# 109
TECHNOLOGY
WHITE PAPER

e-QoS Solutions for Service Providers
using Riverstone Networks Routers

Gary Holland, Riverstone Networks

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. Voice over IP (VoIP) is a major component of the new "converged" data network infrastructure that promises a new set of applications and services. As we have seen in Korea, even games are creating demand for QoS as, all other things being equal, network performance can make the difference between the thrill of victory and the agony of defeat. Riverstone provides complete solution, enabling customers to build networks and to offer QoS using the differentiated service model architecture. In particular, Riverstone Networks offers:

1. High performance switching/routing platforms
2. Wirespeed performance with features enabled
3. A comprehensive QoS framework
4. Policy-based network management

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 customers who want the "platinum plan" and offering enhanced services with enhanced pricing 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 enabling providers to offer consistently-achievable SLAs based on specific customer requirements

River
STONE
NETWORKS™

RIVERSTONE ROUTERS

Riverstone's RS router family delivers QoS by integrating wire-speed Layer 2, 3, and 4 switching with policy-based traffic classification, prioritization, and routing. A non-blocking switch fabric ensures that Riverstone routers will not only keep up with today's bandwidth-hungry, delay-sensitive applications, but will also help future-proof your network equipment investment. 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 in hardware at the application level. This enables the RS product family to provide differentiated service with real-time traffic such as IP telephony, gaming, or running latency-sensitive applications over MPLS VPNs.

Currently, these platforms enable service providers to offer value-added services to their customers:

- **RS 1000/3000** – a metro access and metro edge router with high-port density ideal for delivering intelligent networking services to MTUs and campuses
- **RS 8000/8600** – hot-swappable, chassis-based metro routers for aggregating Point of Presence traffic while providing the flexibility and interface diversity expected in a metro platform optimized to bridge legacy and next-generation Ethernet networks
- **RS 16000** – a next-generation, high-density metro aggregation router optimized for Gigabit and 10-Gigabit Ethernet networks
- **RS 38000** – 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 devices such as the Riverstone RS router family.

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 IP based protocols. RMON2 can be enabled on all ports including Fast Ethernet, Gigabit Ethernet, ATM, and POS/SDH links, thus allowing RMON probes to extend to legacy parts of the network.

A complementary billing protocol for RMON 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 its other subsystems, Riverstone's LFAP supports SNMP, allowing easy access to your billing data.

Riverstone's RS router family provides hardware-based QoS, allowing providers to deliver QoS across an entire Autonomous System (AS). 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 – networking hardware should run at wire speed, even with features enabled.

The Riverstone RS family performs at wire speed by using a non-blocking, dynamic multipoint switch 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 (or fiber), packets flow at wire speed, performance you do not see in software-based solutions. And, since packet lookup and forwarding algorithms are built into hardware, latencies are measured in microseconds, not milliseconds. Furthermore, because Riverstone 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 family divides 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.

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. 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 WRED.

Traffic Classification for Differentiating Service

In the Riverstone RS router family, 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.

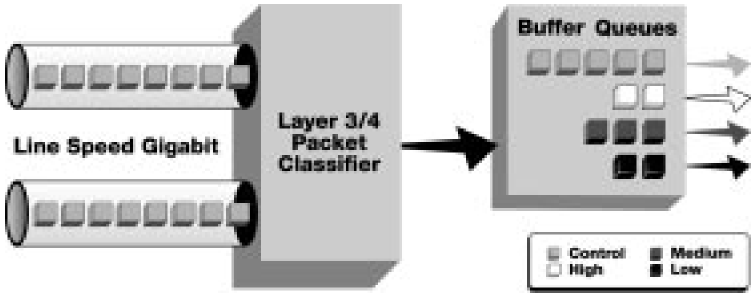

Figure 1. Layer 2-4 traffic remapped to buffer queues

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 service providers 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 use for specific traffic or protocols, 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. Thus a provider could limit P2P file sharing in a customer subnet without impacting Web browsing.
- **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.

