Traffic Control in ATM Networks: Engineering Impacts of Realistic Traffic Processes

Ajit K. Jena‡, Adrian Popescu†, Parag Pruthi†, Ashok Erramilli†,

‡ Computer Center, IIT, Powai, Bombay - 400076, India (ajit@powai.cc.iitb.ernet.in)

† University of Karlskrona/Ronneby, Dept. of Telecommunications and Mathematics, S - 371 79 Karlskrona, Sweden (adrian@itm.hk-r.se)

† Bellcore, 445 South Street, Morristown, NJ 07960-6438, USA (parag@bellcore.com, ashok@bellcore.com)

Keywords: Traffic Control, ATM, Self-Similar, Engineering Impacts, Traffic Shaping, Traffic Policing

Abstract

This paper reviews the current state of the art in the rapidly developing areas of ATM traffic controls and traffic modeling, and identifies future research areas to facilitate the implementation of control methods that can support a desired quality of service without sacrificing network utilizations. Two sets of issues are identified, one on the impacts of realistic traffic on the efficacy of traffic controls in supporting specific traffic management objectives, and the other dealing with the extent to which controls modify traffic characteristics. These issues are illustrated using the example of traffic shaping of individual ON-OFF sources that have infinite variance sojourn times.

1 Introduction

In the recent past two fields of research have evolved almost independently of each other, namely broadband technologies and associated control, and traffic modeling. In the field of broadband technologies great advances have been made in overall capacity increase along with huge advances in transmission speeds. Also, broadband and notably Asynchronous Transfer Mode (ATM) technologies have received much attention in the literature and numerous studies have been performed on many aspects of control, dimensioning, provisioning, monitoring, etc. On the other hand, the discovery of the self-similar or fractal nature of bursty packet traffic has had a great impact on the underlying assumptions of data traffic models. Analysis of data sets from real traffic measurements have brought to light the very different nature of real traffic (i.e. long-range dependence vs. short-range dependence in the autocorrelation function of the arrival process) than that assumed in most of the ATM modeling studies. Such assumptions of the traffic arrival process are a key component in all broadband traffic modeling studies. This raises serious concerns about the applicability of a number of such studies in engineering, control and operations of broadband systems. The fact that packet traffic is inherently fractal or long-range dependent and most broadband studies assume traffic to be short-range dependent leads one to wonder the extent to which the results of these studies are applicable in practice.

Under the present circumstances where current data networks are designed and operated based on models which are not representative of the traffic measured from actual working data networks, our ability to manage and control broadband networks and services in "real time" when subjected to "real traffic" may be very limited. Also, practically useful management knowledge and engineering methods have played a major role in the general acceptance, rapid deployment and sucessful operation of new technologies and services and are expected to impact the growth of broadband services and technologies as well.

These technologies are expected to support, in an integrated fashion, many and diverse broadband applications with different and sometimes conflicting traffic management (such as no loss, high throughput for data vs. acceptable loss, small delays for video), ranging from voice and interactive data to image, full-motion video and bulk data. As such, a broad range of Quality of Service (QoS) guaran-

tees need to be supported simultaneously; and in order to provide the wide spectrum of acceptable QoS required by different services, it is necessary to properly design flexible flow control and bandwidth allocation mechanisms. Examples of such mechanisms include call admission control, cell scheduling control, source policing, etc. which must be designed under realistic assumptions of network conditions such that each communication service is guaranteed an acceptable performance objective.

The main purpose of this paper is to examine the current state-of-the-art in these rapidly developing areas of ATM traffic control and traffic modeling and accordingly to identify future research issues needed to be undertaken in order to practically address issues related to ATM traffic control and management. To this end, we discuss some of the most important implications related to ATM traffic control under realistic assumptions about the incident traffic. Effectiveness expected from different basic control mechanisms (traffic shaping, Leaky Bucket, etc.) as well as from complex traffic management and congestion control are discussed as well. It is our belief that the operations of various proposed control mechanisms as well as rules for the setting of proper parameter values need to be reexamined; i.e. these mechanisms must be adequately characterized under realistic (or self-similar) incident traffic flows. We will provide preliminary evidence in support of our claims.

