Throughput Optimality of Single Queue Priority Schemes.

Piet Van Mieghem and Guido H. Petit
Alcatel
Research Division
Francis Wellesplein 1, B-2018 Antwerp, Belgium
pvmi@rc.bel.alcatel.be and petitg@btmaa.bel.alcatel.be

Bart Steyaert
University of Gent
Lab. for Communications Engineering
Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
bart.steyaert@lci.rug.ac.be

Abstract

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 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 ρ_{max} versus the priority mix α for a PBS and a random push-out buffer (R POB) 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 modeled by a MMPP(2), MMPP(3), and higher order Markov Modulated Processes closely follow the behaviour of the maximum allowable load in the R POB 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.

1. Introduction.

This work focuses on connection admission control (CAC) [2, 7] of a single buffer with a two-type (high and low) priority management [5]. 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 ρ_{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 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, 5, 6] or in a shared buffer for which we further refer to our work [10]. 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.

Among buffer management protocols [5, 6], the push-out buffer (POB) and the partial buffer sharing (PBS) are most well-known. Although these priority schemes have been studied in the literature [5], 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[1] 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

[1] This push-out discipline is FIFO. Other alternatives are LIFO and random push-out.

0-8186-7852-6/97 $10.00 © 1997 IEEE

sequence integrity. A PBS mechanism is somewhat simpler: if the buffer occupancy is below a threshold T, both low and high priority cells are allowed to enter, otherwise only high priority cells are accepted until complete buffer occupation.

The outline is as follows. In Section 2, we investigate the throughput optimality of a priority system in a single buffer and derive two upper bounds. In Section 3, we introduce the R POB and compare for Poisson arrivals the performance of partial buffer sharing to that of the push-out scheme. The main advantage of introducing the R POB is that, first, it serves as an excellent approximation for the conventional FIFO push-out, and second, it allows us to perform exact calculations of the maximum allowable load for very general arrival laws. In the last Section 4, we introduce burstiness in the arrival pattern for the R POB: 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 R POB and PBS are compared for an MMP(3). A literature overview and the detailed derivation of the state equations for the R POB with MMP(N) cell arrivals are presented elsewhere [11].

2. General relations.

2.1. Definitions.

By virtue of the slotted nature of ATM, we concentrate on discrete-time systems where the servers work deterministically. The time unit, further 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

$$\mu_A = \mu_H + \mu_L = 1 \qquad (1)$$

where the subscripts refer to the aggregate (A), the low (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

$$\lambda_A = \lambda_H + \lambda_L \qquad (2)$$

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

Since the system has a finite capacity of K queueing positions with an additional one for the server, in general cell loss will occur. We denote the cell loss ratio *clr* as the mean number of cells lost per time slot over the mean number of arrived cells of that type. Again the total number of lost cells consists of both priorities. From this fact we deduce a useful relation[2],

$$\begin{aligned} \lambda_A\, clr_A &= \lambda_L\, clr_L + \lambda_H\, clr_H \\ clr_A(\alpha) &= (1-\alpha)\, clr_L(\alpha) + \alpha\, clr_H(\alpha) \end{aligned} \qquad (4)$$

The last relation explicitly expresses the dependence on $\alpha = \frac{\lambda_H}{\lambda_A}$). In addition, since we can write the aggregate cell loss ratio as a weighted mean, $clr_A = \frac{\lambda_L\, clr_L + \lambda_H\, clr_H}{\lambda_L + \lambda_H}$, 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, $\widehat{clr_A}$, in the corresponding system without the priority management is exactly described by the loss probability of that corresponding G/D/1/K system. Formally, fixing all other traffic descriptors independent of the load ρ_A, we have

$$\widehat{clr_A} = f_K(\widehat{\rho_A}) \qquad (5)$$

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 a same aggregate cell loss ratio requirement $\widehat{clr_A} = clr_A(\alpha) = clr_A^*$

$$\widehat{\rho_A} \ge \rho(\alpha) \qquad (7)$$

2.2. 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.

