Queueing Systems 38, 397–418, 2001

The Versatility of MMAP[K] and the MMAP[K]/G[K]/1 Queue

QI-MING HE Qi-Ming.He@dal.ca
Department of Industrial Engineering, DalTech, Dalhousie University, Halifax, NS, Canada B3J 2X4

Abstract. This paper studies a single server queueing system with multiple types of customers. The first part of the paper discusses some modeling issues associated with the Markov arrival processes with marked arrivals (MMAP[K], where K is an integer representing the number of types of customers). The usefulness of MMAP[K] in modeling point processes is shown by a number of interesting examples. The second part of the paper studies a single server queueing system with an MMAP[K] as its input process. The busy period, virtual waiting time, and actual waiting times are studied. The focus is on the actual waiting times of individual types of customers. Explicit formulas are obtained for the Laplace–Stieltjes transforms of these actual waiting times.

Keywords: queueing system, waiting time, multiple types of customer, point process, matrix analytic methods

AMS subject classification: primary 60K25

1. Introduction

The objective of this paper is to show how to model point processes with multiple types of customers by using Markov arrival processes with marked arrivals (MMAP[K]) and to study a single server queueing system with multiple types of customers.

This study was motivated by the potential applications of the results obtained in this paper in telecommunications, manufacturing, and service systems where multiple types of customers are present. In telecommunications, systems such as switch centers are required to handle different types of data (e.g., voice, video, and facsimile) simultaneously. The nature of these data processes is usually dramatically different. Some of them are bursty, some are in high volume, some cannot afford any loss, and some cannot have long delays. These data compete for system resource and, thus, have mutual influence when they travel in a network. Therefore, to do system performance analysis, it is useful to understand how each type of data behaves in the network. In manufacturing, a customer order may consist of several suborders (for different components). Thus, the demand processes of individual components are dependent and so should be their replenishment (or production) processes. To gain insights into such production systems, an analysis at the component level is as important as an analysis at the aggregate level. In the service industry, customers can be distinguished into different groups each requiring a particular type of service. It is important to understand how a system serves individual types of customers when they compete for the same resource.

For all the above cases, a formal and convenient formalism of the input process is vital in system modeling. A detailed analysis of the corresponding (queueing) systems considering the behavior or performance of individual types of customers is useful in practice. It is worth pointing out that even for queueing systems with only one type of customer, when the input process is bursty, the waiting times of different customers can be significantly different. Thus, it makes sense to distinguish customers into subgroups and analyze the queueing processes of individual groups of customers respectively (see [8, example 2]).

This paper suggests the use of MMAP[K] to model point processes with multiple types of customers. MMAP[K] is a generalization of Markov arrival processes (MAP) which have been studied and used extensively in queueing theory. MAP was introduced in [18] to model non-Markovian point processes. MAP or some of its special cases are widely used by researchers and practitioners in telecommunications and manufacturing (see [4,14–17,21,22,26,27]). While MAP is a useful tool to model point processes with one type of customer, MMAP[K] is useful when multiple types of customers are present. MMAP[K] was introduced by Neuts (see [11]). Asmussen and Koole [3] introduced MMAP[K] independently. Closely related work can also be found in [24,25], where the term MMAP was introduced for special cases of multivariate Markov additive processes. In this paper, it will be shown that MMAP[K] provides flexibility in modeling correlated point processes with special arrival patterns. It will also be shown that queueing systems with MMAP[K] as input processes are analytically and computationally tractable.

Queueing systems with multiple types of customers have been studied extensively when the input processes are independent Poisson processes and priorities are assigned to different types of customers (see [29–31 and references therein]). Useful results have been obtained for system stability conditions, queue length, waiting time, and busy period distributions. To extend applications of these results, queueing systems with more general input processes were introduced and studied. Takine et al. [34] and Yeung and Sengupta [36] considered Markov modulated Poisson processes and superposition processes of Markov arrival processes. HE [9] and HE and Alfa [10] studied queueing systems with MMAP[K], PH-distributed service times, and a last-come-first-served service discipline. In [8,33], a queueing system with dependent arrival processes of multiple types of customers and a first-come-first-served (FCFS) service discipline was studied. Some results were obtained for the fundamental periods, queue length, and waiting times. This paper generalizes most of the results obtained in [8].

The work of Takine and Hasegawa [33] (also see [32]) is closely related to this paper. In fact, the analysis of the busy period, busy cycle, idle probabilities, virtual waiting times of the queueing system studied in this paper can be carried out by using results obtained in [33]. The differences between this paper and [33] are:

(1) this paper discusses some modeling issues related to MMAP[K];

(2) this paper considers the actual waiting time of any particular type of customer and derives the corresponding Laplace–Stieltjes transform.

Compared to [8], this paper gives a more detailed discussion on the modeling issues of MMAP[K]. Queueing systems with general batch arrivals are considered in this paper, while [8] focused on the case where each batch has only one customer. This paper also considers the service sequence of customers within each batch, an issue that is rarely discussed in the existing literature. In addition, this paper presents some of the proofs given in [8] in a formal way. New insights into why some of the methods work are gained and explained.

The study of the waiting times in a single server queue dates back to the Pollaczek–Khinchin formula for the M/G/1 queue (see [6,12]). For the M/G/1 queue and its related Poisson arrival queueing systems, PASTA [35] guarantees that the actual waiting time and the virtual waiting time have the same distribution. However, for queueing systems with more general input processes (e.g., BMAP), PASTA does not hold, i.e., the virtual waiting time and the actual waiting time have different probability distributions. For the BMAP/G/1 queue, Pollaczek–Khinchin type formulas for virtual and actual waiting times have been found (see [14–17,19,20,26]). When multiple types of customers are present, not only are the virtual waiting time and the actual waiting time different, but also the actual waiting times of individual types of customers are different (see [8]). Therefore, an analysis of the waiting processes of individual types of customers is necessary to gain insights into such queueing systems. Such an analysis is conducted in sections 5 and 6 of this paper.

The rest of the paper is organized as follows. Section 2 gives the definition of MMAP[K] and presents some limiting properties associated with MMAP[K]. Section 3 shows the usefulness of MMAP[K] by presenting a number of interesting examples and introduces the MMAP[K]/G[K]/1 queue with a FCFS service discipline. Section 4 presents some results about the busy period and the Laplace–Stieltjes transform of the virtual waiting time. Sections 5 and 6 study the actual waiting times of individual types of customers. In section 5, the Laplace–Stieltjes transforms of actual waiting times of various types of customers are presented with intuitive proofs. Formal proofs of these results are given in section 6. Finally, in section 7, results obtained in this paper are summarized and some discussion is given to future research.

2. Definition of MMAP[K]

A Markov arrival process with marked arrivals (MMAP[K]) is a stochastic point process with multiple arrivals (batches) occurring in continuous time or discrete time. Each arrival (batch) represents the arrivals of a number of customers into the system of interest. There are K different types of customers. (Throughout of this paper, the words "arrival" and "batch" are equivalent. The word "arrival" is used mostly in the first part of the paper, while "batch" is used mainly in the second part of the paper.) For later use in this paper, a definition of MMAP[K] is given to the continuous time case. This definition was given by Neuts in [11]. Let

$$\aleph = \{\mathbf{h}\colon\ \mathbf{h} = h_1 h_2 \dots h_n,\ 1 \leqslant h_i \leqslant K,\ 1 \leqslant i \leqslant n,\ 1 \leqslant n < \infty\}, \tag{2.1}$$

where K is a positive integer that represents the number of customer types. Each $\mathbf{h} \in \aleph$ represents the types of customers within an arrival and their relative order. The length of $\mathbf{h} \in \aleph$, denoted by $|\mathbf{h}|$, is the number of integers (customers) in $\mathbf{h}$.

Consider an m-state Markov renewal process with an irreducible embedded Markov chain with transition probability matrix $P = (p_{i,j})$ and exponential sojourn time distributions in state i given by $1 - \exp\{-\sigma_i x\}$, for $1 \leqslant i, j \leqslant m$. This is also a Markov chain in continuous time. Let D be the infinitesimal generator of this Markov chain. The matrix D and the parameters P and σ_i, $1 \leqslant i \leqslant m$, of the Markov renewal process are related by $D_{i,i} = -(1 - p_{i,i})\sigma_i$, for $1 \leqslant i \leqslant m$, and $D_{i,j} = p_{i,j}\sigma_i$, for $1 \leqslant i, j \leqslant m, i \neq j$. Let $J(t)$ denote the state of this Markov renewal process at time t. $J(t)$ is called the underlying Markov process with infinitesimal generator D.

