

Scaling Quality of Service in the Enterprise with Quality of Service Policy Manager 2.1(1)



Executive Summary

As applications such as voice and video appear on converging networks, the need for Quality of Service (QoS) from nearly every device in the network becomes crucial. As a result, in large networks the provisioning and administration of QoS quickly become daunting and complex tasks.

Cisco Quality of Service Policy Manager (QPM) is designed to provision QoS for voice, video, and data networks and has many useful features to facilitate scaling. A main feature is the set of QPM 2.1 IP Telephony templates, which embed the Cisco QoS recommendations for voice into predefined policy device-groups. This simplifies an administrator's task to simply assigning the devices and interfaces to the corresponding group and clicking on the Deploy button. Furthermore, these templates can be tailored to customer requirements. Such modifications are included in the test scenarios detailed in this paper.

Additionally, QPM offers features such as rollback (to quickly undo changes that have been pushed out to network devices) and also device verification (which allows an administrator to determine quickly whether the deployed QoS configurations have been tampered with).

The purpose of this paper is to answer quantitatively the question: How well does QPM scale?

To answer this, a scaling metric must first be defined. The metric used in this paper is the number of Cisco IP Phones rolled out. Based on this metric, four levels of scaling are defined:

- Small-scale IP Telephony deployments—fewer than 500 IP Phones
- Medium-scale IP Telephony deployments—500 to 2500 IP Phones
- Large-scale IP Telephony deployments—2500 to 10,000 IP Phones
- Very-large-scale IP Telephony deployments—more than 10,000 IP Phones

Test procedures and scenarios provide typical deployment times for these scaled scenarios. The deployment times (within each scale) vary depending principally on the CPU speed of the QPM server used. QPM running on a 1.7-gigahertz (GHz) single-processor system can yield the following deployment times:

- Small-scale IP Telephony QoS deployments—1 minute
- Medium-scale IP Telephony QoS deployments—3 minutes
- Large-scale IP Telephony QoS deployments—35 minutes
- Very-large-scale IP Telephony QoS deployments—45 minutes per 10,000 phones

Recommendations for QPM database strategies are included, as are performance recommendations.



Introduction

Need for QoS Management

A critical enabler for the convergence of voice, video, and data onto a single network is QoS. QoS technologies ensure that latency- and jitter-intolerant applications, such as voice and video, will receive end-to-end priority service. Achieving end-to-end priority services requires QoS functionality from virtually every device in the network.

Cisco Systems, as the recognized leader in converging networks, continually publishes best-practice and design guides to facilitate the convergence process for customers. The recently released *Cisco IP Telephony QoS Design Guide* is an authoritative reference for the details of why and where QoS is required in an IP Telephony network and how QoS is properly configured to achieve these goals.

The *Cisco IP Telephony QoS Design Guide* is available at:

http://www.cisco.com/en/US/products/sw/cscowork/ps2064/prod_technical_reference09186a0080091bcb.html

As the guide indicates, QoS is required from nearly every device in the network. This requirement, coupled with the size and variation found in typical enterprise networks, establishes a strong case for centralized QoS management.

Cisco QPM is the tool of choice for centralizing, provisioning, and deploying QoS to both LAN and WAN network devices for voice, video, and data.

Metrics and Levels of Scaling

The question is often asked: How well does QPM scale? This is difficult to answer because there is no clearly defined or accepted standard metric for reference. QoS deployments incorporate many different factors, including:

- Number of devices provisioned
- Number of interfaces and ports per device
- Number of policies per interface
- Number of commands required per policy

Simply using the number of devices as a reference is misleading because provisioning a 240-port Cisco Catalyst® 6000 device will require more commands than a Cisco 1750 Router with two interfaces. Totaling the number of interfaces is also insufficient because more complex QoS policies might apply to certain critical interfaces within the network. Additionally, the raw number of interfaces and ports within an enterprise's network is not a well-known number and thus wouldn't prove meaningful.

Because the *Cisco IP Telephony QoS Design Guide* clearly defines many of these deployment factors (for instance, number of commands per policy and number of policies per interface) and additionally describes how devices and interfaces interrelate for an IP Telephony deployment, it provides a simple and meaningful metric for referencing scale: the number of IP Phones (or IP Phone ports, in the case of gradual deployment) rolled out.

This metric could be further broken down into the number of IP Phones deployed in the central campus (LAN) and the number of IP Phones deployed at remote sites (primarily WAN). Examining typical customer deployments reveals that the approximate ratio of campus IP Phones to remote-site IP Phones varies between 65 and 80 percent.



Using the number of IP Phones deployed as a metric, QoS for IP Telephony deployments can be categorized into four distinct levels of scale:

- Small-scale deployments—fewer than 500 IP Phones
- Medium-scale deployments—500 to 2500 IP Phones
- Large-scale deployments—2500 to 10,000 IP Phones
- Very-large-scale deployments—more than 10,000 IP Phones

This white paper presents recommended QoS deployment strategies via QPM with performance results. Performance is measured in the time required to complete each operation for each level of scale.

