

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 7: Optimization.....	172
SLAs and the Scope of Optimization.....	172
The Optimization Cycle.....	173
The Use of Software Tools.....	174
Tool Ubiquity.....	174
Constant Testing and Monitoring.....	175
Exception Reporting and Filtering.....	175
“What If” Scenarios.....	175
Trending and Capacity Planning.....	176
Bandwidth.....	176
Ports/Lines.....	177
Phone Numbers/Numbering Plan.....	179
Grooming.....	180
Gateways.....	181
TDM Grooming.....	181
VLAN Capacity.....	182
VPN Capacity.....	182
Infrastructure.....	182
Hardware.....	182
Repair.....	183
Obsolescence/Refresh.....	184
Software.....	184
Maintenance.....	184
Enhancements.....	185
Upgrades.....	185
Patches and Security Audits.....	185
SLA Improvements.....	186
Delay.....	186
Call Setup.....	187
During Call.....	187
Call Tear Down/Release.....	187
Delay Variation.....	187
Sample and Packet Loss.....	188

Availability	188
Telephony Service Availability	189
GoS	189
Blocking/Non Blocking Access	190
Success Rates of Call Setup	191
Related Gateway Issues	191
Prices and Costs	191
Hardware Costs	191
Service and Support Costs	192
In-Sourcing vs. Outsourcing	193
IP Contact Center Optimization	193
Summary	194

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 7: Optimization

This chapter discusses the process of reviewing the performance of your IPT system and setting new standards and benchmarks. In other words, taking what is already a world-class system and making it even better. This chapter will make the distinction between operations, the objective of which is to keep things as they are, and optimization, the objective of which is to modify the telephony solution in a manner as to improve its operational characteristics.

The chapter will also make a distinction between optimization and troubleshooting and repair, the objective of which is to make something that is not operating or not operating within guidelines to work once again according to the guidelines. In fact, the underlying assumption of troubleshooting and repair is that some capability, system, or feature operated properly at one time and that something has caused it to no longer work as desired. This differs from optimization in that operation in an optimized state is desirable but the system, feature, or capability never operated within that state before but will after the optimization. In fact, the optimized condition will become the new baseline for proper operation and the next review cycle may improve further on prior work in a never-ending loop of improvement.

SLAs and the Scope of Optimization

Because the SLA represents the classes of services that will be provided by the network and because it was, or at least should have been, the focal point of formalization of the needs for the network and IPT services, the discussion will return once again to the SLA. Many of the most important optimization topics can be addressed via the SLA, and those that cannot, become candidates to upgrade the SLA to meet current and anticipated needs.



Figure 7.1: Scope of optimization and the SLA.

When SLAs were originally introduced, they were a tool of traditional service providers that offered services such as Frame Relay, ATM, and Internet access. The “service” for which the level was agreed to was a network-based service. Therefore, the points between which the SLA was defined were within the range of control of the service provider. It was only the rare case in the early days of SLAs when an SLA included the access circuits: SLAs were almost universally measured within “the cloud,” as Figure 7.1 shows.

The view of the SLA covering just the Layer-2 aspects of network services has undergone quite a bit of change since the early days. SLAs today are, increasingly, service-aware IP SLAs or even Multiprotocol Label Switching (MPLS) SLAs that add value to the IP infrastructure. What is required for a comprehensive and effective approach to optimization is to extend the scope of the SLA. The new definition, the domain of optimization, should not include only the Layer-2 services—which, when strung together end to end, create a path over which applications bits flow—but also the emerging concept of the application-aware IP SLA and a new, user-centered view of the end-to-end service. In this manner, the concept of the SLA is being maintained, while the scope is expanding to meet the real, service-oriented needs of optimization so that you can maintain alignment of service-level objectives and the SLA.

There is an entire industry dedicated to application monitoring and optimization. This industry is focused on the use of the actual application such as SAP. There is also an entire industry dedicated to monitoring and optimizing the packets, frames and, ultimately, bits, that flow over circuits. What is being proposed is something somewhere between the optimization of network services, in our case IPT, including the operation of all signaling and other associated functions that support the connection, transmission, and disconnection of packet voice calls.

The Optimization Cycle

By this point, you have set up a world-class packet telephony network, you have assured that, at the very least, the telephony service you are providing meets users' quality expectations and delivers the same basic services as the traditional telephony network it is replacing. It is also quite possible that you have added a number of important features and applications, such as unified communications, to the new network. The new applications make the users more effective in their jobs or make special new capabilities available, which is the real reason to go to all the trouble of implementing a new network. After all, they already had dial tone. By this point, you have also done all the broad, sweeping, system-wide things that can be done to ensure proper operation of your packet telephony services. It is, therefore, time to take a closer look at a number of very specific areas and to optimize what you have created.

Optimization is an ongoing process that is required to keep network costs in check and to avoid the demands of the users running ahead of the capabilities and budgets of the network. IPT networks have a tendency to consume huge amounts of resources and must be continually managed. Once the network is in place and IP-tone has replaced dial tone, the next step might be static image cameras on phones and then full motion, ad hoc, video conferencing, also referred to as video-tone. Who drives the changes? How much control does the internal or external service provider have? To what extent is the SLA considered? Is the SLA changed to reflect new needs or are the new, emerging requirements somehow shoe-horned into current categories? Independent of the answers to these questions, the network and its services must constantly be optimized, and, beyond the administrative questions raised here, the optimization process will be the subject of this chapter.

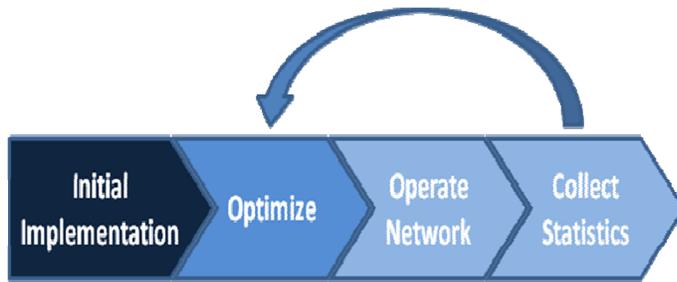


Figure 7.2: The IPT optimization cycle.

The Use of Software Tools

The software tools that will be employed for optimization include both the measurement and monitoring tools used in day-to-day operations as well as specialized modeling and simulation tools last employed at the time of the initial design process. The use of software tools is emphasized within the optimization phase due to the ubiquity of the tools within the network—the fact that they are constantly testing and monitoring many aspects of system operation, the repeatability of the tests and the ability to play “what if” games and simulate different scenarios in order to understand optimization options prior to putting them into operation.

