7

When Is Traffic Dispersion Useful? A Study On Equivalent Capacity

E. Gustafsson
Royal Institute of Technology, Dept. of Teleinformatics
KTH Electrum/204, S-164 40 KISTA, SWEDEN
Phone +46 8 752 14 98, Fax +46 8 751 17 93, Email: evag@it.kth.se

G. Karlsson
Swedish Institute of Computer Science
Box 1263, S-164 28 KISTA, SWEDEN
Phone +46 8 752 15 77, Fax +46 8 751 72 30, Email: gk@sics.se

Abstract

Multi-media and data traffic are anticipated to occupy much of the resources in integrated services networks, based on ATM. These traffic types appear to exhibit strong autocorrelation over long periods, which affects the performance of statistical multiplexing detrimentally. The correlation has most commonly been handled by spreading the traffic in time, so called shaping, which may introduce considerable delay.

We take a different approach, namely spreading the traffic in space over multiple, independent paths. The autocorrelation in the traffic is thereby reduced and bursts are spread out. This alleviates queuing delay and, for a given quality level, lowers the capacity needed for each transmission. We denote this strategy *traffic dispersion*.

In this paper, we focus on how traffic dispersion affects the equivalent capacity needed for a transmission. By studying its behaviour, we can determine under what circumstances spatial traffic dispersion is motivated for different cost functions, when using a certain number of paths in the network. The first cost function is a fixed charge per capacity unit. Next, we add a fixed charge per connection to the previous cost, and lastly, we let the charge per path increase progressively. Our findings show that spatial traffic dispersion alleviates the most troublesome traffic cases, that is, those with a high peak-to-mean ratio and those with a high peak-to-link ratio. Furthermore, the cost benefits due to dispersion seem to justify the extra effort needed to implement it.

This work was in part presented at the IFIP TC6 Third Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, U.K., July 1995.

Keywords

Traffic dispersion, multi-path routing, equivalent capacity, traffic control, ATM networks.

D. D. Kouvatsos (ed.), *ATM Networks*

1 INTRODUCTION

The asynchronous transfer mode (ATM) is the network architecture that the International Telecommunication Union recommends for broadband integrated services digital networks. Succinctly described, the mode combines the circuit switched routing of telephone systems with the statistical multiplexing of packet switching. This is accomplished by establishing a connection (fixed route) through the network before accepting any traffic. The information is then sent over the connection in 53-octet long cells, which are routed according to address information contained in their 5-octet headers.

The capacity of a transmission link is statistically shared among the connections traversing it. When traffic arrives randomly, the capacity offered by the link occasionally becomes insufficient. This could be handled by buffering, but as the arrivals come in longer and longer bursts the buffers will eventually overflow and cells will be lost.

Earlier studies have shown that the probability of cell loss is highly dependent on the correlation in the multiplexed traffic stream, Li (1989). For a given connection, the correlation can be lowered by spreading the traffic in time, so called shaping. This method may however give rise to delays too large to be tolerated by the application. So, statistical multiplexing of traffic streams with strong correlation would, at a low probability of cell loss, require unreasonable low utilization of the network resources and excessively large buffers.

Ever since Maxemchuk's contribution, Maxemchuk (1975), there have been several different suggestions for spreading the traffic from a source in space rather than in time, as a means for load balancing and fault handling in packet-switched networks, Gustafsson (1994:1). Spatial traffic dispersion means that a message is divided into a number of sub-messages, which are transmitted in parallel over disjoint paths in the network, as shown in Figure 1. A large burst of data will consequently be sent as more moderately sized bursts, and the correlation will be reduced without the extra delay that temporal shaping would introduce.

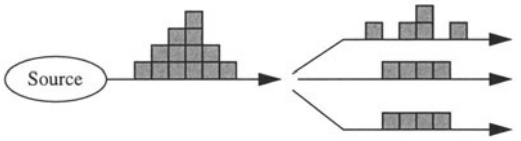


Figure 1 Illustration of spatial traffic dispersion.

The traffic from a source is transmitted in parallel through the network, and resequenced at the receiver. Dispersion should be possible in any network where disjoint paths exist between the source and the destination. Given a number of such paths, the traffic may be spread according to different strategies. One possibility is to spread the packets in the traffic stream cyclically over the paths - a solution which is discussed in Lee and Liew (1993), Maxemchuk (1993). Another way would be to submit the packets in longer sequences on each path, as suggested for the string mode protocol, Déjean et al. (1991), and yet another solution would be to spread the traffic dynamically over the paths, Cheng (1994). The latter variant would however require substantially more overhead. Essentially, a spreading strategy should apply to the traffic characteristics in order to minimize the correlation in the resulting traffic streams, since lowering the correlation is one of the main advantages of traffic dispersion.

Spatial traffic dispersion thus improves statistical multiplexing and it also enhances network security, as eavesdropping on several connections simultaneously may be difficult. Since the dispersion scheme employs disjoint paths, cell loss on one connection is independent of losses on other connections, and forward error correction can successfully be used to correct the losses. Regarding these advantages, the question arises whether traffic dispersion is useful under all circumstances.

To answer this question, or at least give a hint, we have chosen to focus on how traffic dispersion affects the equivalent capacity of a transmission. The equivalent capacity is the predicted capacity, that for certain source characteristics and demands on performance, needs to be allocated on a link. We investigate for what values of source peak rate and source mean rate, and their relation to the link capacity, spatial traffic dispersion is useful.

Equivalent capacity is discussed in Section 2, while Section 3 covers the cost functions used in the evaluations. Our results are presented in Section 4, and Section 5 concludes the paper.

2 EQUIVALENT CAPACITY

2.1 Equivalent capacity without buffering

The ATM concept makes use of statistical multiplexing which allows multiple sources to share a link statistically. This means that the demand for capacity at times may exceed the available resource on the link, and cells will be lost. Given a limit on the cell-loss probability, we can calculate the maximum number of identical sources n which can be multiplexed on a link of capacity C. The equivalent capacity required for one source is then C/n.

If the objective is to maintain the traffic arrival rate below the link capacity, that is, assuming no buffering, the cell-loss probability may be approximated by:

$$\varphi = \frac{E\{(\lambda - C)^+\}}{E\{\lambda\}}, \text{ where } \lambda \text{ is the arrival rate, and } (\lambda - C)^+ \text{ is max } \{0, \lambda - C\}. \quad (1)$$

Define the arrival process to consist of n independent identically distributed on-off sources, each with peak rate h (Figure 2). The model is chosen to get a tractable expression for the equivalent capacity while capturing some of the burstiness that can be anticipated from future traffic sources.

