CHAPTER 1

Principles of Telecommunications Network Architecture

A telecommunications network is a collection of nodes and links that communicate by defined sets of formats and protocols. Within the network there are usually three layers: transmission, switching, and service. The transmission layer consists of transmission systems, for example, cables, radio links, and their related technical equipment. The switching layer consists of switching nodes with generic and application software and data. The service layer, distributed among the switching network elements, consists of special hardware, and their application software and data.

1.1 Broadband Networks

Broadband integrated services digital network (B-ISDN) based on asynchronous transfer mode (ATM) is the technology of choice for transport of information from multimedia services and applications.

ATM is a cell-based, high-bandwidth, low-delay switching and multiplexing technology that is designed to deliver a variety of high-speed digital communication services. These services include LAN (local area network) interconnection, imaging, and multimedia applications as well as video distribution, video telephony and other video applications.

ATM is asynchronous since the recurrence of cells containing information from an individual customer is not necessarily periodic. ATM handles both connection-oriented and connectionless traffic through the use of adaptation layers and operates at either a constant bit rate (CBR) or variable bit rate (VBR) connection. Each ATM cell sent into the network contains addressing information that establishes a virtual connection from source to destination. All cells are then transferred in sequence over this virtual connection. ATM supports permanent virtual connections (PVC) as well as switched virtual connections (SVC).

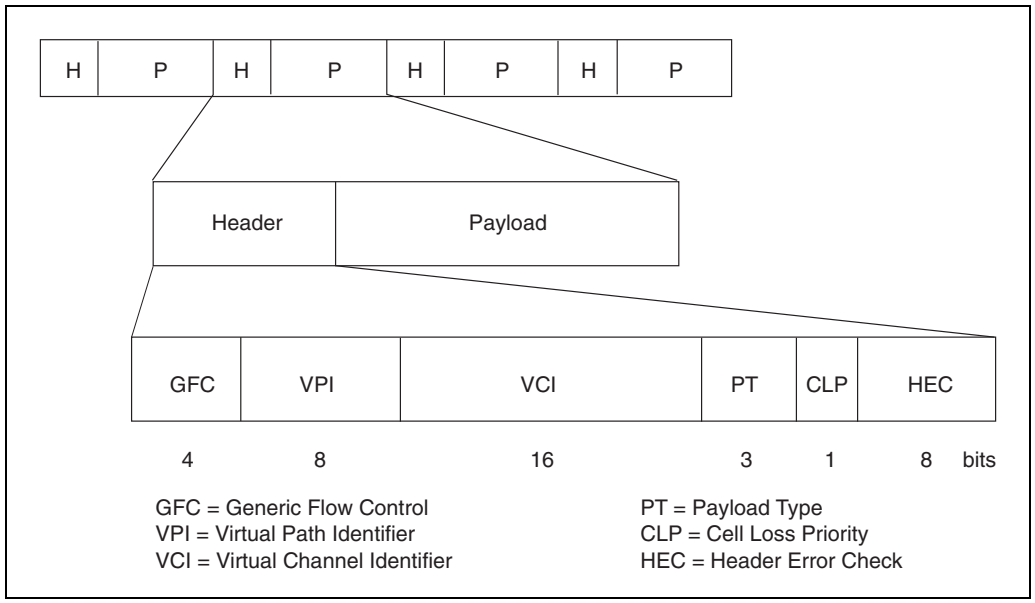

Figure 1–1 ATM Cell Format

1.1.1 Basic ATM Protocols

ATM standards define a fixed-size cell with a length of 53 bytes comprised of a 5-byte header and a 48-byte payload as shown in Figure 1–1. Broadband ISDN supports multimedia applications because of its high performance.

Figure 1–2 depicts the ATM Broadband-ISDN protocol reference model as defined in ITU-T Recommendation I.321. The physical layer (PHY) has two sublayers: transmission convergence (TC) and physical medium-dependent (PMD). The PMD sublayer interfaces with the actual physical medium and passes the recovered bit stream to the TC sublayer. The TC sublayer extracts and inserts the ATM cell with the synchronous digital hierarchy (SDH) time division multiplexed (TDM) frame and then passes them to and from the ATM layer, respectively. The ATM layer performs multiplexing, switching, and controlling actions based upon information in the ATM cell header. It passes cells to and accepts cells from the ATM adaptation layer (AAL). The AAL also has two sublayers: segmentation and reassembly (SAR) and convergence sublayer (CS). The AAL passes protocol data units (PDUs) to and accepts them from higher layers. PDUs may differ in variable or fixed length from the ATM cell length. The physical layer (PHY) corresponds to layer 1 of the PHY in the OSI (Open System Interconnection) model while the ATM layer and adaptation layer (AAL) correspond to parts of OSI layer 2 or data-link layer.

The PHY consists of two logical sublayers: the physical medium-dependent (PMD) sublayer and the transmission convergence (TC) sublayer. PMD includes only PMD functions. It provides bit transmission capability, including bit transfer, bit alignment, line coding, and electrical–optical conversion. TC performs functions required to transform a flow of cells into a flow

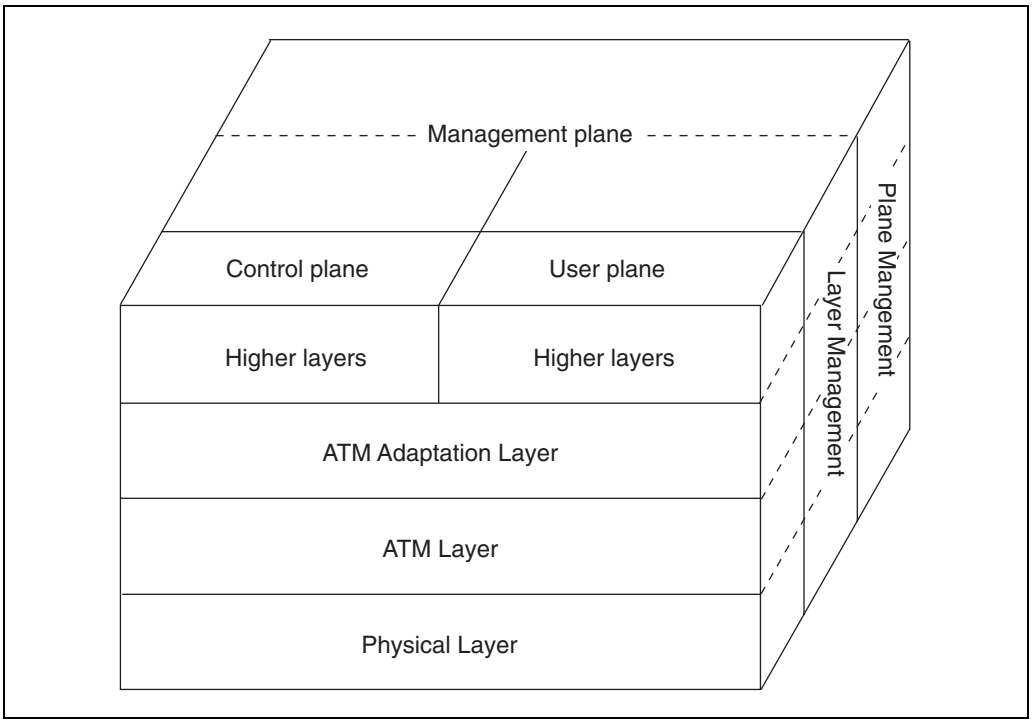

Figure 1–2 ATM B-ISDN Reference Model

of information that can be transmitted and received over a physical medium. TC functions include (1) transmission frame generation and recovery, (2) transmission frame adaptation, (3) cell delineation, (4) header error control (HEC) sequence generation and cell-header verification, and (5) cell-rate decoupling.

The transmission frame adaptation function performs the actions that are necessary to structure the cell flow according to the payload structure of the transmission frame (transmit direction) and to extract this cell flow out of the transmission frame (receive direction). In the United States, the transmission frame requires Synchronous Optical Network (SONET) envelopes for transmission rates higher than 45 Mbps.

SONET, a synchronous transmission structure, is often used for framing and synchronization at the PHY. The basic time unit of a SONET frame is 125 microseconds. The SONET frame structure is depicted in Figure 1–3. The basic building block of SONET is the synchronous transport signal level 1 (STS-1) with a bit rate of 51.84 Mbps. The STS-1 frame structure can be drawn as 90 columns and 9 rows of 8-bit bytes. The first 3 columns of STS-1 contain section and line overhead bytes used for error monitoring, system maintenance functions, synchronization, and identification of payload type. The remaining 87 columns and 9 rows are used to carry the STS-1 synchronous payload envelope. Higher-rate SONET signals are obtained by byte-interleaving *n* frame-aligned STS-1's to form an STS-*n* (e.g., STS-3 has a bit rate of 155.52 Mbps).

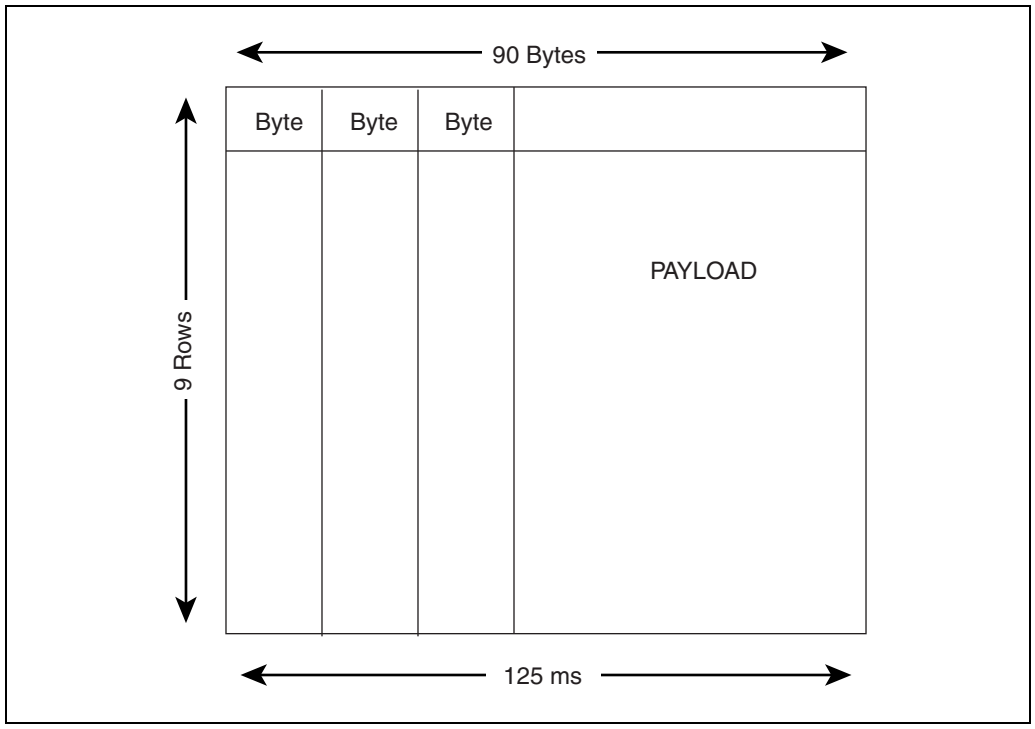

Figure 1–3 SONET STS-1 Payload Envelope

Cell delineation prepares the cell flow to enable the receiver side to recover cell boundaries. In the transmit direction, cell boundaries are identified and confirmed, and the cell flow is descrambled. The HEC mechanism covers the entire cell header, which is available to this layer by the time the cell is passed down to it. The code used for this function is capable of either single-bit correction or multiple-bit error detection. The transmitting side computes the HEC field value.

Due to the PHY framing overhead, the transfer capacity at the user-network interface (UNI) is 155.52 Mbps with a cell-fill capacity of 149.76 Mbps. Since the ATM cell has 5 bytes of overhead, the 48 bytes information field allows for a maximum of 135.631 Mbps of actual user information. A second UNI interface is defined at 622.08 Mbps with the service bit rate of approximately 600 Mbps. Access at these rates requires a fiber-based loop.

For information transport, an ATM uses virtual connections that are divided into two levels: the virtual-path level and the virtual-channel level. ATM layer functions include 1) generic flow control, 2) cell header generation and extraction, 3) cell virtual-path identifier/virtual-channel identifier (VPI/VCI) translation, and 4) cell multiplexing and demultiplexing.