Tool Ubiquity

Software tools can monitor a complex set of metrics across a network simultaneously. The implementation of monitoring tools must, of course, have been accomplished during the implementation phase as retro-fitting an active network with monitoring tools is far more complex, time consuming, and expensive than doing so up front. However, if the proper tools are not in place, it is time to circle back and put them in place. Both intrusive, often called active, and non-intrusive, often called passive, testing will be employed and will provide baseline operational data on the aspects of IPT system operation that will be employed. It is also important to have underlying network infrastructure statistics; however, the two families of tools that you need should be clearly understood and their roles defined.

On the one hand, specialized tools are required that monitor specific characteristics of real-time services; in this case, IPT. For IPT, the types of statistics that must be monitored include calculated MOS, delay, delay variation, E-Model R Value, and other metrics of importance specifically to the real-time voice class of service. It is also necessary for these statistics to be made available in a high-level summary with graphics—in effect, a simple network-wide report card—as well as at the individual call level so that the impact of optimization decisions on individual calls can be assessed.

On the other hand, you still need to understand the operational characteristics of the underlying delivery system: the network. The second set of statistics you are seeking comes from traditional network monitoring and management tools and includes availability of network components, trunk congestion, router buffer and queue management information, and similar types of metrics.

Constant Testing and Monitoring

Constant testing and monitoring is as important to ongoing operations as it is to optimization. In operations, it is desirable to see how the network, and overlaying services such as IPT, is performing in an attempt to ensure that they are performing within established guidelines. Optimization seeks, in contrast, to improve the operational characteristics and, therefore, to reset the baselines to higher standards. One of the aspects that constant testing and monitoring will show is any trends in network operating characteristics that might highlight important changes that must be considered within the context of the optimization process.

Exception Reporting and Filtering

A multimedia network of any size will generate far more information than can reasonably be assimilated by administrators working without automated tools. For this reason, exception reporting is used within the operations process to automate the function of identifying when certain aspects of operations are above or below certain pre-set thresholds. If, for instance, packet loss for VoIP connections exceeds a preset threshold of 3.5%, the operations personnel are alerted of an exception condition. This can be done through a variety of mechanisms, such as SNMP traps that cause red dots to appear on the screens of network operators to emails to beepers or IM messages to key operational personnel—and all this is done in real time.

Because optimization is not done in real time and, in fact, uses historical data, there is not “exception reporting,” per se, in optimization. The optimization equivalent of exception reporting is filtering. Software tools allow optimization engineers to sift through mountains of data using specific filters to identify exceptions and trends and any specific areas in which anomalies might exist that can be exploited by optimization.

“What If” Scenarios

The natural outgrowth of being able to work with mountains of real data from a real, working network is the ability to perform “what if” scenarios that will have outcomes that much more closely match the real operating environment. Having real operational data allows far more accurate prediction than was possible during the design phase of the project and, in fact, almost certainly means that the first optimization phase or two will yield much better results than subsequent optimizations. During the design phase, inputs to the “what if” process were either simulated traffic—because no similar real network previously existed from which to capture traffic—or was based upon an IP network that existed before the current network but supported data only and has been substantially changed in the process of implementing the new IPT solution.

“What if” scenarios can actually be executed using one or both of two approaches. Ideally, both approaches will be used. The first approach for the “what if” exercise is to have a specific list of characteristics, mostly encompassing the SLA metrics and possibly a few others, and to be looking specifically for ways to improve those metrics. This is the most common approach. The second approach is to rummage around in the historical network performance data looking for optimization opportunities. This second direction will often yield surprising and good results. In this second approach, it is also important to keep an open mind to anything you might find and, in fact, to discipline yourself and your team not to try to guess at or pre-determine the outcome.

Trending and Capacity Planning

Optimization includes trending and capacity planning and extends to minimizing infrastructure costs, such as decommissioning excess gateways, lines, and similar un-needed items; maximizing call success rates, security, business continuity; and future proofing the network. Some of the areas to which optimization may most productively be applied are:

- Bandwidth
- Ports and lines
- Phone numbers and numbering plan
- Grooming
- VLAN capacity
- VPN capacity
- Infrastructure

Let's take a closer look at each category and their interdependencies.

Bandwidth

Bandwidth is as inexpensive as it has ever been and the price keeps going down; it is worthwhile to invest in bandwidth to ensure a higher voice quality. But, even considering these two points, there is no reason to waste bandwidth. The optimization phase is the time to reconsider several options relative to bandwidth. The first item to reconsider is the choice of coding scheme. Pulse Code Modulation (PCM) has been strongly recommended throughout this guide as the best choice for a variety of reasons, but PCM also requires the most bandwidth of any of the voice coding choices.

The optimization process should begin by reviewing the reasons for choosing the specific voice coding scheme currently being used and determining whether it is still the proper choice. In this example, let's assume that this is the first optimization and that PCM was chosen. What were the reasons for choosing PCM coding for voice? Are those reasons still valid? The first reason is that PCM delivers the best voice quality under a variety of impairments and network conditions. PCM translates sound, not just voice, into ones and zeros for transmission across a digital circuit or packet network and will deliver a voice and "sound" quality that is as near as possible to the circuit system being replaced. PCM-based systems will often deliver a voice quality in the 4.2 to 4.4 range on a 5.0 MOS scale, within the target range for voice quality over a packet network.

Once the transition is made from circuit PCM to packet PCM, it should be easier to wean users over to other, less traditional-sounding codecs. This process can be accomplished slowly over time; it would be very difficult to do all at once. It is also true that newer digital codecs, particularly broadband codecs such as Global IP Sounds (GIPS), designed specifically for use in packet networks are improving constantly both in terms of better MOS and lower bandwidth utilization.

After considering the effect of codecs on the bandwidth of individual connections, especially on access or LAN connections, the cumulative impact on shared connections, such as shared network access and backbones, must be considered. A small change on individual user connections, such as their Virtual LAN (VLAN) connection, may seem insignificant in the bigger picture but the cumulative impact can be profound when hundreds or thousands of simultaneous connections are considered. Also remember that the major goal of broadband systems, after a large enough pipe has been provided, is enhanced performance and that the instantaneous performance of a broadband system is enhanced by more smaller packets than by fewer larger packets. In addition, although this is at odds with conventional LAN thinking—a frame of reference built on large quantities of cheap bandwidth going short distances—it is a guiding principle for this IPT deployment, especially if the communicating endpoints are somewhat distant.

Ports/Lines

Ports and lines are two hardware units of measure—not to be confused with logical Layer-4 ports—that must be optimized. They are also two terms that have different meanings based on context; the two possible contexts being traditional telephony and IPT.