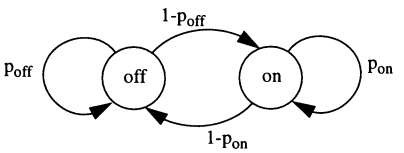


Figure 2 The state diagram of an on-off source. The system stays off with probability p_{off}, and once on, it remains on with probability p_{on}.

While in active state, the source generates traffic at peak rate h. The mean rate of the source is thus given by εh, where ε is the fraction of time that the source spends in active state:

$$\varepsilon = \frac{1 - p_{off}}{2 - p_{on} - p_{off}}. \quad (2)$$

As the number of simultaneously active sources is binomially distributed, the cell-loss probability is given by

$$\varphi = \frac{1}{n \cdot \varepsilon h} \sum_{x=\frac{C}{h}}^{n} (xh - C)\, Pr\{\text{x sources on}\} = \frac{1}{n\varepsilon} \sum_{x=\frac{C}{h}}^{n} \left(x - \frac{C}{h}\right) \binom{n}{x} \varepsilon^x (1-\varepsilon)^{n-x}. \tag{3}$$

The burstiness of a source is in these equations defined as

$$\frac{\text{Source peak rate}}{\text{Source mean rate}} = \frac{h}{\varepsilon h} = \frac{1}{\varepsilon}, \tag{4}$$

and the cell-loss probability is hence dependent on the ratio *C/h* as well as on the source burstiness.

Since the correlation between cells generated by the source in Figure 2 is monotonously decreasing, cyclic dispersion would minimize the correlation on each path. This is thus the dispersion strategy preferred, and it is assumed in the remainder of this paper. A dispersed source, as the link experiences it, may be approximated by another on-off source with the same characteristics except that the peak rate is reduced to *h/N*, *N* being the dispersion factor. The dispersion factor is defined as the number of paths over which the traffic from a source is spread. In the following, a dispersed source represents the traffic that an original source sends over one of the paths. We show the effects of dispersion on the equivalent capacity by replacing each original on-off source by *N* independent dispersed sources.

Essentially, what we do is modelling an on-off source with peak rate *h* as *N* on-off sources, each with peak rate *h/N* (Figure 3). These sources would be completely correlated, since they together represent the original source. With dispersion however, the traffic from each of these *N* sources is sent over a separate path, disjoint from all the other paths. Each link is therefore only affected by the traffic from one of the dispersed sources, and this source can be seen as the fraction of traffic that the original source sends over that specific link. In order to obtain the same load on a link with as without dispersion, we assume that the link instead of carrying the traffic from a number of independent original sources now carries the traffic from *N* times as many independent dispersed sources. That is, one link carries fractions of the traffic from each of *N* times as many original and independent sources. This justifies the independence criterion used in the capacity calculations.

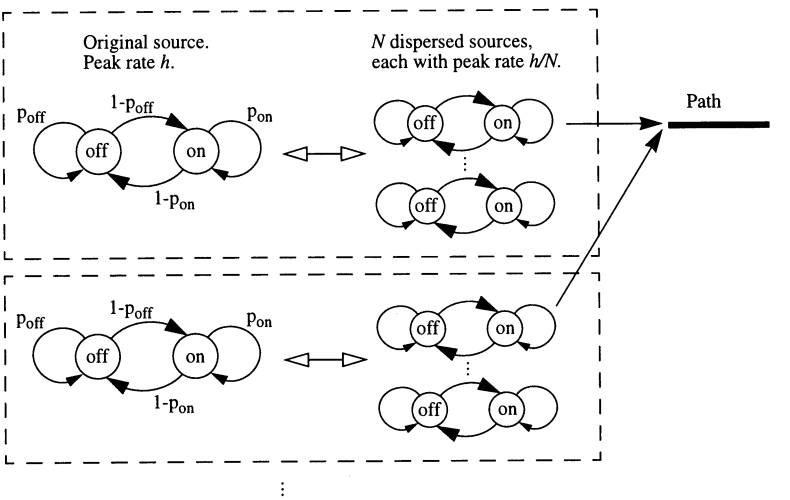


Figure 3 Modelling dispersed traffic sources. Without dispersion, a number of sources are multiplexed on a certain link, while in the case of dispersion, N times as many dispersed sources are multiplexed on the same link. The amount of traffic that the link carries is hence kept constant.

We can now calculate the equivalent capacity for a dispersed source, but we multiply it by N in order to get a fair comparison with the capacity for an original source. From (3), we get the cell-loss probability for n sources, each with peak rate h/N as

$$\varphi = \frac{1}{n\varepsilon} \sum_{x=\frac{NC}{h}}^{n} \left(x - \frac{NC}{h}\right)\binom{n}{x} \varepsilon^x (1-\varepsilon)^{n-x}. \tag{5}$$

For a given cell-loss probability, the capacity required for each source on each path is C/n, and the total capacity required for a source is NC/n. This is the capacity presented in the Figures.

Figure 4 shows for each value of N the aggregated equivalent capacity of N dispersed sources. The equivalent capacity for one source without dispersion is normalized to one. Note that this implies that the graphs for different values of C/h and burstiness are not directly comparable. The graphs only intend to show the multiplexing gain obtained by dispersion for different values of burstiness and source peak rate. The upper left graph shows a small increase in equivalent capacity for N=2 compared to N=1. The equivalent capacity for a dispersed source is in this case lower than the capacity for a non-dispersed source, but it still exceeds fifty percent of that value. When multiplied by two, it hence causes an increase in equivalent capacity. It should be noted that since the peak-to-link ratio is very high, only a few sources fit, given the zero-buffer assumption and the stringent loss requirement (10^{-9}). The peak in the graph is basically due to the fact that the number of sources has to be an integer. The truncation gives proportionally higher effect in this case since the number of multiplexed sources is very low (6 in the case without dispersion).

The figure shows that traffic dispersion decreases the equivalent capacity for a connection. When dealing with statistical multiplexing, the most troublesome traffic sources are those with a high peak-to-mean ratio and those with a high peak-to-link ratio. This is because it becomes extremely difficult to predict the amount of capacity which needs to be allocated in order to fulfil the performance requirements for heavily fluctuating sources. We can see that these sources are those where the benefits of traffic dispersion are most significant.

