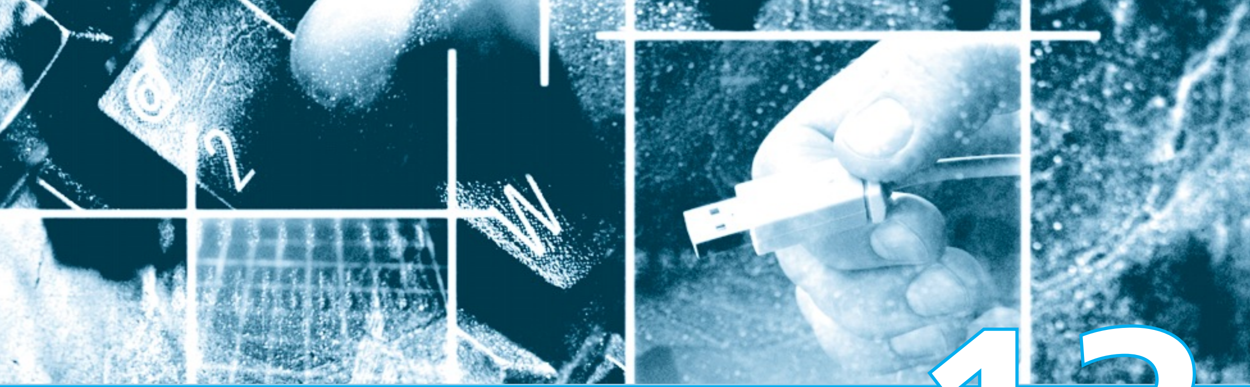


CHAPTER 13

Congestion Control in Data Networks

At St. Paul's a great throng crammed the platform. She saw a sea of faces, each stamped with a kind of purposeful, hungry urgency, a determination to get into this train. As before, when she was on the Northern Line, she thought there must be some rule, some operating law, that would stop more than a limited, controlled number getting in. Authority would appear and stop it.

—*King Solomon's Carpet*, Barbara Vine (Ruth Rendell)

KEY POINTS

- Congestion occurs when the number of packets being transmitted through a network begins to approach the packet-handling capacity of the network. The objective of congestion control is to maintain the number of packets within the network below the level at which performance falls off dramatically.
- The lack of flow control mechanisms built into the ATM and frame relay protocols makes congestion control difficult. A variety of techniques have been developed to cope with congestion and to give different quality-of-service guarantees to different types of traffic.
- ATM networks establish a traffic contract with each user that specifies the characteristics of the expected traffic and the type of service that the network will provide. The network implements congestion control techniques in such a way as to protect the network from congestion while meeting the traffic contracts.
- An ATM network monitors the cell flow from each incoming source and may discard or label for potential discard cells that exceed the agreed traffic contract. In addition, the network may shape the traffic coming from users by temporarily buffering cells to smooth out traffic flows.

A key design issue that must be confronted both with data networks, such as packet-switching, frame relay, and ATM networks, and also with internets, is that of congestion control. The phenomenon of congestion is a complex one, as is the subject of congestion control. In very general terms, congestion occurs when the number of packets[1] being transmitted through a network begins to approach the packet-handling capacity of the network. The objective of congestion control is to maintain the number of packets within the network below the level at which performance falls off dramatically.

To understand the issues involved in congestion control, we need to look at some results from queuing theory.[2] In essence, a data network or

[1] In this chapter we use the term packet in a broad sense, to include packets in a packet-switching network, frames in a frame relay network, cells in an ATM network, or IP datagrams in an internet.

[2] Appendix I provides an overview of queuing analysis.

internet is a network of queues. At each node (data network switch, internet router), there is a queue of packets for each outgoing channel. If the rate at which packets arrive and queue up exceeds the rate at which packets can be transmitted, the queue size grows without bound and the delay experienced by a packet goes to infinity. Even if the packet arrival rate is less than the packet transmission rate, queue length will grow dramatically as the arrival rate approaches the transmission rate. As a rule of thumb, when the line for which packets are queuing becomes more than 80% utilized, the queue length grows at an alarming rate. This growth in queue length means that the delay experienced by a packet at each node increases. Further, since the size of any queue is finite, as queue length grows, eventually the queue must overflow.

This chapter focuses on congestion control in switched data networks, including packet-switching, frame relay, and ATM networks. The principles examined here are also applicable to internetworks. In Part Five, we look at additional congestion control mechanisms in our discussion of internetwork operation and TCP congestion control.

13.1 EFFECTS OF CONGESTION

Consider the queuing situation at a single packet switch or router, such as is illustrated in Figure 13.1. Any given node has a number of I/O ports[3] attached to it: one or more to other nodes, and zero or more to end systems. On each port, packets arrive and depart. We can consider that there are two buffers, or queues, at each port, one to accept arriving packets, and one to hold packets that are waiting to depart. In practice, there might be two fixed-size buffers associated with each port, or there might be a pool of memory available for all buffering activities. In the latter case, we can think of each port having two variable-size buffers associated with it, subject to the constraint that the sum of all buffer sizes is a constant.

In any case, as packets arrive, they are stored in the input buffer of the corresponding port. The node examines each incoming packet, makes a routing decision, and then moves the packet to the appropriate output buffer. Packets queued for output are transmitted as rapidly as possible; this is, in effect, statistical time division multiplexing. If packets arrive too fast for the node to process them (make routing decisions) or faster than packets can be cleared from the outgoing buffers, then eventually packets will arrive for which no memory is available.

When such a saturation point is reached, one of two general strategies can be adopted. The first such strategy is to discard any incoming packet for which there is no available buffer space. The alternative is for the node that is experiencing these problems to exercise some sort of flow control over its neighbors so that the traffic flow remains manageable. But, as Figure 13.2 illustrates, each of a node's neighbors

[3] In the case of a switch of a packet-switching, frame relay, or ATM network, each I/O port connects to a transmission link that connects to another node or end system. In the case of a router of an internet, each I/O port connects to either a direct link to another node or to a subnetwork.

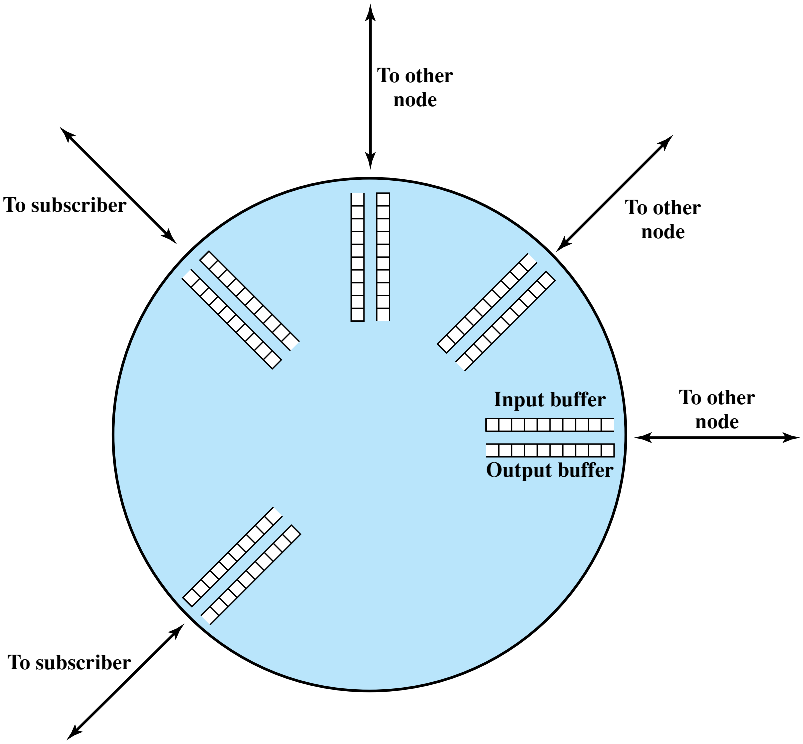


Figure 13.1 Input and Output Queues at Node

is also managing a number of queues. If node 6 restrains the flow of packets from node 5, this causes the output buffer in node 5 for the port to node 6 to fill up. Thus, congestion at one point in the network can quickly propagate throughout a region or the entire network. While flow control is indeed a powerful tool, we need to use it in such a way as to manage the traffic on the entire network.

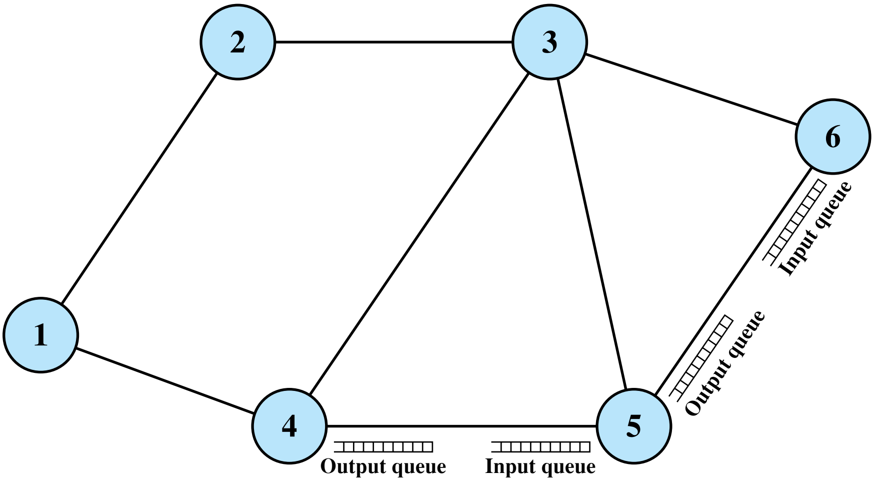


Figure 13.2 Interaction of Queues in a Data Network

Ideal Performance

Figure 13.3 suggests the ideal goal for network utilization. The top graph plots the steady-state total throughput (number of packets delivered to destination end systems) through the network as a function of the offered load (number of packets transmitted by source end systems), both normalized to the maximum theoretical throughput of the network. For example, if a network consists of a single node with two full-duplex 1-Mbps links, then the theoretical capacity of the network is 2 Mbps, consisting of a 1-Mbps flow in each direction. In the ideal case, the throughput of the network increases to accommodate load up to an offered load equal to the full capacity of the network; then normalized throughput remains at 1.0 at higher input loads. Note, however, what happens to the end-to-end delay experienced by the average packet even with this assumption of ideal performance. At negligible load, there is some small constant amount of delay that consists of the propagation delay through the network from source to destination plus processing delay at each node. As the load on the network increases, queuing delays at each node are added to this fixed amount of delay. When the load exceeds the network capacity, delays increase without bound.