2 Role of Controls in Assuring QoS

One of the most important and challenging issues in the design of broadband networks is to develop an efficient and integrated framework within which requested end-to-end QoS guarantees are fully supported. The ATM Forum Traffic Management (TM) working group has recently completed the TM 4.0 specification [1], according to which five service categories have been identified. These are the Constant Bit Rate (CBR), the Real-time Variable Bit Rate (rt-VBR), the Non-real-time Variable Bit Rate (nrt-VBR), the Unspecified Bit Rate (UBR), and the Available Bit Rate (ABR). Many diverse (broadband) applications are expected to be supported within these five classes of services. Each one of these classes of traffic imposes various interactions at various levels of the protocol hierarchy such as interactions between the window-based flow and congestion control of TCP/IP and the preventive congestion control scheme of ATM; and each level may have very different requirements which may make service deployment onerous.

What makes the design and management of broadband networks so complex is that in the presence of disparate service requirements (some of which will be known only as future applications and services are developed) each individual negotiated QoS must be met; examples of applications which lead to these disparate service requirements range from voice and interactive data to image, full-motion video and bulk data [2]. Such different classes of traffic may require deterministic or statistical bounds on the various QoS parameters such as throughput, cell loss, cell delay and cell delay jitter. These parameters represent performance objectives expected from the network for the duration of a connection. The problem of traffic management when integrating communication protocols from both existing and future traffic types and when required to provide bounds on QoS on a per connection basis becomes indeed a very challenging task.

How can a network provider guarantee the potentially wide spectrum of QoS requirements? The ATM Forum [1] has concluded that a number of traffic management mechanisms are necessary to properly provide the expected levels of service from the network. Such management mechanisms include Traffic Shaping, Resource Management, Call Admission Control, Usage Parameter Control, Network Parameter Control, Priority Control, etc. The basic philosophy is that by the use of such control mechanisms the traffic can be sufficiently regulated at the edges of the network (by shaping, feedback control, blocking etc) so as to make the traffic flow within the network conform to behavior such that QoS guarantees can be met. In a nutshell, the wide spectrum of acceptable degrees of QoS can in principle be met by properly designing traffic control and bandwidth allocation mechanisms.

The main goal of a control mechanism in non-linear (e.g. network) systems is to alter the incident stream in such a way to allow trajectories previously unstable to become stabilized or to allow specific trajectories to become possible [11]. Both methods can be seen as a way to synchronize the

original dynamics by specific controls such as to obtain a combined system where a selected form of regular behaviour is stable. These controls may act either on specific parameters or on dynamical variables in the incident stream of the original system, enlarge the number of degrees of freedom in the dynamics and change the evolution of vestiges. In the case of parametric control it is necessary to utilize the properties of the non-linear system to achieve the linearized control in an exponentially rapid fashion. On the contrary, the second control method is more advantageous in the sense that it does not require detailed information about the stable and unstable manifolds of the incident stream.

Such topics are very important subjects in ATM research and have received a lot of attention in the literature and a variety of functional control mechanisms have been proposed. There is generally a trade-off between the levels at which QoS requirements are met and efficiency with which network resources are utilized. For instance, four general categories of services can be distinguished according to this trade-off, namely services with deterministic guarantees (so-called hard guarantees) with worst-case allocation [3], [4], services with statistical guarantees (so-called soft guarantees) with probabilistic allocation [5], [6], [7], prediction-based services with measurement-based admission control [8], and services with feedback-based control [9]. Usually, networks with deterministic guarantees provide the best QoS guarantees but at the expense of lower efficiency in utilization of network resources whereas the other three approaches trade a better resource utilization for a potential QoS degradation.

