An infinite-server queue influenced
by a semi-Markovian environment

Brian H. Fralix · Ivo J.B.F. Adan

Abstract We consider an infinite-server queue, where the arrival and service rates are both governed by a semi-Markov process that is independent of all other aspects of the queue. In particular, we derive a system of equations that are satisfied by various "parts" of the generating function of the steady-state queue-length, while assuming that all arrivals bring an amount of work to the system that is either Erlang or hyperexponentially distributed. These equations are then used to show how to derive all moments of the steady-state queue-length. We then conclude by showing how these results can be slightly extended, and used, along with a transient version of Little's law, to generate rigorous approximations of the steady-state queue-length in the case that the amount of work brought by a given arrival is of an arbitrary distribution.

1 Introduction

Queues with a randomly-varying arrival and service rate have recently received quite a bit of attention in the queueing literature. The typical setting is as follows: there is an external stochastic process, known as the environment, that takes values in some state space $\mathbb{E}$, and these values tend to control various aspects of the queueing system, such as the arrival rate of customers, and the speed at which their work is processed.

Examples of such models in the single-server setting include the recent work of Nain and Nunez-Queija [12], along with the work of Takine [16]. In [12], the authors consider an $M/M/1$ queue that is influenced by a semi-Markovian environment

B.H. Fralix (✉) · I.J.B.F. Adan
EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: fralix@eurandom.tue.nl

which takes values in the state space {0, 1}. The amount of time spent in state 0 has a distribution that possesses a rational Laplace–Stieltjes transform (LST), while the amount of time spent in state 1 is generally distributed, or in particular heavy-tailed, in the sense that it has an infinite moment-generating function. There, the main goal is to compute the z-transform of the steady-state queue-length distribution, by using techniques from complex analysis. However, things are looked at from a different perspective in the work of [16]: in this case the environment process is a continuous-time Markov chain, but multiple customer classes are considered, and both the arrival and service rates change according to the environment. Furthermore, in this model each customer brings an amount of work to the system that is generally distributed, with the distribution depending on the class of the customer. This paper focuses more on properties of the busy period, and the waiting time distribution of a customer that arrives to the system during steady state.

Work has been very recently published in the infinite-server setting as well. In [3], the authors consider an $M/M/\infty$ queue, where only the service rates are governed by an external environment process, which in this case is a continuous-time Markov chain (CTMC) that again takes values in two states. Their main results include showing that the steady-state queue-length can be written as the sum of two independent random variables, where one of these variables is Poisson, and it can also be interpreted as the number of customers present in a standard $M/M/\infty$ queue. In the infinite-server setting, we interpret the term "queue-length" as the number of customers presently in the system. Later, D'Auria wrote a series of papers [4–6] on this topic as well. The main point of [5] is to show that by using basic properties of Poisson processes on $\mathbb{R}^2$, similar stochastic decomposition results can be obtained for the steady-state queue-length, without necessarily assuming that the environment process is a CTMC. In particular, his result allows for virtually any asymptotically stationary environment process, so long as that in steady state the environment process is ergodic. He then obtains the results of [3] by performing a scheme that involves looking at areas of random sets, where these sets are constructed based on the behavior of the environment process. He then uses this same type of technique in the very recent paper [6] to analyze an $M/M/\infty$ queue with a semi-Markovian environment.

Very recently, Falin [7] has also shown that for an $M/M/\infty$ queue, where both the arrival rates and the service rates are influenced by a semi-Markovian environment, it is possible to apply what is known as a supplementary variable technique to compute the mean steady-state queue-length. This technique involves looking at a process that consists of three states: the number of customers in the system, the state of the environment, and the time until the next environment transition (this last state descriptor makes the process Markovian). This allows him to get an expression for the generating function of the number of customers present in the system, and this can be used to compute the steady-state mean queue-length. Older references focusing on the infinite-server case include O'Cinneide and Purdue [13], and Keilson and Servi [11].

The goal of this paper is to show how these results can be extended to the case when all arrivals bring Erlang, or hyperexponential amounts of work to the system. Once these results are established, it is easy to see that the same technique can be used for the case when arrivals bring to the system an amount of work that consists

of a mixture of Erlang distributions. However, it is well known (see, for instance, Asmussen [1]) that these random variables are dense in the space of nonnegative random variables under the Prohorov metric (a weak convergence metric), so we conclude by using this fact, along with a transient version of Little's law, to show that some type of analysis is still possible when the service times have a general distribution.

The main technique we will use in this paper involves looking at the queue-length process at the transition epochs of the environment process. Once we obtain the steady-state behavior of the process at these epochs, we will use the inversion formula (see [2]) to relate this to the steady-state behavior of the process at an arbitrary time. It would also be possible to apply a semi-regenerative argument (see, for example Chap. 7, Sect. 5 of [1]) but in this case we would also have to assume the presence of some sort of non-lattice condition, because these arguments lead to a limiting result. The reader should keep in mind that the Palm approach only says something about the stationary process: however, if the process has a limiting distribution, it will coincide with the marginal distribution of our stationary process.

2 Model description

Our paper will focus on an infinite-server queueing system, where the arrivals and service rates are governed by a stationary semi-Markov process $C := \{C(t); t \in \mathbb{R}\}$, on a finite state space $\mathbb{E}$: here $\mathbb{R}$ is used to denote the set of all real numbers. The transition times of C will be denoted by $\{T_n\}_{n \in \mathcal{Z}}$, where $\mathcal{Z}$ represents the set of all integers. It shall be assumed that these random variables, along with all other random elements found in this paper, exist on a probability space $(\Omega, \mathcal{F}, P)$, where Ω is an arbitrary space, $\mathcal{F}$ is an appropriate σ-field of subsets of Ω, and P is a probability measure that is defined on $\mathcal{F}$.

To completely describe how C evolves through time, it will suffice to give a pathwise description of both how long it spends in each state, and how it makes transitions from one state to the next. Hence, let $A_{i,j}(x) = P(T_{n+1} - T_n \le x, C(T_{n+1}) = j | C(T_n) = i)$. This represents the probability that, given at time T_n the process C has just made a transition to state i, it will make another transition before time $T_n + x$, and that transition will consist of a jump from state i to state j. Moreover, clearly $r_{i,j} := A_{i,j}(\infty)$ represents the probability that the environment jumps directly to state j from state i. We will let $\alpha_{i,j}(s) = \int_0^\infty e^{-sx} \, dA_{i,j}(x)$ denote the LST of $A_{i,j}$, and $\alpha_i(s) = \sum_{j \in \mathbb{E}} \alpha_{i,j}(s)$. Finally, let $\{\pi_i\}_{i \in \mathbb{E}}$ denote the stationary distribution of C at its transition times, and let $\upsilon_i = E[T_{n+1} - T_n | C(T_n) = i]$, and $\upsilon = \sum_{i \in \mathbb{E}} \upsilon_i \pi_i$. Our assumption of a finite $\mathbb{E}$ ensures that the process C is regular, in that the expected number of its transitions in any compact set is finite. Such conditions will not explicitly be needed throughout our analysis, but they are needed in order to ensure that our semi-Markov process does not actually terminate at some point. With that being said, it should be emphasized that a finite state space is not a necessary condition for regularity: indeed, the reason we need the state space of the environment to be finite is because we will eventually be interested in solving a system of equations, where the number of equations and unknowns will equal the number of states.