Define a Markov renewal process $\{(J_n, L_n, \tau_n),\ n \geqslant 0\}$ on the state space $\{[\{1, \ldots, m\} \times \aleph] \times [0, \infty)\}$ with the transition probability matrix, for $1 \leqslant i, j \leqslant m$, $\mathbf{h} \in \aleph, i \neq j,\ x \geqslant 0$,

$$\mathbf{P}\{J_n = j,\ L_n = \mathbf{h},\ \tau_n \leqslant x \mid J_{n-1} = i\} = \left[\int_0^x \exp\{D_0 u\}\, \mathrm{d}u D_{\mathbf{h}}\right]_{i,j}, \tag{2.2}$$

where L_n is the marking variable. The matrices $\{D_{\mathbf{h}},\ \mathbf{h} \in \aleph\}$ are nonnegative. The matrix D_0 has negative diagonal elements and nonnegative off-diagonal elements. D_0 is assumed to be nonsingular. The relationship between the infinitesimal generator D and $\{D_0,\ D_{\mathbf{h}},\ \mathbf{h} \in \aleph\}$ is

$$D = D_0 + \sum_{\mathbf{h} \in \aleph} D_{\mathbf{h}}. \tag{2.3}$$

An arrival is called a type $\mathbf{h}$ arrival (batch) if the arrival is marked by $\mathbf{h} \in \aleph$. The marking rate (matrix) of type $\mathbf{h}$ arrivals is $D_{\mathbf{h}}$. A type $\mathbf{h}$ arrival includes $|\mathbf{h}|$ customers whose types are $\{h_1, h_2, \ldots, h_{|\mathbf{h}|}\}$ and sequenced within the arrival accordingly, i.e., the first customer in the arrival is a type h_1 customer, the second a type h_2 customer, ..., the last a type $h_{|\mathbf{h}|}$ customer. Let $N_k(t)$ be the total number of type k customers who arrived in $(0, t)$. The set $\{N_k(t),\ 1 \leqslant k \leqslant K\}$ consists of the counting processes of individual types of customers. Denote by

$$\begin{aligned} &p_{i,j}(n_1, \ldots, n_K, t) = \mathbf{P}\bigl\{N_1(t) = n_1, \ldots, N_K(t) = n_K,\ J(t) = j \mid J(0) = i\bigr\}, \\ &\quad 1 \leqslant i, j \leqslant m, \end{aligned} \tag{2.4}$$

and $P(n_1, \ldots, n_K, t)$ an $m \times m$ matrix with elements $p_{i,j}(n_1, \ldots, n_K, t)$. Let $P^*(z_1, \ldots, z_K, t)$ be the moment generating function of $P(n_1, \ldots, n_K, t)$. It can be proved that

$$P^*(z_1, \ldots, z_K, t) \equiv \mathbf{E} z_1^{N_1(t)} \cdots z_K^{N_K(t)} = \exp\left\{\left[D_0 + \sum_{\mathbf{h} \in \aleph}\left(\prod_{n=1}^{|\mathbf{h}|} z_{h_n}\right) D_{\mathbf{h}}\right] t\right\}, \tag{2.5}$$

where $\equiv$ represents definition and $\mathbf{E}$ represents mathematical expectation.

Denote by $\boldsymbol{\theta}$ the stationary probability vector of the matrix D, i.e., $\boldsymbol{\theta} D = 0, \boldsymbol{\theta}\mathbf{e} = 1$, where $\mathbf{e}$ is the column vector with all components one (see [7] for more about matrix theory). Using equation (2.5), it is easy to verify that the stationary arrival rate of type $\mathbf{h}$ arrivals is given by $\lambda_{\mathbf{h}} = \boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e}$ for $\mathbf{h} \in \aleph$ (see [23, appendix]). Intuitively, the number of type $\mathbf{h}$ arrivals during the period $(t, t + \delta t)$ is given by $\lambda_{\mathbf{h}}\delta t$, a basic fact that will be used repeatedly. Furthermore, the arrival rates of individual types of customers, a group of customers, and all the customers can be defined and obtained in a similar manner. For instance, the following stationary arrival rates can be obtained easily. For $\mathbf{h} = h_1 \dots h_{|\mathbf{h}|}$ and $1 \leqslant n \leqslant |\mathbf{h}|$, let

$\lambda_{\mathbf{h},k,n} \equiv \boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e}\, I\{h_n = k\}$: the stationary arrival rate of type k customers which come from type $\mathbf{h}$ arrivals and are the nth customer within their corresponding arrivals;

$\lambda_{\mathbf{h},k} \equiv \sum_{n=1}^{|\mathbf{h}|} \boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e}\, I\{h_n = k\} \left(= \sum_{n=1}^{|\mathbf{h}|} \lambda_{\mathbf{h},k,n}\right)$: the stationary arrival rate of type k customers which come from type $\mathbf{h}$ arrivals;

$\lambda_k \equiv \sum_{\mathbf{h}\in\aleph} \sum_{n=1}^{|\mathbf{h}|} \boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e} I\{h_n = k\} \left(= \sum_{\mathbf{h}\in\aleph} \sum_{n=1}^{|\mathbf{h}|} \lambda_{\mathbf{h},k,n}\right)$: the stationary arrival rate of type k customers;

$\lambda_b \equiv \sum_{\mathbf{h}\in\aleph} \boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e} \left(= \sum_{\mathbf{h}\in\aleph} \lambda_{\mathbf{h}}\right)$: the stationary arrival rate of arrivals (batches);

$\lambda \equiv \sum_{\mathbf{h}\in\aleph} |\mathbf{h}|\boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e} \left(= \sum_{\mathbf{h}\in\aleph} |\mathbf{h}|\lambda_{\mathbf{h}}\right)$: the stationary arrival rate of all customers regardless of their types,

where $I\{\cdot\}$ is the indication function. Some conditional probabilities and limiting results of MMAP[K], which shall be used in later sections, can be obtained immediately.

Let $p_{i,j}(\mathbf{h})$ be the probability that the state of the underlying Markov process $J(t)$ right after a type $\mathbf{h}$ arrival is j, given that the arrival is of type $\mathbf{h}$ and the underlying Markov process was in state i just before the arrival at an arbitrary time. Consider the event that there is a type $\mathbf{h}$ arrival in $(t, t + \delta t)$. In steady state, the probability that there is a type $\mathbf{h}$ arrival in $(t, t + \delta t)$ is $\boldsymbol{\theta} D_{\mathbf{h}}\mathbf{e}\delta t = \lambda_{\mathbf{h}}\delta t$. The probability that the state changes from i to j is $(D_{\mathbf{h}})_{i,j}\delta t$. Then

$$p_{i,j}(\mathbf{h}) = \lim_{\delta t\to 0} \frac{(D_{\mathbf{h}})_{i,j}\delta t + \mathrm{o}(\delta t)}{\lambda_{\mathbf{h}}\delta t + \mathrm{o}(\delta t)} = \frac{(D_{\mathbf{h}})_{i,j}}{\lambda_{\mathbf{h}}}. \tag{2.6}$$

Let $p_{i,j}(\mathbf{h}, k, n)$ be the probability that the state of the underlying Markov process $J(t)$ right after the arrival of a type k customer which is the nth customer in a type $\mathbf{h}$

arrival is j, given that the arrival is of type $\mathbf{h}$ and the underlying Markov process was in state i just before the arrival. Then

$$p_{i,j}(\mathbf{h}, k, n) = \frac{(D_{\mathbf{h}})_{i,j} I\{h_n = k\}}{\lambda_{\mathbf{h}}}. \tag{2.7}$$

Similarly, let $p_{i,j}(k)$ be the probability that the state of the underlying Markov process $J(t)$ right after the arrival of a type k customer is j, given that a type k customer just came in and the underlying Markov process was in state i just before the customer arrives. Let $p_{i,j}(\text{an arrival})$ be the probability that the state of the underlying Markov process $J(t)$ right after an arrival is j, given that there is an arrival and the underlying Markov process was in state i just before the arrival at an arbitrary time. Then

$$p_{i,j}(k) = \frac{\sum_{\mathbf{h}\in\aleph} \sum_{n=1}^{|\mathbf{h}|} (D_{\mathbf{h}})_{i,j} I\{h_n = k\}}{\lambda_k} \quad \text{and} \quad p_{i,j}(\text{an arrival}) = \frac{\sum_{\mathbf{h}\in\aleph} (D_{\mathbf{h}})_{i,j}}{\lambda}. \tag{2.8}$$

Property 2.1. For the MMAP[K] defined above, we have the following useful conclusions.

(a) The probability that an arbitrary arrival (batch) is of type $\mathbf{h}$ is $\lambda_{\mathbf{h}}/\lambda_b$.

(b) The probability that an arbitrary type k customer is from a type $\mathbf{h}$ arrival is $\lambda_{\mathbf{h},k}/\lambda_k$.

(c) The probability that an arbitrary type k customer is from the nth position of a type $\mathbf{h}$ arrival is $\lambda_{\mathbf{h},k,n}/\lambda_k$.

(d) The probability that an arbitrary customer is of type k is λ_k/λ.

Finally in this section, the ratio of the number of total customers arrived and the number of a particular type of batch arrived is obtained. Let $\xi_{\mathbf{h}}(n)$ be the number of type $\mathbf{h}$ batches that arrived before and when the nth customer arrives. Then $\lim_{\{n\to\infty\}} n/\xi_{\mathbf{h}}(n)$ is the average number of customers who arrived between two consecutive type $\mathbf{h}$ batches. Let $N(\mathbf{h},\ \mathbf{h})$ be the total number of customers who arrived between two consecutive type $\mathbf{h}$ batches (including one type $\mathbf{h}$ batch). The moment generating function of $N(\mathbf{h},\ \mathbf{h})$ can be obtained as

$$\mathbf{E}z^{N(\mathbf{h},\,\mathbf{h})} = \frac{\boldsymbol{\theta} D_{\mathbf{h}}}{\lambda_{\mathbf{h}}} \left(D_0 + \sum_{\mathbf{L}\neq\mathbf{h},\mathbf{L}\in\aleph} z^{|\mathbf{L}|} D_{\mathbf{L}} \right)^{-1} D_{\mathbf{h}}\mathbf{e} + z^{|\mathbf{h}|}. \tag{2.9}$$

Differentiating expression (2.9) with respect to z and setting $z = 1$, yields $\mathbf{E}N(\mathbf{h}, \mathbf{h}) = \lambda/\lambda_{\mathbf{h}}$. Then it is easy to prove that

$$\lim_{n\to\infty} \frac{\xi_{\mathbf{h}}(n)}{n} = \frac{1}{\mathbf{E}N(\mathbf{h}, \mathbf{h})} = \frac{\lambda_{\mathbf{h}}}{\lambda}. \tag{2.10}$$

More asymptotic and limiting results can be derived for MMAP[K]. Nonetheless, only these to be used later in this paper are presented in this section. Strict proofs of all the conclusions can be obtained following the approaches used for the special case: Markov arrival processes. See [23] for more details.

3. Modeling with MMAP[K] and the MMAP[K]/G[K]/1 queue

In this section, a number of special cases and examples of MMAP[K] are presented and the MMAP[K]/G[K]/1 queue is introduced as well. The objective of this section is to demonstrate the usefulness of MMAP[K] in the modeling of point processes and, hence, the potential applications of MMAP[K] to telecommunications, manufacturing, and service industries.

