No. 109

Technology
White Paper

e·QoS: Solutions for Service Providers using Riverstone Networks' Switch Routers

Abstract

Today's networks need to carry a variety of traffic with expectations of service that vary from application to application. This is commonly called a "differentiated service" (DIFFSERV) Quality of Service (QoS) Strategy. In the past, all data applications used a QoS strategy called "best effort." Using this QoS strategy, network operations were limited to providing Service Level Agreements (SLAs) based solely on general availability of and reachability through the infrastructure, and not on how well applications performed within the infrastructure.

Today's applications range from high-bandwidth file transfers needing networks that can carry large amounts of data, to interactive, low-bandwidth, low latency "chat" applications, and real-time applications like multimedia and IP telephony that require predictable delay characteristics from the network. IP telephony, known as voice over IP (VOIP), is a major component of the new "converged" data network infrastructure that promises a new set of applications and services.

Riverstone provides a complete solution enabling customers to build networks and to offer Quality of Service using the differentiated service model architecture. In particular, Riverstone Networks offers:

1. High performance switching/routing platforms
2. Wire-rate performance at Gigabit link speeds
3. A comprehensive QoS Framework
4. Policy-based network management to baseline your network and configure it

The following sections describe each of these components and show how Class of Service, Capacity Planning and Quality of Service function to provide network-wide, scalable differentiated voice, and data services that help service providers achieve:

- **Improved Profitability** – by attracting high-end business customers and offering higher priced levels of services while reducing overall cost by using bandwidth more efficiently
- **Increased Competitiveness** – by providing customers differentiated and value-added services such as multiple classes of better than "best-effort" service, and providing SLAs based on specific customer requirements

River STONE NETWORKS™

Riverstone Switch Routers

Riverstone's RS Switch Routers deliver Quality of Service by integrating wire-speed Layer 4 switching with policy-based traffic classification and prioritization and routing. A non-blocking switch fabric ensures that Riverstone's Switch Router will keep up with today's bandwidth-hungry, delay-sensitive applications. Because Riverstone's custom ASIC technology can read deeper into the packet, all the way to Layer 4, traffic can be identified, classified, prioritized, and routed at the application level. This enables the RS product family to provide differentiated service to real-time traffic such as IP telephony. Currently, there are five platforms that will enable service providers to offer value-added services to their customers:

- **IA 1100/1200**, a fixed-configuration Web-cache redirector and server-farm load balancer for e-commerce, application or content-hosting environments
- **RS 2000**, a fixed-configuration 10/100 edge router for multi-tenant-building access delivering per-user IP services
- **RS 2100**, an aggregation-edge Gigabit Ethernet switch router and load balancer for space-constrained co-location environments
- **RS 3000**, an access and aggregation edge router with high port density and rich features in a small form factor
- **RS 8000/8600** hot-swap chassis-based metro switch routers for aggregating Point of Presence traffic
- **RS 32000**, a high-density metro core router for aggregating Point of Presence traffic in metro-area network backbones

Prioritization policies can encompass the entire network, groups of users, or specific host-to-host application flows. Knowing what to prioritize requires detailed instrumentation from RMON2 probes or RMON-enabled switches like the Riverstone family of switch routers.

RMON2 measures bandwidth utilization for each Layer 4 flow, allowing real-time network baselining of the entire network infrastructure's bandwidth and traffic patterns. All RS platforms can maintain full RMON2 with a protocol directory of over 500 protocols in the IP family of protocols. RMON2 can be enabled on all ports including Fast Ethernet, Gigabit Ethernet, ATM and POS links, thus allowing RMON probes to be deployed in parts of the network where detailed probe capabilities are better suited to legacy protocols and lower-speed links.

A complementary protocol to RMON for billing is also available. Riverstone products offer pinpoint accounting using Lightweight Flow Accounting Protocol (LFAP, RFC 2124). This protocol provides reliable, TCP-based, resilient flow accounting measuring total time of conversations between end stations in addition to total bandwidth consumed. LFAP complements RMON by providing detailed flow data collection such as source and destination socket identifiers for UDP and TCP protocols. Like other subsystems, LFAP offers SNMP MIB for monitoring your accounting subsystem to let you know when revenue-generating accounting records are being lost.

