A Study of a Loss System with Priorities

Yang, Hang; Fu, Jing; Wu, Jingjin; Zukerman, Moshe

Published in:
Heliyon

Published: 30/08/2024

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1016/j.heliyon.2024.e36109

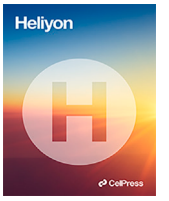
Publication details:
Yang, H., Fu, J., Wu, J., & Zukerman, M. (2024). A Study of a Loss System with Priorities. *Heliyon*, *10*(16), Article e36109. https://doi.org/10.1016/j.heliyon.2024.e36109

Heliyon 10 (2024) e36109

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A study of a loss system with priorities

Hang Yang [a], Jing Fu [b], Jingjin Wu [c,d,*], Moshe Zukerman [a]

[a] Department of Electrical Engineering, City University of Hong Kong, Kowloon, Hong Kong Special Administrative Region of China
[b] School of Engineering, RMIT University, Melbourne, Victoria 3000, Australia
[c] Department of Statistics and Data Science, BNU-HKBU United International College, Zhuhai, Guangdong, 519087, PR China
[d] Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, Guangdong, 519087, PR China

ARTICLE INFO

Keywords:
Loss system
Erlang B system
Preemptive priorities
Multidimensional Markov chain
Insensitivity

ABSTRACT

The Erlang loss formula, also known as the Erlang B formula, has been known for over a century and has been used in a wide range of applications, from telephony to hospital intensive care unit management. It provides the blocking probability of arriving customers to a loss system involving a finite number of servers without a waiting room. Because of the need to introduce priorities in many services, an extension of the Erlang B formula to the case of a loss system with preemptive priority is valuable and essential. This paper analytically establishes the consistency between the global balance (steady state) equations for a loss system with preemptive priorities and a known result obtained using traffic loss arguments for the same problem. This paper, for the first time, derives this known result directly from the global balance equations based on the relevant multidimensional Markov chain. The paper also addresses the question of whether or not the well-known insensitivity property of the Erlang loss system is also applicable to the case of a loss system with preemptive priorities, provides explanations, and demonstrates through simulations that, except for the blocking probability of the highest priority customers, the blocking probabilities of the other customers are sensitive to the service time distributions and that a larger service time variance leads to a lower blocking probability of the lower priority traffic.

1. Introduction

In many applications, there are service resources abstracted as *servers*, and incoming customers that are either served by available service resources, blocked and cleared from the system, or overflowed to another service system. Such systems are called *loss systems* and are characterized by a potential lack of waiting room. Examples of service resources in such a system include telephone circuits, wireless channels, optical wavelengths, and hospital beds of intensive care units. In this paper, we often use the term "customers" to refer to a range of customers or customers' service requests seeking appropriate services, such as phone calls, customers' service requests for wireless or wavelength channels, as well as actual customers, patients, clients, or travelers.

A loss system where arriving customers follow a Poisson process and their service times are independent and exponentially distributed is denoted by M/M/k/k (Kendall notation [1]) where the first k represents the number of servers, and the second k is the maximal number of service requests allowed in the system or the number of buffer places including the buffer places at the servers. Throughout the paper, the terms "service time" and "holding time" are considered synonymous and will be used interchangeably.

* Corresponding author at: Department of Statistics and Data Science, BNU-HKBU United International College, Zhuhai, Guangdong, 519087, PR China.
E-mail address: jj.wu@ieee.org (J. Wu).

https://doi.org/10.1016/j.heliyon.2024.e36109
Received 31 January 2024; Received in revised form 7 August 2024; Accepted 9 August 2024

Available online 14 August 2024

Applications of loss systems in general, and M/M/k/k in particular, include critical systems such as Intensive Care Units (ICU), where the availability of a service resource may have life-and-death consequences [2]. In other applications, such as those in the areas of computers and communications systems and networks, excessive blocking events have adverse consequences on customers' Quality of Service (QoS). Erlang Loss Formula (also known as Erlang B Formula) is a century-old solution for the blocking probability of the arriving customers which has been widely applied in many areas, such as hospital resource allocation [3–8], telephony [9,10], mobile networks [11–16], video on demand [17,18], call centers [19], parallel computing systems [20], self-driving cars [21] and optical networks [22–26]. Over the years, there have been publications that considered extensions of loss systems beyond Erlang B; they include [27–31].

The Erlang Loss Formula is also known to be insensitive to the service time distribution beyond its mean [32]. In other words, the formula gives exact blocking probability results regardless of whether or not the service time is exponentially distributed, so as long as the mean service time is known, its actual distribution does not affect the resulted blocking probabilities, and the formula provides exact results for any service time distribution. This insensitivity property is key to the broad applicability of the Erlang Loss Formula in many practical cases where the service time does not follow an exponential distribution.

In various applications, services are differentiated according to various criteria. For example, in a hospital triage process, patients are differentiated according to the severity of their conditions [33]. In other service systems, such as packet switching networks [34,35] and downlink transmission in 5G mobile networks for ultra-reliable low latency communication (URLLC) traffics [36], customer requests may be differentiated according to the cost and/or their QoS requirements. In such cases, implementation of preemptive priorities is an option [37–39].

Another application of preemptive priorities is to approximate blocking probabilities in overflow loss systems by constructing a surrogate system with preemptive priorities [16,40–43]. However, in this paper, we focus on deriving exact evaluations of blocking probabilities from the global balance equations based on multidimensional Markov Chains and thus have a fundamentally different objective compared to the studies similar to [16,40–43].

We consider a Markovian loss system with a finite number of servers where arriving customers are classified into a finite number of preemptive-priority classes. The arrival process of each class of customers follows a Poisson process, and all the service times are independent and exponentially distributed with a given parameter. In this Markovian loss system, if a customer arrives and all the servers are busy, the arriving customer may preempt a customer in the system that is being served if the arriving customer has higher priority than the preempted customer. Note that the focus of this paper is on loss systems where the blocking probability is the key performance measure. Such systems are different from delay systems such as M/M/1 and M/G/k where arriving customers are never blocked and the key performance measures are delay and queue size statistics. The problem we consider is to find the blocking probability of customers from each of the preemptive priority classes.

The first to consider this problem was Katzschner [44] in 1970. He considered the global balance equations of this loss system with preemptive priorities. Then, he derived the Laplace transform of the joint steady-state distribution of the number of customers of each class for the case of two classes from which the blocking probabilities can be obtained. Finally, he explained how this solution could be extended to the general case of a finite number of customer classes. Vu and Zukerman [45], who were not aware of [44], provided a simple procedure based on rate arguments to derive the blocking probabilities for all priority classes in the context of optical burst switching (OBS) application. Then, Yang and Stol [46] (who were not aware of [44] and [45]) obtained the blocking probabilities numerically from the global balance steady-state equations. None of the previous papers that dealt with this loss system with preemptive priorities established the equivalence between the global balance steady-state equations and the simple solution of [45] analytically. In this paper, we do precisely that; we start from the global balance equations, and rigorously establish their equivalence to the solution of [45].

Another important contribution of this paper is to address the fundamental question of whether or not the system we consider, i.e., a loss system with preemptive priorities, is insensitive to service time distribution beyond its mean (i.e. to the shape of the service time distribution) of the customers of various service classes. As mentioned above, it is well known that the Erlang B formula is insensitive to the shape of the service time distribution [32]. The question of whether or not this insensitivity property is also applicable to the blocking probabilities of the various classes of customers in a loss system with multiple priorities has not been properly addressed for over half a century since this problem was first introduced by Katzschner [44] in 1970. Maslova and Tatashev [47] considered the case of a single channel loss system with priorities and derived the approximated blocking probability of customer requests for each priority class, and the approximated results become exact when the service time distributions of the requests do not depend on their priorities. It is also demonstrated in [47] that the blocking probabilities are not sensitive to service time distributions beyond the mean for the single channel loss system. In this paper, we explain and demonstrate by simulations that, in a multi channel loss system with priorities, for the highest preemptive priority that is not affected by lower priority traffic the insensitivity of the Erlang B system applies. However, for lower preemptive priorities, insensitivity does not hold, and that higher variance of the service time distribution leads to lower blocking probabilities of the lower priority traffic. This has not been demonstrated in the earlier publications on this problem, including [44–46]. In fact, in [45], an incorrect comment was made that implies the insensitivity property for all priority classes.

The remainder of the paper is organized as follows. In Section 2, we describe the model and provide the solution of [45] for the blocking probability of the customers of the different priority classes. In Section 3, we provide a multi-dimensional Markov-chain analysis for our problem and establish consistency with the results of [45] presented in Section 2. In Section 4, we explain and demonstrate by simulations that except for the blocking probability of the highest priority customers, the blocking probabilities of the other customers are sensitive to the holding time distributions. In Section 5, we demonstrate the performance behavior of the system when certain parameters change. Finally, the conclusions are drawn in Section 6.

2. Model description and the solution of [45]

This paper considers the fundamental problem of obtaining the blocking probability in a loss system where the customers are classified into p preemptive priority classes and the arrival process of priority i customers (for $i = 1, 2, 3, \ldots, p$) follows a Poisson process with parameter λ_i. The service time of the customers of all priorities is exponentially distributed with parameter μ. Accordingly, the offered traffic of the customers of priority i is

$$A_i = \frac{\lambda_i}{\mu}, \quad i = 1, 2, 3, \ldots, p.$$

Priority 1 is the highest priority, and priority p is the lowest. If $i > j$, then customers of priority i have lower priority than customers of priority j. In this case, an arriving priority j customer, at the time of its arrival, may preempt a priority i customer already being served. In the case, where there are more than one customer with lower priority than priority j in service, the one with the latest arrival time among the lowest priority in service will be preempted. Note that this is different from the assumptions in some existing studies (e.g., [48]) where the customers arriving earlier will be preempted first.