Special case 3.1. The superposition process of K independent Poisson processes is an MMAP[K]. Suppose that the arrival rates of the K Poisson processes are $\{\lambda_1, \lambda_2, \dots, \lambda_K\}$. The matrix representation of its corresponding MMAP[K] is $D_0 = -(\lambda_1 + \dots + \lambda_K)$, $D_1 = \lambda_1, \dots,$ and $D_K = \lambda_K$.

Special case 3.2. A batch Markov arrival process (BMAP) with matrix representation $\{D_0,\ D_1,\ D_{11},\ D_{111},\ \dots\}$ is an MMAP[K] with $K = 1$. $D_{\mathbf{h}}$ is the arrival rate (matrix) of batches of the size $|\mathbf{h}|$ for $\mathbf{h} = 1 \dots 1$. See [14,15,17,18] for more details about BMAP.

Special case 3.3. An MMAP[K] with matrix representation $\{D_0,\ D_1,\ D_2,\ \dots,\ D_K\}$ describes a point process with K types of customers. Each arrival (batch) consists of a single customer. This type of MMAP[K] has been studied in [11] and used as input processes to queueing systems in [8–10].

Next, a few interesting examples are presented to show how to use MMAP[K] to model stochastic point processes with a special arrival pattern.

Example 3.4. Consider a point process with 2 types of customers. An arrival may consist of a single type 1 customer, a single type 2 customer, or a type 1 customer and a type 2 customer. Such point processes can be modeled by MMAP[2] with a carefully chosen underlying Markov process $J(t)$, and matrices $\{D_0, D_1, D_2, D_{12}\}$. When the order of the two customers in an arrival (batch) must be considered for some reason, D_{12} can then be split into two matrices $\{D_{12}, D_{21}\}$ to distinguish a $\{12\}$ batch from a $\{21\}$ batch.

Example 3.5. When the arrivals of a point process with 2 types of customers possess a cyclic pattern, the process can be modeled by the following MMAP[2]:

$$D_0 = \begin{pmatrix} d_{0,11} & 0 \\ 0 & d_{0,21} \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0 & d_{1,12} \\ 0 & 0 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 0 & 0 \\ d_{2,21} & 0 \end{pmatrix}, \tag{3.1}$$

where $d_{0,11}$, $d_{0,21}$, $d_{1,12}$, and $d_{2,21}$ are matrices with dimensions chosen properly. Matrices $d_{0,11}$ and $d_{0,21}$ have negative diagonal elements and nonnegative off-diagonal elements. Matrices $d_{1,12}$ and $d_{2,21}$ are nonnegative. In this MMAP[2], a type 2 customer follows a type 1 customer and a type 1 customer follows a type 2 customer. With this formulation, not only the sequence of customers is modeled, but also the interarrival times between arrivals of different types can be specified.

Example 3.6. Consider a point process with two types of customers. It is observed that any type 2 customer is followed by at least one type 1 customer. Such a point process can be modeled by an MMAP[2] with a matrix representation:

$$D_0 = \begin{pmatrix} d_{0,11} & 0 \\ 0 & d_{0,21} \end{pmatrix}, \qquad D_1 = \begin{pmatrix} d_{1,11} & d_{1,12} \\ 0 & 0 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 0 & 0 \\ d_{2,21} & 0 \end{pmatrix}, \tag{3.2}$$

where $d_{0,11}$, $d_{0,21}$, $d_{1,11}$, $d_{1,12}$, and $d_{2,21}$ are nonzero matrices with dimensions chosen properly. Matrices $d_{0,11}$ and $d_{0,21}$ have negative diagonal elements and nonnegative off-diagonal elements. Matrices $d_{1,11}$, $d_{1,12}$, and $d_{2,21}$ are nonnegative.

Example 3.7. Consider a point process with three types of customers. Priorities exist among different types of customers within each batch. However, such priorities may change from batch to batch. MMAP[K] can be used to model such point processes. For instance, an MMAP[K] with $\{D_0,\ D_{121},\ D_{112},\ D_{312},\ D_{113},\ D_{32}\}$ describes a point process in which type 3 customers have higher priorities over type 1 customers when a type 2 customer is present; otherwise, type 1 customers have higher priorities over type 3 customers in a batch.

In general, point processes with multiple types of arrivals and more complicated arrival patterns can be modeled by using MMAP[K].

The MMAP[K]/G[K]/1 queue. The MMAP[K]/G[K]/1 queue is a single server queueing system with an MMAP[K] input process. The input process is represented by matrices $\{D_0, D_{\mathbf{h}}, \mathbf{h} \in \aleph\}$. There are K types of customers. The service times of type k customers have a common distribution function $F_k(x)$ with Laplace–Stieltjes transform $f_k^*(x)$ and mean service time $1/\mu_k$, $1 \leqslant k \leqslant K$. We assume that all service times are independent of each other and independent of the input process.

All customers join a single queue and are served based on a first-come-first-served (FCFS) basis. However, service priorities may be assigned to customers in the same batch. That is, within each batch, customers are served as they are sequenced. For example, for a batch of the type $\mathbf{h} = 1231$, the first type 1 customer is served first, then the type 2 customer, then the type 3 customer, and finally the other type 1 customer. Because of the flexibility of MMAP[K] in modeling the input process, a variety of service priorities within batches can be included.

Define the traffic intensity of the queueing system as

$$\rho = \sum_{k=1}^{K} \frac{\lambda_k}{\mu_k}. \tag{3.3}$$

It is assumed that the traffic intensity $\rho < 1$ so that the queueing system can reach its steady state.

To end this section, we present an example to show that the consideration of the correlation or pattern in the arrival process does make a difference in the performance analysis of queueing systems with multiple types of customers.

Example 3.8. Consider a queueing system with an MMAP[2] input process. The service times of the two types of customers have exponential distributions with parameters μ_1 and μ_2, respectively, i.e., $f_1^*(s) = \mu_1/(s+\mu_1)$ and $f_2^*(s) = \mu_2/(s+\mu_2)$. For the input process, we consider two cases:

$$\begin{aligned}
&\text{(a)} \quad D_0 = \begin{pmatrix} -t_1 & 0 \\ 0 & -t_2 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0 & t_1 \\ 0 & 0 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 0 & 0 \\ t_2 & 0 \end{pmatrix}, \\
&\qquad t_1 > 0,\ t_2 > 0; \\
&\text{(b)} \quad D_0 = -(\lambda_1 + \lambda_2), \qquad D_1 = D_2 = \lambda_1 = \lambda_2\ (= t_1 t_2/(t_1 + t_2)).
\end{aligned} \tag{3.4}$$

For the two input processes, the average arrival rates of type 1 and type 2 customers are the same. Using the results obtained in sections 4 and 5, the Laplace–Stieltjes transforms of the waiting times of type 1 customers in steady state (for the two cases) can be obtained as

$$\begin{aligned}
&\text{(a)} \quad \mathbf{w}_1^*(s)\mathbf{e} = \frac{s[y_{01}(s+\mu_2)(s-t_2) - y_{02}t_2\mu_2](s+\mu_1)(t_1+t_2)}{[(s+\mu_1)(s+\mu_2)(s-t_1)(s-t_2) - t_1t_2\mu_1\mu_2]t_2}, \\
&\text{(b)} \quad w_1^*(s) = y_0 \frac{(s+\mu_1)(s+\mu_2)}{s^2 + s(\mu_1+\mu_2-\lambda_1-\lambda_2) + \mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1},
\end{aligned} \tag{3.5}$$

where vector $\mathbf{y}_0 = (y_{01}, y_{02})$ and y_0 are the (vector) probabilities that the queueing system is empty at an arbitrary time when the input processes are given by part (a) and (b) in equation (3.4), respectively. It is easy to see that $\mathbf{y}_0\mathbf{e} = y_0 = 1 - \lambda_1/\mu_1 - \lambda_2/\mu_2$. More details about $\mathbf{y}_0$ can be found in section 4. The difference between expressions (a) and (b) in equation (3.5) shows that the modeling of the input process is useful and may be necessary to obtain good approximations of performance measures.

4. The busy period and the virtual waiting time

The busy period of the MMAP[K]/G[K]/1 queue is only related to the total service time of each batch. Thus, the discussion can be confined at the batch level, not at the individual customer level. It is then convenient to consider individual batches as *super customers*. The service time of a super customer of the type $\mathbf{h} \in \aleph$ is the summation of the service times of all customers within the batch. The distribution function and the corresponding Laplace–Stieltjes transform of the service time of $\mathbf{h} = h_1 \dots h_n \in \aleph$ are given by

$$F_{\mathbf{h}}(x) = F_{h_1} * F_{h_2} * \cdots * F_{h_n}(x) \quad \text{and} \quad f_{\mathbf{h}}^*(s) = f_{h_1}^*(s) f_{h_2}^*(s) \cdots f_{h_n}^*(s). \tag{4.1}$$

For the busy period of the MMAP[K]/G[K]/1 queue, many results obtained in [8,33] can be applied immediately with minor changes. For instance, the exponential equations given in [8, theorem 6.1] for the joint transform of the length of a busy period and the numbers of different types of customers served in the busy period still hold. Using these equations, expressions of the moments of the length of a busy period and numbers of customers served in a busy period can be derived. Since these results are straightforward generalization and are not used in later sections, details are not presented. For later use, the distribution of the length of an arbitrary busy period (regardless of the type of the first arrival of the busy period) is discussed. The results are similar to those obtained in [33] and proofs are omitted.

