CHAPTER 4

Traffic Management

The LightStream 2020 multiservice ATM switch (LS2020 switch) traffic management facility, called ControlStream, allows network administrators to maximize available network resources. ControlStream provides control over network resource allocation and ensures efficient use of resources that have not been explicitly allocated. This traffic management facility controls two key aspects of the quality of service (QoS) on every VCC:

- Delay
- 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 an LS2020 network if you assign it a higher transmit priority.

The remainder of this 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:

- **Rate-based congestion avoidance system**—Keeps the traffic policers on edge modules informed about how much bandwidth is currently available in the network, so that they admit only 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 port.

Transmit Priority

Delay-sensitive traffic (SNA traffic, for example) must get through the network quickly. By setting the transmit priority attribute (also known as forwarding priority or transfer priority), an LS2020 network administrator can control the amount of delay experienced by traffic on a VCC.

When more than one cell or packet is 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 experiences consistently less delay than traffic on lower priority VCCs traversing the same path.

Three transmit priority levels are intended for user data traffic. A fourth (higher) priority level is assigned to internal control traffic (such as VCC setup messages and congestion avoidance updates), and a fifth (highest) priority level is assigned to user CBR traffic. These two additional priority levels ensure that the network remains responsive under all traffic conditions. For details on setting the transmit priority attribute, see the *LightStream 2020 Configuration Guide*.

Bandwidth Allocation

To make efficient use of network resources, an LS2020 network keeps track of the bandwidth available at each of its trunk and edge ports. There are two types of bandwidth in an LS2020 network:

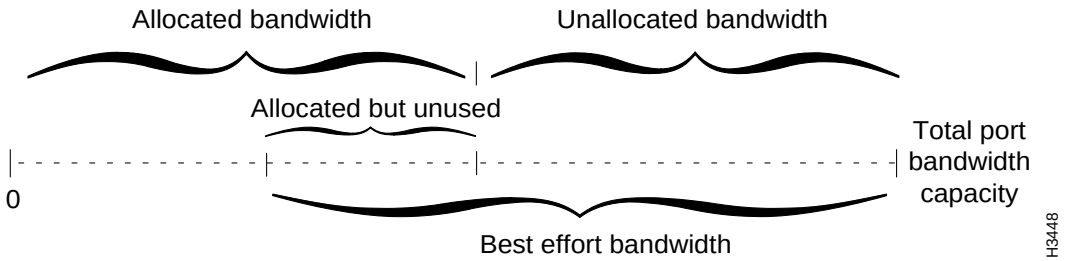
- Allocated bandwidth
- Best effort bandwidth

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 available on a trunk or edge port after the allocated bandwidth needs have been served. You use best effort bandwidth for traffic that can be dropped during network congestion.

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

Figure 4-1 Allocated and Best Effort Bandwidth on a Trunk or an Edge Port



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

- 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 (GID) system described in the chapter entitled "Network Services." Availability of best effort bandwidth is tracked by the rate-based 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 for the VCC and determines 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:

- **Insured Rate**—This parameter specifies the network bandwidth explicitly reserved for a VCC. Traffic for which bandwidth is specifically allocated is referred to as insured traffic. The bandwidth allocated to the VCC becomes part of aggregate network bandwidth.
- **Maximum Rate**—This parameter specifies the highest average transmission rate at which a VCC is allowed to carry network traffic on a sustained basis. Beyond the specified maximum rate, all traffic on the VCC is discarded.

The network rejects a VCC if no path exists with the capacity to accept the full insured rate. However, the network permits a VCC to be built through the use of trunk or edge ports that do not have the capacity to accept the full maximum rate.

The LS2020 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 available trunks. When it reserves bandwidth for a packet interface, the network adjusts the reservation size upward to account for fragmentation that occurs when variable-length packets are segmented into fixed-length cells.

