4OR manuscript No.
(will be inserted by the editor)

Delay Analysis of a Discrete-Time Multiclass Slot-Bound Priority System

Sofian De Clercq*, Bart Steyaert and Herwig Bruneel

Department of Telecommunication and Information Processing, Ghent University, Ghent, Belgium.
e-mail: {sdclercq, bs, hb}@telin.ugent.be

Received: April 14, 2011 / Revised version: September 19, 2011

Abstract This paper introduces a new priority mechanism in discrete-time queueing systems that compromises between first-come-first-served (FCFS) and head-of-line (HoL) priority. In this scheduling discipline - which we dubbed slot-bound priority - customers of different priority classes entering the system during the same time-slot are served in order of their respective priority class. Customers entering during different slots are served on a FCFS basis. In this paper we study the delay in an N-class discrete-time queueing system under slot-bound priority. General independent arrivals and class-specific general service time distributions are assumed. Expressions for the probability generating function of the delay of a random type-j customer are derived, from which the respective moments are easily obtained. The tail behaviour of these distributions is analyzed as well, and some numerical examples show the effect slot-bound priority can have on the performance measures.

Keywords: Queueing Theory, Priority Systems, Discrete Time, Generating Function, Delay

1 Introduction

Multiclass queueing systems, or queueing systems buffering multiple types of customers, have been widely adopted in queueing theory, since they enable the modelling of non-identical behaviour of different types of customers

* Corresponding author: Sofian De Clercq, Department of Telecommunication and Information Processing, Ghent University, St.-Pietersnieuwstraat 41, 9000 Gent, Belgium, Tel. +32-9-264-34-12, Fax +32-9-264-42-95.

that enter the same system. In a multiclass environment, virtually any combination of features with respect to the arrival characteristics, service requirements, buffer management rules that pertain to individual classes of customers can be considered (see f.i. D. Fiems et al. (2007) for an example of very class specific queueing behaviour).

In this paper we study an N-class discrete-time queueing system with infinite waiting room, under the so-called *slot-bound priority rule* (SBP). That is to say, type-1 customers receive preferential treatment over type-2 customers that have arrived during the same slot. In the same way, type-2 customers receive preferential treatment over type-3 customers and so on up until type-N customers, which represent the lowest priority class. In this sense, the higher-priority classes have limited priority over the lower-priority ones. In addition, customers that enter the system during consecutive slots are served on a first-come-first-served (FCFS) basis, regardless of the class they belong to.

Many papers investigate the well know priority scheduling in which priority class customers receive preferential treatment over non-priority class customers of which R.P. Miller Jr. (1960), P.G. Harrison et al. (2005) (in continuous time) and S. Ndreca et al. (2008), J. Walraevens et al. (2002) (in discrete-time) are but some examples. Considering multiple classes of customers is often combined with assigning some kind of priority to either of the classes, but it is not a must. It can be preferable that customers pertaining to different classes are served FCFS. T. Takine (2001) studied such a system in continuous time. In a discrete-time model, when we demand that during a slot only customers pertaining to one class can enter the system, a pure multiclass FCFS policy is the result. However one has the problem of what to do with batches containing customers of different types that arrive during a single slot. The fairest policy could for instance be to serve all customers in a batch in random order regardless of the class they belong to. If however, we prioritize some customer in that batch according to type, we effectively introduce an intermediate priority rule which we dubbed slot-bound priority.

Slot-bound priority can be used to model any system in which batches of for example customers, packets, or tasks arrive which have to be addressed or serviced in a specific order for whatever reason, while the batches themselves need to be served FCFS. For instance a batch of customers may be the traffic that accumulates before a trafficlight. When the light turns green, the faster drivers (high priority customers) will gain an edge and arrive at the next lights sooner, where they will be 'served' once those lights turn green. Moreover SBP can be seen as a polling mechanism where N queues

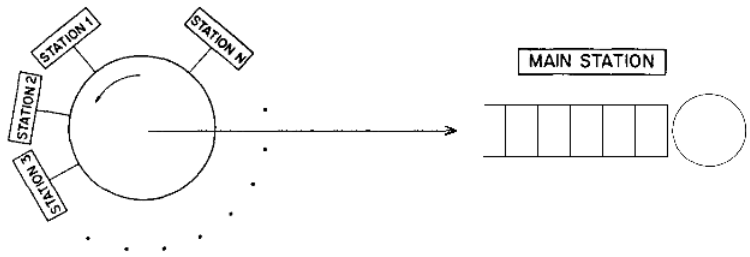

Fig. 1.1. A loop system. The server takes of tour of N stations by which the first station has priority over subsequent stations.

are visited by the server in a cyclic order, and where after each slot, a gate is placed between the arrivals in each of the N queues (see f.i. H. Takagi (1986), O. Boxma et al. (2009)). Also, in a loop system, in the sense explained in Konheim et al. (1972) (see Fig.1.1) where the server takes a tour of N stations each slot and its capacity is unbounded, then the main station behaves as described in our proposed slot-bound priority discipline; Konheim analyzed the N input stations instead of the main station. When the server takes a tour each slot, the interesting queueing phenomena occur at the main station. Finally, by adjusting the slot size, and thus the number of arrivals per slot, one can make slot-bound priority resemble a more general delay differentiating service policy. These observations show that the SBP service mechanism, which to the best of our knowledge has not yet been studied before, potentially has a wide applicability.

