

Realtime  
publishers

"Leading the Conversation"

*The Definitive Guide<sup>™</sup> To*

# Converged Network Management



*Ken Camp*

Chapter 2: Key Considerations in Effective Voice and Data Integration for a Changing IT/IP Landscape .....	20
Quantifiable Business Processes .....	20
Web-Centric Businesses .....	20
Call Centers.....	22
IVR Systems .....	24
CTI .....	24
CTI History .....	24
Application Integration with CRM and Enterprise Resource Planning Systems .....	27
Customer Relations and CRM .....	28
Delivering Call Quality with VoIP .....	30
Traditional Voice Characteristics—the PSTN.....	30
IP Traffic Characteristics .....	31
Design Considerations and Class of Service .....	33
QoS .....	36
QoS Approaches in IP Networks .....	37
QoS—the Signaled Approach.....	40
QoS—the Provisioned Approach.....	43
QoS—the Bypass/Shim Approach.....	45
Summary .....	49

## Copyright Statement

© 2006 Realtimepublishers.com, Inc. All rights reserved. This site contains materials that have been created, developed, or commissioned by, and published with the permission of, Realtimepublishers.com, Inc. (the “Materials”) and this site and any such Materials are protected by international copyright and trademark laws.

THE MATERIALS ARE PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. The Materials are subject to change without notice and do not represent a commitment on the part of Realtimepublishers.com, Inc or its web site sponsors. In no event shall Realtimepublishers.com, Inc. or its web site sponsors be held liable for technical or editorial errors or omissions contained in the Materials, including without limitation, for any direct, indirect, incidental, special, exemplary or consequential damages whatsoever resulting from the use of any information contained in the Materials.

The Materials (including but not limited to the text, images, audio, and/or video) may not be copied, reproduced, republished, uploaded, posted, transmitted, or distributed in any way, in whole or in part, except that one copy may be downloaded for your personal, non-commercial use on a single computer. In connection with such use, you may not modify or obscure any copyright or other proprietary notice.

The Materials may contain trademarks, services marks and logos that are the property of third parties. You are not permitted to use these trademarks, services marks or logos without prior written consent of such third parties.

Realtimepublishers.com and the Realtimepublishers logo are registered in the US Patent & Trademark Office. All other product or service names are the property of their respective owners.

If you have any questions about these terms, or if you would like information about licensing materials from Realtimepublishers.com, please contact us via e-mail at [info@realtimepublishers.com](mailto:info@realtimepublishers.com).

[**Editor's Note:** This eBook was downloaded from Realtime Nexus—The Digital Library. All leading technology guides from Realtimepublishers can be found at <http://nexus.realtimepublishers.com>.]

## Chapter 2: Key Considerations in Effective Voice and Data Integration for a Changing IT/IP Landscape

The integration or convergence of voice, video, and data can provide a business with a competitive edge when effectively implemented. There are several business models and operating environments that present opportunities for strategic consideration when planning for this change. The key factors for success involve leveraging the integrated features to provide the greatest support for existing processes.

This chapter will look first at high-level business models and processes that are often impacted by service convergence. Later, the chapter will delve into the issues of call quality that affect every organization implementing an integrated service solution. Quality is often the single biggest factor in a successful implementation, so this chapter will explore a variety of approaches for delivering total quality network services, with a focus on integrated VoIP.

### Quantifiable Business Processes

Business crosses a wide array of sectors, each having unique business requirements to support aspects of the core business. Call centers may play a central role for many businesses, particularly those in financial services, insurance, or travel. They also play a key role in many other sectors as smaller customer support teams. Interactive voice response (IVR) systems are frequently automated to reduce the requirement for staffing and provide information to customers. Computer telephony integration (CTI) isn't a new concept with the deployment of VoIP, but in many cases, it becomes easier. CTI may provide levels of service and application integration previously outside the financial grasp of some organizations.

### Web-Centric Businesses

In addition to the traditional sectors of business, the integration of the Internet has heightened awareness of four distinct business models in the Web-centric world of e-business. Some of these models fit with large enterprise; others are more amenable to small business and have been used by many an e-business startup company.

In the open market model, anyone can be a buyer, and anyone can be a seller. There's no centralized control, and minimal trust involved. There isn't particularly high value to integration of enterprise systems because relationships may be ephemeral. Market leadership for these e-businesses requires being in the right place at the right time, with the right solution at the right price. OASIS and eBay are good examples of the open market business model.

The alliance model is more common in larger businesses. It embraces a distributed corporate environment with multiple leaders of the pack. The goal of these alliances is frequently optimization of specific solutions to solve identified customer problems. Alliances are often formed among the best and brightest in their respective fields. High levels of integration in services and applications between partners bring tremendous value to an alliance. Sun, IBM, Oracle, and Netscape demonstrate this model via the Java Alliance.

The aggregation model is typically adopted by the leader in a business sector. The aggregator positions itself between producers and consumers, providing access to products. Integration with consumers may be low, but integration with the producers and internally across the aggregator enterprise can add very high value. Wal-Mart represents a perfect example of this model.

The value chain model is adopted by most businesses. Every business is, in some facet, the leader of the pack in its' particular sector. Process optimization within the enterprise is crucial to business success. The leader focuses on optimizing the value chain through service and application integration rather than aggregating buyers and sellers. Cisco Systems, Dell, and Amazon represent value chain leaders in the world of e-business.

There is also an emerging competitive model, driven by competition between clusters, not individual companies. This model is often a matter of survival for businesses that need partnerships to deliver a complete solution set to market. In this model, companies maximize the user of distributors, resellers, and retail sales chains. Integration of services and applications may focus on facilitative technologies such as Web-based Electronic Data Interchange (EDI) to drive e-commerce online.

Business in the new economy of the Internet requires instantaneous reaction to customer behavior. Being nimble and responsive can provide a competitive edge in having the right solution for customers at the right time.



Examples of near-instantaneous change include Wal-Mart and K-Mart. Wal-Mart monitors computer inventories closely in stores. If a customer buys a package of tennis balls, the inventory and shipping systems are automatically updated to ship a replacement set to a specific store. If a store experiences a run on tennis balls, or an unusual trend exceeding normal thresholds for a product occurs, stores all across a geographic area may find they receive increased delivery of tennis balls to support the trend.

K-Mart takes a different approach, using outside data sources to drive product stocking. One often-used example is their use of the National Weather Service (NWS). If K-Mart notes inclement weather headed for a particular part of the country, their stores may receive umbrellas and raincoats for stocking, in support of anticipated demand.

## Call Centers

Businesses implement call centers to effectively administer incoming product support or provide information to customers. Call centers most often interact directly with consumers and are often the lifeline of the customer relationship. Outbound call centers are used for telemarketing, clientele management, and debt collection. The call center might also be a broader contact center, handling postal mail, fax communications, and email for an enterprise or business unit.

Call centers are generally built using large, open work spaces with workstations including computers, telephones with headsets, and supervisory stations. Some businesses build centralized call centers; others distribute call centers at diverse locations, connecting them with voice and data technology. Global companies may have call centers around the world, with each being the primary center as the time of day changes. Location, coupled with time, can provide the advantage of using resources during the local business day.

Call centers are linked to the corporate computer network, including mainframes, microcomputers, and LANs. Increasingly, the voice and data services are linked through CTI.

Most large enterprise businesses use call centers to interact with their customers. What has changed with the widespread deployment of VoIP is the barrier to entry. Now a midsized company can easily deploy a centralized or distributed call center, leveraging integrated technologies and bringing new services to customers that were previously out of reach.

