INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS, VOL. 10, 161–180 (1997)

PERFORMANCE OF CELL LOSS PRIORITY MANAGEMENT SCHEMES IN A SINGLE SERVER QUEUE

PIET VAN MIEGHEM[1]*, BART STEYAERT[2] AND GUIDO H. PETIT[1]
[1]*Alcatel Telecom Research Division, Francis Wellesplein 1, B-2018 Antwerpen, Belgium*
[2]*Laboratory for Communications Engineering, University of Gent, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium*

SUMMARY

The throughput optimality of priority management strategies in a single buffer has been studied for a general aggregate arrival law. The tight upper bounds found are useful to understand optimality in the utilization of specific priority schemes such as push-out buffer (POB) and partial buffer sharing (PBS). This paper further focuses on the maximum allowable load $\rho_{\max}$ versus the priority mix α for a PBS and a random push-out buffer (RPOB) of size K for a wide variety of arrival processes. The role of priorities in a special type of bursty arrivals, the compound Poisson process with constant burst length and random priority assignment within the burst is found to be less pronounced than that of 'pure' Poisson arrivals. On the other hand, the results for ON–OFF cell arrivals modelled by a MMPP(2), MMPP(3), and higher order Markov modulated processes (MMP) closely follow the behaviour of the maximum allowable load in the RPOB with Poisson arrivals, however, scaled to lower loads. The results indicate that the priority mix distribution within the aggregate arrival flow influences the shape of $\rho_{\max}(\alpha)$-curve more than the aggregate arrival distribution itself. © 1997 by John Wiley & Sons, Ltd.

Int. J. Commun. Syst., **10**, 161–180 (1997)

No. of Figures: 12 No. of Tables: 0 No. of References: 52

KEY WORDS: cell loss priority management; push-out buffer; partial buffer sharing

1. INTRODUCTION

This work focuses on connection admission control (CAC)[1, 2] of a single buffer with a two-type (high and low) priority management.[3] The quantity of interest for CAC is the maximum allowable load that a system can bear while still offering the requested quality of services (QOS). The QOS measure considered here is the cell loss ratio. Specifically, subject to the required cell loss ratios for both priorities, clr_L^* and clr_H^*, we determine the maximum allowable traffic intensity $\rho_{\max}$ as a function of the priority mix α and the buffer size K, where α denotes the probability that an arriving cell has high priority.

The literature abounds in suggestions to tackle the CAC problem in asynchrous transfer mode (ATM) switches. A smaller number of articles concentrates on a priority management. Most among those discuss a particular priority scheme and then proceed to evaluate the performance of the priority algorithm in a single buffer[3–13] or in a shared buffer, [14–18] for which we further refer to our work.[19] Generally one finds that the introduction of priorities enhances the number of customers that can be served adequately at the expense of an increased complexity of the control algorithm. However, relatively few (e.g. References 3 and 10) succeed in determining or proposing a concrete CAC algorithm that is optimal given a certain priority scheme. Hardly any paper discusses the trade-off between the gain in performance and the increase in complexity by introducing a priority scheme. Open issues in CAC providing a certain QOS are discussed in a broader scope by Kurose[20], however, omitting the priority problem. The latter topic is given a closer look in this paper.

Among cell loss priority management (CLPM) methods[3, 8, 10], the push-out buffer (POB) and the partial buffer sharing (PBS)[21] are most well-known. Although these priority schemes have been studied in the literature[3], the optimality of a priority scheme for various queue sizes and cell loss ratio requirements has not been discussed in detail. In a POB, the push-out mechanism acts only if the buffer is completely filled and a high priority cell arrives. If there are low priority cells in the buffer, the arriving high priority cell pushes the low priority cell nearest[†] to the server out, all cells behind the pushed-out low priority cell ripple through over one position towards the server, and the arriving high priority cell takes place at the tail of the queue in order to preserve cell sequence integrity. A PBS mechanism is somewhat simpler: if the buffer occupancy is below a threshold T, both low and high priority

[*] Correspondence to: Piet van Mieghem, Alcatel Telecom Research Division, Francis Wellesplein 1, B-2018 Antwerpen, Belgium.

[†] This push-out discipline is first-in/first-out (FIFO). Other alternatives are discussed in Section 2.

CCC 1074–5351/97/040161–20 $17.50

cells are allowed to enter, otherwise only high priority cells are accepted until complete buffer occupation.

Recently, Cidon *et al.*[22] have investigated optimal buffer sharing. They considered a shared buffer with N independent, incoming links each carrying Poisson traffic with load λ_i. The priority mechanism operates on the link index i. Using continuous-time Markov decision theory and dynamic programming[23], they found that the optimal scheme that maximizes the throughput is the threshold push-out strategy explained below. Their formalism involves cumbersome notation and actually requires heuristics for the optimality criteria whereas our approach in Section 3 is particularly transparent and provides attainable upper bounds. Using stochastic fluid models, Elwalid and Mitra[24] have analysed a two-buffer system with prioritized queues that is of particular interest for the emerging available bit rate (ABR)[25, 26] service in ATM. Briefly and only approximately, the CAC problem for their queueing model was touched upon.

Only a few articles present results of priority management involving bursty sources. The published analyses[27–30] are more focussed on the art of obtaining a queueing model with priorities than on a clear study of the benefits or gain of a priority scheme as a function of traffic parameters (such as burstiness, traffic intensity, etc.).

Using a fluid flow model, Garcia and Casals[27] report substantial statistical gains (over 200 per cent) obtained with partial buffer sharing operating on bursty sources. However, their article concludes that statistical gain is very sensitive to small changes in the parameters of the total traffic but rather insensitive to changes in the probability mix α. Hou and Wong[28] present a queueing analysis including delay and loss priorities for mixed continuous-bit-rate (CBR) and bursty traffic. Using a threshold type of priority mechanism, they demonstrate that their recursive model is efficient and flexible, but no details of the benefit of the priority mechanism on bursty sources are mentioned. Mitrou and Pendarakis[30] also touch on the priority problem and present a rather approximate model and an analysis based on a two-dimensional Markov chain. Unfortunately, no clear conclusions regarding the effect of priorities are mentioned. Liao[29] also presents a queueing analysis of partial buffer sharing with Markov modulated Poisson arrivals and briefly points to benefits of priorities. An analytic effective bandwidth method for partial buffer sharing under bursty arrivals was proposed by Kulkarni *et al.*[31], Saito[32] has modelled a push-out buffer with Markov modulated Poisson process (MMPP(N)) arrivals in continuous-time. His method relies on that of Kröner *et al.*[3] in that he concentrates for the description of push-out on the probability that a low priority cell reaches the server. Fonseca and Silvester[33] have proposed a multiclass selective discard mechanism. They have computed the loss probability per class where each class is modelled as a two-state MMPP.

Recently, Chang and Tan[34] have performed a comparison between a push-out and partial buffer sharing scheme with bursty (a three-state discrete-time Markov chain) arrivals. They have investigated the performance of both types as a function of burstiness, buffer size and buffer sharing threshold. It was concluded that partial buffer sharing can be made superior to push-out for high priority traffic at the expense of a dramatic increase in loss for low priority cells. In agreement with Garcia and Casals[27], they point out that the cell loss ratio is very sensitive to traffic details. For example, when fixing the overall traffic load but increasing the number of sources in each priority class, the cell loss ratio was found to increase significantly.

The role of more than two priority types for congestion control was discussed by Bemmel *et al.*[35], Petr *et al.*[36], and Yegani *et al.*[37] Chao *et al.*[38] have proposed a new cell discarding strategy: the self-calibrating push-out. Other dynamic priority queueing approaches were presented by Ren *et al.*[39] and Jun and Cheng[40]. Suri *et al.*[41] have proposed the threshold push-out and the P_{ow} push-out. The idea of both is similar: the push-out mechanism is made dependent on the cell loss ratio of both priority classes. In the threshold push-out, a high priority cell can only push out a low priority cell, if the number of low priority cells exceeds threshold T, otherwise it is discarded. Analogously, a low priority cell can push out a high priority cell if the number of high priority cells is larger than $K-T$ (where K is the buffer size). In the P_{ow} push-out, the push-out mechanism is triggered by a probability P_{ow} that a high priority cell is allowed to push out a low one and similarly, a low priority cell may push out a high priority cell with probability $1-P_{ow}$, provided that the buffer is full. Clearly, varying T and P_{ow}, respectively, provides mechanisms for adjusting the cell loss ratios of the two classes. Their analysis (considering Poisson arrivals) has compared these push-out variants with partial buffer sharing and shows that the push-out schemes could support a significantly higher maximum load subjected to a specific set of cell loss ratio requirements.

Combinations of several QOS metrics are proposed by Jeon *et al.*[42] and by Dailianas and Bovopoulos[43]. Huang and Wu[44, 45] propose a combined loss and delay priority mechanism but conclude that the exact analysis is hardly feasible due to the large dimensionality. Georgiadis *et al.*[46] discuss different non-preemptive policies for various applications. They define a simple analytical model that permits meaningful comparisons and that also allows the derivation of scheduling policies that are optimal in terms of delay and loss requirements. The existence and characteristics of policies that are jointly delay and buffer optimal are studied.

The outline is as follows. First, we discuss different push-out strategies and server disciplines. In

Section 3, we investigate the throughput optimality of a priority system in a single buffer and derive two upper bounds. In Section 4, we introduce the random push-out buffer (RPOB) and compare for Poisson arrivals the performance of partial buffer sharing to that of the push-out scheme. The main advantage of introducing the RPOB is that, first, it serves as an excellent approximation for the conventional first-in/first-out (FIFO) push out, and second, it allows us to perform exact calculations of the maximum allowable load for very general arrival laws. In Section 5, we introduce burstiness in the arrival pattern for the RPOB: we start with a compound Poisson process and then turn to arrivals generated by a Markov modulated process with N states (MMP(N)). The performance of RPOB and PBS are compared for an MMP(3). The detailed derivation of the state equations for the RPOB with MMP(N) cell arrivals are found in Appendix B.

2. PUSH-OUT STRATEGIES AND SERVER DISCIPLINES

In this section, different push-out strategies and server disciplines are compared. For PBS, a similar study is not relevant. The performance measure for the comparison is the cell loss ratio. An additional, influencing and underlying factor is revenue optimization. Clearly, a high priority service is more profitable than a low priority one. The purpose of a CLPM system is just to combine both service categories according to their wishes (in terms of *clr*∗ constraints) to generate as much profit (related to 'load') as possible. Under a push-out scheme operating on a prioritized traffic stream, it is understood that high priority cells remove low priority cells from the queue provided the queue is full. In case there are no low priority cells available for push-out, the arriving high priority cell is lost.

2.1. *Definition of the POB types*

Before embarking on the discussion, it is instructive to recall the difference between p/s POB systems. The first qualifier p specifies the push-out method, whereas the second defines the service discipline, which is always deterministic[†]. For the push-out method, we examine the cases $p = \text{FIFO}$, last-in first-out (LIFO) and R (random) in combination to $s = \text{FIFO}$ and R. We briefly outline the operation of these types.

For $p = \text{FIFO}$, the low priority cell with lowest queue position, or equivalently, closest to the server is pushed out with certainty. A $p = \text{LIFO}$ discipline operates similarly: the latest entered low priority cell is removed with certainty. For $p = R$ (random) an arbitrary low priority cell is removed; each low priority cell is equally likely as a candidate to go.

When $s = \text{FIFO}$, the cell on position 1 (thus next to the server) moves to the server irrespective of its priority. A $s = \text{LIFO}$ strategy serves the last entered cell first. For $s = R$, an arbitrary cell in the queue is taken by the server. The probability that the served cell has high priority equals $n_H[k]/n[k]$ where $n_H[k]$ and $n[k]$ are the number of high priority cells and of the aggregate (high plus low) in the buffer, respectively, at time slot k. A $s = \text{HOL}$ (head of the line) scheme always serves high priority cells in the queue before any low priority cell.

For ATM applications, only the $s = \text{FIFO}$ discipline is allowed since all other disciplines violate the sequence integrity. Here we exclude the $s = \text{LIFO}$ scheme because it fails to offer any attraction over the $s = \text{FIFO}$ strategy from a stochastic theoretical point of view.