The remainder of the paper is organized as follows. In the next section we present the mathematical model of our queueing system. Next, we derive a steady-state expression for the probability generating functions (pgf) of the delay of an arbitrary type-j customer $j \in [1, N]$. Should the reader not be familiar with generating functions, we refer to H.S. Wilf (1994). From this main result, we subsequently derive expressions for the mean values of these random variables as well as their tail probabilities. Finally, for some specific scenarios we see the effect slot-bound priority has on the average delay of high and low-priority customers.

2 Mathematical Model

The discrete-time single server queueing system analyzed hereafter serves arriving customers without interrupting the order in which they enter the system. If two or more customers enter the system during the same slot then the order is arbitrary unless they belong to different customer classes

- N customer classes are assumed. The customers pertaining to the highest priority class are served first after which those pertaining to the second highest priority class and so on, until all customers of the tagged batch are served. Because the priority only has an effect on customers that arrived during the same slot, we call this server policy slot-bound priority.

Let $a_{j,n}$ represent the number of type-j customers ($j \in [1, N]$) entering the system during the slot with index n. In our analysis we assume a general independent and identically distributed (i.i.d.) arrival process such that the joint pgf $A_n(\mathbf{z}) \triangleq E[\prod_{j=1}^{N} z_j^{a_{j,n}}]$ is independent of the slot index (we used $\mathbf{z}$ to represent the vector with j'th element equal to z_j). Therefore we will omit the index for this pgf in the rest of this paper. On a not unimportant sidenote, this model allows that $a_{j,n}$ and $a_{j',n}$ are correlated ($j \neq j'$). We will abbreviate the first and second moments of these discrete random variables (drv's) as

$$\begin{aligned} \lambda_j &\triangleq \frac{\partial}{\partial z_j} A(\mathbf{z})\big|_{\mathbf{z}=\mathbf{1}} & \lambda_T &\triangleq \sum_{j=1}^{N} \lambda_j \\ \lambda_{ij} &\triangleq \frac{\partial^2}{\partial z_i \partial z_j} A(\mathbf{z})\big|_{\mathbf{z}=\mathbf{1}} & \lambda_{TT} &\triangleq \frac{d^2}{dz^2} A(z\mathbf{1})\big|_{z=1} \quad , \end{aligned} \tag{2.1}$$

in which the t'th element of the vector $\mathbf{1}$ is 1. Furthermore, let $s_{j,k}$ represent the service time of the k-th type-j customer entering the system. We assume that service times can only start and end at slot-bounds, making $s_{j,k}$ a discrete random variable. We consider the case that the service times of all type-j customers are mutually independent and follow the same distribution, and the pgf of this distribution will be given by $S_j(z) \triangleq E[z^{s_j}]$ (with mean $\mu_j \triangleq E[s_j]$), in which we omitted the index k, for s_j to represent the service time of an arbitrary type-j customer. Additionally different types of customers can have different service-time distributions.

Lastly, we will call $\rho_j \triangleq \lambda_j \mu_j$ the load offered by type-j customers. When considering an infinite buffer, as we will here, the equilibrium condition requires that the total load $\rho \triangleq \sum_j \rho_j$ be less than 1.

3 Delay Analysis

When an arbitrary type-j customer, hereafter referred to as customer c, enters our system to obtain service, it will spend some time in it before it

leaves. It will at least have to sit out its own service time, which is given by s_j. The rest of its delay is spent waiting on the departures of other customers that are scheduled to be served before it. This waiting time starts at the slot mark following customer c's arrival and ends when the its service is initiated. Hence we can write (with w_j representing the waiting time of customer c)

$$d_j = w_j + s_j, \tag{3.1}$$

in which we introduce the notation d_j for the delay of an arbitrary type-j customer. Relating $D_j(z) \triangleq E[z^{d_j}]$, the pgf of d_j to $S_j(z)$ and $A(\mathbf{z})$ is the purpose of this analysis.

The waiting time of customer c consists of two distinct parts. A first is the waiting time caused by customers already in the system at the beginning of customer c's arrival slot. The second is due to customers entering the system during the same slot as customer c but scheduled before it, which we will denote by the drv p_j. We define the unfinished work of the system as the number of slots it takes to serve all customers present in the system at a certain point in time. Since the arrival of batches of customers is governed by a Bernoulli process (with parameter $1 - A(0,0)$), the unfinished work at the beginning of a random slot (which we'll denote by the drv l) has the same distribution as the unfinished work at the beginning of a random type-j's arrival slot (i.e. customer c), as dictated by the BASTA-property (see S. Halfin (1983)). Because customer c's arrival slot is not included in its delay, we must subtract 1 from l (in case the system was not idle), to account for the first part of customer c's delay. We thus have

$$d_j = p_j + (l-1)^+ + s_j, \tag{3.2}$$

in which $(l-1)^+$ is shorthand for $\max(l-1, 0)$. The three drv's in the right-hand side are independent of one another: p_j and l are independent because of the i.i.d. property of the arrival process. The independence of these two drv's and s_j can be readily deduced from the i.i.d. nature of the service process.

