Guaranteed Frame Rate: A Better Service for TCP/IP in ATM Networks

Olivier Bonaventure, University of Namur
Jordi Nelissen, Colt Telecom

Abstract

Guaranteed frame rate, recently approved by the ATM Forum, is expected to become an important service category to efficiently support TCP/IP traffic in ATM networks. We first describe the GFR traffic contract in details. We then present different types of switch implementations that have been proposed to support GFR. We analyze the performance of three of these switch implementations by simulations in two different network environments. These simulations show that the scheduler-based implementations provide a much better performance than the simple switch implementation. However, we also show that coupling an active packet discard mechanism to a scheduler-based switch implementation does not produce a performance gain when many TCP connections are multiplexed inside one ATM VC.

Asynchronous transfer mode (ATM) was proposed in the 1980s as an evolution of the public networks to support broadband services. Since then, it has evolved into a networking technology which is suitable for both public and private networks. Since the first ITU-T Recommendations on ATM in the late 1980s, a lot of work has been carried out to better support TCP/IP traffic in ATM networks. An ATM network is able to provide different types of services, ranging from guaranteed services comparable to leased lines to best-effort services comparable to the service provided by today's Internet.

The first ITU-T Recommendations on ATM only supported one type of service, constant bit rate (CBR). This service can be considered an ATM version of the leased lines services provided in networks using integrated services digital network (ISDN) or synchronous optical network/synchronous digital hierarchy (SONET/SDH). Shortly after the definition of the CBR service category, the ATM Forum was created to improve the suitability of ATM to be used in data networks. Partially based on the work of the ATM Forum, two new services were defined.

The first one, variable bit rate (VBR), can be considered a natural evolution of CBR service. With CBR service, a fixed amount of bandwidth can be reserved for each virtual circuit (VC). However, this amount of bandwidth is specified roughly as one cell every N µs. This is well suited to applications such as uncompressed voice, video or leased line emulation, but not entirely for data applications, which are much more bursty.

VBR service also allows end systems to reserve some bandwidth inside the network for each VC, but in this case the bandwidth is specified by three parameters: the peak cell rate, which is the highest rate at which the end system is allowed to transmit ATM cells during short periods of time; the maximum burst size, which is used to limit the amount of cells the end system can transmit at the peak cell rate; and the sustainable cell rate, which corresponds to the rate at which the end system is allowed to transmit continuously. These three parameters define a traffic envelope which is used by the network to reserve resources for each VBR VC. Three types of VBR services have been defined by the standardization bodies. They differ in the utilization of the cell loss priority (CLP) bit.

With the VBR.1 service category, the end system is allowed to send CLP = 0 and CLP = 1 cells inside its traffic envelope.

With the VBR.2 service category, the end system is allowed to send CLP = 0 cells inside its traffic envelope and CLP = 1 cells outside this envelope.

The VBR.3 service category is similar to VBR.2 except that the CLP = 0 cells which are sent above the traffic envelope with VBR.2 will be considered nonconformant by the policing unit at the ingress of the network and will usually be discarded, while their CLP bit will be changed from 0 to 1 with the VBR.3 service category. With all ATM service categories, the

0890-8044/01/$10.00 © 2001 IEEE

end system is never allowed to transmit ATM cells at a higher rate than the negotiated peak cell rate.

A second contribution of the ATM Forum was the definition of a best-effort service category, unspecified bit rate (UBR). This service mimics the service provided by today's Internet. Initially, the UBR service category was defined to offer a cell-based service, but it quickly appeared that to efficiently support TCP/IP traffic, it was necessary to provide a service which is at least aware of the ATM Adaptation Layer type 5 (AAL5) frame boundaries. Today, most ATM switches implement frame discard strategies such as early packet discard (EPD) to discard entire AAL5 frames instead of individual ATM cells when congestion occurs.

Although UBR was a good match for today's Internet, it was only suitable for networks where congestion is handled by an upper layer protocol; otherwise, there is a high risk of congestion collapse. The ATM Forum anticipated this potential problem and, as soon as the UBR service category was defined, started to work on a new service category, available bit rate (ABR) [1]. ABR is based on a rate-based congestion control mechanism implemented in the ATM layer. With the ABR service, the end system periodically sends special resource management (RM) cells within its flow of data cells. These RM cells are used to provide feedback to the sources, which are forced to transmit at the rate specified by the RM cells to avoid congestion. The RM cells are modified on the fly by ATM switches based on their congestion level, and the destination returns them to the source. The ABR service can provide both a best-effort service and a service with a minimum guaranteed bandwidth. The end system selects between both types of services by specifying a minimum cell rate (MCR) in addition to the peak cell rate. When MCR is positive, the network must reserve at least this amount of resources for the end system, and the end system is guaranteed to always be allowed to transmit at this rate. The ABR service only supports CLP = 0 cells and, thanks to the utilization of its rate-based congestion control mechanism, provides loss-free operation.

The guaranteed frame rate (GFR) service category was defined as a new service category to better support TCP/IP traffic. We describe the definition of this new service category. We also qualitatively compare it with the existing service categories. The support of GFR in ATM networks will require modifications to the existing ATM switches. We discuss several mechanisms that have been proposed to efficiently support GFR inside ATM switches. To be useful in real networks, it is important that these switch implementations efficiently support TCP/IP traffic. We study the performance of three of these switch implementations by simulations. We consider two different network environments and show the severe limitations of the simplest switch implementation when carrying TCP/IP traffic. We finally summarize our findings and some important lessons when studying the performance of TCP/IP.

The GFR Service Category

The GFR service category is one of the most recent ATM service categories. It was first proposed in December 1996 [2] and quickly became very popular. The GFR specification was recently finalized by the ATM Forum [3].

The main motivation for the introduction of this new service category was to provide a service which is as easy to use as the UBR service category for the end systems while still providing bandwidth guarantees. In [2], the promoters of GFR argued that for most deployed end systems, the only really usable service category defined at that time was the UBR service category because these end systems are either directly connected to the ATM network, but with an ATM adapter which does not provide the shaping mechanisms required by the CBR, VBR, and especially ABR service categories, or not attached directly to the ATM network. In the latter case, which corresponds to most of today's corporate networks, the end systems are fitted with Ethernet or token ring adapters and are connected through intermediate systems (i.e., bridges or routers) to the ATM network. For these end systems, the guarantees offered by the ATM network are completely hidden by the intermediate system.

The GFR service category keeps the simplicity of UBR (from the endsystem's point of view) by allowing the end system to transmit cells at the line rate of their ATM adapter. Another important feature of GFR is that since almost all data traffic is AAL5-based, GFR takes the specific requirements of AAL5 into account. The GFR service category requires the network elements to be aware of the AAL5 frame boundaries and to discard entire AAL5 frames when congestion occurs. Such a strong requirement was not included in the UBR service category, although it is also mainly used for AAL5-based traffic. Another important difference between GFR and UBR is that GFR allows the user to reserve some bandwidth, for each GFR VC, inside the network. This means that the user is assured that she will always be able to transmit at a minimum rate without losses. On the other hand, if the network is not congested, the user will be able to transmit at a higher rate. Furthermore, in case of congestion, the network will drop entire frames instead of dropping individual cells from possible different frames.

