Quality of Service (QoS) for converging Service Technologies

Miroslav Slavov and Pencho Penchev
The Faculty of Electrical Engineering and Electronics
at Technical University of Gabrovo
4 H. Dimitar str
Gabrovo 5300, Bulgaria
miroslav_slavov@mail.bg

ABSTRACT: *QoS isn't something which we configure on our Cisco router. Instead, it's an umbrella term that encompasses a wide range of mechanisms used to control traffic patterns on the network. QoS has already proven to be the go-to technology for converging voice, video and data networks. But as business needs change, so do demands for QoS technologies. This paper will cover the basics of QoS, its needs, and a few types of available QoS mechanisms.*

Keywords: Quality of Service (QoS), Cisco QoS Toolset, QoS Levels

Received: 5 April 2023, Revised 11 June 2023, Accepted 26 June 2023

DOI: 10.6025/jio/2023/13/3/77-84

1. Introduction

Quality of Service (QoS) is a set of capabilities that allows delivering differentiated services for network traffic, thereby providing better service for selected network traffic. QoS expedites the handling of mission-critical applications, while sharing network resources with noncritical applications.

QoS also ensures the available bandwidth and minimum delays required by time-sensitive multimedia and voice applications. This allows using expensive network connections more efficiently, and to establish service level agreements with customers of the network.

QoS features provide better and more predictable network service by:

- Supporting dedicated bandwidth for critical users and applications.
- Controlling jitter and latency (required by real-time traffic).
- Avoiding and managing network congestion.
- Shaping network traffic to smooth the traffic flow.
- Setting traffic priorities across the network [2][3].

2. Defining Quality of Service

2.1. What is QoS?

QoS is the measure of transmission quality and service availability of a network (or internetworks).

Service availability is a crucial foundation element of QoS. The network infrastructure must be designed to be highly available before you can successfully implement QoS. The transmission quality of the network is determined by the following factors:

- Loss – a relative measure of the number of packets that were not received compared to the total number of packets transmitted. Loss is typically a function of availability.

- Delay – the finite amount of time it takes a packet to reach the receiving endpoint after being transmitted from the sending endpoint.

- Delay variation (Jitter) – the difference in the end-to-end delay between packets [1][4].

