

# BPX Congestion Avoidance

## Summary/Abstract

This document discusses traffic management techniques used on ATM WANs and the implementation of traffic management technologies on the Cisco BPX<sup>®</sup> 8600 series wide-area switch. Application performance, service reliability, constrained bandwidth and efficient bandwidth use, and relatively high network latency are the primary forces driving the design of traffic management technologies in ATM WAN switches.

The paper also discusses the basic elements of the ATM Forum Traffic Management 4.0 specification as they apply to practical network implementations. It covers the following traffic management techniques:

- Traffic contract and usage parameter control (UPC)
- Selective cell discard
  - Partial packet discard (PPD)
  - Early packet discard (EPD)
- Explicit forward congestion indication (EFCI)
- Relative rate (RR) mode EFCI
- Available Bit Rate (ABR) flow control with explicit rate
- Virtual source and virtual destination (VS/VD)

## Introduction and Situational Analysis

There is consensus in today's ATM community that sophisticated traffic management is a crucial element of ATM networks. Congestion control—ideally, congestion avoidance—is a critical function, especially in the ATM WAN. Without mechanisms that implement congestion avoidance, fair resource sharing, and firewalling between users and between higher-layer protocols, ATM WAN networks cannot deliver the cell loss ratio (CLR) performance that data applications require. Cell loss severely impacts goodput for data applications such as TCP/IP over ATM, because the loss of even a single cell renders the transmission of a whole IP packet at the higher protocol layer useless, and prompts its immediate retransmission.

Data applications currently represent the justification for ATM deployment in most networks. However, if data applications are not served with predictable quality-of-service (QoS) guarantees and with the negligible CLR they require, ATM will not be viewed as a viable technology that delivers on its initial promise to solve performance issues in dramatically growing data infrastructures. The bottom line is that the explosive growth of data services is the first battleground where ATM can prove itself. And it must win this battle, or network managers will not consider ATM a technology capable of handling other critical applications.

Therefore, for ATM to succeed as a WAN technology, it is crucial that advanced traffic management mechanisms are implemented in ATM switches. This is especially true for ATM service switches in the WAN.

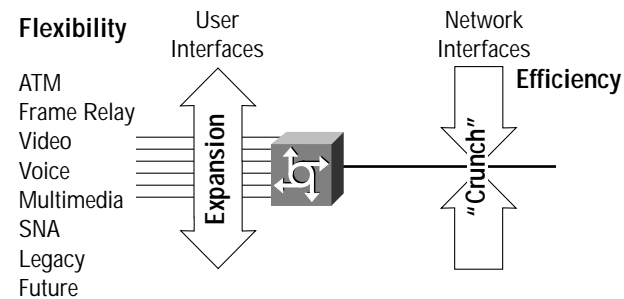
## Technology Drivers

### Why Is Traffic Management Required in the ATM WAN?

For ATM applications in the LAN or campus environment, ATM performance issues can be easily addressed with additional bandwidth. In this environment, bandwidth is a relatively cheap and readily available resource. When increasing traffic patterns lead to cell loss and degraded performance, network managers add additional fiber or implement faster interface speeds.

In the WAN environment, however, bandwidth continues to be a costly resource that must be used as aggressively as possible to boost network efficiency and reduce network operating costs. Thus, mechanisms that use existing capacity more efficiently are crucial, as shown in Figure 1. If those mechanisms deliver predictable performance under all circumstances and enhance application throughput significantly—even better. That's exactly what traffic management mechanisms deliver in the ATM WAN.

Figure 1 ATM WAN Switch Requirements



- Mission: Pack as many services with as much bandwidth as possible into as little bandwidth as possible with the highest quality of service.

In the ATM WAN, the most constricted bandwidth, coupled with a higher overall network latency, drives the development of advanced traffic management mechanisms. The challenge an ATM switch faces is to aggregate a maximum amount of quickly growing user traffic onto a

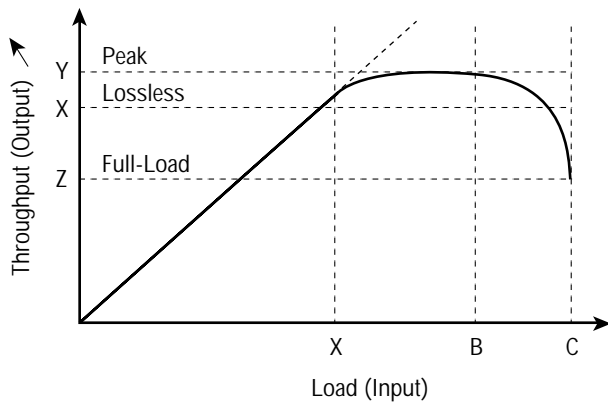
constricted, costly resource: the trunks between the ATM switches in the WAN. The performance of an ATM switch can be measured on how effectively it can handle this user traffic. The mission is clear: to pack as much user traffic as possible over a limited amount of bandwidth with the highest quality of service (QoS)—as perceived by the end user. For data traffic, this implies the avoidance of cell loss. Cisco regards advanced traffic management as a requirement for achieving the desired network performance objectives. These performance standards are quite high for ATM networks, given the expectations in the data community, where ATM must solve all bottleneck problems and increase both performance and efficiency in the usage of resources.

For service providers, the increased network efficiency delivered by a robust traffic management framework means a decisive and strategic competitive advantage over competitors that cannot use their infrastructures as efficiently, while at the same time ensuring that users are going to experience the highest performance and goodput possible. This represents a win-win situation that provides a key business advantage to service providers implementing an ATM infrastructure with ATM switches that support advanced, robust traffic management schemes. Of course for enterprise networks, the cost savings inherent to bandwidth efficiency also represent a compelling competitive advantage.