Call center technology essentially provides a queuing network coupled with workforce planning and management to achieve desired service levels. A common example of call center service levels might require that at least 80 percent of the callers are answered within 20 seconds or no more than 3 percent of the customers hang up, due to their impatience, before being served. Call centers provide information about traffic patterns and business patterns as well. Statistics gathered can help determine whether a single large call center is more effective at providing customer service (answering calls) than several distributed, smaller ones might be.

Centralizing call management into a call center aims at improving business operations and reducing costs while providing a standardized, streamlined, uniform service for consumers. The efficiency of a repeatable process makes this approach ideal for large companies with extensive customer support needs.

Call centers use an array of voice and data technologies and provide a prime candidate for integrated services and applications. These technologies can ensure that customer service agents are kept as productive as possible and that calls are queued and processed as quickly as possible, producing the desired service levels for customers. Some of these technologies include:

- Automatic call distribution (ACD) groups
- Analysis tools to review agent activity
- Optimization analysis for outbound call centers, referred to as Best Time to Call (BTTC)
- Interactive voice response systems (IVR) to improve efficiency by reducing agent time on the phone
- CTI
- Predictive dialing for outbound calling
- Integration with Customer Relationship Management (CRM) systems
- Web collaboration and online chat tools for customer support

Beyond the features and technologies, there is a standard suite of typical performance metrics used in the call center management methodology. When a company invests in a call center, it's crucial to monitor performance levels and ensure return on the investment. The most common metrics include:

- Measuring the average delay callers wait in a queue for an agent to come on the line
- Call duration times, generally referred to as Average Talk Time (ATT)
- The time an agent spends on the total call, including preparation, conversation, and after-call work or wrap-up; this is called Average Handling Time (AHT)
- The percentage of calls that get answered within an established call pickup time; this is typically called the service level (for example, 90 percent of calls are to be answered within 30 seconds)
- The raw number of telephone calls each agent handles per hour; this is used as a measure of agent productivity
- How much time an agent spends after getting off the phone in closing out the customer transaction; this is generally called either Wrap-Up or After Call Work (ACW)
- Calls that completely resolve the customer's question or problem on the first call may be tagged with an indicator of First Call Resolution (FCR); the FCR rate is often used as a broad, overall measure of call center performance
- Either the number or percentage of calls that are abandoned by the customer; the higher the number or percentage, the more indicative this is of long holding times, driving customers to hang up
- Idle time is the percentage of time agents spend either not on the phone with customers or in ACW; high idle time may be an indication that the call center is overstaffed during a particular time period; it's used for trend analysis to balance staffing of work shifts in large call centers
- Quality assurance (QA) monitoring may be performed by either a QA team or supervisor; customers hear a standard "Your call may be monitored for quality purposes." message on a routine basis

The acceptance and widespread deployment of VoIP has enabled the staffing of call centers with remote agents. These agents work from home, often on flexible part-time schedules. In the past, they would use a basic-rate ISDN line, but widespread broadband deployments have made high-speed Internet access easy to couple with VPN and VoIP solutions to provide a fully integrated solution for teleworkers.

Clearly, the call center has been a key resource in large customer service organizations for many years. Today, the reduced barrier to entry in call centers has made the technology more available to a wide set of business environments.

## IVR Systems

IVR is a computerized system that allows a caller to choose options from a voice menu and otherwise interface with a computer system. Typically, the IVR system plays prerecorded voice prompts, and the caller responds by pressing numbers on the telephone keypad to select among the options. Many IVR solutions also allow the caller to speak simple answers such as yes, no, or numbers in response to the prompts. These systems have grown more sophisticated in the past few years, and do a much better job of recognizing human voice than they did when they first gained popularity.

Newer systems use natural language speech recognition to interpret the questions that the person wants answered. There is also a growing trend called *guided speech IVR* that integrates live human agents into the design and workflow of the application to help the speech recognition with human context.

Other innovations include the ability for the system to “read out” complex and dynamic information such as email messages, news reports, and weather information using Text-To-Speech (TTS) conversion tools. TTS is computer-generated synthesized speech that has advanced well beyond the robotic voice people may associate with computerized systems of the past. Human voices are used to create the speech in very small fragments that are assembled to create very real-sounding responses before being played to the caller.

IVR systems are used to create service solutions such as airline ticket booking and banking by phone. Unlike voicemail systems, which are one-way communications, IVR systems provide some level of two-way information exchange between the caller and the company systems. An ACD group in a call center may be the customer’s first point of contact. IVRs are often used to provide the primary front-end to a call center, automating the most common calls for account balances or other easily delivered information.

IVR systems today are built with scripting languages such as VoiceXML or Speech Application Language Tags (SALT), not unlike the way Web pages are constructed. In the case of an IVR, the Web server also acts as the application server, allowing the developer to focus on call flows rather than graphic presentation. Typically, Web developers understand and are familiar with tools in this environment and often don’t require additional programming skills.

## CTI

CTI provides for interaction between the telephone and computer systems. As technology has matured, CTI has expanded to include the integration of all customer contact channels, including voice, email, Web, and fax.

## CTI History

CTI evolved from relatively simple screen population (or *screen pop*) technology. This technique allows data collected from the telephone systems, most often via the touch pad, to be used as input data to query databases with customer information. That data can then be populated instantly to the customer service representative’s screen. When the agent already has the required information on his or her terminal screen before speaking with the customer, overall transaction time is reduced.



This technology began in the closed, proprietary roots of every PBX/ACD vendor in the market space. Most vendors eventually adopted the Computer Supported Telecommunications Applications (CSTA) standard. CSTA originated in the ITU and was adopted as an OSI standard in 2000. Some other accepted service and application integration standards in CTI include:

- Java Telephony API (JTAPI) promoted by Sun
- TSAPI and TAPI—TSAPI was originally promoted by AT&T (later Lucent, then Avaya) and is by far the most widely adopted in large-scale contact centers; Microsoft pushed a separate initiative, thus TAPI was born, with support predominantly from Windows-based applications

Some commonly implemented functions in the CTI environment include:

- Information about the caller—This may include the caller's number, the telephone number called, and in many cases, which prompt in an IVR system the caller selected.
- Screen population (screen pop)—This provides the agent with information about the caller before the agent picks up the call. One example is prompting a customer to enter an account number on the telephone dialpad. This allows the CTI system to retrieve the account and present that data to the agent, who can then be fully prepared to talk to the customer immediately.
- Outbound call centers use on-screen dialing tools to increase productivity. Speed dial options and predictive dialing features increase the rate of outbound call placement.
- A software screen control, such as a VoIP softphone, provides access to telephone features through the computer. Functions such as answering a call, hanging up, placing a caller on hold, and initiating a conference might be performed with the mouse or keyboard shortcuts.
- Call transfer and transfer of the data screen provide the ability to transfer a call to a more senior agent or supervisor when the agent involved needs assistance in resolving a problem.
- Administrative functions and duties, such as logging in or out of an ACD group or completing wrap-up work are common.

CTI takes on two basic forms. The following excerpts from the Wikipedia online encyclopedia article about CTI provide helpful descriptions (Source: [http://en.wikipedia.org/wiki/Computer\\_telephony\\_integration#Forms\\_of CTI](http://en.wikipedia.org/wiki/Computer_telephony_integration#Forms_of_CTII)):

**First-Party Call Control**

*First party call control operates as if there is a direct connection between the user's computer and the phone set. An example of this would be a modem card in a desktop computer, or a phone plugged directly into the computer. Typically, only the computer associated with the phone can control it, by sending command directly to the phone. The computer can control all the functions of the phone, normally at the computer user's direction. First party call control is the easiest to implement but is not suited to large scale applications such as call centers.*

**Third-Party Call Control**

*Third-party call control is more difficult to implement and often requires a dedicated telephony server to interface between the telephone network and the computer network. Third party call control works by sending commands from a user's computer to a telephony server, which in turn controls the phone centrally. Specifically, the user's computer has no direct connection to the phone set, which is actually controlled by an external device. Information about a phone call can be displayed on the corresponding computer workstation's screen while instructions to control the phone can be sent from the computer to the telephone network. Any computer in the network has the potential to control any phone in the telephone system. The phone does not need to be attached directly to the user's computer, although it may physically be integrated into the computer (such as a VoIP soft phone), requiring only a microphone and headset in the circuit, without even a keypad, to connect to the telephone network.*

Like many telecommunications technologies, CTI has made great advances over the past 10 years of evolution. The barrier to entry has lowered to a point that CTI is now available to small and mid-sized businesses and is widely adopted as a competitive tool to improve business processes.