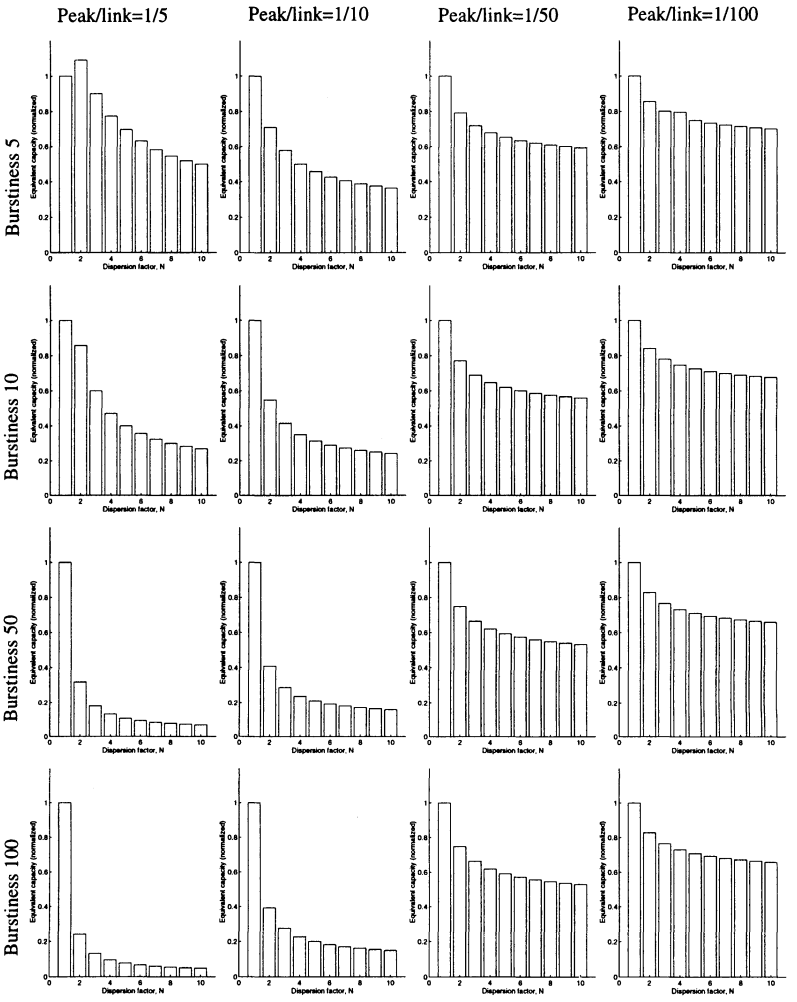


Figure 4 Equivalent capacity for different degrees of dispersion, and different values of source burstiness and peak-to-link ratio. The burstiness is defined as the source's peak rate divided by its mean rate, and 'Peak/link' denotes the source peak rate divided by the link capacity. The cell-loss probability was set to 10^{-9}.

In Figure 4, the values were normalized. Table 1 shows the equivalent capacity without dispersion (N=1) before normalization, given the link capacity C=1. This means that for a given column, the peak rate is constant, while the mean rate decreases for each row in order to make the source more bursty. For a given row, the peak rate as well as the mean rate decreases for each column, to make the source less dominant on the link. Table 2 shows the peak rates h and mean rates εh corresponding to the capacity values in Table 1.

A comparison of the two tables shows that the worst source behaviour is in the lower left box, while the best behaviour is in the upper right box. In the lower left box, the source burstiness is high, causing large fluctuations in the traffic, and the peak rate occupies a large part of the link capacity. The equivalent capacity for such a source is close to the peak rate. On the other hand, the source in the upper right box causes small fluctuations, never demanding more than a small fraction of the link, wherefore the equivalent capacity is close to the mean rate.

Table 1 Equivalent capacity before normalization; C=1, N=1, cell-loss probability 10^{-9}

$1/\varepsilon$	$h = 2.0 \cdot 10^{-1}$	$h = 1.0 \cdot 10^{-1}$	$h = 2.0 \cdot 10^{-2}$	$h = 1.0 \cdot 10^{-2}$
5	$1.7 \cdot 10^{-1}$	$9.1 \cdot 10^{-2}$	$8.3 \cdot 10^{-3}$	$3.3 \cdot 10^{-3}$
10	$1.7 \cdot 10^{-1}$	$7.1 \cdot 10^{-2}$	$4.5 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$
50	$1.4 \cdot 10^{-1}$	$2.3 \cdot 10^{-2}$	$9.6 \cdot 10^{-4}$	$3.6 \cdot 10^{-4}$
100	$1.0 \cdot 10^{-1}$	$1.2 \cdot 10^{-2}$	$4.8 \cdot 10^{-4}$	$1.8 \cdot 10^{-4}$

Table 2 Source mean rate

$1/\varepsilon$	$h = 2.0 \cdot 10^{-1}$	$h = 1.0 \cdot 10^{-1}$	$h = 2.0 \cdot 10^{-2}$	$h = 1.0 \cdot 10^{-2}$
5	$4.0 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	$4.0 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$
10	$2.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$
50	$4.0 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$4.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$
100	$2.0 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$2.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$

It might also be interesting to study the influence of the cell-loss probability on the results. In the calculations discussed above, the cell-loss probability was kept constant and equal to 10^{-9}. Figure 5 shows that if we increase the cell-loss probability to 10^{-3}, traffic dispersion still reduces the capacity like in Figure 4, but not to the same extent as with the lower cell-loss probability. This might be because a higher tolerance of loss allows more sources to be multiplexed on the same link. The increase in number of sources, which dispersion makes possible, hence becomes less significant.

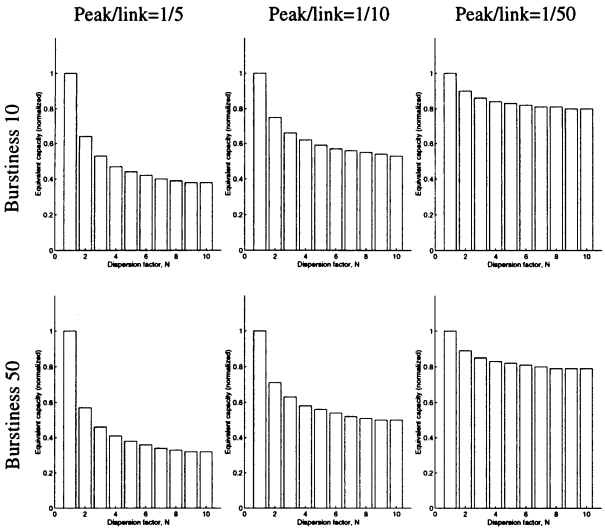


Figure 5 Equivalent capacity for some different values on source burstiness and peak-to-link ratio. The cell-loss probability was set to 10^{-3}.