This section provides the solution of Vu and Zukerman [45] based on rate arguments for the blocking probabilities of customers of each priority class. This solution is also described in [49].

Let $P_{\mathrm{b}}(i)$ denote the blocking probability of the customer class with priority i. Because priority 1 customers may preempt lower priority customers, they can access the loss system as if these lower priority customers do not exist. Therefore, for $i = 1$, we obtain

$$P_{\mathrm{b}}(1) = E_k(A_1),$$

where $E_k(A)$ denotes the blocking probability obtained by the Erlang B formula for an M/M/k/k loss system with k servers and offered traffic A. Specifically,

$$E_k(A) = \frac{A^k/k!}{\sum_{m=0}^{k} \frac{A^m}{m!}}.$$

To obtain $P_{\mathrm{b}}(i)$ for $i = 2, 3, \ldots, p$, the first step is to observe [45,49] that the blocking probability of the total traffic from priority i and higher, which is the traffic generated by customers of priorities $1, 2, \ldots, i$, is equal to:

$$E_k\left(\sum_{j=1}^{i} A_j\right).$$

To explain this observation, we need to consider the fact that the overall blocking probability in the case of a loss system with a given number of servers and with a strict priority regime among multiple classes of traffic streams all having the same mean service time is equal to the blocking probability of a traditional loss system (with a single traffic class) with the same number of servers and the same mean service time as in the case of the multiple traffic classes. This fact holds only because, in both cases, the service time distribution is exponential, and because this distribution is memoryless, a higher priority call that arrives when all servers are busy, which will be blocked in a single class loss system, will preempt a low property priority call in a multiple class loss system and will have the same remaining service time as that of the preempted low priority call. As for both calls, the remaining service time has an exponential distribution with mean $1/\mu$. Accordingly, in both systems, one blocked call will be recorded, and the remaining service times are also equal.

Henceforth, we will call this justification the *displacement/replacement argument*, recalling the term "displacing priorities" of [44]. Notice that using the displacement/replacement argument ignores the potential discrepancy in the priority system by replacing a preempted low-priority call with a high-priority call.

However, the above-described method does not apply to cases where the service times are not exponentially distributed. In other words, except for the highest priority traffic, blocking probabilities results are not insensitive to the shape of holding time distribution (i.e., to the distribution beyond its mean). Unlike the conventional M/M/k/k system, the blocking probabilities of the preempted customers with lower priorities are sensitive to the shape of their holding time distributions. This will be further discussed, explained, and demonstrated by simulations in Section 4. This will correct [45] which implies that the insensitivity property applies to all priority classes.

Next, we observe that the priority i lost traffic, $i = 2, 3, \ldots, p$, denoted $A_L(i)$ is given by the lost traffic of priorities up to i minus the lost traffic of priorities up to $i - 1$, namely,

$$\begin{aligned} A_L(i) = &\left(\sum_{j=1}^{i} A_j\right) E_k\left(\sum_{j=1}^{i} A_j\right) \\ &- \left(\sum_{j=1}^{i-1} A_j\right) E_k\left(\sum_{j=1}^{i-1} A_j\right). \end{aligned} \tag{1}$$

Therefore, the $P_{\mathrm{b}}(i)$ values for $2 \le i \le p$, are obtainable as the ratio of the priority i lost traffic to the priority i offered traffic, that is,

$$P_{\mathrm{b}}(i) = \frac{A_L(i)}{A_i}. \tag{2}$$

3. Multi-dimensional Markov-chain analysis

In this section, we provide a multi-dimensional Markov-chain analysis of our problem of a loss system with priorities. We start with the simpler case of two preemptive priorities. Then, we extend the analysis to the general case of k servers and p preemptive priorities.

3.1. The case of two preemptive priorities

We consider here the case of two classes of customers with two corresponding preemptive priorities. A similar analysis was also provided in [46], but we include it here for completeness. The priority class i customers are assumed to arrive following a Poisson process with parameter λ_i, $i = 1, 2$. We consider a k-server loss system where customers are served by k servers, and there is no waiting room. We assume that the service times of all the customers are exponentially distributed with the parameter μ. The Class 1 customers have preemptive priority over the Class 2 customers. That is, an arriving Class 1 customer that finds all k servers busy may preempt a Class 2 customer in service and is served instead of the preempted Class 2 customer. Accordingly, a Class 1 (higher priority) customer will only be blocked if, upon its arrival, all the k servers are busy serving Class 1 customers, while a Class 2 customer will be blocked on arrival if all k servers are busy serving any of the customer classes, and may also be preempted during its service. Therefore, the offered traffic by Class 1 customers is given by

$$A_1 = \lambda_1/\mu\ ,$$

and the offered traffic by Class 2 customers is

$$A_2 = \lambda_2/\mu.$$

Then, the total offered traffic is

$$A = A_1 + A_2.$$

Let (i, j) be the system's state, where i represents the number of busy servers for Class 1 customers and j is the number of busy servers for Class 2 customers. Thus, in state (i, j), $i = 0, 1, 2, \ldots, k$, $j = 0, 1, 2, \ldots, k$, we must have

$$0 \le i + j \le k.$$

Let $\pi_{i,j}$ be the steady-state probability of the system being in state (i, j). Accordingly,

$$\sum_{i=0}^{k} \sum_{j=0}^{k-i} \pi_{i,j} = 1.$$

Recall that $P_{\mathrm{b}}(1)$ represents the blocking probability of priority 1 customers (Class 1 customers), and $P_{\mathrm{b}}(2)$ is the blocking probability of priority 2 customers (Class 2 customers). Note that $P_{\mathrm{b}}(2)$ is the probability that an arriving priority 2 customer is either blocked on arrival or preempted by a priority 1 customer after it was admitted to service. Since priority 1 customers will only be blocked when all the k servers are busy, and there are no priority 2 customers in service, we have

$$P_{\mathrm{b}}(1) = \pi_{k,0}.$$

Let $P_{\mathrm{boa}}(2)$ denote the probability that an arriving priority 2 customer finds all the servers busy and is blocked on arrival (boa), and let $P_{\mathrm{pre}}(2)$ be the probability that an arbitrary priority 2 customer is lost due to the preemption of its service by an arrival of a priority 1 customer.

To obtain $P_{\mathrm{boa}}(2)$, we observe that since the arriving priority 2 customers follow a Poisson process, and since an arriving priority 2 customer will be blocked on arrival if and only if all k servers are busy, we have

$$P_{\mathrm{boa}}(2) = \sum_{i=0}^{k} \pi_{i,k-i}. \tag{3}$$

An alternative approach to assert (3) is to consider a sufficiently long period of time L. Then, the mean number of priority 2 arrivals during L is $\lambda_2 L$, and the mean number of priority 2 customers that are blocked on arrival during period L is given by $\lambda_2 L \sum_{i=0}^{k} \pi_{i,k-i}$. Accordingly, the ratio of the latter to the former gives the proportion of priority 2 customers that are blocked on arrivals.

To obtain $P_{\mathrm{pre}}(2)$, we again consider a sufficiently long period of time L, and we already know that the mean number of priority 2 customers arriving during period L, is equal to $\lambda_2 L$ out of which $P_{\mathrm{boa}}(2)\lambda_2 L$ were blocked on arrival. We also observe that the mean number of priority 2 customers that are preempted during L is equal to the number of priority 1 arrivals during the periods of