Since $f_K(x)$ is monotonously increasing, the inverse function exists justifying to rewrite (5) as $\widehat{\rho_A} = f_K^{-1}(\widehat{clr_A})$. Further, the inverse function $g^{-1}(x)$ of an increasing function $g(x)$ is increasing. Using (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 (4) and the requirements on the cell loss ratios, we have

$$clr_A(\alpha) \le (1-\alpha)\, clr_L^* + \alpha\, clr_H^* \qquad (8)$$

[2] 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 (8) is decreasing in α due to the fact that $clr_H^* < clr_L^*$, so is (9). The upper bound (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 (9) does not hold for all α emphasizing that (9) forms an unattainable upper bound.

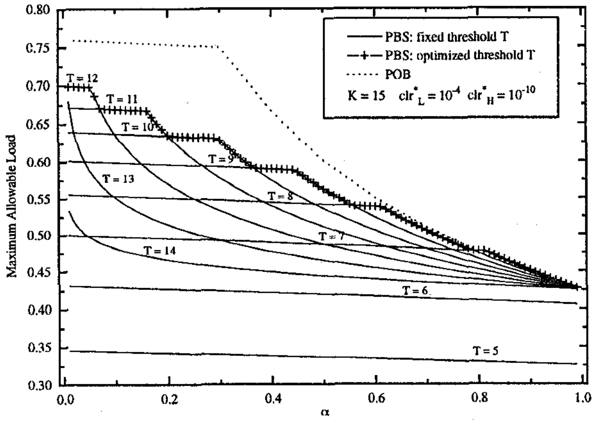

Figure 1. The effect of the threshold T on the performance of PBS in a relatively small buffer of size $K = 15$ for the cell loss ratio couple $(10^{-4}, 10^{-10})$. For comparison purposes, also the performance of the POB is shown in dotted line.

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 $\rho_A(\alpha) \leq \frac{\rho_A(0)}{1-\alpha}$ is always fulfilled by (9) since the left hand side is decreasing in α while the right hand side increases in α. The inequality (10) poses a lower upper bound than (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 (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 (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 and precisely met by a PBS scheme with threshold $T = 0$.

In conclusion, the maximum allowable load ρ_{max} is bounded for low α by (9) and for high α by (10). The upper bounds (10) and (9) coincide at $\alpha = 1$, but have opposite curvatures for $\alpha \leq 1$. In addition around $\alpha \leq 1$ the bound (10) is smaller than (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.

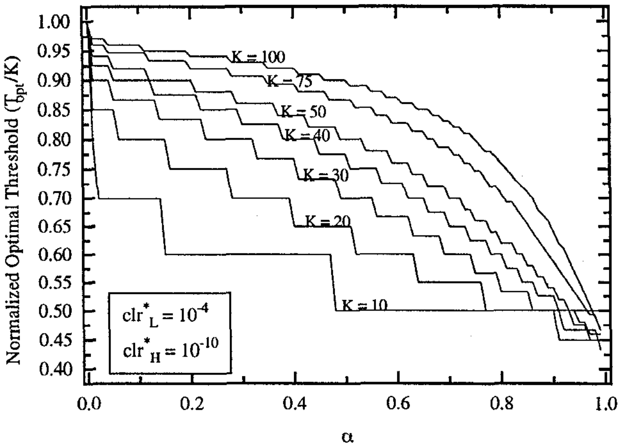

Figure 2. The normalized optimal threshold $\frac{T_{opt}}{K}$ in PBS for various buffer sizes K but fixed cell loss ratio couple $(10^{-4}, 10^{-10})$.

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 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.

3. Poisson arrivals.

This Section compares two standard priority schemes, the push-out buffer (POB) and partial buffer sharing (PBS) for *Poisson arrivals*. The emphasis lies on a new introduced model, the R POB, that is further studied under bursty arrival processes in the next Section.

3.1. Partial Buffer Sharing (PBS).

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 aggregrate load using the discrete-time version of the model of Kröner *et al.* [5]. The effect of the threshold T on the performance is illustrated in Fig. 1. 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 that of the POB. The desired maximum allowable load is the maximum envelope of all these curves and is a concatentation of regions alternatingly dominated by the high and low priority cell loss requirement.

