

Realtime
publishers

"Leading the Conversation"

The Definitive Guide™ To

Successful Deployment of VoIP and IP Telephony

This eBook proudly brought to you by



PROGNOSIS®

Jim Cavanagh

Chapter 4: Design and Pre-Deployment Testing	78
Network Design Considerations	79
The Team	79
Assumptions.....	83
The Voice/Data/Video “Triple Play” and Unified Communications	84
Real-Time	85
Near Real-Time.....	86
Background/Batch.....	86
Interrelationship of Multi-Media Services.....	86
Application Availability.....	86
Delivery/Accuracy/Loss	90
Delay	92
Delay Variation or Jitter.....	92
Differentiation, Prioritization and Queue Management.....	93
Voice QoS.....	95
QoE	96
QoE vs. QoS.....	96
Voice Metrics and Measurements.....	101
Impact on Business Processes.....	103
Standardizing Positive Impacts.....	104
Avoiding Negative Impacts	104
Translating Needs to SLAs	104
The “Proper” SLA.....	105
Metrics	105
Measurement.....	106
Establishing Feedback Loops and Review Cycles.....	106
Validation of Design.....	107
Network Tuning	107
Test Bed Architecture	108
Network Impairments	108
Timing/Synchronization Issues.....	109
Broadband Bandwidth Measurement and Calculation	110
Gateway, Softswitch, Session Border Controller and IP Device Capacity and Performance	111

Testing, Feedback, Network Modification, Testing Loop.....112

 Testing/Proof of Concept Objectives.....113

 Feature Support and End-to-End Operation113

 Closed Lab Environment115

 After Hours/Off Hours Familiarization and Benchmarking115

 Busy Hour Testing in a Live Network.....115

 Evaluation of Results and Ensuring End-User Acceptance.....115

 Fall-Back and Contingency Planning118

 Use of Third-Party Tools118

Summary.....119

Copyright Statement

© 2007 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.

[**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 4: Design and Pre-Deployment Testing

Chapter 3 described the first phase in the life cycle of any successful IPT project—planning and assessment. The information that you gathered and the steps that you took in that phase set the stage for the next step: design & pre-deployment testing. As Figure 4.1 shows, planning and assessment, design and pre-deployment testing and implementation are all initial phases. After these initial phases, you enter the actual cyclical part of the life cycle of your IPT project.

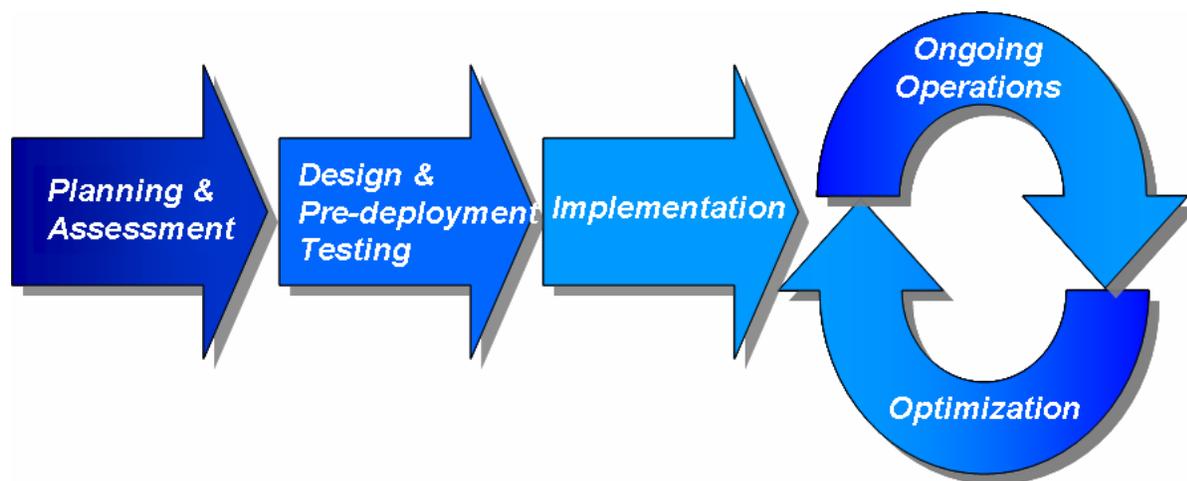


Figure 4.1: IPT project life cycle.

The major areas within Design and Pre-deployment Testing are:

- Network Design Considerations
- Translating Needs to Service Level Agreements (SLAs)
- Validation of Design
- Testing → Feedback → Network Modification → Testing Loop

We will discuss each in great depth in this chapter.

Network Design Considerations

Chapter 3 covered a variety of areas fundamental to network design. Although this chapter will not completely rehash the discussion, it will pull certain assumptions from the previous chapter that will provide a basis for the exploration of network design in this chapter.

The Team

At this point in our project life cycle, the company implementing the IPT project may or may not have engaged, or even considered engaging, outside assistance. Regardless of whether it has come up or not, this is the time to make some difficult decisions. At this point, the company would choose a “do-it-yourself,” “managed service,” or turn-key “outside service provider” approach—or depending upon geographic coverage and other considerations, some hybrid of these solutions. It is also important to distinguish the team that will provide the network transport from the team, or part of the team, responsible for the telephony and multimedia systems themselves. Table 4.1 shows prospective team members, their possible roles in the project, and the pros and cons of selecting a specific prospective team member. Considerations of budget must be balanced with the fact that no other single project is likely to have such a profound impact on the overall operation of the organization as a project involving telephones.

Prospective Team Member	Role	Pros	Cons
Current PBX Vendor Professional Services or Consultant(s)	Provide bridge between current capabilities and capabilities of new system	A lot of knowledge is locked in the current PBX configs, tables, documents, and so on	May not see the importance of their role or may be reluctant to help in the decommissioning of their equipment
Current IP Network Tech	Benchmark current network Assist in testing Get cross-trained in voice to be multi-tech	Know current network Already on payroll No “spin up” time for IP networking	Not familiar with voice May be rivals of voice group May view voice people/app as simple and inferior to data people/app
Current IP Network Mgmt	Collaborate on design Collaborate on testing Get cross-trained in voice	Know current procedures, procurement, budgeting, and so on Know current “players”: techs, management, suppliers, carriers, and so on Already on payroll No “spin up” time for IP networking	Not familiar with voice May be rivals of voice group May view voice people/app as simple and inferior to data people/app
Current Voice Tech	Ensure all voice functionality synchronized Telephony system benchmarking and testing Get cross-trained in IP	Know current voice needs, apps, and users Already on payroll No “spin up” time for telephony	Not familiar with IP May be rivals of IP/data group

Prospective Team Member	Role	Pros	Cons
Current Voice Mgmt	Collaborate on design Collaborate on testing Get cross-trained in IP	Know current procedures, procurement, budgeting, and so on Know current “players”: techs, management, suppliers, carriers, and so on Already on payroll No “spin up” time for telephony	Not familiar with voice May be rivals of IP/data group
Temp/Contract Multi-Tech	Assist in install Go away	Temporary Go away after install	Not familiar with network Will take “lessons learned” and valuable knowledge away when they leave
Current Multi-Tech (<i>if they exist</i>)	Assist in voice and data benchmarking Assist in testing Ensure all voice functionality synchronized Ongoing support of multimedia system	Know current network Already on payroll No “spin up” time Can provide cross-training to IP/data and voice techs	More expensive than IP/data or voice techs Difficult to hire Difficult to retain
New Multi-Tech	Assist in voice and data benchmarking Assist in testing Ensure all voice functionality synchronized Ongoing support of multimedia system	Expand tech staff Mandatory if “do-it-yourself”	Aren’t familiar with network and users More expensive than IP/data or voice techs Difficult to hire Difficult to retain
Network Equipment Vendor	Ensure that the underlying network components—switches, routers, firewalls, and so on—are capable of handling the load of the new services, providing appropriate quality of service (QoS), etc.	They have intimate knowledge of the underlying infrastructure and stand to benefit if upgrades are needed	May be difficult to work with if their IPT solution was not chosen
IT/Network Management	Manage voice and data issues Manage budget Ensure system reliability and security Coordinate integration of voice with VPN	Know current procedures, procurement, budgeting, etc. Know current “players”: techs, management, suppliers, carriers, etc. Already on payroll No “spin up” time Data, voice, and video responsibility	Possible rivalries

Prospective Team Member	Role	Pros	Cons
Consultant	Assist in lab testing and initial implementation Go away	Temporary Go away after install May have Best Practices methodology to leverage	Not familiar with network Does not know voice requirements in advance Will take valuable knowledge away when they leave
Manufacturer Prof Services	Assist in lab testing and initial implementation	Deeper understanding of equipment and software May have Best Practices methodology to leverage Should cross-train/do knowledge transfer Temporary Go away after install	Not familiar with network Will take valuable knowledge away when they leave
Internet Service Provider	Network benchmarking Bandwidth adjustments QoS Issues VPN/tunneling/security SLA-related issues	Familiar with network and infrastructure critical to success ISP provide infrastructure VPN, tunneling, and security are needed services and difficult for "do-it-yourself" SLA coupling to project critical	May see telephony service provider as rival
Data Systems Integrator	Benchmark current network Assist in testing Assess performance impact on data apps Prioritization	Critical supporting skill sets May already know network and applications/users	May see ISP and/or voice systems integrator as rivals
Voice Systems Integrator	Benchmark current network Assist in testing Get cross-trained in voice to be multi-tech Assess performance impact on data apps Prioritization	Critical supporting skill sets NOT needed if using telephony service provider	May not already know network, users, and needs

Prospective Team Member	Role	Pros	Cons
Telephony Service Provider	Turn-key system implementation Coordinate carriers and service providers Ensure SLA performance, telephony features, and system reliability	Turn-key Alleviates need to interface to ISP No need to interface with carriers “One-stop shopping” May be less expensive than “do-it-yourself” Price may include upgrade/technology refresh May include dedicated Help Desk and/or account team and/or on-site personnel No need to keep telephony or multi-tech skill sets on staff	Probably does not know network or users and needs May be more expensive than alternatives
Voice User Representatives	Provide representation of the users in the implementation process	Provide much-needed user viewpoint	Must be in touch with the agreed upon project objectives or they may hinder progress and risk deadlines

Table 4.1: Prospective project team participants.

A big part of the selection of the different elements of the project team is to assign responsibilities, work out any overlaps or conflicts, and test any assumptions regarding the responsibilities of each group. And, in keeping with good project management practice, there should be an actual human responsible for ensuring that each and every task is completed, on time, on budget, and to the standard of quality set for the task.

Project teams will work together with management and user representatives to decide such things as the rollout plan, phases, order of implementation, and schedule. Project teams must also coordinate on such issues as where the IP PBXs (call servers) and gateway(s) will go, call routing configuration, and specifications for other services such as conference bridges, voicemail, auto-attendants, automatic call directors, media servers, and other accessories and appliances. And, if the project is a “do-it-yourself” project, the location, hours of operation, operations procedures, and staffing for Help Desks and Network Operations Centers (NOCs) will have to be determined. If, however, Help Desk and/or NOC services will be handled by an outside provider, the location, staffing, and operating procedures will have to be specified in an SLA. This issue seems a bit more trivial if the network is to be in a single country than if it is a multinational operation.

