Operations Research Letters 37 (2009) 47–50

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Loss probability properties in retrial queues

Chia-Li Wang [a,*], Ronald W. Wolff [b]

[a] Department of Applied Mathematics, National Dong Hwa University, Hualien, Taiwan, ROC

[b] Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 10 July 2008
Accepted 25 September 2008
Available online 8 October 2008

Keywords:
Loss system
Orbiting customers
Arbitrary arrival process
Sample path
Coupling
Convex

ABSTRACT

Under general conditions for c-server loss systems, B_c, the fraction of customers lost, is decreasing and convex. We study the shape of $\{B_c\}$ for retrial queues. We show $B_{c+1} > B_c$ is possible. For arbitrary arrivals and exponential service, we show $\{B_c\}$ is decreasing, and report simulations where it is convex.

1. Introduction

Let customer i arrive at time t_i, $0 \leq t_1 \leq t_2 \cdots$, at a c-server system, where arriving customers who find all servers busy (called *blocked* customers) leave without service. In a conventional *loss system*, these customers never return and are lost. In a *retrial queue*, some of them return later to try again to find an idle server (a *retrial*). For the same customer, this may occur many times. At any time t, customers who have been blocked at least once and will return later are said to be *in orbit*. We assume that eventually, with probability 1, all customers are either served or lost, in the sense that they have not been served and will never return. Assume arrivals occur at finite rate $\lambda > 0$.

Of course, many real systems have a finite buffer (number of queue positions), where arrivals finding all servers busy enter the buffer (join the queue), provided it is not full. On each service completion, a customer in the buffer (if there are any) immediately leaves the buffer and begins service. Customers finding all servers busy and a full buffer are blocked. We have the same possibilities; either all blocked customers are lost (the conventional assumption) or some of them return later to try again. We will not consider buffered systems here.

Let *loss probability* B_c be the (long run) fraction of customers lost, in the sense defined above, when the limit exists. Number the servers 1, 2, . . ., and suppose that each customer's service time does not depend on the server. Without loss of generality, let each customer be assigned to the lowest-numbered server that is free, either on arrival, or if necessary, at a subsequent retrial time.

For loss systems, we assign customers this way to compare B_c as $c \geq 1$ varies. Clearly, $\{B_c\}$ is decreasing (non-increasing) in c because customers served by server $c+1$ are among those that would be lost when there are only c servers. $B_c - B_{c+1}$ is the reduction in the loss probability as we increase the number of servers from c to $c+1$. It has long been of great interest to know conditions under which this quantity also decreases as c increases, or equivalently, that $\{B_c\}$ is convex. Customers served by server $c+1$ account for reduction $B_c - B_{c+1}$, and consequently, convexity is equivalent to this: In the long run, and for every c, server c serves more (at least as many) customers as server $c+1$.

In retrial queues, server $c+1$ may serve customers who, on subsequent retrials, would otherwise be served by an earlier server. Unlike loss systems, adding server $c+1$ may change the fraction of customers served at earlier servers.

For loss systems, we now briefly review the literature on the convexity of $\{B_c\}$. In 1972, Messerli [4] showed that $\{B_c\}$ is *strictly* convex for the $M/G/c$ loss system. In 2002, Wolff and Wang [6] showed that $\{B_c\}$ is convex but not necessarily strictly convex under conditions we call **AI**: Arbitrary arrivals and i.i.d. service times, independent of the arrival process.

For loss systems, it is known that convexity may fail when service times are dependent. We present a simple example because we have not found one in the literature: Arrivals occur at times 0, 2, 4; 10, 12, 14; . . ., with corresponding service times 5, 1, 1; 5, 1, 1; . . .. We have $B_1 = 2/3$ and $B_2 = 0$.

There is a separate literature on loss systems with *ordered entry*, where the order of the servers is fixed, arrivals are served at the

* Corresponding author.
E-mail address: CWang@mail.ndhu.edu.tw (C.-L. Wang).

0167-6377/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2008.09.003

lowest-numbered server that is free, and the service distribution *may depend on the server*. When a customer enters service at server c, the service time is a random draw from some service distribution G_c, independent of all else.