Riverstone's Switch Router provides hardware-specific solutions for implementing QoS across given switch and software solutions to deliver QoS across an entire Autonomous System (AS).

River STONE NETWORKS™

Feature	Benefit
Wire-speed routing on every port	**Removes routing as the bottleneck and allows control over all network resources**
Non-blocking Multipoint Switching Fabric	**Prevents overloaded output wires from clogging the switching hardware, and isolates points of network congestion so that other traffic flows are unaffected**
Large buffering capacity	**Avoids packet loss during transient bursts that exceed output wire capacity. Packet fragmentation for 64k size packets on SONET.**
Traffic classification and prioritization	**Hardware capable of manipulating 802.1p at Layer 2 and DIFFSERV Classification/TOS byte mapping at Layer 3 enables wire-speed differentiated services.**
Layer 4 flow switching	**Provides application-level policy-based routing.**
Resource, COS, QoS SNMP MIBS plus Software Development Kits (SDK)	**Allows powerful QoS policy applications to be implemented and maintained quickly and easily. ISPs can define value-added services on top of a flexible networking switch allowing detailed policies to exist.**
RMON and RMON2 on every port	**Identify existing traffic patterns and baseline every part of your network.**
Flow Accounting	**Facilitates billing and accounting of network resources by source and destination IP Address, port, protocol, AS Source, AS Destination and AS Path.**

QoS technology can neither solve the problem of inadequate network bandwidth, nor can it speed up slow equipment. The first step toward developing a high-performance network is to unleash the potential of the wires: all networking hardware should run at wire speed.

The Riverstone Switch Routers perform at wire speed by using a non-blocking, dynamic multipoint switching fabric and custom ASICs that deliver wire-speed switching and routing of Layer 2, Layer 3, and Layer 4 flows. This means that even with traffic loads up to the full bandwidth of the wire, all data makes it through. And, since packet lookup and forwarding algorithms are built into hardware, latencies are measured in microseconds, not milliseconds. Furthermore, because the Riverstone Switch Router hardware manages Layer 4 flows, this performance is sustained even when features such as Access Control Lists (ACL), RMON, and QoS features are enabled.

When output wires are overloaded and buffers are nearly full, it is time to apply QoS rules so that existing traffic will not be interrupted by new flows entering the system. The Riverstone RS Switch Routers divide all traffic into four internal prioritized classes. Traffic is classified by combinations of Layer 2, Layer 3, and Layer 4 information from a given flow. This provides extremely flexible and powerful packet identification capabilities, which can be as broad as a VLAN or IP subnet, or as specific as a single host-to-host application flow.

River STONE NETWORKS™

At each queuing point in the system, the hardware uses this policy-based classification to make buffering and forwarding decisions. Separate buffer space is allocated to each of the four classes of traffic. Forwarding is done on a prioritized basis, ranking the four classes from highest to lowest priority. The highest-priority class is reserved for router control traffic, which leaves three classes — high, medium, and low — for normal data flows. Buffered traffic in higher-priority classes is sent ahead of pending traffic in lower-priority classes, allowing

latency and throughput demands to be maintained for the higher-priority traffic. To prevent low-priority traffic from waiting indefinitely as higher-priority traffic fills the wire, a "Weighted Fair Queuing" (WFQ) mechanism provides adjustable minimum-bandwidth guarantees, thereby ensuring that some traffic from each priority class always gets through. Weighted Random Early Detection (WRED) can also be applied to keep congestion under control when traffic is predominantly TCP based.

QoS Framework

To achieve Quality of Service goals, network elements and management software must provide the ability to guarantee bandwidth as well as delay characteristics (latency and jitter) per traffic class or flow. To meet QoS goals, the RS Switch Router blends speed traffic classification, queuing, and policing mechanisms into hardware (ASICs) that perform these functions at wire speed.