If we address each unit of work as a typeless customer with service time 1, our system behaves as a $GI-D-1$ queue with an arrival process characterized by the pgf $A(\mathbf{S}(z))$ - we'll abbreviate the parameterlist $S_1(z), \ldots, S_N(z)$

to $\mathbf{S}(z)$. Hence, determining the amount of unfinished work in the system is a straightforward application of the result found in e.g. H. Bruneel (1993). We thus obtain for the pgf $L(z)$ of l

$$L(z) \triangleq E[z^l] = (1-\rho)\frac{(z-1)A(\mathbf{S}(z))}{z - A(\mathbf{S}(z))}, \tag{3.3}$$

and consequently

$$L^-(z) \triangleq E[z^{(l-1)^+}] = (1-\rho)\frac{z-1}{z - A(\mathbf{S}(z))}. \tag{3.4}$$

Let a_t^* denote the number of type-t customers entering the system in the same batch as customer c - we choose to omit j, the type of customer c, for notational convenience; the value of j will become clear from the context. Since we selected a random type-j customer and not a random slot we can't a priori conclude that $(a_1^*, \ldots, a_N^*)$ and $(a_1, \ldots, a_N)$ have the same joint distribution function. In fact, the more type-j customers a batch contains, the more probable it is for it to contain a random type-j customer. This is a well-defined problem known as 'the renewal paradox' (see f.i. L. Kleinrock (1975) or I. Mitrani (1987)) and the relation between the distributions of the two sets of drv's becomes

$$\Pr[a_1^* = k_1, \ldots, a_N^* = k_N] = \frac{k_j}{\lambda_j} \Pr[a_1 = k_1, \ldots a_N = k_N]. \tag{3.5}$$

We can obtain p_j, the delay caused by customers entering the system during the same slot as customer c but queued before it in terms of the number of each type of such customers. Let r_t represent the number of such type-t customers - where again we choose to omit j for notational convenience. The sum of the service times of all these customers then gives us p_j. Let $R_j(\mathbf{x}) \triangleq E[\prod_{t=1}^{N} x_t^{r_t}]$ be the joint pgf of r_1 through r_N, given that customer c's type is j - the t'th element of $\mathbf{x}$ is x_t. We can then write

$$P_j(z) \triangleq E[z^{p_j}] = R_j(\mathbf{S}(z)) \; . \tag{3.6}$$

The slot-bound priority rule dictates that p_j will consist of the service times of all type-t customers with $t < j$ in the selected batch, and some type-j customers service times. Because customer c can be each of the a_j^* type-j customers in the selected batch with equal probability, we can write

$$\Pr[r_1 = i_1, \ldots, r_N = i_N | a_1^* = k_1, \ldots, a_N^* = k_N] = k_j^{-1} \ , \\ \text{if } i_j < k_j,\ i_t = 0 \text{ for } t > j, \text{ and } i_t = k_t \text{ for } t < j. \qquad (3.7)$$

and 0 in all other cases. The joint pgf $R_j(\mathbf{x})$ can now be calculated as follows from (3.5) and (3.7)

$$\begin{aligned} R_j(\mathbf{x}) &= \sum_{\substack{i_1,\ldots,i_N,\\ k_1,\ldots,k_N \geq 0}} \Pr[r_1 = i_1, \ldots, r_N = i_N, a_1^* = k_1, \ldots, a_N^* = k_N] \prod_{t=1}^{N} x_t^{i_t} \\ &= \sum_{k_1,\ldots,k_N \geq 0} \Pr[a_1 = k_1, \ldots, a_N = k_N] \left(\prod_{t=1}^{j-1} x_t^{i_t} \right) \sum_{0 \leq i_j < k_j} \frac{x_j^{i_j}}{\lambda_j} \\ &= \sum_{k_1,\ldots,k_N \geq 0} \Pr[a_1 = k_1, \ldots, a_N = k_N] \left(\prod_{t=1}^{j-1} x_t^{i_t} \right) \frac{x_j^{k_j} - 1}{\lambda_j (x_j - 1)} \ . \end{aligned}$$

When we adopt the shorthand notation

$$C_j(\mathbf{x}) = A(x_1, \ldots, x_{j-1}, x_j, 1, \ldots, 1) - A(x_1, \ldots, x_{j-1}, 1, 1, \ldots, 1) \ , \quad (3.8)$$

we arrive at the following expression for $R_j(\mathbf{x})$.

