Traffic Management

Transmit Priority • Bandwidth Allocation • Call Admission Control • Traffic Policing • Selective Cell Discard • Congestion Avoidance • Traffic Shaping

LightStream traffic management, called ControlStream, allows network administrators to maximize available network resources. It gives them control over network resource allocation, and ensures efficient use of resources that have not been explicitly allocated. The LightStream traffic management system controls two key aspects of the Quality of Service (QoS) on every VCC: delay and bandwidth availability.

The first section of this chapter discusses delay, which is controlled by a mechanism called transmit priority. Delay-sensitive traffic can be given preferential treatment in a LightStream network by assigning it a higher transmit priority.

The remainder of the chapter discusses support for bandwidth availability. Bandwidth availability is controlled by four complementary mechanisms that operate at different levels in the network:

- Bandwidth allocation keeps track of the amount of bandwidth that has been reserved for each VCC.

- Call admission control prevents network users from allocating more bandwidth than the network can provide.

- Traffic policing operates at the edges of the network to ensure that, once a VCC has been established, it does not try to use more bandwidth than the network currently has available.

- Selective cell discard deals with momentary oversubscription of a trunk or edge port. When a traffic surge exceeds the buffer capacity at an output port, this mechanism selectively discards cells, giving preference to different classes of traffic according to parameters set by the network administrator.

The bandwidth availability mechanisms are supported by two additional traffic management features:

- The congestion avoidance system keeps the traffic policers on edge modules informed about how much bandwidth is currently available in the network, so that they only admit traffic that has a high probability of being delivered.

- Traffic shaping meters incoming packet traffic to reduce the occurrence of surges that could exceed the buffer capacity on any output ports.

Transmit Priority

An important aspect of traffic management is the ability to ensure that delay-sensitive traffic (SNA traffic, for example) gets through the network quickly. By setting the transmit priority (also known as forwarding priority or transfer priority) attribute, a LightStream network administrator can control the amount of delay experienced by traffic on a VCC.

When there is more than one cell or packet waiting to be forwarded through a switch, trunk, or edge port, cells or packets on VCCs with a higher transmit priority are always forwarded before cells or packets on VCCs with a lower transmit priority. As a result, traffic on higher priority VCCs will experience consistently lower delay than traffic on lower priority VCCs traversing the same path. Reducing the overall delay on a VCC also generally reduces delay variance.

In addition to the two transmit priority levels available to user traffic, a third (higher) priority level is assigned to internal control traffic such as VCC setup messages and congestion avoidance updates and a fourth (highest) is assigned to CBR traffic. This ensures that the network remains responsive under all traffic conditions. For details on how to set the transmit priority attribute, see the *LightStream 2020 Configuration Guide*.
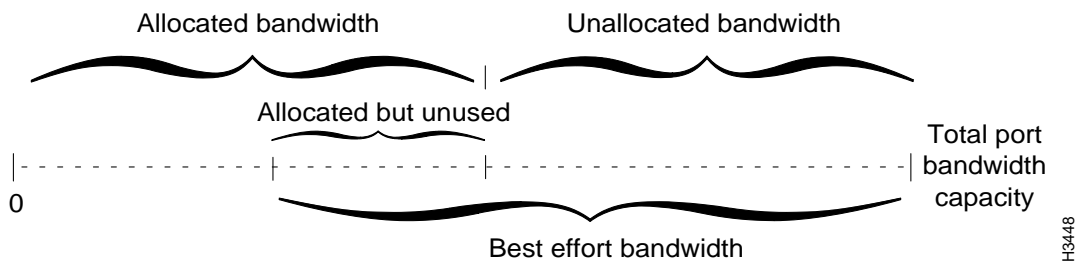
Bandwidth Allocation

To make efficient use of its resources, a LightStream network keeps track of the bandwidth available at each of its trunk and edge ports. There are two types of bandwidth in a LightStream network: *allocated* bandwidth and *best effort* bandwidth. You use allocated bandwidth for traffic that must be passed through the network under all circumstances. This bandwidth is explicitly reserved along the path of a VCC.