Cisco QPM 2.1(1) Features for Scaling

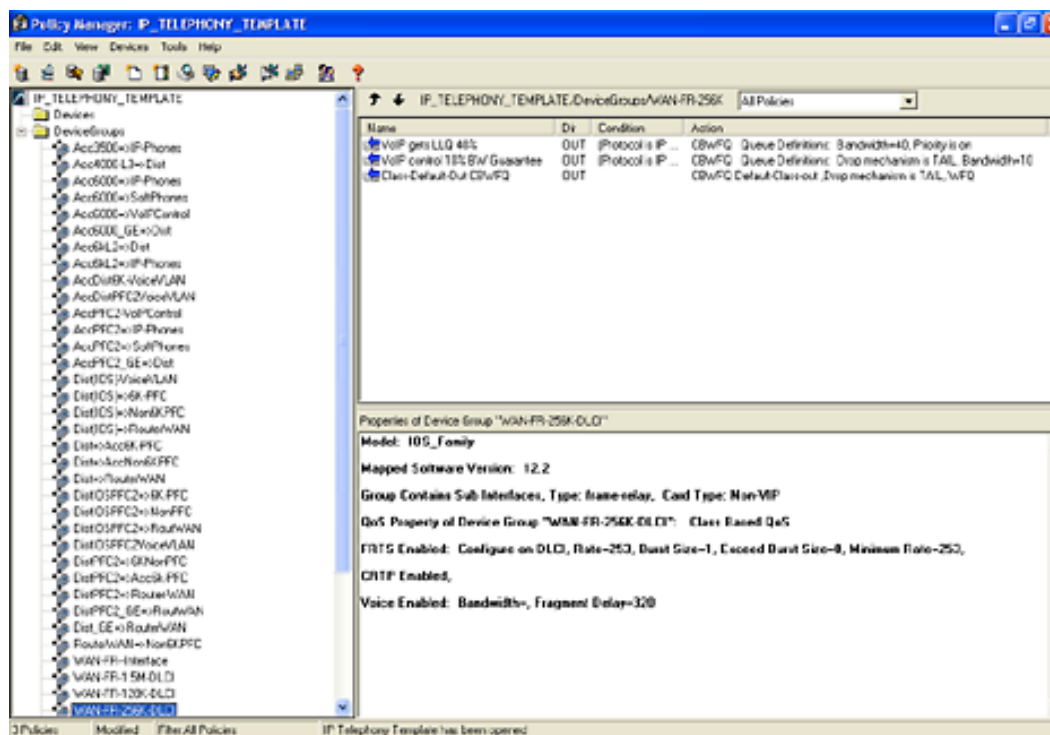
IP Telephony Templates

Although the *Cisco IP Telephony QoS Design Guide* is an excellent resource, it is long (242 pages in PDF). Administrators would need to read the guide thoroughly to understand the QoS features and commands and then correlate these recommendations to their enterprises. However, many administrators do not have the time.

To simplify and expedite the deployment of the recommended QoS features, Cisco has embedded the knowledge contained in the Cisco IP Telephony QoS Design Guide into predefined policy templates that are an integral part of QPM (Figure 1).

These templates contain all the IP Telephony QoS recommendations for classification, trust, queuing, shaping, and fragmentation and correspond to the type and role of the device or interface in the network.

Figure 1 IP Telephony Templates





For instance, the following QoS functionality is recommended for all access-layer Catalyst 6000 ports that are connected to IP Phones:

Trust State—Trust Class of Service (CoS) (trusts the 802.1p CoS value marked by IP Phone)

Trust-Ext State—Untrusted (do not trust 802.1p CoS from anything beyond the IP Phone)

QoS Style—Virtual LAN (VLAN)-Based QoS (access lists are bound to VLAN instead of ports)

Whenever a port is added to the IP Telephony template device group labeled Acc6000 —> IP-Phones, then these policies will be automatically applied, resulting in the following commands being deployed to the device:

```
cat6k-access>(enable) set port qos 5/48 trust-ext untrusted
cat6k-access>(enable) set port qos 5/48 trust trust-cos
cat6k-access>(enable) set port qos 5/48 vlan-based
cat6k-access>(enable) set qos acl ip ACL_IP-PHONES trust-cos ip any any
cat6k-access>(enable) commit qos acl all
cat6k-access>(enable) set qos acl map ACL_IP-PHONES 110
```

WAN QoS is even more complex. For example, to provision QoS on a 256-kbps WAN link with Frame Relay traffic shaping (FRTS) enabled on a sub-interface, the following settings are recommended:

Table 1

QoS Mechanism	Value	Description
FRTS CIR:	256000 kbps	Committed Information Rate
FRTS Bc:	1000 kbps	Committed Burst Rate (a low value results in less delay between transmission intervals)
FRTS Be:	0 kbps	Excess Burst Rate (no "credits" to permit traffic exceeding the burst to be transmitted also)
cRTP		IP RTP Header Compression (to reduce bandwidth and delay by shortening the IP Routing Update Protocol [RTP] header from 40 bytes to 2-4 bytes)
FRF.12:	320 bytes	Frame Relay Fragmentation (To minimize serialization delay, all packets on this link are limited in size to 320 bytes, resulting in a maximum serialization delay at this router of 10 milliseconds.)

Added to these Frame Relay-specific mechanisms are the more general Modular QoS Command-Line Interface (MQC) commands required to provision voice over IP (VoIP) on a WAN link:

- LLQ—Low-Latency Queuing (gives voice a strict-priority service)
- CBWFQ—Class-Based Weighted Fair Queuing (allocates bandwidth for VoIP control traffic, such as Skinny, H.323, and Media Gateway Control Protocol [MGCP] and also shares remaining bandwidth evenly, according to IP Precedence/DSCP levels of flows)



The entire set of commands listed below is deployed to a device added to the FR-256-DLCI device group within the IP Telephony QoS template.

```
class-map VoIP-RTP
    match access-group 100
class-map VoIP-Control
    match access-group 101
!
policy-map QoS-Policy-256k
    class VoIP-RTP
        priority 100
    class VoIP-Control
        bandwidth 8
    class class-default
        fair-queue
!
interface Serial1
    no ip address
    encapsulation frame-relay
    load-interval 30
    frame-relay traffic-shaping
!
interface Serial1.71 point-to-point
    bandwidth 256
    ip address 10.1.71.1 255.255.255.0
    frame-relay interface-dlci 71
    class VoIP-256kbs
!
map-class frame-relay VoIP-256kbs
    frame-relay cir 256000
    frame-relay bc 1000
    frame-relay be 0
    frame-relay mincir 256000
    no frame-relay adaptive-shaping
    service-policy output QoS-Policy-256k
    frame-relay fragment 320
!
access-list 100 permit ip any any precedence 5
access-list 100 permit ip any any dscp ef
access-list 101 permit ip any any precedence 3
access-list 101 permit ip any any dscp 26
```

As these examples illustrate, provisioning such commands throughout the network quickly becomes a challenging, complex, and time-consuming task.

However, with QPM, all an administrator has to do is add or import devices to QPM, put the right interfaces into the corresponding device groups in the IP Telephony template, save the database, and then deploy the policy database. For more details, refer to:

For additional details refer to the Cisco IP Telephony QoS Guide at:

http://www.cisco.com/en/US/products/sw/cscowork/ps2064/products_user_guide_chapter09186a008007ff73.html



Case Study: How Much Time Can QPM IP Telephony Templates Save?

This example illustrates the time-saving potential of QPM and its IP Telephony templates in a typical deployment:

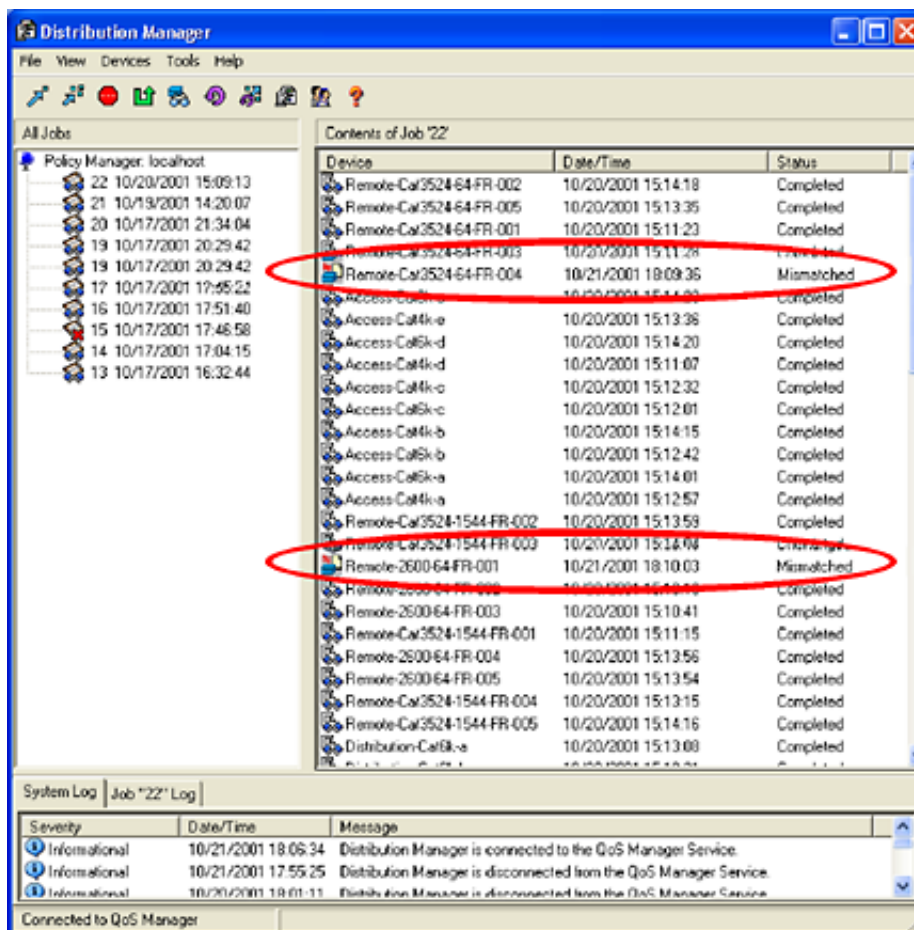
Recently, a consultant was asked to provision QoS for a 400-IP Phone deployment at a northern California college. The consultant had to study the *Cisco IP Telephony QoS Design Guide* in detail and then went about provisioning QoS for this network by hand. This phase of the project took him two weeks to complete. When QPM 2.1 was released with the IP Telephony templates, QoS was redeployed on this network in less than 2 hours. This included installing QPM, and interface assignment to the correct IP Telephony device groups; the actual deployment time was less than 5 minutes. Additionally, QPM picked up some errors made during the original deployment and presented these in a report, providing the option of leaving the configurations intact or correcting the configurations. The correction process was chosen and took less than 1 minute to complete.

Device Rollback and Verification

QPM offers important mechanisms to facilitate large-scale QoS rollout. Device rollback allows an administrator to quickly revert to any previous QPM deployment state. This is similar in function to the Undo button in a word processor.

Device verification is an important tool because sometimes QoS configurations are changed without the administrator's consent or knowledge. Verification is an option that QPM presents to allow for a quick check of the configurations to ensure that they are all intact as originally deployed (Figure 2). If the configurations are not the same, then a mismatch is reported, and the correct commands can be deployed quickly.

Figure 2 Device Verification Quickly Detects Mismatched Configurations





Deployment Target Options

QPM offers the options of sending the deployment commands directly to the device or to a log file (for Trivial File Transfer Protocol [TFTP] or other deployment tool distribution), or to both. These options make it easy to integrate QPM into enterprises that already have their own deployment mechanisms tailored to their specific needs.

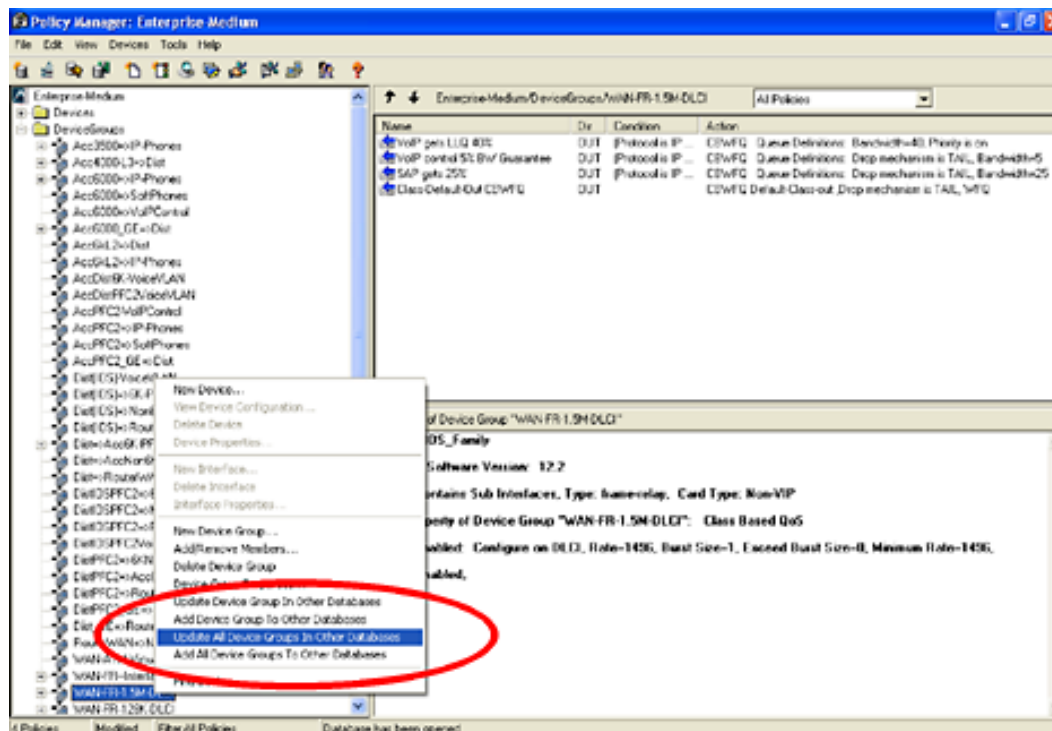
Cross-Database Update

As this paper will show, some scales of deployment are administered better by using multiple QPM databases.

When multiple databases are used, a change in one QPM database is confined to that database only. To prevent having to make the identical changes to every database being administered, QPM provides a cross-database update function. When a user needs to modify a QPM database, he has the option of updating this change on some (or all) of the other databases being administered.

For instance, in the scenarios described in this paper, initially only QoS for IP Telephony is rolled out. Then the enterprise decides to expand its QoS provisioning to include a mission-critical data application. This mission-critical Enterprise Resource Planning (ERP) application is to be colored at both the server and client ends of the network, in addition to being guaranteed bandwidth over the WAN. To make these changes, an administrator only has to manually update one QPM database and then use the cross-database update function to push these new modifications to all other databases (Figure 3).

