Packeteer Technical White Paper Series

Shaping Application Behavior

May 2002

Packeteer, Inc.
10495 N. De Anza Blvd.
Cupertino, CA 95014
408.873.4400
info@packeteer.com
www.packeteer.com

Table of Contents

Shaping Application Behavior

Managing bandwidth allocation for today's traffic diversity is a definite challenge. Network traffic and its associated applications do not share the same characteristics or requirements. We don’t have the same performance expectations for all traffic. A solution must balance organizational goals, users’ expectations, real-time network conditions, competing traffic, applications’ traits, and more.

Packeteer’s PacketShaper gives you the tools to proactively manage bandwidth allocation on congested WAN and Internet links. Using PacketShaper, you can define allocation rules to protect, pace, or block just about any subset of traffic. You can allocate precise bits-per-second rates, minimums and maximums, and/or scaled percentages of whatever bandwidth is available. These allocation rules can apply to each application, each session, each user, each group, or other targets. In short, you have tremendous flexibility and control. The next step is to determine how to use that flexibility and control to bring about your desired performance results.

This paper can help you organize and design a bandwidth-allocation plan. It leads you through the steps of analyzing the relevant characteristics of your traffic and determining appropriate control strategies.

Foremost Objective

First, consider whether you need to manage application performance or if you need to control traffic volume, independent of application or performance. Typically, if you are concerned with keeping customers or employees productive, then you are concerned about application performance. But if you supply bandwidth to users or organizations, and you are not involved with the applications that run over that bandwidth, then you are concerned about capacity and traffic volume.

Examples where Performance is Foremost	Examples where Load is Foremost
• An enterprise providing applications to staff • A service provider offering managed application services to subscribers • A business using B2B or B2C applications to conduct commerce	• A service provider that offers contracted amounts of bandwidth to businesses or individuals • A university that supplies each dormitory room with an equitable portion of bandwidth

If your primary concern is load rather than performance, then you can skip ahead to the section titled *Bandwidth-Allocation Rules* on page 7.

Traffic Characteristics

A good initial approach to managing performance is to manage two traffic categories proactively — traffic that needs to have its performance protected, and traffic that tends to swell to take an unwarranted amount of capacity.

For each type of traffic you want to manage, consider its behavior with respect to four characteristics: importance, time sensitivity, size, and jitter. Each characteristic below has an

associated explanation and question to ask yourself, as well as several examples of applications or protocols that fit the definition for a *YES* answer and that fit the definition for a *NO* answer. Make note of your answers for later.

Although characteristics for the same traffic can vary from one environment to another, you'll find some common traffic types' typical characteristics in an upcoming table.

Important

Sometimes the same application can be crucial to one organization's function and just irritating noise on another's network. Ask yourself: **Is the traffic critical to organizational success?**

Important	Not Important
• SAP to a manufacturing business • Quake to a provider of gaming services • PeopleSoft to a support organization • Email to a business	• Real Audio to a non-related business • Games in a business context • Instant messaging in a classroom

Time Sensitive

Some traffic, although important, is not particularly time sensitive. For example, for most organizations, print traffic is an important part of business. But employees and productivity will probably not be impacted if a print job takes another few seconds to make its way to the printer. In contrast, any critical application that leaves a user poised on top of the Enter key waiting for a response is definitely time sensitive.

Ask yourself: **Is the traffic interactive, time sensitive, or particularly latency sensitive?**

Time Sensitive	Not Time Sensitive
• Telnet • Citrix-based, interactive applications • Oracle	• Print • Email • File transfers

Sizable and Bursty

A traffic session that tends to swell to use increasing amounts of bandwidth and produce large surges of packets is said to be "bursty." TCP's slowstart algorithm creates or exacerbates traffic's tendency to burst. As TCP attempts to address the sudden demand of a bursting connection, congestion and retransmissions occur.

Applications such as FTP, multimedia components of HTTP traffic, and the graphics portion of HTTP traffic (*.gif, *.jpg) are considered bursty since they generate large amounts of download data.

Users' expectations for this traffic depend on the context. For example, if a large multimedia file is being downloaded for later use, the user may not require high speed as much as steady progress and the assurance that the download won't have to be restarted.