### **Application Integration with CRM and Enterprise Resource Planning Systems**

As business advances, there has been enormous acceptance and adoption of an Enterprise Resource Planning (ERP)-driven, Web-centered collaboration in business-to-business (B2B) interaction. For many businesses, this is a competitive move. Companies around the world have focused on improving business processes for many years, spurred early on by W. Edwards Deming's *Out of the Crisis* and his *14 Points for Management*.

Those works, and others, drove business to improve processes, but with advances during the same time period in IT, we've seen growth in the use of ERP systems across businesses of every size and shape, creating more integrated enterprises. Today, efficiency in process drives many organizations to focus on teamwork. ERP systems help dissolve the barriers that exist between different business units or departments. They help break down the *silo effect* so common in the past and change how people work together.

Web-centric business platforms and the success of e-commerce bring new levels of integration, allowing easier support for inter-organization business process integration between trading partners. ERP systems' goal is to integrate all data and process within an organization into a single unified system using computer hardware and software components. One key ingredient to most ERP systems is the use of a single, unified, master database to store data for all the disparate modules involved.

ERP originally implied a system for planning the use of resources across the enterprise, and although it originated in the manufacturing sector, today it is much broader in scope. Today's ERP systems strive to encompass all the basic functions of an organization regardless of the business. ERP systems have spread beyond manufacturing to business, non-profits, governmental organizations, and other large enterprises.

The term ERP system generally refers to an application that replaces two or more independent applications within an organization, eliminating the need for interfaces between systems. This approach provides benefits ranging from standardization and lower maintenance (a single system rather than multiple systems) to improved, dynamic reporting capabilities and enables managers to better monitor the business.

Some examples of different modules that might be available in an ERP system include:

- Manufacturing
- Supply chain
- Financials
- CRM
- Human Resources
- Warehouse management

Each of these discrete application modules support processes require both voice and data communications. Integrating voice services into a CRM module provides a unified customer database, easing customer contact to build relationships. Supply chain modules can build tighter communications processes with vendor partners. Manufacturing and warehouse communications integration can speed time to market. ERP represents the nervous system of an organization and provides an integration point that can tightly couple business process with communications tools to increase productivity and enhance efficiency.

## Customer Relations and CRM

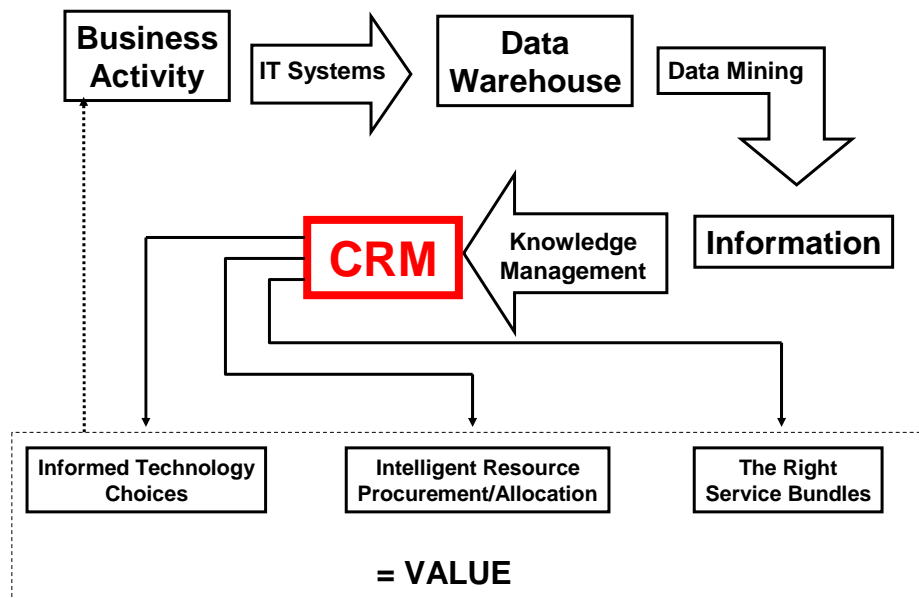
CRM encompasses a broad set of capabilities, methodologies, and technologies that support an enterprise in managing customer relationships. The general purpose of CRM is to enable organizations to better manage their customers through the introduction of reliable systems, processes, and procedures.

CRM is a corporate-level strategy that focuses on creating and maintaining lasting relationships with its customers. CRM tools look at the relationship between a business and its' customers and help manage relationship building and process improvement. Although there are several commercial CRM software packages on the market that support CRM strategy, building value is not about the technologies. CRM often represents a holistic change in an organizational culture and philosophy, placing emphasis on the customer.

To be effective, the CRM process needs to be integrated end to end across marketing, sales, and customer service. An effective CRM program needs to:

- Identify customer success factors
- Create a customer-based culture
- Adopt customer-based measures
- Develop an end-to-end process to serve customers
- Recommend what questions to ask to help a customer solve a problem
- Recommend what to tell a customer with a complaint about a purchase
- Track all aspects of selling to customers and prospects as well as customer support

As Figure 2.1 shows, CRM becomes central to both business process and IT systems. IT systems become a valuable repository of information about the enterprise business, providing insights into further improvements to increase productivity and profitability.



**Figure 2.1: Building value with CRM.**

Customer relationships are managed by a variety of communications tools. There is an old adage in sales that says “People buy from people.” CRM places the focus on the customer, without whom no business can survive. Integration of CRM systems with VoIP technologies provides tighter coupling of several aspects of relationship management.

ERP and CRM solution vendors recognize the criticality of voice communications and are actively working on integration solutions. Because these systems are based on IP technologies, there is a natural synergy with VoIP technologies, which have matured to the point that creative convergence between the network infrastructure, services (voice and data communications), and applications can now provide a seamless working environment.

## Delivering Call Quality with VoIP

Call quality is the single biggest factor leading to user acceptance of VoIP solutions. Later, this guide will talk about managing call quality; this section will compare quality in the public switched telephone network (PSTN) with VoIP, and see how voice conversations are handled.

### ***Traditional Voice Characteristics—the PSTN***

When you implement VoIP, you're trying to use technologies that have evolved in new ways over the past 30 years, specifically, IP. As you start adding voice traffic, a multimedia service, to an IP-based network, you encounter a challenge to networking in general. In the past, networks such as the PSTN were designed to perform specific functions and support a single type of traffic. With VoIP, you're using the Internet, or an IP network, to carry voice traffic.

The problem you face isn't trivial. It actually has ramifications that ripple through every facet of network engineering in both the Internet and in the PSTN. It's a new problem that wasn't anticipated. In the past, when you had a task that required a form of networking, you designed a new network to handle the task because different applications have completely different requirements. IP changed all that because anything that can be digitized can be carried in the payload of an IP packet.

