

Three Steps to Success with QoS

A Riverbed White Paper

Introduction: QoS ensures predictable application performance

QoS is one of the most widely deployed networking technologies. It is a critical capability that makes WANs work for individual applications like VoIP and it's essential when using Internet links for delivering business applications (such as SaaS). Nevertheless, there's a limit to what legacy QoS can do when there are dozens or hundreds of applications competing for the same bandwidth. A similar problem exists when SaaS and public cloud applications fail because of congestion caused by recreational traffic. This is where more advanced QoS solutions are needed to ensure that key applications perform predictably at all times. Riverbed has taken a fresh look at QoS in the light of modern applications and architectures and has delivered a powerful, easy to manage QoS solution that delivers predictable performance for all business applications across today's network architectures.

In this paper, we'll first discuss some basics about QoS including: how to get started, how it's typically deployed, and what it can and can't do. Then we'll discuss the 3 basic steps needed to deploy a QoS solution using Steelhead appliances.

Getting Started

Guiding Principles

QoS provides a way to identify and protect important network traffic. QoS also restricts applications that behave greedily or that are not desirable on the network. When deploying a QoS solution, in general, it's best to first set business goals around what behavior is desired for key applications, and then design the minimum set of policies needed to accomplish those goals. Defining a policy for every application and traffic type is possible, but can lead to an overly complex configuration that is difficult to analyze and troubleshoot. As a guiding principle, aim to protect what's important and contain what's not. Typically, all other traffic that does not have an explicit policy will also benefit because there will be less WAN congestion from unimportant applications.

Shaping vs. Marking

Using Riverbed Steelhead QoS features, network traffic can be shaped, marked, or both:

- **Shaping** means controlling how much minimum and maximum bandwidth can be used and prioritizing how application traffic is sent across the WAN
- **Marking** means "tagging" packets and letting the WAN router/MPLS network shape the traffic based on the tags

The most common use of QoS marking is for Voice over IP (VoIP) traffic. Network Carriers provide one or more Differentiated Services Code Point (DSCP) values that, when applied to packets, determine how they are handled across an MPLS WAN. VoIP is usually assigned to the Expedited Forwarding (EF) class by using the DSCP value of 46. There are other common DSCP values that correspond to varying levels of service across the MPLS cloud. These values typically fall into the Assured Forwarding (AF) category and are used for other applications, each according to their importance and loss/latency-sensitivity.

In addition to marking, it is common to implement shaping especially when multiple applications have to share the same DSCP value or when bad behavior by a single user can negatively affect other users of the same application (e.g. Citrix file transfers). Shaping is also used when more granular control over application performance is needed than MPLS alone can offer. This can reduce the cost of a MPLS link by making the best use of the existing capacity and avoiding the need to add more services to the link.

Compatibility with QoS elsewhere on the network

Network Carriers offer multiple Classes of Service (CoS) on the MPLS links they offer. This provides a way to match the urgency of each application with the right level of service across the WAN. This matching of urgency and service is very well aligned with what Steelhead QoS provides through Hierarchical Fair Service Curve (HFSC) scheduling. When using MPLS, a range of predefined DSCP markings is used to funnel each application into the appropriate class of service (CoS) across the WAN. By default, this is done using Layer 3/4 parameters such as IP address, port, and protocol. This is a difficult configuration to manage because it relies on configuring router access lists (that are often managed by the Carrier). And it is often inaccurate when multiple applications re-use the same port numbers (like HTTP applications).