All the studies on this topic have been based on traditional traffic models. Today, it has been established that traffic flows in packet based networks are self-similar or fractal [13], [17]. These traffic characteristics are very different from those assumed in traditional traffic models. Traditional traffic models assume that correlations in the traffic arrival process span a very limited range of time-scales (short-range dependence), whereas correlations in real traffic arrival processes span a wide range of time-scales (long-range dependence). There is a fundamental conflict between the long-range dependency of a self-similar traffic process and the low-pass filter behaviour of any network control mechanism (such as shapers and FIFO queues). It is inefficient to eliminate variability on many time scales with control mechanisms that behave as low-pass filters. Beyond this fact, there is evidence to indicate that certain controls can infact give rise to or enhance fractal characteristics e.g., the phenomenon of traffic synchronization. The introduction of feedback mechanisms converts networks into a complex ensemble of interacting non-linear systems, with possibilities for a wide range of dynamical systems behavior, such as bistability and chaos. When a fractal process is passed through a low-pass filter (with a limited bandwidth), the resulting process has two states, a transient state (non-linear effect) and a steady state (linear effect). The non-linear part manifests itself as a transient, oscillatory process, with characteristics different from the incident process but, most important, this is a large-deviations process [12]. The characteristics of this process depend, among others, on the power spectrum of the incident process, the characteristics of the low-pass filter and the way the energy (which is constant) is redistributed after passing through the low-pass filter. In other words, there is a disproportionate response for the transient part. There are a large number of ways that a response can be disproportionate (non-linear), but only one way to be proportionate (linear). It is important to notice that non-linearities of a specific type may lead to stability enhanced beyond that expected in a linear world (like, for instance, solitons), whereas non-linearities of another type may lead to instabilities much stronger than those expected in a linear world (for instance, the irreversible behaviour of unstable chaotic trajectories). There is also the possibility to turn estimates about the transient large-deviations bahaviour of the system into estimates about steady state (Freidlin-Wentzell theory [12]), but all in all there is an imperative for mathematical modeling, analysis and simulation of the non-linear processes in ATM networks as well as ways to balance them.

Some recent work in this area has shown that long-range dependent traffic processes drastically impact network performance ([13],[18]) and as such the effectiveness of controls needs to be reexamined. A key observation in recent research on self-similarity is related to the robustness of self-similar features as traffic flows through several queueing stages, or is subjected to diverse network controls, such as traffic shaping and policing [13]. While the short-range correlation structure of a traffic flow can be significantly altered by control mechanisms, the long-range dependency is essentially unaffected.

One particular area realated to traffic shaping and policing has been examined in the context of self-similar traffic flows. Specifically, the extent to which the observed fractal nature of traffic can be altered by shaping and policing controls has been the subject of two recent studies [13] [14]. It has been established in [13] and [14] that shaping and policing are relatively ineffective in eliminating variability on many timescales observed in aggregate network traffic, as well as in self-similar traffic processes such as Fractional Brownian Motion (FBM) model. The conclusion therefore is that diverse control mechanisms would have to incorporate very large buffers and incur extreme delays to shape the low frequency structure of traffic. In the next section, we extend these results to individual sources that behave (as observed in actual packet traffic) as ON/OFF sources with infinite variance sojourn time distributions in the ON and OFF states.

3 Shaping and Policing of Real Traffic

We will illustrate the point that the efficacy of broadband controls sensitively depends on the underlying traffic characteristics by considering the impact of shaping and policing controls on improving network capacities. A widely held belief, based on traditional short-range dependent models, holds that the burstiness of the traffic can be reduced significantly through the use of shaping, thereby improving usable network capacities. Recent studies ([13], [14]) have examined the effect of shaping on aggregate long-range dependent traffic, and found that reductions in the overall variability of the process are not sufficient to result in significant improvements in usable capacity. Specifically, it has been demonstrated that eliminating the low frequency or long-range dependent structure of the traffic will require enormous buffering at the shaper - thereby incurring performance penalties that offset any benefits of improved network capacities.

