is that service providers can offer outbound traffic, which can, in turn, create significant revenue streams for the service provider. Also, subscribers can save a tremendous amount of money on long-distance charges.

With V2L, ITSPs effectively have a local loop to their customers through their modem access and can offer long-distance services through the Internet. Because the ITSP's IP network is less expensive to build than the PSTN, the ITSP can offer lower long-distance rates to the subscriber. Offering long-distance service provides an additional revenue stream for the ITSP.

4.4 Enterprise Case Study

Chapter 2, "Enterprise Telephony Today," discusses ways you can build enterprise telephone networks. This section discusses ways you can use packet voice technologies not only to save money on long-distance charges, but also to decrease capital expenditures on other recurring charges.

The connections between the central site and the remote sites are often called tie-lines. Built to transport voice, tie-lines are permanent 64 Kbps connections that can actually transport voice or data.

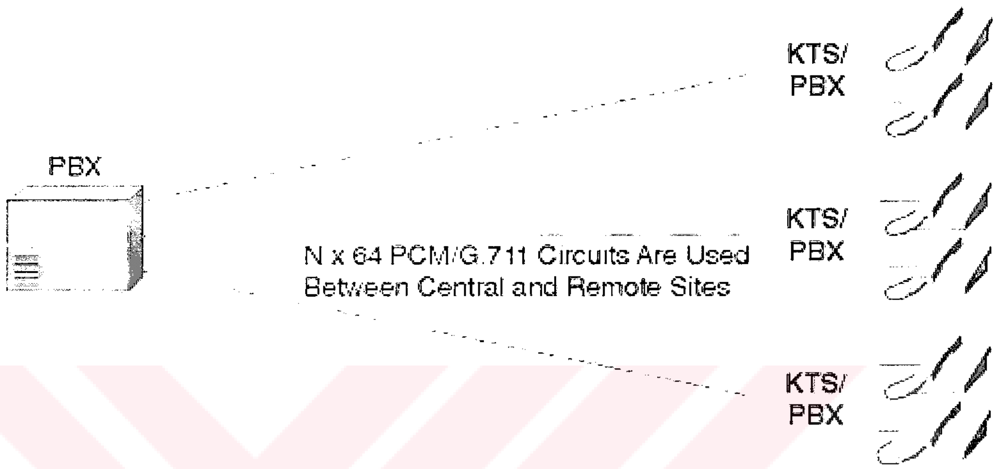

Figure 4. 5 Enterprise Telephony

Most enterprise customers also have data networks, a minor modification and enhancement to the data network can enable voice tie-lines to be replaced simply by moving the voice traffic onto the data infrastructure. This causes the voice and data infrastructure to look something similar to Figure 4.6

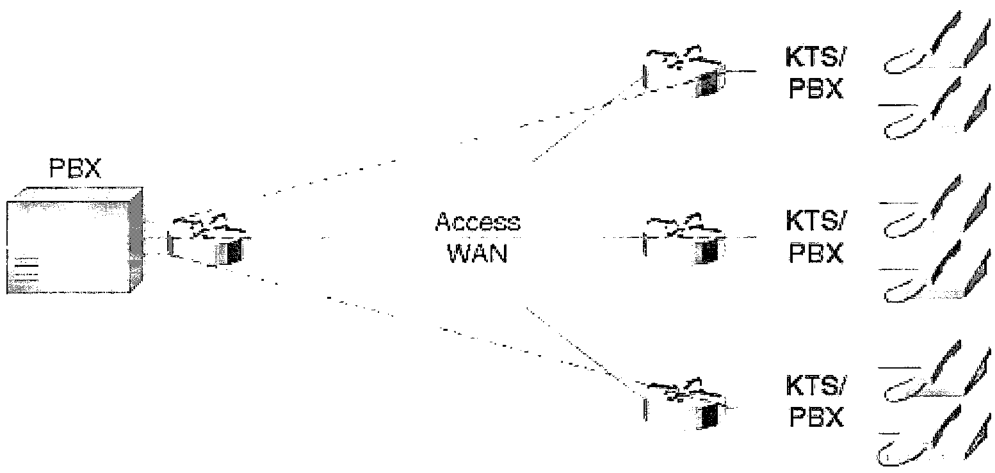

Figure 4. 6 Enterprise Voice and Data Network

Replacing tie-lines with VoIP and leaving the rest of the infrastructure is just the first step to successful voice and data convergence. Many more steps are necessary.

Suppose that a corporation, an enterprise customer, wants to converge its voice and data networks to save money in the short term. Having two separate infrastructures for a voice and data network requires that you have leased lines not only for voice, but also for data paths. Figure 4.7 shows a typical enterprise customer with separate networks.

The voice network uses multiplexers to connect voice and data networks across one T1 circuit. When voice is not being used, however, the voice network is still consuming bandwidth across the leased T1 circuit.

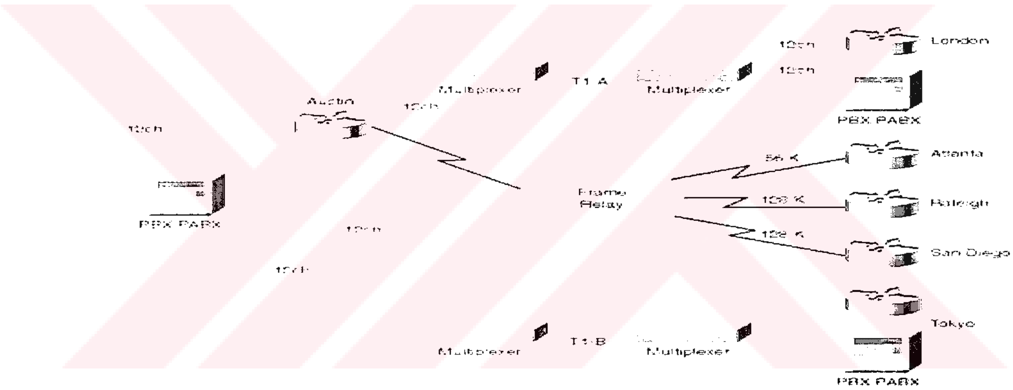

Figure 4. 7 Typical Enterprise Voice and Data Network

4.4.1 Company’s Current Voice and Data Network

Taking a deeper look into the company , you need to understand its current network. Company's headquarters are located in Austin. Company has several remote sales and development offices across the United States, as well as in Tokyo and London, where its two largest offices are located. The remaining offices in the U.S. concentrate mainly on sales. Two of company’s main goals were to cut costs while preparing to deploy a more cost-effective voice network, and to increase bandwidth between sites.

Company has two intercontinental T1 circuits connected to both London and Tokyo. Multiplexers are used on these circuits to separate 12 channels of each T1 to voice and 12 channels of each T1 to data. The U.S. sites run across a Frame Relay network. The Atlanta site houses a small sales office where from two to five people work at any given time. The Raleigh and San Diego sites have slightly larger regional offices employing both sales people and development staff.

The IS department conducted a study and determined that both data and voice bandwidth needs were growing. The IS department decided to research methods for compressing voice and taking advantage of unused time-division multiplexing (TDM) bandwidth currently utilized by the multiplexing configuration.

The IS department also conducted a study to determine calling patterns. It found that most long-distance calls from all sites are clustered around the various regions in which the corporation has branches.

Company asked itself several questions to determine whether a combined voice and data network would provide the expected savings.(Black,1999)

4.4.2 Company's Convergence Plan and Goals

It is important to understand where the customer's network stands today and where the customer wants to be when the data/voice networks have converged. Therefore, ask the following questions:

- What is the total expenditure on voice networks and capital equipment?
- What is the primary application for VoIP (toll bypass, call-center, or ICW)?
- How many remote sites does the company have?
- How many people are at each remote site?
- What is the average phone usage in minutes per user per site?
- How many calls are placed to interoffice locations?
- What is the average cost per minute per location?
- What is the customer's expectation of quality (cellular, toll)?

- What is the total number of long-distance minutes between sites?
- What percentage of traffic is expected to be voice/fax?
- Can the existing IP infrastructure support the necessary quality of service (QoS) for voice?

After these questions are answered, enterprise customers can decide whether they can afford to make the voice/data transition.

Company took the necessary time to plan its network design in phases. Its final network design is shown in Figure 4.8.

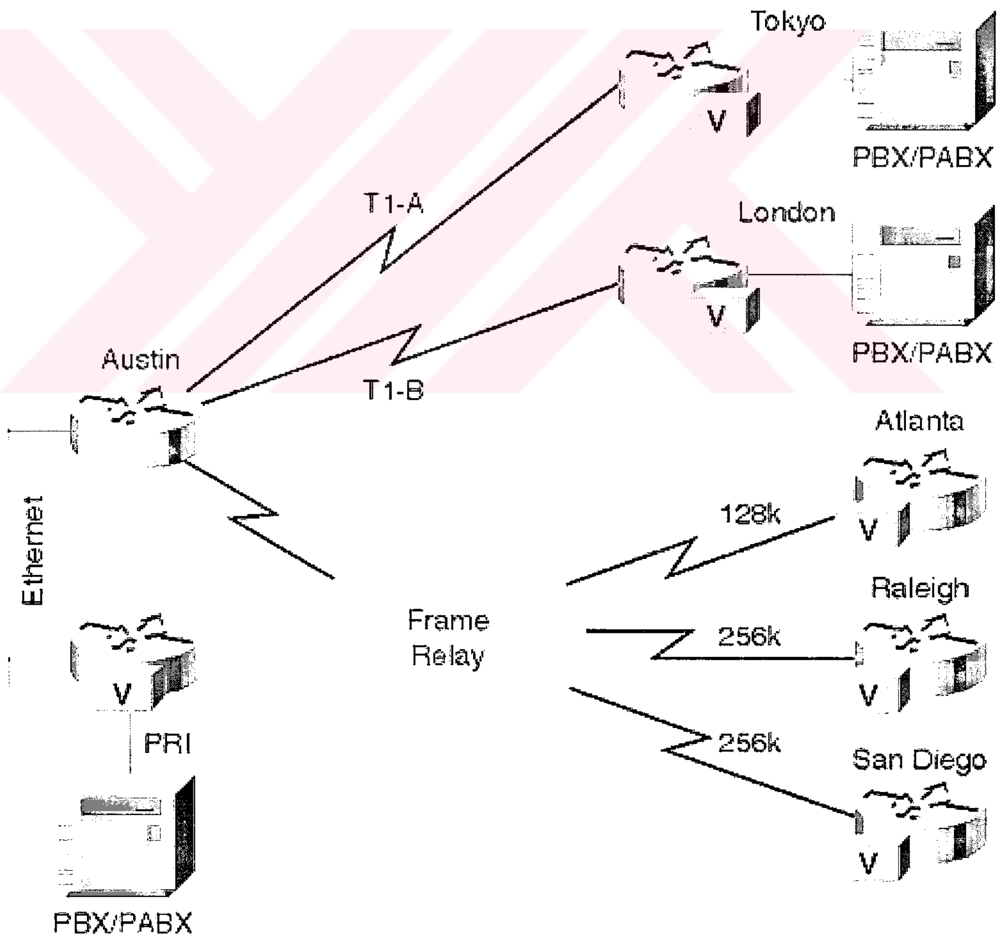

Figure 4. 8 Integrated Voice and Data Network

The network design shown in Figure 4.8 is just one step in the path toward voice and data integration. Company's next step is to slowly replace the key-systems and PBXs at its sites with IP phones. Doing so obviates the need for purchasing additional circuit-switching hardware and provides many additional benefits, including a single infrastructure and support group. (Black,1999)

4.4.3 Integration of Voice and Data Networks

The next step for an enterprise customer is to simplify the local-area network (LAN) by implementing a common voice and data network to the desktop. This is accomplished with IP phones, as shown in Figure 4.9.

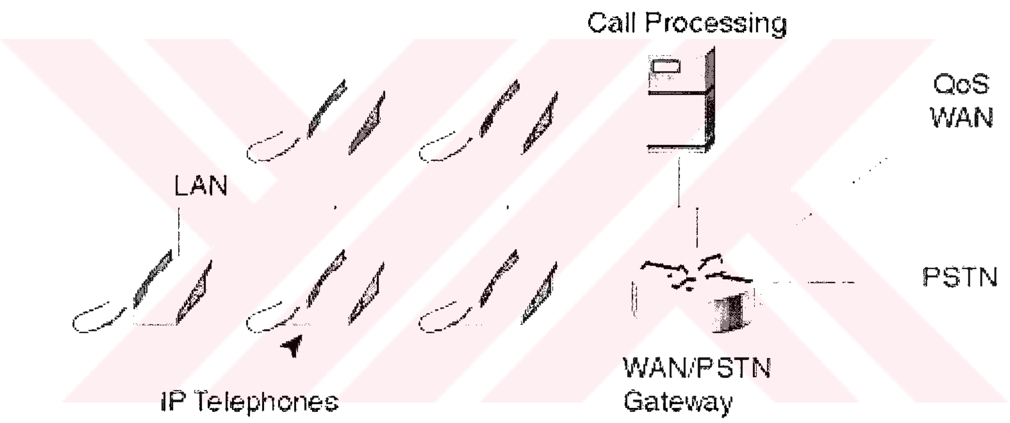

Figure 4. 9 Voice/Data Integration to the Desktop

This type of integrated network provides numerous cost savings:

- Phones use DHCP and keep phone numbers regardless of physical location.
- Cabling to the desktop is easier (everything is Ethernet).
- Call appearance remains the same whether the user is at home or at work. This enables fully transparent telecommuting.
- The call-processing engine is now on a standard platform, which provides the enterprise network with greater flexibility.
- Call Manager is actually configurable through Hypertext Markup Language (HTML), which simplifies administrative overhead as well as PBX administration.

- Call Manager can also support other standards-based interfaces such as Station Message Desk Interface (SMDI) for an interface into a legacy PBX. As an example, you can use this interface to illuminate the message-waiting light.

Putting voice on data networks in enterprise, service provider, and other types of networks affords numerous additional benefits. Some are apparent, and some have yet to be discovered. (Black,1999)

CHAPTER FIVE
IP TUTORIAL

5 IP Tutorial

Many of the benefits of Voice over IP (VoIP) are derived from the use of Internet Protocol (IP) as the transport mechanism. To truly understand these benefits, we will first demonstrate IP . Before we can understand what IP can do for us and ways we can run applications through IP, we must first become familiar with the Open Systems Interconnection (OSI) reference model and how it applies to IP.

5.1 OSI Reference Model

The International Organization for Standardization (ISO) developed the OSI reference model in the early 1980s, and it is now the standard for developing protocols that enable computers to communicate. Although not all protocols follow this model, many people use it to help them develop and teach new protocols.

The OSI reference model breaks up the problem of intermachine communication into seven layers. Each layer is concerned only with talking to its corresponding layer on the other machine as in the Figure 5.1. This means that Layer 5 has to worry only about talking to Layer 5 on the receiving machine, and not what the actual physical medium might be.

In addition, each layer of the OSI reference model provides services to the layer above it (Layer 5 to Layer 6, Layer 6 to Layer 7, and so on) and requests certain services from the layer directly below it (5 to 4, 4 to 3, and so on).

This layered approach enables each layer to handle a small piece of information, make any necessary changes to the data, and add the necessary functions for that layer before passing the data along.

Data becomes less human-like and more computer-like the further down the OSI reference model it traverses, until it becomes 1s and 0s (electrical impulses) at the physical layer. Figure 5.1 shows the OSI reference model.

