Queueing Systems 50, 81–107, 2005

An Exact Solution for the State Probabilities of the Multi-Class, Multi-Server Queue with Preemptive Priorities

ANDREI SLEPTCHENKO * A.Sleptchenko@tue.nl
EURANDOM, Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands

AART VAN HARTEN
MATTHIEU VAN DER HEIJDEN
Department of Technology and Management, University of Twente, P.O. Box 217, 7500AE Enschede, The Netherlands

Received 23 March 2004; Revised 15 January 2005

Abstract. We consider a multi-class, multi-server queueing system with preemptive priorities. We distinguish two groups of priority classes that consist of multiple customer types, each having their own arrival and service rate. We assume Poisson arrival processes and exponentially distributed service times. We derive an exact method to estimate the steady state probabilities. Because we need iterations to calculate the steady state probabilities, the only error arises from choosing a finite number of matrix iterations. Based on these probabilities, we can derive approximations for a wide range of relevant performance characteristics, such as the moments of the number of customers of a certain type in the system en the expected postponement time for each customer class. We illustrate our method with some numerical examples. Numerical results show that in most cases we need only a moderate number of matrix iterations ($\sim$20) to obtain an error less than 1% when estimating key performance characteristics.

Keywords: queueing, multi-class, multi-server, preemptive priority

AMS subject classification: 90B22

1. Introduction

1.1. Model

In this paper, we consider a multi-class $M/M/k$ queue with two priority groups, high and low. That is, each priority group may consist of multiple customer types, each having their own arrival and service rate. Within each priority group, customers are served on a first come, first serve (FCFS) basis. If all servers are occupied upon arrival of a high priority customer, a low priority customer in service can be interrupted to free service capacity for the high priority customer. That is, we assume *preemptive* priorities. If

* Corresponding author.

multiple low priority customers are in service, a customer is selected for postponement randomly. Postponed low priority customers have priority over non-postponed waiting customers of the same priority class. Because of the memoryless property of the exponential distribution, it does not make a difference whether postponed jobs are resumed from the moment of interruption or whether they are restarted completely. If one of the postponed customers can re-enter service due to service completion of one of the customers in service, then the choice of that customer is also made randomly from the available customers. This rule allows us to work with the number of postponed customers of each low priority type in the state space, neglecting the sequence of postponement. Other disciplines for the postponed customers, such as FCFS, are possible at the expense of a larger state space.

We present a method to solve the steady state equations exactly. Because we know all state probabilities, we are able to evaluate a wide range of performance characteristics per customer type, such as the moments of the number of customers waiting and in the system, the waiting probability and the mean number of low priority customers whose service is interrupted. So our analysis facilitates the computation of advanced performance measures beyond standard measures as the mean queue length.

Note that our model covers the $M/H_x/k$ priority queue with no customer subclasses per priority group. In this case, we represent each priority group class with hyperexponential (H_x) service times by x customer subclasses with exponential service times and we use the performance estimators for the total number of customers of each priority group in the system. We will focus on the multi-class model in this paper, because we are interested in performance measures per customer type rather than aggregate performance measures per priority group.

1.2. Motivation

Multi-server priority queueing systems arise in various applications, such as production, computer and telecommunication systems and call centers. We encountered this model during our research on spare parts logistics for repairable items [13,14]. In such situations we aim to minimize holding costs for spare parts stocks given fixed system availability, or to maximize system availability given a fixed budget for spare parts. The queueing models are necessary to model repair facilities.

A repair shop in spare part networks is generally able to handle multiple items and it can use priority setting in its control. Here, we classify priorities statically as either high priority or low priority items. Each item has its own arrival rate and service time distribution. As a consequence, we need to model a repair shop by a (multi-server) priority queueing system with two priority classes, where each class consists of multiple subclasses (item types). An algorithm to determine performance characteristics of such multi-server, multi-class priority queueing systems is not available in the literature as far as we know.

1.3. Literature

There is quite some literature on multi-server priority queueing systems, see e.g. both for preemptive priorities [4,10,11] and for nonpreemptive priorities [3,6–8,10,18,19,20].

Regarding non-preemptive priority queues, Wagner [18–20] analyzes multi-server nonpreemptive priority systems with Markovian arrival process, service times having phase type distributions and both finite or infinite queueing space using matrix-analytic methods. Kao and Narayanan [6] apply a matrix geometric approach to compute the steady-state distribution of the customers in the system. Kao and Wilson [7] apply a power-series approach to a multi-server queue with two priority classes. The power-series approach has been introduced by Hooghiemstra et al. [5] and has been applied before to solve a variety of queueing problems—particularly those with multidimensional state space. The power series approach is interesting, because it can easily be implemented and it can be extended to include more than two priority classes and to preemptive priorities in theory. However these extensions cause an enormous growth of memory requirements and computation time.

Among the papers on preemptive priorities, we mention the approximation approach of Buzen [1] and the generating function approach as proposed by Mitrani and King [11] and by Gail et al. [4]. The basic idea of the Buzen's approximation approach is to replace k servers by a single server that works k times as fast and to use a correction factor, being the ratio of the waiting times when the same trick would be applied to the non-priority multi-server queue. Although this approach is attractive because of its simplicity and its extendibility to general service times, it was done only for the first moments of the number of customers in the system. In contrast to Buzen's idea, the generating function approach gives exact results for the first two moments of the number of customers in the system. However, these approaches [4,11] have only been applied to cases with two classes each with one type of item.

The literature above focuses on priority queues where the priority classes do not consist of multiple customer subclasses (or: customers types) each with their own arrival and service rate. Such a setting can be found in the literature on dynamic scheduling of multi-class queues. However, there the focus is on finding an optimal scheduling policy rather than analytic performance estimation, see Wein [21] and Reinman and Wein [12] for dynamic scheduling of a single-server queue and Maglaras [9] for dynamic scheduling in queueing networks. We could not find any literature on the computation of the steady state probabilities if the priority classes consist of non-identical subclasses. Only for non-priority queues, such an analysis is known, see De Smidt [2] and Van Harten and Sleptchenko [16]. The priority model that we consider in this paper requires a more detailed state description and a different solution scheme.

1.4. Approach

To analyze our model, we proceed as follows. First we construct the equilibrium state equations on a semi-infinite state space (Section 2). The state corresponds with the

numbers of clients waiting for service in queue and the state of service (i.e. how many clients of which type are in service and how many are postponed). We distinguish three areas when solving the equilibrium equations, namely (I) states with at least one high priority customer in the queue, (II) states with only low priority customers in the queue, and (III) states in which the queues are empty. The boundaries between these areas require special attention, as they are critical to solve the steady state equations. We will deal with each area separately.

In Section 3.1, we will solve the equilibrium equations in area I (high priority customers in the queue) combining the generating function approach with the matrix-geometric approach. That is, in the area I the probabilities of the system states will be expressed as derivatives of some power function. In this way we express all the probabilities in this area in state probabilities with only one high priority customer in the queue. The latter states are exactly the boundary between the areas I and II.

Next, in Section 3.2, we will show how to deal with the area II (no high priority customers in the queue), and the boundary between areas I and II (i.e. states with only one high priority customer in the queue). Here, we will apply a generating function approach with respect to the lower priority items only. We can solve the latter equations iteratively only, thereby obtaining a cut-off error. Here again we express all the probabilities in area II in state probabilities with only one high or low priority customer in the queue, which belong to the boundaries between area I and II and between II and III, respectively.

Finally, we solve the remaining state probabilities (area III plus the states with at most one customer in the queue) in Section 3.3.

Altogether, the structure of this problem is a lot more complex than the problem without priorities dealt with in Van Harten and Sleptchenko [16]. The main reason is that we have to cope with a non-hyperexponential structure of the solution. As a consequence, we have to do some tricky transformations to find our solutions. In this respect, our approach resembles the moment generating function approach and is somewhat analogous to Keilson's Laguerre transform approach. We prefer a special scheme in order to get better transparency and numerical efficiency.

In Section 4, we show how we can calculate various performance characteristics from the state probabilities. In Section 5 we summarize the proposed method. In the numerical Section 6, we show the results of several experiments to get insight in the number of matrix iterations required to calculate the performance characteristics with high precision. As examples of performance characteristics, we take the expected number of customers of each type in the queue, being served or postponed, the first two moments of the number of customers in the system for each customer type, and the expected waiting, sojourn and postponement time per customer type. We present our conclusions and we discuss some model extensions in Section 7.

2. Notation and state equations

2.1. Definitions and notation

We denote the number of customer classes with high (low) priority by N^h (N^l). High priority jobs from subclass i arrive according to a Poisson process with rate λ_i^h and low priority jobs from subclass j arrive with rate λ_j^l. We denote the number of parallel service channels by k. The service times of the subclasses are exponentially distributed with rates μ_i^h and μ_j^l for high and low priority customer classes, respectively.

Other general notations used throughout the paper are:

$\Lambda^h,\ \Lambda^l,\ \mu^h,\ \mu^l$—overall arrival and service rates per priority class, i.e. $\Lambda^h = \sum_{i=1}^{N^h} \lambda_i^h$, $\Lambda^l = \sum_{i=1}^{N^l} \lambda_i^l$ and $\mu^h = \Lambda^h / \sum_{i=1}^{N^h} \frac{\lambda_i^h}{\mu_i^h} = \Lambda^h / k\rho^h$, $\mu^l = \Lambda^l / \sum_{i=1}^{N^l} \frac{\lambda_i^l}{\mu_i^l} = \Lambda^l / k\rho^l$, where the overall utilization rates for each priority class are $\rho^h = \Lambda^h / k\mu^h$, $\rho^l = \Lambda^l / k\mu^l$ and the total utilization rate is $\rho = \rho^h + \rho^l$.

a_i^h, a_i^l—customer i arrival rate as fraction of the total arrival rate of its priority class, i.e. $a_i^h = \lambda_i^h / \Lambda^h$, $a_i^l = \lambda_i^l / \Lambda^l$.

$\bar{\mu}(\bar{s}^h, \bar{s}^l)$—total service rates of all customers in service, i.e. $\bar{\mu}(\bar{s}^h, \bar{s}^l) = \sum_{i=1}^{N^h} s_i^h \mu_i^h + \sum_{i=1}^{N^l} s_i^l \mu_i^l$.