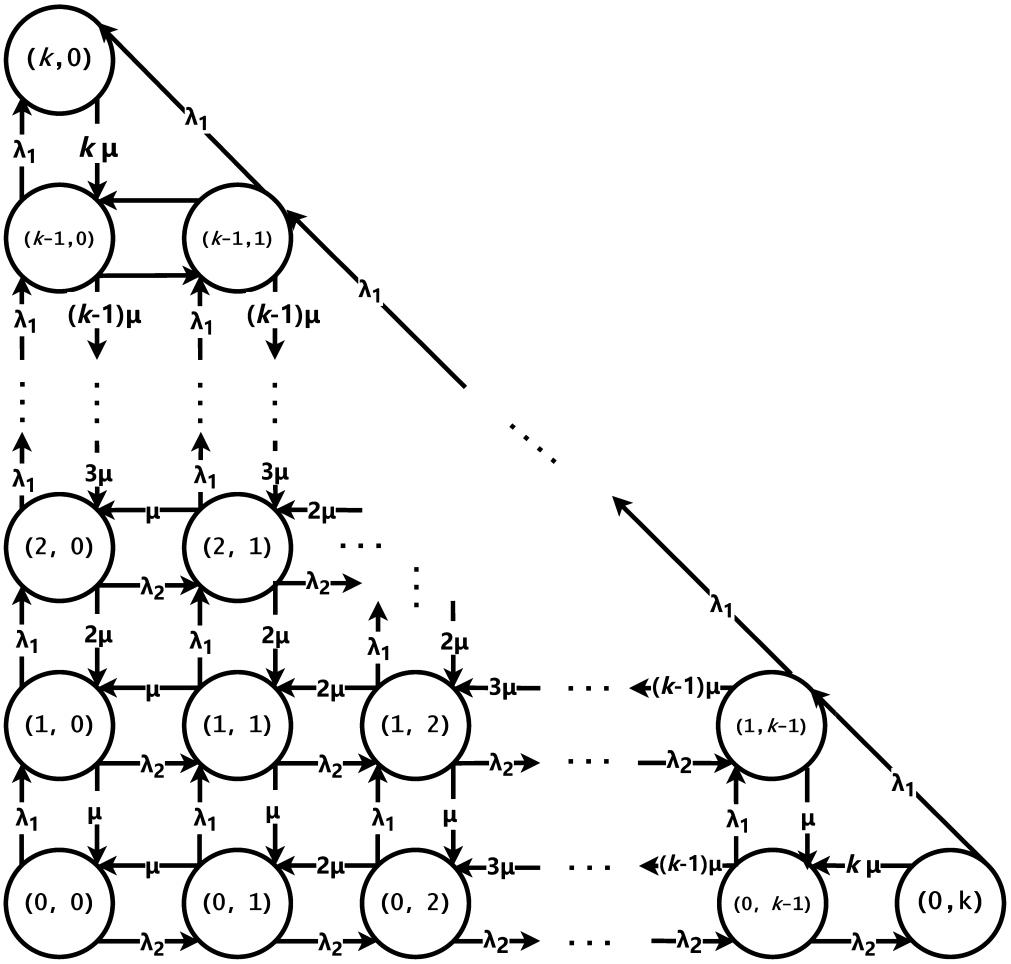


Fig. 1. The state transition diagram for a k-server loss system with 2 preemptive priorities.

times in L that all servers are busy, and there are some priority 2 customers being served because each of these priority 1 arrivals will preempt a priority 2 customer that is being served. This mean is given by

$$\lambda_1 L \sum_{i=0}^{k-1} \pi_{i,k-i}.$$

Therefore, $P_{\text{pre}}(2)$ is given by the ratio of the latter to the mean number of priority 2 customers that arrive during L but are not blocked on arrival, so we obtain

$$P_{\text{pre}}(2) = \frac{\lambda_1}{\lambda_2(1 - P_{\text{boa}}(2))} \times \sum_{i=0}^{k-1} \pi_{i,k-i}.$$

Having obtained $P_{\text{pre}}(2)$ and $P_{\text{boa}}(2)$, the overall blocking probability of a priority 2 customer is the probability that it is blocked on arrival ($P_{\text{boa}}(2)$), plus the probability that it is not blocked on arrival but lost due to the preemption of its service by an arrival of a priority 1 customer ($(1 - P_{\text{boa}}(2))P_{\text{pre}}(2)$). Thus,

$$P_{\text{b}}(2) = P_{\text{boa}}(2) + (1 - P_{\text{boa}}(2))P_{\text{pre}}(2)$$

In the case of a general number of servers (k), the balance equations can be constructed by considering two types of states (refer to Fig. 1): (1) non-congestion states where at least one server is idle, and (2) congestion states where all servers are busy.

In particular, for the non-congestion states the i, j satisfy
$0 \le i < k, 0 \le j < k, 0 \le i + j < k$.

These non-congestion states obey the following global balance equations,

$$\begin{aligned}\pi_{i,j}\times\left(\lambda_1+\lambda_2+(i+j)\mu\right)&=\pi_{i-1,j}\times\lambda_1+\pi_{i,j-1}\times\lambda_2\\&+\pi_{i+1,j}\times(i+1)\mu\\&+\pi_{i,j+1}\times(j+1)\mu.\end{aligned}\tag{4}$$

The congestion states can further be classified into two types: Type 1 and Type 2.

Type 1 congestion states refer to the states in which a new arrival will be blocked only if it is of Priority Class 2. It includes all the congestion states except where $i=k$. In these states, $\{i,j\}$ satisfy $0\le i<k, 0\le j\le k, i+j=k$. These Type 1 states obey the following global balance equations.

$$\begin{aligned}\pi_{i,j}\times(\lambda_1+(i+j)\mu)&=\pi_{i-1,j}\times\lambda_1+\pi_{i,j-1}\times\lambda_2\\&+\pi_{i-1,j+1}\times\lambda_2.\end{aligned}$$

There is only one Type 2 state, namely when $i=k, j=0$, in which a new arrival will always be blocked irrespective of its priority class. The state itself obeys the following global balance equation,

$$\pi_{k,0}\times k\mu=\pi_{k-1,1}\times\lambda_1+\pi_{k-1,0}\times\lambda_1.$$

Note also that $\pi_{i,j}=0$ for all cases where $i<0$ or $j<0$, and all state probabilities should sum to 1, that is,

$$\sum_{i=0}^{k}\sum_{j=0}^{k-i}\pi_{i,j}=1.\tag{5}$$

By assuming statistical equilibrium, we can obtain the state probabilities $\pi_{i,j}$ for all possible pairs of i and j based on (4) and (5).

The priority 1 traffic will only be blocked in the Type 2 state ($i=k, j=0$) where all the k servers are busy serving priority 1 customers. That is,

$$P_{\mathrm{b}}(1)=\pi_{k,0}=\frac{(A_1)^k/k!}{\sum_{i=0}^{k}\frac{A_1^i}{i!}}.$$

For priority 2 customers, we follow similar rationales as in the previous two subsections. That is, a priority 2 customer will be blocked on arrival if the system is in either a Type 1 or Type 2 congestion state. In addition, a priority 2 customer being served will be preempted by a newly arriving priority 1 customer if the system is in a Type 1 congestion state. We again use $P_{\mathrm{boa}}(2)$ and $P_{\mathrm{pre}}(2)$ to represent the probabilities of priority 2 customers being blocked on arrival and preempted after being admitted to service, respectively, and derive the overall blocking probability for them as,

$$\begin{aligned}P_{\mathrm{b}}(2)&=P_{\mathrm{boa}}(2)+(1-P_{\mathrm{boa}}(2))P_{\mathrm{pre}}(2)\\&=\sum_{i=0}^{k}\pi_{i,k-i}+\frac{\lambda_1}{\lambda_2}\times\left(\sum_{i=0}^{k}\pi_{i,k-i}-\pi_{k,0}\right)\\&=\pi_{k,0}+(1+\frac{\lambda_1}{\lambda_2})\times\left(\sum_{i=0}^{k}\pi_{i,k-i}-\pi_{k,0}\right)\\&=\frac{A_1^k/k!}{\sum_{i=0}^{k}\frac{A_1^i}{i!}}+\left(\frac{\lambda_1+\lambda_2}{\lambda_2}\right)\\&\times\left(\frac{(A_1+A_2)^k/k!}{\sum_{i=0}^{k}\frac{(A_1+A_2)^i}{i!}}-\frac{A_1^k/k!}{\sum_{i=0}^{k}\frac{A_1^i}{i!}}\right)\\&=\frac{(A_1+A_2)E_k(A_1+A_2)-A_1E_k(A_1)}{A_2}.\end{aligned}$$

Again, this is consistent with the results in Section 2.

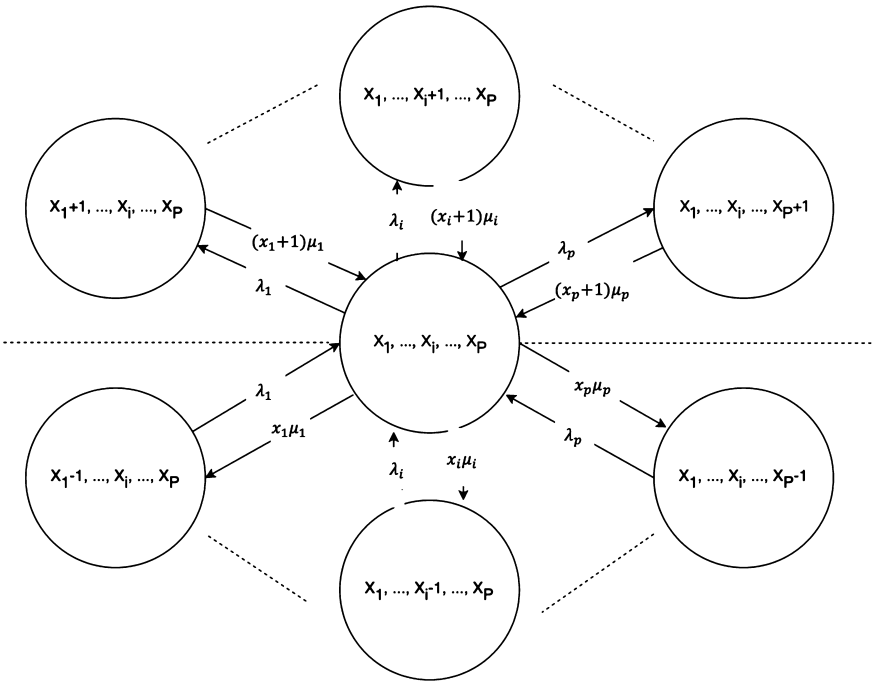


Fig. 2. The transition diagram of non-congestion states for a k-server loss system with p preemptive priorities.