The primary focus of this chapter is to discuss the application, presentation, session, transport, network, data link, and physical layers. Understanding these layers allows you to understand how IP routing works and how IP is transported across various media occurring at Layer 2 and Layer 1. (Tanenbaum,1999)

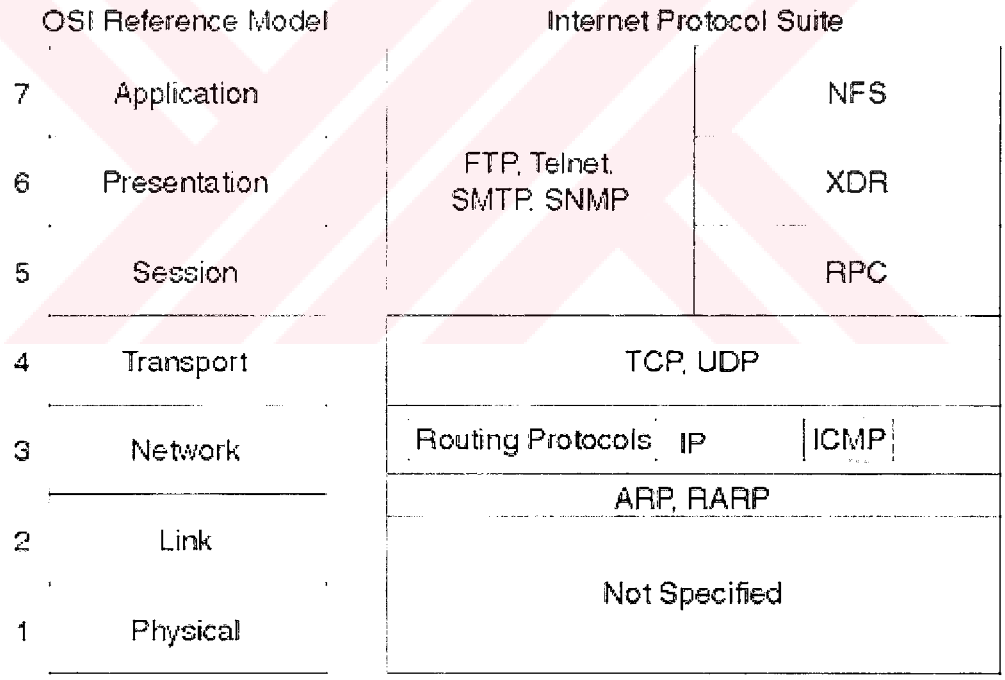

Figure 5. 1 OSI Reference Model

5.1.1 The Application Layer

Most users are familiar with the application layer. Some well-known applications include the following:

- E-mail
- Web browsing
- Word processing

5.1.2 The Presentation Layer

The presentation layer ensures that information sent by the application layer of one system is readable by the application layer of another system. If necessary, the presentation layer translates between multiple data formats by using a common data representation format.

The presentation layer concerns itself not only with the format and representation of actual user data, but also with data structures used by programs. Therefore, in addition to actual data format transformation (if necessary), the presentation layer negotiates data transfer syntax for the application layer.

5.1.3 The Session Layer

As its name implies, the session layer establishes, manages, and terminates sessions between applications. Sessions consist of dialogue between two or more presentation entities (recall that the session layer provides its services to the presentation layer).

The session layer synchronizes dialogue between presentation layer entities and manages their data exchange. In addition to basic regulation of conversations (sessions), the session layer offers provisions for data expedition, class of service (through the use of type of service [ToS] bits), and exception reporting of session-layer, presentation-layer, and application-layer problems.

5.1.4 The Transport Layer

The transport layer is responsible for ensuring reliable data transport on an internetwork. This is accomplished through flow control, error checking (checksum), end-to-end acknowledgments, retransmissions, and data sequencing.

Some transport layers, such as Transmission Control Protocol (TCP), have mechanisms for handling congestion. TCP adjusts its retransmission timer, for example, when congestion or packet loss occurs within a network. TCP slows down the amount of traffic it sends when congestion is present. Congestion is determined through the lack of acknowledgments received from the destination node.

5.1.5 The Network Layer

The network layer provides for the logical addressing which enables two disparate systems on different logical networks to determine a possible path to communicate. The network layer is the layer in which routing protocols reside.

On the Internet today, IP addressing is by far the most common addressing scheme in use. Routing protocols such as Enhanced Interior Gateway Routing Protocol (Enhanced IGRP, or EIGRP), Open Shortest Path First (OSPF), Border Gateway Protocol (BGP), Intermediary System to Intermediary System (IS-IS), and many others are used to determine the optimal routes between two logical subnetworks (subnets).

Traditional routers route IP packets based on their network layer address.Key functions of the network layer include the following:

- Packet formatting, addressing networks and hosts, address resolution, and routing
- Creating and maintaining routing tables

5.1.6 The Data Link Layer

The data link layer provides reliable transport across a physical link. The link layer has its own addressing scheme. This addressing scheme is concerned with physical connectivity and can transport frames based upon the data link layer address.

Traditional Ethernet switches switch network traffic based upon the data link layer (Layer 2) address. Switching traffic based on a Layer 2 address is generally known as bridging. In fact, an Ethernet switch is nothing more than a high-speed bridge with multiple interfaces.

5.1.7 The Physical Layer

The physical layer is concerned with creating 1s and 0s on the physical medium with electrical impulses/voltage changes. Common physical layer communication specifications include the following:

- EIA/TIA-232—Electrical Industries Association/Telecommunications Industry Association specification used for communicating between computer devices. You can use different connectors; this interface is often used for connecting computers to modems.
- V.35—International Telecommunication Union Telecommunication Standardization Sector (ITU-T) signaling mechanism that defines signaling rates from 19.2 Kbps to 1.544 Mbps. This physical interface is a 34-pin connector and also is known as a Winchester Block.
- RS-449—Uses 37 pins and is capable of longer runs than RS-232.

5.2 Internet Protocol

IP itself is a connectionless protocol that resides at Layer 3 (the network layer), which means that no reliability mechanisms, flow control, sequencing, or acknowledgments are present. Other protocols, such as TCP, can sit on top of IP (Layer 4, session) and can add flow control, sequencing, and other features.

Given IP's relative position in the OSI reference model, it doesn't have to deal with common data link issues such as Ethernet, Asynchronous Transfer Mode (ATM), Frame Relay, and Token Ring, or with physical issues such as Synchronous Optical Network (SONET), copper, and fiber. This makes IP virtually ubiquitous.

You can run IP into a home or business through any means necessary (for instance, wireless, broadband, or baseband). This doesn't mean that when you design a network you can ignore the lower two layers. It only means that they are independent of any applications you put on IP.

IP is considered a bursty protocol, which means that the applications residing above IP experience long periods of silence, followed by a need for a large portion of bandwidth. A good example of this is e-mail. If you set your mail package to download e-mail every 20 minutes, about 20 minutes of silence exist during which no bandwidth is needed.

One of the major benefits of IP is the ability to write an application *once* and have it delivered through an assorted type of media *anywhere*, regardless of whether this occurs through a digital subscriber line (DSL) connection in your home or a T1 line in your business.

You can address an IP packet in three general ways: through unicast, multicast, or broadcast mechanisms. Briefly explained, these three mechanisms provide the means for every IP packet to be labeled with a destination address, each in its unique way:

- Unicast is fairly simple, in that it identifies one specific address and only that node is supposed to send the packet to the higher layers of the OSI reference model.

- Broadcast packets are sent to all users on a local subnetwork. Broadcasts can traverse bridges and switches, but they are not passed through routers (unless they are specially configured to do so).
- Multicast packets use a special addressing range that enables a group of users on different subnetworks to receive the same flow. This enables the sender to send only one packet that several disparate hosts can receive.

Unicast, broadcast, and multicast packets each have a significant purpose. Unicast packets enable two stations to communicate with each other, regardless of physical location. Broadcast packets are used to communicate with everyone on a subnetwork simultaneously. Multicast packets enable applications, such as videoconferencing, that have one transmitter and multiple receivers. (Tanenbaum,1999)

Regardless of the type of IP packet used, data link layer addressing is always needed. Data link layer addresses are covered in detail in the next section.

5.3 Data Link Layer Addresses

The two types of addresses are data link layer and network layer addresses. Data link layer addresses—also known as Media Access Control (MAC) addresses and physical layer addresses—are unique to every device. In a local-area network (LAN), for instance, each device has a MAC address which identifies itself on the LAN. This enables computers to know who is sending what message. If you look at an Ethernet frame, the first 12 bytes are the destination and source MAC addresses.

If you use an Ethernet LAN switch, the traffic is routed through the switch based on the data link layer address (the MAC address). If you use a repeater or hub to connect the devices to the LAN, the packet is forwarded to all ports, regardless of the MAC address. This is because forwarding through a hub is based upon the physical layer and not the data link layer.

When traffic is routed based on the MAC layer address, it is generally referred to as being switched or bridged. Before routing became prominent in the late 1980s, many companies developed bridges to connect two disparate networks. This enabled a simple and inexpensive method of connecting two networks at the data link layer. Because these bridges did not look at the network layer address, however, unwanted traffic such as broadcasts and multicasts could be transmitted across the bridge, which consumed a large amount of bandwidth.

Most LANs in the 1980s and early 1990s used a hub to connect their Ethernet workstations. This device was known as a repeater and replicated the Layer 1 information only. So, if a corporation had an eight-port hub and one of the eight ports received a packet, the packet would be repeated (exactly, errors and all) to the other seven ports.

In the early 1990s, companies began developing LAN switches, which were basically a combination of a hub and bridge. In this scenario, the LAN switch learned which Layer 2 addresses were attached to each of its physical interfaces and forwarded traffic based on the Layer 2 address. If the switch did not have a list of a particular destination Layer 2 address in its switching table, or if the packet were a broadcast packet, the packet was repeated to all other interfaces on the switch.

This transition to network switches enabled networks to make better use of the available bandwidth. This saving in bandwidth was accomplished by preventing unnecessary IP packets from being transmitted on a physical port where the receiving device did not reside.

Now that you understand MAC addresses and how networks use them to route packets, it is time to discuss how networks use IP addressing to further route those packets.

5.4 IP Addressing

Of the different addressing schemes, IP addressing is the most important to understand because you must conceptually comprehend how these devices communicate to effectively build networks on top of an IP infrastructure.Many protocols exist, and each has a different addressing scheme.

Network layer addressing is normally hierarchical. As compared to the Public Switched Telephone Network (PSTN) in the North American Numbering Plan Association (NANPA) network of today, each Numbering Plan Area (NPA) includes a region, with a prefix (Nxx) denoting a sub-region and station identifier (xxxx) denoting the actual phone.

Network layer addressing lies at Layer 3 of the OSI model. This enables a group of computers to be given similar logical addresses. Logical addressing is similar to determining a person's address by looking at his or her country, state, ZIP code, city, and street address.

Routers forward traffic based on the Layer 3 or network layer address. IP addressing supports five network classes. The bits at the far left indicate the network class, as follows:

- Class A networks are intended mainly for use with a few large networks because they provide only seven bits for the network address field.
- Class B networks allocate 14 bits for the network address field and 16 bits for the host address field. This address class offers a good compromise between network and host address space.
- Class C networks allocate 21 bits for the network address field. They provide only 8 bits for the host field, however, so the number of hosts per network can be a limiting factor.
- Class D addresses are reserved for multicast groups, as described formally in RFC 1112. In class D addresses, the four highest-order bits are set to 1, 1, 1, and 0.
- Class E addresses also are defined by IP but are reserved for future use. In class E addresses, the four highest-order bits are set to 1, and the fifth bit is always 0.

IP addresses are written in dotted decimal format—for example, 121.10.3.116. Figure 5.2 shows the address formats for class A, B, and C IP networks. An easy way to think of class addressing is that the more networks you have, the fewer hosts you can have on that network.

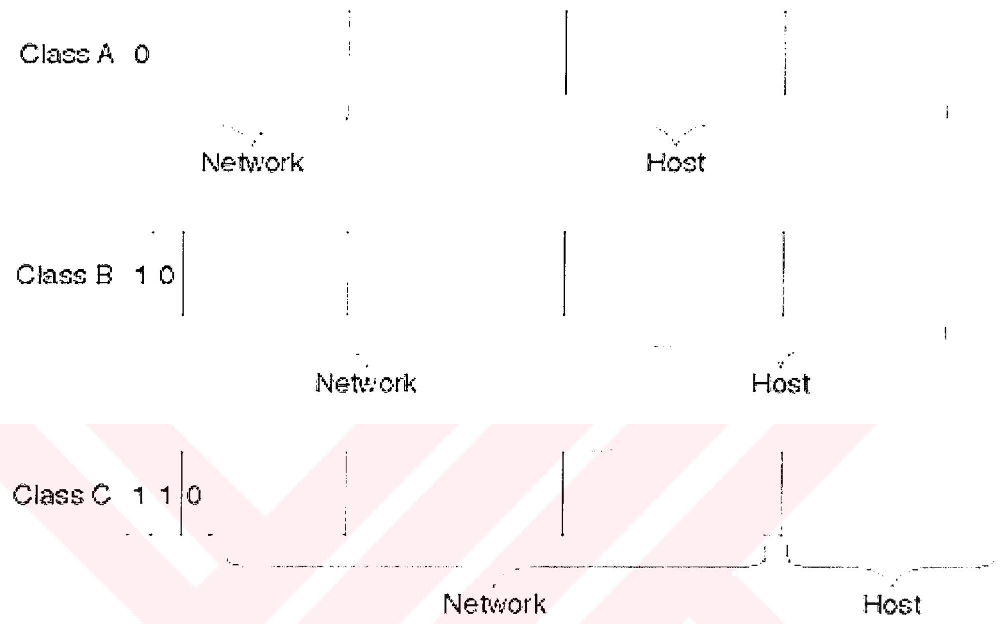

Figure 5. 2 Class A, B, and C Address Formats

You can also divide IP networks into smaller units called subnets. Subnets provide extra flexibility for network administrators. Assume, for example, that a network is assigned a class B address, and all the nodes on the network currently conform to a class B address format. Then assume that the dotted decimal representation of this network's address is 128.10.0.0 (all 0s in the host field of an address specify the entire network).

Rather than change all the addresses to some other basic network number, the administrator can subdivide the network using subnetting. He can do this by borrowing bits from the host portion of the address and using them as a subnet field, as shown in Figure 5.3.

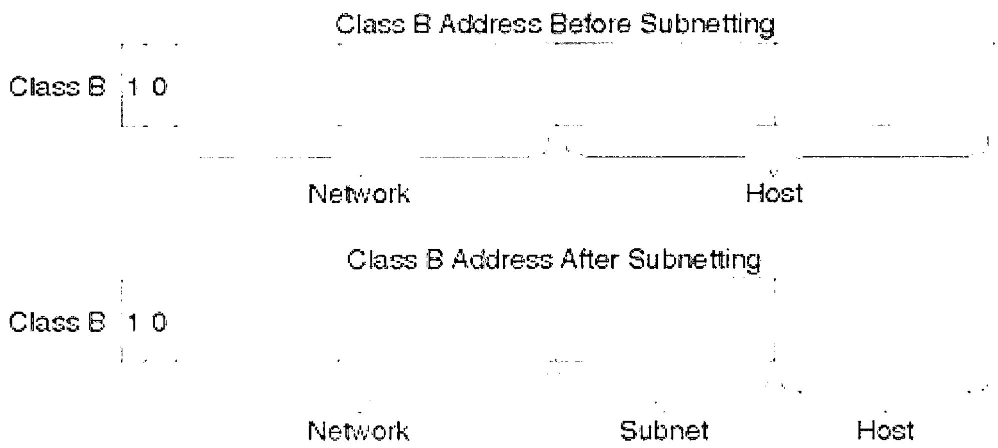

Figure 5. 3 Subnetting a Class B Address

Although this section discusses the makeup of IP addressing, it does not explain how a router knows where to send an IP packet. This is discussed in the next section.