For an ordered-entry loss system with renewal arrivals and exponential service, where service rate μ_c at server c is *decreasing* (non-increasing) in c, Yao [7] in 1986 showed that $\{B_c\}$ is strictly convex. Also see the proof on p. 135 of [3]. In an unpublished 1972 manuscript, Descloux [1] obtained the same result when the μ_c are all the same (and hence whether or not we have ordered entry).

Based on the method in [6], Papier et al. [5] showed in 2007 that $\{B_c\}$ is convex under these assumptions: Arbitrary arrivals that may have batches of customers of arbitrary batch sizes. Service times all have the same distribution, are independent of the arrival process and of each other from batch to batch, but may be dependent within the same batch.

They assumed *partial batch blocking* (PBB): When a batch of size b customers arrives to find $i < b$ idle servers, i customers are served, and $b - i$ are lost. Under *entire batch blocking* (EBB), all b customers are lost. They showed that under EBB, convexity may fail, even under **AI**. PBB was implicitly assumed in [6].

They also greatly extended the convexity result in [7] to an ordered-entry loss system with arbitrary arrivals and general service distributions, where G_c is stochastically increasing (non-decreasing) in c. (The exponential service distributions in [7] are stochastically increasing.)

In real queueing systems, not all blocked customers are lost (with or without a finite buffer). Consequently, there is a large literature on retrial queues, spawned in part by Falin and Templeton [2]. Most of this monograph and literature assume that all blocked customers are (eventually) served, ignoring the trade-off between reducing lost calls and increasing the number of servers. In fact, we are not aware of any results about the shape of $\{B_c\}$ in this literature.

We investigate the shape of $\{B_c\}$ for retrial queues. Under **AI**, we show in Section 2 that when there are c servers, the fraction of customers served at server u is a decreasing function of $u = 1, \ldots, c$. Unlike loss systems, however, this is not equivalent to convexity. By counterexample, we show that $B_{c+1} > B_c$ is possible. We also show that convexity may fail under either (i) constant service and inter-arrival times, or (ii) constant service times and sharply restricted retrial behavior.

In Section 3, we show that $\{B_c\}$ is decreasing for arbitrary arrivals, exponential service, and the usual model of orbiting customers in the literature.

In Section 4, we report simulation results, described there, where $\{B_c\}$ is decreasing and convex. We briefly discuss retrial queues with either batch arrivals or ordered entry in Section 5, and present concluding remarks in Section 6.

In the remainder of this Section, we briefly describe the method introduced in [6] to prove convexity for loss systems because we later adapt it to obtain a result for retrial queues:

For an ordered sequence of servers, arrivals blocked at server 1 *overflow* it, forming an *overflow process*, which is the arrival process at server 2, and so on. Each successive server c *thins* (removes some of the points) from the original arrival process. They compare servers c and $c + 1$, where both are idle initially.

The arrival process at c and the corresponding service times generate the thinned arrival process at $c + 1$ and the following quantities: X_n and Y_n, the respective times of the nth service completion at servers c and $c + 1$, $n \geq 1$, and $J(t)$ and $K(t)$, the respective number of service completions by time t at servers c and $c+1$, $t \geq 0$. They enlarge the sample space generated by the arrival and service times by introducing an i.i.d. sequence of random variables (distributed as service times), $S_1, S_2, \ldots$, independent of all else.

For an arbitrary sample path of arrivals and service times at server c, they generate the thinned arrival process at server $c + 1$ and the quantities above. For the same arrival process at server c and the thinned arrival process at server $c + 1$, they *replace* the service time of the nth customer served at server c and at server $c + 1$ (different random variables) by the *same* random variable S_n, for every $n \geq 1$. The arrival process at each server and $\{S_n\}$ generate the primed stochastic processes $\{X'_n\}$, $\{Y'_n\}$, $\{J'(t)\}$, and $\{K'(t)\}$ corresponding to the unprimed processes above.

They showed: Each primed stochastic process is stochastically equivalent to the corresponding unprimed process, and at every point in the sample space,

$$X'_n \leq Y'_n,\ n \geq 1, \quad \text{and} \quad J'(t) \geq K'(t), t \geq 0.$$

The result follows. The construction of the primed processes is an example of what is called *coupling*. This argument, which combines coupling and thinning, we call **CT**. The extensions in [5] were obtained by modifying this argument.

2. Loss probabilities in retrial queues

For retrial queues, we now investigate the shape of $\{B_c\}$ for a general retrial model. While Eq. (1) is promising, it turns out that without some restrictions on retrial behavior, $\{B_c\}$ may not be convex, and in fact may also *fail to be monotone decreasing*.