We also consider an infinite-server queueing system, which is influenced by the environment in the following way: while the environment is in state i, the system serves the remaining amount of work that is possessed by each customer currently in the system at a rate μ_i, and new arrivals show up in accordance with a Poisson process with rate λ_i, where the nth arrival to the system brings a random amount of work B_n, with distribution function $B(t)$. Here we follow the convention that, if $\{X_n\}$ denotes the arrival times of customers to the system, then $X_0 \leq 0 < X_1$. Throughout the paper, we will assume that $\mu_i > 0$ for all $i \in \mathbb{E}$, but this assumption is not essential for the analysis found within the next three sections; it is only used to make the paper slightly more readable. Finally, we also assume that the service requirements of all customers are independent of one another.

Thus, one could think of the arrival process in this case as being a "semi-Markov-modulated Poisson process," but of course this arrival process is not independent of how the customers are served in the system. To specify exactly what sort of dependence structure is being assumed here among all of the stochastic elements present in this model, it suffices to say that conditional on a given sample-path of the environment process, the customers arrive according to a non-homogeneous Poisson process, and if someone arrives at time t and brings an amount of work W, it will be in the system for an amount of time $T(t)$, where

$$W = \int_t^{t+T(t)} \mu_{C(s)} \, ds.$$

In both [5] and [6], heavy use is made of the following fact from point process theory.

Theorem 2.1 *Suppose N is a non-homogeneous Poisson process on $\mathbb{R}$ with points $\{X_n\}_{n \in \mathcal{Z}}$, where for each Borel set A,*

$$N(A) = \sum_{n \in \mathcal{Z}} \delta_{X_n}(A),$$

where $\delta_x(A) = 1$ if $x \in A$, and $\delta_x(A) = 0$ otherwise. If we associate to each point X_n a random variable Y_n that is independent of all other locations of N and their marks, then the new point process M with points in $\mathbb{R}^2$ that satisfies, for any Borel sets A, B,

$$M(A \times B) = \sum_{n \in \mathcal{Z}} \delta_{(X_n, Y_n)}(A \times B)$$

is a non-homogeneous Poisson process in $\mathbb{R}^2$.

When we say that M is a non-homogeneous Poisson process in $\mathbb{R}^2$, we mean that the number of points in any Borel set $A \in \mathbb{R}^2$ is Poisson-distributed, and for any disjoint collection of Borel sets $A_1, \ldots, A_n \subseteq \mathbb{R}^2$, the random variables $M(A_1), \ldots, M(A_n)$ are independent.

Keeping our model in mind, we see that if we associate with each customer arrival time X_n its service requirement B_n, and we condition on the sample-path of the

environment, the resulting point process on $\mathbb{R}^2$ is Poisson. This implies that while conditioning on the environment, the number of customers in the system at time zero is Poisson as well. Proofs of these well-known facts can be found, for instance, in Serfozo [15].

This interesting property of Poisson processes was used in [5] to obtain a stochastic decomposition of the queue-length into two independent quantities, where one of those quantities represents the stationary distribution of a standard $M/G/\infty$ queue. Indeed, a stochastic decomposition that is similar in type to that discussed in [5] can also be easily generalized to our setting.

Theorem 2.2 *Consider an $M/G/\infty$ system with a semi-Markov-modulated arrival and service rate, where $\lambda := \inf_{i \in \mathbb{E}} \lambda_i > 0$, and $\mu := \sup_{i \in \mathbb{E}} \mu_i > 0$. Then the steady-state number of customers in the system can be decomposed into the sum of two independent random variables*:

$$Q = Q_{\lambda,\mu}^{M/G/\infty} + Q_C,$$

where $Q_{\lambda,\mu}^{M/G/\infty}$ represents the steady-state number of customers in the system in a standard $M/G/\infty$ queue that processes work at a rate μ that is brought by customers arriving in a Poisson manner at rate λ, and Q_C is a randomized Poisson random variable, with a parameter that depends on the environment process C.

The proof of this result is essentially the same as the proof of the decomposition result found in [5], with only a slight modification: a semi-Markov-modulated Poisson process can be represented as an independent sum of a homogeneous Poisson process with rate λ, along with another semi-Markov-modulated Poisson process that at times has a rate of zero. Furthermore, as is done in [5], we will follow with a practical description of the result, by saying that the steady-state population can be decomposed into two classes: those that would have still been in the system, regardless of the increase of arrivals or the decrease in service speed, and the rest.

This result could very well be useful towards determining various properties of the steady-state queue-length distribution, but our approach will not involve its use. Rather, we will first derive a system of equations containing functions that can be interpreted as "parts" of the generating function of the steady-state queue-length, while further assuming that all customers bring an amount of work that is either Erlang or hyperexponentially distributed. These results are then used to show how one can generate a system of equations that consist of unknowns, which once solved for can be used to compute the mean queue-length in steady state. We will then conclude by providing a rigorous approximation of the moments of the queue-length in steady state for the case when the services are allowed to have any type of distribution, by using Little's law, along with the fact that mixtures of Erlang distributions can be used to approximate any type of distribution with nonnegative support, with respect to the weak convergence metric.

Throughout the paper, we will use results from stationary point process theory to relate the stationary distribution of our process at the transition epochs of the environment to the stationary distribution of the process at an arbitrary time. In particular,

this will involve the use of the Palm measure P_0 that is induced by the point process consisting of the transition times of the environment. We will not go into great mathematical details to precisely describe how P_0 is defined: indeed, the interested reader has many references to choose from regarding this topic, with recent ones including [2], Chap. 7, Sect. 6 of [1], and Chap. 6 of [14]. The reader merely needs to be aware of the following facts: (i) for an event A, $P_0(A)$ can be interpreted as the probability of A, given that the environment process changes state at time zero, and (ii) under the measure P_0, the joint distribution of $(Q(T_n), C(T_n))$ is the same for all $n \in \mathcal{Z}$.

3 Exponential services

We will first consider the case when each customer brings an amount of service that is exponentially distributed with rate ν, i.e., for each $t \geq 0$,

$$B(t) = 1 - e^{-\nu t}.$$

The approach we will use to compute all of the factorial moments of the queue-length will basically involve coming up with a system of equations that the functions $m_j(z)$ satisfy, where $m_j(z) = E_0[z^{Q(0)} \mathbf{1}(C(0) = j)]$.

Our first result provides an expression for the generating function of the steady-state queue-length. This will be used to compute the factorial moments $E[(Q(0))_n]$ for $n \geq 1$, where for $x \in \mathbb{R}$, $(x)_n = x(x-1) \cdots (x-n+1)$.

Theorem 3.1 *The functions m_j, $j \in \mathbb{E}$, satisfy the following system of equations*: *for each $j \in \mathbb{E}$,*

$$m_j(z) = \sum_{i \in \mathbb{E}} \int_0^\infty e^{-\frac{\lambda_i (1 - e^{-\mu_i \nu t})(1-z)}{\mu_i \nu}} m_i \left(1 - e^{-\mu_i \nu t}(1-z)\right) dA_{i,j}(t). \tag{1}$$

Proof It was mentioned in the previous section that $m_i(z)$ is equal to $E_0[z^{Q(T_1)} \mathbf{1}(C(T_1) = i)]$, so to compute this we will first condition on $Q(0)$ and $C(0)$.

