18

The Strategy of Traffic Dispersion

E. Gustafsson
Royal Institute of Technology, Dept. of Teleinformatics
KTH-Electrum/204, S-164 40 KISTA, SWEDEN,
Phone +46 8 7521498, Fax +46 8 7511793, Email evag@it.kth.se
G. Karlsson
Swedish Institute of Computer Science
Box 1263 S-164 28 KISTA, SWEDEN
Phone +46 8 7521577, Fax +46 8 7517230, Email gk@sics.se

Abstract

Traffic dispersion means spreading the traffic from a source over multiple independent paths, transmitting it in parallel through the network. The strategy reduces the effects of bursts in the traffic, and may hence improve the network performance in terms of reduced queuing delay. So far, there have been several reports on traffic dispersion, but to our best knowledge, there has been no in-depth investigation of how the strategy affects different types of traffic under various conditions. This paper focuses on the basic properties of traffic dispersion by defining the strategy and investigating it from the source's as well as from the network's point of view. We investigate the influence of dispersion on Poisson traffic, traffic generated by two-state Markov chains and traffic generated by the chaotic FPDI-map. The results indicate a large potential of traffic dispersion to provide fast, secure and fault-tolerant transmission for highly bursty traffic sources.

Keywords

Traffic dispersion, ATM, traffic characteristics, queuing performance, network performance.

1 INTRODUCTION

The broadband integrated services digital network, B-ISDN, is intended to carry various types of traffic, such as voice, data and video. Satisfying the various user performance requirements will put stringent demands on the network to be flexible, reliable and still cost-efficient, that is, to obtain high resource utilization. Some of the difficulties imposed on the network are discerned when considering the traffic behaviour of potential B-ISDN users. While ordinary telephone traffic traditionally has been modelled by Poisson processes, recent studies indicate the characteristics that stem from data and multimedia applications to be vastly different. Essentially, studies have shown this type of traffic to exhibit strong correlation over long time periods, Leland et al. (1994), Paxson and Floyd (1995).

The asynchronous transfer mode, ATM, is the recommended network architecture for B-ISDN. Succinctly described, it combines the circuit-switched routing of telephone systems with the statistical multiplexing of packet switching, by establishing a virtual connection through the network before transmission. The information is then sent over the connection in fixed-size packets called cells. In order to better utilize the network resources, ATM employs statistical multiplexing, which means that the capacity of a transmission link in the network is statistically shared among the connections traversing it. This implies that the demand for capacity

A. Tantawy (ed.), *High Performance Networking VII*

occasionally may exceed the available resources. Surplus traffic could temporarily be buffered at the link access, but would result in buffer overflow and loss of data if continued inordinately.

The probability of information loss due to buffer overflow is highly dependent on the correlation in the multiplexed traffic streams Li (1989). Long-term correlation and extended traffic bursts complicate resource allocation because of the difficulties to guarantee performance. Given a connection, the correlation in a traffic stream may be lowered by spreading the traffic in time, so called shaping. The amount of shaping is however limited by the permissible delay that the information transfer may suffer, and the residual correlation in a shaped traffic stream may consequently still be strong. Multiplexing of such correlated streams at low probability of loss would require unreasonable low utilization of the network capacity, and excessively large buffers.

Another method to reduce the correlation in a traffic stream is to spread the traffic from a source over multiple disjoint paths, thereby sending it in parallel through the network, Figure 1 (a). We call this technique *traffic dispersion*, and we define the dispersion factor N to be the number of paths over which the traffic from a source is spread. Dispersion may be used on several network levels; over paths, links or multiple channels within a link. In this study, we refer to traffic dispersion as dispersion of packets over disjoint paths, that is, paths that do not share links statistically. Given a certain number of paths, there are different ways to spread the traffic. *Cyclic dispersion* is the spreading of packets from a source in round-robin order over the paths (Figure 1 (b)). If the correlation in the traffic stream is monotonously decreasing, this spreading strategy minimizes the correlation in the traffic stream on each path (see Section 2.2.2). *Sequential dispersion* means spreading sequences of L consecutive packets (Figure 1 (c)). A sequence length L=1 corresponds to cyclic dispersion. Sequential dispersion with L>1 does not necessarily reduce the correlation in the traffic stream, and if the sequence length is large, the performance may appear even worse with than without dispersion (see Section 3.1). In both cyclic and sequential dispersion, the order of the paths is known in advance, which facilitates the spreading and resequencing mechanisms. *Dynamic dispersion* spreads the traffic dynamically over the paths, according to the current network load. This makes it possible to avoid congested or heavily loaded paths, but requires continuously updated network information in return. Furthermore, this strategy does not consider the correlation in the traffic stream and may result in strongly correlated traffic on one or several of the paths, possibly congesting paths that were originally clear. Contrary to the two earlier methods, dynamic dispersion is reactive rather than preventive, and if all users were to spread their traffic dynamically, the result could be an unstable operation. The analyses in this study focus on the preventive dispersion strategies.

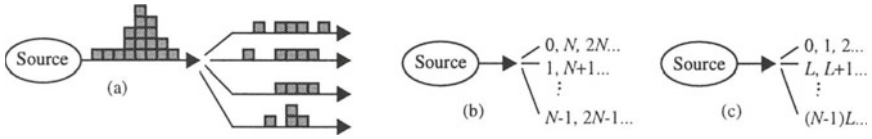


Figure 1 Illustration of traffic dispersion (a), and cyclic (b) and sequential (c) dispersion of the packets generated by a traffic source.