2.2. *The class* s = *FIFO*

When comparing a FIFO/FIFO POB with a LIFO/FIFO and R/FIFO POB we can demonstrate the following:

Property 1. *The best performance in the* s = *FIFO class is achieved by a* p = *FIFO strategy*

- It is sufficient to show that neither a R/FIFO POB nor a LIFO/FIFO POB can serve more high priority cells on average over time than a FIFO/FIFO POB obeying the *clr*∗ constraints. Operating near the 'allowable' constraints implies that push-out actions occur relatively often.
- The difference between the FIFO/FIFO POB and the R/FIFO POB might occur when i out of L low priority cells are at positions 1 to i and a high priority at position $i + 1$ when i push-out actions take place. In the FIFO/FIFO POB, these i low priority cells are certainly removed and the high priority cell in position $i + 1$ is certainly served the next time. In the R/FIFO POB, one can only say that there is a probability of $\binom{L}{i}^{-1}$ that this high priority cell at position $i + 1$ will be served the next time. Comparing both systems, we observe in addition that at the next time slot, probably fewer low priority cells remain as candidates for a push-out in the R/FIFO POB. Hence, every time a low priority cell is served where a high priority cell could have been chosen, this 'wrong choice' influences the performance in both the current and the next time slots badly.
- The situation for $p = \text{LIFO}$ is analogous, and even worse than for $p = R$, because the probability to have low priority cells close to the server is larger than for the $p = R$ push-out mechanism.

[†] Deterministic means that at the end of a slot, k, there is precisely one cell transferred from the queue to the server, provided, of course, that the queue is not empty.

Intuitively, in $p = \text{FIFO}$, the low priority cells are, on average, moved further back in the queue compared to $p = \text{LIFO}$, where they are encouraged to take place in front. The $p = R$ scheme fits somewhere in between these extremes. Now, serving more low priority cells implies, on average, serving fewer high priority cells, supporting property 1 and the following corollary.

Corollary 1. *When* s = *FIFO, the performance for* p = R *is better than for* p = *LIFO*

This analysis suggests that introducing a random server discipline ($s = R$) may boast the performance for $p = R$ in order to achieve a comparable performance to that of a FIFO/FIFO POB. Although apparently only of academic interest, the R/R POB will be studied because the computational burden is much less heavy than for FIFO push-out (see Section 4.2).

2.3. *The class* s = R

Property 2. *There is no difference in performance between the several* p-*schemes*

Proof:

- The $s = R$ discipline uniformly chooses a cell from the queue. The type of the cell served only depends on the number of high and low priority cells.
- For all p-types, the number of push-out operations in each time slot is precisely the same, and so is the number of high and low priority cells. Hence, there is an equal chance for all p-schemes that a high priority cell is served, thus their performance is equal.

On property 2, we will simplify R/RPOB to RPOB in the sequel.

2.4. *The* s = FIFO *class versus the* s = R *class*

We will confine ourselves to a comparison between the FIFO/FIFO POB and a p/RPOB. Since the performance of the last class is independent of the push-out scheme p, we base the discussion on the $p = R$ scheme. The performance of the FIFO/FIFO POB is very close to that of the R/R POB as demonstrated below and in Figure 1. At first glance, this result is surprising because it implies that the sequence integrity is immaterial for our performance standard. However, a closer look reveals that the result is in fact quite natural.

In first order, a push-out action depends on the number of low priorities in the buffer and only in second order to their precise cell position (that, of course, depends on the sequence order). Furthermore, the arrivals of high priorities that are responsible for a possible push-out action are the same for both buffer systems because arrivals at time k are usually independent of the buffer content at that moment. In both cases, a push-out action diminishes the number of low priority cells if there are any in the system. Therefore, if the number of low priority cells at a certain time is the same in both systems, the number of pushed out cells is equal.

In the timeslots k where the buffer is full and high priority cells arrive, at the next timeslot $k + 1$, the $s = \text{FIFO}$ discipline certainly serves a high priority cell, and in these timeslots $s = \text{FIFO}$ is superior to the $s = R$ discipline. On the other hand, at timeslots j where the buffer is not completely filled, the priority mechanism does not interfere. In these timeslots j, there are situations where $s = R$ is never worse (with certainty) than a $s = \text{FIFO}$. For instance, when, at the end of a slot, there are low priority cells located in the FIFO/FIFO POB before high ones, the low ones will be served in the next slot. For the $s = R$ discipline, there is always a chance that a high priority cell is served. In summary, when averaging over all timeslots, the presented qualitative discussion illustrates that the differences between both server disciplines cannot be substantial.

Another argument is that of the self-regulation of a CLPM system subject to *clr* constraints (clr_L^*, clr_H^*). Consider two work-conserving POB strategies, POB1 and POB2, with a service discipline equally fair treating low and high priority cells. Suppose that POB1 systematically handles low priority cells more favourably than POB2. Hence, POB1 will typically contain fewer high priority cells than POB2. In case a push-out action occurs, POB1 possesses fewer low priority cell candidates with a consequence that $clr_{H1} > clr_{H2}$. The *clr* constraints will interfere and generate the relevant feedback with respect to the arrival intensity merely resulting in a small difference in performance.

3. GENERAL RELATIONS

3.1. *Definitions*

By virtue of the slotted nature of ATM, we concentrate on discrete-time systems where the servers work deterministically. The time unit, also called a time slot, equals the time needed to serve precisely one cell. If μ_i denotes the fraction of served i priorities per time slot, we have

$$\mu_A = \mu_H + \mu_L = 1 \tag{1}$$

where the subscripts refer to the aggregate (A), the low priority cells (L) and the high priority cells (H), respectively.

If α denotes the probability that an arriving cell has high probability, the mean number of arrivals per time slot equals

$$\lambda_A = \lambda_H + \lambda_L \tag{2}$$

where $\lambda_H = \alpha\lambda_A$ and $\lambda_L = (1 - \alpha)\lambda_A$. Defining the traffic intensity as usual by $\rho = \lambda/\mu$, we observe that for a deterministic server, it holds that $\lambda_A = \rho_A$.

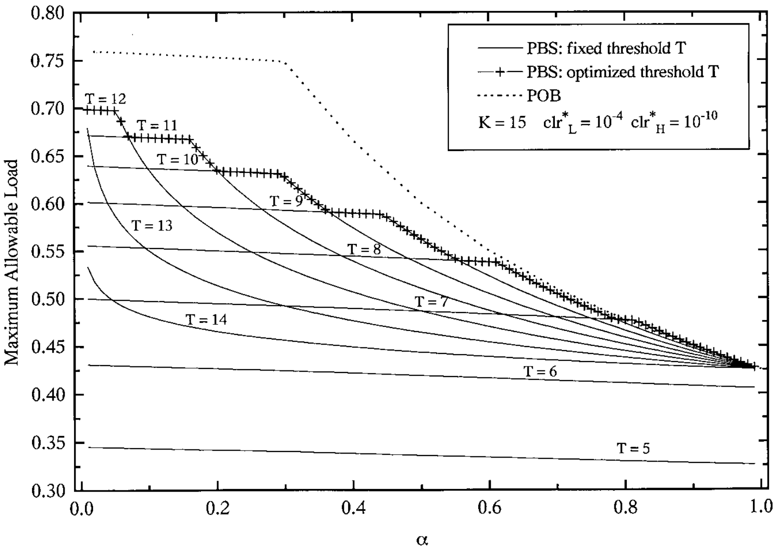

Figure 1. The effect of the threshold T on the performance of PBS in a relatively small buffer of size $K = 15$ for the cell loss ratio couple $(10^{-4}, 10^{-10})$. For comparison purposes, the performance of the POB is shown as a dotted line

Since the system has a finite capacity of K queueing positions with an additional one for the server, in general, cell loss will occur. The cell loss ratio clr is defined as the mean number of cells lost per time slot over the mean number of cells of that type which have arrived. Again, the total number of lost cells consists of both priorities. From this fact we deduce a useful equation†,

$$\lambda_A\, clr_A = \lambda_L\, clr_L + \lambda_H\, clr_H$$

$$clr_A(\alpha) = (1-\alpha) clr_L(\alpha) + \alpha\, clr_H(\alpha) \qquad (4)$$

where

$$\left(\alpha = \frac{\lambda_H}{\lambda_A}\right)$$

The last equation explicitly expresses the dependence on α. In addition, since we can write the aggregate cell loss ratio as a weighted mean, $clr_A = (\lambda_L\, clr_L + \lambda_H\, clr_H)/(\lambda_L + \lambda_H)$, we immediately find that $clr_H(\alpha) \le clr_A(\alpha) \le clr_L(\alpha)$ if we assume that $clr_H(\alpha) \le clr_L(\alpha)$.

The cell loss ratio of the aggregate cell stream, clr_A, in the corresponding system without the priority management is exactly described by the loss probability of that corresponding $G/D/1/K$ system (see e.g. References 26, 47 and 48). Formally, fixing all other traffic descriptors independent of the load ρ_A, we have

$$\widehat{clr}_A = f_K(\hat{\rho}_A) \qquad (5)$$

where $f_K(x)$ is an increasing, continuous and positive function of x bounded by $0 \le f_K(x) \le 1$ and non-increasing in K. A priority mechanism can never lower the aggregate cell loss, hence, we have

$$\widehat{clr}_A \le clr_A(\alpha) \qquad (6)$$

and alternatively, for the same aggregate cell loss ratio requirement $\widehat{clr}_A = clr_A(\alpha) = clr^*_A$

$$\hat{\rho}_{AA} \ge \rho(\alpha) \qquad (7)$$

3.2. *Formal solution*

We are now in a position to treat the problem in more detail: given a priority management protocol, determine the maximal traffic intensity ρ_A subjected to the user's cell loss ratio requirements (clr^*_L, clr^*_H) such that $clr_L(\alpha) \le clr^*_L$ and $clr_H(\alpha) \le clr^*_H < clr^*_L$. The latter inequality means that clr^*_H should be sufficiently smaller than clr^*_L in order for the priority scheme to have impact. Indeed, when $clr^*_H \rightarrow clr^*_L$, and hence, $clr^*_H \rightarrow clr^*_A$, the priority mechanism is abused since it is forced to be independent of α (see footnote on page 166 for a numerical example).

Since $f_K(x)$ is monotonously increasing, the inverse function exists justifying to rewrite equation (5) as $\hat{\rho}_{AA} = f_K^{-1}(\widehat{clr}_A)$. Furthermore, the inverse function $g^{-1}(x)$ of an increasing function $g(x)$ is increasing. Using equation (7), we have $\rho(\alpha) \le f_K^{-1}(clr^*_A)$. Hence, the maximum allowable load $\rho_{\max}(\alpha)$ is found where $clr_A(\alpha)$ is maximal. Specifically, from equation (4) and the requirements on the cell loss ratios, we have

$$clr_A(\alpha) \le (1-\alpha) clr^*_L + \alpha\, clr^*_H \qquad (8)$$

† An alternative relation of the same nature is

$$\lambda_A(1 - clr_A) = (1 - q[0])\mu_A \qquad (3)$$

where $q[0]$ is the probability that the buffer is empty.

offering an upper bound for the maximal allowable load

$$\rho_{\max}(\alpha) \leq f_K^{-1}\left((1-\alpha)clr_L^* + \alpha\, clr_H^*\right) \qquad (9)$$

Since the right-hand side of equation (8) is decreasing in α due to the fact that $clr_H^* < clr_L^*$, so is equation (9). The upper bound equation (9) does not depend on the management protocol and indicates that for every value of $\alpha \in [0,1]$ both requirements, $clr_L(\alpha) = clr_L^*$ and $clr_H(\alpha) = clr_H^* < clr_L^*$ are met. We will now show that the equality sign in equation (9) does not hold for all α emphasizing that equation (9) forms an unattainable upper bound.