Congestion Management Tools

Congestion management tools help packets ride over bursts of traffic in the network without undue loss of data. When the router is receiving more traffic than it has the physical capacity to process, it needs to buffer the data until it can be processed (QUEUE). The bigger the buffer, the better the chances that no traffic gets lost under situations of congestion. Riverstone's products provide several tools to manage the buffers and to service packets waiting to be processed.

The Riverstone Switch Router congestion-management tools include traffic classification (DIFFSERV Classification, 802.1p), rate limiting Committed Access Rate (CAR), queuing policies and Weighted Random Early Detection (WRED).

Traffic Classification for Differentiating Service

In the Riverstone Switch Router, traffic classification is accomplished by mapping Layer 2, 3 or 4 traffic to one of four queues. Each traffic classification is treated as an individual traffic flow in the RS Switch Router. A Layer 2 flow is traffic classified based on 802.1p priority or MAC address or by port of ingress into a switch. A Layer 3 flow is classified based on source/destination IP address. A Layer 4 flow is classified using source/destination TCP/UDP port number in addition to Layer 3 source and destination IP address, TOS byte, protocol type, and incoming interface or port. Once traffic is classified and flows to a queue, the user can apply a rate limit and a queuing policy to the traffic.

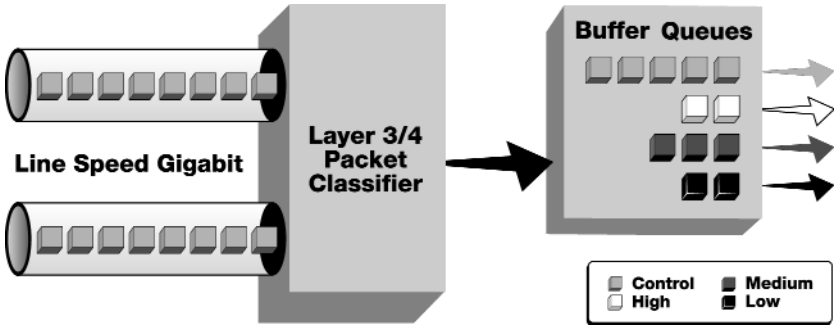

Figure 1: Rate limiting provides precise bandwidth policing.

Riverstone's RS Switch Routers implement the rate-limiting feature in hardware to ensure that other functions such as NAT, ACL, or WAN connections are not affected. All line cards are rate-enabled to help assure controlled, reliable access to your network in the face of denial-of-service situations.

All Riverstone platforms support three different types of Hardware Rate Limiting:

- **Port Rate Limiting** provides bi-directional rate limiting for SPs that want to restrict bandwidth on an inbound or outbound port, regardless of what is transmitted on the physical port.

- **Aggregate Rate Limiting** provides a method for controlling bandwidth consumption on a specific traffic or protocol pattern, based on traffic policy. Traffic policy can be created to control the aggregate traffic to or from a subnet, and the aggregate traffic for the specific application within the subnet. Each traffic policy can consist of multiple applications, allowing great flexibility in specifying the type of traffic to be controlled.

- **Per-Flow Rate Limiting** restricts the bandwidth on a per-flow basis. In order to understand Per-Flow Rate Limit, one must understand the concept of a flow, which is an entry that contains both the senders' and receivers' network addresses, the application port number and the protocol type. The Riverstone RS Switch Router can identify each flow based on the IP header.