Here is a simple intuitive explanation of why delay must go to infinity. Suppose that each node in the network is equipped with buffers of infinite size and suppose that the input load exceeds network capacity. Under ideal conditions, the

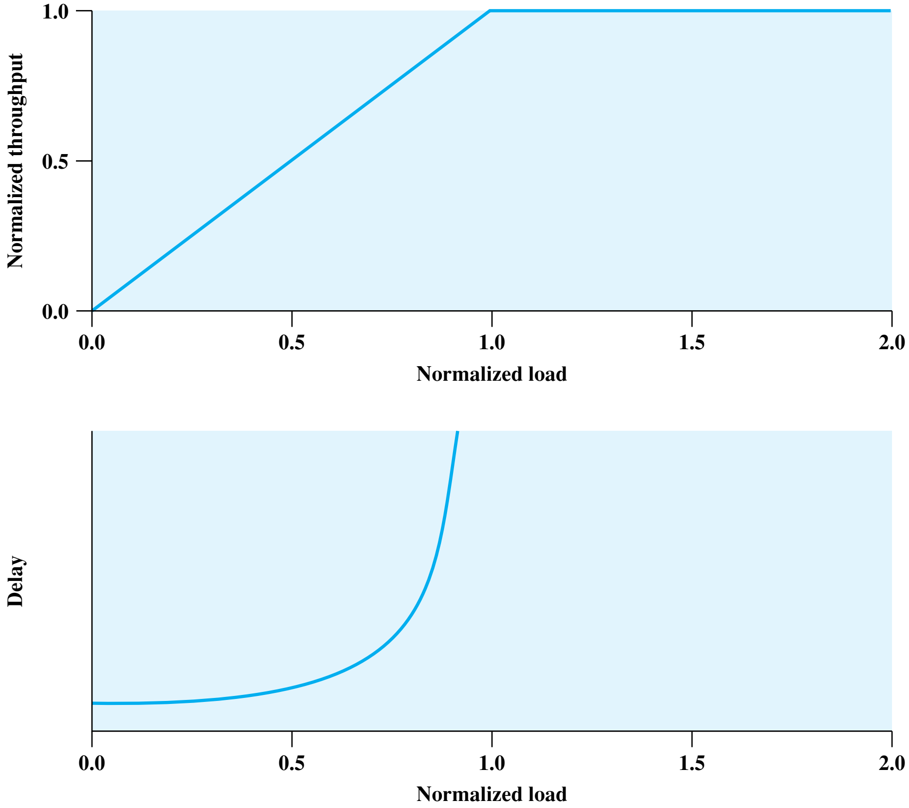


Figure 13.3 Ideal Network Utilization

network will continue to sustain a normalized throughput of 1.0. Therefore, the rate of packets leaving the network is 1.0. Because the rate of packets entering the network is greater than 1.0, internal queue sizes grow. In the steady state, with input greater than output, these queue sizes grow without bound and therefore queuing delays grow without bound.

It is important to grasp the meaning of Figure 13.3 before looking at real-world conditions. This figure represents the ideal, but unattainable, goal of all traffic and congestion control schemes. No scheme can exceed the performance depicted in Figure 13.3.

Practical Performance

The ideal case reflected in Figure 13.3 assumes infinite buffers and no overhead related to congestion control. In practice, buffers are finite, leading to buffer overflow, and attempts to control congestion consume network capacity in the exchange of control signals.

Let us consider what happens in a network with finite buffers if no attempt is made to control congestion or to restrain input from end systems. The details will, of course, differ depending on network configuration and on the statistics of the presented traffic. However, the graphs in Figure 13.4 depict the devastating outcome in general terms.

At light loads, throughput and hence network utilization increases as the offered load increases. As the load continues to increase, a point is reached (point A in the plot) beyond which the throughput of the network increases at a rate slower than the rate at which offered load is increased. This is due to network entry into a moderate congestion state. In this region, the network continues to cope with the load, although with increased delays. The departure of throughput from the ideal is accounted for by a number of factors. For one thing, the load is unlikely to be spread uniformly throughout the network. Therefore, while some nodes may experience moderate congestion, others may be experiencing severe congestion and may need to discard traffic. In addition, as the load increases, the network will attempt to balance the load by routing packets through areas of lower congestion. For the routing function to work, an increased number of routing messages must be exchanged between nodes to alert each other to areas of congestion; this overhead reduces the capacity available for data packets.

As the load on the network continues to increase, the queue lengths of the various nodes continue to grow. Eventually, a point is reached (point B in the plot) beyond which throughput actually drops with increased offered load. The reason for this is that the buffers at each node are of finite size. When the buffers at a node become full, the node must discard packets. Thus, the sources must retransmit the discarded packets in addition to new packets. This only exacerbates the situation: As more and more packets are retransmitted, the load on the system grows, and more buffers become saturated. While the system is trying desperately to clear the backlog, users are pumping old and new packets into the system. Even successfully delivered packets may be retransmitted because it takes too long, at a higher layer (e.g., transport layer), to acknowledge them: The sender assumes the packet did not get through and retransmits. Under these circumstances, the effective capacity of the system declines to zero.

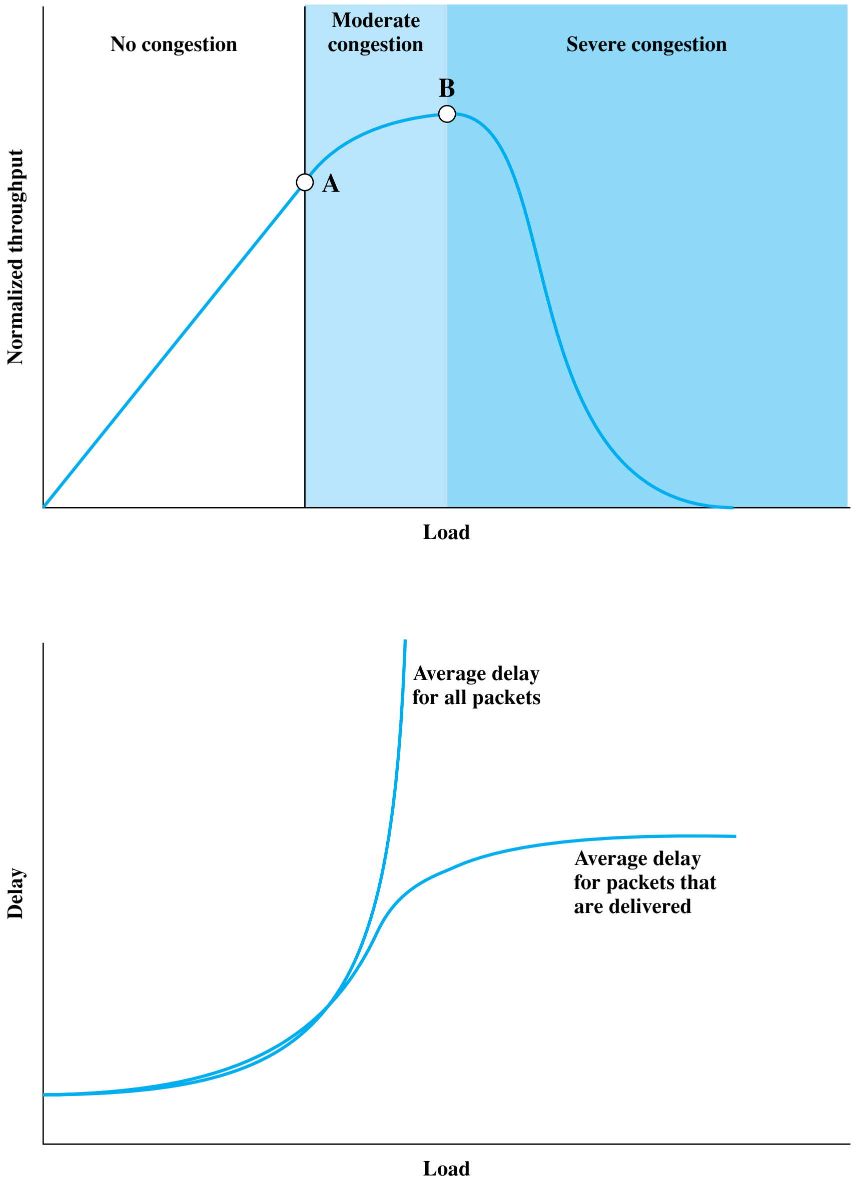


Figure 13.4 The Effects of Congestion

13.2 CONGESTION CONTROL

In this book, we discuss various techniques for controlling congestion in packet-switching, frame relay, and ATM networks, and in IP-based internets. To give context to this discussion, Figure 13.5 provides a general depiction of important congestion control techniques.

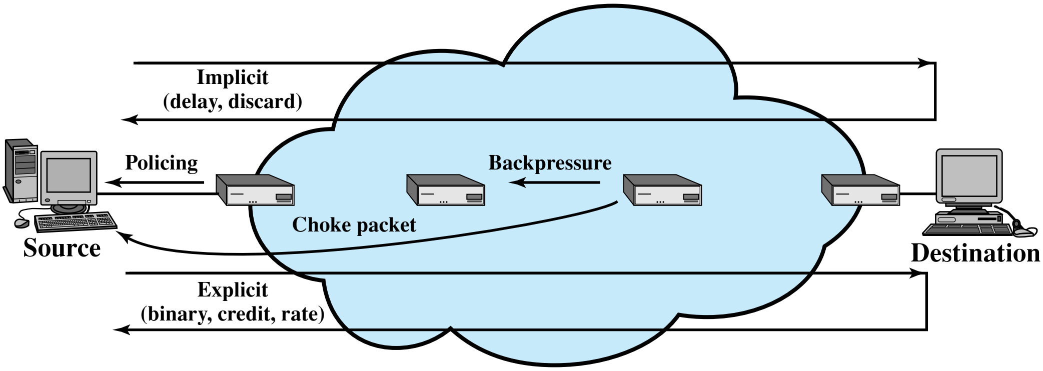


Figure 13.5 Mechanisms for Congestion Control

Backpressure

We have already made reference to backpressure as a technique for congestion control. This technique produces an effect similar to backpressure in fluids flowing down a pipe. When the end of a pipe is closed (or restricted), the fluid pressure backs up the pipe to the point of origin, where the flow is stopped (or slowed).

Backpressure can be exerted on the basis of links or logical connections (e.g., virtual circuits). Referring again to Figure 13.2, if node 6 becomes congested (buffers fill up), then node 6 can slow down or halt the flow of all packets from node 5 (or node 3, or both nodes 5 and 3). If this restriction persists, node 5 will need to slow down or halt traffic on its incoming links. This flow restriction propagates backward (against the flow of data traffic) to sources, which are restricted in the flow of new packets into the network.

Backpressure can be selectively applied to logical connections, so that the flow from one node to the next is only restricted or halted on some connections, generally the ones with the most traffic. In this case, the restriction propagates back along the connection to the source.

Backpressure is of limited utility. It can be used in a connection-oriented network that allows hop-by-hop (from one node to the next) flow control. X.25-based packet-switching networks typically provide this feature. However, neither frame relay nor ATM has any capability for restricting flow on a hop-by-hop basis. In the case of IP-based internets, there have traditionally been no built-in facilities for regulating the flow of data from one router to the next along a path through the internet. Recently, some flow-based schemes have been developed; this topic is introduced in Part Five.

Choke Packet

A choke packet is a control packet generated at a congested node and transmitted back to a source node to restrict traffic flow. An example of a choke packet is the ICMP (Internet Control Message Protocol) Source Quench packet. Either a router or a destination end system may send this message to a source end system, requesting that it reduce the rate at which it is sending traffic to the internet destination. On receipt of a source quench message, the source host should cut back the rate at which it is sending traffic to the specified destination until it no longer receives source quench messages. The source quench message can be used by a router or host that must discard IP datagrams because of a full buffer. In that case, the router or host will issue a source quench message for every datagram that it discards. In addition, a system may anticipate congestion and issue source quench messages when its buffers approach capacity. In that case, the datagram referred to in the source quench message may well be delivered. Thus, receipt of a source quench message does not imply delivery or nondelivery of the corresponding datagram.

The choke package is a relatively crude technique for controlling congestion. More sophisticated forms of explicit congestion signaling are discussed subsequently.

Implicit Congestion Signaling

When network congestion occurs, two things may happen: (1) The transmission delay for an individual packet from source to destination increases, so that it is noticeably longer than the fixed propagation delay, and (2) packets are discarded. If a source is able to detect increased delays and packet discards, then it has implicit evidence of network congestion. If all sources can detect congestion and, in response, reduce flow on the basis of congestion, then the network congestion will be relieved. Thus, congestion control on the basis of implicit signaling is the responsibility of end systems and does not require action on the part of network nodes.

Implicit signaling is an effective congestion control technique in connectionless, or datagram, configurations, such as datagram packet-switching networks and IP-based internets. In such cases, there are no logical connections through the internet on which flow can be regulated. However, between the two end systems, logical connections can be established at the TCP level. TCP includes mechanisms for acknowledging receipt of TCP segments and for regulating the flow of data between source and destination on a TCP connection. TCP congestion control techniques based on the ability to detect increased delay and segment loss are discussed in Chapter 20.