$\bar{e}_i^h(\bar{e}_i^l)$—a vector of dimension N^h (N^l) with component i equal to 1 and all other components equal to 0; this vector is used to indicate the changes in vectors of queue and servers states during transitions from state to state.

$e_{ij}^h(e_{ij}^l)$—the jth component of the vector $\bar{e}_i^h$ ($\bar{e}_i^l$), so e_{ij}^h (e_{ij}^l) = 1 if $i = j$ and 0 otherwise.

We characterize the system states by two vectors of dimension N^h and three vectors of dimension N^l, where the components of each vector refer to the (high and low priority) subclasses. These vectors contain information about the customers in queue, in service and postponed:

$\bar{s}^h$ and $\bar{s}^l$—vectors containing the number of high and low priority customers in service per customer class.

$\bar{w}^h$ and $\bar{w}^l$—vectors containing the number of high and low priority customers in the queue waiting for first service per customer class (the vector $\bar{w}^l$ *excludes* postponed customers).

$\bar{r}^l$—a vector containing the number of postponed low priority customers per customer class.

We denote the systems state probabilities by $P(\bar{w}^h, \bar{s}^h, \bar{w}^l, \bar{s}^l, \bar{r}^l)$. Fortunately, we can reduce the state dimension, because the customers within each priority class are served FCFS and arrive according to independent Poisson processes. Therefore, the conditional distribution of the number of customers of each type in the queue, given the total number of customers in the queue for the particular priority class, has a multinomial

distribution (see van Harten and Sleptchenko [16]). So we can replace the vectors $\bar{w}^h$ and $\bar{w}^l$ by scalars q^h and q^l, denoting the total number of high (low) priority customers in the queue, respectively. So, $q^h = \sum_{i=1}^{N^h} w_i^h$ and $q^l = \sum_{i=1}^{N^l} w_i^l$ and the state probabilities can be written as

$$P(\bar{w}^h, \bar{s}^h, \bar{w}^l, \bar{s}^l, \bar{r}^l) = q^h! \prod_{i=1}^{N^h} \frac{\left(a_i^h\right)^{w_i^h}}{w_i^h!} q^l! \prod_{j=1}^{N^l} \frac{\left(a_j^l\right)^{w_j^l}}{w_j^l!} P_{q^h,q^l}(\bar{s}^h, \bar{s}^l, \bar{r}^l). \tag{1}$$

where $P_{q^h,q^l}(\bar{s}^h, \bar{s}^l, \bar{r}^l)$ denotes the steady state probability that (i) q^h high priority customers are in the queue, (ii) q^l low priority customers (who have not been postponed) are in the queue, and (iii) the vectors of postponed low priority customers and high and low priority customers equal $\bar{r}^l$, $\bar{s}^h$ and $\bar{s}^l$, respectively. In the remainder of this paper, we will use the state probabilities $P_{q^h,q^l}(\bar{s}^h, \bar{s}^l, \bar{r}^l)$ only. When we mention the number of low priority customers in the queue (q^l), we only refer to the customers who have not been served before (i.e., excluding postponed customers).

2.2. *Equilibrium equations*

In this section, we will define the equilibrium equations for the continuous time Markov chain. Without loss of generality, we assume that if multiple servers are available to process a job, each available server has an equal chance to get this job.

Different equilibrium equations hold for different system states, but they have a similar structure. For convenience of notation, we denote the state probabilities as a vector $\mathbf{P}_{q^h,q^l}$, containing all probabilities $P_{q^h,q^l}(\bar{s}^h, \bar{s}^l, \bar{r}^l)$ for fixed q^h and q^l. Then we can write the equilibrium equations in the following generic form:

$$\mathbf{D}_{q^h,q^l}\mathbf{P}_{q^h,q^l} = \mathbf{F}_{q^h,q^l}\mathbf{P}_{q^h-1,q^l} + \mathbf{E}_{q^h,q^l}\mathbf{P}_{q^h,q^l-1} + \mathbf{B}_{q^h,q^l}\mathbf{P}_{q^h+1,q^l} + \mathbf{G}_{q^h,q^l}\mathbf{P}_{q^h,q^l+1} \tag{2}$$

where the operators $\mathbf{D}_{q^h,q^l}$, $\mathbf{F}_{q^h,q^l}$, $\mathbf{E}_{q^h,q^l}$, $\mathbf{B}_{q^h,q^l}$, $\mathbf{G}_{q^h,q^l}$ depend on the area (I, II or III). It is straightforward to derive that these operators are given by:

$$\mathbf{D}_{q^h,q^l} = \mathbf{D}_{1,0} \text{ with } (\mathbf{D}_{1,0}\mathbf{P}_{q^h,q^l})[\bar{s}^h, \bar{s}^l, \bar{r}^l] \stackrel{def}{=} (\Lambda^h + \Lambda^l + \bar{\mu}(\bar{s}^h, \bar{s}^l))\mathbf{P}_{q^h,q^l}[\bar{s}^h, \bar{s}^l, \bar{r}^l], \quad q^h > 0, \quad q^l \geq 0 \tag{3}$$

$$\begin{aligned}\mathbf{D}_{q^h,q^l} = \mathbf{D}_{0,1} \text{ with } (\mathbf{D}_{0,1}\mathbf{P}_{0,q^l})[\bar{s}^h, \bar{s}^l, \bar{r}^l] &\stackrel{def}{=} (\Lambda^h + \Lambda^l + \bar{\mu}(\bar{s}^h, \bar{s}^l))\mathbf{P}_{0,q^l}[\bar{s}^h, \bar{s}^l, 0] \\ &- \Lambda^h \sum_{i=1}^{N^h}\sum_{j=1}^{N^l} \frac{s_j^l + 1}{\sum_{m=1}^{N^l} s_m^l + 1} a_i^h \mathbf{P}_{0,q^l}\left[\bar{s}^h - \bar{e}_i^h, \bar{s}^l + \bar{e}_j^l, \bar{r}^l - \bar{e}_j^l\right] \\ &- \sum_{i=1}^{N^h}\sum_{j=1}^{N^l} \left(s_i^h + 1\right)\mu_i^h \frac{r_j^l + 1}{\sum_{m=1}^{N^l} r_m^l + 1} \mathbf{P}_{0,q^l}\left[\bar{s}^h + \bar{e}_i^h, \bar{s}^l - \bar{e}_j^l, \bar{r}^l + \bar{e}_j^l\right]\end{aligned}$$

$$- \sum_{i=1}^{N^l} \sum_{j=1}^{N^l} \left(s_i^l + 1 - e_{ij}^l\right) \mu_i^l \frac{r_j^l + 1}{\sum_{m=1}^{N^l} r_m^l + 1} \mathbf{P}_{0,q^l}\left[\bar{s}^h, \bar{s}^l + \bar{e}_i^l - \bar{e}_j^l, \bar{r}^l + \bar{e}_j^l\right],$$

$$q^h = 0, \quad q^l > 0 \tag{4}$$

$$\mathbf{D}_{q^h,q^l} = \mathbf{D}_{0,0} \text{ with } (\mathbf{D}_{0,0}\mathbf{P}_{0,0})[\bar{s}^h, \bar{s}^l, \bar{r}^l] \stackrel{def}{=} (\Lambda^h + \Lambda^l + \bar{\mu}(\bar{s}^h, \bar{s}^l))\mathbf{P}_{0,0}[\bar{s}^h, \bar{s}^l, 0]$$
$$- \Lambda^h \sum_{i=1}^{N^h} \mathbf{P}_{0,0}\left[\bar{s}^h - \bar{e}_i^h, \bar{s}^l, 0\right] - \Lambda^l \sum_{i=1}^{N^l} \mathbf{P}_{0,0}\left[\bar{s}^h, \bar{s}^l - \bar{e}_i^l, 0\right]$$
$$- \sum_{i=1}^{N^h} \left(s_i^h + 1\right)\mu_i^h \mathbf{P}_{0,0}\left[\bar{s}^h + \bar{e}_i^h, \bar{s}^l, 0\right] - \sum_{i=1}^{N^l} \left(s_i^l + 1\right)\mu_i^l \mathbf{P}_{0,0}\left[\bar{s}^h, \bar{s}^l + \bar{e}_i^l, 0\right],$$

$$q^h = 0, \quad q^l = 0 \tag{5}$$

$$\mathbf{E}_{q^h,q^l} = \mathbf{E} \text{ with } (\mathbf{E}\,\mathbf{P}_{q^h,q^l-1})[\bar{s}^h, \bar{s}^l, \bar{r}^l] \stackrel{def}{=} \Lambda^l \mathbf{P}_{q^h,q^l-1}[\bar{s}^h, \bar{s}^l, \bar{r}^l], \quad q^h + q^l > 0 \tag{6}$$

$$\mathbf{F}_{q^h,q^l} = \mathbf{F} \text{ with } (\mathbf{F}\,\mathbf{P}_{q^h-1,q^l})[\bar{s}^h, \bar{s}^l, \bar{r}^l] \stackrel{def}{=} \Lambda^h \mathbf{P}_{q^h-1,q^l}[\bar{s}^h, \bar{s}^l, \bar{r}^l], \quad q^h > 0, \quad q^l \geq 0 \tag{7}$$

$$\mathbf{B}_{q^h,q^l} = \mathbf{B} \text{ with } (\mathbf{B}\,\mathbf{P}_{q^h+1,q^l})[\bar{s}^h, \bar{s}^l, \bar{r}^l] \stackrel{def}{=} \sum_{i=1}^{N^h} \sum_{j=1}^{N^h} a_j^h \left(s_i^h + 1 - e_{ij}^h\right)$$
$$\times \mu_i^h \mathbf{P}_{q^h+1,q^l}\left[\bar{s}^h + \bar{e}_i^h - \bar{e}_j^l, 0, \bar{r}^l\right], \quad q^h > 0, \quad q^l \geq 0 \tag{8}$$

$$\mathbf{G}_{q^h,q^l} = 0, \quad q^h > 0, q^l \geq 0 \tag{9}$$

$$\mathbf{G}_{q^h,q^l} = \mathbf{G} \text{ with } (\mathbf{G}\,\mathbf{P}_{0,q^l+1})[\bar{s}^h, \bar{s}^l, \bar{r}^l] \stackrel{def}{=} \sum_{i=1}^{N^l} \sum_{j=1}^{N^l} a_j^l \left(s_i^l + 1 - e_{ij}^l\right)$$
$$\times \mu_i^l \mathbf{P}_{0,q^l+1}\left[\bar{s}^h, \bar{s}^l + \bar{e}_i^l - \bar{e}_j^l, 0\right] + \sum_{i=1}^{N^h} \sum_{j=1}^{N^l} a_j^l \left(s_i^h + 1\right)$$
$$\times \mu_i^h \mathbf{P}_{0,q^l+1}\left[\bar{s}^h + \bar{e}_i^h, \bar{s}^l - \bar{e}_j^l, 0\right], \quad q^h + q^l > 0, \; q^h = 0, \; \sum_{m=1}^{N^l} r_m^l = 0 \tag{10}$$

The dimension of the vector $\mathbf{P}_{q^h,\,q^l}$ can be specified as follows.

