

# Cisco Catalyst **3750** Series Switch Delivers Rate Limiting for Universities and Enterprise Businesses

## Overview

Rate limiting is an important tool for IT managers in universities and enterprises alike. Limiting bandwidth to specific users and ports helps control network congestion, ensure high performance, create efficient networks, and prevent a small number of users from monopolizing network bandwidth. Rate limiting becomes even more critical as more Gigabit Ethernet ports are installed throughout the campus as part of a Gigabit To The Desktop (GTTD) rollout. IT managers will need to manage the bandwidth consumed by these Gigabit Ethernet ports until the distribution and core segments of the network are upgraded to accommodate the added bandwidth demands. All switches in the Cisco® Catalyst® 3750 Series family support an extensive Rate Limiting feature set that can be applied on both Gigabit Ethernet and Fast Ethernet interfaces.

University students are avid network users. Some run Web and FTP servers from their dormitory rooms. Some download large files such as movie and music files from the Internet. Others create videos with their PC video cameras. The traffic multiplies quickly as these large media files are shared with other students, families, and friends within and outside the university campus. Without rate limiting, even a small number of students downloading and sending large files can slow the network to a crawl. The issue is magnified when the network

slowdowns occur during business hours and the faculty members are unable to get critical research data downloaded.

Bandwidth management is important not only in the university environment, but in the enterprise environment as well. Employees downloading large files from the Internet can monopolize a small WAN link to the Internet or to the corporate intranet. Departments with higher-speed switches and PC network interface cards (NICs) can dominate the corporate network backbone and WAN connection at the expense of other departments.

Network congestion resulting from network under-provisioning can cause tremendous degradation to the quality of real-time applications such as IP telephony and video. Unlike data traffic, these real-time applications are especially intolerant of delay and jitter caused by network congestion. Without rate limiting, users pulling large files from a network server can seriously affect the quality of their neighbors' telephone conversation or videoconference. Low network availability and performance can cause a deluge of IT cases, lower employee productivity, and cause reduced network user satisfaction.

## Rate Limiting

With rate limiting, IT managers can control the rate of incoming and outgoing traffic to ensure that no user or application exceeds the maximum transmission rate allotted or monopolizes network bandwidth. This



control applies to Fast Ethernet and more importantly, the higher bandwidth Gigabit Ethernet ports. IT managers can set policies to allocate higher or lower bandwidth to certain users, groups of users, or applications. For example, an IT manager may decide to allocate faculty members connected on Gigabit Ethernet ports only 200 Mbps of bandwidth. Rate limiting, in conjunction with other features such as Time Based Access Control Lists (ACLs), can help IT managers to save money by delaying purchases of additional WAN links by carefully managing how much and when bandwidth is available. Rate limiting is a powerful tool, helping IT managers take control of the network and prevent network problems before they occur.

Important elements of rate limiting of Cisco Catalyst 3750 Series switch are:

- Buffer memory
- Queue threshold
- Queue scheduling
- Traffic shaping
- Traffic policing

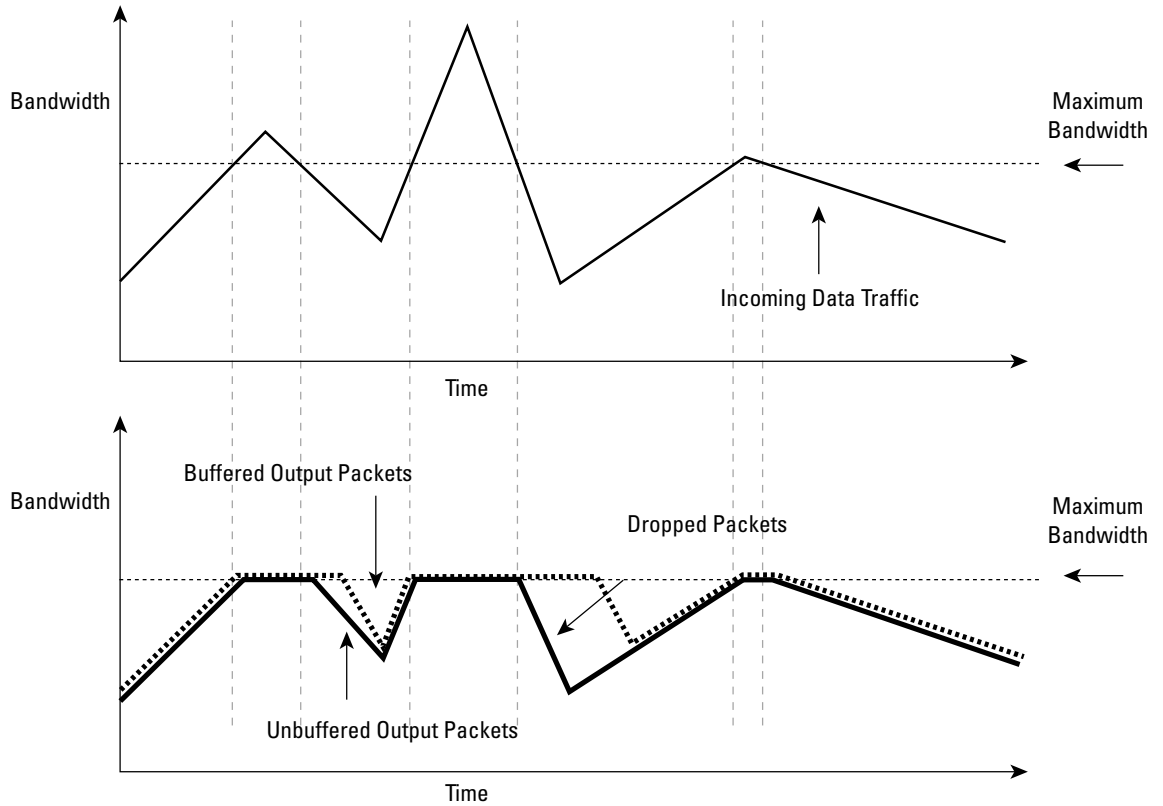
All of these elements work in concert to provide the optimum rate limiting solution to universities and enterprise businesses.