2.2 Equivalent capacity with buffering

The discussion so far has been for a zero-buffer assumption, but it may also be of interest to look at the case where a single on-off source generates input traffic to a link with buffer capacity X. Guérin et al. give an upper bound on the equivalent capacity c of such a connection, Guérin et al. (1991):

$$c = h \cdot \frac{y - X + \sqrt{(y-X)^2 + 4X\varepsilon y}}{2y}, \text{ where } y = T_{on}(1-\varepsilon)h \cdot \ln\frac{1}{\varphi}. \tag{6}$$

T_{on} is the average duration of an active period, φ is the probability of buffer overflow (cell loss), and h and ε are defined as before. Since this case concerns only a single source, it might not be suitable to model dispersion as in the previous section, where each original source was replaced by a number of sources with lower peak rates. We therefore choose to model a dispersed source by an on-off source with a fixed mean but with a correlation function which changes with the dispersion factor.

Recall the on-off source from Figure 2. Define $u(i)$ to be the number of cells generated within the ith time unit. This means that $u(i)$ is either 0 or h. The correlation sequence of the source is given by

$$r(k) = E\{u(i+k)\,u(i)\} = h^2\varepsilon^2\left(1 + \frac{1-p_{on}}{1-p_{off}} \cdot (p_{on}+p_{off}-1)^k\right). \tag{7}$$

The more correlated the traffic is, the more difficult it becomes to handle. When dispersing the traffic, the objective is therefore to minimize the correlation in the cell stream on each path. As the correlation sequence of an on-off source is monotonously decreasing, the minimization is obtained by distributing the generated cells cyclically over the paths, as mentioned before (Figure 6).

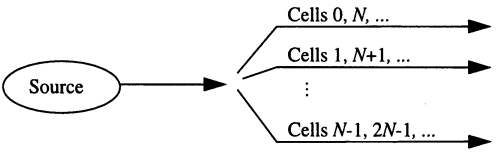


Figure 6 Dispersing the cells cyclically over N disjoint paths.

The correlation sequence between the cells on one of the paths is hence given by Gustafsson (1994:2)

$$r_d(k) = E\{u(iN+kN)\,u(iN)\} = r(kN)\,. \tag{8}$$

In order to study the behaviour of the equivalent capacity for different degrees of dispersion, we model the traffic from a dispersed source on a certain path. This is achieved by keeping the peak and mean rates of an on-off source constant, while reducing the correlation according to (8). The amount of traffic transmitted on a link during a certain time interval hereby remains unaltered, while the traffic behaviour varies under the influence of dispersion.

The fraction of time that the source spends in active state can be written as

$$\varepsilon = \frac{1-p_{off}}{2-p_{on}-p_{off}} = \frac{1}{1+\dfrac{1-p_{on}}{1-p_{off}}}, \tag{9}$$

and keeping ε constant thus means keeping $\dfrac{1-p_{on}}{1-p_{off}}$ constant.

The only part of $r(k)$ varying when the source peak and mean rates are fixed, is hence $(p_{on}+p_{off}-1)^k$.

Given the transition probabilities for a non-dispersed source (N=1), we can calculate the probabilities for dispersion factor N by

$$\frac{1-p_{off}^{(1)}}{2-p_{on}^{(1)}-p_{off}^{(1)}} = \frac{1-p_{off}^{(N)}}{2-p_{on}^{(N)}-p_{off}^{(N)}} \text{ and} \tag{10}$$

$$\left(p_{on}^{(1)}+p_{off}^{(1)}-1\right)^N = p_{on}^{(N)}+p_{off}^{(N)}-1. \tag{11}$$

By adjusting the source characteristics according to (10) and (11), we show the effects of dispersion on the equivalent capacity from (6). We have chosen a numerical example with an on-off source whose T_{on} is about 200 time units.

Previous results have shown that the queue size is highly dependent on the correlation in the traffic, Li and Mark (1990). In order to facilitate a comparison among the different graphs, we therefore keep the correlation fixed in all cases where there is no dispersion, that is, the first bar in each graph. This means that the value of T_{on} is given by

$$T_{on} = \frac{1}{1 - p_{on}}, \tag{12}$$

will vary, because changing the burstiness while keeping the correlation fixed for N=1, means that the value of p_{on} will change as well.

We try to make the comparison among different traffic cases as fair as possible. One solution would be to keep the peak rate of the source constant. This means that the mean rate must be varied when the burstiness varies. The results obtained on these conditions are shown in Figure 7, and the equivalent capacity for N=1 is normalized to be one. The peak rate is constant and set to 100, and φ is 10^{-9}. Since the peak-to-link ratio is not considered in (6), the graphs in Figure 7 are not directly comparable to those in Figure 4. Additional results, which are not presented here, shows that the equivalent capacity behaves similarly if the mean rate is kept constant and the peak rate is changed instead, Gustafsson (1995).

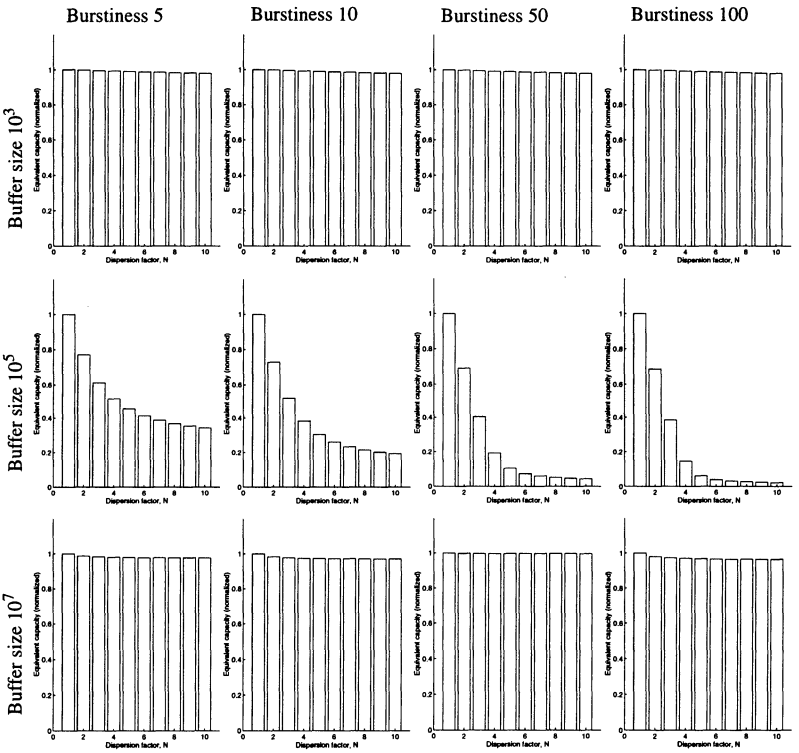


Figure 7 Equivalent capacity for different degrees of dispersion and burstiness, for three different buffer sizes. The source peak rate is constant and set to 100.

