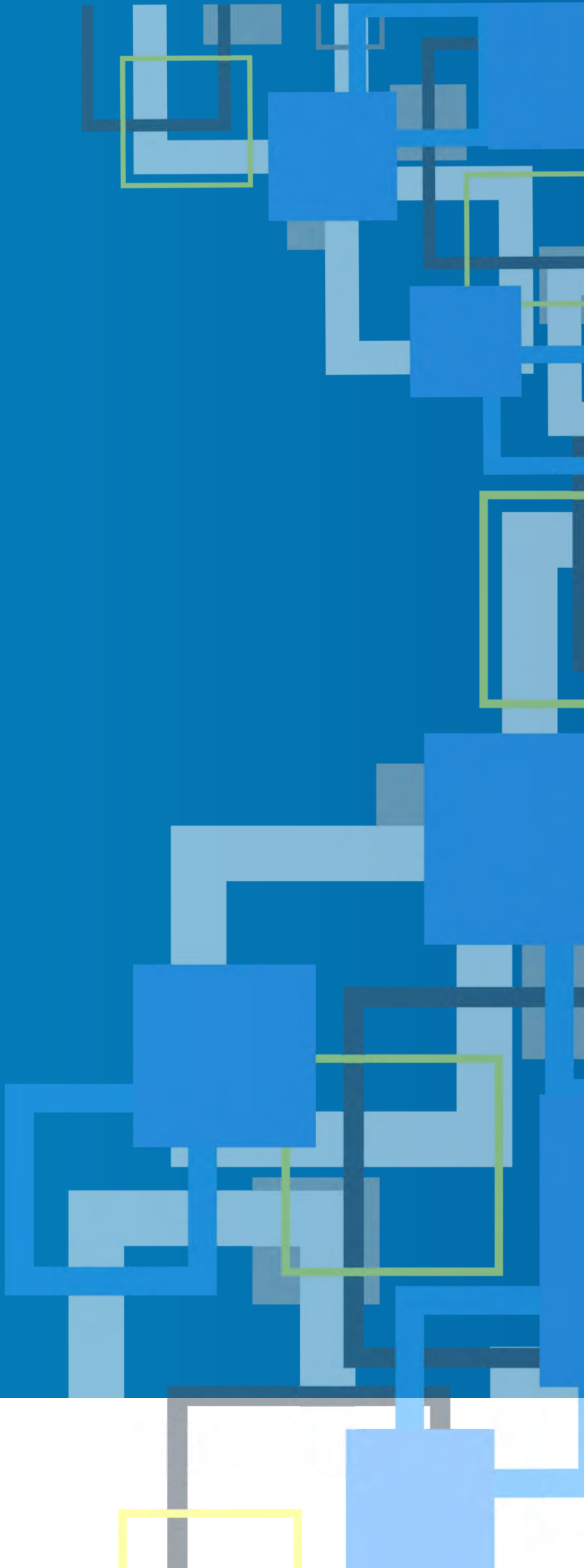
Cisco QoS Handbook

2nd Edition



LiveAction®

Contents

01

How Does QoS Work?

Quality of Service (QoS) is a suite of technologies utilized to manage bandwidth usage as data crosses computer networks. Its most common use is for the protection of real-time and high priority data applications in converged networks. All network infrastructure devices have limits on the amount of traffic that can flow through them. In packet and frame switched networks, this results in data being delayed or dropped during times of congestion. Quality of Service (QoS) management is the collection of mechanisms that control how traffic is prioritized and handled during these times. QoS technologies, or tools, each have specific rolls that are used in conjunction with one another to build end-to-end network QoS policies.

The two most common QoS tools used to handle traffic are classification and queuing. Classification identifies and marks traffic to ensure network devices know how to identify and priorities data as it traverses a network. Queues are buffers in devices that hold data to be processed. Queues provide bandwidth reservation and prioritization of traffic as it enters or leaves a network device. If the queues are not emptied (due to higher priority traffic going first), they overflow and drop traffic.

Policing and shaping are also commonly used QoS technologies that limit the bandwidth utilized by administratively defined traffic types. Policing enforces bandwidth to a specified limit. If applications try to use more bandwidth than they are allocated, their traffic will be remarked or dropped. Shaping defines a software set limit on the transmission bandwidth rate of a data class. If more traffic needs to be sent than the shaped limit allows, the excess will be buffered. This buffer can then utilize queuing to priorities data as it leaves the buffer.

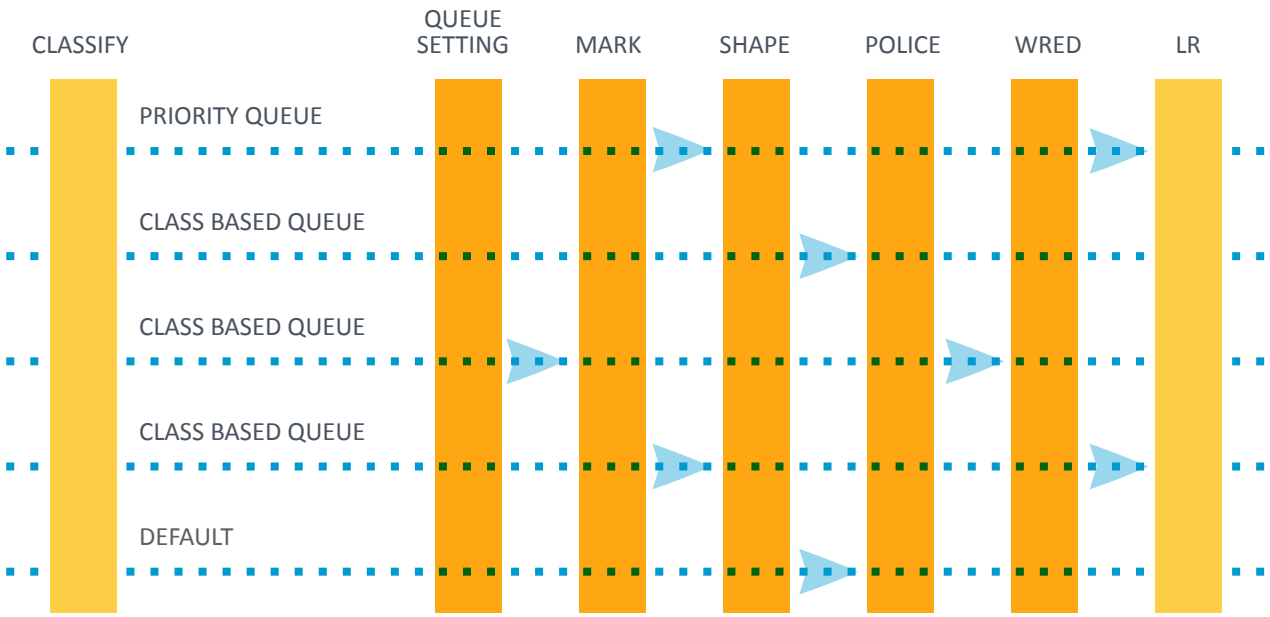
01

How Does QoS Work?

The WRED (Weighted Random Early Discard) technology, provides a congestion avoidance mechanism that will drop lower priority TCP data to attempt to protect higher priority data from the adverse effects of congestion.

Finally, link-specific fragmentation and compression tools are used on lower bandwidth WANs to ensure real-time applications do not suffer from high jitter and delay. ■

Diagram: Packet flow through a typical QoS policy



02

Classification & Marking

The first step in the QoS configuration process is to determine an application's packets and classify them using Layer 2 or Layer 3 QoS markings to identify them within the network.

CLASSIFICATION METHODS

There are several ways to classify traffic. Some common methods are:

- Access Control Lists applied to a class-map (IP address, port, etc.)
- Application Type applied to a class map (via NBAR or NBAR2)
- Switch port trust state

Once a packet is identified, it is typically marked near the source of the traffic flow, allowing the packets to receive differentiated treatment throughout the network. Marking is typically done at the Layer 3 IP header using the Differentiated Services Code Point (DSCP). LAN implementations use the Class of Service (COS) field. See Appendix B later in this document for best practices surrounding DSCP marking values.

02

Classification & Marking

There are two paradigms that are generally accepted as best strategies for how to best classify and mark traffic. These are:

1. THE LAN

LAN Switches such as Cisco's Catalyst portfolio perform hardware-based QoS marking at wire speed. Hence, LAN access switches can be the best place to enable marking to take advantage of hardware-based QoS and ensure packets receive proper treatment throughout the network. Although these hardware-based QoS mechanisms do not significantly impact the CPU load on the switch, they often have limited configuration capabilities like consolidated, basic Access Control Lists for classifying traffic. Another consideration with this strategy is the sheer number of LAN interfaces that must be accurately configured to ensure proper classification and marking are working effectively. This often becomes extremely difficult to manage at scale.

2. THE WAN

WAN routers such as Cisco ISRs and ASRs have more QoS matching and marking capabilities. This often makes it easier to accurately ensure traffic is marked as desired. In addition, there are often significantly fewer interfaces that need to be configured correctly to successfully match and mark traffic at the WAN edge. Finally, the WAN edge is where QoS congestion is most likely to occur, hence matching and marking at the WAN edge ensures traffic is properly identified and subsequently prioritized accurately where it is most needed. Those that choose this strategy must keep in mind that by focusing where QoS matters most and is simplest to deploy, the results are practically "good enough" vs. a more complex, technically accurate strategy. ■

03

Queuing

There are several types of queues built into network devices. For Cisco devices, the two most common are priority queues (PQs) and Class-Based Weighted Fair Queuing (CBWFQ). Priority queues are designed for use with packets that require low-latency and low-jitter treatment such as voice or video over IP. PQs will drop any oversubscription during times of congestion. CBWFQ is designed for bulk and transactional traffic that are not as time-, loss-, or jitter-sensitive. Each CBWFQ queue can specify a reserved bandwidth that is guaranteed to be available for use by that class during times of congestion. Finally, queue depth can also be set to ensure traffic bursts can be handled by a queue.

Congestion avoidance mechanisms, such as Weighted Random Early Detection (WRED), can be used to drop packets as the queues fill up in order to throttle back TCP-based applications. There are also settings for class-based queues that can reduce the possibility of congestion. These settings use DSCP markings to determine the priority of packets to be dropped before the queue becomes full. ■

04

Shaping & Policing

Policing involves checking the packet rate for a particular classification of flows to determine if it conforms to specified rate settings (e.g., a packet flow for a certain source could be set to not exceed 50 Kbps). The policer will check to see if there is excess traffic, and if excess traffic exists, it can drop the packets in order to conform to that specified rate.

The policer can also specify thresholds for exceeding as well as violating and can take actions that include marking down DSCP values rather than simply dropping packets. Markdown involves changing the DSCP value in the IP packet based on flow conditions, which changes the priority of the packets as they traverse the network.

Shaping involves delaying packets that exceed a software defined threshold. If more data is trying to be sent than the shaped threshold, packets are held in the buffer until a later time when they can be sent out. This has the effect of smoothing outbursts, but unlike policers, shaping will attempt to not drop packets in the process. Queuing polices can be assigned to the shaped buffer to prioritize applications as they leave the buffer. If there is congestion or a large amount of traffic; however, the buffer may overflow, causing dropped packets. ■

05

Link Efficiency Mechanisms

1. LLFI FOR MLP—
(Link Fragmentation and Interleaving for Multilink PPP)—MLP splits, recombines, and sequences datagrams across multiple logical data links.

Large delay-sensitive datagrams are multilink encapsulated and fragmented into smaller packets to satisfy delay requirements. Small delay-sensitive packets are not encapsulated but are interleaved between these fragments.

2. LFI FOR FRAME RELAY (FRF.12)—
FRF.12 fragments packets that are larger than the settings specified using the frame-relay fragment_ size command.

Frame Relay cannot distinguish between VoIP and data; therefore configure the fragmentation size on the DLCI so voice frames are not fragmented. The fragment size is specified in bytes (default = 53 bytes). Many variables determine the size of the voice packets.

3. CRTP—
Compressed real time protocol (CRTP) reduces in line overhead for multimedia RTP traffic resulting in a corresponding reduction in delay. CRTP is especially beneficial when the RTP payload size is small such as compressed audio payloads of 20 to 50 bytes.

CRTP should be used on any WAN interface where bandwidth is a concern or there is a high portion of RTP traffic. Note that CRTP requires more processing on the routers. ■

06

QoS Deployment Lifecycle

Defining and deploying QoS can be a daunting task for any IT department, even with highly skilled engineers on staff.

In order to be successful and make the process as manageable as possible, QoS deployments should be broken up into discrete sequential stages that can serve as milestones for the project. A typical implementation would follow these stages:

1. PROJECT PLANNING AND BUY-IN

Understand the current and near future QoS needs of the organization as a whole. Choose an appropriate QoS model and then get departmental buy-in before beginning.