From the definition of the priority mix α and the fact that $\rho_A = \lambda_A$, the following inequality arises

$$\rho_A(\alpha) = \frac{\lambda_H(\alpha)}{\alpha} \leq \frac{\lambda_H(1)}{\alpha} = \frac{\rho_A(1)}{\alpha} \qquad (10)$$

because $\lambda_H(\alpha)$ is increasing in α. Notice that a similar condition for low priority cells $\rho_A(\alpha) \leq \rho_A(0)/(1-\alpha)$ is always fulfilled by equation (9) since the left-hand side is decreasing in α whereas the right-hand side increases in α. The inequality in equation (10) poses a lower upper bound than equation (9) for an α-region near $\alpha = 1$, which can be achieved by one priority management protocol as shown below. Invoking the characteristic property of a deterministic server (equation (1)) we can write

$$\rho_A(\alpha) = \frac{\rho_H(\alpha)\,\mu_H(\alpha)}{\alpha} = \frac{\rho_H(\alpha)}{\alpha}\left(1 - \mu_L(\alpha)\right) \qquad (11)$$

The priority management algorithm that maximizes equation (11) for α close to 1, will minimize the number of served low priority cells. The extreme, of course, is a zero service for the low priority cells $\mu_L = 0$ as almost realized in a head of the line preemptive push-out discipline (HOL POB)† and precisely met by a PBS scheme with threshold $T = 0$.

In conclusion, the maximum allowable load $\rho_{\max}$ is bounded for low α by equation (9) and for high α by equation (10). The upper bounds in equations (10) and (9) coincide at $\alpha = 1$, but have opposite curvatures for $\alpha \leq 1$. In addition around $\alpha \leq 1$ the bound in equation (10) is smaller than in equation (9). Hence, there must exist a certain value of α, α_c, where both upper bounds intersect. A system that closely attains these upper bounds as a HOL POB possesses a maximum allowable load $\rho_{\max}(\alpha)$ that is not differentiable with respect to α at α_c.

Since the cell loss decreases with increasing buffer size K both extremes $\rho_{\max}(0)$ and $\rho_{\max}(1)$ will tend to each other for sufficiently large K. As a consequence, the critical point α_c will tend to unity for large K. This demonstrates that a priority management is almost useless for CAC when large buffers can be utilized (e.g. when time delay constraints are unimportant). Hence, when two cell loss ratio requirements are specified, the role of loss priorities in ATM is questionable for large buffers since the complexity of the control mechanism with priorities is hardly compensated by the gain in performance.

4. POISSON ARRIVALS

This section compares two standard priority schemes, the POB and PBS for Poisson arrivals. The emphasis lies on a newly introduced model, the RPOB, that is further studied under bursty arrival processes in the next section.

4.1. *Partial buffer sharing*

The maximum allowable load for PBS is strongly dependent on the threshold $T \leq K$. We have computed the threshold T_{opt} that maximizes the aggregate load using the discrete-time version of the model of Kröner *et al.*[3] The effect of the threshold T on the performance is illustrated in Figure 2. For small values of T the low priority cell loss ratio requirement clr_L^* is dominating and the opposite is seen for high values of T. The intermediate values clearly introduce two α regions similar to those of the POB. The desired maximum allowable load is the maximum envelope of all these curves and is a concatenation of regions alternately dominated by the high and low priority cell loss requirement. The normalized optimal threshold T_{opt}/K versus α is shown in Figure 3 for various K values. Together with Figure 2, the plot illustrates that, due to the integer character of T, analytic optimization is hardly feasible for small values of K. The longer the buffer size K, the more integer values of T there are available, resulting in a smoother maximum allowable curve. Figure 4 plots the maximum allowable load $\rho_{\max}(\alpha)$ versus α for large values of K and the minimum of the upper bounds in equations (10) and (9). This graph clearly demonstrates how closely

† In a HOL POB, the high priorities are not influenced by the presence of the low ones because they are always served prior to any low priority cell. Hence, a low priority cell is only served if there are no high priorities in the buffer at a time slot. The cell loss ratio for the high priority cells is given by the same $G/D/1/K$ expression that describes the aggregate. We have

$$clr_H(\alpha) = f_K(\alpha\rho_A(\alpha)) \qquad (12)$$

$$clr_L(\alpha) = \frac{f_K(\rho_A(\alpha)) - \alpha\, f_K(\alpha\rho_A(\alpha))}{1-\alpha} \qquad (13)$$

Knowing that $f_K(x)$ is increasing with x, we readily establish that a HOL POB almost attains both discussed upper bounds in equations (9) and (10) provided clr_H^* is sufficiently smaller than clr_L^*. Otherwise, putting $\mu_L = 0$ can violate the low priority cell loss requirement. A simple numerical example illustrates this situation for $clr_L^* = 10^{-5}$, $clr_H^* = 10^{-6}$ and a buffer size $K = 10$. When the aggregate arrivals process is Poisson with $\lambda_A = 0.4755$ and $\alpha = 0.99$, we have for the $M/D/1/K$ queue that $clr_A(0.99) = clr_A = 1.0816\ 10^{-6}$. The RPOB gives $clr_H(0.99) = 1.0\ 10^{-6}$ and $clr_L(0.99) = 9.02\ 10^{-6}$ whereas the HOL POB (equations (12,13)) offers $clr_H(0.99) = 9.30\ 10^{-7}$ and $clr_L(0.99) = 1.61\ 10^{-5}$. Hence, the RPOB meets the requirements whereas the ‘normally superior’ HOL POB fails to obey them.

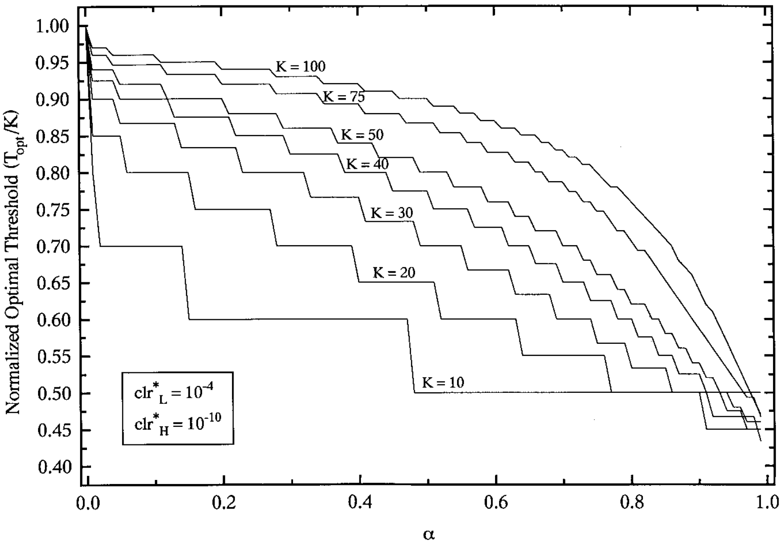

Figure 2. The normalized optimal threshold T_{opt}/K in PBS for various buffer sizes K but fixed cell loss ratio couple (10^{-4}, 10^{-10})

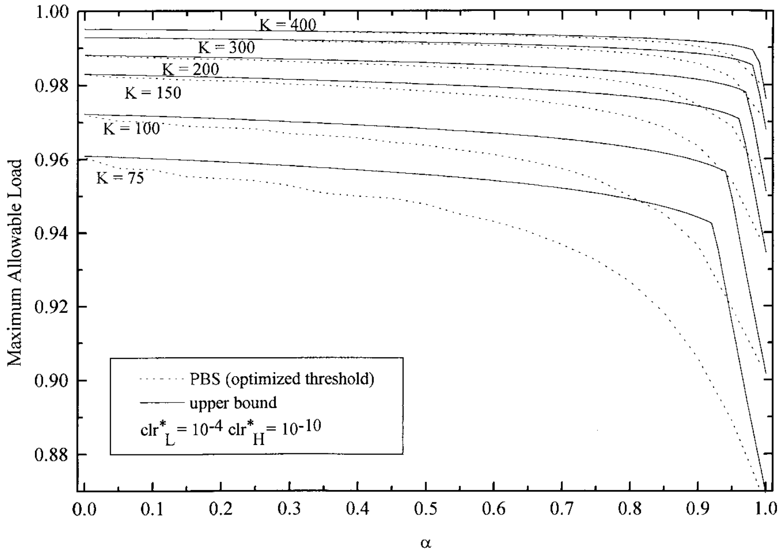

Figure 3. The maximum allowable load in PBS with optimized threshold versus α and the minimum of the upper bounds in equations (10) and (9) for various large buffer sizes K but fixed cell loss ratio couple (10^{-4}, 10^{-10})

PBS (with optimized threshold) approaches the best possible performance for large α, but also that it fails to treat the low priorities in an optimal way.

4.2. *The push-out buffer*

For small α, the aggregate cell loss ratio will be mainly determined by $clr_L(\alpha)$ since there are hardly any high priority cells. Moreover, since generally $clr^*_H \ll clr^*_L$, we have from equation (4) that $clr_A(\alpha) \approx clr^*_L(1-\alpha)$. Invoking equation (9) we conclude that the maximal allowable load is dominated by the clr^*_L requirement. In this region, the cell loss ratio requirement for the low priority cell is precisely met ($clr_L(\alpha) = clr^*_L$), whereas for the high priority cells $clr_H(\alpha) < clr^*_H$. Increasing α or the average number of high priority cells causes $clr_H(\alpha)$ to increase until $clr_H(\alpha) = clr^*_H$. At this point, denoted as α_k, both cell loss ratio requirements are precisely met (and this point is unique as follows by a continuity argument).

The situation is more complex for high values of α. For sufficiently high α, $\rho_{\max}(\alpha)$ follows from equation (11). The problem is how to determine the service rate $\mu_L(\alpha)$ for the low priority cells. For values of α just exceeding α_k, the load will be limited by the high priority requirement such that $clr_H(\alpha) = clr^*_H$ whereas $clr_L(\alpha) < clr^*_L$. However, since $clr^*_H \ll clr^*_L$, we find that $clr_L(\alpha)$ still dominates the aggregate cell loss ratio $clr_A(\alpha)$. When $\alpha > \alpha_k$, the loss in low priority cells will be substantial due to the push-out mechanism leading to $clr_L(\alpha) \approx clr_{\text{Lpo}}(\alpha)$. The calculation of the push-out probability is exceedingly complicated, and we

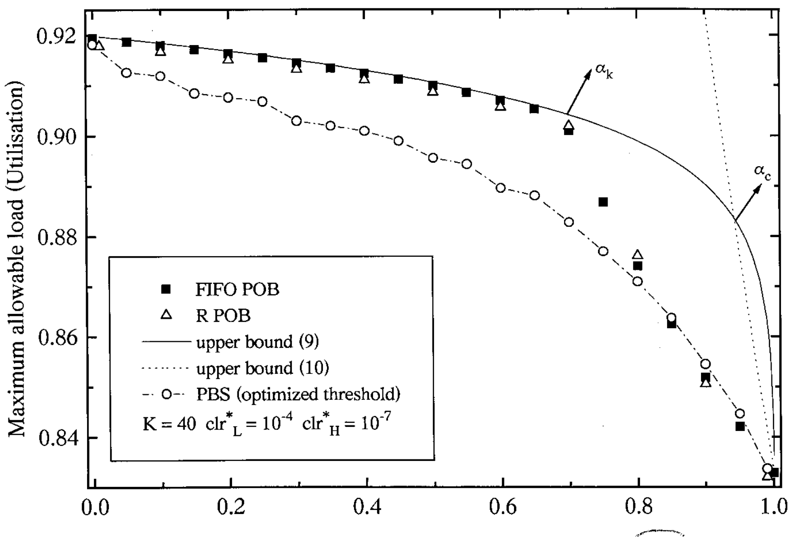

Figure 4. The maximum allowable load versus the priority mix α for a FIFO POB[3] and an RPOB of size $K = 40$ for the cell loss requirements $clr^*_L = 10^{-4}$ and $clr^*_H = 10^{-7}$. We have also drawn both the upper bounds of equations (9) and (10)

believe it is only possible through solving the transition probability matrix.

We have investigated two types of POB: a conventional FIFO POB (as studied by Kröner *et al.* in continuous time[3]) and a RPOB. The delimiter refers to the service discipline as well as to the push-out strategy. Clearly, the RPOB does not obey the sequence integrity. However, as argued above, since the cell loss ratio only weakly depends on the sequence order, the maximum allowable load of the RPOB is expected to approach that of the FIFO POB closely, provided the cell loss ratio requirements are sufficiently stringent ($clr^* < 0{\cdot}1$). Indeed, for both POB types and for Poisson arrivals† the comparison in the maximum allowable load $\rho_{\max}(\alpha)$ versus α shows, as illustrated in Figure 4 that both priority management systems exhibit very similar performances for $\rho_{\max}$.