3.2. Extension to p preemptive priorities

For the M/M/k/k model, the k-server loss system has a Poisson arrival process with rate λ and exponentially distributed service time with intensity μ. With p priorities (the smallest number represents the highest priority), the total offered traffic is

$$A = \sum_{i=1}^{p} A_i.$$

The vector $\mathbf{X} = (..., X_i, ...)$ denotes the state of the system where X_i represents the number of busy servers for priority class i traffic. Accordingly, i and X_i are bounded as follows:

$$1 \le i \le p$$
$$0 \le X_i \le k$$
$$0 \le \sum_{i=1}^{p} X_i \le k.$$

Let $\pi(X_1..., X_i, ..., X_p)$ be the steady-state probability that there are X_i busy servers of priority class i, for $i = 1, 2, ..., p$. Then we have:

$$\sum_{\substack{X_i \in [0,k], i \in [1,p] \\ 0 \le \sum_{i=1}^{p} X_i \le k}} \pi(X_1..., X_i, ..., X_p) = 1.$$

All the states can be divided into two types: (1) *non-congestion states* where for every $i = 1, 2, \ldots, p$, $0 \le X_i < k$, and $0 \le \sum_{i=1}^{p} X_i < k$ (called "normal" states in [46]); and (2) *congestion states* where for every $i = 1, 2, \ldots, p$, $0 \le X_i \le k$, and $\sum_{i=1}^{p} X_i = k$ (called "boundary" states in [46,50]).

In Fig. 2, we provide the transition diagram of the k-server loss system with p preemptive priorities in non-congestion states (a similar figure is provided in [46]). This leads to the following balance equations for the non-congestion states:

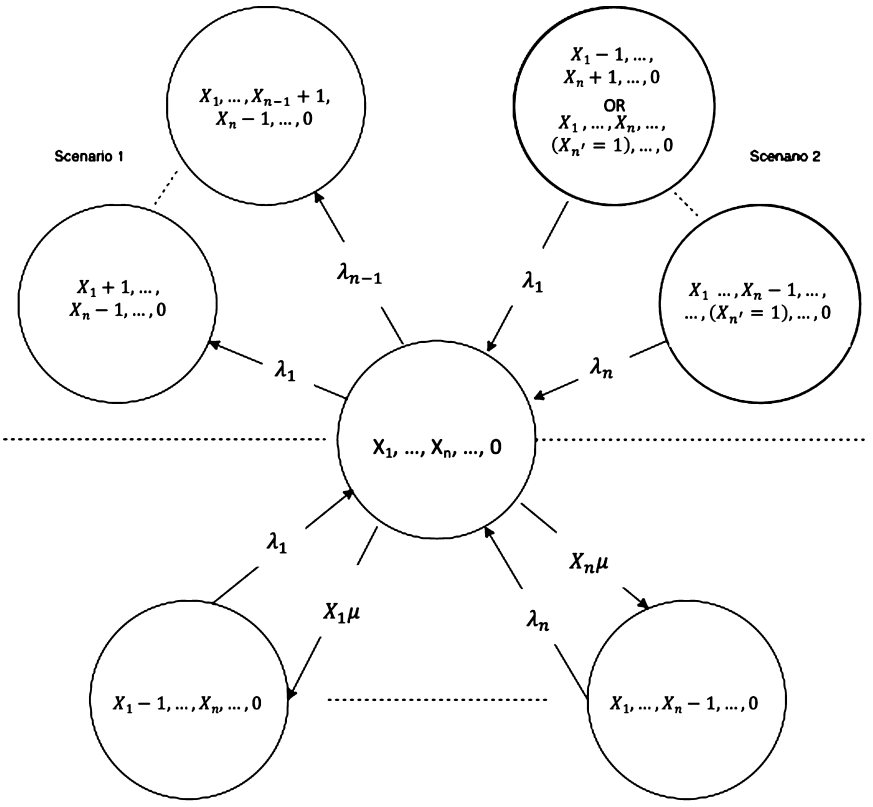


Fig. 3. The transition diagram of a boundary state $\mathbf{X}$, where $1 < l(\mathbf{X}) = n < p$, for a k-server loss system with p preemptive priorities.

$$\pi(..., X_i, ...)\left(\sum_{i=1}^{p}\lambda_i + \sum_{i=1}^{p} X_i\mu\right) = \sum_{i=1}^{p}\pi(..., X_i + 1, ...)(X_i + 1)\mu + \sum_{i=1}^{p}\pi(..., X_i - 1, ...)\lambda_i, \tag{6}$$

where the expression $\sum_{i=1}^{p}\pi(..., X_i + 1, ...)$ denotes all the possible states that can transit in one hop to $(..., X_i, ...)$ by a departure. Similarly, the expression $\sum_{i=1}^{p}(..., X_i - 1, ...)$ represents all the possible states that can transit in one hop to $(..., X_i, ...)$ by an arrival. Note that $\pi(..., X_i, ...) = 0$ if at least one of the X_i values is negative.

Fig. 3 illustrates the transition diagram in boundary states. For illustration, we use $l(\mathbf{X})$ to denote the lowest priority of all customers in service in a boundary state $\mathbf{X}$. For example, $l(X_1, ..., X_n, 0, ..., 0) = n$ if $X_n > 0$ and $X_i = 0$ for all $i > n$. On the top half of the diagram, we point out two preemption scenarios for a boundary state $\mathbf{X}$ when $1 < l(\mathbf{X}) = n < p$. In scenario 1, $\mathbf{X}$ will transit to another state when a Class n user is preempted by an arrival of a higher class. Scenario 2, on the contrary, includes all situations where another state $\mathbf{X}' = (X'_1, ...X'_p)$ transits to $\mathbf{X}$ by a single preemption. This can be further divided to two sub-scenarios: 1) if $l(\mathbf{X}') = n$, then the incoming arrival must have a higher priority than n, and it will preempt an n priority customer in service; 2) if $l(\mathbf{X}') = n' > n$, it implies that $X'_{n'} = 1$ and $X'_i = 0$ for all $i > n, i \neq n'$, as the state will transit to $\mathbf{X}$, where $l(\mathbf{X}) = n$, after the preemption. The incoming arrival must have a priority of n or higher, and it will preempt the only n'th priority customer in service. As a consequence, equations for the boundary states when $1 < n < p$ can be written as:

$$\begin{aligned}&\pi(..., X_n, ...)\left(\sum_{i=1}^{n-1}\lambda_i + \sum_{i=1}^{p} X_i\mu\right)\\ &= \sum_{i=1}^{n}\pi(..., X_i - 1, ...)\lambda_i + \sum_{i=1}^{n-1}\pi(..., X_i - 1, ..., X_n + 1, ...)\lambda_i\\ &\quad + \sum_{i=1}^{n}\sum_{n'=n+1}^{p}\pi(..., X_i - 1, ..., (X_{n'} = 1), ...)\lambda_i.\end{aligned} \tag{7}$$

When $n = 1$, it implies that $X_1 = k$ (as it is a boundary state), and there are no higher priorities that may cause a preemption. Hence:

$$\pi(k,0,...)(k\mu) = \pi(k-1,0,\dots)\lambda_1 + \sum_{n'=2}^{p}\pi(X_1-1,...,(X_{n'}=1),...)\lambda_1.$$

Respectively, when $n=p$, only customers of priority p in service may be preempted by a new arrival. Hence:

$$\begin{aligned}&\pi(...,X_p)\left(\sum_{i=1}^{p}X_i\mu+\sum_{i=1}^{p-1}\lambda_i\right)\\&=\sum_{i=1}^{p-1}\pi(...,X_i-1,...,X_p+1)\lambda_i+\sum_{i=1}^{p}\pi(...,X_i-1,...)\lambda_i.\end{aligned}\tag{8}$$

Based on (6), (7), and (8), we can obtain the probabilities of each state and, in turn, calculate the blocking probability of each priority class, given that the number of priorities p is a constant.

For Class 1 customers with the highest priority, blocking happens when all servers are occupied by only Class 1. Therefore, the blocking probability of Class 1 (p_1) can be written as $p_1=\pi(k,0,0,...)$.

For customers of any other priority Class n ($1<n\le p$), their blocking probabilities consist of two components: (1) the long-run proportion of priority-n customers, among all priority-n arrivals, that are blocked upon arrival because all the servers are busy and there is no user in service with a lower priority than n (denoted $P_{\text{boa}}(n)$); and (2) the long-run proportion of priority-n customers, among those priority-n arrivals admitted to the system, that have been served by the system but preempted by arrivals of customers with higher priorities. The latter corresponds to scenario 1 shown in Fig. 4.

Let $\pi_{\text{p}n}$ be the proportion of time during which the following conditions hold.

- All the k servers are busy.
- There exists at least one customer of Class n.
- There are no customers of priority lower than n in the system.

At such time, any arrival of a customer of priority higher than n will cause preemption of a priority n customer.
Hence,

$$P_{\text{boa}}(n)=\frac{\pi\left(\sum_{i=1}^{n}X_i=k\right)\lambda_n}{\lambda_n},$$

where the steady-state distribution $\pi(\cdot)$ exists for all possible states of an M/M/k/k system, and we introduce the following proposition.