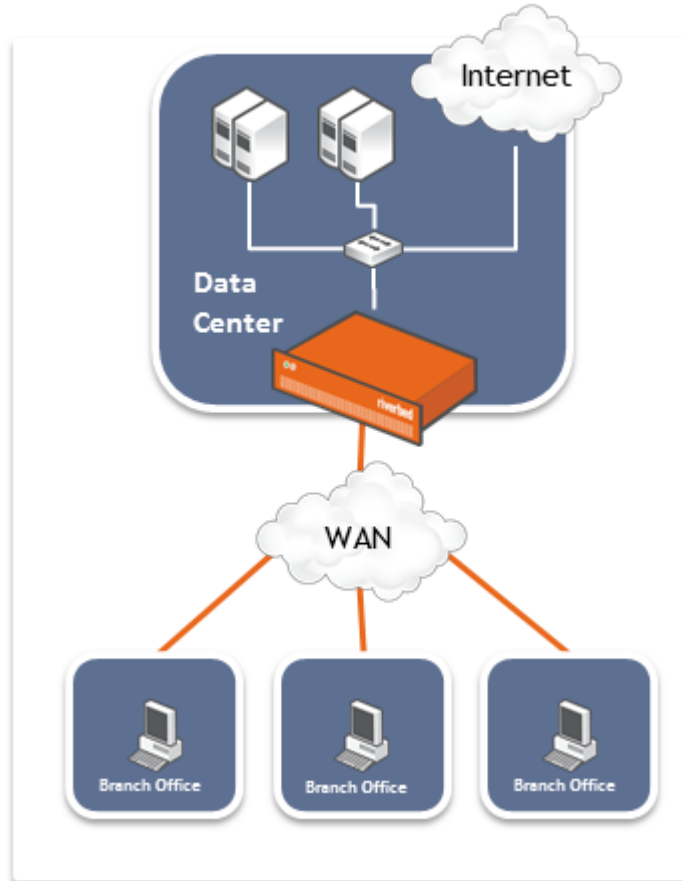
For these reasons, ensuring the best and most predictable levels of performance for applications requires deploying devices with

advanced QoS capabilities. These devices control application traffic more accurately and easily. With its advanced QoS capabilities, the Steelhead appliance is the best device to perform QoS because it can accurately identify all applications (both optimized and pass-through) and apply the right marking and shaping policies.

Typical QoS Deployment Scenarios

This section outlines the most common deployment scenarios as well as what QoS can and cannot do in specific situations.

Data center-only QoS Deployments



QoS is used at data centers to ensure that bandwidth is being allocated properly and to limit outbound traffic so that smaller links at branch offices aren't overwhelmed. Recreational traffic from applications like Facebook and YouTube can be slowed down so that business traffic from applications like Citrix and SharePoint can get the bandwidth they need. Traffic can be allocated properly to avoid overwhelming the remote site's WAN link.

Key Use Cases:

- Controlling bandwidth when Internet access is provided centrally from the data center to the remote sites.
- Conforming with MPLS bandwidth limits for specific types of traffic (such as VoIP) from the data center

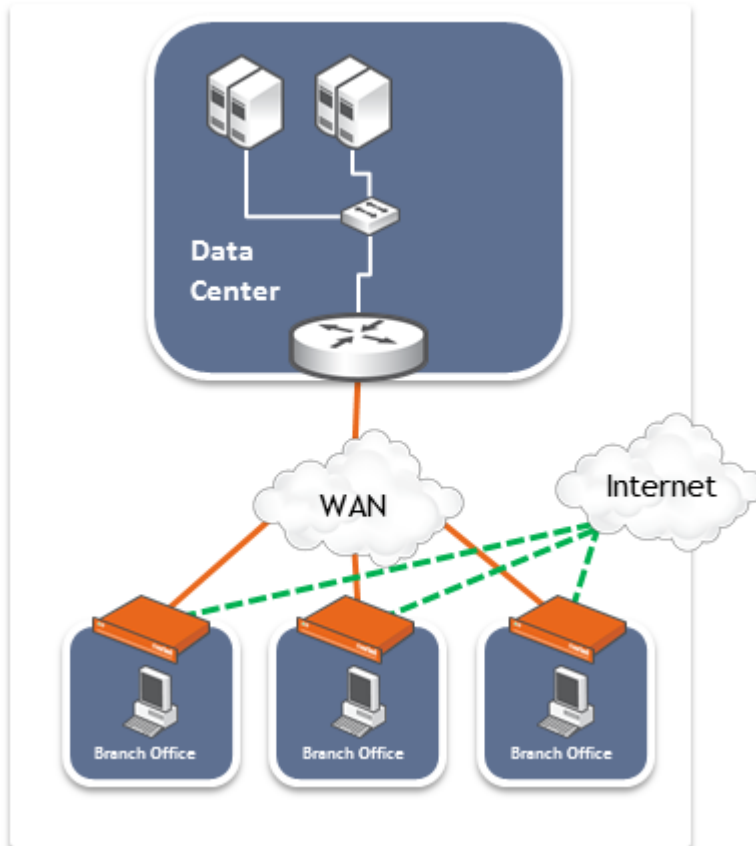
What can QoS do in this scenario?