Everything we understand about voice is based on how it's been handled in the PSTN, so let's examine some basic facts about voice transmission:

- Voice calls require long duration for a conversation. A typical phone conversation lasts 3 to 4 minutes. Voice has historically been connection-oriented.
- Voice calls don't tolerate network delay well. 50 to 100 milliseconds of delay is the norm. Voice traffic is generally what is referred to as real-time traffic.
- Voice calls have traditionally been carried in a 4KHz voice channel, with the actual sound frequency being carried ranging between 300 and 3300Hz.
- Converting analog voice traffic to a digital bit stream, you sample the voice at twice the maximum frequency of the channel (8000 times per second).
- Each sample is coded into an 8-bit word.
- These 8000 samples multiplied by 8 bits give us 64Kbps of bandwidth, which matches the telephone network architecture in the PSTN.

These basic facts bring us to the root of a number of the technical problems encountered today. The public phone network has been consciously designed and optimized over a hundred years to deliver voice traffic in the best way possible. This traffic engineering (in telco-speak) has been a critical component of the management and growth of the PSTN. The telecommunications industry has spent years analyzing the network, monitoring its growth, and designing into the network countless optimizations to carry voice traffic effectively. It is a voice network that guarantees delivery of the traffic within the users' expectations.

### IP Traffic Characteristics

There is a packet switched public data network (PSPDN) that was specifically designed to carry packet data. Private data networks have been built around the world to deliver enterprise data applications. If you compare data networks, specifically IP networks like the Internet or the PSPDN, you will find that they have very different characteristics than the PSTN:

- Long duration isn't necessary because transactions are short in duration, or *bursty* in nature. Packets are small in size and can be routed over different paths. IP packets carry delivery information (addressing) in each individual packet. It is a connectionless environment. At higher layers, TCP may be used to layer on connection capability, but this requires the overhead of a three-way handshake and is commonly avoided unless necessary.
- IP packets aren't generally delay sensitive. Email messages sent over a network can be delayed 30 seconds with no problems whatsoever. Web traffic delivered to a browser can be delayed and loaded in the background. IP traffic is often non-real-time traffic. Delays are expected in an IP network. IP itself doesn't even guarantee delivery of the packet.
- IP uses the available bandwidth when it has data to deliver. There are 300bps modems that send data and data transmitted over an Ethernet LAN at 100Mbps. IP doesn't require dedicated bandwidth. Again, IP comes with no guarantees of performance or delivery.

Table 2.1 makes the differences even clearer.

Voice Traffic on the PSTN	IP Network Traffic
Connection-oriented—A dedicated path is established for each telephone call; calls are long duration (4 minutes on average)	Connectionless—Conversations are packetized and transmitted over the best route based on routing protocols; packets are small, so conversations are cut up into many packets
Delivery is guaranteed once the call path is established	Best efforts are made to deliver traffic but there are no guarantees
Designed to use specific bandwidth; the PSTN uses a 64Kbps voice channel	Uses the bandwidth that is available
Real-time voice traffic is very sensitive to delay	IP data traffic is delay insensitive

**Table 1: Voice and data traffic comparison.**

The chart is small, but it makes one very obvious point. IP was not designed to carry voice traffic. IP was designed specifically to carry bursty data over diverse paths, and make a *best efforts* attempt at delivery. Data can be freely discarded along the way if there are any problems. This is nothing like the PSTN in the very design of the technology.

IP networks today are designed, redesigned, and modified to support carrying any kind of traffic. Most often that is what is today called multimedia traffic—a combination of data, voice, and video. Using digitization, any media can be carried inside an IP packet. The challenge in IPT and video networks is delivering business-quality voice conversations with all the characteristics users have come to expect, such as fidelity, clarity, and near-instantaneous delivery.

If users detect what seems to be unacceptable quality, time has shown they will most likely hang up and try again. If this happens frequently, users become dissatisfied and quit using the service. It's clear that to effectively deliver voice services, tools and mechanisms are required for measuring the quality of service (QoS) and guaranteeing that network resources can support the traffic load without degradation of call quality.

One method to address call quality has been to overprovision the network. In many cases, this simply means adding more bandwidth. Although this approach might work initially, in the long run, it's a road to ruin. First, it requires capital investment. Upgrading the capacity of connections, switches, and routers can be a very expensive undertaking. This approach might work for a time in a small local area network (LAN), but in a large enterprise network, it's often too expensive to be a practical solution. In addition, experience makes it clear that if the bandwidth is available, users will fill it. Data applications can quickly consume the additional bandwidth, leaving you with the same congestion and problems you started with in delivering voice service.

To achieve acceptable voice quality over an IP network, factors such as noise, delay, echo, and jitter must be managed to tolerable thresholds. Delay is present in every IP network due to the statistical multiplexing used in routers and the generally bursty nature of packet data. There will be delay. Jitter is nothing more than variation in delay. As different packets can easily take different paths through a large network such as the Internet, variable delay is fairly common. When transmitting voice, jitter can render the audio signal unintelligible to the human ear. Jitter needs to be tightly controlled. Delay can also be caused by the time required to perform the sampling in the codec used to digitize voice. Each router in the network can also induce packetizing and routing delays as decisions are made about how to route each packet. This is also referred to as nodal delay. When designing and managing an IPT service network, it's important to remember that delay is cumulative. Every place delay occurs, it adds to other delay in an end-to-end service between two people.



## Design Considerations and Class of Service

Networks today are large and complex systems, with a variety of applications running concurrently. They evolve and grow so quickly that they are almost organic in nature. There is a concern for any enterprise about the spiraling complexity of network design. When you design a specific set of criteria to support a given application, you create a *class of service* for that application. The danger in doing so is that every new class of service you create increases network complexity. An enterprise with numerous applications could quickly create an unmanageable set of service classes. Many network designers lean toward simplicity and suggest only a few critical service classes:

- Quick delivery can provide for real-time traffic such as voice and video—Real-time traffic such as voice and video require quick delivery or the services are rendered unusable. Video collaboration tools and VoIP telephone calls don't work if the data experiences needless delays. This class of service is used to tag those services requiring quick delivery for usable service. It's important to note that streaming video from a server, such as watching a stored training video, is not real-time traffic. This traffic can be buffered and delayed with lower QoS requirements. Real-time traffic is most often person-to-person traffic between people. Person-to-system traffic generally doesn't possess the same requirements.
- Guaranteed delivery is suitable for mission-critical traffic—Guaranteed delivery is most often used to support “mission critical” data. The CEO's email isn't what is meant by mission critical. The data is what determines its criticality. Mainframe systems running IBM's SNA architecture require delivery but may be tolerant of delay. If packets are delayed, within reason, SNA continues to function just fine. If they're lost entirely, session timers fail and applications no longer work properly.
- Best efforts delivery is what IP provides for everything else—This is the same delivery mechanism used for all IP traffic today. Email, Web browsing, file transfers, and most applications function perfectly well using best efforts delivery. In networks, most of the network traffic falls into this class.

It's easy to identify more classes of service in almost any network. For the sake of simplicity, these three classes give you adequate prioritization capability in this guide and in many networks. Although the debate continues over how many classes of service are really necessary, the bottom line is that any implementation of a QoS mechanism that distinguishes service classes is a vast improvement over the best efforts approach used by IP.


Classes of service are only one aspect of QoS, but they aren't the only factor involved. Let's think about how enterprise networks evolved over time:

- Businesses implemented standalone mainframes in the past, often to process accounting information.
- Computing power grew, and user terminals were added to connect users to the mainframe.
- With the advent of the PC in the 1980s, businesses started deploying PCs and constructing small LANs. These LANs were disconnected. They were essentially islands of information. Information was shared by carrying a disk from system to system (what used to be referred to as Sneakernet).
- The Internet grew in popularity in the 1990s. Companies added routers to connect LANs to each other and to the Internet.
- Many companies also developed an internal network, or intranet, to share company resources and information.
- Public Web servers were added as the World Wide Web became a widespread resource. E-commerce systems were implemented in many business sectors. Business automation tools such as enterprise resource planning (ERP), customer relationship management (CRM), and sales force automation were implemented.
- Today, real-time streaming voice and video traffic are being added to an already busy network.