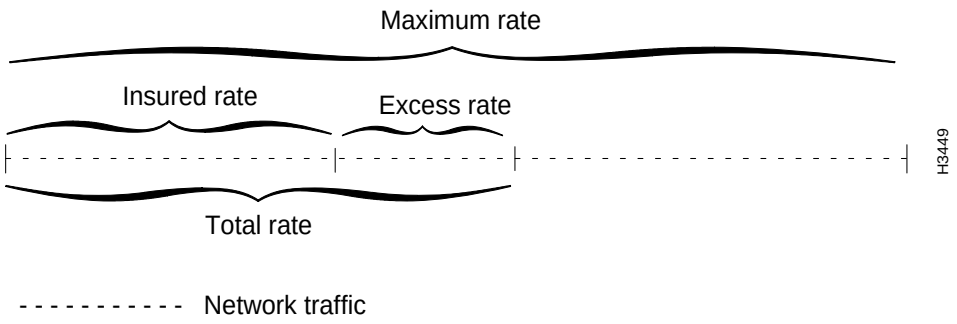
The reservations are implemented through in increase in the allocated bandwidth on each trunk and edge port traversed by the VCC. When a VCC is removed from the network, the bandwidth reserved for it is freed by a reduction in the allocated bandwidth on each trunk and edge port traversed by the VCC.

Traffic Policing

Traffic policing in an LS2020 network is done at the edges of the network for both frame-based and cell-based traffic. The traffic policing 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 an LS2020 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 tate and excess rate (see Figure 4-2).

Figure 4-2 Relationship among VCC Traffic Policing Parameters



The *total rate* is the aggregate bandwidth that the LS2020 network is currently accepting for an individual VCC. This rate varies over time depending on the information received from the rate-based congestion avoidance system. The total rate is never lower than the insured rate, and it is never higher than the maximum rate. The *excess rate* is the difference between the total rate and the 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 items are per-VCC configuration parameters set by the network administrator. They determine how much traffic can be buffered instantaneously for an individual VCC.

As traffic arrives for transmission on a VCC, the LS2020 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 parameters are used to distinguish between insured traffic (using allocated bandwidth) and best effort traffic (using best effort bandwidth). Best effort traffic can be dropped within the network should congestion occur.

Leaky Bucket Algorithm

LS2020 traffic policers use the leaky bucket algorithm required by the ATM Forum UNI specification. The leaky bucket algorithm behaves like a bucket with a hole in its bottom. If data flows into the bucket faster than it flows out, the bucket eventually “overflows,” causing data to be discarded until there is enough room again for new data to be accepted.

The leaky bucket algorithm uses two parameters to control traffic flow:

- **Average rate**—The average number of cells per second that are “drained” from the leaky bucket (allowed to enter the network).
- **Burst**—The rate at which cells are allowed to accumulate in the bucket, expressed in cells per second. For example, if the average rate is 10 cells per second, a burst of 10 seconds allows 100 cells to accumulate in the bucket.