Proposition 1. *The long-run proportion of preempted priority-n customers, among those admitted to the system, exists and is given by*

$$P_{\text{pre}}(n)=\frac{\sum_{i=1}^{n-1}\left(\pi_{\text{p}n}\lambda_i\right)}{(1-P_{\text{boa}}(n))\lambda_n}.$$

The proof of Proposition 1 is provided in A.
Define,

$$\begin{aligned}P_{\text{b}}(n)&=P_{\text{boa}}(n)+(1-P_{\text{boa}}(n))P_{\text{pre}}(n)\\&=\frac{\pi\left(\sum_{i=1}^{n}X_i=k\right)\lambda_n}{\lambda_n}+\frac{\sum_{i=1}^{n-1}\left(\pi_{\text{p}n}\lambda_i\right)}{\lambda_n}\\&=\pi\left(\sum_{i=1}^{n}X_i=k\right)+\frac{\sum_{i=1}^{n-1}\left(\pi_{\text{p}n}\lambda_i\right)}{\lambda_n}.\end{aligned}$$

We generalize the model from the previous two subsections to p priorities class where $A=\sum_{i=1}^{p}A_i$, and obtain

$$\pi\left(\sum_{i=1}^{p}X_i=k\right)=\frac{A^k/k!}{\sum_{i=0}^{k}\frac{A^i}{i!}}.$$

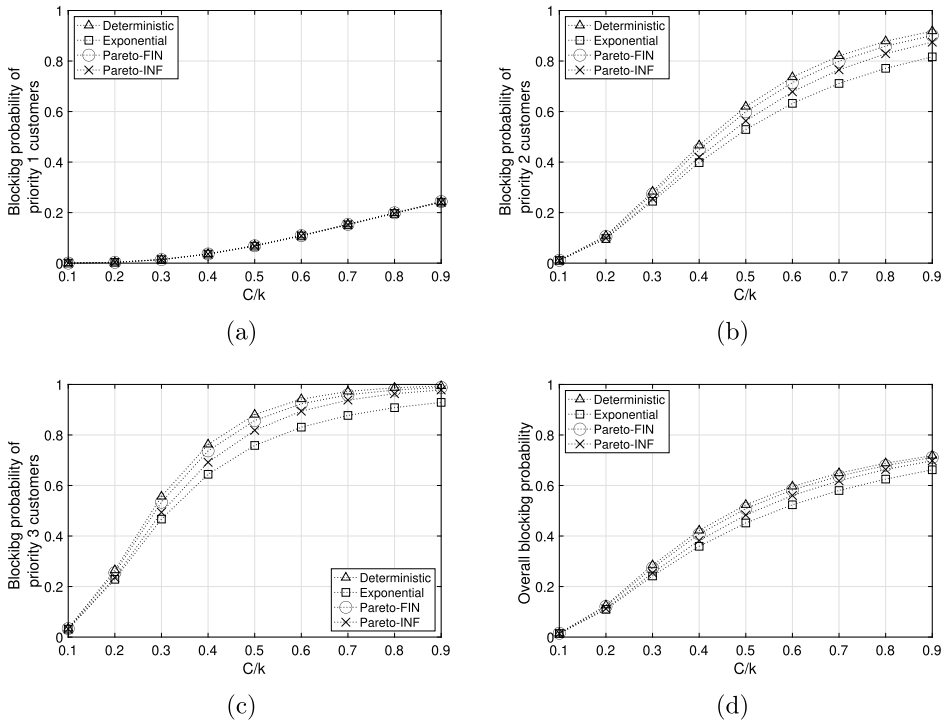


Fig. 4. Blocking probabilities of customers with different priorities and the blocking probability of all customers.

Similarly, by defining the total offered traffic up to priority n,

$$\hat{A}_n = \left(\sum_{i=1}^{n} A_i \right),$$

we have

$$P_{\text{boa}}(n) = \pi\left(\sum_{i=1}^{n} X_i = k \right) = \frac{\hat{A}_n^k / k!}{\sum_{j=0}^{k} \frac{\hat{A}_n^j}{j!}},$$

and

$$\pi_{\text{p}n} = \pi\left(\sum_{i=1}^{n} X_i = k \right) - \pi\left(\sum_{i=1}^{n-1} X_i = k \right).$$

Together with Proposition 1, the probability (proportion) of preemption $P_{\text{pre}}(n)$ is

$$P_{\text{pre}}(n) = \frac{\left(\pi\left(\sum_{i=1}^{n} X_i = k \right) - \pi\left(\sum_{i=1}^{n-1} X_i = k \right) \right) \sum_{i=1}^{n-1} \lambda_i}{(1 - P_{\text{boa}}(n))\lambda_n}.$$

This leads to

$$P_{\text{pre}}(n) = \frac{\left(\frac{\hat{A}_n^k / k!}{\sum_{m=0}^{k} \frac{\hat{A}_n^m}{m!}} - \frac{\hat{A}_{n-1}^k / k!}{\sum_{m=0}^{k} \frac{\hat{A}_{n-1}^m}{m!}} \right) \sum_{i=1}^{n-1} \lambda_i}{(1 - P_{\text{boa}}(n))\lambda_n}.$$

Then, we obtain the following result for the overall blocking probability of priority n customers [47].

$$\begin{aligned} P_{\mathrm{b}}(n) &= P_{\mathrm{boa}}(n) + (1 - P_{\mathrm{boa}}(n))P_{\mathrm{pre}}(n) \\ &= \frac{\hat{A}_n^k/k!}{\sum_{m=0}^{k} \frac{\hat{A}_n^m}{m!}} \\ &\quad + \frac{\sum_{i=1}^{n-1} \lambda_i \left(\frac{\hat{A}_n^k/k!}{\sum_{m=0}^{k} \frac{\hat{A}_n^m}{m!}} - \frac{\hat{A}_{n-1}^k/k!}{\sum_{m=0}^{k} \frac{\hat{A}_{n-1}^m}{m!}} \right)}{\lambda_n}. \end{aligned} \tag{9}$$

3.3. Consistency with the results in Section 2

As expected, the blocking probability results obtained in this paper directly from the multi-dimensional steady-state equations are consistent with the results described in Section 2 based on [45].

By rearranging the terms in (9), we have

$$P_{\mathrm{b}}(n) = \left(\frac{\sum_{i=1}^{n} \lambda_i}{\lambda_n} \right) \frac{\hat{A}_n^k/k!}{\sum_{m=0}^{k} \frac{\hat{A}_n^m}{m!}} - \left(\frac{\sum_{i=1}^{n-1} \lambda_i}{\lambda_n} \right) \left(\frac{\hat{A}_{n-1}^k/k!}{\sum_{m=0}^{k} \frac{\hat{A}_{n-1}^m}{m!}} \right). \tag{10}$$

Recall that the service time of all customers (of all priority classes) is exponentially distributed with parameter μ. By multiplying the numerators and denominators of both terms in (10) by $1/\mu$, we obtain

$$P_{\mathrm{b}}(n) = \left(\frac{\sum_{i=1}^{n} \frac{\lambda_i}{\mu}}{\frac{\lambda_n}{\mu}} \right) \left(\frac{\frac{\hat{A}_n^k}{k!}}{\sum_{m=0}^{k} \frac{\hat{A}_n^m}{m!}} \right) - \left(\frac{\sum_{i=1}^{n-1} \frac{\lambda_i}{\mu}}{\frac{\lambda_n}{\mu}} \right) \left(\frac{\frac{\hat{A}_{n-1}^k}{k!}}{\sum_{m=0}^{k} \frac{\hat{A}_{n-1}^m}{m!}} \right).$$

This leads to an equivalent expression of $P_{\mathrm{b}}(n)$,

$$P_{\mathrm{b}}(n) = \frac{\left(\hat{A}_n\right) E_k\left(\hat{A}_n\right) - \left(\hat{A}_{n-1}\right) E_k\left(\hat{A}_{n-1}\right)}{A_n},$$

which is consistent with (1) and (2) as in Section 2 and [45].

4. Sensitivity to holding time distribution

As discussed in Section 2, the stationary distribution of the number of priority 1 customers in the system is insensitive to the shape of the holding time distribution. However, this insensitivity property does not apply to the preempted customers, that is, to customers of priority i for $i > 1$. This can be explained as follows. Consider, for example, the case where all the service times are deterministic (fixed). When a higher-priority customer preempts a low-priority customer, the remaining service time of the preempted call is shorter than that of the higher-priority customer. In this case, the preemption will increase the load in the system, and since the displacement/replacement argument relies on the exponential service times assumption where the remaining service times of the preempted call and the new high-priority call are the same, we can expect that for the case of deterministic service times, assuming the same mean service time, we will expect to have a higher blocking probability for the preempted priorities traffic than for the case of exponential service times. On the other hand, if the service times have a higher variance than that of the exponential, the opposite will occur. Namely, the remaining service time of the preempted call is longer than that of the higher-priority customer. In this case, the preemption will decrease the load in the system, so we will expect to have a lower blocking probability for the preempted priorities traffic than for the case of exponential service times.

Interestingly, high service time variance improved system performance for low-priority customers. This contradicts the “normal” effects of queueing systems where larger variance adversely affects performance, but we have to remember that when the variance is large, mostly the longer (low priority) jobs are the ones that are being preempted by high priority traffic and this leads to the overall reduction of blocking probability. Note that such behavior, although it is uncommon in queueing systems, in the context of a single server delay system, has been observed before, for example, in [51], where customers with the shortest remaining service time have priority, and in such a system larger variance of the service time improves performance.

Note that this is consistent with similar comments made in [29] based on the numerical results for the loss probability of the $BMAP/PH/N/N$ queue.

We will numerically demonstrate these effects by simulations where we consider four cases involving four holding time distributions: deterministic, exponential, and two cases of Pareto distributions. In particular, we will demonstrate that the higher the variance, the lower the blocking probability of the lower-priority customers.