### Application Throughput and Traffic Management

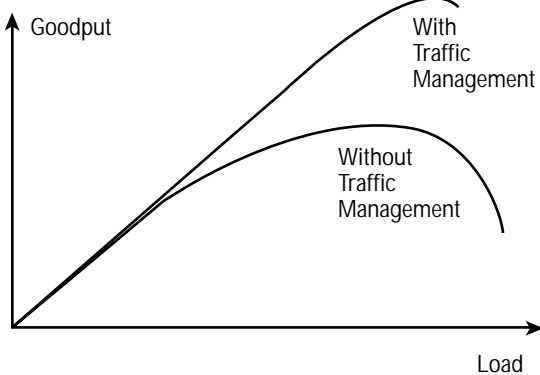
The need for traffic management becomes apparent when visualizing the throughput that data applications can experience over an ATM WAN. Figure 2 shows the benchmarks used when quantifying switch performance. Load refers to the actual input cell rate; throughput—often referred to as “goodput” in this context—refers to the actual effective rate experienced by an application. Ideally, the relationship is linear: the more cells that flow over the physical interface, the higher the IP “goodput” should be. In practice, the relationship without traffic management quickly stops being linear, because sporadic cell loss leads to retransmission of IP packets, which reflects itself in a severely impacted perceived application throughput, or “goodput.” Retransmissions, of course, affect the perceived goodput of the application, because if, for example, 100 packets are sent in one second, but these are all retransmissions of the same packet because of cell loss, the application will only perceive a goodput of 1 packet per second!

Figure 2 Throughput as a Function of Load for Data Over ATM



The relationship between actual cell rate and throughput is linear up to a load X, at which point, the ATM switch starts to drop cells because of buffer overflow, as shown in Figure 3. The lossless throughput (X) is thus achieved at the load point (X) when no cells are being dropped. If the load becomes higher than X, packet retransmissions take place as a consequence of cell loss, and the relationship ceases to be linear. The throughput still increases until reaching its peak (Y) at load (B). If load increases further, many ATM switches run into congestion situations so severe that the throughput starts to drop. At full cell load of the physical interface (C), a very degraded full-load throughput (Z) might be achieved.

Figure 3 Traffic Management and "Goodput"



The curve's shape can change dramatically, depending upon the ATM WAN switch's implementation of internal traffic management algorithms. Small input buffers and more primitive traffic management schemes cause the nonlinear relationship beyond X to be reached at a very modest rate; the peak throughput (Y) becomes significantly lower than the physical line rate would, in theory, allow. Full-load throughput (Z) is a very degraded value in such systems.

The BPX 8600 series has, since its introduction in 1993, implemented very large per-connection ingress buffers and trend-setting advanced traffic management algorithms to deliver an unmatched robustness for data application throughput. This means that with the BPX 8600, the relationship between load and goodput is practically linear; X is pushed toward the right of the diagram, and points B and C come together at the right edge of the diagram, achieving a Z and Y very close to the physical rate. Simply put, applications are going to experience far higher goodput and far more predictable performance across an ATM network implemented with the Cisco BPX 8600 switch.

Cisco's traffic management objective is to deliver maximum application performance; that is, no packet retransmissions and the lowest possible latency across the network, and simultaneously, maximum resource utilization. The following sections describe how this objective has been implemented in full conformance with the ATM Forum's Traffic Management™ 4.0 specification, implementing the specification's most advanced and elaborate traffic management options.

#### ATM Forum Traffic Management 4.0 Specification Options for Congestion Management

The ATM Forum's Traffic Management 4.0 specification lists several possible implementation options for traffic management and congestion control. It is important to note that not all of these techniques are of equal performance and effectiveness. This means that mere compliance with TM 4.0 does not guarantee that the most robust schemes are implemented. So while TM 4.0 provides a framework in which to implement very advanced ATM switches offering outstanding, predictable performance to data traffic, less capable switch platforms with less than optimal data performance can also be designed within the framework.

It is therefore important to understand the different traffic management options specified within the TM 4.0 standard, and to realize that other aspects of switch design (for instance, buffer size and design) are also crucial in determining whether a specific switch design can meet your ATM WAN performance objectives. A white paper focused on the BPX 8600 architecture looks at issues that are not directly related to traffic management in more detail. In the following chapter, we will focus exclusively on traffic management mechanisms and the BPX 8600 implementation.

### Technical Tutorial

The ATM Forum's Traffic Management 4.0 specification defines all traffic aspects within ATM networks, addresses the most important and immediate aspects, and leaves some for further study. Within TM 4.0, the concept of service categories (formerly known as classes of service) is introduced, quality-of-service issues addressed, and traffic contracts and network policing functions discussed. Congestion control and congestion avoidance mechanisms are also addressed. While TM 4.0 discusses a number of issues, it is important to note that when it comes to issues like TCP/IP performance over ATM, a smaller subset of topics within TM 4.0 becomes relevant to the discussion. Unfortunately, this area of discussion has been muddled by many vendors trying to position every aspect of TM 4.0 compliance as a crucial element to achieve high goodput of TCP/IP over ATM networks. Network aspects that in fact have administrative significance, such as policing, are described as cornerstones of traffic management and congestion control in this context. More primitive, merely reactive mechanisms for congestion control are described as if this were state-of-the-art ATM technology, and cell loss were more or less an inevitable side effect of ATM technology deployment. It is important to note the following when it comes to network efficiency and application goodput, the two aspects that ATM technology must successfully address in the WAN:

- Buffer size and architecture (described in the *BPX Architecture* white paper)
- Implementation of proactive congestion avoidance traffic control mechanisms (the BPX 8600 series pioneered this effort)

*In the final analysis, it is only these two aspects that will make all the difference between unacceptable and stellar ATM network performance.*

The traffic contract is negotiated at the time of subscription between the user and the network and gives the network a basis for deciding on the compliance of the user's momentary traffic flow. For instance, if a user subscribes to a constant bit rate (CBR) connection with a peak cell rate (PCR) of 40 Mbps, the network will want to make sure that the user is not intentionally or unintentionally trying to transmit information at a higher rate for that connection. The traffic contract defines the parameters for this purely administrative check, and the usage parameter control (UPC) is the entity that implements and enforces the traffic contract at the network ingress. Given the fact that all the UPC does is check whether the user is complying with his traffic contract, simply forwarding compliant cells into the network, blissfully unaware of the congestion situation within the network itself, it is easy to see that the UPC on its own does not guarantee a low cell loss rate. Its contribution to the network performance aspect is negligible, especially within a large, successful, fast-growing and thus oversubscribed network. The more oversubscribed the network and the more user traffic the network attracts, the less likely it becomes that the UPC is going to ensure a low cell loss rate, acceptable application goodput to users, and satisfactory network resource utilization to network operators.

