Queueing Systems (2019) 91:367–401
https://doi.org/10.1007/s11134-019-09603-4

On the distributions of infinite server queues with batch arrivals

Andrew Daw[1] · **Jamol Pender**[2]

Received: 27 September 2018 / Revised: 28 January 2019 / Published online: 15 February 2019

Abstract
Queues that feature multiple entities arriving simultaneously are among the oldest models in queueing theory, and are often referred to as "batch" (or, in some cases, "bulk") arrival queueing systems. In this work, we study the effect of batch arrivals on infinite server queues. We assume that the arrival epochs occur according to a Poisson process, with treatment of both stationary and non-stationary arrival rates. We consider both exponentially and generally distributed service durations, and we analyze both fixed and random arrival batch sizes. In addition to deriving the transient mean, variance, and moment-generating function for time-varying arrival rates, we also find that the steady-state distribution of the queue is equivalent to the sum of scaled Poisson random variables with rates proportional to the order statistics of its service distribution. We do so through viewing the batch arrival system as a collection of correlated sub-queues. Furthermore, we investigate the limiting behavior of the process through a batch scaling of the queue and through fluid and diffusion limits of the arrival rate. In the course of our analysis, we make important connections between our model and the harmonic numbers, generalized Hermite distributions, and truncated polylogarithms.

Keywords Batch arrivals · Infinite server · General service · Time-varying

Mathematics Subject Classification 60K25 · 90B22 · 60G50

✉ Andrew Daw
amd399@cornell.edu

Jamol Pender
jjp274@cornell.edu

[1] School of Operations Research and Information Engineering, Cornell University, 257 Rhodes Hall, Ithaca, NY 14853, USA

[2] School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, Ithaca, NY 14853, USA

1 Introduction

Queueing systems with batch arrivals have enjoyed a long and rich history of study, at least on the timescale of queueing theory. Researchers have been exploring models of this sort for no less than six decades, based on the April 1958 submission date of Miller Jr [26]. Given this stretch of time, a wide variety of systems and settings have been considered under the banner of batch arrivals. Much of the earliest work focuses on single-server models, including Miller Jr [26], Lucantoni [22], Masuyama and Takine [25], Liu and Templeton [20], and Foster [12], although infinite server models followed soon after, such as work by Shanbhag [36] and Brown and Ross [2]. Later work has expanded the concept into a variety of related models, such as for priority queues [37] and for handling server vacations [19]. Additionally, there is some work that proves heavy traffic limit theorems for queues with batch arrivals. Examples of this include Chiamsiri and Leonard [3], Pang and Whitt [28], and Pender [29]. These papers show that one can approximate the queue length process with Brownian motion and Ornstein–Uhlenbeck processes and also show that one can exploit the approximations even in multi-server and non-Markovian settings.

In this paper, we consider queues with arrivals occurring at times following a Poisson process, with consideration given to both non-stationary and stationary rates. We analyze both general and exponential service as conducted by infinitely many servers. Additionally, this work addresses both fixed and random batch sizes. Our analysis starts with the fixed batch size case. We begin by analyzing the transient behavior of the queue with Markovian service and time-varying arrival rates, providing explicit forms for the moment-generating function, mean, and variance. Then, we show that if the arrival rate is stationary the resulting steady-state distribution can be written as a sum of independent, nonidentical, scaled Poisson random variables. This leads us to uncover connections to the harmonic numbers and generalizations of the Hermite distribution. By viewing the batch arrival queue as a collection of infinite server subqueues that receive solitary arrivals simultaneously, we are able to extend this Poisson sum construction to general service distributions. This perspective also provides an avenue for us to extend to random batch sizes. We also give fluid and diffusion scalings of the queue in the case of random batch sizes, as well as extending many of the results we found for fixed batch sizes.

One can note that the batch arrival queue may not always be given the name "batch," as many authors choose to use the term "bulk" instead. Predominantly, this reflects two leading strands of applications, where "bulk" often gives a connotation of transportation settings, whereas "batch" frequently implies applications in communications. Just as practical by any other name, this family of models has also been studied in a wide variety of applications beyond these two. Perhaps one most distinct from other types of queueing models is particle splitting in DNA caused by radiation, as discussed in Sachs et al. [35]. In this application, primary particles arrive at a cell nucleus and cause DNA double-strand breaks. These double-strand breaks occur in near simultaneity and are thus modeled as arriving in batches of random size, as it is possible that any number of double-strand breaks will be induced. After they are induced, the double-strand breaks are then processed by cellular enzymes, corresponding to service in the queueing model. Another interesting and modern application of these models

is in cloud-based data processing. In this case, the batches arriving to the system are collections of jobs submitted simultaneously. These jobs are then served by each being processed individually and returned. For more discussion, detailed models, and specific analysis for this setting, see works such as Lu et al. [21], Pender and Phung-Duc [31], Xie et al. [38], Yekkehkhany et al. [39] and references therein.

1.1 Main contributions of paper

Our contributions in this work can be summarized as follows:

(i) We show that an infinite server queue with batch arrivals at Poisson process epochs is equivalent in steady-state distribution to a sum of scaled independent Poisson random variables, including for generally distributed service and randomly distributed batch sizes. For exponential service, this reveals a connection to the harmonic numbers and generalized Hermite distributions.
(ii) We derive a limit of the process in which the batch size grows infinitely large and the number of entities in the system is scaled inverse proportionally, yielding a novel distribution characterized by the exponential integral functions. For distributions that meet a divisibility condition, we find that this also holds for random batch sizes.
(iii) In the case of time-varying arrival rates, we give a transient moment-generating function for fixed batch sizes as well as means and variance for both fixed and randomly sized batches.
(iv) We give fluid and diffusion limits of the queue for stationary arrival rates for batches of random size.

1.2 Organization of paper

The body of the remainder of this paper is organized into two main sections: Sects. 2 and 3. In Sect. 2, we consider systems in which the size of the batches is fixed. Similarly, we devote Sect. 3 to the case of randomly distributed batch sizes. At the beginning of each section, we give a detailed overview of the contents within and provide context for the analysis in terms of this project's scope. After these sections, we conclude in Sect. 4.

2 Batches of deterministic size

In this section, we will consider infinite server queues with arrivals occurring in batches of a fixed size. We will assume that the arrival epochs occur according to a Poisson process, including both stationary and non-stationary models. We also will investigate both exponentially and generally distributed service times.

This section starts with studying the case of Markovian arrivals and service in transient state in Sect. 2.1. For a time-varying arrival rate, we give the mean, variance, and moment-generating function. We then use this in Sect. 2.2 to find the steady-state distribution of the queue. Upon observing that this can be represented

as a sum of scaled Poisson random variables, we establish connections to generalized Hermite distributions and to the harmonic numbers. Taking motivation from this, we derive the distribution of the limit of the scaled system as the batch size grows infinitely large. Finally, in Sect. 2.3, we examine the batch queue as a collection of infinite server sub-queues that simultaneously receive solitary arrivals. In doing so, we extend our understanding of the steady-state distribution to the case of general service.

2.1 Transient analysis of the Markovian setting

We begin our analysis with the case of non-stationary Poisson arrival epochs and Markovian service. In Kendall notation, this is the $M_t^n/M/\infty$ queue. We let Q_t represent the number of entities present in the queueing system at time $t \geq 0$, which we often refer to as the "number in system." We will use this notation throughout the remainder of this work, where the precise setting of the queue will be implied by context. In this fully Markovian setting, we can use Dynkin's infinitesimal generator theorem to support our analysis. Specifically, we can note that for a sufficiently regular function $f : \mathbb{N} \to \mathbb{R}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[f(Q_t)\right] = \mathrm{E}\left[\lambda(t)\left(f(Q_t+n) - f(Q_t)\right) + \mu Q_t\left(f(Q_t-1) - f(Q_t)\right)\right], \tag{2.1}$$

for a batch arrival queue with arrival intensity $\lambda(t) > 0$. We will see in this subsection that this infinitesimal generator approach gives us a potent toolkit for exploring this model. Moreover, the insights we find in Markovian settings now and in Sect. 2.2 will provide intuition that will guide our investigation of this system when the Markov property does not hold. To begin, we now derive the moment-generating function of the number in system. We do so for a system with a non-stationary arrival rate given by a Fourier series, allowing these results to hold for all periodic arrival patterns.

Proposition 2.1 *For $\theta \in \mathbb{R}$, let $\mathcal{M}(\theta, t) = \mathrm{E}\left[e^{\theta Q_t}\right]$ be the moment-generating function of the number in system of an infinite server queue with periodic arrival rate $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. Then, $\mathcal{M}(\theta, t)$ is given by*

$$\begin{aligned}\mathcal{M}(\theta,t) = \left(e^{-\mu t}(e^{\theta}-1)+1\right)^{Q_0} & \\ e^{\sum_{j=1}^{n}\binom{n}{j}(e^{\theta}-1)^j\left(\frac{\lambda}{j\mu}\left(1-e^{-j\mu t}\right)+\sum_{k=1}^{\infty}\frac{(a_k j\mu - b_k k)}{k^2+j^2\mu^2}\left(\cos(kt)-e^{-j\mu t}\right)\right)} & \\ \cdot\, e^{\sum_{j=1}^{n}\binom{n}{j}(e^{\theta}-1)^j\sum_{k=1}^{\infty}\frac{(a_k k + b_k j\mu)\sin(kt)}{k^2+j^2\mu^2}} &\end{aligned} \tag{2.2}$$

for all time $t \geq 0$, where Q_0 is the initial number in the system.

Proof From Eq. (2.1), the MGF is given by the solution to the partial differential equation

$$\frac{\partial}{\partial t}\mathcal{M}(\theta,t) = \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)\right)\left(e^{n\theta} - 1\right)\mathcal{M}(\theta,t) + \mu\left(e^{-\theta} - 1\right)\frac{\partial}{\partial\theta}\mathcal{M}(\theta,t),$$

with the initial solution $\mathcal{M}(\theta,0) = e^{\theta Q_0}$. Because $\frac{\mathrm{d}\log(f(x))}{\mathrm{d}x} = \frac{1}{f(x)}\frac{\mathrm{d}f(x)}{\mathrm{d}x}$, we can observe that the partial differential equation for the cumulant generating function $G(\theta,t) = \log\left(\mathrm{E}\left[e^{\theta Q_t}\right]\right)$ is

$$\mu(1-e^{-\theta})\frac{\partial G(\theta,t)}{\partial\theta} + \frac{\partial G(\theta,t)}{\partial t} = \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)\right)(e^{n\theta} - 1),$$

with the initial condition $G(\theta,0) = \log\left(\mathrm{E}\left[e^{\theta Q_0}\right]\right) = \theta Q_0$. We will now solve this PDE by the method of characteristics. We begin by establishing the characteristic ODEs and corresponding initial solutions as follows:

$$\begin{aligned}
&\frac{\mathrm{d}\theta}{\mathrm{d}s}(r,s) = \mu(1-e^{-\theta}), && \theta(r,0) = r,\\
&\frac{\mathrm{d}t}{\mathrm{d}s}(r,s) = 1, && t(r,0) = 0,\\
&\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) = \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt)\right)(e^{n\theta} - 1), && g(r,0) = rQ_0.
\end{aligned}$$

The first two of these initial value problems yield the following solutions.

$$\begin{aligned}
\theta(r,s) = \log(e^{c_1(r)+\mu s} + 1) \quad &\Rightarrow \quad \theta(r,s) = \log\left((e^r - 1)e^{\mu s} + 1\right),\\
t(r,s) = s + c_2(r) \quad &\Rightarrow \quad t(r,s) = s.
\end{aligned}$$

Therefore, we can simplify the remaining characteristic ODE to

$$\begin{aligned}
\frac{\mathrm{d}g}{\mathrm{d}s}(r,s) &= \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(ks) + b_k \sin(ks)\right)\left(\left((e^r - 1)e^{\mu s} + 1\right)^n - 1\right)\\
&= \left(\lambda + \sum_{k=1}^{\infty} a_k \cos(ks) + b_k \sin(ks)\right)\sum_{j=1}^{n}\binom{n}{j}(e^r - 1)^j e^{j\mu s},
\end{aligned}$$

and this produces the general solution of

$$g(r,s) = c_3(r) + \sum_{j=1}^{n} \binom{n}{j} (e^r - 1)^j \left(\frac{\lambda}{j\mu} + \sum_{k=1}^{\infty} \frac{(a_k j\mu - b_k k)\cos(ks)}{k^2 + j^2\mu^2} + \frac{(a_k k + b_k j\mu)\sin(ks)}{k^2 + j^2\mu^2} \right) e^{j\mu s}.$$