Suppose the *virtual* arrival process at server 1 consists of the (initial) arrival times t_i and retrial times. Customer i has a finite (possibly zero) number of ordered *virtual retrial times* $t_i + \delta_{ij}$, where $0 < \delta_{i1} < \cdots < \delta_{ik_i} < \delta_{i,k_i+1} = \infty$. Thus, k_i is the maximum number of retrials that customer i will attempt before leaving forever. A retrial by customer i will occur at time $t_i + \delta_{ij}$, $j \geq 1$, if and only if earlier attempts to find a free server failed. Let $V_i = \{\delta_{i1}, \ldots, \delta_{ik_i}\}$ be the collection of virtual retrial time increments of customer i.

We call this arrival process *virtual* because not all retrial times actually occur. When a customer is served at some station, not only is the corresponding arrival (or retrial) time removed from arrival processes at downstream stations, subsequent retrial times of that customer also are removed from the arrival process at each station, including station 1. Thus, upstream arrival processes also are thinned. To remove ambiguity about which arrivals are removed, we assume that with probability one, initial and virtual arrival (retrial) times are distinct. Every customer, on each attempt, is served by the lowest-numbered server that is free. Number customers in order of their initial arrival times.

When there are c servers, let f_{uc} be the fraction of customers served by server u, $u = 1, \ldots, c$, where we assume these quantities exist. We assume that with probability 1, all customers are (eventually) lost or served. Under these assumptions, the fraction of customers lost (when there are c servers) is well defined as $B_c = 1 - (f_{1c} + \cdots + f_{cc})$, where $B_0 = 1$.

We assume **AI**: Arbitrary initial and virtual arrivals, and i.i.d. service times, independent of the initial and virtual arrivals.

Fix the number of servers c. An arbitrary sample path of initial and virtual arrivals at server 1 and the corresponding service times generate the thinned (*actual*) arrival process and service-completion times at each station. For any $u < c$, the arrival process at station $u + 1$ is a thinning of the arrival process at station u. Applying the **CT** argument to stations u and $u + 1$, we have $f_{uc} \geq f_{u+1,c}$, and

$$f_{1c} \geq \cdots \geq f_{cc}. \tag{1}$$

For loss systems, where f_{uc} is fixed for all $c \geq u$, (1) is equivalent to convexity.

Now let

$$r_c = \sum_{u=1}^{c-1} f_{u,c-1} - \sum_{u=1}^{c-1} f_{uc}$$

be the *reduction* in the fraction of customers served by servers $1, \ldots, c-1$ when the number of servers is increased from $c-1$ to c, where $r_1 = 0$. (Note that we have not shown there is a reduction, i.e., that $r_c \geq 0$.) With this notation,

$$B_{c-1} - B_c = f_{cc} - r_c.$$

We know $f_{12} \geq f_{22}$. If we also have $r_2 \geq 0$, it easily follows that

$$B_1 - B_2 \leq B_0 - B_1.$$

However, having $r_c \geq 0$ for all c is not sufficient for convexity. Even more can go wrong, as the following counterexamples show.

Counterexample 1. Service times are constant, $S = 1$. Customers occur in groups of 5: Initial arrival times are 0, 0.1, 1.3, 1.4, 2.2; 10, 10.1, 11.3, 11.4, 12.2; .... In each group, the middle three customers return at most once; the others do not return at all. Customers 2, 3, 4; 7, 8, 9; ... have respective virtual retrial times 1.1, 3.3, 4.4; 11.1, 13.3, 14.4; .... (Note: A group *is not* a batch.)

For $c = 1$, it is easy to see that every customer is served, in the order 1, 2, 5, 3, 4; 6, 7, 10, 8, 9; .... Hence, $B_1 = 0$.

For $c = 2$, server 1 serves customers 1, 3; 6, 8; ..., and server 2 serves customers 2, 4; 7, 9; .... Customers 5; 10; ... are lost. We have $B_2 = 0.2 > B_1$.

Adding a server *increased* the loss probability! Recall that for loss systems, the monotonicity of $\{B_c\}$ holds without making any stochastic assumptions. For retrial queues, we have shown that monotonicity may fail under **AI**.