The traffic contract defines the behavior of a traffic source, which for an ATM WAN would be an end system (ES) controlling an individual connection—a virtual path (VP) or virtual circuit (VC). The parameters (PCR, sustainable cell rate [SCR], cell delay variation tolerance [CDVT], maximum burst size [MBS], and minimum cell rate [MCR]) used to define a specific connection's traffic contract help network operators in assessing which network resources ought to be reserved. Additionally, they can also be useful input for network design or optimization and billing purposes. For the ES, the traffic contract represents a service guarantee—a means to measure network performance and monitor whether the performance requirement is being met.

Figure 4 TM 4.0 Service Categories and Their Traffic Parameters

	ITU TSS	DBR	SBR	SBR		
	ATMF TM 4.0	CBR	rt-VBR	nrt-VBR	UBR	ABR
Traffic Parameters	PCR, CDVT	Yes	Yes	Yes	Yes	Yes
	SCR, MBS, CDVT	n/a	Yes	Yes	n/a	n/a
	MCR	n/a	n/a	n/a	n/a	Yes
QoS Parameters	Peak-to-peak CDV	Yes	Yes	Unspec.	Unspec.	Unspec.
	Max CTD	Yes	Yes	Unspec.	Unspec.	Unspec.
	CLR	Yes	Yes	Yes	Unspec.	1
Other	Feedback	Unspec.	Unspec.	Unspec.	Unspec.	Yes

DBR—deterministic bit rate  
 SBR—statistical bit rate  
 CDV—cell delay variation  
 CTD—call transfer delay  
 CLR—cell loss ratio

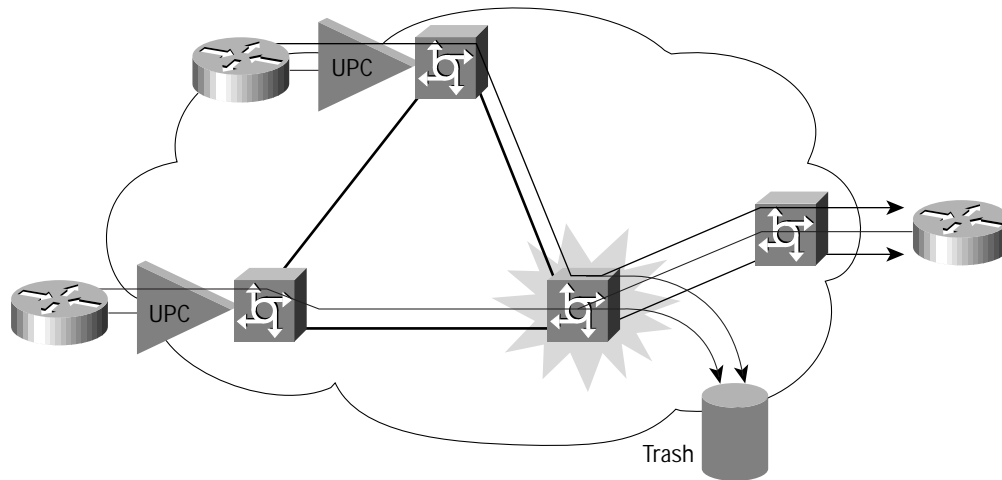
1. Very low when source adjusts cell flow

After the traffic contract has been defined, it must be monitored and enforced by both the ES and the WAN to ensure smooth network operation. On the ES side, the traffic-shaping function smooths momentary bursts that are above the PCR to provide the network with a compliant cell stream.

On the ATM WAN side, the UPC function monitors every incoming cell and determines whether it is within the limits defined by the traffic contract—and thus, whether it is compliant. Depending upon the type of traffic contract and/or the severity of the compliance violation, the UPC function can either tag a noncompliant cell with the cell loss priority (CLP) bit, or discard it immediately.

It is not cost-effective to engineer network resources based upon the simple addition of the traffic parameters of all user traffic contracts. No wide-area network, either a telephone network or any installed data network, is designed to allow all users to access it at exactly the same time, demanding highest quality of service. Instead, ATM WANs are designed to exploit the statistical nature of user traffic to achieve significant cost efficiencies. (The likelihood that every user will simultaneously use the network at maximum contracted throughput is negligible in a large enough network.) Thus, given the nature of WANs, there will always be situations when users as a whole, even though they comply with their individual traffic contracts, contend for network resources that suddenly cannot cope with all requests simultaneously. This means that in addition to the strict UPC policing function, other mechanisms are required within the ATM WAN to resolve contention issues without impacting user quality of service. UPC is simply a network admission policy that sends compliant cells to the network—unaware of the actual congestion situation in the network as shown in Figure 5. It is the responsibility of more sophisticated mechanisms in the ATM network to positively guarantee that the network fulfills its part of the traffic contract bargain and allocates resources in the fairest and most performant way possible.

Figure 5 Congestion Control with UPC, CLP, and/or EFCI Options



- Ingress point has no view of network conditions  
Unconstrained admission leads to heavy discard (UBR)  
"Frame over cell" service goodput reduces dramatically if cells are dropped
- Attempting to avoid discards leads to underloaded trunks (50-70%)
- PPD/EPD are just reactive congestion control measures:  
**Congestion avoidance** in the WAN is the key!


#### "Controlled" Cell Loss Mechanisms

##### Selective Cell Discard Based on CLP

Selective cell discard is a very basic functionality for ATM switches. Strictly speaking, it is not a traffic management mechanism.

During cell assembly or the network's ingress policing function, the convergence sublayer can mark cells as being the first ones that should be discarded under congestion situations (these cells have a *low* cell loss priority). This is achieved by setting the CLP bit in the cell header to 1.

While the cell assembly process in the convergence layer can tag cells with CLP=1 based on real application priority, in practice, the CLP bit is often set based only upon traffic contract compliance. For example, the UPC could tag the CLP bit on all cells that arrive at a rate between the SCR and the PCR. This causes the ATM network to treat cells within the SCR with a higher cell loss priority (cells are kept longer) than cells above the SCR. When that happens, the end application and the ATM network might end up having very different views about the discard priority of the different cells.



In any case, the ATM switch, upon running into a congestion situation, might detect the need to drop some cells to try to alleviate the situation again. Obviously, it seems appropriate that the first cells to be discarded are the cells that have the CLP bit set to 1 and are identified as either low-priority cells or as cells that were transmitted above the SCR or MCR (depending on the service category) by the user.

