

Products Product Literature

Making Voice-over-IP Perform as Advertised with QoS

How Changing Your QoS Strategy Can Ensure Next-Generation Business Benefits and Cost Savings from Voice-over -IP

Enterprises of all types are excited by the potential benefits of converged IP-based voice-data networks. New business applications such as distributed call centers increase customer satisfaction, while enhanced teleconferencing and remote teleworking maximize internal productivity, save money and simplify management.

However, the very nature of IP networks, which are characterized by bursty traffic and "best-effort" delivery, presents significant problems for a real-time application like voice. Best-effort delivery may be acceptable for Web traffic, but voice requires guaranteed delivery in order to achieve acceptable standards for business communications.

At the current time, most companies are evolving their converged voice-data networks, starting first by implementing voice-over-IP (VoIP) for intra-company calls, using gateways that packetize voice and route it over the existing WAN infrastructure (see Figure 1). Over time, companies are expected to take convergence to the next level, replacing traditional handsets and analog PBXs with IP phones and IP-PBXs.

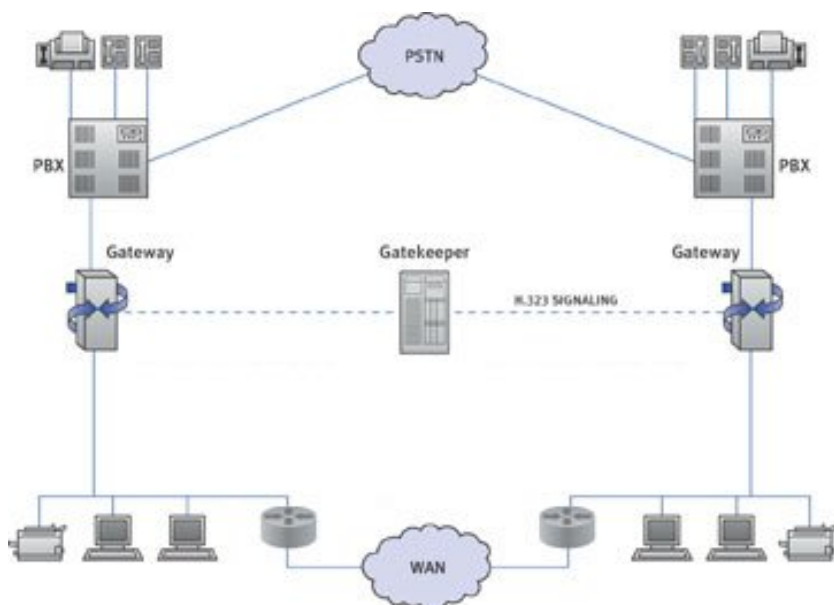


Figure 1. Implementation of VoIP in the enterprise for intra-company calls using VoIP gateways: For many companies, this is a practical first step on the migration path towards the con-verged voice-data network, as it allows them to preserve legacy PBXs and handsets. In such situations, when an employee places an intra-company call, s/he dials a prefix or code identifying it as such. The PBX recognizes the code and forwards the call to the VoIP gateway. The gateway signals the gatekeeper that it would like to place a call over the WAN. The gatekeeper then maps the number dialed to the gateway that has domain control over the called number, and gives instructions to the calling gateway to proceed with the call. The gateway translates the call into packetized voice, then passes it through the router and onto the WAN.

However, these migration plans will continue to move at a snail's pace until companies are able to guarantee Quality of Service (QoS) for VoIP. Why is guaranteeing QoS for VoIP an issue in converged networks? First and foremost, any time voice and data are mixed on a single, uncontrolled network, jitter, packet loss, and excessive delay result. Callers experience voice distortion, loss of portions of words or sentences, echoes, talker overlap, and dropped calls --all unacceptable for a business-critical application in an enterprise environment. Second, because of the different encoding algorithms (codecs) used to translate circuit-switched voice to packetized voice, it can be very difficult to pinpoint voice bandwidth requirements. Finally, the gatekeeper, a typical component of most VoIP deployments, must accomplish call provisioning without any knowledge of conditions on the WAN or topological reference.

Without QoS mechanisms to mitigate these problems:

- Companies that have not yet implemented VoIP will be reluctant to adopt the technology.
- In companies that have adopted, callers will shy away from VoIP due to poor quality, reverting instead to the public-switched telephone network (PSTN) for inter-office calls thus undermining the investment in VoIP.
- These companies will not move to advanced applications based on packetized voice if they can't get basic phone calls to work, further calling into question the initial decision to invest in VoIP.