Suppose $Q(0) = m$ and $C(0) = i$, where $m \geq 0$ and $i \in \mathbb{E}$. After conditioning on the event that $T_1 = t$, we see that each of these m customers will be in the system at time t with probability $e^{-\mu_i \nu t}$. Furthermore, there could also be customers present in the system at time t that originally were not there at time zero. For a given customer that arrives to the system after time 0 but before time t, we see that the probability he or she is still in the system at time t is given by

$$\int_0^t e^{-\mu_i \nu (t-s)} \frac{1}{t} \, ds = \frac{1 - e^{-\mu_i \nu t}}{\mu_i \nu t}.$$

Since the number of customers that arrive in $(0, t]$ is Poisson with rate $\lambda_i t$, we see that

$$E_0\big[z^{Q(T_1)}\mathbf{1}\big(C(T_1)=j\big)\,|\,Q(0)=m,\,C(0)=i\big]$$

$$=\int_0^\infty\sum_{n=0}^\infty\frac{(\lambda_i t)^n e^{-\lambda_i t}}{n!}\big(1-e^{-\mu_i vt}(1-z)\big)^m\left(1-\frac{1-e^{-\mu_i vt}}{\mu_i vt}(1-z)\right)^n dA_{i,j}(t)$$

$$=\int_0^\infty e^{-\frac{\lambda_i(1-e^{-\mu_i vt})(1-z)}{\mu_i v}}\big(1-e^{-\mu_i vt}(1-z)\big)^m dA_{i,j}(t).$$

After unconditioning, we conclude that

$$m_j(z)=\sum_{i\in\mathbb{E}}\int_0^\infty e^{-\frac{\lambda_i(1-e^{-\mu_i vt})(1-z)}{\mu_i v}}m_i\big(1-e^{-\mu_i vt}(1-z)\big)dA_{i,j}(t)$$

which completes the proof.   □

The following corollary then immediately follows from this result.

Corollary 3.1 *The moments* $z_{n,j}=E_0[(Q(0))_n\mathbf{1}(C(0)=j)]$ *satisfy the following system of equations: for* $n\geq 1$ *and* $j\in\mathbb{E}$,

$$z_{n,j}=\sum_{i\in\mathbb{E}}\sum_{k=0}^n\binom{n}{k}\left(\frac{\lambda_i}{\mu_i v}\right)^{n-k}\left[\sum_{l=0}^{n-k}(-1)^l\binom{n-k}{l}\alpha_{i,j}\big((k+l)\mu_i v\big)\right]z_{k,i}. \quad (2)$$

Remark Notice that (2) can be used to compute all embedded factorial moments, in a recursive fashion: indeed, notice that the first factorial moments can be computed by solving a system of linear equations, where the number of equations and unknowns is equal to the cardinality of $\mathbb{E}$. Furthermore, once these are known, they can be plugged into a second system, and we end up with another system that is of the same size as the previous one. This procedure can clearly be repeated to produce higher moments as well.

The reader should also notice that this system of equations will always have a solution, namely the moments $z_{n,j}$. Proving that such a system has a unique solution, however, is an issue. In the special case $|\mathbb{E}|=2$, this appears to be simple (the solution is explicitly given in [7]).

Proof The proof of this result merely involves differentiating (1). From Leibniz rule, we see that for each $n\geq 1$,

$$m_j^{(n)}(z)=\sum_{i\in\mathbb{E}}\int_0^\infty e^{-\frac{\lambda_i(1-e^{-\mu_i vt})(1-z)}{\mu_i v}}\sum_{k=0}^n\binom{n}{k}e^{-k\mu_i vt}\left(\frac{\lambda_i}{\mu_i v}(1-e^{-\mu_i vt})\right)^{n-k}m_i^{(k)}$$

$$\times\big(1-e^{-\mu_i vt}(1-z)\big)^m dA_{i,j}(t). \quad (3)$$

By letting $z=1$ in (3), we observe that

$$E_0\big[(Q)_n\mathbf{1}(C=j)\big]=\sum_{i\in\mathbb{E}}\int_0^\infty\sum_{k=0}^n\binom{n}{k}e^{-k\mu_i vt}\left(\frac{\lambda_i}{\mu_i v}\right)^{n-k}$$

$$\times \sum_{l=0}^{n-k} (-1)^l \binom{n-k}{l} e^{-l\mu_i vt} E_0\big[(Q)_k \mathbf{1}(C=i)\big] dA_{i,j}(t)$$

$$= \sum_{i\in\mathbb{E}} \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\lambda_i}{\mu_i v}\right)^{n-k} \left[\sum_{l=0}^{n-k} (-1)^l \binom{n-k}{l} \alpha_{i,j}\big((k+l)\mu_i v\big)\right]$$

$$\times E_0\big[(Q)_k \mathbf{1}(C=i)\big]. \qquad \square$$

The following formula can be used to relate the moments at a transition epoch to the moments at an arbitrary epoch.

Theorem 3.2 *For each $j \in \mathbb{E}$,*

$$E\big[z^{Q(0)}\mathbf{1}(C(0)=j)\big] = \frac{1}{v} \int_0^\infty e^{-\frac{\lambda_j(1-e^{-\mu_j vt})(1-z)}{\mu_j v}} m_j$$

$$\times \big(1 - e^{\mu_j vt}(1-z)\big)\overline{A}_j(t)\pi_j \, dt, \qquad (4)$$

where $\overline{A}_j(t) = P(T_{n+1} - T_n > t | C(T_n) = j)$.

Proof From the inversion formula (see [2]), we also find that

$$E\big[z^{Q(0)}\mathbf{1}(C(0)=j)\big] = \frac{1}{v} E_0\left[\int_0^{T_1} z^{Q(t)} \mathbf{1}(C(t)=j) \, dt\right]$$

$$= \frac{1}{v} \int_0^\infty E_0\big[z^{Q(t)} | T_1 > t, C(t) = j\big]\overline{A}_j(t)\pi_j \, dt.$$

However, under the measure P_0, $T_0 = 0$ almost surely, and furthermore,

$$E_0\big[z^{Q(t)} | T_1 > t, C(t) = j\big]\pi_j = \pi_j e^{-\frac{\lambda_j}{\mu_j v}(1-e^{-\mu_j vt})(1-z)} m_j\big(1 - e^{-\mu_j vt}(1-z)\big).$$

Therefore,

$$E\big[z^{Q(0)}\mathbf{1}(C(0)=j)\big] = \frac{1}{v} \int_0^\infty e^{-\frac{\lambda_j(1-e^{-\mu_j vt})(1-z)}{\mu_j v}} m_j$$

$$\times \big(1 - e^{\mu_j vt}(1-z)\big)\overline{A}_j(t)\pi_j \, dt.$$

This proves the claim. $\qquad \square$

Corollary 3.2 *The nth factorial moment of the steady-state queue-length distribution is*

$$E\big[(Q(0))_n\big] = \sum_{i\in\mathbb{E}} \frac{\pi_i}{v} \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\lambda_i}{\mu_i v}\right)^{n-k} \sum_{l=0}^{n-k} (-1)^l \binom{n-k}{l}$$

$$\times \left[\frac{1}{(k+l)\mu_i v}\big(1 - \alpha_i\big((k+l)\mu_i v\big)\big)\mathbf{1}(k+l>0)\right.$$

$$\left. + \upsilon_i \mathbf{1}(k+l=0)\right] z_{k,i}.$$

Proof After applying Leibniz rule to (4), we also see that

$$E\big[(Q(0))_n z^{Q(0)-n} \mathbf{1}(C(0)=j)\big]$$

$$= \frac{1}{\upsilon}\int_0^\infty e^{-\frac{\lambda_j(1-e^{-\mu_j vt})(1-z)}{\mu_j v}} \sum_{k=0}^n \binom{n}{k} e^{-k\mu_j vt}\left(\frac{\lambda_j}{\mu_j v}\big(1-e^{-\mu_j vt}\big)\right)^{n-k}$$