$$R_j(\mathbf{x}) = \frac{C_j(\mathbf{x})}{\lambda_j (x_j - 1)} \ . \qquad (3.9)$$

Not surprisingly $R_j(\mathbf{x})$ is not a function of x_t, $t > j$ - as $r_t = 0$ if $t > j$. Because of the independence of the random variables in the right hand side

of equation (3.2), $D_j(z)$ can be obtained rather easily as

$$D_j(z) = P_j(z)L^-(z)S_j(z). \tag{3.10}$$

Substitution of (3.6) and (3.4) into (3.10) rewards us with an explicit formula for the pgf of d_j in our slot-bound priority system.

$$D_j(z) = (1-\rho)\,\frac{(z-1)S_j(z)}{z - A(\mathbf{S}(z))}\,\frac{C_j(\mathbf{S}(z))}{\lambda_j(S_j(z)-1)}\,. \tag{3.11}$$

The asymmetry in $C_j(\mathbf{S}(z))$ is a direct consequence of slot-bound priority scheduling. This can be illustrated by considering the case where $A(\mathbf{z})$ satisfies

$$A(\mathbf{z}) = \sum_{j=1}^{N} E[z_j^{a_j}] - (N-1)A(0,\ldots,0). \tag{3.12}$$

By substitution one can easily verify that the asymmetry in (3.11) vanishes. Indeed, slot-bound priority has no reordering effect if it is impossible that during the same slot customers of different types enter the system, which is exactly what (3.12) enforces.

Evidently, since λ_j/λ_T represents the probability that an arbitrary customer is of type j, the pgf $D(z)$ of an arbitrary customer's delay can be calculated as

$$D(z) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda_T} D_j(z) = (1-\rho)\,\frac{z-1}{\lambda_T(z - A(\mathbf{S}(z)))}\,\sum_{j=1}^{N} \frac{C_j(\mathbf{S}(z))}{1 - S_j(z)^{-1}}. \tag{3.13}$$

When service times of the different priority classes are identically distributed ($S_1(z) = S_2(z) = \ldots$), not entirely unexpected we have that (3.13) is congruent with the results found in H. Bruneel (1993). When the delay distribution of the last customer in each batch is of particular interest, a

simple trick might be to add a customer type (type-$N+1$) with lowest priority type and constant service time equal to zero. The pgf $D_{N+1}(z)$ then represents this delay distribution, if and only if a type-$(N+1)$ customer enters the system only during a slot featuring other arrivals, and only one type-$(N+1)$ customer joins the queue that slot. Of course delay distributions of the last type-1 customer, or the first customer of a batch can be calculated in this way using service times equal to zero as well.

One problem with zero service times is the denominator of $P_j(z)$. It is easy to calculate that the joint pgf of the number of other customers entering the system during the same slot as customer c is given by $\frac{1}{\lambda_j}\frac{\partial}{\partial x_j}A(\mathbf{x})$. Only the customers pertaining to higher priority classes will contribute their service time to p_j (those of equal priority have service time zero if any). Summarizing we find that

$$P_j(z) = \frac{1}{\lambda_j}\frac{\delta}{\delta x_j}A(x_1,\ldots,x_{j-1},1,\ldots,1)\Big|_{\mathbf{x}=\mathbf{S}(\mathbf{z})}, \quad \text{if } S_j(z)=1. \tag{3.14}$$

3.1 Mean Delay

Among others, the first moment of the class-j customer delay can be derived from the obtained respective pgf's as follows:

$$\mathrm{E}[d_j] = \mu_j + \mathrm{E}[(l-1)^+] + \mathrm{E}[p_j] \tag{3.15}$$

$$\mathrm{E}[d] = \frac{\rho}{\lambda_T} + \mathrm{E}[(l-1)^+] + \sum_{j=1}^{N}\frac{\lambda_j}{\lambda_T}\mathrm{E}[p_j], \tag{3.16}$$

where $\mathrm{E}[d]$ represents the mean delay of an arbitrary customer. We can therefore find the first moment of d_j and d by differentiating (3.4), (3.6), resulting in

$$\mathrm{E}[d_j] = \mu_j + \frac{\frac{d^2}{dz^2}A(\mathbf{S}(z))|_{z=1}}{2(1-\rho)} + \frac{\lambda_{jj}\mu_j + 2\sum_{i=1}^{j-1}\lambda_{ij}\mu_i}{2\lambda_j} \tag{3.17}$$

$$\mathrm{E}[d] = \frac{\rho}{\lambda_T} + \frac{\frac{d^2}{dz^2}A(\mathbf{S}(z))|_{z=1}}{2(1-\rho)} + \frac{\lambda_{TT}\rho}{2\lambda_T^2}, \tag{3.18}$$

in which the asymmetry, instigated by the slight priority of higher priority classes, limits itself to the last term in $\mathrm{E}[d_j]$, being $\mathrm{E}[p_j]$. Higher-order moments of the class-j customer delay, such as the variance, can be calculated in a similar way, albeit that the expressions become more complicated.