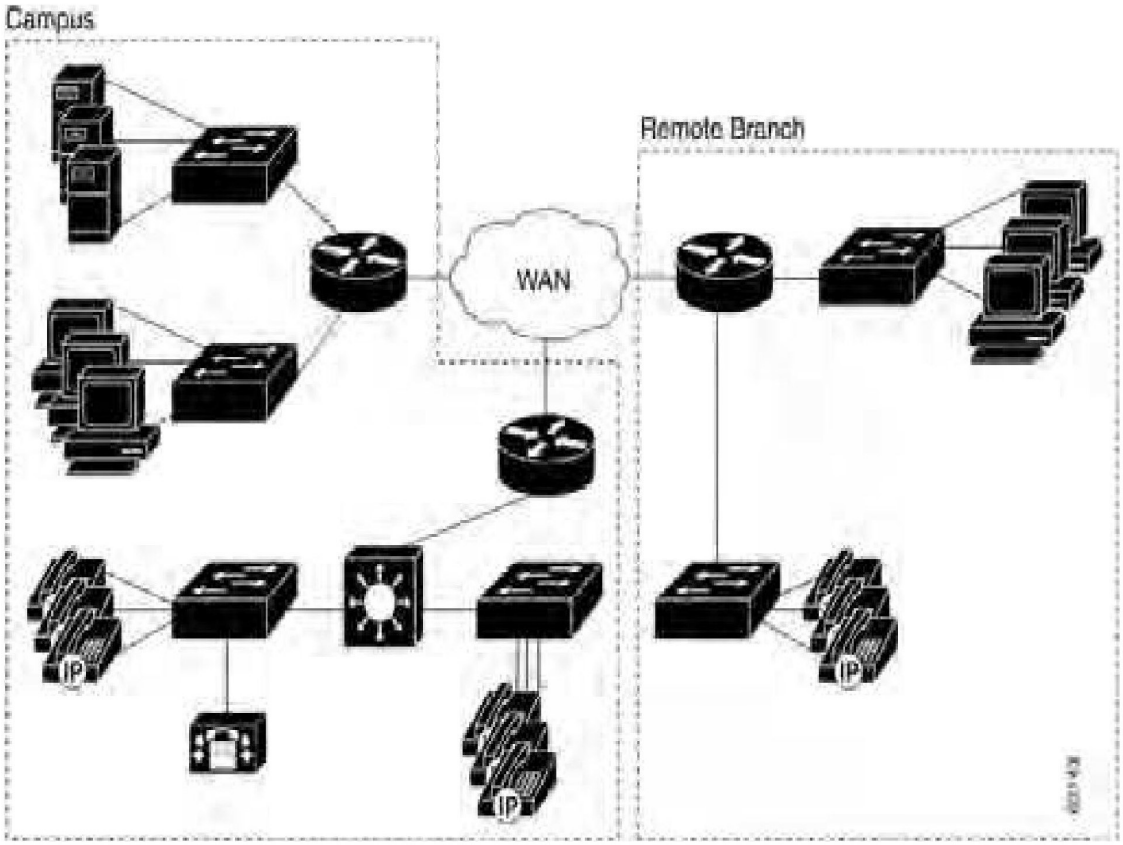

Figure 1. Example of an Enterprise Network

2.2. Why is QoS Important for Enterprise Networks?

QoS technologies refer to the set of tools and techniques to manage network resources and are considered the key enabling technology for network convergence.

QoS tools are not only useful in protecting desirable traffic, but also in providing deferential services to undesirable traffic such as the exponential propagation of worms [4].

The WAN devices can limit the bandwidth available to the traffic, or give the traffic priority, or even change the classification of the traffic. In this way, you can provide endto- end QoS in your network.

Figure 1 shows an example of an enterprise network. Typically, you classify traffic in the LAN before sending it to the WAN. The devices on the WAN then use the classification to determine the service requirements for the traffic [3].

2.3. Which Applications Need QoS?

Understanding the Characteristics of Applications It is important to understand the characteristics of the applications that need protection. Some applications tend to be sensitive to latency or packet loss, while others are considered "aggressive" because they are bursty or consume a lot of bandwidth. If the application is bursty, determine if there is a constant burst or a small burst. Is the packet size of the application large or small? Is the application TCP or UDP based [5]?

Characteristic	Guideline
Application thatis delay– or loss – sensitive. (Voice and Real Time Video)	Do not use weighted random early detection (WRED), traffic shaping, fragmentation (FRF–12 (describes the method of fragmenting Frame Relay frames into smaller frames)), or policing. For this kind of traffic, you should implement Low Latency Queuing (LLQ)and use a priority queue for the delay–sensitive traffic.
Application that is consistently bursty or is a bandwidth hog. (FTP and HTTP)	Use WRED, policing, traffic shaping, or class–based weighted fair queuing (CBWFQ) to guarantee bandwidth.
Application that is TCP–based.	Use WRED since lost packets cause TCP to back off and then ramp up again using the slow–start algorithm. If the traffic is UDP–based and does not change its behaviour when packets are dropped, do not use WRED. Use Policing if you need to rate–limit the application; otherwise just let the packets tail–drop.

Table 1. Applications that Required QoS

3. Cisco QoS Toolset

This section describes the main categories of the Cisco QoS toolset and includes the following topics (Figure 2):

- Classification and Marking tools
- Policing and Markdown tools
- Scheduling tools
- Link-specific tools
- AutoQoS tools
- Call Admission Control tools

3.1. Classification and Marking Tools

Classification and marking tools set this trust boundary by examining any of the following:

- **Layer 2 Parameters**—802.1Q Class of Service (CoS) bits, Multiprotocol Label Switching Experimental Values (MPLS EXP).

- **Layer 3 Parameters**—IP Precedence (IPP), Differentiated Services Code Points (DSCP), IP Explicit Congestion Notification (ECN), source/destination IP address

- **Layer 4 Parameters**— L4 protocol (TCP/UDP), source/destination ports

- **Layer 7 Parameters**— application signatures via Network Based Application Recognition (NBAR)

NBAR is a Cisco proprietary technology that identifies application layer protocols by matching them against a Protocol Description Language Module (PDLM), which is essentially an application signature. The NBAR deep-packet classification engine examines the data payload of stateless protocols against PDLMs. There are over 98 PDLMs embedded into Cisco IOS software 12.3 code.