Best effort bandwidth is the bandwidth that is available on a trunk or edge port after serving the needs of users of allocated bandwidth. You use best effort bandwidth for traffic that can be dropped if the network is congested.

Figure 4-1 shows the relationship between allocated and best effort bandwidth on a trunk or edge port.

Figure 4-1   **Relationship between allocated and best effort bandwidth on a trunk or edge port.**

The allocated bandwidth is the total amount of bandwidth that has been reserved by VCCs passing through the port. It rises and falls as VCCs are added, removed, or modified. The amount of best effort bandwidth is a combination of:

- The amount of unallocated bandwidth (the difference between the capacity of the port and the allocated bandwidth)

- The amount of allocated bandwidth that is not currently being used

Availability of unallocated bandwidth is tracked by the global information distribution system described on page. Availability of best effort bandwidth is tracked by the congestion avoidance system, which is discussed later in this chapter.

Call Admission Control

The call admission control mechanism determines whether the network can support a requested VCC. It looks to see if a path exists between the two designated endpoints and if there is enough bandwidth along the path to support the new VCC.

When a new VCC is created, its bandwidth requirements are determined by configuration parameters set by the network administrator. Two of these parameters are used by the call admission control mechanism:

- The *Insured Rate* specifies the amount of bandwidth that is explicitly reserved for a VCC. This becomes part of the allocated bandwidth in the network. Traffic that uses allocated bandwidth is referred to as insured traffic.

- The *Maximum Rate* specifies the highest rate at which a VCC is allowed to carry sustained traffic. Beyond the Maximum Rate, all traffic on the VCC will be discarded.

The network will reject a VCC if no path exists with the capacity to accept the full Insured Rate. However, the network will permit a VCC to be built using trunk or edge ports that do not have the capacity to accept the full Maximum Rate.

The LightStream network reserves 100 percent of the Insured Rate for each connection it accepts. As a result, insured traffic is never dropped when congestion occurs. The network reserves only a fraction of the difference between the Maximum Rate and the Insured Rate. This ensures that consumers of best effort bandwidth are distributed evenly across the available trunks. When it reserves bandwidth for a packet interface, the network adjusts the size of the reservation upward to account for the fragmentation that occurs when segmenting variable-length packets into fixed-length cells.

The reservations are implemented by increasing the allocated bandwidth on each trunk and edge port traversed by the VCC. When a VCC is removed from the network, the bandwidth reserved for it is freed by decreasing the allocated bandwidth on each trunk and edge port traversed by the VCC.

Traffic Policing

Traffic policing in a LightStream network is done at the edges of the network for both frame-based and cell-based traffic. This mechanism decides whether to accept a unit of incoming traffic (packet or cell), and whether it is to be carried using allocated or best effort bandwidth.

Every VCC in a LightStream network is controlled by a traffic policer at the input edge port. The operation of the policer is governed by the Insured and Maximum Rates discussed in the previous section, plus two additional parameters, Total Rate and Excess Rate, as shown in Figure 4-2.

Figure 4-2   Relationship between VCC traffic policing parameters.
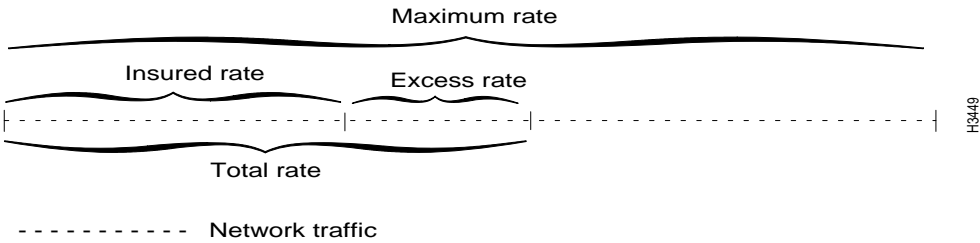
The Total Rate is the total amount of bandwidth that the LightStream network is currently accepting for an individual VCC. This rate varies over time depending on the information received from the congestion avoidance system. The Total Rate will never be lower than the Insured rate, and will never be higher than the Maximum Rate. The Excess Rate is the difference between the Total Rate and Insured Rate.