Implicit signaling can also be used in connection-oriented networks. For example, in frame relay networks, the LAPF control protocol, which is end to end, includes facilities similar to those of TCP for flow and error control. LAPF control is capable of detecting lost frames and adjusting the flow of data accordingly.

Explicit Congestion Signaling

It is desirable to use as much of the available capacity in a network as possible but still react to congestion in a controlled and fair manner. This is the purpose of explicit congestion avoidance techniques. In general terms, for explicit congestion avoidance, the network alerts end systems to growing congestion within the network and the end systems take steps to reduce the offered load to the network.

Typically, explicit congestion control techniques operate over connection-oriented networks and control the flow of packets over individual connections. Explicit congestion signaling approaches can work in one of two directions:

- **Backward:** Notifies the source that congestion avoidance procedures should be initiated where applicable for traffic in the opposite direction of the received notification. It indicates that the packets that the user transmits on this logical connection may encounter congested resources. Backward information is transmitted either by altering bits in a header of a data packet headed for the source to be controlled or by transmitting separate control packets to the source.

- **Forward:** Notifies the user that congestion avoidance procedures should be initiated where applicable for traffic in the same direction as the received notification. It indicates that this packet, on this logical connection, has encountered congested resources. Again, this information may be transmitted either as altered bits in data packets or in separate control packets. In some schemes, when a forward signal is received by an end system, it echoes the signal back along the logical connection to the source. In other schemes, the end system is expected to exercise flow control upon the source end system at a higher layer (e.g., TCP).

We can divide explicit congestion signaling approaches into three general categories:

- **Binary:** A bit is set in a data packet as it is forwarded by the congested node. When a source receives a binary indication of congestion on a logical connection, it may reduce its traffic flow.
- **Credit based:** These schemes are based on providing an explicit credit to a source over a logical connection. The credit indicates how many octets or how many packets the source may transmit. When the credit is exhausted, the source must await additional credit before sending additional data. Credit-based schemes are common for end-to-end flow control, in which a destination system uses credit to prevent the source from overflowing the destination buffers, but credit-based schemes have also been considered for congestion control.
- **Rate based:** These schemes are based on providing an explicit data rate limit to the source over a logical connection. The source may transmit data at a rate up to the set limit. To control congestion, any node along the path of the connection can reduce the data rate limit in a control message to the source.

13.3 TRAFFIC MANAGEMENT

There are a number of issues related to congestion control that might be included under the general category of traffic management. In its simplest form, congestion control is concerned with efficient use of a network at high load. The various mechanisms discussed in the previous section can be applied as the situation arises, without regard to the particular source or destination affected. When a node is saturated and must discard packets, it can apply some simple rule, such as discard the most recent arrival. However, other considerations can be used to refine the application of congestion control techniques and discard policy. We briefly introduce several of those areas here.

Fairness

As congestion develops, flows of packets between sources and destinations will experience increased delays and, with high congestion, packet losses. In the absence of other requirements, we would like to assure that the various flows suffer from congestion equally. Simply to discard on a last-in-first-discarded basis may not be fair. As an example of a technique that might promote fairness, a node can maintain a separate queue for each logical connection or for each source-destination pair. If all of the queue buffers are of equal length, then the queues with the highest traffic load will suffer discards more often, allowing lower-traffic connections a fair share of the capacity.

Quality of Service

We might wish to treat different traffic flows differently. For example, as [JAIN92] points out, some applications, such as voice and video, are delay sensitive but loss insensitive. Others, such as file transfer and electronic mail, are delay insensitive but loss sensitive. Still others, such as interactive graphics or interactive computing applications, are delay sensitive and loss sensitive. Also, different traffic flows have different priorities; for example, network management traffic, particularly during times of congestion or failure, is more important than application traffic.

It is particularly important during periods of congestion that traffic flows with different requirements be treated differently and provided a different quality of service (QoS). For example, a node might transmit higher-priority packets ahead of lower-priority packets in the same queue. Or a node might maintain different queues for different QoS levels and give preferential treatment to the higher levels.

Reservations

One way to avoid congestion and also to provide assured service to applications is to use a reservation scheme. Such a scheme is an integral part of ATM networks. When a logical connection is established, the network and the user enter into a traffic contract, which specifies a data rate and other characteristics of the traffic flow. The network agrees to give a defined QoS so long as the traffic flow is within contract parameters; excess traffic is either discarded or handled on a best-effort basis, subject to discard. If the current outstanding reservations are such that the network resources are inadequate to meet the new reservation, then the new reservation is denied. A similar type of scheme has now been developed for IP-based internets (RSVP, which is discussed in Chapter 19).

One aspect of a reservation scheme is traffic policing (Figure 13.5). A node in the network, typically the node to which the end system attaches, monitors the traffic flow and compares it to the traffic contract. Excess traffic is either discarded or marked to indicate that it is liable to discard or delay.

13.4 CONGESTION CONTROL IN PACKET-SWITCHING NETWORKS

A number of control mechanisms for congestion control in packet-switching networks have been suggested and tried. The following are examples:

1. Send a control packet from a congested node to some or all source nodes. This choke packet will have the effect of stopping or slowing the rate of transmission from sources and hence limit the total number of packets in the network. This approach requires additional traffic on the network during a period of congestion.

2. Rely on routing information. Routing algorithms, such as ARPANET's, provide link delay information to other nodes, which influences routing decisions. This information could also be used to influence the rate at which new packets are produced. Because these delays are being influenced by the routing decision, they may vary too rapidly to be used effectively for congestion control.
3. Make use of an end-to-end probe packet. Such a packet could be timestamped to measure the delay between two particular endpoints. This has the disadvantage of adding overhead to the network.
4. Allow packet-switching nodes to add congestion information to packets as they go by. There are two possible approaches here. A node could add such information to packets going in the direction opposite of the congestion. This information quickly reaches the source node, which can reduce the flow of packets into the network. Alternatively, a node could add such information to packets going in the same direction as the congestion. The destination either asks the source to adjust the load or returns the signal back to the source in the packets (or acknowledgments) going in the reverse direction.

13.5 FRAME RELAY CONGESTION CONTROL

I.370 defines the objectives for frame relay congestion control to be the following:

- Minimize frame discard.
- Maintain, with high probability and minimum variance, an agreed quality of service.
- Minimize the possibility that one end user can monopolize network resources at the expense of other end users.
- Be simple to implement, and place little overhead on either end user or network.
- Create minimal additional network traffic.
- Distribute network resources fairly among end users.
- Limit spread of congestion to other networks and elements within the network.
- Operate effectively regardless of the traffic flow in either direction between end users.
- Have minimum interaction or impact on other systems in the frame relaying network.
- Minimize the variance in quality of service delivered to individual frame relay connections during congestion (e.g., individual logical connections should not experience sudden degradation when congestion approaches or has occurred).

Congestion control is difficult for a frame relay network because of the limited tools available to the frame handlers (frame-switching nodes). The frame relay protocol has been streamlined to maximize throughput and efficiency. A consequence of this is that a frame handler cannot control the flow of frames coming from a subscriber or an adjacent frame handler using the typical sliding-window flow control protocol, such as is found in HDLC.

Table 13.1 Frame Relay Congestion Control Techniques

Technique	Type	Function	Key Elements
Discard control	Discard strategy	Provides guidance to network concerning which frames to discard	DE bit
Backward explicit Congestion Notification	Congestion avoidance	Provides guidance to end systems about congestion in network	BECN bit or CLLM message
Forward explicit Congestion Notification	Congestion avoidance	Provides guidance to end systems about congestion in network	FECN bit
Implicit congestion notification	Congestion recovery	End system infers congestion from frame loss	Sequence numbers in higher-layer PDU

Congestion control is the joint responsibility of the network and the end users. The network (i.e., the collection of frame handlers) is in the best position to monitor the degree of congestion, while the end users are in the best position to control congestion by limiting the flow of traffic.

Table 13.1 lists the congestion control techniques defined in the various ITU-T and ANSI documents. **Discard strategy** deals with the most fundamental response to congestion: When congestion becomes severe enough, the network is forced to discard frames. We would like to do this in a way that is fair to all users.

Congestion avoidance procedures are used at the onset of congestion to minimize the effect on the network. Thus, these procedures would be initiated at or prior to point A in Figure 13.4, to prevent congestion from progressing to point B. Near point A, there would be little evidence available to end users that congestion is increasing. Thus, there must be some **explicit signaling** mechanism from the network that will trigger the congestion avoidance.

Congestion recovery procedures are used to prevent network collapse in the face of severe congestion. These procedures are typically initiated when the network has begun to drop frames due to congestion. Such dropped frames will be reported by some higher layer of software (e.g., LAPF control protocol or TCP) and serve as an **implicit signaling** mechanism. Congestion recovery procedures operate around point B and within the region of severe congestion, as shown in Figure 13.4.

ITU-T and ANSI consider congestion avoidance with explicit signaling and congestion recovery with implicit signaling to be complementary forms of congestion control in the frame relaying bearer service.

Traffic Rate Management

As a last resort, a frame-relaying network must discard frames to cope with congestion. There is no getting around this fact. Because each frame handler in the network has finite memory available for queuing frames (Figure 13.2), it is possible for a queue to overflow, necessitating the discard of either the most recently arrived frame or some other frame.

The simplest way to cope with congestion is for the frame-relaying network to discard frames arbitrarily, with no regard to the source of a particular frame. In that case, because there is no reward for restraint, the best strategy for any individual end system is to transmit frames as rapidly as possible. This, of course, exacerbates the congestion problem.

To provide for a fairer allocation of resources, the frame relay bearer service includes the concept of a committed information rate (CIR). This is a rate, in bits per second, that the network agrees to support for a particular frame-mode connection. Any data transmitted in excess of the CIR are vulnerable to discard in the event of congestion. Despite the use of the term *committed*, there is no guarantee that even the CIR will be met. In cases of extreme congestion, the network may be forced to provide a service at less than the CIR for a given connection. However, when it comes time to discard frames, the network will choose to discard frames on connections that are exceeding their CIR before discarding frames that are within their CIR.

In theory, each frame-relaying node should manage its affairs so that the aggregate of CIRs of all the connections of all the end systems attached to the node does not exceed the capacity of the node. In addition, the aggregate of the CIRs should not exceed the physical data rate across the user-network interface, known as the access rate. The limitation imposed by access rate can be expressed as follows:

$$\sum_i \text{CIR}_{i,j} \leq \text{AccessRate}_j \qquad \textbf{(13.1)}$$

where

$\text{CIR}_{i,j}$ = Committed information rate for connection i on channel j

AccessRate_j = Data rate of user access channel j; a channel is a fixed-data-rate TDM channel between the user and the network

Considerations of node capacity may result in the selection of lower values for some of the CIRs.

For permanent frame relay connections, the CIR for each connection must be established at the time the connection is agreed between user and network. For switched connections, the CIR parameter is negotiated; this is done in the setup phase of the call control protocol.

The CIR provides a way of discriminating among frames in determining which frames to discard in the face of congestion. Discrimination is indicated by means of the discard eligibility (DE) bit in the LAPF frame (Figure 10.16). The frame handler to which the user's station attaches performs a metering function (Figure 13.6). If the user is sending data at less than the CIR, the incoming frame handler does not alter the DE bit. If the rate exceeds the CIR, the incoming frame handler will set the DE bit on the excess frames and then forward them; such frames may get through or may be discarded if congestion is encountered. Finally, a maximum rate is defined, such that any frames above the maximum are discarded at the entry frame handler.