If there are high priority customers in queue, we have:

$$\dim(\mathbf{P}_{q^h,q^l}) = \binom{N^h + k - 1}{k} \sum_{j=0}^{k} \binom{N^l + j - 1}{j}, \quad q^h > 0, \quad q^l \geq 0$$

If there are only low priority customers in queue, we have:

$$\dim(\mathbf{P}_{0,q^l}) = \sum_{i=0}^{k}\left[\binom{N^h+i-1}{i}\binom{N^l+k-i-1}{k-i}\sum_{j=0}^{i}\binom{N^l+j-1}{j}\right],$$
$$q^h = 0, \quad q^l > 0$$

If queues are empty the dimension can be written as:

$$\dim(\mathbf{P}_{0,0}) = \sum_{n=0}^{k-1}\binom{N^h+N^l+n-1}{n} + \sum_{i=0}^{k}\left[\binom{N^h+i-1}{i}\binom{N^l+k-i-1}{k-i}\right.$$
$$\left.\times\sum_{j=0}^{i}\binom{N^l+j-1}{j}\right], \quad q^h+q^l = 0$$

In the next section, we will show how we can solve these equilibrium equations thereby obtaining the exact system state probabilities.

3. Solution of the stationary state equations

3.1. System states with high priority customers in queue ($q^h > 0$)

In this section, we focus on area I, so there is at least one high priority customer in the queue.

$$\mathbf{D}_{1,0}\mathbf{P}_{q^h,q^l} = \Lambda^h\mathbf{P}_{q^h-1,q^l} + \Lambda^l\mathbf{P}_{q^h,q^l-1} + \mathbf{B}\,\mathbf{P}_{q^h+1,q^l}, \quad q^h > 1 \tag{11}$$

where $\mathbf{D}_{1,0}$ and $\mathbf{B}$ are defined by (3) and (8).

Solving this matrix equation, we can express all state probabilities with $q^h > 1$ and $q^l \geq 0$ in the state probabilities $\mathbf{P}_{1,q^l}$. In the next lemma, we explain the structure of the solution to this equation.

Lemma 1. Define the matrix function $\mathbf{Z}(\xi)$ as such solution of

$$\mathbf{D}_{1,0} = \Lambda^h\mathbf{Z} + \Lambda^l\xi + \mathbf{B}\,\mathbf{Z}^{-1}, \tag{12}$$

that satisfies to condition that for all eigenvalues $\alpha(\mathbf{Z}(\xi))$ it holds that

$$|\alpha(\mathbf{Z}(\xi))| > 1, \quad \text{for } \xi = 0, 1$$

Then

$$\mathbf{P}_{q^h,q^l} = \frac{1}{q^l!}\left(\frac{d}{d\xi}\right)^{q^l}(\mathbf{Z}^{-1}(\xi))^{q^h-1}\mathbf{C}(\xi)\Big|_{\xi=0} \tag{13}$$

satisfies all equations for $\mathrm{q}^h \geq 1$ and $q^l \geq 0$.

Note that $\sum_{q^l \geq 0} \mathbf{P}_{q^h,q^l} = (\mathbf{Z}^{-1}(\xi))^{q^h-1}\mathbf{C}(\xi)|_{\xi=1}$ under mild analyticity conditions on $\mathbf{C}(\xi)$.

Proof. This lemma can be easily shown by substitution, see Appendix for details. □

The probabilities of the system states constructed in this section have a differential form. Therefore we need the derivatives of the matrix $\mathbf{Z}(\xi)$. Unfortunately, we have not been able to derive an analytic expression for the matrix $\mathbf{Z}(\xi)$. However, we can use equation (12) to find the derivatives iteratively using following relations:

$$\Lambda^h \frac{d}{d\xi}\mathbf{Z}(\xi) + \mathbf{B}\frac{d}{d\xi}[\mathbf{Z}^{-1}(\xi)] = -\Lambda^l$$

$$\vdots$$

$$\Lambda^h \left(\frac{d}{d\xi}\right)^n \mathbf{Z}(\xi) + \mathbf{B}\left(\frac{d}{d\xi}\right)^n [\mathbf{Z}^{-1}(\xi)] = 0, \quad n > 1$$

and the following relation obtained from the equality $\mathbf{Z}(\xi)\mathbf{Z}^{-1}(\xi) = \mathbf{I}$:

$$\left(\frac{d^n}{d\xi^n}\mathbf{Z}(\xi)\right)\mathbf{Z}^{-1}(\xi) + \mathbf{Z}(\xi)\left(\frac{d^n}{d\xi^n}\mathbf{Z}^{-1}(\xi)\right) = \sum_{i=1}^{n-1}\binom{n}{i}\left(\frac{d^i}{d\xi^i}\mathbf{Z}(\xi)\right)\left(\frac{d^{n-i}}{d\xi^{n-i}}\mathbf{Z}^{-1}(\xi)\right).$$

So, we have found the probabilities of the system states with a nonempty high priority queue ($q^h \geq 1$), expressed via the derivatives of the unknown function $\boldsymbol{C}(\xi)$. However, using (13) it is easy to see that the derivatives of this function for $\xi = 0$ are related to probabilities of the states with $q^h = 1$. So, if we know these probabilities, we can find the probabilities of the other states with $q^h > 1$. In the next sections, we show how to find these probabilities and the other state probabilities (with $q^h = 0$).

3.2. *States with all servers busy and at most one high priority customer in the queue ($q^h \leq 1$, $q^h + q^l > 0$)*

Here we consider the equilibrium equations for the systems states in area II ($q^h = 0$) and the boundary equations of area I and II ($q^h = 1$), which can be written in matrix form as:

$$\begin{pmatrix}\mathbf{D}_{0,1} & \mathbf{B}\\ \mathbf{F} & \mathbf{D}_{1,0}\end{pmatrix}\begin{pmatrix}\mathbf{P}_{0,q^l}\\ \mathbf{P}_{1,q^l}\end{pmatrix} = \begin{pmatrix}\mathbf{E} & \mathbf{0}\\ \mathbf{0} & \mathbf{E}\end{pmatrix}\begin{pmatrix}\mathbf{P}_{0,q^l-1}\\ \mathbf{P}_{1,q^l-1}\end{pmatrix} + \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{P}_{0,q^l+1}\\ \mathbf{P}_{1,q^l+1}\end{pmatrix} + \begin{pmatrix}\mathbf{0}\\ \mathbf{B}\end{pmatrix}\mathbf{P}_{2,q^l}$$

Rewriting these equations in a matrix form, where the vectors $\mathbf{P}_{q^h,q^l}$ with the same number of low priority customers in the system (q^l) are joined into one vector $\mathbf{P}_{q^l}$, we

obtain:

$$D_2 P_{q^l} = E_2 P_{q^l-1} + G_2 P_{q^l+1} + B_2 \frac{1}{q^l!}\left(\frac{d}{d\xi}\right)^{q^l} [\mathbf{Z}^{-1}(\xi)\mathbf{C}(\xi)]\Big|_{\xi=0}, \quad q^l > 1 \quad (14)$$

So, we have an inhomogeneous system of difference equations with a fixed dimension and with fixed coefficients. These fixed matrix coefficients D_2, E_2, G_2 and B_2 were introduced in the previous equation.

We see that the inhomogeneous term has a differential form, so the standard solution of inhomogeneous difference equations (solution of homogeneous equation plus a partial solution of the inhomogeneous equation) can be applied by looking for a solution in a differential form $P_{q^l} = \frac{1}{(q^l-1)!}(\frac{d}{d\xi})^{q^l-1} v(\xi)|_{\xi=0}$. This form is sufficiently general to incorporate, all solutions of the homogeneous equation as well by adding terms to $v(\xi)$ proportional to $(\alpha + \xi)^{-1}$. The substitution of this solution into equation (14) gives:

$$D_2 \frac{1}{q^l!}\left(\frac{d}{d\xi}\right)^{q^l} v(\xi)\Big|_{\xi=0} = E_2 \frac{1}{(q^l-1)!}\left(\frac{d}{d\xi}\right)^{q^l-1} v(\xi)\Big|_{\xi=0} + G_2 \frac{1}{(q^l+1)!}\left(\frac{d}{d\xi}\right)^{q^l+1} v(\xi)\Big|_{\xi=0} + B_2 \frac{1}{(q^l+1)!}\left(\frac{d}{d\xi}\right)^{q^l+1} [\mathbf{Z}^{-1}(\xi)\mathbf{C}(\xi)]\Big|_{\xi=0},$$
$$q^l > 1$$

Here we can apply the following equalities:

$$\left(\frac{d}{dx}\right)^n (xf(x))\Big|_{x=0} = n\left(\frac{d}{dx}\right)^{n-1} f(x)\Big|_{x=0} \text{ and}$$
$$\left(\frac{d}{dx}\right)^n (x^2 f(x))\Big|_{x=0} = n(n-1)\left(\frac{d}{dx}\right)^{t-2} f(x)\Big|_{x=0},$$

which allow us to remove the derivatives from equation (14) and to obtain a new expression of the function $v(\xi)$ for any $q^l > 1$:

$$\frac{1}{(q^l+1)!}\left(\frac{d}{d\xi}\right)^{q^l+1} [\xi D_2 v(\xi) - \xi^2 E_2 v(\xi) - G_2 v(\xi) - B_2 \mathbf{Z}^{-1}(\xi)\mathbf{C}(\xi)]\Big|_{\xi=0} = 0, \quad q^l > 1 \quad (15)$$

Using (13), the function $\mathbf{C}(\xi)$ can be expressed as the part of the vector-function $v(\xi) = \binom{v_0(\xi)}{v_1(\xi)}$, which corresponds to the states with 1 high priority customer in the queue $v_1(\xi)$, i.e. $\mathbf{C}(\xi) = v_1(\xi)$ and for example $v_1(0) = \mathbf{P}_{1,1}$.

