Baltzer Journals Received April 15, 1996

ATM via Satellite: A Framework and Implementation*

Anthony Hung[1], Marie-José Montpetit[2] and George Kesidis[1]

[1]*E.& C.E. Dept., University of Waterloo, Ontario, N2L 3G1, Canada,* E-mail:
`{ahung,kesidis}@odysseus.uwaterloo.ca`

[2]*Spar Aerospace Limited, now with Teledesic Corporation, 2300 Carillon Point, Kirkland, Washington, 98033, USA,* E-mail: `marie@teledesic.com`

This paper describes an ATM-based satellite network, focusing on the networking (ATM) aspects of the design. The ATM requirements and the basic design of the network are outlined. In particular, a simple MAC layer is proposed in which ATM service classes are mapped onto MF-TDMA uplink access methods. The uplink access and resource allocation approaches based on this model are described in detail. As well, this paper shows how different qualities-of-service can be provided by using a combination of different access schemes. This paper also covers scheduling for the uplink portion of the satellite network. The use of Hierarchical Round Robin is argued on the grounds of performance, flexibility and implementability.

1 Introduction

The Broadband Integrated Services Digital Network (B-ISDN) is intended to be the all-purpose network of the future. Primary B-ISDN services will include traditional communications, such as telephony and data, as well as emerging ones, such as multimedia. The market for communications is enormous and much of it will come from the home. To fulfill the large demand for these services, B-ISDN should be readily available at a reasonable cost to everyone.

To implement B-ISDN, the Asynchronous Transfer Mode (ATM) has been proposed as the means of transport. Recently, there has been a tremendous amount of research and development in the design and use of ATM for terrestrial networks with fiber-optic links, but the idea of an ATM implementation in a wireless environment, especially one using a satellite with onboard switching, is quite new. The attention paid to wireless ATM is growing however; a subcommittee has recently been formed within the ATM Forum to investigate wireless ATM.

ATM over satellite has a number of significant advantages over terrestrial ATM: a satellite can cover a large area, is immune to terrestrial disasters, can offer broadband links and can be accessed simply, quickly and at a relatively low cost with very small terminals (for more details see [15]). In the applications described here, the satellite uses the Ka frequency band to give ATM network access to a large number of small ground terminals. Each ground terminal should be capable of supporting multimedia services. In the system that we are considering, small ground terminals can offer 2 Mbps of payload capacity per terminal, which should be sufficient for a home user using multimedia services, although we are aiming at higher rates.

The networking aspect of this paper is concerned with the ATM layer of the B-ISDN protocol stack.

---

*This work was supported by the European Space Agency, Spar Aerospace Limited and NSERC of Canada

The ATM layer provides generally four main classes of service, three of which have *traffic descriptors* [3] which describe the intrinsic characteristics of the traffic source using that service. The sources are policed at the User Network Interface (UNI) to ensure compliance.

The first class is the Constant Bit Rate (CBR) service, which can emulate fixed-bandwidth allotment circuit-switching. It has a peak cell rate (PCR) traffic descriptor which is the bandwidth allocation of the connection.

From our point of view, the Variable Bit Rate (VBR) service allows connections to have a guaranteed rate of service, like CBR, but it also allows the connections to share network resources to take advantage of *statistical multiplexing* resource utilization gains. Three traffic descriptors are most commonly associated with VBR service: PCR, Sustainable Cell Rate (SCR) and Maximum Burst Size (MBS). These traffic descriptors are policed by the Generic Cell Rate Algorithm (GCRA) or "leaky bucket".

The Available Bit Rate (ABR) service being proposed has at least the following traffic descriptors: PCR and Minimum Cell Rate ($\mathrm{MCR} \geq 0$). Based on availability of network resources, a network flow control policy dynamically assigns each ABR connection an Allowed Cell Rate (ACR) satisfying $\mathrm{MCR} \leq \mathrm{ACR} \leq \mathrm{PCR}$. Delay requirements are not specified under ABR service. The ABR class is intended to support non-real-time communications such as data.

Unspecified Bit Rate (UBR) has no traffic descriptors and no quality-of-service (QoS) guarantees. UBR is intended only as a "best-effort" service.

The emphasis of this paper is to describe how a feasible satellite implementation can provide these classes of service efficiently. In view of the fact that satellite communication uses multiple access on a shared medium, a Medium Access Control (MAC) layer, which is not present in traditional ATM networks, would be needed. In this paper we will emphasize the design of a MAC layer which can integrate the satellite functions with the functions provided by the ATM service classes. Others [18] have looked at the implementation of access schemes for ATM-based mobile communications using cellular technology. These papers are typically application specific whereas our goal is to satisfy the more general ATM service classes.

This paper is organized as follows. §2 describes the characteristics of the satellite network that we are considering. §3 describes the MAC implementation. In particular, how to exploit statistical multiplexing gains is discussed in §3.4. In §4 we focus on uplink bandwidth allocation and scheduling: we give the implementation for a Hierarchical Round Robin (HRR) scheduler along with its buffer sizing and delay bounds. §5 concludes this paper with a summary.

2 The Satellite ATM Network

This section gives a high-level coverage of the physical layer implementation and a description of the available multiple access methods. It should be noted that the focus of this paper is on the networking and access schemes discussed in the later sections, and physical layer issues such as bit-error rates and the associated error control are only touched upon lightly because they generally do not affect the access design (given that they are reasonably small). Two papers which discuss satellite ATM networks from a more implementation-specific perspective are [4] and [17]. In the ATM network that we are considering, ground terminals communicate with each other via one satellite. We focus on ground terminals that are directly connected to end-users (typically the home user), although the terminals may also serve as a satellite access point for a LAN. Each terminal may support multiple ATM connections.

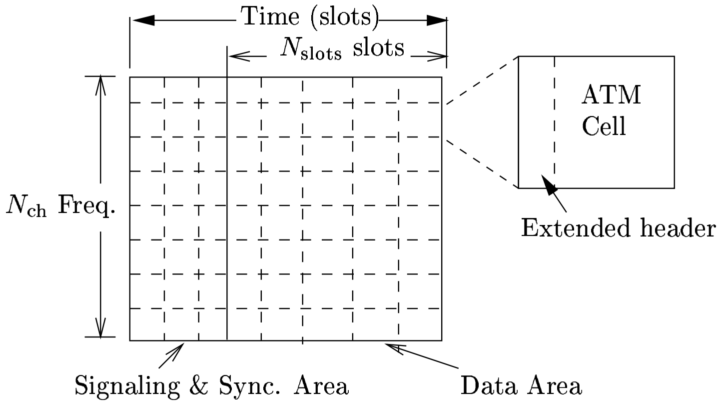

Figure 1: Format of uplink MF-TDMA frames

2.1 The Key Features

The ATM network under consideration uses a satellite which is in a geostationary orbit. A geostationary satellite can cover a large area on the globe, but it has the disadvantage that the one-way propagation delay between a ground terminal and the satellite is about 0.125s. This delay constraint poses challenges for delay-sensitive services. Therefore, transmission and access methods which minimize the added delay must be investigated.

One of the major challenges in the design of a satellite network is the limited transmission power of both the ground terminals and the satellite. Transmission in the satellite network should be designed in such a way that ground terminals at different geographical areas are given access in the most power efficient manner. For these reasons, multibeam systems are currently proposed for both the uplink and downlink. Multibeam systems require switching onboard the satellite because all traffic must be routed from an uplink beam to a downlink beam. This also usually means that queueing onboard is required.

To further save on uplink transmission power, we propose that the satellite ATM network use Multiple Frequency-Time Division Multiple Access (MF-TDMA) as the data link protocol. MF-TDMA has a number of attractive features, including the possibility of "on-demand" allocation of bandwidth. Our design of the MF-TDMA frame structure is shown in Figure 1. As shown in the figure, the MF-TDMA frame is divided into two areas, each of which has a set of fixed-size slots on which terminals may transmit. The signaling and synchronization area allow the terminal to request and receive the timing information necessary for its synchronization, as well as the sending of ATM and satellite signaling for connection establishment and initial entry (c.f. §2.4). The data area of the uplink frame is where ATM cells are transmitted. Each slot contains an ATM cell plus an extended header containing added forward error correction and in-band signaling.