Table 3 shows the values from Figure 7 before normalization, for the cases without dispersion.

Table 3 Equivalent capacity before normalization; N=1, h=100

Buffer size	*Burstiness 5 Mean rate 20* T_{on}=248	*Burstiness 10 Mean rate 10* T_{on}=220	*Burstiness 50 Mean rate 2* T_{on}=202	*Burstiness 100 Mean rate 1* T_{on}=200
10^3	99.8	99.8	99.8	99.8
10^5	77.4	76.5	75.8	75.7
10^7	20.5	10.3	2.01	1.04

The results above show that when the buffer size is extremely small, the equivalent capacity for a connection approaches the source peak rate. Because of the buffer limitation, it becomes difficult not to exceed the allowed probability of overflow, even when dispersion is used. In the example above, the average burst size is about 20 000 cells, while the buffer size is only 1000. When a burst arrives, it hence fills up the buffer rather fast, and in order to keep the cell loss at a low level, the output rate of the buffer must be very high. By increasing the dispersion factor further, we can reduce the average burst size far below the buffer size, to a point where the traffic is almost completely uncorrelated, but it turns out that the decrease in equivalent capacity stays at about 45-50%. An explanation for this might be that the formula only considers a single source. There are hence no capacity gains due to the effects of multiplexing, as can be obtained when several sources share a link. This might also explain why the capacity reductions are not similar to the ones obtained without buffering, since in that case, multiple sources were multiplexed together on a link. In summary, with one single source and a small buffer, dispersion cannot significantly improve the situation, at least not for a dispersion factor smaller than ten.

If on the contrary the buffer size is extremely large, the equivalent capacity approaches the source mean rate. Since the capacity can never be lower than the mean rate, dispersion is of very little help in this case. It should be noted however, that such low capacity values can be obtained because the buffer is large enough to hold entire bursts. The penalty for this is long delays.

When the buffer size lies somewhere between these two extremes, traffic dispersion is useful, and the equivalent capacity under the influence of dispersion follows the same tendency with as without buffering. That is, as the source burstiness increases, the gain obtained by dispersion increases too.

Next, we consider the equivalent capacity of a number of connections, which are multiplexed on the same link. The capacity could be approximated by the sum of the individual capacities, that is

$$C = \sum_{i=1}^{n} c_i. \tag{13}$$

Unless the equivalent capacity of each individual connection is very close to the source mean rate, the capacity according to (13) in many cases overestimates what actually needs to be allocated. This is because the method does not consider the effects of statistical multiplexing.

Guérin et al. (1991) present the following approximation for the equivalent capacity of n multiplexed on-off sources:

$$C = \min\left\{ n \cdot \varepsilon h + \sigma\sqrt{-2\ln\varphi - \ln(2\pi)},\ \sum_{i=1}^{n} c_i \right\}. \tag{14}$$

The term $n \cdot \varepsilon h$ denotes the mean aggregate bit rate of the connections, and σ is the standard deviation of the aggregate bit rate, that is

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^{\,2} = n\sigma_i^{\,2} = n \cdot h^2 \varepsilon (1-\varepsilon). \tag{15}$$

As long as we keep the source peak and mean rates constant, the first part of (14) will not be affected by dispersion. In order to investigate the effects of dispersion on the equivalent capacity in (14), we therefore recall the model of a dispersed source which was used in Section 2.1. By replacing each original source with peak rate h by N sources, each with peak rate h/N, the first part of (14) becomes

$$N \cdot n\varepsilon \cdot \frac{h}{N} + \sigma\sqrt{-2\ln\varphi - \ln(2\pi)} = n \cdot \varepsilon h + \sigma\sqrt{-2\ln\varphi - \ln(2\pi)}\ , \text{ where} \tag{16}$$

$$\sigma^2 = N \cdot n \cdot \left(\frac{h}{N}\right)^2 \cdot \varepsilon(1-\varepsilon) = \frac{n}{N} \cdot h^2 \varepsilon (1-\varepsilon). \tag{17}$$

Figure 8 shows how dispersion affects the equivalent capacity of n multiplexed connections, according to (14). The results presented are for n=10, 100, 1000, and the values are normalized to be one for N=1. We have chosen the buffer size X=100 000, since this was the case where dispersion made significant difference to the results in the previous discussion. Further results, which are not shown here, indicate that we get a capacity reduction as well when reducing the buffer size to about 1000, even though the reduction in that case is slightly smaller than with a larger buffer.

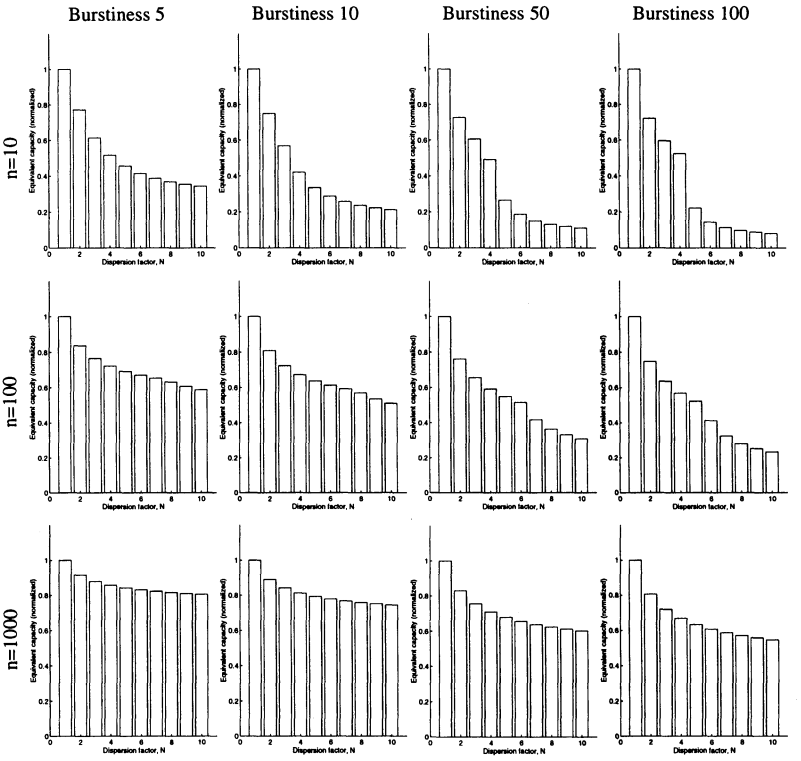


Figure 8 Equivalent capacity for different degrees of dispersion and burstiness. The peak rate of a source is constant and set to 100, the probability of buffer overflow φ is 10^{-9}, and the buffer size is X=100 000.