More precisely, the GFR traffic contract [3] is composed of four main parameters (neglecting the cell delay variation tolerances):

- Peak cell rate (PCR)
- Minimum cell rate (MCR)
- Maximum burst size (MBS)
- Maximum frame size (MFS)

The PCR has the same meaning as in UBR: it is the maximum rate at which the end system is allowed to transmit. It can be expected that the PCR will often be set at the line rate of the ATM adapter of the end system. The MFS is the largest size of AAL5 frame the end systems can send. For GFR switched VCs (SVCs), this parameter will be equal to the AAL5-CPCS SDU size parameter which is negotiated between the source and destination end systems during connection setup.

The end systems request a minimum guaranteed bandwidth by specifying a nonzero MCR and an associated MBS. The MCR, expressed in cells per second, corresponds to the long-term average bandwidth which is reserved for the VC inside the network. It is similar to the sustainable cell rate (SCR) in VBR [3], although the MCR provides a minimum guaranteed bandwidth to entire AAL5 frames, while the SCR provides a minimum guaranteed bandwidth to individual cells. The MBS places an upper bound on the burstiness of the traffic to which the minimum guaranteed bandwidth applies. The value of the MBS is negotiated between the end systems and the network, but according to [3] this parameter must always be at least equal to $1 + [(MFS \times PCR)/(PCR - MCR)]$.

The GFR service category defines a particular utilization of the CLP bit in the ATM cells. Since the logical unit of information is a frame, GFR imposes that all the cells of a frame have the same CLP bit. The CLP = 1 AAL5 frames are considered low-priority frames which should be transmitted by the network on a best-effort basis. The minimum guaranteed bandwidth is only applicable to the CLP = 0 frames.

With this utilization of the CLP bit, the intuitive meaning of the MCR is that if the end system transmits CLP = 0 AAL5 frames at a rate smaller than or equal to the MCR, these frames should be correctly received by the destination.

```
Cell arrival at time t_a:
First cell of an AAL5 frame:                Middle or last cell of an AAL5 frame :
if( ( t_a <TAT – L ) OR (IsCLP1(cell))      if(eligible)
{                                           {
/* non-eligible cell */                     /* eligible cell */
eligible=FALSE;                             TAT = max(t_a,TAT)+T;
}                                           }
else                                        else
{                                           {
/* eligible cell */                         /* non-eligible cell*/
eligible = TRUE;
TAT = max(t_a,TAT)+T;
}                                           }
```

Figure 1. *The simple-F-GCRA(T,L).*

However, GFR does not require the end systems to shape their traffic, and it can be expected that most users of this service category will always transmit at the negotiated PCR. In this case, each frame will appear as a burst of cells transmitted at the PCR.

Formally, the minimum guaranteed bandwidth is specified by F-GCRA(T,L) [3] with parameters T = 1/MCR and L = (MBS – 1) x (1/MCR – 1/PCR). The F-GCRA is an adaptation of the GCRA used in VBR [3]. The main difference between the GCRA and F-GCRA is that the F-GCRA declares entire CLP = 0 frames to be eligible or ineligible for the minimum guaranteed bandwidth. The eligible AAL5 frames are those which should be delivered to the destination to fulfill the minimum guaranteed bandwidth. The CLP = 1 are not eligible for the minimum guaranteed bandwidth. While the F-GCRA is used to specify which frames are eligible for the minimum guaranteed bandwidth, it should be noted that GFR explicitly allows end systems to transmit frames in excess of this minimum guaranteed bandwidth. The GFR service category also expects the network to deliver this excess traffic on a best-effort basis to the destination end systems and to "fairly" distribute the available bandwidth to the active VCs. The exact definition of the fair distribution is implementation-dependent.

The F-GCRA algorithm specified in [3] is an ideal F-GCRA, which may be difficult to implement in real policers, shapers, or schedulers. The Simple-F-GCRA (Fig. 1) defined in [3] is a slightly simplified version that is equivalent to the F-GCRA(T,L) for connections containing only conforming frames (i.e., frames that contain at most MFS cells with the same CLP bit setting and pass the GCRA(PCR,τ_{PCR}) test).

As with other service categories, two GFR conformance definitions have been defined: GFR.1 and GFR.2. The only difference between them is whether an F-GCRA is used to explicitly set the CLP bit to one in the ineligible frames at the ingress of the network or not. With GFR.2 conformance, the policing function at ingress of the network uses an F-GCRA to tag the non-eligible AAL5 frames. When this conformance definition is used, only the eligible AAL5 frames are accepted as CLP = 0 AAL5 frames inside the network. Thus, there is a clear distinction between the eligible (CLP = 0) and ineligible (CLP = 1) AAL5 frames, and the ATM switches may rely on this to decide whether an AAL5 frame must be delivered to fulfill the minimum guaranteed bandwidth or not. As we will see later, a simple switch implementation can be used to support GFR.2 conformance.

With GFR.1 conformance, the network is not allowed to modify the CLP bit of the frames sent by the end systems, but the end systems are still allowed to send CLP = 0 frames in excess of the minimum guaranteed bandwidth (even if only a fraction of these frames are actually eligible for the guaranteed minimum bandwidth). With this conformance definition, there is thus no "visible" distinction between an eligible and an ineligible AAL5 frame inside the network. Thus, to support GFR.1 conformance, each ATM switch in the network must be able to determine, by itself, which CLP = 0 frames must be transmitted to fulfill the minimum guaranteed bandwidth and which AAL5 frames are part of the excess traffic and thus could be discarded if congestion occurs. It should be noted that the GFR service category does not require that the frames found eligible at the ingress of the network are exactly those which must be delivered to the destination to provide the minimum guaranteed bandwidth. The requirement is weaker. GFR only requires the network to deliver enough complete CLP = 0 frames at the destination to provide the minimum guaranteed bandwidth, but does not specify precisely which CLP = 0 frames must be delivered to the destination.

Another particular point of the definition of GFR is the "semantics" of the CLP bit when the GFR.1 conformance definition is used. There are two possible interpretations for the semantics of this bit; after several discussions the ATM Forum did not choose one over the other.

The two possible interpretations are :

- With the strict interpretation, all the CLP = 0 frames, including the ineligible ones, have greater "importance" than the CLP = 1 frames. This interpretation implies that the switches should always first discard CLP = 1 frames before discarding an ineligible CLP = 0 frame.
- With the relative interpretation, a CLP = 1 frame is always less important than a CLP = 0 frame *belonging to the same VC*, but it can be more important than a CLP = 0 frame belonging to a different VC. This interpretation implies that the switches should discard arriving AAL5 frames based not only on the CLP bit of the first cell of the frame but also on the resource consumption of the corresponding VC.