- Guarantee that enough bandwidth is available for specific applications, sites, or classes of traffic
- Allocate bandwidth fairly among all sites, classes, and applications, preventing any group of users or applications from monopolizing bandwidth
- Limit the amount of bandwidth used by specific applications or classes of traffic. This is useful both for restricting recreational traffic and for controlling premium applications like VoIP, where excess usage charges are possible.
- Prioritize and limit the traffic sent to each remote site to avoid overwhelming the (typically) smaller remote site WAN link

What can't QoS do in this scenario?

- QoS can't control traffic that doesn't traverse the data center links. A basic example is branch-to-branch VoIP traffic. No QoS solution can accurately control this scenario without being deployed in at least one of the branch offices.

Branch office-only QoS Deployments



Deployments where QoS is deployed exclusively in branch offices has not traditionally been a common QoS deployment scenario, but is being increasingly considered as internet connectivity is added to the branch and as key applications are being hosted and accessed in public clouds. In this scenario, it's important to control inbound Internet traffic to prevent file transfers and recreational applications like Facebook and YouTube from crowding out SaaS applications and UCC (Unified Communications and Collaboration) applications such as Voice and Videoconferencing.

Key Use Cases:

- Protecting SaaS/Public Cloud applications from recreational traffic on local Internet links
- Making room for VoIP and UCC traffic on fully-meshed networks

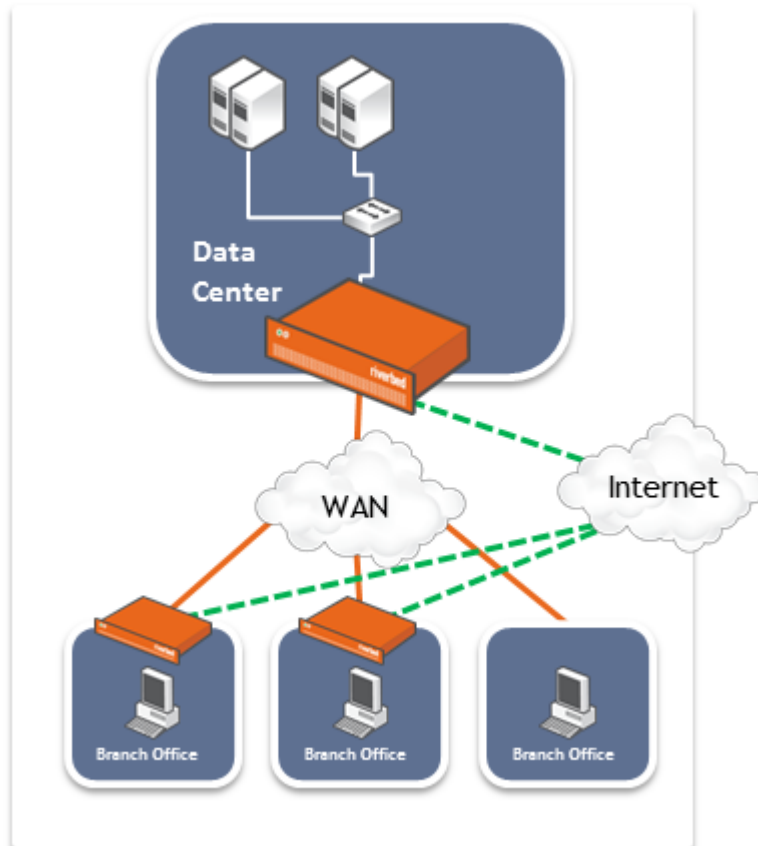
What can QoS do in this scenario?

- Reserve bandwidth for important TCP- and UDP-based applications
- Slow down TCP applications to make room for more important applications

What can't QoS do in this scenario?

- QoS can't control inbound UDP traffic from the Internet (and from the data center, in this case). No QoS technology can do this without being able to control UDP from an upstream location.

Data Center + Multiple Branch QoS Deployments



The most common QoS deployment scenario is where QoS is deployed in the data center as well as in branch offices. From the data center, a basic level of control for all remote sites is provided. Additionally, some of the branch offices with high value applications or highly congested WAN links can implement finer-grained controls to ensure the best possible performance for key applications.