5.5 Routing Protocols

IP is a routed protocol. A routed protocol is a packet that carries data. It is different from a routing protocol, in that the latter updates routers to let them know which path a packet should traverse.

Various routing protocols are used in IP internetworks today. It is important to note that with routing protocols, a well-engineered data network is self-healing and redundant, thus increasing the reliability of the network.

IP networks today use two main types of routing protocols: distance-vector routing and link-state routing. Within these two routing protocols are interior and exterior routing protocols.

Simply explained, distance-vector routing is concerned with how many hops (routers) are traversed, whereas link-state routing is concerned mainly with the state of the interfaces the router supports (in other words, whether they are up or down; hence, the name link state). (Tanenbaum,1999)

Interior routing protocols are usually used to update routers under the control of one administrative authority (autonomous system). Exterior routing protocols are usually used to enable networks in different autonomous systems to pass routing updates. A good example of an exterior routing protocol is the use of BGP on the Internet.

5.5.1 Distance-Vector Routing

Distance-vector routing is an algorithm that routers use to enable them to choose the best route. This algorithm uses the least number of hops (each router is a hop) to determine the best path to the destination.

Broadcasts are periodically sent to update adjacent routers. When the router first starts broadcasting updates, it includes all the reachable networks that are directly connected. The routes received by a router are kept in a routing table, which is then used to forward packets.

This method is bandwidth-intensive because the entire routing update is sent out periodically (usually every 30 seconds).

5.5.2 Link-State Routing

Link-state routing differs from distance-vector routing in that the former transmits routing updates only when the state of an interface changes. This means that traffic is sent and bandwidth is consumed only when an interface goes up or down.

5.5.3 BGP

BGP performs interdomain routing in Transmission Control Protocol/Internet Protocol (TCP/IP) networks. BGP is an Exterior Gateway Protocol (EGP), which means that it performs routing between multiple autonomous systems and exchanges routing and reachability information with other BGP systems.

BGP was developed to replace its predecessor, the now obsolete EGP, as the standard exterior gateway routing protocol used in the global Internet. BGP solves serious problems with EGP and scales to Internet growth more efficiently.

5.5.4 IS-IS

IS-IS is an OSI link-state hierarchical routing protocol. It floods the network with link-state information to build a complete, consistent picture of network topology.
To simplify router design and operation, IS-IS distinguishes between Level 1 and Level 2 Information Services (ISs):

- Level 1 ISs communicate with other Level 1 ISs in the same area.
- Level 2 ISs route between Level 1 areas and form an intradomain routing backbone.

Hierarchical routing simplifies backbone design because Level 1 ISs only need to know how to get to the nearest Level 2 IS. The backbone routing protocol also can change without impacting the intra-area routing protocol.

5.5.5 OSPF

OSPF is a link-state, Interior Gateway Routing Protocol (IGRP). It was designed to operate in TCP/IP networks and to address the shortcomings of the Router Information Protocol (RIP).

OSPF is derived from a number of sources, including the shortest path first (SPF) algorithm developed by Bolt, Beranek, and Newman, Inc. (BBN), an early version of the OSI IS-IS routing protocol, and other research efforts.

5.5.6 IGRP

IGRP is a robust protocol for routing within an autonomous system having arbitrarily complex topology and consisting of media with diverse bandwidth and delay characteristics.

Cisco Systems developed IGRP in the mid-1980s. It is a distance-vector interior gateway protocol that uses a combination of metrics to make routing decisions.

5.5.7 EIGRP

EIGRP is an enhanced version of the IGRP developed by Cisco Systems. EIGRP uses the same distance-vector algorithm and distance information as IGRP. EIGRP's convergence properties and operating efficiency are significantly better than those of IGRP.

EIGRP is a distance-vector interior gateway protocol that has the following features:

- It uses a combination of metrics to make routing decisions.
- It uses the Diffusing Update Algorithm (DUAL) to enable routes to converge quickly.
- It sends partial routing-table updates.
- It implements a neighbor discovery mechanism.

5.5.8 RIP

RIP is a distance-vector protocol that uses hop count as its metric. RIP is an Interior Gateway Protocol (IGP); it performs routing within a single autonomous system.

All these various routing protocols are used in different networks based upon their advantages and disadvantages. This book does not discuss in depth when to choose one over the other, but it is important to understand the basics about each protocol to further understand ways you can assemble IP networks.

It also is important to understand the different transport mechanisms that give IP different characteristics. These transport mechanisms are discussed next.

5.6 IP Transport Mechanisms

TCP and User Datagram Protocol (UDP) have different characteristics that various applications can use. If reliability is more important than delay, for instance, you can use TCP/IP to guarantee packet delivery. UDP/IP does not utilize packet re-transmissions, however. This can lower reliability, but in some cases a late retransmission is of no use.

To compare various transport layer protocols, you must first understand what makes up an IP packet. Figure 5.4 shows the fields of the IP packet.

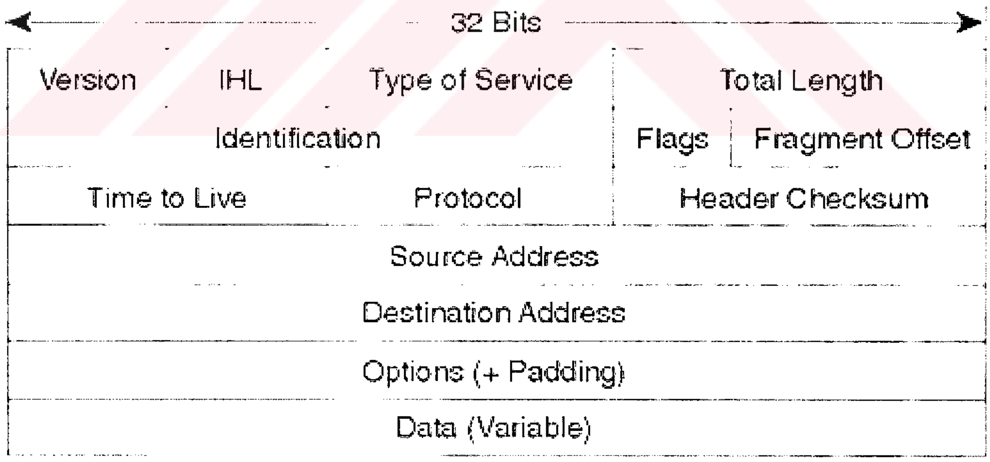

Figure 5. 4 IP Packet Fields

IP packet fields are defined as follows:

- Version—indicates whether IPv4 or IPv6 is being used.
- IP header length (IHL)—Indicates the datagram header length in 32-bit words.
- Type of service—Specifies how a particular upper-layer protocol wants the current datagram to be handled. You can assign packets various quality of service (QoS) levels based on this field.
- Total length—Specifies the length of the entire IP packet, including data and header, in bytes.
- Identification—Contains an integer that identifies the current datagram. This field is used to help piece together datagram fragments.
- Flags—A 3-bit field of which the low-order 2 bits control fragmentation. The high-order bit in this field is not used. One bit specifies whether you can fragment the packet; the second bit specifies whether the packet is the last fragment in a series of fragmented packets.
- Time To Live—Maintains a counter that gradually decrements down to zero, at which point the datagram is discarded. This keeps packets from looping endlessly.
- Protocol—Indicates which upper-layer protocol receives incoming packets after IP processing is complete.
- Header checksum—Verifies that the header is not corrupted.
- Source address—The sending address.
- Destination address—The address to receive the datagram.
- Options—Enables IP to support various options, such as security.
- Data—Contains application data as well as upper-layer protocol information.

5.6.1 TCP

TCP provides full-duplex, acknowledged, and flow-controlled service to upper-layer protocols. It moves data in a continuous, unstructured byte stream where bytes are identified by sequence numbers.

To maximize throughput, TCP enables each station to send multiple packets before an acknowledgment arrives. After the sender receives an acknowledgment for an outstanding packet, the sender slides the packet window along the byte stream and sends another packet. This flow control mechanism is known as a sliding window.

TCP can support numerous simultaneous upper-layer conversations. The port numbers in a TCP header identify an upper-layer conversation. Many well-known TCP ports are reserved for File Transfer Protocol (FTP), World Wide Web (WWW), Telnet, and so on.

Within the signaling portion of VoIP, TCP is used to ensure the reliability of the setup of a call. Due to the methods by which TCP operates, it is not feasible to use TCP as the mechanism to carry the actual voice in a VoIP call. With VoIP, packet loss is less important than latency.

The TCP packet fields are as follows:

- Source port and destination port—Identifies the points at which upper-layer source and destination processes receive TCP services.
- Sequence number—Usually specifies the number assigned to the first byte of data in the current message. Under certain circumstances, it also can be used to identify an initial sequence number to be used in the upcoming transmission.
- Acknowledgment number—Contains the sequence number of the next byte of data the sender of the packet expects to receive.
- Data offset—Indicates the number of 32-bit words in the TCP header.

- Reserved—Reserved for future use.
- Flags—Carry a variety of control information.
- Window—Specifies the size of the sender's receive window (that is, buffer space available for incoming data).
- Checksum—Indicates whether the header and data were damaged in transit.
- Urgent pointer—Points to the first urgent data byte in the packet.
- Options—Specifies various TCP options.
- Data—Contains upper-layer information.

5.6.2 UDP

UDP is a much simpler protocol than TCP and is useful in situations where the reliability mechanisms of TCP are unnecessary. UDP also is connectionless and has a smaller header, which translates to minimal overhead.

The UDP header has only four fields: source port, destination port, length, and UDP checksum. The source and destination port fields serve the same functions as they do in the TCP header. The length field specifies the length of the UDP header and data, and the checksum field enables packet integrity checking. The UDP checksum is optional.

UDP is used in VoIP to carry the actual voice traffic (the bearer channels). TCP is not used because flow control and retransmission of voice audio packets are unnecessary. Because UDP is used to carry the audio stream, it continues to transmit, regardless of whether you are experiencing 5 percent packet loss or 50 percent packet loss.

If TCP were utilized for VoIP, the latency incurred waiting for acknowledgments and retransmissions would render voice quality unacceptable. With VoIP and other real-time applications, controlling latency is more important than ensuring the reliable delivery of each packet.

CHAPTER SIX

VOIP:AN IN-DEPTH ANALYSIS

6 VoIP: An In-Depth Analysis

To create a proper network design, it is important to know all the caveats and inner workings of networking technology. This chapter explains many of the issues facing Voice over IP (VoIP) .

6.1 Delay/Latency

VoIP delay or latency is characterized as the amount of time it takes for speech to exit the speaker's mouth and reach the listener's ear.

Three types of delay are inherent in today's telephony networks: propagation delay, serialization delay, and handling delay.

Propagation delay is caused by the speed of light in fiber or copper-based networks. Handling delay—also called processing delay—defines many different causes of delay (actual packetization, compression, and packet switching) and is caused by devices that forward the frame through the network.Serialization delay is the amount of time it takes to actually place a bit or byte onto an interface. Serialization delay is not covered in depth in this book because its influence on delay is relatively minimal.

6.1.1 Propagation Delay

Light travels through a vacuum at a speed of 186,000 miles per second, and electrons travel through copper or fiber at approximately 125,000 miles per second.

A fiber network stretching halfway around the world (13,000 miles) induces a one-way delay of about 70 milliseconds (70 ms). Although this delay is almost imperceptible to the human ear, propagation delays in conjunction with handling delays can cause noticeable speech degradation.

6.1.2 Handling Delay

As mentioned previously, devices that forward the frame through the network cause handling delay. Handling delays can impact traditional phone networks, but these delays are a larger issue in packetized environments.

Vendors can decide how many speech samples they want to send in one packet. Because G.729 uses 10 ms speech samples, each increase in samples per frame raises the delay by 10 ms.

6.1.3 Queuing Delay

A packet-based network experiences delay for other reasons. Two of these are the time necessary to move the actual packet to the output queue (packet switching) and queuing delay.

When packets are held in a queue because of congestion on an outbound interface, the result is queuing delay. Queuing delay occurs when more packets are sent out than the interface can handle at a given interval.

Packet-based solutions, including PC-based solutions, are not as good at determining packet destination and moving the actual packet to the output queue.
The actual queuing delay of the output queue is another cause of delay. You should keep this factor to less than 10 ms whenever you can by using whatever queuing methods are optimal for your network

The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) G.114 recommendation specifies that for good voice quality, no more than 150 ms of one-way, end-to-end delay should occur, as shown in Figure 6.1.

With vendors of VoIP equipment implementation, two routers with minimal network delay (back to back) use only about 60 ms of end-to-end delay. This leaves up to 90 ms of network delay to move the IP packet from source to destination.

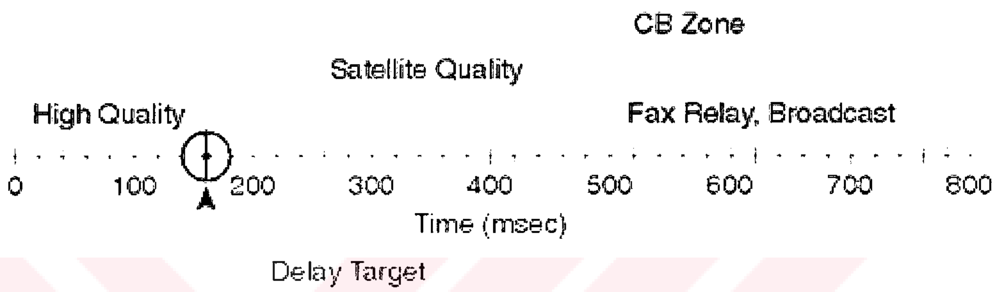

Figure 6. 1 End-to-End Delay

As shown in Figure 6.1 , some forms of delay are longer, although accepted, because no other alternatives exist. In satellite transmission, for example, it takes approximately 250 ms for a transmission to reach the satellite, and another 250 ms for it to come back down to Earth. This results in a total delay of 500 ms.

Although the ITU-T recommendation notes that this is outside the acceptable range of voice quality, many conversations occur every day over satellite links. As such, voice quality is often defined as what users will accept and use.

In an unmanaged, congested network, queuing delay can add up to two seconds of delay (or result in the packet being dropped). This lengthy period of delay is unacceptable in almost any voice network. Queuing delay is only one component of end-to-end delay. Another way end-to-end delay is affected is through jitter.

6.2 Jitter

Simply stated, jitter is the variation of packet interarrival time. Jitter is one issue that exists only in packet-based networks. While in a packet voice environment, the sender is expected to reliably transmit voice packets at a regular interval (for example, send one frame every 20 ms).

These voice packets can be delayed throughout the packet network and not arrive at that same regular interval at the receiving station (for example, they might not be received every 20 ms; see Figure 6.2). The difference between when the packet is expected and when it is actually received is jitter.