Let $\psi_{i,j}(x, y)$ be the probability that the length of a busy period is y or less, the state of the underlying Markov process $J(t)$ is j when the busy period ends, given that the initial state is i and the initial workload is x. $\Psi(x, y)$ is an $m \times m$ matrix with elements $\psi_{i,j}(x, y)$. $\Psi^*(x, s)$ is the Laplace–Stieltjes transform of $\Psi(x, y)$ with respect to y. Define

$$Q^*(s) = -s\mathbf{I} + D_0 + \sum_{\mathbf{h} \in \aleph} D_{\mathbf{h}} \int_0^\infty \Psi^*(x, s) F_{\mathbf{h}}(\mathrm{d}x), \tag{4.2}$$

where $\mathbf{I}$ is the identity matrix. Then it can be shown that matrix $Q^*(s)$ satisfies ([33])

$$Q^*(s) = -s\mathbf{I} + D_0 + \left[\sum_{\mathbf{h} \in \aleph} D_{\mathbf{h}} \int_0^\infty F_{\mathbf{h}}(\mathrm{d}x) \right] \exp\{Q^*(s)x\}. \tag{4.3}$$

Let $Q = \lim_{s \to 0+} Q^*(s)$. It can be proved that matrix Q is the infinitesimal generator of the underlying Markov process that is obtained by excising the busy periods. Matrix Q satisfies

$$Q = D_0 + \left[\sum_{\mathbf{h} \in \aleph} D_{\mathbf{h}} \int_0^\infty F_{\mathbf{h}}(\mathrm{d}x) \right] \exp\{Qx\}. \tag{4.4}$$

Detailed interpretations of Q can be found in [2,33]. A computation method of Q can be found in [33].

System emptiness probabilities are also important functions of the queueing system of interest, especially when they will be needed for the distribution functions of the virtual and actual waiting times. Thus, some basic results about idle probabilities are presented. Let $y_{0,i}$ be the probability that the queueing system is idle at an arbitrary time and the state of the underlying Markov process $J(t)$ is i, $1 \leqslant i \leqslant m$. Let $\mathbf{y}_0$ be an

m-dimensional vector with components $y_{0,i}$. Similar to [33, section 4], it can be proved that $\mathbf{y}_0$ satisfies

$$\mathbf{y}_0 Q = 0. \tag{4.5}$$

Since Q is an irreducible infinitesimal generator, the solution of $\mathbf{y}_0$ is unique up to a constant. The constant shall be determined as $1 - \rho$ in theorem 4.2. Intuitively, equation (4.5) holds since Q is the infinitesimal generator of the underlying Markov process during the idle periods and $\mathbf{y}_0$ is the stationary distribution (up to a constant) during the idle periods.

Equation (4.4) for matrix Q and equation (4.5) for vector $\mathbf{y}_0$ are especially important in computing the distributions of virtual and actual waiting times.

The virtual waiting time is the total workload in the system at an arbitrary time. Apparently, the virtual waiting time depends only on the service times of batches and does not require information about how individual customers in a batch are served. Thus, a relatively simple analysis can be conducted on the virtual waiting time.

Let $v_i^*(s)$ be the Laplace–Stieltjes transform of the workload in the queueing system at an arbitrary time when the state of the underlying Markov process $J(t)$ is i. Let $\mathbf{v}^*(s) = (v_1^*(s), v_2^*(s), \dots, v_m^*(s))$.

Theorem 4.1. When the queueing system of interest can reach its steady state, it has, for $s > 0$,

$$\mathbf{v}^*(s) = s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f_{\mathbf{h}}^*(s)\right)^{-1}, \tag{4.6}$$

where $f_{\mathbf{h}}^*(s)$ is defined in equation (4.1).

Proof. Equation (4.6) can be proved using the result obtained in [33, section 5]. Details are omitted. □

Equation (4.6) gives a complete answer to the Laplace–Stieltjes transform of the distribution of the virtual waiting time except that the vector $\mathbf{y}_0$ needs to be determined. It is known that $\mathbf{y}_0$ is the left invariant vector (up to a constant) of matrix Q. Therefore, $\mathbf{y}_0$ is determined completely when $\mathbf{y}_0\mathbf{e}$ is known.

Theorem 4.2. When the queueing system of interest can reach its steady state, $\mathbf{y}_0\mathbf{e} = 1 - \rho$.

Proof. This proof is similar to the one given in [8, theorem 4]. Details are omitted. □

Note. The results obtained in this section for the busy period and virtual waiting time hold not only for the FCFS case, but also for many work conserving service disciplines.

5. The actual waiting times: results and an informal proof

In this section, the FCFS service discipline is imposed. The actual waiting times of various types of arrivals and customers are investigated. To make it easy for readers to follow, the results about the actual waiting time of batches (or the waiting time of the first customer to be served in a batch) are given first.

Let $w^*_{b,\mathbf{L},j}(s)$ be the Laplace–Stieltjes transform of the actual waiting time $W_{b,\mathbf{L}}$ of the first customer to be served in an arbitrary type $\mathbf{L}$ batch when the state of the underlying Markov process $J(t)$ right after the arrival of the batch $\mathbf{L}$ is j, for $1 \leqslant j \leqslant m$ and $\mathbf{L} \in \aleph$. Let $\mathbf{w}^*_{b,\mathbf{L}}(s) = (w^*_{b,\mathbf{L},1}(s), w^*_{b,\mathbf{L},2}(s), \ldots, w^*_{b,\mathbf{L},m}(s))$. Let $w^*_{b,j}(s)$ be the Laplace–Stieltjes transform of the actual waiting time of the first customer to be served in an arbitrary batch when the state of the underlying Markov process $J(t)$ right after the arrival of the batch is j. Let $\mathbf{w}^*_b(s) = (w^*_{b,1}(s), w^*_{b,2}(s), \ldots, w^*_{b,m}(s))$.

Theorem 5.1. When the queueing system of interest can reach its steady state, the actual waiting time of an arbitrary type $\mathbf{L} \in \aleph$ batch is given by

$$\mathbf{w}^*_{b,\mathbf{L}}(s) = s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\right)^{-1} D_{\mathbf{L}}\frac{1}{\lambda_{\mathbf{L}}}. \tag{5.1}$$

The actual waiting of an arbitrary batch (regardless of the type of the batch) is given by

$$\mathbf{w}^*_b(s) = s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\right)^{-1}\left(\sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}}\right)\frac{1}{\lambda_b}. \tag{5.2}$$

Proof. By theorem 4.1, the workload at an arbitrary time is given by $v^*_i(s)$ when the underlying Markov process $J(t)$ is in state i. According to results in section 2, the probability that a type $\mathbf{L}$ batch arrives when the underlying Markov process $J(t)$ is in state j right after the arrival, given that the underlying Markov process was in state i just before the arrival, is $(D_{\mathbf{L}})_{i,j}/\lambda_{\mathbf{L}}$. Then it is easy to see that

$$w^*_{b,\mathbf{L},j}(s) = \sum_{i=1}^{m} \frac{v^*_i(s)(D_{\mathbf{L}})_{i,j}}{\lambda_{\mathbf{L}}}, \quad 1 \leqslant j \leqslant m. \tag{5.3}$$

This leads to equation (5.1). Equation (5.2) can be proved similarly. This completes the proof. □

For the actual waiting time of individual customers, it is useful to see that the waiting time of a customer is the waiting time of its batch plus the service times of the customers in the same batch but sequenced ahead of it. For instance, consider the waiting time of a type k customer who arrives in a type $\mathbf{h}$ batch and is the nth customer of the batch. Denote by $W_{\mathbf{h},k,n}$ its waiting time (assume that $1 \leqslant n \leqslant |\mathbf{h}|$). Then

$$W_{\mathbf{h},k,n} = W_{b,\mathbf{h}} + v_{1,h_1} + v_{2,h_2} + \cdots + v_{n-1,h_{n-1}}, \quad \text{if } h_n = k, \tag{5.4}$$

where v_{i,h_i} is the service time of the ith customer in the batch $\mathbf{h}$, $1 \leqslant i \leqslant |\mathbf{h}|$. This equation, combining with theorem 5.1, leads to the following elementary result.

Theorem 5.2. Let $w^*_{\mathbf{L},k,n,j}(s)$ be the Laplace–Stieltjes transform of the actual waiting time of an arbitrary type k customer who comes from the nth position of a type $\mathbf{L}$ batch when the state of the underlying Markov process is j right after the customer (or its batch) arrived. Let $\mathbf{w}^*_{\mathbf{L},k,n}(s) = (w^*_{\mathbf{L},k,n,1}(s), w^*_{\mathbf{L},k,n,2}(s), \ldots, w^*_{\mathbf{L},k,n,m}(s))$. When the queueing system of interest can reach its steady state, we have

$$\mathbf{w}^*_{\mathbf{L},k,n}(s) = s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\right)^{-1} D_{\mathbf{L}}\left(\prod_{i=1}^{n-1} f^*_{h_i}(s)\right) I\{h_n = k\}\frac{1}{\lambda_{\mathbf{L},k,n}}. \tag{5.5}$$

Proof. Note that $\lambda_{\mathbf{L},k,n} = \lambda_{\mathbf{L}}$ when $h_n = k$. This completes the proof. □

Theorem 5.2 gives the actual waiting time distribution of a particular type of customer from a particular position of a particular batch. This result is useful in finding the waiting time distributions for many special groups of customers. For instance, let $\mathbf{w}^*_{\mathbf{L},k}(s)$ be the Laplace–Stieltjes transform of the actual waiting time of an arbitrary type k customer which comes from a type $\mathbf{L} \in \aleph$ batch; $\mathbf{w}^*_k(s)$ the Laplace–Stieltjes transform of the actual waiting time of an arbitrary type k customer; and $\mathbf{w}^*(s)$ the Laplace–Stieltjes transform of the actual waiting time of an arbitrary customer.

