Multiplexing gains in ATM networks

Z. Fan
Centre for Communications Systems Research,
University of Cambridge,
10 Downing Street, Cambridge, CB2 3DS, UK.
P. Mars
School of Engineering, University of Durham,
South Road, Durham, DH1 3LE, UK.
Tel: +44 1223 740110 Fax: +44 1223 740099
Email: zhong.fan@ccsr.cam.ac.uk, philip.mars@durham.ac.uk

Abstract

The focus of this paper is to investigate the effects of various traffic and switch characteristics on multiplexing gains and their implications for different bandwidth allocation and admission control algorithms. We show that the total multiplexing gain due to the independent combination of identical sources can be resolved into two factors, one expressing the advantage gained (by means of buffering) from the statistical rate variations *within* a source and the other expressing the efficiency of statistical multiplexing of i.i.d. streams (gain *across* sources). Both simulation and theoretical analysis illustrate that although bursty sources require more bandwidth, multiplexing gains are increasing with burstiness. The effective bandwidth approach is found to work well in the region with high buffer size to burst length ratio, high source utilization and small number of sources, whereas the Gaussian approximation performs well in the region with small buffer size to burst length ratio, high source utilization and large number of sources. Finally, we also give some quantitative information related to self-similar traffic, i.e., FBM. It is shown that even for LRD traffic with high-H values, it is possible to obtain higher multiplexing gains when a large number of independent sources with the same Hurst parameter are multiplexed for combined transmission.

Keywords

ATM, multiplexing gains, long-range dependence, admission control

---

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35353-1_28

D. Kouvatsos (ed.), *Performance Analysis of ATM Networks*

1 Introduction

The emerging high-speed asynchronous transfer mode (ATM) networks are expected to support a wide range of telecommunication services such as voice, data, video and image transfer, with different traffic characteristics and quality of service(QOS) requirements. In ATM, bandwidth allocation deals with determining the amount of bandwidth required by a connection for the network to provide the required QOS. There are two alternative approaches for bandwidth allocation: deterministic multiplexing and statistical multiplexing. In deterministic multiplexing, each connection is allocated its peak bandwidth. Doing so causes large amount of bandwidth to be wasted for bursty connections, particularly for those with large peak-to-average bit rate ratios. This goes against the philosophy of the ATM framework since it does not take advantage of the multiplexing capability of ATM and restricts the utilization of network resources. An alternative method is statistical multiplexing. In this scheme, a multiplexing gain is achieved as the capacity allocated to a group of bursty traffic streams is lower than the sum of their peak rates. Hence, statistical multiplexing allows more connections to be multiplexed in the network than deterministic multiplexing, thereby allowing better utilization of network resources.

Since the resource allocation algorithms will be used by network control functions such as connection admission control and network routing, the real-time requirements necessitate that the complexity of these algorithms should be kept low while still taking into account the characteristics and the desired QOS of the connections. The exact solutions are either intractable or when available, are computationally too complex to meet the real-time requirements. Therefore approximations have to be made. These resource allocation schemes can be divided into two main categories. The first category consists of those which take the buffering in the switches/multiplexers into account. However, in order to meet the real-time computation requirements, they fail to take into consideration the effects of statistical multiplexing across the sources sharing the buffer. A typical example of this kind of bandwidth allocation scheme is the well known effective bandwidth. The second category is of those which treat the switches/multiplexers as bufferless entities but take the statistical multiplexing between the sources into account. The Gaussian approximation approach falls into this category. The focus of this paper is to investigate the effects of various traffic characteristics and switch parameters on multiplexing gains and their implications for different bandwidth allocation and admission control algorithms.

Previous work [15] [16] on this subject have adopted a simplified bufferless model and assumed that the traffic sources are of Gaussian distribution. Although in that case explicit formulas are available, insights into the relationship between multiplexing gains and various traffic and switch characteristics

seem to be restricted. In this paper, we show that the total multiplexing gain due to the independent combination of identical sources can be resolved into two factors, one expressing the advantage gained (by means of buffering) from the statistical rate variations *within* a source and the other expressing the efficiency of statistical multiplexing of i.i.d. streams (gain *across* sources). Both simulation and theoretical analysis illustrate that although bursty sources require more bandwidth, multiplexing gains are increasing with burstiness. The effective bandwidth approach is found to work well in the region with high buffer size to burst length ratio, high source utilization and small number of sources, whereas the Gaussian approximation performs well in the region with small buffer size to burst length ratio, high source utilization and large number of sources.