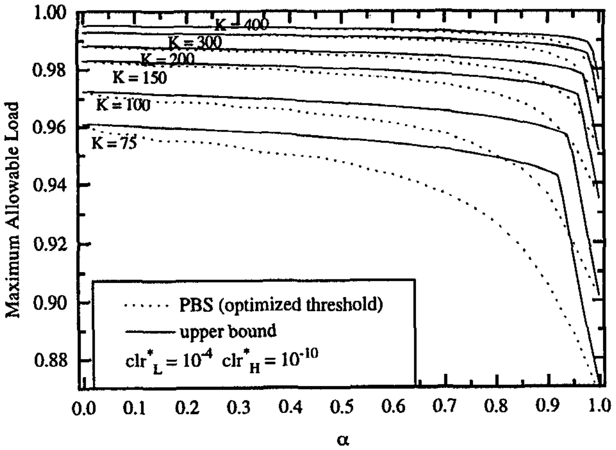

Figure 3. The maximum allowable load in PBS with optimized threshold versus α and the minimum of the upper bounds (10) and (9) for various large buffer sizes K but fixed cell loss ratio couple $(10^{-4}, 10^{-10})$.

The normalized optimal theshold $\frac{T_{opt}}{K}$ versus α is shown in Fig. 2 for various K values. Together with Fig. 1, the plot illustrates that, due to the integer character of T, analytic optimization is hardly feasible for small K. The longer the buffer size K, the more integer values of T there are available resulting in a smoother maximum allowable curve. Fig. 3 plots the maximum allowable load $\rho_{max}(\alpha)$ versus α for large values of K and the minimum of the upperbounds (10) and (9). This graph clearly demonstrates how closely 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.

3.2. The Push-Out Buffer (POB).

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 $clr_H^* << clr_L^*$, we have from (4) approximately that $clr_A(\alpha) \approx clr_L^*(1-\alpha)$. Invoking (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 ($clr_L(\alpha) = clr_L^*$), while 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).

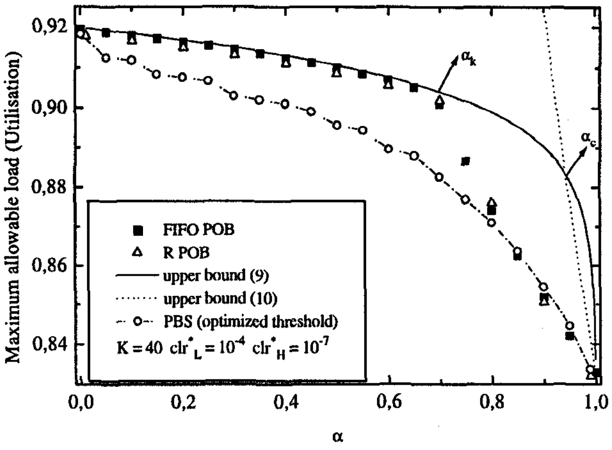

Figure 4. The maximum allowable load versus the priority mix α for a FIFO POB [5] and a R POB of size $K = 40$ for the cell loss requirements $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-7}$. We have also drawn the both upper bounds (9) and (10).

The situation is more complex for high values of α. For sufficienly high α, $\rho_{max}(\alpha)$ follows from (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 $clr_H(\alpha) = clr_H^*$ while $clr_L(\alpha) < clr_L^*$. However, since $clr_H^* << clr_L^*$, 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 $clr_L(\alpha) \approx clr_{Lpo}(\alpha)$. The calculation of the push-out probability is exceedingly complicated and we 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 [5]) and a R POB. The delimiter refers to the service discipline. Thus, R (random) means that all cells available have equal probability to be served as opposed to FIFO where always the cell in the position nearest to the server (or with the longest waiting time) is removed from the queue. Clearly, the R POB does not obey the sequence integrity. However, as the cell loss ratio only weakly depends on the sequence order, the maximum allowable load of the R POB is expected to closely approach that of the FIFO POB, provided the cell loss ratio requirements are sufficiently stringent ($clr^* < 0.1$). Indeed, for both POB types and for

Poisson arrivals[3] the comparison in the maximum allowable load $\rho_{max}(\alpha)$ versus α shows as illustrated in Fig. 4 that both priority management systems exhibit very similar performances for ρ_{max}.