The operation of the policer is also influenced by two parameters not shown in Figure 4-2, called *Insured Burst* and *Maximum Burst*. These are per-VCC configuration parameters set by the network administrator. They determine how much traffic can be instantaneously buffered for an individual VCC.

As traffic arrives for transmission on a VCC, the LightStream network uses the Total Rate and Maximum Burst parameters to determine which traffic, if any, should be dropped before it even enters the network. The Insured Rate and Insured Burst are used to distinguish between insured traffic (using allocated bandwidth) and best effort traffic (using best effort bandwidth), which may be dropped within the network should congestion occur.

The Leaky Bucket Algorithm

LightStream traffic policers use the dual leaky bucket algorithm required by the ATM Forum UNI specification. As implied by the name, a leaky bucket behaves like a bucket with a hole in its bottom. If data flows into the bucket faster than it flows out of the bucket, then the bucket eventually "overflows," causing data to be discarded until there is enough room to accept new data again.

More precisely, a leaky bucket uses two parameters to control the flow of traffic:

- Average rate — the average number of cells per second that are "drained" from the leaky bucket (that is, allowed to enter the network).

- Burst — the number of cells that are allowed to accumulate in the bucket, expressed in seconds. For example, if the average rate is 10 cells per second, a burst of 10 seconds allows 100 cells to accumulate in the bucket.

It also uses two state variables:

- Current time — the current wall clock time.

- Virtual time — a measure of how much data has accumulated in the bucket, expressed in seconds. For example, if the average rate is 10 cells per second and 100 cells have accumulated in the bucket, then the virtual time will be 10 seconds ahead of the current time.

The leaky bucket algorithm operates on each incoming cell as follows:

```
virtual time = max (virtual time, current time)
if (virtual time + 1/average rate > current time + burst)
    drop the incoming cell
else
    put the cell in the bucket
    virtual time = virtual time + 1/average rate
```

If, for example, the average rate is 10 cells per second, and the burst is 50 cells, then the virtual time and current time will remain the same as long as the input rate remains at or below 10 cells per second. If an instantaneous burst of 25 cells is received, the virtual time will move ahead of the current time by 2.5 seconds. If this is followed immediately by a second burst of 30 cells, then the virtual time will move ahead of the current time by 5 seconds, and the last 5 of the 30 cells would be dropped.

For packet traffic, the unit of incoming data is larger than a single cell. For packet interfaces, the leaky bucket algorithm takes the packet size into account as follows:

```
virtual time = max (virtual time, current time)
if (virtual time + (packet size / average rate) >                    current
time + burst)
    drop the incoming packet
else
segment the packet into cells
put the cells in the bucket
virtual time = virtual time + (packet size/average rate)
```

In this version of the algorithm, packet size is the number of cells required to transport the packet across the network, including the overhead imposed by the ATM adaptation layer.

Note that the algorithm drops the entire packet if it does not fit into the space available in the leaky bucket. It is therefore important to make sure the Burst size on a packet interface is large enough to accommodate at least one or two maximum size packets.

The arrangement of the two leaky buckets in a LightStream traffic policer is shown in Figure 4-3.

Figure 4-3   Dual leaky bucket traffic policer.