Figure 3 An Administrator Can Update Several QPM Databases with a Single Click



For more details about QPM rollback, device verification, output target options and cross-database update function, refer to the *QPM 2.1 User Guide*:

http://www.cisco.com/en/US/products/sw/cscowork/ps2064/products_user_guide_book09186a008007ff23.html

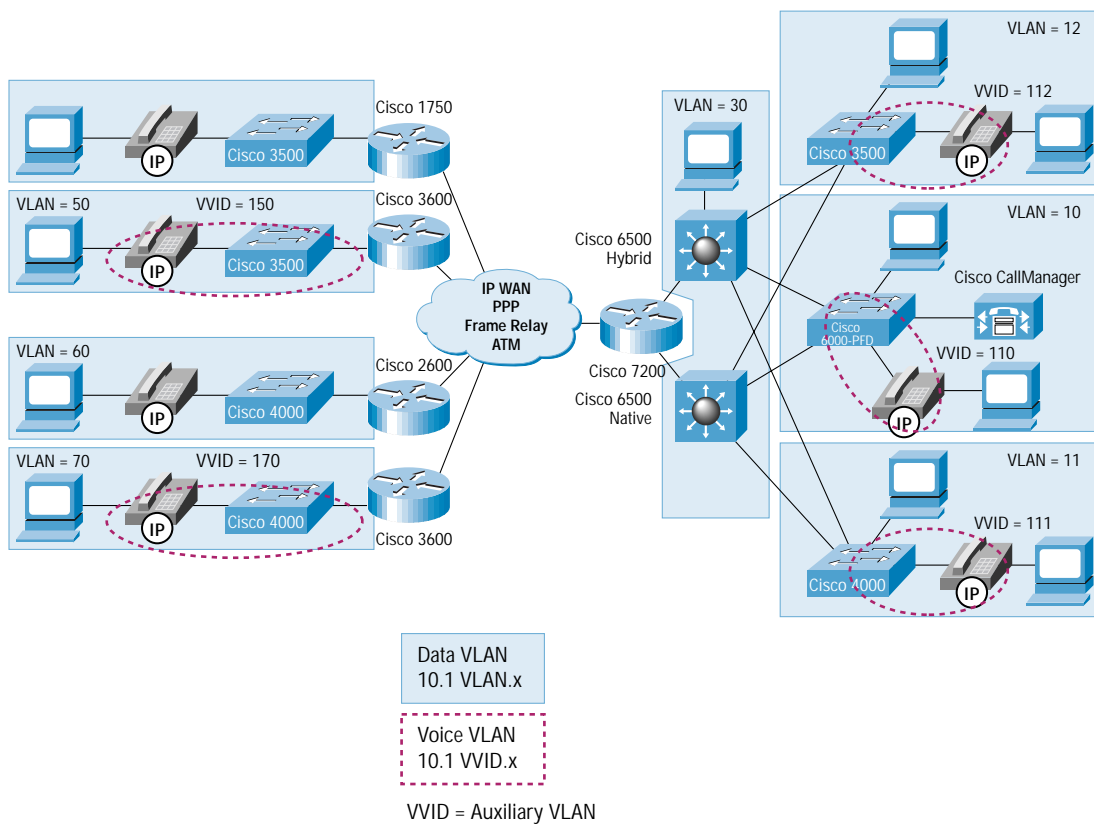


Test Environment and Assumptions

The general network model for IP Telephony, taken from the *Cisco IP Telephony QoS Design Guide*, is shown in Figure 4.

Network

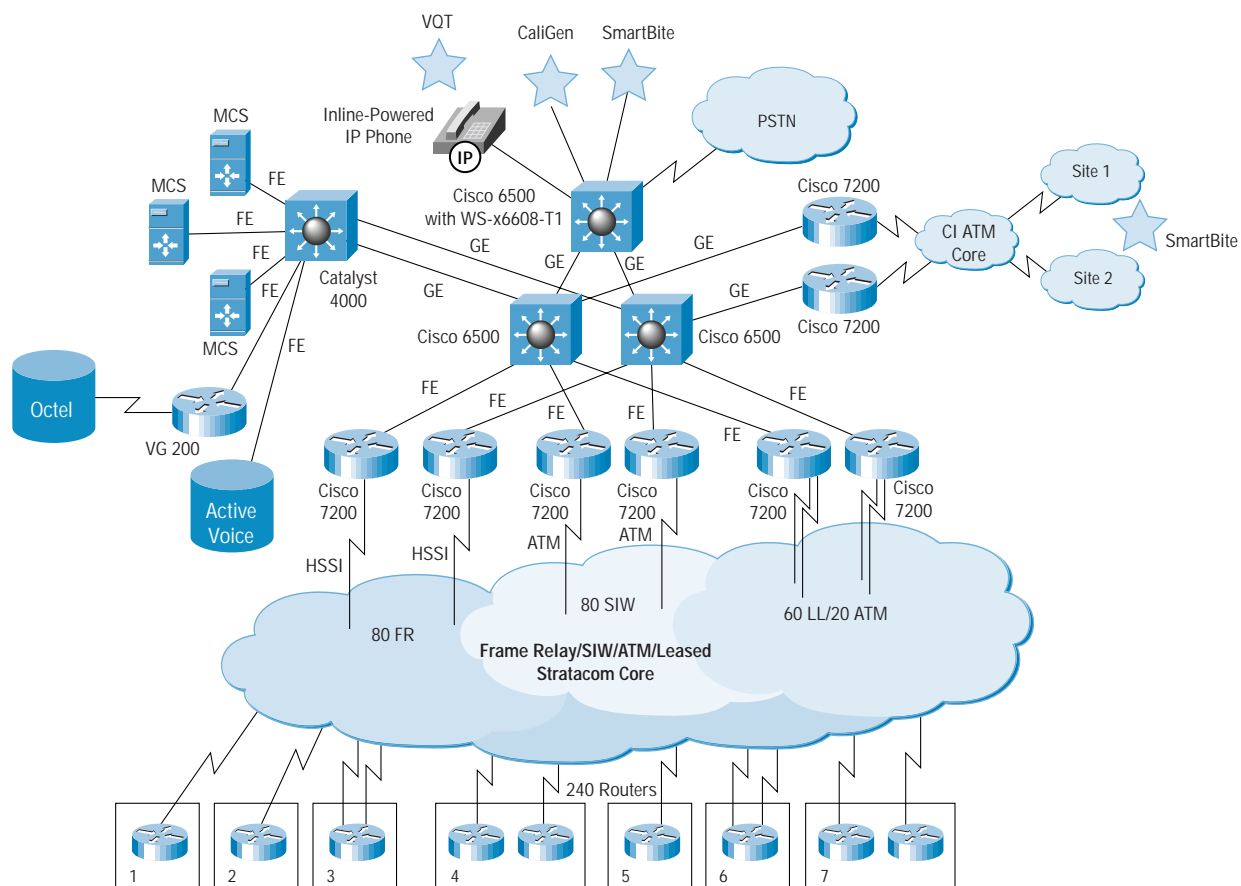
Figure 4 General Network Model for IP Telephony



From this model, the Cisco Enterprise Solutions Engineering (ESE) team built a 250-plus device lab in Research Triangle Park, North Carolina, for large-scale testing (Figure 5). QoS and QPM were both tested extensively in this lab.