To understand the difference between these two semantics, let us consider an example with two VCs with the same MCR and PCR multiplexed on a single link. Let us also consider that the MCR of these VCs is equal to 50 percent of their PCR. Suppose that VC_1 is only sending CLP = 1 frames while VC_2 is only sending CLP = 0 frames. With the strict interpretation for the CLP bit, VC_2 will receive a very large fraction of the bottleneck link, and VC_1 will not be able to efficiently utilize its MCR. With the relative interpretation for the CLP bit, both VC_1 and VC_2 would receive 50 percent of the bottleneck link. Since the ATM Forum did not choose between the two interpretations, each supplier of ATM equipment or network provider will have to choose one interpretation. This may create interoperability problems from a performance point of view in heterogeneous networks.

GFR and the Other Service Categories

Before discussing how GFR can be supported by ATM switches and their respective performance, it is useful to qualitatively compare GFR with the main ATM service categories. An interesting comparison of the various service categories from the viewpoint of an enterprise network may be found in [4].

Compared with UBR, the main advantages of GFR with a zero MCR is that with GFR the network must take the frame boundaries into account. This implies that GFR switches must implement frame discard strategies. Although UBR switches should also implement these strategies, this is not a strong requirement. It can thus be expected that the performance of TCP over GFR with a zero MCR would at least be as good as the performance of TCP over UBR.

Compared to VBR, the main advantage of GFR is that the quality of service (QoS) guarantees are provided at the frame level, while they are only provided at the cell level in VBR.3. The simulation studies [5, 6] which considered both VBR.3 and GFR showed that TCP always achieved better performance with GFR than with VBR.3. It can be expected that networks which today rely on the VBR service category to support TCP/IP traffic will utilize GFR as soon as it becomes readily available.

Compared to ABR, the main advantage of GFR and its initial motivation is its simplicity for the end systems. GFR does not impose the implementation of complex shapers inside the end systems as does ABR. This might be at the expense of more complex switch implementations relying on per-VC scheduling, as we will see in the next section. However, although ABR can be used with simple FIFO switches, it also forces the network operator to select values for a large number of operational parameters (rate increase factor, RIF; rate decrease factor (RDF); etc. [3]) that influence the behavior of the sources and destinations. The correct selection of these parameters in an heterogeneous network is not always a simple task. GFR does not require the specification of such parameters.

A second advantage of GFR is that it does not introduce additional overhead due to the transmission of RM cells. This overhead can be important in practice. Of course, this advantage must be balanced with the fact that GFR does not provide a loss-free service, while ABR avoids losses inside the ATM network.

Since the approval of the GFR specification by the ATM Forum [3], the ATM Forum has been working on two simple improvements for UBR. These two modifications to UBR were in straw ballot within the ATM Forum at the time of this writing.

The first improvement [7] allows the end system to optionally associate a minimum data cell rate (MDCR) to UBR VCs. This MDCR is similar in spirit to the MCR of GFR, but with no QoS commitment attached. This implies that even when the MDCR is specified, UBR still only provides best effort service. The MDCR is thus only an indication that the end system provides to the switches. This contrasts with the utilization of the MCR in GFR since a strong QoS commitment is associated with the MCR.

The second improvement [8] allows the end system to attach a *behavior class* to each UBR VC. This information will then be used by the switches to provide different classes of service to VCs with different behavior classes. This modification is intended to allow ATM switches to better support the Internet Engineering Task Force (IETF) differentiated services model and the IEEE 802.1D user priorities used in 802.x LANs. GFR does not directly support such behavior classes. With GFR, the differentiation between VCs can only be done on the basis of the MCR. Such MCR-based differentiation can be used to support IETF assured forwarding service over ATM networks [9].

Supporting GFR in ATM Switches

Several implementations have been proposed in the literature to support the GFR service category in ATM switches. These implementations can be grouped in three main categories depending on complexity. Another survey of switch implementations may be found in [10].

The simplest switch implementations only need to maintain a few bits of state for each established VC to perform the AAL5 frame discard mechanisms required by GFR service. They usually rely on the tagging of the ineligible frames performed by an F-GCRA at the ingress of the network. The main advantage of these implementations is their low complexity. This group of implementations includes the simple [2] buffer acceptance algorithm described herein and a modified buffer acceptance algorithm proposed in [11]. The main drawback of these implementations is that they only support GFR.2 conformance.

The switch implementations of the second category maintain at least one counter and a few bits of state for each established VC while being suitable for FIFO-based switches. Several implementations of this kind have been proposed in the literature, such as DFBA [12], the virtual queuing technique proposed in [13] and briefly described in this article, and the second and third implementations proposed in [11].

Finally, the more complex switch implementations [2, 14] maintain one counter and a few bits of state for each VC as well as one logical queue for each established VC. These implementations usually rely on a Weighted Fair Queuing (WFQ)-like scheduler to provide the minimum guaranteed bandwidth (i.e., the scheduler is configured to serve the queue corresponding to each VC at least at its MCR). Implementations from this group are discussed below.

A Simple Implementation for GFR.2

The simple switch implementation proposed in [2] is an adaptation of a simple buffer acceptance algorithm frequently used to support VBR service in ATM switches. Intuitively, this switch implementation provides the MCR guarantee by discarding CLP = 1 frames earlier than CLP = 0 frames. This is possible provided that the amount of CLP = 0 frames is bounded (i.e., when GFR.2 conformance is used). The switch provides the MCR guarantee by avoiding to drop CLP = 0 frames.

More precisely, this simple switch implementation is an AAL5-aware buffer acceptance algorithm which relies on two buffer thresholds. The low threshold is used to limit the amount of ineligible (CLP = 1) frames inside the buffer. When the queue occupancy of the buffer is above this threshold, the newly arriving CLP = 1 frames are entirely discarded. The value of the low threshold is chosen as a function of the traffic contract of the established VCs. CLP = 0 frames are entirely discarded when the buffer occupancy reaches the high threshold, but the connection admission control (CAC) algorithm should ensure that this is a rare event. The high threshold is only used to ensure that the switch will not drop individual cells from a frame. Its value will usually be close to the buffer size and should not impact performance. The simple switch implementation is shown graphically in Fig. 2. All the GFR VCs are multiplexed in a single FIFO buffer which is directly attached to the output link.

The main advantage of this switch implementation is that it only requires a global counter for the number of cells in the buffer, and two bits of state information for each VC to discard entire CLP = 1 frames and not individual cells when congestion occurs.