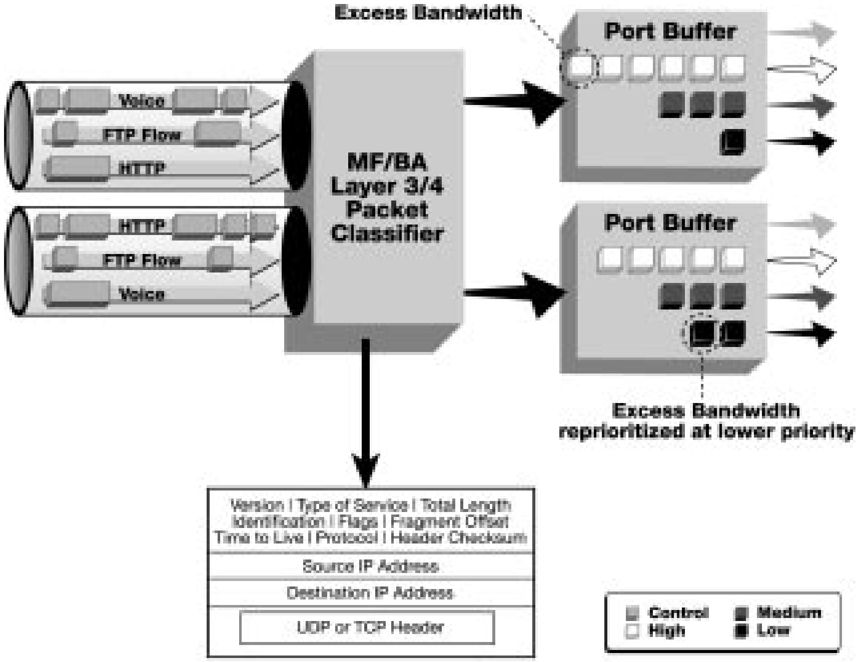
Excess Bandwidth
Port Buffer
Voice
FTP Flow
HTTP
HTTP
FTP Flow
Voice
MF/BA Layer 3/4 Packet Classifier
Port Buffer
Excess Bandwidth reprioritized at lower priority
Version | Type of Service | Total Length
Identification | Flags | Fragment Offset
Time to Live | Protocol | Header Checksum
Source IP Address
Destination IP Address
UDP or TCP Header
Control
High
Medium
Low

Figure 2. Per-flow rate limiting – traffic is shunted to buffers based on flow type

Example of Per-Flow Rate Limiting

Rate limiting allows the provider 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.

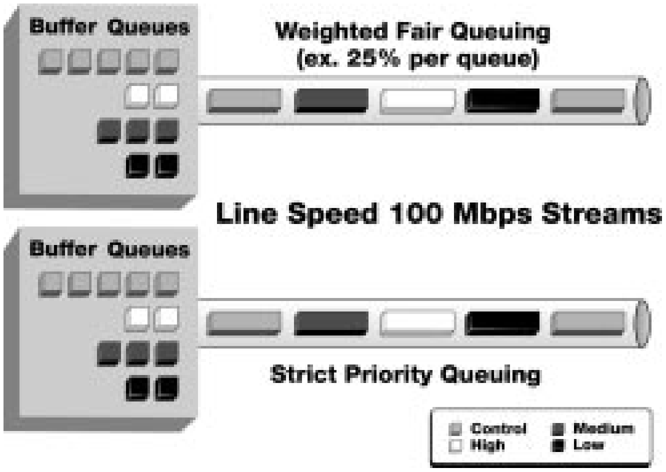

Figure 3. Weighted Fair Queuing vs Strict Priority Queueing

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 puts a significant load on 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 non-critical data. For example, the network should be able to prioritize Oracle traffic over Web traffic generated by network users downloading MP3s and movies.