Figure 5 Large-Scale ESE IP Telephony Lab Network



QPM testing in the ESE labs provided valuable insight into how well QPM scales. However, because some large enterprises are engaged in IP Telephony deployments well beyond even the scale of this ESE lab (one of the largest Cisco labs), another approach to large-scale QPM testing is required.

QoS Manager-Simulator

QPM engineers in Herzliya, Israel, developed a special version of QPM that simulates deployments to network devices without actually requiring these devices to be online. Deployment is simulated by gathering Simple Network Management Protocol (SNMP) information and configurations from real network devices, and then passing these onto QPM for processing. This tool was tested against the shipping version of QPM in the ESE labs to ascertain its performance against the real version. Tests indicated that the simulator tool could simulate deployment metrics and times within 25 percent accuracy of the shipping version of QPM.

The main advantage of the simulator is that the SNMP and configuration files gathered from the ESE network devices could be modified and cloned to represent networks several times larger than the ESE testbed, and then QPM could be tested with relative precision against these new, simulated, very-large-scale networks.



There are two general types of operations that QPM performs: database operations and network operations.

- Database Operations—includes opening, saving, modifying, cross-database updating, and closing QPM databases (simulator tool performance within 2 percent of shipping version of QPM for database operations)
- Network Operations—includes CiscoWorks Resource Management Essentials (RME) device importing, policy deployment, configuration verification, and rollback (simulator tool performance within 25 percent of shipping version of QPM for network operations)

These factors all affect policy deployment:

- Speed of links
- Congestion of links
- Routing updates
- Network-device CPU load or state
- Administration contention (that is, another party simultaneously administering the device via Telnet)

By using a simulator, these values can be held to a minimal constant, thus giving a clearer measurement of the QPM component of policy deployment. However, in an actual policy rollout, these factors should be considered.

Servers

QPM minimal system requirements call for a 266-MHz processor with 64 MB RAM and a 50-MB hard disk. At the time of testing, such a machine was already considered a dinosaur and hardly worthy for such an important network management application. As the tests results will show, the critical component to QPM deployment is CPU speed; therefore, in these testing scenarios, the following machines are used:

- 500-MHz Intel Pentium III IBM running Microsoft Windows 2000 Server + Service Pack 1
- 1-GHz Intel Pentium IV IBM running Microsoft Windows 2000 Server + Service Pack 1
- Dual 1-GHz Intel Pentium IV IBM running Microsoft Windows 2000 Server + Service Pack 1
- 1.7-GHz Intel Pentium IV IBM running Microsoft Windows 2000 Server + Service Pack 1

Scenarios

The scenarios were as follows: each scale of enterprise (small, medium, large, very large) would be provisioned for IP Telephony QoS using QPM on each of the servers used in these tests. These scenarios would include not only deploying the QoS but also rolling back the deployment and performing a worst-case scenario device verification.

Then, the enterprise is reprovisioned to support QoS not only for IP Telephony but also to protect a mission-critical ERP application. (In this case the application is SAP, identified by the well-known TCP port 3200.) The modifications include these changes:

- SAP traffic is marked to Differentiated Services Code Point (DSCP) 28 (AF32) by the servers in the campus data center to mark *server-to-client* traffic.
- SAP traffic is marked to DSCP 28 (AF32) by every single Catalyst 6000 port connected to an IP Phone or PC to mark *client-to-server* traffic.
- SAP traffic is marked to DSCP 28 (AF32) by the Fast Ethernet router interfaces nearest the clients for segments with switches that do not support DSCP marking to mark *client-to-server* traffic.
- SAP traffic is provisioned for CBWFQ with a minimum bandwidth guarantee of 25 percent on all WAN links throughout the enterprise.

These changes are made to the QPM database (and for very-large-scale scenarios cross-database updates are made). When all database operations have been completed, the new policies are deployed to the entire enterprise.



Test Results

Small-Scale IP Telephony Deployments (Fewer Than 500 IP Phones)

Most IP Telephony deployments fall into the small or medium scale. Many of these consist of a single location, where a PBX is being replaced with Cisco CallManager. Few small-scale deployments include remote sites. A representative device inventory for a small-scale IP Telephony enterprise is presented in Table 2:

Table 2 Small-Scale IP Telephony Enterprise Inventory

Device Type	IP Phone Ports	Quantity	Total IP Ports
Catalyst 6000	96	2	192
Catalyst 4000	48	2	96
Catalyst 3524	24	8	192
Total		12	480

Table 3 shows the results of the test scenarios for a small-scale deployment:

Table 3 Results of Small-Scale QoS Deployment Test Scenarios

Operation	500 MHz	1 GHz	Dual 1 GHz	1.7 GHz
Database: Open	0:10.2	0:04.8	0:03.3	0:03.1
Database: Save	0:07.2	0:02.4	0:01.9	0:01.2
Database: Close	0:04.8	0:01.8	0:01.6	0:00.8
RME-Import	1:03.0	0:30.5	0:25.8	0:22.6
IP Telephony QoS Deployment	2:12.5	1:04.2	1:01.2	0:52.8
Rollback	1:57.6	0:58.8	0:52.2	0:45.6
Verification	0:27.0	0:12.4	0:10.8	0:09.6
Database Modification ¹	0:12.2	0:04.8	0:03.6	0:02.4
IP Telephony QoS + ERP Deployment	2:13.8	1:19.8	1:15.6	1:02.8

1. The database modification in this case was adding coloring policies on the Catalyst 6000 IP Phone Ports to mark SAP traffic (TCP Port 3200) to DSCP 28 (AF32).

The results show deployment ranges from 50 seconds to a little more than 2 minutes for IP Telephony QoS for small-scale enterprise deployments. The QPM database size for this scenario was 703 KB.