Key Use Cases:

- Protecting key business applications running at branch offices
- Reducing congestion on small or oversubscribed WAN links (in cooperation with WAN optimization)
- Controlling bandwidth when Internet access is provided centrally from the data center to the remote sites
- Conforming with MPLS bandwidth limits for specific types of traffic (such as VoIP)
- Protecting SaaS/Public Cloud applications from recreational traffic on local Internet links

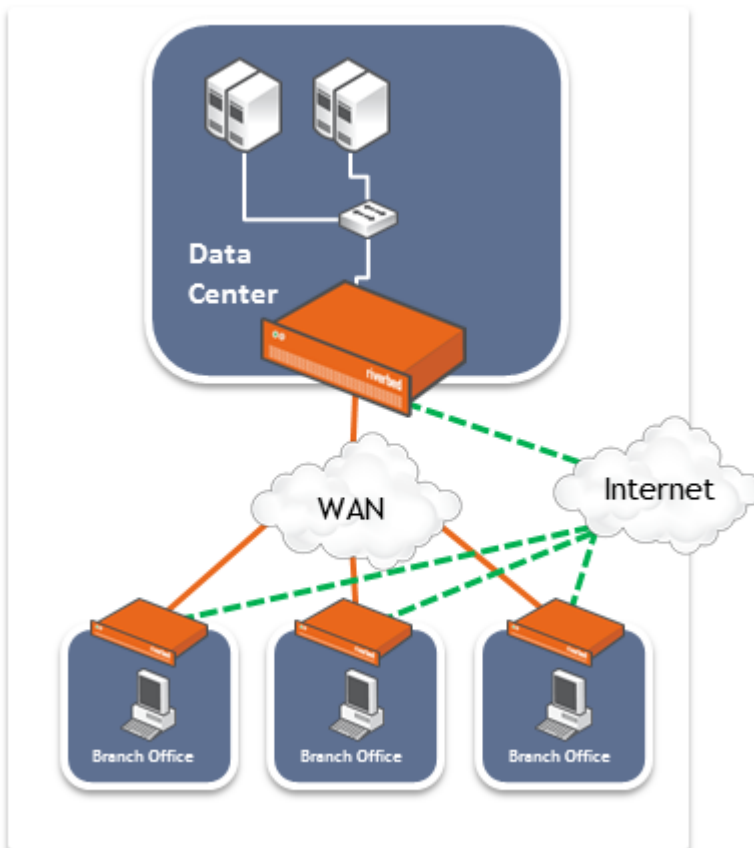
What can QoS do in this scenario?

- Prioritize traffic to/from branch offices where QoS is deployed. This allows branch offices to control applications that may not originate at the data center such as VoIP and locally received Internet traffic.
- Guarantee that enough bandwidth is available for specific applications, sites, or classes of traffic
- Allocate bandwidth fairly among all sites, classes, and applications, preventing any group of users or applications from monopolizing bandwidth
- Limit the amount of bandwidth used by specific applications or classes of traffic. This is useful both for restricting recreational traffic and for controlling premium applications like VoIP, where excess usage charges are possible
- Prioritize and limit the traffic sent to each remote site to avoid overwhelming the (typically) smaller remote site WAN link

What can't QoS do in this scenario?

- QoS can't control traffic between remote sites that doesn't traverse the data center or a branch office with QoS. A basic example would be branch-to-branch traffic such as VoIP. No QoS solution can accurately control this scenario without being deployed in at least one of the branch offices.
- QoS can't throttle inbound UDP traffic from the Internet. No QoS technology can do this without being able to control UDP from an upstream location.

Deployments with QoS at all sites



This scenario, where QoS is deployed at all sites within an organization, provides the ultimate level of control over applications in every case. Combining outbound and inbound controls ensures the most predictable level of application performance in all situations.

Key Use Cases:

- Protecting key business applications running at branch offices
- Reducing congestion on small or oversubscribed WAN links (in cooperation with Optimization)
- Controlling bandwidth when Internet access is being provided centrally through the data center to all of the remote sites
- Conforming with SLAs or bandwidth commitments for specific types of traffic (such as VoIP) from the data center.
- Making room for VoIP traffic on fully-meshed networks
- Protecting SaaS/Public Cloud applications from recreational traffic on local Internet links

What can QoS do in this scenario?