Initial arrival and retrial patterns are "unusual" above. We now show that even with constant initial inter-arrival and service times, convexity may fail.

Counterexample 2. Initial inter-arrival times are constant $T = 1$, and service times are constant $S = 2.4$, where customer i has initial arrival time $t_i = i - 1$, $i \geq 1$. Customers 2, 7; 12, 17; ... return at most once, with respective virtual retrial times 2.5, 7.5; 12.5, 17.5; .... The others do not return at all.

For $c = 1$, server 1 serves customers 1, 2; 6, 7; .... $B_1 = 3/5$.

For $c = 2$, a pattern repeats every 15 customers. Server 1 serves customers 1, 4, 7, 10, and 13. Server 2 serves customers 2, 5, 8, 11, and 12. Customers 3, 6, 9, 14, and 15 are lost. $B_2 = 1/3$.

For $c = 3$, all customers are served; $B_3 = 0$. Thus, $\{B_c\}$ decreases, but $B_0 - B_1 = 6/15$, $B_1 - B_2 = 4/15$, and $B_2 - B_3 = 5/15$. Convexity fails!

In the next case, only the initial arrivals have an irregular pattern.

Counterexample 3. Service times are constant $S = 4$. Customers occur in groups of 5: Initial arrival times are 0, 0.4, 2.8, 3.2, 3.6; 10, 10.4, 12.8, 13.2, 13.6; .... All customers have the same retrial behavior. They return at most once, with constant virtual retrial time increments $\delta_{i1} = 8$.

For $c = 1$, server 1 serves customers 1, 2; 8, 6; 13, 11; 18, 16; ..., in the indicated order. Because of initial conditions, the pattern does not emerge until the second group. We have $B_1 = 1 - 2/5 = 3/5$.

For $c = 2$, server 1 serves customers 1, 6, 11, ..., and server 2 serves customers 2, 7, 12, ... Note that no returning customer is served. As each server serves 1/5 of the customers, we have $B_2 = 3/5$. Convexity fails again!

From the first two counterexamples, it appears that some restriction on retrial behavior is necessary in order to obtain convexity or even monotonicity. For example, suppose that the V_i are i.i.d. We model a version of this assumption in the next section. The third counterexample shows that restricting retrial behavior alone is not sufficient. We comment further about this in Section 6.

3. Monotonicity under restricted retrial model

Now restrict **AI** to **Restricted AI**: Arbitrary initial arrivals, i.i.d. V_i, independent of the initial arrivals, and i.i.d. service times, independent of all else.

In addition, we further restrict retrial behavior to the **Orbit Model**: Each customer's virtual times between successive attempts to find an idle server, given they occur, are i.i.d., with exponential distribution at rate γ. Independent of this, the probability of returning again after j failed attempts is α for every $j \geq 2$, and α_1 for $j = 1$. These assumptions are universal in the retrial-queue literature, usually with $\alpha_1 = \alpha = 1$, or $\alpha_1 < 1$ and $\alpha = 1$. We assume $\alpha_1 > 0$ and $0 \leq \alpha < 1$. At time $t = 0$, all servers are idle and no customers are in orbit.

With this formulation, it is not necessary to keep track of virtual retrial times explicitly. Instead, when there are c servers, let $O_c(t)$ be the number of customers in orbit at time t. Measured from t, the time until the next retrial is exponential at rate $O_c(t)\gamma$. A customer with initial arrival at t enters service if a server is idle; if all servers are busy, the customer enters orbit with probability α_1, increasing $O_c(t)$ by 1, and with probability $1 - \alpha_1$, the customer is lost forever, leaving $O_c(t)$ unchanged. A customer with a retrial at t enters service if a server is idle, decreasing $O_c(t)$ by 1; if all servers are busy, the customer reenters orbit with probability α, leaving $O_c(t)$ unchanged, and with probability $1-\alpha$, the customer is lost forever, decreasing $O_c(t)$ by 1. Notice that the service times of orbiting customers are no longer associated with particular arrivals; instead, they are random draws from the service distribution each time a customer enters service from orbit. We call $\{O_c(t)\}$ an *orbit process*.

For this queue, let $M_c(t)$ be the number of busy servers at time t and $D_c(t)$ be the number of service completions by time t. Initially, $M_c(0) = O_c(0) = 0$. We will prove that $\{B_c\}$ is monotone decreasing by comparing the service-completion processes for retrial queues with c and $c + 1$ servers. To simplify the analysis, we assume exponential service times at rate μ.