The leaky bucket algorithm 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 indicated in the following formula:

```
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, the virtual time and current time 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 moves ahead of the current time by 2.5 seconds. If this is followed immediately by a second burst of 30 cells, the virtual time moves ahead of the current time by 5 seconds, and the last 5 of the 30 cells are 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 shown in the following formula:

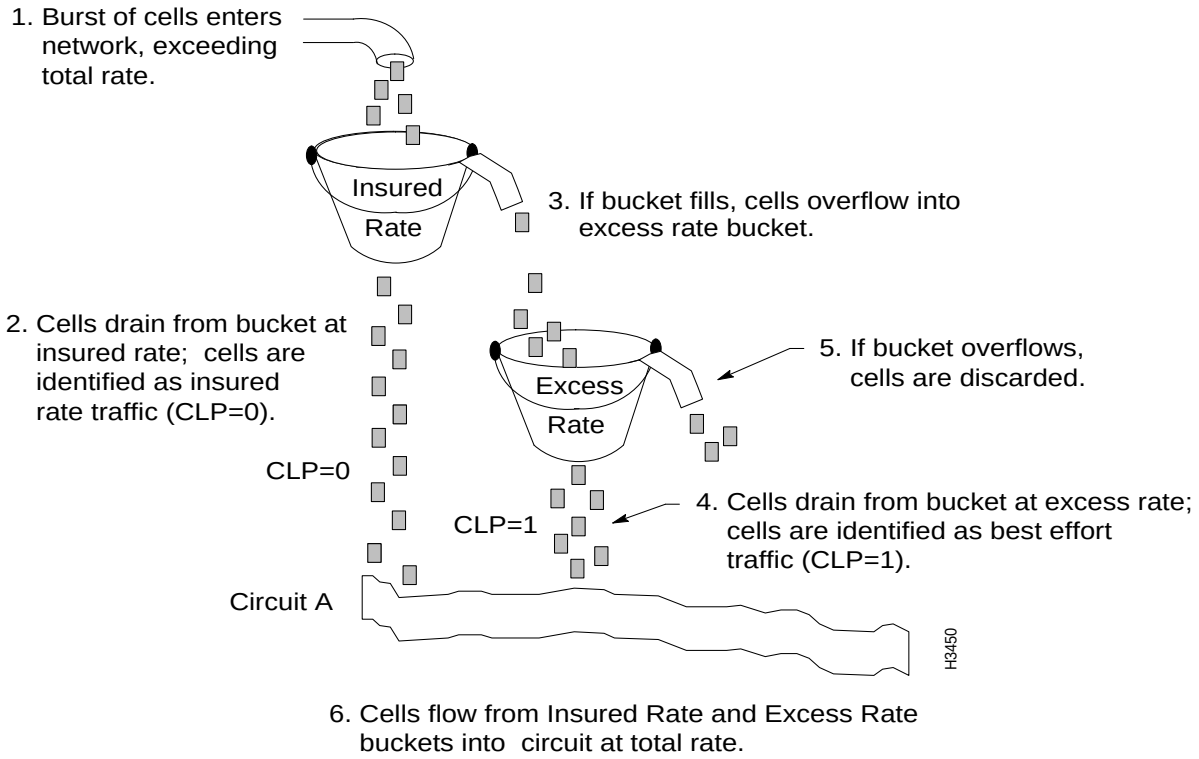
```
virtual time = max (virtual time, current time)
if (virtual time + (packet size / average rate) >   current time + burst)
    drop the incoming packet
else
segment the packet into cells
put the cells in the bucket
virtual time = virtual time + (packet size/average rate)
```

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.

The algorithm drops the entire packet if it does not fit into the space available in the leaky bucket. Therefore, it is 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 leaky buckets in an LS2020 traffic policer is shown in Figure 4-3.

Figure 4-3 Dual Leaky Bucket Traffic Policer



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 LS2020 switch.

The excess rate bucket determines whether enough best effort bandwidth is available 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 is a packet) and prepared for transmission into the LS2020 switch.

All traffic entering the network through the excess rate bucket is tagged by having its cell loss priority (CLP) bit in the cell header set to "1." This allows the selective cell discard mechanism to distinguish between traffic that is using *best effort* bandwidth and 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 can explicitly tag cells by setting the CLP bit in the ATM header. Because 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 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 CLP=1 traffic.

Traffic Policing Examples

The following examples illustrate how LS2020 traffic policers operate in typical operating scenarios.

- **Reliable delivery and predictable flow**—For applications that require reliable delivery of a predictable traffic flow, 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, in which a fixed amount of bandwidth is reserved for each user.