In the traditional telephony context, a port is a physical connection into which one would plug a single telephone device or a device capable of connecting single telephone devices in groups of 1 to 24 per physical port. As the use of IPT grows, the use of traditional telephony ports should decline, though they may not completely go away. It is possible, for instance, that a traditional phone service may be maintained for faxing and/or for use with 9-1-1 calls. One such situation exists at Paulding County School District in Georgia and is detailed in the following sidebar.

VoIP, 9-1-1, and Port Optimization

Before VoIP, telephone numbers have corresponded to a specific physical location within a customer premise, which had a one-to-one correlation to a physical location within the telephone company Central Office (CO) known as a wire center. Each wire center has specific geographic coordinates and a corresponding Public Safety Answering Point (PSAP) from which emergency services such as fire, police, and ambulance are dispatched when a citizen dials 9-1-1.

Since the “break up” of the monopoly phone companies, many competitors have entered the market. And there may not be enough telephone numbers in their geographical areas of operation without purchasing another large block of phone numbers that may go unused. The solution? Assign virtual phone numbers from other areas that do not correspond to nearby wire centers. This is great for the new VoIP phone company, but this arrangement breaks the traditional pairings of wire centers and PSAPs. Is this a problem? For cell phones, no. Cell phones are mobile and can be anywhere. Cell phones also use a different system to report their whereabouts to emergency services. However, this setup can be a big problem for those who must rely on 9-1-1, such as schools and businesses that may need to summon help. In many cases, new VoIP users, especially residential and small business users, may not realize that emergency services will be dispatched by 9-1-1 from dozens or even hundreds of miles away...*until it is too late!* Workarounds exist, but 9-1-1 must be tested while setting up new VoIP systems. A county school district in Georgia solved this problem; let’s explore the workarounds they developed.

Before the implementation of their new VoIP system, Paulding County School District had purchased their voice services exclusively from their Regional Bell Operating Company (RBOC) BellSouth, now AT&T. As a part of the complete re-evaluation of their network, Paulding County’s Chief Technology Officer (CTO) reviewed alternatives and chose USLEC, an alternative carrier offering lower pricing, better installation lead times, and the full range of services available from AT&T but at better prices. The CTO also found the USLEC account team to be more cooperative in handling deadlines and the exceptions that come up during the cutover of a network his size. And, with seven special T1s – ISDN Primary Rate Interface (PRI) lines, Paulding County was an important customer to USLEC but was very much the “small fry” with AT&T. One problem that was encountered during the cutover was with 9-1-1. As a new telecom carrier, USLEC was assigned a group of numbers to provide to their customers that had previously been geographically associated with Fayetteville, Georgia, some distance away.

Ordinarily, the intelligent call routing inside the USLEC network gets all of Paulding County schools’ calls to their proper destinations, but if Paulding County schools dialed 9-1-1, the emergency services would be dispatched by the Fayetteville PSAP and, in many cases, may take 45 minutes or more to arrive at the school. This was clearly not acceptable.

The solution that was adopted was both clever and effective. Each school was to have a FAX/9-1-1 line and because AT&T held the numbers that would allow proper dispatch of emergency services from the proper PSAP, Paulding County had to order the FAX/9-1-1 lines from AT&T. And they did. But, Paulding County wanted USLEC to provide all their voice services. Thus, once the lines were installed and running, Paulding County requested that AT&T transfer the service, with the corresponding correct number for 9-1-1 dispatch, over to USLEC. This is a process known as “porting” and one with which BellSouth is legally bound to comply in less than 30 days from the date of the request.

From the standpoint of optimization, special procedures should be put into place to ensure that the numbers going to special phone numbers or lines, such as traditional analog fax or 9-1-1 lines, not be changed, reconfigured, or optimized in any way, including changing the way the calls are routed, their codecs, or other operational characteristics.

Just as the term port has a special meaning in the traditional telephony context so, too, does the term line. A line is also a physical thing and consists of a two-wire or four-wire twisted pair connection between a subscriber premise—a residential or business location—and the telephone company. It is critical that lines be removed, a common term is “retired,” when they are no longer needed because there is a monthly charge per line that does not always magically disappear from the traditional telephony bill as soon as they are retired. In fact, there are auditing services companies and consultants who do no more than audit the lines that are being used versus the lines are being charged and splitting the savings for the first month or so with the client. This is a big and time-consuming task for an organization but one that must be performed by someone. It is also noteworthy that in many cases, especially if a traditional phone company is being replaced by the organization’s internal IPT service or by a competing VoIP service provider, that they are often remiss in their duty of taking items off the bill and often do not allow retroactive credits for items erroneously charged. Auditing and removing unused lines from the phone bill is a critical element in any optimization exercise and one which, in many cases, automated auditing tools can help.

In the VoIP context, the terms port and line have their own meaning. In the VoIP context, a port is a physical thing, usually an Ethernet port, to which a VoIP phone or other system is attached. The idea of a line in the vocabulary of IPT is a bit less formal. Generally speaking, from a VoIP point of view, optimization of ports has two aspects. The first is that there be enough physical ports on Ethernet devices such as VLAN switches to accommodate the growing population of Ethernet-connected devices. The second thing to monitor is related to bandwidth: to ensure that each of the ports has sufficient bandwidth to give the call and voice quality needed, especially when aggregated with other traditional data, emerging video, and increasingly common Unified Communications demands.

Phone Numbers/Numbering Plan

Although the world of IPT is designed to work without traditional phone numbers (instead with alphanumeric Uniform Resource Identifiers), phone numbers will persist until all ties have been cut with traditional telephony and are on IPT-only telephony infrastructures. Experts suggest that that day is a long way off. In the mean time, optimization of phone numbers and auditing of the numbering plan ensures three things:

- There is a sufficient inventory of new telephone numbers to be assigned to new users.
- Numbers that are un-assigned are returned to the telephone number pool for reassignment.
- Any internal numbering plan that is in place is being followed. Internal numbering plans are used as often for accounting and cost allocation purposes as they are for geographical number routing.

Each of these steps should occur regularly but should be audited and verified as a part of the optimization exercise. This is another area in which automated tools and log analyzers can assist in the hunt for missing or improperly assigned telephone numbers.

Grooming

Grooming is an old term from the fading era of traditional channelized TDM telephony. Grooming refers to ensuring that the least number of T1s and T3s are used in order to optimize costs and network availability. In general, organizations have connected their traditional PBX or other phone systems to the phone company's Central Office switching equipment using T1s that provide 1 to 24 individual voice channels or Integrated Services Digital Network (ISDN) lines, which are specially formatted T1s that provide 1 to 23 channels usable for subscribers' voice connections. In today's IPT world, channelized T1s still exist but, for the most part, connect to softswitches, voice servers, Session Border Controllers (SBCs), and gateways rather than their PBX counterparts.