Within the frame, each terminal may transmit on any one frequency at a given time (if given access). As mentioned earlier, the ATM cell payload capacity on each frequency in the data area is 2 Mbps; also, the system that we are considering would use $N_{\text{ch}} = 32$ frequencies. Hence, there are $N_{\text{ch}} \times 2 = 64$ Mbps total capacity.

A satellite network is controlled on the ground by a special terminal called the Master Control Station (MCS). Assigning functions to an MCS reduces the complexity of the satellite, which is desirable considering that 1) a commercial satellite is essentially irrecoverable, 2) components onboard are expensive and 3)

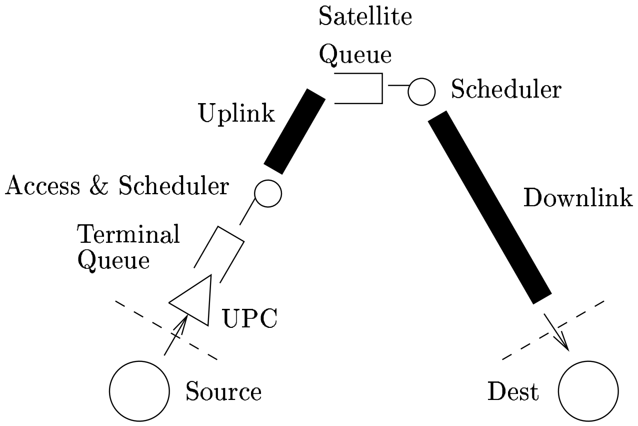

Figure 2: Queueing model of a connection

there is limited power onboard. The disadvantage of using an MCS is that there is an additional round-trip propagation delay required to exchange information with the satellite. So, considering these tradeoffs, the delay-sensitive functions should be put onboard; these include switching, queueing, flow control, and scheduling. The complexity of all processing onboard should be kept simple to reduce cost and power consumption. Two key functions that are not included in the above list are connection admission control and resource allocation, which may be handled either onboard or at the MCS depending on the complexity-delay tradeoff. All remaining (delay-tolerant) functions should be kept on the ground in the MCS; for example, billing and network maintenance.

2.2 A Two-Stage Model

Transmission on the uplink is done via a multiple access scheme that requires scheduling to multiplex ground terminal traffic onto a single uplink beam. In particular, the multiple access scheme, with the help of a scheduler, determines the number of slots in each MF-TDMA frame to assign to connections. This decision is based on the state of the traffic in the network, the resource allocation and the QoS requirements of connections.

Once the cells are on the satellite, they are routed from an uplink beam to a downlink beam. Therefore, the satellite network can be modeled as a network of two output-buffer switches (the choice of output-buffering onboard is assumed). The first switching stage consists of transmit queues on the ground terminal themselves and an uplink access scheme and scheduler residing on the front-end of the satellite. The second switching stage consists of the switch fabric residing on the satellite and the queues that are served by a downlink scheduler (these queues can be associated with output ports of the switch fabric residing on the satellite). This is shown in Figure 2. Each connection therefore sees two tandem queues and, possibly, a policing (usage parameter control (UPC)) device. Each queue of this figure shares its server with other such queues (not depicted) according to an access and/or scheduling policy. Note that the figure has the uplink access and scheduler on the ground, although they physically reside on the satellite because the satellite controls the output of the ground terminal queues.

Figure 3: Satellite-Specific ATM Sublayer

2.3 Methods of Uplink and Downlink Access

The key difference between a satellite ATM network and a point-to-point terrestrial ATM network is the fact that the satellite network uses multiple access while the terrestrial one does not. The choice of the multiple access scheme has a great impact on the performance of the satellite network. The primary goal in the assignment process is to satisfy the users' QoS and to maximize the utilization of the uplink. Once a cell from a connection has been chosen to be assigned to an MF-TDMA frame, it is the job of lower layer functions to choose a specific slot on the MF-TDMA frame for the cell. All these functions constitute the MAC implementation, as shown in Figure 3.

In this section, the term "connection" will be used to mean either a stream of satellite-specific signaling information that is exchanged between the satellite and a given terminal, or a stream of ATM signaling cells, or the data cells of a particular ATM connection. There are five specific uplink access (assignment) schemes to support the connections:

- Random access,
- Fixed assignment,
- Fixed-rate demand assignment,
- Variable-rate demand assignment and
- Free assignment.

The merits of these or similar schemes have been discussed in papers such as [13].

With random access, connections from different terminals may broadcast cells simultaneously resulting in "collisions" (corrupted data) in which case retransmission is necessary. Random access schemes, such as Aloha, obtain reasonable throughput only at low loads and they offer little in terms of performance guarantees. Hence, the use of random access is discouraged.

With fixed assignment, a terminal's connection is permanently assigned a constant number of slots per frame (or some multiple number of frames) for the lifetime of the terminal. This means that when the connection is idle, the slots are not utilized (wasted). The terminal-to-terminal delay when using this access scheme is the propagation delay (0.25s) plus the processing and queueing delays onboard.

Demand assignment (DA) allocates slots on an as-needed basis. There are two basic types of DA: fixed-rate and variable-rate DA. With the fixed-rate variety, a connection is assigned a fixed number of slots per

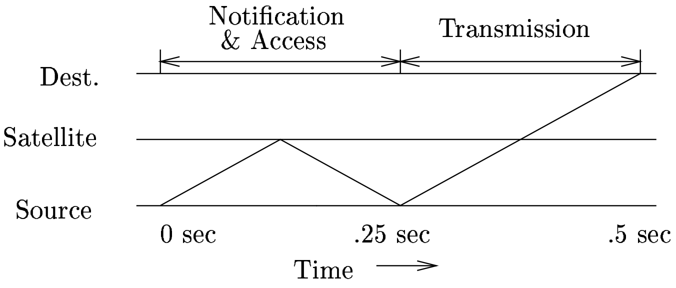

Figure 4: Transmission of a cell using variable-rate demand assignment

frame (or some multiple number of frames) for the duration of the connection. Like the fixed assignment scheme, if the source is idle, the slots will be wasted. So, the fixed-rate DA and fixed assignment schemes are the same except that one is for the duration of the connection while the other is for the lifetime of the terminal.

With the variable-rate DA scheme, slots are only assigned when it is known that there are cells awaiting service at the connection's terminal queue. This works as follows. When a cell arrives at the terminal queue, signaling messages are sent to the satellite notifying it of the arrival. When the satellite receives this information, it dynamically assigns slot(s) to the connection. Variable-rate DA is a guaranteed assignment scheme in that slots are assigned based on previously allocated resources (bandwidth), which is available for use whenever it is needed. It avoids collisions and efficiently uses the uplink capacity because the satellite is aware of the needs of the source and it responds to the need by assigning slots on a frame-by-frame basis. If the connection does not need a slot which has been allocated to it during connection establishment, the satellite may assign the slot to others. The drawback of this scheme is the 0.25s delay from when the signaling information is sent to when the satellite's response (i.e., assignment) is known at the ground terminal. The timing diagram showing the delays associated with variable-rate DA is given in Figure 4 assuming that the processing and queueing delays are small. As can be seen from the figure, the terminal-to-terminal delay is always at least 0.5s.

With variable-rate DA, some slots would be needed for transmitting notifications to the satellite. One way of doing this is to overlay this signaling information on the regular signaling and synchronization slots in an "out-of-band" signaling scheme. A second way, to adapt to the dynamic nature of the traffic, is to use the in-band signaling opportunities in the data slot headers.

Free assignment is concerned with the remaining slots in a frame which have not been assigned by the fixed or demand schemes. These remaining slots are the spare uplink capacity that the network can freely assign to connections in order to increase overall throughput, to relieve congestion at the ground terminal queues, or to reduce the terminal-to-terminal delay. Free assignment could use signaling information to determine the terminal queue's cell occupancy, like variable-rate DA.

On the downlink, transmission is multicast where each downlink beam is a multicast group. This multicast nature makes downlink transmission simple by the use of Time-Division-Multiplexing (TDM).

2.4 Signaling, Synchronization and Connection Establishment

Before a new ground terminal can transmit ATM cells to the satellite, there is a satellite-specific set-up procedure. Initially, when a new terminal is brought into the satellite network, the satellite is not aware of the terminal because it is unregistered. Such a terminal will first synchronize itself onto the uplink frames by listening to the downlink after which it can then register (or "log on") with the satellite. Any new terminal may send registration information to a specific portion of the signaling and synchronization area of Figure 1 that is reserved for this purpose; so, access on these slots is random. The received registration information on the satellite is not processed onboard; instead, it is transmitted down to the MCS. The MCS will acknowledge the registration with a return message. It is true that this registration procedure may be slow but delay is not critical at this point.