$$\times m_j^{(k)}\big(1-e^{-\mu_j vt}(1-z)\big)\overline{A}_j(t)\pi_j \, dt$$

and so, putting $z=1$, gives

$$E\big[(Q(0))_n \mathbf{1}(C(0)=j)\big]$$

$$= \frac{1}{\upsilon}\int_0^\infty \sum_{k=0}^n \binom{n}{k} e^{-k\mu_j vt}\left(\frac{\lambda_j}{\mu_j v}\right)^{n-k} \sum_{l=0}^{n-k}(-1)^l\binom{n-k}{l} e^{-l\mu_j vt}$$

$$\times E_0\big[(Q)_k \mathbf{1}(C=j)\big]\overline{A}_j(t)\pi_j \, dt$$

$$= \frac{\pi_j}{\upsilon}\sum_{k=0}^n \binom{n}{k}\left(\frac{\lambda_j}{\mu_j v}\right)^{n-k}\sum_{l=0}^{n-k}(-1)^l\binom{n-k}{l}$$

$$\times \left[\frac{(1-\alpha_j((k+l)\mu_j v))}{(k+l)\mu_j v}\mathbf{1}(k+l>0) + \upsilon_j\mathbf{1}(k+l=0)\right]$$

$$\times E_0\big[(Q)_k \mathbf{1}(C=j)\big].$$

This concludes the proof.   □

4 Hyperexponential services

We continue by considering the case where the amount of service each customer brings to the system is hyperexponentially distributed, i.e., for each $t \geq 0$,

$$B(t) = \sum_{k=1}^{n_0} p_i\big(1 - e^{-\nu_i t}\big),$$

where n_0 is a positive integer, p_i, $1 \leq i \leq n_0$ are probabilities summing to 1 and $\nu_i > 0$, for $1 \leq i \leq n_0$. In our model, we can say that a given customer that arrives to the system is a type-i customer with probability p_i, and type-i customers bring an exponentially distributed (with rate ν_i) amount of work to the system. For ease of exposition, we will focus on the case when $n_0 = 2$, but it will become clear that the

same type of reasoning can be used for arbitrary n_0. The reader will lately see that the key to calculating all of the moments of the queue-length will depend on solving a system of linear equations, and so in order to gain his confidence in our statement we will provide, in the form of a remark, the linear systems required for an arbitrary n_0.

We will track the customers present in the system by using the processes $\{Q_i(t); t \in \mathbb{R}\}$, for $1 \le i \le n_0$, where $Q_i(t)$ represents the number of customers present in the system at time t that are of type i.

Again, our first step will involve deriving a system of equations, which are satisfied by the functions $m_j(z_1, z_2) = E_0[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j)]$.

Theorem 4.1 *The functions m_j, $j \in \mathbb{E}$, satisfy the following system of equations: for each $j \in \mathbb{E}$,*

$$m_j(z_1, z_2) = \sum_{i \in \mathbb{E}} \int_0^\infty e^{-p_1 \frac{\lambda_i}{\mu_i \nu_1}(1 - e^{-\mu_i \nu_1 t})(1 - z_1)} e^{-p_2 \frac{\lambda_i}{\mu_i \nu_2}(1 - e^{-\mu_i \nu_2 t})(1 - z_2)}$$
$$\times m_i\left(1 - e^{-\mu_i \nu_1 t}(1 - z_1), 1 - e^{-\mu_i \nu_2 t}(1 - z_2)\right) dA_{i,j}(t). \quad (5)$$

Proof We recall that $(Q_1(0), Q_2(0), C(0)) \stackrel{d}{=} (Q_1(T_1), Q_2(T_1), C(T_1))$, so to compute each m_j it will again be helpful to condition on $Q_1(0)$, $Q_2(0)$, and $C(0)$. Once we have conditioned on these variables, it will be useful to also condition on the number of new arrivals in the interval $(0, T_1]$, which we also condition on by assuming that it is of length t. Therefore, assuming $C(0) = i$, there are four different populations to consider: (1) type-1 customers currently in the system at 0, that leave before time T_1 with probability $1 - \eta_i$, (2) type-2 customers currently in the system at 0, that leave before T_1 with probability $1 - \delta_i$, (3) new type-1 customers that arrive in $(0, T_1]$, which leave before T_1 with probability $1 - \beta_i$, and finally (4) new type-2 customers that arrive in $(0, T_1]$, which leave before T_1 with probability $1 - \gamma_i$. Thus, $\eta_i = e^{-\mu_i \nu_1 t}$, $\delta_i = e^{-\mu_i \nu_2 t}$,

$$\beta_i = \int_0^t e^{-\mu_i \nu_1(t-s)} \frac{1}{t} ds = \frac{1 - e^{-\mu_i \nu_1 t}}{\mu_i \nu_1}$$

and

$$\gamma_i = \int_0^t e^{-\mu_i \nu_2(t-s)} \frac{1}{t} ds = \frac{1 - e^{-\mu_i \nu_2 t}}{\mu_i \nu_2}.$$

Now we are ready to begin our computations. Here

$$E_0\left[z_1^{Q_1(T_1)} z_2^{Q_2(T_1)} \mathbf{1}\left(C(T_1) = j\right) \big| Q_1(0) = i_1, Q_2(0) = i_2, C(0) = i\right]$$
$$= \int_0^\infty \sum_{n_1=0}^\infty \sum_{n_2=0}^\infty \frac{(p_1 \lambda_i t)^{n_1} e^{-p_1 \lambda_i t}}{n_1.} \frac{(p_2 \lambda_i t)^{n_2} e^{-p_2 \lambda_i t}}{n_2.} (1 - \eta_i + \eta_i z_1)^{n_1}$$
$$\times (1 - \delta_i + \delta_i z_2)^{n_2} (1 - \beta_i + \beta_i z_1)^{i_1} (1 - \gamma_i + \gamma_i z_2)^{i_2} dA_{i,j}(t)$$
$$= \int_0^\infty e^{-p_1 \lambda_i t \eta_i (1 - z_1)} e^{-p_2 \lambda_i t \delta_i (1 - z_2)} (1 - \beta + \beta z_1)^{i_1} (1 - \gamma + \gamma z_2)^{i_2} dA_{i,j}(t)$$

$$= \int_0^\infty e^{-p_1 \frac{\lambda_i}{\mu_i v_1}(1-e^{-\mu_i v_1 t})(1-z_1)} e^{-p_2 \frac{\lambda_i}{\mu_i v_2}(1-e^{-\mu_i v_2 t})(1-z_2)} \left(1 - e^{-\mu_i v_1 t}(1-z_1)\right)^{i_1}$$

$$\times \left(1 - e^{-\mu_i v_2 t}(1-z_2)\right)^{i_2} dA_{i,j}(t).$$

After unconditioning, we then see that

$$m_j(z_1, z_2) = \sum_{i \in \mathbb{E}} \int_0^\infty e^{-p_1 \frac{\lambda_i}{\mu_i v_1}(1-e^{-\mu_i v_1 t})(1-z_1)} e^{-p_2 \frac{\lambda_i}{\mu_i v_2}(1-e^{-\mu_i v_2 t})(1-z_2)}$$

$$\times m_i \left(1 - e^{-\mu_i v_1 t}(1-z_1), 1 - e^{-\mu_i v_2 t}(1-z_2)\right) dA_{i,j}(t)$$

and so (5) holds.   □

This immediately gives the following corollary.