3.2 Tail Distribution

In this section we derive the tail probabilities using the dominant pole approximation technique (see f.i. Van Mieghem (1996)).

When the delay of a customer of whatever type becomes exceptionally large, there can be three reasons for this. First, its own service time may be exceptionally large. Second, it can be of low priority and the batch of customers in which it arrived featured a lot of high priority customers. And third, the queue size was already very high when it entered. We will see that, not surprisingly, this last scenario is the most probable. Concrete, we will prove that generally speaking, the dominant singularity of $D_j(z)$ is the dominant singularity of $L^-(z)$, and not of $S_j(z)$ or $P_j(z)$. It will also become clear that the dominant singularity is independent of the type of the customer.

We assume that the dominant singularities of all single-variate pgf's under consideration are poles. That way, from Vivanti's theorem, we know they are real and positive. Furthermore, because pgf's are always analytic inside and well defined on the unit disk, these singularities must have a modulus larger than 1. Note that z_{as}, dominant pole of $A(\mathbf{S}(z))$, is smaller than or equal to the dominant pole of $S_j(z)$ unless there are no type-j arrivals ($A(\mathbf{z})$ is independent of its j'th argument), a situation that we exclude from these considerations. The same is true for the dominant pole of $A(S_1(z), \ldots, S_j(z), 1, \ldots, 1)$, which we'll denote z_{as_j}. Also $z_{as_j} \leq z_{as_{j-1}}$ simply because $S_j(z) \geq 1$ when $z \geq 1$.

Possible candidates for the dominant pole of $D_j(z)$, denoted by z_{d_j}, are the dominant pole of $S_j(z)$, that of $z-1$ (which is ∞), or $C_j(\mathbf{S}(z))$, being z_{as_j} which we already reasoned to be smaller than that of $S_j(z)$. The other candidates are zeroes of the denominators of $D_j(z)$, $z - A(\mathbf{S}(z))$ and $S_j(z) - 1$. The latter function has no zeros greater than 1 within its radius of convergence unless $S_j(z) = 1$ in which case $D_j(z)$ does not have $S_j(z)-1$ in its denominator (see (3.14)). The former function has a zero smaller than z_{as}, and is thus the z_{d_j} we searched for. Since z_{d_j} does not depend on the

specific value of j, we set $z_{d_j} \equiv z_d$. Indeed, one can show that under the circumstances previously described, there is a zero for $z - A(\mathbf{S}(z))$ greater than 1 and smaller than z_{as}, provided that $\rho < 1$, where ρ represents the first derivative of $A(\mathbf{S}(z))$ in 1. Summarizing, we find

$$\begin{aligned} z_d &= A(\mathbf{S}(z_d)), \quad z_d > 1 \\ \Pr[d_j = n] &\approx -\theta_j z_d^{-n-1} , \end{aligned} \tag{3.19}$$

where θ_j is the residue of $D_j(z)$ at $z = z_d$ $(= \lim_{z \to z_d}(z - z_d)D_j(z))$, which can be easily calculated from (3.11).

4 Numerical Examples

To help understand the implications of the slot-bound priority model, we present some numerical examples. The number of parameters the average delay of a random customer is dependent on is quite large, and so we illustrate the effects of some of the more important ones. One key parameter, the workload ρ, will have a profound effect on the delay differentiation between customers of different types. Specifically, we will see that a larger delay differentiation is obtained for low workloads, whereas for higher workloads, a more fair queueing delay is obtained (since the second term in (3.17) becomes dominant).

For sake of simplicity we will assume $N = 2$ and see what parameters affect differences in average delays more than others. Previously, we stated that adjusting the slot size, and thus the number of arrivals per slot, one can make slot-bound priority resemble a more general delay differentiating service policy. Let therefore t be the slot size in for instance seconds. Our proposed arrival and service processes are then characterized by the following pgf's:

$$\begin{aligned} A(z_1, z_2) &= e^{\alpha t(\frac{z_1+z_2}{2}-1)} \\ S_j(z) &= \frac{t}{t + \mu(1-z)} . \end{aligned} \tag{4.1}$$

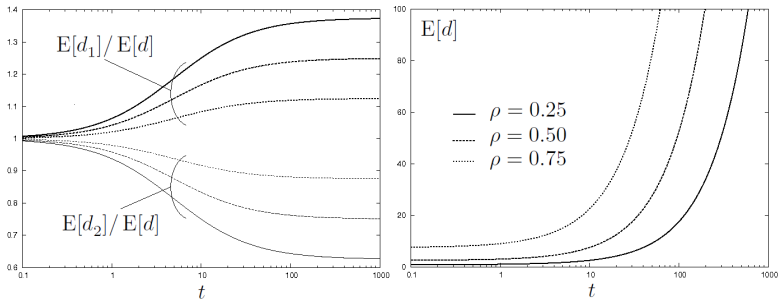

Fig. 4.1. Relative delay (a), and absolute delay (b) against slot-size t. The parameters α and μ from (4.1) were chosen to be 0.4 and ρ/α. Both graphs contain curves for $\rho = 0.25,\ 0.50,\ 0.75$.