Once registered, a terminal synchronizes itself to the frames on the uplink with an internal clock. Over time, the terminal clock will drift away from the reference onboard the satellite and hence synchronization information must be exchanged periodically between the terminal and the satellite. Each terminal is allocated a special synchronization slot once every M frames to transmit synchronization information, which is in the synchronization area of Figure 1. Access for the synchronization slot is granted using fixed assignment. A set of M frames is called a multiframe. The value M should be large in order to increase the number of registered terminals in the network. This synchronization slot may also be used for transmitting other information, such as the out-of-band notifications discussed earlier.

Once a ground terminal is registered, it is capable of supporting one or more ATM connections. All connections begin with a connection establishment stage. The process of connection establishment involves negotiation between the user and the satellite and/or the MCS. This is done by a combination of satellite-specific and ATM signaling, both of which are transmitted in the synchronization slot. Once the connection is established, the satellite/MCS will instruct the satellite to allocate a certain amount of memory and bandwidth onboard to the connection. The satellite will use the bandwidth allocation as a basis for making uplink assignments.

In summary, in order for an unregistered terminal to actually transmit data cells on a data connection, it must do the following:

1. register with the MCS,
2. exchange signaling information with the satellite/MCS,
3. if demand assignment is used, send notifications to the satellite, and
4. transmit data cells.

3 MAC on the Uplink

We now discuss how uplink accesses can be assigned to support ATM given the uplink access methods of §2.3. Satellite access methods and the corresponding resource allocation will be briefly described for different kinds of ATM data connections and service classes. However, we will concentrate on providing CBR and VBR service. A related work on uplink access and performance evaluation for a satellite packet-switched network is [14].

Because signaling information must be exchanged regularly between the ground terminal and the satellite, the signaling and synchronization area of the MF-TDMA frame will use fixed assignment. It is also the scheme used to support ATM signaling cells since they are also in the signaling area. ATM data connections, on the other hand, are dynamic and so cannot use fixed assignment; they must use demand and free assignment.

In order to make the demand assignments (DAs) for an ATM data connection (which from here on will referred to as simply a "connection"), bandwidth must have been allocated to the connection during the connection establishment phase which is consistent with the DA schemes. Let ρ_i^{fx} cells/frame be the

Table 1
Resource allocation for ATM service classes

	ρ_i	ρ_i^{fx}	ρ_i^{v}	Free
CBR	PCR	PCR	—	—
VBR	SCR to PCR	†	†	√
VBR stat-mux	$\rho(n)/n$	‡	‡	√
ABR	MCR to PCR	—	MCR to PCR	√
UBR	0	—	—	√

† QoS dependent; see Figure 5.
‡ QoS dependent; see §3.4.

bandwidth allocated for fixed-rate DA to connection i in a particular uplink beam; i.e., there are ρ_i^{fx} slots assigned to connection i on each MF-TDMA frame for the duration of the connection. Let ρ_i^{v} cells/frame be the same connection's variable-rate DA allocation, the one that can use in-band and out-of-band signaling on a frame-by-frame basis. The assigned number of variable-rate demand slots in a frame will be less than or equal to ρ_i^{v} cells, depending on what the connection needs. Here, we assume that the values ρ_i^{fx} and ρ_i^{v} are integers to simplify matters — in §4 we describe how assignments are made in the case where ρ_i^{fx} and ρ_i^{v} are not integers. The quantity $\rho_i = \rho_i^{\mathrm{fx}} + \rho_i^{\mathrm{v}}$ will be called the total bandwidth allocation for connection i. A key constraint for bandwidth allocation is that

$$\sum_i \rho_i \leq c \tag{1}$$

where c cells/frame is the total data capacity of an uplink beam. That is, a new connection j with a bandwidth requirement of ρ_j cannot be admitted to the network if (1) will be violated.

Table 1 shows the use of demand and free assignment by different ATM classes. Also, it summarizes the bandwidth allocation for fixed-rate and variable-rate DA. Note that many connections would use a combination of access methods. We classify VBR connections into two categories: one utilizes statistical multiplexing gains by bandwidth sharing among connections while the other does not. Note that random access was not included because it cannot offer predictable QoS and reasonable throughput at high loads.

We now analyze the behavior of the terminal queue of Figure 2 given that a connection uses a combination of fixed-rate DA, variable-rate DA and free assignment. This analysis is based on a discrete-time model where the unit of time is the duration of the MF-TDMA frames. Let $B^{\mathrm{g}}(t)$ be the number of cells in the terminal queue of a particular connection at time t frames. Let $A(t)$ be the number of cell arrivals to that terminal queue at t. Let $S^{\mathrm{fx}}(t)$, $S^{\mathrm{v}}(t)$ and $S^{\mathrm{fr}}(t)$ be the number of slots assigned by fixed-rate DA, variable-rate DA and free assignment, respectively, to the connection on the uplink frame at time t. The occupancy of the terminal queue can be defined by the following recursion:

$$\begin{aligned} B^{\mathrm{g}}(t) &= \Big[B^{\mathrm{g}}(t-1) + A(t) - \\ &\qquad S^{\mathrm{fx}}(t) - S^{\mathrm{v}}(t) - S^{\mathrm{fr}}(t)\Big]^{+}, \end{aligned} \tag{2}$$

where t is an integer and $x^{+} = \max(x, 0)$. In the following analysis, we assume that the terminal queue size is large so that there is never overflow.

3.1 Using Fixed-Rate DA

CBR provides a circuit-switched type of service; hence a fixed number of slots per frame corresponding to the bandwidth demand is a natural choice for CBR services. Using fixed-rate DA has the additional advantage that there is no delay for access (also called access delay) after connection establishment; so, it may be used (alone or in conjunction with variable-rate DA) for delay-sensitive VBR connections as well. Resource allocation for CBR is simple: such a connection is given a fixed bandwidth allocation of $\rho^{\mathrm{fx}} = \mathrm{PCR}$. For VBR, the total bandwidth allocation of the connection, ρ, should be at least the SCR because the SCR is, by definition, the minimum bandwidth requirement. However, because the end-to-end delay of the connection is rather large, it may be necessary to give the connection a higher bandwidth than the SCR, but no more than the PCR.

The relationship between the bandwidth allocation and the fixed-rate DA at time t is clearly

$$S^{\mathrm{fx}}(t) = \rho^{\mathrm{fx}}. \tag{3}$$

Recall that ρ^{fx} is in units of cells/frame and it is an integer.

CBR and VBR connections may be bursty which will result in wasted bandwidth. Research is ongoing on means to signal empty slots without compromising QoS of existing connections.

3.2 Using Variable-Rate DA

As discussed previously, if a connection is only given variable-rate DA, then there is an access delay of 0.25s from when the terminal gets a burst of arrivals to when the satellite can respond by assigning more slots to the connection. This delay may be even larger if the terminal does not have an in-band or out-of-band signaling opportunity at every frame, in which case the burst would have to wait for the next signaling opportunity to come along. In this paper, we will assume that all connections can signal in every frame to simplify matters. Note that this delay means that terminal queues for the connections using only variable-rate DA should be sized appropriately to store at least the 0.25s worth of arrivals; this is called the bandwidth-delay product [16].

The concept of using a variable number of slots per frame for a connection can be extended to delay-sensitive connections which can tolerate a delay slightly longer than the one guaranteed by fixed-rate DA. Such connections may use variable-rate DA in conjunction with fixed-rate DA.

In order for a connection to use variable-rate DA, the satellite must be aware of the occupancy of its terminal queue. To send this information to the satellite, the connection notifies the satellite — either by in-band or out-of-band signaling — of the number of cell arrivals in the terminal queue since the last signaling message was sent. The satellite uses this information and the past assignments to calculate the variable-rate DA. The following shows how this is done.