The generic flow control (GFC) function is used only at the UNI. It may assist the customer network in controlling the cell flow towards the network, but it does not perform flow

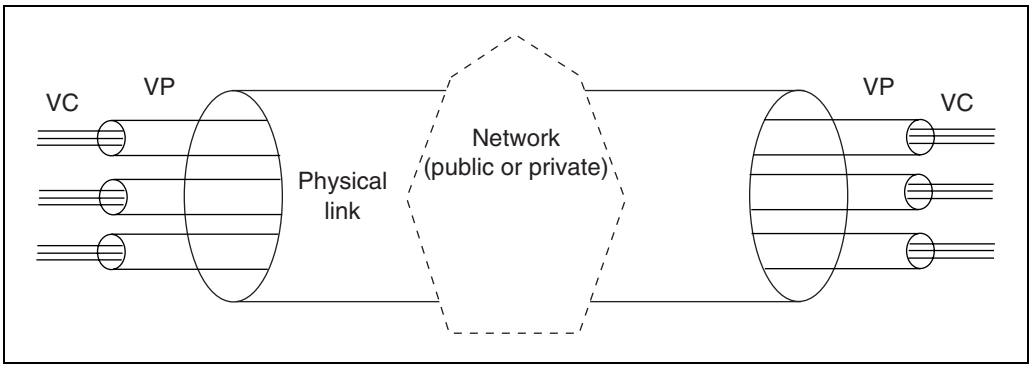

Figure 1–4 Relation Between VCs and VPs

control of traffic from the network. The GFC can also be used within a user's premises to share ATM capacity among the workstations.

Cell-header generation and extraction functions apply at points where the ATM layer is terminated. In the transmit direction, the cell-header generation function receives a cell-information field from a higher layer and generates an appropriate ATM cell header, except for the HEC sequence which is calculated and inserted by the PHY. In the receive direction, the cell-header extraction function removes the ATM cell header and passes the cell information field to the ATM adaptation layer.

The virtual-path identifiers (VPIs) and virtual-channel identifiers (VCIs) are the labels to identify a particular virtual path (VP) and virtual channel (VC) on the link. The switching node uses these values to identify a particular connection and then uses the routing table established at connection setup to route the cells to the appropriate output port. The switch changes the value of the VPI and VCI fields to the new values that are used on the output link.

From the transmitter, the cell-multiplexing function combines cells from the individual VP and VC into one cell flow. On the other hand, in the receive direction the cell demultiplexing function directs individual cells to the appropriate VP or VC. Figure 1–4 depicts the relationship of VPs and VCs.

Both VPIs and VCIs are used to route cells through the network. Note that VPI and VCI values must be unique on a specific transmission path. Each transmission path between network devices such as ATM switches uses VPIs and VCIs independently. This is illustrated in Figure 1–5. Each switch maps an incoming VPI and VCI to an outgoing VPI and VCI.

The ATM AAL performs the necessary mapping between the ATM layer and the next higher layer. The process is done at the terminal equipment or at the terminal adapter (i.e., at the edge of the ATM network).

The ATM network, the part of the network which processes the functions of the ATM layer, is independent from the telecommunications services it carries, which also means that the user payload is carried transparently by the ATM network. The ATM network does not process the user payload nor does it know the structure of the data unit. There is no timing relationship

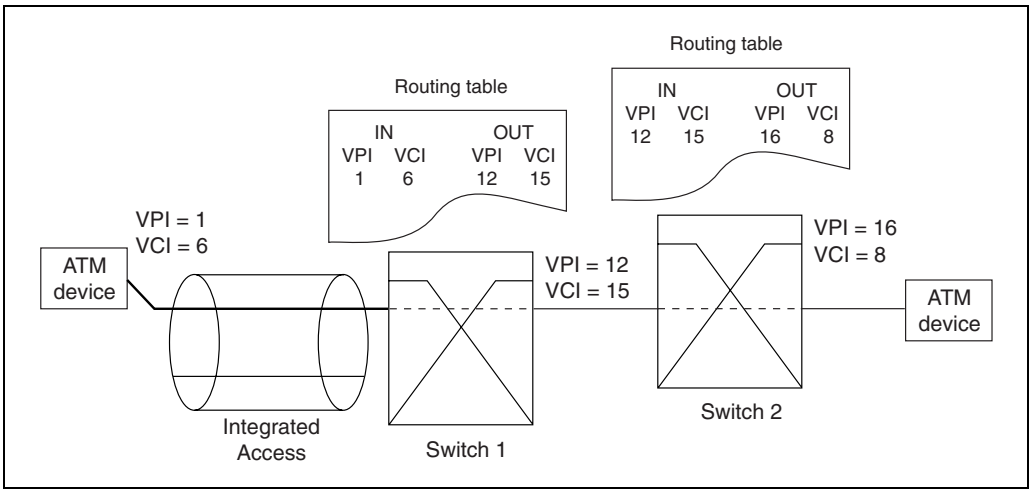

Figure 1–5 Illustration of VPI/VCI Usage

between the clock of application and the clock of network. Therefore, the network is also time independent. The consequence of this time independence infers that all time-dependence functions required by an application are only provided by services within the AAL.

The function of the AAL is to provide the data flow sent by the user to the upper layers at the receiving end by taking into account any effects introduced by the ATM layer. Within the ATM layer, the data flow can be corrupted by errors during the transmission, or it can suffer cell-delay variation as the result of variable delay in buffers, or through congestion in the network. The consequence is the loss of cells or the incorrect delivery of cells, which may have an impact on the application whether it is in data transfer, video, or voice communication. The AAL protocols must cope with these effects. For each telecommunication service, a separate AAL should be developed. However, considering the common factors within possible telecommunication services, it is possible to have a small set of AAL protocols to support these services.

The functions performed in the AAL depend upon the higher layer requirements. Since the AAL supports multiple protocols to fit the needs of different AAL service users, it is, therefore, service dependent. To minimize the number of different AAL protocols required, a telecommunication service classification is defined based on the following parameters: 1) timing relationship between source and destination, 2) bit-rate (constant or variable), and 3) connection mode (connection oriented or connectionless). Using these parameters, five classes of service have been defined. Figure 1–6 depicts different ATM service classes and AALs.

- Class A: Timing required, bit rate constant, connection oriented
- Class B: Timing required, bit rate variable, connection oriented
- Class C: Timing not required, bit rate variable, connection oriented
- Class D: Timing not required, bit rate variable, connectionless

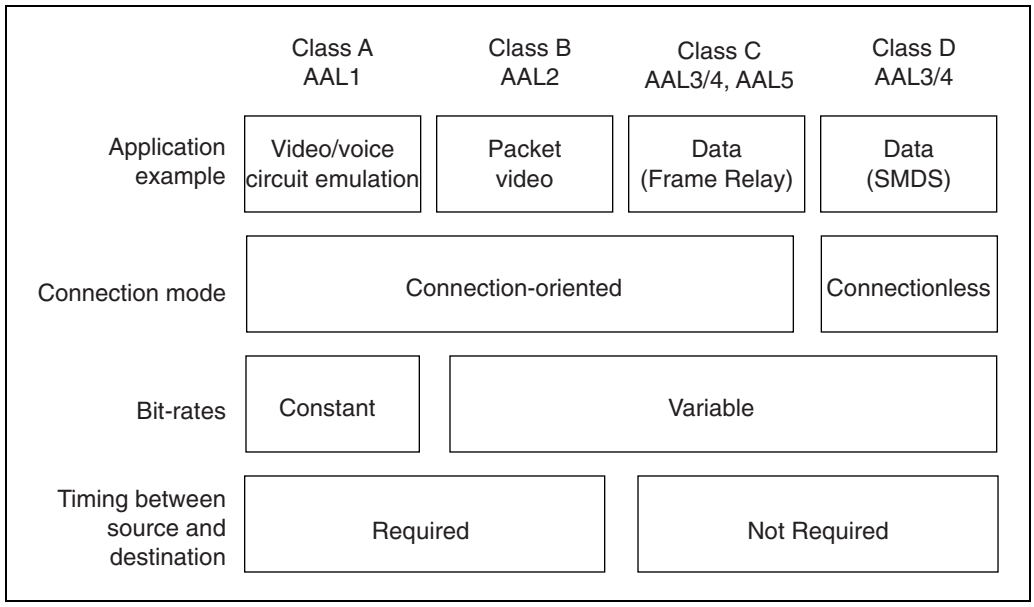

Figure 1–6 ATM Service Classes and AALs

• Class X: Unrestricted (bit rate variable, connection oriented or connectionless)

Class A service is an on-demand connection-oriented, constant bit-rate ATM transport service and has end-to-end timing requirements. Class A may require stringent cell loss, cell delay, and cell-delay variation performance. The user chooses the desired bandwidth and the appropriate QoS during the signaling phase of an SVC (simulated virtual connection) call to establish a Class A connection. (In the PVC (permanent virtual connection) case, this is negotiated in advance.) This service can provide the equivalent of a traditional, dedicated line and may be used for video distribution, video conferencing, multimedia, etc.

Class B service is not currently defined by formal agreements. Class C service is an on-demand, connection-oriented, variable-bit-rate ATM transport service and has no end-to-end timing requirements. The user chooses the desired bandwidth and QoS during the signaling phase of a call to establish a Class C connection. Class D service is a connectionless service and has no end-to-end timing requirements. The user supplies independent data units that are delivered by the network to the destination specified in the data unit. Switched multimegabit data service (SMDS) is an example of a Class D service. Class X service is an on-demand, connection-oriented ATM transport service where the traffic type can be either variable bit rate (VBR) or constant bit rate (CBR). Its timing requirements are user defined (i.e., transparent to the network). The user chooses only the desired bandwidth and QoS during the signaling phase of an SVC call to establish a Class X connection. (In the PVC case, it is negotiated in advance.)

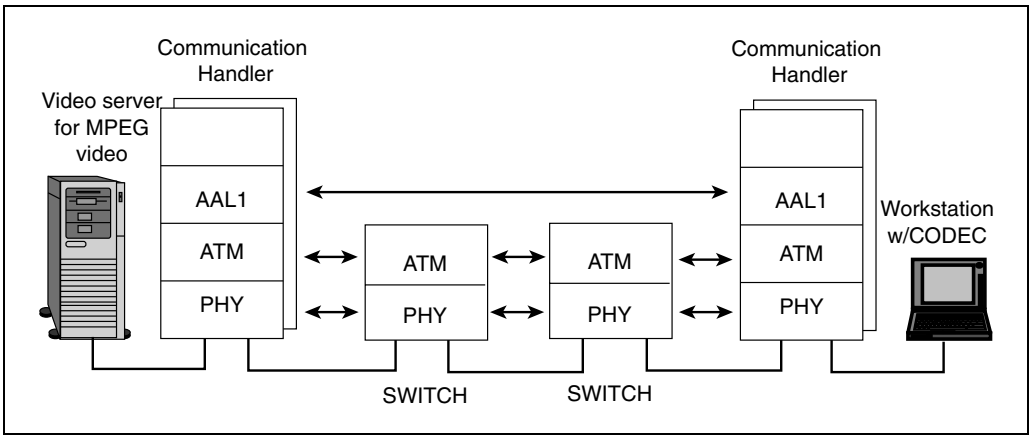

Figure 1–7 AAL Functionality over ATM Network

Figure 1–7 depicts the role played by AAL in end-to-end video transport. A communication handler can be implemented on top of AAL1 to support various video applications such as distribution of a motion picture experts group (MPEG) video from a video server to workstation clients over an ATM network. This communication handler provides video-application interface to establish video sessions according to a desired specification.

1.1.2 Leaky Bucket Model

The leaky bucket (LB) model is defined in the ATM standard to provide QoS to the user and also enforce the reserved rate from the user to the network. The LB model offers this QoS by guaranteeing each admitted call a constant reserved rate. A source is defined as conformance when it does not transmit beyond its reserved rate. The LB algorithm is the key in defining the meaning of conformance checking for an arriving cell stream against traffic parameters in the traffic contract. (A formal definition of the LB algorithm can be found in the ATM Forum UNI specification or International Telecommunications Union–Telecommunication (ITU-T) I.371.) According to the traffic contract, the LB is analogous to a bucket with a certain depth and a hole on the bottom which causes it to leak at a certain rate. Each cell arrival creates a cup of fluid flow that is poured into one or more buckets for use in conformance checking. The funneling of cell arrival fluid into buckets is controlled by the cell loss priority (CLP) bit in the ATM cell header.

