CHAPTER 2
ABR TRAFFIC MANAGEMENT FRAMEWORK

The ABR service is intended for data applications with no delay guarantees. ABR, however, attempts to minimize the cell loss ratio, and gives minimum cell rate guarantees through a flow control mechanism. The network provides feedback to the sources when network load changes, and the sources adjust their transmission rates accordingly. ABR sources share the available bandwidth fairly, and the source is never required to send below its specified MCR.

This chapter provides insights into the development of ABR traffic management and explains reasons behind various decisions.

2.1 ABR RATE-BASED MODEL

ABR mechanisms allow the network to divide the available bandwidth fairly and efficiently among the active traffic sources. In the ABR traffic management framework, the source end systems limit their data transmission to rates allowed by the network. The network consists of switches which use their current load information to calculate the allowable rates for the sources. These rates are sent to the sources as feedback via *resource management* (RM) cells. RM cells are generated by the sources and travel along the data path to the destination end systems. The destinations simply return the RM cells to the sources. The components of the ABR traffic management framework are shown in Figure 2.1. In this chapter, we explain the source and destination end-system behaviors and their implications on ABR traffic management [7].

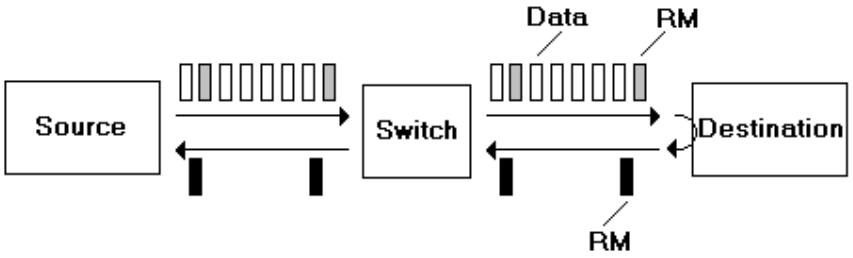


Figure 2.1: ABR Traffic Management Model

The ABR traffic management model is called a "*rate-based end-to-end closed-loop*" model. The model is called "rate-based" because the sources send data at a specified rate. This is different from current packet networks (for example, TCP), where the control is "window based" and the sources limit their transmission to a particular number of packets. The ABR model is called "closed-loop" because there is a continuous feedback of control information between the network and the source. If more sources become active, the rate allocated to each source is reduced. The model used for CBR and VBR traffic, on the other hand, is "open-loop" in the sense that rates are negotiated at the beginning of the connection and do not change dynamically. Finally, the model is called "end-to-end" because the control cells travel from the source to the destination and back to the source. The alternative of "hop-by-hop" control in which each switch would give feedback to the previous switch was considered and not accepted due to its complexity. However, one can achieve the hop-by-hop control in TM4.1 [4] using the virtual source/virtual destination (VS/VD) feature discussed later in this section.

When there is a steady flow of RM cells in the forward and reverse directions, there is a steady flow of feedback from the network. In this state, the ABR control loop has been established and the source rates are primarily controlled by the network feedback (closed-loop control). However, until the first RM cell returns, the source rate is controlled by the negotiated parameters, which may or may not relate to the current load on the network. The virtual circuit (VC) is said to be following an "open-loop" control during this phase. This phase normally lasts for one round-trip time (RTT). As we explain later, ABR sources are required to return to the open-loop control after long idle intervals. Traffic sources that have active periods (bursts) when data is transmitted at the allocated rate and idle periods when no data is transmitted are called "bursty sources". Open-loop control has a significant influence on the performance of bursty traffic particularly if it consists of bursts separated by long idle intervals.

2.1.1 Feedback Techniques

There are three ways for switches to give feedback to the sources:

In the first way, each cell header contains a bit called Explicit Forward Congestion Indication (EFCI), which can be set by a congested switch. Such switches are called "binary" or "EFCI" switches.

In the second way, RM cells have two bits in their payload, called the Congestion Indication (CI) bit and the No Increase (NI) bit, that can be set by congested switches. Switches that use only this mechanism are called "relative rate marking" switches.

In the third way, the RM cells also have another field in their payload called explicit rate (ER) that can be reduced by congested switches to any desired value. Such switches are called "explicit rate" switches.

Explicit rate switches normally wait for the arrival of an RM cell to give feedback to a source. However, under extreme congestion, they are allowed to generate an RM cell and send it immediately to the source. This optional mechanism is called “Backward Explicit Congestion Notification” (BECN).

Switches can use the virtual source/virtual destination (VS/VD) feature to segment the ABR control loop into smaller loops. In a VS/VD network, the switches additionally behave both as a (virtual) destination end system and as a (virtual) source end system. As a destination end system, it turns around the RM cells to the sources from one segment. As a source end system, it generates RM cells for the next segment. This feature can allow

feedback from nearby switches to reach sources faster, and allow hop-by-hop control as discussed earlier.