Let Π be the round-trip propagation delay to the satellite, in terms of MF-TDMA frames; i.e., there are Π frames every 0.25 second. Assuming that each frequency has $N_{\mathrm{slots}} = 32$ ATM (48 byte payload) slots and a payload capacity of 2 Mbps, the value for the frame duration would be $(N_{\mathrm{slots}} \times 48 \times 8)$bits/2Mbps=5.9ms, in which case $\Pi = 42$. At time $(t - \Pi/2)$ frames, the satellite makes the assignments for the uplink frame that will be transmitted at time t ($\Pi/2$ frames is the propagation delay to transmit the assignments to the ground terminals). Ideally, in order to make the assignments for a connection for the uplink frame at time t, the satellite must know the value of $B^{\mathrm{g}}(t-1)$, which is the occupancy of the terminal queue just prior to time t. The value of $B^{\mathrm{g}}(t-1)$ can be derived by (2) only if $A(s), S^{\mathrm{fx}}(s), S^{\mathrm{v}}(s)$ and $S^{\mathrm{fr}}(s)$ are known for all $s \in \{0, 1, 2, ..., t-1\}$. However, the satellite is only aware of $\{A(s)\}_{s=0}^{t-\Pi}$ because only these assignments were made on the satellite at or before time $t - \Pi/2$. In other words, the satellite is not aware of $\{A(s)\}_{s=t-\Pi+1}^{t}$ at time t due to the propagation delay of $\Pi/2$ frames needed to transmit the signaling information.

So, in order to make use of the information at hand, the satellite can assume the extreme case that $A(s) = 0$ for all $s > t - \Pi$, which would give a lower bound on the value of $B^{\mathrm{g}}(t-1)$. Let $\widehat{B}^{\mathrm{g}}(t-1)$ be this lower bound; i.e., $\widehat{B}^{\mathrm{g}}(t-1)$ is the number of cells in the terminal queue at $t-1$ given that $A(s) = 0$ for all $s > t - \Pi$. From (2) we get that

$$\widehat{B}^{\mathrm{g}}(t-1) = \Big[B^{\mathrm{g}}(t-\Pi) - \sum_{s=t-\Pi+1}^{t-1} \big(S^{\mathrm{fx}}(s) + S^{\mathrm{v}}(s) + S^{\mathrm{fr}}(s)\big)\Big]^{+}. \tag{4}$$

The satellite then uses $\widehat{B}^{\mathrm{g}}(t-1)$ and the ATM traffic contract to make decisions regarding slot assignments at frame t. This is due to the fact that the assignments are made based on what is known about the terminal, and when making the assignments for frame t, the satellite only knows that the queue occupancy is at least $\widehat{B}^{\mathrm{g}}(t-1)$.

Recall that ρ^{v} cells/frame is the maximum bandwidth for variable-rate DA; so, ρ^{v} is the maximum number of variable-rate demand slots that can be assigned to the connection in a frame. The satellite can assign the variable-rate demand slots to serve the remaining backlog in the terminal queue after the fixed-rate demand assignments are made; i.e.,

$$S^{\mathrm{v}}(t) = \min\left[\left(\widehat{B}^{\mathrm{g}}(t-1) - S^{\mathrm{fx}}(t)\right)^{+}, \rho^{\mathrm{v}}\right]. \tag{5}$$

Note that by using $\widehat{B}^{\mathrm{g}}(t-1)$ in (5), an assigned demand slot is never wasted.

Because of the access delay of variable-rate DA, it may be desirable to use fixed-rate DA in conjunction with variable-rate DA for delay-sensitive VBR connections. Assuming that a connection's total bandwidth allotment ρ is fixed, the resource allocation scheme determines how to partition ρ into fixed-rate and variable-rate components; i.e., $\rho = \rho^{\mathrm{fx}} + \rho^{\mathrm{v}}$. The terminal-to-terminal delay of the connection depends heavily on how ρ is thus partitioned: basically, the smaller the value of ρ^{v}, the closer the cell delay is to 0.25s and, conversely, the larger the value of ρ^{v}, the closer the cell delay is to 0.5s. To illustrate this, we simulated the assignment algorithm given by (3), (4) and (5) without free assignment (i.e., $S^{\mathrm{fr}}(t) = 0$ for all t) for a monochrome MPEG-1 trace of the movie "Star Wars"[1] which is transmitted as real-time video. Assuming this video trace is transmitted VBR at the encoding rate, the mean rate of this particular trace would be 0.36 Mbps. We set $\rho = 0.9$ Mbps and an MF-TDMA frame duration to be 5.9ms. The average delay of the data cells to the satellite as a function of the bandwidth partition was calculated and the results are plotted in Figure 5. That is, given a fixed total bandwidth allotment ρ, Figure 5 illustrates how QoS (delay) can change as a function of ρ^{v}.

The ABR service class is still in the stages of standardization. We are primarily studying variable-rate DA schemes for ABR service because some ABR schemes currently being proposed (e.g., [16]) use explicit-rate flow control and it is simple to implement in the satellite network.

3.3 Using Free Assignment

All classes of traffic that require variable rates of service can use free assignment. VBR connections would use free assignments to increase QoS by reducing cell loss and delay. ABR connections utilize spare bandwidth in the network; so, they can certainly make use of free assignment to increase throughput. The UBR class

---

[1] This trace was provided by Dr. M. Garrett and Prof. M. Vetterli. It can be obtained by anonymous ftp from thumper.bellcore.com

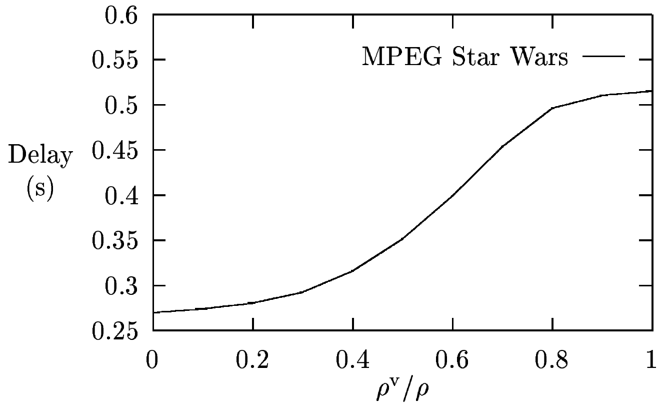

Figure 5: Terminal-to-terminal delay vs. bandwidth allocation

will use free assignment as its only method of access, but it has the lowest priority when being given free assignment slots.

After fixed-rate and variable-rate DAs have been determined for a particular frame, free assignment is used to determine which connections get the remaining data slots of that frame. The purpose of this is to improve the QoS of connections (i.e., relieve congestion) and to increase overall throughput. The amount of free assignment given to a connection is also based on factors such as its QoS requirement and the occupancy of its terminal queue.

Free assignments may be given to VBR connections to relieve congestion in the following manner. The satellite knows that the terminal occupancy after the fixed-rate and variable-rate demand assignments is at least

$$\widehat{B}^{\mathrm{g}}(t-1) - S^{\mathrm{fx}}(t) - S^{\mathrm{v}}(t),$$

but it is actually

$$B^{\mathrm{g}}(t-1) + A(t) - S^{\mathrm{fx}}(t) - S^{\mathrm{v}}(t).$$

Possibly, the difference between these two values, i.e.,

$$B^{\mathrm{g}}(t-1) + A(t) - \widehat{B}^{\mathrm{g}}(t-1),$$

can be estimated via some heuristic. This estimated value can then be used to determine the likelihood that the terminal queue is congested. Using this and other factors, including the availability of the free slots and the QoS requirement of the connection, the scheduler can determine the free assignments, $S^{\mathrm{fr}}(t)$.

3.4 Statistical Multiplexing

Statistical multiplexing (stat-mux) is one of the key benefits of ATM. The most common method of exploiting stat-mux is to merge multiple VBR streams with similar statistical properties into a common first-in-first-out (FIFO) queue, which may be given some constant rate of service. This is called *intra-terminal statistical* *multiplexing*, which is somewhat like transmitting a group of connections on a "trunk". Such a method allows statistical multiplexing to take place because the average bandwidth requirement of each connection is smaller when there are more connections sharing the queue. To elaborate, suppose that $\rho_a(n)$ is the least amount of constant service bandwidth that can be given to the FIFO queue in order to satisfy the QoS of n statistically identical connections. For n connections, the stat-mux "gain" of VBR over CBR is

$$\frac{n\rho_a(1)}{\rho_a(n)}.$$

This definition is intuitive since the gain when multiplexing a single connection is unity. One scheme for exploiting stat-mux in this manner is to group a set of n VBR connections with similar traffic characteristics onto the same network virtual circuit, see [7, 20] for the example of video teleconferencing. An important point to note is that $\rho_a(n)$ cannot be determined by the aforementioned group of three traffic descriptors (PCR,SCR,MBS). "Statistical" traffic descriptors such as effective bandwidths can be used to determine bandwidth allocations for bundles of VBR connections, see [20, 12] and the references therein.