Counter-Based Implementations

The buffer acceptance algorithm proposed in [13] aims to support GFR service with a single FIFO buffer for all the GFR VCs. This implementation maintains one counter for each VC. The value of the counter is used to decide whether a new frame can be accepted inside the FIFO buffer. In addition, a background process updates the counters at a rate function of the MCR of each VC and the current utilization of the output link. When the first cell of a frame arrives, it is accepted inside the buffer provided that the counter associated with its VC is large enough; otherwise, the entire frame is rejected. The update of the per-VC counters is done in order to ensure that each counter receives a number of credits corresponding

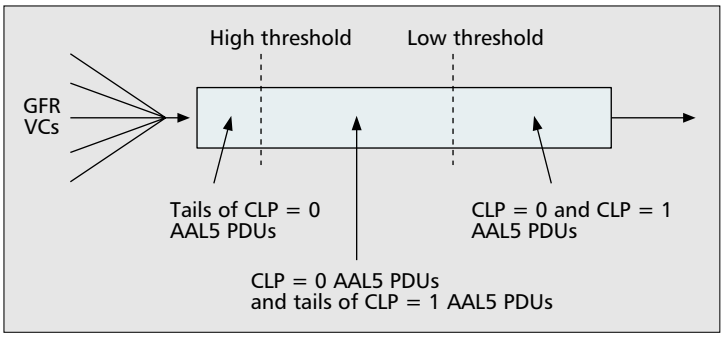


Figure 2. *The simple switch implementation.*

to its MCR. If there is some unused bandwidth on the output link, the counter update algorithm will distribute additional credits to each VC in proportion to their MCR. Additional details on the implementation of this algorithm may be found in [13]. The implementation proposed in [12] also relies on per-VC counters to determine whether an incoming frame should be accepted or not. These counters represent the number of cells from each VC inside the buffer, but they are not updated regularly as in the implementation proposed in [13].

Compared to the simple implementation discussed above, the counter-based implementations have the advantage of being able to support the two GFR conformance definitions. However, this is at the expense of maintaining at least one counter for each VC and implementing a possibly complex counter update algorithm.

Per-VC Threshold and Scheduling

This implementation combines a buffer acceptance algorithm with a per-VC scheduler. It was first proposed in [2]. It provides the bandwidth guarantees required to support the GFR service category by maintaining one logical queue for each GFR VC and by serving these queues with a WFQ-like scheduler at a rate at least equal to their MCR. The utilization of this scheduler guarantees that, when active, each VC will be allocated its reserved bandwidth as well as some fair share of the available excess bandwidth. Many schedulers have been proposed in the literature [15], and several have already been implemented in commercial products. In addition to providing the minimum guaranteed bandwidth, these schedulers often distribute the unreserved bandwidth in proportion to the MCR of each VC, although some schedulers provide a more flexible distribution of the unreserved bandwidth [6].

The scheduler ensures that each VC will be served at a rate at least equal to its MCR, but a switch must also ensure that a single VC will not be able to saturate the switch buffers. For this, the switch must intelligently discard frames when congestion occurs. This is done with a per-VC buffer acceptance algorithm (Fig. 3). This algorithm relies on a global counter for the occupancy of the complete buffer and on one counter for the occupancy of each per-VC queue. It is configured by specifying two buffer thresholds (not shown in Fig. 3) on the total buffer occupancy and one per-VC threshold. The low threshold is used to detect congestion. When the total buffer occupancy is above the low threshold, the switch is considered congested and only the CLP = 0 frames are accepted inside the buffer. Thus, CLP = 1 frames are discarded as soon as the total buffer occupancy is above the low threshold. The high threshold is used together with the per-VC thresholds. When the buffer occupancy is between the low and high thresholds, all incoming CLP = 0 frames are accepted. However, when the buffer occupancy is above the high threshold, a newly arriving CLP = 0 frame is only accepted if the occupancy of its queue is below its per-VC threshold. Otherwise, the arriving CLP = 0 frame is discarded. This is used to ensure that one VC will not be able to saturate the complete buffer and prohibit other VCs from utilizing the buffer. The values of the various thresholds will depend on the traffic contract of the established VCs.

The per-VC threshold and scheduling implementation is able to support the GFR.1 and GFR.2 conformance definitions, although it was designed with the GFR.1 conformance definition in mind.

Per-VC Scheduling and RED

Another example briefly discussed in this article is a modified version of the per-VC threshold and scheduling implementation proposed in [16]. In this switch implementation, the threshold-based frame discard mechanism is replaced by an active frame discard mechanism. Many researchers have argued that the utilization of such mechanisms would be beneficial for TCP/IP traffic [17].

In the per-VC scheduling and random early detection (RED) implementation (Fig. 4), the minimum guaranteed bandwidth is provided by the utilization of a per-VC scheduler. The buffer acceptance mechanism relies on the ATM version of RED described in [18]. Other variants of RED could have been used [19, 20]. This mechanism attempts to accomplish fair sharing of the buffer resources by maintaining accounting information for each per-VC queue. This buffer acceptance algorithm is configured by specifying two global thresholds (min_{th} and max_{th}, not shown in Fig. 4) and two thresholds on each per-VC queue (low[i] and high[i] for queue i). The values of these thresholds depend on the traffic contract of the GFR VCs.

When the occupancy of a per-VC queue is below its low threshold, this VC is considered underloaded, and all incoming frames are accepted inside this per-VC queue.

When the occupancy of a per-VC queue is between its two thresholds, CLP = 0 frames are accepted and the RED buffer acceptance algorithm decides to accept or reject the CLP = 1 frames. This algorithm works as follows. If the average occupancy of the complete buffer is below min_{th}, the incoming

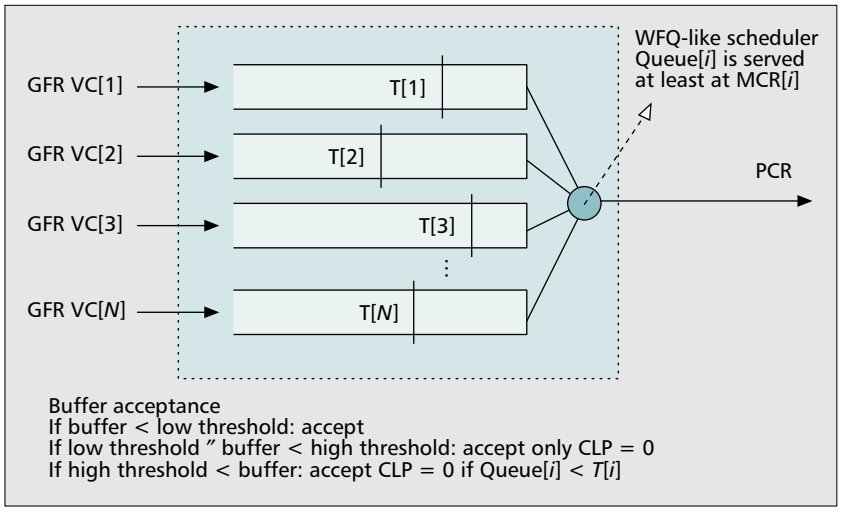


Figure 3. *The per-VC threshold and scheduling switch implementation.*

frame is accepted. If the average occupancy of the complete buffer is between min_{th} and max_{th}, the incoming frame is discarded with a probability that is a function of the average buffer occupancy. When the average buffer occupancy is equal to max_{th}, the incoming frame is discarded with probability max_p. max_p is another parameter of the buffer acceptance algorithm used by this implementation. If the average buffer occupancy is above max_{th}, the incoming frame is discarded.

When the occupancy of a per-VC queue is above its high threshold, it is considered overloaded. In this case, the RED buffer acceptance algorithm decides to accept or reject incoming (CLP = 0 and CLP = 1) frames from this VC.