2. INVESTIGATION

Discover and document current and future network architecture. This includes:

- Gather snapshots existing QoS policies
- Research the QoS capabilities of the network devices
- Baseline the network with flow monitoring and usage analysis

3. DESIGN

Select a QoS model(s) for the traffic classes and define QoS policies for headquarters and campus LANs, WANs and branch offices, and VPNs.

4. PROOF OF CONCEPT (POC)

Test QoS policies and settings first in a non-production environment using real and synthetic traffic to generate controlled conditions. Test separately for each policy and then with all policies combined.

06

QoS Deployment Lifecycle

5. ITERATIVE DEPLOYMENT CYCLE

Roll out QoS policies in a phased approach, either by sections of the network or by QoS functions (classification then queuing). Confirm changes at each iteration for at least 24 hours before continuing to the next step.

6. ONGOING MONITORING AND ANALYSIS

Perform ongoing monitoring and adjust the policies not just for average daily usage but also for monthly, quarterly, and yearly business cycles. ■

07

Project Planning & Buy-in

The first step to any successful project is a good plan.

This plan involves understanding the QoS needs of the organization, choosing the QoS model that fits these needs, designing the implementation, and finally implementing that design.

A good plan starts with understanding the businesses' current and near future needs. This will help to define the size and scope of the implementation. Perhaps voice services (VoIP) are being implemented now, but video teleconferencing will be rolled out in a year. Knowing this business objective, a design could be chosen that supports both technologies from the start. Another overlooked aspect of planning is getting departmental buy-in. While VoIP deployment may be rolled out companywide, there may be some department specific applications that are considered mission critical and other applications that are planned for deployment next year. It's good to get departmental buy-in and understand their future needs before the QoS project gets started. ■

08

Investigation

If making significant hardware or software changes coincide with any QoS rollouts, make these first before adding complexity with new QoS policies. This is also a good time to snapshot existing QoS policies in case any changes need to be rolled back later in the process. The first step in creating a QoS design is to gain an understanding of the current network architecture. This includes understanding the applications used to support the network, and the traffic running over it. If QoS has not been deployed on the network before, this would be the time to research and understand what QoS capabilities are available in the network devices.

If the network's traffic patterns have not been base-lined, or if there have been recent network changes, now is the time to do so. A NetFlow monitoring application will be needed to gather and present network traffic information for analysis. This information will highlight typical utilization patterns of current network applications and may even help spot problems. This data may even indicate that further network infrastructure changes are needed to support any new bandwidth intensive applications such as video teleconferencing. Remember, QoS does not create bandwidth; it can only optimize the bandwidth available. This document will highlight how this can be accomplished using LiveAction's LiveNX software. ■

09

Design

Once a solid understanding of current and future QoS requirements is gathered, it's time to begin the QoS design.

The design may include network changes such as bandwidth or hardware and software upgrades, but it must include the selection of a QoS model, and the different router and switch policies needed to support the model. Policies will need to be tailored to the different areas of the network. Depending on the network's complexity, this may include access, distribution, core, WAN edge, remote offices, and security networks and devices. QoS policies may also be implemented at the Internet edge services for VPN users if mission-critical or voice applications are accessed remotely.

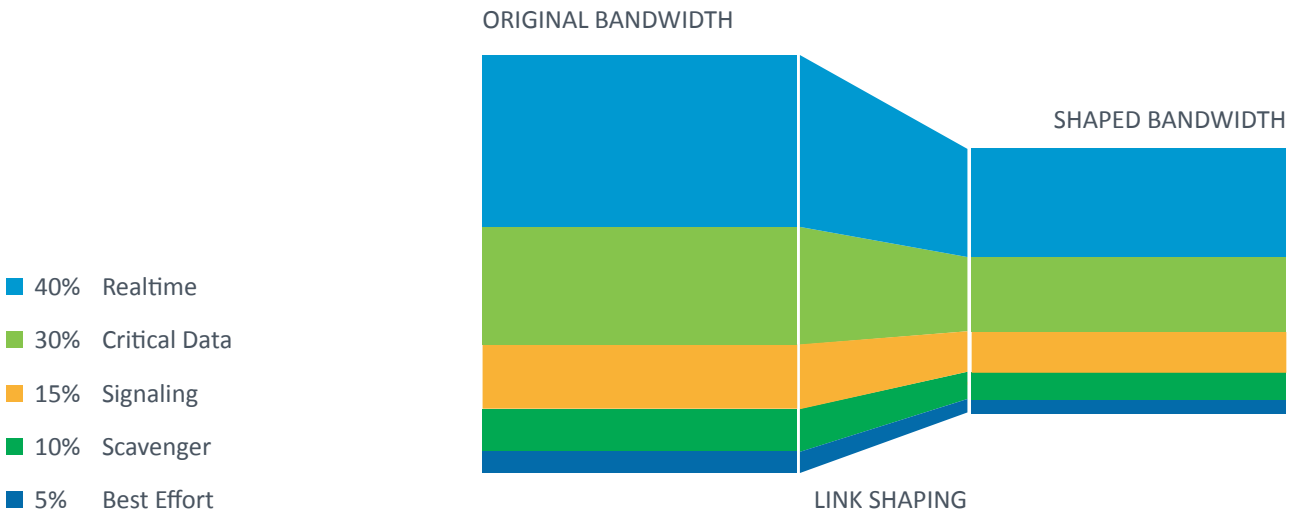
There are several standardized QoS models that can be chosen to support the number of classes required in the network infrastructure. Cisco has defined several models from 3, 4, 5, 8, and 11 classes to the latest 12 class Medianet model. More classes are not always better for the organization due to hardware limitations, MPLS service provider restrictions, or there are not enough IT resources to manage the extra complexity. Many networks implement multiple class models for various sections of their environment. For initial QoS deployment, a smaller 3, 4 or 5 class model can be used to simplify the process. The model can easily be expanded over time as additional applications and requirements arise.

QoS policies should be designed from the high-level requirements gathered in the project planning stage. Once the requirements are understood, they can be translated to QoS best practices for the various application types. See Appendix A – QoS Best Practices for a reference when designing QoS policies.

09

Design

Diagram: QoS Shaping



If the network design requires high-priority traffic to traverse WAN links, consider how QoS will be implemented under these circumstances. This may mean upgrading service levels with the provider to support multiple levels of priority, and then selecting how applications with map to the providers configurations. If the service provider has no QoS policy on their network, they will disregard any QoS markings on their network. Traffic can still be prioritized as it leaves the corporate network, but there is no guarantee how it will be treated.

For many WAN environments, there will be a link-speed mismatch between a WAN link's physical capabilities and the Committed Information Rate (CIR), purchased by the service provider. This is most common on high bandwidth metro Ethernet WANs and MPLS networks where the service providers are not using QoS. In these designs, use hierarchical policies that shape the QoS policy to the WAN link capacity.

09

Design

Hierarchical WAN traffic shaping involves creating a high-level (parent) shaping policy and then associating it with a lower-level (child) policy. When applied to an interface, this policy forces the interface to shape all outgoing traffic to the parent class' rate. Once shaped, the lower level QoS policy queues traffic based on the shaped value and not the raw interface speed itself, thus queuing will result when the shaped value becomes congested.

The committed burst (BC) and excess burst (BE) should be configured to ensure optimum performance for real-time traffic like VOIP and video over IP. The BC should be set to a value less than or equal to 1/100 of the target shaped rate (CIR). This value may be increased to 2/100 for immersive video over IP deployments. The BE should always be configured as 0. ■

10

Proof of Concept

Any significant QoS deployment must start in the test lab. This is where configuration settings can be tested and validated in a non-production environment. Ideally, the test lab will have the same routers and switches and software versions as the production network.

Real and synthetic traffic should be generated in controlled conditions similar to the production network to check the operation of QoS policies and settings. This process must be repeated for each different policy, and finally, together with all other policies deployed. ■

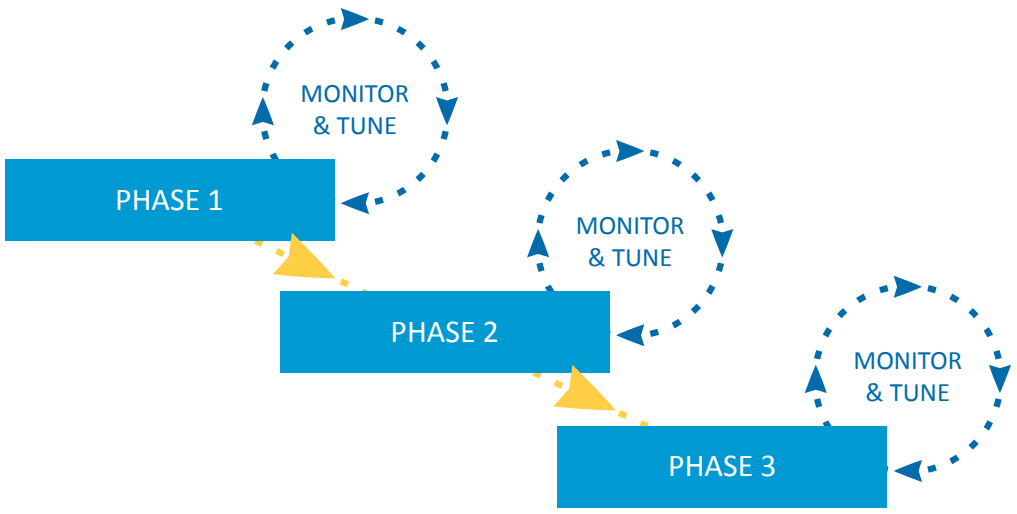
11

Iterative Deployment

Once QoS policies have been validated in a test environment, the process to roll them out in a phased approach can begin.

In order to facilitate the daunting task of implementing such a large change to a network, it should be broken up into implementation phases. At each phase, a change is made, the change is confirmed through monitoring, adjustments are made, and deployment proceeds to the next phase until the project is complete. This method allows the project to proceed on a flexible schedule that can be interrupted at milestones while maintaining business continuity.

Diagram: Phased Approach with monitoring/tuning



11

Iterative Deployment

This phased approach can be carried out in different scenarios. One option is to pick a small portion of a larger network for initial deployment. This can be a remote site with a smaller network. Another option is to deploy the policies by function. For example, the first phase would start with just the classification and marking policies on endpoints and network devices. The traffic can be monitored to determine if the markings are being applied correctly. Once confirmed, queuing, policing, and LFI mechanism can be deployed systematically.

At each stage of policy deployment, the traffic and QoS policies need to be monitored for proper operation. The policies should be fine-tuned as issues are uncovered. This monitoring should occur for at least 24 hours but ideally a week in order to catch different usage patterns that occur throughout the workday and week. This process of monitoring and fine tuning is repeated until the deployment is complete. ■

12

Ongoing Monitoring & Analysis

Once fully implemented, a QoS-enabled network will require ongoing monitoring and maintenance. Over time, changes will occur to the usage patterns, new applications will be deployed, and these changes will need to be incorporated into the existing policies. Another consideration is cyclical usage patterns. The most obvious of these are the work week cycle where traffic occurs in 10-12-hour cycles from Monday to Friday. However, there are other cycles that occur at the end of fiscal periods such as monthly, quarterly, and yearly. During these times, voice calls and ERP usage may peak due to accounting, manufacturing, sales, or internal meeting activities and these usage patterns will need to be accounted for. ■

13

Tools for Successful QoS Implementation

Before deploying QoS, consideration should be given to the tools available to staff for performing the critical functions involved. Implementing QoS requires a wide range of functions from monitoring and analysis to configuration and testing. Budgeting for QoS deployments needs to include funding for network management software tools. These tools will ensure accurate and effective QoS policy creation and enable a deeper understanding of the network environment for capacity planning, tuning and troubleshooting efforts.