2.2 ABR PARAMETERS

Table 2.1: List of ABR Parameters

Label	Expansion	Default Value
PCR	Peak Cell Rate	-
MCR	Minimum Cell Rate	0
ACR	Allowed Cell Rate	-
ICR	Initial Cell Rate	PCR
TCR	Tagged Cell Rate	10 cells/sec
Nrm	Number of cells between FRM cells	32
Mrm	Controls bandwidth allocation between FRM, BRM and data cells	2
Trm	Upper Bound on Inter-FRM Time	100 ms
RIF	Rate Increase Factor	1/16
RDF	Rate Decrease Factor	1/16
ADTF	ACR Decrease Time Factor	500 ms
TBE	Transient Buffer Exposure	16777215
CRM	Missing RM-cell Count	TBE/Nrm
CDF	Cutoff Decrease Factor	1/16
FRTT	Fixed Round-Trip Time	-

At the time of connection setup, ABR sources negotiate several operating parameters with the network. The first among these is the peak cell rate (PCR). This is the maximum rate at which the source will be allowed to transmit on this virtual circuit (VC). The source can also request a minimum cell rate (MCR) which is the guaranteed minimum rate. The network has to reserve this bandwidth for the VC. During the data transmission stage, the rate at which a source is allowed to send at any particular instant is called the allowed cell rate (ACR). The ACR is dynamically changed between MCR and PCR. At the beginning of the connection, and after long idle intervals, ACR is set to initial cell rate (ICR). During the development of the RM specification, all numerical values in the specification were replaced by mnemonics. For example, instead of saying "every 32^{nd} cell should be an RM cell", the specification states "every Nrmth cell should be an RM cell." Here, Nrm is a parameter whose default value is 32. Some of the parameters are fixed while others are negotiated. A complete list of parameters used in the ABR mechanism is presented in Table 1. The parameters are explained as they occur in our discussion.

2.3 RESOURCE MANAGEMENT CELLS

2.3.1 In-Rate and Out-of-Rate RM Cells

Most resource management cells generated by the sources are counted as part of their network load in the sense that the total rate of data and RM cells should not exceed the ACR of the source. Such RM cells are called "*in-rate*" RM cells. Under exceptional circumstances, switches, destinations, or even sources can generate extra RM cells. These "*out-of-rate*" RM cells are not counted in the ACR of the source and are distinguished by having their cell loss priority (CLP) bit set, which means that the network will carry them only if there is plenty of bandwidth and can discard them if congested. The out-of-rate RM cells generated by the source and switch are limited to 10 RM cells per second per VC. One use of out-of-rate RM cells is for BECN from the switches. Another use is for a source, whose ACR has been set to zero by the network, to periodically sense the state of the network. Out-of-rate RM cells are also used by destinations of VCs whose reverse direction ACR is either zero or not sufficient to return all RM cells received in the forward direction. Note that in-rate and out-of-rate distinction applies only to RM cells. All data cells in ABR should have CLP set to 0 and must always be within the rate allowed by the network.

2.3.2 Forward and Backward RM cells

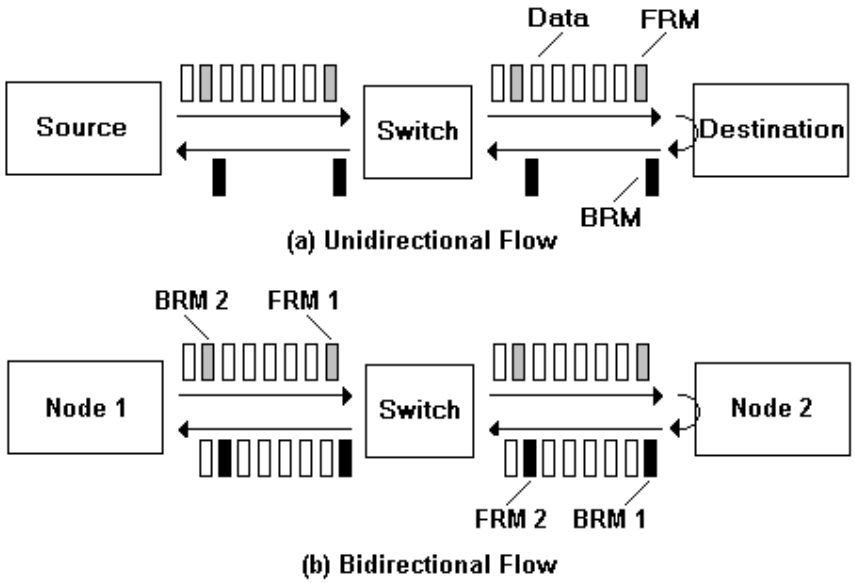


Figure 2.2: Forward and Backward Resource Management Cells (FRMs and BRMs)

Resource Management cells traveling from the source to the destination are called Forward RM (FRM) cells. The destination turns around these RM cells and sends them back to the source on the same VC. Such RM cells traveling from the destination to the source are called Backward RM (BRM) cells. Forward and backward RM cells are illustrated in Figure 2.2. Note that when there is bidirectional traffic, there are FRMs and BRMs in both directions on the Virtual Channel (VC). A bit in the RM cell payload indicates whether it is an FRM or BRM. This direction bit (DIR) is changed from 0 to 1 by the destination.

2.4 RM CELL FORMAT

ATM Header	5 Bytes
Protocol ID	1 Byte
Direction	1 bit
Backward Notification	1 bit
Congestion Indication	1 bit
No Increase	1 bit
Request / Acknowledge	1 bit
Reserved	3 bits
Explicit Rate	2 Bytes
Current Cell Rate	2 Bytes
Minimum Cell Rate	2 Bytes
Queue Length	4 Bytes
Sequence Number	4 Bytes
Reserved	30.75 Bytes
CRC-10	10 bits

Figure 2.3: Resource Management (RM) Cell Fields

The complete format of the RM cells is shown in figure 2.3. Every RM cell has the regular ATM header of five bytes. The payload type indicator (PTI) field is set to 110_2 to indicate that the cell is an RM cell. The protocol ID field, which is one byte long, is set to one for ABR connections. The direction (DIR) bit distinguishes forward and backward RM cells. The backward notification (BN) bit is set only in switch generated BECN cells. The congestion indication (CI) bit is used by relative rate marking switches. It may also be used by explicit rate switches under extreme congestion as discussed later. The no increase (NI) bit is another bit available to explicit rate switches to indicate moderate congestion. The request/acknowledge, queue length, and sequence number fields of the RM cells are for compatibility with the ITU-T recommendation I.371 and are not used by the ATM Forum.