Based on rollout phases and schedule, the data team builds its own plan for the network design to ensure that needed bandwidth, network configuration for quality of service (QoS), and connectivity points are provided for the overlay telephony infrastructure. In parallel, the internal telephony team or systems integrator must build the solution in the lab for the proof of concept phase that is a mandatory step prior to implementation. If a “turn-key” system is being used through a telephony services provider, a very close collaboration must be maintained to ensure the proper interpretation and implementation of user needs and desires prior to a system test. These steps are addressed in more detail later in the chapter.

Assumptions

First, let’s assume that your organization has an existing IP infrastructure—either a traditional intranet or a VPN that uses the Internet—and that any desirable changes to the network have been made prior to, during or after the assessment phase of the project but before the planning phase. Changes, such as a shift from a private IP network to a MPLS VPN service, the implementation of QoS mechanisms in the LAN and WAN to support prioritization of voice traffic, or possibly the correction of bandwidth deficiencies in the network have already been accomplished. Let’s assume that your existing IP network is one of the following:

- Provided and managed completely by an in-house department over telecommunications facilities leased from carriers
- Completely provided by an outside carrier or service provider (via a VPN or similar service)
- A hybrid of these two approaches

In addition, let’s assume that the IP network is designed to be a multi-media network including at least two of the following media types: data, voice, video. If video is not already accommodated and in use today, it may be necessary at some time in the future given the commodification of desktop video and emergence of telepresence and high-definition video systems. Baselines exist for voice quality metrics, such as MOS and other QoS measurements from the assessment effort, and it will be your job in the network design to assure that the baselines are met or exceeded.

Finally, let’s assume that user satisfaction is an important component of the network design and, therefore, you will prioritize voice quality and call quality over bandwidth and processing power of individual systems such as IP phones and gateways. It is understood, of course, that bandwidth and CPU capability are important contributors to both cost and overall network performance, but you will err on the side of the user in this balancing act. Let’s also assume, going into this phase of the project, that a complete inventory and synchronization of the telephony features needed by users has been done, as described in depth in Chapter 3 and that all parties understand the needs of the users in regard to replicating existing functions that are necessary (vs. replicating the dozens of traditional PBX features that aren’t necessary for your organization) as well as their expectations for new functions and features they will gain as a result of the implementation of the new system.

In terms of the actual implementation, let's further assume that the network will require attachment to the Public Switched Telephone Network (PSTN) to allow non-IP and off-net calls and to ensure 911 calling. For this reason, your design will provide gateway access until such a time that the entire world is packetized voice, an eventuality that lies several years, if not decades, into the future. The assumption of interconnection to the PSTN will not only assure connectivity outside the organization but internal connectivity during any phased implementation. PSTN connectivity also puts the network into a different category from a legal and regulatory perspective. In the United States, for instance, interconnecting to the public network can subject VoIP providers to regulations regarding 9-1-1 calls for public safety as well as compliance with the Communications Assistance to Law Enforcement Act (CALEA), also known as the "wire tap law," which has recently been amended to include packet-based networks in addition to the circuit-based networks that it has always covered. These regulations may increase enterprise liability especially in light of other governance regulations and bear consideration by organizations of any size.

Implementation will be phased by a department, geographical area or some other criteria, to be determined so that you can have a phased implementation. A phased implementation plan is crucial to success, allowing tight management of migration with an eye toward minimizing disruption to business operations and isolation of any problems encountered during migration so that they do not cascade through the network and negatively impact users.

The Voice/Data/Video "Triple Play" and Unified Communications

The term "triple play" refers to the provisioning of voice, data and video services over a single network. Although the term triple play originated with residential carriers and service providers, it is also a convenient shorthand for organizational networking. The primary difference between the residential triple play and organizational triple play is the fact that in the residential environment there is a much stronger emphasis on video. Video as a multi-media component has yet to really take off in most organizations (though video is a far more important element in the business environment that it has been in the past).

Another often heard term is unified communications. Although the term *triple play* is a carrier term that is finding increasing acceptance in the service provider and enterprise communities, the term *unified communications* (or *UC* for short) is an enterprise term that is sweeping through the service provider and carrier markets like wildfire. The two terms are not synonyms—are not interchangeable. The term triple play really refers to the basic data, voice, and video services while UC refers to applications that use one, two, or all three of the media. For example, voice and data are two media. Integrated voicemail, one example of dozens of very exciting UC services, can make use of both. For example, IPT is used to deliver a spoken message from the voicemail server to your telephone if you dial in to the voicemail system. However, if you want to retrieve the voicemail messages and download them to your PC or handheld devices, the voicemail files are treated as data files and are downloaded using a data-type session and likely a traditional data protocol such as File Transfer Protocol (FTP) or Trivial File Transfer Protocol (TFTP).

Let's take a closer look at the three "multis" of multi-media—voice, data and video—and break each down further into three different types or flavors of service: real-time, near real-time and background or batch. Real-time and near real-time are distinguished from background or batch by the fact that real-time almost always has a human involved on at least one end of the communication and very often on both ends. What this means is that delay and delay variation are far more important in real-time and near real-time than they are in background or batch. Networks must, therefore, be optimized to accomplish a higher quality of experience for the real-time and near real-time end-user applications. By contrast, background/batch applications do not happen in real-time and, in fact, in many cases are not even attended by a human but rather by a computerized process.

Real-Time

In the case of voice applications, for instance, Table 4.2 shows two sample applications for voice: VoIP and IPT. Both are interactive, bi-directional telephony applications and both usually involve a human on both ends of the connection unless, of course, a call goes to voicemail or an Interactive Voice Response (IVR) system. Such a system may, in fact, allow navigation of a menu for an automated response, such as an account balance or airline arrival time, or may end up being a human-to-human call. VoIP differs, however, from IPT in that VoIP is a simpler voice communication, often part of a Netmeeting or other conference, and IPT includes phone numbers, signaling, and all the other necessary elements of traditional telephony.

Real-time data applications include interactive chat or Instant Messaging (IM), whereby one user types on a computer keyboard or phone keypad and the results of the typing appears on a distant screen; or whiteboard applications, where one either draws on the screen or on a tablet for the benefit of viewing of the human at the other end of the connection. A perfect example of a real-time video application is video conferencing, which is really the same idea as two-way voice but with a moving picture in addition to speech.



One important multi-media consideration, which is beyond the scope of this guide, is the synchronization of the simultaneous voice and video in real-time multi-media conferencing or entertainment video applications.

	Real-Time	Near Real-Time	Background/Batch
Voice	VoIP, IPT	Pre-recorded audio, Voicemail, IVR	Voicemail Synch
Data	Chat, IM, Whiteboard	Browser, FTP, email, SAN, Telemetry	Database Synch, Archiving
Video	Video Conferencing	IPTV, Training Video (One way)	Video Cacheing

Table 4.2: Voice, data and video application examples.

Near Real-Time

Near real-time applications are represented by such applications as prerecorded audio, browser, File Transfer Protocol (FTP), traditional and multi-media electronic mail, storage area networks (SANs), telemetry, IPTV, and one-way training video. Near real-time applications are similar to real-time applications in that a human is watching and waiting for the information to arrive. They are dissimilar in that the timing or pacing of information arrival for near real-time does not need to approach the parameters of low delay and small delay variation required by interactive human speech and video conferencing.

Background/Batch

Examples of background/batch are voicemail and database synchronization and archiving, and video caching. No human is explicitly waiting for the completion of the task, so these applications can have the lowest priority and, therefore, the worst performance characteristics.

Interrelationship of Multi-Media Services

Although there is certainly an inclination to ask each person or department responsible for each application—voice, data and video—to take charge of their own particular area of the design for the new network, this is absolutely the wrong approach. What all parties must realize is that any successful multi-media network is an ensemble performance. In fact, what should really occur is that all political issues, traditional departmental loyalties and other issues that will stand in the way of a successful, combined, multi-media network should be addressed directly and eliminated. This should be done prior to bringing in any carriers, service providers, integrators or any other third parties. Because the voice/telephony project is usually a matter of integrating a new application—voice—or new applications—voice and video—into an existing IP infrastructure, what should occur is that application ambassadors, or ombudsmen, should be designated to act in the best interests of the users of the specific applications. But, instead of pitting one group against another with the outcome being a victory for that specific group, the idea is for ambassadors or ombudsmen to represent their users with the objective of coming up with the very best possible multi-media solution, given time constraints, budgets, objectives, and management and user expectations, keeping in mind that many humans will be users of more than one of the multi-media services.

Application Availability

Application availability is by far one of the most obvious aspects of system operation and the one that will be judged most harshly by the user community. Application availability will drive many other design decisions. In fact, consideration of application availability should make many of the possible design tradeoffs clear and many of the choices obvious.

Application availability has many components, all of which must be considered very carefully during the design phase. Figure 4.2, for instance, highlights some of the key design considerations and decision points. Evaluating each one in terms of application availability will provide guidance to the decision-making process and increase the likelihood of a correct outcome.

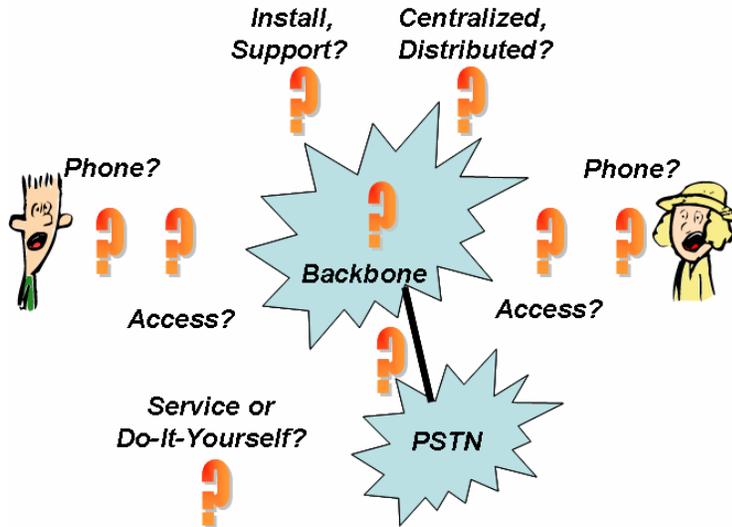


Figure 4.2: IPT design decision points.

For instance, what type of device should a user have to make and receive calls and how should the user be connected to the network? Part of the design process is an inventory of who has what type of phone today. Typically there are many flavors of analog and digital phones among the user base and this will lead to what type of IP phones they get (good, better, best) as well as whether their old analog/digital phone will remain. In one IPT survey, 25% of the respondents plan to continue using analog/digital phones. Careful consideration of this point is important as it's a critical factor in cost. There are many ways to categorize users, but for the sake of this discussion, let's say that the user is an office worker. If you allow users to keep their current telephone devices and use a converter such as an analog terminal adapter to convert their existing signals to the ones used on the new IPT network, you can probably save money over the cost of purchasing a new IP phone.

But, with telephony service availability in mind, you would probably choose a new IP phone. Why? For one thing, a new IP phone will most likely be more reliable than an old phone with an adapter and will also, most likely, provide easier access to a broader range of the capabilities of the new IPT application. The reason for this is that both a new IP phone and the analog terminal adapter that can be used to attach the traditional telephone to the new network can have similar problems, such as out-of-date software or hardware problems, but the IP phones are typically more intelligent and have more processing power and better diagnostics. How about connection to the network?