#### Selective Cell Discard and the Retransmission Problem

An important aspect of selective cell discard is that cell loss is a function of many variables in a congested ATM WAN. Discards happen at the cell level without consideration of higher-layer protocol organization.

For example, a switch could choose to discard ten cells, each from a different user. If these users are running TCP/IP across the ATM WAN, this effectively destroys ten larger frames from higher-level protocols. In TCP/IP, the higher-layer protocol immediately retransmits these frames. If the individual packets have an average length of 256 bytes (six cells), the network would have discarded 10 cells to alleviate a bottleneck, but immediately gets simultaneous requests to retransmit 60 cells. This actually makes the congestion situation worse, as more and more users try to recover from the net effect of network congestion (frame loss) by retransmitting their requests with increasing synchronization.

When the network implements selective cell discard to eliminate congestion, it actually worsens congestion, because the higher-layer protocols on the different ES retransmit even more cells with increasing synchronicity. Dropping data cells, in most cases, leads to worsening network-wide congestion. This can lead to congestive collapse of the ATM network, where the majority of the traffic in the network consists of more and more desperate retransmission attempts that are subsequently discarded within the network. Consequently, selective cell discard must be complemented by other mechanisms that reduce any type of cell loss within the ATM WAN to an absolute minimum.

#### Frame Discard

Frame discard techniques embody the concept that if the network element must discard cells, it is far more effective to do so at the frame level rather than at the cell level. An ATM network can detect frame boundaries by evaluating the cell header. Under severe congestion situations, when a cell is dropped, the switch can immediately drop all cells associated with that specific frame. To improve higher-layer protocol efficiency, the switch should drop all cells except the cell that signals the end of the frame. This way, the higher-layer protocol immediately detects a failure and requests retransmission upon receiving an incomplete protocol data unit (PDU), instead of the source ES having to wait for a higher-layer protocol timeout before retransmitting the PDU, or, even worse, the boundary of the next PDU not being correctly identified and the next PDU also being affected.

#### Partial Packet Discard

Several levels of efficiency are possible within a frame discard environment. With partial packet discard (PPD), the ATM switch, upon dropping a cell, also drops the remainder of cells associated with that PDU except for the last cell comprising the PDU. The last cell of a PDU is identified by setting a bit in the payload type identifier (PTI). If this cell is lost by the network, the segmentation and reassembly function does not know when to start assembling the next PDU, and an error that leads to both PDUs being discarded occurs. While superior to cell-oriented discard, PPD still wastes network resources by allowing partial frames to flow through the network that will subsequently have to be discarded by the destination system anyway.

#### Early Packet Discard

A more performant mechanism can be implemented with early packet discard (EPD). In this case, upon reaching a certain congestion threshold, a switch starts dropping complete PDUs, including the first, all intermediate cells, and the last cell, avoiding less efficient partial drops.

The BPX 8600 implements EPD throughout the ATM network—at ingress, tandem, and egress switches—to maximize network resource protection. On the BPX 8600, the network operator can select, on a per-connection basis, whether CLP or EPD discarding policies are to be used for the connection.

While EPD is implemented in the BPX 8600 series, it is not regarded as a design cornerstone for traffic management. Rather, the whole BPX 8600 system philosophy is geared toward *avoiding any type of cell loss in the first place*—even under very high load. Cisco holds that even a fairly advanced feature such as EPD is a reactive congestion control measure rather than the proactive congestion avoidance that is inherent to the BPX architectural philosophy. So while the feature is implemented, it is viewed as a very last, desperate measure.

Even with EPD, the devastating effect of cell loss on application goodput and the synchronization effect of frame retransmission at higher layers is not resolved. For “frame-over-cell” services to work, the use of EPD must be kept to an absolute minimum, and the ATM network must implement a mechanism that proactively avoids congestion situations and cell loss, thus achieving far more efficient resource utilization and allowing network users to benefit from the superior goodput such a network provides.

**Flow Control Mechanisms**

As established in the previous sections, proactive congestion avoidance is more effective than any cell discard policy, given the fact that higher-layer protocols react allergically to any type of cell loss on the ATM WAN. Thus, mechanisms are called for that intelligently allocate bandwidth and make efficient and effective use of their buffering capabilities under high load situations to avoid cell loss. The TM 4.0 standard defines mechanisms of widely differing complexity and efficiency for flow control.

Flow control mechanisms are implemented using resource management (RM) cells that are inserted into the data cells to convey flow control information, as shown in Figure 6. The RM cells can either simply signal a congestion situation and leave everything else to the ES, or in a more advanced implementation, it can be used to instruct the ES to lower or increase its momentary cell rate to make the most efficient use of network resources.

Figure 6 RM Cell Format

**RM cell format  
ABR and ABT (1371.1)**

		ABR ER	ABT
1-5	ATM Header	VPC: VCI = 6, PTI = 110 VCI: PTI = 110	VPC: VCI = 6, PTI = 110 VCI: PTI = 110
6	Protocol ID	PID = 01	PID = 02/03
7	DIR BN CI NI R/A Resv		
8-9	Explicit Rate	Any Rate	Block Cell Rate
10-11	Current Cell Rate	= ACR	User Block Cell Rate
12-13	Minimum Cell Rate	= MCR	
14-17	Queue Length	0 or see ABT	Block Size
18-21	Sequence Number	0 or see ABT	Sequence Number
22-51	Reserved	0	0
53	CRC	CRC-10	CRC-10

**EFCI**

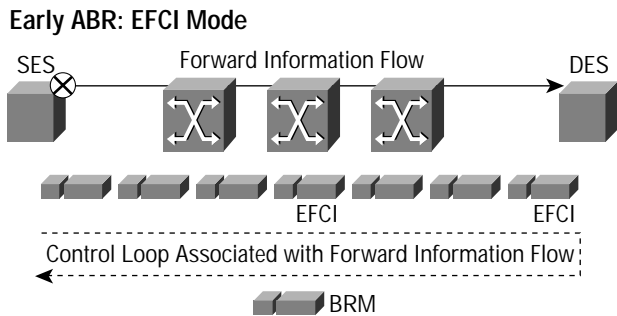
Explicit forward congestion indication (EFCI) is the most basic flow-control mechanism used by ATM switches and takes a more proactive role than dropping cells. With EFCI, upon detecting a congestion situation, network elements set the EFCI bit in the header of user data cells. The ATM network then signals to the ES that a congestion situation is probable, indicating that the user might want to adjust the traffic flow to this circumstance. The EFCI bit is evaluated by the destination ES, which then sends a backward resource management (BRM) cell and sets CI = 1 (congestion indication) to signal the congestion situation to the source ES. If the ES is appropriately configured, it might then lower its allowed cell rate (ACR) by a certain configurable factor and operate at that cell rate until a new change is signaled.