Thus, QoS is the key to making voice-data networks a practical reality. In fact, Dataquest predicts that the market for packetized

voice services, while a multi-billion-dollar market, will still account for less than 10% of the revenue from circuit-switched voice services through 2004. Dataquest and other analysts agree that the stunted growth of the packetized voice segment is due to the inability to guarantee the quality of each voice call carried across the network.

In this paper we look at the specific network problems that adversely affect packetized voice and review the QoS mechanisms, or rather the combination of mechanisms, that are required to deal with these problems. As this paper shows, it is the completeness of the QoS solution (i.e., the integration of multiple QoS mechanisms working together) that determines whether or not companies get the QoS guarantees required to make VoIP perform as advertised.

What Are The Key Challenges for VoIP?

Jitter, packet loss, excessive delay, and poor call provisioning can wreak havoc on call quality. In order to understand how these threats can be handled effectively with QoS, it is first necessary to understand their effects on voice.

Jitter

Voice packets are generated at periodic time intervals by codecs, the encoding algorithms used to packetize and compress voice traffic. The number of bytes in a packet and the time interval between packets are determined by the particular codec that is used. Over a converged network, small voice packets will be interleaved with data packets of varying sizes, causing normally orderly packetized voice to arrive at disorderly intervals. This results in jitter, which, depending on the severity, will make voice sound poor and in some cases unintelligible.

To compensate for this condition, some VoIP equipment manufacturers provide jitter buffers in gateways or handsets. However, if the variability between packets is more than what the jitter buffer can handle, the packet is thrown away and it will be treated as loss. Throwing away packets means throwing away voice content and the result sounds like a bad cell phone call.

The key to minimizing jitter for voice traffic is to deploy a QoS mechanism capable of automatically detecting the required interval between packets, and adjusting queuing parameters in real time to ensure that this interval is maintained.

Packet Loss

The next condition that degrades voice is packet loss. When voice packets are dropped during their journey, the result is a disconcerting "clipped" quality to the human voice. Experts agree that packet loss in excess of 2.5-5% is unacceptable for voice traffic.

Packet loss occurs when queues in the routers begin to overflow during periods of congestion, forcing the routers to drop packets. Without QoS, the router has no way of deciding which packets to drop when the queue fills up, and voice packets will be dropped as randomly as any data packet. On the other hand, the router can't be told to queue all voice packets and never drop any of them, because first, there is only a finite amount of queuing available in the router and second, even if there were an infinite queue the voice delay budget would be exceeded.

To deal with packet loss, some VoIP equipment manufacturers offer a repairing algorithm called silence insertion, which makes up for packet loss by inserting silence packets meant to emulate pauses in human speech. However, silence insertion and other such repairing algorithms do not prevent packet loss, but instead attempt to minimize the problem after the fact.

Repairing algorithms constitute mere damage control, not proactive QoS. Therefore, it is also critical to deploy a QoS mechanism capable of preventing the conditions that lead to packet loss in the first place.

Excessive Delay

In order to have an intelligible conversation, the human voice has to stay within what is called "a delay budget." Raj Jain, Professor of Computer and Information Sciences at Ohio State University, Columbus, Ohio, did a survey of VoIP users and found that 77% of users felt that delay in excess of 150-200 ms was unacceptable.

Keeping voice within the delay budget on a converged network is a difficult task, as there is a certain amount of delay inherent in every VoIP implementation. The network itself adds roughly 50-60 ms of delay to each voice call. Codecs add a minimum of 25 ms on the transmission side. On the receiving end, the jitter buffer also adds approximately 50 ms of delay to the VoIP call. When you add up all of this delay, you're left with approximately 25-75 ms to play with before voice quality becomes unacceptable. Now, consider this scenario: if a voice packet gets stuck behind a single large data packet, the delay budget is obliterated. For example, a single 1,500-byte e-mail packet takes approximately 80 ms to be transmitted on a 128Kbps link. If one voice packet gets stuck behind that e-mail packet, the delay budget has been exceeded. And, more commonly, there would be several big packets queued as part of an e-mail transmission, not just one.

Another factor that affects overall transit delay is transient congestion. Occasionally, more packets arrive in a given period of time than can be sent out over the WAN. This causes congestion and excessive queuing delay. Without proper control mechanisms, all the queued packets will be transmitted based on respective bandwidth allocations and the excessive delay will continue for the duration of the call.