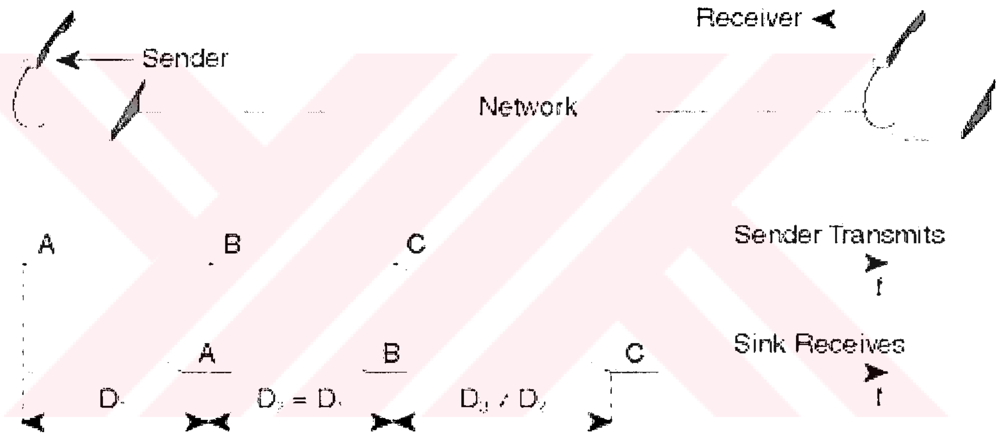

Figure 6. 2 Variation of Packet Arrival Time (Jitter)

In Figure 6.2 , we can see that the amount of time it takes for packets A and B to send and receive is equal (D1=D2). Packet C encounters delay in the network, however, and is received after it is expected. This is why a jitter buffer, which conceals interarrival packet delay variation, is necessary.

Note that jitter and total delay are not the same thing, although having plenty of jitter in a packet network can increase the amount of total delay in the network. This is because the more jitter you have, the larger your jitter buffer needs to be to compensate for the unpredictable nature of the packet network.

If your data network is engineered well and you take the proper precautions, jitter is usually not a major problem and the jitter buffer does not significantly contribute to the total end-to-end delay.

RTP timestamps are used to determine what level of jitter, if any, exists within the network. The jitter buffer found within software is considered a dynamic queue. This queue can grow or shrink exponentially depending on the interarrival time of the RTP packets.

Although many vendors choose to use static jitter buffers, some vendors found that a well-engineered dynamic jitter buffer is the best mechanism to use for packet-based voice networks. Static jitter buffers force the jitter buffer to be either too large or too small, thereby causing the audio quality to suffer, due to either lost packets or excessive delay.

6.3 Pulse Code Modulation

Although analog communication is ideal for human communication, analog transmission is neither robust nor efficient at recovering from line noise. In the early telephony network, when analog transmission was passed through amplifiers to boost the signal, not only was the voice boosted but the line noise was amplified, as well. This line noise resulted in an often-unusable connection.

It is much easier for digital samples, which are comprised of 1 and 0 bits, to be separated from line noise. Therefore, when analog signals are regenerated as digital samples, a clean sound is maintained. When the benefits of this digital representation became evident, the telephony network migrated to pulse code modulation (PCM).

As covered in previous chapters , PCM converts analog sound into digital form by sampling the analog sound 8000 times per second and converting each sample into a numeric code. The Nyquist theorem states that if you sample an analog signal at twice the rate of the highest frequency of interest, you can accurately reconstruct that signal back into its analog form. Because most speech content is below 4000 Hz (4 kHz), a sampling rate of 8000 times per second (125 ms between samples) is required.

6.4 Voice Compression

Two basic variations of 64 Kbps PCM are commonly used: μ-law and a-law. The methods are similar in that they both use logarithmic compression to achieve 12 to 13 bits of linear PCM quality in 8 bits, but they are different in relatively minor compression details (μ-law has a slight advantage in low-level, signal-to-noise ratio performance). Usage is historically along country and regional boundaries, with North America using μ-law and Europe using a-law modulation. It is important to note that when making a long-distance call, any required μ-law to a-law conversion is the responsibility of the μ-law country.

Another compression method used often is adaptive differential pulse code modulation (ADPCM). A commonly used instance of ADPCM is ITU-T G.726, which encodes using 4-bit samples, giving a transmission rate of 32 Kbps. Unlike PCM, the 4 bits do not directly encode the amplitude of speech, but they do encode the differences in amplitude, as well as the rate of change of that amplitude, employing some rudimentary linear prediction.

PCM and ADPCM are examples of waveform codecs—compression techniques that exploit redundant characteristics of the waveform itself. New compression techniques were developed over the past 10 to 15 years that further exploit knowledge of the source characteristics of speech generation. These techniques employ signal processing procedures that compress speech by sending only simplified parametric information about the original speech excitation and vocal tract shaping, requiring less bandwidth to transmit that information.

6.4.1 Mean Opinion Score

We can test voice quality in two ways: subjectively and objectively. Humans perform subjective voice testing, whereas computers—which are less likely to be "fooled" by compression schemes that can "trick" the human ear—perform objective voice testing.

Codecs are developed and tuned based on subjective measurements of voice quality. Standard objective quality measurements, such as total harmonic distortion and signal-to-noise ratios, do not correlate well to a human's perception of voice quality, which in the end is usually the goal of most voice compression techniques.

A common subjective benchmark for quantifying the performance of the speech codec is the mean opinion score (MOS). MOS tests are given to a group of listeners. Because voice quality and sound in general are subjective to listeners, it is important to get a wide range of listeners and sample material when conducting a MOS test. The listeners give each sample of speech material a rating of 1 (bad) to 5 (excellent). The scores are then averaged to get the mean opinion score.

MOS testing also is used to compare how well a particular codec works under varying circumstances, including differing background noise levels, multiple encodes and decodes, and so on. You can then use this data to compare against other codecs.

6.4.2 Perceptual Speech Quality Measurement

Although MOS scoring is a subjective method of determining voice quality, it is not the only method for doing so. The ITU-T put forth recommendation P.861, which covers ways you can objectively determine voice quality using Perceptual Speech Quality Measurement (PSQM). (Rodriguez et al,2001)

PSQM has many drawbacks when used with voice codecs (vocoders). One drawback is that what the "machine" or PSQM hears is not what the human ear perceives. In layman's terms, a person can trick the human ear into perceiving a higher-quality voice, but a computer cannot. Also, PSQM was developed to "hear" impairments caused by compression and decompression and not packet loss or jitter.

6.5 Echo

Echo is on a phone conversation can range from slightly annoying to unbearable, making conversation unintelligible.

Hearing your own voice in the receiver while you are talking is common and reassuring to the speaker. Hearing your own voice in the receiver after a delay of more than about 25 ms, however, can cause interruptions and can break the cadence in a conversation.

In a traditional toll network, echo is normally caused by a mismatch in impedance from the four-wire network switch conversion to the two-wire local loop (as shown in Figure 6.3). Echo, in the standard Public Switched Telephone Network (PSTN), is regulated with echo cancellers and a tight control on impedance mismatches at the common reflection points, as depicted in Figure 6.3.

Echo has two drawbacks: It can be loud, and it can be long. The louder and longer the echo, of course, the more annoying the echo becomes.

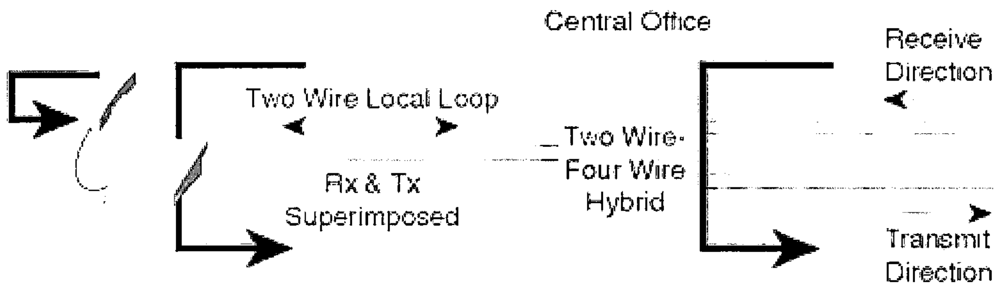

Figure 6. 3 Echo Caused by Impedance Mismatch

Telephony networks in those parts of the world where analog voice is primarily used employ echo suppressors, which remove echo by capping the impedance on a circuit. This is not the best mechanism to use to remove echo and, in fact, causes other problems. You cannot use Integrated Services Digital Network (ISDN) on a line that has an echo suppressor, for instance, because the echo suppressor cuts off the frequency range that ISDN uses.

In today's packet-based networks, you can build echo cancellers into low-bit-rate codecs and operate them on each DSP. In some manufacturers' implementations, echo cancellation is done in software; this practice drastically reduces the benefits of echo cancellation.

To understand how echo cancellers work, it is best to first understand where the echo comes from.In this example, assume that user A is talking to user B. The speech of user A to user B is called G. When G hits an impedance mismatch or other echo-causing environments, it bounces back to user A. User A can then hear the delay several milliseconds after user A actually speaks.

To remove the echo from the line, the device user A is talking through (router A) keeps an inverse image of user A's speech for a certain amount of time. This is called inverse speech (–G). This echo canceller listens for the sound coming from user B and subtracts the –G to remove any echo.

Echo cancellers are limited by the total amount of time they wait for the reflected speech to be received, a phenomenon known as echo tail. Routers have configurable echo tails of 16, 24, and 32 ms.

It is important to configure the appropriate amount of echo cancellation when initially installing VoIP equipment. If you don't configure enough echo cancellation, callers will hear echo during the phone call. If you configure too much echo cancellation, it will take longer for the echo canceller to converge and eliminate the echo.

6.6 Packet Loss

Packet loss in data networks is both common and expected. Many data protocols, in fact, use packet loss so that they know the condition of the network and can reduce the number of packets they are sending.When putting critical traffic on data networks, it is important to control the amount of packet loss in that network.

With protocols such as Systems Network Architecture (SNA) that do *not* tolerate packet loss well, you need to build a well-engineered network that can prioritize the time-sensitive data ahead of data that can handle delay and packet loss.

When putting voice on data networks, it is important to build a network that can successfully transport voice in a reliable and timely manner. Also, it is helpful when you can use a mechanism to make the voice somewhat resistant to periodic packet loss.

Vendors of VoIP equipment, developed many quality of service (QoS) tools that enable administrators to classify and manage traffic through a data network. If a data network is well engineered, you can keep packet loss to a minimum.

Vendors VoIP implementation enables the voice router to respond to periodic packet loss. If a voice packet is not received when expected (the expected time is variable), it is assumed to be lost and the last packet received is replayed, as shown in Figure 6.4. Because the packet lost is only 20 ms of speech, the average listener does not notice the difference in voice quality.

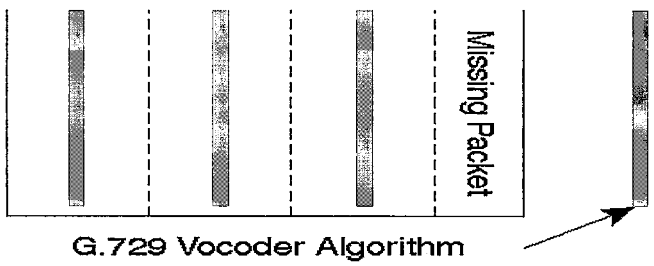

Figure 6. 4 Packet Loss with G.729

Using G.729 implementation for VoIP, let's say that each of the lines in Figure 6.4 represents a packet. Packets 1, 2, and 3 reach the destination, but packet 4 is lost somewhere in transmission. The receiving station waits for a period of time (per its jitter buffer) and then runs a concealment strategy.

This concealment strategy replays the last packet received (in this case, packet 3), so the listener does not hear gaps of silence. Because the lost speech is only 20 ms, the listener most likely does not hear the difference. You can accomplish this concealment strategy only if one packet is lost. If multiple consecutive packets are lost, the concealment strategy is run only once until another packet is received.

Because of the concealment strategy of G.729, as a rule of thumb G.729 is tolerant to about five percent packet loss averaged across an entire call.

6.7 Voice Activity Detection

In normal voice conversations, someone speaks and someone else listens. Today's toll networks contain a bi-directional, 64,000 bit per second (bps) channel, regardless of whether anyone is speaking. This means that in a normal conversation, at least 50 percent of the total bandwidth is wasted. The amount of wasted bandwidth can actually be much higher if you take a statistical sampling of the breaks and pauses in a person's normal speech patterns.

When using VoIP, you can utilize this "wasted" bandwidth for other purposes when voice activity detection (VAD) is enabled. As shown in Figure 6.5, VAD works by detecting the magnitude of speech in decibels (dB) and deciding when to cut off the voice from being framed.

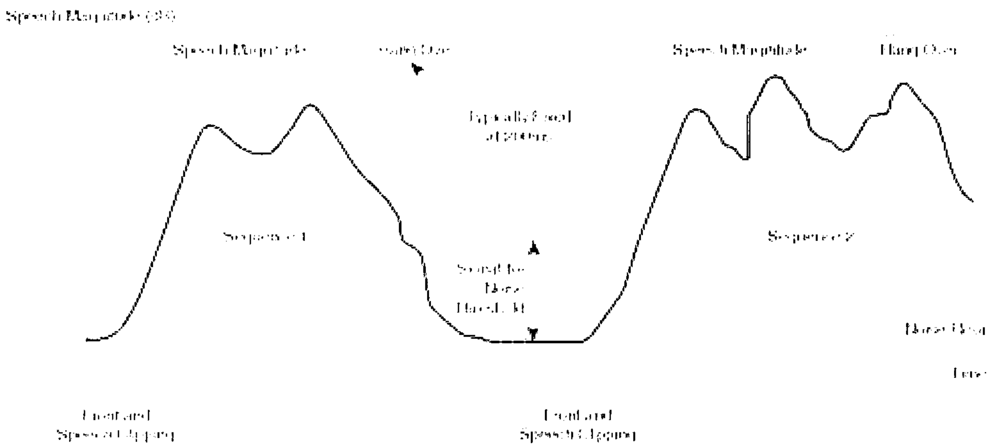

Figure 6. 5 Voice Activity Detection

Typically, when the VAD detects a drop-off of speech amplitude, it waits a fixed amount of time before it stops putting speech frames in packets. This fixed amount of time is known as hangover and is typically 200 ms.

With any technology, tradeoffs are made. VAD experiences certain inherent problems in determining when speech ends and begins, and in distinguishing speech from background noise. This means that if you are in a noisy room, VAD is unable to distinguish between speech and background noise. This also is known as the signal-to-noise threshold . In these scenarios, VAD disables itself at the beginning of the call.

Another inherent problem with VAD is detecting when speech begins. Typically the beginning of a sentence is cut off or clipped (refer to Figure 6.5). This phenomenon is known as front-end speech clipping. Usually, the person listening to the speech does not notice front-end speech clipping. (Rodriguez et al,2001)

6.8 Digital-to-Analog Conversion

Digital to analog (D/A) conversion issues also currently plague toll networks. Although almost all the telephony backbone networks in first-world countries today are digital, sometimes multiple D/A conversions occur.

Each time a conversion occurs from digital to analog and back, the speech or waveform becomes less "true." Although today's toll networks can handle at least seven D/A conversions before voice quality is affected, compressed speech is less robust in the face of these conversions.

It is important to note that D/A conversion must be tightly managed in a compressed speech environment. When using G.729, just two conversions from D/A cause the MOS score to decrease rapidly. The only way to manage D/A conversion is to have the network designer design VoIP environments with as few D/A conversions as possible.

Although D/A conversions affect all voice networks, VoIP networks using a PCM codec (G.711) are just as resilient to problems caused by D/A conversions as today's telephony networks are.