**Relative Rate Mode EFCI**

In addition to setting the EFCI bit in the event of congestion (see Figure 7) and letting the destination ES populate the fields in the BRM cell, some ATM switches modify the cell increase (CI) and the no increase (NI) fields directly. This approach is called the relative rate (RR) mode. The difference

between RR and traditional EFCI mode is very subtle, and the switch can instruct users to not increase their ACR by signaling the NI bit.

Figure 7 Congestion Control by the User with EFCI



- **When congested:**
  - Switch sets EFCI flag
  - Receiver must respond with "marked" RM cells
  - Sender shall slow down upon receiving CI in BRM
  - > Network relinquishes control
- **Only marginal benefits over UBR, requires CPE upgrade.**

Even a simple feedback mechanism such as EFCI, which does not involve any real complexity in the ATM network element, delivers dramatically improved performance to network subscribers. Several tests and papers have shown that ABR with EFCI will always dramatically outperform unspecified bit rate (UBR) with EPD. This is especially true in the WAN, where oversubscription makes network element involvement in the control loop even more important.

**EFCI Shortcomings—User and Protocol Abuse**

Despite the performance benefits of ABR EFCI over simple EPD schemes, and although the switch is involved in the feedback process by signaling congestion situations to the ES, the switch does not exercise direct control over users'

allowed cell rates. Quite the contrary; the switch is dependent upon users' good-will and fair behavior for response to the EFCI-tagged cells. In an ATM WAN, this represents a significant exposure. It is easy to picture a scenario where, during a congestion situation, some users lower their effective cell rates in response to EFCI-tagged cells, while other users—either unintentionally or maliciously—ignore the network's congestion indication and continue to send cells at the same rate. In this case, the network is at the mercy of one user's behavior—a dangerous situation, especially in a WAN environment.

A parallel problem arises with multiple higher-layer WAN protocols and protocol versions in simultaneous operation; each has a different reaction to a congestion situation. If the network does not act to exercise more control over the ES, the protocol that ignores congestion notifications most stubbornly—and operates persistently at a high cell rate—is the one that gets network resources allocated to it.

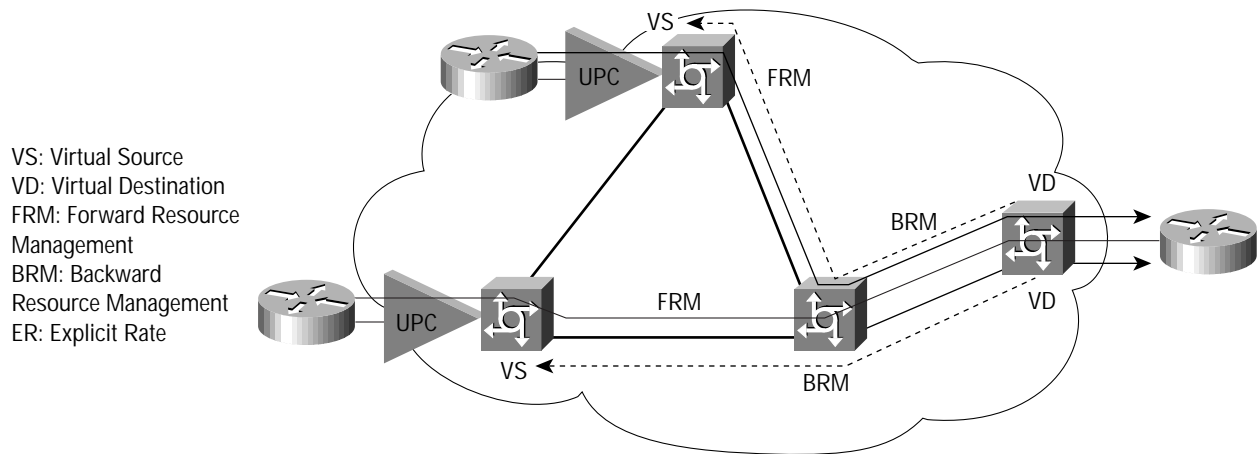
A WAN network that aspires to offer predictable and fair quality of service to its users must be able to firewall applications and users from each other and exercise control over users' information flow, ensuring fairness in the allocation of shared network resources at all times. Clearly, this objective cannot be met with EFCI or RR support—EFCI can penalize all users for network overload caused by only one noncompliant user.

**ABR Flow Control with Explicit Rate**

With the explicit rate (ER) mechanism, ATM switches use resource management (RM) cells to inform users of the exact rate at which they are allowed to operate, as shown in

Figure 8. This provides the ES with accurate information on network utilization and of the resources available for that particular connection in real time. The declared objective of ABR with ER is to avoid cell discard altogether while simultaneously ensuring the most efficient use of network resources.

Figure 8 ABR ER Congestion Avoidance with Closed-Loop Feedback



- With BXM: full TM 4.0-compliant ER implementation for internal (and optionally/additionally external) feedback  
 FRM cells contain PCR in ER field. Network calculates and eventually adjusts the ER value in BRM cells depending on momentary load situation
- Allocates available resources in fair, efficient, and the most performant fashion
- Full backward compatibility to existing cards and Foresight

ABR ER poses a challenge to ATM switch designers, because every switch in the path of a BRM cell must accurately calculate the ER available to that connection and update the BRM cell's ER field accordingly. FRM cells are generated by the source ES at a rate that depends upon the fixed round-trip time (FRTT), usually in the tens of milliseconds range. The source ES initially requests the connection's peak cell rate (PCR) as the ER in its FRM cells. After the FRM reaches the destination ES, it is turned around and sent back as a BRM cell. Every transit ATM switch then calculates the ER that it can support for that particular connection in real time. Typically, in a noncongested environment, the PCR is supported. If, however, the newly calculated ER is lower than the ER already present in the BRM cell, the switch overwrites the ER with the lower value. Thus, the BRM received at the source ES contains the ER value of the most constricted resource along that path. The source ES is then required to adjust its ACR and send cells at the signaled rate until the next BRM arrives with a new ER value.