In this first example, service times (in number of slots) are geometrically distributed with mean μ/t, independent of j, with μ the mean service time (in seconds per customer). The arrival process is composed of two independent Poisson arrival streams with mean $\alpha t/2$ (customers per slot). Hence, $\rho = \alpha\mu$ represents the load of the system, which is independent of t. By increasing t, we effectively increase the average number of arrivals per slot and reduce the average service time in slots. However the average service time in seconds stays the same, as well as the average number of arrivals per second.

Since slot-bound priority has an effect on the slot-level, we obtain more delay differentiation when t is higher. This is illustrated for three different workloads in Fig.4.1a where $\mathrm{E}[d_j]/\mathrm{E}[d]$ is plotted against t. In Fig.4.1b, the absolute value of the delay of a random customer in seconds is plotted against t. The increasing nature of $\mathrm{E}[d]$ as a function of t is due to the variance on s_j and all a_j - note that $\mathrm{E}[d]$ does not ascend as fast as t. Another important observation is that the delay differentiation is limited, in the sense that $\mathrm{E}[d_j]/\mathrm{E}[d]$ has an asymptote for $t \to \infty$. When the slot-size is very short (the scale on the abscissa goes down to 100 ms), then delay differentiation goes to a minimum. In our example, $\mathrm{E}[d_j] \to \mathrm{E}[d]$, but only because $\lambda_{12} \to 0$ for $t \to 0$.

Next, we will show that delay differentiation is limited even for very bursty arrivals. If customers enter the queue typically in very large batches, a lot of reordering is going to take place, and hence the effect of slot-bound priority is going to be maximal. However, for increasingly burstier arrivals, an increasingly larger queue will be formed prior to a newly arriving batch, due to the variance on the arrival process, which is an effect that increases the delay regardless of the class that a customer belongs to. These two effects counteract eachother, so it is interesting to see their combined effect.

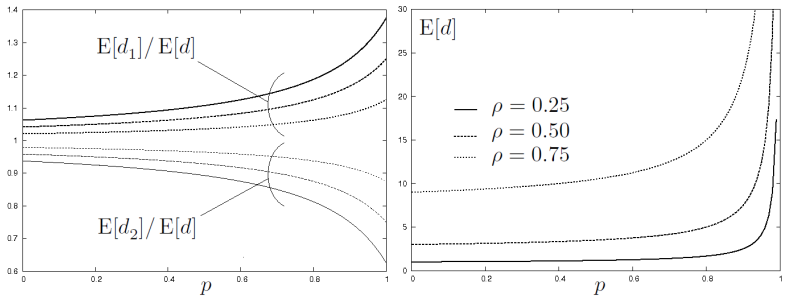

Fig. 4.2. Relative delay (a), and absolute delay (b) against the burstiness factor p (see text). The parameters α and μ from (4.2) were chosen to be 0.4 and ρ/α. Both graphs contain curves for $\rho = 0.25,\ 0.50,\ 0.75$.

A nice way of introducing burstiness into the arrival process is for instance reflected by the following pgf's:

$$\begin{aligned} A(z_1, z_2) &= (1-p)e^{\frac{\alpha}{1-p}(\frac{z_1+z_2}{2}-1)} + p \\ S_j(z) &= \frac{1}{1+\mu(1-z)} \,. \end{aligned} \tag{4.2}$$

For $p = 0$ we have two independent ordinary Poisson arrival streams for each of our customer classes. The larger p becomes, the less batch arrivals we have, but the bigger the batches themselves become to keep the workload a constant. When p approaches 1, arrivals will be very sparsely distributed in time, but once they do occur, a gigantic batch will occupy the queue meaning that the variance of the number of arrivals per slot (i.e., the batch size) becomes infinitely large. In Fig.4.2a we plotted $\mathrm{E}[d_j]/\mathrm{E}[d]$ against p for three different workloads. On this graph we see that delay differentiation grows as p, our burstiness factor, increases but is clearly limited even for $p \to 1$. When looking at (3.17) one can see that increasing the burstiness of the arrival process by increasing all λ_{ij} while keeping all λ_j constant - as we do by increasing p - in general indeed has this effect on the delay differentiation between the different customer classes. Moreover Fig.4.2 shows that by increasing the variance of the arrival process (as we do by increasing p), $\mathrm{E}[d] \to \infty$, as insinuated by (3.18).

Furthermore, allowing the service times of the different customer classes to have different distributions can also have a tremendous impact on the queueing delay. One way of seeing this is realizing that introducing more variation in the service process is effectively increasing the queueing delay.