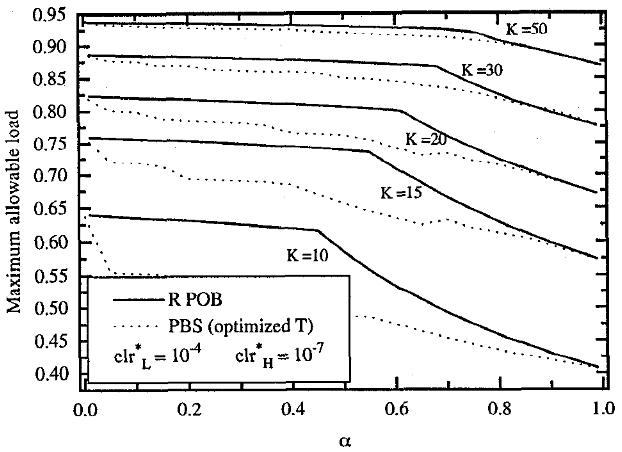

Figure 5. Calculation of the maximum allowable load ρ_{max} for various buffer sizes K versus the priority mix α for the cell loss requirements $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-7}$. The curves are obtained for the R POB

3.3. POB versus PBS.

In Fig. 5 and 6, we present $\rho_{max}(\alpha)$ for the R POB 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 (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[1]. But, once the priority mix $\alpha \leq \alpha_k$, the POB approaches the upper bound (9) and is undoubtedly the better strategy.

As an overall conclusion, the POB offers a better treatment of low priorities, while PBS can be engineered (by properly 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.* [8] 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.

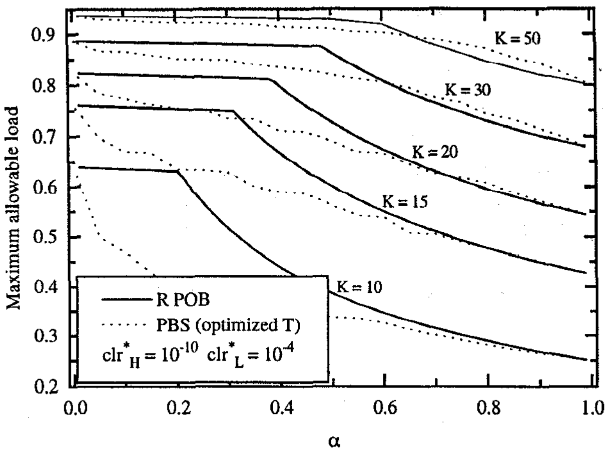

Figure 6. Calculation of the maximum allowable load ρ_{max} for various buffer sizes K versus the priority mix α for the cell loss requirements $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-10}$. The curves are obtained for the R POB

3.4. R POB fit for $\rho_{max}(\alpha)$.

Since $\rho_{max}(\alpha)$ of a R POB in the $[0, \alpha_k]$ interval is sufficiently close approximated by (9) as illustrated in Fig. 4, our objective is to find an estimate in $[\alpha_k, 1]$ accurate to within 1%.

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 R POB is well fitted by

$$\rho_{max}(\alpha) = p_1 + \frac{p_2}{(\alpha + p)^2} \tag{12}$$

Introducing the additional information

$$\begin{aligned} \rho_{max}(1) &= f_K^{-1}(clr_H^*) \\ \rho_{max}(\alpha_k) &= f_K^{-1}((1-\alpha_k)\, clr_L^* + \alpha_k\, clr_H^*) \end{aligned}$$

the equation (12) can be specified as

$$\begin{aligned} \rho_{max}(\alpha) &= \frac{1}{P}\left[\rho_{max}(1)\left(\frac{1}{(\alpha+p)^2} - \frac{1}{(\alpha_k+p)^2}\right) + \right. \\ & \left. \rho_{max}(\alpha_k)\left(\frac{1}{(1+p)^2} - \frac{1}{(\alpha+p)^2}\right)\right] \end{aligned} \tag{13}$$

where $P = \frac{1}{(1+p)^2} - \frac{1}{(\alpha_k+p)^2}$. An elegant approximation for $f_K^{-1}(x)$ in a discrete-time M/D/1/K is found in [11, 9].