Theorem 5.3. When the queueing system of interest can reach its steady state, we have

$$\begin{aligned}
\mathbf{w}^*_{\mathbf{L},k}(s) &= s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\right)^{-1}\left[\sum_{n=1}^{|\mathbf{L}|} D_{\mathbf{L}}\left(\prod_{i=1}^{n-1} f^*_{h_i}(s)\right) I\{h_n = k\}\right]\frac{1}{\lambda_{\mathbf{L},k}};\\
\mathbf{w}^*_k(s) &= s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\right)^{-1}\left[\sum_{\mathbf{h}\in\aleph}\sum_{n=1}^{|\mathbf{h}|} D_{\mathbf{h}}\left(\prod_{i=1}^{n-1} f^*_{h_i}(s)\right) I\{h_n = k\}\right]\frac{1}{\lambda_k};\\
\mathbf{w}^*(s) &= s\mathbf{y}_0\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\right)^{-1}\left[\sum_{\mathbf{h}\in\aleph}\sum_{n=1}^{|\mathbf{h}|} D_{\mathbf{h}} \prod_{i=1}^{n-1} f^*_{h_i}(s)\right]\frac{1}{\lambda}.
\end{aligned} \tag{5.6}$$

Proof. The Laplace–Stieltjes transforms in equation (5.6) can be obtained by conditioning on the type of the arbitrary customer under consideration. According to property 2.1, the probability that a type k customer comes from the nth position of a type $\mathbf{L}$ batch, given that it is a type k customer is $\lambda_{\mathbf{L},k,n}/\lambda_k$. Then

$$\mathbf{w}^*_{\mathbf{L},k}(s) = \sum_{n=1}^{|\mathbf{h}|} \frac{\lambda_{\mathbf{L},k,n}}{\lambda_{\mathbf{L}}} \mathbf{w}^*_{\mathbf{L},k,n}(s), \tag{5.7}$$

which leads to the first expression in equation (5.5). Similarly, the following relationships hold:

$$\begin{aligned} \mathbf{w}_k^*(s) &= \sum_{\mathbf{h}\in\aleph} \frac{\lambda_{\mathbf{h},k}}{\lambda_k} \mathbf{w}_{\mathbf{h},k}^*(s) = \sum_{n=1}^{\infty} \sum_{\mathbf{h}\in\aleph} \frac{\lambda_{\mathbf{h},k,n}}{\lambda_k} \mathbf{w}_{\mathbf{h},k,n}^*(s); \\ \mathbf{w}^*(s) &= \sum_{k=1}^{K} \frac{\lambda_k}{\lambda} \mathbf{w}_k^*(s). \end{aligned} \tag{5.8}$$

These two relationships lead to the second and third expressions in equation (5.6). This completes the proof. □

6. A formal proof of the actual waiting time distributions

As was shown, the key in proving all the formulas obtained in section 5 is to prove theorem 5.1. The objective of this section is to provide a rigorous proof for theorem 5.1.

Let $x_{0,j}$ be the probability that the queueing system is empty after the departure of an arbitrary customer (regardless of its type) when the state of the underlying Markov process $J(t)$ right after the departure is j, $1 \leqslant j \leqslant m$. Denote by $\mathbf{x}_0$ the m-dimensional vector with components $x_{0,j}$.

First, we show that, when the queueing system of interest can reach its steady state, the actual waiting time of an arbitrary type $\mathbf{L}(\in \aleph)$ batch is given by

$$\mathbf{w}_{b,\mathbf{L}}^*(s) = -s\lambda\mathbf{x}_0 D_0^{-1}\left(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f_{\mathbf{h}}^*(s)\right)^{-1} D_{\mathbf{L}} \frac{1}{\lambda_{\mathbf{L}}}. \tag{6.1}$$

We shall then establish a relationship between the vectors $\mathbf{x}_0$ and $\mathbf{y}_0$ to complete the proof of theorem 5.1.

Suppose that the queueing system of interest is in steady state. Let $W_{\mathbf{L}}$ be the waiting time of an arbitrary type $\mathbf{L}$ batch. Between this and the next type $\mathbf{L}$ batch, there might be some other batches, denoted by $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N$, i.e., there are N other batches who arrived between the two consecutive type $\mathbf{L}$ batches. Note that N is a random variable. Then the waiting time of the next type $\mathbf{L}$ batch is given by

$$W_{\mathbf{L},\mathbf{h}_1,\ldots,\mathbf{h}_N,\mathbf{L}} = \left(\left(\ldots\left((W_{\mathbf{L}} + v_{\mathbf{L}} - U_{\mathbf{h}_1})^+ + v_{\mathbf{h}_1} - U_{\mathbf{h}_2}\right)^+ + \cdots\right)^+ + v_{\mathbf{h}_N} - U_{\mathbf{L}}\right)^+, \tag{6.2}$$

where $v_{\mathbf{h}} = \sum_{i=1}^{|\mathbf{h}|} v_{i,h_i}$, i.e., the total service time of all customers in batch $\mathbf{h}$, and $U_{\mathbf{h}}$ is the interarrival time when the next arrival is of type $\mathbf{h}$; and $x^+ = \max\{0, x\}$. (Notice that $v_{\mathbf{h}_i}(U_{\mathbf{h}_i})$ and $v_{\mathbf{h}_j}(U_{\mathbf{h}_j})$ are different for $i \neq j$ even when $\mathbf{h}_i = \mathbf{h}_j$.) To derive the stationary distribution function of $W_{\mathbf{L}}$ from equation (6.2), consider the following simple case first. Let

$$W_{\mathbf{L},\mathbf{h}_1} = (W_{\mathbf{L}} + v_{\mathbf{L}} - U_{\mathbf{h}_1})^+. \tag{6.3}$$

Denote by $W_{\mathbf{L},j}(x)$ the distribution of the waiting time of a type $\mathbf{L}$ batch when the state of the underlying Markov process $J(t)$ is j right after the type $\mathbf{L}$ batch arrived.

Denote by $W_{\mathbf{L},\mathbf{h}_1,j}(x)$ the distribution of the waiting time of a type $\mathbf{h}_1$ batch (following a type $\mathbf{L}$ batch) when the state of the underline Markov process $J(t)$ is j right after the $\mathbf{h}_1$ batch arrived. Then conditioning on the interarrival time $U_{\mathbf{h}_1}$ and the service time of the batch $\mathbf{L}$, yields

$$W_{\mathbf{L},\mathbf{h}_1,j}(x) = \int_0^\infty \int_0^{x+t} \sum_{i=1}^{m} W_{\mathbf{L},i}(x-u+t)\big(\exp\{D_0 t\}D_{\mathbf{h}_1}\big)_{i,j} F_{\mathbf{L}}(\mathrm{d}u)\,\mathrm{d}t. \tag{6.4}$$

Equation (6.4) can be written into vector form as follows:

$$\begin{aligned}\mathbf{W}_{\mathbf{L},\mathbf{h}_1}(x) &= \int_0^\infty \int_0^{x+t} \mathbf{W}_{\mathbf{L}}(x-u+t)\exp\{D_0 t\}D_{\mathbf{h}_1} F_{\mathbf{L}}(\mathrm{d}u)\,\mathrm{d}t \\ &\equiv \int_{-\infty}^{x} \mathbf{W}_{\mathbf{L}}(x-u) H_{\mathbf{L},\mathbf{h}_1}(\mathrm{d}u),\end{aligned} \tag{6.5}$$

where

$$H_{\mathbf{L},\mathbf{h}_1}(u) = \int_0^\infty \exp\{D_0 t\}D_{\mathbf{h}_1} F_{\mathbf{L}}(u+t)\,\mathrm{d}t, \quad -\infty < u < +\infty. \tag{6.6}$$

Then the Laplace–Stieltjes transform of $H_{\mathbf{L},\mathbf{h}_1}(u)$ is

$$\begin{aligned}H^*_{\mathbf{L},\mathbf{h}_1}(s) &= \int_{-\infty}^{\infty} \mathrm{e}^{-su} H^*_{\mathbf{L},\mathbf{h}_1}(\mathrm{d}u) = \int_{-\infty}^{\infty}\int_0^\infty \exp\big\{(s\mathbf{I}+D_0)t\big\}D_{\mathbf{h}_1}\,\mathrm{d}t\,\mathrm{e}^{-su}F_{\mathbf{L}}(\mathrm{d}u) \\ &= -f^*_{\mathbf{L}}(s)(s\mathbf{I}+D_0)^{-1}D_{\mathbf{h}_1}.\end{aligned} \tag{6.7}$$

Extend the function $\mathbf{W}_{\mathbf{L},\mathbf{h}_1}(x)$ to $x<0$ to obtain function $\widehat{\mathbf{W}}_{\mathbf{L},\mathbf{h}_1}(x)$. Then, for $x<0$,

$$\begin{aligned}\widehat{\mathbf{W}}_{\mathbf{L},\mathbf{h}_1}(x) &= \int_{-x}^\infty \int_0^{x+t} \mathbf{W}_{\mathbf{L}}(x-u+t)\exp\{D_0 t\}D_{\mathbf{h}_1} F_{\mathbf{L}}(\mathrm{d}u)\,\mathrm{d}t \\ &= \int_0^\infty \int_0^{t} \mathbf{W}_{\mathbf{L}}(t-u)\exp\big\{D_0(t-x)\big\}F_{\mathbf{L}}(\mathrm{d}u)\,\mathrm{d}t D_{\mathbf{h}_1} \qquad (t := t+x) \\ &= \int_{0-}^\infty \int_0^{\infty} \mathbf{W}_{\mathbf{L}}(y)\exp\big\{D_0(y+u-x)\big\}F_{\mathbf{L}}(\mathrm{d}u)\,\mathrm{d}y D_{\mathbf{h}_1} \quad (y := t-u) \\ &= \int_{0-}^\infty \mathbf{W}_{\mathbf{L}}(y)\exp\{D_0 y\}\,\mathrm{d}y \int_0^\infty \exp\{D_0 u\}F_{\mathbf{L}}(\mathrm{d}u)\exp\{-D_0 x\}D_{\mathbf{h}_1} \\ &= \int_{0-}^\infty \mathbf{W}_{\mathbf{L}}(\mathrm{d}y)\exp\{D_0 y\} \int_0^\infty \exp\{D_0 u\}F_{\mathbf{L}}(\mathrm{d}u)\big(-D_0^{-1}\big)\exp\{-D_0 x\}D_{\mathbf{h}_1} \\ &\equiv \mathbf{C}(\mathbf{L})\big(-D_0^{-1}\big)\exp\{-D_0 x\}D_{\mathbf{h}_1}.\end{aligned} \tag{6.8}$$