6. Cells flow from Insured Rate and Excess Rate
   buckets into circuit at total rate.
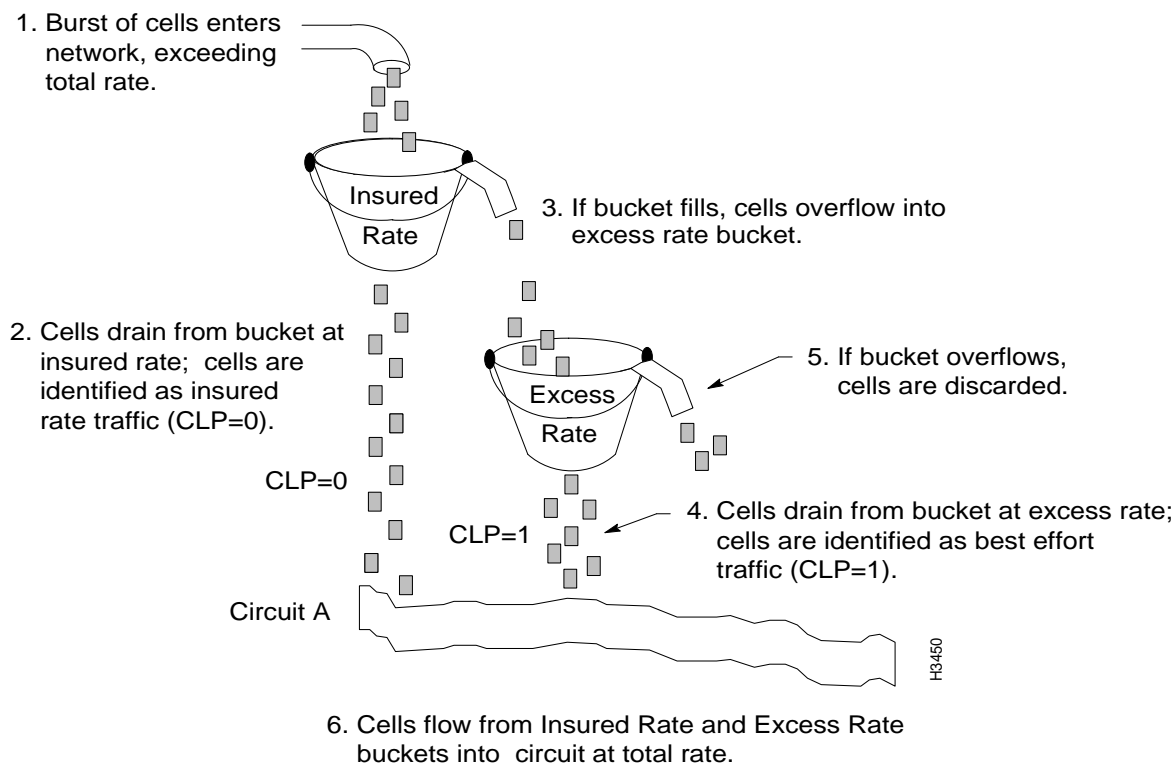
The Insured Rate bucket in Figure 4-3 determines whether an incoming unit of data (packet or cell) can be accommodated by the insured bandwidth for this VCC. The parameters for this leaky bucket are the Insured Rate and the Insured Burst for the VCC. If the test succeeds, the unit of data is segmented into cells (if it is a packet) and prepared for transmission into the LightStream switch.

The Excess Rate bucket determines whether there is enough best effort bandwidth to accommodate the incoming unit of data. The parameters for this leaky bucket are the Excess Rate and Maximum Burst for the VCC. If the test succeeds, the unit of data is segmented into cells (if it's a packet) and prepared for transmission into the LightStream switch.

All traffic entering the network through the Excess Rate bucket is tagged by setting the cell loss priority (CLP) bit in the cell header. This allows the selective cell discard mechanism to distinguish traffic that is using best effort bandwidth from traffic that is using allocated bandwidth.

Note   There is one special case not shown in Figure 4-3. On an ATM user-network interface (UNI), the user device may explicitly tag cells by setting the CLP bit in the ATM header. Since these cells are treated as best effort traffic, they are passed directly to the Excess Rate bucket. Ordinarily, the user device sends enough CLP = 0 traffic to consume the bandwidth that has been reserved for the VCC, and the CLP = 1 cells are regulated by the Excess Rate bucket along with any CLP = 0 traffic that exceeds the reserved bandwidth. In the unusual case where the user does not send enough CLP=0 traffic to consume the reserved bandwidth, *and at the same time* sends more CLP = 1 traffic than the network is currently admitting, the traffic policer assigns the unused reserved bandwidth to the CLP = 1 traffic.

Examples: Traffic Policing in Action

The following examples illustrate the operation of the LightStream traffic policers.