13

Tools for Successful QoS Implementation

The following table summarizes specific functions and technical considerations that should be evaluated when purchasing network management software for QoS:

FUNCTION	USE	TOOL CONSIDERATIONS
Traffic monitoring and analysis	• Initial base-lining and ongoing monitoring of traffic levels. • Troubleshooting of any problems encountered	• Traffic filtering by IP address, application, DSCP, etc. • Long-term recording and reporting • End-to-end traffic path visibility for network understanding • Real-time updating for immediate feedback when troubleshooting • Support for network flow monitoring technology – NetFlow, Sflow or J-Flow
QoS monitoring and analysis	QoS policy validation and ongoing monitoring	• Support for SNMP MIB monitoring such as Cisco CBQoS and NBAR MIBs • Detailed monitoring by class, class drops, application, etc. • Historical reporting of QoS performance
QoS design and configuration	Assist in defining QoS policies and actual configuration of the network devices	• Wizards and templates for easy policy creation • Rules checking and references to ensure error free changes • Intelligent configuration that parses existing router policies and adjusts changes as needed • Load and save policies • Push out policies to multiple devices
Synthetic traffic generation and analysis	Traffic generation and analysis for measuring the impact of QoS policies in a controlled manner	• Ability to leverage built-in Cisco IP SLA technology • Graphically generate different types of traffic and analyze performance in summary and detailed formats

Consideration should be given to an application such as LiveAction's LiveNX. LiveNX is an application-aware network performance management tool that will graphically display how networks and applications are performing using SNMP and the latest advanced NetFlow capabilities now embedded in Cisco devices. In addition to showing application and network performance, LiveNX provides the ability to control application performance via its graphical QoS management capabilities. This software solution integrates all of the

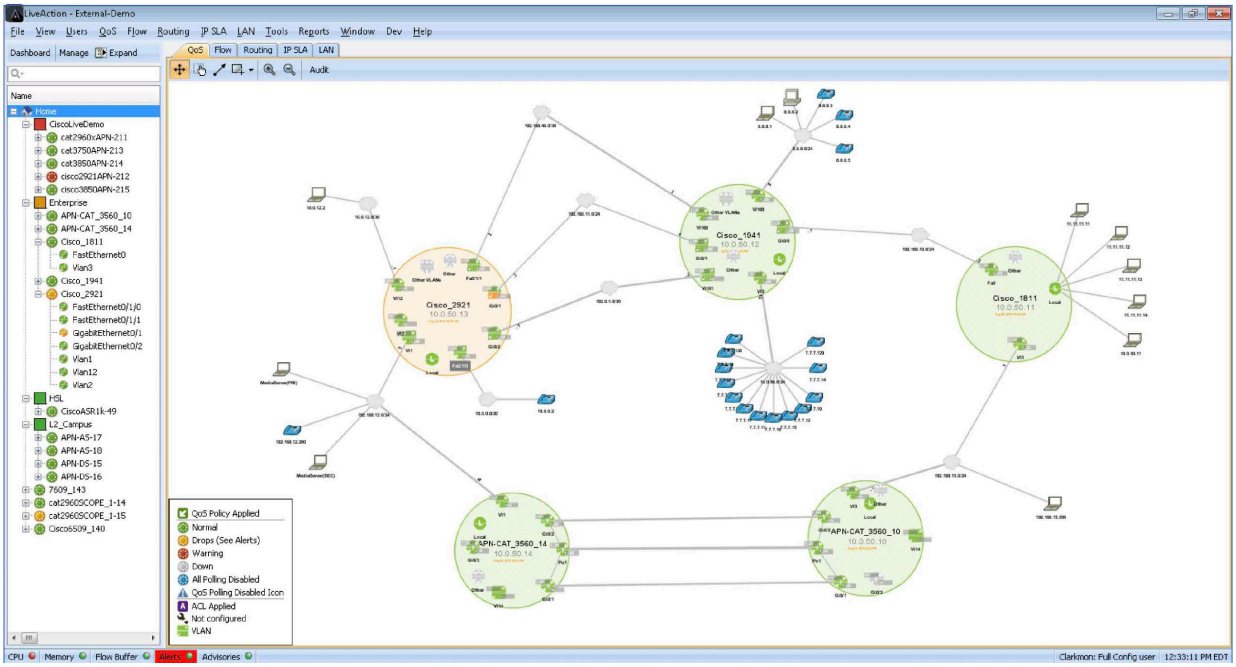
13

Tools for Successful QoS Implementation

functionality required in a QoS deployment into a cohesive, easy-to- manage framework. By combining the entire workflow of analysis, configuration, testing, and tuning into an all-in-one solution, a QoS deployment can greatly benefit from improved efficiency and accuracy. Using LiveNX's built-in expert capability to build and adjust policies graphically, understand network-wide traffic patterns visually, and gain detailed QoS statistics enables a quicker understanding of how policies are working and what adjustments need to be made when tuning QoS policies. ■

The image below is a view of the LiveNX console. It shows a network diagram consisting five network devices and multiple PCs, servers, and IP phones. The five larger green circles represent routers and switches managed by LiveNX. The little circles within the devices represents their interfaces. Four of the devices are green indicating good network performance. One device is amber, indicating that this device is having a network/QoS performance issue.

LiveNX console



14

Traffic Monitoring & Analysis

The early stages of a QoS deployment involves information gathering that is then used to design QoS policies.

For information gathering, NetFlow, NBAR and SNMP monitoring and reporting tools are required. Implementing and analyzing QoS monitor only policies are also very helpful in the information gathering stages of a QoS deployment. The tools used for monitoring these technologies need to support long term traffic analysis and reporting at 1-minute or better views (no averaging) to gain a thorough understanding of the network's data usage prior to any QoS policy design.

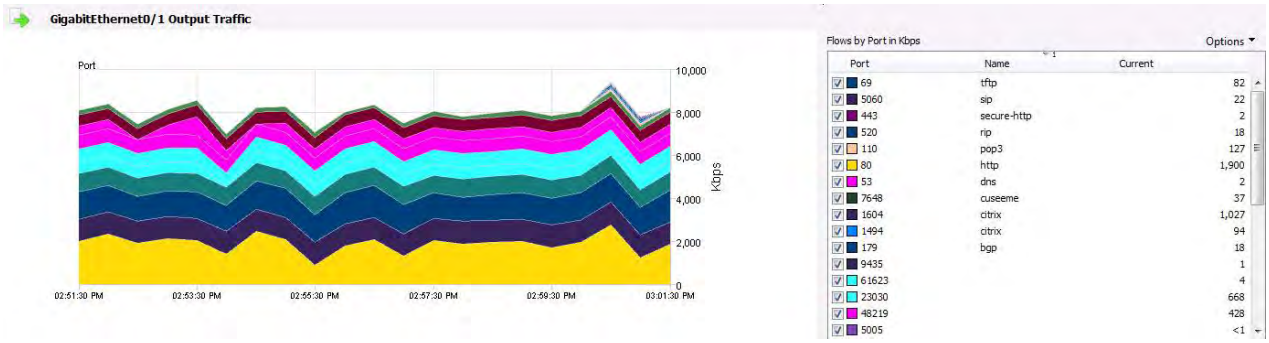
LiveNX can provide both real-time and historical views and reports that are specifically tailored for QoS deployments. The following three LiveNX screenshots show the real-time view of a single interface's performance from four different technologies perspectives. These specific views can be used to highlight the volume of data each application is utilizing on the interface for the last ten minutes. This data can then be used for real-time capacity planning of QoS policies.

14

Traffic Monitoring & Analysis

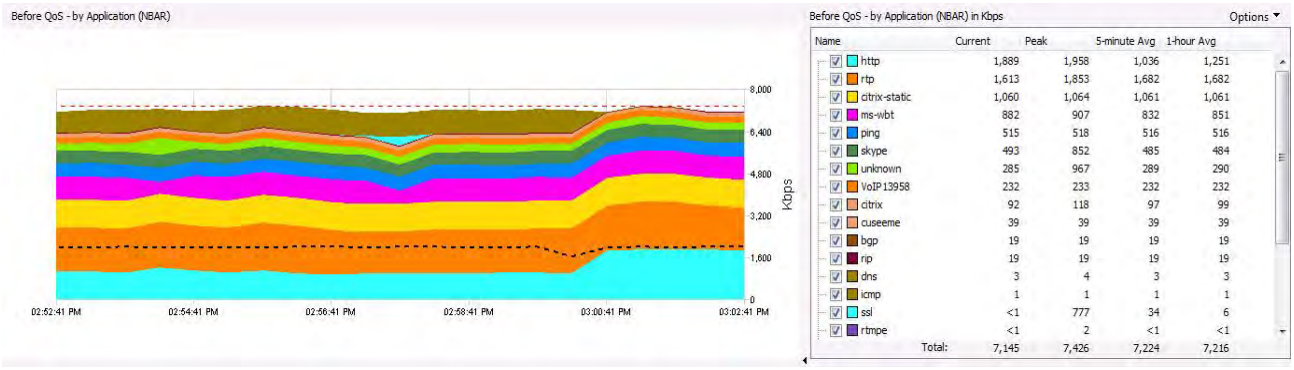
NetFlow

The diagram below shows an interface's network performance from NetFlow's perspective.
This view is detailing bandwidth by TCP/UDP port usage.



NBAR

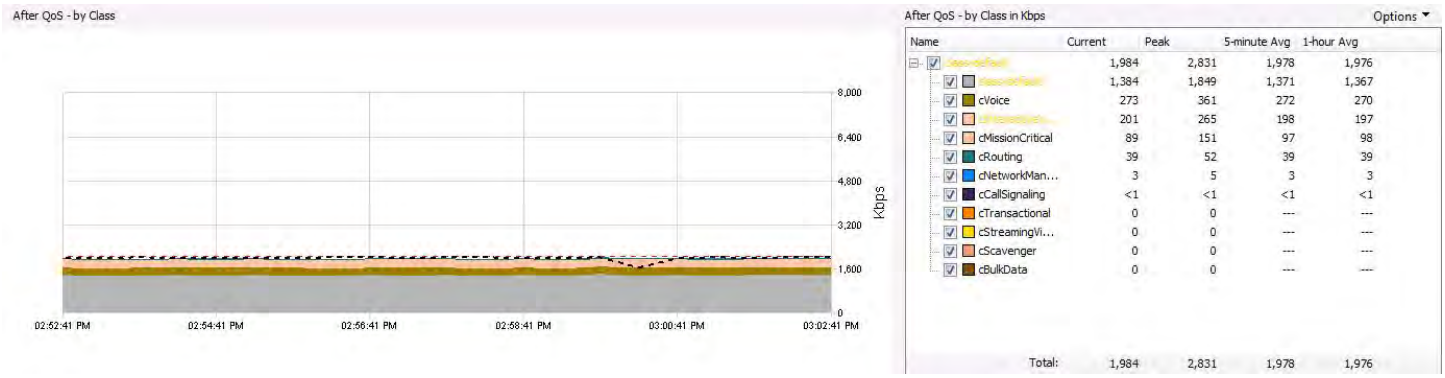
The diagram below shows an interface's network performance from NBAR's perspective.
The view is detailing bandwidth by application usage. The dotted black line shows interface SNMP's bandwidth usage.



14

Traffic Monitoring & Analysis

The diagram below shows an interface's network performance from QoS perspective.
The view is detailing bandwidth usage by queue. The dotted black line shows the interfaces SNMP's bandwidth usage.

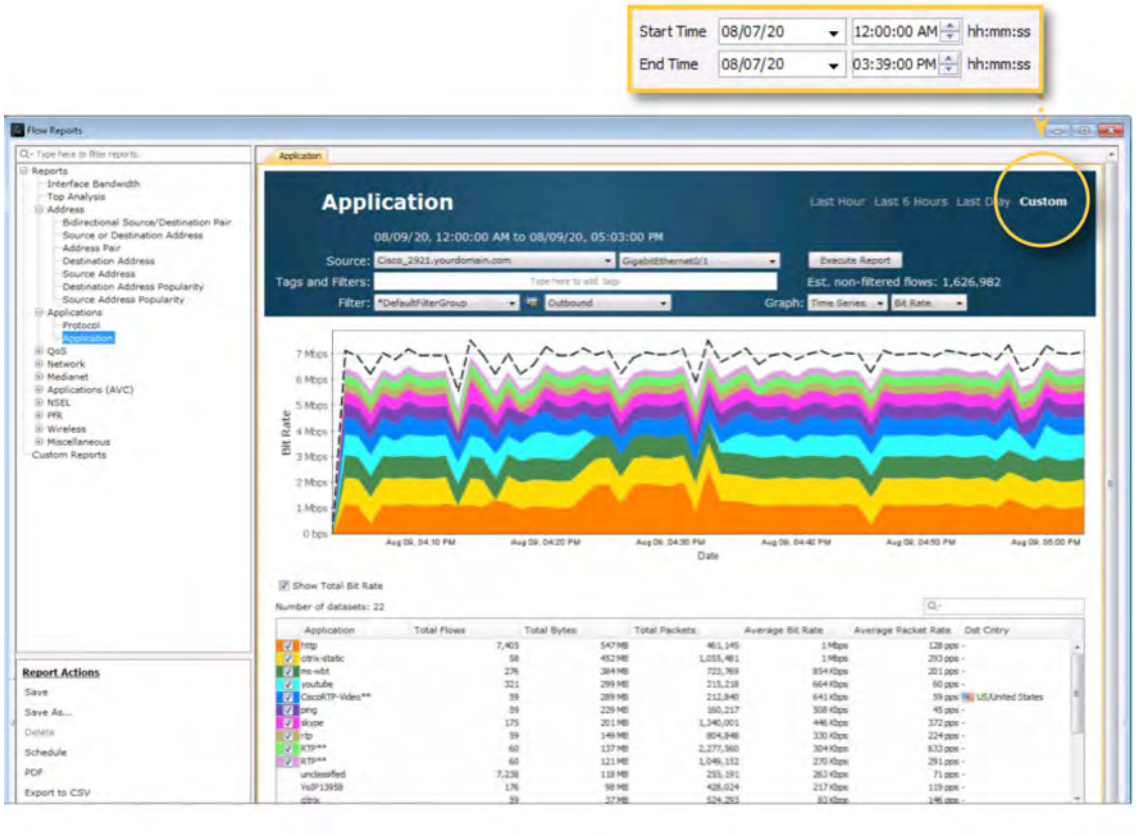


LiveNX can also be used as a Historical reporting tool. Traffic monitoring and analysis for QoS policy creation should typically have at least one week's worth of historical data for accurate policy creation. The following screenshots will highlight a subset of the reports available in LiveNX that would typically be used as part of planning a QoS deployment. The reports shown below should be run on multiple devices and interfaces to understand a network's overall application usage. Note that all historical reports in LiveNX can show data at 1-minute granularity. This ensures the reports have the detail required for accurate QoS policy creation.