The main reason for introducing the RPOB is the drastic simplification of computation. For a RPOB, the computation of the occupancy probability density function in a buffer of size K requires solving a set of $K(K+1)/2$ linear equations (see Appendix B.2), whereas for the FIFO POB, the effort consists of solving a set of 2^{K+1} linear equations. The reason for the difference lies in the sequence integrity. For the FIFO POB in contrast to the RPOB, we have to keep track of the order in which both types of priority cells are queued. This number of possible configurations in the buffer is equivalent to the highest binary number we can form with $K+1$ digits (K for the buffer and one for the server), hence, 2^{K+1}.

4.3. *POB versus PBS*

In Figures 5 and 6, we present $\rho_{\max}(\alpha)$ for the RPOB and PBS with optimized threshold T. We show two sets of cell loss ratios (clr^*_L, clr^*_H): (10^{-4}, 10^{-7}), (10^{-4}, 10^{-10}) as suitable representative priority classes in ATM. For small buffer sizes, K, POB is superior over the whole priority mix region. However, in case K is large, PBS can be controlled closer to the upper bounds in equations (9) and (10) than a POB, and we observe that PBS can guarantee a slightly higher load for the high priorities in an α-region close to unity. This fact was also observed by Chang and Tan[33]. However, once the priority mix $\alpha \le \alpha_k$, the POB approaches the upper bound in equation (9) and is undoubtedly the better strategy. As an overall conclusion, the POB offers a better treatment of low priorities, whereas PBS can be engineered (by adjusting the threshold T) to obtain a higher load for high priorities when $\alpha > \alpha_k$.

This analysis shows that a priority strategy combining the benefits of both POB and PBS such as the threshold push-out proposed by Suri *et al.*[39] can result in a higher performance for all α. However, the implementation of the latter, more refined priority schemes is undoubtedly more complex than that of the conventional POB.

4.4. *RPOB fit for $\rho_{max}(\alpha)$*

Since $\rho_{\max}(\alpha)$ of a RPOB in the $[0,\alpha_k]$ interval is sufficiently closely approximated by equation (9) as illustrated in Figure 4, our objective is to find an estimate in $[\alpha_k,1]$ accurate to within 1 per cent.

Suppose for the moment that the value of α_k is known. We found that the data of the maximum allowable load determined via a matrix solution of the RPOB (see Appendix B) is well fitted by

† Also for MMP(N) arrivals, we found via simulations that the agreement is very good.

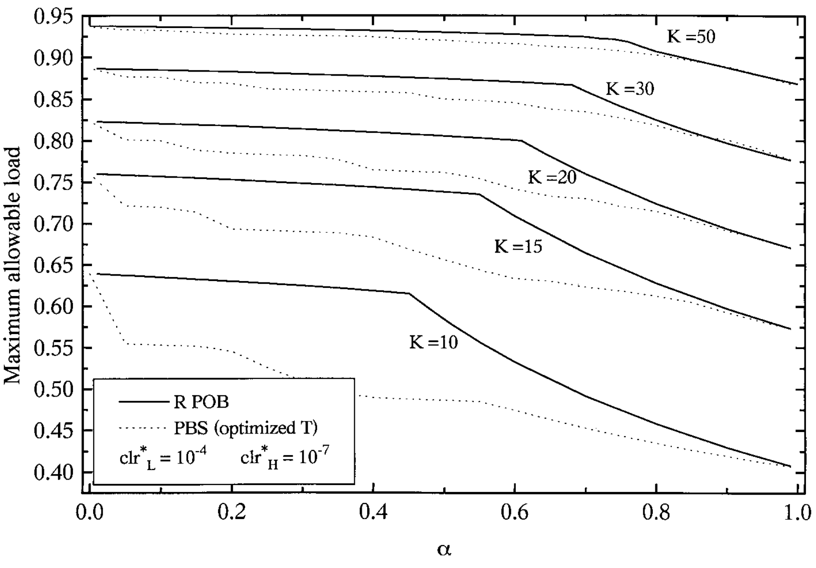

Figure 5. Calculation of the maximum allowable load $\rho_{\max}$ for various buffer sizes K versus the priority mix α for the cell loss requirements $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-7}$. The curves are obtained for the RPOB

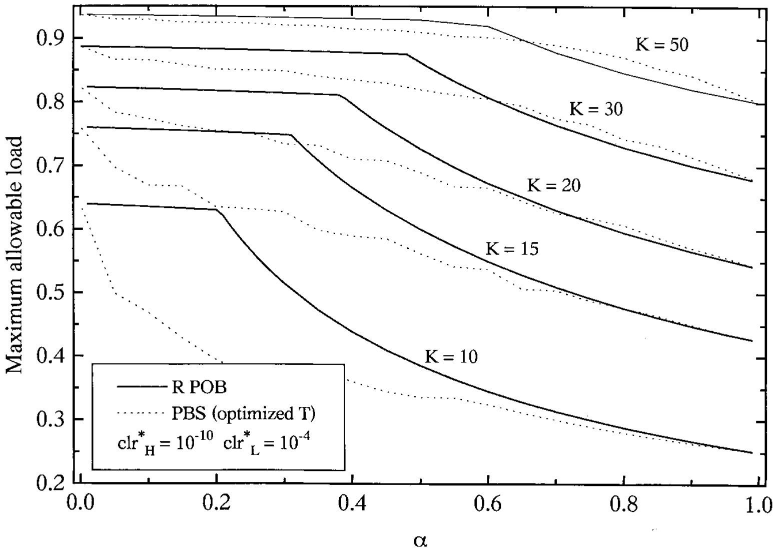

Figure 6. Calculation of the maximum allowable load $\rho_{\max}$ for various buffer sizes K versus the priority mix α for the cell loss requirements $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-10}$. The curves are obtained for the RPOB

$$\rho_{\max}(\alpha) = p_1 + \frac{p_2}{(\alpha + p)^2} \qquad (14)$$

Introducing the additional information

$$\rho_{\max}(1) = f_K^{-1}(clr_H^*)$$

$$\rho_{\max}(\alpha_k) = f_K^{-1}((1-\alpha_k)clr_L^* + \alpha_k clr_H^*)$$

equation (14) can be specified as

$$\rho_{\max}(\alpha) = \frac{1}{D}\left[\rho_{\max}(1)\left(\frac{1}{(\alpha+p)^2} - \frac{1}{(\alpha_k+p)^2}\right) + \rho_{\max}(\alpha_k)\left(\frac{1}{(1+p)^2} - \frac{1}{(\alpha+p)^2}\right)\right] \qquad (15)$$

where $D = 1/(1+p)^2 - 1/(\alpha_k+p)^2$. An elegant approximation for $f_K^{-1}(x)$ in a discrete-time $M/D/1/K$ is given in Appendix A.

The proposed fit in equation (15) is a kind of weighted mean between $\alpha = \alpha_k$ and $\alpha = 1$ with weight function $(\alpha + p)^{-2}$. Apart from α_k, the only unknown is p for which we found $0{\cdot}5 \le p \le 1$. The result is not very sensitive to variations in p (in contrast to α_k) when aiming at an accuracy of 1 per cent. The remainder is therefore devoted to the study of α_k.

For a fixed ratio $\beta = clr_H^*/clr_L^*$ but variable K, we observed that $\log \alpha_k = A/K + B$. On the other hand, for a fixed buffer size K, we found that $\log \alpha_k$ is linear in $\log \beta$ for both the high and low asymptotic values. In practical applications, β is often smaller than 10^{-3} and the low asymptotic regime is adequate to use. After rather extensive fitting this regime can be properly modelled as

$$\alpha_k \approx 10^{\frac{-3}{2K}} (clr_L^*)^{\frac{1}{4K}} \beta^{\frac{1}{K}} \qquad (16)$$

Figure 7 compares the quality of the fit procedure described above with the FIFO POB[3] and the RPOB. This plot exhibits that approximately a 1 per cent accuracy is achieved.

5. INTRODUCING BURSTINESS IN THE ARRIVAL PROCESS

So far, a Poisson arrival law has been considered. Since ATM traffic is very likely to be bursty, inclusion of this characteristic is in order. First, we will confine ourselves to a compound Poisson arrival process, described on a slot-per-slot basis by the generating function $e^{-\lambda(1-B(z))}$, where the generating function $B(z)$ specifies the distribution of the number of cells within a (Poissonean) burst. Then the performance of the RPOB and PBS is investigated for arrivals generated by a Markov modulated process with N states (MMP(N)).

5.1. *Compound Poisson process*

As an example, we take $B(z) = z^B$, meaning that each burst consists precisely of B cells and the bursts arrive according to a Poisson law with parameter λ, hence the load (traffic intensity) equals λB. We have compared, only for the RPOB, two extreme cases of priority distribution within a burst. In the first case, all cells in a burst have precisely the same priority and the probability to have a high priority burst is α. In the second case, the cells within a burst have high priority with probability α, and each cell is assigned a priority independent of the others. Figure 8 plots the maximum allowable λ for both cases.

In the case of random priority assignment, the result shown in Figure 8 demonstrates that introducing 'uncorrelated burstiness' makes $\rho_{\max} = B\lambda$ less dependent on α for burst lengths B small compared to the buffer size K, a conclusion previously drawn by Garcia and Casals.[27] When the burst length B approaches K, the dependence of $\rho_{\max}$ on α increases slightly.

In the case of the same priority assignment in a burst, the performance is, as expected, always lower than in the random priority assignment case. Actually, we found that the performance ($\rho_{\max}$) in the RPOB of size K with a compound Poisson arrival with parameter λ and burst size B (same priority assignment in a burst), is precisely the same as the performance in a RPOB of size K/B, when this fraction is an integer.

5.2. *Markov modulated Poisson process*

We refer to Appendix B for the detailed derivation of the RPOB with MMP(N) arrivals in discrete-time. The MMPP-PBS has been computed by extending the results of Kröner *et al.*[3]

A possible way to relate the characteristics of the actual arrival process to the set of parameters describing an N-state MMP, is to consider the arrival process as a succession of ON and OFF slots. During an OFF slot, no cells are generated, whereas during an ON slot, the number of cell arrivals in each ON-state is assumed to be Poisson distributed, with mean λ. Let σ denote the probability that an arbitrary slot is an ON slot.

In the case $N = 1$, the cell arrival process is i.i.d. and can be described on a slot-per-slot basis by the probability generating function (PGF)

$$A(z) = 1 - \sigma + \sigma\, e^{\lambda(z-1)} \qquad (17)$$

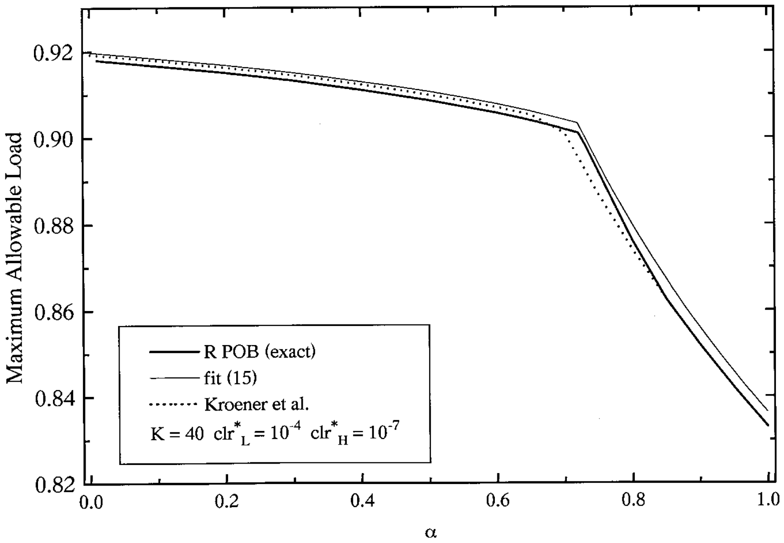

Figure 7. Comparison of the maximum allowable load $\rho_{\max}$ versus the priority mix α computed via different methods: the FIFO POB by Kröner *et al.*,[3] the RPOB and our proposed fit in equation (15) or (53)