It is often assumed that calculating the fair ER on a per-connection basis in real time adds complexity—and thus cost—to an ATM switch. Application-specific integrated circuit (ASIC) technology allows ATM switch designers to create specialized chips to perform particular functions in hardware with extremely high performance. Therefore, implementing traffic management in hardware is a matter of making the initial design and hardware investment. The whole product line can then make economical use of the technology.

ABR ER potentially allows ATM switches to achieve two objectives that have traditionally been mutually exclusive in wide-area networks: maximum real-time utilization of network resources while avoiding cell loss. ABR ER provides applications with an continually predictable, fully deterministic MCR tailored to the user application's requirement, while at the same time allowing applications to rapidly ramp up to fully use any available network resources. With ER and the right design, ATM WAN switches can ensure fairness in the allocation of network resources between different applications and protocol versions.

ABR represents the essence of the true advantages of ATM technology, combining the predictability of traditional time-division multiplexed (TDM) networks with the bandwidth efficiency and performance of statistical multiplexing. Cisco expects that emerging and future

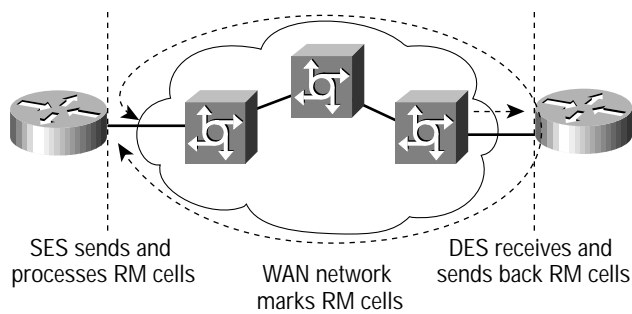
applications will be increasingly bursty and will be designed to be "adaptive," modifying their behavior and QoS expectations to fit the network's current load situation and using the nature of ABR services in the most efficient and performant way possible.

#### Virtual Source and Destination

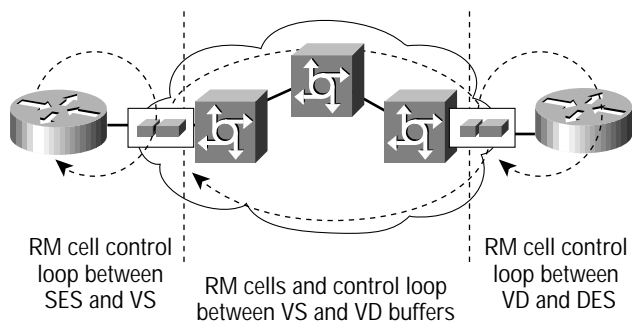
In its basic mode, ABR ER operates the control loop end to end between source and destination. As a further refinement, virtual source, virtual destination (VS/VD) ER segments can be defined within the network. With this architecture, tighter control and enhanced isolation between users and network and different network segments becomes possible. In Figure 9, the upper network drawing represents an implementation where all the control is left to the end systems, and the network simply marks the RM cells to signal congestion. Despite the fact that the network implements the advanced ER option, unfairness could still result if one user does not respect the ER guidelines and decides to ignore the cell rates signaled back by the network.

The lower network drawing shows a VS/VD segmentation that allows the network to implement true control over the internal flows: an ingress and egress per connection buffer implement a virtual source and virtual destination at the WAN network's edge.

Figure 9 VS/VD Implementation in the ATM WAN



**RM Congestion Control Loop without VS/VD**



**RM Congestion Avoidance Loop with VS/VD**

In the simplest, yet highly effective VS/VD implementation, the ATM WAN implements the VS at the ingress switch and the VD at the egress switch. This provides the WAN with far more control over rate allocation. Instead of relying upon the ES to adapt its ACR to the signaled ER, the WAN ensures that misbehaving ESs do not monopolize network resources by operating ABR ER between the VS and VD. Large per-connection queues at the VS are served at the signaled ACR, so user misbehavior does not affect overall network performance or fair resource allocation.

VS/VD creates a clean demarcation between WAN and LAN, and allows optimization of LAN devices for LAN traffic handling and of WAN equipment for WAN traffic handling. Without VS/VD, for example, a LAN switch

connecting to the WAN would have to implement very large buffers to support fast ramp-up times and bursts without cell loss over a connection with higher latency—typically the WAN’s responsibility. With VS/VD, the large buffers need only be implemented on the ATM WAN switch, and the large VS per-connection ingress buffer ensures that fast ramp-ups and bursts are supported without overwhelming network resources and without incurring cell loss.

The LAN and WAN functions are separated. The LAN segment has a low-latency control loop between workstation and WAN access point, which is served with good performance with smaller LAN buffer sizes, and less complex flow-control mechanisms such as EFCI that provide good performance in a LAN environment. The long control loop in the ATM WAN is isolated from the LAN and implements the refinements required for efficient resource utilization and high performance in the WAN: high-performance ER algorithms and large per-connection buffer sizes of tens of thousands of cells.

High-Performance UBR Using VS/VD

VS/VD segmentation allows an ATM network to deliver enhanced performance to ESs that implement UBR. In this case, the ABR VS/VD ER loop operates across the ATM WAN, and none of the traffic management information is passed to the ES. The ES operates over what it assumes to be a UBR connection, but actually experiences much lower cell loss and higher performance than it would over a true UBR service, because the higher-layer protocols take implicit advantage of the additional bandwidth allocated to them. TCP/IP’s slow-start algorithm provides complementary ramp-up/down behavior over UBR in the LAN, but carried as ABR with ER within the ATM WAN VS/VD. As TCP/IP opens its window size (transmitting at incrementally higher and higher rates), ABR ER allocates additional bandwidth, and the large buffers at the network ingress absorb peaks, bursts, or congestion situations. This makes it possible to

maximize performance and investment return of already installed UBR ES while still providing fairness across the WAN.

### **BPX Traffic Management Implementation Details**

Since its introduction in 1993, the BPX series switch has implemented the industry's most advanced and trend-setting traffic-management mechanisms and cell-buffering architectures, evidence that Cisco has a visionary approach to implementing the ATM WAN.