One of the most effective ways to minimize delay for voice traffic is to deploy a QoS solution capable of controlling the size of packets generated by data applications such as e-mail.

Another highly effective way of minimizing delay is to deploy a QoS solution capable of capping queuing delay at a specified

level, and discarding any packets that do not fit within this amount of delay. This will introduce a transitory blip in the voice call, but this is preferable to degrading the call from the time the congestion occurred.

Different Codecs, Different Bandwidth Requirements

No amount of control over jitter, packet loss, and delay will matter if there is insufficient bandwidth for each voice call. However, pinpointing voice bandwidth requirements is not a straightforward task. The amount of bandwidth required per call varies depending on the codec and the overhead associated with it. Therefore, it is essential to deploy a QoS solution capable of monitoring voice bandwidth consumption to determine proper voice bandwidth allocation. This QoS solution must also show the network manager the mix of traffic flowing over the network so that s/he can identify the traffic types impacting voice performance and take steps to control them.

The Challenge of Call Provisioning

In VoIP deployments, the gatekeeper is tasked with performing functions like call admission control. Because the gatekeeper usually sits on the LAN it lacks information about real-time conditions on the WAN. Therefore the gatekeeper does not know whether or not there is capacity on the line for one voice call or two or three.

If the gatekeeper allows one more call on the line than the prevailing network conditions can support, all of the calls in progress can be severely affected. Simply limiting the number of calls that can go through the gatekeeper at any given time isn't efficient because the gatekeeper will deny calls even when there is more than enough capacity on the line.

Yet another problem is that gatekeepers lack topological reference. In an enterprise with several branch offices, where the gatekeeper is deployed at headquarters, the gatekeeper only knows about the total number of calls provisioned in the network as a whole, but does not know how many calls are provisioned for one branch office versus another, the link speeds to the different branches, or the number of users in each branch. Blindly admitting calls leads to total unpredictability since calls to and from one branch may be totally fine at one point in time, while a call to the second branch sounds horrible. This situation may be completely reversed at another instance in time.

Therefore, it is critical to deploy a QoS solution that provides the network-wide intelligence necessary to ensure optimal voice quality for each call.

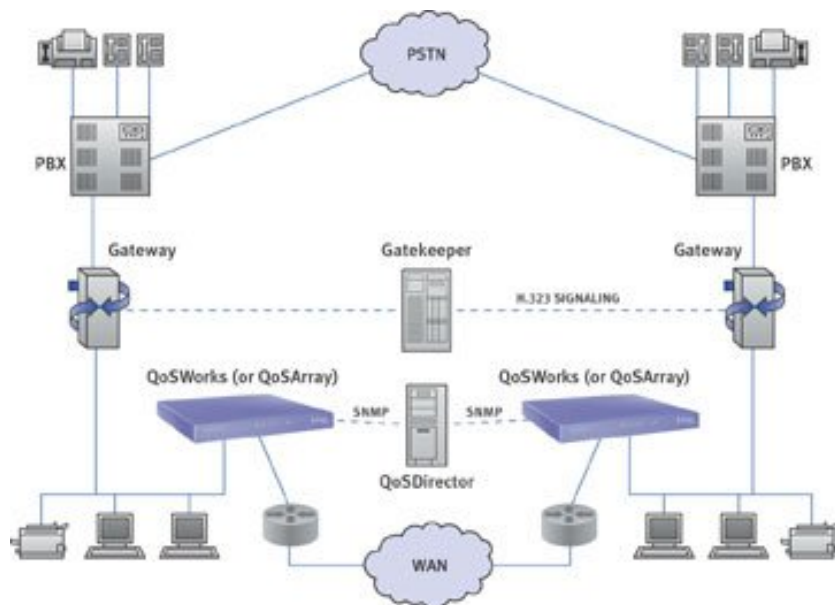


Figure 2. The Sitara solution enabling gateway-to-gateway VoIP for intra-company calls. Sitara's QoS solutions are ideal for ensuring QoS for VoIP in this type of deployment. Sitara QoSWorks (or QoSArray, where a high-availability solution is required) is deployed behind the router at the LAN-WAN interface. Intra-company calls are sent from the PBX to the VoIP gateway, which translates them from circuit-switched to packetized form and compresses them for transmission across the WAN. As the voice packets are sent over the WAN, they pass through the QoSWorks or QoSArray box, which applies policies to protect voice from other traffic sharing the WAN link.