Ask yourself: **Does the traffic tend to burst? Are flows large and bandwidth hungry, expanding to consume all available bandwidth?**

Large and Bursty	Small and Not Bursty
• Music downloads • Email with large attachments • Web browsing	• Telnet • ICMP • TN3270

Prone to Jitter

An application that is *played* (video or audio) as it arrives at its network destination is said to *stream*. A streaming application needs a minimum bits-per-second rate to deliver smooth and satisfactory performance. Streaming media that arrives with stutter and static is not likely to gain many fans. On the other hand, too many fans can undermine performance for everyone, including users of other types of applications.

Ask yourself: **Does the traffic require smooth consistent delivery or it loses value, suffering stutter and static?**

Prone to Jitter	Not Prone to Jitter
• VoIP • WindowsMedia • Real Audio • Distance-learning applications	• Email • Print • MS SQL • TN3270

Common Traffic Behaviors

Now that you have tried characterizing your own traffic, here are some results for common types of traffic. The following table lists common traffic types, their characteristics, typical behavior, and desired behavior.

Traffic Types	Important	Time Sensitive	Sizable / Bursty	Prone to Jitter	Current / Undesired Behavior	Desired Behavior
FTP	√		√		Stuck progress indicators; peaks clog WAN access, slowing more time-sensitive applications	No stalled sessions; sustained download progress; paced bursts
Web graphics: *.gif, *.jpg, etc.	Varies		√		Frozen graphics, stop-and-go display	Smooth, predictable, viewable download
HTML	Varies	√			Unpredictable display and delay times; stuck behind more sizeable graphics	Prompt, consistent display
Telnet	√	√			Slow, inconsistent performance	Immediate transfer for prompt response times; small size won't impact others
Streaming media	Varies	√	Varies	√	Discontinuous sessions, jerky video, low-grade audio	Constant, predictable bit rate for smooth performance; limited number of users
DNS, LDAP	√				Slow infrastructure transactions that impose latency on other applications	Immediate transfer; small size won't impact others
email	√		√		Large email attachments sporadically clog WAN access, slowing down time-sensitive applications	Consistent, timely email transport
Music file-sharing (MP3s) in a business setting			√		Bursts and abundant downloads clog WAN access and slow down time-sensitive applications	Contained, steady downloads using a small portion (or none) of network resources
Oracle, SAP, JD Edwards, and other ERP or CRM applications	√	√			Slow and unpredictable response times	Swift, consistent performance
Prohibited, unsanctioned traffic			√	Varies	Present, active, or even dominant	Blocked

Control Features

PacketShaper provides a variety of control features to change traffic and network behaviors from those described in the Current Behaviors column in the previous table to those in the Desired Behaviors column.

Classification — Control's Prerequisite

Traffic classification is a crucial part of bandwidth management. You can't control a given type of traffic if you can't identify it. The ability to differentiate hundreds and even thousands of different types of traffic gives PacketShaper a strategic advantage in that it can apply its bandwidth-allocation rules to precisely the right traffic.

PacketShaper can differentiate traffic based on:

- Application
- Protocol
- Port number
- URL or wildcard
- Host name
- LDAP host lists
- Diffserv setting
- MPLS labels
- IP precedence bits
- IP or MAC address
- Subnet
- Travel direction (inbound/outbound)
- Source / destination
- Host speed range
- Mime type
- Web browser
- Oracle database
- Citrix published application
- Citrix ICA priority tagging
- VLAN varieties (ISL and 802.1p/q)

PacketShaper peers into packets and headers looking for characteristic markers of specific applications. As a result, it can distinguish multiple applications using the same port, follow an application as it port hops, spot traffic for specific databases, and recognize other traffic that proves illusory for routers and similar solutions. Each traffic category is called a *traffic class*.

Bandwidth-Allocation Rules

PacketShaper's *policies* and *partitions* are rules governing how PacketShaper allocates bandwidth to each traffic class. They enable you to control bandwidth on a flow-by-flow or aggregate basis. Policies determine how an application's individual flows are treated in the context of competing applications and other flows of the same application.