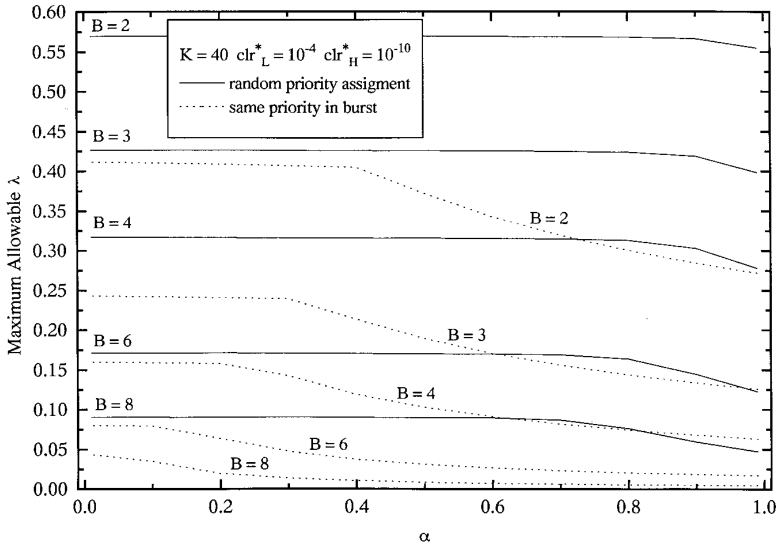

Figure 8. The maximum allowable λ (the load per burst B) in the RPOB versus the priority mix α for various burst lengths B but fixed buffer size $K = 40$ and fixed cell loss ratios $clr^*_L = 10^{-4}$ and $clr^*_H = 10^{-10}$. Curves with random priority assignment within a burst are drawn as an unbroken line, whereas the dotted line represents the case where all cells in a burst have the same priority

Defining an ON (OFF) period as a consecutive number of ON slots, then each ON period is followed by an OFF period (and vice versa), and the length of the respective ON and OFF periods expressed in units of time slots is geometrically distributed with parameter σ and mean $1/\sigma$, respectively parameter $1-\sigma$ and mean $1/(1-\sigma)$. For fixed values of the overall load $\sigma\lambda$, low values of σ means that all cell arrivals are grouped into a relatively small number of slots, whereas values of σ close to 1 imply that the cell arrivals are spread over virtually all slots. Numerical examples (Figures 9 and 10) illustrate the strong impact of σ on the admissible aggregate load.

In a two-state model (Figure 11) with modulator

$$P(2) = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}$$

and Poisson arrival rates $\Lambda(2) = diag\{\lambda, 0\}$ (defined in Appendix B), the length of the ON periods is geometrically distributed with parameter α and mean $1/(1-\alpha)$, whereas the length of the OFF periods is geometrically distributed with parameter β and mean $1/(1-\beta)$. Hence, when $\alpha = 1-\beta = \sigma$, the two-state model reduces to the previous case ($N = 1$) of i.i.d. arrivals. The probability that an arbitrary slot is an ON slot is given by

$$\sigma = \frac{1-\beta}{2-\alpha-\beta} \tag{18}$$

Notice that the steady state vector π of the modulator $P(2)$ equals $\pi_1 = \sigma$ and $\pi_2 = 1-\sigma$. We further define κ as the ratio of the mean length of an ON (OFF) period to the mean length of an ON (OFF) period in the case of i.i.d. arrivals,

$$\kappa \overset{\triangle}{=} \frac{1-\sigma}{1-\alpha} = \frac{\sigma}{1-\beta} \tag{19}$$

The parameter set $(\sigma, \kappa, \lambda)$ can now be used instead of (α, β, λ) to characterize the two-state MMPP. Large values of κ indicate that on average successive ON and OFF periods are long compared to the i.i.d. case ($N = 1$). Therefore, κ can be regarded as a measure for the burstiness in the arrival pattern.

In the three-state MMPP (Figure 12) with modulator

$$P(3) = \begin{pmatrix} \alpha_1 & 0 & 1-\alpha_1 \\ 0 & \alpha_2 & 1-\alpha_2 \\ q(1-\beta) & (1-q)(1-\beta) & \beta \end{pmatrix}$$

and Poisson arrivals rates $\Lambda(3) = diag\{\lambda, \lambda, 0\}$, we confine ourselves to a model with two types of ON periods, represented by ON1 and ON2, both geometrically distributed, with parameter α_1 and α_2, respectively. As before, the length of the OFF periods is geometrically distributed with parameter β. Each OFF period is followed either by an ON1 period, with probability q, or by an ON2 period, with probability $1-q$. The overall distribution of the length of an ON period is a weighted sum of two geometric distributions which allows us to investigate the impact of the variance in the distribution of the length of an ON period on the admissible load. To that extent, we define R as the ratio of the variance of the length of an ON period in this model to the variance of the length of an ON period in the previous case $N = 2$,

$$R \overset{\triangle}{=} \frac{(1-\sigma)^2}{(\kappa+1-\sigma)\kappa}\left[q(1-q)\left(\frac{1}{1-\alpha_1} - \frac{1}{1-\alpha_2}\right) + \frac{q\alpha_1}{(1-\alpha_1)^2} + \frac{(1-q)\alpha_2}{(1-\alpha_2)^2}\right] \tag{20}$$

where

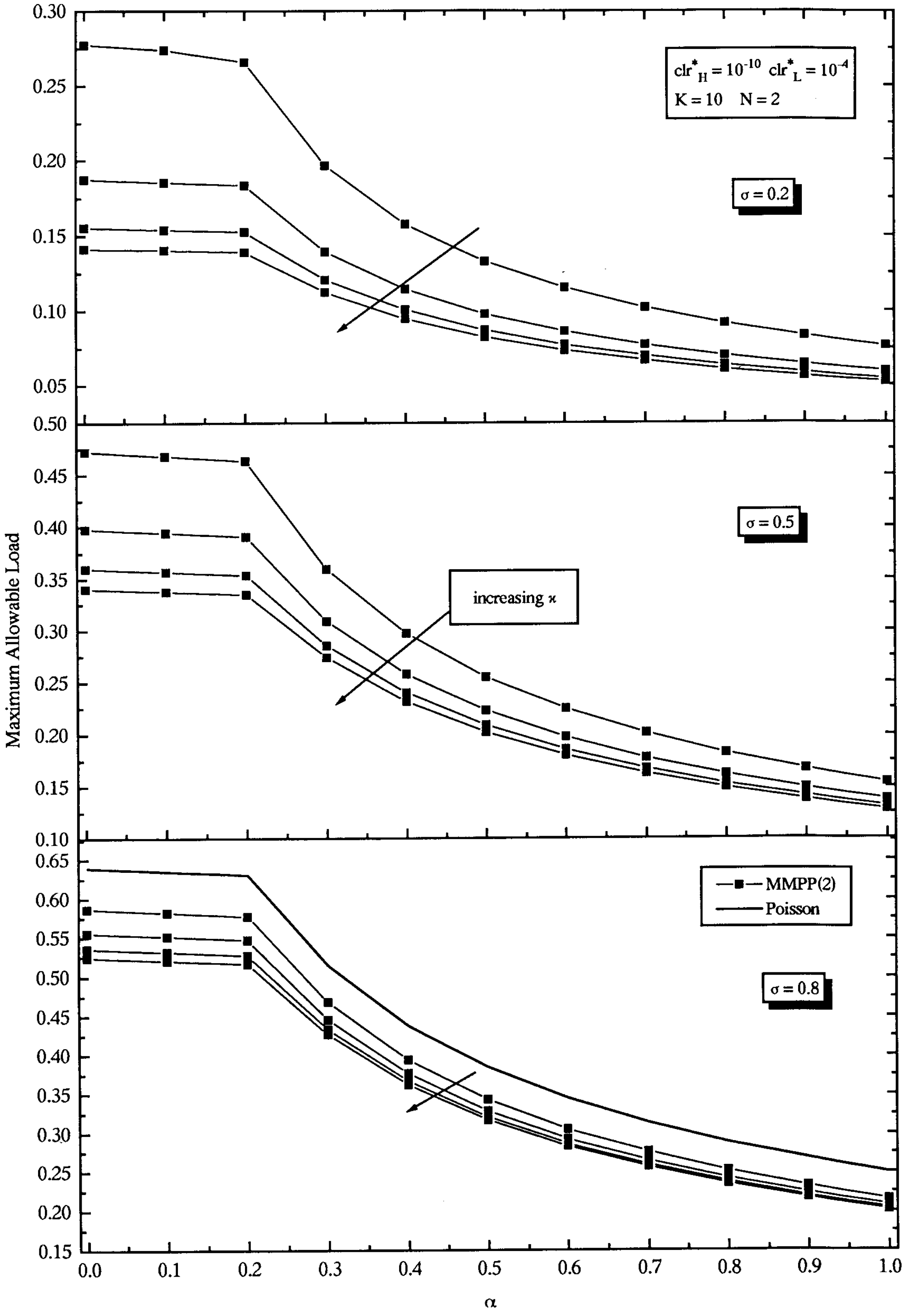

Figure 9. The maximum allowable load $\rho_{\max}$ in the RPOB versus the priority mix α. The arrival process is generated by a MMPP(2) for three values of σ in equation (18). In each plot, κ in equation (19) increases from $\kappa = 1$, 2, 4 to 8. The buffer size $K = 10$ as well as the cell loss ratios $clr^*_L = 10^{-4}$ and $clr^*_H = 10^{-10}$ are the same for all curves

$$\sigma = \frac{\dfrac{q}{1-\alpha_1} + \dfrac{1-q}{1-\alpha_2}}{\dfrac{1}{1-\beta} + \dfrac{q}{1-\alpha_1} + \dfrac{1-q}{1-\alpha_2}} \tag{21}$$

$$\kappa = (1-\sigma)\left[\frac{q}{1-\alpha_1} + \frac{1-q}{1-\alpha_2}\right] \tag{22}$$

Alternatively, the parameter set $(\alpha_1, \alpha_2, \beta, \lambda)$ can be expressed in terms of $(\sigma, \kappa, R, \lambda)$ as

$$\beta = 1 - \frac{\sigma}{\kappa} \tag{23}$$

$$\frac{1}{1-\alpha_1} = \frac{1}{1-\sigma}\left[\kappa + \sqrt{\frac{1-q}{2q}(R-1)\kappa(\kappa+\sigma-1)}\right] \tag{24}$$