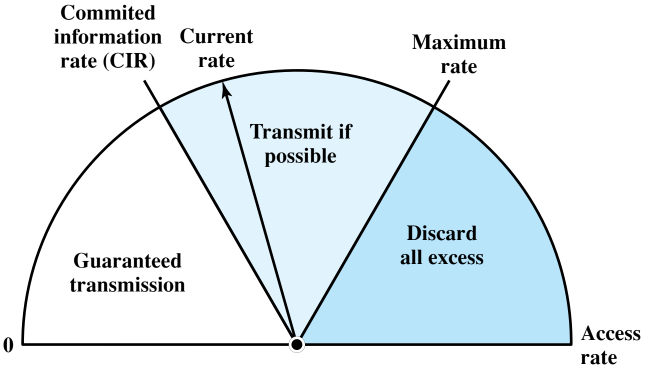


Figure 13.6 Operation of the CIR

The CIR, by itself, does not provide much flexibility in dealing with traffic rates. In practice, a frame handler measures traffic over each logical connection for a time interval specific to that connection and then makes a decision based on the amount of data received during that interval. Two additional parameters, assigned on permanent connections and negotiated on switched connections, are needed. They are

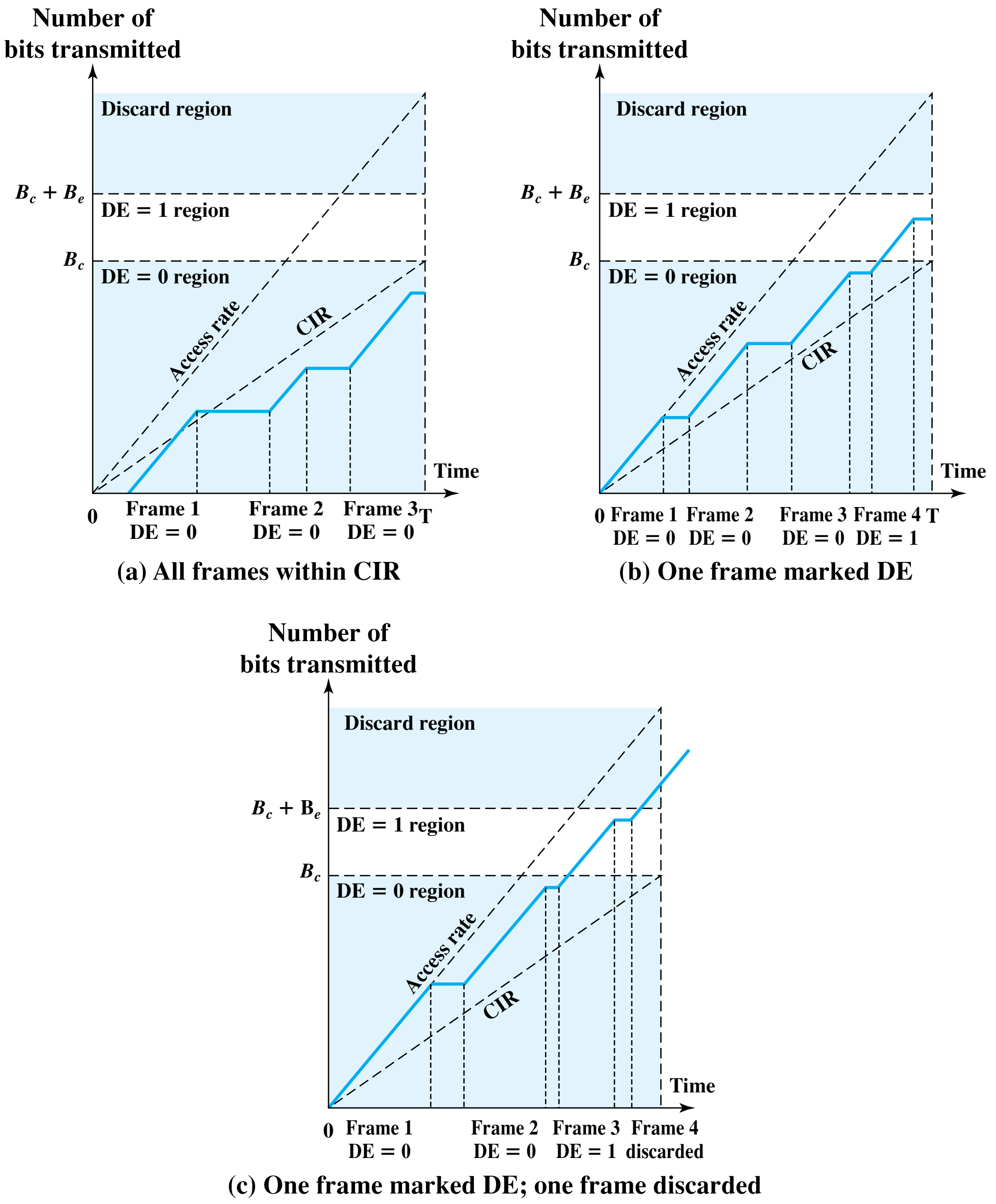
- **Committed burst size (B_c):** The maximum amount data that the network agrees to transfer, under normal conditions, over a measurement interval *T*. These data may or may not be contiguous (i.e., they may appear in one frame or in several frames).
- **Excess burst size (B_e):** The maximum amount of data in excess of B_c that the network will attempt to transfer, under normal conditions, over a measurement interval *T*. These data are uncommitted in the sense that the network does not commit to delivery under normal conditions. Put another way, the data that represent B_e are delivered with lower probability than the data within B_c.

The quantities B_c and CIR are related. Because B_c is the amount of committed data that may be transmitted by the user over a time *T*, and CIR is the rate at which committed data may be transmitted, we must have

$$T = \frac{B_c}{CIR} \tag{13.2}$$

Figure 13.7, based on a figure in ITU-T Recommendation I.370, illustrates the relationship among these parameters. On each graph, the solid line plots the cumulative number of information bits transferred over a given connection since time $T = 0$. The dashed line labeled Access Rate represents the data rate over the channel containing this connection. The dashed line labeled CIR represents the committed information rate over the measurement interval *T*. Note that when a frame is being transmitted, the solid line is parallel to the Access Rate line; when a frame is transmitted on a channel, that channel is dedicated to the transmission of that frame. When no frame is being transmitted, the solid line is horizontal.

Figure 13.7a shows an example in which three frames are transmitted within the measurement interval and the total number of bits in the three frames is less than B_c. Note that during the transmission of the first frame, the actual transmission rate



(a) All frames within CIR

(b) One frame marked DE

(c) One frame marked DE; one frame discarded

Figure 13.7 Illustration of Relationships among Congestion Parameters

temporarily exceeds the CIR. This is of no consequence because the frame handler is only concerned with the cumulative number of bits transmitted over the entire interval. In Figure 13.7b, the last frame transmitted during the interval causes the cumulative number of bits transmitted to exceed B_c. Accordingly, the DE bit of that frame is set by the frame handler. In Figure 13.7c, the third frame exceeds B_c and so is labeled for potential discard. The fourth frame exceeds $B_c + B_e$ and is discarded.

Congestion Avoidance with Explicit Signaling

It is desirable to use as much of the available capacity in a frame relay network as possible but still react to congestion in a controlled and fair manner. This is the purpose of explicit congestion avoidance techniques. In general terms, for explicit congestion avoidance, the network alerts end systems to growing congestion within the network and the end systems take steps to reduce the offered load to the network.

As the standards for explicit congestion avoidance were being developed, two general strategies were considered [BERG91]. One group believed that congestion always occurred slowly and almost always in the network egress nodes. Another group had seen cases in which congestion grew very quickly in the internal nodes and required quick decisive action to prevent network congestion. We will see that these two approaches are reflected in the forward and backward explicit congestion avoidance techniques, respectively.

For explicit signaling, two bits in the address field of each frame are provided. Either bit may be set by any frame handler that detects congestion. If a frame handler receives a frame in which one or both of these bits are set, it must not clear the bits before forwarding the frame. Thus, the bits constitute signals from the network to the end user. The two bits are

- **Backward explicit congestion notification (BECN):** Notifies the user that congestion avoidance procedures should be initiated where applicable for traffic in the opposite direction of the received frame. It indicates that the frames that the user transmits on this logical connection may encounter congested resources.
- **Forward explicit congestion notification (FECN):** Notifies the user that congestion avoidance procedures should be initiated where applicable for traffic in the same direction as the received frame. It indicates that this frame, on this logical connection, has encountered congested resources.

Let us consider how these bits are used by the network and the user. First, for the **network response**, it is necessary for each frame handler to monitor its queuing behavior. If queue lengths begin to grow to a dangerous level, then either FECN or BECN bits or a combination should be set to try to reduce the flow of frames through that frame handler. The choice of FECN or BECN may be determined by whether the end users on a given logical connection are prepared to respond to one or the other of these bits. This may be determined at configuration time. In any case, the frame handler has some choice as to which logical connections should be alerted to congestion. If congestion is becoming quite serious, all logical connections through a frame handler might be notified. In the early stages of congestion, the frame handler might just notify users for those connections that are generating the most traffic.

The **user response** is determined by the receipt of BECN or FECN signals. The simplest procedure is the response to a BECN signal: The user simply reduces the rate at which frames are transmitted until the signal ceases. The response to an FECN is more complex, as it requires the user to notify its peer user of this connection to restrict its flow of frames. The core functions used in the frame relay protocol do not support this notification; therefore, it must be done at a higher layer, such as the transport layer. The flow control could also be accomplished by the LAPF control protocol or some other link control protocol implemented above the frame relay sublayer. The LAPF control protocol is particularly useful because it includes an enhancement to LAPD that permits the user to adjust window size.

13.6 ATM TRAFFIC MANAGEMENT

Because of their high speed and small cell size, ATM networks present difficulties in effectively controlling congestion not found in other types of data networks. The complexity of the problem is compounded by the limited number of overhead bits available for exerting control over the flow of user cells. This area is currently the subject of intense research, and approaches to traffic and congestion control are still evolving. ITU-T has defined a restricted initial set of traffic and congestion control capabilities aiming at simple mechanisms and realistic network efficiency; these are specified in I.371. The ATM Forum has published a somewhat more advanced version of this set in its Traffic Management Specification 4.0. This section focuses on the ATM Forum specifications.

We begin with an overview of the congestion problem and the framework adopted by ITU-T and the ATM Forum. We then discuss some of the specific techniques that have been developed for traffic management and congestion control.

Requirements for ATM Traffic and Congestion Control

Both the types of traffic patterns imposed on ATM networks and the transmission characteristics of those networks differ markedly from those of other switching networks. Most packet-switching and frame relay networks carry non-real-time data traffic. Typically, the traffic on individual virtual circuits or frame relay connections is bursty in nature, and the receiving system expects to receive incoming traffic on each connection in a bursty fashion. As a result,

- The network does not need to replicate the exact timing pattern of incoming traffic at the re exit node.
- Therefore, simple statistical multiplexing can be used to accommodate multiple logical connections over the physical interface between user and network. The average data rate required by each connection is less than the burst rate for that connection, and the user-network interface (UNI) need only be designed for a capacity somewhat greater than the sum of the average data rates for all connections.

A number of tools are available for control of congestion in packet-switched and frame relay networks, some of which are discussed elsewhere in this chapter. These types of congestion control schemes are inadequate for ATM networks. [GERS91] cites the following reasons:

- The majority of traffic is not amenable to flow control. For example, voice and video traffic sources cannot stop generating cells even when the network is congested.
- Feedback is slow due to the drastically reduced cell transmission time compared to propagation delays across the network.

- ATM networks typically support a wide range of applications requiring capacity ranging from a few kbps to several hundred Mbps. Relatively simpleminded congestion control schemes generally end up penalizing one end or the other of that spectrum.
- Applications on ATM networks may generate very different traffic patterns (e.g., constant bit rate versus variable bit rate sources). Again, it is difficult for conventional congestion control techniques to handle fairly such variety.
- Different applications on ATM networks require different network services (e.g., delay-sensitive service for voice and video, and loss-sensitive service for data).
- The very high speeds in switching and transmission make ATM networks more volatile in terms of congestion and traffic control. A scheme that relies heavily on reacting to changing conditions will produce extreme and wasteful fluctuations in routing policy and flow control.