There may be some cost benefits to connecting to the network through an existing PC, but what happens if the PC is out of service, could the phone also be unavailable? VoIP phones are sold in configurations that allows physical connectivity of the phone without the PC needing to be turned on as well as those that use software on the PC to supplement some or all the functions of the telephone. Which does your organization have? Or, what about a situation in which the PC is connected via the phone? Some connections via the phone allow the PC to connect if the phone is not working and some don't. Although providing the same cost and convenience advantages, there is still the question: If the one device connecting the other to the network is out, are they both out? And what about mobility? A wireless connection would be more flexible but would it be as reliable? If telephony service availability were of the utmost importance, the decision would be made to provide a traditional hardwired desk phone.

Should call servers or softswitches be centralized or decentralized? Should they be on your premises or managed by a service provider or integrator? What about the application availability impact of different implementations of PSTN gateways? How do application availability concerns impact the number and placement of gateways? And what is the impact on the outsourcing of gateways or use of gateway services? What if your organization presently has a voice VPN? These and other considerations should be made with application availability, and not just access line availability or similar traditional elements in mind. Don't forget that users do not care why they don't have dial tone, and shouldn't.

Table 4.3 shows actual and SLA application availability targets for real-time, near real-time and background/batch applications. The "actual" number is a realistic target and represents a number that can be used to set expectations of management and users for when they read reports. In reality, your performance should be better, as the "actual" values usually represent a lower limit for individual sites or users and the SLA values show aggregate numbers across all individual sites or users. The SLA application availability numbers are, of course, below 100% and represent target values for purposes of SLAs after factoring in outages and other conditions affecting application availability.

 Crafting the "proper" SLA will be addressed in more detail later in this chapter.

Application	Application Availability	Delivery	Delay (One way)	Delay Variation
Real-Time	Actual/SLA	Actual/SLA		
Telephony	99.7/99.99%	>90%/99.0	40-80ms (reg) 100-250ms (global)	1-20ms
Data	99.7/99.99%	100%/99.0	50-100ms (reg) 150-350ms (global)	1-20ms
Video	99.7/99.99%	>95%/99.0	40-80ms (reg) 100-250ms (global)	1-20ms
Near Real-Time				
Audio	99.5/99.97%	>97%/99.5	50-100ms (reg) 150-350ms (global)	1-20ms
Data	99.5/99.97%	100%/99.5	60-150ms (reg) 150-350ms (global)	1-20ms
Video	99.5/99.97%	>95%/99.5	50-100ms (reg) 150-350ms (global)	1-20ms
Background/Batch				
Audio	99.0/99.95%	100%/99.5	80-200ms (reg) 200-500ms (global)	1-20ms
Data	99.0/99.95%	100%/99.5	80-200ms (reg) 200-500ms (global)	1-20ms
Video	99.0/99.95%	100%/99.5	80-200ms (reg) 200-500ms (global)	1-20ms

Table 4.3: Multi-media service characteristics.

You will also note that application availability measurements will actually decline as one goes from real-time, at 99.7% actual and 99.99% SLA, to background. The reason is that in a contention situation, where available resources are below normal and there is competition for bandwidth, buffers and processing capacity within the network, real-time applications should receive highest priority access, near real-time next priority, and background/batch lowest priority.

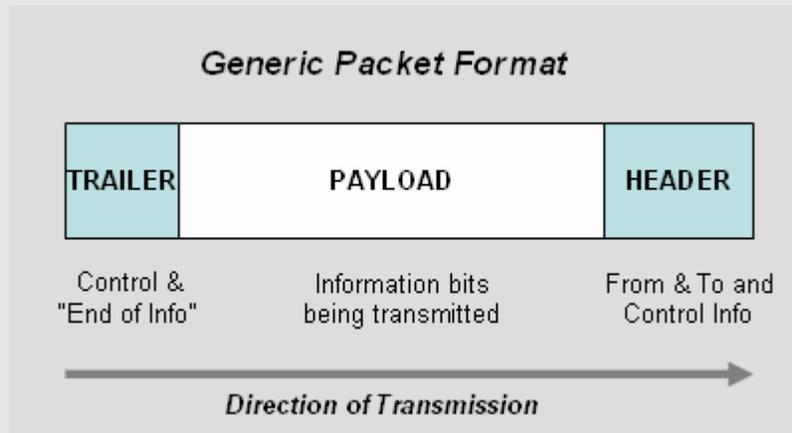
Delivery/Accuracy/Loss

The second consideration of application quality measurements will be a category that can be referred to as delivery, accuracy or loss. If you see the glass as half full, delivery will be your choice of terms as it represents the percentage of packets that were sent from the source that actually make it to the destination. If you see the glass as half empty, your chosen term is loss, as you would think in terms of the percentage of transmitted packets that never make it to the destination. If you really don't think about glasses as being half full or half empty, you might consider this value as the accuracy of the network in delivering packets. Let's take a positive posture and consider the percentage of packets delivered.

As with application availability, you usually see SLA values for telephony and video that are greater than the actual expected minimum values, except in the case of data and all varieties of background and batch services for which expected delivery is 100%. The reason for this is that the later services make use of the Transmission Control Protocol (TCP) as opposed to the telephony and video's use of the User Datagram Protocol (UDP) for sending information across the network. This distinction is important because of the way in which a transmission error is handled in the network. Basically, TCP will retransmit if errors are encountered and UDP will not. The reason why the 100% actual number is less than the SLA value, therefore, is because the SLA value counts the percentage of packets that are sent that arrive on the first attempt. The 100% value may require additional attempts, up to some retry limit, but the 100% number is used for planning purposes: 100% of sent packets arrive at the destination, and the retransmission time for any packets that need to be sent more than once due to errors while traversing the network is adjusted for in the overall delay time. It is noteworthy and important that IPT uses UDP for voice transmission and in most implementations for signaling as well; therefore, UDP never retransmits voice or signaling packets. If they are lost, they are lost.

Error Handling in Packet Networks

Packet networks are not perfect. There is a variety of reasons why packets may be damaged—that is to say, have the value of their 1 and 0 bits changed—or for packets to be discarded and never arrive at their intended destinations. This is as true of data packets as it is of packets carrying voice or video, but the way of handling the problem is a bit different, depending upon the protocol being used to send and receive the information.



Bad Packets

To assure accurate transmission of the packets, the sending system does a special calculation on the bits in the packet. The calculation is usually referred to as a Cyclical Redundancy Check (CRC), block check, Frame Check Sequence (FCS) or Checksum. These are all variations on the same theme used in different transmission protocols. The results of the special calculation are either placed at the end of the packet—in the packet's trailer following the payload bits being transmitted—or in the control bits at the front of the packet called the header. In either case, they are intended as a special part of the message that allows the receiving system to determine whether any 1 bits were changed to 0 bits, or vice versa, which could drastically change the meaning of the information being transmitted. When the packet arrives at its destination—or in many cases at an intermediate relaying system, such as a router—a calculation identical to the original calculation is performed again. The new result is compared with the result of the original calculation stored in the packet by the sender. In the case that the new result and the original are the same, it is assumed that no bits were changed and the packet is either forwarded or used by the system. In the case that the new calculation does not match the original calculation, one of three outcomes is possible. In some cases, the error checking calculation is so sophisticated that it is possible for the receiving system to determine which bit or bits were changed, up to two bits, and fix them. In other cases, the receiving system requests the sending system to resend the packet, up to some number of retransmission attempts as identified by the retry limit. In still other cases, the receiving system discards the inaccurate packet and lets some other part of the network worry about the missing information.

Lost or Missing Packets

It is also possible that packets may be discarded or get lost along the way for other reasons. If, for instance, the intermediate routers between the source and destination become overburdened they might intentionally make a decision to discard packets. This situation, usually referred to as network congestion, is not uncommon at times even in well-architected and managed networks because building a network to handle every possible demand is cost prohibitive. In the case of lost packets, some protocols will provide a sequence number for transmitted packets so that receiving systems can know whether packets are missing. Some systems, however, do not. In the case that a sequence number is missing, packets can be recognized and a request can be made that the missing packets be retransmitted, up to a specified retry limit.

Which Protocols Do What?

The Internet Protocol (IP) is a protocol common to all applications in IP networks. IP is always used in conjunction with a higher-layer protocol, which is virtually always either TCP or UDP. IP has an error-checking mechanism, but only on its header information, and has no sequence number capability. IP relies on the higher-layer protocols to provide sequencing, if needed. TCP is a richer, more robust protocol and has both an error-checking mechanism on its header and a sequence number mechanism, with the tradeoff of more processing and higher bandwidth utilization than UDP. UDP has neither. For these reasons, TCP is used when information accuracy is vital, such as paychecks and browser info, and UDP is used for IPT and similar applications.

Delay

Delay figures are cumulative numbers that include all factors that can impact the time it takes a packet to travel from sender to receiver. In addition to the delay due to the network distance of a transmission over physical links from sender to receiver, there are added delays for the time that packets spend in intermediary relaying systems such as switches and routers, known as nodal delay. Delay is usually a constant factor for any given communication. Reasonable ranges of values, for planning purposes, are shown in Table 4.3. Both regional (same continent) and global (different continent) values are shown. The component of delay due to distance is usually fixed or so very close to fixed for most connections that it can be considered fixed. In addition to planning, these numbers are equally valid for SLA purposes, though some service providers may show more accurate numbers for different continent pairs.

Delay Variation or Jitter

Delay can also have a variable element due to a variety of factors, such as how long the packets spend processing and queuing in intermediate systems on their way from the source to the destination. As has been noted earlier, for the 100% delivery applications, delay variation might also be impacted by the one or more retransmissions needed to achieve the 100% delivery. Delay variation is usually absorbed for data applications and is usually unnoticeable to the end user, while delay variation in voice applications can be very noticeable and contributes to user complaints and dissatisfaction. Delay variation is shown in milliseconds and is rarely less than 1 or more than 20. These values may be used for network design as well as for SLAs.

Fixed delay occurs in traditional telephony systems and can be quite noticeable in transoceanic cable or satellite systems. Delay variation also occurs in traditional telephony but is usually so negligible due to the high-speed circuit switches that provide the foundation for older networks that delay variation due to intermediate switching systems can usually be ignored.

In all the real-time and near real-time applications that Table 4.3 shows, it is assumed that there is a human watching or listening for the information to arrive, and therefore, in terms of delay, near real-time is in a close second place behind real-time. In real-time applications, end-to-end delay must be very short and with as little variation as possible to prevent disruption to the timing and flow of the human conversation. It is interesting to note, almost contrary to what one may think, that very long fixed delays, even exceeding a second and a half, have a less negative impact on MOS values than do much shorter but inconsistent delays. Variable delay is disruptive to the pace of the conversation but once the speakers adjust to a fixed delay, such as that experienced over a satellite link, they are much more comfortable speaking. Near real-time applications, however, can tolerate greater delay, though not the excessive delays possible in background and batch applications. In the case of near real-time applications, buffers may be used to optimize throughput without introducing delay that lowers the QoS perceived by the human user. This is not to say that buffering information for near real-time applications is always an acceptable solution.

Differentiation, Prioritization and Queue Management

The key to optimizing all applications is to provide the proper mix of resources and handling within the underlying IP network, and the first step in this optimization is proper differentiation. By knowing the specific needs of the specific type of traffic contained in each packet, appropriate and different handling can be applied, increasing the likelihood that all packets will be handled optimally.