Figures 1–8 and 1–9 show examples of a conforming and nonconforming cell flow, respectively, and also describe the LB's operation. In both examples, the nominal cell interarrival time is 4 cell times, which is also the bucket increment, with the bucket depth being 6 cell times.

Upon arrival of a cell, an agent checks whether the entire bucket increment for a cell can be added to the current bucket contents without overflowing. If the bucket does not overflow, then the cell is conforming; otherwise it is nonconforming. The agent discards the fluid for nonconforming cells to the floor. Fluid from a cell arrival is added to the bucket only if the cell is conforming. At each cell time, the bucket drains one unit and each cell arrival adds a number of

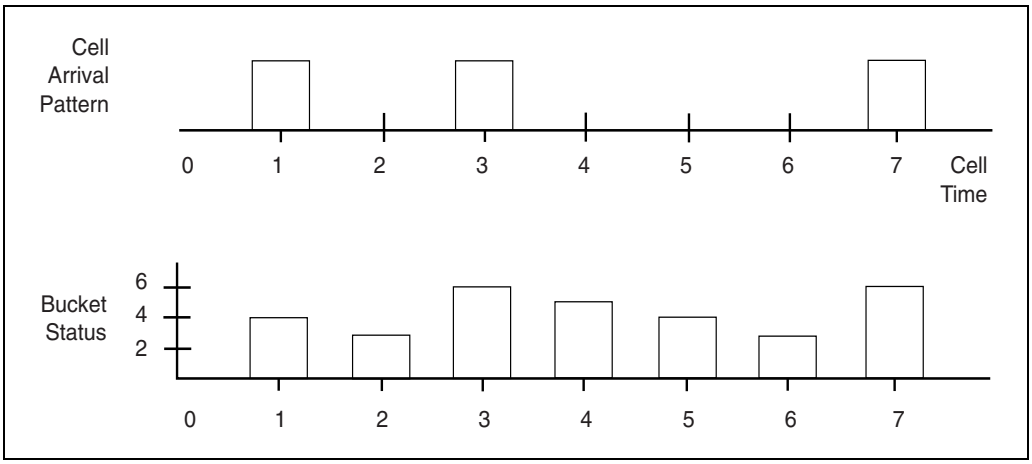

Figure 1–8 Example of Conforming Cell Flow

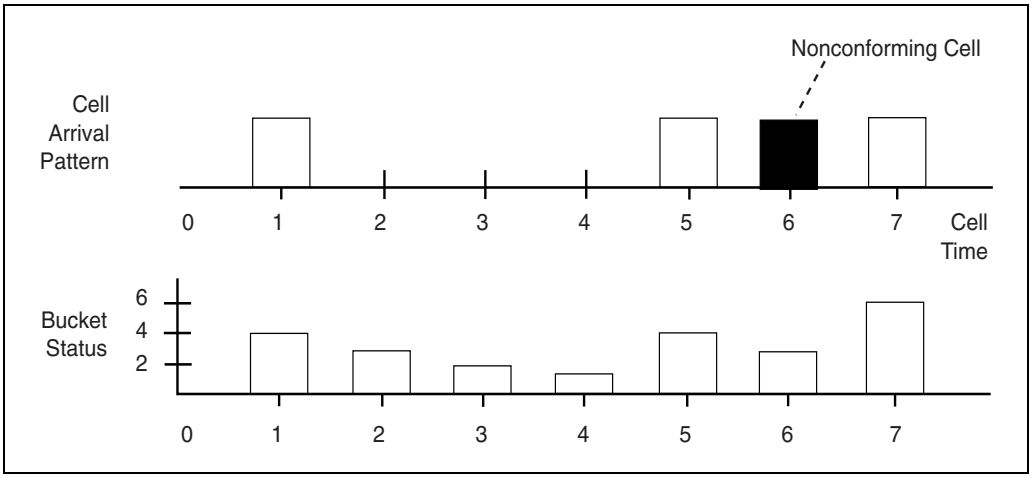

Figure 1–9 Example of Nonconforming Cell Flow

units to the bucket specified by the increment. A cell's fluid is completely drained out after a number of cell times. This depends on both the LB increment and the leak rate of the bucket.

For the conforming cell flow, an example is shown in Figure 1–8. The first cell arrival finds an empty bucket and fills it to a depth of four units. At the third cell time, two units have drained from the bucket, and a new cell arrives. The agent determines whether the fluid from this cell fills the bucket to the rim or to a depth of six units so that it is conforming and then added to the bucket. The earliest conforming cell arrival time is in the next four cell times or at the seven-cell time. The four increments must first be drained from the bucket for a cell arrival not to cause the depth of the bucket (equal to six units) to overflow.

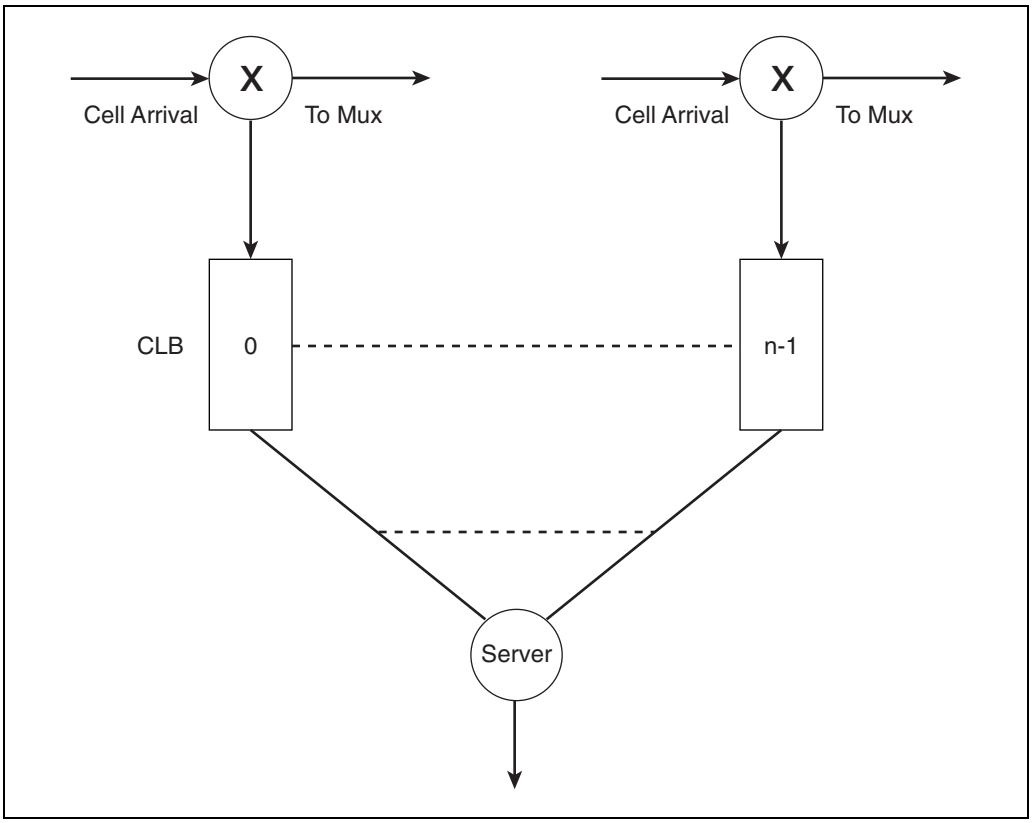

Figure 1–10 Cooperating Leaky Bucket Model

In the nonconforming cell flow example shown in Figure 1–9, the first cell arrival finds an empty bucket and fills it to a depth of four units. On the fifth-cell time, another cell arrives and fills the empty bucket with four increments of fluid. At the sixth-cell time, a cell arrives, and the agent determines that the bucket would overflow if the new arrival's fluid were to be added. Therefore, this definition determines that this cell is nonconforming. The agent then discards the fluid for this cell and this cell is considered to be lost. Since the fluid for the nonconforming cell was not added to the bucket, the next conforming cell can arrive at cell time 7, completely filling the bucket.

1.1.3 Cooperating Leaky Bucket Model

The LB mechanism ensures that the traffic rate entering the network cannot exceed the leak rate of the bucket. Nonbursty traffic is a good candidate for a mean-rate policing scheme such as the LB mechanism. However, as the burstiness of the traffic increases, the LB counter limit has to increase to accommodate the arrival of traffic bursts. There is no consensus on the relationship between the LB counter limit and the burstiness of the traffic source.

The cooperating leaky bucket (CLB) mechanism takes advantage of the fact that multiplexed traffic is much less bursty, but at the same time it makes certain of the fairness in the allocation of bandwidth. It provides fairness to all traffic sources that need additional rate by offering those traffic sources an equal amount of unused bandwidth. The implementation of CLB is similar to that of LB. An LB is located at the entrance to the network. The leak rate is set to the negotiated mean traffic rate and the counter limit is set based on the burstiness of the traffic. There are no standards on the counter limit or the burstiness of the traffic source. The major difference that distinguishes CLB from LB is that once a bucket becomes empty, its leak rate is distributed to other buckets of sources that require an additional rate. When there is no empty bucket, CLB works exactly the same as LB. However, if one or more empty buckets exist, CLB distributes the unused leak rate to other buckets. In the case of LB, the leak rate is wasted if the bucket becomes empty before the arrival of the next burst of cells.

Figure 1–10 depicts the CLB model. The service rate of the server is deterministic and it is set to n times the policed mean rate for identical sources, where n is the number of traffic sources. The traffic sources in this model do not necessarily need to have the same mean rate. The service discipline is round robin with one-limited service. According to this discipline, after serving queue k, the server will then select queue (k+1) modulo n for service. If the bucket selected is empty, the server immediately switches to the next bucket unit in which a nonempty bucket is selected. When a leaky bucket is empty, its allocated leak rate is evenly distributed to the other nonempty buckets.

1.1.4 Compression Technology

MPEG sources generate different traffic rates that stress different network requirements in terms of the expected throughput, delay, and frame loss. A careful analysis of these requirements is necessary in the development of transport models for supporting MPEG services. An MVC model for MPEG transmission demonstrates an efficient method to accommodate the bandwidth requirements of multiple sources. It is beneficial for the client to reserve just the average rate and to require higher rates from other mechanisms. At the same time, the service provider can accommodate more customers and maximize utilization of system resources.

The majority of traffic is contributed by video transmission due to its huge bandwidth demand. Video compression is essential to reduce the bandwidth needed. The compression standards are critically important to the widespread usage of digital video, digital–video distribution, video-on-demand, and other multimedia services. A number of different video compression standards such as Joint Picture Experts Group (JPEG), Motion Joint Picture Experts Group (MJPEG), Motion Picture Experts Group Standard 1 (MPEG-1), Motion Picture Experts Group Standard 2 (MPEG-2), digital video interactive (DVI/Indeo), and compact disk interactive (CD-I) are available. These compression standards have different characteristics and serve different applications. The following are examples of different applications for the JPEG: color facsimile, quality newspaper wirephoto transmission, desktop publishing, and medical imaging. As for DVI/Indeo and CD-I, they are primarily developed to work with CD-ROMs applications. ITU

H.261 is the widely used international video compression standard for video–conferencing markets. The video portion of Recommendation H.320 defines the technical requirements for narrowband visual telephone services for line transmission of nontelephone signals. Compared to other video compression standards, MPEG has a higher compression ratio, better image quality, better ability to handle fast motion pictures, and is more suitable for long-haul network transmission.

A number of proposed and tested ATM systems for video transmission primarily emphasize throughput and cell-loss rate. There are two shortcomings in those systems. First, the proposed and tested systems can only accommodate CBR sources or they require the subscribing sources to adhere to the initial reserved rate throughout the entire session. As a result, the quality of video transmission changes over time since the bandwidth requirement of the video source changes over time. Second, the proposed systems are not designed to understand that not all cells from MPEG transmission are of equal importance. This lack of ability to make a distinction between cells may cause the system to drop certain cells that are more important than others.