There was a long-held perception among vendors and service providers that customers' networks were poorly designed. In truth, they often weren't designed at all. Like some unplanned cyber "big bang," they exploded and grew over time based on the needs of the business they supported. It is important to understand that this happened for good reason. Companies in growth mode were successful. This often led to mergers and acquisitions, producing exponential growth. The economy boomed, and networks connected and merged quickly to meet the needs of rapidly growing business. Sometimes networks were redesigned and new technologies were integrated. In other cases, pressing business needs led to cobbling together the best that the beleaguered IT staff could manage. Some of these networks were kept operating only through long hours of frustration and ongoing reconfiguration.

Earlier, this guide noted that networks have typically been designed and optimized for specific types of services. Voice traffic was the first networked application. The PSTN was tuned and optimized to support narrowband voice channels with very low delay.

Traffic engineering for voice is a highly developed field. Its foundation is built on traffic patterns, the busy hour of the day, and statistical mathematics using something called the *Law of Large Numbers*. This law basically says that large groups are easier to predict than small groups. The larger the group, the more group members are likely to be near the average. For example, if you measure 10 people's height, you may find several people (a high percentage of the group) 5 feet tall. If you measure one thousand people's heights, you'll find that the statistical accuracy of the sample is no longer skewed towards shorter people because the sample size provides for greater statistical accuracy. Both telephone company central offices and enterprise PBX systems are engineered for trunking requirements using a mathematical model known as the *Erlang-B* distribution. Using *Erlang-B*, call loads are measured in centi-call seconds (CCS), which is a unit equivalent to a 100-second call. 36CCS represents a telephony circuit at maximum occupancy with zero idle time.

 This guide won't explore these formulas in any depth. The focus is the practical business of managing IPT. For those interested in further study of traffic engineering, the basic formula is:

$$\text{OfferedLoad} = \frac{\text{Calls / Hour} \times \text{AvgMinutes / Call}}{60} \text{ erlangs}$$

The offered load in the formula isn't a problem in a network such as the PSTN because the network has been optimized to support the load. QoS is a manageable issue when the network is designed to serve a single purpose. In today's multimedia environment, there has been a phenomenon often referred to as convergence grab center stage. Separate parallel networks don't make either technical or economic sense. As all of today's data applications converge onto the IP infrastructure, the logical next step in that integration is, for many, to integrate voice, video, and data on a single, multi-service network.

As you look to IP networks as the next-generation architecture for delivery of multimedia services, the fundamental design of IP leads to some complex issues for engineers. IP, and all packet switched networks in use today, utilize *time division multiplexing* (TDM). This is a statistical multiplexing technology used inside the switches and routers in IP networks. The IP Suite protocols and hardware were developed to hand all traffic types, regardless of payload, in the same manner. Routers and other network nodes have traditionally handled traffic not only on a best efforts basis but often using a first-in first-out (FIFO) approach to processing the packet flow.

## QoS

Classes of service describe what's needed in the network. QoS characteristics look at the technical aspects that make it possible to support those classes of service. There are several characteristics within a network that may need to meet certain performance levels for any given service or application:

- **Availability**—Availability is usually represented as the uptime percentage. It's based on the simple premise that the network is there and available for use whenever it's needed. In commercial telecommunications networks, the industry standard is referred to as 5 nines reliability or 99.999% uptime. This availability percentage equates to roughly 5 minutes of downtime per year. For many business networks, that is a challenging availability threshold to meet.
- **Reliability**—Reliability is tied closely to availability, but it's also a design factor. Reliability includes network architecture features such as redundant paths and duplicated or fault-tolerant equipment to ensure that in the event of a failure, the network remains accessible in the event of an outage.



How necessary is reliability? A consideration is how much effort to put into building reliability into the network. Does the network need redundant paths and fault-tolerant or high-availability equipment to guarantee availability in the event of a failure? For businesses that only operate during the 9-to-5 workday, reliability may be an area of less importance.

- **Throughput**—Bandwidth is the most common measure used for throughput. This is simply a measure of how much data inserted into the network at any given source is successfully transmitted through to the destination on the far end over a specific period of time.
- **Error rate**—Many applications are reasonably tolerant of data loss because the TCP/IP suite relies on higher-layer protocols such as Transmission Control Protocol (TCP) to request retransmission and ensure delivery in the event of errors. Not all applications are delay tolerant.
- **Delay**—As described earlier, delay is simply a reality in IP networks because routers use statistical multiplexing to process traffic and much of the transmitted data on any network has a bursty characteristic. There will be some delay in any transmission network due to the laws of physics. Naturally, a lightly loaded or over-engineered network will have lower delay.

- **Jitter**—Jitter is variation in the delay. Because packets can take different routes across the network, delay variation in IP networks is common. Jitter in VoIP conversations results in unintelligible conversations that sound “jerky.” In video transmission, the video stream can break up and show visual signs of “jerkiness” as well. One common test technicians have used for years is the jumping jack tests. The constant motion of a person doing a few jumping jacks gives a good indication of the perceived quality of a video stream.
- **Scalability**—As companies grow and businesses change, the ability of the network to grow with increased needs is an important consideration.
- **Manageability**—There have been many studies demonstrating that the most expensive component of network services is the ongoing operational support of network management and administration. Additions and changes to the existing network can be very labor intensive, and many system designers now consider this factor very early in the development process. A network with a 5-year life cycle will often cost far more to manage over the course of its lifetime than the initial capital cost to build.

### **QoS Approaches in IP Networks**

As noted several times, IP is a *best efforts* protocol that provides no guarantees. That is not to be misconstrued to mean that the various parameters discussed so far cannot be obtained using IP. However, you must remember that network design becomes a vital factor in any implementation of converged services.

Although the techniques used vary widely, all approaches to QoS share a common characteristic. Regardless of approach, with the exception of over-provisioning, QoS is simply a means to implementing a prioritization scheme. In some cases, this prioritization might involve a routing mechanism that aggregates similar traffic types, the routes each traffic type takes over a path suited to the needs of the traffic. There are many approaches with intricate differences, but every approach, save one, adds some form of overhead to the traffic to provide either a prioritization function, or a “traffic cop” function to direct traffic appropriately. The one exception commonly referred to as gigabandwidth has been seen by many to be the ultimate solution, but this fallacy only defers the need for proper engineering to a future point in time.


Ver	IHL	Type of Service	Total Length	
Identification		M F	D F	Fragment Offset
Time to Live	Protocol	Header Checksum		
Source Address				
Destination Address				
Options			Padding	
Data/User Payload				

**Figure 2.2: The TCP/IP packet structure.**

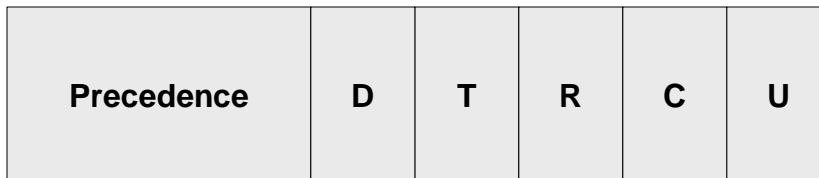
Figure 2.2 serves as a reminder of the format of an IP packet, but it's also shown to review one field in particular. There is a Type of Service (TOS) field in the structure of the IP packet that can and often is used as a QoS mechanism. This prioritization tool was provided in the original specifications for IP.

The TOS field, expanded in Figure 2.3, is one octet, 8 bits in length. It consists of several components:

- Precedence is used purely as a raw prioritization mechanism. The first three bits in binary can represent a precedence value from 0 through 7. These bits provide eight possible levels of prioritization for an IP packet. The higher the precedence value, the higher priority assigned to the traffic.