- For applications that require reliable delivery of a predictable flow of traffic, it is best to reserve enough bandwidth to carry the maximum expected data rate. To accomplish this, the network administrator sets both the Insured Rate and the Maximum Rate to the maximum expected data rate. As long as the data rate stays within the Insured Rate and Insured Burst, all traffic is forwarded. Any traffic that exceeds this rate is dropped. All cells flow into the network through the upper leaky bucket, and the setting of the CLP bit indicates that they are using allocated bandwidth. This mode of operation is similar to (but not identical to) that of a time division multiplexing (TDM) switch, where a fixed amount of bandwidth is reserved for each user.

- For applications such as file transfer where the user wants access to all of the available bandwidth between two points on an irregular basis, one might choose to use best effort bandwidth only. To accomplish this, the network administrator sets the Insured Rate to zero, and the Maximum Rate to the highest expected data rate. In this case, the amount of bandwidth available to the connection is regulated by the congestion avoidance algorithm. All cells flow into the network through the lower leaky bucket, and the setting of the CLP bit will indicate that they are using best effort bandwidth. This mode of operation is similar to that of a packet switch or router, where the user has access to all of the available bandwidth, but no guarantees.

- For some applications, it is useful to reserve enough bandwidth to cover routine traffic, and to allow access to best effort bandwidth during peak usage periods. To accomplish this, the network administrator sets the Insured Rate to accommodate the largest traffic rate expected under routine circumstances, and the Maximum Rate to accommodate the largest non-routine traffic rate. With these settings, all traffic within the Insured Rate and Burst uses allocated bandwidth, and all traffic between the Insured and Maximum Rates uses best effort bandwidth. This mode of operation combines the best of both TDM and packet technology, as each user has access to all of the available bandwidth in the system, with a minimum amount reserved at all times.

Selective Cell Discard

Most of the time, the traffic policers only admit as much traffic as the network can handle. Occasionally, traffic surges may occur at several different sources simultaneously and overload trunk or output ports to the point where cells must be discarded. When this occurs, the cells are selected for discard according to the drop eligibilities that have been assigned to them at the edge of the network.

A cell may be assigned one of three levels of drop eligibility, as shown in the table below. Since insured cells use allocated bandwidth, they are never selected for discard when congestion occurs. Best effort and best effort plus cells consume unused bandwidth, and may therefore be dropped. The two levels of best effort drop eligibility are assigned on a per VCC basis by setting a configuration parameter.

Table 4-1　　　　Drop Eligibility

| Type of Service	Drop Eligibility	Description
Best Effort	Most eligible to be dropped	Dropped first when network congestion occurs
Best Effort Plus		Dropped after best effort when congestion occurs
Insured (also known as guaranteed)	Least eligible to be dropped	Never dropped when congestion occurs

Congestion Avoidance

The LightStream congestion avoidance mechanism provides real-time control for preventing congestion and for reacting to congestion if it occurs. Network congestion occurs when the offered load exceeds the capacity of a network resource. The results of congestion are increased delay and reduced throughput. Since the LightStream network does not permit overallocation of insured bandwidth, insured traffic is never affected by congestion. Therefore, the LightStream congestion avoidance algorithm regulates best effort and best effort plus traffic only.

Congestion occurs principally because the allocation of resources for best effort traffic relies on its statistical nature. Since it is highly unlikely that every source will generate a traffic burst at the same time, networks are always designed with less internal capacity than the total input and output capacity of the attached hosts. This is the basis of the economic advantage of a network over a collection of dedicated lines.

When short bursts of traffic occur that exceed the capacity of a network resource, the selective cell discard mechanism drops best effort traffic inside the network. However, this solution only works for short-lived congestion problems. If the offered traffic continues to exceed the capacity of a network resource, it is more efficient to drop traffic at the edge of the network since this allows more bandwidth to be used by traffic that will reach its destination. The LightStream congestion avoidance system accomplishes this by monitoring resource utilization within the network and periodically updating the traffic policers to admit only as much best effort traffic as the network can transport.