This now equates to

$$g(r,s) = rQ_0 + \sum_{j=1}^{n} \binom{n}{j} (e^r - 1)^j \left(\frac{\lambda}{j\mu} \left(e^{j\mu s} - 1 \right) + \sum_{k=1}^{\infty} \frac{(a_k j\mu - b_k k)}{k^2 + j^2\mu^2} \left(\cos(ks) e^{j\mu s} - 1 \right) + \sum_{k=1}^{\infty} \frac{(a_k k + b_k j\mu)\sin(ks)}{k^2 + j^2\mu^2} e^{j\mu s} \right)$$

as the solution to the initial value problem. We now find the solution to the original PDE by solving for each characteristic variable in terms of t and θ and then substituting these expression into $g(r,s)$. That is, for $s = t$ and $r = \log\left(e^{-\mu t}(e^\theta - 1) + 1\right)$, we have that

$$\begin{aligned} G(\theta,t) &= g\left(\log\left(e^{-\mu t}(e^\theta - 1) + 1\right), t\right) \\ &= \log\left(e^{-\mu t}(e^\theta - 1) + 1\right) Q_0 + \sum_{j=1}^{n} \binom{n}{j} (e^\theta - 1)^j \left(\frac{\lambda}{j\mu} \left(1 - e^{-j\mu t} \right) \right. \\ &\quad + \sum_{k=1}^{\infty} \frac{(a_k j\mu - b_k k)}{k^2 + j^2\mu^2} \cdot \left(\cos(kt) - e^{-j\mu t} \right) \\ &\quad \left. + \sum_{k=1}^{\infty} \frac{(a_k k + b_k j\mu)}{k^2 + j^2\mu^2} \sin(kt) \right). \end{aligned}$$

To conclude the proof, we note that $\mathcal{M}(\theta,t) = e^{G(\theta,t)}$. □

We now extend this analysis through the two following corollaries: First, for systems with a stationary arrival rate, say $\lambda > 0$, we further specify the moment-generating function explicitly in Corollary 2.2. This will be of use when we explore the distribution of the queue in steady state, which we begin in Sect. 2.2. As with Proposition 2.1, the uniqueness of moment-generating functions will aid us in later exploration of the distributions within this model and within generalizations of it.

Corollary 2.2 *For $\theta \in \mathbb{R}$, let $\mathcal{M}(\theta,t) = \mathrm{E}\left[e^{\theta Q_t}\right]$ be the moment-generating function of the number in system of an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. Then, $\mathcal{M}(\theta,t)$ is given by*

$$\mathcal{M}(\theta, t) = \left(e^{-\mu t}(e^{\theta} - 1) + 1\right)^{Q_0} e^{\lambda \sum_{j=1}^{n} \binom{n}{j} \frac{(e^{\theta}-1)^j}{j\mu}(1-e^{-j\mu t})} \tag{2.3}$$

for all time $t \geq 0$, where Q_0 is the initial number in the system.

For the second direct result of Proposition 2.1, we can also give explicit expressions for the transient mean and variance of the queue. We derive these equations from the first and second derivatives, respectively, of the cumulant generating function $\log(\mathrm{E}\left[e^{Q_t}\right])$.

Corollary 2.3 *Let Q_t be an infinite server queue with periodic arrival rate $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. Then, the mean and variance of the queue are given by*

$$\begin{aligned}
\mathrm{E}\left[Q_t\right] &= Q_0 e^{-\mu t} + \frac{n\lambda}{\mu}\left(1 - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{n(a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(a_k k + b_k\mu)}{k^2 + \mu^2} \sin(kt), \qquad (2.4) \\
\mathrm{Var}\left(Q_t\right) &= Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{n\lambda}{\mu}\left(1 - e^{-\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(a_k k + b_k\mu)}{k^2 + \mu^2} \sin(kt) + \frac{n(n-1)\lambda}{2\mu}\left(1 - e^{-2\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(n-1)(2a_k\mu - b_k k)}{k^2 + 4\mu^2} \cdot \left(\cos(kt) - e^{-2\mu t}\right) \\
&\quad + \sum_{k=1}^{\infty} \frac{n(n-1)(a_k k + 2b_k\mu)}{k^2 + 4\mu^2} \sin(kt), \qquad (2.5)
\end{aligned}$$

for all time $t \geq 0$, where Q_0 is the initial number in the system.

In the remainder of this work, we will explore various modifications of this model, including general service and randomized batch sizes. The results of this subsection will serve as a cornerstone throughout much of this upcoming analysis, both supporting the underlying derivation techniques and providing the intuition for new perspectives.

2.2 The Markovian system with stationary arrival rates

Our first departure from our initial model will be modest: Instead of studying the fully Markovian, non-stationary, fixed batch size system in transient time, we will now move to addressing the stationary case, with much of our analysis focused on the system in steady state. This simplified setting will allow us to extract greater intuition from our

prior findings, which in turn will support generalization of the service distribution and randomization of the batch sizes. To begin, we find a representation of the steady-state distribution of the queue length in terms of a sum of independent, scaled Poisson random variables.

Proposition 2.4 *In the steady state, the distribution of the number in system of an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$ is*

$$Q_\infty(n) \stackrel{D}{=} \sum_{j=1}^{n} jY_j, \tag{2.6}$$

where $Y_j \sim \text{Pois}\left(\frac{\lambda}{j\mu}\right)$ are independent.

Proof From Proposition 2.1, we have that the steady-state moment-generating function of the queue is given by

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\lambda \sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta - 1)^k}{k\mu}}.$$

To satisfy our stated Poisson form, we are now left to show that $\sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta-1)^k}{k} = \sum_{k=1}^{n} \frac{e^{k\theta}-1}{k}$ for all $n \in \mathbb{Z}^+$. We proceed by induction. In the base case of $n = 1$, we have $e^\theta - 1 = e^\theta - 1$ and so we are left to show the inductive step. We now assume $\sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta-1)^k}{k} = \sum_{k=1}^{n} \frac{e^{k\theta}-1}{k}$ holds at n. Then, by the Pascal triangle identity $\binom{n}{k} = \binom{n+1}{k} - \binom{n}{k-1}$ and our inductive hypothesis, we can observe

$$\sum_{k=1}^{n} \frac{e^{k\theta} - 1}{k} = \sum_{k=1}^{n} \binom{n}{k} \frac{(e^\theta - 1)^k}{k} = \sum_{k=1}^{n} \left(\binom{n+1}{k} - \binom{n}{k-1} \right) \frac{(e^\theta - 1)^k}{k}.$$

Now, by applying the identity $\binom{n}{k-1} = \frac{k}{n+1}\binom{n+1}{k}$ and distributing the summation, we can further note that

$$\begin{aligned}
&\sum_{k=1}^{n} \left(\binom{n+1}{k} - \binom{n}{k-1} \right) \frac{(e^\theta - 1)^k}{k} \\
&= \sum_{k=1}^{n} \left(\binom{n+1}{k} - \frac{k}{n+1}\binom{n+1}{k} \right) \frac{(e^\theta - 1)^k}{k} \\
&= \sum_{k=1}^{n} \binom{n+1}{k} \frac{(e^\theta - 1)^k}{k} - \frac{\sum_{k=1}^{n} \binom{n+1}{k} (e^\theta - 1)^k}{n+1}.
\end{aligned}$$

Now, we can use the binomial theorem to see that

$$\sum_{k=1}^{n} \binom{n+1}{k} (e^{\theta}-1)^k = (e^{\theta}-1+1)^{n+1} - 1 - (e^{\theta}-1)^{n+1}$$
$$= e^{(n+1)\theta} - 1 - (e^{\theta}-1)^{n+1},$$

and so we can now simplify and find

$$\sum_{k=1}^{n} \binom{n+1}{k} \frac{(e^{\theta}-1)^k}{k} - \frac{\sum_{k=1}^{n} \binom{n+1}{k} (e^{\theta}-1)^k}{n+1}$$
$$= \sum_{k=1}^{n} \binom{n+1}{k} \frac{(e^{\theta}-1)^k}{k} + \frac{(e^{\theta}-1)^{n+1}}{n+1} - \frac{e^{(n+1)\theta}-1}{n+1}.$$

Hence, in conjunction with our initial equation, we have that

$$\sum_{k=1}^{n} \frac{e^{k\theta}-1}{k} = \sum_{k=1}^{n} \binom{n+1}{k} \frac{(e^{\theta}-1)^k}{k} + \frac{(e^{\theta}-1)^{n+1}}{n+1} - \frac{e^{(n+1)\theta}-1}{n+1},$$

and by rearranging terms we now complete the inductive approach:

$$\sum_{k=1}^{n+1} \frac{e^{k\theta}-1}{k} = \sum_{k=1}^{n+1} \binom{n+1}{k} \frac{(e^{\theta}-1)^k}{k}.$$

We can now observe that we have a moment-generating function that is a product of moment-generating functions of scaled Poisson random variables, which yields the stated result. □

While we will continue to explore the stationary arrival rate setting throughout this subsection, we note that this Poisson sum representation will be a leading inspiration in the sequel. Specifically, in Sect. 2.3 we will find intuition for this result by viewing the batch arrival queue as a collection of sub-systems.

Remark In addition to this Poisson sum representation, we can also express the steady-state MGF in terms of the truncated polylogarithm function and harmonic numbers. From the MGF of the queue length in steady state for $\theta < 0$, we can observe that

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\frac{\lambda}{\mu}\sum_{k=1}^{n} \frac{e^{k\theta}-1}{k}} = e^{\frac{\lambda}{\mu}(\mathrm{Li}(e^{\theta}, n, 1) - H_n)},$$

where we have H_n as the nth harmonic number, given by $\sum_{k=1}^{n} \frac{1}{k}$, and where the truncated polylogarithm function $\mathrm{Li}(z, n, s)$ is defined as

$$\mathrm{Li}(z, n, s) = \sum_{k=1}^{n} \frac{z^k}{k^s}.$$

This decomposition into Poisson random variables can be quite useful from a computational standpoint. It allows us to simulate the steady state quite easily since we only need to simulate n Poisson random variables instead of simulating an actual queue, which could be quite expensive. We can now observe that this construction also yields an interesting connection to both the harmonic numbers and Hermite distributions, as suggested in the remark above. To motivate the following analysis, suppose that $n = 2$. Then, the steady-state queue length has steady-state moment-generating function given by

$$\mathcal{M}_n(\theta, \infty) = e^{\frac{\lambda}{\mu}(e^{\theta}-1)+\frac{\lambda}{2\mu}(e^{2\theta}-1)}.$$

We can now observe that this MGF corresponds to a Hermite distribution with parameters $\frac{\lambda}{\mu}$ and $\frac{\lambda}{2\mu}$. This implies that the steady-state CDF of the queue at $n = 2$ is

$$P(Q_\infty(2) \le k) = e^{-\frac{3\lambda}{2\mu}} \sum_{i=0}^{\lfloor k \rfloor} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{\left(\frac{\lambda}{\mu}\right)^{i-2j} \left(\frac{\lambda}{2\mu}\right)^{j}}{(i-2j)!j!} = e^{-\frac{3\lambda}{2\mu}} \sum_{i=0}^{\lfloor k \rfloor} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{\left(\frac{\lambda}{\mu}\right)^{i-j} 2^{-j}}{(i-2j)!j!}.$$

Furthermore, the steady-state PMF of the queue length is given by

$$P(Q_\infty(2) = i) = e^{-\frac{3\lambda}{2\mu}} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{\left(\frac{\lambda}{\mu}\right)^{i-j} 2^{-j}}{(i-2j)!j!}.$$

This observation prompts us to ponder generalizations for $n \ge 3$. The term "generalized Hermite distribution" has taken on slightly varying (yet always interesting) definitions for different authors. For readers interested in the Hermite distribution and popular generalizations of it, we suggest Kemp and Kemp [16], Gupta and Jain [14], and Westcott [27]. In our setting, we note that the coefficients of $\frac{\lambda}{\mu}$ in the MGF for batch size n will be 1, $\frac{1}{2}$, $\frac{1}{3}$,..., $\frac{1}{n}$. For this reason, we think of this particular generalization of Hermite distributions to be the *harmonic Hermite distribution*. We can now note that because of this harmonic structure we can instead fully characterize the distribution simply by n and $\frac{\lambda}{\mu}$. In the following proposition, we find a useful recursion for the probability mass function of this distribution at all $n \in \mathbb{Z}^+$.

Proposition 2.5 *Let $Q_t(n)$ be an infinite server batch arrivals queue with arrival rate $\lambda > 0$, batch size $n \in \mathbb{Z}^+$, and service rate $\mu > 0$. Then, the steady-state distribution of the queue is given by the recursion*

$$\mathbb{P}(Q_\infty(n) = j) = p_j = \sum_{i=1}^{n} i p_{j-i} \frac{\lambda}{ij\mu} = \sum_{i=1}^{n} p_{j-i} \frac{\lambda}{j\mu}, \tag{2.7}$$

where $p_0 = e^{-\frac{\lambda}{\mu}H_n}$ for H_n the nth harmonic number and $p_k = 0$ for all $k < 0$. Thus, we say that $Q_\infty(n)$ follows the "harmonic Hermite distribution" with parameter n.