2 The Model

Consider an ATM multiplexer queue with buffer size B fed by N i.i.d. ON/OFF fluid flow sources. For exponentially distributed ON and OFF periods, a source is completely characterized by three parameters, namely the peak rate R, the utilization p and the mean burst length b. Let x be the normalized buffer size with respect to burst length, i.e., $x = B/b$. This two-state fluid model has been chosen to capture traffic from diverse applications like variable bit rate (VBR) video, voice and data communications [13]. The multiplexing gain G, which is specified by the allowable QOS parameter, is the bandwidth saving due to statistical multiplexing over the case in which peak rate allocation was to be used. It is defined as follows:

$$G = NR/C, \tag{1}$$

where C is the link bandwidth needed to meet desired QOS (cell loss ratio, cell delay, jitter, etc) for the multiplexed stream of N sources. Here we use the buffer overflow probability(or cell loss probability) as the QOS parameter and denote it by ϵ.

The maximum possible value of gain is obtained when admission control is based on average bandwidth assignment, i.e., $C = NRp$. Therefore, the maximum multiplexing gain G is given by $G = 1/p$. Intuitively, this means that for highly bursty traffic, with $p \ll 1, G = 1/p$ can be quite large. However, this maximum gain cannot be attained in reality, because the average bandwidth assignment method is unacceptable in terms of cell loss.

3 Simulation and Analysis

3.1 The Effective Bandwidth Approach

Assume the notion of effective bandwidth is used to determine the bandwidth needed to meet desired QOS for a *single* source. It is denoted by e and satisfies $m \leq e \leq R$, where $m = Rp$ is the mean rate of the source. Then G can be written as:

$$G = \frac{R}{e} \cdot \frac{Ne}{C} = G_1 \cdot G_2 \tag{2}$$

where $G_1 = R/e$ is the statistical gain in a single source, and $G_2 = Ne/C$ is the multiplexing gain across sources.

To compute the effective bandwidth e, Guerin et al. [7] propose a simple and straightforward method which is based on conservative estimates of the buffer overflow probability ϵ. It includes the effect of the access buffer, but ignores the effect of statistical multiplexing between sources. The effective bandwidth for an individual connection is given as:

$$e = \frac{R}{2\alpha(1-p)}(\alpha(1-p) - x + \sqrt{[\alpha(1-p) - x]^2 + 4x\alpha p(1-p)}) \tag{3}$$

where $\alpha = \ln(1/\epsilon)$. For multiple sources, the same expression as in (3) can be used such that the following condition is satisfied:

$$e = \sum_{i=1}^{N} e_i \tag{4}$$

Note that, in general, the effective bandwidth is found to provide a conservative estimate of the bandwidth requirements of various sources, especially when the number of sources is large. That is why G_2 in (2) is expected to be greater than 1. Yet it is a useful tool in many situations because of its additive property, as shown by (4). More details on this method can be found in [7].

Next we present and discuss a number of numerical examples that illustrate the relationship between multiplexing gains and various traffic parameters. The *exact* bandwidth C in (1) is computed by iteratively solving the differential equations associated with the underlying queueing system [2]. There are a number of variables that directly impact the multiplexing gain of connections. We express G_1, G_2 and G as functions of the following parameters: the source utilization p or burstiness $1/p$, the ratio of buffer size to burst length x, and the number of sources N. Without loss of generality, we assume $R = 1$.

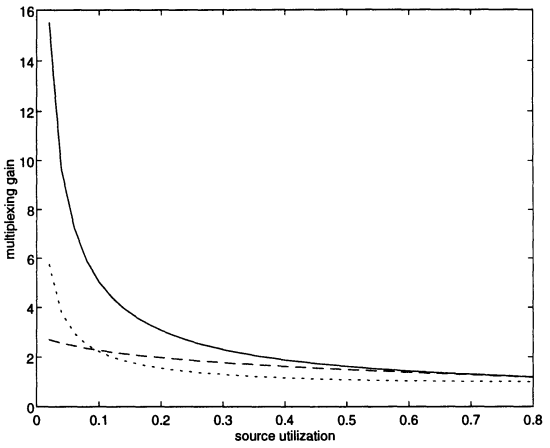

Figure 1: Effect of source utilization on multiplexing gains. $N = 50, x = 7.5, \epsilon = 10^{-5}$. dashed line: G_1, dotted line: G_2, solid line: G

3.1.1 Effect of Source Utilization

The first set of examples illustrate the effect of the source utilization on the multiplexing gains. In Figure 1 and Figure 2 we plot G_1, G_2, G as a function of p varying from 0.02 to 0.8. The other parameters are as shown in the figures. Obviously all the gains are decreasing functions of p. When $p \to 1$, i.e., all N connections have a constant bit rate R, all three gains tend to 1. It is easy to check that as $p \to 1$, the limit of e in (3) is R. However, the case $p \to 0$ exhibits two possible limits, depending on the sign of the quantity $(\alpha R - x)$.

$$\lim_{p \to 0} e = \begin{cases} 0, & \text{if } \alpha R \leq x \\ R(1 - \frac{x}{\alpha R}), & \text{if } \alpha R > x \end{cases} \tag{5}$$

Intuitively, (5) states that as $p \to 0$, the effective bandwidth also goes to 0 only if the buffer is large enough compared to the mean burst size, i.e., the buffer should be able to hold α bursts of average size. When the buffer size is not sufficient, the effective bandwidth has a nonzero limit since, although bursts are less and less frequent, the service rate must still handle large bursts whenever they arrive [7]. In Figure 1, since $\alpha R > x$, G_1 does not grow significantly as $p \to 0$. In Figure 2, since $\alpha R < x$, we observe a sharp increase of G_1 as $p \to 0$.