Putting it All Together: The Sitara Solution for Quality VoIP

It is easy to see why ensuring QoS for VoIP is not an easy problem, which is why the technology isn't being adopted as quickly as it should be. At the same time, it isn't an impossible problem. Sitara Networks' QoS solutions are ideal for ensuring quality of service for voice traffic. Sitara's solution for quality VoIP consists of QoSWorks devices deployed behind the router on both ends of WAN links handling voice traffic (QoSArray can also be used where a highly available QoS solution is desired - see Figure 2). QoSWorks and QoSArray are the first QoS solutions that contain all of the tools necessary to optimize both voice and data traffic. QoSWorks and QoSArray provide the ability to:

- Monitor the network to pinpoint voice bandwidth requirements and determine what data traffic to throttle.
- Prioritize voice traffic and control data traffic to maintain the required interval between voice packets.

- Keep voice packets within the delay budget, intelligently trading off packet loss for delay where necessary.
- Prevent the conditions that lead to packet loss for voice traffic.
- Apply the above mechanisms automatically, intelligently, and in combination..

In addition, Sitara's QoSDirector central policy management software can be used to complement the gateway with the network-wide intelligence necessary for call provisioning

Sitara's Key Voice-Related Features

Monitoring

One of the most difficult challenges in protecting voice traffic is obtaining precise information on bandwidth consumption for both voice and data traffic. Using wire-speed classification, QoSWorks or QoSArray provide a real-time window into all of the traffic traversing the network. User-friendly charts and reports make it easy to see how much bandwidth each user/application is consuming, and how well these users/applications are being served by network resources.

Network managers can use this information to set policies to protect voice traffic while maintaining the proper quality of service levels for the other applications sharing the pipe. The effectiveness of policies in force can be monitored in real time using QoSWorks' or QoSArray's "at a glance" status screens. For example, if voice appears to require more bandwidth than originally predicted, the network manager can adjust the policy in seconds and then evaluate the results to see if the new policy is more effective.

For each group of users and applications within a policy, network managers can set performance thresholds using SNMP mechanisms. These are accessible to QoSDirector (Sitara's central policy management software) and standard network management applications, allowing QoSWorks or QoSArray to react to alarms automatically and update policies accordingly. These monitoring features enable network managers to optimize voice proactively, before callers start complaining about poor voice quality.

Class-Based Queuing

Class-Based Queuing (CBQ) is the most sophisticated form of queuing available, with the flexibility to protect voice from all other types of data traffic. CBQ assigns both priority and bandwidth to ensure the quality of a voice call. Simpler queuing techniques such as priority queuing (PQ) or weighted fair queuing (WFQ) offer either priority or bandwidth control, but not both. Bandwidth management techniques such as TCP rate shaping control only TCP traffic, and only control UDP-based voice indirectly. By their very design these techniques do not provide enough control mechanisms to optimize voice traffic. CBQ, on the other hand, calculates and supplies the different bandwidth and delay budgets for all traffic types carried over both TCP and UDP, making it the ideal tool for ensuring high-quality voice.

CBQ provides the ability to:

- **Prioritize And Allocate Bandwidth Among Different Types of UDP Traffic**

Voice traffic is carried using UDP, the same protocol that is used for applications like audio and video streaming. However, whereas voice consumes relatively little bandwidth and is well-behaved, audio and video streaming consume a great deal of bandwidth and can adversely impact data traffic. Therefore, network managers usually seek to protect data traffic from audio and video streaming, and VoIP from data traffic. In a network running multiple types of UDP traffic, CBQ allows the network manager to prioritize voice running over UDP while controlling other UDP-based traffic.

- **Protect Voice and Signaling Traffic**

It is critical for the quality of the call to protect not only the call itself, but also the call setup (signaling) traffic which is carried by the TCP protocol. TCP and UDP have different characteristics. Signaling and voice traffic have very different bandwidth requirements as well. In addition, signaling traffic must be prioritized over the voice call and all other traffic, as signaling between the gatekeeper and the endpoints is critical to ensure the timely setup and tear down of VoIP connections. Voice traffic should be prioritized just below the signaling traffic, but higher than all other traffic types. CBQ provides the ability to accomplish both of these tasks.

- **Dynamic Optimization of Different Voice Codecs**

CBQ differentiates between the codecs associated with each VoIP conversation, i.e., higher-bit-rate codecs used for toll-quality (high-quality) voice and lower-bit-rate codecs used for sub-toll-quality voice. CBQ can calculate and apply the different bandwidth and delay budgets required for different codecs.