Haćand On (1998a) proposed a system to handle the bandwidth requirements needed by the VBR MPEG sources. It is sensitive to the relative importance of data from the MPEG sources. The system performs compression and selective multiplexing of transmission from MPEG sources. The concatenated traffic is then packetized into cells to a CBR channel. A CBR channel is used to transport the VBR streams as a concatenated group.

There are two basic advantages of this approach. First, compression and selective multiplexing of the MPEG sources are introduced to the system or to the transmitter side so that the decoders of the MPEG streams do not need modifications. Second, the contract between the system and the network is simple. The network is required to guarantee the delivery of all cells as long as the data rate of the combined streams does not exceed the CBR channel bandwidth such as optical carrier signal level 1, 3, and 12 (OC-1, OC-3, and OC-12).

The MPEG compression standard, ISO/IEC 11172, provides video coding for digital storage media with a rate of 2 Mbps or less. H.261 and MPEG-1 standards provide picture quality similar to that obtained with a VCR. Both standards are characterized by low-bit-rate coding and low spatial resolution. H.261 supports 352 pixels per line, 288 lines per frame, and 29.97 noninterlaced frames per second. MPEG-1 typically supports 352 pixels per line, 288 lines per frame, and 29.97 noninterlaced frames per second. The MPEG-1 compression ratio is about 100:1 and the data rate for the MPEG stream is in the range of 1.5 to 2 Mbps.

The International Standards Organization (ISO) MPEG working group has produced a specification to code combined video and audio information. The specification is directed to motion video display as compared to still image. MPEG specifies a decoder and data representation for retrieval of full-motion video information from digital storage media in the 1.5 to 2 Mbps range. The specification is composed of three parts: systems, video, and audio. The system part specifies a system coding layer for combining coded video and audio, and also provides the capability of combining private data streams and other streams that may be defined at a later date. The video part specifies the coded representation of video for digital storage media and

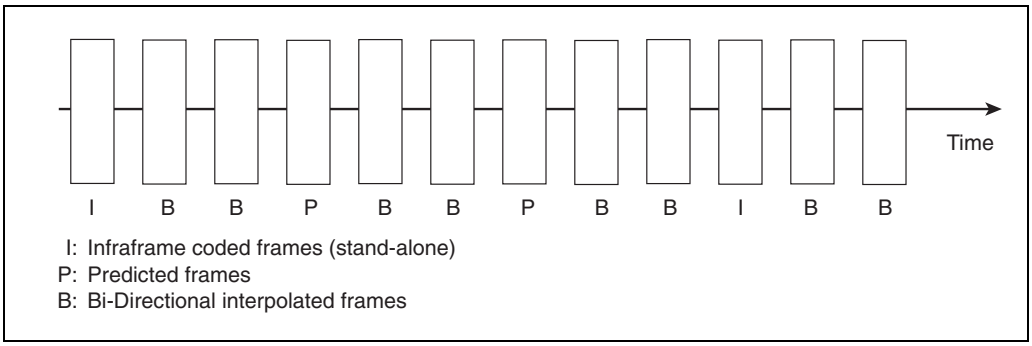

Figure 1–11 Group of MPEG Frames

specifies the decoding process. The audio part specifies the coded representation of high-quality audio for storage media and the method for decoding of high-quality audio signals.

The MPEG standard embodies the concepts of a group of frames and interpolated frames. Each MPEG stream contains frames that are intraframe coded to facilitate random access to different video scenes. Figure 1–11 shows an MPEG stream that consists of key intraframe coded frames or I frames, predicted frames or P frames, and interpolated frames or B frames. The encoding and decoding process falls into two main categories: intraframe and interframe coding. P and B frames use combinations of key motion-predicted and interpolated frames to achieve a high-compression ratio to accommodate the data rate of the transmission channel. This method compresses every frame of video individually. The intraframe or I frame coding offers the advantage of direct editing of key I frames; however, it produces 2 to 10 times more data than the interframe coding.

MPEG's system-coding layer specifies a multiplexing scheme for elementary streams of audio and video, with a syntax that includes data fields directly supporting synchronization of the elementary streams. Figure 1–12 depicts an MPEG encoder at the functional level. The video encoder receives uncoded digitized pictures called video presentation units (VPUs) at discrete time intervals. Similarly, at discrete time intervals, the audio digitizer receives uncoded digitized blocks of audio samples called audio presentation units (APUs). Note that the times of arrival of the VPUs do not necessarily align with the arrival time of the APUs.

The video and audio encoders produce coded pictures called video access units (VAUs) and coded audio called audio access units (AAUs). The outputs are referred to as elementary streams. The system encoder and multiplexer produce a multiplex stream, M(I), containing the elementary streams as well as system-layer coding.

MPEG-1 uses three types of frames: intrapicture (I) frames; predicted (P) frames; and bidirectional (B) frames. I-type frames are compressed using only the information within the frame by the Discrete Cosine Transform (DCT) algorithm. P frames are derived from the preceding I frames or P frames by predicting forward motion in time. B-interpolated frames are derived from the previous or next I or P frames within the stream. Reference [ISO] describes how the I, P, and B frames are coded according the MPEG specification. The B frames are required to

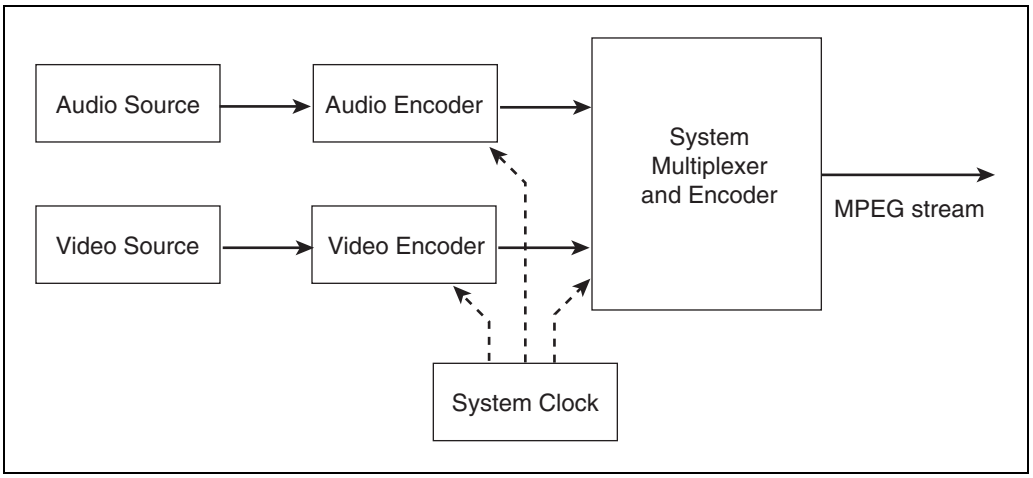

Figure 1–12 MPEG Encoder

achieve a low average data rate. The bandwidth allocated to each type of frame typically conforms to the ratio of 5:3:1 for I, P, and interpolated frames, respectively.

An important aspect of MPEG is synchronization, which is a fundamental issue for multimedia communication. In multimedia, synchronization means that various signal objects comprising the combined signal must be stored, retrieved, and transmitted with precise timing relationships. To achieve synchronization in multimedia systems that decode multiple video and audio signals originating from a storage or transmission medium, there must be a "time master" in the decoding process. MPEG does not specify which entity is the time master. The time master can be (1) any of the decoders, (2) the source stream, or (3) an external time base. All other entities in the system such as decoders and information sources must slave their timing to the master. If a decoder unit is taken as the time master, the time it shows a presentation unit is considered to be the correct time for the use of the other entities. Decoders can implement phase-locked loops or other timing means to ensure proper slaving of their operation to the time master. If the time base is an external entity, all of the decoders and the information sources must slave the timing to the external timing source.

The method of multiplexing the elementary streams of VAUs and AAUs is not directly specified in MPEG. However, there are some constraints that must be followed by an encoder and a multiplexer to produce a valid MPEG data stream. For example, it is required that the individual stream buffers must not overflow or underflow.

1.1.5 Multimedia Virtual Circuit Model

The Multimedia Virtual Circuit Model (MVC) consists of four components: virtual circuit negotiator (VCN), multimedia virtual circuit (MVC) queue, shared LB, and video multiplexing server (VMS) with available bit rate (ABR) pool.

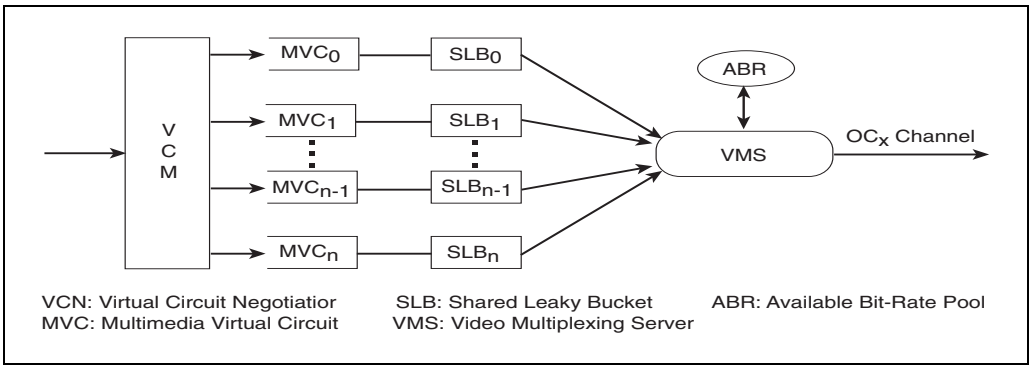

Figure 1–13 Multimedia Virtual Circuit Model

The VCN performs the call admission in the MVC model. The client establishes an MVC call with the VCN. Each client reserves an average rate with the VCN for each call. The VCN can admit a call into the system until the communication link capacity is exhausted. The reservation of an average rate for each MVC call enables this model to accommodate more clients compared to requiring each call to reserve the maximum rate. Figure 1–13 shows four components of the MVC model.

At discrete time intervals, each call submits its frames to the MVC queue waiting to be serviced and transmitted. The MVC stage receives digitized VPUs and digitized APUs from each call. It then performs the video and audio encoding for VPUs and APUs to VAUs and AAUs. The MVC stage performs MPEG video and audio encoding of each call according to the MPEG specification.

The shared leaky bucket (SLB) model is part of the MVC model. The implementation of SLB is similar to that of the LB where the SLB is set to the negotiated average rate. A major difference that distinguishes SLB from LB is that when a bucket becomes empty its leak rate is distributed to the ABR pool. The SLB principle shares the unused leak rate among each other and takes advantage of statistical characteristics of the traffic sources by sharing the unused leak rate with the needed sources. In LB, the leak rate is wasted when the bucket is empty since it does not share the unused leak rate with other LB. SLB works similarly to CLB except for the method it uses to distribute the unused leak rate. The CLB model distributes the unused leak rate evenly among other CLBs that require additional rates. The SLB model distributes the unused leak rate according to the first-come-first-served principle.

The use of the SLB can enforce the reserved rate for each MVC call. There are two options available when the MVC call requires additional rates. The first is to request an extra capacity from the VCN if unallocated capacities are available. The second is to borrow from the ABR pool resulting from the unused capacity from each MVC call. The ABR pool is a storage of the unused leak rate from each MVC call. The ABR pool size changes dynamically during each cycle or frame period and the unused capacity from the previous frame period does not accumulate to the next period. When an MVC call requires capacity beyond the average rate

during which no extra capacity is available from either the VCN or ABR pool, the service provider or telecommunication operator has to perform the frame quality scaling (FQS) method. The FQS method approach reduces the information content of a frame to fit into a particular transmission bandwidth.

The VMS can be represented by a polling system with x SLBs, where x is the number of sessions. The service rate of the VMS is deterministic and the service discipline is round robin. According to this discipline, each time after VMS served SLB k, the VMS will select SLB (k+1) modulo x for service. If the SLB selected is empty, the VMS will immediately switch to the next SLB until a nonempty SLB is selected. This polling model guarantees that the leak rate of each SLB is at least the reserved average rate when it is not empty. When an SLB is empty, its allocated leak rate is distributed to the ABR pool. This ABR pool is managed by the VMS unit. The SLB for a call negotiates with VMS for additional rates from the ABR pool when an MVC call requires an additional rate. When a frame from an MVC call requires more than the reserved rate and there is no rate available, the service provider uses the FQS method.