G_2 increases with decreasing p, since the effective bandwidth approach becomes more and more conservative. This can be readily explained from the

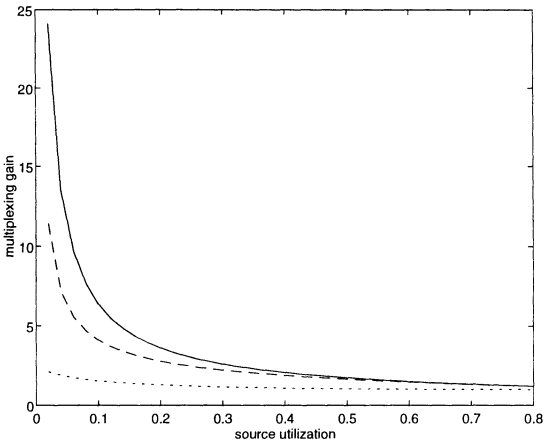

Figure 2: Effect of source utilization on multiplexing gains. $N = 50, x = 8.0, \epsilon = 10^{-3}$. dashed line: G_1, dotted line: G_2, solid line: G

origin of effective bandwidth. It is well known that the probability that the buffer content exceeds the buffer size $x, H(x)$, has the following asymptotic approximation [2]:

$$H(x) \sim D \exp(z_0 x) = A_N (NpR/C)^N \exp(z_0 x) \tag{6}$$

where A_N can be computed from all the negative eigenvalues of the queueing system, and z_0 is the largest negative eigenvalue. In general, D is a value different and sometimes significantly smaller than 1. Because the effective bandwidth method is based on the approximation that the prefactor $D = A_N (NpR/C)^N$ equals 1 and $H(x) = \exp(z_0 x)$, it will become more and more inaccurate as the source utilization decreases. It is clear that G increases when sources become more bursty.

3.1.2 Effect of Number of Sources

In Figure 3, G_1, G_2, and G are plotted as a function of number of sources, N. As e in (3) is independent of N, G_1 remains a constant as N varies. It can be seen from (6) that as N increases, D is expected to drop rapidly below 1. Therefore the effective bandwidth approach becomes more conservative and the gain accrued by combining and smoothing the source cell arrivals at the buffer, G_2, increases. A higher gain G is possible if the number of sources increases.

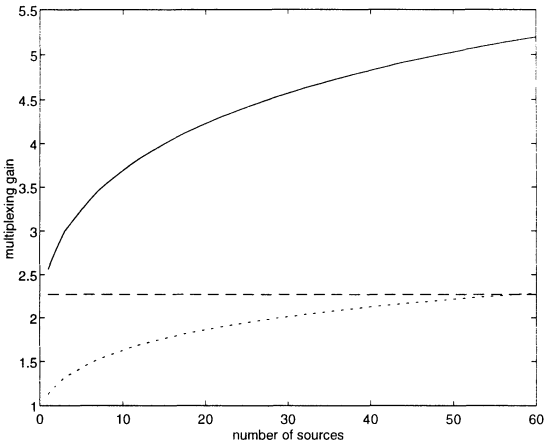

Figure 3: Effect of number of sources on multiplexing gains. $p = 0.1, x = 7.5, \epsilon = 10^{-5}$. dashed line: G_1, dotted line: G_2, solid line: G

3.1.3 Effect of Buffer to Burst Size Ratio

In Figure 4, G_1, G_2, and G are plotted as a function of buffer size to burst length, x. As intuitively expected, e in (3) can be easily found to be the mean bit rate pR and the peak bit rate R when $x \to \infty$ and $x \to 0$, respectively. Therefore G_1 is a monotone increasing function of x, ranging from $1/p$ to 1. On the other hand, since $H(x) \to \exp(z_0 x)$ when x becomes so large that $z_0 x \ll -1$ and $H(x) \ll 1$, the effective bandwidth is more and more accurate as x increases. So G_2 is a monotone decreasing function of x. In general, when we use larger buffers, the total gain achieved, G, is higher.