The Current Cell Rate (CCR) field is used by the source to indicate to the network its current rate. Some switches may use the CCR field to determine a VC's next allocation while others may measure the VC's rate and not trust CCR. The minimum cell rate (MCR) field is redundant in the sense that like PCR, ICR, and other parameters it does not change during the life of a connection. However, its presence in the RM cells reduces number of lookups required in the switch.

The Explicit Rate (ER) field, the CI and the NI fields are used by the network to give feedback to the sources. The ER field indicates the maximum rate allowed to the source. When there are multiple switches along the path, the feedback given by the most congested link is the one that reaches the source. Data cells also have an Explicit Forward Congestion Indication (EFCI) bit in their headers, which may be set by the network when it experiences congestion. The destination saves the EFCI state of every data cell. If the EFCI state is set when it turns around an RM cell, it uses the CI bit to give (a single bit) feedback to the source. When the source receives the RM cell from the network, it adjusts its ACR using the ER, CI, NI values, and source parameters.

All rates (for example, ER, CCR, and MCR) in the RM cell are represented using a special 16-bit floating point format, which allows a maximum value of 4,290,772,992 cells per second (1.8 terabits per second). During connection setup, however, rate parameters are negotiated using a 24-bit integer format, which limits their maximum value to 16,777,215 cells per second or 7.1 Gbps.

2.5 SOURCE END SYSTEM RULES

TM4.1 [4] specifies 13 rules that the sources have to follow. This section discusses each rule and traces the development and implications of certain important rules. In some cases the precise statement of the rule is important.

- **Source Rule 1:** Sources should always transmit at a rate equal to or below their computed ACR. The ACR cannot exceed PCR and need not go below MCR. Mathematically,

$$MCR \leq ACR \leq PCR$$

$$\text{Source Rate} \leq ACR$$

- **Source Rule 2:** At the beginning of a connection, sources start at ICR. The first cell is always an in-rate forward RM cell. This ensures that the network feedback will be received as soon as possible.

- **Source Rule 3:** At any instant, sources have three kinds of cells to send: data cells, forward RM cells, and backward RM cells (corresponding to the reverse flow). The relative priority of these three kinds of cells is different at different transmission opportunities.

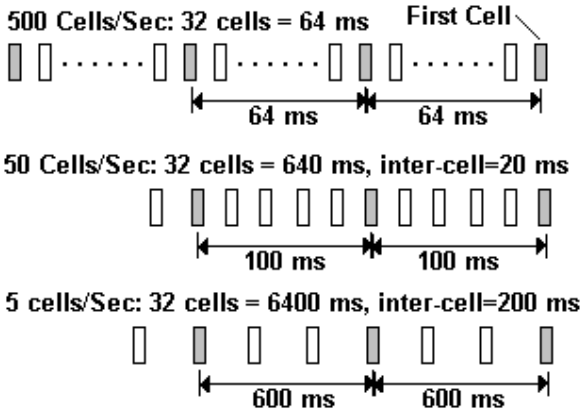


Figure 2.4: Frequency of forward RM cells.

First, the sources are required to send an FRM after every 31 cells. However, if the source rate is low, the time between RM cells will be large and network feedback will be delayed. To overcome this problem, a source is supposed to send an FRM cell if more than 100 ms has elapsed since the last FRM. This introduces another problem for low rate sources. In some cases, at every transmission opportunity the source may find that it has exceeded 100 ms and needs to send an FRM cell. In this case, no data cells will be transmitted. To overcome this problem, an additional condition is added that there must be at least two other cells between FRMs. An example of the operation of the above condition is shown in the figure 2.4. The figure assumes a unidirectional VC (i.e., there are no BRMs to be turned around). The figure has three parts. The first part of the figure shows that, when the source rate is 500 cells/s, every 32nd cell is an FRM cell. The time to send 32 cells is always smaller than 100 ms. In the second part of the figure, the source rate is 50 cells/s. Hence 32 cells takes 640 ms to be transmitted. Therefore, after 100 ms, an FRM is scheduled in the next transmission opportunity (or slot). The third part of the figure shows the scenario when the source rate is 5

cells/s. The inter-cell time itself is 200 ms. In this case, an FRM is scheduled every three slots, i.e., the inter-FRM time is 600 ms.

Second, a waiting BRM has priority over waiting data, given that no BRM has been sent since the last FRM. Of course, if there are no data cells to send, waiting BRMs may be sent.

Third, data cells have priority in the remaining slots.

The second and third part of the this rule ensure that BRMs are not unnecessarily delayed and that all available bandwidth is not used up by the RM cells.

- **Source Rule 4:** All RM cells sent in accordance with rules 1-3 are in-rate RM cells and have their cell loss priority (CLP) bit set to 0. Additional RM cells may be sent out-of-rate and should have their CLP bit set to 1. For example, consider the third unidirectional flow of Figure 2.4. It has an ACR of 5 cells/sec. It is allowed to send only one in-rate RM cell every 600 ms. If necessary, it can send a limited number of out-of-rate RM cells with CLP set to 1.