Two key performance issues that relate to the preceding points are latency/speed effects and cell delay variation, topics to which we now turn.

Latency/Speed Effects

Consider the transfer of ATM cells over a network at a data rate of 150 Mbps. At that rate, it takes $(53 \times 8\text{ bits})/(150 \times 10^6\text{ bps}) \approx 2.8 \times 10^{-6}$ seconds to insert a single cell onto the network. The time it takes to transfer the cell from the source to the destination user will depend on the number of intermediate ATM switches, the switching time at each switch, and the propagation time along all links in the path from source to destination. For simplicity, ignore ATM switching delays and assume propagation at the two-thirds the speed of light. Then, if source and destination are on opposite coasts of the United States, the round-trip propagation delay is about 48×10^{-3} seconds.

With these conditions in place, suppose that source A is performing a long file transfer to destination B and that implicit congestion control is being used (i.e., there are no explicit congestion notifications; the source deduces the presence of congestion by the loss of data). If the network drops a cell due to congestion, B can return a reject message to A, which must then retransmit the dropped cell and possibly all subsequent cells. But by the time the notification gets back to A, it has transmitted an additional N cells, where

$$N = \frac{48 \times 10^{-3}\text{ seconds}}{2.8 \times 10^{-6}\text{ seconds/cell}} = 1.7 \times 10^4\text{ cells} = 7.2 \times 10^6\text{ bits}$$

Over 7 megabits have been transmitted before A can react to the congestion indication.

This calculation helps to explain why the techniques that are satisfactory for more traditional networks break down when dealing with ATM WANs.

Cell Delay Variation

For an ATM network, voice and video signals can be digitized and transmitted as a stream of cells. A key requirement, especially for voice, is that the delay across the network be short. Generally, this will be the case for ATM networks. As we have discussed, ATM is designed to minimize the processing and transmission overhead internal to the network so that very fast cell switching and routing is possible.

There is another important requirement that to some extent conflicts with the preceding requirement, namely that the rate of delivery of cells to the destination user must be constant. It is inevitable that there will be some variability in the rate of delivery of cells due both to effects within the network and at the source UNI; we summarize these effects presently. First, let us consider how the destination user might cope with variations in the delay of cells as they transit from source user to destination user.

A general procedure for achieving a constant bit rate (CBR) is illustrated in Figure 13.8. Let $D(i)$ represent the end-to-end delay experienced by the ith cell. The destination system does not know the exact amount of this delay: there is no timestamp information associated with each cell and, even if there were, it is impossible to keep source and destination clocks perfectly synchronized. When the first cell on a connection arrives at time t_0, the target user delays the cell an additional amount $V(0)$ prior to delivery to the application. $V(0)$ is an estimate of the amount of cell delay variation that this application can tolerate and that is likely to be produced by the network.

Subsequent cells are delayed so that they are delivered to the user at a constant rate of R cells per second. The time between delivery of cells to the target application (time between the start of delivery of one cell and the start of delivery of the next cell) is therefore $\delta = 1/R$. To achieve a constant rate, the next cell is delayed a variable amount $V(1)$ to satisfy

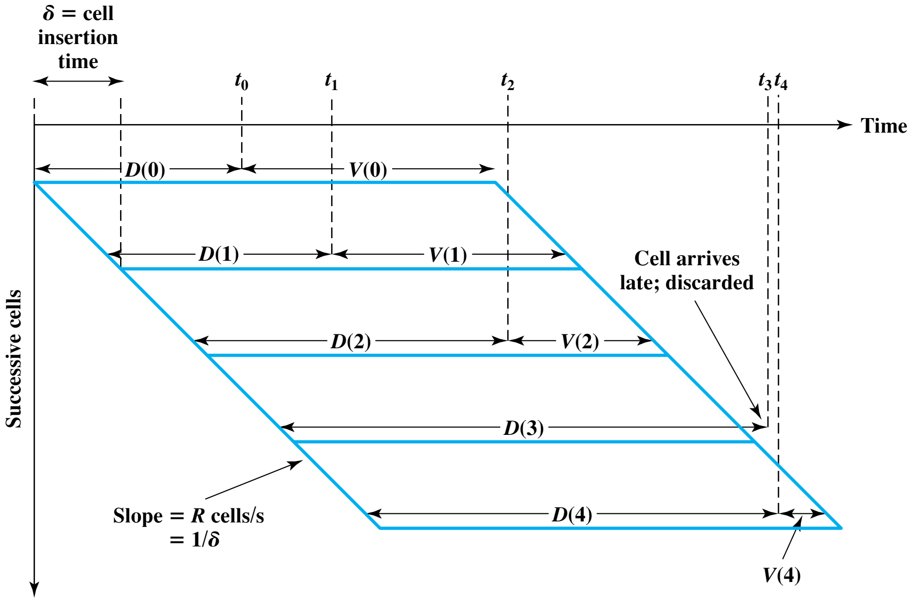


Figure 13.8 Time Reassembly of CBR Cells

$$t_1 + V(1) = t_0 + V(0) + \delta$$

So

$$V(1) = V(0) - [t_1 - (t_0 + \delta)]$$

In general,

$$V(i) = V(0) - [t_i - (t_0 + i \times \delta)]$$

which can also be expressed as

$$V(i) = V(i-1) - [t_i - (t_{i-1} + \delta)]$$

If the computed value of $V(i)$ is negative, then that cell is discarded. The result is that data is delivered to the higher layer at a constant bit rate, with occasional gaps due to dropped cells.

The amount of the initial delay $V(0)$, which is also the average delay applied to all incoming cells, is a function of the anticipated cell delay variation. To minimize this delay, a subscriber will therefore request a minimal cell delay variation from the network provider. This leads to a tradeoff: Cell delay variation can be reduced by increasing the data rate at the UNI relative to the load and by increasing resources within the network.

Network Contribution to Cell Delay Variation One component of cell delay variation is due to events within the network. For packet-switching networks, packet delay variation can be considerable due to queuing effects at each of the intermediate switching nodes and the processing time required to analyze packet headers and perform routing. To a much lesser extent, this is also true of frame delay variation in frame relay networks. In the case of ATM networks, cell delay variations due to network effects are likely to be even less than for frame relay. The principal reasons for this are the following:

- The ATM protocol is designed to minimize processing overhead at intermediate switching nodes. The cells are fixed size with fixed header formats, and there is no flow control or error control processing required.
- To accommodate the high speeds of ATM networks, ATM switches have had to be designed to provide extremely high throughput. Thus, the processing time for an individual cell at a node is negligible.

The only factor that could lead to noticeable cell delay variation within the network is congestion. If the network begins to become congested, either cells must be discarded or there will be a buildup of queuing delays at affected switches. Thus, it is important that the total load accepted by the network at any time not be such as to cause congestion.

Cell Delay Variation at the UNI Even if an application generates data for transmission at a constant bit rate, cell delay variation can occur at the source due to the processing that takes place at the three layers of the ATM model.

Figure 13.9 illustrates the potential causes of cell delay variation. In this example, ATM connections A and B support user data rates of X and Y Mbps, respectively $(X > Y)$. At the AAL level, data are segmented into 48-octet blocks. Note that on a

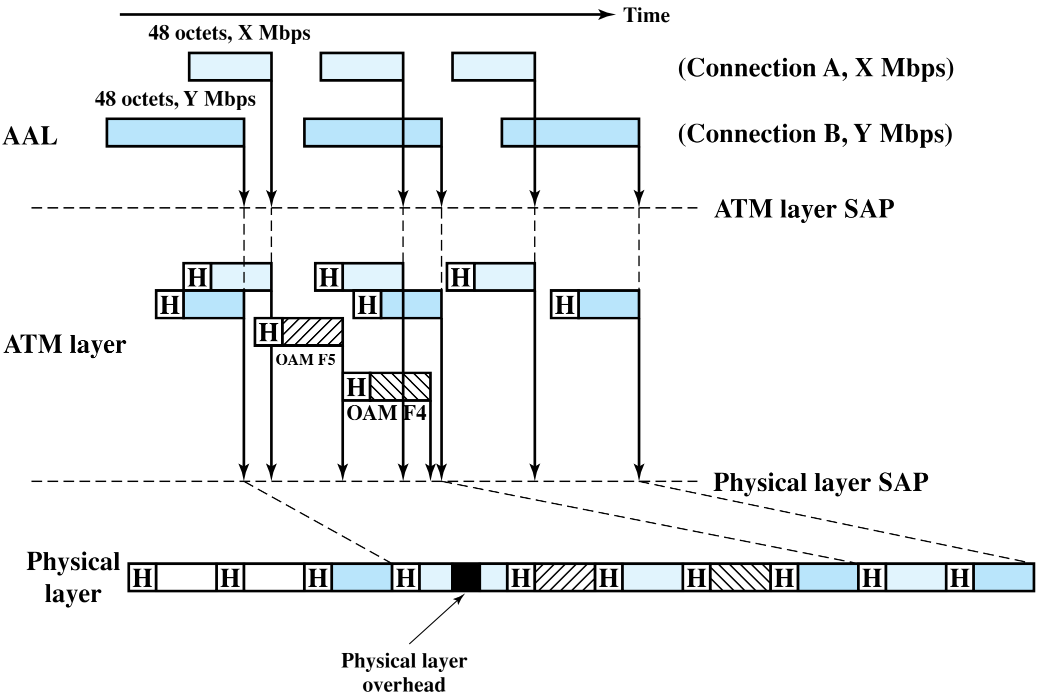


Figure 13.9 Origins of Cell Delay Variation (I.371)

time diagram, the blocks appear to be of different sizes for the two connections; specifically, the time required to generate a 48-octet block of data, in microseconds, is

$$\text{Connection A: } \frac{48 \times 8}{X}$$

$$\text{Connection B: } \frac{48 \times 8}{Y}$$

The ATM layer encapsulates each segment into a 53-octet cell. These cells must be interleaved and delivered to the physical layer to be transmitted at the data rate of the physical link. Delay is introduced into this interleaving process: If two cells from different connections arrive at the ATM layer at overlapping times, one of the cells must be delayed by the amount of the overlap. In addition, the ATM layer is generating OAM (operation and maintenance) cells that must also be interleaved with user cells.

At the physical layer, there is opportunity for the introduction of further cell delays. For example, if cells are transmitted in SDH (synchronous digital hierarchy) frames, overhead bits for those frames will be inserted onto the physical link, delaying bits from the ATM layer.

None of the delays just listed can be predicted in any detail, and none follow any repetitive pattern. Accordingly, there is a random element to the time interval between reception of data at the ATM layer from the AAL and the transmission of that data in a cell across the UNI.

Traffic and Congestion Control Framework

I.371 lists the following objectives of ATM layer traffic and congestion control:

- ATM layer traffic and congestion control should support a set of ATM layer QoS classes sufficient for all foreseeable network services; the specification of these QoS classes should be consistent with network performance parameters currently under study.
- ATM layer traffic and congestion control should not rely on AAL protocols that are network service specific, nor on higher-layer protocols that are application specific. Protocol layers above the ATM layer may make use of information provided by the ATM layer to improve the utility those protocols can derive from the network.
- The design of an optimum set of ATM layer traffic controls and congestion controls should minimize network and end-system complexity while maximizing network utilization.

To meet these objectives, ITU-T and the ATM Forum have defined a collection of traffic and congestion control functions that operate across a spectrum of timing intervals. Table 13.2 lists these functions with respect to the response times within which they operate. Four levels of timing are considered:

- **Cell insertion time:** Functions at this level react immediately to cells as they are transmitted.
- **Round-trip propagation time:** At this level, the network responds within the lifetime of a cell in the network and may provide feedback indications to the source.
- **Connection duration:** At this level, the network determines whether a new connection at a given QoS can be accommodated and what performance levels will be agreed to.
- **Long term:** These are controls that affect more than one ATM connection and are established for long-term use.