14

Traffic Monitoring & Analysis

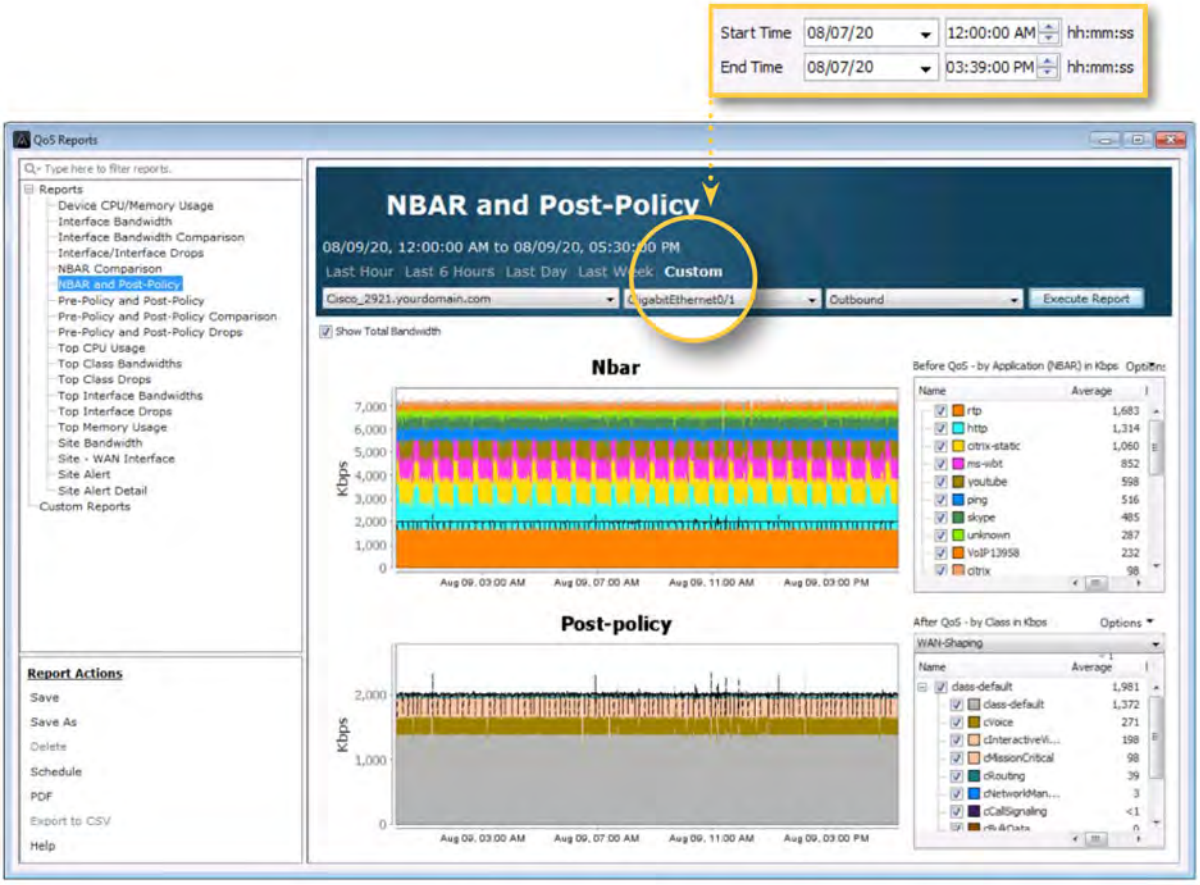
The following diagram shows a historical NetFlow Application report for the same interface as highlighted above. By using the Custom option, the report can be tailored to a desired historical time.



14

Traffic Monitoring & Analysis

The diagram below shows the QoS historical report: NBAR and Post-Policy. This view presents the NBAR and QoS data for the same interface as highlighted above. This report can be used to gather the volume of bandwidth each application is utilizing on an interface for the time period selected. By using the Custom option, the report can be tailored to a desired historical time period. If QoS monitoring policies are utilized, as the screenshot below shows, queue bandwidth sizing can more easily be determined for the QoS design. ■



15

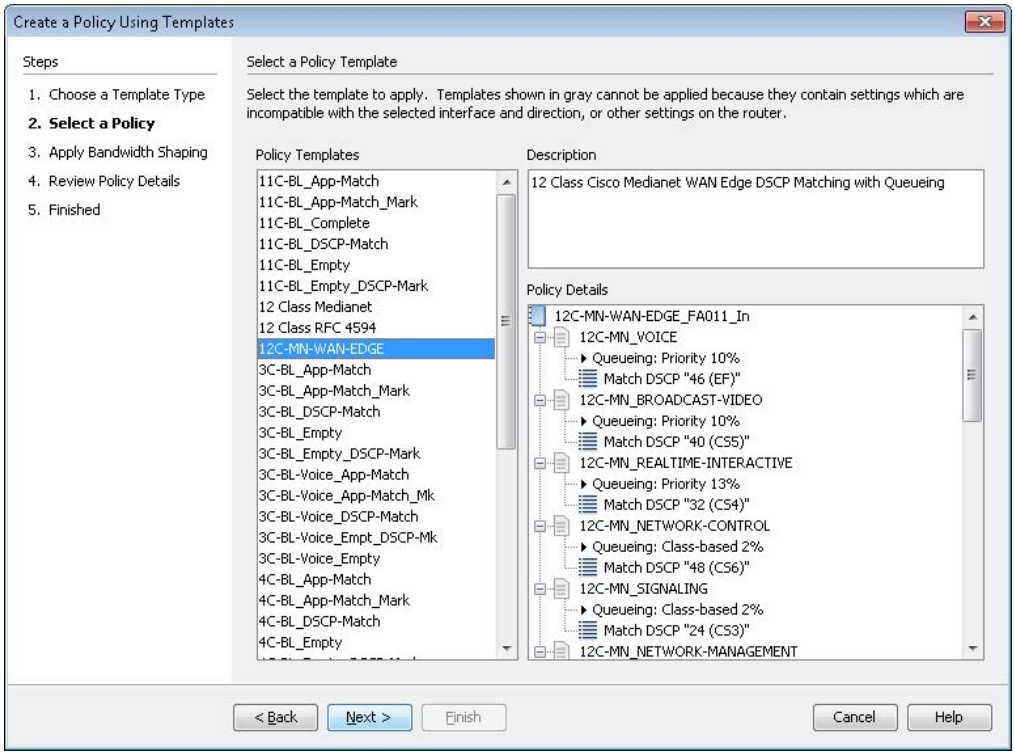
QoS Design & Configuration

Designing QoS policies and configuring them on routers is a time-consuming, complex, and intimidating process that requires trained engineers. This process can be greatly simplified by a QoS tool with built-in expert capability in the form of GUI wizards, templates, and interactive screens for designing, deploying, and adjusting QoS policies. LiveNX provides easy-to-use workflows that enable engineers to deploy, tune, and validate QoS policies across an enterprise network quickly and accurately. LiveNX has built-in advanced QoS policy templates and wizards based on Cisco best practices. Engineers can quickly deploy Cisco recommended 12, 11, 8, 5, 4 or 3 class model QoS policies.

15

QoS Design & Configuration

In the diagram to the right, the Create a Policy Using Templates wizard dialog box is shown. The Cisco 12 Class Medianet WAN Edge policy is selected.

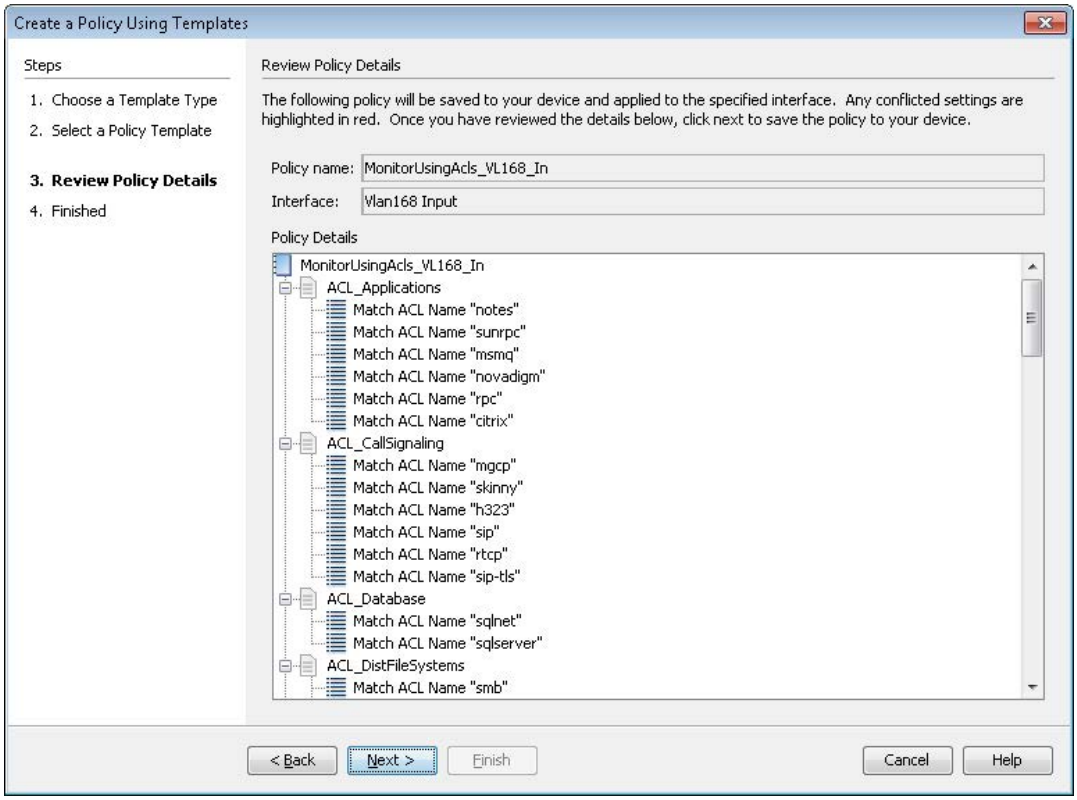


15

QoS Design & Configuration

LiveNX also has built-in QoS monitoring templates using either NBAR or ACLs. These monitor-only templates do not prioritize any applications, but they are extremely helpful for trending and base lining traffic patterns.

To the right is a diagram of an ACL Monitor template wizard dialog box.



15

QoS Design & Configuration

LiveNX has a GUI called Manage QoS Settings that provides engineers a point-and-click interface to quickly create, edit and deploy custom QoS policies. This QoS configuration interface has built-in rules checking and reference information to warn engineers about configurations issues and provide solutions. This tool can replace the use of complex CLI commands associated with QoS.