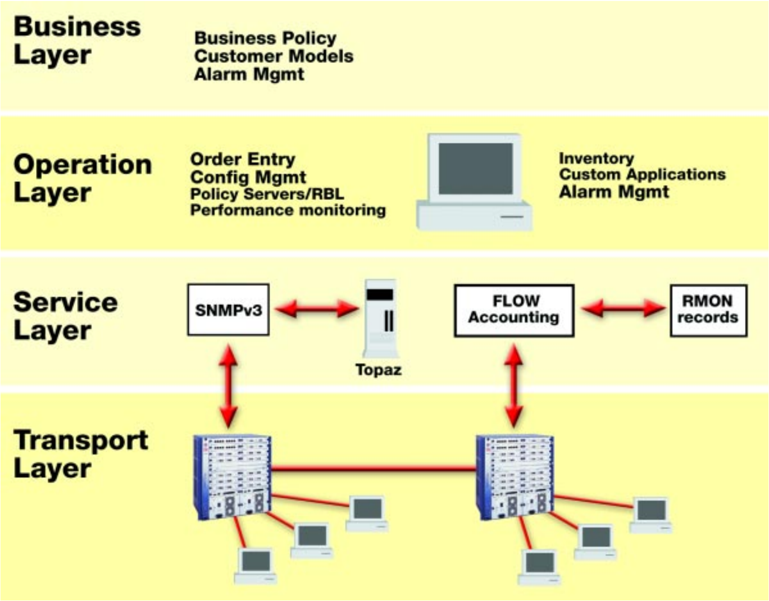

Figure 4. Policy-based network management

In order to implement 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:

A policy-based management system understands the network as a whole, 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 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 be application-aware. It should deliver application data with an understanding of its business-critical nature (or lack thereof in the case of MP3s, etc.), 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, allowing the operator to extract maximum profit not only from new Ethernet-based, next-generation networks but also from existing legacy infrastructure built on SONET/SDH, ATM, or Frame Relay.

It is important to understand how different vendors deliver solutions. Riverstone's award-winning RS 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 router platform 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.

CONCLUSION

The Riverstone RS family, with built-in RMON capabilities, can gather statistics on all ports without compromising performance. It can also generate accounting information, without impacting its wire-speed performance. In order to help unleash the capabilities of the Riverstone RS family, Granite, a C API/SDK, is available from http://www.nmops.org. EMS and other third party solutions are also available to simplify the provisioning and deployment of QoS-based services.

With features such as QoS, MPLS, TLS, and VPNs, Riverstone Networks helps service providers better leverage both new and existing infrastructure by enabling value-added services that translate directly into competitive advantage and enhanced revenue.