Proof We know from our Poisson representation of the steady-state queue length that the steady-state moment-generating function is

$$M(\theta) = \sum_{j=0}^{\infty} \mathbb{P}(Q_\infty(n) = j)\theta^j = \sum_{j=0}^{\infty} p_j \theta^j = \exp\left(\sum_{i=1}^{n} \frac{\lambda}{i\mu}\left(\theta^i - 1\right)\right).$$

If we take the logarithm of both sides, we see that we have

$$\log\left(\sum_{j=0}^{\infty} p_j \theta^j\right) = \sum_{i=1}^{n} \frac{\lambda}{i\mu}\left(\theta^i - 1\right).$$

Now we take the derivative of both sides with respect to the parameter θ, and this yields the following expression:

$$\frac{\sum_{j=1}^{\infty} j p_j \theta^{j-1}}{\sum_{j=0}^{\infty} p_j \theta^j} = \sum_{i=1}^{n} \frac{\lambda}{\mu}\theta^{i-1}.$$

By moving the denominator to the right-hand side, we have that

$$\sum_{j=1}^{\infty} j p_j \theta^{j-1} = \left(\sum_{j=0}^{\infty} p_j \theta^j\right)\left(\sum_{i=1}^{n} \frac{\lambda}{\mu}\theta^{i-1}\right).$$

Finally, by matching similar powers of θ on the left and right sides, we complete the proof. □

From the above result, we see that for the steady-state queue length $Q_\infty(n)$ we can derive the specific probabilities

$$\begin{aligned}
p_0 &= e^{-\frac{\lambda}{\mu}H_n}, \\
p_1 &= \frac{\lambda}{\mu}p_0 = \frac{\lambda}{\mu}e^{-\frac{\lambda}{\mu}H_n}, \\
p_2 &= \frac{\lambda}{2\mu}(p_0 + p_1) = \frac{\lambda}{2\mu}e^{-\frac{\lambda}{\mu}H_n} + \frac{\lambda^2}{2\mu^2}e^{-\frac{\lambda}{\mu}H_n}.
\end{aligned}$$

We can repeat this process as needed for any desired probability. From Proposition 2.4, we can observe that the mean number in the system grows linearly with the batch size, meaning that the mean of the nth harmonic Hermite distribution is

$$\mathrm{E}\left[Q_\infty(n)\right] = \sum_{j=1}^{n} j\mathrm{E}\left[Y_j\right] = \frac{n\lambda}{\mu}. \tag{2.8}$$

We can observe further that the second moment and variance are quadratic functions of n:

$$\mathrm{E}\left[Q_\infty(n)^2\right] = \mathrm{E}\left[\left(\sum_{j=1}^{n} jY_j\right)^2\right] = \frac{n(n+1)\lambda}{2\mu} + n^2\frac{\lambda^2}{\mu^2},$$

$$\mathrm{Var}[Q_\infty(n)] = \mathrm{E}\left[Q_\infty(n)^2\right] - \mathrm{E}\left[Q_\infty(n)\right]^2 = \frac{n(n+1)\lambda}{2\mu}.$$

We note that from Proposition 2.4 and the following remark, the moment-generating function of this distribution is given by

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\frac{\lambda}{\mu}\sum_{k=1}^{n}\frac{e^{k\theta}-1}{k}} = e^{\frac{\lambda}{\mu}(\mathrm{Li}(e^\theta, n, 1) - H_n)}. \tag{2.9}$$

If one is to consider this system as the batch size grows infinitely large, we can see from Eqs. (2.8) and (2.9) that the number in system will grow proportionally, tending to infinity as n does. This leads us to ponder the limiting object of the scaled number in system $\frac{Q_t(n)}{n}$ as the batch size grows.

We begin by using Eq. (2.9) with θ replaced by $\frac{\theta}{n}$ to see that the steady-state moment-generating function of this scaled queue length is

$$\lim_{t\to\infty} \mathcal{M}(\theta, t) = e^{\frac{\lambda}{\mu}\sum_{k=1}^{n}\frac{e^{\frac{k}{n}\theta}-1}{k}}. \tag{2.10}$$

Furthermore, by replacing θ with $\frac{\theta}{n}$ and $Q_0(n)$ with $\frac{Q_0(n)}{n}$ in Proposition 2.1, we can note that the transient moment-generating function for this scaled system with constant arrival rate is given by

$$\mathrm{E}\left[e^{\theta\cdot\frac{Q_t(n)}{n}}\right] \equiv \mathcal{M}_n(\theta, t) = \left(e^{-\mu t}(e^{\frac{\theta}{n}} - 1) + 1\right)^{\frac{Q_0}{n}} e^{\lambda\sum_{k=1}^{n}\binom{n}{k}\frac{\left(e^{\theta/n}-1\right)^k}{k\mu}\left(1-e^{-k\mu t}\right)}.$$

Additionally, we can also observe that the steady-state distribution of the scaled queue can also be interpreted as a sum of Poisson random variables through direction application of Proposition 2.4 or by inspection of Eq. (2.10). This representation is

$$\frac{Q_\infty(n)}{n} \overset{D}{=} \sum_{j=1}^{n}\frac{j}{n}Y_j, \tag{2.11}$$

where again $Y_j \sim \mathrm{Pois}\left(\frac{\lambda}{j\mu}\right)$.

We now consider the limit as $n \to \infty$, in which we are both sending the size of batches of arrivals to infinity while also scaling the size of the queue inversely. We can use this construction to move beyond just the mean and variance and instead explicitly state every cumulant of the scaled queue. In Proposition 2.6, we give exact expressions

of all steady-state cumulants of the scaled queue as functions of the Bernoulli numbers. Further, we find a convenient form of every cumulant of the scaled queue as the batch size grows to infinity.

Proposition 2.6 *Let $\lambda > 0$ be the arrival rate of batches of size $n \in \mathbb{Z}^+$ to an infinite server queue with exponential service rate $\mu > 0$. Then, the kth steady-state cumulant of the scaled queue $\mathcal{C}^k\left[\frac{Q_\infty(n)}{n}\right]$ is given by*

$$\mathcal{C}^k\left[\frac{Q_\infty(n)}{n}\right] = \frac{\frac{n^k}{k} + \frac{1}{2}n^{k-1} + \sum_{j=2}^{k-1}\frac{\mathrm{B}_j}{j!}(k-1)_{j-1}n^{k-j}}{n^k}, \tag{2.12}$$

where $(n)_i = \frac{n!}{(n-i)!}$ is the ith falling factorial of n and B_i is the ith Bernoulli number, which is defined as

$$B_i = \sum_{k=0}^{i}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\frac{(j+1)^i}{k+1}.$$

Moreover, we have that $\lim_{n\to\infty}\mathcal{C}^k\left[\frac{Q_\infty(n)}{n}\right] = \frac{\lambda}{k\mu}$.

Proof From our prior observation that $\frac{Q_\infty(n)}{n} \overset{D}{=} \sum_{j=1}^{n}\frac{j}{n}Y_j$, where $Y_j \sim \mathrm{Pois}\left(\frac{\lambda}{j\mu}\right)$, we have that

$$\mathcal{C}^k\left[\frac{Q_\infty(n)}{n}\right] = \mathcal{C}^k\left[\sum_{j=1}^{n}\frac{j}{n}Y_j\right] = \sum_{j=1}^{n}\mathcal{C}^k\left[\frac{j}{n}Y_j\right] = \sum_{j=1}^{n}\frac{j^k}{n^k}\mathcal{C}^k\left[Y_j\right] = \frac{\lambda}{\mu n^k}\sum_{j=1}^{n}j^{k-1},$$

from the independence of these Poisson distributions. Now, by using Faulhaber's formula as given in Knuth [17], we achieve the stated result. □

Just as we built from inherited expressions for the mean and variance to specify every cumulant in Proposition 2.6, we can also find the limit of the transient-state moment-generating function for the scaled queue given in Eq. (2.9).

Proposition 2.7 *Let Q_t be an infinite server queue with arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and exponential service rate $\mu > 0$. For $\theta \in \mathbb{R}$, let*

$$\mathcal{M}_\infty(\theta, t) = \lim_{n\to\infty}\mathrm{E}\left[e^{\frac{\theta Q_t(n)}{n}}\right].$$

Then, $\mathcal{M}_\infty(\theta, t)$ is given by

$$\mathcal{M}_\infty(\theta, t) = \begin{cases} e^{\frac{\lambda}{\mu}\left(\mathrm{Ei}(\theta)-\mathrm{Ei}(\theta e^{-\mu t})-\mu t\right)} & \text{if } \theta > 0, \\ e^{\frac{\lambda}{\mu}\left(E_1(-\theta e^{-\mu t})-E_1(-\theta)-\mu t\right)} & \text{if } \theta < 0, \\ 1 & \text{if } \theta = 0, \end{cases} \tag{2.13}$$

for all time $t \geq 0$, where the exponential integral functions $\mathrm{Ei}(x)$ *and* $E_1(x)$ *are defined by*

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-s}}{s} \mathrm{d}s, \quad E_1(x) = \int_{x}^{\infty} \frac{e^{-s}}{s} \mathrm{d}s,$$

and are real-valued for $x > 0$.

Proof While conventions may vary by application area, in this work we use the definition of the exponential integral function given by

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-s}}{s} \mathrm{d}s.$$

By taking the limit of the MGF of the scaled queue, we have that

$$\frac{\partial}{\partial t}\mathcal{M}_{\infty}(\theta, t) = \lambda\left(e^{\theta} - 1\right)\mathcal{M}_{\infty}(\theta, t) - \mu\theta\frac{\partial}{\partial \theta}\mathcal{M}_{\infty}(\theta, t),$$

with the initial solution $\mathcal{M}_{\infty}(\theta, 0) = \lim_{n\to\infty} e^{\frac{\theta Q_0}{n}} = 1$. In the same manner as the proof of Theorem 2.1, we solve the PDE of the cumulant generating function through the use of the method of characteristics. We start by establishing the characteristic ODEs:

$$\begin{aligned}
\frac{\mathrm{d}\theta}{\mathrm{d}s}(r, s) &= \mu\theta, & \theta(r, 0) &= r,\\
\frac{\mathrm{d}t}{\mathrm{d}s}(r, s) &= 1, & t(r, 0) &= 0,\\
\frac{\mathrm{d}g}{\mathrm{d}s}(r, s) &= \lambda(e^{\theta} - 1), & g(r, 0) &= 0.
\end{aligned}$$

We now solve the first two initial value problems and find

$$\begin{aligned}
\theta(r, s) = c_1(r)e^{\mu s} \quad &\Rightarrow \quad \theta(r, s) = re^{\mu s},\\
t(r, s) = s + c_2(r) \quad &\Rightarrow \quad t(r, s) = s.
\end{aligned}$$

This allows us to simplify the third characteristic equation to

$$\frac{\mathrm{d}g}{\mathrm{d}s}(r, s) = \lambda(e^{re^{\mu s}} - 1).$$

Because $\theta = re^{\mu s}$, we can note that r and θ will match in sign: $r > 0$ if and only if $\theta > 0$. If $\theta > 0$, the general solution to this ODE is

$$g(r, s) = c_3(r) + \frac{\lambda}{\mu}\left(\mathrm{Ei}(re^{\mu s}) - \mu s\right),$$

whereas if $\theta < 0$, the solution is instead

$$g(r,s) = c_3(r) - \frac{\lambda}{\mu}\left(E_1(-re^{\mu s}) + \mu s\right).$$

This follows from the fact that for $x > 0$ the exponential integral functions are such that $\mathrm{Ei}(x) = -E_1(-x) - i\pi$; that is, the real parts of $E_1(-x)$ and $-\mathrm{Ei}(x)$ are the same. Moreover, for $x > 0$ one can consider $\mathrm{Ei}(x)$ as the real part of $-E_1(-x)$. Additionally, $E_1(x)$ is real for all $x > 0$. Hence, we use each definition of the exponential integral function when appropriate. As an alternative, we could replace each of these functions with `real`$(-E_1(-x))$ to have a single expression for both positive and negative x. For a collection of facts regarding the exponential integral functions, see pp. 228–237 of Abramowitz and Stegun [1].