The right hand side of equation (15) should be a function which becomes zero for any $q^l > 1$, i.e. a polynomial function of second order. Hence, we obtain another

expression for the vector-function $v(\xi)$, that does not contain derivatives, but that contains unknown vectors C_1, C_2 and C_3:

$$\xi \mathrm{D}_2 v(\xi) - \xi^2 \mathrm{E}_2 v(\xi) - \mathrm{G}_2 v(\xi) - \mathrm{B}_2 \mathbf{Z}^{-1}(\xi) v_1(\xi) = C_1 \xi^2 + C_2 \xi + C_3 \tag{16}$$

or

$$\mathrm{H}(\xi) v(\xi) = C_1 \xi^2 + C_2 \xi + C_3 \tag{17}$$

The constants C_1, C_2 and C_3 can easily be expressed in the state probabilities P_2, P_1 and P_0, where P_0 is also obtained by joining vectors $\mathbf{P}_{0,0}$ and $\mathbf{P}_{1,0}$. That is, we have from equation (16) that for $\xi = 0$,

$$-\mathrm{G}_2 v(0) - \mathrm{B}_2 \mathbf{Z}^{-1}(0) \mathbf{C}(0) = C_3$$

and recalling that $v(0) = \mathrm{P}_1$ and $\mathbf{C}(0) = \mathbf{P}_{1,0}$, we obtain an equation for C_3:

$$C_3 = -\mathrm{G}_2 \mathrm{P}_1 + \mathrm{B}_2 \mathbf{Z}^{-1}(0) \mathbf{P}_{1,0} \tag{18}$$

Next, we can take the derivative of equation (16) in the point $\xi = 0$ and we obtain

$$\mathrm{D}_2 v(0) - \mathrm{G}_2 v'(0) - \mathrm{B}_2 \left(\frac{d}{d\xi} \mathbf{Z}^{-1}(0) \mathbf{P}_{1,0} + \mathbf{Z}^{-1}(0) \mathbf{P}_{1,1} \right) = C_2.$$

The left hand side of the last equation is also encountered in equation (14) for $q^l = 1$, if we take into account that $\mathrm{P}_1 = v(0)$, $\mathrm{P}_2 = v'(0)$. Then we find:

$$C_1 = \mathrm{D}_2 \mathrm{P}_1 - \mathrm{G}_2 \mathrm{P}_2 - \mathrm{B}_2 \left(\frac{d}{d\xi} \mathbf{Z}^{-1}(0) \mathbf{P}_{1,0} + \mathbf{Z}^{-1}(0) \mathbf{P}_{1,1} \right). \tag{19}$$

In this way, we have defined a function $v(\xi)$ given the probability vectors P_2, P_1 and P_0, (remembering that $\mathbf{P}_{1,0}$ and $\mathbf{P}_{1,1}$ are parts of the vectors P_1 and P_0).

Next, all probability vectors P_{q^l} for $q^l = 3 \ldots \infty$, and so all state probabilities in the areas I and II follow from P_2, P_1 and P_0. Note that P_1 and P_0 correspond with states at the boundary between area II, III and area I, but P_2 does not. However, up to now we have not used an essential piece of information by which we can eliminate P_2. It is clear that we are looking for decaying solutions P_{q^l} for $q^l \to \infty$. As a consequence, $v(\xi)$ should be analytic on a circle with radius $1 + \varepsilon$ for some $\varepsilon > 0$. Due to (16), extra conditions have to be satisfied at points ξ inside this circle where $\mathrm{H}(\xi)$ is singular.

Next, we shall show that the decay requirement boils down to a relation between the initial condition P_2 and P_1 of the following type:

$$\mathrm{P}_2 = \mathrm{Q}_1 \mathrm{P}_1 \tag{20}$$

It is easy to show (see Appendix B) that:

$$\mathrm{P}_2 = \mathbf{Q}_1^t \mathrm{P}_1 + \mathbf{\Omega}_1^t \mathrm{P}_{t+1} \tag{21}$$

where

$$\mathbf{\Omega}_1^t = -\left(\mathbf{\Theta}_1^2\right)^{-1}\mathbf{\Theta}_0^2 \quad \text{and} \quad \mathbf{Q}_1^t = -\left(\mathbf{\Theta}_1^2\right)^{-1}\mathbf{\Theta}_2^2. \tag{22}$$

with matrices $\mathbf{\Theta}_i^{t^*}$ that can be found recursively from

$$\begin{aligned} \mathbf{\Theta}_i^t &= h_i, \quad i = 0, \ldots t \\ \mathbf{\Theta}_0^{t^*} &= h_0\left(\mathbf{\Theta}_1^{t^*+1}\right)^{-1}\mathbf{\Theta}_0^{t^*+1} \\ \mathbf{\Theta}_i^{t^*} &= \left(h_0\left(\mathbf{\Theta}_1^{t^*+1}\right)^{-1}\mathbf{\Theta}_{i+1}^{t^*+1} - h_i\right), \quad i = 1, \ldots t^* \end{aligned} \tag{23}$$

with $h_i = \frac{1}{i!}\frac{d^i}{d\xi^i}\mathrm{H}(0)$. That is, if $\mathbf{\Omega}_1^t$ remains bounded and taking into account that the vector P_{t+1} decay for $t \to \infty$, we get:

$$\mathrm{Q}_1 = \lim_{t\to\infty} \mathbf{Q}_1^t$$

Note that the matrices $\mathbf{\Omega}_1^t$ and $\mathbf{Q}_1^t$ are computed using a straightforward iteration procedure, which requires $\mathrm{O}(t^2)$ matrix operations. In this iteration, we check the boundedness of $\mathbf{\Omega}_1^t$. Numerical evidence shows that this is true in all cases that we considered. Moreover, we find that

$$\lim_{t\to\infty} \|\mathbf{\Omega}_1^t\| \approx \omega(1-\rho)$$

where ω is some constant depending on other system parameters. Hence, by taking t sufficiently large, we have a good approximation of the matrix Q_1 that will play a role next. Furthermore, $\rho \to 1$ is not a problem for the convergence of the iterations.

3.3. States with at most one customer in the queues ($q^h + q^l \leq 1$)

In this section, we show how to find the state probabilities for $q^h + q^l \leq 1$. We can use these probabilities to find all other states probabilities that were expressed in these remaining probabilities till now.

Using the same techniques as in the previous sections, the equilibrium equations for the states with $q^h + q^l \leq 1$ can be written as:

$$\begin{aligned} \mathrm{D}_1\mathrm{P}_1 &= \mathrm{E}_2\mathrm{P}_0 + \mathrm{G}_2\mathrm{P}_2 + \mathrm{B}_2\left(\frac{d}{d\xi}\mathbf{Z}^{-1}(\xi)\bigg|_{\xi=0}\mathbf{P}_{1,0} + \mathbf{Z}^{-1}(0)\mathbf{P}_{1,1}\right) \\ \mathrm{D}_0\mathrm{P}_0 &= \mathrm{G}_2\mathrm{P}_1 + \mathrm{B}_2\mathbf{Z}^{-1}(0)\mathbf{P}_{1,0} \end{aligned} \tag{24}$$

where the matrices D_0, D_1 are given by

$$D_0 = \begin{pmatrix} \mathbf{D}_{0,0} & \mathbf{B} \\ \mathbf{F} & \mathbf{D}_{1,0} \end{pmatrix}$$
$$D_1 = \begin{pmatrix} \mathbf{D}_{0,1} & \mathbf{B} \\ \mathbf{F} & \mathbf{D}_{1,1} \end{pmatrix}$$

and the operators are defined as in (14). The vectors $\mathbf{P}_{1,i}$ are parts of the vectors P_i, $i = 0,1$ as before.

Using expression (20), we can rewrite these equations by eliminating P_2. We obtain a system of linear equations for P_1 and P_0, which determines them up to a multiplicative constant. Together with normalization condition

$$\sum_{q^h=0}^{\infty} \sum_{q^l=0}^{\infty} \sum_{\bar{s}^h,\bar{s}^l,\bar{r}^l} \mathbf{P}_{q^h,q^l}[\bar{s}^h, \bar{s}^l, \bar{r}^l] = 1 \tag{25}$$

this provides us with P_1, P_0 and all other probabilities.

4. Performance measures

Based on the steady state probabilities, we can calculate a wide range of performance measures for each customer subclass. Examples are the moments of the queue lengths and the correlation between the numbers of customers in the system for two classes. In this section, we concentrate on the performance criteria for the low priority customers in the system, since the performance indicators for the high priority customers can be calculated using the non-priority multi-class, multi-serve queue analysis presented in Van Harten and Sleptchenko [16] as well. The latter is possible, because we assume preemptive priorities. So, low priority customers do not influence high priority customers. Therefore low priority customers can be ignored for the performance concerning high priority customers.

As examples for the calculation of performance measures, we take the mean number of low priority customers of type i in the queue ($E[q_i^l]$), the mean number of the low priority customers of type i in the postponed state ($E[PS_i^l]$) the first two moments of the total number of type i low priority customers in the system (R_i^l) and some others. Such performance indicators play a role in spare part service networks that motivated our research, see the introduction.