Intra-terminal statistical multiplexing gains can be exploited by VBR connections (with similar statistics) that share a terminal FIFO queue using fixed-rate DA. See Figure 6. Fixed-rate DA is constant service; so, in essence, the aggregate traffic is treated as a single connection with $\rho^{\mathrm{fx}} = \rho_a(n)$.

Statistical multiplexing among VBR connections using different terminals is significantly more complex. Here, the queues reside on different terminals; so, the multiplexing must be done by the access scheme using a combination of fixed-rate and variable-rate DA, as well as free assignment. This is called *inter-terminal statistical multiplexing*. Just as for the intra-terminal method of stat-mux, inter-terminal stat-mux takes advantage of sharing among a set of VBR connections with similar statistical characteristics using a scheme that is similar to trunking. To do this, the set of connections uses a common pool of bandwidth, denoted by $\rho_e(n)$. The pool $\rho_e(n)$ is to be divided equally among the connections; i.e., each connection in the set is allocated the same amount of bandwidth, ρ, under fixed-rate and variable-rate DA such that

$$\rho = \rho^{\mathrm{fx}} + \rho^{\mathrm{v}} = \frac{\rho_e(n)}{n}.$$

Note that there is a total bandwidth of $n\rho^{\mathrm{v}}$ for demand assignments and some of this bandwidth may be unassigned. The key feature of inter-terminal stat-mux is that any of this unassigned bandwidth (slots) will always be distributed back to *these* connections via free assignment; hence, it is like a trunk. See Figure 6.

The main objective when using inter-terminal stat-mux is to minimize $\rho_e(n)$. Given a set of n connections of a particular type, the value $\rho_e(n)$ depends on two things: the free assignment algorithm and the ratio ρ^{v}/ρ. A heuristic approach to free assignment will now be presented.

This scheme uses a queue threshold β which depends on the required QoS of the VBR connections involved to determine whether a terminal queue is "congested" at any given time t. For each connection i, the satellite keeps track of the number of cells in the terminal queue which exceeds this threshold; i.e.,

$$w_i(t) = (B_i^{\mathrm{g}}(t) - \beta)^+. \tag{6}$$

The value w_i can be used to indicate the connection's need for available free assignment slots. Recall that when the free assignment for time t is made, however, the satellite is not aware of $\{A_i(s)\}_{s=t-\Pi+1}^{t}$. As a result of this, the satellite is only aware of the terminal queue occupancy up to time $t-\Pi$; i.e., $\ldots, B_i^{\mathrm{g}}(t-\Pi-2), B_i^{\mathrm{g}}(t-\Pi-1), B_i^{\mathrm{g}}(t-\Pi)$. This means that at a particular time t, $\{w_i(s)\}_{s\le t-\Pi}$ is the information available to the satellite. Still, these values can be used to produce a weighting system to determine how free assignments are distributed. Let $W_i(t)$ be the weighting for a particular connection i

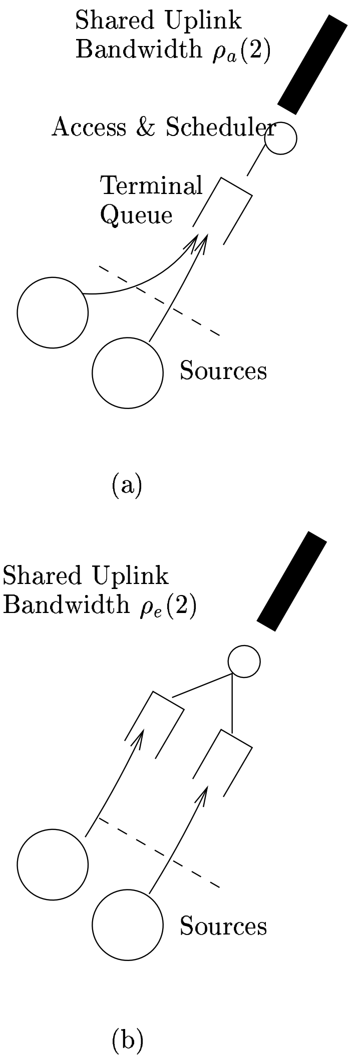

Figure 6: Intra (a) and inter (b) terminal statistical multiplexing

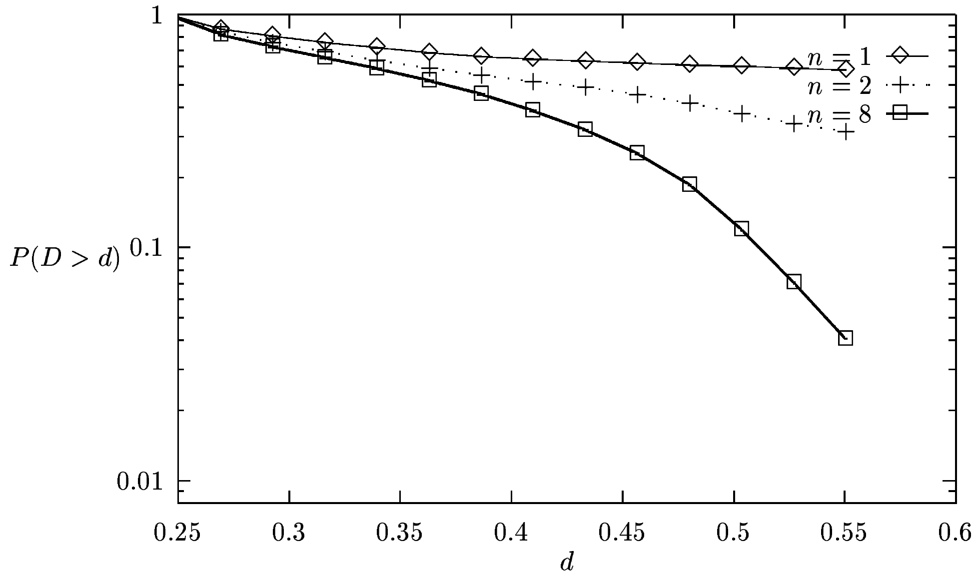

Figure 7: Delay distribution as a function of n with avg. allocated bandwidth 0.45 Mbps

which is used to determine the free assignments at time t. W_i should be an increasing function of w_i; for example, we can take

$$\begin{aligned} W_i(t) &= W_i(t-1)\theta + w_i(t-\Pi) \quad \text{and} \\ W_i(0) &= 0, \end{aligned}$$

where $w_i(t) = 0$ for $t < 0$ and $0 < \theta < 1$ is the "forgetting factor". One recipe for calculating the free assignment could be

$$S_i^{\text{fr}}(t) = F(t)\frac{W_i(t)}{\sum_{j\in V} W_j(t)}, \tag{7}$$

where V is the index set of connections to be multiplexed and $F(t)$ is the total number of unassigned slots from the $\rho_e(n)$ pool of bandwidth. Of course, in practice, a rounding scheme would be required in (7) to make $S_i^{\text{fr}}(t)$ a whole number.

To demonstrate inter-terminal stat-mux, we simulated uplink access for n video traces. Each video trace is a different segment of the aforementioned "Star Wars" trace. The assignment schemes are given by (3), (5) and (7) where we let $\beta = \Pi\rho^{\text{fx}}$ in (6). We simulated $n = 1, 2$ and 8 using inter-terminal stat-mux. As before, the queueing delays onboard were neglected. Figure 7 and 8 show the terminal-to-terminal delay distribution of the connections where each connection is allotted a rate of $\rho = 0.45$ Mbps and $\rho = 0.55$ Mbps, respectively. In both cases, we chose $\rho^{\text{v}}/\rho = 0.5$ and $\theta = 0.98$.

When ρ is fixed, it is clearly seen that delay decreases as n increases. Similarly, when n is fixed, the delay decreases as ρ increases (from 0.45 Mbps to 0.55 Mbps). Stat-mux gains are evident because, given a fixed delay (QoS) requirement, the allocated bandwidth per connection, ρ, would decrease as the number of connections n increase.