3.2 The Gaussian Approximation Approach

When the effect of statistical multiplexing is of significance, the distribution of the stationary aggregate traffic rate can be rather accurately approximated by a Gaussian distribution [7]. Let $M = Nm = NRp$ and $\sigma = R\sqrt{Np(1-p)}$ be the mean and standard deviation of the aggregate traffic of N sources respectively. Then the bandwidth required to meet the desired QOS ϵ is given by

$$g = M + k\sigma \tag{7}$$

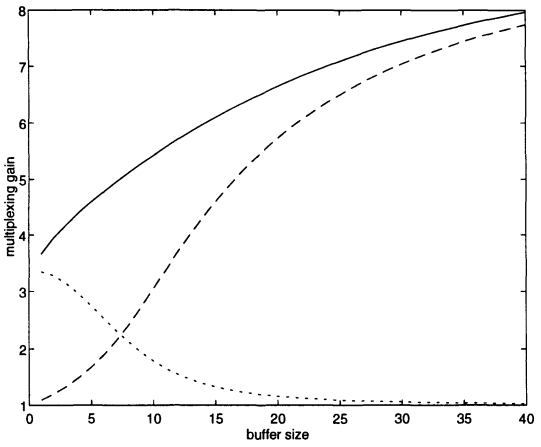

Figure 4: Effect of buffer to burst size ratio on multiplexing gains. $p = 0.1, N = 50, \epsilon = 10^{-5}$. dashed line: G_1, dotted line: G_2, solid line: G

where $k = \sqrt{-2\ln(\epsilon) - \ln(2\pi)}$. It is easy to check that when $1 > p > p_0 = \frac{N}{k^2+N}, g > NR$, which is meaningless. So we modify (7) to the following equation:

$$g = \min(M + k\sigma, NR) \tag{8}$$

Again, G can be resolved into two parts:

$$G = \frac{NR}{g} \cdot \frac{g}{C} = G_3 \cdot G_4 \tag{9}$$

where $G_3 = NR/g$ is the multiplexing gain across sources, and $G_4 = g/C$ is the statistical gain achieved by means of buffering. The problem with the Gaussian approach is that it ignores the multiplexing buffer completely, and relies on conservative bounds on the cell loss probability. That is why G_4 in (9) is expected to be greater than 1.

3.2.1 Effect of Source Utilization

The first set of examples illustrate the effect of the source utilization on the multiplexing gains. In Figure 5 and Figure 6 we plot G_3, G_4, and G as a function of p. The other parameters are as shown in the figures. Interestingly, we can see quite different behaviour of G_3 and G_4 against p in these two figures. As $p \to 1$, the limit of g in (7) is NR and G_3 tends to 1. As $p \to 0$,

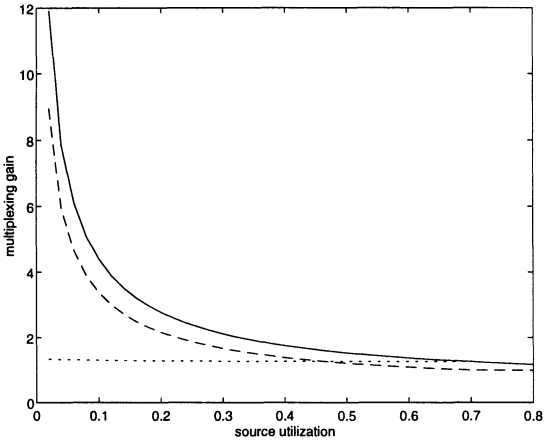

Figure 5: Effect of source utilization on multiplexing gains. $N = 60, x = 5.0, \epsilon = 10^{-6}$. dashed line: G_3, dotted line: G_4, solid line: G

the limit of g in (7) is 0 and G_3 tends to a very big number. Note that when $N \to \infty, p_0 \to 1$, so in this case G_3 is decreasing with p and unlikely to level off at 1. This reflects the fact that the Gaussian approximation is unreliable when the number of sources is not large enough, as is the case in Figure 6($N = 5$). In other words, the Gaussian approximation can only work well when sufficiently large number of sources are multiplexed together. The relationship between G_4 and p seems to be unclear and needs further investigation.

3.2.2 Effect of Number of Sources

In Figure 7, G_3, G_4, and G are plotted as a function of number of sources, N. Obviously, G_3 is a monotone increasing function of N, whereas G_4 is a monotone decreasing function of N. As $N \to \infty, G_3 \to 1/p$. This is consistent with the central limit theorem, i.e., as more and more i.i.d. sources are multiplexed together, the aggregate traffic is more Gaussian.

3.2.3 Effect of Buffer to Burst Size Ratio

In Figure 8, G_3, G_4, and G are plotted as a function of buffer size to burst length, x. As g in (7) is independent of x, G_3 remains a constant as x varies.