- **Source Rule 5:** The rate allowed to a source is valid only for approximately 500 ms. If a source does not transmit any RM cells for this duration, it cannot use its previously allocated ACR particularly if the ACR is high. The source should re-sense the network state by sending an RM cell and decreasing its rate to the initial cell rate (ICR) negotiated at connection setup. If a source's ACR is already below ICR, it should stay at that lower value (and not increase it to ICR). The timeout interval is set by the ACR Decrease Time Factor (ADTF). This parameter can be negotiated with the network at connection setup. Its default value is 500 ms. This simple rule was the cause of a big debate at the Forum. It is intended to solve the problem of ACR retention. If a source sends an RM cell when the network is not heavily loaded, the source may be granted a very high rate. The source can then retain that rate and use it when the network is highly loaded. In fact, a source may set up several VCs and use them to get an unfair advantage. To solve this problem, several so called "*use it or lose it*" (UILI) solutions were proposed. Some of them relied on actions at the source while others relied on actions at the switch. The source based solutions required sources to monitor their own rates and reduce ACR slowly if was too high compared to the rate used. UILI alternatives were analyzed and debated for months because they have a significant impact on the performance of bursty traffic that forms the bulk of data traffic.

  The ATM Forum chose to standardize a very simple UILI policy at the source. This policy provided a simple timeout method (using ADTF as the timeout value) which reduces ACR to

ICR when the timeout expires. Vendors are free to implement additional proprietary restraints at the source or at the switch. A few examples of such possibilities are listed in the Informative Appendix I.7 of the specification [4].

- **Source Rule 6:** If a network link becomes broken or becomes highly congested, the RM cells may get blocked in a queue and the source may not receive the feedback. To protect the network from continuous in-flow of traffic under such circumstances, the sources are required to reduce their rate if the network feedback is not received in a timely manner. Normally under steady state, sources should receive one BRM for every FRM sent. Under congestion, BRM cells may be delayed. If a source has sent CRM FRM cells and has not received any BRM, it should suspect network congestion and reduce its rate by a factor of CDF. Here, CRM (missing RM cell count) and CDF (cutoff decrease factor) are parameters negotiated at the time of connection setup. BECN cells generated by switches (and identified by BN=1) are not counted as BRM. When rule 6 triggers once, the condition is satisfied for all successive FRM cells until a BRM is received. Thus, this rule results in a fast exponential decrease of ACR. An important side effect of this rule is that unless CRM is set high, the rule could trigger unnecessarily on a long delay path. CRM is computed from another parameter called transient buffer exposure (TBE) which is negotiated at connection setup. TBE determines the maximum number of cells that may suddenly appear at the switch during the first round trip before the closed-loop phase of the control takes effect. During this time, the source will have sent TBE/Nrm RM cells. Hence,

$$\text{CRM} = \left\lceil \frac{\text{TBE}}{\text{Nrm}} \right\rceil$$

The fixed part of the round-trip time (FRTT) is computed during connection setup. This is the minimum delay along the path and does not include any queuing delay. During this time, a source may send as many as ICR * FRTT cells into the network. Since this number is negotiated separately as TBE, the following relationship exists between ICR and TBE:

$$\text{ICR} * \text{FRTT} \leq \text{TBE} \text{ or } \text{ICR} \leq \text{TBE} / \text{FRTT}$$

The sources are required to use the ICR value computed above if it is less than the ICR negotiated with the network. In other words:

ICR used by the source = Min (ICR negotiated with the network, TBE/FRTT)

In negotiating TBE, the switches have to consider their buffer availability. As the name indicates, the switch may be suddenly exposed to TBE cells during the first round trip (and also after long idle periods). For small buffers, TBE should be small and vice versa. On the

other hand, TBE should also be large enough to prevent unnecessary triggering of rule 6 on long delay paths. It has been incorrectly believed that cell loss could be avoided by simply negotiating a TBE value below the number of available buffers in the switches. The only reliable way to protect a switch from large queues is to build it in the switch allocation algorithm. The ERICA algorithm [8], which will be explained later, is an example of one such algorithm.

Observe that the FRTT parameter, which is the sum of fixed delays on the path, is used in the formula for ICR. During the development of this rule, an estimate of round trip time (RTT), including the fixed and variable delays was being used instead of FRTT in the ICR calculation. Jain et al [7] argued that RTT estimated at connection setup is a random quantity bearing little relation to the round trip delays during actual operation. Such parameter setting could trigger source Rule 6 unnecessarily and degrade performance. Hence, the Forum decided to use FRTT parameter instead of RTT.

Note that it is possible to disable source Rule 6, by setting CDF to zero.

- **Source Rule 7:** When sending an FRM, the sources should indicate their current ACR in the CCR field of the RM cells.

- **Source Rules 8 and 9:** Source Rule 8 and 9 describe how the source should react to network feedback. The feedback consists of explicit rate (ER), congestion indication bit (CI), and no-increase bit (NI). Normally, a source could simply change its ACR to the new ER value but this could cause a few problems as discussed next.