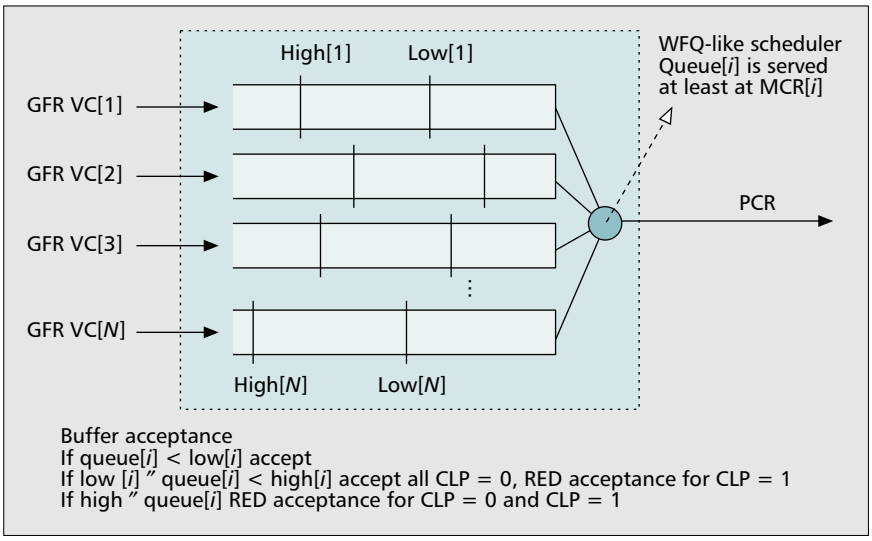


Figure 4. *The per-VC scheduling and RED switch implementation.*

Performance of TCP with GFR

Several researchers have used simulations to study the performance of TCP/IP with GFR service. These studies can be divided in two groups. The first simulations that appeared considered a single TCP connection to be carried by each ATM VC; thus, a minimum guaranteed bandwidth was associated with each TCP connection [11, 21–25]. Such a utilization of the ATM network would occur in a homogeneous ATM network where all the end systems are directly connected to the ATM network. Although this case applies to some existing ATM networks, it is not the most common utilization of ATM networks today. This case will be referred to as *workstation traffic* in the remainder of this section.

Another, more realistic utilization of ATM is as a backbone network [5, 6, 12, 26, 27]. In this case, the end systems are not directly attached to the ATM network but are attached to legacy LANs, and routers are used to multiplex the traffic originating from several end systems into a single ATM VC. In this case, each ATM VC carries the traffic corresponding to a potentially large number of TCP connections. We will use *internetwork traffic* when discussing simulations were each ATM VC carries several TCP connections. This utilization of the ATM network is much closer to the original deployment ideas of the proponents of GFR [2] than the first one.

To evaluate the performance of TCP in these two environments, we relied on simulations with a modified version STCP [28]. STCP is an event-driven simulation tool designed to efficiently carry simulations of TCP in ATM networks. The main advantage of STCP over other simulation tools is that it utilizes the standard BSD 4.4 Lite TCP implementation instead of a model of TCP. We have patched STCP to include the SACK implementation available from [29]. For all our simulations, we consider sources involved in file transfers. Each source continuously transfers files of a fixed size. Once a file has been successfully transferred, a new file is sent, and so on. This allows us to take into account the transients due to the establishment of the TCP connections, such as the initial slow start. The simulation results discussed in this section correspond to average values over a simulated time of 100 s, which is sufficient to obtain stable results.

In the two studied environments, we consider GFR VCs with different MCRs. The simulations are used to evaluate whether the TCP sources using these VCs can efficiently utilize the reserved bandwidth in the ATM layer. Since the performance of TCP is often influenced by the round-trip time, we consider two groups of VCs in all simulations to evaluate whether there is some unfairness for sources with larger round-trip times.

All our simulation results are presented in figures where we show the goodput achieved by the TCP connection(s) using each ATM VC. For each figure, we also plot the calculated expected goodput. This expected goodput is defined as the goodput the TCP connection(s) should achieve under ideal conditions (i.e., without frame losses and retransmissions but taking into account the ATM overhead and assuming that the unreserved bandwidth should be distributed to the established VCs in proportion to their MCRs). Thus, a simulation results close to the expected goodput will indicate that the switch implementation considered is able to efficiently support TCP/IP traffic.

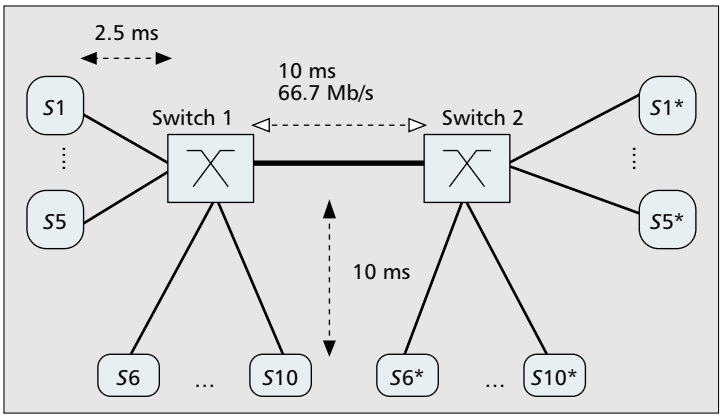
Workstation Traffic with GFR

Our first simulated network is an ATM end-to-end network (Fig. 5). This network is used to model an environment where high-performance workstations are directly connected to ATM switches with ATM adapters. We consider a network with 10 TCP sources (left of Fig. 5) sending large files to 10 TCP destinations. The TCP sources continuously transfer a 1 Mbyte file, and the window size of each workstation is large enough to completely utilize the available bandwidth. A bidirectional GFR VC is established between each pair of workstations. The MTU (maximum packet size) size for all the workstations was set at the default value for IP over ATM, 9180 bytes or 192 ATM cells. The TCP sources and destinations always send CLP = 0 frames.

The characteristics of the GFR VCs are summarized in Table 1. The backbone ATM switches utilize one of the GFR switch implementations described above, and the delay on the link between the two switches is set to 10 ms. The ATM backbone link was chosen so that 90 percent of the available bandwidth of this link is reserved for the GFR VCs. Since the total of the MCRs of the GFR VCs is equal to 60 Mb/s, the bandwidth of the backbone link was set to 66.7 Mb/s.

The results of these simulations are shown in Figs. 6 and 7, where the workstation-to-workstation TCP throughput is plotted for the 10 workstations as a function of their reserved bandwidth (MCR).

We consider GFR.2 conformance for the simulations with the simple switch implementation. The ATM switches were configured with a buffer size of 16,000 cells. This corresponds to buffer sizes of current ATM switches. The low buffer threshold was set to 2000 cells to avoid discarding CLP = 0



■ Figure 5. *A workstation traffic scenario.*

frames. No CLP = 0 frames were discarded during the simulations with this switch implementation.