Grooming is an optimization technique that is performed when you have two or more T1s or their larger counterparts, T3s. Figure 7.3 shows the location of T1s or T3s used for telco network access. Each T3, if they are all used, contains 28 T1s for a total of 24×28 , or 672, voice connections.

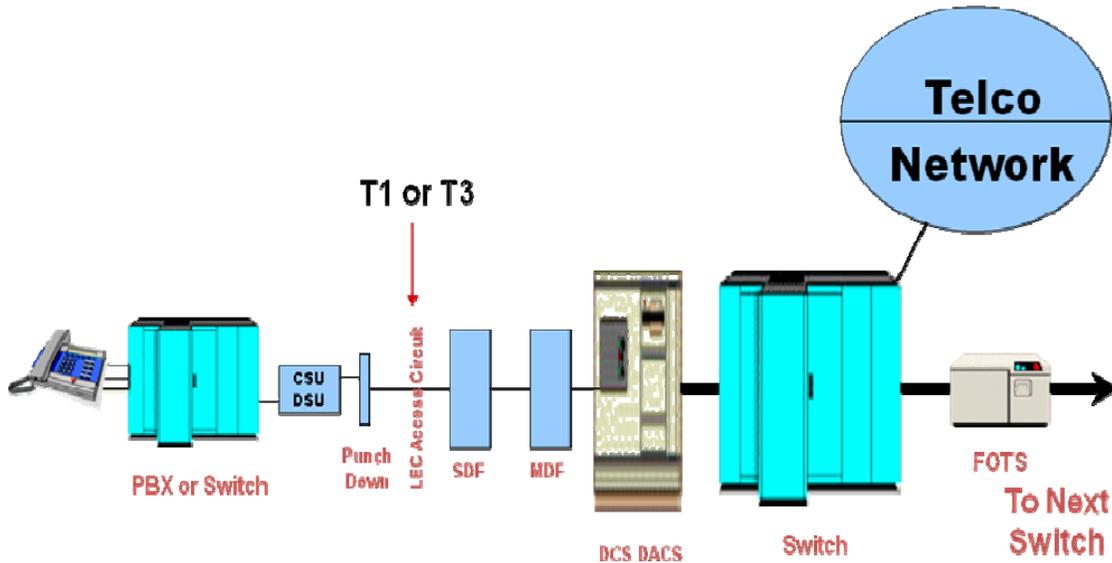


Figure 7.3: Traditional telephony interconnection.

The first consideration is redundancy and reliability. If this is a major consideration, you will have pairs of T1s or T3s, each leaving the building through a different exit and, ideally, traveling over a completely different path once they leave the building. It is even possible to have them ultimately connect to two different central offices of the incumbent phone company or two different Points of Presence (POPs) of competing long-distance carriers. This is all to say that during the grooming process, provisions should be made for redundancy and reliability.

Having addressed reliability and redundancy, let's move to the task of grooming T1s. Each specific channel of a T1 has one or more phone numbers associated with it. These may be individual lines or may be grouped together in a "hunt group" that allows multiple phones to ring in a round-robin cycle before the call is rolled to voicemail or some other method of handling. As the phone numbers associated with the individual channels are moved to VoIP or in some other way released, it is possible to regroup T1s and retire complete T1s. At that point, the billing should cease and further cost savings will be realized.

How does this work? Well, let's say that there are two T1s. One has 24 phone numbers assigned, in the format XXX-YYY-ZZnn where XXX is the area code, YYY is the exchange, and ZZnn represents the subscriber part of the number where nn goes from 01 to 24. Let's say that the other T1 uses the same setup, but the number is in the form XXX-AAA-ZZnn where AAA is a different exchange. Due to IPT activity and number reassignment, 16 numbers are released on T1 number one and 10 numbers are released on T1 number two. That leaves 24-16, or 8 channels, assigned on T1 number one and 24-10, or 14 channels, on T1 number two. 8 + 14 is only 22, which is less than 24, so all the channels can now be combined on—groomed to—one of the T1s. It is not important that the numbers are not numerically contiguous: the hunt group structure takes care of that, if needed, otherwise it is not an issue. At this point, the T1s have been groomed and optimized and one of them may be released from service.

Gateways

VoIP gateways, be they standalone or TDM ports on softswitches or VoIP servers, will be connected to the traditional network and usage will, most likely, be increasing. Grooming, therefore, as it relates to gateways and similar systems, will involve increasing capacity to meet demands. There are many ways to calculate the proper sizing for TDM capacity, but the method used by the phone company involves an Erlang calculation of the number of simultaneous channels used for a population of potential callers based upon the length of their calls and their behavior if they can't make a call. Erlang calculations also involve a determination of the likelihood of blocking of calls using a unit called 'P' where P.1 means that 10% (10 in 100) calls don't get through during the busy hour, P.01 means that 1% (1 in 100) calls don't get through, and so on. The P calculations are intended to predict a Grade of Service (GoS) during the busy hour, which is the time during which the network will be requested to make the maximum number of simultaneous calls.

 Optimization of GoS is specifically addressed later in this chapter.

 Although the specifics of these calculations are outside the scope of this chapter, it is important to point out that these calculations should have been done at the time of the network design and validation and should be performed periodically for optimization.

TDM Grooming

TDM grooming will be a matter of reducing capacity, as described earlier, or entirely removing T1s that are no longer needed because the old equipment is being decommissioned. Optimization includes auditing the bill to ensure that they are removed.

VLAN Capacity

The bandwidth and number of simultaneous connections on the VLAN must also be optimized. The demand on the VLAN will be increasing as more and more IPT traffic is moved off the traditional telephone system and over to the IPT system. In addition to bandwidth and simultaneous calls, intermediate devices, in this case VLAN switches, should be monitored to ensure that they have enough memory and processor capability for the increased load.

VPN Capacity

In conjunction with the internal or external carrier, the capacity of the virtual private WAN network must be optimized. This is yet another area in which automated tools can be very valuable. The specifics of the optimization will be different depending upon the specific underlying technology of the VPN. For instance, a different set of metrics must be optimized if the VPN is based upon MPLS technology than what would be done for a VPN based upon Carrier Ethernet transport.

Infrastructure

In terms of optimization, the “infrastructure” is considered to be the hardware and software that comprises the network. The transport bandwidth interconnecting the hardware components is considered separately for a variety of reasons, the primary one being that each optimization requires a fundamentally different skill set and set of automation tools.