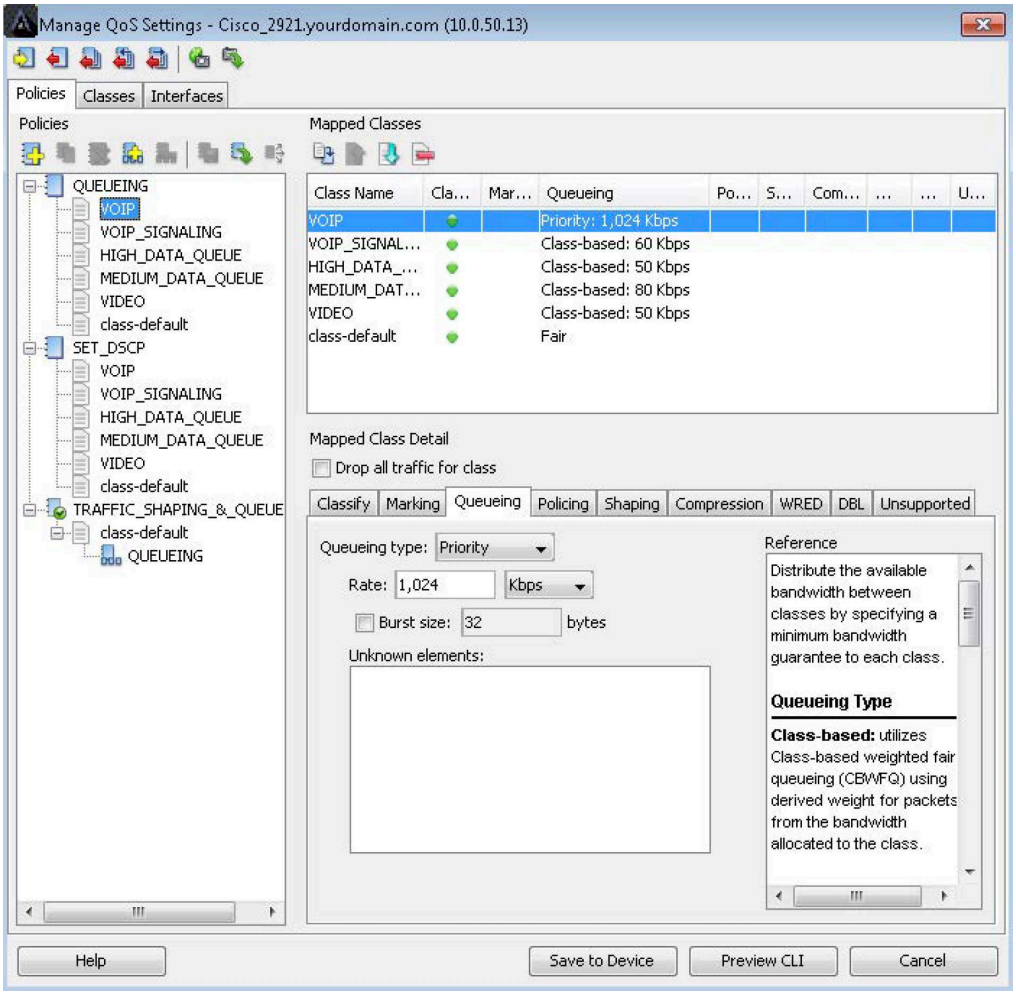
Using LiveNX, engineers can:

- Create advanced QoS queuing polices with WRED congestion avoidance
- Classify and match traffic to classes using ACL, NBAR, DSCP value, and more.
- Implement policing and shaping polices
- Create hierarchical QoS Polices
- Deploy standardized QoS polices to the rest of the network
- Revert changes as needed

15

QoS Design & Configuration

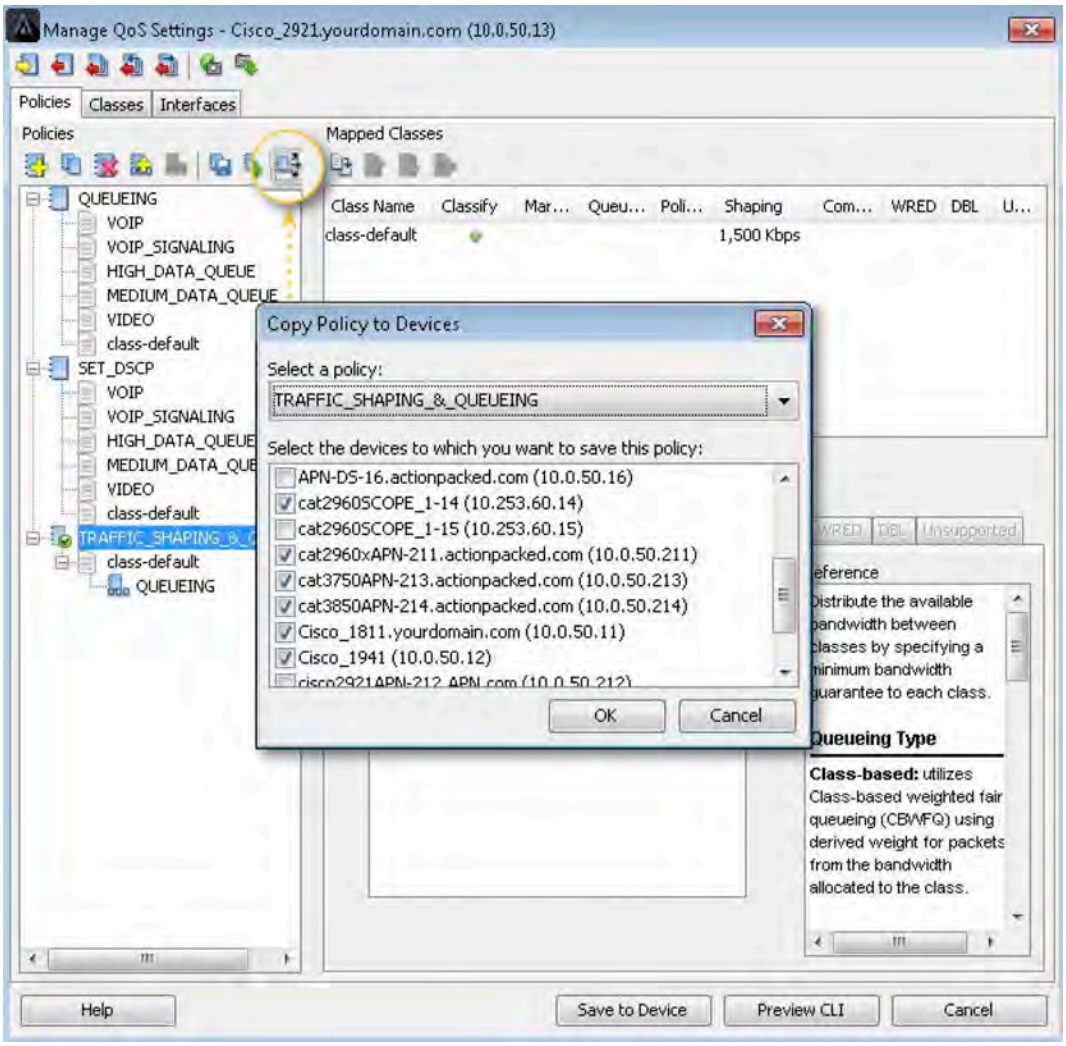
Below is a screenshot of the Manage QoS Settings window. In the example below, the queuing configuration is shown for a class named VOIP. The bandwidth rate of this queue can be determined from the real-time and historical reports collected by LiveNX.



15

QoS Design & Configuration

Below is the same window, but the hierarchical QoS policy TRAFFC_SHAPING_&_QUEUEING is highlighted. If an engineer wanted to deploy this policy to the rest of the network, he could select the Copy Policy to Devices button and use the pop-up window to select and push this policy to the remainder of the enterprise. This feature will ensure uniform and consistent policies are deployed throughout the environment.



15

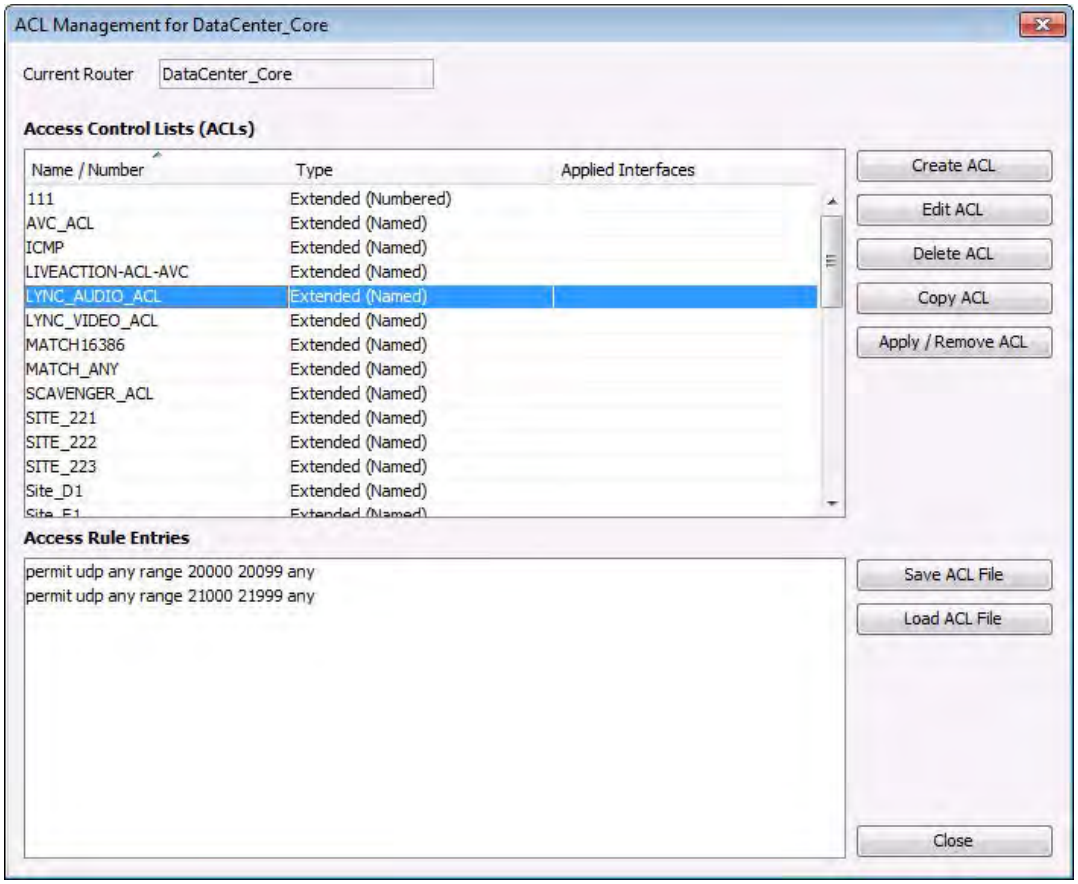
QoS Design & Configuration

LiveNX also has built-in ACL management capabilities that gives administrators a GUI to configure and push access-lists to the devices in the enterprise. This feature provides engineers the ability to centrally manage and deploy ACLs in a uniform, consistent manner.

With LiveNX access-lists can be:

- Created from scratch
- If already in place, discovered automatic and edited
- Created or edited using NetFlow data

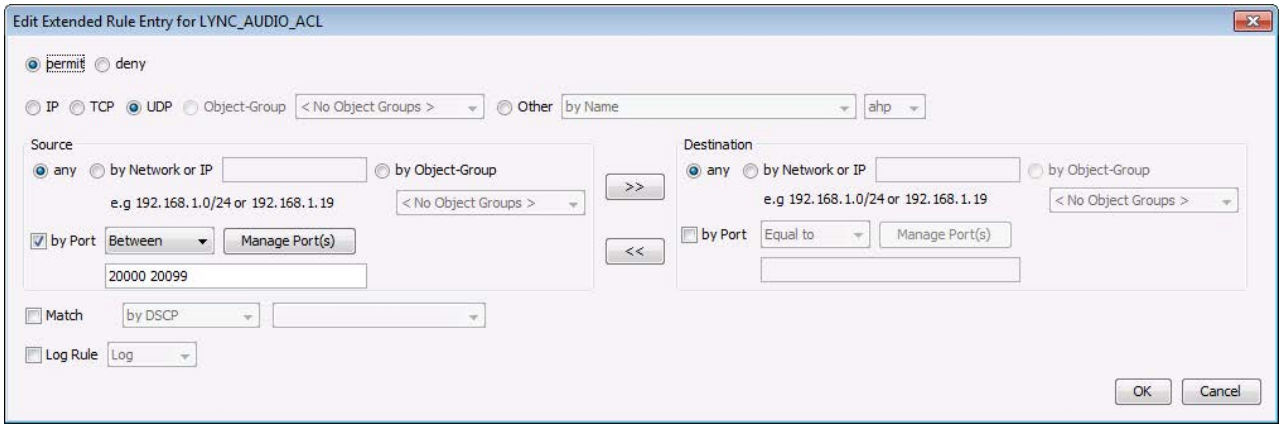
The following diagram shows the ACLs LiveNX discovered on a device.



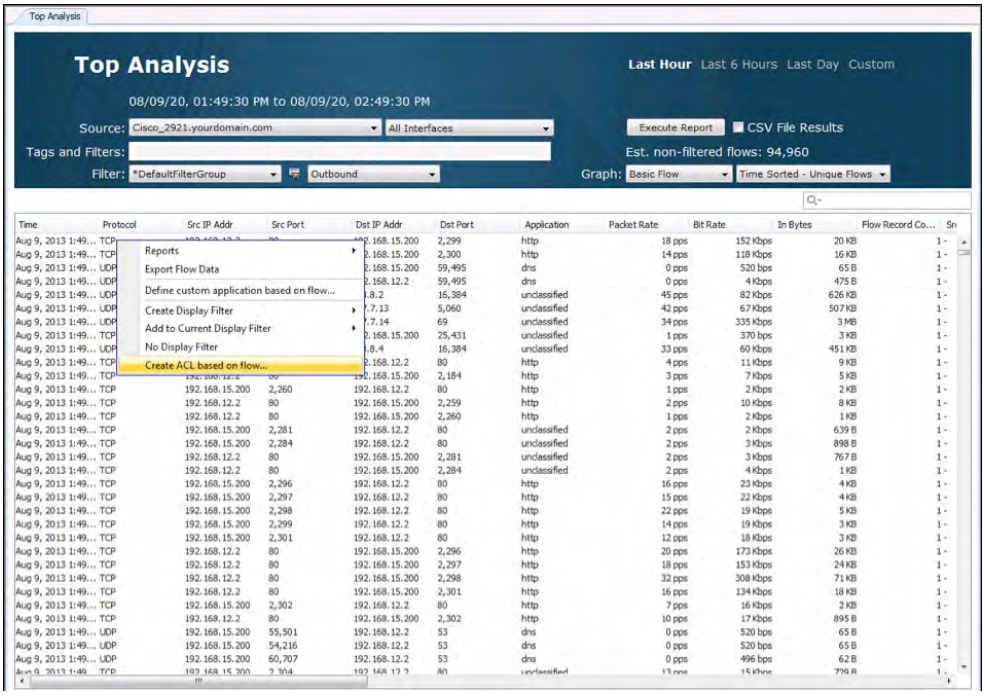
15

QoS Design & Configuration

The following diagram shows how one can create and edit individual lines (ACEs) within an ACL.
In this example, all UDP traffic with a source port in the range of 20000 – 20099 are being permitted.