In general, voice applications are less impacted by loss of information than data applications. Studies have shown that as much as 10% of voice samples can be discarded randomly from a voice conversation with little or no loss of understanding, although a change of a single bit in a data transmission can drastically alter the meaning. Due to differentiation when selective discarding of packets in an overload situation is required, voice packets can be selected and discarded, instead of data packets, up to some limit near 10%. Data applications, in contrast, are less affected by delay and delay variation than voice packets. For this reason, in a situation in which some packets will be delayed and some will be forwarded, differentiation allows voice packets to be prioritized and sent with lower delay, while data packets may be delayed to offset the handling of the voice packets.

There are several mechanisms that can be used in networks to differentiate one type of packet and its preferred treatment, but the mechanism common to most applications is provided by the one protocol common to all applications in IP networks, IP. Figure 4.3 shows the format of the special control bits present in the header portion of the IP packet. The area of most interest to the current discussion is the Type of Service (ToS) Byte, pronounced “toss byte” and more recently described as the Differentiated Services (DiffServ) field. The network may also be configured to use Explicit Congestion Notification (ECN), which uses two bits in the DiffServ field in the IP header. The ToS bits provide four types of information: the precedence or general priority of the packet, and within the precedence, normal or low delay, normal or high throughput, and normal or high reliability. Precedence has eight levels from “routine,” the lowest priority, to “network control,” the highest priority, and six classifications in between. By setting these bits, an application such as FTP or IP telephony, can signal the network as to what special treatment the packet requires. In network design, it is imperative to assure that the ToS bits are set properly during modeling and simulation exercises and to be sure that desired values find their way into the final network implementation. It is also important to note that although ToS bits are defined relatively clearly in “standards,” there are different interpretations of ToS bits from manufacturer to manufacturer and even from one product line to another, even though the IEEE 802.1p and 802.1q standards clearly and unambiguously describe how to map the Layer 3 DiffServ Code Points into Layer 2 Virtual LAN labels, and vice versa. There are, therefore, ramifications for networks in which packets might encounter different interpretations on their path from end to end.

+	Bits 0–3	4–7	8–15	16–18	19–31
0	Version	Header length	Type of Service (now DiffServ and ECN)	Total Length	
32	Identification			Flags	Fragment Offset
64	Time to Live	Protocol		Header Checksum	
96	Source Address				
128	Destination Address				
160	Options				
160 or 192+	Data				

Figure 4.3: IP header structure.

Prioritization in Perspective

In the mid 1990s, a major global telecommunications carrier implemented a new multi-media VPN network for a large multi-national customer. The carrier had sold the customer on using their multi-media VPN service as opposed to either building it themselves or using a competitor's network, based upon the carrier's superior ATM platform and sophisticated prioritization mechanisms. The customer paid a premium for certain service classes. Although the carrier actually offered several gradations of class of service, the new customer settled on three, which they referred to as First, Business and Coach using an airline analogy that is fairly commonplace in networking. Each class of service had specific characteristics associated with it, and each one had a different cost per megabit of information transferred. The user organization accepted this pricing arrangement reluctantly after being sold on the benefits of class of service and the positive impact it would have on application performance and the resulting end-user productivity.

After the network was installed and operating, the customer decided to test the differentiation characteristics to ensure that they were getting what they paid for. They set the appropriate parameters on a packet generator and tried the First Class traffic. Indeed, they did get the promised performance and were very pleased with the outcome of the test. They then set the parameters on the packet generator to simulate Business Class service. The customer was a bit perplexed, and not particularly happy, when they found that the traffic got the same high performance characteristics as the First Class traffic. One might think that they should have been happy to get the same performance for a lower price, but they were unhappy because they felt that the reason for paying the higher price was bogus. They repeated the test for the Coach Class traffic and got the same results.

What were they to do? There were some among the group that proposed sending all traffic as Coach Class, which, as their tests had shown, would give them performance equal to that of First Class but at a Coach Class price. There were others who proposed sending all traffic as Business Class, in case there was some trick they had missed. At least they would not completely ruin the performance but could save some money. The problem was that marking all the traffic as any single class, regardless of the treatment of that class, defeated the purpose of differentiation and made prioritization unnecessary. In the end, they called the carrier, explained their test, and asked the carrier to explain the lack of difference between the classes. The carrier explained that their network backbone was actually over-engineered and that the differentiation, although always present, and the associated prioritization, while always applied, would only become evident if there were a competition for resources that would only occur during a network outage in the backbone.

The customer felt, in effect, that what they were paying for was an insurance policy in the unlikely event of a service-affecting outage in the carrier's network. And, they really were correct. The customer was able to renegotiate their contract with a much smaller price difference between First, Business and Coach classes and is now very happy with their multi-media VPN. It is also noteworthy that the customer has never had to make any "claims" against their "insurance policy," but now that they understand the issues and the cost of the "premiums" is lower, the customer is very pleased with its situation, especially as they migrate more IP-based voice onto their network.

Voice QoS

QoS is a measurement of the treatment of the packets traversing a network and includes delay, delay variation, packet loss and network availability. Classification and marking of packets, in an attempt to assure certain promised levels of quality, is done at or near the network edge and are controlled by Class of Service rules. The calculation of instantaneous QoS in the network, while important, provides only one measure of the user's quality of experience. With its broader definition, QoE is rapidly becoming the more important consideration for voice services.

QoE

QoE is defined as a telephony system user's perception of the quality of the communication being experienced from both the technical QoS rating and other aspects of the call. QoE takes into account the cumulative effect of network characteristics on the human speech being transmitted and received. QoE was first designated as a methodology for assessing the satisfaction of users of network video services in the mid 1990s and has been applied to voice since the late 1990s. QoE is a result of a variety of factors and is best measured by way of a human MOS, which represents the opinion of a panel of human judges, on a five-point scale with five being the highest. Due to the loss in translating analog speech waves into digital values and back, in effect, 5.0 MOS has never been achieved. The best Pulse-code modulation (PCM) scores are in the 4.2 to 4.4 range, depending upon network conditions, but are consistently 4.4 in ideal lab conditions.

In this early phase in the network design, it is often desirable to convene a panel of judges from your user community and perform MOS panels to get user buy-in and to assist in the process of properly setting expectations. This approach is an important tool for setting proper user expectations and assuring that the system will meet user needs but becomes impractical after implementation. After implementation, and to a great extent during design and benchmarking, automated tools will be used to simulate certain types of calls, estimate what the human MOS would be and help isolate the factors that are impacting voice and call quality. Even using PCM in the new IP-based system, the voice will sound different, but the idea is that "different is not bad and, in fact, a properly tuned IP voice system can actually be better than old analog and digital voice. Care should be taken in the selection of panelists to choose from the broadest range of job functions in order to optimize the voice coding selection for your specific network.

However, because it is not possible to empanel human judges whenever an opinion on voice quality in the actual operating network is desirable, a number of measurements and derivative calculations have been developed (as described later in this chapter) that will form the basis not only for design but also for the ongoing network measurement and reporting and will be specified in the SLA. This function must be automated using tools developed specifically for the monitoring and reporting of voice and call quality issues. Tools such as those used to monitor router or switch performance, packet loss, delay and delay variation, and system availability only provide a very general idea of the effect of network activities on voice and call quality.

QoE vs. QoS

Although QoS is an important element for technical issues such as SLA compliance, meeting users' QoE expectations is crucial to a successful IP telephony implementation. QoE can be affected by many factors:

- Human factors
- Voice encoding/compression
- Network issues
- Service/feature support

Human Factors

The biggest factor impacting QoE is user expectations. The users' prior experience and use of telephony will play a large part. Are they used to a highly reliable traditional desk phone and are being asked to use an IP system that occasionally drops calls or a wireless or cordless phone that does not work at certain locations in the building? Is there a critical path or lifeline application such as 9-1-1 emergency calling or do they use the phone for revenue generation, such as an outbound sales call center? People who use the phone for casual conversation or other less stress-inducing purposes are likely to be less harsh critics of any new system.

Speaker and listener age and gender also impact QoE. Age has been shown, for instance, to impact MOS. Older listeners tend to give higher MOS while younger listeners' MOS values tend to be lower, which is almost counter-intuitive. One would think that those who have been listening to phone systems longer would be tougher critics than younger folks, but research shows that older listeners have more highly developed linguistic ability and can get more information from less communication; therefore, they are more satisfied with the same signal than younger people.

Gender also impacts MOS. Males tend to give higher MOS and females give generally lower MOS. The primary reason for this seems to be that higher frequency sounds are more important to female communication, and these signals are often lost or distorted in the analog- to-digital conversion process.

Speaker and listener familiarity also plays a role in QoE. The more familiarity between the speakers, such as a mom and child, the lower the MOS; the less familiarity, such as a first time caller to a call center, the higher the MOS.

Native language of the speaker is also an important consideration. Although PCM-based systems get fairly consistent MOS regardless of the native language of the speaker, due to PCM's analog-to-digital conversion process, systems based on Code Excited Linear Predictive (CELP) are often optimized and get lower MOS values if not optimized to the unique sounds inherent to the native language of speaker. This is because CELP-based systems are voice-optimized and use code books containing sounds that are matched to codes for transmission. In order for a sound to be transmitted, it must have a corresponding code.

These considerations are often very difficult to factor into a design, but are critical to user acceptance and satisfaction and must, at least, be considered. Use of PCM for voice coding and reducing the number of recoding at gateways and other points in the network will improve the MOS and the user's perceived QoE.

Voice Encoding/Compression

One of the underlying assumptions going into this network design effort was that in situations of bandwidth versus voice quality, voice quality will prevail. For this reason, IPT systems should be implemented using the G.711 coding scheme, known to all as PCM. PCM is the standard for voice coding worldwide in traditional telephony networks and although it has not been implemented quite as widely in IPT networks, there are a number of desirable characteristics to recommend it for the job (see Table 4.4).

ITU-T Standard	Coding Scheme	Bit Rate	Sample Size (Bits)	Encoding Delay (Time)	Mean Opinion Score (MOS)
G.711	PCM	64K CBR	8 bits	<1 ms	4.4
G.726	ADPCM	32/24/16K CBR	4/3/2 bits	1 ms	4.2, 3.8, 3.2
G.728	LC-CELP	16K VBR	40 bits	2 ms	4.2
G.729	CS-ACELP	8K VBR	80 bits	15 ms	4.2
G.723.1	ACELP	5.3K VBR	160 bits	37.5 ms	3.5

Table 4.4: Voice coding options.

First and foremost, the CELP systems are optimized for voice transmission and don't perform as well across a wide range of demands. Use of PCM in IPT systems will also sound more natural to users and result in less user resistance. Another reason is that when voice transits traditional gateways, as well as other IPT systems, it will not need to be recoded but rather can stay in native PCM; therefore, it will be less subject to transcoding loss.

The problem in using G.711, however, is with bandwidth. As Table 4.5 shows, with the addition of packet overhead, G.711 PCM well exceeds the 64K channel rate of traditional telephony. In the worst case, G.729A at 50 packets per second over ATM uses only 43kbps, only 40% of the bandwidth of G.729A over ATM, for example with a MOS of 4.2 versus a 4.4 score for PCM. What is not taken into account in this comparison, however, is degradation of G.729A over multiple encodings through multiple gateways, which can be substantial in large intra-national enterprise networks.

Bandwidth Consumption	802.1Q Ethernet	PPP	MLP	Frame-Relay	ATM
G.711 at 50pps	93kbps	84kbps	86kbps	84kbps	106kbps
G.711 at 33pps	83kbps	77kbps	78kbps	77kbps	84kbps
G.729A at 50pps	37kbps	28kbps	30kbps	28kbps	43kbps
G.729A at 33pps	27kbps	21kbps	22kbps	21kbps	28kbps

Table 4.5: Sample bandwidth consumption of PCM and ELP codecs.