First, if the new ER is very high compared to current ACR, switching to the new ER will cause sudden queues in the network. Therefore, the amount of increase is limited. The rate increase factor (RIF) parameter determines the maximum allowed increase in any one step. The source cannot increase its ACR by more than RIF * PCR. Second, if there are any EFCI switches in the path, they do not change the ER field. Instead, they set EFCI bits in the cell headers. The destination monitors these bits and returns the last seen EFCI bit in the CI field of a BRM. A CI of 1 means that the network is congested and that the source should reduce its rate. The decrease is determined by rate decrease factor (RDF) parameter. Unlike the increase, which is additive, the decrease is multiplicative in the sense that

$$ACR \leftarrow ACR * (1 - RDF)$$

It has been shown that additive increase and multiplicative decrease is sufficient to achieve fairness. Other combinations are unfair. The no-increase (NI) bit was introduced to handle

mild congestion cases. In such cases, a switch could specify an ER, but instruct that, if ACR is already below the specified ER, the source should not increase the rate. The actions corresponding to the various values of CI and NI bits are as follows:

NI	CI	Action
0	0	ACR ← Min(ER, ACR + RIF * PCR, PCR)
0	1	ACR ← Min(ER, ACR - ACR * RDF)
1	0	ACR ← Min(ER, ACR)
1	1	ACR ← Min(ER, ACR - ACR * RDF)

ACR ← Max (ACR, MCR)

If there are no EFCI switches in a network, setting RIF to 1 allows ACRs to increase as fast as the network directs it. This allows the available bandwidth to be used quickly. For EFCI networks or a combination of ER and EFCI networks, RIF should be set conservatively to avoid unnecessary oscillations. Once the ACR is updated, the subsequent cells sent from the source conform to the new ACR value. However, if the earlier ACR was very low, it is possible that the very next cell is scheduled a long time in the future. In such a situation, it is advantageous to *reschedule* the next cell, so that the source can take advantage of the high ACR allocation immediately.

- **Source Rule 10:** Sources should initialize various fields of FRM cells as follows. For virtual path connections (VPCs), the virtual circuit id (VCI) is set to 6. For virtual channel connections (VCCs), the VCI of the connection is used. In either case, the protocol type id (PTI) in the ATM cell header is set to 6 (110_2). The protocol id field in the payload of the RM cell is set to 1. The direction bit should be set to 0 (forward). The backward notification (BN) bits should be set to 0 (source generated). Explicit rate field is initialized to the maximum rate below PCR that the source can support. Current cell rate is set to current ACR. Minimum cell rate is set to the value negotiated at connection setup. Queue length, sequence number, and request/acknowledge field are set in accordance with ITU-T recommendation I.371 or to zero. All reserved octets are set to 6A (hex) or 01101010 (binary). This value is specified in ITU-T recommendation I.610 (whose number coincidentally is also 6A in hex). Other reserved bits are set to 0. Note that the sources are allowed to set ER and NI field to indicate their own congestion.

- **Source Rule 11:** The out-of-rate FRM cells generated by sources are limited to a rate below the "tagged cell rate (TCR)" parameter, which has a default value of 10 cells/sec.

- **Source Rule 12:** The EFCI bit must be reset on every data cell sent.

- **Source Rule 13:** Sources can optionally implement additional Use-It-or-Lose-It (UILI) policies (see discussion of source Rule 5).

2.6 DESTINATION END SYSTEM RULES

- **Destination Rule 1:** Destinations should monitor the EFCI bits on the incoming cells and store the value last seen on a data cell.

- **Destination Rule 2:** Destinations are required to turn around the forward RM cells with minimal modifications as follows: the DIR bit is set to "backward" to indicate that the cell is a backward RM-cell; the BN bit is set to zero to indicate that the cell was not generated by a switch; the CCR and MCR field should not be changed. If the last cell has EFCI bit set, the CI bit in the next BRM is set and the stored EFCI state is cleared.
If the destination has internal congestion, it may reduce the ER or set the CI or NI bits just like a switch.

- **Destination Rules 3-4:** The destination should turn around the RM cells as fast as possible. However, an RM cell may be delayed if the reverse ACR is low. In such cases destination rules 3 and 4 specify that old out-of-date information can be discarded. The destinations are allowed a number of options to do this. The implications of various options of destination Rule 3 are discussed in the Informative Appendix I.6 of the TM specification [4]. Briefly, the recommendations attempt to ensure the flow of feedback to the sources for a wide range of values of ACR of the reverse direction VC.

  If the reverse direction ACR is non-zero, then a backward RM cell will be scheduled for in-rate transmission. Transmitting backward RM cells out-of-rate ensures that the feedback is sent regularly even if the reverse ACR is low or zero (for example, in unidirectional VCs). Note that there is no specified limit on the rate of such "turned around" out-of-rate RM cells. However, the CLP bit is set to 1 in the out-of-rate cells, which allows them to be selectively dropped by the switch if congestion is experienced.

- **Destination Rule 5:** Sometimes a destination may be too congested and may want the source to reduce its rate immediately without having to wait for the next RM cell. Therefore, like a switch, the destinations are allowed to generate BECN RM cells. Also, as in the case of switch generated BECNs, these cells may not ask a source to increase its rate (CI bit is set). These BECN cells are limited to 10 cells/s and their CLP bits are set (i.e., they are sent out-of-rate).

- **Destination Rule 6:** An out-of-rate FRM cell may be turned around either in-rate (with CLP=0) or out-of-rate (with CLP=1).

2.7 SWITCH BEHAVIOR

The switch behavior specifies that the switch must implement some form of congestion control, and rules regarding processing, queuing and generation of RM cells.