During the last years, traffic dispersion has been discussed in various terms and for different purposes, Gustafsson and Karlsson (1997). Throughout the studies, traffic dispersion has shown to give performance improvement in terms of better queuing behaviour and reduced loss. The use of independent paths also makes loss of

information on one path independent of losses on other paths, and consequently, forward error correction can be successfully used to increase network reliability. An additional advantage of traffic dispersion is that it enhances the network security by making eavesdropping on parallel connections simultaneously practically impossible. There exist however, to our best knowledge, no thorough investigation of how dispersion affects the behaviour of the traffic from different types of traffic sources, nor a comparison of the queuing performance under different load conditions. In order to motivate more advanced studies on traffic dispersion, its basic properties need to be defined, investigated and reported.

The aim of the work reported in this paper is to do such a study. We start in Section 2 by describing the system model used in the studies, including the different types of traffic sources used. In Section 3 we continue by investing the queuing behaviour for each traffic source. This Section also includes a performance comparison of cyclic and sequential dispersion, and an investigation of how large part of the traffic that must be dispersed in order to make the effects of dispersion show. Section 4 discusses the optimum dispersion factor for a call, and Section 5 concludes the paper.

2 SYSTEM MODEL DESCRIPTION

2.1 One-stage multiplexer model

In order to investigate the effects of traffic dispersion on some different traffic types, we use a one-stage multiplexer model according to Figure 2 (a) and (b). The case without dispersion is illustrated in Figure 2 (a), by a traffic source which generates traffic into a FIFO queue with fixed service rate. Next, the model is extended to spread the traffic from a source over N queues. This strategy will however reduce the amount of traffic entering the queue during a certain time interval. In order to keep the mean arrival rate to each queue constant, independent of N, we assume N independent identically distributed sources, each spreading its traffic over N queues (Figure 2 (b)). The amount of traffic arriving at each queue is thus kept constant, while the traffic behaviour changes due to the effects of dispersion. We define the utilization factor ρ in a queue to be the mean arrival rate multiplied by the queue service time. The buffer size is in the remainder of this paper considered infinite.

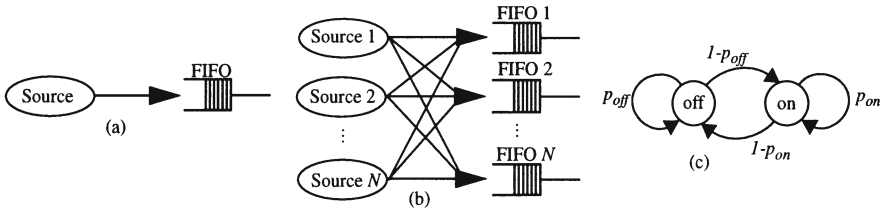


Figure 2 A one-stage multiplexer, consisting of traffic sources generating traffic into a number of FIFO queues. The case of no dispersion is shown in (a), while (b) illustrates dispersion. The state diagram of a two-state Markov chain is shown in (c).

2.2 Traffic source models

2.2.1 Poisson traffic

Poisson traffic is traffic which arrives according to a Poisson process, that is, the interarrival times between packets are independent and exponentially distributed. The Poisson process has been commonly used to model the arrival of telephone calls

during some time interval in a telephone network. A Poisson process with intensity λ is a process $X = \{X(t); t \geq 0\}$, where $X(t)$ denotes the number of arrivals of the process by time t, and

$$Pr\{X(t)=i\} = \frac{(\lambda t)^i}{i!} e^{-\lambda t}, \qquad i = 0, 1, 2... \tag{1}$$

The mean arrival rate is λ packets/time unit, and the variance of the arrival rate is λ. Since the arrivals of a Poisson process are mutually independent, the correlation in the traffic stream is zero.

2.2.2 Two-state Markov chain source

The most commonly used on-off source is based on a two-state Markov chain (Figure 2 (c)). The chain is alternating between an on state and an off state, and while in the on state, it generates traffic at peak rate h packets/time unit. The model may be seen as a crude model for data communications, and has been used with good accuracy to model speech after removal of silence periods.

In the following calculations, the peak rate of the source is set to $h = 1$ packet/time unit. Define $\varepsilon(i)$ as the number of packets generated within the ith time unit; $\varepsilon(i)$ is thus 0 or 1. The mean and variance of the chain are then given by

$$\bar{\varepsilon} = \frac{1 - p_{off}}{2 - p_{on} - p_{off}} \text{ and } \sigma^2 = \bar{\varepsilon}(1 - \bar{\varepsilon}). \tag{2}$$

As in Li and Mark (1990), $\varepsilon(i)$ is normalized: $u(i) = \dfrac{\varepsilon(i) - \bar{\varepsilon}}{\sigma}$. (3)

The correlation between $u(i)$ and $u(i+n)$ is then given by

$$r(n) = E\{u(i+n)u(i)\} = (p_{on} + p_{off} - 1)^{|n|} = \phi^{|n|}. \tag{4}$$

The correlation is monotonously decreasing, and cyclic dispersion, which maximizes the distance between two consecutive packets on each path, hence minimizes the correlation. With cyclic dispersion, the correlation sequence becomes

$$r_d(n) = E\{u(iN + nN)u(iN)\} = r(nN) = \phi^{|nN|}. \tag{5}$$

The sojourn time the two-state Markov chain stays in a state is geometrically distributed with mean

$$\bar{T}_{on} = \frac{1}{1 - p_{on}} \text{ and } \bar{T}_{off} = \frac{1}{1 - p_{off}} \text{ respectively.} \tag{6}$$

Some of the numerical examples in this paper corresponds to transmitting voice traffic over ATM. The transition probabilities of the chain are then assigned values according to the mean length of talkspurts and silent periods, for example 1366 ms and 1802 ms, Brady (1968), and 227 ms and 596 ms, Lee and Un (1986) respectively. Transmitting voice over 64 kbit/s links makes an ATM cell (48 bytes of information) correspond to 6 ms, and the transition probabilities of a corresponding two-state

Markov chain are for the first example p_{on} = 1-6/1366 and p_{off}= 1-6/1802, and for the second p_{on} = 1-6/227 and p_{off}= 1-6/596.