We couple these queues as closely as possible. For example, when one has r busy servers and the other has $s \geq r$, we generate service completions at rate $s\mu$ for the busier queue. A generated service completion is also a service completion at the other queue with probability r/s. The orbit processes are coupled in the same manner. For both queues, the same arrival (coupled retrial) finding all servers busy has identical behavior. Let $A(t) = D_{c+1}(t) - D_c(t)$, $N(t) = M_{c+1}(t) - M_c(t)$, and $P(t) = O_c(t) - O_{c+1}(t)$.

Monotonicity is an easy consequence of the following result:

Theorem 1. *For the coupled retrial queues with arbitrary arrivals and the stochastic assumptions in this section, these sample-path inequalities hold:*

(i) $P(t) \geq 0$,

(ii) $-c \leq N(t) \leq 1$,

(iii) $A(t) + N(t) \geq P(t)$, *and*

(iv) $A(t) \geq 0$, *for all* $t \geq 0$.

Proof. Proof is by induction. Because both queues and orbit processes start empty, $A(0) = N(0) = P(0) = 0$, and (i) through (iv) hold at $t = 0$. Along the sample path, $\{A(t)\}$, $\{N(t)\}$, and $\{P(t)\}$ are constant between jumps at times where one of three types of events occur: an external arrival, a retrial, and a service completion. Suppose an event occurs at time t, where these conditions hold at t^-. We will show that the conditions still hold at t (after the jump) by considering in turn each type of event. As this will occur several times, note that $N(t) = 1$ means that both queues have the same number of idle servers.

Suppose an external arrival occurs. If $N(t^-) = 1$, $N(t)$ and $P(t)$ are the same as $N(t^-)$ and $P(t^-)$, respectively. If $N(t^-) \leq 0$

with $M_c(t^-) < c$, then $N(t)$ and $P(t)$ remain the same; if instead $M_c(t^-) = c$, then $N(t) = N(t^-)+1$ and $P(t-) \leq P(t) \leq P(t^-)+1$ (the arrival may be lost). It is easy to check that under any of these changes, all of the conditions still hold at t.

Suppose a retrial occurs. Since $P(t^-) \geq 0$, a retrial comes either from both orbits or only from the c-server orbit (if $P(t^-) > 0$). When from both, the situation is the same as that of an external arrival. When from one, with $M_c(t^-) = c$, $N(t) = N(t^-)$ and $P(t^-) - 1 \leq P(t) \leq P(t^-)$. If $M_c(t^-) < c$ (note that $M_c(t^-) < c$ implies $N(t^-) > -c$), we have $N(t) = N(t^-) - 1 \geq -c$ and $P(t) = P(t^-) - 1 \geq 0$, and all of the conditions still hold at t.

Finally, suppose a service completion occurs. If it occurs at both queues, $N(t)$ and $A(t)$ do not change. If it occurs only at the $(c+1)$-server queue (only if $N(t^-) = 1$), then $N(t) = N(t^-) - 1$ and $A(t) = A(t^-) + 1$, and all of the conditions hold. If it is from the c-server queue (only if $N(t^-) < 0$, which, in turns, implies $A(t^-) > 0$ by (iii)), then $N(t) = N(t^-) + 1$ and $A(t) = A(t^-) - 1 \geq 0$, and, again, all the conditions hold. □

$A(t) \geq 0$ is the number of *additional* service completions at the $(c + 1)$-server queue by t. When the limits exist, where the B_c do not depend on the coupling,

$$\lambda(B_c - B_{c+1}) = \lim_{t\to\infty} A(t)/t \geq 0, \; c \geq 1, \text{ which is}$$

Theorem 2. *For retrial queues with arbitrary arrivals and the stochastic assumptions in this section, $\{B_c\}$ is monotone decreasing.*

When there are no losses ($\alpha_1 = \alpha = 1$), inequalities (i)–(iv) still hold, where (iii) becomes $A(t) + N(t) = P(t)$ for all t. Service completions occur sooner for the $(c + 1)$-server queue, but when both systems are stable, $B_c = B_{c-1} = 0$, and $A(t)/t \to 0$ as $t \to \infty$.