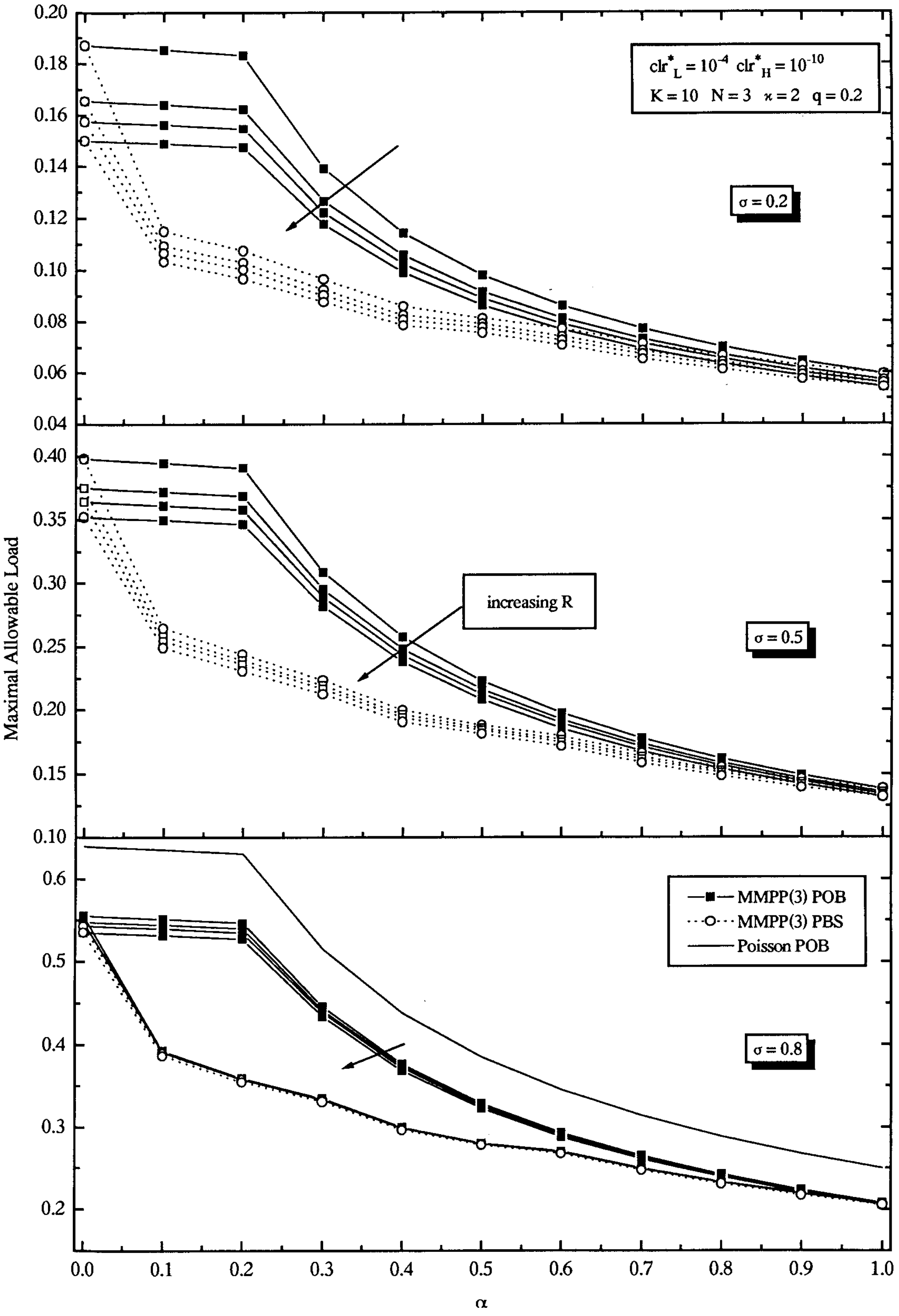

Figure 10. The maximum allowable load $\rho_{\max}$ in the RPOB and PBS versus the priority mix α. The arrival process is generated by a MMPP(3) for three values of σ in equation (21). In each plot, $\kappa = 2$ and $q = 0{\cdot}2$ are constant, whereas R increases as $R = 1$, 2, 3 and 5. The buffer size $K = 10$, as well as the cell loss ratios $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-10}$, are the same for all curves

$$\frac{1}{1-\alpha_2} = \frac{1}{1-\sigma}\left[\kappa - \sqrt{\frac{q}{2(1-q)}(R-1)\kappa(\kappa+\sigma-1)}\right] \tag{25}$$

For fixed values of σ, κ and q, the variance of the ON periods and, hence R, is bounded by

$$R - 1 < 2\,\frac{1-q}{q}\,\frac{\kappa+\sigma-1}{\kappa} \tag{26}$$

By choosing a value of q which is sufficiently small, equation (26) indicates that any value of R can be realized.

Figures 9 and 10 show the behaviour of $\rho_{\max}$ for RPOB versus α for various combinations of the parameters σ, κ and R for a relatively small buffer

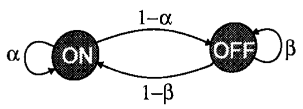

Figure 11. The Markov chain for $N=2$

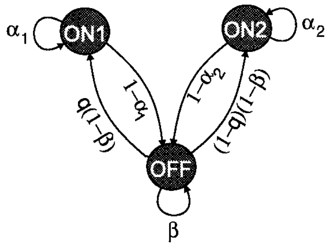

Figure 12. The Markov chain for $N=3$

$K=10$. As the shape is similar to that with pure Poisson arrivals, the results may hint that a MMPP(N) with Bernoulli distribution with parameter α for the priorities can be replaced by a corresponding Poisson process, however, with an adjusted parameter λ. In addition, the scaling rules in K proposed in Section 4.4 seem applicable. For the three-state model, the performance of PBS is also shown (Figure 10) clearly demonstrating a still higher superiority of RPOB as burstiness is involved.

5.3. *Conclusions on priorities and burstiness in the RPOB*

Our study found that the shape of the performance curve of the RPOB was less sensitive to the bursty character of the aggregate arrival process than to the priority distribution process. The compound Poisson arrival process with each arrival consisting of a packet of B cells with random priority assignment in a burst has a definitely different behaviour than that of a Poisson or MMP(N) process. For Markov chains with a larger number of states $N>3$ or with a cell emission process different from Poisson (e.g. state i always emits exactly a_i cells), we found an analogous behaviour as in the MMPP(2) or MMPP(3). The results seem to indicate that for increasing burstiness or correlation in the priority distribution (as in the compound Poisson process), the optimal performance is less influenced by priority information (a flatter behaviour of $\rho_{\max}$ versus α). On the other hand, as expected, the value of $\rho_{\max}$ for a given value of α is very sensitive to the details (e.g. burstiness) of the aggregate arrival process and a Poisson arrival law leads to the best performance.

6. CONCLUSIONS

The optimality of cell loss priority strategies for a single buffer under a general arrival law has been studied. The tight upper bounds found are useful to understand optimality in utilization of specific priority schemes, as illustrated for Poisson arrivals in the case of the POB and PBS.

Furthermore, this paper has focussed on the maximum allowable load $\rho_{\max}$ for RPOB and PBS versus the priority mix α for a wide variety of arrival processes. The priority distribution within bursts and the details of the aggregate arrival process are decisive quantities for the performance. The latter strongly influences (lowers) the value of $\rho_{\max}$ for a certain α, but hardly the shape of $\rho_{\max}$ versus α. The priority assignment within the aggregate cell flow is found to change the form of the $\rho_{\max}$ versus the α-curve.

In Section 3.2, it was shown that for large buffer sizes, CLPM schemes become useless. Hence, CLPM techniques are typically attractive for real-time services that require short buffers (approximately 100 cells). Currently, data services, fuelled by the Internet, are growing very fast, demanding large buffers of several 10,000 cells. Buffer engineering efforts are now concentrating more on intelligent packet discard strategies (tail packet or early packet discard) rather than CLPM techniques and on non-FIFO cell scheduling techniques[49–51] in order to guarantee end-to-end cell delay variation bounds and throughput fairness.

ACKNOWLEDGEMENTS

The authors are much indebted to J. David and are grateful to H. Kröner for the data in Figure 1. This work has been sponsored by the Flemish Institute for the Promotion of Scientific and Technological Research in Industry (IWT 950141).

APPENDIX A. APPROXIMATE EXPRESSION FOR $clr_A(\rho)$

The importance of equation (9) suggests a closer study of the cell loss ratio in a $M/D/1/K$ system in order to find an analytic approximation for the inverse f_K^{-1}. We propose a simple approximation for $clr_A(\rho)$ and the inverse f_K^{-1}. A 1 per cent accuracy is obtained for traffic intensities exceeding $0{\cdot}85<\rho<1$. The detailed derivation is presented elsewhere[52]. The approximative expression for the aggregate cell loss ratio reads

$$clr_A(\rho) \approx (1-\rho)\,\rho^{2K} \tag{27}$$

Our approximate equation (27) is always a lower bound. Equation (27) is particularly well suited to give fast estimates of the required number of buffer positions K given $\rho>0{\cdot}8$ and clr_A as $K=[(\log(clr_A)-\log(1-\rho))/2\log\rho]$ where $[x]$ denotes the integral part of x. The inverse function $f_K^{-1}(w)$ for equation (27) is $f_K^{-1}(w)=\lim_{n\to\infty} f_n(w)$ where

$$f_n(w) = \frac{w^{\frac{1}{2K}}}{[1-f_{n-1}(w)]^{\frac{1}{2K}}}$$

$$f_1(w) = w^{\frac{1}{2K}} \tag{28}$$

This continued fraction converges rapidly.

APPENDIX B. STATE EQUATIONS FOR THE RPOB

B.1. *The Markov modulated process*

A Markov modulated process with N states (MMP(N)) is characterized by the transition probability matrix $P_{N\times N}$ and the emission process in each state. Let $\mathcal{A}_i$ denote the stochastic variable describing the number of cells emitted in state i. The random variable $\mathcal{S}[k]$ denotes the state at discrete time k and the corresponding state vector $s[k] = [s_1[k]\ s_2[k] \dots s_N[k]]$ at discrete time k obeys

$$s[k+1] = s[k]\,.\,P \tag{29}$$

where $s[k].e^T = \sum_{i=1}^{N} 1.s_i[k] = 1$ with $e = [1\,1 \dots 1]$ and, hence $\sum_{j=1}^{N} P_{ij} = 1$ for all i. Written explicitly, we have

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1;N-1} & 1-\sum_{j=1}^{N-1} p_{1j} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2;N-1} & 1-\sum_{j=1}^{N-1} p_{2j} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3;N-1} & 1-\sum_{j=1}^{N-1} p_{3j} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ p_{N-1;1} & p_{N-1;2} & p_{N-1;3} & \cdots & p_{N-1;N-1} & 1-\sum_{j=1}^{N-1} p_{N-1;j} \\ p_{N;1} & p_{N;2} & p_{N;3} & \cdots & p_{N;N-1} & 1-\sum_{j=1}^{N-1} p_{N;j} \end{bmatrix} \tag{30}$$

The transition probability matrix P is composed with $N^2 - N$ probabilities $p_{ij} = \text{Prob}[\mathcal{S}[k+1] = j|\mathcal{S}[k] = i]$.

The general solution of equation (29) is

$$s[k] = s[0]\,.\,P^k \tag{31}$$

The steady state vector $\pi = \lim_{k\to\infty} s[k]$ follows from

$$\pi = \pi\,.\,P \tag{32}$$

with $\pi\,.\,e^T = 1$. Since the steady state vector does not depend on the initial state $s[0]$, it follows from equation (31) that the rows of $A = \lim_{k\to\infty} P^k$ should all be the same so that $\pi = [a_{11}\ a_{12} \cdots a_{1N}]$, because $\pi_j = \sum_{i=1}^{N} s_i[0]a_{ij} = a_{1j} \sum_{i=1}^{N} s_i[0] = a_{1j}$. Hence, $\lim_{k\to\infty} P^k = e^T\,.\,\pi$. On the other hand, equation (32) is an eigenvalue equation with eigenvalue $\lambda = 1$. Except for the trivial case where P is the identity matrix I, the solution of π is obtained from

$$\begin{bmatrix} p_{11}-1 & p_{21} & p_{31} & \cdots & p_{N-1;1} & p_{N1} \\ p_{12} & p_{22}-1 & p_{32} & \cdots & p_{N-1;2} & p_{N2} \\ p_{13} & p_{23} & p_{33}-1 & \cdots & p_{N-1:3} & p_{N3} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ p_{1;N-1} & p_{2;N-1} & p_{3;N-1} & \cdots & p_{N-1;N-1}-1 & p_{N;N-1} \\ 1 & 1 & 1 & \dots & 1 & 1 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \vdots \\ \pi_{N-1} \\ \pi_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \tag{33}$$

The steady state emitting process leads to a load λ,

$$\lambda = \pi\,.\,\Lambda\,.\,e^T = \sum_{i=1}^{N} \lambda_i\,\pi_i \tag{34}$$

where $\lambda_i = \text{E}[\mathcal{A}|\mathcal{S} = i] = \text{E}[\mathcal{A}_i]$ and $\Lambda = \text{diag}\{\text{E}[\mathcal{A}_1], \text{E}\{\mathcal{A}_2], \dots, \text{E}[\mathcal{A}_N]\}$. The distribution of the random variable $\mathcal{A}_k$ is characterized by the set of generating functions, for $(1 \le i \le N)$,

$$A_i(z) \overset{\triangle}{=} \text{E}[z^{\mathcal{A}_k}|\mathcal{S}_k = i] = \sum_{j=0}^{\infty} a_i(j)\, z^j \tag{35}$$

The number of high priority cells arriving during slot k is denoted by $\mathcal{A}_{kH}$. Since each arriving cell has a high priority with probability α, we have

$$\begin{aligned} \gamma_i(j) &\overset{\triangle}{=} \text{Prob}[\mathcal{A}_{kH} = j|\mathcal{A}_k = i] \\ &= \binom{i}{j} \alpha^j (1-\alpha)^{j-i} \end{aligned} \tag{36}$$