2.2.3 The Fixed Point Double Intermittency map

Recent studies have shown tele and data communications traffic to exhibit long-range dependence, Leland et al. (1994), Paxson and Floyd (1995). Modelling traffic with such characteristics requires a new generation of traffic models, like for instance the fixed point double intermittency map, Pruthi (1995). The FPDI-map is a chaotic map in the form of an on-off source, where the distribution of the time spent in each state can be varied from geometric to heavy-tailed. The nature of the distribution depends on the Hurst parameter of the source, and the FPDI-map can be used to generate long-range dependent traffic.

3 QUEUING BEHAVIOUR

In this Section, we present the effects of dispersion in calculated and simulated queuing results. The question of cyclic or sequential dispersion is investigated in Section 3.1, and then the queuing behaviour is studied for the different source models. We start by discussing independent arrivals from a Poisson process in Section 3.2, continue with short-term correlated traffic generated by two-state Markov chains in Section 3.3 and finish in Section 3.4 by considering traffic generated by the chaotic FPDI-map. Section 3.5 deals with the question of how many users must employ dispersion in order to make the positive effects of it show.

3.1 Cyclic or sequential dispersion?

In the following, we use the one-stage multiplexer model from Figure 2 (b) to investigate and compare the effects of cyclic and sequential dispersion. The simulation results in Figure 3 show how the mean queue size changes with increasing sequence length. The results presented in Figure 3 (a), for Poisson traffic sources, show that the mean queue size increases with the sequence length, and the larger dispersion factor, the faster increase. There is always a possibility that several sources will send traffic to the same queue simultaneously, thereby increasing the instantaneous peak arrival rate to the queue. As the sequence length increases, each source generates traffic continuously to each queue during a longer time interval. Several sources sending traffic into the same queue may hence cause traffic bursts of high peak rate. As a consequence of increased instantaneous peak arrival rate and increased correlation in the traffic, the mean queue size increases, even if the queue utilization factor is kept constant (see Section 3.3.2).

Figure 3 (b) shows the results from strongly correlated traffic generated by two-state Markov chains. In this case, the mean queue size increases practically linearly with the sequence length, and the larger the N, the faster the increase. For a large enough sequence length, dispersion aggravates the queuing behaviour, compared to the non-dispersed case. This may be explained by the same discussion as above.

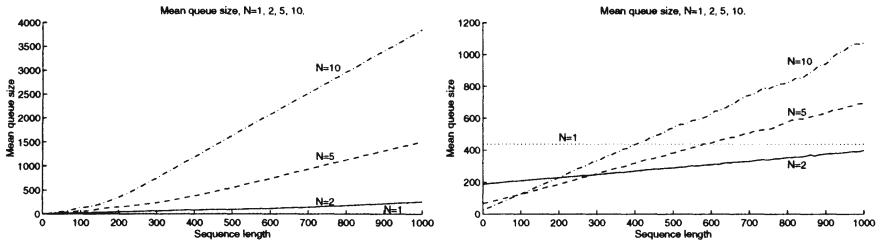


Figure 3 The mean queue size changing with the sequence length for Poisson traffic (a) and traffic generated by Markov chains with p_{on}=1-6/1366, p_{off}=1-6/1802 (b). The results are for N=2, 5 and 10, and the mean queue size without dispersion is shown in (b), while in (a), it is too small to show. The utilization factor is 0.86.

In essence, sequential dispersion may introduce bursts in the traffic, thereby aggravating the queuing behaviour. For a large enough sequence length, sequential dispersion performs worse than non-dispersed traffic, with respect to the mean queue size. Additional results, which are not presented here, show the standard deviation of the queue size to behave similarly. In the remainder of this paper, we therefore assume cyclic dispersion.

3.2 Queuing with Poisson traffic

3.2.1 Calculated mean queue size

To an *M/D/*1 queuing system, traffic arrives according to a Poisson process, and is served in the single server with a constant service rate. If we let λ be the mean arrival rate in packets/time unit, d the service time in time units and $\rho = \lambda d < 1$, the mean queue size, at moments of departure in equilibrium, is, Kleinrock (1975)

$$\bar{q} = \frac{\rho^2}{2(1-\rho)}. \tag{7}$$

Denote the time between arrival *i-1* and *i* to the queue by X_i. The probability density function of X_i is then

$$f_{X_i}(x) = \lambda e^{-\lambda x}, \qquad \forall i > 0 \tag{8}$$

and the mean interarrival time between packets is $1/\lambda$ time units. If the traffic from a source is cyclically dispersed with dispersion factor N, the arrivals to one of the queues corresponds to every Nth arrival from the source. With N=2, the time between two adjacent arrivals to the queue is $Y = X_i + X_{i+1}$, the mean arrival rate is $\lambda/2$ packets/time unit, and the probability density function of Y is gamma distributed according to

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_i}(x) \cdot f_{X_{i+1}}(y-x)dx = \int_0^y \lambda e^{-\lambda x} \cdot \lambda e^{-\lambda(y-x)} dx = \lambda^2 y e^{-\lambda y}. \tag{9}$$

As the arrivals to the queue are still independent, the system becomes a *G/G/*1 queuing system. For this system, the mean waiting time in the queue, *W*, can be calculated using the spectral solution to Lindley's integral equation, described in Kleinrock (1975). In order to make a fair comparison of the queuing behaviour for different degrees of dispersion, ρ is kept constant. The distribution of the interarrival times to one queue is according to (9), and the distribution of service time is given by the Dirac delta function according to $\delta(x-2d)$. Following the calculations in Kleinrock (1975), the mean waiting time and mean queue size are obtained as

$$W = \frac{\rho^2}{2\lambda(1-\rho)} \text{ and } \bar{q} = \frac{\lambda}{2} \cdot W = \frac{\rho^2}{4(1-\rho)}. \tag{10}$$

Repeating the calculations for different *N* gives the mean queue size for different degrees of dispersion. Figure 4 shows the calculated values and the corresponding simulated values.