Table 4 shows the equivalent capacity from Figure 8 before normalization for N=1. In the table, we show the aggregate equivalent capacity for n sources, divided by the number of sources, n. The values in the table can therefore be seen as the equivalent capacity of one of the n sources.

Table 4 Equivalent capacity before normalization; N=1, h=100

Number of sources, n	*Burstiness 5*	*Burstiness 10*	*Burstiness 50*	*Burstiness 100*
10	77.5	69.7	29.9	20.8
100	45.2	28.9	10.8	7.26
1000	28.0	16.0	4.79	2.98

Comparing these results to those in Figure 7 and Table 3, we find that (14) gives significantly lower capacity values than (13) for a large number of sources with high burstiness. This is the situation where the effects of statistical multiplexing show.

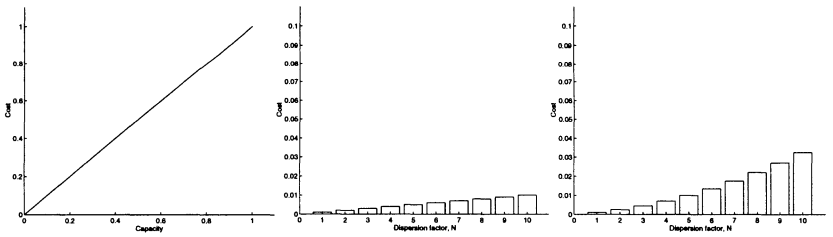
Furthermore, dispersion causes larger capacity reductions in the case with a high source burstiness, and a small number of sources which are multiplexed together. The similarity between this behaviour and the one that appeared in Figure 4 is striking. An increasing number of sources (larger n) means that the ratio between the source peak rate, which in this case is constant, and the aggregate equivalent capacity C decreases. If we let this ratio correspond to the peak-to-link ratio in Figure 4, we get exactly the same tendency with as without buffering. This means that we can make the general conclusion that dispersion improves the equivalent capacity particularly in the case of a small number of sources (high peak-to-link ratio) with high burstiness (high peak-to-mean ratio).

The results presented thus show that for a suitable buffer size, traffic dispersion reduces the equivalent capacity for a connection. For very large buffers dispersion does not affect the equivalent capacity, but will probably reduce the delay, and for very small buffers dispersion over a modest number of paths cannot improve the situation, unless there are enough sources to obtain multiplexing effects. When there is a capacity reduction, it behaves similarly with as without buffering. We have therefore chosen to limit the following discussions to the results without buffering, since they seem to represent a general behaviour.

3 COST FUNCTIONS

The previous section showed that a drastic decrease in equivalent capacity owing to traffic dispersion is possible for bursty sources. Considering only the equivalent capacity might however be somewhat optimistic, since spreading the traffic over several paths requires more virtual circuits to be established, and causes additional signalling overhead. We will therefore weigh the capacity obtained without buffering with three different cost functions, in order to establish under what circumstances traffic dispersion is profitable.

The first cost function is a fixed charge per capacity unit (Figure 9 (a)). This cost is independent of the number of connections used for a transmission, and the cost benefit curve will follow the curves in Figure 4, scaled by a constant cost factor.



(a) (b) (c)

Figure 9 The different costs considered: a fixed cost per capacity unit (a), a fixed cost per connection (b), and a cost increasing with the number of connections (c).

Next, we consider a cost function which is composed of a fixed charge per capacity unit, and a fixed charge per connection used for a transmission (Figure 9 (a) and (b)). The connection charge is motivated by the extra effort needed to set up and maintain several virtual circuits for each transmission.

The last cost function is composed of a fixed charge per capacity unit, and a progressively increasing charge per number of connections used (Figure 9 (a) and (c)). Assume that without dispersion, the virtual circuit follows the shortest path through the network. Since there might only be one path of that length, the additional connections needed for dispersion will have to follow longer paths. The cost increase could therefore be taken as a penalty for using longer and longer paths.

4 WHEN IS TRAFFIC DISPERSION USEFUL?

We relate the different cost functions to the values of equivalent capacity that we obtained in Section 2.1, to see whether dispersion is always motivated. If we only consider a fixed charge per capacity unit and assume that there is no extra cost for using several connections (the first cost function), traffic dispersion is practically always profitable, and the more paths used the better. For sources with a high burstiness and a high peak-to-link ratio, the benefits of dispersion are obvious; by spreading the traffic over only a handful of paths, a cost benefit of about eighty to ninety percent is obtained. Regarding the sources with a low peak-to-link ratio, the benefits are not that large. The gain here is only about thirty percent. However, when considering the values of equivalent capacity without normalization, in Table 1, we find that the cases where the benefits of dispersion are least significant, are those where the cost without dispersion is already very low. Any larger gain would therefore in real values be negligible in comparison to the other cases. In other words: when the gain is needed, it is high.

With this cost situation, traffic dispersion over many paths is consequently always the best solution. The assumption of no extra cost for extra paths may however not be quite realistic, wherefore we move on to the next cost function.

In this case, we have a fixed charge per capacity unit and a fixed charge per connection. The balance between these two charges is very important. If the charge per link is extremely small, the results are the same as above, which means that dispersion is always profitable. If, on the other hand, the charge per link is too large, it will dominate the total cost and result in a cost function which is linearly increasing with the number of paths. Traffic dispersion would hence never be justified.

More realistically, the charge per capacity unit will form the major part of the cost. In our calculations, we have chosen the cost 1 per capacity unit and 0.001 per connection, as a hopefully reasonable proportion. These values apply to a time unit equal to one call. Figure 10 shows the result from applying such a cost function on some of the equivalent capacity values from Figure 4, without normalization.