Hardware

Hardware in the IPT sense potentially covers a lot of different components. Consider end-to-end the devices that may be involved and the optimization effort to determine their continued suitability for use in providing the IPT services of your organization. Consider first the wide range of possible hardware devices, the range and diversity of which is only hinted at in Figure 7.4.

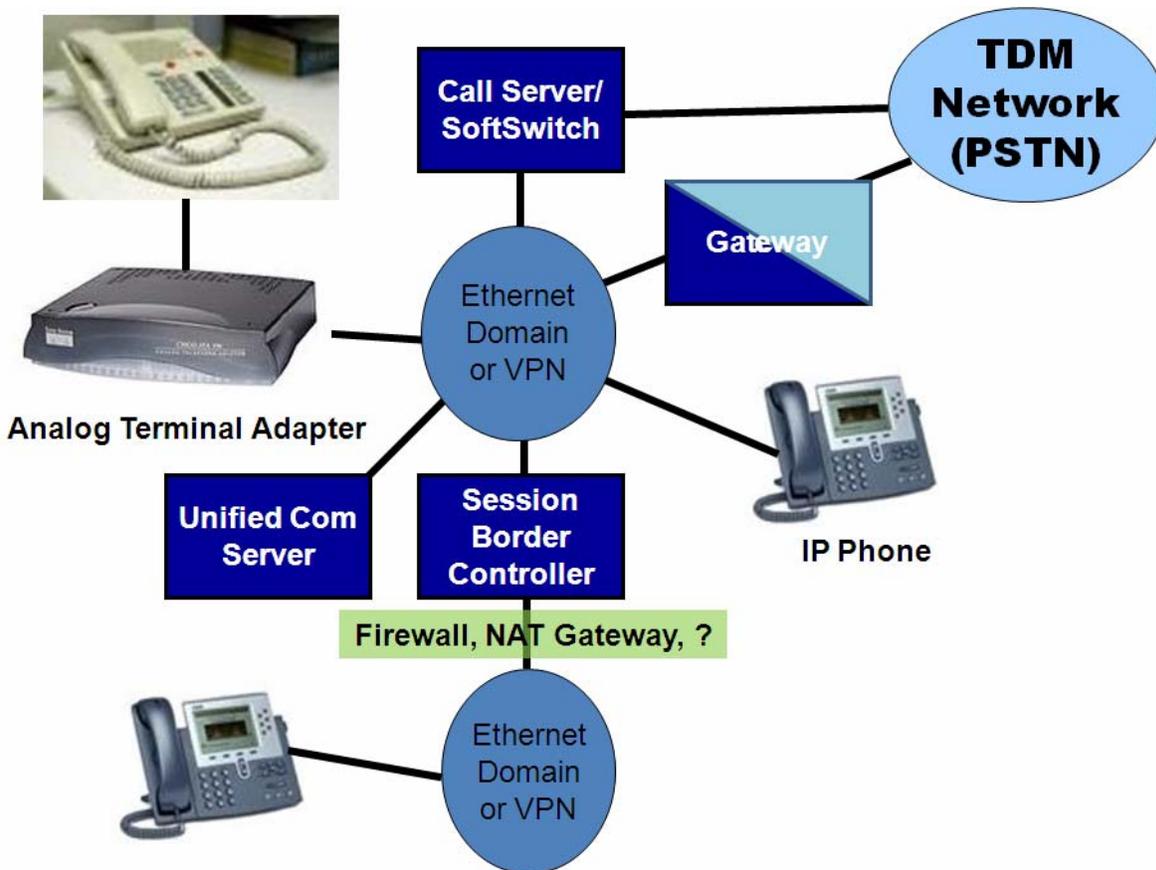


Figure 7.4: Infrastructure in transition.

Beginning in the upper left corner is a device that might be a simple traditional telephone. It may also be a more complicated Private Branch eXchange (PBX) phone. That device, be it a mechanical traditional telephone with a keypad and ringer or of the more sophisticated PBX variety, is connected to the network via an Analog Terminal Adapter. The ATA is connected to an Ethernet network, which, in an organization of any size, employs Ethernet VLAN switching and, most likely, traverses the WAN using one or a combination of VPN technologies. Then, there are, potentially, VoIP servers, softswitches, feature servers, gateways, firewalls, SBCs, NAT gateways and other devices. Each component that plays a role in the end-to-end connection must be evaluated and optimized.

Repair

The optimization is at once both a technical and financial “tune up.” As such, repair procedures must be reviewed for relevance and currency as well as for proficiency of the technicians. It is also important, as a part of the optimization cycle, to audit the repair process to ensure that units that have been repaired are being returned to service.

Obsolescence/Refresh

Optimization also involves the decision to decommission certain devices or to refresh the hardware. Specific guidelines should be in place to determine which devices will be decommissioned, or possibly moved to another part of the network, and which will actually be taken out of service. One effective way to implement a decommissioning policy is to take equipment out of service and replace it with its upgraded counterpart when it comes in for repair. This could, dependent upon the component, be less disruptive and less costly, especially for components used by end users, such as IP phones.

Managing Surplus Equipment

In many cases, especially during a migration from an older technology to a newer technology, there is a great deal of value in the older equipment, especially in a large organization that has an extended timeframe for migration. One of the primary reasons for a migration to a next-generation telephony system—any next-generation system—for many company sites has nothing to do with advanced new capabilities. The real reason for migration is manufacturer discontinuation of older systems and models. When a manufacturer discontinues a product line, availability and rising cost of components for maintenance, repair, and expansion of existing, older, vintage telephony systems becomes a problem. In addition to purchasing upgrades and replacements on the secondary equipment, a manufacturer may market a solution that can extend the lifeline of the older systems and create bottom-line benefits. For example, take equipment that had been decommissioned during the implementation of the new system and send it for refurbishing and placement back in service. This process recycles products either for repair or upgrades, in other locations that are still using the older equipment.

Software

Because almost all the components of the IPT infrastructure are, to some extent, computers, software maintenance, enhancement, upgrade, and the management of patches, licenses and security is of the utmost importance and all these areas must be audited routinely. There is no better time to accomplish this task than during the optimization phase.

Maintenance

Software maintenance is an ongoing, often evolving, process. Optimization of the software maintenance process ensures that all maintenance is performed in the simplest, most secure manner and utilizes as few human and network resources as possible.

Enhancements

Enhancements are differentiated from upgrades in that enhancements bring additional functionality and are voluntary. Upgrades are needed to keep software and firmware within a common operating range and they are mandatory.

When enhancements are made available, the process should ensure that all users to whom the enhancement apply are notified and given the opportunity to download it. Part of this process should be to ensure that users to whom an enhancement does not apply are not “teased” with knowledge of an enhancement they cannot get unless they are also allowed to change their equipment type, OS, and so on.