Data in an MPEG coded video stream are of different importance. In the MPEG compression scheme, the output data stream can be divided into macroblocks. Each macroblock has a certain number of bits or code words that can be dropped if necessary. Associated with each macroblock is a function relating to the number of code words or bits retained and the corresponding image quality. With each macroblock, code words or bits are ordered according to their significance, those of less significance are dropped first when necessary. The header information, Motion Vector (MV) and DC components of an MPEG frame are the most important. Among the DCT AC components, those of lower frequencies are more important than those of higher frequencies for two reasons. First, the amplitude square of the DCT AC components tends to decrease along the zig-zag scanning order. Second, human vision is less sensitive to the high-frequency signals. Figure 1–14 depicts the above ordering concept.

In the FQS method, let N be the total number of macroblocks collected from a frame. Let macroblock $\mathrm{MB}_i(D)$ be the number of code words or bits in macroblock MB_i that must be retained in order to maintain a distortion level of D. The goal is to find a distortion level D' such that

$$\mathrm{MB}_1(D') + \mathrm{MB}_2(D') + \ldots + \mathrm{MB}_N(D') = \mathrm{MB}_t \qquad (1\text{–}1)$$

Macroblock MB_t is defined as the transmission bandwidth available to transmit the N macroblocks within the frame. As a result, when MB_t is insufficient to transport all the code words or bits from all macroblocks, code words are dropped until the above equality can be achieved. The objective is for all macroblocks to achieve the same distortion level D.

When a frame from an MVC call requires an additional rate that is not available from unallocated capacity or ABR, the service provider reduces the number of code words from each macroblock. Each macroblock of the frame has equal quality relative to other macroblocks and at the same time scales down the information contents of the frame to fit into the available rate. This method avoids the dropping of frames when the available rate is insufficient to transmit the

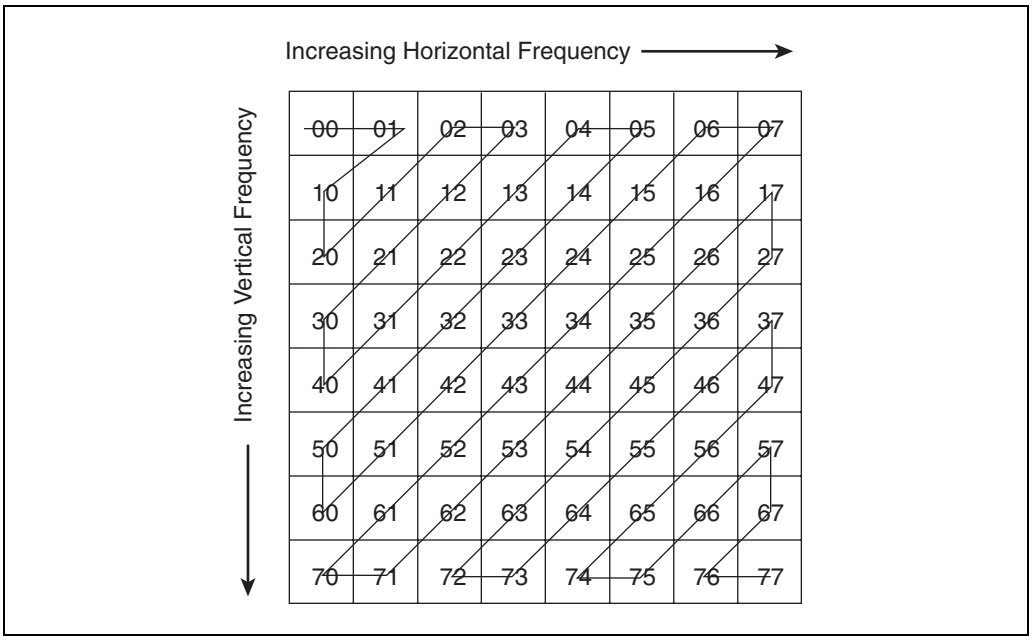

Figure 1–14 Zigzag Scanning Order of DCT Components

frame. It also smooths out the quality of the picture by extracting away only higher frequency code words from each macroblock.

The MVC model provides benefits to both the service provider and clients. The model is beneficial to the client since it is able to reserve just the average rate as well as acquiring more rates from other mechanisms. If the clients reserve the maximum rate, they have to pay for an idle reserved rate during the call session. The service provider can accommodate more customers and enhance the utilization of system resources.

Two key benefits of the MVC model for MPEG are statistical multiplexing and network simplicity. The MVC model employs statistical multiplexing of video sources since VBR sources have different degrees of activity during a connection. Many future ATM applications exhibit such behavior. Statistical multiplexing allows more sources to be admitted when not all VBR sources are expected to generate cells at their peak rates during the entire connection.

1.2 Rate-Based Congestion Control Schemes for ABR Service in ATM Networks

In ATM networks, traffic can be divided into two classes: guaranteed and best effort. Guaranteed traffic requires an explicit guarantee of service given by the network, as in CBR and VBR services. The limit on each connection's usable bandwidth is based on the notion of traffic contract. For these services, congestion control is administered through admission control and bandwidth allocation. However, the bandwidth requirements for data traffic are not likely to be known at

connection set-up time. Dynamically sharing the available bandwidth among all active users is required for this service, referred to as best-effort or ABR service, which has a highly bursty nature of data traffic.

1.2.1 Services in the ATM Network

The capability of ATM networks to provide a large bandwidth and to handle multiple QoS guarantees can be realized by preparing effective traffic management mechanisms. Traffic management includes congestion control, call admission control, and VP / VC routing. Essential for the stable and efficient operation of ATM networks is congestion control, which is performed between ATM end systems. The ATM end system is the point where an ATM connection is terminated, and the connection goes up to the ATM AAL. It is also defined as the point where VC connections are multiplexed, demultiplexed, or both. Therefore, each ATM segment can, depending on its characteristics, adopt a different congestion-control scheme.

An important issue in the selection of a congestion control scheme is the traffic pattern. In CBR and VBR services the traffic parameters are described, for example, in terms of such measures as peak-cell rate, cell-delay variation, sustainable-cell rate and burst-length tolerance. Once the connection request is admitted, the QoS is guaranteed throughout the session. The congestion control for CBR and VBR services is administered through admission control and bandwidth allocation. If the ATM network cannot deliver the resources demanded by the connection request, the request will be rejected at call set-up time. Voice and video are examples of sources that require guaranteed traffic service.

The remaining bandwidth not used by guaranteed bandwidth services must be shared fairly among all active users by using ABR services, or best-effort services. An example is data communication.

1.2.2 Congestion Control for ABR Service

To support ABR traffic, the network requires a feedback mechanism to inform each source how much data to send. The main feedback mechanisms are credit-based and rate-based flow control.

Credit-based flow control schemes make use of hop-by-hop feedback loops. Each link maintains its own independent control loop. The traffic moves across the network through a series of hop-by-hop feedback loops. The receiving end of each link issues “credits” to the transmitting end indicating the number of cells the transmitting station is allowed to send. Source end stations transmit only when they have permission from the network, as shown in Figure 1–15. In this approach, each link in the network runs the flow-control mechanism independently for each virtual circuit. A number of cell buffers are reserved for each virtual circuit at the receiving end of each link. One round-trip worth of cell buffers must be reserved for each connection. The amount of buffering required per connection depends on the propagation delay of the link and the required transmission rate of the virtual connection.

In high-speed Wide Area Network (WAN) based on ATM, the propagation delay is greater then the queuing delay. That is, it takes longer for the data to cross the link than for a switch or

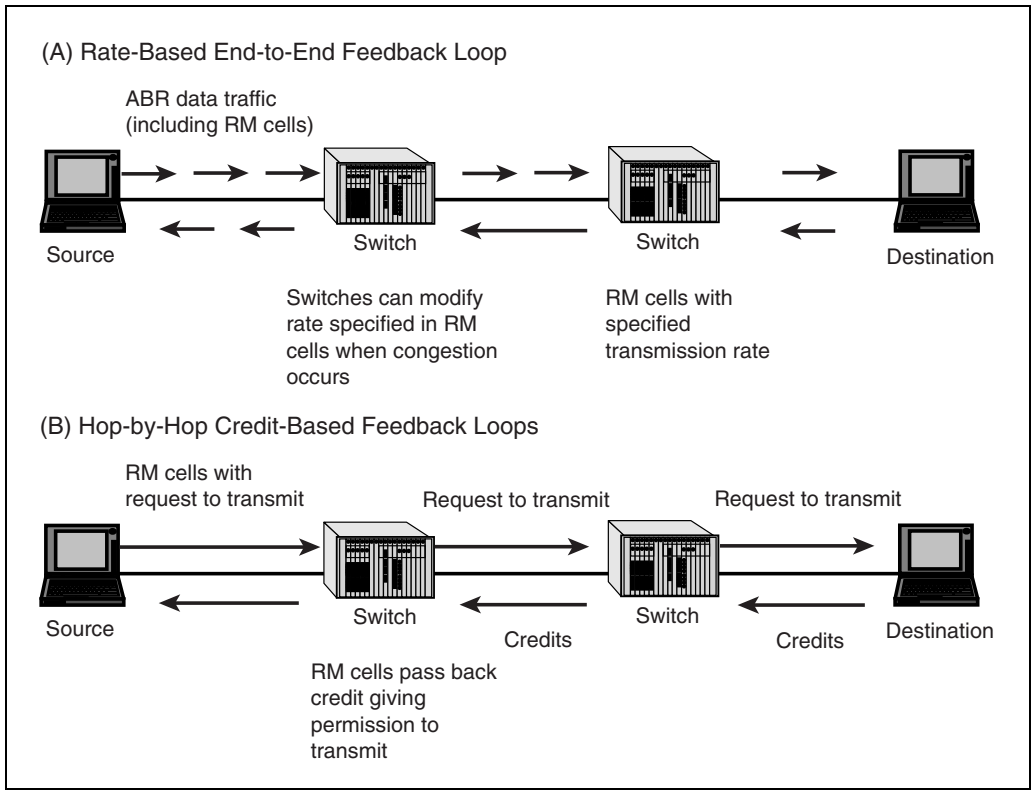

Figure 1–15 Rate and Credit Basics

the end station to process the transmission. Buffer queues fill up more quickly than the network can accommodate the traffic. The buffer sizes required to support a hop-by-hop credit scheme are impractical.

Rate-based flow-control schemes are end-to-end feedback mechanisms. They have one source and one destination station for each feedback loop, as shown in Figure 1–15. Within the feedback loop, the destination end alerts the source end to slow transmission when congestion occurs. If there are ATM switches between the loop's source and destination, these devices simply forward and augment the flow-control information moving between the destination and the source.

In this end-to-end rate-based scheme, a resource management (RM) cell is used, which is a standard 53-byte ATM cell used to transmit flow-control information. This RM cell carries information over the virtual circuit and is therefore allowed to flow all the way to the destination/source end station. The destination reflects or issues the RM cell, with an indicator to show the status of the traffic. The intermediate switches then simply forward or mark down the rate in the RM cell if needed. The source-end system then uses the information in the RM cells for subsequent transmissions until a new RM cell is received.

1.2.3 Rate-Based Control Schemes

There are rate-based congestion control schemes such as the forward explicit congestion notification (FECN) and backward explicit congestion notification (BECN) schemes, proportional rate control algorithm (PRCA), and intelligent congestion control schemes.

FECN and BECN Schemes. FECN schemes make use of the explicit forward congestion indication (EFCI) state carried in the payload type identifier (PTI) field to convey congestion information in the forward direction. When the switch becomes congested, it will mark in each VC the EFCI state of all cells being forwarded to the destination. Upon receiving marked cells, the destination sends an RM cell back to the source along the backward path. Then the source-end system must decrease its cell transmission rate accordingly on each VC. A time interval is defined at the destination-end system, and only one RM cell is allowed to be sent. The source-end system is also provided with an interval timer update interval (UI). When the timer expires without an RM cell received, the source recognizes no congestion in the network and increases the transmission rate.

BECN schemes use similar mechanisms except that the congestion notification is returned directly from the point of congestion to the source by the marked EFCI state of the cells. Thus, the response to congestion is faster.

These approaches require interval timers at the end systems and increase the complexity of the implementation. Furthermore, they use the negative feedback mechanism and could cause a collapse of the network in heavily congested conditions due to RM cells that are delayed or lost and the source increasing its cell emission rate because of the absence of RM cells.