For the extended function $\widehat{\mathbf{W}}_{\mathbf{L},\mathbf{h}_1}(x)$, $-\infty < x < \infty$, its Laplace–Stieltjes transform is given as

$$\widehat{\mathbf{w}}^*_{\mathbf{L},\mathbf{h}_1}(s) = \int_{-\infty}^\infty \mathrm{e}^{-sx} \int_{-\infty}^{x} \mathbf{W}_{\mathbf{L}}(\mathrm{d}x-u)H_{\mathbf{L},\mathbf{h}_1}(\mathrm{d}u) = \mathbf{w}^*_{\mathbf{L}}(s)H^*_{\mathbf{L},\mathbf{h}_1}(s). \tag{6.9}$$

On the other hand, the Laplace–Stieltjes transform of the extended function can be found by

$$\begin{aligned}
&\left[\int_{-\infty}^{0-} + \int_{0-}^{0} + \int_{0}^{\infty}\right] \mathrm{e}^{-sx}\widehat{\mathbf{W}}_{\mathbf{L},\mathbf{h}_1}(\mathrm{d}x) \\
&\quad = \mathbf{C}(\mathbf{L})\bigl(-D_0^{-1}\bigr)\int_{-\infty}^{0-} \mathrm{e}^{-sx}\exp\{-D_0x\}\,\mathrm{d}x(-D_0)D_{\mathbf{h}_1} - \mathbf{C}(\mathbf{L})\bigl(-D_0^{-1}\bigr)D_{\mathbf{h}_1} + \mathbf{w}^*_{\mathbf{L},\mathbf{h}_1}(s) \\
&\quad = -s\mathbf{C}(\mathbf{L})\bigl(-D_0^{-1}\bigr)(s\mathbf{I} + D_0)^{-1}D_{\mathbf{h}_1} + \mathbf{w}^*_{\mathbf{L},\mathbf{h}_1}(s).
\end{aligned} \tag{6.10}$$

Combining equations (6.9) and (6.10) yields

$$\mathbf{w}^*_{\mathbf{L},\mathbf{h}_1}(s) = -s\mathbf{C}(\mathbf{L})D_0^{-1}(s\mathbf{I} + D_0)^{-1}D_{\mathbf{h}_1} + \mathbf{w}^*_{\mathbf{L}}(s)H^*_{\mathbf{L},\mathbf{h}_1}(s). \tag{6.11}$$

Consider the nth batch ($n < N$) after the first type $\mathbf{L}$ batch and before the next type $\mathbf{L}$ batch. The waiting time the type $\mathbf{h}_n$ batch is given as follows:

$$\begin{aligned}
W_{\mathbf{L},\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n} &= \bigl(\bigl(\ldots\bigl((W_{\mathbf{L}} + v_{\mathbf{L}} - U_{\mathbf{h}_1})^+ + v_{\mathbf{h}_1} - U_{\mathbf{h}_2}\bigr)^+ + \cdots\bigr)^+ + v_{\mathbf{h}_{n-1}} - U_{\mathbf{h}_n}\bigr)^+ \\
&= (W_{\mathbf{L},\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_{n-1}} + v_{\mathbf{h}_{n-1}} - U_{\mathbf{h}_n})^+.
\end{aligned} \tag{6.12}$$

Inductively, the Laplace–Stieltjes transform of the distribution function of $W_{\mathbf{L},\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n}$ is given as

$$\begin{aligned}
&\mathbf{w}^*_{\mathbf{L},\mathbf{h}_1,\ldots,\mathbf{h}_n}(s) \\
&\quad = -s\mathbf{C}(\mathbf{L}\mathbf{h}_1\ldots\mathbf{h}_{n-1})D_0^{-1}(s\mathbf{I} + D_0)^{-1}D_{\mathbf{h}_n} + \mathbf{w}^*_{\mathbf{L},\mathbf{h}_1,\ldots,\mathbf{h}_{n-1}}(s)H^*_{\mathbf{h}_{n-1},\mathbf{h}_n}(s) \\
&\quad = \mathbf{w}^*_{\mathbf{L}}(s)H^*_{\mathbf{L},\mathbf{h}_1}(s)\cdots H^*_{\mathbf{h}_{n-1},\mathbf{h}_n}(s) - s\sum_{t=1}^{n}\mathbf{C}(\mathbf{L}\mathbf{h}_1\ldots\mathbf{h}_{t-1})D_0^{-1} \\
&\quad\quad \times (s\mathbf{I} + D_0)^{-1}D_{\mathbf{h}_t}H^*_{\mathbf{h}_t,\mathbf{h}_{t+1}}(s)\cdots H^*_{\mathbf{h}_{n-1},\mathbf{h}_n}(s),
\end{aligned} \tag{6.13}$$

where

$$\begin{aligned}
\mathbf{C}(\mathbf{L}) &= \int_0^{\infty}\exp\{D_0u\}F_{\mathbf{L}}(\mathrm{d}u); \\
\mathbf{C}(\mathbf{L}\mathbf{h}_1\ldots\mathbf{h}_n) &= \int_0^{\infty}\mathbf{W}_{\mathbf{L},\mathbf{h}_1,\ldots,\mathbf{h}_{n-1}}(\mathrm{d}y)\exp\{D_0y\}\int_0^{\infty}\exp\{D_0u\}F_{\mathbf{h}_n}(\mathrm{d}u), \quad n \geqslant 1.
\end{aligned} \tag{6.14}$$

For the actual waiting time of batch $\mathbf{L}$, conditioning on N, the number of batches between two consecutive $\mathbf{L}$ batches, using equations (6.11) and (6.14), yields

$$\begin{aligned}
\mathbf{w}^*_{\mathbf{L}}(s) = \mathbf{w}^*_{\mathbf{L}}(s)&\left[\sum_{n=0}^{\infty}\ \sum_{\{\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n:\ \mathbf{h}_i\neq\mathbf{L},\ 1\leqslant i\leqslant n\}} H^*_{\mathbf{L},\mathbf{h}_1}(s)\cdots H^*_{\mathbf{h}_{n-1},\mathbf{h}_n}(s)H^*_{\mathbf{h}_n,\mathbf{L}}(s)\right] \\
&- s\sum_{n=0}^{\infty}\ \sum_{\{\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_n:\ \mathbf{h}_i\neq\mathbf{L},\ 1\leqslant i\leqslant n\}}\ \sum_{t=1}^{n}\mathbf{C}(\mathbf{L}\mathbf{h}_1\ldots\mathbf{h}_{t-1})D_0^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\times (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_t} H^*_{\mathbf{h}_t,\mathbf{h}_{t+1}}(s) \cdots H^*_{\mathbf{h}_n,\mathbf{L}}(s)\\
\equiv\ & \mathbf{w}^*_{\mathbf{L}}(s) \cdot \text{Part I} - s \cdot \text{Part II}.
\end{aligned} \tag{6.15}
$$

Part I and Part II of equation (6.15) are evaluated next. Part I is evaluated as:

$$
\begin{aligned}
&\sum_{n=0}^{\infty} \sum_{\{\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_n:\ \mathbf{h}_i \neq \mathbf{L},\ 1 \leqslant i \leqslant n\}} (-1)^{n+1} (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_1} f^*_{\mathbf{L}}(s) (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_2} f^*_{\mathbf{h}_1}(s) \cdots\\
&\quad \times (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_n} f^*_{\mathbf{h}_{n-1}}(s) (s\mathbf{I} + D_0)^{-1} D_{\mathbf{L}} f^*_{\mathbf{h}_n}(s)\\
&\quad = \sum_{n=0}^{\infty} \sum_{\{\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_n:\ \mathbf{h}_i \neq \mathbf{L},\ 1 \leqslant i \leqslant n\}} (-1)^{n+1} (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_1} f^*_{\mathbf{h}_1}(s) (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_2} f^*_{\mathbf{h}_2}(s) \times\\
&\quad \cdots (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_n} f^*_{\mathbf{h}_n}(s) (s\mathbf{I} + D_0)^{-1} D_{\mathbf{L}} f^*_{\mathbf{L}}(s)\\
&\quad = -\sum_{n=0}^{\infty} (-1)^n \Bigg(\sum_{\mathbf{h}:\ \mathbf{h} \neq \mathbf{L},\ \mathbf{h} \in \aleph} (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}} f^*_{\mathbf{h}}(s) \Bigg)^n (s\mathbf{I} + D_0)^{-1} D_{\mathbf{L}} f^*_{\mathbf{L}}(s)\\
&\quad = -\Bigg(s\mathbf{I} + D_0 + \sum_{\mathbf{h}:\ \mathbf{h} \neq \mathbf{L},\ \mathbf{h} \in \aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s) \Bigg)^{-1} D_{\mathbf{L}} f^*_{\mathbf{L}}(s).
\end{aligned} \tag{6.16}
$$