Now, using this we have that the corresponding solutions to the initial value problems will be

$$g(r,s) = \begin{cases} \frac{\lambda}{\mu}\left(\mathrm{Ei}(re^{\mu s}) - \mathrm{Ei}(r) - \mu s\right) & \text{if } r > 0, \\ \frac{\lambda}{\mu}\left(E_1(-r) - E_1(-re^{\mu s}) - \mu s\right) & \text{if } r < 0. \end{cases}$$

Hence, for $s = t$ and $r = \theta e^{-\mu t}$, this yields

$$G(\theta,t) = g\left(\theta e^{-\mu t}, t\right) = \begin{cases} \frac{\lambda}{\mu}\left(\mathrm{Ei}(\theta) - \mathrm{Ei}(\theta e^{-\mu t}) - \mu t\right) & \text{if } \theta > 0, \\ \frac{\lambda}{\mu}\left(E_1(-\theta e^{-\mu t}) - E_1(-\theta) - \mu t\right) & \text{if } \theta < 0. \end{cases}$$

By $\mathcal{M}_\infty(\theta,t) = e^{G_\infty(\theta,t)}$, we complete the proof. □

As a consequence, we can also give the moment-generating function in steady state.

Corollary 2.8 *The moment-generating function of the scaled number in system in steady state as $n \to \infty$ is given by*

$$\mathcal{M}_\infty(\theta) = \begin{cases} \theta^{-\frac{\lambda}{\mu}} e^{\frac{\lambda}{\mu}(\mathrm{Ei}(\theta) - \gamma)} & \text{if } \theta > 0, \\ (-\theta)^{-\frac{\lambda}{\mu}} e^{-\frac{\lambda}{\mu}(E_1(-\theta) + \gamma)} & \text{if } \theta < 0, \\ 1 & \text{if } \theta = 0, \end{cases} \tag{2.14}$$

where γ is the Euler–Mascheroni constant.

Proof From Abramowitz and Stegun [1], for $x > 0$ we can expand the exponential integral functions as

$$\mathrm{Ei}(x) = \gamma + \log(x) + \sum_{k=1}^{\infty} \frac{x^k}{kk!}, \quad E_1(x) = -\gamma - \log(x) - \sum_{k=1}^{\infty} \frac{(-x)^k}{kk!}, \tag{2.15}$$

where γ is the Euler–Mascheroni constant. By expanding $\mathrm{Ei}(\theta e^{-\mu t})$ and $E_1(-\theta e^{-\mu t})$ in the respective cases of positive and negative θ and taking the limit as $t \to \infty$, we achieve the stated result. □

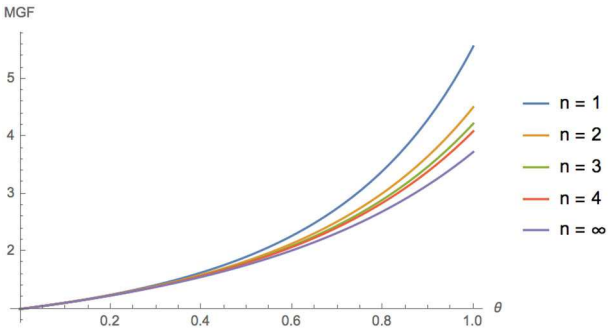

Fig. 1 Steady-state MGF of the scaled queue for increasing batch size where $\frac{\lambda}{\mu} = 1$

As a demonstration of the convergence of the steady-state moment-generating functions of the batch scaled queues to the expression given in Corollary 2.8, we plot the first four cases in comparison with the limiting scenario in Fig. 1.

While it can be argued that even in steady state the form of this moment-generating function is unfamiliar, we can still observe interesting characteristics of it. In particular, for $\theta < 0$ we can uncover a connection back to the harmonic numbers. We now discuss this in the following remark.

Remark Using Eq. (2.15), we can note that for $\theta < 0$ the steady-state moment-generating function of limit of the scaled queue can be expressed

$$M(\theta) = (-\theta)^{-\frac{\lambda}{\mu}} e^{-\frac{\lambda}{\mu}(E_1(-\theta)+\gamma)} = e^{-\frac{\lambda}{\mu}(E_1(-\theta)+\gamma+\log(-\theta))} = e^{-\frac{\lambda}{\mu}\left(-\sum_{k=1}^{\infty}\frac{\theta^k}{kk!}\right)}.$$

From Dattoli and Srivastava [4], we have that $-e^x \sum_{k=1}^{\infty} \frac{(-x)^k}{kk!}$ is an exponential generating function for the harmonic numbers. That is,

$$-e^x \sum_{k=1}^{\infty} \frac{(-x)^k}{kk!} = \sum_{n=1}^{\infty} \frac{x^n}{n!} H_n,$$

where H_n is the nth harmonic number. Thus, for $\theta < 0$ the steady-state moment-generating function of this limiting object can be further simplified to

$$M(\theta) = e^{-\frac{\lambda}{\mu}\left(-\sum_{k=1}^{\infty}\frac{\theta^k}{kk!}\right)} = e^{-\frac{\lambda}{\mu}\sum_{n=1}^{\infty} H_n e^{\theta}\frac{(-\theta)^n}{n!}} = e^{-\frac{\lambda}{\mu}\mathrm{E}[H_N]},$$

where $N \sim \mathrm{Pois}(-\theta)$.

In addition to this remark's connection of the moment-generating function and the harmonic numbers, we can also gain insight into this limiting object through Monte Carlo methods. Using Eq. (2.11), we have a simple and efficient approximate simulation method for this process through summing scaled Poisson random numbers. Furthermore, this approximation of course becomes increasingly precise as n grows. As an example of this, we give the simulated steady-state densities across different

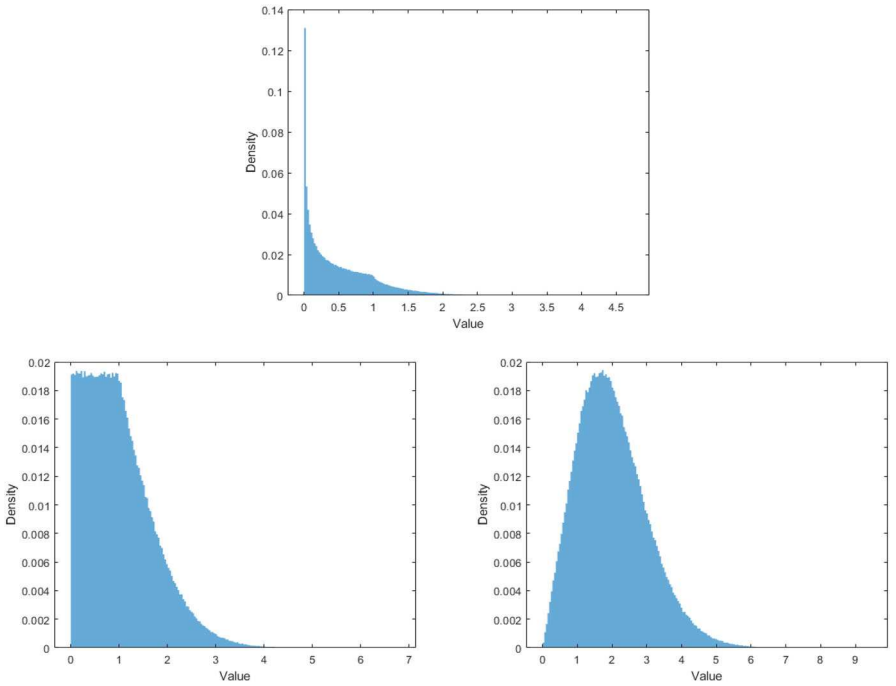

Fig. 2 Approximate steady-state density of the scaled queue limit for size where $\frac{\lambda}{\mu} = \frac{1}{2}$ (top), $\frac{\lambda}{\mu} = 1$ (left), and $\frac{\lambda}{\mu} = 2$ (right), using 1,000,000 simulation replications and $n = 2,000$

relationships of λ and μ in Fig. 2. In addition to the interesting shapes of the densities across the different settings, one can see the limiting form of the relationships given by the recursion in Proposition 2.5 in these plots. We can note that one could also calculate these through a numerical inverse Laplace transform of the steady-state moment-generating function in Corollary 2.8, although this may likely incur significantly more computational costs than the simulation procedure.

So far we have only considered exponentially distributed service. In the next subsection, we will address this and extend this Poisson sum representation of the steady-state distribution to hold for general service. We do this through viewing the n-batch size system as being composed of n sub-systems that experience single arrivals simultaneously.

2.3 Generalizing through sub-system perspectives

Because of the infinite server construction of this model, we can also interpret this system as being a network of sub-systems that also feature infinitely many servers. However, this network's mutuality is not in its services but rather in its arrivals. Specifically, in this subsection we will think of infinite server queues with batch arrivals of size n as being n infinite server queues that all receive individual arrivals simultane-

ously. From this perspective, one can quickly observe that marginally each sub-system will be distributed as a standard infinite server queue.

For example, if the batch system is the $M_t^n/M/\infty$ queue that we first considered in Sect. 2.1, then each of these sub-queues is a $M_t/M/\infty$ system. These sub-systems are coupled through the coincidence of their arrival times but otherwise operate independently from one another. To quantify the relationship between these systems, in Proposition 2.9 we derive the transient covariance between two sub-systems for a general time-varying arrival rate.

Proposition 2.9 *Let the batch arrival queue Q_t with batch size $n \in \mathbb{Z}^+$ be represented as a superposition of n infinite server single arrival queues $\{Q_{t,i} \mid 1 \leq i \leq n\}$ that all receive arrivals simultaneously and each has independent exponentially distributed service, as described above. Let $\lambda(t) > 0$ be the non-stationary rate of simultaneous arrivals, and let $\mu > 0$ be the rate of service. Then, for distinct $i, j \in \{1, \ldots, n\}$, the covariance between $Q_{t,i}$ and $Q_{t,j}$ is given by*

$$\mathrm{Cov}[Q_{t,i}, Q_{t,j}] = e^{-2\mu t} \int_0^t \lambda(s) e^{2\mu s} \mathrm{d}s, \tag{2.16}$$

for all $t \geq 0$.

Proof From Eq. (2.1), we can solve for the product moment of the two sub-systems through the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = \lambda(t) \left(\mathrm{E}\left[Q_{t,i}\right] + \mathrm{E}\left[Q_{t,j}\right] + 1\right) - 2\mu \mathrm{E}\left[Q_{t,i} Q_{t,j}\right].$$

The solution to this differential equation is given by

$$\begin{aligned} \mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = {} & Q_{0,i} Q_{0,j} e^{-2\mu t} \\ & + e^{-2\mu t} \int_0^t \lambda(s) \left(\mathrm{E}\left[Q_{s,i}\right] e^{2\mu s} + \mathrm{E}\left[Q_{s,j}\right] e^{2\mu s} + e^{2\mu s}\right) \mathrm{d}s. \end{aligned}$$

By substituting the corresponding forms of $\mathrm{E}\left[Q_{s,k}\right] = Q_{0,k} e^{-\mu s} + e^{-\mu s} \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u$ in for each of the two means, we have

$$\begin{aligned} \mathrm{E}\left[Q_{t,i} Q_{t,j}\right] = {} & Q_{0,i} Q_{0,j} e^{-2\mu t} \\ & + e^{-2\mu t} \int_0^t \lambda(s) \left(e^{2\mu s} + \left(Q_{0,i} + \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u \right) e^{\mu s} \right. \\ & \left. + \left(Q_{0,j} + \int_0^s \lambda(u) e^{\mu u} \mathrm{d}u \right) e^{\mu s} \right) \mathrm{d}s, \end{aligned}$$

and this simplifies to the following:

$$\begin{aligned}\mathrm{E}\left[Q_{t,i}Q_{t,j}\right] &= Q_{0,i}Q_{0,j}e^{-2\mu t} + e^{-2\mu t}\int_0^t \lambda(s)e^{2\mu s}\mathrm{d}s \\ &\quad + \left(Q_{0,i}+Q_{0,j}\right)e^{-2\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s \\ &\quad + 2e^{-2\mu t}\int_0^t \lambda(s)e^{\mu s}\int_0^s \lambda(u)e^{\mu u}\mathrm{d}u\mathrm{d}s.\end{aligned}$$

We can now use the fact that for a function $F : \mathbb{R}^+ \to \mathbb{R}$ defined such that $F(t) = \int_0^t f(s)\mathrm{d}s$ for a given $f(\cdot)$, integration by parts implies

$$\int_0^t f(s)F(s)\mathrm{d}s = F(t)^2 - \int_0^t F(s)f(s)\mathrm{d}s,$$

and so $\int_0^t f(s)F(s)\mathrm{d}s = \frac{F(t)^2}{2}$. This allows us to simplify to

$$\begin{aligned}\mathrm{E}\left[Q_{t,i}Q_{t,j}\right] &= Q_{0,i}Q_{0,j}e^{-2\mu t} + e^{-2\mu t}\int_0^t \lambda(s)e^{2\mu s}\mathrm{d}s \\ &\quad + \left(Q_{0,i}+Q_{0,j}\right)e^{-2\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s \\ &\quad + e^{-2\mu t}\left(\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right)^2,\end{aligned}$$

and now we turn our focus to the product of the means. Here we distribute the multiplication to find that

$$\begin{aligned}\mathrm{E}\left[Q_{t,i}\right]\mathrm{E}\left[Q_{t,j}\right] &= \left(Q_{0,i}e^{-\mu t} + e^{-\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right) \\ &\quad \left(Q_{0,j}e^{-\mu t} + e^{-\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right) \\ &= Q_{0,i}Q_{0,j}e^{-2\mu t} + (Q_{0,i}+Q_{0,j})e^{-2\mu t}\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s \\ &\quad + e^{-2\mu t}\left(\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s\right)^2,\end{aligned}$$

and by subtracting this expression from that of the product moment, we complete the proof. □

As a consequence of this, we can specify the covariance between sub-systems in the non-stationary and stationary arrival settings we have considered thus far in this report. Further, for stationary arrival rates we capitalize on simplified expressions to also give an explicit expression for the correlation coefficient between two sub-systems.