- **File transfer applications**—For file transfer applications where the user wants access to all 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 rate-based congestion-avoidance algorithm. All cells flow into the network through the lower leaky bucket, and the setting of the CLP bit indicates 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 the available bandwidth, but no guaranteed bandwidth.

- **Bandwidth reservation**—For some applications, it is useful to reserve enough bandwidth to accommodate routine traffic and to provide 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 conditions, and sets the maximum rate to accommodate the largest non-routine traffic rate. With these settings, all traffic within the insured and burst rates 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 technologies, because each user has access to all the available bandwidth in the system, and a minimum amount of bandwidth is reserved at all times.

Selective Cell Discard

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

A cell can be assigned one of three levels of drop eligibility (see Table 4-1). Because 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, therefore, can be dropped. The two levels of best effort drop eligibility are assigned on a per-VCC basis through the setting of a configuration parameter.

Table 4-1 Cell Drop Eligibility

Type of Service	Drop Eligibility	Cell Action
Best effort	Most eligible to be dropped	Dropped first when network congestion occurs
Best effort plus	Less eligible to be dropped	Dropped after best effort when congestion occurs
Insured (also known as guaranteed)	Least eligible to be dropped	Never dropped when congestion occurs

Rate-Based Congestion Avoidance

The LS2020 rate-based congestion avoidance system monitors resource utilization within the network and periodically updates traffic policers to admit only as much best effort traffic as the network can transport.

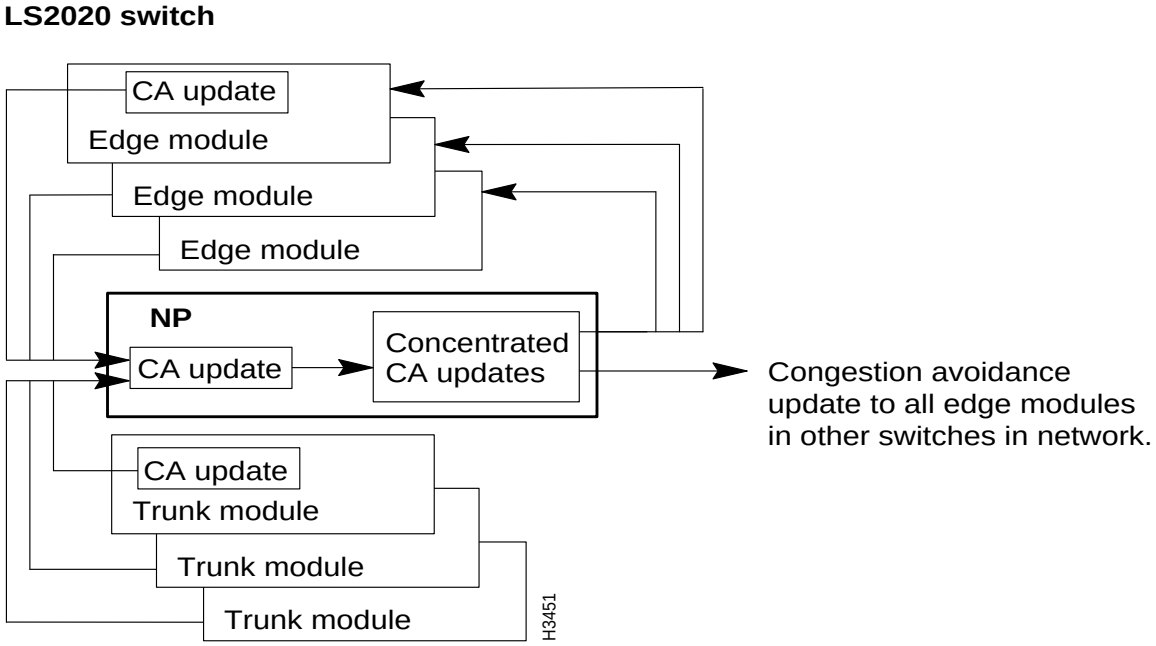
This system provides real-time control for preventing congestion and for reacting to congestion if it occurs. Network congestion occurs when the offered load exceeds the capacity of a network resource. The results of congestion are increased delay and reduced throughput. Because the LS2020 network does not permit over-allocation of insured bandwidth, insured traffic is never affected