Additionally, Cisco IOS software 12.3(4)T introduces the ability to define custom PDLMs which examine user-defined strings within packet payloads.

PDLMs can be added to the system without requiring an IOS upgrade because they are modular. NBAR is dependent on Cisco Express Forwarding (CEF) and performs deep-packet classification only on the first packet of a flow. The remainder of the packets belonging to the flow is then CEF-switched.

Within an enterprise, marking is done at either Layer 2 or Layer 3, using the following fields:

- 802.1Q/p Class of Service (CoS)—Ethernet frames can be marked at Layer 2 with their relative importance by setting the 802.1p User Priority bits of the 802.1Q header. Only three bits are available for 802.1p marking. Therefore, only 8 classes of service (0-7) can be marked on Layer 2 Ethernet frames.

- IP Type of Service (ToS) byte—Layer 2 media often changes as packets traverse from source to destination, so a more ubiquitous classification occurs at Layer 3. The second byte in an IPv4 packet is the ToS byte. The first three bits of the ToS byte are the IPP bits. These first three bits combined with the next three bits are known collectively as the DSCP bits.

- DSCPs and Per-Hop Behaviors (PHBs)—DSCP values can be expressed in numeric form or by special standardsbased names called Per-Hop Behaviors. There are four broad classes of DSCP PHB markings: Best Effort (BE or DSCP 0), RFC 2474 Class Selectors (CS1–CS7, which are identical/backwards-compatible to IPP values 1–7), RFC 2597 Assured Forwarding PHBs (AFxy), and RFC 3268 Expedited Forwarding (EF).

DSCP values can be expressed in decimal form or with their PHB keywords. For example, DSCP EF is synonymous with DSCP 46, and DSCP AF31 is synonymous with DSCP 26.

• IP Explicit Congestion Notification (IP ECN)—IP ECN, as defined in RFC 3168, makes use of the last two bits of the IP ToS byte, which are not used by the 6-bit DSCP markings, as shown in Figure 3.

3.2. Policing and Markdown Tools

Policing tools (policers) determine whether packets are conforming to administratively-defined traffic rates and take action accordingly. Such action could include marking, remarking or dropping a packet.

A basic policer monitors a single rate: traffic equal to or below the defined rate is considered to conform to the rate, while traffic above the defined rate is considered to exceed the rate. On the other hand, the algorithm of a dual-rate policer (such as described in RFC 2698) is analogous to a traffic light.