Session Bandwidth

Session bandwidth is important because voice calls cannot be managed in aggregate, as other data traffic such as Web browsing can. Rather, each voice call must be handled as a single flow, with its own unique characteristics. Otherwise, either bandwidth will be wasted or the voice calls will interfere with each other and get starved. Session bandwidth control guarantees that each voice call will get the rate that is expected.

Packet Size Optimization

Controlling the packet size of lower-priority traffic that could be queued in front of voice packets is a crucial mechanism for

minimizing jitter. There are actually two different ways to control packet size: fragmentation and controlling packet size at the source. QoSWorks' or QoSArray's packet-size optimization feature controls packet size at the source, a method that has numerous advantages over fragmentation.

Fragmentation involves chopping up big IP packets into smaller IP packets. The disadvantage with fragmentation is that the routers downstream have to route dozens of smaller packets instead of one packet, and the receiver has to reassemble all the little packets. In addition, products that do fragmentation don't include the sophistication to specify how and when to fragment.

Controlling packet size at the source (e.g., the e-mail server, the HTTP server, the FTP server) is far more efficient because the source machine never sends packets larger than a specified size. With packet size optimization, the network manager is essentially optimizing the network for voice by specifying, for example, that the source devices never put more than 500 bytes of information in an IP packet. In this way, the network is always more balanced between voice and data (at Sitara, we refer to this as making a data network "voice ready").

Maximum Queuing Delay

Maximum queuing delay allows network managers to limit the amount of wait time for a voice packet in any queue, regardless of how much traffic is queued up. By providing this additional QoS mechanism, network managers prevent voice packets from queuing up in a long line (due to some transitory condition where the packets are queued up elsewhere for a short time and arrive at the QoS device all at once) and exceeding the delay budget.

Because it is preferable, the maximum queue delay will allow the QoS device to drop voice packets at the end of the queue in order to preserve the integrity and continuity of the voice calls. This feature allows for graceful recovery of the call after a transient impairment such as the one mentioned above. This also maintains jitter buffer tolerances on the receiver side.

"WAN-Aware" Call Provisioning

Sitara's QoSWorks or QoSArray and QoSDirector products contain several features that complement the gatekeeper, which usually sits on the LAN, by bringing an awareness of WAN conditions to the call provisioning process.

Admission Control with QoSWorks or QoSArray

Admission control traditionally is an admit/don't admit capability; but with VoIP, admission control includes the signaling intelligence to check the number of voice calls being initiated against the current capacity of the network and then signal the gatekeeper to admit the call or not. Without admission control it is possible that when the new call is admitted, existing calls are adversely affected because the current capacity of the network is insufficient to sustain all the calls. This sort of uncontrolled admission has very severe implications since conversations that were perfectly normal before this new call are all uniformly wiped out.

Central Policy Management with QoSDirector

It can be complicated and time-consuming to figure out how much bandwidth to configure between a dozen or more branches and headquarters. There may be an average of two simultaneous calls going to branch one, four to branch five, and six to branch six. The network manager has to figure out the bandwidth requirements and policies for each branch and the bandwidth required at headquarters to handle the total volume of calls.

QoSDirector can do these calculations automatically, including adding in the statistical probabilities that all calls will be active at the same time. This ensures that network managers don't under-provision bandwidth (which would result in dropped voice calls), or over-provision bandwidth (which would be wasteful).

QoSDirector also makes it more practical to distribute VoIP policies across an enterprise network by allowing the network manager to create template policies for different branches, and then simply apply the same template to all branches that have the same VoIP requirements (see Figure 3).

