

The Case for Quality of Service (QoS)

Implementing QoS as Part of a Complete Application Acceleration Solution

Quality of Service (QoS) is used to optimize performance in the presence of network impairments and to give preferential treatment to certain classes of traffic in the event that demand exceeds available bandwidth.

As each individual enterprise application has its own delivery requirements (e.g., latency and jitter) and enterprises sometimes desire that varying levels of priority be given to different types of traffic, QoS is indispensable to enterprises deploying a mix of applications across a Wide Area Network (WAN). It ensures that each application is treated appropriately as it vies for limited WAN resources.

Application acceleration solutions do not mitigate the need for QoS across the WAN. While these solutions improve bandwidth efficiency and perceived application response times, they do not eliminate the fact that a mix of traffic with varying levels of priority must be delivered over a fixed resource. As a result, QoS is still essential for guaranteeing the quality of application delivery and ensuring predictable behavior across an optimized WAN.

What is the best way to implement QoS in conjunction with application acceleration solutions? What role should the WAN router play vs. the application acceleration appliance? Can enterprises effectively enforce QoS across all applications, including both accelerated and non-accelerated traffic?

To successfully deliver applications across a distributed enterprise, IT staff require a clear QoS strategy that is tightly integrated with application delivery requirements. An advanced application acceleration solution, such as the one offered by Silver Peak, will provide a variety of QoS options to enterprises, giving them the flexibility to support numerous applications and traffic types in an optimized manner.

IMPLEMENTING QoS

QoS involves two functions: 1) classification of packets into traffic classes based on source, destination and/or application and, 2) queuing and service mechanisms that are used to apply service policies based on these classifications.

QoS classification is typically performed either in the WAN router, or by an “upstream” device (e.g., host or Ethernet switch) and then “honored” by the WAN router. Once classified, QoS queuing and enforcement is typically handled within most LAN switches, LAN/WAN routers, and hosts. These devices leverage multiple queues and a variety of traffic management policies that enable them to prioritize traffic and intelligently handle packets during periods of congestion.



Application acceleration appliances introduce a challenge to the way that QoS is implemented. These devices sit on the LAN side of the WAN router, in both the data center and branch offices. As traffic flows from the internal network to the WAN, application acceleration appliances remove repetitive information, compress headers and payload content, modify IP addresses and port numbers, and sometimes even encrypt traffic. By completely obscuring the original data (and its headers), application acceleration appliances prevent downstream devices, such as WAN routers, from applying QoS classification logic based on normal packet inspection.

To account for this, QoS classification can be performed upstream of the application acceleration appliance (by a host, Ethernet switch, or LAN router) and then be honored by the appliance. Or, QoS classification can be performed by the application acceleration appliance itself. This latter approach is often more desirable as enterprises sometimes

do not have intelligent upstream devices in a branch office that are capable of performing QoS classification.

With respect to QoS queuing and service disciplines, most application acceleration appliances can pass existing tags to downstream WAN routers, enabling these devices to participate in the QoS process as they would normally. In addition, queuing and service disciplines can be enforced within the application acceleration appliance itself. This is often desired, because application acceleration appliances are well equipped to collect real-time metrics, like packet loss and delay, and adapt QoS techniques accordingly. (This is usually not part of a typical WAN router's feature set.)

THE SILVER PEAK SOLUTION

Silver Peak provides a variety of QoS options to enterprises. In addition to honoring existing QoS markings, the Silver Peak solution provides native support for advanced QoS, including sophisticated classification logic, a variety of packet marking techniques, queuing, and traffic shaping. These features are described as follows:

Packet Marking

Packet marking provides a way for network elements to provide different levels of service to different packets, based on markings in the IP header. Using packet marking avoids the need to reclassify packets by deep inspection at each hop, and avoids the need to maintain per hop state within the network. By marking tunnel packets, Silver Peak NX Series appliances can provide hints that enable downstream devices, such as WAN routers, to treat each packet appropriately. Silver Peak's packet marking functionality can be used even when other QoS features, such as shaping, are disabled.

IP packets contain an 8-bit Type of Service (TOS) field which may be used to convey type of service / quality of service information. Differentiated services control point (DSCP), or Diffserv as it is commonly called, uses the first six bits of the ToS byte to specify a particular per-hop behavior for each packet. DSCP is used to specify up to 64 different forwarding behaviors.

	Without Application Acceleration	With Application Acceleration
QoS (classification)	<ul style="list-style-type: none"> Performed in WAN router, or Performed in upstream device and then honored by WAN router 	<ul style="list-style-type: none"> Performed upstream, by a host, Ethernet switch, or LAN router; and then honored by the appliance, or Performed by the application acceleration appliance itself
QoS (enforcement)	Performed in LAN switch, router, host, and other network devices	<ul style="list-style-type: none"> Tags transparently passed to downstream WAN routers, or Queuing and service disciplines enforced within the application acceleration appliance



DiffServ relies on packet classification to take place elsewhere in the network. Classification is often applied to “flows” of traffic, with a flow containing 5 basic elements: source IP address, destination IP, source port, destination port, and the transport protocol. Some protocols, however, like the File Transfer Protocol (FTP) and Voice over IP (VoIP), cannot use flow mapping as they dynamically assign ports. These applications require deep packet inspection or proxying to perform QoS classification.