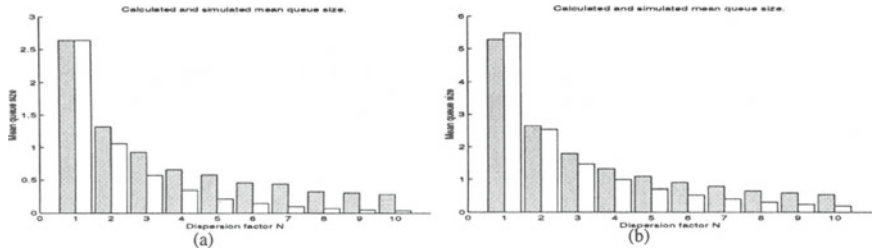


Figure 4 Calculated (shaded bars) and simulated mean queue size on one path, for different degrees of dispersion. The traffic is generated by a dispersed Poisson process, and the utilization factor is 0.86 (a) and 0.92 (b).

3.2.2 Simulated queuing behaviour

The results discussed above were for one source, spreading its traffic cyclically over *N* queues. In the following, the model with *N* sources spreading the traffic over *N* queues is employed (Figure 2 (b)). The simulated mean queue size and the standard deviation of the queue size in such a system are presented in Figure 5. These results show that the minimum mean queue size and standard deviation are obtained with a dispersion factor of approximately five. For a dispersion factor larger than five, even though the queue size of each dispersed source still decreases, as shown in Figure 4, superposing *N* dispersed sources causes the queue size to increase. Additional results, which are not presented here, show that for a lower utilization (smaller ρ), the optimum *N* is smaller than five, while a higher utilization makes the optimum *N* larger.

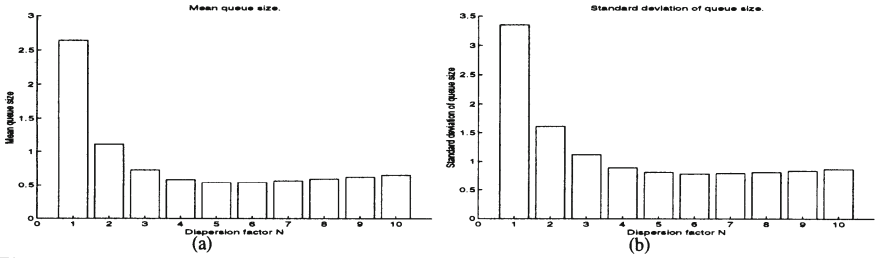


Figure 5 Mean queue size and standard deviation of the queue size for *N* Poisson sources generating traffic into *N* queues. The utilization factor is 0.86.

Figure 6 shows an example of how the queue size varies in time when using dispersion. The decrease in queue size agrees with the results from both Figure 4 and Figure 5. Thus, for Poisson arrivals, dispersion provides a means to reduce the mean queue size as well as the variance of the queue size. The results indicate the optimum dispersion factor to be in the interval [2, 5].

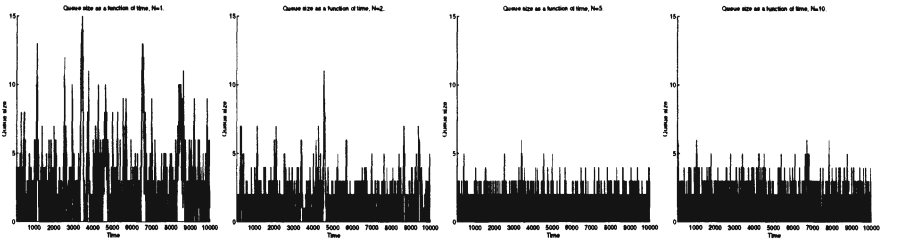


Figure 6 Example of queue size as a function of time for the on-stage multiplexer model with Poisson traffic, for N=1, 2, 5 and 10. The utilization factor is 0.86.

3.3 Queuing with traffic from two-state Markov chains

3.3.1 Calculated mean queue size

The power spectral density of a cyclically dispersed two-state Markov chain source is obtained as the discrete Fourier transform of the correlation sequence (5):

$$S(q) = \sum_{n=-\infty}^{\infty} r_d(n) e^{-j2\pi qn} = \frac{1-\phi^{2N}}{1+\phi^{2N}-2\phi^{N}\cos(2\pi q)}. \tag{11}$$

Since $|\phi| < 1$, $S(q) \to 1$ as $N \to \infty$. That is, when increasing the number of paths, the power spectral density will converge consistently with the definition of white noise. It is thereby clear that cyclic dispersion effectively reduces the correlation in the traffic generated by this type of source. Figure 7 shows the power spectral density of the packets on one path for different degrees of dispersion.

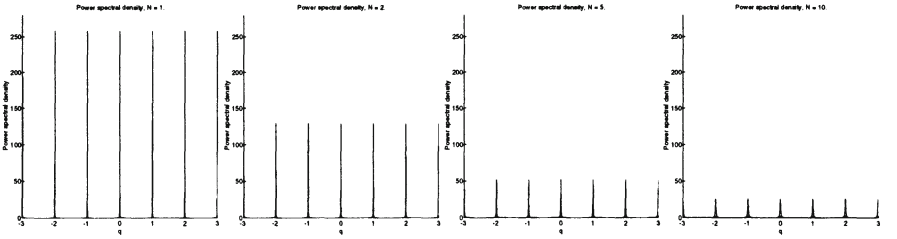


Figure 7 Power spectral density for two-state Markov chain source and dispersion factor N = 1, 2, 5 and 10. The source is characterized by p_{on}=1-6/1366, p_{off}=1-6/1802.