Corollary 2.10 *Let Q_t be an infinite server queue with arrival batch size $n \in \mathbb{Z}^+$ and exponential service rate $\mu > 0$. Further, let $Q_{t,k}$ for $k \in \{1, \ldots, n\}$ be infinite*

server queues with solitary arrivals and exponential service rate $\mu > 0$, so that $\sum_{k=1}^{n} Q_{t,k} = Q_t$ for all $t \geq 0$. Let $i, j \in \{1, \ldots, n\}$ be distinct. Then, if the arrival rate is given by $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, the covariance between $Q_{t,i}$ and $Q_{t,j}$ is

$$\begin{aligned}\mathrm{Cov}[Q_{t,i}, Q_{t,j}] = \frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right) \\ + \sum_{k=1}^{\infty} \frac{a_k}{k^2 + 4\mu^2}\left(2\mu\cos(kt) + k\sin(kt) - 2\mu e^{-2\mu t}\right) \\ + \sum_{k=1}^{\infty} \frac{b_k}{k^2 + 4\mu^2}\left(2\mu\sin(kt) - k\cos(kt) + ke^{-2\mu t}\right), \end{aligned} \tag{2.17}$$

and if the arrival rate is given by $\lambda > 0$, the covariance between $Q_{t,i}$ and $Q_{t,j}$ is

$$\mathrm{Cov}[Q_{t,i}, Q_{t,j}] = \frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right), \tag{2.18}$$

where all $t \geq 0$. Finally, the correlation between two sub-systems in the stationary setting can be calculated as

$$\begin{aligned}&\mathrm{Corr}[Q_{t,i}, Q_{t,j}] \\ &= \frac{\frac{\lambda}{2\mu}\left(1 - e^{-2\mu t}\right)}{\sqrt{\left(Q_{0,i}\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right)\left(Q_{0,j}\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\right)}},\end{aligned}$$

hence for stationary arrival rates, $\mathrm{Corr}[Q_{t,i}, Q_{t,j}] \to \frac{1}{2}$ as $t \to \infty$.

Thus, we find that for a fully Markovian batch arrival queue with stationary arrival rate the correlation among any two sub-systems in steady state is $\frac{1}{2}$, regardless of the arrival or service parameters. In some sense, this seems to capture a balance between the effect of arrivals and of services on an infinite server system, with the latter being independent between these systems and the former being perfectly correlated.

Now, we can pause to note that we have actually made an implicit modeling choice by separating the batch into n identical sub-systems. In this setup, we have decided to route all customers within one batch equivalently, but we are free to make other routing decisions and still maintain the n sub-systems construction. With that in mind, it seems natural to wonder whether we can uncover distributional structure of the full system if we choose our routing procedure carefully. We will now find that not only is this true, but we in fact have already seen a suggestion on what type of routing to consider.

From Proposition 2.4, we have seen that the steady-state distribution of the $M^n/M/\infty$ system is equivalent to that of $\sum_{j=1}^{n} jY_j$ where $Y_j \sim \mathrm{Pois}(\frac{\lambda}{j\mu})$ are independent. We can also note that just as the minimum of the independent sample $S_1, \ldots, S_n \sim \mathrm{Exp}(\mu)$ will be exponentially distributed with rate $n\mu$, for $S_{(i)}$ as the

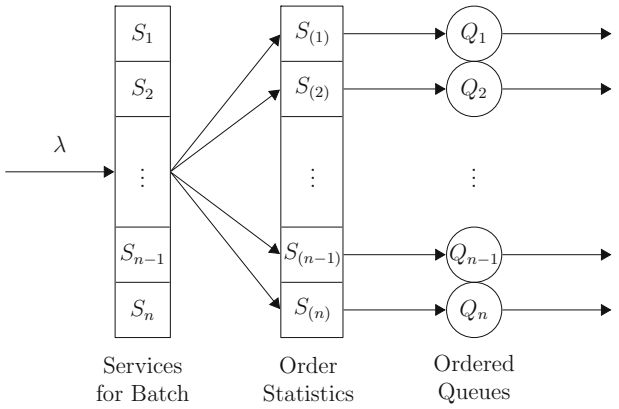

Fig. 3 Queueing diagram for the batch arrival queue with infinite servers, in which the arriving entities are routed according to the ordering of their service durations

ith ordered statistic of the n-sample we have that $S_{(i)} - S_{(i-1)} \sim \text{Exp}((n-i+1)\mu)$. Of course, the sum of these differences will telescope so that $\sum_{j=1}^{i} S_{(j)} - S_{(j-1)} = S_{(i)}$.

Taking this as inspiration, we will now assume that upon the arrival of a batch we can now know the duration of each customer's service. We then take the sub-queues to be such that the first sub-system always receives the service with the shortest duration, the second sub-system receives the second shortest service, and so on. Thus, we will route each batch of customers according to the order statistics within each batch. For reference, we visualize this sub-system construction in Fig. 3.

We can note that while the covariance structure we explored in Proposition 2.9 and Corollary 2.10 does not apply for this new routing, the sub-systems are certainly still correlated. Due to the order-statistic structuring of the service in each queue, we can note that now both the arrival processes and the service distributions will be dependent. However, we can in fact use our understanding of this dependence to not only understand how these systems relate to one another, but also to interpret how they form the structure of the full batch system as a whole. In this way, we will now consider a $M^n/G/\infty$ system. As follows in Theorem 2.11, we will find that the order-statistic-routing inspiration we have used from Proposition 2.4 leads us to a generalized Poisson sum result for general service distributions.

Theorem 2.11 *Let $Q_t(n)$ be an $M^n/G/\infty$ queue. That is, let $Q_t(n)$ be an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batch size $n \in \mathbb{Z}^+$, and general service distribution G. Then, the steady-state distribution of the number in system $Q_\infty(n)$ is*

$$Q_\infty(n) \stackrel{D}{=} \sum_{j=1}^{n} (n-j+1) Y_j, \tag{2.19}$$

where $Y_j \sim \text{Pois}\left(\lambda \text{E}\left[S_{(j)} - S_{(j-1)}\right]\right)$ are independent, with $S_{(1)} \leq \cdots \leq S_{(n)}$ as order statistics of the distribution G and with $S_{(0)} = 0$.

Proof As we have discussed in the paragraphs preceding this statement, we will consider the full queueing system as being composed of n infinite server sub-systems to which we route the arriving customers in each batch. That is, let $Q_1,\ldots, Q_n$ be infinite server queues for which we will consider the steady-state behavior. Upon the arrival of a batch, we order the customers according to the duration of their service. Then, we send the customer with the earliest service completion to Q_1, the customer with the second earliest to Q_2, and so on.

When viewing each sub-system on its own, we see that Q_j is an infinite server queue with single arrivals according to a Poisson process with rate λ and service distribution matching that of $S_{(j)}$, the jth order statistic of G. Thus, we can see that in steady state $Q_j \sim \text{Pois}\left(\lambda \text{E}\left[S_{(j)}\right]\right)$ through the literature for $M/G/\infty$ queues, such as in [9]. While we can further observe that $Q_\infty(n) = \sum_{j=1}^{n} Q_j$, we must take care in re-assembling the sub-queues. In particular, we can note that $S_{(j)}$ shares a similar structure with $S_{(j-1)}$. Each order statistic can be viewed as a construction of the gaps between the lower-ordered quantities:

$$S_{(j)} = \sum_{k=1}^{j} S_{(k)} - S_{(k-1)}.$$

Thus, from the thinning property of the Poisson distribution and the linearity of expectation, we can write the distribution of Q_j as a sum of independent Poisson RVs, as given by

$$Q_j \sim \sum_{k=1}^{j} \text{Pois}\left(\lambda \text{E}\left[S_{(k)} - S_{(k-1)}\right]\right).$$

We can note further that $j-1$ of the Poisson components of Q_j are the exact components of Q_{j-1}, with $j-2$ of these components also shared with Q_{j-2}, $j-3$ with Q_{j-3}, and so on. Then, we see that the Poisson component $\text{Pois}\left(\lambda \text{E}\left[S_{(j)} - S_{(j-1)}\right]\right)$ is repeated $n-j+1$ times across this sub-system construction of $Q_\infty(n)$, as it appears in each of the Poisson sum expressions of Q_j, $Q_{j+1},\ldots, Q_{n-1}$, and Q_n. Assembling $Q_\infty(n)$ in this way, we complete the proof. □

One can also note that this order-statistic sub-system structure also provides some motivation for the occurrence of the harmonic numbers that we observed in Sect. 2.2 when viewing the largest order statistic, which we discuss now in the following remark.

Remark For $S_i \sim \text{Exp}(\mu)$, one can see through the telescoping construction of the order statistics that

$$\text{E}\left[S_{(n)}\right] = \sum_{i=1}^{n} \text{E}\left[S_{(i)} - S_{(i-1)}\right] = \sum_{i=1}^{n} \frac{1}{(n-i+1)\mu} = \frac{1}{\mu} H_n.$$

Now, throughout this section we have operated on the assumption that the batch size is a known, fixed constant. While this may be applicable in some settings, there are certainly many settings where the batch size is unknown and varies between arrivals.

Thus, we address this in Sect. 3 and find that many of the results we have shown thus far can be replicated for models with random batch size.

3 Random batch sizes

We will now consider systems in which the size of an arriving batch is drawn from an independent and identically distributed sequence of random variables. We will treat the distribution of the batch size as general throughout this work. As in Sect. 2, we assume that the times of arrivals are given by a Poisson process, with consideration given to both stationary and non-stationary rates, and we will again analyze both exponential and general service distributions.

We start by giving the mean and variance of the system for time-varying arrival rates with exponential service in Sect. 3.1. Then, in Sect. 3.2 we give three limiting results for the stationary arrivals model: a batch scaling, a fluid limit, and a diffusion limit. Finally, in Sect. 3.3 we extend the Poisson sum construction of the steady-state distribution to hold for random batch sizes.

One can note that many of these results are generalizations or extensions of findings from Sect. 2, thus implying them as a special case and perhaps even building a case for them to be omitted. Rather, these findings are critical to the narrative of this report. As we will see, the results for fixed batch size provide the analytic foundations and conceptual inspirations from which we derive much of the analysis in this section.

3.1 Mean and variance for time-varying, Markovian case

To begin our exploration into random batch size systems, we will start simply: we will look at a fully Markovian (albeit time-varying) system and find the mean and variance, using conditional probability and our results from Sect. 2. Specifically, in this subsection we will consider the $M_t^N/M/\infty$ queue. That is, take an infinite server queue with a general non-stationary arrival rate. We suppose that arrivals occur in batches of random size from a sequence of independent and identically distributed random variables. Furthermore, we suppose that service is exponentially distributed. We now give the mean and variance of this system in Proposition 3.1.

Proposition 3.1 *Let Q_t be an infinite server queue with finite, time-varying arrival rate $\lambda(t) > 0$, exponential service rate $\mu > 0$, and random batch size with finite mean $\mathrm{E}[N]$. Then, the mean number in system is given by*

$$\mathrm{E}[Q_t] = Q_0 e^{-\mu t} + e^{-\mu t}\mathrm{E}[N]\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s, \tag{3.1}$$

for all $t \geq 0$. If the batch size distribution has finite second moment $\mathrm{E}\left[N^2\right]$, the variance of the number in system is given by

$$\begin{aligned}\mathrm{Var}\left(Q_t\right) = Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + e^{-2\mu t}\left(\mathrm{E}\left[N^2\right] - \mathrm{E}\left[N\right]\right)\int_0^t \lambda(s)e^{2\mu s}\mathrm{d}s \\ + e^{-\mu t}\mathrm{E}\left[N\right]\int_0^t \lambda(s)e^{\mu s}\mathrm{d}s, \end{aligned} \tag{3.2}$$

again for all $t \geq 0$.