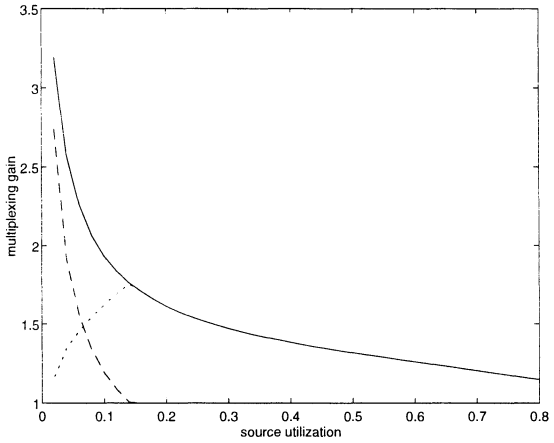

Figure 6: Effect of source utilization on multiplexing gains. $N = 5, x = 5.0, \epsilon = 10^{-7}$. dashed line: G_3, dotted line: G_4, solid line: G

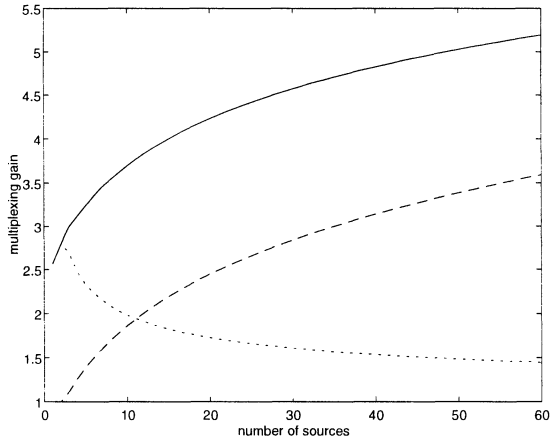

Figure 7: Effect of number of sources on multiplexing gains. $p = 0.1, x = 7.5, \epsilon = 10^{-5}$. dashed line: G_3, dotted line: G_4, solid line: G

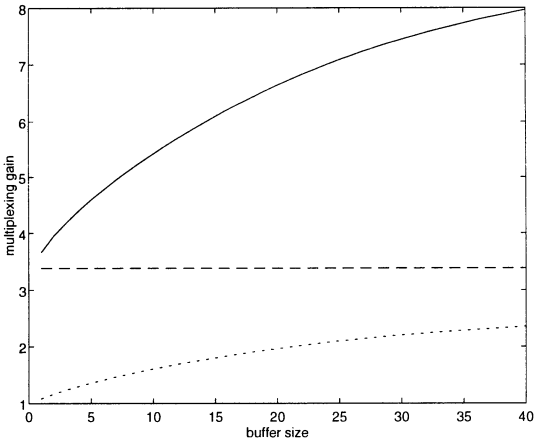

Figure 8: Effect of buffer to burst size ratio on multiplexing gains. $p = 0.1, N = 50, \epsilon = 10^{-5}$. dashed line: G_3, dotted line: G_4, solid line: G

On the other hand, since the Gaussian approximation fails to take into account the effect of access buffer, there is considerable gain of G_4 and this gain is increasing with buffer size.

4 Multiplexing Gains for Traffic with Long-Range Dependence

In this section we summarize some results related to traffic with long-range dependence, which have been published in a previous paper [6]. Readers can refer to that paper for more details.

Recent studies of real traffic data, mainly at Bellcore [11], have shown that Ethernet traffic cannot be sufficiently represented by traditional Markovian models, but instead can be more accurately matched by self-similar (fractal) models. More recently, variable-bit-rate(VBR) video traffic was also found to exhibit self-similar characteristics [3]. An important feature of self-similar processes is their long-range dependence(LRD), that is, their autocorrelation function decays less than exponentially fast. This property of persistent correlation can be characterized by the Hurst parameter H, with $H = 0.5$ for Markovian streams and $H > 0.5$ for streams with LRD. Studies by Norros [12], Erramilli et al. [5] suggest that LRD arrival processes could produce

much higher queue lengths and delays than Markovian sources, and that traffic engineering formulas based on Markovian models could, in some cases, result in under-engineering.

On the other hand, in their studies of the bandwidth needed on an ATM link carrying VBR video teleconference traffic, Heyman et al. [8] and Elwalid et al. [4] have shown that, even though the Hurst parameter of the VBR stream has been determined to be about 0.7 [3], effective bandwidth formulas derived for Markovian models are, in fact, successful in determining the bandwidth required to support LRD VBR streams.

Thus, there exist two somewhat opposing reports on the effect of LRD on traffic engineering, both from Bellcore. Krishnan [9] explains this apparent contradiction with the help of a 'crossover' effect of the Hurst parameter, i.e., when sufficiently large number of independent and identical sources are multiplexed, one can obtain a larger multiplexing gain with high-H sources than with low-H sources. In another work [10], Krishnan and Meempat demonstrate the crossover effect both for infinite and finite buffer queues with data traces of video teleconference.

Considering an infinite buffer queue with fractional Brownian motion(FBM) input, Krishnan [9] utilizes the stationarity and scaling property of the buffer-level process to derive the crossover result. Although the derivation is straightforward, his main results are *qualitative* and the relationship between crossover buffer size, number of sources and the Hurst parameter is not clear. At least two questions remain unanswered: (i) For a specific LRD arrival process with $H > 0.5$, under what buffer sizes(time scale) can an appropriate Markovian model($H = 0.5$) provide good(*conservative*) prediction of the cell loss rate? (ii) How many identical LRD sources need to be multiplexed together to achieve a higher multiplexing gain than the Markovian model? Here we just consider the second issue, while both issues are investigated in [6]. Our analysis is based on the large deviations estimates of the overflow probabilities for the FBM traffic model. Explicit formulas have been derived to give some *quantitative* insights into the impact of the Hurst parameter and its crossover effect on traffic engineering.