Medium-Scale IP Telephony Deployments (500 to 2500 IP Phones)

Remote sites are introduced in the medium-scale enterprise scenarios. The approximate distribution of central campus IP Phones to remote-site IP Phones is 80 percent campus and 20 percent remote. Table 4 shows a representative device inventory for a medium-scale IP Telephony deployment:

Table 4 Medium-Scale IP Telephony Enterprise Inventory

Device Type	IP Phone Ports	Quantity	Total IP Ports
Catalyst 6000	240	5	1200
Catalyst 4000	144	5	720
Catalyst 3524 (Campus)	24	5	120
Catalyst 6000 (Distribution)	N/A	2	N/A
Cisco 7200 (Central WAN)	N/A	2	N/A
Cisco 1700 (Remote)	N/A	4	N/A
Cisco 2600 (Remote)	N/A	8	N/A
Cisco 3600 (Remote)	N/A	8	N/A
Catalyst 3524 (Remote)	24	20	480
Total		59	2520

Table 5 shows the results of the test scenarios for a medium-scale deployment:

Table 5 Results of Medium-Scale QoS Deployment Test Scenarios

Operation	500 MHz	1 GHz	Dual 1 GHz	1.7 GHz
Database: Open	1:03.1	0:30.3	0:29.1	0:20.6
Database: Save	1:01.9	0:27.3	0:26.7	0:18.8
Database: Close	0:15.1	0:04.8	0:04.2	0:03.0
RME-Import	3:31.2	1:41.3	1:38.3	1:21.9
IP Telephony QoS Deployment	7:36.4	3:32.4	3:30.0	3:09.7
Rollback	8:25.6	3:40.3	3:37.3	2:59.0
Verification	3:54.9	1:46.8	1:40.1	1:14.6
Database Modification ¹	1:14.6	0:33.3	0:33.9	0:21.2
IP Telephony QoS + ERP Deployment	9:20.2	4:24.0	4:19.1	3:38.5

1. The database modification in this case included adding coloring policies on the Catalyst 6000 IP Phone Ports to mark SAP traffic (TCP Port 3200) to DSCP 28 (AF32), creating a new device group for all Fast Ethernet interfaces for routers on Catalyst 3524/4000 segments to mark SAP traffic to DSCP 28 (AF32), and also to guarantee a minimum of 25 percent (via CBWFQ) for SAP traffic over all WAN links.

The results show deployment ranges from 3 to 8 minutes for IP Telephony QoS for medium-scale enterprise deployments. The QPM database size for this scenario was 2271 KB.



Large-Scale IP Telephony Deployments (2500 to 10,000 IP Phones)

As the scale of the deployment increases, more remote sites are included in the WAN. In this large-scale IP Telephony deployment scenario, the ratio of campus IP Phones to remote IP Phones is 65 percent to 35 percent, respectively. Table 6 shows a representative device inventory of such a rollout:

Table 6 Large-Scale IP Telephony Enterprise Inventory

Device Type	IP Phone Ports	Quantity	Total IP Ports
Catalyst 6000	240	20	4800
Catalyst 4000	144	15	2160
Catalyst 3524 (Campus)	24	25	600
Catalyst 6000 (Distribution)	N/A	5	N/A
Catalyst 6000 (Distribution-IOS)	N/A	5	N/A
Cisco 7200 (Central WAN)	N/A	5	N/A
Cisco 1700	N/A	40	N/A
Cisco 2600	N/A	60	N/A
Cisco 3600	N/A	5	N/A
Catalyst 3524 (Remote)	24	100	2400
Catalyst 4000 (Remote)	48	5	240
Total		285	9960

In these scenarios, the campus devices are separated from the WAN devices to improve manageability. The central WAN routers are included in the WAN database, even though they physically exist in the central site.

Table 7 shows the results of the test scenarios for a large-scale campus deployment:

Table 7 Results of Large-Scale QoS Deployment Test Scenarios, Part 1: Campus

Operation	500 MHz	1 GHz	Dual 1 GHz	1.7 GHz
Database: Open	9:23.1	4:33.6	4:05.2	2:49.8
Database: Save	8:49.7	4:21.3	4:03.2	2:47.3
Database: Close	3:45.6	1:27.1	1:21.0	0:56.9
RME-Import	17:26.5	8:45.8	8:33.6	6:40.1
IP Telephony QoS Deployment	43:15.6	20:10.1	18:59.8	14:38.7
Rollback	50:45.2	22:34.2	20:14.1	16:38.5
Verification	14:37.9	6:40.6	6:35.4	4:15.4
Database Modification ¹	7:50.8	3:54.3	3:44.2	2:52.4
IP Telephony QoS + ERP Deployment	59:45.7	24:01.2	21:39.1	19:44.1

1. The database modification in this case included adding coloring policies on the Catalyst 6000 IP Phone Ports to mark SAP traffic (TCP Port 3200) to DSCP 28 (AF32) and the inclusion of a device group for data center switches to mark SAP server traffic to DSCP 28 (AF32).



The results show deployment ranges from 15 minutes to 44 minutes for IP Telephony QoS for large-scale campus deployments. The QPM database size for this scenario was 7005 KB.

Table 8 shows the results of the test scenarios for a large-scale remote/WAN deployment:

Table 8 Results of Large-Scale QoS Deployment Test Scenarios, Part 2: WAN

Operation	500 MHz	1 GHz	Dual 1 GHz	1.7 GHz
Database: Open	5:12.9	1:33.1	1:28.4	1:04.5
Database: Save	3:26.0	1:17.0	1:09.7	0:40.8
Database: Close	2:23.4	0:39.7	0:41.7	0:27.6
RME-Import	12:56.7	7:25.7	7:19.7	6:46.1
IP Telephony QoS Deployment	52:35.2	24:39.2	21:57.2	19:33.4
Rollback	1:03:45.2	28:53.8	25:49.8	22:44.8
Verification	19:11.8	8:55.5	8:23.6	6:25.1
Database Modification ¹	3:19.3	1:44.7	1:34.2	1:15.2
IP Telephony QoS + ERP Deployment	1:15:56.1	37:51.3	32:22.4	26:39.8

1. The database modification in this case included creating a new device group for all Fast Ethernet interfaces for routers on Cisco Catalyst 3524/4000 segments to mark SAP traffic to DSCP 28 (AF32) and also to guarantee a minimum of 25 percent (via CBWFQ) for SAP traffic over all WAN links.

The results show deployment ranges from 20 to 53 minutes for IP Telephony QoS for large-scale WAN deployments.

The QPM database size for this scenario was 4227 KB. The database size is of interest because even though the WAN database has more devices than the campus database (215 compared to 70) it is only half the size of the campus database. The reason the WAN database is smaller is that only 2615 interfaces/ports require QoS policies in the WAN database, compared with 7620 interfaces/ports requiring QoS policies in the campus database. This smaller database for the WAN reduces performance times on database operations. However, it is important to note that even though there are about three times fewer interfaces/ports being administered by the WAN database, the QoS commands are significantly more complex, and therefore network operations for the WAN take even longer than for the campus database.

Very-Large-Scale IP Telephony Deployments (More Than 10,000 IP Phones)

Very-large-scale scenarios are to be administered by QPM as modular large-scale scenarios. Each large-scale scenario has a separate campus and WAN QPM database. To increase from large-scale to very-large-scale scenarios, separate QPM databases should be created for every 10,000 IP Phones deployed. The ratio of campus IP Phones to WAN IP Phones within the enterprise should be kept consistent. For example, if the number of campus IP Phones is 75 percent of the total, then a separate campus database should be created for every 7500 campus IP Phones, and a separate WAN database should be created for every 2500 IP Phones.