The proposed fit (13) 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 $0.5 \leq p \leq 1$. The result is not very sensitive to variations in p (in contrast to α_k) when aiming at an accuracy of 1%. The remainder is therefore devoted to the study of α_k.

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

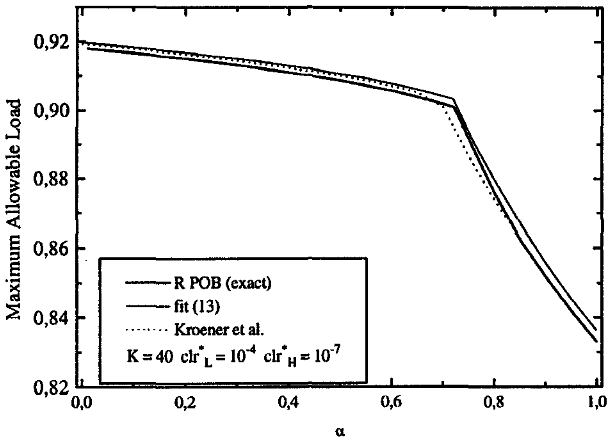

Figure 7. Comparison of the maximum allowable load ρ_{max} versus the priority mix α computed via different methods: the FIFO POB by Kröner *et al.* [5], the R POB and our proposed fit (13).

For a fixed ratio $\beta = \frac{clr_H^*}{clr_L^*}$ 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 as 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

$$\alpha_k \approx 10^{-\frac{3}{2K}} \left(clr_L^*\right)^{\frac{1}{4K}} \beta^{\frac{1}{K}} \tag{14}$$

Figure 7 compares the quality of the fit procedure described above with the FIFO POB [5] and the R POB. This plot exhibits that about a 1% accuracy is achieved.

4. Introducing burstiness in the arrival process.

So far, a Poisson arrival law was 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 R POB and PBS is investigated for arrivals generated by a Markov Modulated Process with N states (MMP(N)).

4.1. Compound Poisson Process.

As an example, we take $B(z) = z^B$, meaning that each burst precisely consists 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 R POB, 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. Fig. 8 plots the maximum allowable λ for both cases.

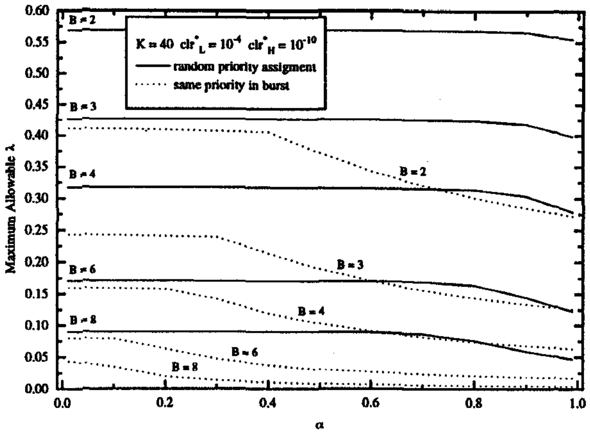

Figure 8. The maximum allowable λ (the load per burst B) in the R POB versus the priority mix α for various burst lenghts B but fixed buffer size $K = 40$ and fixed cell loss ratios $clr_L^* = 10^{-4}$ and $clr_H^* = 10^{-10}$. Curves with random priority assignment within a burst are drawn in full line, while the dotted line represents the case where all cells in a burst have the same priority.

In the case of random priority assignment, the result shown in Fig. 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 [4]. When the burst length B approaches K, the dependence of ρ_{max} on α increases slightly.