LiveNX can also create ACLs based on the NetFlow data it collects. This can be accomplished from both real-time and historical report views. ■



16

QoS Monitoring & Analysis

When implementing Cisco QoS there are a number of useful QoS statistics available in IOS to help understand how well policies are working and guide any needed policy adjustments. However, it's very difficult to use these statistics in a meaningful manner when using the command line interface. Having a software tool that supports comprehensive real-time polling and historical reporting of QoS performance is critical.

Key features should include:

- CBQoS and NBAR MIB support
- Interface graphs for before and after QoS on input, output, or both
- Statistics by Class and Class Drops and Interface Drops. This information can also be useful when troubleshooting network latency issues resulting from application malfunctions or virus outbreaks.
- NBAR application statistics
- Historical graphs with zoom and report generation

16

QoS Monitoring & Analysis

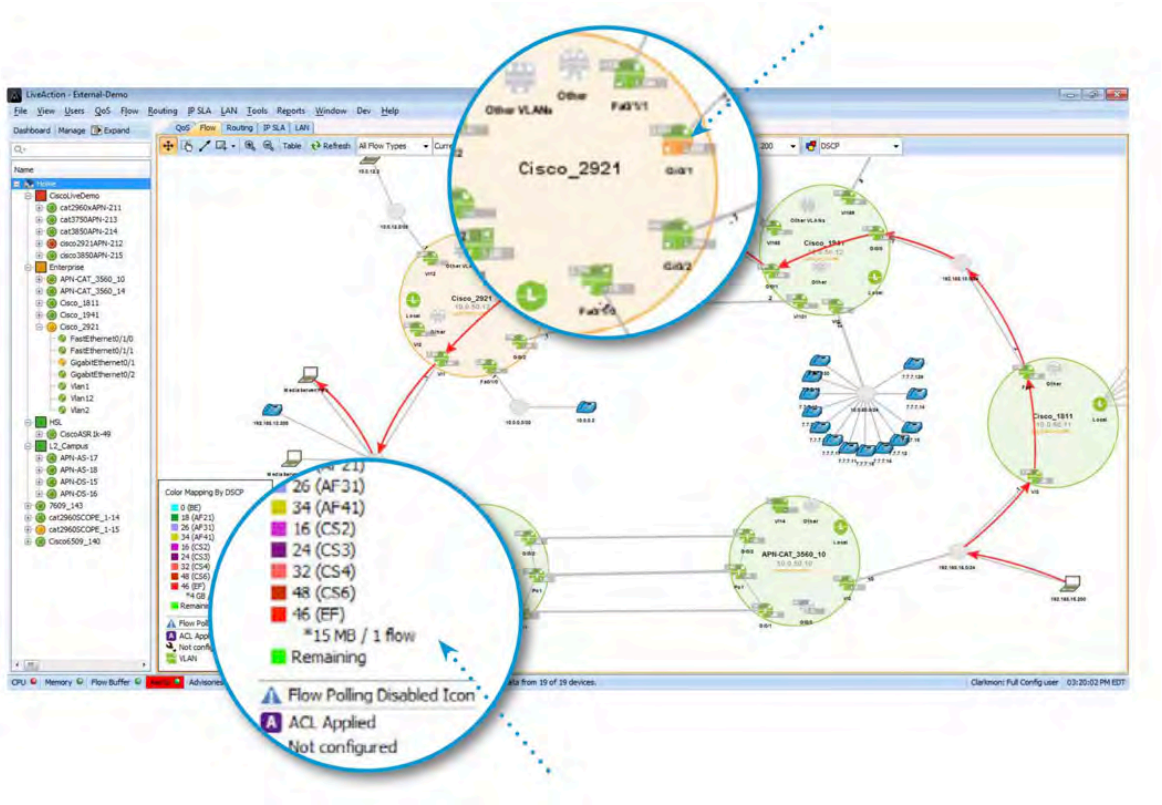
In addition, troubleshooting and implementation will be much easier with system-wide flow visualization. This is critical for quickly gaining end-to-end awareness of the traffic flowing across the network and identifying points of congestion, incorrectly configured services, and invalid traffic.

Since LiveNX is a NetFlow collector and mapping tool too, it has the ability to graphically paint a picture of the traffic that is flowing over a network. In the diagram below, the red arrows are a graphical representation of a conversation between two endpoints. Engineers can quickly analyze the devices in the network path that are part of a conversation and can also determine the QoS marking set on the traffic. The red color in this example, according to the legend, is DSCP 46(EF). This ability to track QoS markings visually facilitates quick and accurate validation of QoS classification polices. Any problems with DSCP markings can be quickly understood and fixed. LiveNX will also highlight any QoS performance issues in real-time.

16

QoS Monitoring & Analysis

- The interface in amber, is indicating a QoS performance issue.

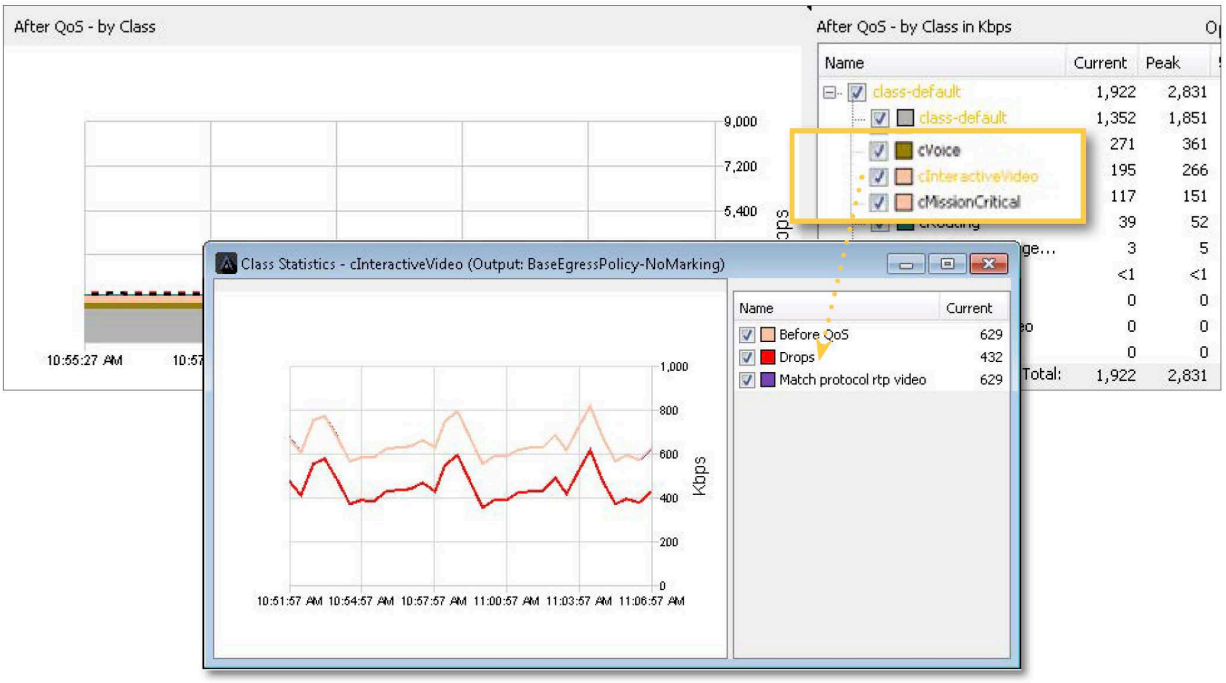


If the amber interface is investigated further with LiveNX's point-and-click interface, the root of the QoS issue can be discovered.

16

QoS Monitoring & Analysis

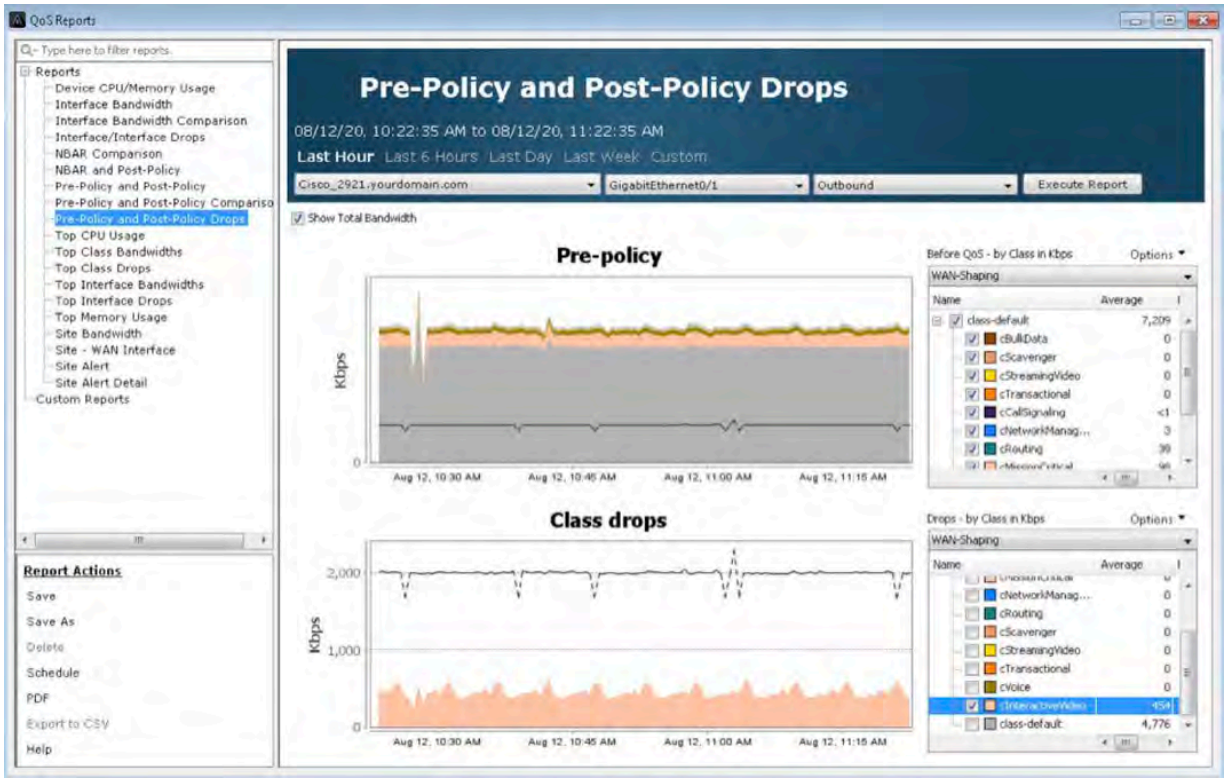
- In the example below, there are drops in the cInteractiveVideo queue.



16

QoS Monitoring & Analysis

If these drops are causing business critical applications to fail, the same real-time and historical reports highlighted in the Traffic Monitoring and Analysis section of this document and the QoS Pre-Policy and Post-Policy Drops historical report can be utilized to determine if any tuning needs to take place with the QoS policy configuration. Below, in the graph title Class drops, the cInteractiveVideo drops are being viewed historically. This will help further determine any changes required to this queue's bandwidth allocations. ■



17

Traffic Generation & Analysis

Whether experimenting in the lab or verifying performance in an operational network, it is useful to have some type of traffic generation and analysis control. Many Cisco devices support built-in IP SLA capabilities that generate synthetic traffic and provide performance statistics from that traffic. Cisco IP SLAs generate synthetic test traffic in a continuous, reliable, and predictable manner to enable accurate measurement of network performance. Simulating network traffic with appropriate QoS markings can validate the operational performance of QoS polices and ensure critical applications are performing as required.