4.1. Mean number of type i customers in the queue

Obviously, $E[q_i^l]$ can be found from

$$E\big[q_i^l\big] = \sum_{\bar{w}^h,\bar{s}^h,\bar{w}^l,\bar{s}^l,\bar{r}^l} w_i^l P(\bar{w}^h, \bar{s}^h, \bar{w}^l, \bar{s}^l, \bar{r}^l).$$

We can simplify this expression via the function $v(\xi)$ and via the matrix $\mathbf{Z}(\xi)$ using the Taylor expansion:

$$\begin{aligned}
E\big[q_i^l\big] &= \sum_{q^h,q^l} \sum_{\bar{w}^h,\ s.t.\ \sum_{i=1}^{N^h} w_i^h = q^h} \sum_{\bar{w}^l,\ s.t.\ \sum_{i=1}^{N^l} w_i^l = q^l} w_i^l q^h! \prod_{i=1}^{N^h} \frac{\left(a_i^h\right)^{w_i^h}}{w_i^h!} q^l! \prod_{i=1}^{N^l} \frac{\left(a_i^h\right)^{w_i^l}}{w_i^l!} \\
&\quad \times \sum_{\bar{s}^h,\bar{s}^l,\bar{r}^l} P_{q^h,q^l}(\bar{s}^h, \bar{s}^l, \bar{r}^l) \\
&= \sum_{q^l} \sum_{\bar{w}^l,\ s.t.\ \sum_{i=1}^{N^l} w_i^l = q^l} \sum_{\bar{s}^h,\bar{s}^l,\bar{r}^l} w_i^l a_i^l \frac{1}{(q^l-1)!} \left(\frac{d}{d\xi}\right)^{q^l} (\mathbf{Z}(\xi)^{-1})^{q^h-1} v_1(\xi)\Big|_{\xi=0} \\
&\quad + \sum_{\bar{w}^l,\ s.t.\ \sum_{i=1}^{N^l} w_i^l = q^l} \sum_{\bar{s}^h,\bar{s}^l,\bar{r}} w_i^l q^l! \prod_{j=1}^{N^l} \frac{\left(a_j^l\right)}{w_j^l!} . \frac{1}{q^l!} \frac{d^{q^l}}{d\xi^{q^l}} v(\xi)\Big|_{\xi=0}
\end{aligned}$$

This finally gives us

$$\begin{aligned}
E\big[q_i^l\big] = &-a_i^l \langle \mathbf{1}, (\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})^{-1} v_1(1)\rangle \\
&+a_i^l \langle \mathbf{1}, (\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}(1) v_1'(1)\rangle + a_i^l \langle \mathbf{1}, v'(1)\rangle
\end{aligned}$$

Here v_1 refers to the vector components of v with one high priority customer in the queue. The notation $\mathbf{1}$ is used for a vector with all components equal to 1 with correspondent dimension.

4.2. Mean number of type i customers postponed

Analogously to the derivation of $E[q_i^l]$, we can find an expression for the mean number of type i customers postponed $E[PS_i^l]$:

$$E\big[PS_i^l\big] = \big\langle \chi^{r_i^l}, (\mathbf{I} - \mathbf{Z}^{-1}(\xi))^{-1} v_1(1)\big\rangle + \big\langle \chi^{r_i^l}, v(1)\big\rangle$$

where $\chi^{r_i^l}$ has as components the number of postponed low priority customers of type i corresponding to the vector component $(\bar{s}^h, \bar{s}^l, \bar{r}^l)$ and $\chi_1^{r_i^l}$ is the part of this vector corresponding to the boundary states with one high priority customer in the queue.

4.3. Moments of the number of type i customers in the system

Clearly, R_i^l is composed of three terms, namely the number of customers in service, in the queue and postponed:

$$E\big[R_i^l\big] = E\big[SR_i^l\big] + E\big[q_i^l\big] + E\big[PS_i^l\big]$$

We have already determined the last two components, so what remains is the mean number of type i customers in service $E[SR_i^l]$. Using the vectors $\chi^{s_i^l}$ and $\chi_0^{s_i^l}$, with the number of low priority customers of type i in service as components and the corresponding dimensions, we can write down an expression for $E[SR_i^l]$:

$$E\big[SR_i^l\big] = \big\langle \chi^{s_i^l}, \mathrm{P}_0 \big\rangle + \big\langle \chi^{s_i^l}, v(1) \big\rangle$$

Note that the mean number of type i customers in service can also be estimated via Little's law, i.e.

$$E\big[SR_i^l\big] = \lambda_i^l / \mu_i^l$$

It is more difficult to find the second moment of R_i^l, since we should take into account the correlations between the numbers of customers in queue, in service and in the postponed states. This can be done analogous to the computations for non-priority systems as presented in Van Harten and Sleptchenko [16]. After a lengthy derivation we obtain:

$$\begin{aligned}
E\big[(R_i^l)^2\big] = {} & \big(a_i^l\big)^2 \langle \mathbf{1}, 2[(\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})^{-1} \\
& -(\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}''(1)(\mathbf{Z}(1)-\mathbf{I})^{-1}]v_1(1)\rangle \\
& - 2\big(a_i^l\big)^2\big\langle \mathbf{1}, (\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})v_1'(1)\big\rangle \\
& + \big(a_i^l\big)^2\langle \mathbf{1}, (\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}(1)v_k''(1)\rangle \\
& - a_i^l\big\langle \mathbf{1}, (\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})^{-1}v_1(1)\big\rangle + a_i^l\big\langle \mathbf{1}, (\mathbf{Z}(1)-\mathbf{I})^{-1}v_1'(1)\big\rangle \\
& - 2a_i^l\big\langle \chi^{r_i^l}, (\mathbf{Z}(1)-\mathbf{I})^{-1}\mathbf{Z}'(1)(\mathbf{Z}(1)-\mathbf{I})^{-1}v_k(1)\big\rangle \\
& + 2a_i^l\big\langle \chi^{r_i^l}, (\mathbf{Z}(1)-\mathbf{I})^{-1}v_k'(1)\big\rangle + \big\langle \chi^{(r_i^l)^2}, (\mathbf{I}-\mathbf{Z}^{-1}(1))^{-1}v_k(1)\big\rangle \\
& + \big(a_i^l\big)^2\big\langle \mathbf{1}, v''(1)\big\rangle + a_i^l(\big\langle \mathbf{1}, v'(1)\big\rangle + 2\big\langle \chi^{r_i^l+s_i^l}, v'(1)\big\rangle) + \big\langle \chi^{(r_i^l+s_i^l)^2}, v(1)\big\rangle \\
& + \big\langle \chi^{(s_i^l)^2}, \mathrm{P}_0\big\rangle
\end{aligned}$$

With an even more lengthy computation, we can derive expressions for backorder calculation in spare part networks as mentioned in Section 1.2.

4.4. *Mean waiting time, postponement time and sojourn time*

Little's law can be applied to calculate the mean waiting time until the *first* time that an customer enters service $E[W_i^l]$, the mean postponement time $E[PsTime_i^l]$ and the mean sojourn time $E[SJTime_i^l]$:

$$\begin{aligned}\lambda_i^l E\left[W_i^l\right] &= E\left[q_i^l\right]\\ \lambda_i^l E\left[PsTime_i^l\right] &= E\left[PS_i^l\right]\\ \lambda_i^l E\left[SJTime_i^l\right] &= E\left[R_i^l\right]\end{aligned}$$

and, of course,

$$E\left[SJTime_i^l\right] = \frac{1}{\mu_i^l} + E\left[W_i^l\right] + E\left[PsTime_i^l\right]$$

Note that $E[PsTime_i^l]$ has to be interpreted as the expected *total* time that an customer of type i spends in the postponed state between the moment it leaves the queue and the moment its service is completed. Further, we should note that the expected total service time equals $1/\mu_i^l$, even though preemption occurs. Because of the memoryless property of the exponential service time distribution, interruptions (by preemption) do not affect the expected service time.

4.5. *Expected number of preemption events*

Let us now focus on another interesting quantity: $E[nrPreemptEvent_i^l]$, the expected number of preemption events per type $i$customer. In order to compute it, we need the arrival rate of a low priority customer into the postponed state. It can be calculated using the state probabilities as derived in the previous sections. The arrival rate of type i low priority customers into the postponed state is equal to the arrival rate of high priority customers multiplied by the probability that customer i is withdrawn from the service:

$$\lambda_i^{ps} = \Lambda^h \sum_{\bar{s}^h,\bar{w}^l,\bar{s}^l,\bar{r}^l} \frac{s_i^l}{\sum_{i=1}^{N^l} s_i^l} P(0,\bar{s}^h,\bar{w}^l,\bar{s}^l,\bar{r}^l) = \Lambda^h \left\langle \chi^{\frac{s_i^l}{\cdot}}, v(1)\right\rangle$$

where the components of the vectors $\chi^{\frac{s_i^l}{\cdot}}$ are equal to $\frac{s_j^l}{\sum_{i=1}^{N^l} s_i^l}$ of the corresponding vector component $(\bar{s}^h,\bar{s}^l,\bar{r}^l)$.

Comparing the number of preemption events with the number of arrivals over a long interval, it is clear that number of preemption events per customer entering the system is equal to:

$$E\left[nrPreemptEvent_i^l\right] = \lambda_i^{ps} \big/ \lambda_i^l$$

It is now also possible to compute the expected time between postponement and resumption moments of type i customer, i.e. the expected re-entrance into service time, $E[ReenterTime_i^l]$. Using Little's law, we obtain

$$\lambda_i^{ps} E\left[ReenterTime_i^l\right] = E\left[PS_i^l\right]$$

Note that the number of preemption events per each customer entering the system can be also calculated as:

$$E\left[nrPreemptEvent_i^l\right] = E\left[PsTime_i^l\right] / E\left[ReenterTime_i^l\right]$$

4.6. On the vectors v(1), v′(1) and v″(1).

To calculate the performance measures, we need to find the values $v(1)$, $v'(1)$ and $v''(1)$. However, the procedure is non-trivial due to the singularity of the matrix $\mathrm{H}(1)$, i.e. we cannot find the vector $v(1)$ just by inverting equation (16) at $\xi = 1$.

$$\mathrm{H}(1)v(1) = \mathrm{F}_k\mathrm{P}_{k-1} - \mathrm{GP}_k. \tag{26}$$

Therefore, we proceed as follows. First, the derivative of (16) at $\xi = 1$ gives:

$$\mathrm{H}(1)v'(1) + \mathrm{H}'(1)v(1) = \mathrm{F}_k\mathrm{P}_{k-1} \tag{27}$$