When priorities get into the mix however some interesting things may occur. For the next graph, we introduce the following pgf's of the arrival and service processes:

$$\begin{aligned} A(z_1, z_2) &= e^{\alpha(\frac{z_1+z_2}{2}-1)} \\ S_j(z) &= \frac{\alpha}{\alpha + 2\rho_j(1-z)} \; . \end{aligned} \tag{4.3}$$

In Fig.4.3 we plotted $\log_{10} \Pr[d_j > 20]$ on a log-scale against ρ_1/ρ, keeping the workload $\rho = \rho_1 + \rho_2$ constant (and hence decreasing ρ_2 when ρ_1 gets larger). We can see that $\Pr[d_1 > 20]$ decreases for increasing ρ_1 (until about $\rho_1 = 0.46$), which is a somewhat counterintuitive result and has to do with the dominant pole extracted from $z - A(\mathbf{S}(z))$. For the above mentioned service time distributions, this decrease can be explained by the reduction in overall service-time variability on which $\mathrm{E}[d_j]$ and conversely $\Pr[d_j > 20]$ is dependent. When service times of type-1 keep increasing, there comes a point at which $\Pr[d_1 > 20] > \Pr[d_2 > 20]$. Whether or not this point exists for a general independent arrival process during consecutive slots, is largely dependent on the degree of correlation between the number of arrivals of the different types of customers that arrive during the same slot. Since z_d is the dominant pole of $D_1(z)$ and $D_2(z)$, the existence of this turnover point can be checked by using only the residue of both functions at z_d.

Up until now we have considered a system in which only two types of customers enter. When we make the number of customer types a variable while keeping the total workload ρ constant, we may consider pgfs $A(z_1, z_2)$ and $S_j(z)$ that take the following form:

$$\begin{aligned} A(\mathbf{z}) &= e^{\frac{\alpha}{N}(\prod_{j=1}^{N} z_j - 1)} \\ S_j(z) &= \frac{1}{1 + \mu(1-z)} \; . \end{aligned} \tag{4.4}$$

We observe from Fig.4.4 that for $N \to \infty$, $\mathrm{E}[d_1]$ evolves to a minimum, and on the other end $\mathrm{E}[d_N]$ evolves to a maximum for a fixed value of ρ. Nonetheless, Fig.4.4 shows that the difference between $\mathrm{E}[d_N]$ and $\mathrm{E}\, d_1$ remains bounded for increasing N. This can be understood as follows. When $N \to \infty$, and the compound Poisson process produces a type-j customer,

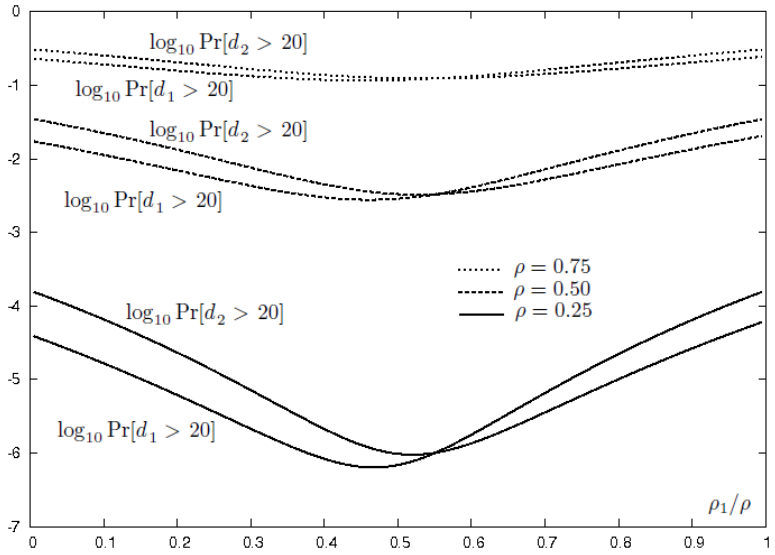

Fig. 4.3. Tail probabilities $\log_{10}\Pr[d_j > 20]$ plotted against $\frac{\rho_1}{\rho}$. The parameter α in (4.3) is kept constant at 0.4 and hence $\frac{\rho_1}{\rho}$ is altered by adjusting μ_1 and μ_2 while keeping the load ρ constant. Both graphs contain curves for $\rho = 0.25,\ 0.50,\ 0.75$.

because of the fixed workload, it will most likely not produce a second type-j customer. However the average arrival rate remains unaltered when we increase N. This means that in the limit $\mathrm{E}[d_1]$ is the average delay of the first customer in a batch given that at least one type-1 customer is in it. The average delay of the last customer in a batch given that at least one type-N customer is in it, is then given by $\mathrm{E}[d_N]$. On a final note, the paragraph following (3.13) explains why $D(z)$ is independent of N as long as $S_j(z)$ and $A(\mathbf{S}(z))$ are independent of j and N respectively. Because of this, $\mathrm{E}[d]$ is constant, and we have therefore omitted $\mathrm{E}[d]$ from Fig.4.4.