We will revisit this issue considering behavior at the individual source level. Note that ([13], [14]) primarily consider aggregate network traffic for which purely second-order characterizations (either in terms of power spectra or variance-time plots) are sufficient to describe the traffic. Traffic to a shaper can in principle behave as an aggregate of several individual traffic streams (e.g., in LAN interconnection services) or as a single ON/OFF source (e.g., in native ATM). In the earlier studies, the input-output behavior of a queue driven by fractal processes was considered, and it was shown that the variance of the output process had the same long-range dependent behavior as the input. We consider the effect of shaping on individual sources that behave in the ON-OFF manner observed in actual traffic traces i.e, the sojourn times in the two states have infinite variance.

There are numerous shaping arrangements and implementations, and we will consider two of these for illustration purposes. In the simplest scenario, we consider a shaper based on the Generic Cell Rate Algorithm (GCRA) acting on an ON-OFF source with infinite variance sojourn times, and the shaped traffic driving a single downstream queue. This scenario is representative of an arrangement in which shaping is done at the network access point, and the downstream queue represents shared network resources. The issues in such an arrangement are whether the queueing backlogs at the downstream queue can be reduced without causing prohibitive backlogs at the shaper. The GCRA virtual scheduling algorithm is parameterized by (I,L) where $1/I$ is the target output rate of the shaper and L is the limit parameter that determines the maximum number of cells that can be transmitted back to back at the peak rate, and we assume that the peak rate of the source is R. The end-to-end performance of such an arrangement is fairly easy to characterize on the basis of earlier results on the queueing behavior of traffic generated by high variability ON-OFF sources [18] - specifically

- in the trivial case of the shaper output rate $1/I > R$, there is never a backlog at the shaper, and the traffic is unchanged
- if $1/I < R$, the distribution of the shaper backlog follows a power law with the decay exponent being less than 1, so that the average shaper backlog is unbounded.
- The queue length at the downstream queue is also analogously either negligible (if the service rate $R_1 > 1/I$) or heavy-tailed (if $R_1 < 1/I$). Depending on the shaper and queue parameters, the net effect of this control arrangement is to relocate the bottleneck (and buffer requirements) from the downstream queue without any improvement in end-to-end performance.

Such catastrophic behavior is not observed in ON-OFF sources which exhibit less variability e.g., with geometrically distributed sojourn times. Given that it is unusual in a network setting to have a single source that can overload network resources, a more realistic scenario is one in which a number of such shaped sources are aggregated onto a network queue. The behavior of individual ON-OFF source is effectively determined by the shaper (more generally, by network flow controls). Consider a high-level abstraction where we ignore the specific details of source-control interactions, and instead focus on the effect of the controls - which is to alter the parameters of the ON-OFF source for e.g., reduce the peak rate *R* of the source, increase the length of the ON period, at least to a first order approximation. This abstraction is representative of an arrangement in which shaping is done within the end systems, and the local secondary storage (e.g., hard disk) in effect functions as the buffers for the shaper. The issue of interest here is whether the poorer performance due to the reduced rate*R* is offset by decreased queueing delays due to smoother traffic flows.