### **Memory Buffer**

A large memory buffer is important to minimize dropped packets in heavily congested networks (Figure 1). By queuing traffic in a buffer, small bursty flows are less likely to cause retransmissions. Without a buffer, bursty traffic can cause retransmissions, which can lead to increased network congestion.



**Figure 1**  
Memory Buffer



In Figure 1, the top diagram shows incoming data traffic over time. Some of the incoming data traffic exceeds the maximum bandwidth allowed. In the bottom diagram, the heavy solid line shows unbuffered output packets. Without rate limiting, packets that exceed the maximum available bandwidth are dropped. The dotted line shows buffered output packets. With rate limiting, packets that exceed the maximum bandwidth available are buffered and transmitted later—no packets are lost, and the buffered output traffic is much smoother. Customers can selectively configure larger buffer memories for queues and ports that have more bursty traffic or are more sensitive to jitter and latency. Typically, queues going to servers or queues dedicated for streaming media require higher buffer memories.

### Queue Thresholds

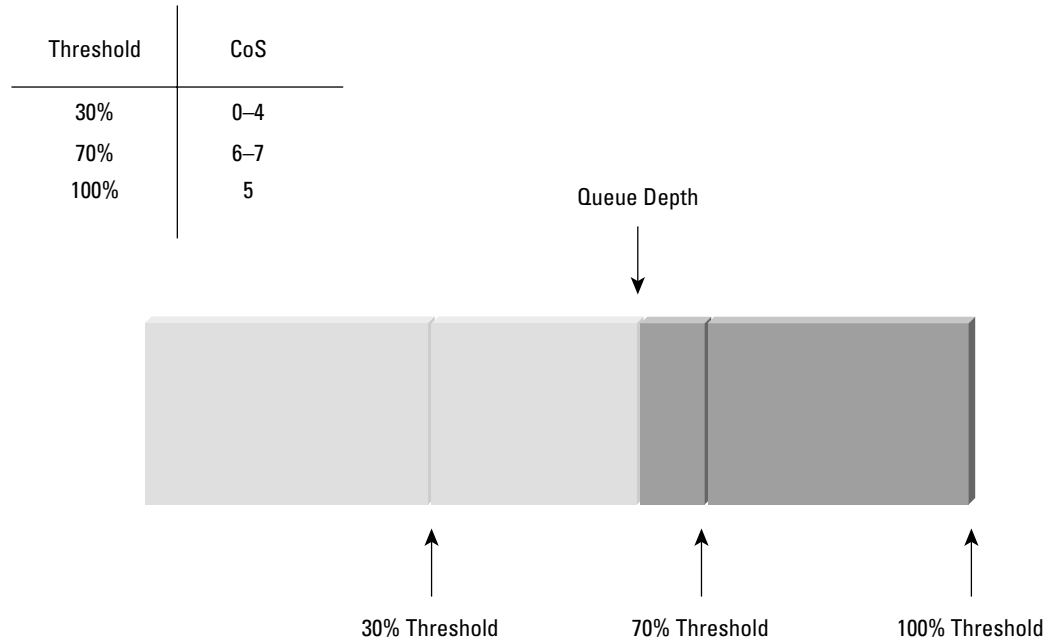
The Cisco Catalyst 3750 Series Switch supports Weighted Tail Drop (WTD) with three thresholds. Two of the thresholds can be configured by the user; the third is set to 100 percent. Packets that exceed the specified thresholds are dropped.