The LightStream network's congestion avoidance feedback mechanism operates as a continuous loop. Trunk and edge modules periodically generate congestion avoidance updates and pass those updates to their associated NPs. Each NP then concentrates this information into a larger update and sends it out to every edge module in the network.

Figure 4-4 **The congestion avoidance (CA) feedback loop.**

.S2020 switch
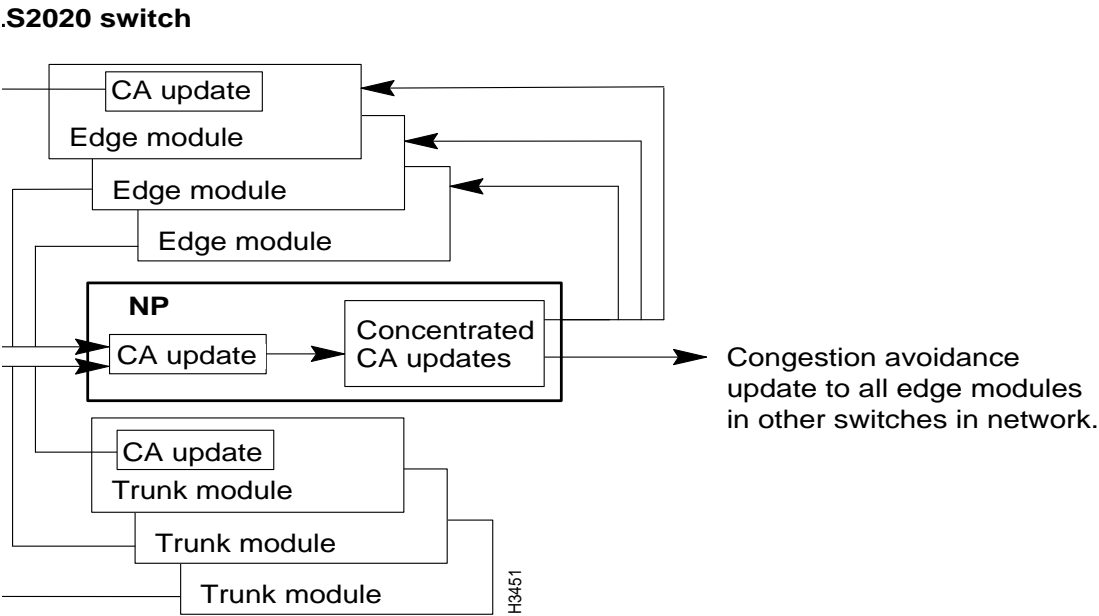
The net result is that the traffic policer for every VCC is continually updated to admit only as much best effort traffic as the network has the capacity to handle. When a surge of traffic hits a trunk or output port, all of the VCCs traversing the port are quickly throttled. When the surge dies out, all of the VCCs are allowed to send at higher rates.

Figure 4-5 shows the effect of the congestion avoidance system on the traffic policers for a VCC that is carrying both best effort and insured traffic.