6.9 Tandem Encoding

As covered in Chapter 1, all circuit-switched networks today work on the premise of switching calls at the data link layer. The circuit switches are organized in a hierarchical model in which switches higher in the hierarchy are called tandem switches.

Tandem switches do not actually terminate any local loops; rather, they act as a higher-layer circuit switch. In the hierarchical model, several layers of tandem circuit switches can exist. This enables end-to-end connectivity for anyone with a phone, without the need for a direct connection between every home on the planet.

If the TDM switches compress voice and the tandem switch must decompress and recompress the voice, the voice quality can be drastically affected. Although compression and recompression are not common in the PSTN today, you must plan for it and design around it in packet networks.

Voice degradation occurs when you have more than one compression/decompression cycle for each phone call. Figure 6.6 provides an example of when this scenario might occur.

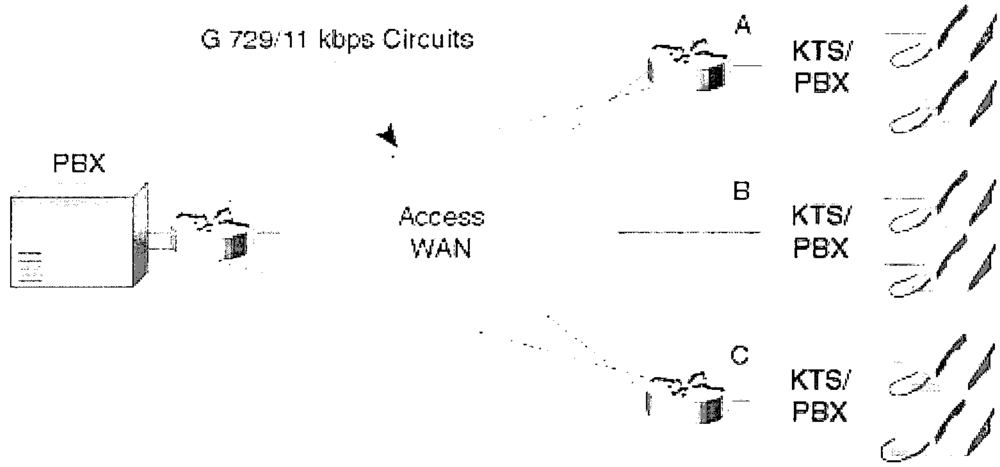

Figure 6. 6 VoIP Tandem Encoding

Figure 6.6 depicts three VoIP routers connected and acting as tie-lines between one central-site PBX and three remote-branch PBXs. The network is designed to put all the dial-plan information in the central-site PBX. This is common in many enterprise networks to keep the administration of the dial plan centralized.

A drawback to tandem encoding when used with VoIP is that, if a telephony user at branch B wants to call a user at branch C, two VoIP ports at central site A must be utilized. Also, two compression/decompression cycles exist, which means that voice quality will degrade.

Different codecs react differently to tandem encoding. G.729 can handle two compression/decompression cycles, while G.723.1 is less resilient to multiple compression cycles.

Assume, for example, that a user at remote site B wants to call a user at remote site C. The call goes through PBX B, is compressed and packetized at VoIP router B, and is sent to the central site VoIP router A, which decompresses the call and sends it to PBX A. PBX A circuit-switches the call back to its VoIP router (router A), which compresses and packetizes the call, and sends it to the remote site C, where it is then decompressed and sent to PBX C. This process is known as tandem-compression; you should avoid it in all networks where compression exists.

It is easy to avoid tandem compression. This customer simplified the router configuration at the expense of voice quality.

Taking the same example of three PBXs connected through three VoIP routers, but configuring the VoIP routers differently, simplifies the call-flow and avoids tandem encoding, as shown in Figure 6.7.

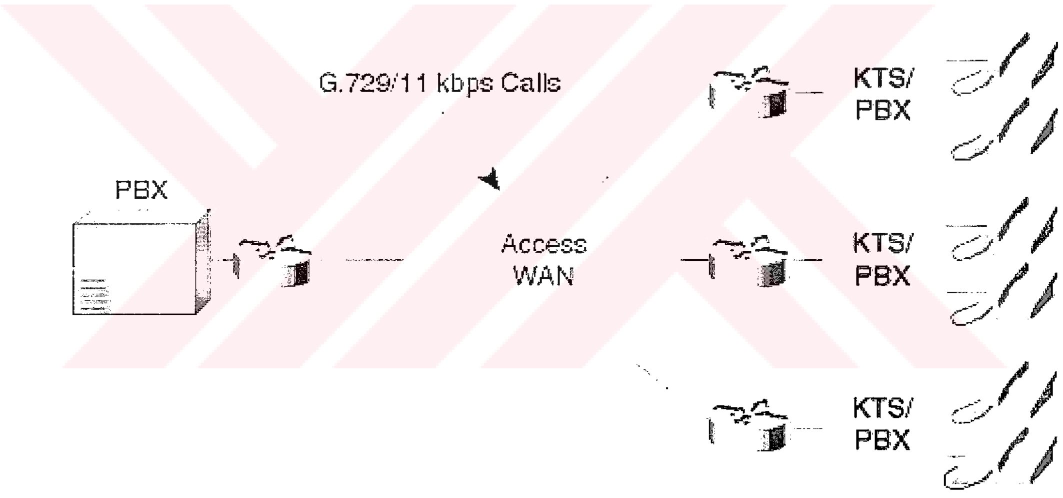

Figure 6. 7 VoIP Without Tandem Encoding

You can see one of IP's strengths in Figure 6.7 : a tie-line does not have to be leased from the telephone company to complete calls between two PBXs. If a data network connects the sites, VoIP can ride across that network.

The dial plan is moved from the central-site PBX to each of the VoIP routers. This enables each VoIP device to make a call-routing decision and removes the need for tie-lines. The major benefit of this change is the removal of needless compression/decompression cycles.

6.10 Transport Protocols

Two main types of traffic ride upon Internet Protocol (IP): User Datagram Protocol (UDP) and Transmission Control Protocol (TCP). In general, you use TCP when you need a reliable connection and UDP when you need simplicity and reliability is not your concern.

Due to the time-sensitive nature of voice traffic, UDP/IP was the logical choice to carry voice. More information was needed on a packet-by-packet basis than UDP offered, however. So, for real-time or delay-sensitive traffic, the Internet Engineering Task Force (IETF) adopted the RTP. VoIP rides on top of RTP, which rides on top of UDP. Therefore, VoIP is carried with an RTP/UDP/IP packet header.

6.10.1 RTP

RTP is the standard for transmitting delay-sensitive traffic across packet-based networks. RTP rides on top of UDP and IP. RTP gives receiving stations information that is not in the connectionless UDP/IP streams. (Rodriguez et al,2001)

As shown in Figure 6.8, two important bits of information are sequence information and timestamping. RTP uses the sequence information to determine whether the packets are arriving in order, and it uses the time-stamping information to determine the interarrival packet time (jitter).

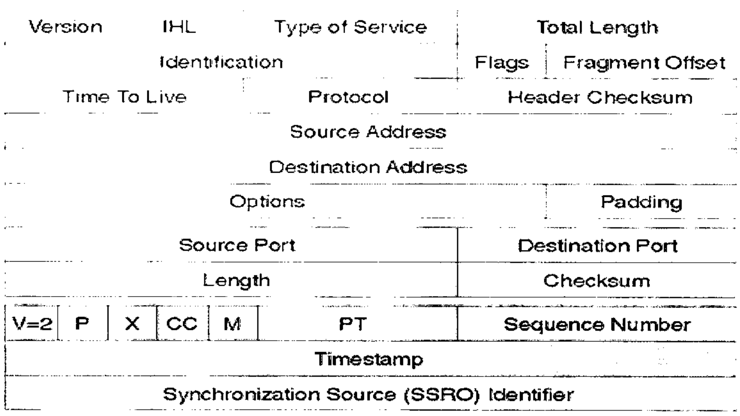

Figure 6. 8 Real-Time Transport Header

You can use RTP for media on demand, as well as for interactive services such as Internet telephony. RTP (refer to Figure 6.8) consists of a data part and a control part, the latter called RTP Control Protocol (RTCP).

The data part of RTP is a thin protocol that provides support for applications with real-time properties, such as continuous media (for example, audio and video), including timing reconstruction, loss detection, and content identification.

RTCP provides support for real-time conferencing of groups of any size within an Internet. This support includes source identification and support for gateways, such as audio and video bridges as well as multicast-to-unicast translators. It also offers QoS feedback from receivers to the multicast group, as well as support for the synchronization of different media streams.

Using RTP is important for real-time traffic, but a few drawbacks exist. The IP/RTP/UDP headers are 20, 8, and 12 bytes, respectively. This adds up to a 40-byte header, which is twice as big as the payload when using G.729 with two speech samples (20 ms).

6.10.2 Reliable User Data Protocol

Reliable User Data Protocol (RUDP) builds in some reliability to the connectionless UDP protocol. RUDP enables reliability without the need for a connection-based protocol such as TCP. The basic method of RUDP is to send multiples of the same packet and enable the receiving station to discard the unnecessary or redundant packets. This mechanism makes it more probable that one of the packets will make the journey from sender to receiver.

This also is known as forward error correction(FEC). Few implementations of FEC exist due to bandwidth considerations (a doubling or tripling of the amount of bandwidth used). Customers that have almost unlimited bandwidth, however, consider FEC a worthwhile mechanism to enhance reliability and voice quality.

6.11 Dial-Plan Design

One of the areas that causes the largest amount of headaches when designing an Enterprise Telephony (ET) network is the dial plan. The causes of these head pains might be due to the complex issues of integrating disparate networks. Many of these disparate networks were not designed for integration.

A good data example of joining disparate networks is when two companies merge. In such a scenario, the companies' data networks (IP addressing, ordering applications, and inventory database) must be joined. It is highly improbable that both companies used the same methodologies when implementing their data networks, so problems can arise.

The same problems can occur in telephony networks. If two companies merge, their phone systems (voice mail, billing, supplementary features, and dial-plan addressing) might be incompatible with each other.

These dial-plan issues also can occur when a company decides to institute a corporate dial plan. Consider Company X, for example. Company X grew drastically in the last three years and now operates 30 sites throughout the world, with its headquarters in Dallas. Company X currently dials through the PSTN to all its 29 remote sites. Company X wants to simplify the dialing plan to all its remote sites to enable better employee communication and ease of use.

Company X currently has a large PBX at its headquarters and smaller PBX systems at its remote sites. Several alternatives are available to this company:

- Purchase leased lines between headquarters and all remote sites.
- Purchase a telephony Virtual Private Network (VPN) from the telephone company and dial an access code from anywhere to access the VPN.
- Take advantage of the existing data infrastructure and put voice on the data network.

Regardless of which option Company X chooses, it must face dial-plan design, network management, and cost issues. Without getting into great detail, most companies must decide on their dial-plan design based on the following issues:

- Plans for growth
- Cost of leased circuits or VPNs
- Cost of additional equipment for packet voice
- Number overlap (when more than one site has the same phone number)
- Call-flows (the call patterns from each site)
- Busy hour (the time of day when the highest number of calls are offered on a circuit)

Depending on the size of the company, the dial plan can stretch from two digits to seven or eight digits. It is important that you not force yourself down a particular path until you address the previous issues.

Company X plans on sustaining 20–30 percent growth and decides on a seven-digit dial plan based on its growth patterns. This choice also cuts down on the number overlap that might be present.

Company X will have a three-digit site code, and four digits for the actual subscriber line. It made this decision because it does not believe it will have more than 999 branch offices.(Murhammer et al,1999)

CHAPTER SEVEN
VOICE OVER IP QoS

7 Quality of Service

Quality of service (QoS) is an often-used and misused term that has a variety of meanings. In this book, QoS refers to both class of service (CoS) and type of service (ToS). The basic goal of CoS and ToS is to achieve the bandwidth and latency needed for a particular application.

A CoS enables a network administrator to group different packet flows, each having distinct latency and bandwidth requirements. A ToS is a field in an Internet Protocol (IP) header that enables CoS to take place. Currently, a ToS field uses three bits, which allow for eight packet-flow groupings, or CoSs (0-7). New Requests For Comments (RFCs) will enable six bits in a ToS field to allow for more CoSs.

Various tools are available to achieve the necessary QoS for a given user and application. This chapter discusses these tools, when to use them, and potential drawbacks associated with some of them.

It is important to note that the tools for implementing these services are not as important as the end result achieved. In other words, do not focus on one QoS tool to solve all your QoS problems. Instead, look at the network as a whole to determine which tools, if any, belong in which portions of your network.

In a well-engineered network, you must be careful to separate functions that occur on the edges of a network from functions that occur in the core or backbone of a network. It is important to separate edge and backbone functions to achieve the best QoS possible.

Voice over IP (VoIP) comes with its own set of problems. QoS can help solve some of these problems—namely, packet loss, jitter, and handling delay.

Some of the problems QoS *cannot* solve are propagation delay, codec delay, sampling delay, and digitization delay.

A VoIP phone call can be equivalent to any other large expense you would plan for. Therefore, it is important to know which parts of the budget you cannot change and which parts you might be able to control, as shown in Figure 7.1.

	Fixed Delay	Variable Delay
Coder Delay G.729 (5 ms Look Ahead)	5 ms	
Coder Delay G.729 (10 ms Per Frame)	20 ms	
Packetization Delay Included in Coder Delay		
Queuing Delay 64 kbps Trunk		6 ms
Serialization Delay 64 kbps Trunk	3 ms	
Propagation Delay (Private Lines)	32 ms	
Network Delay (For Example, Public Frame Relay Svc)		
Dejitter Buffer		2-200 ms
Total - Assuming 50 ms Jitter Buffer	**110 ms**	

Figure 7. 1 End-to-End Delay Budget

The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) G.114 recommendation suggests no more than 150 milliseconds (ms) of end-to-end delay to maintain "good" voice quality. Any customer's definition of "good" might mean more or less delay, so keep in mind that 150 ms is merely a recommendation.

7.1 Edge Functions

When designing a VoIP network, edge functions usually correspond to wide-area networks (WANs) that have less than a T1 or E1 line of bandwidth from the central site. This is not a fixed rule but merely a rule of thumb to follow so that you know when to use edge functions and when to use backbone functions.

7.1.1 Bandwidth Limitations

The first issue of major concern when designing a VoIP network is bandwidth constraints. Depending upon which codec you use and how many voice samples you want per packet, the amount of bandwidth per call can increase drastically. For an explanation of packet sizes and bandwidth consumed, see Table 7.1.

Table 7. 1 Codec Type and Sample Size Effects on Bandwidth

Codec	Bandwidth Consumed	Bandwidth Consumed with cRTP (2-Byte Header)	Sample Latency
G.729 w/ one 10-ms sample/frame	40 kbps	9.6 kbps	15 ms
G.729 w/ four 10-ms samples/frame	16 kbps	8.4 kbps	45 ms
G.729 w/ two 10-ms samples/frame	24 kbps	11.2 kbps	25 ms
G.711 w/ one 10-ms sample/frame	112 kbps	81.6 kbps	10 ms
G.711 w/ two 10-ms samples/frame	96 kbps	80.8 kbps	20 ms

Using G.729 with two 10-ms samples as an example, without RTP header compression, 24 kbps are consumed in each direction per call. Although this might not be a large amount for T1 (1.544-mbps), E1 (2.048-mbps), or higher circuits, it is a large amount (42 percent) for a 56-kbps circuit.