Corollary 4.1 *The first moments* $z_{i,j} = E_0[Q_i(0)\mathbf{1}(C(0) = j)]$ *satisfy the following system of equations*: *for each* $j \in \mathbb{E}$,

$$z_{1,j} = \sum_{i \in \mathbb{E}} \left[\pi_i p_1 \frac{\lambda_i}{\mu_i v_1} \left(r_{i,j} - \alpha_{i,j}(\mu_i v_1)\right) + z_{1,i} \alpha_{i,j}(\mu_i v_1) \right]$$

and

$$z_{2,j} = \sum_{i \in \mathbb{E}} \left[\pi_i p_2 \frac{\lambda_i}{\mu_i v_2} \left(r_{i,j} - \alpha_{i,j}(\mu_i v_2)\right) + z_{2,i} \alpha_{i,j}(\mu_i v_2) \right].$$

Proof After taking partial derivatives in (5) and setting $(z_1, z_2) = (1, 1)$, we end up with the following system of equations:

$$E_0\big[Q_1(0)\mathbf{1}\big(C(0) = j\big)\big] = \sum_{i \in \mathbb{E}} \left[\pi_i p_1 \frac{\lambda_i}{\mu_i v_1} \left(r_{i,j} - \alpha_{i,j}(\mu_i v_1)\right) \right.$$

$$\left. + E_0\big[Q_1(0)\mathbf{1}(C(0) = i)\big]\alpha_{i,j}(\mu_i v_1) \right]$$

and

$$E_0\big[Q_2(0)\mathbf{1}\big(C(0) = j\big)\big] = \sum_{i \in \mathbb{E}} \left[\pi_i p_2 \frac{\lambda_i}{\mu_i v_2} \left(r_{i,j} - \alpha_{i,j}(\mu_i v_2)\right) \right.$$

$$\left. + E_0\big[Q_2(0)\mathbf{1}\big(C(0) = i\big)\big]\alpha_{i,j}(\mu_i v_2) \right],$$

which is the same as the equations given in the second part of the corollary. Notice that these equations are no more difficult to solve than the ones found in [7], which are the same as system (2); as a matter of fact, they can be split into two sets, with each set being of the same form as the system of equations in [7].   □

Remark From this argument, it is easy to see what the system of equations will look like if we consider an arbitrary number n_0 of types of customers. In particular, for each customer type-m, $1 \le m \le n_0$, we have the corresponding system of equations and unknowns:

$$z_{m,j} = \sum_{i \in \mathbb{E}} \left[\pi_i p_m \frac{\lambda_i}{\mu_i v_m} \big(r_{i,j} - \alpha_{i,j}(\mu_i v_m) \big) + z_{m,i} \alpha_{i,j}(\mu_i v_m) \right].$$

Again, we can relate these moments to the moments at an arbitrary time.

Theorem 4.2 *For each* $j \in \mathbb{E}$,

$$E\big[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j) \big]$$

$$= \frac{1}{v} \int_0^\infty e^{-p_1 \frac{\lambda_j}{\mu_j v_1}(1 - e^{-\mu_j v_1 t})(1 - z_1)} e^{-p_2 \frac{\lambda_j}{\mu_j v_2}(1 - e^{-\mu_j v_2 t})(1 - z_2)}$$

$$\times m_j \big(1 - e^{-\mu_j v_1 t}(1 - z_1), 1 - e^{-\mu_j v_2 t}(1 - z_2) \big) \overline{A}_j(t) \, dt.$$

Proof To compute the LST of the steady-state queue-length at an arbitrary instant, we can again use the inversion formula to conclude that

$$E\big[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j) \big]$$

$$= \frac{1}{v} E_0 \left[\int_0^\infty z_1^{Q_1(t)} z_2^{Q_2(t)} \mathbf{1}(C(t) = j) \mathbf{1}(T_1 > t) \, dt \right]$$

$$= \frac{1}{v} \int_0^\infty E_0 \big[z_1^{Q_1(t)} z_2^{Q_2(t)} | T_1 > t, C(t) = j \big] \overline{A}_j(t) \pi_j \, dt.$$

But $T_0 = 0$ almost surely under P_0, and furthermore,

$$E_0 \big[z_1^{Q_1(t)} z_2^{Q_2(t)} | T_1 > t, C(t) = j \big] \pi_j$$

$$= \pi_j e^{-p_1 \frac{\lambda_j}{\mu_j v_1}(1 - e^{-\mu_j v_1 t})(1 - z_1)} e^{-p_2 \frac{\lambda_j}{\mu_j v_2}(1 - e^{-\mu_j v_2 t})(1 - z_2)}$$

$$\times m_j \big(1 - e^{-\mu_i v_1 t}(1 - z_1), 1 - e^{-\mu_i v_2 t}(1 - z_2) \big).$$

Therefore,

$$E\big[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j) \big]$$

$$= \frac{1}{v} \int_0^\infty e^{-p_1 \frac{\lambda_j}{\mu_j v_1}(1 - e^{-\mu_j v_1 t})(1 - z_1)} e^{-p_2 \frac{\lambda_j}{\mu_j v_2}(1 - e^{-\mu_j v_2 t})(1 - z_2)}$$

$$\times m_j \big(1 - e^{-\mu_j v_1 t}(1 - z_1), 1 - e^{-\mu_j v_2 t}(1 - z_2) \big) \overline{A}_j(t) \, dt. \qquad \square$$

We can then get an expression for any moment we would like of the queue-length at an arbitrary time, by taking the appropriate derivatives.

Corollary 4.2 *The mean steady-state queue-length is given by*

$$E[Q(0)] = \sum_{j \in \mathbb{E}} \frac{1}{\upsilon} \left[\pi_j \frac{\lambda_j p_1}{\mu_j \nu_1} \upsilon_j + \left[z_{1,j} - \frac{\lambda_j p_1}{\mu_j \nu_1} \upsilon_j \right] \frac{1 - \alpha_j(\mu_j \nu_1)}{\mu_j \nu_1} \right]$$

$$+ \sum_{j \in \mathbb{E}} \frac{1}{\upsilon} \left[\pi_j \frac{\lambda_j p_2}{\mu_j \nu_2} \upsilon_j + \left[z_{2,j} - \frac{\lambda_j p_2}{\mu_j \nu_2} \upsilon_j \right] \frac{1 - \alpha_j(\mu_j \nu_2)}{\mu_j \nu_2} \right]. \quad (6)$$

Proof After taking derivatives, we see that

$$E[Q_1(0)\mathbf{1}(C(0) = j)] = \frac{1}{\upsilon} \left[\pi_j \frac{\lambda_j p_1}{\mu_j \nu_1} \upsilon_j + \left[E_0[Q_1(0)\mathbf{1}(C(0) = j)] \right. \right.$$

$$\left. \left. - \frac{\lambda_j p_1}{\mu_j \nu_1} \upsilon_j \right] \frac{1 - \alpha_j(\mu_j \nu_1)}{\mu_j \nu_1} \right]$$

and

$$E[Q_2(0)\mathbf{1}(C(0) = j)] = \frac{1}{\upsilon} \left[\pi_j \frac{\lambda_j p_2}{\mu_j \nu_2} \upsilon_j + \left[E_0[Q_2(0)\mathbf{1}(C(0) = j)] \right. \right.$$

$$\left. \left. - \frac{\lambda_j p_2}{\mu_j \nu_2} \upsilon_j \right] \frac{1 - \alpha_j(\mu_j \nu_2)}{\mu_j \nu_2} \right].$$