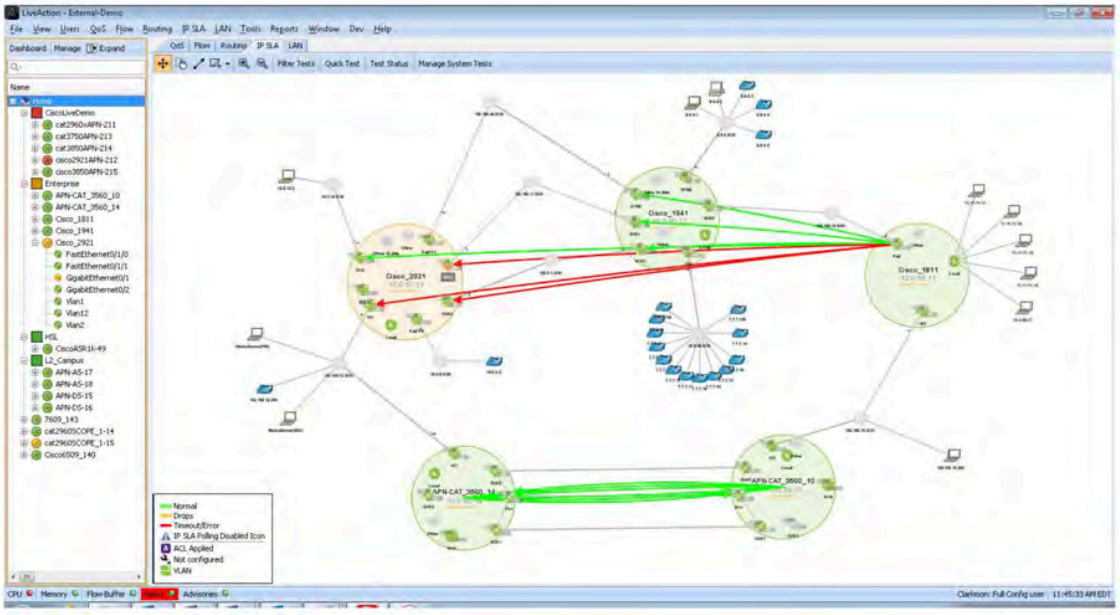
LiveNX's IP SLA features will greatly aid in QoS testing and monitoring via:

- Quick test wizard and advanced GUI setup
- Support for all IP SLA traffic types, including voice, HTTP, UDP
- Graphical results summary tables
- Detailed test results including latency, loss, jitter and MOS
- Real-time, topology views of IP SLA traffic flows

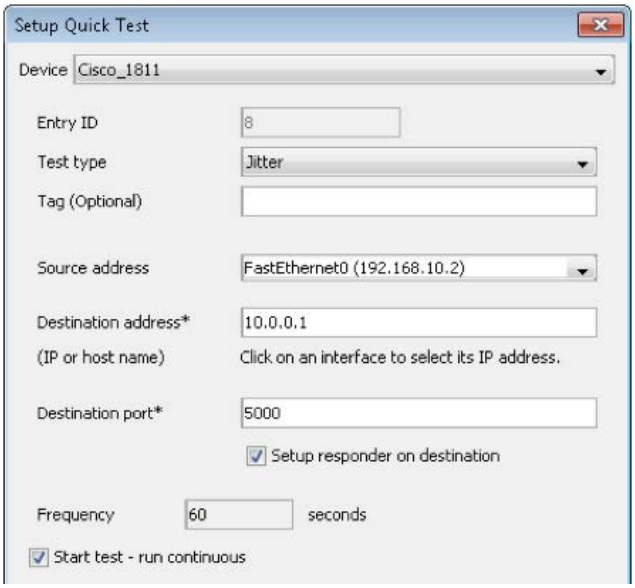
17

Traffic Generation & Analysis

LiveNX will discover the IP SLA test currently running in a network environment. In the screenshot below, LiveNX is visually showing the performance of various IP SLA tests. The green lines are tests that are performing well, the red lines indicate tests that are failing.



Using the Quick Test wizard, IP SLA test can be deployed easily and quickly.

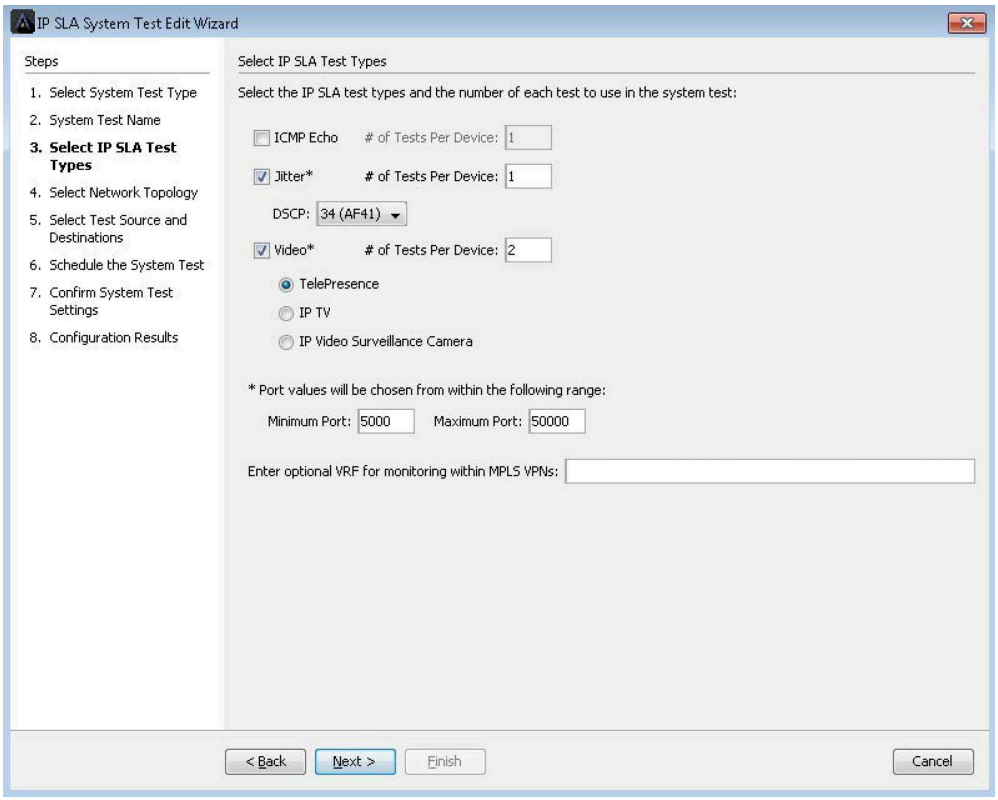


17

Traffic Generation & Analysis

LiveNX also supports the latest and complex IP SLA Video Operation (VO) tests. These tests will turn Cisco network devices into video traffic generators that will simulate video over IP calls. These simulated calls are used to validate a QoS policies' ability to handle video's stringent network requirements on the network.

The built-in wizards simplify this complex set of configurations task and will walk engineers through an easy-to-use set of menus to create and deploy IP SLA VO tests.



17

Traffic Generation & Analysis

LiveNX's IP SLA historical reports will give engineers the data they need to review and analyze the long-term performance of IP SLA and IP SLA VO test and consequently the application performance.

18

Conclusion

Implementing QoS on a network of any size is a complex and time-consuming process. Implementation requires a diverse skill set that encompasses technical and even political skills, project management expertise, patience, and confidence. Proper planning, technical, and project understanding, along with the right network management tools, will help network managers and engineers deploy and maintain converged QoS-enabled networks with less time, money, and stress. In addition, the policies, tools, and processes put into place will be useful in many other aspects of network management. Taking the time to set up these monitoring mechanisms will give administrators a more complete view of the overall health of the network, its weak points, and potential challenges, well before they become issues.

For more information on solutions to help deploy and manage QoS, please visit LiveAction at: www.liveaction.com. ■

Appendix A

QoS Configuration Best Practices

QoS policies should be designed from the high-level requirements gathered in the project planning stage.

Here are some best practices to keep in mind when designing QoS policies.

- Choose a matching and marking strategy that ensure traffic is matched and marked effectively and ensure traffic is protected where it is most needed.
- When possible, follow industry standards when deploying QoS polices to align with service providers and business partners.
- Mark packets as close to the source as possible (i.e. the device itself or switch connected to the device).
- Police recreational or scavenger traffic as close to the source as possible to prevent unnecessary bandwidth usage if traffic exceeds a certain threshold.
- Enable queuing wherever congestion may occur to guarantee service optimum perforce to critical applications. This includes data center and campus LANs as well as public (DMVPN) and private (MPLS WANs.
- Provide at least 25% bandwidth for Best Effort traffic in the class-default.

Appendix A

QoS Configuration Best Practices

- For real-time traffic such as voice and video, use priority queues and be sure to assign ample bandwidth. However, attempt to limit the priority queues to 33% of the overall available bandwidth to prevent starving other traffic.
- If multiple applications use PQ, use discrete priority queues for each applications. ie. VOIP gets its own unique queue, video gets its own queue, etc.
- Deploy Link Efficiencies Mechanisms on WANs less than or equal to 768Kbps.
- Hierarchical traffic shaping should use a BC value of less than or equal to 1/100 of the CIR. This may be increased to 2/100 of CIR for immersive video over IP implementations. The BE value should always equal 0.
- Overprovision video over IP queues by 20% to accommodate bursts.

Appendix B

QoS Reference Tables

Table 1: Medianet DSCP Markings Recommendations

CISCO NAME/ RFC4594 NAME	4 CLASS	8 CLASS	12 CLASS
Voice / IP Telephony	EF (46) CS5(40) CS4 (32)	EF (46)	EF (46)
Interactive Video / Multimedia Conferencing		CS5 (40) CS4 (32)	AF41 (34) AF42 (36) AF43 (38)
Streaming Video		AF31 (26) AF32 (28) AF33 (30	AF31 (26) AF32 (28) AF33 (30
Real-Time Interactive			CS4 (32)
Broadcast Video			CS5(40)
Call Signaling	CS6 (48) CS3 (24) CS2 (16)	CS3 (24)	CS3 (24)
IP Routing / Network Control		CS2 (16) CS6 (48)	CS6 (48)
Network Management / Operations, Administration,Management (OAM)			CS2 (16)
Transactional Data / Low-Latency Data	AF41 (34) AF42 (36) AF43 (38) AF31 (26) AF32 (28) AF33 (30) AF21 (18) AF22 (20) AF23 (22) AF11 (10) AF12 (12) AF13 (14)	AF41 (34) AF42 (36) AF43 (38) AF21 (18) AF22 (20) AF23 (22) AF11 (10) AF12 (12) AF13 (14	AF21 (18) AF22 (20) AF23 (22)
Bulk Data / High Throughput Data			AF11 (10) AF12 (12) AF13 (14)
Scavenger / Low-Priority Data	BE (0)	CS1 (8)	CS1 (8)
Best Effort		BE (0)	BE (0)

Appendix B

QoS Reference Tables

Table 2: Medianet Queue Sizing Recommendations

CISCO NAME/ RFC4594 NAME	4 CLASS	8 CLASS	12 CLASS
Voice / IP Telephony	>33%	10%	10%
Interactive Video / Multimedia Conferencing		23%	10%
Streaming Video		10%	10%
Real-Time Interactive			13%
Broadcast Video			10%
Call Signaling	Remaining	2%	2%
IP Routing / Network Control		5%	2%
Network Management / Operations, Administration,Management (OAM)			2%
Transactional Data / Low-Latency Data		24%	10%
Bulk Data / High Throughput Data	<5%		5%
Scavenger / Low-Priority Data		1%	1%
Best Effort	<25%	25%	25%

Appendix B

QoS Reference Tables

Table 3: VOIP CODEC Bandwidth

CODEC	MOS SCORE	MLPP OR FRF.12	W/CRTP MLPPP OR FRF.12	ETHERNET	L3 (GK AND RSVP)
G.711 64 Kbps	4.1	82.2 Kbps	67.6 Kbps	87.2 Kbps	80 Kbps
G.729 8 Kbps	3.92	26.8 Kbps	11.6 Kbps	31.2 Kbps	24 Kbps
G.723.1 64 Kbps	3.9	18.9 Kbps	8.8 Kbps	21.9 Kbps	23 Kbps
G.723.1 5.3 Kbps	3.8	17.9 Kbps	7.7 Kbps	20.8 Kbps	22 Kbps
G.726 32 Kbps	3.85	50.8 Kbps	35.6 Kbps	55.2 Kbps	48 Kbps
G.728 16 Kbps	3.61	25.8 Kbps	18.4 Kbps	31.5 Kbps	32 Kbps
G.722 64 Kbps	4.13	82.8 Kbps	67.6 Kbps	87.2 Kbps	80 Kbps

Table 4: Voice Video SLA Targets

PARAMETER	VOIP	TRADITIONAL VIDEO OVER IP	HD / IMMERSIVE VIDEO OVER IP
Bandwidth	8-90Kbps	384 -768 kbps + network overhead	1.5 - 12.6 Mbps + network overhead
Latency	150ms	150ms	150ms
Jitter	30ms	30ms	10ms
Loss	1%	1%	0.05%

Appendix C

Validating QoS using Performance Monitor

QoS policies should be designed from the high-level requirements gathered in the project planning stage.