- **Switch Rule 1:** This rule specifies that one or more methods of feedback marking methods must be implemented at the switch. The possible methods include:
  - **EFCI Marking:** This defines the binary (bit-based) feedback framework, where switches may set the EFCI bit in data cell headers. Note that the VC's EFCI state at the destination is set and reset whenever an incoming data cell has its EFCI set or reset respectively.
  - **Relative Rate Marking:** This option allows the switch to set two bits in the RM cell which have a specific meaning to when they reach the source end systems. When the CI bit is set, it asks the source to decrease, while the NI bit tells the source not to increase beyond its current rate, ACR. Observe that the source rate may be further reduced using the explicit rate indication field. These bits allow the switches some more flexibility than the EFCI bit marking. Specifically, the switches can avoid the "beat-down" fairness problem seen in EFCI marking scenarios. The problem occurs because connections going through several switches have a higher probability of their EFCI bits being set, than connections going through a smaller number of switches.
  - **Explicit Rate Marking:** Allows the switch to specify exactly what rate it wants a source to send at. To ensure coordination among multiple switches in a connection's path, the switch may reduce (but not increase) the ER field in the RM-cells (in the forward and/or backward directions). This thesis deals mainly with explicit rate feedback from switches.

- **VS/VD Control:** In this mode, the switch may segment the ABR control loop by appearing as a "virtual source" to one side of the loop and as a "virtual destination" to the other side.

- **Switch Rule 2:** This rule specifies how a switch may generate an RM cell in case it is heavily congested and doesn't see RM cells from the source. Basically, the rule allows such RM cells to only decrease the source rate, and these RM cells are sent out-of-rate. The rate of these backward RM-cells shall be limited to 10 cells/second, per connection. When a switch generates an RM-cell it shall set either CI=1 or NI=1, shall set BN=1, and shall set the direction to backward.

- **Switch Rule 3:** This rule says that the RM cells may be transmitted out-of-sequence, but the sequence integrity must be maintained. This rule allows the switch the flexibility to put the RM cells on a priority queue for faster feedback to sources when congested. However, by queuing RM cells separately from the data stream, the correlation between the quantities declared RM cells and the actual values in the data stream may be lost.

- **Switch Rule 4 and 5:** Rule 4 specifies alignment with ITU-T's I.371 draft, and ensures the integrity of the MCR field in the RM cell. Rule 5 allows the optional implementation of a use-it-or-lose-it policy at the switch [9].

2.8 SWITCH RATE CALCULATION ALGORITHMS

2.8.1 Selection Criteria

The traffic-management working-group was started in the ATM Forum in May 1993. A number of congestion schemes were presented. To sort out these proposals, the group decided to first agree on a set of selection criteria. Since these criteria are of general interest and apply to non-ATM networks as well, we describe some of them briefly here [10].

2.8.1.1 Scalability

Networks are generally classified based on extent (coverage), number of nodes, speed, or number of users. Since ATM networks are intended to cover a wide range along all these dimensions, it is necessary that the scheme be not limited to a particular range of speed, distance, number of switches, or number of VCs. In particular, this ensures that the same scheme can be used for local area networks (LANs) as well as wide area networks (WANs).

2.8.1.2 Optimality

In a shared environment the throughput for a source depends upon the demands by other sources. The most commonly used criterion for what is the correct share of bandwidth for a source in a network environment, is the so called "max-min allocation". It provides the maximum possible bandwidth to the source receiving the least among all contending sources. Mathematically, it is defined as follows. Given a configuration with n contending sources, suppose the i^{th} source gets a bandwidth X_i. The allocation vector $\{X_1, X_2, \ldots, X_n\}$ is feasible if all link load levels are less than or equal to 100%. The total number of feasible vectors is infinite. For each allocation vector, the source that is getting the least allocation is in some sense, the "unhappiest source". Given the set of all feasible vectors, find the vector that gives the maximum allocation to this unhappiest source. Actually, the number of such vectors is also infinite although we have narrowed down the search region considerably. Now we take this "unhappiest source" out and reduce the problem to that of remaining n-1 sources operating on a network with reduced link capacities. Again, we find the unhappiest source among these n-1 sources, give that source the maximum allocation and reduce the problem by one source. We keep repeating this process until all sources have been given the maximum that they could get.

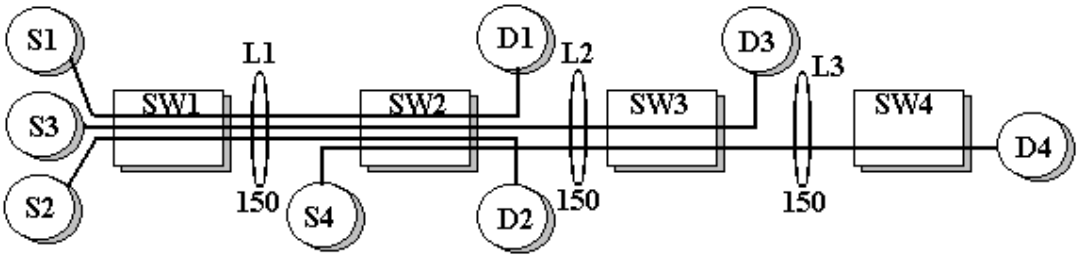
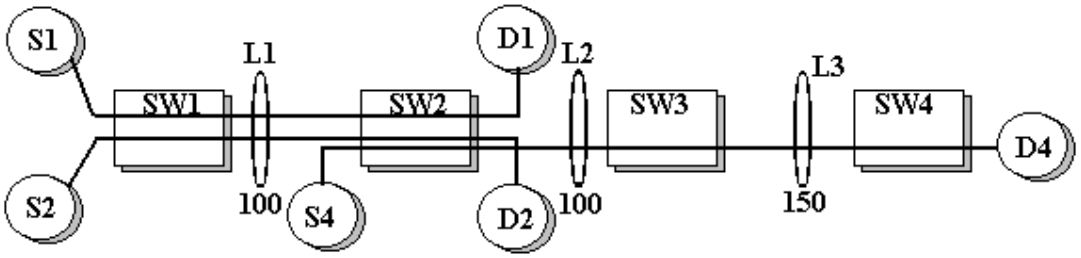