For the one-stage multiplexer in Figure 2 (a), the mean queue size can be calculated. Assuming that the source consists of two independent identically distributed two-state Markov chains, and that the queue service rate is one, the mean queue size is, Li and Mark (1990)

$$\bar{q} = \frac{\bar{\varepsilon}^2}{1-2\bar{\varepsilon}} \cdot \left(\frac{2}{1-\phi} - 1\right), \text{ where } \bar{\varepsilon} \text{ and } \phi \text{ are defined in (2) and (4).} \tag{12}$$

In case of cyclic dispersion, the dispersion may be illustrated by reducing the correlation from $\phi^{|n|}$ to $\phi^{N|n|}$, $N \geq 1$, according to (5), while keeping the peak and mean source rates unchanged. The mean queue size can then be expressed as

$$\bar{q} = \frac{\bar{\varepsilon}^2}{1-2\bar{\varepsilon}} \cdot \left(\frac{2}{1-\phi^N} - 1 \right). \tag{13}$$

ϕ^N in (13) refers to the correlation sequence between adjacent packets from the same source, arriving at a queue. In the multiplexer model, however, the correlation behaves slightly different, since packets originate from different independent sources. The calculated and simulated values might therefore differ. The calculated mean queue size as a function of the dispersion factor is shown together with simulated results in Figure 8, and the results show accordance in behaviour between the calculated and simulated values.

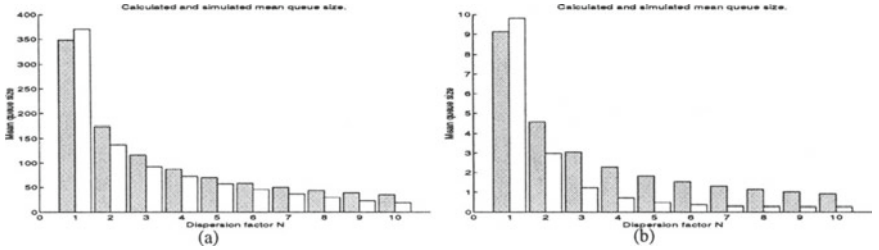


Figure 8 Calculated and simulated mean queue sizes, where each source consists of two Markov chains with p_{on}=1-6/1366, p_{off}=1-6/1802 (a) and p_{on}=1-6/227, p_{off}=1-6/596 (b). The shaded bars show the calculated mean queue size according to (13), and the utilization factor is 0.86.

3.3.2 Simulated queuing behaviour

Figure 9 shows the simulated mean queue size and standard deviation of the queue size, for N two-state Markov chain sources with peak rate 1, using the system model in Figure 2 (b). Both the mean and the deviation of the queue size decrease as N increases, but there is no obvious optimum N, as was the case for Poisson traffic. Nevertheless, one can see from the graphs that for N=1..5 there is a significant change in queue size, while both the mean and the deviation of the queue size start to level out for N>5.

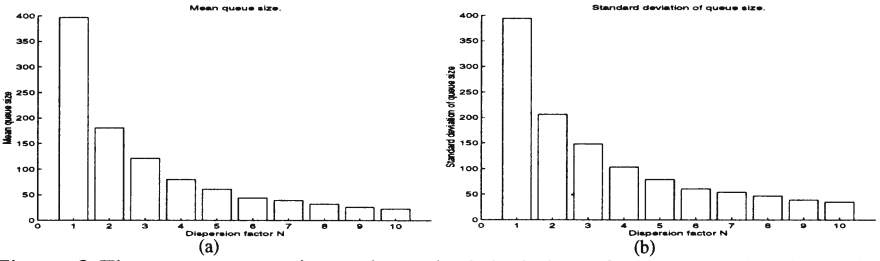


Figure 9 The mean queue size and standard deviation of the queue size for traffic generated by a two-state Markov chain source with peak rate 1, p_{on}=1-6/1366 and p_{off}=1-6/1802. The utilization factor is 0.86.

The sources used to obtain the results above produce strongly correlated traffic streams. In Figure 10, we investigate how the queuing behaviour is affected by the correlation. As the correlation decreases, an optimum dispersion factor shows up. Increasing the dispersion factor above that optimum value results in an increasing mean queue size. For highly correlated traffic, the mean queue size levels out at about N=5, and one may assume that for some larger N than shown in the graphs, the queue size starts to increase again. Additional results, which are not presented here, show that it is actually the case. Further results also show the standard deviation of the queue size to behave similarly to the mean queue size.

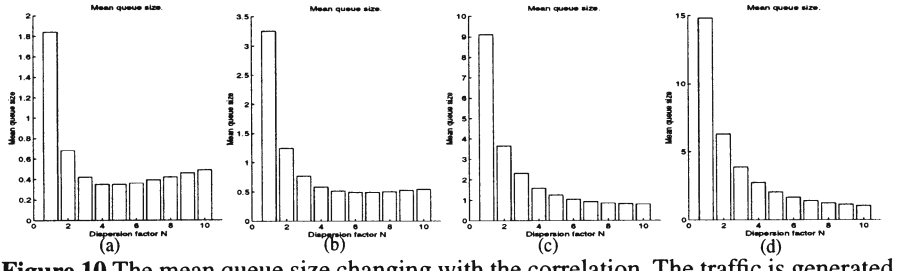


Figure 10 The mean queue size changing with the correlation. The traffic is generated by two-state Markov chains with $|\phi|$ =0.2 (a), $|\phi|$ =0.4 (b), $|\phi|$ =0.7 (c) and $|\phi|$ =0.8 (d). The source peak and mean rates are kept constant, and the queue utilization factor is 0.86.