In the case of 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 (ρ_{max}) in the R POB 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 R POB of size $\frac{K}{B}$, when this fraction is an integer.

4.2. Markov Modulated Poisson Process (MMPP).

We refer to [11] for the detailed derivation of the R POB with MMP(N) arrivals in discrete-time. The MMPP(N)-PBS has been computed by extending the results of Kröner *et al.* [5].

A possible way to relate the characteristics of the actual arrival process to the set of parameters describing an N-state

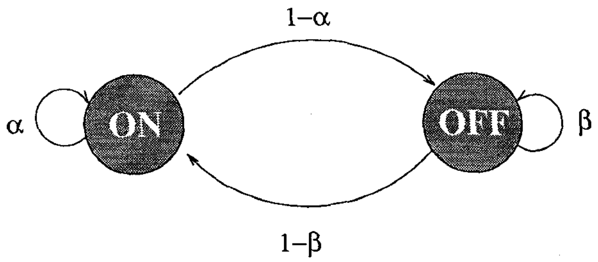

Figure 9. The Markov chain for $N = 2$.

MMP, is to consider the arrival process as a succession of ON and OFF slots. During an OFF slot, no cells are generated, while 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)

$$A(z) = 1 - \sigma + \sigma\, e^{\lambda(z-1)} \qquad (15)$$

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 $\frac{1}{\sigma}$, respectively parameter $1 - \sigma$ and mean $\frac{1}{1-\sigma}$. 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, while values of σ close to 1 imply that the cell arrivals are spread over virtually all slots. Numerical examples (Fig.. 10, 12) illustrate the strong impact of σ on the admissible aggregate load. In a two-state model (Fig. 9) 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 [11]), the length of the ON periods is geometrically distributed with parameter α and mean $1/(1-\alpha)$, while 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

$$\sigma = \frac{1-\beta}{2-\alpha-\beta} \qquad (16)$$

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,

$$\kappa \triangleq \frac{1-\sigma}{1-\alpha} = \frac{\sigma}{1-\beta} \qquad (17)$$

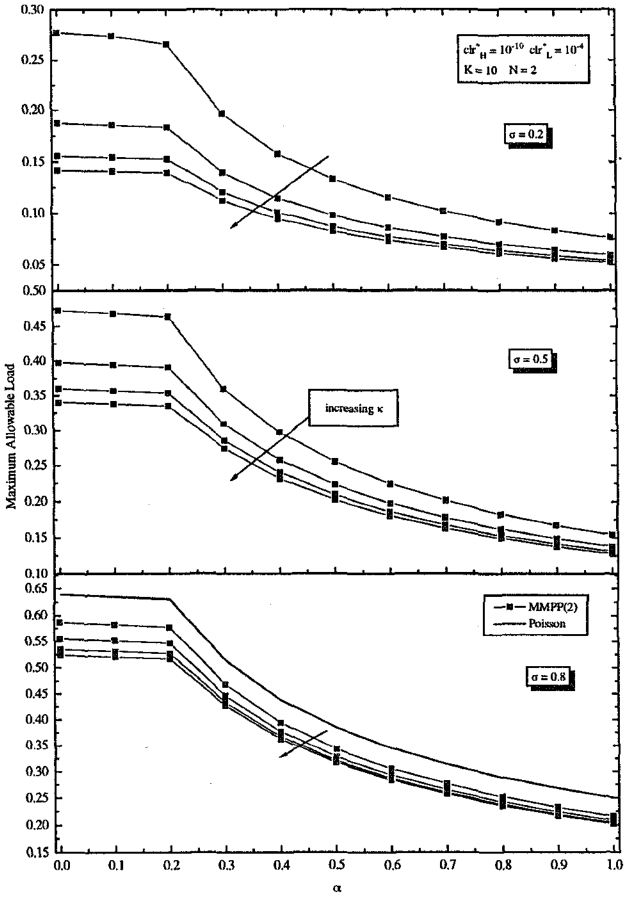

Figure 10. The maximum allowable load ρ_{max} in the R POB versus the priority mix α. The arrival process is generated by a MMPP(2) for three values of σ (16). In each plot, κ (17) increases from κ = 1, 2, 4 to 8. The buffer size K = 10 as well as the cell loss ratios $clr^*_L = 10^{-4}$ and $clr^*_H = 10^{-10}$ are the same for all curves.