Figure 4-5 **Internal services and thresholds when traffic is dropped at the edge of the network.**
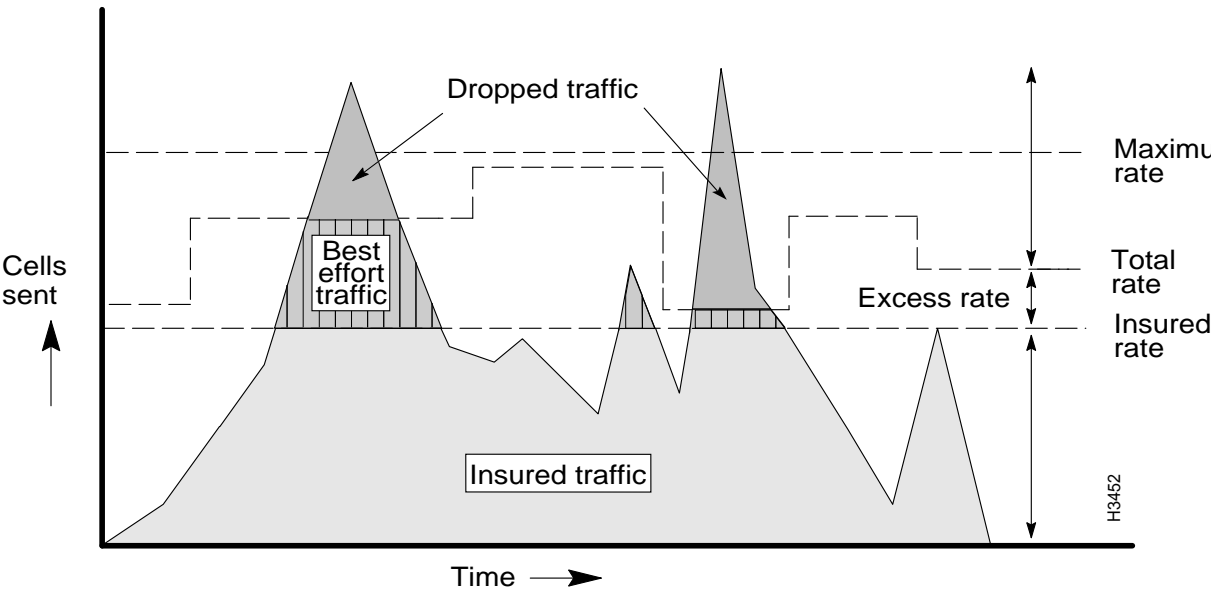

Some important characteristics of the congestion avoidance system include:

- Whenever the insured traffic on a trunk or output edge port does not consume all of the bandwidth reserved for it, the congestion avoidance system makes the remaining bandwidth available to best effort traffic, along with any unreserved bandwidth. This is a key difference between a LightStream switch and a TDM switch, which cannot dynamically reallocate reserved bandwidth.

- The estimates in a congestion avoidance update indicate the total available best effort bandwidth per VCC. Thus, the estimates take into account the number of VCCs traversing the trunk or output line in addition to the amount of traffic.

- For packet traffic, all the cells in a packet are either dropped or sent into the network. This behavior (as contrasted with random cell dropping) has been shown to maximize the "goodput" of TCP/IP traffic when there is network congestion.

Traffic Shaping

Traffic shaping minimizes the occurrence of large bursts of traffic on the network. Traffic is shaped by placing it into buffers and delaying its entry into the network. This metering mechanism causes traffic to enter the network at a more consistent rate.

Figure 4-6   Traffic shaping.
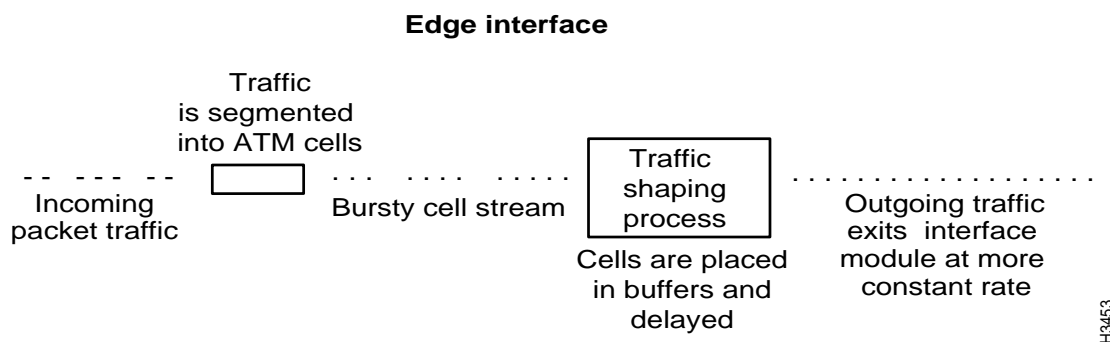
In a LightStream network, traffic shaping is applied at all packet interfaces, as shown in Figure 4-6. Traffic entering ATM UNI interfaces does not need to be shaped, since it obeys the traffic policing parameters set for each virtual circuit.