 IP still doesn't guarantee a level of acceptability but merely a prioritization scheme.

- Delay (flagged as D in the figure) is a field to indicate whether the packet requires low delay or can tolerate higher delay. A one would indicate low delay is required. A zero would indicate more delay is tolerable.
- Throughput (flagged as T in the figure) is a relative indicator, with a one indicating the need for higher throughput or more bandwidth.
- Reliability (flagged as R in the figure) is signified with a one to indicate that a more reliable path is needed.
- Cost (flagged as C in the figure) remains generally undefined and misunderstood in use or intent.
- Unused (flagged as U in the figure) is the last bit that remains available for future use.



## 8 levels of precedence or prioritization

-----

## 4 1-bit fields to further identify packet requirements

*Figure 2.3: The IP TOS field.*

Although the designers of IP built this prioritization capability into the protocol, vendors rarely implemented or took advantage of this feature in the past. For many years, router vendors created OSs that didn't even read the field when processing packets. This was due mainly to a complete lack of standardization. The Internet Engineering Task Force (IETF), the IP standards body, never assigned values to any of the sub-fields in the TOS field. Thus, a precedence of zero might be the highest precedence to one vendor, and the lowest to another. Real-world implementations vary greatly in the absence of a standardized approach, but the field has been used as part of a prioritization scheme more in the past couple of years.

Once you accept that IP provides no guarantees of any kind for either prioritization or delivery, the next step becomes clear. If a user needs to specify some particular network requirement, or QoS, the user must add something to IP to provide it—add it in some other layer of the protocol stack. The following methods are all in use today. They implement QoS in IP by layering some other overhead into the data stream. There are other techniques used to provide QoS, but the following sections explore three common approaches.

## QoS—the Signaled Approach

Integrated Services (IntServ) introduces a signaling protocol to IP. Using IntServ, the user's application sends some sort of call setup signal to the network. This signal is basically a request for a set of service delivery parameters required to complete the call. Using the approach, the network can check resource availability and deliver traffic accordingly. This approach is very similar in process to the circuit setup and teardown associated with a telephone call on the PSTN.

Under IntServ operations, users, through automated software processes, request a particular type of service of the network using Resource Reservation Protocol. RSVP has two key aspects—policy control and admission control. Policy control is used to determine whether the user has authorization to request the specified resources from the network. Assuming the user has the necessary permission, admission control governs the process for allocating and reserving the requested resources or setting up the path for the connection. RSVP passes the user requirements from node to node through the network, requesting resources. If any node along the way is unable to comply, the request is denied and no connection is established.

Real-Time Transport Protocol (RTP) is a widely used supporting protocol in IntServ deployment. It's used to assure the integrity of a session through timestamping and sequencing of UDP segments. Real-time Transport Control Protocol (RTCP) is used in conjunction with RTP and provides some level of monitoring capability.

Although IP theoretically has a rather comprehensive prioritization scheme available, IntServ provides three levels of prioritization. They're familiar because the classes of service identified earlier have been derived from the development of IntServ:

- Best effort—This is exactly what IP provides in existing networks.
- Controlled load, which is described in the standard as providing “data flow with a quality of service closely approximating the QoS that same flow would receive from an unloaded network element.” This is equivalent to the guaranteed delivery class of service described earlier.
- Guaranteed service, which provides a higher level of assurance, which might be required for real-time traffic such as voice or video. This, although called guaranteed delivery in IntServ, mirrors the quick delivery described earlier in this chapter.



## RSVP

RSVP provides a signaling capability that allows a user's application to send a request to the network reserving a particular set of requirements for a voice or video session. These requirements are referred to as the template.



User applications are referred to rather than users because it's important to understand that the person using the computer will not be required to understand QoS or application requirements. The applications will have requirements written into the code to automatically provide the service levels needed.

RSVP allows user applications to identify three of the parameters reviewed earlier that might be necessary for a given application to work properly:

- Throughput or bandwidth
- Delay
- Jitter

In Figure 2.4, the user application makes a request of the network. RSVP policy-control software then evaluates the request against a permissions table and either confirms that the user has the necessary permission to reserve the requested resources or denies access. Assuming the user has permission, RSVP then engages admission control software to determine whether the network has the necessary resources available to complete the request. The network passes the template along the network from node to node in a PATH message until it reaches the receiver. Each node must run an RSVP *daemon* to validate the request for QoS.

In simple form, this specification is a request passed from router to router identifying a need. If an intermediate router can meet the need, the request gets passed to the next node. If not, the request is denied. Intermediate nodes play a vital role in RSVP because every intermediate node must be able to provide the necessary QoS PATH requested in the template. Lack of resources at any point along the path can cause the call to be dropped.

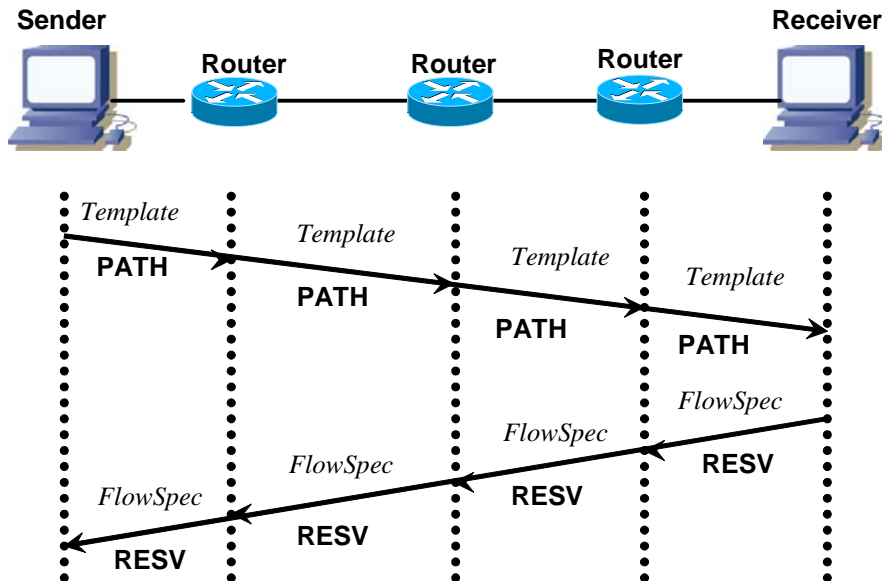


Figure 2.4: Using RSVP to reserve network resources.

Once a path has been identified, each node must pass a flow specification in a reservation (RESV) message back along the network from node to node. As you've gone to the trouble of reserving resources across the network, you now must ensure that you use the resource allocated. This *flowspec* reservation is each router's way of knowing where the next router in the reserved path is to pass packets associated with the call or session.

There are clearly a several factors to consider when implementing RSVP for QoS. These factors become crucial when using the Internet, as those routers are not all under the control of the implementer.

- RSVP is overhead intensive. Call setup isn't something IP was designed for. Adding a signaling protocol in the mix increases the processing CPU cycles required at each router in the path.
- Routing protocols such as RIP and OSPF support only a single routing metric; they don't understand the concept of reservations. RSVP does not provide any solution for this issue. The implication is that to be successful, the network may have to be over-engineered beyond expected carrying capacity.
- RSVP scales very well in the multicast network environment (one-to-many transmissions) but does not scale well for unicast (one-to-one) traffic. As telephony is primarily a one-to-one connection for an end-to-end service, scalability to support VoIP and video conferencing services can quickly cause problems in a large enterprise network. In a multi-provider network such as the Internet, where nobody "owns all the routers, scalability may be unattainable.