Also, keep in mind that the bandwidth in Table 7.1 does not include Layer 2 headers (PPP, Frame Relay, and so on). It includes headers from Layer 3 (network layer) and above only. Therefore, the same G.729 call can consume different amounts of bandwidth based upon which data link layer is used (Ethernet, Frame Relay, PPP, and so on).

7.1.1.1 cRTP

To reduce the large percentage of bandwidth consumed by a G.729 voice call, you can use cRTP. cRTP enables you to compress the 40-byte IP/RTP/UDP header to 2 to 4 bytes most of the time (see Figure 7.1) (Jacobson,1990)

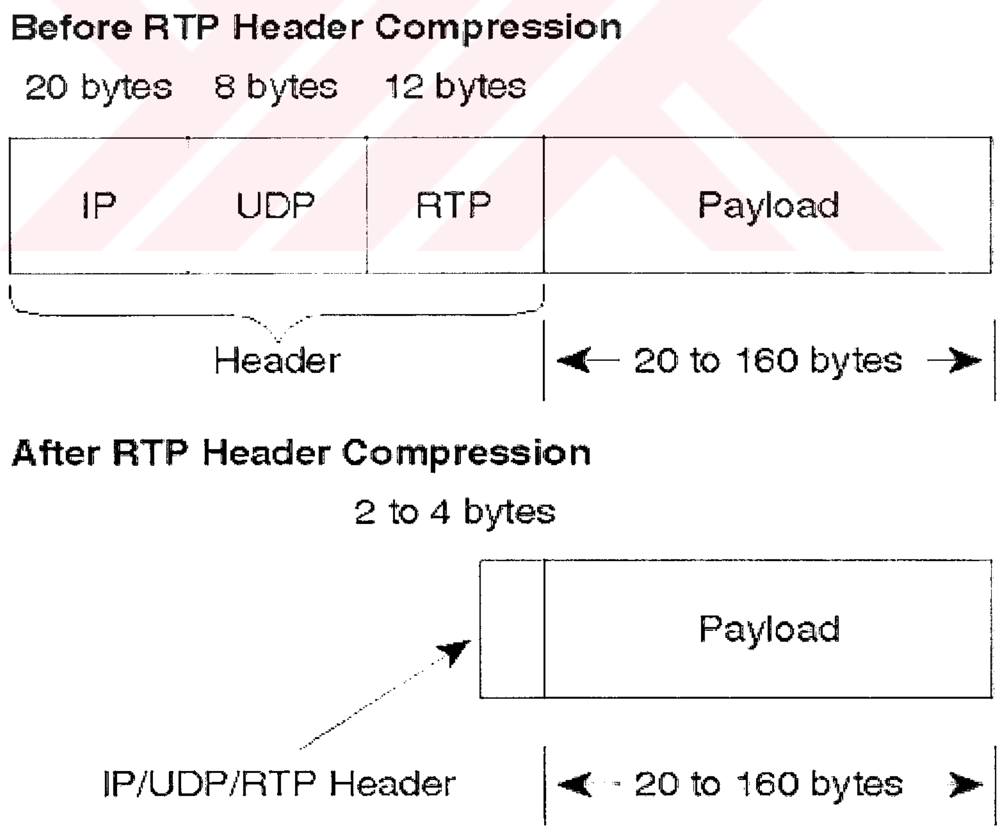

Figure 7. 2 RTP Header Compression

With cRTP, the amount of traffic per VoIP call is reduced from 24 kbps to 11.2 kbps. This is a major improvement for low-bandwidth links. A 56-kbps link, for example, can now carry four G.729 VoIP calls at 11.2 kbps each. Without cRTP, only two G.729 VoIP calls at 24 kbps can be used. (Jacobson,1990)

To avoid the unnecessary consumption of available bandwidth, cRTP is used on a link-by-link basis. This compression scheme reduces the IP/RTP/UDP header to 2 bytes when UDP checksums are not used, or 4 bytes when UDP checksums are used. cRTP uses some of the same techniques as Transmission Control Protocol (TCP) header compression. In TCP header compression, the first factor-of-two reduction in data rate occurs because half of the bytes in the IP and TCP headers remain constant over the life of the connection.

cRTP Caveats

You should not use cRTP on high-speed interfaces, as the disadvantages of doing so outweigh the advantages. "High-speed network" is a relative term: Usually anything higher than T1 or E1 speed does not need cRTP, but in some networks 512 kbps can qualify as a high-speed connection.

As with any compression, the CPU incurs extra processing duties to compress the packet. This increases the amount of CPU utilization on the router. Therefore, you must weigh the advantages (lower bandwidth requirements) against the disadvantages (higher CPU utilization). A router with higher CPU utilization can experience problems running other tasks. As such, it is usually a good rule of thumb to keep CPU utilization at less than 60 to 70 percent to keep your network running smoothly. (Jacobson,1990)

7.1.2 Queuing

Queuing in and of itself is a fairly simple concept. The easiest way to think about queuing is to compare it to the highway system. Let's say you are on the İzmir-

Cesme Turnpike driving at a decent speed. When you approach a tollbooth, you must slow down, stop, and pay the toll. During the time it takes to pay the toll, a backup of cars ensues, creating congestion.

As in the tollbooth line, in queuing the concept of first in, first out (FIFO) exists, which means that if you are the first to get in the line, you are the first to get out of the line. FIFO queuing was the first type of queuing to be used in routers, and it is still useful depending upon the network's topology.

Today's networks, with their variety of applications, protocols, and users, require a way to classify different traffic. Going back to the tollbooth example, a special "lane" is necessary to enable some cars to get bumped up in line. The İzmir-Cesme Turnpike, as well as many other toll roads, has a carpool lane, or a lane that allows you to pay for the toll electronically, for instance.

Likewise, vendors has several queuing tools that enable a network administrator to specify what type of traffic is "special" or important and to queue the traffic based on that information instead of when a packet arrives.

7.1.3 Packet Classification

To achieve your intended packet delivery, you must know how to properly weight WFQ. To achieve the amount of QoS you require ,there are different weighting techniques and ways you can use them in various networks .

7.1.4 Traffic Policing

The previous sections covered ways you can queue different flows of traffic and then prioritize those flows. That is an important part of QoS. Sometimes, however, it is necessary to actually regulate or limit the amount of traffic an application is allowed to send across various interfaces or networks.

Vendors of VoIP equipment have a few tools that enable network administrators to define how much bandwidth an application or even a user can use. These features come in two different flavors: rate-limiting tools such as CAR, and shaping tools such as GTS or FRTS.

The main difference between these two traffic-regulation tools is that rate-limiting tools drop traffic based upon policing, and shaping tools generally buffer the excess traffic while waiting for the next open interval to transmit the data.

CAR and traffic shaping tools are similar in that they both identify when traffic exceeds the thresholds set by the network administrator.

Often, these two tools are used together. Traffic shaping is used at the edge of the network (customer premises) to make sure the customer is utilizing the bandwidth for business needs.

CAR is often used in service provider networks to ensure that a subscriber does not exceed the amount of bandwidth set by contract with the service provider.

7.1.5 Traffic Shaping and Queuing

Traffic shaping smoothes traffic by storing traffic above the configured rate in a queue. When a packet arrives at the interface for transmission, the following happens:

- If the queue is empty, the traffic shaper processes the arriving packet.
  If possible, the traffic shaper sends the packet.
  Otherwise, it places the packet in the queue.
- If packets are in the queue, the traffic shaper sends another new packet in the queue.

When packets are in the queue, the traffic shaper removes the number of packets it can transmit from the queue every time interval.

7.1.6 Fragmentation

Both propagation and queuing delay are discussed in previous chapters. The reasoning behind the need for fragmentation is simple. Large packets (1500-byte MTUs) take a long time to move across low-bandwidth links (768 kbps and less). Fragmentation breaks larger packets into smaller packets. You can accomplish this at either layer 2 or layer 3 of the Open Systems Interconnection (OSI) reference model.

In many data applications, latency caused by low-bandwidth links does not matter to the end user. In real-time applications, however, this can cause many problems (choppy voice quality, missed frames, dropped calls, and so on).

A 1500-byte packet moving across a 56-kbps circuit, for example, takes 214 ms to traverse the circuit. The ITU-T recommendation for uni-directional maximum voice latency is less than 150 ms. Therefore, *one* 56-kbps circuit and *one* 1500-byte packet consume the entire VoIP delay budget.

Fragmentation in itself is not enough to remove the latency problem on low-bandwidth circuits. The router must also be able to queue based upon fragments or smaller packets instead of by the original (prefragmented) packet.

Some vendors of VoIP equipment implementation enables users to modify the number of samples per packet. By default with G.729, two 10-ms speech samples are put into one frame. This gives you a packet every 20 ms. This means you need to be able to transmit a VoIP packet out of the router every 20 ms.

This 20-ms distance between each frame can change based upon the number of speech samples you decide to put in each frame. Also, this number is important because it enables you to determine the size of fragmentation needed. (Jacobson,1990)

As shown in Figure 7.3 , you can determine your fragment size depending on the speed of the link and the samples per frame.

Link Speed		Frame Size						
		1 Byte	64 Byte	128 Byte	256 Byte	512 Byte	1024 Byte	1500 Byte
	56 kbps	143 us	9 ms	18 ms	36 ms	72 ms	144 ms	214 ms
	64 kbps	125 us	8 ms	16 ms	32 ms	64 ms	128 ms	187 ms
	128 kbps	62.5 us	4 ms	8 ms	16 ms	32 ms	64 ms	93 ms
	256 kbps	31 us	2 ms	4 ms	8 ms	16 ms	32 ms	46 ms
	512 kbps	15.5 us	1 ms	2 ms	4 ms	8 ms	16 ms	23 ms
	768 kbps	10 us	640 us	1.28 ms	2.56 ms	5.12 ms	10.24 ms	15 ms
	1536 kbps	5 us	320 us	640 us	1.28 ms	2.56 ms	5.12 ms	7.5 ms

Figure 7. 3 Fixed-Frame Propagation Delay

7.1.7 Edge QoS Wrap-Up

At this point, you should understand the basics of packet classification, fragmentation, queuing, bandwidth, and policing mechanisms. It is important to note that you can use some or many of these mechanisms concurrently in a given environment.

As a rule of thumb, you always need to use packet classification and queuing mechanisms. Based upon bandwidth constraints and administrative policies, you might need to use compression and fragmentation methods as well.

7.2 Backbone Networks

The backbone of the network is completely different than the edge of the network, and you should not treat it with the same QoS mechanisms. Although the classification mechanisms for both might be the same or similar, the queuing, fragmentation, and bandwidth mechanisms are usually either not used or different.

7.2.1 High-Speed Transport

You can define high-speed transport as any interface higher than T1/E1 speed. Although most people refer to digital signal level 3 (DS-3) interfaces and above as high-speed networks, just as with microprocessors what is considered high speed today will be outdated in the future.

It is necessary to focus on different QoS mechanisms with high-speed transports. It is usually not feasible to apply all the same rules and policies on a high-speed interface as you would on a lower-speed interface. This is mainly because the more policies and QoS mechanisms you apply, the longer the router must take to forward a packet. Although this is usually okay on lower-speed interfaces, higher-speed interfaces cannot spend as much time identifying and queuing each packet.

As customers move to higher-speed interfaces (such as STM -1/4/16), however, it is important to provide them with options so that they can determine how to properly enable QoS on their networks.

7.2.1.1 POS

With the capability for IP to travel directly on a Synchronous Optical Network (SONET/SDH) infrastructure came the need to prioritize traffic on this high-speed interface. Before putting IP directly on SONET/SDH , people used ATM as the transport mechanism. Although you can use ATM QoS tools to provide the necessary priority schemes, new IP QoS mechanisms had to be developed, however, to handle QoS on interfaces up to OC-48/STM-1.

7.2.1.2 IP and ATM

Many networks today use ATM as the Layer 2 transport, with IP, TDM, and other traffic traversing a single ATM network. Because of this, it is necessary to map IP QoS onto an ATM network.

You can map IP prioritization onto ATM in two ways. The first method simply maps the precedence values on an IP packet to different ATM PVCs. This enables the network administrator to have different PVCs, allocating important traffic over a variable bit rate (VBR) ATM circuit and less important traffic over an unspecified bit rate (UBR) ATM circuit.

You also can have PVCs of varying speeds and purposely send data down a congested PVC while sending voice and other real-time traffic down a less-congested PVC.

The second method maps IP prioritization onto ATM using queuing techniques such as WFQ to prioritize between different flows per PVC.

7.2.2 Congestion Avoidance

As discussed previously in edge , mechanisms manage existing congestion and prioritize the traffic that is of highest importance.

Congestion avoidance in backbone networks, works on a similar problem from a completely different angle. Instead of managing the existing congestion, congestion avoidance works to avoid congestion to begin with.

In simplistic terms, you avoid congestion by dropping packets from different flows, which causes applications to slow the amount of traffic being sent. This avoids what is known as global synchronization, which occurs when many IP TCP flows begin transmitting and stop transmitting at the same time. This is caused by the lack of QoS in a service provider's backbone.

7.2.3 Backbone QoS Wrap-Up

It is important to note that both edge QoS and backbone QoS must work together to achieve the proper QoS for the various applications which might be traversing a network.

As a rule of thumb, it is wise to use high-speed congestion avoidance techniques in the backbone as well as some form of high-speed transmission, such as POS/Synchronous Digital Hierarchy (SDH) or IP + ATM inter-working. You can achieve IP QoS using several different mechanisms. The actual transport mechanism you choose is not as important as verifying that all the tools you need are present to service your applications.

CHAPTER EIGHT

VOICE OVER IP CONFIGURATION ISSUES

8 Voice over IP Configuration Issues

Signaling, quality of service (QoS), and architectural issues—all of which are covered in previous chapters—are the basic fundamentals of Voice over IP (VoIP) deployment. Now it is time to discuss the configuration considerations that accompany VoIP.

8.1 Dial-Plan Considerations

A dial plan is the method by which you assign individual or blocks of telephone numbers (E.164 addresses) to physical lines or circuits. In Public Switched Telephone Networks (PSTNs), you create a dial plan by partitioning blocks of numbers in a hierarchical manner (10,000 numbers is normal for the PSTN).

To create a dial plan for an enterprise voice network, you also assign individual telephone numbers to individual users. Even in private enterprise voice networks, it is common to adopt hierarchical assignments when creating a dial plan. In such networks, however, dial-plan problems usually surface.

The PSTN uses a specific hierarchy. The International Telecommunication Union Telecommunication Standardization Sector (ITU-T) recommendation E.164 sanctions intercountry calling specifics. The North American Numbering Plan (NANP), for example, builds on the ITU-T recommendation and further specifies how many digits you can use and what you can use them for. Therefore, although dial plans in the PSTN are not simple, they are at least hierarchical. In this way, network service providers can build hierarchical dial plans.(Murhammer et al,1999)

8.1.1 Dial-Plan Problems

Administrators of private enterprise voice networks usually run into dial-plan problems when they launch a companywide dial plan that encompasses multiple remote sites.

In this topology, enterprise voice network administrators must decide on a usable number of digits in the dial plan and ensure that numbers do not overlap. For the network administrator, feature transparency — or the capability of using the same functions across multiple Private Branch eXchanges (PBXs) and throughout all locations—is one of the major requirement.

To illustrate this concept, assume that a company has a headquarters location with a five-digit dial plan. This company has 20 remote sales sites and wants to incorporate a dial plan that enables all its sites to have a cohesive look and feel.