Part II is evaluated as: (Notice that $\mathbf{L} = \mathbf{h}_0$)

$$
\begin{aligned}
&\sum_{t=1}^{\infty} \sum_{n=t-1}^{\infty} \sum_{\{\mathbf{h}_1,\mathbf{h}_2,\dots,\mathbf{h}_n:\ \mathbf{h}_i \neq \mathbf{L},\ 1 \leqslant i \leqslant n\}} \mathbf{C}(\mathbf{L}\mathbf{h}_1 \dots \mathbf{h}_{t-1}) D_0^{-1} (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_t} H^*_{\mathbf{h}_t \mathbf{h}_{t+1}}(s) \cdots\\
&\qquad\qquad \times H^*_{\mathbf{h}_n \mathbf{L}}(s)\\
&\quad = \sum_{t=1}^{\infty} \sum_{n=t-1}^{\infty} \sum_{\{\mathbf{h}_1,\dots,\mathbf{h}_{t-1}:\ \mathbf{h}_i \neq \mathbf{L},\ 1 \leqslant i \leqslant t-1\}} \mathbf{C}(\mathbf{L}\mathbf{h}_1 \dots \mathbf{h}_{t-1}) D_0^{-1}\\
&\quad \times \sum_{\{\mathbf{h}_t,\dots,\mathbf{h}_n:\ \mathbf{h}_i \neq \mathbf{L},\ t \leqslant i \leqslant n\}} (s\mathbf{I} + D_0)^{-1} D_{\mathbf{h}_t} H^*_{\mathbf{h}_t \mathbf{h}_{t+1}}(s) \cdots H^*_{\mathbf{h}_n \mathbf{L}}(s)\\
&\quad = \Bigg(\sum_{t=0}^{\infty} \sum_{\{\mathbf{h}_1,\dots,\mathbf{h}_t:\ \mathbf{h}_i \neq \mathbf{L},\ 1 \leqslant i \leqslant t\}} \mathbf{C}(\mathbf{L}\mathbf{h}_1 \dots\ \mathbf{h}_t) \Bigg) D_0^{-1}\\
&\quad \times \Bigg(\mathbf{I} + \sum_{\mathbf{h}:\ \mathbf{h} \neq \mathbf{L}} (sI + D_0)^{-1} D_{\mathbf{h}} f^*_{\mathbf{h}}(s) \Bigg)^{-1} (s\mathbf{I} + D_0)^{-1} D_{\mathbf{L}}\\
&\quad \equiv \mathbf{C}_{\mathbf{L}} D_0^{-1} \Bigg(s\mathbf{I} + D_0 + \sum_{\mathbf{h}:\ \mathbf{h} \neq \mathbf{L},\ \mathbf{h} \in \aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s) \Bigg)^{-1} D_{\mathbf{L}}.
\end{aligned} \tag{6.17}
$$

Equations (6.16) and (6.17) lead to

$$\begin{aligned}\mathbf{w}^*_{\mathbf{L}}(s) &= -\mathbf{w}^*_{\mathbf{L}}(s)\Bigg(s\mathbf{I} + D_0 + \sum_{\mathbf{h}:\ \mathbf{h}\neq\mathbf{L},\ \mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\Bigg)^{-1} D_{\mathbf{L}} f^*_{\mathbf{L}}(s) \\ &\quad - s\mathbf{C}_{\mathbf{L}} D_0^{-1}\Bigg(s\mathbf{I} + D_0 + \sum_{\mathbf{h}:\ \mathbf{h}\neq\mathbf{L},\ \mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\Bigg)^{-1} D_{\mathbf{L}}. \qquad (6.18)\end{aligned}$$

This leads to

$$\mathbf{w}^*_{\mathbf{L}}(s) = -s\mathbf{C}_{\mathbf{L}} D_0^{-1}\Bigg(s\mathbf{I} + D_0 + \sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} f^*_{\mathbf{h}}(s)\Bigg)^{-1} D_{\mathbf{L}}. \qquad (6.19)$$

The next step is to prove that $\mathbf{C}_{\mathbf{L}} = \mathbf{x}_0\lambda/\lambda_{\mathbf{L}}$. Intuitively, the vector $\mathbf{C}(\mathbf{L}\mathbf{h}_1 \ldots \mathbf{h}_n)$ is the probability that the system becomes idle after the completion of the services of all customers in the nth batch (of the type $\mathbf{h}_n$) after the first $\mathbf{L}$, but before the next type $\mathbf{L}$ batch. Then the vector $\mathbf{C}_{\mathbf{L}}$ is the probability that the queueing system becomes idle at the completion epochs of batches between two consecutive type $\mathbf{L}$ batches. Since $\mathbf{x}_0$ is the probability that the system becomes idle at the completion epoch of an arbitrary batch, it has

$$\mathbf{x}_0 = \frac{\lambda_{\mathbf{L}}}{\lambda}\mathbf{C}_{\mathbf{L}} \quad\Rightarrow\quad \mathbf{C}_{\mathbf{L}} = \frac{\lambda}{\lambda_{\mathbf{L}}}\mathbf{x}_0. \qquad (6.20)$$

Equations (6.19) and (6.20) lead to equation (6.1). A rigorous proof of equation (6.20) is given as follows.

Suppose that the queueing system is in steady state. There is a type $\mathbf{L}$ batch that arrived at time zero (batch zero). Let $\Lambda_{n,j}$ be the event that the nth customer leaves an empty system and the underlying Markov process $J(t)$ is in state j right after the departure.

Let $\eta_{\mathbf{L}}(n)$ be the sequential number of the first customer in the nth type $\mathbf{L}$ batch, i.e., the first customer of the nth type $\mathbf{L}$ batch is the lth customer ($l = \eta_{\mathbf{L}}(n)$) to arrive to the queueing system. It is easy to see that $\eta_{\mathbf{L}}(n+1) \geqslant \eta_{\mathbf{L}}(n) + |\mathbf{L}|$, $n \geqslant 0$. Then we have $\xi_{\mathbf{L}}(n) = \max\{t\colon\ \eta_{\mathbf{L}}(t) \leqslant n\}$, i.e., the number of type $\mathbf{L}$ batches that arrived when the nth customer arrives, which was introduced in section 2. When a type $\mathbf{L}$ batch arrives at zero (called batch zero), we have $\eta_{\mathbf{L}}(0) = 1$. Then according to renewal theory (see [5]), when $\mathbf{x}_0$ exists,

$$\begin{aligned}\mathbf{x}_0 &= \lim_{N\to\infty} \frac{\sum_{n=0}^{N}(I\{\Lambda_{n,1}\}, \ldots, I\{\Lambda_{n,m}\})}{N} \\ &= \lim_{N\to\infty} \frac{\xi_{\mathbf{L}}(N)}{N} \frac{\sum_{l=0}^{\xi_{\mathbf{L}}(N)-1}\sum_{n=\eta_{\mathbf{L}}(l)}^{\eta_{\mathbf{L}}(l+1)-1}(I\{\Lambda_{n,1}\}, \ldots, I\{\Lambda_{n,m}\})}{\xi_{\mathbf{L}}(N)}. \qquad (6.21)\end{aligned}$$

According to equation (2.10), $\xi_{\mathbf{L}}(N)/N$ converges to $\lambda_{\mathbf{L}}/\lambda$. The other part converges to $\mathbf{C}(\mathbf{L})$, which is proved as follows. Consider the embedded stochastic process at the first departure epochs after type **L** batches arrive. It is clear that the embedded stochastic process is a regenerative process (see [28]). Thus

$$\begin{aligned}\lim_{N\to\infty} &\frac{\sum_{l=0}^{\xi_{\mathbf{L}}(N)-1}\sum_{n=\eta_{\mathbf{L}}(l)}^{\eta_{\mathbf{L}}(l+1)-1}(I\{\Lambda_{n,1}\},\ldots,I\{\Lambda_{n,m}\})}{\xi_{\mathbf{L}}(N)}\\ &= \mathbf{E}\sum_{n=1}^{\eta_{\mathbf{L}}(1)-1}\big(I\{\Lambda_{n,1}\},\ldots,I\{\Lambda_{n,m}\}\big).\end{aligned}\tag{6.22}$$

Let $\zeta_{\mathbf{L}}(1)$ be the number of batches that arrived between two consecutive type **L** batches ($\zeta_{\mathbf{L}}(1)\geqslant 0$). Let $\widehat{\Lambda}_{n,j}$ be the event that the nth batch (super customer) leaves an empty system behind and the underlying Markov process $J(t)$ is in state j right after the departure. Then

$$\mathbf{E}\sum_{n=1}^{\eta_{\mathbf{L}}(1)-1}\big(I\{\Lambda_{n,1}\},\ldots,I\{\Lambda_{n,m}\}\big) = \mathbf{E}\sum_{n=0}^{\varsigma_{\mathbf{L}}(1)}\big(I\big\{\widehat{\Lambda}_{n,1}\big\},\ldots,I\big\{\widehat{\Lambda}_{n,m}\big\}\big),\tag{6.23}$$

since only the last customer in a batch is in the position to leave an empty system behind. Then

$$\begin{aligned}\mathbf{x}_0 &= \frac{\lambda_{\mathbf{L}}}{\lambda}\mathbf{E}\sum_{n=0}^{\varsigma(1)}\big(I\big\{\widehat{\Lambda}_{n,1}\big\},\ldots,I\big\{\widehat{\Lambda}_{n,m}\big\}\big)\\ &= \frac{\lambda_{\mathbf{L}}}{\lambda}\mathbf{E}\sum_{n=0}^{\infty}\big(I\big\{\widehat{\Lambda}_{n,1}\cap\big\{\varsigma_{\mathbf{L}}(1)>n\big\}\big\},\ldots,I\big\{\widehat{\Lambda}_{n,m}\cap\big\{\varsigma_{\mathbf{L}}(1)>n\big\}\big\}\big)\\ &= \frac{\lambda_{\mathbf{L}}}{\lambda}\sum_{n=0}^{\infty}\sum_{\{\mathbf{h}_1,\ldots,\mathbf{h}_n:\ \mathbf{h}_i\neq\mathbf{L},\ 1\leqslant i\leqslant n\}}\mathbf{E}\big[\big(I\big\{\widehat{\Lambda}_{n,1}\cap\big\{\varsigma_{\mathbf{L}}(1)>n\big\}\big\},\ldots,\\ &\qquad\qquad I\big\{\widehat{\Lambda}_{n,m}\cap\big\{\varsigma_{\mathbf{L}}(1)>n\big\}\big\}\big)\colon\ \mathbf{h}_1,\ldots,\mathbf{h}_n\big]\\ &= \frac{\lambda_{\mathbf{L}}}{\lambda}\sum_{n=0}^{\infty}\sum_{\{\mathbf{h}_1,\ldots\ ,\mathbf{h}_\mathbf{n}:\ \mathbf{h}_i\neq\mathbf{L},\ 1\leqslant i\leqslant n\}}\mathbf{C}(\mathbf{L}\mathbf{h}_1\ldots\mathbf{h}_n) = \frac{\lambda_{\mathbf{L}}}{\lambda}\mathbf{C}_{\mathbf{L}},\end{aligned}\tag{6.24}$$

which leads to equation (6.1). (Notice that $\mathbf{L}=\mathbf{h}_0$.)