4.1 The Fractional Brownian Motion Model for Traffic with LRD

The fractional Brownian motion model has been used in [12] to successfully characterize the self-similar LAN traffic. Consider an FBM process $Z(t)$ with Hurst parameter $H \in [1/2, 1)$. It is a zero mean non-stationary Gaussian process with stationary increments and covariance structure $\mathrm{Cov}(t, s) = \frac{1}{2}(t^{2H} + s^{2H} - |t - s|^{2H})$. In the special case $H = 1/2$, $Z(t)$ is the stan-

dard Brownian motion. The self-similar property of $Z(t)$ is based on the fact that $Z(\alpha t)$ is identical in distribution to $\alpha^H Z(t)$. The increment process $X(k) = Z(k+1) - Z(k), k \geq 0$ is called fractional Gaussian noise(FGN) and is a stationary(discrete-time) Gaussian process with autocorrelation function $r(k) = 1/2(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}), k \geq 1$. It is easy to see that, asymptotically, $r(k) \sim H(2H-1)|k|^{2H-2}$, i.e., X exhibits LRD [11].

An ATM multiplexer can be modelled as a single-server queue with constant service rate C and buffer size B. Assume there are a large number N of homogeneous self-similar input traffic streams. Let $A(0,t]$ denote the distribution for the cumulative arrivals(cells or work) from each stream over the time interval $(0,t]$. $A(0,t]$ can be constructed as follows [12]:

$$A(0,t] = mt + \sqrt{ma}Z(t), \tag{10}$$

where $m > 0$ is the mean input rate, the scale factor $a > 0$ gives the variance/mean ratio for arrivals over one unit of the chosen time scale, and $Z(t)$ is the above described FBM with Hurst parameter H. Also let b and c denote respectively the amounts of buffer space and bandwidth per source, so that $B = Nb$ and $C = Nc$.

It has been shown in [6] that the buffer overflow probability for the above model can be given by

$$G(B,H) = \Pr(Q > B) \approx \exp[-N^{2H-1}\frac{(c-m)^{2H}B^{2-2H}}{2ma\kappa^2(H)}], \tag{11}$$

where $\kappa(H) = H^H(1-H)^{1-H}$.

4.2 The Hurst Parameter's Crossover Effect on Performance

The Hurst parameter of self-similar traffic has been widely regarded as a measure of *burstiness*, i.e., the higher the Hurst parameter, the burstier the traffic [11]. Starting from this point, one tends to conclude that traffic with LRD($H > 0.5$) may result in much more severe performance degradation than traffic with Markovian structures($H = 0.5$) [5]. Also, it has been claimed in the literature that the buffer behavior of LRD traffic cannot be accurately predicted by simple, parsimonious Markov-based models. However, it can be shown that a curious crossover property with respect to H for FBM traffic models exists and suggests that a high value of H does not necessarily imply that Markovian models will lead to under-engineering of bandwidth on ATM links.

Consider now the gain achieved by statistically multiplexing a large number of independent and identical sources with LRD. We express this multiplexing

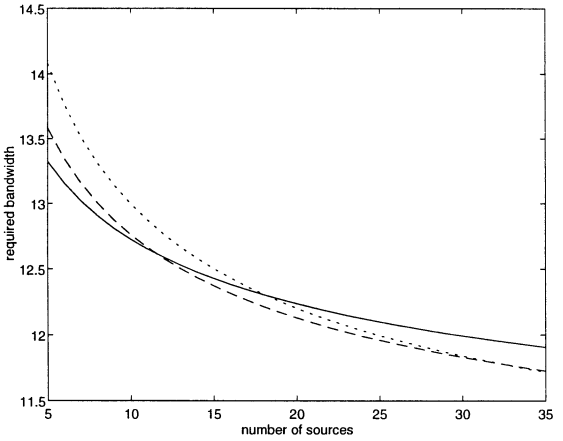

Figure 9: Statistical multiplexing gain. Solid line: $H = 0.7$, dashed line: $H = 0.8$, dotted line: $H = 0.9$

gain in terms of the required bandwidth per stream, c_0. Denote ϵ the target overflow probability $\Pr(Q > B)$. Then from (11), we have

$$c_0(N, H) = m + (-2\log(\epsilon)B^{2H-2}\kappa^2(H)am)^{\frac{1}{2H}} N^{\frac{1-2H}{2H}}. \qquad (12)$$

Note that when $H = 0.5$, $c_0(N, H)$ is independent of N, which is consistent with the concept of effective bandwidth.