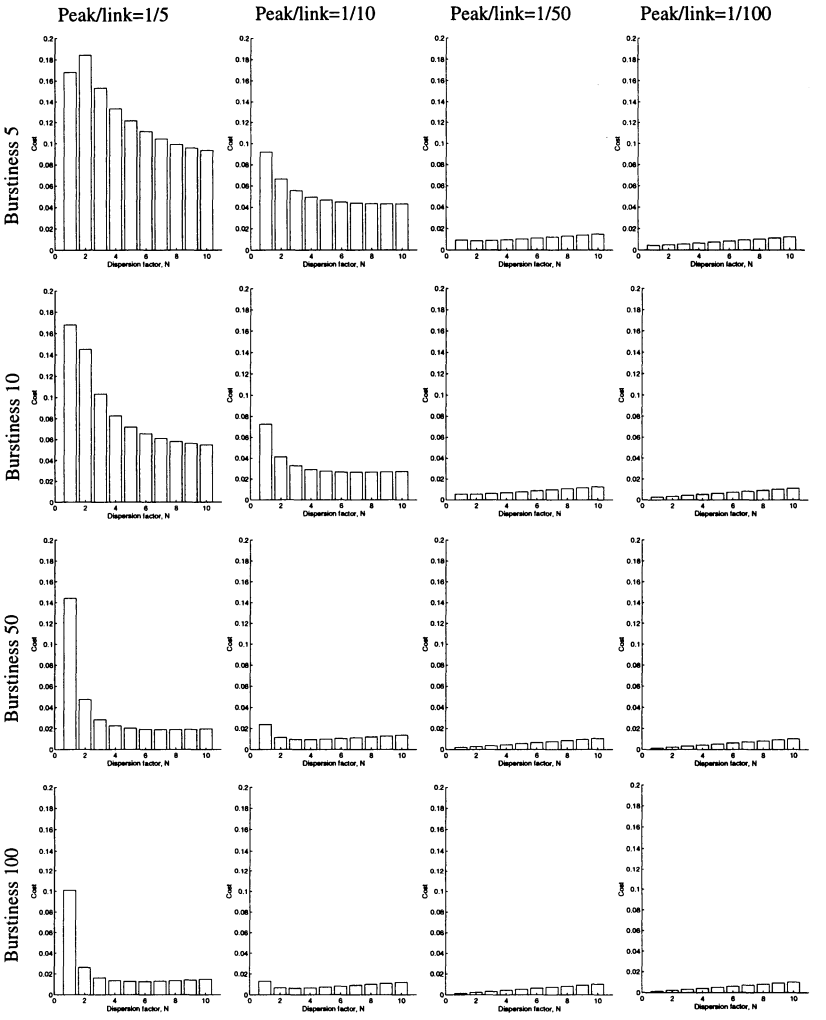


Figure 10 The second cost function related to the equivalent capacity values from Figure 4. The cost is 1 per capacity unit and 0.001 per connection.

These results show again that traffic dispersion is very profitable in the cases where the peak-to-link ratio is high. The maximum gain is about eighty percent for sources with high burstiness. As the peak-to-link ratio decreases, so does the gain, and using a larger number of connections even causes a small cost increase.

The conclusion is that in cases which can be handled well without dispersion, it should not be used. It is then better to allocate resources in a more conventional manner, using only one path for each transmission. In the cases where traffic dispersion does give benefits, it should of course be used. The results show that spreading the traffic over more than about two to five paths does not give any remarkable further benefit, whereas the number of paths should preferably be kept to about this size.

With the chosen proportions on the cost function, the increased cost caused by several connections is however practically negligible compared to the gain in cost obtained on other conditions. In hesitation of whether dispersion should be used or not, it therefore seems better to use it, since the penalty for dispersing when unnecessary is very small compared to the gain obtained when dispersion turns out to be needed. In essence, the benefits from using dispersion in the right place are many times larger than the penalty for using it in the wrong place.

The last cost function is a fixed charge per capacity unit and a charge which increases with the number of connections. The behaviour is as the one we described previously, namely the charge for using several paths soon dominates the total cost, and a transmission will quickly become rather expensive. This is shown in Figure 11. In this example, considerable benefits are still obtained for the higher peak-to-link ratios, but in the other cases there is no gain at all. The best strategy under these circumstances seems to be to disperse the traffic sparingly, and only when an economic gain is guaranteed.

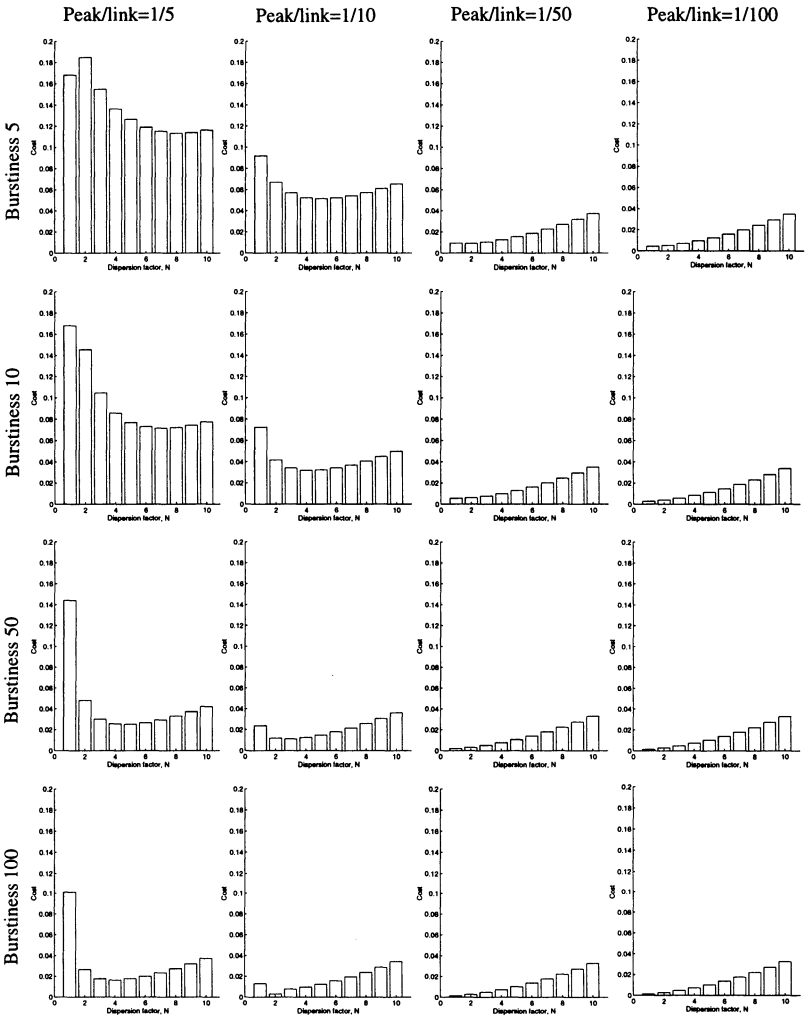


Figure 11 The third cost function related to the equivalent capacity values from Figure 4.