The QPM performance metrics are then simply factors of the large-scale scenarios. The only new elements are the database modification times, for the cross-database update feature will be required to reflect the changes across all campus and WAN databases.



Table 9 shows the results for applying a cross-database update for large campus and WAN databases (detailed in the previous scenario):

Table 9 Results for Cross-Database Updates for Campus and WAN Databases

Operation	500 MHz	1 GHz	Dual 1 GHz	1.7 GHz
Campus Cross-Database Update	8:19.7	4:35.6	4:21.3	3:33.7
WAN Cross-Database Update	4:13.1	2:15.6	1:56.8	1:26.7

Note that these update times reflect the time it takes to update *each* database to which the change is being applied.

Recommendations

QPM Strategy for Scaled QoS Deployments

Small- and medium-scale deployments should be administered using a single QPM database. RME import is optional if fewer than 10 devices are being administered using QPM.

Separate QPM databases are recommended for large-scale IPT deployments: one database for campus (LAN) devices and remote-site (primarily WAN) devices. Central site WAN edge routers should be included in the WAN database (despite their location at the central campus).

Very-large-scale scenarios are to be administered by QPM as modular large-scale scenarios. Each large-scale scenario has a separate campus and WAN QPM database. To increase from large-scale to very-large-scale scenarios, separate QPM databases should be created for every 10,000 IP Phones deployed. The ratio of campus IP Phones to WAN IP Phones within the enterprise should be kept consistent. For example, if the number of campus IP Phones is 75 percent of the total, then a separate campus database should be created for every 7500 campus IP Phones, and a separate WAN database should be created for every 2500 IP Phones.

These recommendations are summarized in Table 10.

Table 10 Recommendations for QoS Scaling Using QPM

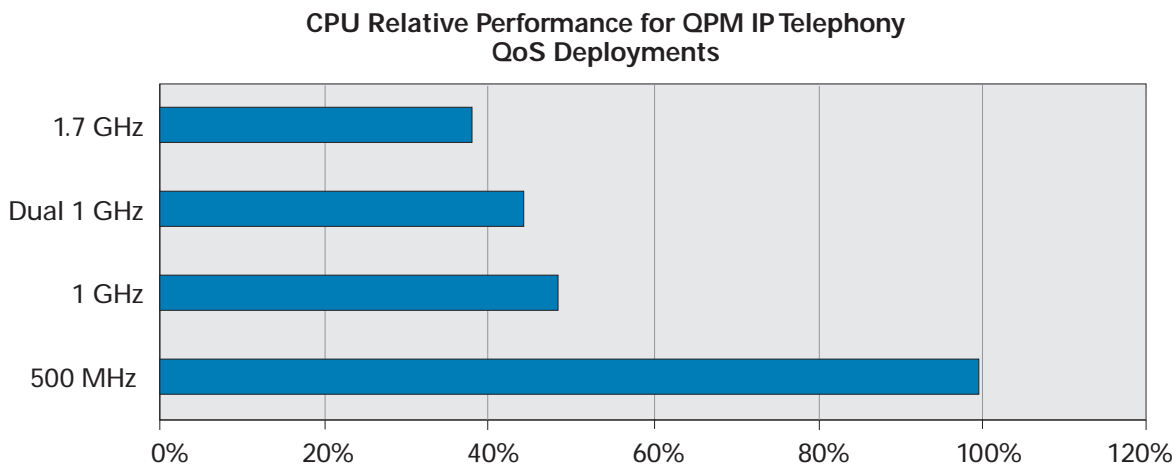
Scale of Deployment	RME Import	QPM Database Strategy	Cross-Database Update
Small Scale (Fewer Than 500 IP Phones)	Optional	Single Enterprise Database	No
Medium Scale (500 to 2500 IP Phones)	Mandatory	Single Enterprise Database	No
Large Scale (2500 to 10,000 IP Phones)	Mandatory	Single Campus DB + Single WAN DB	No
Very Large-Scale (More Than 10,000 IP Phones)	Mandatory	1 Campus Database per 7500 Campus IP Phones	Yes
		1 WAN Database per 2500 Remote IP Phones	Yes



Single-CPU vs. Dual-CPU Servers

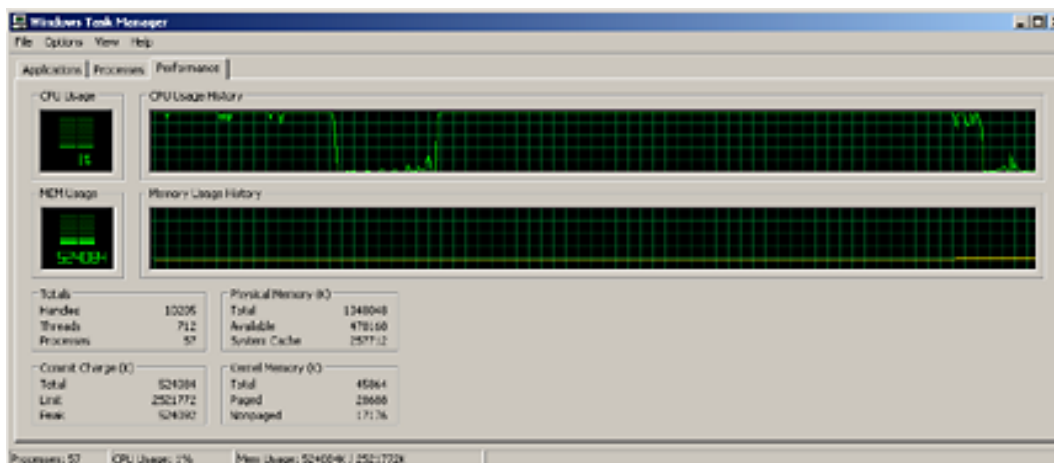
A surprising yet consistent result that surfaced during the course of these tests was the disappointing performance of the dual-CPU server. Rather than reducing the performance time by half, as one might expect, this machine barely reduced time by 5 percent over its identical single 1-GHz processor counterpart. Figure 6 illustrates the overall relative performance of the different CPUs for these operations.

Figure 6 Relative Performance of CPUs Used for Testing



An analysis of the CPU real-time performance reveals why. For instance, the Windows Task Manager provides a real-time graph showing CPU and memory usage. For the 1.7-GHz server, the Task Manager real-time graph is shown in Figure 7, and Figure 8 shows CPU utilization for dual 1-GHz processors.