The optimization process involves looking at enhancements that have been made available, judging user acceptance, and making a determination about whether an enhancement will be supported going forward. It is important to note that the decision to offer, or allow, any enhancement should be made very carefully on the front end because all enhancements have costs—resources such as memory, processing capacity, bandwidth, user training, support—and that once a feature is offered, it is almost impossible to withdraw the offer or to withdraw support for enhancements and features that are already deployed.

Upgrades

An upgrade is involuntary, so it is often given a lower priority by users, is disruptive, or may change the way a system operates, or impact its ease of use. Optimization of upgrades involves first auditing to ensure that all pertinent upgrades have been made and second to assess the impact of the upgrades. Another factor that should be considered as part of upgrade optimization is licensing. As licenses can be either per end-point or time based an audit should ensure that an organization is not paying for unused licenses or breaching licensing agreements.

Patches and Security Audits

Periodically, manufacturers release software “band-aids,” small pieces of computer code that are intended as a temporary fix to a problem, most often a security vulnerability. These “band-aids” are called patches. Very often patches fix one problem or close one hole or vulnerability but create others. The assumption will be made here that proper due diligence was applied to the decision to apply or not apply certain patches and that the optimization process will audit to ensure that all appropriate patches are applied and that they are still needed or, if they have been superseded by upgrades, that the patch is no longer being applied and taking resources.

SLA Improvements

One of the most important aspects of the initial guidelines and baselines in the SLA was to make the transition from traditional telephony as seamless and easy as possible. Admittedly, this was not really necessary for many of the users as their perception of QoE was forged more as a result of recent emphasis on less-than-reliable “can you hear me now?” cell phones rather than historically superior hard-wired land-line phones. Nevertheless, emphasizing quality of the new telephony services relative to traditional “toll quality” is a safe bet as it will optimize the quality of [user] experience for all users.

At this point in the life cycle of the IPT system, the baselines of the new network relative to all of the important SLA metrics—delay, delay variation, packet loss, and availability—are the ones with which the users are familiar day-to-day. The optimization effort should be to improve the metrics but not so quickly as to cause users undue concern. One of the subtleties of “quality” is not that “quality” represents “the best” but rather that to be considered of a high quality, a product or service should be consistent. There is a lot to be said for the confidence one gets from consistency. These observations provide guidelines for the rate of improvement in SLA metrics. Keep in mind that “quality” is relative and that improvement, regardless how well intentioned, represents change and that change, if noticed, always brings concern and sometimes stress and resistance.

The following metrics, often tracked in the SLA and compliance process, are among the measurements of the greatest importance and the ones where third-party measurement and reporting tools can have the biggest impact. Beyond their use in the optimization cycle, the regular tracking of these metrics and exception reporting when the associated targets are not met, comprise the “dash board” or “control panel” of any modern multi-media network and its associated services.

Delay

Delay impacts the user perception of quality in many ways. There are a variety of things that can be done to reduce delay, or the perception of delay, and, therefore, positively impact the user’s QoE and set new bars for future performance. Any call, regardless of whether it is a data call, phone call, video connection, connection oriented, or connectionless, has three distinct phases: call setup, information transfer, and call tear down. There are several points in the process for optimization.

Call Setup

On the downside, if call setup takes too long, an impatient user, more familiar with the near instantaneous call setups of the traditional telephone network, might abandon a call and retry or even attempt to place a service order. For this reason, there are certain standards that must be applied. The first element is the time (or delay) to dial tone. If an audible dial tone, or some other indication that a call can be placed, is not presented to the caller within a very short time, the call may be abandoned. Beyond that, once the digits are pressed or the destination number otherwise signaled to the system, the next delay is the call setup.

Call setup times can be impacted by a variety of factors that can be optimized. For instance, if call servers are centralized, optimization can improve call setup times by increasing bandwidth between the telephony device and the call server or optimizing the placement of other systems that might contend with the call setup for resources. If, for instance, data servers and telephony servers are on the same VLAN, they can be separated. It is also possible to further divide telephony users into multiple VLANs or to upgrade telephony servers to ensure that they are responding as quickly as possible to call setup requests.

During Call

Analysis of traffic patterns, endpoint pairs, time of day, and time of month variations and shifts due to time zones will yield a lot of information about the consistency or lack of consistency of performance during a call. For instance, using automated tools to gather and compare delay, delay variation, and loss statistics for all calls and then plotting them graphically will show times when variations are most likely to occur and the causes of variations can be isolated and optimized. This is an important part of any ongoing optimization effort and, if done properly, can spot problems before they are reported by users.

Call Tear Down/Release

Once the call is completed, the call is released. This is an area of optimization that is most important if calls will be serial, such as in out-bound call centers. If a call is terminated and then the telephony user goes about non-telephony tasks, this delay is buried in the network, invisible to the user and, therefore, far less important to optimize. If, however, one call ends and another call must be made immediately, the call release time becomes important. The good news is that minimizing call setup time and performance during the call will usually have a positive impact on the call tear down delay because the entire end-to-end process has been optimized.

Delay Variation

Delay variation, also known as jitter, is greatly impacted by other traffic. As with delay minimization discussed earlier, delay variation should be tracked regularly and any variations outside of a set range should be reported and researched immediately, independent of the optimization cycle. But, as a part of the optimization cycle, delay variation must also be analyzed, trends noted, and improvements made. The same process applied for delay minimization will yield improvements in delay variation.

Sample and Packet Loss

Sample loss, defined as loss of the actual content representing human speech being transmitted, is one of the network impairments that will have the greatest audible impact on voice quality. If human reports or the results from automated measurement tools indicate that the observed or calculated MOS due to sample loss is outside of established benchmarks, the problem should be addressed through troubleshooting and repair. The optimization process, however, should establish new standards for packet loss.

Sample loss and packet loss are very closely related and should be optimized together. Sample loss refers specifically to the loss of samples of human speech that are created at the speaker's end of the call and must arrive at the listener's end and be played out as a part of the call. Packet loss refers to the loss of the protocol packets containing the human speech samples. If each packet contains a single speech sample, sample loss and packet loss are one and the same. However, for purposes of bandwidth efficiency, it is common that more than one speech sample be placed in each packet. The first task of optimization, therefore, will be to determine the settings for the network being optimized and to understand what, if anything, can be done to improve the current settings.

Research has shown that humans can tolerate a sample loss of up to 10% and in fact this may have been the original benchmark for the network, especially if the IPT system user's frame of reference for voice quality was cell phones. In any case, the goal of optimization is to analyze the loss, optimize the transport network and its components, and to establish new benchmarks.