So, when is traffic dispersion useful? On the condition that the penalty for using several connections does not dominate the total cost for a transmission, dispersion over a moderate number of paths is practically always profitable. The most troublesome traffic sources to be handled by statistical multiplexing - that is, as mentioned before, those with a high burstiness and a high peak-to-link ratio - are those for which the highest gains can be made.

ATM Forum has defined three traffic parameters, namely sustained bit rate, peak bit rate and burst size. As a rough estimation, dispersion should be employed when the relation between the peak bit rate and the sustained bit rate (the source burstiness) is in the order of ten or more, and when the peak bit rate exceeds one tenth of the link capacity. In all cases, the number of paths used for a transmission should stay somewhere between two and five. The burst size is not directly considered in this paper. It is true that the average duration of an active period is obtained directly from the transition probabilities of an on-off source, and so is the probability that the source is in active state. We have however only considered the probability of being in active state, and it is possible to change that probability without affecting the duration of the active period. However, longer bursts (longer active periods) indicate higher correlation in the traffic, and this is where the benefits from dispersion are indisputable, Gustafsson (1994:2).

When dealing with sources having low peak bit rate compared to the sustained bit rate, that is, less than a ratio ten to one, and a link capacity above ten times the peak bit rate, we might just as well do without dispersion. The penalty for using dispersion in vain is however not dramatic, and dispersion may be applied in uncertain cases. Lastly, it should be noted that users may want to pay extra to get the traffic dispersed for reasons of security, or other reasons that are not contained in the results presented in this paper.

5 CONCLUSIONS

This paper presents spatial traffic dispersion as a means for handling difficult traffic sources, and facilitating resource allocation. The use of dispersion shows a large gain in the equivalent capacity, and when relating the capacity to three different cost functions, the benefits are in most cases confirmed.

From the results presented, we conclude that a profit due to dispersion is practically always possible. In the case of a single traffic source, there are no multiplexing effects. On the one hand, a small buffer may limit the gain in equivalent capacity to an extent where dispersion over a modest number of paths cannot improve the values. On the other hand, a large buffer makes the equivalent capacity close to the mean without dispersion, at the expense of long delays. In this case, dispersion could probably reduce the delay, but it is not reflected in the capacity results.

Furthermore, we conclude that the cost benefits from using dispersion are most important for sources with a high peak-to-mean ratio (larger than ten), and a high peak-to-link ratio (larger than one tenth). This is on condition that the charge per capacity unit dominates over the cost for using several connections. The penalty for using dispersion when not necessary turns out to be small compared to the benefits from using dispersion where it is really needed. Traffic dispersion is therefore useful in all cases where its benefits are beyond all doubt, as well as in all uncertain cases.

We may also change our viewpoint from the user to the network operator. If a tariff structure is imposed that erroneously penalizes traffic dispersion, statistical multiplexing may not be used to its full potential in the network. For a user it namely means a charge according to behaviour and not to average use, since the equivalent capacity of the transmission is strongly dependent on the burstiness of the source. The consequence is that the user may choose to keep the connection for a shorter time, and set it up for individual bursts. The operator thus has a situation with low sharing of resources (fewer paying users simultaneously connected) and with substantially more connection requests. As our example of equivalent capacity shows, traffic dispersion basically makes the statistical link sharing immune to source behaviour. We therefore hope that it will be widely employed as an antidote to new bursty traffic sources.

6 REFERENCES

Cheng, T-H. (1994) Bandwidth allocation in B-ISDN. *Computer Networks and ISDN Systems*, **Vol. 26, No. 9**, 1129-42.

Déjean, J.H., Dittmann, L. and Lorenzen, C.N. (1991) String Mode - A New Concept for Performance Improvement of ATM Networks. *IEEE Journal on Selected Areas in Communications*, **Vol. 9, No. 9**, 1452-60.

Guérin, R., Ahmadi, H. and Naghshineh, M. (1991) Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks. *IEEE Journal on Selected Areas in Communications*, **Vol. 9, No. 7**, 968-81.

Gustafsson, E. (1994) *Traffic Dispersion - A Literature Survey*. Technical Report TRITA-IT R 94:35, Royal Institute of Technology, Stockholm.

Gustafsson, E. (1994) *Traffic Dispersion in ATM Networks*. Technical Report TRITA-IT R 94:36, Royal Institute of Technology, Stockholm.

Gustafsson, E. (1995) *When Is Traffic Dispersion Useful? A Study On Equivalent Capacity*. Technical Report TRITA-IT R 95:17, Royal Institute of Technology, Stockholm.

Lee, T.T. and Liew, S.C. (1993) Parallel Communications for ATM Network Control and Management. *Proc. IEEE GLOBECOM*, **Vol. 1**, 442-6.

Li, S-Q. (1989) Study of Information Loss in Packet Voice Systems. *IEEE Trans. on Communications*, **Vol. 37, No. 11**, 1192-1202.

Li, S-Q. and Mark, J.W. (1990) Traffic Characterization for Integrated Services Networks. *IEEE Trans. on Communications*, **Vol. 38, No. 8**, 1231-43.

Maxemchuk, N.F. (1975) Dispersity Routing. *Proc. of ICC '75,* San Fransisco, CA, 41-10-13.

Maxemchuk, N.F. (1993) Dispersity Routing in High-Speed Networks. *Computer Networks and ISDN Systems*, **Vol. 25, No. 6**, 645-61.

7 BIOGRAPHIES

Eva Gustafsson received the M.Sc. degree in electrical engineering from the Royal Institute of Technology, KTH, in 1992. She is currently a Ph.D. student at the Department of Teleinformatics, working on traffic dispersion in high-speed networks.

Gunnar Karlsson received the MS degree from Chalmers University of Technology in 1983, and the Ph.D. from Columbia University in 1989. He was a Fulbright scholar at the University of Massachusetts at Amherst in 1982-83. His Ph.D. thesis was on sub-band coding of video for ATM networks. He joined the IBM Zurich Research Laboratory in 1989 and the Swedish Institute of Computer Science in 1992. He has been the first project leader of the Stockholm Gigabit Network. His research interests include traffic control, switching architectures and packet video. He is Docent at the Department of Teleinformatics at the Royal Institute of Technology (KTH).