Figure 6 illustrates the performance of the simple implementation. This figure corresponds to simulations with a large MBS (2000 cells) for the traffic contract of the GFR VCs. In this case, the simple switch implementation has difficulties efficiently supporting workstation traffic. For example, the workstation attached to a 10 Mb/s VC achieves a goodput of only 7.5 Mb/s when the simple implementation is used in the ATM switches. On the other hand, the 2 Mb/s workstation achieves a higher total goodput than its MCR. The low performance of TCP combined with this implementation is mainly due to poor interactions between the F-GCRA and TCP. TCP traffic is bursty, and the F-GCRA expects smooth traffic. Due to the burstiness of TCP traffic, the F-GCRA may mark a large fraction of the TCP packets as CLP = 1 frames, although the long-term average rate on the ATM VC is smaller than MCR. Furthermore, the F-GCRA has a tendency to mark TCP traffic in bursts. These bursts of CLP = 1 frames are subject to discarding inside the ATM switches, and these burst losses are not recovered by TCP's fast retransmit algorithm. This effect is especially important for sources having to fill a large reserved bandwidth. Simulations carried out with smaller values for the MBS showed that the TCP performance was worse with lower values for the MBS. The simulations did not reveal a significant impact of the low buffer threshold on the performance provided that it is not too small [6].

We consider GFR.1 conformance for the simulations with the two scheduler-based switch implementations. We have implemented the Virtual Spacing [30] scheduler in the STCP simulator. For the per-VC threshold and scheduling implementation, the global thresholds were set to 2000 and 12,000 cells, respectively. We used an MBS of 320 cells for the GFR VCs, and the per-VC thresholds were equal to the MBS. For the per-VC scheduling and RED implementation, $\max_p$ was set to 10 percent, $\min_{th}$ to 2000 cells and $\max_{th}$ to 6000 cells.

The performance of TCP with these two switch implementations is much better, as shown in Fig. 7. The scheduler-based implementations clearly outperform the simple implementation. The goodput achieved by the TCP sources is now much closer to the expected goodput, although the workstations with high MCRs are somewhat penalized.

However, with all switch implementations we notice the unfairness between the TCP connections with 15 ms and 30 ms round-trip times, a problem inherent to TCP's congestion control algorithm. The best bandwidth differentiation is obtained with the RED implementation and, for the 15 ms delay TCP sources, at the expense of the bandwidth differentiation of the 30 ms TCP sources. The slight performance gain of the RED implementation is mainly due to the fact that this switch implementation tries to avoid losing bursts of TCP packets. This has a positive impact on the TCP goodput by reducing the number of expirations of the retransmission timer.

Internetwork Traffic with GFR

Our second simulated network is shown in Fig. 8. This network is representative of many deployed IP over ATM networks in use today where the IP traffic is first aggregated by routers before being sent on an ATM backbone. For this environment, we have considered an ATM network which is used to interconnect 20 switched Ethernet LANs. Each switched Ethernet LAN is modeled as containing 10 workstations attached with 10 Mb/s point-to-point links to an aggregation router. The aggregation router aggregates all the traffic from the workstations and sends it through the ATM backbone network to its corresponding router, which delivers the traffic to the workstations on the remote LANs. All the traffic from one workstation on one LAN on

Workstation pair	PCR	MCR	Delay
1 to 1*	155 Mb/s	2 Mb/s	15 ms
2 to 2*	155 Mb/s	4 Mb/s	15 ms
3 to 3*	155 Mb/s	6 Mb/s	15 ms
4 to 4*	155 Mb/s	8 Mb/s	15 ms
5 to 5*	155 Mb/s	10 Mb/s	15 ms
6 to 6*	155 Mb/s	2 Mb/s	30 ms
7 to 7*	155 Mb/s	4 Mb/s	30 ms
8 to 8*	155 Mb/s	6 Mb/s	30 ms
9 to 9*	155 Mb/s	8 Mb/s	30 ms
10 to 10*	155 Mb/s	10 Mb/s	30 ms

■ Table 1. *Characteristics of the GFR VCs for the workstation scenario.*

Router pair	PCR	MCR	Delay
1 to 1*	34 Mb/s	2 Mb/s	15 ms
2 to 2*	34 Mb/s	4 Mb/s	15 ms
3 to 3*	34 Mb/s	6 Mb/s	15 ms
4 to 4*	34 Mb/s	8 Mb/s	15 ms
5 to 5*	34 Mb/s	10 Mb/s	15 ms
6 to 6*	34 Mb/s	2 Mb/s	30 ms
7 to 7*	34 Mb/s	4 Mb/s	30 ms
8 to 8*	34 Mb/s	6 Mb/s	30 ms
9 to 9*	34 Mb/s	8 Mb/s	30 ms
10 to 10*	34 Mb/s	10 Mb/s	30 ms

■ Table 2. *Characteristics of the GFR VCs for the internetwork scenario.*

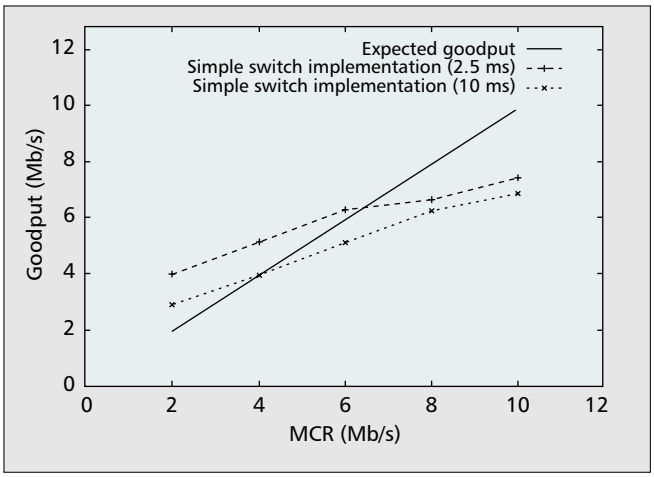


Figure 6. *The performance of workstation TCP traffic with the simple switch implementation.*

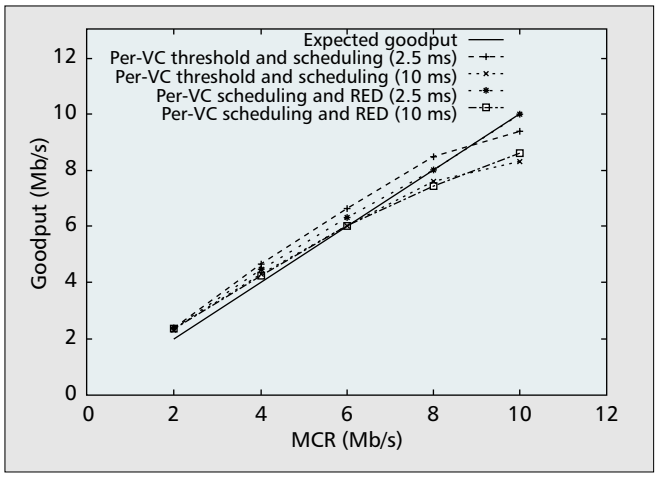


Figure 7. *Performance of workstation TCP traffic with WFQ-based switch implementation.*

the left side of Fig. 8 goes to a corresponding workstation on the corresponding LAN on the right side of Fig. 8. The workstations on the left side of the figure continuously send 1 Mbyte files. The MTU size of the workstations was set to the default MTU size of IP over Ethernet, 1500 bytes or 32 ATM cells. The routers are tail-drop routers with buffers large enough to avoid packet losses. The characteristics of the GFR VCs used in this environment are summarized in Table 2.