Packets are queued based on the quality of service (QoS) label. For example, in Figure 2 below, packets with class of service (CoS) values of 0 to 4 are assigned to a 30-percent threshold, packets with a CoS value of 5 are assigned to a 100-percent threshold, and packets with CoS values of 6 or 7 are assigned to a 70-percent threshold. When packets fill 30 percent of the queue, packets with CoS values of 0 to 4 are dropped. When packets fill 70 percent of the queue,



packets with CoS values of 6 or 7 are dropped. Packets with a CoS value of 5 are not dropped unless the queue is full. Typically, voice packets are assigned a CoS value of 5. Control traffic and video traffic are assigned CoS values of 6 or 7, and low-priority data traffic is assigned CoS values 0 to 4.

**Figure 2**  
Queue Thresholds



Packets with CoS 0–4 are Dropped as the Queue Depth Exceeds 30 Percent  
Packets with CoS 6,7 are Dropped as the Queue Depth Exceeds 70 Percent  
Packets with CoS 5 are Dropped Only if the Queue is Full

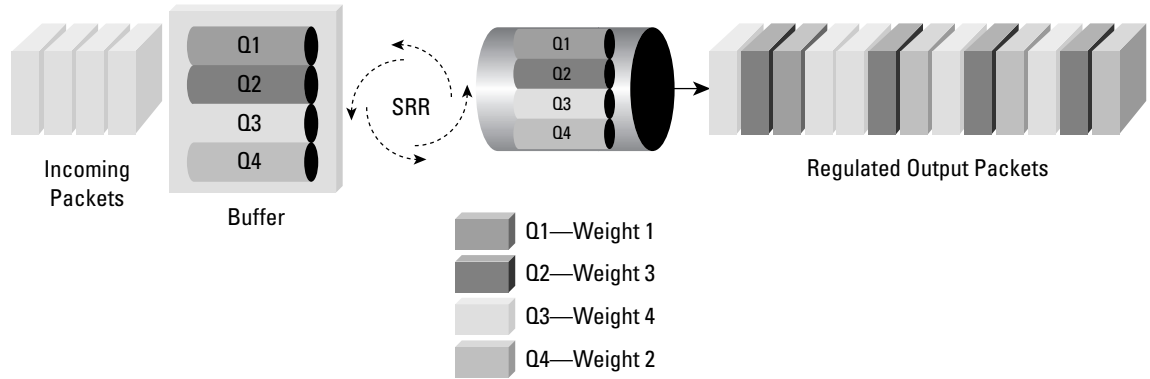
### Shaped Round Robin Queue Scheduling

Shaped Round Robin (SRR) queue scheduling allows for smoother traffic flow, as the queues are served in round-robin fashion in accordance to the specified queue weights. No single queue monopolizes bandwidth, and more packets are transmitted in queues with higher queue weights.

Queue scheduling on the Cisco Catalyst 3750 Series is performed with the SRR algorithm (Figure 3). Each queue can be assigned a weight (the four egress queues have equal weights by default).



**Figure 3**  
SRR Queue Scheduling



In the example in Figure 3, Q1 is assigned queue weight of 1, Q2 is assigned a queue weight of 3, Q3 is assigned a queue weight of 4, and Q4 is assigned a queue weight of 2. Queues are served in round-robin fashion: Q1, Q2, Q3, and Q4. In the second round, only Q2, Q3, and Q4 are served because Q1 only has a weight of 1. In the third round, only Q2 and Q3 are served because Q1 has a weight of 1 and Q4 has a weight of 2. In the fourth round, only Q3 is served because it has the highest weight. The process repeats itself with Q1, Q2, Q3, and Q4 being served in the next round.

### Traffic Shaping

Traffic shaping reduces network congestion and smoothes out traffic flows for more efficient bandwidth utilization. Traffic shaping places packets in a queue with a shaper at the head of the queue to smooth and regulate the rate and volume of traffic admitted into the network. For any traffic that exceeds the specified transmission rate, the transmit queue will not transmit any queued traffic data until the data rate conforms to the specified rate. Traffic shaping eliminates traffic bursts and presents a steady stream of traffic to the network.

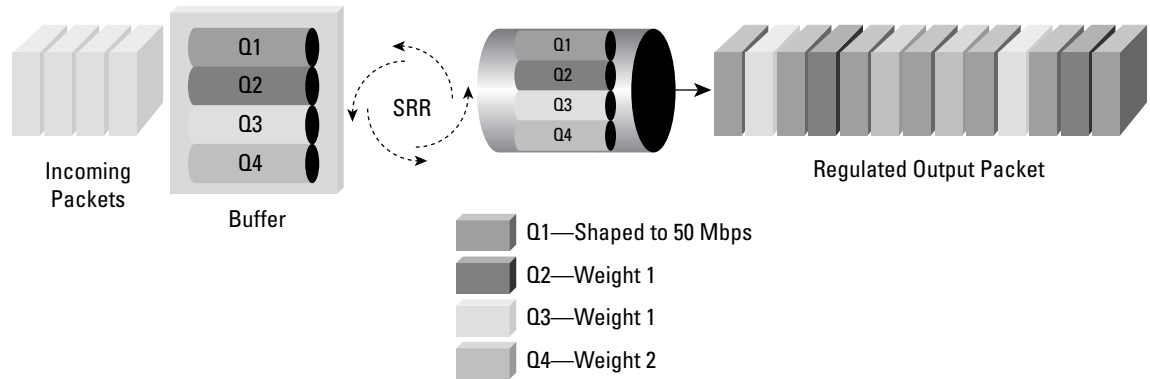
The Cisco Catalyst 3750 Series supports egress traffic shaping with SRR queue scheduling in hardware on all four egress queues. Queuing and buffering are important on the egress ports, primarily for speed mismatch and many-to-one congestion. If a gigabit server sends a burst to a 10/100 port, some of the burst needs to be buffered.