These expressions can then be used to derive (6). $\qquad \square$

Remark As was mentioned in the proof, higher moments of the queue-length can also be computed as well, by taking as many derivatives as needed. For instance, the system of equations that can be used to compute $E[Q_1(0)Q_2(0)]$ is as follows:

$$E_0[Q_1(0)Q_2(0)\mathbf{1}(C(0) = j)]$$

$$= \sum_{i \in \mathbb{E}} \frac{p_1}{\nu_1} \frac{\lambda_i}{\mu_i} \frac{p_2}{\nu_2} \frac{\lambda_i}{\mu_i} \pi_i \left(r_{i,j} - \alpha_{i,j}(\mu_i \nu_1) - \alpha_{i,j}(\mu_i \nu_2) + \alpha_{i,j}(\mu_i(\nu_1 + \nu_2)) \right)$$

$$+ \sum_{i \in \mathbb{E}} \frac{p_2}{\nu_2} \frac{\lambda_i}{\mu_i} \left(\alpha_{i,j}(\mu_i \nu_1) - \alpha_{i,j}(\mu_i(\nu_1 + \nu_2)) \right) E_0[Q_1(0)\mathbf{1}(C(0) = i)]$$

$$+ \sum_{i \in \mathbb{E}} \frac{p_1}{\nu_1} \frac{\lambda_i}{\mu_i} \left(\alpha_{i,j}(\mu_i \nu_2) - \alpha_{i,j}(\mu_i(\nu_1 + \nu_2)) \right) E_0[Q_2(0)\mathbf{1}(C(0) = i)]$$

$$+ \sum_{i \in \mathbb{E}} \alpha_{i,j}(\mu_i(\nu_1 + \nu_2)) E_0[Q_1(0)Q_2(0)\mathbf{1}(C(0) = i)].$$

5 Erlang services

We now consider the case where the services are Erlang-distributed, with $n_0 \geq 1$ phases. In this case,

$$B(t) = 1 - \sum_{k=0}^{n_0-1} \frac{(vt)^k e^{-vt}}{k!}.$$

Again, we will assume that $n_0 = 2$, but it will be clear as to how to proceed for large n_0.

Let $Q_i(t)$ denote the number of queueing customers present in the system that are in phase i, for $i = 1, 2$. Here a phase refers to a characteristic of the customer's service. In particular, each customer's service amount is Erlang-distributed, and this can be broken up into two exponential random variables. A given customer is said to be in phase 1 during the times when the server is processing work from this first exponential amount, and he or she is in phase 2 during the processing of the second amount. Throughout the rest of the paper, customers that are in phase i will be referred to as type-i customers.

Again, we can derive the same type of equations for the functions $m_j(z_1, z_2) = E_0[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j)]$.

Theorem 5.1 *The functions m_j satisfy the following system of equations*: *for each $j \in \mathbb{E}$,*

$$m_j(z_1, z_2) = \sum_{i \in \mathbb{E}} \int_0^\infty e^{-\lambda_i t \eta_{i,1}(1-z_1)} e^{-\lambda_i t \eta_{i,2}(1-z_2)} m_i$$

$$\times (1 - \beta_{i,1} - \beta_{i,2} + \beta_{i,1} z_1 + \beta_{i,2} z_2, 1 - \gamma_{i,2} + \gamma_{i,2} z_2) \, dA_{i,j}(t),$$

where $\eta_{i,n} = \frac{1}{\mu_i vt}[1 - \sum_{k=0}^{n-1} \frac{(\mu_i vt)^k e^{-\mu_i vt}}{k!}]$, $\beta_{i,n} = \frac{(\mu_i vt)^{n-1} e^{-\mu_i vt}}{(n-1)!}$, *and* $\gamma_{i,n} = \frac{(\mu_i vt)^{n-2} e^{-\mu_i vt}}{(n-2)!}$.

Proof Clearly, we see that

$$E_0\big[z_1^{Q_1(T_1)} z_2^{Q_2(T_1)} \mathbf{1}\big(C(T_1) = j\big) | Q_1(0) = i_1, Q_2(0) = i_2, C(0) = i\big]$$

$$= \int_0^\infty E_0\big[z_1^{Q_1(t)} z_2^{Q_2(t)} | Q_1(0) = i_1, Q_2(0) = i_2, C(0) = i, C(t) = j\big] dA_{i,j}(t)$$

$$= \int_0^\infty \sum_{n=0}^\infty E\big[z_1^{Q_1(t)} z_2^{Q_2(t)} | Q_1(0) = i_1, Q_2(0) = i_2, C(0) = i, C(t) = j,$$

$$N_{\lambda_j}(0, t] = n\big] \frac{(\lambda_i t)^n e^{-\lambda_i(t)}}{n.} \, dA_{i,j}(t),$$

where N_{λ_j} represents a homogeneous Poisson process with rate λ_j.

At this point we have to consider three different types of populations: type-1 customers that were present in the system at time 0, type-2 customers present at 0, and new arrivals in $(0, t]$. If, at the time of a transition the environment is in state i, each type-1 customer will, at time t, either be a type-1 customer, a type-2 customer, or it will leave the system, with probabilities $\beta_{i,1}$, $\beta_{i,2}$, and $1 - \beta_{i,1} - \beta_{i,2}$, respectively. A type-2 customer will stay as a type-2 with probability $\gamma_{i,2}$, and a new arrival will be type-1, type-2, or will leave with probabilities $\eta_{i,1}$, $\eta_{i,2}$, or $1 - \eta_{i,1} - \eta_{i,2}$, respectively. The behavior of each customer is independent of all other customers present in the system, and because of this we are able to just use multinomial transforms to compute the above expression. Therefore,

$$
\begin{aligned}
E_0 & \big[z_1^{Q_1(T_1)} z_2^{Q_2(T_1)} \mathbf{1}\big(C(T_1) = j\big) \big| Q_1(0) = i_1,\, Q_2(0) = i_2,\, C(0) = i \big] \\
&= \int_0^\infty \sum_{n=0}^\infty \frac{(\lambda_i t)^n e^{-\lambda_i t}}{n!} (1 - \eta_{i,1} - \eta_{i,2} + \eta_{i,1} z_1 + \eta_{i,2} z_2)^n \\
&\quad \times (1 - \beta_{i,1} - \beta_{i,2} + \beta_{i,1} z_1 + \beta_{i,2} z_2)^{i_1} \\
&\quad \times (1 - \gamma_{i,2} + \gamma_{i,2} z_2)^{i_2} \, dA_{i,j}(t) \\
&= \int_0^\infty e^{-\lambda_i t (\eta_{i,1}(1-z_1) + \eta_{i,2}(1-z_2))} (1 - \beta_{i,1} - \beta_{i,2} + \beta_{i,1} z_1 + \beta_{i,2} z_2)^{i_1} \\
&\quad \times (1 - \gamma_{i,2} + \gamma_{i,2} z_2)^{i_2} \, dA_{i,j}(t).
\end{aligned}
$$

Again, after unconditioning we find that

$$
\begin{aligned}
m_j(z_1, z_2) = \sum_{i \in \mathbb{E}} \int_0^\infty & e^{-\lambda_i t \eta_{i,1}(1-z_1)} e^{-\lambda_i t \eta_{i,2}(1-z_2)} \\
& \times m_i(1 - \beta_{i,1} - \beta_{i,2} + \beta_{i,1} z_1 + \beta_{i,2} z_2,\, 1 - \gamma_{i,2} + \gamma_{i,2} z_2) \, dA_{i,j}(t).
\end{aligned}
$$

All we need to do now is compute the η's, β's and γ's. When computing each $\eta_{i,n}$, we can apply the Campbell–Mecke formula (see [2]) to quickly deduce that, for each n,

$$
\begin{aligned}
\eta_{i,n} &= \int_0^t P\big(N_{\mu_i \nu}(t - s) = n - 1\big) \frac{1}{t} \, ds \\
&= \frac{1}{\mu_i \nu t} \left[1 - \sum_{k=0}^{n-1} \frac{(\mu_i \nu t)^k e^{-\mu_i \nu t}}{k!} \right].
\end{aligned}
$$