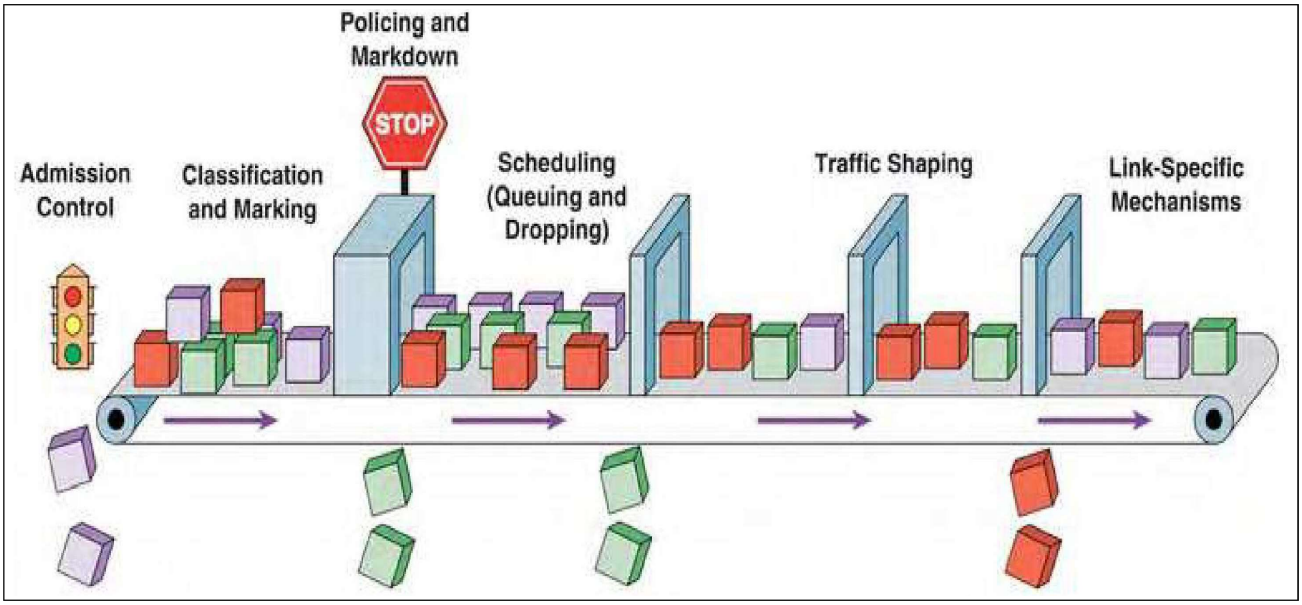

Figure 2. The Cisco QoS Toolset

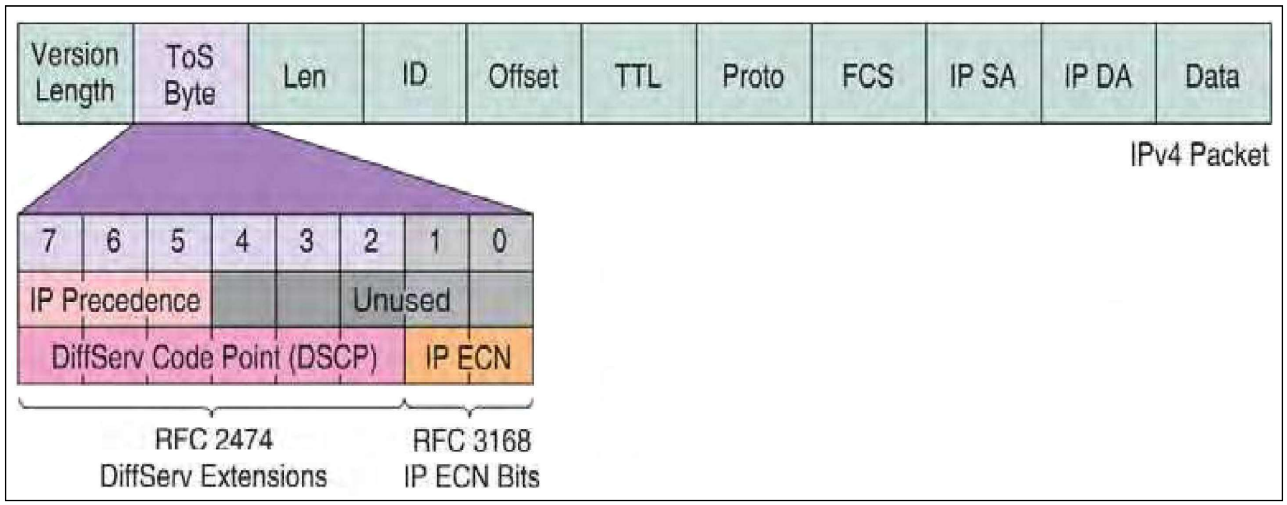

Figure 3. The IP ToS Byte (DSCP and IP ECN)

3.3. Scheduling Tools

Scheduling tools determine how a frame/packet exits a device. Whenever packets enter a device faster than they can exit it, such as with speed mismatches, then a point of congestion, or bottleneck, can occur. Devices have buffers that allow for scheduling higher-priority packets to exit sooner than lower priority ones, which is commonly called queuing.

Queuing algorithms are activated only when a device is experiencing congestion and are deactivated when the congestion clears. Figure 4 shows the Layer 3 and Layer 2 queuing subsystems of the Cisco IOS (Internetwork Operating System) software LLQ/CBWFQ algorithm.