Availability

Availability should be tracked regularly by operations personnel and will usually be shown as a general network availability figure, unless specialty tools are used to monitor availability of IPT as a business service. Basic tools that use simple network functions such as PING will provide sufficiently accurate availability information for IP networks. Although this is fine for most needs, it is insufficient for purposes of optimization of IPT services. For IPT services, availability should go beyond the ability to physically connect to the network and should include the simultaneous availability of all constituent components needed for the end-to-end call, call server, and so on. In other words, the availability of all individual components is important in the context of the availability of the telephony service as was discussed in the section earlier titled Scope of Optimization. Automated tools are very important in assessing service availability and isolating the specific components responsible for any degradation of availability.

Telephony Service Availability

Beyond the availability of the IPT network and its IP infrastructure, it is critical to monitor and optimize the availability of the traditional telephone service. This usually means the connections to the PSTN.

IP networks and traditional telephone networks are different in many regards, one of the most profound being the way in which each handles offered load. In the case of IP networks, which operate on a “best effort” basis, additional load is accepted and, in the absence of the more sophisticated QoS mechanisms that are just now being implemented in IP networks, connections compete for network resources on an equal basis. Basically, IP allows packets to fight it out among themselves on a level playing field.

Telephony networks have specific, clearly identified TDM channels that may be occupied by one and only one call at a time. And, with the exception of certain military and public safety systems, the rule is “first come first served.” The metric within IP networks that must be optimized is Quality of Service (QoS). The metric within the traditional telephony network that must be optimized is Grade of Service (GoS).

GoS

GoS represents the probability that an attempted call will get through. A typical target used by telephone companies is P.02, which means that only two calls out of every hundred attempted during the busy hour would not go through. Depending upon the number of people who might place calls, the length of the calls they might place, the number of telephone devices, and the person’s behavior if their call does not go through, GoS creates some level of oversubscription. That is to say, GoS measurements let telephone companies provide fewer actual dial tones than there are telephone devices because, except in certain rare cases, all the phones will not be in use simultaneously. This design approach is called a blocking system because there is a statistical probability that under certain circumstances, which are specifically predicted and allowed for in the design, calls may not be connected. Just what is meant by the “busy hour” and the calls attempted during that time is detailed in the following sidebar.

The Busy Hour and BHCAs

The “busy hour” is a somewhat mythical and mystical thing referred to by traditional telephony so frequently and so casually that an outsider might draw the conclusion that it actually exists. In fact, there are many busy hours. Because of the natural ebb and flow of the human daily cycle, the first busy hour, both for phone and data, is roughly 10 am to 11 am, plus or minus, local time. This is referred to more specifically as the morning busy hour. The afternoon busy hour occurs at roughly 1:30 to 2:30 pm local time. The evening busy hour occurs at approximately 7 to 8 pm local time. Many things can impact the busy hour.

To begin with, busy hours tend to have more calls occurring in a smaller time window if the calls are between points in the same time zone. Busy hour calls between points in adjacent time zones have the effect of having fewer simultaneous calls. Busy hour calls between points in non-adjacent time zones have the effect of yet fewer simultaneous calls. It is also true that the morning and afternoon busy hours will be of greater interest for business telephony systems while residential telephony systems will focus more on the evening busy hour. It is also true that calls between residential and business systems will more likely occur during the morning and afternoon busy hours. Work at home and Small Office /Home Office (SOHO) will also have an impact on call distribution. All of these variations should be taken into account when considering your own busy hour for optimization and planning purposes.

Another important consideration is Busy Hour Call Attempts (BHCAs). The number of simultaneous calls traversing a gateway, for instance, can be very large and is limited only by the number of TDM circuits interconnecting the gateway to a PBX, PSTN, or a long-distance carrier and the amount of memory and table space in the gateway. The amount of resources taken for converting IP call packets to TDM bits is small compared with the amount of processing resources for setting up calls and maintaining call state information. For this reason, many times gateways can handle a large number of calls of longer duration than they can shorter calls requiring thousands of call setups per second. Most enterprise gateways, and, in fact, many service provider and carrier-class gateways, are not capable of handling the BHCAs needed for large networks. BHCAs are another very important measurement and optimization number.

If the IPT system being optimized is trending toward more on-net calls, meaning more on IP net calls, then GoS on gateways connecting the IP network to the traditional telephone network will be continually improving because the number of simultaneous calls will be decreasing. In this case, the optimization process will be a financial one: decommissioning lines and optimizing access T1s as previously discussed. However, if growth in IPT users means that more calls are flowing across the gateway between the IP network and the PSTN, the number of traditional circuits interconnecting the two will need to be increased.

Blocking/Non Blocking Access

Ideally, the number of circuits available between the IP and traditional parts of a network should be such that call attempts are never blocked at the gateway. Non-blocking access as this is called, may be somewhat expensive to provide under normal circumstances but is a default situation under others. For example, if there are 12 total phones on the IP network and the IP network is connected to the PSTN with a T1, the interface is, by default, non-blocking because the T1 allows 24 simultaneous connections and only 12, at most, will ever be needed. However, if the number of phones exceeded 24 and a T1 were used, there would be a potential for calls to be blocked. If for instance, there were 25 phones and each caller wanted to go through the gateway simultaneously and make a call to a destination on the PSTN, one of them would not be allowed through. However, it is likely that someone might not be in the office, that they don't all need to be on the phone simultaneously, or that one or more people within the office might be calling each other and, therefore, not going across the gateway to the PSTN.

Success Rates of Call Setup

IPT gateway systems and their PSTN counterparts on the other side of the connection have the ability to report the success rates of call setups and to distinguish between call attempts and call completions. These are metrics that should be monitored and optimized as they are a key component of the overall IPT QoS delivered to the caller. Erlang B calculations are used to estimate the actual number of interconnect circuits needed, but observed statistics should form the basis of the actual number installed.

Related Gateway Issues

Not only are circuit connections required between gateways and the PSTN or other TDM devices such as PBXs but also the gateway must be sized to handle the call volumes. BHCAs should be closely monitored and the gateway's resources should be sufficient to ensure that BHCAs will be maximized and the number of blocked calls will be minimized. It is also important to note the distinct differentiation between calls that are intentionally blocked, and reported as such, due to lack of available circuit connections and those that are lost into the black hole of telephony due to lack of memory, buffers, or processor capacity.

Prices and Costs

Cost reduction must certainly be a part of each and every organization's optimization exercise. To clearly understand the component elements for which budget dollars are extended, this optimization should be divided into hardware costs and service and support costs.