We consider an M/M/k/k system with $p = 3$ priorities of customers and $k = 5$ servers where the offered traffic for each priority is fixed and is denoted by C. That is, $A_i = C$ for $i = 1, 2, 3$. Thus, $A = A_1 + A_2 + A_3 = 3C$.

In Fig. 4, we present our simulation results for the blocking probabilities of customers with different priorities ($p = 1, 2, 3$) and the overall blocking probability of all customers, as a function of the ratio $C/k = C/5$. In our simulation results, the 95% confidence intervals based on the Student t-distribution are always maintained within $\pm 3\%$ of the observed mean.

In all the cases we consider that involve all four holding time distributions: deterministic, exponential, and the two Pareto distributions have unit mean. That is, the mean holding time is equal to one, so for the deterministic case, the service times are always equal exactly to one and for the exponential case, the service times are exponentially distributed with parameter $\mu = 1$. We provide the blocking probabilities in a range of C values between 0 and $k = 5$, so that the ratio C/k is between 0 and 1. As we set $\mu = 1$, for any given C value, the arrival rate is given by $\lambda = C\mu = C$ for all four cases.

For the two Pareto cases, we again consider unit mean. In the first of these two cases, the shape parameter of the Pareto distribution of the service time is set at $\gamma = 2.001$, and for the second case, we set $\gamma = 1.98$. It is known that for a Pareto random variable, the variance is infinite for $0 < \gamma \leq 2$. Accordingly, we denote the first case, for which the variance is finite (with $\gamma = 2.001$), as Pareto-FIN and the second case (with $\gamma = 1.98$, where the variance is infinite) as Pareto Pareto-INF. Both cases represent distributions with very large variances.

Also, note that as the mean of the service time is set to be equal to one, the scale parameter denoted δ for each of the two cases can be obtained by

$$\delta = \frac{\gamma - 1}{\gamma}.$$

Accordingly, the values of the Pareto scale parameters are set at $\delta = 0.50025$ and $\delta = 0.495$ for Pareto-FIN and Pareto-INF, respectively.

In Fig. 4(a), we demonstrate that the blocking probabilities of priority 1 customers are insensitive to the shape of the tested holding time distributions; while, in Figs. 4(b) and 4(c), it is demonstrated that the blocking probabilities of other customers are clearly different for different distributions and that high variance of the service time distribution leads to lower blocking probabilities for the lower priority customers. In consistency with the sensitivity of the preempted customers, the curves of the overall blocking probability presented in Fig. 4(d) are also different from each other, and we observe, for the cases we studied, that a larger variance of the service time reduces the overall blocking probability. However, we note that this may not always hold. These numerical results are sufficient to demonstrate the sensitivity of preempted customers to the shape of their holding time distribution.

5. Performance behavior as a function of system parameters

In this section, we provide numerical results to illustrate the behavior of the blocking probability, our performance measure, versus key parameter variations based on the analytical results of Sections 2 and 3. We consider an M/M/k/k system with priorities in the following three cases.

- Case 1: The total offered traffic A varies between 1 and 15, with $p = 3$ priorities of customers and $k = 10$ servers.
- Case 2: The number of priorities varies between 2 and 6, with $k = 10$ servers and the total traffic fixed at $A = 9$.
- Case 3: The number of servers k varies from 5 to 15, while the total offered traffic A varies from 4.5 to 14.25 such that the utilization A/k is maintained at 0.9. The number of priorities is set as $p = 3$.

In each case, we further consider two different scenarios regarding traffic distributions:

- (a) The offered traffic is evenly distributed to different priorities, that is, $A_i = C = A/p$ for all i.
- (b) A significant portion of the offered traffic originates from lower priority classes. Specifically, the proportion of offered traffic across up to six priority classes, ranging from high to low priority, follows the ratio $1 : 2 : 3 : 4 : 5 : 6$. In scenarios where fewer than six priority classes are present, we use a truncated version of this ratio. For example, with three priority classes, the ratio becomes $1 : 2 : 3$.

The blocking probabilities of the different priorities in the three cases are presented in Figs. 5, 6 and 7.

We can observe that, in all cases, the blocking probabilities of lower priorities are more sensitive to the changes in system parameters compared to those of higher priorities. The results can be explained by the fact that customers of the lower priority are affected by the load of the higher priority classes while the higher priority classes are not affected by the load of lower priority classes.

For instance, in Case 1, as we observe in Figs. 5(a) and 5(b), when the offered traffic increases, the entire system becomes more congested in general. However, customers of higher priorities (in particular, priority 1) continue to maintain lower blocking probabilities as they are allowed to preempt the service of lower-priority customers, and therefore, they are not affected at all by

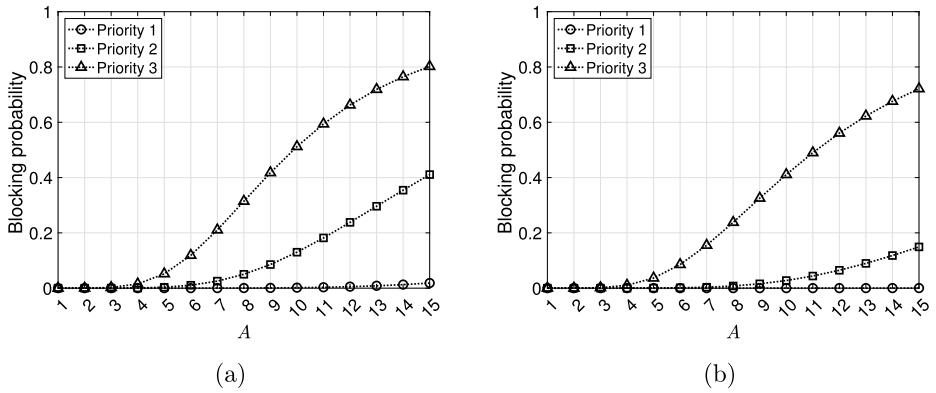


Fig. 5. Blocking probability of the three priority classes versus offered traffic A, with $k = 10$, and (a) evenly distributed traffic to different priorities; (b) heavier traffic from lower priorities.

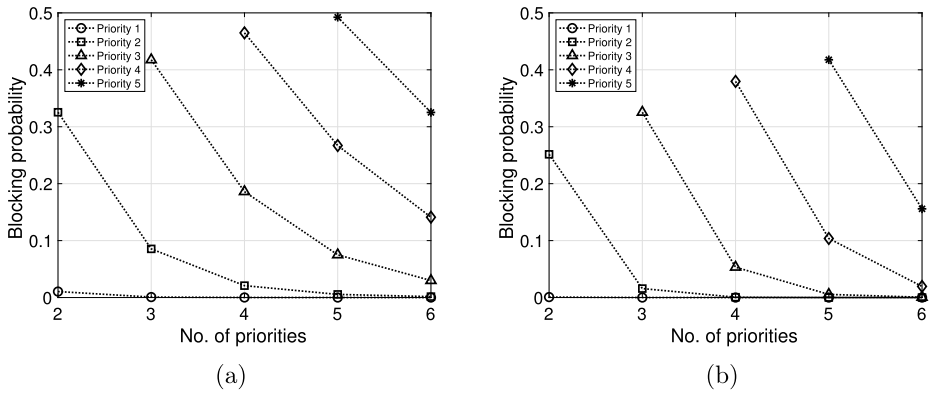


Fig. 6. Blocking probability of the various priority classes versus the number of priorities p, with $k = 10$ and $A = 9$, and (a) evenly distributed traffic to different priorities; (b) heavier traffic from lower priorities.

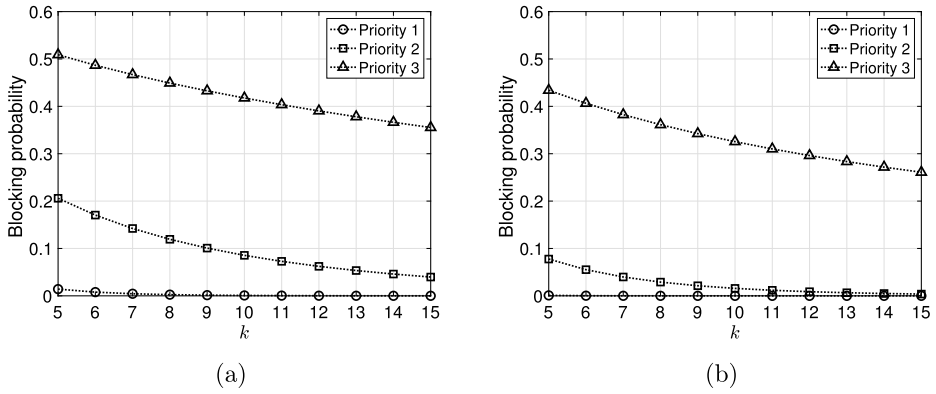


Fig. 7. Blocking probability of the three priority classes versus the number of servers k, with $A/k = 0.9$, and (a) evenly distributed traffic to different priorities; (b) heavier traffic from lower priorities.

the low-priority load. On the other hand, the service availability of lower-priority customers is affected by the increase of customers from the same or higher priorities, leading to a more significant increase in the blocking probability for these customers.

In Figs. 6(a) and 6(b), we further observe that when the number of priorities increases while the total offered traffic remains the same, the blocking probability of customers of a certain higher priority class (e.g., priority 2) decreases. This is because the proportion of lower-priority customers (that can be preempted by customers of that high-priority class) increases.