Voice Compression

Another consideration is the use of voice compression. It has been shown in actual networks that voice compression has minimal benefit in packet-based voice systems (IP, Frame Relay, ATM). Voice compression algorithms usually add nodal processing delay that is unacceptable in very long distance calls when delay budgets are already near their limits. This is not to say, however, that packet voice systems should not use *protocol* compression, such as Van Jacobson header compression on IP or PPP packets. They should, but for compression to be effective, it must occur on the digital form of the voice after packetization.

Data Compression vs. Voice Compression

Voice compression and data compression vary in some important ways. Voice “compression” is usually performed on audio in its analog form and often effectively ‘speeds up’ the voice at one end and slows it down at the other end. This type of voice compression is unsuitable for packet voice applications. Compression of voice in its digital form is also unsuitable for packet voice, though it is seen in many implementations, usually as an option. The reason why it is unsuitable is that it works on voice after it is in its digital form, and there are very few repetitive patterns that can be compressed. What does work with voice over packet, especially on bandwidth-limited links, is traditional data compression that works on the underlying protocols that transport the voice content. One example is called Van Jacobson Compression. Van Jacobson compression can be applied to TCP, IP, PPP and other protocol headers and is described in RFC 1144. For example, Van Jacobson compression can reduce the normal 40-byte IP packet headers down to 3 to 4 bytes for the average case. It does so by saving the state of IP connections at both ends of a link and only sending the differences in the header fields.

Silence Suppression

Silence suppression saves bandwidth, and, when properly tuned for a given network, will yield any where from 2:1 to 4:1 savings. That is to say that with silence suppression—which removes silence at the sending end of the voice connection and reinserts it at the receiving end—two to four more calls can be made using the same bandwidth required for one call without silence suppression. However, the silence suppression algorithm may negatively impact voice quality if not properly tuned.

Echo Cancellation

Another consideration, well understood in the traditional voice network but often not even considered in IPT networks, is echo cancellation. Regardless of what one may hear, echo cancellation is mandatory in IPT systems. Echo cancellation is not needed in IP phone to IP phone configurations over an all-IP backbone with no gateways, but this configuration is highly unlikely. Echo cancellation should occur as close to the endpoints of the call as possible, preferably in the IP handset, softphone, analog terminal adapter and gateway where outbound IP calls jump onto the PSTN and inbound calls jump onto the IP network. It would be wise to ensure that all call servers and gateways that interconnect telephony end systems, be they Analog Terminal Adapters or IP phones, that do not provide their own echo cancellation can, at least optionally, provide echo cancellation.

Tandem Hops/Multiple Encoding

Special attention should be paid in the design to ensure that voice is not recoded more often than needed. Ideally, communication occurs only between IP phones but, this will be a long time coming. In the meantime, the number of gateways should be reduced as should the number of times voice is recoded. Most IPT implementations should consider using PCM on all devices for voice coding and reducing the number of gateways (except in situations in which bandwidth is extremely expensive or delays are very long). This will result in the best possible voice and call quality. This is an issue of importance in both the carrier/service provider environment and in enterprise implementations and has a profound impact on overall perceived quality of communication.

Network Issues

Network issues have already been covered earlier in this guide, so let's simply review the important aspects of network issues that impact QoE and which, therefore, must be considered in your design effort:

- Delay budget is important and includes all sources of fixed delay. Delay less than 150ms one-way is ideal; >300ms will impact the users' QoE. Delay should be specified in the SLA.
- Delay variation (aka jitter) is a bigger QoE problem than delay. Delay is fixed and easier for the listener to adjust to. Delay variation is difficult to adjust for and causes anxiety and stress on the part of listener. Jitter less than 20ms one-way is ideal and should be specified in the SLA.
- Packet loss >1.5 to 3% is where many human listeners can begin to perceive a degradation in voice quality even though the human ear can adjust to packet loss up to about 10%, or more in certain circumstances, such as where the speakers are known to each other or there is a known call context. Packet loss impact can be reduced by shortened voice samples and fewer samples per RTP packet, which is something that the organization responsible for programming the IP phones or adapters has control over. The end-user or end-user organization may or may not exert control, but this should also be addressed in the SLA.
- Network availability directly impacts QoE. No dial tone is a BIG problem and causes loss of confidence in the system. This needs to be a key component of any SLA.

Service/Feature Support

In migrating from an old system to a new one, it is critical to ensure that the new system performs all needed functions of the system being replaced and, to the extent possible, that users are comfortable with and accept any changes to work procedures required by the new system. This may seem like an obvious and simple point, but user reluctance to change is one of the biggest reasons for the failure of IPT systems.

A number of valuable considerations can be made during the network design phase. For instance, is Voice Band Data (VBD), such as analog modems and fax machines, to be supported in the new network? If so, it will be necessary to test modems at all speeds and verify all fax groups work without down-speeding. As simplistic as it may sound, it is necessary to test Dual Tone Multi-Frequency (DTMF—aka Touch Tone) systems. DTMF is critical for many applications and not all voice coding algorithms support all digits. An inventory of all calling features, as described in Chapter 3, must be performed prior to the network design phase to ensure that all-needed calling features are supported. Supervisory tones and intervention, signals that allow the network to perform special functions, such as allowing a person to break out of a voicemail system or make another credit card call, must be supported as well as all call forwarding, routing and blocking features and any other calling features identified in the inventory. Billing and call accounting must not be ignored, abandoned or otherwise neglected during the network design phase. Many organizations use call accounting and billing for a variety of reasons, such as customer billing, verification, tracking of calls, and allocation of internal costs.

 This topic is covered in more depth later in this chapter.

Voice Metrics and Measurements

There are two basic types of voice quality testing, non-intrusive, often called passive, and intrusive, often called active. Non-intrusive does not require special test calls and is based on actual conversations. Intrusive voice quality testing requires a special test call and is based on a comparison of source signal with signal after transmission.

The most desirable situation is to assemble a panel of judges and let them listen to a call and give the call a score from 1 to 5, with 5 being the best. After initial selection of products and technologies that will be put to use in the network, however, this approach is impractical; therefore, a number of calculated values are generally used by automated testing tools, including MOS, J-MOS, PSQM/PSQM+, PESQ/PESQ-LQ and R factor. Non-intrusive methods include MOS, derived MOS, E-model and R value/R factor. Intrusive methods include Perceptual Analysis Measurement System (PAMS), Perceptual Speech Quality Measure (PSQM) and Perceptual Evaluation of Speech Quality (PESQ). The following brief review will allow you to understand their relative importance and value in the network design process.

MOS

MOS is adopted from traditional telephony. The traditional “opinion” of voice quality in the U.S. was from a panel of 16 human judges from Central Illinois consisting of 8 men and 8 women who judged voice quality on a scale from 1 (lowest) to 5 (highest). This type of human-originated MOS is useful for human user QoE analysis but is useless for SLA development and network monitoring. Today, MOS is often calculated by a management tool using the QoS indicators for loss, delay and jitter.

PAMS

PAMS was developed by British Telecom and does not replace MOS, though it was a good first attempt at an automated MOS estimate. PAMS compares original analog voice waves with reproduced/transmitted speech using a complex weighting method intended to take into account characteristics important to the human ear. The scale is from 0 to 6.5, with 0 being “perfect” (no difference between waves). PAMS values estimate MOS +/- 10 to 20% and therefore do not provide enough accuracy for use in IPT networks though PAMS values are excellent for benchmarks and comparisons when both comparison values are calculated using PAMS.

PSQM/PSQM+

PSQM is defined in ITU-T Recommendation P.861. Like PAMS, PSQM and PSQM+ do not replace MOS, though PSQM more closely estimates MOS than PAMS at +/- 10%. Like PAMS, PSQM and PSQM+ are excellent for benchmarks and comparisons and represent improvements on the PAMS algorithm. Like PAMS, the scale is from 0 to 6.5, with 0 being “perfect” (no difference between waves).

PESQ/PESQ-LQ

PESQ is described in ITU-T Recommendation P.862. Like its predecessors, it does not replace MOS (+/- 10%), is excellent for benchmarks and comparisons, and is an improvement on the PSQM algorithm with the same scale from 0 to 6.5, with 0 being “perfect” (no difference between waves).

E-model/Design Tool

E-model, described in the ITU G.107 standard, is a predictive tool that is excellent for use in modeling voice quality in a packet network. The use of design tools incorporating the E-model algorithm combined with network performance data gathered from the live, base network is strongly recommended. E-model predicts average voice quality using a sophisticated and mature mathematical model that takes into account delay, delay variation (jitter), packet loss and codec performance.

The result of the E-model calculation is an R factor/R value that predicts voice quality on a range from 0 to 100, where 0 indicates lowest voice quality and 100 indicates best quality. R factor E-model scores are most useful when based upon measured, rather than hypothetical, parameters. Analysis should, when possible, include Real-Time Protocol (RTP) streams, source and destination addresses, sequence numbers and jitter profile in order to most accurately predict the impact of the IP bearer on MOS.

Figure 4.4 shows the relationship between the MOS value and the R value resulting from the E-model calculation with a description of the user satisfaction level that they represent. A G.107 default value of 94 corresponds with the MOS value of 4.4, which is a target value for many network design efforts.



Figure 4.4: R Value compared with MOS.

Call Integrity/Privacy/Security

Migrating voice calls from traditional telephony networks that are inherently secure, particularly within the backbone of the network, to shared IP networks that are notorious for their insecurity requires a new level of diligence in call integrity, privacy and security not previously seen in telephony. The idea that IPT is just another application is one that has been taken to heart by hackers and cyber-criminals and the archives are beginning to fill with hacker exploits against VoIP, IPT and VoP. The key points to keep in mind during the design phase are that IP phones are not traditional phones with a keypad and a ringer. They are special-purpose computers in a telephone form. They have processors, memory and protocols and can run programs— increasingly programs written in Java or formatted using eXtensible Markup Language (XML) and similar languages that are familiar enough for the hacker to compromise and utilize. In all phases of the network design phase, physical and human security must be considered and, like any IP networking project, the IPT services and the systems that support them should be subject to periodic security audits and penetration tests.

Impact on Business Processes

In the best possible case, the migration to an IP-based voice platform will have an immediate and positive impact on operations, costs may be lower, and users will be ecstatic with the new applications and Unified Communications tools provided by their wonderful employer. In fact, the new system will give them bragging rights beyond their wildest dreams. IP phones on their desk and calls over WiFi phones that attach to their hip or fit in their purse. It can't get much better than this! But, alas, the truth is probably a bit more mundane: this, after all, is dial tone we are talking about. Your reality will, of course, probably be somewhere between these two extremes, and it will be just as important to capture positive impacts and socialize them within the organization as it will to identify negatives and quash them.

Standardizing Positive Impacts

Standardizing positive impacts can be accomplished through two channels: formal training and a grass-roots effort. Formal training can be via classroom, webinar or written and can be conducted all at once or phased over time with practice in between. The grass-roots approach can be as simple as featuring employees who have put the new system to good, company-approved uses to demonstrate their skills at informal “lunch and learn” sessions or in the context of other meetings.

Many organizations really do take the “it’s just dial tone, what could be so complicated?” approach and neglect training, but they do so at their own peril. Absent official training, users will develop their own skills and habits and may or may not make a positive contribution to getting the most productivity out of the new investment. “Over-the-cube wall” help should not be a fall-back to official training.