In order to complete the proof of theorem 5.1, we need to find the relationship between $\mathbf{y}_0$ and $\mathbf{x}_0$. In fact, when the queueing system of interest can reach its steady state, the following relationship between $\mathbf{y}_0$ and $\mathbf{x}_0$ holds:

$$\mathbf{y}_0 = -\lambda\mathbf{x}_0 D_0^{-1} \quad\text{and}\quad -\lambda\mathbf{x}_0 D_0^{-1}\mathbf{e} = 1-\rho.\tag{6.25}$$

To prove equation (6.25), similar to theorem 4.2, we first show that $-\lambda\mathbf{x}_0 D_0^{-1}\mathbf{e} = 1-\rho$. By equation (4.5), vector $\mathbf{y}_0$ satisfies equation $\mathbf{y}_0 Q = 0$. Consider the embedded

Markov chain at the last departures of busy periods. The one step transition matrix of this embedded Markov chain is given by

$$\begin{aligned}(-D_0)^{-1}\left[\sum_{\mathbf{h}\in\aleph} D_{\mathbf{h}} \int_0^{\infty} \Psi^*(0,x)F_{\mathbf{h}}(\mathrm{d}x)\right] &= (-D_0)^{-1}(Q-D_0) \\ &= (-D_0)^{-1}Q+\mathbf{I}. \qquad (6.26)\end{aligned}$$

It is easy to see that $\mathbf{x}_0$ must satisfy equation $\mathbf{x}_0 = \mathbf{x}_0[(-D_0)^{-1}Q+\mathbf{I}]$, which implies that $\mathbf{x}_0(-D_0)^{-1}Q = 0$. Therefore, $-\mathbf{x}_0 D_0^{-1} = c\mathbf{y}_0$. Since $\mathbf{y}_0\mathbf{e} = 1-\rho$ and $-\lambda\mathbf{x}_0 D_0^{-1}\mathbf{e} = 1-\rho$, the constant $c = 1/\lambda$, which leads to equation (6.25).

Combining equations (6.1) and (6.25), we obtain theorem 5.1. This completes the proof. □

This formal proof of theorem 5.1 is long, but it becomes natural when two key facts are observed. The first one is equation (6.2). Once the relationship in equation (6.2) is obtained, a proof following Lindley's approach (see [13]), which is the second key, becomes possible.

All other results given in section 5 can be proved in a similar way. In fact, the proof of equation (5.2) is simpler. Details of those proofs are omitted.

7. Summary and future research

The contribution of this paper is two-fold. First, it shows how to use MMAP[K] to model a variety of point processes with some special arrival patterns such as cyclic, mixed batch, and priorities. Second, it presents a detailed analysis of the actual waiting time processes of individual types of customers in queueing systems with an MMAP[K] input process. Combining the two parts, this paper shows that queueing systems with a complicated input process and multiple types of customers are still analytically tractable when the input process is modeled appropriately.

It is certainly important to develop algorithms for moments of the busy period, virtual waiting time, and actual waiting times. Since moments can be obtained by routine calculations (see [1,8,14,18,22]), details are omitted.

This paper studied the queueing system with an FCFS service discipline. It is interesting to investigate queueing systems with priorities assigned to different types of customers (see [30,31]). Though some of the methodologies used in this paper may not be working for other cases, the formalism adopted in this paper will certainly be useful in that research. Some results along this direction can be found in [33]. This paper does not consider the queue length due to some technical difficulty. For a special case where each batch has only one customer, some results about the queue length have obtained in [8]. In [9,10], the steady state queue length distribution is obtained when the service times have PH-distributions and the service discipline is last-come-first-served (LCFS). The study of the queue length of the queueing system of interest is left as an open problem for future research.

Acknowledgements

The author would like to thank a referee for his/her valuable comments and suggestions that improved the paper. This research is supported by a NSERC research grant.

References

[1] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, Queueing Systems 10 (1992) 5–88.

[2] S. Asmussen, Ladder heights and the Markov-modulated M/G/1 queue, Stochastic Process. Appl. 37 (1989) 319–334.

[3] S. Asmussen and G. Koole, Marked point processes as limits of Markovian arrival streams, J. Appl. Probab. 30 (1993) 365–372.

[4] D.T. Chen and M. Rieders, Cyclic Markov modulated Poisson processes in traffic characterization, Stochastic Models 12 (1996) 585–610.

[5] E. Cinlar, Markov renewal theory, Adv. in Appl. Probab. 1 (1969) 123–187.

[6] J.W. Cohen, *The Single Server Queue* (North-Holland, Amsterdam, 1982).

[7] F.R. Gantmacher, *The Theory of Matrices* (Chelsea, New York, 1959).

[8] Q.-M. HE, Queues with marked customers, Adv. in Appl. Probab. 28 (1996) 567–587.

[9] Q.-M. HE, Quasi-birth-and-death Markov processes with a tree structure and the MMAP[K]/PH[K]/N queue, European J. Oper. Res. 120 (2000) 641–656.

[10] Q.-M. HE and A.S. Alfa, The MMAP[K]/PH[K]/1 queue with a last-come-first-served preemptive service discipline, Queueing Systems 28 (1998) 269–291.

[11] Q.-M. HE and M.F. Neuts, Markov arrival processes with marked transitions, Stochastic Process. Appl. 74 (1998) 37–52.

[12] L. Kleinrock, *Queueing Systems*, Vol. 1: *Theory* (Wiley, New York, 1975).

[13] D.V. Lindley, The theory of queues with a single server, Proc. Cambridge Phil. Soc. 48 (1952) 227–289.

[14] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Stochastic Models 7 (1991) 1–46.

[15] D.M. Lucantoni, The MMAP/G/1 queue: A tutorial, in: *Models and Techniques for Performance Evaluation of Computer and Communications Systems*, eds. L. Donatiello and R. Nelson (Springer, Berlin, 1993) pp. 330–358.

[16] D.M. Lucantoni, G.L. Choudhury and W. Whitt, The transient BMAP/G/1 queue, Stochastic Models 10 (1994) 145–182.

[17] D.M. Lucantoni, K.S. Meier-Hellstern and M.F. Neuts, A single server queue with server vacations and a class of non-renewal arrival processes, Adv. in Appl. Probab. 22 (1990) 676–705.

[18] M.F. Neuts, A versatile Markovian point process, J. Appl. Probab. 16 (1979) 764–779.

[19] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach* (Johns Hopkins Univ. Press, Baltimore, MD, 1981).

[20] M.F. Neuts, Generalizations of the Pollaczek–Khinchin integral equation in the theory of queues, Adv. in Appl. Probab. 18 (1986) 952–990.

[21] M.F. Neuts, The fundamental period of the queue with Markov-modulated arrivals, in: *Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin* (Academic Press, New York, 1989) pp. 187–200.

[22] M.F. Neuts, *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications* (Marcel Dekker, New York, 1989).

[23] M.F. Neuts, D. Liu and N. Surya, Local poissonification of the Markovian arrival process, Stochastic Models 8 (1992) 87–129.

[24] A. Pacheco, Markov-additive processes arising in storage models for communications systems, Technical Report No. 1096, School of Operations Research and Industrial Engineering, Cornell University (1994).

[25] N.U. Prabhu, Markov renewal and Markov-additive processes – a review and some new results, in: *Proc. of KAIST Mathemtics Workshop*, Vol. 6, eds. B.D. Choi and J. Yim, Korea Advanced Institute of Science and Technology, 1991, pp. 57–94.

[26] V. Ramaswami, The $N/G/1$ queue and its detailed analysis, Adv. in Appl. Probab. 12 (1980) 222–261.

[27] V. Ramaswami, From the matrix-geometric to the matrix-exponential, Queueing Systems 6 (1990) 229–260.

[28] S.M. Ross, *Stochastic Processes* (Willey, New York, 1983).

[29] D.A. Stanford and W. Fischer, The interdeparture time distribution for each class in the $\sum_i M_i/G_i/1$ queue, Queueing Systems 4 (1989) 179–191.

[30] H. Takagi, *Queueing Analysis: A Foundation of Performance Evaluation*, Vol. 1: *Vacation and Priority Systems*, Part 1 (Elsevier Science, Amsterdam, 1990).

[31] H. Takagi, Queueing analysis of polling models: progress in 1990–1994, in: *Frontiers in Queueing*, ed. J.H. Dshalalow (1996) pp. 119–146.

[32] T. Takine, A continuous version of matrix-analytic methods with the skip-free to the left property, Stochastic Models 12 (1996) 673–682.

[33] T. Takine and T. Hasegawa, The workload in the MAP/G/1 queue with state-dependent services its application to a queue with preemptive resume priority, Stochastic Models 10 (1994) 183–204.

[34] T. Takine, B. Sengupta and R.W. Yeung, A generalization of the matrix M/G/1 paradigm for Markov chains with a tree structure, Stochastic Models 11 (1995) 411–421.

[35] R.W. Wolff, Poisson arrivals see time average, Oper. Res. 30 (1982) 223–231.

[36] R.W. Yeung and B. Sengupta, Matrix product-form solutions for Markov chains with a tree structure, Adv. in Appl. Probab. 26 (1994) 965–987.