The Cisco Catalyst 3750 Series supports both aggregate and individual rate shaping. With aggregate rate shaping, customers can specify the maximum combined transmission rate of all four queues on a port. For example, if port 5 can transmit no more than 80 Mbps, then the combined bandwidth of Q1, Q2, Q3, and Q4 on port 5 cannot exceed 80 Mbps. For individual rate shaping, customers can specify the bandwidth allocated to a queue. For example, Q1 on port 4 is allotted 20 Mbps and Q2 on port 7 is allotted 5 Mbps.

Customers can configure up to four egress queues to be shaped. Queues that are not shaped are shared and scheduled by SRR. For example, in Figure 4 below, Q1 is shaped to be 50 Mbps. This means that 50 percent of the output packets are from Q1. Q2, Q3, and Q4 then share the remaining 50 Mbps in round-robin fashion according to their queue weights.



**Figure 4**  
Traffic Shaping



### Traffic Policing

Traffic policing is similar to traffic shaping. Traffic policing controls the rate and volume of traffic that enters the network and is processed by the switch. However, in traffic policing, there is no buffer—packet transmissions that exceed the maximum transmission rate are dropped.

The Cisco Catalyst 3750 Series supports ingress policing with SRR in hardware, ensuring stable and predictable forwarding performance. In a single switch, traffic shaping and buffers are not needed on the ingress because the Cisco Catalyst 3750 Series switching fabric is nonblocking. Packets from a user or server port have access to the high-speed fabric immediately.

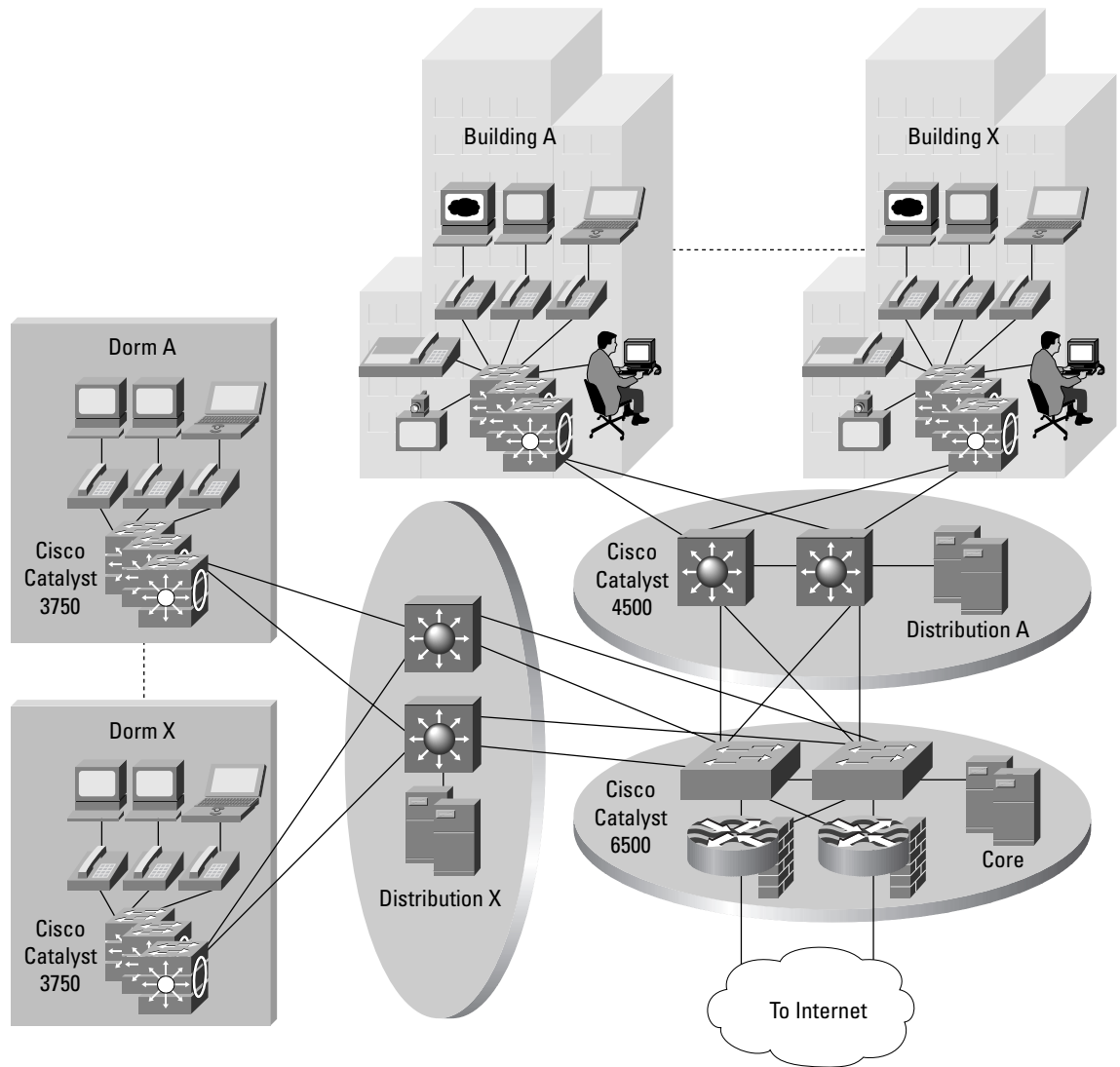
There are two ingress queues on each Cisco Catalyst 3750 Series Switch. One of the ingress queues is dedicated to Strict Priority Queuing. With Strict Priority Queuing, all traffic in this queue is transmitted first. All the high-priority traffic (voice traffic, for example) should be in this queue. All other traffic goes to the other queue.