This company must proceed with the following steps to ensure that its dial plan is consistent:

1. Company must analyze corporate and remote sites to ensure that no overlapping digits with a five-digit plan exist. If overlapping digits do exist, it must decide whether to move to a six- or seven-digit plan or to give each site its own two-digit code.
2. Company must transition the PBX dial plan to translate the new five- or six-digit extensions into valid E.164 addresses so that the call can still traverse the PSTN.
3. After this initial change is made and the users begin to use the new dial plan, Company must transition to VoIP by simply modifying the Automatic Route Selection (ARS) table.

In some circumstances, enterprises with several remote branches must allocate a large number of digits to meet the needs of all possible users. Such enterprises simplify their dial plan by implementing a two-stage dialing procedure.

With two-stage dialing, the caller can dial an access code (similar to using a calling card) that routes to a specific place in the network. The caller is then presented with a second dial tone, at which point he or she can dial the actual number to be called. Two-stage dialing offers two main advantages: the remote PBX's dial plan can be simple, and the network does not need to have a dial-plan outlining the entire network's dial plan. Instead, the network uses a group of access codes, which map to remote switching points.

The limitations of such an approach are that users must follow a two-step procedure, and they must wait for the network to properly prompt them for additional inputs. Despite these limitations, however, private enterprise networks implement two-stage dialing for three main reasons: if they experienced rapid growth, if they merged with another corporation, or if they acquired another corporation that uses another type of PBX technology.

Vendors of VoIP equipment , implemented programs which enables both single- and two-stage dial plans. Using a single-stage dial plan (also known as a translational plan) generally requires that users not change their dialing habits. If a company did not have a dial-plan architecture in the past, imposing a VoIP architecture can introduce some challenges, such as number-overlapping and a lack of call-routing features.

These problems are not necessarily due to VoIP, but they are exacerbated by the fact that no one at CompanyBlue put together a cohesive dial plan in the past that would sustain the company if its branches and main offices were on one central dial plan.

VoIP supports two-stage dialing, but when you use this plan you must be careful for the following reasons:

- You can lose Dual-Tone Multi-Frequency (DTMF) tones as they traverse the Internet Protocol (IP) network if you use inappropriate encodings. Coding a voice stream—carried over a Real-Time Transport Protocol (RTP) and through alternative methods—within a signaling path (such as H.245) enables the transport of DTMF inputs on the network.
- Tandem encodings (dual compressions), which reduce call quality, can now occur due to poor network planning.
- Multiple digital-to-analog (D/A) conversions can occur, which also reduce call quality.

Packet loss is common when using an IP network. If the DTMF tone is carried in a User Datagram Protocol (UDP) stream, however, the packet or tone can be lost or improperly ordered, which causes the wrong sequence of digits to be dialed.

If the VoIP provider uses DTMF relay, which enables DTMF tones to be carried in the Transmission Control Protocol (TCP), the DTMF carriage is just as reliable as the PSTN.

If no single entity controls the voice and data network, however, it is possible in single- and two-stage dialing to have multiple compression cycles, which affect voice quality. You must take great care to make sure that tandem encodings do not occur, as you cannot improve such encodings.

Multiple D/A conversions also can affect voice quality. Where D/A conversions really rear their head, however, is when modems or some other data transmission over voice is handled. When using a 56-kbps modem, for example, you can have only one D/A conversion (at residential facilities). (Murhammer et al,1999)

8.2 Feature Transparency

Switching from time-division multiplexing (TDM) voice networks to packet-based voice solutions also requires that you move and support existing applications and functionality in a similar manner.

Often, many PBXs have proprietary signaling methods that currently have no way to move onto IP. This makes it difficult to have a cost-effective VoIP network that offers limited features. The reason this makes VoIP networks difficult is because of the proprietary nature of inter-PBX signaling protocols. Often, these proprietary signaling methods cannot be carried across a VoIP network.

In an attempt to provide some interoperability between PBXs and vendors using digital signaling, the ETSI Q.Sig standard was developed. This standard is based on the Q.931 signaling stack, but it contains extensions that enable additional signaling information to be passed between the PBXs.

Q.Sig is a standards-based protocol that enables different brands of PBXs, as well as different networks, to interoperate. vendors make Q.Sig available on their VoIP gateways and can complete Q.sig calls, as well as make a Q.Sig tunnel between multiple PBXs.

This enables enterprise customers to do the following:

- Achieve a feature-rich, cohesive telephony network
- Integrate different vendors' PBXs throughout their network

Enterprise customers who are either unwilling to or cannot upgrade to Q.Sig can have a telephony network with only basic voice calls. Often, the cost savings and ability to use new IP-enabled applications are enough to encourage enterprise telephony customers to move to this new network. (Murhammer et al,1999)

8.2.1 PSTN Feature Transparency

Features in the PSTN are based mainly on Signaling System 7 (SS7) and the applications built on top of it. To transparently transport and tunnel features across multiple networks, SS7 must be supported as the mandatory baseline interface.

The H.323 protocol suite was developed assuming Q.931 (Integrated Services Digital Network [ISDN]) interfaces on the voice gateways. The protocol suite has no transparent mechanism to carry and tunnel SS7 messages, including Intelligent Network (IN)-based protocols.

CHAPTER NINE

MARKET FORECASTS AND FUTURE

9 Market Forecasts and Future

The concept of Voice over IP has been in focus for the last couple of years , growing to be a major technology trend during 1997-2000. Voice over IP is the concept of being able to use one single network, the TCP/ IP network, as the common carrier for not only data, but also voice, and eventually any other communication form.

At the moment, the level of interest in Voice over IP is almost entirely driven by expectations regarding its potential. Much of the supporting evidence in this chapter is sourced from the major market analysis companies.(Ericsson Training Book,2000)

9.1 Global Voice / Data Crossover?

The increasing proliferation of data communication in the last two decades is unveiling a new paradigm in communications.

In general terms, the demand for voice services has been increasing at 5% to 10% per year while that of data services has been expanding at 75% to 300% annually. Figure 9.1 provides a graphical representation of this behaviour. (Ericsson Training Book,2000)

A less certain but much more interesting consideration involves the timing of this transformation.

The data presented on the graph suggests that this transformation occured sometime between 2002-2003,and until 2008 data traffic will be at least 5 times of voice traffic.

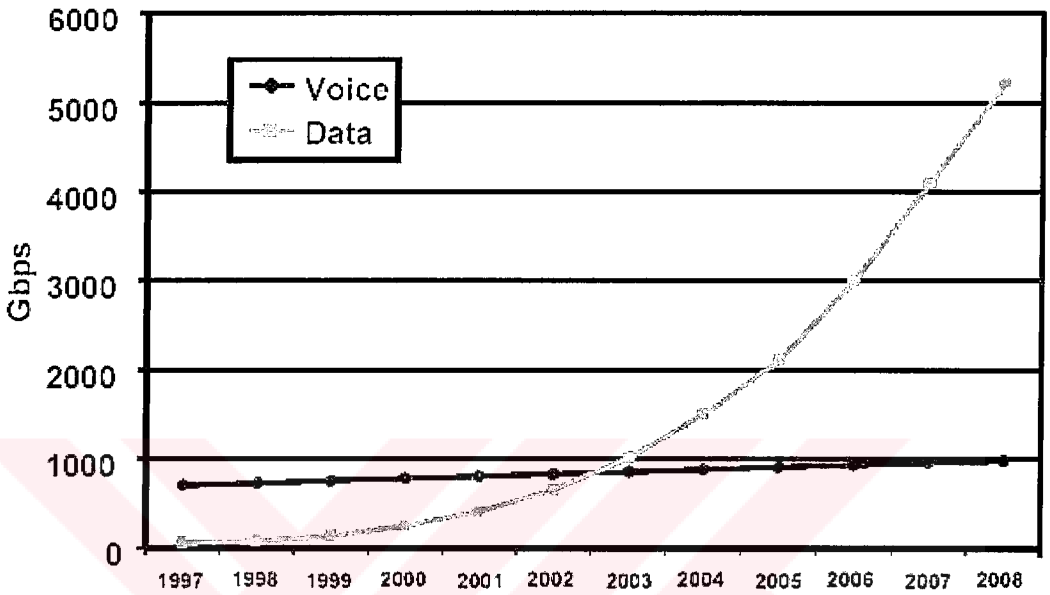

Figure 9. 1 Global Voice/Data Crossover

The revenues forecast for Voice over IP are looking towards an exponential growth in the next years, compared to the "traditional" growth of PSTN with a couple of percent each year (i.e. riding the wave of data traffic growth).

However, the current Voice over IP market size does not correspond to the hype around Voice over IP today, but possibly this hype will be justified by the future market size.

9.2 IP Telephony Voice Projections

As discussed on the previous section , sometime between 2002-2003 data traffic surpassed voice traffic and it is envisaged that Voice over IP will grow at a similar exponential rate as data.

It is predicted that much of the increased traffic carried by the future IP networks will be in the form of advanced telephony applications that have direct ties to Intranets and the Internet - where either a service, like a directory or locator service - or a Voice over IP client or IP telephone is an integral component of the network, that allows that communication to take place.

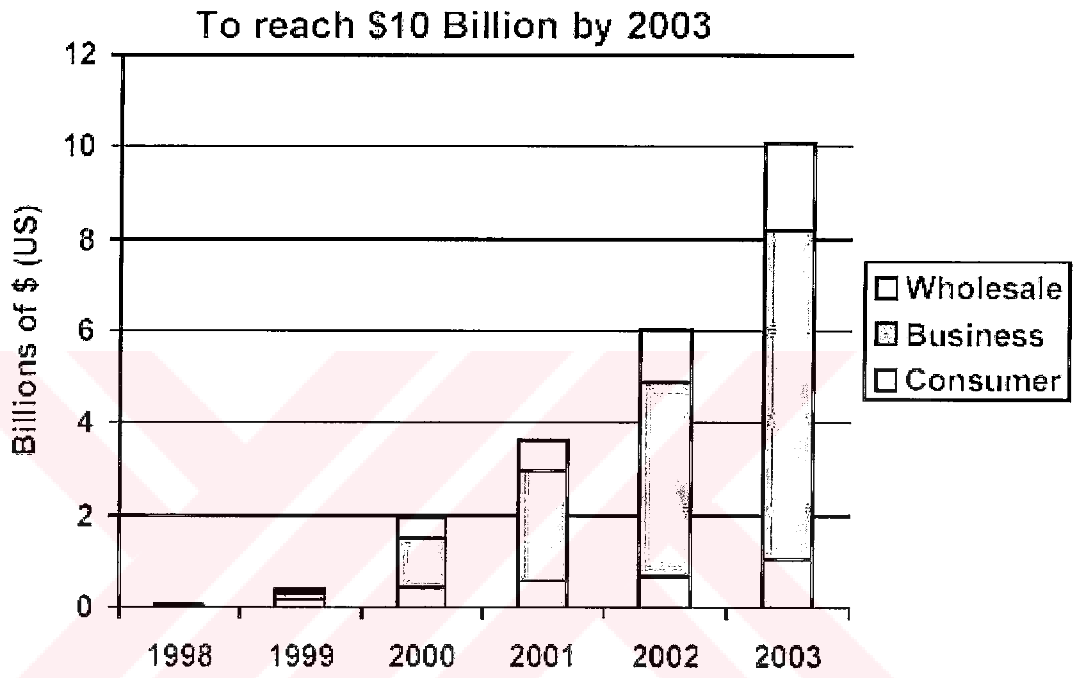

Figure 9. 2 VoIP Worldwide Explosive Growth

The graph shows the predicted explosive growth of VoIP world wide, which is predicted to to reach $US 10 Billion by the year 2003. (Ericsson Training Book,2000)

9.3 What's driving the growth?

We have discussed the exponential growth of the IP Telephony market. Now let us examine what factor are driving this growth.

Costs today are artificially inflated for End-users. Packet voice gets around the regulatory issues (in some countries) because it is currently considered an enhanced data service, and therefore is not subject to regulatory charges, which leads to lower costs.

9.3.1 End-Users

Users today are often involved in multiple communications methods at the same time, for instance surfing the web at the same time we are talking on the phone. This is the benefit behind multimedia and this is also a driving force behind VoIP.

9.3.2 Carriers

Carriers can decrease high capital start-up costs with IP voice solutions. It is no longer necessary to invest in big iron circuit switching gear to be able to call yourself a full featured telephony service provider. Carriers typically are stuck with managing a voice network, a separate private line network, and a separate data network. This involves provisioning three systems, and also monitoring and planning the the systems.

9.3.3 Technology

Just as End-users benefit from new, integrated services made possible by IP voice, carriers benefit by having differentiating services which can be used to attract, retain customers, and also bring higher margins.

Compression allows a carrier to reduce voice calls from 64 kbps down to 8 kbps, resulting in significant savings in bandwidth. Carrier grade data networking equipment is being voice enabled and carrier grade voice switching platforms can now accommodate a native packet network interface to ship voice and data over a common data pipe. Prioritization schemes are improving which will allow voice packets to be identified on an IP network, and receive priority over other types data traffic.

9.4 Technology Trends

On the technology side there is always a lot happening. This section briefly summaries some of the technology trends that are assisting the progress of IP Telephony products.

9.4.1 Specialized hardware solutions

More specialized hardware solutions replacing software solutions (on general purpose computers). More DSP-power is included in various hardware. Routers with voice over IP as add-on card upgrades and combined router and VoIP gateway.

9.4.2 New hardware in PC's

Compact PCI bus gaining popularity offering hot swap as an intrinsic feature.Carrier grade reliable, scalable solutions are beginning to appear offering thousands of ports per system. System footprint is reduced as systems become rack mounted instead of based on general purpose servers.

9.4.3 Better codecs

Voice and video codecs are continuously improved on as is various solutions aimed at lowering latency through improved buffer management. Interoperability Interoperability is the key success factor, both for vendors and for operators. Standards are not enough and both client interoperability and other components are required to interoperate in order to realize the multi-vendor Voice over IP network.

9.4.4 IP Access

On the issue of how the IP access is to evolve the next coming years the prospect is a concept of "Always on - everywhere" but with limitations on bandwidth.

PBX Trends for the enterprise user, the well-known Private Branch Exchange (PBX) is seeing replacement products being offered in the form of IP-based PBXs implemented on standard server(s) and as a feature of standard routers. The IP-based solution also brings promises of doing more than what the PBX could ever do.

Initial IP Telephony solution offerings are "forklift upgrade" replacement concepts, whereas new solutions enabling a migration path is required and emerging. For example, Lucent developing a LAN phone which works with the existing PBX.

9.5 Market Trends

This section summaries some of the market trends influencing the IP Telephony market.

9.5.1 The Next Generation

Telcos , One of the most exiting aspects of the Voice over IP industry is that it has given birth to a new breed of telecom operators, the "Next Generation Telcos" (NGT), the IP- based telephony operator. The NGT is seen seeking co-operation with alternate operators (utility distributors for example).

9.5.2 A place for large players

Since 1998, Voice over IP is also a place for the large players who can not afford to ignore the phenomenon, and in some cases explore it further. Examples are Microsoft (client, API's, etc.) , Cisco with the Architecture for Voice Video and

Integrated Data (AVVID) concept comprising gateways, gatekeepers terminals in a coordinated concept.

9.5.3 Mergers and acquisitions

Mergers and acquisitions are commonplace within the industry and will most likely continue. Some of the major activity included: Nortel + Bay Networks, Ericsson + ACC, Cisco + Selsius, Lucent + Ascend, Alcatel + Newbridge, Intel + Dialogic, Nokia + Vienna Systems.