#### **Driving the Standards: ForeSight**

Cisco's ForeSight<sup>®</sup> product was the ATM industry's first congestion avoidance, closed-loop control loop, introduced along with the BPX's introduction in 1993. ForeSight implemented an internal control loop between BPX network ingress and egress points that was a precursor to relative rate. ForeSight complemented the large per-connection ingress buffers to minimize cell loss. In 1993, the ATM market's reception of an advanced traffic management mechanism like ForeSight was rather cool. It was assumed that the huge bandwidth pool and user traffic's statistical behavior would make implementations that went beyond cell discard superfluous. During 1994, as ATM trial networks took on more and more users and increasing volumes of data, network managers noticed that the performance of ATM for frame-based services was disappointing and downright alarming when congestion situations arose, except in ATM networks that deployed the BPX and ForeSight products. The ATM Forum reacted quickly, and the Traffic Management 4.0 specification was defined with considerable input from the ForeSight operational experience. Thus, the BPX pioneered closed-loop control algorithms in the ATM WAN.

#### **Stratm Technology**

While the original ForeSight operated with fully UNI and NNI standards-compliant cell formats, it required a proprietary implementation within the network. Four bits from the VCI field were "borrowed" in the BPX network internal cell format to convey congestion information and rate-up or -down messages from egress to ingress. When the TM 4.0 framework standardization was completed, however, the need for a proprietary implementation disappeared. With the introduction of Stratm<sup>™</sup> technology

into the BPX 8600 broadband switch modules (BXM) function modules, full compliance with the RM cell format was implemented in hardware, while maintaining full compatibility with older ForeSight modules.

The development of ASICs to deliver full ABR ER support in hardware reduces the cost for ABR ER implementation while delivering unmatched performance. With Stratm technology, the BPX 8600 supports fully standards-compliant ABR ER for all access speeds in modules offering the highest available port densities.

The BPX 8600 support of VS/VD also allows for cost-effective implementation of extremely robust and high-performance LAN interconnection services. The LAN side can be optimized for LAN performance, while the WAN side implements the more sophisticated traffic management functions and large buffers sizes to allow for fast ramp-up and initial burst support.

The BPX 8600 can also deliver all types of ABR service (EFCI/RR or ER notification to ES) and two types of UBR service:

- Traditional UBR — Do not reserve any resources in the network and implement a totally opportunist (if somewhat congestion-exposed) UBR transport service.
- Control-loop Enhanced UBR — Implement an ABR ER control loop within the WAN to ensure efficient utilization of resources and maximum performance. This could be implemented with an MCR of 0 or with a minimal MCR that delivers some minimum performance even to UBR users. This is a service characteristic that might prove to be a competitive advantage for carriers or a planning advantage for enterprise networks.

#### **BPX ER VS/VD Implementation**

In a BPX 8600 switch network with BXMs, the basic operation of the VS/VD ER scheme (see Figure 10) appears to be straightforward and simple: the ingress point into the BPX network is usually configured as the VS, and the egress point out of the network is configured as the VD to take advantage of the system's capabilities. The ingress point is the source BXM that provides UNI access, and as a VS it is the BXM's responsibility to generate FRM cells. Interestingly, the forward direction is not too important for rate control; because one of the objectives is to provide feedback messages that are as accurate as possible, it makes more sense to

perform ER-stamping on the BRM cell flow, because this reduces the time lag between the actual real-time calculation and the receipt of the RM cell by the VS. Therefore, only basic CI marking is performed on FRM cells. When the FRM finally reaches the VD, it is “turned around” and converted into a BRM, and the destination BXM and all of the transit BXMs now perform the ER calculation for every BRM received. When the BRM reaches the source BXM, the ER received with the BRM cell is used to update the ACR for the connection. As ingress buffers build up under very high load, it is important that buffers are used as efficiently and fairly as possible. This explains why the source BXM also has a weighted fair queuing algorithm that may overrule the ER signaled by the network when it determines the ACR.

Figure 10 Basic Principle of VS/VD ER Operation in a BPX Network

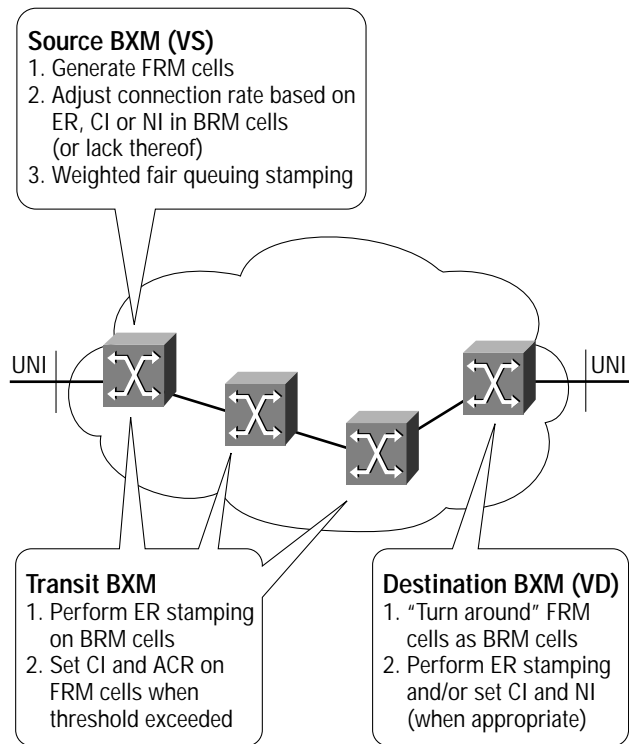
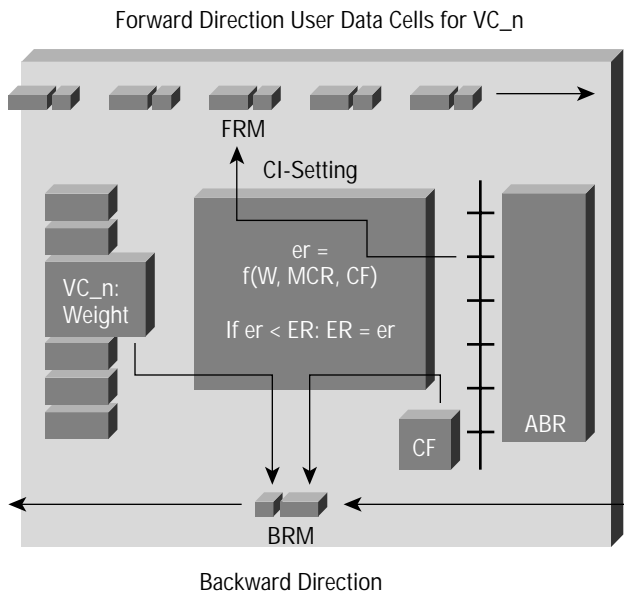


Figure 11 illustrates the operation of the transit BXM cards. In the forward direction, only the CI bit is set. In backward direction, the scheduling and ABR engine (SABRE) ASIC on the BXM card specializes in calculating the fair share of bandwidth and updating the ER field on every

BRM that passes that card. This means that the SABRE can handle BRM cells for up to 64,000 connections every few milliseconds!