### Rate Limiting for Universities

Rate limiting is important in the university environment. With traffic shaping and policing, IT managers can prevent a few heavy users from monopolizing network bandwidth. In addition, with pilot rollouts of GTTD occurring at most universities today, rate limiting provides an important tool for IT managers to manage the bandwidth consumed by these Gigabit Ethernet ports until the distribution and core areas of the network are upgraded to support higher bandwidth technology such as 10 Gigabit Ethernet. Faculty members and students can only send and receive up to the bandwidth individually allocated to them. IT managers can choose to give faculty and research students more bandwidth than undergraduate students, and can allocate greater bandwidth to application servers and workstations.



**Figure 5**  
University Networks



### Dormitories

In the dormitories, IT managers can rate limit the bandwidth each student receives so that no student monopolizes the network. A student running a music sharing server from a dormitory cannot only slow the dormitory network to a crawl, but can also affect the entire campus network. Thousands of students trying to upload and download files to and from a dormitory server can significantly affect the upstream traffic to the distribution as well as the core network. This can ultimately consume a large number of WAN links, typically just a few T1/E1 or T3/E3 lines.



To prevent this scenario from occurring, IT managers can rate limit the bandwidth each student receives to that of a typical student. Rate limiting can be asymmetrical. Normal student usage consists of more downstream traffic than upstream traffic, so the IT manager might provision more downstream traffic than upstream traffic to each student. IT managers can keep track of users and computers by Media Access Control (MAC) addresses, and can rate limit the amount of bandwidth the user or computer receives.

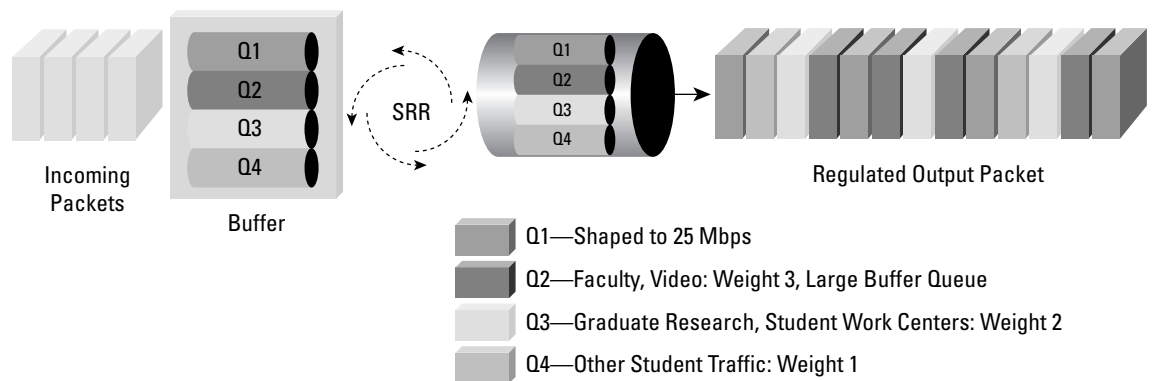
Video is one of the most bandwidth-intensive and jitter-sensitive applications (students download classroom lectures to view the class from their dorm rooms, for example). As IT managers rate limit the downstream and upstream bandwidth, they need to provision enough downstream traffic for a student to watch a video (typically between 300 Kbps to 2 Mbps, depending on the encryption used). In this scenario, the IT manager would allocate more buffers to the video queue and assign the video traffic to the high-priority queues to minimize jitter.