Considering traffic with lower correlation, we also investigate how the optimum dispersion factor depends on the utilization factor in the queue. Figure 11 shows that the optimum dispersion factor increases as the utilization factor increases, so when increasing the utilization factor for low correlation traffic, the result approaches the ones obtained with high correlation.

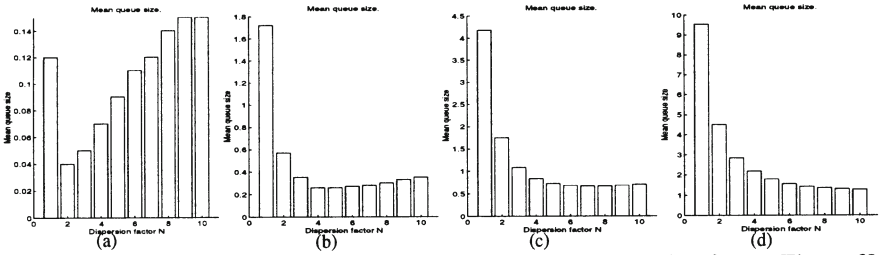
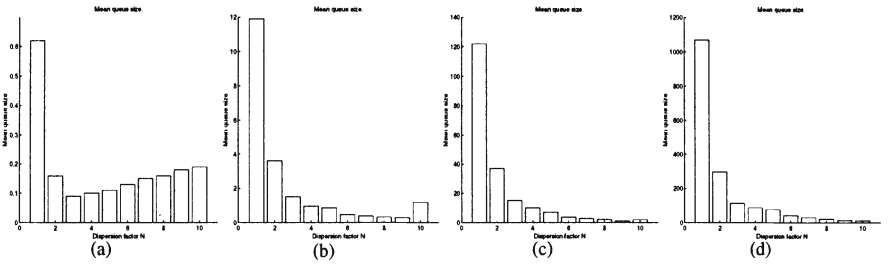


Figure 11 The mean queue size changing with the queue utilization factor. The traffic is generated by two-state Markov chains with p_{on}=0.7, p_{off}=0.8 and peak rate 1. The utilization factor in the different graphs is 0.48 (a), 0.72 (b), 0.84 (c) and 0.92 (d).

Next, we investigate how the queuing behaviour changes with dispersion for different source peak rates (Figure 12). Considering the case with no dispersion, that is, the first bar in each graph, we note that the mean queue size increases approximately proportionally to the source peak rate, when the utilization factor is constant. When looking at the graphs, we also note that dispersion improves the queuing behaviour more as the source peak rate increases. Additional results, which are not shown here, indicate a similar behaviour for the standard deviation of the queue size. The optimum dispersion factor still can be found in the interval [2, 5].



(a) (b) (c) (d)

Figure 12 How the mean queue size changes with the source peak rate. The traffic is generated by two-state Markov chains with p_{on}=0.7 and p_{off}=0.8, and the queue utilization factor is 0.6. The peak rate is 1 (a), 10 (b), 100 (c) and 1000 (d).

Figure 13 gives an example of how the queue size varies in time for the system. Obviously, dispersion reduces the large fluctuations in queue size, hence making it easier to predict the amount of capacity needed for a source. Earlier results accordingly show a decrease in the equivalent capacity for dispersed traffic, Gustafsson and Karlsson (1996). For traffic generated by a two-state Markov chain, dispersion thus reduces the correlation in the traffic stream and thereby also the mean queue size and standard deviation of queue size. The more correlated traffic, and the higher queue utilization factor, the higher the gain obtained by using dispersion. For low correlation, low peak rate and low utilization factor, the dispersion factor should be rather small in order not to aggravate the queuing behaviour. In general, N should be in the interval [2, 5]. With N>5, the surplus queuing benefits may not always justify the overhead caused by the required additional paths.

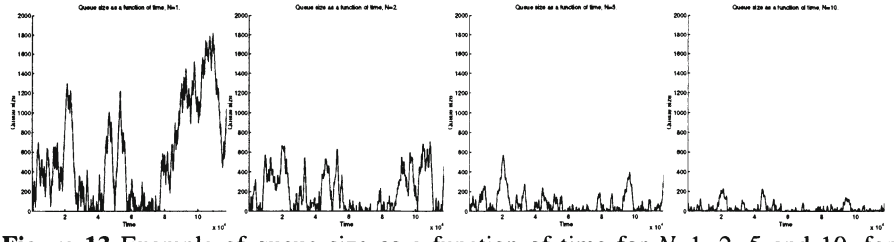


Figure 13 Example of queue size as a function of time for N=1, 2, 5 and 10, for Markov chain sources with peak rate 1, p_{on}=1-6/1366, p_{off}=1-6/1802. The utilization factor is 0.86.

3.4 Queuing with traffic generated by the FPDI-map

3.4.1 Simulated queuing behaviour

For the values on the Hurst parameter considered here, the FPDI-map traffic source generates a queue length distribution with an unbounded mean queue size, Pruthi (1995). The unbounded queue size is due to the fact that the traffic is highly variable, wherefore a single burst occasionally may be large enough to dominate the queuing performance. We can thus not investigate how the mean queue size changes for different degrees of dispersion, as we did in Section 3.2 and Section 3.3. Instead, we look at the queue distribution (Figure 14). The results indicate that dispersion reduces the queue size for this type of traffic as well as for the traffic types discussed earlier. Figure 15 shows an examlpe of how the queue size varies in time when dispersion is

used. These results indicate a similar change in queuing behaviour, due to traffic dispersion, as revealed earlier in this paper.