The essence of the traffic control strategy is based on (1) determining whether a given new ATM connection can be accommodated and (2) agreeing with the subscriber

Table 13.2 Traffic Control and Congestion Control Functions

Response Time	Traffic Control Functions	Congestion Control Functions
Long Term	• Resource management using virtual paths	
Connection Duration	• Connection admission control (CAC)	
Round-Trip Propagation Time	• Fast resource management indication (EFCI)	• Explicit forward congestion • ABR flow control
Cell Insertion Time	• Usage parameter control (UPC) • Priority control • Traffic shaping	• Selective cell discard

on the performance parameters that will be supported. In effect, the subscriber and the network enter into a traffic contract: the network agrees to support traffic at a certain level of performance on this connection, and the subscriber agrees not to exceed traffic parameter limits. Traffic control functions are concerned with establishing these traffic parameters and enforcing them. Thus, they are concerned with congestion avoidance. If traffic control fails in certain instances, then congestion may occur. At this point, congestion control functions are invoked to respond to and recover from the congestion.

Traffic Management and Congestion Control Techniques

ITU-T and the ATM Forum have defined a range of traffic management functions to maintain the quality of service (QoS) of ATM connections. ATM traffic management function refers to the set of actions taken by the network to avoid congestion conditions or to minimize congestion effects. In this subsection, we highlight the following techniques:

- Resource management using virtual paths
- Connection admission control
- Usage parameter control
- Selective cell discard
- Traffic shaping

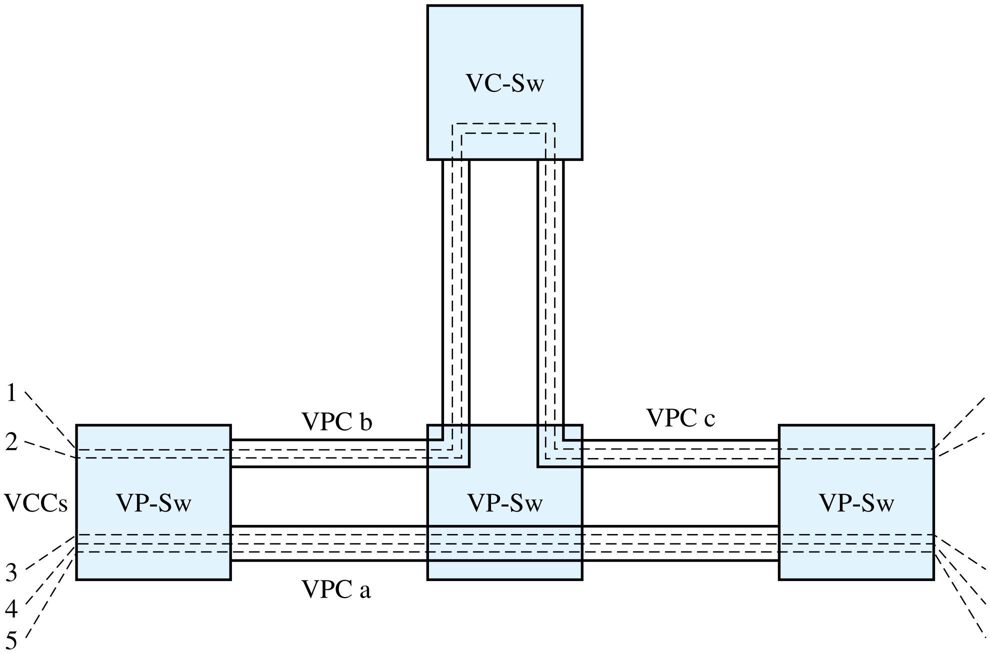
Resource Management Using Virtual Paths

The essential concept behind network resource management is to allocate network resources in such a way as to separate traffic flows according to service characteristics. So far, the only specific traffic control function based on network resource management defined by the ATM Forum deals with the use of virtual paths.

As discussed in Chapter 11, a virtual path connection (VPC) provides a convenient means of grouping similar virtual channel connections (VCCs). The network provides aggregate capacity and performance characteristics on the virtual path, and these are shared by the virtual connections. There are three cases to consider:

- **User-to-user application:** The VPC extends between a pair of UNIs. In this case the network has no knowledge of the QoS of the individual VCCs within a VPC. It is the user's responsibility to assure that the aggregate demand from the VCCs can be accommodated by the VPC.
- **User-to-network application:** The VPC extends between a UNI and a network node. In this case, the network is aware of the QoS of the VCCs within the VPC and has to accommodate them.
- **Network-to-network application:** The VPC extends between two network nodes. Again, in this case, the network is aware of the QoS of the VCCs within the VPC and has to accommodate them.

The QoS parameters that are of primary concern for network resource management are cell loss ratio, cell transfer delay, and cell delay variation, all of which are affected by the amount of resources devoted to the VPC by the network. If a VCC extends through multiple VPCs, then the performance on that VCC depends on the performances of the consecutive VPCs and on how the connection is handled at any



VPC	=	Virtual path connection
VCC	=	Virtual channel connection
VP-Sw	=	Virtual path switching function
VC-Sw	=	Virtual channel switching function

Figure 13.10 Configuration of VCCs and VPCs

node that performs VCC-related functions. Such a node may be a switch, concentrator, or other network equipment. The performance of each VPC depends on the capacity of that VPC and the traffic characteristics of the VCCs contained within the VPC. The performance of each VCC-related function depends on the switching/processing speed at the node and on the relative priority with which various cells are handled.

Figure 13.10 gives an example. VCCs 1 and 2 experience a performance that depends on VPCs b and c and on how these VCCs are handled by the intermediate nodes. This may differ from the performance experienced by VCCs 3, 4, and 5.

There are a number of alternatives for the way in which VCCs are grouped and the type of performance they experience. If all of the VCCs within a VPC are handled similarly, then they should experience similar expected network performance, in terms of cell loss ratio, cell transfer delay, and cell delay variation. Alternatively, when different VCCs within the same VPC require different QoS, the VPC performance objective agreed by network and subscriber should be set suitably for the most demanding VCC requirement.

In either case, with multiple VCCs within the same VPC, the network has two general options for allocating capacity to the VPC:

- **Aggregate peak demand:** The network may set the capacity (data rate) of the VPC equal to the total of the peak data rates of all of the VCCs within the VPC. The advantage of this approach is that each VCC can be given a QoS that accommodates its peak demand. The disadvantage is that most of the time, the VPC capacity will not be fully utilized and therefore the network will have underutilized resources.

- **Statistical multiplexing:** If the network sets the capacity of the VPC to be greater than or equal to the average data rates of all the VCCs but less than the aggregate peak demand, then a statistical multiplexing service is supplied. With statistical multiplexing, VCCs experience greater cell delay variation and greater cell transfer delay. Depending on the size of buffers used to queue cells for transmission, VCCs may also experience greater cell loss ratio. This approach has the advantage of more efficient utilization of capacity and is attractive if the VCCs can tolerate the lower QoS.

When statistical multiplexing is used, it is preferable to group VCCs into VPCs on the basis of similar traffic characteristics and similar QoS requirements. If dissimilar VCCs share the same VPC and statistical multiplexing is used, it is difficult to provide fair access to both high-demand and low-demand traffic streams.

Connection Admission Control

Connection admission control is the first line of defense for the network in protecting itself from excessive loads. In essence, when a user requests a new VPC or VCC, the user must specify (implicitly or explicitly) the traffic characteristics in both directions for that connection. The user selects traffic characteristics by selecting a QoS from among the QoS classes that the network provides. The network accepts the connection only if it can commit the resources necessary to support that traffic level while at the same time maintaining the agreed QoS of existing connections. By accepting the connection, the network forms a *traffic contract* with the user. Once the connection is accepted, the network continues to provide the agreed QoS as long as the user complies with the traffic contract.

The traffic contract may consist of the four parameters defined in Table 13.3: peak cell rate (PCR), cell delay variation (CDV), sustainable cell rate (SCR), and burst tolerance. Only the first two parameters are relevant for a constant-bit-rate (CBR) source; all four parameters may be used for variable-bit-rate (VBR) sources.

As the name suggests, the peak cell rate is the maximum rate at which cells are generated by the source on this connection. However, we need to take into account the cell delay variation. Although a source may be generating cells at a constant peak rate, cell delay variations introduced by various factors (see Figure 13.9) will affect the timing, causing cells to clump up and gaps to occur. Thus, a source may temporarily exceed the peak cell rate due to clumping. For the network to properly allocate resources to this connection, it must know not only the peak cell rate but also the CDV.

The exact relationship between peak cell rate and CDV depends on the operational definitions of these two terms. The standards provide these definitions in terms of a cell rate algorithm. Because this algorithm can be used for usage parameter control, we defer a discussion until the next subsection.

The PCR and CDV must be specified for every connection. As an option for variable-bit rate sources, the user may also specify a sustainable cell rate and burst tolerance. These parameters are analogous to PCR and CDV, respectively, but apply to an average rate of cell generation rather than a peak rate. The user can describe the future flow of cells in greater detail by using the SCR and burst tolerance as well as the PCR and CDV. With this additional information, the network may be able to utilize the

Table 13.3 Traffic Parameters Used in Defining VCC/VPC Quality of Service

Parameter	**Description**	**Traffic Type**
Peak Cell Rate (PCR)	An upper bound on the traffic that can be submitted on an ATM connection.	CBR, VBR
Cell Delay Variation (CDV)	An upper bound on the variability in the pattern of cell arrivals observed at a single measurement point with reference to the peak cell rate.	CBR, VBR
Sustainable Cell Rate (SCR)	An upper bound on the average rate of an ATM connection, calculated over the duration of the connection.	VBR
Burst Tolerance	An upper bound on the variability in the pattern of cell arrivals observed at a single measurement point with reference to the sustainable cell rate.	VBR

CBR = constant bit rate
VBR = variable bit rate

network resources more efficiently. For example, if a number of VCCs are statistically multiplexed over a VPC, knowledge of both average and peak cell rates enables the network to allocate buffers of sufficient size to handle the traffic efficiently without cell loss.

For a given connection (VPC or VCC) the four traffic parameters may be specified in several ways, as illustrated in Table 13.4. Parameter values may be implicitly defined by default rules set by the network operator. In this case, all connections are assigned the same values, or all connections of a given class are assigned the same values for that class. The network operator may also associate parameter values with a given subscriber and assign these at the time of subscription. Finally, parameter values tailored to a particular connection may be assigned at connection

Table 13.4 Procedures Used to Set Values of Traffic Contract Parameters

	Explicitly Specified Parameters		**Implicitly Specified Parameters**
	Parameter Values Set at Connection-Setup Time	**Parameter Values Specified at Subscription Time**	**Parameter Values Set Using Default Rules**
	Requested by User/NMS	**Assigned by Network Operator**	
SVC	signaling	by subscription	network-operator default rules
PVC	NMS	by subscription	network-operator default rules

SVC = switched virtual connection
PVC = permanent virtual connection
NMS = network management system

time. In the case of a permanent virtual connection, these values are assigned by the network when the connection is set up. For a switched virtual connection, the parameters are negotiated between the user and the network via a signaling protocol.

Another aspect of quality of service that may be requested or assigned for a connection is cell loss priority. A user may request two levels of cell loss priority for an ATM connection; the priority of an individual cell is indicated by the user through the CLP bit in the cell header (Figure 11.4). When two priority levels are used, the traffic parameters for both cell flows must be specified. Typically, this is done by specifying a set of traffic parameters for high-priority traffic ($\mathrm{CLP} = 0$) and a set of traffic parameters for all traffic ($\mathrm{CLP} = 0$ or 1). Based on this breakdown, the network may be able to allocate resources more efficiently.