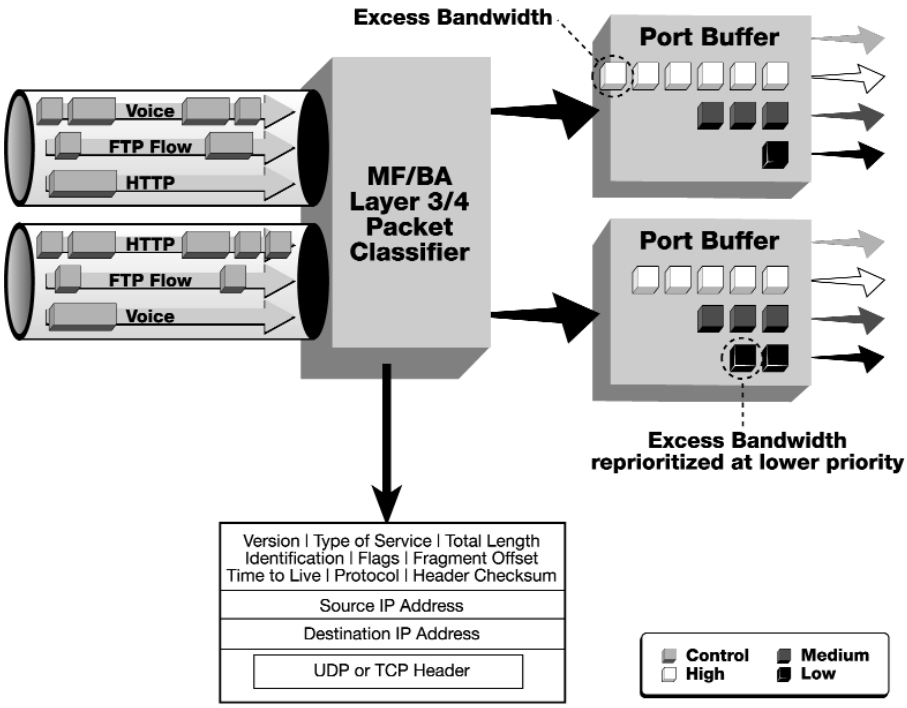

Figure 2: Flow Rate Limiting for the most granular bandwidth control.

Example of Per-Flow Rate Limiting

Rate limiting allows the user to set up a Committed Access Rate (CAR) on traffic that has been classified. Each IP flow can be rate limited at the input. Bandwidth is allocated on a per-flow basis. If the IP flow exceeds the bandwidth limit, then the arriving packet is either dropped or assigned a lower priority.

Queuing Policies for Congestion Management

Within the Riverstone Switch Router, the existing buffer space is divided into four queues according to the importance of the data belonging to a flow. The RS product platform supports low, medium, high, and control queues in order of increasing precedence. You can use one of the two following queuing policies on the RS Switch Router to service requests in the four priority queues:

1. Strict priority – assures the higher priorities of throughput but at the expense of lower priorities (starvation). For example, during heavy loads, low-priority traffic can be dropped to preserve throughput of the higher-priority traffic.

2. Weighted Fair Queuing (WFQ) – distributes priority throughput among the four priorities based on weights specified as percentages. This policy is best for normal Internet and enterprise traffic models.

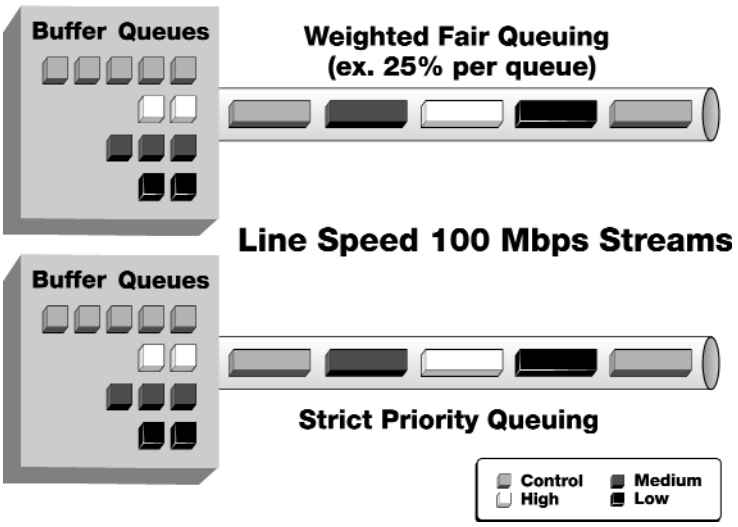

Figure 3: Queuing Policies – Strict Priority and Weighted Fair Queuing.

Congestion Avoidance to Maintain High Link Utilization

Weighted Random Early Detection (WRED) provides a statistical means of maintaining full link utilization when TCP-like protocols are deployed. When output buffers reach key watermarks, packets are randomly chosen to be dropped from a given set of flows. The algorithm will favor flows that account for most of the bandwidth utilization.