Avoiding Negative Impacts

Negative impacts can be avoided by having a simple, clear system for reporting problems and a feedback mechanism to let the users know that the problem is being investigated or has been resolved. Most organizations choose to use their existing Help Desk and trouble reporting and tracking system. Any organization considering this approach should be sure that their Help Desk personnel are properly trained and that their systems are capable of handling the volume and issues.

Translating Needs to SLAs

At this point, it does not matter whether the organization’s new IPT services will be implemented in-house, outsourced or some combination. In any event, an SLA will be needed. The SLA will become, effectively, the voice user’s Bill of Rights and will represent what the user has been promised, how compliance is measured, and any penalties that might exist for non-compliance. In a growing number of cases, users are also held accountable for their actions in terms of number and types of calls, destination, use of gateways into traditional telephony systems and other actions, and are monitored under the agreement.

The “Proper” SLA

There are many different types of SLAs in the wide world of networking and telephony, and many of them are “proper.” In order to be a “proper” SLA, the SLA must not only spell out the requirements of the carrier or service provider (or internal IT organization) and the attendant penalties but also specify similar requirements and penalties for the customer organization (or business unit). It is also common to specify certain benefits or bonuses for exceeding the requirements set forth in the SLA. These provisions acknowledge the fact that in certain cases resources are limited and encourage the carrier, service provider or IT organization to apply those limited resources to the problems of a specific customer or business unit.

An SLA is often an addendum to a service contract or a contract in its own right, and this should be considered from the outset. Because the document may at some time in its future be evaluated and interpreted by a judge or arbitrator, attorneys’ technical terms, acronyms and jargon must be clearly defined or eliminated entirely where doing so does not hurt the meaning of the document. Nothing should be left to interpretation, diagrams should be provided, and, where formulas are used, examples should be provided as to their proper application. Terminology should be clear, crisp and unambiguous, and nothing should be left to chance. If it is understood, for instance, calculations of service availability assume a maintenance window from midnight Saturday to 2:00 am Sunday, that understanding should be explicitly stated and an example of the service availability calculation should be provided.

Metrics

The proper SLA may include any number of items, but at a minimum should include specific details for application availability, packet delivery, delay and delay variation. For numerous guidelines on specific measurement details, refer back to Table 4.3. Chapter 3 also provided some guidelines for other possible aspects for a proper SLA, including throughput, call setup and teardown boundaries, and MOS values.

Availability should include any distinctions for on-net, off-net, wireless access or other variations as well as allowances for time zone differences, maintenance windows and other periods of unavailability that are known in advance and should not be included in down-time calculations. Availability that differs per application should also be noted, as well whether availability is for a single user, class of user, department, geographic region or network-wide. Where possible, the availability should be designated for a site too. In very large networks, it is possible to have a single user or site down for a very long time while hundreds of sites or individuals are up and running.

Packet delivery should include any special considerations for retransmission protocols, variations in requirements by application or class of user, geographic area, wireless vs. wireline and so on. Special consideration may also be given to packet delivery during certain network outages or other special circumstances.

Delay and delay variation targets in SLAs can be very general but should really be at least at a continental and inter-continental level for all appropriate continent pairs to which service will be provided. This type of tiered SLA for delay and delay variation provides more accuracy, better planning and a clearer understanding of voice impairments when they are reported as well as a better understanding for remedies under the SLA in case of non-compliance.

Other items that merit consideration for inclusion in the proper SLA are response times and Mean Time To Repair (MTTR) commitments—not just “mean time to respond”—for systems and users under contract and those not under contract and for those within certain distances of service centers or service personnel. Consideration should also be given for customer-provided on-site stocking of spare components or use of hot-standby components to lower costs, increase uptime, and alleviate some of the pressure on service staff for the health of the network.

Time windows for upgrading existing service and for adding new sites to the network should also be spelled out clearly in the SLA. In all cases, measurements and penalties should be clearly stated and examples shown so that a “Did you comply with section...”-type questions can be answered with an unambiguous “yes” or “no” and the veracity of the answer can be known.

Measurement

Measurement of all compliance metrics, to the extent possible, should be automated and have thresholds for automatic exception reporting that match those found in the actual SLA. Exception reporting is highly desirable as it cuts down on the total number of alarms and alerts and allows you to focus only on those issues that represent non-compliance with the SLA. Care should be taken to align reporting with local time zones and units of measure, as well, to allow the fastest understanding of SLA compliance reports and most intuitive and localized approach to problem resolution. Where possible, SLA penalties should also be automatically calculated and credits applied, as appropriate.

Measurement, specifically for SLA compliance, is a very important area to apply IPT management tools. The reasons for this are numerous. From a business standpoint, costs are reduced and consistently can be assured by using an automated tool that is aligned to the specific, agreed-upon metrics of the SLA. Any manual approach or one based on lower-layer network statistics from which guesstimates of the impact on voice and call quality are calculated lack efficiency and consistency. From a technical standpoint, the sheer volume, range and interaction of measurements needed to develop meaningful reporting and to highlight areas for problem isolation and troubleshooting are daunting and could easily overwhelm any semi-automated or manual system. Combining voice-aware or multimedia-aware IPT management solutions with an exception reporting approach and trend analysis to anticipate problems and eliminate them before they occur, as opposed to simply trying to mitigate them when they do occur, should be the objective of every IPT network design and implementation. It is far easier to design the proper IPT management into the system at the beginning than it is to go back and retrofit a solution later.

Establishing Feedback Loops and Review Cycles

Feedback loops allow operational data to be gathered from the live network and compared with SLA requirements. Adjustments then can be made to the network. Early in the roll-out of a network, after final system test and acceptance, review cycles should be performed frequently, possibly daily, and a conference call or meeting involving operational personnel and management should be held. After the network settles down, weekly then monthly meetings should be conducted. The review meetings are a part of the ongoing network operations and optimization cycle. Review meetings should be held at least monthly to ensure that the network is operating as desired.

Validation of Design

Many organizations prefer to choose a network design and validation tool in the early stages of a project and to use the input forms and screens of the tool to provide guidance as to the information that needs to be collected. Others prefer to determine the important characteristics of their network and then choose a tool that fits their needs. This guide takes the approach that the choice of the tool should be dictated by the need; therefore, it is at this point in the process to select the design and simulation tools that will be used in this design phase.

A report card similar to the one provided in Chapter 3 for assessment tools is provided with this chapter to assist in the choice of network design and validation tools. Very often, the same provider will have tools for assessment and for design and validation that use the same inputs and produce results with little additional effort. The version of the Modeling, Measurement, Monitoring and Management Tools Report Card (4MTRC) for network design and validation will assist you in the selection process of these very important tools. The report card is a set of five Microsoft Excel spreadsheets. There are five identical report card spreadsheets that may be used to evaluate up to five separate 4M tools. The sixth spreadsheet allows you to compare the tools side by side and support an intelligent selection process. The spreadsheets are locked to avoid inadvertent changes to cells containing formulas and labels, but a password is provided so that the spreadsheets may be modified by an organization for its own use. Those knowledgeable in the inner workings of Excel can take a look and see that there is nothing particularly clever about the coding. The real value is the knowledge that is embedded in the categories that were chosen for comparison as well as the weighting factors applied to the different categories. For users with limited resources, these two factors—the comparison list and associated weighting—will be an invaluable source of assistance in the selection process; for the more resource-rich organization, the spreadsheets will provide a convenient starting point for building customized selection report cards for your organization.

The first project will be to gather together all the Excel spreadsheets, Microsoft Word documents, Microsoft PowerPoint presentations, handwritten notes, and the appropriate sections of the SLA that were generated during the early project specification stages and put them into the network design tool. At this point, the design team can create an operational baseline, validate the baseline against real-world observations, then commence the process of network tuning. The cycle will include modifications to the design and service order, the SLA and, possibly, management and user expectations. It should also be possible to begin tracking costs in real-time as network elements are acquired.

Network Tuning

After establishing the baselines and validating the model or simulation results against observations of the real network results, it is possible to perform network tuning, which is an optimization process in which specific characteristics of the network are systematically modified to minimize or maximize certain characteristics. The most important measurement in the network tuning phase will be the E-model R value. Every effort should be made to maximize the R value while keeping cost in balance, both in terms of network costs and resource utilization. Before beginning any serious modeling effort, engineers should familiarize themselves with the tools and just “play around” to gain familiarity, posing hypothetical questions and answering them, and modifying parameters at will. After the actual network tuning exercise begins, each step should be documented and only one change should be made to the model at a time.

Test Bed Architecture

The full network, including all other IP traffic and systems, should be modeled during the simulation phase. If it is impractical to do so due to the gargantuan size of your network or other restrictions, at least model all the elements of a sub-network of your network. After the modeling or simulation on the computer, it will be necessary to build a small test bed in a controlled laboratory and, after appropriate testing, to move the test to a real environment.

Modeling vs. Simulation

Modeling and simulation are closely aligned and the terms are very often used interchangeably, but while the objectives are very similar and the results may look the same on the surface, the processes underlying modeling and simulation are very different. Modeling and simulation are both attempts to predict performance of network-based services given some set of inputs. Inputs can range from wild, “seat of the pants” guesses to sophisticated traffic studies. The fundamental difference between modeling and simulation is that modeling provides a snapshot at a moment in time while simulation is a process over time with changes to network traffic patterns, call volumes, Class of Service (CoS) requirements, and QoS and QoE results. The budget-minded very often use an Excel spreadsheet to develop reasonably accurate snapshots, are emboldened by their results, and use their informal models for actual network design, validation and optimization. This approach appears savvy and does save costs in the short term, but even small variances, which are not even really mistakes or miscalculations, can magnify in the real network and can be very costly, often costing many times the price of a good simulation tool, training and some consulting to get the project kick-started.

There are several issues that need to be considered and tested based upon your environment. Considerations and parameters that are common to most IPT projects include:

- Network impairments
- Timing/synchronization issues
- Buffer management
- Static/dynamic/adaptive jitter buffers
- Broadband bandwidth measurement and calculation
- Gateway performance
- IP device performance
- Feature support and end-to-end operation
- Scalability

Network Impairments

Network impairments should be introduced in the test bed architecture to simulate characteristics that will be encountered in a live network environment. This will be accomplished via software during the design validation phase and via hardware simulators in the lab and limited live beta testing. Impairments that should be considered are testing of extremely long distances, excessive packet loss, excessive delay and delay variation, both individually and combined. All codecs that may be used in the network should be tested as well as all IPT devices, including gateways and phones and adapters.

Timing/Synchronization Issues

Although IP networks are inherently asynchronous systems, sending packets of any length when the packets are ready to be sent, success with voice applications lies in ensuring that packet play-out at the end systems closely mimics the synchronous delivery of serial voice streams from older telephony systems. This steady voice communication can be simulated very effectively, even in a highly asynchronous IP environment, using buffers and, even better, using the Real Time Control Protocol (RTCP) along with the Real Time Protocol (RTP). Doing so provides a closed feedback loop between the receiver and sender based on information about the arrival statistics of multi-media information being sent. RTCP is not widely used today by enterprises, or even many service providers and carriers. Many organizations are simply struggling to get IPT up and running in any condition, but RTCP will be used increasingly as the focus turns to QoS and QoE. The good news is that most manufacturers are ensuring that RTCP is implemented in their products and that RTCP will be available when organizations are ready to exploit it.