PRCA Scheme. PRCA is based on a positive feedback rate-control paradigm. This shift in the rate-control paradigm is intended to remedy the problem of network congestion collapse encountered in FECN and BECN schemes. In PRCA, the source-end system marks the EFCI bit in all data cells except for the first cell of every NRM cells, where NRM is the number of data cells issued between two RM cells emitted. The destination-end system instantly sends an RM cell back to the source when it receives a cell with the EFCI bit cleared. If the EFCI bit is set by an intermediate switch because of congestion, the destination takes no action. By using this mechanism, receiving the RM cell implies that there is no congestion in the network, and therefore, the source-end system is given the opportunity to increase its rate. Otherwise, the source continuously decreases its transmission rate.

There are certain problems with PRCA. When VC is set up through more congested links, the EFCI bit of its data cells will be marked by the congested switches more often than those of other VCs set up through fewer congested links. Consequently, such VC will have a lower allowed cell rate (ACR) than other VCs. This undesirable effect of VC starvation is proportional to the number of congested links on which VC transmits its cells, and is referred to as the ACR beat-down problem.

Intelligent Control Scheme. In intelligent congestion control schemes, each queue of a switch maintains a variable modified ACR (MACR), which is the value of the estimated

optimal cell rate. The RM cells need to contain Current Cell Rate (CCR) and Explicit Rate (ER). Both CCR and ER are modified constantly by the intermediate switches according to the traffic in the network. Although this intelligent scheme shares resources more fairly than the other schemes, there is the expense of having interval timers and VC tables in every switch to detect the congestion status. This intelligent scheme also needs to maintain an MACR for each port in every switch, with the computation to estimate the best possible cell rate on each VC, which can increase the complexity of all switches. In addition, at a switch the effective rate of some connections that experience congestion at another switch can be quite different from the CCR values contained in RM cells; this can lead to misbehavior of the cell-rate control.

A rate-based control scheme proposed by Haćand Ma uses bandwidth fairly and efficiently for all switch connections under the proper congestion control and makes the complexity of switches low. In this scheme, all ATM switches in the network need to keep only one value shared ABR bandwidth (SB) for each output port. When an RM cell is received, the value of SB is modified, no matter to what path the RM cell belongs. Each end-system source modifies the issued cell rate, or ACR, according to the SB value carried in the backward RM cells.

1.3 Multiple Access Protocols for Wireless ATM Networks

Wireless ATM (WATM) is considered the framework for the next generation of wireless communication networks. ATM in the wireless environment has to cope with the low-speed and noisy, wireless medium. An appropriate medium access protocol that can efficiently utilize and share this limited frequency spectrum is essential.

ATM supports multimedia services at any speed from time-bounded voice and multimedia communications to bursty data traffic. Broadband wireless networks support traditional voice service as well as mobile communications with multimedia applications.

ATM networks are fixed (optical) point-to-point networks with high bandwidth and low error rates. These attributes are not associated with the limited bandwidth and error-prone radio medium. While increasing the number of cables (e.g., copper or fiber optics) can increase the bandwidth of wired networks, wireless telecommunications networks experience a more difficult task. Due to limited usable radio frequency, a wireless channel is an expensive resource in terms of bandwidth. For wireless networks to support high-speed networks like ATM, we need a new multiple-access approach for sharing this limited medium in a manner different from the narrowband, along with the means of supporting mobility and maintaining QoS guarantees.

Media Access Control (MAC) is a set of rules that attempt to efficiently share a communication channel among independent competing users. Each MAC uses a different media or multiple access scheme to allocate the limited bandwidth among multiple users. Many multiple-access protocols have been designed and analyzed both for wired and wireless networks. Each has its advantages and limitations based on the network environment and traffic. These schemes can be classified into three categories: fixed assignments, random access, and demand assignment. The demand assignment scheme is the most efficient access protocol for traffic of varying bit rate in the wireless environment.

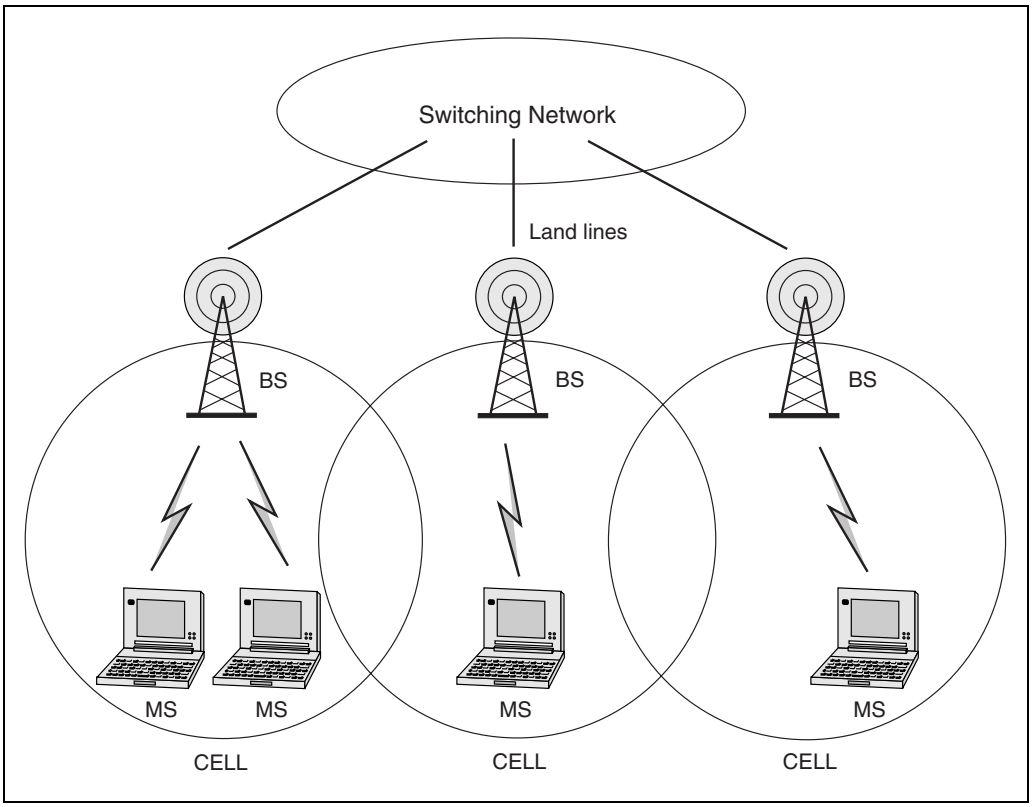

Figure 1–16 Wireless Network Architecture

1.3.1 Wireless Networks

The most widely employed wireless network topology is the cellular network. This network architecture is used in cellular telephone networks, personal communication networks, mobile data networks, and wireless local area networks (WLAN). In this network configuration, a service area, usually over a wide geographic area, is partitioned into smaller areas called cells (Figure 1–16). Each cell, in effect, is a centralized network, with a base station (BS) controlling all the communications to and from each mobile user in the cell. Each cell is assigned a group of discrete channels from the available frequency spectrum, usually a radio frequency. These channels are, in turn, assigned to each mobile user when needed.

Typically, BSs are connected to their switching networks using landlines through switches. The BS is the termination point of the user-to-network interface of a wireless cellular network. In addition, the BS also provides call set-ups, cell handoffs and various network-management tasks, depending on the type of network.

Due to the limited radio frequencies available for wireless communication, wireless networks have to maximize the overall capacity attainable within a given set of frequency channels.

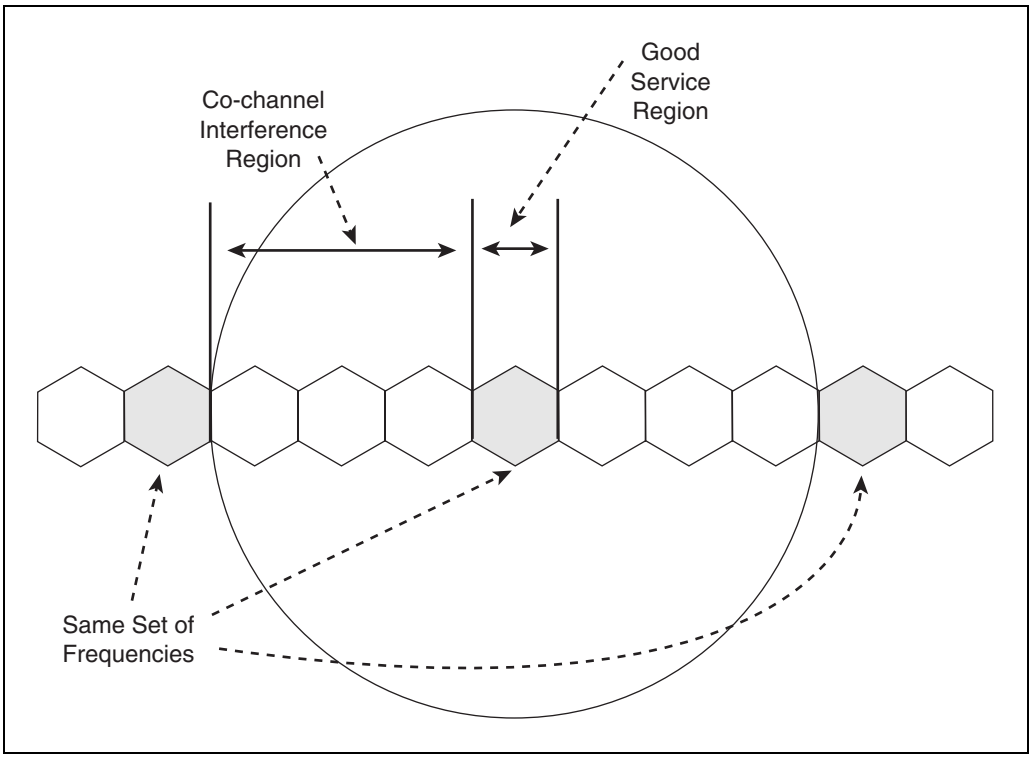

Figure 1–17 Illustration of Channel Reuse

Spectral efficiency describes the maximum number of calls that can be served in a given service area. To achieve high spectral efficiency, cellular networks are designed with frequency reuse, initially proposed by the Bell Telephone Laboratories™. If a channel with a specific frequency covers an area of a radius R, the same frequency can be reused to cover another area (Figure 1–17). A typical cellular service area using frequency reuse is shown in Figure 1–18. A service area is divided into 7-cell clusters. Each cell in the cluster, designated one through seven, uses a different set of frequencies. The same set of frequencies in each cell can be reused in the same service area if it is sufficiently apart from the current cell. Cells using the same frequency channels are called cocells. In principle, by using this layout scheme, the overall system capacity can be increased as large as desired by reducing the cell size, while controlling power levels to avoid cochannel interference. Cochannel interference is defined as the interference experienced by users operating in different cells using the same frequency channel. Smaller size cells called microcells are implemented to cover areas about the size of a city block. Research has been done on even smaller cells called picocells.

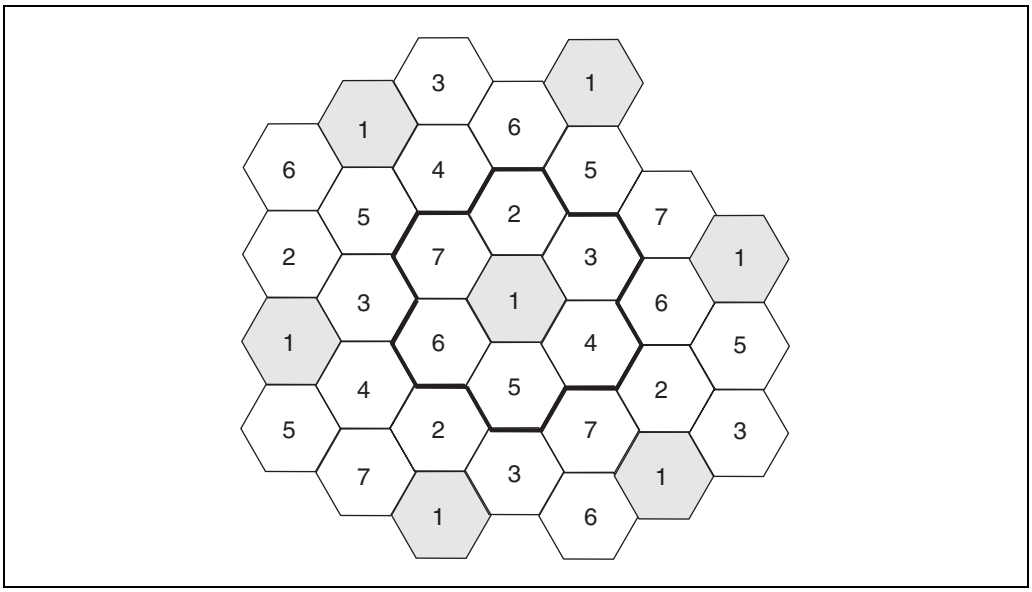

Figure 1–18 Typical Cellular Frequency Reuse Pattern, with Seven-Cell Clusters

1.3.2 Wireless ATM

1.3.2.1 ATM Services

Users request services from the ATM switch in terms of destination(s), traffic type(s), bit rate(s), and QoS. These requirements are usually grouped together and categorized in different ATM traffic classifications. The prototypical ATM services are categorized as follows.