Network Management

Conducting business over geographically dispersed networks places a significant responsibility on your network resources. Data-exchange sessions carrying a variety of traffic– including voice, video and business data–depend on the efficiency, reliability, and predictability of the network infrastructure. The network as a system must be able to identify the applications that are using it and provide services tailored to the needs of the specific application in accordance with IT wishes. The network should be able to guarantee higher throughput for business-critical data and offer lower-priority service to noncritical data. For example, the network should be able to prioritize SAP/R3 traffic vs. Web traffic generated by network users checking show times at the local theater.

In order to realize such a system we need two things:

1. A mechanism to translate business rules to network configuration – Policy-based management

2. A normalized set of APIs to provide a means for implementing the configuration – Policy-based management is the framework for provisioning application-level business rules throughout the system. Here is the conceptual framework:

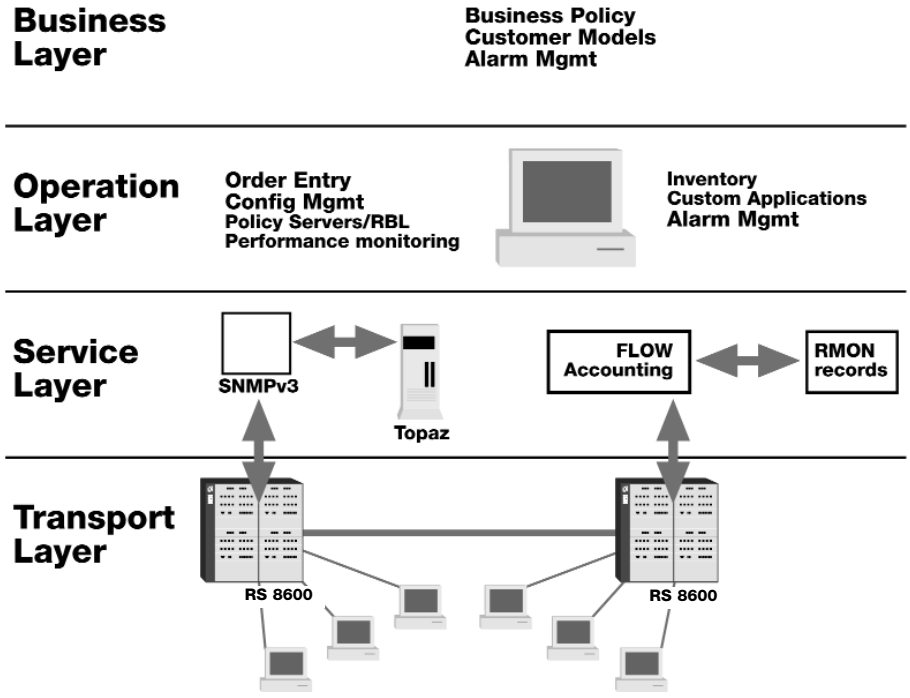

Figure 4: Network Management Infrastructure to allow for greater control and end-to-end policy enforcement.

A policy-based manager understands the network as a system, and can implement logical rules throughout the network without needing an administrator to interact directly with each element or network device within the system (SNMPCONF). For example, a policy-based management system should be able to enforce the rule "no Web surfing between 8 and

11 a.m." across the entire network. This can be done in many ways, including policy protocols or global access-list management. The key, though, is that the rule is created via an intuitive logical policy, and then enforced throughout the entire system via an automated service. Finally, there must be system-wide network and application management, but not to the exclusion of local management. Both element and network configuration models must work together. These vital services provide network administrators and business managers with insight into their network's operation, ideally both in real time and through trend analysis over a historical period.

Being able to view the network as a system and to monitor it proactively is critical to keeping the complexity and cost of managing it to a minimum. Obviously, complex and sophisticated network technology is needed to deliver an application-aware infrastructure. If this technology is implemented without excellent system-level management, its operational costs in terms of time and personnel will be excessive. If this technology is implemented within a framework of complementary network and system-level management, its complexity can be controlled and its operational costs can be dramatically reduced.