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

10GbE:	10-Gigabit Ethernet
AAA:	Authentication, Authorization, & Accounting
ABR:	Available Bit Rate
ACL:	Access Control List
ADM:	Add Drop Muxes
ADSL:	Asymmetric Digital Subscriber Line
ANSI:	American National Standards Institute
AP:	Access Point
API:	Application Program Interface
APS:	Automatic Protection Switching
ARP:	Address Resolution Protocol
ARPU:	Average Revenue per User
AS:	Autonomous System
ASIC:	Application-Specific Integrated Circuit
ASP:	Application Service Provider
ATM:	Asynchronous Transfer Mode
BGP:	Border Gateway Protocol
BOND:	Bandwidth on Demand
BPDU:	Bridge Protocol Data Units
BSS:	Business Support System
CAGR:	Compound Annual Growth Rate
CAPEX:	Capital Expenditure
CAR:	Committed Access Rate
CATV:	Cable Television
CBR:	Constant Bit Rate
CHAP:	Challenge Handshake Authentication Protocol
CIR:	Committed Information Rate
CLEC:	Competitive Local Exchange Carrier (ie: MCI, Sprint, etc.)
CLI:	Command Line Interface
CMTS:	Cable Modem Termination System
CO:	Central Office
CORBA:	Common Object Request Broker Architecture
CoS:	Class of Service
CPE:	Customer Premise Equipment
CR-LDP:	Constraint-Based LDP
CSMA/CD:	Carrier Sense Multiple Access/ Collision Detection
CSP:	Content Service Provider
CSPF:	Constraint-based Shortest Path First
DBS:	Direct Broadcast Satellite
DHCP:	Dynamic Host Configuration Protocol
DiffServe:	Differential Services IETF Standard
DLL:	Data Link Layer
DOCSIS:	Data Over Cable System Interface Specification
DS1/DS3:	Digital Signal, Level 1 (1.54 Mbps) or 3 (44.7 Mbps)
DSCP:	DiffServ Code Point
DSL:	Digital Subscriber Line
DSLAM:	DSL Access Multiplexer
DTV:	Digital TV
DXC:	Digital Cross Connects
E1/E2:	European Trunk 1/2 (2 Mbps/34.3 Mbps)
EAP:	Extensible Authentication Protocol
EAPoL:	Extensible Authentication Protocol over LAN
EBITDA:	Earnings Before Interest, Taxes, Depreciation, and Amortization
ECTA:	European Communities Trademark Association
EFM:	Ethernet in the First Mile
EJB:	Enterprise Java Bean
EMS:	Element Management System
EoATM:	Ethernet over ATM
ERP:	Enterprise Resource Planning
ESP:	Ethernet Service Provider
EtherLEC (or ELEC):	Ethernet Local Exchange Carriers
EXP:	Experimental bits
FCAPS:	Fault, Configuration, Accounting, Performance, and Security
FDDI:	Fiber Distributed Data Interface
FEC:	Forwarding Equivalence Class
FRAD:	Frame Relay access device
GARP:	Generic Attribute Register Protocol
GMPLS:	Generalized MPLS
GVRP:	GARP VLAN
HDSL:	High-data-rate DSL
HDTV:	High Definition TV
HFC:	High Frequency Cable
HPR:	Hitless Protocol Restart
HPS:	Hitless Protection System
HSSI:	High Speed Serial Interface
IANA:	Internet Assigned Numbers Authority
IBGP:	Internal Border Gateway Protocol
IDL:	Interface Definition Language
IDTV:	Interactive Digital TV
IEEE:	Institute of Electrical and Electronic Engineers
IETF:	Internet Engineering Task Force
IGMP:	Internet Group Management Protocol
IGP:	Interior Gateway Protocol
Ion SDK:	Ion Software Developer's Kit
IP:	Internet Protocol
IS-IS:	Intermediate Systems to Intermediate Systems
ISP:	Internet Service Provider
ITU:	International Telecommunications Union
IXC:	Inter-exchange Carrier
Kbps:	Kilobits per second (1 Kilobit = 1,024 binary digits)
LAN:	Local Area Network
LDP:	Label Distribution Protocol
LEC:	Local Exchange Carrier
LER:	Label Edge Router
LFAP:	Lightweight Flow Accounting Protocol
LSP:	Label Switched Path
LSR:	Label Switched Router
MAC:	Media Access Control
MAN:	Metropolitan Area Network
Mbps:	Megabits per second (1 Megabit = 1,048,768 binary digits)
MCDN:	Microcellular Data Network
MDI:	Media Dependent Interface