- Prioritize traffic to/from all branch offices. This allows branch offices to control applications that may not originate at the data center such as VoIP and locally received Internet traffic.
- Guarantee that enough bandwidth is available for specific applications, sites, or classes of traffic
- Allocate bandwidth fairly among all sites, classes, and applications, preventing any group of users or applications from

- monopolizing bandwidth
- Limit the amount of bandwidth used by specific applications or classes of traffic. This is useful both for restricting recreational traffic and for controlling premium applications like VoIP, where excess usage charges are possible
- Prioritize and limit the traffic sent to each remote site to avoid overwhelming the (typically) smaller remote site WAN link

What can't QoS do in this scenario?

- QoS can't throttle inbound UDP traffic from the Internet. No QoS technology can do this without being able to control UDP from an upstream location.


Best Practices for Policies

There are three main steps in creating a QoS configuration:

1. Tell the Steelhead appliance how much bandwidth is available
2. Determine how applications should be categorized and controlled
3. Define the applications that must be controlled

1: Tell the Steelhead appliance how much bandwidth is available

The first step in setting up QoS policy is to set the WAN bandwidth available through each WAN interface. This is a critical step, because QoS policies are set based on percentage of link capacity.



The screenshot shows a configuration window titled "WAN Link". It contains several options with checkboxes: "Enable QoS Classification and Enforcement" (checked), "WAN Bandwidth: 100000 kbps" (text input field), "Enable QoS on primary" (unchecked), "Enable QoS on wan0_0" (checked), and "Enable Local WAN Oversubscription" (unchecked). An "Apply" button is located at the bottom left of the window.

Figure 1: Configure WAN interface bandwidth

2: Determine how applications should be categorized and controlled

Classes are buckets that can hold traffic for one or more rules. A rule simply defines some type of traffic or application (see the next section for more on rules), so it's common for classes to be used as general buckets for traffic such as "Business Critical Applications" or "Recreational Traffic" or "Voice Protocols". If desired, classes can also be 1:1 mapped to applications such that a rule that matches a video application can also have its own class called "Video".

Classes are where link bandwidth is allocated and priorities are set for the applications they represent. The typical parameters are minimum and maximum bandwidth for the entire class (not per application) and one of 6 latency priorities: **real-time, interactive, high priority, normal, low priority, and best-effort**. These latency priorities specify how aggressively the Steelhead appliance sends packets for each class of traffic when there's congestion. A DSCP marking that applies to all traffic within this class is also commonly used.

QoS Classes:

<input type="checkbox"/>	Name
<input type="checkbox"/>	▼ Default-Site\$\$parent_class
<input type="checkbox"/>	Default-Site\$\$Best-Effort
<input type="checkbox"/>	<input checked="" type="checkbox"/> Default-Site\$\$Business-Critical
Shaping Parameters:	
Queue:	sfq
Minimum Bandwidth:	20 %
Maximum Bandwidth:	100 %
Latency Priority:	Business-Critical
Connection Limit:	
Marking Parameters:	
DSCP:	11
<input type="button" value="Apply"/>	
<input type="checkbox"/>	Default-Site\$\$Interactive
<input type="checkbox"/>	Default-Site\$\$Low-Priority
<input type="checkbox"/>	Default-Site\$\$Normal
<input type="checkbox"/>	Default-Site\$\$Realtime

Figure 2: Class configuration

In the above example, the Business-Critical class is a bucket that includes several apps (like Active Directory, Citrix, etc). This class gets 20% min, 100% max and schedules traffic at the Business-Critical latency threshold. The user could create a separate class for another application such as Salesforce and give it the Business-Critical latency threshold too, but that level of granularity should only be used for applications that are uniquely sensitive or that require unique controls.

3: Define the applications that must be controlled

A rule allows the user to identify a specific application or traffic type. Rules are based on a variety of Layer 3-7 parameters such as port, protocol, IP address/range, and/or application signature. These rules do not have policies associated with them other than specific DSCP markings, if desired. They are simply traffic “definitions”. These rules get assigned to classes as described above.