In this equation, we have to get rid of the term with $v'(1)$ to obtain a linear system for $v(1)$ only. It can be done by projecting both sides of (27) onto the null-space of $\mathrm{H}(1)$. Hence, we multiply equation (27) by the matrix Pr_1, defined as:

$$\mathrm{Pr}_1 = S \begin{pmatrix} 1 & & & & \\ & \ddots & & \mathbf{0} & \\ & & 1 & & \\ & & & 0 & \\ & \mathbf{0} & & \ddots & \\ & & & & 0 \end{pmatrix} S^{-1}$$

where S is the matrix of eigenvectors of $\mathrm{H}(1)$ and where the rows with 1 on the diagonal correspond to the zero eigenvalues and the rows with 0 on the diagonal to the non-zero eigenvalues; ε is some number >0 that can be used for scaling purposes. It is easy to

show that multiplication of the matrices Pr_1 and H(1) gives the zero matrix:

$$\mathrm{Pr}_1\mathrm{H}(1) = S\begin{pmatrix} 1 & & & & \\ & \ddots & & \mathbf{0} & \\ & & 1 & & \\ & & & 0 & \\ & \mathbf{0} & & \ddots & \\ & & & & 0 \end{pmatrix} S^{-1}\mathrm{H}(1)$$

$$= S\begin{pmatrix} 1 & & & & \\ & \ddots & & \mathbf{0} & \\ & & 1 & & \\ & & & 0 & \\ & \mathbf{0} & & \ddots & \\ & & & & 0 \end{pmatrix} \begin{pmatrix} 0 & & & & \\ & \ddots & & \mathbf{0} & \\ & & 0 & & \\ & & & * & \\ & \mathbf{0} & & \ddots & \\ & & & & * \end{pmatrix} S^{-1} = 0$$

So, we now have a new system of linear equations:

$$[\mathrm{H}(1) + \mathrm{Pr}_1\mathrm{H}'(1)]v(1) = (\mathbf{I} + \mathrm{Pr}_1)\mathrm{F}_k\mathrm{P}_{k-1} - \mathrm{GP}_k$$

where the matrix $[\mathrm{H}(1) + \mathrm{Pr}_1\,\mathrm{H}'(1)]$ turns out to be non-singular. In all our experiments, it turned out that 0 is a single eigenvalue of H(1), with the corresponding left eigenvector is $\mathbf{1} = (1, \ldots, 1)^{\perp}$. As a consequence, the matrix Pr_1 can be constructed as a matrix with elements of one row (any row) equal to 1.

$$\mathrm{Pr}_1 = \begin{pmatrix} 1 & \cdots & 1 \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

In the same way, we can find the derivatives $v'(1)$, $v''(1)$, etc. That is,

$$[\mathrm{H}(1) + 2\,\mathrm{Pr}_1\mathrm{H}'(1)]v'(1) = -[\mathrm{H}'(1) + \mathrm{Pr}_1\mathrm{H}''(1)]v(1) + \mathrm{F}_k\mathrm{P}_{k-1} \tag{28}$$

$$[\mathrm{H}(1) + 3\,\mathrm{Pr}_1\mathrm{H}'(1)]v''(1) = -[2\mathrm{H}'(1) + 3\,\mathrm{Pr}_1\mathrm{H}''(1)]v'(1) - [\mathrm{H}''(1) + \mathrm{Pr}_1\mathrm{H}'''(1)]v(1)$$

In the same way, we can find the values of the function $v(\xi)$ and the values of the derivatives of these function for other values of ξ. Therefore, we are able to find the derivatives of any order of the function $v(\xi)$, which be needed for estimation of other performance measures of the system

5. Summary of the method and the algorithm

Let us summarize our method to calculate the steady state probabilities for the multi-class, multi-server preemptive priority queue.

First, we use equation (12) to find the derivatives of the matrix-function $\mathbf{Z}(\xi)$ in the points $\xi = 0$ and 1, which allows us to express the state probabilities in area I in terms of the state probabilities with $q^h = 1$, see Lemma 1.

Next, we use the first t derivatives (t is the chosen number of iterations) of the matrix-function $\mathbf{Z}(\xi)$ in the point $\xi = 0$ and relations (21)–(23) to find the matrix Q_1, which expresses the relation (20) between P_1 and P_2. Then, using this relation and the equations (26)–(28), we find the derivatives of the function $v(\xi)$ in the points $\xi = 0$ and 1, which gives us the state probabilities for area II, expressed in the state probabilities for the boundary between the areas II en III (i.e., the states with exactly k customers in the system).

Finally, we express the state probabilities in area III in the probability on an empty system state P_0 using the matrix $\mathbf{Q}_1^t$, the linear equations (24) and condition (25) that the sum of all state probabilities equals 1. Then, we can calculate all system state probabilities and the performance measures based on these probabilities (as in Section 4).

In algorithmic form, we can summarize our procedure as follows:

Step 1. Initialization.
Step 2. Calculate the first $t > 2$ derivatives of the function $\mathbf{Z}(\xi)$ in the point $\xi = 0$ from Lemma 1.
Step 3. For t^*= t, ..., 2 calculate $\mathbf{\Theta}_0^{t^*}, \ldots, \mathbf{\Theta}_{t^*}^{t^*}$ by (23)
Step 4. Calculate Q_1 from (20) using the relations (21)–(23)
Step 5. Calculate P_0 and P_1 using the equations (20) and (24).
Step 6. Calculate the first t_1 derivatives of the function $\mathbf{Z}(\xi)$ in the point $\xi = 1$ using Lemma 1 and the first t_1 derivatives of the function $v(\xi)$ in the point $\xi = 1$ using the equations from Section 4.6, where t_1 is equal to maximum order of the moments we are looking for plus 1 (i.e., $t_1 = 3$ for the variance)
Step 7. Calculate the performance estimators as desired.

6. Numerical experiments

In this section, we discuss the results of three sets of numerical experiments. First, we study the convergence speed and computer time requirements of our iterative procedure. Next, we examine the impact of the system parameters on the key performance characteristics for the low priority customers as presented in the previous section. Finally, we compare the number of customers in the system for priority queues to non-priority queues.

6.1. *Convergence of iterative procedure and computation times*

To examine the computational efforts, we first examine the error in the first and second moment of the number of customers in the system as function of the numbers of iterations t in our algorithm, see formula (21). We chose the following parameter settings for our numerical experiments:

- fixed parameters: $k = 4$; $N^h = N^l = 2$; $\lambda_1^h = 1.75$, $\lambda_2^h = 2.25$, $\lambda_1^l = 2.75$, $\lambda_2^l = 3.25$;
- ρ^h is equal to 0.2, 0.4 or 0.6; ρ is equal 0.75, 0.85 or 0.95
- μ_1^h and μ_1^l are equal to $1.5\mu^h$ and $1.5\,\mu^l$ correspondently, where $\mu^h = \frac{\Lambda^h}{k\rho^h}$ $\mu^l = \frac{\Lambda^l}{k\rho^l}$
- μ_2^h and μ_2^l are such that $\sum_{i=1}^{N^h} \frac{\lambda_i^h}{\mu_i^h} = k\rho^h$ and $\sum_{i=1}^{N^l} \frac{\lambda_i^l}{\mu_i^l} = k\rho^l$;

As performance measures, we focus on the mean and variance of the number of low priority customers in the system per class, because the characteristics of high priority customers are not new (as mentioned in the previous section, they can also be obtained using our method from Van Harten and Sleptchenko [16]).

First, we compared the results of our numerical procedure to some results obtained by discrete event simulation. We found that the deviation between the calculated and simulated performance measures lies within 3% error interval with 95% confidence already after 10 iterations. Further, we have checked whether the test $E[SR_i^l] = \lambda_i^l/\mu_i^l$ from the previous section is satisfied. The experiments have shown that of the results obtained after 20 iterations and after 40 iterations give average relative errors of ∼0.2% and ∼0.04%, respectively.

Next, we examined the convergence speed and computer time requirements of our iterative procedure. In figure 1 we plot the maximum error for the mean and the variance of the numbers of low priority customers in the system obtained by our algorithm as function of the number of iterations. As benchmark, we used the results after 80 iterations, because then the values of the performance measures hardly change anymore in all experiments.

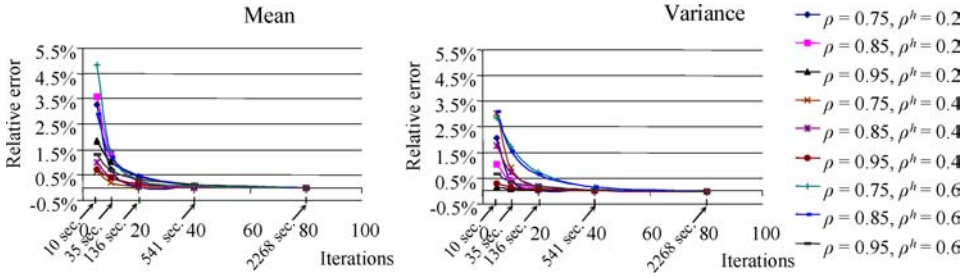

Figure 1. Relation between the error in the mean and variance of the number of low priority customers in the system and the number of iterations.

Under the x-axis, we show the CPU time required for our calculations using a Pentium IV-2.0 PC. We find that the CPU time is approximately a quadratic function of the number of iterations. The reason of this behavior is that for t iterations the algorithm requires $O(t^2)$ matrix operations. As a consequence, we need long run times to obtain extremely accurate results. Figure 1 also shows that the CPU-time requirements are modest if we accept a small error. For example, after 10 iterations the maximum error is less than 1%, and we need only 90 seconds for these calculations.

Figure 1 also shows another interesting characteristic of our algorithm. We see that we can obtain good results for low priority customers, where the number of iterations required decreases with the utilization rate. These cases are most interesting for our application (spare part management). Further, we have an exact method to estimate the performance characteristics of high priority customers that is independent of ρ as for computational effort [16]. Together, we are able to obtain sufficiently accurate results for the performance characteristics required, especially for the most practical cases (high values for ρ)

6.2. *Impact of system parameters on performance characteristics*

Next, we study the influence of the most important parameters of the MCMS priority system on the performance measures. From queueing theory, we know that the total utilization rate ρ and the number of servers k are interesting parameters for any queueing system. However, we have learned from the experiments with the MCMS non-priority queueing system [16] that the arrival rates fractions (a_i) and the perturbations of the service times (δ_i) might also seriously influence the performance characteristics.