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 (Fig. 11) 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.

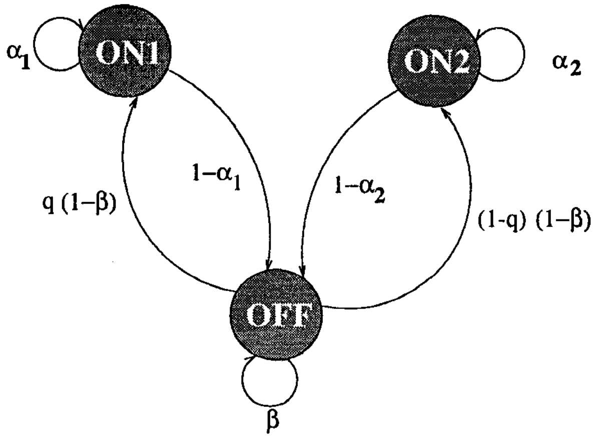

Figure 11. The Markov chain for $N = 3$.

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 $N = 2$,

$$R \triangleq \frac{(1-\sigma)^2}{(\kappa+1-\sigma)\kappa}\left[q(1-q)\left(\frac{1}{1-\alpha_1}-\frac{1}{1-\alpha_2}\right)+\frac{q\alpha_1}{(1-\alpha_1)^2}+\frac{(1-q)\alpha_2}{(1-\alpha_2)^2}\right] \tag{18}$$

where

$$\sigma = \frac{\frac{q}{1-\alpha_1}+\frac{1-q}{1-\alpha_2}}{\frac{1}{1-\beta}+\frac{q}{1-\alpha_1}+\frac{1-q}{1-\alpha_2}} \tag{19}$$

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

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

$$\beta = 1-\frac{\sigma}{\kappa} \tag{21}$$

$$\frac{1}{1-\alpha_1} = \frac{1}{1-\sigma}\left[\kappa+\sqrt{\frac{1-q}{2q}S}\right] \tag{22}$$

$$\frac{1}{1-\alpha_2} = \frac{1}{1-\sigma}\left[\kappa-\sqrt{\frac{q}{2(1-q)}S}\right] \tag{23}$$

where $S = (R-1)\kappa(\kappa+\sigma-1)$. For fixed values of σ, κ and q, the variance of the ON periods and, hence R, is bounded by

$$R-1 < 2\frac{1-q}{q}\frac{\kappa+\sigma-1}{\kappa} \tag{24}$$

By choosing q is sufficiently small, (24) indicates that any value of R can be realized.