To gain insights into this issue, we will once again refer to a number of recent results in the analysis of self-similar traffic processes. The first relates the observed self-similarity in aggregate network traffic first to the high-variability of ON-OFF sources, which are in turn related to the emprirically observed Pareto distribution of computer system files. Consider the case where the filesizes are represented by a shifted Pareto i.e., the probability that a given filesize*B* exceeds *b* ($P(B > b) = (\theta_f / (\theta_f + b))^{\alpha}$). The distribution of the ON period *P(T>t)* of the source is closely related to this distribution through the relation *t = b/R*, and is also a Pareto distribution with a mean value of $\theta_f / (R(\alpha - 1))$, and the same shape parameter. The effect of shaping on aggregate network traffic can be inferred by using results that relate the parameters of the ON-OFF source to aggregate network traffic:

- the exponent of the sojourn time distribution, which is related to the Hurst parameter of the aggregate network traffic ($H = (3 - \alpha)/2$), is unchanged.
- As *R* is decreased, the average ON period increases as $\theta_f / (R(\alpha - 1))$. The impacts on the statistics of the source depend on additional assumptions on how the OFF period is affected by changes in *R*. We assume for purposes of this example that the overall mean rate is unchanged (this implies that average OFF period decreases when *R* is decreased).
- the impact on the variability of the aggregate traffic can be calcualted using results from [15] and [16]. Assuming that the exponent of the OFF period is greater than that of the ON period, and that the peak to mean rates of the sources is high, the variance *V(t)* of the traffic generated in a time interval *(0,t)* from an aggregate of *N* identical ON-OFF sources is $V(t) \propto t^{3-\alpha} R^{2-\alpha} / \sqrt{N}$

Using *V(t)* as a measure of the burstiness of the process, reducing *R* will reduce its burstiness. This is most effective on small timescales. On larger timescales, *V(t)* will scale as before as t^{2H} which will dominate the setting of usable network capacities. Thus while *V(t)* is reduced, it will not be sufficient to significantly improve capacities. In contrast, the reductions in*V(t)* can be significant with short-range dependent processes, once again illustrating the point that the effectiveness of controls is determined by underlying traffic characteristics. An intuitive explanation of this and the findings reported in [13] and [14] is that shapers operate as “low pass filters” and are effective in reducing high-frequency fluctuations. However, long-range dependence is a low frequency characteristic, and is hence relatively unaffected by shapers. Over sufficiently long timescales, the volume of traffic in both the free traffic and the controlled traffic should be the same (if not, this implies that the shaper has enormous backlogs).

A more refined analysis would take into account specific details of the shaper arrangement and implementation, examine the various permutations of impacts on ON and OFF periods, and model the fact that controlled sources may behave in a more general fashion, with multiple ON states corresponding to different rates for the offered traffic. It is anticipated that the qualitative conclusions of this first order analysis are essentially unchanged. For other compelling arguments on the difficulties of using shaping and policing to improve network efficiencies, see [14].

Finally, the conclusions of this section do not imply that shaping and policing controls do not have a role in traffic management. On the contrary, these controls are essential to maintain fairness and

protect performance perceived by users from those with atypically heavy usage. Shaping and policing enable QoS "guarantees", though this may come at the price of reduced utilizations.

4 Conclusions

The research in the area of broadband traffic controls can be organized into two sets of interrelated issues. First, it is important to study the extent to which real traffic characteristics impact controls, and develop and validate control mechanisms on this basis. The second set of issues deals with the extent to which traffic characteristics are modified by controls. These issues not only include the potential for controls to temper the fractal nature of the traffic, but in some cases, create or enhance it. We have illustrated some of these issues with an example on the interactions of realistic ON-OFF source behavior on traffic shaping. We model individual sources using the chaotic map formulation, and model the interactions between the source and a leaky bucket mechanism as a coupled system of non-linear mechanisms. Using a number of analytical results, we show that while the shaping and policing can reduce variability on shorter timescales, the overall reduction is not sufficient to significantly improve network efficiencies. Specifically, the Hurst parameter (which captures the asymptotic behavior of the autocorrelation function) is unchanged, and any attempt to eliminate the long-term variability of the process will incur substantial performance penalties at the shaper, while sharply increasing buffer requirements. The results of this example are significant in several ways. First, consistent with the findings in [13] and [14], the fractal characteristics in traffic are highly robust with respect to a variety of network operations, such as splitting, merging and queueing. Secondly, our results demonstrate that while shaping and policing are essential to maintain fairness in broadband networks, such controls cannot be relied to eliminate the high variability observed in real traffic.