Proof Using the infinitesimal generator method, we have that the first and second moments of this system are given by the solutions to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] &= \lambda(t)\mathrm{E}\left[N_1\right] - \mu\mathrm{E}\left[Q_t\right], \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^2\right] &= \lambda(t)\left(2\mathrm{E}\left[Q_t\right]\mathrm{E}\left[N_1\right] + \mathrm{E}\left[N_1^2\right]\right) - 2\mu\mathrm{E}\left[Q_t^2\right] + \mu\mathrm{E}\left[Q_t\right],\end{aligned}$$

where $\{N_i \mid i \in \mathbb{Z}^+\}$ are the i.i.d. batch sizes that are also independent of the queue. Through noting that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Var}\left(Q_t\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t^2\right] - 2\mathrm{E}\left[Q_t\right]\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[Q_t\right] \\ &= \lambda(t)\mathrm{E}\left[N_1^2\right] + \mu\mathrm{E}\left[Q_t\right] - 2\mu\mathrm{Var}\left(Q_t\right),\end{aligned}$$

we can solve for the stated results. □

In addition to providing a direct comparison to the fixed batch size case in conjunction with Corollary 2.3, Proposition 3.1 also provides a building block for the remainder of this section. In particular, in the following subsection we will develop a series of limiting results for this queueing system, including fluid and diffusion limits. In those cases, we will use this result for added interpretation. To expedite comparison in cases of stationary arrival rates, we now give the mean and variance for such systems in Corollary 3.2. Additionally, to also facilitate comparison to Corollary 2.3, we provide expressions for periodic arrival rates in Corollary 3.3.

Corollary 3.2 *Let* Q_t *be an infinite server queue with stationary arrival rate* $\lambda > 0$*, exponential service rate* $\mu > 0$*, and random batch size with mean* $\mathrm{E}\left[N\right]$*. Then, the mean number in system is given by*

$$\mathrm{E}\left[Q_t\right] = Q_0 e^{-\mu t} + \frac{\lambda\mathrm{E}\left[N\right]}{\mu}\left(1 - e^{-\mu t}\right), \tag{3.3}$$

for all $t \geq 0$*. If the batch size distribution has finite second moment* $\mathrm{E}\left[N^2\right]$*, the variance of the number in system is given by*

$$\begin{aligned}\mathrm{Var}\left(Q_t\right) = Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda\mathrm{E}\left[N\right]}{\mu}\left(1 - e^{-\mu t}\right) \\ + \frac{\lambda}{2\mu}\left(\mathrm{E}\left[N^2\right] - \mathrm{E}\left[N\right]\right)\left(1 - e^{-2\mu t}\right),\end{aligned} \tag{3.4}$$

again for all $t \geq 0$.

Corollary 3.3 *Let Q_t be an infinite server queue with periodic arrival rate $\lambda + \sum_{k=1}^{\infty} a_k \cos(kt) + b_k \sin(kt) > 0$, exponential service rate $\mu > 0$, and random batch size with finite mean $\mathrm{E}[N]$. Then, the mean number in system is given by*

$$\mathrm{E}[Q_t] = Q_0 e^{-\mu t} + \frac{\lambda \mathrm{E}[N]}{\mu}\left(1 - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) + \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k k + b_k \mu)}{k^2 + \mu^2} \sin(kt), \tag{3.5}$$

for all $t \geq 0$. If the batch size distribution has finite second moment $\mathrm{E}\left[N^2\right]$, the variance of the number in system is given by

$$\begin{aligned}\mathrm{Var}(Q_t) = {} & Q_0\left(e^{-\mu t} - e^{-2\mu t}\right) + \frac{\lambda \mathrm{E}[N]}{\mu}\left(1 - e^{-\mu t}\right) \\ & + \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k\mu - b_k k)}{k^2 + \mu^2}\left(\cos(kt) - e^{-\mu t}\right) \\ & + \sum_{k=1}^{\infty} \frac{\mathrm{E}[N](a_k k + b_k \mu)}{k^2 + \mu^2} \sin(kt) \\ & + \frac{\lambda}{2\mu}\left(\mathrm{E}\left[N^2\right] - \mathrm{E}[N]\right)\left(1 - e^{-2\mu t}\right) \\ & + \left(\mathrm{E}\left[N^2\right] - \mathrm{E}[N]\right)\left(\sum_{k=1}^{\infty} \frac{2a_k\mu - b_k k}{k^2 + 4\mu^2}\left(\cos(kt) - e^{-2\mu t}\right)\right. \\ & \left. + \sum_{k=1}^{\infty} \frac{a_k k + 2b_k \mu}{k^2 + 4\mu^2} \sin(kt)\right),\end{aligned} \tag{3.6}$$

again for all $t \geq 0$.

3.2 Limiting results for stationary arrival rates

We will now focus on systems with stationary arrival rates throughout the analysis in this subsection. In doing so, we derive limit theorems for various scalings of this process. To begin, we show a brief technical lemma for the limit of nonnegative random variables that can be represented as sums of independent and identically distributed random variables.

Lemma 3.4 *Let $X(n)$ be any random variable such that $X(n) = \sum_{k=1}^{n} Y_k$, where Y_k are i.i.d. nonnegative, discrete random variables. Then, the moment-generating function of $X(n)$ is such that*

$$\mathrm{E}\left[e^{\frac{\theta X(n)}{n}}\right] \to e^{\mathrm{E}[Y_1]\theta}$$

as $n \to \infty$.

Proof By the strong law of large numbers, we have that

$$\lim_{n\to\infty} \frac{X(n)}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} Y_k \overset{\text{a.s.}}{=} \mathrm{E}\left[Y_1\right],$$

and this implies convergence in distribution, which is equivalent to convergence of moment-generating functions. □

We can note that this condition is a weaker form of infinite divisibility. Thus, in addition to holding for any infinitely divisible and nonnegative random variables such as the Poisson and negative binomial distributions, Lemma 3.4 also holds for some distributions that are not infinitely divisible, such as the binomial. Using this lemma, we can now find our first limit theorem for random batch sizes, a batch scaling result akin to Proposition 2.7.

Theorem 3.5 *For $n \in \mathbb{Z}^+$, let $Q_t(n)$ be an infinite server queue with batch arrivals where the batch size is drawn from the i.i.d. sequence $\{N_i(n) \mid i \in \mathbb{Z}^+\}$. Let $\lambda > 0$ be the arrival rate, and let $\mu > 0$ be the rate of exponentially distributed service. Suppose that for any i and n there is a sequence of i.i.d. nonnegative, discrete random variables $\{B_k \mid k \in \mathbb{Z}^+\}$ such that $N_i(n) = \sum_{k=1}^{n} B_k$. Then, the limiting moment-generating function of the batch-scaled object*

$$\lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta}{n} Q_t(n)}\right] = \begin{cases} e^{\frac{\lambda}{\mu}\left(\mathrm{Ei}(\theta \mathrm{E}[B_1]) - \mathrm{Ei}\left(\theta \mathrm{E}[B_1] e^{-\mu t}\right) - \mu t\right)} & \text{if } \theta > 0, \\ e^{\frac{\lambda}{\mu}\left(E_1(-\theta \mathrm{E}[B_1] e^{-\mu t}) - E_1(-\theta \mathrm{E}[B_1]) - \mu t\right)} & \text{if } \theta < 0, \\ 1 & \text{if } \theta = 0, \end{cases} \tag{3.7}$$

for all $t \geq 0$.

Proof Because this system is Markovian, we can calculate the time derivative of the moment-generating function for a given n as

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{E}\left[e^{\frac{\theta}{n} Q_t(n)}\right] &= \mathrm{E}\left[\lambda \left(e^{\frac{\theta}{n} N_1(n)} - 1\right) e^{\frac{\theta}{n} Q_t(n)} + \mu Q_t(n) \left(e^{-\frac{\theta}{n}} - 1\right) e^{\frac{\theta}{n} Q_t(n)}\right] \\ &= \lambda \left(\mathrm{E}\left[e^{\frac{\theta}{n} N_1(n)}\right] - 1\right) \mathrm{E}\left[e^{\frac{\theta}{n} Q_t(n)}\right] \\ &\quad + n\mu \left(e^{-\frac{\theta}{n}} - 1\right) \mathrm{E}\left[\frac{Q_t(n)}{n} e^{\frac{\theta}{n} Q_t(n)}\right]. \end{aligned}$$

This can then be re-expressed in partial differential equation form as

$$\frac{\partial \mathcal{M}^n(\theta, t)}{\partial t} = \lambda \left(\mathrm{E}\left[e^{\frac{\theta}{n} N_1(n)}\right] - 1\right) \mathcal{M}^n(\theta, t) + n\mu \left(e^{-\frac{\theta}{n}} - 1\right) \frac{\partial \mathcal{M}^n(\theta, t)}{\partial \theta},$$

where $\mathcal{M}^n(\theta, t) = \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right]$. Now, through Lemma 3.4, we see that the limit of this partial differential equation is given by

$$\frac{\partial \mathcal{M}^\infty(\theta, t)}{\partial t} = \lambda\left(e^{\theta \mathrm{E}[B_1]} - 1\right)\mathcal{M}^\infty(\theta, t) - \mu\theta\frac{\partial \mathcal{M}^\infty(\theta, t)}{\partial \theta}.$$

We achieve the stated result through a straightforward update of the method of characteristics approach in Proposition 2.7. □

We can note that a similar batch scaling of infinite server queues is discussed in de Graaf et al. [7], in which the authors show that the limiting process can be interpreted as a shot noise process. However, that work considers a different class of batch size distributions, as the authors define their batch size distribution in terms of the distribution of the marks through use of a ceiling rounding function. In this way, that paper is more oriented around the distribution of the marks in the shot noise process rather than the size of the batches.

From this result, we can identify a relationship between the moment-generating functions of the deterministic and random batch size queues under batch scalings. Let $\mathcal{M}_n^\infty(\theta, t)$ be the limiting moment-generating function of the fixed batch size queue as given in Proposition 2.7, and let $\mathcal{M}_N^\infty(\theta, t)$ be the same for the random batch size queue as we have now seen in Theorem 3.5. Then, we can observe that

$$\mathcal{M}_N^\infty(\theta, t) = \mathcal{M}_n^\infty(\theta \mathrm{E}[B_1], t),$$

whenever the distribution of the random batch sizes meets the "finite divisibility" condition as described in Lemma 3.4. The relationship between these limiting objects provides a direct comparison between the two different batch types.

As two additional limiting results, we now provide fluid and diffusion limits for scaling the arrival rate in Theorems 3.6 and 3.7, respectively. We did not give fluid or diffusion limits for the deterministic batch cases in Sect. 2, so these two limits are built from scratch within this section. Although we did not develop such limits explicitly for the $M^n/M/\infty$ system, we will find that these limits can still be used to draw comparisons between this system and the $M^N/M/\infty$ queue simply by treating the random batch size as deterministically distributed. We now begin with the fluid limit.

Theorem 3.6 *For $n \in \mathbb{Z}^+$, let $Q_t(n)$ be an infinite server queue with batch arrivals where the batch size is drawn from the i.i.d. sequence $\{N_i \mid i \in \mathbb{Z}^+\}$. Let $n\lambda > 0$ be the arrival rate and let $\mu > 0$ be the rate of exponentially distributed service. Then, the limiting moment-generating function of the fluid scaling is given by*

$$\lim_{n\to\infty} \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] = e^{\frac{\lambda \mathrm{E}[N_1]\theta}{\mu}(1-e^{-\mu t})+Q_0\theta e^{-\mu t}}, \tag{3.8}$$

for all $t \geq 0$.

Proof We begin with the infinitesimal generator equation for the time derivative of the moment-generating function at a given n. This is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] &= \mathrm{E}\left[n\lambda\left(e^{\frac{\theta N_1}{n}}-1\right)e^{\frac{\theta}{n}Q_t(n)} + \mu Q_t(n)\left(e^{-\frac{\theta}{n}}-1\right)e^{\frac{\theta}{n}Q_t(n)}\right] \\ &= n\lambda\left(\mathrm{E}\left[e^{\frac{\theta N_1}{n}}\right]-1\right)\mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right] \\ &\quad + \mu n\left(e^{-\frac{\theta}{n}}-1\right)\mathrm{E}\left[\frac{Q_t(n)}{n}e^{\frac{\theta}{n}Q_t(n)}\right],\end{aligned}$$

which can also be expressed in partial differential equation form as

$$\frac{\partial\mathcal{M}^n(\theta,t)}{\partial t} = n\lambda\left(\mathrm{E}\left[e^{\frac{\theta N_1}{n}}\right]-1\right)\mathcal{M}^n(\theta,t) + \mu n\left(e^{-\frac{\theta}{n}}-1\right)\frac{\partial\mathcal{M}^n(\theta,t)}{\partial\theta},$$

where $M^n(\theta,t) = \mathrm{E}\left[e^{\frac{\theta}{n}Q_t(n)}\right]$. By a Taylor expansion of the function $e^{\frac{\theta N_1}{n}}$ and by taking the limit as $n\to\infty$, we can see that this yields

$$\frac{\partial\mathcal{M}^\infty(\theta,t)}{\partial t} = \lambda\theta\mathrm{E}\left[N_1\right]\mathcal{M}^\infty(\theta,t) - \mu\theta\frac{\partial\mathcal{M}^\infty(\theta,t)}{\partial\theta}.$$

Using the initial condition $\mathcal{M}^\infty(\theta,0) = e^{Q_0\theta}$, we can see that the solution to this partial differential equation will be

$$\mathcal{M}^\infty(\theta,t) = e^{\frac{\lambda\mathrm{E}[N_1]\theta}{\mu}(1-e^{-\mu t})+Q_0\theta e^{-\mu t}},$$

and this completes the proof. □

From Corollary 3.2, we see that the mean number in system for the $M^N/M/\infty$ queue is $\frac{\lambda\mathrm{E}[N_1]}{\mu}\left(1-e^{-\mu t}\right)+Q_0e^{-\mu t}$. Thus, this fluid limit moment-generating function is equivalent to $e^{\theta\mathrm{E}[Q_t]}$ for all $t\geq 0$ and all θ, showing that the fluid limit converges to the mean. We now find a connection to both the mean and the variance through a diffusion limit in Theorem 3.7.