Here N_χ is used to represent a homogeneous Poisson process with rate χ, for all $\chi > 0$. Furthermore,

$$
\beta_{i,n} = P\big(N_{\mu_i \nu}(t) = n - 1\big) = \frac{(\mu_i \nu t)^{n-1} e^{-\mu_i \nu t}}{(n - 1)!}
$$

and

$$\gamma_{i,n} = P\big(N_{\mu_i v}(t) = n - 2\big) = \frac{(\mu_i v t)^{n-2} e^{-\mu_i v t}}{(n-2)!}.$$

This concludes the proof.   □

Corollary 5.1 *The first moments $z_{i,j} = E[Q_i(0)\mathbf{1}(C(0) = j)]$ can be found by solving the following systems of equations*: *for $j \in \mathbb{E}$,*

$$z_{1,j} = \sum_{i \in \mathbb{E}}\left[\pi_i \frac{\lambda_i}{\mu_i v}\big(r_{i,j} - \alpha_{i,j}(\mu_i v)\big) + \alpha_{i,j}(\mu_i v)z_{1,i}\right]$$

and

$$z_{2,j} = \sum_{i \in \mathbb{E}}\left[\pi_i \frac{\lambda_i}{\mu_i v}\big(r_{i,j} - \alpha_{i,j}(\mu_i v) + \mu_i v \alpha'_{i,j}(\mu_i v)\big)\right.$$

$$\left. - \mu_i v \alpha'_{i,j}(\mu_i v)z_{1,i} + \alpha_{i,j}(\mu_i v)z_{2,i}\right].$$

Proof After taking derivatives, setting $z = 1$ and plugging-in the correct expressions for η_n, β_n and γ_n, we conclude that

$$E_0\big[Q_1(0)\mathbf{1}\big(C(0) = j\big)\big] = \sum_{i \in \mathbb{E}}\left[\pi_i \frac{\lambda_i}{\mu_i v}\big(r_{i,j} - \alpha_{i,j}(\mu_i v)\big)\right.$$

$$\left. + \alpha_{i,j}(\mu_i v)E_0\big[Q_1(0)\mathbf{1}\big(C(0) = i\big)\big]\right]$$

and

$$E_0\big[Q_2(0)\mathbf{1}\big(C(0) = j\big)\big] = \sum_{i \in \mathbb{E}}\left[\pi_i \frac{\lambda_i}{\mu_i v}\big(r_{i,j} - \alpha_{i,j}(\mu_i v) + \mu_i v \alpha'_{i,j}(\mu_i v)\big)\right.$$

$$\left. - \mu_i v \alpha'_{i,j}(\mu_i v)E_0\big[Q_1(0)\mathbf{1}\big(C(0) = i\big)\big]\right]$$

$$+ \sum_{i \in \mathbb{E}}\alpha_{i,j}(\mu_i v)E_0\big[Q_2(0)\mathbf{1}\big(C(0) = i\big)\big].$$ □

Remark If there are an arbitrary number of phases n_0, it is not difficult to see that for each l, j, where $1 \le l \le n_0$ and $j \in \mathbb{E}$,

$$z_{l,j} = \sum_{i \in \mathbb{E}}\frac{\pi_i \lambda_i}{v \mu_i}\left[r_{i,j} - \sum_{k=0}^{l-1}\frac{(-\mu_i v)^k \alpha_{i,j}^{(k)}(\mu_i v)}{k!}\right]$$

$$+ \sum_{i \in \mathbb{E}}\sum_{m=1}^{l}z_{m,i}\frac{(-\mu_i v)^{l-m}\alpha_{i,j}^{(l-m)}(\mu_i v)}{(l-m)!}.$$

The moments at an arbitrary time can also be computed through the use of exactly the same methods as were used before.

Theorem 5.2 *For each $j \in \mathbb{E}$, we see that*

$$E\left[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j)\right] = \frac{1}{\upsilon} \int_0^\infty e^{-\lambda_j t \eta_{j,1}(1-z_1)} e^{-\lambda_j t \eta_{j,2}(1-z_2)}$$
$$\times m_j(1 - \beta_{j,1} - \beta_{j,2} + \beta_{j,1} z_1 + \beta_{j,2} z_2, 1 - \gamma_{j,2}$$
$$+ \gamma_{j,2} z_2) \overline{A}_j(t) \, dt.$$

Proof In order to convert from the embedded steady-state distribution to the steady-state distribution at an arbitrary time, we will again use the inversion formula:

$$E\left[z_1^{Q_1(0)} z_2^{Q_2(0)} \mathbf{1}(C(0) = j)\right]$$
$$= \frac{1}{\upsilon} E_0 \left[\int_0^\infty z_1^{Q_1(t)} z_2^{Q_2(t)} \mathbf{1}(T > t, C(t) = j) dt \right]$$
$$= \frac{1}{\upsilon} \int_0^\infty e^{-\lambda_j t \eta_{j,1}(1-z_1)} e^{-\lambda_j t \eta_{j,2}(1-z_2)}$$
$$\times E_0 \big((1 - \beta_{j,1} - \beta_{j,2} + \beta_{j,1} z_1 + \beta_{j,2} z_2)^{Q_1(0)} $$
$$\times (1 - \gamma_{j,2} + \gamma_{j,2} z_2)^{Q_2(0)} \mathbf{1}(C(0) = j) \big) \overline{A}_j(t) \, dt$$
$$= \frac{1}{\upsilon} \int_0^\infty e^{-\lambda_j t \eta_{j,1}(1-z_1)} e^{-\lambda_j t \eta_{j,2}(1-z_2)}$$
$$\times m_j(1 - \beta_{j,1} - \beta_{j,2} + \beta_{j,1} z_1 + \beta_{j,2} z_2, 1 - \gamma_{j,2} + \gamma_{j,2} z_2) \overline{A}_j(t) \, dt. \quad \square$$

Again, computing any steady-state moments of interest will involve plugging-in the appropriate values for the $\eta_{j,k}$, $\beta_{j,k}$ and $\gamma_{j,k}$ probabilities, and then differentiating. Thus,

Corollary 5.2 *The mean number of customers in the system during steady state is just*

$$E[Q(0)] = \sum_{j \in \mathbb{E}} \left[E[Q_1(0)\mathbf{1}(C(0) = j)] + E[Q_2(0)\mathbf{1}(C(0) = j)] \right],$$

where

$$E[Q_1(0)\mathbf{1}(C(0) = j)] = \frac{1}{\upsilon} \left[\frac{\lambda_j \pi_j}{\mu_j \upsilon} \left[\upsilon_j - \frac{1 - \alpha_j(\mu_j \upsilon)}{\mu_j \upsilon} \right] + \frac{1 - \alpha_j(\mu_j \upsilon)}{\mu_j \upsilon} z_{1,j} \right]$$

and

$$E\big[Q_2(0)\mathbf{1}\big(C(0)=j\big)\big]$$
$$= \frac{1}{\upsilon}\left[\frac{\lambda_j \pi_j}{\mu_j \upsilon}\left[\upsilon_j - \frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon} - \frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon} + \alpha'_j(\mu_j\upsilon)\right]\right]$$
$$+ \frac{1}{\upsilon}\left[\left(\frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon} - \alpha'_j(\mu_j\upsilon)\right)z_{1,j}\right]$$
$$+ \frac{1}{\upsilon}\frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon}z_{2,j}.$$