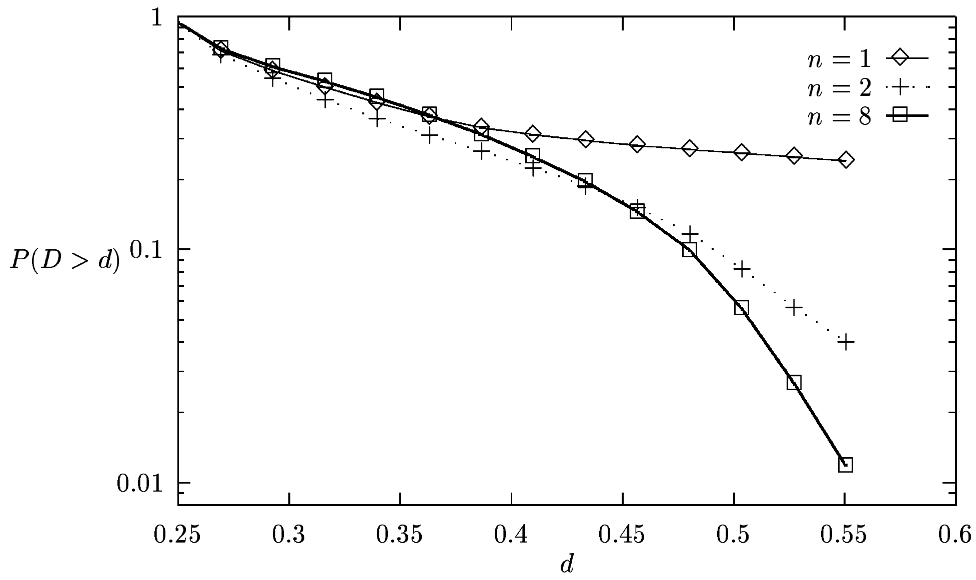

Figure 8: Delay distribution as a function of n with avg. allocated bandwidth 0.55 Mbps

4 Scheduling Implementations for Uplink Access

As discussed above, for each MF-TDMA frame, the satellite grants access to the ground terminal FIFO queues in a manner consistent with their bandwidth allotments. The algorithms given to calculate S^{fx} and S^{v} in (3) and (5) are the theoretical ideal. In practice, however, ρ^{fx} and ρ^{v} cells per MF-TDMA frame cannot be integers for all bandwidth allocations and, of course, it is not possible to assign a fractional number of slots. So, a method of approximating (3) and (5) is necessary. This process is called "bandwidth scheduling". The bandwidth scheduler must be simple to implement and flexible enough to provide for a wide range of bandwidth allocations.

In this section, we will present a practical implementation of a scheduler called Hierarchical Round-Robin (HRR) [11] which can approximate (3) and (5). We will argue that HRR is well suited to scheduling for uplink access because it is simple to implement, can support a range of bandwidth allocations (i.e., fine bandwidth granularity), and does not introduce a large delay jitter.

4.1 HRR Scheduling

The uplink access and scheduler are solely concerned with determining the total number of slot assignments for a particular MF-TDMA frame t, i.e., $S^{\mathrm{fx}}(t) + S^{\mathrm{v}}(t) + S^{\mathrm{fr}}(t)$. This value is then passed on to a lower entity which determines the exact slot locations within the MF-TDMA frame.

The main requirement for the scheduler is that it must be able to guarantee the allocated rate of service. There are generally two ways to provide bandwidth guarantees among the proposed bandwidth schedulers. One method is to service the connection in some round-robin (cyclic) fashion; the other is to use a numerical service deadline (a time-stamp) for each cell and use that to determine which cell to serve. The problems with

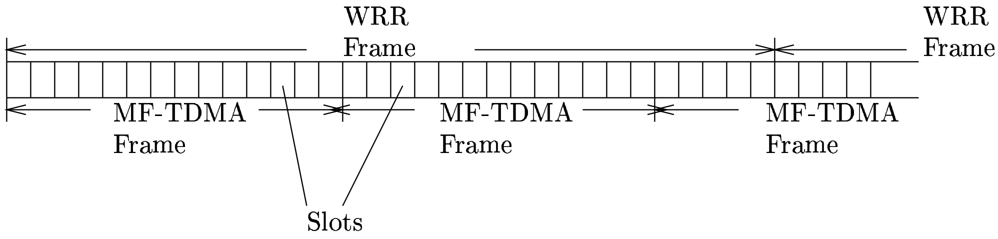

Figure 9: Relating the HRR and MF-TDMA frames

the time-stamping approach are: (large) computationally complexity and (large) storage for time-stamps for each cell. For these reasons, we believe time-stamping is not a practical approach for uplink bandwidth scheduling. So, we turn our attention to the round-robin schedulers.

We first consider a simple TDM-like bandwidth scheduler called Weighted Round Robin (WRR) [21]. Under WRR, connections are given access opportunities in a round-robin fashion. Each round-robin cycle is called a WRR frame (of cell slots). Let f be the number of slots in a WRR frame. In each frame, a connection can be assigned up to its allotted number of slots; the next connection is assigned consecutive slots immediately afterwards. If a connection does not require any of its allotted number of slots, those slots are unassigned; so, this scheduler is *non-work-conserving* or *idling*. The total number of assignments made in a WRR frame is $S^{\rm fx}(t) + S^{\rm v}(t)$. The free assignment algorithm will be invoked after the WRR scheduler has completed an MF-TDMA frame in order to use the unassigned slots to determine $S^{\rm fr}(t)$ for all connections.

Suppose that $f = N_{\rm slots}N_{\rm ch}$ so that the WRR frame corresponds to an MF-TDMA frame. It can be seen that the number of slots which must be allocated in each WRR frame to a connection with a constant bandwidth requirement of ρ cells/frame[2] is $\lceil \rho c^{-1} f \rceil$ cells/frame where we recall that c cells/frame is the total capacity of an uplink beam. Also, recall that we are considering a system using $N_{\rm ch} = 32$ frequencies at 2 Mbps with $N_{\rm slots} = 32$ ATM data slots in each MF-TDMA frame. This means that an MF-TDMA frame is transmitted every 5.9 ms. So, the *minimum* amount of bandwidth that WRR could allocate to a ground terminal queue would be one cell per MF-TDMA frame or (48×8)bits/5.9ms = 64 kbps of payload capacity; this is WRR's *bandwidth granularity*. This means that the bandwidth allocation of connections must be a multiple of 64 kbps. Clearly, many applications may require less than 64 kbps or not a multiple of 64 kbps. To overcome this problem, we might consider a WRR frame that encompasses more than one MF-TDMA frame. In this case, the sequence of assignments chosen are quantized into MF-TDMA frames, i.e., every $N_{\rm ch}N_{\rm slots}$ slots chosen by the scheduler will make up an MF-TDMA frame, as shown in Figure 9. However, since the cells of a particular ground terminal queue are typically served in large clumps when the WRR frame is large, this generally causes greater queueing delay jitter at the ground terminal.

HRR is a more flexible kind of round robin scheduling. HRR attempts to allocate bandwidth more "evenly" over a frame than WRR, and thereby reduce the delay jitter at the terminal queues caused by cell clumping. In HRR, there are $L \geq 1$ *levels* of HRR frames. For the level l HRR frame, $l \in \{1, 2, ..., L\}$, let f_l cells (or "slots") be the frame size. Let $n_l < f_l$ be the number of slots that are reserved for level $l+1$ where $n_L = 0$. Figure 10 shows an example where $f_1 = 14$, $f_2 = 9$, $f_3 = 12$, $n_1 = 3$ and $n_2 = 4$. Queues assigned to level l are served in a WRR fashion on those slots reserved for level l.

The purpose of the multiple levels is to provide different bandwidth granularities. The higher levels,

---

[2] By cells/frame we mean cells per MF-TDMA frame.