Theorem 3.7 *For $n\in\mathbb{Z}^+$, let $Q_t(n)$ be an infinite server queue with batch arrivals where the batch size is drawn from the i.i.d. sequence $\{N_i \mid i\in\mathbb{Z}^+\}$. Let $n\lambda>0$ be the arrival rate and let $\mu>0$ be the rate of exponentially distributed service. Then, the limiting moment-generating function of the diffusion scaling is given by*

$$\lim_{n\to\infty}\mathrm{E}\left[e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right] = e^{\frac{\lambda\theta^2}{4\mu}(\mathrm{E}[N_1]+\mathrm{E}[N_1^2])(1-e^{-\mu t})+\theta Q_0e^{-\mu t}}, \tag{3.9}$$

which gives a steady-state approximation of $X\sim\mathrm{Norm}\left(\frac{\lambda\mathrm{E}[N_1]}{\mu},\frac{\lambda}{2\mu}\left(\mathrm{E}\left[N_1\right]+\mathrm{E}\left[N_1^2\right]\right)\right)$.

Proof Through use of the infinitesimal generator, we have that the time derivative of the moment-generating function for a given n can be expressed as

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\left[e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right] \\
&\quad=\mathrm{E}\left[n\lambda\left(e^{\frac{\theta N_1}{\sqrt{n}}}-1\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right. \\
&\qquad\left.+\mu Q_t(n)\left(e^{-\frac{\theta}{\sqrt{n}}}-1\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right] \\
&\quad=\mathrm{E}\left[\sqrt{n}\lambda\left(\theta N_1+\frac{\theta^2N_1^2}{2\sqrt{n}}+\mathrm{O}\left(\frac{\theta^3N_1^3}{6n}\right)\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right] \\
&\qquad+\mathrm{E}\left[\mu\sqrt{n}\left(\frac{Q_t(n)}{\sqrt{n}}-\frac{n\lambda\mathrm{E}[N_1]}{\sqrt{n}\mu}+\frac{n\lambda\mathrm{E}[N_1]}{\sqrt{n}\mu}\right)\left(e^{-\frac{\theta}{\sqrt{n}}}-1\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right],
\end{aligned}$$

where here we have used a Taylor expansion of the function $e^{\frac{\theta N_1}{\sqrt{n}}}$. Now, for $\mathcal{M}^n(\theta,t)=\mathrm{E}\left[e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right]$, this equation can be written as a partial differential equation as follows:

$$\begin{aligned}
\frac{\partial\mathcal{M}^n(\theta,t)}{\partial t}&=\lambda\theta\sqrt{n}\mathrm{E}[N_1]\mathcal{M}^n(\theta,t) \\
&\quad+\frac{\lambda\theta^2}{2}\mathrm{E}\left[N_1^2\right]\mathcal{M}^n(\theta,t)+\sqrt{n}\lambda\mathrm{E}\left[\mathrm{O}\left(\frac{\theta^3N_1^3}{6n}\right)e^{\frac{\theta}{\sqrt{n}}\left(Q_t(n)-\frac{n\lambda\mathrm{E}[N_1]}{\mu}\right)}\right] \\
&\quad+\sqrt{n}\mu\left(e^{-\frac{\theta}{\sqrt{n}}}-1\right)\frac{\partial\mathcal{M}^n(\theta,t)}{\partial\theta}+n\lambda\mathrm{E}[N_1]\left(e^{-\frac{\theta}{\sqrt{n}}}-1\right)\mathcal{M}^n(\theta,t).
\end{aligned}$$

As we take $n\to\infty$, this PDE becomes

$$\frac{\partial\mathcal{M}^\infty(\theta,t)}{\partial t}=\frac{\lambda\theta^2}{2}\mathrm{E}[N_1]\mathcal{M}^\infty(\theta,t)+\frac{\lambda\theta^2}{2}\mathrm{E}\left[N_1^2\right]\mathcal{M}^\infty(\theta,t)-\mu\theta\frac{\partial\mathcal{M}^\infty(\theta,t)}{\partial\theta},$$

and this yields a solution of

$$\mathcal{M}^\infty(\theta,t)=e^{\frac{\lambda\theta^2}{4\mu}\left(\mathrm{E}[N_1]+\mathrm{E}\left[N_1^2\right]\right)\left(1-e^{-\mu t}\right)+\theta Q_0e^{-\mu t}}.$$

To observe the steady-state distribution, we take the limit as $t\to\infty$ and observe that this produces the moment-generating function for a Gaussian. □

By comparison with the limits of the expressions in Corollary 3.2 as $t\to\infty$, we can now observe that this steady-state approximation is equal in mean and variance to the steady-state queue.

3.3 Extending the order-statistic sub-systems

In Sect. 2.3, we found that the steady-state distributions of infinite server queues with fixed batch size and general service can be written as a sum of scaled Poisson random variables, providing a succinct interpretation of the process and an efficient simulation procedure for approximate calculations. The underlying observation that supported this approach was that we can think of an infinite server queue with batch arrivals as a collection of infinite server queues with solitary arrivals that occur simultaneously. Using the thinning property of Poisson processes, we now extend this result to queues with random batch sizes and general service.

Theorem 3.8 *Let Q_t be a $M^N/G/\infty$ queue. That is, let Q_t an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batches of random size according to the i.i.d. sequence of nonnegative integer-valued random variables $\{N_i \mid i \in \mathbb{Z}^+\}$, and general service distribution G. Then, the steady-state distribution of the number in system Q_∞ is*

$$Q_\infty \stackrel{D}{=} \sum_{n=1}^{\infty} \sum_{j=1}^{n} (n-j+1) Y_{j,n}, \tag{3.10}$$

where $Y_{j,n} \sim \text{Pois}\left(\lambda p_n \text{E}\left[S_{(j,n)} - S_{(j-1,n)}\right]\right)$ are independent, with $S_{(1,n)} \leq \cdots \leq S_{(n,n)}$ the order statistics of the distribution G when $N_i = n$, where $S_{(0,n)} = 0$ for all n and $p_n = \text{P}(N_1 = n)$.

Proof To begin, we suppose that there is some $m \in \mathbb{Z}^+$ such that $\text{P}(N_i \in \{0, \ldots, m\}) = 1$. Then, using the thinning property of Poisson processes, we separate the arrival process into m arrival streams where the nth arrival rate is λp_n. Then, by Theorem 2.11 the steady-state distribution of the number in system from the nth stream is

$$\sum_{j=1}^{n} (n-j+1) \text{Pois}\left(\lambda p_n \text{E}\left[S_{(j,n)} - S_{(j-1,n)}\right]\right).$$

Then, since the m thinned Poisson streams are independent, we have that the full combined system will be distributed as

$$\sum_{n=1}^{m} \sum_{j=1}^{n} (n-j+1) \text{Pois}\left(\lambda p_n \text{E}\left[S_{(j,n)} - S_{(j-1,n)}\right]\right).$$

Through taking the limit as $m \to \infty$, we achieve the stated result. □

We can note that Theorem 3.8 also provides a method for approximate empirical calculation through simulation. This representation can also be simplified if more information is known about the distribution of the batch size or of the service, or both. As an example, we give the distribution for the fully Markovian system in the following corollary.

Corollary 3.9 *Let Q_t be a $M^N/M/\infty$ queue. That is, let Q_t an infinite server queue with stationary arrival rate $\lambda > 0$, arrival batches of random size according to the i.i.d. sequence of nonnegative integer-valued random variables $\{N_i \mid i \in \mathbb{Z}^+\}$, and exponentially distributed service at rate $\mu > 0$. Then, the steady-state distribution of the number in system Q_∞ is*

$$Q_\infty \stackrel{D}{=} \sum_{j=1}^{\infty} jY_j, \tag{3.11}$$

where $Y_j \sim \text{Pois}\left(\frac{\lambda}{j\mu}\bar{F}_N(j)\right)$ are independent, where $\bar{F}_N(j) = \text{P}(N_1 \geq j)$.

One can note that the moment-generating function for this system in steady state is

$$\text{E}\left[e^{\theta Q_\infty}\right] = e^{\sum_{j=1}^{\infty} \frac{\lambda}{j\mu}\bar{F}_N(j)\left(e^{j\theta}-1\right)},$$

and that this also admits a connection to the generalized Hermite distributions we discussed in Sect. 2.2. In particular, this generalized Hermite distribution can be characterized by $\frac{\lambda}{\mu}$, which is again the mean of the distribution, and the complementary cumulative distribution function of the batch size distribution, which dictates the coefficients at each j. For this reason, it may be possible that the steady-state distribution of the queue may be simplified even further for particular batch size distributions.

Because Theorem 3.8 is again built upon an order-statistic sub-queue perspective, it is natural to wonder how the distribution of the batch size would affect those subsystems. In particular, we now consider the following scenario: suppose that the batch size is bounded by some constant, say k, and that we have k sub-systems. For each arriving batch, the customer with the shortest service duration will go to the first subsystem, the second shortest to the second sub-system, and so on, but only up to the number that have just arrived: if this batch is of size $k-1$, the kth sub-queue will not receive an arrival. In this way, the ith sub-queue represents the number in system that were the ith smallest in their batch. In the following proposition, we find the conditions on the batch size distribution under which the distributions of the sub-queues will be equivalent.

Proposition 3.10 *Consider a $M^B/G/\infty$ queueing system in which the distribution of B has support on $\{1, \ldots, k\}$. Let $\phi \in [0,1]^{k-1}$ be such that $\phi_i = \text{P}(B = i)$, yielding $\text{P}(B = k) = 1 - \sum_{i=1}^{k-1} \phi_i$. Let $S_{(i,j)}$ be the ith order statistics in a sample of size j from the service distribution. Furthermore, let Q_i be the steady-state number in system of an infinite server sub-queue to which the customer with the ith smallest service duration in an arriving batch will be routed whenever there are at least i customers in the batch. Let $M \in \mathbb{R}^{k-1\times k-1}$ be an upper triangular matrix such that*

$$M_{i,j} = \frac{\text{E}\left[S_{(i,j)}\right]}{\text{E}\left[S_{(k,k)}\right] - \text{E}\left[S_{(i,k)}\right]},$$

for $i \leq j$, and $M_{i,j} = 0$ otherwise. For $\mathbf{v} \in \mathbb{R}^{k-1}$ as the all-ones column vector, if ϕ is such that

$$\mathbf{v} = \left(M + \mathbf{v}\mathbf{v}^{\mathrm{T}}\right)\phi,$$

then $Q_i \overset{D}{=} Q_j$ for all sub-queues i and j. Moreover, if $1 + \mathbf{v}^{\mathrm{T}} M^{-1}\mathbf{v} \neq 0$, then the distributions of the sub-queues are equivalent if and only if $\phi = (M + \mathbf{v}\mathbf{v}^{\mathrm{T}})^{-1}\mathbf{v}$.