Figure 2.5: Sample configuration for max-min fairness

The following example illustrates the above concept of max-min fairness. Figure 2.5 shows a network with four switches connected via three 150 Mbps links. Four VCs are setup such that the first link L1 is shared by sources S1, S2, and S3. The second link is shared by S3 and S4. The third link is used only by S4. Let us divide the link bandwidths fairly among contending sources. On link L1, we can give 50 Mbps to each of the three contending sources S1, S2, and S3. On link L2, we would give 75 Mbps to each of the sources S3 and S4. On link L3, we would give all 150 Mbps to source S4. However, source S3 cannot use its 75 Mbps share at link L2 since it is allowed to use only 50 Mbps at link L1. Therefore, we give 50

Mbps to source S3 and construct a new configuration shown in Figure 2.6, where Source S3 has been removed and the link capacities have been reduced accordingly. Now we give ½ of the link L1's remaining capacity to each of the two contending sources: S1 and S2; each gets 50 Mbps. Source S4 gets the entire remaining bandwidth (100 Mbps) of link L2. Thus, the fair allocation vector for this configuration is (50, 50, 50, 100). This is the max-min allocation.

Figure 2.6: Configuration after removing VC 3.



Notice that max-min allocation is both fair and efficient. It is fair in the sense that all sources get an equal share on every link provided that they can use it. It is efficient in the sense that each link is utilized to the maximum load possible. It must be pointed out that the max-min fairness is just one of several possible optimality criteria. It does not account for the guaranteed minimum (MCR). Other criterion such as weighted fairness have been proposed to determine optimal allocation of resources over and above MCR.

2.8.1.3 Fairness Index

Given any optimality criterion, one can determine the optimal allocation. If a scheme gives an allocation that is different from the optimal, its unfairness is quantified numerically as follows.

Suppose a scheme allocates $\{Y_1, Y_2, \ldots, Y_n\}$ instead of the optimal allocation $\{X_1, X_2, \ldots, X_n\}$. Then, we calculate the normalized allocations $Z_i = Y_i/X_i$ for each source and compute the fairness index as follows:

$$\text{Fairness} = \frac{\left(\sum_i Z_i\right)^2}{n\sum_i Z_i^2}$$

Since allocations Z_is usually vary with time, the fairness can be plotted as a function of time. Alternatively, throughputs over a given interval can be used to compute overall fairness.

2.8.1.4 Robustness

The scheme should be insensitive to minor deviations. For example, slight mistuning of parameters or loss of control messages should not bring the network down. It should be possible to isolate misbehaving users and protect other users from them.

2.8.1.5 Implementability

The scheme should not dictate a particular switch architecture.

2.8.1.6 Simulation Configurations

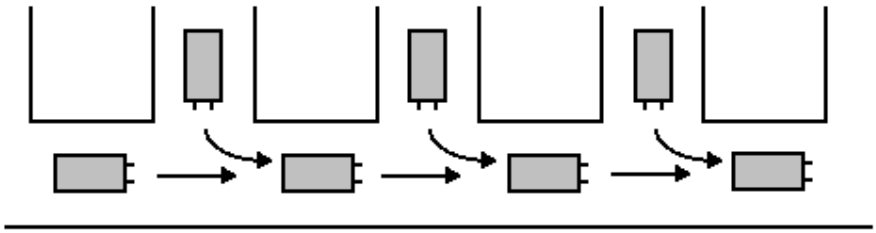


Figure 2.7: Theatre parking lot

A number of network configurations were also agreed upon to compare various proposals. Most of these were straightforward serial connection of switches. The most popular one is the so called "Parking Lot" configuration for studying fairness. The configuration and its name is derived from theatre parking lots, which consist of several parking areas connected via a single exit path as shown in Figure 2.7. At the end of the show, congestion occurs as cars exiting from each parking area try to join the main exit stream.

For computer networks, an n-stage parking lot configuration consists of n switches connected in a series. There are n VCs. The first VC starts from the first switch and goes to the end. For the remaining i^{th} VC starts at the $i\text{-}1^{th}$ switch. A 3-switch parking lot configuration is shown in Figure 2.7.

2.8.1.7 Traffic Patterns

Among the traffic patterns used in various simulations, the following three were most common:

1. *Persistent Sources*: These sources, also known as "greedy" or "infinite" sources always have cells to send. Thus, the network is always congested.

2. *Staggered Source*: The sources start at different times. This allows us to study the ramp-up (or ramp-down) time of the schemes.
3. *Bursty Sources*: These sources oscillate between active state and idle state. During active state, they generate a burst of cells.

2.8.2 Example Switch Algorithm: ERICA

The ATM Forum Traffic Management Specification provides in precise details the rules for the source and destination end system behaviors for the available bit rate (ABR) service for asynchronous transfer mode (ATM) networks as discussed in sections 2.5 and 2.6. The switch behavior, however, is only coarsely specified. This provides the flexibility for various vendors to implement their own switch rate allocation algorithms. Several switch algorithms have been developed. This section describes one of the earliest switch algorithms.

The Explicit Rate Indication for Congestion Avoidance (ERICA) algorithm was presented at the ATM Forum in February 1995. Since then, its performance has been independently studied, and several modifications have been incorporated into the algorithm. This section provides a consolidated description of the scheme [8].