RTP and RTCP

Multi-media information, specifically voice and video over IP, represent digital information or digitized analog information that must get across the IP network in a manner and with performance characteristics that are much more sensitive to delay and delay variation than most of their data counterparts. Multi-media sessions are established using H.323 or, more predominantly in today's carrier and service-provider networks, Session Initiation Protocol (SIP), and use RTP to actually carry the information. The problem is that RTP alone is blindly sending information into the network with no awareness whatsoever of its fate. Did the packet arrive? If so, what about timing and delivery? What was the delay? How about delay variation? If packets are discarded, is it possible to dynamically modify the transmission, including packet size and content fill, to optimize the connection? If a less-desirable codec is being used for bandwidth savings, is it possible to use a higher-quality codec in the presence of greater bandwidth resources? These issues and more are of high value in optimizing the multi-media experience. Though it is not yet utilized in most networks, there is a solution: RTCP. RTCP is used in conjunction with RTP to provide feedback on the status of the packets as they arrive at the destination, and there are versions of RTCP being developed to provide information about the state of the intermediate network and its performance, as well. Network implementers of IPT should tie RTP statistics back to the SLA.

Buffers can be thought of as “holding tanks” for information as it traverses the network in the form of packets. Buffers are good in that they provide a staging area in the case of momentary network overload to keep packets and reduce loss. Buffers are also bad in that improperly managed buffers can add unnecessary delay to connections, especially time-sensitive multi-media traffic. During the network design validation and testing phases, it is important to manage buffers in end systems such as IP phones and gateways as well as in intermediate systems such as routers. Most routers are not optimized by default for multi-media traffic and often require software upgrades to accommodate true multi-media traffic. The likelihood of a successful implementation of IPT is greatly increased if this optimization is done in an earlier, rather than later, phase.

Buffer Management

Most simulation tools will allow some input of information relating to buffer management, such as buffer management algorithms, ingress and egress buffer sizes, and internal switching approaches. Buffers are large areas of memory for storing multiple packets as they transit the network. In some cases, it is valuable to get buffer management performance results from manufacturer-specific simulation tools for input into third-party simulation tools.

Static/Dynamic/Adaptive Jitter Buffers

In addition to the large memory buffers, there are also jitter buffers, which are often as small as three bits in size, that are used to compensate for the timing variations between the circuit and the receiving device. There are three basic “flavors” of jitter buffers: static, dynamic, and adaptive. Static are the least effective for multi-media. They use a single, fixed and predetermined size of buffer, usually optimized for fragmented data packets for all network traffic. Static buffers are very common in older routers, especially the lower-end small office/home office devices. Dynamic buffers are better than static buffers because they dynamically calculate an optimum buffer size based on the first n packets received. Although this approach is preferred over static jitter buffers, it is not an optimum solution. Adaptive jitter buffers represent the state of the art as they can adapt to changing network conditions. Not only should IPT equipment that employs adaptive jitter buffers be chosen for use in the network, the use of adaptive jitter buffers should be simulated to the extent allowed by the chosen testing and validation software tools.

Broadband Bandwidth Measurement and Calculation

Table 4.5 only hints at the wide range of bandwidth utilization between two codecs of very similar MOS performance, but Tables 4.6 and 4.7 strike closer to the heart of the matter. The difference between Tables 4.6 and 4.7 is simply the packetization interval. Table 4.6 shows packets sent every 10ms, resulting in more packets but a smoother flow of voice that is less susceptible to packet loss. Table 4.7 shows a packetization interval of 30ms. With a 30ms packetization interval, bandwidth is consumed less quickly because more voice samples are placed in each packet, thereby reducing the number of overhead bits per sample. However, the connection is impacted more heavily by packet loss, as the loss of a single packet will result in the loss of three times as many voice samples.

Voice Type	over Link Type	Full/Half Duplex	Link Speed (kbps)	10 ms Packetization, VAD off			10 ms Packetization, VAD on		
				G.711	G.729	G.723	G.711	G.729	G.723
				Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)
Digital	TDM	Full	n/a	64,000	8,000	6,300	64,000	8,000	6,300
VoIP	802.3	Half	10,000	252,800	140,800	n/a	176,960	98,560	n/a
VoIP	802.3	Half	100,000	252,800	140,800	n/a	176,960	98,560	n/a
VoIP	802.3	Full	100,000	126,400	70,400	n/a	88,480	49,280	n/a
VoIP	Frame Relay	Full	any	100,800	44,800	n/a	70,560	31,360	n/a
VoIP	PPP	Full	any	102,400	46,400	n/a	71,680	32,480	n/a
VoIP	ATM (AAL-5)	Full	any	127,200	84,800	n/a	89,040	59,360	n/a
VoIP/hc	Frame Relay	Full	any	72,800	16,800	n/a	50,960	11,760	n/a
VoIP/hc	PPP	Full	any	74,400	18,400	n/a	52,080	12,880	n/a
VoFR	Frame Relay	Full	any	71,200	15,200	n/a	49,840	10,640	n/a
VoATM	ATM (AAL-5)	Full	any	84,800	42,400	n/a	59,360	29,680	n/a

Table 4.6: IPT bandwidth estimate with 10ms packetization interval (Source: Matthew Michels, Nortel Networks).

Voice Type	over Link Type	Full/Half Duplex	Link Speed (kbps)	30 ms Packetization, VAD off			30 ms Packetization, VAD on		
				G.711	G.729	G.723	G.711	G.729	G.723
				Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)	Bandwidth demand (bps)
Digital	TDM	Full	n/a	64,000	8,000	6,300	64,000	8,000	6,300
VoIP	802.3	Half	10,000	169,600	57,600	54,400	118,720	40,320	38,080
VoIP	802.3	Half	100,000	169,600	57,600	54,400	118,720	40,320	38,080
VoIP	802.3	Full	100,000	84,800	28,800	27,200	59,360	20,160	19,040
VoIP	Frame Relay	Full	any	76,267	20,267	18,667	53,387	14,187	13,067
VoIP	PPP	Full	any	76,800	20,800	19,200	53,760	14,560	13,440
VoIP	ATM (AAL-5)	Full	any	84,800	28,267	28,267	59,360	19,787	19,787
VoIP/hc	Frame Relay	Full	any	66,933	10,933	9,333	46,853	7,653	6,533
VoIP/hc	PPP	Full	any	67,467	11,467	9,867	47,227	8,027	6,907
VoFR	Frame Relay	Full	any	66,400	10,400	8,800	46,480	7,280	6,160
VoATM	ATM (AAL-5)	Full	any	84,800	14,133	14,133	59,360	9,893	9,893

Table 4.7: IPT bandwidth estimate with 30ms packetization interval (Source: Matthew Michels, Nortel Networks).

Results of this kind can be used to play “what if games” and to develop realistic bandwidth estimates for different types of connections.

Gateway, Softswitch, Session Border Controller and IP Device Capacity and Performance

Gateway, softswitch/call server, session border controller, and IPT device performance and capacity must also be factored into the mix, especially if the IP devices are not dedicated IP phones. In addition to performance metrics, capacity of the devices must be evaluated and simulated. Metrics of importance for this IPT equipment include not only the number of steady-state connections that are possible to support simultaneously but also the number of Busy Hour Call Attempts (BHCAs) and resulting Busy Hour Call Connects (BHCCs). IPT systems are just now beginning to approach the performance of traditional switching systems and very often it is demand—especially at centralized locations such as PSTN-to-IP gateways and traditional and IP-enabled call centers—that overload systems resulting in lost calls and potentially lost revenue.

Testing, Feedback, Network Modification, Testing Loop

Everybody agrees that testing in a laboratory environment, followed by testing in a controlled semi-live “beta” environment, followed by controlled release and then general release, is a good idea. It is remarkable how few organizations actually budget the time and personnel needed to do a thorough job of testing prior to actual live operation. Many organizations simply underestimate the complexity of voice itself—whether packetized or not—and don’t budget enough time or resources to fully understand the system until it is under live load and much more difficult to learn and modify. Organizations will hopefully be willing to learn from the failure of others and budget appropriate resources for the testing cycle as shown in Figure 4.5.

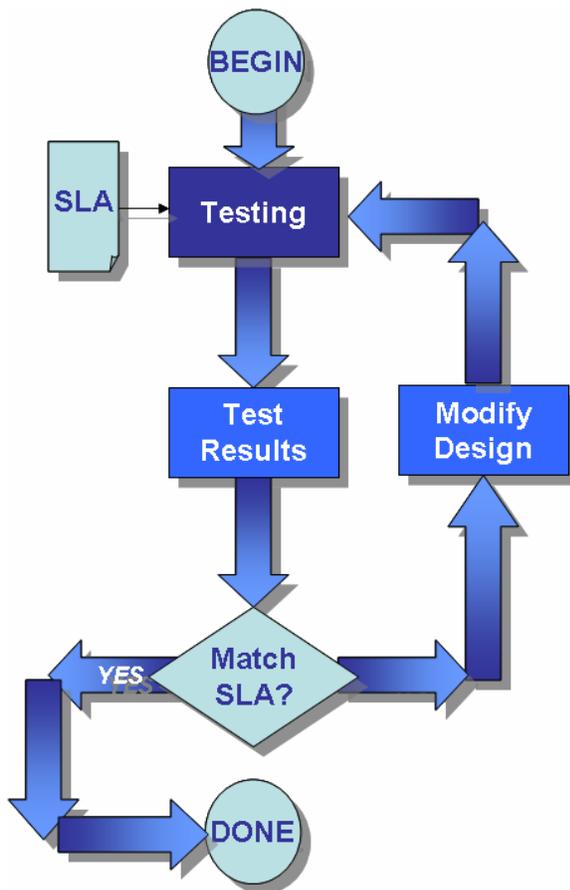


Figure 4.5: Testing, feedback, network modification, testing loop.

Testing/Proof of Concept Objectives

There are two main objectives for testing and they will work if the foregoing design effort is done properly:

- Does the theoretical design meet the actual objective?
- Do chosen products and vendors meet criteria?

And, in case there are either major or minor flaws in the underlying design, this process is set up in a loop to allow modifications to be accomplished at this step that will ensure project success.

Feature Support and End-to-End Operation

Although many of the capabilities and systems referred to earlier in this chapter can reasonably be tested in isolated environments, all features must be tested end to end. This will begin in the “closed lab” environment, then move into the limited production environment, and then on to the live environment as testing moves closer to a realistic environment.

The actual feature set that will be used in your environment will vary, so it is very desirable to establish profiles that include the full set of features a given user profile requires. It is also desirable to have an actual member of the user community who represents other users with the same profile validate the feature set and help test the features. This relationship should have been established well before this point and at this point it should be simply a matter of putting the user to work in helping with final testing. In fact, it is in final testing where the actual users should be put to work. The voice or multi-tech engineers should ensure that features are working correctly and to specifications before actual users are brought in—actual users should be used for validation and certification and not for initial testing and repair. Do not forget that they are the liaison with the other members of the user community—the ultimate judges of the system—and that their opinions on the readiness or non-readiness of the system will filter back quickly.

It might be possible to establish as few as four user profiles that are common to all departments, though it is more likely that there will be far more profiles for any given situation. Let’s look at six basic profiles, just to get an idea of the type of feature sets they might need. Although nowhere near the comprehensive list needed by most organizations, this will provide an example with some points of comparison and differentiation.