4. Simulation results

Retrial queues were simulated under renewal initial arrivals, i.i.d. service times, and the **Orbit Model**, for combinations of exponential, constant, and hyper-exponential inter-arrival and service distributions. In every case, $\{B_c\}$ is convex as well as monotone. For further discussion, see Section 6.

5. Batch arrivals or ordered entry

For retrial queues with batch arrivals, customers from the same batch may be served by the same server. Under the model in [5], **CT** does not apply; it requires that the service times of served customers at each server be i.i.d.

Now suppose we have ordered entry with c servers, where G_u is the service distribution at server u, and the G_u are stochastically increasing. Following [5], we enlarge the sample space by introducing an i.i.d. sequence of random variables $U_1, U_2, \cdots$ that are uniformly distributed on (0, 1), independent of all else. To construct the corresponding primed processes, S_n^u, the service time of the nth customer served at server u, is generated by $S_n^u = G_u^{-1}(U_n)$.

From the stochastic ordering, $S_n^u \leq S_n^{u+1}$ for every n at every sample point, the modified **CT** argument goes through, and we again have (1).

6. Concluding remarks

For loss systems, $\{B_c\}$ is decreasing and convex under **AI** (arbitrary arrivals, i.i.d. service times), very general conditions that are likely to hold in practice. Having arbitrary arrivals allows some customers to arrive in batches and includes renewal arrivals as a special case. A key property in the proof of this result is the downstream thinning of the arrival processes at successive stations.

When the research reported here began, the shape of $\{B_c\}$ for loss systems was well understood, but was a blank slate for retrial queues. All results we report for retrial queues are new; we comment on some of them below.

The upstream thinning of the arrival process greatly complicates the analysis, and more can go wrong. In the first counterexample, which satisfies **AI**, $B_2 > B_1$. For loss systems, this is impossible!

To obtain "positive" theoretical results, we restrict **AI** through retrial behavior. **Restricted AI** has arbitrary initial arrivals, i.i.d. retrial behavior from customer to customer, and i.i.d. service times. Note that in the third counterexample, which satisfies **Restricted AI**, $\{B_c\}$ is decreasing but not convex.

In Section 3, we further restrict retrial behavior to the **Orbit Model**, as described there, and show, under the additional assumption of exponential service, that $\{B_c\}$ is decreasing. While this result holds only for a special case of **Restricted AI**, it is important to note that the initial arrivals are arbitrary.

We also investigated the shape of $\{B_c\}$ through simulation, where it is decreasing and convex. Does this hold for arbitrary service distributions (an important question) and/or for arbitrary inter-arrival distributions (a less important question)? Less important because we doubt that the renewal restriction will be helpful in any proof and also because of shortcomings mentioned below.

Our simulation runs are limited in two ways, by the **Orbit Model**, which may smooth out retrial behavior too much, and by renewal initial arrivals. While a popular assumption in queueing *theory*, (non-Poisson) renewal arrivals is rarely a good approximation of an arrival process in practice. Arrivals are often said to be "bursty", with various meanings given to this term. The third counterexample, where initial arrivals cluster in groups of five, with gaps in between, should serve as a warning that convexity may sometimes fail in practice.

References

[1] A. Descloux, On the Convexity of the Erlang Loss-Function, Bell Laboratories Technical Memorandum #72-3422-3, 1972 unpublished.

[2] G.I. Falin, J.G.C. Templeton, Retrial Queues, Chapman & Hall, London, 1997.

[3] J. Liyanage, J.G. Shanthikumar, Second-order properties of single-stage queueing systems, in: U.N. Bhat, I.V. Basawa (Eds.), Queueing and Related Models, Clarendon Press, Oxford, 1992, pp. 129–160.

[4] E.J. Messerli, Proof of a convexity property of the Erlang B formula, Bell Syst. Tech. J. 51 (1972) 951–953.

[5] F. Papier, R.W. Wolff, U.L. Thonemann, On the convexity of loss probabilities when arrivals occur in batches and servers are heterogeneous, Oper. Res. Lett. 35 (2007) 467–471.

[6] R.W. Wolff, C.-L. Wang, On the convexity of loss probabilities, J. Appl. Probab. 39 (2002) 402–406.

[7] D.D. Yao, Convexity properties of the overflow in an ordered-entry system with heterogeneous servers, Oper. Res. Lett. 5 (1986) 145–147.