- constant bit rate (CBR): Connection-oriented constant bit-rate service such as digital voice and video traffic.
- real-time variable bit rate (rt-VBR): Intended for real-time traffic from bursty sources such as compressed voice or video transmission.
- non-real-time variable bit rate (nrt-VBR): Intended for applications that have bursty traffic but do not require tight delay guarantee. This type of service is appropriate for connectionless data traffic.
- available bit rate (ABR): Intended for sources that accept time-varying available bandwidth. Users are only guaranteed a minimum cell rate (MCR). An example of such traffic is LAN emulation traffic.
- unspecified bit rate (UBR): Best-effort service that is intended for noncritical applications. It does not provide traffic-related service guarantees.

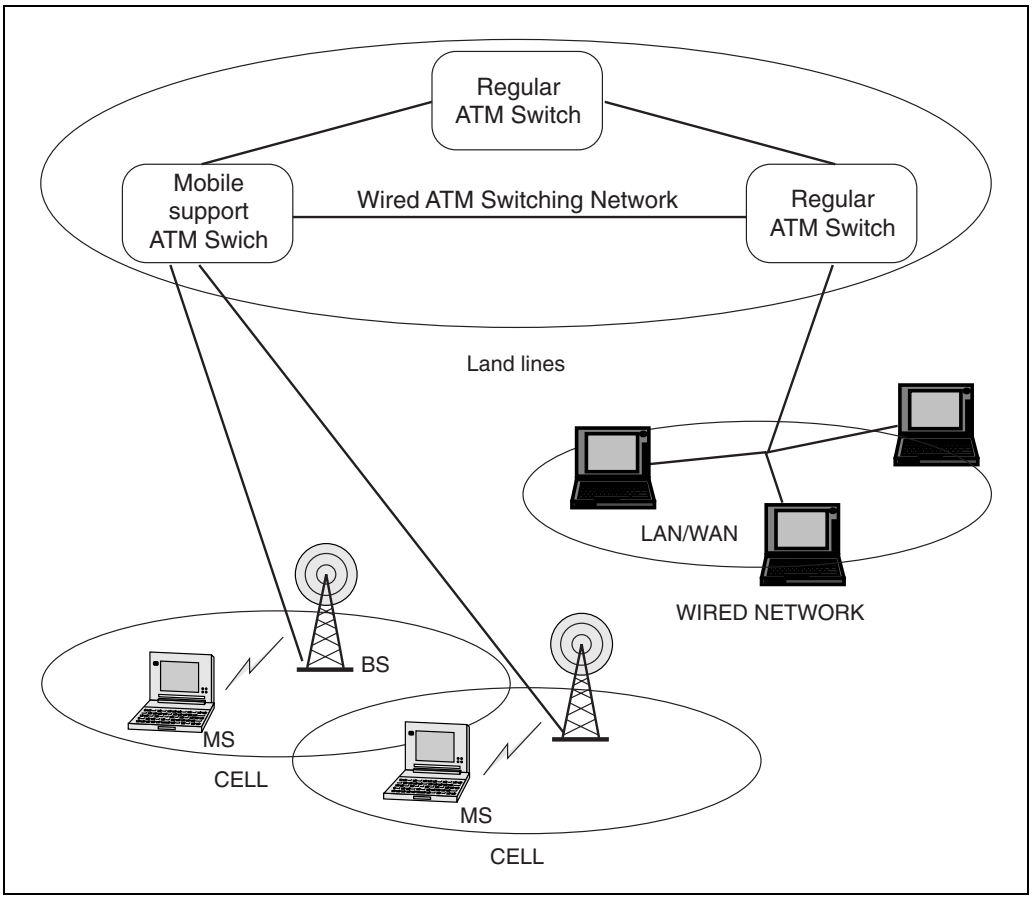

Figure 1–19 Wireless Integration in an ATM Network

1.3.2.2 Wireless Integration

The future of integrated multimedia networks will be dominated by broadband ATM. In addition to providing mobility, wireless ATM networks also allow flexible bandwidth allocation and QoS guarantees that existing wireless LAN is unable to provide. The wired/wireless integration is illustrated in Figure 1–19. Wireless networks are connected to wired networks using high-speed landlines through ATM switches.

This integration raises many serious compatibility issues. First, there is the issue of bandwidth. Wireless medium has a limited (e.g., maximum rate of about 34 Mb/s) and expensive bandwidth, while the ATM was designed for a bandwidth-rich environment. In addition, a wired ATM operates at a very low bit-error rate (BER), whereas wireless medium experiences a noisy and time-varying environment.

1.3.3 Multiple-Access Protocol

A multiple-access protocol is a scheme to control the access to a shared communication medium (a radio frequency in this case) among various users. Although many access protocols have been proposed and each has its advantages and limitations, very few are suitable for integrated wireless communications. Access protocols can be grouped according to the bandwidth allocation mechanism, which can be static or dynamic, according to the type of control mechanism implemented. Multiple-access protocols can be categorized into three classes: fixed assignment, random assignment, and demand assignment.

1.3.3.1 Fixed Assignment

Time-division multiple access (TDMA) and frequency-division multiple access are fixed assignment techniques that incorporate permanent subchannel assignments to each user. These 'traditional' schemes perform well with stream-type traffic such as voice, but are inappropriate for integrated multimedia traffic because of the radio channel spectrum utilization. In a fixed-assignment environment, a subchannel is wasted whenever the user has nothing to transmit. It is widely accepted that most services in the broadband environment are VBR service (e.g., bursty traffic). Such traffic wastes a lot of bandwidth in a fixed-assignment scheme.

1.3.3.2 Random Assignment

Typical random assignment protocols like ALOHA and carrier sense multiple access with collision detection (CSMA/CD) schemes are more efficient in servicing bursty traffic. These techniques allocate the full-channel capacity to a user for short periods on a random basis. These packet-oriented techniques dynamically allocate the channel to a user on a per packet basis.

Although there are a few versions of the ALOHA protocol, in its simplest form it allows users to transmit at will. Whenever two or more user transmissions overlap, a collision occurs and users have to retransmit after a random delay. The ALOHA protocol is inherently unstable due to the random delay. That is, there is a possibility that a transmission may be delayed for an infinite time. Various collision resolution algorithms were designed to stabilize and reduce contention in this scheme.

Slotted ALOHA is a simple modification of the ALOHA protocol. After a collision, instead of retransmitting at a random time, slotted ALOHA retransmits at a random time slot. Transmission can only be made at the beginning of a time slot. Obviously, this protocol is implemented in time-slotted systems. Slotted ALOHA is proven to be twice as efficient as a regular or pure ALOHA protocol.

CSMA/CD takes advantage of the short propagation delays between users in a typical LAN and provides a very high throughput protocol. In a plain CSMA protocol, users will not transmit unless the protocol senses that the transmission channel is idle. In CSMA/CD, the user also detects any collision that happens during a transmission. The combination provides a protocol that has high throughput and low delay. However, carrier sensing is a major problem for radio networks. The signal from the local transmitter will overload the receiver, disabling any attempts to sense remote transmission efficiently. Despite some advances in this area, sensing

still poses a problem due to severe channel fading in indoor environments. Similarly, collision detection proves to be a difficult task in wireless networks. While it can be easily done on a wired network by measuring the voltage level on a cable, sophisticated devices are required in wireless networks. Radio signals are dominated by the terminal's own signal over all other signals in the vicinity, preventing any efficient collision detection. To avoid this situation, a terminal transmitting antenna pattern has to be different from its receiving pattern. This requires sophisticated directional antennas and expensive amplifiers for both the BS and the mobile station (MS). Such requirements are not feasible for the low-powered mobile terminal end.

Code-division multiple access (CDMA) is a combination of both fixed and random assignment. CDMA has many advantages such as near zero channel-access delay, bandwidth efficiency, and excellent statistical multiplexing. However, it suffers from significant limitations such as limited transmission rate, complex BS, and problems related to the power of its transmission signal. The limitation in transmission rate is a significant drawback to using CDMA for integrated wireless networks.

1.3.3.3 Demand Assignment

In this protocol, channel capacity is assigned to users on a demand basis as needed. Demand assignment protocols typically involve two stages: a reservation stage where the user requests access, and a transmission stage where the actual data is transmitted. A small portion of the transmission channel, called the reservation subchannel, is used solely for users requesting permission to transmit data. Short reservation packets are sent to request channel time by using some simple multiple-access schemes, typically, TDMA or slotted ALOHA. Once channel time is reserved, data can be transmitted through the second subchannel contention free. Unlike a random-access protocol where collisions occur in the data transmission channel, in demand assignment protocols, collisions occur only in the small-capacity reservation subchannel.

This reservation technique allows demand-assignment protocols to avoid bandwidth waste due to collisions. In addition, unlike fixed-assignment schemes no channels are wasted whenever a VBR user enters an idle period. The assigned bandwidth will simply be allocated to another user requesting access. Due to these features, protocols based on demand-assignment techniques are most suitable for integrated-wireless networks.

Demand-assignment protocols can be classified into two categories based on the control scheme of the reservation and transmission stages. They can be either centralized or distributed. An example of a centralized controlled technique in demand assignment is polling. Each user is sequentially queried by the BS for transmission privileges. This scheme, however, relies heavily on the reliability of the centralized controller.

An alternative approach is to use distributed control, where MSs transmit based on information received from all the other MSs. Network information is transmitted through broadcast channels. Every user listens for reservation packets and performs the same distributed scheduling algorithm based on the information provided by the MS in the network. Requests for reservation are typically made using contention or fixed-assignment schemes.