Usage Parameter Control Once a connection has been accepted by the connection admission control function, the usage parameter control (UPC) function of the network monitors the connection to determine whether the traffic conforms to the traffic contract. The main purpose of usage parameter control is to protect network resources from an overload on one connection that would adversely affect the QoS on other connections by detecting violations of assigned parameters and taking appropriate actions. Usage parameter control can be done at both the virtual path and virtual channel levels. Of these, the more important is VPC-level control, because network resources are, in general, initially allocated on the basis of virtual paths, with the virtual path capacity shared among the member virtual channels.

There are two separate functions encompassed by usage parameter control:

- Control of peak cell rate and the associated cell delay variation (CDV)
- Control of sustainable cell rate and the associated burst tolerance

Let us first consider the peak cell rate and the associated cell delay variation. In simple terms, a traffic flow is compliant if the peak rate of cell transmission does not exceed the agreed peak cell rate, subject to the possibility of cell delay variation within the agreed bound. I.371 defines an algorithm, the peak cell rate algorithm, that monitors compliance. The algorithm operates on the basis of two parameters: a peak cell rate R and a CDV tolerance limit of τ. Then $T = 1/R$ is the interarrival time between cells if there were no CDV. With CDV, T is the average interarrival time at the peak rate. The algorithm has been defined to monitor the rate at which cells arrive and to assure that the interarrival time is not too short to cause the flow to exceed the peak cell rate by an amount greater than the tolerance limit.

The same algorithm, with different parameters, can be used to monitor the sustainable cell rate and the associated burst tolerance. In this case, the parameters are the sustainable cell rate R_s and a burst tolerance τ_s.

The cell rate algorithm is rather complex; details can be found in [STAL99]. The algorithm simply defines a way to monitor compliance with the traffic contract. To perform usage parameter control, the network must act on the results of the algorithm. The simplest strategy is that compliant cells are passed along and noncompliant cells are discarded at the point of the UPC function.

At the network's option, cell tagging may also be used for noncompliant cells. In this case, a noncompliant cell may be tagged with $\mathrm{CLP} = 1$ (low priority) and passed. Such cells are then subject to discard at a later point in the network, should congestion be encountered.

If the user has negotiated two levels of cell loss priority for a network, then the situation is more complex. Recall that the user may negotiate a traffic contract for high priority traffic ($CLP = 0$) and a separate contract for aggregate traffic (CLP 0 or 1). The following rules apply:

1. A cell with $CLP = 0$ that conforms to the traffic contract for $CLP = 0$ passes.
2. A cell with $CLP = 0$ that is noncompliant for ($CLP = 0$) traffic but compliant for (CLP 0 or 1) traffic is tagged and passed.
3. A cell with $CLP = 0$ that is noncompliant for ($CLP = 0$) traffic and noncompliant for (CLP 0 or 1) traffic is discarded.
4. A cell with $CLP = 1$ that is compliant for ($CLP = 0$ or 1) traffic is passed.
5. A cell with $CLP = 1$ that is noncompliant for (CLP 0 or 1) traffic is discarded.

Selective Cell Discard Selective cell discard comes into play when the network, at some point beyond the UPC function, discards ($CLP = 1$) cells. The objective is to discard lower-priority cells during congestion to protect the performance for higher-priority cells. Note that the network has no way to discriminate between cells that were labeled as lower priority by the source and cells that were tagged by the UPC function.

Traffic Shaping The UPC algorithm is referred to as a form of **traffic policing**. Traffic policing occurs when a flow of data is regulated so that cells (or frames or packets) that exceed a certain performance level are discarded or tagged. It may be desirable to supplement a traffic-policing policy with a **traffic-shaping** policy. Traffic shaping is used to smooth out a traffic flow and reduce cell clumping. This can result in a fairer allocation of resources and a reduced average delay time.

A simple approach to traffic shaping is to use a form of the UPC algorithm known as token bucket. In contrast to the UPC algorithm, which simply monitors the traffic and tags or discards noncompliant cells, a traffic-shaping token bucket controls the flow of compliant cells.

Figure 13.11 illustrates the basic principle of the token bucket. A token generator produces tokens at a rate of ρ tokens per second and places these in the token bucket, which has a maximum capacity of β tokens. Cells arriving from the source

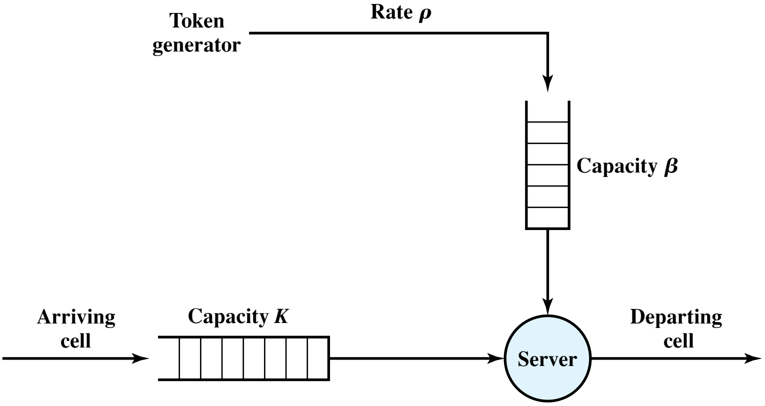


Figure 13.11 Token Bucket for Traffic Shaping

are placed in a buffer with a maximum capacity of K cells. To transmit a cell through the server, one token must be removed from the bucket. If the token bucket is empty, the cell is queued waiting for the next token. The result of this scheme is that if there is a backlog of cells and an empty bucket, then cells are emitted at a smooth flow of ρ cells per second with no cell delay variation until the backlog is cleared. Thus, the token bucket smoothes out bursts of cells.

13.7 ATM-GFR TRAFFIC MANAGEMENT

GFR (guaranteed frame rate) provides a service that is as simple as UBR (unspecified bit rate) from the end system's point of view while placing a relatively modest requirement on the ATM network elements in terms of processing complexity and overhead. In essence, with GFR, an end system does no policing or shaping of the traffic it transmits but may transmit at the line rate of the ATM adapter. As with UBR, there is no guarantee of frame delivery. It is up to a higher layer, such as TCP, to react to congestion that results in dropped frames by employing the window management and congestion control techniques discussed in Part Five. Unlike UBR, GFR allows the user to reserve a certain amount of capacity, in terms of a cell rate, for each GFR VC. A GFR reservation assures an application that it may transmit at a minimum rate without losses. If the network is not congested, the user will be able to transmit at a higher rate.

A distinctive characteristic of GFR is that it requires the network to recognize frames as well as cells. When congestion occurs, the network discards entire frames rather than individual cells. Further, GFR requires that all of the cells of a frame have the same CLP bit setting. The CLP = 1 AAL5 frames are treated as lower-priority frames that are to be transmitted on a best-effort basis. The minimum guaranteed capacity applies to the CLP = 0 frames.

The GFR traffic contract consists of the following parameters:

- Peak cell rate (PCR)
- Minimum cell rate (MCR)
- Maximum burst size (MBS)
- Maximum frame size (MFS)
- Cell delay variation tolerance (CDVT)

Mechanisms for Supporting Rate Guarantees

There are three basic approaches that can be used by the network to provide per-VC guarantees for GFR and to enable a number of users to efficiently use and fairly share the available network capacity [GOYA98]:

- Tagging and policing
- Buffer management
- Scheduling

These approaches can be combined in various ways in an ATM network elements to yield a number of possible GFR implementations. Figure 13.12 illustrates their use.

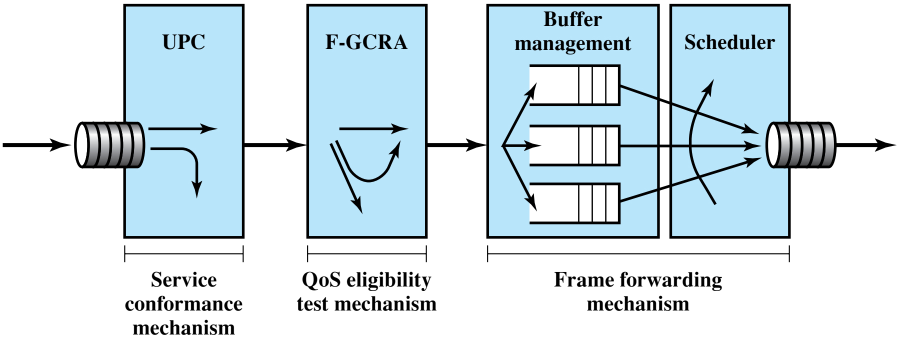

UPC
F-GCRA
Buffer management
Scheduler
Service conformance mechanism
QoS eligibility test mechanism
Frame forwarding mechanism

Figure 13.12 The Fundamental Components of a GFR Mechanism [ANDR99]

Tagging and Policing Tagging is used to discriminate between frames that conform to the GFR traffic contract and those that do not. The network element doing the conformance checking sets CLP = 1 on all cells of each frame that does not conform. Because tagged cells are assumed to be in violation of the traffic contract, they are given a lower quality of service than untagged cells by subsequent mechanisms, such as buffer management and scheduling. Tagging can be done by the network, especially the network element at the ingress to the ATM network. But tagging may also be done by the source end system to indicate less important frames.

The network, at either the ingress network element or at other ATM switching elements, may also choose to discard cells of nonconforming frames (i.e., cells with CLP = 1). Cell discard is considered a policing function.

Buffer Management Buffer management mechanisms have to do with the way in which cells are treated that have been buffered at a network switch or that arrive at a network switch and must be buffered prior to forwarding. When a congestion condition exists, as reflected by high buffer occupancy, a network element will discard tagged cells in preference to untagged cells. In particular, a network element may discard a tagged cell that is already in a buffer to make room for an incoming untagged cell. To provide fair and efficient use of buffer resources, a network element may perform per-VC buffering, dedicating a certain amount of buffer space to individual VCs. Then, on the basis of the traffic contracts for each VC and the buffer occupancy per VC, the network element can make decisions concerning cell discard. That is, cell discard can be based on queue-specific occupancy thresholds.

Scheduling A scheduling function, at minimum, can give preferential treatment to untagged cells over tagged cells. A network can also maintain separate queues for each VC and make per-VC scheduling decisions. Thus, within each queue, a first-come, first-served discipline can be used, perhaps modified to give higher priority for scheduling to CLP = 0 frames. Scheduling among the queues enables the network element to control the outgoing rate of individual VCs and thus ensure that individual VCs receive a fair allocation of capacity while meeting traffic contract requirements for minimum cell rate for each VC.

GFR Conformance Definition

The first function indicated in Figure 13.12 is a UPC function. UPC monitors each active VC to ensure that the traffic on each connection conforms to the traffic contract, and tags or discards nonconforming cells.

A frame is conforming if all of its cells are conforming, and is nonconforming if one or more cells are nonconforming. Three conditions must be met for a cell to be conforming:

1. The rate of cells must be within the cell rate contract.
2. All cells of a frame must have the same CLP value. Thus, the CLP bit of the current cell must have the same value as the CLP bit of the first cell of the frame.
3. The frame containing this cell must satisfy the MFS parameter. This condition can be met by performing the following test on each cell: The cell either is the last cell of the frame or the number of cells in the frame up to and including this cell is less than MFS.

QoS Eligibility Test Mechanism

The first two boxes in Figure 13.12 show what amounts to a two-stage filtering process. First, frames are tested for conformance to the traffic contract. Frames that do not conform may be discarded immediately. If a nonconforming frame is not discarded, its cells are tagged (CLP = 1), making them vulnerable to discard later on in the network. This first stage is therefore looking at an upper bound on traffic and penalizing cells that push the traffic flow above the upper bound.

The second stage of filtering determines which frames are eligible for QoS guarantees under the GFR contract for a given VC. This stage is looking at a lower bound on traffic; over a given period of time, those frames that constitute a traffic flow below the defined threshold are designated as eligible for QoS handling.

Therefore, the frames transmitted on a GFR VC fall into three categories:

- **Nonconforming frame:** Cells of this frame will be tagged or discarded.
- **Conforming but ineligible frames:** Cells will receive a best-effort service.
- **Conforming and eligible frames:** Cells will receive a guarantee of delivery.