Hardware Costs

To some extent as a result of Moore's Law—the observation made in the mid 1960s by Intel cofounder Gordon Moore that products based on computer chips have a predictable increase in capability while their cost goes down—end-user organizations have an expectation that hardware prices will continue to decline. To some extent, this is a realistic expectation, largely because they have and, in the hardware cost category, at least, they should continue to. Keep in mind, however, that declines are percentages of the new cost, not the original cost, so while there will be continuing decreases they will, in actual dollars, be less and less.

The specific prediction upon which Moore's Law is based says that capacity will double and cost will be cut in half approximately every 14 to 18 months. What this means in real terms is that something that cost \$100 in the year 2000 will cost \$1 in 2010. So what does all this mean in terms of actual hardware costs? In terms of the cost of the raw hardware—purchased as a commodity item in an open market and priced independently from any support services—costs will continue to come down. Larger organizations, who have a lot of clout as a result of their total dollar expenditures, often write a 20% year-over-year decline in prices into multi-year purchase contracts. In fact, the sellers are usually happy to have such contracts in place and are willing to include such clauses to reduce the chances of losing the business to a competitor and to lower their costs of constantly selling to the large account.

Hardware should originally have been chosen to yield the greatest possible Return on Investment (ROI) and lowest Total Cost of Ownership (TCO). In other words, it should originally have been selected not only for the lowest price but also for the longest service life, greatest upgrade potential, optimum ease of use and management, and a variety of related factors. Hardware optimization should first consider the extent to which technology refresh or upgrades could be done using existing equipment before purchases of new hardware are considered.

One example of these considerations for an IPT environment is in the choice of user connectivity options. When the IPT service was originally installed, for instance, software upgradeable SIP phones may have been considered and compared with the less expensive option of using a low-cost ATA and keeping the existing traditional analog phone. At the point when that consideration was first made, it is possible that the cost of the software upgradeable SIP phone exceeded \$600 each while the cost of using an ATA with the existing phone was \$120. Now that it is time to optimize the hardware, it might be prudent to review the earlier decision and consider the impact of time on costs. What you might find is that, due in part to Moore's Law and in part to competitive and market pressures, SIP phones costing around \$100 are now available and, even though ATAs now cost less than \$40, it might be wise to begin using SIP phones—even if not a complete replacement, at least for new service and to replace the older ATAs when they come back in for repair.

Service and Support Costs

Anecdotal evidence tells of organizations that are so aware of the constant decrease in prices of technology that they never make a purchase decision for fear of not getting the best deal. This statement may be a bit extreme, but is really not far from the truth in many cases. As true as this constant reduction of price, with increased functionality, is in hardware, the opposite is true in service and support costs.

Service and support costs are based upon the cost of human resources. Savings have been achieved in the past, but service levels have often suffered, by tapping workers in India, Pakistan, Ireland, the Philippines, China, and elsewhere, but even now there is an increase in the price of the best of these resources accompanied by the desire to improve service levels. A qualified VoIP support technician in, India, for instance, might know the technical aspects of the system and be a world-class troubleshooter but if they are unable to clearly communicate verbally with the customer and determine the nature of the problem, service suffers.

Service and support costs must be carefully weighed against the actual cost of an outage or poor support. This is one of the areas in which all costs need to be factored in: What are the acquisition costs for the service or support resource? Will it be outsourced or performed in house? Certain aspects of service cannot, for instance, be outsourced to India. Installation, for example, may require a costly truck roll and in-person visit by a qualified, and expensive, technician, but carefully designed self-installation tools might lower the instance of truck rolls, lower costs, use the qualified technicians more wisely, and get services up and running faster. Assuming, of course, that this process does not frustrate the end user and make their job substantially more difficult.

In-Sourcing vs. Outsourcing

One last consideration that should always be made, which is a hybrid of hardware and software optimization, is consideration of in-sourcing versus outsourcing. If you are presently outsourcing some or all your IPT infrastructure and/or management, it is possible that the assumptions upon which that decision was made have changed and it is prudent to review the outsourcing decision. If, however, you have chosen in-sourcing, or providing some or all of your own components and support in-house, it is also prudent to review that decision. In many cases, there is a compelling business case for outsourcing that is counter-balanced with a desire to more tightly control the mission-critical IPT services, and tighter control often wins because it is driven by fear of the unknown. After the organization is more familiar with the operation of the IPT services, and the fear of the unknown has become familiarity with the known, the organization is better able, and more comfortable, managing those services through an outside organization. At this point, it may be time to take the path of the more compelling business case and the organization has developed the operational maturity to do so.

There is a variety of aspects of the ongoing operation, and in fact the optimization of, an IPT system that could be considered for outsourcing. This is an area in which some creativity and new approaches to running the IPT aspect of the business could pay big dividends.

IP Contact Center Optimization

Although all the foregoing optimization techniques and approaches apply to IP Contact Centers (IPCCs), the IPCC is also very different and has its own special needs and requirements. Optimization of the IPCC requires a combination of technical optimization, as discussed in depth earlier in the chapter, and operator and supervisor training to optimize the key metrics upon which the IPCC is measured.

For example, the maximum number of calls in the queue, average and maximum call waiting times, and call lengths can all be optimized by properly incenting agents and providing improved agent training and more agents, but there are important dependencies to consider. For instance, providing bonuses for shortest call times or most calls handled will likely provide an incentive for quickly closing calls regardless of whether the problem is solved. To ensure that the callers' problems are being addressed satisfactorily, a follow-up method is required. A combination of scheduling optimization to ensure that the proper skills are available when they are most needed will be required.

Agent performance can be reported on a variety of metrics that vary by call center and applications, but typically involve sorting out and reporting an agent or group of agent's performance statistics by skill group, geographic coverage, closed sales, caller satisfaction, or any other performance-based characteristic. Statistics that can be optimized include number of calls handled, average time per call, number of aborted calls, agent downtime, revenue per agent, and even Interactive Voice Response (IVR) statistics including the number of calls handled by the IVR that did not require a human agent, number of IVR calls abandoned, number of IVR calls handed off to a human operator, and similar statistics.

Summary

This chapter has taken a broad brush at many of the most important areas to consider in optimizing the operations and costs of your IPT services. It has gone deeper into topics that are less implementation-specific in nature, such as hardware and service and support costs. The chapter emphasized that optimization is a cycle: optimize your current network, set new operational baselines, operate your “new and improved” network for a period of time, collect data, and begin the optimization cycle all over again. The next, and final, chapter will look at all the details and loose ends that did not fit comfortably into other chapters but that are nonetheless important considerations in order to have a truly successful implementation of VoIP and IPT in your environment.

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>.