We did computations for a large set of instances. Since the effects we want to discuss are already present for small systems, we shall only present the results on experiments for three servers and three customers types. One of customer types has high priority and two types have low priority. The utilization rate ρ is fixed to 95%. First of all, we want to see the influence of the difference between the service rates. This difference is defined by the difference overall service rates $\frac{\mu^l}{\mu^h}$ and difference between service rates of the low priority classes $\frac{\mu_1^l}{\mu_2^l}$. Also, we would like to see the influence of the utilization rate for high priority customers. Hence, we vary in fact three parameters: ρ^h, $\frac{\mu^l}{\mu^h}$ and $\frac{\mu_1^l}{\mu_2^l}$. The other parameters are either fixed (e.g. fractions of arrival rates within group of the low priority customer are equal to $a_1^l = 0.3$ and $a_2^l = 0.7$), or are completely defined by the other parameters (e.g. δ_1^l and δ_2^l). In this way, we have 18 experiments. We choose the values 0.5, 1 and 2 for both $\frac{\mu^l}{\mu^h}$ and $\frac{\mu_1^l}{\mu_2^l}$ and 20% and 60% for ρ^h, thereby obtaining $3 \times 3 \times 2 = 18$ model runs.

Some interesting performance characteristics are presented in Table 1: the expected total number of customers in the system for each low priority subclass $E[R_i^l]$, the

Table 1
Impact of high priority utilization and service rate ratios on the performance of low priority customers; each cell in the table contains the performance measure for both low priority customer classes.

		$\rho^h = 20\%$				$\rho^h = 60\%$			
$\frac{\mu_1^l}{\mu_2^l}$	$\frac{\mu^l}{\mu^h}$	$E[R_i^l]$	$E[q_i^l]$	$E[PS_i^l]$	$\lambda_i^{ps}/\lambda_i^l$	$E[R_i^l]$	$E[q_i^l]$	$E[PS_i^l]$	$\lambda_i^{ps}/\lambda_i^l$
0.5	0.5	6.21	5.03	0.14	0.73	4.34	3.38	0.47	2.95
		13.10	11.73	0.16	0.36	8.98	7.89	0.52	1.43
	1	6.74	5.59	0.11	0.37	5.85	4.96	0.40	1.49
		14.37	13.04	0.12	0.18	12.58	11.58	0.44	0.72
	2	7.78	6.66	0.07	0.18	8.87	8.06	0.33	0.75
		16.84	15.55	0.08	0.09	19.72	18.80	0.35	0.36
1	0.5	5.31	4.54	0.09	0.47	3.80	3.18	0.30	1.89
		12.38	10.59	0.21	0.47	8.86	7.43	0.69	1.89
	1	5.84	5.10	0.07	0.24	5.32	4.76	0.25	0.95
		13.63	11.90	0.15	0.24	12.42	11.11	0.58	0.95
	2	6.89	6.17	0.05	0.12	8.37	7.85	0.20	0.48
		16.08	14.40	0.11	0.12	19.52	18.31	0.47	0.48
2	0.5	5.27	4.83	0.05	0.27	3.65	3.30	0.17	1.08
		13.37	11.26	0.25	0.56	9.39	7.70	0.83	2.23
	1	5.82	5.39	0.04	0.14	5.20	4.88	0.14	0.54
		14.61	12.57	0.19	0.28	12.95	11.38	0.70	1.12
	2	6.88	6.46	0.03	0.07	8.27	7.97	0.11	0.27
		17.06	15.08	0.13	0.14	20.04	18.60	0.57	0.56

expected number of customers in the queue for each low priority subclass $E[q_i^l]$, the expected number of the postponed customers for each low priority subclass $E[PS_i^l]$ and the expected number of preemption events per each low priority customer entering the system $\frac{\lambda_i^{ps}}{\lambda_i^l}$.

From Table 1 we draw the following conclusions:

- The number of customers in the postponed state increases with the utilization rate of the high priority customers ρ^h (this is not a trivial result since $\rho = \rho^h + \rho^l$ is constant, hence increasing of ρ^h means decreasing of ρ^l).
- The total number of postponed customers hardly depends on the ratio $\frac{\mu_1^l}{\mu_2^l}$.
- The dependence of the total number of low priority customers in queue on the ratio $\frac{\mu_1^l}{\mu_2^l}$ is remarkable. Namely, the numbers of customers in queue (hence, the waiting times) are lower when the service times of low priority customers are equal ($\mu_1^l = \mu_2^l$) than when the service times are different. This can be interpreted as a sort of Pollaczek-Khintchine effect [15], i.e. the average waiting time is increases when the variability of the service time increases.

- The number of low priority customers in the queue (hence, the waiting times) is lower when the service rate of high priority customers is higher ($\frac{\mu^l}{\mu^h}$ is smaller).
- The number of low priority customers in the queue decreases with the utilization rate of high priority customers (ρ^h) when the service rate of the high priority customers is bigger than or equal to the average service rate of the low priority customers ($\frac{\mu^l}{\mu^h} \leq 1$) and it increases when the service rate of the high priority customers is smaller than the average service rate of the low priority customers ($\frac{\mu^l}{\mu^h} > 1$).

It is also possible to derive from this table the waiting times in the queue and in the postponed state and the total time spent in the system. This can be done using Little's law as was shown in Section 4.4.

6.3. Comparison between priority and non-priority queues

To conclude this section, we give a sketch of the effect of applying a priority queueing rule. Therefore, we vary ρ in the experiment with $N^h = 1$, $N^l = 2$, $k = 3$, $\mu_1^l/\mu_2^l = 0.5$, $\frac{\mu^l}{\mu^h} = 0.5$ and $\rho^h = 0.6\rho$, $\rho^l = 0.4\rho$. We compare the total numbers of customers in the system for cases with and without priority rules (figure 2).

The picture shows not only the fact that the introduction of priority rules increases (decreases) the total number of low (high) priority customers in the system, but also the scale of this increase (decrease). We see that there is already a significant impact of priority rule usage on the number of customers in the system if the utilization is only moderate (0.6–0.8). This confirms that the appropriate use of priority rules may provide an opportunity for efficiency gain in spare part networks. We will address this issue in our further research.

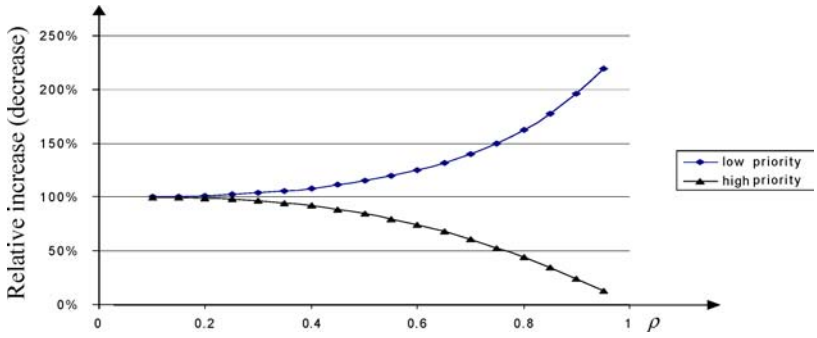

Figure 2. Relative increase (decrease) of the total number of low (high) priority customers caused by introduction of priority rules.

7. Conclusions and generalizations

In this paper, we derived a method to analyze multi-class $M/M/k$ priority queues with preemptive priority and two priority groups (high and low). Each group of priority can contain several classes of customers with different arrival and service rates. The proposed method is based on the solution of the stationary state equations. It is similar to other existing methods for priority queues [10], where it was shown that these systems can be analyzed as QBD systems in the areas with $q^h > 0$ and as $M/G/1$ type systems in the areas with $q^h = 0$. However, using the generating functions along the q^l axis allowed us to reduce the cut-off error appearing normally in analysis of such two-dimensional semi-infinite processes. So, we do not need too many iteration steps to analyze the system (see figure 1). Moreover, the multi-class queueing system presented here is more general than other priority queueing models existing in the literature.

The computational effort to find accurate results depends also on the number of customer types, the number of servers and the utilization rates of high and low priority customers (since higher utilization rates require more iterations). For example, system with 4 subclasses, 4 servers and a utilization rate of 95% needs O(20) matrix operations, where the dimension of each matrix is 238×238. Due to the increase of the size of matrices, the computational effort increases rapidly for large k, N^h and N^l. For example, in case of $k = 5$, $N^h = 2$ and $N^l = 3$, we have to deal with a state space (and matrices) of dimension 1782×1782. Approximations are necessary then. As an approximation, we replace groups of customers with similar characteristic by one customer with average service properties. Other approximations for large systems, applying the method from this paper as basis, are discussed in Van der Heijden et al. [17].

This method can in principle be extended to solve problems with more priority groups. This is possible iteratively due to the preemption property. That is, we can estimate performance estimators for each priority group ignoring all classes with lower priorities and aggregating all classes with high priorities into one high priority group. However, in this case the dimension of the state space becomes extremely large.

Also, our algorithm can in principle be used for multi-class, multi-server priority queues where customers have hyperexponential (H_x) service times. We can deal with these cases by representing each class as x classes with exponential distributed service time and adopting the performance estimators for the total number of customers in the system among these x classes.

Appendix A: Proof of Lemma 1

We will prove this lemma by induction

For $n = 0$, equation (11) can be written as

$$\mathbf{D}_{1,0}\mathbf{P}_{q^h,0} = \Lambda^h \mathbf{P}_{q^h-1,0} + \mathbf{B}\,\mathbf{P}_{q^h+1,0}$$

This equation is similar to the multi-class multi-server equilibrium equation having a solution of the form $\mathbf{P}_{q^h,0} = (\mathbf{Z}^{-1})^{q^h-1}\mathbf{C}$ [16], where $\mathbf{Z}$ with all eigenvalues >1 should satisfy the equation

$$\mathbf{D}_{1,0} = \Lambda^h \mathbf{Z} + \mathbf{B}\,\mathbf{Z}^{-1}, \tag{29}$$

similar to equation (12) with $\xi = 0$. So we have that the solution in the form (13) is the solution of the equation (11) for $q^l = 0$.

For $q^l > 0$, we first define $\tilde{\mathbf{P}}_{q^h}(\xi)$ as the solution of:

$$\mathbf{D}_{1,0}\tilde{\mathbf{P}}_{q^h}(\xi) = \Lambda^h \tilde{\mathbf{P}}_{q^h-1}(\xi) + \Lambda^l \xi \tilde{\mathbf{P}}_{q^h}(\xi) + \mathbf{B}\tilde{\mathbf{P}}_{q^h+1}(\xi) \tag{30}$$

with $\xi \in [0, 1]$, being just a parameter.