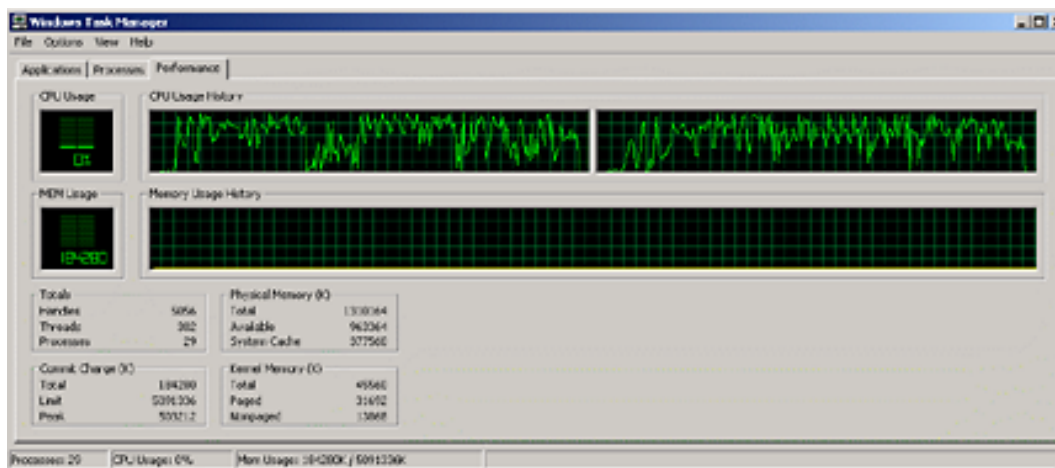
Figure 7 CPU Utilization for 1.7-GHz Single Processor for IP Telephony QoS Deployment



As expected, the CPU maximizes during a computationally-heavy QPM operation (in this case an IPT QoS Deployment); however, the Dual-CPU's Task-Manager graph for the same operation reveals a major difference in CPU utilization.



Figure 8 CPU Utilization for Dual 1-GHz Processor for IP Telephony QoS Deployment



As the graph for each CPU shows, neither CPU is ever being fully utilized. In fact, CPU utilization for IP Telephony QoS deployment averaged a mere 56 percent.

A major inefficiency is apparent in the Windows 2000 Server CPU scheduling algorithms under these scenarios. Therefore, a higher-speed single processor is recommended over dual-processor architectures for QPM deployments.

Modifying the Maximum Number of QPM Deployment Threads

By default, QPM deploys a maximum of 20 threads (20 Telnet sessions to devices) while pushing out QoS policies. This maximum value can be modified by changing the following value in the `cqpm.ini` file. (This file is located by default in the `%SystemRoot%` directory.)

```
[DistributionManager]
```

```
NumberOfThreads=20
```

Although this option does exist, changing this value is not recommended in most scenarios. Exceptions may include small- and medium-scale scenarios with very powerful CPU (more than 2 GHz) QPM servers. As illustrated in Figure 7, QoS policy deployment is a CPU-intensive operation and can maximize CPU usage for a long time with only the default of 20 maximum distribution manager threads.

Scalable QoS Monitoring Technologies

QoS management does not end with QoS policy configuration. Monitoring the effectiveness of the policies is crucial. Monitoring must be conducted immediately after QoS policy deployment (to ensure completeness and functionality) and routinely afterward (to ensure that changing network conditions do not disrupt QoS levels).

It is beyond the scope of this paper to discuss in any depth the technologies that Cisco offers to meet the need for scalable monitoring. These technologies bear mentioning because administrators often simply assume that the QoS is working but rarely verify it.

Cisco scalable QoS monitoring technologies include:

- Class-based QoS Management Information Base (MIB)
- Service Assurance Agent
- NetFlow switching and collecting

It cannot be overemphasized how important monitoring and trending are to QoS management and successful QoS deployments, especially for IP Telephony.



Summary

Small- and medium-scale deployments should be administered using a single QPM database. RME import is optional if fewer than 10 devices are being administered using QPM.

Separate QPM databases are recommended for large-scale IP Telephony deployments: one database for campus (LAN) devices and remote-site (primarily WAN) devices. Central site WAN edge routers should be included in the WAN database (despite being located at the central campus).

Very-large-scale scenarios are to be administered by QPM as modular large-scale scenarios. Each large-scale scenario has a separate campus and WAN QPM database. To increase from large-scale to very-large-scale scenarios, separate QPM databases should be created for every 10,000 IP Phones deployed. The ratio of campus IP Phones to WAN IP Phones within the enterprise should be kept consistent. For example, if the number of campus IP Phones is 75 percent of the total, then a separate campus database should be created for every 7500 campus IP Phones, and a separate WAN database should be created for every 2500 IP Phones. The cross-database update feature should be used when making database modifications in very-large-scale scenarios.

These recommendations are summarized in the Table 11.

Table 11 Recommendations for QoS Scaling Using QPM

Scale of Deployment	RME Import	QPM Database Strategy	Cross-Database Update
Small Scale (Fewer Than 500 IP Phones)	Optional	Single Enterprise Database	No
Medium Scale (500 to 2500 IP Phones)	Mandatory	Single Enterprise Database	No
Large Scale (2500 to 10,000 IP Phones)	Mandatory	Single Campus DB + Single WAN DB	No
Very Large Scale (More Than 10000 IP Phones)	Mandatory	1 Campus Database per 7500 Campus IP Phones	Yes
		1 WAN Database per 2500 Remote IP Phones	Yes

Additionally, higher-speed CPU servers are recommended over dual-CPU systems.

QPM has the ability to increase the number of simultaneous deployments (from the default of 20), but this is not recommended in most scenarios because of the ordinarily high-CPU utilization required by a large QoS deployment using QPM. The only exceptions to this case may be small- and medium-scale scenarios with very fast CPU machines (more than 2 GHz).

QoS management does not end with deploying QoS commands to network devices. Post-deployment monitoring is vital to ensure consistency and completeness in the QoS rollout. Additionally, long-term monitoring is necessary to ensure that changing network conditions and applications do not disrupt provisioned QoS levels.



Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems Europe
11, Rue Camille Desmoulins
92782 Issy-les-Moulineaux
Cedex 9
France
www-europe.cisco.com
Tel: 33 1 58 04 60 00
Fax: 33 1 58 04 61 00

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems Australia, Pty., Ltd
Level 9, 80 Pacific Highway
P.O. Box 469
North Sydney
NSW 2060 Australia
www.cisco.com
Tel: +61 2 8448 7100
Fax: +61 2 9957 4350

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the

Cisco Web site at www.cisco.com/go/offices

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia
Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia
Ireland • Israel • Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru
Philippines • Poland • Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa
Spain • Sweden • Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Copyright © 2001, Cisco Systems, Inc. All rights reserved. Catalyst, Cisco, Cisco IOS, Cisco Systems, and the Cisco Systems logo are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and certain other countries.

All other trademarks mentioned in this document or Web site are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0108R)