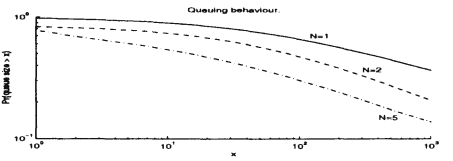


Figure 14 Queue-length distribution for traffic generated by the FPDI-map. The traffic has peak rate 1 and Hurst parameter 0.8333, and the utilization factor for the queue is 0.86. The graph shows results for N=1, 2 and 5, and the axes show x and p(x)=Pr{queue size > x}.

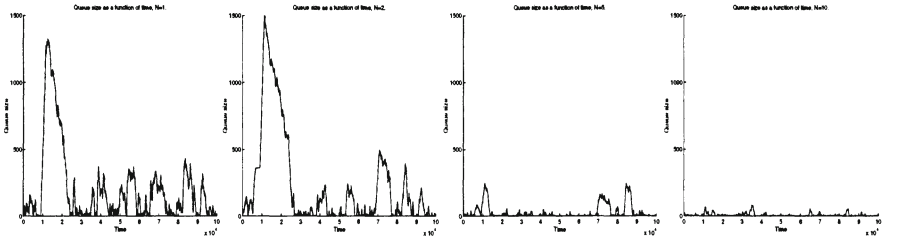
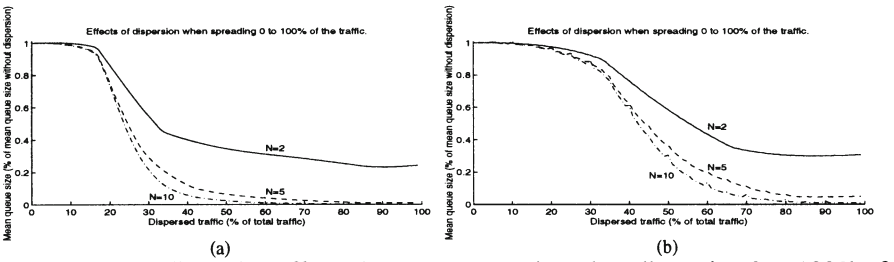


Figure 15 Example of queue size as a function of time for traffic generated by the FPDI-map, with peak rate 1 and Hurst parameter 0.8333. The utilization factor is 0.86, and the results are for N=1, 2, 5 and 10.

3.5 Traffic dispersion - a general or individual decision?

As the previous results show, traffic dispersion may yield large performance improvements. An important remaining question is whether there are any benefits from using dispersion on one source alone, or if every user must employ dispersion to make the benefits show. We try to answer this question by again studying the model in Figure 2 (b), while adding non-dispersed background traffic to the traffic already generated into a queue. All sources are two-state Markov chains, and the sum of the peak rates of the dispersed source and the background traffic source is kept constant, while the dispersed traffic is varied from constituting 0 to 100% of the total traffic.

Figure 16 shows how the mean queue size changes with the amount of dispersed traffic. We note that in Figure 16 (a), the effects of dispersion show when about 20% or more of the traffic is dispersed, while the corresponding limit in Figure 16 (b) is between 20 and 30%. The sources in Figure 16 (b) are less correlated than the sources in Figure 16 (a), and we thus conclude that the stronger correlation in the traffic stream, the earlier the effects of dispersion start to show. In general, dispersion starts showing when employed on about 20% of the traffic, while the large benefits may come later. The results also further emphasize the statement that a dispersion factor larger than 5 does not significantly improve the performance.



(a) (b)

Figure 16 How dispersion affects the mean queue size when dispersing 0 to 100% of the traffic. The total peak rate of the arriving traffic is 100 packets/time unit. In (a), the sources are characterized by p_{on}=0.7, p_{off}=0.8 and the queue utilization factor is 0.6. The corresponding values in (b) are 1-6/1366, 1-6/1802 and 0.52 respectively.

4 OPTIMUM DISPERSION FACTOR

The results in the previous Sections show the optimum dispersion factor to be in the interval N=[2, 5]. The higher dispersion factor, the higher gain, but when N exceeds 5, the additional gain per new path starts to level out. Earlier results have shown the dispersion factor to depend on the ratio of the source peak rate to the link capacity, Gustafsson and Karlsson (1996). The higher ratio, the larger gain, and in general, dispersion is useful as the peak-to-link ratio exceeds about 1/10 or the peak-to-mean ratio exceeds about 10. Evidently, the dispersion factor for a call depends on the number of paths available in the network. As mentioned earlier, the paths should preferably be disjoint and of equal length. An effect of one path being longer or congested would be large resequencing delay at the receiver. The effects of a congested path could be avoided by forward error correction. If for example a single parity check is used, one of the paths would carry redundant information, hereby increasing the tolerance to link failures and information loss at the expense of higher capacity requirements. The effects of redundancy on the capacity can be illustrated by the effective bandwidth of a connection.