LiveNX supports the Cisco's Performance Monitor (PerfMon) Flexible NetFlow template. This will allow administrators to deploy and manage PerfMon in the network infrastructure and gather application performance metrics (packet loss, jitter) for voice and video over IP. This advancement in technology gives administrators the ability to understand how these applications are performing without the use of hardware probes or other costly network appliances. Using the technology now embedded inside Cisco network equipment, voice and video over IP call quality issues can be detected and reported to network administrators before end-users ever complain. These call quality issues can many times be related back to QoS performance issues.

LiveNX can also display real-time PerfMon flow records. Packet loss and jitter measurements are now visible in these flow records. Both high packet loss and high jitter are performance issues that voice and video over IP cannot tolerate. In the example below, two of the flows are being highlighted in pink due to an alarm being triggered by the cells in red.

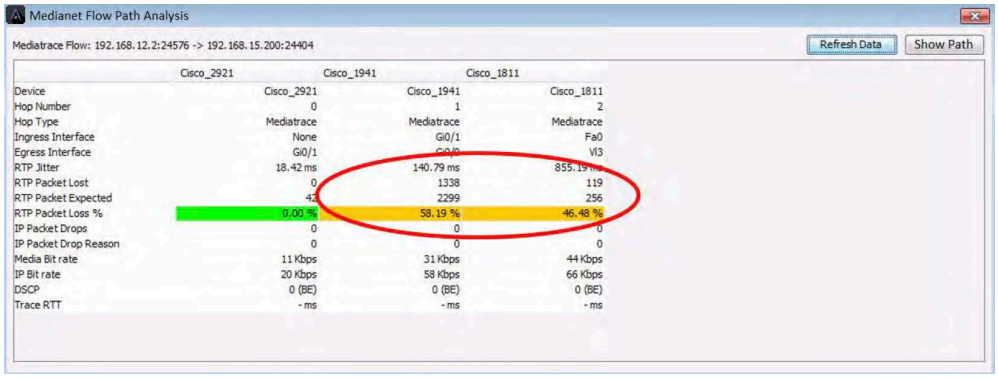
Appendix C

Validating QoS using Performance Monitor

In these example flows; Jitter Max measurements are triggering an alarm.
Network administrators are able to receive these performance alerts via Email or Syslog.

Src IP Addr	Src Port	Dst IP Addr	Dst Port	Application	DSCP and IP...	RTP SSRC	Packet Loss Count	Packet Loss Percentage	Jitter Mean	Jitter Min	Jitter Max
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	46 (EF)	3219829004	0	0.00%	0.00 ms	0.00 ms	0.56 ms
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	46 (EF)	3219829004	0	0.00%	0.00 ms	0.00 ms	0.56 ms
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	46 (EF)	3219829004	0	0.00%	0.00 ms	0.00 ms	1.82 ms
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	46 (EF)	3219829004	0	0.00%	0.00 ms	0.00 ms	1.82 ms
192.168.6.2	20,100	222.222.222.222	20,100	rtp	34 (AF41)	3176870698	0	0.00%	0.01 ms	0.00 ms	1.53 ms
192.168.6.2	20,100	222.222.222.222	20,100	rtp	34 (AF41)	3176870698	0	0.00%	0.01 ms	0.00 ms	1.53 ms
192.168.6.2	20,100	222.222.222.222	20,100	rtp	34 (AF41)	3176870698	0	0.00%	0.01 ms	0.00 ms	[redacted]
192.168.6.2	20,100	222.222.222.222	20,100	rtp	34 (AF41)	3176870698	0	0.00%	0.01 ms	0.00 ms	[redacted]
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	46 (EF)	2254833084	0	0.00%	0.01 ms	0.00 ms	1.04 ms
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	46 (EF)	2254833084	0	0.00%	0.02 ms	0.00 ms	1.06 ms
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	0 (BE)	2254833084	N/A	N/A	N/A	N/A	N/A
192.168.6.2	20,000	222.222.222.222	20,000	ms-lync-media	0 (BE)	2254833084	N/A	N/A	N/A	N/A	N/A

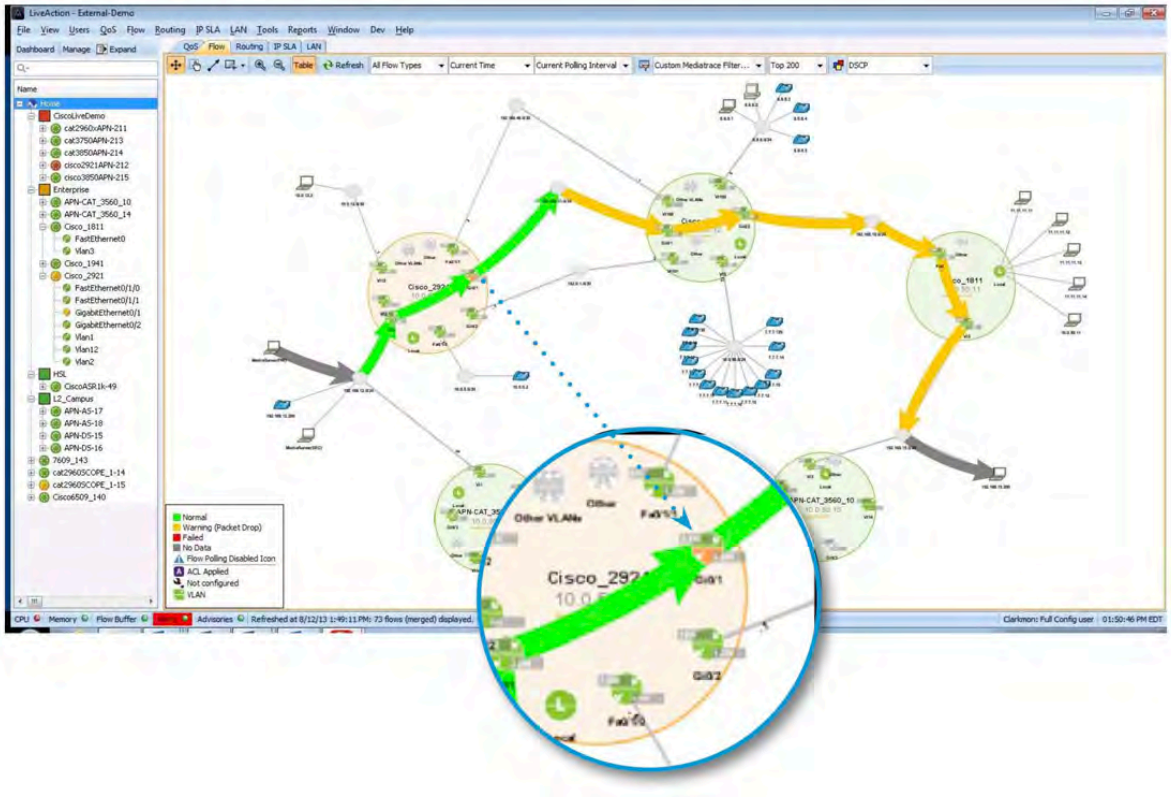
Below is a screenshot of a Flow Path Analysis table view. It is showing a voice/video call's performance through three network devices. The center and right columns are warning of high packet loss being reported to LiveNX by their respective devices.



Appendix C

Validating QoS using Performance Monitor

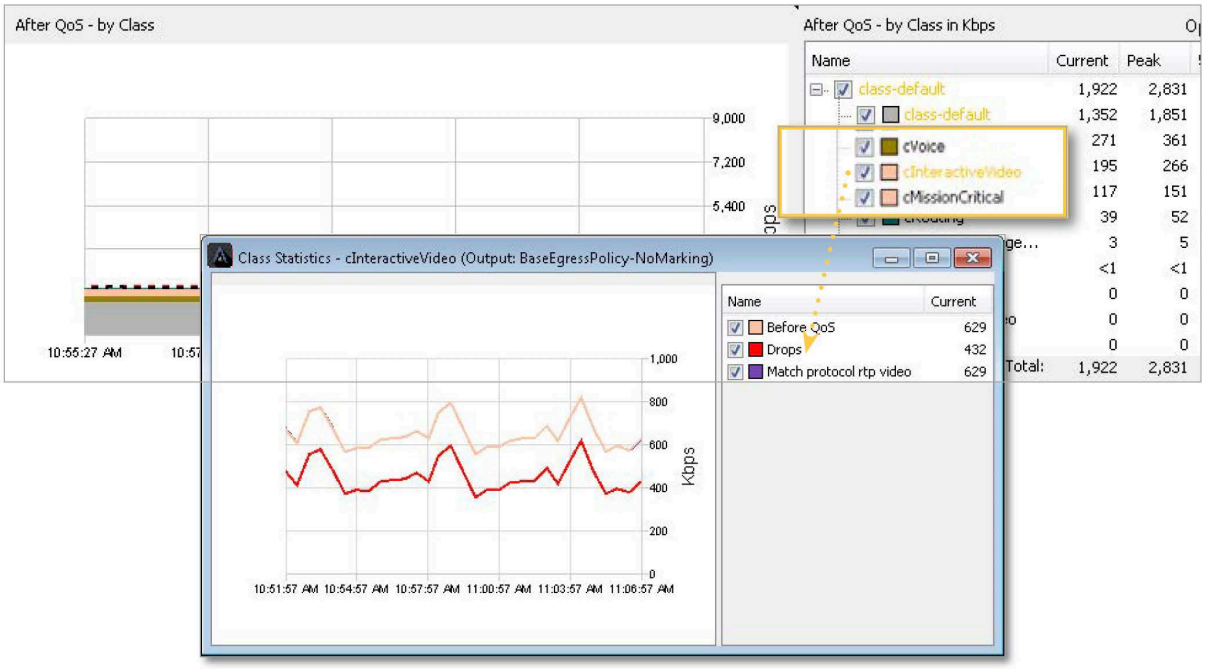
Below is the LiveNX network map of this same Path Analysis highlighting the performance problem visually. The arrows shown in the diagram are one discreet voice/video call through the network. The amber color indicates the call is experiencing performance issues on the center and right devices. Notice how the call passes through an interface reporting performance problems.



Appendix C

Validating QoS using Performance Monitor

If the amber interface is investigated further with LiveNX's point-and-click interface, the root cause being a QoS issue can be discovered. In the example below, there are drops in the cInteractiveVideo queue.



Appendix D

QoS Glossary

CBQoS MIB—Class-Based QoS MIB
Cisco-specific SNMP management information base (MIB) for monitoring QoS behavior on a router. Contains QoS statistics for traffic traversing an interface configured for MQC.

CBWFQ—Class-Based Weighted Fair Queuing
MQC queuing mechanism for handling AF classified bulk and transactional traffic that is not time-, loss-, or jitter-sensitive.

Classification
Identifies traffic by protocol header fields such as IP addresses, DSCP, UDP/TCP port numbers, or higher layer protocol identification using deeper packet inspection.

Congestion Avoidance
Using WRED and packet drop mechanisms to manage traffic before times of congestion.

Header Compression
Removal of redundant header information to improve the overall throughput of the traffic (typically VoIP packets).

Link Fragmentation
The process of splitting large packets into smaller packets that are then reassembled at the end point. Used on low bandwidth links where smaller, high-priority traffic (such as VoIP) can be interleaved with fragments to avoid propagation delays.

LLQ—Low Latency Queuing
MQC queuing mechanism for handling EF classified high priority, time-, loss-, or jitter-sensitive traffic. This is also known as priority queuing.

Appendix D

QoS Glossary

Marking
Uses classification results and applies specific settings to protocol header fields to specify the priority or importance of the traffic.

MQC—Modular QoS Command Line Interface
A common set of functions and their commands for implementing QoS across Cisco routers.

NBAR—Network-Based Application Recognition
MQC classification mechanism using deep packet inspection to identify Layer 7 protocols such as Citrix, Skype, SCCP, SIP, eDonkey and more.

Policing
Checks classified traffic flow rates against pre-determined values and re-marks or drops packets that exceed these thresholds.

Queues
High-performance memory buffers in router where data is held to be processed. Different types of queues are typically used for real-time and non-real-time traffic. Real-time traffic queues will be low latency and low jitter.

Shaping
Checks classified traffic flow rates against pre-determined values and uses queues to buffer and prioritize packets that exceed these thresholds.

More Information

On the Web

To learn more and see how LiveAction delivers unmatched network visibility, visit www.liveaction.com. Follow us on LinkedIn and Twitter to receive regular news and updates on our handbooks, product literature, and more.

Software Tools for Cisco QoS

Visit our website for a free trial download of our LiveAction software for monitoring and configuring Cisco QoS.

About LiveAction

LiveAction provides end-to-end visibility of network and application performance from a single pane of glass. This gives enterprises confidence that the network is meeting business objectives, offers IT administrators full visibility for better decision making, and reduces the overall cost of operations. By unifying and simplifying the collection, correlation and presentation of application and network data, LiveAction empowers network professionals to identify, troubleshoot and resolve issues across increasingly large and complex networks proactively and quickly. ■

LiveAction®

LiveAction
3500 West Bayshore Rd
Palo Alto, CA 94303

Phone + eFAX: +1 888-881-1116
Email: sales@liveaction.com
Website: www.liveaction.com