- As all the intermediate nodes must be able to comply with requests, every router in the Internet would have to be either upgraded to support RSVP or replaced. This would require universal support and acceptance that just doesn't exist. Routers have to maintain state tables containing information about every session. The processing load and memory requirements can drive router cost up significantly.
- RSVP doesn't provide any QoS whatsoever; it's simply a mechanism for sending requests to the network for some specific requirements. It's a signaling protocol, and it still requires help from other protocols or other methods to truly implement QoS.

Many network designers agree that IntServ and RSVP provide a potential solution but only in a network that is completely under the implementer's control. Each and every intermediate router still presents a potential single point of failure. In the Internet, data crosses many provider networks en route to its destination. IntServ can work well in a fully controlled environment, like an enterprise-owned private network. It may also prove useful at the edges of the Internet, either in customer networks or the local metropolitan portion of the Internet provider's network.

Internet service providers have almost universally chosen not to adopt IntServ as a solution to the QoS problem. The value provided just doesn't offset the cost of conversion and implementation. IntServ fails the ROI case for service providers. IntServ has found some application in private networks and is still supported by all the major router vendors. Resource Reservation Protocol for Traffic Engineering (RSVP-TE), a variation, has been deployed with success in conjunction with Multiprotocol Label Switching (MPLS).

### **QoS—the Provisioned Approach**

There is another approach that requires specific routes through the network that are predefined and available for each type or class of traffic called Differentiated Services (DiffServ—also called DiffServ Code Point or DSCP). These paths might be pre-existing as part of the network design or they might be set up on demand in some manner. This second option might include some type of signaling protocol such as RSVP to establish the paths. The DiffServ method is often used as a traffic aggregation approach in order to direct similar traffic types on similar network routes.

This guide won't consider the evolution to IPv6 but will make an observation: In the development of IPv6, the TOS field was the subject of direct discussion. In IPv6, the protocol has been expanded to provide improved granularity in the delivery of QoS. In the IPv6 packet, there is a field referred to as the DiffServ field as a replacement for the current TOS field. It uses six bits of this field as the DSCP to identify how the nodes in the network should handle each packet. Routers handle packets based on a set of forwarding treatments or Per Hop Behaviors (PHB). These rules must be predefined in each network element or node. This is the provisioned approach to QoS in the network.

 DiffServ is also a development of the IETF. The DiffServ working group charter and information is available at <http://www.ietf.org/html.charters/DiffServ-charter.html>. The objective of this working group is to employ “a small, well-defined set of building blocks from which a variety of aggregate behaviors may be built.” This suite is also defined by a series of RFCs:

RFC 2386—Per Hop Behavior Identification Codes

RFC 2474—Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers

RFC 2475—An Architecture for Differentiated Services

RFC 2597—Assured Forwarding PHB Group

RFC 2598—An Expedited Forwarding PHB

RFC 2983—Differentiated Services and Tunnels

RFC 3086—Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification

RFC 3140—Per Hop Behavior Identification Codes


RFC 3246—An Expedited Forwarding PHB

RFC 3247—Supplemental Information for the New Definition of the EF PHB

RFC 3248—A Delay Bound alternative revision of RFC2598

RFC 3260—New Terminology and Clarification for DiffServ

The DiffServ approach is to categorize traffic into classes of service. Similar services are aggregated and treated the same way in the network. Thus, network paths have to be preconfigured to support each class of service. Packets are “tagged” at the edge of the network and the appropriate forwarding treatment for that class of service tag is applied. This approach results in a much coarser granularity at each router and reduces the need for large state tables. This helps control the need for CPU processing power.

 DiffServ has two primary components described in RFC 2475. Packet marking redefines the TOS field of the packet and uses six bits as a coding scheme to classify packets into a class of service. The use of six bits provides for a prioritization scheme that can identify 64 different types of traffic or aggregates.

PHBs govern how an individual class or aggregate is handled. This is defined via behavior aggregates. In essence, the PHB describes the scheduling, queuing, and traffic shaping policies used at a specific node for routing the traffic.

DiffServ can scale to very large enterprise and provider networks. It is widely supported by manufacturers and has been deployed by several large service providers. There is far more to DiffServ than can be addressed in this guide, but information and technical specifications are easily located on the Web.

## QoS—the Bypass/Shim Approach

MPLS is a method of providing QoS that removes the usual hop-by-hop routing in IP from the equation. MPLS adds a “tag” to every packet. This tag shortcuts the delivery by directing the packet to the best available path for the type of traffic associated with the MPLS tag. MPLS is compatible with frame relay and ATM networks. It has been widely adopted in enterprise and service provider networks for adding QoS capabilities to support VoIP and video traffic. MPLS has often been referred to as a bypass or “shim” approach to QoS. Because of the addition of a tag into the data stream, it’s also often referred to as a *Layer 2½ protocol*.

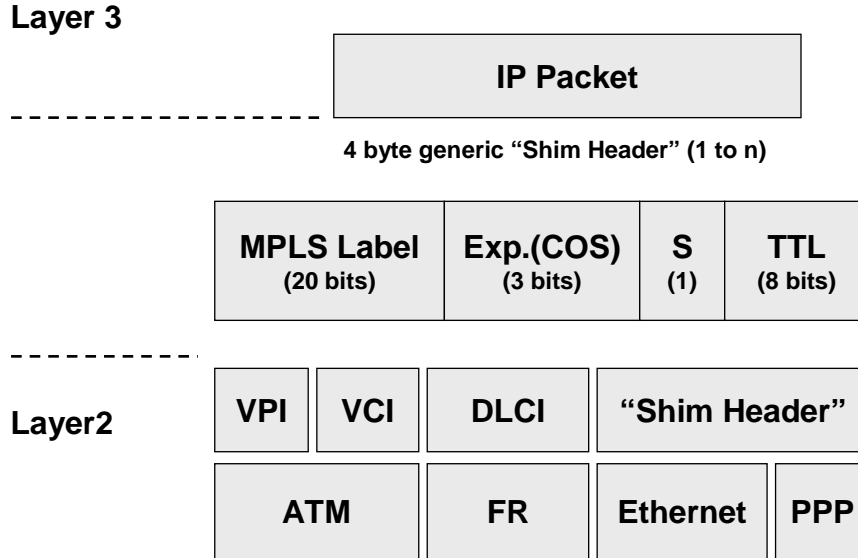
### Using MPLS for QoS

Implementing VoIP makes companies re-evaluate their service networks. Often, it brings some fundamental changes to the way these IP networks operate. As has already been reviewed, IP is a best efforts protocol with no assurances of delivery or quality. When organizations implement VoIP, many businesses discover a compelling need to implement some QoS methodology in order to provide acceptable voice call quality. MPLS is a method that allows packet-based networks to essentially emulate some of the behavioral properties of a circuit-switched network, such as the PSTN. The following section provides a look at MPLS as one method that’s frequently used to deliver QoS assurance in a large-scale enterprise environment.

MPLS evolved from several different but similar techniques. One of the leading approaches was developed by a group of engineers at Cisco Systems. Initially, this proprietary protocol was called *tag switching*. As the technology evolved, industry leaders came together in a consolidated effort under the auspices of an IETF working group pursuing an open standard that became MPLS.

### How MPLS Works

MPLS works by pre-pending packets with an MPLS shim header, or tag, to the beginning of a packet. Although this guide is focused on IP networks, MPLS works equally well with frame-relay frames and ATM cells in those networks. This shim header contains one or more *labels* and is often called a label stack.



**Figure 2.5: The MPLS packet and MPLS encapsulation.**

As Figure 2.5 shows, each label stack entry contains four fields:

- A 20-bit label value
- An experimental field, often used to denote class of service
- A 1-bit flag to signify whether this label is the last label in a stack
- An 8-bit time to live (TTL) field

These MPLS-labeled packets are switched by performing a label lookup/switch instead of a lookup into an IP routing table. Because label switching can be performed within the switching fabric of the hardware, it processes much faster than traditional router-based IP address lookups process.