Figure 9 shows $c_0(N, H)$ vs. N for different values of H. In this experiment, $\epsilon = 10^{-3}, B = 10, m = 10, a = 1$. The graph shows that the bandwidth required per source decreases with increasing number of sources. The crossover effect is obvious: when the number of multiplexed sources is large, the multiplexing gain with high-H sources is larger than that with low-H sources, and the converse is true for smaller number of sources.

Taking a closer look at Figure 9, we find that for different H, the numbers of sources at which the crossover happens are different. Although according to (12) there is no multiplexing gain across sources for $H_0 = 0.5$, we still take H_0 as a reference and this doesn't affect our main results much. We want to investigate how many independent, identical streams with LRD($H > 0.5$) need to be multiplexed together to achieve smaller c_0 than streams with $H_0 = 0.5$. Let's denote this crossover number of sources by N_{cr}. By using the

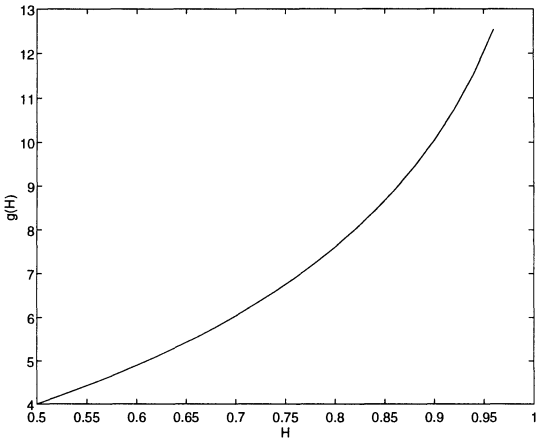

Figure 10: $g(H)$ vs. H

formula (12) and letting $c_0(N, H_0) = c_0(N, H)$, we have

$$N_{cr} = (-2\log(\epsilon)am)^{-1}B^2\frac{[\kappa(H_0)]^{\frac{2H}{H_0-H}}}{[\kappa(H)]^{\frac{2H_0}{H_0-H}}} = (-2\log(\epsilon)am)^{-1}B^2 g(H), \quad (13)$$

where $g(H)$ is a function of H and defined by

$$g(H) = \frac{[\kappa(H_0)]^{\frac{2H}{H_0-H}}}{[\kappa(H)]^{\frac{2H_0}{H_0-H}}}. \quad (14)$$

As shown in Figure 10, $g(H)$ is a monotone non-decreasing function of H. That means, for LRD traffic with a higher value of H, more identical sources should be multiplexed together to achieve possible higher gains. In other words, when we try to use Markov models ($H_0 = 0.5$) to provide a conservative estimate for the bandwidth needed for an LRD traffic with known $H > 0.5$, the higher the Hurst parameter H, the larger the number of multiplexed sources for which this estimate works well. This is confirmed by the results of VBR video traffic in [4]. From the traffic control point of view, we may state that for long-range dependent traffic, increasing the buffer size has little impact on reducing cell loss rate, since an input process with LRD generates occasional bursts of traffic that cannot be absorbed even by very large buffers. On the other hand, statistical multiplexing several streams is a very efficient way to reduce loss while keeping utilization high.

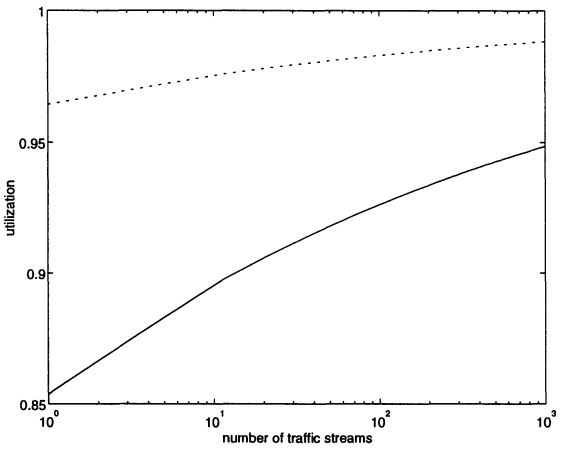

Figure 11: Utilization vs. number of traffic streams, $H = 0.6$. Solid line: $B = 100$, dotted line: $B = 1000$

In Figure 11 and Figure 12, the utilization level of the link as a function of number of multiplexed streams has been depicted. The traffic parameters are chosen as $m = 1, a = 1, \epsilon = 10^{-4}$. It can be seen that there will be quite significant multiplexing gains if the traffic grows by aggregating more and more of the same type of traffic streams. These gains are more apparent for traffic with a larger value of H. As a consequence, there is no reason to believe that the self-similar property of traffic will make it difficult for networks to achieve high levels of utilization. Note that in [1], Addie has got similar conclusions.