The use of tunnels requires special consideration when marking packets for QoS, since a tunnel has a TOS field in the outer (tunnel) packet and a TOS field within the inner packet. To address this, Silver Peak NX Series appliances can perform the following functions at the entrance to the tunnel:

- Mark outer (tunnel) packets based on its own specified criteria
- Mark outer (tunnel) packets based on the inner packet marking
- Mark inner packets based on its own criteria
- Leave inner packet markings alone

Silver Peak enables different DSCP mappings to be used for the service provider and enterprise portions of a network. As there is often no direct correlation between QoS mappings in these portions of the network, Silver Peak provides greater flexibility for end-to-end QoS.

Application Classification

Application acceleration solutions require robust application classification capabilities to enable different priorities and handling instructions to be applied to individual types of traffic. Silver Peak employs a variety of techniques to achieve this, which include:

- **Five-tuple Filters.** Applications can be classified using the five basic elements of a flow: source IP, destination IP, source port, destination port, and protocol. For example, all web traffic can be identified by the fact that it uses Port 80, or application specific traffic can be identified by the IP address of the application server. Silver Peak enables QoS policies to be defined by taking into account specific flow elements for a particular application, including source/destination and protocol.
- **Connection Tracking.** Some applications, like FTP and VoIP, initiate connections that do not use well known ports for data transfer. Stateful classification and proxy functionality are required in order to track these more complex applications.

Queuing and Shaping

The most likely congestion points in a typical enterprise are on the near and far sides of the WAN. There are many QoS mechanisms that can be applied at these congestion points to improve traffic delivery, including:

- **Queuing policies,** including queue by flow and queue by application class
- **Dropping disciplines,** including Random Early Detection (RED), whereby packets are randomly dropped prior to periods of high congestion, and tail drop, where all packets are dropped when output buffers are filled.
- **Service disciplines,** including weighted round robin (WRR) and class-based queuing

However, in many instances, WAN congestion points may not support QoS mechanisms. Or, as is the case when the far end is part of a service provider's network, it may not be possible for an enterprise to control these features. Therefore, significant benefit can be attained by “moving” the congestion point into the Silver Peak appliance, where these mechanisms can be employed. This is achieved by shaping traffic within the Silver Peak appliance to stay below known bottleneck rates. In doing this, packets that would otherwise back up in queues throughout the network (without QoS) are instead aggregated within the Silver Peak appliance, where comprehensive QoS mechanisms can be employed for intelligent packet handling.



The NX Series appliances are equipped with a variety of tools for traffic shaping, which include:

- **Interface Traffic Shaping:** This shapes the outbound WAN-facing interface of the Silver Peak appliance. By shaping the outbound traffic to something less than the WAN bandwidth limit, the Silver Peak appliance is able to apply advanced QoS and congestion control techniques without relying solely on the router's QoS capabilities.
- **Tunnel Traffic Shaping:** This shapes the tunnel that is used to connect any two Silver Peak appliances. By shaping the tunnel to something less than the smallest link or connection limit on the path between the two appliances, the Silver Peak appliance can apply advanced QoS and congestion control techniques without relying solely on the service provider network's QoS.
- **Class-based Service Algorithm:** Specific classes of service can be established within each tunnel. Bandwidth is allocated to these classes taking into account a variety of factors, including minimum service rate (i.e. guaranteed service rate), absolute priorities (e.g. 1 to 10), excess service weight (e.g. 1 to 100), and maximum service rate (e.g., per class shaping).
- **Pass-through Traffic:** In some instances, Silver Peak NX Series appliances will pass through traffic rather than direct it into tunnels. This will occur, for example, when the destination location does not have a Silver Peak appliance in place. In these instances, the WAN router interface can be a bottleneck for both the accelerated and non-accelerated traffic. Silver Peak enables

QoS policies to be created for both of these types of traffic, and ensures that bandwidth is allocated appropriately between tunnel traffic and pass-through traffic.

CONCLUSION

As enterprises move towards a centralized model for branch office servers and storage, new solutions are required to improve application delivery. While these solutions increase bandwidth efficiency and improve perceived application response time, they do not eliminate the need for robust Quality of Service capabilities. Certain types of traffic, such as voice and data, will almost always benefit from QoS when being delivered across a shared resources.

When application acceleration solutions are deployed on both ends of a WAN link, special consideration needs to be given to how QoS is delivered. Traditional methods of performing QoS classification and enforcement in the WAN router are often no longer viable as application acceleration solutions modify headers and obscure payload information.

At a minimum, enterprises should ensure that their application acceleration solution can honor existing QoS schemes by honoring existing markings and transparently passing tags when appropriate. However, significant benefits can be achieved when QoS is implemented directly within the application acceleration appliances. By taking this approach, enterprises can ensure that QoS is implemented in branch offices where other

networking equipment lacks classification and enforcement capabilities. In addition, application acceleration appliances are ideally equipped to collect real-time metrics, like packet loss and delay, which are not collected in other devices. This makes application acceleration appliances uniquely suited to apply a wide variety of QoS techniques to enterprise traffic.

Enterprises should ensure that their application acceleration solution provides different QoS options to accommodate unique traffic requirements. Fundamental QoS capabilities include packet marking, application classification, queuing, and traffic shaping. In addition, QoS must be applied across all traffic – accelerated and otherwise. By satisfying all of these QoS requirements, Silver Peak offers a robust application acceleration platform that ensures consistent and reliable service across an entire distributed enterprise.