Proof We start by considering the mean of each queue and solving for ϕ such that all the means are equal. Let λ be the batch arrival rate. Then, the mean of Q_i is

$$\mathrm{E}\left[Q_i\right] = \sum_{j=i}^{k-1} \lambda \phi_j \mathrm{E}\left[S_{(i,j)}\right] + \lambda \left(1 - \sum_{j=1}^{k-1} \phi_j\right) \mathrm{E}\left[S_{(i,k)}\right],$$

as entities only arrive to Q_i when $B \geq i$. We can note that for Q_k this is

$$\mathrm{E}\left[Q_k\right] = \lambda \left(1 - \sum_{j=1}^{k-1} \phi_j\right) \mathrm{E}\left[S_{(k,k)}\right].$$

Then, we can see that all the queue means will be equal if $\mathrm{E}\left[Q_i\right] = \mathrm{E}\left[Q_k\right]$ for all i. Thus, we want to solve for ϕ such that

$$0 = \sum_{j=i}^{k-1} \lambda \phi_j \mathrm{E}\left[S_{(i,j)}\right] + \lambda \left(1 - \sum_{j=1}^{k-1} \phi_j\right) \mathrm{E}\left[S_{(i,k)}\right] - \lambda \left(1 - \sum_{j=1}^{k-1} \phi_j\right) \mathrm{E}\left[S_{(k,k)}\right],$$

for all i. Rearranging this equation and dividing by $\lambda\left(\mathrm{E}\left[S_{(k,k)}\right] - \mathrm{E}\left[S_{(i,k)}\right]\right)$, we receive

$$\sum_{j=i}^{k-1} \frac{\mathrm{E}\left[S_{(i,j)}\right]}{\mathrm{E}\left[S_{(k,k)}\right] - \mathrm{E}\left[S_{(i,k)}\right]} \phi_j + \sum_{j=1}^{k-1} \phi_j = 1.$$

We can now observe that this forms the linear system $(M + \mathbf{v}\mathbf{v}^{\mathrm{T}})\phi = \mathbf{v}$, and so we have shown that if ϕ satisfies this system then the means of the sub-queues will be equal. We can note moreover that $M + \mathbf{v}\mathbf{v}^{\mathrm{T}}$ is a rank one update of the matrix M. Thus, it is known that $M + \mathbf{v}\mathbf{v}^{\mathrm{T}}$ will be invertible if $1 + \mathbf{v}^{\mathrm{T}} M^{-1}\mathbf{v} \neq 0$; see Lemma 1.1 of Ding and Zhou [8]. In that case, we know that the unique solution to this system is $\phi = (M + \mathbf{v}\mathbf{v}^{\mathrm{T}})^{-1}\mathbf{v}$.

As we noted in the proof of Theorem 3.8, the steady-state distribution of an $M/G/\infty$ queue is $\mathrm{Pois}(\lambda \mathrm{E}\left[S\right])$ when the arrival rate is λ and service distribution is equivalent to the random variables S. We can now note further that $\lambda \mathrm{E}\left[S\right]$ is the steady-state mean of such a queueing system. The distribution of Q_i is then given by $\mathrm{Pois}(\mathrm{E}\left[Q_i\right])$ for each $i \in \{1, \ldots, k\}$ and thus is equivalent across all sub-queues. □

For added motivation, we now consider the two-dimensional case in the following remark.

Remark If $k = 2$, M and ϕ are scalars, given by

$$M = \frac{\mathrm{E}[S]}{\mathrm{E}\left[S_{2,2}\right] - \mathrm{E}\left[S_{1,2}\right]}, \quad \phi = \frac{\mathrm{E}\left[S_{2,2}\right] - \mathrm{E}\left[S_{1,2}\right]}{\mathrm{E}[S] + \mathrm{E}\left[S_{2,2}\right] - \mathrm{E}\left[S_{1,2}\right]}.$$

In this case, we can note that if $\mathrm{P}(B = 1) = \phi$, then in steady state the distribution of the workload in the system from the easier jobs from all batches will be equivalent to that of the harder jobs. If $\mathrm{P}(B = 1) > \phi$, the number of harder jobs will stochastically dominate the number of easier jobs, and vice versa if $\mathrm{P}(B = 1) < \phi$.

This result implies if we have the ability to choose the probability of batch sizes, we can construct each of the sub-systems which are organized by the order statistics to have the same queue length distribution, thus providing equal work to all of the queues.

4 Conclusion and final remarks

In this paper, we have found parallels between infinite server queues with batch arrivals, sums of scaled Poisson random variables, and Hermite distributions. Moreover, we also connect the stochastic objects to analytic quantities and functions of external interest, such as the harmonic numbers, the exponential integral function, the Euler–Mascheroni constant, and the polylogarithm function. In addition to being interesting in their own right, these connections have helped us to specify exact forms of valuable quantities related to this queueing system, including generating functions for the queue and for the limit of the queue scaled by the batch size. Thus, we have gained both insight into the queue itself and perspective on the model's place in operations research and applied mathematics more broadly.

For this reason, we believe continued work on these fronts is merited. For example, while we have some intuition for the harmonic Hermite distribution discussed in Sect. 2.2, we have less of an understanding of the limiting distribution of the scaled queue in that subsection and extended for random batch sizes in Sect. 3.2. Having more knowledge of what distribution might produce a moment-generating function comprised of exponential integral function could not only teach us about this queueing system, it would also likely be worth studying entirely on its own. Additionally, providing further connections of this distribution back to the harmonic numbers and the associated Hermite distribution would also be of interest, such as in the connection of the limiting moment-generating function to the expected value of a harmonic number evaluated at a Poisson random variable that we remarked in Sect. 2.2. One could also consider control problems for the routing of arrivals to sub-systems, like we discuss for the case of random batch sizes in Sect. 3.3.

For future expansion of this work into other areas of queueing, we can group the main themes of potential further investigations in three categories. First, the extension

of our batch model beyond infinite server queues to multi-server queues, queues with abandonment, and networks of infinite server queues, a la Mandelbaum and Zeltyn [23], Massey and Pender [24], Engblom and Pender [10], Gurvich et al. [15], Pender [30], and Daw and Pender [6]. It would be interesting to explore our limit theorems in these cases to understand the impact of having a finite number of servers. Second, it would also be interesting to explore the impact of the batch arrivals in the context of queues with delayed information as in Pender et al. [32–34]. It would be of interest to know whether or not the batch arrivals would influence the Hopf bifurcations or oscillations that occur in the delayed information queues. Additionally, one could explore findings of this work, like the steady-state distribution representation or the batch scaling, in contexts where there is dependence among the service durations within each batch of arrivals, such as those studied in Pang and Whitt [28], and Falin [11]. Finally, we are particularly interested in studying the impact of batch arrivals in the context of self-exciting arrival processes such as Hawkes processes like in the work of Gao and Zhu [13], Koops et al. [18], and Daw and Pender [5]. We intend to pursue the ideas described here as well as other related concepts in our future work.

Acknowledgements We acknowledge the generous support of the National Science Foundation (NSF) for Jamol Pender's Career Award CMMI # 1751975 and Andrew Daw's NSF Graduate Research Fellowship under Grant DGE-1650441.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables, vol. 55. Courier Corporation, North Chelmsford (1965)
2. Brown, M., Ross, S.M.: Some results for infinite server Poisson queues. J. Appl. Probab. **6**(3), 604–611 (1969)
3. Chiamsiri, S., Leonard, M.S.: A diffusion approximation for bulk queues. Manag. Sci. **27**(10), 1188–1199 (1981)
4. Dattoli, G., Srivastava, H.M.: A note on harmonic numbers, umbral calculus and generating functions. Appl. Math. Lett. **21**(7), 686–693 (2008)
5. Daw, A., Pender, J.: Queues driven by Hawkes processes. Stoch. Syst. **8**(3), 192–229 (2018)
6. Daw, A., Pender, J.: New perspectives on the Erlang-A queue. Adv. Appl. Probab. **51**(1), (2019)
7. de Graaf, W.F., Scheinhardt, W.R.W., Boucherie, R.J.: Shot-noise fluid queues and infinite-server systems with batch arrivals. Perform. Eval. **116**, 143–155 (2017)
8. Ding, J., Zhou, A.: Eigenvalues of rank-one updated matrices with some applications. Appl. Math. Lett. **20**(12), 1223–1226 (2007)
9. Eick, S.G., Massey, W.A., Whitt, W.: The physics of the $M_t/G/\infty$ queue. Oper. Res. **41**(4), 731–742 (1993)
10. Engblom, S., Pender, J.: Approximations for the moments of nonstationary and state dependent birth-death queues. arXiv:1406.6164 (2014)
11. Falin, G.: The $M^k/G/\infty$ batch arrival queue by heterogeneous dependent demands. J. Appl. Probab. **31**(3), 841–846 (1994)
12. Foster, F.G.: Batched queuing processes. Oper. Res. **12**(3), 441–449 (1964)
13. Gao, X., Zhu, L.: Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. Queueing Syst. **90**, 161–206 (2018)
14. Gupta, R.P., Jain, G.C.: A generalized Hermite distribution and its properties. SIAM J. Appl. Math. **27**(2), 359–363 (1974)
15. Gurvich, I., Huang, J., Mandelbaum, A.: Excursion-based universal approximations for the Erlang-A queue in steady-state. Math. Oper. Res. **39**(2), 325–373 (2013)

16. Kemp, C.D., Kemp, A.W.: Some properties of the 'Hermite' distribution. Biometrika **52**(3–4), 381–394 (1965)
17. Knuth, D.E.: Johann Faulhaber and sums of powers. Math. Comput. **61**(203), 277–294 (1993)
18. Koops, D.T., Saxena, M., Boxma, O.J., Mandjes, M.: Infinite-server queues with Hawkes input. J. Appl. Probab. **55**(3), 920–943 (2018)
19. Lee, S.S., Lee, H.W., Yoon, S.H., Chae, K.C.: Batch arrival queue with N-policy and single vacation. Comput. Oper. Res. **22**(2), 173–189 (1995)
20. Liu, L., Templeton, J.G.C.: Autocorrelations in infinite server batch arrival queues. Queueing Syst. **14**(3–4), 313–337 (1993)
21. Lu, Y., Xie, Q., Kliot, G., Geller, A., Larus, J.R., Greenberg, A.: Join-idle-queue: a novel load balancing algorithm for dynamically scalable web services. Perform. Eval. **68**(11), 1056–1071 (2011)
22. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Commun. Stat. Stoch. Models **7**(1), 1–46 (1991)
23. Mandelbaum, A., Zeltyn, S.: Service engineering in action: the Palm/Erlang-A queue, with applications to call centers. In: Spath, D., Fähnrich, K.-P. (eds.) Advances in Services Innovations, pp. 17–45. Springer, Berlin (2007)
24. Massey, W.A., Pender, J.: Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. Queueing Syst. **75**(2–4), 243–277 (2013)
25. Masuyama, H., Takine, T.: Analysis of an infinite-server queue with batch Markovian arrival streams. Queueing Syst. **42**(3), 269–296 (2002)
26. Miller Jr., R.G.: A contribution to the theory of bulk queues. J. R. Stat. Soc. Ser. B (Methodological) **21**(2), 320–337 (1959)
27. Milne, R.K., Westcott, M.: Generalized multivariate Hermite distributions and related point processes. Ann. Inst. Stat. Math. **45**(2), 367–381 (1993)
28. Pang, G., Whitt, W.: Infinite-server queues with batch arrivals and dependent service times. Probab. Eng. Inf. Sci. **26**(2), 197–220 (2012)
29. Pender, J.: Poisson and Gaussian approximations for multi-server queues with batch arrivals and batch abandonment. Technical report, Cornell University, Ithaca, NY (2013)
30. Pender, J.: Gram Charlier expansion for time varying multiserver queues with abandonment. SIAM J. Appl. Math. **74**(4), 1238–1265 (2014)
31. Pender, J., Phung-Duc, T.: A law of large numbers for M/M/c/delayoff-setup queues with nonstationary arrivals. In: International Conference on Analytical and Stochastic Modeling Techniques and Applications, pp. 253–268. Springer, Beilin (2016)
32. Pender, J., Rand, R.H., Wesson, E.: Queues with choice via delay differential equations. Int. J. Bifurc. Chaos **27**(04), 1730016 (2017a)
33. Pender, J., Rand, R.H., Wesson, E.: Strong approximations for queues with customer choice and constant delays (under revision)
34. Pender, J., Rand, R.H., Wesson, E.: An analysis of queues with delayed information and time-varying arrival rates. Nonlinear Dyn. **91**(4), 2411–2427 (2018)
35. Sachs, R.K., Chen, P.-L., Hahnfeldt, P.J., Hlatky, L.R.: DNA damage caused by ionizing radiation. Math. Biosci. **112**(2), 271–303 (1992)
36. Shanbhag, D.N.: On infinite server queues with batch arrivals. J. Appl. Probab. **3**(1), 274–279 (1966)
37. Takagi, H., Takahashi, Y.: Priority queues with batch Poisson arrivals. Oper. Res. Lett. **10**(4), 225–232 (1991)
38. Xie, Q., Pundir, M., Yi, L., Abad, C.L., Campbell, R.H.: Pandas: robust locality-aware scheduling with stochastic delay optimality. IEEE/ACM Trans. Netw. **25**(2), 662–675 (2017)
39. Yekkehkhany, A., Hojjati, A., Hajiesmaili, M.H.: GB-PANDAS: throughput and heavy-traffic optimality analysis for affinity scheduling. ACM SIGMETRICS Perform. Eval. Rev. **45**(2), 2–14 (2018)