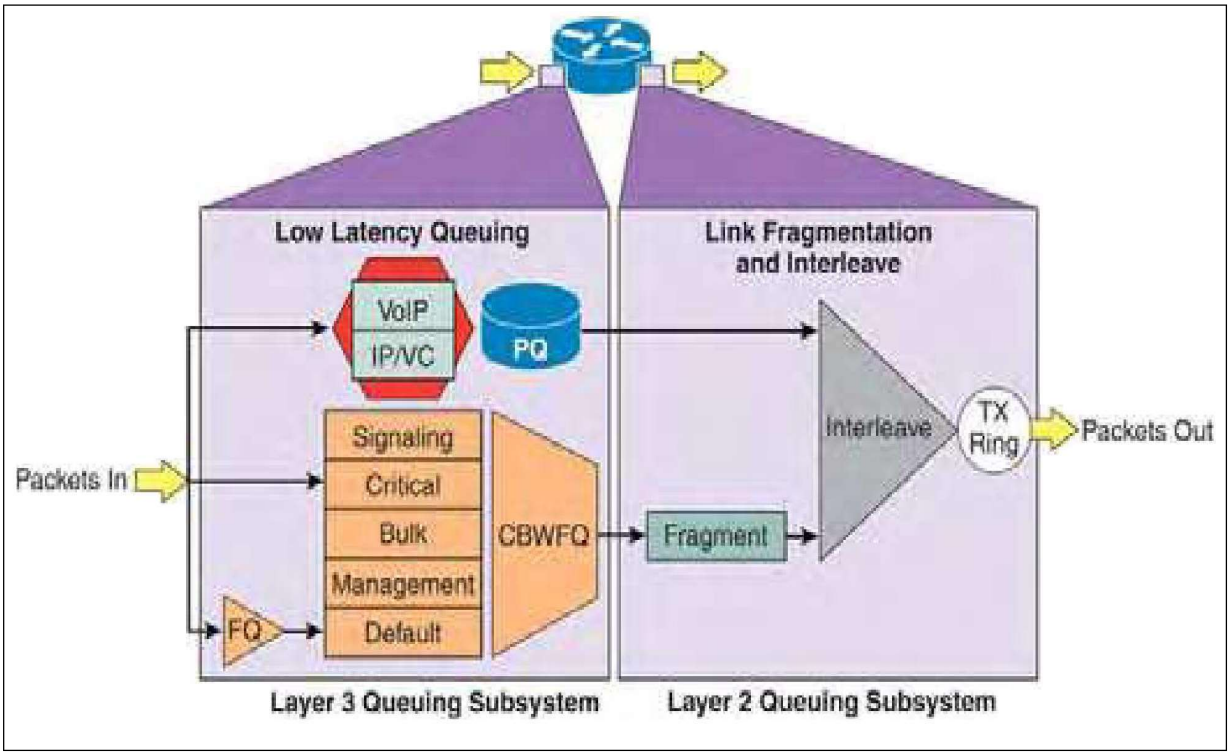

Figure 4. The IP ToS Byte (DSCP and IP ECN)

Selective dropping of packets when the queues are filling is referred to as congestion avoidance. Congestion avoidance mechanisms work best with TCP-based applications because selective dropping of packets causes the TCP windowing mechanisms to “throttle-back” and adjust the rate of flows to manageable rates.

The principle IOS congestion avoidance mechanism is WRED, which randomly drops packets as queues fill to capacity.

3.4. Link-Specific Tools

Link-specific tools include the following:

• Shaping tools—A shaper typically delays excess traffic above an administratively-defined rate using a buffer to hold packets and shape the flow when the data rate of the source is higher than expected.

• Link Fragmentation and Interleaving tools—With slowspeed WAN circuits, large data packets take an excessively long time to be placed onto the wire.

• Compression tools—Compression techniques, such as compressed Real-Time Protocol (cRTP), minimize bandwidth requirements and are highly useful on slow links. At 40 bytes total, the header portion of a VoIP packet is relatively large and

can account for nearly two-thirds or the entire VoIP packet (as in the case of G.729 VoIP).