by congestion. Therefore, the LS2020 rate-based 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 traffic source will generate a traffic burst at the same time, networks are generally designed with less internal capacity than the total input and output capacity of attached hosts. This is 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 works only 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 LS2020 network's rate-based congestion-avoidance feedback mechanism operates as a continuous loop (see Figure 4-4). Trunk and edge modules periodically generate congestion avoidance updates and pass the updates to associated NPs. Each NP then concentrates this information into a larger update and sends it out to every edge module in the network.

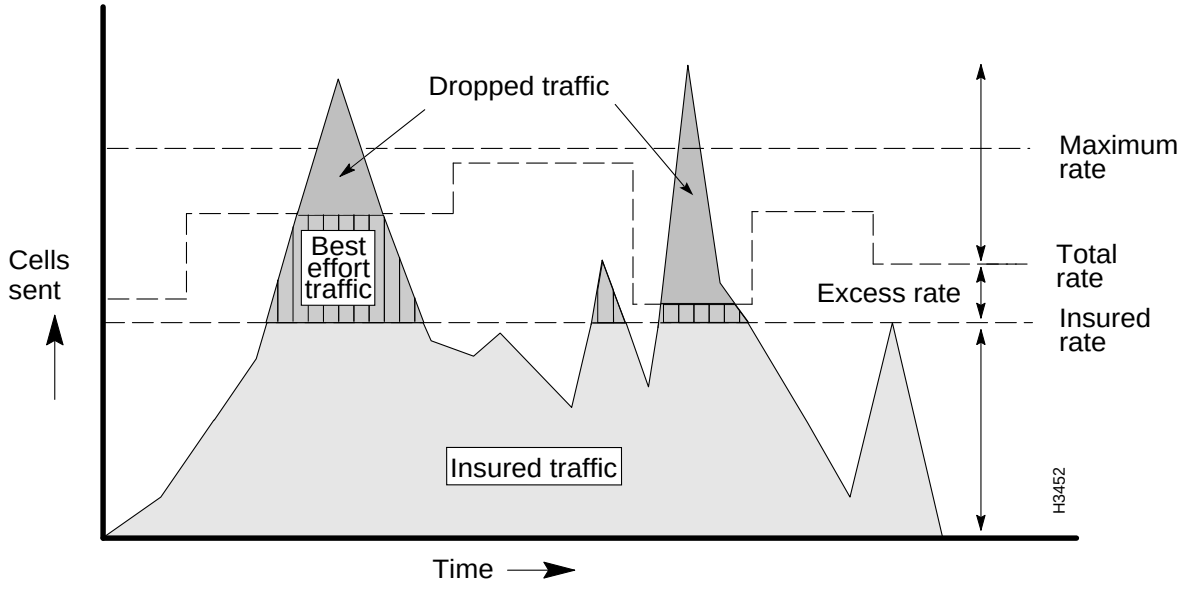
Figure 4-4 Rate-Based Congestion Avoidance Feedback Loop



The traffic policer for every VCC is continually updated and admits only as much best effort traffic as the network has the capacity to handle. When a surge of traffic impacts a trunk or output port, all the VCCs traversing the port are quickly throttled. When the surge abates, all the VCCs are allowed to send at higher rates.

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

Figure 4-5 Congestion Avoidance Thresholds



Three important characteristics of a rate-based congestion system are:

- Whenever the insured traffic on a trunk or output edge port does not consume all 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 an LS2020 switch and a TDM switch, which cannot dynamically reallocate reserved bandwidth.
- The estimates in a congestion avoidance calculation 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 (unlike random cell dropping) maximizes the “goodput” of TCP/IP traffic when network congestion occurs.

Traffic Shaping

Traffic shaping minimizes the occurrence of large bursts of traffic on the network. You can shape traffic by segmenting it, placing it into buffers, and delaying its entry into the network, thereby ensuring a more constant flow of traffic in the network.

In an LS2020 network, traffic shaping is performed at all packet interfaces (see the incoming packet traffic in Figure 4-6). Traffic entering ATM UNI interfaces, however, does not need to be shaped, because it obeys the traffic-policing parameters set for each virtual circuit.

Figure 4-6 Traffic Shaping