Understandably, the actual formula used to derive the ER for the connection cannot be discussed in too much detail. The algorithm considers an administrative weight assigned to the connection, the congestion factor (which is a measure of the utilization of the dedicated ABR service class queue), and the MCR assigned to that particular connection. Based on that formula, the fair ER is calculated. This ER calculation is then compared to the ER already contained in the BRM call. Only if the newly calculated value is lower than the value contained in the BRM does it override the existing ER. This is important because it is crucial that the value that is signaled back to the ingress point is the lowest one and is the one that reflects the status of the most severely utilized resource along the path. Note that the ATM Forum 4.0 specification describes a similar algorithm called congestion avoidance using proportional control (CAPC) in the annex.

Figure 11 ER Marking at Transit BXM

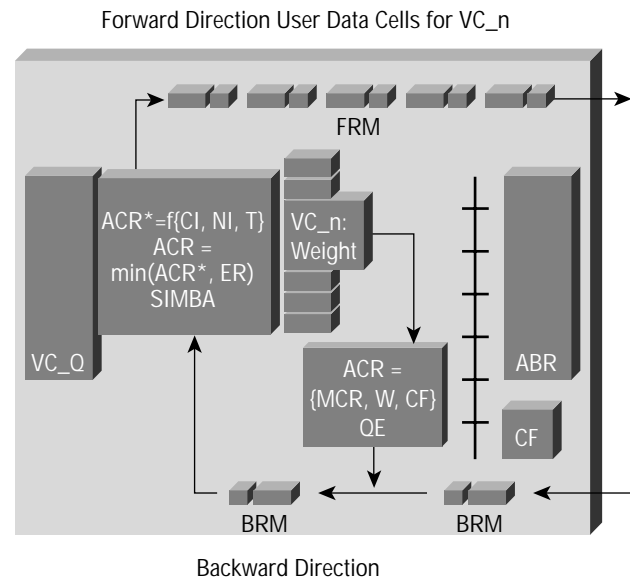


- BXM modules internally support ER stamping
- Each BXM card checks BRM's ER and eventually overwrites it, depending on local load situation
- ER calculation based on Congestion Factor CF, Connection Weight and MCR
- ATM TM 4.0 1.5.2.4  
"Congestion Avoidance using Proportional Control (CAPC)"

After all transit BXMs have performed ER marking, the ingress BXM receives a BRM that accurately reflects the load situation of the network at that exact moment. The VS now typically sets the ACR to be the ER signaled by the network and proceed to empty the per-connection queue at that cell rate. Both network or local congestion situations might lead the local buffers on the ingress cards to build up to the point where a local buffer congestion situation arises. While Stratm technology is capable of up to two million cell buffers per BXM card, this situation is far more unlikely than with any other ATM available switch, a fair policy needs to be implemented for scaling back buffer utilization for the different users. Weighted fair queuing provides for this functionality: a "local" loop on the card considers, very much like in transit BXMs, a congestion factor that reflects the utilization of local buffers, an administrative weight, and the MCR for the connection. Based on this, a "fair cell rate" is calculated that overwrites the ACR value on the BRM cell. The rate scheduler looks at both the signaled ACR and

ER, and takes the lower values to overwrite the ACR at which the per-connection buffer is serviced until the next feedback cycle.

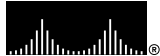
Figure 12 ACR Calculation by the Ingress BXM



### Conclusion—The Stratm Difference

Through implementation of Stratm technology, the BPX once again achieves status as a visionary platform that sets trends in ATM traffic management. The key to the BPX 8600 series' traffic management success is its flexibility; it offers the best possible performance for every possible service and environment. It optimally exploits available resources without allowing congestion situations to spread throughout the network and affect other applications' traffic. In today's environment of exponentially growing data traffic, only an architecture such as the BPX 8600 series has the required robustness and flexibility to ensure the required high quality-of-service levels.

CISCO SYSTEMS



**Corporate Headquarters**

Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134-1706  
USA  
<http://www.cisco.com>  
Tel: 408 526-4000  
800 553-NETS (6387)  
Fax: 408 526-4100

**European Headquarters**

Cisco Systems Europe s.a.r.l.  
Parc Evolic, Batiment L1/L2  
16 Avenue du Quebec  
Villebon, BP 706  
91961 Courtaboeuf Cedex  
France  
<http://www-europe.cisco.com>  
Tel: 33 1 6918 61 00  
Fax: 33 1 6928 83 26

**Americas  
Headquarters**

Cisco Systems, Inc.  
170 West Tasman Drive  
San Jose, CA 95134-1706  
USA  
<http://www.cisco.com>  
Tel: 408 526-7660  
Fax: 408 527-0883

**Asia Headquarters**

Nihon Cisco Systems K.K.  
Fuji Building, 9th Floor  
3-2-3 Marunouchi  
Chiyoda-ku, Tokyo 100  
Japan  
<http://www.cisco.com>  
Tel: 81 3 5219 6250  
Fax: 81 3 5219 6001

**Cisco Systems has more than 200 offices in the following countries. Addresses, phone numbers, and fax numbers are listed on the  
Cisco Connection Online Web site at <http://www.cisco.com>.**

Argentina • Australia • Austria • Belgium • Brazil • Canada • Chile • China (PRC) • Colombia • Costa Rica • Czech Republic • Denmark  
England • France • Germany • Greece • Hungary • India • Indonesia • Ireland • Israel • Italy • Japan • Korea • Luxembourg • Malaysia  
Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal • Russia • Saudi Arabia • Scotland •  
Singapore