Profile	Possible Job Functions	Telephony Features
Basic Telephony User	Office worker, general staff, supervisor, low-level manager	Dial tone In-house (extension) dialing Local dialing Long distance dialing Basic voicemail Call waiting Caller ID
Intermediate Telephony User	Secretary, telemarketer, inside sales person, Help Desk worker	Basic telephony user PLUS: 3-way conference calls Skills-based call routing Call forwarding
Advanced Telephony User	Senior secretary, group admin, senior telemarketer, senior Help Desk worker	Intermediate user PLUS: Enhanced voicemail (add fax storage/retrieval and advanced messaging and retrieval features) Enhanced call forwarding Multi-party conference calls
Power Telephony User	Administrative assistant, senior sales person, senior manager	Advanced user PLUS: Multimedia conferences (Web plus audio) Enhanced call forwarding with remote access
Multi-media User	Senior manager, technician, developer, product marketing manager, senior sales manager	Most of the above PLUS: Call routing management and follow-me features Voice messaging Instant messaging Multi-media conferencing Collaboration
Road Warrior	Sales executive, account manager, senior sales manager, sales VP, sales director, marketing executive	Remote telephony Remote data Remote video Advanced call routing and follow-me features

Table 4.8: Sample user profiles and telephony features.

It is clear that many organizations will have additional requirements, such as various skill levels of call takers, call taker supervisors, operators, executives, administrative assistants, and similar skill groups but this basic mapping of user profiles to features will show a way of simplifying both the testing and implementation.

Closed Lab Environment

Initial testing is performed in a closed lab environment following the process described in Figure 4.5. The process begins with testing to ensure compliance with the SLA. At this point, the SLA should be considered unalterable and should be changed only under the most extreme circumstances as it has become the blueprint that has been agreed to by all parties.

Test results are compared with the SLA. If they match within defined parameters, it is time for the next step. If not, the network is modified and the testing loop is repeated.

After Hours/Off Hours Familiarization and Benchmarking

The next step is to perform the same set of steps as in the lab environment in the real network during non-business/non-busy hours. A good practice is to perform this phase of the testing using a staging area. The staging area is clearly marked out with empty tables surrounded by yellow tape on the floor. Everything that is taken into the staging area is inventoried so you will know what tools, software, connectors, test equipment, network equipment, etc. was needed to make the after-hours test a success. From the list of tools developed in this step, it is possible to put together an optimized toolkit for actual deployment and to write up a series of implementation guidelines and field service notes.

Busy Hour Testing in a Live Network

After success in the non-busy hour testing, it is possible to reproduce the testing during the busy hours in the live network. Many of the principles addressed in Chapter 3 about Network Assessment apply to the live-network testing phase; it may be useful to revisit the “Tips and Tricks for A Successful Network Assessment” section in Chapter 3 at this stage. After successful testing in this controlled environment, it is possible to go to the next step.

Evaluation of Results and Ensuring End-User Acceptance

Results from all prior steps are now evaluated and a “go/no-go” decision is made. If staging and testing results are not conclusive or compelling changes must be made to the network, testing and validation should be repeated. Otherwise, it is possible to move on to the next step—implementation and migration, which will be covered in the next chapter.

One requirement that has been referred to frequently during the first part of this guide as a key to success is end-user acceptance. This point has been discussed peripherally, but it is time to dig in deeper and understand what end-user acceptance really means before you can move forward with the implementation of the telephony project.

Prior chapters discussed the importance of keeping the users involved in the project as it progresses and in final testing and acceptance. Chapter 3 went into a great deal of detail about how to inventory the functions of the traditional telephony system to ensure that they are addressed in the new IPT system. And, by “addressed,” not every function of the traditional telephony system must be replicated, but rather, any discrepancy must be dealt with and the end user’s satisfaction must be weighed against other factors.

Let's assume the laundry list of features has been implemented and checked by the test technicians. At this point, it is time for representative users to be brought in for a final acceptance test: what will they hear, what will they experience and how will they judge the system?

The very first new thing with which the user/tester will have to deal is the new phone device. If the new phone is substantially different than the old phone—if, for instance, the new phone has a screen and softkeys or is a PC-based softphone—quick training and familiarization will be needed. The familiarization is best done by identifying the way in which traditional functions are performed on the new phone and then showing any new features that are important to the specific user's job. What this means is that if someone uses a phone to only make calls, they do not need to be shown how to program the phone with XML.

The next step will be for the user/tester to activate the phone. At this point, the user will judge the phone not only on the time it takes for the dial tone to be heard but also on the actual dial tone. Although most people who have never traveled outside their own country don't realize it, there are actually different dial tones in different parts of the world. The first step to a user's maximum comfort level and ultimate acceptance of a new system is for the new system to operate in key ways like the old system; the first point of comparison will be the dial tone. In implementations of a new system in a single country, it is important that all new phones present a dial tone that sounds correct for that country. In multi-national implementations, it is highly desirable for all phones to present a dial tone that is consistent with what is expected in each country.

The next issue is the time it takes dial tone to be presented to the caller. This is, quite conveniently, called delay-to-dial tone or time-to-dial tone and is one of the most critical elements of a user's comfort in using the system.

For instance, if the user lifts the receiver and waits too long for the dial tone, they will often, almost instinctively, put the handset back on the hook and raise it again to make the call. This time, they get dial tone almost immediately, but it is not a new dial tone. It is the delayed dial tone from their first call attempt. This will happen because dial tone is not automatic, and it is generally not a function of the local phone—dial tone is an audible signal to the caller that they may begin dialing and comes as a result of checking with the switch to determine whether a call would possibly go through if it were dialed. An alternative would be to play an “all circuits are busy now” or other similar announcement.

Consider another case. Consider the case in which a caller lifts the receiver and, without waiting for dial tone, begins to press digits. If they pressed 1-555-984-5800, they might get a message that says “a one is required before this number” or even “55-984-5800 is an invalid number, please try again.” What happened? In this case, it is possible that the caller began pressing digits before the PBX or switch was ready, and the PBX or switch only got some of the digits. Both of these problems are related to delay-to-dial tone.

In terms of system design, the time-to-dial tone issue can be exacerbated if an IPT device must wait on signaling from a remote call server or softswitch before issuing a dial tone, which they often do. If the call server is located locally and connected by a high-speed LAN with low utilization, the issue is not much different from a traditional telco arrangement with a local PBX or switch. However, if the new IPT solution uses a centralized call server that is located in a distant office, especially across a congested network and one that is prone to discarding packets, then the Time-to-Dial Tone might be too long, resulting in user dissatisfaction with the new system.

There is a subtle difference on how and where dial tone is actually produced in some IPT systems. For example, using a SIP-based system the phone itself may actually generate a dial tone whenever the receiver is lifted or the speaker button is pushed. This initial dial tone is considered a ‘comfort’ feature in that the user expects to hear dial tone. Only after the user has entered some recognized pattern of digits, or pressed a ‘send’ button, does the call server actually get involved in the process of call setup. Equivalent to ‘delay to dial tone’, in this instance it would be ‘delay to call (or session) setup’. The impact on the user is the same; delay is unwanted and is a major factor in call quality and QoE.

After the dial tone is received and dialing is underway, the next issue will be any messages and tones that a caller will receive while placing the calls. All messages must be in the proper language and tones must be consistent with what the user expects. This includes any trunk/all circuits are busy and other progress/proceeding tones or network announcements. The most important of these will be the busy or ringing/ringback tone that is received. An issue with the busy or ringing tone involves whether they come from the distant end (remote ringing/busy) or is generated locally (local ringing/busy).

In the first case, it is assured that the caller will always hear the appropriate busy or ringing tone for the location being called. This is not always desirable, however, for two reasons. The first one is that even a busy line or a line that is not answered will use network resources to packetize the ringing or busy tone and send it to the caller. The second issue is that the caller is not always familiar with the common tones for all locations they are calling. In this case, the caller may misinterpret the tones. Both problems are addressed by generating the busy and ringing tones locally, at the caller’s end. In this case, bandwidth is used only for the signaling packets that tell the local end that the phone is ringing or busy, and not for the entire duration of the ringing or busy signal. The second problem is addressed because the locally generated ringing or busy tone is always the same—it is always what the caller expects and does not matter to where the call is being placed.

Beyond actually setting up the call is the conversation itself followed by call termination. The quality of the call will be judged by comparison with what users are used to. It is often a good idea to establish a set of criteria in advance and to emphasize that “different is not bad” and that the objective is to provide a working, quality telephony implementation but not, necessarily, to replicate the old system in every way. Generally speaking, having the user rate voice quality on the Mean Opinion Score (MOS) scale, with 0 being the worst and 5 the best, similar to the scale shown in Figure 4.4, is the best idea. An E-model R value of 94, which corresponds to an MOS of 4.4, will result in about the same level of user satisfaction as the traditional G.107 Pulse-code modulation. Using MOS and R value also allows correlation of the actual user’s opinion to the metrics used to model the network at which time judgment can be passed on the suitability of the modeling tools being realized.

Fall-Back and Contingency Planning

One point that is often overlooked is that success in one phase or step does not absolutely ensure success in the next phase or step; it only makes success more likely. Fall-back and contingency plans should always be a part of the Network Design and Testing phases to ensure that older systems can still be brought back online in the case of a failure. This does not always mean true parallel operation but does mean that nothing permanently fatal should be done to the older system before the operation of the new system has been verified against formal acceptance criteria that have been signed off on by all parties.

Use of Third-Party Tools

Multimedia networks supporting voice applications are complex organisms requiring specialized monitoring intelligence. It is also obvious that manufacturer-provided monitoring tools are a part of the solution but do not take into account the nuances and subtleties of the overall, usually multi-vendor, environment. However, generic monitoring and management tools are insufficient—to be truly effective, third-party tools must be constructed with some knowledge of the individual manufacturer's systems being maintained in order to provide a system-wide, end-to-end view of system availability, quality and performance. What is needed is an approach that makes best use of key aspects of each.

The key areas that monitoring and management tools must cover are:

- Pre-deployment simulation and modeling
- Manufacturer-specific monitoring and management
- Real-time business views
 - Call detail records
 - Calls in progress
 - Delay-to-dial-tone rates
 - Busy Hour Call Attempts (BHCA's)
 - Busy Hour Call Completions (BHCC's)
 - Gateway channel configuration, utilization and loading
- Real-time call monitoring
 - Phone and multi-media device availability and monitoring
 - Successful vs. failed call completion rates
 - Delay and delay variation
 - Packet loss
 - MOS
 - Poorly performing components
 - Service level breaches and SLA compliance

- Real-time interface to Manager of Managers (such as HP OpenView, EMC Smarts, IBM Tivoli NetCool)
- Summary and exception reporting
 - Utilization trends over time
 - Managed devices by company, department and location
- Asset monitoring and tracking
 - Phone registrations and de-registrations
- Capacity planning
 - Incoming and outgoing calls
 - Loading by dial plan, routing rules and gateway
 - Bandwidth utilization
 - Delay and delay variation
 - Packet loss
 - Route patterns, utilization and availability

Use of appropriate, qualified, sophisticated third-party tools will allow documentation of results with consistency and reproducibility and a lack of vendor bias.

Summary

This chapter provided plenty of food for thought as you enter the network design and pre-deployment testing phase of your IPT project. It included considerations both from the IP-packet realm and from the realm of traditional telephony. In addition, another report card has been provided to assist you in the selection of design and validation software tools. This chapter has enabled you to put to good use a lot of the very critical foundational knowledge introduced in prior chapters.

Chapter 5 will go to the next phase of the project: implementation of the system you have benchmarked, validated and optimized, and then, having established a firm foundation, will begin the task of migrating actual users to the new network.

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>.