2.8.2.1 Switch Model

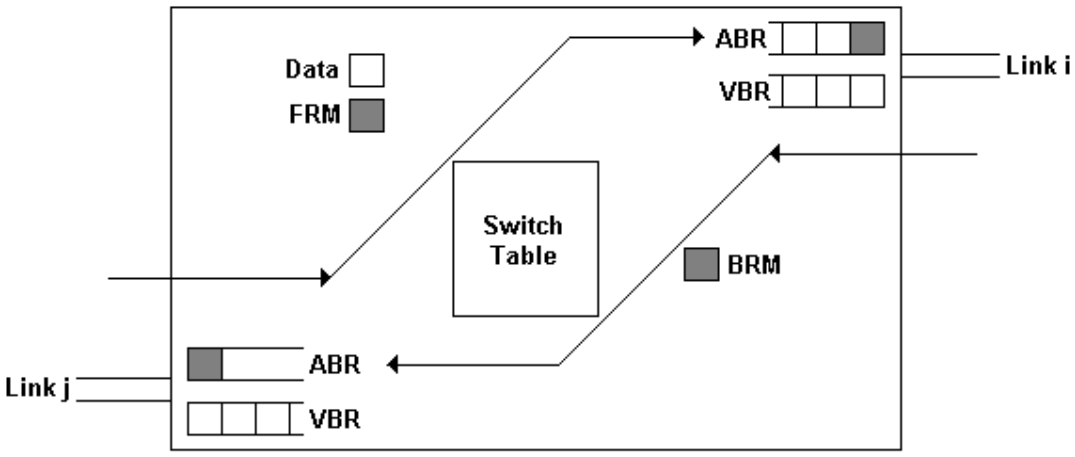
The algorithm switch model is shown in figure 2.8. Every service class has a separate FIFO output queue which feeds to the output link under the control of a scheduling mechanism. The ERICA algorithm (proposed by Jain et al. [8]) works at the ABR output queuing point. The aspects of the scheduling mechanism doesn't be discussed except for the fact that it provides the switch algorithm with the knowledge of the available capacity for the ABR service. It is assumed that there are at most two classes (VBR and ABR) and ABR has the lowest priority, i.e., it gets the leftover capacity after VBR cells are transmitted. In practice, it is desirable to allow some minimum capacity for processing aggregate ABR traffic when there is contention.

In figure 2.8, observe that the RM cells of a connection enter the switch through one port in the forward direction (with the forward data flow) and exit through another port in the reverse direction (with the reverse data flow). In the ERICA algorithm, the forward flow is monitored for metrics, but the feedback is given in the backward RM cells, thus minimizing the latency in delivering feedback to sources. Jain et al. measured certain characteristics of the flow over intervals of time, called "switch measurement intervals" or "switch averaging intervals". The measured quantities are placed in a table for use in the reverse direction when calculating feedback. The feedback calculation may be performed when a backward RM cell

(BRM) is received in the reverse direction, or may be pre-calculated at the end of the previous averaging interval (of the forward direction port) for all active sources. The latter option may also be implemented using lazy evaluation and/or in the background using a dedicated processor.

Figure 2.8: Switch Model

One key feature of the ERICA scheme is that it gives at most one feedback value



per-source during any averaging interval. As a result, it precludes the switch from giving multiple conflicting feedback indications in a single averaging interval using the same control values. Further, since there can be multiple switches on a VC's path, the allocation given to the source is the minimum of all the switch allocations. For performance purposes, it is desirable to have all the switches implement the same switch algorithm, but the traffic management standard [4] does allow switches from multiple vendors to interoperate.

While ERICA gives feedback in the explicit rate field in the RM cell, it is possible to additionally throttle or moderate the sources by setting the CI and NI bits in the RM cell using policies suggested by several other schemes. Also, in Jain et al. studies they set the Rate Increase Factor (RIF) parameter to one allowing maximum increase. Sources/switches can choose to be more conservative and set it to lower values.

2.8.2.2 The Basic Algorithm

The ERICA algorithm operates at each output port (or link) of a switch. The switch periodically monitors the load on each link and determines a load factor (z), the ABR capacity, and the number of currently active virtual connections or VCs (N). A measurement or “averaging” interval is used for this purpose. These quantities are used to calculate the

feedback which is indicated in RM cells. Recall from the discussion in section 2.8.1 that the measurements are made in the forward direction, whereas the feedback is given in the reverse direction. Further, the switch gives at most one new feedback per source in any averaging interval. The key steps in ERICA are as follows:

Initialization:

MaxAllocPrevious ← MaxAllocCurrent ← FairShare

End of Averaging Interval:

1. Total ABR Capacity ← Link Capacity – VBR Capacity
2. Target ABR Capacity ← Fraction * Total ABR Capacity
3. Z ←ABR Input Rate/Target ABR Capacity
4. FairShare ← Target ABR Capacity /Number of Active VCs
5. MaxAllocPrevious ← MaxAllocCurrent
6. MaxAllocCurrent ← FairShare

When an FRM is received:

CCR[VC] ←CCR in RM Cell

When a BRM is received:

1. VCShare ← CCR[VC]/Z
2. IF (Z > 1 + δ) THEN

ER ← Max (FairShare, VCShare)

ELSE ER ← Max (MaxAllocPrevious, VCShare)

3. MaxAllocCurrent ← Max (MaxAllocCurrent,ER)
4. IF (ER > FairShare AND CCR[VC] < FairShare) THEN

ER ← FairShare

5. ER in RM Cell / Min (ER in RM Cell, ER, Target ABR Capacity)

A complete explanation, enhancements, features and pseudo code of the scheme are provided in reference [11].