With policies, you can give each flow of mission-critical traffic the bandwidth it needs for optimum performance, as well as protect it from bandwidth-hungry, less time-sensitive traffic. Partitions manage allocation for the aggregate total of all flows, so that all of the flows for one traffic class are controlled together as one. With partitions, you can both protect and limit one type of traffic to a defined amount of bandwidth.

PacketShaper offers the following policy and partition types:

Type	Description
Priority Policy	Establishes a priority (0, the lowest, to 7, the highest) for traffic without specifying a particular bits-per-second rate. Use priority policies for non-IP traffic types, or traffic that does not burst. *Example* for Telnet: Priority policy at priority 6.
Rate Policy	Smoothes bursty traffic, such as HTTP, using TCP Rate Control[1]. Keeps greedy traffic sessions in line or protects latency-sensitive sessions. Delivers an optional guaranteed (minimum) rate for each individual session of traffic, allows sessions prioritized access to excess bandwidth, and sets an optional limit on the total bandwidth each session can use. *Example* for a critical streaming application requiring jitter-free performance: Rate policy 24 Kbps guaranteed, burstable at priority 6, maximum of 50 sessions
Discard Policy	Tosses all packets for a traffic class, thereby blocking the service. You might use this policy type for an unsanctioned application that you would prefer to not support on your network. *Example*: Discard policy for Internet Radio
Never-Admit Policy	Restricts non-TCP traffic and intelligently rejects web and TCP traffic. Use this policy to redirect certain web users to alternate URLs that could politely explain the restriction. *Example*: Never-admit policy for a popular MP3-download site that redirects users to a web page that explains the performance problems and limitations
Static Partition	Protects or limits all the traffic in one class. You specify the size of the reserved virtual link, choose if it can exceed that size, and optionally cap its growth. Partitions function like frame relay PVCs, but with the added important benefits that they cost less and they share unused bandwidth with other traffic. A partition that protects traffic generally has a minimum size, and a partition that restrains traffic generally has a cap. Partitions that do both have both. *Example* for Microsoft Exchange: Static partition, minimum size of 25 percent of the link, burstable, maximum size of 65 percent of the link.
Dynamic Partition	Creates equally sized, per-user or per-subnet subpartitions dynamically, as needed, when users initiate traffic of a given class. Use dynamic partitions in situations where per-user bandwidth equity is important — more important than the performance of certain applications. *Example* for all traffic to and from a university dormitory: Dynamic partition, per-user, 100 Kbps, maximum of 350 concurrent users, late arrivals refused.

More on Priorities

Use the following priority-guidelines table as a guideline for selecting an appropriate priority for a traffic class as part of a priority policy or rate policy. Of course, different applications are of varying urgencies in different environments, so tailor these suggestions to match your own requirements.

Priority	Description
7	Mission-critical, urgent, important, time-sensitive, interactive, transaction-based. Examples might include SAP, Oracle, a sales web site.
6	

[1] For more about TCP Rate Control, see "*TCP Rate Control and Alternatives*" on the Packeteer web site.

5	Important, needed, less time-sensitive. Examples might include collaboration and messaging systems, such as Microsoft Exchange.
4	
3	Standard service, default, not unusually important or unimportant. Examples might include web browsing and all traffic for those who foremost concern was *Load* at the beginning of this paper.
2	Needed, but low-urgency or large file size. Examples might include FTP downloads, email.
1	Marginal traffic with little or no business importance. Examples might include MP3 music downloads, Internet radio, games.
0	

More on Rate Policies

Creating a rate-based policy consists of determining:

- What the traffic's guaranteed connection rate should be
- How the traffic competes for unused (excess) bandwidth
- What to do when bandwidth runs out
- How to propagate class of service throughout the network

Guaranteed Rates

A guaranteed service rate (in bits per second) is the minimum rate guaranteed to each connection for a given traffic. It's the rate required for minimally acceptable performance. Guaranteed rate is especially useful for applications that are subject to jitter without a sufficient bandwidth stream. Even during heavy competing traffic, flows are guaranteed a smooth connection at a predictable rate.