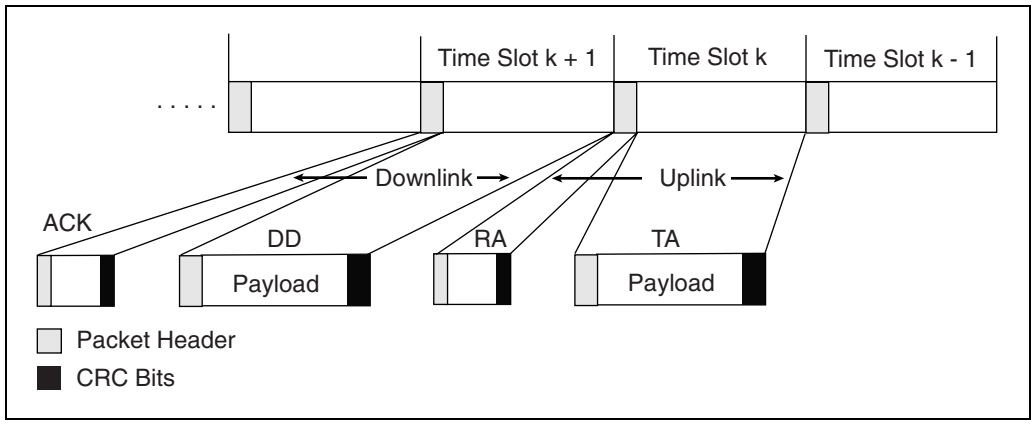

Figure 1–20 Radio Channel Classification: Slot-by-Slot

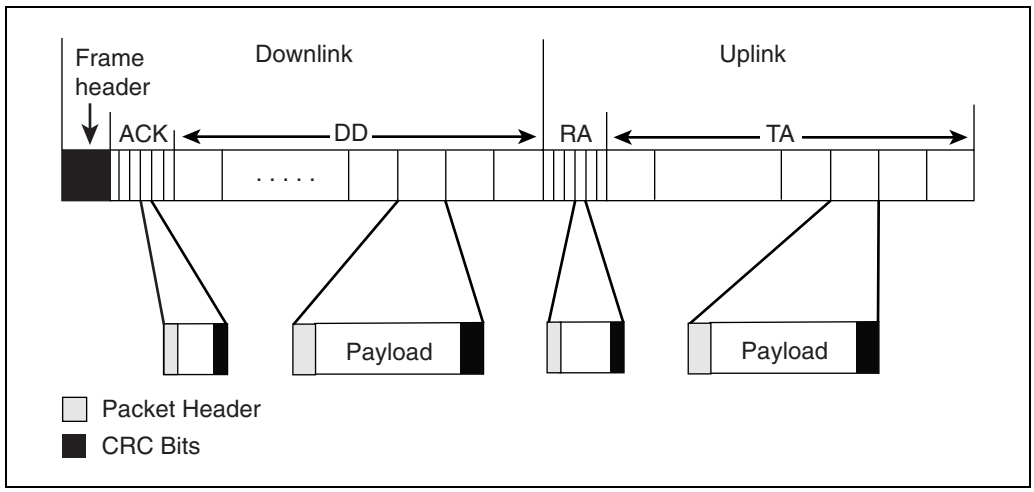

Figure 1–21 Radio Channel Classification: By Period

1.3.4 Demand Assignment Multiple-Access (DAMA) Protocols

Most DAMA protocols use time-slotted channels that are divided into frames. Depending on the transmission rate and the type of services, the channel bandwidth can be represented by a single or multiple frame(s). Each frame is divided into an uplink and a downlink period (i.e., channel). These periods are further divided into two subperiods or slots. They can be partitioned on a slot-by-slot or period basis (Figures 1–20 and 1–21, respectively). In the slot-by-slot method, each uplink and downlink period consist of a single time slot. In the method by period, the uplink and downlink period contains multiple time slots, encapsulated as a frame. The uplink and downlink communications can be physically separated using different frequency channels or dynamically shared using the time-division duplex (TDD) system.

The uplink channel (e.g., mobile-to-base) is divided into the request access (RA) and data transmission access (TA) subperiods. On the other hand, the downlink channel is divided into the acknowledgment (ACK) and the data downstream (DD) subperiods. A user requests bandwidth using the RA subperiods (i.e., uplink). When the BS hears a successful request such as no collision, it will notify the corresponding user through the ACK subperiods (i.e., downlink). Successful users are then assigned bandwidth, if available, in the TA subperiods. The DD subperiods are used by the BS to transmit downstream data to mobiles. These subperiods, also known as slots, vary in length depending on the type and amount of information they carry as determined by the protocol designer. The RA and ACK slots are much smaller than the data slots; hence, their time intervals are called minislots. Depending on the protocol, they may not have equal lengths.

DD transmissions are controlled by the BS and are performed contention free; typically using a time-division multiplexing (TDM) broadcast mode. These transmissions are performed with little delay and are not a crucial performance driver of the system.

In a wireless service area, the number of mobiles a BS covers is much larger than the available channel bandwidth. However, not all the mobiles are active simultaneously. Therefore, mobiles request access using random-access schemes in the RA subperiods. Access methods like ALOHA and its variations are usually used. On the other hand, variations of fixed-assignment schemes are typically used in the TA channel. The methods are the TDMA and CDMA schemes, of which TDMA is easier to implement.

Resource Auction Multiple Access (RAMA). The RAMA protocol was proposed as a fast resource or call-level assignment and handoff mechanism. RAMA is a deterministic DAMA protocol in which mobile users request access by transmitting their b-bit IDs, assigned during call set-up, on a symbol-by-symbol basis. For example, the b-bit ID could be a 10-digit number. Each digit represents a symbol and is transmitted one at a time. After each transmission, the BS acknowledges the symbol with the largest value and ignores the rest of the symbols. Mobiles that do not hear their symbol drop out of the auction. After 10 rounds of these transmission and acknowledgement processes, a single winning mobile remains (e.g., the mobile with the highest ID). The BS then assigns an available communication channel to the winner.

The RAMA protocol provides assignment at a rate that is suitable for many applications, including some statistical multiplexing of voice. For this protocol to be employed in ATM networks, packets need to be used. In that case, RAMA can be implemented on a slot-by-slot basis (i.e., each time slot identifying a mobile to transmit a packet). Like most DAMA, the transmission time slots are never wasted when any mobiles have packets to transmit. However, the overhead experienced during the auction phase is substantial. The total time required for a contention period (i.e., auction) is given by $t = K_m (2t_d + 2T_s)$, where K_m is the number of M-ary digits representing the mobile's ID, T_s is the symbol duration, and t_d is the guard interval after each symbol. The guard interval t_d must be sufficiently long to accommodate the on/off switching of transmitters and to allow for the processing and propagation delays between consecutive uplink and downlink transmissions.

Packet Reservation Multiple Access (PRMA). PRMA was proposed for packet-based voice transmission over wireless networks. PRMA is designed to increase the bandwidth efficiency over fixed assignment TDMA. In this protocol, time slots are grouped into frames and each timeslot is labeled as either reserved or available according to an acknowledgement from the BS at the end of each slot. When a mobile that is generating periodic traffic successfully transmits a packet in an available slot (using the slotted ALOHA protocol), the mobile also reserves that slot in the future frames. There is no subsequent contention with other mobiles in that slot until it is released. A mobile releases a time slot at the end of a burst by leaving the reserved slot empty. Packets from random traffic also contend for available time slots using slotted ALOHA. However, when a random packet is successfully transmitted, the time slot is not reserved in subsequent frames.

Although PRMA can be used in most packet-oriented networks, it proves to be inefficient for wireless ATM. PRMA has a variable-channel access delay (packet or burst level), which does not suit the requirements of the ATM services.

Distributed Queuing Request Update Multiple Access (DQRUMA). The DQRUMA protocol is designed specifically for data-packet (e.g., ATM) networks. It attempts to provide an efficient bandwidth-sharing scheme that can satisfy QoS parameters and support various types of ATM services. DQRUMA is designed for fixed-length packets (e.g., ATM cells) arriving at the mobile at some bursty random rate. The uplink and downlink periods are configured on a slot-by-slot basis. The uplink slot comprises a single data transmission slot (i.e., TA slot) and one or more RA minislots.

DQRUMA introduces the concept of a dynamic uplink slot where the uplink slot may be converted into a whole RA channel filled only with RA minislots. This conversion occurs when the BS senses that there is a significant amount of collisions in the RA channel. It allows as many as 25 RA minislots in a single time slot where mobiles can send their requests. This method drastically reduces request contention in the RA channel. Requests in the RA channel use a random access protocol like slotted ALOHA. The downlink slot contains the typical DD time slot, one or more ACK minislots, and a transmission permission, or perm slot. In situations where the uplink slot is converted into a series of RA minislots, the subsequent downlink slot is converted into a series of corresponding ACK minislots.

When a mobile terminal transmits its RA packet, it listens to the downlink slot for its ACK. ACK only indicates that the request has been received by the BS. Mobile users may not transmit their data until they hear their b-bit access ID in the perm slot. Upon hearing the transmission permission (b-bit ID), users may transmit their data in the next uplink time slot. This is the distributed queuing aspect of the protocol, where packets are queued at the mobile's buffer until the BS services them according to a scheduling policy (e.g., in a round-robin fashion).

DQRUMA also introduces an extra bit called the piggyback (PGBK) bit in the uplink channel. Each time a mobile transmits a packet (uplink), it also includes this PGBK bit to indicate whether it has more packets in the buffer. This bit serves as a contention-free transmission request for a mobile transmitting a packet. The BS checks this bit and updates the appropriate

entry in its request table in the BS, accordingly. When this bit is included, a mobile does not need a request for channel access in the following time slot. The BS knows that the mobile has more data to transmit and will assign a time slot to the mobile accordingly. This is the update portion of the protocol. The PGBK bit drastically reduces contention in the RA channel and greatly improves the overall protocol performance, especially for bursty traffic. Nevertheless, DQRUMA suffers from some channel-access delay problems due to its inability to distinguish different types of traffic.

RAMA performs like an ideal system where transmission time slots are never wasted when mobiles are transmitting. Unfortunately, it produces very significant overhead due to the auction process. Under certain reasonable ATM conditions, RAMA may produce an overhead of approximately 10%. Such substantial overhead makes it very impractical for packet-level access in the limited-bandwidth wireless environment. In addition, unlike DQRUMA where the BS is capable of controlling the order in which mobiles transmit their data based on some scheduling policy, RAMA is solely based on the value of their IDs (i.e., the largest value wins). This is not desirable in the ATM environment where there is an integrated mix of multimedia traffic with different priorities due to different QoS requirements.

PRMA offers a simpler strategy with little overhead and can be easily implemented in any packet network. However, a more sophisticated protocol like DQRUMA provides superior performance in several ways. PRMA wastes a single time slot at the end of each transmission burst by leaving the reserved slot empty (i.e., available slot). In addition, PRMA uses the entire uplink transmission channel for requests (slotted ALOHA), whereas DQRUMA uses only the RA channel. Therefore, when collisions occur, DQRUMA wastes less bandwidth (only a portion of the uplink channel). Finally, as in the comparison with RAMA, DQRUMA allows the BS to control the packet-transmission policy, allowing it to accommodate different types of traffic with various transmission rates. PRMA provides only a limited support for periodic traffic using a reserved packet.

DQRUMA's ability to reduce contention in the request channel with the PGBK bit and the dynamic RA channel conversion significantly improves its performance as a general multiple-access protocol. The flexibility of its BS control channel assignment offers a potential to support different traffic types like those in the ATM network.

Like most protocols, DQRUMA has its drawbacks. Although it is capable of supporting various type of services (e.g., CBR, VBR, etc.), it does not offer any distinction between different types of traffic during the reservation phase. Therefore, time-sensitive application that produces real-time CBR traffic is not given priority over ABR or UBR traffic. A more discriminating request scheme that could assign priority to requests from different traffic types is desired. While DQRUMA claims to support various service types, no detailed procedures for handling different services have been proposed. In addition, admitted but idle users (e.g., VBR) are required to perform a new request when resuming transmission. There are no distinctions between them and other newly arrived services. In order to distinguish between new and recurring VBR traffic, data packets have to carry additional header bits while increasing the complex-

ity of the scheduling policy in the BS. This is a difficult trade-off problem between the channel access delay and the protocol overhead (bandwidth and complexity). The traffic type in a specific network environment plays a significant role in such protocol design.

An Adaptive Request Channel Multiple Access (ARCMA) Protocol. ARCMA is a demand assignment multiple-access protocol with dynamic bandwidth allocation. Its basic architecture is modeled after the DQRUMA protocol. This scheme is designed to function in a cell-based wireless network with many MSs communicating with the BS of their particular cell. Transmissions are done on a slot-by-slot basis without any frames. As with DQRUMA, each slot is divided into a TA slot and a RA minislot. However, the RA channel in ARCMA is capable of carrying additional information for different classes of ATM service (e.g., CBR, VBR, etc.). This additional information is used by the BS to provide better QoS support for different classes of traffic. As in PRMA, transmission from CBR traffic may reserve an incremental series of slots in the duration of their transmission. No further request is needed until the CBR transmission finishes.

The BS maintains a request table to keep track of all successful requests, and assigns permission to mobiles for transmission at different time slots. In ARCMA protocol, the BS inspects the service class of a request and gives transmission priority to delay sensitive data (e.g., CBR). A piggyback bit is used in the uplink channel to reduce contention in the RA channel. This is especially beneficial for bursty traffic.

A dynamic RA channel similar to that of DQRUMA is used where an entire uplink channel can be converted into multiple RA channels. This conversion is done when the request table is empty, which in most cases indicates heavy collisions in the request channel. ARCMA uses a more complex algorithm that takes advantage of the random-access scheme in the RA channel. The slotted ALOHA with binary exponential backoff (BEB) is used as the random access protocol for ARCMA.

ARCMA improves the spectral efficiency by reducing collisions in the RA channel while improving support for the various classes of ATM services.