Finally, the remarkable success achieved in being able to reduce chaotic motions to periodic motions by slightly changing the dynamical system with the addition of new degrees of freedom is quite impressive in a number of areas like, for instance, chaotic lasers, non-linear electric circuits, control in cardiac tissue, etc. [10], [11]. There is no doubt that similar methods of control based on parametric control or carefully selected external forcing can be developed in teletraffic as well.

References

[1] - ATM Forum Traffic Management Specification Version 4.0, 1996.

[2] - Garrett, M. W., "A Service Architecture for ATM: From Applications to Scheduling" IEEE Network, May/June 1996, pp. 8 - 14.

[3] - Knightly, E.W., H-BIND: A New Approach to Providing Performance Guarantees to VBR Traffic" Proc. of IEEE INFOCOM'96, San Francisco

[4] - Wrege, D.E., Knightly, E.W., Zhang, H. and Liebeherr, J.,"Deterministic Delay Bounds for VBR Video in Packet-Switching Networks: Fundamental Limits and Practical Tradeoffs" IEEE/ACM Transactions on Networking, June 1996.

[5] - Elwalid, A. and Mitra, D.,"Analysis, approximations and admission control of a multi-service multiplexing system with priorities," Proc. of IEEE INFOCOM'95, Boston, 1995.

[6] - Kesidis, G., Walrand, J. and Chang, C-S.,"Effective Bandwidths for Multiclass Markov fluids and other ATM Sources," IEEE/ACM Transactions on Networking, August 1993.

[7] - Zhang, H. and Knightly, E.,"Providing End-to-End Statistical Performance Guarantees with Bounding Interval Dependent Stochastic Models," Proc. of ACM SIGMETRICS'94, May 1994.

[8] - Jamin, S., Danzig, P., Shenker, S. and Zhang, L.,"A measurement-based admission control algorithm for integrated services packet networks," Proc. of SIGCOMM'95, Boston 1995.

[9] - Kanakia, H., Mishra, P. and Reibman, A.,"An adaptive congestion control scheme for real-time packet video transport," Proc. of ACM SIGCOMM'94, San Francisco. 1994.

[10] - West, B.J. and Deering, B.,"The Lure of Modern Science Fractal Thinking" Studies of Nonlinear

Phenomena in Life Sciences, No. 3, World Scientific Publishing Co. 1995

[11] - Abarbanel, H.D.L.,"Analysis of Observed Chaotic Data" Springer-Verlag, 1996.

[12] - Weiss, A.,"An Introduction to Large Deviations for Communication Networks" IEEE Journal on Selected Areas in Communications, Vol. 13, No. 6, pp. 938-952, 1995.

[13] A. Erramilli, O. Narayan and W. Willinger, "Experimental Queueing Analysis With Long-Range Dependent Traffic", IEEE Trans on Networking, April 1996.

[14] A.L. Neidhardt and A. Erramilli, "The Role of Shaping and Policing in Traffic Management", in preparation.

[15] M.S. Taqqu and J.B. Levy,"Using Renewal Processes to Generate Long Range Dependence", Dependence in Prob and Statistics, vol. 11, Progress in Prob. and Stat., pp 73-89,eds E. Eberlein and M.S. Taqqu (Birkhauser, Boston, 1986).

[16] F. Brichet, J. Roberts, A. Simonian and D. Veitch, "Heavy Traffic Analysis of Storage Model with Long Range Dependent On/Off Sources", preprint 1995.

[17] W. Willinger, M.S. Taqqu, R. Sherman and D.V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at Source Level" (Extended Version), preprint 1995.

[18] P. Pruthi, " An Application of Chaotic Maps to Packet Traffic Modeling", PhD thesis, Royal Institute of Technology, Stockholm, 1995.