MDU:	Multiple Dwelling Unit
MIB:	Management Information Base
MLPPP:	Multi-Layer Point-to-Point Protocol
MMDS:	Multi-point Multi-channel Distribution System
MMF:	Multi-mode fiber
MPEG:	Moving Picture Experts Group
MPLS:	Multi-Protocol Label Switching
MPPP:	Multi-link PPP
MSN:	Management Network Server
MSP:	Metropolitan Service Provider
MSTP:	Multiple Spanning Tree Protocol
MTBF:	Mean Time Between Failure
MTU:	Multiple Tenant Unit
MVST:	Multiple-VLAN Spanning Tree
NAT:	Network Address Translation
NEBS:	Network Equipment Building Systems
NLP:	Network Layer Protocol
NMS:	Network Management System
OAM:	Operations, Administration, and Maintenance
OC-3/OC-12:	Optical Carrier 3/12 (155 Mbps/622 Mbps)
OPEX:	Operational Expenditure
OSI:	Open System Interconnection
OSPF:	Open Shortest Path First
OSS:	Operation and System Support or Operational Support Systems
PAE:	Port Access Entry
PAT:	Port Address Translation
PCS:	Physical Coding Sublayer
PDU:	Protocol Data Units
PE:	Provider Edge
PEF:	Packet over Ethernet over Fiber
PES:	Packet over Ethernet over SONET-based optical
PEW:	Packet over Ethernet over WDM-based optical
PHY:	Physical Layer
PIM:	Protocol-Independent Multicast; or Personal Information Manager
PMA:	Physical Media Attachment
PMD:	Physical Media Dependent
PON:	Passive Optical Networking
POP:	Point of Presence
POS:	Packet over SONET
PPP:	Point to Point Protocol
PPPoE:	Point to Point Protocol over Ethernet
PSTN:	Public Switch Telephone Network
PVC:	Private Virtual Circuit
PVST:	Per-VLAN Spanning Tree
QoS:	Quality of Service
RADIUS:	Remote Authentication Dial-in User Service
RAS:	Remote Access Services
RBOC:	Regional Bell Operating Company (ie: PacBell, etc.)
RED:	Random Early Discard
RF:	Radio Frequency
RFC:	Request for Comments
RMC:	RapidOS Management Center
RMON:	Remote Monitoring
RPR:	Resilient Packet Ring
RRST:	Rapid Ring Spanning Tree
RSTP:	Rapid Spanning Tree Protocol
RSVP:	Resource Reservation Protocol
RSVP-TE:	RSVP-Traffic Engineering
rt-VBR:	real-time Variable Bit Rate
SAN:	Storage Area Network
SAP:	Session Announcement Protocol
SDH:	Synchronous Data Hierarchy
SDSL:	Symmetric DSL
SFP:	Short Formfactor Pluggable
SLA:	Service Level Agreement
SMF:	Single-mode fiber
SNMP:	Simple Network Management Protocol
SPoF:	Single Points of Failure
SONET:	Synchronous Optical Network
SONET/SDH:	Synchronous Optical Network/ Synchronous Digital Hierarchy
SPoF:	Single Points of Failure
SRP:	Spatial Reuse Protocol
STP:	Spanning Tree Protocol
T1:	Trunk 1 (1.544 Mbps)
TCO:	Total Cost of Ownership
TCP/IP:	Transmission Control Protocol/Internet Protocol
TDM:	Time Division Multiplexing
TLS:	Transparent LAN Service
TNM:	Telecommunication Network Management
TOS:	Terms of Service, or Type of Service
TTL:	Time-to-Live
UBR:	Undefined Bit Rate
UDP:	User Datagram Protocol
UMTS:	Universal Mobile Telecommunication System
UTRAN:	Universal Terrestrial Radio Access Network
VAS:	Value-added Services
VBR:	Variable Bit Rate
VDSL:	Very-high-bit-rate Digital Subscriber Line
VLAN:	Virtual LAN
VoD:	Video on Demand
VoIP:	Voice over IP
VPLS:	Virtual Private LAN Services
VPN:	Virtual Private Network
VRC:	Virtual Router Cluster
VRRP:	Virtual Router Redundancy Protocol
WAN:	Wide Area Network
WCDMA:	Wideband Code Division Multiple Access
WDM:	Wave Division Multiplexing
WEP:	Wired Equivalent Privacy
WFQ:	Weighted Fair Queuing
WRED:	Weighted Random Early Discard
xDSL:	All types of Digital Subscriber Lines (ADSL, SDSL, HDSL, and VDSL)
XGMII:	10G Media Independent Interface

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

877 / 778-9595 or **408 / 878-6500** or **www.riverstonenet.com**

Printed in the USA

v 1.7 6/02