### Campus Buildings

In campus buildings, IT managers typically want to rate limit the faculty traffic to a higher rate and assign faculty traffic to a higher priority queue than student traffic. Faculty traffic on Gigabit Ethernet ports may be rate limited to 200 Mbps, and assigned to the second-highest priority queue. In student classrooms (where lecturing occurs), students connected to the same Gigabit Ethernet ports may be rate limited to just 500 Kbps. However, in graduate student research centers and in student workstation clusters, students may be rate limited to a higher rate, such as 4 Mbps. Engineering graduate students often send large design files to servers to be processed, graphics students need to transfer large video files, and undergraduate students in workstation clusters are working on their assignments and projects.

In both dormitories and campus buildings, IP phones should be rate shaped and placed on Strict Priority Queuing. Strict Priority Queuing ensures that voice traffic is given the highest priority. Strict Priority Queuing provides the dedicated bandwidth that voice traffic needs to minimize latency and jitter. While voice traffic is very latency- and jitter-sensitive, it does not require much bandwidth (less than 100 Kbps). In the example shown in Figure 6, voice traffic is shaped to 25 Mbps. The shaped bandwidth is a function of the volume of voice traffic anticipated.

**Figure 6**  
Rate Limiting in University Networks





IT managers can apply different drop thresholds to a queue. For example, for Q2 in Figure 6, IT managers can assign drop thresholds of 100 percent for video traffic, and drop thresholds of 40 percent for faculty HTTP traffic. Video traffic is a real-time application that is highly sensitive to dropped packets. On the other hand, HTTP traffic is typically composed of short bursts of data traffic that is less sensitive to dropped packets. Waiting one to two seconds longer for HTTP packet retransmission does not affect the user experience as much as waiting one to two seconds for video packet retransmissions.

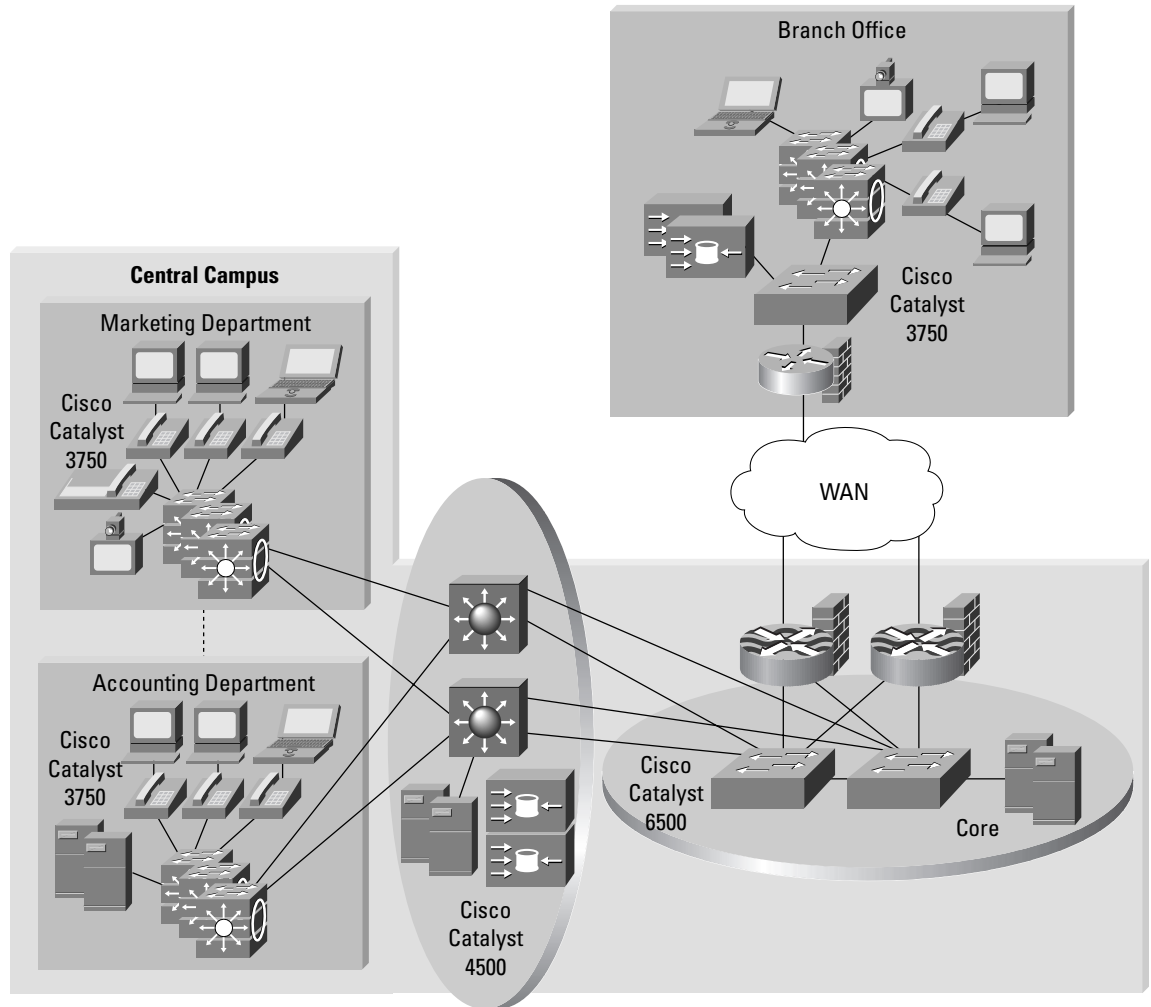
IT managers can preferentially give higher bandwidth to select buildings or distribution centers, depending on bandwidth requirements, or departments that pay more money can be given a higher rate limit. In general, IT managers allocate greater bandwidth from the campus building distribution network to the core network than from the dormitory distribution network to the core. Faculty offices, graduate student research centers, and student work centers are housed in campus buildings. These work-intensive centers typically require higher network bandwidth, and have a greater need for the WAN link to the Internet. Rate limiting allows for more efficient use of network bandwidth.