For the simulations with internetwork traffic, we set the low threshold to 2000 cells for the simple switch implementation, and the MBS of the traffic contract was set to 320 cells (10 maximum size Ethernet packets). For the per-VC threshold and scheduling implementation, the global thresholds were set to 2000 and 12,000 cells, respectively, while the per-VC threshold was equal to the MBS of the GFR VCs. For the RED implementation, $\max_p$ was set to 10 percent, $\min_{th}$ to 2000 cells, $\max_{th}$ to 6000 cells.

The simulations with the simple switch implementation

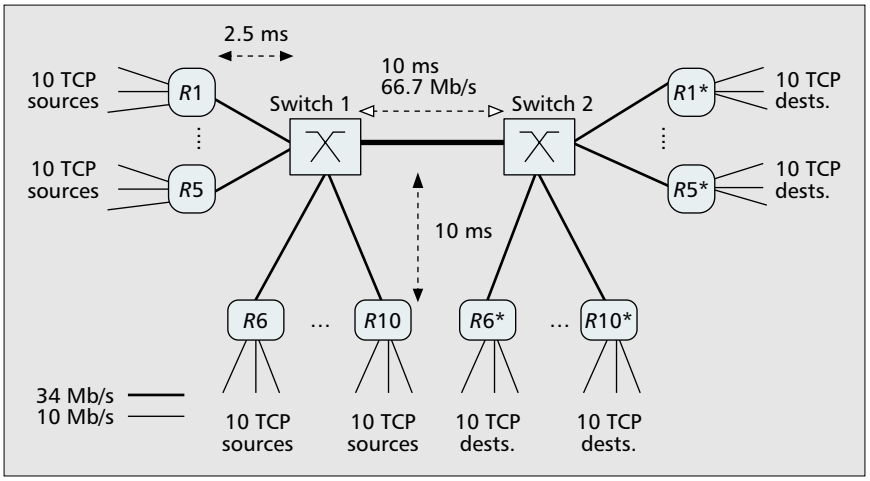


Figure 8. *An internetwork traffic scenario.*

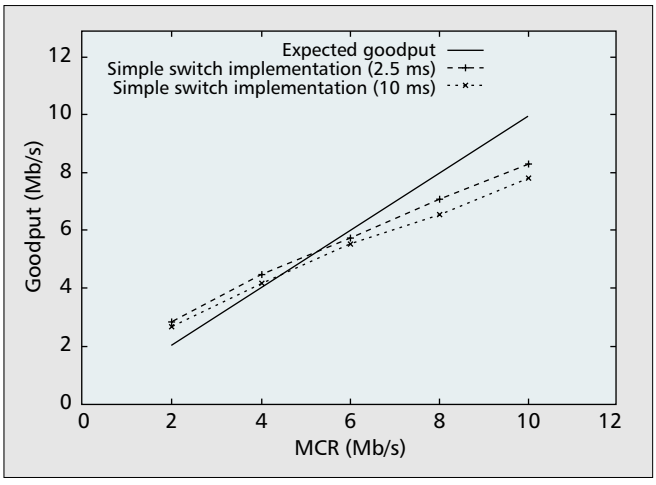
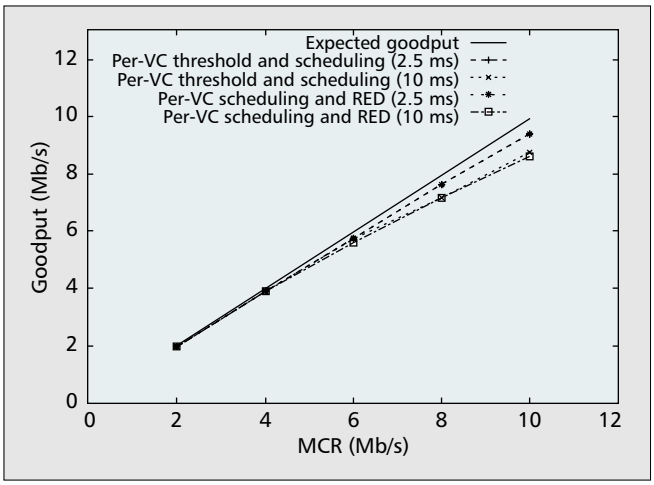


Figure 9. *The performance of internetwork traffic with the simple switch implementation.*

(Fig. 9) reveal that the performance of TCP with internetwork traffic is much better than with workstation traffic, although not perfect. This is explained by two factors. The first is that the burstiness of LAN traffic is smaller than that of workstation traffic. This implies that the F-GCRA will mark a lower percentage of the frames. The second is that internetwork traffic is less sensitive to burst losses. When a burst of packets is marked by the F-GCRA and later on discarded by the ATM switches, only a small number of the workstations attached to the routers are affected by these losses. With internetwork traffic, the influence of the MBS and low threshold on performance was lower than with workstation traffic [6].

As with workstation traffic, the simple switch implementation is clearly outperformed by the scheduler-based implementations (Fig. 10). Also, with internetwork traffic, these per-VC implementations continue to exhibit a bias toward TCP connections with smaller round-trip times. The overall goodput efficiency is 94.3 percent for the two studied scheduler based implementations. Apparently, the utilization of a RED-based buffer acceptance mechanism does not result in better overall TCP goodput. This is probably because the scheduler alone is responsible for most of the good performance, while the influence of the buffer acceptance algorithm is weaker. A potential drawback of the RED-based implementation is that many parameters need to be configured.



■ Figure 10. *The performance of internetwork traffic with WFQ-based switch implementations.*

Conclusion

In this article we first describe the main characteristics of the guaranteed frame rate service category. We show the important elements of the two GFR conformance definitions. We then discuss various switch implementations ranging from the simplest one to complex implementations relying on a per-VC scheduler. We then summarize the results of simulations carried out with three of these implementations. Our simulations in two different environments clearly show the limitations of the simplest switch implementation from a performance point of view with TCP/IP traffic. The two scheduler-based switch implementations we studied produced very good results, with a small advantage for the implementation relying on a RED-based buffer acceptance algorithm when each TCP connection is carried inside a single ATM VC. However, these two switch implementations proved to be equivalent when many TCP connections are carried inside a single ATM VC. From a performance point of view, we expect that switches will rely on per-VC schedulers to efficiently support TCP/IP with the GFR service category.