To determine eligibility, a form of the cell rate algorithm referred to in Section 13.6 is used. A network may discard or tag any cells that are not eligible. However, TM 4.1 states that it is expected that an implementation will attempt to deliver conforming but ineligible traffic is on the basis of available resources, with each GFR connection being provided at each link with a fair share of the local residual bandwidth. The specification does not attempt to define a criterion by which to determine if a given implementation meets the aforementioned expectation.

13.8 RECOMMENDED READING

[YANG95] is a comprehensive survey of congestion control techniques. [JAIN90] and [JAIN92] provide excellent discussions of the requirements for congestion control, the various approaches that can be taken, and performance considerations. An excellent discussion of data network performance issues is provided by [KLEI93]. While somewhat dated, the definitive reference on flow control is [GERL80].

[GARR96] provides a rationale for the ATM service categories and discusses the traffic management implications of each. [MCDY99] contains a thorough discussion of ATM traffic control for CBR and VBR. Two excellent treatments of ATM traffic characteristics and performance are [GIRO99] and [SCHW96].

[ANDR99] provides a clear, detailed description of GFR. Another useful description is [BONA01].

Interesting examinations of frame relay congestion control issues are found in [CHEN89] and [DOSH88]. Good treatments are also found in [BUCK00] and [GORA99].

ANDR99 Andrikopoulos, I.; Liakopoulous, A.; Pavlou, G.; and Sun, Z. "Providing Rate Guarantees for Internet Application Traffic Across ATM Networks." *IEEE Communications Surveys*, Third Quarter 1999. **http://www.comsoc.org/pubs/surveys**

BONA01 Bonaventure, O., and Nelissen, J. "Guaranteed Frame Rate: A Better Service for TCP/IP in ATM Networks." *IEEE Network*, January/February 2001.

BUCK00 Buckwalter, J. *Frame Relay: Technology and Practice.* Reading, MA: Addison-Wesley, 2000.

CHEN89 Chen, K.; Ho, K.; and Saksena, V. "Analysis and Design of a Highly Reliable Transport Architecture for ISDN Frame-Relay Networks." *IEEE Journal on Selected Areas in Communications*, October 1989.

DOSH88 Doshi, B., and Nguyen, H. "Congestion Control in ISDN Frame-Relay Networks." *AT&T Technical Journal*, November/December 1988.

GARR96 Garrett, M. "A Service Architecture for ATM: From Applications to Scheduling." *IEEE Network*, May/June 1996.

GERL80 Gerla, M., and Kleinrock, L. "Flow Control: A Comparative Survey." *IEEE Transactions on Communications*, April 1980.

GIRO99 Giroux, N., and Ganti, S. *Quality of Service in ATM Networks.* Upper Saddle River, NJ: Prentice Hall, 1999.

GORA99 Goralski, W. *Frame Relay for High-Speed Networks.* New York: Wiley 1999.

JAIN90 Jain, R. "Congestion Control in Computer Networks: Issues and Trends." *IEEE Network Magazine*, May 1990.

JAIN92 Jain, R. "Myths About Congestion Management in High-Speed Networks." *Internetworking: Research and Experience*, Volume 3, 1992.

KLEI93 Kleinrock, L. "On the Modeling and Analysis of Computer Networks." *Proceedings of the IEEE*, August 1993.

MCDY99 McDysan, D., and Spohn, D. *ATM: Theory and Application.* New York: McGraw-Hill, 1999.

SCHW96 Schwartz, M. *Broadband Integrated Networks.* Upper Saddle River, NJ: Prentice Hall PTR, 1996.

YANG95 Yang, C., and Reddy, A. "A Taxonomy for Congestion Control Algorithms in Packet Switching Networks." *IEEE Network*, July/August 1995.

13.9 KEY TERMS, REVIEW QUESTIONS, AND PROBLEMS

Key Terms

backpressure cell delay variation choke packet congestion	congestion control explicit congestion signaling implicit congestion signaling	quality of service (QoS) reservations traffic management

Review Questions

13.1. When a node experiences saturation with respect to incoming packets, what general strategies may be used?

13.2. Why is it that when the load exceeds the network capacity, delay tends to infinity?

13.3. Give a brief explanation of each of the congestion control techniques illustrated in Figure 13.5.

13.4. What is the difference between backward and forward explicit congestion signaling?

13.5. Briefly explain the three general approaches to explicit congestion signaling.

13.6. Explain the concept of committed information rate (CIR) in frame relay networks

13.7. What is the significance of cell delay variation in an ATM network?

13.8. What functions are included in ATM usage parameter control?

13.9. What is the difference between traffic policing and traffic shaping?

Problems

13.1 A proposed congestion control technique is known as isarithmic control. In this method, the total number of frames in transit is fixed by inserting a fixed number of permits into the network. These permits circulate at random through the frame relay network. Whenever a frame handler wants to relay a frame just given to it by an attached user, it must first capture and destroy a permit. When the frame is delivered to the destination user by the frame handler to which it attaches, that frame handler reissues the permit. List three potential problems with this technique.

13.2 In the discussion of latency/speed effects in Section 13.5, an example was given in which over 7 megabits were transmitted before the source could react. But isn't a sliding-window flow control technique, such as described for HDLC, designed to cope with long propagation delays?

13.3 When the sustained traffic through a packet-switching node exceeds the node's capacity, the node must discard packets. Buffers only defer the congestion problem; they do not solve it. Consider the packet-switching network in Figure 13.13. Five stations attach to one of the network's nodes. The node has a single link to the rest of the network with a normalized throughput capacity of $C = 1.0$. Senders 1 through 5 are sending at average sustained rates of r_i of 0.1, 0.2, 0.3, 0.4, and 0.5, respectively. Clearly the node is overloaded. To deal with the congestion, the node discards packets from sender i with a probability of p_i.

a. Show the relationship among p_i, r_i, and C so that the rate of undiscarded packets does not exceed C.

The node establishes a discard policy by assigning values to the p_i such that the relationship derived in part (a) of this problem is satisfied. For each of the following policies, verify that the relationship is satisfied and describe in words the policy from the point of view of the senders.

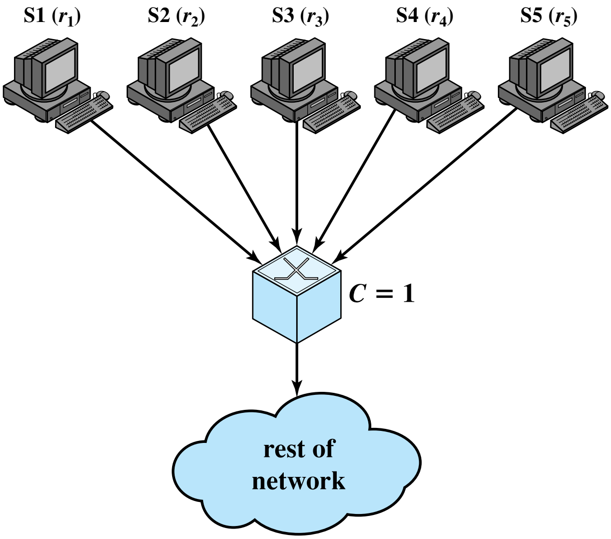


Figure 13.13 Stations Attached to a Packet-Switching Node

b. $p_1 = 0.333; p_2 = 0.333; p_3 = 0.333; p_4 = 0.333; p_5 = 0.333$
c. $p_1 = 0.091; p_2 = 0.182; p_3 = 0.273; p_4 = 0.364; p_5 = 0.455$
d. $p_1 = 0.0; p_2 = 0.0; p_3 = 0.222; p_4 = 0.417; p_5 = 0.533$
e. $p_1 = 0.0; p_2 = 0.0; p_3 = 0.0; p_4 = 0.0; p_5 = 1.0$

13.4 Consider the frame relay network depicted in Figure 13.14. C is the capacity of a link in frames per second. Node A presents a constant load of 0.8 frames per second destined

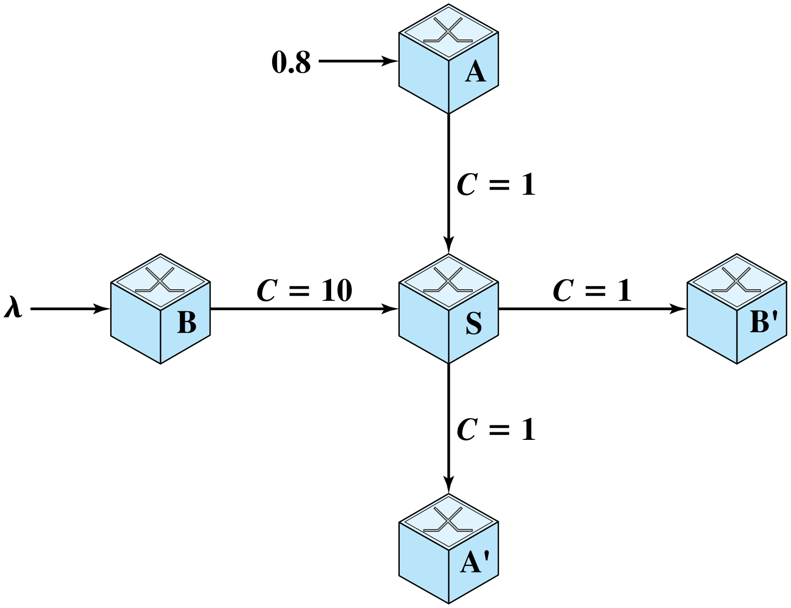


Figure 13.14 Network of Frame Relay Nodes

for A′. Node B presents a load λ destined for B′. Node S has a common pool of buffers that it uses for traffic both to A′ and B′. When the buffer is full, frames are discarded, and are later retransmitted by the source user. S has a throughput capacity of 2. Plot the total throughput (i.e., the sum of A-A′ and B-B′ delivered traffic) as a function of λ. What fraction of the throughput is A-A′ traffic for $\lambda > 1$?

13.5 For a frame relay network to be able to detect and then signal congestion, it is necessary for each frame handler to monitor its queuing behavior. If queue lengths begin to grow to a dangerous level, then either forward or backward explicit notification or a combination should be set to try to reduce the flow of frames through that frame handler. The frame handler has some choice as to which logical connections should be alerted to congestion. If congestion is becoming quite serious, all logical connections through a frame handler might be notified. In the early stages of congestion, the frame handler might just notify users for those connections that are generating the most traffic.

In one of the frame relay specifications, an algorithm for monitoring queue lengths is suggested; this is shown in Figure 13.15. A cycle begins when the outgoing circuit goes from idle (queue empty) to busy (nonzero queue size, including the current frame). If a threshold value is exceeded, then the circuit is in a state of incipient congestion, and the congestion avoidance bits should be set on some or all logical connections that use that circuit. Describe the algorithm in words and explain its advantages.

The algorithm makes use of the following variables:

t = current time
t_i = time of ith arrival or departure event
q_i = number of frames in the system after the event
T_0 = time at the beginning of the previous cycle
T_1 = time at the beginning of the current cycle

The algorithm consists of three components:

1. Update: Beginning with $q_0 := 0$
If the ith event is an arrival event, $q_i := q_{i-1} + 1$
If the ith event is a departure event, $q_i := q_{i-1} - 1$

2.

$$A_{i-1} = \sum_{\substack{i \\ t_i \in [T_0, T_1)}} q_{i-1}(t_i - t_{i-1})$$

$$A_i = \sum_{\substack{i \\ t_i \in [T_1, t)}} q_{i-1}(t_i - t_{i-1})$$

3.

$$L = \frac{A_i + A_{i-1}}{t - T_0}$$

Figure 13.15 A Frame Relay Algorithm

13.6 Compare sustainable cell rate and burst tolerance, as used in ATM networks, with committed information rate and excess burst size, as used in frame relay networks. Do the respective terms represent the same concepts?