B.2. The RPOB

We define Q_k and $\mathcal{H}_k$ as the total number of buffer positions occupied and the number of high priority cells, respectively, evaluated at the beginning of time slot k. These random variables can never exceed $K-1$. The random service discipline, i.e. the random selection of a cell in the queue at each time slot, considerably lowers the state space dimension. The state space can be described by three random variables, Q_k, $\mathcal{H}_k$ and $\mathcal{S}_k$ that constitute a Markov process. We define two additional random variables $\mathcal{R}_k$ and $\mathcal{T}_k$ as the number of cells and the number of high priority cells, respectively, that leave the buffer at the end of time slot k. Both $\mathcal{R}_k$ and $\mathcal{T}_k$ are either zero or one. The deterministic server discipline means that

$$\text{Prob}[\mathcal{R}_k = 1 | Q_k + \mathcal{A}_k = n] = 1 - \delta(n) \tag{37}$$

where $\delta(n) = 1$ if $n = 0$ and zero elsewhere. Due to the random service order, each cell in the buffer has equal probability to be served. Hence,

$$\text{Prob}[\mathcal{T}_k = 1 | Q_k + \mathcal{A}_k = n, \mathcal{H}_k + \mathcal{A}_{kH} = m] = \begin{cases} \dfrac{\min(n, K)}{\min(m, K)} & n > 0 \\ 0 & n = 0 \end{cases} \tag{38}$$

The evolution in time of the number of cell in the buffer is governed by two system equations

$$Q_{k+1} = \min\,[Q_k + \mathcal{A}_k, K] - \mathcal{R}_k \tag{39}$$

$$\mathcal{H}_{k+1} = \min\,[\mathcal{H}_k + \mathcal{A}_{kH}, K] - \mathcal{T}_k \tag{40}$$

Equation (40) indicates that high priority cells will be accepted by possibly pushing low priority cells as long as $\mathcal{H}_k + \mathcal{A}_{kH} \leq K$. Whenever $\mathcal{H}_k + \mathcal{A}_{kH} > K$, high priority cells are lost. We concentrate on the steady state and define

$$q(n, m, l) = \lim_{k \to \infty} \text{Prob}[Q_{k+1} = n \wedge \mathcal{H}_{k+1} = m \wedge \mathcal{S}_{k+1} = l] \tag{41}$$

The state equations are determined as follows. For the empty state, we have,

$$q(0, 0, l) = \sum_{t=1}^{N} p_{tl} a_t(0)[q(1, 0, t) + q(1, 1, t)] + [a_t(0) + a_t(1)]q(0, 0, t) \tag{43}$$

This equation is the same for a conventional POB. Furthermore, as long as the occupied buffer space is less than $K-1$, all arriving cells are accepted. Hence, from the system equations we derive the transition probabilities satisfying, for $0 < n < K-1$,

$$\begin{aligned} &\text{Prob}[Q_{k+1} = n, \mathcal{H}_{k+1} = m, \ \mathcal{S}_{k+1} = l | Q_k = i, \mathcal{H}_k = j, \mathcal{S}_k = t] \\ &= \text{Prob}[\mathcal{A}_{kH} - \mathcal{T}_k = m - j | Q_k = i, \ \mathcal{H}_k = j, \mathcal{A}_k = n + 1 - i]a_t(n + 1 - i) \\ &= \text{Prob}[\mathcal{T}_k = 1 | Q_k + \mathcal{A}_k = n + 1, \ \mathcal{H}_k + \mathcal{A}_{kH} = m + 1]\gamma_{n+1-i}(m + 1 - j) \\ &+ \text{Prob}[\mathcal{T}_k = 0 | Q_k + \mathcal{A}_k = n + 1, \ \mathcal{H}_k + \mathcal{A}_{kH} = m] \\ &\gamma_{n+1-i}(m - j)]a_t(n + 1 - i) \end{aligned}$$

defining $(x)^+ = \max(x, 0)$, we thus find the following state equations

$$\begin{aligned} q(n, m, l) = \sum_{t=1}^{N} p_{tl} \sum_{i=0}^{n+1} \Bigg[&\sum_{j=(m-n+i)^+}^{\min(i, m+1)} \frac{m+1}{n+1} \binom{n+1-i}{m+1-j} \alpha^{m+1-j}(1-\alpha)^{n+j-m-i} q(i, j, t) \\ &+ \sum_{j=(m-n+i-1)^+}^{\min(i, m)} \frac{n-m+1}{n+1} \binom{n+1-i}{m-j} \alpha^{m-j}(1-\alpha)^{n+j+1-m-i} q(i, j, t) \Bigg] \times a(n+1-i, t) \end{aligned} \tag{44}$$

The push-out mechanism is only active if the buffer is entirely filled. Since we describe the process at the beginning of a new time slot while the server just acts at the end of a time slot k, we only have a full buffer if $Q_{k+1} = K - 1$. In this case, low priority cells may occur, whereas high priority cells are only lost if $\mathcal{H}_{k+1} = K - 1$. Therefore, as before, the state equations for the transitions $q(K-1, m, l)$ and $0 \leq m < K - 1$ are

$$q(K-1, m, l) = \sum_{t=1}^{N} p_{tl} \sum_{n=K}^{\infty} \sum_{i=0}^{K-1} \left[\sum_{j=(m-n+i+1)^+}^{\min(i, m+1)} \frac{m+1}{K} \binom{n-i}{m+1-j} \alpha^{m+1-j} (1-\alpha)^{n+j-m-i-1} \times q(i, j, t) \right.$$
$$\left. + \sum_{j=(m-n+i, 0)^+}^{\min(i, m)} \frac{K-m}{K} \binom{n-i}{m-j} \alpha^{m-j} (1-\alpha)^{n+j-m-i} q(i, j, t) \right] \times a(n-i, t) \qquad (45)$$

Finally, the last block row, describing those situations where high priority cells are lost, is given by

$$q(K-1, K-1, l) = \sum_{t=1}^{N} p_{tl} \sum_{n=K}^{\infty} \sum_{m=K}^{n} \sum_{i=0}^{K-1} \sum_{j=(m-n+i)^+}^{i} \binom{n-i}{m-j} \alpha^{m-j} (1-\alpha)^{n+j-m-i} q(i, j, t) a(n-i, t) \qquad (46)$$

The normalization conditions yields the final equation,

$$\sum_{t=1}^{N} \sum_{n=0}^{K-1} \sum_{m=0}^{n} q(n, m, t) = 1 \qquad (47)$$

and reveals that we have a total of $s = NK(K+1)/2$ states. Equations (43), (44), (45) and (46) are sufficient to determine the steady-state probabilities $q(n, m, l)$.

The cell loss ratios have been computed from the mean number of cells rejected due to buffer overflow. This leads to

$$clr_A(\alpha) = \frac{1}{\mathrm{E}[\mathcal{A}]} \sum_{l=1}^{N} \sum_{n=0}^{K-1} \sum_{m=0}^{n} q(n, m, l) \sum_{k=K+1-n}^{\infty} (k+n-K) a_l(k) \qquad (48)$$

$$clr_H(\alpha) = \frac{1}{\alpha\, \mathrm{E}[\mathcal{A}]} \sum_{l=1}^{N} \sum_{n=0}^{K-1} \sum_{m=0}^{n} q(n, m, l) \sum_{k=K+1-n}^{\infty} a_l(k) \sum_{t=K+1-m}^{k} (m+t-K) \gamma_k(t) \qquad (49)$$

Losses only occur if two or more cells arrive, because at the beginning of a slot time, position K in the buffer is always free. Both the cell loss ratio of the high priorities and the aggregate cell loss has been determined in this way. The cell loss ratio of the low priorities is found using $clr_L = \frac{1}{1-\alpha}\,(clr_A(\alpha) - \alpha clr_H(\alpha))$.

B.3. *The compound Poisson process*

We merely give the state equations for the case where all cells in a burst have the same priority. The derivation for the case where the priority of cells is randomly distributed with probability α follows from the previous section where $N = 1$ and $a_1(k) = e^{-\lambda}\, \lambda^{[k/B]}/[k/B]!$

Assuming that $B > 1$ and defining

$$q(n, m) = \lim_{k \to \infty} \mathrm{Prob}[Q_{k+1} = n \wedge \mathcal{H}_{k+1} = m] \qquad (50)$$

$$a(j) = \mathrm{Prob}[\mathcal{A} = j] = e^{-\lambda} \frac{\lambda^j}{j!} \qquad (51)$$

$$\begin{aligned} x^{@} &= \left[\frac{x}{B}\right] + 1 && \left(\frac{x}{B} : \text{not integer}\right) \\ &= \left[\frac{x}{B}\right] && \left(\frac{x}{B} : \text{integer}\right) \\ &= 0 && (x < 0) \end{aligned} \qquad (52)$$

we have

$$q(0, 0) = a(0)[q(0, 0) + q(1, 0) + q(1, 1)] \qquad (53)$$

and for $0 < n < K - 1$,

$$q(n,m)=\sum_{i=0}^{\left[\frac{\mathrm{n+1}}{\mathrm{B}}\right]}\left[\sum_{j=(Bi-n+m)^{@}}^{\left(i,\left[\frac{\mathrm{m+1}}{\mathrm{B}}\right]\right)^{-}}\frac{m+1}{n+1}\binom{i}{j}\alpha^{j}(1-\alpha)^{i-j}q(n+1-Bi,m+1-Bj)\right.$$
$$\left.+\sum_{j=(Bi-n+m-1)^{@}}^{\left(i,\left[\frac{\mathrm{m}}{\mathrm{B}}\right]\right)^{-}}\frac{n-m+1}{n+1}\binom{i}{j}\alpha^{j}(1-\alpha)^{i-j}q(n+1-Bi,m-Bj)\right]a(i) \tag{54}$$

whereas for $n=K-1$ and $0\leq m<K-1$,

$$q(K-1,m)=\sum_{n=K}^{\infty}\sum_{i=(n-K+1)^{@}}^{\left[\frac{\mathrm{n}}{\mathrm{B}}\right]}\left[\sum_{j=(Bi-n+m+1)^{@}}^{\left(i,\left[\frac{\mathrm{m+1}}{\mathrm{B}}\right]\right)^{-}}\frac{m+1}{K}\binom{i}{j}\alpha^{j}(1-\alpha)^{i-j}q(n-Bi,m+1-Bj)\right.$$
$$\left.+\sum_{j=(Bi-n+m)^{@}}^{\left(i,\left[\frac{\mathrm{m}}{\mathrm{B}}\right]\right)^{-}}\frac{K-m}{K}\binom{i}{j}\alpha^{j}(1-\alpha)^{i-j}q(n-Bi,m-Bj)\right]a(i) \tag{55}$$

The last equation for $q(K-1,K-1)$ is replaced by the normalization condition. The relations for the cell loss ratio are

$$\lambda B\,clr_{A}=\sum_{n=0}^{K-1}\sum_{m=0}^{n}q(n,m)\sum_{j=(K-n)^{@}}^{\infty}(Bj+n-K)\,a(j) \tag{56}$$

$$\alpha\lambda B\,clr_{H}=\sum_{n=0}^{K-1}\sum_{m=0}^{n}q(n,m)\sum_{j=(K-m)^{@}}^{\infty}(Bj+m-K)\,a_{H}(j) \tag{57}$$

where $a_H(j)=e^{-\alpha\lambda}\ (\alpha\lambda)^{j}\ j!$. These relations can be simplified when we introduce $q_H(m)$ as the occupancy probability that there are precisely m high priority cells and similarly $q_A(m)$ as the occupancy probability for an aggregate of precisely m cells. Clearly, we have

$$q_{H}(m)=\sum_{n=m}^{K-1}q(n,m) \tag{58}$$

$$q_{A}(m)=\sum_{m=0}^{n}q(n,m) \tag{59}$$

We can immediately substitute equation (59) in (56) whereas the same holds for equation (58) in (57) after reversing the n and m summations. The result is

$$\lambda B\,clr_{A}=\sum_{n=0}^{K-1}q_{A}(n)\sum_{j=(K-n)^{@}}^{\infty}(B_{j}+n-K)\,a(j) \tag{60}$$

$$\alpha\lambda B\,clr_{H}=\sum_{n=0}^{K-1}q_{H}(n)\sum_{j=(K-n)^{@}}^{\infty}(B_{j}+n-K)\,a_{H}(j) \tag{61}$$

REFERENCES

1. M. de Prycker, *Asynchronous Transfer Mode: Solution for Broadband ISDN*. Ellis Horwood, New York, 3rd edn. 1995.
2. H. Saito, *Teletraffic Technologies in ATM Networks*. Artech House, Boston, 1994.
3. H. Kröner, G. Hebuterne, P. Boyer and A. Gravey, 'Priority management in ATM switching nodes'. *IEEE J. Select. Areas Commun.*, **9**, (3), 418–427 (1991).
4. I. F. Akyildiz and X. Cheng, 'Analysis of a finite buffer queue with different scheduling and push-out schemes'. *Performance Eval.*, **19**, 317–340 (1994).
5. Y. Z. Cho and C. K. Un, 'Analysis of the M/G/1 queue under a combined preemptive/non-preemptive priority discipline'. *IEEE Trans. Commun.*, **41**, (1), 132–141 (1993).
6. Z. Dziong, K.-Q. Liao and L. Mason, 'Effective bandwidth allocation and buffer dimensioning in ATM-based networks with priorities'. *Comp. Networks ISDN Systems*, **25**, 1065–1078 (1993).
7. G. Gallassi, G. Rigolio and L. Fratta, 'Bandwidth assignment in prioritized ATM networks'. *Proc. IEEE Globecom'90*, **505**, (2), 852–856 (1990).
8. T.-Y. Huang, J.-L. C. Wu and J. Wu, 'Priority management to improve the QOS in ATM networks'. *IEICE Trans. Commun.*, **E 76-B**, (3), 249–257 (1993).
9. M. Libby and H. Hughes, 'Priority management for selective cell scheduling and discarding in ATM networks', *Proc. Sum. Computer Simulation Conf. (SCSC'93)*, Boston, USA, pp. 572–577 July 1993.
10. A. Y.-M. Lin and J. A. Silvester, 'Priority queueing strategies

and buffer allocation protocols for traffic control at an ATM integrated broadband switching system'. *IEEE J. Select. Areas Commun.*, **9**, (9), 1524–1536 (1991).