### **Rate Limiting for Enterprises**

As in the university environment, rate limiting can be used to intelligently manage bandwidth allocation in the enterprise. It can prevent one department or employee from dominating the available network bandwidth—and it allows IT managers to allocate greater bandwidth to the departments and applications that need it.



**Figure 7**  
Enterprise Networks



### Central Campus

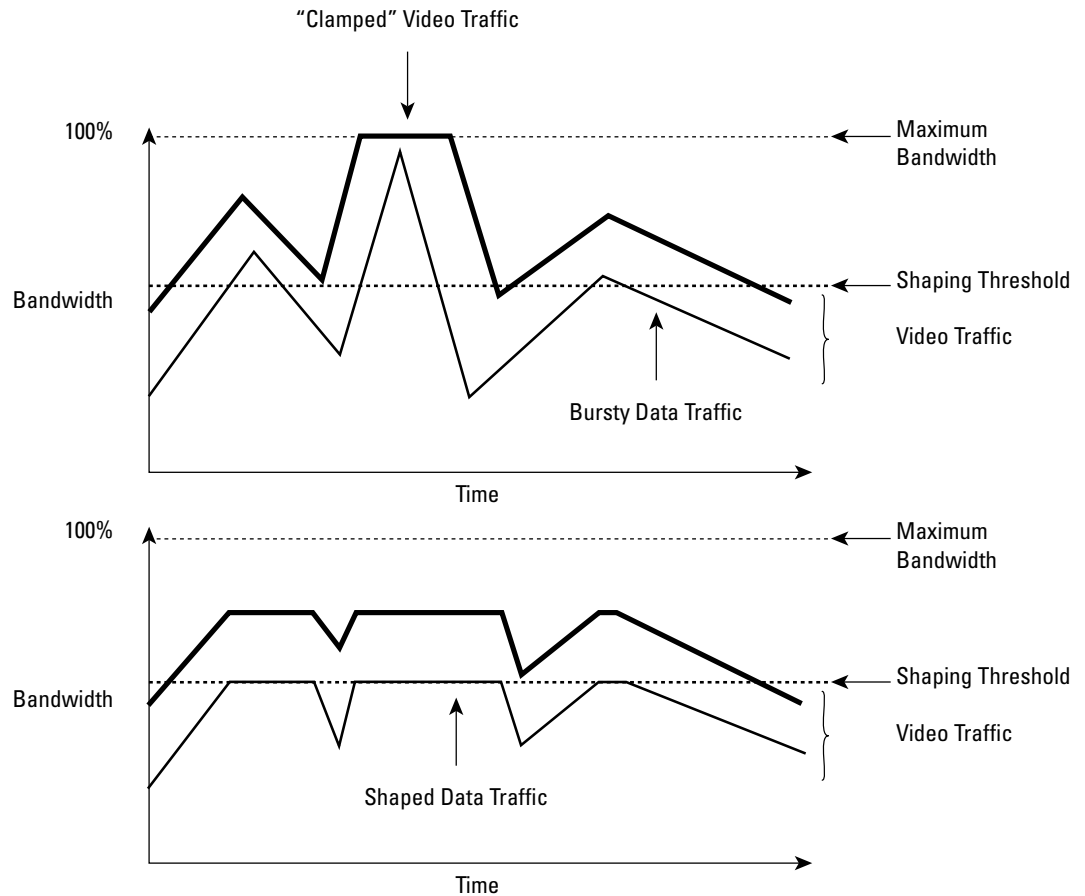
In many organizations, some departments have a higher need for Web or intranet access than others. Marketing departments often search the Internet for competitive data and industry reports. A documentation department might need to download data from a corporate database that is located in another city. From the distribution network, IT managers can rate limit the marketing department to a higher rate for HTTP traffic and rate limit the documentation department to a higher intranet traffic rate from a particular server.