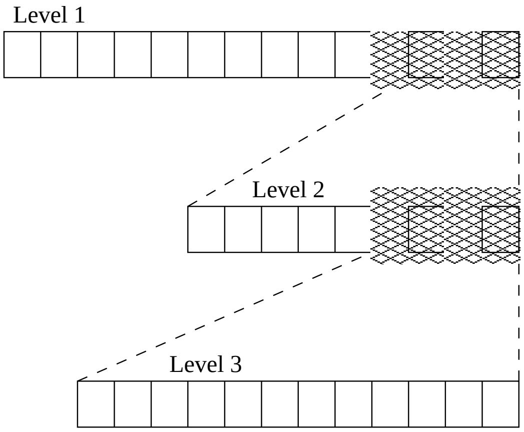

Figure 10: HRR frame structure

i.e., smaller l, are supposed to offer a greater amount of bandwidth but have coarser granularity, and the opposite is true for the lower levels. In addition, it is desirable to provide "regular" (low jitter) service. The following condition simplifies the subsequent analysis:

$$\frac{f_{l+1}}{n_l} \in \{2, 3, 4, ...\} \quad \forall l \in \{1, 2, ..., L-1\}.$$

A connection with a bandwidth requirement of ρ cells/frame would be allocated k_l service slots at each level $l \in \{1, 2, ..., L\}$ such that

$$\sum_{l=1}^{L} k_l \frac{n_1 n_2 ... n_{l-1}}{f_1 f_2 ... f_{l-1} f_l} c \quad \geq \quad \rho. \tag{8}$$

Note that resource allocation is most efficient when the left-hand side of (8) is as small as possible; the inequality is needed because there may not exist a set $\{k_1, k_2, ..., k_L\}$ that achieves (8) with equality. However, for simplicity here, we will assume that (8) is achieved with equality.

Note that HRR, like WRR, leaves unassigned slots for use by the free assignment scheme. In addition, we will show that this non-work-conserving feature of HRR allows for smaller buffer sizing requirements onboard. It should also be noted here that previously proposed implementations of HRR have concentrated on giving each connection access to only one level; i.e., $k_l > 0$ for some l and $k_i = 0$ for all $i \neq l$. However, the above implementation allows for a more flexible set of bandwidth allocations at almost no additional complexity.

For the MF-TDMA system described above, a simple HRR configuration that is appropriate for ATM applications is given in Table 2. Recall that the system described has 32 frequencies, each of which has a payload capacity of 2 Mbps, giving a total of $c = 64$ Mbps on each uplink beam.

The HRR levels are configured in a regular manner. Firstly, $f_1 = N_{\text{slots}} N_{\text{ch}}$ so that a connection which is given a slot in the level 1 HRR frame gets one slot on every MF-TDMA frame. Secondly, it can be seen that a level 2 HRR frame is completely served in exactly 4 MF-TDMA frames; so a connection which is

Table 2
A sample HRR configuration for ATM applications

l	f_l	n_l	Bandwidth for one slot per HRR frame
1	1024	32	64 kbps (1 slot every MF-TDMA frame)
2	128	64	16 kbps (1 slot every 4 MF-TDMA frames)
3	1024	—	1 kbps (1 slot every 64 MF-TDMA frames)

given a slot in the level 2 HRR frame gets exactly one slot every 4 MF-TDMA frames. For level 3 HRR frames, one HRR frame is served in exactly every 16 level 2 frames or every $16 \times 4 = 64$ MF-TDMA frames. So, this configuration can efficiently support bandwidth requirements that are multiples of 64 kbps (e.g., for telephony and certain ISDN services) and it can also support very low bandwidth requirements (down to 1 kbps).

4.2 Performance of HRR

How well HRR allocates bandwidth is indicated by its minimum-bandwidth (or "guaranteed service") property [9, 12, 21, 8]. We begin with some preliminary definitions. Consider a sequence of cell arrival times $\mathbf{A} = \{a_i\}_{i=1}^{\infty}$ where a_i is the arrival time (in MF-TDMA frames) of the i^{th} cell of this sequence. The *service deadlines*[3], $\{F_i\}_{i=1}^{\infty}$, based on the arrival times $\mathbf{A}$ and bandwidth allotment ρ cells/frame are defined recursively as follows:

$$\begin{aligned} F_i &= \max\{F_{i-1}, a_i\} + \rho^{-1} \quad \forall i \geq 1, \text{ and} \\ F_0 &= 0 \end{aligned}$$

Note that F_i is basically the time at which cell i completely departs from a queue with a *constant* service rate of ρ cells/frame.

A bandwidth scheduler handling a set of queues is said to have a minimum bandwidth property of μ for a given queue if, for all arrival processes $\mathbf{A}$ to that queue and for all arrival processes to the other queues handled by the scheduler,

$$d_i \leq F_i + \mu \tag{9}$$

where d_i is the departure time of the i^{th} cell. In general, μ may depend only on the bandwidth allotment of the given queue. Note that the minimum-bandwidth property μ is in units of time (MF-TDMA frames) and is a type of bound on the maximum queueing delay. The minimum-bandwidth property also tells us how well a bandwidth scheduler guarantees a minimum rate of service. This leads to the useful delay bounds and buffer sizing results that are stated below.

Theorem 1
Under HRR, a queue with a bandwidth allotment of ρ cells per MF-TDMA frame has a minimum-bandwidth property of

$$\mu = \rho^{-1} \sum_{l=1}^{L} ((c - \varphi_l)k_l - 1) \qquad \textit{MF-TDMA frames}$$

---

[3] Also known as the VirtualClock Virtual Finishing Times (VC-VFTs).

where

$$\varphi_l = k_l \frac{n_1 n_2 ... n_{l-1}}{f_1 f_2 ... f_{l-1} f_l} c \tag{10}$$

and $\rho = \sum_{l=1}^{L} \varphi_l$.

The proof of this result is given in [10]; see also [12, Chapter 3]. This theorem was used along with another result in [10] to show that HRR has very good delay jitter properties, as well as small buffer size requirements, as shown by the following.

To determine appropriate buffer sizing onboard the satellite, knowledge of the minimum-bandwidth property μ of the uplink scheduler and the so-called "(σ, ρ) constraint" [5, 2, 19] of the flow passing through the UNI is required. The (σ, ρ) constraint is well suited to characterizing ATM traffic; for example, a VBR source which obeys the traffic descriptors (PCR,SCR,MBS) has a constraint of (σ, ρ) =(MBS+1,SCR). A CBR source has a constraint of (σ, ρ) =(1,PCR) and proposed rate-based flow control schemes police the ABR sources so that they obey a (σ, ρ) constraint with a variable ρ. At the UNI, these constraints can be enforced by a leaky bucket or, in the case of CBR, by a cell spacer.

Note in Figure 2 that a connection sees two tandem queues, one for the uplink and one for the downlink, each having a bandwidth scheduler. Let μ^{g} and μ^{sat} be the minimum-bandwidth property of the uplink and downlink access schedulers, respectively.

Theorem 2
Consider a (σ, ρ) constrained flow that enters a terminal queue and is allocated a bandwidth of ρ cells per MF-TDMA frame on the uplink and downlink. Assume that only fixed-rate DA is used; i.e., $\rho^{\mathrm{fx}} = \rho$. If HRR is used as the uplink access scheduler, a queue size of at least

$$\sigma + \rho\mu^{\mathrm{g}} + 1 \quad \textit{cells}$$

at the terminal and a queue of size at least

$$\rho(\mu^{\mathrm{g}} + \mu^{\mathrm{sat}}) + 2 \quad \textit{cells}$$

at the output port queue onboard the satellite are sufficient to ensure that there is never overflow.

The proof of this result is given in [10], see also [12, 6]. A consequence of the non-work-conserving nature of the uplink scheduler is that the queue size requirement onboard the satellite does not depend on the term σ. This is highly desirable because minimizing onboard memory is a key goal. Along these lines, Theorem 2 shows one of the benefits of having a scheduler with a small minimum-bandwidth property. Finally, although Theorem 2 assumes only fixed-rate DA is used, similar results can be derived for the case where both variable-rate and fixed-rate DA are used.

4.3 Implementation Details

The goal of the uplink access scheduler is to best approximate the fixed-rate DA, variable-rate DA, and free assignment schemes of §3 to determine $S_i^{\mathrm{fx}}(t)$, $S_i^{\mathrm{v}}(t)$, and $S_i^{\mathrm{fr}}(t)$ for all connections i. This information is then passed on to a lower layer entity which would make the actual assignments in the MF-TDMA frame (i.e., choosing the exact slot in Figure 1) which are then transmitted on the downlink.