11. A. Y.-M. Lin and J. A. Silvester, 'Priority queueing strategies for traffic control at a multichannel ATM switching system'. *Proc. IEEE Globecom '91*, **8B**, (2), 234–238 (1991).
12. H. Ohta and T. Kitami, 'Simulation study of the cell discard process and the effect of cell loss compensation in ATM networks'. *Trans. IEICE*, **E 73**, (10), 1704–1711 (1990).
13. C. Shim, J. G. Kim, M. Park and S. B. Lee, 'Queueing analysis for ATM switching of mixed traffic with priority'. *IEICE Trans.* **E 74**, (10), 3086–3091 (1991).
14. A. K. Choudhury and E. L. Hahne, 'Space priority management in a shared memory ATM switch'. *Proc. IEEE Globecom'93*, 1375–1383, December 1993.
15. E. J. Hernandez-Valencia and F. G. Bonomi, 'Simulation of a simple loss/delay priority scheme for shared memory ATM fabrics'. *Proc. IEEE globecom'93*, pp. 1389–1394, December 1993.
16. M. Krunz, H. Hughes and Y. Parviz, 'Congestion control in ATM networks using multiple buffers and priority mechanisms'. *Proc. of Sum. Computer Simulation Conf. (SCSC'93)*, Boston, USA, pp. 566–571, July 1993.
17. J. F. Meyer, S. Montagna and R. Paglio, 'Dimensioning of an ATM switch with shared buffer and threshold priority'. *Computer Networks and ISDN Systems*, **26**, 95–108 (1993).
18. K. Rothermel, 'Priority mechanisms in ATM networks'. *Proc. IEEE Globecom'90*, **505**, (1), 847–851 (1990).
19. P. Van Mieghem, J. David and G. H. Petit, 'Performance of cell loss priority management schemes in shared buffers with poisson arrivals'. *Eur. Trans. Telecommun.*, to be published, (1997).
20. J. Kurose, 'Open issues and challenges in providing quality of service guarantees in high-speed networks'. *ACM SIGCOM, Computer Communication Review*, **23**, (1), 6–15 (1993).
21. M. Kwag, S. Seong and C. Kim, 'An analysis of cell loss process in an ATM network under partial buffer sharing policy'. *Proc. 1994 IEEE Region 10's, Ninth Annual International Conference; Singapore 22–26 August*, edited by T. Chan, **1**, 456–460 (1994).
22. I. Cidon, A. Khamisy, L. Georgiadis and R. Guerin, 'Optimal buffer sharing'. *IEEE INFOCOM'95*, Boston, **1**, (1a.4.1), 24–31 (1995).
23. D. M. Bertsekas, *Dynamic Programming, Deterministic and Stochastic Models*. Prentice-Hall, 1987.
24. A. Elwalid and D. Mitra, 'Analysis, approximations and admission control of a multi-service multiplexing system with priorities'. *IEEE INFOCOM'95*, Boston, **2**, (4b.4.1), 463–472 (1995).
25. ITU-T Study Group 13, 'Traffic control and congestion control in B-ISDN'. *Recommendation I.371*, Geneva, p. 1, May 1996.
26. R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Prentice-Hall International Editions, New York, 1989.
27. J. Garcia and O. Casals, 'Performance evaluation of source dependent congestion control procedures in ATM networks'. *Proc. 1991 Singapore Int. Conf. Networks*, pp. 178–182, September 1991.
28. T.-C. Hou and A. K. Wong, 'Queueing analysis for ATM switching of mixed continuous-bit-rate and bursty traffic'. *INFOCOM'90*, **2**, 660–667 (1990).
29. K.-Q. Liao, Queueing analysis of partial buffer sharing with Markov modulated Poisson inputs. *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks: ITC 14*, **1a**, 55–64 (1994).
30. N. M. Mitrou and D. E. Pendarakis, 'Cell-level statistical multiplexing in ATM networks: Analysis, dimensioning and call acceptance control w.r.t. QOS criteria'. *Queueing Performance and Control in ATM (ITC13)*, pp. 7–12, 1991.
31. V. Kulkarni, L. Gun and P. Chimento, 'Effective bandwidth vector for two-priority ATM traffic'. *IEEE INFOCOM'94: Networking for Global Communications*, Toronto, **3**, 1056–1064 (1994).
32. H. Saito, 'Queueing analysis of cell loss probability control in ATM networks'. *ITC-13*, pp. 19–24, 1991.
33. N. L. S. Fonseca and J. A. Silvester, 'Estimating the loss probability in a multiplexer loaded with multi-priority MMPP streams'. *ICC'93*, **2**, 1037–1041 (1993).
34. C. G. Chang and H. H. Tan, 'Queueing analysis of explicit policy assignment push-out buffer sharing schemes for ATM networks'. *IEEE INFOCOM'94: Networking for Global Communications*, Toronto, **2**, 500–509 (1994).
35. V. Bemmel and M. Ilyas, 'A novel congestion control strategy in ATM networks'. *Comput. Ind. Engng.*, **25**, (1–4), 549–552 (1993).
36. D. W. Petr and V. S. Frost, 'Optimized threshold-based discarding for queue overload control'. *Int. J. Digital Analog Commun. Syst.*, **5**, 2, 97–116 (1992).
37. P. Yegani, M. Krunz and H. Hughes, 'Congestion control schemes in prioritized ATM networks'. *IEEE Supercomm/ICC'94: Serving humanity through communications*, New Orleans, **2**, 1169–1173 (1994).
38. H. J. Chao and H. Cheng, 'A new QOS-guaranteed cell discarding strategy: self-calibrating pushout'. *IEEE Globecom '94: The Global Bridge*, San Francisco, **2**, 929–934 (1994).
39. J. F. Ren, J. W. Mark and J. W. Wong, 'A dynamic priority queueing approach to traffic regulation and scheduling in B-ISDN', *IEEE, Globecom'94: The Global Bridge* (San Francisco), **1**, 612–618 (1994).
40. Y. Jun and S. Cheng, 'A novel priority queue scheme for handoff procedure'. *IEEE Supercomm/ICC'94: Serving humanity through communications*, New Orleans, **1**, 182–186 (1994).
41. S. Suri, D. Tipper and G. Meempat, 'A comparative evaluation of space priority strategies in ATM networks'. *IEEE INFOCOM'94: Networking for Global Communications*, Toronto, **2**, 516–523 (1994).
42. Y.-H. Jeon and I. Viniotis, 'Achievability of combined GOS requirements in broadband networks'. *IEEE '93 Conference*, pp. 192–196, 1993.
43. A. Dailianas and A. Bovopoulos, 'Real-time admission control algorithms with delay and loss guarantees in ATM networks'. *IEEE INFOCOM'94: Networking for Global Communications*, Toronto, **3**, 1065–1072 (1994).
44. T.-Y. Huang and J.-L. C. Wu, 'Performance analysis of ATM switches using priority schemes'. *IEE Proc. Commun.*, **141**, (4), 248–254 (1994).
45. J.-L. C. Wu and T.-Y. Huang, 'Dynamic priority schemes to improve the QOS in ATM networks'. *IEEE Globecom'94: The Global Bridge* (San Francisco), **2**, 1195–1206 (1994).
46. L. Georgiadis, R. Guerin and A. Parekh, 'Optimal multiplexing on a single link: delay and buffer requirements'. *IEEE INFOCOM'94: Networking for Global Communications*, Toronto, **2**, 524–532 (1994).
47. L. Kleinrock, *Queueing Systems*, vol. 1: Theory. John Wiley and Sons, NY, 1975.
48. R. Syski, *Introduction to Congestion Theory in Telephone Systems*, vol. 4 of *Studies in Telecommunication*. North-Holland, Amsterdam, 2nd edn, 1986.
49. S. J. Golestani, 'A self-clocked fair queueing scheme for broadband applications'. *IEEE INFOCOM'94: Networking for Global Communications*, Toronto, **2**, 636–646 (1994).
50. A. K. Parekh and R. G. Gallager, 'A generalized processor sharing approach to flow control in integrated services networks: the single-node case'. *IEEE/ACM Trans. Networking*, **1**, (3), 344–357 (1993).
51. J. W. Roberts, 'Virtual spacing for flexible traffic control'. *Int. J. Commun. Syst.*, **7**, 307–318 (1994).
52. P. Van Mieghem, 'The asymptotic behaviour of queueing systems: large deviations theory and dominant pole approximation'. *Queueing Systems*, **23**, 27–55 (1996).

Authors' biographies

Piet Van Mieghem obtained the MS and PhD degree in electrical engineering from the Katholieke Universiteit Leuven (KUL), Belgium, in 1987 and 1991, respectively. In 1987, he joined the Interuniversity Micro Electronics Center (IMEC), Leuven. During 1993, he was a visiting scientist at the Massachusetts Institute of Technology (MIT), USA. From the end of 1993, he has been working in Alcatel Telecom's Research Center in Antwerp, where he is engaged in traffic and network aspects of the current two major network architectures, being ATM and IP. Dr Van Mieghem was awarded the Student Prize in 1987 for the best Belgian Master's thesis in electrical engineering by the Belgian Institute for Control and Automatization (BIRA).

Bart Steyaert was born in Roeselare, Belgium, in 1964. He received the degrees of Licentiate in physics and in computer science in 1987 and 1989, respectively, from the University of Ghent, Gent, Belgium. Since 1990, he has been working as a PhD student at the Laboratory for Communications Engineering, University of Ghent, as a member of the SMACS Research Group. His main research interests include performance evaluation of discrete-time queueing models, applicable in B-ISDN networks.

Guido H. Petit graduated from the University and the Industrial High School of Antwerp, Belgium, with degrees in chemistry and electronic engineering, in 1977 and 1980, respectively. In 1980, he joined Alcatel Bell Telephone, Antwerp, as a traffic engineer and he was awarded a PhD in crystallography from the University of Antwerp in 1984. He is currently Traffic Technologies Manager of the Network Planning Group at the Research Division of Alcatel Telecom in Antwerp. Since 1993, he has been part-time lecturer in a Postgraduate Program on telecommunications and telematics at the University of Antwerp, where he teaches teletraffic management and engineering of B-ISDN networks. His current research interests include performance modelling and analysis of ATM switching systems and of traffic, flow and congestion control algorithms. From 1992 to 1995 he was actively participating in the traffic control and resource management working group of ETSI-NA5. He is a holder of several European patents and has authored and co-authored more than 50 scientific papers. He also acted as Guest Editor in the IEEE Communications Magazine. Dr Petit is an IEEE member.