PacketShaper lets you set guaranteed rates for low-speed and high-speed connections, enabling you to scale the minimum rate to the user's bandwidth expectations. These expectations are strongly determined by the user's access speed. PacketShaper detects the speed of an incoming flow and uses your low-speed and high-speed settings to scale a minimum rate for this connection. PacketShaper continues to monitor the connection speed through the life of the connection and can smoothly adjust bandwidth allocations as the effective connection speed changes.

Admissions Control

If the bandwidth for a traffic class is used up satisfying a guaranteed rate for many connections, you can determine how PacketShaper should treat successive connection requests. If the minimum guaranteed rate is not available, but there is still unsatisfied demand, admissions control is invoked. An admissions control policy instructs PacketShaper to:

- Refuse new connection requests until enough bandwidth is again available. For web traffic, a refusal results in an HTTP error message. For non-web traffic, connections are refused.
- "Squeeze" in the new connection to a fixed "trickle" rate.
- Redirect to a specified URL.

Excess Rate

Unused bandwidth that is not allocated for guaranteed service is known as excess rate. Policies determine which traffic can access how much excess rate. You can:

- Set excess rate parameters in addition to a guaranteed rate. When excess rate is available, a guaranteed connection can compete for the excess rate with all other traffic.
- Set excess rate parameters without setting a guaranteed rate. Most traffic fits this category, as a guaranteed rate is not needed and is set to 0.

There are three excess rate parameters:

- The excess rate priority determines what priority a traffic class has among others competing for excess rate.
- You can assign low-speed and high-speed excess rates to a traffic class. Like the connection speed values you use in guaranteed rates, PacketShaper uses these values and the connection's speed to scale allocation for this connection somewhere between these two values.
- You also can set excess rate limits — the maximum amount of excess rate that can be used by an individual connection at any given time. This protects your excess rate from being used up by high-speed connections. In addition, it can help enforce contractual bandwidth ceilings.

Network-Wide Class of Service

If you deploy one of the industry-standard mechanisms to ensure critical traffic is treated with a consistent class of service throughout your heterogeneous network, you can choose to change, set or clear any of the following markers:

- CoS/ToS bits
- Diffserv bits
- VLAN id
- VLAN priority
- MPLS label

Putting Control to Use

So far, we've examined how to identify and characterize traffic, its current and desired behaviors, and some of the control features at our disposal. Now let's put it all together and determine how you can actually use the control features to turn current behavior into desired behavior. Keep in mind that these tables contain suggestions and guidelines only. Your own context heavily influences the types of policies and partitions that are appropriate for your environment.

By Traffic Characteristics

Traffic Characteristics	Traffic Examples	Control Suggestions
Non-TCP traffic	IPX, SNA, AppleTalk	Priority policy with a priority selected from the priority-guidelines table that reflects importance
Small, time-sensitive, important flows	Telnet, DNS, LDAP, SNMP	Priority policy with a high priority
Large, bursty, important, non-interactive, non-time-sensitive flows	File transfers, email	Rate policy with 0 guaranteed, burstable, medium priority Partition to contain the aggregate of all users
Large, bandwidth-hungry, unimportant flows	Games (in a business setting)	Rate policy with 0 guaranteed, burstable, very low priority, and, if desired, a cap of the appropriate per-session Kbps amount that a particular game needs for acceptable performance Partition to contain the aggregate of all users to less than 5 percent of capacity (or whatever you are willing to devote)
Large, bandwidth-hungry, interactive, time-sensitive flows	Oracle, SAP	Rate policy with 0 guaranteed, burstable, high priority Partition to protect the aggregate for all users
Real-time streaming audio or video flows that need smooth reception and are sensitive to jitter	Streamworks, WindowsMedia	Rate policy with 21 Kbps guaranteed (or the minimum per-session amount needed for acceptable performance), burstable, medium-to-high priority, and a cap, perhaps 60 Kbps, that prevents sessions from getting bandwidth beyond that which improves reception Partition with a size that accommodates the maximum number of allowed users with minimum per-session bandwidth (500 Kbps for 20 users, for example) Admission Control to refuse more than the maximum number of users (or redirect them)
Unimportant, unsanctioned flows that you'd rather not block but want to vigorously contain	Music downloads, URLs of questionable content	Rate policy with 0 guaranteed, burstable, priority 0, and 2 Kbps cap (or the per-session amount you'd like to give) Partition to contain the aggregate of all users to less than 3 percent of capacity (or lower, if desired)
Prohibited, unsanctioned flows	Same as previous, but you want to block it	To block UDP traffic, set a discard policy. To redirect web traffic, set a never-admit policy using the **web-redirect** option with the alternate URL. To block web traffic without redirection, set a never-admit policy using the **web-refuse** option. Otherwise, set a never-admit policy with the **refuse** option.
Flows that are *not* destined for the managed link router	Traffic to and from an intranet server	Ignore policy
Flows that are part of a contracted amount of bandwidth	ISP customers' bandwidth, dormitory students, office tenants	Static partition for all users' bandwidth. Dynamic partition to allocate each user's or each subnet's bandwidth equitably.