Although it's important to note this activity is not entirely driven by IP Telephony. (Ericsson Training Book,2000)

9.5.4 Mobile enablers

There are enablers on the way to give us "Mobile IP", with ever increasing bandwidth, in which voice will eventually be a small stream in the data flow.Wireless technologies such as GPRS, EDGE, UMTS promise to ultimately deliver megabit speeds to the mobile terminal - "always on" - within a couple of years.

On the application side, there are big hopes being put on WAP and SIM Toolkit in the term filling the gap between now and Third generation mobile systems.

9.5.5 IP Telephony as a Catalyst

IP Telephony is in many ways a catalyst in the convergence of wireline, wireless and data networks and eventually also for the convergence of TV, broadcast radio and video/movie distribution into the one multiservice network.

9.6 Expectations and Requirements

As a technology develops, the expectations (and demands) from consumers increases.

Early IP Telephones (IPT) , when IP Telephony was first introduced, the speech quality was in focus, and in many ways it still is. The comparison to wireless telephony has probably assisted consumers to gain an understanding that there is a price/performance/ features mix to telephony where sometimes some advantages outweigh limited functionality and quality. For instance wireless telephony provides mobility, but with a lesser voice quality than the traditional PSTN and also generally more expensive than the PSTN.

Current IPT At the moment, IP-based voice and eventually video is being discussed and considered for large scale deployment. This drives consumers to make many comparisons between circuit switched telephony and IP Telephony, some examples of expectations include:

- 99.99% uptime
- PSTN speech quality (or better!)
- Ease of use
- Valuable services
- Fair and reasonable pricing (or even cheaper than PSTN).

In some sense many of these expectations are what the IP Telephony market is all about and in other cases it may be up to the IP Telephony market players to re-educate consumers regarding their expectations. Certainly the goal of most traditional telecommunication vendors as the new telecommunications market evolves is to bring datacom networks up to the reliability of todays telephone networks.

CHAPTER TEN
APPLICATION OF VOIP

10 Application of VoIP

When we begin thinking about consolidating Dokuz Eylül University's (D.E.U) voice and data networks into a single multiservice network, the initial application we considered is toll-bypass.

Toll-bypass enables businesses to send their intra-office voice and fax calls over their existing TCP/IP network. By moving this traffic off the Public Switched Telephone Network (PSTN), D.E.U can immediately save on local and long-distance charges by using extra bandwidth on their data network without losing existing functionality.

In fact, some businesses like D.E.U with plenty of intra-office calling, both domestic and international, is thought to have a Return On Investment (ROI) in as little as three to six months.

10.1 Consolidating the Networks

To consolidate D.E.U's network , in an effort to reduce costs and maintain functionality ; we laid out the following points for our discussions :

- It wanted to work with existing PBX's from different vendors to provide an end-to-end solution
- It wanted to provide technology to its internal and external customers, but it didn't want to be on the bleeding edge in terms of technology risk.

- It wanted to avoid forklift-upgrades to its existing infrastructure. A forklift-upgrade occurs when most or all of a company's existing network hardware and software needs to be replaced with newer hardware and software. This is not only expensive in terms of capital expenditure, but it also involves physical visits to each site and disruptions of the existing network functionality.
- It wanted the new multiservice network to be cost-effective and expense-reducing.

We determined that Cisco Systems' current multiservice offerings could provide an end-to-end solution that would meet all its stated needs.

10.2 Configuration Issues

We decided not to change the current network topology but to give it a flexibility to be used with VoIP. The current network is shown in figure 10.1:

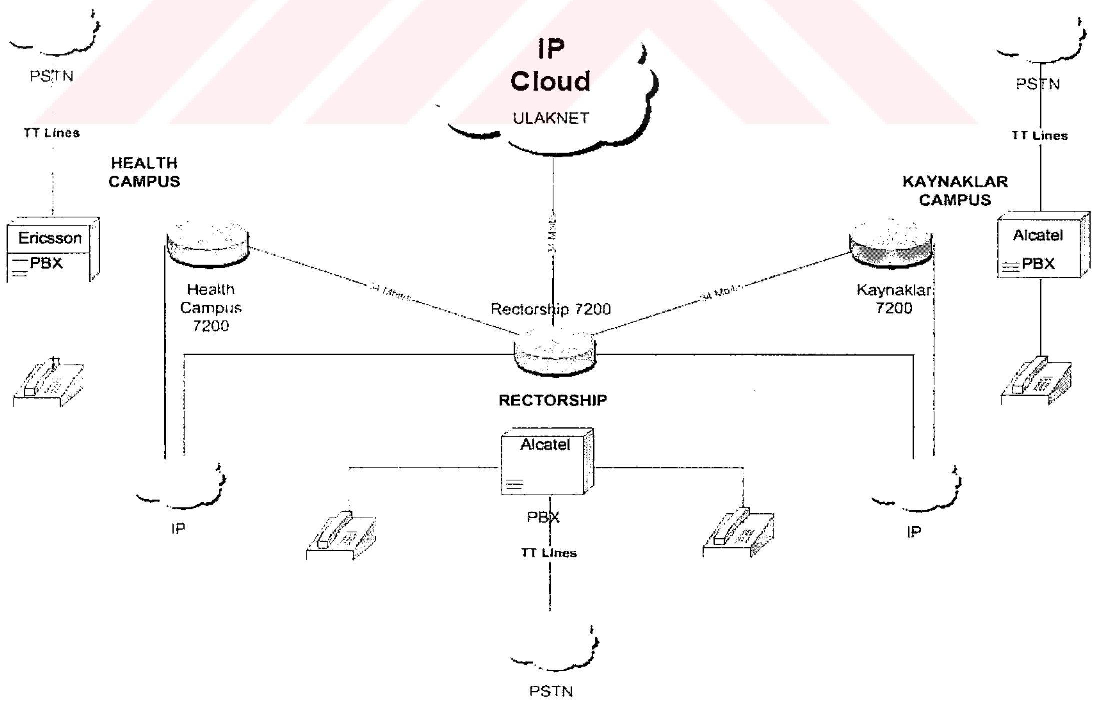

Figure 10. 1 D.E.U's Current Network

In the current network;

- Three Cisco 7200 Routers that will be used to carry voice and data traffic
- In each campus , PBX from different vendors is used for voice traffic
- Routers are connected via 34 Mbit/s leased lines
- In the rectorship building ,the connection to "Ulaknet" is made with a 34 Mbit/s leased line

As shown in the figure , there are two networks exist between the campuses. The first one is the PSTN traffic network , in other words voice network and and the second one is the data network.

To integrate these networks we have to combine Routers and PBXs.

10.2.1 Routers Compatibility

In order to use the routers as VoIP gateways , we found the PA-VXA/VXB/VXC voice port adaptors for the Cisco 7200/7400/7500 router platforms.

These voice adapters are able to combine T1/E1 connectivity and onboard digital signal processor (DSP) resources to provide unparalleled flexibility and power in directly supporting voice services on these gateways. These port adaptors are capable of supporting either T1 or E1 interfaces and depending on the selected model can support up to 60 simultaneous High Complexity (HC) or 120 simultaneous Medium Complexity (MC) codec algorithm voice calls.

In addition,and depending on the selected model, it is also possible to use the onboard DSPs as a DSP farm to provide voice services to port adaptors such as the PA-MCX-nTE1 series of products which can support voice telephony interfaces but have no direct DSP resources of their own.

10.2.2 PBX Integration

Since there are PBX s in each campus , we need to integrate PBXs' to the VoIP network.

Many private branch exchanges (PBXs) use E1 trunks running CAS as the main interface to the publicswitched telephone network (PSTN), and to connect to external peripherals such as voice-mail or interactive voice response (IVR) systems.

Since we have a a voice-capable Cisco router equipped with the E1 Drop and Insert (D&I) Voice port adapter ; selected time slots on one port of a router to be transparently connected to selected time slots on another port of another router.

A E1 trunk consists of 31 individual 64 Kb channels multiplexed together. E1 frame structure allows samples of each time slot to be sent in a repeating pattern. The timing (clocking) on a E1 trunk is embedded in the bit stream with the timing referenced to a central clock source (generally the telco). Because the clocking between E1s is synchronized, it is possible to take (drop) the bits that represent particular time slots on one E1 and insert them into other time slot positions on a different E1.

10.3 Proposed Configuration

After we have choosen the necessary equipment to be used for VoIP networks, we made the proposed configuration as in the Figure 10.2.

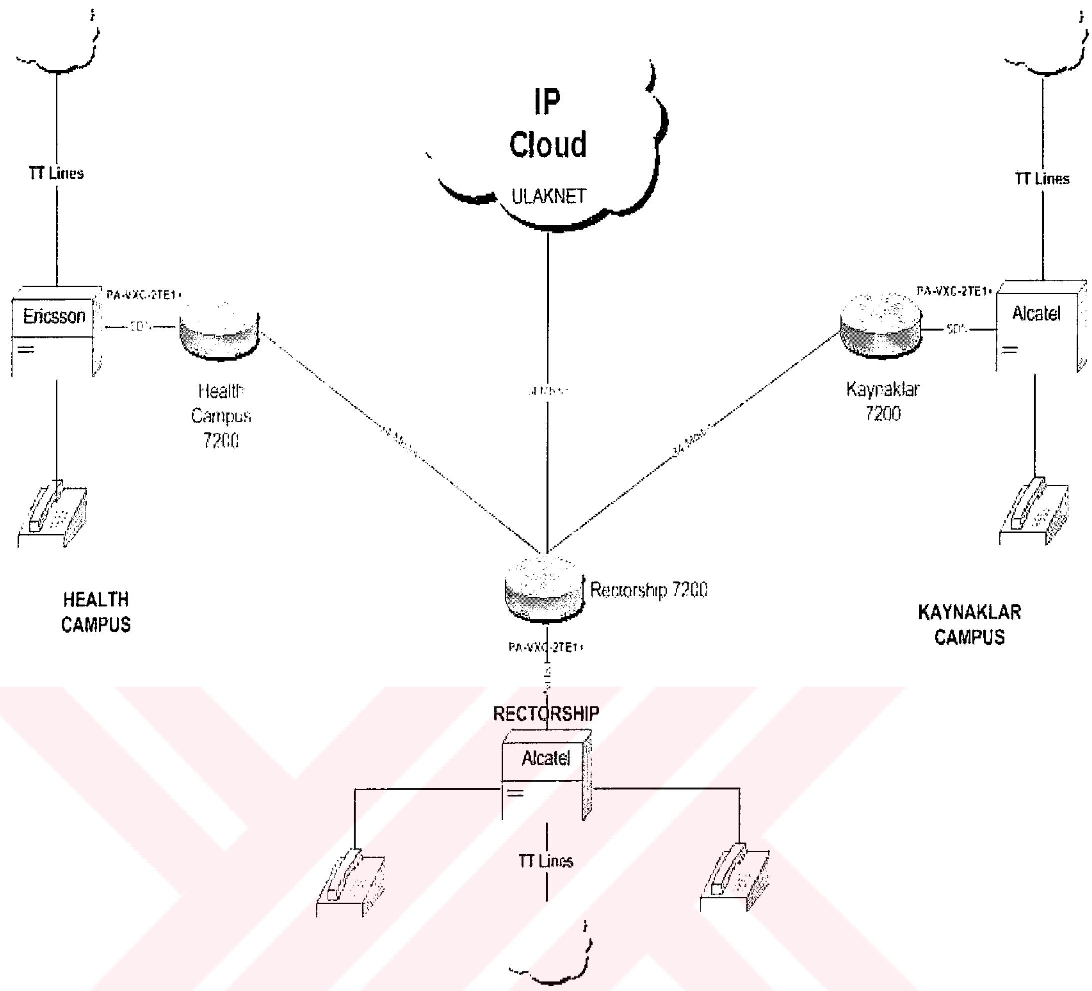

Figure 10. 2 D.E.U's Proposed Network

The highlights of the solution include the following:

- Leveraged D.E.U's existing data network, which was made up of 7200 series routers. 7200 series are modular routers/VoIP gateways. They provide more than 120 LAN and WAN interfaces, from async to optical carrier 3 (OC-3) ATM, as well as analog and digital voice interfaces such as T1/E1, Foreign Exchange Station (FXS), FXO, and recEive and transMit (E&M). Both routers share the same network modules, so stocking, sparing, and consistency across the family of products is maintained.
- Based on open-standard, H.323 protocols.

- Provided an integration path that utilized D.E.U's existing data network and PBX equipment.
- Required neither extensive reconfiguration of existing data and voice equipment, nor a forklift-upgrade of any equipment.
- Enabled the D.E.U IT group to continue utilizing the expertise of both its voice and data support staff.
- Interoperable with other multiservice technologies that Cisco Systems offers, including Cisco IP phones and IP PBXs, as well as H.323-based applications, such as Click-2-Dial and Netmeeting.

After modifying PBXs routing tables and making propriate dialing plan, any employee can call another office by using VoIP.

The most important outcome expected to be immediate cost savings by moving D.E.U's intraoffice voice and fax calls onto its TCP/IP data network. D.E.U will eliminate the leased lines as well as reduced its long-distance charges associated with those calls.

The other outcomes to be , if needed , D.E.U is capable of replacing small-office key-systems with Cisco IP phones and to reduce lease costs when the key-system leases expired. Capability to integrate both Cisco IP phones and existing voice equipment with multiservice applications such as Netmeeting or Intel ProShare video-conferencing using H.323.

CONCLUSIONS

The increasing proliferation of data communication in the last two decades is unveiling a new paradigm in communications. While telecommunications technology has largely been associated with telephony and the Public Switched Telephone Network (PSTN) for the better part of this century, an examination of recent developments and trends suggests that we are in the process of a rapidshift from a voice-centric to a data-centric communication network.

In general terms, the demand for voice services has been increasing at 5% to 10% per year while that of data services has been expanding at 75% to 300% annually.. Granted that these two rates are applied to different base values, it is not surprising that, should this pattern continue, aggregate data traffic will surpass aggregate voice traffic.

A less certain but much more interesting consideration involves the timing of this transformation. It is suggested that this transformation occured sometime between 2002-2003, and until 2008 data traffic will be at least 5 times of voice traffic.

This thesis has tried to lighten this newest method of communication with VoIP which has the highest level of interest entirely driven by expectations regarding its potential.

If it had given the chance to invest in VoIP equipment for D.E.U network as it is depicted in the last chapter, D.E.U can immediately save on local and long-distance charges by using extra bandwidth on their data network without losing existing functionality.

REFERENCES

Adolfo Rodriguez& John Gatrell& John Karas& Roland Peschke ,TCP/IP Tutorial and Technical Overview,IBM,2001

Andrew S. Tanenbaum , Computer Networks,Second Edition, Prentice- Hall PTR, 1999

Ericsson Training Book , Voice Over IP , Telefonaktiebolaget LM Ericsson , 2000

Mani Subramanian , Network Management Principles and Practice, 2000

Martin W. Murhammer&Kok-Keong Lee&Payam Motallebi&Paolo Borghi, IP Network Design Guide , IBM , 1999

S. Keshav ,An Engineering Approach to Computer Networking, , Reading, MA: Addison-Wesley, 1997.

S. Shenker& C. Partridge& R. Guerin , Specification of Guaranteed Quality of Service, RFC 2212.

Uyless Black, Voice over IP , Prentice-Hall PTR,1999

V. Jacobson , Compressing TCP/IP Headers for Low-Speed Serial Links , RFC 1144, 1990.