We consider the continuous-time version of the two-state Markov chain source from Section 2.2.2. With transition rates α and β from off to on and on to off state respectively, the effective bandwidth is given by, Kelly (1996)

$$a(s,t) = \frac{1}{st} \cdot \log\left\{ \begin{bmatrix} \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{bmatrix} \exp\left(\begin{bmatrix} -\beta + hs & \beta \\ \alpha & -\alpha \end{bmatrix} t \right) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}, \tag{14}$$

where s is the space-scale, relating to the expected quality, t is the time-scale and h is the source peak rate. With N paths, of which k carry dispersed information from the source, and N-k carry redundant information, the effective bandwidth can be expressed as

$$a(s,t,N,k) = \frac{N}{st} \cdot \log\left\{ \begin{bmatrix} \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{bmatrix} \exp\left(\begin{bmatrix} -\beta + sh/k & \beta \\ \alpha & -\alpha \end{bmatrix} t \right) \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}. \tag{15}$$

In order to better show the effects of dispersion on the effective bandwidth, we temporarily withdraw the restriction on N to be an integer. Some results are presented

in Figure 17. The graphs show that if k=N-1, a dispersion factor N>3 still requires less capacity than a non-dispersed connection. For k=N-2, this holds for N>5. Provided that there are five disjoint paths of equal length available in the network, a dispersion factor N=5 should be used, with or without redundancy, depending on the quality of service requirements from the user. If the number of available paths is less, a dispersion factor N=2 and N=3 gives a capacity gain without redundancy, while N=4 can tolerate redundant information to be carried on one of the paths. Lastly, it should be noted that the choice of dispersion factor may be different if there are special requirements, such as high security or extremely low capacity utilization on each path.

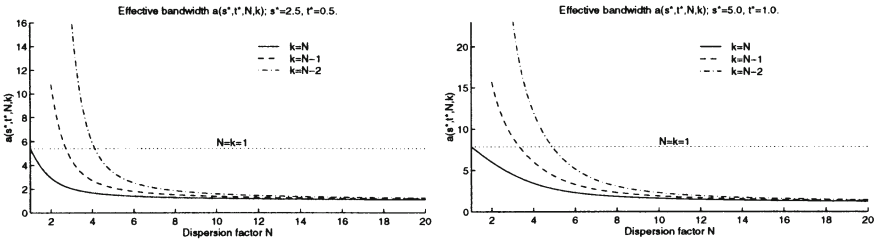


Figure 17 Effective bandwidth for a two-state Markov chain source, as a function of N and k. In the example presented, α =1, β =9 and h=10. The s and t parameters are kept constant in each graph.

5 CONCLUSIONS

In this paper, we have investigated the basic properties of traffic dispersion. We have discussed how dispersion changes the traffic behaviour for different types of traffic sources, and how it affects the queuing performance. We confirmed that cyclic dispersion performs better than sequential dispersion, with respect to the queuing behaviour. When using cyclic dispersion, it is possible to reduce the correlation in the traffic stream and to reduce the peak rate of a source. The influence of a source on each of the paths used is hence reduced, and this makes it possible to fully utilize the benefits of statistical multiplexing in for example an ATM network. Reducing the correlation also gives a smoother queuing behaviour. Traffic dispersion has been shown to decrease the mean and variance of the queue size, and in a more general perspective to reduce the fluctuations in the queue size over time. The benefits from using dispersion are a reduction on the order of five to ten times in the mean and the standard deviation of the queue size, compared to non-dispersed transmissions. In order to make the effects of dispersion show though, the strategy must be employed on at least 20 to 30% of the total amount of multiplexed traffic. In summary, our results show the large potential of traffic dispersion, with an optimum dispersion factor N=5.

Considering the benefits from using traffic dispersion, the technique deserves further research and attention, with focus on studies of larger networks, delays through the networks and resequencing delays at the receiver. What makes the traffic dispersion approach special is that the network does not have to adapt to complicated traffic characteristics. On the contrary, the traffic characteristics are engineered to suit a reasonable network structure. This, we believe, gives us the opportunity of fast, secure and fault-tolerant transfers of data to satisfy high performance demands also for highly bursty traffic sources.

6 ACKNOWLEDGEMENTS

The simulations were performed on the sequential simulator YESS, developed by the Simulation Laboratory at KTH Department of Teleinformatics. The authors thank Robert Rönngren for his valuable help with the simulator and his interest in this work. We also thank Parag Pruthi at Bellcore, for providing us with - and explaining to us - the FPDI map traffic generator.

7 REFERENCES

Brady, P.T. (1968) A Statistical Analysis of On-Off Patterns in 16 Conversations. *The Bell System Technical Journal*, **Vol. 47**, 73-91.

Gustafsson, E. and Karlsson, G. (1996) When Is Traffic Dispersion Useful? A Study On Equivalent Capacity, in *Performance Modelling and Evaluation of ATM Networks* (ed. D. Kouvatsos), Second volume, Chapman & Hall.

Gustafsson, E. and Karlsson, G. (1997) A Literature Survey on Traffic Dispersion. To appear in *IEEE Network*, March 1997.

Kelly, F. (1996) Notes on effective bandwidths, in *Stochastic Networks: Theory and Applications* (ed. F. Kelly, S. Zachary, I. Ziedins), Oxford University Press, 141-168.

Kleinrock, L. (1975) *Queueing Systems, Volume I: Theory*, John Wiley & Sons.

Lee, H.H. and Un, C.K. (1986) A Study of On-Off Characteristics of Conversational Speech. *IEEE Trans. on Communications*, **Vol. COM-34, No. 6**, 630-637.

Leland, W.E. et al. (1994) On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Trans. on Networking*, **Vol. 2, No. 1**, 1-15.

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**, 1231-1243.

Paxson, V. and Floyd, S. (1995) Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Trans. on Networking*, **Vol. 3, No. 3**, 226-244.

Pruthi, P. (1995) *An Application of Chaotic Maps to Packet Traffic Modeling*. Ph.D. Dissertation, TRITA-IT R 95:19, Royal Institute of Technology, Stockholm, Sweden.

8 BIOGRAPHIES

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

Gunnar Karlsson received his master's degree from Chalmers University of Technology in 1983 and his Ph.D. from Columbia University in 1989. He has worked at IBM Zurich Research Laboratory from 1989 to 1992, and at the Swedish Institute of Computer Science since 1992. He is a member of the faculty at the Department of Teleinformatics at KTH, and is currently visiting professor at EPFL, Switzerland.