A heavy network user should not affect the network quality of another employee. For example, engineers often run batch jobs. Although batch jobs increase engineer productivity, an engineer who downloads or uses File Transfer Protocol (FTP) to send 20 1-GB design files from the engineering database server should not affect the training video another employee is running at 500 Kbps. The file download must be rate limited, because real-time applications like



training videos are sensitive to jitter caused by bursty traffic. To help mitigate this problem, IT managers can assign video traffic to higher-priority queues and allocate larger buffers for the video traffic queue. Traffic shaping smoothes out the bursty traffic caused by large file transfers (Figure 8).

**Figure 8**  
Traffic Shaping Smoothes Out Network Traffic



In Figure 8, the data traffic is not shaped and is assigned to a higher-priority queue than the video traffic. For a good user experience, video traffic requires constant bandwidth. Videos that pause every few seconds can be frustrating to the viewer. In the top diagram in Figure 8, there is a large burst of data traffic (the video traffic is "clamped" at the maximum bandwidth). When the total required bandwidth exceeds the maximum available bandwidth, video packets are dropped, and the employee experiences jittery video and incomplete video transmissions. In the bottom diagram in Figure 8, the data traffic is shaped. Data traffic that exceeds the shaping threshold is buffered and transmitted later. Data traffic is transmitted at a smoother rate than in the top diagram. No video packets are dropped, so the video experience is not affected.



## Branch Office

Resources are more limited in branch offices. While some work can be performed independently, branch-office employees often have to transfer data to and from the central campus. Depending on the work, an employee may have to transfer a few large files or many small files over the intranet to the central office. With just a T1 (at 1.5 Mbps) connection to the central campus, a 1-GB file requires 1.5 hours to transmit at full T1 speed (this also assumes that nobody else is using the intranet for 1.5 hours). This can greatly decrease employee productivity at branch offices.

IT managers and employees can address this problem by buying more WAN bandwidth, and by better managing bandwidth allocation and file transfer times. Buying two T1 lines can reduce file transfer time by half. IT managers can also rate limit employees so that no one employee monopolizes the WAN link. For example, no one employee can use more than 250 Kbps of WAN traffic at one time. Employees can also be assigned different rate limits depending on the time of the day. IT managers can assign lower rate limits during the day than in the evenings. Employees can also schedule their file transfers in the evenings, after other employees have gone home for the day.

By intelligently regulating traffic from users and applications, rate limiting lowers an organization's total cost of ownership. It allows enterprise customers to transition to Gigabit Ethernet without having to increase the amount of upstream bandwidth. By enabling IT managers to deploy Gigabit Ethernet switches without having to immediately migrate to 10 Gigabit Ethernet uplinks, companies can maximize network availability and performance while saving money.

## Summary

Rate limiting is a powerful tool that provides IT managers with the benefits of improved network management and reduced traffic congestion—enabling more efficient bandwidth utilization in both university and enterprise environments. Rate limiting puts control in the hands of IT managers, allowing them to proactively manage the network rather than reactively trying to fix a network-congestion-related problem after it occurs. Rate limiting also reduces costs by delaying the need for more WAN bandwidth.



**Corporate Headquarters**

Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134-1706  
USA

www.cisco.com  
Tel: 408 526-4000  
800 553-NETS (6387)  
Fax: 408 526-4100

**European Headquarters**

Cisco Systems International BV  
Haarlerbergpark  
Haarlerbergweg 13-19  
1101 CH Amsterdam  
The Netherlands

www-europe.cisco.com  
Tel: 31 0 20 357 1000  
Fax: 31 0 20 357 1100

**Americas Headquarters**

Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134-1706  
USA

www.cisco.com  
Tel: 408 526-7660  
Fax: 408 527-0883

**Asia Pacific Headquarters**

Cisco Systems, Inc.  
Capital Tower  
168 Robinson Road  
#22-01 to #29-01  
Singapore 068912

www.cisco.com  
Tel: +65 6317 7777  
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the

**Cisco Web site at [www.cisco.com/go/offices](http://www.cisco.com/go/offices)**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia  
Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland  
Israel • Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland  
Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden  
Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

All contents are Copyright © 1992–2003 Cisco Systems, Inc. All rights reserved. Catalyst, Cisco, Cisco Systems, and the Cisco Systems logo are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and certain other countries.

All other trademarks mentioned in this document or Web site are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company.  
(0303R) WH/LW4333 0403