Our work and that done by many other researchers on the performance of TCP with GFR service can be summarized in a few informal lessons that are applicable to the performance of TCP with other technologies with some kind of bandwidth reservation such as the differentiated services under study within IETF:

- Any simulation study of TCP performance should consider at least workstation and internetwork traffic. These two types of traffic have different characteristics, and studying only one type of traffic would produce biased results.
- The traffic generated by one or a group of TCP sources is bursty, and TCP has some difficulties dynamically adapting its traffic pattern to a traffic contract enforced by a leaky-bucket-like mechanism. This imperfect adaptation to the traffic contract enforced by the policing unit usually implies that TCP will have difficulties to fully benefit from the reserved bandwidth unless the burstiness of the traffic is reduced with some kind of shaper, as discussed in [6].
- The performance of TCP is an end-to-end problem and many factors — the source TCP implementation, the shaping and/or policing devices, and the buffer acceptance, queuing, and scheduling strategies of the switches and the destination TCP implementation — influence this end-to-end performance.

Although these lessons are based on research on TCP over ATM, they should also be considered by researchers studying the performance of TCP over other network technologies, including (but not only) pure IP.

Acknowledgments

We would like to thank Sam Manthorpe for STCP. This work was carried out while the two authors were with Alcatel Alsthom Corporate Research Center in Antwerp, Belgium. It was partially supported by the Flemish Institute for the promotion of Scientific and Technological Research in the Industry (IWT).

References

[1] *IEEE Network*, vol. 9, no. 2, Mar./Apr. 1995.
[2] R. Guerin and J. Heinanen, "UBR+ Service Category Definition," ATM Forum ATM96-1598, Dec. 1996.
[3] ATM Forum, "Traffic Management Specification, Version 4.1," af-tm-0121.00, http://www.atmforum.com, Mar. 1999.
[4] S. Fahmy *et al.*, "Quality of Service for Internet Traffic over ATM Service Categories," *Comp. Commun.*, vol. 22, no. 14, 1999, pp. 1307–20.
[5] F. Hellstrand and A. Veres, "Simulation of TCP/IP Router Traffic over ATM using GFR and VBR.3," ATM Forum 98-0087, Feb. 1998.
[6] O. Bonaventure, "Integration of ATM under TCP/IP to Provide Services with Minimum Guaranteed Bandwidth," Ph.D. thesis, Univ. of Liege, Mar. 1999.
[7] ATM Forum, "Addendum to Traffic Management v4.1 for an Optional Minimum Desired Cell Rate indication for UBR,"str-tm-mdcr-01.01, straw ballot, May 2000.
[8] ATM Forum, "Addendum to TM 4.1: differentiated UBR," str-tm-diff-ubr-01.00, straw ballot, May 2000.
[9] F. Cerdan and O. Casals, "Mapping an Internet Assured Service on the GFR ATM Service," *Networking 2000*, May 2000, no. 1815 in Springer Verlag LNCS.
[10] I. Andrikopoulos *et al.*, "Providing Rate Guarantees for Internet Application Traffic Across ATM Networks," *IEEE Commun. Surveys*, vol. 2, no. 3, 1999, http://www.comsoc.org/pubs/surveys
[11] O. Elloumi and H. Afifi, "Evaluation of FIFO-based Buffer Management for TCP over Guaranteed Frame Rate service," *Proc. IEEE ATM '98 Wksp.*, Fairfax, VA, May 1998.
[12] R. Goyal, "Traffic Management for TCP/IP over Asynchronous Transfer Mode (ATM) Networks," Ph.D. thesis, Ohio State Univ., 1999.
[13] K. Siu, Y. Wu, and W. Ren, "Virtual Queuing Techniques for UBR+ Service in ATM with Fair Access and Minimum Bandwidth Guarantee," *GLOBECOM '97*, 1997, pp. 1081–85.
[14] F. Cerdan and O. Casals, "A Per-VC Global FIFO Scheduling Algorithm for Implementing the New ATM GFR Service," *MMNS-98 2nd IFIP/IEEE Int'l. Conf. Mgmt. of Multimedia Networks and Services '98*, Nov. 1998.
[15] H. Zhang, "Service Disciplines for Guaranteed Performance Service in Packet-Switching Networks," *Proc. IEEE*, vol. 83, no. 10, Oct. 1995.
[16] J. Nelissen and S. De Cnodder, "A Versatile RED-based Buffer Management Mechanism for the Efficient Support of Internet Traffic," *Proc. SPIE '99 Conf., Internet II: Quality of Service and Future Directions*, R. Onvural, S. Civanlar, and J. Luciani, Eds., Sept. 1999, pp. 101–12.
[17] B. Braden *et al.*, "Recommendations on Queue Management and Congestion Avoidance," Internet RFC 2309, Apr. 1998.
[18] O. Elloumi and H. Afifi, "RED Algorithm in ATM Networks," *IEEE ATM '97 Wksp.*, Lisboa, Portugal, 1997.
[19] M. Labrador and S. Banerjee, "Packet Dropping Policies for ATM and IP Networks," *IEEE Commun. Surveys*, vol. 2, no. 3, 3rd qtr. 1999, pp. 2–14.
[20] V. Rosolen, O. Bonaventure, and G. Leduc, "A RED Discard Strategy for ATM Networks and its Performance Evaluation with TCP/IP Traffic," *ACM Comp. Commun. Rev.*, July 1999.
[21] O. Bonaventure, "A Simulation Study of TCP with the GFR Service Category," *High Perf. Networks for Multimedia Apps.*, Kluwer, 1998.
[22] S. Pappu and D. Basak, "TCP over GFR Implementation with Different Service Disciplines: A Simulation Study," ATM Forum atm97-0310, Apr. 1997.
[23] R. Goyal *et al.*, "GFR — Providing Rate Guarantees with FIFO Buffer to TCP Traffic," ATM Forum 97-0831, Sept. 1997.
[24] J. Huang, B. Lee, and S. Khorsandi, "Simulation Study of GFR Implementations," ATM Forum 98-1035, Dec. 1997.
[25] D. Wu and H. J. Chao, "TCP/IP over ATM-GFR," *Proc. IEEE ATM'98 Wksp.*, Fairfax, VA, May 1998.
[26] O. Bonaventure, "Providing Bandwidth Guarantees to Internetwork Traffic in ATM networks," *Proc. IEEE ATM '98 Wksp.*, May 1998.
[27] B. Lee, "GFR Dimensioning for TCP traffic," ATM Forum 98-0271, Apr. 1998.
[28] S. Manthorpe, "STCP 3.0 User Manual," Tech. rep., EPFL, Sept. 1996.
[29] J. Mahdavi, "Experimental TCP Selective Acknowledgment Implementation," http://www.psc.edu/networking/tcp.html, 1996.
[30] J. Roberts, "Virtual Spacing for Flexible Traffic Control," *Int'l. J. Commun. Sys.*, vol. 7, 1994, pp. 307–18.

Biographies

OLIVIER BONAVENTURE (Olivier.Bonaventure@info.fundp.ac.be) is professor at the Computer Science Dept. of the Univ. of Namur, Belgium, where he leads the networking group (http://www.infonet.fundp.ac.be). His research interests include traffic engineering, quality of service, and distance learning over the Internet.

JORDI NELISSEN (jnelissen@colt-telecom.be) is working as a network engineer at the Internet Engineering department of COLT Telecom in Brussels, Belgium.