5 Conclusions

This paper introduced a new priority mechanism and dubbed it slot-bound priority, which allows us to assign a limited amount of priority to a particular customer class. We proceeded to study the delay in an N-class discrete-time queueing system with i.i.d. arrivals and class-dependent distributions for the service times under the slot-bound priority rule. We found expressions for the probability distributions of the delay of a random type-j customers ($j \in [1, N]$) in the form of their respective probability generating functions,

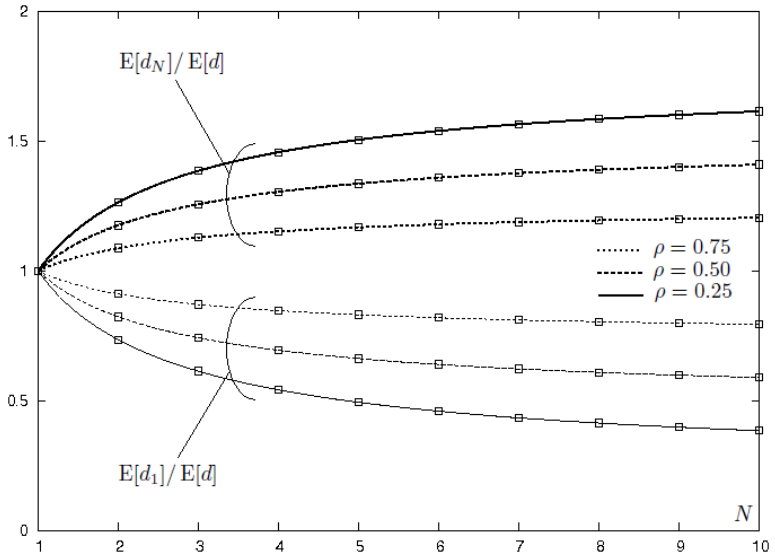

Fig. 4.4. Average relative delay of the highest and lowest priority classes in an N-class SBP system against the number of priority classes N. The parameters α and μ from (4.4) were chosen to be 0.4 and ρ/α. Both graphs contain curves for $\rho = 0.25,\ 0.50,\ 0.75$

and we derived their first moments, as well as calculated the associated tail probabilities. In a couple of numerical examples, we studied the effects of various parameters of the arrival process and service times of customers. We found it instructing to compare the average delays for a type-1 customer and type-N customer for various values of the workload. Most interesting was the observation that for low loads delay differentiation is very stressed, whereas for high loads the average delays tend to approximate the average delay of a random customer regardless of type. Other parameter dependencies were evaluated as well, such as actual the length of the slot-size, and burstiness of the arrival process.

References

S. Halfin (1983) Batch Delays Versus Customer Delays, *The Bell System Technical Journal* 62:2011–2015

H. Bruneel (1993) Performance of Discrete-Time Queueing Systems, *Computer Operations Research* 20:303–320

L. Kleinrock (1975) *Queueing Systems, Volume I: Theory*, Wiley, New York. (ISBN 0-471-49110-1).

I. Mitrani (1987) *Modelling of Computer and Communication Systems*, Cambridge: Cambridge University Press. (ISBN 0-521-31422-4)

D. Fiems, J. Walraevens, H. Bruneel (2007) Performance of a Partially Shared Priority Buffer with Correlated Arrivals, *Proceedings of the 20th International Teletraffic*

Congress (ITC20), Ottawa, 17-21 June 2007, Lecture Notes in Computer Science 4516, pp. 582-593.
R.P. Miller Jr. (1960) Priority queues, *Annals of Mathematical Statistics* 31:86–103
J. Walraevens, B. Steyaert, H. Bruneel, (2002) Delay characteristics in discrete-time GI-G-1 queues with non-preemptive priority queueing discipline, *Performance Evaluation* 50:53–75
T. Takine (2001) Queue Length Distribution in a FIFO Single-Server Queue with Multiple Arrival Streams Having Different Service Time Distributions, *Queueing Systems* 39: 349–375
H.S. Wilf (1994) *Generatingfunctionology*, Department of Mathematics, University of Pennsylvania, AcademicPress, Inc. (ISBN 0-127-51956-4).
S. Ndreca, B. Scoppola (2008) DiscreteTime GI/Geom/1 Queueing System with Priority, *European Journal of Operational Research* 189:1403–1408
P.G. Harrison, Y. Zhang (2005) Delay analysis of priority queues with modulated traffic, *13th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Atlanta (GA)*, pp. 280-287.
H. Takagi, *Analysis of Polling Systems*, MIT Press (ISBN 978-0262200578).
O. Boxma, J. Bruin, B. Fralix (2009) Sojourn times in polling systems with various service disciplines, *Performance Evaluation* 66:621-639.
A.G. Konheim, B. Meister (1972) Service in Loop System, *Journal of the Association for Computing Machinery* 19:92–108
P. Van Mieghem (1996) The asymptotic behavior of queueing systems: Large deviations theory and dominant pole approximation, *Queueing Systems* 23:27–55