In Figs. 7(a) and 7(b), we see that as the number of servers increases, the arriving traffic (from all priority classes) is more likely to be accommodated, leading to lower blocking probabilities for all priority classes. On the other hand, while this benefits all

priorities, the already lower blocked priority 1 customers experience a relatively lower change in blocking probability than customers of priorities 2 and 3.

Finally, by comparing scenarios (a) and (b) for all three cases, we observe that when the lower priority class traffic constitutes a larger proportion of the total offered traffic, the blocking probabilities of all priorities decrease, given that both the total offered traffic and the number of servers remain constant. The decrease is more significant for lower priorities, as they are less likely to be preempted by higher priorities.

6. Conclusion

We have established the consistency between the global balance (steady state) equations for a loss system with preemptive priorities and a result for the various blocking probabilities of the different classes of customers based on traffic loss arguments. This has been done by deriving this known result directly from the global balance equations of the relevant multidimensional Markov chain.

This has been achieved by observing that the blocking probability of a customer of any other priority class consists of two components: (1) the probability of being blocked because all the servers are busy and there is no user in service with a lower priority, and (2) the probability that the customer has been admitted, but after admission, it has been preempted by an arrival of a higher priority customer. Notice that for the top priority class, the second component is equal to zero. After these two probabilities (components) have been derived, we have obtained the steady-state blocking probabilities of all customers, and we have observed the consistency with the previously obtained results based on traffic loss arguments.

We have also provided explanations and demonstrated by simulations that except for the blocking probability of the highest priority customers, the blocking probabilities of the other customers are sensitive to the holding time distributions and that a higher variance of the service time distributions reduces the blocking probabilities of the lower priority customers.

One of the possible future extensions of this work is to consider the partial availability of servers for certain customer classes. This is applicable in practical scenarios such as bed allocations in ICUs, where elective patients (lower priority customers) are barred from accessing the last few beds, which are exclusively reserved for emergency patients (higher priority customers). Future work may also involve the relaxation of Poisson arrivals to other Markovian processes, such as the Interrupted Poisson Process (IPP) or Markov-Modulated Poisson Process (MMPP), which have been demonstrated to better fit certain practical scenarios.

CRediT authorship contribution statement

Hang Yang: Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Jing Fu:** Writing – review & editing, Software, Formal analysis, Data curation. **Jingjin Wu:** Writing – review & editing, Visualization, Validation, Software. **Moshe Zukerman:** Writing – review & editing, Validation, Supervision, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The work described in this paper was supported by the City University of Hong Kong under Projects 7005292, 9610385, 7005435, and by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Project code 2022B1212010006, by Guangdong Higher Education Upgrading Plan (2021-2025) UIC R0400001-22, and by Zhuhai Basic and Applied Basic Research Foundation Grant ZH22017003200018PWC.

The authors would like to thank Prof. Peter Taylor for his helpful comments on the paper. We also thank anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions that improved our manuscript.

Appendix A. Proof of Proposition 1

Lemma 1. *For a Poisson arrival process with arrival rate M, let $M(t)$ be the number of arrivals by time $t > 0$, where the expected number of these arrivals $\mathbb{E}[M(t)] = Mt$. Then, for any $\epsilon > 0$,*

$$\lim_{t\to\infty} \mathbb{P}\left\{\left|\frac{M(t)}{t} - M\right| > \epsilon\right\} = 0. \tag{11}$$

Proof. Denote X_m as the time interval between the $(m-1)$-th and m-th arrivals. By definition, X_m for all $m = 1, 2, \ldots$, are identically and independently distributed random variables. Let $\mathbb{E}\left[X_m\right] = x$. Define $T(m)$ as the time of the m-th arrival. We have, for any $t > 0$, base on the definitions,

$$T(M(t)) \in \left[t - X_{M(t)+1}, t\right]. \tag{12}$$

For a given $M(t)$ value, consider the average number of arrivals per unit time by time t,

$$\bar{M}(t) := \frac{M(t)}{t}, \tag{13}$$

which, based on (12), satisfies

$$\frac{M(t)}{T(M(t)) + X_{M(t)+1}} \leqslant \bar{M}(t) \leqslant \frac{M(t)}{T(M(t))}. \tag{14}$$

We then obtain

$$\lim_{t\to\infty} \bar{M}(t) \leqslant \lim_{t\to\infty} \frac{M(t)}{T(M(t))} = \lim_{M(t)\to\infty} \frac{M(t)}{T(M(t))} = \lim_{M(t)\to\infty} \frac{M(t)}{\sum_{m=1}^{M(t)} X_m}. \tag{15}$$

Similarly, we obtain, from (14),

$$\lim_{t\to\infty} \bar{M}(t) \geqslant \lim_{M(t)\to\infty} \frac{M(t)}{T(M(t)) + X_{M(t)+1}} = \lim_{M(t)\to\infty} \left(\frac{M(t)}{\sum_{m=1}^{M(t)} X_m} - \frac{o(M(t))}{M(t)} \right). \tag{16}$$

From (15) and (16), there exists $T > 0$ such that, for all $t > T$,

$$\bar{M}(t) = \frac{M(t)}{\sum_{m=1}^{M(t)} X_m} - \frac{o(M(t))}{M(t)}, \tag{17}$$

and, based on the law of large numbers, $\mathbb{E}[\bar{M}(t)] = \frac{1}{x}$. Then, for any $\epsilon > 0$,

$$\lim_{t\to\infty} \mathbb{P}\left\{ \left| \bar{M}(t) - \mathbb{E}[\bar{M}(t)] \right| > \epsilon \right\} = \lim_{t\to\infty} \mathbb{P}\left\{ \left| \frac{M(t)}{\sum_{m=1}^{M(t)} X_m} - \frac{1}{x} - \frac{o(M(t))}{M(t)} \right| > \epsilon \right\} \leqslant \lim_{M(t)\to\infty} \mathbb{P}\left\{ \left| \frac{M(t)}{\sum_{m=1}^{M(t)} X_m} - \frac{1}{x} \right| > \epsilon/2 \right\} = 0 \tag{18}$$

where the last equality comes from the law of large numbers. This proves the lemma. □

Based on Lemma 1, the number of priority-n arrivals admitted to the system, represented by $\hat{M}_n(t)$, converges to its expectation $(1 - P_{\text{boa}}(n))\lambda_n t$ as the time horizon $t \to \infty$. That is, for any $\epsilon > 0$,

$$\lim_{t\to\infty} \mathbb{P}\left\{ \left| \hat{M}_n(t)/t - (1 - P_{\text{boa}}(n))\lambda_n \right| > \epsilon \right\} = 0. \tag{19}$$

Based on the definition of the priority-loss system, by time horizon t, the number of priority-n arrivals that are admitted to the system but preempted by customers with higher priorities, denoted by $\tilde{M}_n(t)$, is equal to the number of arrivals that have priorities higher than n and incur preemptions of priority-n customers. The long-run expectation of the latter divided by t exists and is equal to $\pi_{\text{pn}} \sum_{i=1}^{n-1} \lambda_i$, leading to the existence of the long-run expectation of $\mathbb{E}[\tilde{M}_n(t)]/t = \pi_{\text{pn}} \sum_{i=1}^{n-1} \lambda_i$. Together with Lemma 1, it follows that, for any $\epsilon > 0$,

$$\lim_{t\to+\infty} \mathbb{P}\left\{ \left| \frac{\tilde{M}_n(t)}{t} - \pi_{\text{pn}} \sum_{i=1}^{n-1} \lambda_i \right| > \epsilon \right\} = 0. \tag{20}$$

In this context, for any $\epsilon > 0$,

$$\lim_{t\to+\infty} \mathbb{P}\left\{ \left| \frac{\tilde{M}_n(t)}{\hat{M}_n(t)} - \frac{\pi_{\text{pn}} \sum_{i=1}^{n-1} \lambda_i}{(1 - P_{\text{boa}}(n))\lambda_n} \right| > \epsilon \right\} = \lim_{t\to+\infty} \mathbb{P}\left\{ \left| \frac{\pi_{\text{pn}} \sum_{i=1}^{n-1} \lambda_i + o(t)}{(1 - P_{\text{boa}}(n))\lambda_n + o(t)} - \frac{\pi_{\text{pn}} \sum_{i=1}^{n-1} \lambda_i}{(1 - P_{\text{boa}}(n))\lambda_n} \right| > \epsilon \right\} = 0. \tag{21}$$

This proves the proposition. □

Comment. Notice that in the definition of $\tilde{M}_n(t)$, we only consider priority-n arrivals that are admitted to the system before time t and are preempted by customers with higher priorities before time t. We do not consider priority-n arrivals that are admitted to the system before time t and are preempted by customers with higher priorities after time t. As in the proof, we consider $t \to \infty$, we will have the total number of priority-n arrivals that are preempted by customers with higher priorities before time t approaches infinity, but the total number of priority-n arrivals that are preempted by customers with higher priorities after time t cannot be larger than the number of customers in the system at time t which is limited by k. Therefore, the number of such preemptions is negligible so they do not need to be included in the definition of $\tilde{M}_n(t)$.

References

[1] D.G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain, Ann. Math. Stat. (1953) 338–354.