In an MPLS network, routers become *label switching nodes* that add and remove labels as needed. You'll hear the term *label popping* from time to time. Using this method, a large routed network looks like one single routing hop at Layer 3 of the TCP/IP stack. Path selection is based, not on a routing metric, but on the MPLS label. Routers may be thought of as Label Switching Routers (LSR) through the core of the network, with Label Edge Routers (LER) at the border points of the MPLS domain. Routers that provide ingress and egress to the MPLS are often called Provider Edge (PE) routers.

The Class of Service (COS) field is used to assign classes of service much as described earlier in the chapter in a section that looked at three classes. MPLS often adds a fourth class for management traffic. Typical MPLS classes of service are:

- Real-time traffic, such as voice and interactive video, is often given the highest priority to ensure that adequate bandwidth can be provided. The class of service is also designed to provide the delay, packet loss, and jitter requirements suitable for delivery of real-time traffic such as VoIP and video.
- Mission-critical data traffic, such as that from mainframe computers, is often classified in a guaranteed delivery class of service. Timing of delivery is often a more stringent requirement than bandwidth for this type of data.
- Management traffic is commonly aggregated into a management class of its own. Management traffic requires some assurances that it can still be passed regardless of congestion in the network to ensure QoS is maintained.
- All remaining traffic is generally lumped into a best efforts class of service that mirrors how IP networks deliver traffic normally.

### ***The Experimental Bits and QoS***

Using this approach, similar traffic types can be aggregated into the same class of service, not unlike DiffServ. When addressing is applied to the packets, they might be labeled by an application, a router, a switch, or some other mechanism at ingress to the network. This COS header is used to aggregate packets into what is called a forwarding equivalency class (FEC) for switching throughout the network. The undefined experimental bits are used by most MPLS vendors to carry the latest 3 of 5 DSCP priority bits in the IPv4 header, as a simple emulation of traffic prioritization.

When a labeled packet arrives at an MPLS router, the topmost label is “popped” and examined. Based on the contents of that label, a *swap*, *push*, or *pop* operation is performed on the packet’s label stack. Routers can have predefined lookup tables so that they can process the packet very quickly. In a swap operation, the label is swapped with a new label, and the packet is forwarded along whatever network path is associated with that new label.

In a push operation, a new label is pushed on top of the existing label, effectively *encapsulating* the packet in another layer of MPLS. This allows the hierarchical routing of MPLS packets. Notably, this is used by MPLS VPNs.

In a pop operation, the label is removed from the packet, which may reveal an inner label below. If the popped label was the last on the label stack, the packet leaves the MPLS tunnel or domain. This is usually done by the network egress router.

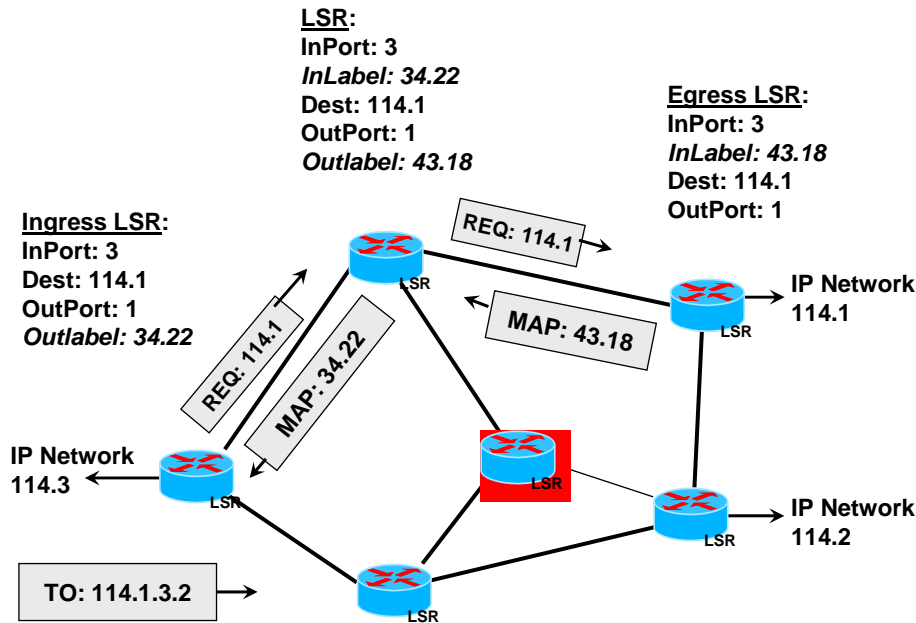


Figure 2.6: Label switching with MPLS.

During these operations, the contents of the packet below the MPLS label stack, the IP packet and payload, are not opened or processed. Intermediate transit routers only need to read the topmost label on the stack. Packet forwarding decisions are based on the contents of the labels, providing *protocol independent packet forwarding*. There is no need to look at a protocol-dependent routing table. This eliminates the cumbersome IP longest prefix match routing protocols require at each hop.

When the packet leaves the network at an edge router, and the last label has been popped, only the payload remains. This can be an IP packet or any of a number of other kinds of payload packet in other networks. The edge router must therefore have routing information for the packet's payload because the router must forward the packet using traditional routing methods. An MPLS transit router has no such requirement.

- 📖 For more information about MPLS, check out the following sites:
- 📖 MPLS Resource Center at <http://www.mplsrc.com/index.shtml>
- 📖 MPLS MFA Forum at <http://www.mplsforum.org/>
- 📖 IETF MPLS Working Group at <http://www.ietf.org/html.charters/mpls-charter.html>
- 📖 IETF RFC 3031 at <http://www.ietf.org/rfc/rfc3031.txt>




### Comparing MPLS to IP

MPLS can't be compared directly with IP as a separate entity. They are complementary protocols. MPLS works with IP and IP's interior gateway (IGP) routing protocols. MPLS brings a level of basic traffic engineering capability to IP networks. MPLS relies on traditional IGP routing protocols to construct the label forwarding table, and the scope of any IGP is usually restricted to a single service provider for stability and policy reasons. There isn't a current standard for interoperable carrier-to-carrier MPLS, so it's not yet practical to span one MPLS service across more than one carrier.

### MPLS Deployment

MPLS is currently in use in large IP only networks and is standardized by IETF in RFC 3031. In practice, MPLS is mainly used to forward IP packets and Ethernet traffic. Major applications of MPLS are telecommunications traffic engineering and MPLS VPNs.

 Traffic engineering considerations with MPLS include:

IGPs mainly use "shortest path" algorithms

Path "overlap" causes congestion in the network

Traffic can overwhelm a short path while others are underutilized

IGPs in large networks present challenges because equal-cost multi-paths (ECMPs) share loads when they really should not, load sharing doesn't happen on multiple paths with different costs, and IGP metrics modified for traffic shifts tend to have side effects

## Summary

This chapter began by considering high-level the business models and processes because they're most often the ones directly affected by VoIP. They're also central focal areas as the concept of unified communications brings network services, such as voice and video, and applications together. The chapter then moved into factors crucial to providing call quality. The network can't effectively converge and thrive if suitable quality isn't present. The next chapter will delve into drivers for change in three areas: business drivers, cost reduction drivers, and strategic drivers.

## Download Additional eBooks from Realtime Nexus!

Realtime Nexus—The Digital Library provides world-class expert resources that IT professionals depend on to learn about the newest technologies. If you found this eBook to be informative, we encourage you to download more of our industry-leading technology eBooks and video guides at Realtime Nexus. Please visit <http://nexus.realtimepublishers.com>.