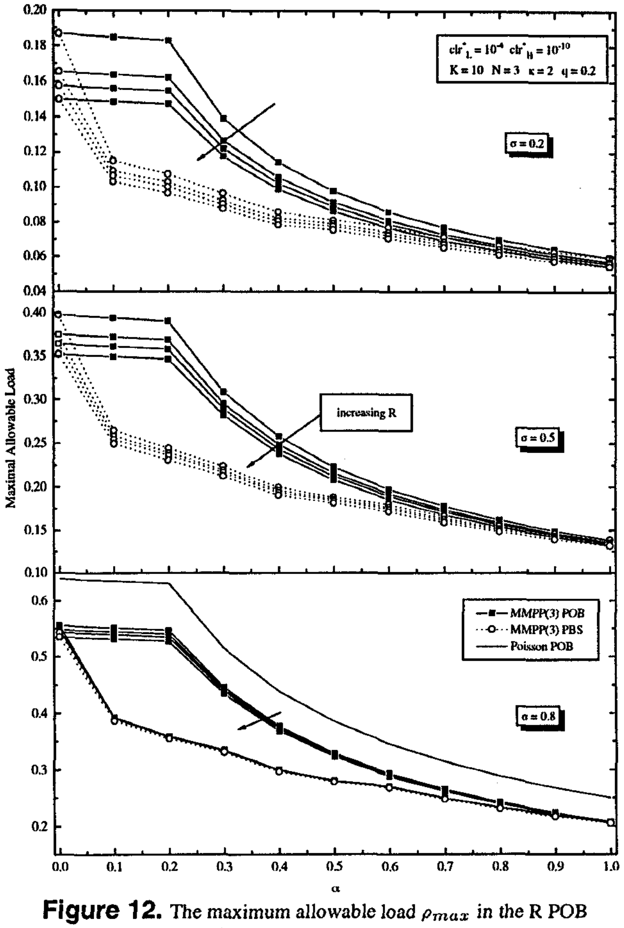

Figure 12. The maximum allowable load ρ_{max} in the R POB and PBS versus the priority mix α. The arrival process is generated by a MMPP(3) for three values of σ (19). In each plot, $\kappa = 2$ and $q = 0.2$ are constant, while R increases as $R = 1, 2, 3$ and 5. The buffer size $K = 10$ as well as the cell loss ratio's $clr^*_L = 10^{-4}$ and $clr^*_H = 10^{-10}$ are the same for all curves.

Figures 10 and 12 show the behaviour of ρ_{max} for R POB versus α for various combinations of the parameters σ, κ and R for a relatively small buffer $K = 10$. As the shape is similar to that with pure Poisson arrivals, the results may hint that a MMPP(N) with Bernouilli 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 3.4 seem applicable. For the three-state model, the performance of PBS is also shown (Fig.12) clearly demonstrating a still higher superiority of R POB as burstiness is involved.

4.3. Conclusions on priorities and burstiness in the R POB.

Our study shows that the shape of the performance curve of the R POB is 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 emits always 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 ρ_{max} versus α). On the other hand, as expected, the value of ρ_{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.

5. Summary.

The optimality of priority management 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 case of the push-out buffer and partial buffer sharing.

Further, this paper has focused on the maximum allowable load ρ_{max} for R POB 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 ρ_{max} for a certain α, but hardly the shape of ρ_{max} versus α. The priority assignment distribution within the aggregate cell flow is found to change the form of the ρ_{max} vs. α-curve.

Acknowledgements. The authors are much indebted to J. David and are grateful to H. Kröner for the data in Fig.4. This work has been sponsored by the Flemish Institute for the Promotion of Scientific and Technological Research in Industry (IWT 950141).

References

[1] 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, June 1994.

[2] M. de Prycker. *Asynchronous Transfer Mode: Solution for Broadband ISDN.* Ellis Horwood, New York, third edition, 1995.

[3] G. Gallassi, G. Rigolio, and L. Fratta. Bandwidth assignment in prioritized ATM networks. *Proc. IEEE Globecom'90*, 505(2):852–856, 1990.

[4] J. Garcia and O. Casals. Performance evaluation of source dependent congestion control procedures in ATM networks. *Proc. 1991 Singapore International Conference on Networks*, pages 178–182, Sept. 1991.

[5] H. Kröner, G. Hebuterne, P. Boyer, and A. Gravey. Priority management in ATM switching nodes. *IEEE J. Select. Areas Commun.*, 9(3):418–427, April 1991.

[6] 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, Dec. 1991.

[7] H. Saito. *Teletraffic Technologies in ATM Networks.* Artech House, Boston, 1994.

[8] 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, June 1994.

[9] P. Van Mieghem. The asymptotic behaviour of queueing systems: Large deviations theory and dominant pole approximation. *Queueing Systems*, 23:27–55, 1996.

[10] P. Van Mieghem, J. David, and G. H. Petit. Performance of cell loss priority management schemes in shared buffers with poisson arrivals. *European Transactions on Telecommunications: to be published*, 1997.

[11] P. Van Mieghem, B. Staeyaert, and G. Petit. Performance of cell loss priority management schemes in a single server queue. *International Journal of Communication Systems: to be published*, 1997.