[2] N. Litvak, M. Van Rijsbergen, R.J. Boucherie, M. van Houdenhoven, Managing the overflow of intensive care patients, Eur. J. Oper. Res. 185 (3) (2008) 998–1010.
[3] A.M. de Bruin, R. Bekker, L. Van Zanten, G. Koole, Dimensioning hospital wards using the Erlang loss model, Ann. Oper. Res. 178 (1) (2010) 23–43.
[4] R. Bekker, A.M. de Bruin, Time-dependent analysis for refused admissions in clinical wards, Ann. Oper. Res. 178 (1) (2010) 45–65.
[5] R. Bekker, G. Koole, D. Roubos, Flexible bed allocations for hospital wards, Health Care Manage. Sci. 20 (4) (2017) 453–466.
[6] A.R. Andersen, B.F. Nielsen, L.B. Reinhardt, Optimization of hospital ward resources with patient relocation using Markov chain modeling, Eur. J. Oper. Res. 260 (3) (2017) 1152–1163.
[7] A.R. Andersen, W. Vancroonenburg, G.V. Berghe, Strategic room type allocation for nursing wards through Markov chain modeling, Artif. Intell. Med. 99 (2019) 101705.
[8] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, W.A. Van Lent, W.H. Van Harten, An exact approach for relating recovering surgical patient workload to the master surgical schedule, J. Oper. Res. Soc. 62 (10) (2011) 1851–1860.
[9] A.H. Adams, Method of producing a balanced telephone exchange, US Patent 1,504,301, Aug. 12 1924.
[10] F.P. Kelly, Blocking probabilities in large circuit-switched networks, Adv. Appl. Probab. 18 (2) (1986) 473–505.
[11] B. Eklundh, Channel utilization and blocking probability in a cellular mobile telephone system with directed retry, IEEE Trans. Commun. 34 (4) (1986) 329–337.
[12] D. Hong, S.S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures, IEEE Trans. Veh. Technol. 35 (3) (1986) 77–92.
[13] L. Zhai, H. Wang, C. Gao, A spectrum access based on quality of service (QoS) in cognitive radio networks, PLoS ONE 11 (5) (2016) e0155074.
[14] D. Everitt, D. Manfield, Performance analysis of cellular mobile communication systems with dynamic channel assignment, IEEE J. Sel. Areas Commun. 7 (8) (1989) 1172–1180.
[15] P.L. Hiew, M. Zukerman, Efficiency comparison of channel allocation schemes for digital mobile communication networks, IEEE Trans. Veh. Technol. 49 (3) (2000) 724–733.
[16] J. Wu, E.W. Wong, J. Guo, M. Zukerman, Performance analysis of green cellular networks with selective base-station sleeping, Perform. Eval. 111 (2017) 17–36.
[17] L. De Giovanni, A. Langellotti, L. Patitucci, L. Petrini, Dimensioning of hierarchical storage for video on demand services, in: Proc. IEEE ICC/SUPERCOMM'94, 1994, pp. 1739–1743.
[18] V.O.K. Li, W. Liao, X. Qiu, E.W.M. Wong, Performance model of interactive video-on-demand systems, IEEE J. Sel. Areas Commun. 14 (6) (1996) 1099–1109.
[19] S. Vakilinia, M.M. Ali, D. Qiu, Modeling of the resource allocation in cloud computing centers, Comput. Netw. 91 (2015) 453–470.
[20] E. Hyytiä, R. Righter, Simulation and performance evaluation of mission critical dispatching systems, Perform. Eval. 135 (2019) 102038.
[21] R.C. Hampshire, S. Bao, W.S. Lasecki, A. Daw, J. Pender, Beyond safety drivers: applying air traffic control principles to support the deployment of driverless vehicles, PLoS ONE 15 (5) (2020) e0232837.
[22] S. Li, E.W.M. Wong, H. Overby, M. Zukerman, Performance modelling of diversity coded path protection in OBS/OPS networks, IEEE/OSA J. Lightwave Technol. 37 (13) (2019) 3138–3152.
[23] M. Wang, S. Li, E.W.M. Wong, M. Zukerman, Performance analysis of circuit switched multi-service multi-rate networks with alternative routing, IEEE/OSA J. Lightwave Technol. 32 (2) (2014) 179–200.
[24] A. Zalesky, H. Vu, Z. Rosberg, E. Wong, M. Zukerman, OBS contention resolution performance, Perform. Eval. 64 (4) (2007) 357–373.
[25] D.H. Hailu, G.G. Lema, B.G. Gebrehaweria, S.H. Kebede, Quality of service (QoS) improving schemes in optical networks, Heliyon 6 (4) (2020).
[26] Y. Khlifi, F.A. Al-Zahrani, Joint resource optimization and flexible QoS provision using hybrid optical core node architecture, Heliyon 10 (2) (2024) e24058.
[27] A. Brandwajn, T. Begin, Multi-server preemptive priority queue with general arrivals and service times, Perform. Eval. 115 (2017) 150–164.
[28] M. Głąbowski, M. Sobieraj, M. Stasiak, A multi-service model of resources with the neighboring choice of allocation units, IEEE Access 9 (2021) 107260–107266, https://doi.org/10.1109/ACCESS.2021.3101412.
[29] V. Klimenok, C. Kim, D. Orlovsky, A. Dudin, Lack of invariant property of Erlang BMAP/PH/N/0 model, Queueing Syst. 49 (2005) 187–213.
[30] V. Klimenok, A. Dudin, V. Vishnevsky, Priority multi-server queueing system with heterogeneous customers, Mathematics 8 (9) (2020) 1501.
[31] I. Moscholios, M. Logothetis, Efficient Multirate Teletraffic Loss Models Beyond Erlang, Wiley, 2019.
[32] B.A. Sevast'yanov, An ergodic theorem for Markov processes and its application to telephone systems with refusals, Theory Probab. Appl. 2 (1) (1957) 104–112.
[33] Y. Ding, E. Park, M. Nagarajan, E. Grafstein, Patient prioritization in emergency department triage systems: an empirical study of the Canadian triage and acuity scale (CTAS), Manuf. Serv. Oper. Manag. 21 (4) (2019) 723–741.
[34] Z. Li, Y. Hu, L. Tian, Z. Lv, Packet rank-aware active queue management for programmable flow scheduling, Comput. Netw. 225 (2023) 109632.
[35] Y. Zhao, Q. Lu, Z. Ye, K. Chen, A communication failure and repair mechanism with adjustable transmission rates for PU packets in CRNs, Heliyon 9 (2) (2023) e13184.
[36] K. Kim, Delay cycle analysis of finite-buffer M/G/1 queues and its application to the analysis of M/G/1 priority queues with finite and infinite buffers, Perform. Eval. 143 (2020) 102133.
[37] M.T. Barros, R.C. Zambon, P.S. Barbosa, W.W.-G. Yeh, Planning and operation of large-scale water distribution systems with preemptive priorities, J. Water Resour. Plan. Manag. 134 (3) (2008) 247–256.
[38] R. Pal, A. Prakash, R. Tripathi, K. Naik, Scheduling algorithm based on preemptive priority and hybrid data structure for cognitive radio technology with vehicular ad hoc network, IET Commun. 13 (20) (2019) 3443–3451.
[39] B. Palit, S.S. Das, Y. Kamavaram, Multiple QoS provisioning with pre-emptive priority schedulers in multi-resource OFDMA networks, Wirel. Netw. (2020) 1–20.
[40] E.W. Wong, A. Zalesky, Z. Rosberg, M. Zukerman, A new method for approximating blocking probability in overflow loss networks, Comput. Netw. 51 (11) (2007) 2958–2975, https://doi.org/10.1016/j.comnet.2006.12.007.
[41] J. Wu, E.W.M. Wong, Y.-C. Chan, M. Zukerman, Power consumption and GoS tradeoff in cellular mobile networks with base station sleeping and related performance studies, IEEE Transactions on Green Communications and Networking 4 (4) (2020) 1024–1036, https://doi.org/10.1109/TGCN.2020.3000277.
[42] Y.-C. Chan, J. Guo, E.W. Wong, M. Zukerman, Surrogate models for performance evaluation of multi-skill multi-layer overflow loss systems, Perform. Eval. 104 (2016) 1–22.
[43] Y.-C. Chan, J. Wu, E.W. Wong, C.S. Leung, Integrating teletraffic theory with neural networks for quality-of-service evaluation in mobile networks, Appl. Soft Comput. 152 (2024) 111208.
[44] L. Katzschner, Loss systems with displacing priorities, in: Proc. 6th International Teletraffic Congress, 1970, p. 224.
[45] H.L. Vu, M. Zukerman, Blocking probability for priority classes in optical burst switching networks, IEEE Commun. Lett. 6 (5) (2002) 214–216.
[46] S. Yang, N. Stol, Performance modeling in multi-service communications systems with preemptive scheduling, J. Commun. 9 (6) (2014) 448–460.
[47] E.A. Maslova, A.G. Tatashev, Approximate calculation of priority service characteristics in a multiserver system with losses, Autom. Control Comput. Sci. 26 (4) (1992) 31.
[48] A.G. Tatashev, O.V. Seleznjev, M.V. Yashina, Approach to evaluating characteristics of multichannel loss system with FCFD preempted priority discipline, arXiv:2207.07123, 2022.
[49] M. Zukerman, Introduction to queueing theory and stochastic teletraffic models, http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf, 2021.
[50] V.B. Iversen, Teletraffic engineering and network planning, DTU Fotonik, https://backend.orbit.dtu.dk/ws/portalfiles/portal/118473571/Teletraffic_34342_V_B_Iversen_2015.pdf, 2015.
[51] A.V. Pechinkin, The MAP/G/1/∞ queue with SRPT service discipline, Theory Probab. Appl. 45 (3) (2001) 532–539.