The screenshot shows a configuration window for a rule. At the top, there is a header bar with a checkbox, the number '33', a checked box next to 'SalesForce', and the text 'Default-Site\$\$Business-Critical'. Below this, the 'Name' field contains 'SalesForce' and the 'Description' field is empty. A section titled 'For Traffic with the Following Characteristics:' contains several fields: 'Local Subnet' and 'Remote Subnet' both set to '0.0.0.0/0', 'Port' set to 'all', 'Protocol' set to 'All', 'VLAN Tag ID' set to 'all', 'DSCP' set to 'All', 'Traffic Type' set to 'All', and 'Application' set to 'Salesforce'. Below this is a section titled 'Apply these QoS Settings:' with 'Service Class' set to 'Default-Site\$\$Business-Critical' and 'DSCP' set to 'Inherit from Service Class'. An 'Apply' button is located at the bottom left of the configuration area.

Figure 3: Configuring a rule

In the example above, the user is mapping **Salesforce** traffic, based on a Layer 7 application signature, to the Business-Critical class.

Pulling it all together

The 3 steps above are all that’s needed to get QoS up and running on the Steelhead appliance. Step 1 is very simple and dictated by the physical WAN link available at the site. Step 2 requires thought about how the total bandwidth should be carved up for each class of application. It’s generally better to start conservatively by protecting key applications like VoIP and Citrix by giving their classes a high enough guarantee to ensure they work properly. Lower priority classes of applications such as recreational traffic require no minimum guarantee at all. It’s also good to put an upper limit on traffic such as unauthorized applications that absolutely shouldn’t use much if any bandwidth.

Step 3 must be repeated for each application that it makes sense to identify and control. This step has the potential for overuse so it’s typically best to start with a few key applications that need to be protected or contained. This simple and cautious approach often results in noticeably improved performance for all applications. This is because most performance issues are caused by overly aggressive traffic that needs to be contained. When there are specific applications that are struggling, simply applying a policy to protect them restores their performance to predictable levels.

The Steelhead appliance offers a Basic QoS UI that uses common, pre-defined classes to speed the creation and implementation of QoS policies. A list of commonly-used applications is also provided and is pre-assigned to these classes by default. This default application list is easily edited and customized. Once the application list is set, the configuration can be applied to all bandwidth equally (i.e. to the default destination 0.0.0.0/0 or “any”) or they can be applied per site as defined by IP subnets.

If more flexibility or granularity is needed, then the Advanced UI is also available. The Advanced UI allows configuration of more granular classes and more specific queuing techniques. It also allows configuration of unique class/rule combinations that don’t align well with the template-based Basic QoS UI.

Validating QoS

Visibility is a key requirement for delivering a reliable QoS solution. It validates whether policies have been set at appropriate levels and provides guidance for updating or creating new policies. The Steelhead appliance offers basic reports that show how much traffic is sent and dropped for each QoS class. This helps to validate that traffic is being set to the right classes and shows how much traffic is being dropped if/when congestion occurs. For a broader view, Riverbed Cascade provides an advanced reporting capability that extends to applications, sites, and groups of users. Cascade provides a high-level view of which applications are using network resources overall and at which specific locations. It also shows how application usage and performance varies for different populations of users even within the same office. This level of visibility allows the configuration of appropriate Steelhead QoS policy settings. If further examination is required to validate a performance issue, Cascade can also drill down to individual user sessions and to individual packets within a transaction.

Conclusion

Ensuring that applications work fast and predictably is a key goal for many organizations. The old way of managing QoS on routers is inaccurate and difficult to manage. 3rd-party solutions fare only slightly better. Riverbed combines best of breed QoS, Optimization, and Network Performance Management to deliver the absolute best combination of speed and predictability for all applications.



Riverbed Technology, Inc.
199 Fremont Street
San Francisco, CA 94105
Tel: (415) 247-8800
www.riverbed.com

Riverbed Technology Ltd.
One Thames Valley
Wokingham Road, Level 2
Bracknell, RG42 1NG
United Kingdom
Tel: +44 1344 31 7100

Riverbed Technology Pte. Ltd.
391A Orchard Road #22-06/10
Ngee Ann City Tower A
Singapore 238873
Tel: +65 6508-7400

Riverbed Technology K.K.
Shiba-Koen Plaza Building 9F
3-6-9, Shiba, Minato-ku
Tokyo, Japan 105-0014
Tel: +81 3 5419 1990