Proof After taking derivatives and letting $z=1$, we get

$$E\big[Q_1(0)\mathbf{1}\big(C(0)=j\big)\big] = \frac{1}{\upsilon}\left[\frac{\lambda_j \pi_j}{\mu_j \upsilon}\left[\upsilon_j - \frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon}\right]\right.$$
$$\left. + \frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon}E_0\big[Q_1(0)\mathbf{1}\big(C(0)=j\big)\big]\right]$$

and

$$E\big[Q_2(0)\mathbf{1}\big(C(0)=j\big)\big]$$
$$= \frac{1}{\upsilon}\left[\frac{\lambda_j \pi_j}{\mu_j \upsilon}\left[\upsilon_j - \frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon} - \frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon} + \alpha'_j(\mu_j\upsilon)\right]\right]$$
$$+ \frac{1}{\upsilon}\left[\left(\frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon} - \alpha'_j(\mu_j\upsilon)\right)E_0\big[Q_1(0)\mathbf{1}\big(C(0)=j\big)\big]\right]$$
$$+ \frac{1}{\upsilon}\frac{1-\alpha_j(\mu_j\upsilon)}{\mu_j\upsilon}E_0\big[Q_2(0)\mathbf{1}\big(C(0)=j\big)\big].$$

This concludes the proof.   □

6 General services

At this point it is also clear that expressions for the steady-state moments of the queue-length can be computed in the case when $B(t)$ is a mixture of Erlang distributions, and from a result in Asmussen [1] we also know that these distributions are dense in the space of distributions with nonnegative support. Our goal now is to prove a continuity theorem, which will allow us to approximate all steady-state moments of the queue-length by approximating arbitrary services with Erlang mixtures. In this section it will be necessary to assume that $\mu_i > 0$, for $i \in \mathbb{E}$; we assumed it as well in previous sections, but there the assumption was not strictly needed.

Theorem 6.1 *Let $\{Q_m(t); t \in \mathbb{R}\}$ denote a stationary version of the queue-length process described above, where each customer brings an amount of work B_n^m, $n \in \mathbb{Z}$. If, for each n, B_n^m converges weakly to B_n as $m \to \infty$, then for each integer k,*

$$\lim_{m \to \infty} E\big[Q_m(0)^k\big] = E\big[Q(0)^k\big].$$

Proof To prove this result, we will first show that, as $m \to \infty$, $E[Q_m(0)|C] \to E[Q(0)|C]$, where $C = \{C(t); t \in \mathbb{R}\}$ represents the entire path of our semi-Markovian environment. Proving this will immediately imply that $E[Q_m(0)^k|C] \to E[Q(0)^k|C]$ for all $k \geq 2$ as well, since conditional on C, the random variables $Q_m(0)$ and $Q(0)$ are all Poisson-distributed (from Theorem 2.1), and the kth factorial moment of a Poisson random variable is just the mean raised to the kth power.

Unfortunately, conditioning on the sample-path of C no longer allows us to use any stationary properties of our system, so we will have to introduce a small amount of notation. Let $\{W(s); s \in \mathbb{R}\}$ denote a stochastic process, where $W(s)$ denotes the amount of work brought to the system by the last customer to arrive at or before time s. Then, from a transient version of Little's law (see Fralix et al. [8]), we see that

$$E\big[Q_m(0)|C\big] = \int_{-\infty}^{0} \mathcal{P}_s\bigg(W_m(s) > \int_{s}^{0} \mu_{C(x)}\, dx \bigg)\lambda_{C(s)}\, ds.$$

Here $\mathcal{P} : \mathbb{R} \times \mathcal{F} \to [0,1]$ is the Palm probability kernel induced by the non-homogeneous Poisson arrival process (keep in mind that we are still conditioning on C). We will not go into great detail to explain how these probabilities are derived: rather, we will simply state that for a given event $A \in \mathcal{F}$, $\mathcal{P}_s(A)$ can be interpreted as the probability of A, given that there is an arrival at time s. Details behind the construction of these measures can be found in, for instance, Chap. 10 of [9].

Clearly, for each s, $\mathcal{P}_s(W_m(s) > \int_{s}^{0} \mu_{C(x)}\, dx)$ represents the probability that the person arriving at time s is still in the system at time zero. It is a simple exercise to verify that the distribution of $W(s)$ under $\mathcal{P}_s$ is indeed the distribution of an arbitrary service time, so under $\mathcal{P}_s$, $W_m(s) \Rightarrow W(s)$ as $m \to \infty$, where $\Rightarrow$ is used to denote weak convergence.

We will eventually want to apply the dominated convergence theorem, so with this in mind we will let $\lambda^* = \sup_{i \in \mathbb{E}} \lambda_i$ and $\mu^* = \inf_{i \in \mathbb{E}} \mu_i$. Thus, since $\mathcal{P}_s(W_m(s) > \int_{s}^{0} \mu_{C(x)}\, dx)\lambda_{C(s)} \leq \mathcal{P}_s(W_m(s) > -\mu^* s)\lambda^*$, the dominated convergence theorem allows us to conclude that the conditional first moments converge. To show that the unconditional moments converge, notice that for each k, $E[Q_m(0)^k|C] \leq E[Q_{m,\lambda^*,\mu^*}(0)^k]$, which represents the kth moment of an $M/G/\infty$ queue with arrival rate λ^* and service speed μ^*, that observes the same work sequence as the mth queue. Furthermore, from a slight extension of the dominated convergence theorem (see Theorem 1.21 of Kallenberg [10]), we can conclude that all moments converge as well. □

7 Conclusions

We have shown that an embedded process approach can be used to both provide yet another method of computing the moments and the generating function of the steady-state queue-length of a semi-Markov-modulated $M/M/\infty$ queue. Moreover, the approach can also be used when mixtures of Erlang services are considered, and this immediately gives an approximation of all moments of the steady-state queue-length in such a system that has generally distributed service times.

References

1. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
2. Baccelli, F., Brémaud, P.: Elements of Queueing Theory: Palm–Martingale Calculus and Stochastic Recurrences. Springer, New York (2003)
3. Baykal-Gursoy, M., Xiao, W.: Stochastic decomposition in $M/M/\infty$ queues with Markov-modulated service rates. Queueing Syst. **48**, 75–88 (2004)
4. D'Auria, B.: $M/M/\infty$ queues with on/off service speeds. Technical report, EURANDOM, Eindhoven, The Netherlands (2005)
5. D'Auria, B.: Stochastic decomposition of the $M/G/\infty$ queue in a random environment. Oper. Res. Lett. **35**, 805–812 (2007)
6. D'Auria, B.: $M/M/\infty$ queues in semi-Markovian random environment. Queueing Syst. **58**, 221–237 (2008)
7. Falin, G.: The $M/M/\infty$ queue in a random environment. Queueing Syst. **58**, 65–76 (2008)
8. Fralix, B.H., Riano, G., Serfozo, R.F.: Time-dependent Palm probabilities and queueing applications. Technical report, EURANDOM, Eindhoven, The Netherlands (2007)
9. Kallenberg, O.: Random Measures. Akademie-Verlag, Berlin (1983)
10. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (2002)
11. Keilson, J., Servi, L.D.: The matrix $M/M/\infty$ system: retrial models and Markov-modulated sources. Adv. Appl. Probab. **25**, 453–471 (1993)
12. Nain, P., Nunez-Queija, R.: An $M/M/1$ queue in a semi-Markovian environment. In: ACM Sigmetrics 2001/Performance 2001, Cambridge, MA, June 17–20, 2001
13. O'Cinneide, C.A., Purdue, P.: The $M/M/\infty$ queue in a random environment. J. Appl. Probab. **23**, 175–184 (1986)
14. Serfozo, R.F.: Introduction to Stochastic Networks. Springer, New York (1999)
15. Serfozo, R.F.: Basics of Applied Stochastic Processes (2008, in progress)
16. Takine, T.: Single-server queues with Markov-modulated arrivals and service speed. Queueing Syst. **49**, 7–22 (2005)