5 Conclusion

In this paper, we show that the total multiplexing gain due to the independent combination of identical sources can be resolved into two factors, one expressing the advantage gained (by means of buffering) from the statistical rate variations within a source and the other expressing the efficiency of statistical multiplexing of i.i.d. streams (gain across sources). The main findings of the paper are summarized in Table 1. The above results indicate that although bursty sources require more bandwidth, multiplexing gains are increasing with burstiness(the reciprocal of utilization). As the effective bandwidth approach is based on large buffer asymptotic and ignores the statistical multiplexing

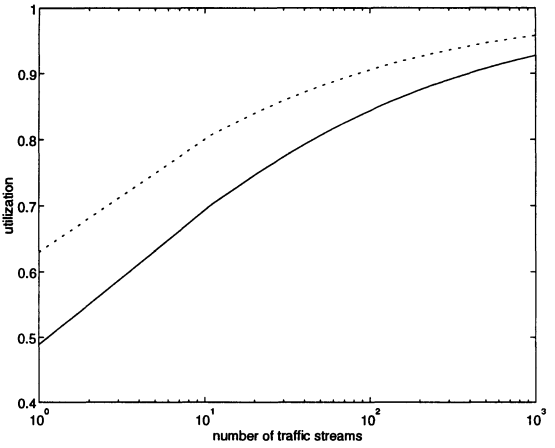

Figure 12: Utilization vs. number of traffic streams, $H = 0.8$. Solid line: $B = 100$, dotted line: $B = 1000$

Gain	Source utilization	Ratio of buffer size to burst length	# sources
G_1	D	I	C
G_2	D	D	I
G_3	D(for large N)	C	I
G_4	N	I	D
G	D	I	I

Table 1: Gains as functions of traffic parameters. I: increasing function; D: decreasing function; C: constant; N: non-monotonic

gain obtained by multiplexing different sources onto a single link, it can only work well in the region with high buffer size to burst length ratio, high source utilization and small number of sources. In the mean time, it is found that the Gaussian approximation performs well in the region where little gain is obtained by having buffers in the system and there is a large number of sources.

Based on the above analysis, resource allocation and admission control schemes should take both sub-gains and the trade-off between them into consideration so as to achieve a higher total multiplexing gain. It has been shown in [14] that a single bandwidth allocation algorithm does not cover the whole region of traffic situations. Therefore a possible solution is the effective combination of several methods which have strengths and limitations within different regions in the traffic space and complement each other. This is a promising direction of further study.

We have extended our discussion to traffic that exhibits LRD. Thanks to large deviations theory, we further Krishnan's work by investigating the crossover property in more detail for the FBM model. The relationship between crossover number of sources and the Hurst parameter is discussed. These results lead us to believe that even for LRD traffic with high-H values, it is possible to obtain higher multiplexing gains when a large number of independent sources with the same Hurst parameter are multiplexed together.

References

[1] R. G. Addie. Traffic will be more Gaussian in future. Technical report, University of Southern Queensland, 1996.

[2] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61:1871–1894, 1982.

[3] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Trans. Commun.*, 43(2):1566–1579, 1995.

[4] A. Elwalid, D. P. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE J. Selected Areas in Commun.*, 13(6):1004–1016, 1995.

[5] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Networking*, 4(2):209–223, 1996.

[6] Z. Fan and P. Mars. The impact of the hurst parameter and its crossover effect on long-range dependent traffic engineering. In *IEE 14th UKTS*, pages 10/1–10/8, 1997.

[7] R. Guerin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Selected Areas in Commun.*, 9(7):968–981, 1991.

[8] D. P. Heyman, A. Tabatabai, and T. V. Lakshman. Statistical analysis and simulation study of video teleconference traffic in ATM networks. *IEEE Trans. Circuits and Systems for Video Tech.*, 2(1):49–59, 1992.

[9] K. R. Krishnan. A new class of performance results for a fractional Brownian traffic model. *Queueing Systems*, 22(3):277–285, 1996.

[10] K. R. Krishnan and G. Meempat. Traffic engineering for VBR video with long-range dependence. In *1st International Conference on Broadband Communications*, pages 467–476, 1996.

[11] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic(extended version). *IEEE/ACM Trans. Networking*, 2(1):1–15, 1994.

[12] I. Norros. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.

[13] R. O. Onvural. *Asynchronous Transfer Mode Networks*. Artech House, 1994.

[14] S. I. A. Shah and T. Yang. ATM resource allocation algorithms: a comparison. In *IEEE GLOBECOM*, pages 365–370, 1995.

[15] I. Sidhu and S. Jordan. Multiplexing gains in bit stream multiplexors. *IEEE/ACM Trans. Networking*, 3(6):785–797, 1995.

[16] P. W. Tse and M. Zukerman. Evaluation of multiplexing gain. In *IEEE GLOBECOM*, pages 1216–1220, 1995.

Biographies

Zhong Fan received the BSc and MPhil degrees in electronic engineering from Tsinghua University, China, in 1992 and 1994, respectively. He also obtained a PhD from the University of Durham, UK in 1997. He is currently working as a Research Fellow at the University of Cambridge.

Philip Mars is Professor of Electronics and Director of the Center for Telecommunication Networks at the University of Durham, UK.