To summarize, since the network infrastructure transports business processes, and since business processes are built on applications, the network should operate based upon application awareness. It should deliver application data with an understanding of its business-critical nature, should provide logical controls to map business policy to a network usage policy, and should deliver these complex services with comprehensive, system-wide management that controls the complexity and costs of operating the network.

It is important to understand how different vendors deliver solutions. Riverstone's award-winning RS Switch Routers deliver application-based classification and prioritization without sacrificing performance. Riverstone's directory-based networking architecture enables an understanding of the physical, protocol, and application-level identities of devices connected to the network – allowing for greater control and end-to-end policy enforcement.

Network Conditioning and Accounting

Riverstone's policy-based solutions provide a complete feedback loop to condition ports and to account for network usage. The high-performance Riverstone Switch Router platforms can monitor network usage, generate accounting information and generate alerts to trigger policies on exceptional events. For a policy-based management system to be complete you need this feedback loop to continuously condition the network infrastructure.

The Riverstone Switch Routers with built-in RMON capabilities can gather statistics on all ports without compromising performance. Riverstone Switch Routers can also generate accounting information, also with no impact on data-switching performance.

Finally, the Ion SDK provides a basic foundation for providing normalized APIs for controlling a network of RSs: QoS, CoS, Resource and Access Control Lists. This API is described in a companion document: RS Ion SDK.

Riverstone Networks helps service providers become more competitive by providing value-added services, and achieve higher profit margins by utilizing bandwidth more effectively.

River STONE NETWORKS™

References

(IP) Internet Protocol. RFC 791, 1981.

(AS) Autonomous System. RFC 1930.1996.

(DIFFSERV) Blake, et al. An Architecture for Differentiated Services. RFC 2474.

(TCP) Transmission Control Protocol. RFC 793. 1981.

(WRED) Sally Floyd and Van Jacobson. "Weighted Random Early Detection Gateways for Congestion Avoidance". IEEE/ACM Transactions on Networking. Vol 1 No. 3. 1993.

Acronyms

ACL	*Access Control List*
ANSI	*American National Standards Institute*
ASIC	*Application-Specific Integrated Circuit*
ASP	*Application Service Provider*
ATM	*Asynchronous Transfer Mode*
CBR	*Constant Bit Rate*
DS1/DS3	*Digital Signal, Level 1 (1.54 Mbps) or 3 (44.7 Mbps)*
DSL	*Digital Subscriber Line*
E1/E2	*European Trunk 1/2 (2 Mbps/34.3 Mbps)*
ERP	*Enterprise Resource Planning*
HSSI	*High Speed Serial Interface*
ISP	*Internet Service Provider*
ITU	*International Telecommunications Union*
LAN	*Local Area Network*
LEC	*Local Exchange Carrier*
MAC	*Media Access Control*
MAN	*Metropolitan Area Network*
MDU	*Multiple Dwelling Unit*
MLPPP	*Multi Layer Point-to-Point Protocol*
MTU	*Multiple Tenant Unit*
OC-3/OC-12	*Optical Carrier 3/12 (155 Mbps/622 Mbps)*
POS	*Packet over SONET*
PPP	*Point-to-Point Protocol*
PVC	*Private Virtual Circuit*
QoS	*Quality of Service*
RED	*Random Early Discard*
SLA	*Service Level Agreement*
T1	*Trunk 1 (1.544 Mbps)*
TCP/IP	*Transport Control Protocol/Internet Protocol*
TDM	*Time Division Multiplexing*
UBR	*Undefined Bit Rate*
VBR	*Variable Bit Rate*
VLAN	*Virtual LAN*
VoD	*Video on Demand*
WAN	*Wide Area Network*
WDM	*Wave Division Multiplexing*
WRED	*Weighted Random Early Discard*

River STONE NETWORKS™

Riverstone Networks, Inc.
5200 Great America Parkway, Santa Clara, CA 95054 USA

408 / 878-6500 *or* ***www.riverstonenet.com***

Printed in the USA *v 1.5 9/00*