• Transmit ring (Tx-Ring) tuning—The Tx-Ring is a final interface First-In-First-Out (FIFO) queue that holds frames to be immediately transmitted by the physical interface. The Tx- Ring ensures that a frame is always available when the interface is ready to transmit traffic, so that link utilization is driven to 100 % of capacity. The size of the Tx-Ring is dependent on the hardware, software, Layer 2 media, and queuing algorithm configured on the interface. E. AutoQoS Tools

The richness of the Cisco QoS toolset inevitably increases its deployment complexity. To address customer demand for simplification of QoS deployment, Cisco has developed the Automatic QoS (AutoQoS) features. AutoQoS is an intelligent macro that allows an administrator to enter one or two simple AutoQoS commands to enable all the appropriate features for the recommended QoS settings for an application on a specific interface.

For Campus Catalyst switches, AutoQoS automatically performs the following tasks:

• Enforces a trust boundary at Cisco IP Phones.

• Enforces a trust boundary on Catalyst switch access ports and uplinks/downlinks.

• Enables Catalyst strict priority queuing for voice and weighted round robin queuing for data traffic.

• Modifies queue admission criteria (CoS-to-queue mappings).

• Modifies queue sizes as well as queue weights where required.

• Modifies CoS-to-DSCP and IP Precedence-to-DSCP mappings.

For Cisco IOS routers, AutoQoS is supported on Frame Relay (FR), Asynchronous Transfer Mode (ATM), High-Level Data Link Control (HDLC), Point-to-Point Protocol (PPP), and FR-to-ATM links.

The AutoQoS Enterprise feature consists of two configuration phases, completed in the following order:

• Auto Discovery (data collection)—Uses NBAR-based protocol discovery to detect the applications on the network and performs statistical analysis on the network traffic.

• AutoQoS template generation and installation— Generates templates from the data collected during the Auto Discovery phase and installs the templates on the interface. These templates are then used as the basis for creating the class maps and policy maps for your network. After the class maps and policy maps are created, they are then installed on the interface.

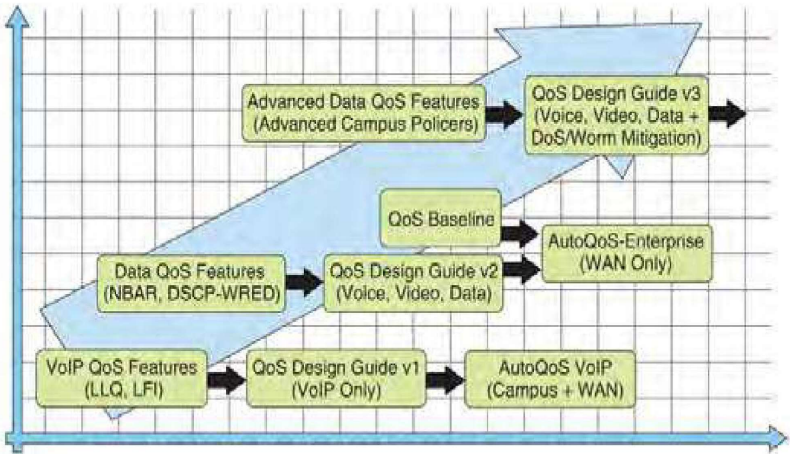

Figure 5. Cisco QoS Feature, Design Guide and AutoQoS Evolution

Figure 5 shows the relationship between Cisco QoS features, Design Guides, and AutoQoS.

4. Conclusion

The purpose of this paper is to describe the terms and the main concept of Quality of Service (QoS). In the paper a set of tools, used for maintaining the QoS are described. These are the basics, needed for future researches of QoS and finding a way to improve it in different types of networks.

Acknowledgement

This paper has been sponsored by E 1102 project of Technical University of Gabrovo.

References

[1] Flannagan, M., et al. (2001). Administering Cisco QoS in IP Networks. Rockland: Syngress. ISBN: 1-928994-21-0.

[2] Szigeti, T. (2005). End-to-End QoS Network Design. Indianapolis: Cisco Press. ISBN: 1-58705-176-1.

[3] User Guide for CiscoWorks QoS Policy Manager Software Version 4.1.5. (2010). Cisco Systems Inc.

[4] Enterprise QoS Solution Reference Network Design Guide Version 3.3. (2008). Cisco Systems Inc.

[5] Implementing Quality of Service. Document ID: 13747. (2008–2009). Cisco Systems Inc.