The role of HRR is to implement fixed-rate and variable-rate DA. A connection with both fixed-rate and variable-rate DA is treated as two connections, one with just fixed-rate DA having an allocated bandwidth of $\rho^{\rm fx}$ and the other with just variable-rate DA having an allocated bandwidth of $\rho^{\rm v}$. So, there would be two sets of allocated slots $\{k_1, k_2, ..., k_L\}$, one would satisfy (8) for $\rho^{\rm fx}$ and the other would satisfy (8) for $\rho^{\rm v}$. These two sets of slots would serve the same queue. The HRR scheduler works in the following manner for a given uplink frame t. First, the scheduler cycles through the next $N_{\rm ch}N_{\rm slots}$ slots in the HRR structure and counts the number of fixed-rate and variable-rate DA slots for all the connections. Let $c_i^{\rm fx}(t)$ and $c_i^{\rm v}(t)$ be the number of fixed-rate and variable-rate DA slots assigned to connection i by HRR. Equations (3) and (5) are then modified for the case of HRR:

$$\begin{aligned} S_i^{\rm fx}(t) &= c_i^{\rm fx}(t) \quad \text{and} \\ S_i^{\rm v}(t) &= \min\left[\left(\widehat{B}_i^{\rm g}(t-1) - S_i^{\rm fx}(t)\right)^+, c_i^{\rm v}(t)\right] \end{aligned}$$

where $\widehat{B}_i^{\rm g}(t-1)$ is defined by (4). Once $S_i^{\rm fx}(t)$ and $S_i^{\rm v}(t)$ have been calculated for all i, $S_i^{\rm fr}(t)$ can be calculated for all i using free assignment as previously discussed.

The next implementation detail is the actual slot assignment in an MF-TDMA frame and its transmission on the downlink. We are currently studying schemes to minimize the amount of downlink transmission for assignments. Another implementation detail is the HRR configuration, examples of which are given in Table 2 and Figure 10. It is possible to reconfigure the HRR frames while the network is active in order to better suit changing bandwidth requirements. We are studying methods of doing this also.

5 Summary and Conclusions

This paper presented an overall framework to integrate satellite uplink access scheduling and current ATM protocols. The satellite considered integrates an onboard ATM fast packet switch and is in line with currently proposed onboard processing systems. An ATM sublayer, in which the bandwidth scheduling is performed, was introduced. In a satellite network, where timing is critical, it was shown that the location of the scheduler and network controller onboard offers many advantages. We described how to use a combination of the free, fixed-rate and variable-rate demand assignment techniques in order to support all ATM traffic classes and their QoS requirements. That is, we mapped the various ATM service classes to satellite access schemes.

To meet individual QoS requirements, a resource allocation scheme for the uplink was outlined. The access scheme also allows for the allocation of remaining capacity using a free assignment strategy. The resource allocation scheme can allow for statistical multiplexing on the uplink so that the network can better cope with the burstiness of VBR and best-effort traffic.

The implementation of HRR as a scheduler for uplink access was also described. HRR was chosen on the grounds that it is simple to implement and not computationally intensive. Also, it can accommodate the required rates of service with a small amount of delay jitter, which in turn creates small buffer requirements onboard the satellite.

The use of ATM via satellite implies an end-to-end system approach in which networking, transmission and onboard processing are related. We are currently working on a number of outstanding issues mentioned in the paper. Others are making progress on ATM via satellite also; for example, current ATM demonstrators have used trunk type systems at OC-3 using large stations and bent-pipe satellites [1]. The onboard switching system is being developed with the ultimate goal of having a simple power efficient architecture [15]. It is our purpose to bring satellite ATM to the end-user by using techniques favored by both the broadband wireless and wireline networks.

Acknowledgments

We would like to thank Peter Takats and Jim Gilderson of Spar Aerospace for their comments. We also thank Professor Jon W. Mark for this comments on the access schemes. Finally, we would like to thank the anonymous reviewers for their helpful comments.

Anthony Hung received the B.A.Sc. degree in Computer Engineering from the University of Waterloo, Canada, in 1994. He plans to complete on his Ph.D. degree at Waterloo in the summer of 1997. His research interests include bandwidth scheduling and design issues for ATM switches, prerecorded video transmission, and satellite ATM. Mr. Hung has also been an undergraduate instructor.

Marie-José Montpetit was born in Rimouski, Canada. She received a Ph.D. in Electrical Engineering from Ecole Polytechnique in Montreal, Canada in 1991. She is currently a Senior Member of the Technical Staff at Teledesic Corporation in Kirkland, Washington. Her main responsibilities are in the development of the multiple access and capacity allocation scheme of the Teledesic network air interface. Her other interests include forward error correction and broadband transmission.

George Kesidis was born in Toronto, Canada, in 1964. He received a B.A.Sc. degree from the University of Waterloo in 1988 and the M.S. and Ph.D. degrees from the University of California at Berkeley in 1990 and 1992, respectively, all in electrical engineering. He is currently an associate professor in the Electrical and Computer Engineering Department of the University of Waterloo, Waterloo, Ontario, Canada. His research interests include design, resource management and performance evaluation of high-speed communication networks.

References

[1] S. Agnelli. The UNOM experiment. In *Proc. IEEE ICC*, Dallas, 1996.

[2] V. Anantharam and T. Konstantopoulos. Optimality and interchangeability of leaky buckets in tandem. Technical Report SCC-95-01, Dept of E.&C.E., U.T. Austin, Jan., 1995.

[3] ATM Forum Technical Committee. *User-Network Interface (UNI) Specification Version 3.1.* Prentice-Hall, Upper Saddle River, NJ, 1995.

[4] D.M. Chitre, D.S. Gokhale, T. Henderson, and N. Mathews. Asynchronous transfer mode (ATM) operation via satellite: Issues, problems, and resolutions. In *Proc. MILCOM*, pages 258–246, 1993.

[5] R. L. Cruz. A calculus for network delay, Part 1: Network elements in isolation. *IEEE Trans. Inform. Theory*, 37:114–131, 1991.

[6] R.L. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE JSAC*, Vol. 13, No. 6:1048–1056, 1995.

[7] A. Elwalid, D. Heyman, T.V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE JSAC*, Vol. 13, No. 6:pages 1004–1016, Aug. 1995.

[8] P. Goyal, S. Lam, and H. Vin. Determining end-to-end delay bounds in heterogeneous networks. In *Proc. 5th International Workshop on Network and Operating System Support for Digital Audio and Video*, pages 273–284, Durham, NH, 1995.

[9] A. Hung and G. Kesidis. Bandwidth scheduling for wide-area ATM networks using virtual finishing times. *IEEE/ACM Transactions on Networking*, Vol. 4, No. 1:49–54, 1996.

[10] A. Hung and G. Kesidis. Performance evaluation of multilevel-assignment hierarchical round-robin bandwidth scheduling for ATM. In *15th International Teletraffic Congress*, Washington, D.C., 1997.

[11] C.R. Kalmanek, H. Kanakia, and S. Keshav. Rate controlled servers for very high-speed networks. In *Proc. IEEE Globecom*, pages 300.3.1–300.3.9, 1990.

[12] G. Kesidis. *ATM Network Performance.* Kluwer Academic Publishers, Boston, MA, 1996.

[13] H. W. Lee and J. W. Mark. Combined random/reservation access for packet-switched transmission over a satellite with on-board processing — part II: Multibeam satellite. *IEEE Trans. Commun.*, vol. 32, no. 8:1093–1104, Oct. 1984.

[14] M.-J. Montpetit and M. Coté. LAN interconnection by onboard switching satellites. In *Proc. Canadian Electrical Engineering Congress, Montreal*, Sept. 1995.

[15] M.-J. Montpetit and A. Hung. Design issues for ATM over satellite. In *18th Biennial Symposium on Communications*, pages 83–86, Kingston, Ont., Canada, 1996.

[16] H. Ohsaki, M. Murata, H. Suzuki, C. Ikeda, and H. Miyahara. Rate-based congestion control for ATM networks. *ACM SIGCOMM, Computer Communication Review*, pages 60–72, 1995.

[17] B.A. Pontano, D.Chitre, L. White, and T. Henderson. Processing and nonprocessing satellite architectures for support of ATM traffic. In *Proc. MILCOM*, 1993.

[18] D. Raychaudhuri and N.D. Wilson. ATM-based transport architecture for multiservices wireless personal communication networks. *IEEE JSAC*, vol. 12, no. 8, 1994.

[19] J.W. Roberts. Virtual spacing for flexible traffic control. *International Journal of Communication Systems*, Vol. 7:307–318, 1994.

[20] G. De Veciana, G. Kesidis, and J. Walrand. Resource management in wide-area ATM networks using effective bandwidths. *IEEE JSAC*, Vol. 13, No. 6:1081–1090, 1995.

[21] H. Zhang. Service disciplines for guaranteed performance service in packet-switching networks. *Proc. of the IEEE*, Vol. 83, No. 10, Oct. 1995.