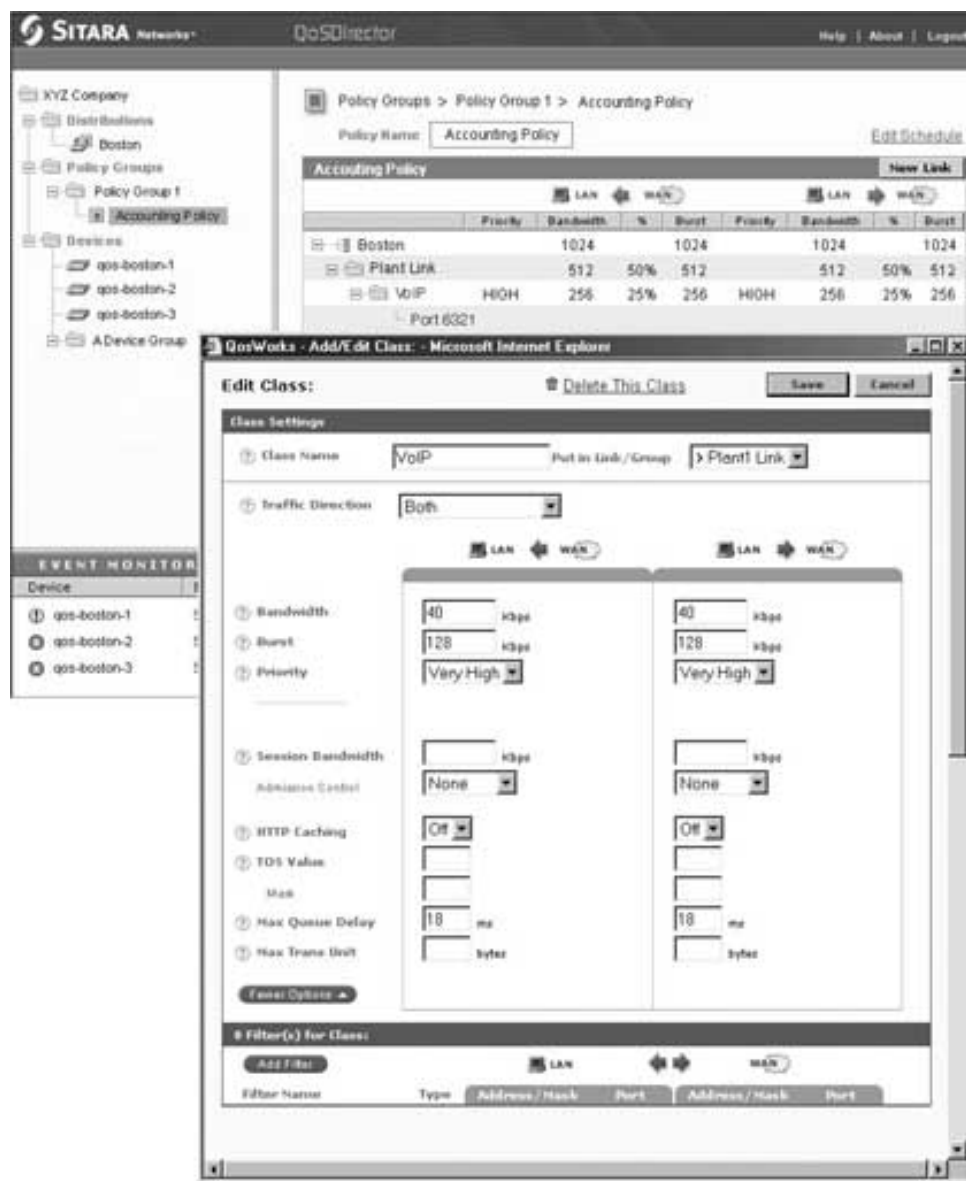


Figure 3. QoSDirector, Sitara's central policy management software, facilitates VoIP across the enterprise. Sitara's QoSDirector central policy management software enables network managers to set up templates to protect voice traffic and distribute them quickly and reliably to all QoSWorks or QoSArray boxes deployed in the network.

In addition, QoSDirector provides the ability to change policies on the fly. For example, if the user needs to add more bandwidth to support a conference call between 4 PM and 5PM on Tuesday, the network manager can set the policy in QoSDirector and it will automatically change that policy on the QoSWorks or QoSArray.

Perhaps most importantly, QoSDirector provides instantaneous feedback, using unsolicited traps, about how policies are performing. For example, if the network manager makes a mistake and provisions a branch for only three calls and that branch is getting four calls at a time, QoSDirector sends an alert that the branch is over-subscribed. Without this automated monitoring and reporting on events that fall above or below thresholds, support staff would have to manually poll each box in the network on an ongoing basis.

Guaranteed Quality Means Guaranteed Returns

Everyone has been talking about the benefits of VoIP for so long that it is tempting to think that VoIP has arrived. While the technology has certainly been embraced by companies because of its numerous compelling benefits, it hasn't really performed to anyone's satisfaction. The irony is that it isn't packetized voice that's the problem ...it's the unruly conditions on the data network that need to be controlled in order for voice to blend seamlessly into the data world. That is where a complete QoS solution, with all of the mechanisms described above, is required.

Copyright © 1999-2001, Sitara Networks Inc. All rights reserved.

Sitara Networks, Inc. Sitara, the Sitara Logo and QoSWorks are registered trademarks and QoSArray and QoSDirector are trademarks of Sitara Networks, Inc. All other trademarks are property of their respective holders.