It follows that $\tilde{\mathbf{P}}_{q^h}(\xi) = (\mathbf{Z}(\xi)^{-1})^{q^h-1}\tilde{\mathbf{C}}(\xi)$. By differentiation of (30) with respect to ξ, we find

$$\mathbf{D}_{1,0}\frac{d\tilde{\mathbf{P}}_{q^h}(\xi)}{d\xi} = \Lambda^h \frac{d\tilde{\mathbf{P}}_{q^h-1}(\xi)}{d\xi} + \Lambda^l \xi \frac{d\tilde{\mathbf{P}}_{q^h}(\xi)}{d\xi} + \mathbf{B}\frac{d\tilde{\mathbf{P}}_{q^h+1}(\xi)}{d\xi} + \Lambda^l \tilde{\mathbf{P}}_{q^h}(\xi)$$

Hence $\frac{d\tilde{\mathbf{P}}_{q^h}(\xi)}{d\xi}|_{\xi=0}$ satisfies equation (11) for $q^l = 1$.

Using the general property $(\frac{d}{dx})^n(xf(x)) = n(\frac{d}{dx})^{n-1}f(x) + x(\frac{d}{dx})^n f(x)$, we find by differentiation of (30) q^l -times with respect to each parameter ξ that:

$$\begin{aligned}\mathbf{D}_{1,0}\left(\frac{d}{d\xi}\right)^{q^l}\tilde{\mathbf{P}}_{q^h}(\xi) = \Lambda^h\left(\frac{d}{d\xi}\right)^{q^l}\tilde{\mathbf{P}}_{q^h-1}(\xi) + \Lambda^l\xi\left(\frac{d}{d\xi}\right)^{q^l}\tilde{\mathbf{P}}_{q^h}(\xi)\\ + \mathbf{B}_{1,0}\left(\frac{d}{d\xi}\right)^{q^l}\tilde{\mathbf{P}}_{q^h+1}(\xi) + \Lambda^l q^l\left(\frac{d}{d\xi}\right)^{q^l-1}\tilde{\mathbf{P}}_{q^h}(\xi)\end{aligned}$$

Transforming this expression back to the $\mathbf{P}_{q^h,q^l}$ yields

$$\mathbf{P}_{q^h,q^l} = \frac{1}{q^l}\left(\frac{d}{d\xi}\right)^{q^l}(\mathbf{Z}^{-1}(\xi))^{q^h-1}\mathbf{C}(\xi)\bigg|_{\xi=0}$$

Finally, it is easy to see that:

$$\sum_{q^l}\frac{1}{q^l!}\left(\frac{d}{d\xi}\right)^{q^l}(\mathbf{Z}^{-1}(\xi))^{q^h-1}\mathbf{C}(\xi)\bigg|_{\xi=0} = (\mathbf{Z}^{-1}(\xi))^{q^h-1}\mathbf{C}(\xi)|_{\xi=1}$$

as a well-known Taylor series expansion of $(\mathbf{Z}(\xi)^{-1})^{q^h-1}\mathbf{C}(\xi)$ around $\xi = 0$, where the value of $\mathbf{Z}(\xi)$ is found from

$$\mathbf{D}_{1,0} = \Lambda^h \mathbf{Z} + \Lambda^l \xi + \mathbf{B}\,\mathbf{Z}^{-1}.$$

with all eigenvalues $\mathbf{Z}(\xi)$ in absolute value larger than 1. This leads to convergence of the Taylor series, assuming analyticity of $\mathbf{C}(\xi)$ for $|\xi| < 1 + \varepsilon$ for some $\varepsilon > 0$. As we have mentioned above, this equation can be solved as in the case of the non-priority multi-class queue [16]. So the solution has the form (13).

Appendix B: Derivation of the relations in (23)

From (17), it follows by differentiating $t > 2$ times with respect to ξ that:

$$\binom{t}{0}\mathrm{H}(0)\frac{d^t}{d\xi^t}v(0) + \binom{t}{1}\mathrm{H}'(0)\frac{d^{t-1}}{d\xi^{t-1}}v(0) + \cdots + \binom{t}{t}\frac{d^t}{d\xi^t}\mathrm{H}(0)v(0) = 0.$$

We can also write this in the form

$$\sum_{i=0}^{t} h_i \mathrm{P}_{t+1-i} = 0$$

where $\mathrm{P}_{i+1} = \frac{1}{i!}\frac{d^i}{d\xi^i}v(0)$ and $h_i = \frac{1}{i!}\frac{d^i}{d\xi^i}\mathrm{H}(0)$.

Now let us show recursively for $t^* = t, t-1, \ldots, 2$ that $\mathbf{\Theta}_0^{t^*}\mathrm{P}_{t+1} + \sum_{i=1}^{t^*}\mathbf{\Theta}_i^{t^*}\mathrm{P}_{t^*+1-i} = 0$.

For $t^* = t$, it is clear from the equation for P_{t+1} that we can take $\mathbf{\Theta}_i^t = h_i$.

For $t^* \leq t$, we have the equation derived in the previous induction step and also the original equation for P_{t^*+1}:

$$\begin{aligned}\mathbf{\Theta}_0^{t^*+1}\mathrm{P}_{t+1} + \mathbf{\Theta}_1^{t^*+1}\mathrm{P}_{t^*+1} + \cdots + \mathbf{\Theta}_{t^*+1}^{t^*+1}\mathrm{P}_1 &= 0\\ h_0\mathrm{P}_{t^*+1} + \cdots + h_{t^*}\mathrm{P}_1 &= 0\end{aligned}$$

Multiplying the first equation by $h_0(\mathbf{\Theta}_1^{t^*+1})^{-1}$ and taking the difference, we can eliminate the term P_{t^*+1}. We obtain

$$\begin{aligned}&h_0(\mathbf{\Theta}_1^{t^*+1})^{-1}\mathbf{\Theta}_0^{t^*+1}\mathrm{P}_{t+1} + (h_0(\mathbf{\Theta}_1^{t^*+1})^{-1}\mathbf{\Theta}_2^{t^*+1} - h_1)\mathrm{P}_{t^*}\\ &+ \cdots + (h_0(\mathbf{\Theta}_1^{t^*+1})^{-1}\mathbf{\Theta}_{t^*+1}^{t^*+1} - h_{t^*})\mathrm{P}_1 = 0\end{aligned}$$

or in other terms

$$\mathbf{\Theta}_0^{t^*}\mathrm{P}_{t+1} + \mathbf{\Theta}_1^{t^*}\mathrm{P}_{t^*} + \cdots + \mathbf{\Theta}_{t^*}^{t^*}\mathrm{P}_1 = 0$$

where the matrices $\mathbf{\Theta}_i^{t^*}$ are equal to

$$\begin{aligned}\mathbf{\Theta}_0^{t^*} &= h_0\big(\mathbf{\Theta}_1^{t^*+1}\big)^{-1}\mathbf{\Theta}_0^{t^*+1}\\ \mathbf{\Theta}_i^{t^*} &= \big(h_0\big(\mathbf{\Theta}_1^{t^*+1}\big)^{-1}\mathbf{\Theta}_{i+1}^{t^*+1} - h_i\big), \quad i = 1, \ldots t^*\end{aligned}$$

Herewith, the relations in (23) are shown.

References

[1] J.P. Buzen and A.B. Bondi, The response times of priority classes under preemptive resume in $M/M/m$ queues, Operations Research 31(3) (1983) 456–465.

[2] J.H.A. de Smit, A numerical solution for the multi-server queue with hyperexponential service time, Operations Research Letters 2(5) (1983) 217–224.

[3] H.R. Gail, S.L. Hantler and B.A. Taylor, Analysis of a non-preemptive priority multiserver queue, Advances in Applied Probability 20(4) (1988) 852–879.

[4] H.R. Gail, S.L. Hantler and B.A. Taylor, On preemptive Markovian queue with multiple servers and two priority classes, Mathematics of Operations Research 17(2) (1992) 365–391.

[5] G. Hooghiemstra, M. Keane and S. van de Ree, Power series for stationary distributions of coupled processor models, SIAM Journal on Applied Mathematics 48(5) (1988) 1159–1166.

[6] E.P.C. Kao and K.S. Narayanan, Computing steady state probabilities of a nonpreemptive priority queue, ORSA Journal on Computing 2 (1990) 211–218.

[7] E.P.C. Kao and S.D. Wilson, Analysis of nonpreemptive priority queues with multiple servers and two priority classes, European Journal of Operational Research 118 (1999) 181–193.

[8] O. Kella and U. Yechiali, Waiting times in the non-preemptive priority $M/M/c$ queue, Communications in Statistics—Stochastic Models 1 (1985) 257–262.

[9] C. Maglaras, Dynamic scheduling in multiclass queueing networks, Queueing Systems 31 (1999) 171–206.

[10] D.R. Miller, Computation of steady-state probabilities for M/M/1 priority queues, Operations Research 29(5) (1981) 945–958.

[11] I. Mitrani and P.J.B. King, Multiprocessor systems with preemptive priorities, Performance Evaluation 1 (1981) 118–125.

[12] M.I. Reiman and L.M. Wein, Dynamic scheduling of a two-class queue with setups, Operations Research 46(4) (1998) 532–537.

[13] C.C. Sherbrooke, *Optimal Inventory Modelling of Systems: Multi-Echelon Techniques*, (Wiley, New York 1992).

[14] A. Sleptchenko, M.C. van der Heijden and A. van Harten, Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems, International Journal of Production Economics, 79 (2002) 109–230.

[15] H.C. Tijms, *Stochastic Models: An Algorithmic Approach*, (John Willey and Sons, Chichester, 1994).

[16] A. van Harten and A. Sleptchenko, On multi-class, multi-server queuing and spare part management, Queueing Systems 43(4) (2003) 307–328.

[17] M.C. van der Heijden, A. Sleptchenko and A. van Harten, Approximations for Markovian multi-class queues with preemptive priorities, working paper, University of Twente, Faculty of Business, Public Administration and Technology (2002) (submitted for publication).

[18] D. Wagner, Waiting time of a finite-capacity multi-server model with non-preemptive priorities, European Journal of Operational Research 102 (1997) 227–241.

[19] D. Wagner, Analysis of mean values of a multi-server model with non-preemptive priorities and non-renewal input, Communications in Statistics—Stochastic Models 13(1) (1997) 67–84.

[20] D. Wagner, A finite-capacity multi-server multi-queueing model with non-renewal input, Annals of Operations Research 79 (1998) 63–82.

[21] L.M. Wein, Dynamic scheduling of a multiclass make-to-stock queue, Operations Research 40(4) (1992) 724–735.