By Traffic Type

Application or Protocol	Traffic Details	Control Suggestions
Email, FTP, Telnet, DNS, SNMP, LDAP, ERP and database applications		For applications and protocols that are included in the examples column of the previous table, consult that table's suggestions.
Web browsing, to a web server (get requests, typically one packet)	Inbound HTTP to an inside server, Outbound HTTP to an outside server	Priority policy with a priority selected from the priority-guidelines table that reflects importance
Web browsing, from a web server (responses, typically large and graphic laden)	Inbound HTTP from an outside server, Outbound HTTP from an inside server	Rate policy with 0 guaranteed, burstable, a middle priority, and a cap of the appropriate per-session amount that would keep high-capacity web users from usurping the bandwidth for other web users (100 Kbps, for example, if capacity is available) Partition to contain the aggregate of all web users to less than 40 percent of capacity (or whatever you are willing to devote)
VoIP (Voice over IP) Voice clients typically use UDP streams. H.323, the industry standard, starts a conversation on one port (H.323), jumps to another port (Q.931), and eventually splits up into a data flow (RTP) and control flow (RTCP).	Setup traffic (H.323 and Q.931), small	Priority policy at a medium-to-high priority, for example 5
	RTCP, small and intermittent	Priority policy at a medium-to-high priority, for example 5
	RTP and other VoIP data protocols, large and data laden	Rate policy with a guaranteed rate (see next comment), burstable at priority at a high priority. For determining the guaranteed rate, use the *Monitor Traffic* window to look at several RTP flows. Observe usage characteristics (current, one-minute average, and peak rate). Typically, if a manufacturer claims that its flow requires 8 Kbps, it will actually need 17 to 21 Kbps due to additional overhead and forward error correction. In addition, it is best to overstate the guarantee of UDP policies by 15 to 20 percent. For example: Rate policy with 24 Kbps guaranteed, burstable at 7. ***Note***: Many links do not deliver full-rated bandwidth. For instance, if you have a 128 Kbps link and are running sustained streaming traffic (such as voice), you will probably need to set the link speed 10 percent lower than actual capacity. This lower rate reflects the overhead for framing and routing updates running on the same line. This is critical for slower (sub-T1) links.
Print	Citrix ICA with tag = 3, LPR, TN5250p, TN3287, NetBIOS IP	Rate policy with 0 guaranteed, burstable, priority 2
Citrix application	Citrix-based ICA applications, classified individually or together	Rate policy: 5 to 20 Kbps guaranteed, burstable, priority 4 or 5. If you have a slow link or a larger number of concurrent Citrix users, use less guaranteed rate. Partition on the Citrix parent class of 25 to 50 percent of the link size, burstable, no limit

Summary

An expectation gap exists today between the current behavior of traffic and what's both desired and expected. The gap widens as industry develops more applications to take advantage of file-sharing, peer-to-peer architectures, web-based user interfaces, and distributed servers. PacketShaper can eliminate the gap and put control of your network resources and performance back into your own hands.

Keep in mind that the control activities described in this paper are part of a broader bandwidth-management process, described in additional resources on Packeteer's web site. The broad perspective is described in "Four Steps to Application Performance" and detailed instructions for all types of PacketShaper solutions are in PacketGuide at support.packeteer.com/documentation. For more information, consult Packeteer's web site at www.packeteer.com or call 408-873-4400 or 800-697-2253.